UNIVERSIDADE DE SÃO PAULO

PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

**Classificação Taxonômica de Sequências obtidas com
Meta-ômicas por meio de Integração de Dados**

Tese de Doutorado

FELIPE PRATA LIMA

ORIENTADOR: JOÃO CARLOS SETUBAL

São Paulo, SP
2019

FICHA CATALOGRÁFICA

L732 Lima, Felipe Prata
Classificação taxonômica de sequências obtidas com meta-ômicas por meio de
integração de dados / Felipe Prata Lima, orientador João Carlos Setubal, São
Paulo, 2019.
94 f.

Tese (Doutorado) - Programa Interunidades de Pós-Graduação em
Bioinformática, Universidade de São Paulo

1. Análise de dados. 2. Bioinformática. 3. Dados de sequência molecular. 4.
Ecologia de comunidades. I. Setubal, João Carlos, orientador. II. Universidade de
São Paulo. III. Título.

Elaborada pelo Serviço de Informação e Biblioteca Carlos Benjamin de Lyra do IME-USP, pela
bibliotecária Eliana Mara Martins Ramalho CRB-8/4819

AGRADECIMENTOS

A Deus.

À minha família, em especial aos meus pais Mirandyr e Liduina, aos meus irmãos Bruno e Diogo e à minha esposa Analice, por todo amor e dedicação incondicionais que sempre me oferecem em todos os momentos da minha vida. Este trabalho é dedicado a eles.

Ao meu orientador João Carlos Setubal por todos os seus ensinamentos, trabalho e paciência dedicados à minha formação e ao desenvolvimento deste trabalho.

Aos professores André Fujita, Alessandro Varani, Carla Taddei, Christian Hoffmann, Emmanuel Dias-Neto, Ronaldo Hashimoto e Welington Araújo pelas participações em minhas bancas de qualificação e defesa e seus valiosos direcionamentos para a melhoria deste trabalho.

Aos membros do Programa Interunidades de Pós-Graduação em Bioinformática da USP, especialmente Patricia, Cris e Márcia, pelo excelente tratamento que sempre recebi durante a realização do doutorado.

Aos amigos oriundos da participação no Laboratório de Bioinformática (Setulab), Ana Carolina, Andrew, Antônio, Bruno, Carlos, David, Deibs, Deyvid, Elton, Fernanda, Fernando, George, Gianluca, Gustavo, Ítalo, Joaquim, Jhonatas, José (Patané), Julio, Leandro, Livia, Lucas, Melline, Ondina, Raquel, Roberta, Rodrigo, Suzana, Thibério, e Victor, por todas as horas de aprendizado e de lazer compartilhadas.

À professora Aline Maria da Silva e a todos os demais participantes do Projeto Metazoo, por todo o seu trabalho.

Aos amigos nascidos de curtas e longas temporadas na cidade de São Paulo, especialmente

Gustavo, Bruno e Letícia, Clarisse, Fernanda e Daniel, Flávio e Marta, Lorena, Marcelo e Juliana, e Pedro e Giovana, pela boa companhia em vários momentos.

Aos amigos do IFAL, em especial Adriano, Alaelson, Davi, Fernando, Gian, Gregor, Heitor, Leonardo, Lucas, Maurício, Osvaldo e Tarsis, pelo apoio e incentivo.

Ao Instituto Federal de Alagoas (IFAL) pelo apoio à formação (Processo: 23041.010066/2014-31), e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo apoio à pesquisa através do Projeto Metazoo (Processo: 2011/50870-6).

Comunidades microbianas possuem papéis importantes em processos que ocorrem em diversos ambientes, tais como solos, oceanos e o trato gastrointestinal dos seres humanos. Portanto, é de interesse a compreensão da estrutura e do funcionamento dessas comunidades. A estrutura dessas comunidades, em termos de organismos componentes, pode ser determinada com o uso do sequenciamento de nova geração em conjunto com as técnicas meta-ômicas e pela análise taxonômica das sequências obtidas com programas de classificação taxonômica. Se por um lado diversos programas estão disponíveis, por outro lado eles cometem erros, como a identificação parcial dos organismos presentes na amostra e a identificação de organismos que não estão presentes na amostra (os falsos positivos – FPs). Algumas abordagens foram propostas para a melhoria das classificações taxonômicas obtidas por esses programas com a redução desses FPs, porém elas abordam apenas um tipo de meta-ômica, a metagenômica. Neste trabalho, propomos uma nova abordagem através da integração de diferentes meta-ômicas - metagenômicas shotgun e de amplicons de 16S, e metatranscritômica. Exploramos os resultados de classificações de dados simulados e *mocks* para a extração de variáveis e desenvolvemos modelos de classificação para discriminação de predições de espécies de bactérias classificadas como corretas ou incorretas. Comparamos o desempenho dos resultados obtidos entre as meta-ômicas individuais e os obtidos através da integração observando o balanceamento entre a precisão e a sensibilidade. De acordo com as medidas calculadas com nossos conjuntos de dados, nossa abordagem demonstrou melhorias na classificação com a redução de FPs e aumentos para a medida F1, quando comparada com abordagens não integrativas, inclusive com o uso de métodos de combinação de classificadores. Para facilitar seu uso, desenvolvemos o Gunga, uma ferramenta que incorpora a abordagem desenvolvida em formato de pacote do R, com funcionalidades para a integração de dados de classificação taxonômica com diferentes meta-ômicas e a classificação das predições incorretas.

Palavras-chave: análise taxonômica. bioinformática. integração de dados. meta-ômicas.

ABSTRACT

Microbial communities play important roles in processes that occur in diverse environments, such as soils, oceans, and the gastrointestinal tract of humans. Therefore, it is of interest to understand the structure and functioning of these communities. The structure of these communities, in terms of component organisms, can be determined by the use of the next generation sequencing in conjunction with the meta-omics techniques and by the taxonomic analysis of the sequences obtained with taxonomic classification programs. If on the one hand several programs are available, on the other hand they make mistakes, such as the partial identification of the organisms present in the sample and the identification of organisms that are not present in the sample (the false positives – FPs). Some approaches have been proposed to improve the taxonomic classifications obtained by these programs with the reduction of these FPs, but they address only one type of meta-omics, the metagenomics. In this work, we propose a new approach by integrating different meta-omics - shotgun and 16S amplicon metagenomics, and metatranscriptomics. We explored the classifications results of simulated data and mocks for variable extraction and developed classification models for discriminating predictions of bacterial species classified as correct or incorrect. We compared the performance of the results obtained between the individual meta-omics and the obtained through the integration observing the balance between precision and sensitivity. According to the measures calculated with our data sets, our approach has shown improvements in the classification with the reduction of the FPs and increases for the F1 measure, when compared to non-integrative approaches, including the use of classifiers combination methods. To facilitate its use, we developed the Gunga, a tool that incorporates the developed approach in R package format, with features for the integration of taxonomic classification data with different meta-omics and the classification of the incorrect predictions.

Palavras-chave: taxonomic analysis. bioinformatics. data integration. meta-omics.

LISTA DE FIGURAS

3.1	<i>Pipeline</i> para simulação de dados meta-ômicos (SH, MT e 16S) “casados”.	32
3.2	Agrupamento dos perfis taxonômicos selecionados para simulação.	35
3.3	Comparação dos perfis taxonômicos das amostras simuladas com NMDS.	35
4.1	Distribuição das medidas de desempenho obtidas com os classificadores taxonômicos em SH.	39
4.2	Dispersão da precisão e da sensibilidade obtidas com os classificadores em SH e MT para espécie, com destaque para o desempenho com os diferentes grupos de amostras.	40
4.3	Agrupamentos dos classificadores com precisão e sensibilidade em todos os níveis taxonômicos.	41
4.4	Correlações de F1 obtidas com os classificadores taxonômicos em SH e MT.	42
4.5	Distribuição das medidas de desempenho obtidas com os classificadores taxonômicos em 16S. O sufixo RepSet se refere ao uso do método de obtenção de OTUs e Slout se refere ao uso do conjunto completo de sequências. Primers é um classificador hipotético, com precisão perfeita (igual a 1) e máximo de sensibilidade para a amostra considerando o par de <i>primers</i> utilizado. Dentre todos, o Spingo (tanto RepSet quanto Slout) é o que constantemente mais se aproxima do Primers.	43
4.6	Números de FNs obtidos com cada ferramenta por nível taxonômico, com e sem o uso da abordagem de OTUs.	44

4.7	Agrupamento dos classificadores de 16S com a precisão e a sensibilidade. O uso ou não da abordagem de OTUs é o principal fator que distingue o desempenho. O SPINGO é o que possui o desempenho mais próximo de um classificador hipotético 100% preciso (Primers).	44
4.8	Distribuição das medidas de desempenho obtidas com as abordagens de combinação em SH e MT.	45
4.9	Distribuição da precisão e da sensibilidade obtidas com as abordagens de combinação em SH e MT para espécie, com destaque para o desempenho com os diferentes grupos de amostras.	46
4.10	Correlações de valores de F1 obtidos com as abordagens de combinação em SH e MT.	46
4.11	Precisão dos modelos de classificação por algoritmo de classificação com as amostras de teste.	48
4.12	Precisão dos modelos de classificação por ferramenta com as amostras de teste.	49
4.13	Sensibilidade e especificidade dos modelos de classificação por algoritmo de classificação com as amostras de teste.	50
4.14	Sensibilidade e especificidade dos modelos de classificação por ferramenta com as amostras de teste.	50
4.15	Comparações de desempenho entre SH, ferramentas de <i>Ensembles</i> em SH e a abordagem integrada com o uso de modelos de AM.	52
5.1	Fluxo genérico de processamento de dados com o Gunga. A fase Pré Gunga corresponde à obtenção das classificações e perfis taxonômicos com um programa independente escolhido pelo usuário, e é executada externamente. As demais fases, do Pré-processamento até a Integração, correspondem aos processamentos realizados internamente pela ferramenta.	54
5.2	Densidades de algumas variáveis do modelo para as classes preditas, com valores dos dados de exemplo.	56
5.3	Desempenho da abordagem baseada em duas meta-ômicas (SH e 16S) em amostras de teste.	58
5.4	Composições taxonômicas das amostras Mix3, usada para a validação da ferramenta, e demais amostras <i>mock</i> Mix1 e Mix2 à esquerda. Distâncias de Bray-Curtis e agrupamento das amostras à direita.	59

5.5	Composições taxonômicas das amostras HC[1-5] do CAMI com as 10 espécies mais abundantes em destaque à esquerda. Distâncias de Bray-Curtis e agrupamento das amostras à direita.	60
5.6	Desempenhos das ferramentas de classificação, de combinação de classificadores e do Gunga com dados do Kraken2 RF para as amostras HC[1-5] do CAMI.	60
5.7	Densidades para valores de alguns preditores usados na classificação dos dados integridos de ZC4D30.	65

LISTA DE TABELAS

1.1	Sequenciadores de nova geração e suas características [8, 55].	3
1.2	Exemplo de dado taxonômico como uma lista de sequências e respectivos táxons associados. Nos referimos ao dado taxonômico nesse formato como classificação taxonômica	6
1.3	Exemplo de dado taxonômico como uma lista de táxons e suas respectivas representações percentuais. Nos referimos ao dado taxonômico nesse formato como perfil taxonômico . Os táxons relacionados no perfil são chamados de predições taxonômicas	7
2.1	Classificadores taxonômicos usados para meta-ômicas SH e MT.	17
3.1	Características das amostras simuladas.	33
4.1	Ferramentas e números de variáveis obtidas a partir da extração de dados. As ferramentas marcadas com Sim no campo Classificação foram as que forneceram saídas com as classificações individuais de cada sequência que foram utilizadas para a extração de variáveis. Para aquelas marcadas com não extraímos variáveis apenas do perfil taxonômico resultante.	48
4.2	Medianas de F1, com IQRs entre parênteses, dos modelos de classificação para cada par (<i>algoritmo, ferramenta</i>).	51
5.1	Funções do pacote Gunga e sua nomenclatura.	55
5.2	Desempenhos das ferramentas de classificação, de combinação de classificadores e do Gunga com dados do Kraken2 RF para Mix3.	59

5.3	Espécies preditas pelo Gunga em ZC4D30, com as também preditas dentre as mais abundantes no metatranscritoma em Antunes et al. (2016) [6] marcadas com * e com as também preditas pelo MetaPhlAn2 destacadas em negrito.	63
-----	--	----

1	Introdução	1
1.1	Meta-ômicas	2
1.1.1	Sequenciadores de Nova Geração	2
1.1.2	Metagenômica	3
1.1.3	Metatranscritômica	3
1.2	Métodos de Classificação Taxonômica em Meta-ômicas	4
1.3	Dados Taxonômicos	5
1.4	Erros de Classificação Taxonômica	7
1.5	Combinação de Classificadores Taxonômicos (<i>Ensembles</i>)	8
1.5.1	WEVOTE	8
1.5.2	MetaMeta	9
1.5.3	McIntyre et al. (2017)	9
1.6	Proposta de uma Abordagem baseada na Integração de Meta-ômicas	10
1.7	Contexto	10
1.8	Organização do Trabalho	10
2	Métodos	12
2.1	Delineamento da Pesquisa	13
2.2	Simulação de Dados Meta-ômicos	13
2.2.1	Jovel et al. (2016)	15
2.2.2	Mande et al. (2012)	15

2.2.3	McIntyre et al. (2017) – Huttenhower	15
2.2.4	McIntyre et al. (2017) – Mavromatis	15
2.2.5	McIntyre et al. (2017)	16
2.2.6	Thompson et al. (2017) – <i>Earth Microbiome Project</i>	16
2.2.7	Singer et al. (2016) – MBARC-26	16
2.2.8	Sczyrba et al. (2017) – <i>Critical Assessment of Metagenome Interpretation</i> . . .	16
2.3	Classificadores Taxonômicos	17
2.4	Abordagens de Combinação de Classificadores Taxonômicos (<i>Ensembles</i>)	23
2.5	Análise do Desempenho dos Classificadores Taxonômicos	25
2.6	Passos no Desenvolvimento da Ferramenta de Análise Integrada	26
2.6.1	Análise de Sequências	26
2.6.2	Uniformização das Classificações Taxonômicas	26
2.6.3	Extração e Seleção de Variáveis	26
2.6.4	Modelos de Classificação	27
3	Simulação de Dados Meta-ômicos	29
3.1	<i>Pipeline</i>	30
3.2	Resultados da Simulação	33
3.3	Comparação das Amostras	34
3.3.1	Amostra de Treinamento, Teste e Validação da Ferramenta	36
4	Análise do desempenho dos classificadores	37
4.1	Desempenhos dos Classificadores Taxonômicos em Meta-ômicas SH e MT	38
4.2	Desempenho dos Classificadores Taxonômicos em 16S	41
4.3	Comparação das Abordagens de Combinação de Classificadores Taxonômicos (<i>Ensembles</i>)	44
4.4	Processo de Escolha dos Modelos de Classificação e Resultados	46
4.4.1	Extração das Variáveis e Junção de Meta-ômicas	47
4.4.2	Comparação dos Algoritmos de Classificação	47
4.4.3	Seleção e Otimização do Modelo de Classificação	52
5	Programa Gunga	53
5.1	Implementação	54
5.2	Validação do Gunga	57

5.2.1	Seleção das Amostras	57
5.2.2	Abordagem para Meta-ômicas SH e 16S	57
5.2.3	Mix3	58
5.2.4	CAMI Toy Data HC[1-5]	59
5.3	Uso do Gunga em Dados Reais	61
5.3.1	Bugios	61
5.3.2	Compostagem	62
6	Conclusões	66
6.1	Contribuições	67
6.2	Trabalhos Futuros	69
A	<i>Template</i> para Execução do Gunga	70
A.1	Preliminares	70
A.2	Execução de Código	70
A.3	Visualização do Resultado	72
B	Junção e Filtragem de Sequências de 16S no Metazoo	73
C	Publicações	75

CAPÍTULO 1

INTRODUÇÃO

Comunidades microbianas possuem papel essencial em diversos processos ambientais, por exemplo no solo [12], nos oceanos [20] e até mesmo na atmosfera [48], assim como implicações na saúde humana, por exemplo em condições como a obesidade [59].

Elas são formadas por micro-organismos de diversas espécies, que podem interagir entre si e com o meio ambiente. A compreensão de sua organização e funcionamento se dá através do estudo de suas características, sendo as principais as diversidades taxonômica e funcional. Essa caracterização ajuda a responder duas das questões fundamentais da ecologia microbiana. A diversidade taxonômica responde a questão (1) “Quais micro-organismos estão presentes nessa comunidade?”, e a diversidade funcional a questão (2) “O que esses organismos fazem?”.

A descrição de diversidade de uma comunidade se dá pela composição, riqueza e abundância. A composição é a lista das diferentes espécies ou funções que a compõe. A riqueza, é uma medida numérica geral, que representa o quão diversa ela é de acordo com o seu número de táxons ou funções. E a abundância relativa, estabelece a proporção de cada organismo ou função em relação ao total de micro-organismos ou elementos funcionais que a constituem [28,49]. Esse trabalho tem como enfoque o contexto da diversidade taxonômica.

1.1 Meta-ômicas

Uma das formas de se estudar os membros de comunidades microbianas é através da análise taxonômica, com a determinação dos conjuntos de táxons que as compõem e suas abundâncias relativas. Essa análise pode ser realizada através de um conjunto de técnicas denominadas meta-ômicas, que consiste no estudo molecular do material e produtos genéticos extraídos diretamente de uma comunidade microbiana amostrada de um ambiente, sem isolamento e cultivo prévios [18, 52].

1.1.1 Sequenciadores de Nova Geração

A disponibilidade das tecnologias de sequenciamento de nova geração permitiu um grande avanço para esses estudos, ao permitir a obtenção de sequências de DNA genômico dos diversos organismos de uma amostra [34]. Dentre as principais plataformas de sequenciamento utilizadas em metagenômica está a tecnologia Illumina. A diferença entre elas estão nas técnicas de sequenciamento utilizadas por cada uma, no rendimento de cada corrida de sequenciamento e nas características das sequências obtidas (Tabela 1.1).

Tabela 1.1: Sequenciadores de nova geração e suas características [8, 55].

