Uma abordagem para trimagem, verificação de contaminação e clusterização de seqüências EST

Christian Baudet

Resumo

Os projetos de seqüenciamento EST podem produzir muitas informações sobre o índice gênico de um organismo, que é a a listagem dos genes existentes em seu genoma. Dentro deste contexto, a trimagem, a verificação de contaminação e a clusterização são procedimentos de grande importância à análise das informações produzidas. No entanto, não existe um conjunto de procedimentos padrão que possa ser utilizado por qualquer projeto. O trabalho aqui proposto tem o objetivo de identificar melhorias a serem aplicadas nestes processos, de modo a obter um protocolo que possa ser adotado de maneira confiável por qualquer projeto EST.

1 Introdução

Este documento tem o objetivo de apresentar o plano de trabalho que executaremos para a escrita da dissertação de mestrado. A seção 2 apresentará os conceitos básicos necessários. As seções 3, 4 e 5 fornecerão detalhes sobre os objetos de estudo deste trabalho. Finalmente, as seções 6 e 7 tratarão da proposta e do cronograma das atividades a serem executadas.

2 Conceitos Básicos

Nesta seção faremos uma breve apresentação do contexto básico em que este trabalho está inserido, através da descrição de uma série de tópicos e termos comumente encontrados na área.

2.1 Genética

Todo organismo vivo apresenta características observáveis tais como cor, estrutura, comportamento, etc. Estas características formam o seu fenótipo, que é determinado pela interação entre o genótipo do organismo e o meio em que ele vive. O genótipo é o conjunto de informações contidas no material genético de um organismo. Estas informações ditam como o organismo será construído e mantido. Elas são replicadas a cada divisão celular e podem ser herdadas no momento da reprodução.

A genética é a área da Biologia que se dedica ao estudo de genes. Os genes são as principais unidades de informação biológica contidas no material genético de um organismo. Um gene armazena as instruções para sintetização de móleculas que participam de reações metabólicas que ocorrem na célula. Na maioria dos organismos, os genes, assim como todo o material genético dos organismos, são compostos de DNA. Existem alguns vírus, denominados retrovírus que possuem material genético composto de RNA.

2.2 DNA & RNA

O DNA ou ácido desoxirribonucléico é um dos dois tipos de ácidos nucléicos encontrados dentro da célula de um organismo. Ele é um polinucleotídeo composto por 4 tipos de nucleotídeos diferentes. Cada nucleotídeo é composto por três partes: uma pentose denominada desoxirribose, um grupo fosfato e uma base nitrogenada que é diferente em cada um dos 4 tipos de nucleotídeos. As bases nitrogenadas podem ser pirimídicas (citosina e timina) e púricas (adenina e guanina). A Figura 1a exibe a estrutura molecular de um nucleotídeo que forma o DNA e as estruturas das possíveis bases.

O outro tipo de acído nucléico é o RNA ou ácido ribonucléico. A sua composição química é similar a do DNA. A diferença está no açúcar pentose que compõem o nucleotídeo, que é uma ribose no lugar da desoxirribose, e na existência do nucleotídeo que possui uma base nitrogenada chamada uracila, substituindo o nucleotídeo que possui a timina. A Figura 1b mostra o açúcar ribose e a base nitrogenada uracila.

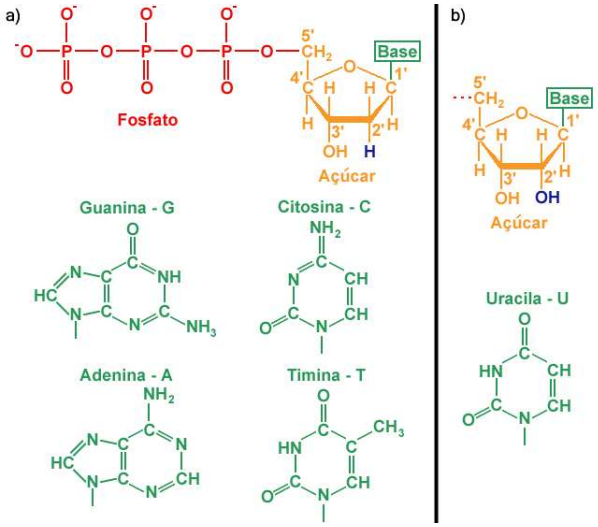

Figura 1: a) Nucleotídeos que formam o DNA. b) Açúcar ribose que substitui a desoxirribose e a base nitrogenada Uracila que substitui a timina na formação do RNA.

O DNA possui uma estrutura de dupla-hélice, que foi descoberta em 1953 por Watson e Crick [46]. Ela é composta por duas fitas de polinucleotídeos que correm em direções opostas. Esta hélice é estabilizada por dois tipos principais de interações químicas. O primeiro tipo é o pareamento de bases, entre as duas fitas, que envolve a formação de pontes de hidrogênio entre uma adenina de uma fita e a timina da outra fita, ou entre a citosina de uma fita e a guanina da outra. O segundo tipo é a interação hidrofóbica que existe entre cada par de bases adjacentes e que adiciona estabilidade à dupla hélice.

A limitação imposta pelos pares de bases possíveis, faz com que a replicação do DNA possa gerar cópias perfeitas de uma molécula pai a partir de fitas pré-existentes, que ditarão as seqüências das novas fitas como se fossem moldes. Esta síntese de DNA dependente de um modelo é utilizada por todas as enzimas celulares de polimerização de DNA.

A possibilidade de se produzir cópias perfeitas a partir de uma molécula de DNA faz com que ela seja perfeita para carregar as informações genéticas de um organismo. A conformação de dupla hélice força a manutenção da seqüência de nucleotídeos existente no genoma do organismo. Se o DNA fosse formado apenas por uma fita de polinucleotídeo, inserções e remoções de nucleotídeos poderiam ocorrer com grande freqüência, e além disso, a fita poderia ser facilmente partida em pedaços.

Por causa da característica citada acima, os genomas da maior parte dos organismo existentes são feitos por DNA (existem alguns vírus, conhecidos como retrovírus, que possuem o material genético composto por RNA).

2.3 Genoma

Todo organismo possui um genoma que contém toda a informação biológica necessária para contruí-lo e mantê-lo vivo. Esta informação é codificada em seqüências de nucleotídeos em suas moléculas de DNA ou RNA e é dividida em unidades discretas chamadas genes.

A informação contida em um gene é lida por proteínas que se ligam ao genoma em posições apropriadas e iniciam uma série de reações bioquímicas conhecidas como expressão gênica. Para organismo com genoma feito de DNA, estas reações são divididas em dois estágios: transcrição e tradução.

2.3.1 Transcrição e tradução

A transcrição é a produção de uma cópia feita de RNA de um gene contido no genoma e que vai ser expresso. Ela se inicia com a ligação da enzima RNA polimerase e outros fatores de transcrição ao genoma, próximo a localização do gene formando um complexo de transcrição. A partir deste complexo a fita de RNA é produzida. Esta fita de RNA é conhecida como mRNA ou RNA mensageiro.

A tradução é a síntese de proteínas a partir das cópias transcritas de RNA. A seqüência de aminoácidos da proteína é determinada com base no código genético do organismo. O código genético é uma tabela que possui a correspondência entre cada tripla de nucleotídeo, denominada códon, e o aminoácido que será utilizado na síntese da proteína. Esta tabela é redundante, pois existem 64 códons possíveis distribuídos entre 20 aminoácidos mais o STOP códon, que é o códon que indica o término da tradução. A tabela não é universal, ou seja, não é a mesma para todos os organismos. Na página do NCBI [23] é possível encontrar 17 tabelas diferentes. A Figura 2 exibe o código genético padrão, utilizado pela maior parte dos organismos. Ela corresponde a tabela número 1 na página do NCBI.

UUU, UUC: phe UUA, UUG: leu	UCU, UCC, UCA, UCG: ser	UAU, UAC: tyr UAA, UAG: stop	UGU, UGC: cys UGA: stop UGG: trp
CUU, CUC, CUA, CUG: leu	CCU, CCC, CCA, CCG: pro	CAU, CAC: his CAA, CAG: gln	CGU, CGC, CGA, CGG: arg
AUU, AUC, AUA: ile AUG: met	ACU, ACC, ACA, ACG: thr	AAU, AAC: asn AAA, AAG: lys	AGU, AGC: ser AGA, AGG: arg
GUU, GUC, GUA, GUG: val	GCU, GCC, GCA, GCG: ala	GAU, GAC: asp GAA, GAG: glu	GGU, GGC, GGA, GGG: gly

Figura 2: Código Genético utilizado pela maioria dos organismos.

Esta descrição do processo de tradução é suficiente para a expressão gênica que ocorre em bactérias e arqueobactérias, pois elas possuem genomas mais simples. Em organismos mais complexos, antes

da tradução ocorre o pré-processamento do mRNA (RNA mensageiro) para remoção dos introns, que são os trechos do gene que não codificam proteínas. O mRNA processado conterá apenas exons, que são os trechos de DNA que serão realmente traduzidos.

Neste pré-processamento da molécula que será traduzida poderá ocorrer também o splice alternativo. O splice aternativo é a produção de um dos vários mRNAs possíveis a partir da combinação dos exons existentes em um gene. Este é um evento comum em organismos mais complexos.

