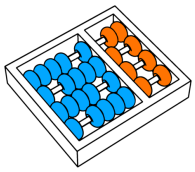
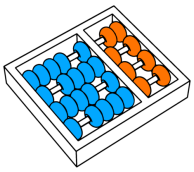


Bruno Malveira Peixoto

"Classificação de Sequências e Análise de Diversidade em Metagenômica"

CAMPINAS
2013



Universidade Estadual de Campinas
Instituto de Computação

Bruno Malveira Peixoto

"Classificação de Sequências e Análise de Diversidade em Metagenômica"

Orientador(es)
Prof. Dr. Zanoni Dias[1]
Prof. Dr. Guilherme Pimentel Telles[2]

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Computação da Universidade Estadual de Campinas para obtenção do título de Mestre em Ciência da Computação.

Este exemplar corresponde à versão final da Dissertação defendida por Bruno Malveira Peixoto, sob orientação de Prof. Dr. Zanoni Dias.

[1]Assinatura do Orientador(a)

[2]Assinatura do Orientador(a)

CAMPINAS
2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

P359c

Peixoto, Bruno Malveira, 1987-
Classificação de sequências e análise de diversidade em metagenômica / Bruno Malveira Peixoto. – Campinas, SP : [s.n.], 2013.

Orientador: Zanoni Dias.
Coorientador: Guilherme Pimentel Telles.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

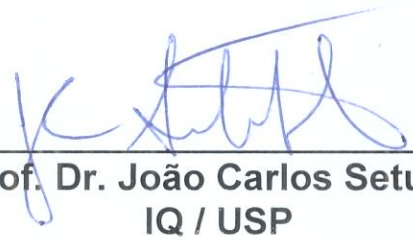
1. Metagenômica. 2. Bioinformática. 3. Biodiversidade. I. Dias, Zanoni,1975-. II. Telles, Guilherme Pimentel,1972-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações para Biblioteca Digital

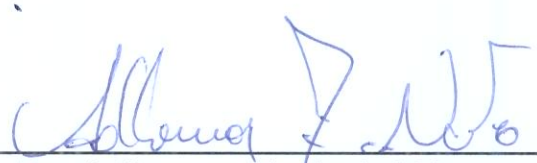
Título em inglês: Classification of sequences and diversity analysis in metagenomics
Palavras-chave em inglês:
Metagenomics
Bioinformatics
Biodiversity
Biodiversity
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Zanoni Dias [Orientador]
João Carlos Setubal
Adhemar Zerlotini Neto
Data de defesa: 19-04-2013
Programa de Pós-Graduação: Ciência da Computação

TERMO DE APROVAÇÃO

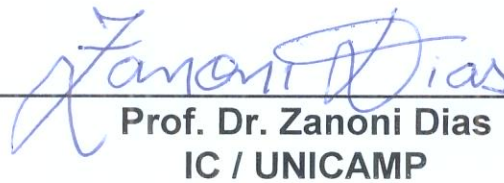
Dissertação Defendida e Aprovada em 19 de Abril de 2013, pela Banca examinadora composta pelos Professores Doutores:



Prof. Dr. João Carlos Setubal
IQ / USP



Dr. Adhemar Zerlotini Neto
Informática Agropecuária / EMBRAPA



Prof. Dr. Zanoni Dias
IC / UNICAMP

Instituto de Computação
Universidade Estadual de Campinas

Classificação de Sequências e Análise de Diversidade em Metagenômica

Bruno Malveira Peixoto[1]

19 de abril de 2013

Banca Examinadora:

- Prof. Dr. Zanoni Dias (Supervisor/*Orientador*)
- Prof. Dr. Guilherme Pimentel Telles (Supervisor/*Orientador*)
- Prof. Dr. João Carlos Setubal
  Instituto de Química - USP
- Dr. Adhemar Zerlotini Neto
  Embrapa Informática Agropecuária
- Dr. Francisco Pereira Lobo
  Embrapa Informática Agropecuária (*Suplente*)
- Prof. Dr. João Meidanis
  Instituto de Computação - UNICAMP (*Suplente*)

[1]Financial support: CNPq scholarship (process 133033/2011-2) 2011–2013

Abstract

Metagenomics is the genetic study of an environmental sample, and allows the analysis of non-culturable organisms. It is a new area of genetic study and has many computational challenges, with few dedicated tools available. The objective of this work is to realize a metagenomic study of the microbial diversity of samples from the composting unit of Fundação Parque Zoológico de São Paulo, analyzing and comparing existing programs and methods. The study of the microbial communities that live there is of great importance to a better understanding of the biological role of these organisms. Simulated data were tested to validate the methods we used with the real data and the results show that even well-known tools are biased by the reference databases it uses.

Resumo

Metagenômica é o estudo genético de uma amostra ambiental, e permite a análise de organismos incultiváveis em laboratório. É uma área de estudo nova e possui muitos desafios computacionais, com poucas ferramentas dedicadas disponíveis. O objetivo deste trabalho é realizar um estudo metagenômico da diversidade de organismos microbianos de amostras da unidade de compostagem da Fundação Parque Zoológico de São Paulo, analisando e comparando os programas e métodos existentes. O estudo das comunidades microbianas que ali vivem é de grande importância para um melhor entendimento do papel biológico desses organismos. Dados simulados foram testados para validar os métodos utilizados com os dados reais e os resultados mostram que mesmo programas conhecidos são sensíveis aos bancos de dados de referência que utilizam.

Agradecimentos

Dedico este trabalho aos meus pais. Sem eles nada disso seria possível, o amor e o apoio incondicionais deles foram o que me fez chegar até aqui.

Agradeço aos professores Zanoni Dias e Guilherme Pimentel Telles pela paciência, disposição que demonstraram durante todo esse tempo, e pelos incentivos que deram nesse período de grande aprendizado.

Agradeço também ao professor João Carlos Setubal, pelo convite para trabalhar no projeto e pelo apoio prestado durante esse tempo.

Agradeço aos meus irmãos e minha irmã, que mesmo distantes estiveram sempre presentes e dispostos a ajudar.

Finalmente, aos amigos que viveram comigo, que me acompanharam, que se tornaram uma segunda família, agradeço por todo o apoio e por tornarem tão proveitosa essa jornada. As amizades que foram construídas durante esse tempo jamais serão esquecidas.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Desde a década de 1970, os estudos de genomas vêm passando por diversos avanços. Tecnologias mais eficientes aceleraram o sequenciamento de milhares de espécies, e essa área continua crescendo. Esse sequenciamento de genomas, no entanto, é feito a partir do isolamento de um organismo com o objetivo de estudá-lo a fundo, descobrir o genoma desse organismo, os genes que o compõe e suas contribuições para o funcionamento desse organismo específico.

Recentemente, uma nova abordagem para o sequenciamento de genomas vem ganhando força, chamada de metagenômica. Em metagenômica estudamos um ambiente. Uma amostra natural é coletada e dessa amostra, o material genético encontrado nela é sequenciado. A partir desse sequenciamento é que estudamos a composição da população de seres vivos desse ambiente, as funções que eles desempenham entre si e como interagem com o ambiente em que vivem. Essa abordagem abre um leque de oportunidades de estudo sobre os mais diversos ambientes.

Motivados por essa nova abordagem, colaboramos com o projeto Estudos da diversidade microbiana do Parque Zoológico do Estado de São Paulo, do departamento de bioquímica do instituto de química da Universidade de São Paulo. Trabalhamos com dois conjuntos de amostras de uma unidade de compostagem do zoológico. A compostagem é um processo que aproveita matéria orgânica de diversas origens, desde excrementos de animais a galhos de árvores, e transforma em fertilizador para áreas agrícolas. Esse material tem uma grande riqueza microbiológica e conhecer as espécies que o compõe pode revelar muito sobre o funcionamento desse processo.

As complexas condições naturais existentes durante a compostagem tornam o ambiente difícil de ser reproduzido em laboratório. Além disso, alguns organismos não podem ser cultivados isoladamente, dependendo de outros para sobreviver. Por isso, analisar amostras extraídas diretamente do ambiente natural é uma melhor opção.

A metagenômica é uma área de estudos extensa e recente, e traz diversas dificulda-

des. O objetivo deste trabalho é focar em uma parte desse estudo, mais especificamente na análise de diversidade, quais organismos compõem as amostras e quais suas populações relativas, bem como avaliar métodos já existentes com dados reais do processo de compostagem.

O texto dessa dissertação está organizado da seguinte maneira. O Capítulo 2 apresenta uma série de conceitos básicos a fim de introduzir melhor o contexto em que o trabalho está inserido. O Capítulo 3 descreve os diversos passos de um projeto de metagenômica e suas diferenças com projetos de genômica tradicionais. O Capítulo 4 mostra alguns programas e ferramentas já desenvolvidas para a classificação e análise de diversidade de metagenomas. O Capítulo 5 traz conceitos mais específicos sobre a análise de diversidade e mostra uma série de estudos de caso em que esses conceitos foram utilizados. O Capítulo 6 discute a análise de diversidade feita em cima dos dados da compostagem do Zoológico de São Paulo. Finalmente, o Capítulo 7 apresenta as conclusões da dissertação e propõe algumas extensões para o trabalho.

Capítulo 2

Conceitos Básicos

Neste Capítulo faremos uma breve descrição de alguns conceitos básicos com os quais lidamos neste trabalho.

2.1 Genética

Um nucleotídeo é uma molécula composta por um açúcar chamado pentose, um grupo fosfato e uma base nitrogenada [25]. Nucleotídeos são diferenciados pela base nitrogenada que os compõem, que pode ser: adenina, citosina, guanina, timina ou uracila. A pentose de um nucleotídeo pode se ligar ao grupo fosfato de um outro, formando uma cadeia. Uma cadeia formada por vários nucleotídeos é chamada de polinucleotídeo.

Existem dois tipos de polinucleotídeos que armazenam informações genéticas: o *DNA* (ácido desoxirribonucleíco) e o *RNA* (ácido ribonucléico). Suas estruturas são representadas na Figura 2.1.

O DNA é formado por duas fitas de nucleotídeos, e a pentose que o constitui é a desoxirribose. As duas fitas do DNA são unidas através de pontes de hidrogênio formadas entre as suas quatro bases nitrogenadas. A adenina sempre forma pontes de hidrogênio com a timina, e a citosina com a guanina. As duas fitas de DNA são ditas complementares, e sempre é possível construir uma fita a partir da outra. A sequência de bases nitrogenadas ao longo da cadeia de DNA constitui a informação genética.

O RNA é composto por apenas uma fita e sua pentose é a ribose. A base nitrogenada timina, exclusiva do DNA, é substituída pela uracila, exclusiva do RNA. Uma fita de RNA pode se dobrar de tal forma que parte de suas próprias bases nitrogenadas se pareiam umas com as outra. Esse pareamento intramolecular é um fator importante no formato tridimensional do RNA, que é capaz de assumir uma variedade maior de formas complexas do que a dupla hélice de DNA.

Genética é um ramo da biologia que estuda os genes. Genes são as unidades orgânicas

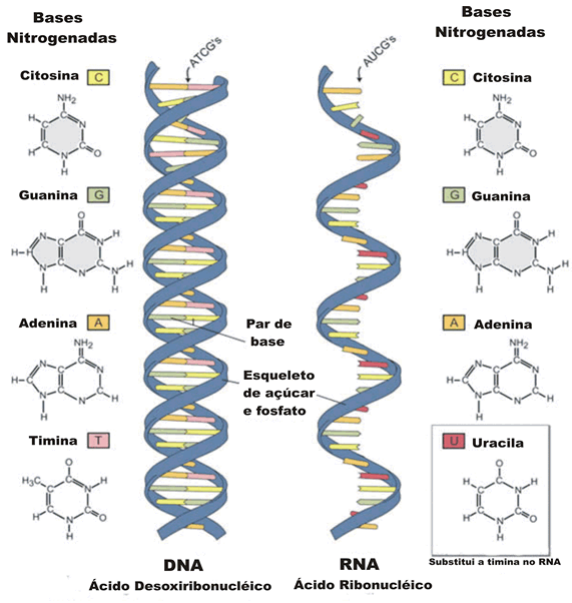


Figura 2.1: Estrutura das moléculas de DNA e RNA.

básicas que contém informações sobre as características físicas e comportamentais de um organismo e são passadas de geração para geração. Essas informações são usadas pra criar proteínas que constroem e regulam o funcionamento das células e do organismo com um todo. Os genes são regiões de moléculas de DNA que se encontram no interior de cada célula.

2.2 Expressão Gênica

Para produzir proteínas a partir do código genético, acontece uma série de transformações bioquímicas conhecida como expressão gênica. Uma sequência de bases do DNA é lida e segmentos do código são traduzidos em aminoácidos que juntos formarão uma proteína [25].

O processo de síntese de RNA a partir de um gene codificado no DNA é chamado de *transcrição*. Uma enzima chamada RNA polimerase se liga ao genoma em um marcador conhecido como promotor e começa a processar o DNA, formando um complexo de transcrição. A partir deste complexo, uma fita de RNA é formada com base na sequência de nucleotídeos do DNA. Esse RNA é conhecido como mRNA ou RNA mensageiro. Existem dois outros tipos de RNA: rRNA (RNA ribossomal) e tRNA (RNA transportador), que também são gerados a partir do DNA, e são utilizados na fase de tradução, descrita mais adiante.

Em organismo procariotos, como as bactérias e arqueobactérias, o mRNA não sofre

mais alterações. Nos eucariotos o mRNA é processado para a adição de uma sequência de bases adenina, chamada de poli-A, que dá maior estabilidade para o RNA. Também é feito o *splicing*, um processo de remoção de íntrons, que são regiões não-codificadoras do DNA, deixando apenas as regiões codificadoras, chamadas de éxons. Nesse processo, nem todos os éxons são preservados, e uma mesma sequência de DNA é capaz de gerar diferentes mRNAs, o que acaba aumentando a variabilidade genética desses organismos.

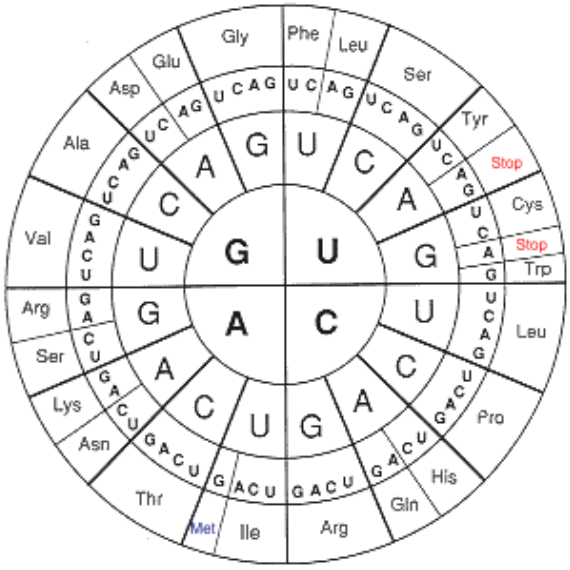


Figura 2.2: Código genético utilizado pela maioria dos organismos. A ordem das bases de um códon é lida do centro para as extremidades. O aminoácido Metionina, em azul, é também o códon de iniciação.

A tradução é o processo de síntese de proteínas a partir de um mRNA, e se baseia em conjuntos de triplas de nucleotídeos chamados *códons*. Existem $4^3 = 64$ códons diferentes, e cada um identifica um aminoácido, no entanto, um aminoácido pode ser identificado por mais de um códon, uma vez que existem 20 aminoácidos distintos. Um códon pode também ter um significado especial, marcando o início ou final de uma região de tradução (*start-codon* e *stop-codon*, respectivamente).

Existem tabelas que descrevem a relação entre um códon e o seu aminoácido correspondente. Na página do NCBI [1] podemos encontrar 17 tabelas. A Figura 2.2 mostra uma representação da Tabela número 1 da página do NCBI, que é o padrão para a maioria dos organismos.

A tradução começa com a ligação do mRNA com o *ribossomo*, que é uma organela no interior da célula responsável por percorrer o mRNA a partir do *start-codon*. Para cada

[1] http://www.ncbi.nlm.nih.gov/Taxonomy

códon, um tRNA é ativado, esse RNA transportador é responsável por ligar um códon específico a um aminoácido, juntando-o à cadeia de aminoácidos já existente e formando assim a proteína codificada.

2.3 Sequenciamento Genômico

O sequenciamento de um genoma é a determinação da cadeia de nucleotídeos que o compõe. Este processo envolve várias etapa, que serão brevemente descritas a seguir.

2.3.1 Preparação do Material Genético

Nos métodos atuais de sequenciamento existe uma limitação física que impede que sejam sequenciadas cadeias de nucleotídeos maiores que 1000 bases. Devido a essa limitação, a cadeia de nucleotídeos a ser estuda é primeiro fragmentada e depois remontada para obter a sequência original. Existem dois métodos principais para fragmentar o DNA: shotgun e digestão.

O método shotgun consiste em submeter o DNA a altas taxas de vibrações, fazendo com que a cadeia de nucleotídeos se quebre em vários pontos aleatórios. O método da digestão utiliza enzimas especiais, chamadas enzimas de restrição, que cortam o DNA em regiões específicas, conhecidas como sítios de restrição.

Os fragmentos obtidos são replicados através de um processo chamado de amplificação. A amplificação pode ser feita tanto por PCR (Polymerase Chain Reaction), que utiliza enzimas que sintetizam novas fitas de DNA a partir dos fragmentos existentes, quanto por clonagem, que insere os fragmentos em bactérias para serem replicados por elas.

2.3.2 Métodos de Sequenciamento

Atualmente existem vários métodos de sequenciamento. O método mais clássico que é eficiente e sujeito a automatização é o da terminação de cadeia. Ele é capaz de gerar centenas de sequências, também chamadas de *reads*, de tamanhos entre 800 e 1000 pares de bases. Recentemente, a grande demanda por sequenciamentos mais baratos acelerou o desenvolvimento de tecnologias de alto rendimento, que paralelizam o processo de sequenciamento e geram milhões de sequências de uma vez. No entanto, essas tecnologias de sequenciamento de segunda geração produzem sequências muito menores, de 50 a 500 pares de bases. A seguir os métodos mais conhecidos são brevemente descritos:

Método da terminação de cadeia [63]: também conhecido como sequenciamento Sanger, gera várias cópias de cadeias de nucleotídeos que diferem de tamanho por apenas uma base, que podem ser separadas e ordenadas.

Primeiro, os fragmentos de DNA são utilizados como moldes para a síntese de uma fita de DNA complementar. Para isso, são utilizados *primers*, que são oligonucleotídeos que se acoplam a regiões específicas do DNA e servem de iniciadores da reação de síntese do DNA. Oligonucleotídeos são fragmentos curtos de uma cadeia simples de ácido nucléico, comumente sintetizados em laboratório.

O processo de síntese do DNA complementar ocorre em uma solução que contém: uma enzima chamada DNA polimerase; os quatro tipos de desoxiribonucleotídeos (contendo as quatro bases do DNA); e alguns dideoxinucleotídeos equivalentes, que servem como terminadores de cadeia, pois quando um desses dideoxinucleotídeos é acoplado na fita de DNA complementar, impede a reação de continuar. Assim, faz com que existam vários fragmentos de DNA complementar de tamanhos diferentes, pois o dideoxinucleotídeo pode ter sido acoplado em qualquer posição, e com um número grande de fitas de DNA e de nucleotídeos, podemos esperar que todos os tamanhos de DNA complementar possíveis sejam construídos pela DNA polimerase.

Os fragmentos são separados por tamanho em um processo chamado eletroforese. Nesse processo, as cadeias são colocadas em um gel e é aplicada um campo elétrico que faz com que os fragmentos migrem para a direção oposta do gel. Os menores fragmentos tendem a migrar com mais facilidade, e ao final do experimento estarão mais próximas do lado oposto do gel.

Inicialmente, as reações com a DNA polimerase eram feitas em quatro recipientes separados, cada um com um dideoxinucleotídeo diferente, e colocadas no gel lado a lado. Após o processo de eletroforese, uma imagem de raio-X era tirada e a partir dessa imagem, observávamos faixas pretas, que representavam as cadeias de DNA que tinha percorrido o gel até uma certa altura. Comparando a altura com que as cadeias tinham chegado e o dideoxinucleotídeo utilizado, era possível identificar a base nitrogenada naquela posição.

Uma técnica mais recente dispensa a utilização de quatro reações diferentes. Cada tipo de dideoxinucleotídeo é ligado a um marcador químico fluorescente diferente, o que permite diferenciar em qual base cada fragmento de DNA termina. Ao final da eletroforese, é aplicada luz de forma a excitar esses marcadores químicos. Como as cadeias estão separadas por tamanho, é possível sequenciar o DNA a partir da última base de cada fragmento. Esse processo fluorescente tornou a automatização do sequenciamento mais viável e aumentou sua eficiência.

A Figura 2.3 mostra uma comparação entre os processos radioativo e fluorescente do sequenciamento Sanger. A parte de baixo da imagem representa o final do gel, e as menores cadeias chegam lá mais rapidamente. No processo radioativo, as alturas são comparadas entre si para identificar a base. No processo fluorescente, cada marcador químico quando excitado produz picos de luz que são registrados pela máquina de sequenciamento. O pico de uma determinada cor identifica a base, mas da mesma forma, as cadeias menores

chegam ao final do gel mais rapidamente.

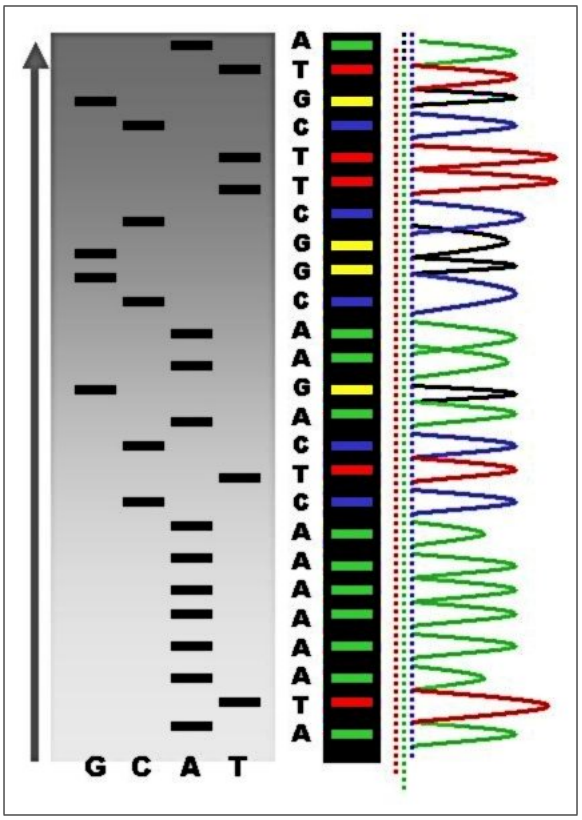


Figura 2.3: Bandas de cadeias de DNA em um gel usado no sequenciamento Sanger comparado com os picos fluorescentes do processo automatizado.

Pirosequenciamento 454 [60]: Pirosequenciamento é um método de sequenciamento paralelizado que detecta a atividade da enzima DNA polimerase com uma outra enzima quimioluminescente.

O DNA é amplificado dentro de gotas de água em uma solução de óleo onde cada gota contém um único molde de DNA acoplado a um *primer*. Cada um dos moldes de DNA é então submetido ao mesmo processo.

Soluções de cada tipo de nucleotídeo (A, C, G e T) e DNA polimerase são adicionadas e removidas em sequência de cada gota de água contendo um molde de DNA. Uma emissão de luz é produzida somente quando a solução com a DNA polimerase e o nucleotídeo correto complementa a primeira base não-emparelhada do molde. A sequência desses sinais quimioluminescentes determina a sequência do molde, uma base de cada vez. Eventualmente, um sinal mais forte é detectado, o que significa que mais de um nucleotídeo do mesmo tipo foi adicionado. Quanto mais forte o sinal, mais nucleotídeos daquele tipo foram adicionados à fita complementar.

Esse método é capaz de gerar milhões de sequências de aproximadamente 400-500 pares de bases.

Sequenciamento Illumina (Solexa) [2]: A Solexa, agora parte da empresa Illumina[2], desenvolveu uma tecnologia de sequenciamento paralelo em que moléculas de DNA são primeiro acopladas a *primers* e amplificadas de modo a formar colônias de clonagem locais.

Quatro tipos de dideoxinucleotídeos com marcadores fluorescentes são adicionados, e os nucleotídeos não incorporados na sequência são removidos da solução. Diferente do pirosequenciamento, o DNA só é estendido um nucleotídeo por vez. Uma câmera fotografa os dideoxinuleotídeos marcados e então o marcador e o terminal que bloqueia a continuação da síntese de DNA são removidos, permitindo um novo ciclo [39].

Esse tipo de sequenciamento gera milhões de sequências de 50-100 pares de base.

SOLiD® sequencing: É uma tecnologia da Applied Biosystems[3], que aplica sequenciamento por ligação. Em vez de usar a DNA polimerase, uma outra enzima, chamada DNA ligase é usada para identificar o nucleotídeo presente em uma dada posição do DNA.

DNA ligase é uma enzima que une as pontas de moléculas de DNA. O sequenciamento por ligação se utiliza da sensitividade dessa enzima para alinhar pares de bases.

