Projeto de Mestrado

Montagem de Seqüências Curtas ou Muito Curtas

Instituto de Computação - UNICAMP

Candidato: Maria Angélica Lopes de Souza
Orientador: Prof. Dr. Zanoni Dias

Resumo

Este projeto visa estudar técnicas de montagem, buscando identificar as deficiências de algoritmos de montagem de genomas e clusterização de ESTs (*Expressed Sequence Tags*), em relação a seqüências curtas ou muito curtas, para o desenvolvimento de metodologias que supram essas deficiências.

Para seqüenciar um genoma (que pode ter bilhões de bases) realiza-se a fragmentação do mesmo e obtem-se as bases de cada parte. Após o seqüenciamento dos fragmentos é preciso montá-los para obter a seqüência completa.

A maioria das estratégias de montagem utiliza a sobreposição das bases de dois fragmentos para identificar se podem ser unidos na formação de um *contig*. No entanto, a montagem completa do genoma esbarra na presença de regiões repetidas.

Métodos de montagem tradicionais não exploram corretamente a presença de *reads* curtos ou muito curtos, levando a necessidade do desenvolvimento de algoritmos que aproveitem melhor as informações contidas nestes tipos de *reads* e ofereçam soluções para o problema de repetição. Além disso, esses métodos são focados na montagem do genoma inteiro, sendo a clusterização de ESTs uma estratégia alternativa e interessante para lidar com seqüências curtas ou muito curtas, e será abordada neste trabalho.

A linguagem de programação Python será usada para o desenvolvimento de programas para realizar a clusterização de *reads* curtos ou muito curtos, no intuito de diminuirmos os problemas apresentados por programas de montagem existentes. Os programas a serem desenvolvidos serão comparados com os principais já existentes e passarão por análises qualitativa e quantitativa. O projeto será desenvolvido em parceria com a Embrapa Recursos Genéticos e Biotecnologia (Cenargen), através do projeto universal CNPq (486303/2006) e do projeto FINEP (0886/2007).

1 Introdução

Este documento tem o objetivo de apresentar um plano a ser executado para o desenvolvimento do trabalho de mestrado. Na seção 2 são apresentados conceitos de biologia necessários para o entendimento do projeto proposto. A seção 3 apresenta a Bioinformática. Projetos de seqüenciamento são descritos na seçao 4. As seções 5 e 6 apresentam, respectivamente, estratégias de seqüenciamento e de montagem. Na seção 7 são apresentados os objetivos do projeto. Na seção 8 são descritos os materiais e métodos que serão usados no projeto, além da forma como os resultados serão analisados. Finalmente, o cronograma das atividades a serem desenvolvidas é indicado na seção 9.

2 Genômica

Genômica é a área da biologia que estuda o genoma de diferentes organismos. Todos organismos vivos apresentam informações que constroem seu material genético, estas constituem o genótipo do organismo e são transmitidas entre diferentes gerações por meio da reprodução. O genótipo de um

organismo aliado ao meio em que vive determina o seu fenótipo. Na maioria dos seres vivos o material genético é o DNA (ácido desoxirribonucleio), em alguns vírus é o RNA (ácido ribonucleico).

O DNA é um polinucleotídeo composto por quatro nucleotídeos (um nucleotídeo é formado por uma pentose, uma base nitrogenada e um grupo fosfato) diferenciados por suas bases nitrogenadas: adenina, citosina, timina e guanina, simbolizadas por A, C, T e G, respectivamente. Auxiliados por um grupo hidroxil os nucleotídeos são unidos pelo terceiro carbono da pentose de um nucleotídeo com o fosfato do nucleotídeo seguinte.

A estrutura do DNA é conhecida como de dupla hélice, pois ele é composto por duas fitas de polinucleotídeos em forma helicoidal ligadas por pontes de hidrogênio. As bases A e T e as bases C e G se complementam em fitas opostas. A figura 1b mostra as bases complementares e suas ligações por pontes de hidrogênio.

As ligações entre nucleotídeos fazem com que cada fita de DNA possua duas extremidades livres, a 3' (grupo hidroxil livre) e a 5' (grupo fosfato livre), de forma que a extremidade 3' de uma fita é correspondente a 5' da outra. Por convenção a escrita (leitura) de moléculas de DNA é feita em apenas uma das fitas na direção 5'→ 3', obtendo-se a outra fita pela inversão da ordem das bases e complementando cada uma delas, o que denomina-se complemento reverso. Na figura 1a é possível observar a estrutura de dupla hélice do DNA, assim como a ligação das bases das fitas complementares em sentidos opostos.

O RNA é um polinucleotídeo como o DNA, mas em lugar da base nitrogenada T possui a base U, uracila, complementar a base A, e, normalmente, formado por fita simples.

A genética é a área da biologia que realiza o estudo dos genes, trechos de DNA que codificam uma proteína ou RNA [7], estas são as principais informações existentes no material genético de um organismo. Tais informações são ativadas através de reações bioquímicas denominadas de expressão gênica, dividida nos processos de transcrição e tradução.

Através do processo de transcrição do gene é produzido o RNA mensageiro, mRNA, como uma cópia de uma das fitas do DNA, trocando T por U (podem ainda serem produzidos a partir de genes os RNAs transportador, tRNA, e ribossômico, rRNA).

O processo de tradução utiliza o mRNA produzido na transcrição para síntese de proteínas, que são formadas por diferentes aminoácidos. As bases do mRNA são lidas como códons (tripla de bases que especifica um aminoácido), existem 64 códons relacionados a 20 aminoácidos, com exceção de três códons que são utilizados para indicar o final da tradução e por isso não se relacionam com nenhum aminoácido. Na figura 2 pode-se observar o Código Genético Padrão, usados pela maioria dos organismos [15].

A tradução do mRNA em proteína é realizada no ribossomo (uma organela que percorre o mRNA) com auxílio de tRNAs, que transportam de um lado o aminoácido e de outro o anti-códon (códon com os nucletídeos complementares), que se acopla ao códon do mRNA, liberando o aminoácido a ser unido com a cadeia de aminoácidos que formará a proteína.

