

FERRAMENTA PARA ANÁLISE DE MONTAGEM DE GENOMAS A PARTIR DE DADOS DE METAGENOMAS

Vivian Mayumi Yamassaki Pereira, Camila Izidio Costa, Geraldo José dos Santos Júnior, Luciano Antonio Digiampietri

Universidade de São Paulo/Escola de Artes, Ciências e Humanidades

vivian.pereira@usp.br

Resumo

Nos últimos 25 anos houve uma imensa revolução no estudo dos micro-organismos, iniciando com equipamentos para a visualização destes organismos, passando pelo sequenciamento de bactérias nos anos 90 até os sequenciadores de alto desempenho utilizados atualmente. Essas novas tecnologias para obtenção de dados possibilitaram uma grande gama de descobertas, mas também trouxeram novos desafios para a organização e análise dos dados gerados. O objetivo deste trabalho é analisar dados dos genomas que fazem parte de sequenciamentos de metagenomas, de modo que foi desenvolvida uma ferramenta para analisar a montagem de genomas identificando o nível taxonômico mais alto em comum que os *reads* de um *contig* possuem, além de identificar os *contigs* cujos *reads* ficaram sem identificação de espécie.

Palavras-chave: montagem de genomas, metagenômica, taxonomia

Abstract

Over the past 25 years, there has been an immense revolution in the study of microorganisms, starting with the equipment for visualization of these organisms to the bacteria's sequencing in the 90s decade until the high performance sequencers used nowadays. These new technologies to obtain data allow a great number of discoveries, but also brought new challenges to organize and analyze the data. The objective of this paper is analyze genomes data that are part of metagenomes sequencing, therefore it was developed a computational tool to analyze the genome assembly by identifying the highest taxonomic level in common that the *reads* of *contigs* have and also identifying the *reads* of *contigs* without species identification.

Key words: genome assembly, metagenomic, taxonomy

Introdução

A grande evolução dos equipamentos para aquisição de dados tem revolucionado a ciência atual (HEY; TANSLEY; TOLLE, 2009). A bioinformática é uma área que está bastante atrelada a esta evolução. Um dos campos mais conhecidos da bioinformática é o relacionado à microbiologia computacional, ou em particular à montagem e anotação de genomas.

A partir da década de 1970 foram desenvolvidas e aprimoradas técnicas para o sequenciamento de DNA. 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 a obtenção de grande volume de DNA (dezenas de milhões de bases) em um único sequenciamento. Mas este volume de dados trouxe também novos desafios, um dos principais está relacionado ao fato de que muitas vezes, em um único sequenciamento, é sequenciado material genético de muitos 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, ou do solo ou da água de locais específicos). Estes projetos são tipicamente chamados de projetos de genômica ambiental ou metagenômica, nos quais 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 muitos 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, 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 (desafio este que pode ser enfrentado realizando-se análises filogenéticas).

Este projeto está contextualizado dentro do Núcleo de Pesquisa em Ciência Genômica (NAP-CG) da Universidade de São Paulo 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* (MARTINS et al, 2013).

Objetivos

O objetivo deste trabalho foi desenvolver uma ferramenta para, a partir da análise da montagem de sequências de metagenomas, identificar quantos *reads* de cada espécie um *contig* possui e quais *reads* ficaram sem identificação de espécie, bem como encontrar, nos casos em que um *contig* possua mais de um *read* com espécie identificada, o nível taxonômico mais alto em comum que os *reads* possuem.

Materiais e Métodos

Primeiramente, foi realizado um estudo de diversos trabalhos relacionados à metagenômica, montagem e anotação de genomas publicados nos últimos anos – focando particularmente naqueles que apresentavam metodologias para a montagem de genomas a partir de metagenomas – por meio de uma revisão bibliográfica.