Sequenciador	Rendimento (Gb)	Sequências por corrida	Tamanho da sequência (pb)
Illumina (MiSeq)	0,3–15	Até 25 milhões	2 x 300
PacBio (Sequel)	5–10	Aproximadamente 365.000	Até 60 mil
Ion Torrent (S5)	0,6–15	Até 130 milhões	Até 600
Nanopore (MinION)	0,1–1	-	Até 100 mil

1.1.2 Metagenômica

A metagenômica é uma meta-ômica que estuda o DNA extraído de uma amostra, que é fragmentado e sequenciado com o uso de sequenciadores de nova geração. As sequências obtidas são submetidas ao processo de classificação taxonômica e a análise taxonômica se dá através das listagem e contagens das associações entre suas sequências e táxons [52].

Dentre as abordagens em metagenômica, estão a *shotgun* e a de *amplicons*. A primeira se caracteriza pelo sequenciamento de trechos quaisquer de DNA genômico fragmentados aleatoriamente da amostra, enquanto na segunda é realizado o sequenciamento de uma biblioteca de fragmentos de uma porção de um gene marcador de interesse amplificado por PCR [64].

Após a obtenção das sequências, uma análise da comunidade microbiana que compõe a amostra pode ser obtida a partir de análises por ferramentas de bioinformática. Um *pipeline* para realização dessas análises em geral deve contemplar as seguintes etapas: (i) pré-processamento, onde sequências são filtradas ou editadas de acordo com as estimações de qualidade de leitura de suas bases; (ii) processamento, onde as sequências remanescentes são classificadas quanto a sua origem taxonômica a partir da comparação de sequências com um banco de dados ou algum método alternativo de classificação; e (iii) análise, onde a composição e a abundância relativa dos táxons que compõem a amostra são quantificadas a partir das classificações obtidas para as sequências, permitindo responder questões sobre a diversidade da comunidade microbiana.

1.1.3 Metatranscritômica

Uma outra meta-ômica é a metatranscritômica, que de forma análoga à metagenômica faz o estudo do RNA da comunidade amostrada. Essa técnica consiste na preparação de bibliotecas a partir da extração de RNA da amostra e sua conversão em cDNA, e posteriormente seu sequenciamento. As sequências obtidas podem então ser analisadas seguindo *pipelines* similares aos da metagenômica *shotgun*, porém podendo-se fazer uso de bancos de dados especializados em transcritos. Assim, enquanto

a metagenômica ajuda a responder a questão sobre quem compõe uma comunidade, a metatranscritômica indica quem está ativo em dados tempo e/ou condições. Dependendo dos objetivos do estudo, diferentes meta-ômicas podem ser usadas em conjunto [1, 7, 18, 52].

Nas próximas seções do texto, iremos usar as siglas SH para a metagenômica *shotgun*, 16S para metagenômica de *amplicons* de 16S e MT para metatranscritômica.

1.2 Métodos de Classificação Taxonômica em Meta-ômicas

Uma vez realizado o pré-processamento e obtidas as sequências a partir das quais serão realizadas as análises, é necessário processá-las para descobrir quais táxons compõem a comunidade microbiana amostrada. Isso é feito a partir da classificação taxonômica de cada sequência, ou seja, associando a cada sequência o táxon do qual ela é originária [32, 45].

Devido à alta complexidade das espécies, às sequências obtidas por sequenciadores de nova geração serem curtas e aos bancos de dados serem incompletos, dado o fato de que a maior parte dos organismos não são cultiváveis isoladamente, essa classificação se dá em níveis de resolução taxonômica variáveis [13, 63].

Vários métodos foram desenvolvidos para classificação taxonômica, e podemos descrevê-los da seguinte forma:

Baseados em busca por similaridade: onde a sequência é mapeada contra sequências de um banco de dados a partir de uma busca por similaridade, usando os programas BLAST [3] ou USEARCH [15] por exemplo. Em seguida, o resultado da busca é analisado e um táxon associado. Duas abordagens se destacam nesses métodos:

Best-hit: o táxon associado à sequência do banco de dados para a qual foi obtida a maior similaridade é associado à sequência da busca.

Lowest Common Ancestor (LCA): cada táxon de cada sequência do resultado da busca é mapeado em uma árvore taxonômica. De acordo com o número de resultados para cada táxon é atribuído a esse uma pontuação. O táxon com maior pontuação na hierarquia da árvore é associado à sequência, se houverem dois táxons com mesma pontuação é atribuído o táxon com maior profundidade na árvore.

Baseados em composição: onde a sequência é classificada através da extração de características

de suas propriedades composicionais, tais como subsequências exatas (chamadas k -mers), e a comparação do seu padrão composicional com padrões aprendidos de conjunto de sequências depositadas em banco de dados com classificações taxonômicas conhecidas, comumente através de métodos de aprendizado de máquina, tais como Máquinas de Vetor de Suporte (*Support Vector Machines* – SVMs) e *Naive Bayes* [60].

Híbridos: onde as duas abordagens anteriores são combinadas para classificar a sequência. Um exemplo de programa dessa abordagem é o PhymmBL [9]. Com as pontuações obtidas por um método de reconhecimento de padrões em sequências com Modelos de Markov Interpolados (*Interpolated Markov Models* – IMM), e uma busca usando BLAST [3], ele combina resultados para construir uma pontuação com maior confiança para a classificação taxonômica da sequência.

Baseados em marcadores: onde a sequência é classificada da mesma maneira baseada na busca por similaridade, se diferenciando apenas pela busca ocorrer em banco de dados específicos de genes marcadores. Algumas abordagens utilizam genes marcadores universais, como o gene que codifica o rRNA 16S em procariotos, e outras utilizam genes considerados específicos de determinados clados [33, 39].

Uma abordagem comum em metagenômica de 16S é a aplicação do conceito de Unidade Taxonômica Operacional (*Operational Taxonomic Unit* – OTU), que de forma geral, para procariotos, descreve que sequências de 16S com mais de 97% de similaridade são de organismos da mesma espécie. Nessa abordagem, as sequências são agrupadas a partir de sua comparação, uma sequência representativa de cada grupo formado é selecionada e classificada taxonomicamente, e por fim a classificação obtida para essa sequência é propagada para todas as demais do seu grupo [41, 64].

Os programas de classificação taxonômica comumente atribuem uma pontuação (*score*) sobre a confiança da classificação de uma sequência. Essa pontuação não é padronizada e pode variar de acordo com o classificador. Por exemplo, no programa Kraken [62], esse valor é dado pela razão entre o número de subsequências que mapearam para o táxon classificado e o número total de subsequências da sequência classificada (<http://ccb.jhu.edu/software/kraken/MANUAL.html>).

1.3 Dados Taxonômicos

Os dados taxonômicos são o resultado da classificação taxonômica. A abundância relativa de um táxon é calculada através da contagem de sequências associadas a este, como a proporção entre essas

sequências e o total de sequências do conjunto.

Dessa forma, os dados taxonômicos possuem duas representações. A primeira como uma lista de associações entre sequências e táxons. A Tabela 1.2 é um exemplo dessa representação. Para um dado taxonômico com n sequências classificadas por um programa qualquer, nele a sequência S_1 foi associada à espécie *Clostridium perfringens*, a sequência S_2 à espécie *Clostridium septicum*, e assim por diante.

Tabela 1.2: Exemplo de dado taxonômico como uma lista de sequências e respectivos táxons associados. Nos referimos ao dado taxonômico nesse formato como **classificação taxonômica**.

Sequência (ID)	Táxon (apenas maior nível de resolução taxonômica)
S_1	<i>Clostridium perfringens</i>
S_2	<i>Clostridium septicum</i>
S_3	<i>Staphylococcus aureus</i>
S_4	<i>Escherichia coli</i>
S_5	<i>Firmicutes</i>
...	...
S_n	<i>Firmicutes</i>

A segunda representação é como uma lista de associações entre os táxons classificados e suas respectivas abundâncias relativas. Embora seja comum o interesse em definir-se a abundância relativa de cada táxon, essa definição pode ser complexa, podendo levar em conta normalizações que necessitam de informações prévias, como o tamanho médio dos genomas [17]. Portanto, para esse trabalho, também será considerada a representação dos dados como uma tabela de percentuais, substituindo a abundância relativa pelo percentual de sequências associadas ao táxon, dentro do total de todas as sequências do dado taxonômico. Por exemplo, na Tabela 1.3 é representado que para 17,4% das sequências foi associada a espécie *Clostridium perfringens* pelo classificador taxonômico.

A classificação taxonômica segue uma estrutura hierárquica, do menor nível de resolução taxonômica para o maior, sendo eles: reino $\rightarrow$ filo $\rightarrow$ classe $\rightarrow$ ordem $\rightarrow$ família $\rightarrow$ gênero $\rightarrow$ espécie. Dados os tamanhos curtos das sequências, os métodos de classificação associam táxons com níveis de resolução variáveis a essas sequências. É de interesse, a obtenção do maior nível de resolução taxonômica possível.

Nas Tabelas 1.2 e 1.3 apresentamos os táxons apenas em seu maior nível de resolução taxonômica, porém implicitamente todos eles possuem essa hierarquia. Por exemplo, para a sequência S_1 o táxon associado é a espécie *Clostridium perfringens*, uma bactéria do filo *Firmicutes*, e sua hierarquia completa é Bacteria $\rightarrow$ Firmicutes $\rightarrow$ Clostridia $\rightarrow$ Clostridiales $\rightarrow$ Clostridiaceae $\rightarrow$ Clostridium $\rightarrow$

Tabela 1.3: Exemplo de dado taxonômico como uma lista de táxons e suas respectivas representações percentuais. Nos referimos ao dado taxonômico nesse formato como **perfil taxonômico**. Os táxons relacionados no perfil são chamados de **predições taxonômicas**.

Táxon (apenas maior nível de resolução taxonômica)	Proporção de sequências
<i>Clostridium perfringens</i>	17,4%
<i>Clostridium septicum</i>	2%
<i>Staphylococcus aureus</i>	7,8%
<i>Escherichia coli</i>	9,75%
<i>Pseudomonas</i>	8,2%
...	...
<i>Firmicutes</i>	1%

Clostridium perfringens. Já a sequência S_n foi classificada como Firmicutes, porém esse é o maior nível de resolução obtido pelo classificador, e para os demais níveis de resolução na hierarquia (exceto reino, que é Bacteria) essa sequência tem sua classificação indefinida.

1.4 Erros de Classificação Taxonômica

Diversos programas foram desenvolvidos para a análise taxonômica em meta-ômicas [21, 23, 35]. Porém, a classificação taxonômica de suas sequências ainda apresenta desafios – esses programas identificam apenas parte da composição real das amostras, e além disso, identificam organismos que não fazem realmente parte da comunidade microbiana analisada [32].

Lindgreen et al. (2016) [32] realizaram uma avaliação sobre acurácia e tempo de execução de 14 ferramentas de classificação taxonômica, em nível de filo e gênero, com o uso de dados simulados de SH. Para o nível de filo, as ferramentas foram capazes de classificar entre 39,91% e 73,68% das sequências em média. E a maioria das ferramentas (12 das 14) apresentaram baixas taxas de falsos positivos, abaixo de 1% em média. CLARK [44], Kraken [62] e One Codex [40] foram as que demonstraram o melhor desempenho, classificando em média mais de 70% das sequências com taxas quase nulas de falsos positivos.

Por outro lado, o trabalho destaca as médias das taxas de falsos positivos obtidas com a ferramenta EBI *metagenomics* [23], que foi de 41,74% (incluindo como falsos positivos sequências classificadas como “*Other bacteria or archaea*”, para os casos de táxons que estão representados em seu banco de dados). A ferramenta classificou em média 0,08% das sequências apenas. Eles destacam também que o Taxator-tk [14] apresentou uma média de 14,07% de falsos positivos.

Para o nível de gênero, os autores demonstram que os programas perdem precisão, mas não forneceram os valores dos percentuais de falsos positivos. Para eles, qualquer resultado de estudo de metagenômica é afetado com erros quanto à sua análise taxonômica. Além disso, eles concluíram que algumas das ferramentas mais utilizadas na literatura não são as mais acuradas, e que as ferramentas mais acuradas não são necessariamente as que demandam maior tempo de processamento.

1.5 Combinação de Classificadores Taxonômicos (*Ensembles*)

Poucas abordagens foram propostas para reduzir esses erros originados da análise taxonômica de meta-ômicas individuais. Sendo as três abordagens descritas a seguir as únicas até o momento, restritas em suas publicações originais à metagenômica do tipo *shotgun*.

1.5.1 WEVOTE

Metwally et al. (2016) [39] desenvolveram para SH o WEVOTE (*WEighted VOting Taxonomic idEntification*), para a combinação de classificadores taxonômicos (*ensemble*) baseado em um sistema de votação (*voting*). Nele, cada classificação taxonômica associada a uma sequência dada por cada programa é interpretada como um “voto” para que tal classificação seja “escolhida”. Em razão dos diferentes níveis de resolução taxonômica em que os programas classificam as sequências, o WEVOTE constrói uma árvore taxonômica para os táxons classificados e atribui um peso para cada táxon igual ao número de vezes em que o táxon é classificado, o táxon com maior peso é escolhido, e no caso de haver um empate o LCA dos táxons empatados é escolhido.

O programa permite a combinação de resultados de três diferentes classificadores, BLASTn [3] (*best hit*), TIPP [42] e Kraken [62], ou cinco, com adição dos resultados obtidos com MetaPhlAn [53] e CLARK [44]. Os resultados obtidos com a avaliação do programa demonstraram que sua estratégia permitiu a redução de falsos positivos, preservando o mesmo nível de sensibilidade apresentado dentre os classificadores considerados com alta sensibilidade.

1.5.2 MetaMeta

Piro et al. (2017) [47] apresentaram o MetaMeta, como um outro método de combinação de classificadores para SH. Diferentemente do WEVOTE, ele executa diferentes classificadores e integra apenas os perfis taxonômicos obtidos. Seu método é baseado na co-ocorrência dos táxons nos diferentes perfis, na filtragem de táxons por abundância relativa (menor que 0,01%), e pondera a presença ou não do táxon nos bancos de dados dos programas utilizados. Em seu resultado final, o programa faz uma normalização da abundância relativa.

O método foi testado com os programas Kraken [62], Clark [44], Kaiju [38], DUDes [46], GOTTECHA [19] e mOTUs [56]. E seus resultados demonstraram que ele é capaz de eliminar erros de classificação, mantendo a maioria das classificações corretas. Embora seja posterior ao WEVOTE [39], o trabalho não comparou seus desempenhos.

1.5.3 McIntyre et al. (2017)

McIntyre et al. (2017) [36] também estudaram a filtragem de resultados e a combinação de classificadores para SH. Eles usaram 35 conjuntos de dados, e concluíram que combinar ferramentas com diferentes abordagens de classificação pode melhorar a precisão do perfil taxonômico obtido, mas nem sempre. Eles testaram várias combinações de classificadores 2 a 2, baseadas na interseção dos táxons. Porém, o MEGAN [24] com alinhamentos com BLAST [3] obteve desempenho superior a várias delas.

Eles apresentaram uma abordagem para combinação mais complexa também, com os classificadores MEGAN [24] (com o uso do BLAST [3] para alinhamentos), GOTTECHA [19], Kraken [62], LMAT [4] e MetaPhlAn [53]. Essa abordagem foi a que apresentou o melhor desempenho, em comparação tanto com outras combinações quanto com todos os classificadores individualmente.

Além das combinações, eles analisaram o uso da filtragem por abundância relativa menor que 0,01% com classificadores individuais. Eles demonstraram que essa filtragem é capaz de melhorar a precisão no caso de alguns classificadores. Novamente, os autores não fizeram uma comparação com o WEVOTE [39].

1.6 Proposta de uma Abordagem baseada na Integração de Meta-ômicas

Embora esses trabalhos recentes tenham proposto abordagens para diminuir os erros da análise taxonômica [36,39,47], nenhum deles estudou a integração de diferentes meta-ômicas para o problema.

Neste trabalho, iremos lidar com esse problema com o desenvolvimento de uma ferramenta para a melhoria dos perfis taxonômicos de amostras de meta-ômicas, com a discriminação de erros de classificação, através da integração dos dados de análise taxonômica de diferentes meta-ômicas.

1.7 Contexto

Esse trabalho foi proposto no contexto do projeto Metazoo, que tem o intuito de caracterizar, através de técnicas meta-ômicas, os perfis taxonômico e funcional de comunidades microbianas de três ambientes do zoológico de São Paulo: uma compostagem, um lago e fezes de macacos que vivem no zoológico. A carência de ferramentas para a análise integrativa de dados em meta-ômicas, e especificamente para a integração de diferentes perfis taxonômicos obtidos por diferentes meta-ômicas, foi inicialmente percebida com o andamento das análises de dados do projeto.

1.8 Organização do Trabalho

Esse trabalho está organizado nos capítulos a seguir da seguinte forma:

- No Capítulo 2 nós apresentamos a metodologia usada para alcançar os resultados, com detalhes sobre a escolha de dados das diferentes meta-ômicas, análise taxonômica com diversas ferramentas, avaliação dessas ferramentas, treinamento e avaliação dos modelos de Aprendizado de Máquina (AM), e a proposta para incorporação desses modelos em uma ferramenta.
- No Capítulo 3 apresentamos nossa abordagem para a simulação de amostras para serem usadas no treinamento e avaliação da abordagem desenvolvida.
- No Capítulo 4 nós apresentamos os resultados obtidos das avaliações dos classificadores taxonômicos em meta-ômicas individuais e das abordagens baseadas na integração de dados.

- No capítulo 5 nós apresentamos a arquitetura da ferramenta desenvolvida, e sua validação, e demonstramos seu uso em dados reais do projeto Metazoo.
- No Capítulo 6 apresentamos as conclusões do trabalho.

CAPÍTULO 2

MÉTODOS

2.1 Delineamento da Pesquisa

Definimos o nosso problema como o desenvolvimento de um modelo de classificação onde uma **predição taxonômica** em uma amostra é classificada como **correta** ou **incorreta**. Consideramos um cenário em que: (1) obtivemos essa predição por análises taxonômicas de duas ou três meta-ômicas dentre SH, MT e 16S; (2) o mesmo classificador taxonômico foi utilizado para as análises das meta-ômicas SH e MT, porém um classificador diferente foi usado para 16S; (3) cada classificador taxonômico fornece diferentes informações sobre as predições taxonômicas; e (4) as diferentes análises podem ter predições taxonômicas comuns às duas ou três meta-ômicas e exclusivas, que ocorreram em apenas uma das meta-ômicas. Em um primeiro passo obtemos um perfil integrado com os dados das diferentes meta-ômicas através de uma junção dos perfis taxonômicos individuais. Em um segundo, aplicamos um modelo de aprendizado de máquina com as características obtidas do perfil integrado para classificar suas predições taxonômicas.