Em seres mais simples como bactérias (procariotos, sem membrana nuclear) em geral não há necessidade de pré-processamento para o processo de tradução, enquanto que em seres mais complexos

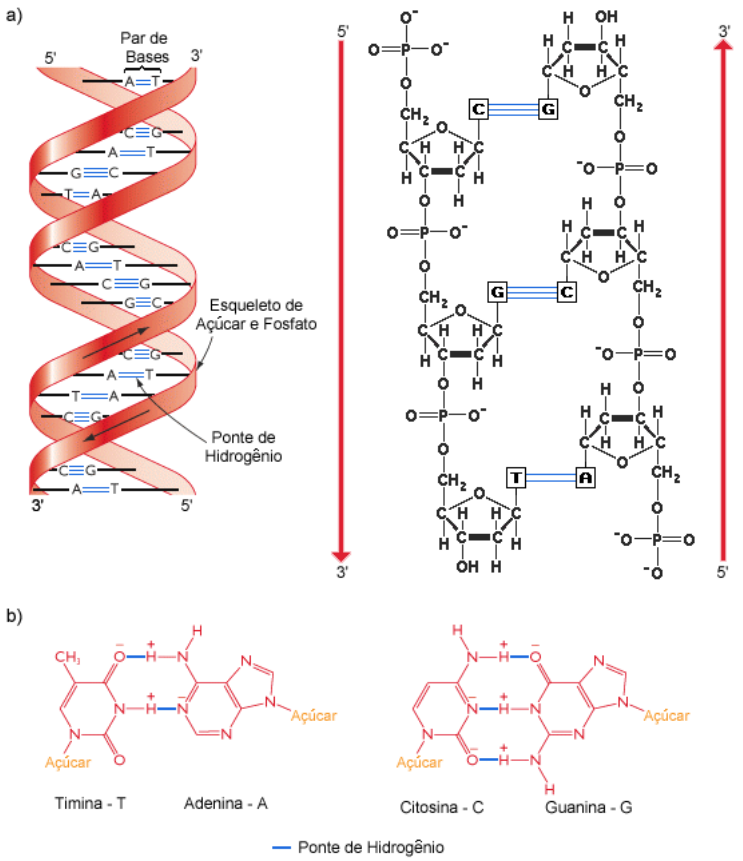

Figura 1: a) Estrutura de dupla hélice do DNA. b) Bases que compões o DNA e ligação por pontes de hidrogênio

Aminoácido	Cód. 3	Cód. 1	codon
Alanina	ALA	A	GCA,GCC,GCG,GCU
Arginina	ARG	R	AGA,AGG,CGA,CGC,CGG,CGT
Asparagina	ASN	N	AAC,AAT
Ácido Aspartico	ASP	D	GAC,GAT
Cysteína	CYS	C	TGC,TGT
Ácido Glutâmico	GLU	E	GAA,GAC
Glutamina	GLN	Q	CAA,CAG
Glycina	GLY	G	GGA,GGC,GGG,GGT,CAG
Histidina	HIS	H	CAC,CAT
Isoleucina	ILE	I	ATA,ATC,ATT
Leucina	LEU	L	CTA,CTC,CTG,CTT,TTA,TTG
Lysina	LYS	K	AAA,AAG
Metyonina	MET	M	ATG (Start-Codon)
Phenilalanina	PHE	F	TTC,TTT
Prolina	PRO	P	CCT,CCC,CCA,CCG
Serina	SER	S	AGT,TCA,TCC,TCT,TCG
Threonina	THR	T	ACA,ACC,ACG,ACT
Tryptophan	TRP	W	TGG
Tyrosina	TYR	Y	TAC,TAT
Valina	VAL	V	GTA,GTC,GTG,GTT
Stop-Codon	.	*	TAA,TAG,TGA

Figura 2: Código Genético utilizado pela maioria dos organismos. A primeira coluna traz o nome dos aminoácidos, a segunda o código com três letras, a terceira coluna traz o código com uma letra, e, finalmente, a quarta tem os códons correspondentes a cada aminoácido.

(como eucariotos, que possuem material genético separado do restante da célula por uma membrana nuclear) ele deve ser realizado. No pré-processamento devem ser removidos os *introns* do mRNA, trechos dos genes que não codificam proteínas, permanecendo apenas os trechos que de fato produzem proteínas, conhecidos como *exons*.

O conjunto de cromossomos (longas moléculas de DNA) de uma célula de um organismo denomina-se genoma [7], que tem uma cópia completa do material genético, produzida a cada divisão celular. A divisão celular se dá com a replicação do DNA, as fitas de uma molécula se separam e cada uma é complementada dando origem a duas moléculas de DNA, o que deve ser feito de forma precisa para evitar mutações.

3 Bioinformática

A Bioinformática é uma área da Ciência da Computação que visa analisar dados e resolver problemas de aplicações biológicas através do desenvolvimento de ferramentas. O avanço tecnológico dos computadores possibilitou à Bioinformática o desenvolvimento de softwares para processar grande volume de dados gerados pelos projetos da Biologia, como projetos de seqüenciamento completos ou de ESTs.

A análise automática de seqüências é muito importante, pois é impraticável analisar manualmente as inúmeras seqüências produzidas em projetos de seqüenciamento. Dentre as atividades desenvolvidas pela Bioinformática para análise de dados de projetos de seqüenciamento, a de montagem e principalmente a clusterização de ESTs (de seqüências curtas ou muito curtas) são o foco do trabalho proposto.

4 Projetos de seqüenciamento

Projetos de seqüenciamento que tem como objetivo a obtenção de genomas completos são conhecidos como Projetos Genomas.

Na página do Entrez Genome [18], em 11 de setembro de 2008, pode-se encontrar o genoma completo de 2000 vírus, 925 bactérias, 1630 eucariotos, além dos genomas de outros organismos.

Alternativamente ao seqüenciamento de genomas completos existem os projetos ESTs (*Expressed Sequence Tag*) [4], que buscam seqüenciar apenas os genes efetivamente expressos do organismo, evitando desta forma o seqüenciamento de *íntrons* ou de regiões entre genes, agilizando o processo de seqüenciamento. Em projetos ESTs seqüencia-se o cDNA obtido a partir do mRNA.

A utilização de ESTs em projetos de seqüenciamento envolve a produção de bibliotecas de cDNA, que são clonados por meio de vetores. O seqüenciamento dos clones é realizado com uma única leitura em uma máquina de seqüenciamento.

