CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

CLASSIFICAÇÃO DE PACIENTES DE ACORDO COM A SENSIBILIDADE QUIMIOTERÁPICA NEOADJUVANTE POR MEIO DA ANÁLISE DE BIMODALIDADE

GUSTAVO HENRIQUE PASSINI SANTOS

Orientador: Prof. Dr. Thiago de Souza Rodrigues
Centro Federal de Educação Tecnológica de Minas Gerais

BELO HORIZONTE
MAIO DE 2019

GUSTAVO HENRIQUE PASSINI SANTOS

CLASSIFICAÇÃO DE PACIENTES DE ACORDO COM A SENSIBILIDADE QUIMIOTERÁPICA NEOADJUVANTE POR MEIO DA ANÁLISE DE BIMODALIDADE

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional

Linha de pesquisa: Métodos Matemáticos Aplicados

Orientador: Prof. Dr. Thiago de Souza Rodrigues
Centro Federal de Educação Tecnológica de Minas Gerais

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL
BELO HORIZONTE
MAIO DE 2019

Santos, Gustavo Henrique Passini
S237c Classificação de pacientes de acordo com a sensibilidade quimioterápica neoadjuvante por meio da análise de bimodalidade / Gustavo Henrique Passini Santos. – 2019.
42 f.

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.
Orientador: Thiago de Souza Rodrigues.
Dissertação (mestrado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Mamas – Câncer – Modelagem matemática – Teses. 2. Terapia neoadjuvante – Teses. 3. Expressão gênica – Teses. 4. Bimodalidade – Teses. I. Rodrigues, Thiago de Souza. II. Centro Federal de Educação Tecnológica de Minas Gerais. III. Título.

CDD 511.8

Elaboração da ficha catalográfica pela Biblioteca-Campus II / CEFET-MG

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

"CLASSIFICAÇÃO DE PACIENTES DE ACORDO COM A SENSIBILIDADE QUIMIOTERÁPICA NEOADJUVANTE POR MEIO DA ANÁLISE DE BIMODALIDADE"

Dissertação de Mestrado apresentada por **Gustavo Henrique Passini Santos**, em 3 de junho de 2019, ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, e aprovada pela banca examinadora constituída pelos professores:

Prof. Dr. Thiago de Souza Rodrigues
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Anton Semenchenko
Aplasys

Prof. Dr. Alisson Marques da Silva
Centro Federal de Educação Tecnológica de Minas Gerais

Visto e permitida à impressão,

Prof. Dr. Thiago de Souza Rodrigues
Coordenador do Programa de Pós-Graduação em
Modelagem Matemática e Computacional

Dedico essa dissertação em primeiro lugar, a Deus, pela força e coragem durante toda esta longa caminhada. Também dedico a minha esposa, meus pais, minha irmã, e a toda minha família que, com muito carinho e apoio, não mediram esforços para que eu chegasse até esta etapa da minha vida. Dedico também ao meu orientador, que sempre esteve ao meu lado indicando o caminho certo a seguir, além dos vários professores do CEFET-MG, que com alta responsabilidade me passaram o conhecimento necessário para finalizar essa jornada. Extendo a dedicatória a MD2 Consultoria LTDA, especialmente a Danilo Soares, Julio Costa e Max Rabello, que atuaram como facilitadores nessa etapa tão árdua de consolidação acadêmica. E por último, mas não menos importante, dedico ao professor Doutor Anton Semenchenko pelo fato de acreditar que sem ele nada disso seria possível.

Agradecimentos

Agradeço a Deus, pela oportunidade de viver e finalizar essa caminhada. Também agradeço a minha esposa, meus pais, minha irmã, e a toda minha família, pelo apoio e companheirismo. Dedico também ao meu orientador, pelo conhecimento passado e conselhos dados. Agradeço também o professor Doutor Anton Semenchenko por toda oportunidade de crescimento acadêmico e formação como aluno.

“O sucesso é a soma de pequenos esforços repetidos dia após dia.” (Robert Collier)

Resumo

Este projeto aborda o desenvolvimento de uma nova metodologia visando à caracterização de pacientes com câncer de mama em PCR (*pathologic complete response*) e NoPCR para tratamento neoadjuvante. Atualmente a comunidade médica enfrenta uma grande dificuldade em encaminhar pacientes com câncer de mama para o tratamento neoadjuvante, quimioterapia realizada com o intuito de diminuir as dimensões do tumor ou curá-lo, pois não existem ferramentas que garantem um nível satisfatório de confiabilidade. Neste contexto, pesquisadores esforçam-se para encontrar métodos que viabilizem uma caracterização mais assertiva dos pacientes, para que dessa forma se evite o desgaste físico e psicológico de uma quimioterapia sem uma redução satisfatória ou cura. Uma tecnologia utilizada para análise genética de pacientes é o *microarray*, técnica que se baseia no emparelhamento de sondas marcadas com quimiluminescência, radioatividade, ou por um anticorpo, para calcular a expressão genética em uma sonda, 22283 para o contexto desta dissertação, que por sua vez representa um gene, e apresentá-lo de forma numérica. Uma vertente nessa linha de pesquisa é a utilização do cálculo da bimodalidade genética para verificar a influência de determinadas sondas na caracterização de pacientes como PCR e NoPCR. A metodologia proposta visa selecionar sondas significativas por meio de ordenação e comparação entre as 3000 primeiras sondas com maior e menor valor de bimodalidade em pacientes das duas classes, PCR e NoPCR, em uma base de dados previamente utilizada em literatura. Após selecionar as sondas, pacientes oriundos de uma base diferente da utilizada para seleção serão classificados por intermédio de três classificadores utilizados previamente em literatura: *Diagonal Linear Discriminant Analisys*, *Particle Swarm Optimization* e *Extreme Learning Machine*, atestando dessa forma a qualidade da metodologia por meio dos resultados utilizando a curva ROC, sensibilidade e especificidade. Os resultados indicam que as sondas selecionadas detêm funções genéticas relacionadas à proliferação do câncer, portanto são significativas. Obteve-se desempenho estatístico satisfatório para as classificações realizadas com *Extreme Learning Machine*, já para os outros algoritmos os resultados não apresentaram significância em comparação a outros trabalhos já realizados e referenciados.

Palavras-chave:Câncer de mama. Tratamento neoadjuvante. Expressão genética. Bimodalidade.

Abstract

This project addresses the development of a new methodology aimed at the characterization of patients with breast cancer in PCR (pathologic complete response) and NoPCR for neoadjuvant treatment. Currently the medical community faces a great difficulty to refer patients with breast cancer for neoadjuvant treatment, chemotherapy performed with the intention of reducing the size of the tumor or cure it, because there are no tools that guarantee a satisfactory level of reliability. In this context, researchers strive to find methods that enable a more assertive characterization of patients, so as to avoid the physical and psychological exhaustion of a chemotherapy without obtaining a satisfactory decrease or cure. One technology used for genetic analysis of patients is the microarray, a technique that relies on the pairing of chemiluminescently labeled probes, radioactivity, or an antibody, to calculate the gene expression in a probe, 22283 for the context of this dissertation, which by its instead represents a gene, and present it numerically. One strand in this line of research is the use of the genetic bimodality calculation to verify the influence of certain probes on the characterization of patients as PCR and NoPCR. The proposed methodology aims to select significant probes by ordering and comparing the 3000 first probes with higher and lower bimodality values in patients of both classes, PCR and NoPCR, in a database previously used in the literature. After selecting the probes, patients from a different base than the one used for selection will be classified by means of three classifiers previously used in literature: DLDA, Particle Swarm Optimization and Extreme Learning Machine, thus attesting the quality of the methodology using the results using the ROC curve, sensitivity and specificity. The results indicate that the selected probes have genetic functions related to cancer proliferation, so they are significant. We obtained satisfactory statistical performance for the classifications carried out with Extreme Learning Machine, and for the other algorithms the results were not significant in comparison to other works already done and referenced.

Keywords: Breast cancer. Neoadjuvant treatment. Genetic Expression. Bimodality.

Lista de Figuras

Sumário

1 Introdução

Neste capítulo os conceitos que embasam essa dissertação serão caracterizados. Na primeira seção será abordado o câncer de mama, a doença da qual sofrem os pacientes analisados nesse estudo, além de se definir o método de tratamento do qual o problema é originado.

A segunda seção discorre sobre os conceitos genéticos envolvidos no estudo realizado e também na representação da análise feita para prever os grupos de resposta patológica. A terceira seção explica o funcionamento da metodologia de expressão genética necessária para que as análises matemáticas e estatísticas sejam realizadas.

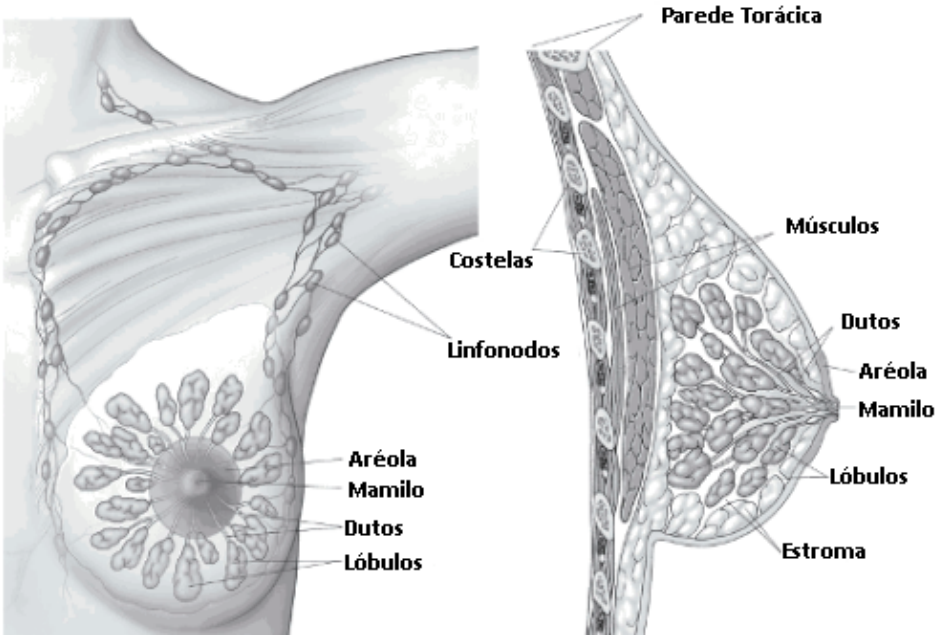
A quarta seção apresenta e define a utilização da bimodalidade no estudo das expressões genéticas, que norteia a metodologia de seleção de genes apresentada nesta dissertação, referenciando os principais autores que caracterizam essa metodologia. A quinta seção caracteriza o problema, apresentando as dificuldades inerentes ao estudo, e também os objetivos gerais e específicos desse trabalho.

1.1 Câncer de Mama

O câncer de mama é uma neoplasia maligna que inicia nos tecidos da mama. Existem dois tipos principais de câncer de mama: Carcinoma Ductal, que começa nos tubos que transportam o leite da mama até o mamilo e representam a grande maioria das ocorrências desse tumor; Carcinoma Lobular, que inicia nos lóbulos, partes da mama que produzem leite. É possível verificar os ductos, lóbulos e toda estrutura da mama por meio da Figura 1. Ainda existem casos raros onde o câncer pode começar em outras áreas da mama (MEDLINE PLUS, 2018).

O crescimento descontrolado das células do seio é o gatilho para a existência do câncer de mama. Tal comportamento observado nessas células geralmente forma um tumor, visível por intermédio de um exame de radiografia, ou por meio de um exame de toque, onde é possível se identificar um nódulo. O tumor é definido como maligno, ou câncer, se as células puderem se transformar em tecidos adjacentes ou se espalharem para áreas distantes do corpo. É possível que o câncer de mama ocorra em pessoas do sexo masculino, porém sua incidência quase que inteira se dá em mulheres (AMERICAN CANCER SOCIETY, 2017).

Figura 1 – Anatomia da Mama (Figura Adaptada)



Fonte: https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html

A partir do momento em que o tumor foi identificado inicia-se uma análise acerca de suas características. O estadiamento é uma técnica que se encaixa nesse contexto, contando com larga utilização na sociedade médica. O estágio do câncer depende do tamanho e da localização de um tumor, se ele se espalhou e até onde se espalhou. A análise do estágio em que o tumor se encontra auxilia na decisão do melhor tratamento, tipo de acompanhamento necessário e previsão da chance de recuperação do paciente (prognóstico). São definidos dois métodos de estadiamento: o estadiamento clínico, que é baseado em testes realizados antes da cirurgia; o estadiamento patológico, que utiliza os resultados dos testes laboratoriais realizados no tecido mamário e nos gânglios linfáticos removidos durante a cirurgia, ajudando a determinar o tratamento adicional e a prever o que esperar após o término do tratamento (MEDLINE PLUS, 2017).

O principal tratamento para o câncer de mama é a quimioterapia. Tal método baseia-se na administração de medicamentos que matam o câncer, seja por via intravenosa (injetados em sua veia) ou pela boca. As drogas são transportadas pela corrente sanguínea alcançando as células cancerígenas na maior parte do corpo. Ocasionalmente, a quimioterapia pode ser administrada diretamente no líquido espinhal que envolve o cérebro e a medula espinhal. Não são todas as mulheres com câncer de mama que precisarão de quimioterapia, porém existem variadas situações em que a quimioterapia pode ser recomendada (AMERICAN CANCER SOCIETY, 2017):

• Após a cirurgia (quimioterapia adjuvante): utilizada na tentativa de matar qualquer célula cancerígena deixada para trás ou que tenha se espalhado, não podendo ser vista

nem com auxílio de exames de imagem. Caso tais células cresçam, novos tumores serão formados em outras partes do corpo. Dessa forma se diminui o risco do câncer de mama voltar;

• Antes da cirurgia (quimioterapia neoadjuvante): utilizada na tentativa de diminuir o tumor, tornando possível sua remoção com uma cirurgia menos extensa. Por esse motivo é frequentemente usada para tratar cânceres que são grandes demais para serem removidos por cirurgia no momento do diagnóstico. Ao realizar a quimioterapia antes da remoção do tumor é possível que o médico verifique como o câncer se comporta;

• Para o câncer de mama avançado: pode ser utilizada como tratamento principal para mulheres cujo câncer se espalhou para fora da área das mamas e das axilas, seja quando diagnosticado ou após os tratamentos iniciais. A duração do tratamento depende de quão bem a quimioterapia está funcionando e quão bem é tolerada.