No contexto do aprendizado de máquina, um modelo de classificação pode ser aprendido a partir de exemplos. Esses exemplos são usados para ajustar os parâmetros de um modelo que associa os valores das variáveis de entrada (atributos) com rótulos conhecidos (classes). Ou seja, para um conjunto de exemplos $\{x_1, x_2, \dots, x_n\}$ com rótulos associados $\{y_1, y_2, \dots, y_n\}$, um modelo de classificação $\hat{M}$ pode ser ajustado para prever $\hat{y}$ de um exemplar x com rótulo desconhecido. Nós aplicamos uma abordagem onde as predições taxonômicas dos perfis integrados são usadas como exemplos para o ajuste de modelos de classificação. Nós avaliamos diferentes variáveis e algoritmos afim de selecionar o melhor algoritmo para o nosso problema.

2.2 Simulação de Dados Meta-ômicos

Usamos dados de sequenciamento de meta-ômicas (SH, MT e 16S) de comunidades sintéticas (*mocks*) e simulados como referência para desenvolver e avaliar a abordagem para integração dessas meta-ômicas. Além disso, usamos os mesmos dados para comparar os desempenhos dessa abordagem e dos programas de classificação taxonômica para meta-ômicas individuais, incluindo aqueles que usam métodos para combinação de classificadores para uma mesma meta-ômica. Usamos dados reais para demonstrar o uso de um programa implementado para viabilizar o uso da abordagem desenvolvida.

Usamos o Grinder [5] para simular os dados de sequenciamento, com a geração de sequências *single-end* com 250 pares de bases e desvio-padrão +/- 50, usando os parâmetros fornecidos pelo

próprio simulador para obtenção de sequências com um perfil de erros compatível com as tecnologias de sequenciamento Illumina. Para SH e MT nós simulamos amostras com 1 milhão de sequências. Para 16S, nós simulamos amostras com 100 mil sequências, baseadas na amplificação da região V4 do gene do rRNA 16S com o par de *primers* 515F-806R (FWD:GTGCCAGCMGCCGCGGTAA; REV:GGACTACHVGGGTWTCTAAT). Os comandos usados foram:

```
# Arquivo de primers:
```

```
>FWD
```

```
GTGCCAGCMGCCGCGGTAA
```

```
>REV
```

```
GGACTACHVGGGTWTCTAAT
```

```
# Comandos:
```

```
# SH
```

```
grinder -reference_file /genomes/path -af sh/ranks/path -total_reads 1000000 \  
-read_dist 250 normal 50 -qual_levels 30 10 -md poly4 3e-3 3.3e-8 \  
-mr 80 20 -length_bias 1 -fastq_output 1 -output_dir sh -random_seed 1
```

```
# MT
```

```
grinder -reference_file /cdss/path -af mt/ranks/path -total_reads 1000000 \  
-read_dist 250 normal 50 -qual_levels 30 10 -md poly4 3e-3 3.3e-8 \  
-mr 80 20 -length_bias 1 -fastq_output 1 -output_dir mt -random_seed 1
```

```
# 16S
```

```
grinder -reference_file /genomes/path -af amp/ranks/path -total_reads 100000 \  
-read_dist 250 normal 50 -qual_levels 30 10 -md poly4 3e-3 3.3e-8 \  
-mr 80 20 -length_bias 0 -fastq_output 1 -output_dir amp -random_seed 1 \  
-fr primers/path -unidirectional 1 -chimera_perc 10 -copy_bias 1
```

Ao total, simulamos o sequenciamento de dados meta-ômicos de SH, MT e 16S de 56 comunidades, com base em composições taxonômicas de amostras descritas em seis estudos.

2.2.1 Jovel et al. (2016)

Sequências de 16S e SH de 3 comunidades *mocks* foram disponibilizados em Jovel et al. (2016) [26], intituladas Mix 1, Mix 2 e Mix 3. As amostras foram formadas com DNA de 11 espécies: *Salmonella enterica*, *Streptococcus pyogenes*, *Escherichia coli*, *Lactobacillus helveticus*, *Lactobacillus delbrueckii*, *Lactobacillus plantarum*, *Clostridium sordelli*, *Bacteroides thetaiotaomicron*, *Bacteroides vulgatus*, *Bifidobacterium breve* e *Bifidobacterium animalis*. Elas foram sequenciadas com a plataforma Illumina MiSeq (<https://www.illumina.com/systems/sequencing-platforms/miseq.html>) e disponibilizadas no SRA [31] (SRP059928). Portanto, para essas comunidades simulamos apenas dados de MT.

2.2.2 Mande et al. (2012)

Nós também simulamos o sequenciamento por SH, 16S e MT de amostras baseadas nos dados de Mande et al. (2012) [33]. Nós usamos composições de 10, 100 e 400 espécies para simular essas amostras. Para essas composições nós simulamos dois tipos de amostras: (i) com percentual igual para todos os táxons que compõem a amostra; e (ii) com percentuais diferentes para os táxons, sendo alguns táxons predominantes, e a maior parte deles em baixa abundância.

2.2.3 McIntyre et al. (2017) – Huttenhower

McIntyre et al. (2017) [37] desenvolveu um dos estudos mais extensos em número de amostras, eles analisaram 35 amostras de dados simulados para SH. Nós nos baseamos nas composições que esse estudo disponibilizou e simulamos os dados para SH, MT e 16S. Nós usamos as composições denominadas Huttenhower LC e HC (que se refere a *Low e High Complexity*) e simulamos 8 amostras de baixa complexidade (LC1, LC2, LC3, LC4, LC5, LC6, LC7 e LC8) e 2 de alta complexidade (HC1 e HC2).

2.2.4 McIntyre et al. (2017) – Mavromatis

Do mesmo estudo nós usamos as composições para o grupo de amostras denominado de Mavromatis, que inclui 3 amostras de baixa, média e alta complexidade.

2.2.5 McIntyre et al. (2017)

Além disso, mais 8 amostras baseadas na composição de diferentes ambientes, como solo ou regiões de uma cidade, foram simuladas – de baixa, média e alta complexidade, com suas comunidades microbianas compostas com diferentes números de genomas entre 14 e 551.

2.2.6 Thompson et al. (2017) – *Earth Microbiome Project*

O *Earth Microbiome Project* (EMP) [57] realizou um levantamento baseado em 16S para diversos ambientes do planeta, incluindo plantas, solo, regiões do corpo de animais, dentre outros. Nós usamos 17 desses ambientes descritos para extrair suas composições e simular amostras com os supostos perfis taxonômicos esperados para esses ambientes. Uma vez que o trabalho só caracterizou as amostras em nível de gênero, nós sorteamos para cada gênero uma espécie, e usamos os genomas e CDSs dessas espécies para simular as sequências.

2.2.7 Singer et al. (2016) – MBARC-26

MBARC-26 é a identificação de uma comunidade sintética composta por 26 microorganismos [54]. Essa composição foi disponibilizada para avaliar os protocolos de sequenciamento e análise das amostras, uma vez que se conhece a realidade sobre a mesma. Porém só há dados de sequenciamento para essas amostras em SH. Nós então nos baseamos em sua composição e simulamos as três meta-ômicas para avaliar como se daria a análise taxonômica dessa amostra com as diferentes meta-ômicas.

2.2.8 Sczyrba et al. (2017) – *Critical Assessment of Metagenome Interpretation*

Nós também simulamos dados baseados nas composições de amostras do projeto *Critical Assessment of Metagenome Interpretation* (CAMI) [51]. Esse projeto tem por objetivo desenvolver amostras para que ferramentas de análise metagenômica possam ser avaliadas de forma isenta, ou seja, sem que se escolha uma amostra que seja favorável para os resultados de uma determinada ferramenta que está sendo desenvolvida. Nós simulamos 8 amostras baseadas nesse projeto.

2.3 Classificadores Taxonômicos

Nós selecionamos um subconjunto dos classificadores taxonômicos usados nos trabalhos de Metwally et al. (2016) [39], Piro et al. (2017) [47] e McIntyre et al. (2017) [37], afim de reduzir ao máximo a influência dos classificadores taxonômicos escolhidos como fator preponderante para resultados baseados na integração e prover comparabilidade de desempenho entre as abordagens propostas nesses trabalhos e a nossa.

A lista dos classificadores para SH e MT com suas características e os estudos a partir dos quais foram selecionados está presente na Tabela 2.1. Além desses, nós também avaliamos os classificadores Kraken 2 (nova versão do Kraken) e o Centrifuge [27], que usa uma abordagem similar a do Kraken, porém possui um esquema de compressão do banco de dados que permite que ele seja executado em um computador com baixo uso de memória (cerca de 4GB).

Tabela 2.1: Classificadores taxonômicos usados para meta-ômicas SH e MT.

Classificador	Tipo	Estudo	Referência
DIAMOND (Best Hit)	Alinhamentos	Metwally	[10]
Kraken	Composicional	Metwally Piro McIntyre	[62]
TIPP	Marcadores	Metwally	[42]
CLARK	Composicional	Metwally Piro McIntyre	[44]
MetaPhlAn2	Marcadores	Metwally McIntyre	[58]
DUDes	Alinhamentos	Piro	[46]
GOTTCHA	Marcadores	Piro McIntyre	[19]
Kaiju	Composicional	Piro	[38]
mOTUs	Marcadores	Piro	[56]
CLARK- <i>S</i>	Composicional	McIntyre	[43]
LMAT	Composicional	McIntyre	[4]
MEGAN (DIAMOND, com LCA)	Alinhamentos	McIntyre	[24]

Para 16S usamos os classificadores QIIME [11] (RDP [61] e UCLUST [15]) e avaliamos o classificador para nível de espécie e gênero SPINGO [2]. A seguir temos uma visão geral de cada classificador.

DIAMOND

O DIAMOND é um alinhador que contém uma heurística para busca em bases de dados de proteínas, uma alternativa mais rápida ao BLASTX ou BLASTP. Nesse contexto nós usamos a busca no modo BLASTX, onde as sequências são traduzidas em sequências de proteínas e uma busca por

similaridade é realizada contra o banco de dados *nr* do NCBI. Em nosso caso, nós usamos o resultado da busca para obter o melhor alinhamento (abordagem *Best-Hit*) e usar o táxon associado à sequência recuperada como classificação taxonômica da sequência consulta.

Os comandos usados foram:

```
diamond blastx -d /work/databases/diamond/nr \
-q qc_filtered.fasta -a diamond
diamond view -a diamond.daa -o diamond.besthit.m8 -k 1
blast2lca --input diamond.besthit.m8 --format BlastTab --topPercent 10 \
--acc2taxa prot_acc2tax-Mar2018X1.abin \
--output megan_classifications.besthit.tsv
```

Kraken (1 e 2)

É um classificador baseado em padrões composicionais. O algoritmo do Kraken divide as sequências em subsequências de tamanho fixo (k-mers) e faz casamentos exatos com um banco de dados pré-compilado com essas subsequências. Com os resultados dos táxons que casam com cada subsequência o Kraken atribui a classificação taxonômica usando o táxon com maior número de casamentos, ou no caso de empate ele soluciona a classificação usando o algoritmo LCA.

Os comandos usados foram:

```
kraken/kraken2 --db [KRAKEN_DB] --threads 12 qc_filtered.fastq --fastq-input \
--output classifications.tsv --report report.tsv # --report apenas em Kraken 2

kraken-filter --db [KRAKEN_DB] classifications.tsv > scores.tsv
kraken-report --db [KRAKEN_DB] classifications.tsv > report.tsv # apenas em Kraken 1
```

Centrifuge

Usa o mesmo procedimento que o Kraken, porém tem um banco de dados que é comprimido para permitir o uso de menos memória.

Os comandos usados foram:

```
centrifuge -q -x p+h+v -U qc_filtered.fastq -S classifications.tsv \
--report-file report.tsv -k 1
```


TIPP

Baseado em marcadores, usa um conjunto de 30 genes marcadores universais de cópia única. Usa o HMMER e pplacer para associar a sequência consulta a uma posição numa árvore filogenética pré-computada e associar a classificação taxonômica baseada nesse posicionamento.

Os comandos usados foram:

```
tipp/sepp/run_abundance.py -f qc_filtered.fasta \  
-c $HOME/.sepp/tipp.config -x 16 -d tipp_profile
```

CLARK

Também usa uma abordagem composicional, mas na construção do seu banco de dados apenas regiões que mapeiam unicamente num genoma são mantidas. O táxon com maior número de casamentos com as subsequências da sequência consulta é usado como classificação.

Os comandos usados foram:

```
set_targets.sh [CLARK_DB] bacteria viruses  
classify_metagenome.sh -O qc_filtered.fastq -R classifications -n 16  
estimate_abundance.sh -D [CLARK_DB] -F classifications.csv > report.tsv
```

MetaPhlAn2

É um classificador baseado em marcadores. O MetaPhlAn2 faz uma busca contra um banco de dados de marcadores por mapeamento da sequência consulta com sequências de genes marcadores, usando o Bowtie2. De acordo com os marcadores encontrados ele relaciona o táxon associado com o marcador, que pode ser em nível de espécie ou qualquer outro nível, no perfil taxonômico – estimando a abundância de acordo com a cobertura desses marcadores inclusive.

Os comandos usados foram:

```
metaphlan2.py qc_filtered.fastq --input_type fastq --nproc 16 \  
--bowtie2out metaphlan2.bowtie.bz2 > metaphlan2_profile.tsv
```

DUDES

Faz o mapeamento da sequência consulta em um banco de dados e usa um algoritmo recursivo para ir aprofundando a associação taxonômica em torno dos táxons mais prováveis através de um

conceito que nomeou como “*deepest uncommon descendent*”.

Os comandos usados foram:

```
bowtie2 -x [DUDES_DB] --no-unal --very-fast -k 10 \  
-U qc_filtered.fastq -S dudes.sam  
DUDes.py -s dudes.sam -d [DUDES_DB].npz -o dudes_profile
```

GOTTCHA

Usa uma abordagem baseada em marcadores. Seu algoritmo divide a sequência consultada em sub-sequências sem sobreposição, e a busca é realizada com casamentos exatos usando BWA.

Os comandos usados foram:

```
gottcha.pl --threads 16 --outdir [OUTDIR] --prefix [PREFIX] \  
--input qc_filtered.fastq --database [GOTTCHA_DB]
```

Kaiju

Abordagem semelhante ao Kraken, porém, usa sequências de proteínas para a construção do seu banco de dados de k-mers e usa também algoritmos para reduzir o uso de memória de seu índice para busca.

Os comandos usados foram:

```
kaiju -v -f [KAIJU_DB] -t [KAIJU_NODES] \  
-i qc_filtered.fastq -o classifications.tsv  
kaijuReport -t [KAIJU_NODES] -n [KAIJU_NAMES] \  
-i classifications.tsv -r species -o report.tsv
```

mOTUs

Usa genes marcadores universais de cópia única. A classificação taxonômica é realizada fazendo o mapeamento da sequência consultada contra esses genes.

Os comandos usados foram:

```
perl mOTUs.pl qc_filtered.fastq --output-directory motus_profile
```

CLARK-S

O mesmo que o CLARK, porém possui um refinamento pois admite o uso de sementes (k-mers) espaçadas (o “s” se refere a “*spaced seeds*”).

Os comandos usados foram:

```
set_targets.sh [CLARK_DB] bacteria viruses
classify_metagenome.sh -O qc_filtered.fastq -R classifications \
-n 16 --spaced
estimate_abundance.sh -D [CLARK_DB] -F classifications.csv > report.tsv
```

LMAT

É um classificador composicional, que possui um banco de dados próprio baseado no RefSeq do NCBI. Assim como outros classificadores baseados em composição ele faz o casamento com as sequências de seu banco de dados e associa a classificação a partir do LCA dos táxons das sequências resultado.

Os comandos usados foram:

```
run_rl.sh --db_file=lmata-4-14.20mer.db --query_file=qc_filtered.fasta \
--threads=20 --odir=profile --overwrite
```

MEGAN

Usa como entrada os resultados de alinhamento do Blast ou DIAMOND. No caso, usamos como entrada os alinhamentos obtidos com DIAMOND. Para resolver a classificação taxonômica, usa os táxons associados às sequências resultado e implementa o algoritmo LCA.

Os comandos usados foram:

```
diamond blastx -d /work/databases/diamond/nr \
-q qc_filtered.fasta -a diamond
diamond view -a diamond.daa -o diamond.m8
blast2lca --input diamond.m8 --format BlastTab --topPercent 10 \
--acc2taxa prot_acc2tax-Mar2018X1.abin \
--output megan_classifications.lca.tsv
```

QIIME – com UCLUST e RDP *Classifier*

É baseado em marcadores, com o gene 16S. O UCLUST classifica uma sequência comparando a similaridade da sequência com sua base de OTUs pré-classificadas. Por isso, é usado geralmente para a classificação de *amplicons*. O RDP é composicional, pois usa k-mers para fazer a busca por casamentos das sequências, porém para resolver a classificação usa os casamentos encontrados e faz a classificação usando o algoritmo *Naive Bayes*, associando o táxon com maior probabilidade. No QIIME, por padrão, ambos classificadores tem as implementações de seus algoritmos baseadas no banco de dados de 16S do GreenGenes.