Nem todos os genes de um organismo são expressos com a mesma freqüência, pois dependem das características, funções ou ainda condições (idade, ambiente, etc) a que estão ligados. É muito importante saber quais genes ocorrem em um organismo, mas também é relevante saber com que

freqüência ocorrem, pois a presença excessiva (ou ausência) de certo gene pode estar relacionada a anomalias no organismo.

Em 11 de julho de 2008 estavam disponíveis 54.495.893 seqüências públicas de ESTs de 1.594 organismos diferentes no dbEST [11]. A lista dos 10 organismos com maior número de seqüências no dbEST pode ser observada na tabela 1.

Organismo	Número de ESTs
Homo sapiens (humano)	8,138,094
Mus musculus + domesticus (camundongo)	4,850,258
Arabidopsis thaliana (herbácea)	1,526,133
Bos taurus (boi)	1,517,053
Sus scrofa (porco)	1,476,546
Zea mays (milho)	1,464,859
Danio rerio (peixe paulistinha)	1,379,829
Xenopus tropicalis (rã)	1,271,375
Oryza sativa (arroz)	1,220,876
Ciona intestinalis (cordado invertebrado)	1,204,893

Tabela 1: Lista dos 10 organismos com maior número de seqüências no dbEST, em 12 de setembro de 2008

Diversos projetos de seqüenciamento, seja de genomas completos ou de ESTs, foram ou estão sendo desenvolvidos. No Brasil podem ser destacados os projetos abaixo, apoiados pela FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) [14], que em 1997 organizou a rede ONSA (*Organization for Nucleotide Sequencing and Analysis*).

- SUCEST: Genoma da Cana-de-açúcar. O projeto teve como objetivo a análise transcriptômica de variedades elite de cana-de-açúcar através da tecnologia de microarrays de cDNAs. A etapa de seqüenciamento já foi concluída, e uniu esforços de diversos pesquisadores de importantes universidades paulistas (USP, UNICAMP e UNESP) [50, 47].
- FORESTs: Projeto de seqüenciamento de eucalipto. O objetivo do Projeto era identificar 15.000 genes através do seqüenciamento de aproximadamente 100.000 ESTs preparadas a partir de bibliotecas de diferentes tecidos, incluindo plântulas, folhas, raízes, caule e madeira. O seqüenciamento teve início em outubro de 2001 e término em fevereiro de 2002 [16].
- Projeto Genomas Agronômicos e Ambientais: seqüenciamento do completo ou de ESTs, de genomas como o do café e de outros organismos com relevância para agronomia durante dois anos [51].
- Projeto Genoma do Câncer Humano: desenvolvido por diversos países incluindo o Brasil, cujo objetivo é a descoberta de genes relacionados a diversos tipos de câncer [49].

O Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [10] aliado ao Ministério da Ciência e Tecnologia [29] também financia diversos projetos em todo o País. No ano

de 2000 criaram a Rede Genoma Nacional [6], que promoveu o seqüenciamento de bactérias como *Chromobacterium violaceum* e *Mycoplasma synoviae*.

4.1 Projeto Genoma Café

O Projeto Genoma Café [33] teve início em fevereiro de 2002, criado em parceria entre o Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café, a Embrapa Café, a FAPESP e a Embrapa Recursos Genéticos e Biotecnologia (Cenargen).

O seqüenciamento total de genomas de organismos como o cafeeiro é uma tarefa de alto custo e trabalhosa, isto fez com que vários projetos de seqüenciamento, como o do café, optassem por ESTs.

Concluído em 2004, o Projeto Genoma Café resultou na construção de um banco de dados com mais de 200 mil seqüências de DNA, o que possibilitou a identificação de mais de 35 mil genes, responsáveis pelos diversos mecanismos fisiológicos de crescimento e desenvolvimento do cafeeiro.

5 Estratégias de Seqüenciamento

Seqüenciar um gene significa determinar a cadeia de nucleotídeos que o compõe. Um genoma pode possuir bilhões de bases, o que faz com que seja complexo seqüenciá-lo inteiramente, pois as tecnologias de seqüenciamento atuais apresentam limitações em relação ao tamanho reduzido (menos de 1000 bases) das seqüências que produzem. Diante disso, realiza-se a fragmentação do genoma em pequenas partes, seqüência-se cada uma dessas partes, e por fim, elas passam pelo processo de montagem.

A fragmentação é a primeira etapa do seqüenciamento e pode ser feita pelo método de digestão, que utiliza de enzimas de restrição para quebrar o DNA, mas é limitado por depender de sítios de restrição apropriados do DNA-alvo e também da possível necessidade de várias enzimas [40]. Uma boa estratégia para o seqüenciamento é a *shotgun*, muito indicada para seqüenciamento em larga escala.

A estratégia *shotgun* [37] submete o DNA a altas taxas de vibração ou nebulização, que promovem a quebra da cadeia de nucleotídeos em vários fragmentos que em geral são únicos. Normalmente, produzem-se fragmentos aleatórios que são clonados para amplificar a quantidade de fitas de DNA disponíveis.

Através de clonagem é possível criar cópias idênticas de moléculas de DNA. Este processo é comumente é realizado utilizando-se de um vetor (bactérias, vírus), por possuírem curto ciclo de vida. O fragmento de DNA a ser clonado é inserido no DNA do vetor e quando este se multiplica, o fragmento de DNA também se multiplica.

Após a divisão do genoma em partes menores inicia-se então o seqüenciamento das bases de cada uma delas, sendo o método de terminação de cadeia, ou de Sanger [38], o mais usado desde a década de 1970.

No método de terminação de cadeia são utilizadas cadeias de nucleotídeos que se diferem em uma base, a elas são acopladas bases de um oligonucleotideo, chamado de primer, para iniciar o processo de

síntese (catalisada pela enzima DNA polimerase) de uma nova fita de DNA complementar. A síntese é interrompida pela adição aleatória de uma pequena quantidade de dideoxinucleotídeo equivalente (um nucleotídeo que não possui o grupo 3'-hidroxil).

