Um sistema para análise de genomas a partir de metagenomas

Vivian M. Y. Pereira[1] **, Geraldo J. dos Santos Júnior**[1] **, Luciano A. Digiampietri**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
Av. Arlindo Béttio, Ermelino Matarazzo – 03828-000 – São Paulo – SP – Brasil

{vivian.pereira,geraldo.jose.santos,digiampietri}@usp.br

Abstract. *This paper presents a system developed to analyze genomes data which is part of metagenomes sequencing. As a result, some tools were developed to identify the species of metagenomics sequences, for helping the assembly process and to analyze the taxonomy of the assembled sequences in order to identify the similarity between microorganisms' genomes in metagenome.*

Resumo. *Este artigo apresenta um sistema desenvolvido com o objetivo de analisar dados dos genomas que fazem parte de sequenciamentos de metagenomas. Como resultado, foram desenvolvidas ferramentas para realizar a identificação das espécies presentes em sequências metagenômicas, auxiliar o processo de montagem e realizar a análise taxonômica das sequências montadas, identificando a similaridade entre os genomas dos micro-organismos presentes no metagenoma.*

1. Introdução

A bioinformática é uma área na qual são aplicadas 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 [Luscombe et al. 2001]. Um dos campos mais conhecidos da bioinformática é o relacionado à microbiologia computacional ou, em particular, à montagem e anotação de genomas.

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 para concluí-la) [Setubal e 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 sua vez, trouxe novos desafios como o fato de que, em muitas vezes, 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 proveniente de um dado habitat específico. Estes projetos são tipicamente chamados de projetos de metagenômica, em que, pelo fato do DNA ser originado de diversas populações, a recuperação dos genomas acaba se tornando uma tarefa complexa [Sharon e Banfield 2013].

Nesses projetos, 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 (*contigs*) a fim de tentar reconstruir a sequência de DNA do genoma completo daquela espécie (processo conhecido como montagem do genoma).

Dois dos principais desafios desta 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 muitos dos genomas possuirão 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, tais como a análise da quantidade e diversidade dos organismos encontrados e a verificação se alguma das espécies sequenciadas provavelmente corresponde a uma nova espécie (nunca sequenciada previamente), algo que pode ser enfrentado realizando-se análises filogenéticas.

O trabalho apresentado neste artigo teve como objetivo 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. Para tanto, foram desenvolvidas técnicas para a identificação automática da provável espécie a que pertence cada *read*, os *reads* de cada espécie foram agrupados e montados e foram desenvolvidas novas ferramentas para automatizar a análise filogenética dos genomas montados tentando utilizar toda a informação disponível do genoma e não apenas genes específicos.

Este trabalho está contextualizado dentro do Núcleo de Pesquisa em Ciência Genômica (NAP-CG) da Universidade de São Paulo[1]. Ele é organizado em projetos motores, que são projetos com base genômica e necessidades sofisticadas de bioinformática. O sistema apresentado neste artigo auxilia na realização de atividades do projeto motor Metagenômica de microbiomas do Zoológico de São Paulo[2] [Martins et al. 2013, Digiampietri et al. 2014].

Este artigo está estruturado de modo que a seção 2 apresenta os trabalhos correlatos; a seção 3, o sistema desenvolvido; e, por fim, na seção 4 estão as conclusões.

2. Trabalhos correlatos

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.

Alguns algoritmos existentes, tais como, SOrt-ITEMS [Haque et al. 2009], MEGAN [Huson et al. 2007], CARMA3 [Gerlach e Stoye 2011] e Genometa [Davenport et al. 2012] apenas realizam o agrupamento de *reads* ou *contigs* para atribuí-los a um mesmo nível taxonômico. Outras ferramentas como a SmashCommunity [Arumugam et al. 2010] e a MetaPhyler [Liu et al. 2010] utilizam genes específicos como marcadores filogenéticos para referência taxonômica. Por fim, o sistema MG-RAST [Meyer et al. 2008] compara dados fornecidos pelo usuário com outros metagenomas ou genomas completos. A ferramenta permite que o usuário escolha entre utilizar o gene 16S rRNA ou os resultados de alinhamentos para realizar a comparação do metagenoma.