Os comandos usados foram:

```
split_libraries_fastq.py -i raw.fastq -o qiime/slout -q 19 \
  --barcode_type 'not-barcoded' --sample_ids [ID] --phred_offset 33

## Open Reference OTUs + UCLUST
pick_open_reference_otus.py -i qiime/slout/seqs.fna \
  -o qiime/open_reference_otus -p qiime_parameters.txt
summarize_taxa.py \
  -i qiime/open_reference_otus/otu_table_mc2_w_tax.biom \
  -o qiime/open_reference_otus/uclust_summary -L 1,2,3,4,5,6,7
biom convert \
  -i qiime/open_reference_otus/otu_table_mc2_w_tax.biom \
  -o qiime/open_reference_otus/otu_table.txt --to-tsv --header-key taxonomy

## Open Reference OTUS + RDP Classifier
assign_taxonomy.py -i qiime/open_reference_otus/rep_set.fna \
  -o qiime/open_reference_otus/rdp_assigned_taxonomy -m rdp
biom add-metadata \
  -i qiime/open_reference_otus/otu_table_mc2.biom \
  --observation-metadata-fp \
  qiime/open_reference_otus/rdp_assigned_taxonomy/rep_set_tax_assignments.txt \
  -o qiime/open_reference_otus/rdp_otu_table_mc2_w_tax.biom \
  --sc-separated taxonomy --observation-header OTUID,taxonomy
```

```

summarize_taxa.py \
-i qiime/open_reference_otus/rdp_otu_table_mc2_w_tax.biom \
-o qiime/open_reference_otus/rdp_summary -L 1,2,3,4,5,6,7

```

SPINGO

Usa um banco de dados baseado na base de sequências de 16S RDP e fornece classificação taxonômica diretamente compatível com a taxonomia do NCBI. Usa subsequências exatas para mapear a sequência contra seu banco de dados e faz uma pontuação para o táxon com maior correspondência.

Os comandos usados foram:

```

spingo -d RDP_11.2.species.fa -i qiime/slout/seqs.fna > spingo.out
spingo_summary -l 2 -p spingo.out > genus.out
spingo_summary -l 3 -p spingo.out > species.out

```

2.4 Abordagens de Combinação de Classificadores Taxonômicos (*Ensembles*)

Nós obtivemos perfis taxonômicos usando abordagens de combinação de classificadores em SH com os programas citados na Seção 1.5:

WEVOTE [39]: nós escolhemos como entrada as classificações taxonômicas obtidas com o Kraken, DIAMOND (*Best-Hit*, com banco de dados nr) e CLARK, com base em seu uso pela plataforma UGENE (<http://ugene.net/>). Essa escolha reflete a complementaridade entre métodos de classificação esperada pelo o WEVOTE, com Kraken e CLARK como representantes de métodos composicionais e DIAMOND como representante dos métodos baseados em alinhamentos [50]. O DIAMOND é usado em detrimento do BLASTX por apresentar menor tempo de execução e manter um grau similar de sensibilidade [10].

Programamos *scripts* para processar os arquivos de saída desses programas e preparar o arquivo de entrada do WEVOTE, e rodamos o programa com parâmetros padrão com os seguintes comandos:

```

./WEVOTE -i /wevote/input/path -d /wetove/db/path -n 16 -p wevote

```

```
./run_ABUNDANCE.sh -i wevote_WEVOTE_Details.txt -p wevote --db /wetove/db/path
```

MetaMeta [47]: nós usamos como entrada arquivos FASTQ com sequências das amostras filtradas por qualidade ($Q \geq 20$). Executamos o programa para obter resultados com seu *pipeline* para a combinação dos resultados de todos os seis classificadores taxonômicos pré-configurados (CLARK, DUDES, GOTTECHA, Kaiju, Kraken e mOTUs). As classificações das sequências com cada classificador separadamente foram realizadas internamente dentro do seu próprio fluxo de trabalho. Com parâmetros padrão, o programa é configurado para a classificação de procariotos através do uso de seus bancos de dados baseados em genomas completos de arqueias e bactérias do RefSeq do NCBI, referenciados pelo programa como **archaea_bacteria_201503**.

Nós realizamos a instalação do programa via Miniconda (<https://conda.io/miniconda.html>) e o rodamos com parâmetros padrão em um ambiente com os seguintes arquivo e comandos:

```
# Arquivo de configuração:
workdir: "/metameta/work/path/"
dbdir: "/metameta/database/path/"
samples:
    "SampleName":
        fq1: "/quality/filtered/fastq/path"
gzipped: 0
threads: 1

# Comandos:
source activate metametaenv
metameta --configfile /metameta/config/path --use-conda --keep-going --cores 16
```

Majority Vote [37]: nós aplicamos o método de voto majoritário (*Majority Vote*) com classificações e perfis taxonômicos obtidos com os programas MEGAN (com DIAMOND, com banco de dados nr e resolução da classificação com o algoritmo LCA), GOTTECHA, Kraken (com filtragem de classificações com pontuação maior ou igual a 0,2), LMAT e MetaPhlAn2, denominado **DiamondEnsemble**. O método considera a predição de um táxon se ele estiver relacionado

nos resultados de ao menos 3 das 5 ferramentas. O **DiamondEnsemble** foi escolhido pois foi o método com maior mediana para a medida $F1$ (que representa um balanceamento entre precisão e sensibilidade, ver a Seção 2.5 a seguir) dentre as abordagens de combinação de classificadores analisadas apresentadas no estudo de McIntyre et al. (2017). Como os autores não forneceram um programa, nós programamos nossa própria implementação dessa abordagem em R (disponível em <https://github.com/felipepratalima/gunga/tree/master/Rmd/MajorityVoteDevelopment.html>).

Experimentamos também essas abordagens em MT, dado que a análise taxonômica dessa meta-ômica ocorreu com o uso dos mesmos programas usados em SH, resultando nos dados de entrada necessários para aplicação dessas abordagens. Em seus estudos originais não há relatos sobre o uso dessas abordagens com MT, portanto esses resultados também foram obtidos afim de verificar a aplicabilidade delas com essa meta-ômica.

2.5 Análise do Desempenho dos Classificadores Taxonômicos

Nós avaliamos o desempenho dos perfis taxonômicos obtidos com a precisão, a sensibilidade e a medida $F1$ para os níveis de filo, gênero e espécie. Há várias medidas para se avaliar o desempenho de modelos de classificação, porém escolhemos a medida $F1$ por ela fazer o balanceamento entre precisão e sensibilidade. Ela também se mostra mais adequada para conjuntos de dados com classes desbalanceadas, nós observamos que a fração de predições incorretas é bem maior do que as predições corretas para diversos classificadores taxonômicos.

Definimos os verdadeiros positivos (VP) como táxons relacionados no perfil que realmente fazem parte da amostra, os falsos positivos (FP) como aqueles relacionados no perfil que não fazem parte, e os falsos negativos (FN) como os táxons que fazem parte da amostra que não foram relacionados no perfil. A precisão é dada por $VP/(VP + FP)$, a sensibilidade por $VP/(VP + FN)$ e a medida $F1$ por $(2 \times \text{precisão} \times \text{sensibilidade})/(\text{precisão} + \text{sensibilidade})$ [30].

2.6 Passos no Desenvolvimento da Ferramenta de Análise Integrada

2.6.1 Análise de Sequências

Filtramos as sequências de SH e MT por qualidade ($Q \geq 20$) com o Sickle (<https://github.com/najoshi/sickle>) e as classificamos com as ferramentas da seção 2.3. Nós analisamos as sequências de 16S com o QIIME [11] usando os *scripts split_libraries_fastq.py* para filtrar as sequências pelo mesmo critério de qualidade e *pick_open_reference_otus.py* com o uso do método UCLUST para a obtenção das OTUs.

2.6.2 Uniformização das Classificações Taxonômicas

Verificamos os nomes das classificações taxonômicas dos classificadores de 16S pela comparação com a taxonomia do NCBI [16]. Um nome é válido se há o mesmo nome científico para o seu nível taxonômico na taxonomia do NCBI. Se o nome for um sinônimo de um nome científico também será válido, e o nome científico deverá ser adotado para uniformizar as classificações.

Por exemplo, o nome *Corynebacterium variabile* é válido pois é um nome científico para uma espécie na taxonomia do NCBI. *Arthrobacter variabilis*, *Caseobacter polymorphus*, *Corynebacterium mooreparkense* e *Corynebacterium variabilis*, são sinônimos de *Corynebacterium variabile* e são válidos também. Todas as classificações para esses sinônimos são uniformizadas com o nome científico *Corynebacterium variabile*.

2.6.3 Extração e Seleção de Variáveis

Extraímos um conjunto de variáveis a partir dos resultados de classificação e perfis taxonômicos obtidos com as ferramentas, as mais comuns foram:

Percentual de Sequências: A abundância relativa foi demonstrada como um fator para identificação de predições incorretas em trabalhos anteriores [47] – táxons mais abundantes tendem a ter uma maior chance de serem predições corretas. Algumas ferramentas, fornecem apenas a abundância relativa como informação sobre suas predições, para essas nós modelamos as variáveis: (i) percentual interno da espécie em relação a outras espécies do mesmo gênero; e (ii) proporção do número de espécies por gênero. Em alguns casos, extraímos também tanto o percentual da

espécie em razão das sequências classificadas com associações taxonômicas como do total de sequências da amostra (incluindo as sequências não classificadas).

Pontuação da Classificação Taxonômica: Analisamos o conjunto das classificações de uma mesma espécie. Exploramos a distribuição das pontuações das classificações das ferramentas para compreender se os grupos de táxons corretamente e incorretamente classificados podem ser discriminados a partir da distribuição desses valores. Usamos diversas medidas estatísticas para resumir essa distribuição: média, desvio padrão, intervalo de confiança, primeiro quartil, mediana, terceiro quartil, desvio absoluto mediano e IQR (*variação interquartil*).

Número de OTUs e Proporção de OTUs: Partimos de uma premissa que quanto maior o número de OTUs classificadas como um determinado táxon, maior a chance de que esse táxon esteja realmente presente. Usamos as informações de OTUs originadas do 16S com a sumarização das informações agrupadas por táxon com as mesmas medidas estatísticas usadas para pontuação, baseadas nos tamanhos das OTUs e na proporção que essas OTUs representam dentro do conjunto geral.

Presença ou Ausência em Diferentes Meta-ômicas: Incluímos a informação da presença ou ausência em cada meta-ômica como um fator favorável para a veracidade da informação, avaliando o princípio de que quando linhas diferentes de análise chegam a uma mesma conclusão ela é fortalecida [18].

Extraímos também algumas medidas específicas de alguns classificadores. A lista completa pode ser conferida em <https://github.com/felipepratalima/gunga>. Seleccionamos as variáveis informativas através de três métodos de filtros de variáveis para remover variáveis não informativas e correlacionadas entre si: CFS (*Correlation-based Feature Selection*), *Information Gain* e *Gain Ratio* [22]. As variáveis remanescentes desses filtros foram removidas em caso de variância zero ou correlação com demais variáveis.

2.6.4 Modelos de Classificação

Usamos combinações de meta-ômicas (SH, MT, 16S) para avaliar os resultados das integrações. Para cada ferramenta, avaliamos diferentes algoritmos de aprendizado de máquina supervisionado para classificação com as variáveis selecionadas para aprender a identificar as predições corretas e

incorretas. Seleccionamos algoritmos que utilizam diferentes paradigmas de aprendizado e representam o conhecimento aprendido de diferentes formas:

Logistic Regression (LR), Linear Discriminant Analysis (LDA) e Quadratic DA (QDA):

Modelam a probabilidade de um objeto pertencer a uma classe por meio de uma função discriminante que envolve a multiplicação de cada valor para cada variável do modelo por um coeficiente estimado para representar seu peso na predição.

Naive Bayes (NB): Modela a probabilidade de um objeto pertencer a uma classe tomando como premissa que todas as suas variáveis são independentes. Com isso o modelo aprende as probabilidades dos valores ocorrerem para cada variável, das classes ocorrerem e desses valores ocorrerem dado essas classes. E então a partir do teorema de Bayes resolve a probabilidade de um objeto pertencer a uma classe dado o conjunto de valores.

Neural Networks (NN): Usam um modelo chamado conexionista, que representa o conhecimento a partir de um conjunto de nós (neurônios) que se conectam entre si, e para cada conexão então é atribuído um peso. Um dos modelos de redes neurais mais utilizados é o MLP com três camadas conectadas. Nele uma camada com neurônios de entrada representa as variáveis de entrada, uma camada intermediária faz um conjunto de transformações dos valores de entrada para a camada de saída que representa as variáveis de saída. O modelo aprende configurando os valores de pesos que satisfazem as relações entre seus neurônios de entrada e de saída com os exemplos de treinamento, em geral a partir de um método chamado *backpropagation*.

Support Vector Machines (SVM): Representam as classes dividindo o espaço entre os pontos através de um hiperplano, onde cada conjunto de pontos separados por esse hiperplano pertence a uma classe.

Classification Trees (CART) e Random Forests (RF): Os algoritmos de árvores de classificação representam o aprendizado como uma árvore de decisão com um conjunto de nós que devem ser percorridos de um nó inicial até um nó final que representa a classe do objeto, tomando decisões de por quais nós percorrer a partir dos valores do objeto classificado. O Random Forest usa o modelo de árvores com reamostragem (*bagging*) para desenvolver diversas árvores de classificação e combiná-las para obter a classificação final. [25, 29].

CAPÍTULO 3

SIMULAÇÃO DE DADOS META-ÔMICOS

3.1 *Pipeline*

Nós desenvolvemos um *pipeline* para simular o sequenciamento de amostras com as três meta-ômicas. Nele, modelamos como entrada a composição microbiana da amostra e um conjunto de *primers* para o 16S. A composição incluiu a lista de táxons (em níveis de cepa, espécie e/ou gênero) e suas abundâncias relativas. A partir dessas informações, automatizamos a recuperação e geração dos dados de entrada necessários para as simulações de cada meta-ômica com o simulador de sequências.

Na Figura 3.1 apresentamos o fluxograma com as etapas que executamos para as simulações, que podem ser descritas da seguinte forma:

Em (1): selecionamos um perfil de uma comunidade apresentada em uma publicação, e geralmente extraímos sua composição com o processamento de arquivos disponibilizados pelos próprios autores.

Em (2), (3) e (4): obtivemos a lista de procariotos que possuíam genoma completo no NCBI. Para isso obtivemos do site do NCBI a tabela *assembly_summary_refseq.txt* que contém informações sobre os genomas montados disponíveis no RefSeq (2) e selecionamos as informações sobre procariotos dessa tabela (4). Usamos a tabela *prokaryotes.txt* para facilitar a identificação dos procariotos para a seleção (3), também obtida no site do NCBI.

Em (5): verificamos para cada organismo da comunidade se havia genoma completo disponível na tabela resultante da Etapa 4. Para as comunidades descritas em nível de cepa, eliminamos diretamente as cepas que não possuíam. Para níveis taxonômicos de espécie e gênero, eliminamos se não houve ao menos 1 genoma disponível para o nível taxonômico descrito.

Em (6), (7) e (8): verificamos se a comunidade estava descrita em nível de cepa. Se sim, usamos a tabela resultante da Etapa 4 para obter diretamente as informações para recuperação dos genomas e CDSs no NCBI, referentes aos organismos da comunidade (7). Se a composição estava descrita no nível de espécie ou gênero, sorteamos um genoma da tabela para cada organismo (8).

Em (9): corrigimos as abundâncias relativas quando houve descarte de organismos do perfil inicial na Etapa 5, e obtivemos o perfil final para simulação.

Em (10) e (11): obtivemos os arquivos FASTA dos genomas e CDSs de cada organismo. Usamos a biblioteca **rentrez** e o *Assembly Name* para recuperar informações auxiliares para baixá-los do

site do NCBI. A primeira opção foram os arquivos do genoma no subdiretório GCA (*Assembly*), baixamos os arquivos com final “_genomic.fna.gz”, “_cds_from_genomic.fna.gz”. O arquivo de CDSs ocasionalmente não estava disponível no subdiretório GCA, então a segunda opção foi obtê-lo do subdiretório GCF (*RefSeq*). Tanto para os arquivos FASTA de genomas quanto para os de CDSs, para cada organismo, renomeamos todas as sequências para um mesmo nome – o que foi necessário para que o Grinder reconhecesse sequências diferentes como sendo do mesmo organismo.

Em (12): criamos os arquivos de referência que foram usados para simulação de sequências no Grinder: um arquivo FASTA com todos os genomas de todos os organismos (“genomes.fna”), e um arquivo FASTA com todas as CDSs (“cdss.fna”). Além disso, criamos o arquivo que descreveu a comunidade, uma tabela com os identificadores de cada organismo nos arquivos FASTA e suas abundâncias relativas. Esses arquivos foram usados para a simulação dos dados de SH e MT. O arquivo FASTA com genomas também foi usado para a simulação dos dados de 16S.

Em (13), (14) e (15): obtivemos o perfil para os dados de 16S. Avaliamos a capacidade do par de *primers* 16S escolhido extrair os *amplicons* com o método **Bio::Tools::AmpliconSearch** do pacote **Bioperl** (13) e em seguida obtivemos o perfil de 16S com a eliminação dos organismos que não possuíam sequências extraídas e com a correção dos valores de abundância (14). Em seguida, criamos o arquivo específico que descreveu a comunidade para a simulação dos dados de 16S com o Grinder (15).

Em (16): usamos os arquivos produzidos nas etapas anteriores para simular os dados com o Grinder.

Adaptamos o *pipeline* durante o trabalho para também:

1. Manter organismos com genomas classificados como incompletos – dispensando a verificação ocorrida na Etapa 5.
2. Obter as informações sobre genomas e CDSs disponíveis no NCBI através do uso da biblioteca **rentrez** – modificando as Etapas 7 e 8.
3. Predizer CDSs com o Prodigal. Somente para organismos para os quais não encontramos os arquivos FASTA com suas CDSs, nomeados com final “_cds_from_genomic.fna.gz” no NCBI – modificando a Etapa 10.

Figura 3.1: Pipeline para simulação de dados meta-ômicos (SH, MT e 16S) “casados”.

3.2 Resultados da Simulação

Simulamos dados de sequenciamento com as três meta-ômicas (SH, MT e 16S) a partir das composições de organismos e abundâncias relativas das comunidades extraídas dos trabalhos descritos na seção 2.2 e obtivemos 56 amostras descritas na Tabela 3.1.

Tabela 3.1: Características das amostras simuladas.

