UNIVERSIDADE DE SÃO PAULO

PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Caravela: um navegador para metagenomas

Dissertação de Mestrado

GIANLUCA MAJOR MACHADO DA SILVA

ORIENTADOR: JOÃO CARLOS SETUBAL

São Paulo, SP
Junho, 2017

Resumo

Metagenômica é a técnica que permite analisar os genomas de microorganismos que habitam determinados nichos do ambiente sem a necessidade de isolar e cultivar cada um separadamente. Ao conjunto de microorganismos que habita um determinado nicho se dá o nome de microbioma. Análises do perfil da diversidade taxonômica e funcional de comunidades microbianas em microbiomas são comuns em estudos de metagenômica. No entanto, atualmente as plataformas de uso geral (como MG-RAST e IMG/M) tendem a separar as análises baseadas em reads (sequências não montadas) das baseadas em contigs (sequências montadas), isto dificulta as análises destes dados. Motivado por esta separação, desenvolvemos uma plataforma web, batizada de CARAVELA, que facilita a conexão entre os resultados de análises de diversidade taxonômica e funcional baseadas em reads e contigs respectivamente. Uma das principais funções da plataforma CARAVELA é associar a identificação taxonômica de cada read com o contig que este read faz parte e, anotações funcionais do contig, quando existirem. Essa função deve permitir a rápida identificação de contigs potencialmente quiméricos bem como contigs taxonomicamente bem resolvidos. Também é possível fazer buscas, tais como: listar todos os contigs que tenham um ou mais reads classificados como "*Pseudoxanthomonas suwonensis*" em sua composição e ainda, é possível navegar nos contigs de maneira similar a navegadores de metagenomas tradicionais. Podem ser utilizados como arquivos de entrada a saída de outros programas, desde que o formato atenda certos padrões. A plataforma CARAVELA foi desenvolvida com Java, HTML, CSS, Javascript e Mysql, e com o fim de testar a ferramenta, utilizamos o conjunto de dados metagnômicos obtidos a partir da operação de compostagem do Parque Zoológico de São Paulo.

O ser humano preferirá ainda querer o
nada a nada querer.

Friedrich Nietzsche

Agradecimentos

Em primeiro lugar, gostaria de agradecer a todos os membros da minha família, que de forma direta ou indireta contribuíram para minha formação, principalmente a meus pais, Maurílio e Silvana, que durante todo meu caminho me proporcionaram oportunidades e me suportaram em diversos sentidos. Em especial, gostaria de agradecer minha mãe, Silvana Major Machado, por ser uma mulher com visão à frente do seu tempo, e me apresentar lições fundamentais na formação dos valores que hoje habitam a minha essência.

Quero agradecer a minha namorada, Vanessa Stefanuto, que ao longo dos últimos anos (melhor não dizer quantos) tem sido uma companheira excepcional e uma mulher maravilhosa, que mesmo em momentos difíceis sempre se fez presente e autêntica.

Quero agradecer a todos os colegas e amigos, alguns em especial. Rodrigo Montanholi, por estar sempre presente nas encruzilhadas da vida, por compartilhar suas ideias, valores e cervejas desde os tempos de faculdade, por ser mais que um amigo, por ser um irmão. Wellington Nagasawa, por compartilhar comigo sua visão sobre engenharia de software, política, filosofia, religião e vida, dentre tantos outros temas que debatemos, muitas vezes tendo como testemunhas apenas algumas latas de cerveja e a Av. Luiz Carlos Berrini, ambas já vazias.

Gostaria de agradecer a todos os integrantes do laboratório de bioinformática do Prof. João Carlos Setubal, ou Setulab, com o lab é carinhosamente conhecido. Um obrigado especial aos Alunos de Doutorado Thibério Rangel, por compartilhar pensamentos críticos e de alto nível sempre provocando nossa capacidade de argumentação. Rodrigo Guarischi, por toda paciência e dedicação nas explicações abastadas de detalhes, também por diversas oportunidades me apresentar o contraditório sempre com pleno espaço para o debate e evolução dos pensamentos.

Deyvid Amgarten, por me ajudar a testar a Caravela, a ferramenta desenvolvida neste trabalho e por ser um entusiasta nato, capaz de motivar todos a sua volta.

Por fim, gostaria de agradecer ao Prof. João Carlos Setubal, por me orientar e prover toda estrutura necessária para pleno desenvolvimento deste trabalho. Quero agradecer-lo por toda a dedicação e atenção dispensada durante o período em que estive em seu laboratório e sob sua orientação. Laboratório, inclusive, que é um ambiente de trabalho muito colaborativo e agradável. Obrigado por manter a porta da sua sala sempre entre aberta, por responder e-mails com a agilidade necessária e, especialmente, pelo cuidado na comunicação com seus alunos, sempre nos motivando.

Este trabalho foi realizado com o apoio financeiro da Fundação de amparo à pesquisa do estado de São Paulo (FAPESP) através do projeto temático “Estudos da diversidade microbiana no Parque Zoológico do Estado de São Paulo” (Processo 2011/50870-6) e de uma bolsa de treinamento técnico TT-5 (processo 2013/00536-8); e de uma bolsa de mestrado da Coordenação de aperfeiçoamento de pessoal de nível superior (CAPES), concedida pelo projeto BIGA (Bioinformática, Genômica e Associados) (processo 3385/2013). Esse apoio foi imprescindível para a realização deste trabalho.

Sumário

1	Introdução	1
1.1	O problema	2
1.2	Motivação	2
1.3	Contexto	3
1.4	Relacionamento entre os capítulos	3
2	Fundamentos	4
2.1	Sequenciamento de nova geração	4
2.2	Montagem (<i>genome assembly</i>)	4
2.3	Metagenômica	5
2.4	Análise do perfil taxonômico	6
2.5	Análise do perfil funcional	6
3	Trabalhos relacionados	8
4	Caravela	11
4.1	Arquitetura	12
4.2	Arquivos e formatos	14
4.3	Processamento dos arquivos no PIER	17
4.4	Carregamento de metagenomas na CARAVELA	17
4.5	Persistência de <i>reads</i> com táxon associado	18
4.6	Índice de identificação taxonômica geral	18

4.7	Resolução de nível taxonômico e definição de consistência taxonômica	19
4.8	Distribuição dos táxons ao longo do contig. Definindo cobertura, regiões, cobertura exclusiva, consistência taxonômica vertical e fronteiras	21
4.9	Classificação taxonômica de <i>reads</i> por contexto	26
4.10	Contigs taxonomicamente bem resolvidos e potencialmente quiméricos	28
4.11	Gerando relatório através do filtro de contigs Taxonomicamente Bem Resolvido (TBR) e Potencialmente Quimérico (PQ)	29
4.12	Busca de contigs por táxon ou função	31
4.13	Navegação visual no contig	33
5	Casos de uso do projeto Metazoo	37
5.1	Processamento dos dados metagenômicos do projeto Metazoo	37
5.1.1	Identificação taxonômica dos reads	38
5.1.2	Filtro de qualidade e montagem com SOAPdenovo2	39
5.1.3	Anotação funcional dos contigs com Integrated Microbial Genomes with Microbiome Samples (IMG/M)	40
5.1.4	Mapeamento dos reads nos contigs	40
5.2	Análise visual de contigs formados por reads associados a organismo de interesse	41
5.2.1	Explorando contig associado ao gênero <i>Paracoccus</i>	42
5.2.2	Explorando contig associado ao gênero <i>Geobacillus</i>	43
5.3	Analisando qualidade de contigs montados com <i>SoapDeNovo 2</i> partir das amostras de Zoo Composting 4 (ZC4) e Zoo Composting 3 (ZC3).	47
5.4	Comparando resultado de diferentes montadores	50
5.4.1	Comparando percentual de contigs TBR e PQ formados pelos montadores <i>SoapDeNovo 2</i> e <i>Mira</i>	51
5.4.2	Comparando total de bases TBR e PQ geradas pelos montadores <i>SoapDeNovo 2</i> e <i>Mira</i>	53
5.5	Classificação taxonômica de <i>reads</i> por contexto	56
5.5.1	Aplicação aos <i>reads</i> da compostagem ZC4 e ZC3	57
5.6	Análise do contig <i>lake-12-11_rep_c113861</i> do Lago São Francisco	60

6	Conclusão	63
6.1	Contribuições deste trabalho	64
	Apêndice	65
A	Tabelas	65
A.1	Tabelas utilizadas para produzir gráficos apresentados na seção 5.3.	65
A.2	Tabelas utilizadas para produzir gráficos apresentados na seção 5.4.1.	69
A.3	Tabelas utilizadas para produzir gráficos apresentados na seção 5.4.2.	76
B	Banco de dados	78
B.1	MER	79

Introdução

Os microrganismos, como seu nome diz, são organismos tão pequenos que não podem ser observados a olho nu. Existem microorganismos em todos os grandes ramos da vida: Bactérias, Archaea, Eucariotos e Virus. Os microrganismos se encontram em toda a biosfera: ambientes aquáticos, solo, ar, animais e em especial o oceano, onde eles podem ser responsáveis por até 90% da biomassa total [50]. As atividades de comunidades complexas de microrganismos afetam o ciclo biogeoquímico de diversos ecossistemas.

Os microrganismos desempenham um variedade de funções ambientais e são responsáveis pelo ciclo de muitos nutrientes em nosso ambiente. Por exemplo, eles desempenham um papel importante no ciclo do carbono[22], e também são fundamentais para o ciclo de nitrogênio, que permite que as plantas cresçam. Bactérias são capazes de pegar o nitrogênio do ar e convertê-lo em uma forma que ele possa ser utilizado pelas plantas no solo [18, 13].

Microrganismo também desempenham um papel importante na limpeza de águas de efluentes. Eles transformam materiais químicos e orgânicos que podem ser tóxicos para seres humanos em outras substâncias não tóxicas, promovendo uma limpeza da água [27]. Não é incomum notícias sobre bactérias que “comem” petróleo e foram utilizadas para ajudar a limpar sujeiras produzidas por desastres causados por humanos[19].

A microbiota humana, nome dado ao conjunto de microrganismo que habita o corpo humano, desempenha funções importantes. Por exemplo, os microrganismos que vivem no intestino de uma pessoa podem determinar ou não se alguns medicamentos são tóxicos para o fígado [17]. Alguns problemas de saúde como a obesidade, que já é considerada uma epidemia mundial,

pode estar associada com a composição da microbiota do intestino [49, 33]. Outros problemas como a doença de *Crohn* [36] e câncer [48, 37] também já foram relacionados a comunidade de microrganismo que vivem em nosso corpo.

As pesquisas em ecologia de comunidades microbianas cresceram muito na era da genômica funcional de alto rendimento. Essa tendencia, em grande parte, é devido aos avanços no sequenciamento de DNA, que permitem agora aos pesquisadores investigar a composição e funções das comunidades microbianas em alta resolução, sem a necessidade de cultura [24]. Inclusive porque grande parte desses microrganismos não são cultiváveis.

Elementos importantes para pesquisa em ecologia de comunidades microbianas incluem a análises de vias funcionais e o entendimento dos mecanismos de interação entre as populações microbianas e meio ambiente [31].

1.1 O problema

Análises do perfil da diversidade taxonômica e funcional de comunidades microbianas são comuns em estudos de metagenômica. Entretanto, plataformas/ferramentas de uso comum, como por exemplo o IMG-M [38] e MG-RAST [26], costumam separar as análises baseadas em contigs (IMG-M) das análises baseadas em reads (MG-RAST). Desta forma, em nenhum momento há uma junção dessas análises baseadas em contigs (sequências montadas) com as baseadas em reads (sequências não montadas) e isto dificulta as análises destes dados.

1.2 Motivação

Em estudos de comunidades microbianas é importante associar organismos a funções para que as interações entre as populações microbianas e o meio ambiente seja melhor compreendido. Não temos conhecimento de nenhuma ferramenta que integre as análises baseadas em contigs e reads de maneira fácil e que permita responder perguntas como: “dada uma função anotada em um contig, quais organismos foram identificados nas reads que compõem esse contig?”, ou ainda, “dado um organismo X identificado em reads, quais são as funções anotadas nos contigs que foram montados com reads identificados como sendo do organismo X?”

1.3 Contexto

Este trabalho foi concebido dentro do projeto temático Metazoo que tem por objetivo geral coletar, analisar e prospectar dados moleculares de três microbiomas existentes no Parque Zoológico do Estado de São Paulo (PZSP): compostagem vegetal da mata atlântica, lago, e fezes de macacos bugio. O projeto faz uso de diversas metodologias, principalmente metagenômica e bioinformática. O projeto Metazoo é financiado pela Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) e coordenado pelo Prof. João Setubal, orientador do aluno.

1.4 Relacionamento entre os capítulos

Os capítulos seguintes desta dissertação estão organizados da seguinte forma: Primeiramente, apresentamos no capítulo 2 alguns fundamentos necessários sobre o estudo de comunidades microbianas. No capítulo 3, discorremos sobre diversas ferramentas que representam o estado-da-arte em metagenômica. No capítulo 4, descrevemos a ferramenta construída. No capítulo 5, apresentamos resultados obtidos a partir do uso da ferramenta Caravela com dados do projeto Metazoo. Por fim, apresentamos nossas conclusões no capítulo 6.

Fundamentos

2.1 Sequenciamento de nova geração

O estudo de genomas mudou rapidamente desde o surgimento do chamado sequenciamento de nova geração, conhecido como *next-generation sequencing* (*NGS*). Esses sequenciadores têm a capacidade de produzir um enorme número de reads com custo muito baixo, esses sequenciadores estão acelerando a pesquisa biológica em muitas áreas, tais como genômica, transcritômica, metagenômica, análise de expressão gênica, descoberta de RNA não codificadores, entre outras [21, 24].

As *reads* são sequências curtas de DNA, neste trabalho supomos que as reads têm aproximadamente 250 pares de base, no entanto, esse tamanho pode variar de acordo com a tecnologia de sequenciamento empregada. O conjunto de *reads* resultantes de um sequenciamento pode ser pareado, conhecido na literatura como *paired-end*, o que significa que todo read A deve corresponder a um outro read B, sendo A e B extremidades opostas do mesmo fragmento de DNA.

2.2 Montagem (*genome assembly*)

A necessidade de montar o genoma surge da impossibilidade de sequenciar o genoma completo utilizando apenas uma única *read*. O método de sequenciamento *shotgun* fragmenta o genoma em *reads* de forma aleatória e gera cada *read* independentemente. O processo de re-

construção de um genoma completo, unindo o conjunto de *reads* em um único cromossomo, é conhecido como montagem do genoma [21].

As abordagens de montagem supõem que os fragmentos de DNA com alta similaridade são originários da mesma posição dentro do genoma, de modo que, a similaridade entre as sequências de DNA é utilizada para unir os fragmentos menores (*reads*), em sequências contíguas maiores (*contigs*). Desta forma, recupera-se a informação perdida durante o processo de sequenciamento. O processo de montagem é complicado, pois em muitos casos essa suposição inicial está incorreta. Por exemplo, podem existir repetições genômicas - segmentos de DNA repetidos de forma quase idêntica ao longo do genoma - trazendo dificuldades ao processo automático de montagem [41].

Em metagenomas, o processo de montagem enfrenta dificuldades adicionais, pelo fato de não se saber, de forma antecipada, a origem de cada fragmento de DNA que foi sequenciado. Nesse cenário, o montador pode acabar juntando sequências de diferentes organismos e, dessa maneira, formando *contigs* artefatuais, ou seja, sequências que não existem na natureza, e que são apenas artefatos do processo de montagem, também conhecidos como quimeras artefatuais.

2.3 Metagenômica

Metagenômica é a aplicação de técnicas modernas de genômica para o estudo de comunidades de organismos microbianos diretamente nos seus ambientes naturais, esse processo evita a necessidade de isolamento e cultivo em laboratório de espécies individuais [15]. Atualmente, metagenômica compreende o sequenciamento de milhões de fragmentos genômicos aleatórios, amostrados a partir de uma comunidade microbiana. As sequências de DNA resultantes desse processo são comumente utilizadas para tentar entender dois aspectos da comunidade microbiana. O primeiro é em relação a “*quem são os organismos presentes na comunidade?*”, essa questão pode ser elucidada com a análise de perfil **taxonômico**. O outro aspecto é o funcional, onde tenta-se responder “*o que os organismos da comunidade podem estar fazendo?*”, para responder essa questão utiliza-se a análise de perfil **funcional** [24].

2.4 Análise do perfil taxonômico

A análise de perfil taxonômico tem por objetivo responder a pergunta “*quais são os organismos presentes e qual sua abundância na comunidade microbiana?*”. Esta análise pode ser feita através do sequenciamento de genes marcadores (normalmente se utiliza o gene 16S RNA ribossomal) e/ou sequenciamento do DNA total, amplamente conhecido na literatura como *whole-metagenome shotgun sequencing*.

No caso do sequenciamento do DNA total do metagenoma, tanto as (*reads*), quanto os contigs, podem ser utilizadas para determinar os membros da comunidade microbiana. As sequências obtidas podem ser individualmente comparadas com sequências presentes em um catálogo de genes ou genomas microbianos já conhecidos ou com um catálogo pré-computado de sequências marcadoras de táxon específico.

Por fim, as sequências obtidas podem ser definidas como sendo de um determinado táxon com base em modelos de composição do genoma ou por correspondência exata de k-mers (todas as subsequências de tamanho k possíveis de uma sequência), isso permite que sequências, montadas ou não, sejam classificadas mesmo quando não existem genomas correspondente de referência disponível, algo comum em comunidades mal caracterizadas [24].

