Bioinformática de Projetos Genoma de Bactérias

Vagner Katsumi Okura

Dissertação de Mestrado

Instituto de Computação
Universidade Estadual de Campinas

Bioinformática de Projetos Genoma de Bactérias

Vagner Katsumi Okura [1]

Fevereiro de 2002

Banca Examinadora:

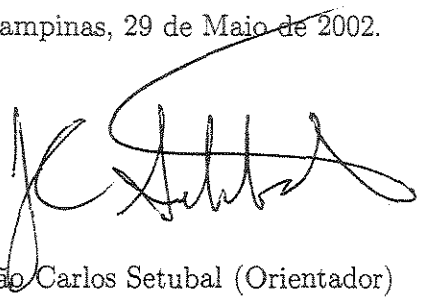
- João Carlos Setubal (Orientador)
- João Paulo F. W. Kitajima
  CBMEG - UNICAMP
- Paulo Arruda
  CBMEG - UNICAMP
- João Meidanis (Suplente)
  IC - UNICAMP

[1]**Suporte parcial da CAPES e FAPESP.**

Bioinformática de Projetos Genoma de Bactérias

Este exemplar corresponde à redação final da Dissertação devidamente corrigida e defendida por Vagner Katsumi Okura e aprovada pela Banca Examinadora.

Campinas, 29 de Maio de 2002.



João Carlos Setubal (Orientador)

Dissertação apresentada ao Instituto de Computação, UNICAMP, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP

Okura, Vagner Katsumi

Ok7b Bioinformática de projetos Genoma de bactérias / Vagner Katsumi Okura – Campinas, [S.P. :s.n.], 2002.

Orientador : João Carlos Setubal

Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Computação.

1. DNA. 2. Biologia molecular. 3. Genoma. 4. Sistemas de recuperação da informação – Sequência de nucleotideo. 5. Sequência de nucleotideo – Processamento de dados. I. Setubal, João Carlos. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Dedicatória

A minha querida família - meu pai Mikio, minhas irmãs Elaine e Gislene, meu irmão Márcio, e ao meu adorável sobrinho Luan - e em memória de minha saudosa mãe Toshie.

Agradecimentos

Agradeço ao meu orientador João Carlos Setubal pela compreensão, incentivo e auxílio nesse trabalho.

Agradeço aos meus amigos biólogos André Vettore, Edson Luis Kemper e Felipe Rodrigues da Silva pelo auxílio com explicações, correções e figuras do capítulo sobre conceitos de biologia molecular e seqüenciamento de DNA.

Agradeço aos meus colegas de trabalho, que fizeram parte da equipe que desenvolveu a bioinformática dos projetos: Prof. Dr. João Meidanis, Prof. Dr. João Paulo W. Kitajima, Marcos Renato R. Araujo, Nalvo Franco de Almeida Jr., Guilherme Pimentel Telles, Zanoni Dias, Lin Tzy Li, Marilia Dias Vieira Braga e Renato Fonseca Furquim Werneck.

Agradeço à família Oscar, Lúcia, Juliana e Jaqueline pelo acolhimento em sua casa no início dessa jornada, pelo carinho e pela amizade.

Agradeço aos meus amigos pelo apoio.

Agradeço à Capes, pela concessão da bolsa de mestrado no período de Maio de 1996 à Março de 1997. À Fapesp pela bolsa de treinamento técnico III, concedida no período de Novembro de 1998 à Abril de 1999. Ao Fundecitrus, do qual recebi salário para trabalhar no projeto *Xanthomonas* no período de Abril de 2000 à Setembro de 2001.

Agradeço, enfim a Deus pela vida, pela saúde e pela graça de chegar até aqui, e por tantas outras etapas vencidas.

Resumo

Este trabalho apresenta a bioinformática desenvolvida e usada nos projetos genoma *Xylella fastidiosa* e *Xanthomonas*. O objetivo geral da bioinformática nesses projetos é armazenar, organizar, analisar e disponibilizar os dados biológicos oriundos dos laboratórios de seqüenciamento. Em particular, são apresentados dois sistemas de software usados para a montagem e para a anotação de genomas de bactérias. A dissertação contém também um capítulo detalhando fundamentos de biologia molecular e de técnicas de seqüenciamento de DNA.

Conteúdo

Dedicatória vi

Agradecimentos vii

Resumo viii

1 Introdução 1

2 Conceitos Básicos de Biologia Molecular e de Seqüenciamento de DNA 3
- 2.1 Células . . . 3
- 2.2 Gene, Cromossomo e Genoma . . . 4
- 2.3 Ácidos Nucleicos . . . 5
  - 2.3.1 Síntese de Ácidos Nucleicos . . . 8
- 2.4 Proteínas . . . 9
- 2.5 Expressão de Genes . . . 10
  - 2.5.1 Transcrição . . . 10
  - 2.5.2 Tradução: Síntese de Proteínas . . . 12
- 2.6 Seqüenciamento de DNA . . . 13
  - 2.6.1 Fragmentação . . . 14
  - 2.6.2 Amplificação . . . 15
    - 2.6.2.1 Clonagem de DNA . . . 15
    - 2.6.2.2 Detalhes da Obtenção de Clones de DNA . . . 18
    - 2.6.2.3 PCR . . . 20
  - 2.6.3 Leitura de DNA . . . 21
  - 2.6.4 Visão Global de Seqüenciamento . . . 23

3 Projetos Genoma 28
- 3.1 Projetos de Seqüenciamento de Genomas Completos . . . 28
  - 3.1.1 Seqüenciamento . . . 29
  - 3.1.2 Montagem . . . 31

3.1.3 Anotação 31
3.2 Projetos de Seqüenciamento de EST 32
3.3 Definição do Papel da Bioinformática para Projetos Genoma 33
3.4 Projetos Genoma de Procariotos 33
3.4.1 O Projeto Genoma *Xylella fastidiosa* 34
3.4.2 O Projeto Genoma *Xanthomonas* 36
3.4.3 O Projeto Genoma *Xylella fastidiosa Pierce's Disease* 37
3.4.4 O Projeto Genoma *Agrobacterium tumefaciens* 37
3.5 LBI 38

4 Bioinformática de Base **39**
4.1 Montagem 39
4.2 Ferramentas 42
4.2.1 O Software phred/phrap/consed 43
4.3 Pipeline de Montagem do LBI 48
4.3.1 Organização e Caracterização das Seqüências 48
4.3.2 Submissão de Cromatogramas 49
4.3.3 Submissão de Seqüências Montadas 53
4.3.4 Montagem Geral 55

5 Anotação **59**
5.1 Objetivos e Descrição dos Problemas 59
5.1.1 Anotação de Genes 60
5.1.2 Anotação de Vias Metabólicas 61
5.2 Ferramentas 61
5.2.1 BLAST 61
5.2.2 Glimmer 63
5.2.3 tRNAscan-SE 65
5.2.4 Bancos de Seqüências Públicos 65
5.3 Pipeline de Anotação do LBI 66
5.3.1 O Banco de dados de Anotação 67
5.3.2 Interação Montagem e Anotação 69
5.3.3 Anotação Automática 71
5.3.4 Interface Web 72
5.3.5 Outras Ferramentas de Anotação 77

6 Outras Ferramentas **80**
6.1 Digestão Eletrônica 80
6.2 PCR Eletrônico 82

6.3 Miscelânea 82

7 Conclusão 86

A Tabelas Adicionais 89

B Relatórios e Sumários 91
B.1 Relatório de Submissão de Reads 91
B.2 Relatório de Submissão de Seqüências Montadas 93
B.3 Sumário de Cosmídeos 94
B.4 Sumário de Contigs 95

C Endereços WEB 97

Bibliografia 98

Lista de Tabelas

A.1 Código genético: mapeamento de codons para aminoácidos. 90

Lista de Figuras

2.1 Esquema do nucleotídeo. O nucleotídeo possui três partes: fosfato (F), ribose e uma base. O fosfato liga-se ao carbono número 5 da pentose, enquanto que a base liga-se ao carbono número 1. 6

2.2 Representação linear da hélice dupla. As fitas possuem orientação inversa e ligam-se através de suas bases complementares: A com T e C com G. . . 7

2.3 Estrutura química geral dos aminoácidos. 9

2.4 Estrutura de um gene de eucarioto. Genes de procarioto (em geral) não possuem *introns*. 11

2.5 Esboço de autoradiograma de um gel. Os traços horizontais indicam a presença de DNA. As linhas verticais pontilhadas representam as canaletas onde são colocados os fragmentos. 15

2.6 Inserto e vetor ligados pelo mesmo sítio de restrição. 19

2.7 Ilustração de uma placa contendo colônias de bactérias. Os pontos brancos referem-se às colônias com insertos e os pontos sombreados referem-se às colônias sem insertos. 20

2.8 Esboço de uma imagem de gel de seqüenciamento de DNA, mostrando a seqüência de bases obtida, que no caso é lida de baixo para cima, como mostrado à esquerda da figura. 22

2.9 Figura de um arquivo de gel obtido com seqüenciamento automático de DNA. Na vertical estão as canaletas. As cores representam cada uma das bases: verde como `A`, azul como `C`, amarelo como `G`, vermelho como `T`. À esquerda, está a visualização do cromatograma da seqüência correspondente a canaleta 77. 24

2.10 Diagrama ilustrando o processo global de seqüenciamento. 25

3.1 Exemplo de estratégia de seqüenciamento envolvendo fragmentação e sub-fragmentação. Primeiramente são gerados fragmentos grandes da molécula alvo. A seguir, cada um desses fragmentos grandes é sub-fragmentado, gerando fragmentos de comprimento intermediário (ou médio), ainda longos demais para serem seqüenciados. Assim, é feito mais um passo, onde cada fragmento intermediário é sub-fragmentado, de onde são gerados fragmentos pequenos e possíveis de serem seqüenciados. 30
3.2 Seqüenciamento de um inserto em cosmídeo, combinando a abordagem de shotgun e a baseada em pontas. A linhas mais espessas representam os segmentos seqüenciados. Os números são a identificação dos clones e as letras a e b referem-se qual a ponta do clone seqüenciada. 31
3.3 Organização da rede ONSA, tendo LBI como nó central. 36

4.1 Exemplo de uma região sem cobertura. Linha pontilhada representa trecho não coberto. 42
4.2 Repetições em uma molécula de DNA levando a montagens ambíguas. . . . 42
4.3 Exemplo de repetição colapsada, levando à uma solução com dois contigs, ao invés da solução correta de um único contig. 43
4.4 Janela do Consed mostrando uma montagem. 47
4.5 Visualisação parcial de um cromatograma no Consed. 47
4.6 Principais processos e bancos de seqüências do pipeline de montagem. As elipses representam os processos, os retângulos abertos representam os repositório de dados, o retângulo fechado representa uma entidade externa ao sistema e as linhas orientadas indicam o fluxo dos dados. 50
4.7 Versão gráfica de um relatório de submissão de reads. Nesse relatório podem ser visualizados as regiões de baixa (< 20) e alta (≥ 20) qualidade, e as regiões com e sem vetor. 52
4.8 Corte do mapa físico construído a partir da montagem. 58

5.1 Saída gráfica do BLAST. 64
5.2 Pipeline da anotação do Projeto *Xanthomonas*. As elipses representam os processos, os retângulos abertos representam os repositórios de dados, os retângulos fechados representam entidades externas ao sistema e as linhas orientadas indicam o fluxo dos dados. 68
5.3 Esboço do esquema do banco de anotação. Uma linha simples indica a relação um para um. Uma linha com orientação única indica a relação um para muitos. 69
5.4 Interface do editor de genes. 74

5.5 Interface do visualizador de *chunks*. No quadro azul, as quatro barras superiores delimitadas por pequenos traços verticais correspondem aos segmentos e sub-segmentos escolhidos para visualização. As seis barras inferiores de cor vermelha representam os seis quadros de leitura, junto com os codons de início (traços azuis) e codons de parada (traços pretos). Abaixo de cada quadro de leitura são mostrados os respectivos genes. No quadro laranja, são escolhidos os genes a serem removidos. 76
5.6 Inferface do modificador de codon de início. 78
5.7 Mapa de genes mostrado parcialmente. 79

6.1 Mapa de restrição produzido pela digestão eletrônica. 81
6.2 Resultado de uma pesquisa usando o PCR eletrônico. 83

Capítulo 1

Introdução

Este trabalho apresenta a bioinformática para projetos genoma de bactérias, baseado nos projetos *Xylella fastidiosa* [33] e *Xanthomonas* [10]. Em linhas gerais, a bioinformática a ser descrita refere-se aos sistemas de software desenvolvidos nesses projetos para armazenar, organizar, analisar e disponibilizar os dados biológicos. Como contribuições desta dissertação temos:

- apresentação detalhada e sistematizada do trabalho de bioinformática realizado nesses projetos;
- descrição de alguns programas específicos realizados pelo autor;
- um capítulo sobre biologia molecular, útil para pessoas da área de computação.

São apresentados neste trabalho os dois principais sistemas desenvolvidos e usados nos projetos acima mencionados: o sistema de montagem e o sistema de anotação. Vale notar que esses sistemas também foram utilizados nos projetos genoma da bactéria *Xylella fastidiosa* [30] que ataca videiras e da bactéria *Agrobacterium tumefaciens* [40]. Como resultados importantes, tivemos participação nos trabalhos científicos desses dois projetos.

Os sistemas foram desenvolvidos por uma equipe, da qual faço parte, que pertence ao Laboratório de Bioinformática (LBI) do Instituto de Computação da Unicamp. No caso de programas que foram escritos por outros ou em colaboração com outros, uma nota de rodapé indica o nome dos colaboradores.

Este trabalho está inserido na área de Biologia Computacional e é direcionado primariamente aos profissionais de computação. Assim, no Capítulo 2, apresentamos algumas noções básicas de Biologia Molecular e de tecnologia de seqüenciamento de DNA, necessárias para um melhor entendimento da dissertação.

No Capítulo 3, apresentamos de uma maneira geral o que são projetos genoma, como são suas abordagens e quais são suas etapas básicas. Além disso, apresentamos uma

definição sucinta do papel da bioinformática dentro dos projetos genoma, uma breve descrição dos projetos genoma onde foram aplicados os sistemas de montagem e anotação, além de uma descrição também sumária do LBI.

Nos Capítulo 4 e 5, apresentamos, respectivamente, o sistema de montagem e o sistema de anotação desenvolvidos pelo LBI, descrevendo primeiramente os objetivos e os problemas motivadores do desenvolvimento desses sistemas, e algumas das ferramentas desenvolvidas fora do LBI e usadas pelos sistemas.

No Capítulo 6, apresentamos duas ferramentas específicas desenvolvidas e usadas como apoio ao sistema de montagem, e uma miscelânea de outros programas desenvolvidos durante os projetos.

No Capítulo 7, apresentamos as conclusões dessa dissertação, descrevendo as principais contribuições e os trabalhos futuros relacionados principalmente a melhoria dos sistemas.

Por último, nos apêndice A e B são apresentados algumas tabelas e relatórios referenciados na dissertação.

Capítulo 2

Conceitos Básicos de Biologia Molecular e de Seqüenciamento de DNA

Apresentamos neste capítulo alguns conceitos básicos de Biologia Molecular e de seqüenciamento de DNA, começando com uma breve descrição das células, onde são encontradas as principais moléculas de estudo - os ácidos nucleicos e as proteínas. Em seguida, são apresentados alguns conceitos que estão ligados com a área de Genética - genes, cromossomo e genoma, mostrando como eles se relacionam. Fechando então a parte sobre Biologia Molecular, são descritos os ácidos nucleicos e as proteínas. Como parte integrante e que vem trazendo grandes benefícios aos estudos em biologia molecular, apresentamos a tecnologia de seqüenciamento de DNA, descrevendo as principais etapas envolvidas nesse processo. Este capítulo foi escrito com base nas seguintes referências da literatura de biologia, bioquímica e genética: [24, 39, 41, 36, 23, 27].

2.1 Células

As células são consideradas as unidades fundamentais dos organismos. Nelas realizam-se a maior parte dos processos metabólicos dos seres vivos e nelas está contido o material genético.

As células são constituídas por uma membrana envolvente, que separa sua massa interior (chamada de citoplasma) do ambiente onde está inserida. Apesar de ser considerada impermeável (não permite que o meio externo misture-se ao meio interno) a membrana celular contém alguns poros por onde determinadas substâncias entram e saem da célula, mantendo-se assim uma interação com o ambiente, de onde recebe os nutrientes para desenvolver-se e para onde expele substâncias.

A organização interna de uma célula divide o mundo dos seres vivos em dois domínios de organismos: eucariotos e procariotos. Nas células dos eucariotos são encontrados compartimentos internos bem definidos como o núcleo e organelas. Exemplo de organelas são a mitocôndria, encontrada em células de animais, e o cloroplasto, encontrado em células de plantas. Já as células de procariotos não possuem compartimentos internos e em particular não possuem núcleo. Nas células de organismos eucariotos o material genético reside no núcleo, e nas células dos procariotos o material genético permanece livre no citoplasma. Nos eucariotos, além da informação genética contida no núcleo, organelas como as mitocôndrias e cloroplastos têm seu próprio material genético.

A propriedade fundamental de uma célula está na sua capacidade de crescer e replicar-se, gerando células descendentes contendo cópias do seu material genético. Isso é resultado de uma série de processos metabólicos desencadeados dentro da célula. Parte desses fenômenos químicos estão ligados a fabricação (síntese) de moléculas, como os aminoácidos, as proteínas e os ácidos nucleicos. Embora haja muitos elementos envolvidos nesse processo de síntese, eles podem ser agrupados nas seguintes componentes:

1. matéria-prima: elementos que serão usados na constituição das moléculas. Ex.: carbono, oxigênio, hidrogênio, nitrogênio e também outras pequenas moléculas.

2. energia externa: necessária para realização das reações químicas, e pode ser proveniente da decomposição de moléculas de alimento (células de animais), ou da energia solar (células fotossintéticas em plantas).

3. enzimas: agentes que realizam as reações químicas, também denominadas de catalisadores.

As moléculas contidas e fabricadas nas células podem ser moléculas simples como os açúcares e os aminoácidos, ou podem ser moléculas mais complexas, chamadas macromoléculas. As macromoléculas são polímeros, formados pelo encadeamento de várias moléculas simples e semelhantes (chamadas monômeros). Exemplos de macromoléculas são os ácidos nucleicos e as proteínas. Os primeiros são compostos por unidades chamadas nucleotídeos e as últimas são compostas por unidades chamadas aminoácidos.

2.2 Gene, Cromossomo e Genoma

Considerado um dos marcos históricos da genética, os experimentos do austríaco Gregor Mendel o levaram a identificar fatores responsáveis pelos traços hereditários dos organismos vivos, os quais foram chamados **genes**. Passado algum tempo, descobriu-se que havia estruturas dentro da célula chamadas **cromossomos**, que duplicavam-se durante

a divisão da célula. Associando então esses dois fatos lançou-se a hipótese, confirmada posteriormente, de que a herança genética é carregada pelos cromossomos arranjada nos genes. Restava então saber qual era a composição dos cromossomos, reconhecidos posteriormente como sendo moléculas de DNA. Assim, sob a ótica molecular, os **genes** podem então ser considerados como segmentos contíguos e discretos de uma molécula de DNA onde estão armazenadas as informações genéticas.

A informação genética contida nas moléculas de DNA na célula de um ser vivo compreende o seu **genoma**.

Os genomas dos procariotos e eucariotos possuem algumas diferenças importantes. Uma das diferenças está relacionada ao número de cromossomos presentes em cada célula. Uma célula de procarioto em geral possui apenas um cromossomo, enquanto que em uma célula eucariótica esse número pode ser por exemplo, 23 pares de cromossomos (46 cromossomos) para uma célula humana.

Outra diferença se refere à organização dos genes entre os genomas desses dois domínios de organismos. Além dos genes, os cromossomos possuem regiões intergênicas, que não possuem nenhuma função particularmente conhecida e são chamadas "lixo de DNA" (*junk DNA*). Nos procariotos os genes compreendem quase todo o cromossomo, havendo poucas e em geral curtas regiões intergênicas. Em números, isso poderia ser aproximadamente colocado na proporção de 90% de genes e 10% de regiões intergênicas. No caso dos eucariotos (principalmente os superiores) o volume de genes é bem menor em relação às regiões intergênicas, numa proporção contrária a dos procariotos. No caso de um humano, menos que 5% do genoma é compreendido pelos genes.

2.3 Ácidos Nucleicos

Ácidos nucleicos são moléculas que armazenam as informações relativas ao desenvolvimento e divisão das células, as quais formam os organismos vivos. Na natureza há dois tipos de ácidos nucleicos: **DNA** ou ácido desoxiribonucleico e **RNA** ou ácido ribonucleico. Analogamente a um sistema de comunicação, essas informações são mantidas dentro da célula em forma de código, que no caso denomina-se **código genético**.

Em sua estrutura primária, os ácidos nucleicos podem ser vistos como uma cadeia linear composta de unidades químicas simples chamadas **nucleotídeos**. Um nucleotídeo é um composto químico e possui três partes: um grupo fosfato, uma pentose (molécula de açúcar com cinco carbonos) e uma **base** orgânica (Figura 2.1).

Nas moléculas de DNA a pentose é uma desoxiribose enquanto que nas moléculas de RNA a pentose é uma ribose. A base orgânica, também conhecida como base nitrogenada, é quem caracteriza cada um dos nucleotídeos, sendo comum o uso tanto do termo seqüência de nucleotídeos quanto o termo seqüência de bases. As bases são adenina (A), guanina

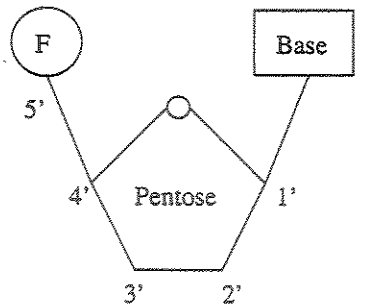


Figura 2.1: Esquema do nucleotídeo. O nucleotídeo possui três partes: fosfato (F), ribose e uma base. O fosfato liga-se ao carbono número 5 da pentose, enquanto que a base liga-se ao carbono número 1.

(G), citosina (C), timina (T) e uracila (U), sendo as duas primeiras chamadas de purinas e as três últimas chamadas de pirimidinas. No DNA são encontradas as bases A, G, C e T. No RNA encontra-se a base U ao invés da base T.

A ligação entre os nucleotídeos na cadeia linear é feita entre o grupo químico chamado hidroxil (OH), ligado ao terceiro carbono da pentose de um nucleotídeo, e o fosfato do nucleotídeo seguinte. Dessa série de ligações fosfato-pentose forma-se então a coluna dorsal (*backbone*) dos ácidos nucleicos. Essa cadeia, também chamada de **fita**, possui uma orientação química denotada por uma extremidade 3′, onde há um grupo hidroxil livre ligado ao carbono número 3 da pentose, e por uma extremidade 5′, que possui um fosfato livre ligado ao carbono número 5 da pentose (Figura 2.2). Por convenção, as seqüências são representadas na orientação $5' \rightarrow 3'$.

Moléculas de DNA compõem-se de duas fitas, que ligam-se entre si formando uma estrutura helicoidal, conhecida como **hélice dupla**. As duas fitas unem-se pela ligação regular das bases de seus nucleotídeos. A base A sempre liga-se a base T e a base G sempre liga-se a base C (Figura 2.2). Esse pareamento direcional, exclusivo e único entre as bases (ditas bases complementares) deve-se em geral ao seu tamanho, a sua forma e a sua composição química. As duas fitas são **anti-paralelas**, ou seja, as fitas possuem orientação $5' \rightarrow 3'$ opostas uma em relação a outra.

Utiliza-se como unidade de medida de comprimento de uma molécula de DNA o número de **pares de bases** que a formam, denotado por bp (*base pair*). Por exemplo, a molécula A tem comprimento de 800 bp. Cifras maiores em geral são representadas por kb (1000 bp) e mb (10^6 bp).

As moléculas de DNA podem ser circulares (não possuem extremidades livres) ou podem ser lineares. Bactérias são exemplos de organimos que contém moléculas de DNA circular.

Moléculas de RNA têm fita simples e são bem curtas se comparadas às moléculas de DNA. Existem três tipos distintos de moléculas de RNA nas células: RNA mensageiro

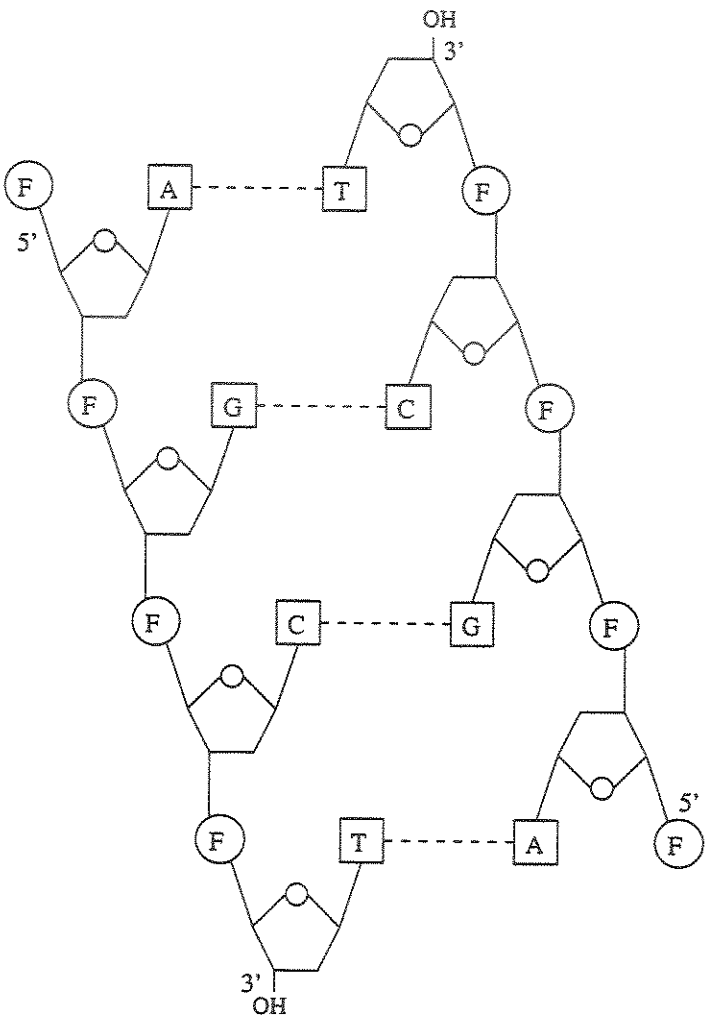


Figura 2.2: Representação linear da hélice dupla. As fitas possuem orientação inversa e ligam-se através de suas bases complementares: A com T e C com G.