O DNA a ser sequenciado passa por um processo de amplificação PCR e cada fita única de DNA gerada é processada paralelamente. Cada fita é acoplada a uma sequência conhecida através de uma pequena sequência "âncora". Uma mistura de oligonucleotídeos (com oito a noves bases) são adicionado à reação, marcados com substâncias fluorescentes de acordo com a posição que será sequenciada. Essas moléculas se unem ao DNA desconhecido nas proximidades da sequência âncora, e a DNA ligase preferencialmente une as moléculas quando as bases casam com às do DNA desconhecido. Baseado na fluorescência produzida pela molécula, é possível inferir qual nucleotídeo foi unido naquela posição da sequência.

É possível construir oligonucleotídeos de tal forma que se possa remover o marcador e permitir um novo ciclo de ligação, para gerar sequências mais longas. Esse método sequencia toda n-ésima base, onde n é o tamanho do oligonucleotídeo utilizado. Para sequenciar as posições intermediárias, a âncora e os oligonucleotídeos ligado são separados da fita de DNA e uma outra rodada de sequenciamento pode ser iniciada com uma âncora mais curta.

Esse método gera bilhões de sequências de aproximadamente 50 pares de bases.

[2]http://www.illumina.com/technology/sequencing_technology.ilmn
[3]http://www.appliedbiosystems.com

Capítulo 3

Fluxo de Trabalho de um Projeto Metagenômico

Os primeiros estudos de genomas que utilizaram procedimentos de sequenciamento foram feitos no meio da década de 1970. Desde então foram desenvolvidas técnicas de sequenciamento mais eficientes, os genomas de milhares de espécies se encontram em bancos de dados e ferramentas como o BLAST (Basic Local Alignment Search Tool) [1] permitem buscar e comparar sequências com as já existentes nesses bancos de dados.

Ainda assim, o sequenciamento de genomas de organismos isolados tem seus limites. Primeiro, limitações tecnológicas impõem que o organismo deve ser antes mantido em cultura e clonado para que se faça o sequenciamento de seu genoma completo [77]. No entanto, apenas uma pequena porcentagem de micróbios na natureza podem ser cultivados, o que significa que os dados genômicos existentes são altamente tendenciosos e não representam uma imagem verdadeira das espécies microbianas [54]. Segundo, raramente micróbios vivem em comunidades isoladas de outras espécies: espécies interagem tanto com outras quanto com o seu habitat, incluindo organismos hospedeiros [77].

Novas tecnologias e a drástica redução dos custos do sequenciamento ajudaram a superar essas limitações. Atualmente é possível obter informação genômica diretamente de comunidades microbianas em seus habitats naturais. Esses dados de sequências de DNA retirados diretamente do ambiente são chamados de Metagenoma [77].

Por não ser necessário obter culturas puras para o sequenciamento, a metagenômica promete revelar genomas da maioria dos microrganismos para os quais não se pode fazer culturas [31]. Além disso, como as amostras são obtidas de comunidades e não de populações isoladas, metagenômica pode servir para estabelecer hipóteses sobre a interação entre os membros da comunidade [38].

O processo de um projeto metagenômico típico é descrito brevemente a seguir, mostrando suas diferenças em relação a uma análise genômica tradicional, algumas tecnologias

existentes e os desafios computacionais associados.

3.1 Amostragem

O primeiro passo no estudo metagenômico é a obtenção de amostras ambientais. As amostras devem representar a população de onde elas foram retiradas. O problema é descobrir quantas amostras são necessárias para obter uma boa representatividade [77]. Curvas de rarefação são utilizadas para estimar a fração de espécies sequenciadas. Essas curvas relacionam o número de espécies observadas em função do número de amostras coletadas.

A complexidade da comunidade idealmente deve ser avaliada antes do sequenciamento [38]. A complexidade da comunidade é uma função do número de espécies em uma comunidade (riqueza) e sua abundância relativa (igualdade). Uma comunidade com muitas espécies que são próximas em abundância é mais complexa que uma comunidade com menos espécies que tem abundâncias desiguais.

A presença ou ausência de uma população dominante afeta bastante o tipo de análise que pode ser feita, independentemente do número total de espécies [38]. Populações dominantes que abrangem mais que uma pequena porcentagem do número total de células em uma comunidade terão uma representação maior no conjunto de dados, e podem ser estudadas mais a fundo.

3.2 Anotação de Metadados

Coletar dados associados com uma amostra ambiental melhora bastante a habilidade de interpretar as sequências, particularmente para uma análise comparativa de uma série temporal ou espacial [15, 74]. Tais metadados incluem dados bioquímicos, geográficos, data de coleta, método de extração de DNA, dentre outros. Metadados podem se tornar muito importantes quando dados suficientes são gerados para comparar comunidades [38].

Quanto mais condições sobre a amostragem dos metagenomas forem descritas, mais detalhadas podem ser as inferências das características ambientais sobre eles [51]. Para auxiliar na descrição de metadados existe um padrão, chamado *'Minimum Information about a Metagenomic Sequence (MIMS)'*, que pode ser seguido quando os dados forem submetidos a bases públicas [18].

3.3 Sequenciamento

Em metagenômica, o sequenciamento por *shotgun* é feito da mesma forma que em genomas de culturas clonadas. No entanto, o material genético original não é de apenas um organismo, mas de uma comunidade. Dependendo da amostra, o DNA fornece apenas um fragmento do genoma dos organismos daquele ambiente.

Métodos de sequenciamento de segunda geração como os descritos na Seção 2.3.2 estão substituindo o sequenciamento Sanger principalmente para comunidades virais e bacterianas [38]. Algumas vantagens do pirosequenciamento sobre o sequenciamento Sanger, por exemplo, incluem o custo mais baixo por base sequenciada e o fato de não ser preciso passar por uma etapa de clonagem [59]. No entanto, o tamanho médio do read é bem menor, dificultando o trabalho de montagem e de predição de genes, explicados em mais detalhes posteriormente. Muitos estudos que usaram o pirosequenciamento para análises metagenômicas não tentaram fazer montagem ou predição de genes, confiando apenas na busca de similaridades dos reads pequenos com bancos de dados de referência como base para as suas análises [22, 69, 74].

3.4 Montagem

Montagem é o processo de combinar sequências lidas em trechos contínuos de DNA, chamados *contigs*, baseado nas similaridades das sequências [38]. Um contig é uma *sequência consenso* que representa um alinhamento múltiplo de reads. Cada nucleotídeo da sequência consenso vem de sua posição correspondente dos reads alinhados e pode ser ou o nucleotídeo de maior qualidade ou o tipo de nucleotídeo mais frequente dentre todos os que estão naquela determinada posição. A *cobertura* de uma sequência consenso é o número médio de reads que a representa. Quanto maior a cobertura, mais confiável é a posição de um determinado nucleotídeo no consenso.

Quando se está sequenciando um genoma completo, os reads são montados em sequências ou contigs cada vez maiores, e finalmente no genoma completo [77]. Lidando com dados genômicos, é comum analisar grandes sequências. Em contraste, na maioria dos metagenomas, uma montagem completa não é possível. Primeiro, porque a amostragem é incompleta e muitos, se não todos, os genomas das espécies estão apenas parcialmente amostrados. Segundo, porque as informações das espécies em si é incompleta, e é difícil mapear reads individuais a suas espécies de origem. Existe também o perigo de se montar sequências a partir de reads de espécies diferentes, gerando quimeras.

Geralmente é possível montar a maior parte dos genomas de ambientes que tenham um pequeno número de espécies dominantes [20], porém, amostras com alta riqueza de espécies, como solo [72], dificilmente podem ser montadas.

Montadores atuais como Phrap [17], Arachne [35], CAP3 [29], e Celera [45] estão sendo adaptados e usados para montar metagenomas. A montagem de metagenomas é um processo problemático e qualquer programa de montagem produzirá vários erros [38]. Idealmente, toda montagem metagenômica deve ser inspecionada manualmente. Erros de montagem podem ser identificados com ferramentas de visualização, como o Consed [24], que são utilizados para facilitar a finalização de genomas. Ainda assim, a grande quantidade de dados metagenômicos impossibilita a inspeção manual, e por sua vez a correção de todos os erros de montagem identificados.

Uma abordagem para aliviar os problemas de montagem é o uso de sequências de referência. No entanto, o número de genomas de referência ainda é insuficiente para montagens de metagenomas complexos e o processo de classificação só parece ser satisfatório em comunidades muito simples [51].

Outra estratégia para lidar com essa limitação é fazer duas ou mais montagens dos mesmos dados utilizando diferentes montadores [20] para facilitar a identificação de montagens errôneas.

Existem diversas técnicas utilizadas pelos algoritmos de montagem, e muitos deles, por exemplo, utilizam um modelo baseado em grafos. Cada read é representado como um vértice e cada sobreposição entre reads é representado como uma aresta entre os vértices que se sobrepõem. O problema de encontrar a montagem correta é então reduzido a um conhecido problema de grafos, que é o de encontrar um caminho hamiltoniano em um grafo, isto é, um caminho onde cada vértice é visitado exatamente uma vez [77]. Para reads curtos, no entanto, essa técnica não é adequada. Para estabelecer uma cobertura adequada, reads curtos devem ser produzidas em grandes quantidades. Com grandes quantidades de reads, o grafo correspondente tem um grande número de vértices e arestas, e o tempo necessário para resolver o problema do caminho hamiltoniano cresce exponencialmente com o número de vértices. Com as grandes quantidades de sequências geradas pelos sequenciadores de segunda geração, o problema se torna intratável.

Uma solução adotada pelo montador Velvet [79] é a de usar os vértices para representar palavras, isto é, trechos de sequências, e os reads em si serem as arestas conectando os vértices. Assim, o grande número de reads e suas redundâncias não afetam o número de nós, e o problema é reduzido para outro problema de grafos, o de encontrar um caminho euleriano, ou seja, um caminho onde cada aresta é visitada exatamente uma vez. A vantagem dessa abordagem é que existem algoritmos mais eficientes para encontrar caminhos eulerianos do que para encontrar caminhos hamiltonianos.

Recentemente, foi desenvolvida uma extensão do montador Velvet, chamada MetaVelvet [46], para a montagem de reads curtos de metagenomas. A idéia é utilizar o grafo construído pelo Velvet, que no caso de metagenomas é uma mistura de sequências de várias espécies, e decompô-lo em subgrafos individuais, que servem de esqueleto para a

diferenciação de espécies dentre as sequências.

3.5 Predição de Genes

Genes são a unidade básica funcional em um genoma, que podem compor grandes unidades funcionais como *operons*, unidades de transcrição e redes funcionais [77].

Predição de genes é o procedimento de identificar proteínas e sequências de RNA codificantes nas amostras de DNA [38]. Duas abordagens diferentes são aplicadas para predição de genes: métodos intrínsecos e extrínsecos. Métodos intrínsecos analisam propriedades de sequências de genomas para diferenciar entre sequências codificadoras e regiões não-codificadoras [37]. As ferramentas que utilizam esses métodos são, em sua maioria, baseadas em aprendizado supervisionado e métodos estatísticos de reconhecimento de padrões. Uma grande vantagem de métodos intrínsecos é que eles permitem a identificação de genes sem homólogos em bases de dados disponíveis [38].

Métodos extrínsecos predizem genes procurando por fragmentos de DNA que foram conservados durante a evolução. A maioria dos novos genes são formados por eventos de duplicação, rearranjos e mutações de genes existentes [9]. Esses métodos usam ferramentas como o BLAST para identificar genes similares aos observados e informar a existência de famílias de genes dentre o metagenoma. Entretanto, o BLAST por si só não pode ser usado para encontrar novas famílias e novos genes [77].

Novamente, a natureza incompleta e fragmentada dos dados metagenômicos apresenta desafios na identificação de genes. Muitos reads permanecem intocados, em vez de serem unidos a outros e formarem contigs ou ainda são quimeras devido a erros de montagem [77].

Os métodos de predição de genes tem de ser adaptados para lidar com a enorme quantidade de genes fragmentados em sequências curtas, bem como com a diversidade filogenética das amostras, que dificultam o uso de conjuntos de treinamento específico de espécies, e com a baixa qualidade de sequências [51].

3.6 Classificação

Após a predição de genes, é importante associá-los às suas espécies de origem (ou grupos taxonômicos de alto nível). Essa análise é chamada *binning* ou classificação [77].

Uma forma de classificar sequências é encontrar similaridades com sequências de referência que podem ser usadas para construir uma árvore [77]. Essa técnica é útil quando a maioria das sequências da amostra possuem similaridades significantes com sequências de referência conhecidas. Sequências de genes preditas, quando não tem homólogos, são adicionadas em seu próprio nó isolado na árvore. O quadro resultante das sequências na

árvore de espécies pode mostrar uma visão geral das espécies dominantes da amostra.

No entanto, existem várias limitações para esse tipo de abordagem. Primeiro, a base de dados do genoma de referência é atualmente incompleta e altamente tendenciosa para apenas três filos de bactérias (*Proteobacteria, Firmicutes* e *Actinobacteria*) dentre pelo menos 50 filos [31]. Segundo, genes derivados de dados metagenômicos, particularmente aqueles com montagem mínima, são normalmente fragmentados e produzem alinhamentos incompletos. Terceiro, vários genes, particularmente de proteínas ribossomais, às vezes não são detectados por preditores de genes automáticos pelo seu tamanho reduzido [41]. Finalmente, genes informativos conservados filogeneticamente representam apenas uma pequena fração do total de dados metagenômicos [38].

Mesmo com tantas limitações, o algoritmo SOrt-ITEMS [26] consegue bons resultados. O algoritmo adota uma abordagem exaustiva para julgar primeiro a qualidade do alinhamento de uma sequência com seus *hits* e chega a um nível apropriado na árvore taxonômica para o qual a sequência pode ser atribuída. O algoritmo então usa uma abordagem de genes ortólogos para identificar hits que mostram uma similaridade recíproca com a sequência consultada. Genes ortólogos são genes de diferentes espécies que evoluíram de um ancestral comum através de especiação. Normalmente esses genes tem a mesma função através da evolução. O algoritmo avalia os alinhamentos obtidos entre o read e seus *hits* correspondentes na saída do BLASTx [1] para criar subconjuntos específico de *hits* que compartilham uma relação ortóloga com o read dado.

Uma outra abordagem é a classificação baseada em composição, que usa análises estatísticas das sequências [77]. Modelos de Markov baseados em frequências de k-mers se mostraram bem poderosos para análises estatísticas [64]. Esses métodos, porém, não são livres de erros. Quanto mais próximas filogeneticamente são as espécies estudadas no metagenoma, e quanto mais numerosas elas são, maior é a frequência de erros de classificação. A força da classificação baseada em k-mers é que sequências de referências não são necessárias para a classificação em si, toda a informação é obtida da própria sequência. Isso permite classificar sequências que tem poucos ou nenhum homólogo e portanto nenhuma função conhecida.

Outros métodos de classificação baseados em composição seguem o conhecimento de que processos celulares como uso de códons, sistemas de restrição-modificação, e mecanismos de reparo de DNA produzem assinaturas de composição de sequências, primariamente frequências de oligonucleotídeos (palavras), que são distintas em diferentes genomas [16].

Métodos baseados em composição podem ser divididos em procedimentos (de clusterização) supervisionados e não-supervisionados [38]. Procedimento não-supervisionados agrupam fragmentos metagenômicos em um espaço de assinaturas de composições sem a necessidade de um modelo de treinamento com sequências de referência. Uma vantagem da classificação não-supervisionada é que novas populações podem ser classificadas

por compartilharem semelhanças de características de composição, embora a identificação de fragmentos ainda necessite da similaridade de sequências para referenciar organismos. Uma desvantagem é que esses métodos tendem a focar em classes majoritárias de um conjunto de dados e não forneceram bons resultados em populações de baixa abundância.

Métodos supervisionados classificam fragmentos metagenômicos contra modelos treinados em sequências de referência classificadas e, a princípio, podem classificar fragmentos de populações de baixa abundância, se há um modelo para treinar. Como são capazes de aprender características relevantes que distinguem uma população em particular das outras, métodos supervisionados geralmente tem uma acurácia (sensitividade e especificidade) maior que modelos não-supervisionados, mas dependem de um conhecimento prévio sobre as espécies sequenciadas [38].

3.7 Análise de Dados

Os primeiros passos da análise de qualquer metagenoma envolvem comparar as sequências lidas de uma amostra com bases de dados de sequências conhecidas. Essa tarefa computacionalmente intensa provê os tipos básicos de dados para várias análises posteriores [44].

A análise de comunidades de baixa complexidade é, em vários aspectos, similar à análise de genomas isolados. Genomas de populações dominantes tem cobertura suficiente e contexto genético para permitir uma reconstrução metabólica razoavelmente compreensível [38]. Uma combinação de composição de sequências, classificação e montagem parece ser suficiente para sequenciar quase que completamente os membros da comunidade. Isso permite a atribuição de algumas atividades metabólicas para membros individuais do ecossistema [51]. Se mais de uma população dominante é sequenciada, a potencial interação metabólica dessas populações também pode ficar aparente [38].

Os sequenciamentos de comunidades microbianas de alta complexidade resultam em pouca ou nenhuma montagem de fragmentos [72]. A abordagem mais comum para a anotação funcional é baseada no BLAST, utilizando sequências de referência, principalmente de bactérias.

A alta densidade de áreas codificadoras em genomas de bactérias e o tamanho médio de um gene significa que a maioria dos reads obtidos pelo sequenciamento de bactérias irão capturar uma sequência codificadora [38]. Assim, é possível analisar os dados observando apenas os genes ali existentes e tratando a comunidade como um agregado, ignorando quais espécies contribuíram com qual gene.

Dessa forma, pode-se estudar o potencial funcional da comunidade microbiana do qual o metagenoma foi derivado [77]. Primeiro são atribuídas funções biológicas para os genes. Depois são descobertos genes que constituem redes biológicas, como vias metabólicas, nos dados. Vários estudos foram feitos e levaram à descoberta de vias metabólicas comple-

mentares de micróbios que constituem a comunidade.

Entretanto, essa técnica baseada em sequências de referência só permite anotar funcionalmente 25-50% das proteínas por metagenoma publicado [52].

3.8 Metagenômica Comparativa

A análise de sequências de genomas mostrou que muito se ganha com abordagens comparativas, uma vez que elas geram contexto para amostras individuais [51]. Comparações de amostras diferentes do mesmo ambiente ou um ambiente similar, podem revelar a influência de fatores ambientais particulares em comunidades microbianas. A comparação de metagenomas de diferentes habitats permite a descoberta de tendências gerais que ligam metagenomas e propriedades das comunidades com características fenotípicas dos ambientes.

O framework SEED [48] foi desenvolvido para genômica comparativa e foi usado como base para funcionalidades do servidor RAST [44]. São construídos mapas de taxonomias e subsistemas para encapsular diferenças entre as amostras e comparar suas similaridades.

Apesar do grande potencial de abordagens metagenômicas comparativas, elas devem ser aplicadas com cautela [51]. Vários fatores biológicos específicos do ambiente e vários problemas técnicos dificultam a comparação direta de ambientes, pois eles influenciam um ao outro e a maioria dos resultados derivados. Diferenças no tamanho médio de genomas das amostras implicitamente levam a diferenças da composição funcional relativa das amostras. A complexidade filogenética das amostras influenciam fortemente a análise, e diferentes características funcionais dos ambientes podem resultar em diferentes taxas evolucionárias, distorcendo a detecção de genes e funções. A cobertura limitada das amostras e a diversidade filogenética podem dificultar a comparação direta de parâmetros genéticos populacionais, uma vez que estimativas robustas baseadas em poucos dados são difíceis e as espécies abundantes podem ocultar a real estrutura populacional das amostras. Além dos vários fatores biológicos, muitos problemas técnicos relacionados a amostragem, sequenciamento e anotação influenciam toda a análise.

Capítulo 4

Programas para Análise de Metagenomas

Muitos programas foram desenvolvidos para auxiliar a análise de metagenomas. Montadores foram adaptados para lidar com a maior fragmentação dos dados, outros montadores foram desenvolvidos especificamente para trabalhar com metagenomas. Novos métodos de predição de genes também surgiram, bem como uma grande quantidade de classificadores que se utilizam de diversos métodos para conseguir um retrato da comunidade estudada. Neste Capítulo faremos um breve resumo de alguns programas, com ênfase em mostrar o que eles se propõem a fazer e os diferentes métodos que utilizam para tal. Os seis programas estudados são:

- MEGAN [34] - Utiliza uma série de buscas baseadas no BLAST.
- SOrt-ITEMS [26] - Evolui os resultados de buscas do BLAST identificando genes ortólogos.
- NBC [61] - Utiliza um classificador Bayesiano ingênuo.
- CARMA [36] - Constrói uma árvore filogenética baseada em comparações com famílias de proteínas.
- Phymm e PhymmBL [4] - Utiliza modelos de Markov interpolados e faz um híbrido com resultados do BLAST.
- PhylOTU [68] - Constrói uma árvores filogenética e utiliza as distâncias dessa árvore para clusterização. Este método utiliza apenas rRNA 16S para classificação.

4.1 MEGAN

O MEGAN (Metagenome Analyser) [34] é uma abordagem para o estudo inicial de um conjunto de dados metagenômicos e permite a análise de grandes conjuntos de dados por um único cientista. Em uma etapa de pré-processamento, os reads (ou contigs) são comparados com uma base de dados conhecida utilizando o BLAST. A principal aplicação do MEGAN é processar os resultados dessas comparações, através de estimativas e explorações interativas do conteúdo taxonômico do conjunto de dados. A taxonomia do NCBI é usada para resumir e ordenar os resultados.

O programa usa um algoritmo simples que atribui a cada leitura o *lowest Common Ancestor* (LCA) do conjunto de taxa que foram encontrados na comparação. Como resultado, reads mais específicos, mais semelhantes a sequências de espécies conhecidas, são atribuídas a taxa próximas às folhas da árvore NCBI, e sequências amplamente conservadas são atribuídas a taxa de alta ordem, próximas à raiz. Essa abordagem usa diversos valores limitantes:

- **min-score** é um limitante para a pontuação que um alinhamento deve alcançar para ser considerado nos cálculos.
- **top-percent** é um filtro que retém apenas aqueles hits para um dada leitura r que tenham uma pontuação dentro de uma porcentagem da maior pontuação envolvendo r. Isso ajuda a distinguir entre hits de identidade e de homologia.
- **win-score** pode ser determinado de forma que para aquela leitura, apenas hits com pontuação maior que esse limite sejam aceitas.
- **min-support** é um filtro usado para reduzir falsos-positivos, marcando um número mínimo de leituras que deve ser atribuído a um certo táxon ou qualquer um de seus descendentes na árvore taxonômica.

O resultado do algoritmo LCA é apresentada ao usuário como uma taxonomia parcial induzida pelo conjunto de taxa que foi identificado. O programa permite ao usuário explorar os resultados interagindo com a árvore taxonômica.

Esse processo foi aplicado a um subconjunto dos dados do Mar de Sargasso de Venter *et al.* [74]. A análise feita pelo MEGAN usou uma abordagem estatística independente e alcançou resultados muito similares aos de Venter *et al.* para a distribuição das espécies. Foi utilizado todo o conjunto de dados das amostras 1-4 do Mar de Sargasso, selecionadas algumas leituras e feitos dois conjuntos, um com leituras da amostra 1 e outro com leituras das amostras 2-4. O conjunto de leituras de DNA ou contigs foram comparados contra um banco de dado de sequências conhecidas, NCBI-NR, usando o BLAST.

Para o conjunto de reads da amostra 1, 83% de todas os reads foram atribuídos a taxa mais específicas que o nível de reino, a maioria sendo atribuído para grupos bacterianos. Para as amostras 2-4, o mesmo aconteceu para 59% dos reads.

Para validar o programa com tecnologias de sequenciamento recentes, que produzem reads menores, de até 100 bp, uma abordagem simples foi coletar reads de um genoma conhecido e processar como dados metagenômicos. Sequências de dois organismos foram usadas: *E. coli K12* e *B. bacteriovorus HD100*. *E. coli* foi escolhido por ser usado como hospedeiro de clonagem na maioria dos projetos de sequenciamento, e é mais provável que ocorra em várias sequências das bases de dados por engano. O segundo organismo de teste, *B. bacteriovorus*, tem sequências bem distintas de outras Proteobactérias e não tem parentes próximos atualmente representados nas bases de dados.

Foram simulados reads aleatórios e comparados com o banco de dados NCBI-NR usando o BLAST, e processados os resultados com o MEGAN. Não foram detectados falsos-positivos, demonstrando que leituras curtas em geral podem ser utilizadas em análises metagenômicas, mesmo com o custo de uma baixa taxa de predição.

Uma análise similar à do MEGAN foi obtida usando uma cópia da base de dados NCBI-NR na qual todas as sequências representando o genoma do *B. bacteriovorus HD100* foram removidas. Isso simula o caso de reads que não tem genomas representados em bases de dados. 65% dos reads não tiveram hits, e 13% não tiveram atribuições taxonômicas. Um pequeno número de falsos-positivos ocorreu por toda a árvore taxonômica, até o nível de filo.

Enquanto esses experimentos realizados com organismo de distâncias filogenéticas conhecidas demonstram a robustez do algoritmo LCA, a sua performance não é conhecida, sequências de distâncias maiores podem ser apenas estimadas.

