UNIVERSIDADE DE SÃO PAULO

Escola de Artes, Ciências e Humanidades

Vivian Mayumi Yamassaki Pereira

Montagem e análise de genomas a partir de metagenomas

São Paulo
Novembro de 2014

Universidade de São Paulo
Escola de Artes, Ciências e Humanidades

Vivian Mayumi Yamassaki Pereira

Montagem e análise de genomas a partir de metagenomas

Monografia a ser apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, como parte dos requisitos exigidos para aprovação na disciplina ACH2018 – Projeto Supervisionado ou de Graduação II, do curso de Bacharelado em Sistemas de Informação.

Modalidade: TCC Longo (1 ano) – individual.

Orientador:

Prof. Dr. Luciano Antonio Digiampietri

São Paulo

Novembro de 2014

Universidade de São Paulo
Escola de Artes, Ciências e Humanidades

Vivian Mayumi Yamassaki Pereira

Montagem e análise de genomas a partir de metagenomas

Monografia a ser apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, como parte dos requisitos exigidos para aprovação na disciplina ACH2018 – Projeto Supervisionado ou de Graduação II, do curso de Bacharelado em Sistemas de Informação.

Modalidade: TCC Longo (1 ano) – individual.

Data de Aprovação:

19/11/2014

Banca Examinadora:

Profª Drª Karina Valdivia Delgado
EACH-USP

São Paulo
Novembro de 2014

Glossário

Análise filogenética: estudo da relação entre os organismos, no qual se pode verificar quão próximos, evolutivamente, eles estão uns dos outros. Para isso, são analisados os dados genéticos desses organismos.

Contig: sequências maiores de DNA que são montadas por meio de sobreposição de *reads*.

Read: resultado obtido após o sequenciamento de genomas por sequenciadores de alto desempenho. Corresponde a uma pequena sequência de poucos pares de base.

Resumo

A microbiologia foi impactada por dois fatores que têm permitido um número cada vez maior de descobertas. O primeiro deles foi a descoberta da importância que o estudo da metagenômica possui para a ciência, visto que grande parte da diversidade de micro-organismos encontra-se no meio ambiente e não é possível realizar a sua cultura em laboratório. O segundo diz respeito ao avanço tecnológico iniciado com equipamentos para a visualização dos micro-organismos, passando pelo sequenciamento de bactérias nos anos 90 até os sequenciadores de alto desempenho utilizados atualmente. O uso desses equipamentos para obtenção de dados biológicos, apesar de permitir que muitas descobertas fossem e sejam realizadas, provocou desafios com relação à organização e análise dos dados gerados. O objetivo deste trabalho foi analisar dados dos genomas que fazem parte de sequenciamentos de metagenomas. Em particular, abordou-se a montagem de genomas a partir de sequências de metagenomas e a análise comparativa destes genomas com outros mais similares para completar montagens parciais e/ou identificar a provável filogenia de novas espécies. Como resultado, foram desenvolvidas ferramentas para realizar a identificação das espécies presentes em sequências metagenômicas, ferramentas para auxiliar no processo de montagem e para realizar a análise taxonômica das sequências montadas de modo a identificar a similaridade entre os genomas dos micro-organismos presentes no metagenoma.

Palavras-Chave: montagem e análise de genomas, análise filogenética, metagenômica, taxonomia.

Abstract

The microbiology was impacted by two factors that allowed an increased number of new discoveries. The first one was the discover of the importance in metagenomics study to science, since great part of microorganisms diversity lies in the environment and it can not be cultured on laboratories. The second one was the technology advance which started with the visualization equipments of microorganisms to the bacteria's sequencing in the 90s decade until the high performance sequencers used nowadays. The use of these equipments to obtain biologic data, despite have allowed a great number of discoveries has been made, brought challenges like data's organization and its analysis. The objective of this paper was analyse genomes data which is part of metagenomes sequencing. In particular, the approach was the genomes assembly by metagenomes sequencies and a comparative analysis of these genomes with similaries ones in order to complete partial assembly and/or identify new species probably philogeny. As a result, some tools were developed to identify the species of metagenomics sequences, for helping in the assembly process, and to analyse the taxonomy of the assembled sequences in order to identify the similarity between microorganisms genomes in metagenome.

Keywords: assembly and genome analysis, phylogenetic analysis, metagenomics, taxonomy.

Lista de Figuras

Sumário

1 Introdução

A ciência atual tem sido revolucionada pelo avanço tecnológico dos equipamentos utilizados, cuja capacidade de aquisição de dados cresce cada vez mais (HEY; TANSLEY; TOLLE, 2009). Um dos campos da ciência afetados por esse avanço foi a biologia, visto que a evolução dos equipamentos permitiu que dados biológicos fossem produzidos rapidamente e em grande volume. Por este motivo, o uso de computadores para manipular e analisar esses dados tornou-se imprescindível nas pesquisas biológicas.

A área de bioinformática está atrelada a este contexto, uma vez que trata-se da aplicação de técnicas provenientes dos campos da matemática, estatística e da ciência da computação para entender e organizar, em larga escala, a informação associada aos dados biológicos. Em resumo, pode-se dizer que bioinformática é um sistema de informação de gestão para a biologia molecular e que possui diversas aplicações práticas (LUSCOMBE; GREENBAUM; GERSTEIN, 2001). Um dos campos mais conhecidos da bioinformática é o relacionado à microbiologia computacional ou, em particular, à montagem e anotação de genomas.

