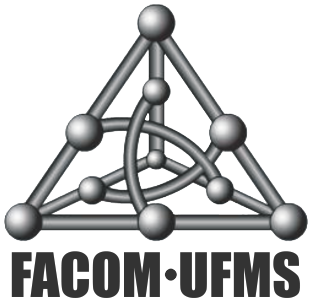
Remontagem de genomas: uma avaliação das técnicas *de novo* e por comparação

Dissertação de Mestrado

Sérgio Ronaldo Alves de Sousa Júnior

Orientação: Prof. Dra. Luciana Montera Cheung

Área de concentração: Biologia Computacional



Faculdade de Computação
Universidade Federal de Mato Grosso do Sul

Campo Grande, Agosto de 2012

Resumo

As técnicas de sequenciamento de nova geração são alternativas ao método Sanger de sequenciamento. Essas técnicas produzem um grande volume de *reads* por ciclo de execução. Muitas vezes, esses *reads* precisam ser remontados seguindo alguma técnica de remontagem a fim de se obter um genoma completo. Abordamos neste trabalho as técnicas de remontagem *de novo* e remontagem por comparação, avaliando os resultados produzidos por ambas. Como a anotação de um genoma é um processo que depende da remontagem do genoma, deve ser criteriosa a escolha de uma ferramenta para tal fim.

Palavras-chave: remontagem de genomas; remontagem *de novo*; remontagem por comparação; mapeamento de *reads*.

Abstract

The Next-Generation Sequencing techniques are an approach for reads assembly and differs from Sanger method. Those techniques produces a big amount of reads per execution. In many occasions, those reads must be assembled to form a whole genome. We give an approach about Comparative-Assembling and *de novo* Assembling in this project. Besides, we compare the results from both assembly techniques. Its important to follow some rules on choosing a software to assembly reads, giving the important of this step to determine a whole genome.

Sumário

Capítulo 1

Introdução

O sequenciamento de DNA é o processo que consiste na determinação de sequências de nucleotídeos. As técnicas NGS, do inglês *Next-Generation Sequencing*, são técnicas de sequenciamento de pequenos fragmentos de DNA (*reads*) alternativas ao método Sanger de sequenciamento. Essas técnicas de nova geração produzem um grande volume de fragmentos por ciclo de execução. Os pequenos fragmentos precisam ser remontados a fim de que seja obtido o genoma inteiro.

Abordamos neste trabalho as técnicas de remontagem *de novo* e remontagem por comparação. A primeira técnica alinha os *reads* a um genoma chamado genoma-referência enquanto que a segunda baseia-se apenas nas informações contidas nos próprios *reads*. Para a técnica *de novo* foram utilizadas as ferramentas SSAKE e SOAPdenovo. Já para a técnica por comparação, as ferramentas utilizadas foram AMOScmp, Blat e Bowtie. Todas as ferramentas foram submetidas a um conjunto de testes, composto por dados reais e dados simulados.

Os *reads* de dados simulados foram gerados pelo simulador GenFrag e variam na cobertura e tamanho dos fragmentos, bem como na taxa de erro da geração dos mesmos. Os dados simulados mostram que os resultados produzidos pelas ferramentas sofrem variações conforme as características dos *reads* submetidos como entrada mudam. A corretude dos resultados com dados simulados chegou a 98% e 86,43% para ferramentas de remontagem por comparação e ferramentas de remontagem *de novo*, respectivamente.

Os dados reais refletem situações reais que as ferramentas são submetidas. Esses dados são compostos de *reads* de projetos reais contidos no GenBank. Os resultados com esses dados confirmam tendências evidenciadas nos resultados com dados simulados como: quanto maior a cobertura dos *reads* a serem remontados, menor o número de *contigs* formados; quanto menor o número de *contigs* formados, maiores eles são em comprimento; dentre outras.

A remontagem dos *reads* para a obtenção de um genoma não é um processo simples em termos computacionais, visto que um grande volume de fragmentos, representados por sequências de caracteres escritas em um alfabeto específico, precisam ser alinhados uns aos outros. O alinhamento de sequências é um problema recorrente na Biologia Com-

putacional e demanda algoritmos robustos para serem executados em tempo satisfatório utilizando os recursos disponíveis.

Como a anotação é um processo que depende da remontagem do genoma, deve ser criteriosa a escolha de uma ferramenta para tal fim. As informações sobre os *reads* submetidos como entrada para as ferramentas de remontagem são de extrema importância para tal escolha. Logo, esse trabalho oferece diretrizes na escolha de uma ferramenta de remontagem de um genoma com base nos resultados das execuções das ferramentas com dados reais e simulados.

O texto esta organizado da seguinte forma. No Capítulo 2 estão descritos os conceitos básicos sobre Biologia Computacional necessários para o entendimento do trabalho. O Capítulo 3 descreve as técnicas de remontagem abordadas neste trabalho: remontagem por comparação e remontagem *de novo*, nesta ordem. A metodologia utilizada para os estudos deste trabalho, bem como os critérios de avaliacão dos resultados estão descritos no Cápítulo 4. A análise sobre os resultados das simulações e execuções é feita no Capítulo 5 e uma breve conclusão sobre este estudo é realizada no Capítulo 6. Como complemento de entendimento do conteúdo deste trabalho foram escritos dois Apêndices: A e B. Por fim, este texto encerra-se com as referências utilizadas.

Capítulo 2

Biologia Molecular e sequenciamento de DNA

Este capítulo traz, de forma breve, um apanhado sobre conceitos de Biologia Molecular necessários à contextualização e compreensão deste trabalho de pesquisa. Sua escrita foi baseada principalmente nas referências [29], [22] e [19].

2.1 Ácidos nucléicos

Os organismos vivos são compostos por células, as quais, por sua vez, são constituídas por vários elementos, entre eles os ácidos nucléicos, que podem ser de dois tipos: ácido ribonucléico (RNA) e ácido desoxirribonucléico (DNA). O RNA e o DNA constituem-se da união, em sequência, de nucleotídeos, formando cadeias. Um nucleotídeo é formado por três componentes básicos:

- um açúcar (Pentose);
- uma base nitrogenada (Base); e
- um grupo fosfato (P).

Uma visão geral da estrutura de um nucleotídeo é apresentada na Figura 2.1.

As bases nitrogenadas existentes são Adenina (A), Citosina (C), Guanina (G), Timina (T) e Uracila (U). Apenas as bases nitrogenadas A, T, C e G estão presentes no DNA, enquanto as bases A, U, C e G compõem o RNA.

O DNA, que tem função de codificar as informações genéticas dos indivíduos, é uma molécula de cadeia (ou fita) dupla, onde os nucleotídeos de cada cadeia são ligados entre si por pontes de hidrogênio entre pares específicos de bases nitrogenadas. Os pares de bases, ou bp (*base pairs*) possíveis de se ligar são ditos complementares, e ligam A com T e C com G, utilizando 2 e 3 pontes de hidrogênio, respectivamente, em cada ligação. Ligadas

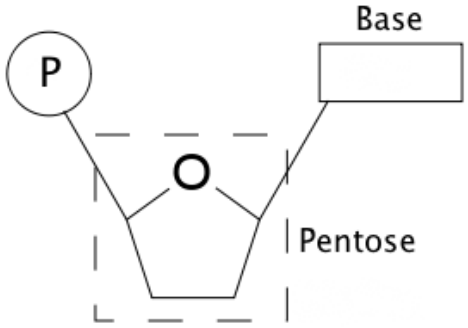


Figura 2.1: Visão geral da estrutura de um nucleotídeo.

por pontes de hidrogênio, as duas fitas complementares do DNA estão estruturadas em uma forma chamada dupla hélice (como pode ser visto na Figura 2.2).

Cada uma das fitas do DNA possui uma orientação específica, que é determinada pelas extremidades livres de cada cadeia de nucleotídeos. Uma extremidade é dita livre quando esta pode ser ligada, por meio de um átomo de carbono, a um outro nucleotídeo. O açúcar presente na molécula de DNA é chamado de desoxirribose e possui 5 carbonos (e por isso é chamado de pentose) em sua estrutura, referidos como: C_1, C_2, C_3, C_4 e C_5. A Figura 2.3 detalha a estrutura de um nucleotídeo, ilustrando os elementos de sua composição, bem como as ligações entre eles.

Uma das fitas do DNA, chamada de fita +, possui orientação $5' \rightarrow 3'$, o que indica que sua cadeia de nucleotídeos se inicia pelo carbono C_5 livre e termina com o carbono C_3 livre. A outra fita do DNA, dita fita -, tem orientação $3' \rightarrow 5'$, contrária à orientação da fita +. Dessa forma, além de complementares, as fitas também são chamadas de fitas anti-paralelas. Por convenção, uma molécula de DNA pode ser representada apenas pela cadeia de nucleotídeos que tem orientação $5' \rightarrow 3'$ (Veja Figura 2.2).

A ordem entre os nucleotídeos de uma cadeia representa, de fato, uma fita de DNA (ou RNA). Deste modo, por exemplo, a molécula de DNA, mostrada na Figura 2.2, pode ser representada apenas pela sequência de bases da sua fita +: ATGCTGC.

Além da ligação por pontes de hidrogênio, que ligam nucleotídeos de fitas complementares, dois nucleotídeos de uma mesma fita se ligam através de ligações fosfodiésteres. Essas são ligações que ocorrem entre o carbono C_3 da pentose de um nucleotídeo e o grupo fosfato do nucleotídeo seguinte. As enzimas DNA polimerase, que participam da síntese da molécula de DNA, são responsáveis pela catalisação das ligações fosfodiésteres. A Figura 2.4 mostra uma ligação fosfodiéster entre dois nucleotídeos.

Embora composto apenas por um único filamento, a estrutura do RNA é similar a do DNA. Como já foi dito, o RNA não possui Timina no seu possível conjunto de bases, mas a Uracila. As diferenças e semelhanças entre as moléculas de DNA e RNA podem ser vistas na Figura 2.5. O RNA desempenha funções variadas dentro da célula. Essa

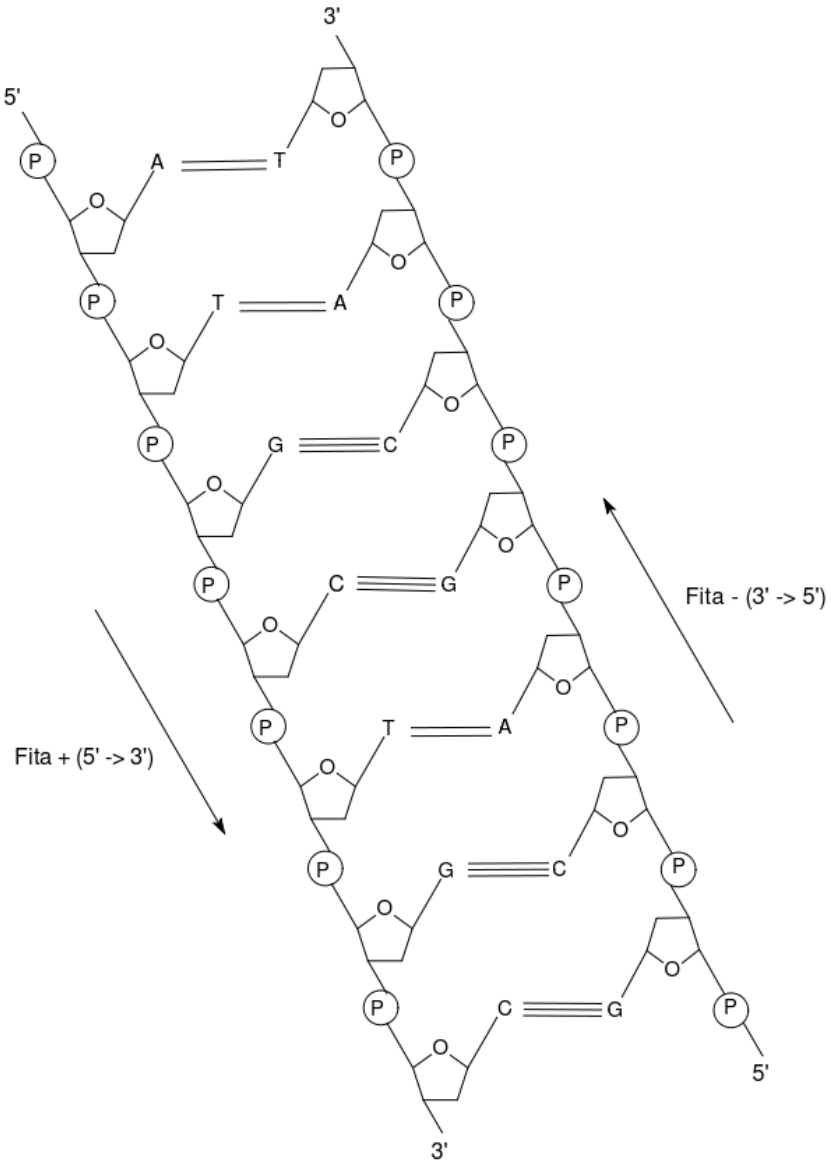


Figura 2.2: Fragmento de uma molécula DNA.

diversidade de papéis reflete-se na grande variedade de RNAs existentes. De acordo com as funções que desempenham, os RNAs podem ser divididos em duas classes principais: os RNAs codificantes (cRNAs) e os RNAs não-codificantes (ncRNAs). Segundo [32], existe apenas um tipo de RNA codificante e a ele dá-se o nome de RNA mensageiro (mRNA). O mRNA contém a informação para codificação de proteínas e é uma molécula sintetizada a partir do DNA por um mecanismo complexo do qual participam, entre outros, enzimas do tipo RNA polimerase. Sobre os RNAs não-codificantes, os mais conhecidos são: o RNA transportador (tRNA), responsável pelo transporte de aminoácidos (que são elementos constituintes das proteínas) e o RNA ribossômico (rRNA), que tem papel estrutural e formam os ribossomos, que são estruturas citoplasmáticas que compõem o mecanismo de síntese de proteínas.

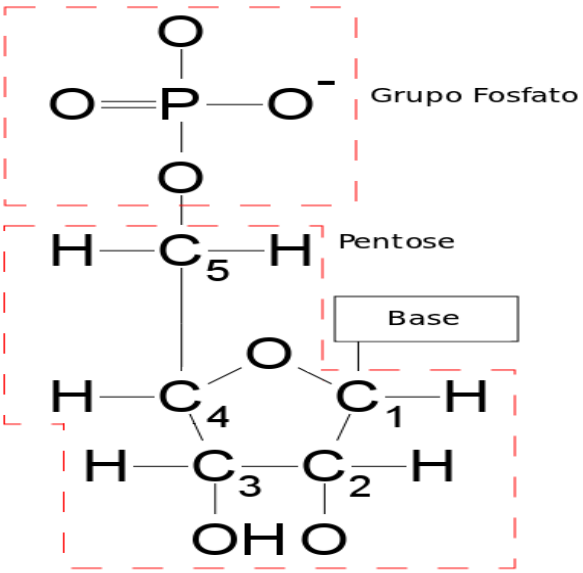


Figura 2.3: Visão detalhada da estrutura de um nucleotídeo.

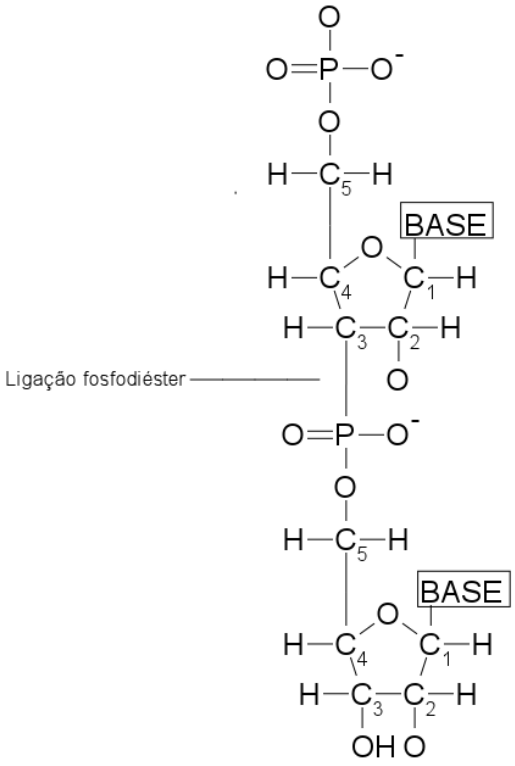


Figura 2.4: Dois nucleotídeos de uma mesma fita são ligados por uma ligação fosfodiéster.

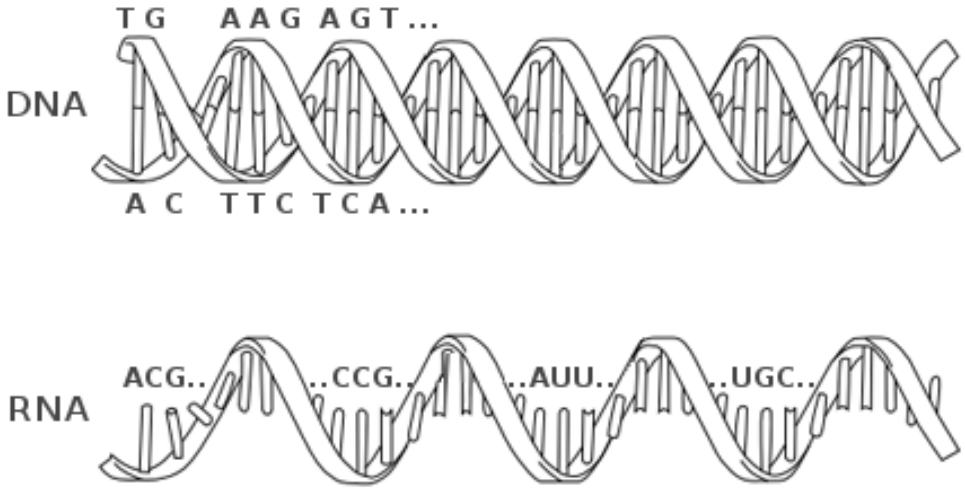


Figura 2.5: Estruturas de uma molécula de DNA (dupla hélice) e uma molécula de RNA.

2.2 Genes, proteínas e síntese de proteínas

Ao longo da molécula de DNA de um organismo, existem informações organizadas em unidades chamadas genes, responsáveis pela produção de macromoléculas chamadas proteínas. A síntese de proteínas a partir dos genes é regida pelo mecanismo chamado Dogma Central da Biologia Celular. Este mecanismo estabelece que o fluxo da informação genética é: "DNA faz RNA, que faz proteína que, por sua vez, facilita os dois passos prévios bem como a replicação do DNA". A síntese de proteínas envolve dois processos: Transcrição e Tradução. O DNA é transcrito em RNA com atuação de enzimas RNA polimerases, capazes de identificar a posição de um gene dentro da molécula de DNA e iniciar a transcrição a partir do ponto identificado. Após a síntese da molécula de RNA, esta é traduzida para que uma proteína seja produzida. A tradução é feita considerando triplas de nucleotídeos chamadas códons, onde cada tripla traduzida corresponde a um aminoácido. São conhecidos 20 aminoácidos conhecidos, cada um associado a um (ou mais) códon [19]. Os vinte aminoácidos, seus nomes, abreviaturas e seus códons correspondentes podem ser visualizados na Tabela 2.1.

Uma proteína é uma molécula formada pelo encadeamento de moléculas menores chamadas aminoácidos, formando uma estrutura denominada polímero. As proteínas possuem diferentes funções nos seres vivos, dentre elas: transporte de nutrientes, eliminação de resíduos tóxicos e construção de estruturas complexas como a parede celular. Algumas proteínas, chamadas enzimas, são responsáveis por catalisar, ou acelerar, reações químicas necessárias à vida.

Um aminoácido é formado por um carbono central, onde se liga um grupo amina, um

Tabela 2.1: Tabela de código genético. Os aminoácidos existentes e os códons correspondentes. Os códons de parada são indicados pelo caractere *.

Nome do Aminoácido	**Abreviatura**	**Códons correspondentes**
Alanina	Ala	GCU, GCC, GCA, GCG
Cisteína	Cys	UGU, UGC
Ácido Aspártico	Asp	GAU, GAC
Ácido Glutâmico	Glu	GAA, GAG
Fenilalanina	Phe	UUU, UUC
Glicina	Gly	GGU, GGC, GGA, GGG
Histidina	His	CAU, CAC
Isoleucina	Ile	AUU, AUC, AUA
Lisina	Lys	AAA, AAG
Leucina	Leu	UUA, UUG, CUU, CUC, CUA, CUG
Metionina	Met	AUG
Asparagina	Asn	AAU, AAC
Prolina	Pro	CCU, CCC, CCA, CCG
Glutamina	Gln	CAA, CAG
Arginina	Arg	CGU, CGC, CGA, CGG, AGA, AGG
Serina	Ser	UCU, UCC, UCA, UCG, AGU, AGC
Treonina	Thr	ACU, ACC, ACA, ACG
Valina	Val	GUU, GUC, GUA, GUG
Triptofano	Trp	UGG
Tirosina	Tyr	UAU, UAC
-	-	UAA*, UAG*, UGA*

grupo carboxila e uma cadeia lateral. Os aminoácidos diferem entre si pela estrutura da cadeia lateral, que pode variar entre um único átomo de hidrogênio até anéis carbônicos. Uma representação esquemática de um aminoácido pode ser vista na Figura 2.6.

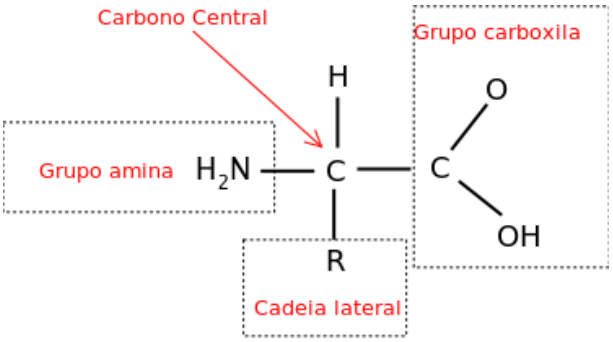


Figura 2.6: Estrutura esquemática de um aminoácido.

O mecanismo responsável pela síntese de uma proteína precisa realizar o reconhecimento do início da sequência de um gene dentro de uma molécula de DNA. O início de um gene é marcado por uma sequência específica no DNA que sinaliza esse início. Tal região é denominada região promotora. Após a localização do início do gene, a enzima

do tipo RNA polimerase copia a informação do gene, criando uma molécula de RNA complementar. Essa fita complementar é uma molécula especial chamada de pré-RNA mensageiro, ou pré-mRNA, que só se tornará RNA mensageiro maduro, ou mRNA, após a fase conhecida como *splicing*[1]. Dessa forma, o pré-mRNA possui uma sequência complementar a uma das fitas do DNA, exceto pela presença da base U no lugar da base T. Essa sequência descrita de passos chamamos de processo de transcrição.

Os genes dos seres eucariotos são compostos pelos íntrons e éxons: os íntrons são regiões eliminadas do pré-mRNA na fase *splicing*, ou seja, não codificarão proteínas; os éxons são regiões que permanecem no pré-mRNA após a fase de *splicing*. Após a eliminação dos íntrons, têm-se o mRNA maduro (ou somente mRNA) e, na última fase do processo de síntese, a molécula de mRNA é finalmente traduzida para uma proteína num processo denominado tradução.

A proteína é sintetizada em estruturas celulares chamadas ribossomos. Os ribossomos são compostos de proteínas e uma molécula de rRNA, funcionando como linhas de montagens de proteínas, onde as informações contidas no mRNA são lidas e traduzidas pelos tRNAs. Os tRNAs são moléculas responsáveis por efetuar a conexão entre códons e os aminoácidos correspondentes, efetuando, de fato, a tradução. Cada tRNA é composto por duas partes: uma possui afinidade química a um códon e a outra liga-se com o aminoácido correspondente a esse códon. Uma enzima catalisa a ligação entre os aminoácidos, parando somente quando um códon de parada é encontrado. O códon de parada é uma tripla específica de bases que encerra o processo de tradução. De forma simplificada, o procedimento de síntese de proteínas pode ser visto na Figura 2.7, retirada de [32].