Sabe-se que a quimioterapia pode ser útil para muitos pacientes com câncer de mama. Mas prever quem irá se beneficiar e em qual escala esses benefícios ocorrerão ainda está sendo estudado. Às vezes há efeitos colaterais significativos (longo e curto prazo) da quimioterapia, portanto, avaliações que podem determinar com alto grau de assertividade quem realmente precisa de quimioterapia são de grande valia. Muitas pesquisas são realizadas para avaliar diferentes testes que podem demonstrar com mais precisão quais pacientes se beneficiariam da quimioterapia e quais pacientes poderiam evitá-la (AMERICAN CANCER SOCIETY, 2018).

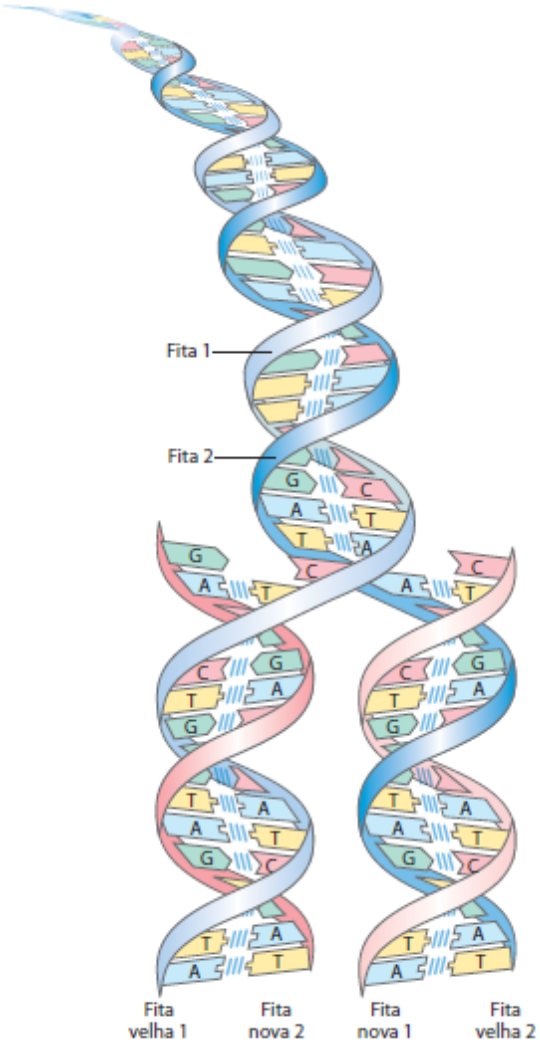
1.2 Conceitos Básicos de Biologia Molecular

O DNA é definido como uma cadeia de moléculas formando uma dupla hélice. Nele está contida toda a informação de que os processos celulares necessitam. O DNA atua como um modelo para que a criação do ácido ribonucleico (RNA) e as proteínas ocorram, sendo que a produção de proteínas é coordenada pelo RNA.

Quatro tipos de moléculas constituem a cadeia de DNA: guanina (G), adenina (A), timina (T), e citosina (C). Tais moléculas são definidas como bases. As bases organizam-se em pares, sendo que a citosina se conecta apenas com a guanina, e a timina se conecta apenas com a adenina. Para que uma hélice conecte-se a outra para formar uma dupla hélice é necessário que sejam complementares, observando as regras citadas anteriormente. Um exemplo é: a hélice CTAGTC se conectará somente com a hélice GATCAG, pois tal conexão corresponde à regra de pareamento, sendo que todas as bases das hélices são complementares entre si. É possível verificar a complementaridade entre fitas de DNA na Figura 2. As duas hélices irão se repelir caso ao menos uma das bases não for complementar na segunda hélice (NELSON; COX, 2014).

O processo de polimerização do DNA é definido por três etapas: iniciação, alongamento e término. O desenrolar e a separação das hélices do DNA ocorrem na etapa de iniciação. A helicase atua na divisão da hélice de DNA, produzindo duas fitas separadas que atuam como molde para formação das novas moléculas nas etapas posteriores. A principal enzima da segunda etapa é o DNA polimerase, sendo responsável pela ligação de um nucleotídeo por vez nas fitas preparadas pela ação realizada na etapa de iniciação. São formados pares de bases entre os nucleotídeos das fitas de DNA e nucleotídeos trifosfatos livres que se encontram ao redor dos DNAs no ambiente, gerando dessa forma novas hélices (CRAIG et al., 2014).

Figura 2 – Complementaridade entre Fitas do DNA

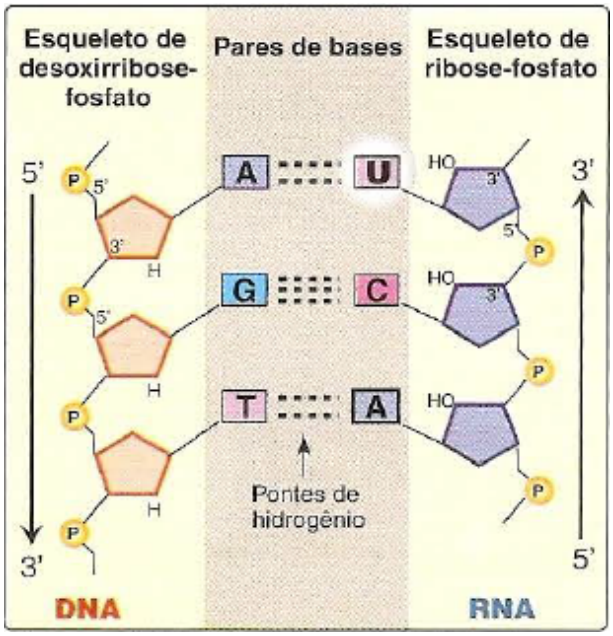


Fonte: Princípios de bioquímica de Lehninger. 6. ed.

O DNA e o RNA diferenciam-se basicamente pela existência da uracila (U) como item formador de sua hélice, tal base substitui a timina (T), como é observado na Figura 3. O RNA funciona como um mensageiro, transportando a informação da qual a produção de

proteína necessita por meio de uma cadeia simples de bases, sendo assim, o RNA atua como um gabarito para a produção de uma proteína. Participam do processo de síntese proteica três tipos principais de RNA: RNA ribossomal (RNAr), RNA transportador (RNAt) e RNA mensageiro (RNAm). O RNAr é um componente dos ribossomos. O RNAt atua como uma molécula de adaptação, que carrega um específico aminoácido para o sítio de síntese da proteína. O RNAm transporta a informação do DNA nuclear para o citosol, onde é utilizado como molde para a síntese de proteína. O processo de síntese do RNA é definido como transcrição e seus substratos são ribonucleosídeos trifosfatos (CHAMPE; HARVEY; FERRIER, 2006).

Figura 3 – Pares de bases complementares entre o DNA e o RNA



Fonte: Bioquímica Ilustrada. 3. ed.

Define-se gene como uma parte do DNA, contendo informações necessárias para a produção proteica. Para a criação de uma proteína primeiramente as duas cadeias simples da dupla hélice do DNA são separadas. Moléculas de RNA são atraídas para a cadeia e se conectam a ela originando um modelo similar a cadeia, porém com uracila no lugar da timina. Denomina-se tal processo como transcrição. Após concluir a duplicata, o RNA se desconecta da cadeia e se encaminha ao ribossomo para realizar a síntese proteica no processo denominado translação. Após isso, as cadeias do DNA se unem novamente e retornam para a forma original de dupla hélice (CHAMPE; HARVEY; FERRIER, 2006).

1.3 Microarray

Dois importantes métodos relacionados à manipulação de DNA são:

• Polymerase Chain Reaction: definido como uma técnica de multiplicação da fita de DNA, sendo utilizada para produzir um número desejado cópias de um determinado fragmento de DNA. Tal técnica utiliza tubos de ensaio contendo o DNA e outros componentes, como a enzima DNA polimerase (enzima que faz a replicação do DNA) e os *primers* (DNAs iniciadores). Para execução do processo utiliza-se uma máquina de *Polymerase Chain Reaction*, que controla a temperatura no tubo de ensaio, garantindo que os passos da replicação do DNA ocorram;

• Hibridização: utiliza sondas para permitir a caracterização e identificação do DNA, baseando-se em suas propriedades, por meio do emparelhamento de bases por pontes de hidrogênio, e a desnaturação e renaturação das cadeias complementares. Tal técnica é utilizada na análise de genes expressos assim como genes não expressos. Para visualizar os genes de interesse emprega-se o emparelhamento de sondas que se encontram marcadas com quimiluminescência, radioatividade, ou por um anticorpo.

A partir da união das duas técnicas definidas anteriormente surge o *microarray* de DNA, que permite a triagem rápida e simultânea de grande quantidade de genes. Segmentos do DNA de genes que compõem alguma sequência conhecida são amplificados por meio de *Polymerase Chain Reaction*. A partir daí, dispositivos robóticos depositam precisas quantidades de nanolitros de soluções de DNA em uma superfície sólida medindo poucos centímetros quadrados, em um arranjo predeterminado, tendo cada um dos pontos contendo sequências derivadas de um determinado gene. Uma estratégia cada vez mais utilizada na síntese do DNA diretamente sobre uma superfície sólida é a fotolitografia. O arranjo resultante é na maioria das vezes definido como chip. O microarranjo é normalmente sondado com mRNA ou cDNA a partir de um determinada cultura de células ou tipo celular para identificar os genes expressos nessas células (NELSON; COX, 2014).

Foi observada uma rápida evolução na tecnologia de *microarray*. Juntamente ao avanço da tecnologia, verifica-se o aumento da capacidade de armazenamento e análise de dados e a redução dos custos para produção de *arrays*, dessa forma tornam-se viáveis aplicações práticas como a análise dos genes que estão correlacionados com doenças como o câncer (STEKEL, 2003). Em 1995 iniciou-se uma inovação onde foi apresentada uma técnica baseando-se em DNA complementar (cDNA) na medição de expressão genética. Foram inseridos, com a utilização de um robô, cDNA de variados genes do organismo Arabidopsis thaliana em uma lâmina de microscópio. Após isso, para cada gene a ser analisado foram preparadas sondas fluorescentes de mRNA (SCHENA et al., 1995).

Foi previsto em Schena et al. (1995) que os *arrays* comportariam facilmente mais de 20000 genes, algo possível com os *microarrays* affymetrix. Tal tecnologia tem a capacidade de analisar variados genes em paralelo, possibilitando um entendimento maior de mecanismos que são responsáveis por doenças ou por características (LOCKHART, 1996). A tecnologia detém larga utilização em estudos acerca de vários tipos de câncer, incluindo o

câncer de mama.

Como é possível verificar pela Figura 4, a utilização dos *microarrays* se dá por meio dos seguintes passos (STEKEL, 2003):

• Inicia-se com a escolha de quais tecidos sofrerão comparações. A tecnologia de *microarray* consiste na comparação da expressão genética de dois tecidos, sendo um o tecido doente, a ser analisado, e um tecido normal;

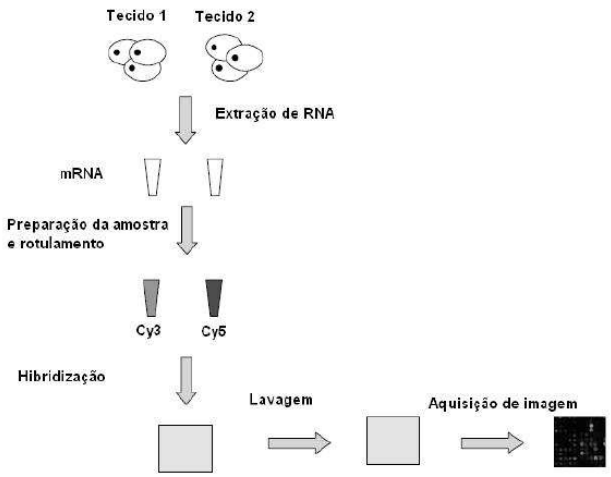
• O segundo passo consiste em extrair o RNA dos dois tecidos das amostras, aplicando após isso um marcador fluorescente. O marcador Cy5 fornece a cor vermelha à amostra e o marcador Cy3 fornece a cor verde;

• A partir daí as amostras sofrem hibridização ao serem inseridas no *microarray*. A hibridização consiste em combinar o RNA com a hélice do DNA presente no *array*. Foi definido na Seção 1.2 que o RNA só se combinará com o DNA de bases complementares. Ao ocorrer à combinação, uma dupla hélice é formada e o marcador fluorescente continua conectado na nova hélice;

• Com o intuito de retirar as amostras que não se combinaram, não sofrendo hibridização, com o DNA presente no *array*, é realizada a lavagem da lâmina.

• O *array* passa então por um *scanner*, que emitirá feixe de laser verde e um feixe de laser vermelho, com intuito de ativar a fluorescência dos marcadores. O *scanner* realizará a aquisição da imagem tratada posteriormente, por intermédio de técnicas de análise de imagem, com a normalização ao final do processo resultando na criação de um vetor numérico que representará a expressão dos genes.