[1]http://www.iq.usp.br/napcg/

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

Alguns diferenciais das ferramentas desenvolvidas e apresentadas neste artigo para identificação de espécie em relação às ferramentas similares são o uso de todas as sequências para a identificação taxonômica e 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 a MEGAN, só permitem que o usuário parametrize o *bit-score* (a validação da ferramenta SOrt-ITEMS demonstrou que ela atingiu melhores resultados comparados com a MEGAN porque permitia que três atributos fossem parametrizados para filtrar alinhamentos considerados insignificantes); e a possibilidade da classificação em níveis superiores, no caso de classificações incertas/ambíguas: por exemplo, se um *read* satisfaz os critérios de similaridade para duas ou mais espécies diferentes então ele não será classificado como pertencente às espécies, mas ao nível taxonômico que essas espécies compartilham (gênero, família, 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 permitir que seja escolhida apenas a primeira espécie (identificada pelo melhor alinhamento) ou todas as espécies (cujos alinhamentos com um determinado *read* passaram pelos filtros determinados pelo usuário) 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, elas 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 a similaridade entre os micro-organismos seja erroneamente estimada.

3. Sistema

O desenvolvimento das ferramentas que compõem o sistema, no qual utilizou-se linguagem de programação Java, foi realizado em duas partes. Primeiramente, foi especificada, implementada e testada uma ferramenta para a identificação da espécie a que pertencia cada sequência e 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*)[3]. As ferramentas de alinhamento local utilizadas foram o USEARCH[4]

[3]http://www.ncbi.nlm.nih.gov

[4]http://www.drive5.com/usearch

e BLAST[5]. Já a montagem das sequências foi realizada pela ferramenta Newbler[6]. 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 e que são utilizados como entrada para as ferramentas de identificação de espécie. Os campos são: *query name* (nome da sequência de consulta); *subject name* (nome da sequência cujo alinhamento foi encontrado); *percent identities* (porcentagem de bases idênticas); *aligned length* (tamanho do alinhamento); *number of mismatched positions* (número de alinhamentos entre bases diferentes); *number of gap positions* (número de espaços em branco (buracos) no alinhamento); *query sequence start* (posição inicial na sequência de consulta, ou seja, onde o alinhamento começou); *query sequence end* (posição final na sequência de consulta, ou seja, onde o alinhamento terminou); *subject sequence start* (posição inicial na sequência encontrada); o *subject sequence end* (posição final na sequência encontrada); *e-value* (probabilidade do alinhamento ser uma coincidência); e o *bit-score* (nota dada ao alinhamento).

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.

Uma visão geral do funcionamento do sistema pode ser visualizada na Figura 1. Nela, pode-se observar a interação das ferramentas desenvolvidas, destacadas em amarelo, com ferramentas de terceiros e com o conjunto de dados utilizado.

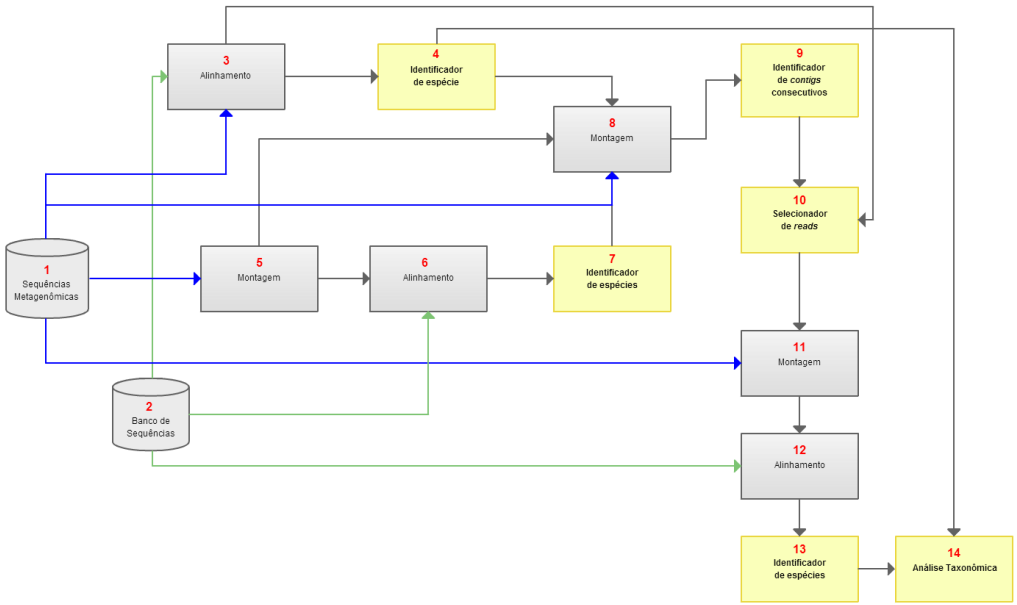

Figura 1. Visão geral com todas as ferramentas desenvolvidas (em amarelo) e utilizadas