4.2 SOrt-ITEMS

Haque *et al.* [26] apresentam um algoritmo de *binning* baseado em similaridade, chamado SOrt-ITEMS (Sequence ORTholog based approach for binning and Improved Taxonomic Estimation of Metagenomic Sequences). Esse algoritmo adota uma abordagem exaustiva para julgar primeiro a qualidade do alinhamento de uma sequência com seus hits e chega a um nível apropriado na árvore taxonômica para o qual a sequência pode ser atribuída. O algoritmo então usa uma abordagem baseada em genes ortólogos, que são genes de diferente espécies que evoluíram de um ancestral comum através de especiação. Ortólogos normalmente -mas nem sempre- tem a mesma função através da evolução. A identificação de hits que mostram uma ortologia significativa é tratada como uma identificação de hits que tenham similaridade recíproca com a sequência consultada. Esses hits são então identificados como ortólogos e utilizados para a atribuição final da sequência.

A atribuição de um read a um táxon em uma árvore filogenética pelo SOrt-ITEMS pode ser dividido em duas fases principais. A primeira fase envolve uma avaliação dos alinhamentos obtidos entre a leitura e seus hits correspondentes na saída do BLAST [1]. Os parâmetros de alinhamento usados para a avaliação da significância de um alinhamento incluem bit-score, tamanho do alinhamento, porcentagem de identidade, e positivos. Enquanto 'identidades' indicam a porcentagem de resíduos idênticos no alinhamento, 'positivos' correspondem à porcentagem de resíduos no alinhamento para os quais as pontuações de alinhamento tem valores positivos. Leituras com alinhamentos insignificantes são primeiro identificadas e categorizadas como 'Não-avaliadas'.

Supondo que sequências de diferentes organismos tenham se diferenciado (evoluíram) de ancestrais comuns, a identidade entre sequências de organismos diferentes decresce progressivamente de acordo com o aumento da distância entre esses organismos. O alinhamento entre duas sequências pertencentes a organismos que se diferenciaram em níveis relativamente altos da árvore taxonômica irá mostrar valores relativamente baixos de identidade e um grande número de mismatches, e vice-versa. Supondo uma taxa uniforme de evolução, os valores obtidos irão então indicar o provável nível taxonômico de onde as sequências divergiram. A atribuição do read é então limitada a esse nível taxonômico, já que esse nível não só seria conservativo, evitando assim falsos-positivos, como também é o nível mais específico para o qual a atribuição poderia ser feita.

Uma vez que o nível taxonômico apropriado é determinado, o algoritmo entra na segunda fase. Os passos dessa fase tem o objetivo de melhorar a especificidade da atribuição de cada read, baseado na identificação de um subconjunto específico de hits que compartilham uma relação ortóloga com o read dado. Dois parâmetros são utilizados pelo SOrt-ITEMS para identificar o subconjunto específico, são eles (1) o grau de ortologia entre o read e seus hits e (2) o grau de ortologia entre cada hit.

Para determinar os valores de limite para vários parâmetros de alinhamento para os reads obtidos das tecnologias de sequenciamento 454 (100-400 pares de bases) e Sanger (800-1100 pares de base), dois conjuntos de leituras simuladas correspondentes a essas duas tecnologias foram geradas usando o software MetaSim [56]. Seis variantes da base de dados NCBI-NR foram criadas removendo sequências correspondentes a espécie, gênero, família, ordem, classe e filo dos organismos de origem das leituras, respectivamente. Usando BLAST com os parâmetros padrões, todas as leituras simuladas pertencentes a um conjunto foram consultadas contra o NCBI-NR e contra todas as seis variantes que foram criadas a partir dele da forma anteriormente descrita. De cada uma das sete saídas do BLAST, parâmetros de alinhamento correspondentes aos melhores hits de cada read foram analisados.

Se o nome do táxon correspondente ao melhor hit pertence à mesma espécie que o organismo de origem do read, o read é marcado como 'diverge de espécie', indicando que

o nível taxonômico mais baixo de diferenciação do organismo de onde veio o read e aquele que se encontra na base da dados é o de espécie. Similarmente, as marcações de gênero, família, ordem, classe e filo são atribuídas aos reads.

Reads com marcações similares são agrupados e por sua vez, reads de cada grupo são sub-agrupados de acordo com os valores de identidade entre o read e o hit. Desses resultados, os limites de vários parâmetros de alinhamento são identificados para restringir a atribuição de reads à níveis taxonômicos específicos, aplicáveis para reads de 454 e Sanger.

Em resumo, os passos do algoritmo para identificação e atribuição taxonômica de leituras são:

- **Analisar informações de saída do BLAST**: Vários parâmetros de alinhamentos são obtidos a partir do BLAST. Esses parâmetros são recomputados depois da remoção de regiões de baixa complexidade mascaradas do alinhamento. Reads que não geraram nenhum hit são classificadas na categoria 'No Hits'.

- **Categorizar reads 'Não-atribuídos'**: Hits correspondentes a um read que tenham pontuações menores que 35 ou tamanho de alinhamento menor que 25 são removidos. Se o read não tem mais hits depois desse passo, é categorizado como 'não-atribuído'.

- **Reduzir a ocorrência de hits insignificantes**: Para os outros reads, apenas os hits que tenham pontuações dentre 10% do bit score correspondente ao melhor hit BLAST para aquele read são mantidos. Isso reduz o número de hits insignificantes.

- **Obter um nível taxonômico apropriado para atribuir**: Vários parâmetros de alinhamento obtidos para o melhor hit do read são usados pelo algoritmo para chegar a um nível taxonômico apropriado, onde a atribuição do read pode ser restringida. Uma vez que o nível taxonômico é identificado, o nome do táxon que ocorre nesse nível é obtido, e usado como substituto para o nome do táxon de melhor hit. Os nomes dos táxons de outros hits também são substituídos com seus respectivos nomes que ocorrem no nível taxonômico.

- **Identificar o ortólogo e atribuição taxonômica final do read**: Uma busca BLAST recíproca do read e seu melhor hit é feita contra a base de dados que contém o read e as sequências correspondentes aos outros hits. A ordem dos hits na saída BLAST dessa busca recíproca é obtida e os hits que precedem aquele obtido com o read são identificados como compartilhadores da mesma relação ortóloga com o read. Os nomes dos táxons substituídos correspondentes a esses hits são obtidos e usados como atribuição final do read.

Os resultados do SOrt-ITEMS foram validados utilizando sequências conhecidas e bases de dados controladas. Foram feitos testes também em dados metagenômicos reais, utilizando a base de dados do Mar de Sargasso [74]. Foi observado que geralmente o SOrt-ITEMS tem uma porcentagem maior de atribuições corretas comparada com o MEGAN. A única exceção é para o cenário de 'espécies conhecidas' correspondentes às leituras de 454. Entretanto, quando as leituras correspondem a novas espécies ou gêneros, o SOrt-ITEMS atribui corretamente um número relativamente maior de leituras que o MEGAN, indicando a precisão melhorada do SOrt-ITEMS.

Para ambos os métodos, entre 91% e 93% dos reads de Sanger originários de espécies conhecidas são atribuídas corretamente no nível de espécie. Para os reads de 454, SOrt-ITEMS atribuiu 64.4% corretamente no nível de espécie, comparado com 83.2% do MEGAN. No entanto, para os cenários em que as sequências dos bancos de dados correspondentes a espécie ou gênero do read foram retiradas do banco de dados, o número de atribuições corretas do SOrt-ITEMS é equivalente ao do MEGAN. Cerca de 5% dos reads são atribuídos corretamente em níveis específicos quando sequências correspondentes ao gênero são removidas da base de dados. A maior parte das atribuições nesse caso ocorrem em níveis intermediários da árvore taxonômica, ficando entre 15% e 20% para reads de 454 e entre 40% e 60% para reads Sanger, tanto com o SOrt-ITEMS quanto com o MEGAN. A grande vantagem do SOrt-ITEMS no entanto, é na quantidade de atribuições erradas, que é de 3 a 11 vezes menor que as do MEGAN para os reads de 454 e de 4 a 5 vezes menor para os reads Sanger.

4.3 Naive Bayes Classifier (NBC)

Métodos de classificação tradicionalmente se baseiam no alinhamento de sequências para comparar sua similaridade. Rosen *et al.* [61] apresentaram um método que usa um classificador de Bayes ingênuo (naive Bayes) que é capaz de identificar características significantes do DNA sem comparações diretas par a par. O método utiliza a frequência de N-mers, ou palavras, de uma sequência como característica chave para classificação.

O termo 'classificador' é usado no sentido de uma ferramenta estatística, treinado usando todos os dados genômicos, para discriminar entre classes. Cada classe é uma cepa, espécie, família, e assim por diante, dependendo da definição do rótulo. Nesse trabalho, os casos estudados foram com cepas, espécies e gêneros.

O classificador de Bayes ingênuo (NBC, do inglês naive Bayes classifier) é baseado na aplicação do teorema de Bayes supondo que cada característica na classificação é independente. Essa forte suposição é a ingenuidade do algoritmo, mas esse classificador já mostrou boa performance em situações complexas [57].

No caso estudado, as características das sequências são compostas de palavras de

DNA(N-mers). N-mers são palavras de tamanho N de DNA que podem ou não se sobrepor. O fundamento da análise feita por Rosen *et al.* é correlacionar a frequência desses N-mers numa sequência com sua identidade geral. É análogo a predizer o gênero de um livro pelas palavras contidas nele. Por exemplo, um livro sobre leis tem mais chances de conter altas frequências de palavras como 'lei', 'tribunal', 'veredicto' do que um texto sobre metagenômica, que dentre as palavras mais frequentes estão 'genoma', 'fragmento', 'sequência'.

O classificador de Bayes escolhe uma classe predita C, como a classe de maior probabilidade, dado um vetor de características observadas $\mathbf{w}$. A probabilidade de uma dada sequência pertencer a uma classe C_i é calculada usando a fórmula de Bayes:

$$P(C_i|\mathbf{w}) = \frac{P(\mathbf{w}|C_i) \cdot P(C_i)}{P(\mathbf{w})}.$$

Em outras palavras, a probabilidade de um fragmento f pertencer a uma classe C_i dada seu vetor de palavras $\mathbf{w}$ é igual à probabilidade dessas palavras aparecerem, dada uma classe C_i, vezes a probabilidade do genoma observado conter essa classe C_i, dividido pela probabilidade incondicional de observar as palavras $\mathbf{w}$ que compõem o fragmento f.

A probabilidade do genoma observado conter uma certa classe C_i é dada previamente pela amostra ambiental hipotética. Rosen *et al.* fizeram uma suposição de que a amostra era uniforme, ou seja, cada genoma é igualmente provável de pertencer à amostra. Com um conhecimento prévio do ambiente, uma estimativa melhor pode ser obtida.

Como é preciso conhecer as palavras (N-mers) como características para o algoritmo, uma implementação eficiente foi pensada para calcular as frequências de todos os N-mers. Um método geral que serve para qualquer tamanho de sequência é gerar todos os possíveis N-mers que se sobrepõem por $N-1$ nucleotídeos. Uma vez que todos os possíveis N-mers são gerados, eles são ordenados e suas cardinalidades e repetições são contadas.

O classificador foi testado em todas as sequências microbianas completas do Genbank do NCBI de fevereiro de 2008, o que eram 635 diferentes cepas, pertencentes a 470 espécies distintas e 260 gêneros. Os tamanhos dos fragmentos foram escolhidos como 500bp, 100bp, e 25bp. O tamanho de N-mer variou entre 3,6,9,10,11,12,13,14, e 15-mer.

Para validar o método, foram escolhidos 100 fragmentos aleatórios de cada genoma do conjunto de treinamento, totalizando 63500 fragmentos. Aumentar o tamanho de N aumenta consideravelmente a taxa de acerto do algoritmo, com fragmentos de 25bp, para $N = 9$, a maioria das cepas apresentaram uma taxa de classificação entre 0 e 5%, mas essa taxa sobe para 50% com $N = 12$. Como esperado, fragmentos de 500bp mostraram resultados bem melhores, conseguindo taxas de acerto próximas a 90% para 12-mers. Fragmentos de 25bp conseguem índices melhores de acerto com 15-mers, chegando a 75.8%. A taxa de acerto para gênero também é a melhor, e mantém a tendência de

aumentar de acordo com N. Essa taxa chega a um ponto de inflexão e nivela em 99.8% para 500bp, 99.3% para 100bp, e 97.5% para fragmentos de 25bp.

Rosen *et al.* também fizeram comparações com o BLAST. Para conduzir esse experimento, foi rodado o BLAST com os 63500 fragmentos contra o banco de dados de 635 genomas usado como base e considerados os resultados corretos no nível de cepa. Os resultados foram comparados com o caso de $N = 15$ do NBC. O BLAST calcula a significância de um alinhamento através de um e-value, que é o número de HSPs (highest scoring pairs) esperados ao acaso. Nos testes, desejaram que o BLAST retornasse o maior número possível de hits, portanto, o ideal seria um E-value infinito. Porém, muitos hits foram produzidos pelo programa para E-values acima de 3000, causando erros de memória, por isso, o E-value foi limitado a 3000.

Mesmo com o alto E-value, 287, ou 0.5% dos fragmentos não tiveram hits do BLAST. Muitos desses fragmentos são encontrados apenas uma vez em apenas um genoma em todo o banco de dados. Por causa dessa unicidade, o NBC foi capaz de classificar o genoma correto que o originou, 71% das vezes. O BLAST classificou apenas 66% dos fragmentos com um hit único com alta pontuação, e acertou todos eles. Comparativamente, o classificador de Bayes acertou 99% desses fragmentos. Nas situações em que existem vários hits com alta pontuação, o BLAST errou 13 deles, ou seja, há vários hits com alta pontuação, mas o correto não está nessa lista. O restante dos fragmentos tem a classificação correta em uma lista que varia entre 2 e 200 hits de alta pontuação. Se escolhido aleatoriamente um desses hits, o genoma correto pode ser adivinhado 29% das vezes em média. Desse conjunto, o NBC escolhe o genoma correto 31% das vezes.

Em resumo, o BLAST consegue encontrar o genoma correto (mesmo que ambíguo) de 63200 fragmentos, mas só resolve 41641 deles unicamente. Com os top hits e escolhendo aleatoriamente os hits ambíguos, o BLAST conseguiria 47889 (75.4%) corretos. O NBC conseguiu 48118 (75.8%) corretos. Com um N maior, o NBC potencialmente consegue resultados melhores no nível de cepa. Para fragmentos de 25bp, o NBC tem uma performance no mínimo tão boa quanto o BLAST.

Enquanto o BLAST pontua vários organismos com o mesmo valor, o NBC faz um ranking dos resultados e nunca reportou um empate em todos os 63500 fragmentos.

Também foram feitos experimentos com os dados sobre o Mar de Sargasso de Venter *et al.* [74]. Nessa análise, foram selecionados 10000 reads da amostra 1, o mesmo utilizado por Huson *et al.* na análise do MEGAN [34]. No caso, é feita uma classificação da cepa exata, com $N = 9$ e $N = 15$ e comparada com os resultados do MEGAN.

A análise de Venter do gênero *Burkholderia* na amostra 1 é de cerca de 38.5% da população. Com os mesmos 10000 reads da amostra 1, MEGAN reportou *Burkholderia* como 25.2% da amostra. O NBC mostrou 21% para 9-mers e 24.6% para 15-mers. Venter *et al.* estimaram 14.4% para o gênero Shewanella. O MEGAN encontrou 17.4% da amostra

e o NBC classificou 11.4% da amostra como Shewanella para 9-mers e 17.4% para 15-mers.

O classificado de Bayes ingênuo funciona bem no conjunto de treinamento, é comparável ao BLAST, e consegue classificar genomas de amostras ambientais. Os 9-mers tem uma baixa taxa de acerto para fragmentos de 25bp, mas essa taxa aumenta consideravelmente para 15-mers. No conjunto de treinamento, o classificador alcança uma taxa de acerto de 89% em cepas e 99.8% em gêneros para fragmentos de 500bp.

4.4 CARMA

Krause *et al.* [36] desenvolveram um algoritmo para classificação filogenética chamado CARMA em 2008. Desde então o algoritmo foi otimizado e sofreu algumas alterações, inclusive adaptações para a Web e em 2011 Gerlach e Stoye apresentaram o CARMA3 [21]. Aqui, iremos nos concentrar na ideia original de Krause *et al.* que mostra os principais conceitos por trás desse classificador.

O algoritmo proposto usa todos os domínios e famílias de proteínas do Pfam [19] como marcadores filogenéticos para identificar os organismos de fragmentos de DNA de uma amostra ambiental tão pequenos quanto 80 bp. Cada família do Pfam é representada por um alinhamento múltiplo completo de todos os membros da família, bem como um Pfam profile hidden Markov model (pHMM), que pode ser usado numa busca por membros novos desconhecidos. O método tem dois componentes: o primeiro identifica o domínio e fragmentos de famílias de proteínas em reads não montados de uma amostra usando pHMMs. Esses modelos são bons para detectar sinais funcionais fracos e pequenas sequências funcionais conservadas. Nesse estudo, o domínio ambiental e os fragmentos de famílias de proteínas identificados nos reads de uma amostra ambiental são definidos como *environmental gene tags* (EGTs), que podem ser usados para caracterizar um metagenoma quantitativamente. No segundo componente do método, uma árvore filogenética é reconstruída para cada família Pfam encontrada. EGTs são classificadas em taxonomias de alto nível baseado em suas relações filogenéticas com os membros da família com afiliações taxonômicas conhecidas, chamados de membros *taxaknown*.

Para avaliar o algoritmo, foram usados 77 genomas completos do GenBank. Um metagenoma sintético foi construído fragmentando os 77 genomas em sequências de tamanho entre 80 e 120 bp, com o tamanho médio sendo determinado em 100 bp.

No primeiro passo do algoritmo, uma busca por similaridade de cada read é feita contra o banco de dados do Pfam usando BLAST. Reads sem um hit do BLAST ou E-value menor que 10 são excluídos de análises futuras. Esse pré-processamento reduz o esforço computacional que precisa ser feito nas buscas com os pHMMs. Em seguida, todos os reads são analisados pelos Pfam pHMMs em busca de famílias de proteínas conservadas. Cada read é traduzido de acordo com o seu melhor hit do BLAST. No caso

de um read ter hits em mais de uma família Pfam, ele é separadamente traduzido para cada família. Depois, as sequências traduzidas são alinhadas com suas respectivas famílias usando o seu pHMM local, o que faz com que famílias de proteínas que são cobertas apenas parcialmente por um read possam ser identificadas. As sequências de todos os fragmentos de famílias Pfam identificados (EGTs) são adicionadas ao alinhamento múltiplo de sua respectiva família.

Os alinhamentos múltiplos de membros *taxaknown* e os EGTs de cada família Pfam são usados para calcular a distância par a par de todas as combinações entre membros *taxaknown* e seus respectivos EGTs. A distância entre duas sequências é definida como o percentual de identidade de sequências (PID). Uma árvore filogenética sem raiz é reconstruída a partir dessas distâncias usando o método neighbor-joining. Assim, EGTs são classificados dependendo de suas relações filogenéticas com respeito aos membros *taxaknown*. Se um EGT g está localizado em um grupo de membros *taxaknown* que dividem um mesmo taxon t, então g é classificado como t. Caso contrário, é classificado como 'taxon desconhecido'

O conjunto de dados simulados representa uma comunidade microbiana com fragmentos de sequências de Archaea e Bacteria, 10 filos, 11 classes, 29 ordens e 62 gêneros. Um EGT foi encontrado em aproximadamente 15% de cerca de 2.7 milhões de fragmentos analisados. Na média, a origem taxonômica foi corretamente predita para 84% no nível de domínio e 61% no nível de ordem dos EGTs identificados. Enquanto a proporção de EGTs classificados corretamente cai de domínio para ordem, a proporção de EGTs classificados erroneamente se mantém em 7% para todos os níveis taxonômicos. Por outro lado, a proporção de EGTs que não foram atribuídos a nenhum grupo taxonômico (táxon desconhecido), aumenta de 10% no nível de domínio para 31% no nível de ordem.

Os resultados mostram que fragmentos de famílias de proteínas Pfam são viáveis como marcadores filogenéticos para inferir afiliações taxonômicas em pequenos fragmentos de DNA ambiental. É uma boa alternativa a métodos que utilizam poucos marcadores, como o rDNA 16S, pois o uso de todas as famílias Pfam proporciona um retrato melhor da composição taxonômica da amostra ambiental.

No CARMA3, Gerlach e Stoye utilizam várias técnicas para otimizar os resultados do CARMA, e uma delas é a busca recíproca feita pelo SOrt-ITEMS. O primeiro passo do CARMA3 é usar o BLAST para buscar homólogos na base dados NCBI NR e a busca recíproca é usada para eliminar hits.

4.5 Phymm e PhymmBL

Phymm [4] é um programa que usa modelos de Markov interpolados (IMMs) para caracterizar oligonucleotídeos de tamanhos variáveis típicos de um grupo filogenético. Brady e

Salzberg apresentaram também um método híbrido PhymmBL, que utiliza informações tanto do Phymm quanto do BLAST e produz resultados melhores que cada um deles isoladamente.

Brady e Salzberg conduziram experimentos com reads metagenômicos sintéticos e associaram rótulos taxonômicos utilizando o Phymm, BLAST e PhymmBL. O Phymm contem IMMs treinados com cromossomos e plasmídeos de organismos obtidos do banco do NCBI.

Quando utilizado para pontuar uma sequência x de DNA, um IMM calcula uma pontuação correspondente à probabilidade da IMM ter gerado aquela sequência x, o que pode ser usado para estimar a probabilidade da sequência x pertencer à classe de sequências com as quais a IMM foi treinada.

No experimento com o Phymm, cada read foi pontuado por cada IMM na biblioteca de referência. O read então foi classificado usando o rótulo pertencente ao organismo cuja IMM gerou a melhor pontuação para aquele read. No experimento com o BLAST, cada read foi submetido a uma busca com o BLAST contra um banco de dados construído com os mesmos genomas usados para gerar os IMMs e o rótulo foi atribuído a cada read usando o rótulo do melhor hit do BLAST.

Já o PhymmBL pontuou cada read usando uma combinação dos métodos. A fórmula, determinada empiricamente, usada para pontuar foi:

$$Score = IMM + 1.2(4 - log(E))$$

Onde IMM é a pontuação do melhor IMM correspondente ao read e E é o menor (melhor) E-value retornado pelo BLAST. O IMM usado por Brady e Salzberg retorna pontuações logarítmicas, geralmente entre -500 e -100, com os valores maiores representando melhores pontuações. A constante 4 foi determinada experimentalmente como ótima através de uma busca binária em pequenos inteiros positivos (entre 0 e 5) e o peso 1.2 foi então determinado como ótimo através de uma busca binária nos valores entre 1 e 3. Os alcances dessas buscas foram estabelecidos identificando os valores nos quais as predições de um método completamente dominava as do outro. Por exemplo, pesos multiplicativos menores que 1 geraram pontuações combinadas essencialmente idênticas às obtidas pelos IMMs sozinho, enquanto aqueles maiores que 3 geraram pontuações combinadas que eram as mesmas das obtidas apenas pelo BLAST. Esses parâmetros podem representar apenas um ótimo local, mas valores diferentes dos pesos tiveram um efeito marginal no acerto geral do algoritmo.

O objetivo central do trabalho era o de modelar o problema de classificação de sequências para espécies nunca antes observadas. Por definição, dado um read de um organismo que nunca foi sequenciado, nenhum método de classificação pode predizer corretamente o rótulo de espécie, pois o mesmo não existe. Para classificações filogenéticas de alto nível,

no entanto, é possível existir o gênero, família ou outro nível taxonômico mais alto. Cada grupo de experimentos foi então repetido diversas vezes, com cada iteração configurada para explicitamente excluir comparações de cada read de entrada com espécies relacionada a ele em níveis cada vez mais altos.

Os resultados da classificação dos dados simulados com o Phymm mostraram que a classificação no nível de gênero é a tarefa mais difícil. Os experimentos foram feitos mascarando matches de nível de espécie, e repetidos com reads de tamanho 100, 200, 400, 800 e 1000 bp. No nível de gênero, o algoritmo classificou corretamente 32.8% dos reads de tamanho 100, mas o acerto pula para 89.8% com reads de tamanho 1000. Com 400 bp, o tamanho de read médio de um pirosequenciador 454, o acerto do algoritmo é de 60.3%. Esses resultados são grandes melhoras comparadas com métodos anteriores como PhyloPythia [43], método baseado em SVM (Support Vector Machine), que reportou um acerto de apenas 7.1% para o nível de gênero em reads de tamanho 1000 e CARMA [36], que classificou corretamente 6% dos reads de 100 bp no nível de gênero.

Na maior parte dos casos, o BLAST sozinho foi superior ao Phymm, mas para sequências de tamanho 800 e 1000, O Phymm supera os resultados do BLAST nos níveis de classe e filo.