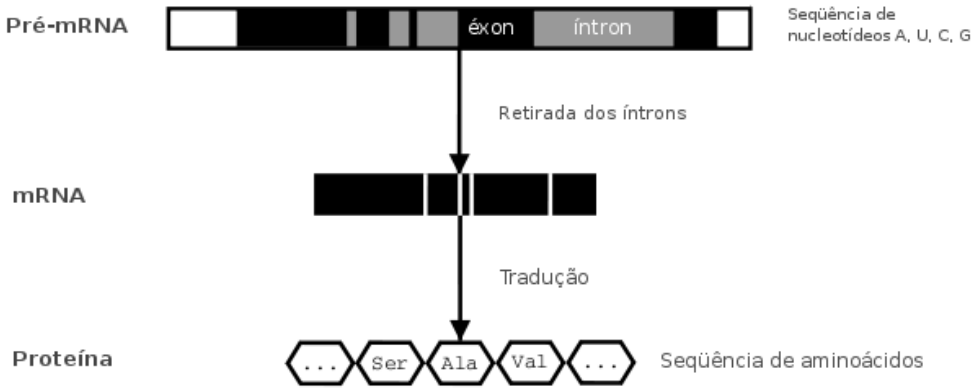


Figura 2.7: Síntese de proteínas.

2.3 Sequenciamento de DNA (RNA)

O sequenciamento de DNA consiste em um processo onde os nucleotídeos presentes em uma molécula de DNA são identificados, bem como a sua sequência. Este processo é execu-

[1]Na fase de *splicing* algumas regiões do pré-mRNA são eliminadas da molécula e as regiões mantidas compõem o mRNA maduro. Essa fase é específica dos organismos eucariontes, pois não existe pré-RNA nos organismos procariontes. Maiores detalhes sobre o mecanismo de *splicing* podem ser obtidos em [19].

tado por uma máquina denominada sequenciador. O processo de sequenciamento produz um arquivo contendo sequências de caracteres correspondentes às bases dos núcleotídeos que formam os fragmentos sequenciados.

O método Sanger é uma técnica clássica e pioneira de sequenciamento de DNA [33]. Este método pode ser utilizado no sequenciamento de moléculas de DNA de no máximo 200 mil pares de bases (bp). Entretanto, o processo de sequenciamento é capaz apenas de determinar a composição de pequenos fragmentos de DNA, o que torna o sequenciamento de todo o conteúdo genético de um organismo um processo complexo e composto por várias etapas.

Novas técnicas de sequenciamento começaram a surgir por volta de 2005. Essas novas tecnologias, conhecidas como tecnologias de nova geração (*Next-Generation Sequencing*, ou simplesmente NGS), são capazes de produzir, em menos tempo e com menor custo, um volume bem maior de informação se compararadas ao método Sanger [24]. Entretanto, o tamanho dos fragmentos sequenciados no método Sanger têm tamanho de até 900 bp, enquanto que os gerados pelos sequenciadores NGS estão entre 25 bp e 600 bp.

As seções 2.3.1 e 2.3.2 a seguir apresentam, respectivamente, uma descrição do método clássico de sequenciamento (Sanger) e de algumas das técnicas de *Next-Generation Sequencing*.

2.3.1 Método Sanger de sequenciamento

A maior parte do conteúdo desta seção foi compilada da seção 4.1 do livro referência [21] e do artigo [36].

Na década de 70 surgiram os primeiros procedimentos eficientes e rápidos para sequenciamento de moléculas de DNA: o método do término de cadeia, posteriormente conhecido como método Sanger, e o método da degradação química. Mesmo com a popularidade de ambos, o segundo método causava danos aos pesquisadores devido ao uso de substâncias tóxicas em seus procedimentos. Além disso, os procedimentos do método Sanger se mostraram mais fáceis de automatizar, o que difundiu ainda mais o seu uso.

O método Sanger consiste na síntese de um conjunto de cadeias a partir do fragmento de DNA a ser sequenciado. Em linhas gerais o processo segue os seguintes passos:

1. O fragmento duplo de DNA é desnaturado. O processo de desnaturação consiste no aquecimento da molécula de fita dupla de DNA (a temperatura de aproximadamente 94 °C) a fim de que elas se soltem uma da outra, formando cadeias simples, ditas cadeias *templates*. O objetivo é realizar o sequenciamento destes *templates*.

2. Para cada cadeia *template*, um *primer* é posicionado, marcando o início da síntese de uma cadeia de DNA complementar ao *template*. Um *primer* é um oligonucleotídeo, ou uma pequena cadeia de nucleotídeos, complementar a uma região do *template* e é responsável por marcar o início do sequenciamento do *template*, pois é a partir dele, e tendo como molde o *template*, que a enzima DNA polimerase é capaz de realizar o processo de síntese de uma nova cadeia de DNA.

3. São realizadas as reações para o sequenciamento das *templates*.

No passo 1, após a desnaturação, as moléculas de fita simples do DNA original são clonadas em vetores de clonagem apropriados a fim de que milhares de cópias do DNA original sejam produzidas (ver Figura **??**). Em seguida, no passo 2, *primers* são colocados juntos com o DNA alvo a fim de iniciar o sequenciamento.

O sequenciamento de fragmentos do DNA *template*, mencionados no passo 3, é possível pela utilização de nucleotídeos que não possuem o grupo 3'-OH, os chamados ddNTPs (didesoxirribonucleosídeos trifosfatados, veja Figura 2.8). São realizadas 4 reações distintas contendo:

- Adenina (dATP), Citosina (dCTP), Guanina (dGTP) e ddTTP;
- Adenina (dATP), Citosina (dCTP), Timina (dTTP) e ddGTP;
- Adenina (dATP), Guanina (dGTP), Timina (dTTP) e ddCTP;
- Citosina (dCTP), Guanina (dGTP), Timina (dTTP) e ddATP.

Quando um ddNTP é adicionado a uma cadeia de DNA sendo sintetizada, o processo de síntese pela DNA polimerase é interrompido devido a inexistência do grupo 3'-OH no último nucleotídeo (ddNTP) incorporado, resultando em uma cadeia truncada em relação ao *template*. A Figura 2.9 (construida a partir dos conceitos apresentados no livro referência [21]) mostra um esquema das quatro reações de sequenciamento descritas.

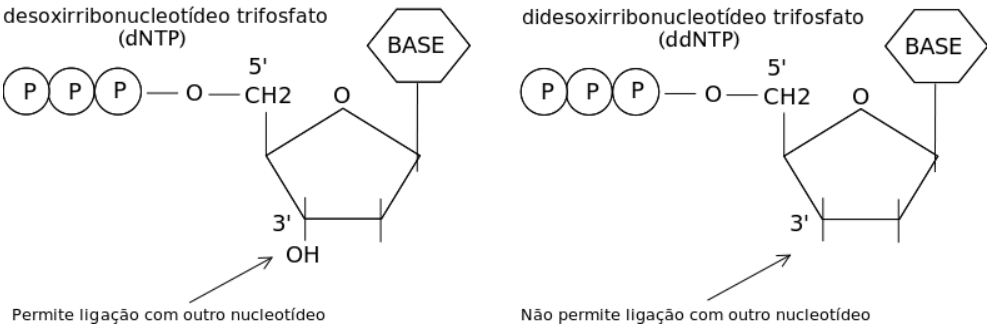


Figura 2.8: Diferença entre um ddNTP e um dNTP.

O procedimento de sequenciamento descrito é realizado por máquinas sequenciadoras que não geram diretamente um arquivo contendo as bases correspondentes às cadeias de nucleotídeos sequenciados e sim gráficos de cores chamados de cromatogramas.

Um cromatograma (Ver Figura 2.10) é o registro gráfico de uma análise por um método cromatográfico. A cromatografia é uma técnica eminentemente quantitativa que tem por princípio a separação de componentes de uma mistura. O princípio básico da quantificação é que a área dos picos registradas no cromatograma é proporcional à massa do composto injetada. Logo, é fundamental para a confiabilidade da análise que as áreas dos picos sejam medidas com a máxima precisão possível. Assim, em um cromatograma é possivel

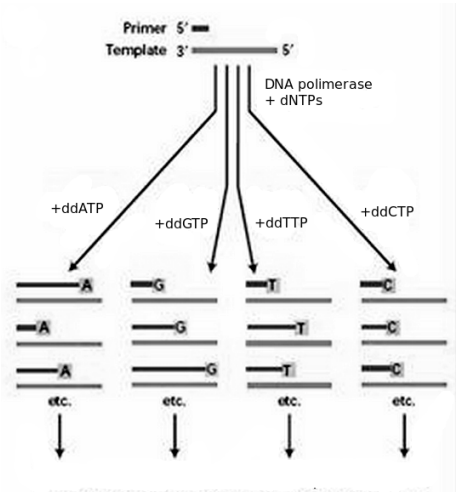


Figura 2.9: Conjunto de cadeias truncadas resultantes do processo de síntese. São realizadas 4 reações de síntese, uma para cada tipo de ddNTP, juntamente com os 4 dNTPs (dATP, dTTP, dCTP, dGTP). A partir do momento que a DNA polimerase insere um ddNTP, a síntese não é mais possível. Dessa forma, o conjunto de cadeias truncadas resultantes das 4 reações é formado por fragmentos que foram truncados em posições e tipos de nucleotídeos aleatórios em relação à cadeia original.

visualizar a separação dos componentes da mistura, bem como deteminar a concentração de cada susbstância pelas áreas dos picos.

Nos cromatogramas gerados pelas máquinas sequenciadoras, cada tipo de nucleotídeo (A, C, T e G) é representado por uma curva colorida distinta das demais. O eixo X corresponde à sequência de nucleotídeos lidos. No eixo Y são encontrados os picos formados pelas curvas coloridas.

Após gerado o cromatrograma, um *software* de computador efetua sua leitura e determina a probabilidade de erro do sequenciamento de bases da molécula a partir dos picos formados pelas curvas coloridas. Em outras palavras, como mais de um pico pode ser formado para a leitura de um nucleotídeo, a cada nucleotídeo lido no cromatograma, o *software* de computador determina qual foi o nucleotídeo interpretado e a probabilidade de este não ter sido realmente o nucleotídeo sequenciado.

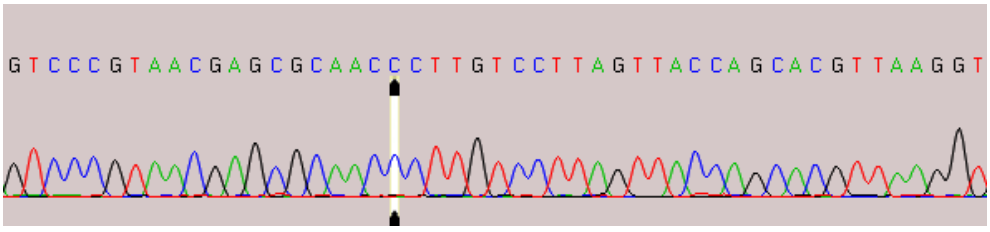


Figura 2.10: Exemplo de Cromatograma. Para representar os nucleotídeos A, C, T, G são usadas, respectivamente, as cores verde, azul, vermelha e preta. A linha branca indica um pico de curva azul e, logo, a base C é a base correspondente à posição indicada.

Usualmente, as ferramentas e arquivos de entrada para ferramentas referem-se à qua-

lidade de uma base. A relação entre a qualidade (Q) de uma base lida e a probabilidade (p) de ela estar errada é definida da seguinte forma:

$$Q = -10 \log_{10} p \tag{2.1}$$

Para uma base lida com probabilidade $p = 100\% = 1$ de estar errada, a qualidade desta base é $Q = 0$. Da mesma forma, para $p = 0,1$, $Q = 10$; para $p = 0,01$, $Q = 20$; para $p = 0,001$, $Q = 30$; e assim sucessivamente. Logo, é fácil notar que quanto menor a probabilidade de uma base lida estar errada, maior sua qualidade.

Denominamos cada pequeno fragmento do DNA original de *read*. Após todo o processo de sequenciamento e análise dos cromatogramas dos *reads* gerados pelo processo descrito anteriormente, as informações pertencentes a esses *reads* (bases que o compõem e suas respectivas qualidades) são armazenadas em dois arquivos de saída: um arquivo contém as informações sobre as bases que compõem cada *read* e outro as respectivas qualidades.

O formato padrão para o arquivo de saída das bases é o formato FASTA. Esse arquivo marca cada entrada de uma nova sequência com o sinal '>'. A identificação do *read* vem em seguida, na mesma linha. Na linha seguinte a sequência de nucleotídeos é escrita e pode ser quebrada em várias linhas, se necessário. Um exemplo de arquivo FASTA é mostrado na Figura 2.11. Arquivos deste tipo têm extensão .fasta.

```
>read1
CACAATGTGAAGTCTTATTTCATTCGTCTCAATGAATATTTTCCATAAAATTTGCATTGCATTG
AACAATTATTGATTTTTTTT
>read2
AAAAATAACGAAAAAACAAAGTTCATGGAGCAGGAAATTTACAATTTGGATCATTTAGAATTTA
GATGAATAATGAACCTTAATATAATAAGTATATGAGAAGTCATGGGCCCAGATTCAGAAC
>read3
AAACCACGTCAAATAATCAATTATGACGCAGGTATCGTATTAATTGATCTGCATCAACTTAACG
CAATACAAATCAGCGACACTGAATACGGGGCAACCTC
>read4
ACGTCGGCTGGAAATGTATTAAGATAATATCTGGAGATCAGAAGACTAACGTTCAATTTTATAC
ACTCAGTAAACATTTTCCACAAGACAATC
>read5
GAAATGTATTCATTTTCCACAAGATCAGAAGACTA
```

Figura 2.11: Conjunto de *reads* em um arquivo FASTA.

O arquivo que armazena a qualidade dos *reads* geralmente tem extensão QUAL. Um exemplo de arquivo QUAL para os nucleotídeos lidos no cromatograma da Figura 2.10, expressando a qualidade de um nucleotídeo sequenciado com números entre 0 e 99 (menor e maior qualidade, respectivamente), é o seguinte:

```
>id_read
80 93 92 45 50 13 99 87 87 92 31 48 54 59 61 78 83 94 92
91 82 65 99 27 35 93 88 82 89 63 77 69 83 94 96 91 63 52
```

58 59 79 21 16 8 11 10 26 36 17

O conteúdo armazenado nos arquivos não está livre de erros, pois a análise do cromatograma pode não ser exata. Para um nucleotídeo representado no cromatograma, duas ou mais linhas podem formar picos parcialmente ou totalmente sobrepostos, impossibilitando a determinação de qual é o real nucleotídeo sequenciado (ver exemplo na Figura 2.12). Pode ocorrer também duas ou mais leituras consecutivas do mesmo tipo de nucleotídeo, e por isso as curvas correspondentes a esses nucleotídeos formam uma sucessão de picos, cuja quantidade de cumes se confunde, dificultando a determinação da quantidade de nucleotídeos do mesmo tipo que foram realmente lidos (ver Figura 2.13).

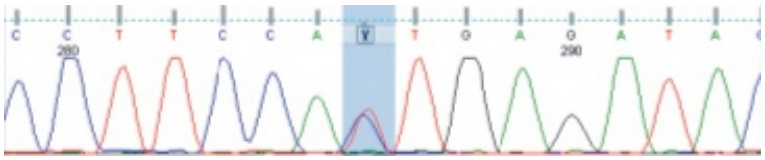


Figura 2.12: Sobreposição de curvas correspondentes a diferentes tipos de nucleotídeo no cromatograma. A linha azul corresponde ao nucleotídeo C e a linha vermelha corresponde ao nucleotídeo T. A área destacada no cromatograma mostra a sobreposição de picos impossibilitando a determinação do nucleotídeo sequenciado naquele ponto. Essa incerteza é representada, nesta figura, pelo caractere Y.

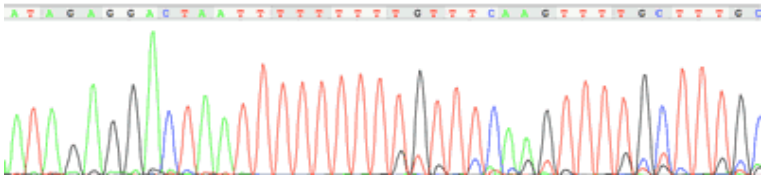


Figura 2.13: Picos consecutivos no cromatograma correspondentes ao mesmo tipo de nucleotídeo.

O método do término de cadeia, ou método Sanger, foi utilizado como guia em muitos projetos de sequenciamento de DNA, inclusive em iniciativas para sequenciar o genoma humano completo [31]. Entretanto, por ser um método caro e que consome muito tempo, além de muito trabalhoso, novas técnicas de sequenciamento surgiram com a promessa de solucionar esses problemas [30], as chamadas técnicas de nova geração.

2.3.2 *Next-Generation Sequencing* (NGS)

Os avanços nas técnicas de sequenciamento de DNA substituiram recentemente o método clássico Sanger e trouxeram como melhoria um aumento considerável no custo-benefício, principalmente no que diz respeito ao número de *reads* sequenciados por unidade de tempo. É válido lembrar que, por outro lado, as sequências provenientes de técnicas NGS possuem um comprimento menor se comparadas às sequências geradas pelo método Sanger. Os próximos parágrafos descrevem detalhes de algumas das técnicas de sequenciamento de DNA NGS. As principais referências utilizadas para elaboração desta seção foram [31], [39] e [30].

Algumas das tecnologias NGS disponíveis para uso no mercado atual incluem o 454, Illumina e o ABI. Uma comparação entre essas tecnologias e o método Sanger no que diz respeito ao tamanho dos *reads* e ao volume de dados produzidos estão sumarizadas na Tabela 2.2, adaptada de [31].

Tabela 2.2: Tamanho dos *reads* produzidos pelas tecnologias *Next-Generation* e pelo método Sanger.

Tecnologia	**Tamanho dos *reads* (bp)**	**Número de sequências geradas**
Sanger	≈1000	300K
454	100 - 600	1M - 1,5M
Illumina	20 - 100	10M - 15M
ABI	≈35	1M - 1,5M

Como pode ser observado na Tabela 2.2 as tecnologias NGS geram *reads* de tamanho bem menores que a metodologia Sanger.

Cada tecnologia possui suas peculiariades e procedimentos específicos. As subseções seguintes descrevem sucintamente as tecnologias NGS citadas na Tabela 2.2.

454

A tecnologia *Roche 454* foi a primeira NGS a aparecer no mercado, por volta de Outubro de 2005 [1]. Ao invés de usar clonagem de DNA por organismos (hospedeiros), a *Roche 454* usa um método de amplificação de DNA *in vitro* altamente eficiente conhecido como *emulsion PCR*.

Reação de Extensão por Polimerase, ou simplesmente PCR (*Polymerase Chain Reaction*), é um procedimento de replicação de cadeias de DNA *in vitro*. Uma cadeia de DNA desnaturada (de fita simples) é unida por complementaridade de bases a um *primer*. A enzima DNA polimerase é utilizada para a extensão do *primer* tendo como molde a molécula de DNA de fita simples. Assim, ao final do processo, uma nova cópia complementar ao molde é gerada.

A Reação de Emulsão PCR (*Emulsion PCR*) é um procedimento onde as moléculas de DNA são desnaturadas e copiadas rapidamente. Os fragmentos de DNA são misturados com gel de agarose[2] contendo sequências adaptadoras (*primers* aleatórios), que são complementares a trechos dos fragmentos a serem sequenciados. As moléculas de DNA são preparadas e agrupadas em conjuntos denominados gotículas de emulsão. Cada gotícula de emulsão é formada por uma estrutura de agarose ligada a um único fragmento de DNA e é capaz de gerar um grande número de cópias do DNA por meio de um ciclo de *Emulsion*

[2]A agarose é um polímero composto por subunidades de galactose que, quando dissolvida em água, torna-se um gel muito utilizado em diferentes experimentos de Biologia Molecular.

PCR [2]. As cópias do DNA são feitas por um processo que consiste em um grande número de reações do tipo *Pyrosequencing reactions*, procedimento que determina a ordem dos nucleotídeos pela síntese [31], onde cada base é identificada à medida que é adicionada à cadeia recém-formada de DNA. A essas cópias do DNA dá-se o nome de *templates*. Ao término do processo, moléculas de DNA são representadas por um gráfico chamado *Pyrogram*, correspondendo à ordem dos nucleotídeos que foram incorporados durante a reação de *Emulsion PCR*. Esta tecnologia é capaz de gerar 1M - 1,5M de sequências em uma execução de apenas 4 horas.

As informações contidas nos gráficos *Pyrogram* são também armazenadas em um arquivo texto. Esse arquivo é então o resultado final do sequenciamento contendo o conjunto de nucleotídeos e suas respectivas qualidades e tem extensão SFF.

De forma alternativa, a *Roche 454* também gera arquivos nos formatos FASTA e QUAL. Exemplos desses arquivos podem ser vistos no Apêndice A.1.

Illumina

A idéia principal dessa tecnologia é unir, duas-a-duas, várias moléculas de fita simples em uma única molécula de DNA. A tecnologia utilizada pelo Illumina não tem a etapa de clonagem para a amplificação do DNA a ser sequenciado, o que pode ser considerado uma vantagem, já que durante o processo de clonagem podem ocorrer contaminações que podem interferir nos resultados, além do tempo gasto com o mesmo. O processo de amplificação do DNA é feito pela ligação das moléculas de DNA a serem sequenciadas até que seja formada uma fita única para ser replicada. A fita única é obtida da seguinte forma:

- Os fragmentos de DNA são desnaturados para que fitas simples de DNA sejam obtidas;
- Duas fitas simples aleatórias, geradas pela desnaturação dos fragmentos, de DNA são ligadas uma à outra por adaptadores na extremidade $5'$ da primeira e na extremidade $3'$ da segunda, criando assim o conceito de "ponte" entre as duas fitas e formando uma nova cadeia de DNA composta pelas duas fitas anteriores;
- A cadeia resultante da junção das duas fitas anteriores é ligada da mesma forma, através de um adaptador, a outra fita de DNA.

O processo descrito anteriormente é repetido até que seja formada uma única cadeia de DNA composta de todas as fitas e adaptadores entre elas. A única molécula, formada das moléculas iniciais e suas pontes, forma a cadeia *template* para o sequenciamento. A cadeia *template* possui tamanho correspondente à soma dos tamanhos de todas as cadeias menores e adaptadores utilizados para sua confecção. Após a amplificação da *template*, mais de 40 milhões de *clusters*, contendo uma cópia de cada cadeia *template*, são formados e, através de um processo enzimático, similar a uma reação de PCR envolvendo a enzima DNA polimerase, os *templates* são replicados novamente dentro de cada *cluster*. Em uma

execução de 2 a 3 dias, essa tecnologia é capaz de gerar 10M - 15M de *reads* de tamanho aproximado de 100 - 200 bp.

A tecnologia Illumina gera arquivos com os caracteres correspondentes ao formato chamado de Illumina FASTQ. Esse formato é parecido com o formato FASTA, mas além de conter as informações dos nucleotídeos sequenciados, também contém, para cada nucleotídeo, um valor correspondente à sua qualidade.

Detalhes do formato *Illumina* FASTQ podem ser vistos no Apêndice A.2.

ABI

A tecnologia *Applied Biosystems SOLiD* também constrói sua biblioteca de sequências através da técnica de *Emulsion PCR*, similar ao processo usado na *Roche 454*. Um esquema de cores é previamente definido: são 4 cores e 16 combinações de dinucleotídeos (dNTPs), resultantes do agrupamento de dois nucleotídeos consecutivos ligados em cadeia. Os produtos das amplificações (resultantes da *Emulsion PCR*) são transferidos para uma superfície de vidro e sofrem turnos de hibridização (ou ligação) de dinucleotídeos. No sequenciamento, cada nucleotídeo é identificado pela cor da reação resultante de reações de ligação em uma posição do *read* sequenciado. Todo esse processo é capaz de produzir 1M - 1,5M de *reads* de tamanho aproximado de 35 bp em uma execução de 8 horas.

A saída gerada por esta tecnologia é detalhada no Apêndice A.3.

2.3.3 Aplicações das tecnologias NGS

Os *reads* produzidos pelas técnicas de nova geração são menores que os *reads* produzidos pelo método Sanger. Apesar do menor conjunto de informação em cada *read*, essa tecnologia gera uma quantidade muito maior de dados e essa característica é utilizada na construção de novas aplicações. Em [26] as tecnologias *Next-Generation* transformaram o campo da análise de sequenciamento de DNA. Os *reads* gerados pelas novas tecnologias foram utilizados em diversos projetos de pesquisa como ressequenciamento do genoma humano [20], identificação de novos genes no genoma humano em [41], identificação de variações genéticas em um genoma não cobertas pelo processo de sequenciamento [28], além de pesquisas utilizando metagenômica [34].

