Pontifícia Universidade Católica
do Rio de Janeiro

Sarah Hannah Lucius Lacerda de Góes Telles Carvalho Alves

Agrupamento *fuzzy* aplicado à integração de dados multi-ômicos

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós–graduação em Engenharia Elétrica, do Departamento de Engenharia Elétrica da PUC-Rio.

Orientador : Prof.ª Marley Maria Bernardes Rebuzzi Vellasco
Coorientador: Prof.ª Karla Tereza Figueiredo Leite
Coorientador: Mariana Lima Boroni Martins

Rio de Janeiro
Setembro de 2020

Pontifícia Universidade Católica
do Rio de Janeiro

Sarah Hannah Lucius Lacerda de Góes Telles Carvalho Alves

Agrupamento *fuzzy* aplicado à integração de dados multi-ômicos

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós–graduação em Engenharia Elétrica da PUC-Rio. Aprovada pela Comissão Examinadora abaixo:

Prof.ª Marley Maria Bernardes Rebuzzi Vellasco
Orientador
Departamento de Engenharia Elétrica – PUC-Rio

Prof.ª Karla Tereza Figueiredo Leite
UERJ

Mariana Lima Boroni Martins
INCA

Nicolas Carels
Fiocruz

Israel Tojal da Silva
A.C. Camargo Cancer Center

Eduardo Krempser da Silva
Fiocruz

Rio de Janeiro, 30 de Setembro de 2020

Sarah Hannah Lucius Lacerda de Góes Telles Carvalho Alves

Graduada em engenharia metalúrgica pela Universidade Federal do Rio de Janeiro, atua profissionalmente há 9 anos nas áreas de engenharia e gestão.

Ficha Catalográfica

Alves, Sarah Hannah L. L. de G. T. C.

Agrupamento *fuzzy* aplicado à integração de dados multi-ômicos / Sarah Hannah Lucius Lacerda de Góes Telles Carvalho Alves; orientador: Marley Maria Bernardes Rebuzzi Vellasco; coorientadores: Karla Tereza Figueiredo Leite, Mariana Lima Boroni Martins. – 2020.

89 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2020.

Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Inteligência computacional – Teses. 3. Integração de dados multi-ômicos. 4. Agrupamento fuzzy. 5. Seleção de atributos . I. Vellasco, Marley Maria Bernardes Rebuzzi. II. Leite, Karla Tereza Figueiredo. III. Martins, Mariana Lima Boroni. IV. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. V. Título.

CDD: 621.3

À minha família e amigos, pelo incentivo
e apoio.

Agradecimentos

À professora Karla Figueiredo pelos ensinamentos, disponibilidade e preciosas sugestões.

À professora Marley Vellasco, pelo apoio e contribuições no desenvolvimento desta pesquisa.

À pesquisadora Mariana Boroni, pelo incentivo e notáveis sugestões no desenvolvimento deste trabalho.

Às professoras e aos professores da banca pela disponibilidade para participar da Comissão Examinadora.

Aos meus sobrinhos Maria Eduarda e João Gabriel pelo apoio e afeto.

Aos meus irmãos, Marianne Sthéphanie, Tabatha Etienne e Estevam pelo apoio e carinho ainda que distantes.

À meus pais, Estevam e Miriam, pelo apoio contínuo em todos os momentos.

Aos meus amigos, Carlos Eduardo Valinoti, Aline Soares, Gabriel Vignoli, Cristiane Oliveira e Heloísa Melino.

À todos meus colegas do INCA que foram fundamentais no desenvolvimento do trabalho de forma direta ou indireta, Cristovão Lanna, Jéssica Cruz, Daniel Moreira, Caroline Poubel e Nicole Scherer.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

Alves, Sarah Hannah L. L. de G. T. C.; Vellasco, Marley Maria Bernardes Rebuzzi; Leite, Karla Tereza Figueiredo; Martins, Mariana Lima Boroni. **Agrupamento *fuzzy* aplicado à integração de dados multi-ômicos**. Rio de Janeiro, 2020. 89p. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Os avanços nas tecnologias de obtenção de dados multi-ômicos têm disponibilizado diferentes níveis de informação molecular que aumentam progressivamente em volume e variedade. Neste estudo, propõem-se uma metodologia de integração de dados clínicos e multi-ômicos, com o objetivo de identificar subtipos de câncer por agrupamento *fuzzy*, representando assim as gradações entre os diferentes perfis moleculares. Uma melhor caracterização de tumores em subtipos moleculares pode contribuir para uma medicina mais personalizada e assertiva. Os conjuntos de dados ômicos a serem integrados são definidos utilizando um classificador com classe-alvo definida por resultados da literatura. Na sequência, é realizado o pré-processamento dos conjuntos de dados para reduzir a alta dimensionalidade. Os dados selecionados são integrados e em seguida agrupados. Optou-se pelo algoritmo *fuzzy C-means* pela sua capacidade de considerar a possibilidade dos pacientes terem características de diferentes grupos, o que não é possível com métodos clássicos de agrupamento. Como estudo de caso, utilizou-se dados de câncer colorretal (CCR). O CCR tem a quarta maior incidência na população mundial e a terceira maior no Brasil. Foram extraídos dados de metilação, expressão de miRNA e mRNA do portal do projeto *The Cancer Genome Atlas* (TCGA). Observou-se que a adição dos dados de expressão de miRNA e metilação a um classificador de expressão de mRNA da literatura aumentou a acurácia deste em 5 pontos percentuais. Assim, foram usados dados de metilação, expressão de miRNA e mRNA neste trabalho. Os atributos de cada conjunto de dados foram selecionados, obtendo-se redução significativa do número de atributos. A identificação dos grupos foi realizada com o algoritmo *fuzzy C-means*. A variação dos hiperparâmetros deste algoritmo, número de grupos e parâmetro de fuzzificação, permitiu a escolha da combinação de melhor desempenho. A escolha da melhor configuração considerou o efeito da variação dos parâmetros nas características biológicas, em especial na sobrevida global dos pacientes. Observou-se que o agrupamento gerado permitiu identificar que as amostras consideradas não agrupadas têm características biológicas compartilhadas entre grupos de diferentes prognósticos. Os resultados obtidos com a combinação

de dados clínicos e ômicos mostraram-se promissores para melhor predizer o fenótipo.

Palavras-chave

Integração de dados multi-ômicos; Agrupamento fuzzy; Seleção de atributos .

Abstract

Alves, Sarah Hannah L. L. de G. T. C.; Vellasco, Marley Maria Bernardes Rebuzzi (Advisor); Leite, Karla Tereza Figueiredo (Co-Advisor); Martins, Mariana Lima Boroni (Co-Advisor). **Fuzzy clustering applied to multi-omics data**. Rio de Janeiro, 2020. 89p. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

The advances in technologies for obtaining multi-omic data provide different levels of molecular information that progressively increase in volume and variety. This study proposes a methodology for integrating clinical and multi-omic data, which aim is the identification of cancer subtypes using fuzzy clustering algorithm, representing the different degrees between molecular profiles. A better characterization of tumors in molecular subtypes can contribute to a more personalized and assertive medicine. A classifier that uses a target class from literature results indicates which omic data sets should be integrated. Next, data sets are pre-processed to reduce high dimensionality. The selected data is integrated and then clustered. The fuzzy C-means algorithm was chosen due to its ability to consider the shared patients characteristics between different groups. As a case study, colorectal cancer (CRC) data were used. CCR has the fourth highest incidence in the world population and the third highest in Brazil. Methylation, miRNA and mRNA expression data were extracted from The Cancer Genome Atlas (TCGA) project portal. It was observed that the addition of miRNA expression and methylation data to a literature mRNA expression classifier increased its accuracy by 5 percentage points. Therefore, methylation, miRNA and mRNA expression data were used in this work. The attributes of each data set were pre-selected, obtaining a significant reduction in the number of attributes. Groups were identified using the fuzzy C-means algorithm. The variation of the hyperparameters of this algorithm, number of groups and membership degree, indicated the best performance combination. This choice considered the effect of parameters variation on biological characteristics, especially on the overall survival of patients. Clusters showed that patients considered not grouped had biological characteristics shared between groups of different prognoses. The combination of clinical and omic data to better predict the phenotype revealed promissing results.

Keywords

Multi-omic data integration; Fuzzy clustering; Feature selectio.

Sumário

Lista de Abreviaturas

TCGA	–	do inglês, *The Cancer Genome Atlas*
CNVs	–	Variação do número de cópias (do inglês, *Copy number variation*)
HTS	–	Tecnologias de sequenciamento de larga escala (do inglês, *High Throughput Sequencing Technologies*)
CMS	–	do inglês *Consensus Molecular Subtypes*
MCFS	–	Seleção de atributos multi-cluster (do inglês, *Multi-Cluster Feature Selection*)
FCM	–	do inglês, *Fuzzy C-Means*
FPC	–	Coeficiente de partição *fuzzy* (do inglês, *fuzzy partition coefficient*)
OOB	–	do inglês, *out-of-bag*
MCIA	–	do inglês, *Multiple Co-Inertia Analysis*
MFA	–	Análise de múltiplos fatores (do inglês, *Multiple factor analysis*)
CPT	–	Carcinoma papilífero da tireóide
PCA	–	Análise de componentes principais (do inglês, *Principal Component Analysis*)
CCR	–	Câncer colorretal
EVR	–	Taxa de variância explicada (do inglês, *explained variance ratio*)

"Se queres prever o futuro, estuda o passado."

Confúcio, *Os Anacletos.*

1
Introdução

O avanço das tecnologias de larga escala, e a consequente relativa diminuição de preços na geração de dados, promoveram o acesso a uma grande quantidade de dados de um único paciente, como dados ômicos de alto rendimento, entre outros, associados a registros clínicos (por exemplo, idade, sexo, histórico, patologias e terapêutica) [1].

Dados ômicos são conjuntos de informações sobre diferentes níveis de regulação celular, oriundos da genômica (conjunto de informações sobre o DNA), epigenômica (conjunto de informações sobre as modificações epigenéticas como a metilação do DNA, por exemplo), transcriptômica (conjunto dos transcritos de genes expressos, tais como mRNAs e miRNAs), proteômica (conjunto de proteínas) e metabolômica (conjunto de metabólitos) [2].

A geração de dados ômicos requer esforços colaborativos dentro de um grande consórcio, como por exemplo o *The Cancer Genome Atlas* (TCGA) [3]. Este repositório mapeia dados multidimensionais e disponibiliza ferramentas para examinar as principais mudanças genômicas associadas a diferentes tipos tumorais nos fenótipos analisados [4].

O estudo de dados ômicos considera que as informações de um organismo estão dispostas no seu respectivo genoma, e que o funcionamento deste baseia-se no dogma central da biologia molecular, o qual caracteriza o fluxo da informação genética. Este princípio, enunciado em 1952 por Francis Crick, descreve como o processo de produção de proteínas necessárias ao funcionamento das células se desenvolve. Este processo é composto de duas etapas, transcrição do DNA em RNA, e tradução do RNA em proteína [5].

1.1
Motivação

A integração de diferentes conjuntos de dados e métodos pode oferecer suporte a interpretações mais significativas das relações genótipo-fenótipo do que análises usando apenas um único conjunto de dados, apesar destes serem bastante informativos individualmente [2]. Os conjuntos ômicos analisados devem refletir a complexidade do fenótipo estudado.

A integração de diferentes ômicas favorece portanto a complementaridade mútua, causalidade e heterogeneidade de respostas. No entanto, isto aumenta quase exponencialmente o volume, variedade, redundância e complexidade das informações. Os conjuntos de dados resultantes diferem em seus tipos e origens, podendo seguir diferentes distribuições estatísticas, ter diferentes níveis de imprecisões e incertezas [1].

A seleção de um conjunto de atributos conciso facilita a implementação de metodologias que relacionam o genótipo e o fenótipo. Uma metodologia de seleção de atributos que seja mais eficaz na eliminação de atributos redundantes, irrelevantes e pouco informativos, torna mais acessível a identificação das características moleculares do paciente, uma vez que é necessária uma quantidade menor de atributos para relacionar o fenótipo com base nos padrões moleculares do paciente.

A correta identificação do perfil molecular do paciente é importante porque a sobrevida, por exemplo, dos pacientes de câncer varia muito, mesmo para um mesmo tipo de câncer. Este comportamento é decorrente da diferença entre os subtipos moleculares de cada câncer. Por isso, a identificação destes subtipos moleculares é uma oportunidade de se melhor caracterizar o prognóstico e de proporcionar um tratamento personalizado [6].

Na identificação dos subtipos moleculares é importante utilizar uma técnica que permita informar sobre possíveis gradações e tendências de evolução das características dos pacientes, possibilitando identificar aqueles pacientes que possuem características de diferentes estágios da doença. Por exemplo, um paciente que tenha características moleculares tanto de um grupo com bom prognóstico, com estadiamento menos grave, como de outro grupo de prognóstico ruim, de estadiamento mais avançado, será considerado como pertencente a apenas um destes grupos ao utilizar-se técnicas clássicas de agrupamento, incorrendo numa aproximação inexata da condição real do paciente.

Deste modo, se faz necessária a escolha assertiva dos seguintes itens para uma adequada análise de dados multiômicos: (i) conjuntos de dados ômicos a serem utilizados; (ii) atributos destes conjuntos que devem ser selecionados; (iii) técnica de identificação de subtipos moleculares que consiga representar a gradação das informações do perfil molecular dos pacientes.

As técnicas de aprendizado de máquina têm sido amplamente usadas na integração de dados multiômicos. Estas são algoritmos computacionais que proporcionam a automatização no processo de reconhecimento de padrões complexos com base em dados [7]. O objetivo dos métodos de aprendizado de máquina é permitir que um algoritmo aprenda a identificar um padrão a partir de um conjunto de dados e use esse conhecimento para fazer previsões

ou tomar decisões para novos dados [2].

A medicina de precisão, um campo emergente da medicina, é um importante conjunto de práticas médicas com perspectivas de customização dos cuidados de saúde, tornando as decisões, práticas e tratamentos médicos personalizados para cada paciente [8]. O uso de dados multiômicos, como biomarcadores genômicos, tem desempenhado um papel importante na medicina de precisão, por exemplo, na área da oncologia. Neste tipo de abordagem, os pacientes são divididos em grupos de acordo com a variabilidade genética e outros biomarcadores para que os medicamentos possam ser adaptados para os pacientes com características genéticas semelhantes ou relacionadas [9]. Deste modo, as abordagens multiômicas identificam moléculas e genes de relevância para a condição analisada, permitindo a interação com sistemas biológicos em nível molecular e com alto nível de precisão.

A combinação de técnicas de aprendizado de máquina e dados multiômicos através da medicina de precisão é, portanto, um grande desafio para o melhor uso desses dados para necessidades específicas.

A utilização de dados multiômicos com técnicas apropriadas de aprendizado de máquina fornece uma oportunidade sem precedentes para compreender um sistema biológico complexo de diferentes ângulos e níveis, permitindo observar, por exemplo, interações genótipo-fenótipo em estudos de câncer, e tornar mais precisas previsões baseadas em dados, como a previsão de resposta à medicação, por exemplo.

1.2 Objetivos

O objetivo principal deste trabalho é o desenvolvimento de uma metodologia, baseada em agrupamentos *fuzzy*, que permita utilizar os dados moleculares de pacientes para identificar subtipos moleculares baseados na relação entre o genótipo e o fenótipo do paciente. Isto é possível a partir da utilização de dados de múltiplas ômicas, métodos de seleção de variáveis e técnicas de agrupamento que indiquem o grau de pertinência das amostras não classificadas a cada um dos grupos obtidos.

Este trabalho também busca:

- Avaliar a contribuição individual de diferentes ômicas para a classificação molecular através da adição de informações moleculares e da comparação de desempenho de classificadores da literatura;
- Definir quais os métodos de seleção de atributos são mais adequados à análise de dados multiômicos;

• Elaborar uma metodologia capaz de identificar pacientes que tenham características de mais de um subtipos molecular;

• Desenvolver abordagem que diminua o compartilhamento de características dos perfis de subtipos moleculares adicionando dados além dos já utilizados;

• Relacionar os grupos deste trabalho com grupos já definidos na literatura e caracterizar os resultados com base em características biológicas.

1.3 Descrição da dissertação

Esta dissertação foi elaborada seguindo-se as seguintes etapas:

• Pesquisa sobre a área de integração de dados multiômicos, tratamento de dados de elevada dimensionalidade e técnicas de aprendizado de máquina para a identificação de grupos *ab initio*;

• Análise pormenorizada de trabalhos que envolviam o estudo de dados de diferentes ômicas na área de oncologia, em específico, de câncer colorretal, e técnicas de seleção de atributos que viabilizassem a integração destes dados;

• Definição dos métodos a serem utilizados, com respectivos parâmetros, e proposição da abordagem a ser desenvolvida;

• Avaliação da metodologia proposta com relação a resultados anteriores da literatura.

Na primeira etapa do trabalho foram estudadas as diferentes ômicas disponíveis e as diversas possibilidades de integração destes dados, considerando suas vantagens e desvantagens com relação à identificação de subtipos moleculares.

Também foram realizadas pesquisas sobre os métodos de identificação de subtipos moleculares, como estas técnicas auxiliam na definição do fenótipo e suas respectivas aplicações. Foram observados os principais desafios na área e oportunidades de melhoria dos resultados que caracterizam o fenótipo da doença. A partir disso, buscou-se elaborar a metodologia de análise de dados adequada as características de dados ômicos.

Em seguida, na etapa de estudo de caso, foi escolhido o fenótipo a ser estudado com base nas oportunidades de se ampliar o conhecimento sobre o fenômeno analisado. Foram priorizados dados disponíveis da doença e técnicas de aprendizado de máquina consolidadas. Os parâmetros do modelo foram testados e avaliados com relação a métricas de desempenho adequadas. Os resultados foram caracterizados com relação as características biológicas associadas à doença.

1.4 Organização do Trabalho

Este trabalho está organizado em mais cinco capítulos a seguir:

• Capítulo 2 – São apresentados conceitos fundamentais de biologia molecular, as técnicas de extração e processamento de dados biológicos e a disponibilidade dos mesmos.

• Capítulo 3 - São analisados os métodos de seleção de atributos e demais métodos de pré-processamento de dados. Na sequência, são revistos métodos de identificação de classes através de técnicas de aprendizado de máquina e as estratégias de integração de dados multiômicos. O capítulo é encerrado analisando os trabalhos anteriores que utilizaram técnicas de aprendizado de máquina para integrar dados ômicos.

• Capítulo 4 - Apresenta-se o modelo *fuzzy* de integração de dados ômicos, iniciando-se pela técnicas que podem auxiliar na escolha das variáveis a serem estudadas. Em seguida são destacadas as vantagens e desvantagens das diferentes técnicas de agrupamento utilizadas para identificar os subtipos dos dados. Posteriormente detalha-se a etapa de avaliação do modelo. Em cada uma das etapas são destacados os algoritmos utilizados.

• Capítulo 5 - Descreve-se o estudo de caso de aplicação do modelo proposto. É escolhido um fenótipo de interesse, isto é, que possibilite novas descobertas da sua relação com seu genótipo, para se analisar. São definidos os parâmetros globais e apresentados os resultados obtidos com as abordagens propostas neste trabalho. São comparados os resultados obtidos neste trabalho com os de métodos que utilizaram os conjuntos de dados do mesmo fenótipo e da mesma base de dados. Métricas de desempenho são aplicadas para comparar as abordagens propostas e seus parâmetros. O modelo é também avaliado com relação a sua capacidade de relacionar o perfil molecular dos pacientes com o fenótipo selecionado, isto é, com as características biológicas do paciente.

• Capítulo 6 - Neste capítulo são abordadas as principais conclusões e propostas de trabalho futuros.

2
Fundamentos da biologia molecular

2.1
Dados ômicos

As antigas técnicas de sequenciamento do genoma eram baseadas em um conjunto restrito de genes e poucas amostras, enquanto as novas técnicas de sequenciamento possibilitam a coleta de informações de milhares de genes e de maiores quantidades de amostras [10]. A coleta de informações do genoma nesta magnitude é denominada por genômica, caracterizada pelo seu sufixo '-ômica' [11], referente à avaliação em larga escala de um conjunto de moléculas [12]. Há portanto uma maior disponibilidade de informações sobre o genoma, o que possibilita o estudo e compreensão das complexas interações biológicas em seus diferentes níveis de forma holística. Este campo de estudo é denominado por integração multiômica de dados e tem como objetivo relacionar o comportamento genotípico com o fenotípico [13].

Todas as informações de um organismo são dispostas no seu respectivo genoma, que é armazenado na forma de DNA. O genoma humano contém aproximadamente 30 mil genes. Para realizar seu objetivo de armazenar informação, o DNA deve também expressar as informações contidas nele. Estas informações orientam a replicação de moléculas, determinando as propriedades das células, e devem ocorrer através da transcrição do DNA em RNA, denominada expressão de mRNA, e na subsequente tradução do RNA em proteína, expressão de microRNA(miRNA). Contudo, apenas aproximadamente 21 mil genes do DNA são codificantes, isto é, realizam a conversão do RNA em proteína. Os demais genes têm como função a regulação da expressão de mRNA e constituem uma grande diversidade de genes [5].

O controle da expressão dos genes (codificados pelo DNA) em uma célula pode ser regulado de diferentes formas. Dentre eles, temos a regulação epigenética que, dentre outras formas, se dá através da adição de grupos metila ao DNA. A adição de grupos metila a regiões promotoras de genes está geralmente associada à inativação (reversível) dos mesmos. Esta alteração é conhecida como metilação. Quando há uma variação, natural ou induzida, na composição de nucleotídeos de um cromossomo, ou seja, quando a alteração

é hereditária, devido à ação de um agente mutagênico, considera-se que esta é uma mutação, uma alteração genética que pode levar à inativação do gene, ou à produção de uma proteína com perda ou ganho de função [5].

2.2 Técnicas para obtenção de dados ômicos

Em geral, as análises de bioinformática consistem em alinhar as leituras de sequências de nucleotídeos a um genoma de referência, reunir as leituras por genes e detectar diferenças na expressão dos transcritos entre os grupos [14]. A ordem da sequência de DNA e sua variações genéticas ditam os processos de desenvolvimento humano, identificam cada pessoa de forma única e codificam nossa suscetibilidade a doenças [14].

As tecnologias de sequenciamento de larga escala (HTS, do inglês, High Throughput Sequencing Technologies, HTS) têm possibilitado o acesso a diferentes dados ômicos. Estas tecnologias permitem também a compreensão mais ampla sobre as assinaturas genômicas e transcriptômicas de células acometidas por várias doenças e em diferentes estágios de desenvolvimento [14]. Há diferentes técnicas de HTS, cada uma com um objetivo específico. Por exemplo, o sequenciamento de RNA (RNA-seq) pode ser utilizado para analisar como o transcriptoma muda; o sequenciamento completo do exoma (conjunto de exons que codificam os genes que determinam a produção de proteínas) pode ser usado para identificar novas variantes e mutações; o sequenciamento das regiões da cromatina imunoprecipitada (ChIP-seq) e o sequenciamento das sequências metiladas (Metil-seq) podem indicar alterações epigenéticas; e, por fim, o sequenciamento dos RNAs traduzidos pelos ribossomos (Ribo-seq) pode indicar quais transcritos de mRNA estão sendo traduzidos ativamente [14].

No sequenciamento de DNA de última geração produzido, por exemplo, pelas plataformas Illumina, o DNA é primeiramente fragmentado por enzimas, ou sonicação, em partes menores. As extremidades desses fragmentos são reparadas e adaptadores específicos são ligados as extremidades dos fragmentos, permitindo que ocorra a hibridização desses fragmentos em uma lâmina. Estes equipamentos realizam a amplificação em ponte para aumentar o sinal dos fragmentos. Uma fita de DNA é removida e cada grupo é exposto a nucleotídeos marcados com fluorescência. A imagem da lâmina é armazenada para cada ciclo e um computador verifica qual nucleotídeo foi incorporado a cada coordenada de cada agrupamento de fita de DNA. O marcador fluorescente é clivado e as fitas molde são expostas novamente a um novo grupo de nucleotídeos e o processo se repete.

Cada ciclo resulta na leitura de uma base em cada fragmento de DNA.

Findado o sequenciamento dos fragmentos de DNA, essas leituras são então alinhadas a um genoma de referência e permitem a detecção de alterações moleculares como a presença de mutações pontuais ou inserções e deleções (INDELs), variantes estruturais, assim como variação do número de cópias de genes (duplicação ou deleção).

Essa tecnologia também permite sequenciar as moléculas de RNA (RNA-seq) de uma célula, e através da quantificação da quantidade de transcritos, é possível inferir diferenças na expressão de mRNA entre pacientes de condições diversas [15].

O RNA-seq é utilizado principalmente na análise do estado atual de uma célula ou tecido e dos possíveis efeitos no transcriptoma decorrentes de uma doença ou condições de tratamento [14].

Inúmeras variações de preparações de biblioteca de RNA-seq foram desenvolvidas, cada uma com seus benefícios e limitações em termos de custos relativos e requisitos de utilização. As principais diferenças nas várias preparações de biblioteca são os métodos de purificação e isolamento do RNA de interesse (mRNA, uRNA, transcrições completas, etc.). O isolamento do mRNA poliadenilado e, em seguida, a transcrição reversa é o método convencional de preparação de uma amostra de RNA-seq [14].

O nível de confiança, isto é, o quão precisas são as informações do sequenciamento, é dado pela "profundidade" deste. Esta profundidade é definida como a média de observações de cada nucleotídeo no genoma [16]. A profundidade de leitura é importante para interpretar as variações estruturais, uma vez que, para um dado intervalo, um aumento na quantidade de leituras em uma determinada profundidade de leitura pode indicar um aumento no número de cópias, enquanto uma diminuição na quantidade de leituras em uma determinada profundidade de leitura pode indicar deleções. Em geral, conforme o nível de profundidade de leitura aumenta, o sequenciamento de informações torna-se mais confiável [17].

Além disso, dado que em cada corrida do sequenciador apenas um número limitado de fragmentos de sequência pode ser lido, pode haver melhor relação custo-benefício em analisar apenas a sequência do material genético que é transcrito em mRNA (exons), dependendo das necessidades experimentais. Erros no mapeamento de leituras curtas também podem ocorrer devido as ambiguidades de regiões genômicas altamente repetitivas e de famílias de genes homólogos [18]. Na maioria dos casos, aumentar a profundidade de leitura é mais caro e métodos alternativos, direcionados e mais econômicos, podem ser preferidos em relação ao HTS. Um exemplo de métodos alternativos são os chips de hibridização personalizados.

Esses chips são microarranjos projetados especialmente para avaliar sequências de DNA. Estes microarranjos têm possibilitado a medição dos níveis de metilação do DNA de partes específicas do genoma, sendo cada uma destas representada por sondas específicas no microarranjo. Uma das principais vantagens dessa abordagem é o menor custo por amostra em comparação com o sequenciamento do genoma inteiro utilizando as plataformas de sequenciamento descritas anteriormente.

Além disso, microarranjos podem ser processados com a mesma infraestrutura para diferentes técnicas, por exemplo para genotipagem e conversão por bissulfito. Os procedimentos experimentais são um pouco mais fáceis e rápidos de realizar do que a preparação de bibliotecas de sequenciamento de DNA. As desvantagens são a cobertura genômica limitada, resolução limitada, e o alto custo de instalação para projetar microarranjos personalizados [19].

O processamento bioinformático destes dados compreende processamento de imagem e normalização de dados como etapas principais [20].

No caso dados de microarranjo de metilação, o resultado da normalização é uma tabela de valores β (e, opcionalmente, valores M) que serve como ponto de partida para análises posteriores. Os valores β são conceitualmente equivalentes aos níveis absolutos de metilação do DNA calculados a partir de dados de sequenciamento de bissulfito, enquanto os valores M são valores β logisticamente transformados e exibem uma distribuição que é mais adequada para uso com alguns testes estatísticos comuns.

