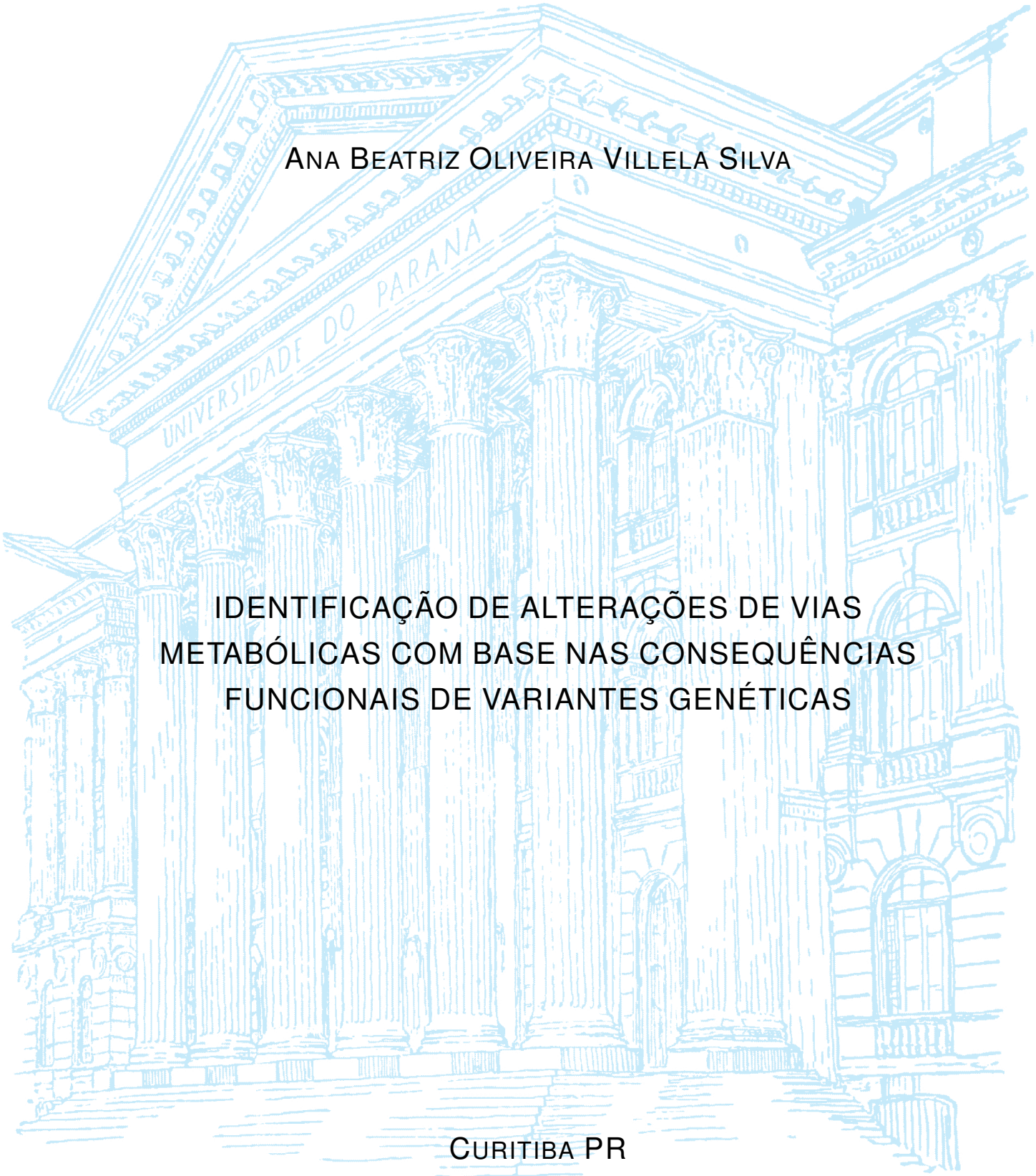
UNIVERSIDADE FEDERAL DO PARANÁ

ANA BEATRIZ OLIVEIRA VILLELA SILVA



IDENTIFICAÇÃO DE ALTERAÇÕES DE VIAS METABÓLICAS COM BASE NAS CONSEQUÊNCIAS FUNCIONAIS DE VARIANTES GENÉTICAS

CURITIBA PR

2018

ANA BEATRIZ OLIVEIRA VILLELA SILVA

IDENTIFICAÇÃO DE ALTERAÇÕES DE VIAS METABÓLICAS COM BASE NAS CONSEQUÊNCIAS FUNCIONAIS DE VARIANTES GENÉTICAS

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Informática Biomédica, setor de Ciências Exatas, da Universidade Federal do Paraná.

Orientador: Eduardo Jaques Spinosa.

Coorientadora: Angélica Beate Winter Boldt.

CURITIBA PR

2018

Agradecimentos

Ao meu orientador Eduardo Spinosa, que sempre foi um exemplo de professor e pessoa e me apoiou em todas as minhas loucuras nesses cinco anos de graduação.

A minha co-orientadora Angélica, por ser a pessoa que tem o nome que mais combina com a personalidade de alguém que eu conheci. Por todos os conselhos, por todo o carinho e por todo apoio nesse trajeto de aprender genética e bioinformática.

A minha mãe e meus familiares, por me providenciar todo o suporte emocional e financeiro necessários nessa jornada de ser adulta.

A todos os meus amigos e colegas que me aturam na UFPR desde 2014, o que convenhamos que não foi uma tarefa muito fácil. Em especial, agradeço ao Gabriel por ter sido uma pessoa mais didática que o *Google* e ter tirado todas as dúvidas possíveis que uma pessoa poderia ter em relação à *deep learning*. Também agradeço a Julia por escutar todos os choros e reclamações possíveis, pela revisão do texto e por estar sempre presente.

Agradeço ao meu namorado Yago, pois esse trabalho não existiria sem ele e seu apoio e cuidados incondicionais nos últimos seis anos.

E a todos os professores e pesquisadores que contribuíram de forma direta ou indireta para a minha jornada, espero poder retribuir de alguma forma no futuro.

Resumo

Variantes genéticas podem contribuir para o desenvolvimento de diferentes fenótipos, por afetarem diferentes vias metabólicas no corpo humano. As variantes podem ter efeito estrutural, alterando a conformação da proteína ou do RNA funcional que resulta da sua transcrição, ou afetando mecanismos de regulação da expressão gênica. Neste trabalho, tem-se o objetivo de avaliar variantes genéticas organizadas em um arquivo do tipo *Variant Call Format* (VCF) e informar em quais vias metabólicas essas possuem um impacto. Para a análise de variantes codificadoras de proteínas, a ferramenta ANNOVAR foi escolhida. Para variantes não codificantes, foi utilizada uma rede neural de aprendizado profundo convolucional, que a partir de uma sequência de DNA, pretende classificar uma variante como possuindo ou não possuindo um efeito eQTL (*Expression Quantitative Trait Loci*). As bibliotecas *Tensorflow* e *Keras* foram utilizadas para a implementação. Por fim, a busca de quais genes estão relacionados com quais vias metabólicas é feita por meio de requisições *web* pela plataforma KEGG. Entre os 53 tecidos humanos disponibilizados pelo portal GTEx, foram utilizados como dados de treinamento e validação da rede neural informações provenientes de 5 tecidos: fígado, hipocampo, sangue, tecido subcutâneo exposto e não exposto ao sol. As taxas de acerto do conjunto de validação das redes neurais obtidas foram entre $74,17\%$ e $90,48\%$, com uma área sob a curva variando entre 0.817 e 0.948. O F_1 *score* foi calculado, obtendo-se valores entre 0.7465 e 0.9040. Conclui-se que a abordagem utilizada tornou possível automatizar e facilitar o processo de identificação do impacto funcional de variantes genéticas na expressão gênica em seres humanos.
Palavras-chave: bioinformática,genética,aprendizado profundo.

Abstract

Genetic variants may contribute to the development of different phenotypes by affecting different metabolic pathways in the human body. They may have a structural effect, altering the conformation of the functional protein or RNA that results from its transcription, or may affect mechanisms of gene expression regulation. In this work, the main goal is to evaluate genetic variants from a Variant Call Format (VCF) file and to inform in which metabolic pathways these may have an impact. For the selection of protein coding variants, the ANNOVAR software was chosen. For non-coding variants, a deep convolutional neural network was used, in which a DNA sequence is an input, and the output is the classificaton whether the variants have or not an eQTL effect. The Tensorflow and Keras libraries were used to implement the network. Finally, the relationship of which genes are related to which metabolic pathways is made by web requests on the KEGG platform. Among the 53 human tissues available in the GTEx portal, information from 5 tissues were used to train and validate the nerwork: liver, hippocampus, blood, subcutaneous tissue exposed and not exposed to the sun. The accuracy of the validation datasets were between 74.17% and 90.48%, with an area under the curve ranging from 0.817 to 0.948. The F_1 score was calculated, obtaining values between 0.7465 and 0.9040. The conclusion is that the approach used showed that it is possible to automate and facilitate the process of identifying the functional impact of genetic variants in humans.

Keywords: bioinformatics, genetics, deep learning.

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

ANN	Artificial Neural Network
AUC	Area under curve
pb	Pares de base de nucleotídeos
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DINF	Departamento de Informática
DNA	Ácido desoxirribonucleico
eQTL	Expression quantitative trait loci
GPU	Graphics Processing Unit
GWAS	Genome-wide Association Studies
kb	Milhares de pares de base
KEGG	Kyoto Encyclopedia of Genes and Genomes
MLP	Multilayer Perceptron
ROC	Receiver operating characteristic
RNA	Ácido ribonucleico
ReLU	Recfitied Linear Unit
SNP	Single Nucleotide Polymorphism
UFPR	Universidade Federal do Paraná
VCF	Variant Call Format

Sumário

Capítulo 1

Introdução

A Biologia como campo de estudo vem sofrendo drásticas mudanças desde o fim do século XX. Devido ao grande avanço impulsionado pelo projeto Genoma Humano nas técnicas de sequenciamento de DNA (1000GENOMES, 2012), hoje existem diversos de estudos envolvendo a análise de variantes genéticas.

Eventualmente, estas variantes podem contribuir para o desenvolvimento de diferentes fenótipos, por afetarem diferentes vias metabólicas no corpo humano. As variantes podem ter efeito estrutural, alterando a conformação da proteína ou do RNA funcional que resulta da sua transcrição, ou afetando mecanismos de regulação da expressão gênica.

Dentre os diversos tipos de variantes genéticas, os polimorfismos de nucleotídeo único (*SNPs*), constituem 90% de todas as variações genômicas humanas (GRIFFITHS et al., 2005). Os SNPs, contudo, nem sempre apresentam impactos funcionais, mesmo quando causam substituições de aminoácido no produto final[1]. De fato, a maior parte dos SNPs são considerados como variantes neutras, por não causarem substituição de aminoácidos[2] ou ocorrerem em regiões sem função reguladora reconhecida (1000GENOMES, 2012).

Para SNPs não sinônimos, ferramentas como ANNOVAR, desenvolvida por Wang, Li e Hakonarson (2010), utilizam-se de algoritmos de predição de estrutura e função de proteínas como os desenvolvidos por Adzhubei et al. (2010) e Sim et al. (2012), para estimar o impacto funcional de um SNP. Esses algoritmos baseiam-se em características como o grau de conservação evolutiva, propriedades bioquímicas de aminoácidos, entre outras.

A gama de ferramentas disponível para predizer especificamente alterações de expressão gênica decorrentes de SNPs é ainda mais escassa, haja vista a maior complexidade dos mecanismos de regulação. Neste sentido, técnicas de redes neurais de aprendizado profundo têm sido apresentadas como potenciais soluções para problemas de bioinformática. Soluções como as desenvolvidas por Zhou e Troyanskaya (2015) e Angermueller et al. (2017) são capazes de extrair características específicas relacionadas ao controle da expressão gênica.

Neste trabalho pretende-se, de forma automatizada, avaliar variantes genéticas provenientes de um arquivo do tipo *Variant Call Format* (VCF) e informar em quais vias metabólicas estas podem possuir um impacto.

No capítulo 2, conceitos relacionados à forma de herança de variantes genéticas, assim como a estrutura do DNA são apresentados. Também é feita uma descrição de técnicas de aprendizado de máquina, com o foco em redes neurais profundas convolucionais. No capítulo 3, o problema de identificar atributos relacionados à regulação da expressão gênica em seres humanos é abordado, descrevendo três exemplos de métodos que foram utilizados na literatura.

[1] substituições não sinônimas

[2] substituições sinônimas ou silenciosas

Nos capítulos 4 e 5 serão discutidos detalhes sobre o desenvolvimento do projeto, juntamente a descrição dos resultados obtidos. No capítulo 6, discutem-se os resultados e perspectivas para trabalhos futuros.

1.1 Objetivos

1.1.1 Objetivo geral

Identificar SNPs com impacto funcional em vias metabólicas, por meio da avaliação da sua associação com alterações na expressão gênica e / ou na estrutura do produto gênico e identificar em qual via estas variantes podem interferir.