Capítulo 3

O problema da remontagem de um genoma

As técnicas de sequenciamento NGS produzem um conjunto de *reads*, geralmente de 35 a 600 pares de bases. Estes *reads* precisam, de alguma maneira, ser "organizados" a fim de se obter o genoma do organismo sequenciado (DNA alvo), uma vez que um *read* representa apenas um fragmento do DNA que se deseja conhecer.

A remontagem de um genoma consiste na organização de um conjunto de *reads* com intuito de se obter um genoma de um organismo sequenciado. Existem diferentes técnicas para a remontagem de genomas e esse trabalho aborda duas delas: **remontagem por comparação** e **remontagem *de novo***. A primeira é capaz de remontar um genoma por meio da comparação dos *reads* com um genoma referência, enquanto a segunda utiliza apenas as informações contidas nos próprios *reads*.

As duas técnicas de remontagem de genoma acima citadas serão abordadas nas seções 3.1 e 3.2, respectivamente.

3.1 Remontagem por comparação (*Comparative-Assembling*)

Muitos dos atuais projetos de remontagem de genomas utilizam um ou mais genomas conhecidos, ditos referências, ou ainda, **genomas-referência** para que o mapeamento dos *reads* seja feito [37]. Remontar um genoma tendo como referência um ou mais outros genomas é a técnica conhecida como **Remontagem por comparação**, do inglês *Comparative-Assembling*.

Mapear um *read* em um genoma-referência é o processo onde a região do genoma-referência mais semelhante ao genoma do *read* é encontrada. Mais especificamente, mapear um *read* é um processo recorrente da Biologia Computacional: o alinhamento entre sequências. Não é foco deste trabalho o tema de alinhamento entre sequências, mas detalhes podem ser encontrados em [19]. O mapeamento dos *reads* em um genoma-referência

é um problema que pode ser definido resumidamente da seguinte maneira: Dado um conjunto de *reads* R e um genoma-referência S, para cada *read* r pertencente a R, determinar o melhor alinhamento de r com a sequência S.

Define-se um *match* de r em S como sendo um possível alinhamento de r com S. Nem todo alinhamento de r com S representa semelhança considerável entre as duas sequências, logo é necessário definir uma similaridade λ entre elas. Dessa forma, quanto maior o valor de λ, maior a similaridade entre r e S. Assim, para cada *read* r é necessário encontrar o conjunto de *matches* de r em S onde cada *match* tem $\lambda > k$, sendo que k é um valor inteiro definido como similaridade mínima aceita entre duas sequências.

Uma vez definido o conjunto de *matches* para cada *read* r, o melhor *match* de r é escolhido como o mapeamento de r em S. A escolha do melhor *match* não é trivial e cada ferramenta que executa o mapeamento de *reads* utiliza uma heurística própria para esta escolha. Algumas dificuldades na escolha do melhor *match* pode ser vista em [26].

Quando um *read* é mapeado em uma região específica de um genoma, diz-se que este *read* **representa** esta região do genoma. A Figura 3.1 exemplifica o mapeamento de um *read* em um genoma.

Figura 3.1: Um *read* que representa uma região específica no genoma, chamada de região mapeada. A região mapeada é a região que possui maior semelhança entre o genoma e o *read*. O sinal '*' indica que duas bases correspondentes no alinhamento não são iguais.

Para mapear um conjunto de *reads* em um genoma-referência, este deve ser escolhido estrategicamente observando suas características. Alguns dos critérios para escolha de um genoma-referência são:

1. O genoma-referência deve ser previamente conhecido, proveniente, preferencialmente, de um projeto já finalizado.
2. O genoma-referência deve ser semelhante ao genoma a ser remontado, ou seja, de um organismo filogeneticamente próximo ao organismo de onde os *reads* são provenientes.

A utilização do genoma-referência proveniente de projetos já finalizados agrega qualidade e confiabilidade ao mapeamento dos *reads*, pois estes serão mapeados em um genoma previamente estudado. A utilização de genoma-referência semelhante ao genoma a ser remontado tem por objetivo facilitar o processo de mapeamento devido à semelhança entre as sequências dos *reads* e a sequência do genoma-referência.

Como já foi dito, as tecnologias NGS geram um grande volume de dados em forma de *reads*. Para a remontagem de um genoma, os pequenos fragmentos de DNA (*reads*) são submetidos a ferramentas computacionais capazes de mapear estes *reads* em um ou mais genomas-referência. A remontagem de *reads* por comparação não é um processo trivial e existem vários softwares que se propõem a efetuar tal tarefa. Uma lista de alguns desses *softwares* pode ser vista na Tabela 3.1.

Tabela 3.1: *Softwares* para mapeamento de *reads* em genomas-referência. O sinal "-" significa que o *software* não possui limitação quanto ao tamanho máximo do *read*.

Software	**Código aberto**	**Tamanho máximo do *read***
Bowtie	Sim	-
Maq	Sim	127
Mosaik	Não	-
SOAP3	Não	60
MUMmer	Sim	-
Blat	Sim	-

Remontar genomas por comparação não é um processo simples. Segundo [37], duas questões principais estão envolvidas nesta tarefa:

1. O genoma-referência é muito grande e é necessário mapear bilhões de *reads*, quão rapidamente é possível fazer isso?

2. Um *read* pode ser igualmente semelhante a mais de uma região do genoma-referência e, por isso, pode ser mapeado em mais de uma região deste genoma. Esse *read* deve ser mapeado em todos lugares onde é semelhante ao genoma-referência ou é necessário escolher um dos possíveis mapeamentos? Como fazer esta escolha?

3. Como identificar um SNP?

O primeiro desafio demonstra com clareza a importância de algoritmos poderosos em termos de tempo e espaço de armazenamento para o problema de alinhamento de sequências. Cada *read* deve ser alinhado à referência de forma que o resultado seja obtido em tempo satisfatório.

Como um *read* pode ser igualmente semelhante a duas ou mais regiões da referência (ver Figura 3.2), o *software* de mapeamento precisa decidir qual será a estratégia (geralmente uma heurística) utilizada para mapear este *read*.

Além dos desafios anteriormente citados, deve-se considerar o fato da orientação do *read* (se $3' \rightarrow 5'$ ou se $5' \rightarrow 3'$) ser desconhecida, já que um fragmento pode ser o resultado do sequenciamento de qualquer uma das fitas de uma molécula de DNA. O *software* precisa escolher (heuristicamente) se faz o mapeamento do *read* considerando que a orientação deste é $5' \rightarrow 3'$ ou se calcula o complemento reverso do *read* e realiza o mapeamento de ambos.

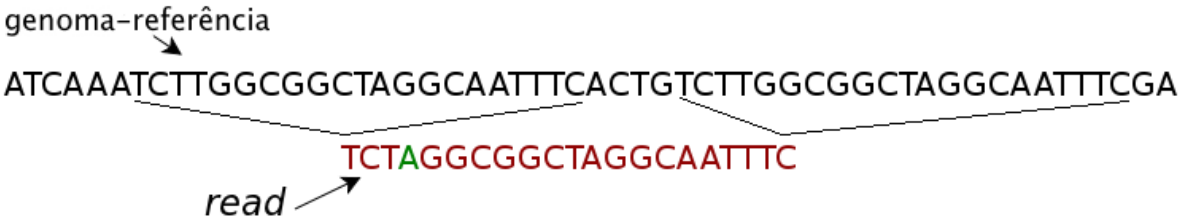


Figura 3.2: *Read* igualmente semelhante a duas regiões do genoma-referência.

Durante o processo de mapeamento, uma mesma região do genoma-referência pode ser representada por mais de um *read*. Assim, os *reads* mapeados em um genoma-referência podem sobrepor-se uns aos outros formando ***contigs***, que são conjuntos contíguos de *reads* sobrepostos. A Figura 3.3 mostra a formação de *contigs* em um mapeamento.

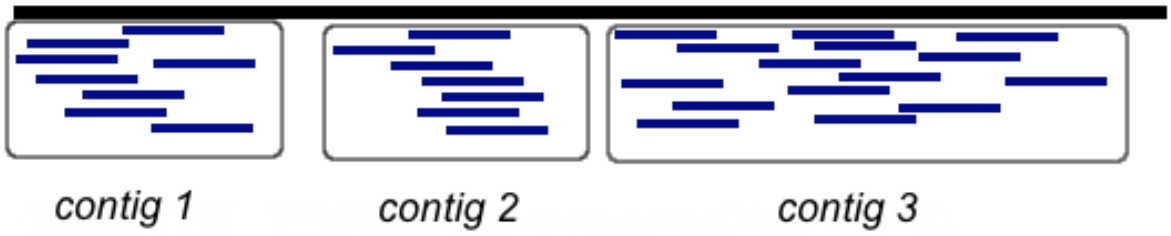


Figura 3.3: O alinhamento (mapeamento) dos *reads*, representados em azul, em um genoma-referência, representado pelo segmento na cor preta, formando três regiões contíguas (*contigs*) de fragmentos de DNA sobrepostos.

A sequência **consenso** é determinada a partir da sobreposição dos *reads* em um *contig*. Para cada nucleotídeo mapeado no genoma-referência, as sobreposições dos nucleotídeos dos *reads* que o mapeiam são avaliadas. A média ponderada das qualidades dos nucleotídeos, considerando de modo isolado cada tipo (A, C, T e G), é computada, sendo que a maior determina o nucleotídeo correspondente na sequência consenso. Um exemplo simplificado de determinação de sequências consensos pode ser visto na Figura 3.4.

O resultado do mapeamento pode ser formado um único *contig* ou por vários (dois ou mais) *contigs*. Quando um único *contig* é formado e seus *reads* representam todo o genoma-referência, é possível determinar uma sequência consenso única que, neste caso, corresponde a uma possível remontagem do genoma por completo. Quando uma região do genoma-referência não é representada por nenhum *read*, ou ainda por poucos *reads*, diz-se que a região não foi coberta, de forma que a determinação de uma única sequência consenso não é possível. Assim, duas ou mais sequências consensos são obtidas a partir da sobreposição dos *reads* em cada região coberta na referência e as regiões não mapeadas pelos *contigs* no genoma-referência são chamadas *gaps*.

Como visto, um único nucleotídeo do genoma-referência pode (e deve) ser mapeado por diversos *reads*. À quantidade de *reads* mapeados em um nucleotídeo do genoma-referência

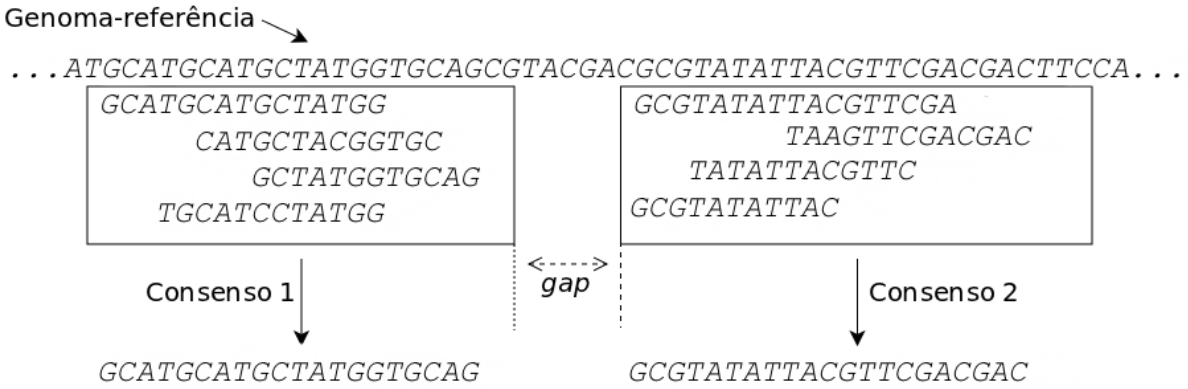


Figura 3.4: Obtenção da sequência consenso de cada *contig* resultante do mapeamento dos *reads*.

dá-se o nome de **cobertura do nucleotídeo**.

3.2 Remontagem *de novo*

Outra técnica de remontagem de um genoma a partir de um conjunto de *reads* é a técnica conhecida como *de novo*. Diferente da remontagem com base em um genoma-referência, a técnica *de novo* não conta com informação de nenhum genoma previamente conhecido e, supostamente, semelhante ao genoma que se deseja remontar. A idéia principal dessa técnica é remontar o genoma com base apenas nas informações contidas nos *reads*.

Para utilizar as informações contidas nos *reads*, a técnica *de novo* de remontagem compara, por meio de alinhamento de sequências, cada *read* com os demais, formando sobreposições entre esses *reads*. Assim, grupos contíguos de *reads* sobrepostos são formados, também denominados *contigs*.

O problema da remontagem *de novo* pode ser definido da seguinte maneira: seja R um conjunto de *reads* e, para cada *read* pertencente ao conjunto R, é necessário alinhá-lo aos demais *reads* de R. Quando r se sobrepõe a um *read* s, também pertencente a R, ambos pertencem ao mesmo conjunto sobreposto (*contig*) de *reads*. Em outras palavras, quando existe uma sobreposição entre r e s, ambos pertencem ao mesmo *contig*. Logo, quando um terceiro *read* t possui uma sobreposição com r ou s, este também pertence ao mesmo *contig* de r e s. Os alinhamentos são feitos para todos os *reads* pertencentes a R e ao final desses alinhamentos podem ser formados um ou mais *contigs*.

Da mesma forma que na remontagem por comparação, caso um único *contig* seja formado, é possível determinar uma única sequência consenso, que pode corresponder à remontagem completa do genoma. Caso sejam formados dois ou mais *contigs*, não será possível determinar uma única sequência consenso.

Os *contigs* formados pela técnica de remontagem *de novo* não possuem ordem entre si, ou seja, a posição de um *contig* em relação a outro não é definida. Supondo X e Y

dois *contigs* obtidos a partir dos *reads* remontados por meio da técnica de remontagem *de novo*, as seguintes perguntas devem ser respondidas:

- No genoma remontado, X precede Y ou Y precede X?
- Qual a orientação dos *contigs*? X ou Y foram sequenciados pelo comprimento reverso?
- Se X precede Y, quão distantes, em número de nucleotídeos, X está de Y?
- Se só um *contig* for obtido, a sequência consenso correspondente representa uma remontagem completa do genoma ou apenas uma parte dele, já que regiões do genoma podem não ter sido cobertas por nenhum *read*?

Um procedimento que estabelece a ordem e a orientação entre os *contigs* é chamado de *Scaffolding*. Este procedimento usa as informações contidas nos *reads* para responder as perguntas anteriores. Pop, Kosack e Salzberg em [33] descrevem situações onde o pareamento de dois *reads* e a sobreposição entre *contigs* podem ajudar no processo de ordenamento de todos os *contigs*. O sequenciamento do tipo *paired-end*[1] pode ser utilizado como informação complementar sobre os *contigs* obtidos [31]. Neste tipo de sequenciamento, as duas extremidades de um fragmento são sequenciadas gerando dois *reads*: *readA* e *readB* de orientações opostas.

Os *reads readA* e *readB* são chamados de *clone-mates* e cada um corresponde a uma extremidade do fragmento original sequenciado. O sequenciamento das extremidades de um fragmento pode não resultar no sequenciamento do fragmento por completo, como ilustrado na Figura 3.5.

Supondo que o *readA* está em um *contig*, o *readB* está em outro *contig* distinto do primeiro e o tamanho (em número de bases) do fragmento que deu origem a esses dois *reads* é conhecido. É possível neste caso estimar a quantidade de bases existentes entre *readA* e *readB* para formar o fragmento original de onde foram derivados. Logo, pareando os dois *reads* é possível determinar a distância (tamanho do *gap*) entre esses dois *contigs* (ver Figura 3.6).

Figura 3.5: *Clone-mates* podem não resultar no sequenciamento do fragmento por completo.

Após descobrir a presença de *clone-mates* entre *contigs* e estimar o *gap* entre eles, esses *contigs* podem ser ordenados. Em uma forma geral, a representação de precedência

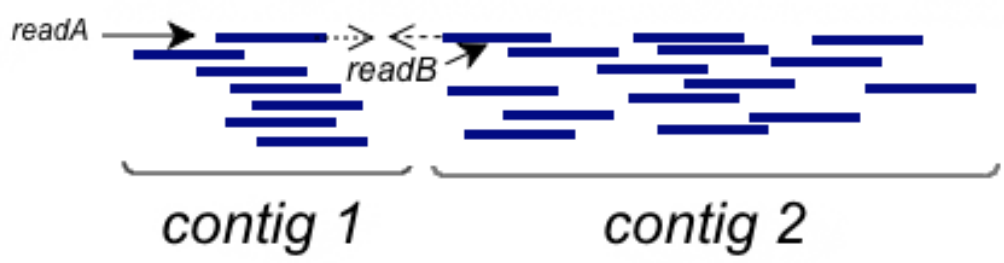


Figura 3.6: Informações de *clone-mates* ajudam a determinar o tamanho do *gap* entre *contigs*.

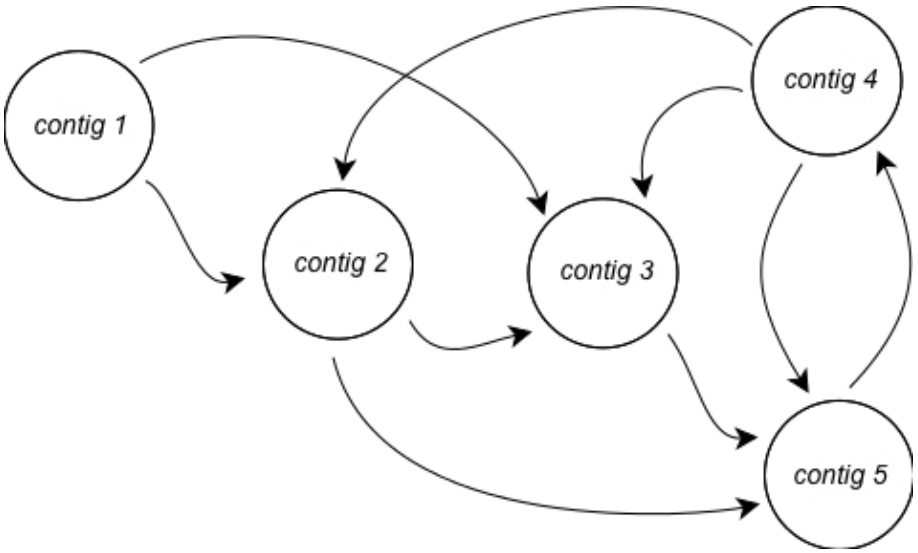


Figura 3.7: Precedência entre *contigs*. Uma seta de um *contig* **X** para um *contig* **Y** significa a possibilidade de precedência de **X** em relação a **Y**

entre um *contig* e outro é representada por um grafo orientado, como exemplificado na Figura 3.7.

É possível notar que existem diversas possibilidades de se estabelecer uma linha de precedências entre os *contigs*. A partir da Figura 3.7, por exemplo, possíveis ordens entre os *contigs* são, dentre outras: *contig* 1 - *contig* 3 -*contig* 5 - *contig* 4 - *contig* 2 e *contig* 1 - *contig* 2 - *contig* 3 - *contig* 5 - *contig* 4. Dessa forma, um *software* que esteja sendo utilizado para a remontagem dos *reads* e seus *contigs* deve estabelecer alguma técnica ou heurística para determinar qual das possíveis opções de precedência entre os *contigs* corresponderá ao ordenamento correto entre eles. Uma vez determinadas as relações de distância entre os *contigs* e a ordem entre eles, uma possibilidade da sequência do genoma remontado é obtida.

Embora aproveite as informações contidas nos *reads*, a técnica de remontagem *de novo* encontra limitações computacionais para sua utilização, por exemplo, a disponibilidade

[1] *Paired-end* é uma forma de sequenciamento onde um *read* de tamanho definido é sequenciado em ambas orientações, a fim de prover mais informação sobre o fragmento.

de ter recursos computacionais suficientes e computar o resultado em tempo hábil. O alinhamento entre os milhares de pequenos fragmentos de DNA exige um algoritmo robusto e tempo para sua execução. Outra dificuldade desta técnica é que os *softwares* desenvolvidos têm um desempenho bom quando se trata de bactérias e eucariotos pequenos, mas ainda é um desafio elaborar um *software* eficiente para um genoma de um vertebrado de maior tamanho [38].

Alguns dos *softwares* que realizam a remontagem de um genoma apenas com base nas informações contidas nos *reads* (*de novo*) são apresentados na Tabela 3.2, adaptada de [33].

Tabela 3.2: *Softwares* que utilizam a técnica *de novo* para remontagem de *reads*.

Software	**Suporte ao Illumina**	**Suporte ao 454**	**Suporte ao SOLiD**
Abyss	Sim	Não	Não
ALLPATHS-LG	Sim	Sim	Não
MIRA3	Não	Sim	Não
Newbler	Não	Sim	Não
SSAKE	Sim	Sim	Não
SOAPdenovo	Sim	Sim	Não
SHARCGS	Sim	Não	Não
VCAKE	Sim	Não	Sim
Velvet	Sim	Não	Não

Capítulo 4

Metodologia e critérios de avaliação

A remontagem de um genoma é um processo que envolve desde o sequenciamento de pequenos fragmentos de DNA até a remontagem destes por alguma técnica específica. Os capítulos anteriores mostraram técnicas de sequenciamento de DNA, clássica e NGS, e formas como os fragmentos resultantes do sequenciamento podem ser remontados via comparação com genomas de referência ou pela técnica de remontagem *de novo*.

Dado um conjunto de *reads*, o genoma que deu origem a estes é único. Entretanto, diferentes técnicas de remontagem podem resultar, para um mesmo conjunto de fragmentos, em genomas, ou *contigs*, distintos. Este fato decorre das diferentes técnicas empregadas para a solução das questões envolvidas no processo de remontagem, como por exemplo: algoritmo de alinhamento, escolha de posicionamento dos *reads* quando mais do que uma posição para este é possível, entre outros.

Frente às diferenças existentes nos resultados produzidos pelos *softwares* e o fato de o genoma original ser desconhecido (na maioria das vezes), a escolha da ferramenta a ser utilizada é feita, muitas vezes, ao acaso. Neste sentido, o objetivo deste estudo é o de facilitar, ou colaborar com esta escolha por meio da avaliação de um conjunto de ferramentas de remontagem.

As comparações foram realizadas considerando ferramentas de remontagem por comparação e *de novo*. Os resultados das remontagens foram avaliados comparando-os uns aos outros quanto à mesma técnica de remontagem, bem como os resultados entre as duas diferentes técnicas. O processo de avaliação das ferramentas seguiu a metodologia descrita na seção seguinte.

4.1 Metodologia e ferramentas selecionadas

A seguinte metodologia foi empregada durante o desenvolvimento deste trabalho:

- Levantamento das ferramentas de remontagem de genoma baseadas nas técnicas *de novo* e remontagem por comparação;

- Estudo sobre as ferramentas encontradas: disponibilidade, funcionamento, recursos computacionais necessários para sua execução, tipos de arquivos de entrada e saída, parâmetros de execução, etc.;
- Escolha das ferramentas a serem utilizadas e instalação das mesmas;
- Determinação do conjunto de testes a serem submetidos às ferramentas;
- Execução das ferramentas sobre os dados do conjunto de testes;
- Avaliação dos resultados; e
- Comparação entre as ferramentas.

A escolha das ferramentas existentes foi baseada na capacidade das mesmas em efetuar a remontagem de fragmentos oriundos de NGS, tanto pela técnica de comparação quanto pela técnica *de novo*. Em [15] é possível consultar uma lista de ferramentas que executam diversas tarefas em bioinformática como: remontagem de fragmentos, alinhamento de fragmentos, visualização de fragmentos, dentre outras. Os conceitos de alinhamento de fragmentos e remontagem de fragmentos pela técnica de comparação são muitas vezes utilizados como sinônimos. Porém, na utilização da maioria das ferramentas descritas como alinhadoras/remontadoras de fragmentos utilizando a técnica de remontagem por comparação, é possível verificar que a etapa de alinhamento é executada, mas não há geração de *contigs*, nem de sequências consensos, como é o caso das ferramenta Blat [27] e Bowtie [3]. A análise sobre os *contigs* gerados é importante para comparar os resultados obtidos nas remontagens entre as técnicas *de novo* e por comparação. Para a solução deste problema foi estudada a ferramenta AMOScmp.