Figura 4 – Funcionamento Microarray



Fonte: Microarray Bioinformatics

Similar a todos os tipos de *microarray*, o Affymetrix baseia-se na hibridização para

calcular a expressão genética. Tal técnica é capaz de medir a expressão de todos os genes humanos conhecidos, e também de outros organismos que se deseje. O primeiro passo é inserir um trecho de DNA, definida como probe ou sonda, na superfície do slide de vidro. Tal trecho tem 25 bases de comprimento, que representam uma pequena parte de um gene maior. A sonda de sequência de 25 bases é comparada com o restante do genoma humano, garantindo que este trecho genético não se repita em nenhum outro gene. No momento em que uma molécula de RNA associa-se a sonda, é possível validar que o gene foi expresso, pois tal sequência ocorre apenas para o gene especificado. Desta forma, tal trecho de DNA expressa o gene completo. Utilizam-se 11 seções diferentes do mesmo gene para sua identificação, aumentando a exatidão dos resultados gerados pelos *microarrays*, implicando que serão 11 sondas para cada gene medido. De modo diferente de outros tipos de *microarray*, o Affymetrix é monocolor, não necessitando de um tecido de referência, pois as sondas contêm os genes do genoma humano. Portanto o nível de expressão de cada gene do tecido estudado é o resultado da análise, e não a relação entre os níveis de expressão entre o tecido de referência e o tecido alvo (HORTA, 2008).

1.4 Bimodalidade

A identificação de genes com padrões de expressão bimodal a partir de dados de perfil de expressão em larga escala é uma importante vertente de pesquisa na busca pela caracterização de pacientes PCR e NoPCR. Padrões de expressão bimodal podem resultar naturalmente da expressão diferencial, com os dois modos centrados na expressão média de um gene em dois subgrupos distintos de amostras. No contexto do câncer, os padrões de expressão bimodal podem resultar de lesões genômicas que ocorrem em alguns pacientes, mas não em outros (WANG et al., 2009).

A expressão genética no nível do RNA mensageiro frequentemente opera em dois modos: linha de base, e subexpressão ou superexpressão. Embora os mecanismos subjacentes que conduzem estes dois modos de expressão (expressão bimodal) possam se diferenciar, na alteração do número de cópias de DNA, na regulação de microRNA, na metilação do DNA, na atividade de fatores de transcrição e assim por diante, o impacto da expressão bimodal é crítico para o funcionamento celular. Por exemplo, a expressão específica de tecido para certos genes são necessários para a diferenciação celular; a manutenção e regulação da expressão bimodal são essenciais para a sinalização celular; a mudança do modo de expressão, especialmente para oncogenes e supressores de tumor, podem levar a proliferação celular descontrolada e câncer maligno (TONG et al., 2013).

A definição de uma distribuição bimodal pode ser vaga: o termo refere-se tipicamente a uma mistura de duas populações com meios distintos. Na estimativa de densidade, uma distribuição bimodal pode ser reconhecida pela presença de dois modos, cada um com um

pico característico. Determinar o fator que caracteriza as amostras que pertencem a cada uma das duas distribuições pode ser difícil. No domínio do perfil de expressão genética em larga escala, encontrar padrões de expressão bimodal é um importante processo analítico. Em oncologia, esse processo pode ser parte da busca de alvos terapêuticos clinicamente importantes nos tumores. O processo também pode revelar assinaturas moleculares que distinguem os subtipos de tumor, o que contribui para a compreensão das características clínicas e biológicas do câncer (WANG et al., 2009).

Devido à sua importância, a identificação sistemática de genes expressos bimodalmente tem recebido grande atenção, especialmente em pesquisas sobre o câncer. Vários métodos foram propostos durante a última década. Os métodos atuais podem ser colocados em duas categorias: modelos de mistura não paramétricos e normais. Por exemplo, o *cancer outlier profile analysis* é um método não paramétrico para pesquisar explicitamente genes com padrões de expressão "discrepantes". Métodos na segunda categoria postulam uma mistura de distribuições normais para níveis de expressão genética (TONG et al., 2013).

1.5 Caracterização do Problema

A quimioterapia neoadjuvante é um tratamento utilizado antes da cirurgia, com o intuito de reduzir ou acabar com o tumor. A adversidade é que tal tratamento nem sempre produz o resultado esperado, causando um sofrimento inútil ao paciente. Quando pacientes tem o câncer completamente erradicado são tratados como PCR (*pathologic complete response*), caso continuem existindo resíduos do tumor, os pacientes são caracterizados como NoPCR. Observou-se que a tentativa de prever a eficácia do tratamento utilizando somente dados clínicos não gerava bons resultados. Essa foi a motivação para pesquisadores buscarem na tecnologia de *microarrays* uma nova vertente para essa previsão. Tal abordagem traz como vantagem a possibilidade de compreender a nível genético o que motiva a ineficiência desse tipo de quimioterapia (American Cancer Society, 2018).

Existem basicamente duas formas de previsão de pacientes PCR, por intermédio das informações clínicas dos pacientes ou por meio de informações de níveis de expressão genética dos pacientes. Uma dificuldade para realizar a previsão de PCR por meio de dados de níveis de expressão genéticas é a gigante dimensão das bases de dados, para cada paciente são geradas expressões contendo mais de 22000 sondas, sendo que mais de uma sonda são utilizadas para caracterizar um gene. O custo para a obtenção dos dados também é definida como uma dificuldade, pois para gerar um *microarray* de um paciente é gasto uma alta quantia. Assim, tal problema tem poucos dados para a análise e um grande espaço de entrada. Existe também uma terceira dificuldade, as classes (PCR e NoPCR) são totalmente desbalanceadas, pois em aproximadamente 70 % dos casos não há resposta patológica completa (NoPCR) apresentada pelos pacientes, dificultando dessa forma a

utilização de classificadores, que ficam super ajustados.

Pode-se inferir que a redução do espaço de entrada facilitará as análises, possibilitando que no futuro *microarrays* menores possam ser utilizados para a verificação da sensibilidade à quimioterapia neoadjuvante, barateando o custo da produção do *array*, pois a quantidade de sondas utilizadas no processo de construção do *microarray* será menor.

1.6 Objetivos

Objetivo Geral:

• O projeto possui como objetivo geral propor um classificador para sensibilidade quimioterápica baseado na seleção dos genes de acordo com sua bimodalidade.

Objetivos Específicos:

• Pesquisar por bases de dados recentes de *microarray* que tratam do problema de sensibilidade quimioterápica;

• Propor uma metodologia de seleção de genes para a classificação de pacientes PCR/ NoPCR, baseada no cálculo da bimodalidade;

• Testar diferentes classificadores utilizando os genes selecionados;

• Análise dos resultados de acordo com a curva ROC.

1.7 Organização do Trabalho

O próximo capítulo busca apresentar os principais trabalhos relacionados a área de conhecimento estudado, tratando da predição de grupos PCR e NoPCR entre pacientes com câncer de mama. O terceiro capítulo traz a fundamentação teórica observada para a realização do trabalho, os métodos de classificação, as bases utilizadas e os métodos estatísticos para avaliar os experimentos.

O quarto capítulo caracteriza a metodologia utilizada para seleção, classíficação e análise estatística e também descreve os experimentos realizados. O quinto capítulo apresenta a análise estatística dos resultados obtidos nos experimentos e também comparações com trabalhos referenciados pelo autor. Por fim o sexto capítulo traz a conclusão de todo o estudo realizado, apresentando as opiniões do autor e ideias para trabalhos futuros.

2 Trabalhos Relacionados

O capítulo que se segue objetiva apresentar os trabalhos mais significantes relacionados a área de conhecimento e linha de estudos seguidas nesta dissertação, como também seus resultados para a literatura e suas contribuições para literatura.

2.1 Principais Trabalhos

2.1.1 Hess et al.

Em Hess et al. (2006) os autores desenvolveram um preditor multigênico visando caracterizar pacientes com resposta patológica completa (PCR) ao tratamento neoadjuvante utilizando Paclitaxel e Fluorouracil-doxorrubicina-ciclofosfamida (T/ FAC), além de avaliar a precisão preditiva em casos independentes. A escolha pelo regime de tratamento multidrogas se deu pelo fato de ser um tratamento clinico padrão para pacientes que necessitavam de tratamento citotóxico sistêmico.

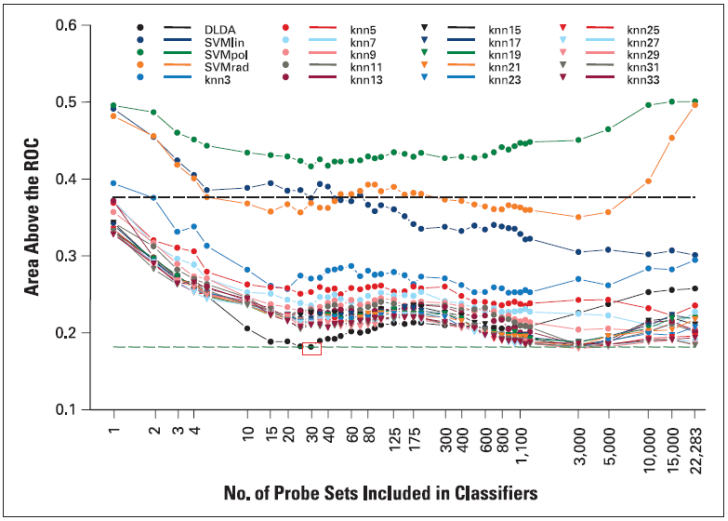
Foi utilizado um tamanho de amostra maior para a descoberta do preditor (n=82), e também uma tecnologia de expressão genética padrão no mercado, o *microarray*. Foi examinado sistematicamente o desempenho de um grande número de potenciais preditores por meio da técnica de validação cruzada nos dados de treinamento e selecionando um preditor final para validação independente em 51 novos casos. O estudo também examina a reprodutibilidade de resultados de previsão em 31 experiências replicadas.

O software dCHIP (Li, 2008) V1.3 foi usado para gerar medidas de qualidade, incluindo intensidade mediana e porcentagem de sonda definindo *outliers*. Este programa normaliza todas as matrizes para uma matriz padrão que representa uma fatia com intensidade global mediana. Valores de expressão gênica normalizados foram transformados para escala log10 por motivo de análise. Para identificar genes diferencialmente expressos entre casos com PCR (*pathologic complete response*) e aqueles com genes de RD (*residual disease*), de duas amostras, testes-t (CUI; CHURCHILL, 2003) foram realizados e a classificação dos genes foi ordenada pelos p-valores.

Os classificadores multigênicos foram construídos usando combinações dos genes mais informativos e vários algoritmos de previsão de classes diferentes, incluindo Support Vector Machines com núcleo linear, radial e polinomial (SEMOLINI, 2002), Análise diagonal Linear Discriminante (DLDA), e K-Nearest Neighbor (FRIEDMAN; HASTIE; TIBSHIRANI, 2001) usando distância euclidiana. A validação cruzada foi realizada para estimar o desempenho de previsão dos diferentes classificadores nos dados de treinamento e facilitar a seleção de um único classificador para validação independente.

Conforme é possível avaliar na Figura 5, o desempenho do classificador nos dados de validação foi avaliado usando a área acima da curva (AAC) característica de operação do receptor (ROC), complemento da área sob a curva ROC (AUC), calculada com 1-AUC. O critério foi avaliar o classificador que apresentou o menor valor de AAC. Além de variar o tipo de classificador eles também variaram a quantidade de sondas utilizadas em cada classificador. Segundo este critério o melhor classificador foi o DLDA com 30 sondas tendo acurácia sensibilidade de 0.92 e especificidade de 0.71 sobre o conjunto de validação.

Figura 5 – Resultados dos Classificadores - Hess et al. (2006).



Fonte: Pharmacogenomic Predictor of Sensitivity to Preoperative Chemotherapy With Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide in Breast Cancer

2.1.2 Horta

Horta (2008) nos apresenta uma nova metodologia que visa selecionar os genes mais informativos quanto à sensibilidade de um determinado paciente ao tratamento neoadjuvante, propondo classificadores com o intuito de prever a resposta patológica dos pacientes ao tratamento. O método utilizando DLDA foi comparado a dois importantes trabalhos da literatura, Natowicz (2008) e Hess et al (2006), mostrando resultados superiores para especificidade e inferiores para sensibilidade, portanto indicando prever melhor os pacientes caracterizados como RD e pior os PCR. A quantidade de sondas selecionadas também foi menor.

Como principais características da metodologia utilizada na seleção dos genes encontra-se o uso da técnica Volcano Plot, um *scatter-plot* de $-\log_1 0$ dos p-valores calculados pelo teste-t para cada sonda, pelo *fold-change* (CUI; CHURCHILL, 2003) calculado utilizando o $\log_2$ da média das razões entre as expressões dos genes de cada classe, com o objetivo de selecionar um primeiro conjunto significativo de genes/ sondas que detenham informações significativas quanto à influência da quimioterapia pré-operatória nos paci-

entes, além da proposta dos classificadores: Voto Majoritário (NATOWICZ, 2008), *Naive Bayes* (BISHOP, 2006), DLDA, *Suport Vector Machine* (SEMOLINI, 2002), Redes Neurais *Multi-Layer Perceptron* (BRAGA; CARVALHO; LUDERMIR, 2000), Redes Neurais Artificiais Multi-Objetivo (TEIXEIRA, 2001) e Redes Neurais Artificiais *least absolute shrinkage and selection operator* (TIBSHIRANI, 1996). A utilização dos classificadores tinha o intuito de prever a resposta patológica dos pacientes ao tratamento (PCR ou NoPCR).

Outro resultado representativo foi a obtenção de 2 conjuntos de sondas, um com 18 sondas, e outro com 11 sondas, subconjunto do primeiro, sendo utilizados para classificação por meio do classificador *Naive Bayes*. Salienta-se que os resultados foram próximos a Natowicz (2008), porém utilizando 1/3 das sondas.

