Tahila Andrighetti

# Qualificação de Doutorado: Aprendizado de máquina para análises funcionais e taxonômicas de dados de metagenômica

Universidade Estadual Paulista "Júlio de Mesquita Filho"

Câmpus de Botucatu

Instituto de Biociências

Orientador: Prof. Dr. Ney Lemke

Botucatu - SP

Outubro de 2017

# Resumo

Estudos de metagenômica têm evidenciado a importância da composição das comunidades microbianas nos mais diversos ambientes. A partir da metagenômica é possível sequenciar e analisar o material genético de uma comunidade microbiana sem a necessidade de cultura dos micro-organismos. Uma vez que 99% dos micro-organismos não são cultiváveis, a metagenômica se tornou a metodologia padrão para investigar a dinâmica e composição das comunidades microbianas. Entretanto, os metagenomas são compostos por diversos fragmentos de DNA originados de diferentes micro-organismos. Além disso, a falta de genomas de referência nos bancos de dados dificulta a identificação taxonômica de organismos desconhecidos. Nesse trabalho, avaliamos o poder de predição de uma ferramenta de classificação taxonômica de *reads* de metagenômica desconhecidos desenvolvida a partir do algoritmo de aprendizagem de máquina *Support Vector Machine* (SVM). Para simular a identificação de micro-organismos desconhecidos, utilizamos sequência de Gammaproteobacteria excluindo as sequências da espécie *Escherichia coli* como conjunto de treinamento para a SVM. Do modelo treinado, classificamos as sequências de *E. coli* e computamos as corretamente identificadas como Gammaproteobacteria. Executamos os testes com sequências de 100, 400 e 1000 pares de base para avaliar a influência do tamanho na predição. As simulações foram executadas utilizando as seguintes características de sequência como dados de entrada para o SVM: conteúdo GC, entropia de di, tri e tetrapletes, frequências de di, tri e tetranucleotídeos (2, 3 e 4-mers), abundância de dinucleotídeos e correlações de tetranucleotídeos derivadas de *z-score* (TETRA). Nós testamos conjuntos de medidas compostas por todos parâmetros excluindo um de cada vez para comparar o impacto relativo de cada medida. Observamos que os grupos que excluíam TETRA apresentam menos poder de predição para a maioria dos tamanhos testados. Os outros grupos mostraram valores AUC maiores do que 0.7 para a predição de sequências desconhecidas. Os resultados mostram que a utilização das características de sequência é uma abordagem interessante para caracterizar sequências de organismos cujas sequências não estão disponíveis em bancos de dados e para a realização de caracterização taxonômica de comunidades microbianas.

Várias análises meta-ômicas têm sido executadas para caracterizar a microbiota intestinal em casos de doenças relacionadas ao microbioma. Um exemplo é a doença de Crohn (DC), um subtipo de doença inflamatória intestinal. Ela é causada por uma manifestação atípica da resposta imune à presença de proteínas microbianas alteradas na interface da mucosa intestinal. Entretanto, devido à limitações das tecnologias e desvantagens inerentes aos métodos experimentais existentes, os mecanismos mediados pelas proteínas do microbioma alterado de DC ainda não foram exploradas. Para estudar a relação entre microbioma e DC, executamos uma abordagem computacional. Combinamos conjuntos de dados de metaproteômica publicados de um estudo de pares de gêmeos com predições de interação entre proteínas microbianas e humanas baseados em características estruturais, inferências de regiões desordenadas, vias de sinalização e redes de interação para determinar as possíveis funções mediadas pelas proteínas microbianas.

# Abstract

The acknowledgement of the importance of microbiota composition is increasing steadily after the advent of metagenomics. This approach allows sequencing and analyzing genetic material from a microbial community without the need of microbial culture. Since 99% of microorganisms are not culturable, metagenomics is the standard methodology to investigate microbiomes composition and dynamics. However, the actual output data of metagenome sequencing consists of a bunch of DNA fragments originated from various microorganisms. Moreover, the lack of reference genomes in databases challenges taxonomic identification of unknown organisms in these samples. In this work, we evaluated the predictive power of Support Vector Machine (SVM) learning tool on taxonomic classification of unknown metagenomics DNA reads. To simulate the identification of unknown microorganisms, we used Gammaproteobacteria sequences excluding *Escherichia coli* sequences as the training set in SVM. From the trained model, we classified the sequences of *E. coli* and analyzed if they were correctly assigned on Gammaproteobacteria group. The tests were performed for 100, 400 and 1000 bp test sequences to evaluate the influence of size on the prediction. The simulations were performed using the following DNA measurements as SVM input: GC content, di, tri and tetraplet entropy, di, tri and tetranucleotides frequencies (2, 3 and 4-mers), dinucleotide abundance and tetranucleotide derived z-score correlations (TETRA). We tested sets of measurements composed by all parameters but excluding one to compare the relative impact of each measure. We found that the groups which excluded TETRA shows less predictive power for the most of sizes tested, specially for 100 bp. The other groups showed AUC values higher than 0.7 for prediction of unknown sequences. The use of sequence features is an interesting approach to characterize sequences of not fully sequenced organisms and characterizing the taxonomic composition of different envoironments.

Various meta-omic based analyses have been carried out to profile and characterize the gut microbiota in microbiome-related diseases. As an example, there is Crohn's disease (CD), an inflammatory bowel disease subtype. The disease is a manifestation of the unchecked immune response to the presence of altered microbial proteins on the mucosal luminal interface and which otherwise does not elicit such a response. However, due to the limitations of the technologies and inherent drawbacks of existing experimental methods, the actual mechanisms mediated by the proteins of the altered CD microbiota have not yet been explored. To address this gap in knowledge, we carried out an integrated computational approach. We combined published metaproteomic datasets from a twin-pair CD cohort study, structural feature based interaction prediction between microbial and host receptor proteins, disordered region inferences, signalling pathways and interaction networks to determine the possible functions mediated by the microbial proteins.

# Preâmbulo

Como poderá ser observado nessa pré-tese de qualificação de doutorado, a tese completa será dividida em duas partes. A primeira parte foi iniciada em agosto de 2015, juntamente com o início do doutorado pelo programa de Pós-Graduação em Ciências Biológicas (Genética). Nessa qualificação estão inseridos os resultados preliminares obtidos até março de 2017. Os resultados finais estão programados para serem processados nos dois últimos anos do doutorado para serem inseridos na versão final da tese.

A realização da segunda parte teve início em março de 2017, quando obtive 6 meses de bolsa de doutorado sanduíche da CAPES PDSE. Durante esse período participei de pesquisas no grupo do Dr. Tamas Korcsmaros do Instituto Earlham instituído na cidade de Norwich, Inglaterra. O trabalho visa analisar a influência da microbiota intestinal no desenvolvimento da doença de Crohn. O estudo está sendo conduzido a partir de uma abordagem de biologia de sistemas, onde utilizamos dados metaproteômicos de pacientes acometidos pela doença de Crohn, metaproteomas de induvíduos saudáveis e vias de sinalização humanas. Os dados são submetidos a ferramentas de bioinformática desenvolvidas pelo grupo do Dr. Korcsmaros para analisar como a interação entre proteínas bacterianas e humanas pode conduzir ao desenvolvimento da doença de Crohn. O desenvolvimento e análise dos resultados dessa parte do doutorado, inclusive a redação e submissão de artigo científico, estão programados para serem realizados a partir de novembro de 2017.

# Sumário

# Parte I

# Parte 1: Ferramenta de *binning* de metagenomas

# 1 Introdução

O avanço da genômica esteve ancorado em dois fatores críticos: a produção crescente de dados experimentais e a nossa capacidade de processá-los. Essa simbiose foi possível, pois em linhas gerais o volume de dados cresceu na mesma velocidade em que aumentou nossa capacidade de processá-los.

Nos anos 1970, Frederick Sanger desenvolveu o primeiro método de sequenciamento de DNA. Nessa década, o termo "bioinformática" foi cunhado, sendo definido como a utilização de ferramentas computacionais para a análise de informações genéticas (HAGEN, 2000; OUZOUNIS; VALENCIA, 2003). Até meados dos anos 2000, os dados genéticos disponíveis eram provenientes do método de sequenciamento de Sanger. Essa técnica foi paulatinamente sendo substituída pelas plataformas de sequenciamento de nova geração que foram lançadas primeiramente em 2005, pela *Life Sciences* (Roche). Esses novos métodos permitiram o sequenciamento de genomas inteiros a baixos custos e em alta velocidade. A partir de então, os sequenciadores popularizaram-se entre os laboratórios de pesquisa e ciências "ômicas", como a genômica (estudo dos genes), difundiram-se entre os pesquisadores (DIJK et al., 2014; WANICHTHANARAK; FAHRMANN; GRAPOV, 2015).

Outro aspecto importante é que com o aumento da escala surgiram também desafios qualitativos, a análise dos dados obtidos é muito mais desafiadora. Agora temos que considerar não apenas genomas de um único organismo, mas também considerar como esses dados complexos evoluem no tempo e como se organizam no espaço, sem contar como os organismos interagem nos diferentes ecossistemas.

As armas que temos disponíveis para seguir avançando são arquiteturas computacionais massivamente paralelas, como as GPUs ou mesmo chips com muitos processadores organizados em arquiteturas mais complexas. E por outro lado ferramentas computacionais que explorem de forma eficiente essas novas arquiteturas. A aprendizagem de máquina é uma das maiores apostas da indústria para realizar esse avanço. Os algoritmos de aprendizagem de máquina realizam predições em conjuntos de dados a partir de padrões reconhecidos nos próprios conjuntos. Essas ferramentas são adequadas para análise de grandes quantidades de dados com profusão de ruídos e, muitas vezes, com padrões imperceptíveis para outras técnicas (BALDI; BRUNAK, 2001). A análise de dados genômicos está inserida nesse contexto.

Metagenômica é um campo da genômica que analisa a composição genética de comunidades microbianas presentes em um determinado ambiente. O advento dessa abordagem revolucionou os métodos de estudos de comunidades microbianas, pois dis-

pensa a necessidade de cultura dos micro-organismos. Deste modo, permitiu a revelação da diversidade microbiana e genética de sistemas biológicos, como novas enzimas e biocatálises, relações genômicas entre função e filogenia de organismos não cultiváveis e perfis evolucionários de comunidades (THOMAS; GILBERT; MEYER, 2012).

## 1.1 Comunidades microbianas

Os micro-organismos são os seres-vivos mais abundantes da Terra. Bactérias, arqueias, vírus e micro-eucariotos (fungos e protozoários) fazem parte de todos os ecossistemas terrestres que têm condições de suportar vida, desde os mais amenos – como solo, tecidos animais e vegetais e oceanos – até ambientes extremos, como fumarolas, minas ácidas e geleiras, onde muitas vezes são os únicos habitantes. Comunidades microbianas cumprem papéis cruciais na dinâmica dos ecossistemas, decompondo matéria morta e disponibilizando novamente nutrientes como enxofre, carbono, nitrogênio e oxigênio, para serem adquiridos por outros organismos (COUNCIL, 2007; WOOLEY; GODZIK; FRIEDBERG, 2010).

Em solos, a associação da composição microbiana com plantas é indispensável, desempenhando papéis na qualidade do solo e produtividade e saúde das plantas hospedeiras através de mecanismos diretos ou indiretos, como na mineralização da matéria orgânica do solo, ativação dos mecanismos de defesa de plantas e produção de antibióticos contra patógenos; o melhoramento de microbiomas do solo pode auxiliar para maior rendimento agrícola e controle de pestes, bem como aprimoramento de alimentos como vinhos e queijos (ZARRAONAINDIA et al., 2015). Em oceanos, podem ser observadas diferenças significativas nas comunidades microbianas em diferentes profundidades, influenciadas por características ambientais como oxigenação, salinidade e temperatura; em ambientes marinhos poluídos observou-se a presença de genes de resistência a arsênico e a metais pesados e de redução de sulfato, refletindo a alta capacidade de adaptação dos micro-organismos (BIK, 2014).

A habilidade de reciclagem de nutrientes torna os micro-organismos indispensáveis para a vida na Terra e atrai o interesse humano para aplicações que podem ser úteis em diversas áreas. Comunidades microbianas associadas a outros organismos influenciam na fisiologia do hospedeiro e contribuem para sua saúde e crescimento. A microbiota no intestino de bovinos produz enzimas para a digestão de celulose; o entendimento sobre a relação entre a digestão e as enzimas produzidas pelos microrganismos, fornecem informações que podem servir de embasamento para a melhoria da produção de leite e carne e também para a diminuição do impacto ambiental causado pela criação de gado (MORGAVI et al., 2013). No corpo de seres humanos, há mais de 100 trilhões de células de bactérias, dez vezes mais do que a quantidade das células do próprio

corpo (BELLA et al., 2013). Esses micro-organismos são indispensáveis para a vida do ser humano, habitam muitos de seus tecidos garantindo imunidade e aquisição de nutrientes a seu corpo. A microbiota da pele auxilia na imunidade e proteção dos humanos. A composição microbiológica do trato gastrointestinal influencia na aquisição de nutrientes, no rendimento de energia e em diversas vias metabólicas e seu desequilíbrio pode facilitar a indução de doenças como diabetes tipo 2, obesidade e doenças inflamatórias do intestino, como doença de Crohn. A a partir de estudos dessas microbiotas, podem ser desenvolvidos meios alternativos de tratamento e de prevenção dessas doenças e de outras mais (DEVARAJ; HEMARAJATA; VERSALOVIC, 2013).

### 1.1.1 Taxonomia microbiana

A relação das comunidades microbianas com a saúde humana tem aumentado a demanda por estudos de composição de microbiomas e impulsionou o desenvolvimento de novas tecnologias de identificação e descrição de micro-organismos. Com o avanço de tecnologias de sequenciamento de DNA e novas descobertas genéticas, características genotípicas têm sido consideradas cada vez mais importantes na organização dos táxons de micro-organismoss (MADIGAN et al., 2016).

Atualmente, micro-organismos são organizados de acordo com suas características fenotípicas, genotípicas e filogenéticas, dando origem a diversos níveis de táxons (MADIGAN et al., 2016). Táxons são "grupos progressivamente mais inclusivos" onde os seres-vivos são organizados de acordo com características em comum, que hoje incluem genótipo, fenótipo e filogenia (MADIGAN et al., 2016; THOMPSON et al., 2013). O nível taxonômico mais alto existente é domínio, seguido de filo, classe, ordem, família, gênero e espécie (MADIGAN et al., 2016).

O principal método de classificação taxonômica de micro-organismos a partir de características genotípicas é o 16S rRNA (CLARRIDGE, 2004). O gene 16S rRNA tem regiões hiperconservadas entre os micro-organismos e variáveis ao longo de sua sequência (Figura 1). As regiões conservadas são quase idênticas entre os organismos e seu alto índice de conservação permite o desenvolvimento de primers universais que possibilitam o isolamento dos genes da amostra. As outras regiões variam de acordo com a proximidade evolutiva entre as espécies. As sequências das regiões variáveis dos organismos a ser identificados são comparados com as disponíveis em banco de dados e a identidade entre eles é calculada (NIKOLAKI; TSIAMIS, 2013). Com um índice de identidade de 99%, dois micro-organismos podem ser considerados de mesma espécie. Índices de identidade entre 97% e 99% identificam micro-organismos de mesmo gênero. Espécies potencialmente novas devem apresentar identidade menor do que 97% comparado à espécies já catalogadas (DRANCOURT; BERGER; RAOULT, 2004).

A comparação de genes de RNA ribossômico acarretou na divisão dos seres

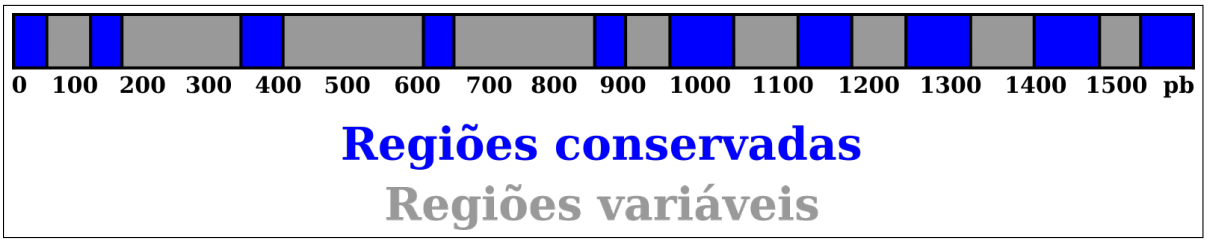


Figura 1 – Ilustração das regiões conservadas (em azul) e variáveis (em cinza) do gene 16S rRNA.

vivos nos domínios Eukarya, Archaea e Bacteria 2. Nos domínios Bacteria e Archaea estão presentes micro-organismos procariotos: organismos formados por uma célula sem organelas e nem núcleo. O domínio Eukarya é formado por organismos pluri e unicelulares, com célula(s) eucariótica(s), ou seja, que apresentam organelas e núcleo. Embora ambos Bacteria e Archaea sejam formados por organismos procariotos, o domínio Archaea é molecularmente mais próximo do domínio Eukarya do que do Bacteria. O último ancestral universal comum de todas as células (LUCA, sigla originada do inglês "*Last Universal Common Ancestor*") originou duas vertentes, sendo uma delas o domínio Bacteria. A segunda vertente divergiu entre Archaea e em Eukarya, como podemos observar na Figura 2. Os domínios Bacteria e Archaea são divididos em filos e o domínio Eukarya apresenta o nível de reino acima de filos (MADIGAN et al., 2016; WOESE; KANDLER; WHEELIS, 1990).

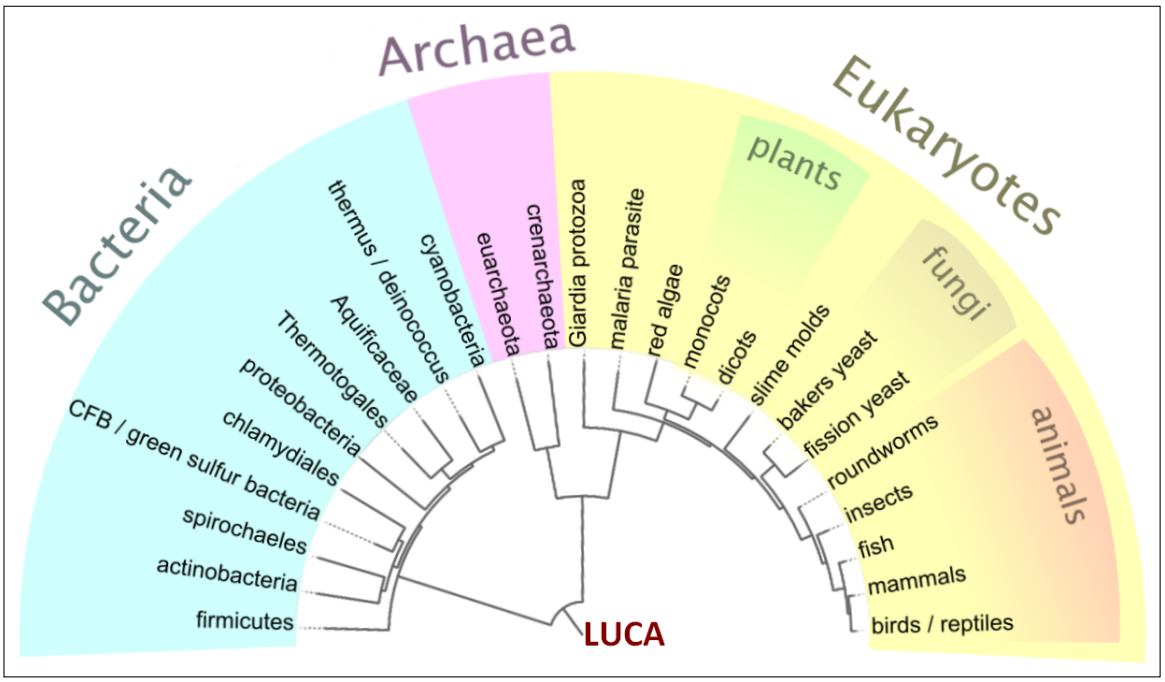


Figura 2 – Árvore de classificação dos três Domínios: Bacteria, Archaea e Eukaryotes e seus respectivos Reinos (no caso de Eukaryotes) e principais filos a partir do último ancestral comum (LUCA).