Os dados de microarranjo tem alguns vieses como, por exemplo, ligações não-específicas entre os fragmentos, que podem ser contornados adequando-se o número de leituras (sondas). Apesar do uso de algoritmos de normalização para reduzir esses vieses técnicos, ainda há várias fontes de enviesamento dos dados [19].

Deste modo, apesar dos microarranjos terem trazido grandes avanços ao estudo da transcriptômica e terem se mostrado úteis na determinação dos perfis de expressão de mRNA, a técnica de RNA-seq é mais sensível e mais confiável. Esta fornece níveis de quantidade absolutos, e não é afetada por vieses de sequência do chip, disponibilizando informações adicionais sobre os níveis de expressão de mRNA e variantes de junção de *splice* [21].

2.3 Normalização de dados ômicos

O resultado do sequenciamento do RNA para estimar a expressão deve ser normalizado em relação à contagem de leituras por duas razões principais. Uma destas é que a fragmentação do DNA realizada para a construção da sequência

do genoma induz uma maior quantidade de leituras em transcritos longos quando comparados a transcritos pequenos presentes na mesma quantidade na amostra (normalização intra amostra). O outro motivo é a variabilidade no número de leituras produzidas em cada amostragem, induzindo variações no número de fragmentos mapeados entre as amostras (normalização inter amostras).

Para atender a ambas demandas, pode-se normalizar a expressão em leituras por kilobase de transcrito por milhão de leituras mapeadas (RPKM), normalizando o transcrito tanto pelo seu comprimento quanto pelo número de leituras por amostra.

Quando os dados usados são gerados por sequenciamento *paired-end*, a normalização análoga é a fragmentos por kilobase de transcritos por milhão de leituras mapeadas (FPKM) [22]. Com relação à normalização de microRNA-seq, atualmente, a maioria dos datasets são normalizados em leituras por milhão de leituras de microRNA(miRNA). Esta normalização, denominada RPM, é realizada apenas inter amostra pois os miRNAs apresentam como característica o pequeno tamanho de cerca de 21pb.

Esta abordagem é semelhante à FPKM usada para leituras de expressão de mRNA; no entanto, não há a necessidade de se normalizar com relação ao tamanho para miRNAs. Como o número de leituras de miRNAs também pode variar substancialmente entre projetos, observa-se o uso de contagens brutas para os miRNAs. A desvantagem desta abordagem é que esta depende do valor normalizado, de forma que qualquer alteração na contagem das leituras irá ajustar todos os demais valores de miRNA sem que o valor absoluto da expressão de mRNA tenha de fato sido alterado. Métodos como microarranjo não têm os mesmos desafios, uma vez que este se utilizam de observações independentes [23].

2.4 Análise de expressão diferencial

Em análises que buscam comparar resultados de sequenciamento de alto rendimento, é importante considerar a contagem de leituras por gene realizada no sequenciamento. A consideração do número de contagens de leituras auxilia na avaliação de diferenças quantitativas entre as condições experimentais, além das especificidades dos dados de contagem, como a não normalidade e a dependência do desvio padrão.

Um desafio central é, em geral, o pequeno número de amostras em experimentos HTS, o que indica a necessidade de se tratar os dados com uma abordagem estatística que minimize estes efeitos. Os métodos diferenciais que

tratam cada gene separadamente perdem sua confiabilidade nesse caso, devido à alta incerteza das estimativas de variação dentro de cada grupo.

Em HTS, essa limitação pode ser superada combinando informações entre genes e explorando suposições sobre a similaridade das variâncias de diferentes genes medidos no mesmo experimento.

O método de análise diferencial de dados de contagem DESeq2 realiza este tipo de análise através da estimativa de normalização das dispersões e *fold change* (número de vezes que a expressão se altera em relação ao grupo controle), para melhorar a estabilidade e interpretabilidade das estimativas. Isso permite uma análise mais quantitativa focada na magnitude da expressão diferencial ao invés apenas da observação da diferença da expressão de qualquer ordem [24].

2.5 Análise de sobrevida

A análise de sobrevida pelo método de Kaplan-Meier é uma metodologia bastante utilizada para verificar a variação do prognóstico dos pacientes que compartilham características semelhantes, representando assim um grupo.

O tempo de sobrevida é o tempo até o evento, incluindo morte e início da doença, por exemplo. A principal característica da curva de sobrevida é que esta contém dados "censurados", nos quais o tempo até o evento não pode ser completamente observado, como por exemplo no caso de pacientes em que foi perdido o contato durante o período observado. Nesse caso, apenas o tempo de censura é observado.

Esta abordagem tem como premissa que o tempo de sobrevida dos indivíduos censoreados é no mínimo maior que o tempo censoreado. Muitos métodos estatísticos tradicionais têm sido usados com eficácia para analisar dados de sobrevida com observações censuradas [25].

3
Análise computacional de dados biológicos

3.1
Redução de dimensionalidade

A alta dimensionalidade dos dados biológicos é um problema crítico. Este problem consiste na elevada esparsidade dos dados em um espaço de alta dimensão, o que afeta adversamente algoritmos projetados para espaços de baixa dimensão [26].

Dados com esta característica também podem ter a interpretabilidade dificultada e aumentar significativamente os requisitos de armazenamento de memória e custos computacionais para sua análise.

A redução da dimensionalidade é uma estratégia para se reduzir os problemas descritos anteriormente. Esta pode ser realizada através da extração ou seleção de atributos. Tanto a extração quanto a seleção de atributos têm vantagens de melhorar o desempenho do aprendizado, aumentar a eficiência computacional, diminuir o armazenamento de memória e construir melhores modelos de generalização. Portanto, ambas são consideradas técnicas eficazes de redução de dimensionalidade [27].

3.1.1
Extração de atributos

A extração de atributos projeta os atributos originais de alta dimensão para um novo espaço de atributos, de baixa dimensionalidade. O novo espaço de atributos é geralmente uma combinação linear ou não linear dos atributos originais [27].

Em aplicações cujos dados de entrada não necessitam de atributos compreensíveis, estas técnicas podem ser amplamente utilizadas, como no caso da visualização de dados em uma dimensão reduzida da original.

Um exemplo destas técnicas é a análise de componentes principais (PCA, do inglês *Principal Component Analysis*). Esta consiste em um mapeamento linear que não demanda a determinação da classe-alvo. O PCA preserva a quantidade máxima de variância dos dados originais. O objetivo deste método é identificar as Componentes Principais (PCs) dos dados analisados.

A primeira componente contém a maior quantidade de variação dos dados originais, enquanto a segunda PC, ortogonal à primeira, representa a segunda maior variação, e assim por diante. As primeiras k PCs mantêm a maior variação dos dados originais e reduzem a dimensão original dos dados para k dimensões. Indica-se usar PCs que contenham uma quantidade relevante de informação (variância) para realizar a referida visualização de dados em espaço de menor dimensão.

Portanto, uma vez que a extração de atributos cria um conjunto de novos atributos, isto é, não preserva o significado físico destes, o uso destas técnicas é limitada.

3.1.2 Seleção de atributos

A seleção de atributos escolhe diretamente um subconjunto de atributos relevantes para a construção do modelo [28]. Isto é importante porque muitos atributos são irrelevantes, ou redundantes, seja por não contribuírem para o processamento dos dados, ou por conter as mesmas informações [29]. As técnicas de seleção de atributos têm como objetivo portanto a redução da dimensionalidade dos dados e a melhoria do desempenho dos preditores ao construir modelos simples e mais compreensíveis [28].

Estes métodos podem ser avaliados tanto pela eficiência quanto pela eficácia. A eficiência é composta do tempo requerido para se definir o subconjunto de atributos, enquanto a eficácia é relacionada à acurácia do subconjunto de atributos selecionado.

Estas técnicas, enquanto estratégia de pré-processamento de dados, têm se mostrado eficientes e efetivas no preparo de dados (em especial para os de alta dimensionalidade) para vários problemas de mineração de dados e aprendizado de máquina [28]. A maior disponibilidade de dados de alta dimensionalidade tem trazido importantes desafios para a seleção de atributos [27].

Ao manter alguns dos atributos originais, a seleção de atributos mantém os significados físicos desses atributos e oferece aos modelos melhor legibilidade e interpretabilidade. Portanto, a seleção de atributos é frequentemente preferida na área de bioinformática [27].

Como observado na Seção 2.1, o estudo de dados ômicos envolve o tratamento e modelagem de grandes quantidades de atributos (genes) para cada amostra. Entretanto, é importante determinar quais os atributos de maior relevância para o modelo, evitando que este tenha problemas em sua generalização.

Espera-se também que os modelos desenvolvidos possam sempre atender ao critério da parcimônia com uma quantidade de genes adequada, a fim de se facilitar a utilização do modelo. Um modelo mais simples, com menos atributos, demanda menos investimento para a sua utilização futura do que um modelo com número elevado de genes [30].

Como resultado destas técnicas, em geral, obtém-se um ranking de importância de cada atributo, ou o conjunto de atributos com maior importância, a partir de uma quantidade de atributos previamente definida [30].

Os algoritmos de seleção de atributos podem ser classificados quanto à disponibilidade de informação referente aos valores a serem preditos pelo modelo, isto é, quanto à disponibilidade de informação da classe-alvo, quanto à estratégia de busca, ou ainda, com relação ao critério que define a relevância dos atributos, como mostrado na Figura 3.1.

Quando os dados a serem analisados por métodos de seleção de atributos apresentam informações referentes aos valores a serem preditos pelo modelo, isto é, quando os dados têm rótulos, é possível utilizar um modelo supervisionado, em que a relevância dos atributos é determinada pela avaliação destes em relação a sua classe correspondente.

No caso da informação referente aos valores preditos não estar disponível, pode-se utilizar um modelo não-supervisionado. Neste caso, objetiva-se avaliar a importância dos atributos através da identificação de estruturas e relações desconhecidas entre os dados, ou seja, baseia-se na variância e na separabilidade dos dados.

Havendo apenas uma parte dos dados com rótulos, pode-se utilizar técnicas semi-supervisionadas, que aproveitam as vantagens de ambas as abordagens [30].

Com relação à estratégia de busca, os métodos de seleção de dados podem ser distribuídos em 4 grupos: *wrapper*, filtro, *embedded* e métodos híbridos. No entanto, este último representa apenas a composição dos demais métodos.

Métodos de filtragem são implementados como uma etapa de pré-processamento para a escolha do subconjunto de atributos adequados, sendo independente do algoritmo de aprendizado. Métodos de filtro ajudam a reduzir o espaço de busca e são mais adequados quando há elevado número de atributos. Sua complexidade computacional é baixa, contudo, a acurácia destes algoritmos nem sempre é satisfatória [28].

Métodos *wrapper* consideram os subconjuntos de atributos de acordo com a sua utilidade para um determinado preditor. O método realiza uma busca por um subconjunto adequado de dados usando o próprio algoritmo de aprendizado como parte de uma função de avaliação, ou, validação. A etapa

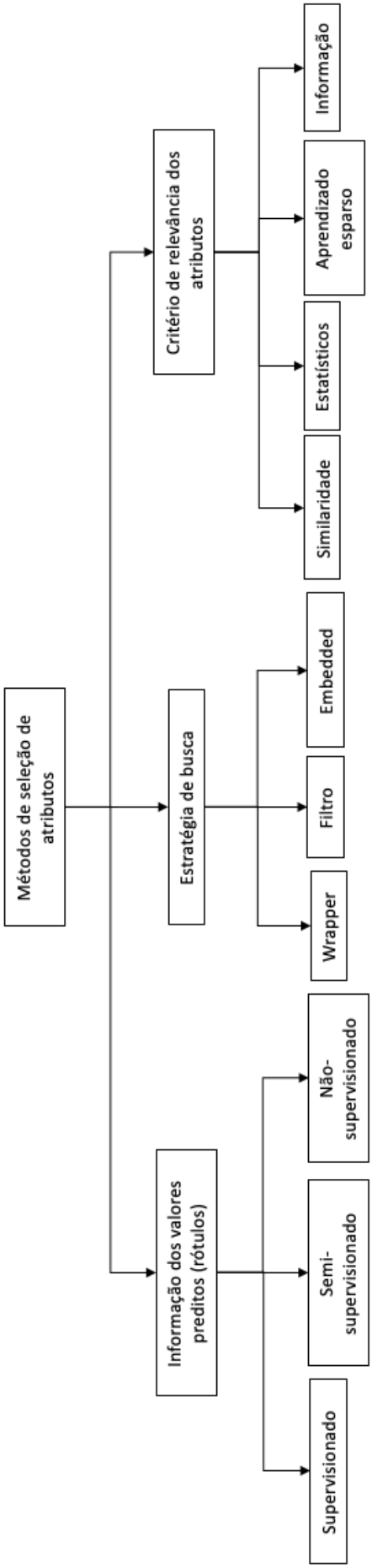


Figura 3.1: Classificação de métodos de seleção de atributos, adaptado de Miao[31].

seguinte pode ser, por exemplo, uma regressão linear. Este método apresenta alta acurácia. Contudo, exige alta capacidade computacional.

Nos métodos *embedded*, a seleção de atributos é parte do aprendizado do algoritmo. Por isso, estes métodos são, em geral, direcionados para um método de aprendizado de máquina específico. Métodos de aprendizado de máquina como árvores de decisão ou redes neurais artificiais são exemplos de abordagens *embedded*. Estes métodos são menos intensivos computacionalmente do que *wrapper*. Contudo, métodos *embedded* têm sua utilização limitada ao algoritmo de aprendizado.

Os métodos de seleção de atributos também podem ser categorizados de acordo com o critério utilizado para definir a relevância de cada atributo entre: baseados em similaridade, informação, aprendizado esparso e métodos estatísticos [27].

Os algoritmos baseados em informações exploram diferentes critérios de filtro heurísticos para medir a importância dos atributos. Muitos critérios de informação são projetados de forma específica para maximizar a relevância e minimizar a redundância dos atributos. A relevância de um atributo geralmente é medida por sua correlação com os rótulos de classe. Além disso, a maioria dos conceitos teóricos da informação só podem ser aplicados a variáveis de saída discretas. Portanto, os algoritmos de seleção de atributos nesta família só podem funcionar com dados discretos. Para valores de atributos contínuos, algumas técnicas de discretização de dados são necessárias a priori [27].

Os algoritmos baseados em estatística dependem de medidas estatísticas em vez de algoritmos de aprendizagem para avaliar a relevância dos atributos. A maioria destes métodos é baseada em filtros, analisando os atributos individualmente. Portanto, a redundância é elevada. Contudo, estes algoritmos são simples e diretos e os custos computacionais costumam ser muito baixos. Por isso, estes são frequentemente usados como uma etapa de pré-processamento antes de se aplicar outros algoritmos sofisticados de seleção de atributos. Adicionalmente, do mesmo modo que os algoritmos baseados em informação, a maioria dos algoritmos desta família pode funcionar apenas com dados discretos. Deste modo, técnicas convencionais de discretização de dados são necessárias para pré-processar variáveis de entrada numéricas e contínuas [27].

Os algoritmos baseados em similaridade e aprendizado esparso são os algoritmos que serão estudados nesse trabalho, por isso serão abordados em maior detalhe nas seções seguintes.

3.1.2.1 Algoritmos baseados em similiaridade

Os métodos baseados em similaridade são métodos que avaliam a habilidade dos atributos em manter a similaridade dos dados. Uma forma de se avaliar esta semelhança é através da utilização de diferentes métricas de distância. Quando a informação do rótulo esta disponível, a distância entre os dados pode ser atribuída de acordo com a informação fornecida pelo rótulo [27].

O algoritmo ReliefF se baseia em similaridade para avaliar a qualidade dos atributos. Este método é a extensão do algoritmo Relief para classificação de múltiplos rótulos [31]. O ReliefF se utiliza de uma função de avaliação baseada em distância que pondera cada característica com base em sua relevância (correlação) em relação à classe-alvo [29]. Este é capaz de estimar corretamente a qualidade dos atributos em problemas de classificação com fortes dependências entre os atributos [32]. O algoritmo ReliefF é baseado em amostragem aleatória, por isso fornece uma análise global da qualidade dos atributos analisados. Este algoritmo também considera as interações entre os atributos próximos, portanto, é apropriado para problemas em que há fortes dependências entre os atributos [32]. O critério de avaliação do método ReliefF é fornecido pela Equação 3-1.

$$SC(f_i) = \frac{1}{p}\sum_{t=1}^{p}\left(-\frac{1}{m_{x_t}}\sum_{x_j \in NH(x_t)} d(f_{t,i} - f_{j,i}) + \sum_{y \neq y_{x_j}} \frac{1}{m_{x_t,y}} \frac{P(y)}{1 - P(y_{x_t})} \sum_{x_j \in NM(x_t,y)} d(f_t, i - f_j, i)\right) \tag{3-1}$$

onde y_{x_t} é o rótulo da classe da amostra x_t e *P(y)* é a probabilidade de uma amostra ser da classe *y*. O número de amostras é denotado por *p*, $f_{t,i}$ indica o valor da amostra x_t para o atributo f_i, a função d(·) é uma medida de distância. *NH(x_t)* são os pontos mais próximos a *x* com a mesma classe de *x*, e *NM(x_t, y)* os pontos mais próximos a *x* com a classe diferente. Os termos m_{x_t} e $m_{x_t,y}$ são os tamanhos dos conjuntos *NH(x_t)* e *NM(x_t, y)*, respectivamente. Normalmente, o tamanho de *NH(x_t)* e *NM(x_t, y)* são pré-definidos como uma constante k para todo $y \neq y_{x_t}$.

Portanto, o ReliefF é um método que seleciona atributos que preservam a matriz de similaridade de dados. Métodos baseados em similaridade são simples e diretos, pois o cálculo se concentra na construção de uma matriz de afinidade, sendo, na sequência, obtidas as pontuações dos atributos. Além disso, esse método é independente de quaisquer algoritmos de aprendizagem e

os atributos selecionados são adequados para muitas tarefas de aprendizagem subsequentes. No entanto, uma desvantagem desse método é que este não consegue lidar com a redundância de atributos. Em outras palavras, este pode encontrar repetidamente características altamente correlacionadas durante a fase de seleção [27].

3.1.2.2 Algoritmos baseados em aprendizado esparso

Métodos baseados em aprendizado esparso são algoritmos que buscam a melhor relação entre a métrica de ajuste apropriada e a esparsidade dos resultados [33]. Alguns métodos baseados em análise espectral são oriundos dos métodos de decomposição espectral (SPEC, do inglês, *SPECtrum decomposition*) [33]. Exemplos destes são o método de seleção de atributos multi-grupos (MCFS, do inglês *Multi-Cluster Feature Selection* [34], FixedSPEC e GenericSPEC, que são variações do algoritmo SPEC [30].

A família de métodos SPEC considera um conjunto de instâncias par a par de similaridade S, a partir do qual se obtém um grafo G que representa estas similaridades. A classe-alvo, neste caso, pode ser considerada como a própria estrutura do grafo G. Um atributo que seja aderente à estrutura do grafo atribui valores semelhantes as amostras próximas umas das outras no grafo. Conforme mostrado na Figura 3.2, o atributo F atribui valores a amostras de forma consistente com a estrutura do grafo, mas F' não.

Assim, F tem maior similaridade com o alvo e pode separar melhor os dados (isto é, forma grupos com instâncias similares de acordo com o conceito de grafo). De acordo com a teoria dos grafos, as informações da estrutura de um grafo podem ser obtidas a partir de seu espectro. O método de seleção de atributos espectral estuda como selecionar características de acordo com as estruturas do grafo induzido por S [30]. Portanto, o algoritmo usa a Teoria do Grafo Espectral para encontrar atributo com a melhor separabilidade.

Logo, cada algoritmo desta família cria uma representação gráfica da distribuição de amostras em um espaço de alta dimensão. As amostras tornam-se vértices e a distância RBF (*Radial Basis Function*) entre eles torna-se o peso da aresta. Os algoritmos usam a Teoria Espectral dos Grafos para calcular as pontuações dos atributos. O algoritmo seleciona os k principais atributos de acordo com esta pontuação.

O algoritmo GenericSPEC encontra o autovetor trivial do Laplaciano do grafo e o usa para normalizar as pontuações calculadas usando a função objetivo *normalized cut* formulada por Zhao [30]. Essa normalização ajuda a melhorar a precisão da seleção de atributos. Essa função tenta encontrar o corte

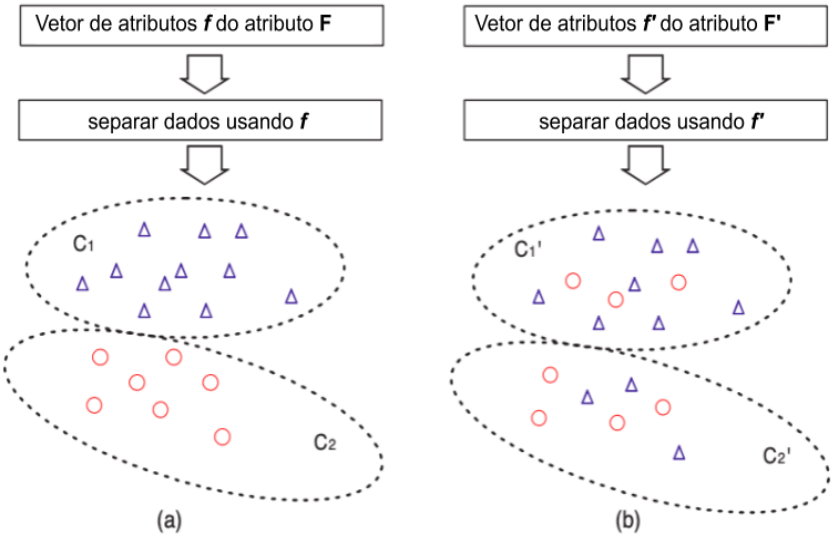


Figura 3.2: Comparação da aderência dos atributos à estrutura. O conceito de classe-alvo é representado pela estrutura do grafo (onde os grupos são indicados pelas elipses). Formas diferentes denotam valores diferentes atribuídos por um atributo, adaptado de Zhao [30].

mínimo neste grafo. O algoritmo seleciona k atributos que foram separados por este corte, pois são considerados os melhores para a explicação do conjunto de dados [30].

O algoritmo FixedSPEC, também é da família de algoritmos SPEC, usa a Teoria do Grafo Espectral para encontrar atributos com a melhor separabilidade, assumindo que os pontos são separados de acordo com um número predefinido de grupos. Para isso, o algoritmo encontra os autovetores correspondentes aos K menores autovalores do Laplaciano do grafo, exceto o trivial, e usa a distância de cosseno entre eles e os vetores de atributos para detectar as características mais explicativas. O algoritmo seleciona k atributos usando esta pontuação [30].

O algoritmo MCFS [34] foi um dos primeiros não-supervisionados de aprendizado esparso. Este é composto de três etapas:

1) análise espectral

2) aprendizado dos coeficientes esparsos

3) seleção de atributos

Na primeira etapa, de análise espectral, são detectadas as estruturas de grupos nos dados. Em seguida, na segunda etapa, uma vez conhecida a estrutura de grupos dos dados, através dos k primeiros autovetores da matriz do laplaciano, o MCFS mede a importância dos atributos através de um modelo

de regressão linear com regularização de norma l1. Finalmente, na terceira etapa, após resolver o problema de regressão, o MCFS seleciona os atributos, baseado no maior valor absoluto do coeficiente obtido através do problema de regressão. Este se mostrou eficiente na diferenciação de classes para problemas com poucos grupos ou binários [34].

3.2 Classificação

Os métodos de classificação de dados são um conjunto de técnicas baseadas em aprendizado supervisionado. Para a utilização destes, deve-se conhecer a priori os parâmetros de classificação, tal como os valores a serem preditos (rótulos) e o número exato de grupos [35].

A classificação é um problema que tem como objetivo identificar a que categoria pertence uma nova observação. A tarefa de classificação pode ser descrita formalmente da seguinte forma: dado um conjunto de exemplos de treinamento composto de pares x_i, y_i, deve-se obter uma função *f(x)* que mapeia cada vetor de atributo x_i para sua classe associada y_i, sendo i = *1, 2,. . . ,* n e n o número total de amostras de treinamento. Os algoritmos de classificação recebem como entrada um conjunto de dados rotulados. Nesse caso, assume-se que um conjunto de dados contendo amostras que representam todas as situações possíveis está disponível a priori [36].

Modelos de classificação tem sido amplamente utilizados em importantes aplicações na área de biomedicina [37]. Estes modelos podem fornecer informações sobre os mecanismos moleculares que resultam em determinados fenótipos [38].

Há diferentes tipos de modelos de classificação. Dentre estes o algoritmo floresta aleatória tem sido usado com diferentes finalidades [37]. Este algoritmo é um classificador que atribui uma importância a cada preditor do modelo, construindo uma infinidade de árvores de decisão. Cada nó de uma árvore considera um subconjunto de preditores selecionados aleatoriamente, dos quais o melhor preditor é selecionado e dividido. Um critério usado para determinar o melhor preditor é a diminuição da impureza do nó, estimado pela variância da resposta, como por exemplo, o índice de Gini [27]. Deste modo, cada árvore é construída usando uma amostra *bootstrap* aleatória, que consiste, em geral, de dois terços das observações totais e é usada como um conjunto de treinamento para prever os dados nas amostras *out-of-bag* (OOB) restantes, ou conjunto de teste. As previsões para cada variável são agregadas em todas as árvores e o erro quadrático médio (MSE) das estimativas OOB é calculado [39]. Darst e outros (2018) usaram o modelo floresta aleatória para integrar múltiplos dados

ômicos para a previsão de quatro características fenotípicas [39]. Os resultados indicaram haver correlações entre as variáveis nos conjuntos de dados.

Acharjee e colegas (2016) integraram múltiplos dados ômicos, entre estes, dados de expressão de mRNA, metabólitos, proteômicos e um conjunto selecionado de dados fenotípicos através de um modelo floresta aleatória. Neste estudo, foi observado que a integração de conjuntos relativamente pequenos de variáveis ômicas inter-relacionadas podem prever, com maior precisão, traços fenotípicos [40].

3.3 Agrupamento

Os métodos de agrupamento são baseados em treinamento não-supervisionado, no qual não é necessário conhecer a priori as características dos grupos (rótulos) nem, em alguns casos, o número de grupos [35].

Algoritmos de agrupamento podem ser subdivididos, de acordo com a abordagem utilizada para separar os grupos, entre: particionado, hierárquico, baseado em modelo, e outros. A categoria 'outros' engloba vários algoritmos de agrupamento que não se encaixam nesses grupos ou não são relevantes para este trabalho [35].

Métodos de agrupamento baseados em modelo assumem que um conjunto de dados corresponde a um modelo, que em muitos casos, é uma distribuição estatística. Os modelos são geralmente definidos pelo usuário, e são em geral escolhidos para lidar com resultados de agrupamento indesejáveis caso modelos inadequados (ou seus parâmetros) sejam escolhidos. Algoritmos de agrupamento baseados em modelo são geralmente mais lentos do que algoritmos particionais [41].

Os métodos de agrupamento hierárquico criam uma organização dos grupos de forma ordenada, de cima para baixo (ou de baixo para cima). Ou seja, o agrupamento é realizado de forma indireta e sucessiva, subdividindo-se (ou reagrupando) os grupos remanescentes. Estes algoritmos são constituídos dos seguintes elementos:

(i) Matriz de similaridade - É construída encontrando a similaridade entre cada par de pontos de dados. A escolha da métrica de similaridade para construir a matriz de similaridade tem grande influência no formato dos grupos obtidos;

(ii) Critério de ligação - determina a distância entre conjuntos de observações em função das distâncias par a par das observações.