AMOS [16] é um conjunto de ferramentas, conversores de tipos de arquivos, *scripts* e outros utilitários para uso em soluções de bioinformática. AMOScmp é uma ferramenta capaz de executar a remontagem de fragmentos utilizando a técnica por comparação. Os *scripts* de execução desta ferramenta fazem parte do conjunto de ferramentas AMOS. A remontagem feita pela ferramenta AMOScmp é feita em dois passos principais: primeiro os fragmentos são alinhados à referência utilizando a ferramenta MUMmer [23]. Em seguida, um *script* chamado *make-consensus* gera, a partir do alinhamento, *contigs*, bem como as respectivas sequências consensos.

Como dito anteriormente, as ferramentas Blat e Bowtie executam apenas o processo de alinhamento dos *reads* ao genoma-referência. Os resultados dos alinhamentos das ferramentas Blat e Bowtie são convertidos para o padrão de entrada do *script make-consensus* e os *contigs* são gerados. Dessa forma é possível incluir ferramentas, com diferentes heurísticas de alinhamento, para comparação de resultados quanto ao processo de remontagem de um genoma. Além disso, há uma padronização dos resultados obtidos pelas ferramentas, visto que o *script* que gera os resultados é o mesmo para todas elas.

A avaliação da remontagem pela técnica *de novo* foi feita utilizando duas ferramentas: SSAKE [40] e SOAPdenovo [17]. Estas ferramentas recebem como entrada um conjunto de *reads* e devolvem como saída arquivos contendo os *contigs* formados e suas respectivas sequências consensos. Para os *contigs* formados, ambas as ferramentas executam o

processo de *scaffolding*, estabelecendo a ordem entre esses *contigs*. As informações nos arquivos de saída são semelhantes em relação às informações nos arquivos de saída para as ferramentas que utilizam a técnica por comparação após estes terem sido submetidos ao *script make-consensus*. Assim, em ambas as técnicas é possível avaliar os dados sobre os *contigs* gerados e comparar os resultados entre si.

Em suma, as ferramentas utilizadas foram AMOScmp, Blat e Bowtie para a técnica de remontagem por comparação e SSAKE e SOAPdenovo para a técnica *de novo*. Uma breve descrição das principais ferramentas estudadas está no Apêndice B.

As ferramentas selecionadas foram instaladas e configuradas para funcionar em ambiente Unix 64 bits, 22 Gb RAM e 2 processadores Intel(R) Xeon(R) 2.13GHz de 4 núcleos. As execuções consideraram conjuntos de testes simulados e também para conjuntos de testes reais. A seguir, tais conjuntos são descritos.

4.2 Conjunto de testes

São descritos os conjuntos de testes, compostos por dois tipos de dados: simulados e reais.

4.2.1 Dados simulados

Para a construção dos dados simulados foi gerada uma sequência aleatória de DNA denominada SEQ-ALE de tamanho 642519 pb, próximo ao tamanho do genoma usado como referência nas execuções com dados reais. A sequência SEQ-ALE foi fragmentada utilizando o simulador de *reads* GenFrag [25]. Um simulador de *reads* é capaz de simular a produção de pequenos fragmentos de DNA a partir de uma sequência maior dada como entrada, assim como ocorre no sequenciamento de moléculas pelos métodos Sanger ou NGS. O simulador GenFrag foi executado diversas vezes, alternando os valores de seus parâmetros, a fim de gerar dados simulados com diferentes características como: diferentes tamanhos de *reads*, altas e baixas taxas de cobertura e diferentes taxas de erros na geração dos fragmentos.

Os fragmentos gerados pelo simulador GenFrag a partir da SEQ-ALE possuem tamanhos 35, 50 e 120 pb, semelhantes aos tamanhos de *reads* gerados pelas técnicas NGS, respeitadas as limitações deste trabalho. Além disso, para cada tamanho de *read*, foram gerados fragmentos considerando taxas de erro semelhantes às taxas de erro da enzima DNA polimerase: 1 erro a cada 10^9 bases ($1 * 10^{-9}$), 2,5 erros a cada 10^9 bases ($2,5 * 10^{-9}$) e 4 erros a cada 10^9 bases ($4 * 10^{-9}$). Para cada tamanho de fragmento e taxa de erro, foram gerados fragmentos considerando as coberturas de 7, 16 e 20 pb, escolhidas com base em [18]. Por fim, para cada combinação de tamanho, taxa de erro e cobertura, foram feitas 10 execuções diferentes do simulador de fragmentos. Dessa forma, o conjunto de dados simulados é formado por 27 conjuntos de testes (27 arquivos de tamanho médio 84 Mb e número médio de bases de 34,5 M), cada um com 10 conjuntos de fragmentos, totalizando 270 conjuntos de fragmentos, formados por todas as combinações entre os tamanhos, taxas de erro, coberturas e execuções descritos anteriormente.

Como a sequência completa a ser remontada SEQ-ALE é conhecida, os resultados obtidos das remontagens puderam ser comparados. Como resultado ótimo, espera-se que a remontagem do genoma a partir dos *reads* simulados resulte na sequência fornecida como entrada ao simulador GenFrag, mas devido às características dos *reads* (tamanhos e erros) a serem remontados, bem como a técnica de remontagem empregada, nem sempre é possível que a remontagem resulte em uma sequência única, igual a sequência original. Desta maneira, conhecendo o resultado ótimo, é possível verificar a corretude de cada ferramenta na tarefa de remontagem dos *reads*, considerando as variações de características dos *reads* gerados pelo GenFrag.

4.2.2 Dados reais

Para a simulação com os dados reais foi utilizado um organismo cujo projeto genoma já foi finalizado e cuja sequência completamente remontada está depositada no Genbank. Sendo assim, esta foi utilizada como referência para remontagem por comparação. Note, entretanto, que esta sequência foi obtida por algum método de remontagem e, por isso, pode não representar a sequência ótima. O conjunto de *reads* foi composto por *reads* de organismos filogeneticamente próximos ao genoma-referência, depositados no banco de sequências SRA, do inglês *Sequence Read Archive*. O banco de sequências SRA [4] armazena conjuntos de pequenos fragmentos resultantes de sequenciamentos de tecnologias NGS.

Um projeto de sequênciamento de fragmentos de um determinado organismo no banco SRA pode conter vários *runs*. Um *run* corresponde a uma execução da máquina sequenciadora para este organismo, gerando *reads*. Cada projeto possui um numeral identificador (ID) e cada *run* possui uma identificação alfanumérica. Por exemplo, o projeto identificado pelo ID 29669 possui três *runs*, quais sejam identificados por SRR066395, SRR066396 e SRR066397. Cada conjunto de *reads* utilizados como entrada para os testes reais é formado por todos os *runs* de um projeto escolhido. Assim, todos os *reads* gerados no sequenciamento de um projeto são utilizados na remontagem com dados de reais.

Como genoma-referência foi escolhida a bactéria *Bacillus cereus* depositada no Genbank, identificada neste banco por DQ889677.1. Como pode ser visto em [35], este genoma foi remontado utilizando *reads* provenientes de técnica clássica de sequenciamento. A escolha de um genoma que não tenha sido remontado com *reads* provenientes de tecnologias NGS é importante pois evita que haja vício nos testes de remontagem deste trabalho, visto que uma das ferramentas testadas pode ser a mesma (ou de algoritmo semelhante) que remontou o genoma em questão. Os *reads* submetidos como entrada para as ferramentas foram coletados de projetos envolvendo também a bactéria *Bacillus cereus*, porém cada projeto isolou este organismo em condições particulares, como região geográfica ou procedimento de laboratório distintos. A Tabela 4.1 contém informações sobre os *reads* utilizados.

Tabela 4.1: Conjuntos de *reads* submetidos como entrada para as ferramentas. Os tamanhos e número de bases correspondem ao somatório de todos os *reads* de todos os *runs* do mesmo projeto.

ID do projeto no SRA	Número de bases	Tamanho	Sequenciamento
29655	127,3M	307,4Mb	Roche 454
29657	295,2M	699,9Mb	Roche 454
29649	276,8M	653,7Mb	Roche 454
29659	285,6M	673,5Mb	Roche 454
70391	2,9G	1,8Gb	Illumina
70301	2,4G	1,44Gb	Illumina
70307	2,8G	1,6Gb	Illumina
70321	2,6G	1,6Gb	Illumina

Capítulo 5

Experimentos e comparações entre as ferramentas

Com o objetivo de avaliar os resultados dos experimentos realizadas, algumas informações foram consideradas nos resultados produzidos pelas ferramentas de remontagem de *reads* avaliadas:

- Número de *contigs*;
- Cobertura dos *reads* em relação aos *contigs*;
- Tamanho dos *contigs*;
- Heurísticas utilizadas pelas ferramentas;
- Percentual de acerto dos resultados gerados pelas ferramentas de remontagem quando comparados com a remontagem ótima (SEQ-ALE).

As ferramentas de remontagem foram submetidas aos conjuntos de testes reais e simulados. Como cada conjunto de teste dos dados simulados é composto por 10 execuções de cada combinação de parâmetros (taxa de erro, tamanho e cobertura dos fragmentos) do simulador GenFrag, os resultados para dados simulados foram apresentados em termos das médias da quantidade, cobertura, tamanho e corretude dos *contigs* gerados. Para as execuções das ferramentas com dados reais, cada conjunto de fragmentos selecionado do banco de sequências SRA foi submetido uma vez como entrada para cada ferramenta.

5.1 Resultados com dados simulados

Quando uma ferramenta de remontagem é executada tomando como entrada um conjunto de *reads*, a saída desta ferramenta corresponde ao resultado da remontagem desses fragmentos utilizando a técnica de remontagem e as heurísticas próprias dessa ferramenta.

Em cada execução de uma ferramenta, foi possível coletar informações sobre os resultados obtidos. Tais informações foram a quantidade de *contigs* gerados, a cobertura dos *reads* em relação a cada *contig*, o tamanho de cada *contig* e a corretude da sequência consenso em relação à sequência ótima a ser remontada. As tabelas 5.1, 5.2, 5.3, 5.4 e 5.5 a seguir consolidam os dados dos resultados obtidos nas execuções das ferramentas de remontagem avaliadas, considerando as particularidades nos conjuntos de *reads* fornecidos como entrada.

Cada tabela agrupa os resultados para uma única ferramenta. Com base nesses resultados, foram feitas análises considerando as informações de número, cobertura, tamanho e corretude dos *contigs*. Tais análises são descritas logo após a apresentação das tabelas.

Tabela 5.1: Resultados da remontagem com dados simulados para a ferramenta AMOScmp

AMOScmp						
Conjunto de *reads*			Dados sobre os *contigs* resultantes da remontagem			
Cobertura	Tamanho	Taxa de erro	Quantidade	Cobertura	Tamanho	Corretude (%)
7	35	$1 * 10^{-9}$	368,40	7,77	1744,08	79,32
		$2,5 * 10^{-9}$	368,40	7,52	1744,07	79,18
		$4 * 10^{-9}$	368,10	7,51	1745,48	79,19
	50	$1 * 10^{-9}$	251,80	7,31	2547,91	82,93
		$2,5 * 10^{-9}$	251,90	7,37	2560,35	82,57
		$4 * 10^{-9}$	251,80	7,35	2546,49	82,03
	120	$1 * 10^{-9}$	170,70	6,32	3741,50	89,12
		$2,5 * 10^{-9}$	172,10	6,31	3731,60	88,38
		$4 * 10^{-9}$	172,90	6,32	3729,22	86,90
16	35	$1 * 10^{-9}$	134,20	15,69	4797,63	88,22
		$2,5 * 10^{-9}$	134,10	15,66	4802,66	85,77
		$4 * 10^{-9}$	134,10	15,52	4802,70	84,10
	50	$1 * 10^{-9}$	133,90	14,69	4796,05	89,31
		$2,5 * 10^{-9}$	134,20	14,81	4776,29	87,56
		$4 * 10^{-9}$	134,10	14,74	4784,13	86,24
	120	$1 * 10^{-9}$	129,70	13,33	4947,20	91,66
		$2,5 * 10^{-9}$	130,50	13,35	4966,10	91,34
		$4 * 10^{-9}$	131,00	13,37	4981,64	91,17
20	35	$1 * 10^{-9}$	134,00	19,73	4802,12	92,15
		$2,5 * 10^{-9}$	134,10	19,77	4800,13	90,99
		$4 * 10^{-9}$	134,10	19,75	4800,15	89,93
	50	$1 * 10^{-9}$	134,00	18,79	4800,68	95,32
		$2,5 * 10^{-9}$	134,10	18,78	4800,44	95,11
		$4 * 10^{-9}$	134,10	18,79	4800,46	93,84
	120	$1 * 10^{-9}$	130,00	17,97	4939,31	96,69
		$2,5 * 10^{-9}$	130,00	17,92	4939,26	96,42
		$4 * 10^{-9}$	130,00	17,96	4939,28	95,88

Tabela 5.2: Resultados da remontagem com dados simulados para a ferramenta Blat

Blat						
Conjunto de *reads*			Dados sobre os *contigs* resultantes da remontagem			
Cobertura	Tamanho	Taxa de erro	Quantidade	Cobertura	Tamanho	Corretude (%)
7	35	$1 * 10^{-9}$	298,20	6,98	2221,71	83,01
		$2,5 * 10^{-9}$	298,10	6,97	2228,49	81,56
		$4 * 10^{-9}$	298,20	6,98	2221,60	80,22
	50	$1 * 10^{-9}$	210,60	6,89	3050,59	83,20
		$2,5 * 10^{-9}$	210,60	6,90	3050,46	82,48
		$4 * 10^{-9}$	210,60	6,90	3050,77	82,34
	120	$1 * 10^{-9}$	136,10	6,83	4720,93	89,03
		$2,5 * 10^{-9}$	136,20	6,81	4715,51	88,97
		$4 * 10^{-9}$	136,20	6,82	4721,02	87,95
16	35	$1 * 10^{-9}$	113,40	16,03	5665,95	89,55
		$2,5 * 10^{-9}$	113,40	16,01	5665,93	88,91
		$4 * 10^{-9}$	113,30	15,98	5666,63	85,38
	50	$1 * 10^{-9}$	108.90	15,63	5900,08	90,15
		$2,5 * 10^{-9}$	108,90	15,66	5900,21	89,36
		$4 * 10^{-9}$	108,90	15,65	5900,09	84,70
	120	$1 * 10^{-9}$	102,60	15,20	6262,36	92,87
		$2,5 * 10^{-9}$	102,50	15,22	6264,44	91,66
		$4 * 10^{-9}$	102,60	15,16	6262,31	89,73
20	35	$1 * 10^{-9}$	111,30	19,83	5772,85	90,12
		$2,5 * 10^{-9}$	111,30	19,82	5771,90	90,04
		$4 * 10^{-9}$	113,20	19,79	5780,02	89,59
	50	$1 * 10^{-9}$	109,50	19,02	5867,75	95,92
		$2,5 * 10^{-9}$	109,40	19,07	5867,98	95,18
		$4 * 10^{-9}$	109,40	19,14	5868,04	94,09
	120	$1 * 10^{-9}$	99,80	18,33	6438,06	98,00
		$2,5 * 10^{-9}$	99,80	18,26	6437,95	97,67
		$4 * 10^{-9}$	99,80	18,12	6438,11	96,90

Tabela 5.3: Resultados da remontagem com dados simulados para a ferramenta Bowtie

Bowtie						
Conjunto de *reads*			Dados sobre os *contigs* resultantes da remontagem			
Cobertura	Tamanho	Taxa de erro	Quantidade	Cobertura	Tamanho	Corretude (%)
7	35	$1 * 10^{-9}$	311,00	6,35	2065,97	82,88
		$2,5 * 10^{-9}$	311,00	6,42	2065,80	82,44
		$4 * 10^{-9}$	311,10	6,50	2064,77	79,93
	50	$1 * 10^{-9}$	284,40	6,83	2259,20	84,11
		$2,5 * 10^{-9}$	284,30	6,84	2261,13	83,16
		$4 * 10^{-9}$	285,00	6,79	2263,74	82,64
	120	$1 * 10^{-9}$	167,20	6,91	3842,81	89,20
		$2,5 * 10^{-9}$	167,20	6,89	3842,69	89,09
		$4 * 10^{-9}$	167,20	6,93	3842,85	89,05
16	35	$1 * 10^{-9}$	121,00	14,98	5310,07	88,34
		$2,5 * 10^{-9}$	121,10	15,36	5309,45	87,99
		$4 * 10^{-9}$	121,10	15,44	5308,99	84,93
	50	$1 * 10^{-9}$	110,80	15,60	5798,90	90,02
		$2,5 * 10^{-9}$	110,90	15,61	5798,29	89,75
		$4 * 10^{-9}$	111,00	15,65	5801,02	88,66
	120	$1 * 10^{-9}$	104,50	15,61	6148,50	92,41
		$2,5 * 10^{-9}$	104,50	15,70	6148,50	92,09
		$4 * 10^{-9}$	104,70	15,68	6146,86	90,89
20	35	$1 * 10^{-9}$	114,10	19,26	5631,39	89,97
		$2,5 * 10^{-9}$	114,20	19,22	5631,14	89,38
		$4 * 10^{-9}$	114,20	19,35	5631,17	88,87
	50	$1 * 10^{-9}$	109,50	19,14	5853,29	94,76
		$2,5 * 10^{-9}$	109,50	19,22	5855,20	94,74
		$4 * 10^{-9}$	109,80	19,19	5850,91	93,68
	120	$1 * 10^{-9}$	96,20	19,62	6679,25	96,90
		$2,5 * 10^{-9}$	96,30	19,57	6678,99	96,59
		$4 * 10^{-9}$	96,90	19,55	6674,12	95,44

Tabela 5.4: Resultados da remontagem com dados simulados para a ferramenta SSAKE

SSAKE						
Conjunto de *reads*			Dados sobre os *contigs* resultantes da remontagem			
Cobertura	Tamanho	Taxa de erro	Quantidade	Cobertura	Tamanho	Corretude (%)
7	35	$1 * 10^{-9}$	1048,90	6,14	511,98	42,90
		$2,5 * 10^{-9}$	1048,30	6,15	613,29	39,92
		$4 * 10^{-9}$	1047,90	6,14	615,16	38,72
	50	$1 * 10^{-9}$	229,00	6,78	2804,72	55,84
		$2,5 * 10^{-9}$	229,10	6,80	2806,64	54,08
		$4 * 10^{-9}$	229,00	6,78	2804,74	54,06
	120	$1 * 10^{-9}$	16,70	6,98	38715,49	62,43
		$2,5 * 10^{-9}$	16,60	6,99	38721,10	62,11
		$4 * 10^{-9}$	16,60	7,01	38721,20	62,10
16	35	$1 * 10^{-9}$	642,00	15,24	1012,95	38,47
		$2,5 * 10^{-9}$	642,00	15,24	1012,38	38,46
		$4 * 10^{-9}$	642,10	15,86	1013,27	38,52
	50	$1 * 10^{-9}$	92,85	16,00	6924,35	43,49
		$2,5 * 10^{-9}$	92,00	16,01	6989,13	42,39
		$4 * 10^{-9}$	92,00	16,01	6988,44	42,33
	120	$1 * 10^{-9}$	4,80	16,09	133668,95	74,04
		$2,5 * 10^{-9}$	4,80	16,09	133668,35	73,75
		$4 * 10^{-9}$	4,80	16,11	133668,63	73,28
20	35	$1 * 10^{-9}$	167,50	18,40	3838,00	78,54
		$2,5 * 10^{-9}$	167,40	18,48	3838,10	78,54
		$4 * 10^{-9}$	167,40	19,42	3836,95	78,45
	50	$1 * 10^{-9}$	32,70	20,20	19648,83	81,20
		$2,5 * 10^{-9}$	32,70	20,18	19648,81	81,19
		$4 * 10^{-9}$	32,70	20,14	19648,83	81,14
	120	$1 * 10^{-9}$	4,20	21,40	152980,71	86,42
		$2,5 * 10^{-9}$	4,20	21,40	152978,25	86,43
		$4 * 10^{-9}$	4,20	21,20	152980,41	86,39

Tabela 5.5: Resultados da remontagem com dados simulados para a ferramenta SOAPdenovo

SOAPdenovo						
Conjunto de *reads*			Dados sobre os *contigs* resultantes da remontagem			
Cobertura	Tamanho	Taxa de erro	Quantidade	Cobertura	Tamanho	Corretude (%)
7	35	$1 * 10^{-9}$	1025,00	7,44	623,67	41,36
		$2,5 * 10^{-9}$	1025,10	7,43	622,84	40,19
		$4 * 10^{-9}$	1025,10	7,38	623,04	40,19
	50	$1 * 10^{-9}$	314,90	7,08	2042,46	49,13
		$2,5 * 10^{-9}$	314,90	7,03	2042,51	48,49
		$4 * 10^{-9}$	314,90	6,99	2042,64	48,45
	120	$1 * 10^{-9}$	27,30	5,59	23544,29	57,26
		$2,5 * 10^{-9}$	27,20	5,57	23546,02	57,04
		$4 * 10^{-9}$	27,30	5,60	23545,20	56,88
16	35	$1 * 10^{-9}$	723,00	16,91	888,22	59,52
		$2,5 * 10^{-9}$	723,10	16,89	887,87	59,41
		$4 * 10^{-9}$	723,10	16,74	887,89	59,23
	50	$1 * 10^{-9}$	134,30	16,18	4786,35	64,11
		$2,5 * 10^{-9}$	134,35	16,22	4786,12	64,12
		$4 * 10^{-9}$	134,35	16,20	4786,12	64,06
	120	$1 * 10^{-9}$	7,00	15,62	91787,26	69,02
		$2,5 * 10^{-9}$	7,00	15,86	91787,15	68,97
		$4 * 10^{-9}$	7,00	15,84	91787,17	69,08
20	35	$1 * 10^{-9}$	165,35	19,36	3889,13	66,85
		$2,5 * 10^{-9}$	165,40	19,55	3884,70	66,39
		$4 * 10^{-9}$	165,40	19,79	3885,03	66,35
	50	$1 * 10^{-9}$	39,40	17,98	16302,61	74,40
		$2,5 * 10^{-9}$	39,50	18,45	16303,76	74,27
		$4 * 10^{-9}$	39,50	18,37	16303,60	74,31
	120	$1 * 10^{-9}$	5,10	18,26	125976,50	80,04
		$2,5 * 10^{-9}$	5,10	18,30	125977,77	80,06
		$4 * 10^{-9}$	5,20	18,65	125980,34	79,92

Quantidade de *contigs*

A cada conjunto de fragmentos fornecido como entrada para a execução de uma ferramenta, o número de *contigs* gerados pode variar. Nos testes com dados simulados o número de *contigs* teve variação significante de acordo com os parâmetros de entrada fornecidos para o GenFrag. Por exemplo, na execução da ferramenta SOAPdenovo há execuções em que a média do número de *contigs* gerados não ultrapassa 5,20 *contigs*, mas para esta mesma ferramenta há execuções em que a média chega a 1025,10 *contigs*, nos casos de fragmentos gerados pelo GenFrag possuirem cobertura 20 e tamanho 120, e cobertura 7 e tamanho 35, respectivamente.

Para as ferramentas que utilizam a técnica *de novo*, é notável a variação no número de *contigs* gerados conforme as variações das configurações dos *reads* fornecidos como entrada. Considerando ainda a ferramenta SOAPdenovo (ver Tabela 5.5), observe que, para um certo tamanho fixo de *reads* e valores crescentes de cobertura, o número de *contigs* no resultado das execuções passa de 1025,10 para 165,35 (ver Figura 5.1). Resultados semelhantes podem ser observados na Tabela 5.4, bem como na Figura 5.1 para a ferramenta SSAKE que, para fragmentos com tamanho 120 e cobertura 7, gerou uma média de 16,60 *contigs*, enquanto que para fragmentos de cobertura 20 e mesmo tamanho, a média foi de 4,20. As Figuras 5.2 e 5.3 confirmam o comportamento citado, bem como a proximidade dos resultados das ferramentas que utilizam a mesma técnica de remontagem entre si, conforme o tamanho dos fragmentos sofre variação.