O classificador híbrido PhymmBL produz resultados ainda melhores. Para todos os tamanhos de reads e em todos os níveis taxonômicos, PhymmBL é melhor que o Phymm e o BLAST usados independentemente, mostrando aproximadamente 6% a mais de acerto que o BLAST para todos os níveis taxonômicos com reads de tamanho 1000. Esses resultados indicam que os dois métodos são complementares e que o PhymmBL é capaz de usar informação de ambos. Tanto o Phymm quanto o PhymmBL retornaram resultados robustos, em todos os casos, o desvio padrão da taxa de acerto foi menor que 1%.

Os métodos também foram testados com um metagenoma real extraído de uma drenagem ácida de mina (DAM) [73] composta por três populações dominantes, uma de Archaea e duas de Bacteria. O PhymmBL acertou o filo da população de Archaea em 61% das vezes e as populações de Bacteria em 80% dos casos.

Uma das vantagens de usar modelos de Markov interpolados é que eles usam informações de múltiplos oligonucleotídeos de diferentes tamanho e integram os resultados. Então, em vez de ter que escolher entre 5-mers e 6-mers para classificação no nível de classe ou filo (como é feito pelo PhyloPythia), Phymm pode usar os dois. Nos experimentos, foram considerados k-mers variando de 1 a 12 e o Phymm automaticamente seleciona aqueles que melhor caracterizam cada espécie.

O BLAST é um método melhor por si só, mas quando integrado ao Phymm é capaz de gerar resultados ainda melhores.

4.6 PhylOTU

Sharpton *et al.* [68] argumentaram que o RNA ribossomal 16S é tradicionalmente amplificado por PCR usando primers universais. Cada produto do PCR é então sequenciado individualmente. Uma das grandes desvantagens disso é que o PCR tem mostrado viés relacionado a sequências de primers. Além disso, os primers ditos universais falham ao amplificar sequências suficientemente distintas daquelas utilizadas para criá-los. O resultado disso é que algumas espécies podem ser amplificadas desproporcionalmente ou até mesmo perdidas. Métodos metagenômicos eliminam esse viés com o sequenciamento de fragmentos aleatórios de uma amostra ambiental, e mesmo tendo seus próprios vieses, ainda podem fornecer uma caracterização mais realista da diversidade microbiana.

Por causa da natureza fragmentada, as sequências de metagenomas raramente apresentam sobreposições. PhylOTU é um programa que permite a identificação de OTUs microbianas diretamente de sequências metagenômicas não-sobrepostas.

A partir de sequências completas como referência, o programa constrói um perfil probabilístico do rRNA 16S. Esse perfil é usado para alinhar reads metagenômicos com as sequências de referência e esse alinhamento é então utilizado para computar a distância filogenética entre cada par de reads como entrada para um algoritmo de clusterização.

Primeiro, modelos probabilísticos que codificam a diversidade evolucionária e estrutura secundária do rRNA 16S de Bacterias e Archaeas são construídos com alinhamentos de sequências completas de referência através do Infernal [47].

Para um dado conjunto metagenômico, reads homólogos de 16S são identificados através de uma busca por BLAST contra o banco de dados STAP [78], que é um banco de dados não-redundante especializado em sequências de rRNA 16S. Essa busca permite diferenciar Archaea de Bacteria, o que acelera o alinhamento e a análise filogenética.

Alinhamentos múltiplos de reads metagenômicos são criados entre cada read de 16S com o respectivo perfil, de Archaea ou de Bacteria. Esse alinhamento é então mapeado nos alinhamentos de referência construídos com o Infernal. O passo final desse processo é um filtro de controle de qualidade que garante que apenas as sequências homólogas de rRNA 16S do domínio filogenético apropriado sejam incluídas no alinhamento final e mascara colunas do alinhamento com muitos buracos.

Sharpton *et al.* usam esses alinhamentos para construir uma árvore filogenética totalmente resolvida e para determinar a relação evolucionária entre os reads. As sequências de referência também são incluídas nesse estágio da análise para guiar a posição de reads muito curtos.

Após a construção da árvore, as sequências de referência são removidas e a filogenia resultante é utilizada para calcular uma matriz de distâncias filogenéticas, em que a distância entre um par de reads é definida como a distância total do caminho entre eles

na árvore. Essa matriz é então utilizada como entrada para o Mothur [66], para a geração de clusters.

A inovação do método é justamente o cálculo da distância filogenética (PD - Phylogenetic Distance) ser utilizado como medida de distância para clusterização, em vez do tradicional percentual de identidade de sequências (PID - Percent Sequence Identify). Quando aplicados os mesmos limitantes de distância de cluster para as distâncias PD e PID, a clusterização com a matriz PID produz uma estimativa de riqueza maior, ou seja, um maior número de OTUs, que a clusterização feita com a distância PD. No entanto, a composição geral dos clusters é muito similar.

O método do phylOTU foi testado com dados simulados, foram gerados 25 conjuntos de dados, cada um gerado a partir de 50 sequências de rRNA 16S escolhidas aleatoriamente de um total de 508. Em cada simulação, as 50 sequências usadas para gerar os dados foram retiradas do banco de dados de referência. Com um limitante de distância PD de 0.15, que de acordo com as análises de simulação, corresponde a uma distância PID de 0.03, ou seja, 97 de identidade entre as sequências completas, o programa obteve uma taxa de acerto de 80% com reads metagenômicos.

O phylOTU também foi usado para analisar os reads do Global Ocean Survey [62]. Com um limitante de distância PD de 0.15, foram identificados 1078 OTUs, comparados com os 811 OTUs identificados pelo projeto original com 97% de identidade.

Capítulo 5

Análise de Diversidade

A partir do estudo de diversidade de macro-organismos, foram adaptadas abordagens de documentação e análise de padrões de diversidade ambiental para serem aplicadas a micro-organismos [12, 75]. Comparações de comunidades em que os macro-organismos são tão diversos quanto numa comunidade microbiana sugerem que métodos de estimativa de riqueza de comunidade desenvolvidos para macro-organismos podem ser utilizados para amostras microbianas [32].

Em qualquer amostragem de uma comunidade, o número de tipos de organismos observados aumenta com o número de amostras, até que todos os tipos sejam observados. Essa relação fornece informações sobre a diversidade total da comunidade amostrada. Esse padrão pode ser visualizado através de uma curva cumulativa ou de uma curva de abundância.

Uma curva cumulativa é o gráfico do número acumulado de tipos observados pelo número de indivíduos amostrados. Como todas as comunidades contêm um número finito de espécies, se os pesquisadores continuassem a coletar amostras, as curvas eventualmente alcançariam uma reta no número real de riqueza da comunidade. Assim, essas curvas contêm informações sobre quão boa é a amostragem da comunidade estudada. Quanto mais a curva se aproximar de uma reta horizontal, melhor é a amostragem [32].

Uma forma mais pontual de se analisar a diversidade de uma comunidade é construindo um histograma de abundância das amostras. As espécies são ordenadas de mais para menos abundante no eixo x e a abundância de cada tipo observado marcada no eixo y.

5.1 Rarefação e estimadores de diversidade

A rarefação compara a riqueza biológica observada entre amostras desiguais de um mesmo habitat. Uma curva de rarefação é calculada a partir do número de espécies observadas em função do número de amostras coletadas [27]. No entanto, essa curva não fornece

uma medida confiável sobre a real riqueza da comunidade, sendo apenas uma forma de descrever a riqueza das amostras coletadas.

Estimadores de diversidade, por sua vez, estimam a riqueza total de uma comunidade a partir de uma amostra e podem ser comparados entre diferentes amostras. Duas abordagens principais são usadas para esse cálculo: estimadores paramétricos e estimadores não-paramétricos.

Métodos paramétricos usam modelos de abundância relativa de espécies e tentam encaixar os dados observados em algum modelo. Dentre os modelos usados, estão os modelos lognormal, Poisson lognormal e a Gaussiana inversa. Por exemplo, Pielou [49] derivou um estimador que supõe que a abundância das espécies é distribuída de acordo com uma lognormal, ou seja, se as espécies são atribuídas a classes de abundâncias logarítmicas, a distribuição das espécies por essas classes é normal. Ao ajustar os dados observados a essa distribuição, os parâmetros da curva podem ser analisados, e o estimador de Pielou usa esses parâmetros para estimar o número de espécies que não foram observados pela amostra e com isso estimar o número total de espécies na comunidade.

A vantagem desses métodos é que, dadas algumas suposições simplificadoras, o modelo pode ser usado para estimativas a partir de amostras relativamente pequenas de indivíduos de um ambiente, o que torna essa abordagem ideal para estimativas de ambientes com grande diversidade [3].

No entanto, existem várias limitações no uso de métodos paramétricos para estimar comunidades microbianas. O primeiro deles é que não existem muitos dados sobre diversidade microbiana para apoiar o uso de vários modelos de abundância existentes. Na falta de dados empíricos, apenas argumentos teóricos podem ser feitos para defender um modelo específico sobre os outros, como foi feito com o modelo lognormal [14]. Ainda assim, existem controvérsias sobre qual modelo é o mais apropriado para representar uma comunidade [33, 76].

Finalmente, mesmo que um modelo específico seja uma boa aproximação da abundância relativa das espécies em uma comunidade, estimadores paramétricos precisam de grandes conjuntos de dados para calcular os parâmetros de distribuição [32].

Métodos não-paramétricos estimam a diversidade sem supor um modelo específico de abundância. Muitos desses estimadores são adaptados de estatísticas de marca-recaptura (do inglês *mark-release-recapture*, ou MRR) para estimar o tamanho de populações animais [67]. Essas abordagens consideram a proporção de OTUs que foram observadas anteriormente (recapturadas) relativa àquelas que foram observadas apenas uma vez. Em uma comunidade muito diversificada, a probabilidade de uma mesma espécie ser observada mais de uma vez é baixa, e a maioria das espécies serão representadas por apenas um indivíduo. Em comunidades menos diversificadas, a probabilidade de uma espécie ser observada mais de uma vez é maior [3].

Estimadores como Chao1 [6] e estimadores baseados em cobertura de abundância (ACE) usam essa proporção de marca-recaptura para calcular a riqueza, adicionando algum fator de correção ao número de espécies observadas. O estimador Chao1 calcula o total de diversidade de espécies como:

$$S_{Chao1} = S_{obs} + \frac{n_1^2}{2n_2}$$

onde S_{obs} é o número de espécies observadas, n_1 é o número de singletons (espécies capturadas apenas uma vez), e n_2 é o número de doubletons (espécies capturadas duas vezes). Esse índice é particularmente útil para conjuntos de dados inclinados a terem muitas classes com baixa abundância, o que é a tendência geral de comunidades microbianas.

O ACE [8] considera dados de todas as espécies com menos de 10 indivíduos, em vez de usar apenas singletons ou doubletons. A diversidade é então estimada como:

$$S_{ACE} = S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{F_1}{C_{ACE}}\gamma^2_{ACE}$$

onde S_{abund} é o número de espécies abundantes (abundância amostrada maior que 10), e S_{rare} é o número de espécies raras (abundância amostrada menor ou igual a 10). $C_{ACE} = 1 - F_i/N_{rare}$ estima a cobertura da amostra, onde F_i é o número de espécies com i indivíduos e $N_{rare} = \sum_{i=1}^{10} iF_i$. Finalmente,

$$\gamma^2_{ACE} = \max\left[\frac{S_{rare}\sum\limits_{i=1}^{10} i(i-1)F_i}{C_{ACE}(N_{rare})(N_{rare}-1)} - 1, 0\right]$$

estima o coeficiente de variação de cada F_i [11].

Chao1 é um estimador particularmente útil por causa de uma fórmula de variância que foi derivada para ele [7]. Essa variância é uma estimativa da precisão de Chao1, ou seja, estima uma variância das estimativas de diversidade caso várias amostras diferentes sejam retiradas de uma mesma comunidade [3]. A variância de S_{Chao1} é calculada como:

$$Var(S_{Chao1}) = n_2\left(\frac{m^4}{4} + m^3 + \frac{m^2}{2}\right)$$

onde $m = \frac{n_1}{n_2}$ e pode ser usada para calcular intervalos de confiança da estimativa, que por sua vez determinam se a diferença de diversidade entre duas amostras é estatisticamente significante. Ainda não foi derivada uma fórmula para a variância do ACE.

Abordagens não-paramétricas também tem suas desvantagens: comparações de diversidade exigem definições claras do que é uma OTU, do inglês, *Operational Taxonomic Unit*. Uma OTU pode se referir a qualquer nível da hierarquia taxonômica, incluindo indivíduos de uma espécie, ou espécies diferentes, ou gêneros diferentes, e assim por diante.

Geralmente uma suposta espécie ou OTU é definida a partir de um percentual de similaridade dentre as sequências do conjunto de amostras, levando a diferentes interpretações do que seria uma espécie ou OTU [70].

A maioria dos métodos não-paramétricos utilizam informações sobre as frequências relativas de OTUs, porém muitos estudos revelam que tendências geradas durante a amostragem podem se propagar para as análises de diversidade. A abundância de genes ampliados por PCR pode não refletir a abundância do DNA original por causa de diferenças nas ligações dos primers e a eficiência de elongamento [53, 55, 71]. Somente quando as unidades de medida são bem definidas e mantidas constantes, comparações usando métodos não-paramétricos podem ser feitas [32].

Outra desvantagem de estimadores não-paramétricos é que eles fornecem apenas um limitante inferior para a diversidade de OTUs quando o número de amostras é baixo. Por exemplo, o valor máximo de S_{Chao1} é $(S_{obs}^2 + 1)/2$, que acontece quando uma espécie na amostra é um doubleton ($n_2 = 1$) e todas as outras são singletons ($n_1 = S_{obs} - 1$). Assim, S_{Chao1} vai se correlacionar com o tamanho da amostra até S_{obs} ser pelo menos do tamanho da raiz quadrada de duas vezes a riqueza total [12].

Métodos não-paramétricos utilizam apenas informações sobre OTUs observadas e não fazem suposições sobre a distribuição de suas abundâncias, como fazem os métodos paramétricos. Isso significa que eles não contam classes de OTUs raras, que podem não ter aparecido nas amostras, e estão ligados ao próprio tamanho da amostra [3].

5.2 Estudos de Caso

Muitos estudos microbiológicos usam estimadores de diversidade para mostrar a grande riqueza de espécies nos mais diversos ambientes.

5.2.1 Microbial diversity in the deep sea and the underexplored rare biosphere [69].

Neste trabalho, Sogin *et al.* analisaram comunidades oceânicas do Atlântico Norte e de um vulcão submarino no nordeste do Pacífico. Oito amostras foram extraídas e feito o sequenciamento de rRNA delas. As sequências obtidas foram então clusterizadas em grupos definidos pela similaridade das sequências. Por exemplo, definir uma clusterização com uma distância genética de 10% significa dizer que todas as sequências pertencentes a um determinado cluster não são mais que 10% distintas entre si. Dessa forma, foram gerados diferentes conjuntos de clusters, definidos como tendo distâncias genéticas desde 0% (sequência única) até 10%. Cada cluster dentro de uma clusterização define uma

OTU, e foram usados para gerar curvas de rarefação e para os cálculos de estimadores de diversidade como ACE e Chao1.

Uma das amostras do vulcão submarino (identificada como FS396) com 6.326 OTUs identificadas a partir de clusters definidos com distância de 3% é estimada pelo ACE como tendo uma diversidade total de 23.315 OTUs, e pelo Chao1 como tendo 20.949 OTUs. As curvas de rarefação, como por exemplo a da Figura 5.1, indicaram que mesmo quando os clusters são definidos com grandes distâncias (10% de distância), amostras adicionais aumentariam significativamente as estimativas de diversidade total da comunidade.

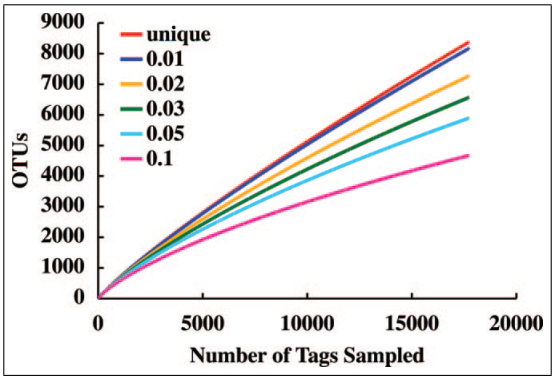


Figura 5.1: Curva de rarefação para a amostra FS396 baseada em distância entre pares. A rarefação é apresentada para OTUs que contem sequências única e para OTUs com diferenças de até 1%, 2%, 3%, 5% ou 10% [69].

5.2.2 Molecular Analysis of Bacterial Community Structure and Diversity in Unimproved and Improved Upland Grass Pastures [42].

Neste trabalho o ambiente estudado é o solo. McCaig *et al.* coletaram amostras de duas pastagens, uma previamente fertilizada e a outra não, com o intuito de comparar a riqueza de espécies desses ambientes. O DNA ribossomal (16S rDNA) de bactérias foi ampliado por PCR e clonado, sequenciando um total de 137 clones do solo fertilizado e 138 do não-fertilizado. Pela definição de OTU utilizada por eles, de pelo menos 97% de similaridade, foram identificadas 113 OTUs no habitat fertilizado e 117 no não-fertilizado. O estimador Chao1 estabiliza em ambos habitats quando o número de clones analisados alcança aproximadamente 70, isto é, a partir de amostras com mais de 70 clones, o resultado dos estimadores não apresenta mudanças significativas. A riqueza estimada aparenta ser maior no solo não-fertilizado, um total de 590 OTUs contra 467 no solo fertilizado, mas a diferença não é significativa.

Estimadores de diversidade também diminuem significativamente o tamanho da amostra necessária para se ter uma boa estimativa da riqueza de uma comunidade.

5.2.3 Toward a Census of Bacteria in Soil [65].

Schloss e Handelsman analisaram o solo do Alaska. A partir de uma biblioteca de 1.033 sequências de rRNA 16S, eles usaram uma clusterização baseada em distância entre pares para atribuir OTUs às sequências. Cada grupo representa uma OTU e é definido com sequências com pelo menos 97% de similaridade, representando o nível de espécie. Com esse método, eles observaram 633 OTUs, sendo que o filo mais abundante era representado por 23 sequências, e o segundo mais abundante por 17 sequências.

Eles não tiveram sucesso em obter um modelo paramétrico de distribuição de frequências que se ajustasse bem com os dados observados, mas usando um modelo de distribuição de frequências lognormal, eles heurísticamente identificaram uma média para a normal, um desvio padrão e uma riqueza de OTU para uma distribuição simulada cuja distribuição de amostras se assemelhava à distribuição observada com os dados coletados do solo do Alaska. Outras métricas foram analisadas para confirmar a plausibilidade dessa comunidade simulada, como por exemplo a porcentagem do total da comunidade representada pela OTU mais abundante (2.9% na comunidade simulada, contra 2.2% na observada). O total de OTUs para essa comunidade simulada foi de 5.000 OTUs, definidas com 97% de similaridade.

Com o conhecimento da riqueza total dessa comunidade simulada, foi possível calcular que para se amostrar todas as OTUs dessa comunidade pelo menos duas vezes com 95% de confiança, seriam necessários 480.000 fragmentos de rRNA 16S. Para se amostrar 95% da riqueza total, seriam necessários 71.000 fragmentos de rRNA16S. No entanto, partindo de amostras bem menores, de 18.000 e 39.000 fragmentos de rRNA 16S, os estimadores ACE e Chao1, respectivamente, incluem o valor de 5 mil OTUs nos intervalos de confiança de suas estimativas para a riqueza total da comunidade. Com 18.000 fragmentos de rRNA 16S, são efetivamente amostradas 65% da riqueza total, mostrando que esses estimadores conseguem bons resultados trabalhando em cima de aproximadamente dois terços da riqueza total de uma comunidade. O Chao1 alcançou a estimativa de 5 mil OTUs com os 39.000 fragmentos de rRNA 16S, que efetivamente representa 85% da riqueza total.

Muitos estudos comparam estimadores em busca de uma melhor representação da diversidade do ambiente. Métodos paramétricos também foram considerados e mostraram números mais elevados para a estimativa da riqueza de um ambiente.

5.2.4 Predicting microbial species richness [28].

Hong *et al.* analisaram amostras de um banco de areia da Baía de Massachussets. As sequências de rRNA 16S foram ampliadas e clonadas, gerando 556 sequências, e depois agrupadas em OTUs definidas por valores de 50%, 60%, 70%, 80%, 90%, 95%, 96%, 97%, 98% e 99% de similaridade. Seis modelos paramétricos foram testados para descrever a distribuição de probabilidades das frequências das OTUs: Poisson, gamma, Gaussiana inversa, lognormal, Pareto e o misto de 2 exponenciais. Esses modelos foram ajustados aos dados observados pelo método ML (maximum likelihood).

Geralmente nenhum modelo paramétrico atual se ajusta a um conjunto de dados completamente, então eles separaram seus dados em conjuntos de espécies 'raras' e 'abundantes', ou seja, todas as espécies que tinham abundância menor que um determinado ponto de truncamento eram consideradas 'raras' e as que estavam acima desse ponto, 'abundantes'. Todos os modelos foram ajustados a todas as possíveis coleções de frequências 'raras', e calculados o ML-SE (sequence variation estimation), que são dois métodos para calcular os parâmetros do modelo que mais se ajustam aos dados observados. Finalmente foram selecionados os melhores parâmetros, considerando a menor SE, o maior ponto de truncamento, e o melhor ajuste dado pelo ML.

Os valores dos estimadores não-paramétricos ACE e ACE1 também foram usados na comparação. ACE1 é uma modificação do ACE para comunidades muito diversificadas [8]. Foram calculados os valores desses estimadores e suas respectivas SEs para a coleção de frequências 'raras' correspondentes aos pontos de truncamento escolhidos como melhores ajustes dos modelos paramétricos. Como os valores de ML não se aplicam a estimadores não-paramétricos, foi selecionado o 'melhor' estimador baseado na literatura.

Com isso, os resultados apontam que o modelo lognormal foi melhor quando o número de OTUs foi máximo, no agrupamento definido com 99% de similaridade. No agrupamento definido com 98% e 97%, a distribuição Pareto forneceu uma melhor estimativa, e para os agrupamentos com 96%, 95% e 90% de similaridade, o modelo exponencial misto mostrou resultados melhores. Os estimadores ACE e ACE1 mostraram valores entre 10% e 50% abaixo dos métodos paramétricos. No agrupamento com 97% de similaridade, por exemplo, com 380 OTUs detectadas, o melhor modelo paramétrico estimou uma riqueza de 2.434 OTUs, enquanto o ACE estimou 1.385 e o ACE1 estimou 2.296.

5.2.5 The rational exploration of microbial diversity [50].

Quince *et al.* compararam estimadores paramétricos com o Chao e chegaram a resultados similares. Eles utilizaram dois conjuntos de sequências de rDNA 16S, um de águas oceânicas profundas do trabalho de Huber *et al.* [30], e outro de solos derivados de quatro locais diferentes, estudado por Roesch *et al.* [58]. OTUs foram definidas como conjuntos

de sequências com pelo menos 97% de similaridade. O estimador Chao forneceu valores entre 27% e 97% abaixo dos obtidos por métodos paramétricos. Na amostra de bactérias FS396 do conjunto de águas oceânicas, com 5.853 OTUs observadas, o Chao estimou uma diversidade de 10.569 OTUs e o modelo lognormal estimou 306.589 OTUs.

Capítulo 6

Análise de diversidade das amostras de compostagem do Zoológico de São Paulo

A Fundação Parque Zoológico de São Paulo (FPZSP) possui uma unidade de compostagem que aproveita matéria orgânica de várias origens, como excrementos de animais, folhas e galhos de árvores e restos de alimentos. Esse material, atualmente utilizado como fertilizador de áreas agrícolas do próprio zoológico, tem uma grande riqueza microbiológica, podendo conter várias espécies ainda não descritas.

Estima-se que a maior parte dos microrganismos que habitam esse material não seja cultivável, devido às diversas condições naturais desconhecidas, impossibilitando a reprodução do habitat em laboratório, e também pela dificuldade de se isolar organismos que dependem de outros para sobreviver. Assim, a abordagem metagenômica é a mais indicada para esse estudo.

Para explorar a diversidade das amostras extraídas, utilizamos as sequências obtidas com o método de pirosequenciamento 454 e conduzimos vários experimentos com elas.

As análises foram feitas em dois conjuntos independentes de amostras, Zoo Compost 1 (ZC1) e Zoo Compost 2 (ZC2), retiradas de estágios diferentes do processo de compostagem. As duas amostras de DNA foram submetidas a um pirosequenciamento, seguindo protocolos padrões da Roche 454 GS FLX Titanium (Roche Applied Science). Cada amostra passou por quatro corridas de sequenciamento e os reads submetidos a um filtro de qualidade e montados pelo software 454 Newbler assembler, versão 2.5.3. A Tabela 6.1 mostra um conjunto de métricas calculadas pelo Newbler para as duas amostras.

Os dados das amostras ZC1 e ZC2 de metagenômica da compostagem do parque zoológico de São Paulo foram obtidos graças ao projeto "Estabelecimento de um laboratório de Microbiologia Aplicada no Zoológico de São Paulo: Identificação e isolamento

de microrganismos que produzam enzimas e seus inibidores de aplicação nas áreas médica, veterinária e industrial", coordenado pelo profesor Luiz Juliano Neto (UNIFESP), e financiado pela FAPESP (processo número 09/52030-5).