A maioria dos algoritmos de agrupamento hierárquico tem uma grande complexidade de tempo e de memória que aumentam exponencialmente de

acordo com número de dados.

3.3.1 Algoritmos particionais

A família de algoritmos mais estudada em agrupamento de dados são os chamados algoritmos particionais [41]. Os métodos de particionamento tentam agrupar os dados diretamente. Eles dividem os dados em k grupos homogêneos, sendo k definido a priori. A partir da definição de k, são definidos aleatoriamente os centros desses k grupos. Em seguida, os dados são atribuídos ao centro de grupo mais próximo em razão da função de similaridade baseada na distância (como o algoritmo *k-means* e suas variantes) ou em probabilidade (como no caso dos algoritmos Modais de Mistura Gaussiana e variantes), e, então um algoritmo iterativo otimiza os grupos até a convergência [35] [41]. Um exemplo deste tipo de algoritmo é o *k-means*, que usa a distância euclidiana e tem por objetivo a minimização da distância dentro do mesmo grupo e a maximização da distância entre os grupos, minimizando sua função objetivo. De forma geral, as etapas que constituem os algoritmos particionais são:

1) Escolha centro dos grupos para representar características numéricas e/ou categóricas;

2) Seleção de função que indique a distância para refletir características numéricas e/ou categóricas;

3) Definição da função de objetivo a ser minimizada [41].

Considerando as três premissas citadas, a maioria dos algoritmos de agrupamento particional otimiza a seguinte função de custo iterativamente de acordo com a Equação 3-2:

$$\sum_{i=1}^{n} \xi(d_i, C_i) \tag{3-2}$$

sendo, n o número de pontos de dados no conjunto de dados, C_i o centro do grupo mais próximo do ponto de dados d_i, e ξ é uma medida de distância entre d_i e C_i.

Uma razão importante para a popularização destes métodos é que são escaláveis para conjuntos de dados grandes e podem ser adaptados para fluxos de dados paralelizados [41]. A questão mais importante para este tipo de agrupamento é que o resultado do agrupamento depende fortemente da inicialização dos centros e do número de grupos [35]. Uma forma simples de contornar estes problemas é através da realização de diversos experimentos, para a definição do número de grupos mais adequado.

Os métodos clássicos de agrupamento apresentados até aqui são ampla-

mente difundidos e utilizados devido à simplicidade de implementação e ao baixo tempo de execução principalmente para conjuntos de dados limpos, pequenos e sintéticos. Entretanto, estes métodos clássicos apresentam muitos limites quando usados para resolver agrupamentos de conjuntos de dados da vida real, que são ruidosos, incompletos, e para conjuntos de dados grandes e sobrepostos, ou seja, muitos dados podem não pertencer de forma estanque a apenas um grupo [35]. A Lógica *fuzzy*, criada por Lofti Zadeh [42], estende a noção tradicional de lógica para resolver problemas reais. Esta abordagem utiliza regras simples de lógica para resolver problemas complexos e não lineares. Ruspini foi o primeiro a implementar um sistema de agrupamento *fuzzy* [43].

Os métodos de agrupamento *fuzzy* são baseados na associação difusa dos dados, isto é, enquanto nos métodos clássicos de agrupamento rígido (*crisp*) os dados são atribuídos exclusivamente a um único grupo, no agrupamento *fuzzy* os dados podem ter diferentes graus de pertinência a mais de um agrupamento, de forma que os dados podem pertencer a vários grupos ao mesmo tempo. Isto permite analisar instâncias que possam conter características compatíveis com mais de um grupo, com diferentes graus de compatibilidade.

Existem três categorias de métodos de agrupamento *fuzzy*: baseados na relação *fuzzy*, baseados na regra dos k-vizinhos mais próximos e baseados na função objetivo. A última categoria é a mais usada no agrupamento *fuzzy*.

Esses métodos são usados para melhorar os resultados de agrupamento quando os limites são sobrepostos. Um dos principais modelos de agrupamento *fuzzy* é o algoritmo *fuzzy C-means* (FCM) [35]. O algoritmo *fuzzy C-means* foi desenvolvido por Dunn [44] e depois implementado por Bezdek [45].

Neste algoritmo a função de pertinência não é apenas 0 ou 1, mas um valor real entre 0 e 1, de modo que o vetor de pertinência (para *k-means*) é substituído por uma matriz de pertinência (para *C-means*) [35]. A pertinência é representada por uma matriz de c por n, onde c é o número de grupos *fuzzy* e n é o número de dados. Cada linha representa a associação de todos os n objetos a um determinado grupo *fuzzy* e cada coluna representa a associação de um objeto a todos os c grupos *fuzzy*. O parâmetro de fuzzificação m controla a sobreposição dos conjuntos, e é usado para reduzir a sensibilidade dos centros de cada classe ao ruído. Este parâmetro pode ser escolhido de acordo com a aplicação. Um maior valor de m resulta em grupos mais difusos, e um valor pequeno, próximo de 1, resulta em um agrupamento *crisp*, semelhante ao algoritmo *k-means*. Bezdek provou que o grau de fuzzificação m deve ser definido entre 1.5 e 2.5. Em geral, m é definido como 2 [35] [45]. Portanto, o algoritmo *C-means* é muito semelhante ao algoritmo *k-means*, mas com nova função objetivo J' dada pela Equação 3-3.

$$J'(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\|X_j - V_i\|^2 \tag{3-3}$$

Nesta equação $u_{i,j}$ representa o grau de pertinência da j-ésima amostra para o i-ésimo centro. O grau de fuzzificação é indicado por m, o número de grupos é indicado por c. V_i é o vetor que representa o centro de cada grupo. E X_j é a j-ésima amostra.

As etapas do algoritmo *fuzzy C-means* são semelhantes ao algoritmo *k-means*:

1) Escolher os C centros aleatoriamente (vetor V)

2) Calcular a matriz de partição U

3) Atualizar os centros V_i

4) Computar a função objetivo J'

5) Repetir as etapas 2 a 4 até a convergência ($\|var(J') \leq \epsilon\|$)

As questões mais importantes para o algoritmo FCM são o tempo computacional, que é maior que o *k-means* devido à natureza iterativa, escolha da métrica de distância (distância euclidiana nem sempre é a melhor), escolha dos centróides iniciais, escolha do número de grupos e escolha do parâmetro de fuzzificação m [46].

Considerando que o objetivo de algoritmos de agrupamento é separar as amostras de modo a observar padrões semelhantes em uma classe e padrões diferentes em classes disjuntas, a otimização da função objetivo consiste em tentar maximizar a separação entre os grupos de um conjunto de dados e minimizar a distância entre os elementos dos grupos. Pode-se avaliar se o objetivo foi alcançado através de métricas de desempenho. Idealmente, essas métricas devem avaliar os seguintes aspectos do particionamento [47]:

1) **Coesão**: Os padrões dentro do mesmo grupo devem ser tão semelhantes quanto possível para um bom agrupamento. Esta é uma medida da compactação dos pontos de dados em um agrupamento.

2) **Separação**: Os grupos devem ser bem separados entre si. A distância entre os centros dos grupos é uma medida eficaz da separação de dois grupos.

3) **Estabilidade parcial**: os resultados de agrupamentos de dois conjuntos de dados diferentes que foram gerados pela mesma fonte devem ser semelhantes. Por exemplo, os resultados gerados por um algoritmo de agrupamento e por um classificador treinado utilizando um conjunto de dados agrupado por este algoritmo devem ser semelhantes [47].

Essas métricas são funções matemáticas que também podem ser usadas para comparar o resultado de diferentes resultados de agrupamento [48] e avaliar o número ótimo de grupos de um conjunto de dados [49].

O coeficiente de partição *fuzzy* (FPC, do inglês *fuzzy partition coefficient*) é uma métrica *fuzzy* que foi projetada para medir a quantidade de sobreposição entre os grupos [48]. A formulação matemática dessa métrica é dada pela Equação 3-4.

$$FPC = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^2 \tag{3-4}$$

onde μ_{ij} denota o valor de pertinência do *j-ésimo* ponto ao *i-ésimo* grupo. Dessa forma, FPC é inversamente proporcional à sobreposição geral entre pares de subconjuntos difusos. As desvantagens do coeficiente de partição são a falta de uma conexão direta com uma propriedade geométrica e sua tendência decrescente monotônica com c, o número de grupos.

Outra métrica *fuzzy* popular foi proposta por Fukuyama e Sugeno [48], explorando a coesão e a separação. A formulação matemática é apresentada na Equação 3-5, onde o primeiro termo é uma medida de compactação e o segundo termo é o grau de separação entre cada grupo e a média (V) dos centróides do grupo [48].

$$FS = \sum_{j=1}^{n}\sum_{i=1}^{c}(\mu_{ij})^m\|x_j - V_i\|^2 - \sum_{i=1}^{c}\|V_i - \bar{V}\|^2 \tag{3-5}$$

Onde,

$$\bar{V} = \frac{1}{c}\sum_{i=1}^{c}V_i \tag{3-6}$$

Quanto mais separação e menor a medida de compactação, melhor é o agrupamento de acordo com esta métrica.

Xie e Beni (1991) [49] também propuseram uma métrica *fuzzy* que foca na compactação e separação [49]. A função objetivo S proposta é apresentada na Equação 3-7:

$$XB = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c}(\mu_{ij})^2\|x_j - V_i\|^2}{n[\|V_i - V_k\|^2]} \tag{3-7}$$

Analisando as métricas de agrupamento fuzzy apresentadas, percebe-se que a métrica proposta por Fukuyama e Sugeno (FS) é mais sensível ao número de grupos. Isto se deve ao fato da média dos centróides do grupo ser inversamente proporcional ao número de grupos. Portanto, o aumento deste parâmetro resulta na diminuição da média dos centróides e em um menor valor

para esta métrica. Deste modo, espera-se que esta seja mais sensível à variação deste parâmetro auxiliando na escolha do valor deste parâmetro.

Considerando que a métrica FPC é um valor médio do coeficiente de partição u, por isso deve mostrar ser mais sensível à variação do parâmetro de fuzzificação m.

A métrica proposta por Xie Beni (XB), demonstra relação com a compacidade do grupo, considerando também a variação do número de grupos, porém menos que a métrica FS.

3.4 Estratégias de integração de dados biológicos

Os métodos de integração de dados podem ser categorizados em dois tipos de abordagens: análise de múltiplos estágios e análise multi-dimensional. Na análise de múltiplos estágios, os modelos são construídos usando apenas duas escalas diferentes de cada vez, de maneira gradativa, linear ou hierárquica. Escalas são as características numéricas e categóricas dos dados. Por exemplo, as variáveis SNP (do inglês, *single nucleotide polymorphism*) são variáveis categóricas, indicando presença ou ausência de um determinado SNP, e as variáveis de expressão de mRNA contêm valores contínuos, indicando o nível de expressão dos genes. Estes por sua vez podem ser classificados em categorias de acordo com o perfil de expressão em genes superexpressos ou subexpressos. A análise meta-dimensional, ou fusão de escalas, é uma abordagem na qual todas as escalas de dados são combinadas simultaneamente para identificar modelos complexos e meta-dimensionais com múltiplas variáveis de diferentes tipos de dados [50]. O principal objetivo da abordagem de múltiplos estágios é dividir a análise em várias etapas para encontrar associações primeiro entre os diferentes tipos de dados e, posteriormente, entre os tipos de dados e o traço ou fenótipo de interesse [50]. Entretanto, esta abordagem não considera possíveis interações entre todas as variáveis de forma simultânea, refletindo o ambiente interativo que estas descrevem.

A análise meta-dimensional combina vários tipos de dados em uma análise simultânea. A integração baseada em concatenação e a integração baseada em transformação são dois exemplos desse tipo de análise. Estas definem a forma como os dados são manipulados antes de aplicar o algoritmo de integração de dados.

A integração baseada em concatenação é realizada combinando dados ômicos em uma única matriz, para combinar todos os dados em uma única base de dados [2] [50] [51]. Deste modo, depois de se determinar como combinar as variáveis em uma matriz, é relativamente fácil usar qualquer método estatístico

para analisar os dados contínuos e categóricos [2]. Outra vantagem desta abordagem é que a integração baseada em concatenação é particularmente útil para considerar interações entre diferentes tipos de dados [50]. Um exemplo de método baseado em concatenação é o *Multiple Co-Inertia Analysis* (MCIA) [52], uma extensão do *Co-Inertia Analysis* [53] para mais de dois tipos de dados. Este método considera a otimização da covariância global da matriz concatenada de dados de mRNA, miRNA e proteômica, e consegue distinguir perfis de linhagens celulares de melanoma, leucemia e sistema nervoso central [54]. Além deste, outro exemplo é a análise de múltiplos fatores (MFA) [51] que é um método baseado em concatenação cuja estratégia é a análise de componente principal (do inglês, *principal component analysis*, PCA) da matriz concatenada. O MFA foi aplicado por Tayrac e colaboradores (2009) [51] para medições de número de cópias e expressão de mRNA de um conjunto de dados de glioma para estudar diferenças entre diferentes subtipos de tumor.

Os métodos baseados em transformação integram dados ômicos após sua transformação em uma representação intermediária dos dados, como um grafo ou uma matriz de kernel. A principal vantagem de uma etapa de transformação é preservar características ômicas individuais que podem ser perdidas ao serem integradas sem esta representação. Por exemplo, o *Similarity Network Fusion* (SNF), descrito por Wang e colaboradores (2014) [55], cria redes de similaridade de pacientes a partir de dados ômicos. O método reconheceu três subtipos de glioblastoma multiforme com diferentes perfis de sobrevida a partir da integração da metilação do DNA, expressão de mRNA e miRNA.

3.5 Trabalhos relacionados

As abordagens de integração de dados se diferenciam também com relação à natureza do fenótipo que se pretende observar. Neste sentido, existem trabalhos prévios na literatura que utilizaram dados ômicos de apenas um sítio tumoral para identificar subtipos moleculares [56], [57], [58], [59], [60], enquanto outros trabalhos foram capazes de identificar subtipos moleculares de câncer independente do sítio anatômico, analisando diferentes tipos de cânceres simultaneamente [6], [61], [62], [63].

No trabalho desenvolvido por Guinney e colaboradores [56], foram identificados 4 subtipos CMS (CMS, do inglês *Consensus Molecular substypes*) de CCR (Câncer Colorretal) através da observação da equivalência do resultado de 6 agrupamentos considerando somente dados de expressão de mRNA através da distância de Jaccard [56]. Os dados agrupados eram provenientes de diferentes técnicas de mensuração deste dado (microarranjos e RNA-Seq), e o

número de grupos variou entre 3, 5 e 6.

É importante observar que as técnicas usadas no agrupamento foram muito semelhantes, todas basearam-se em agrupamento hierárquico.

Posteriormente, com as amostras identificadas como pertencentes a algum grupo na etapa anterior, foi realizado o treinamento de um classificador baseado em floresta aleatória (*random forest*). As amostras que não foram identificadas como pertencentes a algum grupo na etapa de agrupamento, foram classificadas com este classificador. Entretanto, mesmo se utilizando dessa estratégia, houve uma alta quantidade de amostras não classificadas, no geral 13% dos pacientes não foram identificados como pertencentes a algum grupo.

É importante destacar a necessidade de se diminuir a quantidade de pacientes identificados de forma equivocada devido à gravidade que o diagnóstico incorreto da doença pode acarretar, como a postergação da realização de tratamento adequado.

Ressalta-se ainda que, apesar de serem indicados biomarcadores no trabalho de Guinney [56], devido ao alto compartilhamento de características, os indicadores identificados ainda não podem ser considerados como indicadores inequívocos do subtipo da doença. Ou seja, os resultados deste trabalho, da perspectiva clínica, ainda permanecem sem esclarecer quais as características capazes de elaborar a melhor ferramenta de subtipificação para esta doença.

Lu e colaboradores (2018) integraram dados de expressão de mRNA e outros dados moleculares para identificar genes *drivers*, os quais conferem vantagem no crescimento clonal das células tumorais. Foram usados métodos estatísticos de aprendizado de máquina para realizar a pré-seleção de atributos e uma análise modular de rede para identificar potenciais candidatos a genes *driver* [58]. A seleção final de genes foi realizada através do cálculo da distância entre estes nos subtipos. Para validar a especificidade dos genes *driver*, dados de expressão de mRNA foram usados para classificar amostras de pacientes com validação cruzada. Também foram realizadas análises de enriquecimento nos genes identificados. Os resultados mostraram que o método de integração proposto pode identificar os genes potencialmente *driver* e o classificador destes genes demonstrou melhor desempenho do que com outros genes identificados por outros métodos [58].

Neuroblastoma de alto risco é uma doença muito agressiva, com excesso de crescimento tumoral e prognóstico ruim. Zhang e colaboradores (2018) usaram um *autoencoder* combinado com agrupamento *k-means* para realizar a integração multiômica dos dados de expressão de mRNA e CNV [59]. Foram identificados dois subtipos moleculares com sobrevidas significativamente diferentes. Com relação à classificação baseada na integração multiômica de dados,

foi observado que a classificação com autoencoder tem melhor desempenho do que as abordagens alternativas. Os resultados mostraram que os subtipos identificados pelas técnicas de aprendizados de máquina permitiram não apenas aprimorar o entendimento sobre os mecanismos moleculares mas também auxiliar as decisões clínicas [59].

Dentre estes, o trabalho de Liu e colaboradores (2016) [60] se destaca, pois este realizou o agrupamento de dados de miRNA, metilação, expressão de mRNA e número de cópias para 256 amostras do carcinoma hepatocelular obtidas do repositório do TCGA utilizando a técnica *"Cluster of a Cluster"*, que é uma técnica de agrupamento hierárquico. Os resultados dos agrupamentos de cada ômica foram posteriormente analisados. Portanto, foi realizado um agrupamento para cada ômica e estes foram integrados utilizando a distância de Jaccard para verificar a coincidência entre cada grupo. Os resultados foram caracterizados de acordo com as variações nos dados de mutação e expressão de proteínas, identificando-se 5 subtipos. As curvas de sobrevida e os dados relacionados aos subtipos também foram distinta entre si [60].

Ramazzotti e colaboradores (2018) realizaram a integração de dados de câncer através de aprendizado multi-kernel e uma nova metodologia de subtipificação de dados de câncer [6]. Esta abordagem foi aplicada a um conjunto de dados correspondendo a 36 diferentes tipos de câncer e mostrou melhorias significativas tanto no aspecto de eficiência computacional como na habilidade de extrair subtipos de câncer. Os subtipos encontrados mostraram diferenças significativas nas curvas de sobrevida de 27 dos 36 tipos tumorais. As análises mostraram a existência de padrões de expressão de mRNA, metilação, mutação, e variações no número de cópias CNV de vários cânceres, e mostraram padrões específicos associados a sobrevidas ruins [6].

No trabalho de Liu e colaboradores (2018) foram analisadas 921 amostras de adenocarcinomas de esôfago, estômago, cólon e reto [64]. O trabalho realizou o agrupamento hierárquico de dados de metilação para identificar as principais diferenças e semelhanças entre os tumores do trato gastrointestinal. Esta análise foi realizada também entre estas amostras e amostras de outros cânceres.

Lyu e colaboradores (2018) utilizaram dados do *Pan-Cancer Atlas*, que fornece informação de 33 tipos de tumores, como conhecimento primário para gerar biomarcadores específicos [62]. Foram combinados dados de RNA-Seq em imagens 2-D e uma rede neural convolucional foi usada para classificar os 33 tipos de câncer. A acurácia final foi de 95.59%, que é maior que a de outros trabalhos que utilizaram algoritmos genéticos e K-Nearest Neighbor para o mesmo conjunto de dados. Foi gerado um mapa de calor relativo à

significância dos genes para cada classe, utilizando o algoritmo *Guided Grad Cam*. Através de uma análise funcional dos genes, as grandes intensidades nos mapas de calor validaram que estes top genes eram relacionados a vias tumorais específicas, sendo que alguns destes já foram indicados em outros trabalhos como biomarcadores, o que prova a efetividade do método desenvolvido. Este foi o primeiro trabalho a aplicar uma CNN a dados de *Pan-Cancer Atlas* para classificação, e também foi o primeiro trabalho a combinar a significância da classificação com a importância dos genes. Os resultados do experimento desenvolvido mostraram que a abordagem tem bons resultados e pode ser usada em outros dados genômicos [62].

A classificação de pacientes com perfis de mutação diferentes pode ajudar na identificação de subtipos moleculares que podem se beneficiar de tipos específicos de tratamento. No entanto, a classificação de dados de mutação é complexa devido à esparsidade e heterogeneidade dos dados. Kuijjer e colaboradores (2018) utilizaram a análise de vias biológicas para tornar os dados de mutação somática menos esparsos [63]. Este método foi aplicado a 23 tipos de câncer do TCGA, incluindo as amostras de 5.805 tumores primários [63]. Os resultados para a maioria dos tipos de câncer mostram que os dados não-esparsos de mutação estão associados a perfis fenotípicos. Foram identificados prognósticos ruins para 3 tipos de câncer, associados a padrões de enriquecimento de vias específicos, para os quais tratamentos específicos estão disponíveis. Foi realizada uma análise com os dados do *Pan-Cancer Atlas* e foram identificados 9 subtipos de câncer com perfis de mutação associados a quatro grupos de vias biológicas [63].

Com relação ao CCR não foram realizados estudos integrando diferentes conjuntos de dados ômicos. Dos trabalhos que integraram diferentes ômicas, estes utilizaram-se de técnicas de agrupamento ou classificação rígidas, que não refletiam o alto compartilhamento de características dos dados. Por isso, este trabalho buscou preencher estas lacunas utilizando as técnicas que serão analisadas no Capítulo 4.

4
Integração fuzzy de dados multiômicos

4.1
Introdução

Este capítulo apresenta a abordagem proposta de integração multiômica de dados que utiliza técnicas de aprendizado de máquina para identificar relações entre características genotípicas e fenotípicas.

O fluxograma de trabalho descrito é ilustrado na Figura 4.1. A primeira etapa consiste na escolha dos conjuntos de dados ômicos a serem analisados, a qual é realizada através da avaliação do desempenho da abordagem multiômica. Em seguida, é definida a combinação de métodos de seleção de atributos, a partir dos quais são identificados os atributos mais informativos associados a cada conjunto de dados. Os dados selecionados são então integrados, através de concatenação em uma base de dados única. Por fim, é realizado o agrupamento *fuzzy* destes dados, o qual é avaliado por métricas de desempenho adequadas que qualificam os parâmetros e atributos escolhidos.

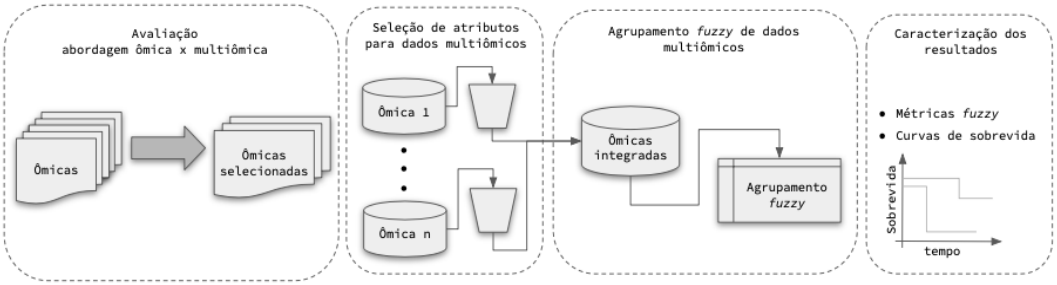


Figura 4.1: Fluxograma da metolodogia de integração fuzzy de dados ômicos

Nas seções a seguir será detalhada cada etapa do fluxograma da metodologia citada e ilustrada na Figura 4.1.

4.2
Pré-processamento

4.2.1
Abordagem ômica x multiômica

Conforme visto na Seção 2.2, as técnicas de sequenciamento são mais sensíveis, fornecendo níveis de quantificação absoluta. Entretanto, essas técnicas têm como uma de suas maiores desvantagens o alto custo em comparação às técnicas de microarranjo. Por outro lado, as técnicas de microarranjo têm desvantagens que podem ser contornadas com o devido pós-processamento do resultado. Por isso, ambas as formas de obtenção de dados serão usadas neste trabalho, condicionando-se à disponibilidade das mesmas.

É importante observar que os conjuntos de dados a serem analisados devem ser primeiramente conciliados, isto é, que as amostras não constantes nas bases de dados simultaneamente devem ser desconsideradas.

As variáveis escolhidas devem ser filtradas, de forma breve e preliminar com relação a seus atributos, eliminando-se atributos pouco informativos. Por exemplo, atributos com todos valores iguais a zero, para todas as amostras, ou com divergências decorrentes de inconsistências, como amostras com atributos com valor não especificado (NA), ou ainda amostras duplicadas, devem ser excluídos da base.

A metodologia proposta considera o conhecimento de trabalhos anteriores para avaliar a escolha dos conjuntos de dados analisados. Isto é possível pela utilização de um método de classificação supervisionado com a classe-alvo baseada em trabalhos da literatura.

A comparação do resultado deste classificador multiômico com outros classificadores da literatura permite analisar o efeito da utilização destes conjuntos de dados. A observação de um desempenho melhor ou pior que o observado na literatura é indicativo de que os atributos dos conjuntos de dados multiômicos adicionados têm maior ou menor identidade, respectivamente, com os agrupamentos que originaram as classes-alvo dos classificadores destes trabalhos. Do mesmo modo, entende-se que a obtenção de um desempenho do classificador multiômico próximo ao encontrado na literatura, denota identidade semelhante dos atributos dos conjuntos de dados multiômicos àqueles atributos usados nos classificadores da literatura.

Nesse contexto, o classificador floresta aleatória, conforme pontuado na Seção 3.2, apresenta-se como uma técnica capaz de avaliar a adição dos conjuntos de dados com relação a resultados da literatura. Deste modo, as etapas para se realizar a avaliação do desempenho da abordagem multiômica em relação à abordagem de uma ômica individualmente aqui descrita podem ser ilustradas como na Figura 4.2.

Observa-se que na 1a etapa os conjuntos de dados ômicos devem ter

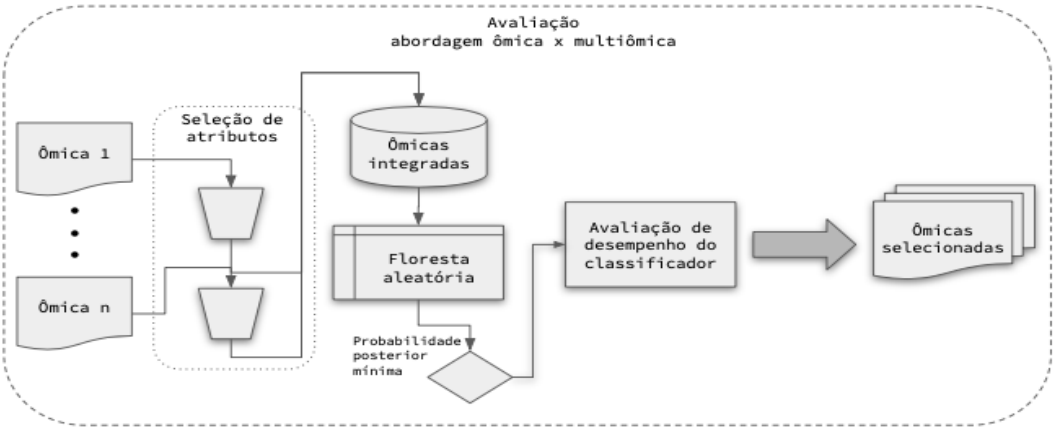


Figura 4.2: Fluxograma da avaliação do desempenho da abordagem multiômica em relação à abordagem de uma ômica individualmente

seus atributos selecionados. Os dados selecionados são então concatenados em uma única matriz, conforme mencionado na Seção 3.4. Em seguida uma parte dos conjuntos de dados, aproximadamente 2/3 das amostras desta matriz, devem ser utilizados para treinar o modelo de floresta aleatória, e 1/3 deve ser utilizado para testar o desempenho do modelo. Para se facilitar a comparação com resultados da literatura, deve-se definir uma probabilidade posterior mínima entre as resultantes do classificador.