## 1.2 Metagenômica

O maior fator limitante no estudo da composição de comunidades microbianas é que apenas 1% dos micro-organismos podem ser cultivados em laboratório (**??**), restringindo consideravelmente a extensão a que estudos de microbiomas podem ser conduzidos a partir de meios de cultura. Essa dificuldade foi superada com o advento das tecnologias de sequenciamento de DNA que possibilitaram o estabelecimento de um novo campo de estudo inserido na genômica: a metagenômica. O termo, cunhado por Handelsman em 1998 (HANDELSMAN et al., 1998), define o estudo dos genomas de comunidades microbianas presentes em um determinado habitat a partir do DNA extraído desse ambiente, sem a necessidade de cultivo dos micro-organismos. Deste modo, permitiu a revelação da diversidade microbiana e genética de diversos sistemas biológicos, relações genômicas entre função e filogenia de organismos não cultiváveis e perfis evolucionários de comunidades, além de outras interações biomoleculares (MARCO, 2011; THOMAS; GILBERT; MEYER, 2012).

Como exemplos de grandes iniciativas baseadas nessa tecnologia, há o Projeto Microbioma Humano (Human Microbiome Project), financiado pelo NIH (National Institutes of Health), e o consórcio europeu MicroWine. O primeiro tem como objetivo sequenciar o metagenoma de partes do corpo humano, como cavidade gastro-intestinal, olhos, pele, vias aéreas, trato urogenital e sangue, para esclarecer o papel do microbioma na saúde e desenvolver novas ferramentas que possam ser utilizados posteriormente em prol de outras pesquisas (PETTERSSON; LUNDEBERG; AHMADIAN, 2009). Já o MicroWine, explora comunidades de micro-organismos que desempenham papéis importantes em todos os estágios da vinicultura, auxiliando desde o acesso das plantas a nutrientes do solo e na sua imunidade contra patógenos até os processos de vinificação, que influenciam nos sabores e aromas característicos de cada vinho (MICROWINE, A MARIE CURIE INITIAL TRAINING NETWORK, 2016).

A primeira etapa de qualquer estudo metagenômico envolve a retirada das amostras do ambiente de estudo e posterior isolamento, fragmentação e sequenciamento do material genético dos micro-organismos relacionados àquele meio. Há três gerações de métodos de sequenciamento. Os métodos de primeira e segunda geração fragmentam o DNA em segmentos (*reads*) cujos comprimentos variam entre 35 e 700 pares de base. Devido à sua natureza, a análise dos dados metagenômicos resultantes dessas técnicas através de ferramentas computacionais torna-se bastante complexa. O método de primeira geração, também chamado de sequenciamento de Sanger, ainda é utilizado devido à sua baixa taxa de erros e *reads* relativamente longos, com mais de 700 pb, facilitando a análise pós-sequenciamento. Entretanto, seu custo é mais elevado do que das plataformas de nova geração: U$ 400 mil por gigabase, e limita-se a até 96 Kb de informação por sequenciamento. Em contrapartida, as plataformas de segunda geração podem chegar a custar

U$ 50,00 por gigabase e rendem mais de 1 Gb por sequenciamento. Consequentemente, essas tecnologias vêm substituindo o sequenciamento de Sanger, principalmente através de plataformas como Illumina/Solexa, 454/Roche e Applied Biosystems SOLiD. Por sua vez, o sequenciamento de terceira geração, também conhecido como sequenciamento de molécula única, propõe o rendimento de mais dados a menores custos e *reads* de tamanho maior do que 10 mil pb; as duas tecnologias de sequenciamento de molécula única mais utilizados são Pacific Biosciences e Oxford Nanopore. Apesar de suas vantagens, sequenciamentos de terceira geração ainda são pouco utilizados na metagenômica devido a sua alta taxa de erros (Tabela 1) (MOROZOVA; MARRA, 2008; THOMAS; GILBERT; MEYER, 2012; LAND et al., 2015; OULAS et al., 2015; LEE et al., 2016).

Tabela 1 – Lista de plataformas de sequenciamento, o tamanho de seus *reads* e seu custo por GB (MOROZOVA; MARRA, 2008; THOMAS; GILBERT; MEYER, 2012; LEE et al., 2016).

| Geração | Tecnologia de sequenciamento | Tamanho dos *reads* | Custo por GB (aprox.) | Rendimento por corrida |
|---|---|---|---|---|
| 1ª | Sanger | >700 pb | U$ 400 000 | 96 kb |
| 2ª | 454 / Roche | 400 - 700 pb | U$ 20 000 | 80 - 120 Mb |
| 2ª | Illumina | 100 - 150 pb | U$ 50 | 1 Gb |
| 2ª | Life Technologies / SOLiD | 35 - 75 pb | U$ 130 | 1 - 3 Gb |
| 3ª | Pacific Biosciences | 10 - 15 kpb | U$ 500 | 5 Gb |
| 3ª | Oxford Nanopore Technologies | 5 - 10 kpb | U$ 1000 | >40 Gb |

A diminuição de preço das tecnologias de sequenciamento de segunda e terceira geração (NGS, do inglês *new generation sequencing*, ou sequenciamento de nova geração) permitiu a popularização da metagenômica entre os pesquisadores, e, consequentemente, o aumento na quantidade de dados disponibilizada em bancos de dados. Entretanto, o poder computacional e o desenvolvimento de algoritmos de análise de metagenomas não acompanha o crescimento na quantidade de dados produzidos. O primeiro problema está relacionado com a disponibilidade dos metagenomas: os sistemas de armazenamento de sequências não suportam quantidades de dados tão massivas e o formato dos dados não é padronizado. Outro obstáculo está relacionado às características dos dados produzidos: *reads* muito curtos e grande quantidade de erros gerados pelas plataformas de nova geração fazem com que a análise de metagenomas demande algoritmos mais complexos e custosos computacionalmente. Deste modo, é evidente a necessidade de novas ferramentas de análise de metagenomas para a maioria das etapas do processamento de dados pós-sequenciamento (KIM et al., 2013; KUMAR et al., 2015).

Existem duas categorias para o sequenciamento de metagenomas: na primeira, um gene marcador, mais frequentemente o 16S rRNA, é isolado através de PCR e sequenciado; já na metagenômica por WGS, do inglês, *whole genome shotgun*, todo o DNA dos micro-organismos presentes na amostra é sequenciado (OULAS et al., 2015). A escolha da técnica

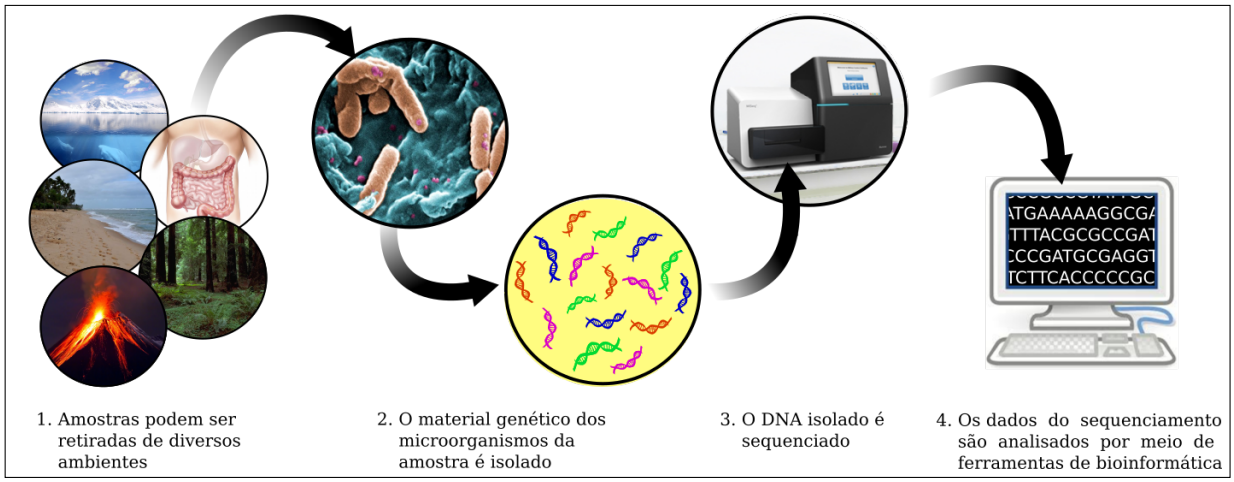


Figura 3 – Figura ilustrando as etapas de amostragem e análises de metagenômica. a) *Amostragem*: É retirada uma amostra de micro-organismos de um ambiente (e.g. geleiras, oceano, solo e intestino). Seu material genético é extraído e sequenciado. Os dados de saída consistem em *reads* de diversos organismos misturados. b) *Binning*: a análise taxonômica por *binning* consta no agrupamento dos *reads* do metagenoma a partir de características das sequências encontradas por ferramentas de bioinformática. A partir de então, é possível classificar os grupos de fragmentos em organismos cujo nível taxonômico dependerá da ferramenta utilizada. b) *Análise funcional*: busca-se a similaridade das sequências dos *reads* com genes e proteínas conhecidos disponíveis em bancos de dados para identificar genes e caracterizar funcionalmente o metagenoma.

mais adequada depende do objetivo de análise dos dados.

Há métodos de análise e ferramentas para trabalhar especificamente com os dados de cada abordagem. Portanto, entraremos em detalhes sobre as ferramentas utilizadas dentro das seções abaixo que discorrem separadamente sobre metagenômica a partir de 16S rRNA e de WGS.

### 1.2.1 Metagenômica a partir do gene 16S rRNA

Embora os estudos a partir de WGS estejam sendo desenvolvidos com frequência cada vez mais alta, a análise dos 16S rRNA ainda é amplamente aceita e é uma ferramenta poderosa para o estudo das comunidades microbianas em alta resolução (SUN et al., 2011). A utilização do gene 16S rRNA como marcador taxonômico possibilitou o desenvolvimento de um método de identificação de micro-organismos de uma amostra sem a necessidade de cultivo dos micro-organismos. A primeira execução bem sucedida ocorreu em 1991, quando foram registrados novas espécies a partir da análise dos genes 16S rRNA de amostras de oceano (SCHMIDT; DELONG; PACE, 1991; RIESENFELD; SCHLOSS; HANDELSMAN, 2004).

A inclusão da análise taxonômica de microbiomas a partir do gene 16S rRNA no

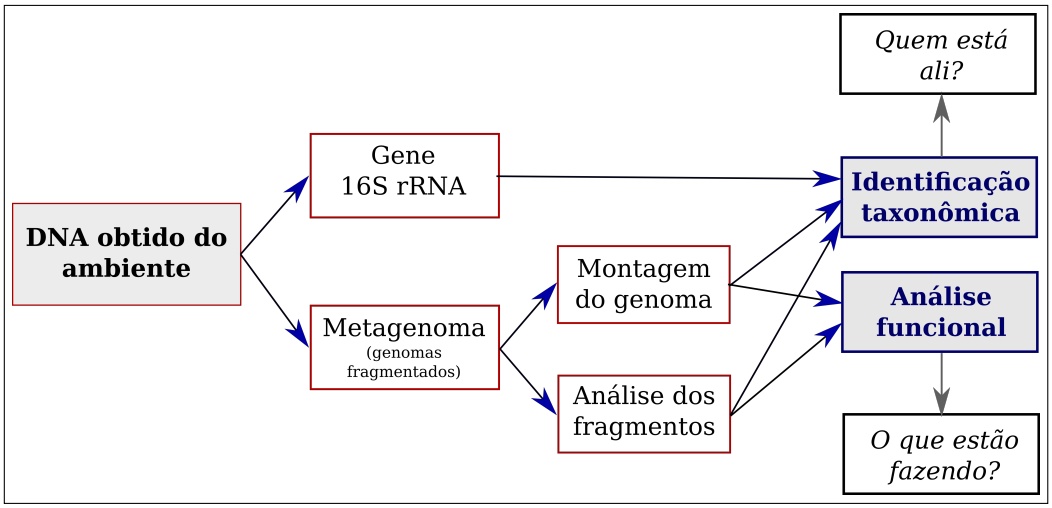


Figura 4 – O DNA obtido do ambiente pode ser analisado de duas formas: i) a partir da informação de somente um gene, sendo o mais utilizado o 16S rRNA, ou ii) a partir da informação de todo o DNA sequenciado, ou metagenoma.

conceito de "metagenômica" ainda é uma controvérsia entre os pesquisadores. Muitos deles sugerem que esse tipo de análise seja denominada "metagenética", por utilizar apenas um gene e não todo o genoma (ESPOSITO; KIRSCHBERG, 2014).

O gene 16S rRNA codifica a subunidade pequena do RNA ribossômico de Archaeas e Bactérias e mostrou-se adequado por apresentar regiões hiperconservadas intercaladas com regiões variáveis ao longo de sua sequência. As regiões conservadas são quase idênticas dentre os micro-organismos, portanto são utilizadas para desenvolver *primers* universais para o isolar os genes da amostra. As outras regiões variam proporcionalmente à proximidade filogenética entre os táxons, sendo portanto utilizadas como parâmetros de comparação para a identificação dos micro-organismos (Figura 1) (KIM et al., 2013; NIKOLAKI; TSIAMIS, 2013).

Entretanto, as regiões variáveis do gene 16S rRNA apresentam baixa resolução entre as espécies, permitindo a classificação eficiente dos micro-organismos de um metagenoma somente até o nível de gênero. Outro obstáculo da técnica 16S rRNA na metagenômica é a alta susceptibilidade a vieses, pois os *primers* podem apresentar mais afinidade por sequências de determinadas espécies do que de outras. Essa tendência pode favorecer a seleção dos genes alguns organismos em detrimento de outros em uma amostra e impedir que novos táxons sejam representadas se não forem compatíveis aos *primers* utilizados (NIKOLAKI; TSIAMIS, 2013; PORETSKY et al., 2014).

O primeiro passo para a execução da metagenômica a partir do gene 16S rRNA é a obtenção do DNA dos micro-organismos de um meio ambiente. Depois de isolado, esse DNA passa pelo processo de PCR, onde ocorre a amplificação dos genes a partir de *primers* que identificam a localização do gene 16S rRNA e o replica diversas vezes. O produto dessa amplificação é submetido à técnica de eletroforese, onde o gene pode ser identificado em

um gel por meio de bandas. As bandas que correspondem ao gene 16S rRNA são selecionadas e o DNA contido é purificado e submetido ao sequenciamento (SANSCHAGRIN; YERGEAU, 2014; ZHOU et al., 2015).

Depois de sequenciados, os dados resultantes (*output*) são armazenados em um computador e devem ser processados e analisados a partir de ferramentas de bioinformática.

Para a análise computacional, primeiramente, realiza-se o pré-processamento dos dados oriundos do sequenciamento. O *output* do sequenciamento de DNA é um conjunto de dados brutos com erros de replicação e sequências de baixa qualidade que prejudicam a análise dos genes sequenciados. Para a análise mais precisa dos metagenomas, é necessário realizar uma etapa denominada, em inglês, *denoising*, que consiste no pré-processamento dos dados brutos. No *denoising*, realiza-se uma filtragem das sequências para que reste somente as que representam a comunidade microbiana com qualidade (KIM et al., 2013; OULAS et al., 2015). As ferramentas de *denoising* mais utilizadas são PyroNoise, Amplicon-Noise, QIIME e DADA (CAPORASO et al., 2010; QUINCE et al., 2011; ROSEN et al., 2012; GASPAR; THOMAS, 2013).

As sequências de baixa qualidade a serem retiradas dos dados incluem quimeras. Quimeras são recombinantes artificiais que se formam entre duas ou mais sequências durante a amplificação do DNA por PCR. Elas normalmente formam-se quando fragmentos de DNA que terminam prematuramente a amplificação anelam-se a outros. Essas moléculas artificiais dificultam a diferenciação das sequências originais das recombinantes, resultando na superestimação do nível de diversidade microbiana presente na amostra (KIM et al., 2013). Entretanto, a detecção das quimeras na amostra não é um processo trivial, uma vez que a união das moléculas ocorre em posições aleatórias e as plataformas de NGS geram bases de dados de 16S rRNA e *reads* curtos, dificultando a localização das sequências originais que possuam informação taxonômica suficiente (KIM et al., 2013). As ferramentas mais utilizadas para a remoção de quimeras dos dados obtidos por NGS são ChimeraSlayer (HAAS et al., 2011), Perseus (QUINCE et al., 2011), Decipher (WRIGHT; YILMAZ; NOGUERA, 2012) e UCHIME (EDGAR et al., 2011).

Depois que a filtragem dos dados for executada, realiza-se a classificação taxonômica dos genes 16S rRNA. Os pesquisadores que optam por realizar seus estudos a partir do 16S rRNA normalmente visam a obtenção de um perfil taxonômico ou filogenético da comunidade microbiana em questão, para responder a pergunta *"quem está no ambiente?"*. A partir dessas informações é possível realizar estudos relacionados à evolução da comunidade microbiana, associação entre micro-organismos e comparação, composição microbiótica de diferentes meios, entre outros.

A classificação dos genes podem ocorrer basicamente de duas maneiras: de forma taxonomicamente dependente ou independente. Nas análises taxonomicamente dependen-

tes, as sequências de 16S rRNA desconhecidas são comparadas com sequências conhecidas disponíveis em bancos de dados e atribuídas aos táxons das que apresentam maior similaridade. Essas comparações podem ser feitas a partir de ferramentas: i) de alinhamento de sequencias, como BLAST (ALTSCHUL et al., 1990), ii) de composição de sequencias, como RDP (WANG; HUYCKE, 2007; KIM et al., 2013), ou iii) de acordo com sua alocação em uma árvore filogenética guia baseada em modelos evolucionários, como SEPP (MIRARAB; NGUYEN; WARNOW, 2012), EPA (BERGER; KROMPASS; STAMATAKIS, 2011), pplacer (MATSEN; KODNER; ARMBRUST, 2010), QIIME (CAPORASO et al., 2010) e AMPHORA2 (WU; SCOTT, 2012; KIM et al., 2013). Posteriormente, as sequências alinhadas são submetidas à análises em ferramentas que processam os *outputs* e apresentam as atribuições taxonômicas ao pesquisador, como MEGAN (HUSON et al., 2007) e TANGO (CLEMENTE; JANSSON; VALIENTE, 2011; KIM et al., 2013).