Origem	Amostra	# Genomas	# Filos	# Gêneros	# Espécies
cam	CamiToyDataHC1	449	21	207	375
cam	CamiToyDataHC2	409	21	195	344
cam	CamiToyDataHC3	396	21	192	333
cam	CamiToyDataHC4	388	21	193	330
cam	CamiToyDataHC5	388	21	193	330
cam	CamiToyDataLC	25	9	12	25
cam	CamiToyDataMC1	225	19	115	200
cam	CamiToyDataMC2	225	19	115	200
emp	AerosolNonSaline	164	19	161	164
emp	AnimalCorpus	123	15	121	123
emp	AnimalDistalGut	149	17	145	148
emp	AnimalProximalGut	142	15	139	142
emp	AnimalSecretion	167	18	163	166
emp	AnimalSurface	182	18	178	181
emp	HypersalineSaline	36	10	36	36
emp	PlantCorpus	142	15	139	142
emp	PlantRhizosphere	133	17	130	133
emp	PlantSurface	109	12	107	109
emp	SedimentNonSaline	152	18	149	151
emp	SedimentSaline	180	18	177	180
emp	SoilNonSaline	181	17	177	181
emp	SurfaceNonSaline	184	18	180	183
emp	SurfaceSaline	151	18	148	151
emp	WaterNonSaline	186	18	182	185
emp	WaterSaline	161	18	158	161
jovel	Mix1	12	4	7	11
jovel	Mix2	12	4	7	11
jovel	Mix3	12	4	7	11
mavromatis	SimHC	111	9	80	105
mavromatis	SimLC	112	9	81	106

mavromatis	SimMC	112	9	81	106
mbarc	MBARC	26	10	24	26
mcintyre	Ds7	551	25	334	516
mcintyre	DsBuccal	14	2	4	12
mcintyre	DsCityParks	52	5	44	48
mcintyre	DsGut	25	5	17	20
mcintyre	DsHous1	32	2	10	29
mcintyre	DsHous2	20	3	5	20
mcintyre	DsNYCSM	22	2	13	20
mcintyre	DsSoil	52	5	44	48
mcintyre_huttenhower	HC1	99	14	78	99
mcintyre_huttenhower	HC2	98	15	72	98
mcintyre_huttenhower	LC1	25	8	23	25
mcintyre_huttenhower	LC2	25	11	25	25
mcintyre_huttenhower	LC3	25	8	21	25
mcintyre_huttenhower	LC4	25	9	25	25
mcintyre_huttenhower	LC5	24	8	22	24
mcintyre_huttenhower	LC6	25	11	24	25
mcintyre_huttenhower	LC7	25	10	25	25
mcintyre_huttenhower	LC8	25	5	23	25
mande	100SpeciesEven	100	14	66	85
mande	100SpeciesStaggered	100	14	66	85
mande	10SpeciesEven	10	4	10	10
mande	10SpeciesStaggered	10	4	10	10
mande	400SpeciesEven	400	22	243	398
mande	400SpeciesStaggered	400	22	243	398

3.3 Comparação das Amostras

Comparamos os perfis usados nas simulações para avaliar a diversidade das amostras selecionadas. Fizemos um agrupamento hierárquico para os níveis de espécie e gênero usando a medida de Bray-Curtis para o cálculo das distâncias. Não encontramos nenhum padrão significativo nos agrupamentos, aparentemente eles se deram em maior parte em razão das origens dos perfis (publicação) e seus números de táxons. Pela visualização da matriz com um *heatmap*, notamos que predominantemente houve uma maior dissimilaridade entre os perfis, ainda que esses perfis tivessem mesma origem. (Figura 3.2).

Figura 3.2: Agrupamento dos perfis taxonômicos selecionados para simulação.

Uma análise com *Non-Metric Multidimensional Scaling* (NMDS) com duas dimensões apresentou novamente a ausência de formação de grupos precisos de um modo geral e a origem como principal fator de aproximação dos perfis. As comparações entre grupos baseados na origem com o método *Permutational Multivariate Analysis of Variance Using Distance Matrices* (ADONIS) confirmaram essas diferenças como significativas em ambos os níveis ($R^2 = 0,31649$ e $P = 0,001$ para gênero; $R^2 = 0,30165$ e $P = 0,001$ para espécie) (Figura 3.3).

Figura 3.3: Comparação dos perfis taxonômicos das amostras simuladas com NMDS.

3.3.1 Amostra de Treinamento, Teste e Validação da Ferramenta

Observando essas dissimilaridades entre as amostras, selecionamos o conjunto de amostras (Mix2, LC6, MBARC, DsCityParks, HC1, PlantRhizosphere, SedimentNonSaline, WaterSaline, CamiToyDataMC2 e Ds7) para avaliar os modelos de classificação desenvolvidos com o método *hold-out*. Com o algoritmo de melhor desempenho construímos e otimizamos o modelo final usando o método de validação cruzada. As amostras Mix3 e CamiToyDataHC[1-5] foram usadas apenas para a validação da ferramenta com o modelo final.

CAPÍTULO 4

ANÁLISE DO DESEMPENHO DOS CLASSIFICADORES

A seguir apresentamos as análises dos resultados das classificações taxonômicas dos dados meta-ômicos *mock* e simulados com as ferramentas e com as meta-ômicas individuais e a abordagem integrada. Fizemos uma avaliação baseada nas medidas de desempenho de classificadores (precisão, sensibilidade e F1), estabelecendo critérios objetivos para compará-los de acordo com cada tipo de meta-ômica e ferramenta. Os resultados com meta-ômicas individuais são importantes como referência para compará-las entre si e com a abordagem integrada, desse modo verificando se de fato há melhoria da classificação taxonômica por meio da abordagem integrada.

Iniciamos o capítulo estabelecendo as meta-ômicas individuais como referência, incluindo uma avaliação sobre o uso das abordagens de combinação de classificadores taxonômicos (*ensembles*) como uma alternativa para a obtenção de melhorias em SH. Dada a existência de semelhanças entre SH e MT, avaliamos o uso das mesmas ferramentas e abordagens de SH em MT.

Em seguida, apresentamos os resultados da aplicação da nossa metodologia no desenvolvimento da nossa abordagem para melhoria da classificação taxonômica com a integração das meta-ômicas. Medimos o desempenho de diferentes modelos de classificação para eliminar predições incorretas e prover um perfil taxonômico mais robusto, com diferentes ferramentas e algoritmos. Concluímos mostrando como fizemos a seleção de um algoritmo para a construção dos modelos a serem incorporados em uma ferramenta para a integração das meta-ômicas.

4.1 Desempenhos dos Classificadores Taxonômicos em Meta-ômicas SH e MT

Os classificadores taxonômicos em SH tiveram desempenho variável. Os resultados mostram bem o balanço entre precisão e sensibilidade, ou seja, quando se ganha em precisão se perde em sensibilidade, e vice-versa. Os métodos baseados em marcadores foram os mais precisos, porém menos sensíveis, enquanto que os métodos composicionais e baseados em alinhamentos foram mais sensíveis, porém menos precisos. Para todos os classificadores, quanto menor foi a profundidade taxonômica avaliada melhor foi o desempenho em todas as medidas (Figura 4.1).

O MetaPhlAn2 foi o que teve a maior precisão em espécie dentre todos os classificadores, porém, o GOTCHA também apresentou precisão próxima e maior sensibilidade, o que lhe conferiu um valor melhor na medida F1. O TIPP destacou-se negativamente com baixa precisão para um classificador de marcadores. Dentre os métodos de alinhamentos, o DUDes teve o comportamento mais distante, com

Figura 4.1: Distribuição das medidas de desempenho obtidas com os classificadores taxonômicos em SH.

precisão e sensibilidade mais próxima dos de marcadores. O Kraken2 (com nr) foi o composicional mais sensível, porém mais impreciso também, e o LMAT apesar da menor sensibilidade foi o mais preciso.

O desempenho dos classificadores foi similar em SH e MT. Para cada um deles os valores de precisão e sensibilidade obtidos para as duas meta-ômicas nas amostras correspondentes foram próximos para ambas medidas. Esperávamos encontrar maiores diferenças para essas medidas de desempenho entre essas meta-ômicas, uma vez que SH contém sequências de quaisquer trechos dos genomas dos organismos sequenciados e MT contém apenas sequências codificadoras. Porém, verificamos que a meta-ômica representa menor influência nos resultados de análise taxonômica do que os classificadores utilizados (Figura 4.2).

Fizemos um agrupamento hierárquico usando a precisão e a sensibilidade para obtenção das distâncias entre os classificadores e obtivemos que classificadores de mesmas categorias tiveram desempenhos mais similares entre si. Observamos a ocorrência de dois grupos principais, formados majoritariamente (1) por classificadores baseados em marcadores e (2) por classificadores baseados em alinhamentos e composicionais. Dentro do grupo (2) ainda percebemos a divisão entre dois sub-grupos, (2a) com menor precisão e maior sensibilidade e outro (2b) ao contrário. As exceções nos grupos foram os

Figura 4.2: Dispersão da precisão e da sensibilidade obtidas com os classificadores em SH e MT para espécie, com destaque para o desempenho com os diferentes grupos de amostras.

classificadores TIPP (marcadores), que teve precisão tão baixa quanto os composicionais, e o DUDes (alinhamentos), com precisão e sensibilidade similar aos de marcadores. Os pares SH e MT para mesmo classificador representaram as menores distâncias na maioria dos casos, sendo os desempenhos de diferentes meta-ômicas com mesmo classificador mais similares do que os desempenhos de mesma meta-ômica com diferentes classificadores (Figura 4.3).

Com a análise das correlações com a nossa medida geral de desempenho F1, encontramos correla-

Figura 4.3: Agrupamentos dos classificadores com precisão e sensibilidade em todos os níveis taxonômicos.

ções positivas significativas ($R \geq 0,7$; $P < 0,05$) das duas meta-ômicas em todos os níveis taxonômicos (Figura 4.4), e apenas o CLARK em filo apresentou uma coeficiente menor do 0,7.

4.2 Desempenho dos Classificadores Taxonômicos em 16S

Os desempenhos dos classificadores RDP-classifier e Uclust com as sequências representativas das OTUs obtidas com o método *open-reference OTUs* em 16S foram similares para a maior parte das amostras em filo e gênero, mas variaram em espécie, onde o RDP foi mais sensível e o Uclust mais preciso. O SPINGO se diferenciou principalmente em razão da maior precisão em todo os níveis, obtendo o máximo para a medida (igual a 1) em mais da metade das amostras em filo e gênero. Além disso, sua sensibilidade também foi maior em espécie, o que lhe conferiu melhor F1 dentre todos (Figura 4.5).

A precisão com sequências representativas em 16S foi mediana em comparação com os classificadores em SH e MT, maior em relação aos classificadores composicionais e de alinhamentos e menor

Figura 4.4: Correlações de F1 obtidas com os classificadores taxonômicos em SH e MT.

em relação aos de marcadores. Porém, a sensibilidade foi predominantemente mais baixa. Já esperávamos essas diferenças, tanto em razão da dificuldade de classificação de uma região restrita do 16S em níveis taxonômicos mais profundos, quanto pela limitação imposta pelo par de *primers* utilizado, que necessariamente implica na possibilidade de detecção apenas parcial dos organismos presentes em uma amostra. No entanto, esperávamos que essas diferenças ocorressem principalmente e até de forma mais acentuada em gênero e espécie, mas não em filo, onde foram muito acentuadas com uma mediana

Figura 4.5: Distribuição das medidas de desempenho obtidas com os classificadores taxonômicos em 16S. O sufixo RepSet se refere ao uso do método de obtenção de OTUs e Slout se refere ao uso do conjunto completo de sequências. Primers é um classificador hipotético, com precisão perfeita (igual a 1) e máximo de sensibilidade para a amostra considerando o par de *primers* utilizado. Dentre todos, o Spingo (tanto RepSet quanto Slout) é o que constantemente mais se aproxima do Primers.

de detecção de apenas 50% dos filos em 16S contra a mediana de sensibilidade perfeita (igual a 1) da maioria dos classificadores em SH e MT.

Obtivemos melhoria na sensibilidade com a classificação direta dos conjuntos completos de sequências de 16S em todos os níveis taxonômicos – com uma detecção média aproximada de 65% dos táxons dentre aqueles que foram selecionados pelo par de *primers* no nível de espécie, e mais de 90% em gênero e filo.

A menor sensibilidade com a abordagem de OTUs provavelmente se deu muito mais pela redução do conjunto a sequências representativas que não capturam toda a diversidade dos táxons presentes¹ do que pelas capacidades de seleção dos *primers* e de classificação das ferramentas (Figura 4.6). Nos chamou atenção que as diferenças entre a quantidade de FNs para os classificadores com e sem o método de OTUs foi atenuada em espécie. A perda de precisão somada ao menor número de classificações que chegam a esse nível podem ser os principais fatores para essa atenuação.

Por outro lado, ressaltamos que embora com a classificação sem a abordagem de OTUs ocorra

¹O que justifica a introdução do conceito de *amplicon sequence variants* adotado na versão 2 do pacote QIIME.

Figura 4.6: Números de FNs obtidos com cada ferramenta por nível taxonômico, com e sem o uso da abordagem de OTUs.

essa perda na precisão com a introdução de muitos FPs, principalmente em espécie, ela foi menos significativa e obtivemos melhores valores para F1 em todos os níveis taxonômicos em comparação com a abordagem de OTUs (Figura 4.7).

Figura 4.7: Agrupamento dos classificadores de 16S com a precisão e a sensibilidade. O uso ou não da abordagem de OTUs é o principal fator que distingue o desempenho. O SPINGO é o que possui o desempenho mais próximo de um classificador hipotético 100% preciso (Primers).

4.3 Comparação das Abordagens de Combinação de Classificadores Taxonômicos (*Ensembles*)

O MetaMeta e o Majority Vote tiveram desempenhos mais similares entre si, e ambos mais dissimilares com o WEVOTE. Os dois primeiros se destacaram por obterem maiores valores em precisão e F1. O terceiro se destacou por ter superado os demais em sensibilidade obtendo o maior valor possível

(igual a 1) para mais da metade das amostras em todos os níveis taxonômicos (Figura 4.8). Testes de Wilcoxon em SH mostraram que não há diferença significativa entre os desempenhos do MetaMeta e do Majority Vote em espécie ($P = 0,36$ para precisão; $P = 0,05$ para sensibilidade; e $P = 0,06$ para F1), enquanto que para o WEVOTE os testes mostraram que há diferenças significativas com ambos (com $P < 0,001$ para todas as comparações).

Figura 4.8: Distribuição das medidas de desempenho obtidas com as abordagens de combinação em SH e MT.

Como com os classificadores individuais, as abordagens tiveram medidas de desempenho similares para as mesmas amostras em SH e MT, demonstrando por comparação a validade do uso dessas abordagens também para MT. Os valores das medidas de desempenho aparentemente sofreram mais influência da composição taxonômica da amostra e da abordagem do que da meta-ômica (Figura 4.9).

E novamente os desempenhos obtidos em SH e MT foram similares em cada amostra para todos os níveis taxonômicos quando consideramos o balanceamento entre precisão e sensibilidade. A análise da correlação com o método de Spearman com as medidas de F1 quando usada a mesma abordagem em ambas meta-ômicas mostrou correlações positivas significativas ($R \geq 0,7$; $P < 0,05$) em todos os níveis taxonômicos (Figura 4.10).

Figura 4.9: Distribuição da precisão e da sensibilidade obtidas com as abordagens de combinação em SH e MT para espécie, com destaque para o desempenho com os diferentes grupos de amostras.

Figura 4.10: Correlações de valores de F1 obtidos com as abordagens de combinação em SH e MT.

4.4 Processo de Escolha dos Modelos de Classificação e Resultados

Para desenvolver e selecionar o modelo final de classificação do Gunga, fizemos uma análise preliminar de modelos treinados com 8 diferentes algoritmos. A seguir apresentamos os resultados das etapas para a escolha das variáveis e avaliação desses modelos.

4.4.1 Extração das Variáveis e Junção de Meta-ômicas

Extraímos as variáveis das classificações e perfis taxonômicos em espécie e selecionamos aquelas com maior potencial para a construção dos classificadores usando a abordagem de seleção por filtro com as amostras de treinamento. Para cada ferramenta, tivemos um número de variáveis extraídas de acordo com suas saídas, e um subconjunto de variáveis selecionadas de acordo com a meta-ômica. Conseguimos extrair entre 5 e 133 de acordo com a ferramenta e selecionar entre 1 e 36 variáveis de acordo com a ferramenta e a meta-ômica. Ferramentas que fornecem apenas o perfil taxonômico tiveram menor número de variáveis extraídas, uma vez que a abundância relativa foi frequentemente a única informação sobre suas predições taxonômicas.

Comparamos as distribuições de todas variáveis entre grupos de táxons corretos (VPs) e incorretos (FPs) para analisar seus potenciais como preditoras e para avaliar o resultado da nossa seleção em relação a um número máximo de variáveis que poderiam ser selecionadas. Fizemos um teste de Wilcoxon para cada uma delas e as classificamos como potenciais preditoras quando a diferença entre os grupos foi significativa com $P < 0,05$. Notamos que o número de variáveis selecionadas foi até 6 vezes menor do que o número de significativas, dependendo da ferramenta, o que provavelmente se deu em razão das técnicas de seleção eliminarem variáveis redundantes.

Obtivemos o **perfil juntado** para cada ferramenta de classificação em SH, com a junção das predições em MT para a mesma ferramenta e em 16S com o Spingo e o QIIME com RDP. Apresentamos um resumo dos resultados do processo de seleção e junção na Tabela 4.1.

4.4.2 Comparação dos Algoritmos de Classificação

Treinamos um modelo de classificação para cada par (*algoritmo, ferramenta*) e comparamos seus desempenhos através das amostras de teste. Calculamos as medidas de desempenho desconsiderando os táxons das amostras que não foram preditos, afim de anular o efeito das diferentes taxas de sensibilidade das ferramentas e compararmos o desempenho médio entre os algoritmos em si. Eliminamos o MEGAN no desenvolvimento dos modelos em razão da seu desempenho que foi próximo ao obtido com o DIAMOND.

4.4.2.1 Precisão

Para os modelos com o mesmo algoritmo obtivemos uma precisão mediana acima de 0,9, com exceção do LDA, NB e QDA. Os modelos com RF foram os mais precisos com mediana 0,96 e mais

Tabela 4.1: Ferramentas e números de variáveis obtidas a partir da extração de dados. As ferramentas marcadas com **Sim** no campo **Classificação** foram as que forneceram saídas com as classificações individuais de cada sequência que foram utilizadas para a extração de variáveis. Para aquelas marcadas com **não** extraímos variáveis apenas do perfil taxonômico resultante.