Caso identifique-se que o desempenho do classificador tenha uma acurácia mais elevada ao utilizar-se uma abordagem multiômica, considera-se que as ômicas adicionadas tem maior identidade com o agrupamento que gerou as classes-alvo do classificador, por isso, estas ômicas podem auxiliar no melhor entendimento da relação genótipo-fenótipo. No entanto, dada a alta dimensionalidade destes dados, estes devem ser submetidos à métodos de redução de dimensionalidade que serão apresentados na próxima seção.

4.2.2 Seleção de atributos para dados multiômicos

Como apresentado na Seção 3.1.2, é necessário utilizar-se de técnicas que possam diminuir o número de atributos antes de se realizar a integração dos dados. A fim de se obter as vantagens de diferentes técnicas, propõe-se nesse trabalho que sejam realizadas análises com diferentes métodos de seleção de atributos. Em seguida, os resultados destas técnicas devem ser normalizados e sumarizados em um *ranking* que represente a importância de cada atributo conjugando-se os diferentes métodos.

Considerando-se as vantagens e desvantagens dos métodos de seleção supervisionados e não-supervisionados, e considerando-se a disponibilidade de

informações de classe-alvo na literatura, este trabalho propõe que métodos de seleção de atributos supervisionados e não-supervisionados sejam combinados de modo a obter os benefícios de ambos os tipos de métodos. Assim, considera-se os resultados de estudos já realizados anteriormente, com objetivo de restringir o quantitativo de atributos, mas espera-se ser possível descobrir novos atributos importantes para a caracterização do fenótipo.

Quanto à estratégia de busca, observa-se que os métodos de filtro são os mais adequados ao tratamento de dados ômicos pela flexibilidade de não estarem atrelados a um algoritmo de aprendizado, como os métodos *embedded*, e serem computacionalmente menos custosos que os algoritmos *wrapper*, por exemplo.

Devido à característica esparsa das matrizes de dados ômicos, propõe-se utilizar técnicas de seleção de atributos de aprendizado esparso. É necessário evitar a redundância dos atributos em razão do elevado número destes, utilizando também métodos baseados no critério de seleção da semelhança. Adicionalmente, métodos estatísticos podem fornecer informações importantes sobre a magnitude da variação da expressão.

Portanto, a associação de algoritmos que tenham estas características, como os algoritmos MCFS, FixedSPEC, GenericSPEC (apropriados para matrizes esparsas), o algoritmo reliefF (atribui importância aos atributos com relação à semelhança entre estes), e de análise de expressão diferencial (método estatístico DESeq2), tem como objetivo obter os benefícios oferecidos por cada método.

O resultado de cada método de seleção de atributos é fornecido em diferentes ordens de grandeza. Deste modo, estes devem ser normalizados antes de serem combinados. A normalização entre os mínimos e máximos do conjunto de dados permite que todos estejam no intervalo de 0 a 1, o que possibilita a comparação e combinação entre os diferentes métodos.

Mesmo obtendo-se o *ranking* de importância dos atributos, resta ainda a tarefa de elencar um valor mínimo de pontuação deste, ou uma quantidade específica de atributos a serem selecionados como mais informativos.

Para isso, propõe-se a utilização da técnica *k-means*, mencionada na Seção 3.3.1, de modo a verificar-se qual o grupo de atributos a ser selecionado. Também propõe-se que esta seleção seja feita pelo ajuste de uma reta ao *ranking*, de modo a identificar de forma direta o ponto de inflexão deste e separar quais são os atributos mais informativos.

Ainda, realiza-se, de forma heurística, a escolha de diferentes valores mínimos de pontuação para obter-se o melhor conjunto de atributos, selecionando os atributos com diferentes pontuações: acima de 0.90, 0.80, 0.70, 0.60 e 0.50.

Para refinar os resultados, propôs-se testar também a utilização de atributos com pontuações maiores que 0.875, 0.85 e 0.825, visando a investigar qual a melhor quantidade de atributos a ser utilizada.

As etapas descritas nesta metodologia para a seleção dos atributos informativos referentes a cada conjunto de dados são ilustradas na Figura 4.3.

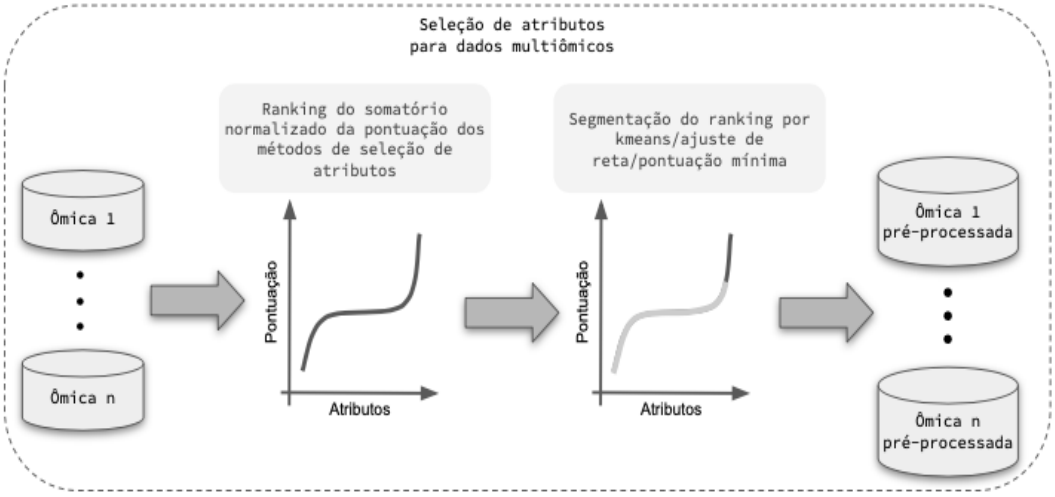


Figura 4.3: Fluxograma da seleção dos atributos informativos referentes à cada conjunto de dados

4.3 Agrupamento *fuzzy* de dados multiômicos

Com a redução do número de atributos na etapa de seleção, aplica-se o método de concatenação de dados para realizar a integração de dados ômicos. Conforme mostrado na Seção 3.4, a abordagem de concatenação consegue, ao juntar os diferentes conjuntos de dados, obter diferentes níveis de informações biológica na mesma matriz e considerar assim as interações entre estes diferentes conjuntos de dados.

Esta abordagem também mantém as características dos dados de cada ômica, que poderiam se perder caso os atributos fossem integrados antes de serem selecionados em cada conjunto de dados. Com esta tabela de dados concatenados, pode-se realizar uma análise dos dados como um todo.

Esta metodologia busca identificar a relação das gradações dos dados moleculares com relação aos diferentes grupos e o aspecto difuso que estas características podem apresentar entre os grupos, como exposto na Seção 3.3.1. Por isso, o agrupamento *fuzzy C-means* apresenta-se como uma técnica promissora para atender à complexidade do problema proposto e ampliar o conhecimento em relação a estes aspectos dos dados.

Este modelo tem dois parâmetros, como mostrado na Seção 3.3.1: o grau de fuzzificação m e o número de grupos c.

Conforme os limites para o valor de m sugeridos na Seção 3.3.1, neste trabalho serão utilizados valores de m de 1.5, 1.75 e 2.

O número de grupos, como observado em trabalhos anteriores abordados na Seção 3.5, varia em geral entre 4 e 8 grupos. Este intervalo de grupos independe do tipo de dado ômico; por isso, esta metodologia propõe a utilização de números de grupos entre 4 e 8.

A biblioteca *skfuzzy* em *python* disponibiliza uma implementação do algoritmo *fuzzy C-means*, tendo sido selecionada neste trabalho para se realizar o agrupamento dos dados. Nesta implementação, a métrica de desempenho considerada para o atingimento da função objetivo é a métrica FPC, citada na Seção 3.3.1.

4.4 Caracterização dos resultados

A avaliação dos resultados com relação ao agrupamento das amostras deve considerar que as amostras não devem se concentrar significativamente em apenas um grupo, uma vez que objetivo de toda abordagem de agrupamento é subdividir as amostras entre diferentes grupos. O resultado do agrupamento *fuzzy* é uma matriz de pertinência na qual, para cada amostra, é definido um valor de pertinência relativo a cada grupo, como apontado na Seção 3.3.1. Este grau de pertinência foi avaliado de duas formas:

1) grau mínimo de pertinência: para facilitar a análise dos resultados, foi aferido um grau mínimo de pertinência de modo que as amostras que não atingirem este grau mínimo foram agrupadas em um grupo denominado 'não-grupo' (NG). Este grau mínimo varia para cada número de grupos, conforme mostrado na Figura 4.4

2) maior de grau de pertinência: as amostras serão consideradas pertencentes ao grupo de maior grau de pertinência.

A análise do resultado deste agrupamento, com relação aos dados de entrada, através de técnicas da técnica PCA, apresentada na Seção 3.1, permite a visualização dos resultados em baixa dimensionalidade, como mostrado na Seção 3.1.

A sobrevida global também auxilia na investigação dos resultados, conforme mostrado na Seção 2.5. Propõe-se a utilização do método de Kaplan-Meier com um teste de log-rank [25]. O pacote ggsurvplot em linguagem R foi usado para implementar esta análise.

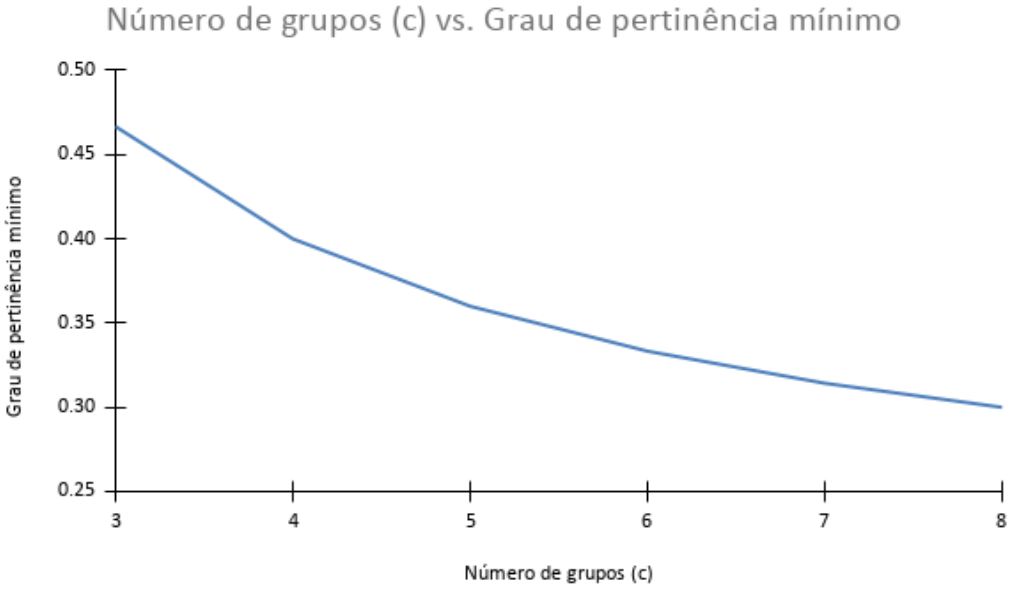


Figura 4.4: Grau mínimo de pertinência de acordo com número de grupos

As diferentes combinações de parâmetros do modelo *fuzzy* devem ser avaliadas através de métricas *fuzzy* apropriadas. As métricas apresentadas na Seção 3.3.1, por exemplo, consideram de formas diferentes as características do modelo. Algumas das métricas apresentadas tem maior ênfase de algum parâmetro específico em sua formulação matemática, conforme mencionado na Seção 3.3. A métrica FS por exemplo, é mais sensível ao número de grupos, por isso, auxiliará a definir qual o melhor número de grupos entre os valores testados. A métrica, FPC é mais afetada pelo valor do grau de fuzzificação, por isso auxiliará na escolha deste parâmetro. Por fim, a métrica XB, por ponderar ambos os parâmetros, número de grupos e grau de fuzzificação, auxiliará no entendimento geral do modelo.

A metodologia descrita neste capítulo foi avaliada em relação a dados de câncer colorretal. O resultado para diferentes combinações de parâmetros do modelo são apresentados no capítulo a seguir.

5
Estudo de caso

5.1
Câncer colorretal

O câncer colorretal (CCR) é o quarto tipo de tumor maligno mais letal no mundo [65]. O maior fator de risco é a idade, seguido de histórico familiar de primeiro grau de CCR [66]. Fatores demográficos, comportamentais e ambientais têm sido associados ao aumento do risco do CCR [67] [68] [69].

As estratégias para reduzir a incidência e mortalidade do CCR incluem as etapas de prevenção primária, como alterações na dieta ou aumento da atividade física, e prevenção secundária, que tem como exemplo o acompanhamento através de exames.

Fatores modificáveis associados com maior risco de CCR incluem ingestão de álcool, obesidade, tabagismo e consumo de carne processada e vermelha. De outro modo, os fatores modificáveis que têm sido associados à redução do câncer colorretal são o aumento da atividade física, terapia hormonal pós-menopausa, uso de anti-inflamatórios não esteroides e ingestão de vegetais e frutas.

No entanto, os resultados dos estudos que mostram a associação da ingestão de frutas e vegetais com a diminuição do risco de câncer colorretal têm se mostrado inconsistentes.

Os fatores não modificáveis que aumentam o risco de desenvolvimento da doença incluem doença inflamatória intestinal, história familiar de câncer colorretal e idade [67].

A correta identificação do subtipo molecular da doença pode ajudar na escolha do tratamento mais adequado e consequente aumento da sobrevida dos pacientes. Uma abordagem oncológica de precisão, que caracteriza o tumor com base nas particularidades genéticas do paciente, pode contribuir com maior assertividade na designação de terapias.

A identificação dos subtipos moleculares de cânceres com base nas propriedades moleculares é uma abordagem que vem sendo cada vez mais estudada, devido à maior disponibilização de informações relacionadas à taxonomia molecular de diferentes tumores.

Com isso, dados moleculares têm sido fonte de informação relevante para a classificação do subtipos de CCR. Atualmente, a falta de evidências suficientes de quais são os biomarcadores capazes de identificar corretamente os subtipos de CCR [56] indica que há grande oportunidade de se realizar avanços com relação ao estudo desta doença.

Neste trabalho, foram analisados dados de câncer colorretal para identificação de subtipos moleculares. Primeiramente, realizou-se a escolha dos conjuntos de atributos a serem agrupados. Em seguida, se realizou a seleção de atributos de cada conjunto de dados, individualmente. Esta etapa foi realizada por diferentes técnicas e estas, devidamente normalizadas, foram então combinadas no *ranking* de importância dos atributos, permitindo o teste de diferentes subconjuntos de atributos do *ranking*.

Realizou-se em seguida o agrupamento dos dados através do algoritmo de agrupamento não supervisionado *fuzzy C-means*.

A caracterização dos resultados considerando as diferentes combinações de parâmetros e métodos de seleção de atributos foi validada através de métricas apropriadas e da identificação dos fatores biológicos relevantes como curva de sobrevida, análise do local de origem do tumor, estadiamento da doença, frequência de mutações *driver* e sexo biológico.

5.2 Pré-processamento

5.2.1 Escolha dos conjuntos de dados de CCR

Alguns conjuntos de dados, como por exemplo, metilação, expressão de mRNA e miRNAs, foram analisados separadamente em outros estudos, como mostrado na Seção 3.5. Deste modo, a abordagem apresentada neste trabalho se utiliza destes conjuntos de dados para avaliar a integração de dados multiômicos.

A utilização de dados de expressão de mRNAs e miRNA deve ser seguida das normalizações FPKM e RPM, respectivamente, como pontuado na Seção 2.3. Para dados de metilação deve-se utilizar os valores normalizados β que refletem os níveis absolutos de metilação do DNA, como mostrado na Seção 2.2.

Os dados moleculares e clínico-patológicos foram extraídos do repositório do TCGA. Os dados de expressão de mRNA e miRNA são provenientes de sequenciamento e os de metilação de técnicas de microarranjo.

Os dados de expressão mRNA e miRNA foram normalizados de acordo com a metodologia proposta na Seção 4.2.1. Os dados de metilação foram interpretados conforme também mencionado na Seção 4.2.1.

Foi aplicada a remoção de atributos pouco informativos e conciliação dos conjuntos de dados, como mencionado na Seção 4.2.1, em todas as bases de dados. Portanto, considerou-se apenas amostras de tumor primário.

Verificou-se a aderência dos conjuntos de dados selecionados aos resultados da literatura. Para isto foi elaborado um classificador multiômico análogo ao apresentado por Guinney [56], conforme metodologia apresentada na Seção 4.2.1.

Deste modo, foram usados os mesmos parâmetros que o classificador baseado em expressão de mRNA elaborado por Guinney [56], a saber: método de seleção de atributos baseado na mediana dos atributos, 500 árvores, 70 nós por árvore, e divisão de amostras de teste e treino de 1/3 e 2/3, respectivamente. A classe-alvo usada no classificador multiômico foram os resultados do trabalho de Guinney [56].

Portanto, a avaliação do desempenho do classificador multiômico permite observar se a utilização de um conjunto de dados multiômicos fornece maior nível de informação do que a análise de apenas um conjunto de dados ômicos, como a realizada por Guinney [56].

O classificador de dados multiômicos obteve uma acurácia de 96%, conforme mostrado na Figura 5.1, considerando a probabilidade posterior de 0.50, a mesma escolhida no trabalho de Guinney [56], que obteve acurácia inferior, 85%.

O classificador multiômico obteve especificidade de 98%, como pode ser observado na Figura 5.2, enquanto o classificador baseado em expressão de mRNA obteve 90% [56], considerando a probabilidade posterior de 0.5. De forma geral, observa-se que o classificador multiômico obtém maiores valores de especificidade para todos os grupos, e no geral (98%), quando comparado ao classificador baseado em expressão que tem especificidade geral e por grupo menor (80%), considerando probabilidade posterior 0,50.

Este ganho na aderência das amostras às suas classes se reflete em um pequeno aumento na proporção de não classificados da abordagem multiômica(0.162) que é maior que a da abordagem baseada em expressão de mRNA(0.13).

Entretanto, das amostras analisadas neste trabalho e não-classificadas no classificador baseado em expressão, apenas 0,46%(3 amostras) não foram classificadas pelo classificador multiômico. As amostras com suas classificações definidas pelo classificador multiômico estão na Figura 5.3.

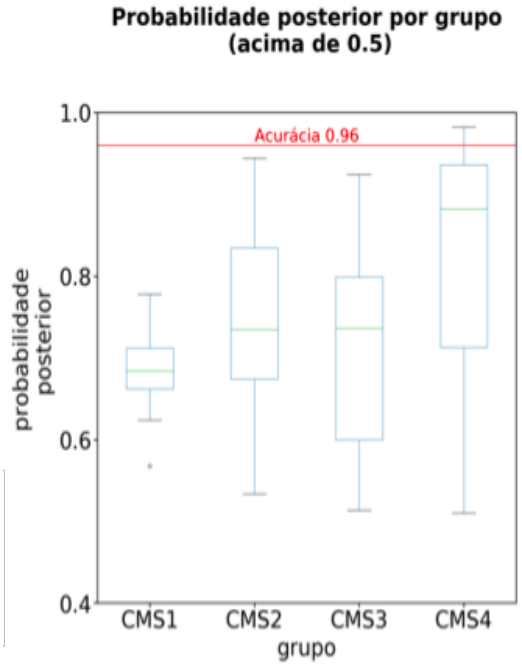


Figura 5.1: Probabilidade posterior do classificador multiômico e acurácia considerando probabilidade posterior igual a 0.50, considerando mesmos parâmetros usados por Guinney [56]

O desempenho do resultado do classificador multiômicos pode ser analisado através da matriz de confusão, mostrada na Figura 5.4. O objetivo desta análise, é verificar se a elevada acurácia do classificador multiômico é referente às classes que estavam verdadeiramente certas ou verdadeiramente equivocadas. A observação desta diferença se insere na premência de se identificar pacientes classificados equivocadamente como pertencentes à classe negativa.

No caso de problemas multi-classe, como o abordado neste estudo de caso, a classe negativa são todas as demais que não a verdadeira. Para o problema de câncer, é interessante observar se o paciente classificado equivocadamente-pertence a uma classe de estadiamento menos grave que a classe que este de fato pertence (classes positivas).

Pode-se observar na Figura 5.4 que o classificador multiômico identificou de forma equivocada pacientes dos grupos CMS1 e CMS2 como sendo do grupo CMS4. Também identificou equivocadamente pacientes do grupo CMS3, como sendo do grupo CMS2. Considerando que no trabalho de Guinney[56] o grupos CMS4 é o que tem o pior prognóstico, a incorreta identificação das amostras dos grupos CMS 1 e 2 como pertencentes ao grupos CMS4 não é tão prejudicial como a incorreta identificação das amostras verdadeiramente pertencentes ao grupo CMS3 como amostras pertencentes ao grupo CMS2, uma vez o grupo CMS2 tem o melhor prognóstico de todos os grupos.

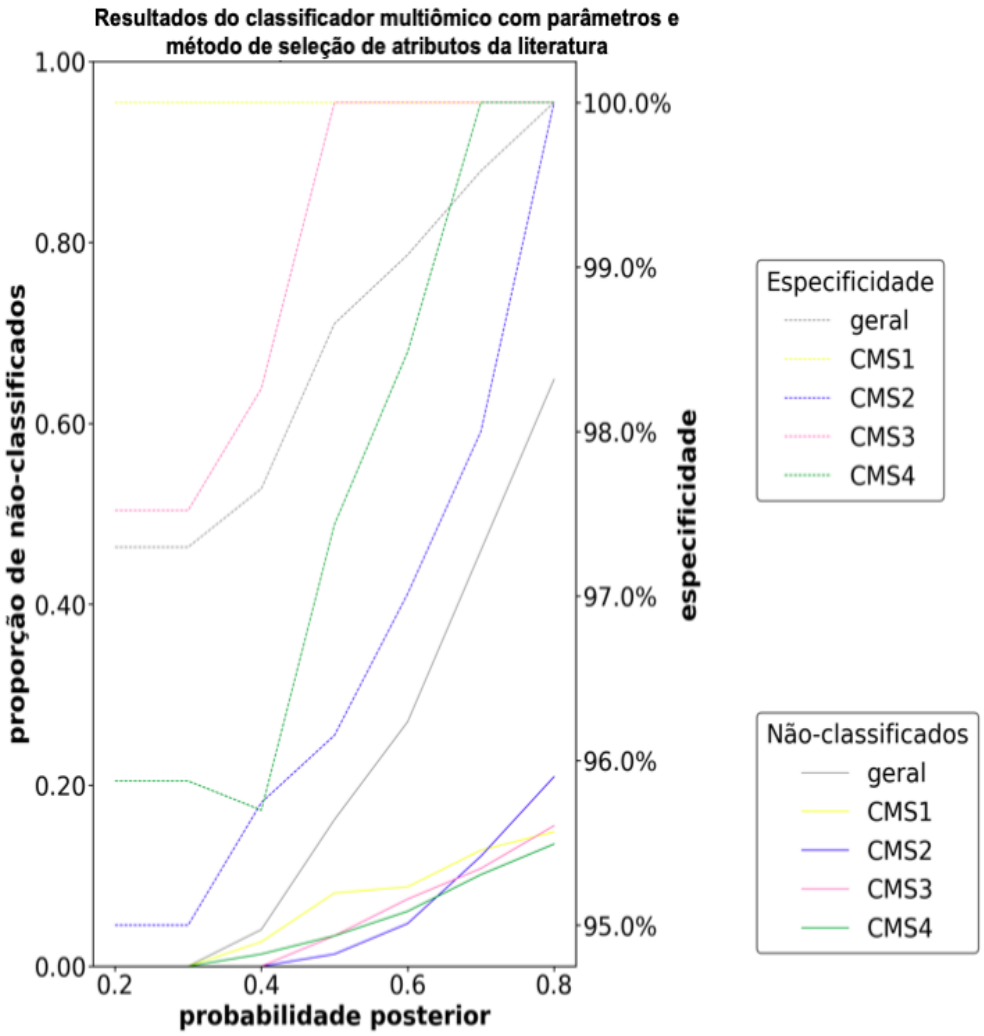


Figura 5.2: Especificidade do classificador multiômico e mesmos parâmetros usados por Guinney[56]

Entretanto, a alta acurácia e melhores resultados de especificidade, indicam que a adição de dados de expressão de miRNA e metilação ao estudo de CCR auxilia no desempenho da classificação de subtipos de CCR. Por isso, as análises subsequentes foram realizadas considerando-se as bases de dados de expressão de mRNA, miRNA e metilação.

5.2.2 Seleção de atributos de dados de CCR

Foram aplicados às bases de dados de expressão de mRNA, miRNa e metilação os algoritmos de pré-seleção de atributos FixedSPEC, GenericSPEC ReliefF.

Estes algoritmos têm como parâmetro de entrada o número de atributos que se deseja selecionar. Estes parâmetros foram definidos para cada conjunto

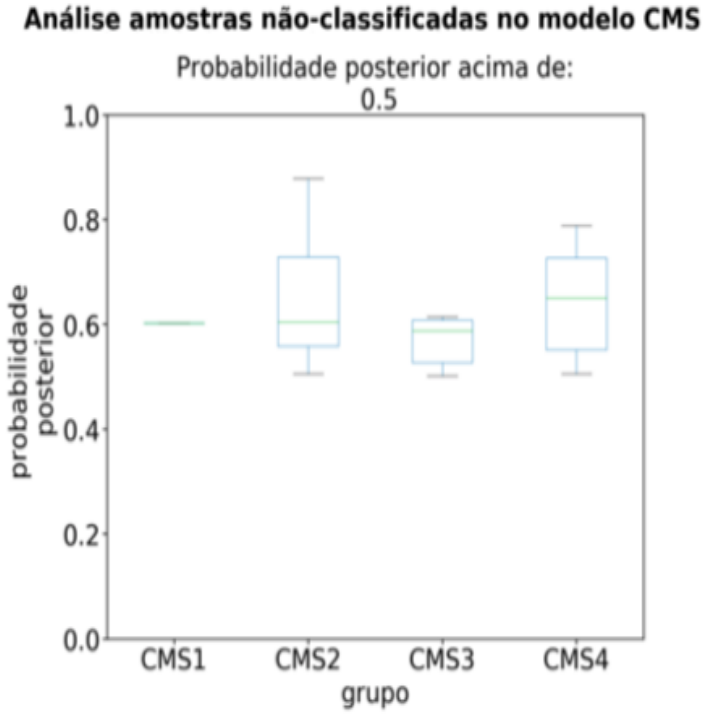


Figura 5.3: Análise da classificação multiômica de amostras não-classificadas no modelo CMS[56] considerando probabilidade posterior igual a 0.50 e mesmos parâmetros usados por Guinney[56]

de dados ômicos com base no trabalho de Liu[60], a saber: 80 atributos para miRNA, 10000 para metilação e 6000 para expressão de mRNA. O algoritmo FixedSPEC usou a mesma quantidade de grupos a ser utilizada no agrupamento realizado na Seção 3.5.

A implementação dos algoritmos FixedSPEC e GenericSPEC usada foi a da biblioteca FSFC[70] em python.