A compostagem do Zoológico de São Paulo é supervisionada pelo Dr. João Batista Cruz, diretor científico da Fundação Parque Zoológico de São Paulo. A coleta das amostras e a extração do DNA metagenômico foi realizada por Luciana Principal Antunes, na época aluna de iniciação científica da professora Renata Pascon, da UNIFESP. O sequenciamento foi realizado na máquina Roche GS/FLX do Centro Avançado de Tecnologia Genômica (CATG) do Departamento de Bioquímica do Instituto de Química da USP.

O sequenciamento propriamente dito foi realizado pela Dra. Layla Farage Martins, pesquisadora do CATG sob a supervisão dos professores Aline Maria da Silva e Sergio Verjovski Almeida (coordenador do CATG), ambos do Departamento de Bioquímica do Instituto de Química da USP. As sequências obtidas foram imediatamente carregadas na plataforma MG-RAST, limitando-se o acesso exclusivamente aos participantes do projeto.

O projeto do professor Luiz Juliano foi seguido em 2012 de um segundo projeto, ainda vigente, intitulado "Estudos da diversidade microbiana do Parque Zoológico de São Paulo", coordenado pelo professor João Carlos Setubal do Departamento de Bioquímica do Instituto de Química da USP, e financiado pela FAPESP (processo número 2011/50870-6).

Outros pesquisadores que ajudaram na disponibilização dos dados são o professor Luciano Digiampietri (EACH/USP) e o professor Julio Cezar Franco de Oliveira (UNIFESP). Todos os professores e pesquisadores mencionados foram ou são colaboradores dos dois projetos.

Os resultados das análises das amostras ZC1 e ZC2 realizados pelas equipes dos dois projetos foram publicados em abril de 2013 [40].

Parâmetro	Zoo Compost 1	Zoo Compost 2
Número total de reads	3,167,044	2,966,244
Tamanho médio do read	276 nt	299 nt
Tamanho do Metagenoma (reads não-montados)	836 Mbp	842 Mbp
Tamanho do Metagenoma (reads montados)	506.0 Mbp	433.7 Mbp
Número de reads em contigs	1,178,578 (37.2%)	1,448,502 (48.8%)
Número de contigs	52,953	52,182
Reads/contig	22.26	27.76
Maior contig (bp)	39,861	65,988
Tamanho médio do contig (bp)	1,384	1,332
Número de singletons	1,842,944	1,404,679

Tabela 6.1: Métricas calculadas pelo Newbler para os conjuntos de amostras sequenciadas ZC1 e ZC2.

6.1 MG-RAST

Em uma abordagem preliminar, foram utilizados os serviços do MG-RAST [44]. Com ele, pudemos ter uma primeira versão da análise dos dados. Os dados do MG-RAST aqui apresentados foram obtidos de uma análise feita com o M5NR (M5 non-redundant protein database), que é uma composição de vários bancos de dados. M5 é um projeto de iniciativa do Genomics Standards Consortium e visa soluções para as necessidades computacionais da comunidade científica de metagenômica [23]. O nome M5 representa "Metagenômica, Metadados, Meta-análises, Modelos e Meta-infraestrutura". O M5NR é composto por bancos de dados de RNA, de proteínas e ontologias, dentre eles: 16S rRNA Greengenes Database, SILVA rRNA Database Project, Ribossomal Database Project (RDP), NCBI GenBank, Joint Genome Institute Integrated Microbial Genomes (JGI/IMG), Kyoto Encyclopedia of Genes and Genomes (KEGG), Uniprot Knowledgebase SwissProt e TrEMBL, Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups (eggNOG) e o SEED Project Database.

As Figuras 6.1 e 6.2 mostram os gêneros mais abundantes nas amostras ZC1 e ZC2, ordenadas do mais abundante para o menos abundante. Só os 20 mais abundantes são mostrados, juntamente com a porcentagem de abundância relativa de cada um. No conjunto ZC1, os 20 gêneros mais abundantes representam 50% das sequências identificadas. No conjunto ZC2, o gênero *Lactobacillus* corresponde sozinho a 33% de todas as sequências identificadas. O MG-RAST classificou 2.668.964 de sequências em ZC1 e 2.786.392 de sequências em ZC2.

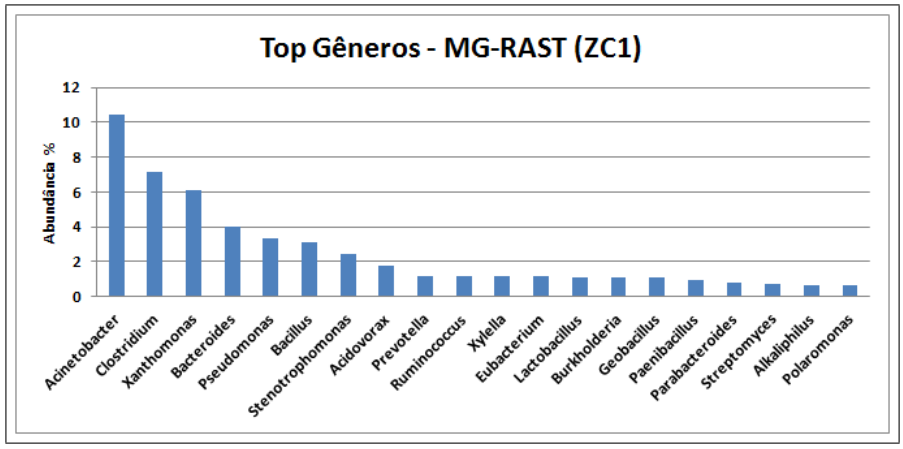


Figura 6.1: Histograma de abundância de gêneros das sequências da amostra ZC1 obtidas pelo MG-RAST.

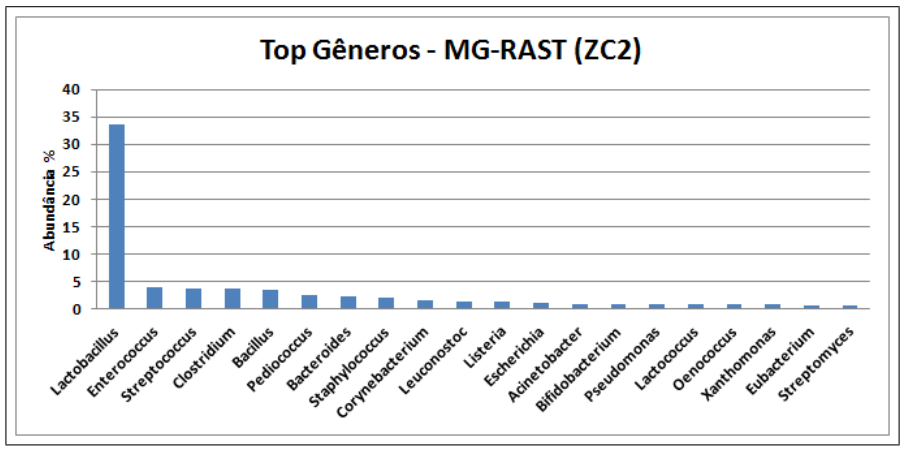


Figura 6.2: Histograma de abundância de gêneros das sequências da amostra ZC2 obtidas pelo MG-RAST.

A Figura 6.3 mostra a curva de rarefação das espécies anotadas. Esse é um gráfico do número total de espécies distintas em função do número de sequências amostradas. As primeiras amostras costumam conter muitas espécies diferentes, e conforme o número de amostras cresce, o número de novas espécies descobertas diminui, por isso geralmente as curvas de rarefação sobem rapidamente no início e tendem a se estabilizar quando o número de amostras cresce. Essas curvas são calculadas a partir da Tabela de abundância de espécies e representam o número médio de espécies distintas para cada sub-amostra do conjunto completo de dados.

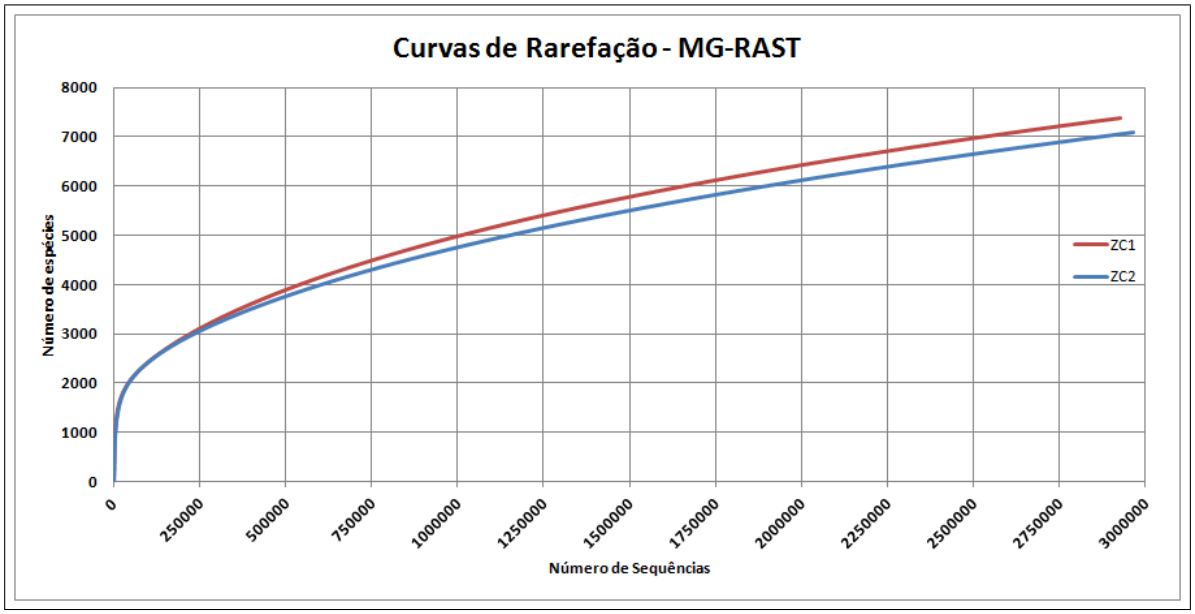


Figura 6.3: Curvas de rarefação obtidas pelo MG-RAST.

6.2 Naive Bayes Classifier (NBC)

Uma alternativa ao método do MG-RAST é utilizar os programas discutidos no Capítulo 4. Em específico, testamos o classificador Bayesiano com os dados do Zoológico. Para isso, utilizamos uma máquina Intel® Core™ i7-2600K, 3.40GHz, com 8 núcleos e 16GB de memória RAM, rodando por aproximadamente 250 horas. Antes de rodar com os dados do Zoológico, no entanto, o classificador precisa ser treinado com um banco de dados de referência. Treinamos com o banco de dados padrão sugerido pelos desenvolvedores, o de genomas completos de bactérias do NCBI, que na versão utilizada, contava com 2262 genomas.

O programa compara cada sequência de entrada com o banco de dados que foi treinado e dá uma pontuação para cada um dos alinhamentos, dando uma pontuação de 'infinito' para os melhores alinhamentos, mas o programa não especifica as restrições para pontuar um alinhamento como tal. Os resultados são organizados de acordo com cada cepa de espécie utilizada no treinamento, e quais as pontuações dos alinhamentos daquela cepa com cada uma das sequências de entrada. Em sua maioria, um genoma recebeu uma pontuação 'infinita' com diversas sequências de entrada. Além disso, muitas sequências tiveram essa pontuação com vários genomas distintos.

Para simplificar a análise, a quantidade de sequências que obtiveram pontuação 'infinita' para cada genoma foram somadas, e esse número é uma representação da abundância daquela espécie na amostra. Muitas sequências da entrada, no entanto, obtiveram pon-

tuações 'infinitas' com todas as sequências pertencentes a um mesmo gênero. Para evitar redundâncias, a espécie mais abundante dentro de cada gênero foi selecionada e a sua abundância passou a representar a abundância do gênero correspondente.

As figuras 6.4 e 6.5 mostram os gêneros mais abundantes das amostras ZC1 e ZC2, respectivamente. O gênero mais abundante de ZC1 conta com 3.144.972 sequências, que são quase todas as 3.167.069 sequências da amostra, mas ao mesmo tempo, isso corresponde a apenas 0.53% de todas as sequências identificadas. De forma semelhante, o gênero mais abundante de ZC2 conta com 2.938.090 de um total de 2.966.260 sequências, mas corresponde a apenas 0.48% das sequências identificadas. No total, o NBC identificou 590.612.405 sequências em ZC1, e 640.500.036 sequências em ZC2, números muito acima das 3 milhões de sequências originais. Isso indica que mesmo entre diferentes gêneros, existe uma grande intersecção de sequências.

Existe uma grande diferença entre a ordenação dos gêneros mais abundantes identificados pelo NBC com relação ao MG-RAST. A grande diferença, tanto do método quanto do banco de dados de referência, pode explicar essa disparidade. Mesmo que poucos gêneros mais abundantes ainda sejam comuns, como *Bacteroides* no caso de ZC1 e *Lactobacillus* no caso de ZC2, a distribuição é muito diferente, em especial com o gênero *Lactobacillus*, uma vez que 33% das sequências de ZC2 foram atribuídas a esse gênero pelo MG-RAST.

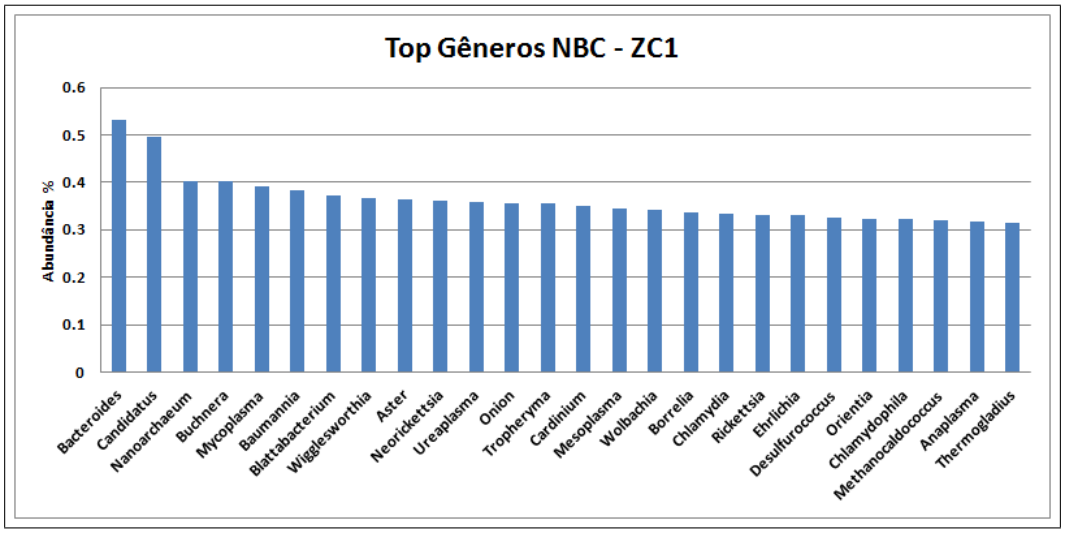


Figura 6.4: Gêneros mais abundantes da amostra ZC1 identificados pelo NBC.

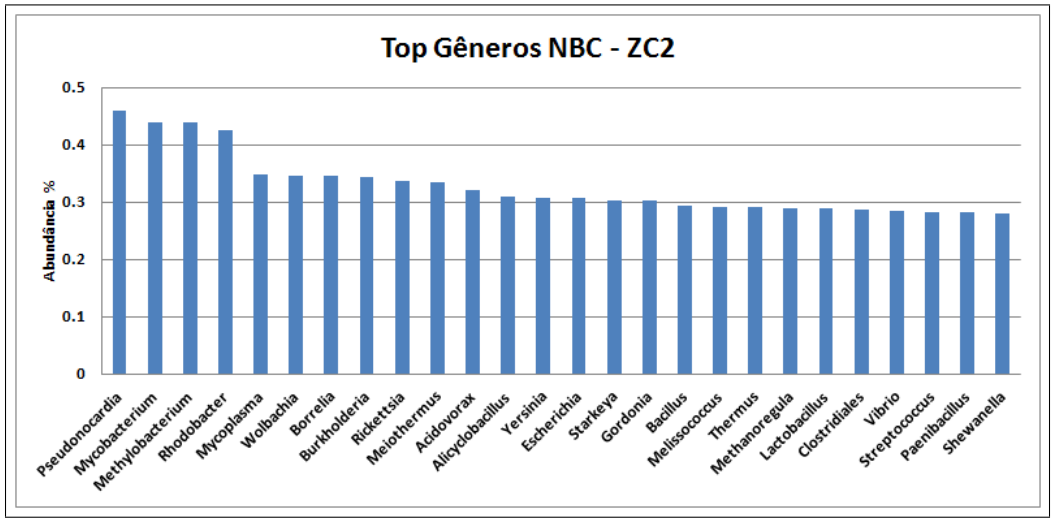


Figura 6.5: Gêneros mais abundantes da amostra ZC2 identificados pelo NBC.

O grande número de sequências com pontuações 'infinitas' pode ter atrapalhado o resultado do NBC. Os alinhamentos são feitos entre sequências geralmente curtas contra genomas completos, o que pode explicar a grande quantidade de sequências de entrada que tenham uma boa pontuação com cada um dos genomas do banco de dados.

6.3 Outros programas

Além do NBC, outros programas citados no Capítulo 4 foram testados, mas sem sucesso. O programa PhymmBL apresentou diversos erros de compilação no nosso ambiente, e como os próprios desenvolvedores afirmam que a instalação é bem demorada, decidimos abandonar os testes com esse programa.

O programa CARMA também foi testado, mas erros de execução impediram a continuação dos testes, investigamos se o problema era do nosso ambiente, dos nossos dados ou de alguma incompatibilidade, mas não chegamos a uma conclusão e decidimos abandonar os testes com esse programa também.

O PhylOTU apresentou incompatibilidade com a versão do BLAST que utilizamos na comparação dos bancos de dados. Até o momento de conclusão deste trabalho, o programa não foi atualizado.

O SOrt-ITEMS forneceu resultados intermediários, e em nosso ambiente não foi possível completar a análise através dele.

Diante desses problemas, resolvemos retomar uma análise própria e independente de programa prontos específicos para classificação de metagenomas.

6.4 Ribosomal Database Project (RDP)

Para nos aprofundarmos com relação às análises feitas pelo MG-RAST, fizemos uma mais detalhada, utilizando a base de dados do Ribosomal Database Project (RDP) [10], release 10, update 27, que contém sequências anotadas de rRNA.

Para comparar esse banco de dados com as sequências do Zoológico, utilizamos o Infernal [47]. Essa ferramenta monta modelos de covariância para representar consensos de estruturas secundárias de RNA e os utiliza para buscar RNAs homólogos em outros bancos de dados. É possível treinar o Infernal para construir os modelos de covariância a partir de qualquer banco de dados, o que fizemos com o do RDP. Com esses modelos treinados, buscamos RNAs homólogos nas sequências do Zoológico e classificamos os resultados com o próprio classificador do RDP, versão 2.4.

O classificador utiliza limites de confiança de 50%, 60%, 70%, 80%, 90% e 95% para atribuir um determinado taxon a uma sequência de busca, por isso nem todas as sequências foram classificadas até o nível de gênero. A Tabela 6.2 indica quantas sequências/indivíduos foram classificadas até no máximo cada nível taxônomico para as amostras ZC1 e ZC2, com limite de confiança de 95%, bem como quantas OTUs foram identificadas em cada nível taxonômico. As Figuras 6.6 e 6.7 mostram histogramas de frequências de sequências que foram classificadas no nível de gênero com 95% de confiança, tanto para as amostras de ZC1 quanto para ZC2. Em ZC1, o RDP classificou 1569 sequências, e o ZC2 classificou 3540.

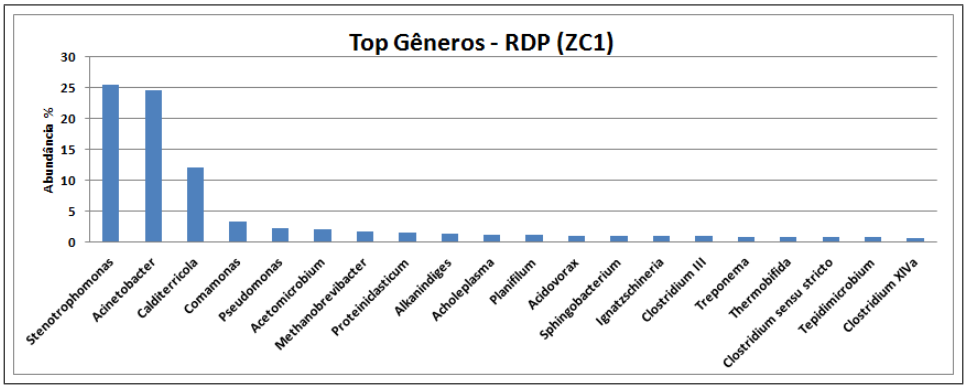


Figura 6.6: Histograma de abundância de gêneros das sequências da amostra ZC1 encontradas pelo Infernal contra o banco de Dados do RDP, extraídos da classificação com um limite de confiança de 95%.

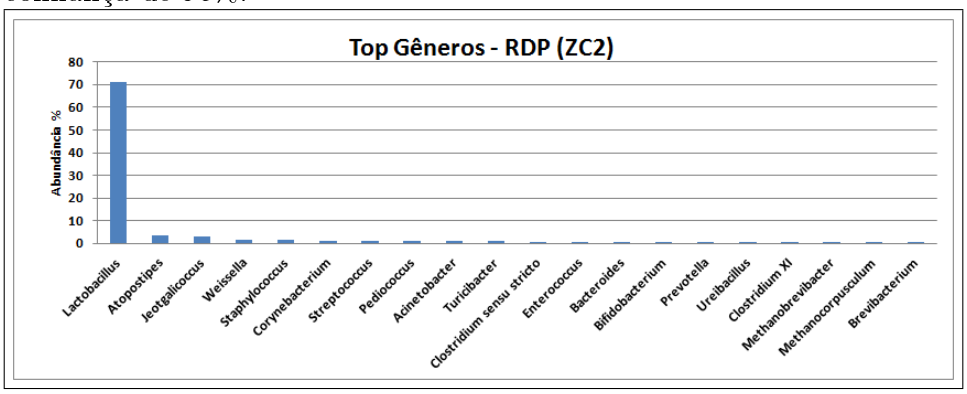


Figura 6.7: Histograma de abundância de gêneros das sequências da amostra ZC2 encontradas pelo Infernal contra o banco de Dados do RDP, extraídos da classificação com um limite de confiança de 95%.

	Raiz	Domínio	Filo	Classe	Ordem	Família	Gênero
ZC1							
OTUs	-	2	17	120	48	81	109
Indivíduos	265	1675	596	330	456	1292	1569
ZC2							
OTUs	-	2	14	118	46	83	138
Indivíduos	347	1198	423	439	877	798	3540

Tabela 6.2: Número de OTUs identificadas em cada nível taxonômico e número de sequências não classificadas além do táxon correspondente

Para gerar os gráficos de rarefação e os estimadores de diversidade, foi utilizado o mothur [66] versão 1.23.0. O mothur é uma ferramenta que calcula distância entre pares

de sequências e gera clusters baseados em limitantes para essa distância. Por exemplo, se limitarmos a distância a 0.05, uma sequência só fará parte de um cluster caso a distância dela para qualquer outra dentro desse cluster seja menor ou igual a 0.05, ou seja, todas as sequências no cluster possuem 95% de similaridade entre si, e cada cluster é considerado uma OTU.

Com esses clusters, é possível gerar curvas de rarefação e calcular estimadores de diversidade, criando sub-amostras aleatórias em cima dos clusters encontrados. As Figuras 6.8 e 6.9 mostram curvas de rarefação feitas considerando distâncias de 0.0 (Unique), caso onde cada cluster (OTU) possui apenas uma sequência, distâncias de 0.03, 0.05 e 0.10. a Tabela 6.3 mostra os resultados de dois estimadores de diversidade não-paramétricos, ACE e Chao1, para os clusters definidos pelas distâncias de 0.03, 0.05 e 0.10.

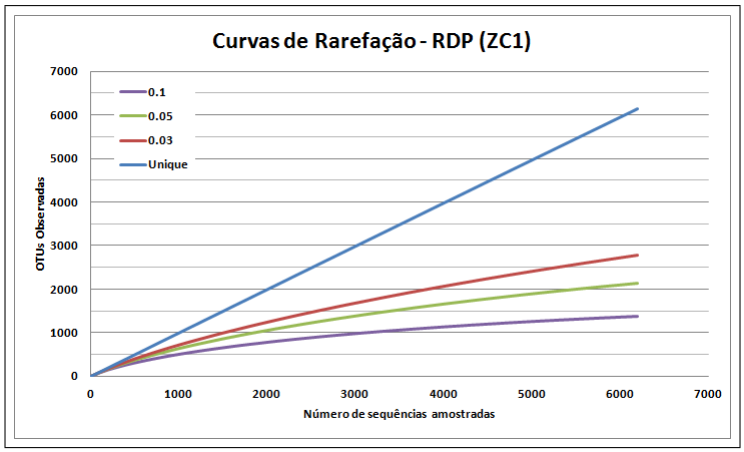


Figura 6.8: Curvas de rarefação da amostra ZC1.