O avanço nos processos de montagem e anotação de genomas, portanto, também está intimamente ligado ao avanço tecnológico dos equipamentos para sequenciamento de DNA, cujas técnicas foram aprimoradas e desenvolvidas a partir de 1970. Até o final da década de 1990, o sequenciamento, montagem e anotação do genoma de uma única bactéria, cujo genoma é tipicamente composto por poucos milhões de pares de bases, era uma tarefa custosa (tanto financeiramente quanto no tempo necessário para ser realizada) (SETUBAL; MEIDANIS, 1997). Com os sequenciadores de alto desempenho desenvolvidos nos últimos anos, tornou-se possível, em um único sequenciamento, a obtenção de grande volume de DNA (dezenas de milhões de bases) que, por tratar-se de um grande volume de dados, também trouxe novos desafios. Um dos principais está relacionado ao fato de que muitas vezes, em um único sequenciamento, é sequenciado material genético de centenas de indivíduos de milhares de espécies diferentes e o objetivo desse tipo de estudo é realizar a análise da genômica de um dado nicho específico (por exemplo, do estômago de um animal, do solo ou da água de locais específicos). Estes projetos são tipicamente chamados de projetos de genômica ambiental ou metagenômica, em

que, pelo fato do DNA ser originado de diversas populações, a recuperação dos genomas acaba se tornando uma tarefa complexa (SHARON; BANFIELD, 2013).

Apesar dessa complexidade, o estudo da metagenômica é de extrema importância pois, durante muito tempo, cientistas se dedicaram ao estudo de micro-organismos cuja cultura podia ser feita em laboratório, devido ao poder e precisão dos estudos de bactérias nessas condições. Entretanto, percebeu-se que a cultura não capturava a grande diversidade microbiana existente, que se encontra principalmente entre os micro-organismos em que não é possível fazer a cultura. Portanto, neste contexto, a metagenômica tornou-se uma peça central para se obter conhecimento sobre a fisiologia e genética desses micro-organismos (HANDELSMAN, 2004).

Nos projetos de genômica ambiental, o resultado do sequenciamento corresponde a milhões de pequenos pedaços de DNA (chamados de *reads*, com dezenas ou poucas centenas de pares de bases) e sem identificação de qual organismo este DNA foi extraído. Assim, um problema inicial é tentar agrupar os *reads* de acordo com a espécie a que pertencem e, utilizando-se algoritmos baseados na sobreposição de sequências de DNA, sobrepor estes *reads* em sequências maiores (chamadas de *contigs*) a fim de tentar, se possível, reconstruir a sequência de DNA do genoma completo daquela espécie (processo este conhecido como montagem do genoma).

Dois dos principais desafios relacionados a esta atividade são a identificação das sequências (*reads*) que pertencem a cada espécie e a montagem em si dos genomas, pois provavelmente existirão partes faltantes (que não permitirão uma montagem completa) e muito dos genomas possuem regiões repetitivas (o que dificulta a montagem por sobreposição).

Após a identificação das prováveis espécies a que pertence cada *read* e a montagem destes genomas, surgem diversos desafios na análise destes dados. Dentre eles estão: (i) a análise da quantidade e diversidade dos organismos encontrados, tarefa que se torna mais interessante quando são feitos diferentes sequenciamentos de um mesmo nicho ecológico, por exemplo, um sequenciamento por mês ou por estação do ano, pois é possível realizar uma análise comparativa entre as informações de cada sequenciamento e tentar identificar as razões bioquímicas para as variações encontradas; (ii) verificação se alguma das espécies sequenciadas provavelmente corresponde a uma nova espécie (nunca sequenciada previamente); este desafio pode ser enfrentado realizando-se análises filogenéticas, em que se pode obter estimativas mais robustas (SUNAGAWA et al., 2013) e ter maior poder de discriminação de espécies quando realizada a concatenação de diversos genes (DEVULDER; MONTCLOS; FLANDROIS, 2005); (iii) identificação de genes de interesse (para alguma aplicação industrial, por exemplo); (iv) identificação dos processos metabólicos que estão ocorrendo ao longo do tempo naquele nicho (utilizando, por exemplo, informações sobre genes expressos).

Este trabalho está contextualizado dentro do Núcleo de Pesquisa em Ciência Genômica (NAP-CG) da Universidade de São Paulo[1] que reúne pesquisadores em biociências e pesquisadores em bioinformática para pesquisa e trabalho conjunto em projetos de ciências genômicas. O princípio organizador do núcleo são projetos motores, que são projetos com base genômica, com necessidades sofisticadas de bioinformática e liderados por participantes do núcleo. Especificamente, este projeto irá auxiliar na realização das atividades do projeto motor Metagenômica de microbiomas do Zoológico de São Paulo[2] (MARTINS et al., 2013).

[1] http://www.iq.usp.br/napcg/
[2] http://www.iq.usp.br/setubal/metazoo.html

2 Objetivos

2.1 Objetivo Geral

O objetivo geral deste trabalho foi desenvolver ferramentas para auxiliar no processo de montagem de genomas a partir de metagenomas e realizar uma análise filogenética comparando as informações dos genomas montados (total ou parcialmente) em relação aos genomas mais próximos.

2.2 Objetivos Específicos

Os objetivos específicos deste trabalho foram:

- Desenvolver/estender técnicas para a identificação automática da provável espécie a que pertence cada *read*;
- Agrupar os *reads* de cada espécie e realizar a montagem destes *reads*;
- Especificar uma metodologia e desenvolver novas ferramentas para automatizar a análise filogenética dos genomas montados tentando utilizar toda a informação disponível do genoma e não apenas genes específicos.

3 Revisão bibliográfica

Por conta da importância científica que o estudo de metagenomas possui, diversas ferramentas e metodologias para realização da montagem de genomas a partir de metagenomas e para análise filogenética foram desenvolvidas nos últimos anos. Nesse capítulo, serão abordadas algumas dessas ferramentas, que foram encontradas por meio da revisão bibliográfica.