Também foi realizada uma análise de expressão diferencial (LDF) com intuito de se selecionar os atributos de maior variação. A análise de expressão diferencial considerou como genes diferencialmente expressos aqueles com p valor ajustado pelo método de BH menor que 0.001 na análise

O algoritmo ReliefF, por ser supervisionado, foi testado com três tipos de classe-alvo diferentes: dias até a morte/consulta de acompanhamento (DTD); resultado do trabalho de Guinney[56] (CMS) e do trabalho de Liu[64] (Pancancer). Para se realizar a segmentação da classe-alvo DTD, usou-se o algoritmo *k-means*, conforme citado na Seção 4.2.2 e mostrado na Figura 5.5.

Utilizou-se a implementação deste algoritmo da biblioteca scikit-learn 0.23.2, disponível no repositório Pypi[71]. O algoritmo *k-means* separou os dados em dois grupos.

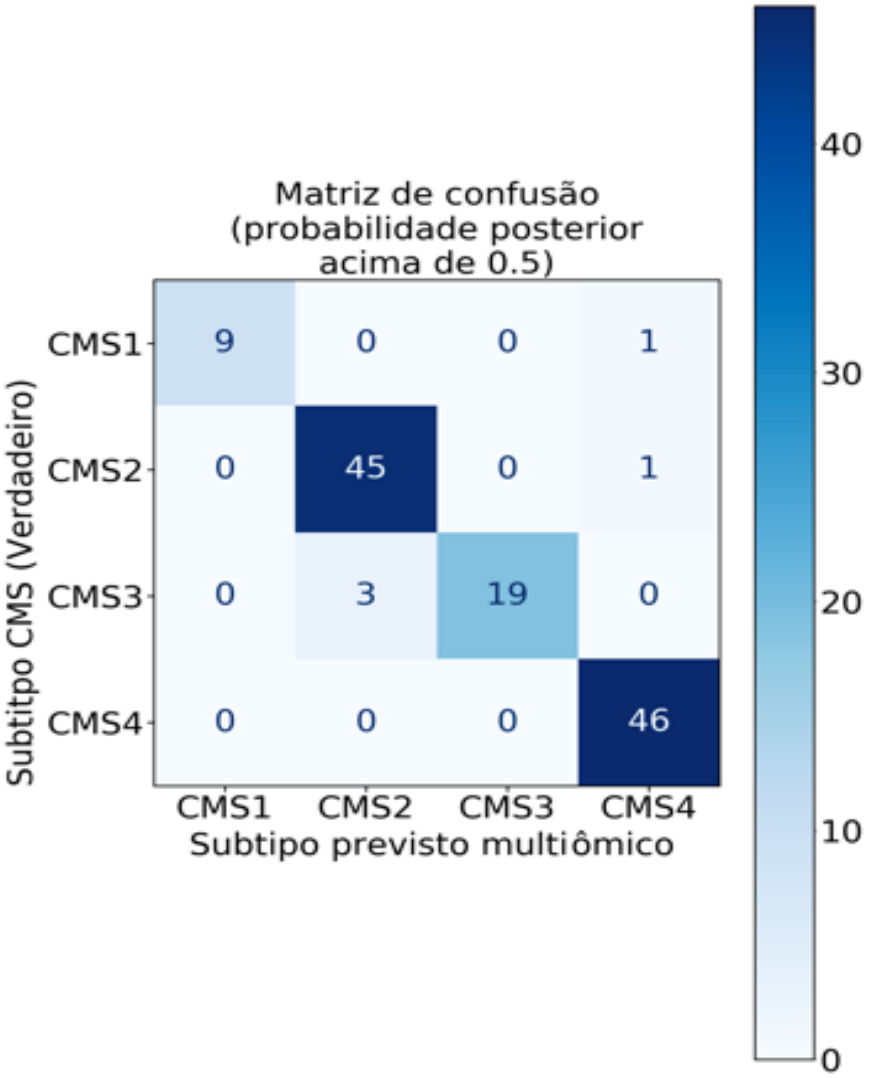


Figura 5.4: Matriz de confusão

O algoritmo ReliefF, utilizou 5 k vizinhos como parâmetro. Este algoritmo foi implementado na versão 0.6 da biblioteca skrebate[72], disponível no repositório Pypi[71].

Os resultados da aplicação destes algoritmos normalizados entre 0 e 1 para os conjuntos de dados de expressão de mRNA, miRNA e metilação são apresentados nas Figuras 5.6, 5.7 e 5.8.

Observa-se, nas Figuras 5.6b e 5.6f, respectivamente, que o algoritmo fixedSPEC, considerando o número de grupos igual a 5, e a análise de expressão diferencial não conseguiram separar os dados de miRNA de forma satisfatória, uma vez que consideraram que a maior parte dos atributos não tinha significância.

Para as demais técnicas, observa-se nas Figuras 5.6a, 5.6c, 5.6d e 5.6e

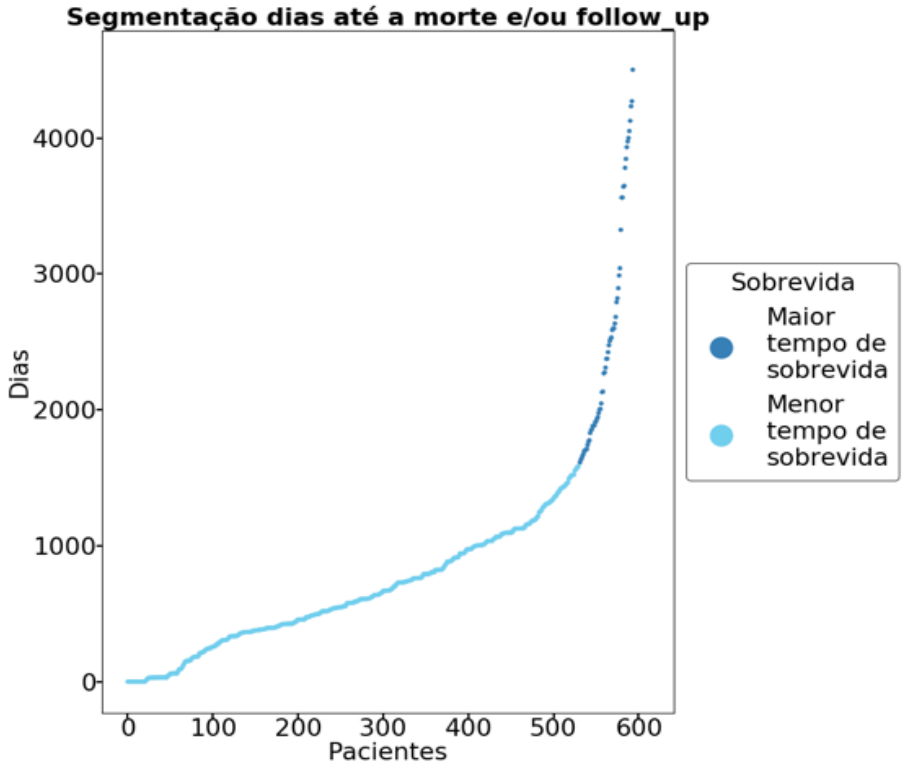


Figura 5.5: Segmentação dias até a morte e/ou follow up

que os métodos Relief, com diferentes classes-alvo, e GenericSPEC conseguiram separar satisfatoriamente os atributos, obtendo-se um *ranking* onde há diferenciação entre os atributos, como mostrado pela inflexão da curva destes *rankings*.

A aplicação do algoritmo GenericSPEC aos dados de metilação, como mostrado na Figura 5.7d, também não conseguiu separar estes dados satisfatoriamente. No entanto, nas Figuras 5.7a, 5.7b, 5.7c, e 5.7e observa-se que os dados de metilação foram separados satisfatoriamente com as técnicas reliefF, para diferentes classes-alvo, e FixedSPEC, gerando um *ranking* onde alguns atributos tem uma pontuação significativamente maior que a dos demais.

Para o conjunto de dados de expressão de mRNA, os resultados das técnicas de seleção de atributos são mostrados na Figura 5.8. Pode-se observar que todas as técnicas conseguiram diferenciar os atributos em sua maior parte, atribuindo a estes pontuações com diferenças significativas no *ranking* de cada método.

Portanto, das análises preliminares, foi possível observar que o ReliefF conseguiu gerar *rankings* que distinguem os atributos das 3 variáveis sem que haja uma excessiva concentração de atributos como melhores ou piores. Também se observou que, ao realizar o agrupamento como na Seção 4.3,

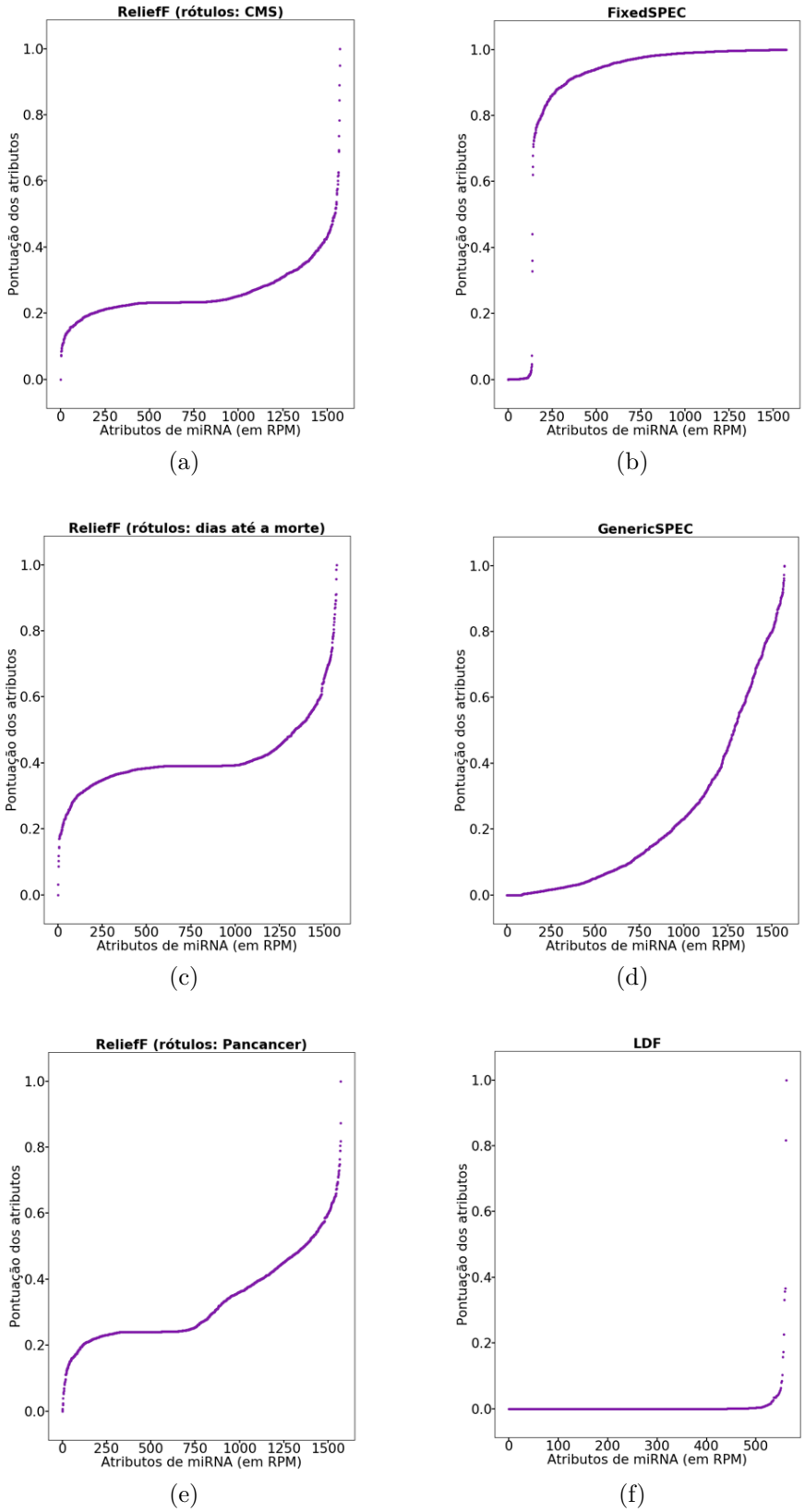
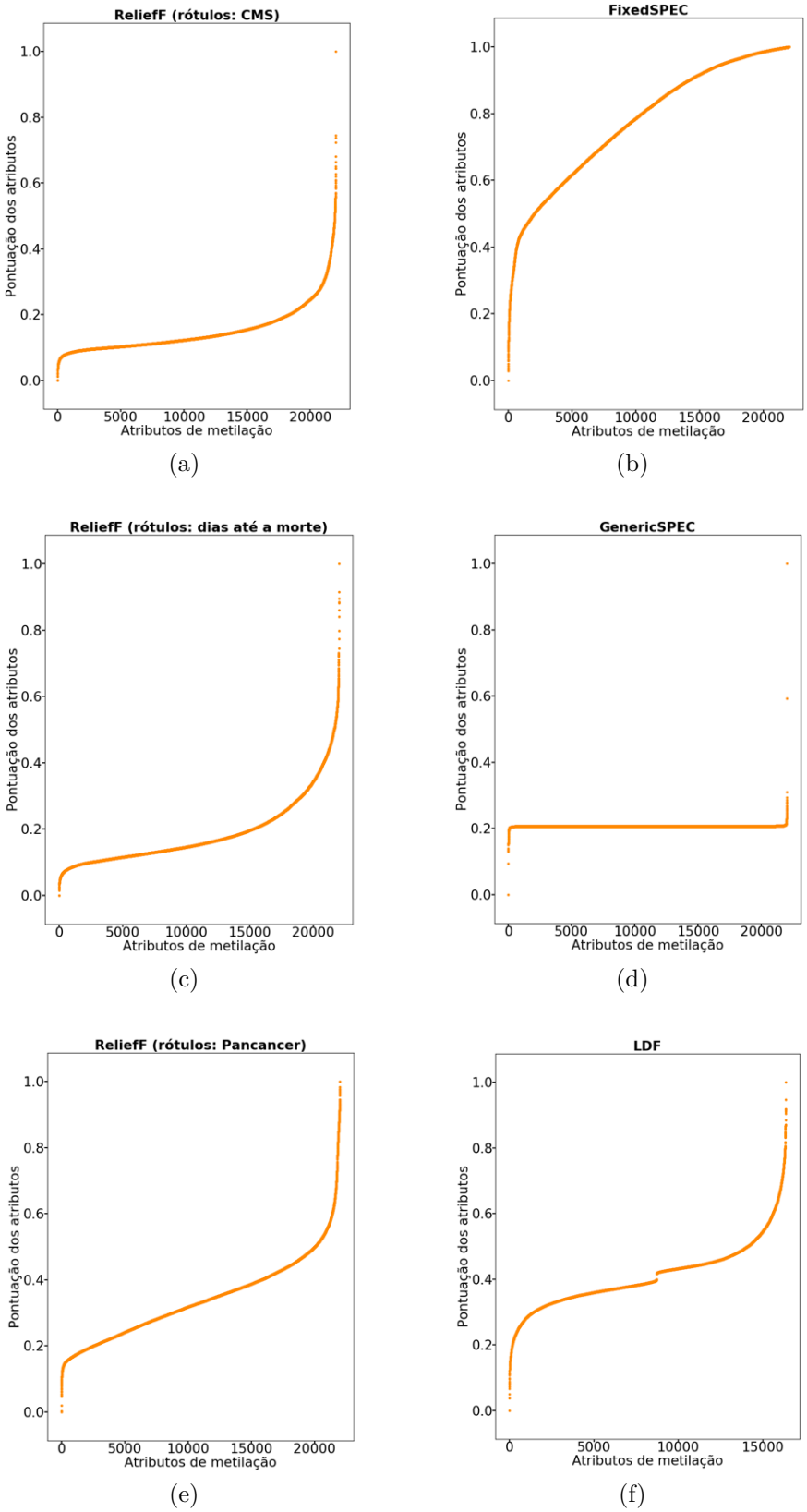


Figura 5.6: Métodos de seleção de atributos para expressão de miRNA, em (a) (e) (c) método reliefF com diferentes rótulos; em (b) (d) métodos baseados em análise espectral, FixedSPEC e GenericSPEC; e em (f) análise de expressão diferencial de genes(LDF)



(a) (b) (c) (d) (e) (f)

Figura 5.7: Métodos de seleção de atributos para metilação, em (a) (e) (c) método reliefF com diferentes rótulos; em (b) (d) métodos baseados em análise espectral, FixedSPEC e GenericSPEC; e em (f) análise de expressão diferencial de genes(LDF)

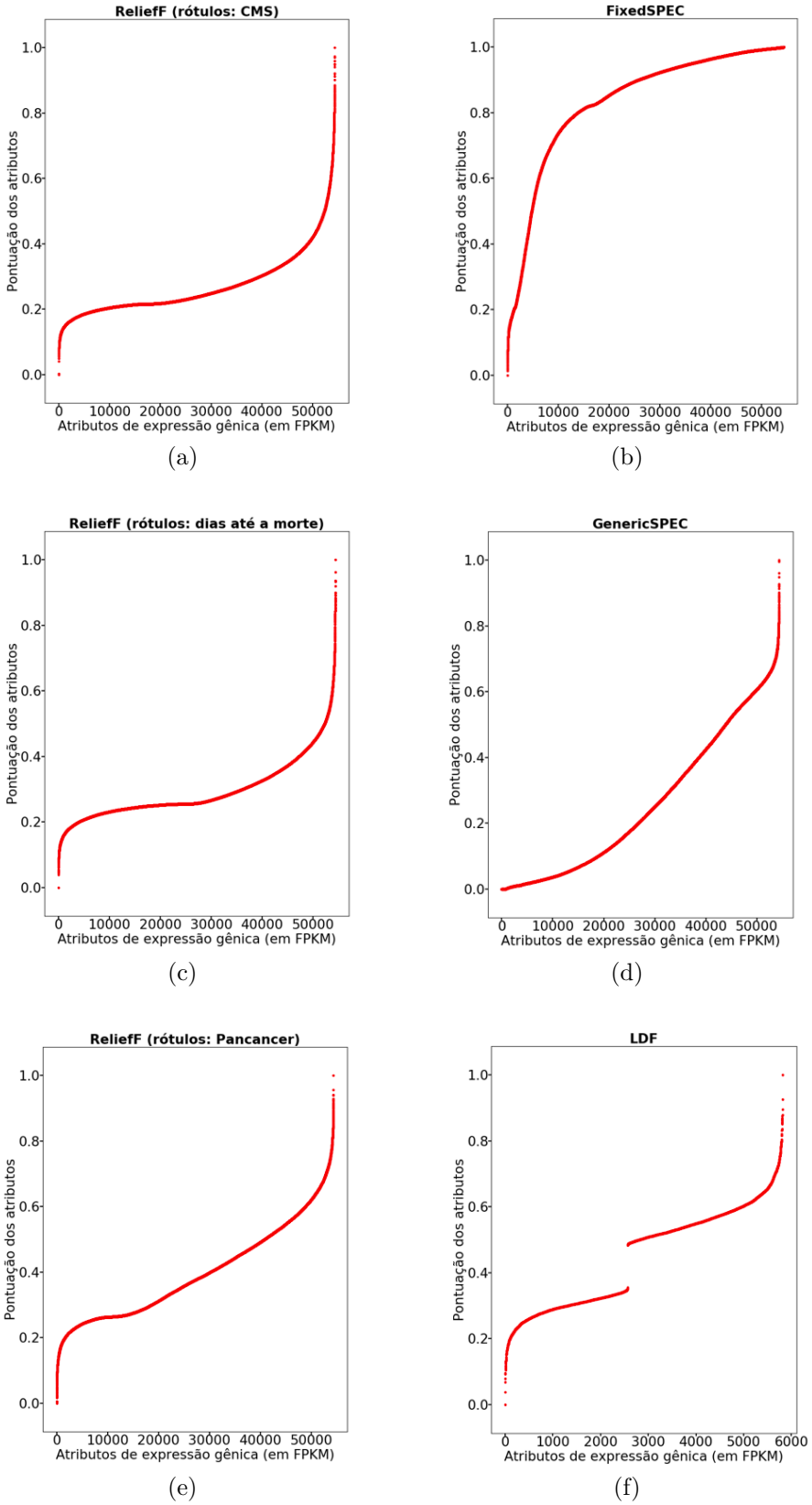


Figura 5.8: Métodos de seleção de atributos para expressão de mRNA, em (a) (e) (c) método reliefF com diferentes rótulos; em (b) (d) métodos baseados em análise espectral, FixedSPEC e GenericSPEC; e em (f) análise de expressão diferencial de genes(LDF)

considerando todos os métodos de seleção de atributos, havia, para todas as combinações dos parâmetros analisados, um grupo com uma quantidade muito pequena de amostras (em torno de 3% no máximo) em um dos grupos. Isto mostra que a seleção de atributos não estava adequada, optando-se por realizar as análises apenas com o método reliefF.

Foi obtido então um *ranking* final com a importância de cada atributo, como mostrado na Figura 5.9. Observa-se que, para os dados de metilação (Figura 5.9a), a curva do *ranking* final tem uma variação dos maiores valores muito maior do que a curva de miRNA (Figura 5.9c) e que a inclinação da curva de expressão de mRNA (Figura 5.9b) se situa de forma intermediária. Portanto, a variável de miRNA não terá tantos atributos selecionados como as demais variáveis.

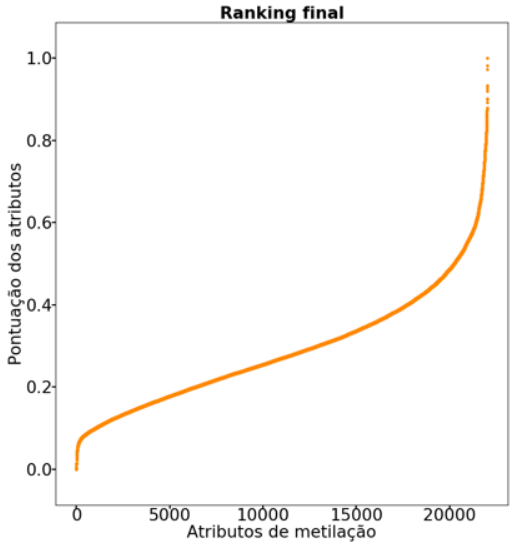
Os dados dos *rankings* obtidos para cada conjunto de dados na Seção 5.2.1 foram separados pelo algoritmo *k-means* (Figura 5.10b, 5.10d e 5.10f) pelo ajuste de uma reta aos dados, conforme descrito na Seção 4.2.2, e que pode ser observado nas Figuras 5.10a, 5.10c e 5.10e.

Contudo, observou-se que os atributos selecionados ainda eram de grande ordem e geraram agrupamentos em que mais de 80% das amostras ficavam concentradas em apenas um grupo. Assim, optou-se por selecionar, de forma heurística, diferentes valores mínimos de pontuação para obter-se o melhor conjunto de atributos.

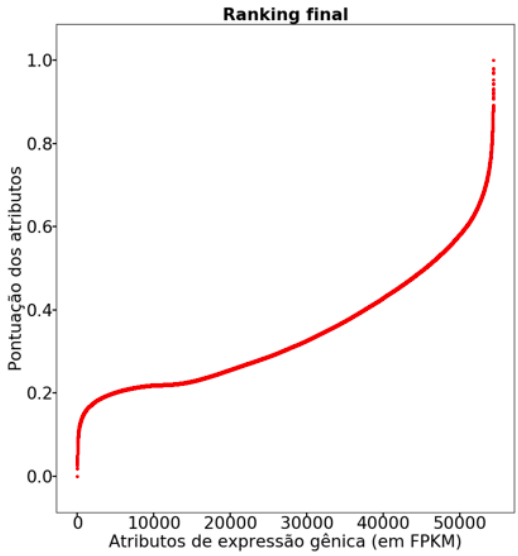
Foi observado que a seleção de atributos de 0.50 até 0.70 de importância no *ranking* resultava em um subdivisão de amostras em que um dos grupos permanecia com um número muito reduzido de amostras, em torno de 2%. Deste modo, optou-se por observar apenas os atributos com importância acima de 0.80 até 0.90 com um incremento de segmentação de 0.25. Portanto, foram observados, para cada combinação de parâmetros, conjuntos de atributos de até 0.80, 0.825, 0.85, 0.875 e 0.90.

Considerando estes intervalo de atributos, foi observado que a concentração de amostras variava entre aproximadamente 30% a 40% no grupo de maior concentração, quantidade de amostras próxima à observada por Guinney[56]. Este comportamento foi observado nas análises considerando-se um valor mínimo para o grau de pertinência de uma amostra ao grupo, conforme elaborado na Seção 4.4, e nas análises realizadas de forma *fuzzy*, isto é, que não consideravam um valor mínimo para o grau de pertinência de amostras aos grupos.

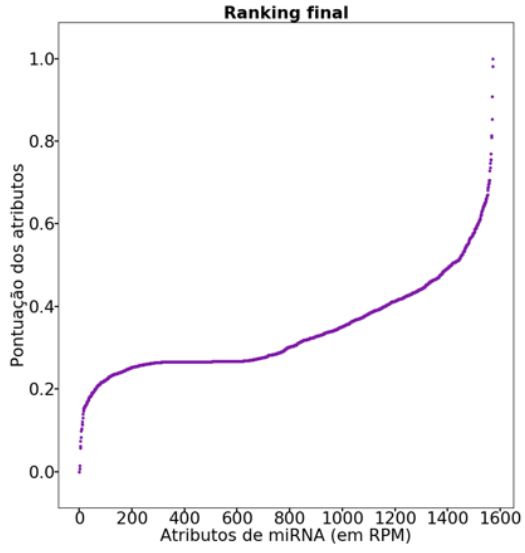
Ao utilizar os métodos de seleção de dados e os parâmetros aqui identificados como mais adequados ao fenótipo estudado para o classificador multiômico, observou-se que a acurácia do classificador foi de 90%, 5 pontos percentuais a



(a)



(b)



(c)

Figura 5.9: *Ranking* final dos métodos de pré-seleção de atributos para metilação, em (a), (b) (c) método reliefF com diferentes rótulos

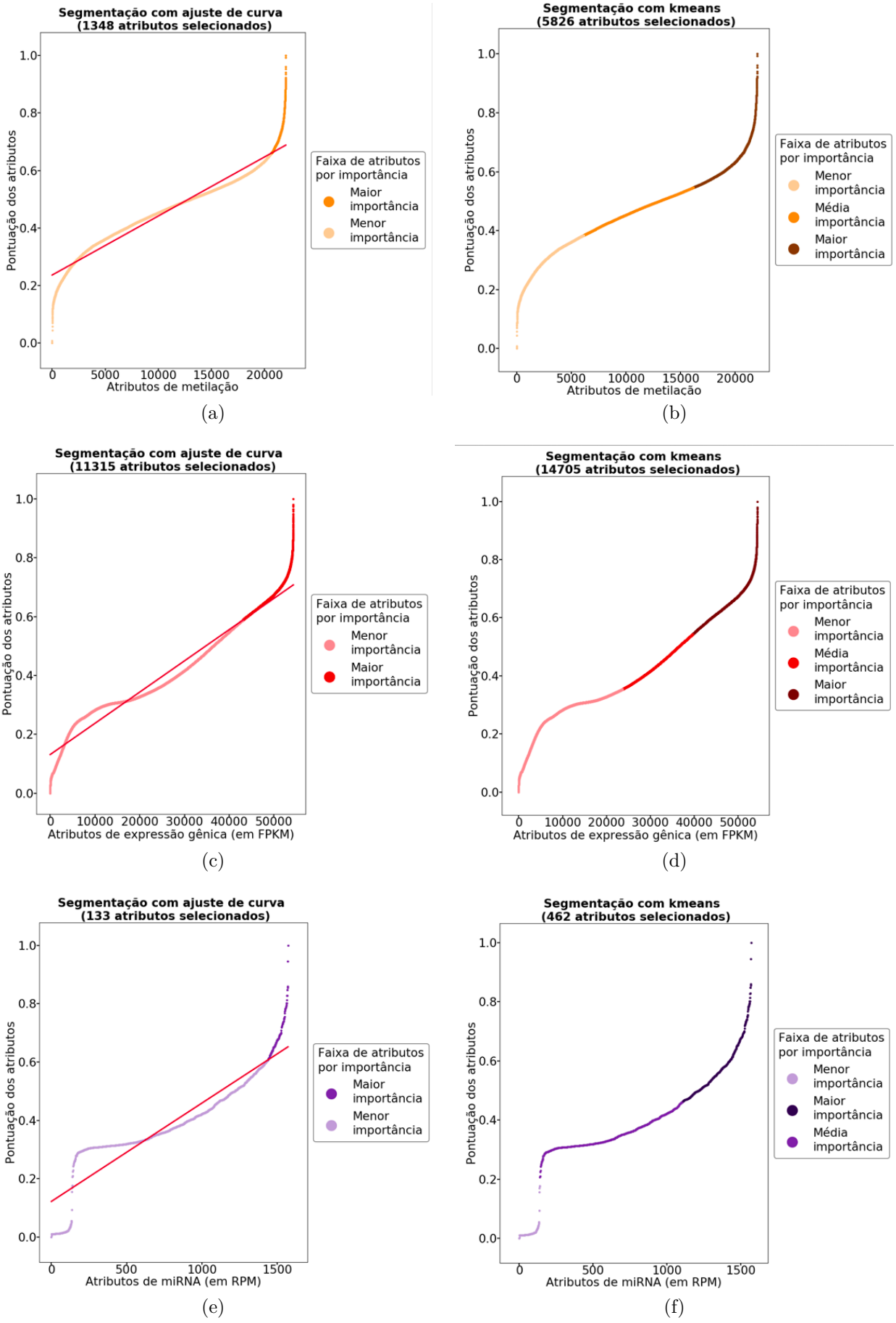


Figura 5.10: *Ranking* final dos métodos de seleção de atributos para metilação, em (a), (b) (c) método reliefF com diferentes rótulos

mais do que a obtida por Guinney[56], como mostrado na Figura 5.11.

