ANDRÉ LUIZ DE OLIVEIRA

**GenSeed-HMM: desenvolvimento de uma
plataforma para reconstrução de sequências e
sua aplicação em dados de sequenciamento de
nova geração**

Dissertação apresentada ao Programa de Pós-Graduação em Biologia da Relação Patógeno-Hospedeiro do Instituto de Ciências Biomédicas da Universidade de São Paulo, para obtenção do Título de Mestre em Ciências.

São Paulo
2012

ANDRÉ LUIZ DE OLIVEIRA

**GenSeed-HMM: desenvolvimento de uma
plataforma para reconstrução de sequências e
sua aplicação em dados de sequenciamento de
nova geração**

Dissertação apresentada ao Programa de Pós-Graduação em Biologia da Relação Patógeno-Hospedeiro do Instituto de Ciências Biomédicas da Universidade de São Paulo, para obtenção do Título de Mestre em Ciências.

Área de concentração: Biologia da Relação Patógeno-Hospedeiro

Orientador: Prof. Dr. Arthur Gruber

Versão original

São Paulo
2012

DADOS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
Serviço de Biblioteca e Informação Biomédica do
Instituto de Ciências Biomédicas da Universidade de São Paulo

reprodução não autorizada pelo autor

Oliveira, André Luiz de.

GenSeed-HMM: desenvolvimento de uma plataforma para reconstrução de sequências e sua aplicação em dados de sequenciamento de nova geração / André Luiz de Oliveira. -- São Paulo, 2012.

Orientador: Prof. Dr. Arthur Gruber.

Dissertação (Mestrado) – Universidade de São Paulo. Instituto de Ciências Biomédicas. Departamento de Parasitologia. Área de concentração: Biologia da Relação Patógeno-Hospedeiro. Linha de pesquisa: Bioinformática aplicada ao estudo de parasitos.

Versão do título para o inglês: GenSeed-HMM: development of a platform for sequence reconstruction and application on next-generation sequencing data.

1. Bioinformática 2. Reconstrução de sequências 3. Sequenciamento de nova geração 4. Perfis de HMM 5. Metagenômica 6. Montagem de DNA I. Gruber, Prof. Dr. Arthur II. Universidade de São Paulo. Instituto de Ciências Biomédicas. Programa de Pós-Graduação em Biologia da Relação Patógeno-Hospedeiro III. Título.

ICB/SBIB0102/2012

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Candidato(a): André Luiz de Oliveira.

Título da Dissertação: GenSeed-HMM: desenvolvimento de uma plataforma para reconstrução de sequências e sua aplicação em dados de sequenciamento de nova geração.

Orientador(a): Prof. Dr. Arthur Gruber.

A Comissão Julgadora dos trabalhos de Defesa da **Dissertação de Mestrado**,
em sessão pública realizada a/...../.....,

(☐) **Aprovado(a)**

(☐) **Reprovado(a)**

Examinador(a): Assinatura:
Nome:
Instituição:

Examinador(a): Assinatura:
Nome:
Instituição:

Presidente: Assinatura:
Nome:
Instituição:

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Cidade Universitária "Armando de Salles Oliveira"
Av. Prof. Lineu Prestes, 2415 – CEP. 05508-000 São Paulo, SP – Brasil
Telefone : (55) (11) 3091-7733 - telefax : (55) (11) 3091-8405
e-mail: cep@ich.usp.br

“
Comissão de Ética em Pesquisa

CERTIFICADO DE ISENÇÃO

Certificamos que o Protocolo CEP-ICB Nº **393/10** referente ao projeto intitulado: *“GenSeed-HMM: desenvolvimento de uma plataforma para a reconstrução de sequências e aplicação na descoberta de novos vírus”* sob a responsabilidade de **André Luiz de Oliveira**, foi analisado na presente data pela CEUA - COMISSÃO DE ÉTICA NO USO DE ANIMAIS e pela CEPSh- COMISSÃO DE ÉTICA EM PESQUISA COM SERES HUMANOS, tendo sido deliberado que o referido projeto não envolve manipulação animal ou humana que justifique uma aprovação quanto aos princípios éticos exigidos por ambas as Comissões.

São Paulo, 21 de junho de 2010.

PROF. DR. WOTHAN TAVARES DE LIMA
Coordenador da CEUA - ICB/USP

PROF. DR. PAOLO M.A. ZANOTTO
Coordenador da CEPsh - ICB/USP

**UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS**

Cidade Universitária "Armando de Salles Oliveira"
Av. Prof. Lineu Prestes, 2415 - cep. 05508-000 São Paulo, SP - Brasil
Telefone: (55) (011) 38130900 - e-mail: cep@icb.usp.br

Decl. CEP.026/2012.

DECLARAÇÃO

Em adendo ao Protocolo de Isenção CEP-ICB nº 393/10, aprovado em 21 de junho de 2010, informo que o título do Projeto foi alterado para "*GenSeed-HMM: desenvolvimento de uma plataforma para a reconstrução de sequências e sua aplicação em dados de sequenciamento de nova geração*", sem modificações de seu conteúdo.

São Paulo, 18 de junho de 2012.

Prof. Dr. WOTHAN TAVARES DE LIMA
Coordenador da CEUA - ICB/USP

Prof. Dr. PAULO M.A. ZANOTTO
Coordenador da CEPISH- ICB /USP

AGRADECIMENTOS

Aos meus pais Oswaldo e Marcia por todo apoio e amor incondicional. Por mais cliché que seja, sem eles nada disso seria possível ou mesmo imaginável. À minha família: Tio Fernando, Tio Pedro (in memoriam), Tia Magda, Tia Marley, Tia Zezé, e aos meus primos Felipe, Livia, Marcos, Matheus e Enzo.

Ao Arthur Gruber, pela orientação e dedicação ao longo destes três anos. Sou muito grato pela oportunidade, lições, puxões de orelha, por todas as nossas conversas e por ajudar no meu processo de amadurecimento e formação científica. Ao Alan Durham, pela co-orientação e ajuda na parte computacional do meu projeto.

Ao Marcelo Arruda Fiuza de Toledo, Prof. Dr. Paolo Marinho de Andrade Zanotto, e Tiago José Paschoal Sobreira, pela ajuda durante os testes iniciais e prova de principio do programa desenvolvido nesse trabalho.

Ich danke meinem Deutschlehrer Erwin Wiesenhofer. Vielen Dank für Ihre Geduld, Freundschaft, und dass Sie mir zeigten, das Leben sehr kurz ist um Deutsch zu lernen.

Agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP – número do processo 2010/04609-1) e à CAPES o suporte financeiro.

“Antes, ainda na escada, no descendimento, ele mirou, melhor, a multidão, deogenésica, diogenista. Vindo o quê, de qual cabeça, o caso que já não se esperava. Deu-nos outra cor. Pois, tornavam a endoidá-lo? Apenas proclamou: – ***“Viva a luta! Viva a liberdade!”*** – nu, adão, nado, psiquiartista.”

João Guimarães Rosa *in* Darandina

RESUMO

Oliveira AL. GenSeed-HMM: desenvolvimento de uma plataforma para reconstrução de sequências e sua aplicação em dados de sequenciamento de nova geração. [Dissertação (Mestrado em Biologia da Relação Patógeno-Hospedeiro)]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2012.

O sequenciamento de DNA se tornou uma técnica universal e de ampla aplicação para desvendar a complexidade do genoma dos organismos vivos. Com o desenvolvimento das novas plataformas de sequenciamento, a quantidade de dados gerados cresceu em várias ordens de grandeza. Como consequência, muitos projetos de sequenciamento genômico apenas disponibilizam dados parcialmente montados ou mesmo sequências não montadas. Como muitas das questões biológicas estão dirigidas a alvos moleculares simples, a reconstrução de sequências específicas é uma tarefa muito comum. A maioria dos programas convencionais de montagem utilizam a abordagem sobreposição-leiaute-consenso. Nesse método, o processo se inicia com uma comparação das leituras todas-contra-todas, de forma a determinar as sobreposições entre elas e computar uma estrutura de grafo. Na fase do leiaute, o grafo é simplificado removendo-se as informações redundantes, e um esquema consistindo da posição relativa das leituras é estabelecido. Finalmente, o programa constrói um múltiplo alinhamento de sequências a partir da qual a sequência consenso é obtida. Essa abordagem computacional é de complexidade quadrática e dispendiosa em termos de memória alocada, podendo se tornar impraticável para alguns grupos de pesquisa que não dispõem de grandes recursos computacionais. Esse problema nos motivou a desenvolver um método de montagem progressiva dirigida por semente, o qual permite reconstruir sequências de DNA para montagem alvo-específicas partindo-se de sequências iniciais curtas de DNA ou proteína. O programa *GenSeed*, que implementa esse método, foi descrito pelo nosso grupo e pode ser aplicado para a reconstrução de fragmentos genômicos, extracromossômicos e cDNAs. Embora esse método tenha se mostrado eficiente para a reconstrução de sequências usando sementes e bases de dados derivados do mesmo organismo, a reconstrução de sequências de amostras heterólogas pode se tornar problemática. Modelos probabilísticos têm sido utilizados em perfis de alinhamentos múltiplos de sequências proteicas para estabelecer relações entre proteínas ortólogas. Um HMM de perfil transforma o alinhamento múltiplo de sequências em um sistema de pontuação posição-específico, adequado para encontrar sequências homólogas remotas em bases de dados de proteína. Com base nisso, esse trabalho teve como objetivo o desenvolvimento do *GenSeed-HMM*, uma versão do *GenSeed*, capaz de utilizar HMMs de perfis como sementes para a reconstrução de sequências específicas, e de processar dados gerados pelas novas plataformas de sequenciamento, incluindo leituras curtas. O presente trabalho relata a implementação do programa *GenSeed-HMM*, e sua validação utilizando dados reais de diferentes plataformas de sequenciamento, originados de procariotos, eucariotos, bem como de amostras metagenômicas.

Palavras-chave: Bioinformática. Reconstrução de sequências. Sequenciamento de nova geração. HMMs de perfis. Metagenômica.

ABSTRACT

Oliveira AL. GenSeed-HMM: development of a platform for sequence reconstruction and application on next-generation sequencing data. [Master's thesis in Biology of Host-Pathogen Relationship]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2012.

DNA sequencing has become a mainstream high-throughput technique to unravel the complexity of the genome of living organisms. With the development of new sequencing platforms, the amount of generated data has increased by several orders of magnitude. As a consequence, many genome sequencing projects only provide partially assembled data or even unassembled sequences. Since many biological questions rely on simple molecular targets, sequence reconstruction of specific targets is a very common task. Most of the currently available DNA assembly programs implement the overlap-layout-consensus approach. Briefly, the process starts with an all-versus-all comparison to determine all overlaps among reads and to compute a graph structure. In the layout phase, the graph is simplified by redundancy elimination, and a layout consisting of the relative position of the reads is established. Finally, the program builds a multiple sequence alignment that permits to derive a consensus sequence. This computational approach is highly demanding in terms of allocated memory and processing capacity, and may become unfeasible for research groups that do not have large computational resources. This problem motivated us to develop a seed-driven progressive assembly method for target-specific assembly of DNA sequences, starting from short DNA or protein sequences. The program GenSeed, that implements this method, has been described by our group and can be applied to the reconstruction of genomic fragments, extrachromosomal genomes, and cDNAs. Although this method has proven effective for sequence reconstruction using seeds and databases derived from the same organism, sequence reconstruction from heterologous samples may become problematic. Probabilistic models have been used in multiple sequence alignment profiles of protein sequences to establish relationships among orthologous proteins. A profile HMM transforms a multiple sequence alignment into a position-specific scoring system suitable to find homologous sequences in protein databases. On this basis, we aimed at developing GenSeed-HMM, a new version of GenSeed program that can use profile HMMs as seeds for the reconstruction of specific sequences, and incorporates the ability to work with data generated by the new sequencing platforms, including short reads. The present work reports the implementation of GenSeed-HMM program and its validation using real life data produced by different next-generation sequencing platforms, and originated from prokaryotic, eukaryotic and metagenomic samples.

Keywords: Bioinformatics. Sequence reconstruction. Next-generation sequencing. Profile HMM. Metagenomics.

LISTA DE FIGURAS

Figura 1 -	Fluxo de trabalho do algoritmo do programa GenSeed.....	29
Figura 2 -	Fluxo de trabalho do programa GenSeed-HMM.....	40
Figura 3 -	Captura de tela do programa Seeder.....	43
Figura 4 -	Reconstruções de sequências de <i>C. jejuni</i> utilizando-se o GenSeed-HMM e o montador ABySS, ou montagem global apenas com o montador. Os HMMs de perfis utilizados como sementes estão apresentados. O efeito do uso das rotinas de otimização está apresentado nas barras vermelhas.....	51
Figura 5 -	Reconstruções de sequências de <i>C. jejuni</i> utilizando-se o GenSeed-HMM e o montador SOAPdenovo, ou montagem global apenas com o montador. Os HMMs de perfis utilizados como sementes estão apresentados. O efeito do uso das rotinas de otimização está apresentado nas barras vermelhas.....	51
Figura 6 -	Reconstruções de sequências de <i>C. jejuni</i> utilizando-se o GenSeed-HMM e o montador Velvet, ou montagem global apenas com o montador. Os HMMs de perfis utilizados como sementes estão apresentados. O efeito do uso das rotinas de otimização está apresentado nas barras vermelhas.....	52
Figura 7 -	Reconstruções de sequências de <i>E. coli</i> utilizando-se o GenSeed-HMM e o montador ABySS, ou montagem global apenas com o montador. Os HMMs de perfis utilizados como sementes estão apresentados. O efeito do uso das rotinas de otimização está apresentado nas barras vermelhas.....	54
Figura 8 -	Reconstruções de sequências de <i>E. coli</i> utilizando-se o GenSeed-HMM e o montador SOAPdenovo, ou montagem global apenas com o montador. Os HMMs de perfis utilizados como sementes estão apresentados. O efeito do uso das rotinas de otimização está apresentado nas barras vermelhas.....	55
Figura 9 -	Reconstruções de sequências de <i>E. coli</i> utilizando-se o GenSeed-HMM e o montador Velvet, ou montagem global apenas com o montador. Os HMMs de perfis utilizados como sementes estão apresentados. O efeito do uso das rotinas de otimização está apresentado nas barras vermelhas.....	55
Figura 10 -	Mapeamento das sequências reconstruídas através de montagem progressiva com GenSeed-HMM/ABYSS (contig A), ou de montagem global com o programa ABySS (contigs B e C) contra a sequência genômica de referência (linha tracejada na parte superior da imagem). As sequências correspondentes aos genes de aspartato (D) e ornitina carbamoiltransferase (E) também estão representadas.....	57

Figura 11 -	Captura de tela do programa <i>Artemis</i> (Rutherford et al., 2000) mostrando uma região de 15.103 pb (assinalada com a cor verde), reconstruída a partir de leituras da plataforma Illumina de <i>D. melanogaster</i>	62
Figura 12 -	Montagem progressiva e o efeito de erros de sequenciamento nos grafos de de Bruijn.....	76

LISTA DE TABELAS

Tabela 1 -	Programas de terceiros utilizados pelo GenSeed-HMM.....	41
Tabela 2 -	Bases de dados utilizadas para validação do programa GenSeed-HMM....	44
Tabela 3 -	HMMs de perfis utilizados para validação do programa GenSeed-HMM...	45
Tabela 4 -	Reconstruções realizadas utilizando o GenSeed-HMM e o montador Newbler.....	59
Tabela 5 -	Reconstruções de sequências de dados metagenômicos de microbioma intestinal humano utilizando o programa GenSeed-HMM com os montadores ABySS, SOAPdenovo e Velvet.....	66

LISTA DE ABREVIATURAS E SIGLAS

ABYSS – Assembly By short sequences (Montagem por sequências curtas)

BLAST – Basic Local Alignment Search Tool (Ferramenta de busca de alinhamento local básica)

CAP3 – Contig Assembly Program (Programa de montagem de contigs)

DNA – Ácido desoxirribonucleico

HMM – Hidden Markov models (Modelos ocultos de Markov)

HTML – Hypertext markup language (Linguagem de marcação de hipertexto)

NCBI – National Center for Biotechnology Information (Centro Nacional de Informação Biotecnológica – EUA)

RAM – Random access memory (Memória de acesso aleatório)

PCR – Polymerase chain reaction (reação em cadeia da polimerase)

Pfam – Protein Family database (Bases de dados de famílias de proteínas)

SUMÁRIO

1 INTRODUÇÃO	17
1.1 O sequenciamento de DNA	18
1.1.1 O sequenciamento do tipo Sanger	18
1.1.2 As novas plataformas de sequenciamento de DNA	20
1.2 Aplicações do sequenciamento de DNA na biologia	21
1.3 Montagem de sequências	23
1.3.1 Abordagem da sobreposição-leiaute-consenso	24
1.3.2 Abordagem dos grafos de de Bruijn.....	25
1.4 GenSeed e a montagem progressiva dirigida por semente	26
1.5 Modelos Ocultos de Markov (HMMs) e HMMs de perfis	30
1.6 Um novo programa para reconstrução de sequências: GenSeed-HMM.....	31
2 OBJETIVOS	32
3 MATERIAL E MÉTODOS	34
3.1 Algoritmo do GenSeed-HMM e implementação.....	35
3.1.1 Sementes e bases de dados	35
3.1.2 A montagem progressiva e integração com montadores de DNA	36
3.1.3 Processo final de execução.....	39
3.2 Implementação e requerimentos.....	41
3.3 Seeder: o configurador gráfico do GenSeed-HMM	42
3.4 Validação do programa GenSeed-HMM.....	43
3.4.1 Bases de dados	44
3.4.2 HMMs de perfis	44
4 RESULTADOS	46
4.1 Prova de princípio para o desenvolvimento do programa GenSeed-HMM.....	47
4.2 Reconstruções de loci utilizando HMM de perfis e bases de dados originadas de organismos procariotos.....	49
4.2.1 Montagem progressiva com leituras de <i>Campylobacter jejuni</i>	49
4.2.2 Montagem progressiva com leituras de <i>Escherichia coli</i>	53
4.2.3 Montagem progressiva com leituras de <i>Gardnerella vaginalis</i>	58
4.3 Reconstruções de genomas extracromossomais utilizando HMM de perfis e bases de dados originadas de organismos eucariotos.....	60
4.3.1 Reconstrução do genoma mitocondrial de <i>Drosophila melanogaster</i>	61
4.3.2 Reconstrução do genoma mitocondrial de <i>Plasmodium vivax</i>	63

