CONTAGEM E RELAÇÃO ENTRE GENES E ESPÉCIES: FERRAMENTAs PARA O AUXÍLIO NA ANÁLISE DE DADOS DE METAGENOMAS

Geraldo José dos Santos Júnior
Camila Izidio Costa
Vivian Mayumi Yamassaki Pereira
Orientador: Luciano Antonio Digiampietri
Universidade de São Paulo, Escola de Artes, Ciências e Humanidades
EACH-USP

geraldo.jose.santos@usp.br

Resumo

A Bioinformática é uma área essencialmente interdisciplinar envolvendo as Ciências Biológicas, Ciência da Computação, Estatística, Química, Farmácia, Matemática, entre outras, no desenvolvimento de métodos para armazenamento e recuperação de dados biológicos e na construção de modelos e algoritmos para a solução de problemas biológicos. Metagenômica corresponde ao estudo de comunidades microbianas em nichos ecológicos específicos, por exemplo, o estômago de animais, o fundo do mar, as nuvens e o interior de cavernas. Neste projeto, estamos inicialmente analisando dados de metagenomas da compostagem que ocorre no Zoológico Municipal de São Paulo. A pesquisa desenvolvida neste projeto visa a caracterizar e quantificar os genes existentes no DNA coletado em um dado nicho ecológico e identificar de quais organismos estes genes são oriundos. Para isso, desenvolveu-se uma ferramenta que estabelece as possíveis relações entre genes e espécies em um determinado nicho.

Palavras Chaves: DNA, bioinformática, algoritmo

Abstract

Bioinformatics is an essentially interdisciplinary field involving the Biological Sciences, Computer Science, Statistics, Chemistry, Pharmacy, Mathematics, among others, the development of methods for storage and retrieval of biological data and the construction of models and algorithms for solving problems biological. Metagenomics corresponds to the study of microbial communities in specific ecological niches, for example, the stomach of animals, the deep sea, and inside a cave. In this project, we initially analyzed metagenome data from composting operating at the Municipal Zoo of Sao Paulo. The research undertaken in this project aims to characterize and quantify the existing genes in the DNA collected in a given ecological niche and identify from which organisms these genes are derived. Thereunto, we developed a tool to down possible relationship between genes and species in a given niche.

Keywords: DNA, bioinformatics, algorithm

Introdução

O contexto do estudo de genomas e metagenomas encontra-se na bioinformática, que é uma área interdisciplinar e busca o desenvolvimento de métodos para armazenamento e recuperação de dados biológicos e na construção de modelos e algoritmos para a solução de problemas biológicos (SETUBAL e MEIDANIS, 1997). Com o grande crescimento na produção de dados em diferentes áreas e em especial nas áreas de Biologia, Medicina e Química (HEY et al, 2009), a bioinformática tem se tornado cada mais fundamental.

Uma das áreas ligadas à bioinformática é a metagenômica que corresponde ao estudo de comunidades microbianas em nichos ecológicos específicos, por exemplo, o estômago de animais, o fundo do mar, as nuvens e o interior de cavernas. Recolhendo-se amostras de estudo, analisa-se os dados coletados e gera-se um arquivo contendo as informações obtidas através da análise. Neste projeto, estamos inicialmente analisando dados de metagenomas da compostagem que ocorre no Zoológico Municipal de São Paulo (MARTINS et al, 2013).

Busca-se desenvolver algoritmos que facilitem e automatizem parcialmente o processo de análise de dados de metagenomas de foram a se obter uma diminuição no tempo para a realização de algumas atividades, bem como a praticidade em reunir e apresentar dados. Essas ferramentas são criadas com base na necessidade e dificuldades encontradas durante atividades de pesquisas. Para este trabalho, desenvolveu-se, utilizando a linguagem de programação Java, três ferramentas para a organização e apresentação de dados sobre a relação entre genes e espécies.

Objetivos

A pesquisa desenvolvida neste projeto visou a auxiliar na análise de dados de metagenomas a partir da quantificação e caracterização dos genes existentes no DNA coletado em um dado nicho ecológico. Objetivou-se o desenvolvimento de ferramentas que possibilitem apresentar: (i) o número de genes diferentes para cada espécie, (ii) a co-ocorrência de genes entre espécies; e (iii) e a lista de genes encontrados para cada espécie.

Estas informações servem de base para diversas análises de dados de metagenomas e especialmente quando diferentes sequenciamentos são feitos em amostras de um mesmo nicho (por exemplo, um sequenciamento por mês ou por estação do ano) estas informações podem ser utilizadas para auxiliar na análise da variação da composição destas amostras.

Materiais e Métodos

A metodologia deste trabalho foi composta pelo estudo da literatura correlata em artigos e textos técnicos relacionados, especificação, desenvolvimento e teste das ferramentas.

As ferramentas desenvolvidas fazem uso de algumas bibliotecas básicas da linguagem Java, tais como vetores, mapas, iteradores, tabelas, além da leitura/escrita em arquivos. Os dados de entrada correspondem a resultados de alinhamentos locais nos quais um conjunto de dados básicos do sequenciamento (isto é, milhões de pequenas sequências de DNA) foram comparados com o banco de nucleotídeos do NCBI (*National Center for Biotechnology Information*) usando-se as ferramentas *blast* ou *usearch*. As ferramentas desenvolvidas, inicialmente verificam se cada um dos alinhamentos de entrada deve ser analisado (de acordo com uma série de filtros sobre a qualidade dos alinhamentos gerados) e, caso seja, executa-se uma série de comandos a fim de armazenar e organizar as informações encontradas (que indicam a possível existência de um gene na sequência de entrada).

