Bioinformática aplicada a um projeto de metagenômica

Bruno Malveira Peixoto

25 de agosto de 2011

Resumo

Metagenômica é o estudo genético de uma amostra ambiental, e permite a análise de organismos não-cultiváveis em laboratório. É uma área de estudo nova e possui muitos desafios computacionais, com poucas ferramentas dedicadas disponíveis. O objetivo deste trabalho é realizar um estudo metagenômico de amostras da unidade de compostagem da Fundação Parque Zoológico de São Paulo, analisando os programas existentes. O estudo das comunidades microbianas que ali vivem é de grande importância para um melhor entendimento do papel biológico desses organismos.

1 Introdução

Este documento tem o objetivo de apresentar o plano de trabalho a ser desenvolvido durante o mestrado. A Seção 2 apresentará os conceitos básicos necessários. A Seção 3 fornece detalhes sobre o objeto de estudo deste trabalho. Finalmente, as Seções 4 e 5 tratam da proposta e do cronograma das atividades a serem executadas.

2 Conceitos Básicos

Nesta seção faremos uma breve descrição de alguns conceitos básicos que contextualizam o cenário em que o trabalho está inserido.

2.1 Genética

Um nucleotídeo é uma substância química composta por um açúcar chamado pentose, um grupo fosfato e uma base nitrogenada. Nucleotídeos são diferenciados pela base nitrogenada que os compõem, que pode ser: adenina, citosina, guanina, timina ou uracila. A pentose de um nucleotídeo pode se ligar ao grupo fosfato de um outro, formando uma cadeia. Uma cadeia formada por vários nucleotídeos é chamada de polinucleotídeo.

Existem dois tipos de polinucleotídeos que armazenam informações genéticas: o *DNA* (ácido desoxirribonucleíco) e o *RNA* (ácido ribonucleíco). Suas estruturas são representadas na figura 1.

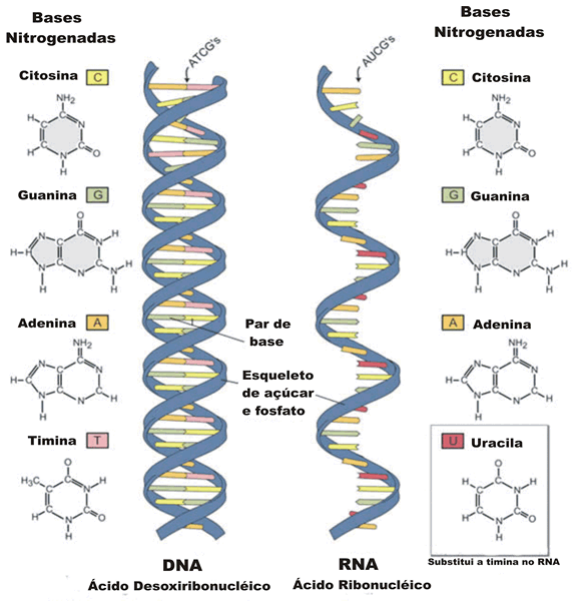

Figura 1: Estrutura das moléculas de DNA e RNA

O DNA é formado por duas fitas de nucleotídeos, e a pentose que o constitui é a desoxirribose. As duas fitas do DNA são unidas através de pontes de hidrogênio formadas entre as bases nitrogenadas. A adenina sempre forma pontes de hidrogênio com a timina, e a citosina com a guanina. As duas fitas de DNA são ditas complementares, e sempre é possível construir uma fita a partir da outra. A sequência de bases nitrogenadas ao longo da cadeia de DNA constitui a informação genética.

O RNA é composto por apenas uma fita e sua pentose é a ribose. O RNA possui uma base nitrogenada chamada uracila, que substitui a timina do DNA. Uma fita de RNA pode se dobrar de tal forma que parte de suas próprias bases nitrogenadas se pareiam umas com as outra. Esse pareamento intramolecular é um fator importante no formato tridimensional do RNA, que é capaz de assumir uma variedade maior de formas complexas do que a dupla hélice de DNA.

Genética é um ramo da biologia que estuda os genes. Genes são as unidades orgânicas básicas que contém informações sobre as características físicas e comportamentais de um organismo e são passadas de geração para geração. Essas informações são usadas pra criar proteínas que constroem e regulam o funcionamento das células e do organismo com um todo. Os genes são regiões de moléculas de DNA que se encontram no interior de cada célula.

2.2 Expressão Gênica

Para produzir proteínas a partir do código genético, acontece uma série de transformações bioquímicas conhecida como expressão gênica. Uma sequência de bases do DNA é lida e segmentos do código são traduzidos em aminoácidos que juntos formarão uma proteína.

O processo de síntese de RNA a partir de um gene codificado no DNA é chamado de *transcrição*. Uma enzima chamada RNA polimerase se liga ao genoma em um marcador conhecido como promotor e começa a processar o DNA, formando um complexo de transcrição. A partir deste complexo, uma fita de RNA é formada com base na sequência de nucleotídeos do DNA. Esse RNA é conhecido como mRNA ou RNA mensageiro. Existem dois outros tipos de RNA: rRNA (RNA ribossomal) e tRNA (RNA transportador), que também são gerados a partir do DNA, e são utilizados na fase de tradução, descrita mais adiante.