1.1.2 Objetivos específicos

1. Identificar variantes genéticas raras (frequência inferior a 1%) em arquivos de sequenciamento de DNA humano;

2. Dentre estes SNPs, identificar os que potencialmente alteram a estrutura do produto gênico associado original e indicar a dimensão do seu impacto, utilizando ferramentas publicamente disponíveis;

3. Implementar uma rede neural de aprendizado profundo capaz de predizer variantes que funcionem como eQTLs *(Expression quantitative trait loci)*.

4. Identificar a(s) via(s) metabólica(s) na(s) qual(is) as variantes identificadas com impacto positivo estão inseridas, utilizando a base de dados *KEGG PATHWAYS*.

Capítulo 2

Fundamentação Teórica

Nesse capítulo, são apresentados os conceitos relacionados à estrutura do DNA como molécula, bem como o impacto de variantes na estrutura do produto proteico e/ou expressão genética. Além disso, descrevem-se definições na área de aprendizado de máquina, especificamente em redes neurais de aprendizado profundo, que serão utilizados para desenvolver o classificador que prediz variantes reguladoras da expressão gênica.

2.1 Genética molecular

A estrutura dos seres vivos e seus processos fisiológicos associados são baseados, de modo geral, em proteínas. A informação genética para a síntese dessas proteínas pelas células está contida no DNA, o ácido desoxirribonucleico. O DNA é composto de uma dupla fita que contém bases nitrogenadas de diferentes tipos: adenina (A), citosina (C), guanina (G) e timina (T). Cada fita contém uma sequência complementar à fita oposta, em que A se pareia com T e C se pareia com G. Uma unidade de medida para se referir ao comprimento de uma molécula de DNA é o número de pares de bases (pb). Um conjunto único e completo de informação genética de um organismo é chamado de genoma (GRIFFITHS et al., 2005).

A definição do Dogma Central da Biologia molecular foi publicada nos anos 1950 por Watson e Crick. No trabalho, uma descrição do caminho da informação genética ao longo da vida de um organismo foi proposta, relacionando DNA, ácido ribonucleico (RNA) e proteínas. O DNA tem a capacidade de se replicar, assim como transcrever uma molécula de RNA mensageiro (mRNA). O mRNA, por sua vez é traduzido em proteínas, que são elementos fundamentais para a manutenção da vida. O DNA também é capaz de produzir moléculas de RNA que não são traduzidas, sendo elas o produto final da transcrição. O processo pode ser visto na figura 2.1.

Um gene pode ser entendido como uma sequência de DNA composta de promotor, exons e introns em um local (em latim *locus*) específico do DNA, que codifica uma característica genética capaz de ser herdada. Além disso, genes podem apresentar diversas formas, conhecidas como alelos (PIERCE, 2012). Em organismos eucariotos como o ser humano, genes possuem uma estrutura complexa, como a apresentada no exemplo da figura 2.2.

O modo como o DNA codifica proteínas específicas é por meio do código genético: a sequência de aminoácidos a ser traduzida é especificada por uma sequência de unidades codificantes elementares em um gene. Essas unidades codificantes são trinucleotídeos adjacentes presentes no mRNA chamados *codon* (SIMMONS; SNUSTAD et al., 2006).

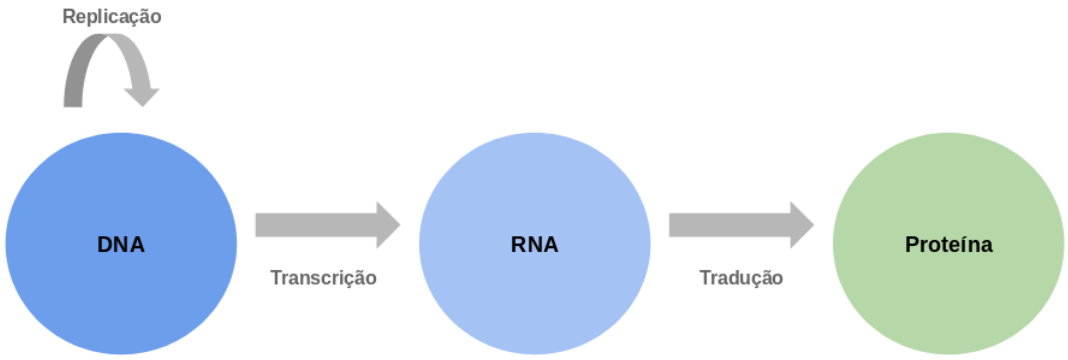


Figura 2.1: O dogma central da Biologia Molecular. O DNA é capaz de se replicar ou de ser transcrito em RNA, este por sua vez pode ser traduzido. Fonte: autoria própria

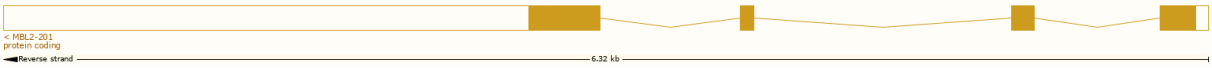


Figura 2.2: Representação da estrutura de um gene eucarioto. As linhas representam regiões intrônicas, enquanto os retângulos preenchidos representam exons. Os retângulos vazios são as regiões não traduzidas (UTR). A flecha indica o sentido da transcrição. Disponível em: <https://goo.gl/75PKpV>. Acesso em 27 nov. 2018

2.1.1 Mutações

O DNA pode sofrer mutações, em diferentes regiões com potenciais efeitos diferentes. Mutações podem ser somáticas ou germinativas, sendo no último caso, capazes de serem transmitidas para a prole de um indivíduo. Mutações somáticas podem apenas ser transmitidas para gerações celulares subsequentes.

Mutações também podem ser subdivididas entre mutações induzidas ou espontâneas. Mutações induzidas são aquelas que são geradas por um agente externo ao organismo, como fatores físicos, químicos e biológicos. Mutações espontâneas acontecem por fatores internos do organismo, e ambos os tipos permanecem na célula quando existe falha nos mecanismos de reparação do DNA (SIMMONS; SNUSTAD et al., 2006).

Um exemplo comum de mutação espontânea é o acontecimento de tautomerias nas bases nitrogenadas. Tautomeria é um caso particular de isomeria funcional onde as bases nitrogenadas ficam em equilíbrio químico dinâmico (SIMMONS; SNUSTAD et al., 2006). As tautomerias fazem com que o pareamento entre as bases aconteça de forma errada, como uma Citosina se pareando com uma Adenina e uma Timina se pareando com uma Guanina.

A quantidade de informação genética que sofre mutação também é variável. Estas variações podem ser mutações gênicas, afetando nucleotídeos específicos ou pequenas porções de DNA, cromossômicas quando o número de cromossomos da célula está desbalanceado (trissomias, por exemplo), quando o número de cópias do conjunto total de cromossomos é alterado (triploidias, tetraploidias), ou quando grandes porções do DNA são alteradas (deleções, duplicações, inversões, translocações).

A natureza química das bases nitrogenadas é um dos fatores que influenciam as mutações. Bases nitrogenadas podem ser purinas (A e G) ou pirimidinas (C e T) e mutações entre bases de mesmo tipo, conhecidas como transições, são mais comum do que entre os dois tipos, as transversões. Essa relação pode ser visualizada na figura 2.3.

As mutações gênicas são denominadas mutações de ponto quando ocorrem em um único par de bases do DNA ou em um pequeno número de bases. Quando estas ocorrem em

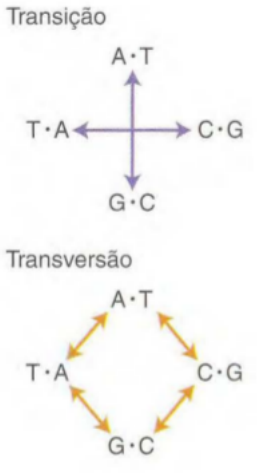


Figura 2.3: Transições e transversões (GRIFFITHS et al., 2005). Transições ocorrem entre bases nitrogenadas com as mesmas propriedades químicas, enquanto transversões ocorrem na situação oposta.

apenas um nucleotídeo, são denominadas *(Single Nucleotide Polymorphisms)* (SNPs). Outros tipos de mutações comuns são as deleções e inserções e as repetições de nucleotídeos em *tandem*. Alguns exemplos de impacto de SNPs estão na tabela 2.1.

Como porção de genes codificantes do DNA é aproximadamente 1% do total, a maior parte das variantes ocorre em regiões reguladoras que não são codificantes, como nos introns, regiões promotoras ou até mesmo entre dois genes diferentes.

A maior parte dos transcritos não-codificantes existentes possui um papel na regulação da expressão gênica, de modo que uma mutação nessas regiões não impacta diretamente no gene em si, apenas na quantidade do produto gênico que será transcrito ou traduzido.

2.1.2 Regulação da expressão gênica

Considerando o produto final do dogma central da biologia molecular como uma proteína, a regulação da expressão gênica pode ocorrer durante as diversas etapas do fluxo gênico. Ocorre regulação em nível transcricional, no processamento de introns, durante a exportação do mRNA do núcleo ao citoplasma, na tradução por miRNAs nas regiões UTR a 3', ou até mesmo no controle do nível de degradação do mRNA ou no controle da atividade proteica (ALBERTS et al., 2017).

Os organismos multicelulares contêm muitos tipos especializados de células organizados em tecidos e órgãos; porém, o material genético é idêntico em todas as células somáticas de um mesmo ser vivo. Isso ocorre porque determinado gene pode estar sendo expresso em células do sangue, por exemplo, mas não em células nervosas. A regulação da expressão gênica é que cria essas diferenças da quantidade e tipos de produto gênico em cada tecido. Existem porém diversos genes que garantem o funcionamento basal de toda as células, e estes são chamados de genes de *housekeeping*.

Um dos mecanismos primários de regulação da expressão gênica é a variação na densidade e nível de compactação da cromatina, que é o conjunto de cromossomos de uma célula. Seções mais compactas são chamadas de heterocromatina e seções mais frouxas, eucromatina. Sabe-se que a transcrição ocorre em níveis maiores na eucromatina, pela facilidade

Tipo de alteração	Impacto
Substituição sinônima	Não altera o produto proteico final, mas pode ser reconhecida por tRNAs raros, dificultando a tradução
Substituição não sinônima	Altera um aminoácido do produto proteico final, mas este pode ter um efeito neutro caso a mudança seja conservativa
Perda de codon de início	Causa o encurtamento da porção N-terminal da proteína
Perda de codon de parada	O produto proteico final é estendido
Perda de um ou mais codons	O produto proteico é apenas encurtado, com a perda de alguns aminoácidos
Adição de um ou mais codons	O produto proteico é apenas alongado, mas a sequência de aminoácidos em si não muda
Adição de um codon de parada	O produto proteico é encurtado
Alteração de fase ou quadro de leitura	Altera a sequência de aminoácidos a ser traduzida
Alteração na região UTR a 5' ou a 3'	Pode afetar o reconhecimento por proteínas reguladoras e RNAs não codificantes
Alteração nas regiões de início ou fim de splicing	Pode alterar quais introns vão passar pelo processo de *splicing* alternativo
Alteração na região promotora	Pode alterar a ligação entre moléculas ativadoras e alterar a expressão do gene em questão
Intrônica	Variável

Tabela 2.1: Impactos possíveis de variantes em exons (Adaptado de: <http://www.ensembl.org/info/genome/variation/prediction/predicted_data.html>. Acesso em 27 nov. 2018)

de acesso na sequência alvo de fatores de transcrição e outras proteínas auxiliares do processo (PIERCE, 2012).

Segundo Simmons, Snustad et al. (2006), dos aproximadamente 3 bilhões de pares de bases nitrogenadas no genoma típico de mamífero, cerca de 40% são pares de bases G:C, e cerca de 2% a 7% desses são modificados pelo acréscimo de um grupo metil (CH_3) à citosina, quando ocorrem em dinucleotídeos 5'CpG3'[1] (figura 2.4). O DNA metilado normalmente está associado à repressão da transcrição. Segmentos ricos em dinucleotídeos 5'CpG3', geralmente com 1 a 2kb de extensão são denominados ilhas CpG. No genoma humano, há aproximadamente 30 mil ilhas, a maioria delas não é metilada e ocorre próxima a sítios de regulação da transcrição de genes de economia doméstica ou genes de expressão de tecidos específicos.

Para a regulação de características multifatoriais em seres humanos (como a altura ou a cor da pele), *Expression quantitative trait loci* (eQTLs) são *loci* específicos em que variantes podem descrever total ou parcialmente a variação da expressão gênica dos RNA mensageiros. eQTLs podem ter efeito *cis* ou *trans*, dependendo da natureza da interação com o gene alvo e da distância do gene que é regulado (NICA; DERMITZAKIS, 2013). No efeito *cis*, as variantes estão associadas à expressão do cromossomo no qual se encontram localizadas, enquanto em efeito *trans* existe a associação da expressão de outro cromossomo.