Programa	Classificação	Meta-ômica	# Extração	# Signif.	# Seleção	# Junção
Centrifuge	Sim	SH / MT	99	59 / 63	15 / 10	56
CLARK	Sim	SH / MT	24	24 / 23	11 / 11	53
CLARK-S	Sim	SH / MT	96	75 / 93	36 / 28	95
DIAMOND	Sim	SH / MT	133	113 / 115	29 / 41	101
DUDes	Não	SH / MT	5	4 / 4	4 / 4	39
GOTTCHA	Não	SH / MT	9	8 / 7	2 / 5	38
Kaiju	Sim	SH / MT	21	21 / 21	7 / 6	44
Kraken	Sim	SH / MT	38	36 / 38	16 / 14	61
Kraken2 (Nr)	Sim	SH / MT	38	36 / 38	6 / 5	42
Kraken2 (Nt)	Sim	SH / MT	38	36 / 38	10 / 10	51
Kraken2 (Standard)	Sim	SH / MT	38	36 / 36	11 / 8	50
LMAT	Não	SH / MT	13	11 / 11	6 / 4	41
Metaphlan2	Não	SH / MT	5	4 / 4	4 / 1	36
mOTUs	Não	SH / MT	5	5 / 5	4 / 1	36
TIPP	Não	SH / MT	5	4 / 5	1 / 1	33
SPINGO	Sim	16S	81	68	11	-
QIIME com RDP (E)	Sim	16S	62	16	6	-
QIIME com RDP (G)	Sim	16S	62	16	9	-

estáveis quanto a obtenção de valores próximos da sua mediana com IQR 0,09. Modelos com NB, NN e QDA foram os mais instáveis com IQRs 0,3, 0,21 e 0,23 respectivamente (Figura 4.11).

Figura 4.11: Precisão dos modelos de classificação por algoritmo de classificação com as amostras de teste.

Para os modelos com a mesma ferramenta também obtivemos uma mediana maior que 0,9 para a maioria deles. Os modelos com as ferramentas baseadas em marcadores tiveram melhor precisão

e os com métodos composicionais, pior precisão. Os modelos com Kraken2 se destacaram tanto negativamente quanto positivamente, com a pior precisão para os modelos com banco de dados *nr* e a melhor para os modelos com *standard*, com medianas 0,78 e 1 e IQRs 0,39 e 0,06 respectivamente (Figura 4.12).

Figura 4.12: Precisão dos modelos de classificação por ferramenta com as amostras de teste.

4.4.2.2 Sensibilidade e Especificidade

Uma medida de interesse que avaliamos em conjunto com a sensibilidade foi a especificidade, que resume a proporção de acertos em todas as classificações de táxons como incorretos. Ela é definida por $VN/(VN + FN)$, onde VN são os táxons preditos pelas ferramentas que não fazem parte da amostra e que foram classificados como incorretos pelos modelos.

Do ponto de vista dos algoritmos, a sensibilidade mediana dos modelos foi acima de 0,9 para quase todos eles; apenas os modelos com LDA e LR tiveram uma mediana menor, 0,88 e 0,89 respectivamente. Os modelos com RF foram novamente os que tiveram melhor desempenho com mediana 0,92 e menor IQR 0,11. A especificidade mediana foi maior que 0,97 para os modelos com a maioria dos algoritmos, sendo 1 para os modelos com RF (Figura 4.13). Portanto, os algoritmos foram eficazes para a criação de modelos que identificam os táxons incorretos, porém com o custo de uma perda mediana de até 10% dos táxons corretos preditos pelas ferramentas.

Para maior parte das ferramentas seus conjuntos de modelos também tiveram uma sensibilidade mediana menor do que a especificidade. Os modelos com o Kraken2 e bancos Standard e nt tiveram um desempenho destacado, com medianas de sensibilidade 0,97 e 0,98 respectivamente e de especificidade 1 para ambos (Figura 4.14).

Os modelos para as ferramentas de marcadores e o DUDes (que individualmente teve precisão e

Figura 4.13: Sensibilidade e especificidade dos modelos de classificação por algoritmo de classificação com as amostras de teste.

sensibilidade mais próximas das ferramentas baseadas em marcadores) foram menos específicos que sensíveis. De forma geral, as ferramentas de marcadores produzem perfis mais precisos e menos sensíveis, gerando mais exemplos de predições corretas do que incorretas. Isso diminui a base de treinamento para a identificação de VNs em comparação com as ferramentas composicionais, que produzem mais exemplos de predições incorretas.

Figura 4.14: Sensibilidade e especificidade dos modelos de classificação por ferramenta com as amostras de teste.

4.4.2.3 F1

Resumimos o desempenho geral dos modelos com a medida F1 na Tabela 4.2. Considerando os algoritmos, os modelos com RF tiveram o melhor desempenho, com medianas superiores ou iguais para quase todas as ferramentas. A exceção se deu com o DUDes, que obteve desempenho melhor para a medida com o SVM, com maior mediana apesar da mesmo IQR. Os modelos com NB tiveram o pior desempenho de forma geral, sendo o desempenho com o DIAMOND o pior. Considerando as ferramentas, os modelos com o Kraken2 e banco de dados *standard* tiveram o melhor desempenho, com medianas maiores que 0,95 e IQRs menores que 0,05. Já os modelos com o TIPP tiveram o pior desempenho de forma geral.

Tabela 4.2: Medianas de F1, com IQRs entre parênteses, dos modelos de classificação para cada par (*algoritmo*, *ferramenta*).

	Cart	Lda	Lr	Nb	Nn	Qda	Rf	Svm (Linear)
Centrifuge	0,93 (0,11)	0,9 (0,11)	0,95 (0,07)	0,87 (0,15)	0,68 (0,38)	0,87 (0,1)	0,95 (0,06)	0,94 (0,06)
Clark	0,86 (0,11)	0,75 (0,13)	0,87 (0,11)	0,75 (0,18)	0,86 (0,1)	0,75 (0,2)	0,89 (0,09)	0,86 (0,09)
ClarkS	0,91 (0,05)	0,89 (0,1)	0,94 (0,08)	0,84 (0,09)	0,92 (0,07)	0,8 (0,13)	0,95 (0,06)	0,93 (0,08)
Diamond	0,9 (0,09)	0,65 (0,09)	0,64 (0,21)	0,48 (0,3)	0,83 (0,09)	0,78 (0,22)	0,91 (0,1)	0,82 (0,15)
Dudes	0,94 (0,06)	0,92 (0,08)	0,93 (0,06)	0,84 (0,19)	0,94 (0,07)	0,86 (0,15)	0,93 (0,06)	0,94 (0,06)
Gottcha	0,94 (0,07)	0,92 (0,15)	0,94 (0,08)	0,87 (0,19)	0,94 (0,1)	0,89 (0,17)	0,95 (0,09)	0,93 (0,09)
Kaiju	0,84 (0,17)	0,76 (0,11)	0,9 (0,1)	0,66 (0,34)	0,9 (0,09)	0,68 (0,17)	0,91 (0,09)	0,9 (0,09)
Kraken	0,89 (0,11)	0,89 (0,14)	0,92 (0,09)	0,85 (0,08)	0,93 (0,09)	0,88 (0,06)	0,95 (0,07)	0,93 (0,1)
Kraken2 Nr	0,84 (0,18)	0,82 (0,17)	0,87 (0,14)	0,58 (0,33)	0,82 (0,14)	0,6 (0,36)	0,91 (0,09)	0,88 (0,15)
Kraken2 Nt	0,97 (0,06)	0,98 (0,02)	0,98 (0,02)	0,91 (0,07)	0,98 (0,02)	0,91 (0,05)	0,99 (0,02)	0,98 (0,02)
Kraken2 Standard	0,97 (0,02)	0,98 (0,04)	0,98 (0,01)	0,95 (0,05)	0,98 (0,02)	0,96 (0,01)	0,99 (0,02)	0,99 (0,01)
Lmat	0,89 (0,09)	0,87 (0,07)	0,89 (0,05)	0,74 (0,16)	0,92 (0,07)	0,78 (0,15)	0,94 (0,06)	0,9 (0,08)
Metaphlan2	0,93 (0,06)	0,91 (0,05)	0,93 (0,05)	0,88 (0,23)	0,92 (0,07)	0,85 (0,17)	0,94 (0,05)	0,92 (0,06)
Motus	0,88 (0,09)	0,79 (0,26)	0,88 (0,08)	0,76 (0,09)	0,88 (0,09)	0,81 (0,13)	0,91 (0,09)	0,88 (0,08)
Tipp	0,82 (0,14)	0,76 (0,14)	0,76 (0,14)	0,7 (0,2)	0,81 (0,21)	0,68 (0,19)	0,83 (0,15)	0,78 (0,16)

Avaliamos o desempenho geral das abordagens SH e dos modelos de classificação (com perfis juntados) nas amostras de teste. Calculamos as medidas de desempenho considerando os táxons das amostras que não foram preditos afim de uniformizá-los para ambas as abordagens em relação às amostras. Comparamos as abordagens então observando as distribuições das medidas para as diferentes ferramentas e as ferramentas de combinação de classificadores (*ensembles*) (Figura 4.15).

Notamos que em razão da diversidade de desempenho dos modelos para as diferentes combinações de algoritmos e ferramentas, não foi possível observar uma melhoria expressiva na medida F1 para a abordagem de integração considerando qualquer das ferramentas, em comparação com o desempenho individual de cada uma sobre os dados SH. Em particular chegamos a dois padrões de desempenho: (i) para modelos com ferramentas composicionais e de alinhamentos houve um aumento visível de F1 através de um maior aumento na precisão e uma menor perda na sensibilidade; (ii) para modelos com ferramentas de marcadores a melhora foi inexpressiva, com ganhos de sensibilidade e perdas de precisão menores e quase proporcionais. Dessa forma, a abordagem pode ser considerada mais interessante para os modelos com ferramentas composicionais e de alinhamentos. Em comparação com as ferramentas de combinação de classificadores (*ensembles*) os modelos com AM foram consistentemente melhores do que o WEVOTE, porém para as outras duas ferramentas os modelos tiveram um desempenho melhor ou pior dependendo da combinação de algoritmo e ferramenta.

A ferramenta para qual a abordagem com AM ofereceu maior melhoria foi o Kraken2 com banco de

dados *standard*, onde os modelos tiveram uma pequena perda na sensibilidade e um ganho expressivo na precisão, conferindo aos modelos os melhores valores para F1 de forma constante com todos os algoritmos. Essa constância pode ser um indicador de que as variáveis usadas para a construção dos modelos do Kraken2 foram mais discriminativas dentre as predições taxonômicas corretas e incorretas da ferramenta.

Figura 4.15: Comparações de desempenho entre SH, ferramentas de *Ensembles* em SH e a abordagem integrada com o uso de modelos de AM.

4.4.3 Seleção e Otimização do Modelo de Classificação

Dado seu melhor F1 para a maioria das ferramentas em amostras de teste (Tabela 4.2), selecionamos o algoritmo *Random Forest (RF)* para otimização. Selecionamos os modelos com melhor desempenho para cada ferramenta através do método de validação cruzada (*leave-one-out cross-validation* – *LOOCV*) e diferentes parâmetros para o algoritmo, escolhidos automaticamente pelo pacote *caret* do R. Os modelos selecionados foram incorporados à ferramenta desenvolvida neste trabalho. —

A ferramenta que desenvolvemos com base nas ideias expostas no capítulo anterior chama-se *Gunga*. Neste capítulo apresentamos uma descrição da arquitetura da ferramenta, assim como resultados de validação e de aplicação do Gunga em dados reais.

5.1 Implementação

Implementamos um protótipo do programa Gunga como um pacote em R. O pacote inclui uma família de funções para lidar com o carregamento, o pré-processamento e a junção dos resultados de classificação e perfilamento taxonômico de diversos programas para meta-ômicas SH, MT e 16S, além da aplicação dos modelos de aprendizado de máquina desenvolvidos para discriminação dos táxons corretamente e incorretamente preditos por esses programas. Na Figura 5.1 apresentamos a modelagem de um fluxo genérico que projetamos para o processamento desses dados com a ferramenta.

Figura 5.1: Fluxo genérico de processamento de dados com o Gunga. A fase Pré Gunga corresponde à obtenção das classificações e perfis taxonômicos com um programa independente escolhido pelo usuário, e é executada externamente. As demais fases, do Pré-processamento até a Integração, correspondem aos processamentos realizados internamente pela ferramenta.

Esse fluxo sofre variações de acordo com o programa que é usado para análise taxonômica de cada meta-ômica individual, visto que alguns programas fornecem apenas as classificações ou os perfis taxonômicos como saída. Por exemplo, o programa Kraken produz saídas que incluem tanto as classificações individuais de cada sequência quanto perfis que incluem a sumarização das abundâncias

relativas de cada táxon predito. Por outro lado, as saídas do Diamond (com o uso do LCA do Megan para associação taxonômica) e do MetaPhlAn2 incluem, respectivamente, somente as classificações das sequências individuais e somente o perfil.

Nomeamos as funções do pacote de forma padronizada, de modo que os nomes escolhidos representem o conjunto de atividades descritas no fluxo genérico. Por exemplo, o padrão de nomenclatura *loadRaw[X]Classifications*, onde *[X]* designa o classificador taxonômico utilizado, representa a implementação para atividade **carregar classificações**. Para cada classificador então há uma função disponível, como *loadRawKrakenClassifications*, *loadRawCentrifugeClassifications*, e assim por diante. Descrevemos todas as principais funções do pacote na Tabela 5.1.

Tabela 5.1: Funções do pacote Gunga e sua nomenclatura.

Atividade	Nomenclatura	Descrição
Carregar classificações (Pré-processamento)	<i>loadRaw[X]Classifications</i>	Carregam os dados brutos da classificação taxonômica para um objeto do tipo Data Frame do R, preservando ao máximo a correspondência com o formato dos arquivos de saída originais de cada ferramenta.
Processar classificações (Pré-processamento)	<i>process[X]Classifications</i>	Processam os dados brutos da classificação com o objetivo principal de extrair e/ou aprimorar a informação sobre a classificação taxonômica de cada sequência.
Sumarizar classificações (Sumarizações)	<i>summarize[X]Classifications</i>	Sumarizam os dados das classificações, principalmente com medidas de estatísticas de tendência central e de dispersão para cada táxon presente.
Carregar perfil (Pré-processamento)	<i>loadRaw[X]Profile</i>	O mesmo que <i>loadRaw[X]Classifications</i> , porém para os dados do perfil.
Processar perfil (Pré-processamento)	<i>process[X]Profile</i>	Processam os dados brutos do perfil para extrair e/ou aprimorar a informação da classificação taxonômica para cada táxon predito.
Carregar OTU table (Pré-processamento)	<i>loadRawOtuTable</i>	Carregam os dados da OTU table em formato tabular para um objeto do tipo Data Frame, que é usado no processamento dos dados de classificações do 16S.
Juntar perfil e sumários (Junção)	<i>assign[X]ProfileAs[Y]</i>	Associa sua meta-ômica de origem aos dados processados. <i>[Y]</i> pode ser <i>Shotgun</i> , <i>Metatranscriptomics</i> ou <i>Amplicons</i> .
Juntar meta-ômicas (Integração)	<i>joinProfiles</i>	Junta os perfis das diferentes meta-ômicas.
Aplicar modelo de classificação (Integração)	<i>apply[X]ClassifierModel</i>	Aplicam os modelos de classificação para discriminação dos táxons corretos e incorretos nos dados dos perfis juntados, de acordo com classificador, associando seus resultados e produzindo assim o perfil integrado final.

Além das funções para execução das atividades individualmente, implementamos também um

grupo de funções que executam todo o fluxo a partir de uma única chamada. Elas seguem o padrão *gungaWorkflowFor[X]QiimeAndSpingo*. Para os exemplos do Kraken e Centrifuge, elas foram nomeadas *gungaWorkflowForKrakenQiimeAndSpingo* e *gungaWorkflowForCentrifugeQiimeAndSpingo*.

O código desenvolvido foi depositado no repositório *GitHub* e pode ser acessado através da URL <https://github.com/felipepratalima/gunga>. O Gunga pode ser instalado com o auxílio do pacote *devtools* através do comando `devtools::install_github("felipepratalima/gunga")`. O usuário deve interagir com o Gunga dentro da própria plataforma R. Com conhecimento da linguagem o usuário pode explorar os resultados e as variáveis do modelo e produzir desde saídas tabulares até gráficos que exploram as diferenças entre os grupos de táxons corretos e incorretos, como no exemplo da Figura 5.2.

Figura 5.2: Densidades de algumas variáveis do modelo para as classes preditas, com valores dos dados de exemplo.

Para execução do Gunga os dados meta-ômicos devem ter sido previamente classificados com um classificador de escolha de acordo com um conjunto de opções de modelos fornecidos pelo programa. Atualmente os classificadores disponíveis para escolha devem ser aqueles descritos nesta tese. Chamamos a atenção para o fato de que o classificador deve ser o mesmo para dados de SH e MT. Para

16S deve-se ter obrigatoriamente a classificação com SPINGO e QIIME com método de *open reference* OTUs e classificador RDP.

5.2 Validação do Gunga

Nesta seção apresentamos a validação do Gunga e da abordagem de integração com a sua aplicação em novas amostras de teste ainda não apresentadas aos modelos de classificação. Para mensurar a melhoria no perfil integrado obtido para as amostras comparamos seus resultados com os obtidos com o Centrifuge, Kraken2 e MetaPhlAn2, e também com as abordagens de combinação de classificadores WEVOTE, MetaMeta e Majority Vote em SH.

5.2.1 Seleção das Amostras

Escolhemos as amostras Mix3 e CamiToyDataHc[1-5] para validar a ferramenta, uma vez que elas se mostraram mais dissimilares às outras amostras dos seus grupos – Mix3 entre *mocks* e CAMI formando o grupo mais externo dentre todas as amostras simuladas, como mostramos no Capítulo 3. Além disso, consideramos as seguintes razões:

1. Mix3 é uma comunidade *mock* com composição conhecida, sequenciada com meta-ômicas SH e 16S com a plataforma Illumina. Ela representa sequências reais, com perfil de erros originados na preparação da amostra e no sequenciamento, e foi tomada por nós como o tipo de amostra que nos permitiria chegar mais próximo de uma avaliação com uma amostra real.
2. As amostras do CAMI foram desenvolvidas para representar um desafio aos classificadores taxonômicos, e as amostras HC (de *High Complexity*) foram tomadas por nós como o tipo de amostra que nos permitiria chegar mais próximo de uma avaliação com uma amostra com uma composição complexa.