O programa MEGAN (HUSON et al., 2007) é uma ferramenta para análise de dado de metagenomas. Para realizar tal análise, utiliza a taxonomia proveniente do NCBI e ferramentas, como o BLAST, para realizar a comparação dos *reads* de entrada com sequências já conhecidas. Com esses dados, a ferramenta utiliza a abordagem do menor ancestral em comum para realizar a atribuição taxonômica do *read* de acordo com o resultado da comparação. A classificação de um *read* em um determinado táxon (unidade taxonômica, que pode ser um reino ou espécie, por exemplo) se dá após a análise de 6 marcadores filogenéticos (rRNA, RecA/RadA, HSP70, RpoB, EF-Tu e Ef-G). A ferramenta realiza a montagem de *reads* provenientes de metagenomas e não estima a abundância quantitativa de organismos, pois só conta os *reads* mapeados para espécies e genes conhecidos, medida que é afetada pelo comprimento do gene ou pelo tamanho do genoma. Além disso, utiliza somente o *bit-score* como parâmetro para filtrar alinhamentos considerados insignificantes.

O SmashCommunity (ARUMUGAM et al., 2010) realiza a análise e anotação de metagenomas. As amostras fornecidas para a ferramenta podem ser classificadas filogeneticamente por meio da identificação dos *reads* que contenham sequências do gene 16S rRNA ou utilizando apenas as melhores correspondências encontradas pelo BLAST para referenciar os genomas taxonomicamente. O número de perfis filogenéticos gerados é calculado ao contar os *reads* classificados e corrigindo o tamanho do genoma ou da variação do número de cópias do gene 16S rRNA. A análise é feita com sequências geradas pelos sequenciadores de alto desempenho Sanger e 454 e, além disso, a ferramenta não realiza a montagem de metagenomas, que é feita por meio das ferramentas Arachne e Celera.

O algoritmo SOrt-ITEMS (HAQUE et al., 2009) realiza o processo de agrupamento *reads*

ou *contigs* para atribui-los a um mesmo nível taxonômico. Primeiramente, verifica a qualidade do alinhamento entre o *read* e a correspondente sequência alvo no BLAST e depois encontra o nível taxonômico ao qual o *read* pode ser atribuído. A qualidade é verificada por meio dos valores *bit-score*, *alignment length* e *percentage of identities* de cada alinhamento. Em seguida, o algoritmo identifica sequências que apresentam similaridade significante com o *read* de consulta. Essas sequências identificadas são então utilizadas para a atribuição final de um *read*. Para encontrar o nível taxonômico ao qual o *read* deve ser atribuído, o algoritmo também utiliza o conceito de menor ancestral em comum. Uma análise realizada indicou que esse algoritmo realizou menos erros ao atribuir os níveis taxonômicos do que a ferramenta MEGAN, o que indica que é importante levar em consideração outros parâmetros do alinhamento, além do *bit-score*, para atribuir um determinado nível taxonômico a um *read*.

O CARMA3 (GERLACH; STOYE, 2011) é um algoritmo inspirado em outros, como o SOrt-ITEMS, que realiza a classificação taxonômica de sequências metagenômicas montadas ou não. Para tanto, utiliza o BLAST e o HMM3, que é uma ferramenta para análise de sequências através do uso do modelo oculto de Markov, além de também utilizar a abordagem de menor ancestral em comum para realizar a classificação. O método utilizado por este algoritmo assume um modelo de evolução no qual os diferentes famílias de genes apresentam diferentes taxas de mutação, mas dentro de cada família essas taxas não variam demasiadamente.

A ferramenta MetaPhyler (LIU; GIBBONS; POP, 2010) baseia-se em 31 genes como marcadores filogenéticos para referência taxonômica. O classificador baseia-se no BLAST e utiliza diferentes limiares para cada um dos parâmetros que são automaticamente aprendidos a partir da estrutura da base de dados de referência. Com isso, a ferramenta pode identificar novos organismos ou táxons. Testes realizados com a ferramenta demostraram que o seu desempenho foi superior às ferramentas CARMA3 e MEGAN.

Genometa (DAVENPORT et al., 2012) é uma ferramenta que realiza a atribuição taxonômica a pequenos *reads* de metagenomas de procariontes. Isto permite que os usuários verifiquem a distribuição de *reads* pelo genoma de cada espécie identificada para que possam remover os alinhamentos que considerem atrapalhar a classificação. Ela só atribui *reads* às espécies cujos genomas estão contidos nas sequências de referência utilizadas. Com isso, a ferramenta possui uma boa performance ao analisar dados de metagenomas com a microbiota bem caracterizada em termos de genomas de referência, mas não analisa tão bem amostras provenientes do meio ambiente.

Por fim, há o sistema MG-RAST (MEYER et al., 2008) que possui algumas funcionalidades, como comparar os dados fornecidos pelo usuário com outros metagenomas ou genomas

completos ou comparar o metabolismo e anotações de um ou mais genomas ou metagenomas. A ferramenta permite que o usuário escolha a abordagem (utilização do 16S rRNA ou a partir de resultados do BLAST) para realizar a comparação do metagenoma. Assim como a ferramenta MEGAN, não estima a abundância quantitativa de organismos.

4 Metodologia

A metodologia deste trabalho consistiu, primeiramente, no estudo de diversos trabalhos relacionados à metagenômica, montagem e anotação de genomas e análises filogenéticas publicados nos últimos anos, por meio de uma revisão bibliográfica. Este estudo focou-se nos trabalhos que apresentavam metodologias para a montagem de genomas a partir de metagenomas e para a análise filogenética. Para tanto, as pesquisas pelos artigos foram realizadas nas bibliotecas eletrônicas da ACM (Association for Computing Machinery)[1], IEEE (Institute of Electrical and Electronics Engineers)[2] e PLOS ONE[3]. Adicionalmente, foram feitas pesquisas em outras revistas científicas que possuíssem artigos a respeito de ferramentas ou metodologia citadas pelos artigos encontrados nas três bibliotecas digitais utilizadas.