[1] Os nucleotídeos estão em sequência, na mesma fita. O "p" entre as duas bases significa o fosfato que existe entre as mesmas

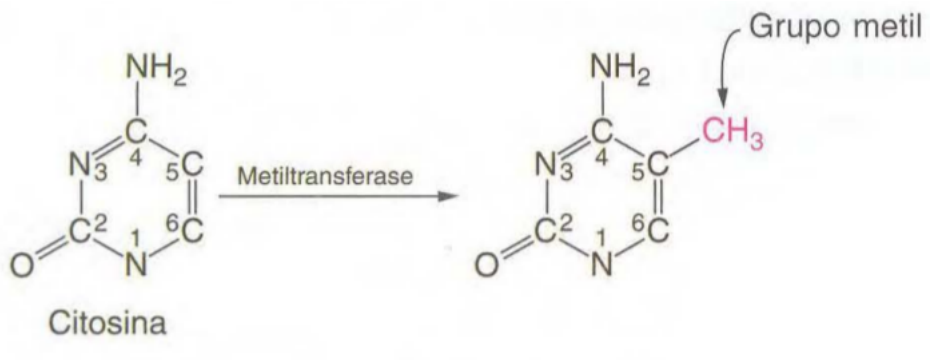


Figura 2.4: Metilação de um nucleotídeo de Citosina (GRIFFITHS et al., 2005). A metilação no DNA tende à aumentar o seu nível de compactação

2.2 Aprendizado de Máquina

Aprendizado de máquina é um ramo de estudo da inteligência artificial que estuda algoritmos capazes de abstrair conhecimento de forma autônoma. (MICHALSKI; CARBONELL; MITCHELL, 2013). Uma das formas de dividir os diferentes tipos de algoritmos é entre aprendizado supervisionado, não supervisionado e aprendizado por reforço. A figura 2.5 mostra um diagrama com as subdivisões.

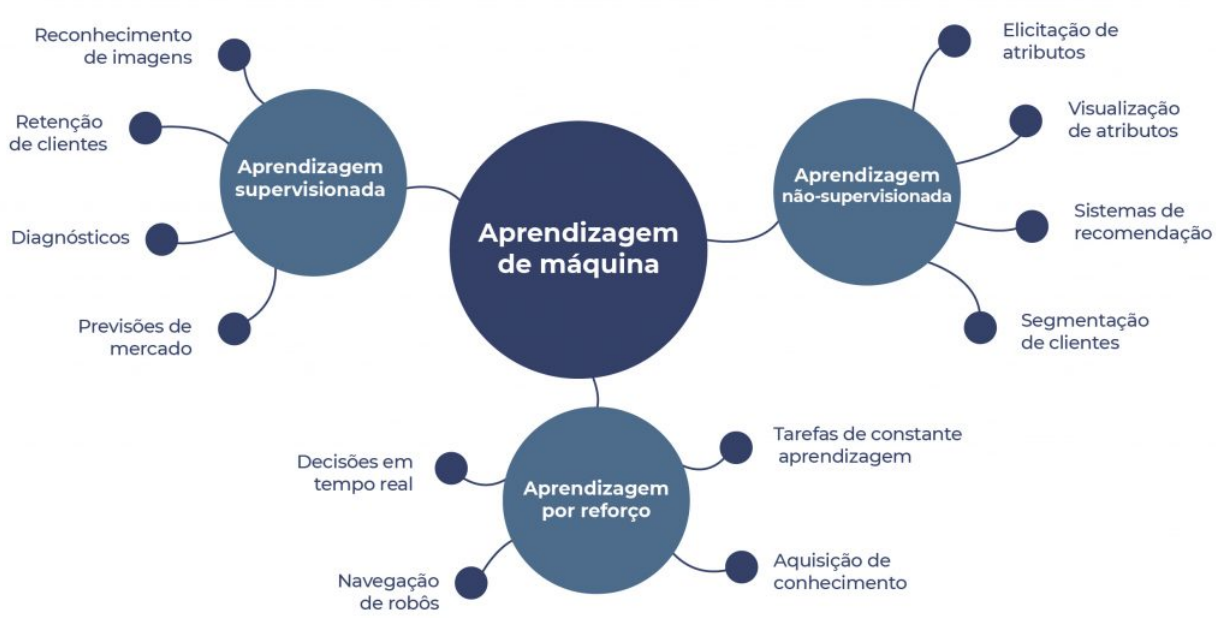


Figura 2.5: Principais áreas de aprendizado de máquina e exemplos de aplicação (ISI-TICS, 2018)

Nas soluções em que os dados de entrada são divididos entre *treinamento* e *teste*, em que os dados de treinamento possuem os valores no espaço vetorial juntamente com sua classificação prévia fazem parte dos métodos de *aprendizado supervisionado*. Esse conjunto pode se subdividir ainda entre problemas de *classificação* ou *regressão*. Problemas de classificação têm o objetivo de identificar classes de um conjunto de objetos desconhecidos. Problemas de

regressão pretendem descrever a relação estatística entre uma ou mais características. Um exemplo de problema de classificação é o reconhecimento de dígitos manuscritos, enquanto para um problema de regressão seria a predição da produção de um certo componente químico no futuro, considerando reagentes, temperatura e pressão (BISHOP, 2006).

Já a categoria de aprendizado não supervisionado pode ser definida como a classe de problemas em que o objetivo é baseado na descoberta de similaridades entre grupos de dados (*clustering*) ou determinar qual a distribuição de um conjunto de valores no espaço de entrada (estimativa de densidade).

Por fim, a técnica de aprendizado por reforço (*reinforcement learning*) foca no problema de encontrar ações adequadas dada uma determinada situação com o objetivo de maximizar uma recompensa. Em contraste com os algoritmos de aprendizado supervisionado, o programa não recebe exemplos de entradas com saídas ótimas, mas descobre a relação entrada-saída em um processo de tentativa e erro. Uma alegoria do mundo real seria o processo de adestramento de um cachorro.

Para o problema de identificação de eQTLs baseados nas variantes genômicas, há a possibilidade de se buscar informações em bases de dados biológicas para treinamento para algoritmos de classificação.

2.2.1 Redes Neurais Artificiais

O cérebro humano pode ser visto como um computador de alta complexidade e que atua de forma massivamente paralela. Com essa inspiração, foram modeladas as redes neurais artificiais: uma máquina inspirada no cérebro humano com o objetivo de executar uma tarefa ou função de interesse. Uma rede é geralmente implementada usando componentes eletrônicos ou é simulada em um *software* (HAYKIN et al., 2009). A rede simulada lembra o comportamento do cérebro humano em dois aspectos:

1. O conhecimento é adquirido pela rede dentro de seu ambiente por meio de um processo de aprendizado.

2. Conexões entre neurônios, conhecidas como pesos sinápticos são utilizados para armazenar o conhecimento adquirido.

Segundo Haykin et al. (2009), entre as vantagens do uso de sistemas baseados em redes neurais está sua estrutura massivamente paralela e distribuída por natureza, além da capacidade de aprender a inferir informações sobre os dados de entrada. A combinação dessas duas propriedades torna as redes neurais artificiais capazes de encontrar soluções aproximadas para problemas complexos em larga-escala que são intratáveis por outros métodos de aprendizagem de máquina.

O elemento estruturante de uma rede neural é um neurônio artificial. Ele é composto de uma entrada formada por um conjunto de pesos sinápticos e um viés (*bias*), passa por uma função de ativação e produz uma saída. O viés é uma variável que pode influenciar no aumento ou na diminuição da entrada na função de ativação dependendo se é positivo ou negativo. Um neurônio artificial pode ser descrito usando a imagem 2.6 ou a equação 2.1.

$$Y = \varphi\left(\sum(\text{weight} \times \text{input}) + \text{bias}\right) \tag{2.1}$$

A função de ativação, descrita por $\varphi(v)$, define a saída de um neurônio artificial. Existem vários tipos de função de ativação, sendo algumas entre as mais comuns: função limiar, função

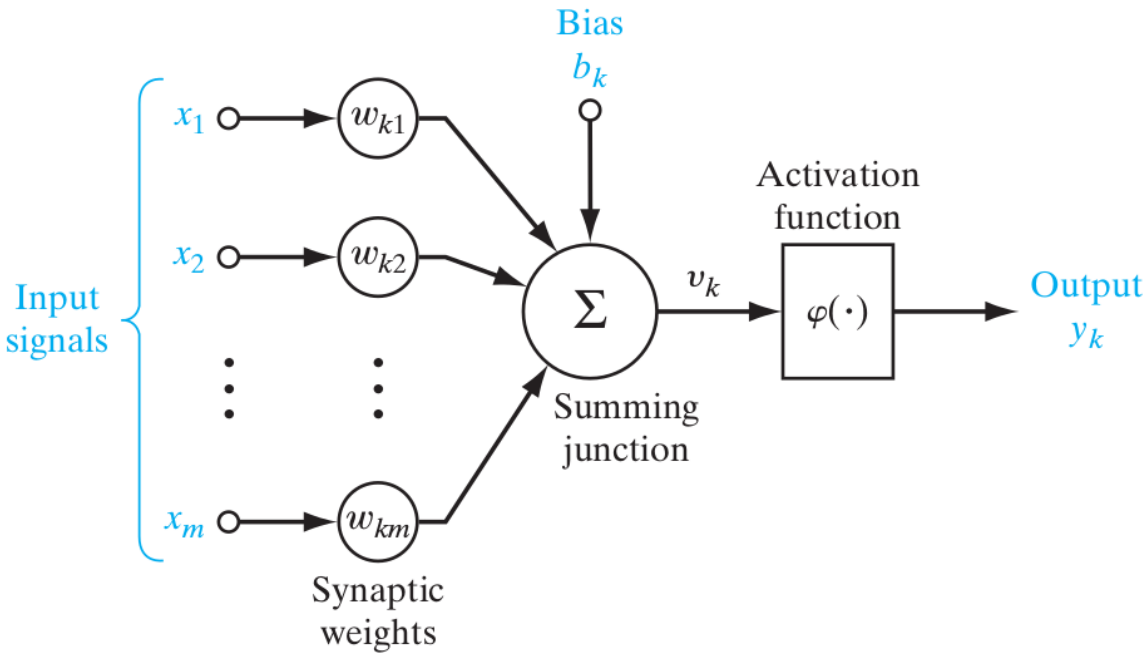


Figura 2.6: A estrutura de um neurônio artificial (HAYKIN et al., 2009). x_i representam as entradas, w_i representam os pesos, b é o viés e y é a saída esperada para o conjunto da iteração k

sigmoide e função *Rectified Linear Unit* (ReLU). A função limiar (*threshold*) está entre as mais simples utilizadas em redes neurais, descrita na equação 2.2, em que o valor de v_k é binarizado caso o valor de entrada ultrapasse esse limite. Um exemplo de função sigmoide está na equação 2.3, sendo essa uma das funções mais comumente utilizadas por possuir um equilíbrio entre a linearidade e não-linearidade da saída. Por fim, a função *ReLU*, descrita na equação 2.4 é a função identidade para valores positivos, enquanto valores negativos se transformam em zero e não ativam o neurônio.

$$\varphi(v) = \begin{cases} 1 \text{ se } v_k \geq 0 \\ 0 \text{ se } v_k < 0 \end{cases} \tag{2.2}$$

$$\varphi(v) = \frac{1}{1 + \text{ exp } (-av)} \tag{2.3}$$

$$\varphi(v) = \text{max}(0, v) \tag{2.4}$$

Quando todo o conjunto de treinamento foi utilizado como uma entrada na rede e produziu uma saída, têm-se uma época ou iteração. Esse processo pode ser subdividido em *batches*, para que conjuntos menores passem pelo neurônio artificial ao mesmo tempo, ao invés de todo o conjunto de dados.

O processo de atualização dos pesos sinápticos entre duas épocas é feito em um processo conhecido como *back propagation*. Para calcular o erro, a diferença entre a saída esperada e a obtida na rede é calculada, com uma função chamada *loss*. Diferentes métodos são usados para o cálculo da *loss*, sendo um dos mais utilizados o *mean-squared error* (MSE), que calcula o quadrado da diferença entre resultado observado e esperado.

No processo de propagação dos erros para as camadas anteriores da rede, os pesos são modificados de forma a tentar minimizar a *loss* por meio de uma função de otimização. Essa função calcula o gradiente, isto é, a derivada parcial da *loss* levando em consideração os pesos. Os pesos então são atualizados no sentido oposto do gradiente calculado. Esse ciclo se repete até que se encontre o valor mínimo da função *loss* e uma representação gráfica pode ser vista na figura 2.7.