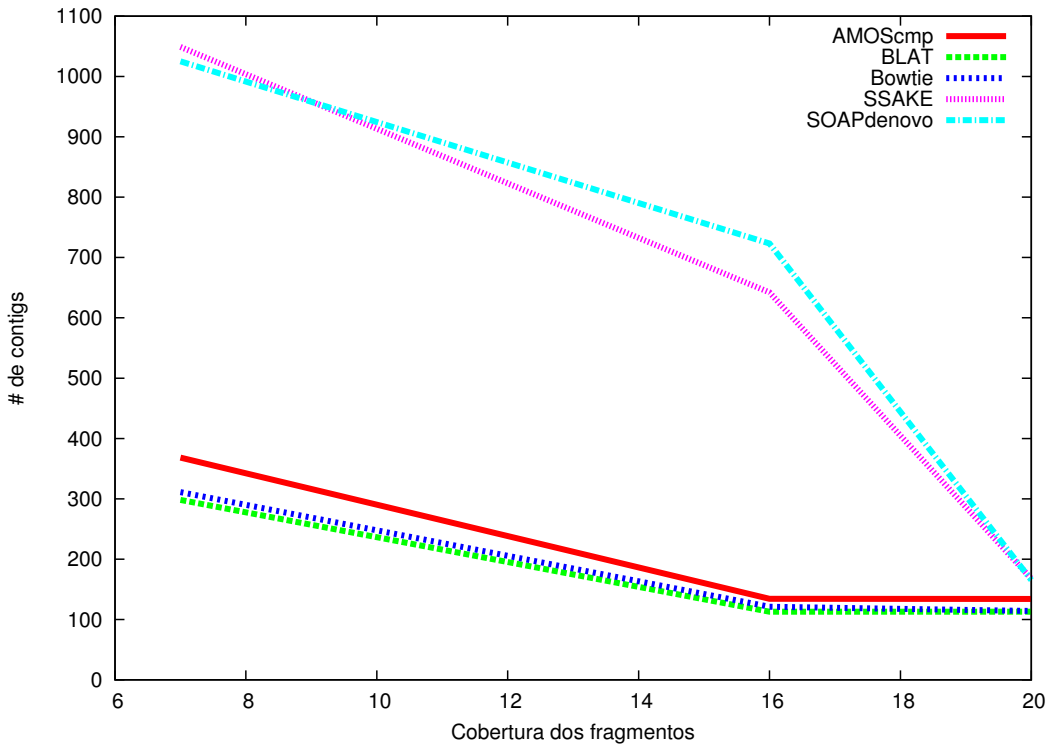


Figura 5.1: Variação na maior média de número de *contigs* para fragmentos de tamanho 35.

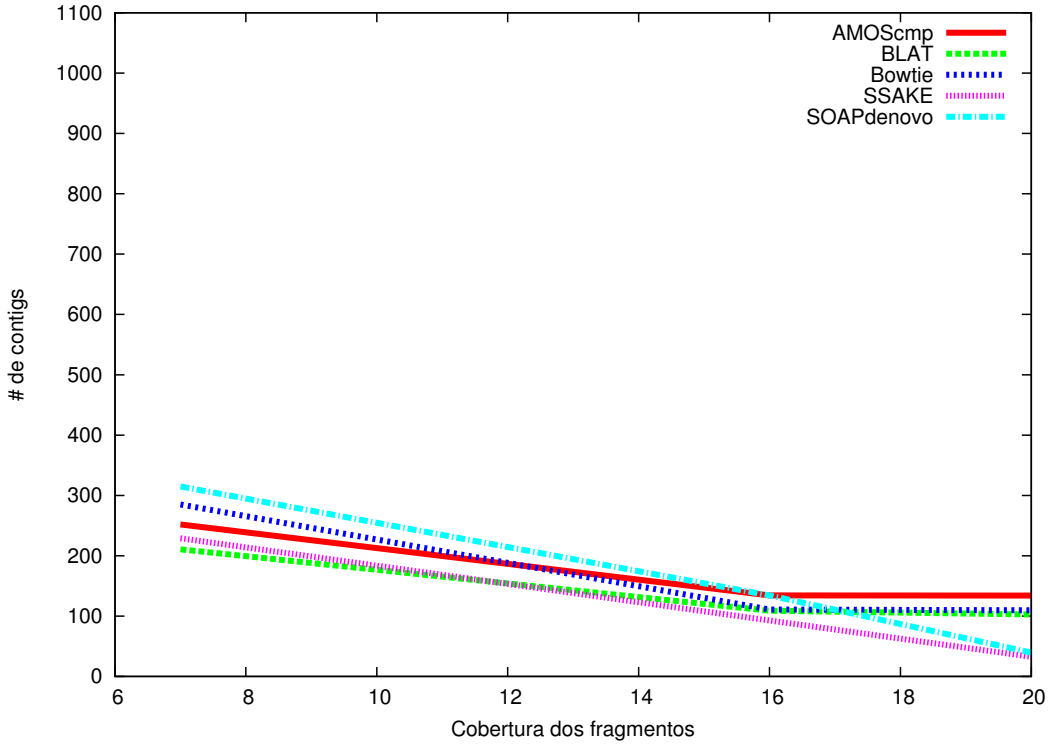


Figura 5.2: Variação na maior média de número de *contigs* para fragmentos de tamanho 50.

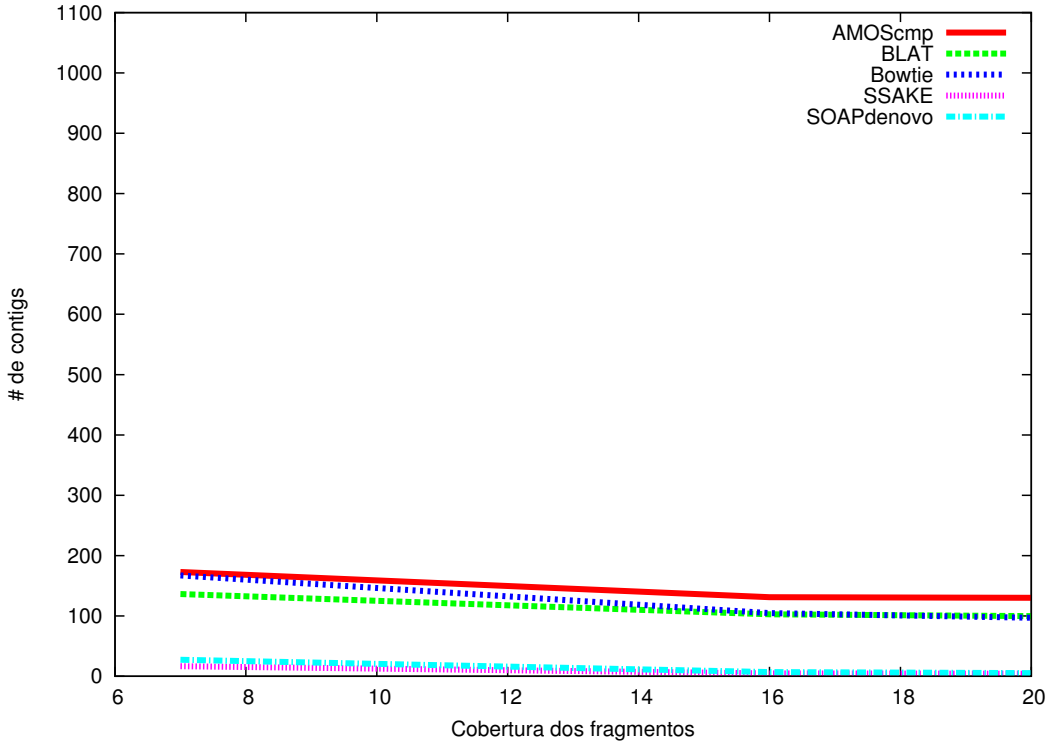


Figura 5.3: Variação na maior média de número de *contigs* para fragmentos de tamanho 120.

Conforme pode ser visto nas Figuras 5.4, 5.5 e 5.6, a variação no número de *contigs*, para *reads* de uma mesma cobertura, nos resultados das ferramentas *de novo* é mais evidente quando considera-se *reads* de tamanhos diferentes, enquanto as ferramentas de comparação praticamente não alteram o seu resultado. Por exemplo, para *reads* com cobertura 16 (Figura 5.5), a ferramenta SSAKE obteve média de 642,00 *contigs* para fragmentos de tamanho 35; média de até 92,85 para tamanho 50 e 4,8 para tamanho 120. Para a ferramenta SOAPdenovo e fragmentos de cobertura 20 e tamanho 35, a média do número de *contigs* gerados é por volta de 165,40. A média é de 39,50 para fragmentos de tamanho 50 e mesma cobertura e não ultrapassa 5,20 para fragmentos de tamanho 120.

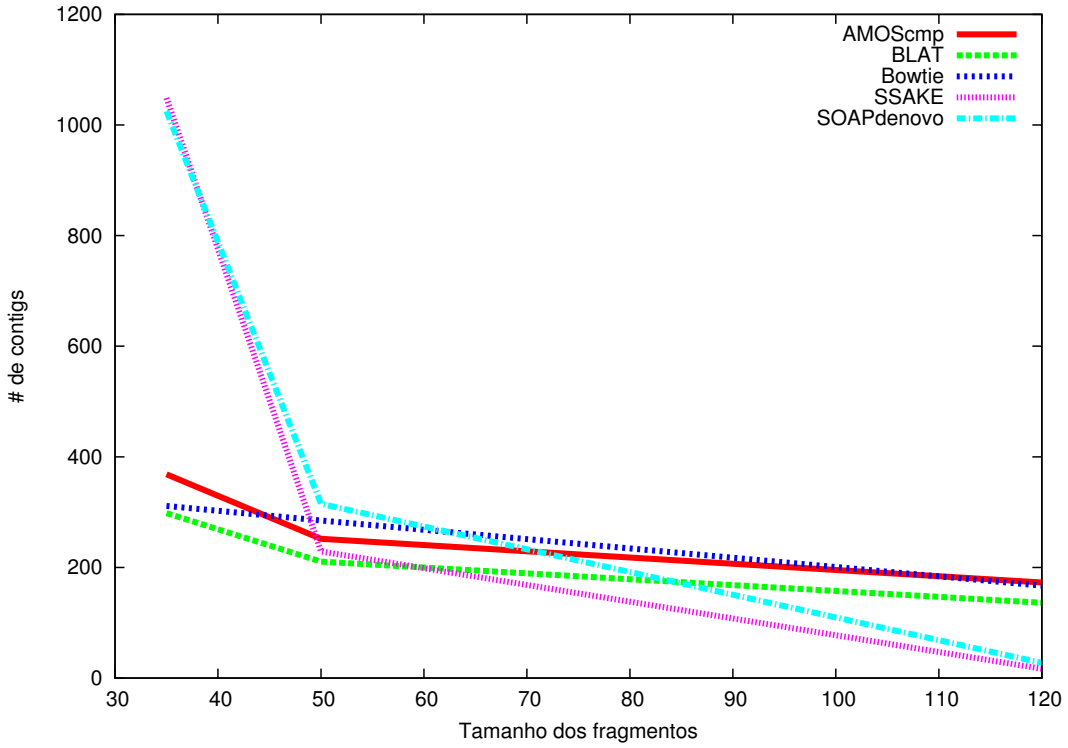


Figura 5.4: Variação na maior média de número de *contigs* para fragmentos de Cobertura 7.

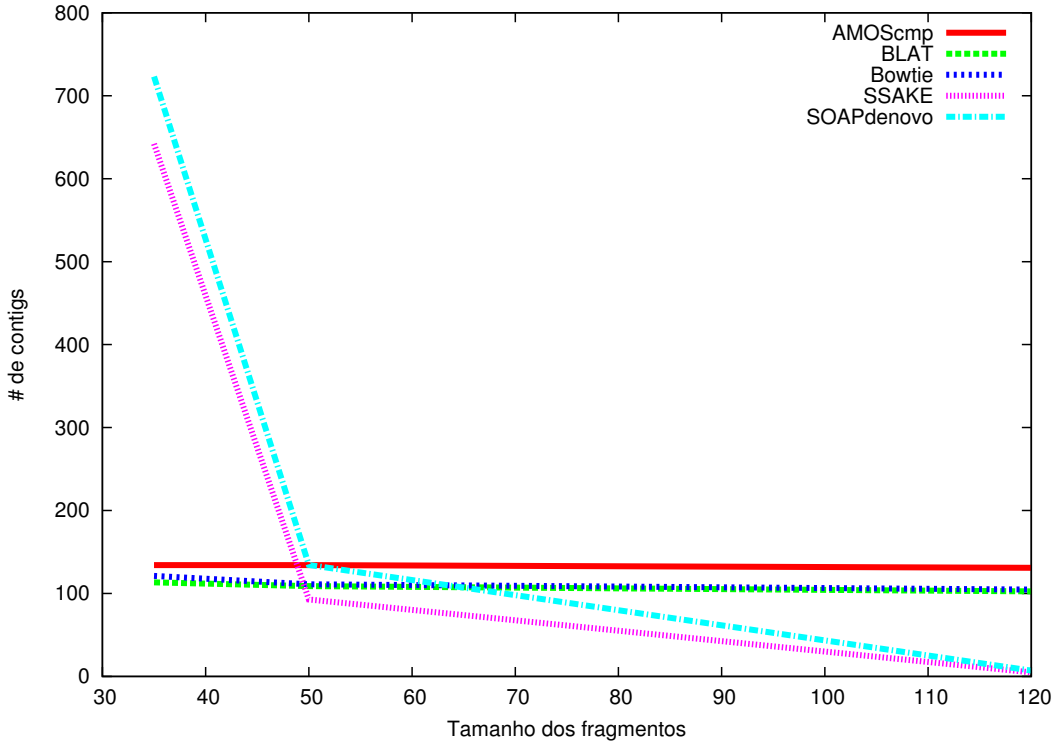


Figura 5.5: Variação na maior média de número de *contigs* para fragmentos de Cobertura 16.

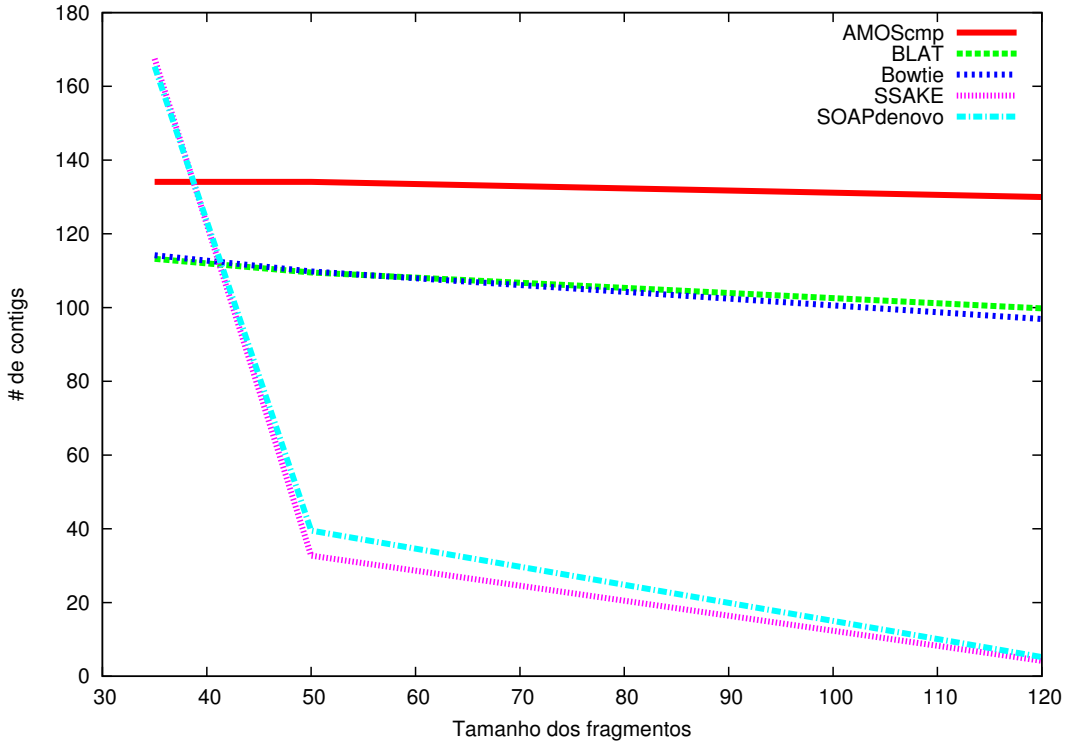


Figura 5.6: Variação na maior média de número de *contigs* para fragmentos de Cobertura 20.

As ferramentas que utilizam a técnica de comparação também apresentam variações na quantidade de *contigs* gerados quando as configurações dos fragmentos dados como entrada são alteradas, porém, como pode ser observado nas Figuras 5.1 e 5.5, essa variação é menos acentuada se comparada com a variação dos resultados das ferramentas de técnica *de novo*. Conforme os resultados para a ferramenta AMOScmp na Tabela 5.1, para fragmentos de entrada de tamanho 35 e cobertura 7, a média do número de *contigs* gerados é aproximadamente 368,40 *contigs*. Para o mesmo tamanho de *reads* e com cobertura 16, a média é de 134,20 *contigs*. Para a cobertura 20 a média foi sempre de 130 *contigs*, mesmo com alteração da taxa de erro do GenFrag.

As ferramentas Blat e Bowtie também utilizam a técnica de remontagem por comparação e a variação na média do número de *contigs* também não é tão acentuada. De acordo com a Tabela 5.2, para a ferramenta Blat e valor de cobertura 20 para os fragmentos de entrada, os tamanhos 35, 50 e 120 produzem médias de número de *contigs* de 111,30, 109,40 e 99,80 respectivamente. Para a ferramenta Bowtie a variação na média de *contigs* gerados para uma cobertura fixa também foi baixa se comparado com as ferramentas que não utilizam a técnica por comparação. Para a cobertura 7 e variação nos tamanhos de 35, 50 e 120, os valores respectivos das médias foram aproximadamente de 311,00, 284,30 e 167,20.

Para as variações nas taxas de erros na geração dos fragmentos não foram observadas relações de variações significativas na quantidade média de *contigs* gerados nas execuções de nenhuma das ferramentas de remontagem avaliadas.

Tendo como comparação o número médio de *contigs* gerados, a técnica *de novo* atingiu a menor média de número de *contigs* nos testes executados: 4,20 *contigs* para fragmentos de cobertura 20 e tamanho 120, utilizando a ferramenta SSAKE, enquanto que entre as ferramentas de remontagem por comparação, a menor média é de 96,20 utilizando a ferramenta Bowtie e fragmentos de mesmo tamanho.

Cobertura dos *reads* em relação aos *contigs*

Outra informação obtida dos resultados das ferramentas de remontagem é a cobertura média dos *contigs*. Esta cobertura se refere ao número médio de *reads* que representam cada nucleotídeo na sequência consenso correspondente a um *contig*.

É importante ressaltar que dois conceitos de coberturas são utilizados neste trabalho: o primeiro conceito refere-se à cobertura relacionada aos fragmentos gerados pelo simulador GenFrag em relação à SEQ-ALE e o segundo se refere à cobertura média dos *contigs*. Como nesta análise abordaremos com frequência os dois conceitos, chamaremos o primeiro de cobertura-genfrag e o segundo apenas de cobertura.

De um modo geral, as ferramentas apresentaram como resultados coberturas médias próximas das coberturas-genfrag. Além disso, nas Figuras 5.7, 5.8 e 5.9 também é possível observar que as coberturas obtidas como resultados de remontagem são, com poucas discrepâncias, próximas umas das outras independente da ferramenta utilizada. Em alguns casos a cobertura média é maior que a cobertura-genfrag dos fragmentos de entrada.

Como exemplo, na Tabela 5.1, para *reads* de cobertura-genfrag 7, em todos os tamanhos (35, 50 e 120) a cobertura média foi de 7,77. Este mesmo comportamento é observado nas ferramentas *de novo*. Veja, por exemplo, a Tabela 5.4, onde para a cobertura-genfrag 20 e para os tamanhos 50 e 120, as coberturas médias foram de até 20,20 e 21,40, respectivamente.

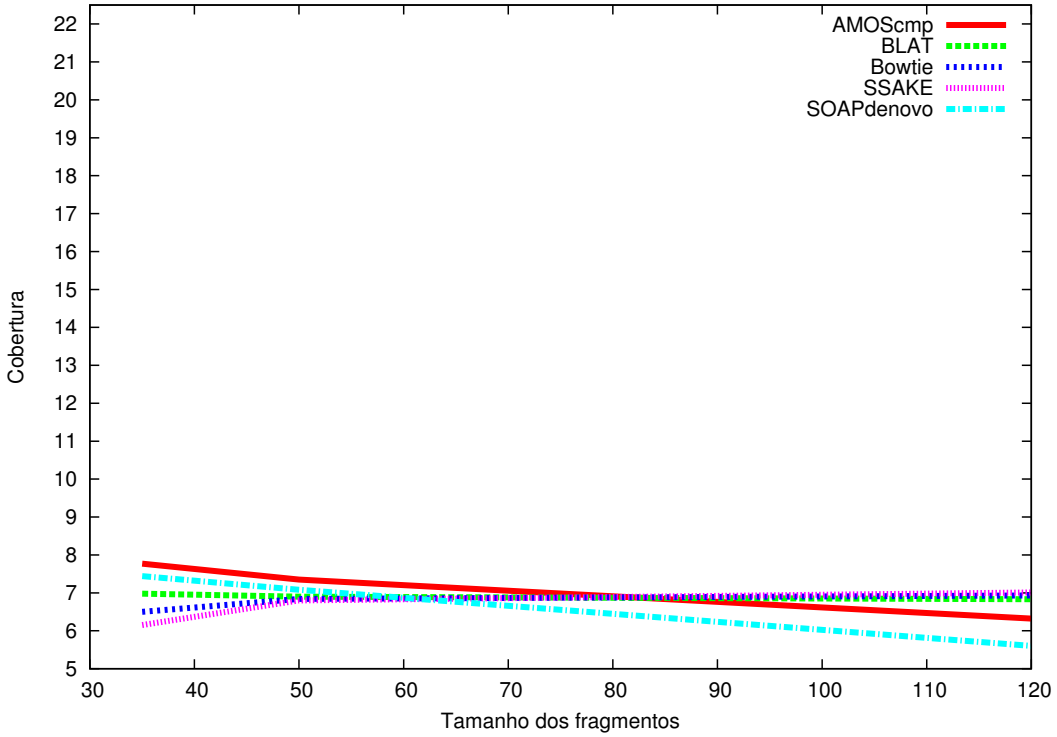


Figura 5.7: Variação na média de coberturas dos resultados de remontagem para fragmentos de cobertura-genfrag 7.

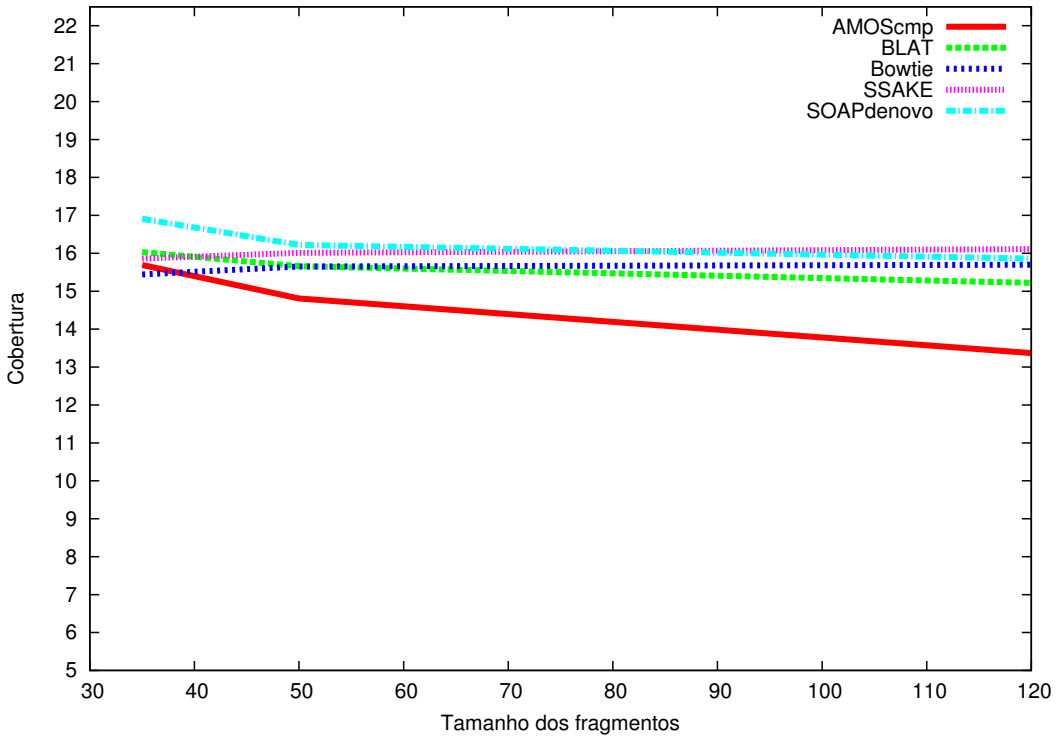


Figura 5.8: Variação na média de coberturas dos resultados de remontagem para fragmentos de cobertura-genfrag 16.