4.3.3 Reconstrução do genoma mitocondrial de <i>Schizosaccharomyces japonicus</i>	63
4.4 Reconstruções de loci utilizando HMM de perfis e bases de dados originadas de amostras metagenômicas	64
4.4.1 Genoma mitocondrial do mamute lanoso extinto (<i>Mammuthus primigenius</i>)	65
4.4.2 Microbioma intestinal humano.....	65
5 DISCUSSÃO	68
5.1 Características do programa GenSeed-HMM e Seeder.....	69
5.2 Desenvolvimento e desempenho do programa GenSeed-HMM.....	71
5.3 Rotinas de otimização para eliminação de erros de sequenciamento	74
5.4 Perspectivas futuras e aplicações	76
5.4.1 Reconstrução de sequências a partir de DNA fóssil.....	76
5.4.2 Uma nova abordagem para a detecção de agentes etiológicos.....	77
5.4.3 GenSeed-HMM como uma ferramenta auxiliar para o fechamento de genomas.....	79
6 CONCLUSÕES	80
REFERÊNCIAS	82

1 INTRODUÇÃO

1.1 O sequenciamento de DNA

1.1.1 O sequenciamento do tipo Sanger

O primeiro genoma a ser totalmente sequenciado foi o do bacteriófago phi-X174 (Sanger et al., 1977) no qual utilizou-se uma técnica denominada plus-and-minus (Sanger e Coulson, 1975). Em 1977, duas novas metodologias de sequenciamento de DNA foram propostas: o método de sequenciamento químico (Maxam e Gilbert, 1977) e o método de terminação de cadeia (Sanger et al., 1977). Embora ambos os métodos fossem bastante eficientes, o de terminação de cadeia acabou se tornando o método padrão para o sequenciamento de DNA por toda a comunidade internacional, graças principalmente à sua maior facilidade de execução (Schuster, 2008). O método de terminação de cadeia se baseia no uso de inibidores sintéticos 2'3'-didesoxirribonucleotídeos durante a etapa de polimerização, atuando como competidores dos análogos naturais 2'-desoxirribonucleotídeos. Os inibidores sintéticos, por não apresentarem a hidroxila no carbono 3', quando incorporados na cadeia nascente de DNA, não permitem a adição de novos nucleotídeos, visto que se torna impossível para a polimerase catalisar a reação de síntese da ligação fosfodiéster. Como a reação ocorre sob a presença de ambos os tipos de nucleotídeos, em concentrações cuidadosamente planejadas, as fitas nascentes de DNA são interrompidas em diferentes pontos de sua síntese, de acordo com a incorporação ou não dos terminadores. A etapa final consiste numa corrida eletroforética que permite a separação dos produtos de síntese de diferentes tamanhos, para cada tipo de terminador utilizado (T, C, G ou A). A técnica original utilizava uma marcação isotópica de um dos desoxirribonucleotídeos (geralmente o α [^{32}P]dATP). Posteriormente utilizaram-se terminadores marcados radioativamente. Finalmente, passou-se a empregar terminadores conjugados a moléculas fluoróforas, que sob a incidência de um feixe de laser, emitem luz em diferentes comprimentos de onda, correspondentes às cores vermelho, azul, verde e amarelo. Todo o sistema de detecção passou a ser automatizado, permitindo assim aumentar muito a quantidade de dados gerados. Finalmente, a corrida eletroforética que era realizada em géis muito finos de poliacrilamida passou a ser feita por eletroforese capilar empregando-se múltiplos capilares em sistemas robotizados.

Utilizando-se o método de Sanger, centenas de organismos tiveram o seu genoma sequenciado. Entre os marcos do sequenciamento genômico, podemos citar os

seguintes: o sequenciamento genômico do primeiro organismo celular, a bactéria *Haemophilus influenzae* (Fleischmann et al., 1995), do primeiro eucarioto unicelular, a levedura *Saccharomyces cerevisiae* (Goffeau et al., 1996), do primeiro eucarioto multicelular, o helminto *Caenorhabditis elegans* (*C. elegans* Sequencing Consortium, 1998), da mosca *Drosophila* (Adams et al., 2000) e, finalmente, a publicação da versão rascunho (*draft*) do genoma humano (Lander et al., 2001). Em paralelo a esses esforços de sequenciamento, novas tecnologias foram incorporadas aos equipamentos de sequenciamento permitindo a sua automatização, como a marcação fluorescente dos nucleotídeos e sua detecção automática.

Apesar do método de Sanger ter sido amplamente utilizado para o sequenciamento genomas, ele apresenta várias limitações:

- a) é necessária uma grande quantidade inicial de amostra de DNA para realizar o sequenciamento;
- b) a metodologia apresenta várias etapas susceptíveis a erros e vieses biológicos, pois depende da clonagem e amplificação do material que se deseja sequenciar em vetores;
- c) mesmo com as mais modernas máquinas de sequenciamento a capacidade de geração de dados é baixa; e finalmente;
- d) o custo por base sequenciada e infraestrutura necessária para realizar um projeto de sequenciamento são muito elevados.

Dessa maneira, Metzker (2005), estimou que o tempo e o custo necessários para gerar um rascunho com alta qualidade de um genoma da ordem de bilhões de pares de bases (como o de mamíferos), é respectivamente, de cerca de seis meses e 12 milhões de dólares. As limitações anteriormente descritas, associadas com as novas tecnologias emergentes de sequenciamento, levaram à produção de novas máquinas capazes de gerar uma quantidade de dados dramaticamente superior e a um custo muito menos elevado, se comparadas à metodologia tradicional de Sanger (Pop e Salzberg, 2008). Esses novos equipamentos, denominados de sequenciadores de nova geração ou sequenciadores maciçamente paralelos, estão revolucionando o campo do sequenciamento genômico, permitindo explorar e responder questões biológicas com uma velocidade e precisão nunca antes atingida.

1.1.2 As novas plataformas de sequenciamento de DNA

Cada uma das novas plataformas de sequenciamento apresenta a sua própria metodologia de sequenciamento, porém, uma característica em comum entre todas essas novas máquinas, é a capacidade de gerar uma grande quantidade de dados em um tempo muito curto. O fluxo de trabalho para se produzir as bibliotecas de DNA para o sequenciamento pode ser dividida em três principais etapas: na primeira os fragmentos de DNA genômico sofrem um processo de reparo em suas pontas e adaptadores específicos são ligados em ambas às extremidades desses fragmentos, em seguida os adaptadores específicos ligados aos fragmentos de DNA são hibridizados com adaptadores complementares fixos a uma superfície (lâmina de sílica ou uma microesfera), e finalmente uma PCR ocorre. As reações de sequenciamento ocorrem nas superfícies anteriormente citadas e as bases nucleotídicas são detectadas utilizando um sensor que captura a emissão de luz liberada pelo processo de pirosequenciamento, ou por uma emissão fluorescente (Mardis, 2006).

A plataforma de sequenciamento 454 (Roche, Connecticut, Estados Unidos da América) foi a primeira plataforma de sequenciamento maciçamente paralela a ser introduzida no mercado em meados de 2004. Esse sequenciador trabalha com o princípio do pirosequenciamento introduzido por Ronaghi et al. (1996), onde uma cascata de reações enzimáticas iniciadas pela liberação de um pirofosfato durante reação de incorporação de um nucleotídeo na cadeia nascente pela DNA polimerase, gera ao final do processo um sinal de luz que é capturado por uma câmera que faz parte do sequenciador (Margulies et al., 2005). Em 2006, a plataforma de sequenciamento Illumina (Illumina, San Diego, Estados Unidos da América) foi lançada no mercado apresentando a capacidade de gerar simultaneamente milhões de leituras curtas de DNA, permitindo o sequenciamento de um bilhão de bases nucleotídicas em uma única corrida. Recentemente, foi lançada a primeira plataforma de sequenciamento que realiza a detecção de bases sem se apoiar em emissões de luz geradas durante o processo de identificação das bases. Essa detecção “não óptica” das sequências de DNA é feita pela plataforma Ion Torrent PGM (Life Technologies, São Francisco, Estados Unidos da América), e se apoia na modificação do pH que ocorre no chip de sílica no qual acontece todo o processo de sequenciamento (Rothberg et al., 2011).

1.2 Aplicações do sequenciamento de DNA na biologia

Com o desenvolvimento das autodenominadas novas plataformas de sequenciamento, várias aplicações foram aprimoradas ou originalmente desenvolvidas no campo do sequenciamento de DNA como:

- a) sequenciamento de novos genomas;
- b) identificação de nucleotídeos de polimorfismo único (SNPs), inserções e deleções em projetos de ressequenciamento;
- c) caracterização do transcriptoma de genomas inteiros (RNA-seq);
- d) identificação de regiões de interação entre DNA-proteína (ChIP-seq);
- e) estudos metagenômicos;
- f) descobertas de RNAs não codificantes (nc-RNAs);
- g) paleogenômica (estudo de amostras de DNA fóssil).

Graças ao Projeto Genoma Humano e ao esforço de centenas de grupos de pesquisa, o primeiro rascunho do genoma humano foi publicado (Lander et al., 2001), representando um ponto de virada para o entendimento da genética humana e o ponto de partida para os estudos genômicos humano em larga escala (Gonzaga-Jauregui et al., 2012). Em meados do ano de 2007, foi criado o Projeto Microbioma Humano (HMP), que consiste em uma extensão lógica e conceitual do Projeto Genoma Humano (Turnbaugh et al., 2007). Os microrganismos que vivem em associação com os humanos nos diferentes ambientes (nasal, oral, pelos, gastrointestinal e urogenital), e que formam a microbiota, superam em número as próprias células somáticas e germinativas humanas em várias ordens de grandeza. Se os humanos são compostos por células humanas que convivem diretamente com células microbianas, cria-se uma situação na qual os genes humanos e microbianos definem as nossas características metabólicas. Assim, a compreensão de nossa diversidade fisiológica, passa pelo conhecimento da distribuição desses microrganismos e seu papel biológico. Esse tipo de estudo envolve não apenas um único genoma de um organismo conhecido, mas múltiplos genomas de organismos não necessariamente identificados ou isolados, constituindo-se no que se denomina metagenômica. Essa ciência tem se expandido e transformado nossa visão sobre o mundo microbiano (Wooley et al., 2010). De fato, cerca de 99% dos taxa não são cultiváveis em laboratório (Amann et al., 1995). A metagenômica é uma ferramenta que permite transcender essas limitações técnicas,

além de nos permitir estender nossos conhecimentos muito além do escopo laboratorial. Com as novas plataformas de sequenciamento, os estudos metagenômicos foram se tornando economicamente mais viáveis, e as comunidades microbianas humanas passaram a ser mais estudadas.

Ao lado dos estudos envolvendo bactérias, a metagenômica viral tem permitido vislumbrar melhor a complexidade de vários ecossistemas, não somente pela descoberta de novos vírus, como também pelo aumento do nosso conhecimento sobre a diversidade viral e a ecologia dos sistemas. Está bem estabelecido que os vírus, tanto de RNA quanto de DNA, estão presentes em ambientes aquáticos e, portanto, é importante conhecer a biomassa viral total do ambiente e seus impactos na estruturação das comunidades (Djikeng et al., 2009). Aproximadamente 75% das sequências obtidas em estudos metagenômicos virais não apresentam correspondência com quaisquer genes da base de dados não redundante (nr) do GenBank, sugerindo que a maioria da diversidade viral permanece ainda não caracterizada. Esse fato se reveste da maior importância se atentarmos que as comunidades virais são as mais diversas entre todas as entidades biológicas estudadas até o momento (Breitbart e Rohwer, 2005). Os desafios técnicos relacionados ao cultivo de novos sistemas vírus-hospedeiros têm representado um impedimento primário para a descoberta de novos vírus. Se na abordagem tradicional é requerido que células do hospedeiro e seus vírus sejam passíveis de cultivo, a metagenômica, viral surgiu para contornar essas dificuldades, acessando diretamente o vírus e seu hospedeiro em seu local de origem, sem requerer um cultivo laboratorial durante o processo (Schoenfeld et al., 2009). Além do mais, numerosas doenças crônicas e agudas com etiologia desconhecida podem ser causadas por vírus não identificados, e a busca sistemática por vírus desconhecidos é uma tarefa científica de máxima importância. Esse aspecto, ainda que tecnicamente muito desafiador, é um ponto crucial para estudos de doenças infecciosas (Allander et al., 2005).

Um caso emblemático da aplicação de técnicas de metagenômica para a elucidação da etiologia de uma virose foi reportada por Palacios et al. (2008). Três mulheres na Austrália receberam órgãos viscerais (fígado e dois rins) transplantados de um único homem doador, e faleceram de uma doença febril, em quatro a seis semanas após o transplante. O doador havia falecido de hemorragia cerebral três meses após ter visitado a área rural da antiga Iugoslávia. Todos os testes realizados foram incapazes de apontar a etiologia da doença. Esses testes incluíram a cultura, PCR e ensaios sorológicos, bem como uma análise usando um arranjo de oligonucleotídeos contra uma

variedade de agentes infecciosos. O isolamento do RNA dos órgãos transplantados seguido de sequenciamento em larga escala com um equipamento 454 permitiu identificar sequências de arenavirus do Velho Mundo, associadas a vírus de coriomeningite linfocítica. Testes sorológicos complementares comprovaram a infecção das pacientes receptoras. Ainda, nessa mesma vertente, Chen et al. (2011) utilizando o sequenciamento de nova geração (Illumina), conseguiram associar a morte (devido a uma pneumonia aguda) de vários macacos do Novo Mundo (*Cebus cupreus*), a um adenovírus nunca antes identificado. Esses trabalhos demonstram de forma inequívoca o potencial uso de técnicas de metagenômica e sequenciamento em larga escala no diagnóstico etiológico de surtos de doenças.

Concomitantemente aos estudos acima descritos, vários outros foram realizados focando-se em loci genéticos únicos, como o DNA mitocondrial, e até em genomas completos de espécies e populações extintas (Stoneking e Krause, 2011). Esses estudos são importantes, pois podem ajudar a responder perguntas de longas datas relacionadas à evolução molecular dos organismos, as bases moleculares da especiação e evolução dos genes (Poinar et al., 2006). Nos últimos anos, vários trabalhos relacionados ao DNA fóssil têm sido publicados, entre eles podemos citar o sequenciamento do DNA nuclear do urso da caverna (Noonan et al., 2005), do mamute-lanoso (Miller et al., 2008; Poinar, et al., 2006), o sequenciamento e produção do rascunho do genoma do neandertal (Green et al., 2010), e de uma cepa do organismo *Yersinia pestis*, agente causador da peste negra, de cerca de 600 anos atrás (Bos et al., 2011).

1.3 Montagem de sequências

Como mencionado nas seções 1.1 e 1.2, o sequenciamento de DNA se tornou uma técnica padrão para investigar o genoma dos organismos. Contudo, as atuais metodologias de sequenciamento genômico conseguem apenas sequenciar fragmentos de dezenas ou centenas de pares de bases. Assim, fragmentos maiores de DNA precisam ser subdivididos em fragmentos menores, que por sua vez são então sequenciados. As leituras de DNA originadas através desse processo são então montadas com a ajuda de programas computacionais especializados, denominados montadores de DNA (*DNA assemblers*). A maioria dos programas convencionais de montagem como o Phrap¹ e o

¹ Phil Green, não publicado.

CAP3 (Huang e Madan, 1999) utiliza a abordagem sobreposição-leiaute-consenso para a montagem de sequências. Porém, com a introdução dos sequenciadores de nova geração, a quantidade de dados gerados cresceu em várias ordens de grandeza, ao passo que o tamanho das leituras diminuiu significativamente, dificultando a montagem de sequências e as análises bioinformáticas subsequente (Kunin et al., 2008). Dessa maneira, a montagem das leituras de DNA foi complicada pela introdução de novos tipos de erros e vieses, e pelo tamanho pequeno e grande quantidade de leituras produzidas (Pop, 2009). Esse aspecto incentivou e levou ao desenvolvimento de novos montadores destinados a lidar com esses novos tipos de dados.

1.3.1 Abordagem da sobreposição-leiaute-consenso

A abordagem de sobreposição-leiaute-consenso é tipicamente utilizada para realizar a montagem de leituras longas do tipo Sanger. Os montadores que utilizam essa abordagem realizam a montagem dos genomas seguindo três principais etapas. A primeira consiste em uma comparação das leituras todas-contra-todas, de forma que ao final desse processo todas as sobreposições existentes em cada uma dessas leituras são encontradas. As sobreposições encontradas são sensíveis ao tamanho e ao percentual mínimo de identidade permitido entre as sobreposições, que são definidos pelo usuário. O programa CAP3, por exemplo, antes de iniciar as comparações entre as leituras remove regiões com baixa qualidade nas extremidades das leituras de DNA, eliminando falsas sobreposições e simplificando o processo de montagem. As informações obtidas através das comparações são então utilizadas para construir um grafo de sobreposições onde as arestas correspondem às sobreposições encontradas entre as leituras, e os nós são as próprias leituras de DNA. Na segunda etapa, a fase do leiaute, o grafo de sobreposições é simplificado pela eliminação da redundância entre as leituras e um esquema consistindo da posição relativa das leituras é estabelecido. A montagem dos segmentos do genoma é feita pela determinação do caminho no grafo onde cada vértice é visitado apenas uma única vez, sendo esse procedimento é conhecido como caminho Hamiltoniano. Encontrar esse circuito é um problema NP-completo, o que significa que é efetivamente impossível calcular uma solução ótima (Paszkiwicz e Studholme, 2010). Por isso, a abordagem de sobreposição-leiaute-consenso não fornece necessariamente a melhor solução para o problema de montagem, mas sim uma aproximação ou uma heurística muito próxima da solução “ótima”. Finalmente, todas as

leituras encontradas ao longo do caminho Hamiltoniano são recrutadas, e o programa constrói um alinhamento múltiplo de sequências a partir da qual a sequência consenso é obtida. A abordagem descrita acima é computacionalmente intensa e pode se tornar inviável quando aplicada a conjuntos de dados compostos por milhões de leituras curtas, como as originadas pelas novas plataformas de sequenciamento. A falta de um algoritmo eficiente para a solução desse problema levou ao desenvolvimento de montadores baseados em uma abordagem diferente, a qual utiliza grafos de de Bruijn (ou caminhos Eulerianos).

1.3.2 Abordagem dos grafos de de Bruijn

A abordagem Euleriana para montagem de sequências foi descrita pela primeira vez por Pevzner et al. (2001), se transformando em uma alternativa à abordagem da sobreposição-leiaute-consenso. Esses autores definiram essa nova abordagem de uma maneira muito ilustrativa:

Nosso novo algoritmo [...] é muito diferente da abordagem tradicional (sobreposição-leiaute-consenso) – nós nunca tentamos encontrar a sobreposição entre os pares de fragmentos, e muito menos temos a etapa de sobreposição. Ao invés disso, nós fazemos uma coisa muito contra intuitiva (alguns dirão infantil): nós cortamos os pedaços existentes de um quebra cabeça em pedaços ainda menores de tamanho definido. Embora essa operação pareça infantil e irresponsável, nós fazemos isso de propósito [...] Essa operação transposta à montagem do quebra cabeça do complicado mundo do leiaute para dentro do mundo Euleriano, com algoritmos polinomiais para realizar essa montagem (Pevzner et al., 2001, p. 1).