O processo de classificação taxonômica de sequências de DNA total de metagenomas, como descrito acima, tende a dar melhores resultados se aplicados nas reads, ao invés dos contigs. Isto porque a read é a leitura de um fragmento de DNA originário de um único genoma. No entanto, as reads precisam ser suficientemente grandes para poder ser classificadas, uma vez que as reads curtas (menores que 50 pares de bases) dificultam a classificação taxonômica.

2.5 Análise do perfil funcional

A pergunta a ser respondida por esta análise é “*O que os organismos presentes podem estar fazendo?*”. Para responder esta questão, pode-se realizar anotações de *reads* ou anotação de contigs. O processo de anotação de sequências de metagenomas é semelhante ao processo de anotação de sequências de genomas isolados. Em uma primeira etapa são preditas as *Coding DNA Sequence (CDS)s*, em uma segunda etapa as sequências das *CDSs* preditas são

processadas por algum pipeline que procura associar uma função molecular a cada *CDS*. Os pipelines geralmente utilizam bases de dados como: Clusters of Orthologous Groups (COGs) [46], Non-supervised Orthologous Groups (NOGs) [43], Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology [30], Pfam [44] e UniProt Reference (UniRef) [45] e produzem perfis de presença e ausência de família de genes, bem como a abundância relativa de cada família dentro de uma mesma amostra metagenômica.

O processo de anotação como descrito acima dá melhores resultados se for realizado sobre os contigs, ao invés das *reads*. Isto ocorre porque os contigs tendem a ser maiores do que as *reads* (pois são produto de montagem das *reads*) e, portanto, contém mais informação. Porém, há o risco de contigs serem quimeras, ou seja, serem junções artefatuais de pedaços de genomas distintos.

Trabalhos relacionados

Na literatura, existem muitos trabalhos relacionados à ferramentas e plataformas para processamento, análise e visualização de dados metagenômicos, algumas aceitam *reads* e/ou contigs como arquivos de entrada, outras necessitam que o usuário execute algum pré-processamento antes de submeter os dados.

O Joint Genome Institute (JGI) disponibiliza a plataforma IMG/M [38] e espera contigs como dados de entrada, o IMG/M processa os dados metagenômicos utilizando um pipeline próprio de anotação, que resulta na análise funcional dos contigs. Os resultados podem ser visualizados, explorados e baixados utilizando a própria plataforma do IMG/M.

Já a plataforma *Metagenomics RAST (MG-RAST)* [26] tem pipeline próprio tanto para análise do perfil funcional quanto taxonômico e, apesar de também aceitar sequências montadas, o padrão de dados de entrada são *reads* e não contigs. O MG-RAST também disponibiliza ferramentas para análise exploratória dos resultados e, como o IMG/M, permite baixar os resultados.

A ferramenta MEGAN[40] é um software *stand-alone* e não uma plataforma web e espera receber como dados de entrada o resultado de um BLAST[9], o que exige do usuário pré-processamento dos dados de entrada, *reads* ou contigs. Essa ferramenta oferece recursos para análise taxonômica e funcional.

Há ainda, ferramentas como Tablet [39], MGAviwer [52] e IVG [47], que surgiram para facilitar a visualização da enorme quantidade de dados gerados por NGS, facilitando a visualização entre contigs, genoma de referência (quando existe) e *reads*. Em alguns casos também é

possível adicionar dados de análise funcional e visualizá-los.

O “*Environmental Laboratory Visualization (ELVIZ)*” [14] é uma ferramenta web interativa que permite explorar visualmente metagenomas montados (contigs) e seus metadados, apresentando uma visão geral da comunidade microbiana, para isso a ferramenta utiliza um gráfico interativo em duas dimensões, como pode ser visto na figura 3.1. Também possui recurso de busca por táxons e por nome de função como, por exemplo, nome de *COG* ou *Pfam*.

É possível analisar as informações tanto do perfil funcional quanto taxonômico do contig na mesma área de visualização, além disso, muitos metagenomas anotados pelo IMG/M já foram automaticamente disponibilizados para análise no ELVIZ e também é possível carregar arquivos em formato tabular e General Feature Format (GFF). Esses arquivos devem vir de um pré-processamento feito pelo usuário. Porém, o ELVIZ não disponibiliza nenhuma ferramenta para ajudar o usuário a gerar esses dados.

Figura 3.1: Interface da ferramenta ELVIZ. Por padrão, o eixo Y do gráfico representa o conteúdo GC, e o eixo X a cobertura média dos contigs. Os contigs são apresentados como círculos, com tamanho que varia de acordo com o número de nucleotídeos de cada contig. Além disso, cada círculo tem uma cor que é associada a um grupo taxonômico previamente atribuído ao contig.

As ferramentas encontradas na literatura oferecem funcionalidades de análise do perfil taxonômico e funcional, baseada em reads ou contigs. No entanto, estas ferramentas não fazem a conexão dos resultados destas duas análises e, portanto, não são capazes de associar funções a táxons e vice-versa. O ELVIZ é capaz de fazer essa associação entre funções e táxons, porém, ambas as análises são baseadas em sequências montadas (contigs), não sendo possível utilizar a classificação taxonômica das sequências não montadas (reads).

Caravela

Neste trabalho apresentaremos a CARAVELA, uma plataforma web capaz de integrar dados de análises de perfil taxonômico e funcional, baseados em *reads* e contigs respectivamente. Uma de suas principais características é associar classificação taxonômica de cada *read* com o contig que esta *read* faz parte. Esta capacidade permite a identificação automatizada de contigs potencialmente quiméricos e também de contigs taxonomicamente coerentes.

Através de sua interface interativa e amigável, a CARAVELA permite navegar nos contigs e visualizar suas anotações funcionais, *reads* que fazem parte de sua composição e táxon associados a cada *read*. Outro recurso importante é a busca por funções ou táxon, onde é possível buscar por todos os contigs que a função *"Reverse transcriptase"* tenha sido anotada, por exemplo, ou ainda, buscar todos os contigs que tenham um ou mais reads classificados como *"Pseudoxanthomonas"* em sua composição.

É importante mencionar que o objetivo da CARAVELA não é de simplesmente facilitar a visualização de dados. Mas também permitir a “curadoria” dos dados ao indicar potenciais erros de montagem e/ou classificação taxonômica.

Notar finalmente que a ferramenta não permite edição manual de qualquer dado. Será de responsabilidade do usuário fazer tais correções com base nos resultados apresentados pela ferramenta.

A fase de projeto da ferramenta foi desenvolvida com base na literatura, principalmente nas publicações de ferramentas para análise de dados metagenômicos, tais como *EBI metagenomics—a new resource for the analysis and archiving of metagenomic data* [29], *Seed: a*

user-friendly tool for exploring and visualizing microbial community data [11] e *Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling* [23], além de todas as outras ferramentas já citadas na seção 3.

Também tomamos como inspiração as interfaces do IMG/M, MG-RAST, EBI e *Taxonomer* e, para melhorar o entendimento da natureza dos dados e funcionalidades pretendidas, fizemos entrevistas e discussões com a equipe do projeto Metazoo.

A seguir, descreveremos os resultados da fase de projeto (que portanto podem ser entendidas tanto como metodologia quanto como resultados).

4.1 Arquitetura

A ferramenta CARAVELA é uma plataforma web e sua arquitetura tem três camadas: interface web, processamento e repositório de dados. A interface web é responsável pela interação com o usuário, recebendo dados de entrada e exibindo o resultado de consultas e relatórios. Qualquer interação na interface web pode gerar requisições para camada de processamento, que por sua vez, cria um processo para tratar a requisição.

O processo poderá ler os dados submetidos, fazer verificações, aplicar regras de alguma funcionalidade específica e fazer requisições à camada de repositório.

A camada de repositório recebe e responde somente requisições feitas pela camada de processamento. Aliás, a camada repositório é a única que tem acesso aos dados, sendo de sua responsabilidade a recuperação e armazenamento dos dados.

O desenvolvimento da CARAVELA foi feito sobre a plataforma Java¹ e faz uso de *frameworks* como Vraptr[8], que além de ajudar no desenvolvimento também implementa o padrão de projeto de software Model View Controller (MVC). Especificamente para o desenvolvimento da interface web, fizemos uso das linguagens HTML5² e JavaScript³, além dos *frameworks* e bibliotecas jQuery [6], Feature Viewer [2] e Bootstrap [1]. Para armazenamento dos dados usamos o sistema gerenciador de banco de dados Mysql Server⁴ que é acessado através da Java

¹<http://www.oracle.com/technetwork/java/javase/>

²<http://www.w3.org/TR/html5/>

³<http://www.w3schools.com/js/>

⁴<http://www.mysql.com/>

Persistence API (JPA) implementada por *Hibernate* [4].

Por se tratar do desenvolvimento de uma plataforma web, fazemos uso do contêiner web Apache Tomcat⁵ para abrigar a ferramenta desenvolvida, além disso, a implantação foi feita em uma máquina virtual (VM) do parque computacional do laboratório do Prof. João Carlos Setubal e, atualmente, a VM está configurada com 4 núcleos de processamento, 12 GB de memória, 150 GB de disco e sistema operacional Ubuntu Server⁶.

Para auxiliar os usuários que desejam submeter dados para a CARAVELA, desenvolvemos o PIER. Este é um pequeno módulo *stand-alone* que também está sobre a plataforma *Java*. No entanto, com uma arquitetura mais simples de apenas uma camada. O PIER é um *Java Archive (JAR)* que pode ser executado a partir de uma estação de trabalho comum⁷ e que tenha *Java 8 (JRE)* instalado. Sua principal função é ler os arquivos resultantes de análises metagenômicas prévias (taxonômica e funcional), recuperar dados de interesse, e gerar um arquivo final que poderá ser enviado ao CARAVELA.

Os arquivos submetidos à CARAVELA devem estar no formato *JavaScript Object Notation (JSON)* com atributos pré-estabelecidos, que serão descritos na seção 4.2. Escolhemos o formato *JSON* por ser um padrão aberto, leve e amplamente utilizado em aplicações web. Além disso, diferentes ferramentas de análises metagenômicas geram diferentes formatos de arquivos de saída e o PIER não será capaz de lidar com todos os formatos.

Contudo, visto que utilizamos *JSON* como formato de arquivo de entrada e, dado que grande parte das linguagens de programação têm bibliotecas disponíveis para manipulação e criação de arquivos neste formato, a adoção do padrão *JSON* deve facilitar que usuários com alguma habilidade em programação escrevam seus próprios *scripts*/programas para lidar com os diversos formatos.

⁵<http://tomcat.apache.org/>

⁶<http://www.ubuntu.com/>

⁷Testamos em uma estação de trabalho com 4 núcleos de processamento e 8 GB de memória

Figura 4.1: Caravela - Visão geral da arquitetura. À direita, os arquivos de taxonômica, contigs, mapeamento, e genes de onde a ferramenta auxiliar Pier extrai informações, gerando um arquivo .json (*sample 1*) ao centro. À esquerda, aos dados da amostra (*sample 1*) são efetivamente submetidos a ferramenta Caravela.

4.2 Arquivos e formatos

A CARAVELA utiliza como dados de entrada o resultado de análises de perfil taxonômico e funcional de comunidades microbianas vindas de outras ferramentas. Espera-se que, para cada amostra, o usuário já tenha os seguintes dados:

1. Reads (resultado do sequenciamento das amostras microbianas);
2. Contigs (resultado da montagem do reads);
3. Identificação taxonômica associada a cada read;
4. Anotação funcional dos contigs (genes encontrados em cada contig e função associada);
5. Mapeamento dos reads nos contigs.

O único dado que não é comumente gerado nas análises que precedem o uso da ferramenta é o Item 5, mapeamento dos reads nos contigs, contudo, gerar esses dados é um processo razoavelmente simples.

Os usuários podem fazer uso do PIER para transformarem alguns formatos de arquivos gerados por outras ferramentas em um arquivo que possa ser submetido à CARAVELA no formato *JSON*.

O PIER está habilitado a receber arquivos nos seguintes formatos:

- *Identificação taxonômica dos reads*: Arquivo tabular com duas colunas; a primeira coluna deve conter a identificação da *read* e a segunda a identificação taxonômica, que será o *taxonomy id* de acordo com a base de dados taxonômica⁸ do National Center for Biotechnology Information (NCBI).
- *Contigs*: Arquivo em formato FASTA com as sequências de DNA dos contigs (sequências montadas);
- *Alinhamento dos reads nos contigs*: Esse arquivo deve conter cada contig (apenas as identificações) e os reads (sequencia e identificação) que mapeiam no contig, além das coordenadas de alinhamento padrão do formato Sequence Alignment/Map (SAM). É importante que o alinhador utilizado reporte múltiplos alinhamentos válidos para a mesma *read*. Saber que uma *read* possui mais de um alinhamento válido significa que o mesmo read pode ter sido utilizado para formar mais de um contig e essa informação é importante para a Caravela.

O formato padrão para alinhamento/mapeamento é o SAM, contudo, o formato que deve ser apresentado ao PIER é o BAM, versão binária do SAM. O arquivo BAM deve estar ordenado e acompanhado do arquivo de índices (.BAI) com o mesmo nome. Buscas e leituras em um arquivo BAM ordenado são mais rápidas que em um arquivo SAM;

- *Anotações funcionais*: Anotações funcionais comumente são feitas por pipelines e podem resultar em diversos arquivos de formatos diferentes. O formato GFF é utilizado para descrever genes e outras características em sequências de DNA, RNA e proteínas. Portanto, costuma ser um padrão nos pipelines de anotação funcional. No entanto, por se tratar de um formato flexível as informações dentro do arquivo podem estar arranjadas de diversas maneiras e, além disso, é possível que identificadores dentro do GFF apontem para outros arquivos externos. Todas essas características torna muito difícil, senão impraticável, a definição de um único arquivo, formato e arranjo que compreenda mais que um pipeline de anotação funcional. Diante de tal contexto, o PIER foi habilitado a trabalhar com a

⁸<http://www.ncbi.nlm.nih.gov/taxonomy>

estrutura de arquivos resultante do pipeline do IMG/M. Tais arquivos estão nos formatos GFF, fasta e tabular.

No final da execução do PIER é gerado um arquivo no formato *JSON*, onde cada linha representa um contig. Os principais atributos de cada contig são:

- *Reference*: Identificador do contig no metagenôma;
- *Sequence*: Sequência de DNA do contig;
- *Features*: Lista de características encontradas no contig (CDSs, rRNA, tRNA);
 - *Type*: Tipo da característica. Por exemplo, CDS e tRNA.
 - *Start*: Posição de início da *feature* na sequência do contig.
 - *End*: Posição final da *feature* na sequência do contig.
 - *phylo distribution*: Linhagem taxonômica atribuída a *feature*.
 - *Gene Product*: Função molecular associada a *feature*.
- *Reads On Contig*: Lista de reads que foram mapeados no contig;
 - *Reference*: Identificador do read no metagenôma.
 - *Start*: Posição de início do alinhamento do read em relação ao contig.
 - *End*: Posição final do alinhamento do read em relação ao contig.
 - *Flag*: *bit flag* do alinhamento. Permite verificar é o *pair* 1 ou 2 do read, por exemplo.
 - *Taxon*: Táxon associado ao read.

O arquivo gerado estará no formato apropriado e com todas as informações necessárias, pronto para ser submetido à CARAVELA.

Algumas características e atributos do arquivo JSON acima apresentado foram omitidas, afim de minimizar a quantidade de informações técnicas não necessárias nesta seção.

4.3 Processamento dos arquivos no PIER

Para o processamento dos arquivos no PIER fizemos uso da *Java API for high-throughput sequencing data (HTS) format (HTSJDK)* [5], biblioteca Java que possibilita ler, navegar e fazer buscas em arquivos no formato FASTA, FASTQ e BAM. A partir do arquivo de contigs, para cada contig, através de seu identificador de referência, buscamos por todas as características funcionais nos arquivos de anotações funcionais, CDSs por exemplo. Buscamos também, agora no arquivo de alinhamento, todos os reads que alinharam no contig e, para cada read, buscamos por táxon a ele associado no arquivo de identificação taxonômica. Todas as informações sobre o contig, como anotações funcionais, reads, mapeamento e táxons serão resumidas e transformadas em uma única entidade no formato *JSON*. Fazemos uso da biblioteca *GSON* [3] nas transformações de objetos JAVA em entidades *JSON*. Esse processo será feito para todos os contigs da amostra. No final desse processamento, todos os contigs da amostra metagenômica submetida, estarão resumidos e transformados em entidades *JSON* em um único arquivo por amostra. Pronto para ser submetido à CARAVELA.

4.4 Carregamento de metagenomas na CARAVELA

Antes que um arquivo de amostra seja submetido à CARAVELA, é preciso ter um projeto e um *treatment* cadastrado na plataforma, que podem ser previamente criados. O processo de carregamento de uma amostra é razoavelmente simples, já que o usuário deve estar logado na plataforma, selecionar o projeto e o *treatment*, cadastrar a amostra e submeter o arquivo gerado pelo PIER para amostra em questão. Imediatamente depois do *upload*, a plataforma inicia o processo de persistência dos dados enviados. Em resumo, esse processo consiste em ler o arquivo submetido, fazendo uso novamente da biblioteca *GSON*, agora para transformar entidades *JSON* em objetos JAVA. Logo após a criação dos objetos *contigs*, *features*, *reads* e *taxons*, tais objetos são persistidos no banco de dados em suas respectivas tabelas, de forma a manter seus relacionamentos. Dentro do processo de persistência dessas entidades, existem dois subprocessos que merecem destaque, a **persistência de reads com táxon associado** e o **cálculo do índice de identificação taxonômica geral** de cada contig.