Em análises taxonomicamente independentes, as sequências são agrupadas de acordo com determinados índices de similaridade apenas comparando umas com as outras, sem utilizar bases de dados como referência (SUN et al., 2011). Embora muitas ferramentas apliquem os métodos separadamente, essas abordagens podem ser utilizadas concomitantemente para uma análise mais prática e aprofundada. A realização de análises taxonomicamente independentes normalmente baseiam-se na clusterização de OTUs. OTU, sigla em inglês *operational taxonomic unit*, significa "unidades taxonômicas operacionais" em português. OTUs são os agrupamentos dos genes 16S rRNA de acordo com sua similaridade. Níveis de similaridade maiores do que 97% para bactérias e archaeas correspondem à mesma espécie (PATIN et al., 2013; OULAS et al., 2015).

Algoritmos para clusterização de sequências de 16S rRNA utilizam basicamente duas estratégias: a partir de alinhamento de sequências, na qual as sequências desconhecidas de 16S rRNA são alinhadas entre elas, podendo haver ou não a utilização de sequências de referência, e estratégias independentes de alinhamento. Dentre elas estão NAST (DESANTIS et al., 2006), SINA aligner (PRUESSE; PEPLIES; GLÖCKNER, 2012), Infernal (NAWROCKI; KOLBE; EDDY, 2009), UCLUST (EDGAR et al., 2011), CD-HIT (FU et al., 2012) e ESPIRIT-Tree (CAI; SUN, 2011).

### 1.2.2 *Whole Genome Shotgun*

Embora a metagenômica a partir do gene 16S rRNA seja um modo eficaz para estudar a diversidade microbiana das comunidades, é possível obter somente informações sobre a composição dos micro-organismos dos meios analisados. A partir de metagenômica por WGS, é possível adquirir as informações de biodiversidade relacionadas à composição funcional do meio, possibilitando a responder as questões *"quem está presente no ambiente?"*, *"o que esses micro-organismos podem fazer?"* e *"como eles interagem?"* (FORDE; O'TOOLE, 2013; OULAS et al., 2015).

O procedimento para metagenômica por WGS também inclui o isolamento do material genético da amostra, assim como na metagenômica por 16S rRNA, embora no caso da WGS todo o DNA é submetido ao sequenciamento. Deste modo, o resultado do sequenciamento consiste em diversos segmentos de DNA dos genomas de micro-organismos presentes no meio. Como o genoma é aleatoriamente fragmentado, os *reads* do metagenoma pertencem a qualquer parte do genoma: desde genes taxonomicamente informativos, como o gene 16S rRNA, até sequências codificadoras que fornecem informações sobre funções que os micro-organismos podem realizar no ambiente (SHARPTON, 2014).

Por apresentar maior complexidade do que os metagenomas de 16S rRNA, o desenvolvimento de ferramentas para análise de metagenomas por WGS também é mais desafiador. Primeiramente, porque é mais difícil determinar a origem taxonômica dos *reads*. Depois, na maioria das vezes não é possível obter a representação de todos os micro-organismos do ambiente por causa de sua diversidade. Além disso, a etapa de filtragem é mais complicada do que a do 16S rRNA: a identificação e remoção de sequências contaminantes é problemática devido à dificuldade de diferenciar as sequências dos micro-organismos das sequências de organismos indesejados. Também pode haver a necessidade de montar os genomas, etapa desafiadora para os bioinformatas devido ao alto custo computacional e demanda por grandes quantidades de informações (SHARPTON, 2014).

Essas limitações tem esmaecido com o avanço das tecnologias de informática, que vem apresentando computadores mais potentes ao longo do tempo, e com as tecnologias de sequenciamento, que prezam por maiores tamanhos de *reads*. Abaixo, estão detalhadas as etapas de análise de bioinformática para metagenômica por WGS e ferramentas utilizadas para cada finalidade.

#### 1.2.2.1 Controle de qualidade das sequências

O primeiro passo a ser realizado depois que o metagenoma é sequenciado, é o controle de qualidade, ou filtragem dos dados retornados para retirar sequências de baixa qualidade do metagenoma. Essa é uma importante etapa a ser realizada, pois erros e sequências de organismos não desejados dificultam a montagem dos reads e sua análise, especialmente quando o contaminante é altamente abundante ou tem um grande genoma (BRAGG; TYSON, 2014; SHARPTON, 2014). Alguns dos programas utilizados para o controle de qualidade são plataforma-específicos, sendo os mais utilizados FASTX toolkit (FASTX-TOOLKIT..., 2016), FASTQC (BABRAHAM BIOINFORMATICS, 2016), ngs backbone (BLANCA et al., 2011), Pyrobayes (QUINLAN et al., 2008) e Shore (OSSOWSKI et al., 2008).

#### 1.2.2.2 Montagem de fragmentos

Depois de pré-processados, os *reads* dos metagenomas podem ser submetidos à de montagem. Nessa etapa, os fragmentos são unidos com outros originados do mesmo genoma para formar sequências maiores e, assim, facilitar a análise. É praticamente impossível realizar a montagem de genomas inteiros a partir de dados de metagenômica, pois o genoma da maioria dos micro-organismos representados na amostra não é completamente sequenciada e é difícil atribuir exatamente a qual espécie cada *read* pertence. Entretanto, em alguns casos, é possível montar grande parte dos genomas para realizar estudos que requerem a estrutura do genoma, como por exemplo em análises funcionais que busquem por regiões codificantes. (WOOLEY; GODZIK; FRIEDBERG, 2010; KIM et al., 2013; SHARPTON, 2014).

Há dois tipos de montagem de genomas que pode ser executados: montagem a partir de genomas de referência e montagem de novo. Na primeira, utiliza-se um ou mais genomas de referência como "mapas" onde os *reads* podem ser posicionados, formando fragmentos maiores quando dois ou mais *reads* dispõem-se um ao lado do outro; MetaAMOS (TREANGEN et al., 2013), Newbler (montagem de *reads* da 454-Roche) e MIRA4 (CHEVREUX et al., 2004) são as ferramentas que utilizam genomas de referência mais utilizadas. Montagem de novo une os fragmentos sem a utilização de genomas de referência utilizando algoritmos mais complexos, como por exemplo grafos de-Bruijin (COMPEAU; PEVZNER; TESLER, 2011); as ferramentas de montagem de novo mais frequentemente utilizadas são Abyss (SIMPSON et al., 2009), Velvet (ZERBINO; BIRNEY, 2008), SOAP (LI et al., 2008) e EULER (PEVZNER; TANG; WATERMAN, 2001). Há ainda uma nova geração de ferramentas de montagem especificamente para metagenomas. A maioria dos algoritmos citados acima foram desenvolvidos para a montagem de *reads* sequenciados a partir de genomas únicos e adaptados para a aplicação em metagenomas. Dois exemplos de *softwares* de nova geração são Meta-Velvet-SL (NAMIKI et al., 2012) e Meta-IDBA (PENG et al., 2011).

#### 1.2.2.3 Análise taxonômica e *binning*

Para analisar a diversidade taxonômica de uma amostra de metagenômica, pode ser realizado um processo denominado *binning* (MANDE; MOHAMMED; GHOSH, 2012). A partir desse procedimento os *reads*, que até então não têm sua origem taxonômica determinada, são agrupados em táxons de acordo com diferentes características da sequência. Além de quantificar os micro-organismos de diferentes táxons que estão presentes no meio, a partir do *binning* é possível reduzir a complexidade do conjunto de dados para facilitar posteriores fases do estudo, como a montagem de fragmentos ou análise funcional (SHARPTON, 2014).

A partir de análises taxonômicas é possível estudar o papel da composição das

comunidades microbianas nos ecossistemas em que fazem parte. Zarraonaindia mostrou que a diversidade de espécies associadas aos órgão das parreiras (folhas, flores, frutos e raízes) e ao solo onde estão plantadas são importantes para definir o sabor final dos vinhos produzidos (ZARRAONAINDIA et al., 2015). Outro estudo demonstrou que a microbiota intestinal pode influenciar no desenvolvimento da obesidade (RIDAURA et al., 2013). Resultados como os citados evidenciam a importância da composição microbiana nos ambientes em diversas áreas.

Apesar da importância das análises taxonômicas, esse procedimento representa grandes desafios para os pesquisadores e desenvolvedores de *software* principalmente por causa do pequeno tamanho dos *reads* obtidos pelas NGS. Por serem muito pequenos, muitas vezes não apresentam informações suficientes para que seja possível classificá-los. Consequentemente, muitos fragmentos acabam sendo excluídos do agrupamento. Para minimizar esses problemas, pode ser realizada a pré-montagem dos fragmentos e deve ser utilizada a ferramenta de análise mais adequada para o tipo de dados que se está trabalhando (KIM et al., 2013).

Existem basicamente duas categorias de ferramentas que utilizam diferentes abordagens para a classificação taxonômica dos *reads* de metagenômica: similaridade de sequências e de composição de sequências.

**Ferramentas de similaridade de sequência:** Classificam a sequência desconhecida de acordo com sua similaridade com as armazenadas em bancos de dados. Primeiramente, os dados são alinhados com ferramentas como BLAST. Posteriormente, os *softwares* de análise de metagenômica utilizam as informações para fazer inferências taxonômicas e filogenéticas (KIM et al., 2013).

O método de similaridade de sequências para a classificação taxonômica dos metagenomas fornece maior resolução e acurácia da análise do que o *binning* a partir da composição de sequências. Entretanto, seu custo computacional é maior e aumenta exponencialmente com a diminuição do comprimento dos *reads* (LIU et al., 2013; SHARPTON, 2014). Abaixo, estão citadas ferramentas de análise taxonômica por similaridade popularmente utilizadas (SHARPTON, 2014):

- **MEGAN:** utiliza BLAST para comparar os *reads* de metagenômica com banco de dados de sequências que são anotadas com a taxonomia do NCBI. Depois, o *software* infere a taxonomia da sequência colocando o fragmento em um nó da árvore taxonômica do NCBI correspondente ao último ancestral comum (LCA, do inglês *last commom ancestor*) de todos os táxons que contém homologia com o *read* (HUSON et al., 2007).
- **MG-RAST:** utiliza reconstrução filogenética de sequências de banco de dados que sejam similares a cada *read* para classificá-lo taxonomicamente (MEYER

et al., 2008).

- **CARMA:** utiliza os melhores alinhamentos recíprocos entre as sequências disponíveis em banco de dados e os *reads* de metagenômica e modelos de índices de evolução específicos gene-família para inferir o rank taxonômica apropriado a cada fragmento de DNA (GERLACH; STOYE, 2011).

**Ferramentas de composição da sequência:** Utiliza as características intrínsecas da composição das sequências (e.g. conteúdo GC, frequência de oligonucleotídeos, utilização de códons, assinaturas periódicas) para classificá-las taxonomicamente. Essas características, também chamadas de *assinaturas genômicas*, são moldadas em cada grupo taxonômico ao longo da evolução de acordo com as pressões evolutivas às quais os micro-organismos estão sujeitos. Levando em conta que organismos filogeneticamente mais próximos apresentarão mais similaridade na composição de suas sequências, é possível agrupar ou classificar os *reads* de metagenômica de acordo com essas características (CAMPBELL; MRAZEK; KARLIN, 1999; THOMAS; GILBERT; MEYER, 2012).

As ferramentas que utilizam assinaturas genômicas são mais rápidas e custam menos computacionalmente do que as ferramentas de similaridade. Entretanto, apresentam dificuldade na identificação de fragmentos muito pequenos (i.e., 150 pb), pois eles não contém informação suficiente para uma classificação eficiente (MANDE; MOHAMMED; GHOSH, 2012).

Para agrupar e classificar os *reads*, os programas de composição de sequências frequentemente utilizam aprendizagem de máquina supervisionados ou não supervisionados. Os supervisionados utilizam genomas conhecidos para que o algoritmo reconheça os padrões de cada grupo taxonômico; os *reads* são atribuídos a um táxon de acordo com as características reconhecidas pela máquina. Algoritmos não supervisionados não utilizam sequências de referência, eles comparam um fragmento com outro reconhecendo padrões comuns entre eles e os agrupando em conjuntos de acordo com suas características compartilhadas. *Binning* a partir de algoritmos não supervisionados reúnem os *reads* em grupos taxonomicamente distintos, mas é necessária a utilização de ferramentas adicionais para atribuir táxons a cada agrupamento (BRAGG; TYSON, 2014).

Os *Softwares* mais popularmente utilizados para o *binning* por composição de sequências são (KIM et al., 2013; SHARPTON, 2014):

- **PhyloPithia e PhyloPithiaS:** utilizam o algoritmo de aprendizagem supervisionada de máquina *support vector machines*, que analisa sequências de treinamento já identificadas taxonomicamente para construir modelos de

frequências de oligonucleotídeos que determinam se o *read* é um membro do grupo (MCHARDY et al., 2007).

- **Phymm:** ferramenta supervisionada que utiliza modelos Markovianos interpolados que combinam probabilidades de predição derivadas de sequências de treinamento com diversos tamanhos. Opcionalmente, aplica alinhamento com BLAST para a classificação dos *reads*. Adequado para sequências pequenas originadas de NGS (BRADY; SALZBERG, 2009; BRADY; SALZBERG, 2011).
- **NBC:** utiliza classificador supervisionado de *Naive Bayes* baseando-se nos perfis de frequência de *k-mers* de cada grupo taxonômico. Adequado para sequências pequenas originadas de NGS (ROSEN et al., 2012).
- **TACOA:** ferramenta não supervisionada que utiliza a "regra do vizinho mais próximo" (*k-nearest neighbor*) para agrupar os *reads* (DIAZ et al., 2009).

**Ferramentas híbridas:** Para compensar as vantagens e desvantagens das categorias de algoritmos citados acima, há ferramentas que utilizam tanto a abordagem de composição de sequências quanto o alinhamento para a classificação taxonômica dos *reads* de metagenômica. Alguns exemplos estão descritos abaixo:

- **PhymmBL:** combina a ferramenta Phymm, descrita anteriormente, com alinhamentos em BLAST pata aumentar a precisão da classificação (BRADY; SALZBERG, 2009).
- **RITA:** combina BLAST com a ferramenta NBC (descrita anteriormente), mas considera os resultados do BLAST com mais peso (MACDONALD; PARKS; BEIKO, 2012).
- **SPHINX:** na primeira fase, SPHINX compara a composição de tetranucleotídeos do *read* com a dos genomas de referência para realizar uma pré-filtragem dos grupos taxonômicos a que ele possa pertencer. Depois, utiliza algoritmos de alinhamento de sequência para classificar o fragmento mais restritamente (MOHAMMED et al., 2010).

### 1.2.3 Análise funcional

A análise funcional dos metagenomas fornece informações sobre funções codificadas nos genomas dos micro-organismos da comunidade, respondendo à pergunta "o que os micro-organismos podem fazer no ambiente estudado?". A partir da caracterização dos genes do metagenoma é possível traçar um perfil funcional da comunidade microbiana que pode ser utilizado para comparar metagenomas de diferentes ambientes, revelar a presença de novos genes ou fornecer informações do mesmo ambiente em diferentes condições (SHARPTON, 2014).

Para realizar a análise funcional do metagenoma, primeiramente identifica-se os fragmentos que contém sequências codificadoras e depois as compara com sequências de banco de dados de genes, proteínas, famílias de proteínas ou vias metabólicas com funções conhecidas para identificar qual a função dos genes desconhecidos (SHARPTON, 2014). Os bancos de mais utilizados para a busca de informações de genes conhecidos incluem (BELLA et al., 2013): KEGG (*Kyoto Encyclopedia of Genes and Genomes*) (KANEHISA; GOTO, 2000), COG (*Clusters of Orthologous Groups system*) (TATUSOV et al., 2003), Pfam (BATEMAN et al., 2004), CDD (*Conserved Domains Database*) (MARCHLER-BAUER et al., 2005), SEED (OVERBEEK et al., 2005), TIGRFAM (SELENGUT et al., 2006) e eggNOG (MULLER et al., 2009). Para as análises funcionais recomenda-se utilizar metaproteômica e metatranscriptômica para uma análise mais precisa dos genes que de fato estão sendo expressos pelos micro-organismos.

#### 1.2.3.1 Ferramentas generalizadas

Há ferramentas que foram desenvolvidas para executar a análise funcional dos metagenomas de forma mais simplificada, promovendo a interação de algoritmos de identificação de genes, alinhamento de sequencias e bancos de dados. Muitas delas também processam outras etapas como filtragem dos dados e comparação de metagenomas e fornecem visualização acessível dos resultados (THOMAS; GILBERT; MEYER, 2012; BELLA et al., 2013; OULAS et al., 2015). Dentre eles estão MG-RAST (KENT, 2002; MEYER et al., 2008; GLASS et al., 2010), IMG/MER 4 (MARKOWITZ et al., 2013), EBI Metagenomics service (HUNTER et al., 2013), CAMERA (SESHADRI et al., 2007) e MEGAN (HUSON et al., 2007).

## 1.3 Aprendizagem de máquina

As ferramentas de aprendizagem de máquina são alternativas eficazes quando há necessidade de analisar uma grande quantidade de dados. Diversas áreas além da genética, como financeira, marketing, médica, robótica, e.g., utilizam algoritmos para o processamento de dados que humanos são incapazes de analisar sozinhos (MARSLAND, 2014).

Os algoritmos de aprendizagem de máquina podem ser classificados como algoritmos de aprendizagem supervisionada ou não supervisionada:

**Aprendizagem supervisionada:** Um conjunto de dados rotulados, chamados de exemplos de treinamento, são fornecidos à ferramenta. O algoritmo encontra padrões e constrói um modelo para cada classificação possível. Quando os novos dados são apresentados ao modelo aprendido, o algoritmo retorna uma predição da classificação baseado no conceito assimilado (MAHAMUDA; U; RASHEED, 2010).

**Aprendizagem não supervisionada:** Não são fornecidos dados classificados. Para o processamento, o algoritmo identifica similaridades entre os dados de entrada e agrupa em classes aqueles que apresentam características em comum (MARSLAND, 2014).

### 1.3.1 *Support Vector Machine*

*Support Vector Machines* (SVM) são ferramentas de aprendizagem supervisionada que analisam dados e reconhecem padrões. Seus algortimos foram introduzidos em 1992, e desde então são bastante utilizados por seu potencial de obter classificadores com boa generalização e aplicáveis a grandes conjuntos de dados. São populares em áreas como categorização de textos, análise de imagens e bioinformática (LORENA; CARVALHO, 2007; MARSLAND, 2014).

Para formar um modelo de SVM, os dados de treinamento categorizados são distribuídos em um espaço em forma de pontos. Cada grupo de pontos é posicionada de modo que fique claramente separada das outras. Quando dados desconhecidos são inseridos, eles são posicionados no mesmo espaço, e classificados conforme sua localização dentre as categorias mapeadas anteriormente (KAPETANOVIC; ROSENFELD; IZMIRLIAN, 2004) (Figura 5).