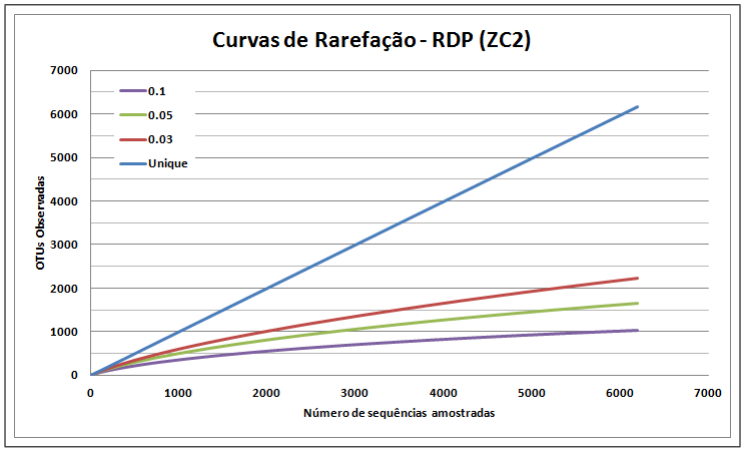


Figura 6.9: Curvas de rarefação da amostra ZC2.

	Distância definida para o Cluster											
	unique			0.03			0.05			0.10		
	OTUs	ACE	Chao1	OTUs	ACE	Chao1	OTUs	ACE	Chao1	OTUs	ACE	Chao1
ZC1	6133	327322	318841	2774	10377	6277	2129	3956	3606	1367	1955	1824
ZC2	7553	405521	396253	2559	11695	6408	1856	5264	3351	1133	2161	1680

Tabela 6.3: OTUs baseadas em similaridade e estimadores de diversidade de espécies.

6.5 Swiss-prot

Complementando as análises feitas com o banco de dados do RDP, analisamos as amostras do zoológico de um ponto de vista mais abrangente, comparando-as com bancos de dados que não fossem exclusivamente de rRNA. Um desses bancos é o Swiss-prot [13]. Nossas análises foram feitas com a versão 2012_03 do Swiss-prot.

Para gerar os gráficos de abundância, consideramos o primeiro hit do BLAST de cada sequência com um alinhamento de pelo menos 40 bases e 95% de identidade. As Figuras 6.10 e 6.11 mostram os gêneros mais abundantes encontrados nas amostras ZC1 e ZC2 respectivamente. O Swiss-prot classificou 32.132 sequências em ZC1 e 50.534 sequências em ZC2.

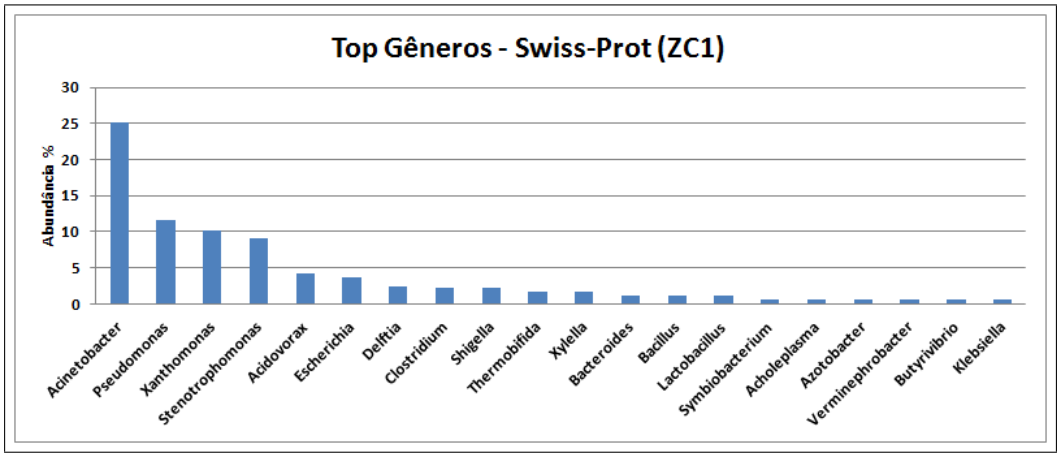


Figura 6.10: Gêneros mais abundantes da amostra ZC1 contra o banco Swiss-prot.

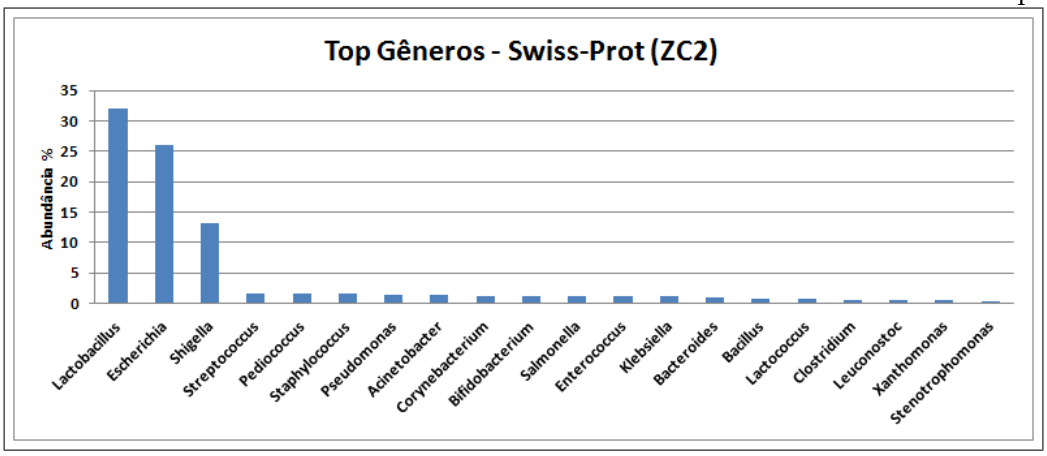


Figura 6.11: Gêneros mais abundantes da amostra ZC2 contra o banco Swiss-prot.

Para gerar as curvas de rarefação e calcular os estimadores de diversidade, fizemos uma série de sub-amostragens aleatórias. Tomando como base os primeiros hits do BLAST utilizados para gerar os histogramas de abundância, fizemos dois conjuntos de sub-amostragens aleatórias: uma deles com 1000 sub-amostras de 300 hits aleatórios cada, e outro com 500 sub-amostras de 300 hits aleatórios cada. Com isso, geramos uma Tabela de entrada para a ferramenta EstimateS [11], versão 8.2.0. O EstimateS calcula várias estatísticas de diversidade, bem como pontos da curva acumulada do número de organismos identificados por amostra, que foram utilizados para gerar as curvas de rarefação da Figura 6.13.

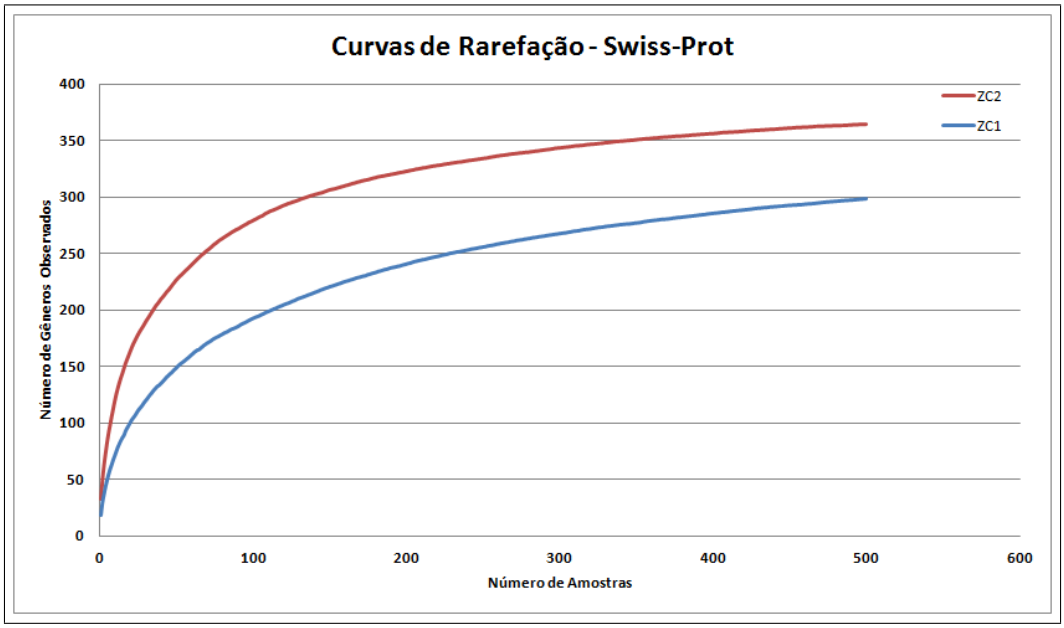


Figura 6.12: Curvas de rarefação do conjunto de 500 subamostras de ZC1 e ZC2 com os resultados do BLAST contra o banco de dados Swiss-prot.

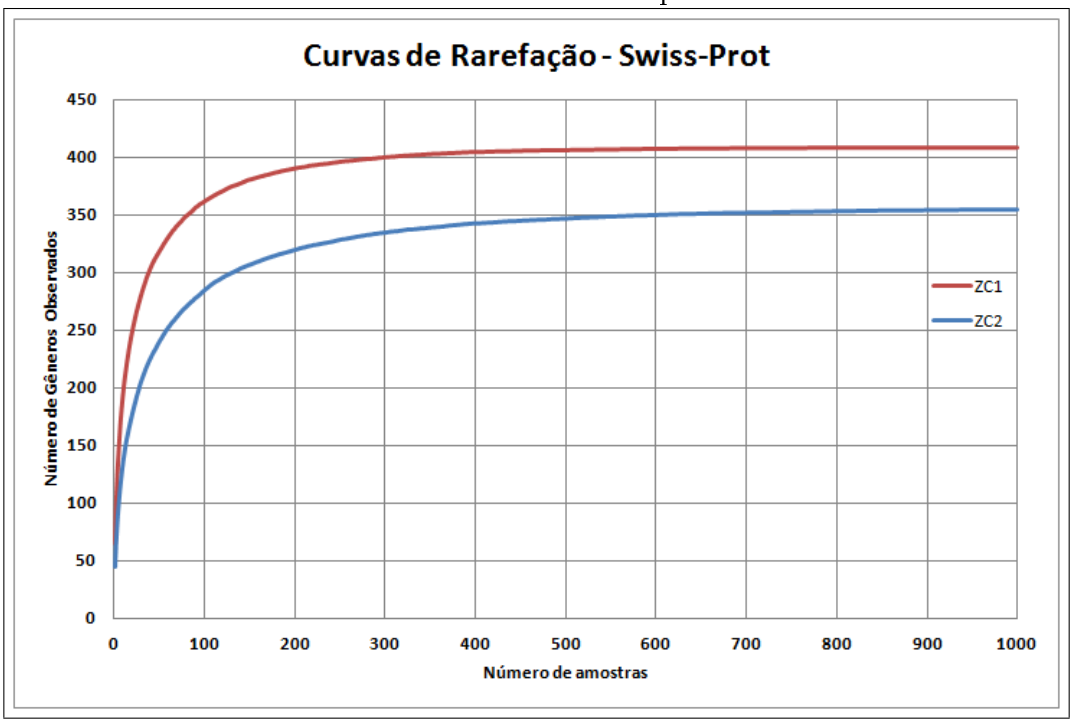


Figura 6.13: Curvas de rarefação do conjunto de 1000 subamostras de ZC1 e ZC2 com os resultados do BLAST contra o banco de dados Swiss-prot.

É possível perceber que as curvas com 1000 amostras se aproximam de uma reta

horizontal, estabilizando no número de organismos identificados pelo Swiss-prot.

O número de OTUs identificadas com essa análise contra o banco de dados Swiss-prot foi consideravelmente menor que os identificados contra o banco de dados do RDP por termos considerados apenas os hits do BLAST com pelo menos 95% de identidade. O classificador do RDP permite contar uma OTU com uma porcentagem menor de identidade com um determinado gênero, mas ainda classificá-lo como integrante de um nível taxonômico mais elevado. Essa diferença na abordagem também reflete nos cálculos feitos com os estimadores de diversidade, mostrados na Tabela 6.4, que inclui também os números feitos quando foram considerados os hits com pelo menos 90% de identidade. Ainda assim, considerando o limite de 95% de identidade, o BLAST com o banco do Swiss-Prot identificou 409 e 357 gêneros distintos para o ZC1 e ZC2 respectivamente, contra 109 e 138 identificados pelo RDP, como mostrado na Tabela 6.2.

	Porcentagem de identidade							
	90%				95%			
Amostra	Hits	OTUs	ACE	Chao1	Hits	OTUs	ACE	Chao1
ZC1	69841	545	636	639	32132	409	475	477
ZC2	83548	459	562	565	50534	357	473	478

Tabela 6.4: OTUS identificadas pelo banco de dados Uniprot/Swiss-prot com o corte na porcentagem de identidade e estimadores de diversidade de espécies correspondentes.

6.6 NCBI-NR

Além do banco Uniprot/Swiss-prot, fizemos um BLAST contra o banco de dados NR. Devido ao grande número de sequências que temos, aproximadamente 3 milhões em cada amostra e ao alto tempo de execução do BLAST, não foi possível processar os dados com todo o NR. Com quatro máquinas Intel® Core™ 2 Duo, 2.13GHz, rodando em paralelo, levou 40 dias para processar 10% das sequências. Portanto, utilizamos o resultado do BLAST com apenas esses 10% das sequências amostradas pelo projeto do Zoológico e fizemos os mesmos procedimentos das análises realizadas em cima dos resultados com o banco do Uniprot.

As Figuras 6.14 e 6.15 mostram os 20 gêneros mais abundantes identificados pelo NR, considerando 95% de identidade como critério mínimo para identificação de um organismo. Comparando com o NR, foram classificadas 7.718 sequências em ZC1 e 20.380 sequências em ZC2.

A Figura 6.16 mostra as curvas de rarefação geradas para ZC1 e ZC2. A Tabela 6.5 mostra os valores dos estimadores de diversidade não-paramétricos ACE e Chao1 para os organismos identificados pelo NCBI-NR considerando pelo menos 90% e 95% de identidade.

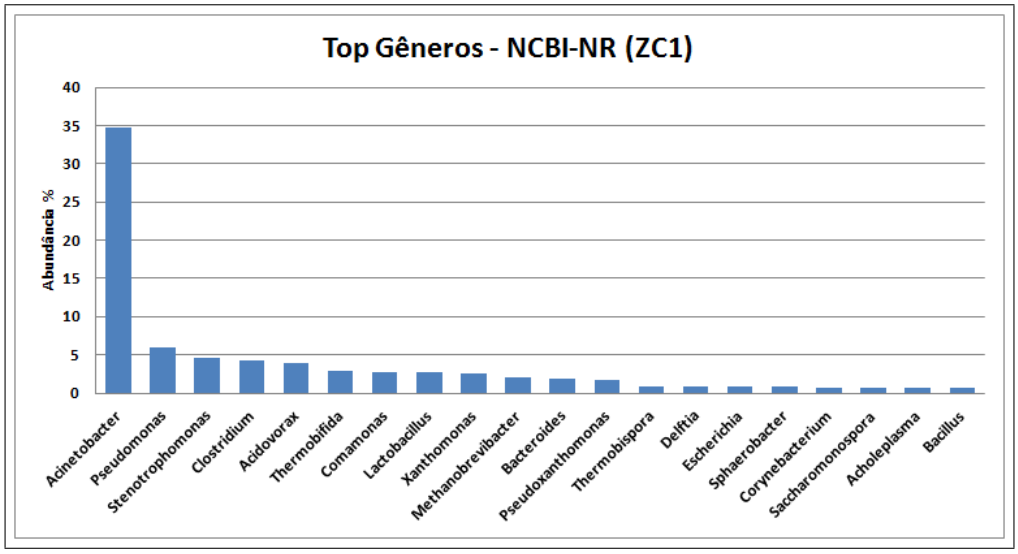


Figura 6.14: Gêneros mais abundantes da amostra ZC1 contra 10% do banco NR.

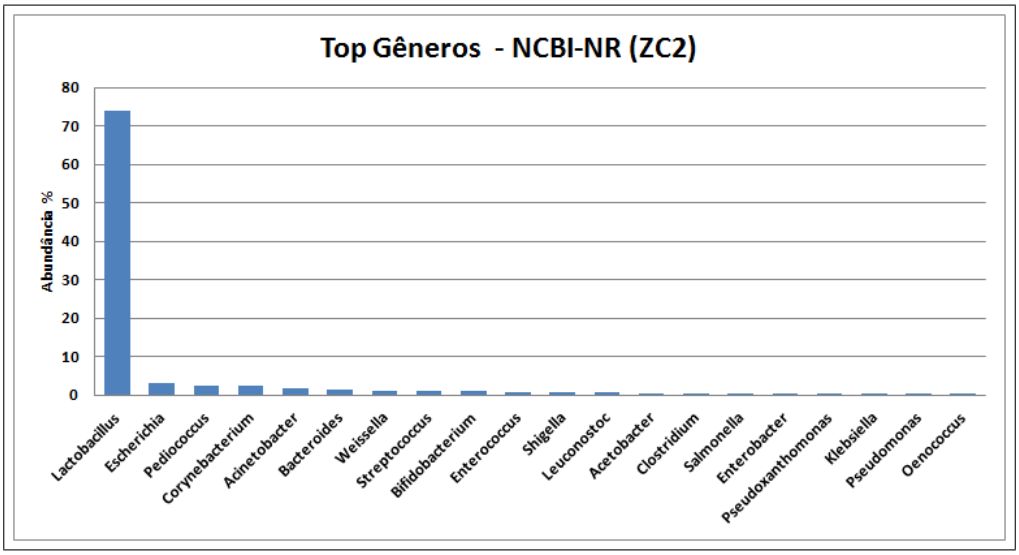


Figura 6.15: Gêneros mais abundantes da amostra ZC2 contra 10% do banco NR.

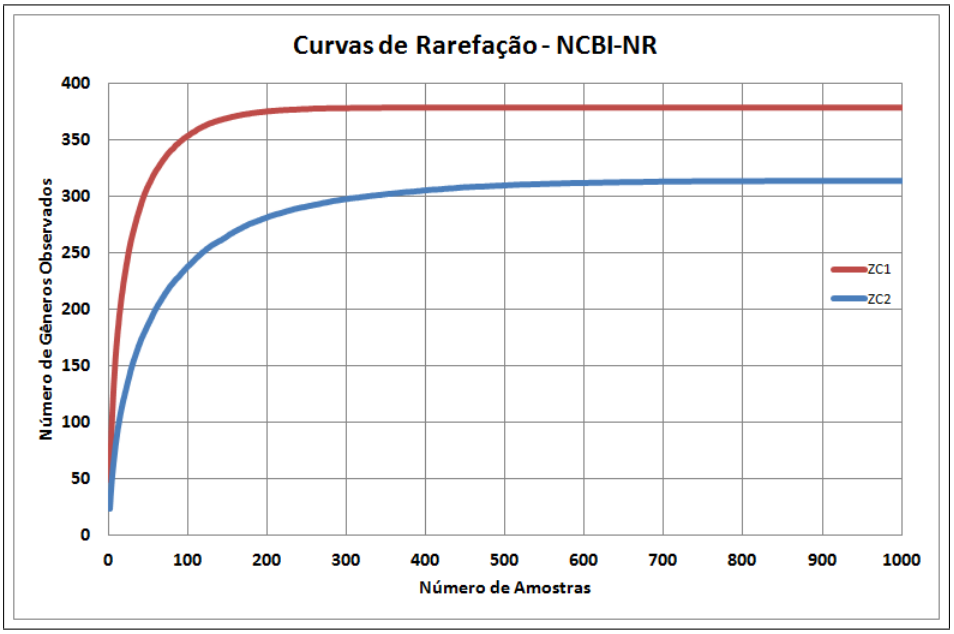


Figura 6.16: Curvas de rarefação gerada a partir de 1000 subamostras de ZC1 e ZC2 com os resultados do BLAST contra o banco de dados NR.

	Porcentagem de identidade							
	90%				95%			
Amostra	Hits	OTUs	ACE	Chao1	Hits	OTUs	ACE	Chao1
ZC1	15520	519	670	673	8190	379	508	511
ZC2	31579	429	583	587	21654	314	452	457

Tabela 6.5: OTUS identificadas pelo banco de dados NCBI-NR com o corte na porcentagem de identidade e estimadores de diversidade de espécies correspondentes.

Com 10% das sequências analisadas, considerando o limite de 95% de similaridade, o NCBI-NR identificou, respectivamente em ZC1 e ZC2, 25% e 43% do número total de hits identificados pelo Swiss-prot, mas com 93% e 88% das OTUs identificadas com os resultados do Swiss-prot.

Em comparação com os resultados do MG-RAST, considerando o ZC1 com 90% de similaridade, com os resultados do NCBI-NR foram identificados 56 gêneros não-identificados pelo MG-RAST. Ao se considerar 95% de similaridade, o número de gêneros identificados pelo NCBI-NR mas não pelo MG-RAST passa para 41. No caso de ZC2, com 90% de similaridade, o NCBI-NR identificou 38 gêneros que o MG-RAST não identificou e ao se considerar 95% de similaridade esse número passa para 22.

6.7 MetaSim - Simulador de dados metagenômicos

Para auxiliar na comparação dos resultados, foi utilizado um conjunto de dados simulados, baseados na distribuição do metagenoma de rumen de vaca descrito por Brulc *et al.* [5]. Para gerar os dados, utilizamos o software MetaSim [56]. O MetaSim gera sequências simuladas de metagenoma a partir de um conjunto de espécies pré-definidas pelo usuário utilizando modelos de distribuição e erros de diversas tecnologias de sequenciamento, inclusive de pirosequenciamento 454, que foi a utilizada nas sequências do Zoológico.

A partir da análise de abundância de gêneros do trabalho de Brulc *et al.* feita pelo MG-RAST, selecionamos os 50 gêneros mais abundantes, escolhemos aleatoriamente um genoma completo de uma espécie de cada gênero e demos como entrada para o MetaSim. A Tabela 6.6 mostra o número de reads e de qual espécie foi derivado o genoma de cada gênero utilizado na simulação, bem como o número de identificação no GenBank, o número do Taxonomy ID do NCBI, o tamanho do genoma (todos eles circulares), bem como o número de cópias que o simulador utilizou para gerar os dados. Um Taxonomy ID de -1 indica que aquela determinada espécie não tem um ID associado a ele.

O MetaSim foi configurado para fragmentar os genomas de entrada e gerar um conjunto de 100.000 reads com uma representatividade de cada gênero proporcional à abundância identificada pelo MG-RAST para o projeto de Brulc *et al.*. Com isso, temos um metagenoma simulado baseado em abundâncias derivadas de dados reais. Devido às diferenças dos bancos de dados utilizados entretanto, nem todos os gêneros dentre os 50 mais abundantes do rumen de vaca aparecem no conjunto simulado. Assim, gêneros além do 50º mais abundante foram considerados, de forma que o conjunto simulado contem 50 gêneros.

O conjunto de reads simulados foi então submetido ao mesmo pipeline de processamento que as sequências do Zoológico, usando as mesmas versões dos bancos de dados e os mesmos parâmetros.