Em organismo procariotos, como as bactérias e arqueobactérias, o mRNA não sofre mais alterações. Nos eucariotos o mRNA é processado para a adição de uma sequência de bases adenina, chamada de poli-A, que dá maior estabilidade para o RNA. Também é feito o *splicing*, um processo de remoção de íntrons, que são regiões não-codificadoras do DNA, deixando apenas as regiões codificadoras, chamadas de exons. Nesse processo, nem todos os exons são preservados, e uma mesma sequência de DNA é capaz de gerar diferentes mRNAs, aumentando a diversidade genética desses organismos.

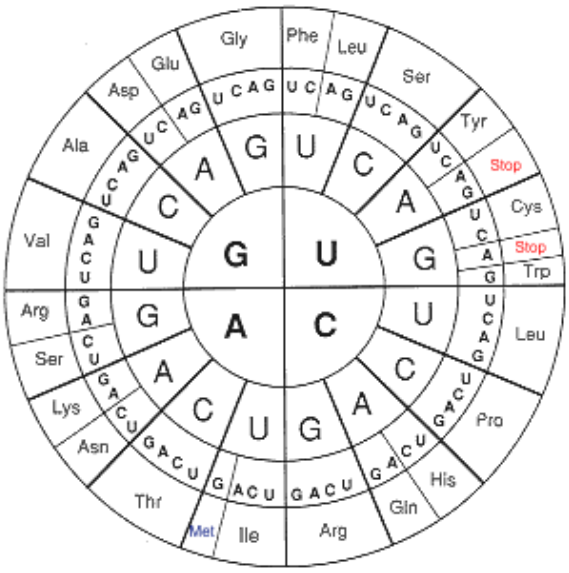

Figura 2: Código genético utilizado pela maioria dos organismos. A ordem das bases de um codon é lida do centro para as extremidades. O aminoácido Metionina, em azul, é também o start-codon.

O tradução é o processo de síntese de proteínas a partir de um mRNA, e se baseia em conjuntos de triplas de nucleotídeos chamados *codons*. Existem $4^3 = 64$ codons diferentes, e cada um identifica um aminoácido, no entanto, um aminoácido pode ser identificado por mais de um codon, uma vez que existem 20 aminoácidos distintos. Um codon pode também ter um significado especial, marcando o início ou final de uma região de tradução (*start-codon* e *stop-codon*, respectivamente).

Existem tabelas que descrevem a relação entre um codon e o seu aminoácido correspondente. Na página do NCBI [13] podemos encontrar 17 tabelas. A Figura 2 mostra uma representação da tabela número 1 da página do NCBI, que é o padrão para a maioria dos organismos.

A tradução começa com a ligação do mRNA com o *ribossomo*, que é uma organela no interior da célula responsável por percorrer o mRNA a partir do start-codon. Para cada codon, um tRNA é ativado, esse RNA transportador é responsável por ligar um codon específico a um aminoácido, juntando-o à cadeia de aminoácido já existente e formando assim a proteína codificada.

2.3 Sequenciamento Genômico

O sequenciamento de um genoma é a determinação da cadeia de nucleotídeos que o compõe. Este processo envolve várias etapa, que serão brevemente descritas a seguir.

2.3.1 Preparação do Material Genético

Nos métodos atuais de sequenciamento existe uma limitação física que impede que sejam sequenciadas cadeias de nucleotídeos maiores que 1000 bases. Devido a essa limitação, a cadeia de nucleotídeos a ser estuda é primeiro fragmentada e depois re-montada para obter a sequência original. Existem dois métodos principais para fragmentar o DNA: shotgun e digestão.

O método shotgun consiste em submeter o DNA a altas taxas de vibrações, fazendo com que a cadeia de nucleotídeos se quebre em vários pontos aleatórios. O método da digestão utiliza enzimas especiais, chamadas enzimas de restrição, que cortam o DNA em regiões específicas, conhecidas com sítios de restrição.

Os fragmentos obtidos são replicados através de um processo chamado de amplificação. A amplificação pode ser feita tanto por PCR (Polymerase Chain Reaction), que utiliza enzimas que sintetizam novas fitas de DNA a partir dos fragmentos existentes, quanto por clonagem, que insere os fragmentos em bactérias para serem replicados por elas.

2.3.2 Métodos de Sequenciamento

Atualmente existem vários métodos de sequenciamento. O método mais clássico que é eficiente e sujeito a automatização é o da terminação de cadeia.

Ele é capaz de gerar centenas de sequências de tamanhos entre 800 e 1000 pares de bases. Recentemente, a grande demanda por sequenciamentos mais baratos acelerou o desenvolvimento de tecnologias de alto rendimento, que paralelizam o processo de sequenciamento e geram milhões de sequências de uma vez. No entanto, essas tecnologias de sequenciamento de segunda geração produzem sequências muito menores, de 50 a 500 pares de bases. A seguir os métodos mais conhecidos são brevemente descritos:

Método da terminação de cadeia [31]: também conhecido como sequenciamento Sanger, gera várias cópias de cadeias de nucleotídeos que diferem de tamanho por apenas uma base, que podem ser separadas e ordenadas.

Primeiro, os fragmentos de DNA são utilizados como moldes para a síntese de uma fita de DNA complementar. Para isso, são utlizados primers, que são oligonucleotídeos que se acoplam a regiões específicas do DNA e servem de iniciadores da reação de síntese do DNA.