A tradução começa, de fato, com a ligação da fita de RNA já processada em um ribossomo, uma organela existente no interior da célula, e termina com a produção da proteína desejada. Esta proteína passará por um processamento e após isso adquirirá a sua conformação final.

Além dos genes que produzem proteínas, existem aqueles que produzem seqüências de rRNA (RNA ribossomal) e tRNA (RNA transportador).

2.3.2 Replicação do genoma

Uma cópia completa do genoma é feita toda vez que uma célula se divide. A replicação do DNA precisa ser altamente precisa para evitar a ocorrência de mutações. Contudo, algumas podem ocorrer devido a erros na replicação ou a efeitos de agentes mutagênicos químicos ou físicos que alteram diretamente a estrutura do DNA. Enzimas de reparação de DNA corrigem a maior parte dos erros, mas alguns escapam.

Se o organismo que sofreu mutação sobreviver, esses erros que escaparam do processo de correção podem se tornar características permanentes nas linhagens que descenderem deste organismo.

As mutações em conjunto com os eventos de rearranjo de genoma resultantes da recombinação de trechos do material genético permitem a evolução molecular, que dirige a evolução dos organismos vivos.

2.3.3 As diferenças entre procariotos e eucariotos

Os biólogos dividem os organismos vivos em dois grupos: procariotos e eucariotos. Os procariotos são os organismos que não possuem um núcleo celular organizado de maneira que o seu material genético se encontra espalhado dentro do citoplasma. Os eucariotos são os organismos que possuem um núcleo onde o material genético fica armazenado, de modo que ele fique separado do citoplasma.

Os procariotos possuem um genoma organizado de forma diferente dos eucariotos. Em geral, eles possuem uma única molécula de DNA, e esta molécula é circular. Além disso, alguns procariotos podem conter pequenas moléculas circulares ou lineares de DNA chamadas plasmídeos.

O genoma dos eucariotos é divido em duas ou mais moléculas lineares de DNA, cada uma contida em um cromossomo. Adicionalmente, todos os eucariotos possuem um pequeno genoma mitocondrial, que é usualmente circular. As plantas e outros organismos fotossintéticos possuem um terceiro genoma localizado nos cloroplastos.

Uma grande variação de tamanhos pode ser encontrada entre os genomas dos eucariotos. O fungo *Saccharomyces cerevisiae* possui um genoma com tamanho de 12,1 Mbp (Mbp - milhões de pares de bases), enquanto a planta ornamental *Fritillaria assyriaca* possui um genoma com tamanho de 120.000 Mbp. O ser humano possui um genoma com aproximadamente 3.200 Mbp. Os procariotos possuem genomas menores que giram em torno de 0,58 Mbp (como na bactéria *Mycoplasma genitalium*) a 30 Mbp (como na bactéria *Bacillus megaterium*). Esta diferença de tamanho de genoma entre eucariotos e procariotos tem grande relação com a diferença de complexidade de genomas dos dois grupos.

Os eucariotos possuem uma organização do genoma mais complexa que os procariotos. Os seus genes geralmente são localizados distantes um dos outros e a maior parte do genes é compostas por

exons separados por grandes introns. Além disso, no momento em que o mRNA de um gene eucarioto é sintetizado, ele sofre a adição de uma cauda poli-A, um trecho formado por dezenas de nucleotídeos do tipo Adenina, para aumento da estabilidade da molécula, que será transportada do núcleo para o citoplasma.

Os genes presentes nos genomas de organismos eucariotos costumam ocupar uma baixa porcentagem de toda a seqüência de DNA. No caso do ser humano, por exemplo, os genes ocupam somente 3% de todo o genoma nuclear. Além disso, os organismos eucariotos costumam ter a presença de um grande número de elementos repetitivos em seus genomas.

Os procariotos possuem menos genes que os genomas dos eucariotos. Além disso os seus genes costumam se localizar próximos uns dos outros e a imensa maioria dos genes não possuem introns. Os genes ocupam a maior parte do genoma de um procarioto. Outra característica é a baixa freqüência de seqüências repetitivas.

Estas diferenças nas características dos genomas destes dois grupos de organismos fazem com que diferentes estratégias sejam adotadas pelos projetos de seqüenciamento de genoma de cada tipo de organismo.

2.4 Seqüenciamento de DNA

Os primeiros procedimentos rápidos e eficientes para seqüenciamento de DNA foram desenvolvidos no meio da década de 1970. Dentre os métodos existentes, o método de terminação de cadeia [32] é o mais utilizado.

2.4.1 Método de terminação de cadeia

O método de terminação de cadeia é o mais utilizado porque permite maior automatização. O seu princípio básico é que moléculas de DNA que diferem em comprimento por apenas um nucleotídeo podem ser separadas umas das outras através da eletroforese em gel de poliacrilamida. Neste experimento, o primeiro passo é preparação de fitas únicas de DNA idênticas. Para isso, um pequeno oligonucleotídeo é ligado a uma mesma posição em cada fita de DNA disponível. A função deste oligonucleotídeo é de atuar como um primer (iniciador) para a síntese da fita complementar.

A síntese da fita complementar é catalisada pela enzima DNA polimerase e necessita da presença dos quatro tipos de nucleotídeos como substratos. Em condições normais, a síntese ocorreria até que milhares de nucleotídeos fossem polimerizados, no entanto, neste tipo de experimento também são adicionados para um dos 4 tipos de nucleotídeos (A, T, C, ou G) uma pequena quantidade do dideoxinucleotídeo equivalente, que é um nucleotídeo que não possui a extremidade 3' e que portanto, impede que a enzima continue a aumentar a fita de DNA complementar. Ou seja, a fita complementar cresce enquanto um dideoxinucleotídeo não é incorporado. Como um grande número de fitas é produzido, teremos fitas de diversos tamanhos, pois a incorporação do dideoxinucleotídeo é aleatória. Quatros experimentos são gerados, cada um contendo um tipo de dideoxinucleotídeo. Neste ponto, a eletroforese em gel de poliacrilamida entra em ação ao separar as fitas de diferentes tamanhos.

O gel é dividido em quatro faixas e no topo de cada faixa são colocados as fitas complementares de cada um dos 4 experimentos. Quando uma diferença de potencial é aplicada ao gel, as seqüências menores tendem a ir para direção oposta com mais facilidade do que as maiores. Ao final do experimento, veremos no gel uma série de bandas em cada uma das faixas do gel. A menor seqüência será aquela que mais se distanciou do ponto de partida. Se esta seqüência estiver na faixa do nucleotídeo A, significa que a seqüência a ser determinada se inicia com o nucleotídeo A. A segunda banda mais distante do ponto de partida é equivalente a seqüência que possui o tamanho de um nucleotídeo a

mais que a menor seqüência. Se esta banda estiver na faixa do nucleotídeo T, significa que temos até aqui a seqüência AT. Realizando esta análise sucessivamente, até que não seja mais possível distingüir as bandas, é possível determinar a seqüência complementar da fita original. A Figura 3a mostra um exemplo de um experimento com gel de eletroforese. Na figura, cada uma das faixas equivalem a um nucleotídeo e a seqüência de DNA pode ser determinada através da leitura das faixas de baixo para cima.

O método original de terminação de cadeia utilizava marcas radioativas para que os padrões de bandas no gel pudessem ser visto por autoradiografia. O método mais moderno utiliza marcadores fluorescentes, um para cada tipo de dideoxinucleotídeo. A utilização de marcadores fluorescentes foram fundamentais para a automatização do processo de seqüenciamento, pois permitem que as quatro reações com dideoxinucleotídeos possam ocorrer em um mesmo tubo, e que a leitura possa ser feita em uma única faixa de gel, já que o detector consegue diferenciar os sinais. A Figura 3b exibe um exemplo de um experimento com a mesma seqüência utilizada na Figura 3a. A Figura 3c mostra a representação gráfica da leitura dos sinais fluorescentes emitidos pelas bandas e captados pela máquina de seqüenciamento.

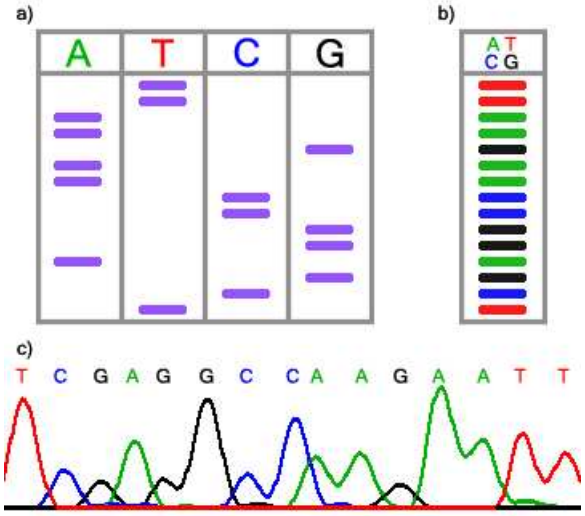

Figura 3: a) Gel de eletroforese feito com um segmento de DNA que contém a seqüência TCGAGGCCAAGAATT. b) Experimento de eletroforese que utiliza marcadores fluorescentes, realizado com o mesmo segmento de DNA. c) Representação gráfica da leitura dos sinais emitidos pelos marcadores fluorescentes, captados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

2.4.2 Clonagem

A clonagem é um procedimento que permite que cópias idênticas de uma molécula de DNA sejam produzidas. Como a técnica de terminação de cadeia necessita de uma grande quantidade de fitas únicas de DNA para funcionar, a clonagem é utilizada para amplificar a quantidade de fitas disponíveis.

