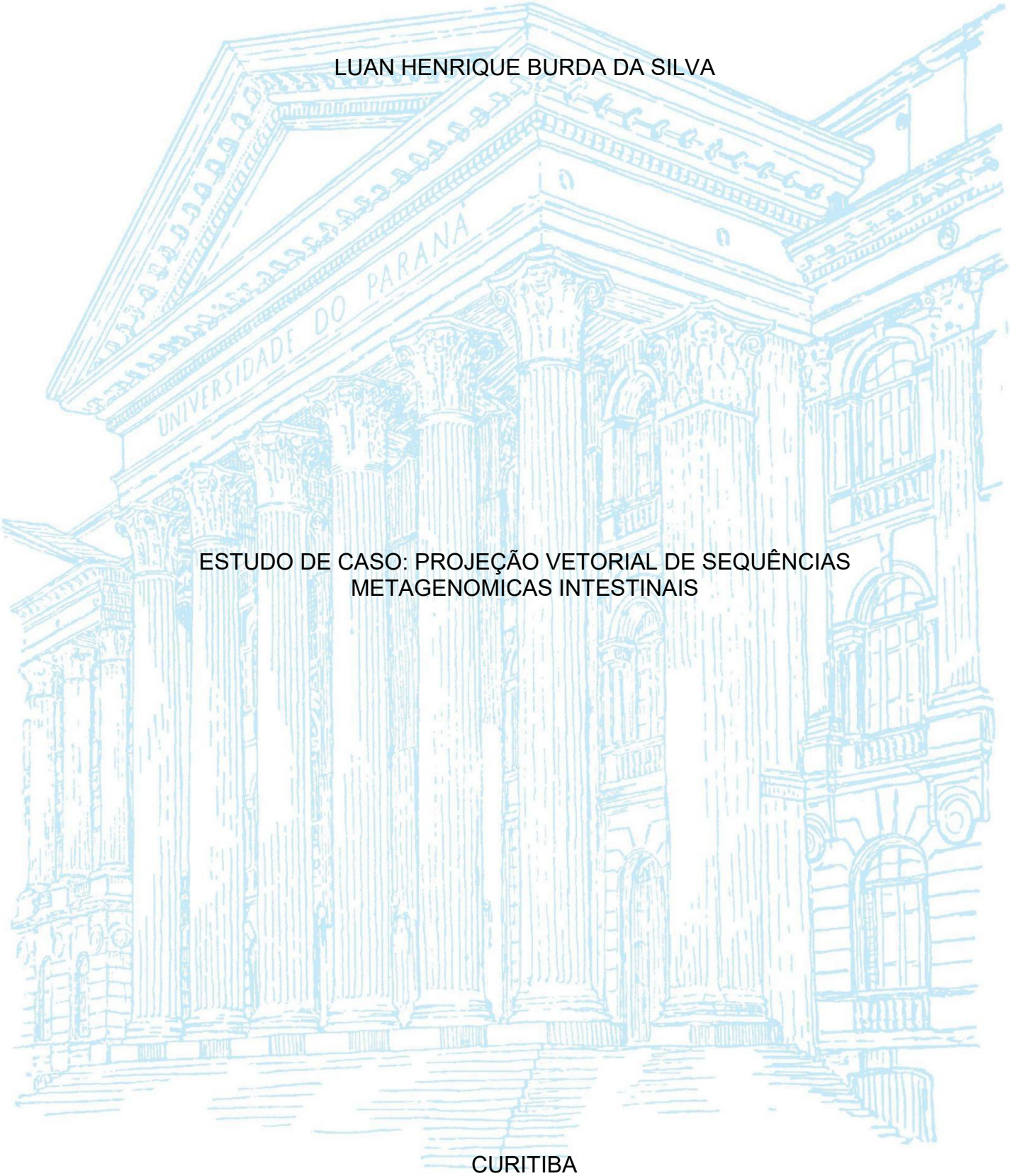
UNIVERSIDADE FEDERAL DO PARANÁ

LUAN HENRIQUE BURDA DA SILVA



ESTUDO DE CASO: PROJEÇÃO VETORIAL DE SEQUÊNCIAS METAGENOMICAS INTESTINAIS

CURITIBA

2019

LUAN HENRIQUE BURDA DA SILVA

ESTUDO DE CASO: PROJEÇÃO VETORIAL DE SEQUÊNCIAS METAGENÔMICAS INTESTINAIS

Monografia apresentada ao curso de Ciências Biológicas, Setor de Ciências Biológicas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Ciências Biológicas.

Orientador: Profº. Dr. Roberto Tadeu Raittz

Coorientadora: Ma. Camilla Reginatto de Pierri

CURITIBA

2019

TERMO DE APROVAÇÃO

LUAN HENRIQUE BURDA DA SILVA

ESTUDO DE CASO: PROJEÇÃO VETORIAL DE SEQUÊNCIAS METAGENÔMICAS INTESTINAIS

Monografia apresentada ao curso de Ciências Biológicas, Setor de Ciências Biológicas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Ciências Biológicas.

_____

Profº. Drº Roberto Tadeu Raittz

Orientador – SEPT - Bioinformática, Universidade Federal do Paraná

_____

Profª. Drª. Ana Cláudia Bonatto

Departamento de Genética, Univesidade Federal do Paraná

_____

Profª. Drª. Jeroniza Nunes Marchaukoski

SEPT - Bioinformática, Universidade Federal do Paraná

Curitiba, 22 de novembro de 2019.

Dedico este trabalho aos meus pais, amigos, aos que estiveram presentes da minha família e principalmente àqueles que buscam incessantemente aprender com os erros na ciência.

AGRADECIMENTOS

Agradeço aos meus pais, que sempre me deram apoio nas minhas decisões e me ampararam quando estive em situações de profunda tristeza. Além de me darem estrutura e amor em todos os dias da minha vida.

Agradeço a todos os meus amigos, principalmente meus também colegas de curso: Livia, Fernando e Yane pela companhia e amizade que tornou os dias dentro da universidade mais esperançosos. Também agradeço aos meus outros amigos que com certeza tornaram ao menos um dia mais alegre no decorrer do curso.

Aos colegas de laboratório anterior e atual, pelas conversas, dicas e conselhos diários que me auxiliaram na experiência dentro da Academia, mas que com certeza, também levarei para a vida.

Aos professores do Setor de Ciências Biológicas que me ajudaram na formação como bacharel, em principal aos meus orientadores anteriores de PVA Erika Amano e Marcos B. Carlucci que me introduziram na experiência em laboratórios e me deram muitos conselhos acerca do ambiente científico. Também em especial, agradeço a Monique e a Andressa por estarem presente ao longo dos melhores momentos de proveito acadêmico nos estágios.

Agradeço ao meu orientador de IC e de estágio supervisionado Profº Drº Roberto Tadeu Raittz pela ajuda na proposta deste trabalho e disponibilização das ferramentas necessárias, além da descontração que proporciona a todos do laboratório e agradeço também a minha Coorientadora Ma. Camilla Reginatto de Pierri pela paciência comigo, disponibilidade de tempo para me ajudar em revisões deste trabalho e pela companhia e parceria desde a Iniciação Científica.

Agradeço também a coordenação do curso, aos técnicos de laboratório e aos zeladores que auxiliaram nos trâmites de documentações, que proporcionaram aulas práticas com qualidade e que mantiveram os locais da universidade sempre nas melhores condições de zelo para as atividades acadêmicas, respectivamente.

Obrigado a todos que me acompanharam nesta jornada de 5 anos dentro do maravilhoso curso de ciências biológicas.

Truth is after all a moving target

Hairs to split and pieces that don't fit

How can anybody be enlightened?

Truth is after all so poorly lit

(**Neil Elwood Peart**, 1987)

RESUMO

Microrganismos presentes no lúmen intestinal são fatores característicos de algumas doenças, tumores, diferenças morfológicas e endócrinas em seres humanos. O metagenoma intestinal, caracterizado pela microbiota total encontrada neste habitat, é obtido por meio de sequenciamentos de genomas completos (WGS). Estes sequenciamentos geram dados com quantidades grandes de *reads* que podem ultrapassar facilmente 40 milhões de trechos de sequências, o que resulta em uma alta porcentagem de informações de organismos desconhecidos. Devido à dificuldade de comparação destes dados, algumas metodologias em Bioinformática são definitivas para resolver os problemas associados à manipulação de grandes quantidades de dados, como é o caso de metagenomas. Existem diversas ferramentas que auxiliam no processamento de dados de larga escala, entretanto, até o momento não existem modelos específicos aplicados à análises de metagenomas que apresentem acurácia, agilidade e baixo custo computacional. Nesta perspectiva, a ferramenta SWeeP, que consiste em um modelo computacional aplicado à análises de *Machine Learning* baseado em representação vetorial de genes, permite a redução de dimensionalidade das sequências biológicas, apresentado potencial para análises de grandes quantidades de dados. Dessa forma, devido à importância dos estudos relacionados ao microbioma intestinal humano e a dificuldade de manipulação destes dados, este estudo tem como propósito a exploração do modelo SWeeP para análises metagenômicas. Para isso, foram selecionados para estudo de caso sequências biológicas da microbiota intestinal dentre 2103 indivíduos, selecionados a partir de 11 estudos científicos, um único estudo com 7 indivíduos e 70 amostras. Após filtragem e pré-processamento destes dados, foi realizada a montagem dos metagenomas utilizando o software MetaSPAdes, no intuito de obter contigs mais consistentes. Estes contigs foram representados vetorialmente utilizando o SWeeP com parâmetros default. Testes de performance, redes neurais artificiais e análises filogenéticas foram conduzidas para explorar a capacidade do método na comparação de metagenomas, bem como a agilidade na construção dos vetores. As árvores filogenéticas indicaram uma proximidade entre as amostras dos mesmos indivíduos, apesar de se encontrar algumas amostras deslocadas. Por este motivo, foi treinado uma rede neural artificial identificando parcialmente algumas amostras como conjunto de treinamento e testada em seguida. Obtivemos uma acurácia média de 96,43%, obtendo correlação de Pearson média de 0,83.

Palavras-chave: Projeção Vetorial; Metagenoma; Filogenia; Bioinformática; Microbioma intestinal;

ABSTRACT

Microorganisms present in the intestinal lumen are characteristic factors of some diseases, tumors, morphological and endocrine differences in humans. Intestinal metagenome, characterized by the total microbiota found in this habitat, is obtained by complete genome sequencing (WGS). These sequences generate data with too much reads that can easily exceed 40 million passages of sequences, resulting in a high percentage of information from unknown organisms. Due to the difficulty of comparing these data, some methodologies in Bioinformatics are definitive to solve the problems associated with the manipulation of large amounts of data, such as metagenomes. There are several tools that assist in the processing of large-scale data, however, so far there are no specific models applied to metagenome analysis that present accuracy, agility and low computational cost. In this perspective, the SWeeP tool, which consists of a computational model applied to Machine Learning analysis based on vector representation of genes, allows the reduction of dimensionality of biological sequences, presenting potential for analysis of large amounts of data. Thus, due to the importance of studies related to the human intestinal microbiome and the difficulty of manipulating these data, this study aims to explore the SWeeP model for metagenomic analyzes. For this, biological sequences of the intestinal microbiota were selected for case study among 2103 individuals, selected from 11 scientific studies, a single study with 7 individuals and 70 samples. After filtering and preprocessing these data, the metagenomas were assembled using the MetaSPAdes software in order to obtain more consistent contigs. These contigs were represented vectorily using SWeeP with default parameters. Performance tests, artificial neural networks and phylogenetic analyzes were conducted to explore the method's ability to compare metagenomes as well as the agility in vector construction. Phylogenetic trees indicated a proximity between samples from the same individuals, although some misplaced samples were found. For this reason, an artificial neural network was partially identified by identifying some samples as a training set and then tested. We obtained an average accuracy of 96.43%, obtaining an average Pearson correlation of 0.83.