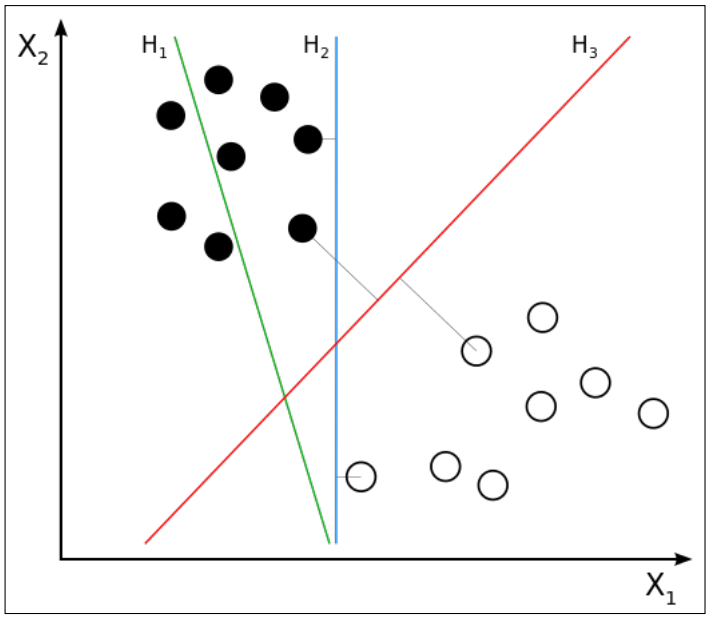


Figura 5 – O algoritmo do SVM tenta separar as classes presente no plano da forma mais eficiente possível. A figura ilustra três exemplos de retas que separam as classes de um plano. A reta H1 não separa as classes, portanto não é válida para predição por SVM. A reta H2 separa as classes, mas mantendo uma margem muito próxima a uma das classes, portanto não é tão eficiente. A reta H3 é a mais eficiente, pois separa as classes com a maior distância possível entre elas, portanto seria a utilizada pelo SVM.

As SVMs são utilizadas por algumas ferramentas de análise de dados de metagenômica. O primeiro algoritmo de classificação de sequências de metagenômica a utilizar SVM foi a ferramenta PhyloPythia. Esse algoritmo permite classificar fragmentos de metagenômica com mais de 1 kb para diversos níveis taxonômicos (domínio, classe, ordem e gênero) (MCHARDY et al., 2007). Outros estudos como o de Cui e Zhang (CUI; ZHANG, 2013) e de Garbarine e colaboradores (GARBARINE et al., 2011) também apresentam resultados que demonstram potencial que o SVM tem para a identificação taxonômica dos *reads*.

Entretanto, algoritmos de aprendizagem de máquina tradicionais - como as SVMs - apresentam alguns obstáculos para uma execução eficiente. A principal dificuldade é chamada de "maldição da dimensionalidade" (AREL; ROSE; KARNOWSKI, 2010): a complexidade computacional aumenta exponencialmente com o aumento linear do número de características (dimensionalidade dos dados). Para driblar essa dificuldade, os dados de entrada desses algoritmos devem ser pré-processados antes de inseridos na rede. Essa etapa permite a redução da dimensionalidade dos dados. A maioria dos *softwares* de análise de sequências utilizam os valores de frequências de oligonucleotídeos (e.g. *k-mers*) como dados de entrada. A seleção desses parâmetros não é uma tarefa trivial, uma vez que interferem de maneiras diferentes na eficácia da classificação (GARBARINE et al., 2011).

## 1.4 Proposta

A importância da composição microbiótica nos mais diversos ambientes se torna cada vez mais evidente com a o avanço das pesquisas e estimula o crescimento contínuo de dados de metagnômica. Entretanto, as análises desses dados esbarram em dificuldades computacionais que seriam dribladas com algoritmos que compensem rapidez e eficácia na identificação dos *reads* metagenômicos sem montagem prévia. Além disso, a maioria das ferramentas disponíveis falham em identificar sequências de micro-organismos desconhecidos, pois descartam os *reads* que não apresentarem similaridade com as sequências dos bancos de dados.

Deste modo, propomos o desenvolvimento de uma ferramenta computacional que realize o *binning* de metagenomas a nível de filo sem a necessidade de que sejam montados e que incluam sequências de micro-organismos ainda não identificados.

Desenvolvemos uma ferramenta de *binning* de composição de sequências com o algoritmo de aprendizagem de máquina SVM. Como parâmetros de entrada, testamos *k-mers* de 2, 3 e 4 nucleotídeos e outros parâmetros não tradicionais como entropias de di, tri e tetrapletes e abundância de dinucleotídeos.

# 2 Objetivos

O principal objetivo do projeto é o desenvolvimento de uma ferramenta que determina a origem taxonômica de *reads* obtidos por metagenômica a nível de filo sem a necessidade de pré-montagem dos metagenomas, utilizando aprendizado de máquina e sua relação funcional com os filos a que pertencem. Os objetivos secundários são:

1. identificação taxonômica dos fragmentos de metagenômica a partir de aprendizado de máquina;
2. identificação das assinaturas genômicas mais relevantes para a identificação taxonômica das sequências;
3. identificação dos filos a que pertencem sequências de espécies de micro-organismos ainda não catalogadas.

# 3 Métodos

## 3.1 Desenvolvimento da ferramenta

O algoritmo de *binning* dos *reads* de metagenômica foi desenvolvido na ferramenta Mathematica 8.0, com a linguagem de programação Wolfram Language e a ferramenta de aprendizagem de máquina Support Vector Machine (SVM).

### 3.1.1 *Support Vector Machine*

*Support Vector Machine* (SVM) são algoritmos de aprendizagem de máquina recomendados para a classificação de grandes quantidades de dados que podem ser linearmente separados e representados na forma de classificação binária. Por ser considerado um método supervisionado, classificará os fragmentos de metagenômica considerando modelos construídos a partir de sequências de referência (KUNIN et al., 2008; LIU et al., 2013).

Os SVM são considerados uma boa opção para a execução do *binning* por serem capazes de lidar com uma grande quantidade de informação discimilar. Além disso, a função discriminante é caracterizada por somente um pequeno conjunto comparativo do conjunto de dados inteiros, então realiza as computações notavelmente rápidas (KAPETANOVIC; ROSENFELD; IZMIRLIAN, 2004).

### 3.1.2 Escolha do Kernel

Na aprendizagem de máquina supervisionada, podemos considerar, em termos didáticos, que os dados de entrada fornecidos como treinamento são dispostos em uma superfície de acordo com os valores atribuídos a eles. Tradicionalmente, o SVM procura traçar uma linha reta que separa os dados conforme seus rótulos e permite que novos dados sejam alocados em uma das divisões, classificando-os de acordo com os dados similares. Essa é a chamada classificação linear de SVM (CRISTIANINI; SHAWE-TAYLOR, 2000).

Entretanto, em alguns casos, os dados seriam mais precisamente separados se a margem de separação entre as classes pudesse ser flexibilizada 6. Deste modo, foi desenvolvida uma nova estratégia que possibilitou a criação de diferentes classificadores não lineares, chamados *kernels*.

Existem diferentes tipos de *kernels*. Diferentes dados são melhor classificados com diferentes *kernels*. Testamos os kernels linear, radial (Radial Basis Function, RBF), polinomial e quadrático.

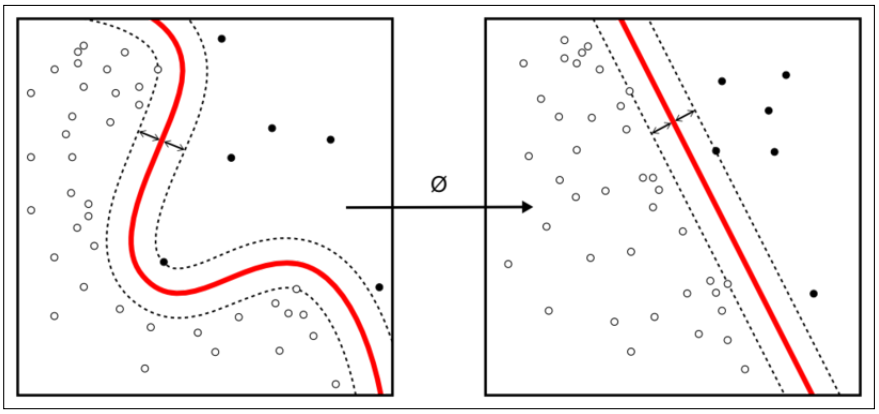


Figura 6 – A figura mostra a flexibilização de um *kernel* para que este possa ajustar-se de modo a separar os dados mais eficientemente.

### 3.1.3 Medidas

Para o treinamento das SVM supervisionado, é necessário fornecer para o algoritmo um conjunto de dados que contenha informações rotuladas para que ele possa reconhecer os padrões característicos de cada classe de entrada; esses dados são chamados de **dados de treinamento**. A partir do aprendizado obtido, a ferramenta tem a capacidade de classificar os *reads* desconhecidos dos metagenomas de acordo com os padrões identificados.

Utilizamos 9 parâmetros para a classificação dos *reads* de metagenômica e testamos a importância de cada um para a predição das sequências. Cada um está detalhado abaixo:

#### *Conteúdo GC*

A quantidade de guanina (G) e citosina (C) varia entre os genomas microorganismos de acordo com a pressão evolutiva no ambiente que vivem. A ligação G-C é mais estável do que a A-T por conter três pontes de hidrogênio, enquanto a ligação A-T apresenta somente duas. Deste modo, uma maior concentração de GC no genoma confere uma maior estabilidade do DNA, proporcionando maior vantagem aos habitantes de determinados ambientes como por exemplo, habitantes de altas temperaturas (CAMPBELL; MRAZEK; KARLIN, 1999; PASSEL et al., 2006; YAKOVCHUK; PROTOZANOVA; FRANK-KAMENETSKII, 2006). Entretanto, uma molécula mais estável pode ser desvantagem em casos em que os organismos necessitam replicar-se rapidamente, como é o caso de alguns vírus; nesses casos, seus genomas podem ter adquirido evolutivamente maior quantidade de AT pela maior facilidade de separar moléculas de duas pontes de hidrogênio do que três. Outros vírus patógenos, como por exemplo o influenza H1N1, apresentam supressão de GC como tentativa de assemelhar-se mais ao corpo humano, e assim driblar mais facilmente o sistema imunológico (GREENBAUM et al., 2008; GERHARDT et al., 2013).

Essas propriedades da porcentagem de guanina e citosina nos genomas a torna uma

característica apropriada para a diferenciação entre micro-organismos. Deste modo, conteúdo GC é uma medida de identificação de sequências amplamente utilizada (GERHARDT; CORSO, 2006) e também usaremos como dados de entrada do SVM para classificar as sequências de metagenômica.

### *Frequência de oligonucleotídeos (k-mers)*

As frequências de oligonucleotídeos são parâmetros com alto potencial para distinguir genomas, pois seus valores variam significativamente entre as sequências dos diferentes organismos (sendo eles procariotos ou eucariotos). Essas frequências são biologicamente relevantes, pois podem influenciar em fatores como na estabilidade do DNA, restrições na helicidade, modificações na metilação, pressões seletivas dependente do contexto e mecanismos de replicação e reparo. Além disso, o fato de que os valores são constantes ao longo do genoma permite que essas características possam ser utilizadas na identificação de fragmentos de DNA obtidos por metagenômica (KARLIN; MRAZEK; CAMPBELL, 1997; ABE et al., 2003; TAKAHASHI; KRYUKOV; SAITOU, 2009).

Utilizamos frequências dinucleotídeos, trinucleotídeos e tetranucleotídeos como parâmetro de identificação taxonômica das sequências.

### *Abundância de Dinucleotídeos*

A frequência de dinucleotídeos tem grande influência sobre a estrutura da molécula de DNA, podendo interferir na sua interação com proteínas, principalmente as que atuam no processo de replicação e tradução (KARLIN; BURGE, 1995). Estes processos são vitais para a célula, portanto sofrem grande pressão de seleção, o que passa a acentuar a diferença dos padrões entre as espécies e entre os organismos de diferentes ambientes (DUTTA; PAUL, 2012), aumentando sua eficiência como parâmetro de identificação de sequências.

A abundância relativa de dinucleotídeos é a medida mais utilizada para calcular frequência. É representada pela equação 3.1 (KARLIN; BURGE, 1995; KARLIN, 1998):

$$\rho_{XY} = \frac{f_{XY}}{f_X f_Y} \tag{3.1}$$

onde $f_{XY}$ representa a frequência do dinucleotídeo $XY$ na sequência, $f_X$ representa a frequência do nucleotídeo $X$ e $f_Y$ a frequência do nucleotídeo $Y$. Essa equação resulta na diferença entre a frequência observada do dinucleotídeo $XY$ na sequência e sua frequência esperada em uma sequência aleatoria que possua a mesma quantidade dos nucleotídeos $X$ e $Y$ (KARLIN; BURGE, 1995; KARLIN, 1998; DUTTA; PAUL, 2012).

O valor da abundância de dinucleotídeos é obtido pela média da abundância relativa de cada dinucleotídeo, como representado na equação 3.2.

$$din = \frac{1}{16} \sum_X \sum_Y \frac{f_{XY}}{f_X f_Y},\ X \in \{\mathrm{A, C, T, G}\},\ Y \in \{\mathrm{A, C, T, G}\} \tag{3.2}$$

### *N-entropia*

A palavra *entropia* conota desordem ou incerteza. A métrica de entropia de um sistema é valor da desordem ou incerteza inerente neste sistema (HANKERSON; JOHNSON; HARRIS, 1998). Neste caso a entropia representa o padrão de desordem na distribuição de oligonucleotídeos em uma sequência. Neste trabalho aplicamos medidas de entropia de dipletes ($S_2$), tripletes ($S_3$) e tetrapletes ($S_4$).

Para o cálculo da entropia, é necessário primeiramente estimar a probabilidade de cada oligonucleotídeo possível na sequência ($P_n$). Para a obtenção desse valor, conta-se a quantidade de vezes que cada oligonucleotídeo em uma janela deslizante aparece na sequência e divide-se esse número pela quantidade total de nucleotídeos do segmento (GERHARDT; CORSO, 2006; SANTOS; RYBARCZYK-FILHO; GERHARDT, 2011).

O cálculo de entropia é realizado a partir da equação da entropia de Shannon (Equação 3.3) (SHANNON, 1948):

$$S = -\frac{1}{\ln N} \sum_{n=1}^{N} P_n \log P_n, \tag{3.3}$$

$P_n$ simboliza a probabilidade do e-nésimo oligonucleotídeo presente em uma sequência de tamanho $W$. $N$ representa a quantidade de oligonucleotídeos possíveis, sendo 16 para dinucleotídeos, 64 para trinucleotídeos e 256 para tetranucleotídeos.

### *TETRA*

Como discorrido anteriormente, pode-se encontrar padrões espécie-específicos em frequências de DNA dos micro-organismos. Dentre os vários tamanhos de oligonucleotídeos dos quais podem ser calculadas as frequências, os de tamanho 4 (tetranucleotídeos) são propriados para a busca de padrões nos genomas por compensarem uma resolução apropriada com efetividade computacional (TEELING et al., 2004). Entretanto, a medida da frequência de tetranucleotídeos como apresentada originalmente sofre a influência de vieses das frequências de oligonucleotídeos menores, como mono, di e trinucleotídeos. Para desprezar esses vieses, foi desenvolvida uma medida alternativa: a frequência de tetranucleotídeos derivada de correlações de Z-score, ou, TETRA (TEELING et al., 2004).

Para calcular o TETRA, primeiramente calcula-se as frequências observadas dos 256 tetranucleotídeos possíveis e as frequências esperadas correspondentes. As diferenças entre os valores esperado e observado foram transformados em *z-scores* de cada tetranucleotídeo. Considerando $N_{(n1n2n3)}$ a frequência de um tetranucleotídeo, a equação de *z-scores* está representada em 3.4:

$$Z_{(n1n2n3n4)} = \frac{N_{(n1n2n3n4)} - E_{(n1n2n3n4)}}{\sqrt{var(N_{(n1n2n3n4)})}} \tag{3.4}$$

sendo que:

$$E_{(n1n2n3n4)} = \frac{N_{(n1n2n3)} N_{(n2n3n4)}}{N_{(n2n3)}} \tag{3.5}$$

$$var(N_{(n1n2n3n4)}) = E_{(n1n2n3n4)} * \frac{[N_{(n2n3)} - N_{(n1n2n3)}][N_{(n2n3)} - N_{(n2n3n4)}]}{{N_{(n2n3)}}^2} \tag{3.6}$$

## 3.2 Análise do desempenho da ferramenta

### 3.2.1 Genomas

Foi realizado o download de genomas completos de Bacteria, Archaea e Fungi do banco de dados GenBank (maio de 2015) (DATABASE. . . , 2013).

Os genomas de Bacteria foram agrupados em filos de acordo com a classificação taxonômica do NCBI. O filo Proteobacteria apresenta um alto número de espécies disponíveis, portanto, foi dividido entre as classes: Alphaproteobacteria, Betaproteobacteria, Deltaproteobacteria, Epsilonproteobacteria e Gammaproteobacteria. Fungi e Archaea apresentavam sequências de poucos filos disponíveis, portanto suas sequências só foram classificadas a nível de "Fungi" e "Archaea". A divisão dos genomas dos micro-organismos pode ser observado na tabela 2. Os filos que continham menos que 2 espécies representantes foram excluídos.

### 3.2.2 Grupos de simulações

Fragmentamos aleatoriamente os genomas obtidos em tamanhos de 100, 400 e 1000 pares de base, para avaliar a eficácia da predição para os três tamanhos de sequência, e os dividimos em grupos de treinamento e teste conforme a tabela 3.

Nas simulações realizadas com o grupo 1 e 2, avaliamos a capacidade do preditor em identificar corretamente o filo de sequências de micro-organismos conhecidos pelo preditor. Na simulação realizada com o grupo 3, avaliamos o desempenho da ferramenta em identificar o filo de sequências não conhecidas pelo algoritmo. Esse teste tem o intuito de analisar o comportamento do algoritmo em situações em que os micro-organismos a serem identificados não são conhecidos ou não tem sua sequência disponível em banco de dados.

Para cada grupo, realizamos simulações que excluíam uma medida de cada vez para selecionar as medidas que influenciam positivamente a predição e descartar as que reduzem sua eficácia. Foram realizadas 120 simulações com 500 sequências de exemplo positivo e 500 de exemplo negativo para cada grupo com cada medida excluída. Os valores de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos foram computados para a avaliação estatística a partir de valores AUC.

Tabela 2 – Tabela de classificação dos *reads*.

| Domínio / Reino | Filo | Classe |
|---|---|---|
| Eucaria / Fungi | - | - |
| Archaea | - | - |
| Bacteria | Acidobacteria | - |
| | Actinobacteria | - |
| | Aquificae | - |
| | Bacteroidetes | - |
| | Chlamydiaceae | - |
| | Chlorobi | - |
| | Cyanobacteria | - |
| | Deferribacteres | - |
| | Deinococcus-Thermus | - |
| | Firmicutes | - |
| | Fusobacteria | - |
| | Nitrospirae | - |
| | Planctomycetes | - |
| | Proteobacteria | Alphaproteobacteria |
| | | Betaproteobacteria |
| | | Deltaproteobacteria |
| | | Epsilonproteobacteria |
| | | Gammaproteobacteria |
| | Spirochaetes | - |
| | Synergistetes | - |
| | Tenericutes | - |
| | Thermotogae | - |
| | Verrumiccobia | - |

Tabela 3 – Tabela dos grupos de treinamento e teste.

| **EXEMPLOS POSITIVOS** | Grupo 1 | Grupo 2 | Grupo 3 |
|---|---|---|---|
| Treinamento | Sequências de Gammaproteobacteria | Sequências de Gammaproteobacteria | Sequências de Gammaproteobacteria exceto *E. coli* |
| Teste | Sequências de *E. coli* | Sequências de Gammaproteobacteria | Sequências de *E. coli* |
| **EXEMPLOS NEGATIVOS** | Os exemplos negativos para treinamento e teste dos três grupos foram sequências aleatórias de micro-organismos de outros táxons que não Gammaproteobacteria. | | |

### 3.2.3 Utilização do valor AUC como parâmetro de comparação entre os classificadores

Utilizamos o valor de AUC como critério para determinar a eficácia da ferramenta. AUC (sigla para *Area Under Curve*, ou, "área abaixo da curva" em português), refere-se ao valor da área abaixo da curva ROC. A curva ROC é representada por um gráfico cujo eixo

Y apresenta valores do índice de verdadeiros positivos e o eixo X contém o índice de falsos positivos. Os valores obtidos em cada simulação são plotados entre os eixos, dando origem ao gráfico de curva ROC (FAWCETT, 2006). A curva ROC é amplamente utilizada na avaliação da acurácia de preditores de aprendizagem de máquina por fornecer informações sobre o custo-benefício de cada classificador (HAJIAN-TILAKI, 2013). O valor AUC é a quantificação da curva ROC para que seja possível comparar diferentes classificadores.