Após a síntese, as cadeias são colocadas em uma placa com gel de poliacrilamida, dividida em quatro faixas (cada uma representa um nucleotídeo do DNA), para passarem por eletrosforese, o que faz com que as menores se afastem com mais facilidade do ponto de partida que as maiores. Ao final do processo a faixa em que está a seqüência (visualizado por marcas radioativas) indica o nucleotídeo correspondente, por exemplo, se a primeira está na faixa de T, então o primeiro nucleotídeo é o T. A organização das bases de um segmento de DNA pode ser observado na figura 3a.

Na figura 3b mostra-se a análise em uma mesma caneleta, o que é possível através de marcadores que emitem luz fluorescente para cada dideoxinucleotídeo. Ainda é possível observar as bases da seqüência graficamente, através da leitura de cada base como uma curva, na qual os picos indicam as posições em que a base ocorre, o que pode ser visto na figura 3c.

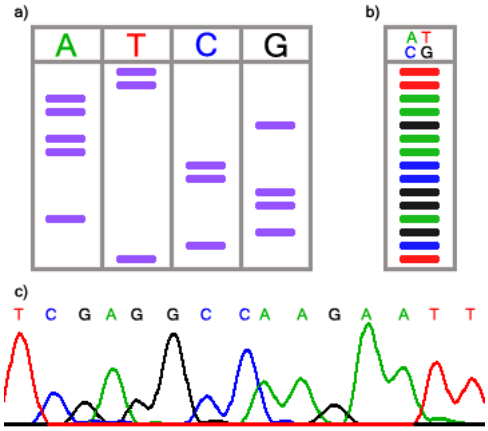

Figura 3: a) Gel de eletroforese feito com um segmento de DNA que contém a seqüência TCGAGGCCAAGAATT. b) Experimento de eletroforese que utiliza marcadores fluorescentes, realizado com o mesmo segmento de DNA. c) Representação gráfica da leitura dos sinais emitidos pelos marcadores fluorescentes, captados pela máquina de seqüenciamento.

Atualmente pesquisadores têm se esforçado para desenvolver um método de seqüenciamento mais simples e robusto que o de Sanger [38]. Um método interessante e que está sendo muito utilizado é o pirosseqüenciamento [35, 36, 34, 45].

O pirosseqüenciamento é uma estratégia de seqüenciamento por síntese de DNA em tempo real, sem a necessidade de eletrosforese. O método baseia-se na detecção do pirofosfato, PPi, liberado na síntese de DNA de forma proporcional a incorporação de dioxinucletídeo. O PPi é convertido em ATP pela sulfurilase e ATP é usado para produção de luz pela enzima luciferase. Um luminômetro ou uma câmera CCD (*charge-coupled device*), sensíveis a luz, são usados para medí-la.

O fragmento de DNA a ser seqüenciado é incubado com DNA polimerase, ATP sulfurilase, luciferase e apirase (enzima de degradação, usada na estratégia líquida). Uma alternativa é a estratégia sólida, sem a enzima de degradação, em seu lugar realiza-se uma etapa de limpeza para remover o

excesso de substrato após a adição de cada nucleotídeo.

O piroseqüenciamento é realizado de forma cíclica, inicialmente adiciona-se dATP, em seguida realiza-se o processo de limpeza, posteriormente adiciona-se dGTP, dCTP e dTTP. A cada nucleotídeo adicionado mede-se através da reação acoplada (descrita anteriormente) o PPi liberado, o processo segue até que se obtenha as informações necessárias sobre a seqüência. As atividades enzimáticas podem ser observadas através de um gráfico, as curvas de ascensão são determinadas por polimerase e ATP sulfurilase, a altura do sinal por luciferease e a descendência pela remoção de nucleotídeos (é necessário ter uma concentração de nucleotídeos suficiente para que a reação de polimerase seja relativamente baixa).

Outra interessante tecnologia de seqüenciamento desenvolvida nos últimos anos é a Illumina [43]. Nela fragmentos de DNA são colocados randomicamente em uma superfície plana e transparente, estes são estendidos para formação de mais de 50 milhões de *clusters*, cada um com 1000 cópias do mesmo modelo. Cada modelo é seqüenciado usando a tecnologia de seqüenciamento por síntese que emprega a detecção de corantes florescentes. Detecção que é ativada usando excitação por laser e reflexão óptica interna. *Reads* curtos produzidos são alinhados em relação a um genoma de referência e as diferenças são observadas usando um *software pipeline* para análise de dados. Após a produção do primeiro *read*, o modelo pode ser regenerado, permitindo a formação de um segundo *read* (com mais de 36 bp) na extremidade oposta dos fragmentos. Illumina possui ainda sistemas para aplicações em expressão gênica, descoberta de pequenos RNAs e interações entre proteínas e ácidos nucleicos.

O sistema SOLID [44] possui uma unidade de seqüenciamento que analisa o genoma do início ao fim, ainda lida com a formação de *clusters* e armazenamento de dados. O seqüenciamento é baseado na ligação e detecção de oligonucleotídeo, a tecnologia empregada permite a geração de dados de alta qualidade para diversas aplicações como: seqüenciamento de genomas inteiros, descoberta de pequenos RNAs, avaliação de expressão gênica e resseqüenciamento microbial e de eucariotos. Os *reads* produzidos são de aproximadamente 35 bases, com taxa de erro zero. A ausência de erros ocorre devido ao sistema de detecção de erros utilizado, que é capaz de distinguir erros randômicos de erros sistemáticos, ligados a mudanças reais nas bases ou SNPs (polimorfismo de base única).

Como pôde ser notado, o processo de seqüenciamento é trabalhoso e a utilização de computadores para realização do processo é justificável quando o processo trabalha com dados em larga escala, mas não exclusivamente. Desta forma, mesmo em se tratando de problemas que poderiam ser resolvidos manualmente é mais conveniente que sejam resolvidos por computadores, um exemplo foi o trabalho realizado por Smith, Waterman e Fitch [42] e citado por Meidanis e Setubal [40]. Os autores observaram o alinhamento de 12 bases iguais através de comparação de seqüências de dois bacteriófagos realizada por computador. Anteriormente estes mesmos bacteriófagos haviam sido comparados de forma manual, quando foram registradas 11 bases iguais, certamente os pesquisadores neste primeiro experimento não perceberam todos os alinhamentos.

Uma vez realizado o seqüenciamento dos *reads* é extremamente importante escolher uma boa estratégia de montagem para que se atinja uma alta cobertura da seqüência original.