Reads	GI	Tax Id	Nome	Cópias	Tamanho
56738	93004831	335284	Psychrobacter cryohalolentis K5	17619	3059876
14860	15893298	272562	Clostridium acetobutylicum ATCC 824	7960	3940880
13943	375356399	862962	Bacteroides fragilis 638R	6588	5373121
4530	302344773	553174	Prevotella melaninogenica ATCC 25845 chromosome I	6560	1796408
2686	213155370	480119	Acinetobacter baumannii AB0057	3302	4050513
2152	317054731	697329	Ruminococcus albus 7	3089	3685408
1259	238915976	515620	Eubacterium eligens ATCC 27750	3136	2144190
947	302669374	515622	Butyrivibrio proteoclasticus B316 chromosome 1	2117	3554804
688	154684518	326423	Bacillus amyloliquefaciens FZB42	1730	3918589
367	218888746	557722	Pseudomonas aeruginosa LESB58	961	6601757
232	150006674	435591	Parabacteroides distasonis ATCC 8503	943	4811379
192	296112228	749219	Moraxella catarrhalis RH4	1364	1863286
145	22536185	208435	Streptococcus agalactiae 2603V/R	1085	2160267
120	337744364	1036673	Paenibacillus mucilaginosus KNP414	476	8663821
119	347530298	585394	Roseburia hominis A2-183	794	3592125
84	325955721	891391	Lactobacillus acidophilus 30SC	841	2078001
76	257062754	471855	Slackia heliotrinireducens DSM 20476	661	3165038
75	119773142	326297	Shewanella amazonensis SB2B	544	4306142
70	390945347	-1	Alistipes finegoldii DSM 17242	518	3734239
68	288559258	634498	Methanobrevibacter ruminantium M1	544	2937203
55	219666071	272564	Desulfitobacterium hafniense DCB-2	430	5279134
53	29374661	226185	Enterococcus faecalis V583	537	3218031
45	336122587	882102	Vibrio anguillarum 775 chromosome I	515	3063912
41	317151727	643562	Desulfovibrio aespoeensis Aspo-2	393	3629109
38	150387853	293826	Alkaliphilus metalliredigens QYMF	409	4929566
37	115350056	339670	Burkholderia ambifaria AMMD chromosome 1	384	3556545
27	119025018	367928	Bifidobacterium adolescentis ATCC 15703	463	2089645
27	328956382	208596	Carnobacterium sp. 17-4	344	2635294
27	162960844	227882	Streptomyces avermitilis MA-4680	227	9025608
24	56418535	235909	Geobacillus kaustophilus HTA426	315	3544776
21	222528057	521460	Caldicellulosiruptor bescii DSM 6725	350	2919718
21	332299201	879243	Porphyromonas asaccharolytica DSM 20707	415	2186370
20	284800255	653938	Listeria monocytogenes 08-5578	289	3032288
20	59800473	242231	Neisseria gonorrhoeae FA 1090	435	2153922
19	197116402	404380	Geobacter bemidjiensis Bem	252	4615150
18	317131008	663278	Ethanoligenens harbinense YUAN-3	275	3008576
18	19703352	190304	Fusobacterium nucleatum subsp. nucleatum ATCC 25586	388	2174500
17	347534971	1034807	Flavobacterium branchiophilum FL-15	251	3559884
16	257789778	479437	Eggerthella lenta DSM 2243	281	3632260
15	284047386	591001	Acidaminococcus fermentans DSM 20731	280	2329769
13	253771435	866768	Escherichia coli 'BL21-Gold(DE3)pLysS AG'	243	4570938
13	120552944	351348	Marinobacter aquaeolei VT8	269	4326849
11	330837866	546271	Selenomonas sputigena ATCC 35185	297	2568361
11	82749777	273036	Staphylococcus aureus RF122	284	2742531
10	258513366	485916	Desulfotomaculum acetoxidans DSM 771	224	4545624
10	333992987	545695	Treponema azotonutricium ZAS-9	221	3855671
9	227831830	548476	Corynebacterium aurimucosum ATCC 700975	222	2790189
5	257783814	521095	Atopobium parvulum DSM 20469	265	1543805
4	33151282	233412	Haemophilus ducreyi 35000HP	212	1698955
4	289577265	580331	Thermoanaerobacter italicus Ab9	240	2451061

Tabela 6.6: Espécies utilizadas para gerar os dados simulados, cada uma com seus respectivos números de identificação no GenBank, Taxonomy ID, o número de reads gerados para o conjunto simulado, o número de cópias de cada espécie usada na simulação, e o tamanho do genoma de cada espécie.

A princípio, submetemos os dados simulados ao NBC. Os gêneros mais abundantes encontrados, mostrados na figura 6.17, são bem diferentes dos dados gerados pelo MetaSim. Dessa forma, podemos concluir que para os nossos dados, o NBC não se mostrou uma ferramenta confiável para a análise de diversidade. Essa diferença pode ser devido ao diferente banco de dados de referência que foi utilizado com o NBC, ou devido ao próprio método utilizado para comparação, que não se trata de uma comparação direta com as sequências de entrada.

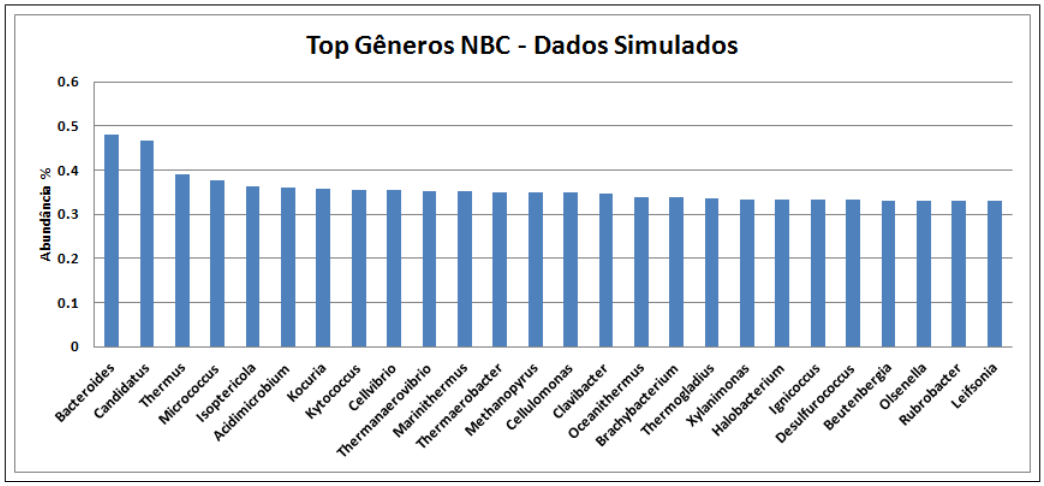


Figura 6.17: Gêneros mais frequentes identificados pelo NBC para os dados simulados.

A Tabela 6.7 mostra todos os 50 gêneros do conjunto simulado e os mais abundantes identificados pelo NCBI-NR, pelo SwissProt e pelo MG-RAST, juntamente com as porcentagens de abundância relativa de cada gênero. Observamos que sete dos gêneros mais abundantes estão entre os dez mais do Swissprot, com os três primeiros aparecendo na mesma ordem, enquanto que os resultados obtidos com o NCBI-NR mostram os oito gêneros mais abundantes exatamente na mesma ordem que os dados simulados. O MG-RAST também identificou oito dentre os dez gêneros mais abundantes, mas só o mais abundante, *Psychrobacter*, aparece na mesma posição, e mesmo assim com uma porcentagem de abundância bem distinta. enquanto na simulação, o gênero *Psychrobacter* representa em torno de 56% da amostra, o MG-RAST identificou esse gênero como tendo 34% de representatividade.

Os resultados com o NCBI-NR mostram porcentagens de abundância muito próximas do conjunto original, o gênero *Psychrobacter*, chegou a ser identificado em 56.753% das sequências pelo NCBI-NR, bem próximo dos 56.738% originais. O gênero *Clostridium* também tem uma porcentagem de abundância próxima do original, identificado em 13.242% das sequências de um original de 14.860%. Apesar de ter número mais distantes dos originais, a distribuição dos gêneros mais abundantes identificados pelo Swiss-Prot é

bem próxima da distribuição original, com o gênero mais abundante tendo 63.379% de representação, e o segundo com 17.166%.

O gênero *Acinetobacter* foi o gênero mais abundante que mais se aproximou da abundância relativa real nos resultados do Swiss-Prot, com 2.726% de representação, comparado com os originais 2.686%. Os resultados do NCBI-NR, no entanto, não ficaram longe, apontando 3.134% de representação. O MG-RAST, no entanto, se distancia mais e chega a 6.601% de representação para esse gênero.

A diferença da distribuição de gêneros do MG-RAST pode ser observada na Figura 6.18, que mostra os dez gêneros mais abundantes dos dados simulados e suas respectivas porcentagens de abundância. Em conjunto, podemos ver a distribuição de abundância desses mesmos dez gêneros em cada um dos três bancos de dados testados. O NCBI-NR sempre se mantém próximo aos resultados originais. Embora o Swiss-Prot não tenha identificado os gêneros *Prevotella* e *Ruminococcus* dentre os mais abundantes (apenas uma sequência de cada um desses gêneros foi identificada pelo Swiss-Prot), todos os outros se mantêm próximos do original, com a maior diferença sendo no gênero mais abundante, *Psychrobacter*. Já o MG-RAST mostra uma distribuição mais diferenciada, com alguns gêneros menos abundantes (*Psychrobacter* e *Clostridium*), alguns mais abundantes (*Bacteroides* e *Acinetobacter*), e outros mais próximos do original (*Prevotella* e *Ruminococcus*).

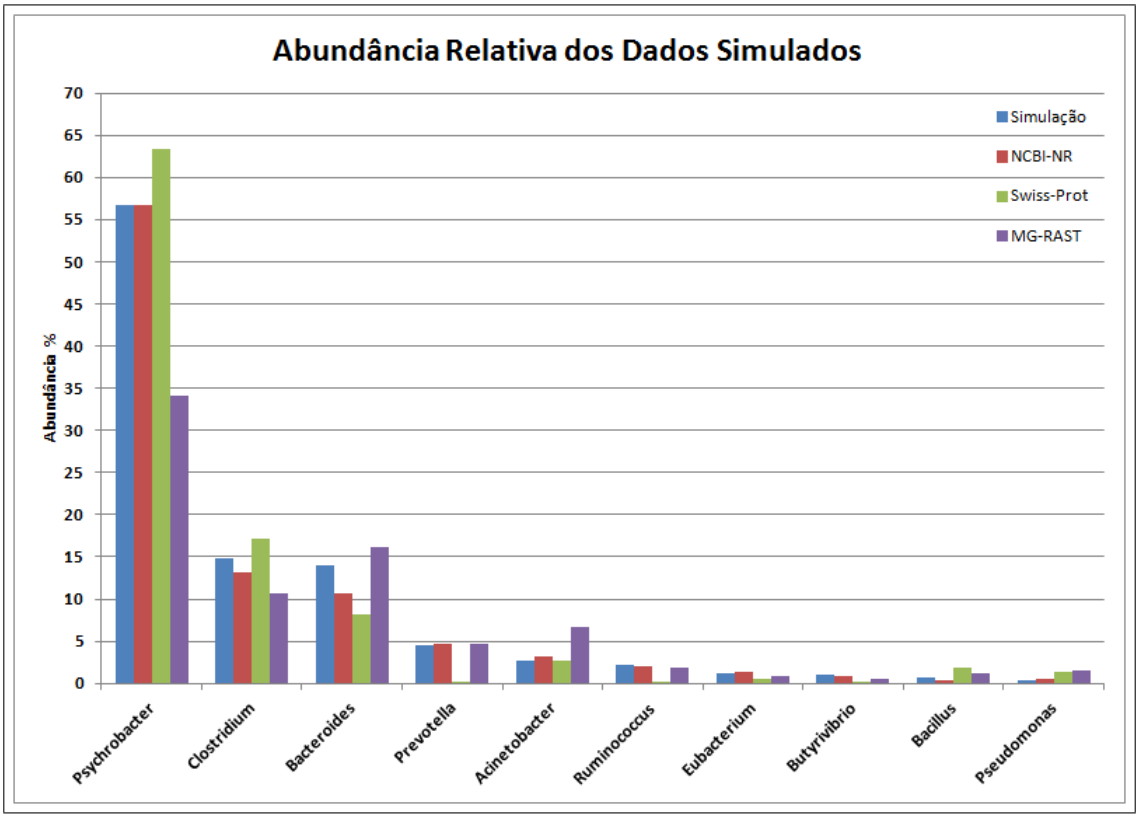


Figura 6.18: Os dez gêneros mais abundantes dos dados simulados, e suas respectivas porcentagens de abundância, comparadas com as identificadas pelo NCBI-NR, Swiss-Prot e MG-RAST.

O MG-RAST também identificou um número maior de hits do que o que realmente existia, como é notado no número de hits do gênero *Bacteroides*, o segundo mais abundante segundo o MG-RAST. Na simulação, 13.943 reads eram do gênero *Bacteroides*, e o MG-RAST identificou 19.351 reads para esse mesmo gênero. Essa diferença pode ser devido à grande quantidade de bancos de dados de referência utilizados pelo MG-RAST para identificação, algumas sequências de entrada são identificadas como hits em mais de um banco de dados. No total, o MG-RAST identificou 119.968 sequências das 100.000 originais, distribuídas em 843 gêneros. O Swiss-Prot identificou 2.971 sequências e 72 gêneros, e o NCBI-NR 14.424 sequências e 96 gêneros.

Mesmo com um número de hits inferior ao do MG-RAST, 96% dos gêneros simulados foram identificados pelo NCBI-NR, e 62% foram identificados pelo Swiss-Prot. Todos os gêneros identificados corretamente pelo Swiss-Prot foram também identificados pelo NCBI-NR, apenas os gêneros *Treponema* e *Atopobium* não foram identificados por nenhuma das duas bases. O MG-RAST identificou todos os gêneros simulados, como era de se esperar, porém o último gênero identificado, *Atopobium*, ficou na 327ª posição. O

outro gênero não-identificado pelas outras duas bases, *Treponema*, ficou na 179ª posição do MG-RAST.

Também é interessante observar que o gênero *Enhydrobacter*, dito como o sexto mais abundante pelo MG-RAST e 14º do NCBI-NR, não fez parte do conjunto de dados simulados, mas apareceu dentre as primeiras posições nesses bancos de dados.

Os resultados da Tabela 6.7 consideram o primeiro hit do BLAST, no entanto, a saída padrão do BLAST oferece até 250 hits para cada sequência de entrada. É interessante verificar os resultados caso mais de um hit seja considerado. Por exemplo, se um gênero G aparece frequentemente como o segundo hit de diversas sequências de entrada e poucas vezes como o primeiro hit, ao considerar apenas o primeiro hit, podemos concluir que G não é tão abundante na amostra, mesmo ele sendo considerado segundo hit pelo BLAST, o que ainda é significativo.

Com esse pensamento em mente, consideramos todos os n primeiros hits do BLAST, con n variando de 1 a 250, e calculamos duas medidas de comparação:

- **Erro de abundância:** Para cada n, o erro de abundância e_a foi calculado como:

$$e_a = \left(\sum_{i=1}^{k} \left[\left(\frac{Ai_{obs} * 100}{At_{obs}} \right) - \left(\frac{Ai_{sim} * 100}{At_{sim}} \right) \right]^2 \right)^{\frac{1}{2}}$$

  Onde k é o número total de gêneros, Ai_{obs} é o valor de abundância observada para o gênero i, At_{obs} é a soma das abundâncias de todos os gêneros observados, Ai_{sim} é a abundância simulada do gênero i e At_{sim} é a soma das abundâncias de todos os gêneros simulados. É importante notar que cada gênero do conjunto observado só foi comparado com o seu correspondente do conjunto simulado. Para os gêneros que não tinham correspondentes em algum conjunto, a abundância deles foi considerada zero no conjunto em que eles não existiam.

- **Erro de posição relativa:** De forma semelhante, com o objetivo de verificar quão distantes os gêneros estão dos dados simulados em termos de posição, calculamos o erro de posição relativa, e_p, baseado nas diferenças de posições de cada gênero. Por exemplo, a posição do gênero *Pseudomonas* nos resultados observados do NCBI-NR para $n = 1$ é 9, enquanto que nos dados simulados, a posição real dele é 10, ou seja, o erro de posição relativa nesse caso é 1. Assim, para cada n, seu e_p correspondente é dado como:

$$e_p = \left(\sum_{i=1}^{k} (Pi_{obs} - Pi_{sim})^2 \right)^{\frac{1}{2}}$$

Reads Simulados		Abundância NCBI-NR		Abundância Swiss-Prot		Abundância MG-RAST	
Gênero	(%)	Gênero	(%)	Gênero	(%)	Gênero	(%)
Psychrobacter	56.738	Psychrobacter	56.753	Psychrobacter	63.379	Psychrobacter	34.033
Clostridium	14.860	Clostridium	13.242	Clostridium	17.166	Bacteroides	16.130
Bacteroides	13.943	Bacteroides	10.670	Bacteroides	8.112	Clostridium	10.734
Prevotella	4.530	Prevotella	4.659	Acinetobacter	2.726	Acinetobacter	6.601
Acinetobacter	2.686	Acinetobacter	3.134	Bacillus	1.885	Prevotella	4.603
Ruminococcus	2.152	Ruminococcus	2.017	Pseudomonas	1.380	Enhydrobacter	1.995
Eubacterium	1.259	Eubacterium	1.317	Eubacterium	0.505	Ruminococcus	1.891
Butyrivibrio	0.947	Butyrivibrio	0.922	Escherichia	0.370	Moraxella	1.825
Bacillus	0.688	Pseudomonas	0.499	Parabacteroides	0.370	Pseudomonas	1.442
Pseudomonas	0.367	Bacillus	0.388	Vibrio	0.303	Bacillus	1.139
Parabacteroides	0.232	Moraxella	0.263	Shigella	0.303	Eubacterium	0.863
Moraxella	0.192	Streptococcus	0.215	Streptococcus	0.269	Shewanella	0.581
Streptococcus	0.145	Parabacteroides	0.146	Shewanella	0.236	Parabacteroides	0.540
Paenibacillus	0.120	Enhydrobacter	0.132	Haemophilus	0.202	Vibrio	0.534
Roseburia	0.119	Roseburia	0.118	Neisseria	0.168	Butyrivibrio	0.528
Lactobacillus	0.084	Lactobacillus	0.104	Pseudoalteromonas	0.168	Burkholderia	0.464
Slackia	0.076	Slackia	0.090	Ruminococcus	0.101	Marinobacter	0.436
Shewanella	0.075	Burkholderia	0.076	Alcanivorax	0.101	Neisseria	0.433
Alistipes	0.070	Listonella	0.069	Azotobacter	0.101	Pseudoalteromonas	0.350
Methanobrevibacter	0.068	Desulfitobacterium	0.069	Methylobacillus	0.067	Streptococcus	0.312
Desulfitobacterium	0.055	Shewanella	0.069	Staphylococcus	0.067	Lactobacillus	0.263
Enterococcus	0.053	Streptomyces	0.062	Serratia	0.067	Escherichia	0.203
Vibrio	0.045	Methanobrevibacter	0.055	Coxiella	0.067	Alcanivorax	0.183
Desulfovibrio	0.041	Alistipes	0.055	Geobacter	0.067	Haemophilus	0.178
Alkaliphilus	0.038	Paenibacillus	0.049	Chromobacterium	0.067	Porphyromonas	0.175
Burkholderia	0.037	Eggerthella	0.035	Photobacterium	0.067	Marinomonas	0.159
Bifidobacterium	0.027	Ethanoligenens	0.035	Psychromonas	0.067	Paenibacillus	0.157
Carnobacterium	0.027	Enterococcus	0.035	Salmonella	0.067	Enterococcus	0.138
Streptomyces	0.027	Neisseria	0.035	Burkholderia	0.067	Alistipes	0.137
Geobacillus	0.024	Geobacillus	0.035	Actinobacillus	0.067	Xanthomonas	0.136
Caldicellulosiruptor	0.021	Desulfovibrio	0.035	Aeromonas	0.034	Actinobacillus	0.133
Porphyromonas	0.021	Tannerella	0.028	Ralstonia	0.034	Streptomyces	0.131
Listeria	0.020	Paraprevotella	0.028	Caldanaerobacter	0.034	Desulfovibrio	0.129
Neisseria	0.020	Fusobacterium	0.028	Marinomonas	0.034	Chromohalobacter	0.122
Geobacter	0.019	Bifidobacterium	0.028	Enterococcus	0.034	Hahella	0.116
Ethanoligenens	0.018	Listeria	0.028	Bifidobacterium	0.034	Yersinia	0.112
Fusobacterium	0.018	Caldicellulosiruptor	0.028	Porphyromonas	0.034	Legionella	0.111
Flavobacterium	0.017	Geobacter	0.021	Alkaliphilus	0.034	Salmonella	0.105
Eggerthella	0.016	Staphylococcus	0.021	Listeria	0.034	Photobacterium	0.103
Acidaminococcus	0.015	Caenorhabditis	0.021	Fusobacterium	0.034	Halomonas	0.103
Escherichia	0.013	Salmonella	0.021	Corynebacterium	0.034	Geobacillus	0.101
Marinobacter	0.013	Marinobacter	0.021	Geobacillus	0.034	Psychromonas	0.101
Selenomonas	0.011	Escherichia	0.021	Streptomyces	0.034	Ralstonia	0.101
Staphylococcus	0.011	Vibrio	0.021	Lactobacillus	0.034	Geobacter	0.100
Desulfotomaculum	0.010	Carnobacterium	0.021	Flavobacterium	0.034	Slackia	0.096
Treponema	0.010	Alkaliphilus	0.021	Yersinia	0.034	Acidovorax	0.093
Corynebacterium	0.009	Thermoanaerobacter	0.014	Prevotella	0.034	Staphylococcus	0.092
Atopobium	0.005	Haemophilus	0.014	Chlamydomonas	0.034	Roseburia	0.091
Haemophilus	0.004	Pseudoalteromonas	0.014	Buchnera	0.034	Alkaliphilus	0.088
Thermoanaerobacter	0.004	Gallibacterium	0.014	Myxococcus	0.034	Cupriavidus	0.086

Tabela 6.7: Comparação entre o número de reads simulados dos 50 gêneros simulados e os gêneros identificados com os bancos de dados SwissProt, NCBI-NR e MG-RAST, com suas respectivas abundâncias relativas.

Onde Pi_{obs} é a posição do gênero i observada e Pi_{sim} é a posição do gênero i de acordo com os dados simulados. A maior diferença desse cálculo para o do erro de abundância, é no caso de um determinado gênero não ter correspondente em um dos conjuntos. Como a posição de um gênero não pode ser traduzida em porcentagens, consideramos que a posição do gênero que não possui correspondente é $maxPos + 1$, onde $maxPos$ é calculada como o máximo entre a maior posição dos gêneros simulados e a maior posição entre os gêneros observados. Por exemplo, caso o experimento com o SwissProt tenha identificado 80 gêneros, um gênero que não pertence a um dos conjuntos (simulados ou observados), é dito estar na posição 81 daquele conjunto em que ele não pertence.

As figuras 6.19 e 6.20 mostram gráficos com os erros de abundância e de posição relativa respectivamente para os resultados com o Swiss Prot. O menor erro de abundância é alcançado quando os 9 primeiros hits são considerados, 8.72, sendo que o erro considerando apenas o primeiro hit foi de 9.49. O menor erro de posição é alcançado considerando os 2 primeiros hits, de 151.00, e o erro considerando um hit é de 159.15. Além desses pontos mais baixos, os erros aumentam junto com o número de hits considerados.

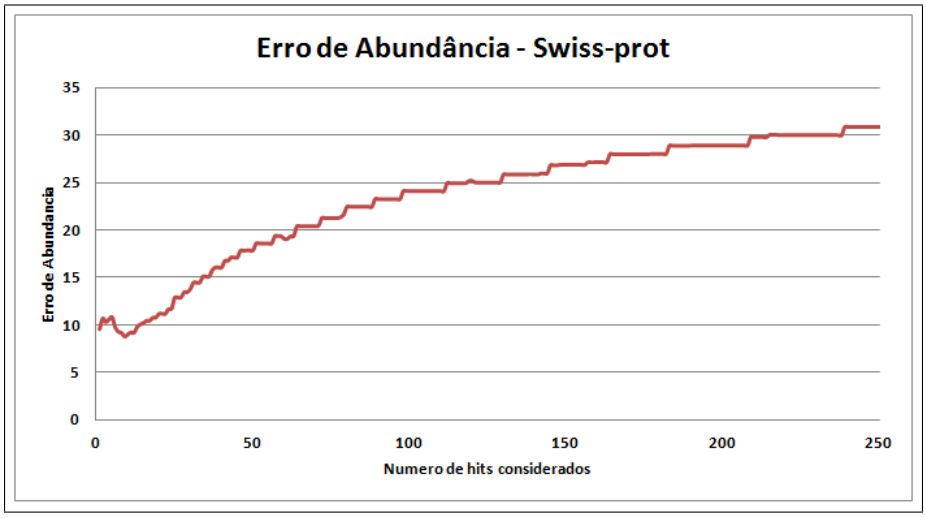


Figura 6.19: Erro de abundância considerando os n primeiros hits do Swiss Prot.

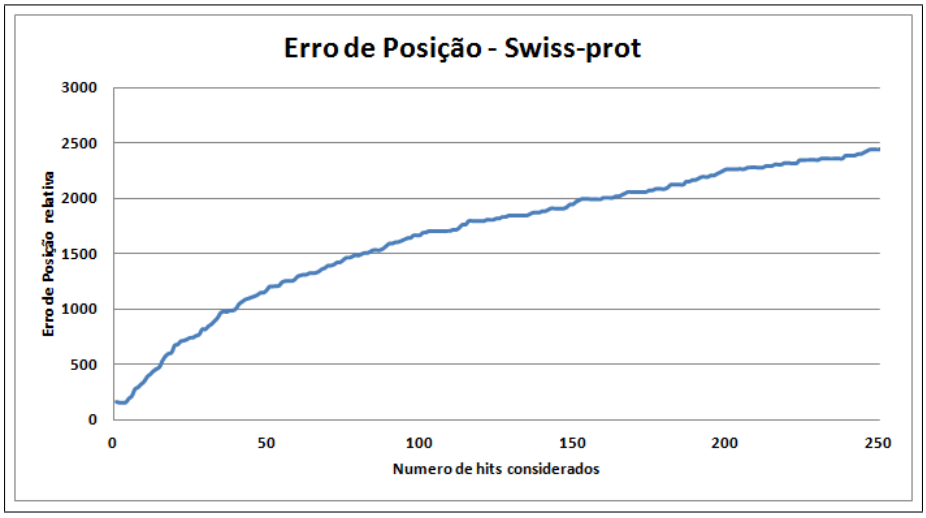


Figura 6.20: Erro de posição relativa considerando os n primeiros hits do Swiss Prot.

As figuras 6.21 e 6.22 mostram gráficos com os erros para os resultados com o NCBI-NR. O menor erro de abundância nesse caso é alcançado considerando os 3 primeiros hits, de 4.47, muito próximo do erro de 5.48, que foi o calculado considerando apenas o primeiro hit. O menor erro de posição relativa ocorre quando consideramos apenas o primeiro hit, que é de 101.00. Da mesma forma que o Swiss-Prot, além desses mínimos, o erro só aumenta com o número de hits considerados.