Verificou-se que certamente as sondas mais significativas foram 205548_s_at, 204825_at, 219051_x_at, sendo apresentadas nos métodos de seleção dos trabalhos utilizados na comparação, Natowicz (2008) e Hess et al (2006). Em sua maior parte, as sondas verificadas pelos métodos de seleção estão relacionadas com processos biológicos ligados ao comportamento do câncer.

2.1.3 Wang et al.

Em Wang et al. (2009) os autores procuram identificar genes com padrões de expressão bimodal, que podem caracterizar lesões genômicas que ocorrem em alguns pacientes e em outros não. O objetivo é propor um novo critério, o índice de bimodalidade, que não só identifica, mas também classifica padrões bimodais significativos e confiáveis.

Como proposta surge o índice de bimodalidade (BI), um critério para pontuar e classificar os genes pela bimodalidade de suas expressões. A modelagem pressupõe que duas distribuições normais têm o mesmo desvio padrão. Os autores aplicaram o BI em conjuntos de dados de expressão gênica de *microarray* mRNA: A distribuição da expressão de um gene foi modelado por uma distância *inter-cluster* padronizada e pelo equilíbrio do modelo. A distância padronizada entre os grupos foi $\delta = \frac{|\mathrm{u}_1 - \mathrm{u}_2|}{\sigma}$, onde u_1 e u_2 são os valores médios dos modelos e σ foi o desvio padrão do modelo. O equilíbrio do modelo foi $\sqrt{\pi(1-\pi)}$, onde π e $1-\pi$ foram proporções das expressões em dois modelos. Desta forma a definição de BI foi

$$\delta\sqrt{\pi(1-\pi)}, \quad \text{onde} \quad \delta = \frac{|\mathrm{u}_1 - \mathrm{u}_2|}{\sigma} \tag{1}$$

Tal função de duas variáveis aumenta na distância *inter-cluster* δ e no equilíbrio do modelo. Conclui-se que o índice de bimodalidade foi desenvolvido para favorecer genes que detém alta distância *inter-cluster* e modelos bem balanceados.

O método proposto foi aplicado a um conjunto de dados contendo perfis de expressão de genes estudados relacionados ao câncer de mama. Diferentemente dos dados simulados,

a expressão gênica real dos dados foi muito mais complexa e continha um alto nível de ruído. Ao aplicar a técnica de agrupamento baseada em modelo de mistura com BIC, mais de 35% dos genes foram identificados como bimodal. Os autores tentaram remover algum ruído antes de analisar, definindo várias condições de filtragem. Ao aplicarem o método proposto aos conjuntos de dados filtrados, foram produzidos resultados razoáveis, sugerindo uma abordagem útil para a análise de conjuntos de dados reais. Os genes identificados como mais fortemente bimodais pareciam estar relacionados com o receptor hormonal (ERS1 e PGR) e status de HER2 (proteína ligada à presença do tumor) do câncer de mama pacientes. Foram descritos também dois genes adicionais (CKB e BST2) que exibiram fortes padrões de expressão bimodal dentro de amostras do câncer de mama.

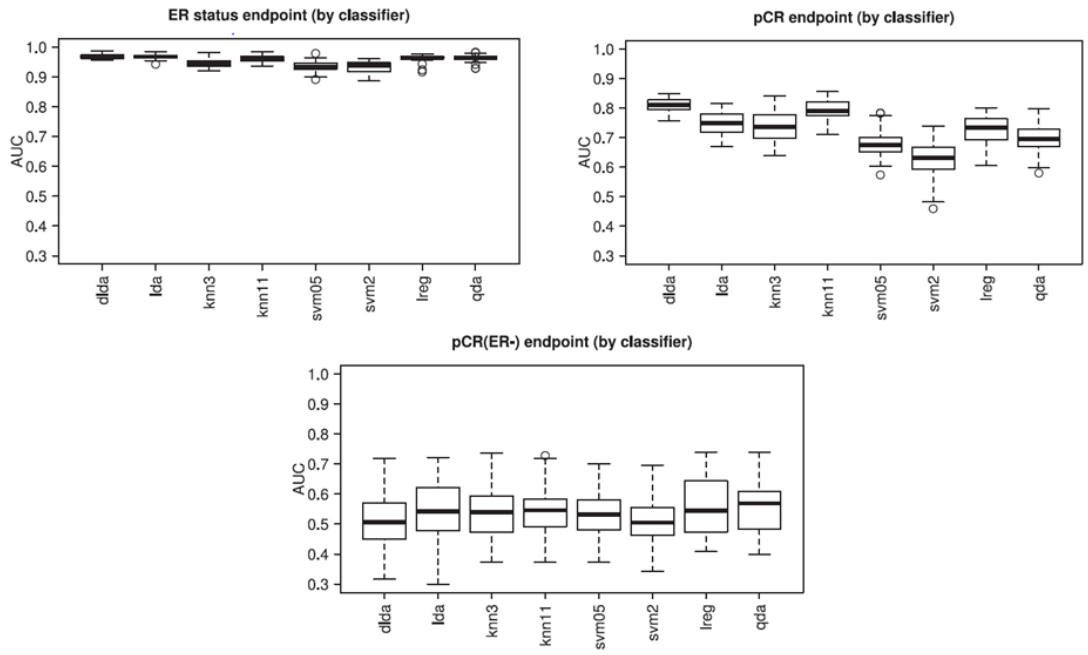
2.1.4 Popovic et al.

Em Popovic et al. (2010) apresenta-se uma tentativa de analisar como a escolha de variados métodos de seleção de sondas e posterior classificação, e suas diversas interações podem influenciar na qualidade de preditores genéticos em três problemas clínicos relacionados a sensibilidade a quimioterapia neoadjuvante: predição de *estrogen receptor status*, predição de pacientes PCR dentro de uma população contendo *estrogen receptor status* diferentes, e predição de pacientes PCR dentro de uma população contendo *estrogen receptor status* negativo, sendo o primeiro tratado como problema menos complexo e o último o mais complexo.

A base de dados de análise era composta por 230 pacientes, divididos em dois grupos: um de treinamento contendo 130 pacientes e um de validação contendo 100 pacientes. Para os autores, um preditor genético é caracterizado pela união entre um método de seleção e um classificador, partindo dessa definição 40 preditores foram avaliados, sendo cinco métodos de seleção, onde três utilizavam resultados de testes-t em seu cálculo e dois não, e oito classificadores utilizados, dentre eles o DLDA e K-Nearest Neighbor (HASTIE; TIBSHIRANI; FRIEDMAN, 2001).

Como melhores resultados obteve-se a seleção de 1502 genes para a predição de *estrogen receptor status*, 252 genes para a predição de pacientes PCR dentro de uma população contendo *estrogen receptor status* diferentes, e 5 genes para a predição de pacientes PCR dentro de uma população contendo *estrogen receptor status* negativo. É possível observar por meio da Figura 6 os resultados dos experimentos com todos preditores, é possível perceber que quanto menor o número de sondas utilizadas para execução dos algoritmos de classificação, pior era a performance.

Figura 6 – Resultados dos Preditores - Popovic et al. (2010).



Fonte: Effect of training-sample size and classification difficulty on the accuracy of genomic predictors

2.1.5 Hatzis et al.

Hatzis et al. (2011) nos apresenta um estudo realizado a partir de junho 2000 a março de 2010 no M. D. Anderson Cancer Center para desenvolver e testar preditores para resposta e sobrevivência à quimioterapia neoadjuvante em pacientes com câncer de mama recém-diagnosticado. Os pacientes detinham câncer de mama tratado com quimioterapia contendo regimes sequenciais de taxano e antraciclina.

Assinaturas preditivas diferentes para resistência e resposta à quimioterapia pré-operatória (neoadjuvante), estratificada de acordo com *estrogen receptor status*, foram desenvolvidas a partir de *microarrays* de expressão genética de câncer de mama. A sensibilidade do tratamento do câncer de mama foi então prevista usando a combinação de assinaturas para: (1) sensibilidade à terapia endócrina, (2) quimiorresistência, e (3) quimiossensibilidade. Como saída da análise obtém-se a sobrevida livre de recidiva à distância (DRFS), se o tratamento previsto leva a redução de risco sensível e absoluta (ARR), diferença em DRFS entre dois grupos de previsões, no período de 3 anos.

Concluiu-se que um preditor genético baseado no *estrogen receptor status* foi capaz de realizar uma previsão satisfatória para a quimiorresistência, a quimiossensibilidade e a sensibilidade endócrina identificando pacientes com alta probabilidade de sobrevida após quimioterapia com taxano e antraciclina.

2.1.6 Tong et al.

Em Tong et al. (2013) os autores atuaram na mesma linha de pesquisa do trabalho citado na subseção anterior, tendo objetivos similares. A inovação na proposta do artigo refere-se à percepção dos autores em notar que os métodos apresentados até então para o cálculo da bimodalidade sempre pressupunham distribuições normais, características dos dados oriundos de *microarrays*. Avaliando o surgimento de novas tecnologias que compartilhariam mercado com os *microarrays*, tendo por sua vez dados de natureza discreta, os autores propuseram uma generalização do modelo de Wang et al. (2009).

A modelagem proposta não assume que os desvios padrão ocorrem modo igual. O índice de bimodalidade generalizado foi

$$\frac{|u_1 - u_2|}{(1-\pi)\sigma_1^2 + \pi\sigma_2^2}\sqrt{\pi(1-\pi)}, \tag{2}$$

GBI é uma generalização do método BI, pois sua distância padronizada $\frac{|u_1-u_2|}{(1-\pi)\sigma_1^2+\pi\sigma_2^2}$ é responsável por distribuições bimodais de desvios padrão que ocorrem de modo diferente.

2.1.7 Carlson e Roth

Em Carlson e Roth (2013) os autores apresentam o Oncotype DX (ODX), o teste mais comum utilizado para investigação de proliferação do câncer de mama baseado em análise de expressão genética, desenvolvido pela Genomic Health. O teste avalia a expressão de 21 genes, dos quais 16 estão relacionados ao câncer e 5 tem função de referência, sendo associados à sinalização de estrógeno e proliferação do tumor. As informações acerca das expressões genéticas são combinadas com um algoritmo matemático, gerando dessa forma uma pontuação de recorrência, que é analisada pelos médicos.

Foi realizada uma revisão sistemática e meta-análise sobre o uso e o impacto do ODX nos estágios iniciais câncer de mama. A revisão concentrou-se na proporção de pacientes classificados com risco de recorrência alta, intermediária e baixa, na proporção geral e específica da categoria de pacientes que receberam tratamento adjuvante, na proporção de médicos que alteraram a recomendação do tratamento adjuvante após o teste ODX, no uso comparativo de tratamento adjuvante entre as categorias e a probabilidade dos pacientes receberem a recomendação de tratamento sugerido pelo teste ODX. A revisão resultou em 23 estudos (abrangendo sete anos) incluídos.

O estudo resultou em indagações acerca da assertividade de caracterização para pacientes com risco intermediário de recorrência do câncer, pois a quantidade de pacientes existentes nessa categoria não seguiu um padrão dentre os trabalhos analisados. Sendo assim, as decisões de estratégias de tratamento sugeridas pelo ODX e o impacto econômico

foi mais claro para casos em que os pacientes foram categorizados como de alto e baixo risco de recorrência.

Outro resultado interessante foi que pacientes com baixo risco de recorrência foram significativamente mais propensos a seguir o tratamento sugerido pelo ODX em relação aos pacientes com alto risco de recorrência. Tal informação sugere uma tendência para tratamento menos agressivo, apesar de existirem pacientes com alto risco de recorrência. Os autores entendem que houve uma falta de estudos sobre o impacto do ODX no uso do tratamento adjuvante contra outros padrões de abordagens, sugerindo mais estudos a partir do apresentado.

2.1.8 Beumer et al.

Em Beumer et al. (2016) os autores apresentam o MammaPrint, um diagnóstico por assinatura multigênica baseado em *microarray*, patenteado pela US Food and Drug Administration (FDA). Tais tipos de testes genômicos avaliam os níveis de expressão genética nos tecidos com o intuito de fornecer informações sobre o estado de uma doença e/ou prognóstico. O MammaPrint auxilia os médicos na tomada de decisões de tratamento adjuvante nos estágios iniciais do câncer de mama por meio de definição da expressão de 70 genes. O teste é adequado tanto para tecidos frescos quanto para congelados.

É informado que desde 2004 o MammaPrint é utilizado para determinar o risco de recorrência de câncer em pacientes. A expressão genética dos 70 genes é compilada em um índice com o intuito de estabelecer resultados qualitativos: Baixo Risco ou Alto Risco para recorrência. Pacientes designados como clinicamente de baixo risco podem ser tratados por terapia hormonal adjuvante, pacientes com câncer de mama clinicamente designados como alto risco são mais adequados para receber uma combinação de quimioterapia e terapia hormonal, ou seja, um tratamento neoadjuvante.

Amostras de controle são utilizadas normalmente como controles técnicos e experimentais dentro de cada lote de amostras. Dados do MammaPrint, cada uma composta de um pool de tumores de câncer de mama representativos, são continuamente monitorados no diagnóstico clínico, e foram disponibilizados para avaliação de estabilidade dos índices de teste MammaPrint no estudo em questão, totalizando n = 11.333 pontos de dados para análises.

Dados clínicos patológicos e dados de resultados clínicos de 345 amostras de pacientes foram analisados por meio do pacote SPSS 22.0 para Windows (SPSS Inc, Chicago, US). Análises de sobrevida foram realizadas para comparar desempenho de tecidos frescos e tecidos fixados em formalina e parafina (FFPE). Análises Kaplan-Meier foram utilizadas para comparar as distribuições MammaPrint para tecido fresco e FFPE para metástase em longo prazo como primeiro evento (DMF) e intervalo livre de recorrência

(DRFI). DMF foi definido como o tempo da cirurgia até o diagnóstico de uma metástase. Para DMF, pacientes que apresentam recidiva local, recidiva regional, ou segundo tumor primário antes do diagnóstico de uma metástase em longo prazo foi censurado em evento, na morte ou no final do seguimento. DRFI foi definido como o tempo desde a cirurgia até o diagnóstico de primeira metástase em longo prazo ou morte relacionada ao câncer de mama.