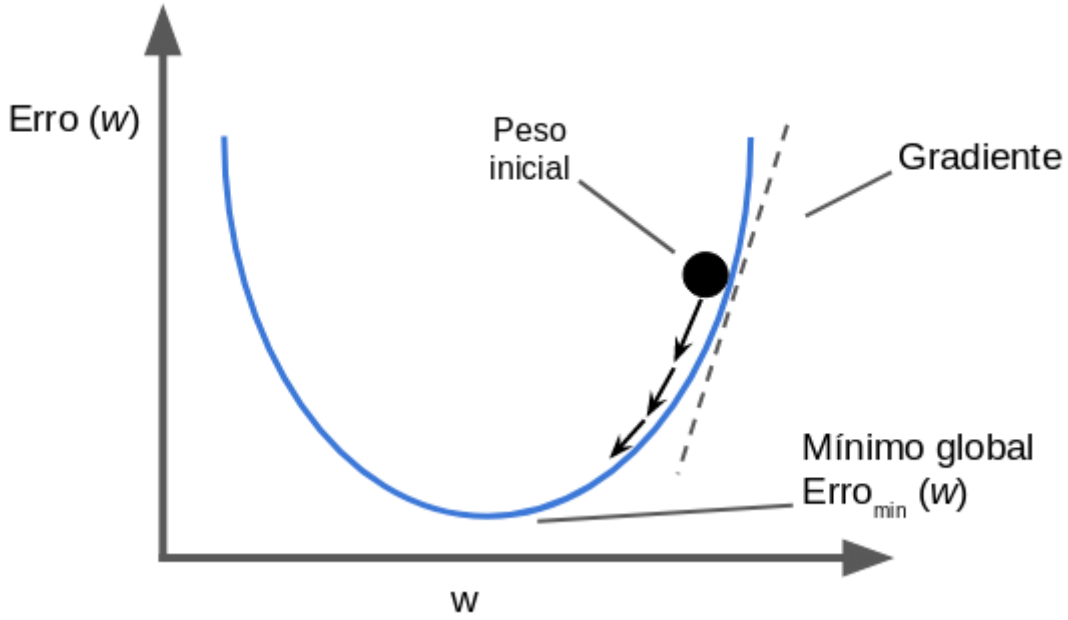


Figura 2.7: O cálculo do gradiente do erro de um conjunto de pesos w. O objetivo de se seguir o gradiente é chegar no que se chama de mínimo global da função *loss*. Fonte: autoria própria

Uma rede neural artificial pode ser organizada em conjuntos de neurônios agrupados em camadas. Frank Rosenblatt na década de 50 definiu um modelo conhecido como *Multi-Layer-Perceptron* (MLP), que é composto de três tipos de camadas: a camada de entrada, a camada de saída e uma ou mais camadas intermediárias. O uso de camadas intermediárias, que também são conhecidas como camadas escondidas, torna o MLP um modelo capaz de separar dados que não são linearmente separáveis (BARRETO, 2002). Uma MLP que possui mais de uma camada escondida pode ser chamada de rede de aprendizado profunda, ou *deep learning network*.

2.2.2 *Deep Learning*

O conceito de *deep learning* vem da premissa de se adicionar mais camadas escondidas e mais unidades por camada do que uma rede neural comum de forma hierárquica (GOODFELLOW et al., 2016). Fazendo isso, uma rede neural de aprendizado profundo é capaz de representar informações de alta complexidade.

Visto o aumento da quantidade de camadas e neurônios, é necessário se atentar às questões da arquitetura de uma rede de aprendizado profundo. Essas redes possuem uma estrutura em cadeia, em que a entrada de uma camada é resultado da saída da camada anterior. Nesses casos, é necessário avaliar duas propriedades: a largura (número de unidades por camada) e a profundidade da rede (número de camadas escondidas) (GOODFELLOW et al., 2016).

Dentro da área de *deep learning*, existem redes com características específicas; entre elas, a CNN (*Convolutional Neural Network*) é um tipo especializado para o processamento de dados que possuem um conjunto de características que podem ser representadas em forma de matriz. Exemplos de dados de entrada em que uma CNN pode ser aplicada são imagens e sequências de DNA ou RNA.

Convolutional Neural Networks

O nome “Rede neural Convolucional” descreve uma rede neural que implementa a operação matemática chamada convolução. Uma convolução é um tipo especial de operador linear que atribui o valor a uma célula de acordo com os valores da sua vizinhança matricial.

Cada camada de convolução é composta por um conjunto de filtros (*kernels*), que multiplicam por meio de uma janela deslizante ao longo da matriz de entrada. Os *kernels* são matrizes pequenas compostas por valores reais. Após a multiplicação do *kernel* pela matriz de entrada, o resultado obtido é um mapa de características. Um exemplo de convolução aplicada a uma matriz de duas dimensões pode ser vista em 2.8.

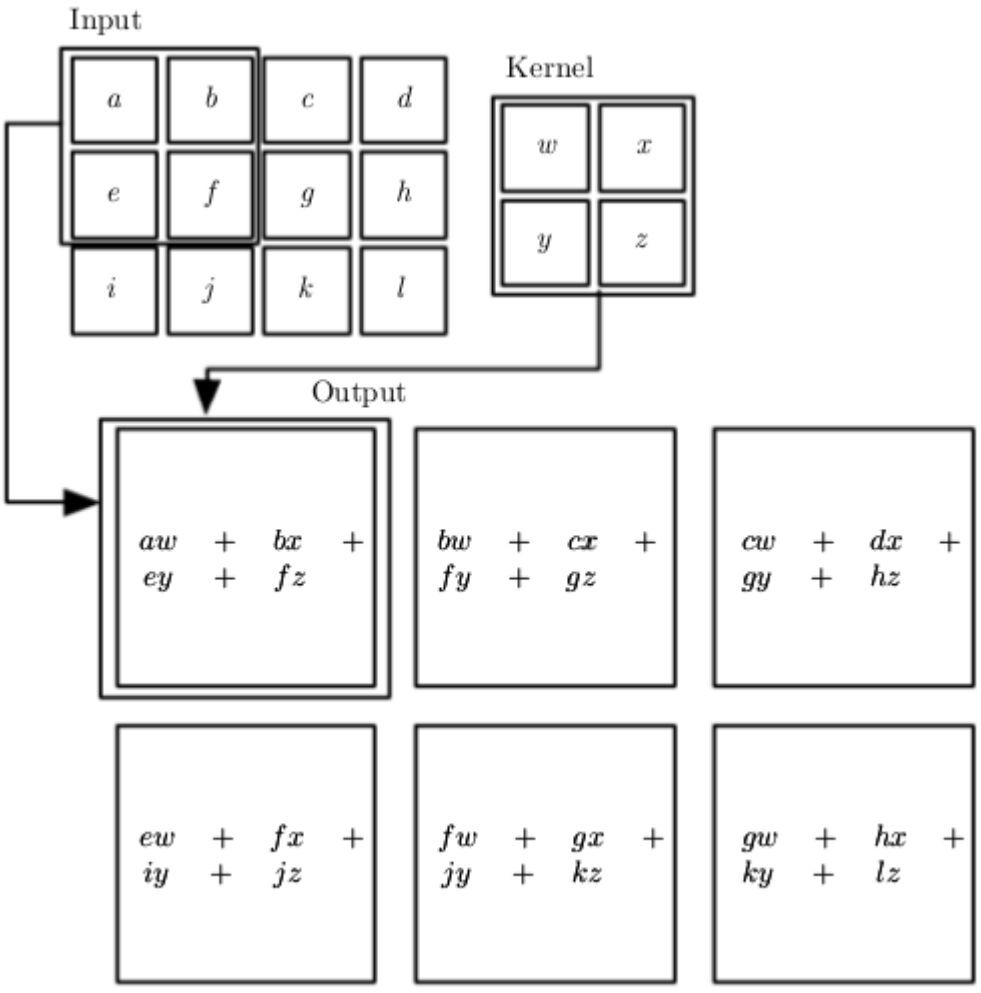


Figura 2.8: Uma convolução em uma matriz de 2 dimensões (GOODFELLOW et al., 2016). O processo de convolução é uma multiplicação de matrizes em uma janela deslizante que produz como saída um mapa de características.

Entre as vantagens de se utilizar camadas convolucionais está a capacidade de fazer interações esparsas. Uma rede neural tradicional utiliza, na multiplicação de matrizes, parâmetros que descrevem a interação de cada atributo de entrada com um atributo de saída. Redes convolucionais, por outro lado, tipicamente tem conectividade esparsa, graças ao *kernel* que tem um tamanho menor do que os dados de entrada (GOODFELLOW et al., 2016). Na prática, isso garante que não há a necessidade de haver um peso para cada atributo do conjunto de entrada.

Após a camada de convolução, uma função de ativação é utilizada como visto na seção 2.2.1. A escolha mais comum é a função ReLU da equação 2.4, baseado no estado da arte.

A camada seguinte na rede neural é o *pooling*, que substitui a saída da camada anterior em um certo local por um resumo estatístico das saídas próximas. Por exemplo, a operação de *max pooling* retorna a saída máxima dentro de uma janela retangular de vizinhança, e pode ser vista na figura 2.9. De modo geral, a função de *pooling* torna a representação dos dados mais invariante às pequenas mudanças da entrada (GOODFELLOW et al., 2016).

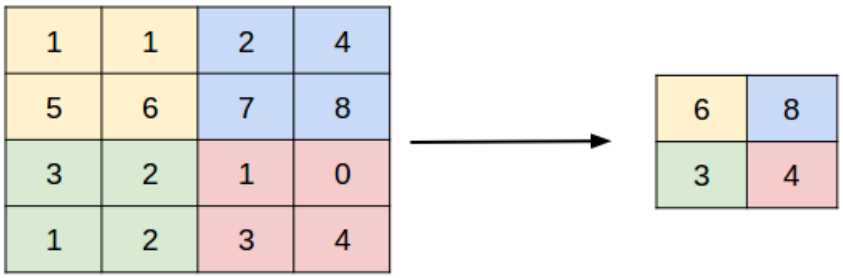


Figura 2.9: Operação de *max pooling* sendo aplicada em uma matriz de duas dimensões. Fonte: autoria própria

2.2.3 Hiperparâmetros e regularização da rede

O desafio central dentro da área de aprendizado de máquina é obter bons resultados com dados fora do conjunto de treinamento. A capacidade de bom desempenho em dados desconhecidos é chamada generalização. A não generalização da rede significa que os dados estão sob o efeito de *overfitting* ou *underfitting*.

O *underfitting* ocorre quando o modelo da rede não é capaz de obter uma taxa de erro baixa o suficiente nos dados de treinamento. Isso pode ser corrigido aumentando a profundidade ou a largura da rede. Por outro lado, o *overfitting* acontece quando a taxa de erro é baixa no conjunto de treinamento comparada com a do conjunto de teste. Um dos motivos para tal é quando a rede é muito profunda ou larga, e o modelo torna-se capaz de “memorizar” cada entrada do conjunto de treinamento, perdendo a capacidade de generalização.

Para prevenir casos de *overfitting*, existem diversas medidas de regularização de uma rede neural. Segundo Goodfellow et al. (2016), a camada intermediária de *dropout* é uma medida que funciona para diversos tipos de topologia. A função dessa camada é desativar alguns dos neurônios da camada associada com alguma probabilidade p. Um neurônio desativado sempre tem como saída o valor 0.

Outros parâmetros utilizados são a regularização L^1 e L^2. Ambos são utilizados para fazer o decaimento dos pesos da rede ao longo das épocas, e seu uso previne que os pesos de um modelo tenham uma granularidade muito alta. A diferença entre os dois se resume na regularização L^1 ser composta da soma atual dos pesos enquanto a regularização L^2 utiliza a soma dos pesos ao quadrado.

Considerando que redes neurais não geram necessariamente problemas com solução convexa (em que o gradiente do erro sempre vai na direção da solução global), a possibilidade do modelo de treinamento estagnar em um mínimo local existe. Uma das estratégias para aumentar a variabilidade do espaço de busca é alterar a taxa de aprendizado (*learning rate*) do modelo ao longo das épocas. Taxas de aprendizado mais altas significam passos mais largos dentro do espaço de busca, permitindo a rede “escapar” de mínimos locais, enquanto taxas mais baixas permitem que o modelo faça o fino ajuste de atualização dos pesos para chegar o mais próximo possível do mínimo global, porém a convergência tende a ser mais lenta.

Capítulo 3

Trabalhos Relacionados

No presente capítulo, são apresentadas diferentes abordagens para o problema de identificar o impacto de variantes em uma característica epigenética específica utilizando redes neurais convolucionais, em busca de se obter mais informações do potencial impacto funcional de variantes. Todas as estratégias apresentadas utilizam uma sequência de DNA como entrada.

3.1 Predição de efeitos epigenéticos utilizando *Deep Learning*

Variantes de um único nucleotídeo do DNA em um cromossomo já podem ter um impacto significativo em diversas vias metabólicas. Variações localizadas nos exons de um gene podem alterar a estrutura final do seu produto de acordo com os casos descritos na tabela 2.1. Em todos os casos, a proteína resultante tem chances de não ser funcional, trazendo impactos negativos em uma ou mais vias metabólicas.

