Proposta de Dissertação de Mestrado

Bioinformática aplicada a um projeto de metagenômica

Bruno Malveira Peixoto e Zanoni Dias

Instituto de Computação, UNICAMP

11 de Novembro de 2011

Agenda

O Que é Metagenômica

Genômica: Estudo do material genético dos organismos.

O Que é Metagenômica

Genômica: Estudo do material genético dos organismos.

Metagenômica: Estudo do material genético obtido diretamente de amostras ambientais.

Motivação

- Muitos organismos não são passíveis de cultivo em laboratório.

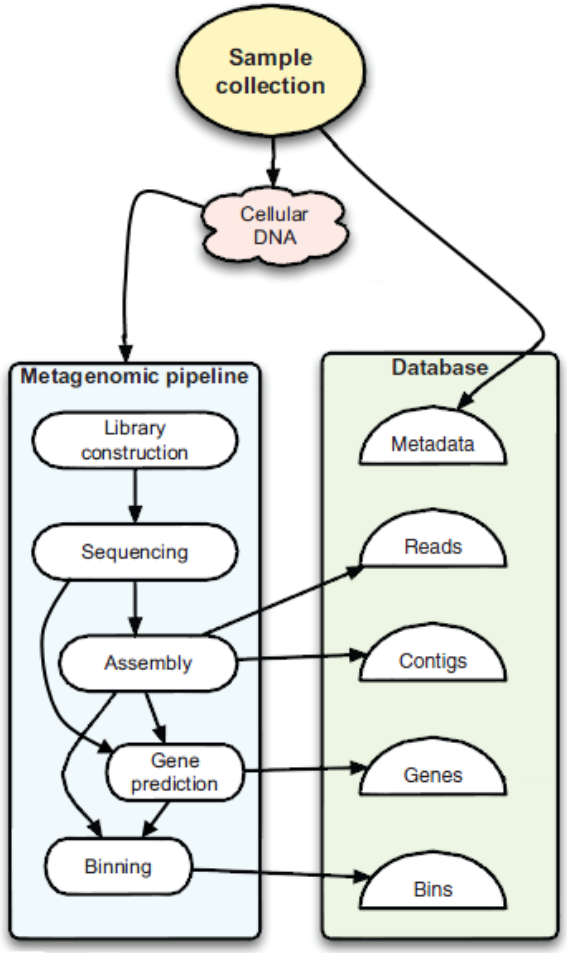
Motivação

- Muitos organismos não são passíveis de cultivo em laboratório.
- Descoberta de novos genes e espécies.

Motivação

- Muitos organismos não são passíveis de cultivo em laboratório.
- Descoberta de novos genes e espécies.
- Análise das interações entre organismos e ambiente.

Fluxo de Trabalho de um Projeto Metagenômico



Amostragem

- Amostras devem representar bem a população de onde elas foram retiradas.

Amostragem

- Amostras devem representar bem a população de onde elas foram retiradas.
- A presença ou ausência de uma população dominante afeta bastante o tipo de análise que pode ser feita.

Amostragem

- Amostras devem representar bem a população de onde elas foram retiradas.
- A presença ou ausência de uma população dominante afeta bastante o tipo de análise que pode ser feita.
- Coletar dados associados com uma amostra ambiental melhora bastante a habilidade de interpretar as sequências.

Amostragem

- Amostras devem representar bem a população de onde elas foram retiradas.
- A presença ou ausência de uma população dominante afeta bastante o tipo de análise que pode ser feita.
- Coletar dados associados com uma amostra ambiental melhora bastante a habilidade de interpretar as sequências.
- Foi proposto na comunidade um padrão de *'Minimum Information about a Metagenomic Sequence (MIMS)'*.

Montagem

- Combinar sequências lidas em *contigs* - trechos contínuos de DNA.
- Baseado em similaridades.

Desafios

- Amostragem incompleta - Genomas parcialmente amostrados.

Desafios

- Amostragem incompleta - Genomas parcialmente amostrados.
- Difícil mapear leituras individuais a suas espécies de origem.

Desafios

- Amostragem incompleta - Genomas parcialmente amostrados.
- Difícil mapear leituras individuais a suas espécies de origem.
- Possibilidade de se montar sequências de espécies diferentes.

Abordagens

- Habitats com pequeno número de espécies dominantes possibilita uma abordagem clássica.

Abordagens

- Habitats com pequeno número de espécies dominantes possibilita uma abordagem clássica.
- Montagem dos dados com diferentes ferramentas para minimizar erros.

Abordagens

- Habitats com pequeno número de espécies dominantes possibilita uma abordagem clássica.
- Montagem dos dados com diferentes ferramentas para minimizar erros.
- Utilizar genomas de referência.