A reprodutibilidade foi medida em termos de estabilidade, calculada como 100 menos o desvio padrão relativo, onde o desvio padrão foi medido como um percentil do intervalo total de testes MammaPrint. A análise incluiu avaliações de precisão e repetitividade de Amostras de alto risco e baixo risco do MammaPrint. A precisão foi determinada calculando a precisão relativa de medições repetidas de mama de alto risco e baixo risco de câncer durante um período de 20 dias. A repetitividade foi determinada pelo cálculo da estabilidade relativa entre execuções duplicadas das medições realizadas cada dia. Tanto a estabilidade relativa quanto à precisão relativa são calculadas como 100 menos o desvio padrão relativo.

Os resultados do estudo mostraram que o MammaPrint é um teste estável, independentemente do tipo de *array* usado ser um *mini-array* personalizado ou *array* de genoma inteiro personalizado, ou o tipo de amostra ser fresca ou FFPE. A excelente reprodutibilidade dos índices MammaPrint de amostras de controle nos últimos 10 anos e a excelente precisão e repetitividade de amostras de alto risco e baixo risco do MammaPrint é mais um certificado da qualidade e robustez do teste. A principal força do estudo é o tamanho da amostra. Quase 2000 pares de amostras demonstraram quase perfeita concordância nos valores do índice MammaPrint entre os tipos de *mini-array* e genoma completo. Além disso, 552 pares de amostras (com dados clínicos disponíveis para 345 pares de amostras) contribuíram para a comparação entre tecido fresco e FFPE, com os resultados indicando excelente acordo. Avaliações que ultrapassam 11.000 amostras de controle exemplificam a alta estabilidade dos resultados do MammaPrint ao longo do tempo.

2.1.9 Carvalho

Em Carvalho(2017) o autor busca criar um classificador para a primeira etapa do tratamento do câncer de mama, de forma a identificar se o paciente irá apresentar Resposta Patológica Completa (PCR), ou seja, eliminação total da doença, ou se ele irá apresentar Doença Residual (RD), que ocorre quando o tumor não é eliminado totalmente.

Os métodos de seleção utilizados foram *Volcano Plot*, além de dois modelos de regressão, *stepwise regression* e *stepwise generalized linear model*, para recuperar as sondas mais significativas dentre as expressadas via tecnologia *microarray*. Para classificação o autor optou por utilizar redes neurais artificiais com 20 nodos seguindo os algoritmos *Backpropagation*, *Levenberg-Marquardt*, *Conjugate Gradient*, *Conjugate Gradient with Powell/Beale Restarts*, *Fletcher-Powell Conjugate Gradient* e *Polak Ribiere Conjugate*

Gradient (GARDNER; DORLING, 1998), além de um classificador baseado em *Particle Swarm Optimization* contendo 200 partículas e 2 clusters, e por fim um classificador que baseia-se em *Extreme Learning Machine* contendo 75 neurônios na camada escondida, dois agrupamentos (PCR e RD, com respostas 1 e 0 respectivamente), utilizando as funções de ativação Sigmoidal e Hardlim.

Como resultados a seleção via *Volcano Plot* apresentou 31 genes/ sondas, o modelo *stepwise regression* retornou 110 genes, e o modelo *stepwise generalized linear model* 151 genes/ sondas. O melhor resultado encontrado nos experimentos foi oriundo da união do 151 genes/ sondas selecionados no último método analisado e o classificador *Extreme Learning Machine*, que foi capaz de performar 83% de acertos na validação para a classe RD.

O autor conseguiu observar um futuro interessante para continuidade dos trabalhos no algoritmo de *Extreme Learning Machine*, devido a sua velocidade e também aos resultados apresentados, porém verificou que o desbalanceamento das classes poderia afetar o rendimento da classificação, visto que os percentuais de acerto foram diferentes entre as duas classes. Também constatou que não era interessante continuar os estudos de melhorias nos algoritmos baseados em redes neurais artificiais, pois os resultados apresentados não foram satisfatórios.

3 Fundamentação Teórica

Esse capítulo apresenta informações apoiados na literatura acerca dos métodos matemáticos e estatísticos utilizados para os experimentos relacionados a seleção, classificação e análise dos resultados, além de caracterizar as bases utilizadas nos dois experimentos realizados.

3.1 Cross Validation

A validação cruzada é uma técnica que tem como objetivo medir a capacidade de generalização de um modelo, a partir de um conjunto de dados. Tal técnica é largamente utilizada para resolução de problemas onde o objetivo da modelagem é a predição. Procura-se estimar a precisão deste modelo na prática, em outras palavras, o seu desempenho para um novo conjunto de dados (KOHAVI, 1995).

O particionamento do conjunto total de dados em subconjuntos mutualmente exclusivos é o principal pilar das técnicas de validação cruzada, e posteriormente, alguns destes subconjuntos são utilizados para estimar os parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos (dados de validação ou de teste) são empregados na validação do modelo.

Existem variadas formas para realizar o particionamento de dados, as principais são:

• *Holdout*: Baseia-se em dividir o conjunto de dados em dois subconjuntos definidos como treinamento e teste. Normalmente divide-se a informação em 70% para treinamento e 30% para teste.

• *K-fold*: Baseia-se em normalmente dividir o conjunto de dados em 5-10 grupos, a partir daí um dos grupos é escolhido para teste e o restante para treinamento do seu modelo. Repete-se o método até que todos grupos sejam utilizados como conjunto de teste.

Em qualquer um dos métodos de particionamento, a precisão final do modelo é obtida por:

$$AC_f = \frac{1}{v}\sum_{i=1}^{v}\epsilon_{yi,\hat{y}i} = \frac{1}{v}\sum_{i=1}^{v}(yi - \hat{y}i), \tag{3}$$

onde v é o número de dados de validação e $\epsilon_{\mathrm{yi},\hat{\mathrm{y}}\mathrm{i}}$ é o resíduo dado pela diferença entre o valor real da saída i e o valor predito. Desta forma, é possível inferir de forma quantitativa a capacidade de generalização do modelo.

No projeto em questão utilizamos uma mistura dos dois métodos definidos nesta seção. Onde utilizamos dez grupos diferentes, por meio de distintas permutações, também optando por executar experimentos utilizando um percentual de 70% para treinamento e 30% para validação, além de também realizar execuções com um percentual de 60% para treinamento e 40% para validação.

3.2 Classificador DLDA

O classificador DLDA (*Diagonal Linear Discriminant Analysis*) é definido como uma regra utilizada para resolução de problemas referentes à distribuição normal multivariada, sendo de máxima verossimilhança, onde a matriz diagonal de variância-covariância é a mesma para as distribuições das classes, sendo assim, a variância para cada variável é a mesma em todas as classes e as variáveis são não correlacionadas (SIMON et al., 2003).

O padrão x pertencerá à classe 1 caso

$$\sum_{i=1}^{S} \frac{(x_i - \bar{x}_i^{(1)})^2}{\sigma_i^2} \leq \frac{(x_i - \bar{x}_i^{(2)})^2}{\sigma_i^2}, \tag{4}$$

caso contrário pertencerá à classe 2. Na equação acima S é a quantidade de variáveis que compõem o padrão, $\bar{\mathrm{x}}^{(1)}$ é o valor médio da variável i para a classe 1, e $\bar{\mathrm{x}}^{(2)}$ é o valor médio da variável i para a classe 2, σ_{i}^2 é a variância da variável i sobre o conjunto formado pelas duas classes.

3.3 Particle Swarm Optimization com Clustering

O algoritmo em questão inspira-se na interação de bandos de pássaros, o modo como se comportam enquanto voam e também ao procurar por alimentos, que caracterizam por si só um mecanismo de otimização. O espaço de busca pode ser representado pela área sobrevoada, o ótimo local ou a solução ótima de um problema pode ser comparado com um local onde existe uma grande quantidade de alimento. Pássaros utilizam suas experiências e também de seu bando com o intuito de atingir um alvo. Para denominar tal experiência utiliza-se o termo pBest, para determinar o conhecimento de todo bando utiliza-se o termo gBest. Cada participante de um bando é representado como uma partícula no algoritmo, cada uma caracterizando uma possível solução para o problema em questão (AL-DUJAILI; TANWEER; SURESH, 2015).

Existe um vetor de posição em cada partícula, indicando a localização no espaço de busca, de forma similar a coordenadas. Outro atributo das partículas é a velocidade, com o objetivo de guiar as mudanças de posição das partículas durante a execução do processo. Com base nas experiências do bando as velocidades e posições são modificadas

dinamicamente. Em resumo, a variável gBest indica o melhor valor global, sendo utilizada como referência para o ajuste das partículas, e o pBest é o valor de cada partícula, que será ajustado baseado na variável indicativa de velocidade e na direção do gBest.

Da mesma forma que em Carvalho (2017) o algoritmo foi utilizado baseando-se em bibliotecas para o MatLab. O código está disponível em <https://drive.google.com/file/d/1uP8kP5DpsbY2MAtOdp8juXEZ-gSGizPd/view?usp=sharing>.

3.4 Extreme Learning Machine

Método também denominado Máquina de Aprendizado Extremo, é tratado como uma alternativa a utilização de redes neurais, buscando aprimorar o tempo de processamento para convergência dos pesos. Sua lógica de treinamento não necessariamente ajusta todos os pesos com grande frequência, podendo convergir de forma mais rápida (HUANG et al., 2008). O algoritmo se utiliza de duas camadas básicas, uma de entrada e outra oculta, mais simples. Os pesos são determinados de forma aleatória entre as camadas de entrada e oculta, e de forma analítica entre as camadas oculta e de saída. Esse método tem larga aplicação em problemas de classificação e identificação de padrões graças a sua velocidade de aprendizagem.

Da mesma forma que em Carvalho (2017) o algoritmo foi utilizado baseado-se em bibliotecas para o MatLab (NTU, 2013).

3.5 Curva ROC e Matriz de Confusão

Dado um classificador e uma instância, há quatro resultados possíveis: Se a instância for positiva e classificada como positiva, é contada como um verdadeiro positivo; se for classificado como negativo, é contado como falso negativo. Se a instância for negativa e for classificada como negativa, é contado como um verdadeiro negativo; se for classificado como positivo, é contado como um falso positivo. De posse das informações anteriores podemos calcular duas taxas: A taxa de verdadeiros positivos, que é definida por meio da divisão do número de verdadeiros pelo número total de positivos dentro do conjunto de testes; e a taxa de falsos positivos, definida por meio da divisão do número de falsos positivos pelo número total de negativos dentro do conjunto de testes (FAWCETT, 2004).

Tendo em mãos o conjunto de instâncias e seus respectivos resultados, é possível criar uma matriz de confusão 2x2, que é base para o cálculo das métricas mais comuns. A Figura 7 apresenta uma matriz de confusão, é possível observar que os números dispostos na diagonal principal representam as decisões corretas tomadas por um determinado classificador, e as duas taxas restantes demonstram as decisões erradas, ou confusões, daí o nome dado a matriz. Conseguimos verificar também na Figura 7 a forma de calcular a

sensibilidade, que indica a taxa de predições corretas para pacientes PCR; a especificidade, que indica a taxa de predições corretas para pacientes NoPCR; e a acurácia, que indica a taxa de predições corretas em toda população.

Figura 7 – Matriz de Confusão

		Classes Reais	
		p	n
Resultado Classificador	p	Verdadeiros Positivos	Falsos Positivos
	n	Falsos Negativos	Verdadeiros Negativos
Total		P	N

Taxa VP/ Sensibilidade:	$\frac{VP}{P}$	Taxa FP:	$\frac{FP}{N}$
Especificidade:	$\frac{VN}{N}$	Acurácia:	$\frac{VP+VN}{P+N}$

Fonte: ROC Graphs: Notes and Practical Considerations for Researchers

O gráfico ROC (*Receiver Operating Characteristics*) é um espaço bidimensional onde a taxa de verdadeiros positivos é plotada no eixo Y e a taxa de falsos positivos no eixo X. A ferramenta em questão possibilita visualizar um comparativo entre os benefícios (taxa de verdadeiros positivos) e os custos (taxa de falsos positivos) referentes à classificação. Um classificador discreto é aquele que apresenta uma classe de saída, classificadores com essa característica produzem um par (taxa de verdadeiros positivos, taxa de falsos positivos) correspondente a um único ponto no gráfico ROC. Quanto mais a noroeste do gráfico um ponto estiver melhor será o desempenho do classificador, pois maior será a taxa de verdadeiros positivos e menor será a taxa de falsos positivos. Pontos na diagonal onde X=Y representam classificadores aleatórios.