A partir deste estudo foi especificada, implementada e testada uma ferramenta para a identificação das espécie a que pertencia cada sequência e mesmo o agrupamento de sequências (que potencialmente pertençam a mesma espécie, por mais que ainda seja desconhecida/não sequenciada) para que fosse realizada a montagem destas sequências. A identificação das espécies baseou-se nos resultados de ferramentas de alinhamento local que compararam as sequências de entrada com o banco de sequência de nucleotídeos não-redundantes do NCBI (*National Center for Biotechnology Information*)[4]. As ferramentas de alinhamento local utilizadas foram o USEARCH[5] e BLAST[6]. Já a montagem das sequências foi realizada pela ferramenta Newbler[7]. Além disso, foi necessária a utilização da taxonomia proveniente também do NCBI para realizar a identificação e classificação das sequências.

As ferramentas de alinhamento foram configuradas para gerar arquivos no formato m8, que são arquivos tabulados com 12 campos com informações sobre o alinhamento. Esses campos são: *query name*, que é o nome da sequência de consulta; *subject name*, que é nome da

[1] http://dl.acm.org/
[2] http://ieeexplore.ieee.org/Xplore/
[3] http://www.plosone.org/
[4] http://www.ncbi.nlm.nih.gov
[5] http://www.drive5.com/usearch
[6] http://blast.be-md.ncbi.nlm.nih.gov/Blast.cgi
[7] http://454.com/products/analysis-software/index.asp

sequência cujo alinhamento foi encontrado; o campo *percent identities* é porcentagem de bases idênticas; *aligned length* é o tamanho do alinhamento; *number of mismatched positions*, que é o número de alinhamentos entre bases diferentes; *number of gap positions*, que se trata do número de espaços em branco (buracos) no alinhamento; *query sequence start*, que é a posição inicial na sequência de consulta, ou seja, onde o alinhamento começou; *query sequence end*, que é a posição final na sequência de consulta (onde o alinhamento terminou); *subject sequence start*, que é a posição inicial na sequência encontrada; o *subject sequence end* é a posição final na sequência encontrada; *e-value*, que é a probabilidade do alinhamento ser uma coincidência; e o *bit-score*, que é a nota dada ao alinhamento. Portanto, são esses os dados de entrada para a ferramenta de identificação de espécie.

Com base nos genomas montados (parcialmente ou totalmente) foram especificadas, desenvolvidas e testadas ferramentas para a realização de análises filogenéticas objetivando-se utilizar a maior quantidade possível de informações para esta análise e não apenas os genes específicos tipicamente utilizados nesta análise.

Todas as ferramentas foram desenvolvidas utilizando-se a linguagem de programação Java e houveram chamadas das ferramentas desenvolvidas para ferramentas já implementadas em Java e Perl.

Além disso, foram criados scripts em R para a geração de diagramas que possibilitaram a análise filogenética. Para que fosse possível criar tais diagramas, foi necessário instalar o pacote ape [8].

As implementações resultantes foram testadas com dados reais e os resultados foram analisados e validados com a ajuda de um especialista do domínio e com base em outras análises já realizadas e curadas manualmente. Esses dados reais fornecidos como entrada para as ferramentas foram sequências metagenômicas da compostagem realizada no Zoológico de São Paulo, provenientes de amostras coletadas que foram sequenciadas pelo sequenciador de alto desempenho Roche 454 GS FLX Titanium (MARTINS et al., 2013).

[8] http://cran.r-project.org/web/packages/ape/

5 Resultados

Como resultado, foram desenvolvidas ferramentas que tratam da montagem e análise de genomas a partir de metagenomas. A Figura 5.1 apresenta a interação com as ferramentas desenvolvidas, destacadas em amarelo, com outras ferramentas já desenvolvidas e com o conjunto de dados utilizado.