Keywords: 1Vector Projection. 2.Metagenomes. 3.Bioinformatics. 4.Phylogeny. 5.Intestinal Microbiome.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS OU SIGLAS

SRA	- Sequence Read Archive
NCBI	- Nacional Center of Biological Information
ORF’s	- Open Reading Frames
WGS	- Whole Genome Sequencing
NGS	- New Generation Sequencing
SWeeP	- Spaced Words Projection
HGF	- Hybrid Gene Finder
IA	- Inteligência Artificial
APC	- Análise de Componentes Principais
RNA	- Redes Neuronais Artificiais

LISTA DE SÍMBOLOS

@ - Arroba

® - Marca Registrada

SUMÁRIO

1 INTRODUÇÃO

A prática laboratorial de sequenciamento e cultivo de microrganismos isolados se tornaram inviáveis ao passo que a maior parte do material genético foi descoberta como “incultiváveis”. O suprimento desta falta de informação, é fornecida pela metagenômica, que consiste em sequenciamentos de amostras ambientais sem o cultivo.

A Metagenômica é definida como um ramo da genômica aplicada a amostras ambientais (HANDELSMAN, 2004) que busca desvendar a parcela não cultivável de microorganismos e de seus genes (ex. amostras de solo, amostras de ar, amostras de água, amostras de pele, intestinal, oral e vaginal), onde sequenciamentos geralmente com resultado de curtas leituras são utilizadas ou para aproximação da diversidade de espécies conhecidas em tais ambientes, pela amplificação de RNA 16s ou para a obtenção de diversos genes e/ou transcritos, obtidos pela técnica shotgun (WGS).

Através da aplicação das técnicas de sequenciamento de genoma completo (WGS) sobre os metagenomas, foi revelado uma diversidade diferente sob forma de um novo espectro de proteínas expressas em ambientes diferentes, como no rio dos Sargaços em São Francisco, EUA (VENTER et al, 2004) onde foi evidenciado grandes lacunas de informação dentro dos ambientes que nos contornam.

A análise de amostras ambientais se tornou cada vez mais relevantes após tais descobertas, havendo enfoques em sítios cada vez mais interessantes e trazendo esperança de um aumento de produtividade industrial ou de pesquisa, assim como uma nova tecnologia para diagnósticos clínicos de indivíduos ou a possibilidade de tratamento de alguns tipos de doenças (TURNBAUGH *et al*, 2009). A aproximação metagenômica se tornou mais interessante à alguns pesquisadores, especificamente tratando de microbioma intestinal (HUTTENHOWER *et al*, 2012) expandindo estudos na perspectiva de diversas regiões e nacionalidades, além de consequentemente aumentar a quantidade de dados biológicos em diferentes bancos de dados.

O NCBI, como uma destas fontes de informação, tem disponibilizado acessos à dados de genomas e metagenomas, sequenciamentos, taxonomia, proteínas, e diversos outros atributos recorrentes em análises de seres vivos (O'LEARY *et al.*, 2016). Além de poder reproduzir essencialmente os estudos de outras pessoas, novas análises são possíveis através de arquivos de leituras de sequências (Sequence Read Archive - SRA), os quais são essencialmente arquivos nativos de sequenciamentos sem alterações e dispostos em várias sequências em um único FASTA.

Porém, análises comparativas utilizando metagenomas não são triviais, devido à grande quantidade de dados de sequências que são gerados a partir dos sequenciamentos e montagens referência. Assim, as metodologias em Bioinformática são uma estratégia definitiva para resolver os problemas associados à manipulação de grandes quantidades de dados (NURK *et al,* 2017). Entretanto, até o momento as ferramentas específicas voltadas à análises de metagenomas não têm sido suficientes para auxiliar no processamento de dados de larga escala.

Em decorrência das quantidades crescentes de pesquisas que utilizam o microbioma intestinal humano como estudo de caso (QIN, J. et al, 2010; KARLSSON et al, 2013; NIELSEN et al, 2014; NISHIJIMA et al, 2014; ZELLER et al, 2014) e o consequente aumento de dados biológicos provenientes destes estudos, a utilização de métodos de representação de sequências biológicas em espaços vetoriais como alternativa para a redução de dimensionalidade de tais dados é uma opção favorável. A ferramenta SWeeP, que consiste em um modelo computacional aplicado ao aprendizado de máquina, contempla todos os aspectos necessários para análises de grandes quantidades de dados, podendo vir a ser uma importante aliada nas análises de metagenomas.

O aprendizado de máquina é uma área da Inteligência Artificial que busca definir um conjunto de abordagens para identificar padrões em dados, envolvendo métodos e algoritmos que possuem a capacidade prever comportamentos e extrair informações automaticamente (ALLAHYARI *et al.*, 2017). Portanto, a proposta deste trabalho é viabilizar a utilização do modelo de projeção vetorial (SWeeP) para a compactação de grandes quantidades de dados utilizando sequências de metagenomas intestinais humanos como estudo de caso, bem como explorar

árvores filogenéticas e abordagens de aprendizado de máquina nas representações vetoriais resultantes.

Nesta perspectiva, a questão que norteou este trabalho foi: É possível a avaliação por um método rápido de representatividade (projeção vetorial) de sequências de metagenoma intestinal humano de modo a obter uma identificação consistente dentre um conjunto de amostras contendo diferentes indivíduos?

2 OBJETIVOS

2.1 OBJETIVO GERAL

- ✓ Aplicar o modelo SWeeP para análises de metagenoma, utilizando como estudo de caso o microbioma intestinal humano.

2.2 OBJETIVOS ESPECÍFICOS

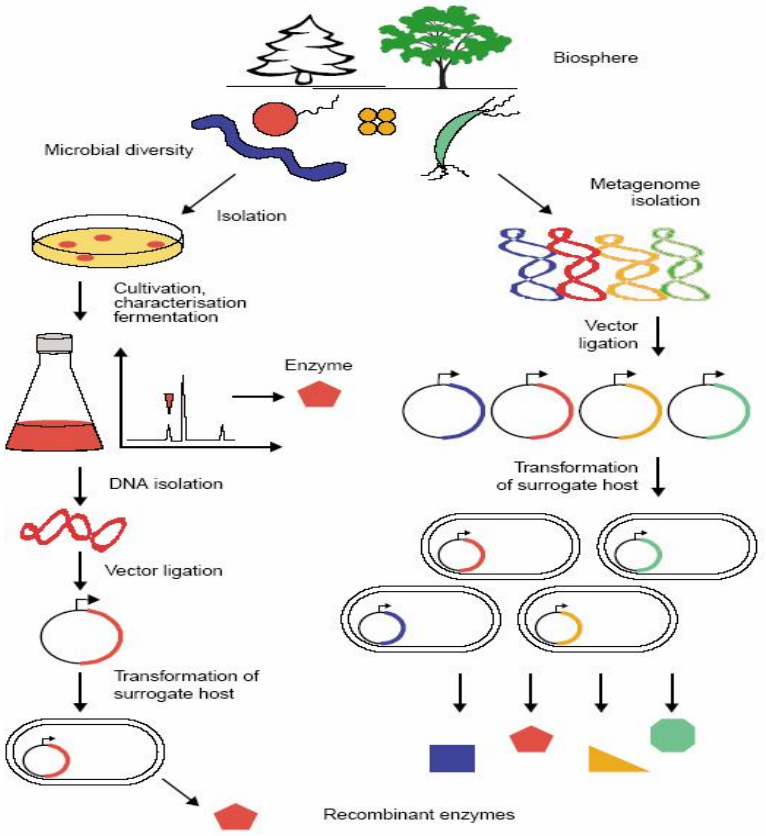
- ✓ Obter sequências de metagenomas de microbioma intestinal de bancos de dados públicos, selecionando dados de diferentes localidades (países);
- ✓ Realizar a montagem de sequências, utilizando ferramentas especializadas em metagenomas;
- ✓ Padronizar as informações obtidas, convertendo em um único padrão de dado;
- ✓ Utilizar a ferramenta HGF para a descoberta de ORF's nas leituras de metagenoma para obter possíveis genes contidos nas sequências;
- ✓ Padronizar os dados e aplicar ao modelo SWeeP;
- ✓ Realizar análise filogenética utilizando os vetores gerados pelo SWeeP;
- ✓ Desenvolver uma abordagem de aprendizado de máquina para demonstrar a capacidade de classificação do modelo SWeeP para metagenomas;

3 REVISÃO DE LITERATURA

3.1 METAGENÔMICA INTESTINAL

A metagenômica é um ramo da genômica inicialmente estudada no fim do século passado caracterizando-se como o sequenciamento de amostras ambientais. Esses sequenciamentos não recebem nenhum tratamento prévio no que diz respeito ao cultivo, tendendo a ser um desafio na análise do reconhecimento de organismos (HANDELSMAN, 2004; SLEATOR et al, 2007). Existem sequenciamentos utilizando-se de bibliotecas de RNA 16s que geram amplicons e são muito efetivos para utilização usual na busca de diversidade de ambientes. Por outro lado, sequenciamentos WGS (*Whole-Genome Sequencing*) geram dados com bibliotecas randômicas que trazem um aglomerado de fragmentos da amostra, trazendo informações amplas, porém misturadas e dispersas. A seguir, os. processos diferenciais entre genômica e metagenômica são representados (FIGURA 1).

FIGURA 1 - PROCESSOS DIFERENCIAIS ENTRE GENÔMICA E METAGENÔMICA



Fonte: Oliveira, Sette e Fantinatti-Garboggini (2006)

O aumento de publicações e de dados biológicos já é algo facilmente visível em sites de buscas e até mesmo pelo desenvolvimento de ferramentas que auxiliam/realizam uma procura cada vez mais profunda em dados publicados e armazenados em repositórios públicos, majoritariamente (MA; PRINCE; AAGARD, 2014). Não diferentemente, análises metagenômicas acabaram se tornando comuns dentro da ciência visto que, embora todas elas tragam novas informações, fundamentalmente explicam separadamente inexistindo protocolos para obter padrões no ramo.