Dentre as técnicas de clonagem, a mais utilizada é a clonagem por vetor, e dentre os possíveis vetores, os mais utilizados são os plasmídeos. Este método consiste na inserção do trecho de DNA que se deseja replicar em um ponto específico de um plasmídeo, que será inserido dentro de uma bactéria. Uma vez dentro da célula bacteriana, o plasmídeo será duplicado e cada cópia ficará com um

descendente produzido pela divisão celular. Como as bactérias possuem um ciclo de vida muito curto e se reproduzem em grande velocidade, em pouco tempo é possível obter uma grande quantidade de cópias do DNA desejado.

2.5 Estratégias de seqüenciamento

Independente da técnica de clonagem ou de seqüenciamento utilizada, as seqüências produzidas pelos experimentos são muito pequenas se comparadas ao genoma de um organismo. Em geral as seqüências possuem algumas centenas de bases, enquanto um genoma pode ter vários milhões de bases. Por isso, um projeto de seqüenciamento trabalha de forma a obter as seqüências pequenas e realizar a montagem destas seqüências para conseguir a seqüência genômica completa. A montagem é a união de trechos de seqüências segundo critérios como, por exemplo, as sobreposições que os trechos possuem, para obtenção de uma seqüência maior. Um exemplo de estratégia de seqüenciamento é o shotgun.

2.5.1 Shotgun

Este método consiste na quebra do DNA genômico em fragmentos através da sonicação, que é o emprego de ondas sonoras de alta freqüência para cortar a molécula de DNA. Os fragmentos são então separados para obtenção de trechos que tenham entre 1,6 e 2,0 kbp. Os pedaços selecionados são então clonados para aumento da redundância, que é necessária para que o DNA genômico seja todo coberto por eles. Após a clonagem, o seqüenciamento dos fragmentos é feito e a etapa seguinte é o processamento destas seqüências por computadores que operam de forma a unir os pedaços que se sobrepõem. O próximo passo é a obtenção dos trechos referentes ao buracos que não foram seqüenciados, para que o genoma possa ser completado.

A obtenção do genoma completo de organismo complexos como, por exemplo, o homem ou o milho é uma tarefa extremamente difícil. Por isso, outras técnicas são utilizadas para obtenção de informações do genoma sem a necessidade de seu seqüenciamento completo. Um exemplo é o seqüenciamento de ESTs, que visa obter as seqüências dos genes expressos pelo organismo e será explicado na Seção 2.7.

2.6 Projetos de Seqüenciamento de Genoma

Os projetos de seqüenciamento completo de genomas, conhecidos como Projetos Genomas, têm como principal objetivo a descoberta de toda a seqüência genômica de um organismo. Diversos projetos Genomas já foram e estão sendo realizados. Graças a estes projetos, hoje conhecemos o genoma de diversas espécies, incluindo o do ser humano.

A conclusão do seqüenciamento do genoma da bactéria *Xyllela fastidiosa* em 1999 [35, 50], em particular, colocou o Brasil em posição de destaque mundial, pois foi o primeiro país do mundo a concluir a montagem do genoma de um fitopatógeno, organismo causador de doenças em plantas.

Na página Entrez Genome [10] mantida pelo NCBI [22] é possível encontrar os genomas completos de 1358 vírus, 19 arqueobactérias, 176 bactérias e de diversos eucariotos como o homem e o rato (versão de 06 de Setembro de 2004). Além disso, há a lista de todos os projetos de seqüenciamento que estão em andamento.

Mas qual a utilidade da descoberta da seqüência genômica de um organismo? De posse da seqüência genômica os cientistas podem determinar quais são os genes existentes no organismo, o que cada gene produz, quais genes são relacionados a uma determinada característica boa ou ruim.

Em um projeto genoma, a seqüência genômica é seqüenciada e analisada em busca de regiões que apontem início dos genes. Além disso, no caso de organismos eucariotos, uma análise extra precisa ser feita para identificação dos exons e introns dos genes.

O estudo da seqüência genômica permite também que mutações que ocorreram ao longo da evolução do organismo possam ser analisadas para que informações filogenéticas ou relacionadas com doenças possam ser obtidas.

Curas ou mecanismos de prevenção de doenças poderão ser descobertas com a análise dos genes que estão relacionadas com elas. Por exemplo, no caso da *Xylella fastidiosa* a obtenção de seu genoma poderá auxiliar na prevenção e na cura da doença do amarelinho que ela provoca na laranja, uma planta economicamente importante para o Brasil.

Estas são apenas algumas das muitas atividades que podem ser derivadas do estudo do genoma de um ser vivo.

2.7 Projetos de Seqüenciamento de ESTs

Os projetos de seqüenciamento de ESTs (*Expressed Sequence Tag*) [1] são realizados com o objetivo de rapidamente obter uma boa aproximação do índice gênico de um organismo, que é a a listagem dos genes existentes em seu genoma.

A estratégia adotada por esta técnica é a de realizar o seqüenciamento de segmentos de cDNA (DNA complementar), que é uma fita de DNA produzida a partir do complemento do mRNA com a utilização da enzima transcriptase reversa.

Como já dito anteriormente, o mRNA é a molécula de RNA produzida pela célula, a partir da transcrição do gene contido no DNA, e que será utilizada para produção de proteínas na fase de tradução. Assim, o cDNA nada mais é que a seqüência de nucleotídeos de um gene existente no genoma do organismo.

O processo de seqüenciamento com a utilização de ESTs envolve a produção de bibliotecas de cDNA, a clonagem dos cDNAs com a utilização vetores (em geral bactérias), e o seqüenciamento dos clones através de uma única leitura em uma máquina de seqüenciamento, que torna esta técnica de baixo custo, em relação às outras técnicas existentes.

Os genes de um organismos não são expressos com igual freqüência. Existem genes que são expressos a toda hora pois produzem proteínas ligadas às vias metabólicas que regem as reações químicas que ocorrem no organismo, e existem genes que são expressos apenas quando o organismo é submetido à condições especiais, e, além disso, tecidos diferentes expressam genes diferentes. Graças a estas características, este tipo de seqüenciamento necessita que sejam produzidas bibliotecas de cDNA com origem em diversos tecidos, extraídos sob diferentes condições, tais como, idade, ambiente, presença de doenças, etc.

Na versão 090304 de 03 de Setembro de 2004 do dbEST [8] estavam disponíveis 23.416.084 seqüências públicas de ESTs de 741 organismos diferentes. A Tabela 1 exibe a lista dos 10 organismos com maior quantidade de seqüências ESTs depositadas no dbEST nesta mesma versão.

2.7.1 Problemas encontrados nos Projetos ESTs

Os projetos baseados em EST também possuem problemas. Devido ao fato do seqüenciamento ser feito em apenas uma leitura, os ESTs possuem uma taxa de erro tão alta quanto 3% [36]. Além disso, por causa da limitação da técnica, apenas as pontas 3' e 5' são seqüenciadas no método padrão e, normalmente, não conseguem cobrir todo o gene por possuírem apenas algumas centenas de bases de comprimento. Uma técnica de seqüenciamento denominada ORESTES [5, 41] pode ser utilizada

Organismo	Número de ESTs
Homo sapiens (homem)	5.679.423
Mus musculus + domesticus (camundongo)	4.246.846
Ciona intestinalis (cordado invertebrado)	684.280
Rattus sp. (rato)	683.238
Danio rerio (peixe paulistinha)	575.250
Triticum aestivum (trigo)	561.713
Gallus gallus (galinha)	495.092
Bos taurus (touro)	493.329
Xenopus laevis (rã)	432.424
Xenopus tropicalis (rã)	423.107

Tabela 1: Lista dos 10 organismos com maior número de seqüências no dbEST (versão 090304 de 3 de Setembro de 2004).

como complemento à técnica EST pois produz seqüências que tendem a se concentrar na parte central do gene.

Além dos erros de seqüenciamento, as seqüências sofrem vários tipos diferentes de contaminação, dependendo, em parte, de qual dos muitos protocolos foi utilizado na construção das bibliotecas de cDNA.

2.8 Projetos de Seqüenciamento no Brasil

No Brasil existem diversos grupos de pesquisa realizando projetos de seqüenciamento completo ou ESTs. Os resultados produzidos por estes grupos colocaram o país em posição de destaque no cenário científico internacional.

Os principais projetos realizados no Brasil recebem apoio financeiro da FAPESP, do MCT e do CNPq, e serão citados a seguir.