4.5 Persistência de *reads* com táxon associado

Quando existir um táxon associado a uma *read* nos dados submetidos à Caravela, deve existir um *taxonomy id* para representar o táxon. O *taxonomy id* deve vir da base de dados taxonômica do NCBI. Essa base de dados é formada por um conjunto de nomes e classificações para todos os organismos que estão representados no GenBank⁹.

A Caravela faz uso de uma versão dessa bases de dados taxonômica, carregada durante a fase de configuração da plataforma. Durante o processo de carregamento do metagenoma, quando existe um táxon associado a uma *read*, é feita uma busca na base de dados taxonômica da caravela através do *taxonomy id* e, se a busca tiver sucesso, o táxon é associado a *read*, senão, nenhum táxon é associado a *read* e o *taxonomy id* não encontrado é gravado no *log* de alerta.

É importante dizer que um *taxonomy id* identifica um táxon em um dado nível taxonômico, por exemplo, o *taxonomy id*: 314722 representa um táxon a nível de espécie, *Pseudoxanthomonas suwonensis*, no caso deste exemplo. Em alguns casos, é preciso navegar na estrutura dos táxons, ou seja, se temos a informação de espécie, deveria ser possível saber qual é o gênero.

A Caravela está habilitada a navegar na estrutura taxonômica da base de dados do NCBI, essa capacidade se mostrará especialmente útil na execução algumas funcionalidades descritas ao longo deste documento.

4.6 Índice de identificação taxonômica geral

Através da capacidade de integração entre as sequências montadas (contigs) com as sequências que foram utilizadas na montagem (*reads*) e seus respectivos táxons, calculamos o índice de Identificação Taxonômica Geral (ITG) para cada contig. Esse processo acontece ainda no carregamento do metagenoma e exatamente antes que as informações do contig sejam gravadas na base de dados da CARAVELA.

Definição 1 A *Identificação Taxonômica Geral* do contig é um valor entre 0 e 1 dada

⁹<https://www.ncbi.nlm.nih.gov/genbank/>

pela seguinte fórmula:

$$ITG_i = \frac{rc_i}{r_i} \quad (4.1)$$

Onde: i = contig, rc_i = número de *reads* do contig i classificadas em algum táxon e r_i = número de *reads* do contig.

O *ITG* do contig é importante pois é comum que grande parte dos reads que formam um contig não tem táxon atribuído e através desse índice podemos encontrar contigs que tenham maior razão de número de reads classificados em sua formação. Para fim do cálculo do *ITG*, o nível do táxon associado a *read* não é considerado, não importando se táxon foi classificado a nível de espécie ou reino.

4.7 Resolução de nível taxonômico e definição de consistência taxonômica

Partimos do princípio de que um contig, no qual a grande maioria dos reads que o compõem foram associados ao mesmo táxon, a nível de gênero ou mais específico, é um contig **taxonomicamente consistente** e tem maior chance de ser um fragmento genômico real, portanto, ser um contig bem montado.

Por outro lado, um contig cuja maioria dos reads não esteja associada ao mesmo táxon, tem maior possibilidade de ser um contig artefactual, ou seja, ser fruto de um erro de montagem, também chamado de **quimera artefactual**. Vale ressaltar que também poderia ser erro do classificador taxonômico, ou ainda, eventos de transferência lateral de genes.

O NCBI organiza os táxons em uma árvore hierárquica com diversos níveis taxonômicos. Trabalhamos com sete desses níveis, do mais abrangente para o mais específico, que são: **Reino, Filo, Classe, Ordem, Família, Gênero e Espécie**. Para um contig ser considerado perfeito, da perspectiva de consistência taxonômica, 100% dos reads que o formam devem estar associado ao mesmo táxon a nível de espécie. Porém, esse não é o caso da maior parte dos contigs metagenômicos, principalmente devido a erros de montagem e/ou de classificação taxonômica

dos reads, especialmente a nível de espécie, onde os classificadores taxonômicos costumam errar mais.

Diante deste contexto, adotamos os três níveis taxonômicos mais específicos (espécie, gênero e família) como padrões para nossas análises. Quando é necessário especificar apenas um nível, especificamos gênero, por ser o nível mais próximo de espécie e menos sujeito a erros de classificação taxonômica.

Para encontramos contigs taxonomicamente consistentes precisamos analisar a classificação taxonômica das *reads* que o formam. Nesse sentido, calculamos o índice de consistência taxonômica com base nos táxons associados as *reads* do contig e no número de reads atribuídos a cada táxon. Esse índice nos permite encontrar o táxon com maior número de *reads* atribuídas e também a fração que esse táxon representa dentro do total de reads formadores do contig.

Definição 2 A *Consistência Taxonômica (CT)* do contig é um valor entre 0 e 1 dada pela seguinte fórmula:

$$CT_i^l = \max_{t \in T_l} \left(\frac{rc_i^t}{r_i} \right) \quad (4.2)$$

Onde: i = contig, l = nível taxonômico específico, T_l = é o conjunto de todos os táxons existentes no nível l , t = táxon presente em i no nível l , rc_i^t = número de *reads* do contig i classificadas no táxon t e r_i = número de *reads* do contig.

Para exemplificar a aplicação do índice de *CT*, considere um contig hipotético formado por 100 *reads*, das quais 60 foram classificadas a nível de espécie, foram 40 *reads* associadas a *Pseudomonas putida* e 20 *Pseudomonas aeruginosa*. Das *reads* iniciais, 30 foram classificadas apenas a nível de gênero, todas como *Pseudomonas*, e 10 não foram classificadas. Ao verificar a *CT* a nível de espécie neste contig hipotético, observa-se o valor de 0,4 associado ao táxon *Pseudomonas putida* - (A). Portanto, com base na contagem das *reads* classificadas a nível de espécie, 40% do contig é composto *Pseudomonas putida*. Ao verificar a *CT* a nível de gênero, esse valor passa a ser de 0,9 para o gênero *Pseudomonas* - (B). Sendo assim, 90% do contig é taxonomicamente consistente para o gênero *Pseudomonas*, com base na contagem dos reads associados ao táxon.

A)

$$CT_{\text{hipotetico}}^{\text{especie}} = \frac{40}{100} = 0,4 \quad (4.3)$$

B)

$$CT_{\text{hipotetico}}^{\text{genero}} = \frac{90}{100} = 0,9 \quad (4.4)$$

4.8 Distribuição dos táxons ao longo do contig. Definindo cobertura, regiões, cobertura exclusiva, consistência taxonômica vertical e fronteiras

A definição de contigs taxonomicamente consistentes não é suficiente para caracterizar plenamente um contig, uma vez que é preciso levar em conta também a distribuição dos reads ao longo do contig. Por exemplo, seria perfeitamente possível termos um contig em que 70% dos reads associados a um táxon estão de um lado do contig e 30% dos reads associados a um outro táxon estão de outro lado. Esse contig seria um ótimo candidato a ser considerado quimérico, apesar de ser 70% consistente de acordo com nossa definição 2.

Figura 4.2: Representação gráfica do contig (ilustração). **a.** A barra vermelha representa um contig de 1.5k, o tamanho e posição relativa ao contig são apresentados na parte superior. **b.** Alinhamento dos reads que compõe o contig de referência, os reads com a mesma cor estão associados com a mesma classificação taxonômica. **c.** Exibe a associação entre táxon e cor, a nível de gênero, por exemplo: todos os reads da cor azul foram associados à *Rhodothermus* enquanto os de cor cinza não foram classificados. **d.** CT - índice de consistência taxonômica do contig.

O oposto também pode ocorrer, pois um contig pode ter baixo valor de consistência taxonômica ($CT < 0.5$) e, entretanto, as *reads* encontram-se bem distribuídos ao longo de toda extensão do contig, sugerindo que se trata de um contig genuíno (ver exemplo na Fig. 4.2).

Com o propósito de identificar quanto cada táxon representa ao longo da extensão do contig, verificamos quantas bases do contig foram associadas a cada táxon e, para esse fim, definimos o conceito de **cobertura**.

Definição 3 A **Cobertura (CO)** é um valor entre 0 e 1 calculada para cada táxon presente no contig. É dada pela seguinte fórmula:

$$CO_i^l = \frac{b_i^t}{b_i}, t \in T_l \quad (4.5)$$

Onde: i = contig, l = nível taxonômico específico, T_l = é o conjunto de todos os táxons existentes no nível l , t = táxon presente em i no nível l , b_i^t = total de bases associadas ao táxon t no contig i , b_i = total de bases do contig i .

O conceito de cobertura, da forma que está definido, apresenta uma visão geral sobre a presença de táxons no contig, contudo, é uma visão limitada. Visto que, quando existe mais de um táxon no contig, não é possível detectar a presença e tamanho das sobreposições entre tais táxons.

Para lidar com estas situações, ao analisamos a distribuição dos táxons ao longo da extensão do contig, além de verificamos qual táxon foi atribuído a cada *read* formador de cada base do contig, também consideramos que cada base do contig é formada por uma ou mais *reads*, e cada *read* pode estar ou não associada a um táxon. Como resultado, quando a base do contig é formada por mais de uma *read*, é possível que existam diferentes táxons formando a mesma base. Dessa forma, definimos que um contig pode ser formado por três tipos de base, **não classificada (n)**, **classificada (c)** e **indefinida (i)**.

Definição 4 **base não classificada (Bn)** é formada exclusivamente por *reads* sem táxon associado.

Definição 5 **base classificada (Bc)** é formada por uma ou mais *reads* associadas ao mesmo táxon. Uma base será considerada classificada mesmo que haja *reads* formadores não classificados, desde que exista ao menos uma *read* classificada.

Definição 6 *base indefinida (Bi)* é formada por dois ou mais reads associados a diferentes táxons. Permite-se bases não classificadas.

Figura 4.3: Ilustração da formação dos tipos de base no contig - a) Contig. b) reads que formam o contig. As cores verde e azul representam diferentes táxons. A cor cinza representa reads não classificados. c) Os três tipos de base, **não classificada (n)**, **classificada (c)** e **indefinida (i)**

É importante notar que as definições de base **não classificada**, **indefinida** ou **classificada**, bem como todas as definições subsequentes que delas dependem, estão diretamente conectadas ao nível taxonômico que estamos analisando as reads formadoras do contig. Portanto, a mesma base pode ser tipificada de forma diferente em níveis taxonômicos distintos.

Através das definições [4, 5 e 6] podemos definir **regiões taxonômicas** no contig.

Definição 7 *região taxonômica (RT)* é um segmento de bases consecutivas, composto exclusivamente pelo mesmo tipo de base dentro do contig. As regiões taxonômicas podem ser tipificadas de acordo com as bases que as formam e portanto:

- (i) Uma região formada apenas por bases não classificadas é uma **região taxonômica não classificada (RTn)**;
- (ii) Uma região formada apenas por bases indefinidas é uma **região taxonômica indefinida (RTi)**;
- (iii) Uma região formada apenas por bases classificadas é uma **região taxonômica classificada (RTc)**.

O tamanho das regiões taxonômicas indefinida e não classificada podem ajudar na caracterização do contig, dado que, o tamanho da **região indefinida** indica o quanto do contig é

confuso, ou seja, foi associado a mais de um táxon, tornando-o menos consistente taxonomicamente. Já o tamanho da **região não classificada** indica quanto do contig é desconhecido, com base na classificação taxonômica das *reads* formadoras do contig. Neste cenário, podemos calcular a fração que cada uma dessas regiões taxonômicas representa no contig.

Definição 8 A **Fração da Região Taxonômica Não Classificada (FRTN)** é um valor entre 0 e 1 dada pela seguinte formula:

$$FRTN_i^l = \frac{BN_i^l}{b_i} \quad (4.6)$$

Onde: i = contig, l = nível taxonômico específico, BN_i^l = total de **bases não classificadas** (definição 4) no contig i em nível l e b_i = total de bases do contig i .

Definição 9 A **Fração da Região Taxonômica Indefinida (FRTI)** é um valor entre 0 e 1 dada pela seguinte formula:

$$FRTI_i^l = \frac{BI_i^l}{b_i} \quad (4.7)$$

Onde: i = contig, l = nível taxonômico específico, BI_i^l = total de **bases indefinidas** (definição 6) no contig i em nível l e b_i = total de bases do contig i .

No caso da **região taxonômica classificada**, não é suficiente observar apenas o seu tamanho, pois toda região classificada está associada a um táxon e o contig pode conter mais de uma região classificada que não estejam, necessariamente, associadas ao mesmo táxon. Portanto, afim de verificar qual o tamanho da região classificada associada a cada táxon, definimos **cobertura exclusiva**.

Definição 10 A **Cobertura Exclusiva (CE)** é um valor entre 0 e 1 calculada para cada táxon presente no contig. É dada pela seguinte fórmula:

$$CE_i^l = \frac{BC_i^t}{b_i}, t \in T_l \quad (4.8)$$

Onde: i = contig, l = nível taxonômico específico, T_l = é o conjunto de todos os táxons existentes no nível l , t = táxon presente em i no nível l , BC_i^t = total de **bases classificadas** (definição 5) associadas ao táxon t no contig i e b_i = total de bases do contig i .

Um contig pode conter nenhuma, uma ou mais regiões classificadas e, por definição, tais regiões classificadas no mesmo contig devem estar separadas por uma região **não classificada** ou **indefinida**. Quando encontramos regiões não classificadas ou indefinidas que separam duas regiões classificadas associadas a diferentes táxons, tal região é chamada de **fronteira**.

Definição 11 *fronteira é uma região taxonômica não classificada ou indefinida flanqueada por regiões classificadas associadas a diferentes táxons.*

As fronteiras são indicadores de contigs potencialmente quiméricos. Sua presença evidencia a existência de duas regiões taxonômicas associadas a diferentes táxons, tornando o contig mais próximo de uma quimera que de um fragmento genômico real.

Figura 4.4: Representação gráfica da formação do contig por sobreposição de reads (ilustração). a) Sequência de DNA do contig. b) Conjunto de reads formadores do contig. As cores representam a associação taxonômica. c) Regiões taxonômicas não classificadas, indefinidas e classificadas. d) Região rotulada como fronteira. e) Associação entre táxon e cor, total de reads e bases do contig. Nível de gênero.

Por fim, podemos identificar quantos diferentes táxons foram associados as **regiões taxonômicas classificadas** e contabilizar quanto do contig, o táxon com maior número de bases classificadas representa. Espera-se que um contig bem montado tenha um único táxon associado a todas as suas bases ao longo do seu comprimento e, para mensurar essa característica, calculamos o índice de **consistência taxonômica vertical**.

Definição 12 *Consistência Taxonômica Vertical (CTV)* é um valor entre 0 e 1 dada pela seguinte fórmula:

$$CTV_i^l = \max_{t \in T_l} \left(\frac{BC_i^t}{b_i} \right) \quad (4.9)$$

Onde: i = contig, l = nível taxonômico específico, T_l = é o conjunto de todos os táxons existentes no nível l , t = táxon presente em i no nível l , BC_i^t = total de **bases classificadas** (definição 5) associadas ao táxon t no contig i e b_i = total de bases do contig i .

A CTV revela qual a fração do total de bases do contig foi atribuída ao táxon com maior número de **bases classificadas** presente no contig, em outras palavras, dizer que um dado contig tem $CTV = 0.7$ em nível de gênero, significa que 70% das bases desse contig foram associados exclusivamente a um táxon a nível de gênero. Por exemplo, se calculássemos o índice de CTV da figura 4.4, o índice de CTV seria 0.508 para gênero (*Nocardiopsis*).

Ao analisarmos a distribuição dos táxons ao longo do contig e identificarmos o tamanho das regiões **não classificadas**, **indefinidas**, presença de **fronteiras** e a **consistência taxonômica vertical**, podemos caracterizá-lo de forma mais completa, evidenciando características que podem ser utilizadas para rotular o contig como mais próximo de ser uma quimera ou um fragmento genômico real.

4.9 Classificação taxonômica de *reads* por contexto

Contigs com **regiões taxonômicas classificadas** podem ser fontes para identificação de *reads* sem classificação taxonômica, mas que poderiam ser classificados por contexto.

Classificadores taxonômicos geralmente analisam cada read de maneira independente e dependem da composição e sequência de bases em cada *read*, além da base de dados de referência. É comum que muitos reads não sejam classificados por falta de informação na sequência, seja por conta do tamanho do read muito curto ou por se tratar de um organismo mal caracterizado. Neste último caso, com pouca ou nenhuma informação na base de dados de referência.

As **regiões taxonômicas classificadas** (definição 7) permitem a presença de **bases não classificadas** (definição 4), portanto, é perfeitamente possível que existam *reads* não clas-

Figura 4.5: Contig com *reads* classificáveis por contexto. **a)** Contig. **b)** *Reads* formadores do contig associados a seus respectivos táxon por cores. **c)** Regiões taxonômicas. **d)** *Reads* não classificadas completamente contidas em região classificada e associada ao gênero *Rhodothermus* **e)** Táxons, suas respectivas cores e total de *reads* associadas.

sificados completamente contidos dentro de regiões classificadas. Uma *read* nessa condição, certamente se alinha a uma ou mais *reads* associadas a um único táxon e, neste contexto, é razoável associá-la ao táxon da região classificada.