O processo de síntese do DNA complementar ocorre em uma solução que contem: uma enzima chamada DNA polimerase; os quatro tipos de desoxiribonucleotídeos (contendo as quatro bases do DNA); e alguns dideoxinucleotídeos equivalentes, que servem como terminadores de cadeia, impedindo que a reação continue após ele ter sido acoplado à fita de DNA complementar. Isso faz com que existam vários fragmentos de DNA complementar de tamanhos diferentes.

Os fragmentos são separados por tamanho em um processo chamado eletroforese. Nesse processo, as cadeias são colocadss em um gel e é aplicada uma corrente elétrica que faz com que os fragmentos migrem para a direção oposta do gel. Os menores fragmentos tendem a migrar com mais facilidade, e ao final do experimento estarão mais próximas do lado oposto do gel.

É então aplicada uma radiação de forma a excitar marcadores químicos ligados ao dideoxinucleotídeos. Cada tipo de dideoxinucleotídeo é ligado a um marcador químico fluorescente diferente, o que permite diferenciar em qual base cada fragmento de DNA termina. Como as cadeias estão separadas por tamanho, é possível sequenciar o DNA a partir da última base de cada fragmento.

Esse foi

Pirosequenciamento 454 [30]: é um método de sequenciamento paralelizado que detecta a atividade da enzima DNA polimerase com uma outra enzima quimioluminescente.

O DNA é amplificado dentro de gotas de água em uma solução de óleo onde cada gota contém um único molde de DNA acoplado a um primer. Cada um dos moldes de DNA é então submetido ao mesmo processo:

Soluções de cada tipo de nucleotídeo (A, C, G e T) e DNA polimerase são adicionadas e removidas em sequência de cada gota de água contendo um molde de DNA. Luz é produzida somente quando a solução de nucleotídeo complementa a primeira base não-emparelhada do molde. A sequência desses

sinais quimioluminescentes determinam a sequência do molde, uma base de cada vez. Quanto mais forte o sinal, mais nucleotídeos daquele tipo foram adicionados à fita complementar.

Esse método é capaz de gerar milhões de sequências de aproximadamente 400-500 pares de bases.

Sequenciamento Illumina (Solexa) [16]: A Solexa, agora parte da empresa Illumina, desenvolveu uma tecnologia de sequenciamento paralelo em que moléculas de DNA são primeiro acopladas a primers e amplificadas de modo a formar colônias de clonagem locais.

Quatro tipos de dideoxinucleotídeos com marcadores fluorescentes são adicionados, e os nucleotídeos não incorporados na sequência são removidos da solução. Diferente do pirosequensiamento, o DNA só é extendido um nucleotídeo por vez. Uma câmera fotografa os dideoxinuleotídeos marcados e então o marcador e o terminal que bloqueia a continuação da síntese de DNA são removidos, permitindo um novo ciclo [20].

Esse tipo de sequenciamento gera milhões de sequências de 50-100 pares de base.

SOLiD sequencing [5]: é uma tecnologia que aplica sequenciamento por ligação. Em vez de usar a DNA polimerase, uma outra enzima, chamada DNA ligase é usada para identificar o nucleotídeo presente em uma dada posição do DNA.

DNA ligase é uma enzima que une as pontas de moléculas de DNA. O sequenciamento por ligação se utiliza da sensitividade dessa enzima para alinhar pares de bases.

O DNA a ser sequenciado passa por um processo de amplificação PCR e cada fita única de DNA gerado é processada paralelamente. Cada fita é acoplada a uma sequência conhecida através de uma pequena sequência "âncora". Uma mistura de oligonucleotídeos (com oito a noves bases) são adicionado à reação, marcados com substâncias fluorescentes de acordo com a posição que será sequenciada. Essas moléculas se unem ao DNA desconhecido nas proximidades da sequência âncora, e a DNA ligase preferencialmente une as moléculas quando as bases casam com às do DNA desconhecido. Baseado na fluorescência produzida pela mólecula, é possível inferir qual nucleotídeo foi unido naquela posição da sequência.

É possível construir oligonucleotídeos de tal forma que se possa remover o marcador e permitir um novo ciclo de ligação, para gerar gsequências mais longas. Esse método sequencia toda n-ésima base, onde n é o tamanho do oligonucleotídeo utilizado. Para sequenciar as posições intermediárias, a âncora e os oligonucleotídeos ligado são separados da fita de DNA e uma outra rodada de sequenciamento pode ser iniciada com uma âncora mais curta.

Esse método gera bilhões de sequências de aproximadamente 50 pares de bases.

3 Fluxo de Trabalho de um Projeto Metagenômico

Os primeiros estudos de genomas que utilizaram procedimentos de sequenciamento foram feitos no meio da década de 1970. Desde então foram desenvolvidas técnicas de sequenciamento mais eficientes, os genomas de milhares de espécies se encontram em bancos de dados e ferramentas como o BLAST [2] permitem buscar e comparar sequências com as já existentes nesses bancos de dados.

Ainda assim, o sequenciamento de genomas de organismos isolados tem seus limites. Primeiro, limitações tecnológicas impõem que o organismo deve ser antes mantido em cultura e clonado para que se faça o sequenciamento de seu genoma completo [35]. No entanto, apenas uma pequena porcentagem de micróbios na natureza podem ser cultivados, o que significa que os dados genômicos existentes são altamente tendenciosos e não representam uma imagem verdadeira das espécies microbianas [29]. Segundo, raramente micróbios vivem em comunidades isoladas de outras espécies: espécies interagem tanto com outras quanto com o seu habitat, incluindo organismos hospedeiros [35].