2.8.1 FAPESP

A FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo [11] em parceria com outras intituições financia uma série de projetos. Em 1997 ela organizou a rede ONSA (Organization for Nucleotide Sequencing and Analisys), um instituto virtual de genômica formado inicialmente por 30 laboratórios ligados a instituições de pesquisa do Estado de São Paulo. A FAPESP participa de diversos projetos Genomas e de seqüenciamento de ESTs e alguns deles são listados abaixo:

- *Xyllela fastidiosa* - Primeiro projeto realizado por esta rede e que já teve a etapa de seqüenciamento concluída, tendo como resultado a publicação de um artigo que foi capa da revista Nature, por ser o primeiro organismo fitopatógeno a ser seqüenciado [35].
- Genoma Cana-de-Açúcar (SUCEST) - Segundo projeto da rede. Projeto de seqüenciamento de ESTs que já teve a etapa de seqüenciamento concluída [39, 45, 37].
- *Xanthomonas citri* e *Xanthomonas campestri* - Duas bactérias fitopatógenas que também tiveram o seqüenciamento concluído pela ONSA [49].
- Projeto Genoma Humano do Câncer - Projeto EST realizado internacionalmente e que tem como objetivo a descoberta de genes relacionados a diversos tipos de câncer. A parcela brasileira deste projeto é formada por diversos grupos, incluindo o Instituto Ludwig para Pesquisa do Câncer [41].
- Projeto FORESTs - Projeto EST de seqüenciamento do eucalipto [12].

- Projeto *Schistossoma mansoni* - Projeto EST de seqüenciamento do organismo causador da esquistossomose [34].
- Projeto Genomas Agronômicos e Ambientais - Projeto multi-genômico de seqüenciamento completo ou EST de organismos ligados ao ambiente e à agronomia (*Xylella fastidiosa*/Doença de Pierce, *Leifsonia xyli subsp. xyli*, Genoma Café e *Leptospira interrogans*) [13].

2.8.2 MCT e CNPq

O Ministério da Ciência e Tecnologia (MCT) [21] e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [7] financiam uma série de projetos em todo o país.

Genoma Nacional A Rede Genoma Nacional [3] ou Genoma Brasileiro é formada por 25 laboratórios de seqüenciamento, 1 laboratório de processamento de DNA e 1 centro de Bioinformática, incluindo o trabalho de aproximadamente 100 pesquisadores. Esta rede, criada em 2000, desenvolveu o projeto de seqüenciamento da bactéria *Chromobacterium violaceum* que já foi concluído [9], e desenvolve atualmente o projeto de seqüenciamento da bactéria *Mycoplasma synoviae*.

Redes Regionais Além da Rede Genoma Nacional o MCT e o CNPq promovem a implantação de diversas redes gênomicas, espalhadas por todas as regiões do país e que realizam vários projetos:

- Rede Genoma do Estado de Minas - EST - *Schistosoma mansoni* [15].
- Programa Genoma Nordeste (ProGeNe) - EST - *Leishmania chagasi* [28].
- Rede Genoma Centro-Oeste - EST - *Paracoccidioides brasiliensis* [24].
- Rede Genoma do Consórcio do Instituto de Biologia Molecular do Paraná, FIOCRUZ e Universidade de Mogi das Cruzes - EST - *Trypanossoma cruzi* [42].
- Rede Genômica do Estado da Bahia e São Paulo - EST - *Crinipellis perniciosa* [19].
- Rede Genoma do Rio de Janeiro (RioGene) - Genoma completo - *Gluconacetobacter diazotrophicus* [31].
- Programa Genoma do Estado do Paraná (GenoPar) - Genoma completo - *Herbaspirillum seropedicae* [16].
- Rede Sul de Análise de Genomas e Biologia Estrutural (PROGENESUL) - Genoma completo - *Mycoplasma hyopneumoniae* [29].
- Rede da Amazônia Legal de Pesquisas Genômicas (REALGENE) - EST - *Paullinia culpana* [30].

O MCT financia ainda o Projeto Genolyptus [14] que é uma parceria público-privada para o seqüenciamento EST do eucalipto executado pela Rede Brasileira de Pesquisa do Genoma do Eucalipto, formada por 12 empresas, 7 universidades e pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa).

2.9 Bioinformática

A resolução dos problemas apresentados pelos projetos de seqüenciamento é uma importante área de atuação da Bioinformática. A Bioinformática é a área da Computação destinada a desenvolver ferramentas para análise de dados e resolução de problemas em aplicações biológicas.

A evolução que ocorreu na capacidade de processamento dos computadores permitiu que a Bioinformática desenvolvesse softwares para lidar com o imenso volume de dados produzidos pelos diversos

projetos na área de Biologia, especialmente os projetos de seqüenciamento completos ou ESTs. Tais projetos produzem inúmeras seqüências que precisam ser processadas de forma automática e com a menor taxa de erros possível. A automatização é extremamente importante pois o trabalho de processamento e análise dos dados é impossível de ser feito manualmente.

Entre as diversas atividades desenvolvidas pela Bioinformática na análise de dados de projetos de seqüenciamento, podemos citar a trimagem, a verificação de contaminação e a clusterização, que serão os objetos de estudo do trabalho aqui proposto.

3 Trimagem

A Trimagem é o processo de limpeza das seqüências produzidas pelo processo de seqüenciamento. Ela é responsável pela remoção de regiões que apresentem baixa qualidade ou que são indesejadas por causarem incidência de erros nas análises dos dados. Neste trabalho, tais regiões serão denominadas artefatos.

3.1 As origens dos artefatos

O processo de obtenção da seqüência de nucleotídeos de um segmento de DNA envolve a realização de uma série de experimentos biológicos. O DNA alvo precisar ser replicado para que haja quantidade de material suficiente para processá-lo e isto é normalmente feito com a utilização de vetores. O DNA alvo é inserido em um local específico na molécula de DNA do vetor. Devido ao fato da técnica de seqüenciamento EST produzir trechos nas extremidades dos genes, é normal a presença de vetor nas seqüências lidas pela máquina de seqüenciamento. Estes trechos de vetor são artefatos.

Seqüências com baixa complexidade são consideradas artefatos, e as seqüências de poli-A e poli-T são exemplos de seqüências deste tipo. Como vimos na seção anterior, no momento da produção do mRNA, nos organismos eucariotos, uma cauda poli-A é ligada ao seu final. O mRNA é utilizado para produção do cDNA, que será seqüenciado. Conforme a direção do seqüenciamento, isto é, conforme a fita do cDNA que for seqüenciada, trechos de poli-A ou poli-T podem aparecer.

Alguns tipos de vetores utilizados para clonagem necessitam da utilização de um pequeno segmento de DNA chamado adaptador para que a inserção do DNA alvo possa ser feita. Esta seqüência que não pertence nem ao vetor e nem ao organismo estudado é um artefato.

Relacionados ao processo de leitura realizado pelas máquina de seqüenciamento temos os artefatos de baixa qualidade. O valor de qualidade de uma base indica a probabilidade dela estar correta. Quanto menor o valor de qualidade, maior a probabilidade de erro. A precisão da leitura realizada pela máquina de seqüenciamento depende da intensidade dos sinais emitidos pelos marcadores fluorescentes, que tendem a ser mais fracos nas extremidades do gel e mais fortes na porção central.

Os sinais fracos emitidos por algumas bases podem gerar erros, mas os sinais fortes também podem acarretar em erros. Quando a máquina de seqüenciamento encontra uma região com picos de sinais muito altos, o fenômeno de derrapagem pode acontecer. Este fenômeno se caracteriza pela repetição de bases de maneira anormal. Devido aos fortes sinais, a máquina pode interpretar a existência de mais de uma base, onde na verdade só existe uma, como se a seqüência estivesse borrada.

3.2 Problemas causados pelos artefatos

A presença de artefatos nas seqüências podem influenciar negativamente os resultados das análises dos dados produzidos pelo projeto.

As seqüências de baixa qualidade são seqüências que possuem taxas de erros muito altas. A manuntenção de seqüências deste tipo seria uma atitude imprudente, pois não pode se dizer com

uma boa margem de segurança que um trecho de baixa qualidade realmente represente a seqüência determinada pelo programa de base calling (software que realiza a determinação das bases a partir da saída produzida pela máquina de seqüenciamento).

Em um processo de clusterização, as seqüências de vetores e as seqüências de baixa complexidade podem forçar a criação de clusters através do agrupamento errôneo de seqüências por causa da adição de similaridade não relevante ao processo [40]. Como o critério utilizado na montagem dos clusters é a sobreposição das seqüências, os trechos de baixa complexidade poderiam gerar sobreposições válidas no critério do software de clusterização, mas que na realidade não existem.

Os artefatos que representam vetores, adaptadores e derrapagens são seqüências que não pertencem ao organismo alvo do projeto. A presença destas seqüências podem ocasionar erros na identificação de genes e na montagem do genoma do organismo.

Para a eliminação de possíveis problemas os projetos de seqüenciamento operam de forma a remover estas seqüências utilizando diversas técnicas.

3.3 Técnicas de Trimagem

Cada tipo de artefato possui formas de detecção e remoção apropriadas que discutiremos a seguir. Normalmente, após a aplicação destas técnicas o tamanho da seqüência restante é verificado. Se ela tiver tamanho menor que um certo valor mínimo, a seqüência é descartada de futuras análises.

3.3.1 Trimagem de artefatos de baixa qualidade

A remoção de trechos de baixa qualidade pode ser atacada de várias maneiras, que podem ser simples ou mais elaboradas. As soluções mais simples são obviamente as mais rápidas, um fator importante quando o volume de dados a ser processado é muito grande. Assim, a decisão da estratégia a ser utilizada depende do tempo que se deseja gastar com esta tarefa.

O valor de qualidade de uma base determinada por um programa de base calling como o phred [17] ou o TraceTuner [43] é baseada na probabilidade de erro que essa base possui e é dada pela fórmula $Q = -10 \times \log_{10}(\textit{probabilidade de erro})$ [6]. Assim, quanto maior a probabilidade de erro, menor a qualidade.