Definição 13 *Read classificado por contexto* é um *read* não classificado, mapeado em um único contig, completamente contido em uma região taxonômica classificada do contig.

A Caravela pode gerar relatório analítico dos *reads* classificados por contexto para cada amostra processada. Esse relatório é sempre a nível taxonômico de **gênero** e são reportados as seguintes informações: *referência do read (id)*, *nome científico*, índices de *CT*, *CTV* e número de *fronteiras* do contig que o *read* faz parte. Essas informações pode ajudar o usuário a decidir se aceita ou não essa sugestão de classificação taxonômica dada pela ferramenta. A Tabela 4.1 exemplifica as estrutura do relatório com algumas linhas.

Tabela 4.1: Exemplo de relatório da classificação taxonômica de reads por contexto

read reference	scientific name	CT	CTV	borders	contig reference
M01677:13-A6YLB:1:1112:9428:10513	Pseudomonas	0.577	0.802	1	Ga0131843_100891
M01677:8-A4JA9:1:2114:13811:6472	Thermobispora	0.579	0.5	2	Ga0131843_100893
M01677:13-A6YLB:1:2113:26647:21417	Acinetobacter	0.023	0.439	0	Ga0131843_101340
M01677:8-A4JA9:1:2109:23007:5514	Pseudomonas	0.577	0.802	1	Ga0131843_100891
M01677:8-A4JA9:1:2110:9042:12591	Streptomyces	0.112	0.379	2	Ga0131843_101344

4.10 Contigs taxonomicamente bem resolvidos e potencialmente quiméricos

Através de arranjos apropriados nos valores dos índices de *ITG* (definição 1), *CT* (definição 2), *CTV* (definição 12) e número de **fronteiras** (definição 11), podemos usar a Caravela para encontrar contigs com maior consistência taxonômica e, portanto, mais próximos de serem um fragmento genômico real. Esses serão chamados de contigs **taxonomicamente bem resolvidos**. O contrário também é possível, ou seja, ajustando os valores dos mesmos índices, também é possível encontrar contigs com baixa consistência taxonômica e, nesse caso, **potencialmente quiméricos**. Para encontrarmos tais contigs, é preciso definir valores aos índices e para esse objetivo definimos:

Definição 14 *o contig **TBR** deve ter os índices de $ITG \geq 0.5$, $CT \geq 0.4$, $CTV \geq 0.7$ e **fronteira** = 0*

Definição 15 *o contig **PQ** deve ter os índices de $ITG \geq 0.5$, $CT \leq 0.2$, $CTV \leq 0.4$ e **fronteira** ≥ 1*

Os conceitos TBR e PQ estão diretamente associados ao nível taxonômico em que analisamos os contigs, de forma que, é perfeitamente possível que o mesmo contig seja PQ ao nível de espécie, entretanto, TBR ao nível de gênero.

É importante notar que os valores das definições 14 e 15 foram alcançados de forma empírica, com base na experiência da utilização da Caravela em versões preliminares com dados metagenômicos do projeto *Metazoo*.

Consideramos que essas definições são conservadoras e podem ser tomadas como referência para outros conjuntos de dados metagenômicos. Contudo, deve-se considerar o ajuste desses valores frente ao conjunto de dados e/ou métodos de classificação taxonômica diferentes daqueles usados neste trabalho.

4.11 Gerando relatório através do filtro de contigs TBR e PQ

Uma das formas de explorar a amostra metagenômica na Caravela é através da funcionalidade de filtro de contigs por valores de *ITG*, *CT*, *CTV*, número de fronteiras e de *features*, permite à Caravela encontrar contigs TBR e PQ. Por padrão temos dois filtros, TBR e PQ, com valores já definido conforme a seção 4.10. A figura 4.6 é um screenshot do formulário de filtro com os valores padrões já preenchidos. É perfeitamente possível que o usuário mude esses valores, flexibilizando ou restringindo as definições de contigs TBR e PQ.

The image shows two overlapping screenshots of a web interface for filtering contigs. The top panel is titled 'Filter By contigs TBR:' and contains the following fields: ITG: >= 0,5; Genes: >= 0; Rank: GENUS (dropdown); Borders: <= 0; CT: >= 0,4; CTV: >= 0,7. The bottom panel is titled 'Filter By contigs PQ:' and contains: ITG: >= 0,5; Genes: >= 0; Rank: GENUS (dropdown); Borders: >= 1; CT: <= 0,2; CTV: <= 0,4. Both panels have a green 'Filter' button at the bottom right.

Figura 4.6: Screenshot do formulário de filtro para contigs **TBR** e **PQ**.

O campo *Rank* do formulário refere-se ao nível taxonômico que o filtro deve atuar. São três as opções: Espécie, Gênero e Família. Por padrão, utilizamos o nível de Gênero. Embora o campo *genes* (número de *features*) do contig não faça parte das definições de TBR e PQ, quando os dados funcionais são submetidos à Caravela, também é possível utilizá-lo como filtro. O resultado do filtro pode ser uma lista de contigs navegáveis ou um relatório estático

em formato texto.

O relatório estático faz parte da capacidade analítica da ferramenta, ele é gerado em formato tabular e reporta valores dos índices de CT , CTV e número de fronteiras para os níveis taxonômicos de espécie, gênero e família. A Tabela 4.2 exemplifica a estrutura do relatório com algumas linhas de exemplo. Algumas colunas (gênero e família) foram omitidas por questão de espaço.

Tabela 4.2: Exemplo de relatório de contig taxonomicamente bem resolvidos. Apenas as colunas com índices CT , CTV e número de fronteira (ct|ge, ctv|ge e border|ge) a nível de gênero estão sendo exibidas neste exemplo. As colunas com essas mesmas informações para espécie e família foram omitidas por questão de espaço. As colunas nr e nrc significam número de reads e reads classificadas respectivamente.

reference	size	...	ct ge	ctv ge	border ge	...	itg	nr	nrc	features
100270	9091	...	0.803	0.942	0	...	0.908	478	434	6
100532	6465	...	0.548	0.726	0	...	0.688	352	242	10
100765	5346	...	0.792	0.921	0	...	0.833	168	140	7
100830	5117	...	0.677	0.836	0	...	0.735	291	214	7
100838	5091	...	0.698	0.764	0	...	0.831	308	256	6

Além do relatório de contigs **TBR** (Tabela 4.2), também é possível gerar um relatório complementar, o relatório de táxons no contig. Através desse relatório é possível identificar a quantidade de reads e percentual de número de bases (*cobertura*) que foram associadas a cada táxon a nível(*rank*) de espécie, gênero e família para cada contig.

Tabela 4.3: Exemplo de relatório de táxons no contig taxonomicamente bem resolvidos.

reference	number of reads	coverage	rank	scientific name
101303	4	0.131	species	<i>Bifidobacterium thermophilum</i>
101303	41	0.426	species	<i>Bifidobacterium animalis</i>
101303	9	0.178	species	<i>Bifidobacterium pseudolongum</i>
101303	79	0.872	genus	<i>Bifidobacterium</i>
101303	79	0.872	family	<i>Bifidobacteriaceae</i>
101340	2018	1	species	<i>Acinetobacter baumannii</i>
101340	4	0.126	species	<i>Acinetobacter sp. 826659</i>
101340	8	0.08	species	<i>Acinetobacter oleivorans</i>

É importante que ambos os relatórios, contigs TBR (Tabela 4.2) e táxons no contig (Ta-

bela 4.3), sejam gerados com os mesmos valores no filtro. Caso isso não seja verdade, pode ocorrer que um contig reportado em um relatório não esteja reportado no outro. Por exemplo, caso os valores definidos no filtro do relatório de táxons no contig sejam mais restritivos, em relação aos valores geradores do relatório de contigs TBR, é possível que nem todos os contigs reportados em contigs TBR tenham seus táxons no relatório de táxons no contig.

Através desses relatórios analíticos, a Caravela entrega ao usuário todas as informações geradas para cada contig. Informações tais como índices de *CT*, *CTV*, número de fronteiras, número de reads e cobertura de cada táxon para três níveis taxonômicos (espécie, gênero e família). Além do índice de *ITG*, tamanho do contig em pares de bases, número de reads totais e classificadas, e número de *features*. Essas informações podem ser baixadas, possibilitando que o usuário possa fazer novas análises e/ou processamentos.

4.12 Busca de contigs por táxon ou função

a.1 Search by taxon on contig

Scientific name:

Taxon coverage - (between 0 and 1) - Less or equals:

☐ Exclusively

a.2 Find

b Taxons found

Rank	Scientific name	Total
genus	Geobacillus	57
species	Geobacillus sp. WCH70	22
species	Geobacillus sp. Y4.1MC1	18
species	Geobacillus thermoglucosidasius	14

c Search by contigs where at least 50% of contig was mapped by reads assigned to taxon **Geobacillus**.
Number of contig found: 57

Top 10 - functions found in contigs features

Total	Product name	Product source
5	ABC-type Fe3+-siderophore transport system, permease component	COG0609 Q
4	Transcriptional regulators	COG1846 Q
3	Acyl-CoA dehydrogenases	COG1960 Q
2	Amino acid transporters	COG0531 Q
2	Predicted hydrolase of the alpha/beta superfamily - COG2819	COG2819 Q

Contigs

Reference	Size	reads	reads classified	features	TII	Action
ZCTH0002_134575	8947	1658	1212	7	0,731	Q

Figura 4.7: Screenshot Caravela, busca de contigs por táxon: **a.1)** Formulário de busca com os campos nome científico e cobertura (definição 4.5) mínima do táxon a ser buscado. **a.2)** Opção *Exclusively* quando marcada, exigem que o valor definido na cobertura (*taxon coverage*) seja exclusivo (cobertura exclusiva, definição 4.8), ou seja, não pode existir nenhum outro táxon se sobrepondo na mesma região do contig. **b)** Lista resultante da busca pelo táxon *Geobacillus* com cobertura mínima de 50%. Foram encontrados táxons em diferentes níveis. **c)** Resultado da busca pelo gênero **Geobacillus**. As *10 funções que mais aparecem associadas a *features* em contigs com 50% das suas bases associada a esse gênero **Geobacillus**. Lista de *contigs resultantes, essa lista é navegável, sendo possível clicar e visualizar cada contig.

*A lista de funções e contigs foi limitada (cortada) para caber na imagem.

Na Caravela, uma das formas de explorar a amostra de microorganismos é através da busca

de contigs por táxon. Devido a sua capacidade de identificar **regiões taxonômicas classificadas** (definição 7) no contig, a Caravela possui uma funcionalidade poderosa de busca por táxon. É possível encontrar contigs que tenham regiões taxonômicas associadas a qualquer táxon nos níveis de espécie, gênero e família. Além do nome científico do táxon, é possível definir um valor de cobertura mínima do táxon ao longo do contig. Essa busca é especialmente poderosa pois, além de gerar uma lista com os contigs que atendem os requisitos da busca, também é capaz de responder rapidamente ao usuário, quais funções estão associadas ao táxon buscado. Essa funcionalidade se deve a capacidade de integrar informações taxonômicas e funcionais da ferramenta.

No exemplo, na figura 4.7 (a), estamos buscando por contigs em que ao menos 50% das suas bases foram associados ao táxon *Geobacillus*. O resultado dessa busca poder ser observado na figura 4.7 (b), indicando que existem táxons em diferentes níveis (gênero e espécie) que atendem aos requisitos da busca. Ao clicarmos no gênero *Geobacillus*, somos levados a uma nova tela de resultados, figura 4.7 (c), onde são exibidas duas listas; a primeira, refere-se às 10 funções que mais foram associadas a *features* presentes em contigs que atendem aos requisitos da busca. A segunda, refere-se a lista de contigs que atendem os requisitos da busca, sendo possível acessar cada um dos contigs através do botão visualizar.

Quando as *features* e suas respectivas anotações funcionais são enviadas à Caravela, essas informações são associadas aos seus respectivos contigs e por meio dessas associações é possível buscar por contigs que tenham *features* associadas a uma dada função molecular. Tais funções são procedentes de bancos de dados como COGs, NOGs, KEGG Orthology, Pfam e UniRef. A busca pode ser feita pelo nome ou código da função. O resultado dessa busca é, além da lista de contigs que atendem os requisitos da busca, uma lista com os nomes científicos dos 10 táxons que mais vezes foram associados aos contigs resultantes da busca. Essa é uma forma da Caravela informar ao usuário quais táxons estão mais associados a função buscada. Um exemplo de busca por função está ilustrado na figura 4.8.

The screenshot displays the Caravela web interface with three main sections labeled a.1, a.2, and b. Section a.1 shows a search form for 'deaminase' with source 'COG0550'. Section a.2 shows the results for 'deaminase', listing 'Cytosine deaminase and related metal-dependent hydrolases' (COG0402) and 'L-serine deaminase' (COG1760). Section b shows a search form for 'Topoisomerase IA' with source 'COG1760'. Section c shows the results for 'COG1760', including the number of contigs found (1), a list of top 10 taxa found in contig features, and a list of contigs with associated features.

a.1 Search by function on contig

Name: Find

Source: Find

a.2 Function found

Product Name	Product Source
Cytosine deaminase and related metal-dependent hydrolases	COG0402 Q
L-serine deaminase	COG1760 Q

b Search by function on contig

Name: Find

Source: Find

c Search by function: COG1760

Number of contigs found with this function: 1

Top 10 - taxons found in contig features

Total	Rapk	Scientific Name
240	species	Geobacillus thermoglucosidasius Q
184	species	Geobacillus sp. Y4.1MC1 Q
110	genus	Geobacillus Q
10	species	Geobacillus sp. WCH70 Q
6	species	Bacillus anthracis Q
4	species	Enterococcus faecalis Q
2	species	Tetragenococcus halophilus Q

Contigs

Reference	Size	reads	reads classified	features	TII	Action
ZCTH0002_134539	6098	948	556	15	0,586	Q

Figura 4.8: Screenshot Caravela - **a.1** Formulário de busca por contigs que tenham alguma *feature* associada a função de nome *deaminase*, o quadro **a.2** exibe uma lista com o resultado dessa busca e botões em forma de lupa, que levam ao resultado final. O quadro **b** exibe o formulário de busca por código da função. **c)** Resultado final da busca por função, exibindo o código da função, número de contigs, lista com lista de 10 táxons que mais vezes apareceram nos reads desses contigs, e lista dos contigs com *features* associadas a função buscada, além de um botão em de visualização que possibilita a inspeção do contig.

4.13 Navegação visual no contig

É possível que o usuário queira inspecionar a composição de um contig de forma individual, explorando e analisando como os reads estão distribuídos e quais foram as *features* e anotações funcionais associadas ao contig. Ferramentas para esse fim não são novidade: Tablet [39], MGAviwer [52] e IVG [47] são exemplos. No entanto, a visualização de contigs da Caravela traz possibilidades diferentes. A maior delas, é evidenciar visualmente através de diferentes cores, a distribuição taxonômica associada aos reads ao longo do contig e, identificar visualmente as **regiões taxonômicas classificadas, indefinidas, não classificadas e fronteiras** existentes no contig, sendo possível alternar a visualização em diferentes níveis taxonômicos.

A área de visualização do contig é interativa, começando pelo painel de resolução taxonômica exibido na figura 4.9b, onde é possível verificar o tamanho total da região taxonômica não classificada (*Region Size Unclassified*) e indefinida (*Region Size Undefined*), além dos índices de *CT* e *CTV*. Todas essas informações são relativas ao nível taxonômico corrente, no caso

Figura 4.9: Screenshot Caravela - **Visualização do contig na Caravela ao nível de gênero.** **a.** Painel com informações gerais sobre o contig. **b.** Painel de índices relativo a taxonomia das *reads* do contig ao nível de gênero. **c.** Grupo de botões que permitem alterar o nível de resolução taxonômica da visualização do contig. **d.** Informações sobre o nível de *zoom*, posição e base (DNA) do contig. **e.** *Features* (CDSs) e anotações funcionais. **f.** A barra vermelha representa o contig, seu tamanho e posição relativa ao contig são apresentados na parte inferior da imagem. **g.** Alinhamento dos reads que compõe o contig, os reads com a mesma cor estão associados com a mesmo táxon. A cor cinza é reservada para reads não classificados. **h.** Área de sobreposição de reads associados a diferentes táxons. Para cada táxon sobreposto é adicionado uma nova coluna, na região exata de sobreposição, tornando fácil a identificação de áreas com mais de dois táxons sobrepostos. **i.** Regiões indefinidas. **j.** Regiões não classificadas. **k.** Fronteiras.

do exemplo da figura 4.9, gênero. Um pouco abaixo, figura 4.9c, apresenta os botões espécie, gênero, família, ordem, classe, filo e reino que possibilitam alternar a visualização do contig entre esses diferentes níveis. Ainda na área de visualização do contig, também é possível interagir com as *features*, figura 4.9e. Quando clicada, a *feature* exibe seus detalhes, que dizem respeito ao processo de anotação funcional como descrito na seção 2.5, quando existem anotações, é comum que sejam funções moleculares e táxon associados a *feature*. Um exemplo de detalhe de *feature* por ser visto na figura 4.10

Para exibirmos as reads ao longo do contig utilizamos a posição de alinhamento de cada read em relação ao contig e, também, associamos diferentes cores a cada táxon para evidenciar quais reads foram associados a cada táxon. Dessa forma, ao observamos a área de visualização dos reads no contig, figura 4.9g, fica claro que existem dois diferentes táxons no nível de gênero,

representados pelas cores verde e rosa. A cor cinza é reservada para representar reads sem nenhum táxon associado. Sendo assim, é possível fazer uma análise da distribuição taxonômica dos reads que formam o contig de maneira rápida e intuitiva. Além disso, apenas clicando no read, é possível visualizar detalhes sobre o seu alinhamento, CIGAR¹⁰ por exemplo, e atribuição taxonômica, como pode ser visto no exemplo da figura 4.11

A representação gráfica do contig provida pelo Caravela permite ao usuário identificar visualmente de forma rápida regiões de sobreposição de reads associadas a diferentes táxons, ou seja, regiões indefinidas, assim como as regiões não classificadas e fronteiras em diferentes níveis taxonômicos, figura 4.9h, i, j e k. Essa capacidade transforma a visualização de contigs da Caravela em um poderoso recurso visual para ajudar a caracterizar contigs como sendo fragmentos genômicos reais ou artefatuais.