Na medida que a ciência lentamente avança com qualidade, os registros são de muita importância para a replicação e para a validação dos resultados encontrados. Apesar disto, paralelamente a grande quantidade de publicações (SIMON et al, 2019), a ciência da computação com a produção de ferramentas de forma especialista nas últimas décadas, proporcionou métodos que auxiliam a pesquisa até mesmo na revisão bibliográfica, auxiliando na busca de referenciais de metagenomas. Esta ascensão traz estudos relevantes e “ocultos” (seja em revistas de menor impacto, pela quantidade de citações pequena ou cientistas iniciantes) para a utilização prática em trabalhos, permitindo com que projetos de pesquisas consigam obter dados e revisá-los pontualmente, sem com que haja tempo perdido ou dúvidas quanto a busca (OLIVEIRA, SETTE & FANTINATTI-GARBOGGINI, 2006), no caso, de estudos de metagenomas intestinais.

O microbioma intestinal se mostrou característico em diversos aspectos dos fenótipos de indivíduos, influenciando na quantidade de tecido adiposo acumulada e demonstrando alterações (ainda sem conclusões mais apuradas) em pacientes com diabetes, cirrose hepática e algumas doenças inflamatórias no próprio órgão (TURNBAUGH et al, 2009). Por esta razão existem estudos baseados na função, na busca de diversidade e outras abordagens que demonstram as variações do universo de microorganismos presentes no intestino humano através de abordagens metagenômicas, que como citado anteriormente, podem tanto trazer informações taxonômicas como informações de transcritos presentes no local da amostra (TYSON, 2004).

3.1.1 Microbioma intestinal humano

O ser humano resguarda grande valor em sua saúde junto aos microorganismos que lhe acompanham. Por permitirem a homeostase e proteção em diversas interações entre hospedeiro, a disbiose entre seu microbioma podem alterar o metabolismo, causar doenças leves ou graves e podendo até influenciar na incidência e progressão de carcinomas e outros tumores (LOPETUSO et al, 2017; SIEGEL et al, 2019). Fatores diferentes também podem influenciar nestes microbiomas, e na região intestinal já é fato que além do microbioma propriamente dito, fatores genéticos, ambientais, hábitos alimentares, entre outros podem entram neste rol no diagnóstico de doenças e alterações.

A saúde humana relacionada a região intestinal tem sido citada em diferentes estudos desde o início da década e foi abordada uma metodologia de forma consistente dentro do Projeto Microbioma Humano (Human Project Consortium, 2012), iniciado em meados de 2007, perdurando aproximadamente 10 anos e deixando um vasto conhecimento à respeito.

3.2 OBTENÇÃO DE METAGENOMAS

3.2.1 Sequenciamento

Para o reconhecimento de transcritos dentro de sequências biológicas são necessárias inicialmente informações do processo de sequenciamento. Diferentes métodos visam propostas diferentes nos dados (SLEATOR et al, 2008). No caso de métodos de sequenciamento de genoma completo, estes visam a busca de genes, OTU’s (unidade taxonômica operacional) através de loci alvos sem ou com processamento através de montadores específicos com intuito de sobrepor as leituras recorrentes da fragmentação do material genético pelos diferentes sequenciadores (LOZUPONE et al, 2013).

3.3.2 Montagem

Após processos de sequenciamentos, a bioinformática implementa diversos recursos para análise e isto inclui vários montadores de sequências com o intuito de buscar genomas ou “contigs” de organismos isolados e sequenciados

separadamente. Infelizmente quando tratamos de metagenomas, tais sequências possuem uma vasta informação genética misturada em formatos de leituras separadas, as quais variam de qualidade em efeito ao tamanho dos fragmentos . O processo de montagem de genomas isolados está presente dentro da biologia molecular a muito tempo, incluindo montagens de metagenomas. Entretanto diversos novos processos foram incorporados na montagem de metagenomas e ascendendo montadores dedicados, trazendo maiores informações à respeito das sequências (AYLING; D CLARK; LEGGETT, 2019).

3.2.3 Anotação

A busca por similaridade de sequências em banco de dados é uma das etapas mais importantes para a obtenção de genomas de maneira geral. A anotação é realizada em 3 etapas, sendo elas: Anotação à nucleotídeos; anotação à nível de proteínas; e anotação à nível de funcionalidade (para verificar função dos genes/proteínas) (PIERRI, 2017 apud STEIN, 2001). Estes processos devem ser realizados com muita atenção, pois anotações mal executadas podem arruinar metagenomas inteiros.

3.3 SRATOOLKIT

Um banco de dados biológicos importante para os estudos de metagenômica é o NCBI, que armazena de cerca de 1.213.375 informações brutas de sequenciamentos (consultado em novembro de 2019). O repositório *Sequence Read Archive* (SRA) trouxe um aumento na quantidade de dados disponíveis, se tornando uma forma de obter dados e analisá-los independentemente.

Dentre as diversas ferramentas disponíveis pelo NCBI, o SRA ToolKit fornece uma maneira fácil de adquirir sequências de leitura, permitindo a busca de dados de sequências biológicas armazenados de diversos estudos em formato “.sra” e outros formatos com capacidades maiores de compressão (ex. “.srf”, “.bam”). A possibilidade de conversão de formatos de arquivos é algo que otimiza as análises de outros pesquisadores, através da função “fastq-dump”, onde formatos nativos de sequenciadores ABI SOLiD e Illumina podem ser convertidos em formatos

convencionais como “fasta” e “fastq” ou também para menos usuais “sff” e “sam” (Bethesda (MD), 2011).

3.4 METAGENÔMICA COMPARATIVA

As análises de metagenoma são geralmente realizadas de maneira simples, por meio de comparação de sequências. Estas comparações são eficazes, porém, levando em consideração o grande volume de dados que um metagenoma contém, acabam demandando grande esforço computacional (PACCHIONI, 2010).

Uma das maneiras mais usuais de se realizar comparações entre sequencias biológicas é por meio da ferramenta BLAST (Basic Local Alignment Search Tool), que encontra similaridade entre regiões de sequencias biológicas contra o banco de dados NCBI e realiza cálculos de significância estatística entre os resultados encontrados (ALTSCHUL *et al.*, 1990).

3.4.1 Metodologias livre de alinhamento

Para lidar com os problemas associados a falta de agilidade dos métodos de alinhamento de sequências, surgiram as técnicas livre de alinhamento que estão sendo usadas com sucesso mostrando-se ser superior em muitos casos (ZIELEZINSKY *et al.*, 2017). Como exemplo destas metodologias, destacam-se KMACS (HORWEGE *et al.*, 2014), Prot-Spam (LEIMEISTER *et al.*, 2019) e Feature Frequence Profile (FFP) (SIMS *et al.,*2009).

3.4.2 K-mers

A maioria ferramentas livre de alinhamento são baseadas em k-mer, que significa a ocorrência de palavras exatas de comprimento fixo “k” em um conjunto de sequências de DNA (PIERRI, 2017 apud VINGA; ALMEIDA, 2003). O mapeamento das sequências em vetores, onde a resolução dos vetores é determinada pelo comprimento da palavra. A partir disso, uma medida de distância é aplicada entre espaços dos vetores, as quais refletem em uma matriz de distância par-a-par (*pairwise*) (ZIELEZINSKY *et al.*, 2017). Várias medidas de distância podem ser aplicadas, como Cossenos, Spearman, Pearson, porém a medida mais usual é a Euclidiana (BODEN *et al.*, 2013).

Pode-se usar ainda K-mers espaçados, que são representados em um vetor binário, a partir das frequências relativas, onde o cálculo da distância dos pares é aplicado (BODEN *et al.*, 2013; LEIMEISTER *et al.*, 2014; HORWEGE *et al.*, 2014).

3.5 PROJEÇÃO VETORIAL

Neste trabalho, a linguagem de programação utilizada para desenvolver e executar todos os algoritmos foi o MatLab®. O elemento base nesta linguagem de programação é o cálculo entre matrizes, o que permite a resolução de uma série de problemas de Bioinformática em alta performance. Para entender a aplicação do algoritmo SWeeP, que é o elemento principal deste trabalho, é necessário compreender as definições de espaço vetorial, matrizes e base ortonormal.

O espaço vetorial é um conjunto de objetos denominados vetores, que podem ser somados e multiplicados por números escalares. Existem espaços vetoriais onde a multiplicação pode ser realizada utilizando números complexos, porém os números reais são mais utilizados (SANTOS, 2010). Uma matriz é uma tabela de elementos que são dispostos em linhas e colunas, podendo seus elementos ser números, funções ou ainda outras matrizes (BOLDRINI *et al.*, 1980).

A grande contribuição do método SWeeP é a possibilidade de redução de dimensionalidade de vetores que contém grandes quantidades de dados. A projeção de vetores em uma base ortonormal (uma base ortonormal é um conjunto de vetores ortogonais) gera uma perspectiva da representação desse conjunto na base. A base é obtida pelo o produto da matriz de vetores a serem projetados. Existe ainda o conceito de base quasi-ortomormal, que é aplicado quando há necessidade de obter bases de projeção satisfatórias, levando em conta a ortogonalidade, sendo que o produto interno dos vetores da base deve ser muito pequeno, mas não necessariamente zero (PIERRI, VOYCEIK, RAITTZ *et al.*, 2019).

3.6 ANÁLISE FILOGENÉTICA

As representações filogenéticas podem assumir diferentes topologias dependendo de como os ramos se encontram ao longo da árvore, sendo que diferentes topologias podem representar diferentes eventos evolutivos, ficando a critério dos algoritmos e softwares de filogenia estimar a topologia que melhor representa o conjunto de dados (SAKAMOTO, 2016).

As ferramentas para análise filogenética são classificadas em duas categorias; Análise baseada em alinhamento e análise livre de alinhamento, podendo ser utilizadas para genomas completos ou não (JUN *et al*., 2010).

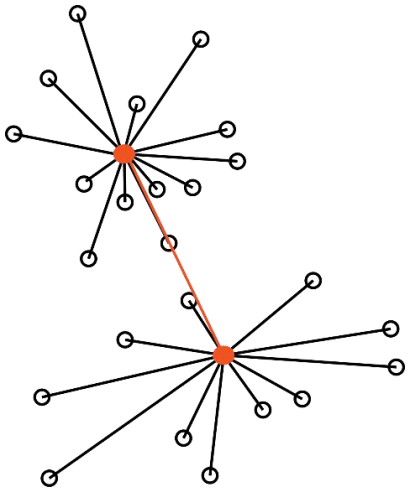
Os métodos livres de alinhamento para análise filogenética levam em consideração a matriz de distância para reconstruírem a filogenia. São exemplos destes métodos; Neighbor-joining (SAITOU, NEI, 1987), UPGMA (MICHENER; SOKAL, 1957) e Ward (WARD, 1963). Neste trabalho, as reconstruções filogenéticas serão realizadas usando o algoritmo de clusterização hierárquica Ward.

3.6.1 Método Ward

Ao criar um cluster hierárquico, além distâncias entre todos objetivos dos grupos, precisa-se calcular as distâncias entre os agrupamentos. Existem diferentes opções de como defini-las, sendo que cada opção resulta em um tipo diferente de cluster hierárquico (HOLMES, HUBER, 2019).