Os programas que utilizam essa abordagem realizam a montagem dos genomas seguindo algumas etapas principais. A primeira etapa consiste em se extrair as palavras nucleotídicas de tamanho definido (kmer ou câmero) a partir leituras de DNA, e realizar a construção de uma tabela (*hash*) que contém todas as palavras nucleotídicas criadas. A fase de construção do grafo de câmeros (ou de de Bruijn) é realizada rapidamente, apoiando-se nas informações armazenadas na tabela anteriormente descrita, onde os câmeros são representados por nós dentro do grafo, e as sobreposições entre os câmeros são representadas pelas arestas. A sobreposição é sempre definida como o tamanho da palavra nucleotídica menos um (câmero - 1), ou seja, se a palavra nucleotídica tiver

tamanho 25, o grafo é construído com todos os câmeros de tamanho 25 que se sobrepõem uns aos outros por 24 nucleotídeos. Esse procedimento pode consumir grandes quantidades de memória, porém o grafo representa cada câmero apenas uma única vez, não importando quantas vezes esse mesmo câmero apareceu nas leituras (Miller et al., 2010). Dessa maneira, é possível lidar com as grandes quantidades de leituras extremamente redundantes (cobertura elevada) originadas pelas novas plataformas de sequenciamento sem aumentar a complexidade do grafo e o custo computacional. Após a construção dos grafos, esses são simplificados através de uma correção dos erros que são introduzidos durante o processo de sequenciamento. Esses erros de sequenciamento são identificados através de mudanças na topologia dos grafos, que são facilmente localizadas pelos programas. Alternativamente, outra maneira de se localizar erros de sequenciamento é através da frequência dos câmeros extraídos das leituras de DNA. As palavras nucleotídicas “verdadeiras”, aquelas livres de erros de sequenciamento, apresentam teoricamente uma frequência muito mais elevada dentro da amostra sequenciada do que aquelas que contêm erros. Assim, utilizando um limiar delimitado pelo usuário, os programas conseguem remover palavras nucleotídicas com baixa frequência e, conseqüentemente, os erros de sequenciamento. Finalmente, após os grafos serem construídos, simplificados e eliminados de erros, os montadores encontram o melhor caminho Euleriano, visitando cada aresta apenas uma única vez e, a partir desse caminho, a sequência reconstruída é originada.

A maior desvantagem da abordagem dos grafos de de Bruijn é a perda da informação causada pela decomposição das leituras em palavras nucleotídicas. Comparados aos montadores convencionais, onde uma leitura representa um único nó em um grafo de sobreposição, os montadores que utilizam os grafos de de Bruijn criam múltiplos nós para uma mesma leitura, e esses nós podem não formar um caminho linear entre uma leitura e outra (Schatz et al., 2010).

1.4 GenSeed e a montagem progressiva dirigida por semente

Como visto nas seções anteriores (1.3.1 e 1.3.2), a abordagem computacional para montagem de genomas é dispendiosa em termos de memória alocada e processamento computacional, podendo se tornar impraticável para a maioria dos grupos de pesquisa se as bases de dados utilizadas forem muito grandes (problema que é potencializado com as novas plataformas de sequenciamento). Além disso, muitas das

questões biológicas estão dirigidas a alvos moleculares simples, e a reconstrução de sequências específicas é uma tarefa muito comum. Existem duas maneiras possíveis de se identificar uma sequência alvo ou região específica em projetos de sequenciamento:

- a) identificá-la com programas que realizam uma busca de similaridade caso as leituras sequenciadas já estejam montadas;
- b) realizar a montagem global dessas leituras para somente depois identificar o alvo específico que se quer encontrar.

Esses problemas e limitações motivaram o nosso grupo a criar um método para a montagem alvo-específica de sequências de DNA. O método, denominado montagem progressiva dirigida por semente, consiste na reconstrução de sequências específicas a partir de dados não montados, partindo-se de sequências iniciais curtas de DNA ou proteína, as quais denominamos de sementes.

O algoritmo do programa *GenSeed*, desenvolvido pelo nosso grupo (Sobreira e Gruber, 2008), consiste em um processo iterativo de seleção e montagem de sequências que apresentam sobreposições com a sequência semente. O processo inicia-se com uma busca de similaridade da semente contra uma base de dados de leituras não montadas. As leituras que apresentam sobreposições com a semente e que, portanto, estendem as extremidades dessa semente, são selecionadas. Essas leituras são então montadas com a própria semente, resultando numa sequência consenso mais longa. As extremidades dessa sequência consenso são então utilizadas como sementes secundárias na próxima rodada de extensão, e o processo é iterado certo número de vezes, resultando numa extensão progressiva da sequência semente original. O programa aceita tanto sequências nucleotídicas como proteicas como sementes (Figura 1A). Em termos de base de dados, qualquer conjunto de sequências em formato FASTA pode ser utilizado, sem necessidade de se fornecer os arquivos de eletroferogramas, ou dados de qualidade. No caso de a base de dados não estar formatada, o *GenSeed* invoca o programa *formatdb*, do pacote *BLAST* (Altschul et al., 1997), gerando automaticamente os arquivos de índices (Figura 1B). O programa automaticamente detecta a natureza da sequência semente. Se a semente for nucleotídica, ela é submetida a uma busca de similaridade por *BLASTN* contra a base de dados (Figura 1C). No caso de ser proteica, o programa *TBLASTN* é executado. Todas as leituras positivas na busca são recuperadas com o programa *fastacmd*, também do pacote *BLAST*. Caso as sequências recrutadas

necessitem serem mascaradas contra vetores, o programa `Cross_match`² é executado. As sequências são então montadas com o programa `CAP3`, os arquivos de saída da montagem são inspecionados, e somente as sequências que contém a sequência semente original são armazenadas e continuam no processo de montagem progressiva. O processo iterativo de montagem prossegue até sua exaustão, detectada pela ausência de novas leituras que estendam a sequência, ou se não houver aumento do comprimento da sequência consenso por duas rodadas consecutivas. Alternativamente, o usuário pode definir dois critérios adicionais de término: o número máximo de rodadas e o comprimento máximo da sequência consenso. Após o término do processo de montagem progressiva, o `GenSeed` executa uma etapa de montagem final na qual todas as leituras recrutadas ao longo de todos os ciclos são utilizadas, gerando assim uma sequência consenso final (Figura 1D). O programa armazena arquivos de todas as etapas intermediárias, incluindo as saídas de `BLAST`, montagem do `CAP3`, além de relatórios gráficos em formato `HTML`, que demonstram a reconstrução progressiva da sequência.

² Phil Green, não publicado.

Figura 1 – Fluxo de trabalho do algoritmo do programa GenSeed

O GenSeed detecta automaticamente o tipo de sequência biológica (A) utilizada como semente e (B) formata, se necessário, o arquivo FASTA da base de dados para gerar os arquivos de índices do BLAST. A sequência semente é introduzida num processo iterativo de montagem dirigida por semente (C), o qual resumidamente consiste numa busca de similaridade, recuperação de sequências e montagem. O processo de caminhada é concluído quando um critério de término é preenchido (ver texto para detalhes). Todas as leituras selecionadas durante os ciclos de extensão são então montadas e a sequência reconstruída é armazenada (D).

Fonte: Sobreira e Gruber (2008)

Dois outros programas descritos na literatura implementam conceitos similares aos do GenSeed: Genotrace (Berezikov et al., 2002) e Tracembler (Dong et al., 2007), mas apresentam um conjunto de limitações que impedem seu uso no escopo de aplicações que o GenSeed apresenta. O programa GenSeed quando comparado aos dois outros programas apresenta várias vantagens como, a capacidade de realizar reconstruções uni ou bidirecionais, checagem iterativa das sequências reconstruídas, visando manter no processo de reconstrução progressiva apenas as sequências que contem a semente original, e finalmente, a flexibilidade de poder ser aplicado em qualquer bases de leituras em formato FASTA.

O método desenvolvido e implementado no GenSeed se mostrou bastante eficiente para a reconstrução de sequências usando sementes de bases de dados derivadas do mesmo organismo. Contudo, a reconstrução de sequências de organismos filogeneticamente mais distantes pode se tornar problemática. Modelos probabilísticos têm sido utilizados em perfis de alinhamentos múltiplos de sequências para estabelecer relações entre proteínas de mesma função, ainda que pertencentes a organismos distantes. Assim, a base Pfam (Punta et al., 2012) consiste numa grande coleção de HMMs de perfis de milhares de diferentes famílias proteicas, onde cada família consiste em um conjunto de proteínas ortólogas. A similaridade entre as sequências protéicas e os HMMs de perfis é detectada utilizando programas presentes no pacote HMMER3 (Eddy, 2009).

1.5 Modelos Ocultos de Markov (HMMs) e HMMs de perfis

Os métodos tradicionais de alinhamento pareado (como o BLAST) assumem que todas as posições dos aminoácidos representam caracteres independentes uns em relação aos outros, e não consideram as informações relativas à posição-específica de cada resíduo na sequência. Um alinhamento de múltiplas sequências proteicas, por outro lado, pode ser útil para se deduzir quais resíduos são mais conservados em cada posição, bem como para identificar as regiões de inserções e deleções mais frequentes. Dessa maneira, definir um perfil proteico a partir de um alinhamento de múltiplas sequências é uma maneira mais intuitiva, eficaz e flexível de se caracterizar proteínas ou famílias de proteínas, do que o simples alinhamento par a par (Eddy, 1996).

Krogh et al. (1994) relataram uma arquitetura baseada em modelos ocultos de Markov (HMM) muito adequada para representar perfis de alinhamento de múltiplas sequências. Nessa arquitetura, atribui-se um estado de correspondência ou *match* para a maioria das colunas (caracteres) de um alinhamento de múltiplas sequências. Esses estados de correspondência modelam a distribuição de um resíduo de aminoácido para cada uma das colunas, refletindo a probabilidade de se observar um dado resíduo específico nessa posição. Cada um dos estados de correspondência é acompanhado por dois outros estados, os de inserção e deleção, permitindo, assim, modelar a inserção e a deleção de um ou mais resíduos de aminoácidos entre uma coluna e as demais (Eddy, 1998). Assim, um HMM de perfil é um modelo probabilístico derivado de um alinhamento de múltiplas sequências que captura informações posição-específicas de cada uma das colunas, levando em consideração que colunas são mais conservadas e quais os resíduos mais prováveis em cada uma delas.

Uma das vantagens de se utilizar HMMs é a base probabilística formal associada ao método, que serve como guia para definir os parâmetros durante a construção dos HMMs de perfis, além de se embasar numa teoria consistente para o ajuste dos escores de inserção e deleção. Além disso, o método é robusto e automatizável, permitindo a criação de bases de dados de contendo centenas de HMMs de perfis, e a sua aplicação em projetos de grande escala como a análise de genomas inteiros (Eddy, 2010).

1.6 Um novo programa para reconstrução de sequências: GenSeed-HMM

A implementação do uso de HMMs como sementes dentro da utilização do programa GenSeed poderia incrementar muito o escopo de aplicações do programa, especialmente no que diz respeito à reconstrução de sequências filogeneticamente mais distantes em relação às sementes. Diante do que foi exposto nas seções anteriores, decidimos aumentar a abrangência e a aplicação do programa original GenSeed, incorporando o uso de sementes de HMMs de perfis e a capacidade de realizar reconstruções sitio-dirigidas em bases de dados originados pelas novas plataformas de sequenciamento. O desenvolvimento desse novo programa, denominado GenSeed-HMM, é um dos objetivos do presente trabalho.

2 OBJETIVOS

- Desenvolver uma ferramenta genérica, denominada GenSeed-HMM, para a reconstrução de sequências utilizando HMMs de perfis como sementes e leituras geradas por plataformas de sequenciamento de nova geração;
- validar o GenSeed-HMM utilizando-se conjuntos de dados de organismos procariotos, eucariotos e amostras metagenômicas.

3 MATERIAL E MÉTODOS

3.1 Algoritmo do GenSeed-HMM e implementação

3.1.1 Sementes e bases de dados

A Figura 2 apresenta o fluxo de trabalho da nova versão do programa GenSeed, denominada a partir desse momento de GenSeed-HMM. A capacidade de se utilizar sementes proteicas e nucleotídicas foi preservada no GenSeed-HMM (Figura 2A). Caso a semente seja identificada como sendo proteica ou nucleotídica, sua introdução no processo iterativo de montagem é similar àquela do algoritmo original descrito por Sobreira e Gruber (2008). Em adição ao fluxo de trabalho original do GenSeed (Figura 1), foi implementada uma rotina de detecção de HMMs de perfis, os quais são agora processados e utilizados como sementes para o recrutamento de leituras destinadas ao processo de montagem progressiva (Figura 2A). Em relação à base de dados de leituras, qualquer conjunto de múltiplas sequências nucleotídicas em formato FASTA pode ser utilizado dentro do GenSeed-HMM (Figura 2B), seja esse conjunto composto por sequências curtas (derivadas das plataformas Illumina) ou longas (plataforma 454 e sequenciamento tipo Sanger).

A versão original do GenSeed utiliza o pacote BLAST (Altschul et al., 1997) para realizar as buscas de similaridade a partir de sementes nucleotídicas e proteicas. O GenSeed-HMM utiliza o pacote BLAST+ (Camacho et al., 2009), o qual apresenta muito maior versatilidade de formatos de saída, bem como um melhor desempenho de processamento e memória utilizada. O GenSeed-HMM incorpora em seu fluxo de trabalho uma rotina de checagem para verificar se a base de dados está formatada para o BLAST+, e caso não esteja, o programa `makeblastdb` do pacote BLAST+ é invocado (Figura 2B). No caso de a semente ser um HMM de perfil, o GenSeed-HMM executa o `transeq`, um aplicativo em C do pacote EMBOSS (EMBOSS Package, 2012), para realizar a tradução conceitual das sequências nucleotídicas da base de dados nas seis fases de leitura possíveis. Esse procedimento cria automaticamente outra base de dados, de sequências proteicas traduzidas (Figura 2B). O HMM de perfil é então submetido a uma busca de similaridade contra a base de dados proteica com o programa `hmmsearch`, do pacote HMMER3 (Eddy, 2009). As leituras positivas são então recuperadas a partir da base de dados nucleotídica usando-se o `blastdbcmd`, outro programa do pacote BLAST+, e destinadas ao montador escolhido pelo usuário (Figura

2C). Em adição ao uso do montador CAP3 (Huang e Madan, 1999), já empregado pela versão original do GenSeed, implementamos o uso dos montadores ABySS (Simpson et al., 2009), Velvet (Zerbino e Birney, 2008), SOAPdenovo (Li et al., 2010) e Newbler (Margulies et al., 2005).

3.1.2 A montagem progressiva e integração com montadores de DNA

As sequências nucleotídicas recuperadas com o programa `blastdbcmd` (vide seção anterior) são introduzidas no processo de montagem progressiva (Figura 2C). Foram desenvolvidas rotinas em Perl que encapsulam os cinco programas de montagem listados na seção anterior, permitindo sua utilização no fluxo de trabalho do programa GenSeed-HMM de maneira modular. As sequências positivas recrutadas são encaminhadas para o montador definido pelo usuário em parâmetro específico do programa, dando continuidade ao processo de montagem progressiva.

Em algumas ocasiões, especialmente quando se utiliza bases de dados de leituras curtas e os montadores ABySS, Velvet e SOAPdenovo, o processo de montagem progressiva pode não se iniciar pelo fato do montador não conseguir realizar a montagem a partir das primeiras leituras recrutadas. Esse tipo de problema em geral se deve à ocorrência de erros de sequenciamento das leituras, em especial em suas pontas. Alternativamente, parâmetros muito rigorosos utilizados pelo montador podem impedir a junção das leituras em um *contig*. Para otimizar o processo e reduzir essas ocorrências, o GenSeed-HMM incorpora algumas rotinas específicas. Assim, caso seja detectado que não houve sucesso na iniciação (nenhuma sequência reconstruída), o tamanho do câmero utilizado por esses montadores é reduzido em duas unidades e uma nova montagem é tentada. Persistindo o insucesso, esse procedimento é repetido por mais três vezes, com reduções subsequentes do tamanho do câmero até um máximo de oito unidades.

Caso o procedimento acima não seja bem sucedido, o programa segue uma nova rotina que consiste em aparar a extremidade 3' terminal das leituras positivas em duas em duas bases. A montagem é tentada novamente e, caso necessário, esse procedimento é repetido mais três vezes, até o um máximo de 8 pb de corte. O racional dessa abordagem é eliminar possíveis erros de sequenciamento, os quais são mais frequentes nas pontas 3' das leituras, e que podem impedir a sobreposição adequada com outras

leituras, prejudicando o processo de montagem. Para compensar a redução do comprimento das leituras, faz-se concomitantemente uma redução equivalente do tamanho do câmero utilizado pelo respectivo montador. Finalmente, se todas as tentativas anteriores fracassarem, uma última estratégia é adotada. Com base em resultados de várias análises, verificamos que o `ABYSS` é o montador menos suscetível a problemas de iniciação. Assim, quando o montador escolhido pelo usuário tiver sido o `Velvet` ou `SOAPdenovo`, e uma falha de iniciação for detectada, mesmo com as rotinas descritas acima, o `GenSeed-HMM` realiza apenas a etapa de iniciação (primeira rodada de montagem) utilizando o montador `ABYSS`. Caso haja sucesso, o `GenSeed-HMM` volta a usar o montador escolhido pelo usuário nos ciclos subsequentes de montagem progressiva até a exaustão do processo.