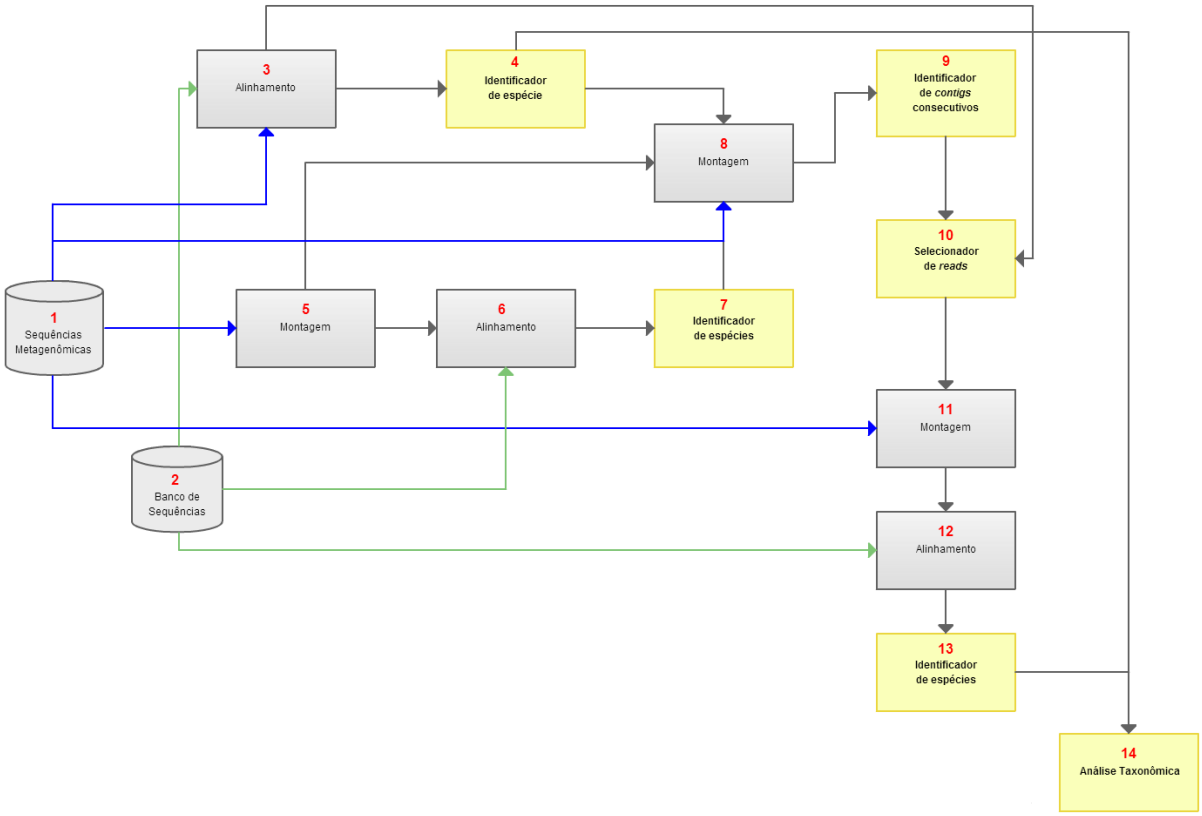

Figura 5.1 *– Visão geral com todas ferramentas desenvolvidas e utilizadas*

De acordo com a Figura 5.1 têm-se em 1 as sequências metagenômicas geradas a partir das amostras coletadas da compostagem realizada no Zoológico de São Paulo que foram sequenciadas pelo sequenciador de alto desempenho Roche 454 GS FLX Titanium. Os *reads* gerados por esse sequenciamento, ou seja, as sequências metagenômicas, foram então comparados com as sequências de nucleotídeos não-redundantes do banco do NCBI (representado na Figura 5.1 como 2 - banco de sequências) por meio das ferramentas de alinhamento lo-

cal USEARCH e BLAST, em 3. Os alinhamentos gerados por essas ferramentas foram então utilizados como entrada pela ferramenta desenvolvida neste trabalho, em 4, que identifica a classificação taxonômica mais provável com base nos alinhamentos por meio da utilização da taxonomia proveniente do NCBI.

A ferramenta de identificação de espécies possui filtros que são parametrizados pelo usuário, de modo que só classifica as sequências que estão acima do limiar escolhido. Os filtros dizem respeito aos valores mínimos e máximos que 6 dos 12 campos do arquivo m8 devem satisfazer para que o alinhamento seja considerado bom e são os seguintes: *percent identities* mínimo, *aligned length* mínimo, *number of mismatched positions* máximo, *number of gap positions* máximo, *e-value* máximo e o *bit-score* mínimo. Se todos os filtros forem satisfeitos para um determinado alinhamento, então é realizada a identificação da espécie daquela sequência.

Além disso, também foi realizada uma montagem (em 5), com a ferramenta Newbler, utilizando todos os *reads* (sem a prévia separação dos *reads* por genoma) e, com os *contigs* montados a partir da sobreposição dos *reads*, foi realizado o alinhamento com as sequências do banco do NCBI, na etapa identificada como 6. Em seguida, na etapa 7, uma ferramenta desenvolvida analisa a montagem e os alinhamentos realizados identificando-se quantos *reads* de cada espécie um *contig* possui e quais *reads* ficaram sem identificação de espécie.

Com os alinhamentos identificados por espécie, o resultado da montagem dos *reads* (sem a prévia separação dos *reads* por genoma) e a identificação de quantos *reads* de cada espécie um *contigs* possui e quantos ficaram sem identificação, os *reads* foram agrupados de acordo com a espécie ao qual provavelmente pertenciam e, após esse agrupamento, foi realizado o processo de montagem novamente, que ocorre em 8.

As sequências montadas serviram de entrada para a ferramenta, identificada como 9, que identifica os *contigs* consecutivos[1], ou seja, verifica se há sequências que poderiam ser sobrepostas ou justapostas de modo a gerar sequências maiores, mas que o montador não conseguiu fazê-lo. Para tanto, essa ferramenta implementa um algoritmo de *backtracking* para que sempre encontre as melhores combinações de *contigs* que possam formar as maiores sequências possíveis para cada espécie.

Após essa identificação dos *contigs* consecutivos, foi utiliza outra ferramenta desenvolvida, em 10, que procura por *reads* que resolvam os conflitos e formem "pontes" para juntar os *contigs* encontrados pela ferramenta anterior de modo que a montagem possa gerar as sequências maiores. Essa procura pelos *reads* foi realizada por meio da observação dos campos *query sequence start* e *query sequence end*, presentes no arquivo m8 fornecido pelas ferramentas de

[1] Ferramenta desenvolvida pelo aluno de iniciação científica Geraldo José dos Santos Júnior

alinhamento local.

Com os dados gerados por essa ferramenta de seleção de *reads*, é realizada a montagem das sequências em 11. Essa montagem resultou em sequências maiores, com menos lacunas do que as montagens realizadas anteriormente.

Em seguida, foi realizado o alinhamento dessas sequências maiores em 12. As sequências maiores e com menos lacunas geradas pela montagem em 11 permitem um identificação mais precisa de qual espécia a sequência pertence ou mesmo se a sequência possivelmente pertence a uma espécie nova (nunca antes sequenciada). Posteriormente, a ferramenta de identificação de espécies foi utilizada para verificar a taxonomia desses alinhamentos gerados.