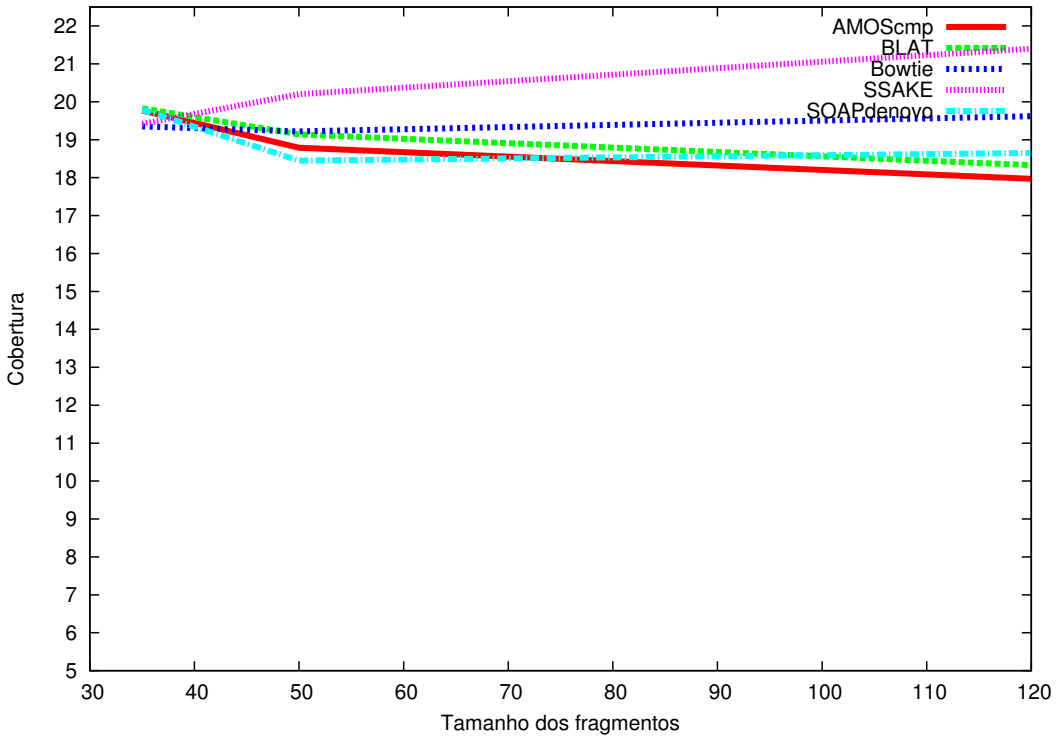


Figura 5.9: Variação na média de coberturas dos resultados de remontagem para fragmentos de cobertura-genfrag 20.

Entre os resultados das ferramentas *de novo* não houve um padrão na variação da média de cobertura dos *contigs* em relação à variação dos tamanhos dos fragmentos de entrada. Para a ferramenta SSAKE e cobertura 20 (ver Figura 5.9), há uma tendência de aumento da cobertura média com o aumento do tamanho dos *reads*. Para a cobertura-genfrag 20 e tamanho 35, a maior cobertura média foi de 19,42; para o tamanho 120, a maior cobertura média foi de 21,40. A cobertura média para o tamanho 50 foi de, no máximo, 20,20. Porém, para a ferramenta SOAPdenovo, não ocorreu o mesmo. Para a mesma cobertura-genfrag 20 e para o tamanho 35, a cobertura média foi de até 19,79. Para o tamanho 50, a cobertura atingiu o valor de 18,45. A cobertura média para o tamanho 120 teve mínimo de 18,26.

As ferramentas de remontagem pela técnica por comparação diferem-se pouco umas das outras na média de cobertura gerada como resultado. Bowtie é a única que a cobertura média não ultrapassa nenhuma vez o valor da cobertura-genfrag, isto pode ser verificado na Tabela 5.3. Isso não ocorre com a ferramenta Blat que para a cobertura-genfrag 16 e tamanho 35 atinge valor de 16,03. O mesmo é válido para a ferramenta SOAPdenovo que para a cobertura-genfrag 7, como dito anteriormente, a cobertura média obtida atingiu valores maiores que 7.

Assim como a análise dos resultados quanto ao número medio de *contigs* gerados, não foi possível notar variações significativas na cobertura média dos resultados devido à variações na taxa de erros.

De um modo geral as duas técnicas de remontagem obtiveram médias de coberturas, em seus resultados, próximas às respectivas coberturas-genfrag dos fragmentos submetidos como entrada. Assim, consideramos que ambas as técnicas tiveram um desempenho satisfatório na análise dos resultados com relação ao critério cobertura.

Tamanho dos *contigs*

Como é de se esperar, os tamanhos médios dos *contigs* gerados pelas ferramentas sofrem variações conforme o número médio de *contigs* varia. Nos testes com dados simulados, conforme o número médio de *contigs* aumenta, o tamanho médio dos *contigs* diminui, ou seja, estas duas variações são grandezas inversamente proporcionais.

Esta proporção ocorre tanto nas remontagens utilizando a técnica *de novo* quanto nas remontagens por comparação. Por exemplo, quando a ferramenta AMOScmp gerou média de 368,10 *contigs* (fragmentos de cobertura 7 e tamanho 35), o tamanho médio desses *contigs* é 1745,48 e quando gerou média de 172,90 *contigs* (fragmentos de cobertura 7 e tamanho 120), o tamanho médio desses *contigs* é 3729,22. A ferramenta SSAKE gerou média de 1047,90 *contigs* com tamanho médio de 615,16 para fragmentos de cobertura 7 e tamanho 35, enquanto que para fragmentos de mesma cobertura, tamanho 120 e quando a média de *contigs* é 16,70, o tamanho médio dos mesmos é 38715,49. A proporção entre a quantidade média de *contigs* gerados e seus tamanhos médios pode ser visto na Figura 5.10.

Como visto, o tamanho médio dos *contigs* guarda relação muito próxima com o número

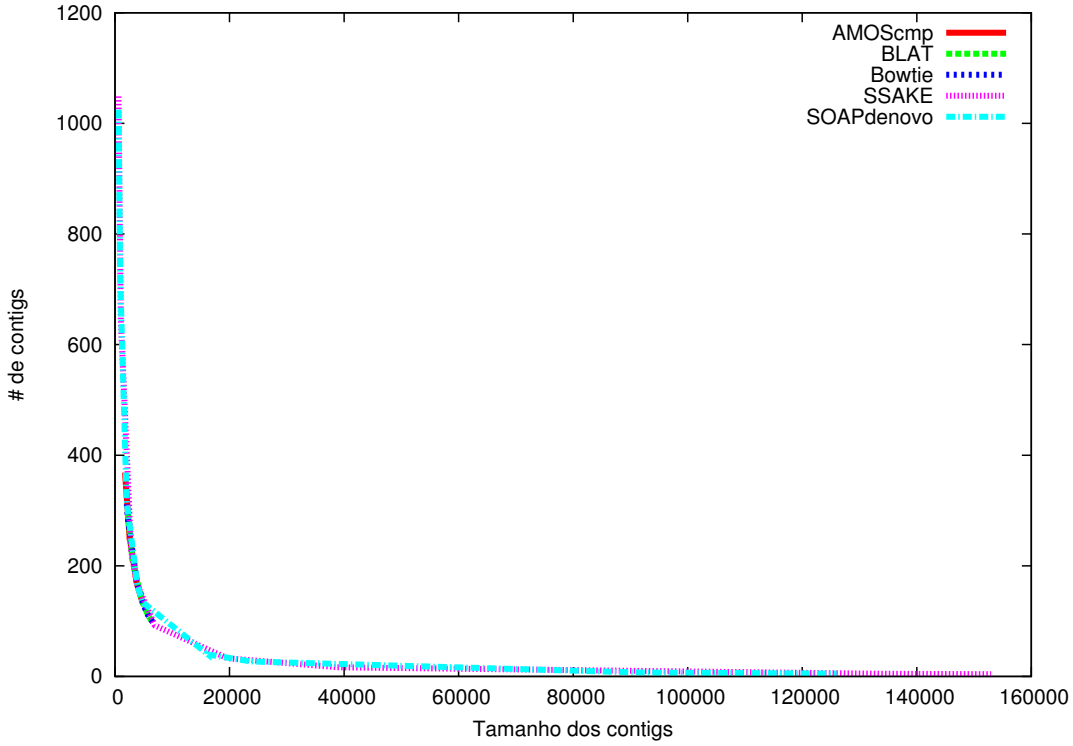


Figura 5.10: Proporção entre o número médio de *contigs* gerados e o tamanho médio de *contigs*.

médio de *contigs* gerados. Logo, para a análise quanto ao tamanho médio dos *contigs*, aplica-se a mesma análise feita quanto ao número médio de *contigs* gerados, porém seguindo a relação de proporcionalidade explicada anteriormente.

Corretude

A corretude é uma verificação, ou comparação, do resultado da remontagem de uma ferramenta com a sequência SEQ-ALE. Como a sequência SEQ-ALE é conhecida, comparamos os resultados produzidos pelas ferramentas de remontagem e, em forma de porcentagem, calculamos a corretude destes resultados, ou seja, a correspondência de igualdade entre as bases das sequências consensos dos resultados com as bases da sequência ótima a ser remontada.

Para as ferramentas que usam a técnica por comparação, a sequência SEQ-ALE auxilia no processo de remontagem, servindo como parâmetro para a escolha do posicionamento dos *reads*. Para o cálculo da corretude, comparamos base a base as sequências consensos de cada *contig* resultante dos posicionamentos dos *reads* e a sequência SEQ-ALE. A cada *gap* ou não correspondência exata entre as bases, foi computado um erro de remontagem e atribuido valor -1 e, caso contrário, um acerto de remontagem, atribuido valor +1. Ao final, o resultado da soma de acertos e erros de remontagem em relação ao tamanho da sequência SEQ-ALE corresponde à corretude dessa remontagem.

Como na remontagem por técnica *de novo*, as ferramentas não utilizam uma referência, os *contigs* e as respectivas sequências consensos são ordenados pelo processo de *scaffolding*. Os processos de *scaffolding* das ferramentas testadas geram como saída uma sequência contendo os *contigs* ordenados entre si e os *gaps* entre eles. Mais detalhadamente, o resultado do processo de *scaffolding* dessas ferramentas é uma sequência única, formada pela concatenação dos *contigs* e dos *gaps* entre eles. Os *gaps* entre esses *contigs* são representados por cadeias de caracteres especiais (diferente dos caracteres representantes das bases nitrogenadas A, C, T e G) idênticos, escolhidos na programação da respectiva ferramenta. Por exemplo, um *gap* de tamanho 4 caractere especial "X" seria representado por "XXXX". Como no processo de remontagem *de novo* não há relação de alinhamento entre o resultado do processo de *scaffolding* com o resultado ótimo de remontagem esperado (SEQ-ALE), foi utilizado para o cálculo da corretude o melhor alinhamento[1] entre as duas sequências. Assim como na técnica por comparação, os *gaps* e não correspondências exatas entre as bases das duas sequências foram considerados como erro de remontagem. O percentual de acertos de remontagem corresponde à corretude de remontagem.

Os resultados nas execuções das ferramentas com dados simulados apresentam variações quanto à média de corretudes. Para ferramentas que utilizam a técnica de remontagem por comparação, os valores de médias de corretudes variaram de 79,18% a 98,00%. As ferramentas *de novo* sofreram uma maior variação na média de corretudes dos resultados, o percentual mínimo e máximo são, respectivamente de 38,46% e 86,43%.

Para as técnicas de comparação, conforme os valores de cobertura dos fragmentos aumentam, a maior média de corretudes tende a aumentar (ver Figura 5.11). Para a ferramenta AMOScmp, a maior média de corretudes para fragmentos de cobertura 7 é 89,12%; para cobertura 16, a maior média é 91,66%; para cobertura 20 é 96,69%. A maior média para as coberturas 7, 16 e 20 para a ferramenta Blat são, respectivamente, 89,03%, 92,87% e 98,00% e para a ferramenta Bowtie são 89,20%, 92,41% e 96,90%.

Quando variados os tamanhos dos fragmentos para uma determinada cobertura, as ferramentas que utilizam a técnica de remontagem por comparação têm em seus resultados variações da maior média de corretudes, conforme exemplo na Figura 5.12. Para a ferramenta Blat e cobertura 20, a maior média de coberturas para fragmentos de tamanho 120 é 98,00%. Para fragmentos de tamanho 50, a maior média é 95,92%. Os fragmentos de tamanho 35 resultaram na maior média de 90,12%. Considerando a mesma cobertura de fragmentos e os tamanhos 35, 50 e 120, as maiores médias de corretudes para a ferramenta Bowtie são 89,97%, 94,76% e 96,90%. Nos resultados das execuções da ferramenta AMOScmp com fragmentos de mesma cobertura, as maiores médias de corretudes são 92,15% para fragmentos de tamanho 35, 95,32% para tamanho 50 e 96,69% para tamanho 120.

Além das variações descritas, a técnica de remontagem por comparação tem médias de corretudes diferentes para taxas de erros diferentes (ver Figura 5.13). Os fragmentos de cobertura 16, tamanho 50 e taxa de erro $1 * 10^{-9}$ ocasionaram médias de corretudes de 90,02% para a ferramenta Bowtie, 89,31% para a ferramenta AMOScmp e 90,15%

[1] O alinhamento entre as sequências foi feito apenas para posicioná-las a fim de computar a corretude do resultado. Portanto, não foram estabelecidos *gaps* de alinhamento entre sequências.

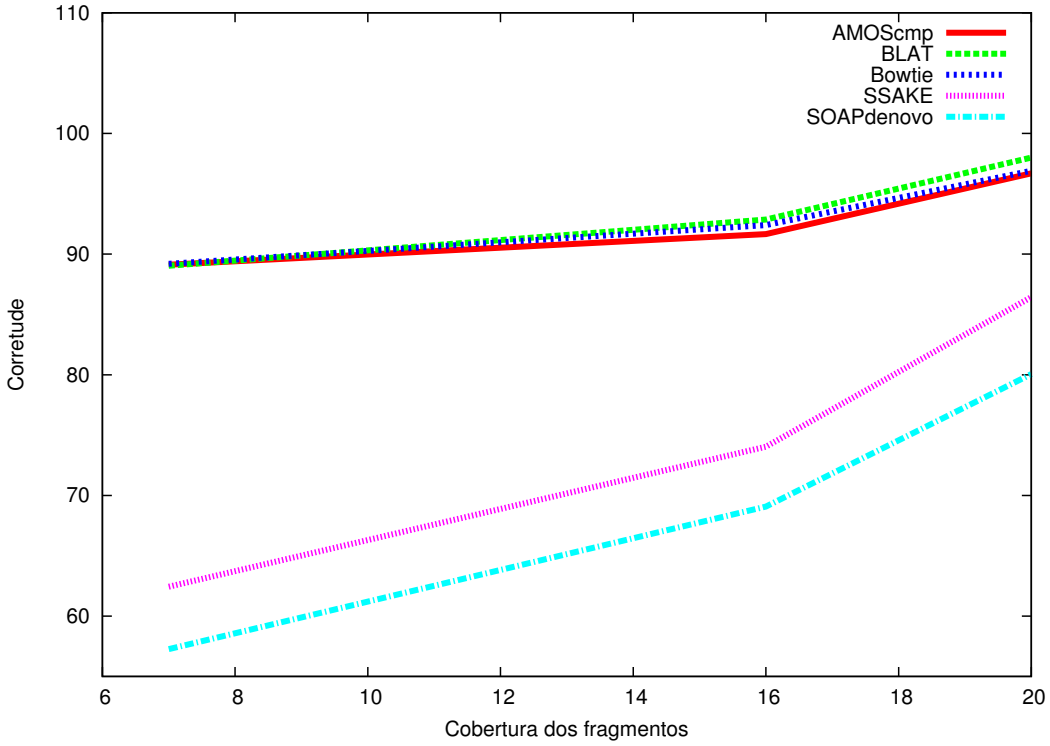


Figura 5.11: Porcentagens máximas de médias de corretudes dos resultados conforme a variação de cobertura dos fragmentos.

para a ferramenta Blat. Quando a cobertura e o tamanho dos fragmentos permanecem os mesmos, porém a taxa de erros é elevada para $4 * 10^{-9}$, as médias de corretudes para as ferramentas Bowtie, AMOScmp e Blat diminuem, respectivamente, para 88,66%, 86,24% e 84,70%. De forma geral, o mesmo ocorre para os outros tamanhos e coberturas de fragmentos, ou seja, quanto maior a taxa de erro, menor a média de corretudes no resultado obtido.

A média de corretudes nos resultados das ferramentas *de novo* também varia. Quando a cobertura dos fragmentos aumenta, a maior média de corretudes aumenta. Para a cobertura 7, a ferramenta SSAKE tem maior média de corretudes de 62,43% em seus resultados; a maior média de corretudes para a ferramenta SOAPdenovo é 57,26%. Os resultados com fragmentos de cobertura 16 e 20 geram resultados com as respectivas maiores médias de corretudes de 74,04% e 86,43% para a ferramenta SSAKE e 69,08% e 80,06% para a ferramenta SOAPdenovo.

Para os tamanhos 35, 50 e 120 e cobertura 16 dos fragmentos, a ferramenta SSAKE tem como respectivos resultados as maiores médias de corretudes de 38,47%, 43,49% e 74,04%. Considerando, respectivamente, as mesmas variações, os resultados para a ferramenta SOAPdenovo possuem maiores médias de corretudes de 59,52%, 64,12% e 69,08%. Assim ocorre para as outras coberturas, ou seja, com o aumento dos tamanhos dos fragmentos, as médias de corretudes tendem a aumentar. Para a cobertura 16 citada, essa tendência pode ser verificada na Figura 5.14.

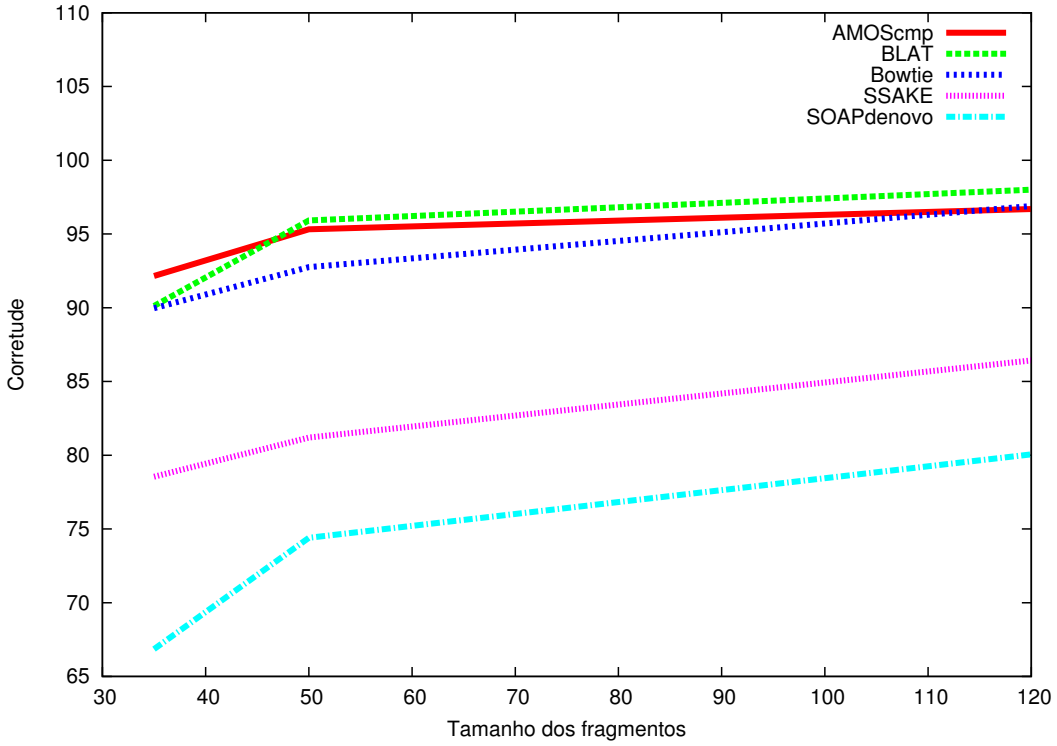


Figura 5.12: Porcentagens máxima s de médias de corretudes dos resultados para cobertura 20.

A variação na taxa de erro na geração dos fragmentos, em geral, representou proporcionalidade de quanto maior a taxa de erros, menores as médias de corretudes dos resultados produzidos com as ferramentas de tecnologia *de novo*. Mesmo assim, em algumas situações, essa relação não ocorreu. Por exemplo, conforme pode ser visto na Tabela 5.4, para a cobertura 16, tamanho 35 e taxa de erro $1 * 10^{-9}$, a média de corretudes para a ferramenta SSAKE é 38,47%. Para a taxa de erro $4 * 10^{-9}$, a média de corretudes é 38,52%.

A partir da análise sobre a perspectiva da corretude dos resultados, os resultados das ferramentas que efetuam a remontagem por comparação atingem maiores porcentagens se comparado com os resultados das ferramentas de remontagem *de novo*. Para a técnica de remontagem por comparação, as médias de corretude obtiveram variação entre 79,18% (AMOScmp) a 98,00% (Blat). Mesmo com médias de corretudes menores, a técnica *de novo* produziu resultados satisfatórios, obtendo valores entre 38,46% e 86,43%, ambos com a ferramenta SSAKE. Conforme dados da Tabela 5.4, é possível notar que, para alguns casos, a ferramenta SSAKE apresenta baixas taxas de corretudes (menores que 40%). Isso se deve ao fato de que no processo de *scaffolding* utilizado por essa ferramenta, os *contigs* foram ordenados de forma que o melhor alinhamento entre o resultado desse processo (*contigs* + *gaps*) com a sequência SEQ-ALE resultou em muitos erros de alinhamento entre as duas sequências.

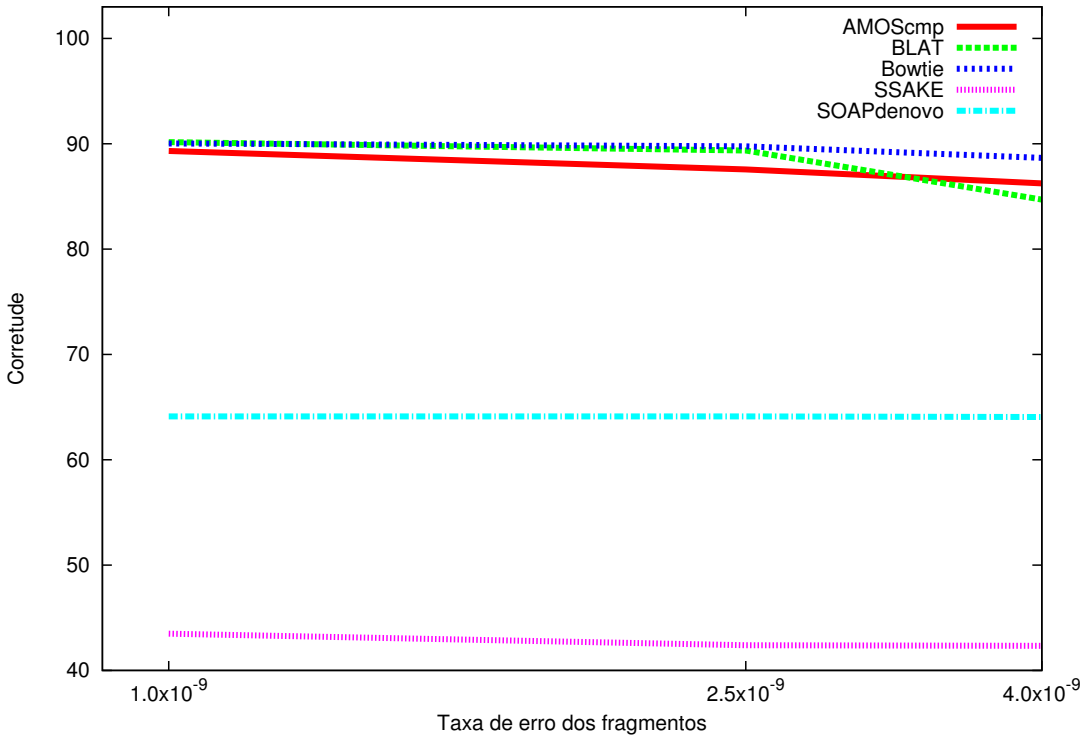


Figura 5.13: Porcentagem de médias de corretudes dos resultados para cobertura 16 e tamanho 50 dos fragmentos.