Abordagens

- Habitats com pequeno número de espécies dominantes possibilita uma abordagem clássica.
- Montagem dos dados com diferentes ferramentas para minimizar erros.
- Utilizar genomas de referência.
- Representar leituras em grafos.

Predição de Genes

- Genes são a unidade básica funcional em um genoma.
- Predição de Genes é o procedimento de identificar proteínas e sequências de RNA codificantes nas amostras de DNA.

Predição de Genes

- Genes são a unidade básica funcional em um genoma.
- Predição de Genes é o procedimento de identificar proteínas e sequências de RNA codificantes nas amostras de DNA.
- Dependendo da aplicação e do sucesso da montagem, essa etapa pode ser feita em:
  - Contigs montados.
  - Metagenoma não-montado.
  - Mistura de contigs e leituras não-montadas.

Abordagens

- Duas abordagens diferentes são aplicadas para predição de genes:

Abordagens

- Duas abordagens diferentes são aplicadas para predição de genes:
- **Métodos Intrínsecos** analisam propriedades de sequências e se baseiam em aprendizado supervisionado e métodos estatísticos de reconhecimento de padrões.

Abordagens

- Duas abordagens diferentes são aplicadas para predição de genes:
- **Métodos Intrínsecos** analisam propriedades de sequências e se baseiam em aprendizado supervisionado e métodos estatísticos de reconhecimento de padrões.
- **Métodos Extrínsecos** usam ferramentas como o BLAST (Basic Local Alignment Search Tool) para identificar genes similares aos observados.

Desafios

- A natureza incompleta e fragmentada dos dados metagenômicos dificulta a identificação de genes.

Desafios

- A natureza incompleta e fragmentada dos dados metagenômicos dificulta a identificação de genes.
- A diversidade filogenética das amostras dificultam o uso de conjuntos de treinamento específicos de espécies.

Classificação

- Classificação é a associação de sequências ou genes à sua espécie de origem.

Classificação

- Classificação é a associação de sequências ou genes à sua espécie de origem.
- Altamente desejado para uma boa interpretação do ambiente, capaz de fornecer estatísticas de composição da comunidade.

Classificação

- Classificação é a associação de sequências ou genes à sua espécie de origem.
- Altamente desejado para uma boa interpretação do ambiente, capaz de fornecer estatísticas de composição da comunidade.
- Duas estratégias mais conhecidas:
  - Classificação baseada em similaridade.
  - Classificação baseada em composição.

Abordagens

- **Classificação por similaridade** consiste em fazer buscas em sequências de referência e recriar árvores filogenéticas a partir das similaridades com as sequências observadas.

Abordagens

- **Classificação por similaridade** consiste em fazer buscas em sequências de referência e recriar árvores filogenéticas a partir das similaridades com as sequências observadas.
- **Classificação por composição** utiliza o conhecimento de que processos celulares produzem assinaturas no DNA diferentes em cada genoma. Uma análise estatística pode identificar características que diferem uma população em particular das outras.

Limitações

- A base de dados do genoma de referência é atualmente incompleta e altamente tendenciosa para apenas três filos de bactérias (Proteobacteria, Firmicutes e Actinobacteria) dentre pelo menos 50 filos.

Limitações

- A base de dados do genoma de referência é atualmente incompleta e altamente tendenciosa para apenas três filos de bactérias (Proteobacteria, Firmicutes e Actinobacteria) dentre pelo menos 50 filos.
- Genes conservados filogeneticamente representam apenas uma pequena fração do total de dados metagenômicos.

Análise de Dados

- Atribuição de algumas atividades metabólicas para membros individuais do ecossistema.

Análise de Dados

- Atribuição de algumas atividades metabólicas para membros individuais do ecossistema.
- Se mais de uma população dominante é sequenciada, a potencial interação metabólica dessas populações também pode ficar aparente.

Análise de Dados

- Atribuição de algumas atividades metabólicas para membros individuais do ecossistema.
- Se mais de uma população dominante é sequenciada, a potencial interação metabólica dessas populações também pode ficar aparente.
- O servidor metagenômico MG-RAST é um sistema desenvolvido para processar automaticamente as sequências de metagenomas, fornecendo reconstruções filogenéticas e diversas estatísticas sobre a amostra.

Metagenômica Comparativa

- Comparações de amostras diferentes do mesmo ambiente ou similar, podem revelar a influência de fatores ambientais particulares em comunidades microbianas.

Metagenômica Comparativa

- Comparações de amostras diferentes do mesmo ambiente ou similar, podem revelar a influência de fatores ambientais particulares em comunidades microbianas.
- A comparação de metagenomas de diferentes habitats permite a descoberta de tendências gerais que ligam metagenomas e propriedades das comunidades com características fenotípicas dos ambientes.