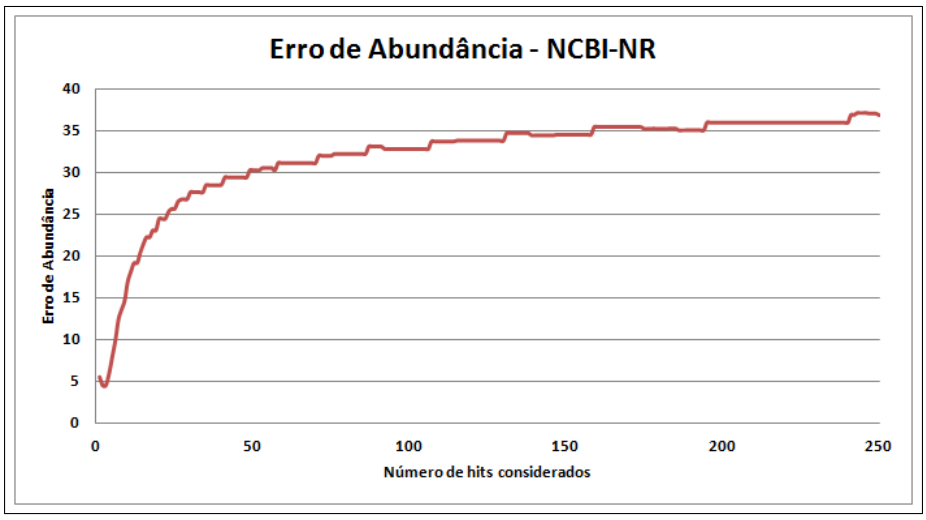


Figura 6.21: Erro de abundância considerando os n primeiros hits do NCBI-NR.

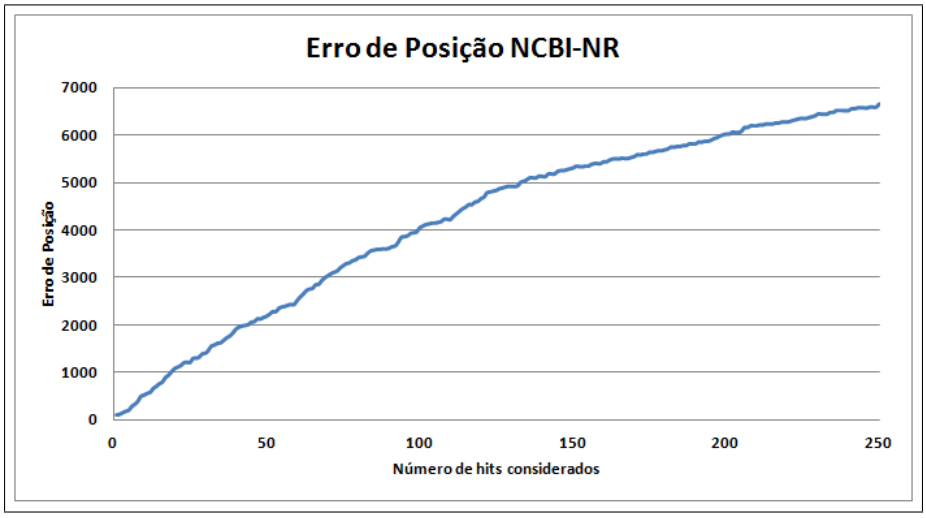


Figura 6.22: Erro de posição relativa considerando os n primeiros hits do NCBI-NR.

Os cálculos desses erros indicam que considerar muitos hits do BLAST não necessariamente gera resultados melhores, e a proposta de considerar apenas o primeiro hit funciona bem e tem um erro baixo. Os resultados do MG-RAST não foram considerados para a construção dos gráficos de erros pois a ferramenta só fornece resultados com o melhor hit. Usando a fórmula, o erro de posição para o MG-RAST com o melhor hit ficou em 10822.62 e o erro de abundância em 23.04. Ambos são números bem maiores que os calculados para Swiss-Prot e NCBI-NR, o que leva à conclusão de que o MG-RAST, que utiliza o método de melhor hit, ou usa mais do que apenas o primeiro hit, ou a grande quantidade de bancos de dados usadas pela ferramenta causa esse desvio.

6.8 Análise comparativa

As sequências do Zoológico foram comparadas com vários bancos de dados diferentes, e as dificuldades de se tratar dados metagenômicos se torna aparente. Não existe um método 100% confiável, mas as comparações entre os resultados pode servir para indicar uma composição mais provável das populações.

Os métodos utilizados para comparar as sequências com os bancos NCBI-NR e Swiss-Prot são os mesmos, mas o método utilizado com o RDP é significativamente diferente. As sequências são agrupadas em clusters que identificam uma OTU genérica, não necessariamente uma espécie. As comparações são feitas com os resultados obtidos com 95% de identidade, que foi a maior porcentagem de similaridade utilizada em todos os métodos.

Considerando os dez gêneros mais abundantes identificados pelo MG-RAST para ZC1, observamos que seis deles aparecem entre os dez primeiros identificados somente pelo

NCBI-NR. Além disso, o gênero mais abundante, *Acinetobacter*, é o mesmo. Esses mesmos seis gêneros aparecem entre os dez mais abundantes na análise feita com o Swiss-prot, com poucas alterações na ordem, mas *Acinetobacter* continua sendo o mais abundante. Já com o RDP, apenas três desses estão entre os dez mais abundantes, e *Acinetobacter* aparece como o segundo gênero mais abundante.

Considerando ZC2, todas as análises identificaram o gênero *Lactobacillus* como o mais abundante, com significativa diferença para o segundo colocado. Dos dez mais abundantes identificados pelo MG-RAST, seis deles estão também entre os dez mais abundantes nas análises feitas com o NCBI-NR, e cinco deles nas realizadas com o Swiss-prot. Os mesmos cinco identificados pelo Swiss-prot estão também entre os dez mais abundantes do RDP.

De forma mais geral, comparando o número de OTUs identificados, no caso do MG-RAST, foram identificados 2103 gêneros em ZC1 e 1928 gêneros em ZC2. Considerando uma porcentagem de identidade de 95%, o RDP se aproxima desses números, com 2129 OTUs identificados em ZC1 e 1856 OTUs em ZC2. No caso do Swiss-prot, poucos hits alcançaram 95% de identidade, apenas 409 em ZC1 e 357 em ZC2. Já com o NCBI-NR, foram identificados 379 OTUs com 95% de identidade para ZC1 e 314 para ZC2. É importante ressaltar que as sequências do Zoológico só foram comparadas com 10% do NCBI-NR, resultados mais interessantes podem vir de uma comparação mais completa.

É interessante notar que nos dois casos em que foram identificados mais gêneros, as curvas de rarefação não atingem um platô, sugerindo que o número de sequências amostradas não foi suficiente para identificar uma quantidade significativa de espécies que representam o ambiente. O mesmo não acontece nos casos do Swiss-prot e do NCBI-NR, as curvas de rarefação dos dois casos atingem um claro plateau.

Mesmo com a diferença numérica, todos os resultados apontam uma maior diversidade da amostra ZC1 em relação à ZC2, o que pode ser devido à presença de um organismo dominante na amostra ZC2.

De todos os métodos utilizados, os resultados do NBC são os que mais divergem do que foi observado nos outros, mesmo com os dados simulados, os resultados foram consideravelmente diferentes do esperado. Os resultados das simulações com os bancos de dados Swiss-Prot e NCBI-NR no entanto, sugerem que as análises com esses dois bancos são confiáveis, principalmente os resultados com este último.

O MG-RAST também identificou bem os gêneros utilizados na simulação, mas em contrapartida, identificou uma quantidade enorme de gêneros que não existiam no conjunto original. Enquanto o NCBI-NR, identificou hits para 96 gêneros a partir de um conjunto com 50, o MG-RAST identificou hits para 843 gêneros para o mesmo conjunto. Isso leva à conclusão de que a análise do MG-RAST por melhor hit é insatisfatória, tanto no número de hits de cada gênero como na quantidade de gêneros identificados.

A Tabela 6.8 mostra uma comparação entre o número de reads classificados por cada

método. Podemos observar a enorme quantidade de reads identificados pelo NBC. Também podemos observar que em todas as situações, mais reads foram identificados em ZC2.

	Reads Totais	MG-RAST	RDP	Swiss-prot	NCBI-NR	NBC
ZC1	3.167.044	2.668.964	1.569	32.132	7.718	590.612.405
ZC2	2.966.244	2.786.392	3.540	50.534	20.380	640.500.036

Tabela 6.8: Comparação entre o número de reads totais e identificados com os bancos de dados RDP, SwissProt, NCBI-NR e MG-RAST e pelo NBC.

Capítulo 7

Conclusões e Trabalhos Futuros

Motivados pelo crescimento da metagenômica, buscamos trabalhar mais a fundo nesse importante passo que é a análise de diversidade. O MG-RAST é um boa ferramenta para um estudo inicial, mas acaba por gerar muito ruído nos resultados, talvez pela grande quantidade de bancos de dados utilizados. O nosso trabalho foi refinar esses resultados, usando diferentes métodos e bancos de dados.

Os programas para análise de diversidade estudados fornecem diversas formas para abordarmos o problema. Infelizmente, o único programa que pôde ser testado não apresentou resultados satisfatórios com os dados do Zoológico, e os resultados com os dados simulados reforçaram a falta de confiança do Naive Bayes Classifier com os nossos dados.

O RDP proporcionou um estudo mais focado no rRNA 16S, além de ter um processamento prévio com o Infernal, que analisou a estrutura secundária do RNA. Os resultados com o RDP foram processados pelo mothur, que formou clusters de OTUs. Apesar desses clusters não representarem um único grupo taxonômico, eles são agrupados por porcentagem de similaridade, e é essa porcentagem que é o parâmetro para classificar uma sequência em uma determinada espécie, gênero, família ou outro grupo.

As análises com o Uniprot Swiss-Prot e NCBI-NR foram as mais próximas entre si, e os resultados com os dados simulados foram bastantes satisfatórios para aumentar a confiança dos resultados com os dados reais. A simulação foi fundamental para mostrar que é possível obter resultados confiáveis utilizando um método simples de comparação com o melhor hit, e embora esse seja um método amplamente difundido na literatura, não passou por esse teste de forma a validar sua eficácia.

Uma comparação mais completa das sequências do Zoológico com o banco de dados NCBI-NR pode revelar resultados mais precisos. Esse foi o banco que mais se aproximou da composição real dos dados simulados e que, por consequência, é o melhor indicador da composição da população nos dados reais.

O contexto que nosso trabalho se insere se preocupa em obter uma análise de diversi-

dade confiável, pois esse conhecimento da composição da população é muito importante para análises futuras de funcionamento da população e interação com o ambiente, bem como para um maior entendimento da dinâmica microbiológica no processo de compostagem. As amostras coletadas e utilizadas ao longo desse trabalho não possuem relações diretas entre si, e trabalhos futuros podem analisar amostras coletadas ao longo de um mesmo processo, a fim de montar uma série temporal da compostagem e tirar mais conclusões a respeito do processo. Por exemplo, a dominância de *Lactobacillus* na amostra ZC2 pode fazer parte da dinâmica microbiológica da compostagem, mas apenas com os dados que tivemos disponíveis não é possível fazer essa conclusão.

A metagenômica é uma área que está crescendo aceleradamente e poucos estudos sobre isso são feitos no Brasil. Apesar das dificuldades inerentes ao problema, é possível obter resultados interessantes, buscando sempre um melhor entendimento do mundo em que vivemos.

Referências Bibliográficas

[1] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997.

[2] R. D. Bentley. Whole-genome re-sequencing. *Current Opinion in Genetics & Development*, 16(6):545–552, 2006.

[3] B. Bohannan and J. Hughes. New approaches to analyzing microbial biodiversity data. *Current Opinion in Microbiology*, 6(3):282–287, June 2003.

[4] A. Brady and S. Salzberg. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, 6(9):673–676, 2009.

[5] J. Brulc, D. Antonopoulos, M. Berg Miller, M. Wilson, A. Yannarell, E. Dinsdale, R. Edwards, E. Frank, J. Emerson, P. Wacklin, P. Coutinho, B. Henrissat, K. Nelson, and B. White. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proceedings of the National Academy of Sciences (PNAS)*, 106(6):1948–1953, 2009.

[6] A. Chao. Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics*, 11:265–270, 1984.

[7] A. Chao. Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. *Biometrics*, 43:783–791, 1987.

[8] A. Chao and S.-M. Lee. Estimating the Number of Classes via Sample Coverage. *Journal of the American Statistical Association*, 87:210–217, 1992.

[9] C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann. Evolution of the Protein Repertoire. *Science*, 300(5626):1701–1703, June 2003.

[10] J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. The

Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37(suppl 1):D141–D145, 2009.

[11] R. Colwell. EstimateS: Statistical estimation of species richness and shared species from samples. Version 8.2. University of Connecticut.

[12] R. Colwell and J. Coddington. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B*, 345:101–118, 1994.

[13] The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 40(D1):D71–D75, 2012.

[14] T. Curtis, W. Sloan, and J. Scannell. Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 99:10494–10499, 2002.

[15] E. F. DeLong, C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N-U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl. Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science*, 311(5760):496–503, January 2006.

[16] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Molecular Biology and Evolution*, 16(10):1391–1399, October 1999.

[17] B. Ewing and P. Green. Base-Calling of Automated Sequencer Traces Using Phred II. Error Probabilities. *Genome Research*, 8(3):186–194, 1998.

[18] D. Field, G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M. J. Allen, S. V. Angiuoli, M. Ashburner, N. Axelrod, S. Baldauf, S. Ballard, J. Boore, G. Cochrane, J. Cole, P. Dawyndt, P. De Vos, C. dePamphilis, R. Edwards, N. Faruque, R. Feldman, J. Gilbert, P. Gilna, F. O. Glockner, P. Goldstein, R. Guralnick, D. Haft, D. Hancock, H. Hermjakob, C. Hertz-Fowler, P. Hugenholtz, I. Joint, L. Kagan, M. Kane, J. Kennedy, G. Kowalchuk, R. Kottmann, E. Kolker, S. Kravitz, N. Kyrpides, J. Leebens-Mack, S. E. Lewis, K. Li, A. L. Lister, P. Lord, N. Maltsev, V. Markowitz, J. Martiny, B. Methe, I. Mizrachi, R. Moxon, K. Nelson, J. Parkhill, L. Proctor, O. White, S.-A. Sansone, A. Spiers, R. Stevens, P. Swift, C. Taylor, Y. Tateno, A. Tett, S. Turner, D. Ussery, B. Vaughan, N. Ward, T. Whetzel, I. San Gil, G. Wilson, and A. Wipat. The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, 26(5):541–547, May 2008.

[19] R. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. Eddy, E. Sonnhammer, and A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Research*, 34(suppl 1):D247–D251, 2006.

[20] H. García Martín, N. Ivanova, V. Kunin, F. Warnecke, K. W. Barry, A. C. Mchardy, C. Yeates, S. He, A. A. Salamov, E. Szeto, E. Dalin, N. H. Putnam, H. J. Shapiro, J. L. Pangilinan, I. Rigoutsos, N. C. Kyrpides, L. L. L. Blackall, K. D. Mcmahon, and P. Hugenholtz. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nature Biotechnology*, 24(10):1263–1269, October 2006.

[21] W. Gerlach and J. Stoye. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Research*, 39(14):e91, 2011.

[22] S. R. Gill, M. Pop, R. T. DeBoy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson. Metagenomic Analysis of the Human Distal Gut Microbiome. *Science*, 312(5778):1355–1359, 2006.

[23] E. Glass, F. Meyer, J. Gilbert, D. Field, S. Hunter, R. Kottmann, N. Kyrpides, S. Sansone, L. Schriml, P. Sterk, O. White, and J. Wooley. Meeting Report from the Genomic Standards Consortium (GSC) Workshop 10. *Standards in Genomic Sciences*, 3(3), 2010.

[24] D. Gordon, C. Abajian, and P. Green. Consed: A graphical tool for sequence finishing. *Genome Research*, 8(3):195–202, March 1998.

[25] A.J.F. Griffiths. *Introdução à genética*. Guanabara Koogan, 2009.

[26] M. Haque, T. Ghosh, D. Komanduri, and S. Mande. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, 25(14):1722–1730, July 2009.

[27] K. Heck, G. van Belle, and D. Simberloff. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology*, 56:1459–1461, 1975.

[28] S.-H. Hong, J. Bunge, S.-O. Jeon, and S. Epstein. Predicting microbial species richness. *Proceedings of the National Academy of Sciences of the United States of America*, 103(1):117–122, 2006.

[29] X. Huang and A. Madan. CAP3: A DNA Sequence Assembly Program. *Genome Research*, 9(9):868–877, September 1999.

[30] J. Huber, D. Mark Welch, H. Morrison, S. Huse, P. Neal, D. Butterfield, and M. Sogin. Microbial Population Structures in the Deep Marine Biosphere. *Science*, 318(5847):97–100, 2007.

[31] P. Hugenholtz. Exploring prokaryotic diversity in the genomic era. *Genome Biology*, 3(2):reviews0003.1–reviews0003.8, 2002.

[32] J. Hughes, J. J Hellmann, T. H Ricketts, and B. Bohannan. Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity. *Applied and Environmental Microbiology*, 67(10):4399–4406, 2001.

[33] R. Hughes. Theories and Models of Species Abundance. *The American Naturalist*, 128(6):879–899, 1986.

[34] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. MEGAN analysis of metagenomic data. *Genome research*, 17(3):377–386, March 2007.

[35] D. B. Jaffe, J. Butler, S. Gnerre, E. Mauceli, K. Lindblad-Toh, J. P. Mesirov, M. C. Zody, and E. S. Lander. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Research*, 13(1):91–96, January 2003.

[36] L. Krause, N. Diaz, A. Goesmann, S. Kelley, T.. Nattkemper, F. Rohwer, R.. Edwards, and J. Stoye. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research*, 36(7):2230–2239, 2008.

[37] L. Krause, N. N. Diaz, D. Bartels, R. A. Edwards, A. Pühler, F. Rohwer, F. Meyer, and J. Stoye. Finding novel genes in bacterial communities isolated from the environment. *Bioinformatics*, 22(14):e281–e289, July 2006.

[38] V. Kunin, A. Copeland, Al. Lapidus, K. Mavromatis, and P. Hugenholtz. A Bioinformatician's Guide to Metagenomics. *Microbiology and Molecular Biology Reviews*, 72(4):557–578, December 2008.

[39] E. R. Mardis. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, 9(1):387–402, June 2008.

[40] L. Martins, L. Antunes, R. Pascon, J. de Oliveira, L. Digiampietri, D. Barbosa, B. Peixoto, M. Vallim, C. Viana-Niero, E. Ostroski, G.. Telles, Z. Dias, J. da Cruz, L. Juliano, S. Verjovski-Almeida, A. da Silva, and J. Setubal. Metagenomic analysis of a tropical composting operation at the são paulo zoo park reveals diversity of biomass degradation functions and organisms. *PLoS ONE*, 8(4):e61928, 04 2013.

[41] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, 4(6):495–500, April 2007.

[42] A. McCaig, L. Glover, and J. Prosser. Molecular Analysis of Bacterial Community Structure and Diversity in Unimproved and Improved Upland Grass Pastures. *Applied and Environmental Microbiology*, 65(4):1721–1730, 1999.

[43] A. McHardy, H. Martín, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos. Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, 4(1):63–72, 2007.

[44] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1):386–8, December 2008.

[45] E. Myers, G. Sutton, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, S. Kravitz, C. Mobarry, K. Reinert, K. Remington, E. Anson, R. Bolanos, H. Chou, C. Jordan, A. Halpern, S. Lonardi, E. Beasley, R. Brandon, L. Chen, P. Dunn, Z. Lai, Y. Liang, D. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. Rubin, M. Adams, and J. Venter. A whole-genome assembly of drosophila. *Science*, 287(5461):2196–2204, March 2000.

[46] T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara. Metavelvet: An extension of velvet assembler to *de novo* metagenome assembly from short sequence reads. In *The 2011 ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM-BCB 2011)*, pages 116–124, August 2011.

[47] E. P. Nawrocki, D. L. Kolbe, and S. R. Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10):1335–1337, 2009.

[48] R. Overbeek, T. Begley, R. Butler, J. Choudhuri, H.-Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, Roy. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Rückert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33(17), October 2005.

[49] E.C. Pielou. *Ecological diversity*. Wiley, 1975.

[50] C. Quince, T. Curtis, and W. Sloan. The rational exploration of microbial diversity. *International Society for Microbial Ecology*, 2(10):997–1006, 2008.

[51] J. Raes, K. U. U. Foerstner, and P. Bork. Get the most out of your metagenome: computational analysis of environmental sequence data. *Current Opinion in Microbiology*, 10(5):490–498, October 2007.

[52] J. Raes, E. D. D. Harrington, A. H. H. Singh, and P. Bork. Protein function space: viewing the limits or limited by our view? *Current Opinion in Structural Biology*, 17(3):362–369, June 2007.

[53] L. Raeymaekers. A commentary on the practical applications of competitive PCR. *Genome Research*, 5(1):91–94, 1995.

[54] M. S. Rappé and S. J. Giovannoni. The uncultured microbial majority. *Annual Review of Microbiology*, 57(1):369–394, 2003.

[55] A. Reysenbach, L. Giver, G. Wickham, and N. Pace. Differential amplification of rRNA genes by polymerase chain reaction. *Applied and Environmental Microbiology*, 58(10):3417–1418, 1992.

[56] Daniel C. Richter, Felix Ott, Alexander F. Auch, Ramona Schmid, and Daniel H. Huson. MetaSim-A Sequencing Simulator for Genomics and Metagenomics. *PLoS ONE*, 3:e3373, 2008.

[57] I. Rish. An empirical study of the naive bayes classifier. *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 41–46, 2001.

[58] L. Roesch, R. Fulthorpe, A. Riva, G. Casella, A. Hadwin, A. Kent, S. Daroub, F. Camargo, W. Farmerie, and E. Triplett. Pyrosequencing enumerates and contrasts soil microbial diversity. *International Society for Microbial Ecology*, 1(4):283–290, 2007.

[59] M. Ronaghi. Pyrosequencing sheds light on dna sequencing. *Genome Research*, 11(1):3–11, 2001.

[60] M. Ronaghi, M. Uhlén, and P. Nyrén. A sequencing method based on real-time pyrophosphate. *Science*, 281(5375):363–365, July 1998.

[61] G. Rosen, E. Garbarine, D. Caseiro, R. Polikar, and B. Sokhansanj. Metagenome Fragment Classification Using N-Mer Frequency Profiles. *Advances in Bioinformatics*, 2008:Article ID 205969, 2008.

[62] D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y.-H. H. Rogers, L. I. Falcón, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, 5(3):e77+, March 2007.

[63] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of The National Academy of Sciences of the United States of America (PNAS)*, 74(12):5463–5467, December 1977.

[64] S. Schbath, B. Prum, and E. de Turckheim. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *Journal of Computational Biology*, 2(3):417–437, 1995.

[65] P. Schloss and J. Handelsman. Toward a Census of Bacteria in Soil. *PLoS Compu-tacional Biology*, 2(7):e92, 2006.

[66] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, 2009.

[67] G. Seber. A review of estimating animal abundance. *Biometrics*, 42:267–292, 1986.

[68] T. Sharpton, S. Riesenfeld, S. Kembel, J. Ladau, J. O'Dwyer, J. Green, J.. Eisen, and K. Pollard. PhylOTU: A High-Throughput Procedure Quantifies Microbial Community Diversity and Resolves Novel Taxa from Metagenomic Data. *PLoS Computational Biology*, 7(1):e1001061, 2011.

[69] M. Sogin, H. Morrison, J. Huber, D. Welch, S. Huse, P. Neal, J. Arrieta, and G. Herndl. Microbial diversity in the deep sea and the underexplored rare biosphere. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 103(32):12115–12120, 2006.

[70] J. Staley. Biodiversity: Are microbial species threatened? *Current Opinion in Biotechnology*, 8(3):340–345, 1997.

[71] M. Suzuki and S. Giovannoni. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology*, 62:625–630, 1996.

[72] S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. Comparative Metagenomics of Microbial Communities. *Science*, 308(5721):554–557, April 2005.

[73] G. Tyson, J. Chapman, P. Hugenholtz, E. Allen, R. Ram, P. Richardson, V. Solovyev, E. Rubin, D. Rokhsar, and J. Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, 2004.

[74] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, April 2004.

[75] C. Webb, D. Ackerly, M. McPeek, and M. Donoghue. Phylogenies and Community Ecology. *Annual Review of Ecology and Systematics*, 33:475–505, 2002.

[76] M. Williamson and K. Gaston. The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution. *Journal of Animal Ecology*, 74:409–422, 2005.

[77] J. C. Wooley, A. Godzik, and I. Friedberg. A primer on metagenomics. *PLoS Computacional Biology*, 6(2):e1000667, February 2010.

[78] D. Wu, A. Hartman, N. Ward, and J. Eisen. An Automated Phylogenetic Tree-Based Small Subunit rRNA Taxonomy and Alignment Pipeline (STAP). *PLoS ONE*, 3(7):e2566, 2008.

[79] D. R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, May 2008.