Quando as variantes estão em regiões não codificantes, as mudanças possíveis tendem a ser mais sutis, e geralmente são epigenéticas (SIMMONS; SNUSTAD et al., 2006). A adição ou remoção de uma citosina ou guanina por exemplo pode alterar sítios de metilação ou até mesmo tornar possível o surgimento ou desaparecimento de Ilhas CpG.

Para identificar potenciais sítios de metilação em diferentes tipos de células Angermueller et al. (2017) combinaram dados de diferentes tipos de classificadores com diferentes dados de entrada; a própria sequência de DNA foi utilizada em conjunto com dados de expressão gênica, a fim de identificar potenciais sítios reguladores na sequência. A sequência de DNA que foi utilizada era composta 1001 pares de base de comprimento e contendo exatamente no meio um dinucleotídeo CpG e foi a informação de entrada em uma CNN. Os dados de perfis de expressão gênica foram obtidos de células por meio de técnicas específicas como *scBS-seq* e *scRRBS-seq*. Locais que possuíam ilhas CpG metiladas eram identificadas com o número 1, sendo as não metiladas identificadas com zeros. Quando não se sabia o estado de metilação de uma ilha CpG identificava-se com um ponto de interrogação. Esses dados foram então utilizados para treinar uma rede neural recorrente (RNN) bidirecional, que é capaz de analisar as relações de vizinhança dos estados CpG de um número variável de células em um vetor de características de tamanho fixo. Após o treinamento em ambas as redes, os resultados foram combinados por meio de uma MLP densa, que retorna a predição de cada estado de metilação nos locais sequenciados para cada célula. A arquitetura da rede pode ser vista na figura 3.1.

Com o objetivo de predizer qual entre os 2002 possíveis estados a cromatina pode estar em um tipo celular, Zhou e Troyanskaya (2015) utilizaram dados de sequências contendo

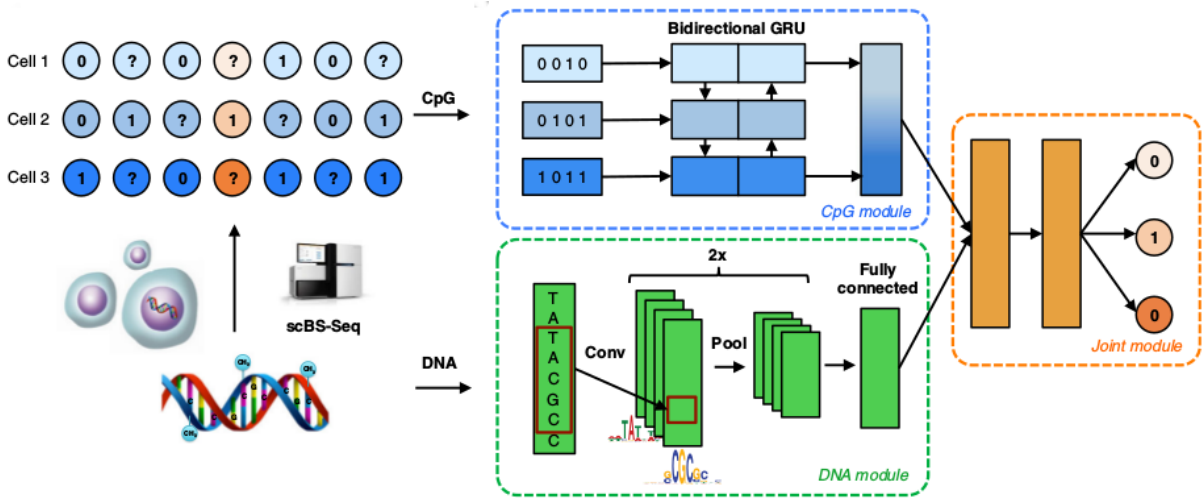


Figura 3.1: Topologia das redes neurais utilizadas por Angermueller et al. (2017). A primeira rede é uma CNN que recebe como entrada uma sequência de DNA; a segunda rede é uma RNN bidirecional que recebe dados de sobre a metilação dessas sequências. As duas redes são combinadas e predizem se o sítio central da sequência de DNA de entrada está ou não metilado.

variantes genéticas em conjunto com informações epigenéticas provenientes do projeto ENCODE (CONSORTIUM et al., 2007) bem como o projeto *Roadmap Epigenomics* (BERNSTEIN et al., 2010). As informações foram utilizadas em conjunto, treinando uma mesma rede neural convolucional que contém três camadas de convolução e duas de *pooling*. O fluxo de trabalho geral pode ser visto na figura 3.2.

Os dados epigenéticos usados foram a marcação de histonas, um teste de DNAse que indica qual o nível de compactação da cromatina em uma determinada sequência e se a sequência possuia sítio de ligação a fatores de transcrição ativos em cada linhagem celular associada a um dos tecidos disponíveis no portal GTEx (LONSDALE et al., 2013). A sequência de DNA propriamente dita possui 1001 pares de base, contendo um SNP exatamente na posição do meio. O resultado é a predição do estado da cromatina, e com esta informação os autores conseguem priorizar os potenciais impactos funcionais de variantes não codificantes.

As principais diferenças descritas no trabalho mais recente de Zhou et al. (2018) são a inclusão de mais camadas convolucionais e de *pooling*. O objetivo das alterações foi aumentar a precisão da predição, que agora também inclui variantes codificantes. O tamanho da sequência foi alterado para 2001 pares de bases.

Todos os trabalhos citados anteriormente possuem em comum o objetivo de prever o impacto de variantes genéticas na regulação da expressão gênica em seres humanos, além da abordagem que utiliza redes neurais convolucionais com sequências de DNA como entrada da rede. Os resultados, que combinam vários tipos de informação adicional além da sequência, são promissores e têm alta taxas de acerto superiores à 80% . O trabalho de Angermueller et al. (2017) possui uma área sob a curva (AUC) de 83% e os trabalhos Zhou obtiveram 95,8% e 81,5%. Porém, a grande quantidade de dados que são de origens diferentes e a complexidade da topologia das redes torna a replicação destes métodos inviável para o presente trabalho , que possui o objetivo de predizer se existe uma variante com efeito eQTL em uma sequência. Como este é um problema de classificação binário, acredita-se que uma topologia de rede mais simples possa obter resultados similares.

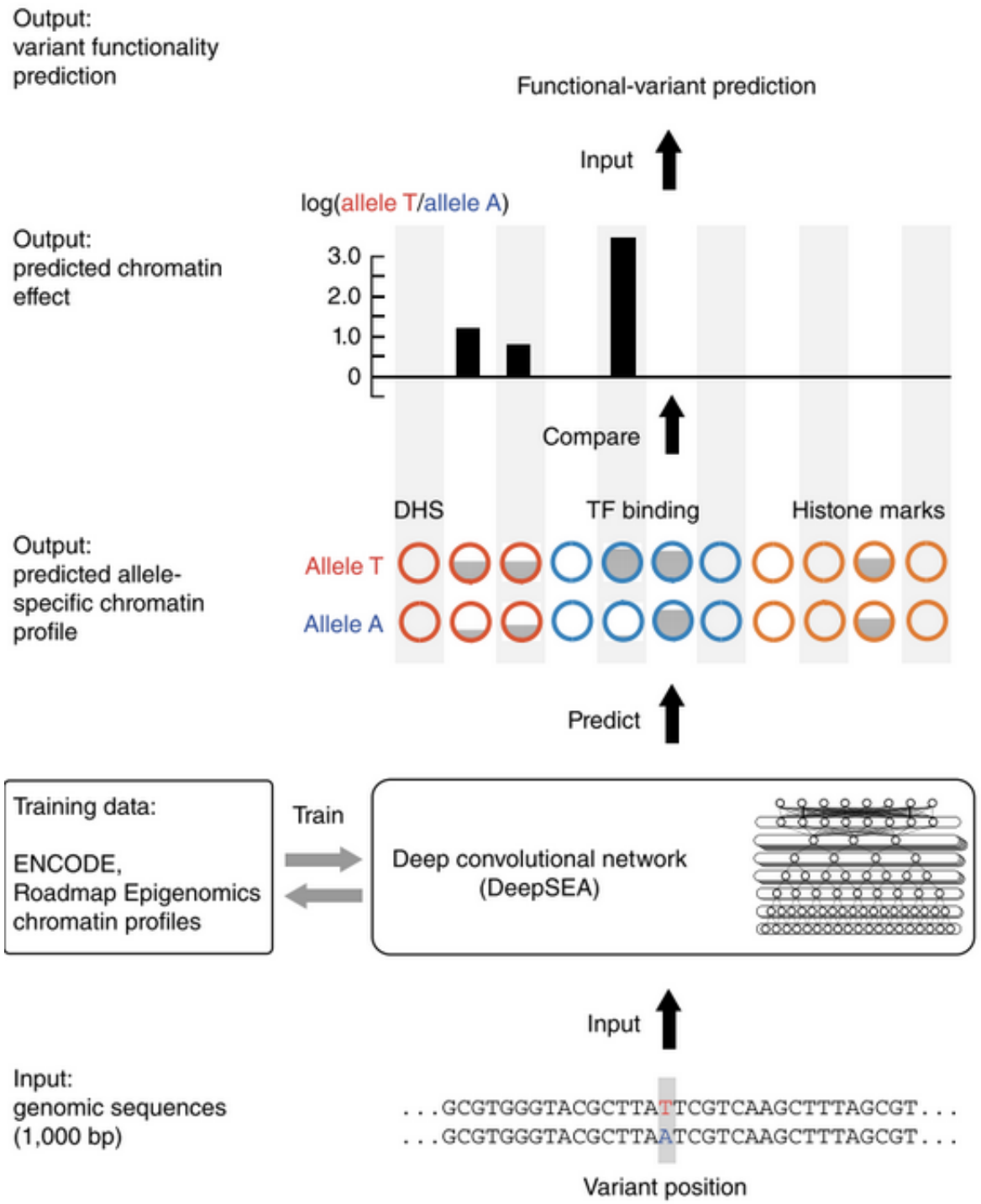


Figura 3.2: *Workflow* do trabalho de Zhou e Troyanskaya (2015). A entrada da CNN é composta de uma sequência de DNA e de dados epigenéticos de diferentes tipos celulares. A saída esperada é qual o perfil de cromatina aquelase encontra naquela região do DNA.

Capítulo 4

Materiais e Métodos

O presente capítulo é subdividido em três partes: a análise de variantes codificadoras de proteínas, a predição da existencia ou não de variantes com efeito eQTL e a indicação de possíveis vias metabólicas nas quais variantes encontradas podem ter impacto.

4.1 Variantes codificadoras de proteínas

Identificar potenciais efeitos em variantes genéticas localizadas em sequências codificadoras de proteínas é um problema bem estabelecido na literatura. Entre as ferramentas já disponibilizadas de forma aberta, a solução ANNOVAR (WANG; LI; HAKONARSON, 2010) foi a escolhida para ser utilizada.

ANNOVAR é um conjunto de ferramentas que busca em bancos de dados genômicos informações para fazer a anotação de variantes genéticas. O *software* é capaz de identificar se SNPs ou outros tipos de variantes são codificadores de proteínas ou não, assim como seus subtipos (variante intrônica, variante de *splicing*, entre outros). Para variantes codificantes, a ferramenta também prediz com base em diferentes indicadores qual o grau do impacto funcional da proteína resultante.

O ANNOVAR é capaz de receber uma alta gama de tipos de entrada diferentes, e a escolhida para o projeto são arquivos VCF. O formato é utilizado para guardar informações de sequências genéticas que contêm variantes, mas não qual a sequência de nucleotídeos em si. É necessário junto com o arquivo informar qual genoma de referência foi utilizado para o sequenciamento do DNA do organismo.

Arquivos VCF são tabelados. Além das informações de cabeçalho que dizem sobre o método de sequenciamento e o organismo de referência, existem outros campos fixos e variáveis. Os campos a serem utilizados são os descritos na tabela 4.1.

Campo	**Informação**
CROM	O cromossomo em que se localiza a variante encontrada
POS	A posição da variação dentro do cromossomo indicado
ID	Se já existe um ou mais identificadores para essa variante na base de dados dbSNP. Se desconhecido, "."
REF	O alelo de referência na posição da variante
ALT	Os possíveis alelos alternativos na posição da variante

Tabela 4.1: Campos de dados de um arquivo VCF

Após a leitura do arquivo VCF pelo ANNOVAR, o resultado esperado de cada variante pode ser dos seguintes tipos:

- ***exonic***: variante localizada na região codificante de um gene (exon);
- ***splicing***: variante localizada numa região em que pode ocorrer eventos de *splicing* alternativo;
- ***ncRNA***: variante se sobrepõe a um transcrito que não codifica proteínas;
- ***UTR5***: variante localizada na região reguladora 5' UTR;
- ***UTR3***: variante localizada na região reguladora 3' UTR;
- ***intronic***: variante localizada entre dois exons (intron);
- ***upstream***: variante localizada à até 1kb antes do sítio de início da transcrição;
- ***downstream***: variante localizada à até 1kb depois do sítio de fim da transcrição;
- ***intergenic***: variante localizada entre dois genes.