Durante a primeira rodada de processamento, caso a semente seja um HMM de perfil, as sequências reconstruídas são verificadas quanto à presença do motivo proteico representado no HMM de perfil (etapa não mostrada na Figura 2). Para isso, todos os contigs resultantes da montagem são traduzidos conceitualmente utilizando-se o `transeq` nas seis fases de leitura, e comparados contra o HMM de perfil com o programa `hmmsearch`. Apenas as sequências que contém o motivo proteico prosseguem no processo de montagem progressiva. No caso de sementes proteicas ou nucleotídicas, realiza-se uma busca de similaridade das sequências montadas contra a sequência semente, utilizando-se respectivamente os programas `tblastn` ou `blastn`. Para economizar recursos computacionais, o `GenSeed-HMM`, semelhantemente ao algoritmo original, não utiliza toda a sequência montada como semente para a próxima rodada. Ao invés disso, o usuário pode definir o tamanho das extremidades da sequência que irão ser empregadas como sementes nas rodadas subsequentes. No caso dos montadores de leituras longas, `CAP3` e `Newbler`, que geram arquivos de montagem com informações de qualidade base a base, o `GenSeed-HMM` tem uma rotina que permite selecionar como semente o trecho de sequência que apresenta um padrão mínimo de qualidade das bases. Essa rotina é baseada no uso de uma janela deslizante na qual se analisa a porcentagem de bases que apresentam qualidade inferior à qualidade mínima especificada pelo usuário. É importante ressaltar que as qualidades discutidas acima não são aquelas definidas classicamente pelo programa `Phred` (Ewing e Green, 1998; Ewing et al., 1998). Toda a vez que o `Newbler` ou o `CAP3` utilizam sequências em formato FASTA sem valores de qualidade, esses montadores atribuem um valor padrão

para todas as bases. Entretanto, durante o processo de montagem, a redundância de algumas bases pode implicar em uma alta confiança, e esses montadores atribuem um valor maior para essas bases. Alternativamente, discrepâncias entre as bases presentes em uma mesma posição, porém em leituras diferentes, podem diminuir o valor inicial atribuído a essas bases. Assim, os valores atribuídos às bases nas etapas de montagens presentes no GenSeed-HMM apenas representam o nível de redundância e discrepância entre as leituras que participam desse processo. No caso dos montadores montadores de sequências curtas (ABYSS, Velvet e SOAPdenovo), as sequências semente são simplesmente extraídas das sequências reconstruídas. Terminado o primeiro ciclo de montagem, todas as sementes utilizadas no GenSeed-HMM são sempre nucleotídicas, e os demais ciclos de montagem são executados com todas as leituras positivas recrutadas na rodada, mais a(s) sequência(s) reconstruída(s) inteira(s) geradas na rodada anterior.

Todas as sequências montadas são submetidas a buscas de similaridade realizadas com o programa BLAST+ contra as sequências da rodada anterior (Figura 2C). Esse passo permite checar se a sequência atual contém a sequência da rodada anterior, evitando assim a reconstrução de sequências não relacionadas com a semente original. Além disso, detecta-se se houve incremento no tamanho da sequência montada em relação à da rodada anterior. Em caso negativo, um procedimento adicional é executado sempre que não houver a extensão da sequência entre duas rodadas consecutivas (Figura 2C). A sequência que não tenha sido estendida é submetida a uma rotina de aparamento de suas extremidades e reintrodução no ciclo de montagem progressiva. Se após a execução do processo mencionado acima a sequência montada voltar a aumentar de comprimento, ela prossegue no processo recursivo de reconstrução, caso contrário, o processo de montagem para esta sequência é concluído. Essa rotina visa reduzir as ocorrências de paradas prematuras de progressão da montagem causadas por erros pontuais de sequenciamento de algumas leituras, os quais são mais frequentes nas extremidades dessas leituras.

A iteração é concluída com a exaustão do processo, detectada pela ausência de leituras na base de dados que possam levar à extensão da sequência montada, ou se o tamanho da sequência montada não for incrementado em duas rodadas sucessivas. Alternativamente, o usuário pode definir dois critérios adicionais para a finalização da montagem progressiva: número máximo de rodadas e tamanho máximo em pares de bases da sequência reconstruída. Todas as sequências montadas que terminaram o

processo de montagem progressiva antes da última rodada de execução são armazenadas em um arquivo final de sequências.

3.1.3 Processo final de execução

Uma vez concluído o processo de montagem progressiva (Figura 2D), o GenSeed-HMM realiza um passo final de checagem. Caso a semente original seja identificada como um HMM de perfil, o arquivo final de sequências é traduzido com o `transeq` e faz-se uma busca com o `hmmsearch` do HMM de perfil contra esse mesmo arquivo. Com isso, garante-se que a(s) sequência(s) final(ais) montadas apresentem o motivo proteico detectado pelo HMM de perfil. Porém, se as sementes originais forem nucleotídicas ou proteicas, realiza-se uma busca de similaridade utilizando-se respectivamente os programas `blastn` ou `tblastn`, garantindo-se assim que a semente original esteja presente nas sequências finais reconstruídas.

Finalmente, se o usuário tiver definido o parâmetro `-mapping yes`, é invocado o programa `Bowtie-2` (Langmead e Salzberg, 2012) para alinhar todas as leituras de DNA recrutadas ao longo dos ciclos de montagem contra a(s) sequência(s) montadas. O produto desse processo é um arquivo de mapeamento em formato SAM (Sequence alignment/MAP - Li et al., 2009). O arquivo SAM, juntamente com o arquivo final de sequências positivas produzido pelo GenSeed-HMM, podem ser visualizados graficamente utilizando-se o programa `Tablet` (Milne et al., 2010). Ao final do processo, o usuário pode analisar todas as etapas ocorridas no processo de montagem progressiva do GenSeed-HMM, através de um arquivo `log`. O usuário também pode definir se todos os resultados intermediários gerados durante o processo de montagem serão armazenados, ou simplesmente as sequências finais reconstruídas.

Figura 2 - Fluxo de trabalho do programa GenSeed-HMM.

O GenSeed-HMM detecta qual o tipo de semente fornecida pelo usuário (A) e a seguir realiza a verificação da base de dados (B). Caso necessário, a base de dados é formatada para poder ser utilizada nas buscas de similaridade pelo BLAST+. Caso a semente seja de um HMM de perfil, a base é traduzida conceitualmente nas seis fases de leitura com o programa *transeq* e uma nova base proteica é criada. Realiza-se uma busca de similaridade entre a semente e a base de dados, e as leituras positivas são introduzidas no processo iterativo de montagem (C). O processo é concluído quando algum critério de término é atingido (ver texto). As sequências finais reconstruídas são checadas quanto à presença da semente original e, finalmente, um arquivo de mapeamento em formato SAM e os demais resultados finais são armazenados (D).

Fonte: Oliveira (2012)

3.2 Implementação e requerimentos

O programa GenSeed-HMM, desenvolvido nesse trabalho, foi inteiramente implementado na linguagem Perl. Para seu pleno funcionamento, o programa utiliza um conjunto de programas de terceiros, conforme listado na Tabela 1.

Tabela 1 – Programas de terceiros utilizados pelo GenSeed-HMM.

Programa(s)	Pacote	Função	Requerimento
blastn	BLAST+	Buscas de similaridade com sequências de DNA	Obrigatório
tblastn	BLAST+	Buscas de similaridade com sequências protéicas	Opcional (requerido apenas para sementes protéicas)
hmmsearch	HMMER3	Buscas de similaridade com HMMs de perfis	Opcional (requerido apenas para sementes de HMMs de perfis)
makeblastdb	BLAST+	Formatação da base de dados	Obrigatório
blastdbcmd	BLAST+	Recuperação de sequências da base de dados	Obrigatório
SOAPdenovo-127mer/63mer/31mer	SOAPdenovo	Montador de DNA (leituras curtas)*	Opcional (requerido apenas se definido pelo usuário)
abyss	ABYSS	Montador de DNA* (leituras curtas)	Opcional (requerido apenas se definido pelo usuário)
velvethe velvetg	Velvet	Montador de DNA* (leituras curtas)	Opcional (requerido apenas se definido pelo usuário)
runAssembly	Newbler	Montador de DNA* (leituras longas)	Opcional (requerido apenas se definido pelo usuário)
cap3	CAP3	Montador de DNA* (leituras longas)	Opcional (requerido apenas se definido pelo usuário)
transeq	EMBOSS	Tradução de sequências	Opcional (requerido apenas para sementes de HMMs de perfil)
splitter	EMBOSS	Fragmentação de sequências	Opcional (requerido apenas para o uso do Newbler)
Bowtie e bowtie-build	Bowtie	Alinhamento de sequências	Opcional (requerido apenas se o parâmetro <code>-mapping</code> for usado)

*Ao menos um montador de DNA deve estar instalado para bases de dados de leituras curtas ou longas. Recomendamos instalar o ABYSS mesmo quando outro montador for utilizado (veja texto).

Fonte: Oliveira (2012)

O programa GenSeed-HMM roda em qualquer plataforma com sistema operacional UNIX/Linux, incluindo o MacOS X e requer um interpretador Perl instalado. O espaço em disco requerido depende do tamanho das bases de dados utilizadas. Em termos de memória RAM, o requerimento é ditado pelos montadores utilizados, mas como a montagem é feita apenas com as leituras recrutadas pelo GenSeed-HMM ao longo de cada ciclo, quantidade de memória exigida é bastante pequena (menor do que 4 GB para a maioria dos casos). O GenSeed-HMM pode ser executado com uma grande variedade de parâmetros, os quais podem ser diretamente especificados no comando de execução em linha, ou definidos em um arquivo de configuração em formato texto. Para facilitar o uso do GenSeed-HMM, foi construído um programa gráfico de configuração (Seeder – ver seção 3.3), o qual requer a instalação da biblioteca Perl Tk (Tk, 2012). Os programas GenSeed-HMM e Seeder estão disponíveis para download no endereço <http://www.coccidia.icb.usp.br/genseed-hmm/> (Oliveira, 2012). A documentação relativa aos programas também está disponível no mesmo sítio.

3.3 Seeder: o configurador gráfico do GenSeed-HMM

Como a versão atual do programa GenSeed-HMM suporta o uso três tipos de sementes, cinco montadores distintos e duas ferramentas para buscas de similaridade, o processo de criação e arquivos de configuração pode se tornar muito complexo para o usuário iniciante. O programa Seeder (Oliveira, 2012) é uma ferramenta gráfica com interface amigável, escrita em Perl utilizando a biblioteca Tk, (ver Figura 3), que visa auxiliar o usuário na criação dos arquivos de configuração "otimizados", fornecendo assim um ponto de partida para a reconstrução das leituras com o GenSeed-HMM. Dentre as funções do programa Seeder podemos citar:

- a) verificação automática da disponibilidade dos programas que serão utilizados pelo GenSeed-HMM;
- b) auxílio na construção dos arquivos de configuração;
- c) verificação da consistência dos parâmetros informados pelo usuário, e;
definição automática dos melhores parâmetros para uma dada combinação de tipo de semente e base de dados (se definido pelo usuário).

Figura 3 - Captura de tela do programa Seeder.

O usuário pode definir o número de tarefas de processamento (1) e o comprimento médio das leituras da base de dados (2) nos campos de entrada de valores. O menu de opções permite escolher o montador (3), o tipo de semente (4) e se a base de dados é de sequências curtas ou longas (5). Todos os parâmetros são opcionais, e se o usuário escolher as opções *Choose for me* e *Guess for me*, o programa *Seeder* irá definir automaticamente, através de diversas rotinas de checagem, as melhores opções de execução para a combinação de semente e base de dados.

Fonte: Oliveira (2012).

3.4 Validação do programa GenSeed-HMM

Para validar o programa GenSeed-HMM, foram utilizados vários conjuntos de dados derivados das plataformas de sequenciamento 454 e Illumina, e diversas sementes de HMMs de perfis.

3.4.1 Bases de dados

Todas as bases de dados foram obtidas a partir do banco Sequence Read Archive (SRA - National Center for Biotechnology Information, 2012a) exceção da base de dados de *Gardnerella vaginalis* cepa 5-1, que foi gentilmente cedida pelo Dr. José Paulo Alves. As bases de dados obtidas no SRA foram convertidas do seu formato nativo (sra ou sra-lite) para o formato FASTQ utilizando-se o SRA ToolKit (National Center for Biotechnology Information, 2012b). Essas bases então foram convertidas para o formato FASTA utilizando-se o programa `fastq_to_fasta`, presente no pacote FASTX Toolkit (Hannon Lab, 2012). A base de dados de *G. vaginalis* foi convertida do formato nativo de saída SFF (Standard Flowgram Format), padrão da plataforma 454, para o formato FASTA, com o uso do programa `sff_extract_0_2_10.py` (Comav Institute, 2012). A Tabela 2 sumariza as informações relativas a cada uma das bases de dados utilizadas.

Tabela 2 – Bases de dados utilizadas para validação do programa GenSeed-HMM.

Código de acesso	Organismo	Plataforma	Nº de leituras	Tipo de leitura
ERR019412	<i>C. jejuni</i>	Illumina	20.282.728	Curta (54 pb)
SRR001666	<i>E. coli</i>	Illumina	14.095.336	Curta (36 pb)
Não disponível*	<i>G. vaginalis</i> *	454	203.827	Longa (> 200 pb)
SRR063698	<i>D. melanogaster</i>	Illumina	56.124.204	Curta (94 pb)
ERR023042	<i>P. vivax</i>	Illumina	76.507.524	Curta (75 pb)
ERR024711	<i>S. japonicus</i>	454	687.627	Longa (> 500 pb)
SRR04654 a SRR041657	Metagenômica de intestino humano	Illumina	161.745.112	Curta (100 pb)
SRR006598 a SRR006751	Metagenômica de mamute lanoso	454	31.144.387	Longa (> 177 pb)

*Cedida pelo Dr. José Paulo Alves
Fonte: Oliveira (2012)

3.4.2 HMMs de perfis

Foram obtidos a partir da base Pfam, todos os HMMs de perfis que foram utilizados como sementes durante a validação do programa. Os HMMs de perfis utilizados originam-se domínios proteicos únicos quanto de domínios presentes em

proteínas parálogas. A Tabela 3 sumariza as informações relativas a cada um dos HMMs de perfis utilizados.

Tabela 3 – HMMs de perfis utilizados para validação do programa GenSeed-HMM.

Nº de acesso	Descrição
PF03485	Arginil-tRNA sintetase (domínio N terminal)
PF00689	Cátion APTase (domínio C terminal)
PF00137	ATP sintase (subunidade C)
PF00115	Citocromo C e quinol oxidase polipeptídio I
PF00116	Citocromo C oxidase (subunidade C), domínio periplasmático
PF00510	Citocromo C oxidase (subunidade III)
PF00986	DNA girase B subunidade, carboxi-terminal
PF00145	DNA metiltransferase
PF00044	Gliceraldeído-3-fosfato desidrogenase
PF00288	GHMP quinases (domínio N terminal)
PF00534	Glicosiltransferase (grupo I)
PF03738	Glutathionil espermidina sintetase
PF00009	Fatores de elongação ligados ao GTP (subfamília F-Tu/EF-1A)
PF06580	Histidina quinase
PF00012	Proteínas de choque térmico (70 quilodaltons)
PF05538	Proteína principal da membrana externa de <i>Campylobacter</i>
PF00146	NADH desidrogenase
PF09112	N-glicanase (domínio N terminal)
PF02729	Aspartato/Ornitina carbamoiltransferase (domínio de ligação ao carbamoil-P)
PF00162	Fosfoglicerato quinase
PF00224	Piruvato quinase (domínio “barril”)
PF00085	Tioredoxina
PF01131	DNA topoisomerase
PF00905	Proteínas de ligação à penicilina (domínio transpeptidase)

Fonte: Oliveira (2012)

4 RESULTADOS

4.1 Prova de princípio para o desenvolvimento do programa GenSeed-HMM

A partir da ideia de se utilizar HMMs de perfis como sementes para a reconstrução de sequências de DNA, surgiram várias questões fundamentais para a viabilidade do método, como segue:

- a) a semente de HMM teria que ser usada contra uma base de proteínas e não mais de DNA como na versão original do GenSeed. Que programa permitiria traduzir as sequências de DNA em todas as fases de leitura com desempenho suficientemente alto para lidar de forma viável com bases de dados de milhões de leituras?;
- b) que programa poderia ser usado para se fazer a busca de similaridade do HMM de perfil contra a base de proteínas traduzidas?;
- c) sementes de HMMs de perfis, construídas a partir de proteínas de um conjunto de organismos, poderiam ser capazes de permitir a reconstrução de sequências heterólogas, derivadas de um organismo não incluído na construção do perfil?;
- d) qual seria o comprimento mínimo das leituras de DNA para permitir a tradução e detecção por similaridade com um HMM de perfil? Essa pergunta é particularmente crítica em se considerando o uso de bases de leituras curtas como as da plataforma Illumina.

Para responder a todas essas perguntas, iniciamos um conjunto de testes e simulações, procurando gerar uma prova de princípio de que o método proposto era viável e exequível. Para os primeiros testes, foi utilizado o HMM de perfil RVT_1, de transcriptase reversa (PF00078), obtido a partir da base Pfam. Esse HMM foi então utilizado para se buscar motivos proteicos em uma base de dados fragmentada de *Eimeria tenella*, um protozoário parasita de galinha doméstica. Para realizar a busca de similaridade entre o HMM de perfil e as sequências do parasita, decidiu-se utilizar o pacote HMMER3 (Eddy, 2009). Como os programas desse pacote não realizam automaticamente a tradução das leituras de DNA para a comparação contra o HMM de perfil, foi necessária a utilização de um programa para traduzir conceitualmente todas as sequências. Dentre os programas disponíveis para essa tarefa, acabamos escolhendo o transeq (Questão a – EMBOSS Package, 2012). Este programa consegue lidar com bases de dados grandes com ótimo desempenho, permitindo traduzir conjuntos de dados compostos de dezenas de milhões de leituras em poucos minutos, utilizando um

computador PC com processador Intel i7 (Intel, Santa Clara, Estados Unidos da América). Por exemplo, nesse equipamento, a tradução de uma base de dados de câncer mamário humano, contendo cerca de 13 milhões de leituras nas seis possíveis fases de leituras, demorou 9,5 minutos.

Utilizando-se então o programa `hmmsearch` (Questão b) para a busca de similaridade entre o HMM de perfil RVT_1 e a base traduzida de sequências de *E. tenella*, foi possível se identificar um conjunto de nove leituras positivas. Essas nove sequências foram utilizadas como sementes dentro do programa GenSeed original, utilizando-se a base de dados traduzida de *E. tenella*. Obteve-se como resultado a reconstrução de um gene que codifica a enzima transcriptase reversa, gene esse incorporado no genoma nuclear do parasita. Esse primeiro ensaio confirmou, portanto, que o uso de um HMM de perfil, construído a partir de sequências de um conjunto de organismos, poderia ser utilizado para reconstruir sequências de outro organismo não relacionado (Questão c).

Finalmente, foi realizado um teste com o propósito de se avaliar o limite mínimo de comprimento das leituras para sua detecção pelo HMM de perfil. Essa pergunta é particularmente crítica considerando-se que muitas das bases de leituras disponíveis são de sequências curtas, como aquelas geradas pela plataforma Illumina. Para isso, foi desenvolvido um programa que realiza a fragmentação de leituras de DNA em sequências menores, com comprimento e sobreposição definidos pelo usuário. Esse último parâmetro teve como intenção minimizar a chance de que as leituras não fossem reconhecidas pelo HMM de perfil pelo fato do motivo proteico codificado estar presente justamente na quebra entre duas sequências. Foram feitas fragmentações com tamanhos de 100, 75, 50 e 35 pb. Em todos os casos foi possível se identificar os fragmentos positivos utilizando diferentes HMMs de perfis. De fato, pode-se constatar que sequências tão curtas quanto 33 pb (11 aas) puderam ser reconhecidas com valores de escore de alinhamento relativamente confiáveis. Conclui-se, portanto, que a busca em bancos de dados de leituras curtas traduzidas, utilizando-se HMMs de perfis, é viável dentro do contexto da montagem progressiva realizada pelo programa GenSeed (Questão d).