Uma estratégia simples para a trimagem de baixa qualidade é a utilização de um algoritmo para determinação da subseqüência máxima [20]. A seqüência determinada por esse algoritmo seria a seqüência com as pontas de baixa qualidade removidas. O próprio programa phred possui um parâmetro que faz com que ele indique qual é essa subseqüência, em uma seqüência que ele acabou de determinar. O algoritmo implementado pelo phred converte o valor de qualidades em probabilidades de erros e tenta minimizar a probabilidade de erro da subseqüência. Antes de executar o algoritmo cada base tem sua probabilidade de erro subtraída de 0,05. Este valor equivale ao valor de qualidade 13, a mínima aceitável segundo esta implementação do algoritmo.

Muitos projetos realizam a análise através de janelas deslizantes. Em geral, a seqüência é percorrida base a base, nas duas direções a partir das extremidades, por uma janela de um determinado tamanho em busca de trechos que possuam um número máximo de bases com qualidades menores que a mínima. No trabalho desenvolvido por Telles e da Silva [40], por exemplo, utilizou-se uma janela de tamanho 20, que devia ter no máximo 12 bases com qualidade abaixo de 10.

Alguns projetos de seqüenciamento utilizam programas específicos para o processo de trimagem como, por exemplo o ESTprep [33].

Este programa faz a trimagem de qualidade em duas etapas. Na primeira etapa, o programa verifica se o trecho inicial de 20 bases da seqüência possui menos que 8 bases com qualidade maior

que 20, caso em que será removido. Nesta mesma etapa, o software verifica se a qualidade média das 200 primeiras bases é menor que 20, caso em que toda a seqüência é descartada.

Na segunda etapa, após a trimagem de outros tipos de artefatos, o ESTprep utiliza uma janela deslizante de 20 bases para identificar a primeira região com no máximo 8 bases com qualidade menor que 10 para determinação do ponto de trimagem na extremidade 3', que será o início da região encontrada.

Outro software de trimagem é o LUCY [6], que é utilizado pelo TIGR - The Institute of Genomic Research. Ele possui uma estratégia de análise mais complexa.

Como o início e o final da seqüência são em geral de baixa qualidade, o LUCY age de forma a identificar estes trechos primeiro. A partir da ponta esquerda da seqüência uma janela de tamanho 10 percorrerá a seqüência até encontrar um trecho que tenha uma probabilidade de erro menor ou igual a 2%. O mesmo será feito na ponta direita da seqüência. Estes trechos identificados são removidos e o que sobrar passará pelo processo restante. Se a seqüência inteira não passar no teste, ela será inteiramente descartada.

O próximo passo é a identificação de trechos que possuem taxas de erros altas para que possam ser eliminados. Neste passo, duas janelas são utilizadas. A primeira tem tamanho 50 e um valor limite de probabilidade de erro igual a 8% e elimina os trechos grandes de baixa qualidade. A segunda tem tamanho 10 e um valor limite de probabilidade de erro igual a 30% e elimina os trechos pequenos que não são removidos pela primeira.

A primeira janela percorre a seqüência resultante do primeiro passo de limpeza. A partir do início desta seqüência, o programa calcula para a janela o valor médio de probabilidade de erro. Se o valor estiver dentro do valor limite a janela será adicionada a seqüência candidata, que continuará crescendo enquanto as janelas que percorrerem a seqüência estiverem com o valor dentro do limite. Se o valor estiver fora do limite, a seqüência candidata será terminada e separada para o próximo passo. A janela continuará a percorrer a seqüência até o final, se uma nova janela voltar a ter o valor dentro do limite, uma nova seqüência candidata será iniciada.

Cada seqüência candidata será percorrida pela segunda janela seguindo o mesmo critério. Após este processo, todas as seqüências candidatas com tamanho menor que o mínimo serão descartadas. Dentre as seqüências restantes, aquela que tiver um uma probabilidade de erro geral menor ou igual que 2,5% e probabilidades de erros nas extremidades menor que 2% será a seqüência final. A probabilidade de erros nas extremidades é avaliada com as duas últimas bases de cada ponta. No caso raro de mais de uma seqüência atender ao critério, a maior será mantida.

3.3.2 Trimagem de vetores e adaptadores

Uma maneira de se realizar a remoção de vetores e adaptadores é através da análise da seqüência com o programa cross_match [17]. Este programa recebe como entrada a seqüência do vetor ou do adaptador e a seqüência a ser processada e realiza o mascaramento (substituição das letras das bases por Xs) dos artefatos encontrados. Assim, analisando as regiões marcadas com X é possível indentificar os trechos de vetores e adaptadores e removê-los.

O programa LUCY também realiza a trimagem de vetores e adaptadores. Para este serviço, ele necessita das seqüências dos trechos do vetor e adaptor no ponto onde o inserto é fixado (splice sites upstream e downstream, ou seja, as regiões vizinhas ao ponto onde a seqüência do vetor foi cortada para inserção da seqüência a ser replicada).

Como os artefatos relacionados aos vetores costumam se localizar no início da seqüência onde a qualidade das bases é geralmente baixa, uma comparação simples que busca pelo alinhamento mais longo pode não encontrar todos os trechos de vetor devido aos erros de base-calling. Assim, o software realiza uma busca adaptativa pelos valores médios de qualidades das bases. Nas regiões de

baixa qualidade o programa permite que pequenos trechos de vetor sejam identificados enquanto em regiões de melhor qualidade apenas trechos maiores são identificados. Devido as diferentes regiões de qualidades existentes no início da seqüência o software considera três critérios diferentes. A busca é feita em áreas de 40, 60 e 100 bases com comprimentos mínimos de alinhamento de 8, 12 e 16 bases. Um alinhamento local ótimo dentro de cada área deve ter pelo menos o comprimento mínimo para ser considerado vetor. Estas janelas são colocadas no início da seqüência original para evitar que fragmentos de vetores sejam perdidos em seqüências que possuem um trecho de baixa qualidade muito longo no início.

O splice site upstream será procurado nas primeiras 200 bases. LUCY procurará pelo maior alinhamento com pelo menos 3 bases corretas para cada base incompatível, o que não significa que haverá 25% de erro porque apenas o alinhamento com maior pontuação local será utilizado. Alinhamentos menores a esquerda podem ser ignorados. Se ainda existirem bons alinhamentos após o melhor, o programa continuará a busca até que todos os fragmentos sejam identificados. Depois de terminar a busca pelo splice site upstream, o dowstream é também procurado, pois a seqüência pode conter um inserto pequeno. O splice site downstream é procurado utilizando-se o critério de alinhamento mínimo de 16 bases.

3.3.3 Trimagem de poli-A e poli-T

A buscas por caudas poli-A e poli-T também variam em complexidade. O programa LUCY, por exemplo, utiliza um esquema simples. Ele realiza a busca por caudas poli-A/T utilizando uma janela de 50 bases para identificação de trechos que possuam no mínimo 10 bases Ts ou As. Nesta busca, são permitidos no máximo três bases incompatíveis entre cada trecho de 10 Ts ou As.

O procedimento de trimagem de seqüências desenvolvido por Telles e da Silva realiza a remoção destes artefatos após a remoção dos trechos de vetores, que foi feita com a utilização do programa cross_match. Os trechos de poli-A/T são identificados através do alinhamento da seqüência sem vetores com seqüências de prova compostas apenas por As ou Ts, conforme o tipo de cauda procurado. Um trecho de seqüência é considerado poli-A/T se possuir um alinhamento com pontuação de pelo menos 8 e se localizar a no máximo 10 bases de distância das extremidades da seqüência. O alinhamento é feito com o programa swat [17] utilizando o seguinte esquema de pontuação: 1 para cada coincidência, -2 para cada diferença e -8 para cada buraco aberto.

O ESTprep utiliza uma estratégia diferente. Em primeiro lugar, ele percorre a seqüência em busca do primeiro nucleotídeo da cauda após a identificação do sítio de restrição. A partir desta posição, uma seqüência maximal formada apenas por A/Ts é construída de tal forma que ela possua similaridade maior que um limite pré-estabelecido (95%) em relação a sequëncia original. Desta maneira, a região mais rica em A/T é encontrada. Se esta região não termina com um A/T, ela é retraída em uma base, o que é repetido até que a última base seja um A/T. Se a cauda de poli-A/T não tiver tamanho suficiente (10 bases), a busca é refeita começando uma base à esqueda/direita do ponto de início original. Se a cauda é identificada, a busca é repetida utilizando-se um limite menor (94% do limite original) para evitar o truncamento de caudas poli-A/T grandes. Se depois de todos estes passos a cauda não foi identificada, o programa analisa uma janela com o tamanho da distância média entre o sítio de restrição e o trecho que identifica o tecido (18 bases) em busca de uma densidade suficiente de A/Ts (65%). Após a identificação do poli-A, o programa tenta localizar os sinais de poliadenilação. Os sinais procurados podem ser canônicos (AAUAAA ou AUUAAA) ou alternativos. Eles devem estar dentro de 11 a 30 nucleotídeos a partir do final da cauda poli-A.

3.3.4 Trimagem de trechos derrapados

No estudo que fizemos apenas Telles e da Silva citam a trimagem de trechos derrapados. Tais trechos eram removidos com a utilização do programa swat que analisava as seqüências em buscas de padrões de repetições anormais.