(**mRNA**), RNA transportador (**tRNA**) e RNA ribossomal (**rRNA**). A descrição das moléculas de RNA está na Seção 2.5.1.

2.3.1 Síntese de Ácidos Nucleicos

A síntese de uma cadeia de DNA ou RNA nas células, também conhecido como polimerização de ácidos nucleicos, se faz a partir de uma fita simples de DNA molde e da regra de pareamento de bases complementares, de onde uma fita nova é gerada a partir do encadeamento dos nucleotídeos com base complementar aos nucleotídeos da fita molde.

Os principais agentes envolvidos nesse processo são enzimas especiais chamadas polimerases. Enzimas que fazem a síntese de DNA são chamadas de polimerases de DNA e enzimas que fazem a síntese de RNA são chamadas de polimerases de RNA. Essas enzimas encarregam-se de adicionar um a um os nucleotídos à fita nova, bastando que haja na sua extremidade $3'$ um grupo hidroxil (OH) livre, onde é ligado o próximo nucleotídeo pelo seu grupo fosfato. É importante, então, notar que a síntese de uma fita nova de um ácido nucleico acontece na direção $5' \rightarrow 3'$.

O início da ação de uma polimerase de DNA requer que haja um fragmento curto de DNA ou RNA em fita simples ligado à fita molde, onde a enzima possa concatenar os nucleotídeos. Esse fragmento, chamado de **primer**, é sintetizado por uma outra enzima. Já as polimerases de RNA iniciam a polimerização ao encontrar o sítio iniciador apropriado na fita de DNA.

Dentro da célula, há dois momentos importantes quando acontece a síntese de ácidos nucleicos: na replicação das moléculas de DNA durante a duplicação da célula e na transcrição.

No primeiro caso, cada uma das fitas de uma molécula de DNA é usada individualmente como molde, e cada uma das duas fitas duplas resultantes é composta por uma fita molde e por uma fita nova sintetizada. Nesse caso, o início da síntese das fitas se faz a partir de um ponto na molécula chamado origem de replicação, onde começa a separação das fitas.

Na transcrição, uma fita de RNA é sintetizada a partir de uma fita de DNA molde. Mais detalhes sobre transcrição estão na Seção 2.5.1.

Outro tipo de síntese de ácidos nucleicos conhecido é chamado de **transcrição reversa**, da qual se produz moléculas de DNA a partir de uma molécula de RNA. Moléculas assim produzidas são chamadas de **cDNA** ou DNA complementar.

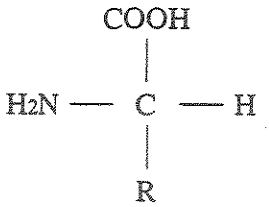


Figura 2.3: Estrutura química geral dos aminoácidos.

2.4 Proteínas

Proteínas são macromoléculas que têm funções específicas dentro de um organismo. Elas podem ter um cárater estrutural (como a keratina presente nos cabelos e o colágeno presente nos tendões e cartilagens), ou podem estar ligadas a determinadas atividades, como é caso dos anticorpos (como imunoglobina), de hormônios (como a insulina), e das enzimas (como àquelas encontradas no estômago e relacionadas a digestão de alimentos).

Uma proteína é formada por unidades conhecidas como **aminoácidos**. Essas unidades ligam-se linearmente resultando uma cadeia conhecida como **polipeptideo**. Um aminoácido é composto por um carborno central (C_α), um hidrogênio (H), um grupo amino (H_2N), um grupo carboxil (COOH) e uma cadeia lateral (R) (Figura 2.3). A cadeia lateral distingue cada um dos 20 aminoácidos diferentes que existem na natureza. Os aminoácidos ligam-se através de ligações peptídicas, formadas pela junção do grupo carboxil do primeiro aminoácido com grupo amino do segundo, havendo a liberação de uma molécula de água (H_2O). A cadeia resultante compõem-se então de **resíduos** dos aminoácidos.

A seqüência linear de aminoácidos representa a estrutura primária das proteínas. Essas moléculas dobram-se e empacotam-se até um quarto nível, formando diferentes formas tridimensionais, que estão diretamente associadas a função bioquímica da proteína. As dobras se fazem em ângulos variados, decorrentes das ligações peptídicas. A existência de 20 diferentes aminoácidos propicia uma grande e complexa variedade de formas irregulares, que determinam a sua ligação com outras moléculas. Cada proteína, com a sua forma, liga-se a tipos específicos de moléculas, ou seja àquelas que possuem uma forma complementar à da proteína. Uma analogia simplista (em forma dimensional) disso são as formas apresentadas pelas peças de um jogo de quebra-cabeça, onde somente peças específicas podem se encaixar.

2.5 Expressão de Genes

Como visto na Seção 2.2, um gene é um segmento de DNA que carrega informação genética. Essa informação torna-se disponível para célula pela expressão gênica. Quando isso acontece, uma cópia do gene é sintetizada em uma molécula de mRNA (transcrição), que por sua vez é usada na fabricação de uma proteína (tradução). Esse fluxo de informação envolvendo DNA, RNA e proteína foi descrito por Francis Crick como o dogma central, e é esboçado na figura a seguir.

Um fato interessante tirado dessa dogma é que uma proteína está associada a um determinado gene. Maiores detalhes sobre transcrição e tradução são apresentados nas seções abaixo.

A expressão dos genes é altamente regulada (controlada). Com isso, em organismos multicelulares (como o ser humano), células presentes em diferentes tecidos (como estômago e pele) apresentam um conjunto de genes ativos distintos entre si. E mesmo em organimos unicelulares, como as bactérias, nem todos os genes são ou estão ativos em um dado momento.

Em sua estrutura, os genes possuem uma região chamada de promotor, que é responsável por sua ativação ou não. Um promotor é um segmento de DNA ao qual a polimerase liga-se e inicia a síntese da molécula de mRNA. Os promotores possuem seqüências de nucleotídeos comuns (conservadas), o que explica como a polimerase reconhece o lugar onde se ligar.

Além do promotor, os genes possuem em sua estrutura, uma região codificadora e um terminador, como esboçado na Figura 2.4. A região codificadora é o segmento do gene que contém a informação usada para sintetizar uma proteína e o terminador é o segmento de DNA que sinaliza o final da síntese da molécula de mRNA.

2.5.1 Transcrição

Transcrição é o processo no qual é produzida uma fita de RNA a partir de uma fita de DNA. O produto de uma transcrição pode ser um mRNA, um rRNA ou um tRNA. Uma molécula de mRNA ou RNA mensageiro (*messager RNA*) contém a informação para produzir proteínas. As moléculas de rRNA ou RNA ribossomal (*ribosomal RNA*) combinam-se com proteínas ribossomais constituindo os ribossomos, que funcionam então como um aparato estrutural para a montagem da cadeia de aminoácidos. Moléculas de tRNA ou RNA transportador (*transfer RNA*) agem como adaptadores entre a seqüência codificante

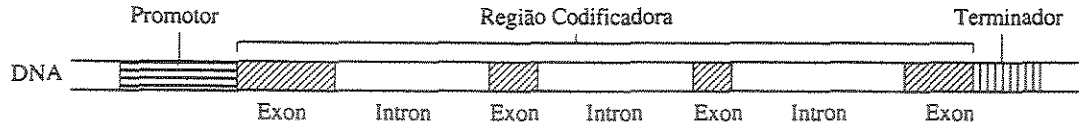


Figura 2.4: Estrutura de um gene de eucarioto. Genes de procarioto (em geral) não possuem *introns*.

dos nucleotídeos do mRNA e o aminoácido que é codificado. Uma ponta dessa molécula carrega o aminoácido e uma outra ponta consiste de uma seqüência de três nucleotídeos conhecida como **anticodon**.

O processo de transcrição dos procariotos e dos eucariotos apresentam várias diferenças entre si. Dentre essas diferenças há duas principais. A primeira é que em procariotos o processo de tradução de um mRNA inicia-se antes que a transcrição tenha se encerrado. Já nos eucariotos, a tradução é iniciada somente após a finalização da transcrição, e nesse caso o mRNA sai do núcleo para o citoplasma onde é realizada a tradução.

A segunda diferença é que nos eucariotos o mRNA sofre modificações antes de ser traduzido. O produto da transcrição, chamado de transcrito primário, sofre mudanças e resulta em um transcrito chamado maduro. As três mudanças sofridas pelo transcrito primário são: remoção de seqüências intragênicas, modificação da extremidade $5'$ e modificação da extremidade $3'$.

Nos eucariotos há segmentos gênicos que são transcritos em mRNA mas não são usados na tradução de proteínas. Esses segmentos são chamados de *introns* e são removidos do mRNA. Os segmentos dos genes entre os *introns* que são transcritos e traduzidos são chamados de *exons*. O processo de remoção de *introns* e ligação dos *exons* é chamado de *splicing*. Veja a ilustração do arranjo de *introns* e *exons* de um gene na Figura 2.4.

Após o *splicing*, o transcrito recebe na sua extremidade $3'$ uma cadeia de adeninas chamada cauda de *poly-A*, e na sua extremidade $5'$ recebe uma molécula chamada de *cap*. A cauda de *poly-A*, com comprimento entre 20 a 200 adeninas, dá estabilidade a molécula no seu transporte para fora do núcleo. O *cap* permite que o ribossomo reconheça o início do mRNA.

Além das duas grandes diferenças acima, uma diferença mais peculiar é que em eucariotos em geral, um transcrito contém apenas um gene, enquanto que em procariotos como bactérias, um transcrito pode conter mais de um gene. A esse fato está relacionado o conceito de *operon*, que refere-se a uma seqüência de genes adjacentes sob o controle transcricional do mesmo promotor.

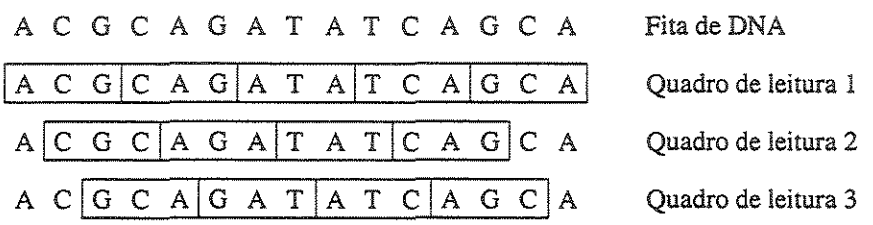
2.5.2 Tradução: Síntese de Proteínas

Tradução é o processo de síntese ou fabricação de proteínas. Para a fabricação das proteínas é necessário que a informação armazenada no DNA (gene) seja transmitida até estruturas celulares chamadas ribossomos que residem no citoplasma. Essa transmissão é realizada pelas moléculas de mRNA (RNA mensageiro) que levam em sua estrutura a mensagem a ser interpretada nos ribossomos, onde é sintetizada a cadeia de aminoácidos que constituirá a proteína.

O ponto chave da tradução está no deciframento ou decodificação da cadeia de nucleotídeos na cadeia de aminoácidos. O deciframento está baseado em triplas de nucleotídeos, chamadas **codons**, que são usados para especificar os aminoácido. A correspondência entre uma tripla de nucleotídeos e um aminoácido é chamada de código genético, e é apresentada em forma de tabela no Apêndice A.1. Combinando os 4 nucleotídeos em triplas obtém-se 64 combinações. Embora esse número seja superior aos 20 aminoácidos, mais do que um codon pode representar um mesmo aminoácido. Dentre os codons possíveis, 3 não especificam aminoácidos, e referem-se a sinais de terminação da síntese de um cadeia de aminoácidos. Esses codons são chamados de codons de parada (*stop codons*). O código genético estabelece também um codon de início (*start codon*), pelo qual começa o processo de tradução do mRNA. Na maioria das proteínas o codon de início especifica o aminoácido metionina, que também está presente no interior das cadeias.

Sumariamente, o processo de tradução é realizado da seguinte maneira: ao combinar-se com os ribossomos, o mRNA tem sua seqüência de codons lida, e para cada codon o respectivo tRNA (Seção 2.5.1) é atraído até os ribossomos, e pela complementariedade de bases é feita a ligação entre o codon (do mRNA) e o anticodon (do tRNA), liberando o aminoácido carregado pelo tRNA que é concatenado à cadeia crescente do polipeptideo. A síntese da proteína é encerrada ao ser encontrado um codon de parada.

Um outro aspecto referente a tradução é o conceito de quadro de leitura (*reading frame*). Um quadro de leitura especifica uma das três formas de agrupar nucleotídeos em codons. Veja a figura a seguir.



O primeiro quadro de leitura começa no nucleotídeo 1 da fita de DNA, o segundo no nucleotídeo 2 e o terceiro no nucleotídeo 3. Se iniciado no nucleotídeo 4, os codons resultam um sub-conjunto dos codons obtidos com início no primeiro nucleotídeo, sendo

então tratados como o mesmo quadro de leitura. Quando se trabalha com DNA é comum falar em seis quadros de leitura, sendo três obtidos em uma fita e outros três obtidos da fita complementar.

Um quadro de leitura que inicie com um codon de início e que não contenha codons de parada é denominado de **open reading frame** ou **ORF**. Embora seja comum na prática o uso dos termos ORF e gene indistintamente, é importante frisar sua diferença. Toda região codificadora de um gene é uma ORF, mas nem toda ORF é um gene. Essa última afirmativa justifica-se pelo fato de que em uma fita de DNA pode conter muitas ORFs, mas nem todas referem-se a um gene.

2.6 Seqüenciamento de DNA

A biologia molecular tem direcionado seus esforços para obter a informação básica de uma molécula de DNA, ou seja, a sua seqüência de nucleotídeos ou a sua seqüência de bases. Determinar cada uma das bases de uma molécula de DNA é o processo chamado **seqüenciamento**.

Apesar dos avanços tecnológicos em manipular moléculas de DNA, há ainda um limitante no número de bases que podem ser lidas pelos procedimentos laboratoriais. Nesses experimentos, mesmo havendo a possibilidade de se trabalhar com segmentos longos de DNA, somente as 700 primeiras bases (aproximadamente) podem ser lidas ou seqüenciadas diretamente. Por outro lado, moléculas de DNA em geral, como os cromossomos, são muito longas, variando de milhares a milhões de pares de bases.

A alternativa usada para lidar com essa diferença de grandezas consiste em fragmentar aleatoriamente a molécula de DNA, seqüenciar individualmente os fragmentos gerados e usar as seqüências obtidas na reconstituição da seqüência completa da molécula. Isso é conhecido como montagem de DNA.

Uma alternativa anterior à montagem de DNA, conhecida como **seqüenciamento direto**, tem como princípio seqüenciar um segmento a partir de uma das extremidade da molécula de DNA, usando então a seqüência obtida para seqüenciar o segmento seguinte. Executando-se iterativamente esses passos, são obtidas uma a uma as sub-seqüências que constituem a seqüência completa. Um exemplo desse tipo de seqüenciamento é conhecido como **primer walking**, no qual os segmentos são obtidos a partir da síntese (artificial) de primers obtidos da seqüência obtida anteriormente (mais datalhes sobre primer estão na Seção 2.6.2.3). A principal desvantagem dessa alternativa é que o processo é totalmente seqüencial e lento.

O processo de seqüenciamento de uma molécula de DNA envolve várias manipulações experimentais. Em particular estão envolvidas as tarefas de fragmentação, de cópia e de leitura de DNA, que são descritas individualmente a seguir. Mais adiante será apresentado

o processo global de seqüenciamento de DNA, partindo de uma molécula de DNA até chegar nas seqüências de seus fragmentos.

2.6.1 Fragmentação

Atualmente, é relativamente fácil quebrar uma molécula de DNA e separar os fragmentos obtidos de acordo com seus comprimentos. Existem basicamente dois métodos que realizam a fragmentação de uma molécula de DNA: digestão e shotgun.

No método de digestão são usadas enzimas de restrição, que agem como tesouras, cortando o DNA em segmentos específicos chamados sítios de restrição, que são reconhecidos pelas enzimas de restrição. Essas enzimas agem nas células bacterianas como defensoras, cortando (digerindo) e degradando DNAs estrangeiros ou invasores. Os sítios de restrição têm sua seqüência conhecida e são curtas (4, 6 ou 8 bp). Em grande parte das enzimas, o corte é feito de tal forma que as extremidades dos dois segmentos de DNA resultantes tenham a fita $5'$ um pouco mais curta que a fita $3'$, ou a fita $3'$ seja um pouco mais curta que a fita $5'$. Isso possibilita que fragmentos cortados com a mesma enzima sejam ligados por tais extremidades (veja um exemplo na Figura 2.6). Essas pontas são conhecidas como *sticky ends*. Há também enzimas que cortam uma molécula e deixam pontas retas (*blunt ends*), ou seja não têm a extremidade $5'$ nem a $3'$ protuberantes.

No método shotgun, uma solução contendo DNA (várias cópias de uma molécula) é submetida à elevadas taxas de vibração, fazendo com que as moléculas sejam quebradas aleatoriamente em diferentes pontos. O processo de vibração pode ser acionado com ar ou com ultra-som. O primeiro processo é chamado de nebulização e o último é chamado de sonicação.

Após a quebra, os fragmentos de DNA são separados conforme seu comprimento em um processo conhecido como **eletroforese em gel**. Nesse processo os fragmentos são colocados em uma solução gelatinosa chamada gel, e sobre a ação de um campo elétrico movem-se do polo negativo (onde são colocados inicialmente) para o polo positivo. No gel os fragmentos mais curtos movem-se mais facilmente do que os fragmentos mais longos, e após um determinado período de tempo estarão mais próximos do polo positivo do que os mais longos. Antes de serem colocados no gel, os fragmentos são rotulados com isótopos radioativos ou com corantes fluorescentes (*dyes*), para que possam ser localizados após a corrida do gel. No primeiro caso, após sua corrida, o gel é fotografado, produzindo-se um registro (autoradiograma) das localizações onde foram identificados os fragmentos de DNA. A Figura 2.5 apresenta um esboço de um gel fotografado, onde as pequenas barras horizontais indicam a presença de DNA. De maneira análoga, o gel contendo fragmentos marcados com os corantes fluorescentes é submetido à ação de uma luz ultra-violeta, sendo então destacados pelo brilho os pontos do gel onde há DNA. Uma das maneiras de

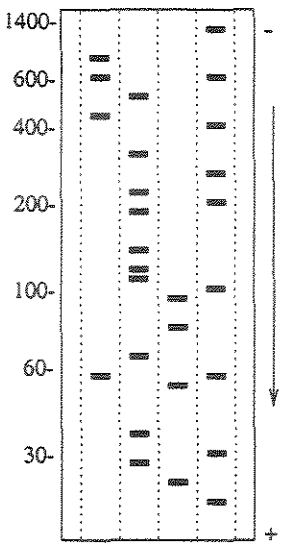


Figura 2.5: Esboço de autoradiograma de um gel. Os traços horizontais indicam a presença de DNA. As linhas verticais pontilhadas representam as canaletas onde são colocados os fragmentos.

estimar o comprimento dos fragmentos é fazer o processo em paralelo com fragmentos de comprimento conhecido. No caso, são usados canais (ou canaletas) separados para cada uma das amostras de fragmentos.

2.6.2 Amplificação

Para fazer experimentos com DNA, é necessário uma quantidade razoável do mesmo material. Isso é conseguido pelo processo chamado de amplificação de DNA, pelo qual são produzidas cópias de uma molécula ou fragmento de DNA. Para isso, existem as técnicas de clonagem e de PCR, que são descritas a seguir.

2.6.2.1 Clonagem de DNA

Uma das maneiras de amplificar um DNA, é usar a natureza para isso. Nesta abordagem, o DNA de interesse (fonte) é introduzido no DNA de uma célula hospedeira, que ao multiplicar-se, replica seu material genético para seus descendentes junto com o DNA inserido (inserto). Assim, após um determinado período de incubação, uma quantidade exponencial de células é produzida, com uma cópia do DNA fonte em cada célula. As cópias ou os clones de DNA são retirados (extraídos) das células pelo processo de purificação.

A molécula de DNA que recebe o inserto é chamada de **vetor**, e contém segmentos reconhecidos pela maquinaria de replicação da célula. O DNA resultante da combinação

entre vetor e inserto é chamado de DNA recombinante. Abaixo são descritos alguns dos principais tipos de vetores usados na clonagem de DNA.

Plasmídeos

Plasmídeos são moléculas de DNA circular autônomas encontradas em organismos como bactérias e leveduras. Seu comprimento geralmente varia entre 1 e 10 kb. No processo de divisão da célula, assim como o cromossomo, os plasmídeos são também replicados e passados para células descendentes.

Os vetores de plasmídeos têm em geral 3 kb de comprimento, e possuem uma origem de replicação e genes de resistência a antibióticos (ver Seção 2.6.4). Além disso, podem conter também um *polylinker*, ou seja, um segmento sintético composto de diferentes sítios de restrição, que possibilita uma aplicação mais ampla do vetor, no sentido de permitir, por exemplo, que fragmentos gerados por diferentes enzimas de restrição sejam clonados pelo mesmo vetor.

Embora seja possível insertos de até 15 kb em plasmídeos, na prática é comum o uso de insertos mais curtos (até 5 kb), pois vetores de plasmídeos com insertos mais longos estão menos propensos a terem sucesso na replicação.

Apesar de ser comum o uso de plasmídeos como vetores, eles apresentam uma limitação quanto ao comprimento dos insertos. Além disso, pode acontecer de determinados insertos não conseguirem sucesso na clonagem, pois pode haver genes presentes no inserto que codificam proteínas tóxicas para a bactéria hospedeira.

Bacteriófagos

Bacteriófagos são vírus que infectam bactérias. Vírus são caracterizados como parasitas, que para se reproduzirem usam a maquinaria celular. Os vírus possuem uma estrutura simples, composta em geral de uma molécula de DNA ou RNA envolvida por uma cápsula formada de proteínas. Ao infectar uma célula, um vírus pode exibir um comportamento ativo (ciclo lítico), do qual são produzidas cópias de DNA virais, que após a degradação da célula infectam outras células. Os vírus podem, por outro lado, assumir um comportamento passivo (ciclo lisogênico), ficando incubados na célula até que eventualmente assumam seu estado ativo.

Os bacteriófagos, também chamados fagos, por terem seu comportamento e estrutura molecular bem conhecidos, são usados também como vetores de clonagem. Um exemplo típico desse tipo de vetor é o fago λ. Esse fago possui uma cabeça em forma hexagonal que envolve o DNA genômico viral de aproximadamente 50 kb, e uma cauda cuja função é aderir-se à parede celular da bactéria, permitindo que o seu DNA seja introduzido na célula. O DNA viral é replicado dentro da célula e suas cópias são empacotadas para

formar novos vírus.

Para ser usado como vetor de clonagem, o DNA do fago λ foi modificado, sendo removidos os genes relacionados ao ciclo lisogênico do vírus, juntamente com alguns outros segmentos. Com isso, foi retirado cerca de 25 kb do DNA viral, cujo comprimento corresponde ao comprimento máximo do inserto permitido nesse tipo de vetor. O vírus contendo o DNA recombinante, ao infectar uma bactéria terá então um comportamento ativo, disparando o ciclo de replicação viral e conseqüente degeneração da célula. Com o rompimento celular, as partículas virais liberadas infectam outras células, reiniciando o ciclo. Após um determinado período, inúmeras cópias do inserto serão produzidas.

Cosmídeos

Cosmídeo é o nome dado aos vetores fabricados com elementos do fago λ e do plasmídeo. Tais vetores permitem receber insertos de 35 a 45 kb.

O DNA recombinante formado por um vetor de cosmídeo e um inserto é reconhecido pela presença de segmentos do fago λ (chamados **cos**) e empacotado como um típico DNA viral. Vale notar que o comprimento do DNA recombinante não excede o comprimento de 50 kb possível ao fago λ. Ao ser injetado em uma célula pelos mecanismos do fago, o DNA recombinante circulariza-se, formando um grande plasmídeo. De maneira análoga aos plasmídeos, o cosmídeo replica-se durante processo de divisão da célula, e essa réplica é passada juntamente com restante do material genético da célula originária para as células descendentes.

Outros Vetores

Outros tipos de vetores usados para clonagem visam aceitar insertos mais longos. Dentro dessa categoria existem os vetores de YACs, do bacteriófago P1, de BACs e de PACs.

Atualmente, os vetores chamados de YAC ou cromossomo artificial de levedura (*yeast artificial chromosome*) permitem receber insertos muito longos, que chegam a aproximadamente 1000 kb. Esse vetor é incubado em células da levedura *Saccharomyces cerevisae*, que diferente dos vetores descritos acima são incubados em células da bactéria *Escherichia coli*.

Além dos componentes básicos de replicação de um cromossomo natural, o vetor YAC contém alguns marcadores de seleção e pelo menos um sítio de restrição onde entrará o inserto. Os marcadores de seleção são genes que estão inativos na célula hospedeira e estão ligados à fabricação de nutrientes essenciais para sua sobrevivência. Assim, somente células contendo o DNA recombinante de YAC conseguem sobreviver, o que facilita a recuperação dos clones de DNA. Um problema com esse tipo de vetor é que pode haver mudanças na seqüência de nucleotídeos do inserto.

Os vetores de bacteriófagos P1 são similares aos vetores de fago λ. A principal diferença é que o genoma do fago P1 é maior do que o genoma do fago λ. Assim, um vetor P1 pode aceitar insertos bem mais longos, que chegam até 125 kb.

BACs ou cromossomos artificiais de bactéria (*bacterial artificial chromosome*) são vetores baseados também em plasmídeos. No entanto, BACs baseiam-se em um plasmídeo diferente, que é maior que os plasmídeos descritos acima, aceitando insertos de 300 kb ou mais longos.

Os vetores de PACs ou cromossomos artificiais derivados do fago P1 (*P1-derived artificial chromosome*) são formados por caraterísticas oriundas dos vetores do fago P1 e dos BACs. PACs aceitam insertos de até 300 kb.