Dentre essas opções, todas os SNPs classificados como algo diferente de *exonic* serão direcionadas como entradas de teste da rede neural descrita na próxima seção. As variantes localizadas em regiões codificantes apresentam já na saída do programa a predição do impacto funcional do seu produto gênico associado. Os algoritmos descritos por Adzhubei et al. (2010) e Sim et al. (2012) já implementados no ANNOVAR fazem referência às mudanças em codons e qual o grau de perda de funcionalidade da proteína essas alterações de aminoácidos podem causar.

4.2 Variantes não-codificantes

A base de dados principal utilizada para o treinamento da rede a ser proposta foi obtida do consórcio *Genotype-Tissue Expression* (GTEx) (LONSDALE et al., 2013). O projeto consiste numa base de dados pública de informações de expressão gênica específica para diferentes tecidos do corpo humano obtidos a partir da quantificação de mRNA. Amostras foram obtidas de 53 tecidos saudáveis diferentes a partir de aproximadamente 1000 indivíduos.

A proposta do presente trabalho é utilizar uma rede neural convolucional, que a partir de uma sequência de DNA contendo o SNP, pretende classificar o SNP como possuindo ou não possuindo efeito eQTL. As informações obtidas pela base de dados são a classificação acurada de variantes e suas posições genômicas que possuem efeito eQTL ou não. Para saber qual é a sequência de DNA em torno do SNP, foi usada a montagem genômica de hg19 disponibilizada pelo *National Center for Biotechnology Information* (NCBI).

Em humanos, a maioria dos perfis de expressão gênica são específicos para cada tecido do corpo (ALBERTS et al., 2017). Portanto, para estudar se existem variantes que possuem um efeito eQTL, é necessário analisar cada tipo de tecido de forma separada. A estrutura de uma amostra de entrada da rede é constituída de acordo com a tabela 4.2. O tamanho escolhido para a sequência de DNA (201) foi baseada nas limitações de espaço e processamento dos dados, bem como no trabalho de Zhou e Troyanskaya (2015), que obteve resultados positivos já com sequências deste tamanho.

Característica	Descrição
REF	Alelo de referência segundo hg19
ALT	Alelo alternativo
SEQ	Sequência de DNA contendo 201 nucleotídeos, no qual o nucleotídeo da posição 101 é sempre igual a ALT

Tabela 4.2: Características escolhidas como entrada da CNN

4.2.1 Pré-processamento

O pré-processamento dos dados consistiu em transformar os nucleotídeos da sequência da DNA ao redor da variante em uma matriz de *one-hot encoding*. Essa matriz consiste na transformação de variáveis categóricas (no caso, os nucleotídeos A, C, G e T) em uma representação que contém o tamanho da entrada $\times$ o número de características, com apenas uma coluna de cada linha possuindo o número um e o restante possuindo zeros. A figura 4.1 ilustra esse processo.

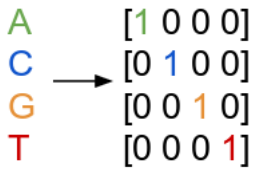


Figura 4.1: *One-hot encoding* para sequências de DNA. Cada letra é interpretada como um vetor binário equivalente. Fonte: autoria própria

A topologia da rede escolhida foi similar a de Zhou e Troyanskaya (2015), por apresentar uma estrutura simples e de baixo custo computacional. Ela está organizada de forma geral segundo a figura 4.2 e a sua estrutura detalhada pode ser vista na figura 4.3.

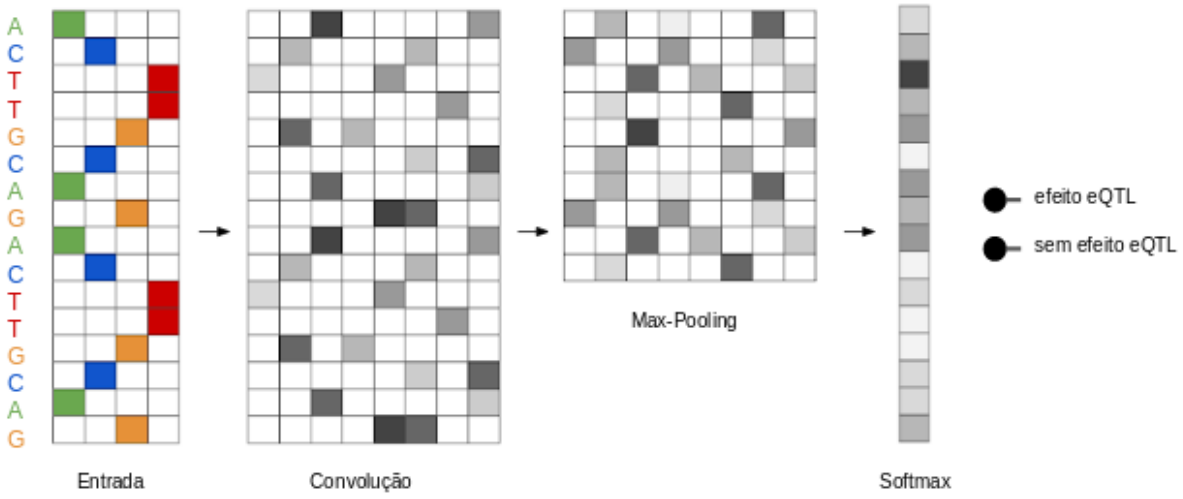
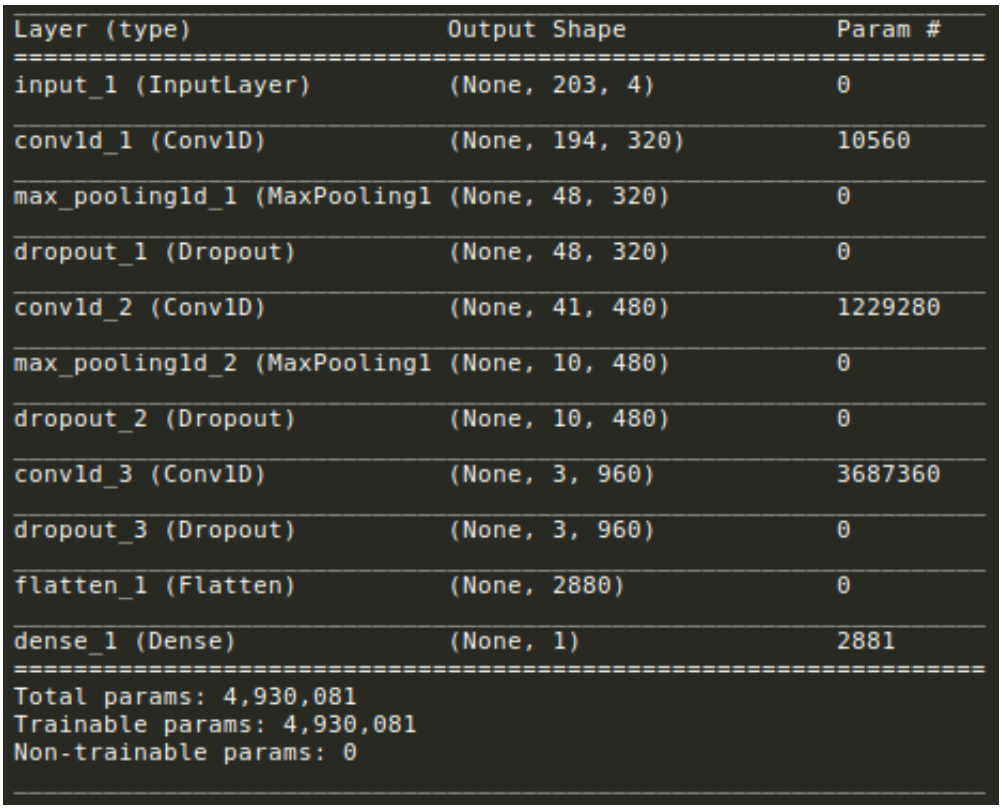


Figura 4.2: Topologia geral da rede neural proposta. A sequência de DNA de entrada sofre sucessivos processos de convolução e pooling, para no fim existir uma camada densa que classifica se a entrada possui ou não efeito eQTL. Fonte: autoria própria



Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 203, 4)	0
conv1d_1 (Conv1D)	(None, 194, 320)	10560
max_pooling1d_1 (MaxPooling1	(None, 48, 320)	0
dropout_1 (Dropout)	(None, 48, 320)	0
conv1d_2 (Conv1D)	(None, 41, 480)	1229280
max_pooling1d_2 (MaxPooling1	(None, 10, 480)	0
dropout_2 (Dropout)	(None, 10, 480)	0
conv1d_3 (Conv1D)	(None, 3, 960)	3687360
dropout_3 (Dropout)	(None, 3, 960)	0
flatten_1 (Flatten)	(None, 2880)	0
dense_1 (Dense)	(None, 1)	2881

Total params: 4,930,081
Trainable params: 4,930,081
Non-trainable params: 0

Figura 4.3: Topologia detalhada da rede neural proposta. Fonte: autoria própria

A entrada é constituída de N amostras de 201 + 2 nucleotídeos $\times$ 4 bases possíveis codificadas como *one-hot*. O primeiro nucleotídeo é REF, o segundo ALT e todos os seguintes são a sequência de DNA contendo ALT na posição do meio. A proporção entre o número de amostras para treinamento e teste da rede foi 80% para treinamento e 20% para teste.

No total, a rede é composta de três camadas de convolução com a função de ativação ReLU (equação 2.4) , sendo duas dessas seguidas por camadas de *max-pooling*. As camadas de convolução possuem filtros de tamanho 320, 480 e 960 respectivamente, todas com um *kernel* de tamanho 8. As camadas de *max-pooling* consideram uma vizinhança de tamanho 4.

Para evitar *overfitting*, três camadas de *Dropout* foram incluídas, após cada *pooling* e após a última convolução. Elas possuem probabilidade de desativarem os neurônios de 20%, 20% e 50%, respectivamente.

Após isso, há o achatamento da rede convolucional para que ela possa passar por uma camada Densa. A camada Densa é composta de apenas um neurônio para classificação binária e possui a função de ativação *softmax* que pode ser descrita utilizando a equação 4.1. É uma função que calcula a distribuição de probabilidades de um evento sobre N classes.

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{4.1}$$

Para o cálculo da função *loss*, a métrica escolhida foi a entropia cruzada binária. A entropia cruzada é usada como uma medida de erro quando as saídas de uma rede neural podem ser pensadas como independentes, e as ativações dos neurônios podem ser entendidas como representando a probabilidade de cada uma das hipóteses ser verdadeira. A saída representa uma distribuição de probabilidades, enquanto a entropia cruzada indica a distância entre o que a

rede acredita que essa distribuição deve ser e o que realmente deveria ser (GOODFELLOW et al., 2016).

O otimizador que visa minimizar a função *loss* utilizado é o gradiente descendente estocástico. Este é um método numérico iterativo que supõe que a solução é convexa. Ele utiliza equações diferenciais de primeira ordem para calcular o gradiente e vai em sua direção oposta. A versão estocástica do método utiliza não todo o conjunto de dados, mas um subconjunto gerado com reamostragem dos dados originais.

4.2.2 Implementação

Com o crescimento da popularidade de algoritmos de *Deep Learning*, bibliotecas que facilitam o seu uso estão surgindo e sendo atualizadas constantemente. Uma delas, desenvolvida pela Google, é o *Tensorflow*. É uma *Application Programming Interface* (API) desenvolvida para as linguagens *Python* e *C/C++*, e sua versão mais atual é a 1.12. Juntamente ao *Tensorflow*, que fornece uma base para operações com tensores e compatibilidade com processamento em *Graphics Processing Unit* (GPU), a API *Keras* na versão $2.2.2$ também foi utilizada por implementar por padrão vários tipos de camadas comuns em redes neurais, incluindo camadas Densas, Convolucionais, *Pooling*, além de possuir ferramentas de regularização da rede como a redução automática da *learning rate* quando o algoritmo não converge mais, entre outras ferramentas.

Callbacks da biblioteca *Keras* são um conjunto de funções aplicadas durante o processo de treinamento. As utilizadas no trabalho são:

- **Tensorboard**: é uma ferramenta de visualização do grafo computacional da rede, além de gerar gráficos sobre as métricas da execução do treinamento e validação, entre outros;
- ***ReduceLRonPlateau***: função que diminui a taxa de aprendizado durante o treinamento caso uma medida escolhida não melhore em um número determinado de épocas. No caso, a medida monitorada foi a `val_loss`, que caso permanecesse estável por um período de 10 épocas, a taxa de aprendizado era reduzida em 50%.
- ***ModelCheckpoint***: Os pesos sinápticos são salvos a cada melhora da taxa de `val_loss`.

No fim de toda a execução, os pesos da rede neural são salvos e a chamada de `model.predict()` é aplicada às variantes selecionadas pelo ANNOVAR descritas na seção anterior. As variantes não exônicas que foram preditas como possuindo efeito eQTL são selecionadas junto com as variantes exônicas que não foram utilizadas na rede neural para a próxima etapa.