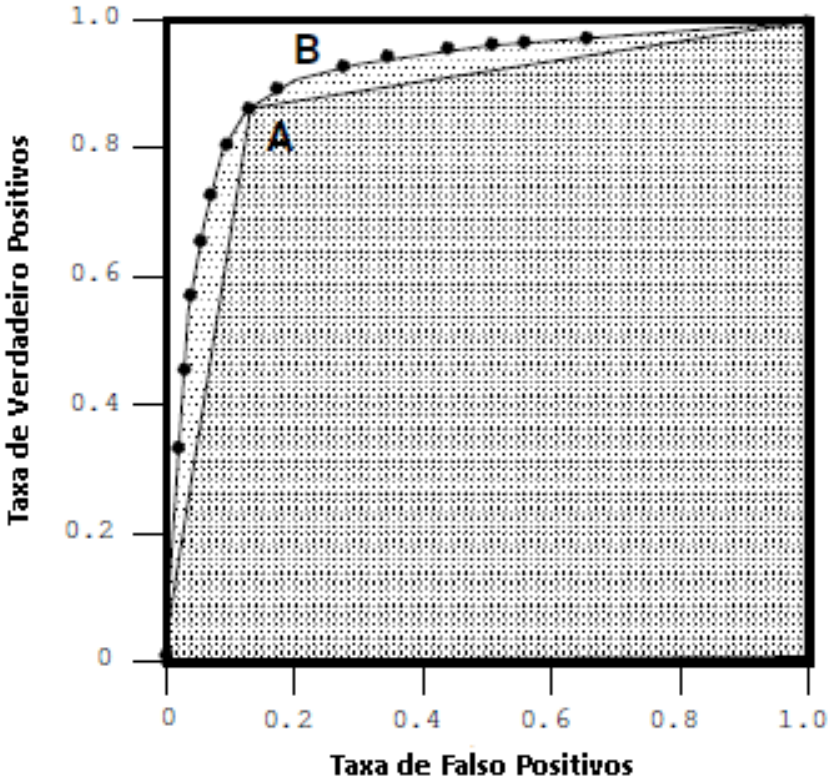
Existem classificadores que produzem uma pontuação ou probabilidade de instância, um valor numérico para representar o grau no qual uma instância está inserida uma classe. Estes valores podem seguir padrões e teoremas de probabilidade ou podem ser pontuações

gerais e não calibradas, nesse caso o único método de avaliação é que uma maior pontuação indica uma probabilidade maior.

Tal classificação ou classificador de pontuação pode ser utilizado como um limite para produzir um classificador discreto (binário): dependendo da saída do classificador pode se produzir um Y, senão um N. Cada valor limite produz um ponto diferente no espaço ROC. Podemos então traçar uma curva variando de $-\infty$ a $+\infty$ no gráfico ROC, tal curva é chamada de curva ROC.

Com a intenção de comparar classificadores, pode ser interessante reduzir o resultado do ROC para um único valor escalar que represente o desempenho esperado. Um método largamente utilizado é calcular a área sob a curva ROC. O valor resultante do cálculo em questão sempre estará entre 0 e 1, sendo que um classificador realista, ou seja, não aleatório, deve deter área resultante maior que 0.5. É possível verificar dois modelos de áreas sob a curva ROC na Figura 8.

Figura 8 – Gráfico Ilustrando Duas Curvas ROC e a área abaixo delas (Figura Adaptada)



Fonte: ROC Graphs: Notes and Practical Considerations for Researchers

3.6 Bases de Dados

No primeiro experimento utilizou uma base de pacientes prospectivamente fornecidos por consentimento escrito que possibilita a obtenção uma amostra de biópsia do tumor por

aspiração ou biópsia central antes de qualquer terapia sistêmica para estudos genéticos e testes de preditores de tratamento. Todos os *microarrays* de expressão genética foram perfilados no Departamento de Patologia do M. D. Anderson Cancer Center (MDACC), em Houston, Texas. A base contém 488 pacientes e foi utilizada em Hatzis et al. (2011).

No segundo experimento utilizou-se a base GSE20194, intitulada de *MAQC-II Project: human breast cancer*, publicada em fevereiro de 2010 e atualizada em Agosto de 2018, sendo uma contribuição da *University of Texas M. D. Anderson Cancer Center*, contendo um total de 268 pacientes. Os pacientes receberam 6 meses de quimioterapia pré-operatória (neoadjuvante), incluindo paclitaxel, 5-fluorouracil, ciclofosfamida e doxorrubicina, seguida de ressecção cirúrgica do câncer. A extração de RNA e perfil de expressão genética foram realizados em múltiplos lotes ao longo do tempo usando *microarrays* Affymetrix U133A (NCBI, 2018). A primeira versão da base GSE20194 foi utilizada em Popovic et al. (2010).

4 Metodologia

O capítulo que se segue apresenta a metodologia proposta, que caracteriza-se por realizar a seleção de sondas e posteriores classificações utilizando diferentes algoritmos: DLDA, *Particle Swarm Optimization* e *Extreme Machine Learning*.

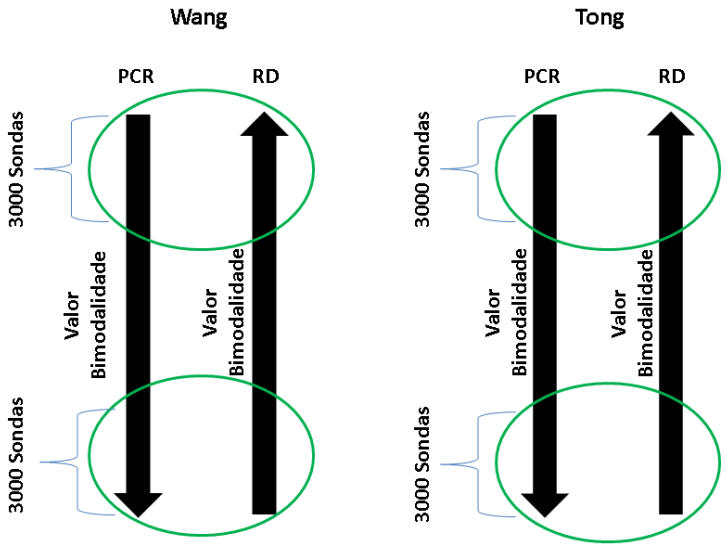
4.1 Seleção de Sondas

A primeira série de experimentos, ilustrada pela Figura 9, se baseia em selecionar as sondas mais significativas por meio da execução dos métodos de bimodalidade Tong (2013) e Wang (2009) para pacientes caracterizados na base de dados como doença residual (RD) e resposta patológica completa (PCR). Após obter o valor do índice de bimodalidade das 22283 sondas expressas, ordena-se as mesmas de forma crescente para cada conjunto de pacientes de cada método utilizado (RD_WANG, PCR_WANG, PCR_TONG e RD_TONG). A partir daí recupera-se as 3000 primeiras sondas em cada conjunto e as armazena. O próximo passo é ordenar de forma decrescente cada conjunto de pacientes de cada método utilizado e também recuperar as 3000 primeiras sondas e armazená-las.

Com as sondas selecionadas observando as regras do parágrafo anterior iniciam-se comparações entre os conjuntos caracterizados como divergentes dentro de cada método buscando sondas pertencentes a um conjunto de interseção entre os mesmos, as 3000 primeiras sondas ordenadas de forma crescente para pacientes RD executados com o método Wang e as 3000 primeiras sondas ordenadas de forma decrescente para pacientes PCR executados com o método Wang, e assim por diante:

- RD_WANG_Decrescente X PCR_WANG_Crescente;
- RD_TONG_Crescente X PCR_TONG_Decrescente;
- RD_TONG_Decrescente X PCR_TONG_Crescente.

Figura 9 – Metodologia Proposta para Seleção de Sondas.



Fonte: Elaborado pelo Autor

4.2 Classificação Baseada em Sondas Selecionadas

Conhecendo as sondas selecionadas por intermédio dos experimentos realizados na Seção 4.1, utilizou-se três classificadores distintos para predizer a resposta dos pacientes ao tratamento neoadjuvante para o câncer de mama. Os classificadores escolhidos foram DLDA, *Particle Swarm Optimization* com clustering e *Extreme Learning Machine*.

É importante destacar a utilização do *cross validation* como uma tentativa de tratar o desbalanceamento entre as classes, realizando execuções contendo diferentes permutações de pacientes nos arquivos de treinamento e validação, além de particionar os pacientes em grupos de 60% para treinamento e 30% para validação ou 70% para treinamento e 30% para validação.

A partir daí, contendo os dados de verdadeiros positivos, falsos positivos, verdadeiros negativos, falsos negativos, sensibilidade e especificidade resultantes do classificador, calcula-se a área sob a curva ROC, possibilitando a análise quanto ao desempenho do classificador utilizado.

4.2.1 DLDA

A classificação por meio do DLDA (*Diagonal Linear Discriminant Analysis*) justifica-se por tratar de um método amplamente utilizado em literatura, como em Hess et al. (2006) e Horta (2008), e de simples configuração, onde a matriz de agrupamento caracterizou 0 como PCR e 1 como RD.

4.2.2 Particle Swarm Optimization

Como em Carvalho (2017), ajustou-se o algoritmo de *Particle Swarm Optimization* para conter 200 partículas e 2 *clusters*, cada um representando um valor provável para a saída. Como função objetivo utilizada optou-se pela distância Euclidiana, onde a cada iteração os possíveis *clusters* são verificados com relação aos dados, na tentativa de encontrar os que melhor separam os dois grupos de classificação. O algoritmo foi ajustado para não utilizar os dados dos melhores globais, garantindo que cada partícula possa convergir para um melhor local, gerando assim os *clusters* necessários para fazer a posterior classificação. Serão gerados dois *clusters*, um relativo a saída PCR e outro relativo à saída RD, e então é feito o cálculo da distância euclidiana dos pacientes para cada um dos *clusters*. O que possuir a menor distância será utilizado para associar o paciente ao seu grupo de referência.

4.2.3 Extreme Learning Machine

Também observando Carvalho (2017), O algoritmo de *Extreme Learning Machine* foi configurado para a opção "executar classificação", os arquivos de entrada contendo uma coluna identificando duas classes distintas, 0, indicando PCR, e 1, indicando RD, dessa forma orientando o algoritmo na realização da classificação. Como parâmetros foram passados 75 neurônios na camada escondida e as funções de ativação Sigmoidal e Hardlim, salientando que só é possível passar uma delas por execução. A tática de *cross validation* utilizada nas execuções do algoritmo DLDA também é realizada para as execuções do algoritmo em questão, adicionando execuções contendo tentativas de balancear as classes mesclando os pacientes PCR de diferentes bases com o intuito de igualar a quantidade de pacientes das duas classes analisadas no arquivo de treinamento. Essa tática adicional de *cross validation* justifica-se na suposição de Carvalho (2017) onde infere-se que o desbalanceamento das classes no arquivo de treinamento poderia afetar na assertividade do algoritmo classificador em questão.

5 Análise e Discussão dos Resultados

O capítulo a seguir apresenta os resultados dos experimentos realizados, e também as análises estatísticas referentes a cada um, buscando caracterizar os dados e também compará-los aos principais trabalhos referenciados nesta dissertação.

5.1 Seleção de Sondas

O experimento de seleção de sondas ocorreu com a utilização da base de Hatzis et al. (2011). A metodologia resultou em quatro sondas selecionadas para cada método de bimodalidade utilizado, tendo a sonda 205509_at presente nos dois conjuntos de seleção. No contexto dessa dissertação a interseção informada não tem significado nas análises, visto se tratar de dois métodos de bimodalidade comparados entre si. Para um trabalho futuro cabe entender se a interseção entre os dois métodos pode indicar uma sonda de alta representatividade na classificação de pacientes.

De acordo com a NetAffx Query (Thermo Fisher Scientific, 2017), a sonda 208442_s_at, selecionada para o método de bimodalidade de Tong et al. (2013), é relacionada às funções do DNA, exercendo participação na proliferação do câncer no corpo humano, e podendo indicar significância na área de pesquisa em questão. Quanto as outras sondas, verifica-se funções moleculares de ligação proteica e peptidase, podendo, por meio de testes de bancada indicar alguma importância no contexto estudado. Na Figura 10 é possível verificar o gene relacionado às sondas e suas respectivas funções moleculares.

Figura 10 – Informações acerca das Sondas Selecionadas na Metodologia Proposta.

Método Proposto - Wang			Método Proposto - Tong		
Sonda	**Gene**	**Função Molecular**	**Sonda**	**Gene**	**Função Molecular**
215733_x_at	CTAG2	Ligação proteica	221424_s_at	OR51E2	Atividade de transdutor de sinal Atividade de receptor acoplado à proteína G
213240_s_at	KRT4	Atividade da molécula estrutural Ligação proteica	205509_at	CPB1	Atividade de carboxipeptidase Atividade de peptidase Atividade de hidrolase
205509_at	CPB1	Atividade de carboxipeptidase Atividade de peptidase Atividade de hidrolase	220922_s_at	LOC105369237	Ligação proteica
207885_at	S100G	Ligação de vitamina D Ligação de ion cálcio Ligação de íons de metal	208442_s_at	ATM	Complexo de reparo de DNA

Fonte: Elaborado pelo autor

Na Figura 11 é possível verificar as sondas selecionadas pela metodologia proposta, dividindo por método de bimodalidade aplicado, e também uma comparação com as sondas recuperadas em outros trabalhos citados nesse projeto. Verifica-se que a quantidade de

sondas significativas selecionadas é relativamente pequena se compararmos com os trabalhos de Hess et al. (2006) e Horta (2008).

Figura 11 – Comparação de Sondas Selecionadas por Métodos da Literatura e o Método Proposto.

Hess et al		Horta		Método Proposto - Wang		Método Proposto - Tong	
Sonda	Gene	Sonda	Gene	Sonda	Gene	Sonda	Gene
203929_s_at	MAPT	205548_s_at	BTG3	215733_x_at	CTAG2	221424_s_at	OR51E2
203930_s_at	MAPT	204825_at	MELK	213240_s_at	KRT4	205509_at	CPB1
212745_s_at	BBS4	204913_s_at	SOX11	205509_at	CPB1	220922_s_at	LOC105369237
203928_x_at	MAPT	219051_x_at	METRN	207885_at	S100G	208442_s_at	ATM
212207_at	THRAP2	210147_at	ART3				
217542_at	MBTPS1	217028_at	CXCR4				
206401_s_at	MAPT	201508_at	IGFBP4				
215304_at	PDGFRA	203929_s_at	MAPT				
219741_x_at	ZNF552	205225_at	ESR1				
204916_at	RAMP1	214164_x_at	CA12				
208945_s_at	BECN1	205044_at	GABRP				
213134_x_at	BTG3	208370_s_at	DSCR1				
219197_s_at	SCUBE2	220559_at	EN1				
204825_at	MELK	202342_s_at	TRIM2				
205548_s_at	BTG3	217838_s_at	EVL				
202204_s_at	AMFR	203628_at	IGF1R				
209617_s_at	CTNND2	203789_s_at	SEMA3C				
205354_at	GAMT	221728_x_at	XIST				
204509_at	CA12						
214124_x_at	FGFR1OP						
213234_at	KIAA1467						
219051_x_at	METRN						
219044_at	FLJ10916						
203693_s_at	E2F3						
214053_at	ERBB4						
215616_s_at	JMJD2B						
209773_s_at	RRM2						
219438_at	FLJ12650						
205696_s_at	GFRA1						
201508_at	IGFBP4						

Fonte: Elaborado pelo autor

5.2 Classificação Baseada em Sondas Selecionadas

5.2.1 DLDA

Dois experimentos foram realizados para o classificador DLDA. O primeiro utilizando a base de Hatzis et al. (2011), que contém 488 pacientes, dos quais 99 são classificados como PCR e 389 como RD. O segundo experimento utilizou a base GSE20194, contendo 268 pacientes, dos quais 55 são classificados como PCR e 213 como RD.