5.2.2 Abordagem para Meta-ômicas SH e 16S

Desenvolvemos um modelo de classificação baseado no Centrifuge, Spingo e QIIME para avaliar a abordagem do Gunga com a integração das metagenômicas SH e 16S apenas, motivados pela previsão do uso desse modelo em dados do projeto Metazoo que possuíam apenas essas duas meta-ômicas. Além disso, a disponibilidade da amostra *mock* Mix3 com SH e 16S nos permitiu testar seu desempenho

em uma amostra de sequências completamente reais. Previamente, também testamos seu desempenho com as dez amostras de teste usadas na avaliação dos modelos com AM para validar o modelo e observamos melhorias significativas de F1 em relação à meta-ômica individual e com uma pequena diferença para o modelo com a mesma ferramenta e as três meta-ômicas (Figura 5.3).

Figura 5.3: Desempenho da abordagem baseada em duas meta-ômicas (SH e 16S) em amostras de teste.

5.2.3 Mix3

5.2.3.1 Descrição da Amostra

Mix3 possui a mesma composição de Mix1 e Mix2, porém com abundâncias relativas diferentes. Dentre as amostras *mock*, ela foi a amostra para a qual obtivemos a maior dissimilaridade com a distância de Bray-Curtis em espécie (Figura 5.4). Complementamos o conjunto com dados de MT simulados para a avaliação do Gunga com as três meta-ômicas.

5.2.3.2 Desempenho dos Classificadores Taxonômicos

O Gunga com a integração das meta-ômicas com Kraken2, Spingo e QIIME obteve um desempenho perfeito para a amostra, porém com a integração de duas meta-ômicas com Centrifuge, Spingo e QIIME teve o segundo melhor desempenho, mantendo uma predição incorreta e eliminando duas corretas do perfil final. A vantagem do Gunga nessa amostra foi a baixa presença de predições incorretas em comparação com as demais ferramentas – com exceção do MetaPhlAn2, todas tiveram uma boa sensibilidade mas uma precisão ruim com a presença de muitas predições incorretas.

Figura 5.4: Composições taxonômicas das amostras Mix3, usada para a validação da ferramenta, e demais amostras *mock* Mix1 e Mix2 à esquerda. Distâncias de Bray-Curtis e agrupamento das amostras à direita.

Tabela 5.2: Desempenhos das ferramentas de classificação, de combinação de classificadores e do Gunga com dados do Kraken2 RF para Mix3.

5.2.4 CAMI Toy Data HC[1-5]

As amostras de alta complexidade do CAMI são bastante parecidas na composição de espécies, sendo HC1 a mais diferente com 31 exclusivas. Dessa forma, consideramos escolher apenas uma das amostras para a validação da ferramenta com amostras complexas. Porém, ao examinarmos as dissimilaridades entre as amostras com a distância de Bray-Curtis, percebemos que há maiores diferenças entre elas no aspecto da abundância relativa do que no da presença e ausência das espécies (Figura 5.5). Decidimos manter a análise com as 5 amostras então, considerando que essas diferenças poderiam influenciar os resultados das ferramentas de classificação taxonômica, por diferenças na amostragem de sequências de uma espécie e nas distribuições de *scores* de suas sequências classificadas

por exemplo.

Figura 5.5: Composições taxonômicas das amostras HC[1-5] do CAMI com as 10 espécies mais abundantes em destaque à esquerda. Distâncias de Bray-Curtis e agrupamento das amostras à direita.

5.2.4.2 Desempenho dos Classificadores Taxonômicos

As variações dos desempenhos das ferramentas com as diferentes amostras foram baixas, fazendo-nos acreditar que as abundâncias relativas foram menos importantes do que a presença ou ausência das espécies para influenciar os resultados para essas amostras. O Gunga com a integração baseada no Kraken2 foi novamente o melhor em F1, com uma pequena vantagem sobre o Majority Vote com dados de SH – o Majority Vote foi a abordagem mais precisa, porém menos sensível que o Gunga (Figura 5.6).

Figura 5.6: Desempenhos das ferramentas de classificação, de combinação de classificadores e do Gunga com dados do Kraken2 RF para as amostras HC[1-5] do CAMI.

Novamente observamos um pior desempenho para as meta-ômicas individuais. Porém dentre essas ferramentas o SPINGO obteve o melhor F1, principalmente por causa de uma boa precisão, menor apenas do que as das ferramentas baseadas em marcadores, abordagens de combinação de classificadores e o Gunga, e de uma sensibilidade mediana entre as ferramentas de marcadores e composicionais.

5.3 Uso do Gunga em Dados Reais

Nesta seção apresentamos os resultados da aplicação do Gunga em dados reais do projeto Metazoo. Selecionamos amostras de dois ambientes estudados pelo projeto para descrever seus resultados com o Gunga: uma amostra de fezes de um macaco *bugio* criado em cativeiro e uma amostra de um dia de uma compostagem.

5.3.1 Bugios

5.3.1.1 Descrição da Amostra

Os *bugios* são um sub-projeto do Metazoo para compreender o impacto que os estilos de vida livre e em cativeiro tem na composição microbiana do intestino desses animais. Sua abordagem é identificar as principais diferenças entre esses dois grupos com SH e 16S, com sucessivas coletas de fezes de indivíduos dos dois grupos, sequenciamento e classificação taxonômica. Para esta análise escolhemos as primeiras coletas dos animais de vida livre, intituladas como **Vlp**.

As sequências das amostras foram obtidas com sequenciamento *pair-ended* com a plataforma Illumina, com a seleção e amplificação das regiões V3-V4 em 16S com o par de *primers* 341F-805R (FWD:CCTACGGGNGGCWGCAG; REV:GACTACHVGGGTATCTAATCC). Os mesmos parâmetros para a análise das sequências foram empregados com uso do *read* 1 em SH, e *reads* 1 e 2 em 16S – com a adição do uso do *fastq-join* para junção dos *pairs* e a modificação do uso do método *split_libraries.py* do QIIME para demultiplexação e filtragem por qualidade (ver Apêndice B).

As sequências obtidas das diferentes coletas foram analisadas em conjunto. Com a execução das filtrações iniciais obtivemos conjuntos com 1.431.643 sequências em SH e 95.421 em 16S.

5.3.1.2 Resultados da Classificação Taxonômica

Dentre 2.706 espécies obtidas com a integração das meta-ômicas apenas 3 foram classificadas pelo Gunga como corretas, sugerindo que ele não teve uma boa sensibilidade para essa amostra. Por outro

lado, o MetaPhlAn2 identificou apenas 4 espécies – *Bacteroides ovatus*, *Bacteroides cellulosilyticus* e *Roseburia hominis* em comum com o Gunga, e *Bacteroides finegoldii* exclusivamente. Esperávamos uma melhor sensibilidade do Gunga em comparação com uma ferramenta de marcadores, porém, acreditamos que há dois fatores principais para o baixo desempenho nessa amostra: (i) percebemos anteriormente que a abordagem com duas meta-ômicas levam a redução do desempenho; e (ii) a composição da microbiota intestinal de primatas não humanos é pouco conhecida, e há poucos genomas de suas espécies depositados nos bancos de dados usados pelas ferramentas.

5.3.2 Compostagem

5.3.2.1 Descrição da Amostra

O sub-projeto da compostagem estuda as composteiras do zoológico de São Paulo para compreender a atividade da comunidade microbiana nesse ambiente de degradação de biomassa, e suas variações durante o processo decorrente do impacto das mudanças de temperatura e outras propriedades físico-químicas do meio. O processo de compostagem do zoológico de São Paulo dura cerca de 3 meses. Para uma composteira específica, chamada de ZC4, diversas coletas foram feitas para acompanhar o processo [6]. Escolhemos a amostra do dia 30, cujo rótulo é **ZC4D30**. Para essa amostra o projeto metazoo obteve dados de 16S, SH, e MT, sendo portanto um conjunto de dados propício para teste da ferramenta desenvolvida nesta tese.

As sequências de 16S foram analisadas com os mesmos métodos e parâmetros usados em **Vlp**, porém as de SH e MT foram classificadas usando o Kraken2 ao invés do Centrifuge. Após filtragem ficamos com um total de 4.818.327 sequências em SH, 4.469.967 em MT e 548.865 em 16S.

5.3.2.2 Resultados da Classificação Taxonômica

O Gunga identificou 4722 espécies como predições incorretas dentre 4774 possíveis advindas da integração das três meta-ômicas, sobrando um perfil final com 52 espécies. Dessas, 10 também estavam presentes dentre 30 espécies mais abundantes preditas no metatranscritoma de ZC4D30 em Antunes et al. (2016) [6], e 13 também estavam presentes dentre 19 espécies preditas pelo MetaPhlAn2 (Tabela 5.3). Muitas das espécies mantidas pelo Gunga foram bactérias termofílicas, que condizem com a característica do processo da compostagem, que ocorre em altas temperaturas.

Em contraponto com a amostra anterior, o Gunga com a integração das três meta-ômicas conseguiu manter uma maior sensibilidade em relação ao MetaPhlAn2 em ZC4D30. Acreditamos que isso se deu

tanto pela comunidade ser mais complexa como pelo modelo que usamos baseado no Kraken2 para SH e MT ter um melhor desempenho em sensibilidade. Na Figura 5.7 apresentamos uma comparação das densidades entre os grupos de predições corretas e incorretas para algumas variáveis do modelo.

Tabela 5.3: Espécies preditas pelo Gunga em ZC4D30, com as também preditas dentre as mais abundantes no metatranscritoma em Antunes et al. (2016) [6] marcadas com * e com as também preditas pelo MetaPhlAn2 destacadas em negrito.

Gunga
<i>Aeribacillus pallidus</i>
<i>Cupriavidus gilardii</i>
<i>Staphylococcus xylosus</i>
<i>Staphylococcus sciuri</i>
<i>Aerococcus urinaeequi</i>
<i>Bacillus coagulans</i> *
<i>Bacillus thermoamylovorans</i>
<i>Methanothermobacter wolfeii</i>
<i>[Clostridium] thermosuccinogenes</i>
<i>Lactobacillus acidipiscis</i>
<i>Pseudopedobacter saltans</i>
<i>Sphingopyxis granuli</i>
<i>Thermogutta terrifontis</i>
<i>Novibacillus thermophilus</i>
<i>Tistrella mobilis</i>
<i>Corynebacterium glutamicum</i>
<i>Corynebacterium variabile</i>
<i>Corynebacterium stationis</i>
<i>Rhodococcus rhodochrous</i>
<i>Saccharomonospora viridis</i>
<i>Saccharomonospora glauca</i>
<i>Mycobacterium thermoresistibile</i>
<i>Parageobacillus thermoglucosidasius</i>
<i>Gordonia bronchialis</i> <i>Sphingobacterium mizutaii</i>
<i>Pseudomonas pseudoalcaligenes</i>
<i>Leadbetterella byssophila</i>
<i>Castellaniella defragrans</i>
<i>Methylobacillus flagellatus</i>
<i>Acetobacter pasteurianus</i>
<i>Acetobacter pomorum</i>
<i>Legionella pneumophila</i>

Weissella paramesenteroides
Acinetobacter lwoffii
Acinetobacter baumannii
Thermomicrobium roseum
Providencia stuartii
Providencia rettgeri
Geobacillus thermodenitrificans*
Bacillus smithii
Thermobispora bispora*
Ruminiclostridium thermocellum*
[Clostridium] stercorarium
Mycobacterium phlei
Thermomonospora curvata*
Sphaerobacter thermophilus*
Symbiobacterium thermophilum*
Rhodothermus marinus*
Thermobacillus composti*
Thermobifida fusca*
Nitrosomonas europaea

Figura 5.7: Densidades para valores de alguns preditores usados na classificação dos dados integrados de ZC4D30.

CAPÍTULO 6

CONCLUSÕES

As análises com a integração das meta-ômicas SH, MT e 16S com o uso dos modelos de classificação de AM desenvolvidos neste trabalho demonstraram que esta abordagem fornece melhorias quanto à identificação de predições de espécies corretas e incorretas originadas em diferentes ferramentas de classificação taxonômica. Essas melhorias foram validadas sob a percepção de que o melhor modelo é aquele que oferece balanceamento entre precisão, com o maior número possível de espécies corretas dentre as preditas, e sensibilidade, com o maior número possível de espécies corretas detectadas na amostra.

6.1 Contribuições

As principais contribuições deste trabalho são as seguintes:

Novo conjunto de dados para análise de ferramentas de classificação taxonômica meta-ômicas: O conjunto de dados desenvolvido compreende um extenso levantamento de amostras procurando minimizar o viés da composição das amostras de teste nas avaliações das ferramentas com a simulação do sequenciamento de amostras “casadas” para as três meta-ômicas com diversidade entre elas.

Comparação dos desempenhos das ferramentas em meta-ômicas individuais: Verificamos que as ferramentas para meta-ômicas individuais SH e MT possuem um desnivelamento entre precisão e sensibilidade. Em particular, ferramentas de classificação taxonômica baseadas em método composicionais e alinhamentos são de maneira geral mais sensíveis que as ferramentas baseadas em marcadores, e por outro lado, as ferramentas baseadas em marcadores são as mais precisas. Esses perfis de desempenho justificam a busca por melhorias nos métodos de classificação taxonômica com a finalidade de se chegar a resultados mais equilibrados.

Comparação de *ensembles* para meta-ômicas individuais: Uma abordagem que conseguiu melhorias no equilíbrio entre precisão e sensibilidade foi as ferramentas de combinação de classificadores (*ensembles*) para dados de uma mesma meta-ômica. Porém, nenhum desses trabalhos ofereceu uma comparação entre elas, deixando uma lacuna sobre as diferenças de desempenho que poderíamos esperar delas. Fizemos essa comparação e demonstramos que uma delas é a mais sensível e menos precisa (WEVOTE) e duas delas são mais equilibradas (MetaMeta e Majority Vote).

Validação do uso das ferramentas de SH em MT: Nossas análises demonstraram que os resultados obtidos com essas ferramentas são similares em dados de SH e MT, sendo o resultado mais influenciado pela escolha da ferramenta do que da meta-ômica em si, e validam assim o uso das mesmas ferramentas de dados SH em dados de MT.

Comparação de métodos e ferramentas em 16S: A análise dos dados de 16S nos revelou que a análise taxonômica com sequências representativas de OTUs prejudica a sensibilidade na detecção de espécies. Além disso, dois dentre os classificadores mais usados com essa meta-ômica (UCLUST e *RDP Classifier*) possuem uma baixa precisão na classificação de espécies – desencorajando a classificação dessas sequências nesse nível. Porém, nosso estudo com o classificador SPINGO demonstrou que ele possui um desempenho superior a essas ferramentas, inclusive em gênero também, e é capaz de ter um desempenho mais equilibrado em precisão e sensibilidade que diversos classificadores de dados SH.

Comparação de ferramentas e algoritmos para o desenvolvimento de um abordagem baseada na integração de dados meta-ômicos: Avaliamos diferentes ferramentas e algoritmos de classificação de AM para o desenvolvimento de modelos. O resultado de integração das meta-ômicas foi superior a todas as outras baseadas em meta-ômicas individuais, com maior vantagem sobre as ferramentas individuais e menor sobre os *ensembles*. Os modelos baseados em ferramentas com baixa precisão e maior sensibilidade (principalmente composicionais) tiveram as melhorias mais significativas no desempenho e indicam que as variáveis extraídas e selecionadas dessas ferramentas são melhores preditores para a construção desses modelos. Comparando os desempenhos dos diversos modelos, escolhemos o algoritmo *Random Forest* para o desenvolvimento dos modelos finais, por demonstrar uma maior estabilidade com diferentes ferramentas.

Uma nova ferramenta para a integração de meta-ômicas e melhoria dos resultados de classificação taxonômica: Desenvolvemos o programa Gunga como a primeira ferramenta para facilitar o uso da abordagem desenvolvida no trabalho e incorporar nossos modelos treinados em nossos conjuntos de dados. Além disso, validamos a abordagem com o uso da ferramenta em dados simulados, *mock* e reais.

6.2 Trabalhos Futuros

Apesar do bom desempenho da abordagem desenvolvida, ainda há espaço para a continuação dos estudos em busca de desempenhos ainda melhores. Novos experimentos devem ser realizados afim de aprimorar os modelos nos vários aspectos relevantes ao seu desenvolvimento, como a avaliação de novos algoritmos de classificação de AM, integração das meta-ômicas a partir de novas ferramentas ou modificações na abordagem de seleção de variáveis.

APÊNDICE A

TEMPLATE PARA EXECUÇÃO DO GUNGA

A.1 Preliminares

O Gunga tem como dependência os arquivos **nodes.dmp**, **names.dmp** e **merged.dmp** do banco de dados de taxonomia do NCBI e do pacote **taxdumpr**, criado por nós para manipular os arquivos de taxonomia de NCBI com maior desempenho através de uma implementação das operações baseada em *hashmaps* carregados diretamente na memória, que são necessárias para os processamentos de taxonomia internos do Gunga. O **taxdumpr** pode ser instalado com o comando **devtools::install_github("felipepratalima/taxdumpr")**.