Novas tecnologias de sequenciamento e a drástica redução dos custos do sequenciamento ajudaram a superar essas limitações. Atualmente é possível obter informação genômica diretamente de comunidades microbianas em seus habitats naturais. Esses dados de sequências de DNA retirados diretamente do ambiente são chamados de Metagenoma [35].

Por não ser necessário obter culturas puras para o sequenciamento, a metagenômica promete revelar genomas da maioria dos microorganismos os quais não se podem fazer culturas [15]. Além disso, como as amostras são obtidas de comunidades e não de populações isoladas, metagenômica pode servir para estabelecer hipóteses sobre a interação entre os membros da comunidade [19].

O processo de um típico projeto metagenômico é brevemente descrito a seguir, mostrando suas diferenças em relação a uma análise genômica tradicional, algumas tecnologias existentes e os desafios computacionais associados.

3.1 Amostragem

O primeiro passo no estudo metagenômico é a obtenção de amostras ambientais. As amostras devem representar a população de onde elas fora retiradas. O problema é descobrir quantas amostras são necessárias para obter uma boa representatividade [35]. Curvas de raridade são utilizadas para estimar a fração de espécies sequenciadas. Essas curvas relacionam o número de espécies como uma função do número de indivíduos na amostra.

A complexidade da comunidade também deve ser avaliada antes do sequenciamento [19]. Complexidade da comunidade é uma função do número de espécies em uma comunidade (riqueza) e sua abundância relativa (igual-

dade). Uma comunidade com muitas espécies que são próximas em abundância é mais complexa que uma comunidade com menos espécies que tem abundâncias desiguais.

A presença ou ausência de uma população dominante afeta bastante o tipo de análise que pode ser feita, independente do número total de espécies [19]. Populações dominantes que abrangem mais que uma pequena porcentagem do número total de células em uma comunidade terão uma representação maior no conjunto de dados, e uma maior chance de montagem e recuperação de contigs.

3.2 Anotação de Metadados

Coletar dados associados com uma amostra ambiental melhora bastante a habilidade de interpretar as sequências, particularmente para uma análise comparativa de uma série temporal ou espacial [7, 34]. Tais metadados incluem dados bioquímicos, geográficos, data de coleta, método de extração de DNA, dentre outros. Metadados podem se tornar muito importantes quando dados suficientes são gerados para comparar comunidades [19].

Quanto mais condições de amostragem para metagenomas forem reportados, mais detalhadas podem ser as inferências das características ambientais [27]. Foi proposto na comunidade um padrão de *'Minimum Information about a Metagenomic Sequence (MIMS)'* para que sejam aplicadas medidas essenciais quando os dados forem submetidos a bases públicas [9].

3.3 Sequenciamento

Em metagenômica, o sequenciamento por *shotgun* é feito da mesma forma que em genomas de culturas clonadas. No entanto, o material genético original não é de apenas um organismo, mas de uma comunidade. Dependendo da amostra, o DNA fornece apenas um fragmento do genoma dos organismos daquele ambiente.

Métodos de sequenciamento de segunda geração como os descritos na Seção 2.3.2 estão substituindo o sequenciamento Sanger para genomas e metagenomas pequenos [35].

3.4 Montagem

Montagem é o processo de combinar sequências lidas em trechos contínuos de DNA, chamados *contigs*, baseados na similaridades das sequências lidas [19]. A *sequência consenso* para um contig é ou o nucleotídeo de maior qualidade em qualquer leitura em cada posição, ou o nucleotídeo mais frequente em cada posição. O número de leituras em que a sequência consenso foi baseada é chamado de *cobertura*.

Quando se está sequenciando um genoma completo, as leituras são montadas em sequências ou contigs cada vez maiores, e finalmente no genoma

completo [35]. Lidando com dados genômicos, é comum analisar grandes sequências. Em contraste, na maioria dos metagenomas, uma montagem completa não é possível. Primeiro, porque a amostragem é incompleta e muitos, se não todos, os genomas das espécies estão apenas parcialmente amostrados. Segundo, porque as informações das espécies em si é incompleta, e é difícil mapear leituras individuais a suas espécies de origem. Existe também o perigo de se montar sequências a partir de leituras de espécies diferentes, gerando quimeras.

Usualmente é possível montar a maior parte dos genomas de ambientes que tenham um pequeno número de espécies dominantes [10], porém, amostras com alta riqueza de espécies, como solo [33], dificilmente podem ser montadas.

Montadores atuais como Phrap [12], Arachne [17], CAP3 [14], e Celera [3] estão sendo adaptados e usados para montar metagenomas. A montagem de metagenomas é um processo problemático e qualquer programa de montagem produzirá vários erros [19]. Idealmente, toda montagem metagenômica deve ser inspecionada manualmente. Erros de montagem podem ser identificados com ferramentas de visualização, como o Consed [11], que são utilizados para facilitar a finalização de genomas. Ainda assim, a grande quantidade da maioria dos dados metagenômicos impossibilitam a inspeção manual, e por sua vez a correção de todos os erros de montagem identificados.

Uma abordagem para essa limitação é fazer duas ou mais montagens dos mesmos dados utilizando diferentes montadores [10] para facilitar a identificação de montagens errôneas.

Outra estratégia para aliviar os problemas de montagem é o uso de sequências de referências. No entanto, o número de genomas de referências ainda é insuficiente para montagens de metagenomas complexos e o processo de classificação só parece ser satisfatório em comunidades muito simples [27].