Após a leitura dos artigos, foi desenvolvida a ferramenta para fazer a análise da montagem a partir dos dados de metagenomas. Antes, porém, foi necessário identificar a espécie a qual pertencia cada *read*, que se baseou nos resultados de ferramentas de alinhamento local. Essas ferramentas compararam cada um dos *reads* utilizados com o banco de sequências de nucleotídeos não-redundantes do NCBI (*National Center for Biotechnology Information*). Como entrada, foi utilizado o resultado de alinhamentos locais no formato conhecido como m8, que consiste em um arquivo texto separado por tabulações contendo 12 campos informativos sobre cada alinhamento. Nos testes, foram utilizados resultados de duas ferramentas: *blast*⁵ e *usearch*⁶. Além disso, foi utilizada a taxonomia provida também pelo NCBI. Com estes dados, desenvolveu-se a primeira parte da ferramenta que faz uso da taxonomia completa (e não apenas algumas sequências de referência), possibilita ao usuário parametrizar a similaridade exigida para os diferentes parâmetros do alinhamento (cobertura, score, evalue, tamanho do alinhamento, etc) e, além disso, permite resolver classificações redundantes em que, se um *read* satisfaz critérios de similaridade para duas ou mais espécies diferentes, ele é classificado pelo nível taxonômico mais alto que essas espécies têm em comum. Após essa identificação, os *reads* foram agrupados de acordo com a espécie a que provavelmente pertenciam e, após esse agrupamento, foi também realizada a montagem dos genomas, utilizando todos os *reads* (sem prévia separação dos *reads* por genoma).

A partir desses dados, foi desenvolvida a segunda parte da ferramenta, utilizando-se a linguagem de programação JAVA, para analisar este tipo de montagem na qual se identificou quantos *reads* de cada espécie um *contig* possuía e quais *reads* ficaram sem identificação de espécie. Nos casos em que um *contig* possuísse *reads* associados a mais de uma espécie, encontrou-se o nível taxonômico mais alto que os *reads* compartilhavam. Para tanto, foram consideradas apenas as sequências de *reads* que entraram na montagem (integralmente ou parcialmente). Em seguida, verificou-se se tais sequências constavam no arquivo m8. Se um *contig* não possuísse nenhum *read* com identificação de espécie (ou seja, que constasse no arquivo m8), então ele era classificado como “sem identificação”; caso contrário, verificou-se qual era o nível taxonômico mais alto compartilhado pelos *reads*. A ferramenta ainda permite que o usuário escolha entre utilizar apenas a primeira espécie que alinhou com cada *read* ou todas as espécies que alinharam com um dado *read* para fazer a comparação. A ferramenta foi então testada com dados reais.

Resultados

A ferramenta desenvolvida analisa os *reads* montados em cada *contig*, de modo que contabiliza as suas respectivas ocorrências em cada nível taxonômico. Em seguida, é encontrado o nível taxonômico mais alto que todos os *reads* montados no *contig* têm em comum. Um pedaço da saída fornecida pela ferramenta é apresentado na Figura 1. Nesse caso, o nível taxonômico mais alto em comum compartilhado pelos 8 *reads* montados de um dado *contig* é a ordem *Clostridiales* e, portanto, o *contig* é classificado com essa ordem.

```

Bacteria[8]
  Firmicutes[8]
    Clostridia[8]
      Clostridiales[8]
        Ruminococcaceae[3]
          Ruminococcus[2]
            Ruminococcus sp. 5_1_39BFAA[2]
          Faecalibacterium[1]
            Faecalibacterium prausnitzii[1]
              Faecalibacterium prausnitzii L2-6[1]
        Lachnospiraceae[2]
          Roseburia[2]
            Roseburia hominis[2]
              Roseburia hominis A2-183[2]
        Eubacteriaceae[3]
          Eubacterium[3]
            Eubacterium rectale[3]
              Eubacterium rectale M104/1[2]
              Eubacterium rectale DSM 17629[1]

```

Figura 1: Exemplo de saída da ferramenta

Em um dos testes realizados com a ferramenta, em que foi levada em conta apenas a primeira espécie de cada *read*, obteve-se o nível taxonômico mais alto em comum dos *reads* de cada *contig*, conforme se pode

observar na Figura 2. Os dados utilizados como entrada fazem parte de um projeto de metagenoma da compostagem realizada no Zoológico de São Paulo (MARTINS et al, 2013).