4.2 Reconstruções de loci utilizando HMM de perfis e bases de dados originadas de organismos procariotos

Para se verificar a capacidade de reconstrução locus-específica usando HMMs de perfis, utilizamos inicialmente três bases de dados de organismos procariotos distintos (ver Tabela 1). Duas dessas bases são compostas por leituras curtas originadas da plataforma Illumina (*Campylobacter jejuni* NCTC 11168 - código de acesso ERR019412 e *Escherichia coli* K-12 MG 1655 – código de acesso SRR001666). O terceiro conjunto de dados procariotos, da bactéria *Gardnerella vaginalis* cepa 1-5, é composto por leituras longas originadas pela plataforma 454. A acurácia da reconstrução os genes realizadas pelo GenSeed-HMM foi determinada através de buscas por similaridade contra as sequências de referência disponibilizadas em sítios públicos. Os montadores ABySS, Velvet e SOAPdenovo foram utilizados para a montagem das bases de dados de leituras curtas, e o montador Newbler para a montagem da base de leituras longas. As reconstruções realizadas na base de leituras longas (*G. vaginalis*) foram feitas limitando-se o tamanho da(s) sequência(s) reconstruída(s) (parâmetro – `max_contig_length` do GenSeed-HMM) para um máximo 30.000 pares de bases. As reconstruções realizadas a partir dos conjuntos de leituras curtas foram feitas até a exaustão do processo. Para cada conjunto de dados, foram utilizados dez HMMs de perfis distintos como sementes. Em função dos resultados obtidos, foram incorporadas rotinas de otimização (ver seção 3.1.2) no programa GenSeed-HMM e as análises foram repetidas com a nova implementação.

4.2.1 Montagem progressiva com leituras de *Campylobacter jejuni*

A bactéria gram-negativa *Campylobacter jejuni* é reconhecida como um importante patógeno humano de distribuição mundial, responsável por causar infecções diarreicas causadas geralmente pelo consumo de comida ou água contaminadas (Blaser, 1997). Foram selecionados da base Pfam um total de dez HMMs de perfis (PF00137, PF00986, PF00009, PF00012, PF05538, PF02729, PF00162, PF00224, PF00085 e PF03485), construídos a partir de proteínas codificadas por genes de cópia única e também de famílias de proteínas. Como conjunto de dados, utilizou-se cerca de 20 milhões de leituras da plataforma Illumina (Tabela 1) derivadas de uma cepa virulenta de *C. jejuni* (National Center for Biotechnology Information, 2012c). As reconstruções

de sequências foram testadas com os montadores Velvet, SOAPdenovo e ABySS. Todas as sequências finais reconstruídas por montagem progressiva com o GenSeed-HMM foram comparadas por similaridade (`blastn`) com aquelas obtidas por montagem global com os respectivos montadores, na qual se utilizou o conjunto inteiro de leituras. Finalmente, as sequências reconstruídas por ambos os métodos (montagem progressiva e global) foram comparadas por similaridade contra o genoma de referência da bactéria (código de acesso NC_002163.1), disponível publicamente.

Todas as montagens realizadas com o GenSeed-HMM/ABYSS atingiram um tamanho similar ao obtido por montagem global. Ao contrário, as montagens utilizando os montadores SOAPdenovo e Velvet, mesmo com o uso das rotinas de otimização (descritas na seção 3.1.1), acabaram gerando *contigs* mais curtos. Para melhorar o desempenho das montagens com esses dois montadores, toda a base de dados de *C. jejuni* foi submetida a uma pré-filtragem de qualidade antes de ser inserida no processo de montagem progressiva. Utilizando-se o programa `fastq_quality_trimmer` do pacote FASTX-TOOLKIT (Hannon Lab, 2012), e a base de dados de *C. jejuni* em formato FASTQ, as bases com baixa qualidade presentes na extremidade final de cada uma das leituras foram aparadas (qualidade mínima Phred de 28). As leituras com tamanho inferior a 52 pb, após o processo de corte das pontas, foram excluídas da base de dados. Após essa filtragem, a base de dados FASTQ foi convertida para o formato FASTA, excluindo-se ainda as leituras que apresentassem nucleotídeos desconhecidos (representados pela letra N). Esse processo de filtragem eliminou 2.905.005 leituras (~14% do total). Assim, essa base filtrada e convertida para o formato FASTA foi utilizada no processo progressivo de montagem realizado pelo GenSeed-HMM. Ao final desse processo, em apenas um único caso (HMM de perfil PF00162 – fosfoglicerato quinase), não se atingiu com os montadores Velvet e SOAPdenovo o tamanho de *contig* observado com a montagem global. As Figuras 4, 5 e 6 mostram os resultados das reconstruções de sequências utilizando-se diferentes HMMs de perfis como sementes, e os três montadores citados anteriormente, além da contribuição das rotinas de otimização (em pares de bases) no tamanho da sequência final reconstruída.

Figura 4 - Reconstruções de sequências de *C. jejuni* utilizando-se o GenSeed-HMM e o montador ABySS, ou montagem global apenas com o montador. Os HMMs de perfis utilizados como sementes estão apresentados. O efeito do uso das rotinas de otimização está apresentado nas barras vermelhas.

Fonte: Oliveira (2012)

Figura 5 - Reconstruções de sequências de *C. jejuni* utilizando-se o GenSeed-HMM e o montador SOAPdenovo, ou montagem global apenas com o montador. Os HMMs de perfis utilizados como sementes estão apresentados. O efeito do uso das rotinas de otimização está apresentado nas barras vermelhas.

Fonte: Oliveira (2012)

Figura 6 - Reconstruções de sequências de *C. jejuni* utilizando-se o GenSeed-HMM e o montador Velvet, ou montagem global apenas com o montador. Os HMMs de perfis utilizados como sementes estão apresentados. O efeito do uso das rotinas de otimização está apresentado nas barras vermelhas.

Fonte: Oliveira (2012)

Os HMMs de perfis PF00009 e PF02729 foram escolhidos para validar a capacidade do GenSeed-HMM de reconstruir vários *loci* simultaneamente, visto que os respectivos genes apresentam várias cópias no genoma bacteriano. O HMM de perfil PF00009 corresponde a um conjunto de proteínas pertencentes à família dos fatores de alongação presentes em organismos eucariotos e procariotos, e existem no genoma de *C. jejuni* seis genes que codificam essas proteínas: fator de alongação G (código de acesso CAL34641), fator de alongação 4 (CAL35148), fator de alongação SelB (CAL35491), fator de alongação Tu (CAL34618), proteína de ligação ao GTP typA (CAL34219) e fator de iniciação da transcrição IF-2 (CAL34307). O HMM de perfil PF02729 corresponde a uma família de proteínas que contém duas enzimas relacionadas evolutivamente: aspartato carbamoiltransferase (CAL35112) e ornitina carbamoiltransferase (CAL35215). Como previsto, as montagens realizadas com GenSeed-HMM/ABYSS utilizando-se as sementes PF00009 e PF02729 geraram múltiplas sequências finais, respectivamente quatro e duas. Por outro lado, com o GenSeed-HMM/SOAPdenovo, essas mesmas sementes resultaram na reconstrução de apenas uma e duas sequências finais, respectivamente. Com o GenSeed-HMM/Velvet

foi gerada apenas uma sequência por semente. Quando utilizados os montadores ABySS, SOAPdenovo e Velvet em processos de montagem global, todos conseguiram reconstruir todos os parálogos.

A montagem progressiva com a semente PF00224 (piruvato quinase), utilizando os montadores ABySS e SOAPdenovo resultou na reconstrução de duas sequências finais em ambos os montadores. Porém, as sequências finais reconstruídas nos dois casos cobrem um único locus genômico onde se localiza o gene da piruvato quinase. Isso demonstra que as sequências reconstruídas não representam genes parálogos da piruvato quinase no genoma de *C. jejuni*, mas, sim um único gene que foi coberto por duas sequências distintas originadas pelo processo de montagem. De fato, existe apenas uma única cópia do gene codificador da piruvato quinase no genoma de referência dessa bactéria (CAL34542). Esse resultado errado de montagem também foi reproduzido com o uso de montagem global.

Diante dos resultados expostos acima, conclui-se que com o uso das rotinas adicionais, todos os montadores foram capazes de produzir reconstruções de tamanho similar às observadas em montagens globais desses mesmos montadores (excetuando-se a semente PF00162 com os montadores Velvet e SOAPdenovo). Por outro lado, o montador ABySS demonstrou melhor capacidade de montagem, resultando em sequências maiores do que os outros dois montadores, seja por montagem global como progressiva. Finalmente, esse montador também demonstrou uma melhor capacidade para a reconstrução de múltiplos genes de famílias proteicas dentro da montagem progressiva, sendo menos sensível a problemas de iniciação (ver seção 3.2.2) se comparado aos montadores SOAPdenovo e Velvet.

4.2.2 Montagem progressiva com leituras de *Escherichia coli*

Utilizou-se um conjunto de dados gerados na plataforma Illumina para o genoma da bactéria *E. coli* K12, cepa MG 1655 (National Center for Biotechnology Information, 2012d). Para a reconstrução das sequências, foram utilizados dez HMMs de perfis (PF02729, PF00009, PF00044, PF00288, PF03738, PF00012, PF00162, PF00224, PF00085, PF0348). As sequências finais reconstruídas localmente foram comparadas com as sequências correspondentes obtidas a partir da montagem global e com a sequência genômica de referência (código de acesso NC_000913.2), como descrito na

seção 4.1.1. Foi possível se observar que as rotinas de otimização descritas nas seções 3.1.1 foram fundamentais para a iniciação e manutenção da maioria das reconstruções (Figuras 7, 8 e 9). De fato, observa-se que em muitos casos a reconstrução nem sequer foi iniciada sem as rotinas, resultado caracterizado nos gráficos das Figuras 7 a 9 pela ausência de barras azuis. Esse resultado pode ser possivelmente atribuído a uma menor qualidade média das leituras desse conjunto de dados, quando comparado àquele de *C. jejuni*.

Os montadores ABySS, SOAPdenovo e Velvet, reconstruíram via GenSeed-HMM dois *loci* quando o HMM de perfil PF00044 foi utilizado como semente. As duas sequências finais reconstruídas correspondem aos genes da gliceraldeído-3-fostato desidrogenase (conhecido como *gapA*, número de acesso EIJ02961.1) e D-eritrose-4-fostato desidrogenase (conhecido como *gapB*, número de acesso EIJ15989.1). Esses dois genes são relacionados evolutivamente e apresentam papel fundamental na glicólise e no metabolismo da vitamina B6, respectivamente.

Figura 7 - Reconstruções de sequências de *E. coli* utilizando-se o GenSeed-HMM e o montador ABySS, ou montagem global apenas com o montador. Os HMMs de perfis utilizados como sementes estão apresentados. O efeito do uso das rotinas de otimização está apresentado nas barras vermelhas.

Fonte: Oliveira (2012)

Figura 8 - Reconstruções de sequências de *E. coli* utilizando-se o GenSeed-HMM e o montador SOAPdenovo, ou montagem global apenas com o montador. Os HMMs de perfis utilizados como sementes estão apresentados. O efeito do uso das rotinas de otimização está apresentado nas barras vermelhas.

Fonte: Oliveira (2012)

Figura 9 - Reconstruções de sequências de *E. coli* utilizando-se o GenSeed-HMM e o montador Velvet, ou montagem global apenas com o montador. Os HMMs de perfis utilizados como sementes estão apresentados. O efeito do uso das rotinas de otimização está apresentado nas barras vermelhas.

Fonte: Oliveira (2012)

No caso do HMM de perfil PF00009 (fatores de alongação ligados ao GTP), o GenSeed-HMM com os montadores ABySS e SOAPdenovo, geraram cada um cinco sequências finais, ao passo que o GenSeed-HMM com Velvet gerou apenas uma única sequência final. Essa diferença no número de sequências reconstruídas entre o montador Velvet e os outros dois programas de montagem é explicada pelas rotinas de iniciação da montagem progressiva. Como explicado na seção 3.1.3 três procedimentos são realizados para se iniciar a montagem progressiva. No caso do programa Velvet, após ser detectado que não houve sucesso na iniciação da montagem progressiva, o tamanho do câmero foi reduzido em duas bases e a montagem foi realizada novamente. Com a redução do câmero o GenSeed-HMM/Velvet foi capaz de reconstruir uma única sequência inicial que prosseguiu na montagem progressiva até o processo de exaustão. Porém, com o montador SOAPdenovo, os dois primeiros procedimentos executados pelas rotinas de iniciação não conseguiram reconstruir nenhuma sequência inicial. Assim, o montador ABySS foi invocado para realizar a reconstrução das leituras positivas somente na primeira rodada de execução. Por esta razão, as reconstruções com o GenSeed-HMM/ABySS e GenSeed-HMM/SOAPdenovo acabaram gerando cinco sequências finais reconstruídas, ao passo que o GenSeed-HMM/Velvet gerou apenas uma. Esse resultado, observado com vários outros conjuntos de dados, sugere que o ABySS é muito mais eficiente para realizar a primeira montagem de leituras recrutadas pela semente de HMM.

No caso dos HMMs de perfis PF00224, PF00085 e PF00012, quando utilizados como sementes no GenSeed-HMM/ABySS, resultaram, cada um, na montagem de duas sequências finais, correspondentes aos genes de piruvato quinase I e II (NP_416191.1 e NP_416368.1), tioredoxina I e tioredoxina II (AAC76786.2 e AAC75635.1), e proteínas de choque térmico Hsc62 e HscB (NP_415183.1 e NP_417022.1), respectivamente. No caso das montagens com o GenSeed-HMM/Velvet e GenSeed-HMM/SOAPdenovo utilizando os HMMs de perfis anteriormente descritos, apenas um único parólogo foi reconstruído para cada semente.

Um exemplo interessante pode ser observado com a montagem realizada com o GenSeed-HMM/ABySS e a semente PF02729. Esse HMM de perfil, como citado na seção anterior, apresenta um domínio proteico compartilhado por dois genes, o aspartato carbamoiltransferase (NP_41866.1) e a ornitina carbamoiltransferase (NP_418675.1). Esses dois genes aparecem de forma contígua dentro do genoma de *E. coli*, separados

um do outro por aproximadamente 6 kb (ver Figura 10D e 10E), a montagem global, com o montador ABySS, reconstruiu ambos os genes em duas sequências distintas (ver Figura 10B e 10C), sendo que o gene da ornitina carbamoiltransferase foi reconstruído parcialmente. Por outro lado, a montagem progressiva, via GenSeed-HMM/ABYSS, conseguiu reconstruir totalmente ambos os genes em uma única sequência final (ver Figura 10A).

Concluindo, os resultados observados com os dados de *E. coli* mais uma vez demonstraram o melhor desempenho do montador ABySS na montagem progressiva, quando comparados com os montadores SOAPdenovo e Velvet. Na maioria dos casos, o mesmo também se observa nas montagens globais. Em alguns casos, a montagem local conseguiu estender algumas regiões genômicas onde a montagem global foi incapaz, reconstruindo completamente genes que na montagem global permaneceram fragmentados. Foi observado ainda, em alguns casos, uma extensão adicional do comprimento de sequências reconstruídas de modo local, quando comparadas ao processo global de montagem, sem perda de especificidade (GenSeed-HMM/ABYSS com a semente PF00012).

Figura 10 - Mapeamento das sequências reconstruídas através de montagem progressiva e global GenSeed-HMM/ABYSS (contig A), ou de montagem global com o programa ABySS (contigs B e C) contra a sequência genômica de referência (linha tracejada na parte superior da imagem). As sequências correspondentes aos genes de aspartato (D) e ornitina carbamoiltransferase (E) também estão representadas.

Fonte: Oliveira (2012)

As rotinas de iniciação se mostraram fundamentais para o início da montagem progressiva, especialmente quando realizadas com os montadores SOAPdenovo e Velvet. Contudo, é sempre aconselhável uma análise a posteriori das sequências reconstruídas, pois em alguns casos as sequências geradas ao final do processo incorporaram regiões fisicamente não relacionadas, quando comparadas contra o

genoma de referência, resultado observado, por exemplo, na montagem com a semente PF00009 e o GenSeed-HMM/SOAPdenovo.

4.2.3 Montagem progressiva com leituras de *Gardnerella vaginalis*

Para as análises de *Gardnerella vaginalis*, foi utilizado um conjunto de leituras da plataforma 454 (cedido pelo Dr. José Paulo Alves), correspondente a uma cobertura de 130x do genoma da cepa patogênica 5-1 de *G. vaginalis* (Harwich et al., 2010). Para todas as montagens, utilizou-se o montador Newbler de forma global ou acoplado ao GenSeed-HMM. Nesse último caso, as reconstruções foram limitadas para um comprimento máximo de 30.000 pb.

O uso do programa Newbler no contexto do GenSeed-HMM exigiu algumas adaptações no processamento dos dados de montagem. O Newbler ignora leituras individuais com comprimento maior do que 1.999 pb no processo de montagem. Isso representa um grande problema dentro do fluxo de trabalho do GenSeed-HMM, pois a cada ciclo de montagem, a sequência reconstruída da rodada anterior é fornecida ao montador, juntamente com todas as leituras positivas recrutadas na rodada atual. Para contornar esse problema, foi desenvolvida uma abordagem similar à descrita por Wurm et al. (2011), que consiste em dividir as sequências com tamanho maior que 1.999 pb em subsequências menores com tamanho e sobreposições definidas (1,5 kb de tamanho e 1,3 kb de sobreposição). Para isso, incorporou-se ao GenSeed-HMM uma rotina que invoca o programa splitter (EMBOSS Package, 2012), presente no pacote EMBOSS. Esse programa realiza a fragmentação das sequências e, em seguida, as mesmas são encaminhadas para o montador, juntamente com as leituras positivas recrutadas. Essa abordagem, apesar de permitir o uso do montador Newbler dentro do fluxo de trabalho do GenSeed-HMM, apresenta uma desvantagem. Na medida em que a(s) sequência(s) consenso(s) aumenta(m) de tamanho, mais subsequências são criadas, e o tempo de execução em cada ciclo de iteração vai aumentando, causando um decréscimo da velocidade de processamento. A Tabela 4 mostra os resultados de reconstrução de sequências, obtidos com o uso das leituras 454 de *G. vaginalis* e diferentes HMMs de perfis. Todas as análises foram feitas com a montagem local e utilizando as rotinas de otimização descritas na seção 3.1.1.