# 4 Resultados e Discussão

## 4.1 Escolha do Kernel

Podemos observar na figura 7 que os resultados das simulações com o kernel RBF apresentaram os valores de AUC mais elevados do que dos outros kernels, indicando que as simulações de SVM que utilizam esse kernel classificam mais sequências corretamente do que as simulações com outros. Levando em conta que as diferenças entre o kernel RBF e os outros é significativa, como mostra a Tabela 4, chegamos à conclusão que este é o kernel mais apropriado para nosso conjunto de dados.

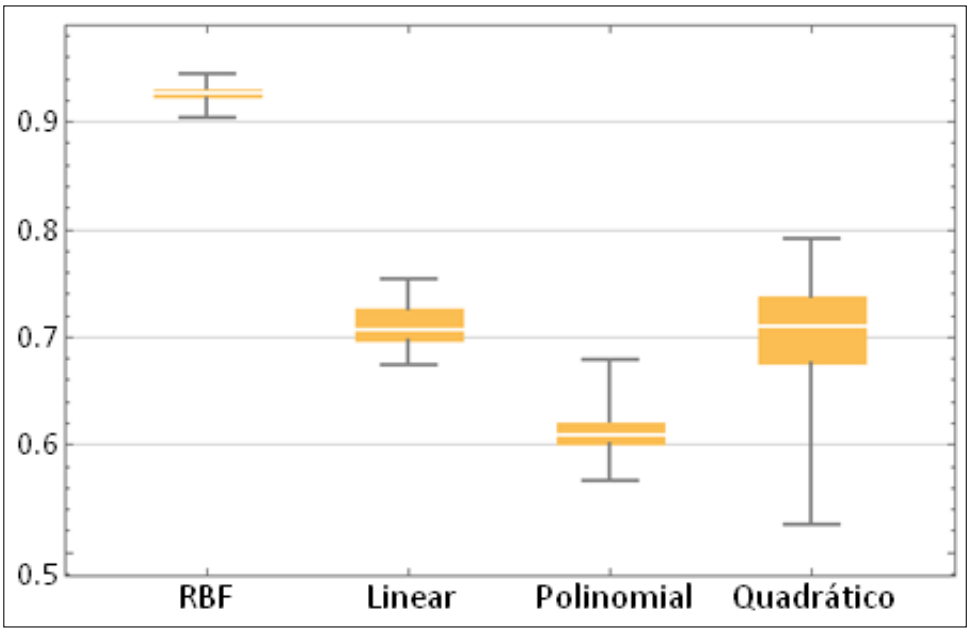


Figura 7 – Distribuição dos valores de AUC para os kernels RBF, linear, polinomial e quadrático. Observamos que os valores para o kernel RBF estão mais elevados do que os valores dos kernels linear, polinomial e quadrático.

Tabela 4 – Valores do teste de normalidade para os resultados das simulações realizadas com cada um dos kernels. Consideramos diferenças significativas as que apresentaram valores menores do que 0.05 no teste de normalidade. Os valores em verde-claro representam diferenças significativas, e os valores em cinza representam que não há diferença estatística entre os kernels.

| | RBF | Linear | Polinomial | Quadrático |
|---|---|---|---|---|
| RBF | | $2.56 \times 10^{-34}$ | $2.70 \times 10^{-149}$ | $2.56 \times 10^{-34}$ |
| Linear | $2.48 \times 10^{-34}$ | | $3.06 \times 10^{-34}$ | 0.819 |
| Polinomial | $2.70 \times 10^{-149}$ | $2.99 \times 10^{-34}$ | | $6.14 \times 10^{-17}$ |
| Quadrático | $2.48 \times 10^{-34}$ | $6.14 \times 10^{-17}$ | 0.821 | |

## 4.2 Comparação entre tamanhos de *reads*

Executamos simulações de predição de filos para três tamanhos de sequência: 100, 400 e 1000 pares de base para analisar o desempenho da ferramenta ao predizer *reads* de diferentes sequenciadores. Podemos observar no gráfico da Figura 8 que a ferramenta teve um índice de AUC de quase 100% para todos os tamanhos testados no Grupo 1, cujas sequências de teste de exemplo positivo pertencem à espécie *Escherichia coli*. As SVM são conhecidas por seu alto poder de predição de dados de classificação binária, portanto não teve dificuldade nem de identificar que todas as sequências positivas de teste, originadas de *E. coli* no caso do Grupo 1, pertenciam ao táxon Gammaproteobacteria e nem que as de outros filos não pertenciam.

Os resultados do Grupo 2, que contém como exemplo positivo de teste sequências do táxon Gammaproteobacteria, mostram que o poder preditivo para um grupo mais amplo também é alto, para os três tamanhos de sequência testados, com uma média de, aproximadamente, 0,85 de AUC.

O Grupo 3 testa o poder de predição da ferramenta para sequências de espécies desconhecidas pelo preditor. Nesse caso, utilizamos como exemplos positivos sequências de todas espécies micro-organismos do táxon Gammaproteobacteria exceto da espécie *E. coli*. Deste modo, utilizando sequências de *E. coli* como exemplos positivos a ser identificados, testamos o poder preditivo da ferramenta ao classificar sequências de espécies desconhecidas no filo correto. No gráfico da Figura 8 observamos que as simulações obtiveram um poder preditivo com média de mais de 85% para sequências de 1000 pb e de mais de 75% para sequências de 100 pb.

Os valores de AUC das simulações com sequências do Grupo 3 mostram que a ferramenta é capaz de classificar os *reads* corretamente mesmo para sequências de menor tamanho, como as de 100 pb, que apresentam resolução mais baixa.

## 4.3 Teste de exclusão de medidas

As simulações realizadas para comparação dos tamanhos de *reads* foram executadas com as medidas de sequência citadas na seção de Métodos. Para avaliar se a medida está realizando influência positiva ou negativa na predição, executamos simulações que excluem uma medida de cada vez e comparamos os resultados com os valores de AUC obtidos quando utilizadas todas as medidas. Os resultados para todos os grupos de sequência e tamanhos estão ilustrados na Figura 9.

Observamos nos gráficos da Figura 9 que, assim como nos resultados descritos na seção anterior, os valores de AUC para sequências do Grupo 1 dos três tamanhos permanecem em torno de 1,0.

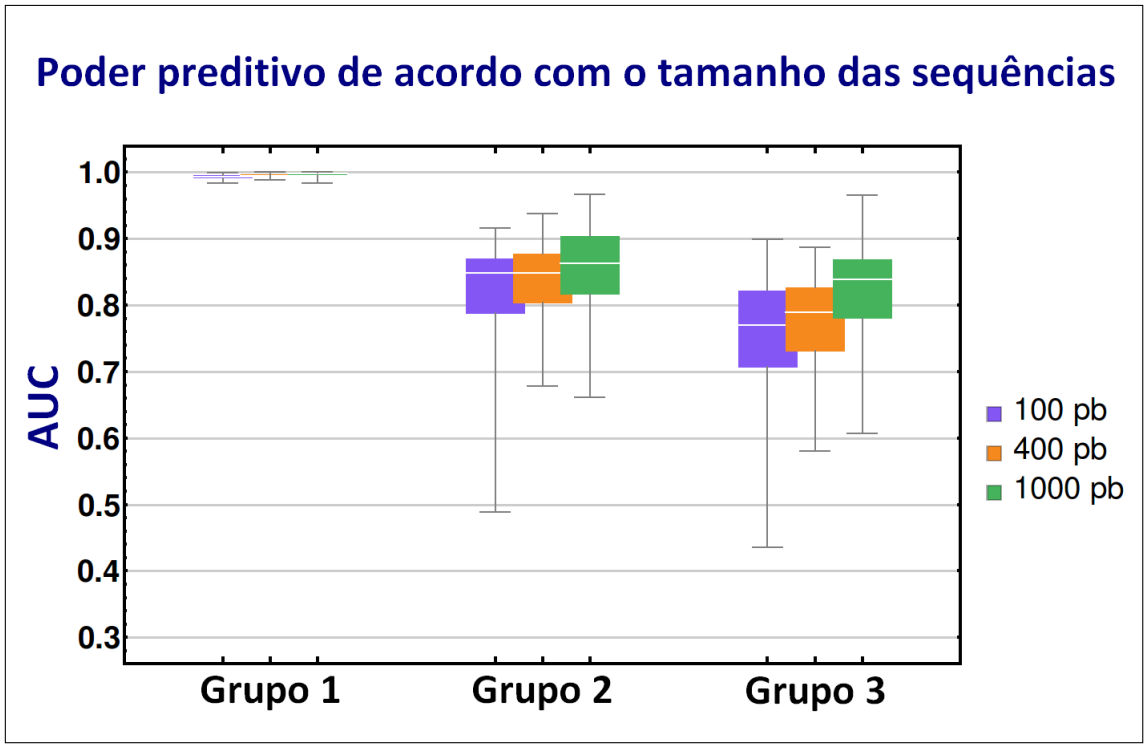


Figura 8 – Gráfico de comparação do poder de predição da ferramenta. Para os três grupos de sequências foram testados os tamanhos de 100, 400 e 1000 pares de base.

Para sequências de 100 pb do Grupo 2, observamos que quando a medida TETRA não é utilizada o poder preditivo da ferramenta cai consideravelmente: de uma média de 0,85 para, em torno de 0,55. Para o Grupo 3, a identificação de sequências de 100 pb também tem sofre influência positiva de TETRA, pois a média de seu valor de AUC cai para aproximadamente 0,45 quando retiramos a medida.

Essa mesma influência positiva é observada também para os Grupos 2 e 3 com sequências de tamanhos 400 e 1000 pb: embora menos acentuada, a queda no poder de predição com a exclusão da medida TETRA ainda é visível. Em contrapartida, observamos um aumento nos valores de AUC dessas simulações com a exclusão da medida AB.

A Figura 10 ilustra os resultados dos valores de AUC para os três grupos para sequências de 400 pb com a comparação estatística dos resultados. Realizamos o teste de Mann-Whitney e comparamos cada grupo de medidas excluídas com o grupo com todas as medidas. Consideramos estatisticamente diferentes aqueles que apresentaram valor de $p$ maior do que 0.05. Os mesmos estão representados na Figura 10 por um asterisco ($*$). A partir da análise estatística observamos que quando retiradas as medidas AB e E2 os valores de AUC aumentam para os grupos 2 e 3. Também foi estatisticamente comprovada a influência positiva das medidas GC para o grupo GC, K3 para o grupo 4 e TETRA e K4 para ambos os grupos 3 e 4.

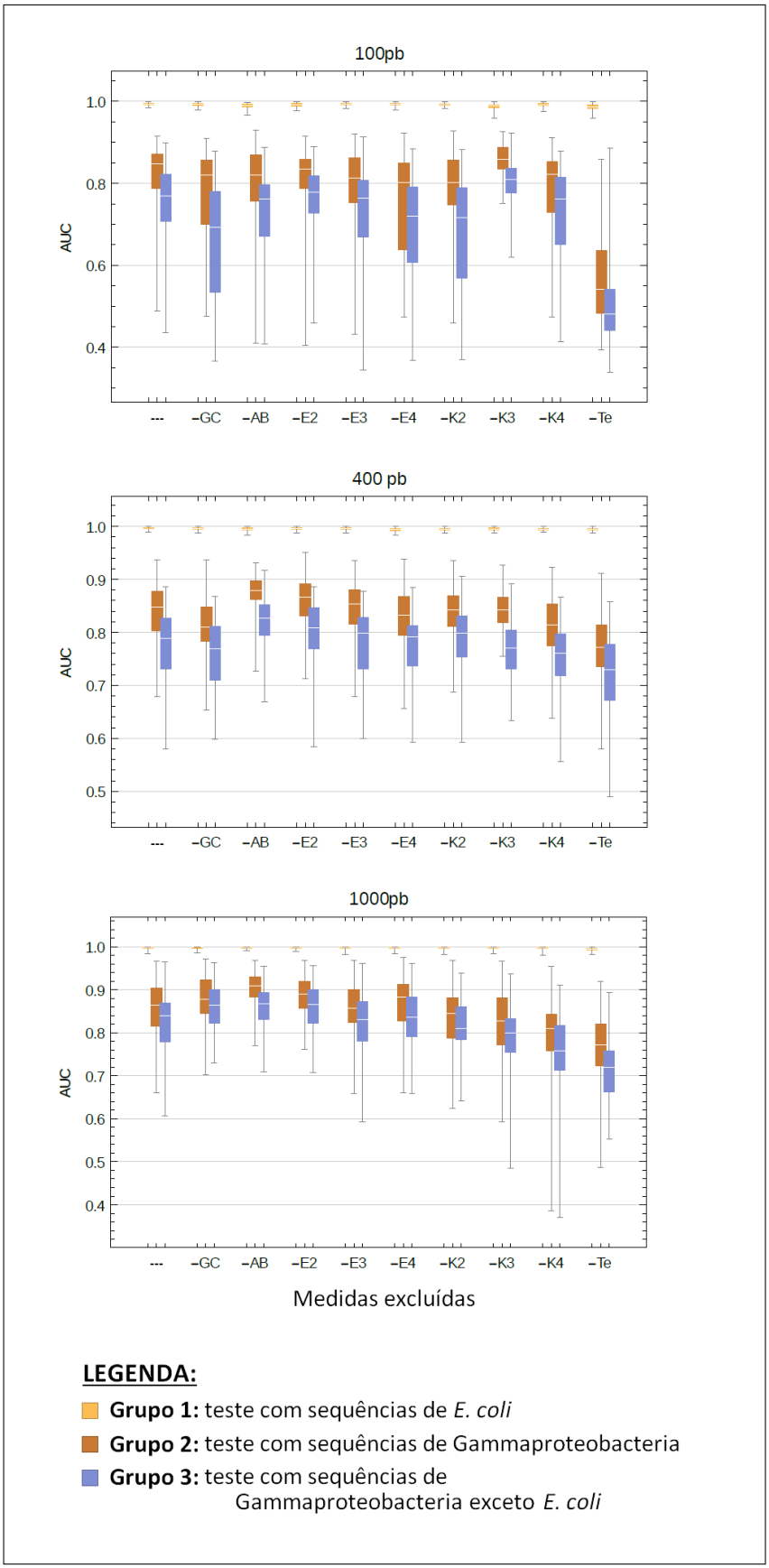


Figura 9 – Gráficos com os valores de AUC com a exclusão de uma medida de cada vez para sequências de tamanhos 100, 400 e 1000 pb. Os gráficos indicados com "—" representam os resultados das simulações que não excluíram nenhuma medida. Os gráficos seguintes são indicados pelas medidas que foram excluídas.

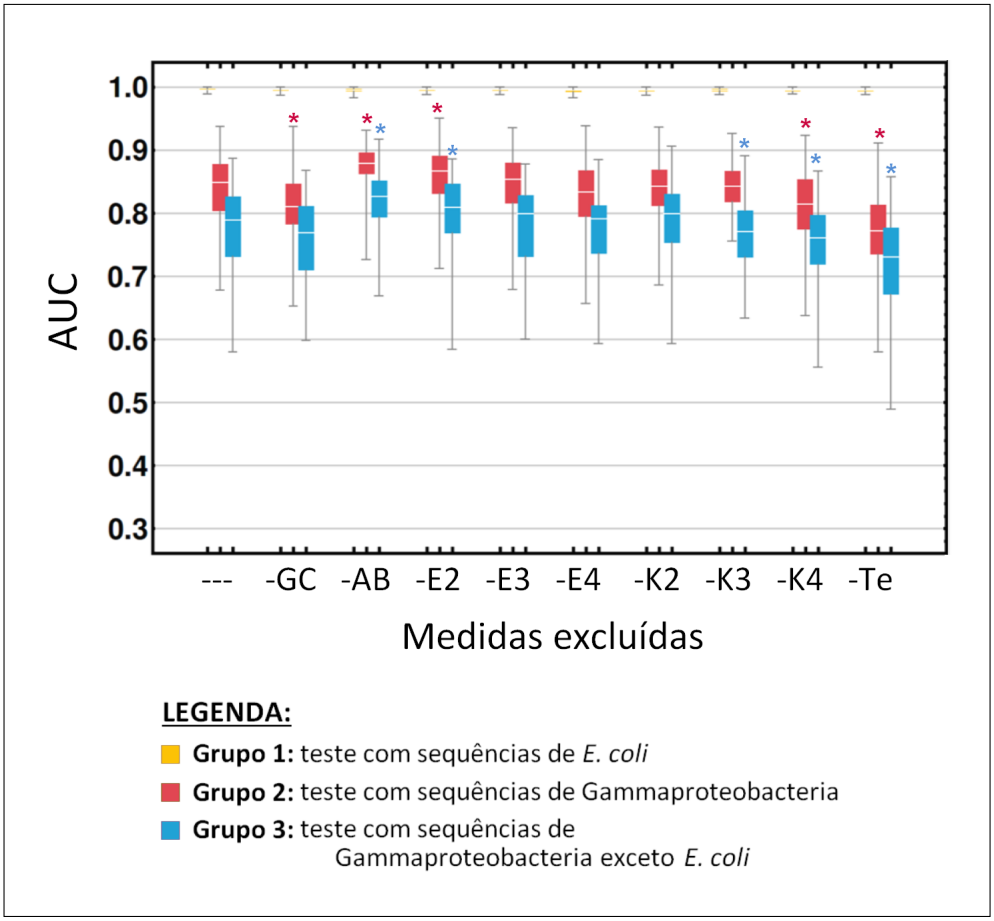


Figura 10 – Poder preditivo dos três grupos de treinamento para sequências de 400 pares de base. Os testes foram executados no SVM utilizando todas as medidas de DNA ("–") ou excluindo uma medida por vez para testar a relevância de cada uma para a predição. O teste estatístico Mann-Whitney foi aplicado nos valores para detectar diferenças estatísticas entre o grupo que não exclui medidas e os testes com medidas excluídas. O $*$ indica os grupos que apresentaram diferenças estatísticas ($p > 0.05$) em comparação com o primeiro grupo.