Após salvar os dados encontrados, ou seja, após verificar todo o arquivo utilizado como entrada, que possui as informações sobre as espécies e seus respectivos genes, o algoritmo, então, exibe de maneira organizada os dados obtidos como conclusão do processamento realizado. Esta apresentação depende da ferramenta utilizada e pode ser: número de genes por espécie; número de ocorrências de cada gene em toda a amostra (incluindo co-ocorrências entre espécies); e lista de genes por espécie.

Resultados

As ferramentas desenvolvidas foram aplicadas a diferentes sequenciamentos relacionados ao metagenoma da compostagem realizada no Zoológico de São Paulo. Os resultados destas ferramentas estão sendo utilizados por especialistas do domínio para auxiliar na compreensão e esclarecimento de hipóteses sobre os micro-organismos presentes neste nicho.

A seguir são apresentados três pequenos exemplos do resultado do uso das ferramentas. Tem-se, na ordem em que são apresentadas as cópias de tela da execução das ferramentas: o nome da espécie e o número de genes diferentes da espécie, a quantidade de espécies diferentes em que os genes aparecem e, por fim, a lista de genes que apareceram na respectiva espécie.

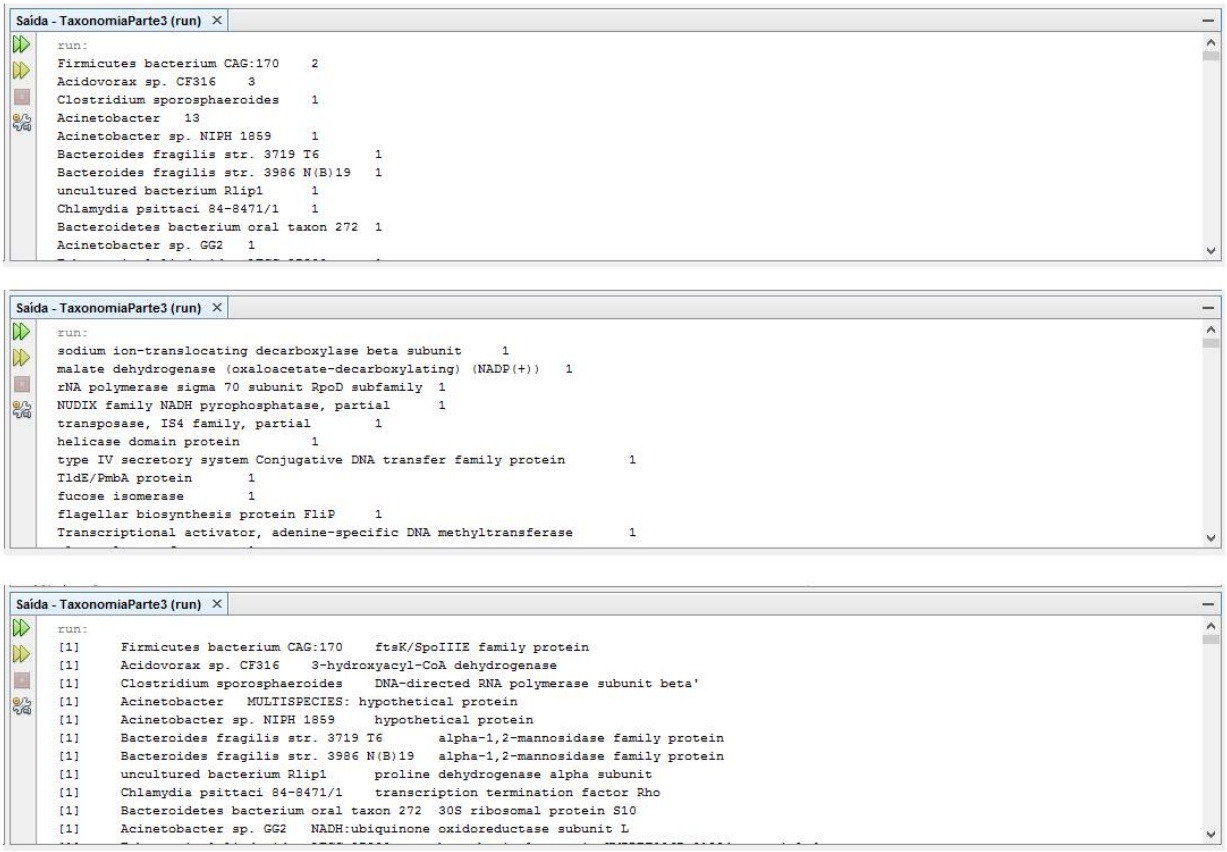

Figura 1: Saídas produzidas pelas ferramentas para um determinado conjunto de dados

Conclusões

Neste projeto foram desenvolvidas três ferramentas para auxiliar na análise de dados relacionados a metagenomas. Em especial, estas três ferramentas realizam, de maneiras distintas, análises sobre a possível presença de genes nas amostras de DNA coletadas.

As ferramentas foram testadas e validadas usando dados reais e seus resultados estão sendo utilizados para auxiliar a identificar as características de dados de metagenoma e a variação destes dados ao longo das diferentes amostras coletadas.

Como trabalhos futuros, pretende-se identificar relações entre os diferentes genes de forma a se estimar quais vias metabólicas estão ativas nas diferentes amostras coletadas.

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.