Tabela 3 – Reconstruções realizadas utilizando o GenSeed-HMM e o montador Newbler

HMM de perfil utilizado como semente	Nº de sequências reconstruídas	Tamanho da maior sequência reconstruída (pb)
PF00009	4	30272
PF00288	2	30153
PF00137	1	30077
PF00044	1	30218
PF02729	1	30263
PF00012	1	30160
PF00162	1	30141
PF00224	1	30128
PF00085	1	30410
PF03485	1	30392

Fonte: Oliveira (2012)

Ao contrário das montagens realizadas com as bases de dados de leituras curtas, não foi necessário empregar as rotinas de iniciação (seção 3.1.1). Isso se explica pelo fato de que leituras longas certamente permitem uma detecção muito mais ampla da região correspondente ao perfil de HMM do que leituras curtas. Com isso, o processo de detecção é muito menos sensível a erros de sequenciamento pontuais do que quando realizado com sequências de 35-50 bp. Além disso, o algoritmo de montagem utilizado pelo Newbler (sobreposição-leiaute-consenso) é menos sensível a erros de sequenciamento do que os algoritmos usados pelos montadores ABySS, SOAPdenovo e Velvet (grafos de de Bruijn) (Li et al., 2011). Por outro lado, as rotinas que efetuam o aparamento iterativo das extremidades das sequências que estão sendo reconstruídas foram vitais para que algumas sequências atingissem o tamanho delimitado pelo GenSeed-HMM de 30 kb. Isso pode ser explicado porque as leituras da plataforma 454 tendem a acumular um número significativo de erros no final da sequência (extremidade 3'terminal), o que pode impedir sua sobreposição com leituras contíguas.

Como pode ser visto na Tabela 4, os genes correspondentes às sementes PF00009 (fatores de alongação ligados ao GTP) e PF00288 (GHMP quinases), resultaram na reconstrução de quatro (fator de alongação 2 – EFH71327.1, fator de alongação – EFH71872.1, proteína de ligação ao GTP lepA – EFH71327.1 e elemento

de resposta ao estresse – EFH71652.1) e duas sequências (gene da mevalonato quinase - EFH71573.1 e homoserina quinase - EFH72120.1), as quais de fato correspondem aos respectivos parálogos no genoma da bactéria. O HMM de perfil PF02729 representa uma família de proteínas geralmente codificadas por dois genes distintos (aspartato carbamoiltransferase e ornitina carbamoiltransferase), como já foi apresentado para *E. coli* (Figura 10). *G. vaginalis* apresenta em seu genoma apenas o gene da aspartato carbamoiltransferase (ZP_06976462.1) e, portanto, o resultado que obtivemos com a montagem progressiva refletiu a realidade deste genoma.

Concluindo, todas as montagens realizadas com o GenSeed-HMM/Newbler, atingiram o tamanho especificado de 30 kb, apresentando uma excelente correspondência (entre 99% e 100% de similaridade) com as sequências originadas através do processo global de reconstrução. Isso demonstra que a incorporação do montador Newbler no GenSeed-HMM foi realizada com sucesso, não havendo diferença entre os comprimentos dos contigs obtidos pelas montagens local e global. Um aspecto importante, contudo, que torna o uso do GenSeed-HMM/Newbler vantajoso, é que o uso desse programa para a reconstrução específica de poucos loci apresenta um tempo de execução mais curto do que o uso do Newbler em uma montagem global, seguido da identificação dos contigs correspondentes aos respectivos genes de interesse.

4.3 Reconstruções de genomas extracromossomais utilizando HMM de perfis e bases de dados originadas de organismos eucariotos

Para avaliar a capacidade do GenSeed-HMM de reconstruir genomas extracromossomais de organismos eucariotos usando sementes de HMM de perfil, foram utilizados três conjuntos de dados obtidos do NCBI Sequence Read Archive (ver Tabela 1), sendo dois conjuntos de leituras curtas Illumina, originadas do sequenciamento do genoma completo de *Drosophila melanogaster* (National Center for Biotechnology Information, 2012e) e de *Plasmodium vivax* (National Center for Biotechnology Information, 2012f), e um conjunto de dados de leituras longas de 454, obtido da levedura *Schizosaccharomyces japonicus* (National Center for Biotechnology Information, 2012g). Foram usados como sementes HMMs de perfis construídos a partir de proteínas codificadas exclusivamente por genes mitocondriais. As bases de dados de leituras curtas foram submetidas a um processo de filtragem semelhante ao descrito na

seção 4.2.1, tendo sido descartadas 44% das leituras de *D. melanogaster* e 9% de *P. vivax*. No caso das leituras de 454, não foi necessário o uso de filtro de qualidade.

4.3.1 Reconstrução do genoma mitocondrial de *Drosophila melanogaster*

O genoma mitocondrial de *D. melanogaster* tem 19.517 pb e apresenta 13 genes codificadores de proteínas. Para a reconstrução desse genoma extracromossômico, foram utilizados como sementes dois HMMs de perfis correspondentes a citocromo oxidase c subunidade II (PF00116) e NADH desidrogenase (PF00146). Além disso, para validar a capacidade de reconstrução de um mesmo locus genômico partindo de dois HMMs de perfis iniciais, as duas sementes acima foram concatenadas em um único arquivo, gerando um “Multi-HMM” de perfil. O gene da citocromo oxidase c e NADH desidrogenase localizam-se respectivamente, próximos à extremidade 5’ (posição 3.083 a 3.767) e 3’ (posição 12.669 a 12.733) do genoma mitocondrial de referência (código de acesso NC_001709), separados por aproximadamente 8 kb. Esse teste foi realizado para validar a capacidade do GenSeed-HMM de reconstruir e unir dois loci contíguos. Para todas as montagens foram utilizados o montador SOAPdenovo para montagens local e global. As duas sementes de HMM de perfil geraram uma única sequência final com 15.103 pb, idêntica à obtida por montagem global e à sequência do genoma de referência. Uma análise mais detalhada da reconstrução revelou que o processo de montagem progressiva foi interrompido em torno da posição 15.000 do genoma mitocondrial de referência, correspondendo à região do D-loop, composta por unidades repetitivas seriadas com cerca de 330 pb (Figura 11).

Figura 11 - Captura de tela do programa Artemis (Rutherford et al., 2000) mostrando uma região de 15.103 pb (assinalada com a cor verde), reconstruída a partir de leituras da plataforma Illumina de *D. melanogaster*.

A reconstrução foi feita com o GenSeed-HMM usando o montador SOAPdenovo e uma semente contendo dois HMMs de perfis (PF00116 e PF00146). Os genes codificadores das proteínas mitocondriais representados pelos HMMs de perfis estão assinalados nos quadros em cor vermelha, e as demais regiões codificadoras de proteínas estão representadas pelos quadros azuis. Notar que a extremidade 3' terminal da sequência reconstruída coincide com a presença de uma região altamente repetitiva, em torno da posição 15.000.

Fonte: Oliveira (2012)

4.3.2 Reconstrução do genoma mitocondrial de *Plasmodium vivax*

Os parasitas do gênero *Plasmodium* apresentam um genoma mitocondrial contendo unidades repetitivas seriadas (tandem repeats) com 5.990 pb de comprimento. Esse genoma é composto por apenas três genes codificadores de proteínas: citocromo oxidase 3 (cox3), citocromo oxidase 1 (cox1) e citocromo b (cytb) (Wilson e Williamson 1997). Para reconstrução desse genoma extracromossômico, foram utilizados dois HMMs de perfis obtidos da base Pfam, correspondentes ao citocromo oxidase 1 (PF00115) e ao citocromo oxidase 3 - subunidade III (PF00510). As reconstruções utilizando as duas sementes descritas anteriormente foram realizadas utilizando-se o montador Velvet até a exaustão do processo. As reconstruções feitas por montagem local e global geraram contigs idênticos de 6.059 pb. Essa sequência apresentou 99% de similaridade (blastn) contra o genoma mitocondrial de referência (NC_007243.1). A diferença de 69 pb de comprimento entre a sequência reconstruída e a do genoma de referência, na verdade corresponde às extremidades duplicadas, por se tratar de um genoma repetido em série, ou seja, cada ponta da sequência apresenta algumas bases da unidade repetitiva imediatamente anterior ou posterior.

Esses resultados, assim como aqueles apresentados para *D. melanogaster* (seção 4.3.1), sugerem que a abordagem utilizada pelo GenSeed-HMM pode ser tão eficiente quanto a montagem global, especialmente se as leituras apresentam uma alta qualidade. Contudo, como o GenSeed-HMM é usado para a reconstrução de alvos específicos, o custo computacional tende a ser muito menor, especialmente se as bases de dados forem muito grandes. De fato, as montagens locais realizadas com a base de dados de *P. vivax*, foram executadas em torno de 12 minutos e necessitaram de menos de dois gigabytes de memória RAM, ao passo que a montagem global foi executada em aproximadamente 31 minutos e necessitou de cerca de 20 gigabytes de memória RAM em determinadas etapas durante o processo de montagem.

4.3.3 Reconstrução do genoma mitocondrial de *Schizosaccharomyces japonicus*

As leveduras do gênero *Schizosaccharomyces* formam um amplo e antigo clado dentro dos fungos ascomicetos, e são utilizados como modelo para o estudo da biologia dos organismos eucariotos (Rhind et al., 2011). Esse clado engloba os fungos *Schizosaccharomyces pombe*, *S. octosporus*, *S. cryophilus* e *S. japonicus*. Como alvo da

montagem progressiva, escolhemos um conjunto de dados de sequenciamento do genoma completo de *S. japonicus*. A reconstrução do genoma mitocondrial foi realizada utilizando-se um HMM de perfil de citocromo oxidase 2 e o montador *Newbler*. Após 178 rodadas de execução, foi obtida uma sequência final reconstruída com 80.048 pb, deduzida a partir de 28.642 leituras longas da plataforma 454. A sequência reconstruída, quando comparada com a do genoma mitocondrial de referência (total de 80.059 pb, código de acesso AF547983.1), apresentou 99% de similaridade, cobrindo quase toda a extensão do genoma mitocondrial. Ao se realizar a comparação entre a montagem realizada pelo *GenSeed-HMM/Newbler* e a montagem global realizada por esse mesmo montador (utilizando o programa *blastn*), verificou-se que a montagem local foi superior a montagem global em termos de contiguidade da sequência reconstruída. Em termos de desempenho computacional a montagem local via *GenSeed-HMM* foi realizada em menos de 18 minutos, ao passo que a montagem global realizada pelo montador *Newbler* levou cerca de quatro horas. Nesse teste, as rotinas que realizam o aparaamento das pontas e os parâmetros para seleção e montagem das leituras, tiveram papel fundamental no sucesso da reconstrução. Esse fato deve-se a baixa qualidade das bases localizadas na extremidade final das leituras, que possivelmente foram responsáveis pelo baixo desempenho da montagem global (em termos de comprimento das sequências reconstruídas e de processamento computacional).

4.4 Reconstruções de loci utilizando HMM de perfis e bases de dados originadas de amostras metagenômicas

Para validar a capacidade do *GenSeed-HMM* de ser utilizado em estudos metagenômicos, vários conjuntos de dados foram obtidos a partir do NCBI Sequence Read Archive (ver Tabela 1). Foram utilizadas no total quatro bases de dados meteganômicos de leituras curtas da plataforma Illumina, derivadas de microbioma intestinal humano (National Center for Biotechnology Information, 2012h). Essas leituras foram submetidas a um processo de filtragem semelhante ao descrito na seção 4.1.1, porém dessa vez com o programa *fastq_quality_filter* (Hannon Lab, 2012), excluindo as leituras com menos do que 90% de bases com qualidade inferior a 30 (parâmetros *-q 30 -p 90*). Após o processo de filtragem, foi gerado um conjunto de dados de alta qualidade com 107.258.490 leituras. Além desse conjunto de dados,

foram utilizados nos ensaios de validação um total de 154 conjuntos de dados de leituras longas (plataforma 454) do mamute lanoso extinto (National Center for Biotechnology Information, 2012i).

4.4.1 Genoma mitocondrial do mamute lanoso extinto (*Mammuthus primigenius*)

Por diversas razões, o DNA mitocondrial tem sido alvo de estudos em vários projetos envolvendo DNAs fósseis, visto que apresenta um número muito maior de cópias do que o genoma nuclear, e tem, portanto, maior chance de ser recuperado (Ho e Gilbert, 2010). Para a reconstrução da sequência do genoma mitocondrial de mamute, utilizou-se o HMM de perfil PF00510 (subunidade III do citocromo c oxidase) como semente, e o programa GenSeed-HMM/Newbler sem restrição do número de ciclos de montagem. Após 70 rodadas e o recrutamento de 10.736 leituras, foi obtida uma sequência final de 16.863 pb. Essa sequência apresentou 99% de similaridade com o genoma mitocondrial de mamute (tamanho de 16.770 pb, código de acesso NC_007596.2). A diferença de 93 pb entre as duas sequências é devida à região de D-loop (posição 15.422 – 16.770), que resulta na incorporação de regiões redundantes em ambas as extremidades da sequência reconstruída. Esse resultado é uma prova de princípio de que a reconstrução de loci genéticos, e até mesmo de genomas extracromossômicos inteiros, pode ser realizada utilizando-se o programa GenSeed-HMM. Um importante aspecto desse tipo de aplicação é que o uso de montagem global utilizando o montador Newbler ou CAP3, para um conjunto de dados tão grande (cerca de 31 milhões de leituras) requereria computadores de alto desempenho. Isso inviabilizaria essa tarefa para laboratórios pequenos e sem grande infraestrutura computacional. Em nosso laboratório, usando-se um computador com processador Intel i7 (Intel, Santa Clara, Estados Unidos da América) e 16 GB de memória RAM, a tarefa de reconstrução desse genoma mitocondrial a partir desse conjunto de dados demorou menos que 18 minutos.

4.4.2 Microbioma intestinal humano

O conteúdo bacteriano presente na microbiota humana tem sido objeto de vários estudos nesses últimos anos, liderados por grandes projetos de pesquisa como o Projeto Microbioma Humano (Peterson et al., 2009; Turnbaugh et al., 2007), e o

catálogo de genes do microbioma humano estabelecido por sequenciamento metagenômico, MetaHIT (Qin et al., 2010). Para as nossas reconstruções, foram utilizados oito HMMs de perfis como sementes (PF00689, PF00145, PF00534, PF06580, PF00012, PF09112, PF01131 e PF00905 – vide Tabela 3), os quais correspondem às proteínas mais abundantes encontradas em amostras de microbioma humano (Arumugam et al. 2011). Para as montagens, foi utilizado o GenSeed-HMM e os montadores de leituras curtas ABySS, SOAPdenovo e Velvet. Todas as reconstruções realizadas com os HMMs dos perfis acima, deram origem a diversas sequências finais, com exceção da montagem realizada com o GenSeed-HMM/Velvet e a semente PF00905, que resultou em apenas uma sequência reconstruída. Todas as sequências reconstruídas foram analisadas através de buscas de similaridade contra a base de dados nr (não redundante) do NCBI, visando identificar os seus respectivos organismos de origem. Além disso, todas as sequências foram testadas quanto à presença dos motivos proteicos nas respectivas proteínas codificadas. Conforme pode ser visto na Tabela 5, o montador ABySS resultou na reconstrução do maior número de sequências, seja na forma de gene inteiro, seja em termos de montagem parcial. Além disso, esse montador resultou na maior diversidade de espécies de origem das sequências. Os resultados das montagens local e global foram bastante semelhantes em termos do comprimento das sequências. Existem alguns fatores que dificultam a reconstrução de contigs muito longas em bases de dados metagenômicas, principalmente quando originadas de bases de leituras curtas, como, polimorfismos entre subespécies similares, regiões genômicas similares que são compartilhadas entre diferentes espécies, e proporções desiguais entre as espécies presentes na amostra (Peng et al., 2011).

Tabela 5 – Reconstruções de sequências de dados metagenômicos de microbioma intestinal humano utilizando o programa GenSeed-HMM com os montadores ABySS, SOAPdenovo e Velvet.

Montador	Nº total de sequências reconstruídas	Nº total de organismos distintos	Nº total de genes reconstruídos totalmente	Nº total de genes reconstruídos parcialmente
ABySS	256	71	47	209
SOAPdenovo	67	37	21	46
Velvet	60	35	15	45

Fonte: Oliveira (2012)

Em termos de desempenho, o tempo de processamento computacional foi maior com o GenSeed-HMM (chegando em até oito horas), quando comparado com os respectivos montadores em montagens globais (cerca de um hora). Isso se explica porque os montadores de DNA que utilizam a abordagem de grafos de de Bruijn, a montagem das leituras é realizada de maneira muito rápida. Como cada ciclo de montagem realizado pelo GenSeed-HMM consiste em buscas de similaridade, recrutamento de leituras, montagem e extensão progressiva das sequências, o tempo de processamento computacional pode ser muito elevado em bases de dados muito grandes. Além disso, quanto mais longas as sequências reconstruídas, maior o número de ciclos de montagem necessários. Esse aspecto é ainda mais crítico com conjuntos de dados de sequências curtas, onde cada ciclo resulta na extensão de apenas alguns poucos pares de bases.

Os resultados obtidos demonstraram claramente que um único HMM de perfil permite reconstruir genes ortólogos e parálogos de uma base de dados metagenômicos. Como exemplo, podemos citar o caso do HMM de perfil PF00534 com o montador ABYSS, a partir do qual foram reconstruídas 99 sequências pertencentes a 44 organismos diferentes. O GenSeed-HMM pode portanto ser utilizado para a prospecção de organismos em bancos de dados metagenômicos, bem como para a reconstrução de famílias proteicas específicas a partir desse tipo de dado.

5 DISCUSSÃO

5.1 Características do programa GenSeed-HMM e Seeder

O programa GenSeed-HMM é uma versão atualizada do programa original GenSeed. Foram incorporadas novas características a essa atual versão, mas toda a funcionalidade da versão original foi preservada. A versão original do programa é restrita ao uso de leituras longas geradas por sequenciamento Sanger, e sementes de DNA ou proteínas. A versão atual do GenSeed-HMM, por outro lado, suporta três tipos de sementes (DNA, proteína e HMM de perfil), e permite ao usuário realizar montagens progressivas utilizando tanto bases de dados de leituras curtas (Illumina) como de leituras longas (Sanger e 454). Enquanto o GenSeed utiliza apenas o CAP3 como montador, o GenSeed-HMM é também compatível com os montadores ABySS, SOAPdenovo, Velvet e Newbler. Os montadores foram incorporados ao programa GenSeed-HMM de maneira modular, onde cada módulo funciona como um envoltório que interage diretamente com o programa GenSeed-HMM. Assim, a incorporação de novos montadores é realizada de maneira simples e rápida. Isto é particularmente importante, visto que novas plataformas de sequenciamento estão sendo introduzidas constantemente no mercado. Com a emergência dessas tecnologias, as dificuldades impostas para montagem dos genomas e a metodologias utilizadas pelos montadores podem se modificar, necessitando assim, de novas estratégias e diferentes programas de montagem para lidar com esses novos tipos de dados (Compeau et al., 2012). Dessa maneira, o programa GenSeed-HMM pode incorporar em seu fluxo de trabalho novos montadores com extrema facilidade, adequando o programa rapidamente às eventuais modificações que possam ocorrer dentro do campo de sequenciamento de DNA.