2.6.2.2 Detalhes da Obtenção de Clones de DNA

As subseções acima descrevem os conceitos de clonagem de DNA enfatizando a manipulação de unidades de moléculas. Na prática trabalha-se com milhares de moléculas ao mesmo tempo, de maneira que os eventos biológicos entre elas ocorram. A seguir, são apresentados os passos para obtenção de clones em plasmídeo, que em linhas gerais valem para os outros vetores.

A combinação de insertos e vetores é feita misturando-se em uma única solução a amostra de fragmentos e os vetores preparados para receberem um inserto, juntamente com enzimas de restrição e ligases. As enzimas de restrição são usadas para abrir os plasmídeos. As ligases são enzimas que fazem a ligação entre segmentos de DNA por suas extremidades, e no caso juntam os fragmentos de DNA com os vetores.

Fragmentos gerados por digestão podem ser diretamente inseridos nesses vetores que também tenham sido digeridos com as mesmas enzimas de restrição (Figura 2.6). Para inserir fragmentos oriundos do método shotgun em plasmídeos, os fragmentos sofrem um tratamento enzimático que torna suas pontas retas (*blunt ends*). O vetor é também linearizado pela digestão com uma enzima de restrição que corta a fita de DNA deixando as extremidades retas, sendo então fragmento e vetor ligados pela ação das ligases.

Após a combinação entre as moléculas, elas são introduzidas em bactérias de *Escherichia coli*, por um processo chamado de transformação. É importante relatar que quando um vetor entra em uma bactéria, ele inibe a entrada de outros, condicionando então essa bactéria a receber um único inserto. Assim, a solução resultante do processo de transformação passa a conter as seguintes classes de bactérias:

1. bactérias com fragmentos de DNA;
2. bactérias com vetores abertos;
3. bactérias com vetores fechados sem inserto;

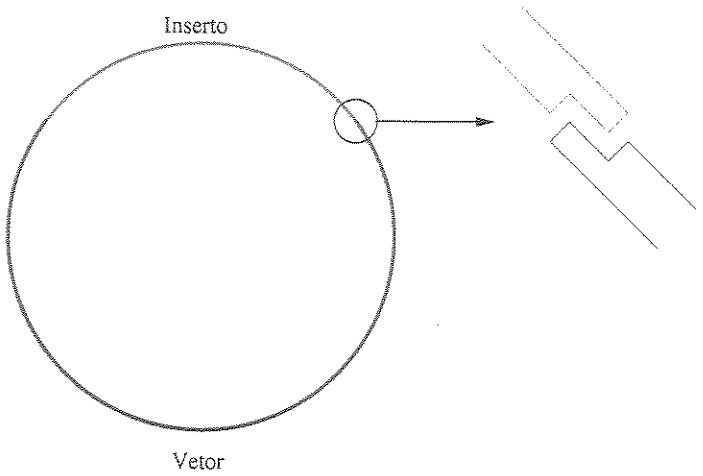


Figura 2.6: Inserto e vetor ligados pelo mesmo sítio de restrição.

4. bactérias com vetores fechados com inserto;

A classe de interesse é claramente a última. Para separar ou distinguir essa classe de interesse das demais, utiliza-se algumas estratégias. Veja a seguir como isso é feito.

As bactérias são espalhadas em placas (prato *dish*) contendo um meio adequado para seu crescimento e também um tipo de antibiótico. Aqui, graças a um tipo especial de diluição da solução, cada uma das células fica bastante separada das demais. Isso é essencial para garantir que cada colônia (isto é, a bactéria original sub-dividida em milhões de cópias) contenha dentro de cada célula constituinte uma cópia do mesmo inserto que entrou inicialmente (e não vários insertos diferentes).

As bactérias da classe 1 não mantêm o gene de resistência a antibiótico, e na presença deste não sobrevivem. Nas bactérias da classe 2, o vetor é reconhecido como um DNA estrangeiro, e então é degradado pela célula. De forma análoga a classe 1, essas bactérias em meio ao antibiótico não sobrevivem. As bactérias das classes 3 e 4 crescem, formando então colônias que são identificadas na placa por pequenas porções em forma circular (Figura 2.7). A distinção entre colônias de bactérias da classe 4 das demais, é baseada em um outro gene presente no vetor que codifica a enzima β-Galactosidase, que degrada um substrato que produz uma pigmentação azul. A abertura dos vetores de plasmídeos é feita pontualmente nesse gene. Assim, nas bactérias cujo vetor de plasmídeo não tenha recebido o inserto, a enzima é ativa, apresentando a colônia respectiva uma pigmentação azul, que a difere das demais colônias (cor branca) em que o gene da enzima foi interrompido pelo inserto.

As colônias com insertos são então arranjadas em placas de 8×12 posições. Cada posição nessas placas corresponde então aos clones dos insertos, os quais passam a ter uma identificação. A passagem de bactérias de uma placa para outra é feita com o uso

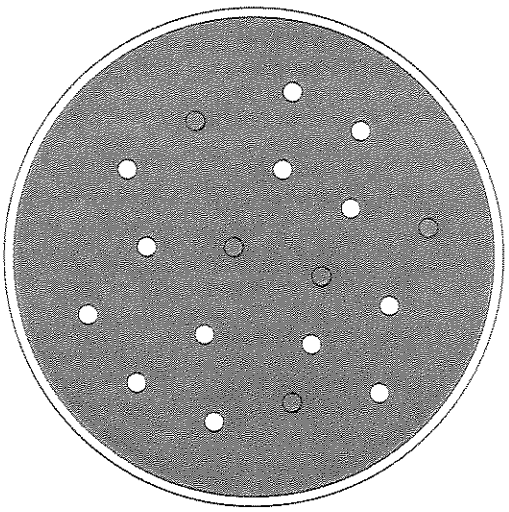


Figura 2.7: Ilustração de uma placa contendo colônias de bactérias. Os pontos brancos referem-se às colônias com insertos e os pontos sombreados referem-se às colônias sem insertos.

de pontas esterelizadas, que são espetadas uma única vez em uma colônia e a seguir colocadas individualmente em cada posição da outra placa. Esse processo é chamado de "picar colônia".

Para obter somente os clones de DNA deve ser feita uma purificação. Esse processo envolve remover o DNA das células e separar o DNA recombinante do DNA da bactéria. Para isso, a colônia é colocada em um tubo contendo uma solução com detergente que causa o rompimento da membrana das bactérias, liberando o DNA bacterial e o DNA do vetor. Utiliza-se então uma centrífuga para ajudar a separar os elementos contidos na solução, segundo a sua densidade. Com a centrifugação, os vetores com inserto que são bem mais curtos, ficam em suspensão na solução, enquanto que o DNA bacterial, proteínas e restos celulares vão para o fundo do tubo.

2.6.2.3 PCR

PCR (*Polymerase Chain Reaction*) é um processo laboratorial baseado na reação de enzimas, de onde são sintetizadas fitas novas de DNA a partir de uma molécula de DNA molde. O princípio básico do processo é que em cada reação, duas novas moléculas de DNA são produzidas a partir do DNA molde. As novas moléculas sintetizadas são então usadas na próxima reação como o DNA molde. Assim, após repetidos ciclos, é gerada uma quantidade exponencial de cópias. Numa reação de PCR estão envolvidos os seguintes componentes:

- DNA molde: molécula ou fragmento de DNA a ser copiado.

- Polimerases de DNA: as polimerases usadas no PCR trabalham em ambientes de alta temperatura.
- primers: um par de primers sintetizados quimicamente, que delimitam o segmento de DNA a ser copiado;
- Nucleotídeos livres: uma quantidade de nucleotídeos que serão usados pelas polimerases na síntese das novas fitas de DNA.

O processo de PCR baseia-se em três fases: denaturação, anelamento de primers e extensão. Na denaturação, pelo aquecimento da reação, ocorre a desestabilização da hélice dupla e sua separação em fitas simples de DNA, capazes de serem copiadas pelas polimerases. A seguir, na fase de anelamento de primers, a reação é resfriada permitindo que os primers liguem-se às fitas denaturadas. Por fim, aquecendo novamente a reação (a uma temperatura que não possibilite uma denaturação), as polimerases tornam-se ativas e, orientadas pelos primers, iniciam a síntese das novas fitas de DNA, adicionando nucleotídeos complementares aos da fita usada como molde.

Como vantagem sobre a clonagem, o PCR tem a rapidez. No entanto, tem como desvantagens a necessidade de conhecer as seqüências delimitadoras do segmento a ser amplificado, e é aplicável somente para segmentos de DNA curtos, tipicamente menores que 5 kb.

2.6.3 Leitura de DNA

Os procedimentos que envolvem a leitura de bases de DNA são baseados na técnica de **eletroforese em gel** (Seção 2.6.1). No caso, os fragmentos envolvidos correspondem a segmentos do DNA alvo (a ter sua seqüência de bases determinada) e são produzidos de tal forma que tenham comprimento diferindo em apenas uma base. Tomando-se por exemplo um pedaço de DNA que tenha a seqüência de bases `TAGCTGACTC`, teríamos os seguintes fragmentos (as letras maíusculas apenas enfatizam a base terminal):

```
T
tA
taG
tagC
tagcT
tagctG
tagctgA
tagctgaC
tagctgacT
tagctgactC
```

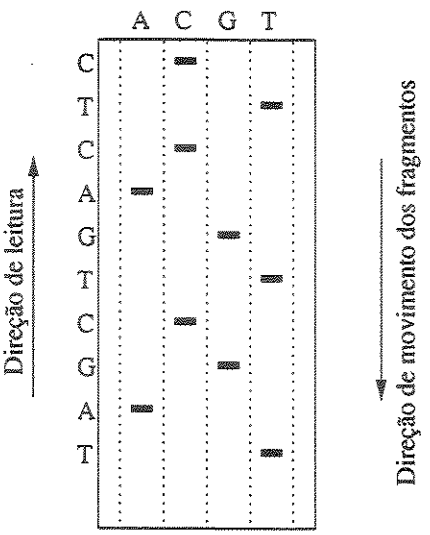


Figura 2.8: Esboço de uma imagem de gel de seqüenciamento de DNA, mostrando a seqüência de bases obtida, que no caso é lida de baixo para cima, como mostrado à esquerda da figura.

Um dos métodos de seqüenciamento baseado nessa idéia é conhecido como Dideoxi ou Sanger. Nesse caso, o DNA alvo é submetido a um processo de amplificação baseado em PCR (Seção 2.6.2.3), chamado reação de seqüenciamento. Além de cópias do DNA alvo, nucleotídeos livres e primers, esse processo usa como componente adicional nucleotídeos terminadores de cadeia chamados dideoxi-nucleotídeos. Esses nucleotídeos especiais não possuem o grupo hidroxil no carbono número 3 da pentose. Assim, quando concatenados pela polimerase à cadeia crescente de nucleotídeos, encerra-se a síntese da fita de DNA. Nessa reação é usado apenas um tipo de primer, que indica o início do segmento a ser amplificado.

Fazendo separadamente a reação para cada um dos quatro dideoxi-nucleotídeos (correspondentes aos quatro nucleotídeos), e submetendo os quatro conjuntos de fragmentos amplificados à eletroforese, obtém-se um gel de onde é feita a leitura das bases, como ilustrado na Figura 2.8.

Desde algum tempo, surgiram máquinas de seqüenciamento que automatizam o processo de leitura de bases de DNA. Nesses instrumentos, o processo de eletroforese é combinado com um emissor de raio laser, que excita os fragmentos rotulados, causando a emissão de florescência que é detectada em padrões que correspondem as bases do DNA alvo. Esses padrões são passados para um computador, onde é gerada a imagem do gel. Mais recentemente, surgiram máquinas capilares, que em linhas gerais têm o mesmo princípio, e não serão descritas aqui.

Em uma das abordagens do seqüenciamento automatizado, adotada pelas máquinas

ABI [3], os fragmentos são rotulados com corantes fluorescentes (*dyes*) de acordo com a sua base terminal e, em seguida, são colocados conjuntamente em uma mesma canaleta. Durante a corrida do gel, os corantes emitem um sinal de luz com comprimento de onda distinto quando excitados pelo raio laser. A imagem do gel revela então na seqüência de cores de cada uma de suas canaletas, a seqüência de bases de cada uma das amostras de DNA colocadas para leitura (Figura 2.9). As máquinas de seqüenciamento atuais processam até 96 seqüências de uma vez, e são capazes de obter leituras de até 1000 bp, que representam um ganho de 3 a 5 vezes em relação ao seqüenciamento manual.

No computador, a imagem do gel passa por um processo chamado de *tracking*. Nesse processo cada uma das canaletas é analisada via software, produzindo-se no final as seqüências individuais em forma de eletroferogramas ou **cromatogramas**. Cromatogramas são arquivos binários com tamanho aproximado de 250 kbytes e correspondem ao dado "cru" (*raw data*) de uma seqüência. Visualmente, um cromatograma refere-se a um conjunto de quatro linhas com amplitude distinta para cada padrão de comprimento de onda detectado pelas máquinas de seqüenciamento. Veja no lado esquerdo da Figura 2.9 a visualização do cromatograma referente a canaleta 77.

O processo de *tracking* está sujeito a erros, oriundos diretamente de problemas no gel, como sinais fracos ou mesmo ausentes em suas canaletas (como a canaleta 69 da Figura 2.9), e deslocamento de canaletas vizinhas. Os problemas no gel podem ser causados por uma série de fatores [3], como por exemplo, falhas em algum componente do próprio seqüenciador ou na reação de seqüenciamento. Logicamente, os cromatogramas são diretamente afetados pelos problemas no gel. Um exemplo crítico disso são cromatogramas sem dado rastreado (*no trace data*), ou seja, não contém dado útil sobre a seqüência de bases.

Cromatogramas são interpretados por um processo chamado de *base-calling*, pelo qual é obtida efetivamente cada uma das letras que constituem a seqüência de DNA. A seqüência de letras `A`, `G`, `C`, `T` resultante da interpretação dos cromatogramas é chamada de **read**. Em decorrência da imprecisão dos dados de um cromatograma, o processo de *base-calling* também está sujeito a erros, que caracterizam-se por substituições, inserções e remoções de bases. Esses erros podem acontecer na prática em até 5% das bases de uma seqüência.

2.6.4 Visão Global de Seqüenciamento

As seções acima descrevem individualmente as principais etapas envolvidas no processo de seqüenciamento de DNA. Nesta seção, será descrito o processo global para genomas de bactérias, onde é suficiente geralmente o uso apenas de vetores de plasmídeo e cosmídeo. Serão mostrados os passos dentro de cada etapa, partindo da molécula de DNA alvo até

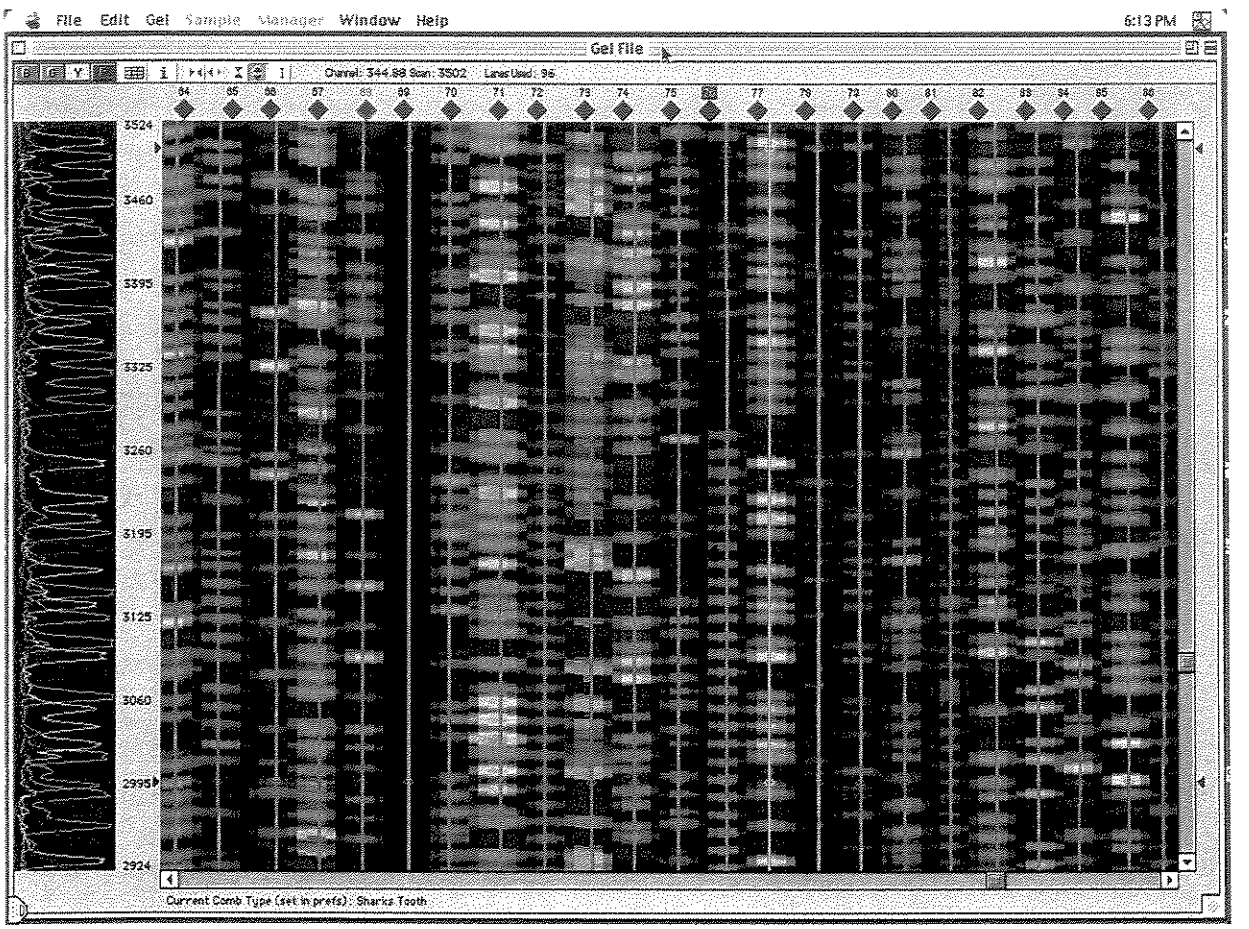


Figura 2.9: Figura de um arquivo de gel obtido com seqüenciamento automático de DNA. Na vertical estão as canaletas. As cores representam cada uma das bases: verde como `A`, azul como `C`, amarelo como `G`, vermelho como `T`. À esquerda, está a visualização do cromatograma da seqüência correspondente a canaleta 77.

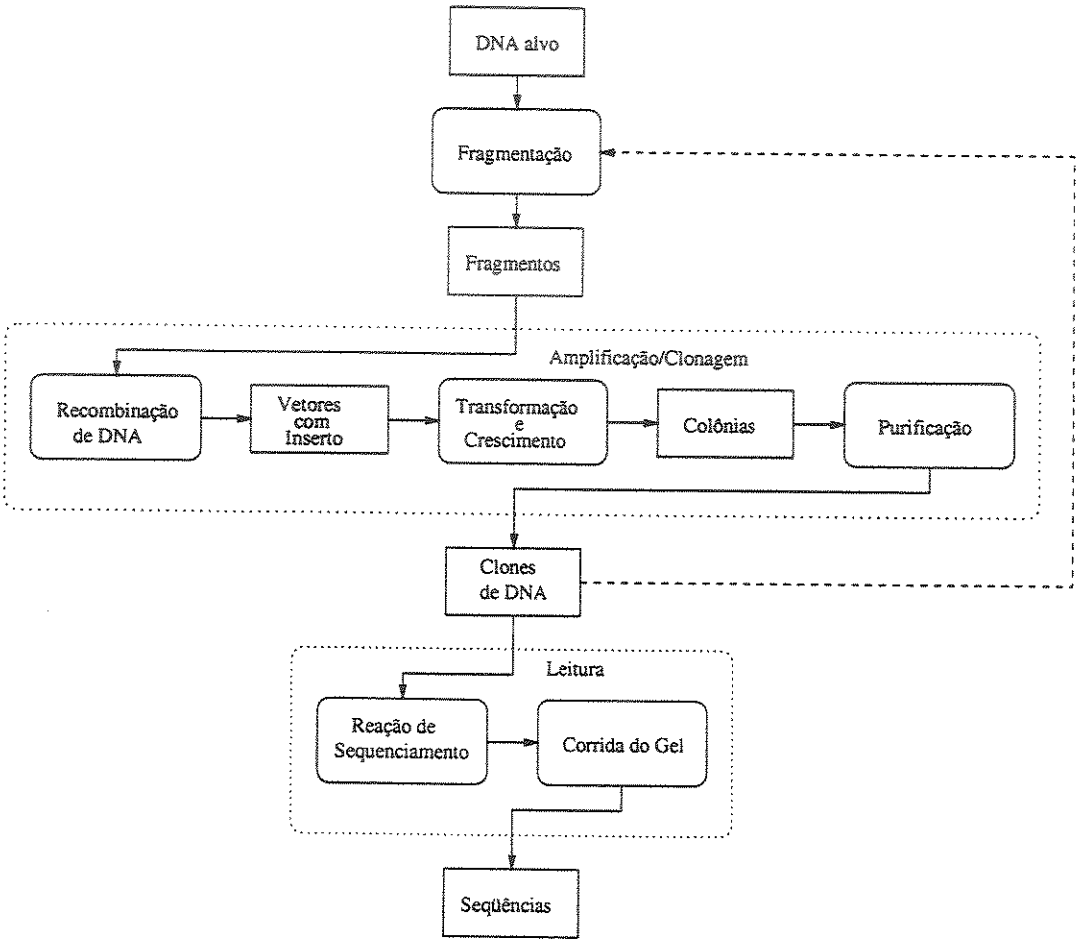


Figura 2.10: Diagrama ilustrando o processo global de seqüenciamento.

chegar na seqüência de seus fragmentos, como ilustrado no diagrama da Figura 2.10.

A primeira etapa do seqüenciamento dentro desse esquema está associada à fragmentação do DNA original em fragmentos curtos o suficiente de serem seqüenciados diretamente. O DNA alvo corresponde a várias cópias do DNA original.

Da etapa de fragmentação (Seção 2.6.1) do DNA alvo (obtida por digestão, nebulização ou sonicação), obtém-se uma grande quantidade de fragmentos com comprimentos variados e de diferentes partes do DNA original. Em geral os fragmentos são divididos e selecionados segundo seu comprimento, podendo-se amostrar mais de um conjunto de fragmentos. Uma outra possibilidade é fazer amostras distintas de fragmentos para as diferentes técnicas de fragmentação. Essas amostras de fragmentos quando submetidas a clonagem, formam as chamadas bibliotecas de clones (ou *clone library*).

As amostras de fragmentos são passadas então para a etapa de amplificação (Seção 2.6.2), onde a técnica de clonagem (Seção 2.6.2.1) é predominantemente usada. Nesta etapa, os fragmentos são inseridos em vetores, e em seguida colocados em bactérias pelo processo de transformação (Seção 2.6.2.2). As bactérias então multiplicam-se e formam colônias. Para recuperar somente os clones de DNA, é feita então a purificação.

Os insertos podem ser ainda longos demais para serem seqüenciados diretamente, requerendo sua sub-clonagem (sub-fragmentação e clonagem) em fragmentos de DNA menores, como ilustrado pela linha tracejada na Figura 2.10.

A próxima etapa do esquema refere-se a leitura de DNA. Os clones de DNA são usados na reação de seqüenciamento, de onde são produzidos fragmentos de DNA como descrito na Seção 2.6.3. O material de DNA resultante dessa reação é colocado nas máquinas de seqüenciamento, onde é automaticamente feita a corrida do gel e então geradas as seqüências na forma de cromatogramas.

Um aspecto importante sobre a leitura de DNA é que somente as extremidades do fragmento, inserto ou clone de DNA são lidas. Assim, de acordo com o seu comprimento, o fragmento poderá ter sua seqüência totalmente determinada ou não. Ou seja, se um fragmento tiver até 2 kb, a leitura das duas extremidades deve ser suficiente para determinar toda a sua seqüência. No entanto, para fragmentos mais longos, a sua leitura ou o seu seqüenciamento completo depende de outras alternativas, como o *primer walking* ou ainda uma sub-clonagem.

Como foi visto na Seção 2.6.3, a leitura de DNA depende da reação de seqüenciamento. Essa reação usa um primer que indica o início do segmento a ser amplificado. Como visto na Seção 2.6.2.3, primers são sintetizados artificialmente a partir de seqüências conhecidas. No entanto, não se conhece a seqüência do inserto em questão. Porém, as seqüências dos vetores são conhecidas e o DNA recombinante (vetor e inserto) entra na reação de seqüenciamento. Assim, os primers são sintetizados usando a seqüência dos vetores, havendo um primer para para cada um dos dois segmentos adjacentes às

extremidades do inserto.

Capítulo 3

Projetos Genoma

O estudo de um genoma é uma tarefa complexa, e envolve uma série de etapas até que se consiga chegar a resultados mais concretos, que possam então ser aplicados na prática (por exemplo, prevenção ou tratamento de doenças). A primeira etapa nessa direção corresponde determinar a seqüência de bases do genoma e, identificar e caracterizar seus genes. Os projetos genoma ou mais especificamente os projetos de seqüenciamento de genoma atuam exatamente nessa etapa e seguem em geral duas abordagens. A primeira visa determinar a seqüência completa do genoma alvo e consequentemente a obtenção de seus genes, e a outra, conhecida como seqüenciamento EST, procura determinar somente segmentos do genoma que expressam genes. Neste capítulo, apresentaremos essas duas abordagens, descrevendo suas principais etapas e diferenças. Além disso, apresentaremos uma descrição sucinta de alguns dos projetos genoma mencionados neste trabalho, da rede ONSA e do LBI.

3.1 Projetos de Seqüenciamento de Genomas Completos

Uma das abordagens empregadas para seqüenciar um genoma é fazê-lo por completo, ou seja, determinar a sua seqüência de bases, podendo então serem identificados os seus genes. Seqüenciar um genoma compreende o seqüenciamento de cada uma de suas moléculas de DNA. O seqüenciamento de uma molécula de DNA é dividido em duas etapas. A primeira etapa, essencialmente biológica, visa fragmentar a molécula e seqüenciar individualmente cada um dos fragmentos (Seção 2.6). A etapa seguinte, que envolve a área da computação, procura determinar a ordem dos fragmentos, de onde pode ser então deduzida a seqüência de bases completa da molécula. Com a seqüência de DNA determinada, pode-se então passar para o processo de identificação e análise dos genes, conhecido como anotação.