4.3 Indicação das vias metabólicas

A busca de quais genes estão relacionados com quais vias metabólicas é feita por meio de requisições pela plataforma KEGG PATHWAY. A *Kyoto Encyclopedia of Genes and Genomes* é uma coleção de vias metabólicas manualmente curadas. Cada mapa metabólico é identificado por uma combinação de um prefixo que indica qual o tipo da informação (gene, via metabólica, proteína, entre outros) juntamente de um código identificador composto de cinco números.

A base de dados possui uma API que de forma geral é uma série de regras de comunicação de forma que outros programas sejam capazes de receber mensagens da base de dados. As requisições nesse caso utilizam o padrão cliente-servidor do modelo HTTP.

Cada requisição é feita a partir da seção KEGG GENES da base de dados, com os parâmetros de entrada sendo o cromossomo e posição cromossômicas de cada variante identificada como possuindo efeito eQTL ou o nome do gene caso ele tenha sido documentado pelo ANNOVAR para as variantes codificantes.

O retorno esperado é um arquivo estruturado com o padrão *JavaScript Object Notation* (JSON) contendo as informações em formato de texto de forma semelhante à pesquisa pelo navegador, exemplificada pela figura 4.4. Entre as informações retornadas, são algumas relevantes:

KEGG **Homo sapiens (human): 672** Help

Entry	672 CDS T01001
Gene name	BRCA1, BRCAI, BRCC1, BROVCA1, FANCS, IRIS, PNCA4, PPP1R53, PSCP, RNF53
Definition	(RefSeq) BRCA1, DNA repair associated
KO	K10605 breast cancer type 1 susceptibility protein [EC:2.3.2.27]
Organism	hsa Homo sapiens (human)
Pathway	hsa01524 Platinum drug resistance hsa03440 Homologous recombination hsa03460 Fanconi anemia pathway hsa04120 Ubiquitin mediated proteolysis hsa04151 PI3K-Akt signaling pathway hsa05206 MicroRNAs in cancer hsa05224 Breast cancer
Module	hsa_M00295 BRCA1-associated genome surveillance complex (BASC)
Disease	H00027 Ovarian cancer H00031 Breast cancer H01554 Fallopian tube cancer H01665 Primary peritoneal carcinoma

Figura 4.4: Exemplo de busca na base KEGG com o gene *BRCA1*

- **Entry**: identificador do gene na base de dados;
- **Gene name**: nomes possíveis para o gene;
- **Organism**: organismo de referência;
- **Pathway**: vias metabólicas associadas ao gene em questão;
- **Disease**: doenças que podem estar associadas ao gene;
- **Motif**: motivos associados àquela posição ou gene;
- **Other DBs**: identificadores desse mesmo gene em outros bancos de dados;
- **Position**: posição cromossômica do início do gene;

- **AA seq**: sequência de aminoácidos, caso o gene seja codificante de proteínas;
- **NT seq**: sequência de nucleotídeos do gene;

Capítulo 5

Experimentos

Este capítulo apresenta os resultados obtidos na implementação do projeto referentes à rede neural convolucional proposta na seção 4.2.

Entre os 53 tecidos disponibilizados pelo GTEx, foram utilizados dados de efeitos eQTL em cinco tecidos: fígado, hipocampo cerebral, sangue, tecido subcutâneo exposto ao sol e não exposto ao sol. Devido à proporção desigual entre o número de amostras classificadas como positivas (possuindo efeito eQTL) e negativas (não possuindo efeito eQTL), foi realizada uma reamostragem aleatória dentre os dados classificados como negativo, de forma que os dados ficassem balanceados em 50% para cada classificação.

Dois filtros de seleção foram aplicados ao conjunto total da base de dados. O primeiro foi remover variantes da amostra que se localizam em regiões próximas (1kb) as teloméricas, pois a sequência de nucleotídeosno início e fim de cada cromossomo possui muitos elementos desconhecidos. Também houve um filtro para selecionar apenas variantes em que o tamanho dos alelos alternativo e referência fossem 1 base nitrogenada, para conter apenas SNPs. No total, o número de amostras que foram utilizadas por cada tecido segue de acordo com a tabela 5.1.

Tecido	**Dados de treinamento**	**Dados de validação**
Fígado	1.594.832	398.708
Hipocampo	1.153.128	288.282
Sangue	5.080.624	1.270.156
Tecido subcutâneo exposto ao sol	8.388.608	2.097.152
Tecido subcutâneo não exposto ao sol	6.327.552	1.581.888

Tabela 5.1: Número de sequêcias obtidas para treinamento da rede neural

O treinamento das cinco redes neurais foram realizados em uma placa *NVIDIA TITAN XP* com 12GB de memória RAM e 3.840 núcleos CUDA, disponibilizado pelo laboratório de Visão, Robótica e Imagem do departamento de informática da UFPR. Cada época de treinamento demorou em torno de 30 minutos e todos os tecidos treinaram até entrar em convergência (quando a função *loss* para de diminuir), num limite máximo de 5000 épocas.

Todos os testes foram realizados com uma taxa de aprendizado inicial de 10^{-3} e taxa de decaimento dos pesos ao longo do gradiente descendente de $1e^{-6}$.

A principal métrica utilizada para avaliar o desempenho da rede, além do valor da função *loss* no conjunto de validação (`val_loss`), foi a taxa de acerto. Além disso, foram gerados os gráficos de área sob a curva de *Recieving Operating Characteristic* (ROC). A curva ROC é

um método de avaliar o desempenho de classificadores binários, pois é a relação da taxa de verdadeiros positivos (TPR) sobre a taxa de falsos positivos (FPR) encontrados na classificação. Esses valores são encontrados a partir da matriz de confusão, que pode ser vista na tabela 5.1.

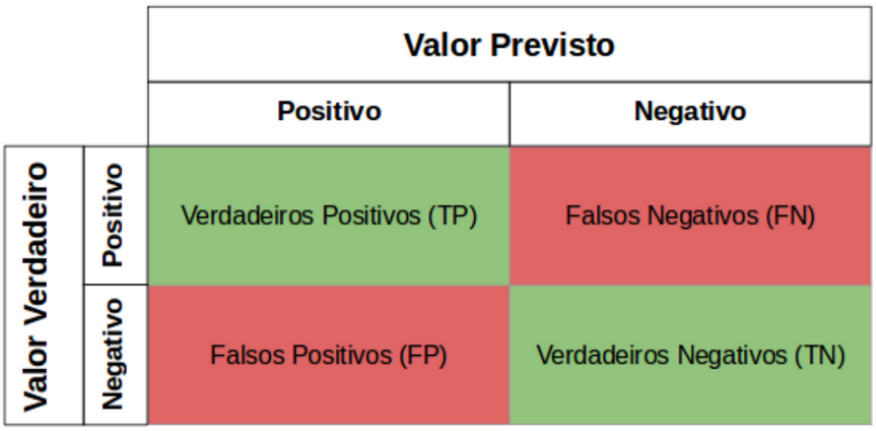


Figura 5.1: Matriz de confusão dos resultados. Fonte: autoria própria

A partir dessa matriz, é possível calcular os valores da taxa de verdadeiro positivo (TPR) ou sensitividade da equação 5.1; a taxa de verdadeiro negativo (TNR) ou especificidade da equação 5.2; a taxa de falso positivo (FPR) da equação 5.3 ou a taxa de falso negativo (FNR) da equação 5.4.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.1}$$

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5.2}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{5.3}$$

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{5.4}$$

5.1 Análise dos resultados

As curvas ROC bem como a área sob as curvas do conjunto de validação podem ser vistas nas imagens abaixo, onde a linha azul são os valores da curva ROC e a linha tracejada indica a posição de 0.5 de área sob a curva. Quanto mais próximo de 1 for a área sobre a curva, mais precisa é considerada a classificação. Todas as estatísticas geradas podem ser visualizadas de forma resumida na tabela 5.2.

5.1.1 Um exemplo no gene *SLX4*

O gene *SLX4* codifica uma proteína que funciona como um dos componentes de montagem de múltiplas estruturas específicas em endonucleases. Esses complexos em endonucleares são necessários para o reparo de tipos de lesões específicas de DNA e é crítico para a resposta celular à falha na replicação. O gene está localizado no braço curto do cromossomo 16 e é transcrito na fita de DNA reversa (3' $\rightarrow$ 5') (PRUITT; TATUSOVA; MAGLOTT, 2005). Segundo a

Tecido	Taxa de acerto média	F_1 *score*	AUC
Fígado	$87,07\%$	0.8728	0.948
Hipocampo	$74,17\%$	0.7465	0.817
Pele não exposta ao sol	$90,48\%$	0.9040	0.914
Pele exposta ao sol	$83,66\%$	0.8468	0.923
Sangue total	$85,46\%$	0.8598	0.873

Tabela 5.2: Resumo dos resultados obtidos

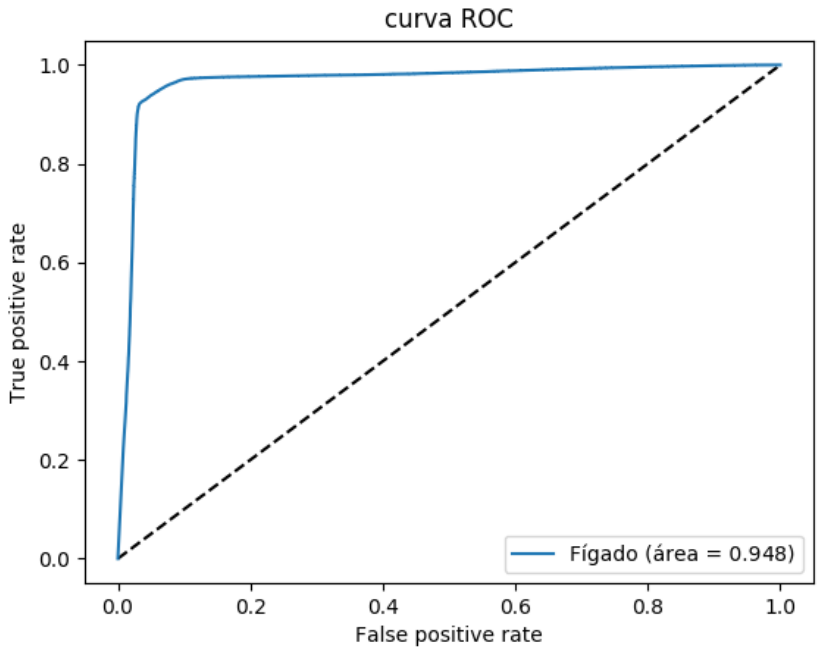


Figura 5.2: Área sob a curva ROC - Fígado. Fonte: autoria própria

plataforma KEGG, esse gene está associado com a anemia Fanconi, estando presente nesta via metabólica.

O SNP *rs78637028* desse gene foi um dos apontados como que possuem efeito eQTL pelas redes neurais nos tecidos do fígado e sangue. Fazendo a busca no portal GTEx (LONSDALE et al., 2013), foi verificado que este SNP de fato possui efeito eQTL no sangue, mas não no fígado. Isso indica um exemplo de amostra verdadeiramente positiva no treinamento da rede neural do sangue, e uma amostra falso positiva no treinamento da rede neural do fígado.