Outros montadores descritos na literatura como o ALLPaths (Butler et al., 2008), Edena (Hernandez et al., 2008) e o Mira (Chevreux et al., 1999) foram testados para uso integrado ao GenSeed-HMM. Contudo, devido a diferentes razões, o uso desses montadores no contexto do GenSeed-HMM não foi possível. O montador ALLPaths necessita de no mínimo duas bibliotecas distintas de sequenciamento para realizar a montagem de sequências. Essas bibliotecas, conhecidas como *paired-end* ou *mate-pair*, são construídas sequenciando-se ambas as extremidades dos fragmentos de DNA. Como o tamanho médio desses fragmentos é conhecido, é possível se obter a distância que separa as duas leituras pareadas. As informações obtidas a partir dessas bibliotecas auxiliam o processo de montagem de genomas, e a identificação de

rearranjos estruturais em projetos de ressequenciamento. O uso do montador `ALLPaths` integrado ao `GenSeed-HMM` é impossível, pois o processo de montagem local progressiva é incompatível com o funcionamento desse programa.

O programa `Edena`, por sua vez, somente é capaz de montar sequências que apresentam exatamente o mesmo comprimento. Uma vez que o `GenSeed-HMM` utiliza nos ciclos de montagem e extensão as leituras recrutadas juntamente com a sequência reconstruída da rodada anterior, é impossível empregar o `Edena` para as montagens progressivas do `GenSeed-HMM`. Finalmente, o programa `Mira`, semelhantemente a outros montadores de leituras curtas, baseia-se na cobertura da amostra sequenciada. Contudo, ao contrário dos outros montadores, o `Mira` não permite ajustar esse parâmetro. Assim, o programa assume que a cobertura é uniforme ao longo de toda a sequência, e regiões com baixa densidade física de leituras não são consideradas no processo de montagem. Dessa maneira, quando o `GenSeed-HMM` recruta as leituras a partir das sementes e as encaminha para o montador, juntamente com a sequência reconstruída da rodada anterior, introduz-se uma grande diferença na cobertura de leituras nesse conjunto de dados. Como consequência, o `Mira` realiza apenas a montagem das regiões com alta cobertura (relativa às pontas da sequência), excluindo a sequência reconstruída da rodada anterior, impossibilitando a extensão das sequências por montagem progressiva.

Com relação às buscas de similaridade, o `GenSeed` original utilizava o pacote `BLAST`. Na nova versão, foi implementado o uso do pacote `BLAST+`, o qual apresenta melhoras significativas nos tempos de processamento e memória utilizada. Além disso, no `GenSeed` original, o arquivo de saída do `BLAST` tinha que ser processado com rotinas específicas para se extrair as informações de interesse a partir da saída tabular. No caso do `BLAST+`, a saída é totalmente configurável, permitindo gerar arquivos que já contém exatamente todas as informações requeridas, sem necessidade de processamento adicional.

Finalmente, o programa `GenSeed-HMM` incorporou o uso de HMMs de perfis como semente no seu fluxo de trabalho. Para isso, utiliza-se o `HMMER3`, um pacote que realiza buscas de similaridade usando HMMs de perfis como consulta contra bases de dados proteicas. O `HMMER3` permite executar buscas de similaridade com excelente desempenho, comparável até mesmo a buscas de similaridade do `BLAST+`.

O programa GenSeed-HMM, permite utilizar uma grande variedade de arquivos de entrada como sementes (nucleotídicas, proteicas e HMMs de perfis) e bases de dados (leituras curtas e leituras longas), bem como vários montadores distintos (CAP3, Newbler, ABySS, SOAPdenovo e Velvet). Essa grande diversidade de combinações de arquivos de entrada e montadores torna a configuração dos parâmetros do GenSeed-HMM uma tarefa bastante desafiadora para usuários iniciantes. Para facilitar essa tarefa, foi desenvolvido o *Seeder* (Figura 3) um programa gráfico para a construção assistida de arquivos de configuração otimizados para o GenSeed-HMM. O arquivo de configuração é um arquivo em formato texto quem contém todas as informações necessárias para a execução do programa. O *Seeder* permite ao usuário selecionar numa interface gráfica os parâmetros desejados e gerar um arquivo de configuração. Alternativamente, o programa pode determinar automaticamente o tipo de semente e base de dados escolhidos pelo usuário e sugerir automaticamente os parâmetros mais recomendados. É importante ressaltar que os arquivos de configuração construídos pelo *Seeder* não contém necessariamente o conjunto de parâmetros mais apropriados, mas representam um ótimo ponto de partida para o usuário.

5.2 Desenvolvimento e desempenho do programa GenSeed-HMM

O decréscimo do custo da geração de dados produzidos pelas novas plataformas de sequenciamento não foi acompanhado na mesma escala por uma queda dos custos de infraestrutura computacional requeridos para a sua análise (Sboner, et al., 2011). Levando isso em consideração, cada etapa do fluxo de trabalho do programa GenSeed-HMM foi analisada e otimizada individualmente antes de ser incorporada ao programa como uma rotina. Assim, o programa GenSeed-HMM foi desenvolvido com o objetivo de realizar a reconstrução de genes ou famílias gênicas sem requerer a complexidade computacional que seria necessária para se realizar a montagem global de todos os dados.

Para evitar grandes demandas computacionais, o GenSeed-HMM não utiliza durante a montagem progressiva toda a sequência reconstruída da rodada anterior como semente para a rodada seguinte. Ao invés disso, o programa utiliza apenas as extremidades das sequencias reconstruídas como sementes na rodada seguinte. Essa abordagem torna o GenSeed-HMM muito mais rápido ao longo do processo

intermediário de montagem. O tamanho das extremidades utilizadas como sementes nas rodadas de montagem é definido pelo próprio usuário, e tem relação direta com o tipo de base de dados utilizada no processo de montagem progressiva. O tamanho da semente precisa ser definido de modo que exista uma sobreposição razoável entre si e as leituras presentes na base de dados, de forma a permitir a extensão da sequência que está sendo reconstruída. Esse tamanho, contudo, não pode ser muito pequeno, para evitar que muitas sequências não relacionadas sejam incorporadas e perturbem o processo de montagem.

Em termos comparativos, a reconstrução de um locus genômico de tamanho similar, realizada utilizando-se bases de dados de leituras curtas e longas, apresenta grande diferença no tempo de processamento e número de rodadas de execução no GenSeed-HMM. Essa diferença é devida principalmente ao comprimento das leituras e tamanho da base de dados utilizadas na reconstrução. O pequeno comprimento das leituras curtas, que implica em menor informação, tem que ser compensado por uma grande redundância de leituras (alta cobertura), para se deduzir com confiança a sequência reconstruída (Miller et al., 2010). Além disso, leituras curtas adicionam poucas bases a cada ciclo de montagem progressiva do GenSeed-HMM, exigindo um número muito maior de ciclos para se cobrir uma dada extensão de DNA. É importante salientar que em cada uma das rodadas de montagem do GenSeed-HMM há rotinas específicas de buscas de similaridade, recuperação de sequências, extensão e verificação das sequências reconstruídas, o que exige certo grau de processamento computacional. Dessa forma, montagens locais em bases de dados de leituras curtas frequentemente exigem maior tempo de processamento do que uma montagem global realizada pelo respectivo montador. No que diz respeito às bases de dados de leituras longas, elas apresentam como característica uma cobertura em geral muito inferior à de bases de leituras curtas, porém com comprimento médio de leituras muito maior. Considerando-se o GenSeed-HMM, bases de leituras longas requerem muito menor número de ciclos de montagem para se reconstruir um determinado locus genômico, uma vez que cada rodada incorpora uma informação maior. Além disso, como as bases têm menos cobertura (são menores), as buscas de similaridade e o recrutamento de leituras são mais rápidas. Como consequência, as reconstruções locais via GenSeed-HMM, quando comparadas com montagens globais de leituras longas, são realizadas de uma maneira proporcionalmente mais rápida. Outro aspecto importante a ser considerado é que os

montadores de leituras longas como o CAP3 e o Newbler são proporcionalmente muito mais lentos do que os montadores de leituras curtas, que utilizam grafos de de Bruijn. Assim, montar todo o conjunto de leituras de uma base de dados é frequentemente muito mais lento e computacionalmente intensivo do que reconstruir por montagem progressiva alguns loci específicos com o GenSeed-HMM. Outra vantagem importante do uso do GenSeed-HMM em relação a montagens globais, é o baixo uso de recursos de memória. Assim, montagens globais de bases de dados grandes, sejam elas de leituras curtas ou longas, exigem recursos de memória que podem facilmente chegar a dezenas e mesmo centenas de gigabytes. Como o GenSeed-HMM somente realiza montagens com as leituras recrutadas em cada rodada, a memória requerida é de poucos gigabytes.

Concluindo, o desempenho do programa GenSeed-HMM para a reconstrução de loci grandes em bases de dados de leituras curtas, é pior em termos de processamento do que o observado em montagens globais, utilizando-se os montadores ABySS, Velvet e SOAPdenovo, mas os resultados em termos de comprimento das sequências é na maioria das vezes semelhantes. Por outro lado, o desempenho do GenSeed-HMM em montagens locais é superior ao de montagens globais usando-se somente o montador Newbler, tanto em termos de tempo de processamento como, em alguns casos, na extensão das sequências reconstruídas.

Em termos gerais de desempenho computacional, as montagens feitas pelo GenSeed-HMM são realizadas em poucos minutos. Contudo, o tempo de execução das reconstruções pode variar devido aos seguintes fatores: tamanho da base de dados utilizada; número de loci reconstruídos simultaneamente, que depende do número de cópias dos genes, ou da diversidade de organismos em amostras metagenômicas; comprimento das leituras, que depende da plataforma de sequenciamento; e montador selecionado para realizar as montagens. Embora o desempenho do GenSeed-HMM para leituras curtas não justifique seu emprego, quando comparado a montagens globais. Contudo, para bases de dados de leituras longas, a vantagem do programa se torna muito mais evidente. De qualquer forma, é importante ressaltar que a versão final do GenSeed-HMM permite o uso de montadores de ambas, leituras curtas e longas, oferecendo grande versatilidade ao usuário final.

5.3 Rotinas de otimização para eliminação de erros de sequenciamento

Várias rotinas destinadas à otimização do processo de montagem progressiva foram incorporadas no fluxo de trabalho do programa *GenSeed-HMM* (ver seção 3.1.2). Os procedimentos realizados por essas rotinas têm como finalidade a eliminação de erros presentes nas leituras durante o processo de reconstrução. Os erros em leituras de DNA são específicos e inerentes a cada uma das diferentes plataformas de sequenciamento. A plataforma 454, por exemplo, gera leituras com muitos erros de inserção e deleção de nucleotídeos associados a regiões homopoliméricas (Huse et al., 2007), ao passo que a plataforma Illumina produz leituras com erros de substituição na região 3'-terminal de suas leituras (Dohm et al. 2008). Assim, quando a reconstrução de sequências é realizada pelo *GenSeed-HMM*, essas leituras podem inviabilizar o início da montagem progressiva, assim como levar à incorporação de bases erradas nas extremidades da sequência que está sendo reconstruída. Essas sementes nucleotídicas criadas a partir das extremidades da sequência irão ser utilizadas para buscas de similaridade contra a base de dados, recrutando por sua vez, leituras que também apresentam erros de sequenciamento. Essas leituras, por apresentarem baixa representatividade na base de dados, podem levar a uma interrupção prematura do processo de montagem progressiva, mesmo quando há cobertura física de leituras naquela região. As rotinas de otimização checam se há término de progressão da reconstrução, caracterizado pela ausência de incorporação de novas bases na sequência. Quando a sequência não aumenta de comprimento em duas rodadas consecutivas, um corte arbitrário das pontas é realizado (trimming). Dessa forma, possíveis erros de sequenciamento, geralmente presentes em extremidades de leituras, podem ser eliminados, permitindo assim retomar o processo de reconstrução.

Os montadores que utilizam a abordagem dos grafos de de Bruijn (*ABYSS*, *SOAPdenovo* e *Velvet*) realizam a remoção e correção de erros apoiados na redundância de leituras curtas presente nas bases de dados (ver seção 1.3.2). Assim, o usuário pode definir uma cobertura média dos câmeras, e aqueles que apresentem uma cobertura inferior à definida, são excluídos do processo de montagem. Nesse ponto, é importante lembrar que a cada ciclo de montagem realizado pelo *GenSeed-HMM*, somente as extremidades das sequências (sementes) são utilizadas para se realizar a busca por similaridade (Figura 12A). Dessa maneira, quando as leituras recrutadas,

correspondentes às extremidades da sequência reconstruída da rodada anterior, são fornecidas ao montador juntamente com a própria sequência, introduz-se uma enorme diferença de cobertura. Assim, os montadores estimam uma cobertura média de câmeros que não corresponde à realidade, tornando inviável a remoção de erros através desse procedimento. Como demonstrado por Pevzner e Tang (2001), a remoção de erros de sequenciamento é um passo indispensável para o bom desempenho das montagens que utilizam metodologia de grafos de de Bruijn. Portanto, os montadores ABySS, SOAPdenovo e Velvet, por não conseguirem remover os erros de sequenciamento de maneira eficiente dentro do contexto de uso do GenSeed-HMM (Figura 12B), levam frequentemente a paradas prematuras durante o processo de montagem.

Alguns procedimentos podem ser adotados para a eliminação de leituras com erros de sequenciamento, antes de essas serem fornecidas aos montadores. O primeiro procedimento consiste em se realizar uma filtragem por qualidade das leituras presentes nas bases de dados, antes de inseri-las no processo de montagem progressiva (ver seções 4.2.1, 4.3 e 4.4). O segundo, consiste em se realizar uma seleção rigorosa das leituras que irão ser destinadas para o montador, através de parâmetros da busca de similaridade com o BLAST+ e do próprio GenSeed-HMM. Finalmente, pode-se utilizar preferencialmente o programa ABySS para a reconstrução de leituras curtas de DNA. Inserido nesse programa existe um parâmetro que não utiliza o valor da distribuição média de câmeros para a remoção dos erros. O parâmetro `-e` remove câmeros que apresentem uma cobertura inferior à delimitada pelo usuário, apenas das extremidades das sequências que estão sendo reconstruídas, não levando em consideração a frequência global dos câmeros. Dessa maneira, o programa ABySS consegue diminuir a quantidade de erros introduzidos dentro do processo de montagem progressiva de maneira muito mais eficiente que os montadores Velvet e SOAPdenovo. Essa característica do programa ABySS pode explicar em grande parte os resultados muito melhores que se observou na reconstrução de sequências com este programa, quando comparado ao SOAPdenovo e o Velvet (seções 4.2.1, 4.2.2 e Figura 4 a 10). De fato, o configurador Seeder sugere no modo automático o uso do ABySS como montador padrão para leituras curtas.

Na abordagem tradicional da sobreposição-leiaute-consenso, utilizada pelos montadores CAP3 e Newler, a correção de erros está implicitamente contida nos parâmetros utilizados para se validar a sobreposição entre as leituras (Zerbino, 2009).

Esses parâmetros levam em consideração o tamanho mínimo da sobreposição entre as leituras, assim como a porcentagem mínima de identidade entre elas (ver seção 1.3.1). Dessa forma, erros de sequenciamento pouco numerosos não chegam a impedir o alinhamento entre duas leituras, tornando essa abordagem mais robusta para a montagem. Esse aspecto também foi observado nas montagens realizadas pelo GenSeed-HMM, onde mesmo sem a pré-filtragem por qualidade das bases de dados, nenhuma reconstrução foi interrompida devido a erros de sequenciamento (seções 4.2.3, 4.3.3 e 4.4.1).

Figura 12 - Montagem progressiva e o efeito de erros de sequenciamento nos grafos de de Bruijn.

O GenSeed-HMM utiliza as extremidades das sequências reconstruídas como sementes (barras verdes) para as rodadas de montagem. Somente as leituras que apresentam sobreposição e estendem as extremidades das sementes (barras azuis) são recrutadas (A). Esse processo faz com que os montadores não possam determinar a verdadeira cobertura média de câmeros no conjunto de toda a sequência, impedindo assim a correta eliminação das leituras que contêm erros de sequenciamentos. O grafo de de Bruijn (B) resultante do uso de leituras contendo erros de sequenciamento. Os círculos representam câmeros originados a partir das leituras recrutadas com (azul) e sem erros de sequenciamento (vermelho) e os originados a partir da sequência reconstruída na rodada anterior (laranja).

Fonte: Oliveira (2012)

5.4 Perspectivas futuras e aplicações

5.4.1 Reconstrução de sequências a partir de DNA fóssil

Recentemente, Agoni et al. (2012) identificaram inserções de retrovírus endógenos (HERV-K) em vários loci pertencentes a hominídeos ancestrais, através de leituras de DNA (Illumina) originadas do sequenciamento do homem de Neandertal e

do hominídeo de Denisova. Para buscar regiões de junção entre os HERV-K e os genomas dos hominídeos, foram utilizadas sequências correspondentes às extremidades de vários HERV-K como consulta, em buscas de similaridade contra as bases de leituras. Esse trabalho permitiu identificar 14 provirus pertencentes ao hominídeo de Denisova e ao Neandertal. Outro trabalho utilizando DNA fóssil foi relatado por Campbell et al. (2010), que conseguiram amplificar e sequenciar regiões codificadoras do gene da hemoglobina a partir de amostras de DNA provenientes de ossos do mamute lanoso (cerca de 43.000 anos atrás). De forma similar, outros trabalhos têm sido descritos utilizando-se DNAs fósseis obtidos a partir de pelos, coprólitos, ossos, cascas de ovos e solo, possibilitando o entendimento de aspectos evolutivos e fisiológicos de organismos extintos (Huynen et al., 2012). Os trabalhos descritos acima, são exemplos de caracterização de loci, ou genes individuais, presentes em genomas de organismos extintos, a partir de bases de dados fragmentadas. Essa mesma abordagem poderia ser realizada de forma automática com o GenSeed-HMM. De fato, testes de validação demonstraram que a ferramenta pode ser aplicada em uma ampla variedade de bases de dados, inclusive bases de dados provenientes de organismos extintos (seção 4.1.1), tendo reconstruído com êxito a quase totalidade da sequência do genoma mitocondrial do mamute lanoso (seção 4.4.1). Assim, para ambos os trabalhos exemplificados acima, HMMs de perfis representando proteínas ortólogas de retrovírus endógenos ou hemoglobinas, respectivamente, poderiam ser construídos e utilizados como sementes para a reconstrução de sequências nessas bases de dados fragmentadas. Além disso, o programa permite a reconstrução simultânea de vários loci relacionados. Finalmente, o uso de HMMs de perfis é particularmente útil para a detecção de sequências ortólogas e parálogas.