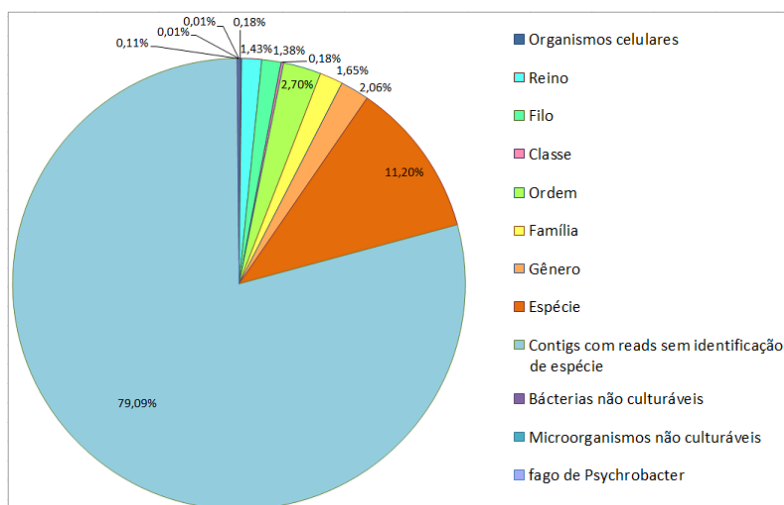


Figura 2: gráfico com os níveis taxonômicos mais altos dos *reads* de cada *contig*

Conforme se pode observar na Figura 2, grande parte dos *contigs* possui apenas *reads* sem identificação de espécie. Para os que possuem *reads* com identificação de mais de uma espécie, destacam-se os *contigs* cujos *reads* possuem a ordem como nível taxonômico mais alto em comum. Isto indica que muitos dos microrganismos provavelmente são de espécies novas (ainda não sequenciadas). Adicionalmente, para aqueles que possuem alinhamentos com *reads* com mais de uma espécie, o fato do nível taxonômico mais alto ser para a ordem indica que estas potenciais novas espécies correspondem a organismos cuja filogenia possui características similares a organismos já sequenciados de famílias diferentes.

Conclusões

De acordo com o observado nos testes da ferramenta, esta possui grande utilidade no que diz respeito à análise taxonômica da montagem de genomas a partir de metagenomas, uma vez que permite averiguar quão próximos, taxonomicamente, os *reads* que foram montados em um mesmo *contig* estão. Isto permite uma primeira análise sobre a identificação de potenciais novas espécies e da similaridade entre estas espécies e espécies cujos genomas já foram sequenciados. Como continuidade deste trabalho, será realizada a análise filogenética dos genomas (parcialmente ou completamente) montados.

Agradecimentos

Agradecemos ao Programa de Educação Tutorial (MEC/SESu) que financiou esse trabalho.

Referências Bibliográficas

HEY, T.; TANSLEY, S.; TOLLE, K. The Fourth Paradigm: Data-Intensive Scientific Discovery, p. 287, Microsoft, 2009.

MARTINS, L. F. ; ANTUNES, L. P. ; PASCON, R. ; OLIVEIRA, J. C.; DIGIAMPIETRI, L. A. ; BARBOSA, D. ; PEIXOTO, B. M. ; VALLIM, M. A. ; VIANA-NIERO, C. ; OSTROSKI, E. H. ; TELLES, G. ; DIAS, Z. ; CRUZ, J. B. ; JULIANO NETO, L. ; VERJOVSKI-ALMEIDA, S. ; SILVA, A. M. ; SETUBAL, J. C. 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.

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