O roteiro geral seguido pelos projetos de genoma completo compreende as etapas de seqüenciamento, de montagem e de anotação, e são descritas a seguir.

3.1.1 Seqüenciamento

Esta etapa consiste de dois passos fundamentais. A quebra da molécula e o seqüenciamento propriamente dito dos fragmentos obtidos. O ponto chave desta etapa está no primeiro passo, onde são determinadas as estratégias de seqüenciamento. Uma estratégia de seqüenciamento define como deve ser dividido e, possivelmente, sub-dividido o DNA alvo até que se consiga fragmentos de comprimento que possam ser seqüenciados diretamente. A idéia geral seguida pelas diferentes estratégias é criar uma hierarquia baseada no comprimento dos fragmentos, começando com fragmentos longos e terminando com fragmentos curtos (Figura 3.1). No processo de fragmentação, dois pontos devem ser notados. Primeiro, ao quebrar uma molécula de DNA (ou um fragmento), a ordem dos fragmentos (ou sub-fragmentos) é perdida, ou seja, não é conhecida a localização dos fragmentos no DNA original. Outro ponto é que somente uma amostra dos fragmentos obtidos são seqüenciados.

A ordem entre os fragmentos pequenos (aproximadamente 700 bp) pode ser obtida comparando-se suas seqüências diretamente. Porém, para fragmentos maiores a informação de sobreposição deve ser obtida de outros modos. Nesse caso, devem ser utilizados os chamados mapas de DNA [32]. Esses mapas contém informações sobre a localização de certos marcadores em uma molécula de DNA. Marcadores são seqüências curtas, únicas e conhecidas, que se sabe existirem na molécula. Comparando-se dois fragmentos entre si em termos de seus marcadores, é possível determinar se eles têm ou não sobreposição e desse modo ordená-los. Existem dois tipos de mapas de DNA: mapas de restrição e mapas de hibridização. Os primeiros são construídos a partir de experimentos com enzimas de restrição, que indicam a presença de sítios de restrição existentes no DNA. Os últimos são construídos a partir da análise de ligação (hibridização) entre segmentos de DNA rotulados (chamados sondas) e o DNA.

Estratégias de seqüenciamento comumente utilizadas são o seqüenciamento shotgun, seqüenciamento por cosmídeos e seqüenciamento baseado em pontas. O seqüenciamento shotgun é o método de seqüenciamento mais utilizado atualmente. Nessa abordagem, várias cópias ou clones de um DNA são quebradas aleatoriamente (Seção 2.6.1), gerando fragmentos de comprimentos diversos entre 1 e 5 kb. Como o processo de quebra é aleatório, com grande probabilidade os fragmentos de diferentes clones terão sobreposição entre si e todo segmento da molécula original estará representado em algum fragmento. Seqüenciando-se os fragmentos individualmente, e subsequentemente analisando as sobreposições entre eles, pode-se deduzir então a seqüência completa da molécula original de

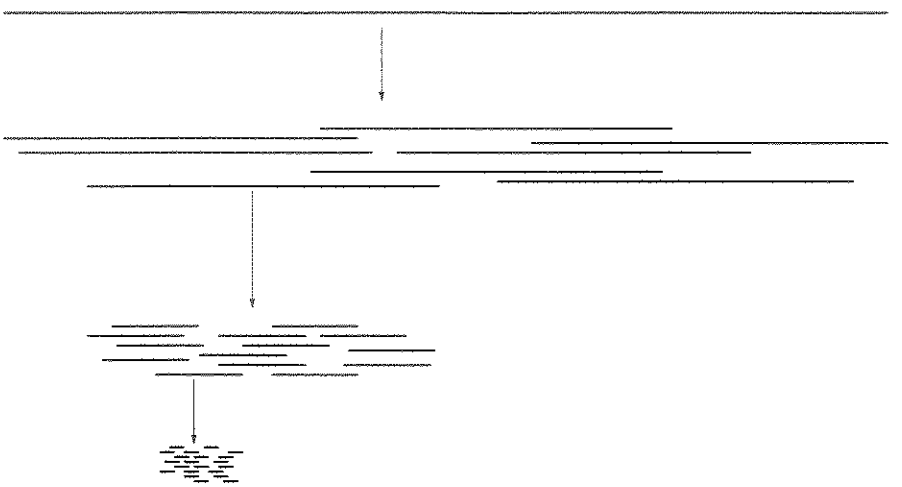


Figura 3.1: Exemplo de estratégia de seqüenciamento envolvendo fragmentação e sub-fragmentação. Primeiramente são gerados fragmentos grandes da molécula alvo. A seguir, cada um desses fragmentos grandes é sub-fragmentado, gerando fragmentos de comprimento intermediário (ou médio), ainda longos demais para serem seqüenciados. Assim, é feito mais um passo, onde cada fragmento intermediário é sub-fragmentado, de onde são gerados fragmentos pequenos e possíveis de serem seqüenciados.

DNA.

No seqüenciamento por cosmídeos, também denominado estratégia por cosmídeos ordenados, várias cópias de uma molécula de DNA são também submetidas a fragmentação. Nesse caso, são escolhidos fragmentos com comprimento estimado de 40 kbp, inseridos em vetores de cosmídeos. A escolha dos insertos a serem sequenciados é baseada em dados obtidos pela construção de mapas de DNA. Cada fragmento escolhido deve então ser seqüenciado individualmente, o que corresponde a uma sub-clonagem. Isso leva a fragmentos pequenos inseridos em vetores de plasmídeo. Constituída a seqüência de cada fragmento inserido em cosmídeo, é possível assim reconstruir a seqüência completa do DNA original de forma análoga ao seqüenciamento shotgun.

O seqüenciamento baseado em pontas compreende de certa forma uma estratégia complementar as outras. O princípio dessa abordagem está em seqüenciar apenas as extremidades de um inserto (mais longos que 1 kb), não necessariamente requerendo todo o seu seqüenciamento. A sua combinação com o seqüenciamento shotgun é um exemplo de aplicação prática que traz algumas vantagens. Uma delas é aumentar a amostragem de clones que podem ser submetidos ao seqüenciamento, não se restringindo apenas a clones com insertos pequenos (que podem ser inteiramente seqüenciados). A Figura 3.2 apresenta o exemplo de seqüenciamento de um inserto em cosmídeo, usando a abordagem de shotgun combinada com seqüenciamento de pontas. Nesse exemplo, faltaria apenas seqüenciar o miolo do clone identificado com o número 5 para obtenção da seqüência

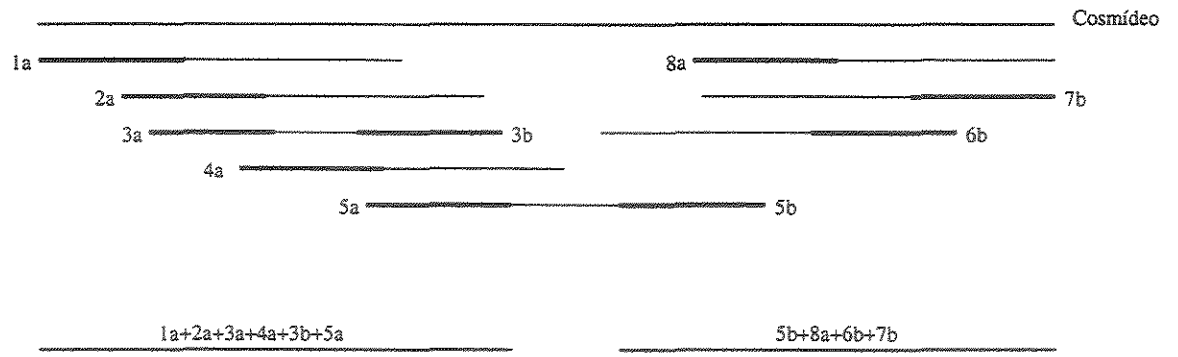


Figura 3.2: Seqüenciamento de um inserto em cosmídeo, combinando a abordagem de shotgun e a baseada em pontas. A linhas mais espessas representam os segmentos seqüenciados. Os números são a identificação dos clones e as letras a e b referem-se qual a ponta do clone seqüenciada.

completa do inserto.

Vale notar que mesmo sendo uma estratégia de seqüenciamento bem definida, ela está sujeita a erros provenientes da natureza experimental do processo laboratorial de seqüenciamento, levando a problemas na etapa posterior de montagem de fragmentos (ver Seção 4.1).

3.1.2 Montagem

Com o seqüenciamento dos fragmentos, começa a etapa de montagem, que usa então as seqüências individuais de cada fragmento na reconstrução da seqüência completa da molécula original ou dos fragmentos longos. Assim, de acordo com a estratégia usada, pode-se distinguir uma montagem global, visando a seqüência completa, e montagens individuais de fragmentos longos. Fazer uma montagem não é uma tarefa elementar. Existem vários fatores que complicam o processo (ver Seção 4.1), exigindo que haja várias iterações até que se obtenha a seqüência final. Baseado no resultado de cada montagem, fragmentos podem ser adicionados ou talvez retirados do conjunto de entrada.

3.1.3 Anotação

Dado que a seqüência completa de DNA de um cromossomo ou do genoma esteja determinada, pode-se iniciar o processo para determinar e analisar os genes dessa seqüência. Esse processo é conhecido como anotação. A anotação envolve dois passos: a identificação e a descrição (ou categorização) dos genes. O primeiro passo pode ser realizado com o auxílio de ferramentas computacionais, que automatizam o processo de reconhecimento de genes em uma seqüência. Já o segundo passo, pode ser automatizado em parte, pois

requer muito a intervenção humana, no sentido de julgar e atribuir a função bioquímica de cada seqüência gênica.

3.2 Projetos de Seqüenciamento de EST

Uma outra abordagem utilizada por projetos de seqüenciamento de genoma é conhecida como EST (*Expressed Sequence Tag*) [9]. Seu objetivo é seqüenciar regiões de maior interesse, ou seja, segmentos do genoma que codificam proteína. O fundamento para essa abordagem está no seqüenciamento de cDNAs (Seção 2.3.1), que no caso são cópias de mRNAs, ou seja, dos genes transcritos.

Esse tipo de tecnologia torna mais rápido o processo de determinação de genes, se comparada com a abordagem de genoma completo, principalmente para genomas de organismos altamente complexos, como animais e plantas, onde uma parcela geralmente bem pequena de seu genoma é compreendida por genes. Uma característica desses organismos vivos é que embora todos os genes estejam presentes nas células, somente uma porção é expressa em determinados tecidos, como discutido na Seção 2.5. Com isso, a abordagem EST possibilita domínios distintos de estudo dos genes, ou seja, estudar genes relacionados a um determinado tecido.

Basicamente, um projeto EST pode ser dividido nas mesmas etapas que um projeto de genoma completo (seqüenciamento, montagem e anotação) embora possua características bem distintas. Na etapa de seqüenciamento, por exemplo, a meta está em capturar os mRNAs de tecidos distintos, transcrevê-los em cDNAs, montar bibliotecas (de clones) distintas de cDNAs conforme o tecido, e por fim seqüenciar individualmente os clones de cDNA. Embora haja uma etapa de montagem em projetos EST sua finalidade está em agrupar em *clusters* as seqüências derivadas do mesmo gene. Essa etapa é conhecida como *clustering*. Por último, a etapa de anotação visa descrever a função de cada uma das seqüências gênicas ou cada um dos clusters recuperados com essa técnica.

A diferença fundamental entre um projeto de genoma completo e um projeto genoma de ESTs está no fato de que num projeto EST nada se pode inferir sobre a ausência de genes, visto que não há garantia de que todos os genes de interesse serão seqüenciados. Além disso, os projetos de genoma completo fornecem informações valiosas sobre seqüências regulatórias que não são transcritas no mRNA.

3.3 Definição do Papel da Bioinformática para Projetos Genoma

A imensa soma de dados produzidos em um projeto genoma requer claramente a aplicação de processos automatizados, que envolvem o armazenamento, organização e análise dos dados. Nesse âmbito, a bioinformática tem o papel de fornecer softwares e ferramentas computacionais que são utilizados nas principais etapas dos projetos genoma (Seções 3.1 e 3.2):

- Seqüenciamento: entre as tarefas computacionais de análise realizadas nesta etapa, está o processo de *base-calling*, que visa converter os dados produzidos pelas máquinas seqüenciadoras (*raw data*) em seqüências de letras.
- Montagem: essa etapa é essencialmente realizada por um programa de computador (montador), que compara as seqüências si, encontra a sobreposição entre elas, e então reproduz a seqüência de uma molécula fragmentada ou agrupa em clusters os ESTs relativos a um mesmo gene.
- Anotação: dessa etapa fazem parte programas de computador usados na identificação dos genes, envolvidos principalmente em encontrar seqüências codificadoras (ORFs), e programas de busca de seqüências similares em bancos de dados de seqüências conhecidas, para ajudar na classificação funcional das novas ORFs.

3.4 Projetos Genoma de Procariotos

Apesar do genoma humano ser o mais importante alvo de estudos da área genômica, vários outros projetos genoma de organismos satélites têm sido iniciados e finalizados (http://www.tigr.org/tdb/mdb/mdbcomplete.html), e muitos ainda continuam em andamento (http://www.tigr.org/tdb/mdb/mdbinprogress.html). Esses projetos têm seu papel de importância em relação ao projeto humano, pois têm ajudado no aperfeiçoamento e surgimento de novas técnicas e tecnologias de seqüenciamento, bem como na disseminação do conhecimento especializado para trabalhar nesse novo campo de pesquisa.

Dentre os diversos organismos alvo de estudo, é representativa a quantidade de projetos genomas de procariotos, ou seja, pequenos microorganismos como as bactérias. O genoma desses organismos é simples, quando comparados aos genomas de organismos mais evoluídos como aminais e plantas. Isso representa uma das principais motivações pelo qual se tem estudado essa classe de organismos. Outras motivações estão relacionadas a aspectos ambientais, agronômicos e patológicos do organismo alvo.

O Brasil entrou no ramo da genômica recentemente, com o seqüenciamento completo do genoma da bactéria *Xylella fastidiosa*, firmando-se no ramo com a realização de outros projetos genoma, como o projeto *Xanthomonas* e o projeto *Xylella* PD (*Pierce's disease*). Além disso, outros projetos genoma brasileiros estão em andamento (http://watson.fapesp.br/genoma3.htm e http://www.brgene.lncc.br/index.html) e mostram o resultado do grande investimento nessa área.

3.4.1 O Projeto Genoma *Xylella fastidiosa*

O projeto teve como objetivo principal seqüenciar o genoma completo da bactéria *Xylella fastidiosa* [33], e também determinar seus genes. Outra finalidade do projeto era disseminar a cultura genômica no país, promovendo treinamentos na área de biotecnologia, e incentivando grupos de pesquisa na área de biologia molecular a aplicar e desenvolver essas técnicas em suas próprias pesquisas. Com a finalização do projeto, a bactéria *Xylella fastidiosa* passou a ser o primeiro patógeno (agente causador de doenças) de planta a ter seu genoma completamente seqüenciado.

O projeto apresenta-se com características diferentes em relação a sua organização, se comparado à maioria dos projetos genoma, onde toda infraestrutura de seqüenciamento e processamento de dados é centralizada em uma mesma instituição, como é o caso da TIGR (*The Institute for Genomic Research*), uma das mais conhecidas instituições de pesquisa na área genômica, e responsável por grande parte dos genomas seqüenciados. No caso do projeto *Xylella*, criou-se uma rede virtual (rede **ONSA** - *Organization for Nucleotide Sequencing and Analysis*) composta por laboratórios de biotecnologia geograficamente distantes e por um centro de bioinformática, e que possibilita o compartilhamento de dados e de informações entre os laboratórios. Uma das vantagens desse tipo de organização é a possibilidade da troca de conhecimento entre os pesquisadores pertencentes a diferentes instituições, a áreas de aplicação diferentes (plantas, bactérias, etc), ou mesmo de disciplinas diferentes (computação e biologia). Outra vantagem é o aproveitamento de recursos materiais e humanos já existentes na composição da infra-estrutura, possibilitando uma economia nos custos do projeto. Considerando a tecnologia de seqüenciamento e o número de laboratórios envolvidos, esse tipo de estrutura distribuída pode prover uma alta taxa de produção de seqüências em um espaço menor de tempo. Veja mais detalhes sobre a rede ONSA abaixo.

Xylella fastidiosa é uma bactéria que causa doenças em plantas. No Brasil, essa bactéria tem infectado plantações de citros causando a CVC (Clorose Variegada de Citros) [13], doença popularmente conhecida como "amarelinho". O impacto dessa doença é economicamente sentido em plantações de laranja, em especial no estado de São Paulo, que é responsável por aproximadamente 87% da produção brasileira, e possui grande par-

cela na produção mundial [26]. Com a finalização do projeto, observou-se que o genoma dessa bactéria é formado por três moléculas de DNA: pelo cromossomo principal com 2679305 bp, e por dois plasmídeos, sendo um com 51158 bp e outro com 1285 bp. Foram identificados 2904 genes. Isso representa apenas um passo inicial na busca de meios de combate a doença do amarelinho. Os passos seguintes farão uso das informações obtidas com o seqüenciamento do genoma em estudos mais aprofundados dos genes (projeto genoma funcional), em particular os que estão relacionados a função de patogenicidade da bactéria. Explorando-se os mecanismos desses genes, talvez seja então possível encontrar maneiras práticas de combater a doença.

A Rede ONSA

A rede ONSA (*Organization for Nucleotide Sequencing and Analysis*) é constituída por diferentes laboratórios de pesquisa de diferentes instituições de ensino e pesquisa do estado de São Paulo. Ela foi criada com o apoio da FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), visando inicialmente o seqüenciamento do genoma da bactéria *Xylella fastidiosa* e posteriormente de outros genomas. Antes mesmo da finalização desse primeiro projeto, outros já tinham sido iniciados, e atualmente a rede ONSA abriga outros projetos (http://watson.fapesp.br/genoma3.htm). Fizeram parte do projeto *Xylella* 35 laboratórios oriundos das seguintes instituições e cidades:

- Instituto Agronômico de Campinas (IAC) - Campinas
- Instituto Agronômico de Campinas (IAC) - Cordeirópolis
- Instituto Biológico - São Paulo
- Instituto Butantan - São Paulo
- Instituto Ludwig de Pesquisas sobre o Câncer - São Paulo
- Universidade de Mogi das Cruzes - Mogi das Cruzes
- Universidade de Ribeirão Preto (UNAERP) - Ribeirão Preto
- Universidade de São Paulo (USP) - Piracicaba
- Universidade de São Paulo (USP) - Riberão Preto
- Universidade de São Paulo (USP) - São Paulo
- Universidade Estadual de Campinas (UNICAMP) - Campinas
- Universidade Estadual Paulista (UNESP) - Jaboticabal

o que desperta um grande interesse por parte da indústria de biotecnologia, que usa essa bactéria para transferir genes de interesse para plantas geneticamente modificadas, como por exemplo o algodão, o milho e o tomate.

O genoma dessa bactéria possui quatro moléculas de DNA, que totalizam juntas 5.67 milhões de pares de base. Essas moléculas se dividem em: um cromossomo circular com 2.84 mb, um cromossomo linear com 2.07 mb, um plasmídeo com 542 kb e um plasmídeo com 214 kb. Ao todo, o genoma possui 5419 genes que codificam proteína.

3.5 LBI

O LBI (Laboratório de Bioinformática) foi inicialmente criado para fazer parte do projeto genoma da *Xylella fastidiosa*, provendo a estrutura computacional para armazenamento, processamento e análise dos dados gerados pela rede ONSA. Mesmo antes da finalização desse projeto pioneiro no Brasil, o LBI já estava envolvido com os projetos genoma *Xanthomonas* e SUCEST, e posteriormente com outros projetos (http://www.lbi.ic.unicamp.br).

Começando com três pessoas e apenas uma estação de trabalho, o LBI cresceu e atualmente possui diversas máquinas (http://www.lbi.ic.unicamp.br/lbi_hardware.html), entre estações de trabalho e servidoras, além de ter aumentado os recursos humanos (http://www.lbi.ic.unicamp.br/lbi_team.html).

Além das máquinas e pessoas, o LBI possui três sistemas de software principais voltados para área de bioinformática. Os dois primeiros referem-se aos sistemas empregados em projetos de genoma completo e estão descritos neste trabalho. O outro sistema está associado a projetos genoma de ESTs e é capaz de organizar os dados gerenciais do projeto, receber e analisar as seqüências, fazer o *clustering* das seqüências, e diponibilizar várias informações sobre os dados processados. Esse sistema foi desenvolvido no projeto SUCEST e está sendo utilizado (com algumas modificações) no projeto EST *Schistosoma mansoni* (http://verjo18.iq.usp.br/schisto).

Capítulo 4

Bioinformática de Base

Por bioinformática de base entende-se as seguintes tarefas:

1. recebimento, armazenamento e organização dos fragmentos de DNA gerados pelos laboratórios;
2. processamento preliminar dos fragmentos;
3. montagem dos fragmentos.

A tarefa tecnicamente mais complexa é a terceira, e suas complexidades influenciam nas duas primeiras. Por esses motivos, iniciamos os capítulo descrevendo o processo de montagem. Seguem-se outras seções que dão detalhes sobre as demais tarefas tais como realizadas no projeto *Xylella fastidiosa*.

4.1 Montagem

Esta seção se baseia no capítulo 4 de Setubal e Meidanis [32]. A montagem de fragmentos visa determinar a ordem de uma coleção de seqüências de fragmentos de um mesmo DNA, com base na sobreposição que exista entre elas, para assim obter a seqüência de bases do DNA original. Veja a seguir um exemplo simples do problema. Dado o conjunto de seqüências, o objetivo é produzir um arranjo (*layout*), colocando uma seqüência sobre a outra, inserindo espaços (representado pelo caracter '-') de tal forma que os segmentos de sobreposição fiquem alinhados. Baseado no arranjo ou alinhamento obtido, deduz-se então a seqüência final conhecida como **consenso**. A construção desse arranjo, ou alinhamento, corresponde a uma extensão do problema básico de alinhamento múltiplo de seqüências.

Fragmentos		Alinhamento
ACGGAGCA		ACGGAGCA----------
GAGCACTTG		---GAGCACTTG--------
CTTGAGTC		--------CTTGAGTC----
GGAGCACT		--GGAGCACT----------
TGAGTCAAAC		----------TGAGTCAAAC
GCACTTG		-----GCACTTG--------
	Consenso	ACGGAGCACTTGAGTCAAAC

Este exemplo apresenta apenas um caso ideal do problema. Na prática existem vários fatores que dificultam o processo de montagem. Entre as principais complicações, descritas abaixo, estão os erros nas seqüências, orientação desconhecida, repetições e falta de cobertura. Mesmo para o caso ideal, formalismos apresentados para problema da montagem de fragmentos [32] não servem para aplicações práticas. Assim, nas soluções práticas para o problema o que existe atualmente são programas heurísticos que incorporam diversas regras para tratamento de casos específicos.

Uma das grandes melhorias no processo de montagem foi obtida recentemente pela incorporação e utilização da qualidade das bases (Seção 4.2.1), que foi conseguida graças ao surgimento de algoritmos de *base-calling* mais precisos [16, 15]. O principal benefício do fator qualidade está no tratamento mais preciso das bases com erros de *base-call*. Ou seja, no caso das bases terem uma qualidade muito baixa e estarem nas extremidades da seqüência, elas são retiradas ou ignoradas no processo de montagem. Para o caso das bases com baixa qualidade que se encontram no meio da seqüência, tais bases passam ter um peso menor do que as bases que tenham alta qualidade, quando é feita a construção do alinhamento e obtenção das bases da seqüência consenso.

Erros

Um dos problemas comuns nas seqüências são os erros de *base-call* (Seção 2.6.3). A principal conseqüência desses erros é que as seqüências passam a ter divergências entre suas bases, e com com isso, sobreposições aproximadas entre as seqüências devem ser consideradas.

Outros problemas são a contaminação de seqüências e as seqüências quiméricas. As seqüências contaminadas devem passar pelo processo de *screening* antes de serem inseridas no conjunto de entrada de uma montagem. No processo de *screening*, os segmentos contaminados de uma seqüência são identificados e marcados (mascarados) trocando cada uma de suas bases pela letra X. A presença de seqüências contaminadas pode influenciar erroneamente no cálculo de sobreposição entre as seqüências. Por exemplo, as seqüências podem

apresentar uma falsa sobreposição por estarem contaminadas com a mesma seqüência de um vetor.

Seqüências quiméricas ou quimeras aparecem quando dois fragmentos de partes distintas do DNA juntam-se por suas extremidades e formam um único fragmento que não é contíguo no DNA original. As seqüências desses fragmentos devem ser identificadas e eliminadas da montagem, pois sua presença pode resultar montagens confusas e erradas.

Orientação do Fragmento

Como visto, uma molécula de DNA possui duas fitas. Embora se conheça a orientação do fragmento ($5' \rightarrow 3'$), não há como saber a qual das fitas ele pertence. Porém é conhecido que as duas fitas são complementares e possuem orientação oposta. Com isso, a seqüência complementar reversa dos fragmentos também deve ser considerada na montagem.

Falta de Cobertura

Quando uma amostragem de fragmentos é produzida, pode acontecer que determinadas regiões da molécula alvo não estejam representadas por um desses fragmentos (Figura 4.1). Essas regiões são ditas não possuirem cobertura. Nos casos em que existam regiões sem cobertura, a reconstrução da seqüência completa fica comprometida, originando então várias regiões contíguas distintas, chamadas **contigs**. Os segmentos sem cobertura em uma montagem são denominados de buracos ou *gaps*. Há dois tipos de buraco: buraco de seqüência e buraco físico. Um buraco de seqüência é resolvido com o seqüenciamento completo de um inserto que está na biblioteca de clones mas que não foi seqüenciado completamente ainda. Um buraco físico ocorre devido a não haver nenhum inserto na biblioteca de clones que venha a resolver a região descoberta.