5.4.2 Uma nova abordagem para a detecção de agentes etiológicos

Outra aplicação para o programa GenSeed-HMM é a reconstrução e identificação de novas entidades virais, haja visto a grande quantidade de dados metagenômicos produzidos recentemente e sua disponibilidade em sítios públicos. Um exemplo do uso de dados metagenômicos para a identificação de novos vírus está descrito por Lauck et al. (2011). Esse trabalho relata a identificação de duas novas variantes de um arterivírus (SFHV) a partir de amostras de sangue de macacos de cativeiro que morreram devido a um surto fatal de febre hemorrágica em Uganda. O

diagnóstico foi feito a partir de pirosequenciamento, seguido por montagem de novo das leituras e buscas de similaridade. De fato, a abordagem de sequenciamento de DNA, montagem global das leituras e, finalmente, a comparação das sequências reconstruídas contra bases de dados públicas, é a mais adotada (Mokili et al., 2012).

O programa GenSeed-HMM oferece uma alternativa a essa abordagem tradicional através de um fluxo de trabalho simples e automatizado, no qual a reconstrução de genomas virais poderia ser realizada com a utilização de HMMs de perfis derivados de sequências virais já conhecidas. De fato, uma das aplicações que vislumbramos para o GenSeed-HMM, é a sua possível utilização para a detecção de agentes etiológicos. Atualmente, as abordagens tradicionais como PCR competitivo ou métodos de microarranjos, se limitam a identificar sequências de organismos já conhecidos. Esses métodos requerem que uma lista de patógenos a serem considerados seja feita, limitando, assim, a busca por novos patógenos ou organismos relativamente distantes daqueles já conhecidos. Em contrapartida, o sequenciamento metagenômico é imparcial e nos fornece a oportunidade de considerar um amplo espectro de organismos, o que é ideal para aplicações onde bactérias, fungos, parasitas, e vírus desconhecidos possam estar envolvidos (Whitley, 2008).

Os sistemas convencionais de diagnóstico biológico sempre se basearam na premissa de detectar um alvo conhecido, seja ele um antígeno ou uma sequência nucleotídica. Essa abordagem, contudo, falha em detectar patógenos emergentes, a não ser que eles apresentem grande semelhança com aqueles previamente conhecidos e, portanto, resultem em uma reatividade cruzada. Mesmo assim, o diagnóstico obtido é falho devido à inespecificidade do resultado, visto que o que se diagnostica mais uma vez é a entidade previamente conhecida, e não a nova. Uma forma de quebrar esse paradigma é tentar utilizar métodos mais refinados que possibilitem a detecção de organismos mais distantes, sem abrir mão da especificidade da detecção. O uso de HMMs de perfis de proteínas essenciais ao funcionamento de vírus pode contribuir para a detecção de novos vírus, desde que eles compartilhem algumas características estruturais e/ou funcionais semelhantes. A detecção de novos vírus a partir de amostras de dados metagenômicos com o uso de HMMs de perfis pode criar um novo paradigma diagnóstico, em que a caracterização prévia desses vírus já não será uma condição *sine qua non*. Essa nova modalidade diagnóstica que esperamos alcançar com o programa GenSeed-HMM denominamos de diagnóstico de novo.

5.4.3 GenSeed-HMM como uma ferramenta auxiliar para o fechamento de genomas

Com as novas plataformas de sequenciamento, o processo de fechamento de genomas tornou-se mais complexo, principalmente quando se utilizam leituras curtas. Apesar de genomas inteiros poderem ser sequenciados em poucos dias, gerando uma grande quantidade de informação, a montagem desses genomas frequentemente é muito fragmentada. Assim, o fechamento de lacunas entre as sequências reconstruídas pode ser uma tarefa muito desafiadora e demorada em projetos de sequenciamento, requerendo métodos específicos de biologia molecular (Nagarajan et al., 2010). Entretanto, em alguns casos, análises *in silico* podem contribuir para o fechamento, sem que se tenha necessariamente que voltar aos experimentos de bancada. Os testes de validação que obtivemos com a montagem local via GenSeed-HMM, permitiram identificar vários casos nos quais as sequências reconstruídas pelo GenSeed-HMM eram mais longas do que as obtidas por montagens globais (Figura 10 e seção 4.3.3). Isso se deve principalmente às rotinas de otimização (seção 3.1.2) para montagem progressiva, e do critério altamente rigoroso para seleção das leituras e montagem das sequências. Dessa maneira, o GenSeed-HMM pode funcionar como uma ferramenta auxiliar em projetos genômicos, visando ajudar no fechamento de lacunas e extensão de regiões de sequências, desde obviamente que haja cobertura física de leituras.

Considerando-se a versatilidade do programa GenSeed-HMM e sua ampla validação, confirmada através de reconstruções realizadas em um amplo conjunto de bases de dados e diferentes sementes, podemos assumir que a atual versão do programa se encontra em um grau de desenvolvimento bastante maduro. Dessa maneira, a reconstrução de genomas de novos vírus e genes de organismos extintos, são tarefas que podem se beneficiar do emprego do GenSeed-HMM. Da mesma forma, projetos de sequenciamento genômico podem ser auxiliados por esse programa, seja na reconstrução de famílias de genes específicos, seja como ferramenta acessória no fechamento de sequências. Finalmente, o GenSeed-HMM requer recursos computacionais modestos, o que permite seu amplo uso tanto por grupos de pesquisa com infraestrutura computacional grande ou pequena, em uma ampla variedade de aplicações.

6 CONCLUSÕES

- Foi desenvolvido o GenSeed-HMM, um programa para a montagem progressiva de sequências a partir do uso de sementes de sequências nucleotídicas, proteicas e de HMMs de perfis;
- o programa permite utilizar dados de sequenciamento de diferentes plataformas, incluindo leituras do tipo Sanger, Illumina e 454;
- foi desenvolvido o Seeder, ferramenta gráfica para a geração assistida de arquivos de configuração do GenSeed-HMM;
- testes de validação demonstraram que o GenSeed-HMM pode ser aplicado de forma eficiente na reconstrução de loci específicos em bases de dados originadas de organismos procariotos, eucariotos e de amostras metagenômicas;
- o GenSeed-HMM permite reconstruir famílias de genes codificadores de proteínas, a partir de uma semente de HMM de perfil;
- o GenSeed-HMM abre a perspectiva de se poder realizar prospecções automáticas de novos genomas virais e do levantamento da diversidade biológica a partir de dados metagenômicos.

REFERÊNCIAS*

- Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The genome sequence of *Drosophila melanogaster*. *Science*. 2000;287(5461):2185-95.
- Agoni L, Golden A, Guha C, Lenz J. Neandertal and Denisovan retroviruses. *Current Biology*. 2012;5;22(11):R437-8.
- Allander T, Tammi M, Eriksson M, Bjerkner A, Tiveljung-Lindell A, and Andersson B. Cloning of a human parvovirus by molecular screening of respiratory tract samples. *Proceedings of the National Academy of Sciences*. 2005;102:12891-6.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 1997;25(17):3389-402.
- Amann R, Ludwig W, Schleifer K. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiology and Molecular Biology Reviews*. 1995;59:143-69.
- Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. *Nature*. 2011;473(7346):174-80.
- Blaser MJ. Epidemiologic and clinical features of *Campylobacter jejuni* infections. *Journal Infectious Diseases*. 1997;176(Suppl 2):S103-5. Review.
- Berezikov E, Plasterk RH, Cuppen E. GENOTRACE: cDNA-based local GENOME assembly from TRACE archives. *Bioinformatics*. 2002;18(10):1396-7.
- Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, McPhee JB, DeWitte SN, Meyer M, Schmedes S, Wood J, Earn DJ, Herring DA, Bauer P, Poinar HN, Krause J. A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature*. 2011;478(7370):506-10.
- Breitbart, M and Rohwer, F. Here a virus, there a virus, everywhere the same virus? *TRENDS in Microbiology*. 2005;13:278–84.
- Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Research*. 2008;(5):810-20.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.
- Campbell KL, Roberts JE, Watson LN, Stetefeld J, Sloan AM, Signore AV, Howatt JW, Tame JR, Rohland N, Shen TJ, Austin JJ, Hofreiter M, Ho C, Weber RE, Cooper A. Substitutions in woolly mammoth hemoglobin confer biochemical properties adaptive for cold tolerance. *Nature Genetics*. 2010;42(6):536-40.
- C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*. 1998;282(5396):2012-8.

*De acordo com:

International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to *Biomedical Journal*: sample references. Available from: <http://www.icmje.org> [2011 Jul 15].

Chen EC, Yagi S, Kelly KR, Mendoza SP, Tarara RP, Canfield DR, Maninger N, Rosenthal A, Spinner A, Bales KL, Schnurr DP, Lerche NW, Chiu CY. Cross-species transmission of a novel adenovirus associated with a fulminant pneumonia outbreak in a new world monkey colony. *PLoS Pathogens*. 2011;7(7):e1002155.

Chevreur B, Wetter T, Suhai S. Genome sequence assembly using trace signals and additional sequence information. *Computer science and biology: Proceedings of the German Conference on Bioinformatics*. 1999;99:45-56.

COMAV Institute. SFF extract. [página da internet]. Disponível em: <http://bioinf.comav.upv.es/sff_extract/>. [2012 Jun 23].

Compeau PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*. 2011;29(11):987-91.

Djikeng A, Kuzmickas R, Anderson NG, Spiro DJ. Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One*. 2009;4(9):e7264.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*. 2008;36(16):e105.

Dong Q, Wilkerson MD, Brendel V. Tracembler - software for in-silico chromosome walking in unassembled genomes. *BMC Bioinformatics*. 2007;8:151.

Eddy SR. Hidden Markov models. *Current Opinion in Structural Biology*. 1996;6(3):361-5.

Eddy SR. Profile hidden Markov models. *Bioinformatics*. 1998;14(9):755-63.

Eddy SR. A new generation of homology search tools based on probabilistic inference. *Genome Informatics*. 2009;23(1):205-11.

Eddy SR. HMMER User's Guide – Biological sequence analysis using profile hidden Markov models.[Manual]. Ashburn; Janelia Farm Research Campus; 2010.

EMBOSS Package. [ftp]. Disponível em <<ftp://emboss.open-bio.org/pub/EMBOSS/>>. [2012 Jun 23].

Ewing B, Green P: Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*. 1998;8:186-94.

Ewing B, Hillier L, Wendl M, Green P: Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*. 1998;8:175-85.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 1995;269(5223):496-512.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. Life with 6000 genes. *Science*. 1996;274(5287):563-7.

Gonzaga-Jauregui C, Lupski JR, Gibbs RA. Human genome sequencing in health and disease. *Annual Review of Medicine*. 2012;63:35-61.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. *Science*. 2010;328(5979):710-22.

Hannon Lab. FASTX TOOLKIT [página da internet]. Disponível em: <http://hannonlab.cshl.edu/fastx_toolkit/download.html>. [2012 Jun 23].

Harwich MD, Jr. Alves JM, Buck GA, Strauss JF 3rd, Patterson JL, Oki AT, Girerd PH, Jefferson KK. Drawing the line between commensal and pathogenic *Gardnerella vaginalis* through genome analysis and virulence studies *BMC Genomics*. 2010;11:375.

Hernandez D, François P, Farinelli L, Osteras M, Schrenzel J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research*. 2008;(5):802-9.

Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Research*. 1999;9(9):868-77.

Ho SY, Gilbert MT. Ancient mitogenomics. *Mitochondrion*. 2010;10(1):1-11.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*. 2007;8(7):R143.

Huynen L, Millar CD, Lambert DM. Resurrecting ancient animal genomes: the extinct moa and more. *Bioessays*. 2012. doi: 10.1002/bies.201200040.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921.

Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012;9:357-9.

Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology*. 1994;4;235(5):1501-31.

Kunin V, Copeland A, Lapidus A, Mavromatis K, and Hugenholtz P. A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews*. 2008;72(4):557-78.

Lauck M, Hyeroba D, Tumukunde A, Weny G, Lank SM, Chapman CA, O'Connor DH, Friedrich TC, Goldberg TL. Novel, divergent simian hemorrhagic fever viruses in a wild Ugandan red colobus monkey discovered using direct pyrosequencing. *PLoS One*. 2011;22;6(4):e19056.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*. 2010;20(2):265-72.

Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B, Yang B, Fan W. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Briefings in Functional Genomics*. 2011;11(1):25-37.

Mardis ER. Anticipating the 1,000 dollar genome. *Genome Biology*. 2006;7(7):112.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005; 15;437(7057):376-80.

Maxam AM, Gilbert W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of United States of America*. 1977;74(2):560-4.

Metzker ML. Emerging technologies in DNA sequencing. *Genome Research*. 2005;15(12):1767-76.

Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, et al. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*. 2008;20;456(7220):387-90.

Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics*. 2010;95(6):315-27.

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F and Marshall D. Tablet next generation sequence assembly visualization. *Bioinformatics*. 2010;26(3), 401-2.

Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology*. 2012;2(1):63-77.

Nagarajan N, Cook C, Di Bonaventura M, Ge H, Richards A, Bishop-Lilly KA, DeSalle R, Read TD, Pop M. Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. *BMC Genomics*. 2010;16;11:242.

National Center for Biotechnology Information. Sequence Read Archive [página da internet]. Disponível em: <<http://www.ncbi.nlm.nih.gov/sra>>. [2012a Jun 23].

National Center for Biotechnology Information. Sequence Read Archive [página da internet]. Disponível em: <<http://trace.ncbi.nlm.nih.gov/Traces/sra/>>. [2012b Jun 23].

National Center for Biotechnology Information. Base de dados de *C. jejuni* [página da internet]. Disponível em: <<http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=viewer&m=data&s=viewer&run=ERR019412>>. [2012c Jun 23].

National Center for Biotechnology Information. Base de dados de *E. coli* [página da internet]. Disponível em: <<http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=viewer&m=data&s=viewer&run=ERR019412>>. [2012d Jun 23].

National Center for Biotechnology Information. Base de dados de *D. melanogaster* [página de internet]. Disponível em: <<http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=viewer&m=data&s=viewer&run=SRR063698>>. [2012e Jun 23].

National Center for Biotechnology Information. Base de dados de *P. vivax* [página de internet]. Disponível em: <<http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=viewer&m=data&s=viewer&run=ERR023042>>. [2012f Jun 23].

National Center for Biotechnology Information. Base de dados de *D. japonicus* [página de internet]. Disponível em: <<http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=viewer&m=data&s=viewer&run=ERR024711>>. [2012g Jun 23].

National Center for Biotechnology Information. Base de dados de microbioma humano. [ftp]. Disponível em: <<ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX019/SRX019702>>. [2012h Jun 23].

National Center for Biotechnology Information. Base de dados de mamute lanoso extinto. [página de internet]. Disponível em: <<http://www.ncbi.nlm.nih.gov/sra?term=Mammuthus%20primigenius>>. [2012i Jun 23].

Noonan JP, Hofreiter M, Smith D, Priest JR, Rohland N, Rabeder G, Krause J, Detter JC, Pääbo S, Rubin EM. Genomic sequencing of Pleistocene cave bears. *Science*. 2005;309(5734):597-9.

Oliveira AL. GenSeed-HMM: desenvolvimento de uma plataforma para reconstrução de sequências e sua aplicação em dados de sequenciamento de nova geração. Dissertação (Mestrado em Biologia da Relação Patógeno-Hospedeiro). São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2012.

Palacios G, Druce J, Du L, Tran T, Birch C, Briesse T, Conlan S, et al. A new arenavirus in a cluster of fatal transplant-associated diseases. *New England Journal of Medicine*. 2008;358(10):991-8.

Paszkiwicz K, Studholme DJ. De novo assembly of short sequence reads. *Briefings in Bioinformatics*. 2010;11(5):457-72.

Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, et al. The NIH Human Microbiome Project. *Genome Research*. 2009;19(12):2317-23.

Pevzner PA, Tang H. Fragment assembly with double-barreled data. *Bioinformatics*. 2001;17(Suppl 1):S225-33.

Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of United States of America*. 2001;14;98(17):9748-53.

Peng Y, Leung HC, Yiu SM, Chin FY. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics*. 2011;27(13):i94-101.

- Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, Rampp M, Miller W, Schuster SC. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*. 2006;20;311(5759):392-4.
- Pop M. Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics*. 2009;10(4):354-66.
- Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. *Trends in Genetics*.(2008);24(3):142-9.
- Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD. The Pfam protein families database. *Nucleic Acids Research*. 2012;40 (Database issue):D290-301.
- Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al.. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010;464(7285):59-65.
- Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, Habib N, et al. Comparative functional genomics of the fission yeasts. *Science*. 2011;20;332(6032):930-6.
- Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyren P. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*. 1996;1;242(1):84-9.
- Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 2011;20;475(7356):348-52.
- Rothberg JM, Leamon JH. The development and impact of 454 sequencing. *Nature Biotechnology*. 2008;26(10):1117-24.
- Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. Artemis: sequence visualization and annotation. *Bioinformatics*, 2000. 16(10):944-5.
- Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*. 1977;24;265(5596):687-95.
- Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*. 1975;25;94(3):441-8.
- Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of United States of America*. 1977;74(12):5463-7.
- Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! *Genome Biology*. 2011;12(8):125.

Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. *Genome Res.* 2010;20(9):1165-73.

Schuster SC. Next-generation sequencing transforms today's biology. *Nature Methods.* 2008;5(1):16-8.

Schoenfeld T, Liles M, Wommack K, Polson S, Godiska R, Mead D. Functional viral metagenomics and the next generation of molecular tools. *Trends in Microbiology.* 2009;18:20-9.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. *Genome Research.* 2009;19(6):1117-23.

Sobreira TJ, Gruber A. Sequence-specific reconstruction from fragmentary databases using seed sequences: implementation and validation on SAGE, proteome and generic sequencing data. *Bioinformatics.* 2008;24;(15):1676-80.

Stoneking M, Krause J. Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics.* 2011;12(9):603-14.

Tk. Perl Tk [página da internet]. Disponível em <<http://www.perltk.org/>> [2012 Jun 23].

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. *Nature.* 2007;449(7164):804-10.

Whitley, R. The new age of molecular diagnostics for microbial agents. *New England Journal of Medicine.* 2008;358:988-9.

Wilson RJ, Williamson DH. Extrachromosomal DNA in the Apicomplexa. *Microbiology and Molecular Biology Reviews.* 1997;61(1):1-16.

Wooley J, Godzik A, and Friedberg I. A Primer on Metagenomics. *PLoS Computational Biology.* 2010;6:e1000667.

Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, et al. The genome of the fire ant *Solenopsis invicta*. *Proceedings of the National Academy of Sciences of United States of America.* 2011;108(14):5679-84.

Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008;18(5):821-9.

Zerbino, DR. Genome assembly and comparison using de Bruijn Graphs . [Ph. D. thesis]. Hixton: University of Cambridge; 2009.