Gene Product assigned: Transcriptional regulator **Source:** COG0583

IMG PhiloDist - Lineage: Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Alcaligenaceae;Bordetella;Bordetella bronchiseptica;Bordetella bronchiseptica Bbr77 **Identity:** 87.5

Type: CDS **Start:** 675 **End:** 1622 **Strand:** 1

a

Annotations

Type	Name	Identity	Align Length	Query Start	Query End	E-value	Bit Score
KO	KO:K13634	87.5	312	1	312	1.46e-155	550
COG	COG0583	29.22	308	1	306	4e-31	128
PFAM	pfam03466	100	214	88	301	2.7e-40	128
PFAM	pfam00126	100	62	3	64	2.4e-15	46.4

b

Figura 4.10: Screenshot Caravela - **Detalhes da feature** - **a)** Informações como tipo de *feature* e posição de início e fim em relação ao contig. Quando existe função molecular e táxon associados a *feature*, essas informações também são exibidas. **b)** Outras anotações e detalhes que a *feature* pode ter recebido.

A biblioteca *Feature Viewer*[2] facilitou a implementação do modelo visual do contig, visto que permite exibir características de sequências de DNA e proteína gerando componentes Scalable Vector Graphics (SVG) de fácil integração com HTML e CSS. Utilizamos essa capacidade para criar elementos visuais que representam as *features*, anotações, reads, táxons e regiões.

¹⁰<https://samtools.github.io/hts-specs/SAMv1.pdf>

Alignment Read
Start: 897 **End:** 1146 **CIGAR:** 250M **FLAG:**
read paired
mate unmapped
read reverse strand
second in pair

Taxon Read
Scientific Name: Bordetella parapertussis **Rank:** species

Figura 4.11: Screenshot Caravela - **Detalhes da read** - **a)** Detalhes sobre o read em relação ao contig, tais como posição de início, fim e qualidade do alinhamento. **b)** Nome científico do táxon atribuído no nível mais profundo, ou seja, o mais próximo de espécie.

Casos de uso do projeto Metazoo

5.1 Processamento dos dados metagenômicos do projeto Metazoo

Testamos a ferramenta utilizando conjunto de dados gerados a partir do processo de compostagem do projeto Metazoo. Esse processo costuma durar por volta de 100 dias, durante esse período foram coletadas amostra seriadas de diferentes estágios, para duas diferentes composteiras, ZC3 dias (01, 30, 64, 78, 99) e ZC4 dias (01, 03, 07, 15, 30, 64, 67, 78, 99). Para o sequenciamento utilizamos a técnica de sequências pareadas (paired-end), com reads de 250 bp na plataforma Illumina MiSeq, os totais de reads resultante do processo de sequenciamento das amostras das duas composteiras podem ser visto nas Tabelas 5.1 e 5.2. O desenho experimental, coletas, extração de DNA e preparação de bibliotecas para o sequenciamento foram principalmente realizados por Luciana Antunes Principal (então aluna de doutorado da profa. Aline Maria da Silva). No fim de 2016, publicamos um artigo intitulado: "*Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and meta-transcriptomics*" na revista *Scientific Reports* descrevendo os principais resultados do projeto metazoo até então [10]. Neste trabalho, do qual o aluno é coautor, são descritos em detalhe todo os processos, materiais e métodos relacionados ao sequenciamento da compostagem do parque Zoológico de São Paulo.

A partir dos reads sequenciados, fizemos uso de algumas ferramentas para obter a monta-

gem do metagenoma, identificação taxonômica dos reads, análise de perfil funcional dos contigs gerados pela montagem e o mapeamento dos reads nos contigs. Por fim, depois destes processamentos, que serão descritos a seguir, obtivemos todos os arquivos de entrada necessários para submeter as amostras da compostagem ZC3 e ZC4 à Caravela. A figura 5.1 ilustra o fluxo de processamento do conjunto de dados metagenômicos.

Figura 5.1: Fluxo padrão de processamento dos dados metagenômicos. **a.** Todos os reads providos pelo sequenciamento em formato FASTQ. **b.** Processo de identificação taxonômica que recebe todos os reads como entrada e gera arquivo em formato tabular (5.1.1). **c.** Processo que filtra os reads de entrada por *score* de qualidade e tamanho mínimo gerando FASTQ (5.1.2). **d.** Processo de montagem que recebe os reads já filtrados como entrada e tenta formar sequências maiores através da sobreposição dos reads, o resultado é um arquivo com sequências formadas por junção de reads chamados de contig em formato FASTA (5.1.2). **e.** Processo que busca por possíveis sequências codificadoras de proteínas, as CDSs, quando uma CDS é encontrada ela é comparada com banco de dados de famílias de proteínas e pode ser anotada como sendo uma proteína com função já conhecida, no final é gerando um arquivo GFF com todas as CDSs encontradas (5.1.3). **f.** Processo de alinhamento para mapear os reads nos contigs, definindo quais reads formam um determinado contig, gerando um arquivo de mapeamento no formato BAM (5.1.4).

5.1.1 Identificação taxonômica dos reads

Utilizamos o MyTaxa[34], classificador de sequências genômicas e metagenômicas baseado em homologia para classificar taxonomicamente os reads do conjunto de dados ZC3 e ZC4. Primeiro, fizemos uma busca por similaridade usando USEARCH[20] (algoritmo ublast) filtrando por *e-value* máximo de $1e-5$, utilizamos como sequências *query* os reads e, como banco de dados o NCBI-NR [12]. O resultado do USEARCH foi então submetido ao MyTaxa. O arquivo final deste processamento foi formatado de maneira a gerar um arquivo tabular com duas colunas,

onde a primeira é a identificação do read e a segunda o *taxonomy id* de acordo com a base de dados de taxonomia do NCBI. Em média, cerca de 37% e 26% dos reads de ZC3 e ZC4 respectivamente, puderam ser classificados em algum nível taxonômico (tabelas 5.1 e 5.2).

Tabela 5.1: Totais de reads e reads classificados por Mytaxa para ZC3

ZC3					
Dia da coleta	01	30	64	78	99
Total de reads	2.3 M	2.1 M	1.3 M	2.2 M	1.5 M
Classificadas por MyTaxa	422 K	757 K	570 K	938 K	694 K
%	18.57	35.28	45.08	43.25	45.97

Tabela 5.2: Totais de reads e reads classificados por Mytaxa para ZC4

ZC4									
Dia da coleta	01	03	07	15	30	64	67	78	99
Total de reads	4.1 M	4.7 M	4.6 M	7.2 M	4.8 M	7.2 M	4.2 M	11.2 M	8.3 M
Classificadas por MyTaxa	1.3 M	1.6 M	1.4 M	1.6 M	1.4 M	2.2 M	928 K	2. M	1.6 M
%	32.68	34.58	31.07	22.81	28.90	30.38	22.19	18.15	18.82

5.1.2 Filtro de qualidade e montagem com SOAPdenovo2

Antes da montagem os reads foram submetidos a um controle de qualidade utilizando o software Sickle[25], foram removidos todos os reads com *Phred score* abaixo de 20 ou menores que 50 pares de base, então, os reads foram montados com SOAPdenovo2 [35], montagem *de novo* (sem referência). O número total de contigs de gerados para cada amostra a partir desse processo de montagem pode ser visto nas Tabelas 5.3 e 5.4 na linha *contigs*.

Esse tipo de filtro não é empregado na classificação taxonômica de reads. No entanto, é importante dizer que os adaptadores e filtros, utilizados na preparação da biblioteca e sequenciamento, foram removidos.

5.1.3 Anotação funcional dos contigs com IMG/M

A anotação funcional dos contigs foi feita utilizando o *pipeline* do IMG/M. Antes de submetermos os contigs fizemos um filtro removendo todas as sequências menores que 300 pares de base. Resumidamente, o *pipeline* do IMG/M tem três passos: 1) Filtro de qualidade das sequências. 2) Predição das features (CDSs, rRNA, tRNA) do metagenoma. 3) Busca por similaridade das features preditas, na sua maioria CDSs, com base de sequências de proteínas como COGs e *Pfam*. As CDS do metagenoma podem ser associadas com termos KEGG, Enzyme Commission number (EC) e filogenia [38]. O resultado deste processamento foi recuperado através de *download* de arquivos tabulares (*COG*, *Pfam*, *EC*, *KO*, *phylodist*), arquivos no formato GFF3 com *CDSs* e *tRNA*, por exemplo. Além de arquivos no formato FASTA com sequências dos contigs (nucleotídeos) e CDSs (aminoácidos).

Tabela 5.3: Número de contigs e features para ZC3 - Contigs com ao menos 300pb gerados a partir do processo de montagem com SoapDeNovo 2. Features, na sua enorme maioria CDSs, encontradas e anotadas pelo *pipeline* do IMG/M.

Total de contigs e features - ZC3					
Dias	01	30	64	78	99
Contigs	193 K	173 K	163 K	348 K	229 K
Features	243 K	211 K	201 K	423 K	285 K

Tabela 5.4: Número de contigs e features para ZC4 - Contigs com ao menos 300pb gerados a partir do processo de montagem com SoapDeNovo 2. Features, na sua enorme maioria CDSs, encontradas e anotadas pelo *pipeline* do IMG/M.

Total de contigs e features - ZC4									
Dias	01	03	07	15	30	64	67	78	99
Contigs	569 K	488 K	337 K	392 K	421 K	884 K	389 K	517 K	597 K
Features	649 K	566 K	417 K	454 K	489 K	1007 K	437 K	604 K	692 K

5.1.4 Mapeamento dos reads nos contigs

O mapeamento dos reads nos contigs é importante para determinar quais reads foram utilizados para montar cada contig. Para este fim, fizemos uso do *software* Bowtie 2 [32]. O

conjunto de contigs de cada amostra foi utilizado para criar uma base de dados, subsequentemente utilizamos os reads *paired-end* formadores dos contigs em questão como *query* para o mapeamento. É importante que o alinhador reporte múltiplos alinhamentos válidos para a mesma *read*. o Bowtie 2 não reporta essa informação por padrão, para habilita-lo é preciso utiliza o parâmetro *-a* na execução. Saber que um read possui mais de um alinhamento válido significa que o mesmo read pode ter sido utilizado para formar mais de um contig, essa informação é importante para a Caravela, especialmente para funcionalidade de classificação de reads por contexto. Além do o Bowtie 2, também fizemos uso do *software* SAMTools [7] para converter o arquivo de saída do formato SAM para BAM (versão binária do SAM), ordenar e criar os índices. A taxa de alinhamento geral para cada amostra das duas compostagem (ZC3 e ZC4) podem ser vistas na Tabela 5.5.

Tabela 5.5: Taxa de alinhamento geral dos reads nos contigs

Taxa de alinhamento de reads nos contigs									
Dias	01	03	07	15	30	64	67	78	99
ZC4	31.36%	30.05%	35.34%	29.58%	31.87%	33.00%	26.17%	29.33%	29.20%
ZC3	41.17%	-	-	-	29.93%	33.48%	-	33.07%	36.94%

5.2 Análise visual de contigs formados por reads associados a organismo de interesse

Selecionamos dois táxons para serem buscados e analisados através da ferramenta Caravela. Os gêneros selecionados foram *Paracoccus* e *Geobacillus*, ambos foram encontrados através da funcionalidade de busca de contigs por táxon, apresentada na seção 4.12. As buscas foram feitas na amostra ZC4 dia 01 e, para cada um dos dois gêneros, inúmeros contigs foram encontrados. O objetivo deste caso de uso é apresentar as possibilidades de visualização, navegação e análise que a ferramenta possibilita. Desta forma, separamos apenas um contig para cada táxon selecionado e os apresentamos de forma a maximizar as possibilidades da ferramenta.

5.2.1 Explorando contig associado ao gênero *Paracoccus*

A Caravela é uma ferramenta com interface interativa, capaz de possibilitar a análise visual de contigs em diferentes níveis taxonômicos e, para tentar ilustrar a experiência do usuário ao alternar entre diferentes níveis taxonômicos do contig *Ga0070450_1566305*, adicionamos a figura 5.2, que junto com as Tabelas 5.6 (táxons associados aos reads formadores do contig em diferentes níveis taxonômicos) e 5.7 (*features* e anotações funcionais associadas ao contig) tem por objetivo trazer para este documento parte das informações encontradas na funcionalidade de visualização de contig da ferramenta.

O contig *Ga0070450_1566305*, figura 5.2, foi encontrado na amostra do dia 01 da compostagem ZC4, através da ferramenta foi possível identificar que o contig é composto por 22 reads dos quais 20 foram associados à espécie *Paracoccus denitrificans* e 2 não foram associados a nenhum táxon. Apesar dessas 2 reads não classificadas, todas as bases do contig foram cobertas por um único táxon, portanto o índice de *CTV* é 1, enquanto a *CT* é igual a 0.9. Esses índices são os mesmos para os três níveis taxonômicos, família, gênero e espécie. Além disso, não existem regiões não classificadas ou de sobreposição entre táxons e, por consequência, também não existem fronteiras ao longo do contig. Todas essas características evidenciam que se trata de um contig taxonomicamente bem resolvido, inclusive ao nível de espécie.

Figura 5.2: Visualização em múltiplos níveis taxonômicos do contig *Ga0070450_1566305*

a. Anotações funcionais (*KO*, *PFAM*, *COG*, *EC*) e distribuição dos reads a nível taxonômico de família (*Rhodobacteraceae*) ao longo do contig. **b.** Reads a nível taxonômico de gênero (*Paracoccus*). **c.** Reads a nível taxonômico de espécie (*Paracoccus denitrificans*).

As informações contidas na Tabela 5.7 vieram do processo de anotação funcional, pipeline do IMG/M aplicado nos contigs, enquanto os dados da Tabela 5.6 vieram do processo de classificação taxonômica efetuada com o software *mytaxa*. Embora tais processos sejam independentes, ambos associaram seus componentes, *reads* e CDSs, ao mesmo gênero *Paracoccus*. A reunião de todas essas informações na mesma área de visualização nos permite sugerir que, embora se trate de um contig com poucas *reads* formadoras, o contig *Ga0070450_1566305* apresenta características de um fragmento genômico real do gênero *Paracoccus*.

5.2.2 Explorando contig associado ao gênero *Geobacillus*

O contig *Ga0070450_1568123*, figura 5.3, também foi encontrado na amostra do dia 01 da compostagem ZC4. Ao explorarmos visualmente o contig, podemos verificar que é composto por 43 *reads*, das quais 39 foram classificados em algum nível taxonômico. Ao inspeciona-lo nos níveis de família (figura 5.3 **a**) e gênero (figura 5.3 **b**), descobrimos que existe apenas um táxon para cada um desses níveis, *Bacillaceae* e *Geobacillus* (Tabela 5.8), além disso, os índices de

Tabela 5.6: Táxons associados ao contig *Ga0070450_1566305* em diferentes níveis taxonômicos
 * Cobertura exclusiva é o percentual de bases associados exclusivamente ao táxon.

Táxons no contig - Ga0070450_1566305 - ZC4 dia 01				
Nível taxonômico	Nome científico	reads	Cobertura	* Cobertura exclusiva
Espécie	Paracoccus denitrificans	20	1.00	1.00
Gênero	Paracoccus	20	1.00	1.00
Família	Rhodobacteraceae	20	1.00	1.00

Tabela 5.7: Features e anotações funcionais associados ao contig *Ga0070450_1566305*

Features no contig - Ga0070450_1566305 - ZC4 dia 01				
Tipo	início	fim	Anotações	
CDS	1	354	Predicted S-adenosylmethionine-dependent methyltransferase involved in bacterial cell division COG0357	
			Paracoccus sp. J56	
CDS	347	1123	ATPases involved in chromosome partitioning COG1192	
			Paracoccus sp. TRP	
CDS	1133	1555	Predicted transcriptional regulators COG1475	
			Paracoccus sp. J56	

CT e *CTV* são 0.90 e 0.99 respectivamente. Ainda, foi possível observar uma pequena fração de bases do contig que não foram classificadas e só ficam evidente quando utilizamos o recurso de *zoom* da ferramenta, como pode ser visto na figura 5.4 a.

Figura 5.3: Visualização em múltiplos níveis taxonômicos do contig *Ga0070450_1568123* **a.** Anotações funcionais (*KO*, *PFAM*), distribuição dos reads a nível taxonômico de família (*Bacillaceae*) ao longo do contig e região não classificada (*unclassified*). **b.** reads a nível de gênero (*Geobacillus*) e região não classificada. **c.** Reads associados às espécies *Geobacillus thermodenitrificans* (azul claro), *Geobacillus sp. GHH01* (azul escuro), *Geobacillus kaustophilus* (roxo) e *Geobacillus sp. Y412MC61* (verde claro), além das regiões de sobreposições, indefinidas, não classificadas e fronteiras.