Muitos algoritmos de montagem representam cada leitura como um vértice e cada sobreposição como uma aresta entre os vértices que se sobrepõem. Encontrar a montagem correta é então reduzida a um problema de encontrar um caminho hamiltoniano no grafo, um caminho onde cada vértice é visitado apenas uma vez [35]. Para leituras curtas, no entanto, essa técnica não é adequada. Para estabelecer uma cobertura adequada, leituras curtas devem ser produzidas em grandes quantidades. Isso faz com que o grafo tenha um grande número de vértices e arestas. O tempo necessário para resolver o problema do caminho Hamiltoniano cresce exponencialmente com o número de vértices, e o problema se torna intratável com as grandes quantidades de sequências geradas pelos sequenciadores de segunda geração.

Uma solução é usar os vértices para representar palavras e as leituras em si serem as arestas conectando os vértices. Assim, o grande número de leituras e suas redundâncias não afetam o número de nós, e o problema é reduzido para o de encontrar um caminho Euleriano, onde cada aresta é visitada apenas uma vez. Existem algoritmos mais eficientes para encontrar

caminhos Eulerianos. O montador Velvet [36] é um exemplo que utiliza essa abordagem.

Recentemente, foi desenvolvida uma extensão do montador Velvet, chamada MetaVelvet [24], para a montagem de leituras curtas de metagenomas. A idéia é decompor o grafo construído de uma mistura de sequências de várias espécies em sub-grafos individuais, que servem de esqueleto para a diferenciação de espécies dentre as sequências.

3.5 Predição de Genes

Genes são a unidade básica funcional em um genoma, que podem compor grandes unidades funcionais como *operons*, unidades de transcrição e redes funcionais [35].

Predição de genes é o procedimento de identificar proteínas e sequências de RNA codificantes nas amostras de DNA [19]. Duas abordagens diferentes são aplicadas para predição de genes: métodos intrínsecos e extrínsecos. Métodos intrínsecos analisam propriedades de sequências de genomas para diferenciar entre sequências codificadoras e regiões não-codificadoras [18]. As ferramentas que utilizam esses métodos são, em sua maioria, baseadas em aprendizado supervisionado e métodos estatísticos de reconhecimento de padrões. Genemark.hmm [4] é um programa que usa modelos de Markov não-homogêneos baseados em análise de frequência de monocodons para a predição de genes. Uma grande vantagem de métodos intrínsecos é que eles permitem a identificação de genes sem homólogos em bases de dados disponíveis [19].

Métodos extrínsecos predizem genes procurando por fragmentos de DNA que foram conservados durante a evolução. A maioria dos novos genes são formados por eventos de duplicação, rearranjos e mutações de genes existentes [6]. Esses métodos usam ferramentas como o BLAST (Basic Local Alignment Search Tool) [1] para identificar genes similares aos observados e informar a existência de famílias de genes dentre o metagenoma. Entretanto, o BLAST por si só não pode ser usado para encontrar novas famílias e novos genes [35].

Krause e seus colegas [18] apresentaram um algoritmo que não usa comparações de pares de sequências, mas combina informações de todos os *hits* do BLAST ao mesmo tempo. Calculando a taxa de substituições sinônimas é possível diferenciar sequências codificadoras de regiões não-codificadoras, pois em proteínas funcionais os genes codificadores mostram um número muito maior de substituições sinônimas.

Novamente, a natureza incompleta e fragmentada dos dados metagenômicos apresenta desafios na identificação de genes. Muitas leituras permanecem intocadas, em vez de serem unidas a outras e formarem contigs ou ainda são quimeras devido a erros de montagem [35].

Os métodos de predição de genes tem de ser adaptados para lidar com a

enorme quantidade de genes fragmentados em sequências curtas, bem como com a diversidade filogenética das amostras, que dificultam o uso de conjuntos de treinamento específico de espécies, e com a baixa qualidade de sequências [27].

3.6 Classificação

Após a predição de genes, é importante associá-los às suas espécies de origem (ou grupos taxonômicos de alto-nível). Essa análise é chamada *binning* ou classificação [35].

Uma forma de classificar sequências é encontrar similaridades com sequências de referência que podem ser usadas para construir uma árvore [35]. Essa técnica é útil quando a maioria das sequências da amostra possuem similaridades significantes com sequências de referências conhecidas. Sequências de genes preditas, quando não tem homólogos, são adicionados em seu próprio nó isolado na árvore. O quadro resultante das sequências na árvore de espécies podem mostrar uma visão geral das espécies dominantes da amostra.

No entanto, existem várias limitações para esse tipo de abordagem. Primeiro, 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 [15]. Segundo, genes derivado de dados metagenômicos, particularmente aqueles com montagem mínima, são normalmente fragmentados e produzem alinhamentos incompletos. Terceiro, predição errônea de genes, particularmente de proteínas ribossomais, às vezes não são detectadas por preditores de genes automáticos pelo seu tamanho reduzido [21]. Finalmente, genes informativos conservados filogeneticamente representam apenas uma pequena fração do total de dados metagenômicos [19].