A.2 Execução de Código

O código a seguir inclui um conjunto completo de comandos para executar a abordagem do Gunga passo-a-passo com dados do Centrifuge para metagenômica *shotgun* (SH) e metatranscritômica (MT), e Spingo e QIIME com RDP para metagenômica de *amplicons* de rRNA 16S.

```
1 library(taxdumpr)
2 library(gunga)
3
4 ## Cria o objeto da classe Taxdumpr
5 taxdumprObject <- Taxdumpr(nodesDmpLocation = "nodes.dmp",
6   namesDmpLocation = "names.dmp", mergedDmpLocation = "merged.dmp")
7
8 ## Fluxo de comandos para SH
```

```

9 raw.classifications.sh <- loadRawCentrifugeClassifications("data-raw/sh/centrifuge_
  classifications.tsv")
10 classifications.sh <- processCentrifugeClassifications(raw.classifications.sh, taxdumpObject)
11 summaries.sh <- summarizeCentrifugeClassifications(classifications.sh)
12 raw.profile.sh <- loadRawCentrifugeProfile("data-raw/sh/centrifuge_profile.tsv")
13 profile.sh <- processCentrifugeProfile(raw.profile.sh)
14 sh <- assignCentrifugeProfileAsShotgun(profile.sh, summaries.sh)
15
16 ## Fluxo de comandos para MT
17 raw.classifications.mt <- loadRawCentrifugeClassifications("data-raw/mt/centrifuge_
  classifications.tsv")
18 classifications.mt <- processCentrifugeClassifications(raw.classifications.mt, taxdumpObject)
19 summaries.mt <- summarizeCentrifugeClassifications(classifications.mt)
20 raw.profile.mt <- loadRawCentrifugeProfile("data-raw/mt/centrifuge_profile.tsv")
21 profile.mt <- processCentrifugeProfile(raw.profile.mt)
22 mt <- assignCentrifugeProfileAsShotgun(profile.mt, summaries.mt)
23
24 ## Fluxo de comandos para o SPINGO
25 raw.classifications.spingo <- loadRawSpingoClassifications("data-raw/spingo/seqs.results.out")
26 classifications.spingo <- processSpingoClassifications(raw.classifications.spingo,
  taxdumpObject)
27 summaries.spingo <- summarizeSpingoClassifications(classifications.spingo)
28 raw.profile.species.spingo <- loadRawSpingoProfile("data-raw/spingo/seqs.species.out")
29 raw.profile.genus.spingo <- loadRawSpingoProfile("data-raw/spingo/seqs.genus.out")
30 profile.spingo <- processSpingoProfile(raw.profile.species.spingo, raw.profile.genus.spingo,
  taxdumpObject)
31 spingo <- assignSpingoProfileAsAmplicon(profile.spingo, summaries.spingo)
32
33 ## Fluxo de comandos para o QIIME com RDP
34 otu.table <- loadRawOtuTable("data-raw/qiime/otu_table.txt")
35 raw.classifications.qiime <- loadRawQiimeClassifications("data-raw/qiime/rep_set_tax_
  assignments.txt", "rdp")
36 classifications.qiime <- processQiimeClassifications(raw.classifications.qiime, otu.table,
  taxdumpObject)
37 summaries.qiime <- summarizeQiimeClassifications(classifications.qiime)
38 raw.profile.qiime <- loadRawQiimeProfile("data-raw/qiime/rdp_otu_table_mc2_w_tax_L7.txt")
39 profile.qiime <- processQiimeProfile(raw.profile.qiime, taxdumpObject)
40 qiime <- assignQiimeProfileAsAmplicon(profile.qiime, summaries.qiime)
41
42 ## Fluxo de comandos para juntar perfis e aplicar modelo para identificar t xons incorretos
43 gunga.profile <- joinProfiles(sh, mt, spingo, qiime, "species", taxdumpObject)
44 gunga.profile$classification <- applyCentrifugeClassifierModel(gunga.profile)

```

A.3 Visualização do Resultado

Se usarmos o código anterior como *template* iremos obter o resultado do Gunga através da coluna **classifications**, que criamos para indicar as predições de táxons corretas e incorretas. Usando o a biblioteca padrão do R, podemos então usar **write.table** para exportar o resultado armazenado no objeto **gunga.profile** (um *data frame* do R) para um arquivo em formato tabular, que permitirá a exploração do resultado através de vários programas, como o Excel. O código a seguir exporta o resultado para um arquivo **resultado.gunga.tsv**.

```
1 write.table(gunga.profile, "resultado.gunga.tsv", sep = "\t")
```

Uma segunda alternativa é usar o objeto dentro do próprio ambiente do R, e explorar suas colunas para obter as informações desejadas, como a relação entre as variáveis do modelo e as classes preditas nos dados do perfil juntado. O *template* de código a seguir foi usado para produzir as Figuras 5.2 e 5.7.

```
1 table(gunga.profile$classifications) # 15 corretas contra 1497 incorretas
2 gunga.profile %>%
3   reshape2::melt(.) %>%
4   subset(variable %in% c("abundanceSumForSh", "directScoreMADForSh", "directScoreMADForMt",
5     "directHitLengthQ1ForMt", "spingoSpeciesRelativeAbundanceForAmp", "
6     directSpingoSpeciesScoreMeanForAmp", "g_mappedQiimePercentageOfReadsForAmp" "g_
7     mappedQiimeOtuSizeProportionMeanForAmp")) %>%
8   rename(Valor = value) %>%
9   rename(Classe = classifications) %>%
10  ggplot(aes(log10(Valor), fill = Classe)) +
11  geom_density(alpha = 0.5) +
12  facet_wrap(~variable, scales = "free", ncol = 2) +
13  theme_minimal() +
14  theme(legend.position = "top")
```

APÊNDICE B

JUNÇÃO E FILTRAGEM DE SEQUÊNCIAS DE 16S NO

METAZOO

O *script* a seguir foi desenvolvido e usado pela equipe do projeto para a junção, demultiplexação, remoção dos *primers* e filtragem de sequências por qualidade em 16S:

```
1 # Arquivo AmpliconZc4.sh:
2
3 #!/bin/sh
4 ARQUIVO1=$1
5 ARQUIVO2=$2
6 AMOSTRA=$3
7 MAP=$4
8
9 mkdir $AMOSTRA/
10 cd $AMOSTRA/
11
12 fastq-join ../$ARQUIVO1 ../$ARQUIVO2 -o $AMOSTRA-
13
14 mv $AMOSTRA-join $AMOSTRA-join.fastq
15
16 convert_fastaqual_fastq.py -f $AMOSTRA-join.fastq -c fastq_to_fastaqual -o Fasta-Qual/
17
18 split_libraries.py -f Fasta-Qual/*.fna -q Fasta-Qual/*.qual -l 100 -s 20 -a 2 -H 10 \
19     -w 50 -g -M 1 -b 8 -e 1 -o $AMOSTRA-Split-Libraries/ -m ../$MAP
20
21 truncate_reverse_primer.py -f $AMOSTRA-Split-Libraries/*.fna -m ../$MAP \
22     -z truncate_remove -M 2 -o $AMOSTRA-Reverse-Primer-Split/
23
24 mv $AMOSTRA-Reverse-Primer-Split/*.fna ../$AMOSTRA-Filtered.fna
```

ARQUIVO1 e ARQUIVO2 são os arquivos FASTQ dos *pairs* 1 e 2, AMOSTRA é o nome do diretório de saída e MAP é um arquivo do QIIME com contém descrições das amostras, inclusive os *barcodes* necessários para a identificação de cada uma delas.

APÊNDICE C

PUBLICAÇÕES

Ao longo do doutorado, até o momento, o aluno foi autor nas seguintes publicações:

- A.M. Thomas, F.P. Lima, L.M.S. Moura, A.M. da Silva, E. Dias-Neto, J.C. Setubal. Comparative Metagenomics. *Methods in Molecular Biology*, v. 1704, pp. 243-260. Springer New York, 2018. Trata-se de um review sobre métodos de comparação para dados metagenômicos. A experiência conseguida ao longo do doutorado foi importante para permitir contribuição a este trabalho.
- G. Kinker, A. M. Thomas, V. Carvalho, F. P. Lima, A. Fujita. Deletion and low expression of NFKBIA are associated with poor prognosis in lower-grade glioma patients. *Scientific Reports*, v. 6, p. 24160, 2016. Este artigo é resultado de trabalho numa disciplina do programa de pós-graduação em bioinformática.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] Vanessa Aguiar-Pulido, Wenrui Huang, Victoria Suarez-Ulloa, Trevor Cickovski, Kalai Mathee, and Giri Narasimhan. Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis. *Evolutionary bioinformatics online*, 12(Suppl 1):5, 2016.
- [2] Guy Allard, Feargal J Ryan, Ian B Jeffery, and Marcus J Claesson. Spingo: a rapid species-classifier for microbial amplicon sequences. *BMC bioinformatics*, 16(1):324, 2015.
- [3] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.
- [4] Sasha K Ames, David A Hysom, Shea N Gardner, G Scott Lloyd, Maya B Gokhale, and Jonathan E Allen. Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics*, 29(18):2253–2260, 2013.
- [5] Florent E Angly, Dana Willner, Forest Rohwer, Philip Hugenholtz, and Gene W Tyson. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic acids research*, page gks251, 2012.
- [6] Luciana Principal Antunes, Layla Farage Martins, Roberta Verciano Pereira, Andrew Maltez Thomas, Deibs Barbosa, Leandro Nascimento Lemos, Gianluca Major Machado Silva, Livia Maria Silva Moura, George Willian Condomitti Epamino, Luciano Antonio Digiampietri, et al. Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics. *Scientific reports*, 6:38915, 2016.
- [7] Stavros Bashiardes, Gili Zilberman-Schapira, and Eran Elinav. Use of metatranscriptomics in microbiome research. *Bioinformatics and biology insights*, 10:19, 2016.
- [8] John Besser, Heather A Carleton, Peter Gerner-Smidt, Rebecca L Lindsey, and Eija Trees. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clinical microbiology and infection*, 24(4):335–341, 2018.
- [9] Arthur Brady and Steven L Salzberg. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nature methods*, 6(9):673–676, 2009.
- [10] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59, 2014.
- [11] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336, 2010.

- [12] Lilia C Carvalhais, Paul G Dennis, Gene W Tyson, and Peer M Schenk. Application of meta-transcriptomics to soil environments. *Journal of microbiological methods*, 91(2):246–251, 2012.
- [13] National Research Council et al. The new science of metagenomics: Revealing the secrets of our microbial planet. 2007.
- [14] Johannes Dröge, Ivan Gregor, and Alice Carolyn McHardy. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics*, 31(6):817–824, 2014.
- [15] Robert C Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.
- [16] Scott Federhen. The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143, 2011.
- [17] Jeremy A Frank and Søren J Sørensen. Quantitative metagenomic analyses based on average genome size normalization. *Applied and environmental microbiology*, 77(7):2513–2521, 2011.
- [18] Eric A Franzosa, Tiffany Hsu, Alexandra Sirota-Madi, Afrah Shafquat, Galeb Abu-Ali, Xochitl C Morgan, and Curtis Huttenhower. Sequencing and beyond: integrating molecular’omics’ for microbial community profiling. *Nature Reviews Microbiology*, 13(6):360–372, 2015.
- [19] Tracey Allen K Freitas, Po-E Li, Matthew B Scholz, and Patrick SG Chain. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic acids research*, 43(10):e69–e69, 2015.
- [20] Jorge Frias-Lopez, Yanmei Shi, Gene W Tyson, Maureen L Coleman, Stephan C Schuster, Sallie W Chisholm, and Edward F DeLong. Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences*, 105(10):3805–3810, 2008.
- [21] Elizabeth M Glass, Jared Wilkening, Andreas Wilke, Dionysios Antonopoulos, and Folker Meyer. Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols*, (2010):pdb–prot5368, 2010.
- [22] Zena M Hira and Duncan F Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015, 2015.
- [23] Sarah Hunter, Matthew Corbett, Hubert Denise, Matthew Fraser, Alejandra Gonzalez-Beltran, Christopher Hunter, Philip Jones, Rasko Leinonen, Craig McAnulla, Eamonn Maguire, et al. Ebi metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic acids research*, 42(D1):D600–D606, 2014.
- [24] Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. Megan analysis of metagenomic data. *Genome research*, 17(3):377–386, 2007.
- [25] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.
- [26] Juan Jovel, Jordan Patterson, Weiwei Wang, Naomi Hotte, Sandra O’Keefe, Troy Mitchel, Troy Perry, Dina Kao, Andrew L Mason, Karen L Madsen, et al. Characterization of the gut microbiome using 16s or shotgun metagenomics. *Frontiers in microbiology*, 7, 2016.
- [27] Daehwan Kim, Li Song, Florian P Breitwieser, and Steven L Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research*, 26(12):1721–1729, 2016.
- [28] Allan Konopka. What is microbial community ecology&quest. *The ISME journal*, 3(11):1223–1230, 2009.

- [29] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.
- [30] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, Second edition edition, 2014.
- [31] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. The sequence read archive. *Nucleic acids research*, page gkq1019, 2010.
- [32] Stinus Lindgreen, Karen L Adair, and Paul P Gardner. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific reports*, 6, 2016.
- [33] Sharmila S Mande, Monzoorul Haque Mohammed, and Tarini Shankar Ghosh. Classification of metagenomic sequences: methods and challenges. *Briefings in bioinformatics*, page bbs054, 2012.
- [34] Elaine R Mardis. Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402, 2008.
- [35] Victor M Markowitz, Natalia N Ivanova, Ernest Szeto, Krishna Palaniappan, Ken Chu, Daniel Dalevi, I-Min A Chen, Yuri Grechkin, Inna Dubchak, Iain Anderson, et al. IMG/M: a data management and analysis system for metagenomes. *Nucleic acids research*, 36(suppl 1):D534–D538, 2008.
- [36] Christopher Mason, Alexa McIntyre, Rachid Ounit, Ebrahim Afshinnkoo, Robert Prill, Elizabeth Henaff, Noah Alexander, Sam Minot, David Danko, Jonathan Foox, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *bioRxiv*, page 156919, 2017.
- [37] Alexa BR McIntyre, Rachid Ounit, Ebrahim Afshinnkoo, Robert J Prill, Elizabeth Hénaff, Noah Alexander, Samuel S Minot, David Danko, Jonathan Foox, Sofia Ahsanuddin, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome biology*, 18(1):182, 2017.
- [38] Peter Menzel, Kim Lee Ng, and Anders Krogh. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature communications*, 7, 2016.
- [39] Ahmed A Metwally, Yang Dai, Patricia W Finn, and David L Perkins. Wevote: Weighted voting taxonomic identification method of microbial sequences. *PloS one*, 11(9):e0163527, 2016.
- [40] Samuel S Minot, Niklas Krumm, and Nicholas B Greenfield. One codex: a sensitive and accurate data platform for genomic microbial identification. *bioRxiv*, page 027607, 2015.
- [41] José A Navas-Molina, Juan M Peralta-Sánchez, Antonio González, Paul J McMurdie, Yoshiki Vázquez-Baeza, Zhenjiang Xu, Luke K Ursell, Christian Lauber, Hongwei Zhou, Se Jin Song, et al. Advancing our understanding of the human microbiome using qiime. *Methods in enzymology*, 531:371, 2013.
- [42] Nam-phuong Nguyen, Siavash Mirarab, Bo Liu, Mihai Pop, and Tandy Warnow. TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics*, 30(24):3548–3555, 2014.
- [43] Rachid Ounit and Stefano Lonardi. Higher classification sensitivity of short metagenomic reads with clark-s. *Bioinformatics*, 32(24):3823–3825, 2016.
- [44] Rachid Ounit, Steve Wanamaker, Timothy J Close, and Stefano Lonardi. Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC genomics*, 16(1):236, 2015.

- [45] Michael A Peabody, Thea Van Rossum, Raymond Lo, and Fiona SL Brinkman. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC bioinformatics*, 16(1):1, 2015.
- [46] Vitor C Piro, Martin S Lindner, and Bernhard Y Renard. Dudes: a top-down taxonomic profiler for metagenomics. *Bioinformatics*, 32(15):2272–2280, 2016.
- [47] Vitor C. Piro, Marcel Matschkowski, and Bernhard Y. Renard. Metameta: integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome*, 5(1):101, Aug 2017.
- [48] Paraskevi N Polymenakou. Atmosphere: a source of pathogenic or beneficial microbes? *Atmosphere*, 3(1):87–102, 2012.
- [49] Gurdeep Rastogi and Rajesh K Sani. Molecular techniques to assess microbial community structure, function, and dynamics in the environment. In *Microbes and microbial technology*, pages 29–57. Springer, 2011.
- [50] Rebecca Rose, Olga Golosova, Dmitrii Sukhomlinov, Aleksey Tiunov, and Mattia Prosperi. Flexible design of multiple metagenomics classification pipelines with ugene. *Bioinformatics*, 2018.
- [51] Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature methods*, 14(11):1063, 2017.
- [52] Nicola Segata, Daniela Boernigen, Timothy L Tickle, Xochitl C Morgan, Wendy S Garrett, and Curtis Huttenhower. Computational meta’omics for microbial community studies. *Molecular systems biology*, 9(1):666, 2013.
- [53] Nicola Segata, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811–814, 2012.
- [54] Esther Singer, Bill Andreopoulos, Robert M Bowers, Janey Lee, Shweta Deshpande, Jennifer Chiniquy, Doina Ciobanu, Hans-Peter Klenk, Matthew Zane, Christopher Daum, et al. Next generation sequencing data of a defined microbial mock community. *Scientific data*, 3:160081, 2016.
- [55] Barton E Slatko, Andrew F Gardner, and Frederick M Ausubel. Overview of next-generation sequencing technologies. *Current protocols in molecular biology*, 122(1):e59, 2018.
- [56] Shinichi Sunagawa, Daniel R Mende, Georg Zeller, Fernando Izquierdo-Carrasco, Simon A Berger, Jens Roat Kultima, Luis Pedro Coelho, Manimozhiyan Arumugam, Julien Tap, Henrik Bjørn Nielsen, et al. Metagenomic species profiling using universal phylogenetic marker genes. *Nature methods*, 10(12):1196–1199, 2013.
- [57] Luke R Thompson, Jon G Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J Locey, Robert J Prill, Anupriya Tripathi, Sean M Gibbons, Gail Ackermann, et al. A communal catalogue reveals earth’s multiscale microbial diversity. *Nature*, 551(7681), 2017.
- [58] Duy Tin Truong, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature methods*, 12(10):902, 2015.
- [59] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunencko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *nature*, 457(7228):480–484, 2009.

- [60] Kévin Vervier, Pierre Mahé, Maud Tournoud, Jean-Baptiste Veyrieras, and Jean-Philippe Vert. Large-scale machine learning for metagenomics sequence classification. *Bioinformatics*, 32(7):1023–1032, 2015.
- [61] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, 73(16):5261–5267, 2007.
- [62] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46, 2014.
- [63] John C Wooley, Adam Godzik, and Iddo Friedberg. A primer on metagenomics. *PLoS computational biology*, 6(2):e1000667, 2010.
- [64] Iratxe Zarraonaindia, Daniel P Smith, and Jack A Gilbert. Beyond the genome: community-level analysis of the microbial world. *Biology & philosophy*, 28(2):261–282, 2013.