Tempo de execução

Nos testes realizados, as ferramentas de remontagem por comparação obtiveram médias de tempo menores em comparação com as médias de tempo de execução das ferramentas de remontagem *de novo*. As ferramentas de remontagem por comparação obtiveram médias muito próximas entre si, enquanto que a ferramenta SOAPdenovo foi, em média, 1 minuto mais demorada do que a ferramenta SSAKE. O gráfico na Figura 5.15 mostra as variações no tempo de execução entre as ferramentas.

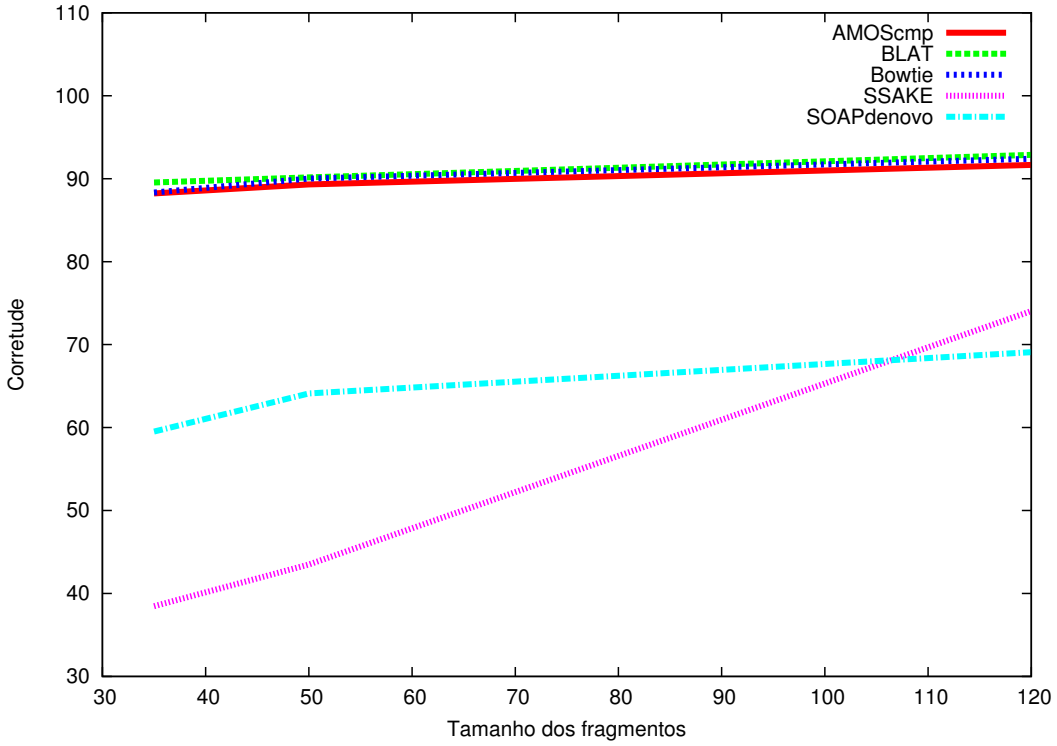


Figura 5.14: Porcentagens máximas de médias de corretudes dos resultados para cobertura 16.

5.2 Resultados com dados reais

Após as execuções das ferramentas com dados simulados, foram feitas novas execuções das ferramentas, porém com dados reais. Cada conjunto de *reads* foi submetido como entrada para todas as ferramentas de remontagem *de novo* e por comparação. Os resultados obtidos das remontagens estão sumarizados na Tabela 5.6 e é com base neles que são feitas as análises e comparações de desempenho das ferramentas e técnicas de remontagem.

Os fragmentos dos projetos cujos IDs no banco de dados são 29655, 29657, 29649 e 29659 foram gerados pela tecnologia de sequenciamento Roche 454. Como já dito anteriormente, os fragmentos gerados utilizando esta técnica de sequenciamento possuem tamanhos entre 200 e 300 pb. Os projetos 70391, 70301, 70307 e 70321 foram gerados pela tecnologia Illumina e possuem fragmentos de tamanhos de aproximadamente 35 pb. Para simplificar a leitura, denotamos fragmentos curtos os fragmentos gerados para tecnologia Illumina e fragmentos longos os gerados pela tecnologia Roche 454.

O menor número de *contigs* formados na remontagem por comparação de fragmentos longos é 38. Este resultado foi obtido com a ferramenta AMOScmp. As ferramentas Blat e Bowtie obtiveram mínimos respectivos de 44 e 50 *contigs*, também para fragmentos longos. A técnica de remontagem *de novo* obteve o número mínimo de *contigs* de 103 para a ferramenta SSAKE e 155 para a ferramenta SOAPdenovo com os fragmentos longos. Para os fragmentos curtos, o menor número de *contigs* gerados com a técnica de remontagem por comparação é 19 utilizando a ferramenta Blat. As ferramentas AMOScmp e Bowtie

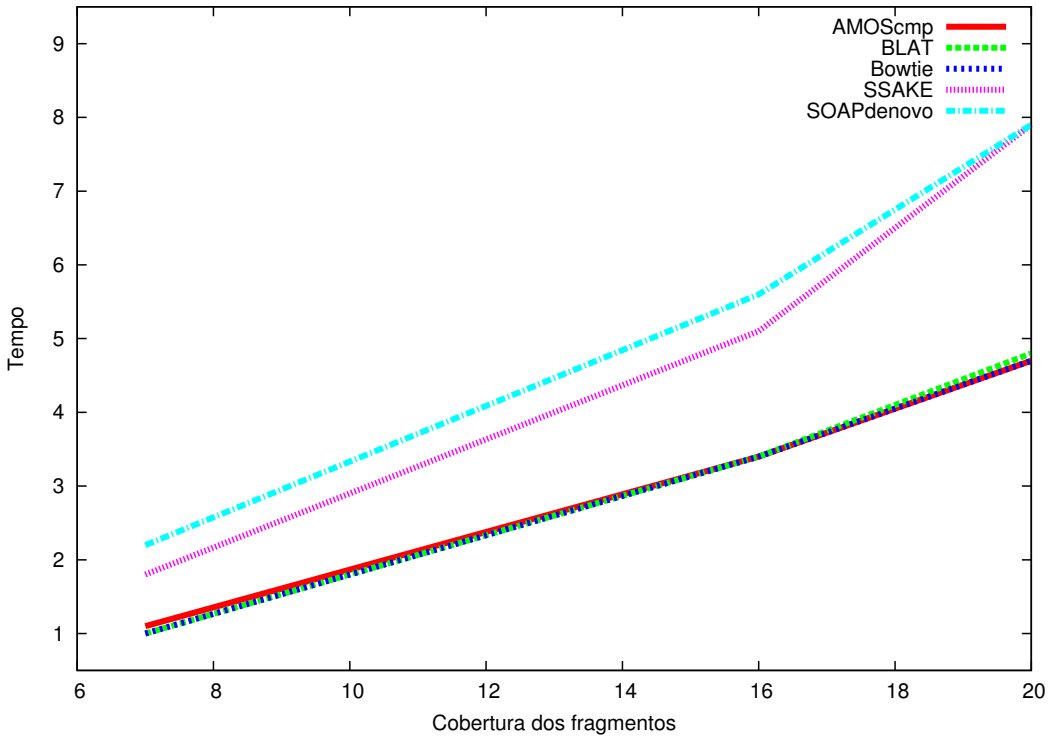


Figura 5.15: Variação em minutos no tempo de execução das ferramentas conforme valores das coberturas dos fragmentos.

geraram, para fragmentos curtos, os respectivos menores números de *contigs*: 21 e 28. Os menores números de contigs gerados pelas ferramentas de remontagem *de novo* são 408 para a ferramenta SSAKE e 485 para a ferramenta SOAPdenovo. As médias de *contigs* gerados pelas ferramentas de remontagem, tanto para fragmentos longos quanto para fragmentos curtos, podem ser vistas na Figura 5.16, sendo possível observar que a técnica de remontagem *de novo* gera mais *contigs* em ambos os casos.

Para os dados reais, as ferramentas que utilizam a técnica de remontagem por comparação produziram menores números de *contigs* para os fragmentos curtos se comparado com os resultados com os fragmentos longos. Nos resultados das ferramentas de remontagem *de novo*, ocorre o contrário, o número de *contigs* formados é maior com os fragmentos curtos. Nos resultados com dados simulados os números de *contigs* formados diminui conforme o aumento do tamanho dos fragmentos.

Se a análise dos resultados com dados reais for feita somente considerando o número de *contigs* gerados, estes resultados se confrontariam com os resultados com dados simulados. Por isso, é necessário que o desempenho das ferramentas seja avaliado considerando outros fatores, como, por exemplo, a cobertura de remontagem. As coberturas obtidas como resultado para os fragmentos longos e ferramentas de remontagem por comparação são baixas se comparadas com as coberturas obtidas para os mesmos fragmentos e ferramentas de remontagem *de novo*. Para os *reads* do projeto 29655, a cobertura obtida é de 2,98 utilizando a ferramenta AMOScmp, 2,92 utilizando a ferramenta Blat e 2,84 utilizando a ferramenta Bowtie. Isto ajuda a explicar a situação descrita anteriormente onde as

Tabela 5.6: Resultados obtidos das ferramentas de remontagem para dados reais.

Dados sobre os *contigs* resultantes da remontagem					
Ferramentas		ID SRA	Quantidade	Cobertura	Tamanho
Remontagem por comparação	AMOScmp	29655	56	2,98	1868,09
		29657	38	5,22	4049,96
		29649	41	4,97	3411,12
		29659	44	4,76	3163,08
		70391	23	15,01	7605,89
		70301	26	13,33	6188,18
		70307	21	13,90	7438,66
		70321	21	14,45	7907,79
	Blat	29655	57	2,92	1917,44
		29657	46	4,86	3368,10
		29649	44	5,03	3466,15
		29659	46	5,11	2940,95
		70391	19	15,74	9977,73
		70301	22	11,28	7601,53
		70307	19	12,90	8324,80
		70321	20	13,70	8022,86
	Bowtie	29655	62	2,84	1708,81
		29657	50	4,85	2898,90
		29649	51	4,81	2787,63
		29659	50	4,90	2704,04
		70391	28	15,37	5844,34
		70301	32	14,99	5095,08
		70307	29	15,29	5634,36
		70321	32	15,06	5157,17
Remontagem *de novo*	SSAKE	29655	103	7,02	637,09
		29657	176	8,57	1073,89
		29649	175	8,13	989,83
		29659	178	7,79	890,13
		70391	412	18,29	498,32
		70301	427	18,32	439,83
		70307	401	16,96	492,78
		70321	398	18,04	470,65
	SOAPdenovo	29655	167	8,17	698,13
		29657	155	9,15	1219,39
		29649	159	9,12	1089,44
		29659	158	8,86	1002,81
		70391	492	21,60	433,28
		70301	512	19,93	383,77
		70307	485	21,41	432,96
		70321	489	20,77	397,81

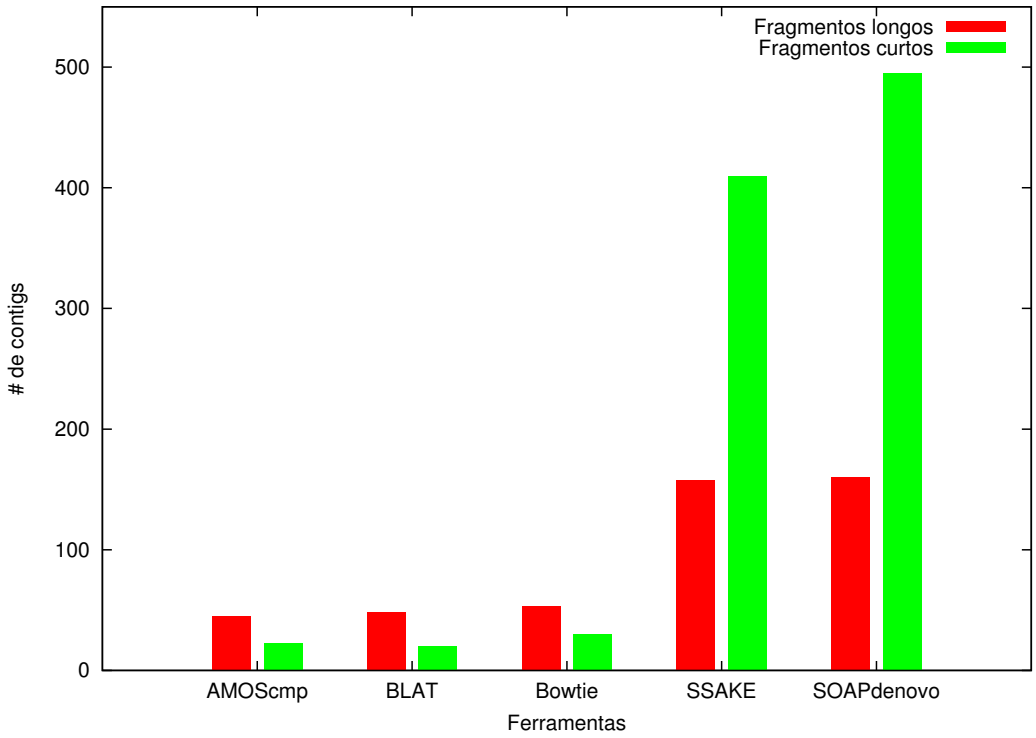


Figura 5.16: Média de *contigs* gerados para os fragmentos longos e curtos.

coberturas dos *contigs* são baixas para fragmentos longos e ferramentas de remontagem por comparação, pois, para os fragmentos curtos, estas mesmas ferramentas obtiveram, respectivamente, coberturas mínimas de 13,33, 11,28 e 14,99. A cobertura baixa se deve ao fato de existirem menos *reads* a serem alinhados já que estes são maiores do que os fragmentos curtos, bastando notar a diferença no número de bases e tamanho dos arquivos na Tabela 4.1.

A figura 5.17 representa as médias de coberturas de remontagem obtidas pelas ferramentas com os fragmentos longos e curtos, respectivamente. Em ambos os tipos de fragmentos, as ferramentas de remontagem *de novo* produziram valores maiores para a cobertura de remontagem se comparado com os valores dos resultados para as ferramentas de remontagem por comparação.

Dessas execuções, podemos concluir que as ferramentas de remontagem *de novo* geram mais *contigs* e com maior cobertura de remontagem. Isto leva a questionar se as ferramentas de remontagem por comparação não utilizaram um volume maior de fragmentos de entrada na geração de *contigs* do que as ferramentas de remontagem *de novo*. Porém, embora as ferramentas de remontagem por comparação obtenham como resultado menos *contigs* com cobertura menor, os tamanhos dos *contigs* gerados por esta técnica de remontagem são maiores que tamanhos dos *contigs* gerados pela outra técnica de remontagem. Os tamanhos médios de *contigs* gerados nos resultados com fragmentos longos e curtos são mostrados na Figura 5.18.

Infelizmente, a maioria das ferramentas não produzem como resultado informações

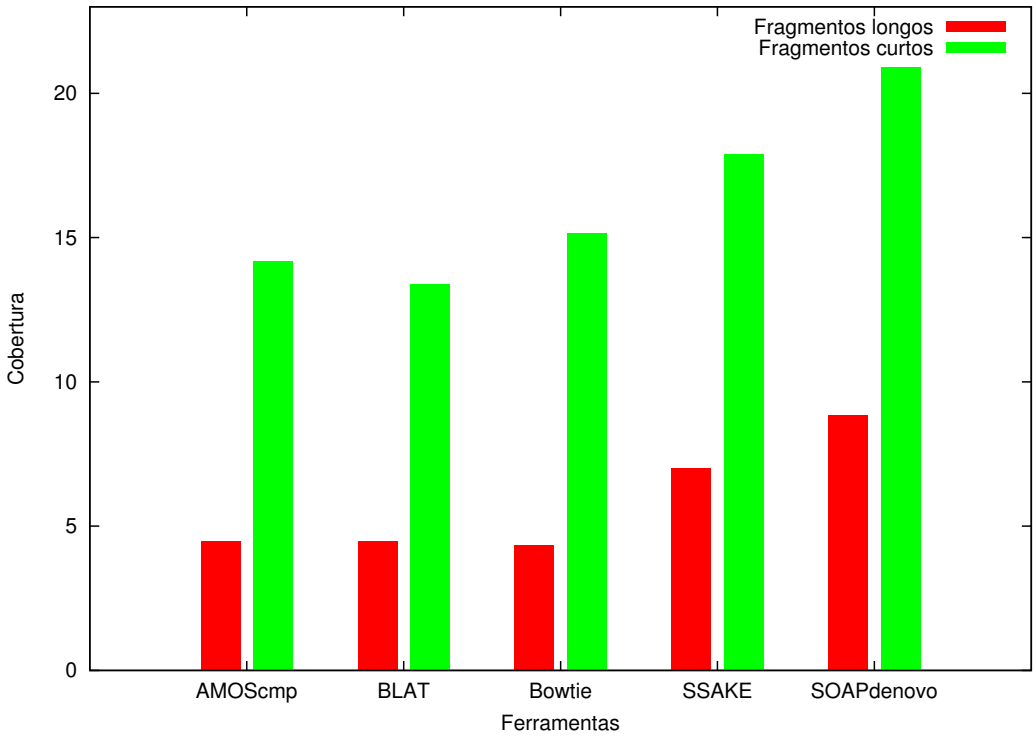


Figura 5.17: Média de coberturas dos *contigs* gerados para os fragmentos longos e curtos.

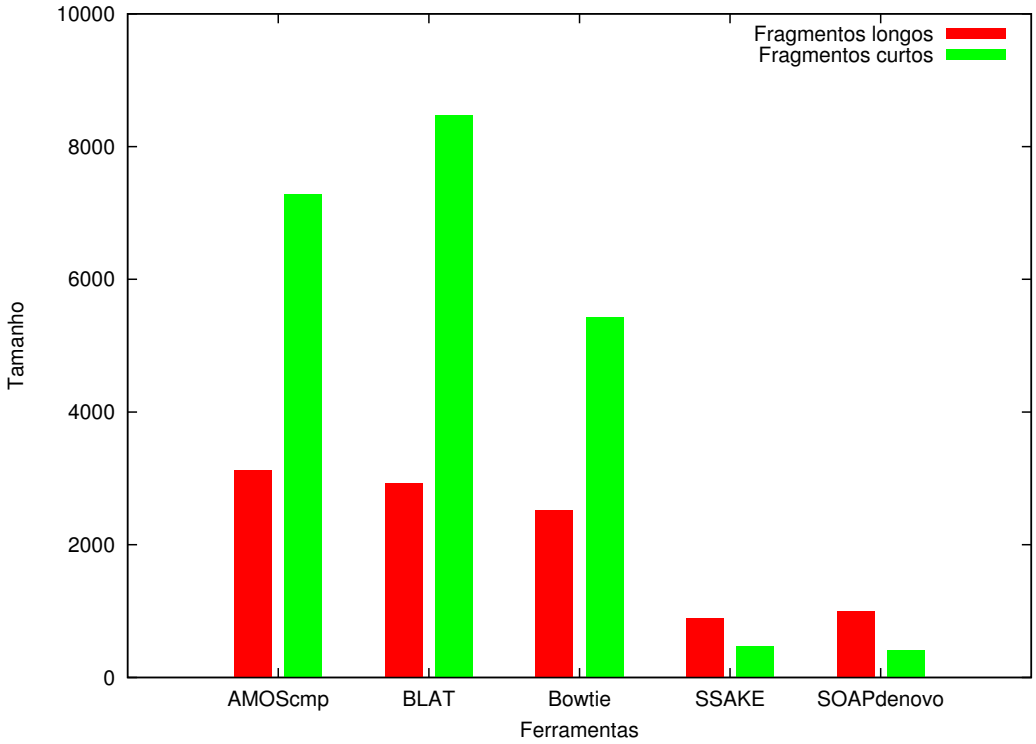


Figura 5.18: Média de tamanhos dos *contigs* gerados para os fragmentos longos e curtos.

sobre o quantitativo de fragmentos não utilizados na remontagem, impossibilitando uma comparação mais precisa sobre este aspecto.

Para efetuar a remontagem de *reads*, primeiramente é necessário escolher a técnica de remontagem. A remontagem *de novo* gerou, nos testes com dados reais, maiores números de *contigs*, tanto para fragmentos Illumina, quanto para fragmentos Roche 454, se comparado com os resultados utilizando a técnica de remontagem por comparação. Assim como nos resultados com dados simulados, o número de *contigs* para os resultados com dados reais é inversamente proporcional ao tamanho destes.

As ferramentas que remontam *reads* por comparação obtiveram resultados semelhantes. Isto ocorre, pois o procedimento que efetua a geração de *contigs*, pós mapeamento dos fragmentos, é o mesmo, utilizando o *script make-consensus*. Como o tempo médio de execução das ferramentas, bem como a cobertura dos *contigs* gerados por elas também são semelhantes, não há discrepância entre os resultados que priorize o uso de uma ferramenta em detrimento da outra. Na remontagem pela técnica *de novo*, a ferramenta SSAKE produziu menos *contigs*, porém com tamanhos menores, se comparado com os resultados produzidos com a ferramenta SOAPdenovo. A cobertura entre os resultados é ligeiramente maior para a ferramenta SOAPdenovo.

Quando uma única sequência consenso é produzida como resultado, espera-se que o genoma em questão tenha sido remontado completamente. Contudo, é precipitado dizer que quanto menos *contigs* foram gerados, mais próximo da remontagem completa está o resultado, pois nada garante que a porcentagem de erros de remontagem em um resultado que tenha menos ou mais *contigs* seja menor ou maior, além de que o resultado pode ter sido influenciado pelo genoma-referência, caso a remontagem tenha sido feita por comparação. Portanto, como nos resultados com dados simulados não é possível aferir a corretude dos mesmos, a escolha de uma ferramenta deve ser guiada pelas informações que se deseja obter como resultado de remontagem. Por exemplo, caso seja de maior utilidade que os *contigs* formados tenham o maior tamanho possível, a técnica de remontagem que obteve maiores *contigs* nestes testes foi a técnica de remontagem por comparação. A ferramenta que atingiu o maior tamanho médio de *contigs* foi a ferramenta Blat com tamanho médio de 9977,73.

5.3 Limitações e dificuldades

Em todo o processo de estabelecer o comparativo dos resultados obtidos pelas ferramentas existiram limitações respeitadas e dificuldades superadas.

A principal limitação quanto à escolha das ferramentas a serem utilizadas neste estudo se deu pelo fato que a maioria delas foram desenvolvidas para receber como entrada arquivos de *reads* de tecnologias de sequenciamento específicas. Como precisávamos submeter um mesmo conjunto de *reads* a diferentes ferramentas, o principal parâmetro de escolha foi a capacidade da ferramenta em remontar *reads* oriundos das tecnologias de sequenciamento Illumina e Roche 454.

Uma vez escolhidas as ferramentas, foi necessário lidar com as diferenças na forma

como os resultados são apresentados por cada uma delas. Enquanto as ferramentas AMOScmp, SSAKE e SOAPdenovo retornam os *contigs* e os *gaps* entre eles, as ferramentas Blat e Bowtie retornam as posições onde os *reads* foram mapeados no genoma-referência. Desta maneira, foi necessário ainda processar as saídas destas ferramentas para que os resultados pudessem então ser comparados.

Uma grande dificuldade encontrada na realização dos testes foi o espaço em disco necessário para armazenar os resultados obtidos. Quanto maior o tamanho e cobertura desejados para os *reads* gerados pelo GenFrag, o espaço necessário de armazenamento dos resultados das ferramentas de remontagem aumenta significativamente. Por exemplo, o menor espaço de armazenamento necessário para executar todas as ferramentas testadas foi com fragmentos gerados com cobertura 7 e tamanho 35. A execução das ferramentas de remontagem para tal configuração de *reads* precisou ser dividida em 3 etapas, cada uma gerando aproximadamente 40 Gb de dados e sobrecarregando o espaço de diretório disponibilizada para uso na máquina de teste. Para coberturas e tamanhos maiores dos fragmentos, as execuções das ferramentas de remontagem precisaram ser divididas em um número maior de etapas. Essa dificuldade impossibilitou que fossem testados os resultados para outros tamanhos e coberturas de fragmentos interessantes de serem analisados, como coberturas e tamanhos muito grandes na geração do fragmentos.