6 Estratégias de Montagem

Montagem é o processo de reconstrução da seqüência original e ocorre em projetos de seqüenciamento de genomas completos. O princípio de montagem inicialmente adotado foi o de alinhamento de seqüências por sobreposição.

Técnicas de sobreposição identificam a semelhança, correspondência de bases, entre um *read* e outro. Deve ser observado se ambos *reads* provêm da mesma fita de DNA ou de fitas diferentes, pois neste último caso a semelhança será observada entre uma das seqüências e o complemento reverso da outra.

Um dos métodos mais antigos de montagem foi proposto entre as décadas de 80 e 90 por S. Dean e R. Staden [12]. A idéia básica do algoritmo é o processamento dos fragmentos um a um, unindo em *contigs* os grupos de fragmentos relacionados. A cada novo fragmento observa-se se ele se une a um *contig* existente, forma um novo *contig* ou ainda promove a união de dois *contigs*, a junção de vários *contigs* formará a seqüência completa.

Posteriormente J. Kececioglu apresentou um outro sistema [24], que utiliza o grafo de sobreposições. Neste grafo, os nós são os fragmentos seqüenciados e a uma aresta entre dois nós indica que há sobreposição entre eles e que formarão *contigs*.

No entanto, a construção de *contigs* não é perfeita na maioria das vezes, isto porque muitos erros podem estar presentes nos *reads*. Estes podem ser classificados de diferentes formas e serem causados, por exemplo, por mutações, contaminação por vetor no processo de clonagem, ou por problemas na identificação das fitas de DNA. Além disso, outro problema é o fato de se desprezar o uso de *reads* curtos (em torno de 100 bases) ou muito curtos (com aproximadamente 25 bases), também chamados de *microreads*, na fase de seqüenciamento.

A utilização de *reads* muito curtos é justificada por possuírem informações genéticas importantes ligadas a transformações nos cromossomos ou a diferenciação entre um genoma desconhecido e outro já seqüenciado [9]. Infelizmente é mais complexo trabalhar com a montagem destes tipos de *reads* (com tamanho, por exemplo, de 25 a 50 bases), pois há mais possibilidade de ocorrer sobreposição uma vez que estas regiões se repetem, o que diminui a probabilidade de unir corretamente estas sobreposições. She e colegas [41] mostraram que 5 a 6 % do genoma humano é composto por repetições.

A complexidade das seqüências aumenta por um fator de quatro a cada base adicionada [52], então é intuitivo pensar que é melhor trabalhar sempre com *reads* muito curtos, mas isto esbarra no aumento da probabilidade de se encontrar *reads* redundantes quando possuem tamanho reduzido. É óbvio que há maior possibilidade de montagem quando as sobreposições potenciais são únicas e que a porcentagem de unicidade dos *reads* aumenta quando estes aumentam de tamanho [53].

Assim, tem-se um impasse em relação à utilização de um *read* de um determinado tamanho, o que varia de acordo com o genoma a ser seqüenciado. Por exemplo, em experimentos realizados por N. Whiteford e colegas [53], para se obter *contigs* de 10.000 nt foram necessários *reads* de 18 nt a 20 nt em genomas virais para uma cobertura de 99,6%, enquanto que no genoma humano precisou-se de *reads* de 500 nt, para uma cobertura equivalente.

O sucesso das tecnologias de seqüenciamento não depende apenas da rapidez do processo de produção dos *reads*, mas também da solução computacional do problema de montagem de fragmentos em genomas completos. A dificuldade em analisar as repetições faz com que a montagem de genomas longos utilizando *reads* curtos seja um problema computacionalmente difícil [31].

Às vezes é interessante realizar pré e pós-processamento das amostras a serem montadas. O pré-processamento pode ser feito pela comparação dos fragmentos, selecionando apenas os que têm grande probabilidade de se sobreporem para o processo de montagem. O pós-processamento pode ser feito através da comparação dos *contigs* produzidos para gerar uma seqüência consenso (formada por caracteres que aparecem com mais freqüência em um alinhamento) [28].

Tecnologias que usam pirosseqüenciamento, como a 454 [39], que teve sucesso no resseqüenciamento do adenovírus e é referência de seqüenciamento ultra-rápido de DNA tem seu potencial prejudicado devido as repetições, limitando, atualmente, o tamanho dos *reads* em 100 bases, o que pode não ser bom o bastante para grandes genomas [25]. A 454 não está pronta para substituir tecnologias de seqüenciamento tradicional, sendo portanto melhor aproveitada de forma complementar a tais métodos, proporcionando uma abordagem híbrida (Sanger aliado ao pirosseqüenciamento) [19].

Métodos de montagem tradicionais não exploram corretamente a presença de *reads* muito curtos, pois este aparecem em grande número e repetidamente, levando a ambigüidades [54]. Desta forma, é relevante o desenvolvimento e testes de algoritmos que aproveitem melhor as informações contidas nos *reads* curtos ou muito curtos, ofereçam melhores soluções para o problema de repetição e ainda que utilizem a clusterização de ESTs como alternativa para identificação dos genes presentes em um organismo.

6.1 Algoritmos para montagem de seqüências

Existem diversos algoritmos que proporcionam uma solução aos problemas propostos pelos projetos de seqüenciamento, incluindo a montagem. Um deles é o Phrap [20]. O Phrap é um software para montagem de DNA que utiliza a idéia do *shotgun*, não específico para determinado tamanho de *read*. Produz informações de alta qualidade a partir dos *reads* mesmo que existam repetições.

Outro software importante é o CAP3 [22], que utiliza sobreposição de bases para montagem de seqüências e algoritmos para correção de erros e regiões de baixa qualidade. Esta é a terceira geração de softwares CAP [21], a exemplo do Phrap não é específico para determinados tamanhos de *reads*. Ele é gratuito para uso acadêmico e possui 3 fases principais. A primeira consiste na identificação e remoção de pontas de baixa qualidade, cálculo das sobreposições e eliminação das falsas sobreposições. Na segunda os *contigs* são criados através da união de seqüências em ordem decrescente de pontuação das sobreposições. Por fim, na terceira faz-se a construção do alinhamento das seqüências e determinação da qualidade das mesmas.