4 Verificação de contaminação

A contaminação de seqüências é um problema extremamente sério em projetos de seqüenciamento. Ocorrências embaraçosas têm acontecido com freqüência, como, por exemplo, projetos de seqüenciamento em larga-escala que utilizaram bibliotecas de clones altamente contaminadas e tiveram que descartar uma quantidade enorme de seqüências. Outro exemplo, foi o anúncio, em 1995, de que DNA havia sido extraído com sucesso a partir de ossos de um dinossauro [48]. Hoje em dia, este anúncio é visto como, no mínimo, prematuro. As seqüências "extraídas" se mostraram, através de buscas realizadas em bancos de seqüências de DNA, muito mais semelhantes às seqüências de mamíferos, do que de aves ou crocodilos, sugerindo que o DNA utilizado na análise fosse, na verdade, uma contaminação humana e não DNA de dinossauros [51, 4].

4.1 Tipos de contaminação

Existem vários tipos de contaminação que variam conforme o protocolo utilizado na produção de bibliotecas e na clonagem das seqüências [36]. As contaminações podem ser separadas em 2 grupos diferentes: contaminações causadas por seqüências de outros organismos e contaminações causadas por seqüências do próprio organismo.

4.1.1 Contaminação por seqüências de outros organismos

O vetor utilizado na clonagem pode ser uma fonte de contaminação. Devido a eventos de rearranjo de genoma, seqüências do vetor podem ser inseridas no meio do inserto, formando uma seqüência híbrida. Neste trabalho, os casos de contaminação por vetor serão tratados na fase de trimagem.

Em um laboratório de seqüenciamento, é comum a execução de experimentos com organismos diferentes. Acidentalmente, é possível que uma biblioteca de clones de seqüências de um organismo seja contaminada com seqüências de outro organismo estudado no mesmo laboratório.

Existem projetos de seqüenciamento que lidam com tecidos que podem estar contaminados. Por exemplo, é comum a preparação de ESTs de tecidos atacados por alguma doença para obtenção dos genes que são expressos quando um organismo está doente. No meio do conjunto de seqüências do organismo podem existir ESTs originários em mRNAs do organismo patógeno.

Outra possibilidade de contaminação ocorre quando se estuda organismos que vivem relações simbiontes. Existe a possibilidade de contaminação por seqüências do organismo que vive com o organismo estudado, pois durante a coleta de material existe a possibilidade da obtenção de DNA de ambos.

4.1.2 Contaminação por seqüências do próprio organismo

Seqüências do próprio organismo também podem causar contaminação. Ela ocorre quando existe a formação de ESTs contendo trechos de seqüência que não possuem origem no mRNA processado.

O EST é uma seqüência produzida através do mRNA, mas durante o processo de produção das bibliotecas, pode ocorrer do rRNA ser utilizado pela enzima transcriptase reversa para produção do cDNA.

As células eucariotas tem mitocôndrias e as células eucariotas de vegetais também possuem cloroplastos. Estas organelas possuem um genoma próprio e conforme a natureza do projeto, o seqüenciamento de genes destas organelas pode ser desnecessário ou, até mesmo, indesejado.

Como vimos anteriormente, o mRNA, após a transcrição, é processado para remoção dos introns. Pode acontecer de um mRNA prematuro originar cDNA, que apesar de ter origem em um mRNA, contém trechos que não pertecem à porção codificante do gene.

Problemas no protocolo de produção de bibliotecas podem gerar cDNAs contendo trechos de DNA genômico, que não fazem parte de genes, algo indesejado quando se deseja obter o indice gênico de um organismo.

Finalmente, eventos de rearranjo de genoma podem gerar seqüências quiméricas. Estas seqüências se caracterizam por conter trechos de dois ou mais genes que possuem origens em pontos diferentes do genoma.

4.2 Técnicas de detecção de contaminação

A maior parte dos projetos utilizam a similaridade para a detecção de contaminação [26]. Normalmente o programa BLAST [2] é utilizado neste método, que consiste na comparação da seqüência a ser analisada com as seqüências existentes em um banco de contaminantes. Este banco contém seqüências dos possíveis organismos contaminantes, de rRNA e de genes de cloroplatos e mitocôndrias. Os critérios para detecção de contaminação através da similaridade podem variar entre diferentes projetos.

Ao indicar que uma seqüência é similar a de um contaminante existente no banco, este método pode dizer que ela é um possível contaminante, contudo, nada pode-se dizer das seqüências que não apresentaram similaridade com nenhuma do banco. Existe a possibilidade de algumas seqüências pertencerem a organismos não existentes no banco de contaminantes. Além disso, no caso do banco de contaminantes ser muito grande, a busca por similaridade pode ser muito custosa.

Além da detecção de contaminação através da similaridade, existem as técnicas que aplicam as características encontradas nos genomas dos organismos como critério. A abordagem destas metodologias é classificar as seqüências em dois grupos (seqüências pertencentes ao organismo alvo e seqüências não pertencentes ao organismo alvo) de acordo com as características obtidas pela análise delas em comparação com as obtidas através de um conjunto de treino formado por seqüências do próprio organismo e, opcionalmente, dos organismos contaminantes.

Diversas características diferentes podem ser utilizadas. O trabalho desenvolvido por White *et al.* [47], por exemplo, utiliza a composição de hexâmeros. O trabalho realizado por Piazza e Setubal [26, 27] emprega uma gama maior de características com o objetivo de melhorar a precisão da detecção de contaminantes.

Este tipo de metodologia é mais indicado para detecção de contaminações por DNA do próprio organismo. A análise é feita comparando-se as assinaturas apresentadas pelos ESTs contra as verificadas em genes do organismo. As assinaturas dos genes costumam ser bastante diferentes do restante do genoma, e esse fato pode auxiliar na detecção de contaminação por DNA genômico ou por mRNA prematuro.

As metodologias baseadas em características são menos utilizadas que as baseadas em similaridade. Elas apresentam uma taxa de erros maior e possuem a desvantagem de necessitarem de treino com seqüências do organismo, o que nem sempre é possível.

A detecção de seqüências quiméricas requer uma análise cuidadosa, pois trata-se de seqüências que são formadas pela fusão de dois ou mais genes. Como as quimeras são formadas geralmente por concatenação de genes de diferentes regiões do genoma, a maneira mais apropriada seria a utilização da comparação com o genoma completo do organismo, o que nem sempre é possível, especialmente no

caso de projetos ESTs. Uma alternativa é a utilização de algoritmos sofisticados capazes de modelar as propriedades dos ESTs. Quimeras de genes conhecidos podem ser identificadas ao observar-se a concatenação de trechos de genes não relacionados.

5 Clusterização

Outro procedimento de interesse da Bioinformática é o processo de clusterização de seqüências. A clusterização consiste no agrupamento de seqüências semelhantes em um grupo chamado cluster. A idéia básica é unir as seqüências provenientes do mesmo gene, para evitar a predição incorreta dos genes existentes no genoma de um organismo e, assim, obter índices gênicos mais confiáveis. Isto é preciso ser feito porque os projetos de seqüenciamento produzem muita redundância de dados. É comum um gene muito expresso pelo organismo ser seqüenciado várias vezes, e dessa maneira, todos os clones que tiveram origem nele devem ser agrupados no mesmo cluster.

5.1 Técnicas de clusterização

As técnicas de clusterização utilizam a similaridade como principal ferramenta. Através das comparações entre cada par de seqüências é possível determinar aquelas que se sobrepõem e que, portanto, podem ter origem no mesmo gene.

Diferentes implementações de processos de clusterização estão disponíveis e elas diferem na estratégia adotada para otimizar o desempenho sem comprometer a qualidade dos clusters produzidos.

Os programas de clusterização normalmente realizam uma avaliação inicial para identificação das seqüências que podem se sobrepor. Feita a identificação, as seqüências são alinhadas para construção dos clusters.

O ideal seria que todas as seqüências pudessem ser unidas através de um alinhamento múltiplo, no entanto, este é um problema NP-completo. Isso explica a necessidade de se fazer uma avaliação inicial em busca de seqüências que se sobrepõem e a aplicação de uma série de heurísticas realizada por diversos softwares.

Os clusters produzidos podem ter ou não associados a eles seqüências consensos. A seqüência consenso é derivada através da análise das seqüências que formam o cluster e é aceita como a seqüência com maior probabilidade de ser a do gene existente no organismo.

Alguns programas de clusterização podem realizar o alinhamento das seqüências existentes nos clusters com o objetivo de produzir melhores consensos.

Um dos programas mais utilizados é o CAP3 [18], mas existem muitos outros, como o Phrap [17], o TIGR Assembler [38], o UIcluster [44] e o TGICL [25].

6 Proposta

Diante da importância dos projetos de seqüenciamento, este trabalho tem como objetivo o estudo dos processos de trimagem, verificação de contaminação e clusterização aplicados em projetos ESTs.

Normalmente, cada projeto de seqüenciamento executa o seu próprio protocolo de trimagem, verificação de contaminação e clusterização, que pode ser mais ou menos complexo. Assim, este trabalho pretende propor uma metodologia que possa ser aplicada de maneira confiável pelos diversos projetos.

Durante a pesquisa que fizemos, foi possível observar que alguns projetos realizavam limpeza de seqüências sem muito rigor como, por exemplo, a remoção somente da ponta de baixa qualidade.