A Figura 2, mostra como funcionado o método Ward de maneira simples. O algoritmo Ward maximiza a soma de quadrados entre grupos, minimizando as somas de quadrados dentro de grupos.

FIGURA 2 – MODELO DO AGRUPAMENTO HIERÁRQUICO WARD



FONTE: HOLMES, HUBER (2019)

LEGENDA: A linha em vermelho representa a maximização da soma de quadrados entre os grupos. As linhas pretas, representa a minimização entre as somas de quadrados dentro dos grupos. Os grupos são representados pelos círculos.

Este método adota uma abordagem de análise de variação, em que o objetivo é minimizar a variação nos clusters, sendo muito eficiente para dados extensos, pois quebra os agrupamentos em tamanhos menores (WARD, 1963). Quando existe o conhecimento prévio de que os clusters têm tamanhos similares, utilizar o método Ward para diminuir a variação dentro da classe é a melhor opção (HOLMES, HUBER, 2019).

3.7 APRENDIZADO DE MÁQUINA

A inteligência artificial é uma ciência recente. Ela visa compreender e criar entidades inteligentes, abrangendo uma variedade de subcampos (RUSSELL, NORVIG, 2004). Os métodos e ferramentas desenvolvidos em IA são aplicados à simulação de experiências humanas para resolver os mais diversos problemas.

3.7.1 Análise de Componentes Principais

A análise de componentes principais é uma metodologia estatística que se baseia em transformar sob a mesma dimensão um conjunto de observações

variáveis em outro conjunto de observações que não são correlacionadas, denominadas de componentes principais (VARELLA, 2008).

Os componentes principais são calculados com o objetivo de obter o máximo de informação em termos da variação contida nos dados, associando concepção de redução de volume, com perda mínima de informação (ABDI, WILLIANS, 2010). Como resultado, os dados são agrupados de acordo com seu comportamento dentro da população. Basicamente, a técnica aglomera os dados de uma população de acordo com a variação de suas características (VARELLA, 2008).

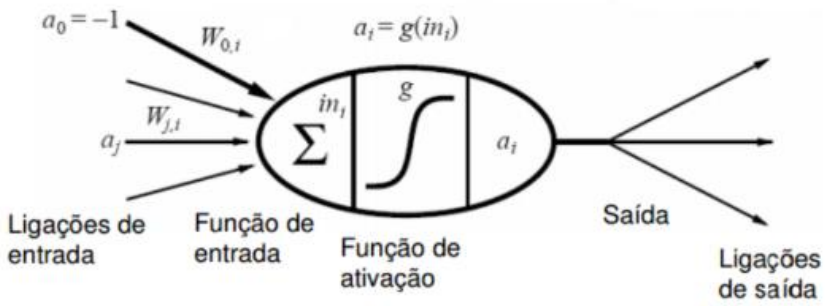
3.7.2 Redes Neuronais Artificiais

As redes neuronais artificiais (RNAs) é uma técnica baseada na modelagem de funcionamento do cérebro que compilam seus conhecimentos por meio da identificação de padrões dos dados apresentados, sendo capaz de processar uma grande quantidade de dados e realizar previsões, que muitas vezes são extremamente precisas (AGATONOVIC-KUSTRIN, BERESFORD, 2000).

Em 1943, McCullosh e Pitts criaram a lógica das RNAs, tendo como inspiração o comportamento dos neurônios, que são células que se comunicam por meio de sinapses, ou seja, pela transmissão de impulsos nervosos.

As RN’s são compostas por unidades conectadas por vínculos que possuem um peso numérico, o qual corresponde a intensidade e o sinal da conexão. Esses vínculos possuem a função de propagação da ativação que deve atender a duas metas: Primeiramente, deseja-se que a unidade seja ativa quando a entrada recebida for adequada, e inativa quando for inadequada; em segundo lugar, a ativação não pode ser linear (RUSSELL, NORVIG, 2004). Abaixo, a Figura 3 representa o modelo matemático simples do neurônio

FIGURA 3 - MODELO MATEMÁTICO SIMPLES DO NEURÔNIO PROPOSTO POR MCCULLOCH E PITTS (1943)



FONTE: Russell e Norvig (2004)

LEGENDA: A ativação é propagada de a_j de j até i. O peso do desvio $W_{j,i}$ determina o sinal da conexão, calculando primeiramente uma soma ponderada de suas entradas. A função de ativação g é aplicada à soma ponderada, derivando a saída.

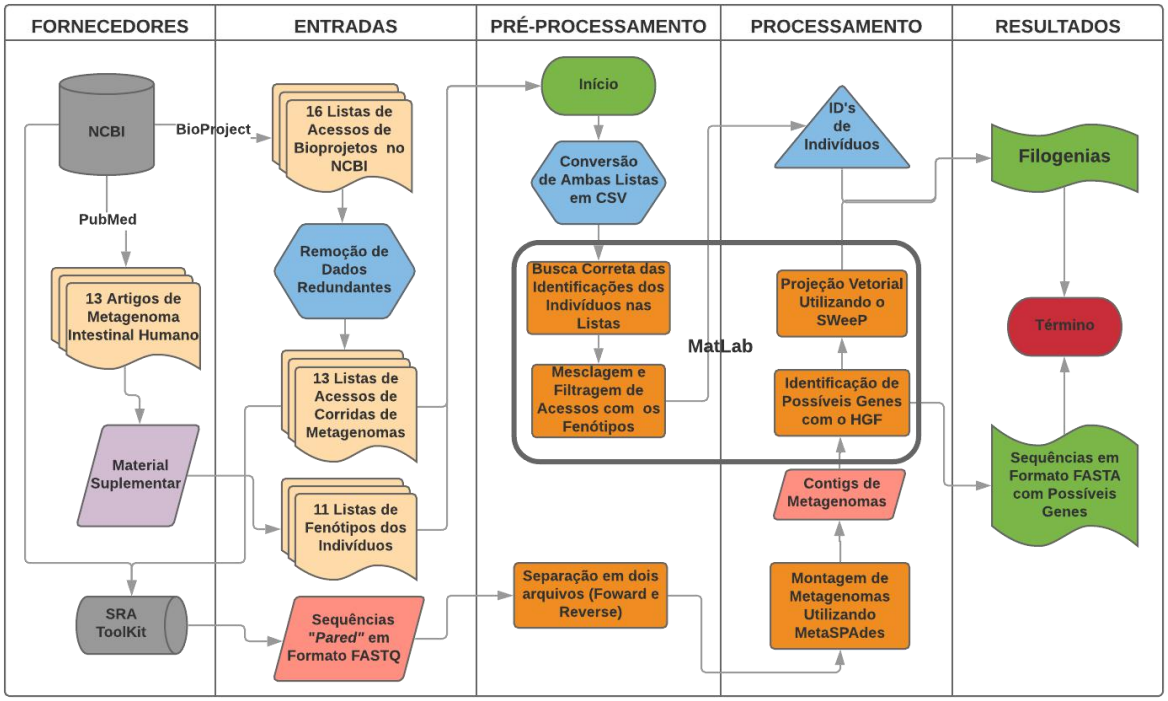
As redes neurais são organizadas em camadas, sendo de dois tipos: Redes neurais de única camada (que não possui unidades ocultas) e Redes neurais de múltipla camada (que tem uma ou mais camadas ocultas) (RUSSELL, NORVIG, 2004). Neste trabalho, utilizamos a Rede neural de múltipla camada (MLP).

A *multilayer perceptron* ou MLP, é a rede neural mais comumente utilizada. Possui capacidade de ignorar ruídos e modelar funções complexas, tendo alta capacidade de adaptação. Para o treinamento da rede, é necessário um conjunto de dados de entrada e vetores de saída associados. Durante o processo, os dados de treinamento são repetidamente apresentados ao MLP, onde os pesos na rede são ajustados até que o mapeamento de entrada e saída ocorra. Os pesos são ajustados de acordo com a magnitude e do sinal de erro, dado pela diferença entre a saída desejada e a real. Importante ressaltar que ao projetar uma MLP não se deve utilizar muitos neurônios na camada oculta. Muitos neurônios torna a rede propensa à *overfitting*, ou seja, pode levar a rede a memorizar os dados de treinamento (BASHEER, HAJMEER, 2000).

4 METODOLOGIA

Este estudo foi executado em cinco etapas principais (Figura 4), sendo elas: Etapa 1 - Obtenção de dados; Etapa 2 - Pré-processamento de dados; Etapa 3 - Aplicação de algoritmos de Inteligência Artificial; Etapa 4 – Exploração de filogenia e abordagens de Aprendizado de máquina; Etapa 5 – Análise de dados.

FIGURA 4 – FLUXOGRAMA DE PROCESSOS DA MONOGRAFIA. LEGENDA: FORNECEDORES - Etapa 1; ENTRADAS - Etapa 2; PRÉ-PROCESSAMENTO - Etapa 3; PROCESSAMENTO - Etapa 4; RESULTADOS - Etapa 5.



Fonte: O Autor (2019)

4.1 OBTENÇÃO DOS DADOS

As sequências metagenômicas foram extraídas do banco de dados NCBI, do repositório de arquivos de sequências de leituras (SRA), utilizando o *SRA Toolkit*. Com prévia seleção foram selecionadas e catalogadas de acordo com o *pipeline* do banco (Disponível em: https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/).

Utilizou-se leituras de metagenoma do microbioma intestinal humano, atendendo como critério apenas leituras WGS (*Whole Genome Shotgun*). Estas leituras passaram por uma etapa de pré-processamento, onde foram concatenados em uma única sequência, de acordo com a origem do metagenoma (nacionalidade), em formato multiFASTA.

4.2 PRÉ-PROCESSAMENTO DOS DADOS

Com intuito de reduzir a quantidade de dados e conseguir realizar inicialmente em tempo hábil com todas os estudos propostos, as sequências obtidas foram processadas com o modo de metagenomas pelo montador do pacote SPAdes 3.13.0 (BANKEVICH *et al.*, 2012) utilizando o algoritmo principal metaSPAdes disponibilizado gratuitamente.

4.2.1 Montagem utilizando o algoritmo metaSPAdes

Através da utilização de 3 k-mers diferentes, todas as sequências de múltiplas leituras obtidas anteriormente foram transformadas em sequências de múltiplos contigs. Por conta disto, as sequências de leituras seguiram um pipeline descrito pelos desenvolvedores do pacote, iniciando com a correção de erros nas leituras com a ferramenta BayesHammer seguido da montagem utilizando o metaSPAdes (NURK et al, 2017).

Para a montagem dos metagenomas, introduzimos sequências “paired-end” e utilizamos os k-mers padrões e automáticos para a seleção de contigs, sendo os tamanhos dos k-mers de 21, 33 e 55.

Correção de disparidades ao fim do pipeline com a ferramenta MismatchCorrector não foi realizada visando a preservação da identidade dos metagenomas (AYLING; CLARK; LEGGETT, 2019). Os núcleos de processamento do computador foram limitados (por padrão) em 16 e a memória RAM utilizada foi limitada a 160 GB.