Esses resultados nos mostram que a medida TETRA é indispensável para uma classificação acurada das sequências. A medida K4 também ajuda a elevar o AUC das predições, mas menos evidentemente. O TETRA utiliza valores de frequências de tetranucleotídeos, mas despreza os ruídos intrínsecos das frequências de mono, di e tri nucleotídeos. A desconsideração desses fatores nas medidas de tetranucleotídeos mostrou-se importante para a predição principalmente para as sequências de tamanho 100 pb. As sequências de menor tamanho naturalmente apresentam menos informação contida, portanto o cálculo de medidas sem ruídos mostraram-se importantes para o aproveitamento das informações contidas nela pelo SVM. Em compensação, a inclusão dos valores de abundância de dinucleotídeos (AB) na predição prejudicou a classificação das sequências, e o resultado é mais claro para as sequências de 1000 pb. Deste modo, os valores de AB não devem ser utilizados para garantir uma classificação mais precisa das sequências de metagenômica.

# Parte II

# Parte 2: Análise da interação hospedeiro-microbioma do metaproteoma da doença de Crohn

# 5 Introdução

## 5.1 Doença de Crohn

A doença de Crohn (DC) é uma das duas doenças inflamatórias intestinais (DII), juntamente com colite ulcerosa, causada por inflamação na mucosa do trato gastrointestinal, principalmente no íleo terminal, porção distal do intestino delgado (HENDERSON; STEVENS, 2012). O desenvolvimento da doença é atribuído por diversos fatores genéticos, imunológicos e ambientais que, atuando juntos, causam a inflamação crônica das células intestinais(BAUMGART; SANDBORN, 2012). Embora fatores genéticos intrínsecos foram considerados os maiores responsáveis pelo DC nas primeiras décadas de pesquisa, eles só explicavam uma pequena fração da hereditabilidade da doença (MONDAL; KUGATHASAN, 2017), o que ressaltou a importância dos fatores extrínsecos na etiologia de DC. Não obstante, dependendo do grau de predisposição genética caracterizada pela ocorrência de mutações e da profundeza dos fatores extrínsecos, a severidade da doença evolui para uma forma extrema.

Fatores extrínsecos como hábitos alimentares, drogas, poluição, radiação, estresse, exposição a antibióticos e higiene modulam a funcionalidade do microbioma intestinal, influenciando na suscetibilidade do indivíduo de desenvolver DC e em sua intensidade (BAUMGART; SANDBORN, 2012; SHEEHAN; SHANAHAN, 2017). O fato de que os protagonistas no desenvolvimento de DC são os micro-organismos é corroborado pelo fato de que ratos livres de micro-organismos não desenvolvem nenhuma das DII (STANGE; WEHKAMP, 2016). Além disso, há uma notória diferença entre a microbiota intestinal de pacientes saudáveis dos com DC. Indivíduos com doença de Crohn apresentam uma comunidade microbiana disbiótica, com uma baixa diversidade $\alpha$ comparada à indivíduos saudáveis (PASCAL et al., 2017). Essa alteração desencadeia uma resposta imune agressiva à microbiota modificada, resultando na inflamação (MANICHANH et al., 2012). Deste modo, podemos considerar que a disbiose intestinal é o elo entre os elementos extrínsecos e a resposta imune do hospedeiro (MANICHANH et al., 2012).

### 5.1.1 Microbiota intestinal na doença de Crohn

Uma das diferenças mais significativas entre a microbiota intestinal de indivíduos saudáveis e com DC é a quantidade diminuída de bactérias fibrolíticas em pacientes com doença de Crohn. Essas espécies realizam a fermentação de polissacarídeos em ácidos graxos de cadeia curta (AGCC): moléculas com propriedades imunomodulatórias utilizadas

pelo epitélio do cólon como principal fonte de energia. Dentre as bactérias produtoras de AGCC estão *Faecalibacterium prausnitzii*, *Roseburia* e *Oridobacter*. Essas espécies são conhecidas também pelas suas propriedades anti-inflamatórias e são sub-representadas em pacientes com DC (MORGAN et al., 2012; SHEEHAN; SHANAHAN, 2017). Outra diferença significativa entre os microbiomas é a alta concentração da bactéria patogênica *Salmonella* e linhagens aderentes e invasivas de *Escherichia coli* (AIEC) em intestinos de pacientes com DC. Essas bactérias adquirem vantagem competitiva na presença de estresse oxidativo ocasionado pela mucosa inflamada e não são combatidas pelo sistema imune por causa dos defeitos genéticos presentes em indivíduos com DC (MORGAN et al., 2012; SHEEHAN; SHANAHAN, 2017). Firmicutes, Bacteroidetes e outras espécies do gênero *Fecalibacterium* também estão presentes em números reduzidos no microbioma intestinal de pacientes com doença de Crohn; Proteobacteria, *Shigella* e Fusobacteriaceae são encontradas em maior abundância (MANICHANH et al., 2006; REHMAN et al., 2010; RICANEK et al., 2012; THORKILDSEN et al., 2013; GEVERS et al., 2014; CRUZ et al., 2015).

A disbiose do microbioma intestinal pode ser ocasionada por defeitos nas células de Paneth, células responsáveis por controlar a composição e densidade da microbiota intestinal (YANO; KURATA, 2009). As células de Paneth são células epiteliais presentes na cripta do intestino delgado com a função de proteger o epitélio contra a entrada de micro-organismos. Elas são especializadas em produzir peptídeos antibacterianos, $\alpha$-defensina HD5, um peptídeo com propriedades antibióticas e $\alpha$-defensina HD6, que forma redes na cripta do instestino pra restringir a mobilidade das bactérias (SHI, 2007; STANGE; WEHKAMP, 2016). As células de Paneth de indivíduos com doença de Crohn apresentam sua função comprometida, possibilitando o contato dos micro-organismos com o epitélio intestinal e ocasionando a inflamação (STANGE; WEHKAMP, 2016).

### 5.1.2 Susceptibilidade genética à doença de Crohn

Os defeitos na funcionalidade das células de Paneth estão associados a variações em genes relacionados ao reconhecimento microbiano, sistema imune inato e autofagia. A partir de estudos de associação ampla do genoma (GWAS), observou-se que mutações nesses genes são apresentadas por indivíduos com alta susceptibilidade genética à doença de Crohn (YANO; KURATA, 2009; HENDERSON; STEVENS, 2012; MICHAIL; BULTRON; DEPAOLO, 2013). Nod2 é um dos genes cujo variante representa aumento no risco de desenvolvimento de DC, e está envolvido na secreção de peptídeos antibicrobianos em resposta a estímulos bacterianos por ele reconhecidos (SHI, 2007; STANGE; WEHKAMP, 2016). Esse gene é altamente expresso pelas células de Paneth, cuja secreção de $\alpha$-defensina é comprometida em pacientes com mutação em Nod2 (SHI, 2007; YANO; KURATA, 2009).

Outro gene importante para o aumento da susceptibilidade de DC é o Atg15L1, proteína relacionada à autofagia, processo importante no controle da microbiota intestinal (SHI, 2007). Ratos com mutações no gene Atg15L1 apresentam anormalidades nas células de Paneth incluindo diminuição na produção de lisozimas, grânulos reduzidos e desorganizados, mitocôndria degenerada e falta de microvilosidades apicais (YANO; KURATA, 2009; HENDERSON; STEVENS, 2012). Além disso, macrófagos deficientes em Atg16L1 produziram altas quantidades das interleucinas pró-inflamatórias interleucina 1-beta (IL-1$\beta$) e interleucina 18 (IL-18), contribuindo para o aumento da inflamação do tecido intestinal (SAITOH et al., 2008). A mutação do gene Atg15L1 também impede a realização de autofagia pelos macrófagos (SHI, 2007; YANO; KURATA, 2009). A autofagia é um processo celular homeostático responsável pela degradação de moléculas, organelas e bactérias. A autofagia basal tem baixa atividade em células saudáveis, mas é ativada em situações de estresse como infecção de patógenos, estresse oxidativo e baixa disponibilidade de nutrientes. Em pacientes com doença de Crohn, os macrófagos não realizam autofagia nem no nível mais basal, favorecendo o desbalanço da microbiota intestinal (YANO; KURATA, 2009).

Além dos variantes citados, existem mais de 163 genes de risco para DC identificados. Alguns dos mais relevantes clinicamente estão apresentados na tabela 5.

### 5.1.3 Metadados

A maioria dos genes de susceptibilidade à doença de Crohn foram obtidos por estudos de associação ampla do genoma (GWAS) (HENDERSON; STEVENS, 2012). GWAS é a análise de como variantes genéticos estão distribuídos entre indivíduos de diferentes populações. A partir dessas análises, é possível determinar associações entre os variantes e diferentes doenças (NORRGARD, 2008), como no caso dos genes de susceptibilidade à doença de Crohn citados na seção anterior.

Enquanto as informações sobre a susceptibilidade genética do indivíduo com doença de Crohn são obtidas a partir de GWAS, informações do microbioma intestinal são obtidas pela exploração de tecnologias como 16S rRNA, metagenômica *whole genome shotgun* (WGS) e metatranscriptômica. A identificação das características relacionadas à microbiota de indivíduos com DC são úteis para inferências primárias associadas à doença. Entretanto, análises de metagenômica e metatranscriptômica excluem informações essenciais sobre a presença ou ausência de proteínas microbianas, que são essenciais para o estudo completo da interação microbioma-hospedeiro e desempenham papel fundamental no desenvolvimento de DC. Por outro lado, estudos metaproteômicos fornecem um panorama mais completo a respeito das proteínas expressas na comunidade microbiana do indivíduo, como por exemplo informações sobre a diferença entre as proteínas microbianas de indivíduos saudáveis e indivíduos com doença de Crohn (ERICKSON et al., 2012;

Tabela 5 – Lista de genes de risco para doença de Crohn (MICHAIL; BULTRON; DEPAOLO, 2013).

| Gene | Função |
|---|---|
| Nod2 | ⋆ Resposta do sistema imune inato |
| Atg16L1 | ⋆ Homeostase da mucosa e degradação de componentes celulares (autofagia) |
| IL23R | ⋆ Receptor de citocina tipo 1 |
| IBD5 | ⋆ Citocina associada ao risco de desenvolvimento de doença de Crohn |
| TLR4 | ⋆ Receptor *Toll-like*. Detecta lipopolissacarídeos de bactérias gram-negativas e ativa o sistema imune |
| OCTN1 | ⋆ Proteína presente na membrana plasmática responsável pelo cotransporte de íons de sódio e ergotioneína |
| IRGM | ⋆ Regula a autofagia em resposta a patógenos intracelulares |
| DLG5 | ⋆ Proteína importante em sítios de contato intercelular |
| LRRk2 | ⋆ Substrato para autofagia mediada por chaperonas |
| PTPN22 | ⋆ Influencia na resposta dos receptores de celulas T e B |
| IL10R | ⋆ Media o sinal imunosupressor da interleucina 10, inibindo a síntese de citocinas pró-inflamatórias |
| TNF | ⋆ Citocina envolvida na inflamação sistêmica |

PRESLEY et al., 2012; JUSTE et al., 2014). Entretanto, apesar desses avanços tecnológicos, não há esclarecimento suficiente dos mecanismos moleculares e vias de sinalização do hospedeiro que são mediadas por proteínas microbianas, seja de indivíduos saudáveis ou doentes.

## 5.2 Proposta

Nesse estudos propomos a utilização de uma abordagem computacional integrada que acessa o impacto potencial de proteínas microbianas diferencialmente presentes em pacientes com doença de Crohn e indivíduos saudáveis. Utilizaremos assinaturas de interação baseadas nas características estruturais das proteínas para predizer a ligação entre receptores de proteínas humanas extracelulares e proteínas microbianas expressadas unicamente em indivíduos com DC ou saudáveis.

O conjunto de dados de proteínas microbianas foi obtido de um estudo sueco (ERICKSON et al., 2012) que identifica diferenças metaproteômicas entre duplas de

gêmeos dos quais um deles fora diagnosticado com DC e o outro não. Comparando os pares de gêmeos entre si, será possível desprezar a variabilidade genética como principal fator determinístico contribuindo para a doença. Também determinaremos vias de sinalização afetadas pela ligação das proteínas microbianas nos receptores humanos. Utilizando testes de enriquecimento e informação *a priori*, seremos capazes de identificar cadeias sinalizadoras canônicas e não canônicas em particular relacionadas a processos modulados unicamente pelas proteínas encontradas em indivíduos com DC.

# 6 Objetivos

O principal objetivo do presente projeto é utilizar diferentes tipos de dados incluido metaproteômica de indivíduos com DC, vias de sinalização e redes de interação, para identificar e caracterizar os possíveis mecanismos moleculares mediados pela microbiota disbiótica em DC. Os objetivos secundários são:

1. analisar se e como as proteínas bacterianas de pacientes de DC podem modular a autofagia;
2. analisar se as mutações genéticas de indivíduos com DC influenciam nas interações entre as proteínas microbianas e humanas;
3. identificar proteínas microbianas chaves que podem induzir o desenvolvimento de DC.

# 7 Métodos

## 7.1 Metaproteomas e dados de proteoma humanos de indivíduos com doença de Crohn e saudáveis

Os conjuntos de dados foram obtidos do estudo "Integrated Metagenomics Metaproteomics Reveals Human Host-Microbiota Signatures of Crohn's Disease" (ERICKSON et al., 2012). Nesse estudo, o grupo de Erickson obtiveram metaproteomas e metagenomas de amostras de fezes de seis pares de pacientes com doença de Crohn ileal (IDC) ou colônica (CDC). Um dos gêmeos do par tinha doença de Crohn e o outro era saudável. Foram observadas variações entre o metabolismo bacteriano, interações microbioma-hospedeiro e enzimas secretadas pelo hospedeiro quando as amostras de indivíduos com DC foram comparadas com as de indivíduos saudáveis.

## 7.2 Obtenção das vias de sinalização

Para avaliar como proteínas bacterianas podem potencialmente interferir nas proteínas de autofagia humana através de uma via de sinalização, primeiramente compilamos redes de interações proteína-proteína (PPIs) e de interação através de regulação transcricional (TRIs)

### 7.2.1 Bancos de dados de PPIs e TRIs humanos

As PPIs e TRIs foram obtidas dos seguintes bancos de dados:

**Autophagy Regulatory Net (ARN):** Um banco de dados contendo informação manualmente curada sobre vias de sinalização de autofagia. Em ARN, é possível obter quatro tipos de interação: PPIs, TRIs, interações miRNA-mRNA e regulação transcricional de miRNAs (TUREI et al., 2015).

**OmniPath:** uma coleção de PPIs humanas curadas manualmente (TUREI; KORCSMAROS; SAEZ-RODRIGUEZ, 2016).

**SignaLink2:** Banco de dados de vias de sinalização manualmente curadas de *Homo sapiens*, *Drosphila melanogaster* e *Caernorhabditis elegans*. Contém uma estrutura de multi-camadas que permite o usuário acessar diferentes tipos de interações, como PPIs, TRIs e regulações pós-transcricionais e pós-translacionais (FAZEKAS et al., 2013).

**HTRI e TRRUST:** Bancos de dados que contém interações de regulação transcricional.

## 7.3 Filtragem dos dados

Os dados de interação obtidos pelos bancos de dados citados acima foram filtrados pelos seguintes parâmetros:

**PPIs diretas:** foram selecionadas somente interações diretas, ou seja, cuja direção da interação entre duas proteínas eram conhecidas. E.g.: sabe-se que as proteínas A e B interagem, a interação será considerada direta se a informação de qual proteína é a alvo estiver disponível. Se não, será uma interação indireta. Essa informação está incluída nas interações obtidas pelos bancos de dados.

**Proteínas humanas expressadas no intestino e em contato com o exterior celular:** utilizamos os bancos de dados ComPPI, MatrixDB e Human Protein Atlas (HPA) para obter as informações sobre as localizações das proteínas nas células e em qual tecido elas são expressadas. Selecionamos apenas interações contendo proteínas do exterior da célula (pois são acessíveis às proteínas bacterianas) e aquelas que são expressas em tecidos intestinais.

## 7.4 Predição de interações entre proteínas humanas e bacterianas

### 7.4.1 Predição de interações baseadas em domínios e motifs

Utilizamos o método de predição baseado na interação entre domínios e motifs das proteínas para predizer a interação entre proteínas humanas e microbianas. Essa etapa foi executada a partir de um algoritmo em Python desenvolvido pelo grupo do Dr. Tamas Korcsmaros que utiliza a informação de entrada (domínios das proteínas bacterianas, sequências FASTA das proteínas humanas desejadas, lista de motivos ELM e a lista de interações domínio-motivo conhecidas) para predizer as potenciais interações entre as proteínas microbianas e humanas.

### 7.4.2 Filtragem de região estrutural

Embora uma interação domínio-motif seja predita, os domínios das proteínas necessitam estar acessíveis na molécula para que o motif o alcance e eles possam interagir. Portanto, para reduzir os número de falsos positivos, excluímos interações envolvendo proteínas humanas cujo motif está localizado dentro de regiões globulares e fora de regiões desordenadas.

# 8 Resultados preliminares

A partir do estudo "*Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease*" (ERICKSON et al., 2012) obtivemos dados de metaproteomas e proteomas de indivíduos saudáveis e com doença de Crohn. Desses dados, realizamos a análise de quais proteínas microbianas estavam presentes no metaproteoma de indivíduos com DC e executamos a predição de quais proteínas humanas elas interagiam. A partir desses dados, analisamos quais delas levavam à interação com as proteínas de autofagia a partir de dados de sinalização dos bancos de dados.

A rede da Figura 11 apresenta resultados preliminares da interação entre as proteínas bacterianas, ilustradas em azul, e proteínas humanas, representadas por rosa e verde, sendo que as verdes são proteínas das vias de autofagia.

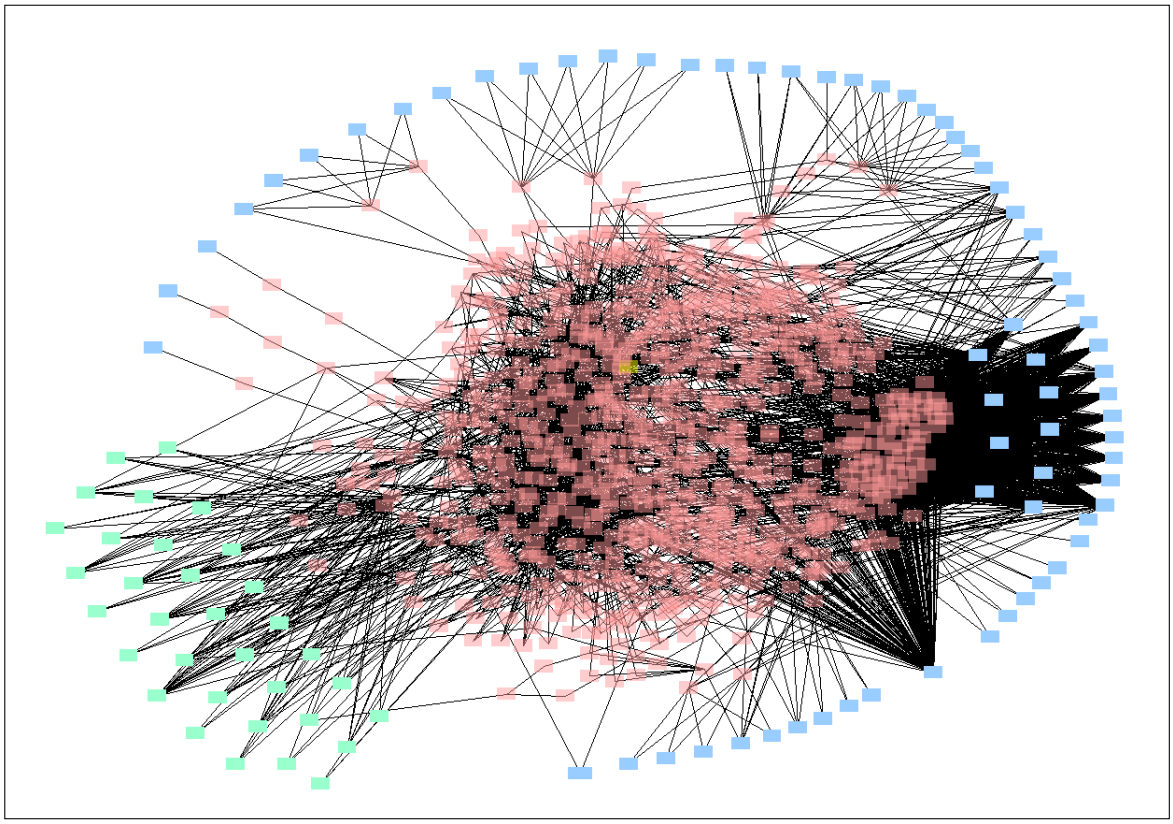

Figura 11 – Ilustração da rede de interação entre as proteínas bacterianas (em azul) e as humanas. As proteínas envolvidas no processo de autofagia estão coloridas de verde e as intermediárias de rosa.