Mesmo com tantas limitações, o algoritmo SOrt-ITEMS [23] consegue bons resultados. O algoritmo adota uma abordagem exaustiva para julgar primeiro a qualidade do alinhamento de uma sequência com seus *hits* e chega a um nível apropriado na árvore taxonômica para o qual a sequência pode ser atribuída. O algoritmo então usa uma abordagem de genes ortólogos para identificar hits que mostram uma significante similaridade recíproca com a sequência consultada. Genes ortólogos são genes de diferentes espécies que evoluíram de um ancestral comum através de especiação. Normalmente esses genes tem a mesma função através da evolução. O algoritmo avalia os alinhamentos obtidos entre a leitura e seus *hits* correspondentes na saída do BLASTx [2] para criar subconjuntos específico de *hits* que compartilham uma relação ortóloga com a leitura dada.

Uma outra abordagem é a classificação baseada em composição, que usa análises estatísticas das sequências [35]. Modelos de Markov baseados em frequências de k-mers se mostraram bem poderosos para análises estatísticas [32]. Esses métodos, porém, não são livres de erros. Quanto mais pró-

ximas filogeneticamente são as espécies estudadas no metagenoma, e quanto mais numerosas elas são, maior é a frequência de erros de classificação. A força da classificação baseada em k-mers é que sequências de referências não são necessárias para a classificação em si, toda a informação é intrínseca. Isso permite classificar sequências que tem poucos ou nenhum homólogo e portanto nenhuma função conhecida.

Outros métodos de classificação baseado em composição seguem o conhecimento de que processos celulares como uso de códons, sistemas de restrição-modificação, e mecanismos de consertos de DNA produzem assinaturas de composição de sequências, primariamente frequências de oligonucleotídeos (palavras), que são distintas em diferentes genomas [8].

Métodos baseados em composição podem ser divididos em procedimentos (de clusterização) supervisionados e não-supervisionados [19]. Procedimento não-supervisionados agrupam fragmentos metagenômicos em um espaço de assinaturas de composições sem a necessidade de um modelo de treinamento com sequências de referência. Uma vantagem da classificação não-supervisionada é que novas populações podem ser classificadas por compartilharem semelhanças de características de composição, embora a identificação de fragmentos ainda necessite da similaridade de sequências para referenciar organismos. Uma desvantagem é que esses métodos tendem a focar em classes majoritárias de um conjunto de dados e não tem bons resultados em populações de baixa abundância.

Métodos supervisionados classificam fragmentos metagenômicos contra modelos treinados em sequências de referência classificadas e, a princípio, podem classificar fragmentos de populações de baixa abundância, se há um modelo para treinar. Como são capazes de aprender características relevantes que distinguem uma população em particular das outras, métodos supervisionados geralmente tem uma acurácia (sensitividade e especificidade) maior que modelos não-supervisionados, mas dependem de um conhecimento prévio sobre as espécies sequenciadas [19].

3.7 Análise de Dados

Os primeiros passos da análise de qualquer metagenoma envolvem comparar as sequências lidas de uma amostra com bases de dados de sequências conhecidas. Essa tarefa computacionalmente intensiva provê os tipos básicos de dados para várias análises posteriores [22].

A análise de comunidades de baixa complexidade é, em vários aspectos, similar à análise de genomas isolados. Genomas de populações dominantes tem cobertura suficiente e contexto genético para permitir uma reconstrução metabólica razoavelmente compreensível [19]. Uma combinação de composição de sequências, classificação e montagem parece ser suficiente para quase que completamente sequenciar os membros da comunidade. Isso permite a atribuição de algumas atividades metabólicas para membros individuais do

ecossistema [27]. Se mais de uma população dominante é sequenciada, a potencial interação metabólica dessas populações também pode ficar aparente [19].

Sequenciamentos de comunidades microbianas de alta complexidade resultam em pouca ou nenhuma montagem de fragmentos [33]. A abordagem mais comum para a anotação funcional é baseada no BLAST, entretanto essa técnica só permite anotar funcionalmente 25-50% das proteínas por metagenoma publicado [28].

Duas metodologias adicionais estão sendo usadas para melhorar esse número: buscas homólogas e contexto de genes [27]. Enquanto abordagens baseadas em homologias são úteis para traçar novas subfamílias funcionalmente distintas dentro de superfamílias conhecidas, abordagens baseadas em contexto são particularmente úteis para descobrir e anotar proteínas completamente novas associadas a processos conhecidos.

O servidor metagenômico RAST [22] é um sistema de código aberto desenvolvido para processar automaticamente as sequências de metagenomas, fazer comparações com bases de dados existentes e computar reconstruções filogenéticas e classificar funcionalmente os potenciais genes encontrados na amostra. Ele está disponível através da web para todos os pesquisadores.

A alta densidade de áreas codificadoras em genomas de bactérias e o tamanho médio de um gene significam que a maioria das leituras irão capturar uma sequência codificadora [19]. Isso permite uma análise dos dados centrada em genes, que trata a comunidade como um agregado, ignorando a contribuição de espécies individuais.

Assim, pode-se estudar o potencial funcional da comunidade microbiana do qual o metagenoma foi derivado [35]. Primeiro são atribuídas funções biológicas para os genes. Depois são descobertos genes que constituem redes biológicas, como vias metabólicas, nos dados. Vários estudos foram feitos e levaram à descoberta de vias metabólicas complementares de micróbios que constituem a comunidade.

A força do método está na comparação relativa de famílias de genes ou abundâncias de subsistemas entre metagenomas para realçar diferenças funcionais. Como a determinação de frequências relativas de famílias de genes dentro e entre conjuntos de metagenomas é um aspecto fundamental desse método, é importante que as frequências não sejam mascaradas pela montagem. Ou a análise é feita em leituras não montadas, ou as coberturas dos contigs devem ser levadas em consideração [21].