O software CAP3 foi comparado com o Phrap [20] com a utilização de quatro conjuntos de BACs e formou menor quantidade de *contigs* e com menor índice de erros. Telles e da Silva [48], realizaram clusterização da cana-de-açúcar e também compararam os dois softwares (CAP3 e Phrap) encontrando resultados equivalentes aos apresentados anteriormente.

Atualmente diversos algoritmos estão sendo propostos para realização da montagem utilizando *reads* muito curtos. Como o ALLPATHS [9], que utiliza de um grafo ordenado denominado grafo unipath para montar as seqüências. Os nós do grafo são as unipaths (seqüências que não podem ser mais quebradas) e as arestas são indicativos de que há sobreposição entre as k-1 bases do último k-mer (seqüência de k-nucleotídeos) de um nó x com as k-1 bases do primeiro k-mer do nó y. O algoritmo trabalha com correção de erros e com a numeração dos k-mers, de forma que k-mers iguais tenham a mesma numeração, realizando a montagem baseada nesta numeração de forma local e posteriormente global (unindo-se as montagens).

Um outro algoritmo é o SSAKE [52] (*Short Sequence Assembly by progressive K-mer search and 3' read Extension*) que realiza a montagem de milhões de seqüências curtas de DNA. Ele utiliza uma tabela de hash que indexa as seqüências de acordo com o número de vezes em que aparecem no conjunto das seqüências e uma árvore de prefixo onde as seqüências são ordenadas de forma decrescente em relação ao número de ocorrências, diminuindo a extensão de *reads* que contém erros de seqüenciamento. Gera-se possíveis 3' k-mers, a partir de *reads* não montados, e utiliza-os na busca de um 5' k-mer que seja seu complemento perfeito de um *read*. Quando encontra-se o 5' k-mer, o read não montado é estendido pela extremidade 3' e então remove-se o *read* com o 5' k-mer da tabela de hash e da árvore de prefixo. O processo se repete até que não se consiga mais extensões de 3', quando finalmente estende-se o *contig* montado com um *read* 5'.

O algoritmo VCAKE (*Verified Consensus Assembly by k-mer Extension*) [23] aprimora a extensão de k-mers realizada pelo SSAKE [52]. Diferentemente do SSAKE, o VCAKE encontra todos os 3' k-mers complementares, usando as onze primeiras bases do k-mer na busca e retornando as chaves comparadas com o restante do k-mer. A montagem dos *contigs* é feita através de um vetor de seqüência, adicionando-se ao *contig* a base que excede um limiar de representação. Em seguida remove-se da árvore e da tabela as seqüências que aparecem no *contig*, o algoritmo é então aplicado ao complemento reverso do *contig*. Em testes realizados com bactérias mostrou-se melhor que SSAKE.

Um quarto algoritmo interessante é o VELVET [54], na verdade é um conjunto de algoritmos para realizar a montagem de *reads* curtos utilizando-se de grafos de Bruijn, útil na eliminação de erros e repetições. O grafo é constituído de blocos, cada bloco é composto por um vértice N e seu complemento reverso Ñ. Nós adjacentes possuem k-mers com sobreposição de k-1 nucleotídeos. O grafo é construído utilizando uma tabela de hash dos *reads* indexadas por um k-mer. Os *contigs* são produzidos pelo mapeamento dos *reads* e são formados seguindo as transições do grafo, unindo as que forem ambíguas. O VELVET possui mecanismos para remoção de erros ou repetições que prejudicam o percurso no grafo. Apesar de utilizar mais memória produziu *contigs* maiores que SSAKE e VCAKE.

O algoritmo SHARCGS [13] (*Short-read Assembler based on Robust Contig Extension for Genome Sequencing*) foi desenvolvido para realizar a montagem de milhões de *reads* muito curtos. Executa inicialmente a filtragem dos *reads* com erros de seqüenciamento, em seguida montagem de *contigs*, que é estendida pela extremidade 3', e por fim a fusão de *contigs* de várias montagens parciais e a geração de medidas de qualidade. Com a utilização de seqüências de 25-40 mer os autores do

algoritmo observaram melhor montagem quando ambigüidade era menor.

Uma idéia diferente é abordada pelo software EULER [32]. O algoritmo utiliza o grafo de Bruijn e o percorre através de um caminho Euleriano, no qual visita-se cada aresta do grafo apenas uma vez. Para atingir a precisão desejada, EULER restaura informação sobre *reads* seqüenciados que foram perdidas na construção do grafo de Bruijn, encontrando um super caminho Euleriano que contenha todos os subcaminhos encontrados anteriormente no grafo e analisando as ambigüidades. Apesar de produzir bons resultados, EULER [32] é limitado pela necessidade da correção de erros dos *reads* antes que seja aplicado.

Diante da tentativa de superar a limitação apresentada por EULER [32], foi criado um algoritmo de EULER modificado [31], que utilizou de programação dinâmica para correção de erros e montagem de seqüências. No entanto, os autores do algoritmo observaram que os erros complicaram o grafo de Bruijn, o que levou a um maior esforço para montagem de *contigs*.

O protocolo de seqüenciamento e de montagem para seqüenciar genomas de mamíferos, SHARP (*Short Read Assembly Protocol*) [45], visa a montagem fiel de genomas a partir de *reads* curtos gerados pelo Pirosseqüenciamento [35]. SHARP é uma variação do seqüenciamento hierárquico (que lida com a clusterização de *reads* em pequenos conjuntos locais que representam uma seqüência), possui alto potencial de automação e paralelização do processo de montagem.

No processo de montagem SHARP realiza a detecção de sobreposições e correção de erros. Além da criação de conjuntos de *reads*, criação de grandes conjuntos de contigs, união dos contigs (montagem final).

Apesar do grande número de algoritmos desenvolvidos para montagem de *reads* curtos ou muito curtos, estes se preocupam apenas com o genoma completo. Sendo portanto, a clusterização de ESTs uma estratégia interessante para ser abordada.

6.2 Clusterização de ESTs

Genes expressos representam um grande número de seqüências biológicas redundantes e parciais, podendo conter regiões de baixa qualidade, quiméricas ou ainda contaminadas por vetores e adaptadores.