Um exemplo de via que contém uma proteína bacteriana influenciando numa proteína de autofagia por meio de outras intermediárias presente na rede da figura 11 está ilustrado na figura 12. A proteína bacteriana CLST014522, expressa em indivíduos com DC, pode interagir com as duas proteínas intermediárias coloridas de cor-de-rosa,

que interagem com a proteína de autofagia ATG16L1.

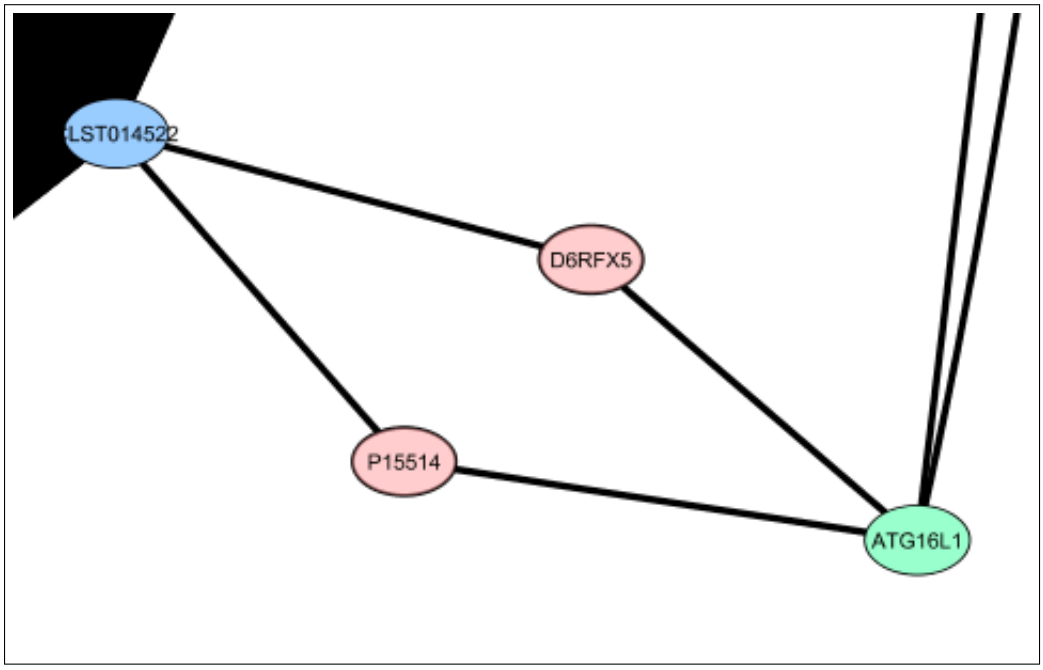


Figura 12 – Ilustração de uma via de interação rede de interação que mostra como a proteína bacteriana CLST014522 pode interagir com a proteína de autofagia ATG16L1 por meio de proteínas intermediárias.

As informações aprofundadas de como essa e as outras interações ocorre, em quais situações elas acontecem, e suas consequências e quais micro-organismos estão envolvidos serão analisadas nas próximas etapas do estudo.

# 9 Perspectivas futuras

A doença de Crohn (DC) é uma doença inflamatória intestinal causada pela resposta imune anormal contra a microbiota intestinal. Experimentos mostram que ratos sem microbiota intestinal não desenvolvem DC, fator que comprova que o microbioma é o principal responsável pela manifestação da doença. Além disso, a diferença entre as composições microbióticas de indivíduos saudáveis e com doença de Crohn já é bastante estudada. Entretanto, os mecanismos moleculares pelos quais os micro-organismos interagem com as vias de sinalização humanas para induzir o desenvolvimento da doença e sua relação com a composição microbiótica ainda não foi elucidada.

Já obtivemos as informações dos metaproteomas e proteomas de indivíduos com DC e executamos predições de como estão relacionados à vias de autofagia que podem levar à manifestação da doença. Planejamos analisar melhor esses dados para lançar luz nas explicações biológicas. Utilizaremos a ferramenta de identificação taxonômica desenvolvida para relacionar as informações funcionais com as taxonômicas do microbioma dos indivíduos com DC a partir do metagenoma obtido no mesmo estudo.

O plano de trabalho está programado de acordo com o cronograma mostrado na seção abaixo e na tabela 6.

## 9.1 Cronograma

1. Revisão bibliográfica
2. Obtenção dos genomas para teste do algoritmo
3. Organização dos dados
4. Preparação dos códigos
5. Teste dos códigos
6. Teste do poder preditivo da ferramenta
7. Análise da influência das medidas para a predição

DS Doutorado sanduíche no Instituto Earlham

8. Revisão bibliográfica
9. Busca de dados de metaproteômica de doença de Crohn (DC)

10. Busca e filtragem das vias de sinalização humanas
11. Predição dos motifs presentes nas proteínas bacterianas
12. Predição das interações entre proteínas bacterianas e humanas
13. Preparação do texto e apresentação de qualificação
14. Análise da interação entre as proteínas humanas e bacterianas em DC
15. Redação e submissão do artigo
16. Adaptação da ferramenta para detecção de composição microbiotica característica da doença de Crohn
17. Teste da ferramenta em metagenomas de DC
18. Análise da relação da composição microbiótica com sua influência em DC
19. Redação do artigo científico
20. Redação da tese de doutorado

Tabela 6 – Cronograma da perspectiva de execução do plano de trabalho.

| PLANO DE TRABALHO | 2015 | | 2016 | | | | 2017 | | | | 2018 | | | | 2019 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Jul Ago Set | Out Nov Dez | Jan Fev Mar | Abr Mai Jun | Jul Ago Set | Out Nov Dez | Jan Fev Mar | Abr Mai Jun | Jul Ago Set | Out Nov Dez | Jan Fev Mar | Abr Mai Jun | Jul Ago Set | Out Nov Dez | Jan Fev Mar | Abr Mai Jun |
| 1. | | | | | | | | | | | | | | | | |
| 2. | | | | | | | | | | | | | | | | |
| 3. | | | | | | | | | | | | | | | | |
| 4. | | | | | | | | | | | | | | | | |
| 5. | | | | | | | | | | | | | | | | |
| 6. | | | | | | | | | | | | | | | | |
| 7. | | | | | | | | | | | | | | | | |
| DS | | | | | | | Mar | | Ago | | | | | | | |
| 8. | | | | | | | | | | | | | | | | |
| 9. | | | | | | | | | | | | | | | | |
| 10. | | | | | | | | | | | | | | | | |
| 11. | | | | | | | | | | | | | | | | |
| 12. | | | | | | | | | | | | | | | | |
| 13. | | | | | | | | | Set | Out | | | | | | |
| 14. | | | | | | | | | | | | | | | | |
| 15. | | | | | | | | | | | | | | | | |
| 16. | | | | | | | | | | | | | | | | |
| 17. | | | | | | | | | | | | | | | | |
| 18. | | | | | | | | | | | | | | | | |
| 19. | | | | | | | | | | | | | | | | |
| 20. | | | | | | | | | | | | | | | | |
| | Já realizado | | | | | | Realizado no DS | | | | Será realizado | | | | | |

# Referências

ABE, T. et al. Informatics for unveiling hidden genome signatures. *Genome research*, Cold Spring Harbor Lab, v. 13, n. 4, p. 693–702, 2003. Citado na página 29.

ALTSCHUL, S. F. et al. Basic local alignment search tool. *Journal of molecular biology*, Elsevier, v. 215, n. 3, p. 403–410, 1990. Citado na página 17.

AREL, I.; ROSE, D. C.; KARNOWSKI, T. P. Deep machine learning-a new frontier in artificial intelligence research [research frontier]. *Computational Intelligence Magazine, IEEE*, IEEE, v. 5, n. 4, p. 13–18, 2010. Citado na página 25.

BABRAHAM BIOINFORMATICS. *FASTQC, A quality control tool for high throughput sequence data.* 2016. Data de acesso: 10 jun. 2016. Disponível em: <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>. Citado na página 18.

BALDI, P.; BRUNAK, S. *Bioinformatics: the machine learning approach*. [S.l.]: MIT press, 2001. Citado na página 8.

BATEMAN, A. et al. The pfam protein families database. *Nucleic acids research*, Oxford University Press, v. 32, n. suppl_1, p. D138–D141, 2004. Citado na página 23.

BAUMGART, D. C.; SANDBORN, W. J. Crohn's disease. *The Lancet*, Elsevier, v. 380, n. 9853, p. 1590–1605, 2012. Citado na página 40.

BELLA, J. M. D. et al. High throughput sequencing methods and analysis for microbiome research. *Journal of microbiological methods*, Elsevier, v. 95, n. 3, p. 401–414, 2013. Citado 2 vezes nas páginas 10 e 23.

BERGER, S. A.; KROMPASS, D.; STAMATAKIS, A. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic biology*, Oxford University Press, v. 60, n. 3, p. 291–302, 2011. Citado na página 17.

BIK, H. M. Deciphering diversity and ecological function from marine metagenomes. *The Biological Bulletin*, Marine Biological Laboratory Woods Hole, Massachusetts, v. 227, n. 2, p. 107–116, 2014. Citado na página 9.

BLANCA, J. M. et al. ngs_backbone: a pipeline for read cleaning, mapping and snp calling using next generation sequence. *BMC genomics*, BioMed Central, v. 12, n. 1, p. 285, 2011. Citado na página 18.

BRADY, A.; SALZBERG, S. Phymmbl expanded: confidence scores, custom databases, parallelization and more. *Nature methods*, Nature Research, v. 8, n. 5, p. 367–367, 2011. Citado na página 22.

BRADY, A.; SALZBERG, S. L. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nature methods*, Nature Publishing Group, v. 6, n. 9, p. 673–676, 2009. Citado na página 22.

BRAGG, L.; TYSON, G. W. Metagenomics using next-generation sequencing. *Environmental Microbiology: Methods and Protocols*, Springer, p. 183–201, 2014. Citado 2 vezes nas páginas 18 e 21.

CAI, Y.; SUN, Y. Esprit-tree: hierarchical clustering analysis of millions of 16s rrna pyrosequences in quasilinear computational time. *Nucleic acids research*, Oxford University Press, v. 39, n. 14, p. e95–e95, 2011. Citado na página 17.

CAMPBELL, A.; MRAZEK, J.; KARLIN, S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial dna. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 96, n. 16, p. 9184–9189, 1999. Citado 2 vezes nas páginas 21 e 28.

CAPORASO, J. G. et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, Nature Publishing Group, v. 7, n. 5, p. 335–336, 2010. Citado 2 vezes nas páginas 16 e 17.

CHEVREUX, B. et al. Using the miraest assembler for reliable and automated mrna transcript assembly and snp detection in sequenced ests. *Genome research*, Cold Spring Harbor Lab, v. 14, n. 6, p. 1147–1159, 2004. Citado na página 19.

CLARRIDGE, J. E. Impact of 16s rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev*, v. 17, n. 4, p. 840–62, table of contents, Oct 2004. Disponível em: <http://dx.doi.org/10.1128/CMR.17.4.840-862.2004>. Citado na página 10.

CLEMENTE, J. C.; JANSSON, J.; VALIENTE, G. Flexible taxonomic assignment of ambiguous sequencing reads. *BMC bioinformatics*, BioMed Central, v. 12, n. 1, p. 8, 2011. Citado na página 17.

COMPEAU, P. E.; PEVZNER, P. A.; TESLER, G. How to apply de bruijn graphs to genome assembly. *Nature biotechnology*, Nature Publishing Group, v. 29, n. 11, p. 987–991, 2011. Citado na página 19.

COUNCIL, N. R. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. The National Academies Press, 2007. ISBN 978-0-309-10676-4. Disponível em: <http://www.nap.edu/catalog/11902/the-new-science-of-metagenomics-revealing-the-secrets-of-our>. Citado na página 9.

CRISTIANINI, N.; SHAWE-TAYLOR, J. *An introduction to SVM*. [S.l.]: Cambridge University Press, 2000. Citado na página 27.

CRUZ, P. D. et al. Association between specific mucosa-associated microbiota in crohn's disease at the time of resection and subsequent disease recurrence: A pilot study. *Journal of gastroenterology and hepatology*, Wiley Online Library, v. 30, n. 2, p. 268–278, 2015. Citado na página 41.

CUI, H.; ZHANG, X. Alignment-free supervised classification of metagenomes by recursive svm. *BMC genomics*, BioMed Central Ltd, v. 14, n. 1, p. 641, 2013. Citado na página 25.

DATABASE resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, v. 41, n. Database issue, p. D8–D20, Jan 2013. Citado na página 31.

DESANTIS, T. Z. et al. Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, Am Soc Microbiol, v. 72, n. 7, p. 5069–5072, 2006. Citado na página 17.

DEVARAJ, S.; HEMARAJATA, P.; VERSALOVIC, J. The human gut microbiome and body metabolism: implications for obesity and diabetes. *Clinical chemistry*, Clinical Chemistry, v. 59, n. 4, p. 617–628, 2013. Citado na página 10.

DIAZ, N. N. et al. Tacoa–taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC bioinformatics*, BioMed Central, v. 10, n. 1, p. 56, 2009. Citado na página 22.

DIJK, E. L. V. et al. Ten years of next-generation sequencing technology. *Trends in genetics*, Elsevier, v. 30, n. 9, p. 418–426, 2014. Citado na página 8.

DRANCOURT, M.; BERGER, P.; RAOULT, D. Systematic 16s rrna gene sequencing of atypical clinical isolates identified 27 new bacterial species associated with humans. *Journal of Clinical Microbiology*, Am Soc Microbiol, v. 42, n. 5, p. 2197–2202, 2004. Citado na página 10.

DUTTA, C.; PAUL, S. Microbial lifestyle and genome signatures. *Curr Genomics*, v. 13, n. 2, p. 153–162, Apr 2012. Citado na página 29.

EDGAR, R. C. et al. Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, Oxford University Press, v. 27, n. 16, p. 2194–2200, 2011. Citado 2 vezes nas páginas 16 e 17.

ERICKSON, A. R. et al. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of crohn's disease. *PloS one*, Public Library of Science, v. 7, n. 11, p. e49138, 2012. Citado 4 vezes nas páginas 42, 43, 46 e 48.

ESPOSITO, A.; KIRSCHBERG, M. How many 16s-based studies should be included in a metagenomic conference? it may be a matter of etymology. *FEMS Microbiology Letters*, v. 351, p. 145–146, 2014. Citado na página 15.

FASTX-TOOLKIT, A short-reads pre-processing tools. 2016. Data de acesso: 10 jun. 2016. Disponível em: <http://hannonlab.cshl.edu/fastx_toolkit/index.html>. Citado na página 18.

FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006. Citado na página 33.

FAZEKAS, D. et al. Signalink 2 – a signaling pathway resource with multi-layered regulatory networks. v. 7, p. 7, 01 2013. Citado na página 46.

FORDE, B. M.; O'TOOLE, P. W. Next-generation sequencing technologies and their impact on microbial genomics. *Briefings in functional genomics*, Oxford University Press, v. 12, n. 5, p. 440–453, 2013. Citado na página 17.

FU, L. et al. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, Oxford University Press, v. 28, n. 23, p. 3150–3152, 2012. Citado na página 17.

GARBARINE, E. et al. Information-theoretic approaches to svm feature selection for metagenome read classification. *Computational biology and chemistry*, Elsevier, v. 35, n. 3, p. 199–209, 2011. Citado na página 25.

GASPAR, J. M.; THOMAS, W. K. Assessing the consequences of denoising marker-based metagenomic data. *PloS one*, Public Library of Science, v. 8, n. 3, p. e60458, 2013. Citado na página 16.

GERHARDT, G. J. L. et al. Triplet entropy analysis of hemagglutinin and neuraminidase sequences measures influenza virus phylodynamics. *Gene*, v. 528, n. 2, p. 277–281, Oct 2013. Disponível em: <http://dx.doi.org/10.1016/j.gene.2013.06.060>. Citado na página 28.

GERHARDT, N. L. G. J.; CORSO, G. Network clustering coefficient approach to dna sequence analysis. *Chaos, Solitons and Fractals*, v. 28, p. 1037–1045, 2006. Citado 2 vezes nas páginas 29 e 30.

GERLACH, W.; STOYE, J. Taxonomic classification of metagenomic shotgun sequences with carma3. *Nucleic acids research*, Oxford University Press, v. 39, n. 14, p. e91–e91, 2011. Citado na página 21.

GEVERS, D. et al. The treatment-naive microbiome in new-onset crohn's disease. *Cell host & microbe*, Elsevier, v. 15, n. 3, p. 382–392, 2014. Citado na página 41.

GLASS, E. M. et al. Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols*, Cold Spring Harbor Laboratory Press, v. 2010, n. 1, p. pdb–prot5368, 2010. Citado na página 23.

GREENBAUM, B. D. et al. Patterns of evolution and host gene mimicry in influenza and other rna viruses. *PLoS pathogens*, Public Library of Science, v. 4, n. 6, p. e1000079, 2008. Citado na página 28.

HAAS, B. J. et al. Chimeric 16s rrna sequence formation and detection in sanger and 454-pyrosequenced pcr amplicons. *Genome research*, Cold Spring Harbor Lab, v. 21, n. 3, p. 494–504, 2011. Citado na página 16.

HAGEN, J. B. The origins of bioinformatics. *Nature Reviews Genetics*, Nature Publishing Group, v. 1, n. 3, p. 231–236, 2000. Citado na página 8.

HAJIAN-TILAKI, K. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, Babol University of Medical Sciences, v. 4, n. 2, p. 627, 2013. Citado na página 33.

HANDELSMAN, J. et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, Elsevier, v. 5, n. 10, p. R245–R249, 1998. Citado na página 12.

HANKERSON, D.; JOHNSON, P. D.; HARRIS, G. A. *Introduction to Information Theory and Data Compression*. 1st. ed. Boca Raton, FL, USA: CRC Press, Inc., 1998. ISBN 0849339855. Citado na página 30.

HENDERSON, P.; STEVENS, C. The role of autophagy in crohn's disease. *Cells*, Molecular Diversity Preservation International, v. 1, n. 3, p. 492–519, 2012. Citado 3 vezes nas páginas 40, 41 e 42.