3.8 Metagenômica Comparativa

A análise de sequências de genomas mostrou que muito se ganha com abordagens comparativas, uma vez que elas provêm contexto para amostras individuais [27]. 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.

O framework SEED [25] foi desenvolvido para genômica comparativa e foi usado como base para funcionalidades do servidor RAST [22]. São construídos mapas de taxonomias e subsistemas para encapsular diferenças entre as amostras e comparar suas similaridades.

Apesar do grande potencial de abordagens metagenômicas comparativas, elas devem ser aplicadas com cautela [27]. Vários fatores biológicos específicos do ambiente e vários problemas técnicos dificultam a comparação direta de ambientes, pois eles influenciam um ao outro e a maioria dos resultados derivados. Diferenças no tamanho médio de genomas das amostras implicitamente levam a diferenças da composição funcional relativa das amostras. A complexidade filogenética das amostras influenciam fortemente a análise, e diferentes características funcionais dos ambientes podem resultar em diferentes taxas evolucionárias, distorcendo a detecção de genes e funções. A cobertura limitada das amostras e a diversidade filogenética podem dificultar a comparação direta de parâmetros genéticos populacionais, uma vez que estimativas robustas baseadas em poucos dados são difíceis e as espécies abundantes podem ocultar a real estrutura populacional das amostras. Além dos vários fatores biológicos, muitos problemas técnicos relacionados a amostragem, sequenciamento e anotação influenciam toda a análise.

4 Proposta

O estudo de metagenomas é recente e contem muitos desafios computacionais. A maior parte das ferramentas existentes não foram projetadas para lidar com dados metagenômicos. A proposta deste trabalho é fazer um estudo metagenômico a partir de dados coletados na Fundação Parque Zoológico de São Paulo (FPZSP).

A FPZSP possui uma 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, devido às diversas condições naturais desconhecidas, impossibilitando a reprodução do habitat em laboratório. Assim, a abordagem metagenômica é a mais indicada para esse estudo.

Com os dados do sequenciamento de amostras desse material, realizaremos testes com as diversas ferramentas de montagem disponíveis. Com isso pretendemos obter a melhor montagem possível para o conjunto de dados da FPZSP.

A predição de genes é uma outra fase importante do estudo metagenômico, e também realizaremos testes com ferramentas destinadas a esse fim. Pretendemos analisar os métodos utilizados e propor melhorias para esse procedimento.

As ferramentas disponíveis para classificação também serão analisadas de forma que suas limitações sejam estressadas, e diversas combinações de parâmetros sejam testadas. Os resultados desses testes auxiliarão num melhor entendimento sobre as amostras estudadas e qual conjunto de softwares é o mais indicado atualmente para esse tipo de estudo metagenômico.

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.

5 Cronograma

As atividades a serem realizadas durante a execução deste trabalho segue descrita na Tabela 1.

Tabela 1: Cronograma de Atividades

	2011										2012											2013	
	mar	abr	mai	jun	jul	ago	set	out	nov	dez	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez	jan	fev
1	I																							
2					II			III																
3							IV			V	VI													
4											VII			VIII			IX							
5																	X		XI		XII			
6																							XIII	XIV

1. Créditos:
    - I - Obtenção dos créditos obrigatórios.
2. Proposta:
    - II - Escrita da proposta de qualificação de mestrado.
    - III - Exame de qualificação de mestrado.
3. Montagem:
    - IV - Análise dos métodos existentes e identificação de melhorias nos métodos de montagem metagenômica.
    - V - Montagem metagenômica das sequências do Zoológico.
    - VI - Escrita dos resultados obtidos nos testes de montagem metagenômica.

4. Predição de Genes:

    - VII - Análise dos métodos existentes e identificação de melhorias nos métodos de predição de genes em dados metagenômicos.
    - VIII - Predição de genes dos dados metagenômicos do Zoológico.
    - IX - Escrita dos resultados obtidos nos testes de predição de genes em dados metagenômicos.

5. Classificação e Análise de Dados:

    - X - Estudo dos métodos e ferramentas de classificação de dados metagenômicos existentes.
    - XI - Testes com os programas de classificação e valores de parâmetros.
    - XII - Análise e escrita dos resultados obtidos nos testes de classificação em dados metagenômicos.

6. Dissertação:

    - XIII - Revisão final do texto da dissertação.
    - XIV - Defesa da dissertação.

Referências

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3): 403–410, October 1990.

[2] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997.

[3] Celera assembler. http://sourceforge.net/projects/wgsassembler/.

[4] R. K. Azad and M. Borodovsky. Probabilistic methods of identifying genes in prokaryotic genomes: Connections to the HMM theory. *Briefings in Bioinformatics*, 5(2):118–130, June 2004.

[5] Applied Biosystems. Solid - next generation sequencing. http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html.

[6] C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann. Evolution of the Protein Repertoire. *Science*, 300(5626):1701Ű1703, June 2003.

[7] E. F. DeLong, C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N-U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl. Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science*, 311(5760):496–503, January 2006.

[8] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Molecular Biology and Evolution*, 16(10):1391–1399, October 1999.