Tivemos a chance de analisar seqüências trimadas com a técnica proposta por Telles e da Silva e vimos que algumas falhas ocorriam. Os critérios utilizados na identificação de vetores, por exemplo, eram capazes de detectar o vetor na maior parte dos casos, mas observamos várias situações em que ele era removido parcialmente. A remoção parcial deste acabava por induzir erros na detecção de adaptadores, que dependia fortemente da detecção precisa do vetor na seqüência.

Baseados nestas observações, realizaremos testes com os diversos procedimentos de trimagem existentes para detecção de pontos positivos e negativos de cada um deles. Com base nos resultados dos testes, pretendemos desenvolver uma metodologia de limpeza de seqüências mais completa e precisa.

Em relação a verificação de contaminação, realizaremos um estudo para identificar qual critério de seleção baseado em similaridade se aplica melhor ao contexto de um projeto EST. Faremos testes variando parâmetros como percentual de similaridade, valor de e-value e tamanho da seqüência, em busca de um conjunto balanceado que diminua os números de falsos positivos e de falsos negativos.

Realizaremos também um estudo para verificar a validade de se adotar alguma característica como, por exemplo, a composição de hexâmeros na análise de contaminação.

A clusterização será utilizada principalmente como meio de validação dos resultados obtidos nos processos de trimagem e verificação de contaminação. Os clusters serão utilizados para avaliação se houve melhora ou piora nos resultados obtidos com as novas metodologias propostas.

Contudo, este trabalho também realizará a avaliação de diversos programas disponíveis em domínio público. Através da variação de parâmetros de execução, pretendemos avaliar qual o melhor conjunto para cada programa. O objetivo final, é determinar qual o software mais adequado para o serviço de clusterização de ESTs.

7 Cronograma

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

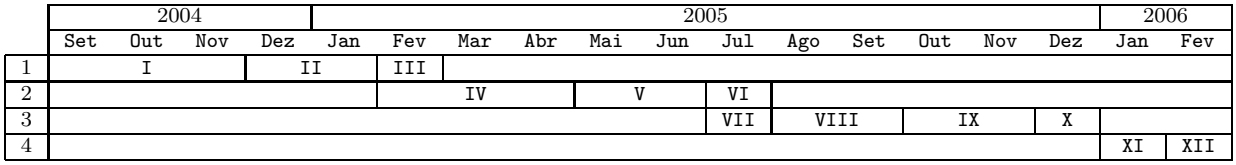

Tabela 2: Cronograma de atividades.

1. Trimagem:
   - `I` - Estudo e identificação de melhorias em métodos de trimagem.
   - `II` - Testes com os novos métodos de trimagem desenvolvidos.
   - `III` - Escrita dos resultados obtidos nos testes de trimagem.
2. Verificação de Contaminação:
   - `IV` - Estudo e identificação de melhorias em métodos de verificação de contaminação.
   - `V` - Testes com os novos métodos de verificação de contaminação desenvolvidos.
   - `VI` - Escrita dos resultados obtidos nos testes de verificação de contaminação.
3. Clusterização:
   - `VII` - Escolha dos programas de clusterização.

- `VIII` - Avaliação dos melhores valores de parâmetros dos programas de clusterização.
- `IX` - Testes com os programas e valores de parâmetros escolhidos.
- `X` - Escrita dos resultados obtidos nos testes de clusterização.

4. Dissertação:
   - `XI` - Revisão final do texto da dissertação.
   - `XII` - Defesa da Dissertação.

Referências

[1] 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.

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

[3] Brazilian Genome (BrGene) – The Virtual Institute of Genomic Research, September 2004. http://www.brgene.lncc.br.

[4] M. Browne. Critics see humbler origin in "dinosaur" DNA, June 1995. New York Times.

[5] A. A. Camargo, H. P. B. Samaia, E. Dias-Neto, D. F. Simão, I. A. Migotto, M. R. S. Briones, F. F. Costa, M. A. Nagai, S. Verjovski-Almeida, M. A. Zago, L .E. C. Andrade, H. Carrer, H. F. A. El-Dorry, E. M. Espreafico, A. Habr-Gama, D. Giannella-Neto, G. H. Goldman, A. Gruber, C. Hackel, E. T. Kimura, R. M. B. Maciel, S. K. N. Marie, E. A. L. Martins, M. P. Nóbrega, M. L. Paçó-Larson, M. I. M. C. Pardini, G. G. Pereira, J. B. Pesquero, V. Rodrigues, S. R. Rogatto, I. D. C. G. da Silva, M. C. Sogayar, M. F. Sonati, E. H. Tajara, S. R. Valentini, F. L. Alberto, M. E. J. Amaral, I. Aneas, L. A. T. Arnaldi, A. M. de Assis, M. H. Bengston, N. A. Bergamo, V. Bombonato, M. E. R. de Camargo, R. A. Canevari, D. M. Carraro, J. M. Cerutti, M. L. C. Corrêa, R. F. R. Corrêa, M. C. R. Costa, C. Curcio, P. O. M. Hokama, A. J. S. Ferreira, G. K. Furuzawa, T. Gushiken, P. L. Ho, E. Kimura, J. E. Krieger, L. C. C. Leite, P. Majumder, M. Marins, E. R. Marques, A. S. A. Melo, M. B. de Melo, C. A. Mestriner, E. C. Miracca, D. C. Miranda, A. L. T. O Nascimento, F. G. Nóbrega, E. P. B. Ojopi, J. R. C. Pandolfi, L. G. Pessoa, A. C. Prevedel, P. Rahal, C. A. Rainho, E. M. R. Reis, M. L. Ribeiro, N. da Rós, R. G. de Sá, M. M. Sales, S. C. Sant'anna, M. L. dos Santos, A. M. da Silva, N. P. da Silva, W. A. Silva Jr., R. A. da Silveira, J. F. Sousa, D. Stecconi, F. Tsukumo, V. Valente, F. Soares, E. S. Moreira, D. N. Nunes, R. G. Correa, H. Zalcberg, A. F. Carvalho, L. F. L. Reis, R. R. Brentani, A. J. G. Simpson, and S. J. de Souza. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *PNAS*, 28(21):12103–12108, October 2001.

[6] H. Chou and M. H. Holmes. DNA sequence quality trimming and vector removal. *Bioinformatics*, 17:1093–1104, 2001.

[7] CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico, July 2004. http://www.cnpq.br.

[8] dbEST – The International Expressed Sequence Tags Database, July 2004. http://www.ncbi.nlm.nih.gov/dbEST.

[9] A. T. R. de Vasconcelos, D. F. de Almeida, M. Hungria, C. T. Guimarães, R. V. Antônio, F. C. Almeida, L. G. P. de Almeida, R. de Almeida, J. A. Alves-Gomes, E. M. Andrade, J. Araripe, M. F. F. de Araújo, S. Astolfi-Filho, V. Azevedo, A. J. Baptista, L. ArturMendesBataus, J. S. Batista, A. Beló, C. den Berg, M. Bogo, S. Bonatto, J. Bordignon, M. M. Brigido, C. A. Brito, M. Brocchi, H. A. Burity, A. A. Camargo, D. D. P. Cardoso, N. P. Carneiro, D. M. Carraro, C. M. B. Carvalho, J. C. M. Cascardo, B. S. Cavada, L. M. O. Chueire, T. B. Creczynski-Pasa, N. C. da Cunha-Junior, N. Fagundes, C. L. Falcão, F. Fantinatti, I. P. Farias, M. S. S. Felipe, L. P. Ferrari, J. A. Ferro, M. I. T. Ferro, G. R. Franco, N. S. A. de Freitas, L. R. Furlan, R. T. Gazzinelli, E. A. Gomes, P. R. Gonçalves, T. B. Grangeiro, D. Grattapaglia, E. C. Grisard, E. S. Hanna, S. N. Jardim, J. Laurino, L. C. T. Leoi, L. F. A. Lima, M. F. Loureiro, M. C.

C. P. de Lyra, H. M. F. Madeira, G. P. Manfio, A. Q. Maranhão, W. S. Martins, S. M. Z. di Mauro, S. R. B. de Medeiros, R. V. Meissner, M. A. M. Moreira, F. F. do Nascimento, M. F. Nicolás, J. G. Oliveira, S. C. Oliveira, R. F. C. Paixão, J. A. Parente, F. de O. Pedrosa, S. D. J. Pena, J. O. Pereira, M. Pereira, L. S. R. C. Pinto, L. S. Pinto, J. I. R. Porto, D. P. Potrich, C. E. Ramalho-Neto, A. M. M. Reis, L. U. Rigo, E. Rondinelli, E. B. P. do Santos, F. R. Santos, M. P. C. Schneider, H. N. Seuanez, A. M. R. Silva, A. L. C. da Silva, D. W. Silva, R. Silva, I. C. Simões, D. Simon, C. M. A. Soares, R. de B. A. Soares, E. M. Souza, K. R. L. de Souza, R. C. Souza, M. B. R. Steffens, M. Steindel, S. R. Teixeira, T. Urmenyi, A. Vettore, R. Wassem, Arnaldo Zaha, and A. J. G. Simpson. The complete genome sequence of *Chromobacterium violaceum* reveals remarkable and exploitable bacterial adaptability. *PNAS*, 100(20):11660–11665, September 2003.