Figura 5.4: Zoom na região 1800-2020 do *Ga0070450_1568123* ao nível de gênero.
a. Representação visual da região não classificada do contig, ou seja, bases do contig foram exclusivamente formadas por reads sem nenhum táxon associado.

Ao analisamos o contig a nível espécie (figura 5.3 **c**) percebemos que os índices *CT* e *CTV* são diferentes, 0.23 e 0.29 respectivamente. Além disso, novos elementos visuais como regiões indefinidas, regiões de sobreposição e fronteiras são apresentados. Isto acontece pois quatro

diferentes espécies de *Geobacillus* foram associados aos reads neste nível taxonômico, *Geobacillus thermodenitrificans*, *Geobacillus sp. Y412MC61*, *Geobacillus sp. GHH01* e *Geobacillus kaustophilus*, todos listados na Tabela 5.8. A existência de diferentes táxons no mesmo nível taxonômico possibilita a sobreposição entre os mesmos e, de fato isso acontece, sendo possível observar tal sobreposição em dois elementos da visualização. Na região de sobreposição (*overlap taxa*) e na região indefinida (*undefined*). A região de sobreposição, inclusive, é capaz de informar visualmente que existe sobreposição de dois diferentes táxons entre a posição 800-1200 e de três táxons entre a posição 1200-1600 do contig. Outra característica é o aumento do número de reads sem táxon associado (na cor cinza), algo que acontece pois um read associado a um gênero não está obrigado a ter também uma espécie atribuída e, por consequência, resulta na formação de regiões não classificadas (*unclassified*) a nível de espécie, que não existiam a nível de gênero. As regiões de indefinidas e não classificadas são premissas para formação de fronteiras, algo que acaba se concretizando através das duas regiões indefinidas do contig, pois por estarem ladeadas por diferentes táxons, foram também identificadas como fronteiras (*boundaries*).

Tabela 5.8: Táxons associados ao contig *Ga0070450_1568123* em diferentes níveis taxonômicos
 * Cobertura exclusiva é o percentual de bases associados exclusivamente ao táxon.

Táxons no contig - Ga0070450_1568123 - ZC4 dia 01				
Nível taxonômico	Nome científico	reads	Cobertura	* Cobertura exclusiva
Espécie	<i>Geobacillus thermodenitrificans</i>	10	0.42	0.29
	<i>Geobacillus sp. Y412MC61</i>	2	0.22	0.12
	<i>Geobacillus sp. GHH01</i>	6	0.28	0.04
	<i>Geobacillus kaustophilus</i>	2	0.10	0.00
Gênero	<i>Geobacillus</i>	39	1.00	1.00
Família	Bacillaceae	39	1.00	1.00

Embora exista uma discordância entre os táxons associados aos reads a nível de espécie, é fato que a nível de gênero o contig é taxonomicamente bem resolvido para *Geobacillus*. Ainda na tentativa de definir o contig a nível de espécie podemos utilizar as informações da Tabela 5.9 que mostra que três das quatro CDSs encontradas no contig foram associadas a espécie *Geobacillus thermodenitrificans* DSM 465, reforçando as evidências da Tabela 5.8, que informa a espécie

Geobacillus thermodenitrificans como correspondente a 42% das bases do contig. Por fim, podemos dizer que existem fortes evidências que se trata de um fragmento genômico real do gênero *Geobacillus* e possivelmente da espécie *Geobacillus thermodenitrificans*.

Tabela 5.9: Features e anotações funcionais associados ao contig *Ga0070450_1568123*

Features no contig - Ga0070450_1568123 - ZC4 dia 01			
Tipo	início	fim	Anotações
CDS	2	151	-
			-
CDS	483	803	-
			<i>Geobacillus thermodenitrificans</i> DSM 465
CDS	862	1797	Predicted metal-dependent enzyme COG3872
			<i>Geobacillus thermodenitrificans</i> DSM 465
CDS	1984	2256	Conserved membrane protein YqhR pfam11085
			<i>Geobacillus thermodenitrificans</i> DSM 465

5.3 Analisando qualidade de contigs montados com *SoapDeNovo 2* partir das amostra de ZC4 e ZC3.

Ao final do processo de montagem é comum que se queira analisar os contigs resultantes e, para esse fim, pode-se fazer uso de ferramentas como Quast [28], por exemplo. Essa ferramenta é capaz de gerar métricas tais como número total de contigs montados, maior contig, total de bases montadas e N50. Tais métricas auxiliam na interpretação do resultado da montagem, especialmente de genomas, porém, essas informações são apenas quantitativas e, quando se trata de metagenomas, informações sobre a qualidade dos contigs montados podem ser mais relevantes que as informações quantitativas. Por exemplo, é bom que um montador de metagenomas gere contigs longos, contudo, é mais importante que tais contigs não sejam apenas artefatos da montagem.

Utilizamos os conceitos de contigs **taxonomicamente bem resolvidos** (TBR, definição 14) e **potencialmente quiméricos** (PQ, definição 15), apresentados nesse trabalho e

implementados na ferramenta Caravela, para produzirmos análises de contigs metagenômicos da perspectiva qualitativa, como mostraremos a seguir.

Para cada amostra (dia) das compostagens ZC3 e ZC4 apresentadas nas Tabelas 5.3 e 5.4 respectivamente, na página 40, separamos apenas os contigs com índice $ITG \geq 0.5$ e, a esse grupo de contigs chamamos de **candidatos**. Esse primeiro filtro é importante pois contigs que não têm ao menos metade dos seus reads associados a algum táxon não atendem aos requisitos mínimos para serem caracterizados como TBR ou PQ. Os contigs **candidatos** foram então classificados em TBR ou PQ de acordo com as definições 14 e 15, enquanto os contigs que não se encaixaram em nenhuma das duas definições foram chamados de **outros**.

Ao analisarmos os resultados das amostras da compostagem ZC4, figura 5.5, verificamos que mesmo a nível de espécie, onde é esperado uma maior número de contigs com menor consistência taxonômica, nenhuma das amostras apresentou índice de contigs PQ maior que 3%. Por outro lado, a maior parte dos contigs de todos as dias, por volta de 60% em média, foram classificas como TBR.

Figura 5.5: Análise de qualidade dos contigs de ZC4 utilizando conceitos TBR e PQ - Os dias (amostras) no eixo X e percentual de contigs no eixo Y - **a)** Nível de gênero. **b)** Nível de família **c)** Nível de espécie.

Outros são contigs candidatos que não se encaixaram em nenhuma das duas definições, TBR ou PQ.

Ainda nas amostras de ZC4, como já era esperado, o percentual de contigs TBR aumenta enquanto o de PQ diminui conforme subimos na hierarquia taxonômica. Por exemplo, na

amostra do dia 01 ao nível de espécie, o percentual de contigs TBR não chega a 70%, enquanto o percentual de contigs PQ é de 1.21%. Na mesma amostra ao nível de família, o percentual de contigs TBR se aproxima dos 80%, enquanto o percentual de contigs PQ é de apenas 0.062%. Do mesmo modo, esse padrão se repete em todas as amostras.

Nas amostras da compostagem ZC3, figura 5.6, o perfil dos resultados é bem próximo ao de ZC4. Para os níveis taxonômicos de espécie, gênero e família, o percentual de contigs PQ também é baixo, no entanto, ligeiramente superior ao da composteira ZC4. Assim como nas amostras de ZC4, o percentual de contigs TBR aumenta enquanto o de PQ diminui ao subirmos na hierarquia taxonômica. Além disso, mesmo ao nível de espécie, todas as amostras apresentam mais de 60% de contigs TBR.

Esses resultados sugerem que os contigs montados com *SoapDeNovo 2*, a partir das amostras ZC3 e ZC4 foram bem montados, pois apresentam baixo percentual de contigs potencialmente quiméricos e um grande percentual de contigs bem resolvidos taxonomicamente.

Figura 5.6: Análise de qualidade dos contigs de ZC3 utilizando conceitos TBR e PQ - Os dias (amostras) no eixo X e percentual de contigs no eixo Y - a) Nível de gênero. b) Nível de família. c) Nível de espécie.

Outros são contigs candidatos que não se encaixaram em nenhuma das duas definições, TBR ou PQ.

As tabelas com valores absolutos e percentuais a partir dos quais foram gerados os gráficos da figura 5.5 e 5.6 podem ser encontradas no apêndice A, Tabelas A.1, A.2, A.3, A.4, A.5 e A.6.

5.4 Comparando resultado de diferentes montadores

É comum em estudos de comunidades de microorganismos testar diferentes programas de montagem antes de definir qual montador será efetivamente utilizado. Isso porque, ainda que recebam exatamente o mesmo conjunto de sequências como entrada, por implementarem técnicas de montagem distintas, os montadores apresentam resultados diferentes. Nesta seção, comparamos os resultados de montagens efetuadas por dois diferentes montadores, *SoapDeNovo 2* e *Mira* [16]. Para esse fim, exploramos a capacidade da ferramenta Caravela em analisar qualitativamente, da perspectiva de consistência taxonômica, os contigs metagenômicos e, desta forma, tentamos esclarecer qual montador apresentou melhores resultados. Os dados utilizados como entrada foram as sequências das amostras da compostagem ZC4.

As sequências das amostras da compostagem ZC4 foram montadas com o *software SoapDeNovo 2*, como já foi descrito nesse documento. No entanto, essas mesmas amostra também foram montadas *software Mira* [16]. As montagens com o *Mira* foram feitas por Deyvid Emanuel Amgarten, dentro do seu projeto de mestrado intitulado *Análise computacional da diversidade viral presente na comunidade microbiana do processo de compostagem do Zoológico de São Paulo*. O montador *Mira* foi configurado no modo “genome, accurate” e demais parâmetros *default*, além disso, apenas contigs ≥ 500 pb foram mantidos no resultado da montagem.

A descrição do processo de montagem com o *software SoapDeNovo 2* já foi apresentada na seção 5.1.2, porém, para podermos comparar os resultados das duas montagens filtramos o resultados da montagem obtida com *SoapDeNovo 2* para que apenas contigs ≥ 500 pb fossem considerados, os resultados das duas montagens, com *Soap 2* e *Mira*, ambas contendo apenas contigs ≥ 500 pb estão na Tabela 5.10.

Tabela 5.10: Total de Contigs ≥ 500 pb montados com SoapDeNovo 2 e Mira - 1K=1000

Contigs ≥ 500 pb montados com SoapDeNovo 2 e Mira									
Dias	01	03	07	15	30	64	67	78	99
Soap 2	89 K	100 K	68 K	85 K	89 K	155 K	55 K	117 K	129 K
Mira	52 K	58 K	56 K	77 K	58 K	101 K	43 K	117 K	111 K

5.4.1 Comparando percentual de contigs TBR e PQ formados pelos montadores *SoapDeNovo 2* e *Mira*

Fizemos uso da Caravela e novamente exploramos os conceitos de TBR e PQ, dessa vez para comparar resultados de diferentes montadores, *SoapDeNovo 2* e *Mira*. A partir do conjunto de contigs apresentados na Tabela 5.10, buscamos por contigs **candidatos**, aqueles que apresentam índice $ITG \geq 0.5$, então, classificamos os contigs candidatos em TBR, PQ e, no caso do contig não se encaixe em nenhuma dessas duas definições, foi classificado como **outros**. Executamos esse processo para três diferentes níveis taxonômicos, espécie, gênero e família. Os resultados podem ser observados nos gráficos das figuras 5.7 e 5.8.

Ao analisarmos os resultados, é notável que em todos níveis taxonômicos apresentados, a montagem feita pelo *SoapDeNovo 2* apresenta maior percentual de contigs TBR e menor percentual de contigs PQ, quando comparada aos resultados do montador *Mira*. A título de exemplo, ao nível de gênero, a média de contigs TBR, **outros** e PQ montados com *Soap 2* é de 35.03%, 58.87% e 6.10% respectivamente, enquanto a do *Mira* é de 25.04%, 58.80% e 16.96%. Em resumo, é possível dizer que ao nível de gênero, o *SoapDeNovo 2* gerou em torno de 10% a mais de contigs **taxonomicamente bem resolvidos** e por volta de 10% a menos de contigs **potencialmente quiméricos** quando comparado com *Mira*.

Figura 5.7: Análise de qualidade dos contigs de ZC4 com ao menos 500 bp montados com *SoapDeNovo 2* - Os dias (amostras) no eixo X e percentual de contigs no eixo Y - a) Nível de gênero. b) Nível de família. c) Nível de espécie.

Outros são contigs candidatos que não se encaixaram em nenhuma das duas definições, TBR ou PQ.

Tais resultados, figuras 5.7 e 5.8, mostram que o montador *SoapDeNovo 2* gerou contigs melhores, não apenas gerando um maior percentual de contigs taxonomicamente bem resolvidos, mas também gerando menor percentual de contigs potencialmente quiméricos, quando comparados aos resultados do montador *Mira*. Isso ocorre para todas as amostras da compostagem ZC4 nos três níveis taxonômicos apresentados.

Figura 5.8: Análise de qualidade dos contigs de ZC4 com ao menos 500 bp montados com *Mira* - Os dias (amostras) no eixo X e percentual de contigs no eixo Y - a) Nível de gênero. b) Nível de família. c) Nível de espécie.

Outros são contigs candidatos que não se encaixaram em nenhuma das duas definições, TBR ou PQ.

As tabelas com valores absolutos e percentuais a partir dos quais foram gerados os gráficos da figura 5.7 e 5.8 podem ser encontradas no apêndice A.2, Tabelas A.7, A.8, A.9, A.10, A.11 e A.12.

5.4.2 Comparando total de bases TBR e PQ geradas pelos montadores *SoapDeNovo 2* e *Mira*

Uma outra métrica comum em ferramentas de avaliação de montagens é o **total de bases montadas** da amostra, de maneira geral, valores maiores indicam montagem melhores. O total de bases montadas com *SoapDeNovo 2* e *Mira*, dos contigs da Tabela 5.10, página 50, estão representados no gráfico da figura 5.9. Nesse gráfico, é possível observar que exceto pelas amostras dos dias 01 e 30, onde existe uma ligeira vantagem para o montador *SoapDeNovo 2*, em todas as outras amostras, o montador *Mira* gera números maiores de bases montadas. Esse resultado revela que, embora o número de contigs do montador *SoapDeNovo 2* seja maior para todas as amostras da compostagem ZC4 (Tabela 5.10), isso não se reflete no total de bases montadas, sinalizando que os contigs formados pelo *Mira* são maiores. De forma geral, produzir

contigs maiores é bom, contudo, isso aumenta a possibilidade de gerar contigs quiméricos.

Figura 5.9: Total de bases montadas com *SoapDeNovo 2* e *Mira* a partir de contigs $\geq 500pb$ das amostras da compostagem ZC4.

O total de bases montadas de uma amostra também é uma métrica quantitativa, não apresentando informação sobre a qualidade das bases montadas. Com o objetivo de verificar qual montador gera o maior número de bases montadas onde a qualidade também possa ser verificada, somamos o total de base montadas apenas para os contigs TBR para ambas as montagens. Dessa forma, os resultados apresentados no gráfico da figura 5.10, podem ser chamados de *total de bases taxonomicamente bem resolvidas montadas*.

Quando observamos o total de bases montadas apenas a partir dos contigs TBR, figura 5.10, a montagem feita com *SoapDeNovo 2* leva vantagem em todos as amostras da compostagem ZC4, demonstrando que, apesar do montador *Mira* ter apresentado vantagem no total de bases montados geral (figura 5.9), quando observamos apenas bases montadas a partir de contigs taxonomicamente bem resolvidos, o *SoapDeNovo 2* apresenta melhores resultados.

Figura 5.10: Total de bases montadas com *SoapDeNovo 2* e *Mira* a partir de contigs $TBR \geq 500pb$ das amostras da compostagem ZC4.

Por outro lado, avaliamos também qual montador gerou maior número de bases considerando apenas contigs PQ. Nesse caso, o *total de bases potencialmente quiméricas montadas*. Os resultados estão apresentados no gráfico da figura 5.11, demonstrando que o montador *Mira*, em todas as amostra da compostagem ZC4, gerou ao menos o dobro de bases potencialmente quiméricas montadas, quando comparado com o *SoapDeNovo 2*.

Figura 5.11: Total de bases montadas com *SoapDeNovo 2* e *Mira* a partir de contigs $PQ \geq 500pb$ das amostras da compostagem ZC4.

Os resultados apresentados nesta seção sugerem que embora o montador *Mira* gere contigs mais longos, ao menos do ponto de vista de consistência taxonômica, o conjunto de contigs gerado pelo montador *SoapDeNovo 2* é melhor que os do *Mira*.

Os dados que foram utilizados para gerar os gráficos das figuras 5.9, 5.10 e 5.11 apresentados nesta seção, estão nas Tabelas A.13, A.14 e A.15 do apêndice A.3.