O tempo de execução das ferramentas também pode ser considerado um limitante. Muitas etapas de execução dos testes realizados terminavam a execução após mais de 12 horas de processamento. Por exemplo, para fragmentos de cobertura 16, o tempo médio de remontagem foi de 3,4 minutos para as 3 ferramentas de remontagem por comparação (AMOScmp, Blat e Bowtie). Para os mesmos fragmentos, o tempo médio de remontagem foi de 5,1 minutos para a ferramenta SSAKE e 5,6 minutos para a ferramenta SOAP-denovo. Portanto, somando o tempo de execução das 5 ferramentas, a execução de um conjunto de fragmentos de cobertura 16 é realizada em tempo médio de 20,9 minutos. Como, para cada valor de cobertura, existem conjuntos de fragmentos de 3 tamanhos, 3 taxas de erros e 10 execuções distintos, a remontagem com todas ferramentas e com todos os fragmentos de cobertura 16 é feita em 1881 minutos, ou ainda, 31,35 horas.

Capítulo 6

Conclusão

A fim de avaliar os resultados das técnicas de remontagem *de novo* e por comparação, as execuções das ferramentas de remontagem foram feitas submetendo como entrada conjuntos de testes simulados e reais. Para criação dos dados simulados foi utilizada a ferramenta GenFrag, capaz de simular, a partir de uma sequência de DNA dada como entrada, conjuntos de *reads* sob condições específicas de cobertura, tamanho e taxa de erro. Tais variações foram úteis na avaliação do comportamento das ferramentas estudadas.

As execuções das ferramentas de remontagem com dados simulados mostraram algumas tendências nos resultados produzidos. Quanto maior a cobertura dos *reads* a serem remontados, menor o número de *contigs* formados. Para um tamanho fixo e diferentes coberturas dos *reads*, as ferramentas que efetuam a remontagem por comparação geraram, em média, menos *contigs* que as ferramentas de remontagem *de novo*. Porém, para uma determinada cobertura fixa e aumentando os tamanhos dos *reads*, as ferramentas de remontagem *de novo* tendem a gerar menos *contigs* se comparado à técnica de remontagem por comparação.

Além disso, foi observada uma relação de proporcionalidade entre o número dos *contigs* e seus tamanhos. Neste relação, as duas grandezas, tamanho e quantidade de *contigs*, são inversamente proporcionais, ou seja, quanto mais *contigs* são formados, notou-se que menores eles são, independente da técnica de remontagem utilizada. As coberturas médias dos *contigs* variam de acordo com diferentes parâmetros de geração de fragmentos passados para o simulador GenFrag, porém não foi possível constatar tendências de comportamento do valor da cobertura dos *contigs* em relação às técnicas de remontagem.

A partir do pressuposto de que o resultado ótimo de remontagem para dados simulados era conhecido, foi estabelecida uma forma de avaliar a corretude dos resultados gerados pelas ferramentas. A corretude dos resultados foi calculada com base nos erros e acertos de remontagem, sofrendo variações conforme os parâmetros dos *reads* iam sendo alterados. Uma maior a cobertura de geração dos *reads* implicou em aumento das porcentagens máximas de médias de corretudes. Mesmo que de forma pouco expressiva, a taxa de erro influencia, em alguns casos, a corretude do resultado, de forma que quanto maior a taxa de erros, a média de corretudes dos resultados tende a diminuir.

O tempo médio de execução das ferramentas para dados simulados é muito próximo entre as ferramentas de remontagem por comparação, enquanto que as ferramentas de remontagem *de novo* possuem diferenciação no tempo de execução, em média, de 1 minuto.

Os fragmentos do conjunto de dados reais permitiu analisar situações reais de execução das ferramentas de remontagem. Os resultados das remontagens foram avaliados conforme as características dos fragmentos de entrada: curtos (fragmentos gerados pela técnica de sequenciamento Illumina) e longos (fragmentos gerados pela técnica Roche 454). As ferramentas de remontagem por comparação produziram mais *contigs* com fragmentos longos, enquanto que as ferramentas que utilizam a técnica *de novo* produziram mais *contigs* com fragmentos curtos. A proporcionalidade entre a quantidade de *contigs* e seus tamanhos observada nas execuções com dados simulados também ocorre quando as ferramentas efetuam a remontagem com dados reais. Por fim, também foi possível observar que, para dados reais, a cobertura dos *contigs* são maiores para fragmentos curtos, independentemente da técnica de remontagem.

Ambas as análises dos resultados com dados simulados e reais levaram à conclusão de que é necessário avaliar um conjunto de fatores sobre os fragmentos de entrada para escolher uma ferramenta que efetue sua remontagem. Como nem sempre é possível obter a remontagem completa de um genoma, a escolha de uma ferramenta de remontagem deve ser guiada pelos critérios que se deseja obter como resultado de remontagem. Por exemplo, no caso de não ser possível realizar uma remontagem completa do genoma (mais de um *contig*), na quantidade ou tamanho dos *contigs* gerados, algumas ferramentas, para um mesmo conjunto de fragmentos geram menos *contigs*, porém de maior tamanho, que outras ferramentas. Se o interesse do estudo não é no número de *contigs* formados ou em seus tamanhos, mas sim na cobertura destes, a escolha da ferramenta deve se basear nesse critério.

Com base nas execuções das ferramentas de remontagem, tanto com dados reais quanto com dados simulados, as Tabelas 6.1 e 6.2 sumarizam os resultados com base em critérios como quantidade mínima, cobertura máxima, tamanho máximo e corretude máxima dos *contigs* formados. As informações contidas nessas Tabelas visam auxiliar o processo de escolha de uma ferramenta de remontagem.

Tabela 6.1: Crítérios de avaliação das ferramentas que utilizam remontagem por comparação

Dados sobre os *contigs* resultantes da remontagem					
Tipo de dados	Ferramenta	Quantidade mínima	Cobertura máxima	Tamanho máximo	Corretude máxima (%)
Simulados	AMOS-cmp	130,00	19,77	4981,64	96,69
	Blat	99,80	19,83	6438,11	98,00
	Bowtie	96,20	19,62	6679,25	96,90
Reais	AMOS-cmp	21	15,01	7907,79	–
	Blat	19	15,74	9977,73	–
	Bowtie	28	15,37	5844,34	–

Tabela 6.2: Crítérios de avaliação das ferramentas que utilizam remontagem *de novo*

Dados sobre os *contigs* resultantes da remontagem					
Tipo de dados	Ferramenta	Quantidade mínima	Cobertura máxima	Tamanho máximo	Corretude máxima (%)
Simulados	SSAKE	4,20	21,40	152980,71	86,43
	SOAP-denovo	5,10	19,79	125980,34	80,60
Reais	SSAKE	103,00	18,32	1073,89	–
	SOAP-denovo	155,00	21,60	1219,39	–

Embora as dificuldades e limitações encontradas dificultem realizar mais comparações entre os resultados das ferramentas e técnicas de remontagem, os testes foram suficientes para construir uma base de conhecimento empírico sobre o tema. Espera-se que esse projeto de pesquisa contribua no processo de escolha de uma técnica de remontagem de fragmentos, bem como na escolha da ferramenta que desempenhe tal tarefa. Além disso, os conceitos apresentados ao longo de todo esse trabalho podem ser utilizados como contribuição de outros estudos em bioinformática ou áreas afins.

Apêndice A

Tecnologias de sequenciamento de DNA/RNA

A.1 *Roche 454*

O arquivo FASTA, na *Roche 454*, corresponde a um arquivo com extensão .fna e contém um identificador para o fragmento sequenciado, bem como a sequência de bases lidas de cada fragmento. O arquivo .qual contém a qualidade para cada uma das bases lidas de cada fragmento descrito no arquivo .fna. Seguem exemplos de arquivos FASTA e QUAL na *Roche 454*.

FASTA (.fna)

```
>E6PIHNP01B74B0
AACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACA
AAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGT
>E6PIHNP01DZVD8
GGGGTTGATCTTTCGCGCGTCACCGTTGGTCACTGCGAT
TCTCGGCGCGTACGTGCAGTTCGACACCATCGGCAAGAA
ATGCGCTACGGTTGGAACCGAAAGGGTTTGAATTCAAAC
GATAGCTTTGGCGTAGG
```

QUAL (.qual)

```
>E6PIHNP01B74B0
34 27 28 26 34 28 28 35 28 25 28 28 28
28 28 27 28 28 28 32 25 28 28 25 27 27
27 31 22 28 31 24 28 27 27 27 27 27 25
28 27 28 34 26 27 32 25 27 31 22 25 24
28 20 27 31 23 33 25 27 32 25 22 28 28
27 34 27 27 24 27 25 25 25 25 25 27 31
>E6PIHNP01DZVD8
34 24 14 3 34 27 28 27 27 28 36 32 13
```

```
28 28 28 27 27 27 27 27 28 34 26 27 34
27 34 26 28 27 28 28 28 28 28 28 27 28
27 28 27 27 28 34 27 27 28 39 35 22 9
28 34 27 27 28 27 28 27 27 28 27 28 28
28 28 27 32 25 28 28 27 28 28 28 27 27
28 27 28 28 34 27 28 35 28 28 34 27 27
28 27 34 28 27 27 27 28 37 33 15 34 27
27 28 28 38 34 22 8 28 28 27 27 28 33
28 27 28 25 27 28 28 35 28 32 24 33 25
33 26 35 28
```

A.2 *Illumina*

Um arquivo no formato *Illumina* FASTQ é mostrado abaixo.

```
@HWI-EAS255_4_FC2010Y_1_43_110_790
TTAATCTACAGAATAGATAGCTAGCATATATTT
+
hhhhhhhhhhhhhhdhhhhhhhhhhhdRehdh
```

Na primeira linha, o símbolo '@' inicia o nome ou identificador do fragmento ou *read*. A segunda linha contém a cadeia de nucleotídeos. O caractere '+' na terceira linha separa os caracteres correspondentes aos nucleotídeos dos caracteres correspondentes às qualidades desses nucleotídeos. Por fim, a última linha contém as qualidades dos nucleotídeos.

É válido lembrar que há mais de um formato padrão de arquivo FASTQ e eles diferem-se do Illumina FASTQ pelo sistema de codificação para representar as qualidades dos nucleotídeos. O Sanger FASTQ usa ASCII/33 a 126, correspondendo a valores de qualidades de 0 a 93. O Illumina FASTQ possui qualidades de 0 a 62 usando ASCII/64 a 126. Como exemplo, considerando o arquivo mostrado anteriormente, o caractere 'R' (ASCII 82) representa a qualidade 18 (82-64), 'e' representa a qualidade 37, 'h' representa a qualidade 40 e 'd' representa a qualidade 36.

Existem ainda outros formatos, menos utilizados, de arquivos gerados por essa tecnologia. Esses formatos são: SCARF, contendo também as informações dos nucleotídeos e suas qualidades em um arquivo único; e dois arquivos nos formatos SEQ e PRB, contendo, respectivamente, as informações dos nucleotídeos e as probabilidades (qualidades) de cada um deles.

A.3 *ABI SOLiD*

O arquivo de saída da tecnologia *ABI/SOLiD* tem o formato ColorSpace Fasta, ou CSFASTA (.csfasta). O arquivo contém uma indicação da coloração das bases a partir da

última base do *primer*. Cada *read* é identificado por um nome, ou identificador, precedido pelo símbolo '>'. Na linha seguinte ao identificador do *read*, o primeiro caractere corresponde à última base do *primer*, podendo ser identificado com uma das quatro letras: A, C, G e T. Cada caractere seguinte corresponde à coloração atribuida ao dNTP formado pelo nucleotídeo sequenciado e o nucleotídeo sequenciado imediatamente antes dele. Assim, a primeira coloração corresponde ao dNTP formado pelo último nucleotídeo do *primer* e o primeiro nucleotídeo sequenciado; a segunda corresponde ao dNTP formado pelo segundo e primeiro nucleotídeos sequenciados e assim por diante. A última coloração corresponde ao dNTP formado pelo último e penúltimo nucleotídeos sequenciados. Segue um exemplo de arquivo na extensão .csfasta.

CSFASTA (.csfasta)

```
>1_51_64_F3
T1030103123033323320333300021122223
>1_51_127_F3
T2010323233203132310110100200310 3102
```

O indicador único de cada *read* é gerado automaticamente pela ferramenta e é chamado de TAG ID. Existem 4 colorações diferentes, indicadas com os números de 0 a 3. Uma coloração é atribuida a cada par (após a identificação da última base do *primer*) de nucleotídeos em uma cadeia e, para tal fim, a Tabela A.1 é utilizada.

Tabela A.1: Tabela para determinar valores de coloração de dinucleotídeos na tecnologia *Applied Biosystems SOLiD instrument*.

	A	**C**	**G**	**T**
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

A coloração é determinada para cada par de nucleotídeos em sequência. O eixo X corresponde ao primeiro nucleotídeo e o eixo Y corresponde ao segundo nucleotídeo. Dessa forma, por exemplo, se a última base do *primer* for A e os núcleotídeos a serem sequenciados forem ACGT, o conjunto de nucleotídeos AA (última base do *primer* junto com núcleotídeo A) é codificado como 0, AC como 1, CG como 3 e GT como 1. Assim, AACGT seria codificado como A0131.

Da mesma forma que nas outras tecnologias, um arquivo contendo a qualidade de cada coloração atribuida as bases também é gerado e tem extensão ".QV.qual".

QUAL (.QV.qual)

```
>1_51_64_F3
```

```
12 7 21 16 6 2 25 5 25 26 6 7 2 8 5
2 3 2 6 21 5 2 3 9 4 2 2 2 17 6 2 2
2 5 3
>1_51_127_F3
3 18 15 4 11 2 6 4 4 6 2 7 2 9 4
3 2 6 18 2 2 4 3 2 2 2 2 2 2 4 2
3 4 4 2
```

Apêndice B

Ferramentas de remontagem de *reads*

B.1 Ferramentas de remontagem por comparação

AMOScmp

AMOScmp é uma ferramenta que efetua a remontagem de fragmentos utilizando a técnica de remontagem por comparação. A remontagem é efetuada em dois passos: alinhamento dos fragmentos ao genoma-referência utilizando a ferramenta de alinhamento MUMMer. Logo após, o mapeamento é processado gerando *contigs*. Mais informações sobre essa ferramenta podem ser encontradas em [16].

Bowtie

Bowtie é um *software* que realiza o mapeamento de *reads* em genomas referência com base na comparação. Segundo [3], Bowtie é capaz de mapear pequenos fragmentos de DNA de forma rápida e com uso eficiente de memória. Por exemplo, Bowtie é capaz de mapear 25 milhões de *reads* (na ordem de 35 bp) por hora tendo como referência o genoma humano. A saída gerada por este *software* pode ser utilizada para gerar sequências consensos de remontagem de fragmentos, bem como detecção de SNP's e indels.

Maq

Maq é uma ferramenta que realiza o alinhamento e a remontagem de *reads* mapeando-os a genomas-referências. Após o mapeamento, o algoritmo do Maq faz o passo de remontagem onde são geradas sequências consensos correspondentes aos *contigs* formados pelo mapeamento. O manual desse *software* pode ser encontrado em [5].

Mosaik

O manual do Mosaik pode ser encontrado em [6] e mostra que este *software* é composto por subprogramas que têm funções como transformar entradas para entradas suportadas pelo Mosaik, mapear *reads* por comparação e remontagem dos fragmentos. Mosaik possui código aberto e é composto por um fluxo de trabalho que realiza desde o mapeamento de pequenos fragmentos de DNA até a remontagem deles em um genoma completo.

SOAP3

SOAP3 é uma evolução de seus predecessores SOAP2 e SOAP. Este *software* é cerca de dez vezes mais rápido que o SOAP2 e toma como entrada um conjunto de pequenos fragmentos de DNA e um genoma referência. De acordo com [7], é capaz de mapear os fragmentos no genoma referência, gerar sequências consensos, bem como fazer a remontagem pela técnica *de novo*.

BLAT

BLAT foi desenvolvido para alinhar sequências de 25 ou mais pares de bases em genomas-referências. Esta ferramenta se assemelha com o BLAST, porém possui suas peculiaridades. A principal referência sobre o BLAT, bem como as diferenças entre esta ferramenta e a ferramenta BLAST podem ser vistas em [27].

B.2 Ferramentas de remontagem *de novo*

Abyss

Abyss pode ser encontrado em [8], utiliza o algoritmo do Bowtie para alinhar um *read* aos outros e é capaz de fazer a remontagem desses *reads* utilizando a técnica *de novo*.

ALLPATHS-LG

ALLPATHS-LG é uma ferramenta desenvolvida para os *reads* produzidos pelas tecnologias de nova geração (NGS) como Illumina/Solexa, por exemplo. Infelizmente, ALLPATHS-LG não é capaz de lidar com *reads* produzidos pela tecnologia Roche 454 nem pelo método clássico Sanger. É uma ferramenta otimizada para *reads* de 100 bp e seu manual e descrição podem ser encontrados em [9].

MIRA3

MIRA é um projeto que tem sendo executado com sucesso por 12 anos. MIRA3 é a última versão do *software* desse projeto onde é capaz de remontar fragmentos de DNA utilizando a técnica denovo. Segundo [10], esse *software* é capaz de remontar tanto fragmentos das tecnologias NGS quanto fragmentos da metodologia clássica.

Newbler

Software que remonta fragmentos de DNA utilizando a técnica *de novo* de remontagem. Foi desenvolvido pela Roche. Publicações ao seu respeito podem ser encontrados em [11].

SSAKE

Como pode ser visto em [12], SSAKE é um *software* robusto em termos de tempo de execução, foi desenvolvido em PERL e pode ser executado na plataforma UNIX. É capaz de alinhar e remontar *reads* pela técnica *de novo* de remontagem.

SHARCGS

Assim como outros *softwares* descritos anteriormente, SHARCGS é uma ferramenta capaz de fazer a remontagem de *reads* pela técnica *de novo* de remontagem. Em [24], essa ferramenta foi testada e se mostrou como uma alternativa robusta e eficaz em gerar sequências consensos, resultantes do processo de mapeamento de *reads* entre si e formação de contigs.

VCAKE

Capaz de executar o processo de remontagem a partir das informações presentes nos próprios *reads*, VCAKE é mais um *software* que utilizada a técnica *de novo* de remontagem. Sua descrição e código fonte podem ser localizados em [13].

Velvet

Velvet foi desenhado especificamente para remontar *reads* gerados pelas tecnologias NGS. Essa ferramenta gera *contigs* e suas respectivas sequências consensos. De acordo com [14], seu código fonte é livre e apresentará melhores resultados se executado em um sistema operacional compativel com Linux 64 bits.

SOAPdenovo

SOAPdenovo [17] é uma ferramenta de remontagem de genomas que utiliza a técnica *de novo*. O seu uso é otimizado quando *reads* provenientes da técnica de sequenciamento Illumina são utilizados como entrada.

Referências Bibliográficas

[1] http://www.454.com/about-454/index.asp, Feb. 2011. 2.3.2

[2] http://www.454.com/products-solutions/how-it-works/index.asp, Apr. 2011. 2.3.2

[3] http://bowtie-bio.sourceforge.net/manual.shtml#what-is-bowtie, Aug. 2011. 4.1, B.1

[4] http://trace.ncbi.nlm.nih.gov/Traces/sra/, Feb. 2011. 4.2.2

[5] http://maq.sourceforge.net/maq-man.shtml, Aug. 2011. B.1

[6] http://code.google.com/p/mosaik-aligner/downloads/detail?name=Mosaik%201.0%20Documentation.pdf, Aug. 2011. B.1

[7] http://soap.genomics.org.cn/, Aug. 2011. B.1

[8] http://www.bcgsc.ca/platform/bioinfo/software/abyss, Aug. 2011. B.2

[9] http://www.broadinstitute.org/science/programs/genome-biology/crd, Aug. 2011. B.2

[10] http://sourceforge.net/apps/mediawiki/mira-assembler/index.php?title=Main_Page, Aug. 2011. B.2

[11] http://www.454.com, Aug. 2011. B.2

[12] http://www.bcgsc.ca/platform/bioinfo/software/ssake, Aug. 2011. B.2

[13] http://sourceforge.net/projects/vcake/, Aug. 2011. B.2

[14] http://www.ebi.ac.uk/~zerbino/velvet/, Aug. 2011. B.2

[15] http://seqanswers.com/forums/showthread.php?t=43, Jan. 2012. 4.1

[16] http://sourceforge.net/apps/mediawiki/amos/index.php?title=AMOS, Jan. 2012. 4.1, B.1

[17] http://soap.genomics.org.cn/soapdenovo.html, Jan. 2012. 4.1, B.2

[18] http://www.illumina.com/truseq/quality_101/coverage/coverage_distribution.ilmn, Jan. 2012. 4.2.1

[19] Adi, S. S. *Identificação de Genes por Comparação de Seqüências*. PhD thesis, Universidade de São Paulo (USP), 2005. 2, 2.2, 1, 3.1

[20] Bentley, D. R. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, Nov. 2008. 2.3.3

[21] Brown, T. A. *Genomes*. 1999. 2.3.1, 2.3.1

[22] Cheung, L. M. *SimAffling - um ambiente computacional para suporte e simulação do processo de DNA shuffling*. PhD thesis, Universidade Federal de São Carlos (UFSCAR), 2008. 2

[23] A. L. Delcher. Alignment of whole genomes. *Nucleic Acids Research*, 27(11):2369–2376, Jan. 1999. 4.1

[24] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Research*, 17(11):1697–1706, Nov. 2007. 2.3, B.2

[25] M. L. Engle and C. Burks. Genfrag 2.1: new features for more robust fragment assembly benchmarks. *Computer Applications in the Biosciences*, 10(5):567–568, 1994. 4.2.1

[26] M. Holtgrewe, A. K. Emde, D. Weese, and K. Reinert. A novel and well-defined benchmarking method for second generation read mapping. *BMC Bioinformatics*, 12(1):210+, 2011. 2.3.3, 3.1

[27] J. J. Kent. BLAT - the BLAST-like alignment tool. *Genome research*, 12(4):656–664, Apr. 2002. 4.1, B.1

[28] McKernan, K. J. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, 19(9):1527–1541, Sept. 2009. 2.3.3

[29] Meidanis, J. and Setubal, J. C. *Introduction to Computational Molecular Biology*. PWS Publishing Co., 1997. Instituto de Computação - Universidade de Campinas (UNICAMP). 2

[30] M. L. Metzker. Emerging technologies in DNA sequencing. *Genome Research*, 15(12):1767–1776, Dec. 2005. 2.3.1, 2.3.2

[31] O. Morozova and M. A. Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264, Nov. 2008. 2.3.1, 2.3.2, 2.3.2, 3.2

[32] Neto, H. A. and Sousa, S. R. A. Classificação de Sequências de RNAs não-codificantes através da Distância de Compressão Normalizada., 2008. 2.1, 2.2

[33] M. Pop, D. S. Kosack, and S. L. Salzberg. Hierarchical scaffolding with Bambus. *Genome research*, 14(1):149–159, Jan. 2004. 2.3, 3.2, 3.2

[34] Qin, J. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, Mar. 2010. 2.3.3

[35] A. D. Rasko. Complete Sequence Analysis of Novel Plasmids from Emetic and Periodontal Bacillus cereus Isolates Reveals a Common Evolutionary History among the B. cereus-Group Plasmids, Including Bacillus anthracis pXO1. *Journal of Bacteriology*, 189, Jan. 2007. 4.2.2

[36] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, Dec. 1977. 2.3.1

[37] C. Trapnell and S. L. Salzberg. How to map billions of short reads onto genomes. *Nature Biotechnology*, 27(5):455–457, May 2009. 3.1, 3.1

[38] D. Turner, T. Keane, I. Sudbery, and D. Adams. Next-generation sequencing of vertebrate experimental organisms. *Mammalian Genome*, 20(6):327–338, June 2009. 3.2

[39] M. J. Wakefield. Genomics - from neanderthals to high-throughput sequencing. *Genome Biology*, 7:326, Aug. 2006. 2.3.2

[40] R. L. Warren, G. G. Sutton, S. J. Jones, and R. A. Holt. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 23(4), Feb. 2007. 4.1

[41] Wheeler, D. A. The complete genome of an individual by massively parallel DNA sequencing. 452:872–876, Apr. 2008. 2.3.3