Os processos foram realizados através do acesso remoto do servidor comum da equipe do laboratório de Inteligência Artificial aplicada a Bioinformática cujas

configurações combinam um processador Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz e 256 GB de memória RAM.

4.3 ALGORITMOS DE INTELIGÊNCIA ARTIFICIAL

Após o pré-processamento e montagem dos dados, os possíveis genes nas sequências de metagenomas foram identificados pela ferramenta HGF (*Hybrid Gene Finder*), para ser realizado o estudo dos genes presentes nas montagens das sequências. Em seguida, estas sequências foram vetorizadas utilizando o SWeeP, com o intuito de agilizar e facilitar as análises dos metagenomas.

As ferramentas HGF (ainda não submetida) e SWeeP (RAITTZ, 2019 - em revisão) foram desenvolvidas pelo Laboratório de Inteligência Artificial Aplicada à Bioinformática e suas respectivas publicações estão em fase de revisão no momento.

4.2.1 HGF

O HGF é uma ferramenta de anotação genômica que realiza a busca por ORFs (Open Reading Frames) em sequências biológicas com o intuito de identificá-las e facilitar a análise de dados genômicos. A ferramenta foi utilizada através da entrada de dados metagenômicos (sequências de nucleotídeos) em forma de diversos contigs (montados) em um único arquivo por amostra, além da rede neural treinada para identificação de possíveis genes (inclusa da ferramenta), o qual foi gentilmente disponibilizada pelo Profº Drº Roberto Tadeu Raittz.

4.2.2 SWeeP

SWeeP é um modelo de representação de sequências de proteínas em vetores de tamanho relativamente baixo, que preserva as informações para comparação, tornando mais ágeis as análises de sequências biológicas (PIERRI, VOYCEIK, RAITTZ *et al.*, 2019). Esta abordagem utiliza o conceito de palavras espaçadas (BODEN; SCHÖNEICH; HORWEGE, 2013; LEIMEISTER *et al.*, 2014) para varrer sequências que geram índices para criar um vetor de alta dimensão contendo uma rica representação de dados, que é projetado para uma dimensão muito menor, de acordo com o lema Johnson-Lindenstrauss (JONSON, LINDENSTRAUSS, 1986).

A entrada para o SWeeP consiste em um arquivo multiFASTA contendo seqüências de aminoácidos ou nucleotídeos. Neste trabalho, todas as leituras para cada indivíduo foram concatenadas formando uma única sequência para representar o metagenoma como um todo. Cada metagenoma é representado por uma matriz bidimensional, que quando representada em colunas, é um vetor que reflete dados de sequência altamente representativos.

4.4 EXPLORAÇÃO DA FILOGENIA E REDE NEURONAL ARTIFICIAL

Para gerar as árvores filogenéticas contendo as amostras selecionadas foi utilizado distância euclidiana como parâmetro para o cálculo dos ramos e a filogenia foi baseada no método Ward. As matrizes utilizadas para as comparações foram projeções de 1600 coordenadas.

Para o treinamento de cada rede neural artificial foi utilizado 5 camadas internas e 1 saída, totalizando no treinamento de 7 redes para cada indivíduo. Foi repetido o treinamento das redes 50 vezes, onde aleatoriamente foram reordenados com o intuito de separar o conjunto treino (71%) e o conjunto teste (29%) uniformemente. Uma coluna identificando o indivíduo de forma binária era introduzida na sequência da coluna final da matriz de projeção, gerando as redes com uma matriz de 70x1601 como entrada.

4.5 ANÁLISE DOS RESULTADOS

Após o pré-processamento e seleção das amostras, obtivemos matrizes de distância correspondentes a projeções geradas pelo SWeeP, as quais foram analisadas quanto a performance e dessa forma utilizadas para averiguar árvores filogenéticas e posteriormente para o treinamento de redes neurais com o intuito de obter a qualidade da projeção nas amostras que correspondem aos respectivos indivíduos. Assim conseguimos obter valores significativos nas redes para a validação

5 RESULTADOS E DISCUSSÃO

5.1 FILTRAGEM E DOS DADOS

Em princípio a busca de dados foi iniciada seguindo uma “query” pelo NCBI, no repositório SRA. Através de palavras-chave como “metagenoma”, “human-gut”, “feacal sample” embora sem muita clareza nos dados encontrados. Após algumas buscas com diferentes combinações de identificadores, a observação mais recorrente foi de má organização de dados e planilhas, por vezes desalinhadas, senão com informações incoerentes (FIGURA 5).

FIGURA 5 – DADOS DO REPOSITÓRIO SRA

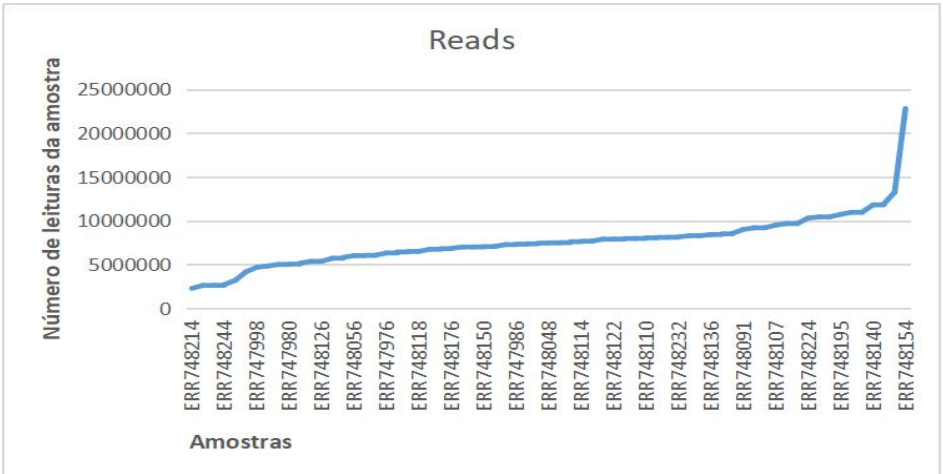
LibrarySource	MBases	MBytes	Organism
GENOMIC	1442	838	human gut metagenome
GENOMIC	1315	882	human gut metagenome
GENOMIC	468	295	human gut metagenome
GENOMIC	311	247	human gut metagenome

LibrarySource	MBases	MBytes	Organism
METAGENOMIC	50	29	Homo sapiens
METAGENOMIC	115	68	Homo sapiens
METAGENOMIC	58	34	Homo sapiens
METAGENOMIC	53	31	Homo sapiens

Fonte: NCBI, 2019.

Por este motivo o estudo de Huttenhower (2017) foi analisado como início de seleção por estudos com o mesmo alvo de metagenomas (intestinais humanos).

GRÁFICO 1 QUANTIDADE DE LEITURAS EM CADA AMOSTRA



Fonte: O Autor, 2019

O filtro em comum dentre o total de 13 estudos selecionados para busca das sequências, baseou-se na estratégia WGS de sequenciamento, para obter sequências além de amplificações de um único gene, sequências “Paired-End” que possuem informações de ambos sentidos de leitura, visando, além da melhor qualidade de dados, entradas coerentes na etapa de montagem e tudo isso dentre estudos de metagenoma de intestino humano.

TABELA 1. PROJETOS UTILIZADOS PARA A GLOBALIZAÇÃO DOS DADOS DE METAGENOMAS DE MICROBIOMA HUMANO.

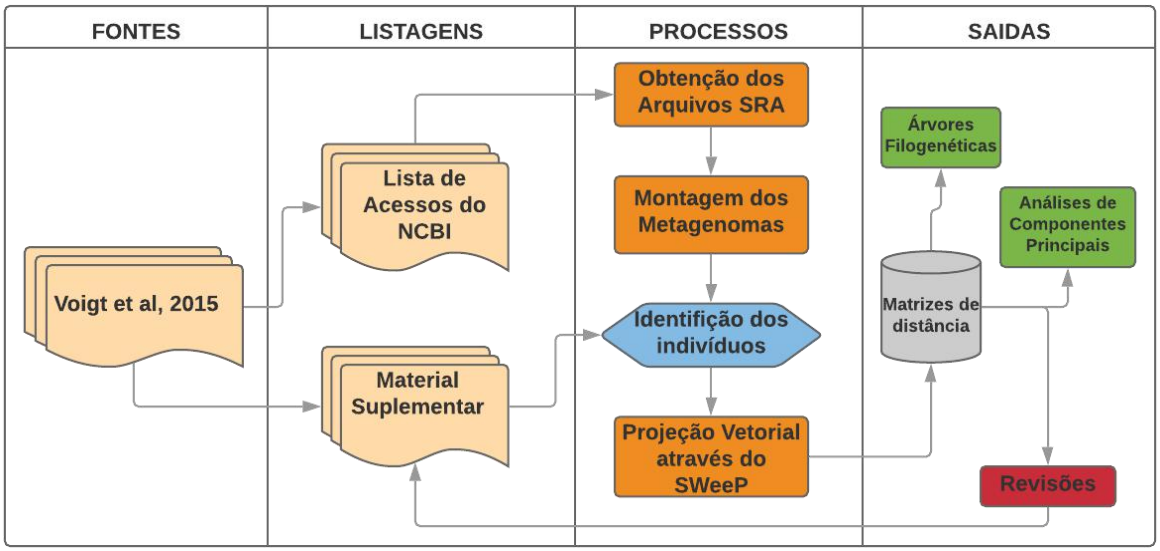
BioProject	Acesso SRA	Países	Enfoque	Indivíduos	Biblioteca	Refer.
PRJDB3601	DRP003048	JAP	Diversidade Global	106	Random	30
PRJEB1220	ERP002061	DIN	Binning	318	Other	29
PRJEB17632	ERP019502	ALE KAZ	Sub-espécies	283	-	10
PRJEB2054	ERP000108	DIN ESP	Catálogo de genes	124	Random	38
PRJEB4336	ERP003612	DIN	Obesidade	292	Other	20
PRJEB5224	ERP004605	DIN ESP CHI	Catálogo de Genes	249	Other	24
PRJEB6070	ERP005534	FRA ALE	Câncer Colorretal	38	Random	57
PRJEB6337	ERP005860	CHI	Cirrose	237	S.F.	40
PRJEB7774	ERP008729	AUS	Câncer Colorretal	156	Other	12
PRJEB8347	ERP009422	ALE	Avaliação Temporal	7	Random	54
PRJNA422434	SRP008047	CHI	Diabetes	368	Random	39

Legenda: ALE=Alemanha; AUS=Austria; CHI=China; DIN=Dinamarca; ESP=Espanha; FRA=França; JAP=Japão; KAZ=Cazaquistão. Biblioteca: Random= Aleatório; Other= Outros; S.F.: Size Fractionation. Em vermelho o estudo selecionado para as análises de amostras.

A partir dos estudos citados acima, foi obtido informações fenotípicas de um total de 2103 indivíduos distintos nos respectivos materiais suplementares de 11 artigos (TABELA 1). A compilação foi realizada de 22 planilhas, 11 obtidas através do NCBI e 11 obtidas dos materiais suplementares dos estudos, contendo códigos de acesso correspondentes às sequências e amostras disponíveis e as identificações fenotípicas correspondentes nos estudos, respectivamente.