Podemos definir cobertura de uma posição p como sendo número de fragmentos que possuem uma de suas bases passando em p. A aplicação desse cálculo é possível apenas dispondo-se de uma montagem. Utiliza-se em geral calcular a cobertura média, que corresponde a soma do comprimento dos fragmentos a serem seqüenciados dividido pelo comprimento estimado da molécula alvo. Quanto maior a cobertura média, menor a chance de haver regiões não cobertas. Esse cálculo ajuda a fazer um balanço entre o esforço/custo de seqüenciamento e a possibilidade de fechamento da seqüência completa.

Repetições

Repetições são segmentos que aparecem duas ou mais vezes em uma molécula de DNA. Na montagem, são problemáticas as repetições longas para as quais não há fragmentos suficientemente longos capazes de cobrir toda a extensão da região repetida. Dentre as

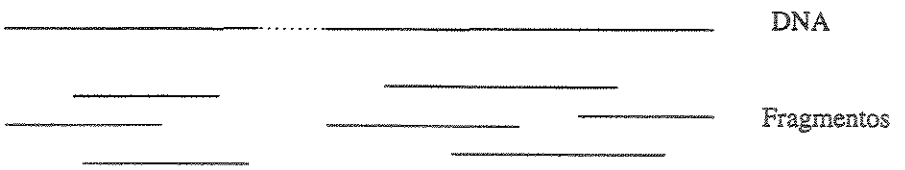


Figura 4.1: Exemplo de uma região sem cobertura. Linha pontilhada representa trecho não coberto.

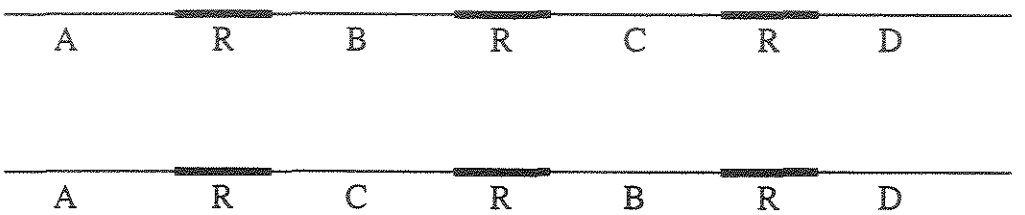


Figura 4.2: Repetições em uma molécula de DNA levando a montagens ambíguas.

possíveis conseqüências dessa característica em genomas, a principal recai na ambiguidade do arranjo dos fragmentos, e consequentemente leva a mais de uma seqüência consenso. Na Figura 4.2 é ilustrado um exemplo dessa ambiguidade, onde os dois segmentos (B e C), ladeados por três cópias de uma mesma repetição (R), podem permutar-se e produzir duas montagens distintas.

Um outro tipo de confusão é causado por repetições colapsadas. Nesse caso os segmentos repetidos são montados como um único segmento. Um exemplo de repetição colapsada é ilustrado na Figura 4.3, onde são montados dois contigs ao invés de um. Na figura é ilustrado como os fragmentos contidos dentro da repetição são aglutinados, fazendo com que a quantidade de fragmentos na região repetida seja maior que a quantidade média nas demais regiões (supondo distribuição uniforme e aleatória).

Existe também a distinção entre repetições exatas e repetições aproximadas, ou seja, as primeiras tem exatamente a mesma seqüência enquanto que as últimas têm seqüências com algumas divergências. Por último, existem as repetições *tandem*, que são repetições concatenadas, como por exemplo a seqüência CGACTCGACTCGACTCGACTCGACTCGACT, composta de consecutivas sub-seqüências CGACT.

4.2 Ferramentas

Nesta seção descrevemos algumas ferramentas usadas no pipeline de montagem, e em geral são bastante utilizadas por outros centros de bioinformática. Juntamente com descrição das ferramentas, apresentamos alguns conceitos relacionados.

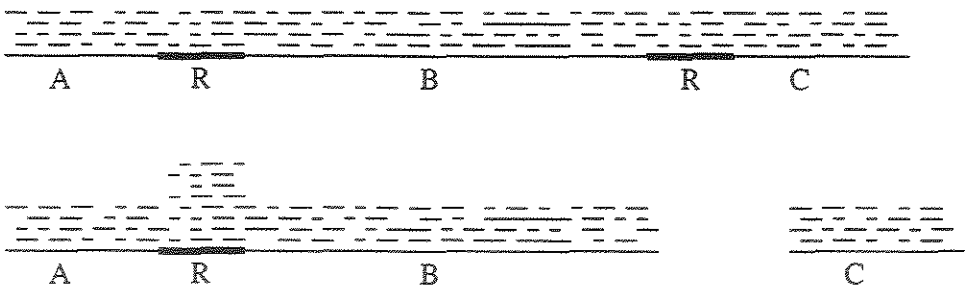


Figura 4.3: Exemplo de repetição colapsada, levando à uma solução com dois contigs, ao invés da solução correta de um único contig.

4.2.1 O Software phred/phrap/consed

Esse software corresponde a um pacote de programas usado para fazer montagens de seqüências de DNA. Tem como característica principal o uso de qualidade para as bases, um fator adicional que possibilita a obtenção de montagens mais precisas. O pacote é formado pelos programas phred [16, 15], swat, cross_match, phrap [18] e consed [17].

Phred

O programa phred é um *base-caller*. Ele lê os dados gerados pelo processo de *tracking* (cromatogramas) e escreve a sua seqüência de bases. Além disso, ele atribui valores de qualidade para cada uma das bases. O valor de qualidade está relacionado a probabilidade de erro na leitura da base. Quanto menor a probalidade de erro, maior é o valor da qualidade. O valor da qualidade de uma base é dado pela fórmula

$$VQ = -10 \times \log(PE)$$

onde PE é a probabilidade de erro, e log é o logaritmo na base 10. Assim, se a probabilidade de erro for 1 em 10^2, a qualidade tem valor 20. Se a probabilidade de erro for 1 em 10^3, a qualidade tem valor 30. E assim por diante.

Para calcular a probabilidade de erro, e então o valor de qualidade de cada uma das bases de uma seqüência, o programa phred usa um certo número de parâmetros extraídos do cromatograma. Os valores dos parâmetros de uma base são usados para procurar em uma tabela de consulta uma linha cujos valores sejam tão grandes quanto cada um dos respectivos valores dos parâmetros de consulta. A probabilidade de erro associada a essa linha da tabela é então atribuída a base. O processo de construção dessa tabela de consulta é chamado de calibração de probabilidade de erro e é feito empiricamente sobre um conjunto de cromatogramas de treinamento.

Swat e Cross_match

O programa swat é uma implementação eficiente do algoritmo de Smith-Waterman [34] - também conhecido como algoritmo para alinhamento ou comparação local - que visa comparar duas seqüências procurando identificar sub-seqüências contíguas (ou segmentos) similares entre si. O resultado dessa comparação é obtido a partir do preenchimento de uma matriz $m \times n$ (m e n são os comprimentos das seqüências) usando programação dinâmica. Com algumas otimizações em relação ao algoritmo original, o programa swat consegue reduzir o número de instruções executadas para preencher a matriz, diminuindo assim o tempo de execução. A aplicação comum do programa swat é comparar uma ou mais seqüências de DNA ou de proteínas contra um conjunto de seqüências procurando encontrar (sub-)seqüências similares às de consulta.

O cross_match é um programa de aplicação geral para comparação de dois conjuntos seqüências de DNA baseado no mesmo algoritmo que o programa swat. Em relação ao swat, o programa cross_match possui outras características que visam diminuir o seu tempo de execução. Uma das características permite que o cálculo da comparação entre as seqüências se limite as bandas da matriz Smith-Waterman, levando a uma diminuição no tempo de execução para grandes conjuntos de seqüências, sem comprometer significativamente a sensibilidade. Quanto menor a sensibilidade, maior é a possibilidade de que seqüências fracamente parecidas não sejam reconhecidas, não havendo o alinhamento entre elas. As bandas definem uma porção da matriz a ser preenchida, delimitada por regiões em volta da diagonal principal. Outra técnica usada pelo programa baseia-se em submeter à programação dinâmica somente determinados pares de seqüências. No caso, somente os pares de seqüências que tenham uma ou mais palavras exatas comuns de tamanho mínimo são comparadas.

Exemplos de aplicação do programa cross_match são: comparar um conjunto de reads com um conjunto de seqüências de vetores e produzir versões dos reads com as seqüências dos vetores mascaradas (*screened reads*), comparar seqüências produzidas por dois programas de montagens distintos, comparar uma seqüência de consulta contra um conjunto de seqüências. Para esse último caso, existem programas mais eficientes como é o caso do BLAST [1, 2, 43].

Diferente do cross_match, que usa um algoritmo baseado programação dinâmica para calcular a similaridade, o BLAST é uma heurística desenvolvida para ser uma alternativa mais rápida para comparar seqüências. A motivação para o aparecimento do BLAST é o problema da comparação de uma seqüência contra um banco contendo milhares de seqüências. A grande diferença entre esses dois programas é que BLAST é muito mais rápido, embora seja menos sensível que o cross_match. Ambos os programas aproveitam-se do conceito de que um alinhamento representando uma relação de similaridade verdadeira entre duas seqüências contém pelo menos uma palavra exata que é comum em

ambas seqüências [6]. A busca de palavras comuns pode ser feita de forma muito rápida pré-indexando-se as palavras contidas na seqüência de consulta e na seqüência alvo. Aplicando essa técnica, os programas conseguem salvar tempo, ou seja, impedem que esforço computacional seja dispendido na comparação de seqüências que claramente não estejam relacionadas. O cross_match faz a pré-indexação das seqüências no momento da consulta, portanto é mais adequado para comparação 2 a 2, enquanto que o BLAST faz uso de um banco pré-indexado. O tempo de execução dos dois programas é quadrático $O(n)$, supondo a comparação de duas seqüências com comprimento médio n.

Phrap

Phrap (*Phil's assembly program*) [18] é um programa para montagem de seqüências de DNA. Assim como outras implementações, o phrap divide o problema em três partes: detecção de sobreposição, alinhamento das seqüências e determinação do consenso.

A primeira fase visa encontrar as sobreposições existentes entre cada par de seqüências (inclusive os complementos reversos), em particular sobreposições entre o sufixo da primeira seqüência com o prefixo da segunda seqüência. Isso corresponde a $2k(2k-1)$ combinações para um conjunto de k seqüências. Dado que o algoritmo para achar a sobreposição aproximada entre duas seqüências é da ordem $O(n^2)$ (n é o tamanho médio das seqüências), o tempo gasto por essa fase torna-se o gargalo para os programas de montagem. O phrap usa o cross_match para fazer essas comparações, e como visto acima, o cross_match possui mecanismos que diminuem o tempo gasto nessa etapa. Para as seqüências comparadas o phrap atribui uma pontuação para o alinhamento produzido entre duas seqüências. Essa pontuação qualifica o nível de similaridade entre as seqüências e é usada com a qualidade das bases para produzir uma outra pontuação a ser utilizada na fase seguinte.

A segunda fase visa construir o alinhamento das seqüências a partir das informações de sobreposição obtidas pela fase anterior. No phrap isso é feito de forma incremental, sendo em cada passo feita a junção de uma seqüência ou de um par de seqüências com um outro grupo de seqüências já alinhadas. Toma-se primeiramente pares de seqüências cuja pontuação seja a melhor. Diferente de outras implementações que seguem essa abordagem (gulosa), o phrap utiliza uma pontuação modificada que considera a qualidade das bases.

Na última fase é feita a determinação da seqüência consenso dos alinhamentos (ou contigs) obtidos. Em outros programas de montagem isso é feito por votação, ou seja, é tomada a base mais freqüente na coluna do alinhamento. No phrap, cada base da seqüência consenso é determinada pela base com qualidade mais alta entre as bases de cada coluna do alinhamento.

Consed

O consed [17] é um programa usado para visualizar e editar montagens produzidas pelo phrap. Além disso, o consed ajuda no processo de *finishing* ou finalização de seqüências. A visualização de uma montagem visa mostrar suas seqüências e como elas estão dispostas no alinhamento, além da seqüência consenso gerada. O programa destaca nas seqüências exibidas, as regiões ou bases de alta qualidade (identificadas por uma cor de fundo mais clara) e as regiões ou bases de baixa qualidade (identificadas por uma cor de fundo mais escura), como pode ser visto na Figura 4.4. O consed permite também a visualização de cromatogramas (Figura 4.5).

A habilidade de editar uma montagem permite a correção de erros de *base-call* e também a correção de erros na própria montagem. A correção dos erros de *base-call* nos reads da montagem é feita por uma operação de substituição, remoção ou inserção de uma base. Em geral isso pode ser feito desde que haja um ou mais reads que tenham bases adequadas com qualidade alta para inferir a correção da base incorreta. Os erros de montagem podem ser corrigidos no consed pela marcação das seqüências montadas incorretamente com determinados *tags*, sendo as modificações efetuadas com uma re-montagem envolvendo tais seqüências. Há *tags* que possibilitam separar seqüências que não deveriam estar juntas, e *tags* que possibilitam juntar seqüências que aparecem separadas na montagem.

O processo de *finishing* visa colocar a seqüência (ou a montagem) em um nível satisfatório de qualidade e precisão medidos por algum critério objetivo. Usa-se comumente basear tal critério no número esperado de erros encontrados na seqüência, sendo considerada uma seqüência finalizada aquela que possui um número máximo de erros (por exemplo 0,0001 ou 1 erro a cada 10000 bases).

As qualidades das bases produzidas pelo phred e usadas pelo phrap servem como um bom critério objetivo para guiar o *finishing* [17]. Em especial, a precisão alcançada pelo phrap deve-se ao uso dessas qualidades. Alguns dos pontos que são definidos como problemas e que contribuem para taxa de erros com uso de qualidade de bases são descritos na Seção 4.3.3.

O consed ajuda no *finishing* identificando as regiões nas seqüências que contribuem para o aumento da taxa de erro, possibilitando assim a correção dos segmentos com erros, pela edição de bases no próprio consed como dito acima, ou pela produção de dados de seqüências adicionais a serem inseridos na montagem posterior. O intuito dessas correções é levar a diminuição do número de erros, até chegar ao patamar desejado.

O consed possui ainda algumas ferramentas que podem ser úteis no fechamento de buracos. Uma dessas ferramentas calcula para uma determinada região, os segmentos que podem ser usados para fabricação de *primers*, que podem ser utilizadas pela técnica de *primer walking* (Seção 2.6), objetivando resolver principalmente os buracos físicos.

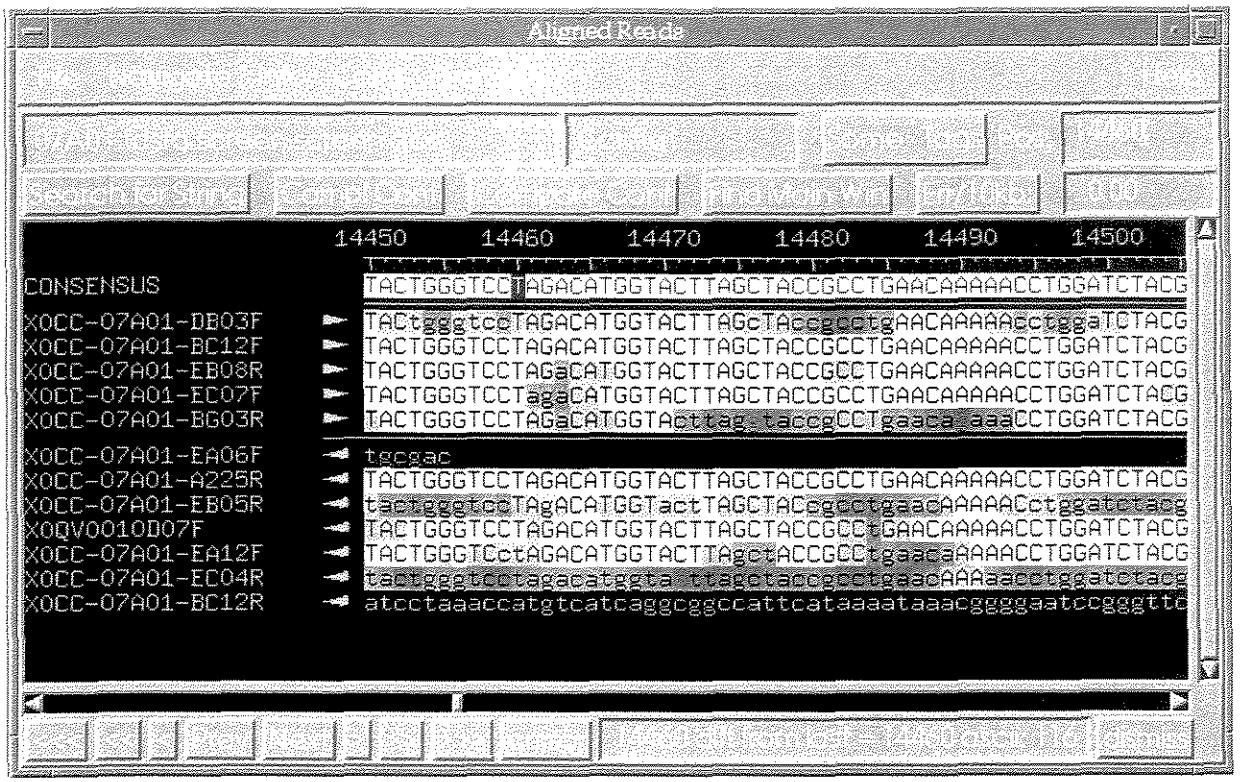


Figura 4.4: Janela do Consed mostrando uma montagem.

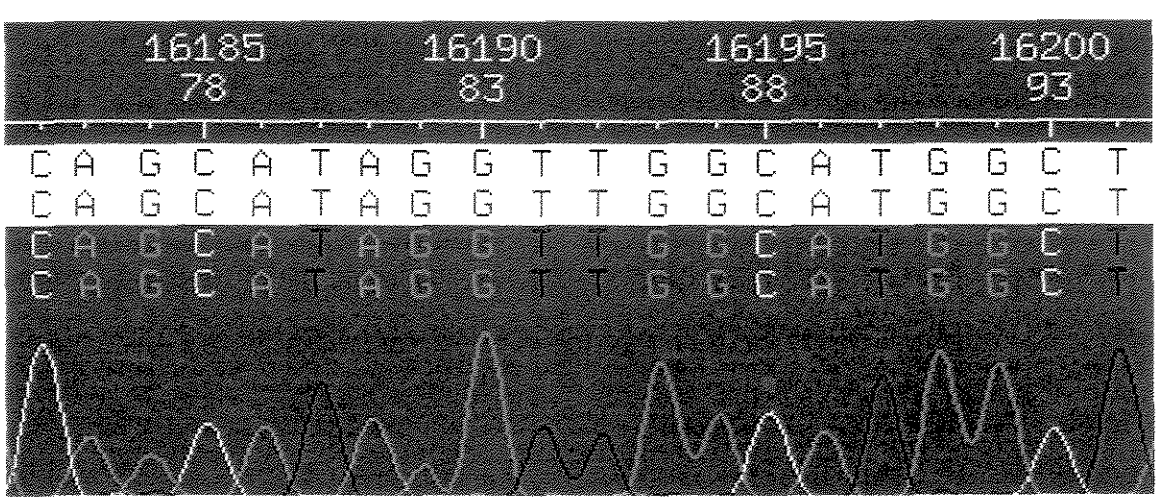


Figura 4.5: Visualisação parcial de um cromatograma no Consed.

4.3 Pipeline de Montagem do LBI

Para armazenar, organizar e analisar o montante de seqüências produzidas pelos laboratórios de seqüenciamento no projeto *Xylella*, foi montado um sistema de software composto por um banco de dados de seqüências e vários programas.

O banco de dados de seqüências é baseado em arquivos, mais especificamente em *flat files* do sistema operacional Unix. A organização da estrutura desse banco de dados e os mecanismos usados para manter sua integridade foram implementados nos próprios programas, principalmente naqueles encarregados pelo recebimento das seqüências, ou seja, pelo programa de submissão de cromatogramas e pelo programa de submissão de seqüências montadas, descritos abaixo.

Os programas de submissão juntamente com o montador geral formam o pipeline geral de montagem, como ilustrado na Figura 4.6. O fluxo de dados ocorre na seguinte ordem. Primeiramente, os laboratórios de seqüenciamento enviam suas seqüências para serem analisadas e armazenadas pelo sistema, que aceita ou rejeita tais seqüências, enviando um relatório ao laboratório submissor. Essas seqüências podem ser os reads na forma de cromatogramas, e no caso de seqüências montadas, é enviada a lista de reads (cujos cromatogramas já devem estar depositados no sistema) a ser usada pelo sistema para montar a seqüência e assim analisá-la. Em seguida, conforme o depósito de novas seqüências no banco, o sistema realiza a montagem geral usando os reads e as seqüências montadas, gerando então os contigs.

Os programas foram escritos utilizando a linguagem de programação Perl [38]. Os programas de submissão provêem uma interface WEB baseada em scripts CGI (*Common Gateway Interface*) [14]. Além dos três programas principais citados acima, há vários outros programas auxiliares, relacionados a geração de relatórios WEB para os dados de montagem e relatórios sobre a contabilização das seqüências totais e de cada laboratório, além de programas de consulta e recuperação das seqüências, que não serão discutidos aqui.

Nas seções abaixo são apresentados e detalhados os programas principais do sistema. Mas antes, será apresentado como é a organização das seqüências.

4.3.1 Organização e Caracterização das Seqüências

Os dados de seqüências estão organizados basicamente em três conjuntos: os reads, as seqüências montadas e os contigs. Os reads são os objetos principais de armazenamento do sistema. Como visto na Seção 2.6.3, um read é resultado do seqüenciamento de um fragmento ou inserto (apesar de serem usados na prática indistintamente, os termos fragmento e inserto têm uma ligeira diferença: um fragmento é um pedaço qualquer de DNA, e um inserto é um fragmento que foi clonado através da inserção em vetores). Com

exceção dos fragmentos com comprimento menor do que 800 bp, todos os reads se referem a pontas de fragmento. Os reads são caracterizados segundo o tipo de clonagem de seus fragmentos, sendo chamados de reads de plasmídeo, reads de cosmídeo e reads de fago.

Os fragmentos inseridos em plasmídeos foram obtidos a partir da molécula original pelo método de shotgun, e têm comprimento entre 800 bp a 4.5 kb. Por isso, reads oriundos desses fragmentos foram também chamados de reads de shotgun (além de reads de plasmídeo). Fragmentos inseridos em cosmídeos têm comprimento entre 35 a 45 kb. Os fragmentos inseridos em fago têm comprimento até 25 kb e foram amostrados no projeto *Xylella* com o propósito principal ajudar no fechamento de buracos.

Seqüência montada é o objeto do sistema que corresponde à seqüência completa de um fragmento, obtida pela montagem de seus sub-fragmentos (na verdade, os reads desses sub-fragmentos). Uma seqüência montada pode ser de um fragmento inserido em cosmídeo, em plasmídeo ou em fago. Para os dois últimos vetores, as respectivas seqüências são referenciadas como GFS's (*gap filling sequences*), pois o seqüenciamento completo de seus fragmentos está relacionado ao fechamento de buracos. Reads dos sub-fragmentos desses fragmentos são chamados de reads de GFS.

Os contigs são objetos do sistema que têm como principais atributos a sua seqüência consenso e a lista de seqüências (reads e seqüências montadas) que o constituem. Além disso, um contig tem associado consigo um alinhamento e a disposição em coordenadas das suas seqüências.

Assim como acontece na prática, usaremos os termos plasmídeo, cosmídeo e fago para referir-se aos fragmentos inseridos nesses vetores.

4.3.2 Submissão de Cromatogramas

O programa de submissão de cromatogramas[1], ou submissão de reads, pode ser dividido em três etapas: recepção, verificação e gravação dos cromatogramas no banco de reads. Embora haja uma distinção entre cromatograma e read, trataremos os dois termos indistintamente. Na primeira etapa, o programa recebe dos laboratórios um lote de cromatogramas. Esses lotes são rotulados e guardados em uma área temporária (*pool*), de onde são enviados para a etapa de verificação. São desconsiderados os lotes vazios ou incompletos, que em geral são decorrentes de problemas no envio. A área temporária permite que haja processos paralelos de lotes de cromatogramas, possibilitando servir diferentes laboratórios simultaneamente, e também o envio de mais de um lote por laboratório.

A etapa de verificação visa analisar cada cromatograma em um lote, procurando identificar problemas e extraindo informações sobre a sua seqüência. Os principais pontos problemáticos avaliados são os erros de nomenclatura, a duplicação de reads, erros de

[1]Desenvolvido por João Meidanis e João Carlos Setubal

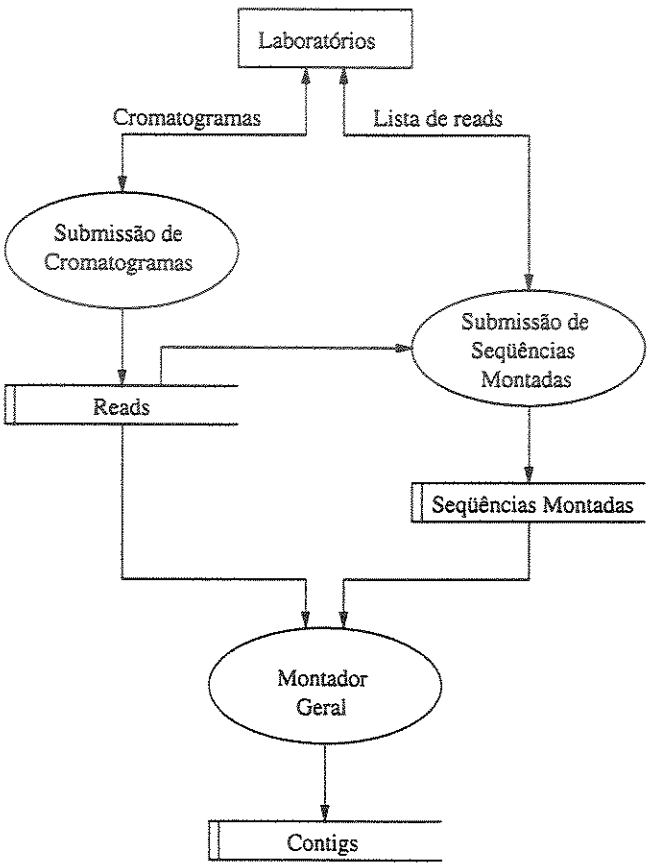


Figura 4.6: Principais processos e bancos de seqüências do pipeline de montagem. As elipses representam os processos, os retângulos abertos representam os repositório de dados, o retângulo fechado representa uma entidade externa ao sistema e as linhas orientadas indicam o fluxo dos dados.

tracking e contaminações.