Os ESTs podem apresentar algumas particularidades como SNP (polimorfismo de base única), domínios conservados (trechos curtos idênticos presentes em diferentes proteínas) e famílias multigênicas (grupo de genes relacionados, originados a partir de um gene ancestral comum a todos).

Programas para montagem de seqüências não são bem adaptados para ESTs, podendo levar a montagens incorretas. Para superar esta dificuldade utiliza-se o processo de clusterização (agrupamento de seqüências semelhantes em um grupo chamado cluster) dos ESTs e a obtenção da seqüência consenso, formada pela análise das seqüências que formam o cluster e que é aceita como a seqüência com maior probabilidade de ser a do gene existente no organismo.

Diversos programas têm sido desenvolvidos para clusterizar ESTs. Como o TGICL [30], o d2_cluster [8], o TIGR Assembler [46], o *xtract* [27] e o UIcluster [26].

No Instituto de Computação da UNICAMP foram desenvolvidos trabalhos que envolveram a clusterização de ESTs [5] e análise de SNPs [17].

7 Objetivos

Este trabalho tem como objetivo estudar técnicas de montagem e clusterização, buscando identificar as deficiências de algoritmos de montagem considerando *reads* curtos e muito curtos. A partir destas análises desenvolveremos metodologias para montagem de seqüências e principalmente a clusterização de ESTs utilizando *reads* curtos e muito curtos.

8 Materiais e Métodos

Os resultados do Projeto Genoma Café [33] colocaram o Brasil na posição de líder mundial das pesquisas do genoma cafeeiro (30% das seqüências de sua base de dados nunca foram descritas em base de dados mundiais).

O Projeto FINEP (0886/2007), propõe a geração de uma base de dados (será agregada à já existente do Genoma Café) que constituirá uma importante fonte de conhecimento para o entendimento das bases do café. A formação desta base de dados só será possível através do seqüenciamento do genoma do café, sendo imprescindível o desenvolvimento de boas estratégias de montagem e clusterização.

O organismo alvo do estudo será o do café, pois além dos 200 mil ESTs gerados pelo projeto Genoma Café [33] (Embrapa/FAPESP), há mais de 100 mil ESTs depositados no dbEST [11] e ainda serão gerados mais de um milhão de seqüências genômicas curtas, graças a um projeto recentemente aprovado pela FINEP (0886/2007). Neste projeto FINEP serão usadas plantas *Coffea arabica L. cv.* cultivadas em diferentes estações experimentais (como IAPAR em Londrina-PR, Embrapa Cerrados em Planaltina-DF e EPAMIG em Lavras-MG), coletadas em diferentes fases do desenvolvimento e seqüenciadas pelas tecnologias 454 [39] ou Illumina [43], que produzem *reads* curtos e muito curtos.

Apesar do grande número de softwares desenvolvidos para montagem de seqüências nenhum deles é perfeito, pois montam muitos *contigs* mas nem sempre a seqüência completa. No entanto, estes servem como base e podem ser utilizados em um estudo comparativo para o desenvolvimento de novas ferramentas. Neste sentido que serão utilizados alguns dos softwares apresentados na seção 6.1.

Devido a natureza deste projeto, é desejável a utilização de linguagens de programação que ofereçam suporte para resolução de problemas de bioinformática. Uma delas é a Python [3], linguagem interpretada que possui uma série de bibliotecas úteis à bioinformática, além disso é gratuita e compatível com a plataforma Linux [2], por isso foi a linguagem escolhida para o desenvolvimento do projeto.

Serão realizadas análises qualitativa, com validação por biólogos, e quantitativa, utilizando métodos estatísticos (consistência interna, consistência externa, discrepância de bases etc) para avaliar a metodologia a ser desenvolvida.

9 Plano de Trabalho e Cronograma

A Tabela 2 descreve a distribuição das atividades a serem realizadas durante a execução deste trabalho.

	2008				2009												2010	
	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	X	X	X	X														
2	X																	
3		X																
4	X	X	X	X														
5					X	X	X											
6							X	X	X									
7									X	X	X							
8											X	X	X					
9													X	X	X			
10											X	X	X	X	X	X		
11																	X	
12																		X

Tabela 2: Cronograma de atividades.

1. Disciplinas obrigatórias do programa de mestrado.
2. Escrita do projeto de mestrado.
3. Exame de qualificação.
4. Estudo da linguagem de programação Python e da biblioteca BioPython [1], específica para bioinformática.
5. Testes de algoritmos existentes.
6. Identificação de problemas.
7. Construção da metodologia.
8. Implementação.
9. Testes.
10. Escrita da dissertação.
11. Revisão final do texto da dissertação.
12. Defesa da dissertação.

Referências

[1] Biopython Homepage, August 2008. http://www.biopython.org.

[2] Linux homepage, July 2008. http://www.linux.org.

[3] Python Homepage, July 2008. http://www.python.org.

[4] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. R. McCombie, and J. C. Venter. Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. *Science*, 252:1651–1656, June 1991.

[5] C. Baudet. Uma abordagem para detecção e remoção de artefatos em seqüências ests. Master's thesis, Universidade Estadual de Campinas, Brasil, Dezembro 2006.

[6] Brazilian Genome - The Virtual Institute of Genomic Research, September 2008. http://www.brgene.lncc.br.

[7] T. A. Brown. *Genomes*. John Wiley and Sons, Inc, 1999.

[8] J. Burke, D. Davison, and W. Hide. d2_cluster: A Validated Method for Clustering EST and Full-Length cDNA Sequences. *Genome Research*, 9:1135–1142, 1999.

[9] J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, Eric S. Lander, C. Nusbaum, and D. B. Jaffe. Allpaths: De novo assembly of whole-genome shotgun microreads. *Genome Research*, 20:810–820, 2008.

[10] Conselho Nacional de Desenvolvimento Científico e Tecnológico, July 2008. http://www.cnpq.br.

[11] dbEST – The International Expressed Sequence Tags Database, July 2008. http://www.ncbi.nlm.nih.gov/dbEST.

[12] S. Dean and R. Staden. A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Research*, 19:3907–3911, 1991.

[13] 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*, 17:1697–1706, 2007.

[14] Fundação de Amparo à Pesquisa do Estado de São Paulo, July 2008. http://www.fapesp.br.