5.5 Classificação taxonômica de *reads* por contexto

A classificação taxonômica de sequências curtas (*reads*) é um dos grandes desafios nos estudos de comunidades microbianas. Na literatura, existem muitas ferramentas que se dedicam a classificação taxonômica de sequências metagenômicas, tais como *MyTaxa* [34], *Kraken*[51], *Taxonomer* [23] e *Clark-s* [42], por exemplo. Essas ferramentas podem utilizar diferentes métodos e/ou base de dados para tentar associar uma sequência a um táxon, no entanto, é comum que deixem parte das sequências de uma amostra metagenômica sem classificação.

Como já foi visto ao longo desse documento, a ferramenta Caravela faz uso das informações

taxonômicas dos *reads* formadores de um contig para classifica-los em contigs TBR ou PQ. Por outro lado, a Caravela também pode fazer uso dos contigs para identificar *reads* não classificados que esteja totalmente alinhados em uma região classificada do contig, permitindo a classificação taxonômica de *read* por contexto, como foi apresentado na seção 4.9.

5.5.1 Aplicação aos *reads* da compostagem ZC4 e ZC3

Utilizamos o recurso de classificação taxonômica por contexto da Caravela com o propósito aumentar o número de *reads* já classificadas com a ferramenta *MyTaxa*. Fizemos uso dos contigs formados pelo montador *SoapDeNovo 2*, Tabelas 5.3 e 5.4, e a classificação taxonômica das *reads* dada pela ferramenta *MyTaxa*, Tabelas 5.1 e 5.2 das compostagens ZC3 e ZC4.

Como resultado, que poder ser visto figura 5.12 (em verde) e na Tabela 5.11, foi possível aumentar o número de *reads* classificados pela ferramenta *MyTaxa* para todas as amostras da compostagem ZC4. Em especial, dia 78, foi possível classificar por contexto mais de 450 mil *reads*, o que representou um aumento de mais de 48% de *reads* classificados inicialmente pelo software *MyTaxa*.

Figura 5.12: Gráfico (cumulativo) de *reads* classificados por contexto pela ferramenta Caravela para amostras da compostagem ZC4. - Em cinza, o total de *reads* não classificados, em azul, o total de *reads* classificados com *MyTaxa*. Em verde, o total de *reads* classificados pela Caravela utilizando a classificação por contexto.

Tabela 5.11: ZC4 - Total de reads não classificados, classificados com *MyTaxa* e classificados por contexto na ferramenta Caravela

A última coluna, classificadas por contexto Caravela (%), refere-se ao percentual de reads que foi possível classificar a mais utilizando a classificação por contexto da Caravela, em relação ao número de reads classificados com *MyTaxa*

Total de reads - ZC4				
Dia	Não classificados	Classificadas	Classificadas por contexto	Classificadas por contexto
		MyTaxa	Caravela	Caravela (%)
01	618,655	614,768	74,846	12.17
03	663,260	687,142	120,911	17.60
07	829,019	661,759	144,872	21.89
15	1,215,486	754,792	232,519	30.81
30	825,519	635,556	124,014	19.51
64	1,274,076	1,021,510	121,817	11.93
67	611,891	441,628	58,825	13.32
78	2,023,580	929,242	450,993	48.53
99	1,523,799	707,167	240,150	33.96

Para as amostras da compostagem ZC3, figura 5.13 e Tabela 5.12, os resultados foram similares. Também foi possível aumentar o número de reads classificados para todos os dias. A classificação por contexto aumentou em torno de 17%, em média, o total de reads classificados inicialmente com *MyTaxa*. Ainda na compostagem ZC3, o destaque foi o dia 01, que sozinho teve um aumento de 38.15% em relação ao total de reads classificados inicialmente.

Figura 5.13: Gráfico (cumulativo) de reads classificados por contexto pela ferramenta Caravela para amostras da compostagem ZC3. - Em cinza, o total de reads não classificados, em azul, o total de reads classificados com *MyTaxa*. Em verde, o total de reads classificados pela Caravela utilizando a classificação por contexto.

Tabela 5.12: ZC3 - Total de reads não classificados, classificados com *MyTaxa* e classificados por contexto na ferramenta Caravela

A última coluna, classificadas por contexto Caravela (%), refere-se ao percentual de reads que foi possível classificar a mais utilizando a classificação por contexto da Caravela, em relação ao número de reads classificados com *MyTaxa*.

Total de reads - ZC3				
Dia	Não classificadas	Classificadas	Classificadas por contexto	Classificadas por contexto
		MyTaxa	Caravela	Caravela (%)
01	714,190	192,775	73,549	38.15
30	336,626	288,320	60,190	20.88
64	231,629	234,083	25,570	10.92
78	410,300	368,851	34,322	9.31
99	291,006	330,805	32,046	9.69

A classificação por contexto efetuada pela Caravela é sempre feita a nível de gênero e os resultados são apresentados para o usuário em forma de relatório, como descrito na seção 4.9. É importante dizer que esse relatório é uma sugestão de classificação taxonômica, cabendo ao usuário aceitar ou não tais sugestões.

5.6 Análise do contig *lake-12-11_rep-c113861* do Lago São Francisco

Um dos ambientes estudados no projeto Metazoo é o Lago São Francisco (na verdade uma represa) do Parque Zoológico de São Paulo. Foram coletadas amostras de água durante doze meses entre os anos de 2012 e 2013. A aluna de Doutorado, Lívia Moura e o bolsista TT5, Antônio Diaz Tula, têm trabalhando na recuperação de genomas a partir de amostras metagenômicas do lago. Recentemente, o também aluno de Doutorado, Deyvid Amgarten, se juntou aos dois para ajudar nessa tarefa.

Quando o Deyvid se juntou ao grupo, com o objetivo de recuperar genomas a partir de amostras do lago, ele passou a fazer uso de alguns montadores metagenômicos já conhecidos como o Mira [16]. Em uma das amostras do lago, *lake-12-11*, de forma surpreendente, esse processo conseguiu recuperar contigs longos (251 kb, 203 kb, 194 kb, 147 kb, 131 kb, 130 kb), algo incomum para amostras de comunidade microbiana. Um contig em especial, o *lake-12-11_rep-c113861*, chamou a atenção de todos no grupo pelo seu tamanho de 503 kb.

Em um primeiro momento, a principal suspeita foi que se tratava de um contig artefactual, visto que já tínhamos indícios que o montador Mira gera um número considerável de contigs potencialmente quiméricos (ver Seção 5.4). Entretanto, também tínhamos evidências que embora parte dos contigs gerados pelo Mira fossem potencialmente quiméricos, ele também é capaz de gerar contigs taxonomicamente consistentes. Neste ponto, fui convidado pelo grupo a adicionar os dados do contig *lake-12-11_rep-c113861* na Caravela para analisarmos. Em outra frente, o Deyvid iria fazer análises complementares sobre o contig.

Inicialmente fizemos o mapeamento das *reads* da amostra *lake-12-11* no contig *lake-12-11_rep-c113861* com Bowtie 2 [32], neste processo verificamos que em torno de 95 mil *reads* mapearam no contig. Além disso, utilizamos a classificação taxonômica das *reads* dada pelo Clark-S [42]. A escolha foi feita com base em resultados preliminares efetuados com metagenomas sintéticos onde o Clark-S apresentou resultados melhores em relação a outras ferramentas. Juntamos todos os dados, carregamos e processamos na Caravela.

O resultado: a Caravela mostrou que 91% ($ITG=0.91$) das *reads* do contig foram classi-

ficadas em algum nível taxonômico. Quando analisado ao nível de gênero, o contig foi dado como potencialmente quimérico pela caravela. Porém, por ser um contig tão longo, e por existir a suspeita de se tratar de um fragmento de genoma de um organismo ainda desconhecido, analisamos o contig em um nível menos específico.

Ao nível de filo, figura 5.14, o contig apresentou 41% ($CT=0.41$) das *reads* associada ao mesmo filo, contudo, menos de 2% ($CTV=0.016$) das bases do contig foram formadas por *reads* associadas a um único filo, as reads formadoras do contig foram divididas, em sua maior parte, entre *Actinobacteria* e *Proteobacteria*. Por consequência, em torno de 98% ($Region\ Size\ Undefined=0.981$) do contig foi considerado indefinido ao nível de filo. Essas características apontaram para a confirmação da suspeita inicial, de tratar-se de um contig artefactual.

Figura 5.14: Visualização em modo reads consenso do contig *lake-12-11_rep_c113861* - Nesse modo de visualização não são apresentadas as reads do contig e sim um consenso entre as reads do mesmo táxon. **a)** Botões de controle do contig. **b)** Visualização consenso entre reads do mesmo táxon, em sua maior parte *Actinobacteria* (marrom), *Proteobacteria* (roxo). **c)** Regiões de sobreposição entre táxons, indefinidas, não classificadas e fronteiras.

Por outro lado, as análises complementares efetuadas em paralelo pelo doutorando Deyvid, apontavam justamente para o contrário, dando a entender que poderia sim se tratar de um

fragmento genômico real. Neste ponto, diante de resultados contrastantes, levantou-se a hipótese de que o classificador taxonômico (Clark-S) estava classificando as reads de forma errada. Essa hipótese ganhou força ao constataremos que o Clark-S classificou por volta de 90% de todos as *reads* da amostra *lake-12-11*, um percentual muito alto em relação ao que costuma ser classificados para amostras ambientais.

A hipótese de erro do classificador taxonômico nos levou revisão do processo de identificação taxonômica efetuada pelo Clark-S. Ao longo dessa revisão, o Antônio percebeu que a forma de interpretar o arquivo de saída do Clark-S não estava muito clara, sendo assim, ele escreveu para os autores da ferramenta com o objetivo de esclarecer como o tal arquivo deveria ser interpretado. Ao receber a resposta, ficou claro que estávamos interpretando o arquivo de saída do Clark-S de forma equivocada, o que estava inflando o total de *reads* classificadas da amostras e consequentemente do contig *lake-12-11_rep-c113861* com associações taxonômicas espúrias. O percentual de *reads* classificadas da amostras passou de 90% para menos de 20% depois da correção.

Embora até o momento ainda não exista uma definição sobre a natureza do contig *lake-12-11_rep-c113861*, a Caravela se mostrou útil tanto para sua análise quanto para sua exploração. Especialmente em relação a distribuição dos táxons em diferentes níveis taxonômicos ao longo do contig. Essa capacidade, junto com análises complementares, ajudaram a descobrir um erro na interpretação do arquivo de saída do classificador taxonômico Clark-S. A descoberta desse erro foi particularmente importante pois teve implicações em outros processos do projeto Metazoo que faziam, ou ainda fazem uso dos resultados do programa Clark-S.

Conclusão

A Caravela apresenta uma interface interativa e amigável, permitindo ao usuário a exploração dos dados a partir de uma nova perspectiva. Por exemplo, a visualização de contigs evidencia, através de diferentes cores, a distribuição taxonômica associada as *reads* ao longo da extensão do contig. Além disso, é possível identificar visualmente as **regiões taxonômicas classificadas, indefinidas, não classificadas e fronteiras** existentes no contig.

Ainda do ponto de vista exploratório, é possível efetuar buscas por contigs através dos táxons associados as *reads* ou funções atribuídas as *features*. Esse recurso de busca é poderoso, pois permite identificar de maneira rápida quais funções estão associadas a um dado organismos, ou o contrário, quais organismos estão associados a uma dada função.

Da perspectiva analítica, a ferramenta Caravela foi capaz de reportar listas de contigs TBR e PQ, a partir das quais foi possível efetuar uma análise qualitativa dos contigs resultantes do processo de montagem de diferentes montadores. Essa capacidade é uma novidade, pois, embora existam ferramentas para avaliação de montagens metagenômicas na literatura, em geral, tais ferramentas são apenas quantitativas ou utilizam métodos diferentes dos empregados neste trabalho.

A atribuição taxonômica de sequências originárias de comunidade microbianas certamente é um dos grandes desafios da metagenômica, por consequência, ferramentas para esse fim estão constantemente sendo publicadas na literatura. A capacidade da Caravela em classificar *reads* com base no contexto, pode ser entendida como um método de melhoria da atribuição taxonômica, muito embora, também poderia ser considerada um novo método de classificação

taxonômica. Em ambos os casos, certamente é mais um contribuição deste trabalho.

6.1 Contribuições deste trabalho

- **A ferramenta Caravela:** uma plataforma capaz de integrar dados metagenômicos de análises de perfil taxonômico e funcional baseados em *reads* e contigs respectivamente.
- **Índices e Definições:** Os índices de Identificação Taxonômica Geral (***ITG***), Consistência Taxonômica (***CT***), Consistência Taxonômica Vertical (***CTV***), Cobertura (***CO***) e Cobertura Exclusiva (***CE***). As definições de **Região Indefinida**, **Região Não Classificada** e **Fronteira**.
- **Resultados de aplicação:** O resultado de aplicação da seção 5.6 descreve um resultado da efetiva utilização da ferramenta Caravela dentro do projeto Metazoo.

Tabelas

A.1 – Tabelas utilizadas para produzir gráficos apresentados na seção 5.3.

Para as Tabelas A.1, A.2 e A.3 foram utilizados os dados da Tabela 5.4, página 40, como conjunto de contigs **inicial**.

Candidatos: Conjunto de contigs **inicial** com índice de $ITG \geq 0.5$.

Outros: Conjunto de contigs **candidatos** que não se encaixam nas definição de TBR ou PQ.

Tabela A.1: Análise qualitativa de contigs $\geq 300pb$ compostagem ZC4 montados com *SoapDeNovo 2* - Espécie.

ZC4 - <i>SoapDeNovo 2</i> ($\geq 300pb$) - Espécie							
Dia	valores percentuais			Valores absolutos			Candidatos
	PQ	Outros	TBR	PQ	Outros	TBR	
01	1.21%	29.00%	69.78%	2778	66417	159813	229008
03	2.45%	34.88%	62.66%	4792	68177	122465	195434
07	3.00%	36.52%	60.48%	3888	47283	78298	129469
15	3.00%	38.42%	58.58%	3750	48081	73305	125136
30	2.67%	39.56%	57.77%	3884	57634	84156	145674
64	2.10%	35.33%	62.57%	6984	117562	208210	332756
67	1.95%	33.47%	64.57%	2782	47661	91941	142384
78	2.62%	39.11%	58.27%	3467	51734	77070	132271
99	2.79%	39.24%	57.98%	4742	66808	98710	170260

Tabela A.2: Análise qualitativa de contigs $\geq 300pb$ compostagem ZC4 montados com *SoapDeNovo 2* - Gênero.

ZC4 - SoapDeNovo 2 ($\geq 300pb$) - Gênero							
Dia	valores percentuais			Valores absolutos			Candidatos
	PQ	Outros	TBR	PQ	Outros	TBR	
01	0.62%	24.50%	74.88%	1543	61181	186992	249716
03	1.39%	31.84%	66.76%	3031	69383	145465	217879
07	1.88%	33.74%	64.39%	2701	48481	92528	143710
15	2.11%	35.79%	62.10%	2909	49452	85801	138162
30	1.92%	36.66%	61.42%	3115	59631	99904	162650
64	1.34%	32.38%	66.28%	4944	119879	245391	370214
67	1.39%	31.45%	67.16%	2230	50396	107628	160254
78	1.80%	36.24%	61.96%	2651	53348	91197	147196
99	2.01%	37.14%	60.85%	3790	70072	114802	188664

Tabela A.3: Análise qualitativa de contigs $\geq 300pb$ da compostagem ZC4 montados com *SoapDeNovo 2* - Família.

ZC4 - SoapDeNovo 2 ($\geq 300pb$) - Família							
valores percentuais				Valores absolutos			
Dia	PQ	Outros	TBR	PQ	Outros	TBR	Candidatos
01	0.57%	21.14%	78.29%	1407	51985	192489	245881
03	1.35%	28.93%	69.72%	2872	61643	148531	213046
07	1.79%	31.45%	66.76%	2493	43783	92958	139234
15	1.88%	33.39%	64.74%	2514	44710	86692	133916
30	1.76%	34.33%	63.91%	2768	53987	100491	157246
64	1.11%	29.68%	69.21%	4003	107402	250439	361844
67	1.18%	29.68%	69.14%	1835	46175	107566	155576
78	1.60%	33.38%	65.02%	2291	47816	93155	143262
99	1.65%	34.76%	63.59%	3014	63657	116459	183130

Para as Tabelas A.4, A.5 e A.6 foram utilizados os dados da Tabela 5.3, página 40, como conjunto de contigs **inicial**.

Tabela A.4: Análise qualitativa de contigs $\geq 300pb$ da compostagem ZC3 montados com *SoapDeNovo 2* - Espécie.

ZC3 - SoapDeNovo 2 - Espécie							
valores percentuais				Valores absolutos			
Dia	PQ	Outros	TBR	PQ	Outros	TBR	Candidatos
01	2.02%	28.74%	69.24%	842	11976	28855	41673
30	2.44%	27.65%	69.90%	1592	18007	45524	65123
64	4.09%	32.84%	63.07%	2620	21043	40408	64071
78	2.70%	35.10%	62.20%	3614	47064	83393	134071
99	3.79%	37.59%	58.61%	3402	33705	52551	89658

Tabela A.5: Análise qualitativa de contigs $\geq 300pb$ da compostagem ZC3 montados com *SoapDeNovo 2* - Gênero.

ZC3 - SoapDeNovo 2 - Gênero							
valores percentuais				Valores absolutos			
Dia	PQ	Outros	TBR	PQ	Outros	TBR	Candidatos
01	1.71%	25.98%	72.30%	754	11425	31795	43974
30	2.18%	26.35%	71.48%	1580	19134	51910	72624
64	3.60%	31.65%	64.75%	2598	22817	46685	72100
78	2.08%	32.53%	65.39%	3144	49151	98806	151101
99	3.04%	35.55%	61.41%	3050	35660	61595	100305

Tabela A.6: Análise qualitativa de contigs $\geq 300pb$ da compostagem ZC3 montados com *SoapDeNovo 2* - Família.