Em 13, foi realizada a identificação das espécies desses alinhamentos, que foi feita de modo que, se o *contig* não possui nenhum *read* com identificação de espécie, ou seja, se não houve alinhamento de seus *reads*, então o *contig* é classificado como "sem identificação"; se o *contig* foi montado com *reads* de uma única espécie, o *contig* é identificado como pertencente a essa espécie; e, no caso do *contig* possuir *reads* de diferentes espécies, ele é identificado pelo nível taxonômico em comum que as diferentes espécies compartilham. Essa abordagem é chamada de menor ancestral em comum (do inglês, *Lowest Common Ancestor* - LCA), que também foi utilizada por outras ferramentas como o MEGAN e o SOrt-ITEMS para classificar o *read* pelo nível taxonômico mais alto em comum que os organismos que alinharam com o *read* possuem. Um problema dessa abordagem é que alinhamentos insignificantes podem resultar na atribuição de níveis relativamente altos (classificação no reino Bacteria, por exemplo) ao *read*, o que acaba reduzindo a especificidade da montagem dos *reads* e *contigs* (HAQUE et al., 2009). Por isso, essa ferramenta desenvolvida também possui os mesmos filtros das ferramentas anteriores e que são parametrizados pelo usuário, de modo que apenas os alinhamentos considerados relevantes serão utilizados para fazer a classificação dos *contigs*. A ferramenta ainda permite que o usuário opte por utilizar apenas a primeira espécie que alinhou com cada *read* (isto é, o melhor alinhamento) ou todas as espécies que alinharam com um dado *read* para fazer a comparação descrita acima.

É importante destacar que o montador retorna o *status* da montagem realizada. Para cada *read* utilizado na montagem, esse *status* pode ser *Singleton*, o que significa que a sequência tinha condições de entrar na montagem, mas não "juntou" com nada; *TooShort*, indicando que a sequência foi considerada pequena demais para ser utilizada na montagem; *PartiallyAssembled*, ou seja, apenas parte sequência fez parte da montagem; e *Assembled*, o que indica que a sequência entrou na montagem. Com bases nessas informações fornecidas pelo montador, a ferramenta de identificação de espécies considera apenas os *reads* cujo *status* seja *PartiallyAssembled* ou

Assembled.

A partir dos dados gerados pela ferramenta de identificação de espécie, foi realizada a análise taxonômica por meio de ferramentas desenvolvidas em 14. Uma delas calcula a distribuição dos micro-organismos na amostra de entrada, de modo que a ferramenta retorna, em ordem decrescente, as espécies mais frequentes existentes na amostra. Essa frequência pode ser calculada tanto em termos de número de *reads* ou *contigs* pertencentes a uma determinada espécie ou pela soma do tamanho dos alinhamentos de sequências que foram identificados como sendo de uma determinada espécie. Os dados gerados por esta ferramenta permitem que sejam gerados gráficos para analisar visualmente a distribuição de micro-organismos e, com isso, averiguar quão similares as distribuições de espécies de determinados nichos são, como pode ser observado na Figura 5.2, cujas informações dizem respeito à amostra coletada da compostagem do Zoológico de São Paulo. Neste caso, a maior parte (cerca de 79%) dos *contigs* possuem apenas *reads* sem identificação de espécie. Para os que possuíam, com exceção dos casos em que um *contig* possuísse apenas *reads* de uma mesma espécie, houve um maior número de *reads* de um mesmo *contig* que tiveram a *Ordem* como nível taxonômico mais alto em comum, o que é um indício de que muitos dos muitos dos micro-organismos presentes na amostra da compostagem do Zoológico de São Paulo pertencem a espécies novas, ou seja, que ainda não foram sequenciadas. Além disso, pelo fato da ordem ser o nível taxonômico mais alto, é provável que essas possíveis novas espécies possuam uma filogenia com características similares a de outros micro-organismos já sequenciados, mas que pertencem a famílias diferentes.

Adicionalmente, também foi desenvolvida outra ferramenta que gera uma "matriz de proximidade" a partir dos dados dos alinhamentos das sequências metagenômicas para que seja possível realizar a análise filogenética das sequências. A métrica atualmente utilizada para preenchimento dessa matriz é: quanto maior soma do tamanho dos alinhamentos existentes entre dois organismos, mais próximos os dois estão. Com a matriz preenchida, ela pode ser utilizada para a geração de cladogramas, que são diagramas que mostram a ancestralidade dos organismos e que permitem verificar, visualmente, quão próximas as espécies estão evolutivamente. Um cladograma gerado por meio dos dados resultantes da ferramenta, com os alinhamentos de três diferentes espécies de *Mycobacterium* utilizados como entrada, pode ser observado na Figura 5.3. Pode-se notar que, segundo esta abordagem, *NC018150.1* e *CU458896.1* estão mais próximos evolutivamente do que *ENC01*.