Para o primeiro experimento é possível verificar na Figura 12 a comparação dos resultados estatísticos referentes à classificação de pacientes do experimento realizado e os trabalhos de Hess et al. (2006) e Horta (2008), nos quais também foi utilizado o classificado DLDA. Por meio da tabela observa-se que os resultados referentes à sensibilidade, ou seja, o quanto os pacientes verdadeiro-positivo representam no contexto de todos os casos de RD, e especificidade, ou seja, os pacientes considerados verdadeiro-negativo em relação aos

indivíduos diagnosticados como RD, são piores para os experimentos realizados seguindo a metodologia proposta.

Figura 12 – Comparação entre os Resultados de Classificação Utilizando DLDA.

Metodologia	**Hess et al**	**Horta**	**Método Proposto**			
Treinamento - Validação			**70-30(%)**		**60-40(%)**	
Método Aplicado			**Wang et al**	**Tong et al**	**Wang et al**	**Tong et al**
Sensibilidade	0.92308	0.84615	0,726495726	0,521367521	0,6794872	0,49359
Especificidade	0.71053	0.86842	0.7	0.733333333	0.575	0.725

Fonte: Elaborado pelo autor

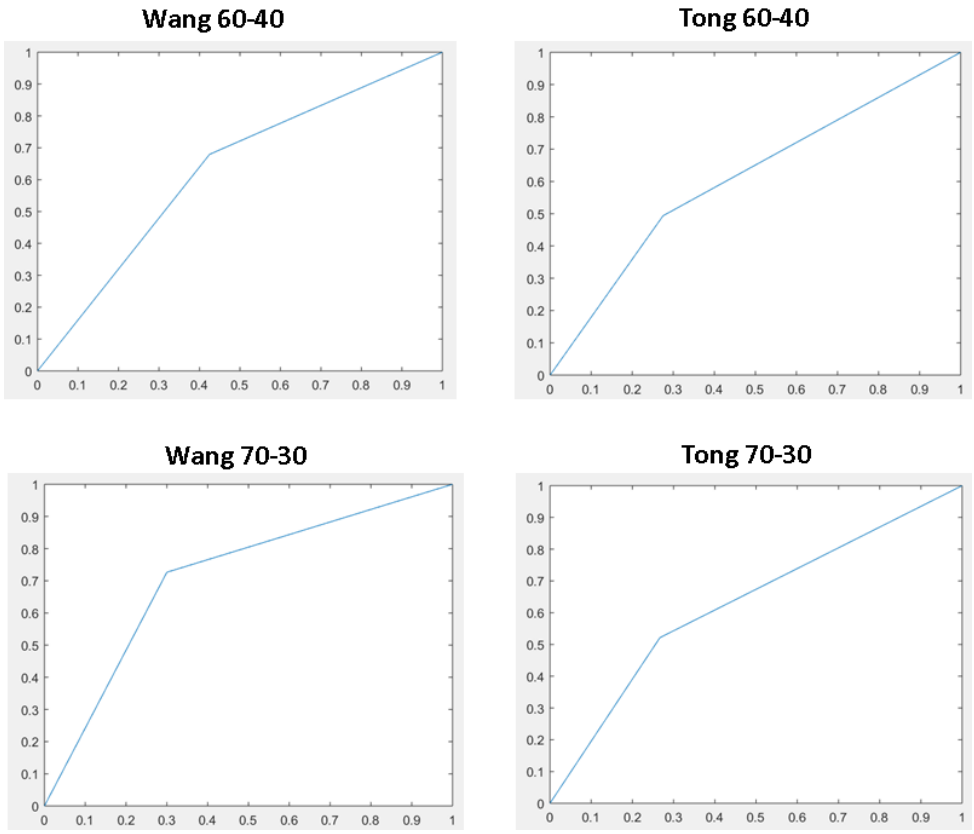
Na Figura 13 é possível verificar a área sob a curva ROC resultante do experimento 01. Os dados não são satisfatórios, visto que espera-se algo acima de 0.8 para que a classificação seja considerada significativa. É possível verificar os gráficos ROC plotados em Figura 14.

Figura 13 – Área sob a Curva ROC - DLDA - Base Hatzis et al. (2011).

		AUC
Tong	Exp. 60-40(%)	0.609294871794872
Wang	Exp. 60-40(%)	0.627243589743590
Tong	Exp. 70-30(%)	0.627350427350428
Wang	Exp. 70-30(%)	0.713247863247863

Fonte: Elaborado pelo autor

Figura 14 – Curvas ROC DLDA - Base Hatzis et al. (2011).



Após a realização dos experimentos com a base GSE20194 observou-se que as métricas estatísticas continuaram abaixo das encontradas nos trabalhos de Hess et al. (2006) e Horta (2008) se comparados no tocante a utilização de tal classificador, essa afirmação pode ser verificada comparando-se a Figura 12 e a Figura 15. O fato pode ser explicado por se tratar de um classificador padrão e simples, e o número de sondas selecionadas para ambos os métodos de bimodalidade e em ambos experimentos ser pequeno em relação aos outros trabalhos comparados.

Verifica-se em Popovic et al. (2010) que utilizando a primeira versão da base GSE20194, visando a predição do problema analisado nessa dissertação, foram selecionadas 252 sondas, observando valores de área sob a curva ROC próximos a 0,8 utilizando DLDA. Ao realizar a predição de um problema similar ao anterior, porém com uma população restrita, o número de sondas selecionadas diminuiu para 5, e a área sob a curva ROC para o classificador DLDA cai para valores entre 0,45 e 0,57 na maioria dos experimentos, se aproximando da realidade dos resultados apresentados nessa dissertação.

Figura 15 – Métricas Estatísticas - DLDA - Base GSE20194.

Método Proposto		AUC	Especificidade	Sensibilidade
60-40	Wang	0,629144	0,576471	0,681818
	Tong	0,627540	0,482353	0,772727
70-30	Tong	0,642004	0,578125	0,705882
	Wang	0,642004	0,578125	0,705882

Fonte: Elaborado pelo autor

5.2.2 Particle Swarm Optimization

Após a realização de execuções respeitando a variedade dos métodos de bimodalidade propostos observou-se resultados de clusterização não satisfatórios, pois um percentual majoritário dos pacientes de ambas as classes contidos nos arquivos de treinamento foram encaminhados para um cluster em especial. A partir dessa informação infere-se que esse método de clusterização falho, que em seguida embasaria a tentativa de classificação dos pacientes existentes no arquivo de validação, torna essa metodologia não significativa.

Segundo Carvalho (2017), existe uma ideia de que a quantidade de expressões genéticas, ou seja, sondas, utilizadas para caracterizar os pacientes no arquivo de treinamento poderiam afetar os resultados. Em relação a afirmação anterior, entende-se que o fato de quatro sondas serem selecionadas por método de bimodalidade interferiu nos resultados. Como o objetivo desta dissertação é encontrar o menor número possível de sondas, infere-se que o algoritmo de *particle swarm optimization* com clusterização não é significativo no contexto do projeto.

5.2.3 Extreme Learning Machine

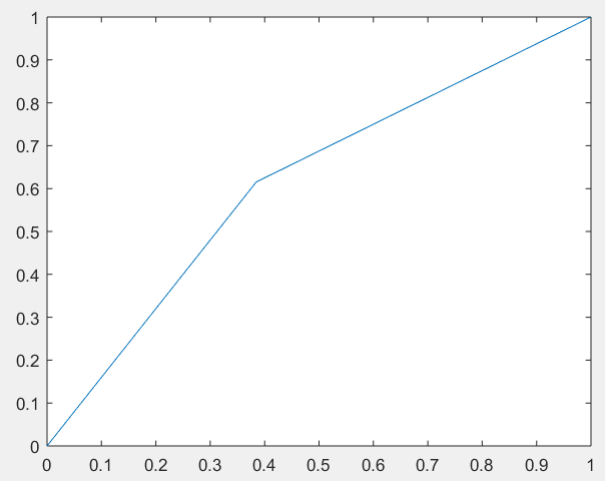
Como já citado na Seção 4.2, Carvalho (2017) menciona que a abordagem relacionada ao algoritmo *Extreme Learning Machine* foi a que obteve melhores resultados, porém o autor observou que a qualidade das métricas atribuia-se somente a uma das classes, RD, ou NoPCR no contexto da dissertação aqui apresentada. Como justificativa para tal comportamento ele propõe o desbalanceamento entre as classes. Ao realizar os experimentos observou-se que o comportamento verificado em Carvalho (2017) ainda se fazia presente, desta forma optou-se por também realizar execuções contendo a quantidade de pacientes em cada classe no arquivo de treinamento de forma balanceada.

Os resultados foram satisfatórios para a classe NoPCR em ambos métodos de bimodalidade, inclusive detendo resultados melhores que em Carvalho (2017), porém para a classe PCR não foram obtidos resultados satisfatórios, conseguindo como melhor resultado um experimento de classificação utilizando o método proposto em Wang et al. (2009) 0,6154

de acurácia para classe PCR e 0,9294 em NoPCR. Os resultados em questão apresentam uma melhora nas métricas em relação aos experimentos com DLDA, além de observar um valor interessante de *area under the curve*, apresentando valor acima de 0,6, demonstrando não se tratar de um classificador aleatório.

Conforme afirmado em Fawcett (2014), quanto mais a noroeste um ponto estiver no gráfico ROC, melhor será seu desempenho de classificação, pois maior será a taxa de verdadeiros positivos e menor será a taxa de falsos positivos. Observando a Figura 16 pode-se observar que ainda existe uma quantidade significativa de falsos positivos.

Figura 16 – Curva ROC Extreme Learning Machine - Bases Hatzis et al. (2011)/ GSE20194 - Wang.



Por tudo informado nessa seção, observa-se um classificador que se apresenta interessante dentro do contexto analisado nesta dissertação, demonstrando entregar resultados interessantes com uma quantidade pequena de sondas. Também é possível verificar que as sondas selecionadas por meio de Wang et al. (2013) foram mais significativas como base para uma posterior classificação.

5.2.3.1 Seleção de Sondas e Classificação Adicional

Percebendo o potencial do método de classificação por meio do ELM, decidimos realizar mais experimentos utilizando a metodologia de seleção proposta utilizando uma interseção maior de sondas e realizando uma posterior classificação utilizando o ELM para verificar se encontraríamos resultados mais significativos dos que observados até então.

Foram realizados experimentos observando uma interseção de 3500, 4000, 4500 e 5000 sondas. No tocante a classe NoPCR nenhum experimento se mostrou mais significativo do que foi verificado nos experimentos originais, porém no que se refere a classe PCR observou-se um melhor resultado significativo para classificação utilizando-se a seleção por meio da interseção de 4500 sondas. O experimento em questão utilizou-se do método de

Tong et al. (2013) e resultou em uma acurácia de 0,8077 para a classe PCR e 0,7736 para NoPCR.

É possível verificar quais foram as sondas selecionadas por meio da interseção das 4500 sondas do método Tong et al. (2013) e informações acidicionais acerca das mesmas na Figura 17.

Figura 17 – Informações acerca das Sondas Selecionadas no experimento adicional.

Sondas Selecionadas		
Sonda	**Gene**	**Função Molecular**
204600_at	EPHB3	Ligação de nucleótidos Ligação proteica Ligação de ATP
205509_at	CPB1	Atividade de carboxipeptidase Atividade de peptidase Atividade de hidrolase
207885_at	S100G	Ligação de vitamina D Ligação de ioncálcio Ligação de ions de metal
208442_s_at	ATM	Complexo de reparo de DNA
210540_s_at	B4GALT4	Ligação de ións de metal
213240_s_at	KRT4	Atividade da molécula estrutural Ligação proteica
214884_at	MCF2	Ligação proteica
220922_s_at	LOC105369237	Ligação Proteica
221424_s_at	OR51E2	Atividade de transdutor de sinal Atividade de receptor acoplado à proteína G

Fonte: Elaborado pelo autor

Os resultados apresentados nessa seção criam uma discussão em torno do que é mais importante: Classificar corretamente os PCRs ou os NoPCRs. Independente da situação em questão é de extrema importância atingir um equilíbrio entre a acurácia referente as duas classes, que deve ser o objetivo dos trabalhos futuros nessa área de conhecimento.

6 Conclusão

É possível concluir analisando os resultados do trabalho em questão que a proposta de metodologia é promissora, visto o número de sondas selecionadas e suas funções bioquímicas no organismo humano, uma relacionada à proliferação do câncer e outras que por meio de testes de bancada podem se mostrar significantes no contexto desse trabalho. Como foi descrito anteriormente, trata-se de um problema que afeta milhões de pessoas e está relacionado a um processo de desgaste físico e psicológico desnecessário, acarretando grande sofrimento em casos onde uma caracterização é feita de maneira errônea.