[5]http://blast.be-md.ncbi.nlm.nih.gov/Blast.cgi
[6]http://454.com/products/analysis-software/index.asp

De acordo com a Figura 1, em **1** estão as sequências metagenômicas. Elas são geradas a partir do sequenciamento de material genético por um sequenciador de alto desempenho. Os *reads* gerados pelo sequenciamento, ou seja, as sequências metagenômicas, são comparados com as sequências de nucleotídeos não-redundantes do banco do NCBI (representado em **2**) por meio das ferramentas de alinhamento local USEARCH e BLAST, em **3**. Os alinhamentos gerados são então utilizados como entrada por uma das ferramentas desenvolvidas, 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, ou seja, só realiza a identificação da espécie se todos os filtros são satisfeitos. 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.

Além disso, é 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*, realiza-se o alinhamento com as sequências do banco do NCBI (em **6**). Em seguida, na etapa **7**, uma ferramenta desenvolvida analisa a montagem e os alinhamentos realizados identificando-se quantos *reads* pré-classificados de cada espécie um *contig* possui e quais *reads* ficaram sem identificação de espécie.

Outro conjunto de atividades de montagem é realizado. Neste processo, os *reads* são inicialmente agrupados por espécies (de acordo com a classificação já realizada) e para cada espécie é realizada uma montagem (atividade **8** da Figura 1).

As sequências montadas servem de entrada para a ferramenta, enumerada como **9**, que identifica os *contigs* consecutivos, 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, utiliza-se 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* é realizada por meio da observação dos campos *query sequence start* e *query sequence end*, presentes no arquivo *m8* fornecido pelas ferramentas de alinhamento local. Com os dados gerados por essa ferramenta, é realizada a montagem das sequências em **11**, que resulta em sequências maiores, com menos lacunas do que as montagens realizadas anteriormente.

Em seguida, é feito 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 espécie a sequência pertence ou mesmo se a sequência possivelmente pertence a uma espécie nova. Posteriormente, a ferramenta de identificação de espécies é utilizada para verificar a taxonomia desses alinhamentos gerados.

Em **13**, realiza-se a identificação das espécies desses alinhamentos, em 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) e é utilizada 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 (por exemplo, no reino *Bacteria*) ao *read*, o que acaba reduzindo a especificidade da montagem dos *reads* e *contigs* [Haque et al. 2009]. Por isso, essa ferramenta também possui os mesmos filtros das ferramentas anteriores e que são parametrizados pelo usuário para que apenas os alinhamentos considerados relevantes sejam 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, é feita a análise taxonômica, em **14**, por meio de ferramentas desenvolvidas. Uma delas calcula a distribuição dos micro-organismos na amostra de entrada e retorna, em ordem decrescente, as espécies mais frequentes existentes na amostra. Essa frequência pode ser calculada tanto pelo número de *reads* ou *contigs* pertencentes a uma determinada espécie quanto pela soma do tamanho dos alinhamentos de sequências que foram identificados como sendo de uma determinada espécie. Os dados gerados por esta ferramenta permitem a geração de gráficos para analisar visualmente a distribuição de micro-organismos e averiguar quão similares as distribuições de espécies de determinados habitat são.

Outra ferramenta 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 utilizada para preenchimento dessa matriz foi a proporção entre o tamanho do alinhamento entre dois organismos e o tamanho do genoma sequenciado. Esta matriz pode ser utilizada para a geração de cladogramas, que são diagramas que mostram a ancestralidade dos organismos e que permitem (de maneira aproximada) verificar visualmente quão próximas as espécies estão evolutivamente.

As ferramentas 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 eram sequências metagenômicas da compostagem realizada no Zoológico de São Paulo, que foram sequenciadas pelo sequenciador de alto desempenho Roche 454 GS FLX Titanium [Martins et al. 2013].

Uma análise taxonômica realizada com a ferramenta que calcula a distribuição dos micro-organismos sobre os dados provenientes da compostagem pode ser observada na Figura 2. 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 pertencem a espécies novas (ainda não sequenciadas) e é provável que essas possíveis novas espécies possuam uma filogenia com características similares a de outros micro-organismos já sequenciados, mas pertencentes a famílias diferentes.