Os reads possuem uma identificação única, estabelecida por uma nomenclatura padrão, que também os diferencia conforme seu tipo (cosmídeo, plasmídeo e fago). Além disso, permite que outras informações sejam embutidas no nome de um read. Por exemplo, o nome `X0JJ001204R` refere-se a um read de plasmídeo. Analisando-se o nome da esquerda para direita, temos: a letra `X` (de *Xylella*) faz parte de todos os reads; o digito `0` (zero) indica o código dado a fonte (cepa) da *Xylella*, que pode ser 0 (zero) ou 1; `JJ` é o código do laboratório seqüenciador; `00` (zero, zero) é o código da biblioteca; `12` é o número da placa; `04` é a posição do clone na placa; `R` é a direção de seqüenciamento do clone (F para *forward* e R para *reverse*). O nome `X0QR-07H03-AA01F` refere-se a um read de um sub-clone do cosmídeo identificado por `07H03`, seqüenciado pelo laboratório `QR`, na direção forward (F). O sub-clone é identificado pelo códido `AA01`, onde a primeira letra `A` indica o código da placa e a cadeia `A01` indica sua posição na placa. Os reads de fago são nominados da mesma maneira que os reads de cosmídeo, trocando-se apenas a identificação do cosmídeo pela identificação do fago.

Como a quantidade de reads em um laboratório é bastante grande, e sua identificação é feita praticamente de forma manual, os nomes dos reads estão sujeitos a erros. Alguns erros comuns no nome dos reads são consertados automaticamente pelo programa. Demais reads cujo nome não segue a nomenclatura são rejeitados e não são inseridos no banco de reads.

A duplicação de reads acontece quando dois reads têm o mesmo nome ou a mesma identificação. No banco de seqüências não deve haver esse tipo de problema, ou seja, caso um nome esteja sendo usado pelo banco para identificar um determinado read, um outro read com o mesmo nome não pode ser inserido no banco. Esse tipo de problema acontece devido a erros de atribuição de nomes, erros de digitação, ou mesmo pela atribuição do mesmo nome para uma re-leitura do mesmo clone. Uma re-leitura (que requer um re-seqüenciamento), pode ocorrer uma ou mais vezes, visando melhorar a qualidade das bases do read. Em suma, reads que já tiverem seu nome usado pelo banco de reads devem ser rejeitados.

Um ponto importante sobre o problema de duplicação de reads está no fato de haver nomes corretos sintaticamente, mas errados no conteúdo. Esse problema é difícil de ser detectado e tem conseqüências na contabilização dos reads e no processo de montagem. Para efeito de contabilização, os reads em questão deveriam ser considerados apenas uma vez, visto que o conteúdo é o mesmo.

No caso da montagem, pode haver inconsistência entre a informação inferida entre as pontas e a sua real localização nos contigs. Ou seja, pelos seus nomes, as duas pontas de um mesmo inserto deveriam estar nos contigs a uma distância correspondente ao comprimento de tal inserto, mas pelo seu conteúdo, essa distância é completamente diferente.

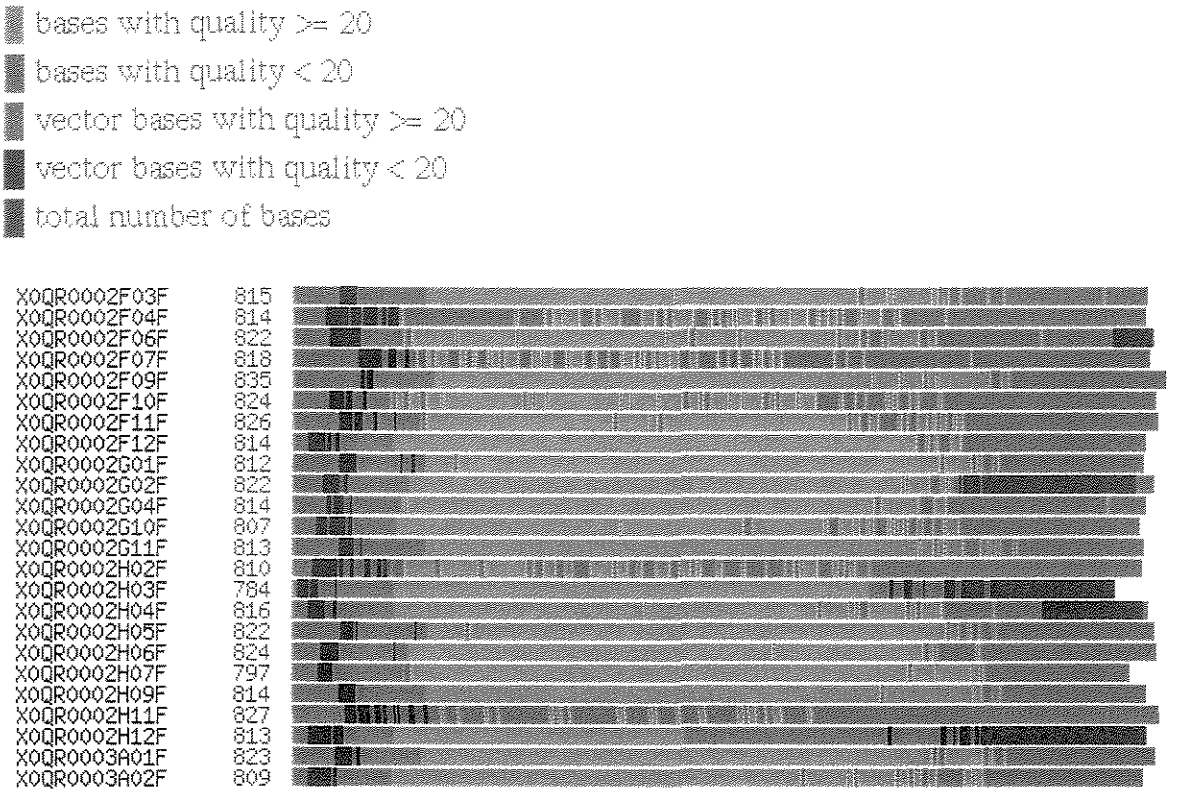


Figura 4.7: Versão gráfica de um relatório de submissão de reads. Nesse relatório podem ser visualizados as regiões de baixa (< 20) e alta (≥ 20) qualidade, e as regiões com e sem vetor.

Para extrair informações sobre as seqüências são usadas algumas ferramentas, que automaticamente identificam os erros de *tracking* e as contaminações. O principal problema de *tracking* são os cromatogramas sem dado útil e tal problema é detectado pelo *base-caller* phred. Já as contaminações são detectadas e mascaradas com o cross_match. Assim, após a execução dessas ferramentas, são obtidas a seqüência de bases dos reads, a qualidade de cada base e os segmentos que contém seqüências de vetor. As saídas da execução dessas ferramentas são analisadas pelo programa de submissão, que gera o relatório para o lote processado, contendo para cada read as informações encontradas. Um exemplo desse relatório está no Apêndice B.1. Uma versão gráfica do relatório de submissão de reads também é gerada. Nesse tipo de relatório, a qualidade geral dos reads submetidos é mais facilmente analisada, como pode ser visto na Figura 4.7.

Terminada a verificação, o próximo passo do programa é fazer o armazenamento dos

reads. Os cromatogramas mais outros dados gerados pela etapa de verificação são então inseridos adequadamente no banco de reads. No final do processo, o relatório gerado na etapa anterior é enviado ao laboratório submissor.

Um aspecto importante a ser levantado é como os lotes transitam da primeira para a segunda etapa. O que acontece é que somente um lote é processado por vez. Isso evita danos de integridade no banco de reads. Um exemplo de ocorrência danosa, sem essa exclusividade, acontece quando lotes distintos que tenham reads em comum sejam processados paralelamente. Isso ocorre devido a falta de controle dos laboratórios sobre a submissão de cromatogramas, colocando em lotes distintos os mesmos cromatogramas. Outro exemplo mais comum, acontece quando o mesmo lote é submetido mais de uma vez. Isso acontece quando o submissor (inadvertidamente) aperta o botão de submissão mais de uma vez em um intervalo muito curto de tempo, fazendo com que o browser ative mais de um processo. Assim, se um dos processos consulta o banco de reads no intervalo entre a consulta e a escrita do outro processo, os dois irão escrever os mesmos reads no banco. Para tratar disso, o programa usa um mecanismo de *lock* de escrita exclusivo ao banco de reads. Os processos competem entre si para pegar o direito ao *lock*. O processo que toma a posse do *lock* segue para a etapa de verificação, impossibilitando os outros de fazerem o mesmo. Após a verificação, e o armazenamento dos dados das seqüências, o processo libera o *lock* para que demais lotes sejam processados.

4.3.3 Submissão de Seqüências Montadas

Seqüências montadas são aquelas que são constituídas pela montagem de seus fragmentos seqüenciados, visto que não podem ser seqüenciadas diretamente. Fazem parte do conjunto de seqüências montadas os cosmídeos e as GFS's. Os laboratórios responsáveis por tais seqüências encarregavam-se não apenas da fragmentação e do seqüenciamento, mas também de fazer a sua montagem, submetendo então o produto para ser analisado pelo programa de submissão.

O programa de submissão de seqüências montadas[2] divide-se basicamente nas mesmas etapas que o de submissão de reads, apesar de possuirem características próprias. A etapa de recepção de seqüências montadas é relativamente simples. O laboratório passa para o programa uma lista contendo os nomes dos reads usados na montagem da seqüência. Cada uma das submissões é rotulada e colocada em uma área temporária, para ser então repassada a etapa posterior.

Nesse caso não há exigência para que as submissões sejam verificadas seqüencialmente, a menos no instante de alterar o banco de seqüências montadas. Esse programa utiliza-se também de um mecanismo de *lock* de escrita, permitindo somente um processo por vez

[2]Desenvolvido em parceria com Marcos Renato R. Araujo

escrever nesse banco. A etapa da verificação consiste de três passos:

1. verificar se os todos os reads estão depositados;
2. montar localmente a seqüência;
3. certificar a seqüência.

Para que o programa possa fazer a montagem local, todos os reads usados pelo laboratório na montagem da seqüência devem estar no banco de reads. Caso haja reads na lista que não estejam no banco de reads, a submissão é cancelada e o laboratório submissor notificado. Utiliza-se para fazer a montagem o phrap (Seção 4.2.1). Após seu término, a montagem é analisada pelo programa de submissão de maneira a certificar a seqüência obtida. Essa certificação é realizada sobre os seguintes critérios de qualidade:

1. todas as bases do consenso devem ter valor de qualidade 20 (pelo menos) dado pelo programa phred;
2. não deve haver discrepâncias de alta qualidade (≥ 40) não justificadas entre a base do consenso e as bases dos reads (qualidade dada pelo programa phred) - o critério automático de justificação é: a base discrepante não pode estar a mais do que 10 bp do extremo do read que a contém;
3. cada base do consenso deve ser confirmada por pelo menos uma seqüência em cada fita;
4. a taxa de erro global estimada deve ser menor que 1 a cada 10000 bases; a taxa de erro global é calculada pela fórmula $(\sum_i^L pe_i)/L \times 10000$, onde pe_i é a probabilidade de erro da base i e L é o comprimento da seqüência consenso.

As seqüências que satisfazem todos esses critérios são consideradas finalizadas (*finished*). Além de *finished*, são também atribuídas às seqüências outros estados: não submetido (*not submitted*), submetido (*submitted*), um contig (*one contig*) e próximo de finalizado (*close to finished*). Uma seqüência está *close to finished* se a soma de bases com problemas segundo os itens 1, 2, 3 acima é menor que 0.1% das bases do consenso, mas não é igual a zero, pois nesse caso o estado passa a ser (*finished*).

Como produto dessa análise é gerado um relatório, onde são apontados os erros que impossibilitaram a certificação da seqüência, ou onde é notificado que a seqüência foi certificada. Um exemplo do relatório gerado para uma seqüência de cosmídeo é apresentado no Apêndice B.2.

Por último é feita a inserção ou alteração do banco de seqüências montadas. Além da seqüência propriamente dita, são guardados o seu estado quanto a certificação, o relatório e a lista de reads usados.

O programa permite que uma mesma seqüência seja submetida mais de uma vez, até que ela tenha sido certificada. A cada nova re-submissão, são realizadas atualizações do banco de seqüências montadas, mantendo-se as versões anteriores.

Para acompanhar o andamento da finalização das seqüências montadas (cosmídeos e GFS) foram elaborados sumários, em forma de tabela, contendo algumas informações básicas sobre cada seqüência. O sumário de cosmídeos é mostrado no Apêndice B.3.

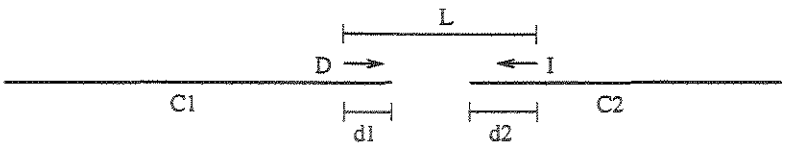
4.3.4 Montagem Geral

Antes de descrever o programa montador geral, apresentaremos brevemente a seguir, a metodologia de montagem seguida no Projeto *Xylella*.

No projeto *Xylella*, foi utilizado o software phred/phrap/consed (Seção 4.2.1) para fazer montagem da seqüência genômica. A montagem (chamada montagem geral) dividiu-se em duas fases. A primeira fase estava focada sobre montagem dos fragmentos inseridos em cosmídeos juntamente com os reads de shotgun, e a segunda estava focada no fechamento dos buracos.

Na primeira fase, a medida que novos reads de shotgun eram depositados no banco, e as seqüências montadas de cosmídeos tornavam-se finalizadas, eles eram inseridos na montagem. Esta fase continuou até a finalização de todos os cosmídeos escolhidos para serem montados. O ideal seria que ao final desta fase, os reads de shotgun (seqüências mais curtas) pudessem juntar os cosmídeos (seqüências mais longas), fechando-se a seqüência completa.

Mesmo com a finalização da amostra de cosmídeos escolhidos, restaram ainda muitos contigs, passando-se então para a fase seguinte, visando a identificação e o preenchimento dos buracos. Identificar um buraco é saber que dois contigs se ligam, embora a seqüência que os une não esteja presente na montagem. A maneira prática para determinar essa informação de ligação entre dois contigs, é fazer a verificação da localização, orientação e distância das pontas de insertos não seqüenciados totalmente. Ou seja, estando uma das pontas (D) a uma distância $d1$ da extremidade de um contig C1, e estando outra ponta (I) a uma distância $d2$ da extremidade de outro contig ($C2$), com D e I apontando para fora de seus respectivos contigs, é possível confirmar a ligação desses dois contigs com a validade da equação $d1 + d2 \leq L$, onde L é o comprimento do inserto. Veja a figura a seguir.



Identificado o buraco dessa maneira, basta então fazer o seqüenciamento completo do inserto, e colocar a sua seqüência (uma GFS) na montagem para fechar (ou preencher) o buraco respectivo.

Após exaustivo seqüenciamento de diversas pontas de insertos das bibliotecas de clones disponíveis, muitos buracos foram identificados com a técnica acima. No entanto, ainda restaram os buracos físicos (Seção 4.1). Foram usadas então algumas abordagens possíveis para fechar tais buracos, em particular foi utilizada a técnica de *primer walking*. O princípio dessa abordagem é tentar prolongar os contigs a partir de suas extremidades, com a adição de novos reads, intencionando-se que em um momento ocorrerá uma sobreposição com outro contig. Esses novos reads são obtidos através de PCR (Seção 2.6.2.3) com a utilização de primers apropriados, sintetizados a partir de seqüências do contig em questão.

Montador Geral

O programa montador geral[3] realiza três etapas básicas:

1. seleção das seqüências;
2. montagem das seqüências;
3. análises pós-montagem.

Na primeira etapa o programa faz uma filtragem entre as seqüências que estão no banco, deixando algumas de fora da montagem. Exemplos dessas seqüências são todos os reads de shotgun oriundos de uma dada biblioteca, ou reads específicos, chamados de reads excluídos, que causam problemas na montagem. Para as seqüências montadas existe também uma seleção. No caso, são incluídas na montagem somente seqüências que estejam finalizadas ou próximas deste estado. Além disso, o programa possibilita a inclusão de seqüências em outros estados, bastando que ser forneça o parâmetro correspondente.

Selecionadas as seqüências, o programa realiza a montagem propriamente dita, executando o software phred/phrap. Esse processo leva algumas horas. Embora o tempo gasto de uma montagem esteja diretamente relacionado ao número e ao comprimento das seqüências de entrada, existem casos particulares que influenciam no tempo de execução.

[3]Desenvolvido por Marília D. V. Braga, com módulos específicos desenvolvidos por Marcos Renato R. Araujo e Renato F. F. Werneck

Um exemplo disso é quando existem repetições na própria seqüência completa sendo montada, e assim entre as seqüências no conjunto de entrada da montaem, o que contribui para um aumento no tempo de execução, para o caso desse software.

Por fim, na última etapa o programa realiza algumas análises sobre a montagem, gerando alguns relatórios e figuras. Entre os resultados, citamos o relatório geral da montagem, os mapas dos contigs, o relatório sobre a localização das pontas e o sumário de contigs.

O relatório geral de montagem é um resumo da montagem realizada, contendo o número de contigs, o comprimento do maior contig, e uma tabela com o número de contigs que caem em uma determinada faixa de comprimento (10, 30, 70, 100 e 200 kb). Juntamente com esse resumo, é criado um sumário contendo a lista dos contigs, mostrando para cada um, o seu comprimento, o número de reads de shotgun, a lista de cosmídeos com a sua localização, além de outros dados.

Os mapas (físicos) dos contigs apresentam a disposição das seqüências dos cosmídeos, das pontas dos cosmídeos e das pontas dos fagos (Figura 4.8) dentro do contig. No mapa, os cosmídeos são representados por barras coloridas, onde cada cor indica o seu estado. As pontas dos cosmídeos e dos fagos são representadas por figuras semelhantes aos sinais de maior (>) e menor (<). Se as duas pontas de um mesmo inserto estão contidas no contig, então uma linha é desenhada para ligá-las entre si. Ao contrário, se apenas uma ponta está presente, a identificação do inserto na figura vem acompanhada de um ponto de interrogação. Um círculo envolvendo uma ponta indica que ela pode estar situada em uma região repetida do contig. Essa informação é obtida da saída do programa phrap. Além do propósito de identificação de buracos, as pontas desses insertos foram usadas também para confirmar a integridade e orientação dos contigs.

A análise sobre a localização das pontas visa identificar quais pontas de insertos (principalmente dos plasmídeos) estão nas extremidades dos contigs (a menos de 2 kb). O resultado é dividido em duas tabelas: uma contém a lista de reads em que apenas uma ponta foi depositada no banco, e outra contém a lista de reads em que as duas pontas foram depositadas. As tabelas eram analisadas manualmente, procurando identificar pontas ou insertos candidatos para seqüenciamento. Na primeira tabela, analisava-se quais pontas deveriam ter outra ponta seqüenciada baseada em contigs sólidos, que basicamente eram aqueles que continham seqüências de cosmídeos. De forma análoga, a segunda tabela era analisada para confirmar o seqüenciamento do inserto.

Para acompanhar o andamento da montagem geral, foi elaborado um sumário para os contigs, mostrado no Apêndice B.4.

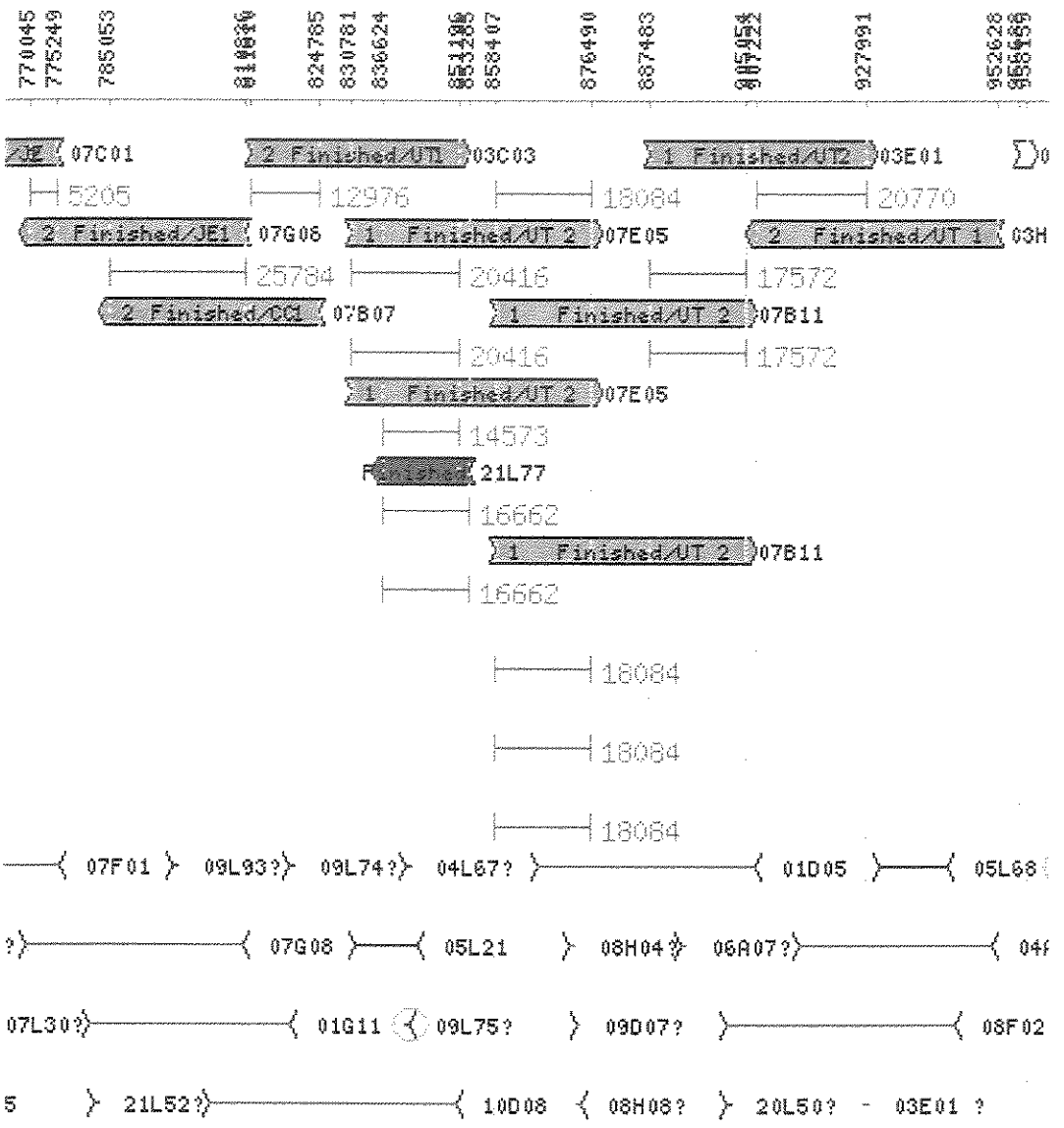


Figura 4.8: Corte do mapa físico construído a partir da montagem.

Capítulo 5

Anotação

Este capítulo apresenta os conceitos gerais sobre anotação e enfatiza o processo aplicado para genomas de procariotos, utilizando como exemplo o sistema de anotação usado no projeto *Xanthomonas*. São apresentados a seguir, os objetivos e a descrição dos problemas que envolvem a anotação de maneira geral, a descrição de algumas das principais ferramentas usadas no projeto e, por fim é descrito o sistema de anotação.

5.1 Objetivos e Descrição dos Problemas

O processo de anotação tem por finalidade identificar os genes em uma seqüência de DNA e caracterizar suas funções bioquímicas ou estruturais, possilitando o conhecimento e entendimento do componentes funcionais básicos de um organismo. Quando se fala em anotação deve-se ter claro a distinção entre o domínio dos procariotos e domínio dos eucariotos (Seção 2), que apresentam diferenças na sua organização celular e nos processos de transcrição e de tradução. Essas diferenças refletem no processo de anotação, havendo metodologias e ferramentas específicas aplicadas nos genomas de cada um dos domínios. Neste trabalho, será abordado somente o caso dos procariotos.

A anotação de maneira geral pode ser caracterizada como possuindo dois níveis: anotação de genes e anotação de interações gênicas. No primeiro nível são identificados e caracterizados os genes, sendo a análise baseada em seqüências. No segundo nível, com o conhecimento das funções dos genes, o objetivo recai em estudar como as proteínas interagem entre si, formando o complexo sistema biológico de um ser vivo. Dentre as interações gênicas, as vias metabólicas são as mais conhecidas e estudadas. A seguir são apresentados cada um desses níveis.

5.1.1 Anotação de Genes

Para fazer a anotação dos genes em uma nova seqüência de DNA, é preciso em primeiro lugar identificá-los. Tratando-se de um cadeia de caracteres (com alfabeto de quatro letras), identificar um gene corresponde definir suas posições de início e de fim dentro dessa cadeia. Apesar dessa visão um pouco simplista, existem fatores relacionados ao processo de expressão gênica que não permitem saber com exatidão onde começa e onde termina um gene. De maneira geral um gene é composto por uma região codificadora, que é traduzida na seqüência de aminoácidos, e por regiões que flanqueiam a região codificadora e são encarregadas de ativar ou inibir a expressão do gene [39], como visto na Seção 2.5. Há também genes que não codificam proteínas e são usados no processo de síntese de proteínas, como os tRNA's e rRNA's.

As ferramentas computacionais de identificação de genes que codificam proteína são em geral baseados na procura de ORFs, com propriedades estatísticas especiais. Exemplos desses programas são Glimmer [31, 12] e Genmark [8]. Esses dois programas usam uma abordagem baseada em modelos estatísticos. Uma breve descrição dessa abordagem é apresentada juntamente com a descrição do programa Glimmer (usado no projeto *Xanthomonas*) na Seção 5.2.2.