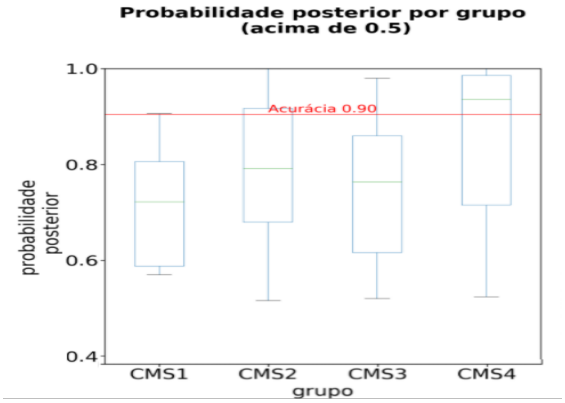


Figura 5.11: Boxplot considerando 0.825 do *ranking* de importância dos atributos e considerando amostras com probabilidade posterior de 0.50

Além da maior acurácia, o classificador multiômico obteve uma proporção menor de amostras não classificadas, 0.081, enquanto a de Guinney[56] foi de 0.13.

Por isso, concluímos que a adição destas variáveis ao estudo do fenótipo citado adiciona informações relevantes sobre a doença.

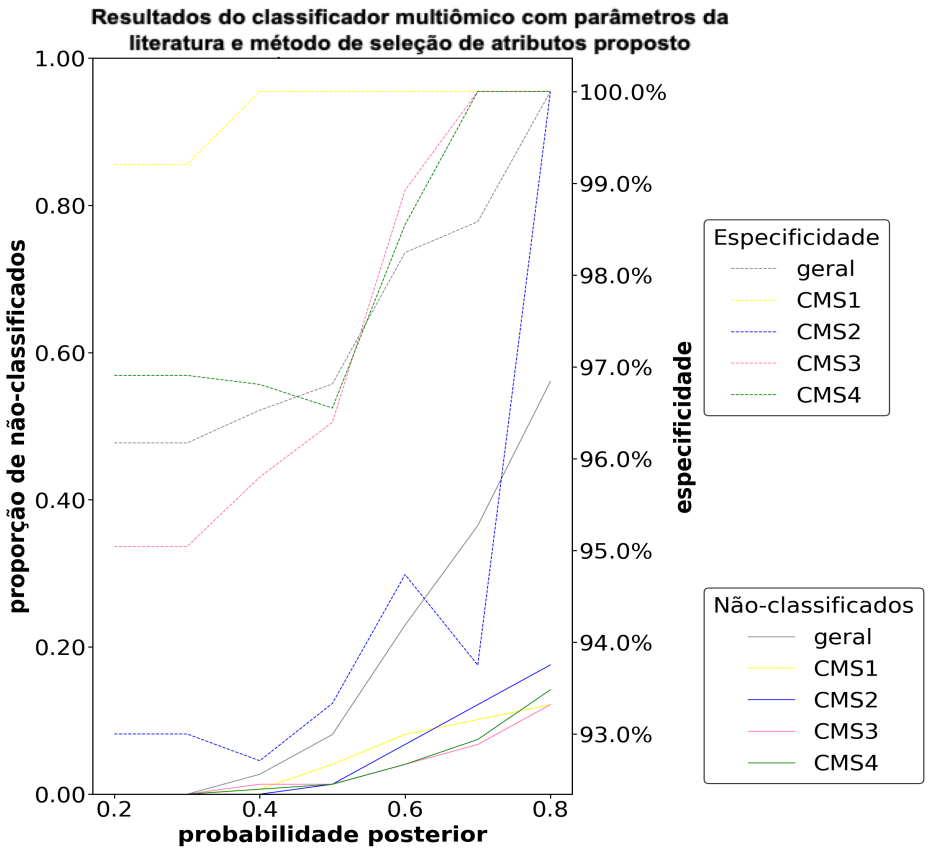


Figura 5.12: Análise da especificidade do classificador considerando 0.825 do *ranking* de importância dos atributos expressão de mRNA, miRNA e metilação.

5.3 Agrupamento *fuzzy* de dados de CCR

O modelo de agrupamento *fuzzy* para a identificação de classes proposto na metodologia deste trabalho teve o número de grupos c variados de 3 a 8. Já para o grau de fuzzificação m, foram testados os valores 1.5, 1.75 e 2, para cada conjunto de atributos mencionado na Seção 3.3.1.

Estes agrupamentos foram caracterizados considerando-se um valor mínimo do grau de pertinência, ou somando-se os graus de pertinência de cada paciente para representar as características dos grupos. Contudo, a consideração do grau de pertinência através desse somatório obteve resultados onde um dos grupos sempre permanecia com uma quantidade muito pequena (em torno de 3%) de amostras. Por isso, foram analisados apenas os resultados considerando um grau de pertinência mínima para cada amostras pertencer a algum grupo.

A análise da curva do tempo de sobrevida foi calculada pelo método de Kaplan-Meier e usada para caracterizar os resultados dos agrupamentos, relacionando as características genotípicas e fenotípicas [25].

Na Figura 5.13 pode-se observar o p-valor global das curvas de sobrevida para cada combinação dos parâmetros do agrupamento e do conjunto de atributos utilizados considerando um grau de pertinência mínimo, como estabelecido na Seção 4.4. Os menores p-valores globais foram obtidos para 5 grupos e m igual a 2, independentemente da quantidade de atributos entre 0.80 e 0.90, sendo o menor destes p-valores, 0.045, obtido utilizando importância mínima de atributos igual a 0.825.

Observa-se que as métricas *fuzzy* mostradas nas Figuras 5.14a e 5.14c indicam uma maior dessas métricas com a variação do número de grupos quando o grau de fuzzificação é igual a 2.

A Figura 5.14b mostra que, até 5 grupos, o grau de fuzzificação não afeta tanto o valor da métrica, pois pode-se observar que estas têm valores similares até 5 grupos. Observa-se ainda que, para mais de 5 grupos, o grau de fuzzificação 2 é o menos afetado pelo número de grupos.

A métrica estabelecida por Xie-Beni[49], da qual se esperava obter uma ponderação entre os resultados das métricas FS, NPC e FPC, mostra que, até 6 grupos, os resultados se mantêm estáveis, tendo seu mínimo (melhor desempenho), para o número de grupos igual a 5.

Considerando os resultados das métricas *fuzzy* com os resultados de p-valor, pode-se concluir que o resultado do p-valor conjuga o resultado de todas as métricas, uma vez que mostra uma tendência de se obter valores relativamente estáveis para o número de grupos de 4 até 6,e que os agrupamentos com

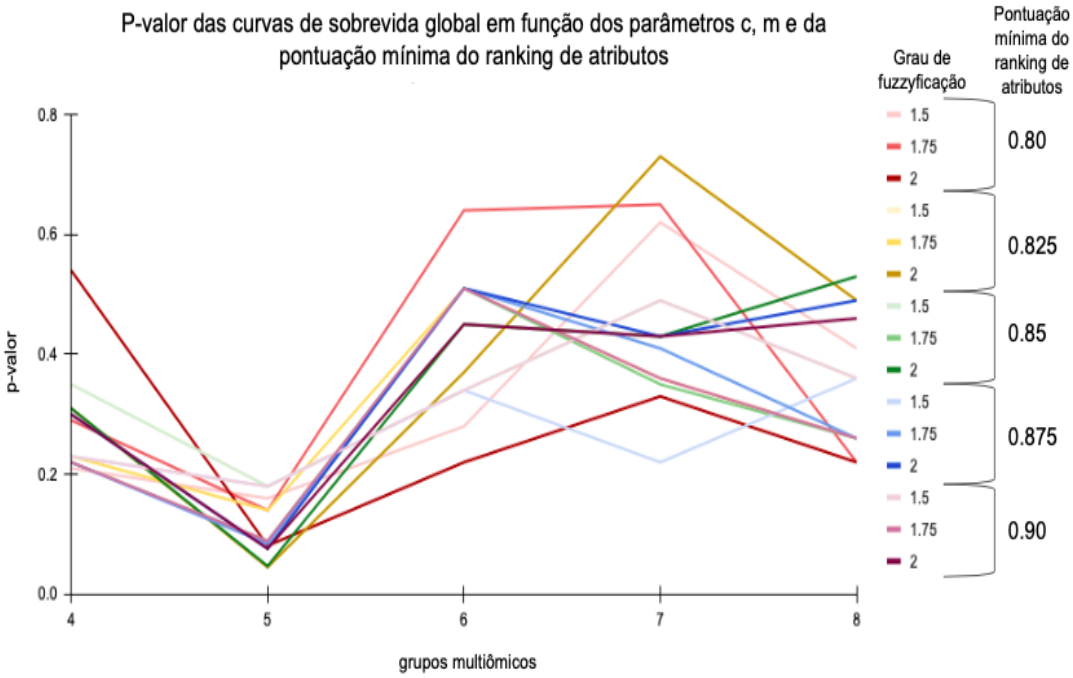


Figura 5.13: P-valor global em relação ao número de grupos, grau de fuzzificação e número de atributos selecionados

número de grupos maiores que 6 indicam perda de aderência das características biológicas. Além disso, dentre os resultados de número de grupos 4, 5 e 6, o número de grupos 5 é o que obtém o desempenho mais adequado. Com relação ao grau de fuzzificação, estes resultados também se corroboram, no sentido de que observa-se que o grau de fuzzificação igual a 2 é o que mais afeta positivamente o desempenho do aumento de grupos, como observa-se na curva para importância de 0.825. Verifica-se também que o p-valor diminuiu mais com o aumento dos grupos de 4 para 5, com m igual a 2, do que com m igual a 1.75.

Por isso, os resultados dos agrupamentos foram avaliados considerando as seguintes combinações de parâmetros: número de grupo igual a 5, grau de fuzzificação igual a 2 e importância mínima do *ranking* de atributos igual a 0.825.

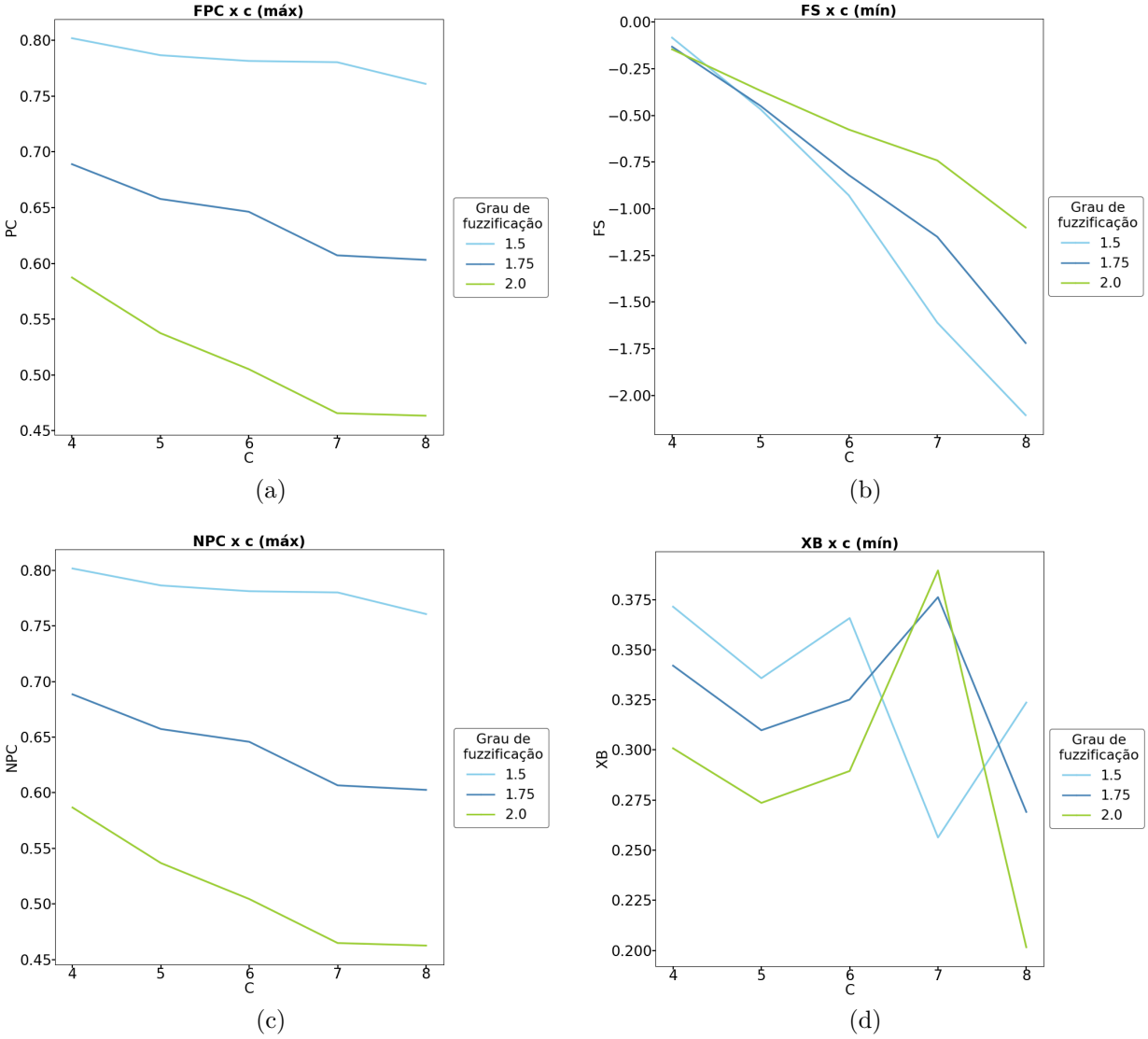


Figura 5.14: Métricas *fuzzy* de avaliação do modelo em relação à variação do número de grupos e ao grau de fuzzificação, (a), (c) (d) método reliefF com diferentes rótulos

5.4 Resultados

A comparação do resultado deste agrupamento com outros agrupamentos da literatura pretende demonstrar a equivalência entre os grupos deste trabalho e da literatura, indicando concordâncias e disparidades.

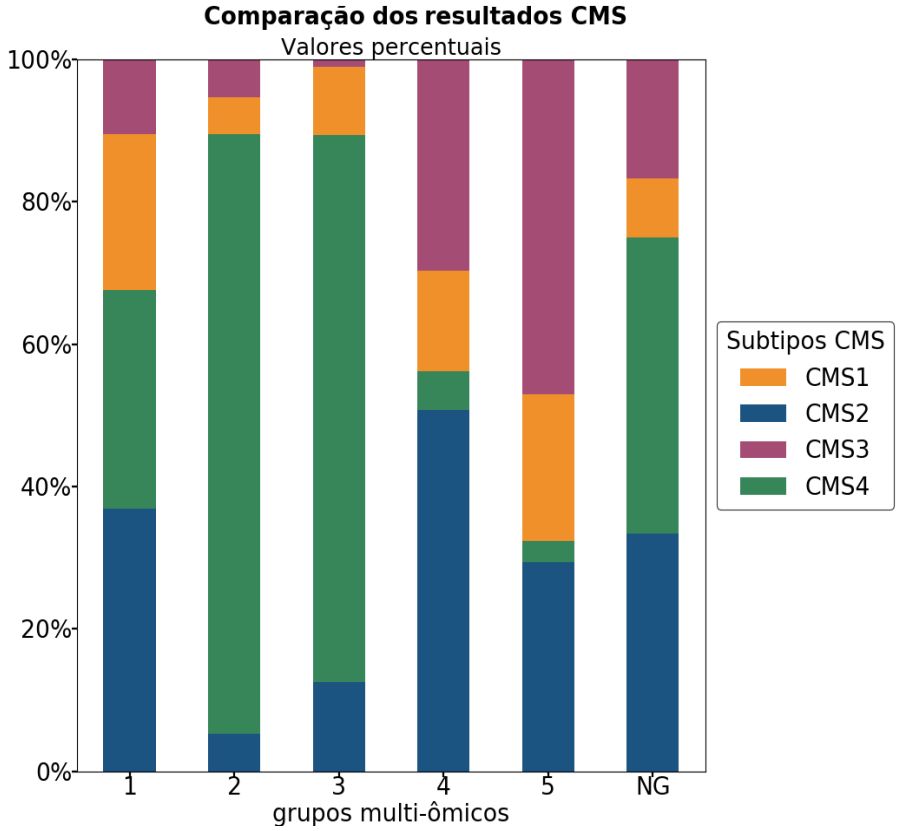
Nas Figuras 5.15a e 5.15b é possível observar que os grupos 2 e 3 multiômicos têm maior identidade com o grupo CMS4, chegando a ter, respectivamente, 84,2% e 76,9% de suas amostras pertencentes ao grupo CMS4. Entretanto, o valor absoluto mostra que a diferença no percentual de amostras do total de cada grupo, 6,8% e 19,9% respectivamente, faz com que a a identidade do grupo 3 com o CMS4 seja mais significativa. Pode-se observar ainda que o grupo 4 da metodologia proposta tem maior identidade com o grupo CMS2. O grupo 5 é formado majoritariamente pelo grupo CMS3. O grupo 1 multiômico tem maior quantidade de amostras do CMS2 e CMS4, com o restante se subdividindo entre o CMS1, e uma menor parte pelo CMS3. De forma semelhante, as amostras consideradas não agrupadas(grupo NG), devido a terem um grau de pertinência menor que o mínimo estabelecido para este número de grupos, também tem mais amostras do CMS2 e CMS4.

A análise PCA, Figura 5.16, dos resultados mostra que o grupo 1 multiômico está entre os grupos 3 e 4 multiômicos. Considerando que o grupo 4 tem maior identidade com o grupo CMS2, o grupo 3 tem maior identidade com o grupo CMS4, e que o grupo 1 tem características compartilhadas entre o CMS 2 e CMS4, fica evidente a gradação entre os grupos observados por esta metodologia.

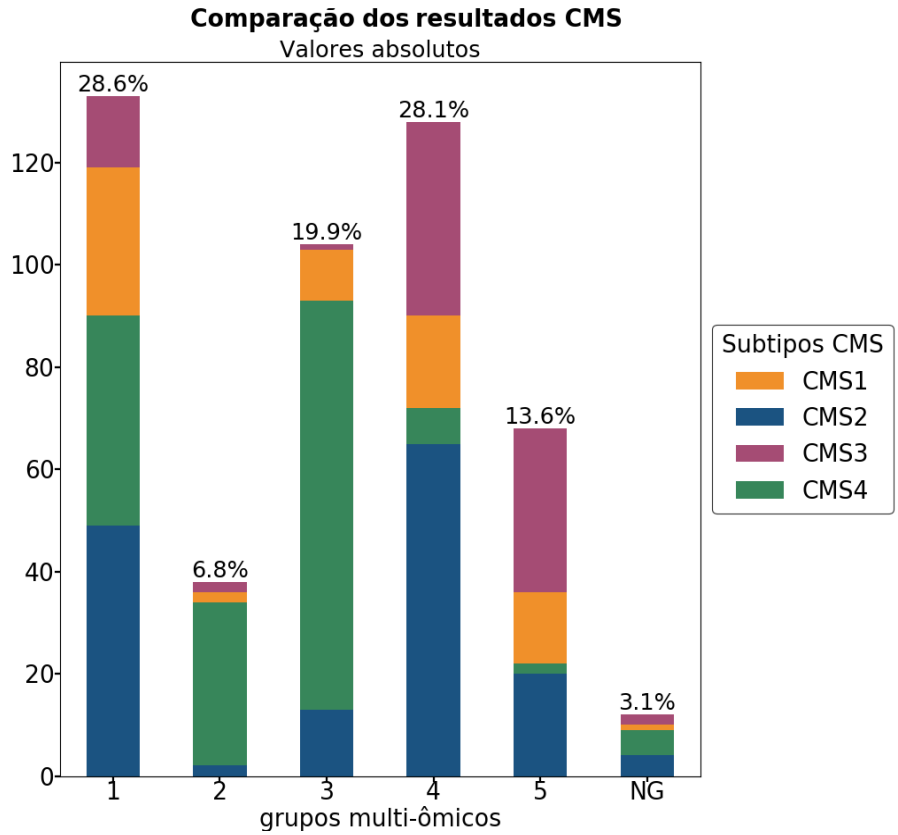
Observando as figuras da curva de sobrevida e da análise de PCA, Figuras 5.17 e 5.16, respectivamente, nota-se que os grupos multiômicos mais distantes entre si, o grupo 2 e 5, têm também suas curvas de sobrevida em posições distintas, isto é, o grupo multiômico 5 tem uma sobrevida maior e o grupo 2 tem uma sobrevida menor.

É interessante notar também que, entre os pacientes não classificados(NG), algumas amostras estão no limiar entre os grupos 1, 3 e 4. Logo, com relação à classificação CMS, estes seriam equivalentes aos CMS2, uma combinação de CMS2 e CMS4, e CMS4 respectivamente. Portanto, era de se esperar que estes tivessem uma curva de sobrevida entre ou próxima aos grupos multiômicos 1, 3 e 4. Contudo, como nem todos os pacientes se situam neste limiar, a curva de sobrevida das amostras não agrupadas (NG) tem os menores p-valores par-a-par considerando as demais curvas de sobrevida, como pode ser observado na Figura 5.17.

Observa-se ainda que a sobrevida do grupo NG difere da sobrevida do



(a)



(b)

Figura 5.15: Comparação dos resultados do agrupamento multiômico com trabalho de Guinney[56], em (a)valores percentuais e (b)absolutos

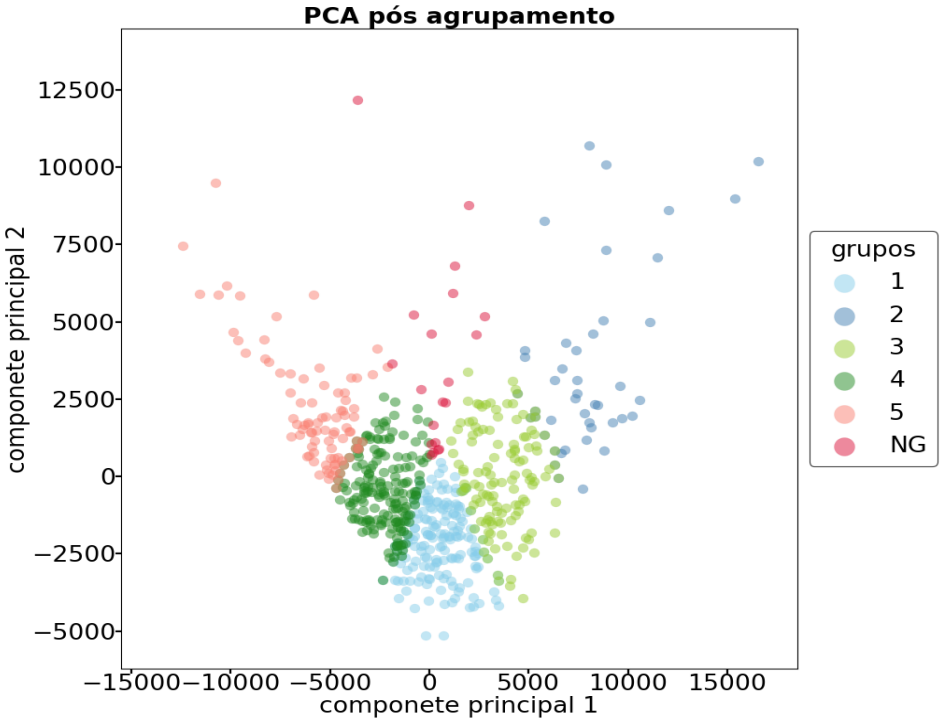


Figura 5.16: Análise PCA das amostras analisadas ressaltando o pertencimento de cada uma a um grupo distinto

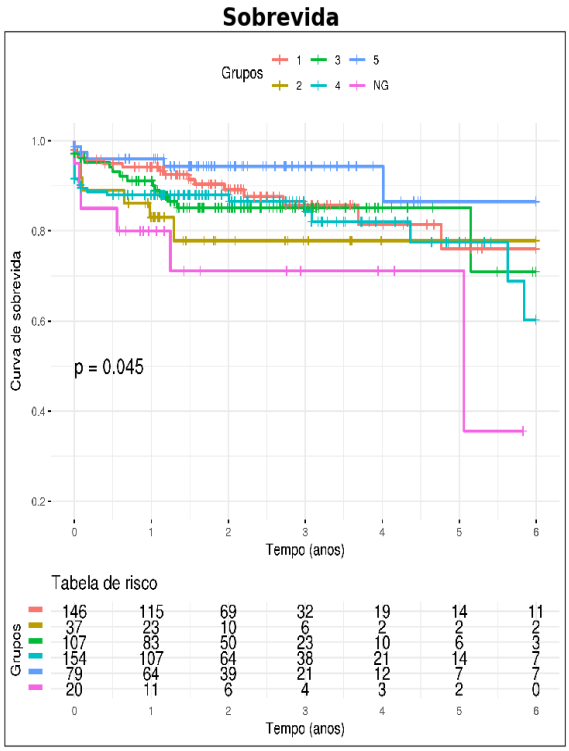


Figura 5.17: Curva de sobrevida obtida com os parâmetros: c=5, m=2, importância mínima da pontuação dos atributos = 0.825.

grupo 5 multiômico. Observando a análise do PCA, nota-se que estas se situam no mesmo patamar considerando a componente principal 2, e diferem apenas na componente 1, sendo o grupo 5 negativo e o grupo NG próximo a zero ou positivo para os valores da componente principal 1. Logo, pode-se concluir que os atributos agrupados pela componente 1 contribuem de forma efetiva para separar as curvas de sobrevida. A análise EVR (explained variance ratio), Figura 5.18, do PCA mostra que os atributos agrupados por essa dimensão correspondem a 71,8% da variância.