Metagenômica Comparativa

- Comparações de amostras diferentes do mesmo ambiente ou similar, podem revelar a influência de fatores ambientais particulares em comunidades microbianas.
- A comparação de metagenomas de diferentes habitats permite a descoberta de tendências gerais que ligam metagenomas e propriedades das comunidades com características fenotípicas dos ambientes.
- No servidor MG-RAST são construídos mapas de taxonomias e subsistemas para encapsular diferenças entre as amostras e comparar suas similaridades.

Proposta

- Estudo metagenômico a partir de dados coletados na Fundação Parque Zoológico de São Paulo (FPZSP).
- Unidade de compostagem que aproveita matéria orgânica de várias origens, desde excrementos de animais a colunas de água do lago.
- Este material, atualmente utilizado como fertilizador de áreas agrícolas do próprio zoológico, tem uma grande riqueza microbiológica, podendo conter várias espécies ainda não descritas.
- Estima-se que a maior parte dos microrganismos que habitam este material não seja cultivável, assim a abordagem metagenômica é a mais indicada para esse estudo.

Proposta

- As ferramentas disponíveis para montagem, predição de genes e classificação também serão analisadas de forma que suas limitações sejam estressadas, e diversas combinações de parâmetros sejam testadas.
- A análise dos dados obtidos nos permitirá ter uma visão mais detalhada sobre a composição e o funcionamento da comunidade microbiana amostrada.
- Além disso, há expectativas de obtenção de informação de genomas ainda não explorados, constituindo dados genéticos que podem ser estudados por toda a comunidade científica.

Testes preliminares

- Para realizarmos testes preliminares com alguns programas de montagem, predição de genes e classificação, utilizamos a base de dados do projeto Lauber 88 Soils, disponível publicamente no servidor MG-RAST.
- O projeto possui disponíveis cerca de 150 mil sequências, totalizando mais de 34 milhões de pares de bases.

Março de 2011 a Fevereiro de 2013

	2011										2012												2013	
	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	●	●	●	●	●																			
2					●	●																		
3							●	●	●	●	●	●												
4												●	●	●	●	●	●							
5																	●	●	●	●	●	●		
6												●					●				●	●		
7																							●	
8																								●

1. Obtenção dos créditos obrigatórios.

Março de 2011 a Fevereiro de 2013

	2011										2012												2013	
	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	●	●	●	●	●																			
2					●	●																		
3							●	●	●	●	●	●												
4												●	●	●	●	●	●							
5																	●	●	●	●	●	●		
6												●					●				●	●		
7																							●	
8																								●

2. Escrita da proposta de qualificação de mestrado.

Março de 2011 a Fevereiro de 2013

	2011										2012											2013	
	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	●	●	●	●	●																			
2					●	●																		
3							●	●	●	●	●	●												
4												●	●	●	●	●	●							
5																	●	●	●	●	●	●		
6												●					●				●	●		
7																							●	
8																								●

3. Análise dos métodos de montagem e aplicação nas sequências do Zoológico.

Março de 2011 a Fevereiro de 2013

	2011										2012												2013	
	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	●	●	●	●	●																			
2					●	●																		
3							●	●	●	●	●	●												
4												●	●	●	●	●	●							
5																	●	●	●	●	●	●		
6												●					●				●	●		
7																							●	
8																								●

4. Análise dos métodos de predição de genes e aplicação nas sequências do Zoológico.

Março de 2011 a Fevereiro de 2013

	2011										2012												2013	
	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	●	●	●	●	●																			
2					●	●																		
3							●	●	●	●	●	●												
4												●	●	●	●	●	●							
5																	●	●	●	●	●	●		
6												●					●				●	●		
7																							●	
8																								●

5. Análise dos métodos de classificação e aplicação nas sequências do Zoológico.

Março de 2011 a Fevereiro de 2013

	2011										2012												2013	
	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	●	●	●	●	●																			
2					●	●																		
3							●	●	●	●	●	●												
4												●	●	●	●	●	●							
5																	●	●	●	●	●	●		
6												●					●				●	●		
7																							●	
8																								●

6. Escrita da dissertação.

Março de 2011 a Fevereiro de 2013

	2011										2012												2013	
	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	●	●	●	●	●																			
2					●	●																		
3							●	●	●	●	●	●												
4												●	●	●	●	●	●							
5																	●	●	●	●	●	●		
6												●					●				●	●		
7																							●	
8																								●

7. Revisão final do texto da dissertação.

Março de 2011 a Fevereiro de 2013

	2011										2012												2013	
	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	●	●	●	●	●																			
2					●	●																		
3							●	●	●	●	●	●												
4												●	●	●	●	●	●							
5																	●	●	●	●	●	●		
6												●					●				●	●		
7																							●	
8																								●

8. Defesa da dissertação.