Programas para achar genes de RNA são de certa forma menos complicados do que aqueles que procuram genes que codificam proteína, pois possuem mecanismos de transcrição mais simples, como é o caso de ferramentas para achar tRNAs. No projeto *Xanthomonas* foi usado o programa tRNAscan-SE [25] para achar genes de tRNA, que é descrito abaixo.

Identificados os genes, é preciso ainda saber qual o seu papel no organismo, ou seja, caracterizar a função bioquímica da proteína codificada. Para se obter efetivamente essa informação são necessários experimentos laboratoriais, que dispendem bastante esforço e levam muito tempo para serem finalizados. Uma outra alternativa de caracterização, mais rápida embora hipotética e usada na prática, é procurar em bancos de proteínas conhecidas alguma seqüência que seja similar à nova seqüência. Caso seja encontrada, a hipótese gerada é de que o gene de interesse tem a mesma função do que o gene do banco com o qual tem similaridade. O programa popularmente usado para esse fim é o BLAST. A comparação também pode ser feita contra bancos de famílias de proteínas, que em geral são mais precisas ao revelar a função da proteína. Ver abaixo, na Seção 5.2.4 alguns dos bancos de seqüências e de famílias de seqüências usadas na classificação da função de proteínas desconhecidas.

5.1.2 Anotação de Vias Metabólicas

Após catalogar os genes e seus produtos, o passo seguinte fixa-se no entendendimento de como as proteínas codificadas pelos genes funcionam em um organismo vivo. As proteínas correspondem aos componentes individuais do sistema biológico, e interagem entre si ou com outras moléculas formando simples ou complexas redes de interações. Na área de computação esse campo de estudo ainda é pouco investigado, não havendo ferramentas de software que permitam a reconstrução automática e confiável de vias metabólicas. As tentativas nesse sentido concentram-se em catalogar computacionalmente o conhecimento e informações adquiridas com os genomas seqüenciados, formando então uma base mais sólida para serem desenvolvidas e aplicadas ferramentas. Exemplos disso são o KEGG (*Kyoto Encyclopedia of Genes and Genomes*) [29, 20], um banco de dados de vias metabólicas de propósito mais geral, e o Ecocyc [21, 22] que é específico para a bactéria *Escherichia coli*, e contém algumas ferramentas para consultar e visualizar dados de metabolismos.

5.2 Ferramentas

5.2.1 BLAST

BLAST [1, 2, 43] (*Basic Local Alignment Search Tool*), já mencionado na Seção 4.2.1, é uma ferramenta bastante conhecida e usada, que visa comparar seqüências quanto a sua similaridade, em particular comparar uma seqüência de consulta contra um banco de seqüências. A saída do programa apresenta uma lista contendo os *hits* encontrados e em seguida cada um dos respectivos alinhamentos, ou acusa se não foi encontrado nenhum *hit*. Cada linha da lista contém a identificação junto com uma breve descrição da seqüência alvo, a pontuação (*Score*) do alinhamento e o *E value*, valor que indica a significância estatística do *hit*. A pontuação reportada em *bits* corresponde a uma pontuação normalizada, que leva em conta, por exemplo, o tamanho do banco e o sistema de pontuação empregado, permitindo que haja uma comparação entre consultas feitas contra diferentes bancos de seqüências ou consultas feitas usando diferentes sistemas de pontuação. O *E value* corresponde ao número estimado de alinhamentos com pontuação igual ou superior ao do *hit* que poderiam ter sido encontrados ao acaso. Assim, quanto menor esse valor, mais estatisticamente significativo é o alinhamento obtido. Tomando o primeiro *hit* no exemplo abaixo, temos um *E value* de 10^{-122} ou $1/10^{122}$.

```
                                                          Score     E
Sequences producing significant alignments:               (bits)  Value

gi|544173|sp|P35891|DNAA_SALTY CHROMOSOMAL REPLICATION INITIATO...   439  e-122
```

```
gi|118708|sp|P22837|DNAA_PROMI CHROMOSOMAL REPLICATION INITIATO...   434  e-121
gi|72915|pir||IQECDA dnaA protein - Escherichia coli                  433  e-120
gi|399394|sp|P03004|DNAA_ECOLI CHROMOSOMAL REPLICATION INITIATO...   433  e-120
gi|1352278|sp|P29440|DNAA_SERMA CHROMOSOMAL REPLICATION INITIAT...   432  e-120
gi|1706462|sp|P49996|DNAA_VIBHA CHROMOSOMAL REPLICATION INITIAT...   428  e-119
gi|290550 (L10328) dnaA (CG Site No. 851) [Escherichia coli]         422  e-117
gi|231998|sp|P29434|DNAA_BUCAP CHROMOSOMAL REPLICATION INITIATO...   409  e-113
gi|144149 (M80817) dnaA [Buchnera aphidicola]                        390  e-107
gi|1169364|sp|P43742|DNAA_HAEIN CHROMOSOMAL REPLICATION INITIAT...   389  e-107
...
```

O BLAST possui variantes, ou seja, programas distintos que são usados de acordo com os tipos de seqüências sendo comparados. As variantes são:

- blastn: compara seqüências de nucleotídeos contra seqüências de nucleotídeos;
- blastp: compara seqüências de aminoácidos contra seqüências de aminoácidos;
- blastx: traduz seqüências de nucleotídeos (nos seis quadros de leitura) e compara contra seqüências de aminoácidos;
- tblastn: compara seqüências de aminoácidos contra seqüências traduzidas de nucleotídeos (nos seis quadros de leitura);
- tblastx: traduz seqüências de nucleotídeos (nos seis quadros de leitura) e compara contra seqüências traduzidas de nucleotídeos (nos seis quadros de leitura);

Essas variantes têm aplicações distintas, como por exemplo, o uso do blastx na identificação de genes em uma nova seqüência genômica, e o uso do blastp para comparar proteínas hipotetizadas por programas de identificação de genes (como o Glimmer) contra um banco de proteínas conhecidas.

BLAST é uma heurística que tenta otimizar velocidade do cálculo da medida de similaridade. Diferente da solução ótima que usa programação dinâmica, o BLAST possui um algoritmo diferente, que segue os seguintes passos:

1. procura palavras no banco de seqüências com um comprimento definido (W) e que tenham pontuação pelo menos T quando casadas com alguma palavra de mesmo comprimento da seqüência de consulta;
2. a partir de cada uma das palavras encontradas no passo anterior, estende o alinhamento em ambas as direções até que a pontuação obtida caia abaixo de um valor definido pela diferença entre a máxima pontuação conseguida e um fator limitante X.

3. reporta os alinhamentos que tenham uma pontuação mínima M.

Das variáveis usadas pelo BLAST e que podem ser mudadas pelo usuário, vale destacar o parâmetro T, que permite fazer um balanço entre rapidez e sensibilidade [2]. Um valor mais alto para T aumenta a rapidez, embora também aumente a probabilidade de perder similaridades fracas. O tempo de execução do programa é proporcional ao produto entre o comprimento da seqüência de consulta e tamanho do banco alvo.

Em sua interface WEB o BLAST reporta uma saída gráfica mostrando os alinhamentos encontrados, como exemplificado na Figura 5.1. A seqüência de consulta é representada pela barra superior, na qual é exibida a escala de tamanho e abaixo dela, seguem os alinhamentos representados por linhas coloridas, onde cada cor indica a faixa de pontuação do alinhamento. O posicionamento do ponteiro do mouse sobre uma linha, faz com que sejam exibidas na pequena janela acima a descrição e a pontuação do respectivo alinhamento. O clique sobre um linha, exibe o alinhamento correspondente entre a seqüência de consulta e a seqüência alvo.

5.2.2 Glimmer

Uma das abordagens usadas para identificar genes baseia-se em detectar individuamente alguns de seus elementos. Existem métodos que procuram detectar pequenos sítios funcionais (sinais), como por exemplo, os codons de início e de parada, promotores e terminadores de transcrição e sítios de ligação ribossomal (RBS - *ribossomal binding site*). Há também métodos que tentam reconhecer as outras partes do gene, como os *introns* e os *exons* (conteúdo). As metodologias aplicadas na análise de seqüências de procariotos e eucariotos têm em geral características diferentes, devido as diferenças que existem no processo de transcrição dos genes e tradução de proteínas entre esses dois domínios, como apresentado na Seção 2. Dentre os métodos, os que mais têm se destacado são os que estão relacionados a análise de conteúdo dos genes, em particular os que procuram identificar regiões codificadoras ou ORFs.

O programa Glimmer [31, 12] é usado para idenficar ORFs em um genoma. Para isso, o programa realiza um série de passos. Primeiramente, o programa extrai todas as ORFs longas (o valor padrão de comprimento mínimo é 500 bp). Tais ORFs, com grande probabilidade são genes autênticos, pois genomas de procariotos, conforme já discutido, em geral têm regiões intergênicas pequenas.

O passo seguinte é montar um modelo oculto de Markov baseado nas ORFs longas. A idéia é que a grande maioria dos genes de um organismo obedece a um mesmo padrão. O modelo oculto de Markov captura a essência desse padrão.

Em seguida, o genoma é varrido, e cada ORF é analisada sob o ponto de vista do modelo construído, recebendo uma pontuação que reflete a probabilidade daquela ORF

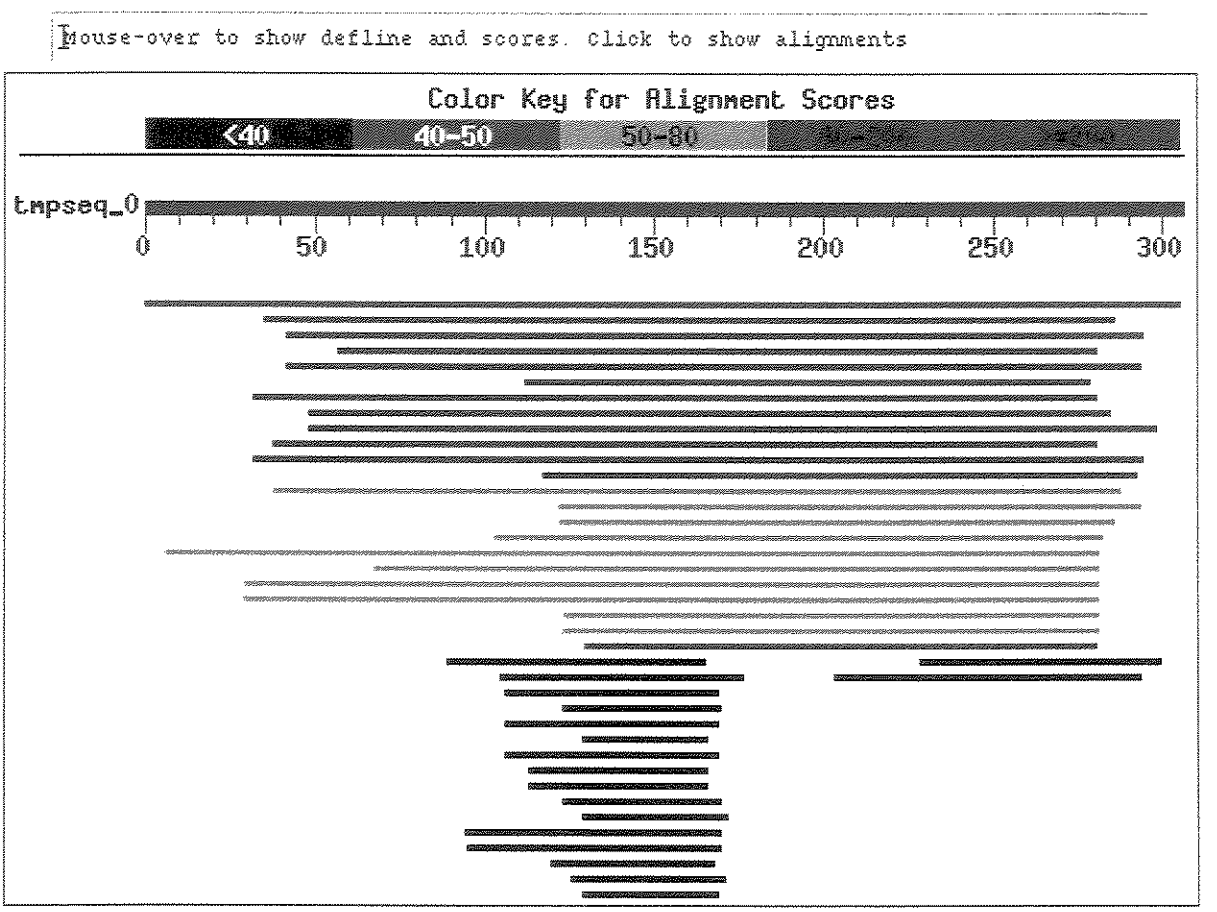


Figura 5.1: Saída gráfica do BLAST.

ser de fato um gene. O programa então apresenta as ORFs com pontuação acima de um valor mínimo. O Glimmer encontra corretamente cerca de 98% dos genes em um genoma.

5.2.3 tRNAscan-SE

O programa tRNAscan-SE é usado para predizer genes de tRNA (RNA transportador) (Seção 2.5.2) em uma seqüência de DNA ou RNA, e pode ser usado para qualquer seqüência de eucarioto quanto de procarioto, bastando escolher os devidos parâmetros. Na verdade o programa é um script escrito em Perl [38] que combina três programas independentes e gerencia o fluxo entre eles, de maneira a usar o que cada programa tem de mais forte quanto a detecção de genes de tRNA, obtendo assim um resultado muito mais confiável se comparado com os resultados individuais de cada um dos três programas. O primeiro desses programas é o *tRNAscan*, que embora seja capaz de identificar quase que 100% dos tRNAs em uma seqüência, gera muitos falsos positivos. O segundo programa, conhecido como *EufindtRNA* possui algumas características a mais que lhe permitem identificar outros genes diferentes do primeiro, mas também gera muitos falsos positivos. O terceiro programa, chamado *covels*, é o que apresenta os melhores resultados quanto a taxa de falsos positivos. Porém o programa consome muito processamento, tornando-se muito lento para seqüências mais longas. O tRNAscan-SE organiza o fluxo entre os três programas da seguinte forma. Como são mais rápidos em relação ao *covels*, os programas *tRNAscan* e o *EufintRNA* são usados primeiramente para gerar uma lista de tRNA's candidatos. Esta saída é então modificada para otimizar o processamento do programa *covels*, que recebe cada um dos tRNA's candidatos e gera uma lista final, onde é minimizado o número de falsos positivos.

5.2.4 Bancos de Seqüências Públicos

O principal banco de seqüências conhecido é o Genbank [7], construído e mantido pelo NCBI (*National Center for Biotechnology Information*), que tem como colaboradores o EMBL, situado na Europa, e o DDBJ, situado no Japão. Esses três sites trocam dados diariamente, assegurando a mesma coleção de sequências. Atualmente, mais que 55000 diferentes espécies estão representadas no Genbank, em aproximadamente 3.4 bilhões de bases de nucleotídeos. As seqüências podem ser submetidas ao NCBI através dos programas BankIt ou Sequin. O primeiro é uma ferramenta de submissão de dados baseado em WEB, e serve para submissões de poucas seqüências. O Sequin é um programa de submissão que roda em várias plataformas, mais robusto e com diversas capacidades, e pode tratar tanto com seqüências simples quanto seqüências muito longas, uma característica ausente em outros programas. Com a submissão, as seqüências passam a ter uma identificação única, ou seja um código de acesso (*accession code*). O principal meio de consulta

e recuperação de dados de seqüências depositados do Genbank é feito via BLAST, que como visto na Seção 5.2.1, é usado para procurar seqüências similares à uma seqüência de consulta.

Além do Genbank, existem outros bancos de dados, que tentam organizar as seqüências sobre diferentes perspectivas, como por exemplo o SWISS-PROT - um banco específico de seqüências de proteínas, o PFAM [35, 4, 5] - um banco de famílias de domínios (que são regiões bem conservadas entre seqüências) e o COG [37] - um banco de seqüências de genes ortólogos. Genes ortólogos são aqueles derivados de um acestral comum, e que por isso, em geral, tem a mesma função.

A diferença entre comparar a seqüência de um gene contra um banco de dados como o Genbank, que contém diversas seqüências sem nenhuma informação de relação entre elas, e comparar a seqüência de um gene contra um banco de dados específico como o PFAM ou COG, é que no segundo caso os resultados obtidos são mais concisos e simplificam a predição funcional da seqüência de consulta [35].

5.3 Pipeline de Anotação do LBI

É comum que a anotação de um genoma comece somente após a fase de montagem, com a sua seqüência de DNA determinada, ou mesmo quando a montagem está próxima de seu final. Porém, no projeto *Xanthomonas* o processo de anotação dos genes começou bem antes do término da fase de montagem, uma metodologia herdada e aperfeiçoada do processo de anotação do projeto *Xylella*. Neste projeto, a análise dos genes foi feita inicialmente sobre as seqüências dos cosmídeos que tinham passado pela certificação. A escolha dessa abordagem foi baseada primeiramente no fato dessas seqüências serem consideradas sólidas, ou seja, que não teriam sua seqüência modificada. Posteriormente, com o mapeamento das seqüências dos cosmídeos nas seqüências dos contigs gerados pela montagem geral, a anotação dos cosmídeos foi migrada para os contigs, que passaram a ser as entidades principais de anotação. Assim, a cada nova montagem, os genes anotados nos contigs velhos eram migrados para os contigs novos, além de serem adicionados os novos genes encontrados nos contigs mais recentes. A partir dessa experiência, adotou-se no projeto *Xanthomonas* essa mesma metodologia na anotação do genoma.

Um lição importante tirada da anotação do projeto *Xylella* é que um contig não é uma boa entidade para anotação, devido às mudanças que ocorriam na seqüência dos contigs entre uma montagem e outra. No projeto *Xanthomonas*, a entidade sólida de anotação escolhida foi nomeada de *chunk*, pedaço ou segmento contíguo de um contig com um critério estrito de qualidade. Dependendo então da qualidade das bases da seqüência do contig, este pode ter um ou mais *chunks*.

Um esboço do sistema de anotação é apresentado na Figura 5.2. Seguindo o esque-

ma, o sistema recebe como entrada os contigs de uma montagem e extrai os *chunks* (ver Seção 5.3.2), que são armazenados e passam pelo identificador de genes. Os genes identificados são então comparados com os genes atuais do banco pelo migrador, que atualiza no banco de anotação os genes computados como sendo os mesmos entre uma montagem e outra, e inclui os novos genes encontrados. Os novos genes passam então pelo anotador automático que cria uma anotação preliminar de cada gene baseado na comparação automática com bancos de seqüências públicos. O anotador humano pode então acessar, analisar e modificar as anotações automáticas usando o editor de genes, mudar a seqüência usando o modificador de codon de início, além de visualizar as seqüências dos *chunks* e os respectivos genes usando o visualizador de *chunks*. O sistema ainda provê a capacidade de gerar arquivos de submissão ao Genbank usando o gerador de ASN.

As características gerais do sistema de anotação são:

1. independente se a seqüência completa esta fechada ou não;
2. facilidade para executar todo ou parte do pipeline;
3. facilidade de uso pelo anotadores;
4. portabilidade para outras plataformas;
5. facilidade de customização para outros projetos genoma de procariotos.

Maiores detalhes dos processos e do sistema como um todo são discutidos nas seções abaixo.

5.3.1 O Banco de dados de Anotação

Os dados de anotação são armazenados e organizados pelo Mysql [42]. Além dos dados dos genes e dos *chunks*, o banco de anotação abriga outros dados, como ilustrado na Figura 5.3. Pelo esquema, um *chunk* tem várias *features* e contém vários genes. Um gene tem uma categoria e pode sofrer alterações. O anotador pode anotar uma ou mais categorias. A categoria de um gene designa a função que o gene tem dentro do organismo. *Features* são regiões de interesse em um *chunk*, como por exemplo as repetições e os *operons* (Seção 2.5.1).

O esquema implementado permite que haja um controle de versão sobre cada atualização no banco de anotação conforme a montagem usada para alimentar o sistema com os contigs. Assim, quando o banco de anotação é atualizado, o sistema passa a tomar como referência a data da última montagem, embora os dados (genes e *chunks*) das montagens anteriores ainda permaneçam no banco. Outras características do esquema de anotação são:

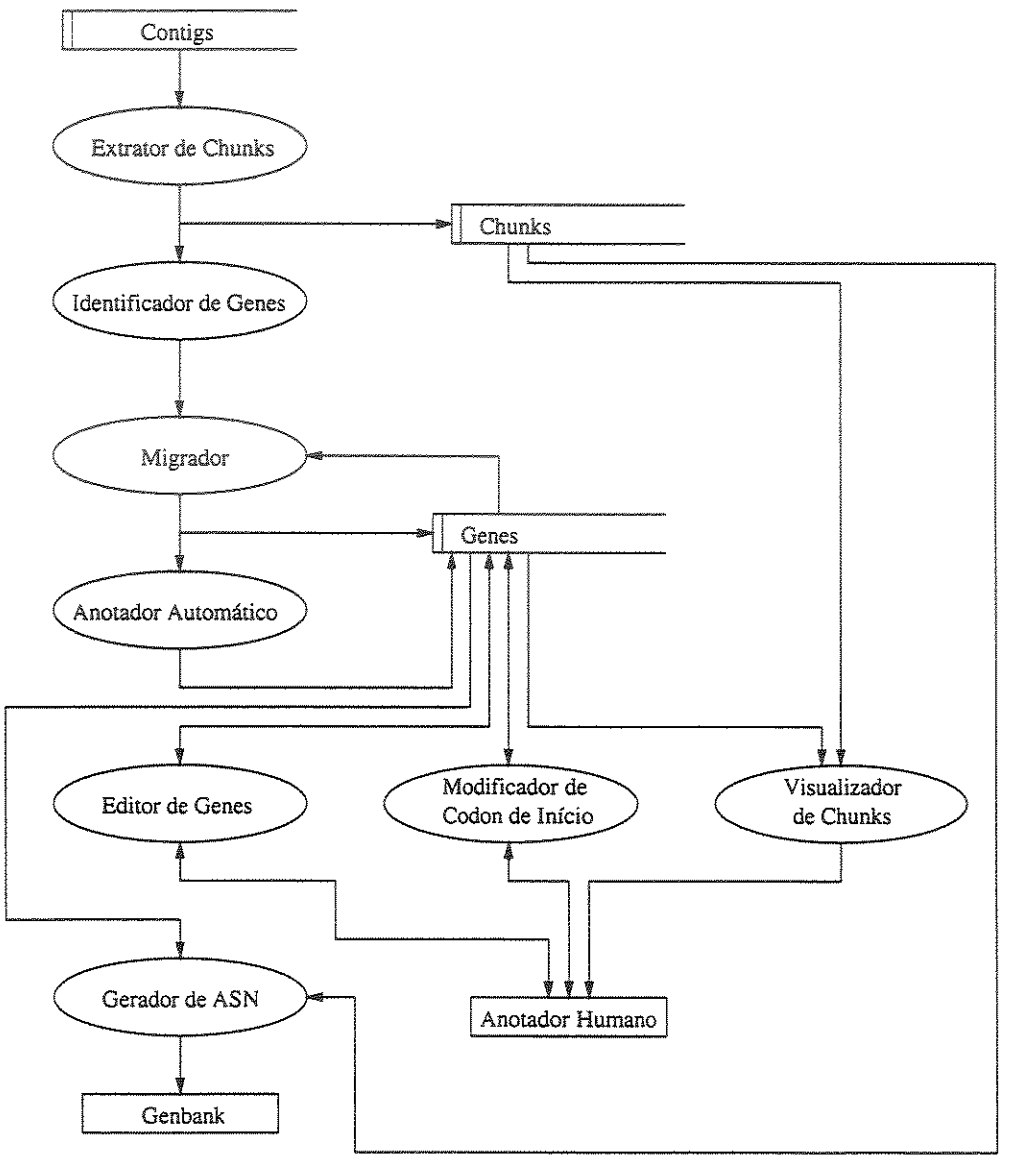


Figura 5.2: Pipeline da anotação do Projeto *Xanthomonas*. As elipses representam os processos, os retângulos abertos representam os repositórios de dados, os retângulos fechados representam entidades externas ao sistema e as linhas orientadas indicam o fluxo dos dados.

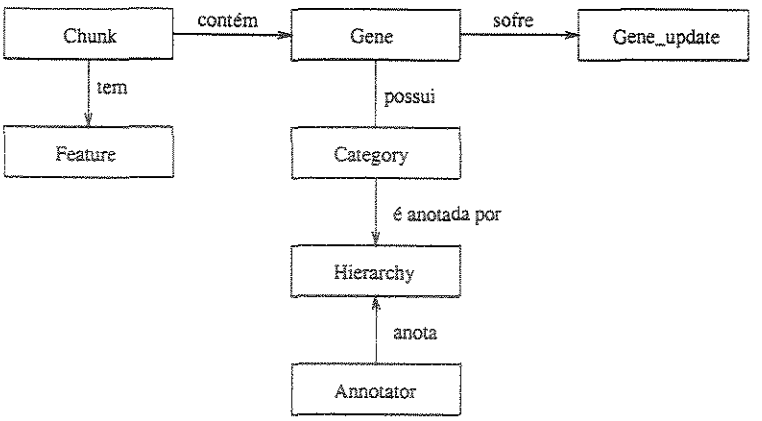


Figura 5.3: Esboço do esquema do banco de anotação. Uma linha simples indica a relação um para um. Uma linha com orientação única indica a relação um para muitos.

- permite o histórico de alterações dos genes, sabendo quem fez a mudança e quais foram os campos modificados;
- faz controle de edição de genes baseado na sua categoria, restringindo um grupo de pessoas a modificar somente os genes pertencentes a sua categoria de anotação;
- armazena as informações de qualidade das bases dos *chunks*, possibilitando conseqüentemente resgatar a qualidade de bases de cada gene.

5.3.2 Interação Montagem e Anotação

Uma das preocupações principais do sistema de anotação está em como se dá a sua interação com a montagem, principalmente na extração de *chunks* dos contigs novos e na migração dos genes de uma versão para outra da montagem. Além disso, é preciso descobrir se existem novos genes entre os contigs novos. Veja a seguir, como essas tarefas são realizadas pelo sistema.