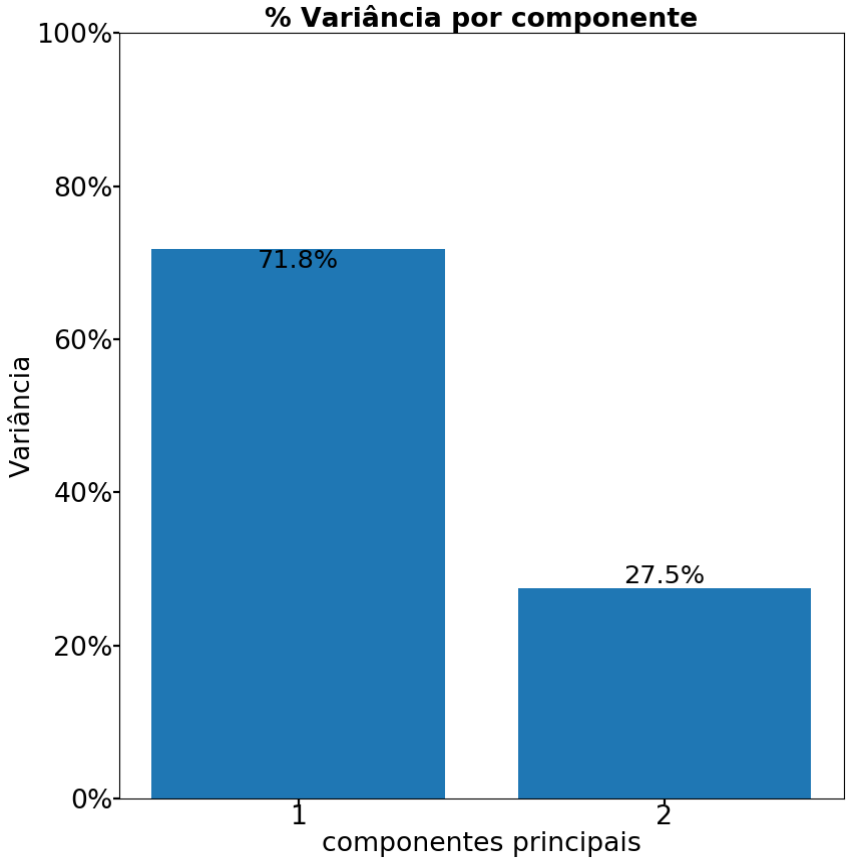


Figura 5.18: Percentagem explicada por cada componente da análise PCA

Para se avaliar os resultados com relação à mutação, ilustrados na Figura 5.19, devemos considerar a prevalência de mutações no gene RAS no câncer colorretal, conforme observado por Vogelstein e outros [73]. Estes indicaram que a exclusão do gene Deleted in Colon Cancer (DCC) junto com o locus DCC envolvendo uma porção do cromossomo 18q foi acompanhada por suscetibilidade ao carcinoma colorretal não polipose hereditário através de uma análise de parentesco[74].

As mutações do gene p53 também são associadas à neoplasia colorretal por meio da inativação da função supressora do tumor do gene *wild-type* p53[75]. A presença e alta prevalência da mutação da polipose coli adenomatosa

(APC) em pacientes com polipose coli familiar (FAP) e CCR esporádico também são reconhecidas[76].

Esses estudos resultaram na observação de que a progressão de lesões moleculares acumuladas ocasionava câncer colorretal ou "Vogelgram" inicial[77]. Estes e outros estudos corroboraram na associação destas e outras mutações em um grupo de mutações denominadas *drivers* do CCR, de acordo com o estudo do trabalho de Liu[64].

Por isso, a incidência de mutação nos grupos deste trabalho será analisada com relação as 11 mutações definidas por Liu[64].

O grupo CMS1 é associado a BRAF e hipermutação. No agrupamento *fuzzy*, essa mutação pode ser observada em valores absolutos no grupo multiômico 1 que tem maior identidade com o CMS1 em valores absolutos.

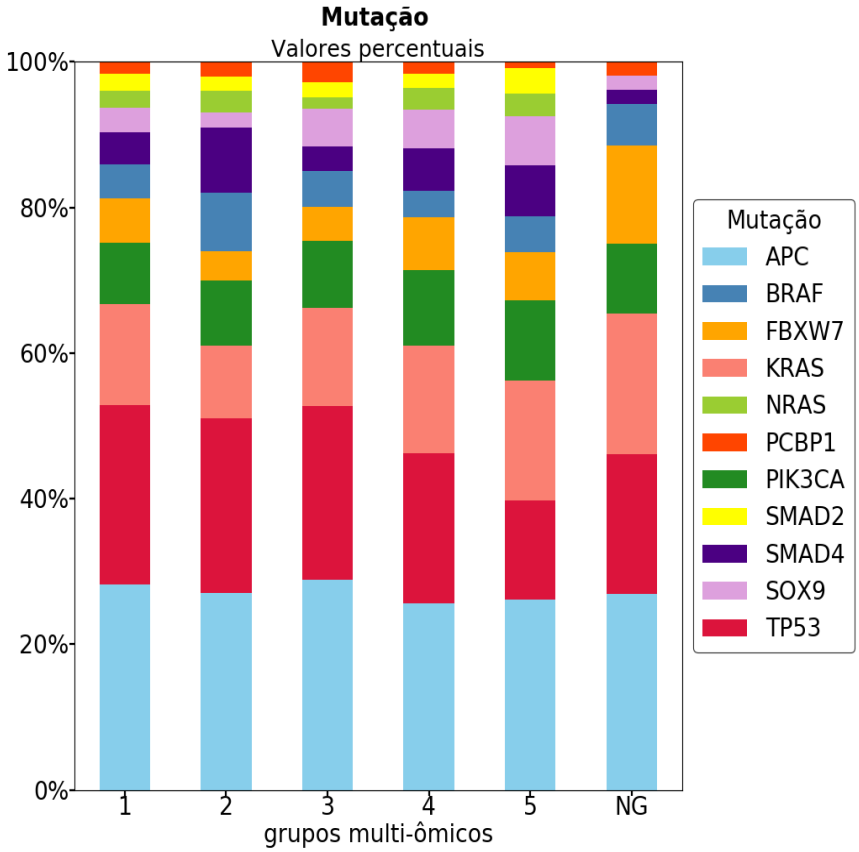
Observa-se ainda que o CMS 3, equivalente ao grupo multiômico 5, tem maiores valores de KRAS percentualmente (16,4%). Porém, em valores absolutos, uma vez que o grupo 5 tem apenas 13,6% das amostras, este fica com menor quantidade de KRAS que o grupo multiômico 4 em valor absolutos. Entretanto, observa-se que todos os grupos, exceto o grupo multiômico 2, tem percentual próximo de KRAS.

Com relação à mutação APC, o trabalho de Guinney[56] aponta que todos os grupos têm a presença desta mutação/ Porém, o CMS2 é o que se espera ter maior quantidade proporcional. Em nosso trabalho, os grupos com mais identidade com o grupo CMS2 é o grupo multiômico 4, porém não é este o grupo com maior nível de APC.

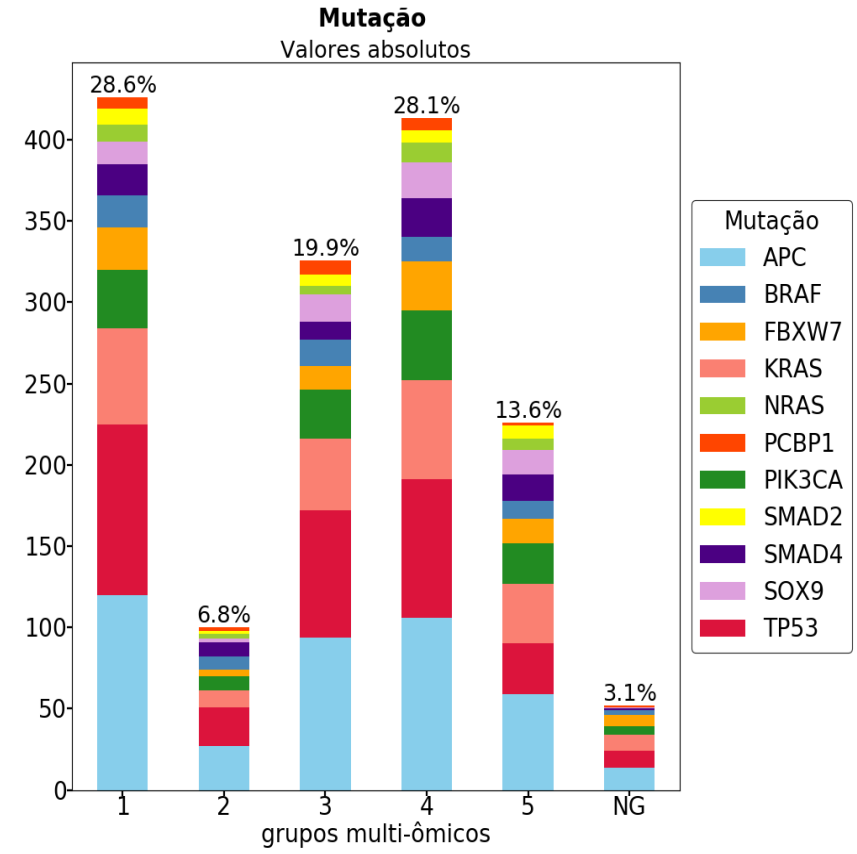
Com relação à mutação TP53, no resultado do agrupamento multiômico, os grupos 1, 2 e 3 apresentaram os maiores níveis de TP53, sendo 2 e 3 relacionados aos CMS4, e o 1 sendo uma combinação entre o CMS4 e 2.

Guinney[56] observou que o CMS1 é o que tem mais pacientes do sexo feminino proporcionalmente e que estas têm mais lesões no lado direito, e de maior grau histopatológico. Esta observação foi verificada em parte nos resultados do agrupamento multiômico. Os grupos 1 e 5 são os que mais têm amostras do CMS1. Os grupos 3 e 4 têm um percentual maior da sua composição de mulheres. Em termos absolutos, o grupo 1 tem um quantitativo próximo ao grupo 4 de pacientes do sexo feminino.

Com relação ao local de resseção ou biópsia, os dados foram agrupados em direito (ceco, cólon ascendente, flexão hepática e cólon transverso), esquerdo (flexura esplênica, colon sigmóide e descendente) e Reto. O grupo CMS2 tem maior quantidade de amostras do lado esquerdo como local de resseção. O CM2 está contido nos nossos grupos 3 e 4, que têm como local de resseção o lado esquerdo em maiores valores absolutos e percentuais, conforme observado

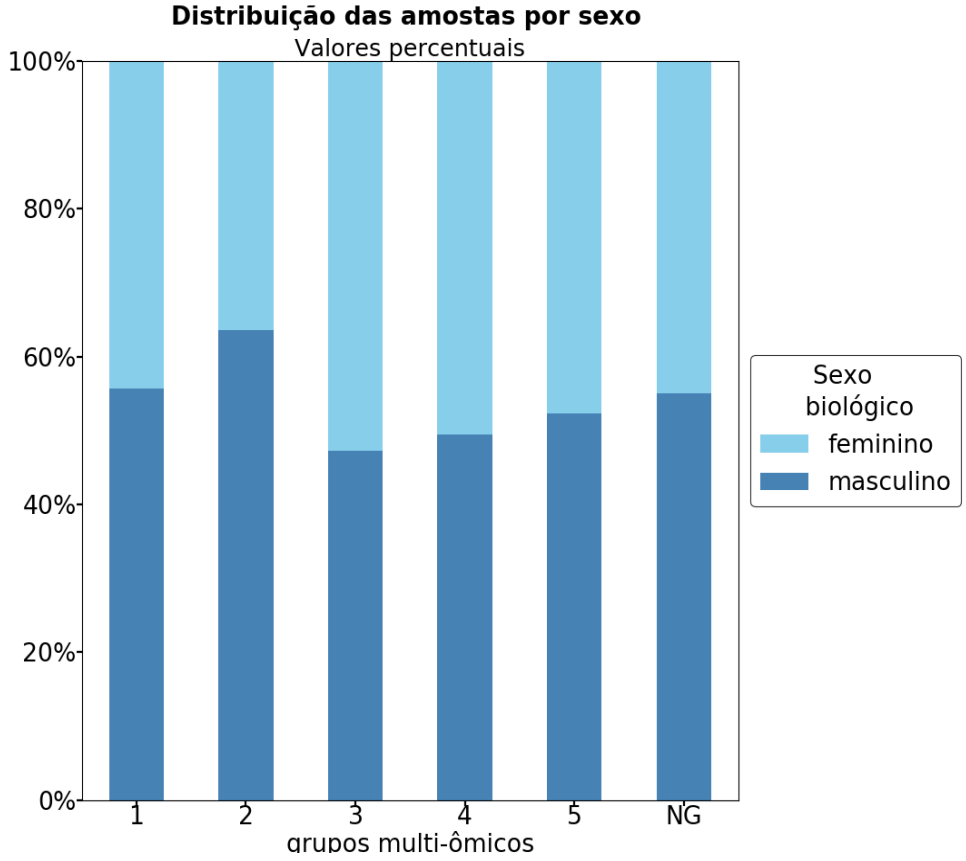


(a)

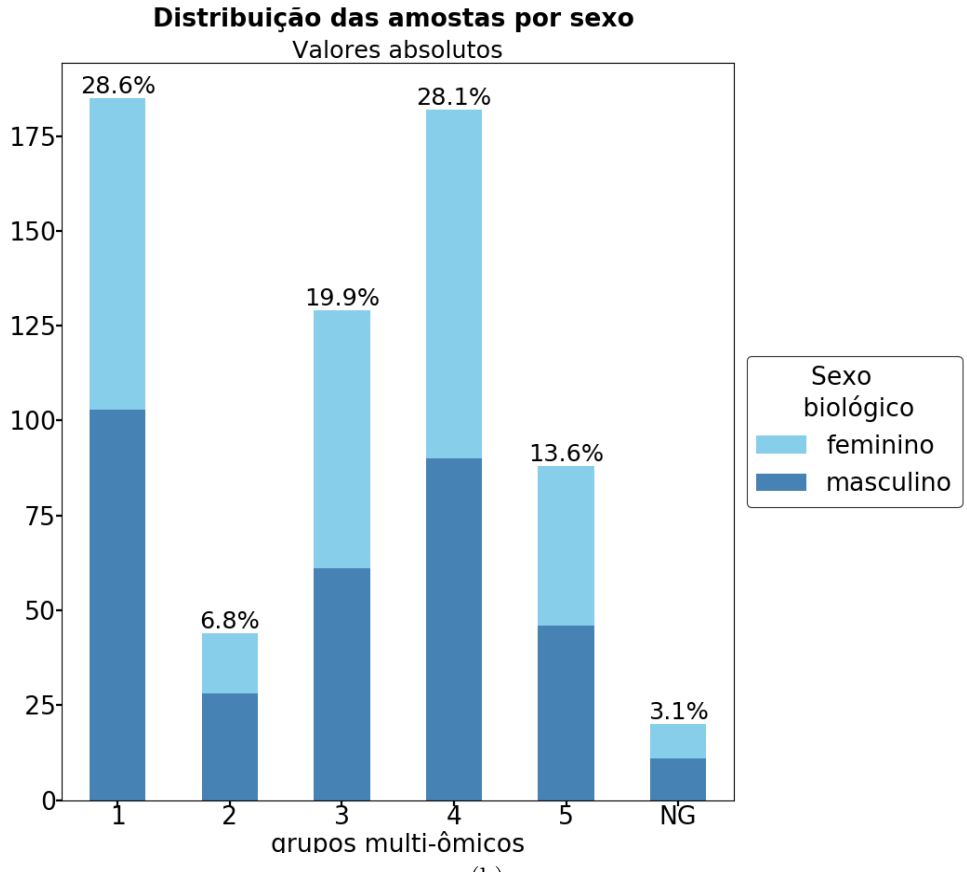


(b)

Figura 5.19: Caracterização dos resultados do agrupamento multiômico por mutação em (a)valores percentuais e (b)absolutos



(a)



(b)

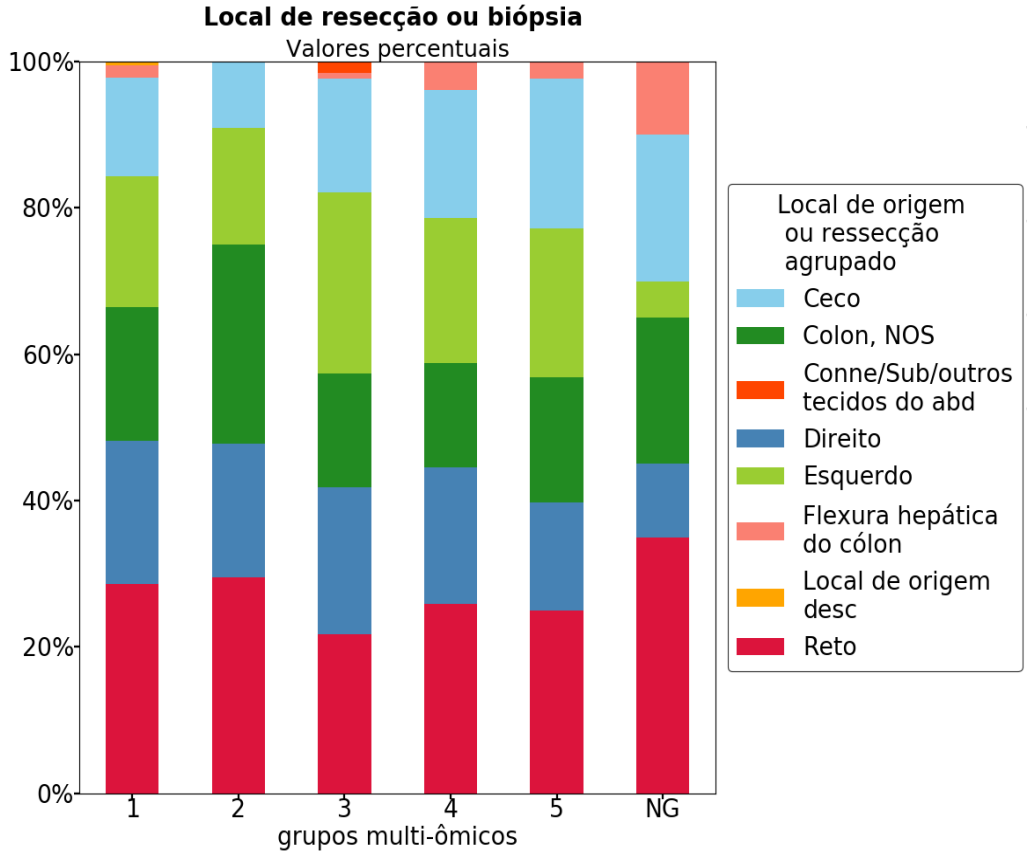
Figura 5.20: Caracterização dos resultados do agrupamento multiômico por sexo biológico em (a)valores percentuais e (b)absolutos

na Figura 5.21.

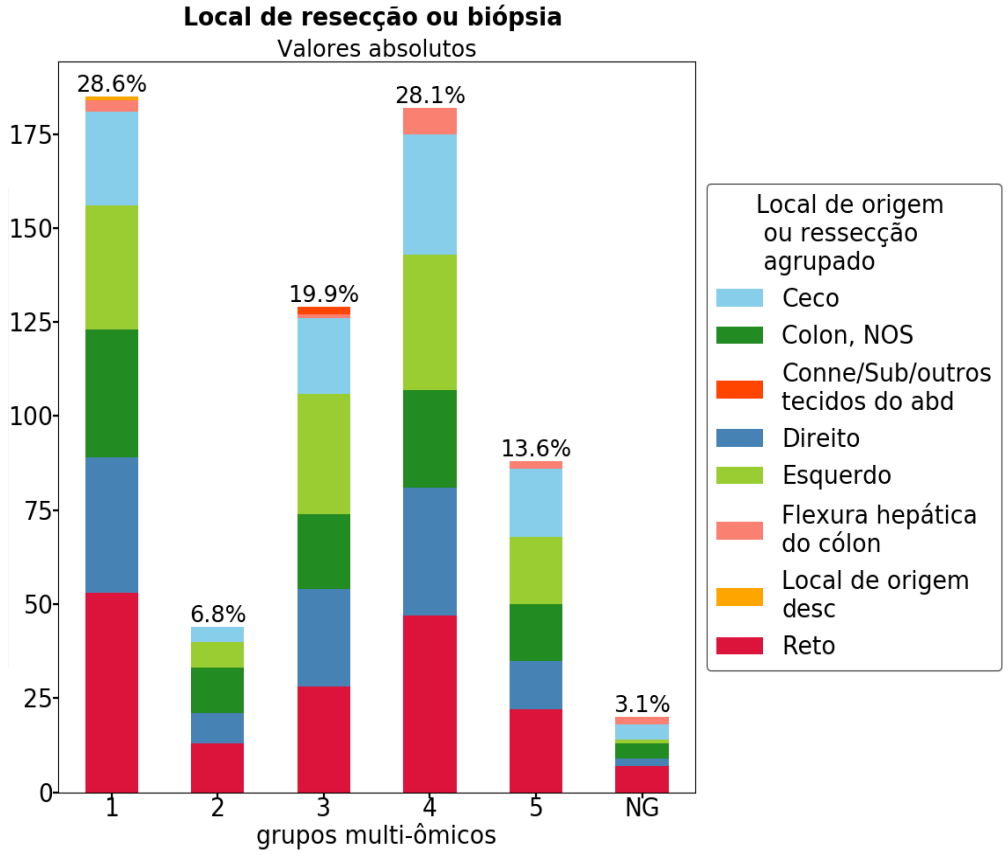
Com relação ao estágio da doença, Guinney observou que os estágios III e IV estavam associados ao subtipo CMS4. Entretanto, não foi identificada essa relação em nosso resultado, conforme observado na Figura 5.22.

Observa-se que as amostras dos grupos 1 e 4 têm maior aderência, isto é, maior média dos graus de pertinência a seus grupos. Já os grupos, 2, 3 e 5 têm uma maior quantidade de *outliers*, o que diminui a média de cada grupo.

Esta característica também é mostrada no gráfico *upset* dos graus de pertinência na Figura 5.24a, ou considerando o valor mínimo (0.36) para este como mostrado na Figura 5.24b.

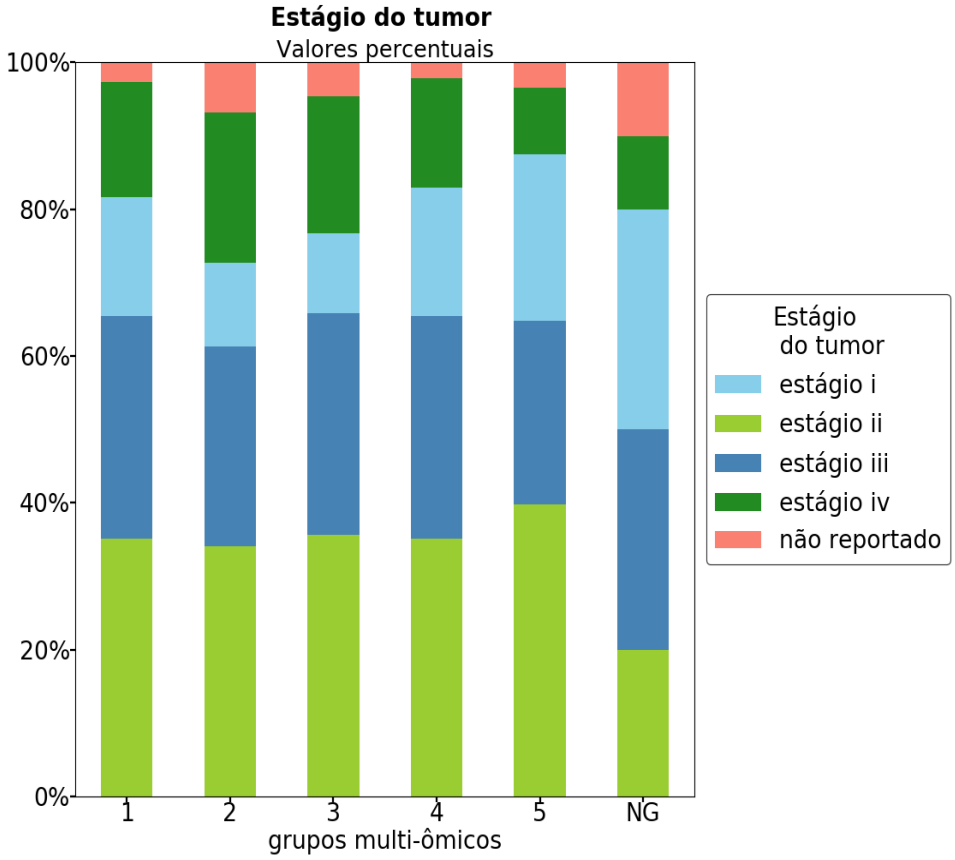


(a)

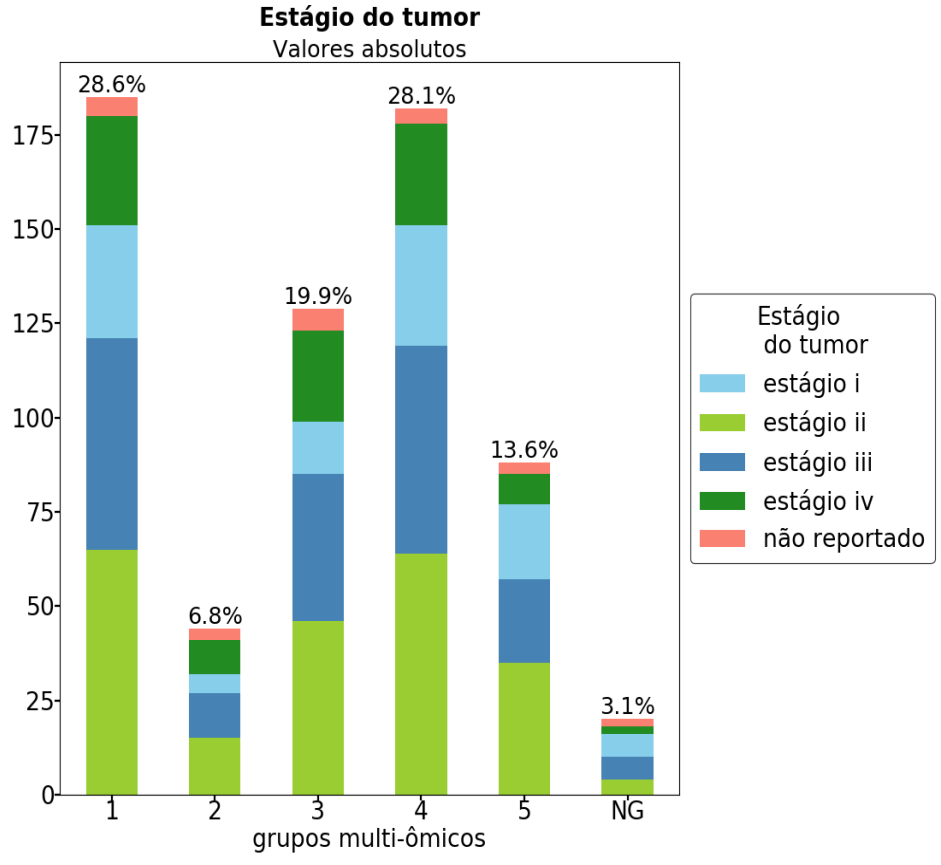


(b)

Figura 5.21: Caracterização dos resultados do agrupamento multiômico por local de resseção e biópsia agrupados (em direito/esquerdo) por (a)valores percentuais e (b)absolutos