[15] National Center for Biotecnology Information. http://www.ncbi.nlm.nih.gov, July 2008.

[16] FORESTs: Eucalyptus Genome Sequencing Consortium, July 2008. https://forests.esalq.usp.br/.

[17] M. Galves. Uma abordagem computacional para a determinação de polimorfismo de base única. Master's thesis, Universidade Estadual de Campinas, Brasil, Dezembro 2006.

[18] Genome, September 2008. http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome.

[19] S. M. D. Goldberg, J. Johson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S. A. Kravitz, F. M. Lauro, K. Li, Y. Rogers, R. Strausberg, G. Sutton nd L. Tallon,

T. Thomas, E. Venter, M. Frazier, and J. C. Venter. A sanger/pyrosequncing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 103:11240–11245, 2006.

[20] P. Green. Phrap Homepage: phred, phrap, consed, swat, cross_match and RepeatMasker Documentation, July 2008. http://www.phrap.org.

[21] X. Huang. A contig assembly program based on sensitive detection of fragments overlap. *Genomics*, 14:18–25, 1992.

[22] X. Huang and A. Madan. CAP3: a DNA sequence assembly program. *Genome Research*, 9:868–877, 1999.

[23] W. R. Jeck, J. A. Reinhardt, D. A. Baltrus, M. T. Hickebotham, V. Magrini, E. R. Mardis, J. L. Dangl, and C. D. Jones. Extending assembly of short dna sequences to handle error. *Bioinformatics*, 23(21):2942–2944, 2007.

[24] J. D. Kececioglu and E. W. Myers. Combinatorial algorithms for dna sequence assembly. Technical Report TR92-37, The University of Arizona, October 1992.

[25] J. Kling. Ultrafast dna sequencing. *Nature Biotechnology*, 21:1425–1427, 2003.

[26] K. Malde, E. Coward, and I. Jonassen. Fast sequence clustering using a suffix array algorithm. *Bioinformatics*, 19:1221–1226, 2003.

[27] K. Malde, E. Coward, and I. Jonassen. A graph based algorithm for generating est consensus sequences. *Bioinformatics*, 21:1371–1375, 2005.

[28] J. Meidanis. A simple toolkit for dna fragment assembly. *American Mathematical Society*, 47:271–288, 1999.

[29] Ministério da Ciência e Tecnologia, September 2008. http://www.mct.gov.br.

[30] G. Pertea, X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, and J. Quackenbush. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, 19(5):651–652, 2003.

[31] P. A. Pevzner, H. Tang, and M. Chaisson. Fragment assembly with short reads. *Bioinformatics*, 20:2067–2074, 2004.

[32] P. A. Pevzner, H. Tang, and M. S. Waterman. An eulerian path approch to dna fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 14:9748–9753, 2001.

[33] Projeto Genoma Café, August 2008. http://www.cenargen.embrapa.br/biotec/genomacafe/.

[34] M. Ronaghi. Pyrosequencing sheds light on dna sequencing. *Genome Research*, 11:3–11, 2001.

[35] M. Ronaghi, S. Kramohamed, B. Pettersson, M. Uhlen, and P. Nyren. Real-time dna sequencing using detection of pyrophosphate release. *Analitical Biochemistry*, 242:84–89, 1996.

[36] M. Ronaghi, M. Uhlen, and P. Nyren. Dna sequencing: A sequencing method based on real-time pyrophosphate. *Science*, 281:363–365, 1998.

[37] F. Sanger and et al. Nucleotide sequence of bacteriophage lambda DNA. *Journal of Molecular Biology*, 162:729–773, 1982.

[38] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain termination inhibitors. *Proceedings of the National Academy Science, USA*, 74:5463–5467, 1977.

[39] 454 Life Sciencs. 454 Homepage, July 2008. http://www.454.com.

[40] J. C. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology.* PWS Publishing Company, 1997.

[41] X. She, Z.Jiang, R. A. Clark, G. Liu, Z. Cheng, E. Tuzun, D. M. Church, G.Sutton, A. L. Halpern, and E. E. Eichler. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*, 431:927–930, 2004.

[42] T. Smith, M. Waterman, and W. Fitch. Comparative biosequence metrics. *J. Mol. Evol.*, 18:38–46, 1981.

[43] Illumina Solexa. Technology: illumina sequencing technology, July 2008. http://www.illumina.com/pages.ilmn?ID=250.

[44] SOLID. The SOLID generation delivers, September 2008. http://solid.appliedbiosystems.com/.

[45] A. Sundquist, M. Ronaghi, H. Tang, P. Pevzner, and S. Batzoglou. Whole-genome sequencing and assembly with high-throughput, short-read technologies. *Publishing Science*, 5:e484, 2007.

[46] G. G. Sutton, O. White, M. D. Adams, and A. R. Kerlavage. TIGR assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science Technology*, 1:9–19, 1995.

[47] G. P. Telles, M. D.V. Braga, Z. Dias, L. T. Li, J. A. A. Quitzau, F. R. da Silva, and J. Meidanis. Bioinformatics of the Sugarcane EST Project. *Genetics and Molecular Biology*, 24(1-4):9–15, December 2001.

[48] G. P. Telles and F. R. da Silva. Trimming and clustering sugarcane ESTs. *Genetics and Molecular Biology*, 24(1-4):17–23, December 2001.

[49] The Human Cancer Genome Project, July 2008. http://www.ludwig.org.br/ORESTES.

[50] A. L. Vettore, F. R. da Silva, E. L. Kemper, and P. Arruda. The libraries that made SUCEST. *Genetics and Molecular Biology*, 406:151–157, 2001.

[51] L. G. E. Vieira et al. Brazilian coffee genome project: an est-based genomic resource. *Brazilian Journal Plant Physiology*, 15, 2006.

[52] R. L. Warren, G. G. Sutton, S. J. M. Jones, and R. A. Holt. Assembling millions of short dna sequences using ssake. *Bioinformatics*, 23(4):500–501, 2007.

[53] N. Whiteford, N. Haslam, G. Weber, A. Prugel-Bennett, J. W. Essex, P. L. Roach, M. Bradley, and C. Neylon. An analysis of the feasibility of short read sequence. *Nucleic Acids Research*, 33:171–177, 2005.

[54] D. R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 20:821–829, 2008.