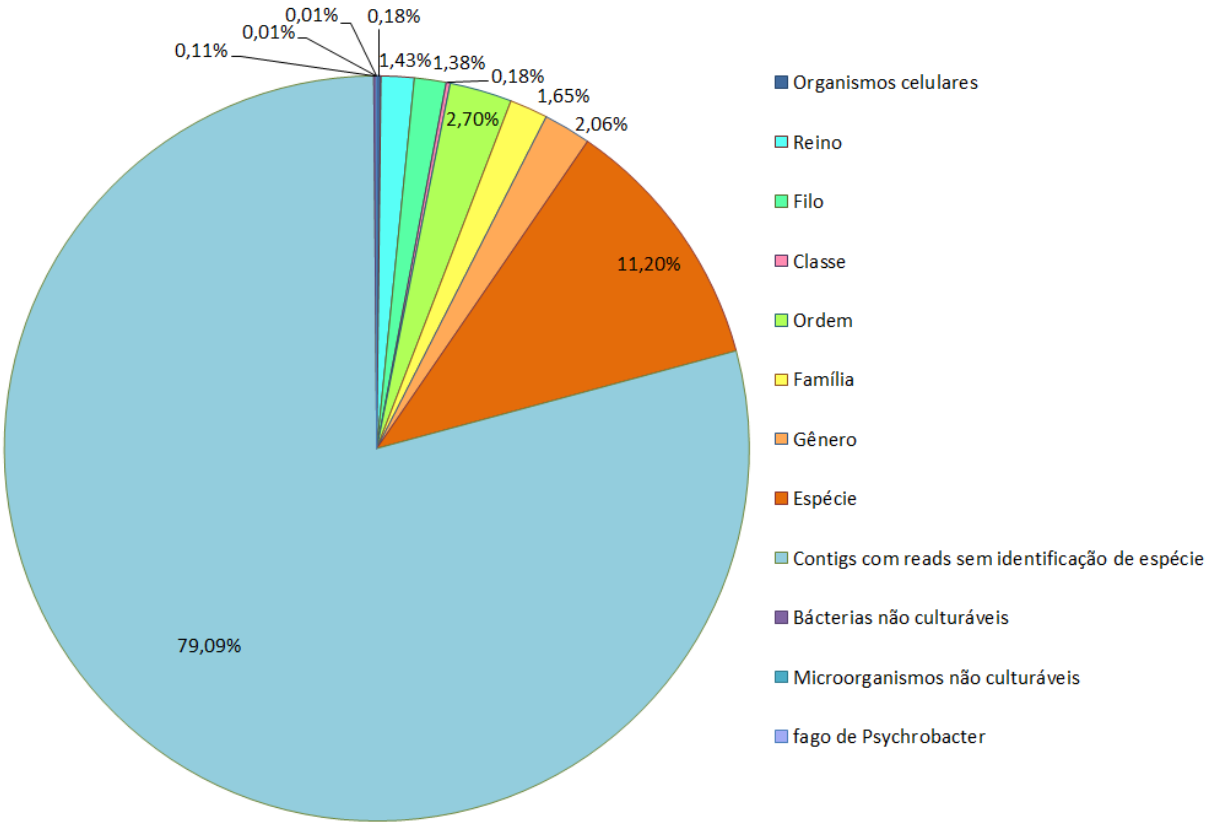

Figura 5.2 *– Gráfico com os níveis taxonômicos mais altos dos reads de cada contig*

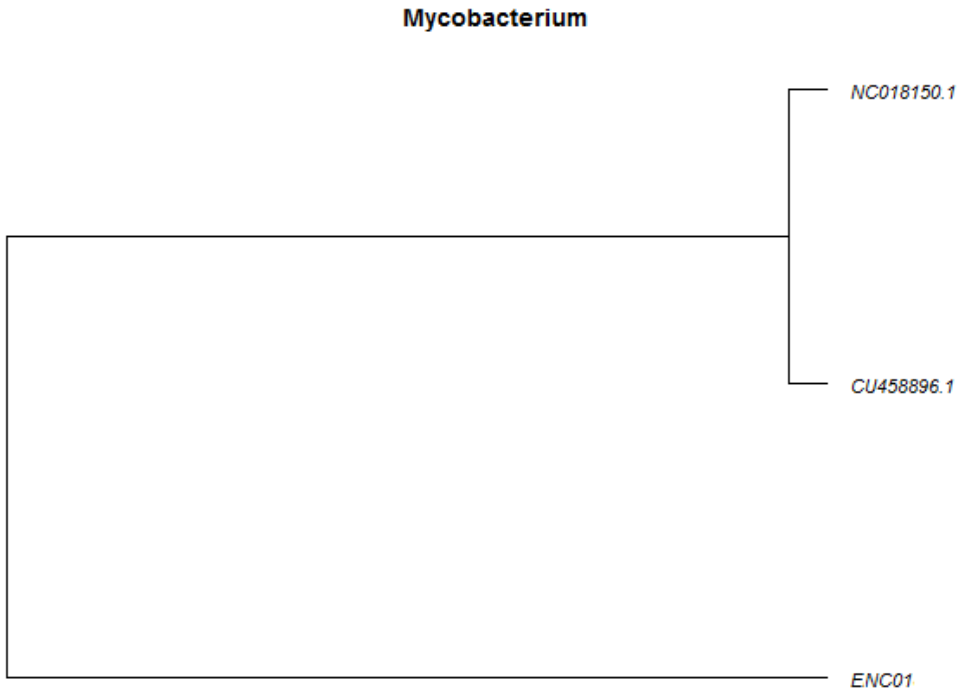

Figura 5.3 *– Cladograma horizontal gerado a partir dos dados do alinhamento do Mycobacterium*

6 Discussão

Alguns diferenciais das ferramentas desenvolvidas para identificação de espécie em relação às ferramentas similares que foram encontradas por meio da revisão bibliográfica são o uso da taxonomia completa, não somente algumas sequências de referência; a possibilidade de o usuário parametrizar a similaridade exigida para os diferentes parâmetros do alinhamento (*bit-score*, *e-value*, *aligned length*, *percent identities*, *number of mismatched positions* e *number of gap positions*), enquanto outras ferramentas, por exemplo o MEGAN, só permitem que o usuário parametrize o *bit-score* (a validação da ferramenta SOrt-ITEMS demonstrou que ela atingiu melhores resultados comparados com o MEGAN porque permitia que alguns dos atribuitos do alinhamento fossem parametrizados); e a possibilidade da classificação em níveis superiores, no caso de classificações redundantes: por exemplo, se um *read* satisfaz os critérios de similaridade para duas ou mais espécies diferentes então ele não será classificado como pertencente a nenhuma das espécies, mas sim ao nível taxonômico que essas espécies compartilham (gênero, família, ordem, etc).