HUNTER, S. et al. Ebi metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic acids research*, Oxford University Press, v. 42, n. D1, p. D600–D606, 2013. Citado na página 23.

HUSON, D. H. et al. Megan analysis of metagenomic data. *Genome research*, Cold Spring Harbor Lab, v. 17, n. 3, p. 377–386, 2007. Citado 3 vezes nas páginas 17, 20 e 23.

JUSTE, C. et al. Bacterial protein signals are associated with crohn's disease. *Gut*, BMJ Publishing Group, v. 63, n. 10, p. 1566–1577, 2014. Citado 2 vezes nas páginas 42 e 43.

KANEHISA, M.; GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, Oxford University Press, v. 28, n. 1, p. 27–30, 2000. Citado na página 23.

KAPETANOVIC, I. M.; ROSENFELD, S.; IZMIRLIAN, G. Overview of commonly used bioinformatics methods and their applications. *Annals of the New York Academy of Sciences*, Wiley Online Library, v. 1020, n. 1, p. 10–21, 2004. Citado 2 vezes nas páginas 24 e 27.

KARLIN, S. Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol*, v. 1, n. 5, p. 598–610, Oct 1998. Citado na página 29.

KARLIN, S.; BURGE, C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*, v. 11, n. 7, p. 283–290, Jul 1995. Citado na página 29.

KARLIN, S.; MRAZEK, J.; CAMPBELL, A. M. Compositional biases of bacterial genomes and evolutionary implications. *Journal of bacteriology*, Am Soc Microbiol, v. 179, n. 12, p. 3899–3913, 1997. Citado na página 29.

KENT, W. J. Blat—the blast-like alignment tool. *Genome research*, Cold Spring Harbor Lab, v. 12, n. 4, p. 656–664, 2002. Citado na página 23.

KIM, M. et al. Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics & informatics*, v. 11, n. 3, p. 102–113, 2013. Citado 7 vezes nas páginas 13, 15, 16, 17, 19, 20 e 21.

KUMAR, S. et al. Metagenomics: Retrospect and prospects in high throughput age. *Biotechnology research international*, Hindawi Publishing Corporation, v. 2015, 2015. Citado na página 13.

KUNIN, V. et al. A bioinformatician's guide to metagenomics. *Microbiology and molecular biology reviews*, Am Soc Microbiol, v. 72, n. 4, p. 557–578, 2008. Citado na página 27.

LAND, M. et al. Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics*, Springer, v. 15, n. 2, p. 141–161, 2015. Citado na página 13.

LEE, H. et al. Third-generation sequencing and the future of genomics. *bioRxiv*, Cold Spring Harbor Labs Journals, p. 048603, 2016. Citado na página 13.

LI, R. et al. Soap: short oligonucleotide alignment program. *Bioinformatics*, Oxford Univ Press, v. 24, n. 5, p. 713–714, 2008. Citado na página 19.

LIU, Y. et al. Gene prediction in metagenomic fragments based on the svm algorithm. *BMC bioinformatics*, BioMed Central Ltd, v. 14, n. Suppl 5, p. S12, 2013. Citado 2 vezes nas páginas 20 e 27.

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007. Citado na página 24.

MACDONALD, N. J.; PARKS, D. H.; BEIKO, R. G. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic acids research*, Oxford University Press, v. 40, n. 14, p. e111–e111, 2012. Citado na página 22.

MADIGAN, M. T. et al. *Microbiologia de Brock-14ª Edição*. [S.l.]: Artmed Editora, 2016. Citado 2 vezes nas páginas 10 e 11.

MAHAMUDA, V.; U, M. C.; RASHEED, K. Application of machine learning algorithms for binning metagenomic data. p. 68–74, 2010. Citado na página 23.

MANDE, S. S.; MOHAMMED, M. H.; GHOSH, T. S. Classification of metagenomic sequences: methods and challenges. *Briefings in bioinformatics*, Oxford University Press, v. 13, n. 6, p. 669–681, 2012. Citado 2 vezes nas páginas 19 e 21.

MANICHANH, C. et al. The gut microbiota in ibd. *Nature Reviews Gastroenterology and Hepatology*, Nature Publishing Group, v. 9, n. 10, p. 599–608, 2012. Citado na página 40.

MANICHANH, C. et al. Reduced diversity of faecal microbiota in crohn's disease revealed by a metagenomic approach. *Gut*, BMJ Publishing Group, v. 55, n. 2, p. 205–211, 2006. Citado na página 41.

MARCHLER-BAUER, A. et al. Cdd: a conserved domain database for protein classification. *Nucleic acids research*, Oxford University Press, v. 33, n. suppl_1, p. D192–D196, 2005. Citado na página 23.

MARCO, D. *Metagenomics: Current innovations and future trends*. [S.l.]: Horizon Scientific Press, 2011. Citado na página 12.

MARKOWITZ, V. M. et al. Img 4 version of the integrated microbial genomes comparative analysis system. *Nucleic acids research*, Oxford University Press, v. 42, n. D1, p. D560–D567, 2013. Citado na página 23.

MARSLAND, S. *Machine learning: an algorithmic perspective*. [S.l.]: CRC press, 2014. Citado 2 vezes nas páginas 23 e 24.

MATSEN, F. A.; KODNER, R. B.; ARMBRUST, E. V. pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, BioMed Central, v. 11, n. 1, p. 538, 2010. Citado na página 17.

MCHARDY, A. C. et al. Accurate phylogenetic classification of variable-length dna fragments. *Nature methods*, Nature Publishing Group, v. 4, n. 1, p. 63–72, 2007. Citado 2 vezes nas páginas 22 e 25.

MEYER, F. et al. The metagenomics rast server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, BioMed Central, v. 9, n. 1, p. 386, 2008. Citado 2 vezes nas páginas 21 e 23.

MICHAIL, S.; BULTRON, G.; DEPAOLO, R. W. Genetic variants associated with crohn's disease. *The application of clinical genetics*, Dove Press, v. 6, p. 25, 2013. Citado 2 vezes nas páginas 41 e 43.

MICROWINE, A MARIE CURIE INITIAL TRAINING NETWORK. *MicroWine, A Marie Curie Initial Training Network*. 2016. Data de acesso: 10 jun. 2016. Disponível em: <http://www.microwine.eu/>. Citado na página 12.

MIRARAB, S.; NGUYEN, N.; WARNOW, T. Sepp: Saté-enabled phylogenetic placement. In: *Biocomputing 2012*. [S.l.]: Citeseer, 2012. p. 247–258. Citado na página 17.

MOHAMMED, M. H. et al. Sphinx—an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics*, Oxford University Press, v. 27, n. 1, p. 22–30, 2010. Citado na página 22.

MONDAL, K.; KUGATHASAN, S. Ibd: Genetic differences in crohn's disease susceptibility and outcome. *Nature Reviews Gastroenterology & Hepatology*, Nature Research, v. 14, n. 5, p. 266–268, 2017. Citado na página 40.

MORGAN, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome biology*, BioMed Central, v. 13, n. 9, p. R79, 2012. Citado na página 41.

MORGAVI, D. P. et al. Rumen microbial (meta)genomics and its application to ruminant production. *Animal*, v. 7, p. 184–201, 2013. Citado na página 9.

MOROZOVA, O.; MARRA, M. A. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, Elsevier, v. 92, n. 5, p. 255–264, 2008. Citado na página 13.

MULLER, J. et al. eggnog v2. 0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic acids research*, Oxford University Press, v. 38, n. suppl_1, p. D190–D195, 2009. Citado na página 23.

NAMIKI, T. et al. Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, Oxford University Press, v. 40, n. 20, p. e155–e155, 2012. Citado na página 19.

NAWROCKI, E. P.; KOLBE, D. L.; EDDY, S. R. Infernal 1.0: inference of rna alignments. *Bioinformatics*, Oxford University Press, v. 25, n. 10, p. 1335–1337, 2009. Citado na página 17.

NIKOLAKI, S.; TSIAMIS, G. Microbial diversity in the era of omic technologies. *BioMed research international*, Hindawi Publishing Corporation, v. 2013, 2013. Citado 2 vezes nas páginas 10 e 15.

NORRGARD, K. Genetic variation and disease: Gwas. *Nature Education*, v. 1, n. 1, p. 35, 2008. Citado na página 42.

OSSOWSKI, S. et al. Sequencing of natural strains of arabidopsis thaliana with short reads. *Genome research*, Cold Spring Harbor Lab, v. 18, n. 12, p. 2024–2033, 2008. Citado na página 18.

OULAS, A. et al. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and Biology insights*, Libertas Academica, v. 9, p. 75, 2015. Citado 4 vezes nas páginas 13, 16, 17 e 23.

OUZOUNIS, C. A.; VALENCIA, A. Early bioinformatics: the birth of a discipline - a personal view. *Bioinformatics*, Oxford University Press (OUP), v. 19, n. 17, p. 2176–2190, nov 2003. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btg309>. Citado na página 8.

OVERBEEK, R. et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic acids research*, Oxford University Press, v. 33, n. 17, p. 5691–5702, 2005. Citado na página 23.

PASCAL, V. et al. A microbial signature for crohn9s disease. *Gut*, BMJ Publishing Group, p. gutjnl–2016, 2017. Citado na página 40.

PASSEL, M. W. J. van et al. The reach of the genome signature in prokaryotes. *BMC Evol Biol*, v. 6, p. 84, 2006. Citado na página 28.

PATIN, N. V. et al. Effects of otu clustering and pcr artifacts on microbial diversity estimates. *Microbial ecology*, Springer, v. 65, n. 3, p. 709–719, 2013. Citado na página 17.

PENG, Y. et al. Meta-idba: a de novo assembler for metagenomic data. *Bioinformatics*, Oxford University Press, v. 27, n. 13, p. i94–i101, 2011. Citado na página 19.

PETTERSSON, E.; LUNDEBERG, J.; AHMADIAN, A. Generations of sequencing technologies. *Genomics*, v. 93, n. 2, p. 105–111, Feb 2009. Citado na página 12.

PEVZNER, P. A.; TANG, H.; WATERMAN, M. S. An eulerian path approach to dna fragment assembly. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 98, n. 17, p. 9748–9753, 2001. Citado na página 19.

PORETSKY, R. et al. Strengths and limitations of 16s rrna gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS One*, Public Library of Science, v. 9, n. 4, p. e93827, 2014. Citado na página 15.

PRESLEY, L. L. et al. Host–microbe relationships in inflammatory bowel disease detected by bacterial and metaproteomic analysis of the mucosal–luminal interface. *Inflammatory bowel diseases*, Wiley Online Library, v. 18, n. 3, p. 409–417, 2012. Citado 2 vezes nas páginas 42 e 43.

PRUESSE, E.; PEPLIES, J.; GLÖCKNER, F. O. Sina: accurate high-throughput multiple sequence alignment of ribosomal rna genes. *Bioinformatics*, Oxford University Press, v. 28, n. 14, p. 1823–1829, 2012. Citado na página 17.

QUINCE, C. et al. Removing noise from pyrosequenced amplicons. *BMC bioinformatics*, BioMed Central, v. 12, n. 1, p. 38, 2011. Citado na página 16.

QUINLAN, A. R. et al. Pyrobayes: an improved base caller for snp discovery in pyrosequences. *Nature methods*, Nature Publishing Group, v. 5, n. 2, p. 179–181, 2008. Citado na página 18.

REHMAN, A. et al. Transcriptional activity of the dominant gut mucosal microbiota in chronic inflammatory bowel disease patients. *Journal of medical microbiology*, Microbiology Society, v. 59, n. 9, p. 1114–1122, 2010. Citado na página 41.

RICANEK, P. et al. Gut bacterial profile in patients newly diagnosed with treatment-naïve crohn's disease. *Clinical and experimental gastroenterology*, Dove Press, v. 5, p. 173, 2012. Citado na página 41.

RIDAURA, V. K. et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science*, American Association for the Advancement of Science, v. 341, n. 6150, p. 1241214, 2013. Citado na página 20.

RIESENFELD, C. S.; SCHLOSS, P. D.; HANDELSMAN, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.*, Annual Reviews, v. 38, p. 525–552, 2004. Citado na página 14.

ROSEN, M. J. et al. Denoising pcr-amplified metagenome data. *BMC bioinformatics*, BioMed Central, v. 13, n. 1, p. 283, 2012. Citado 2 vezes nas páginas 16 e 22.

SAITOH, T. et al. Loss of the autophagy protein atg16l1 enhances endotoxin-induced il-1 [beta] production. *Nature*, Nature Publishing Group, v. 456, n. 7219, p. 264, 2008. Citado na página 42.

SANSCHAGRIN, S.; YERGEAU, E. Next-generation sequencing of 16s ribosomal rna gene amplicons. *Journal of visualized experiments: JoVE*, MyJoVE Corporation, n. 90, 2014. Citado na página 16.

SANTOS, L. dos; RYBARCZYK-FILHO; GERHARDT, G. J. L. Triplet entropy in h1n1 virus. *Tendências em Matemática Aplicada e Computacional*, v. 12, p. 253–261, 2011. Citado na página 30.

SCHMIDT, T. M.; DELONG, E.; PACE, N. Analysis of a marine picoplankton community by 16s rrna gene cloning and sequencing. *Journal of bacteriology*, Am Soc Microbiol, v. 173, n. 14, p. 4371–4378, 1991. Citado na página 14.

SELENGUT, J. D. et al. Tigrfams and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic acids research*, Oxford University Press, v. 35, n. suppl_1, p. D260–D264, 2006. Citado na página 23.

SESHADRI, R. et al. Camera: a community resource for metagenomics. *PLoS biology*, Public Library of Science, v. 5, n. 3, p. e75, 2007. Citado na página 23.

SHANNON, C. E. A mathematical theory of communication. *Bell System Technical Journal*, v. 27, n. 3, p. 379–423, 1948. Citado na página 30.

SHARPTON, T. J. An introduction to the analysis of shotgun metagenomic data. *Frontiers in plant science*, Frontiers Media SA, v. 5, 2014. Citado 6 vezes nas páginas 18, 19, 20, 21, 22 e 23.

SHEEHAN, D.; SHANAHAN, F. The gut microbiota in inflammatory bowel disease. *Gastroenterology Clinics*, Elsevier, v. 46, n. 1, p. 143–154, 2017. Citado 2 vezes nas páginas 40 e 41.

SHI, J. Defensins and paneth cells in inflammatory bowel disease. *Inflammatory bowel diseases*, Wiley Online Library, v. 13, n. 10, p. 1284–1292, 2007. Citado 2 vezes nas páginas 41 e 42.

SIMPSON, J. T. et al. Abyss: a parallel assembler for short read sequence data. *Genome research*, Cold Spring Harbor Lab, v. 19, n. 6, p. 1117–1123, 2009. Citado na página 19.

STANGE, E. F.; WEHKAMP, J. Recent advances in understanding and managing crohn's disease. *F1000Research*, Faculty of 1000 Ltd, v. 5, 2016. Citado 2 vezes nas páginas 40 e 41.

SUN, Y. et al. A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Briefings in bioinformatics*, Oxford Univ Press, p. bbr009, 2011. Citado 2 vezes nas páginas 14 e 17.

TAKAHASHI, M.; KRYUKOV, K.; SAITOU, N. Estimation of bacterial species phylogeny through oligonucleotide frequency distances. *Genomics*, Elsevier, v. 93, n. 6, p. 525–533, 2009. Citado na página 29.

TATUSOV, R. L. et al. The cog database: an updated version includes eukaryotes. *BMC bioinformatics*, BioMed Central, v. 4, n. 1, p. 41, 2003. Citado na página 23.

TEELING, H. et al. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental microbiology*, Wiley Online Library, v. 6, n. 9, p. 938–947, 2004. Citado na página 30.

THOMAS, T.; GILBERT, J.; MEYER, F. Metagenomics-a guide from sampling to data analysis. *Microb Inform Exp*, v. 2, n. 3, p. 1–12, 2012. Citado 5 vezes nas páginas 9, 12, 13, 21 e 23.

THOMPSON, C. C. et al. Microbial genomic taxonomy. *BMC genomics*, BioMed Central, v. 14, n. 1, p. 913, 2013. Citado na página 10.

THORKILDSEN, L. T. et al. Dominant fecal microbiota in newly diagnosed untreated inflammatory bowel disease patients. *Gastroenterology research and practice*, Hindawi Publishing Corporation, v. 2013, 2013. Citado na página 41.

TREANGEN, T. J. et al. Metamos: a modular and open source metagenomic assembly and analysis pipeline. *Genome biology*, BioMed Central, v. 14, n. 1, p. R2, 2013. Citado na página 19.

TUREI, D. et al. Autophagy regulatory network — a systems-level bioinformatics resource for studying the mechanism and regulation of autophagy. v. 11, 01 2015. Citado na página 46.

TUREI, D.; KORCSMAROS, T.; SAEZ-RODRIGUEZ, J. Omnipath: Guidelines and gateway for literature-curated signaling pathway resources. v. 13, p. 966–967, 11 2016. Citado na página 46.

WANG, X.; HUYCKE, M. M. Extracellular superoxide production by enterococcus faecalis promotes chromosomal instability in mammalian cells. *Gastroenterology*, Elsevier, v. 132, n. 2, p. 551–561, 2007. Citado na página 17.

WANICHTHANARAK, K.; FAHRMANN, J. F.; GRAPOV, D. Genomic, proteomic, and metabolomic data integration strategies. *Biomarker insights*, Libertas Academica, v. 10, n. Suppl 4, p. 1, 2015. Citado na página 8.

WOESE, C. R.; KANDLER, O.; WHEELIS, M. L. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 87, n. 12, p. 4576–4579, 1990. Citado na página 11.

WOOLEY, J. C.; GODZIK, A.; FRIEDBERG, I. A primer on metagenomics. *PLoS computational biology*, Public Library of Science, v. 6, n. 2, p. e1000667, 2010. Citado 2 vezes nas páginas 9 e 19.

WRIGHT, E. S.; YILMAZ, L. S.; NOGUERA, D. R. Decipher, a search-based approach to chimera identification for 16s rrna sequences. *Applied and environmental microbiology*, Am Soc Microbiol, v. 78, n. 3, p. 717–725, 2012. Citado na página 16.

WU, M.; SCOTT, A. J. Phylogenomic analysis of bacterial and archaeal sequences with amphora2. *Bioinformatics*, Oxford University Press, v. 28, n. 7, p. 1033–1034, 2012. Citado na página 17.

YAKOVCHUK, P.; PROTOZANOVA, E.; FRANK-KAMENETSKII, M. D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res*, v. 34, n. 2, p. 564–574, 2006. Citado na página 28.

YANO, T.; KURATA, S. An unexpected twist for autophagy in crohn's disease. *Nature immunology*, Nature Publishing Group, v. 10, n. 2, p. 134–136, 2009. Citado 2 vezes nas páginas 41 e 42.

ZARRAONAINDIA, I. et al. The soil microbiome influences grapevine-associated microbiota. *mBio*, American Society for Microbiology, v. 6, n. 2, p. e02527–14, mar 2015. Citado 2 vezes nas páginas 9 e 20.

ZERBINO, D. R.; BIRNEY, E. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, Cold Spring Harbor Lab, v. 18, n. 5, p. 821–829, 2008. Citado na página 19.

ZHOU, J. et al. High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. *MBio*, Am Soc Microbiol, v. 6, n. 1, p. e02288–14, 2015. Citado na página 16.