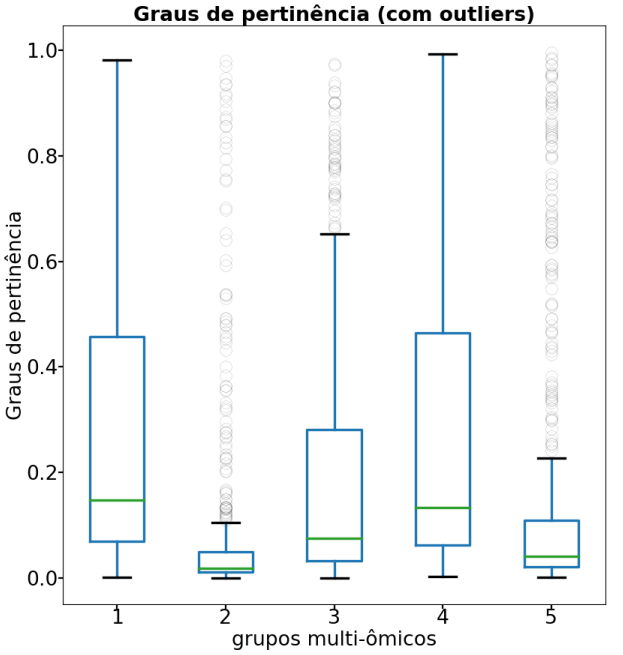


(a)

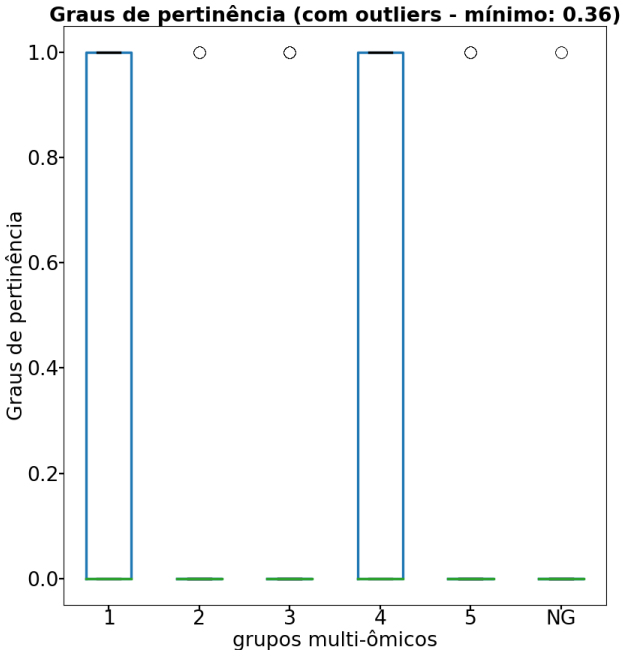


(b)

Figura 5.22: Caracterização por estadiamento da doença em (a)valores percentuais e (b)absolutos

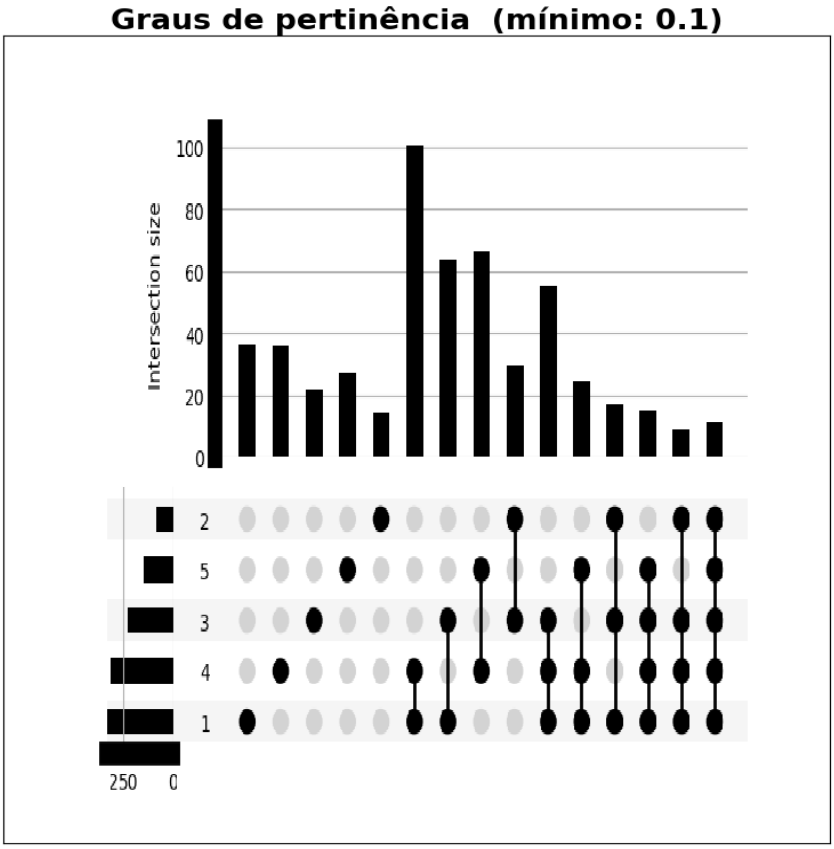


(a)

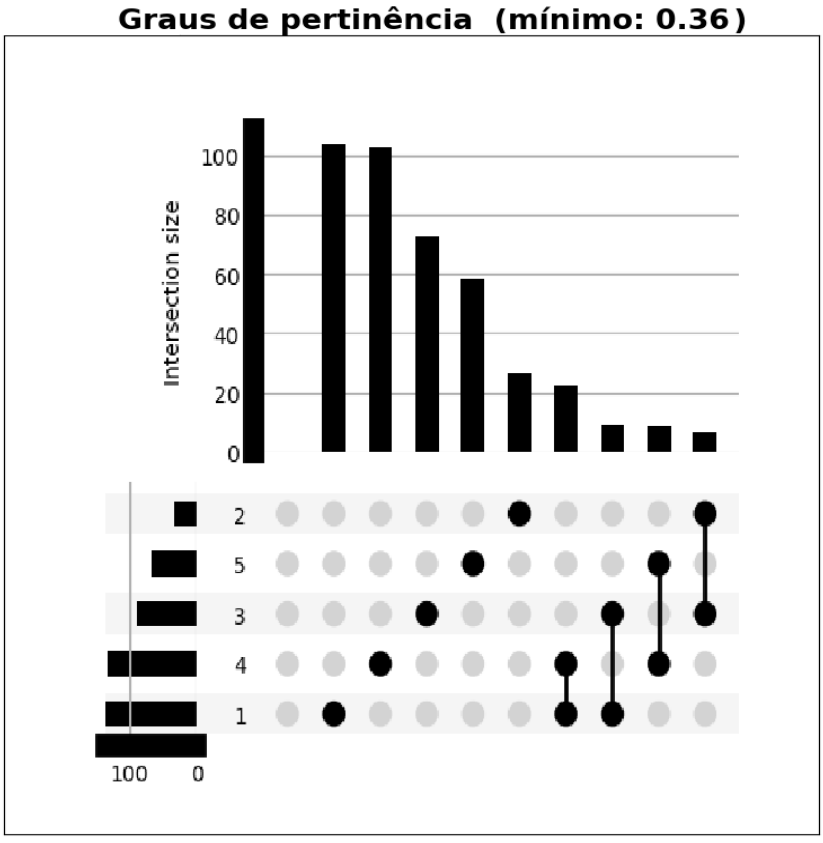


(b)

Figura 5.23: Boxplot dos graus de pertinência (a) sem e (b) com valor mínimo para pertencimento aos grupos



(a)



(b)

Figura 5.24: Gráfico upset do grau de pertinência (a) sem e (b) com valor mínimo para pertencimento aos grupos

6
Conclusões e trabalhos futuros

Este trabalho teve como objetivo o desenvolvimento de uma metodologia para se agrupar dados multiômicos permitindo a identificação da relação entre subtipos moleculares e o fenótipo estudado através da combinação de técnicas de *machine learning*, seleção de atributos e agrupamento *fuzzy C-means*. Além disso, foi realizado um estudo de caso aplicando-se esta metodologia ao estudo de dados moleculares de câncer colorretal.

Os resultados da etapa de pré-processamento indicaram que a adição de diferentes conjuntos de dados moleculares pode auxiliar no melhor entendimento da relação genótipo-fenótipo, também foi possível observar que o quantitativo de genes foi significativamente menor do que o que vem sendo utilizado na literatura.

Destaca-se ainda que dos resultados obtidos da etapa de pré-processamento, pode-se observar que ainda há espaço para se refinar o método de seleção da importância mínima dos atributos a ser considerada. É também importante observar que as técnicas de seleção de atributos se comportaram de forma distinta para as distintas variáveis, caracterizando o comportamento diferente de cada ômica. Portanto, pode-se ainda estudar a escolha do método de seleção de forma individualizada para cada ômica, tornando mais precisa a seleção de atributos.

Com relação à etapa de agrupamento, a técnica utilizada, *fuzzy C-means*, conseguiu evidenciar que um número elevado de amostras, que seria considerada apenas "não-classificada"por métodos rígidos de agrupamento, encontra-se na fronteira entre diferentes subgrupos moleculares.

Esta técnica também se mostrou importante ao possibilitar um maior grau de liberdade, isto é, de refino na escolha do melhor agrupamento. Contudo, a diferença entre a densidade dos grupos observados, como mostrado pela análise de PCA, indica que uma técnica que considere esta diferença pode ser mais adequada ao problema, como por exemplo o método de agrupamento *fuzzy* hierárquico.

Considerando-se os resultados das métricas *fuzzy* observa-se que o comportamento do resultado da métrica de Xie-Beni (XB) obteve um comportamento muito similar ao do resultado da curva de p-valor da sobrevida global,

obtendo melhor desempenho para um número de grupos variando entre 4 e 6, e um pior desempenho para 7 grupos. Por isso, há de se observar a oportunidade de se utilizar a função objetivo do modelo *fuzzy* como sendo a métrica XB ao invés da utilizada neste trabalho, a função PC. Esta alteração pode ajudar o modelo a buscar parâmetros que resultarão em agrupamentos com grupos com curvas de sobrevida mais distintas entre si.

A adição de ômicas também mostrou ser importante na obtenção de resultados que possibilitem maior compreensão do problema, o que mostra que a metodologia utilizada colabora para que o modelo desenvolvido reflita a enorme complexidade do problema apresentado.

E, por fim, a aplicação desta metodologia a outros tipos de cânceres, pode ampliar o entendimento da relação genótipo-fenótipo, possibilitando a compreensão inclusive de mecanismo comuns aos diferentes tipos de cânceres.

Bibliografia

[1] Yifeng Li, Fang-Xiang Wu e Alioune Ngom. "A Review on Machine Learning Principles for Multi-View Biological Data Integration". Em: *Briefings in Bioinformatics* 19 (out. de 2016), accpted. DOI: 10.1093/bib/bbw113.

[2] Eugene Lin e Hsien-Yuan Lane. "Machine learning and systems genomics approaches for multi-omics data". Em: *Biomarker Research* 5 (dez. de 2017). DOI: 10.1186/s40364-017-0082-y.

[3] Daniel Koboldt et al. "Comprehensive molecular portraits of human breast tumours". Em: *Nature* 490 (out. de 2012), pp. 61–70.

[4] Hojoon Lee et al. "The Cancer Genome Atlas Clinical Explorer: A web and mobile interface for identifying clinical-genomic driver associations". Em: *Genome Medicine* 7 (out. de 2015). DOI: 10.1186/s13073-015-0226-3.

[5] Alberts B et al. *Biologia Molecular da célula*. 2002.

[6] Daniele Ramazzotti et al. "Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival". Em: (fev. de 2018). DOI: 10.1038/s41467-018-06921-8.

[7] Hsien-Yuan Lane, Guochuan Tsai e Eugene Lin. "Assessing Gene-Gene Interactions in Pharmacogenomics". Em: *Molecular diagnosis & therapy* 16 (fev. de 2012), pp. 15–27. DOI: 10.2165/11597270-000000000-00000.

[8] Eugene Lin. "Novel Drug Therapies and Diagnostics for Personalized Medicine and Nanomedicine in Genome Science, Nanoscience, and Molecular Engineering". Em: *Pharmaceutical Regulatory Affairs: Open Access* 01 (jan. de 2012). DOI: 10.4172/2167-7689.1000e116.

[9] Eugene Lin e Shih-Jen Tsai. "Novel diagnostics R&D for public health and personalized medicine in Taiwan: current state, challenges and opportunities". Em: *Curr Pharmacogenomics Person Med* (2012).

[10] Gun Jung, Kwang-Pyo Kim e Kwoneel Kim. "How to interpret and integrate multi-omics data at systems level". Em: *Animal Cells and Systems* 24 (jan. de 2020), pp. 1–7. DOI: 10.1080/19768354.2020.1721321.

[11] Ali Torkamani, Nathan Wineinger e Eric Topol. "The personal and clinical utility of polygenic risk scores". Em: *Nature Reviews Genetics* 19 (mai. de 2018), p. 1. DOI: 10.1038/s41576-018-0018-x.

[12] Yehudit Hasin-Brumshtein, Marcus Seldin e Aldons Lusis. "Multi-omics Approaches to Disease". Em: *Genome Biology* 18 (dez. de 2017). DOI: 10.1186/s13059-017-1215-1.

[13] Indhupriya Subramanian et al. "Multi-omics Data Integration, Interpretation, and Its Application". Em: *Bioinformatics and Biology Insights* 14 (jan. de 2020), p. 117793221989905. DOI: 10.1177/1177932219899051.

[14] Jared Churko et al. "Overview of High Throughput Sequencing Technologies to Elucidate Molecular Pathways in Cardiovascular Diseases". Em: *Circulation research* 112 (jun. de 2013), pp. 1613–23. DOI: 10.1161/CIRCRESAHA.113.300939.

[15] Hane Lee et al. "Improving the efficiency of genomic loci capture using oligonucleotide arrays for high throughput resequencing". Em: *BMC genomics* 10 (dez. de 2009), p. 646. DOI: 10.1186/1471-2164-10-646.

[16] C. Buerkle e Zachariah Gompert. "Population genomics based on low coverage sequencing: How low should we go?" Em: *Molecular ecology* 22 (nov. de 2012). DOI: 10.1111/mec.12105.

[17] Daniel Koboldt et al. "Challenges of sequencing human genomes". Em: *Briefings in bioinformatics* 11 (set. de 2010), pp. 484–98. DOI: 10.1093/bib/bbq016.

[18] Andreas Gnirke et al. "Solution Hybrid Selection with Ultra-long Oligonucleotides for Massively Parallel Targeted Sequencing". Em: *Nature biotechnology* 27 (fev. de 2009), pp. 182–9. DOI: 10.1038/nbt.1523.

[19] Christoph Bock. "Analysing and interpreting DNA methylated data". Em: *Nature reviews. Genetics* 13 (set. de 2012), pp. 705–19. DOI: 10.1038/nrg3273.

[20] Nizar Touleimat e Jörg Tost. "Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation". Em: *Epigenomics* 4 (jun. de 2012), pp. 325–41. DOI: 10.2217/epi.12.21.

[21] Paweł Łabaj et al. “Characterization and Improvement of RNA-Seq Precision in Quantitative Transcript Expression Profiling”. Em: *Bioinformatics (Oxford, England)* 27 (jul. de 2011), pp. i383–91. DOI: 10.1093/bioinformatics/btr247.

[22] Manuel Garber et al. “Computational methods for transcriptome annotation and quantification using RNA-seq”. Em: *Nature methods* 8 (jun. de 2011), pp. 469–77. DOI: 10.1038/nmeth.1613.

[23] Kenneth Whitaker Witwer e Marc K Halushka. “Toward the promise of microRNAs – Enhancing reproducibility and rigor in microRNA research”. English (US). Em: *RNA Biology* 13.11 (nov. de 2016), pp. 1103–1116. ISSN: 1547-6286. DOI: 10.1080/15476286.2016.1236172.

[24] Michael Love, Wolfgang Huber e Simon Anders. “Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2”. Em: *Genome biology* 15 (dez. de 2014), p. 550. DOI: 10.1186/PREACCEPT-8897612761307401.

[25] Seungyeoun Lee e Heeju Lim. “Review of statistical methods for survival analysis using genomic data”. Em: *Genomics & Informatics* 17 (dez. de 2019), e41. DOI: 10.5808/GI.2019.17.4.e41.

[26] Trevor Hastie et al. “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”. Em: *Math. Intell.* 27 (nov. de 2004), pp. 83–85. DOI: 10.1007/BF02985802.

[27] Jundong Li et al. “Feature Selection: A Data Perspective”. Em: *ACM Computing Surveys* 50 (jan. de 2016). DOI: 10.1145/3136625.

[28] S. S. Gandhi e S. S. Prabhune. “Overview of feature subset selection algorithm for high dimensional data”. Em: *2017 International Conference on Inventive Systems and Control (ICISC)*. 2017, pp. 1–6.

[29] Asir Antony Danasingh, Suganya Balamurugan e JEBAMALAR LEAVLINE EPIPHANY. “Literature Review on Feature Selection Methods for High-Dimensional Data”. Em: *International Journal of Computer Applications* 136 (fev. de 2016). DOI: 10.5120/ijca2016908317.

[30] Zheng Zhao e Huan Liu. “Spectral feature selection for supervised and unsupervised learning”. Em: vol. 227. Jan. de 2007, pp. 1151–1157. DOI: 10.1145/1273496.1273641.

[31] Jianyu Miao e Lingfeng Niu. "A Survey on Feature Selection". Em: *Procedia Computer Science* 91 (2016). Promoting Business Analytics and Quantitative Management of Technology: 4th International Conference on Information Technology and Quantitative Management (ITQM 2016), pp. 919–926. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2016.07.111. URL: http://www.sciencedirect.com/science/article/pii/S1877050916313047.

[32] Marko Robnik-Sikonja e Igor Kononenko. "An adaptation of Relief for attribute estimation in regression". Em: *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning* (fev. de 2000).

[33] Saúl Solorio-Fernández, J. Carrasco-Ochoa e José Francisco Martínez-Trinidad. "A review of unsupervised feature selection methods". Em: *Artificial Intelligence Review* (jan. de 2019). DOI: 10.1007/s10462-019-09682-y.

[34] Deng Cai, Chiyuan Zhang e Xiaofei He. "Unsupervised feature selection for Multi-Cluster data". Em: jul. de 2010, pp. 333–342. DOI: 10.1145/1835804.1835848.

[35] Abdelkarim Ben Ayed, Mohamed Ben Halima e Adel Alimi. "Survey on clustering methods : Towards fuzzy clustering for big data". Em: (ago. de 2015).

[36] Igor Škrjanc et al. "Evolving Fuzzy and Neuro-Fuzzy Approaches in Clustering, Regression, Identification, and Classification: A Survey". Em: *Information Sciences* 490 (mar. de 2019). DOI: 10.1016/j.ins.2019.03.060.

[37] Guillermo de Anda-Jáuregui e Enrique Hernández-Lemus. "Computational Oncology in the Multi-Omics Era: State of the Art". Em: *Frontiers in Oncology* 10 (abr. de 2020). DOI: 10.3389/fonc.2020.00423.

[38] Syed Abdul Shabbir, Usman Iqbal e Yu-Chuan Li. "Predictive Analytics through Machine Learning in the clinical settings". Em: *Computer Methods and Programs in Biomedicine* 144 (jun. de 2017), A1–A2. DOI: 10.1016/S0169-2607(17)30552-7.

[39] Burcu Darst, Kristen Malecki e Corinne Engelman. "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data". Em: *BMC Genetics* 19 (set. de 2018). DOI: 10.1186/s12863-018-0633-8.

[40] Animesh Acharjee et al. “Integration of multi-omics data for prediction of phenotypic traits using random forest”. Em: (jan. de 2016).

[41] Amir Ahmad e Shehroz Khan. “A survey of state-of-the-art mixed data clustering algorithms”. Em: (mar. de 2019). DOI: 10.13140/RG.2.2.17863.55209.

[42] L.A. Zadeh. “Fuzzy set theory”. Em: *Information Science* 2 (1965), pp. 338–352.

[43] Enrique H. Ruspini. “Numerical methods for fuzzy clustering”. Em: *Information Sciences* 2.3 (1970), pp. 319–350. ISSN: 0020-0255. DOI: https://doi.org/10.1016/S0020-0255(70)80056-1. URL: http://www.sciencedirect.com/science/article/pii/S0020025570800561.

[44] J. C. Dunn. “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters”. Em: *Journal of Cybernetics* 3.3 (1973), pp. 32–57. DOI: 10.1080/01969727308546046. eprint: https://doi.org/10.1080/01969727308546046. URL: https://doi.org/10.1080/01969727308546046.

[45] James Bezdek. *Pattern Recognition With Fuzzy Objective Function Algorithms*. Jan. de 1981. DOI: 10.1007/978-1-4757-0450-1.

[46] Karim Kalti e Mohamed Mahjoub. “Image Segmentation by Gaussian Mixture Models and Modified FCM Algorithm”. Em: *International Arab Journal of Information Technology* 11 (jan. de 2014), pp. 11–18.

[47] Sanghamitra Bandyopadhyay. “Multiobjective Simulated Annealing for Fuzzy Clustering With Stability and Validity”. Em: *IEEE Transactions on Systems, Man, and Cybernetics, Part C* 41 (set. de 2011), pp. 682–691. DOI: 10.1109/TSMCC.2010.2088390.

[48] Sailik Sengupta et al. “An improved fuzzy clustering method using modified Fukuyama-Sugeno cluster validity index”. Em: *2011 International Conference on Recent Trends in Information Systems, ReTIS 2011 - Proceedings* (dez. de 2011). DOI: 10.1109/ReTIS.2011.6146880.

[49] Xuanli Xie e Gerardo Beni. “A Validity Measure for Fuzzy Clustering”. Em: *IEEE Trans. Pattern Anal. Mach. Intell.* 13 (ago. de 1991), pp. 841–847. DOI: 10.1109/34.85677.

[50] Marylyn Ritchie et al. “Methods of integrating data to uncover genotype-phenotype interactions”. Em: *Nature reviews. Genetics* 16 (fev. de 2015), pp. 85–97. DOI: 10.1038/nrg3868.

[51] Marie Tayrac et al. “Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach”. Em: *BMC genomics* 10 (fev. de 2009), p. 32. DOI: 10.1186/1471-2164-10-32.

[52] Chen Meng et al. “A multivariate approach to the integration of multi-omics datasets”. Em: *BMC bioinformatics* 15 (mai. de 2014), p. 162. DOI: 10.1186/1471-2105-15-162.

[53] Aedín Culhane, Guy Perrière e Desmond Higgins. “Culhane AC, Perrière G, Higgins DG.. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. BMC Bioinformatics 4: 59”. Em: *BMC bioinformatics* 4 (dez. de 2003), p. 59. DOI: 10.1186/1471-2105-4-59.

[54] Meng Chen et al. “Dimension reduction techniques for the integrative analysis of multi-omics data”. Em: *Briefings in Bioinformatics* 17 (mar. de 2016), bbv108. DOI: 10.1093/bib/bbv108.

[55] Bo Wang et al. “Similarity network fusion for aggregating data types on a genomic scale”. Em: *Nature methods* 11 (jan. de 2014). DOI: 10.1038/nmeth.2810.

[56] Justin Guinney et al. “The consensus molecular subtypes of colorectal cancer”. Em: *Nature medicine* 21 (out. de 2015). DOI: 10.1038/nm.3967.

[57] Murilo Geraldo, Helder Nakaya e Edna Kimura. “Down-regulation of 14q32-encoded miRNAs and tumor suppressor role for miR-654-3p in papillary thyroid cancer”. Em: *Oncotarget* 8 (dez. de 2016). DOI: 10.18632/oncotarget.14162.

[58] Xinguo Lu et al. “The Integrative Method Based on the Module-Network for Identifying Driver Genes in Cancer Subtypes”. Em: *Molecules* 23 (jan. de 2018), p. 183. DOI: 10.3390/molecules23020183.

[59] Li Zhang et al. “Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma”. Em: *Frontiers in Genetics* 9 (out. de 2018), p. 477. DOI: 10.3389/fgene.2018.00477.

[60] Gang Liu, Chuanpeng Dong e Lei Liu. “Integrated Multiple “-omics” Data Reveal Subtypes of Hepatocellular Carcinoma”. Em: *PLOS ONE* 11 (nov. de 2016), e0165457. DOI: 10.1371/journal.pone.0165457.

[61] Hyeongmin Kim e Yong-Min Kim. "Pan-cancer analysis of somatic mutations and transcriptomes reveals common functional gene clusters shared by multiple cancer types". Em: *Scientific Reports* 8 (dez. de 2018). DOI: 10.1038/s41598-018-24379-y.

[62] Boyu Lyu e Anamul Haque. "Deep Learning Based Tumor Type Classification Using Gene Expression Data". Em: ago. de 2018, pp. 89–96. DOI: 10.1145/3233547.3233588.

[63] Marieke Kuijjer et al. "Cancer subtype identification using somatic mutation data". Em: *British Journal of Cancer* 118 (mai. de 2018). DOI: 10.1038/s41416-018-0109-7.

[64] Yang Liu et al. "Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas". Em: *Cancer Cell* 33 (abr. de 2018). DOI: 10.1016/j.ccell.2018.03.010.

[65] *Incidência de CCR no brasil e no mundo*. https://gco.iarc.fr/today/home. Accessed: 2020-08-30.

[66] J. Ferlay et al. "Cancer incidence and mortality worldwide: IARC CancerBase No". Em: *International Agency for Research on Cancer.* 2 (jan. de 2013).

[67] Constance Johnson et al. "Meta-analyses of Colorectal Cancer Risk Factors". Em: *Cancer causes & control : CCC* 24 (abr. de 2013). DOI: 10.1007/s10552-013-0201-5.

[68] Claudia Allemani et al. "Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries". Em: *The Lancet* 391 (mar. de 2018).

[69] Rebecca Siegel et al. "Colorectal cancer statistics, 2017". Em: *CA: A Cancer Journal for Clinicians* 67 (mar. de 2017). DOI: 10.3322/caac.21395.

[70] Danil Kolikov. *Methods for generic data*. 2018. URL: https://github.com/danilkolikov/fsfc.

[71] *Repositório skrebate no pip*. https://pypi.org/project/skrebate/. Accessed: 2020-08-30.

[72] Ryan J. Urbanowicz et al. *Benchmarking Relief-Based Feature Selection Methods*. arXiv e-print. https://arxiv.org/abs/1711.08477. 2017.

[73] JL Bos et al. “Prevalence of ras gene mutations in human colorectal cancers”. Em: *Nature* 327.6120 (1987), pp. 293–297. ISSN: 0028-0836. DOI: 10.1038/327293a0. URL: https://doi.org/10.1038/327293a0.

[74] Päivi Peltomäki et al. “Evidence Supporting Exclusion of the DCC Gene and a Portion of Chromosome 18q as the Locus for Susceptibility to Hereditary Nonpolyposis Colorectal Carcinoma in Five Kindreds”. Em: *Cancer research* 51 (set. de 1991), pp. 4135–40.

[75] Patricia Muller e Karen Vousden. “P53 mutations in cancer”. Em: *Nature cell biology* 15 (jan. de 2013), pp. 2–8. DOI: 10.1038/ncb2641.

[76] Hanan Lamlum et al. “APC mutations are sufficient for the growth of early colorectal adenomas”. Em: *Proceedings of the National Academy of Sciences* 97 (mar. de 2000). DOI: 10.1073/pnas.040564697.

[77] Eric R. Fearon e Bert Vogelstein. “A genetic model for colorectal tumorigenesis”. English (US). Em: *Cell* 61.5 (jun. de 1990), pp. 759–767. ISSN: 0092-8674. DOI: 10.1016/0092-8674(90)90186-I.