[9] D. Field, G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M. J. Allen, S. V. Angiuoli, M. Ashburner, N. Axelrod, S. Baldauf, S. Ballard, J. Boore, G. Cochrane, J. Cole, P. Dawyndt, P. De Vos, C. dePamphilis, R. Edwards, N. Faruque, R. Feldman, J. Gilbert, P. Gilna, F. O. Glockner, P. Goldstein, R. Guralnick, D. Haft, D. Hancock, H. Hermjakob, C. Hertz-Fowler, P. Hugenholtz, I. Joint, L. Kagan, M. Kane, J. Kennedy, G. Kowalchuk, R. Kottmann, E. Kolker, S. Kravitz, N. Kyrpides, J. Leebens-Mack, S. E. Lewis, K. Li, A. L. Lister, P. Lord, N. Maltsev, V. Markowitz, J. Martiny, B. Methe, I. Mizrachi, R. Moxon, K. Nelson, J. Parkhill, L. Proctor, O. White, S.-A. Sansone, A. Spiers, R. Stevens, P. Swift, C. Taylor, Y. Tateno, A. Tett, S. Turner, D. Ussery, B. Vaughan, N. Ward, T. Whetzel, I. San Gil,

G. Wilson, and A. Wipat. The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, 26(5):541–547, May 2008.

[10] H. García Martín, N. Ivanova, V. Kunin, F. Warnecke, K. W. Barry, A. C. Mchardy, C. Yeates, S. He, A. A. Salamov, E. Szeto, E. Dalin, N. H. Putnam, H. J. Shapiro, J. L. Pangilinan, I. Rigoutsos, N. C. Kyrpides, L. L. L. Blackall, K. D. Mcmahon, and P. Hugenholtz. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nature biotechnology*, 24(10):1263–1269, October 2006.

[11] D. Gordon, C. Abajian, and P. Green. Consed: A graphical tool for sequence finishing. *Genome Research*, 8(3):195–202, March 1998.

[12] P. Green. Phrap assembler. www.phrap.org, August 2011.

[13] NCBI Taxonomy Homepage. The genetic codes. http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c, August 2011.

[14] X. Huang and A. Madan. CAP3: A DNA Sequence Assembly Program. *Genome Research*, 9(9):868–877, September 1999.

[15] P. Hugenholtz. Exploring prokaryotic diversity in the genomic era. *Genome Biology*, 3(2):reviews0003.1–reviews0003.8, 2002.

[16] Illumina. Solexa sequencing technology. http://www.illumina.com/technology/solexa_technology.ilmn.

[17] D. B. Jaffe, J. Butler, S. Gnerre, E. Mauceli, K. Lindblad-Toh, J. P. Mesirov, M. C. Zody, and E. S. Lander. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome research*, 13(1): 91–96, January 2003.

[18] L. Krause, N. N. Diaz, D. Bartels, R. A. Edwards, A. Pühler, F. Rohwer, F. Meyer, and J. Stoye. Finding novel genes in bacterial communities isolated from the environment. *Bioinformatics*, 22(14):e281–e289, July 2006.

[19] V. Kunin, A. Copeland, Al. Lapidus, K. Mavromatis, and P. Hugenholtz. A Bioinformatician's Guide to Metagenomics. *Microbiology and Molecular Biology Reviews*, 72(4):557–578, December 2008.

[20] E. R. Mardis. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, 9(1):387–402, June 2008.

[21] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, 4(6):495–500, April 2007.

[22] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1):386–8, December 2008.

[23] M. Monzoorul Haque, T. S. S. Ghosh, D. Komanduri, and S. S. Mande. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, 25(14): 1722–1730, July 2009.

[24] T Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara. Metavelvet: An extension of velvet assembler to *de novo* metagenome assembly from short sequence reads. In *The 2011 ACM Conference on Bioinformatics, Computational Biology an Biomedicine*, pages 116–124, August 2011.

[25] R. Overbeek, T. Begley, R. M. Butler, Chuang H.-Y. Choudhuri, J. V., M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, Roy. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Rückert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33(17), October 2005.

[26] G. J. Porreca, J. Shendure, and G. M. Church. *Polony DNA sequencing.*, chapter Chapter 7. Hoboken (New Jersey): John Wiley and Sons, Inc, 2006.

[27] J. Raes, K. U. U. Foerstner, and P. Bork. Get the most out of your metagenome: computational analysis of environmental sequence data. *Current opinion in microbiology*, 10(5):490–498, October 2007.

[28] J. Raes, E. D. D. Harrington, A. H. H. Singh, and P. Bork. Protein function space: viewing the limits or limited by our view? *Current opinion in structural biology*, 17(3):362–369, June 2007.

[29] M. S. Rappé and S. J. Giovannoni. The uncultured microbial majority. *Annual review of microbiology*, 57(1):369–394, 2003.

[30] M. Ronaghi, M. Uhlén, and P. Nyrén. A sequencing method based on real-time pyrophosphate. *Science*, 281(5375):363–365, July 1998.

[31] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of The National Academy of Sciences of the United States of America*, 74(12):5463–5467, December 1977.

[32] S. Schbath, B. Prum, and E. de Turckheim. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *Journal of computational biology*, 2(3):417–437, 1995.

[33] S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. Comparative Metagenomics of Microbial Communities. *Science*, 308(5721):554–557, April 2005.

[34] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, April 2004.

[35] J. C. Wooley, A. Godzik, and I. Friedberg. A primer on metagenomics. *PLoS Computacional Biology*, 6(2):e1000667, February 2010.

[36] D. R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, May 2008.