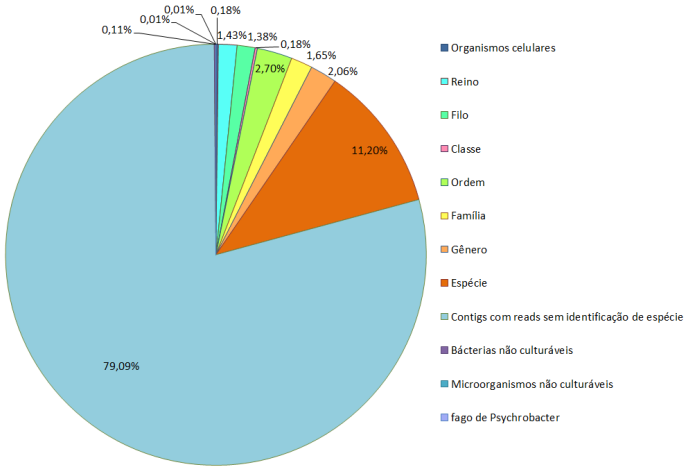

Figura 2. Gráfico com os níveis taxonômicos mais altos dos reads de cada contig

4. Conclusões

Neste artigo, apresentou-se um sistema desenvolvido para automatizar parte significativa do processo de análise de genomas a partir de metagenomas e sua análise filogenética. As ferramentas de identificação de espécies que compõem o sistema possuem grande utilidade para a análise da montagem de genomas, permitindo a averiguação da proximidade taxonômica dos *reads* montados em um mesmo *contig*, auxiliando numa primeira análise sobre a identificação de potenciais novas espécies e de suas similaridades com 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 habitat, pode-se gerar gráficos para visualização dessa distribuição e que também podem ser utilizados para verificar as mudanças ocorridas na metagenômica de um habitat específico ao longo do tempo.

A ferramenta para geração da "matriz de proximidade" possibilita estimar a proximidade entre os micro-organismos cujos genomas compõem o metagenoma e, com isso, permite a análise se há uma potencial nova espécie presente. 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". Com isso, a identificação da espécie utilizando esta sequência maior terá maiores chances de ser realizada corretamente do que se fosse realizada com pequenas sequências e que apresentam mais "buracos".

Agradecimentos

Ao Programa de Educação Tutorial do Ministério da Educação (MEC/SESu) e à CAPES.

Referências

Arumugam, M., Harrington, E. D., Foerstner, K. U., Raes, J., e Bork, P. (2010). Smashcommunity: a metagenomic annotation and analysis tool. *Bioinformatics*, 26(23):2977–2978.

Davenport, C. F., Neugebauer, J., Beckmann, N., Friedrich, B., Kameri, B., Kokott, S., Paetow, M., Siekmann, B., Wieding-Drewes, M., Wienhöfer, M., Wolf, S., Tümmler, B., Ahlers, V., e Sprengel, F. (2012). Genometa - a fast and accurate classifier for short metagenomic shotgun reads. *PLOS ONE*, 7(8):e41224.

Digiampietri, L. A., Pereira, V. M., Costa, C. I., dos Santos-Junior, G. J., Stefanini, F. M., e Santiago, C. R. (2014). An extensible framework for genomic and metagenomic analysis. In *Advances in Bioinformatics and Computational Biology*, volume 8826 of *Lecture Notes in Computer Science*, pages 1–8. Springer.

Gerlach, W. e Stoye, J. (2011). Taxonomic classification of metagenomic shotgun sequences with carma3. *Nucleid Acids Research*, 39(14).

Haque, M. M., Ghosh, T. S., Komanduri, D., e Mande, S. S. (2009). Sort-items: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, 25(14):1722–1730.

Huson, D. H., Auch, A. F., Qi, J., e C., S. S. (2007). Megan analysis of metagenomic data. *Genome Research*, 17(3):377–386.

Liu, B., Gibbons, T., e Pop, M. G. M. (2010). Metaphyler: Taxonomic profiling for metagenomic sequences. In *International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 95–100.

Luscombe, N. M., Greenbaum, D., e Gerstein, M. (2001). What is bioinformatics? a proposed definition and overview of the field. *Methods of information in medicine*, 40(4):346–358.

Martins, L. F., Antunes, L. P., Pascon, R. C., de Oliveira, J. C. F., Digiampietri, L. A., Barbosa, D., Peixoto, B. M., Vallim, M. A., Viana-Niero, C., Ostroski, E. H., Telles, G. P., Dias, Z., da Cruz, J. B., Juliano, L., Verjovski-Almeida, S., da Silva, A. M., e Setubal, J. C. (2013). Metagenomic analysis of a tropical composting operation at the são paulo zoo park reveals diversity of biomass degradation functions and organisms. *Plos One*, 8:1–13.

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

Setubal, J. C. e Meidanis, J. (1997). *Introduction to Computational Molecular Biology*. PWS Publishing Company, Boston, EUA.

Sharon, I. e Banfield, J. F. (2013). Genomes from metagenomics. *Science*, 342:1057–1058.