[10] Entrez Genome – Whole Genomes Page, September 2004. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome.

[11] FAPESP – Fundação de Amparo à Pesquisa do Estado de São Paulo, May 2004. http://www.fapesp.br.

[12] FORESTs: Eucalyptus Genome Sequencing Consortium, July 2004. https://forests.esalq.usp.br/.

[13] Agronomical & environmental genomes, July 2004. http://watson.fapesp.br/AEG/agro.htm.

[14] Genolyptus, July 2004. http://www.lge.ibi.unicamp.br/eucalyptus/.

[15] Genome Network of the State of Minas Gerais, May 2004. http://www.cpqrr.fiocruz.br/genoma/.

[16] Genopar - genoma do paraná, May 2004. http://www.genopar.org/.

[17] P. Green. Phrap Homepage: phred, phrap, consed, swat, cross_match and RepeatMasker Documentation, March 2004. http://www.phrap.org.

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

[19] LGE - *Crinipellis perniciosa* - projeto vassoura de bruxa, September 2004. http://www.lge.ibi.unicamp.br/vassoura/.

[20] U. Manber. *Introduction to Algorithms*. Addison-Wesley, 1989.

[21] Ministério da Ciência e Tecnologia, July 2004. http://www.mct.gov.br.

[22] NCBI - National Center for Biotechnology Information, May 2004. http://www.ncbi.nlm.nih.gov/.

[23] NCBI Taxonomy Homepage – The Genetic Codes, July 2004. http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c.

[24] Projeto genoma Pb, May 2004. https://www.biomol.unb.br/Pb/.

[25] 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.

[26] J. P. Piazza and J. C. Setubal. New ways for automatic detection of contaminants in EST projects. In S. Lifschitz, editor, *Proceedings of Workshop of Bioinformatics (WOB'2003)*, Macaé - RJ, Brazil, December 2003.

[27] J. P. Piazza and J. C. Setubal. EST contaminant detection by combination of multiple classifiers. January 2004.

[28] Progene - Programa Genoma Nordeste, May 2004. http://www.progene.ufpe.br/index.jsp.

[29] Rede Sul de Análise de Genomas e Biologia Estrutural - Programas de Investigação de Genomas Sul, May 2004. http://www.sct.rs.gov.br/index.htm.

[30] Rede da Amazônia Legal de Pesquisas Genômicas – REALGENE, July 2004. https://www.biomol.unb.br/GR/body.html.

[31] RIOGENE - Virtual Institute of Genomic Research, May 2004. http://www.riogene.lncc.br/.

[32] 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.

[33] T. E. Scheetz, N. Trivedi, C. A. Roberts, T. Kucaba, B. Berger, N. L. Robinson, C. L. Birkett, A. J. Gavin, B. O'Leary, T. A. Braun, M. F. Bonaldo, H. P. Robinson, V. C. Sheffield, M. B. Soares, and T. L. Casavant. ESTprep: preprocessing cDNA sequence. *Bioinformatics*, 19(11):1318–1324, November 2003.

[34] *Schistosoma mansoni* EST Genome Project, July 2004. http://verjo18.iq.usp.br/schisto.

[35] A. J. G. Simpson, F.C. Reinach, P. Arruda, F. A. Abreu, M. Acencio, R. Alvarenga, L. M. C. Alves, J. E. Araya, G. S. Baia, C. S. Baptista, M. H. Barros, E. D. Bonaccorsi, S. Bordin, J. M. Bove, M. R. S. Briones, M. R. P. Bueno, A. A. Camargo, L. E. A. Camargo, D. M. Carraro, H. Carrer, N. B. Colauto, C. Colombo, F. F. Costa, M. C. R. Costa, C. M. Costa-Neto, L. L. Coutinho, M. Cristofani, E. Dias-Neto, C. Docena, H. El-Dorry, A. P. Facincani, A. J. S. Ferreira, V. C. A. Ferreira, J. A. Ferro, J. S. Fraga, S. C. França, M. C. Franco, M. Frohme, L. R. Furlan, M. Garnier, G. H. Goldman, M. H. S. Goldman, S. L. Gomes, A. Gruber, P. L. Ho, J. D. Hoheisel, M. L. Junqueira, E. L. Kemper, J. P. Kitajima, J. E. Krieger, E. E. Kuramae, F. Laigret, M. R. Lambais, L. C. C. Leite, E. G. M. Lemos, M. V. F. Lemos, S. A. Lopes, C. R. Lopes, J. A. Machado, M. A. Machado, A. M. B. N. Madeira, H. M. F. Madeira, C. L. Marino, M. V. Marques, E. A. L. Martins, E. M. F. Martins, A. Y. Matsukuma, C. F. M. Menck, E. C. Miracca, C. Y. Miyaki, C. B. Monteiro-Vitorello, D. H. Moon, M. A. Nagai, A. L. T. O. Nascimento, L. E. S. Netto, A. Nhani, F. G. Nobrega, L. R. Nunes, M. A. Oliveira, M. C. De Oliveira, R. C. De Oliveira, D. A. Palmieri, A. Paris, B. R. Peixoto, G. A. G. Pereira, H. A. Pereira, J. B. Pesquero, R. B. Quaggio, P. G. Roberto, V. Rodrigues, A. J. De M. Rosa, V. E. De Rosa, R. G. De Sá, R. V. Santelli, H. E. Sawasaki, A. C. R. Da Silva, A. M. Da Silva, F. R. Da Silva, W. A. Silva, J. F. Da Silveira, M. L. Z. Silvestri, W. J. Siqueira, A. A. De Souza, A. P. De Souza, M. F. Terenzi, D. Truffi, S. M. Tsai, M. H. Tsuhako, H. Vallada, M. A. Van Sluys, S. Verjovski-Almeida, A. L. Vettore, M. A. Zago, M. Zatz, J. Meidanis, and J. C. Setubal. The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature*, 406(6792):151–159, July 2000.

[36] R. Sorek and H. M. Safer. A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Research*, 31(3):1067–1074, 2003.

[37] The Sugar Cane EST Genome Project, March 2004. http://www.sucest.lad.ic.unicamp.br.

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

[39] 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.

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

[41] The Human Cancer Genome Project, September 2002. http://www.ludwig.org.br/ORESTES.

[42] *Trypanosoma cruzi*, May 2004. http://www.dbbm.fiocruz.br/TcruziDB/index.html.

[43] TraceTuner. http://www.paracel.com/sas/tt.htm.

[44] N. Trivedi, J. Bischof, S. Davis, K. Pedretti, T. E. Scheetz, T. A. Braun, C. A. Roberts, N. L. Robinson, V. C. Sheffield, M. B. Soares, and T. L. Casavant. Parallel Creation of Non-redundant Gene Indices from Partial mRNA Transcripts. *Future Generation Computer Systems*, 18(6):863–870, 2002.

[45] A. L. Vettore, F. R. da Silva, E. L. Kemper, G. M. Souza, A. M. da Silva, M. I. T. Ferro, F. Henrique-Silva, A. Giglioti, M. V. F. Lemos, L. L. Coutinho, M. P. Nobrega, H. Carrer, S. C. Fran, M. Bacci Jr., M. H. S. Goldman, S. L. Gomes, L. R. Nunes, L. E. A. Camargo, W. J. Siqueira, M. A. V. Sluys, O. H. Thiemann, E. E. Kuramae, R. V. Santelli, C. L. Marino, M. L. P. N. Targon, J. A. Ferro, H. C. S. Silveira, D. C. Marini, E. G. M. Lemos, C. B. Monteiro-Vitorello, J. H. M. Tambor, D. M. Carraro, P. G. Roberto, V. G. Martins, G. H. Goldman, R. C. de Oliveira, D. Truffi, C. A. Colombo, M. Rossi, P. G. de Araujo, S. A. Sculaccio, A. Angella, M. M. A. Lima, V. E. de Rosa Jr., F. Siviero, V. E. Coscrato, M. A. Machado, L. Grivet, S. M. Z. Di Mauro, F. G. Nobrega, C. F.M.Menck, M. D. V. Braga, G. P. Telles, F. A. A. Cara, G. Pedrosa, J. Meidanis, and P. Arruda. Analysis and functional annotation of an expressed sequence tag collection for the tropical crop sugarcane. *Genome Research*, 13:2725–2735, 2003. Submitted: 12/May/2003. Acepted: 08/September/2003.

[46] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[47] O. White, T. Dunning, G. Sutton, M. Adams, J. C. Venter, and C. Fields. A quality control algorithm for DNA sequencing projects. *Nucleic Acids Research*, 21:3829–3838, 1993.

[48] S. R. Woodward, N. J. Weyand, and M. Bunnel. DNA sequences from cretaceous period bone fragments. *Science*, 266:1229–1232, 1994.

[49] *Xanthomonas axonopodis pv. citri* and *Xanthomonas campestris pv. campestris* Genomes Project, September 2004. http://cancer.lbi.ic.unicamp.br/xanthomonas/.

[50] *Xylella fastidiosa* Genome Project, May 2004. http://aeg.lbi.ic.unicamp.br/xf/.

[51] H. Zischler, M. Hoss, O. Handt, A. von Haeseler, A. C. van der Kuyl, J. Goudsmit, and S. Paabo. Detecting dinosaur DNA. *Science*, 268:1191–1193, 1995.