Apesar de outras ferramentas também realizarem a classificação em níveis superiores, por não deixarem que o usuário escolha os critérios para o qual um alinhamento é considerado relevante, essas ferramentas podem permitir que alinhamentos insignificantes acabem sendo utilizados no momento da classificação de *reads* e *contigs* e, com isso, perde-se a especifidade da montagem de genomas e metagenomas.

Além disso, o fato da ferramenta de identificação de espécie desenvolvida neste trabalho permitir que seja escolhida apenas a primeira espécie ou todas cujo alinhamento foi realizado com um determinado *read* possibilita que seja feita uma comparação com os dois resultados (utilizando apenas a primeira espécie do *read* e utilizando todas as espécies em que sua sequência alinhou com o *read*), algo que não é possível com as outras ferramentas.

Por fim, as ferramentas encontradas por meio da revisão bibliográfica não realizam a identificação da existência de *contigs* consecutivos que poderiam formar sequências maiores se juntados, mas que o montador não conseguiu montar, nem a seleção de *reads* para "juntar"

esses *contigs* consecutivos antes da realização da análise filogenética. O fato dessas ferramentas não realizarem essa etapa, pode fazer com que similaridade entre os micro-organismos seja erroneamente calculada.

7 Conclusão

Neste trabalho, foram desenvolvidas ferramentas que automatizam parte significativa do processo de análise de genomas a partir de metagenomas e a análise filogenética desses genomas.

Com o que se pode observar nos testes realizados com as ferramentas de identificação de espécies, pode-se notar que elas possuem grande utilidade para a análise da montagem de genomas, pois possibilitam a averiguação da proximidade, em termos de taxonomia, dos *reads* montados em um mesmo *contig*, o que permite uma primeira análise sobre a identificação de potenciais novas espécies e da similaridade entre essas possíveis novas espécies e espécies cujos genomas já foram sequenciados.

Essa primeira análise é facilitada pelo uso das ferramentas que fazem a análise taxonômica, pois a partir dos dados sobre a distribuição dos micro-organismos em um determinado nicho, é possível gerar gráficos informativos sobre essa distribuição que, além de permitirem a visualização dessa primeira análise, também podem ser utilizados para verificar as mudanças ocorridas na metagenômica de um nicho específico ao longo do tempo. A ferramenta para geração da "matriz de proximidade" também demonstrou ser de grande utilidade, pois os cladogramas gerados a partir dessa matriz possibilitam que seja verificado a proximidade que há entre os diversos micro-organismos cujos genomas compõem o metagenoma. Essa proximidade ou não também pode ser utilizada para analisar se há uma potencial nova espécie presente no metagenoma analisado.

Além disso, as ferramentas desenvolvidas que realizam a identificação de *contigs* consecutivos que não foram montados inicialmente e a seleção de *reads* que possam juntar esses *contigs* são muito úteis no processo de montagem de genomas, uma vez que permitem a geração de sequências maiores e com menos "buracos" e, com isso, a identificação da espécie dessa sequência maior terá maiores chances de ser realizada corretamente do que se fosse realizada com pequenas sequências e que apresentam mais "buracos".

Referências Bibliográficas

ARUMUGAM, M. et al. Smashcommunity: a metagenomic annotation and analysis tool. **Bioinformatics**; v. 26; n. 23; p. 2977–2978; 2010. 3

DAVENPORT, C. F. et al. Genometa - a fast and accurate classifier for short metagenomic shotgun reads. **PLOS ONE**; v. 7; n. 8; p. e41224; 2012. 3

DEVULDER, G.; MONTCLOS, M. P. D.; FLANDROIS, J. P. A multigene approach to phylogenetic analysis using the genus mycobacterium as a model. **International journal of systematic and evolutionary microbiology**; v. 55; n. 1; p. 293–302; 2005. 1

GERLACH, W.; STOYE, J. Taxonomic classification of metagenomic shotgun sequences with carma3. **Nucleid Acids Research**; v. 39; n. 14; 2011. 3

HANDELSMAN, J. Metagenomics: Aapplication of genomics to uncultured microorganisms. **Microbiology and molecular biology reviews**; v. 68; n. 4; p. 669–685; 2004. 1

HAQUE, M. M. et al. Sort-items: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. **Bioinformatics**; v. 25; n. 14; p. 1722–1730; 2009. 3, 5

HEY, A. J. G.; TANSLEY, S.; TOLLE, K. M. The fourth paradigm: data-intensive scientific discovery. Microsoft Research Redmond, WA; 2009. 1

HUSON, D. H. et al. Megan analysis of metagenomic data. **Genome Research**; v. 17; n. 3; p. 377–386; 2007. 3

LIU, B.; GIBBONS, T.; POP, M. G. M. Metaphyler: Taxonomic profiling for metagenomic sequences. In: **Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on**. [S.l.: s.n.]; 2010. p. 95–100. 3

LUSCOMBE, N. M.; GREENBAUM, D.; GERSTEIN, M. What is bioinformatics? a proposed definition and overview of the field. **Methods of information in medicine**; v. 40; n. 4; p. 346–358; 2001. 1

MARTINS, L. F. et al. Metagenomic analysis of a tropical composting operation at the são paulo zoo park reveals diversity of biomass degradation functions and organisms. **Plos One**; v. 8; p. 1–13; 2013. 1, 4

MEYER, F. et al. The metagenomics rast server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. **BMC Bioinformatics**; v. 9; n. 1; p. 386; 2008. 3

SETUBAL, J. C.; MEIDANIS, J. **Introduction to Computational Molecular Biology**. Boston, EUA: PWS Publishing Company; 1997. 296 p. 1

SHARON, I.; BANFIELD, J. F. Genomes from metagenomics. **Science**; v. 342; p. 1057–1058; 2013. 1

SUNAGAWA, S. et al. Metagenomic species profiling using universal phylogenetic marker genes. **Nature methods**; v. 10; n. 12; p. 1196–1199; 2013. 1