ZC3 - SoapDeNovo 2 - Família							
valores percentuais				Valores absolutos			
Dia	PQ	Outros	TBR	PQ	Outros	TBR	Candidatos
01	1.44%	25.04%	73.51%	632	10962	32176	43770
30	1.84%	23.68%	74.49%	1303	16806	52875	70984
64	3.23%	29.62%	67.14%	2281	20886	47345	70512
78	1.68%	30.23%	68.09%	2481	44551	100350	147382
99	2.44%	32.74%	64.82%	2401	32232	63808	98441

A.2 – Tabelas utilizadas para produzir gráficos apresentados na seção 5.4.1.

Para as Tabelas A.7, A.8, A.9, A.10, A.11 e A.12 foram utilizados os dados da Tabela 5.10, página 50, como conjunto de contigs **inicial**.

Candidatos: Conjunto de contigs **inicial** com índice de $ITG \geq 0.5$.

Outros: Conjunto de contigs **candidatos** que não se encaixam nas definição de TBR ou PQ.

Tabela A.7: Análise qualitativa de contigs $\geq 500pb$ da compostagem **ZC4** montados com *SoapDeNovo 2* - Espécie.

ZC4 - SoapDeNovo 2 ($\geq 500pb$) - Espécie							
Dia	valores percentuais			Valores absolutos			Candidatos
	PQ	Outros	TBR	PQ	Outros	TBR	
01	5.41%	60.53%	34.06%	2366	26480	14898	43744
03	8.24%	61.87%	29.89%	3732	28013	13533	45278
07	10.28%	60.86%	28.86%	2906	17208	8159	28273
15	10.22%	57.74%	32.04%	2693	15217	8445	26355
30	9.23%	63.09%	27.68%	3066	20956	9195	33217
64	8.47%	64.77%	26.76%	5607	42856	17708	66171
67	9.48%	62.16%	28.36%	2083	13651	6228	21962
78	8.59%	60.60%	30.80%	2233	15751	8006	25990
99	9.63%	60.74%	29.63%	3329	21000	10244	34573

Tabela A.8: Análise qualitativa de contigs $\geq 500pb$ da compostagem ZC4 montados com *SoapDeNovo 2* - Gênero.

ZC4 - SoapDeNovo 2 ($\geq 500pb$) - Gênero							
Dia	valores percentuais			Valores absolutos			Candidatos
	PQ	Outros	TBR	PQ	Outros	TBR	
01	2.82%	51.54%	45.64%	1289	23566	20868	45723
03	4.84%	58.39%	36.78%	2336	28186	17754	48276
07	6.66%	58.94%	34.39%	1982	17532	10230	29744
15	7.42%	56.06%	36.52%	2050	15486	10090	27626
30	7.04%	60.56%	32.40%	2475	21280	11383	35138
64	5.63%	62.16%	32.21%	3933	43455	22516	69904
67	7.11%	62.33%	30.56%	1654	14500	7110	23264
78	6.19%	59.49%	34.32%	1686	16194	9343	27223
99	7.18%	60.34%	32.49%	2607	21920	11803	36330

Tabela A.9: Análise qualitativa de contigs $\geq 500pb$ da compostagem ZC4 montados com *SoapDeNovo 2* - Família.

ZC4 - SoapDeNovo 2 ($\geq 500pb$) - Família							
Dia	valores percentuais			Valores absolutos			Candidatos
	PQ	Outros	TBR	PQ	Outros	TBR	
01	2.61%	43.73%	53.67%	1169	19588	24041	44798
03	4.58%	53.12%	42.30%	2167	25110	19995	47272
07	6.48%	55.52%	38.00%	1858	15931	10905	28694
15	6.72%	52.08%	41.20%	1797	13924	11015	26736
30	6.48%	56.56%	36.96%	2205	19233	12567	34005
64	4.61%	56.87%	38.53%	3162	39037	26448	68647
67	5.99%	59.49%	34.52%	1361	13507	7838	22706
78	5.52%	54.18%	40.30%	1463	14369	10689	26521
99	5.82%	56.47%	37.72%	2066	20052	13394	35512

Tabela A.10: Análise qualitativa de contigs $\geq 500pb$ da compostagem ZC4 montados com *Mira* - Espécie.

ZC4 - Mira ($\geq 500pb$) - Espécie							
Dia	valores percentuais			Valores absolutos			Candidatos
	PQ	Outros	TBR	PQ	Outros	TBR	
01	13.34%	63.57%	23.08%	3832	18256	6628	28716
03	19.39%	61.73%	18.88%	5737	18265	5586	29588
07	22.05%	55.82%	22.13%	5307	13435	5325	24067
15	21.73%	51.89%	26.38%	5365	12811	6513	24689
30	21.26%	59.82%	18.93%	5130	14436	4568	24134
64	23.35%	59.91%	16.74%	11434	29340	8197	48971
67	26.61%	51.47%	21.92%	4604	8904	3793	17301
78	21.52%	55.31%	23.16%	4679	12026	5036	21741
99	29.60%	53.64%	16.75%	9388	17012	5313	31713

Tabela A.11: Análise qualitativa de contigs $\geq 500pb$ da compostagem ZC4 montados com *Mira* - Gênero.

ZC4 - Mira ($\geq 500pb$) - Gênero							
Dia	valores percentuais			Valores absolutos			Candidatos
	PQ	Outros	TBR	PQ	Outros	TBR	
01	7.16%	53.73%	39.11%	2076	15572	11335	28983
03	12.36%	64.69%	22.95%	3733	19545	6934	30212
07	15.64%	58.58%	25.77%	3810	14268	6277	24355
15	16.84%	53.80%	29.36%	4207	13440	7333	24980
30	16.98%	60.18%	22.84%	4174	14798	5616	24588
64	16.12%	63.57%	20.32%	8007	31582	10093	49682
67	20.80%	56.83%	22.36%	3654	9982	3928	17564
78	15.45%	58.62%	25.93%	3389	12856	5686	21931
99	22.30%	59.20%	18.50%	7180	19058	5956	32194

Tabela A.12: Análise qualitativa de contigs $\geq 500pb$ da compostagem ZC4 montados com *Mira* - Gênero.

ZC4 - Mira ($\geq 500pb$) - Família							
Dia	valores percentuais			Valores absolutos			Candidatos
	PQ	Outros	TBR	PQ	Outros	TBR	
01	7.06%	41.03%	51.91%	2020	11741	14857	28618
03	11.43%	57.64%	30.93%	3411	17205	9234	29850
07	15.84%	54.01%	30.15%	3779	12883	7191	23853
15	16.82%	50.81%	32.37%	4080	12327	7853	24260
30	17.60%	56.33%	26.07%	4201	13448	6225	23874
64	13.79%	57.24%	28.96%	6764	28068	14201	49033
67	18.17%	56.00%	25.84%	3161	9744	4496	17401
78	14.77%	52.31%	32.92%	3172	11235	7071	21478
99	18.91%	56.72%	24.37%	6003	18001	7735	31739

A.3 – Tabelas utilizadas para produzir gráficos apresentados na seção 5.4.2.

Tabela A.13: Total geral de bases montadas - em pares de bases - Todos os contigs $\geq 500pb$ gerados pelos montadores *SoapDeNovo 2* e *Mira* a partir das amostras da compostagem ZC4

ZC4 - Total geral de bases montadas		
Montador	SoapDeNovo 2	Mira
01	64,791,721	61,146,829
07	72,957,424	77,991,053
15	51,494,757	74,945,241
30	63,809,094	84,449,027
64	66,435,113	65,120,098
67	111,647,487	115,845,660
78	38,017,680	45,575,039
99	87,383,425	143,589,065

Tabela A.14: Total de bases TBR montadas - em pares de bases - Todos os contigs TBR $\geq 500pb$ gerados pelos montadores *SoapDeNovo 2* e *Mira* a partir das amostras da compostagem ZC4.

ZC4 - Total de bases TBR montadas - Gênero		
Montador	SoapDeNovo 2	Mira
01	14,546,777	10,176,631
07	11,997,788	6,688,146
15	7,213,242	6,066,409
30	7,327,996	6,589,001
64	8,268,433	5,311,825
67	14,960,509	8,228,006
78	4,541,763	3,248,601
99	6,671,347	5,407,257

Tabela A.15: Total de bases PQ montadas - em pares de bases - Todos os contigs PQ $\geq 500pb$ gerados pelos montadores *SoapDeNovo 2* e *Mira* a partir das amostras da compostagem ZC4.

ZC4 - Total de bases PQ montadas - Gênero		
Montador	SoapDeNovo 2	Mira
01	1,328,368	3,650,273
07	2,197,184	6,555,713
15	1,754,047	6,267,971
30	1,683,286	4,298,560
64	2,291,061	5,046,251
67	3,754,681	13,739,989
78	1,358,912	4,750,984
99	1,303,974	3,264,184

Apêndice **B**

Banco de dados

B.1 – MER

79

Figura B.1: Modelo Entidade Relacionamento (MER) - Caravela

Referências Bibliográficas

- [1] Bootstrap. <http://getbootstrap.com/>.
- [2] Feature viewer. <https://github.com/calipho-sib/feature-viewer>.
- [3] Gson. <https://github.com/google/gson>.
- [4] Hibernate. <http://hibernate.org/>.
- [5] Htsjdk - a java api for high-throughput sequencing data (hts) formats. <https://samtools.github.io/htsjdk/>.
- [6] jquery. <https://jquery.com/>.
- [7] Sam tools. <https://samtools.sourceforge.net/>.
- [8] Vraptor. <http://www.vraptor.org/>.
- [9] S. Altschul. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, Jan 1997.
- [10] Luciana Principal Antunes, Layla Farage Martins, Roberta Verciano Pereira, Andrew Maltez Thomas, Deibs Barbosa, Leandro Nascimento Lemos, Gianluca Major Machado Silva, Livia Maria Silva Moura, George Willian Condomitti Epamino, Luciano Antonio Digiampietri, et al. Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics. *Scientific Reports*, 6, 2016.
- [11] Daniel Beck, Christopher Dennis, and James A Foster. Seed: a user-friendly tool for exploring and visualizing microbial community data. *Bioinformatics*, 31(4):602–603, 2015.
- [12] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. Genbank. *Nucleic Acids Research*, 36(Database), 2007.
- [13] Andrew Bissett, Alan E Richardson, Geoff Baker, and Peter H Thrall. Long-term land use effects on soil microbial community structure and function. *Applied Soil Ecology*, 51:66–78, 2011.
- [14] Michael Cantor, Henrik Nordberg, Tatyana Smirnova, Matthias Hess, Susannah Tringe, and Inna Dubchak. Elviz – exploration of metagenome assemblies with an interactive visualization tool. *BMC Bioinformatics*, 16(1), 2015.
- [15] Kevin Chen and Lior Pachter. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Computational Biology PLoS Comp Biol*, 1(2), 2005.
- [16] Bastien Chevreux. Mira: an automated genome and est assembler. 2007.

- [17] T Andrew Clayton, David Baker, John C Lindon, Jeremy R Everett, and Jeremy K Nicholson. Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism. *Proceedings of the National Academy of Sciences*, 106(34):14728–14733, 2009.
- [18] MJ Colloff, SA Wakelin, D Gomez, and SL Rogers. Detection of nitrogen cycle genes in soils for measuring the effects of changes in land use and management. *Soil Biology and Biochemistry*, 40(7):1637–1645, 2008.
- [19] Victor de Lorenzo. Blueprint of an oil-eating bacterium. *Nature biotechnology*, 24(8):952–953, 2006.
- [20] R. C. Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, Dec 2010.
- [21] Sara El-Metwally, Taher Hamza, Magdi Zakaria, and Mohamed Helmy. Next-generation sequence assembly: Four stages of data processing and computational challenges. *PLoS Computational Biology PLoS Comput Biol*, 9(12), Dec 2013.
- [22] Mallory Embree, Joanne K Liu, Mahmoud M Al-Bassam, and Karsten Zengler. Networks of energetic and metabolic interactions define dynamics in microbial communities. *Proceedings of the National Academy of Sciences*, 112(50):15450–15455, 2015.
- [23] Steven Flygare, Keith Simmon, Chase Miller, Yi Qiao, Brett Kennedy, Tonya Di Sera, Erin H Graf, Keith D Tardif, Aurélie Kapusta, Shawn Ryneerson, et al. Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mrna expression profiling. *Genome biology*, 17(1):1, 2016.
- [24] Eric A. Franzosa, Tiffany Hsu, Alexandra Sirota-Madi, Afrah Shafquat, Galeb Abu-Ali, Xochitl C. Morgan, and Curtis Huttenhower. Sequencing and beyond: integrating molecular ‘omics’ for microbial community profiling. *Nat Rev Micro Nature Reviews Microbiology*, 13(6):360–372, 2015.
- [25] Najoshi G. Sickel - a windowed adaptive trimming tool for fastq files using quality. <https://github.com/najoshi/sickel>.
- [26] Elizabeth M. Glass and Folker Meyer. The metagenomics rast server: A public resource for the automatic phylogenetic and functional analysis of metagenomes. *Metagenomics and Complementary Approaches Handbook of Molecular Microbial Ecology I*, page 325–331, Mar 2011.
- [27] Rani Gupta and Harapriya Mohapatra. Microbial biomass: an economical alternative for removal of heavy metals from waste water. *Indian journal of experimental biology*, 41(9):945–966, 2003.
- [28] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. Quast: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013.
- [29] Sarah Hunter, Matthew Corbett, Hubert Denise, Matthew Fraser, Alejandra Gonzalez-Beltran, Christopher Hunter, Philip Jones, Rasko Leinonen, Craig McAnulla, Eamonn Maguire, et al. Ebi metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic acids research*, 42(D1):D600–D606, 2014.
- [30] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic Acids Research*, 42(D1), Jul 2013.
- [31] Allan Konopka. What is microbial community ecology&quest. *The ISME journal*, 3(11):1223–1230, 2009.

- [32] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.
- [33] Ruth E Ley. Obesity and the human microbiome. *Current opinion in gastroenterology*, 26(1):5–11, 2010.
- [34] C. Luo, L. M. Rodriguez-R, and K. T. Konstantinidis. Mytaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Research*, 42(8), Mar 2014.
- [35] Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, and et al. Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience Giga Sci*, 1(1):18, 2012.
- [36] Chaysavanh Manichanh, Lionel Rigottier-Gois, Elian Bonnaud, Karine Gloux, Eric Pelletier, Lionel Frangeul, Renaud Nalin, Cyrille Jarrin, Patrick Chardon, Phillipe Marteau, et al. Reduced diversity of faecal microbiota in crohn’s disease revealed by a metagenomic approach. *Gut*, 55(2):205–211, 2006.
- [37] Julian R Marchesi, Bas E Dutilh, Neil Hall, Wilbert HM Peters, Rian Roelofs, Annemarie Boleij, and Harold Tjalsma. Towards the human colorectal cancer microbiome. *PloS one*, 6(5):e20447, 2011.
- [38] V. M. Markowitz, I.-M. A. Chen, K. Chu, E. Szeto, K. Palaniappan, M. Pillay, A. Ratner, J. Huang, I. Pagani, S. Tringe, and et al. IMG/m 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research*, 42(D1), 2013.
- [39] I. Milne, M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright, and D. Marshall. Tablet—next generation sequence assembly visualization. *Bioinformatics*, 26(3):401–402, Apr 2009.
- [40] Suparna Mitra, Mario Stärk, and Daniel H Huson. Analysis of 16s rRNA environmental sequences using megan. *BMC Genomics*, 12(Suppl 3), 2011.
- [41] Niranjan Nagarajan and Mihai Pop. Sequence assembly demystified. *Nat Rev Genet Nature Reviews Genetics*, 14(3):157–167, 2013.
- [42] Rachid Ounit and Stefano Lonardi. Higher classification sensitivity of short metagenomic reads with clark-s. *Bioinformatics*, 32(24):3823–3825, 2016.
- [43] S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, and et al. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Research*, 40(D1), 2011.
- [44] M. Punta, P. C. Coghill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, and et al. The pfam protein families database. *Nucleic Acids Research*, 40(D1), 2011.
- [45] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.
- [46] R. L. Tatusov. A genomic perspective on protein families. *Science*, 278(5338):631–637, 1997.
- [47] H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, 2012.
- [48] Harold Tjalsma, Annemarie Boleij, Julian R Marchesi, and Bas E Dutilh. A bacterial driver–passenger model for colorectal cancer: beyond the usual suspects. *Nature Reviews Microbiology*, 10(8):575–582, 2012.

- [49] Peter J Turnbaugh, Ruth E Ley, Michael A Mahowald, Vincent Magrini, Elaine R Mardis, and Jeffrey I Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *nature*, 444(7122):1027–131, 2006.
- [50] Marcel GA Van Der Heijden, Richard D Bardgett, and Nico M Van Straalen. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecology letters*, 11(3):296–310, 2008.
- [51] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*, 15(3):R46, 2014.
- [52] Z. Zhu, B. Niu, J. Chen, S. Wu, S. Sun, and W. Li. Mgaviewer: a desktop visualization tool for analysis of metagenomics alignment data. *Bioinformatics*, 29(1):122–123, Aug 2012.