Através dos mesmos materiais suplementares foram retirados informações comuns a todos os estudos, incluindo os seguintes fatores: sexo; idade; índice de massa corporal (IMC) e estado de saúde da pessoa (com diabetes, câncer, cirrose ou outro estado que desconfigura uma pessoa saudável). Embora as análises tenham se iniciado no agrupamento dos 13 estudos como citado, o trabalho de análise revelou uma demanda de tempo maior, e por este motivo, recorremos ao processamento e análise de dados de um único estudo. Além da demanda de tempo, por precaução nos possíveis erros de parametrizações dentre os dados de todos os estudos e com a intenção exploratória dos métodos sob metagenomas, a escolha foi única e com componentes variáveis controlados. O esquema de análises do estudo escolhido é mostrado na Figura 6.

FIGURA 6 - ESQUEMA SIMPLIFICADO DA ANÁLISE DO ESTUDO ESCOLHIDO



FONTE: O Autor, 2019

NOTA: Todos os processos presentes no fluxograma da FIGURA 4 foram realizados igualmente.

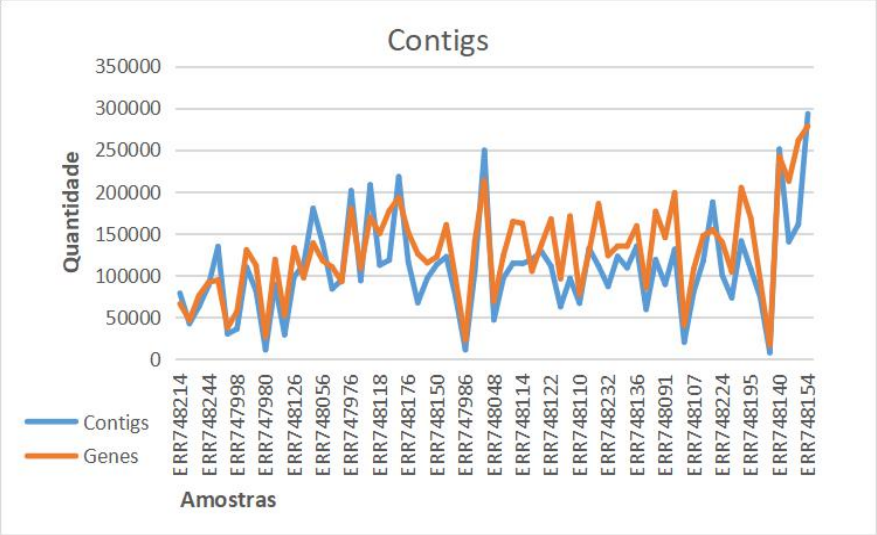
O estudo escolhido, portanto, para realizar as identificações e comparações entre projeções vetoriais é uma avaliação temporal de 7 indivíduos durante períodos “mensais” e “semanais” ao longo de um total máximo de 2 anos (VOIGT et al, 2015).

Dentro deste estudo 285 sequências foram retiradas das listas de acesso do NCBI, correspondendo a 70 diferentes amostras (GRÁFICO 1) (diferenciando no método de preservação, a contagem de dias e os indivíduos), as quais foram caracterizadas através do material suplementar do mesmo estudo (FIGURA 2).

5.2 MONTAGEM E ANOTAÇÃO DE ORF´S

O processo de montagem com o algoritmo metaSPAdes seguiu os parâmetros estabelecidos na metodologia, resultando diferentemente para cada arquivo de sequência de leitura “.fastq”, um arquivo correspondente com vários contigs. Dentre as 70 amostras houve diferenças na quantidade de leituras presentes por amostra (GRÁFICO 2). A relação do tempo de montagem com o tempo do processo seguinte (HGF) também está representada no Gráfico 3.

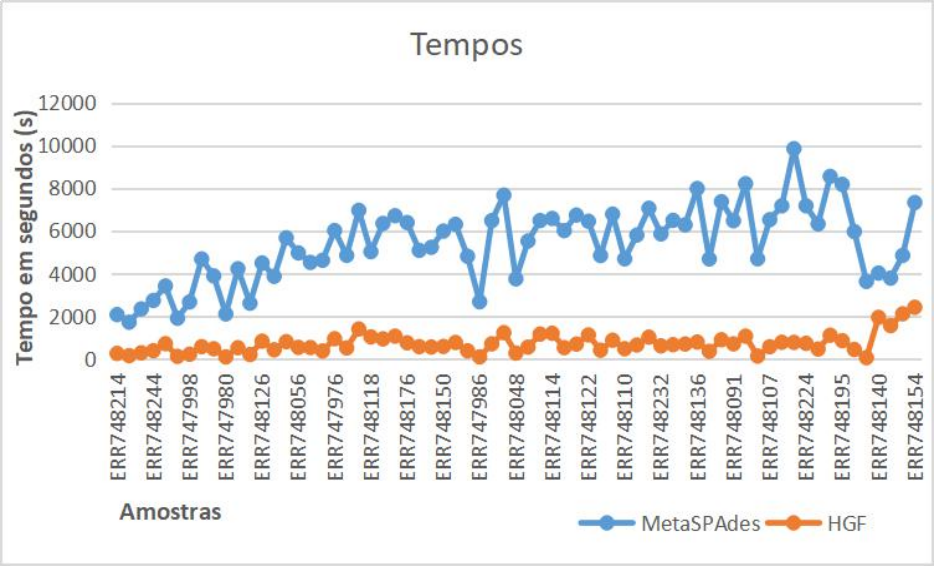
GRÁFICO 2 – QUANTIDADE DE LEITURAS NAS 70 AMOSTRAS



Fonte: O Autor, 2019

As 70 sequências contendo contigs decorrentes das montagens (múltiplos contigs em um único arquivo .fasta) foram processadas pelo algoritmo HGF. Para isto foi utilizado linguagem MatLab R2017b ® para criar rotinas e tornar possível leituras e saídas de forma particionada.

GRÁFICO 3 – TEMPO DE MONTAGEM DAS AMOSTRAS



Fonte: O Autor, 2019

O algoritmo em si, concatenou os múltiplos contigs estabelecendo um espaçamento de 21 aminoácidos na separação destes. Assim as ORF's encontradas não seriam identificadas com início em um contig e término em outro, evitando a possibilidade da formação de quimeras entre contigs de organismos distintos. O resultado final desta etapa foram 70 sequências, com suas respectivas ORF'S convertidas em aminoácidos e agrupadas novamente em um único arquivo, agora com múltiplas sequências de possíveis genes (FIGURA 7).

5.3 PROJEÇÃO VETORIAL UTILIZANDO O MODELO SWeeP

Seguido da etapa de busca de ORF's, foram realizadas diversas projeções vetoriais, utilizando a ferramenta SWeeP, com 70 amostras (ANEXO I) em 4 formatos de máscara distintas, sendo a primeira projeção em 100 coordenadas e a última com 1600, criadas em intervalos equidistantes de 100.

FIGURA 7 – ORF´s CONVERTIDAS EM AMINOÁCIDOS

Fonte: O autor, 2019

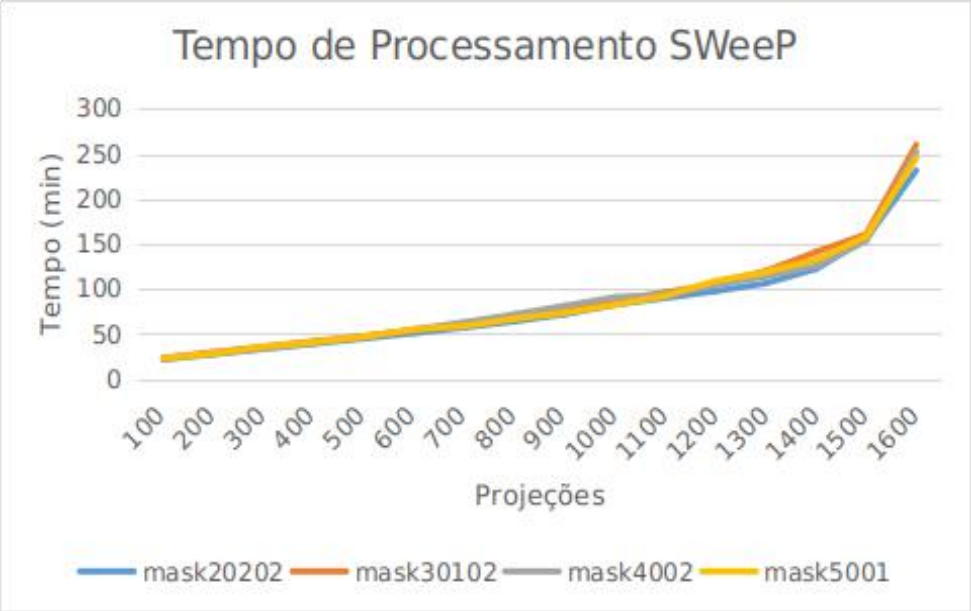
LEGENDA: Sequência de entrada (contigs) na parte superior da imagem, convertidas em possíveis genes pelo algoritmo HGF na parte inferior.

As máscaras citadas acima, dizem respeito a janela deslizante utilizada no SWeeP, as quais pela maior representatividade e por limitação computacional foram enquadradas em 6 pb, variando exclusivamente na localização dos espaços sem informações.

5.3.1 Teste de performance

Os tempo de processamento do SWeeP para cada projeção, de acordo com as máscaras, estão representadas abaixo (GRÁFICO 4).