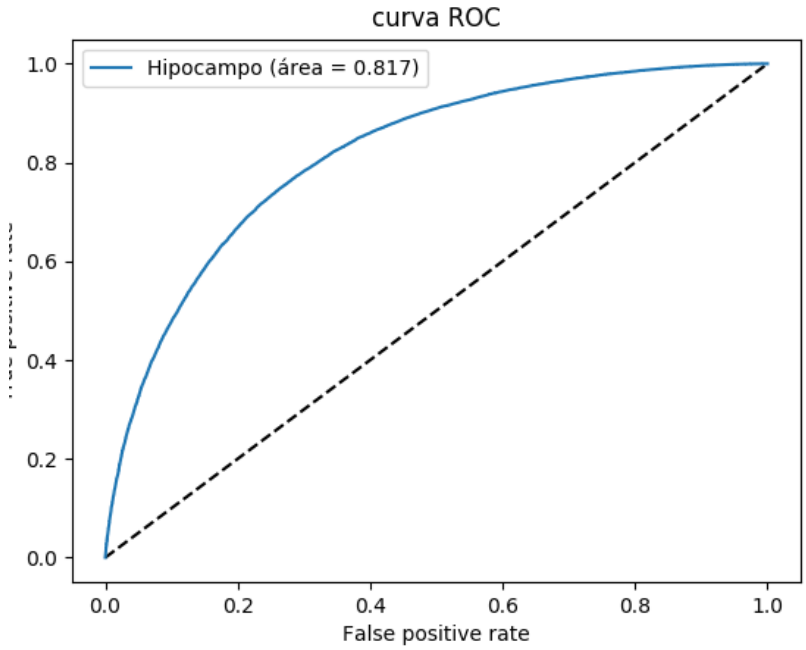


Figura 5.3: Área sob a curva ROC - Hipocampo. Fonte autoria própria

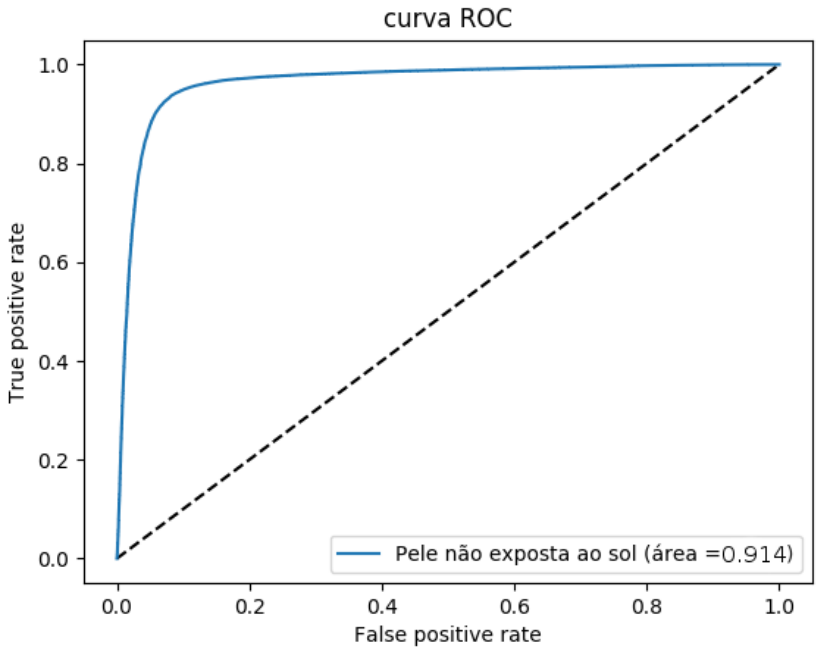


Figura 5.4: Área sob a curva ROC - Pele não exposta ao sol. Fonte: autoria própria

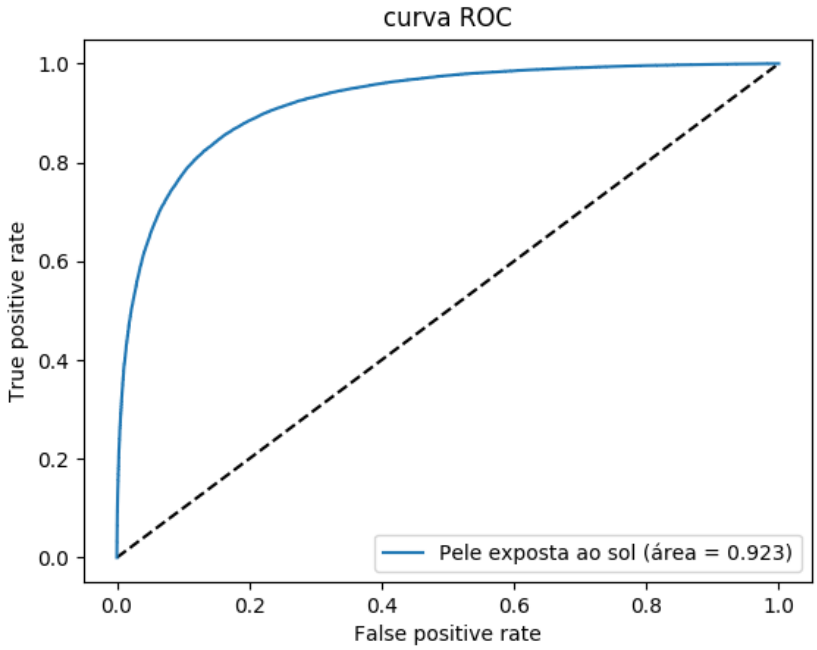


Figura 5.5: Área sob a curva ROC - Pele exposta ao sol. Fonte: autoria própria

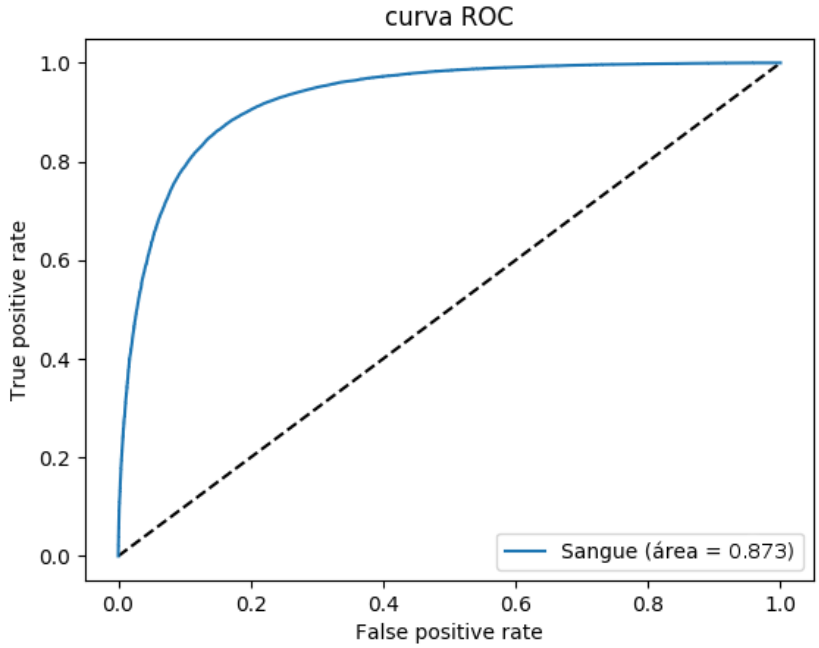


Figura 5.6: Área sob a curva ROC - Sangue. Fonte: autoria própria

Capítulo 6

Conclusão

Este trabalho apresentou uma possível abordagem para automatizar e facilitar o processo de identificação do impacto funcional de variantes genéticas em seres humanos. O processo não exime a validade de pesquisas de SNPs específicos em estudos funcionais.

A utilização de redes neurais profundas é o estado da arte para a predição variantes que possuem efeitos epigenéticos. O uso da sequência de DNA como entrada da rede torna possível a utilização de redes convolucionais, conhecidas por serem capazes de extrair características de matrizes. A topologia da rede proposta por Zhou e Troyanskaya (2015) possui poucas camadas convolucionais, tendo um bom desempenho computacional e sendo adequada para os experimentos realizados neste trabalho.

Os resultados obtidos se apresentam como promissores, mas há ainda espaço para aprimoramento da técnica. Os resultados indicam que existe uma sequência curta de DNA pode não ser capaz de descrever com alta precisão uma variante em efeito eQTL. Dados que informam sobre o estado epigenético da sequência, como o padrão de metilação, possuem grande potencial para melhorar os resultados obtidos. Além disso, o tamanho da sequência de entrada provavelmente não é capaz de descrever variantes que possuem efeito eQTL em *trans* que atuam em outros genes, devido à distância da região alvo da regulação.

Um grande desafio do projeto desde o início da sua concepção foi o tratamento dos dados de entrada. O volume inicial de dados diretamente extraído do portal GTEx era de aproximadamente 800GB, e o processamento desses dados para extrair as sequências de DNA referentes à cada variante tomou boa parte do tempo do projeto. Além disso, mesmo após o filtro das variantes escolhidas, a quantidade de dados para o treinamento da rede neural tornou esse volume intratável para a realização de testes em CPUs e para realizar ajustes nos hiperparâmetros de regularização da rede.

6.1 Trabalhos Futuros

A gama de projetos futuros a partir deste trabalho é extensa. Uma primeira abordagem futura possível seria testar outras topologias de redes neurais, além de utilizar outros otimizadores da função *loss*, visto que a probabilidade da solução no espaço de busca não ser convexa é grande.

Outra proposta seria tentar predizer com mais precisão qual o efeito da variante em relação ao aumento ou diminuição da expressão gênica. Para isso, talvez seja necessário utilizar informações provenientes de outras bases de dados como por exemplo do projeto ENCODE.

Um acontecimento biológico que não foi levado em consideração neste projeto foram variantes potencialmente em desequilíbrio de ligação com outras variantes, o que faz com que seja mais difícil identificar apenas uma variante causal para a alteração do fenótipo em questão.

Por fim, seria importante predizer variantes que possuem efeito eQTL em *trans*, como mencionado. Para isso, talvez seja necessário obter informações que indicam o local no núcleo celular em que a variante se encontra, pois variantes em efeito eQTL *trans* comumente interagem com genes muito distantes e até mesmo em outros cromossomos.

Referências Bibliográficas

1000GENOMES. An integrated map of genetic variation from 1,092 human genomes. *Nature*, Nature Publishing Group, v. 491, n. 7422, p. 56, 2012.

ADZHUBEI, I. A. et al. A method and server for predicting damaging missense mutations. *Nature methods*, Nature Publishing Group, v. 7, n. 4, p. 248, 2010.

ALBERTS, B. et al. *Molecular Biology of the Cell*. [S.l.: s.n.], 2017.

ANGERMUELLER, C. et al. Deepcpg: accurate prediction of single-cell dna methylation states using deep learning. *Genome biology*, BioMed Central, v. 18, n. 1, p. 67, 2017.

BARRETO, J. M. Introduçao as redes neurais artificiais. *V Escola Regional de Informática*, Sociedade Brasileira de Computaçao, 2002.

BERNSTEIN, B. E. et al. The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, Nature Publishing Group, v. 28, n. 10, p. 1045, 2010.

BISHOP, C. *Pattern Recognition And Machine Learning*. [S.l.]: Springer, 2006.

CONSORTIUM, E. P. et al. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, Nature Publishing Group, v. 447, n. 7146, p. 799, 2007.

GOODFELLOW, I. et al. *Deep learning*. [S.l.]: MIT press Cambridge, 2016. v. 1.

GRIFFITHS, A. J. et al. *An introduction to genetic analysis*. [S.l.]: Macmillan, 2005.

HAYKIN, S. S. et al. *Neural networks and learning machines*. [S.l.]: Pearson Upper Saddle River, 2009. v. 3.

ISI-TICS. *Principais sub divisões e aplicabilidade da aprendizagem de máquina*. 2018. <https://isitics.com/2018/05/10/principais-sub-divisoes-e-aplicabilidade-da-aprendizagem-de-maquina/>. Acessado em 06/11/2018.

ISKANDAR, J. *Normas da ABNT Comentadas para Trabalhos Científicos*. [S.l.]: Editora Champagnat (PUCPR), 2000.

KUMAR, P.; HENIKOFF, S.; NG, P. C. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, Nature Publishing Group, v. 4, n. 7, p. 1073, 2009.

LI, Y.; SHI, W.; WASSERMAN, W. W. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. *BMC bioinformatics*, BioMed Central, v. 19, n. 1, p. 202, 2018.

LONSDALE, J. et al. The genotype-tissue expression (gtex) project. *Nature genetics*, Nature Publishing Group, v. 45, n. 6, p. 580, 2013.

MAZIERO, C. *Modelo PPGInf UFPR para teses e dissertações*. 2015. <http://www.inf.ufpr.br/maziero>. Acessado em 30/11/2015.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. *Machine learning: An artificial intelligence approach*. [S.l.]: Springer Science & Business Media, 2013.

MIN, S.; LEE, B.; YOON, S. Deep learning in bioinformatics. *Briefings in bioinformatics*, Oxford University Press, v. 18, n. 5, p. 851–869, 2017.

NICA, A. C.; DERMITZAKIS, E. T. Expression quantitative trait loci: present and future. *Phil. Trans. R. Soc. B*, The Royal Society, v. 368, n. 1620, p. 20120362, 2013.

PIERCE, B. A. *Genetics: A conceptual approach*. [S.l.]: Macmillan, 2012.

PRUITT, K. D.; TATUSOVA, T.; MAGLOTT, D. R. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, Oxford University Press, v. 33, n. suppl_1, p. D501–D504, 2005.

SAWICKI, M. P. et al. Human genome project. *The American journal of surgery*, Elsevier, 2016.

SIM, N.-L. et al. Sift web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research*, Oxford University Press, v. 40, n. W1, p. W452–W457, 2012.

SIMMONS, M. J.; SNUSTAD, D. P. et al. *Principles of genetics*. [S.l.]: John Wiley & Sons, 2006.

SPRINGENBERG, J. T. et al. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

UFPR, B. *Manual de normalização de documentos científicos de acordo com as normas da ABNT*. Curitiba PR, 2015.

WANG, K.; LI, M.; HAKONARSON, H. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, Oxford University Press, v. 38, n. 16, p. e164–e164, 2010.

XIE, R. et al. A predictive model of gene expression using a deep learning framework. In: IEEE. *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. [S.l.], 2016. p. 676–681.

ZHOU, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, Nature Publishing Group, v. 50, n. 8, p. 1171, 2018.

ZHOU, J.; TROYANSKAYA, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, Nature Publishing Group, v. 12, n. 10, p. 931, 2015.