Outro ponto de observação é o grau de dificuldade em se executar a metodologia, que não é tão alto e custoso. Portanto podemos inferir que a metodologia proposta pode vir a ser um caminho interessante a se seguir na área de pesquisas médicas, pois quanto menos sondas forem necessárias para avaliação menor será o custo em processos laboratoriais, como a expressão genética via Microarray.

Ao realizar os experimentos com uma segunda base, mais recente, verificou-se que o desempenho dos algoritmos DLDA e *Particle Swarm Optimization* não foi satisfatório. O comportamento indicado nesse parágrafo explica-se pela quantidade de sondas selecionadas, e portanto utilizadas na caracterização dos pacientes nos arquivos de treinamento, tal quantidade é pequena em relação aos trabalhos referenciados e que obtiveram sucesso. Tal raciocínio leva a inferência de que os dois classificadores em questão não tem relevância no contexto apresentado.

Validando as curvas ROC e as métricas estatísticas verificou-se que o desempenho do algoritmo *Extreme Learning Machine* como classificador se mostrou superior aos encontrados em Carvalho (2017) no tocante a classe NoPCR, ou RD para o trabalho referenciado, também foi possível obter bons resultados para a classe PCR, por meio de um balanceamento de classes realizado utilizando pacientes das duas bases para igualar os dados das duas classes de caracterização no arquivo de treinamento e a utilização do método de seleção baseado na interseção de 4500 sondas.

Portanto a metodologia, unindo a proposta de seleção e o classificador *Extreme Learning Machine*, e também o método de bimodalidade de Tong et al. (2013), resultaram em um desempenho interessante para classe PCR utilizando o método de seleção baseado na interseção de 4500 sondas. Desta forma avaliou-se que a proposta pode vir a se tornar uma referência na literatura, sempre salientando que mais testes estatísticos e com bases maiores, talvez outras uniões entre bases diferentes das utilizadas.

6.1 Trabalhos Futuros

Conforme já informado anteriormente, ainda é necessário a utilização de outras técnicas estatísticas para validar o desempenho do melhor classificador encontrado nesse trabalho, além de encontrar novas bases, e se possivelmente mais volumosas, para realizar outros balanceamentos entre as classes de pacientes no arquivo de treinamento, e também uniões entre tais bases buscando tal objetivo.

Outro trabalho futuro é a verificação de classificadores mais robustos que o ELM, buscando apresentar uma metodologia na qual exista diversidade de possíveis classificadores, podendo também combinar algoritmos na busca de resultados melhores. Sempre realizando testes e comparações com os resultados já encontrados nesta dissertação para validar a qualidade de cada classificador.

No tocante a metodologia de seleção, é interessante buscar novos meios de calcular a bimodalidade utilizados na literatura, além de procurar outros métodos matemáticos para filtrar as sondas encontradas e validar a significância das mesmas no contexto avaliado.

Também será proposto a utilização dessa metodologia para tratar o contexto de outras oncologias, validando assim a solução como um método interdisciplinar e acessível para os vários casos relacionados a diferentes estados clínicos, e até problemas relacionados a classificação que resultará em dois conjuntos distintos.

Referências

AL-DUJAILI, A.; TANWEER, M.; SURESH, S. On the performance of particle swarm optimization (pso) algorithms in solving cheap problems. 2015.

AMERICAN CANCER SOCIETY. **What Is Breast Cancer**. 2017. [Online; Acesso em: 09 de Julho de 2018]. Disponível em: <https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html>.

AMERICAN CANCER SOCIETY. **What's New in Breast Cancer Research?** 2018. [Online; Acesso em: 09 de Julho de 2018]. Disponível em: <https://www.cancer.org/cancer/breast-cancer/about/whats-new-in-breast-cancer-research.html>.

BEUMER, I. et al. Equivalence of mammaprint array types in clinical trials and diagnostics. **Breast cancer research and treatment**, Springer, v. 156, n. 2, p. 279–287, 2016.

BISHOP, C. M. **Pattern recognition and machine learning**. [S.l.]: springer, 2006.

BRAGA, A. d. P. **Redes neurais artificiais: teoria e aplicações**. [S.l.]: Livros Técnicos e Científicos, 2000.

CARLSON, J. J.; ROTH, J. A. The impact of the oncotype dx breast cancer assay in clinical practice: a systematic review and meta-analysis. **Breast cancer research and treatment**, Springer, v. 141, n. 1, p. 13–22, 2013.

CARVALHO, P. **Classificadores para Pacientes com Câncer de Mama de Acordo com a Sensibilidade Quimioterápica Neoadjuvante**. Dissertação (Mestrado) — Centro Federal de Educação Tecnológica de Minas Gerais, 2017.

CHAMPE, P. C. et al. **Bioquímica**. [S.l.: s.n.], 2008.

CRAIG, N. L. et al. **Molecular biology: principles of genome function**. [S.l.]: Oxford University Press, 2014.

CUI, X.; CHURCHILL, G. A. Statistical tests for differential expression in cdna microarray experiments. **Genome biology**, BioMed Central, v. 4, n. 4, p. 210, 2003.

FAWCETT, T. Roc graphs: Notes and practical considerations for researchers. **Machine learning**, v. 31, n. 1, p. 1–38, 2004.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **The elements of statistical learning**. [S.l.]: Springer series in statistics New York, 2001. v. 1.

GARDNER, M. W.; DORLING, S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. **Atmospheric environment**, Elsevier, v. 32, n. 14-15, p. 2627–2636, 1998.

HATZIS, C. et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. **Jama**, American Medical Association, v. 305, n. 18, p. 1873–1881, 2011.

HESS, K. R. et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. **Journal of clinical oncology**, American Society of Clinical Oncology, v. 24, n. 26, p. 4236–4244, 2006.

HORTA, E. G. **Previsores para a eficiência da quimioterapia neoadjuvante no câncer de mama**. Dissertação (Mestrado) — Universidade Federal de Minas Gerais (UFMG), 2008.

HUANG, G.-B. et al. Incremental extreme learning machine with fully complex hidden nodes. **Neurocomputing**, Elsevier, v. 71, n. 4-6, p. 576–583, 2008.

INC, T. F. S. **NetAffx Query**. 2013. [Online; Acesso em: 10 de Setembro de 2018]. Disponível em: <https://www.affymetrix.com/analysis/netaffx/xmlquery.affx?netaffx=netaffx4_annot>.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. **Ijcai**. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

LI, C. Automating dchip: toward reproducible sharing of microarray data analysis. **BMC bioinformatics**, BioMed Central, v. 9, n. 1, p. 231, 2008.

LOCKHART, D. J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. **Nature biotechnology**, Nature Publishing Group, v. 14, n. 13, p. 1675, 1996.

MEDLINE PLUS. **Breast cancer staging**. 2017. [Online; Acesso em: 14 de Julho de 2018]. Disponível em: <https://medlineplus.gov/ency/patientinstructions/000911.htm>.

MEDLINE PLUS. **Breast cancer**. 2018. [Online; Acesso em: 14 de Julho de 2018]. Disponível em: <https://medlineplus.gov/ency/article/000913.htm>.

NATOWICZ, R. et al. Prediction of the outcome of preoperative chemotherapy in breast cancer using dna probes that provide information on both complete and incomplete responses. **BMC bioinformatics**, BioMed Central, v. 9, n. 1, p. 149, 2008.

NCBI. **Gene Expression Omnibus**. 2018. [Online; Acesso em: 04 de Fevereiro de 2018]. Disponível em: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20194>.

NELSON, D. L.; COX, M. M. **Princípios de Bioquímica de Lehninger-6**. [S.l.]: Artmed Editora, 2014.

NTU. **MATLAB Codes of ELM Algorithm**. 2013. [Online; Acesso em: 10 de Fevereiro de 2019]. Disponível em: <http://www.ntu.edu.sg/home/egbhuang/elm_random_hidden_nodes.html>.

POPOVICI, V. et al. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. **Breast Cancer Research**, BioMed Central, v. 12, n. 1, p. R5, 2010.

SCHENA, M. et al. Quantitative monitoring of gene expression patterns with a complementary dna microarray. **Science**, American Association for the Advancement of Science, v. 270, n. 5235, p. 467–470, 1995.

SEMOLINI, R. et al. Support vector machines, inferência transdutiva e o problema de classificação. [sn], 2002.

SIMON, R. M. et al. **Design and analysis of DNA microarray investigations**. [S.l.]: Sprin. ger Science & Business Media, 2003.

STEKEL, D. **Microarray bioinformatics**. [S.l.]: Cambridge University Press, 2003.

TEIXEIRA, R. de A. Treinamento de redes neurais artificiais através de otimização multi-objetivo: Uma nova abordagem para o equilíbrio entre a polarização e a variância. UFMG, 2001.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.

TONG, P. et al. Siber: systematic identification of bimodally expressed genes using rnaseq data. **Bioinformatics**, Oxford University Press, v. 29, n. 5, p. 605–613, 2013.

WANG, J. et al. The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. **Cancer informatics**, SAGE Publications Sage UK: London, England, v. 7, p. CIN–S2846, 2009.

Apêndices

APÊNDICE A – Texto Submetido e aceito para o CNMAC

Segue texto submetido e aceito no XXXIX Congresso Nacional de Matemática Aplicada e Computacional (CNMAC).

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Caracterização de Pacientes com Câncer de Mama em PCR e NoPCR para Tratamento Neoadjuvante por Meio da Análise de Bimodalidade

Gustavo Henrique Passini Santos [1]
Centro Federal de Educação Tecnológica de Minas Gerais
Thiago de Souza Rodrigues [2]
Centro Federal de Educação Tecnológica de Minas Gerais

1 Introdução e Proposta

O câncer de mama é uma neoplasia maligna que se inicia nos tecidos da mama, tendo o crescimento descontrolado das células do seio como gatilho. Tal comportamento observado nessas células geralmente forma um tumor, visível por exame de radiografia ou de toque, onde é possível identificar um nódulo [2]. O principal tratamento para o câncer de mama é a quimioterapia, onde ocorre a administração de medicamentos com o intuito de diminuir ou eliminar o tumor, ocorrendo por via intravenosa ou oral. A quimioterapia realizada antes da cirurgia de retirada do tumor é denominada neoadjuvante, tendo o ideal de curar o paciente sem a necessidade de enfrentar uma intervenção cirúrgica [1]. O objetivo do projeto apresentado é prever os pacientes que terão o câncer completamente erradicado, Patologic Complete Response ou PCR, e quais pacientes não serão curados, Residual Disease ou RD, com a quimioterapia neoadjuvante.

A vertente principal da metodologia de seleção de sondas aplicada no projeto é a análise da bimodalidade na expressão genética, caracterizada via microarray [5], de cada uma. Padrões de expressão bimodal podem resultar naturalmente da expressão diferencial, com os dois modos centrados na expressão média de um gene em dois subgrupos distintos de amostras. No contexto do câncer, os padrões de expressão bimodal podem resultar de lesões genômicas que ocorrem em alguns pacientes, mas não em outros [7].

A metodologia proposta visa selecionar sondas significativas por meio de ordenação e comparação entre as 3000 primeiras sondas com maior e menor valor de bimodalidade nos modelos propostos em [6] e [7] para pacientes das duas classes, PCR e NoPCR. Após selecionar as sondas, os pacientes serão classificados por intermédio de classificadores utilizados previamente em literatura, atestando dessa forma a qualidade da metodologia por meio dos resultados utilizando o gráfico e a curva ROC.

[1] passini.gustavo@gmail.com
[2] thiagothiagodcc@gmail.com

2 Conclusão

Como resultado preliminar foram selecionadas 4 sondas por método de bimodalidade utilizado. Foi verificado que todas sondas selecionadas relacionam-se a genes que desempenham funções protéicas relacionadas a proliferação do câncer no corpo humano [3]. Utilizando tais sondas para realizar a classificação dos pacientes por meio do classificador DLDA [4] foi obtida uma sensibilidade de 0,72 para Tong [6] e 0,52 para Wang [7], e uma especificidade de 0,7 para Tong [6] e 0,73 para Wang [7].

O grau de dificuldade em se executar a metodologia e seu custo são baixos. Portanto podemos inferir que a metodologia proposta apresenta um caminho interessante a se seguir na área de pesquisas médicas, pois quanto menos sondas forem necessárias para avaliação menor será o custo em processos laboratoriais, como a expressão genética via Microarray.

Espera-se que por meio de experimentos realizados em uma base mais recente e utilizando classificadores mais robustos, seja possível verificar um percentual de sensibilidade e especificidade maior, garantindo maior confiabilidade na caracterização de pacientes PCR e NoPCR utilizando a metodologia proposta.

Referências

[1] American Cancer Society. Chemoterapy for Breast Cancer. 2017. Disponível em: <https://www.cancer.org/cancer/breast-cancer/treatment/chemotherapy-for-breast-cancer.html>. Acesso em: 08 de Julho de 2018.

[2] Medline Plus. Breast cancer. 2018. Disponível em <https://medlineplus.gov/ency/article/000913.htm>. Acesso em 14 de Julho de 2018.

[3] Thermo Fisher Scientific Inc. NetAffx Query. 2017. Disponível em: <https://www.affymetrix.com/analysis/netaffx/xmlquery.affx?netaffx=netaffx4_annot>. Acesso em 10 Setembro de 2018.

[4] R. M. Simon, E. L. Korn, L. M. McShane, M. D. Radmacher,G. W. Wright, and Y. Zhao. *Design and analysis of DNA microarray investigations*. Springer Science & Business Media, 2003

[5] D. Stekel. *Microarray bioinformatics*. Cambridge University Press, 2003

[6] P. Tong, Y. Chen, X. Su and K. R. Coombes. SIBER: systematic identification of bimodally expressed genes using RNAseq data, *Bioinformatics*, 29:605–613, 2013

[7] J. Wang, S. Wen, W. F. Symmans, L. Pusztai and K. R. Coombes The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data, *Cancer informatics*, 7:CIN–S2846, 2009