GRÁFICO 4 – AVALIAÇÃO DE PERFORMANCE DO SWeeP



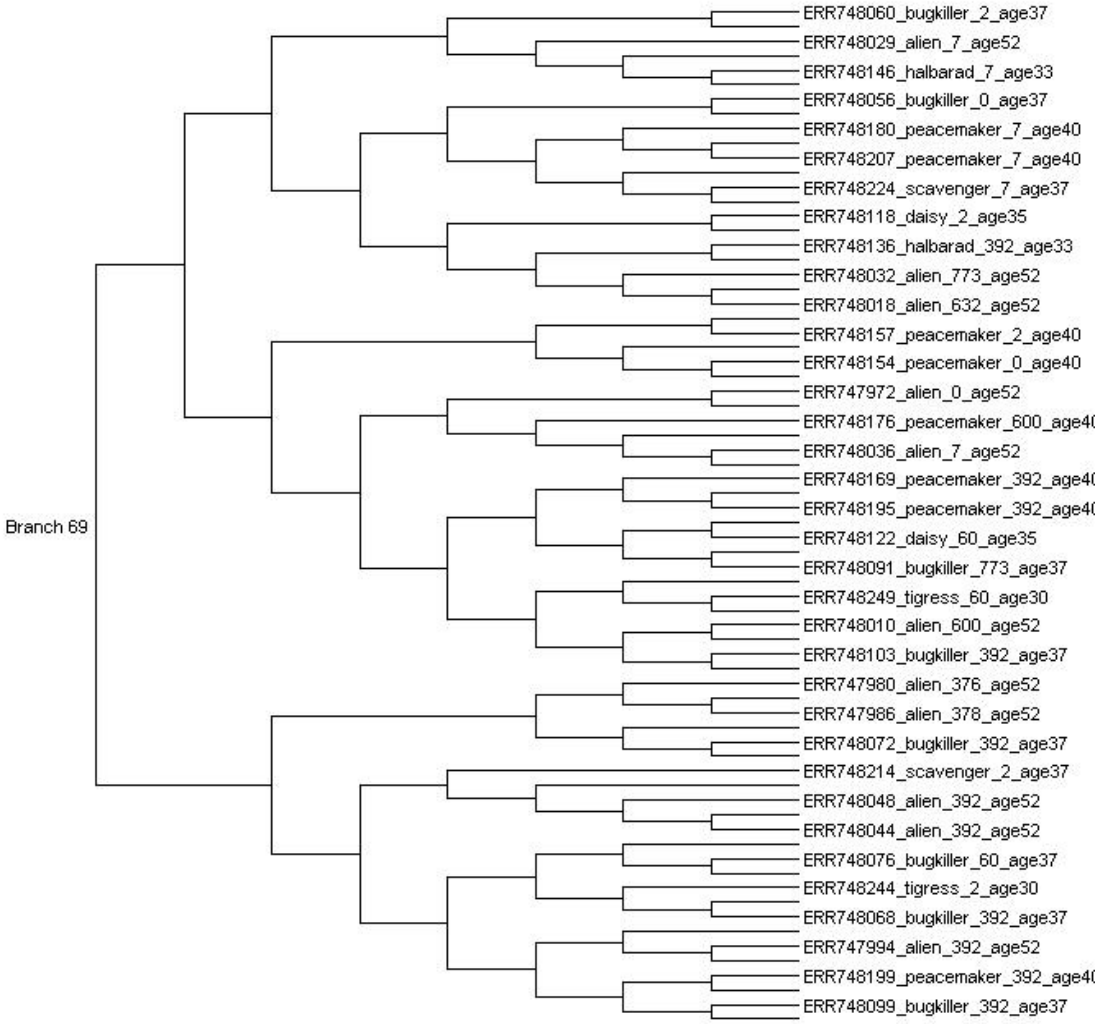
Fonte: O Autor, 2019

5.4 ANÁLISE FILOGENÉTICA

Através dos resultados de cada projeção obtivemos uma matriz de distância e uma árvore filogenética, construída pelo método Ward (WARD, 1963), resultado da seleção de uma matriz de distância em um total de 60 matrizes, sendo elas de projeções e máscaras distintas, que incluem um total de 70 amostras.

Tendo em vista as metodologias distintas entre o estudo selecionado e a metodologia que utilizamos, para a busca de correlação nas dendrologias, através da comparação visual de ambas figuras (FIGURA 8 e 9) ficou evidente que algumas características na árvore, obtidas da projeção vetorial, foram suficientes para identificar amostras de um mesmo indivíduo.

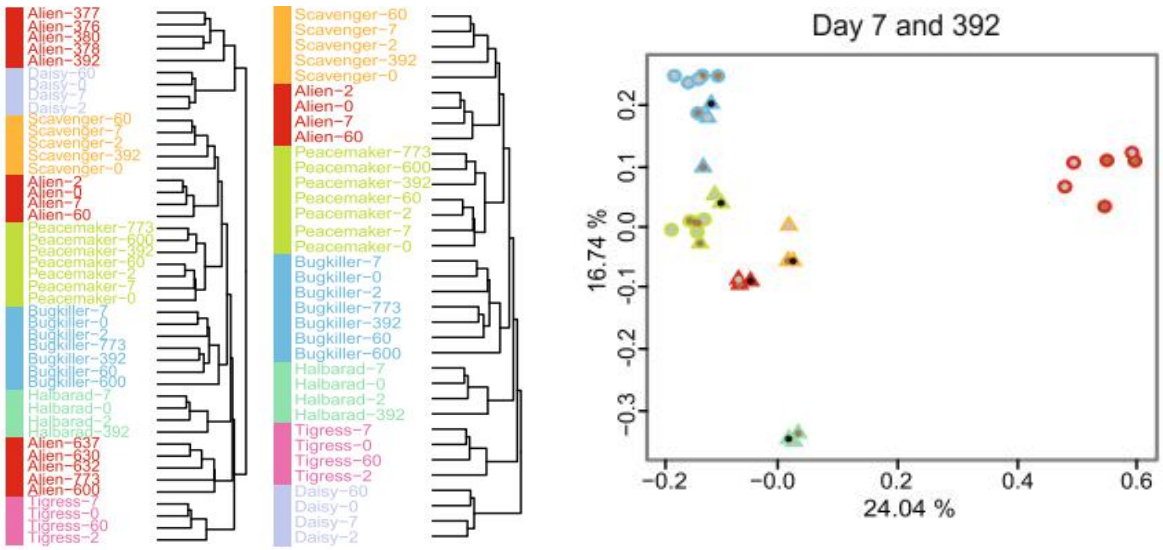
FIGURA 8. ÁRVORE FILOGENÉTICA OBTIDA DE UMA PROJEÇÃO VETORIAL DE 1600 COORDENADAS, COM MÁSCARA TAMANHO 6 E UTILIZANDO-SE DE 70 AMOSTRAS



Fonte: O autor (2019)

Entretanto o agrupamento do conjunto todo das amostras dos indivíduos, tornou a análise um pouco inconsistente, para que a filogenia fosse mais contundente.

FIGURA 9 - DENDROLOGIA E ANÁLISE DE COMPONENTES PRINCIPAIS REALIZADOS NO ESTUDO ESCOLHIDO



FONTE: Adaptado de Voigt *et al*., (2015)

LEGENDA: Cada cor representando um indivíduo diferente (n=7), sendo os métodos de preservação das amostras representados pelos formatos diferentes.

5.5 ABORDAGENS DE APRENDIZADO DE MÁQUINA

Foi realizado uma rotina com sucesso e obtido os resultados do treinamento das rede neurais e em seguida 50 repetições foram realizadas e por motivos de falhas estruturais nos resultados obtidos, somente uma delas foi retirada para calcular a média dos valores (TABELA 2). A correlação Pearson do teste obtida pela rede neural artificial foi de 0,96 no conjunto treino e 0,83 no conjunto teste.

TABELA 2. Média de 2 testes utilizando rede neural artificial nos conjuntos de treino (n=50) e teste (n=20).

	Acurácia	Precisão	Sensibilidade	F-Score	Correlação	P-value
TREINO	98,71%	96,29%	95,57%	95,58%	0,96	0,00
TESTE	96,43%	83,93%	82,74%	82,50%	0,83	0,00

Fonte: O Autor, 2019

6 CONSIDERAÇÕES FINAIS

Através dos resultados da RNA, conseguimos responder a questão norteadora, representando cada indivíduo dentro de cada sequência de metagenoma, que por conjuntura, conseguimos revelar características incluídas nas projeções vetoriais de metagenomas, onde através de IA foi possível categorizar as amostras indivíduos corretamente. Portanto, conseguimos efetivamente reduzir a dimensionalidade dos dados, realizar análises relativamente rápidas e identificar, com uma acurácia e correlação alta, cada indivíduo independente do tempo amostral.

Com este trabalho, estabelecemos alguns parâmetros para manipular os dados de metagenomas WGS e desta maneira conseguimos entender as necessidades e relevâncias dadas. A monografia foi um passo inicial para obter conhecimentos múltiplos, de biologia molecular, à respeito de sequenciadores, bancos de dados e principalmente programação e maior entendimento da aplicação de conceitos estatísticos.

6.1 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Dito isto, logo em seguida pretendemos aumentar a quantidade de amostras utilizadas para treinar a rede e verificar novas projeções para gerar árvores filogenéticas mais congruentes. A ideia principal mais a frente é dar prosseguimento com o projeto em um mestrado na bioinformática para recolher mais análises sob manipulação de dados públicos de metagenomas, trazendo novos resultados com intuito de aprimorar e estender a metodologia e o fluxograma para uma quantidade maior de dados e de ambientes.

REFERÊNCIAS

1. ABDI, H.; WILLIAMS, L.J. Principal component analysis. **Computational Statistics**, v. 2, p. 433-459, 2010.

2. AGATONOVIC-KUSTRIN, S; BERESFORD, R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. **Journal of Pharmaceutical and Biomedical Analysis**, V. 22(5), p. 717-727, 2000.

3. ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of molecular biology**, v. 215, n. 3, p. 403–10, 1990

4. AYLING, M., CLARK, M. D. & LEGGETT, R. M. New approaches for metagenome assembly with short reads. Briefings in Bioinformatics. bbz020. 2019.

5. BANKEVICH, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. ***J Comput Biol****.* v.19 p.455–477, 2012.

6. BASHEER I. A., HAJMEER M. Artificial neural networks: Fundamentals, computing, design, and application. **J. Microbiol. Meth**. v.43, p.3–31, 2000.

7. BETHESDA (MD): National Center for Biotechnology Information (US). **SRA Knowledge Base [Internet]**. EUA: NCBI, 2011. E-book. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK56551/>. Acesso em: 30 out. 2019.

8. BODEN, M.; SCHÖNEICH, M.; HORWEGE, S. Alignment-free sequence comparison with spaced k-mers. **German Conference on Bioinformatics** 2013, v. 34, p. 24–34, 2013.

9. BOLDRINI, J. L. *et al.* **Álgebra linear**. Harper & Row: Rio de Janiero, 1980.

10. COSTEA P. *et al.* Subspecies in the global human gut microbiome. **Molecular Systems Biology**. V. 13:960 p. 1-11. Novembro. 2017.

11. ERICKSON A. R. *et al.* Integrated Metagenomics/Metaproteomics Reveals Human Host-Microbiota Signatures of Crohn's Disease. **PLoS One**. V. 7:11, p. 1:14. Novembro. 2012.

12. FENG, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. **Nature Communications**. V. 6:6528, p. 1-13. Março. 2015.

13. HANDELSMAN, Jo. Metagenomics: application of genomics to uncultured microorganisms. **Microbiol. Mol. Biol**. V. 68, n. 4, p. 669-685. 2004.

14. HOLMES, S., HUBER, W. **Modern Statistics for Biology**. DeNBI: German network for Bioinformatics infrastructure, Alemanha, 2019.

15. HORWEGE, S. *et al.* Spaced words and kmacs: Fast alignment-free sequence comparison based on inexact word matches. ***Nucleic Acids Res****.,* v. **42**, p.7–11 2014.

16. HUTTENHOWER C., GEVERS D., KNIGHT, R. *et al.* Structure, function and diversity of the healthy human microbiome. **Nature**. V. 486, p. 207–214. Junho. 2012.

17. JOHNSON, W. B.; LINDENSTRAUSS, J. Extensions of Lipschitz mappings into a Hilbert space. **Contemp. Math***.,* v. 26, p. 189–206, 1984.

18. JUN, S.R.; SIMS, G. E.; W.U., G. A; KIM, S.H. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. **Proceedings of the National Academy of Sciences of the United States of America**, v. 107, n. 1, p. 133–138, 2010.

19. KARLSSON F. H., TREMAROLI V., NOOKAEW I., BERGTRÖM G., BEHRE C. J., FAGERBERG B., NIELSEN J. & BÄCKHED F. Gut metagenome in European women with normal, impaired and diabetic glucose control. **Nature**. V. 498, p. 99-103. Junho. 2013.

20. LE CHATELIER, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. **Nature**. V. 500 p. 541-546. Agosto. 2013.

21. LEIMEISTER, C. A. *et al.* Prot-SpaM: fast alignment-free phylogeny reconstruction based on whole-proteome sequences. ***Gigascience*** v. 8, p. 1–14, 2018.

22. LEIMEISTER, C. A.; BODEN, M.; HORWEGE, S.; LINDNER, S.; MORGENSTERN, B. Fast alignment-free sequence comparison using spacedword frequencies. **Bioinformatics**, v. 30, n. 14, p. 1991–1999, 2014.

23. LEINONEN, R., SUGAWARA, H. & SHUMWAY, M. The Sequence Read Archive. **Nucleic Acids Research.** V. 39, p. D19–D21. January. 2011.

24. LI, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. **Nature Biotechnology**. V.32:8, p. 834-841. Agosto. 2014.

25. LLOYD-PRICE, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. **Nature**. V. 550, p. 61-66. Outubro. 2017.

26. LOPETUSO L. R., PETITO V., GRAZIANI C., SCHIAVONI E. *et al.* Gut Microbiota in Health, Diverticular Disease, Irritable Bowel Syndrome, and Inflammatory Bowel Diseases: Time for Microbial Marker of Gastrointestinal Disorders. **Dig Dis**. V. 36, n. 1, p. 56-65. 2018.

27. McCULLOCH, W.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, v.5, p.115-133, 1943.

28. MICHENER, D., SOKAL, R.R. A quantitative approach to a problem of classification. **Evolution**, v. 11(2), p. 130-162, 1957.

29. NIELSEN H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. **Nature Biotechnology**. V. 32:8, p. 822-828. Julho. 2014.

30. NISHIJIMA, S., SUDA, W., OSHIMA K., KIM S., HIROSE Y., MORITA H. & HATTORI M. The gut microbiome of healthy Japanese and its microbial and functional uniqueness. **DNA Research**. P. 1-9. Março. 2016.

31. NURK, S., MELESHKO, D., KOROBEYNIKOV, A. & PEVZNER, P. A. metaSPAdes: a new versatile metagenomic assembler. **Genome Research**. V. 27 p. 824–834. 2017.

32. MA, J., PRINCE, A. & AAGAARD, K. M. Use of whole genome shotgun metagenomics: a practical guide for the microbiome-minded physician scientist. In: Seminars in reproductive medicine. **Thieme Medical Publishers**. p. 005-013. 2014.

33. OLIVEIRA, V. M; SETTE, L. D. FANTINATTI-GARBOGGINI, F. Preservação e Prospecção de Recursos Microbianos. **Multiciência**. Universidade Estadual de Campinas, 19p. 2006.

34. OLSEN, L. R. BioReader: a text mining tool forperforming classification of biomedical literature. **BMC Bioinformatics**. V. 19(Suppl 13):57, p. 165-170. 2019.

35. PACCHIONE, R. G. Metagenômica comparativa de solo de regiões da mata Atlântica e Caatinga do Estado do Rio Grande do Norte - Brasil. Dissertação de mestrado. Universidade do Rio Grande do Norte, 111p. Natal, 2010.

36. PIERRI, C.R. Representações vetoriais de proteomas: Um estudo de caso com sequências mitocondriais. **Dissertação de mestrado**. Universidade Federal do Paraná, Curitiba, 2017.

37. PIERRI, C.R.; VOYCEIK, R.; RAITTZ, R.T. *et al.* SWeeP: representing large biological sequences datasets in compact vectors. **Scientific Reports**, 2019 (em revisão).

38. QIN J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. **Nature**. V. 464 p. 59-65. Março. 2010.

39. QIN J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. **Nature**. V. 490, p. 55-60. Outubro. 2012.

40. QIN N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. **Nature**. V.513, p. 59-64. Setembro, 2014.

41. RUSSEL, S.; NORVIG, P. **Artificial Inteligence: A Modern Approach**. 3ed. Prentice hall: Rio de Janeiro, 2004.

42. SAITOU N, NEI M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Molecular Biology and Evolution**, v.4(4), p.406–425, 1987.

43. SANTOS, R. J. **Geometria Analítica e Álgebra Linear**. Universidade Federal de Minas Gerais, 2010.

44. SAKAMOTO, T. Ferramentas para análise filogenética e de distribuição taxonômica de genes ortólogos. **Tese de doutorado**. Universidade Federal de Minas Gerais: Belo horizonte. 2016.

45. SIEGEL R. L., TORRE L. A., SOERJOMATARAM I. *et al.* Global patterns and trends in colorectal cancer incidence in young adults. **Gut**. V. 0 p. 1-7. Setembro. 2019.

46. SIMON, C., DAVIDSEN, K., HANSEN, C. *et al.* BioReader: a text mining tool for performing classification of biomedical literature. **BMC Bioinformatics.** V.19, n. 13 p. 57. 2019.

47. SIMS, G. E., JUN, S.-R., WU, G. A., KIM, S.H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. **Proc. Natl. Acad. Sci. U. S. A.** V.106, p. 2677–2682. 2009.

48. SLEATOR, R. D.; SHORTALL, C.; HILL, C. Metagenomics. **Letters in applied microbiology.** V. 47, n. 5, p. 361-366, 2008.

49. STEIN, L. Genome annotation: from sequence to biology**. Nature reviews.** Genetics, v. 2, n. 7, p. 493–503, 2001.

50. TURNBAUGH J., Peter *et al.* The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. **Science translational medicine**. V. 1, n. 6, p. 6ra14-6ra14. 2009.

51. TYSON, G. W. et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. **Nature.** V. 428, n. 6978, p. 37. 2004.

52. VARELLA, C. A. A. Análise Multivariada Aplicada as Ciências Agrárias: Análise de Componentes Principais. Universidade Federal Rural do Rio de Janeiro, 12p. Seropédica. 2008.

53. VINGA, S. Editorial: Alignment-free methods in computational biology. ***Brief. Bioinform.*** v 15, p. 341–342 ,2014.

54. VOIGT, A., COSTEA P., KULTIMA J. R., LI S. S, ZELLER G., SUNAGAWA S. & BORK P. Temporal and technical variability of human gut metagenomes. **Genome Biology**. V. 16:73, p. 1-12. Abril. 2015.

55. WARD, J. H. Hierarchical grouping to optimize an objective function. J**. Am. Stat. Assoc**. v.58, p.236–244, 1963.

56. YU J., FENG Q., WONG S. H. *et al.* Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. **Gut**. V. 66 p. 70–78. Setembro. 2015.

57. ZELLER G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. **Molecular Systems Biology**. V. 10:766 p. 1-18. Novembro. 2014.

58. ZIELEZINSKI, A., VINGA, S., ALMEIDA, J., KARLOWSKI, W. M. Alignment-free sequence comparison: Benefits, applications, and tools. **Genome Biol**., v. 18, p. 1–17, 2017.

ANEXO 1 – LISTA DE ACESSOS DAS 70 AMOSTRAS UTILIZADAS PARA A MONTAGEM E PROJEÇÃO VETORIAL

Acesso	ID Indivíduo	Dias da Amostragem	MBytes	Sexo	IMC
ERR747972	alien	0	719	male	26,3
ERR747976	alien	2	664	male	26,3
ERR748026	alien	7	1585	male	26,3
ERR748036	alien	7	978	male	26,3
ERR748052	alien	7	1132	male	26,3
ERR748006	alien	60	792	male	26,3
ERR747980	alien	376	576	male	26,3
ERR747984	alien	377	1284	male	26,3
ERR747986	alien	378	831	male	26,3
ERR747990	alien	380	934	male	26,3
ERR747994	alien	392	1006	male	26,3
ERR747998	alien	392	529	male	26,3
ERR748002	alien	392	910	male	26,3
ERR748040	alien	392	1208	male	26,3
ERR748044	alien	392	610	male	26,3
ERR748048	alien	392	857	male	26,3
ERR748010	alien	600	856	male	26,3
ERR748014	alien	630	724	male	26,3
ERR748018	alien	632	767	male	26,3
ERR748022	alien	637	893	male	26,3
ERR748032	alien	773	1089	male	26,3
ERR748056	bugkiller	0	635	male	23,3
ERR748060	bugkiller	2	836	male	23,3
ERR748085	bugkiller	7	1798	male	23,3
ERR748095	bugkiller	7	1270	male	23,3
ERR748110	bugkiller	7	956	male	23,3
ERR748076	bugkiller	60	610	male	23,3
ERR748064	bugkiller	392	844	male	23,3
ERR748068	bugkiller	392	601	male	23,3
ERR748072	bugkiller	392	482	male	23,3

ERR748099	bugkiller	392	851	male	23,3
ERR748103	bugkiller	392	1387	male	23,3
ERR748107	bugkiller	392	1095	male	23,3
ERR748080	bugkiller	600	251	male	23,3
ERR748091	bugkiller	773	1005	male	23,3
ERR748114	daisy	0	844	female	19,4
ERR748118	daisy	2	746	female	19,4
ERR748126	daisy	7	615	female	19,4
ERR748122	daisy	60	904	female	19,4
ERR748130	halbarad	0	957	male	31,2
ERR748134	halbarad	2	1270	male	31,2
ERR748140	halbarad	7	1336	male	31,2
ERR748146	halbarad	7	916	male	31,2
ERR748150	halbarad	7	812	male	31,2
ERR748136	halbarad	392	852	male	31,2
ERR748154	peacemaker	0	2463	male	25
ERR748157	peacemaker	2	1534	male	25
ERR748180	peacemaker	7	964	male	25
ERR748191	peacemaker	7	810	male	25
ERR748207	peacemaker	7	750	male	25
ERR748173	peacemaker	60	1496	male	25
ERR748161	peacemaker	392	989	male	25
ERR748165	peacemaker	392	300	male	25
ERR748169	peacemaker	392	807	male	25
ERR748195	peacemaker	392	1231	male	25
ERR748199	peacemaker	392	577	male	25
ERR748203	peacemaker	392	592	male	25
ERR748176	peacemaker	600	746	male	25
ERR748187	peacemaker	773	1156	male	25
ERR748211	scavenger	0	2152	male	21,1
ERR748214	scavenger	2	212	male	21,1
ERR748224	scavenger	7	1191	male	21,1
ERR748232	scavenger	7	951	male	21,1
ERR748236	scavenger	7	705	male	21,1
ERR748221	scavenger	60	2130	male	21,1

ERR748217	scavenger	392	557	male	21,1
ERR748240	tigress	0	1058	female	19,8
ERR748244	tigress	2	307	female	19,8
ERR748253	tigress	7	734	female	19,8
ERR748249	tigress	60	873	female	19,8