Extrator de *Chunks*

Um *chunk*, como visto acima, é um pedaço de seqüência escolhido para ser a entidade básica de anotação. De maneira mais formal, um *chunk* em um contig é um segmento contendo somente bases com qualidade pelo menos QC e tenham no mínimo um comprimento LC. Além disso, existe o parâmetro DC que refere-se ao número máximo de bases com qualidade menor que QC permitido em um *chunk*, para evitar que haja um excesso de *chunks* curtos. Embora isso pareça influir na solidez da seqüência, pequenos segmentos

de bases de baixa qualidade entre segmentos de boa qualidade não são geralmente regiões críticas e podem ter suas qualidades melhoradas em uma próxima montagem com adição de outros reads de boa qualidade que passem por essa região. A valor máximo de QC aceito pelo programa é 50. O programa extrator de *chunks*[1] utiliza os seguintes valores como padrão: $QC = 20$, $LC = 5000$ e $DC = 10$.

Embora o ideal seja extrair seqüências longas estritamente de boa qualidade, aumentar a estringência significa proporcionalmente perder informação. E de maneira geral, por mais que se queira ajustar os parâmetros, algumas coisas sempre passam. O mais importante a ressaltar é que o mais seguro é anotar precisamente os genes no final do processo de montagem, quando os contigs estão mais estáveis. Aconselha-se elevar o critério de qualidade para anotação no início da montagem.

Identificador de Genes

O programa de identificação de genes[2] usa dois softwares: o glimmer e o tRNAscan-SE. O primeiro, com visto acima, procura achar as ORFs da seqüência de entrada, e o outro acha os tRNAs. Ambos os softwares são usados sobre cada um dos *chunks* de entrada do programa identificador, que lê os resultados na forma de coordenadas e em seguida extrai a seqüência das ORFs e dos tRNAs a partir dos *chunks*.

O programa identificador provê ainda a capacidade de treinar o glimmer. Para isso, são utilizadas as seqüências dos próprios *chunks* dados como entrada. Como o glimmer aprende com as seqüências analisadas, é importante que esporadicamente haja um retreinamento para capturar com mais precisão as ORFs nas seqüências.

Migrador

A migração de genes entre *chunks* de uma montagem para outra não é uma tarefa trivial, pois existem fatores que complicam o processo. De forma geral, migrar um gene é localizar sua seqüência entre os novos *chunks* e alterar suas coordenadas e a identificação do novo *chunk* em que está contido. Um fator complicador é saber exatamente como identificar um gene entre uma montagem e outra com base em sua seqüência. Uma alternativa óbvia é considerar o casamento exato da seqüência do gene com o *chunk*. Porém, existem casos em que genes têm sua seqüência alterada de uma montagem para outra, em geral pela substuição de bases.

Para fazer a migração o programa[3] primeiramente compara as seqüências dos genes atuais com as seqüências dos *chunks* novos, usando o cross_match. Essa comparação

[1] Desenvolvido por João Paulo F. W. Kitajima
[2] Desenvolvido por João Paulo F. W. Kitajima
[3] Desenvolvido por João Paulo F. W. Kitajima

resulta em três sub-conjuntos:

1. genes que casaram perfeitamente;
2. genes que casaram com discrepâncias, somente com substituições entre as bases;
3. genes que casaram com discrepâncias de inserção e/ou deleção de bases.

O genes do primeiro sub-conjunto são automaticamente migrados. Para os genes do segundo sub-conjunto deve-se fazer uma análise (manual) verificando se, com a substituição de bases, foi obtido algum codon de parada no meio da seqüência. Os genes fora dessa condição são então migrados com a seqüência corrigida. E por fim, os genes do terceiro conjunto não são migrados.

Após esse mapeamento entre os genes para os novos chunks e a atualização do banco de anotação, o migrador compara os genes encontrados pelo programa identificador com os genes migrados, para determinar quais são os novos genes entre o *chunks* da nova montagem. Essa comparação é baseada na posição do codon de parada das seqüências, ou seja, se não houver nenhum gene no banco com a mesma posição de um gene encontrado pelo programa identificador, então este último é considerado um gene novo.

5.3.3 Anotação Automática

O anotador automático de seqüências[4] é uma ferramenta usada para catalogar minimamente informações sobre as ORFs passadas como entrada. Como resultado, o programa gera anotações preliminares ou anotações automáticas das ORFs. Para isso, o programa compara cada seqüência de entrada com o banco de seqüências Genbank usando o BLAST, e atribui para cada uma os mesmos dados da seqüência do banco que teve melhor *hit* (caso haja *hits*). Na maioria dos casos, o primeiro *hit* (*hit* mais significativo) traz a resposta correta para a consulta. No entanto, há outros casos em que a resposta correta não é o melhor *hit*, mas sim outro que está um pouco atrás. Nesse caso, é dito que o *hit* mais significativo não é o *hit* mais relevante. De qualquer forma, essas anotações automáticas já contribuem bastante com relação a diminuição de esforço e trabalho, principalmente no primeiro caso em que o *hit* mais relevante é o mais significativo. Há um outro caso, em que não há o *hit* mais relevante, por não haver nenhum *hit*, ou pelos *hits* serem fracos segundo o BLAST, ou seja tem um *E value* ruim (ver Seção 5.2.1). Em resumo, uma anotação preliminar visa apenas servir como um ponto inicial de referência para o anotador humano, além de lhe poupar tempo em tarefas repetitivas que a anotação envolve, como por exemplo, comparar os genes contra bancos de seqüências.

[4]Desenvolvido por Felipe Rodrigues da Silva e Guilherme Pimentel Telles

No projeto *Xanthomonas* esse programa foi melhorado, sendo adicionado um módulo chamado categorizador automático[5]. Esse programa compara cada seqüência de entrada com um banco de seqüências muito específico, composto basicamente de seqüências de três genomas próximos evolutivamente ao genoma sendo estudado, e que possuem categorias bem parecidas: *Xylella fastidiosa*, *Escherichia coli* e *Xanthomonas campestris*. Desse último estão apenas as seqüências públicas conhecidas depositadas no Genbank. Com base no resultado do BLAST, o categorizador então atribui a seqüência de entrada a anotação da seqüência com melhor *hit* dentre os três genomas, inclusive a categoria, não considerada pela versão anterior no anotador automático.

Havendo ou não *hits* com esses três bancos específicos, é realizada ainda a comparação dos genes com outros bancos de seqüências como o Genbank, PFAM e COG. O propósito dessas últimas comparações é agregar de forma compacta o máximo de informação relacionado à nova seqüência gênica, de maneira que possa ser consultada e analisada pelo anotador humano para ajudá-lo na decisão de confirmar ou alterar a categoria atribuída automaticamente e, também preencher ou mudar outros dados sobre a anotação de um gene.

5.3.4 Interface Web

Para acessar o banco de anotação foram criadas algumas interfaces, orientadas principalmente para auxiliar na edição dos genes. A interface principal de edição é o editor de genes, onde são feitas as modificações dos campos (como categoria, nome do gene, nome do produto, *remarks*, etc), a menos da seqüência do gene que é alterada com o modificador de codon de início. Outra ferramenta de auxílio é o visualizador de *chunks*, onde podem ser vistos os genes mapeados nas seqüências dos *chunks*. Veja mais detalhes desses programas abaixo.

Para acessar os dados de anotação usando essa interface, o anotador deve estar cadastrado no banco de anotação. O acesso pode ser somente de leitura ou de escrita, sendo que para este último, a permissão se limita aos genes da categoria na qual o anotador está responsável.

Essas interfaces fazem o controle de acesso, distinguindo três níveis de usuários: os anotadores convencionais, os chefes de categoria e o(s) curador(es) do banco anotação. Os anotadores convencionais são aqueles que têm permissão de editar somente os genes de sua categoria, e são inibidos de remover genes ou mudar os genes de sua categoria para outra. Os chefes de categoria têm a capacidade de fazer qualquer operação nos genes de sua categoria, inclusive aquelas operações não permitidas ao anotator convencional. Por fim, o curador do banco pode realizar qualquer operação em qualquer um dos genes nas

[5]Desenvolvido por João Paulo F. W. Kitajima

diferentes categorias.

Editor de Genes

O editor de genes permite ao anotator visualizar e modificar os dados de anotação dos genes. A deleção de um gene é feita junto ao visualizador de *chunks*. Além da possibilidade de mudança dos campos, o editor de genes possui as seguintes características:

- fornece o mecanismo de registro e visualização do histórico das mudanças das anotações suportado pelo esquema do banco de dados;
- permite fazer consultas ao banco usando palavras chave, retornando uma lista com a identificação de cada gene que tenha algum de seus campos casado com a palavra chave de consulta;
- a interface do editor é implementada usando frames da linguagem HTML [11], facilitando o acesso a cada gene listado pela consulta ao banco;
- permite a visualização da qualidade das bases da seqüência dos genes;
- permite a navegação entre os genes (próximo e anterior) segundo a posição no *chunk* ou segundo a categoria.

A Figura 5.4 apresenta a interface do editor de genes. No lado esquerdo da figura, é apresentada a lista de genes resultantes da consulta sobre a palavra `protease`. Dessa lista, foi escolhido o gene 153.1, cujos dados são apresentados no lado direito. Na parte em laranja (abaixo da figura) estão os campos que são editáveis através da interface.

Visualizador de *Chunks*

O visualizador de *chunks* (*Chunk Viewer*) permite que um segmento da seqüência de um *chunk* seja escolhido e mostrado nos seis quadros de leitura, junto com os genes localizados no respectivo segmento. Veja um exemplo da visualização de um *chunk* na Figura 5.5. No quadro azul, as quatro barras superiores indicam como foi feita a escolha do segmento a ser exibido. Os seis quadros de leitura, representados pelas barras vermelhas, são mostrados somente se o comprimento do segmento escolhido cabe dentro do tamanho janela de exibição (5500 Kb). Caso o segmento tenha um comprimento maior que o da janela, então uma barra com a mesma cor do segmento escolhido é mostrada imediatamente abaixo com seus sub-segmentos. Isso continua até que o comprimento do segmento escolhido caiba dentro da janela de exibição. Uma outra maneira de acessar um segmento é fornecer a posição a ser mostrada ao programa de visualização, que então exibe diretamente a figura nos mesmos moldes como se a escolha fosse feita interativamente. Sobre os seis quadros de

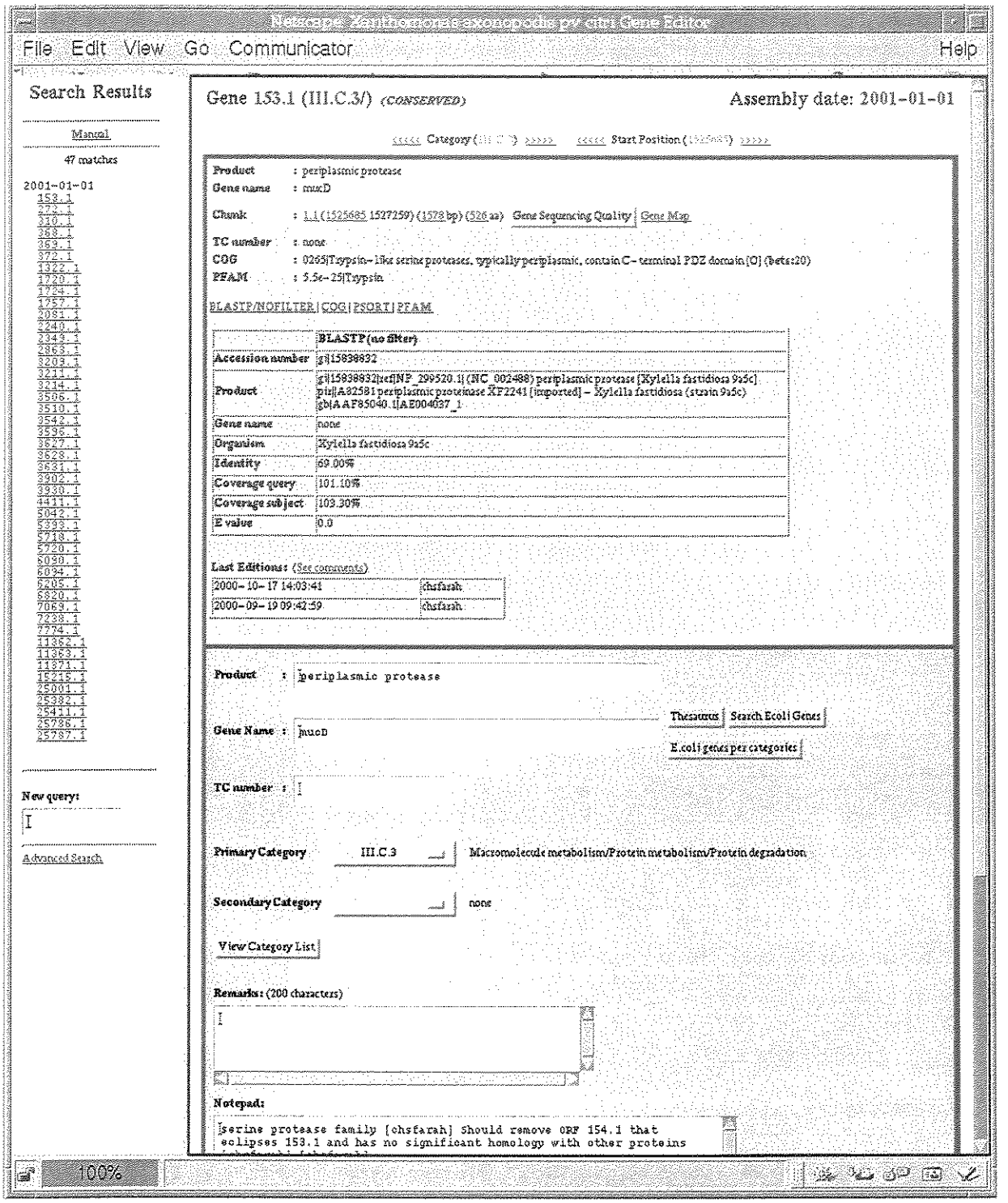


Figura 5.4: Interface do editor de genes.

leitura, são mostrados os codons de início - representados pelos traços azuis, e os codons de parada - representados pelos traços pretos. Abaixo de cada quadro de leitura aparecem os respectivos genes, representados por barras da cor da respectiva categoria, que são clicáveis e apontam para o editor de genes. No quadro com a cor de fundo laranja, é mostrada a lista de genes de cada quadro de leitura, juntamente com o nome da categoria. Os genes a serem removidos devem ser marcados, e após apertar o botão "Submit", é pedido a confirmação de remoção. Os nomes dos genes são clicáveis e apontam para o modificador de codon de início.

A escolha por implementar a operação de remoção de genes por essa interface e não pelo editor de genes, deve-se ao fato de que visualmente é mais fácil identificar os candidados à remoção, pois em geral isso é dependente da análise de sobreposição entre os genes, sendo removidos na maioria das vezes os genes totalmente cobertos ou os genes curtos que tenham sobreposição com outros genes longos.

Uma outra característica desse programa é possibilidade de mostrar, junto aos quadros de leitura, os genes encontrados por outros programas de identificação. Esses genes são representados por linhas horizontais acima de cada quadro de leitura. A incorporação desse tipo de informação ajuda de maneira geral o anotador a confirmar a existência de um dado gene. No projeto *Xanthomonas* foi usado o programa Genmark [8], mas as linhas respectivas aos genes encontrados por esse programa não estão na figura mostrada como exemplo (Figura 5.5).

Modificador de Codon de Início

Dentro do contexto do sistema de anotação, a operação de modificar a seqüência de um gene, feita por um anotador, corresponde modificar o seu codon de início. O programa modificador de codon de início (*Start Codon Changer*) provê os mecanismos e a interface pelo qual o anotador pode fazer essa operação de maneira fácil e intuitiva.

Em sua interface, o programa exibe uma figura onde os codons de início são representados por traços coloridos segundo seu tipo, desenhados sobre uma barra que representa a seqüência. Veja um exemplo na Figura 5.6. Os traços são clicáveis e selecionam o respectivo codon de início. O codon de início escolhido é indicado na figura por uma pequena seta, e abaixo aparecem o codon escolhido, as coordenadas da seqüência, a seqüência de nucleotídeos e a seqüência de aminoácidos. O anotador pode usar o botão "BlastP without filter" para fazer a comparação da seqüência de aminoácidos contra o banco de seqüências de proteínas usando o programa BLAST, de onde ele pode identificar se o codon foi ou não uma boa escolha. O botão "Change Start Codon" permite que o anotador mude a seqüência do gene para o respectivo codon de início escolhido.

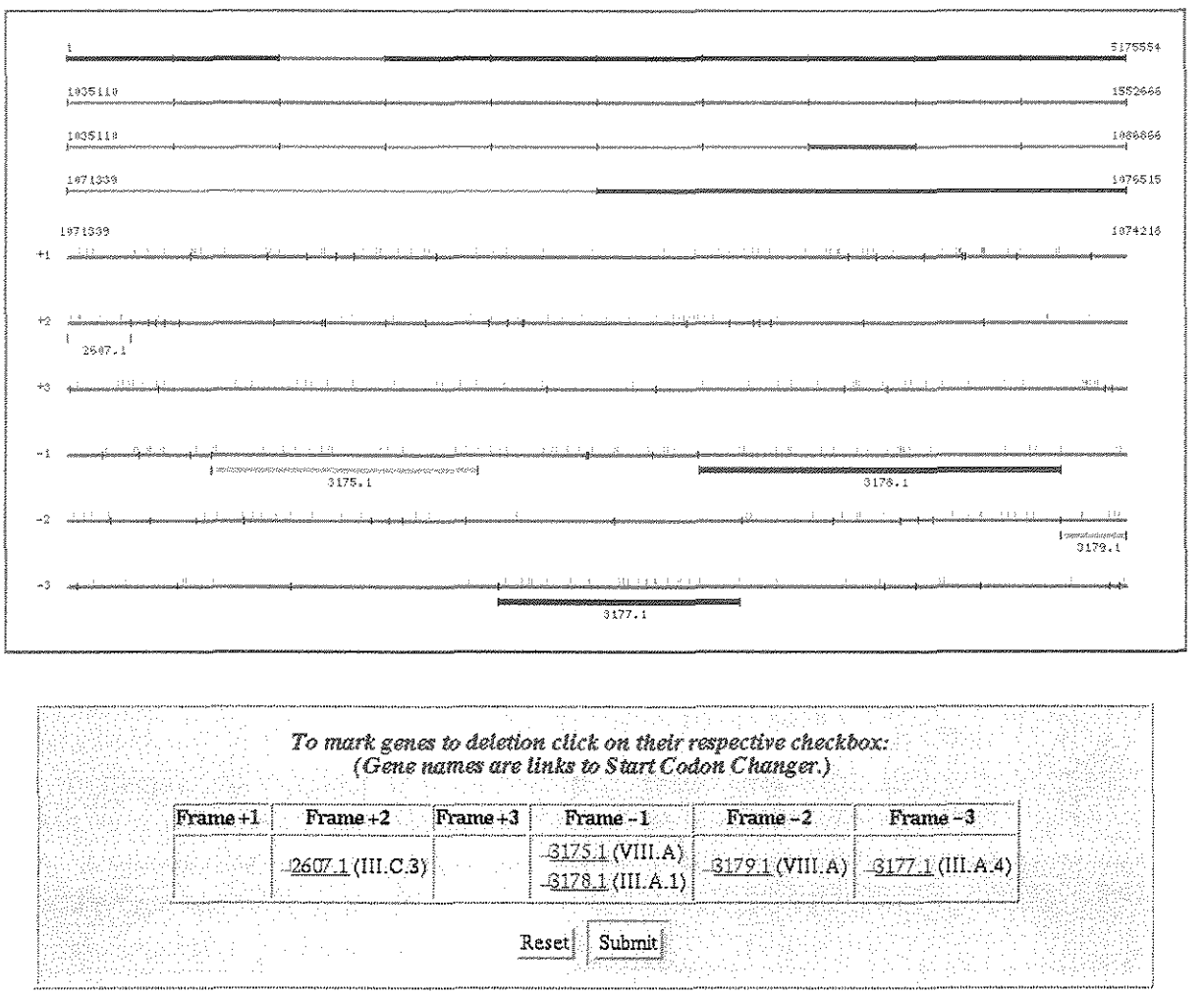


Figura 5.5: Interface do visualizador de *chunks*. No quadro azul, as quatro barras superiores delimitadas por pequenos traços verticais correspondem aos segmentos e sub-segmentos escolhidos para visualização. As seis barras inferiores de cor vermelha representam os seis quadros de leitura, junto com os codons de início (traços azuis) e codons de parada (traços pretos). Abaixo de cada quadro de leitura são mostrados os respectivos genes. No quadro laranja, são escolhidos os genes a serem removidos.

5.3.5 Outras Ferramentas de Anotação

Gerador de ASN

Ao final de um projeto de seqüenciamento de genoma, é habitual tornar público os dados e as informações de seqüências coletadas. A maneira usual de publicação dos dados de seqüência é conseguida com o seu depósito junto ao banco de seqüências público Genbank.

A submissão das seqüências ao Genbank deve ser feita em um formato específico, no caso, o ASN.1 (*Abstract Syntax Notation 1*). Esse formato é reconhecido pelo programa Sequin (ver Seção 5.2.4). O programa gerador de ASN[6] recebe como entrada a seqüência de nucleotídeos e os dados de seus genes (como nome, coordenadas, nome do produto, tipo de RNA, entre outros), lidas diretamemente do banco, e produz um arquivo no formato ASN.

Gerador do Mapa de Genes

O mapa de genes de uma seqüência de DNA mostra graficamente a disposição geral de seus genes, além da orientação e da categoria de cada gene. A seqüência de DNA é representada por várias linhas, delimitadas por espaços uniformes que estabelecem uma escala (cada espaço equivale a 1000 bp). Abaixo dessas linhas, são desenhadas barras coloridas com orientação que representam os genes. Essa orientação indica em qual das duas fitas está o gene. As cores dos genes indicam a categoria a que pertencem, mostrado pela legenda na parte inferior do mapa, como ilustrado na Figura 5.7. As barras dos genes são clicáveis e apontam para o editor de genes. Outras características sobre os genes podem ser ressaltadas no mapa, como a marca em forma de X sobre alguns genes. No caso, essa marca indica que há um problema na seqüência do gene (frameshift ou ponto de mutação).

Além de possibilitar uma visão geral da disposição dos genes, o mapa de genes permite a facilidade de varredura visual dos genes por parte do anotador, podendo fazer uma varredura seqüencial ou uma varredura por categoria.

O programa gerador do mapa de genes lê o banco de anotação e produz a figura representando o mapa. Durante o processo de anotação, os genes podem sofrer alterações, ou mesmo podem ser removidas do banco. Essas mudanças afetam o mapa de genes, que deve então ser atualizado com uma re-execução do programa gerador do mapa de genes. De acordo com o nível de mudanças do banco, a atualização do mapa era feita diariamente no perído de poucas mudanças, e feita com mais freqüência no período de muitas mudanças.

[6] Desenvolvido em parceria com Guilherme Pimentel Telles

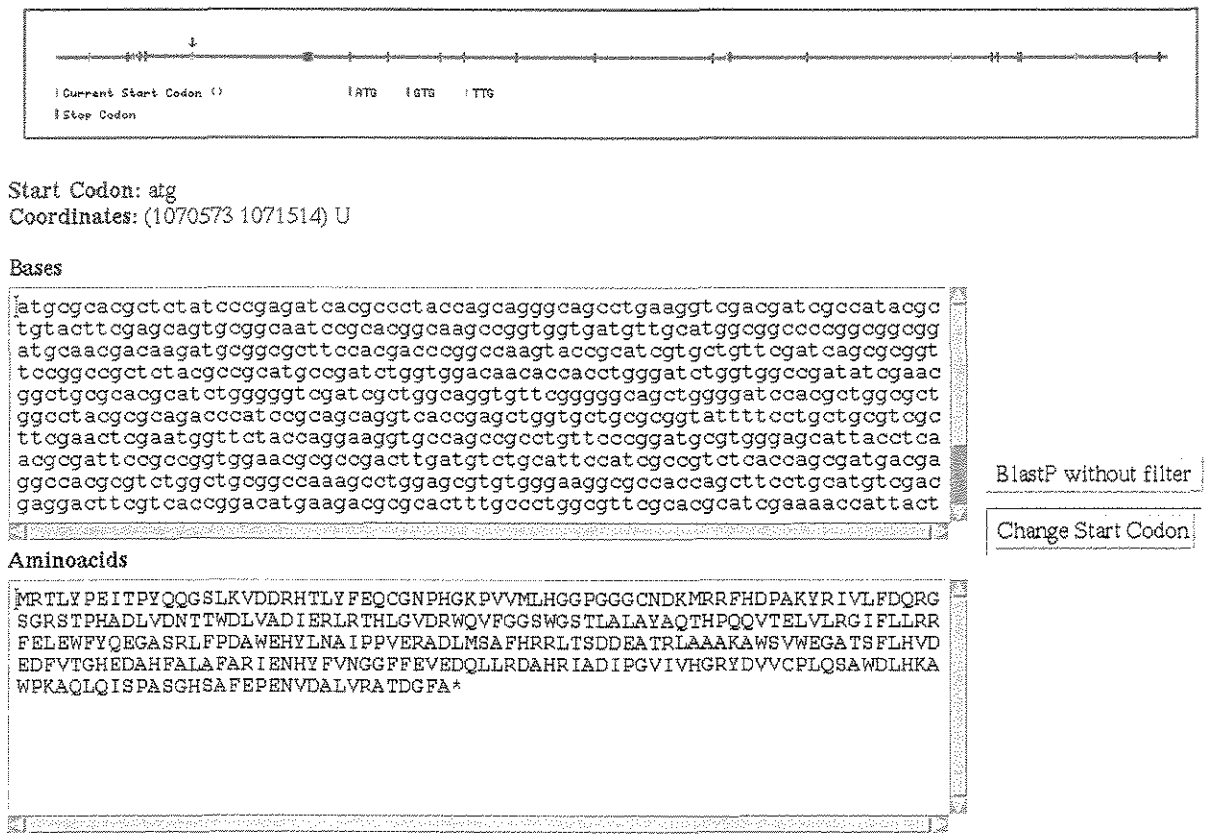


Figura 5.6: Inferface do modificador de codon de início.

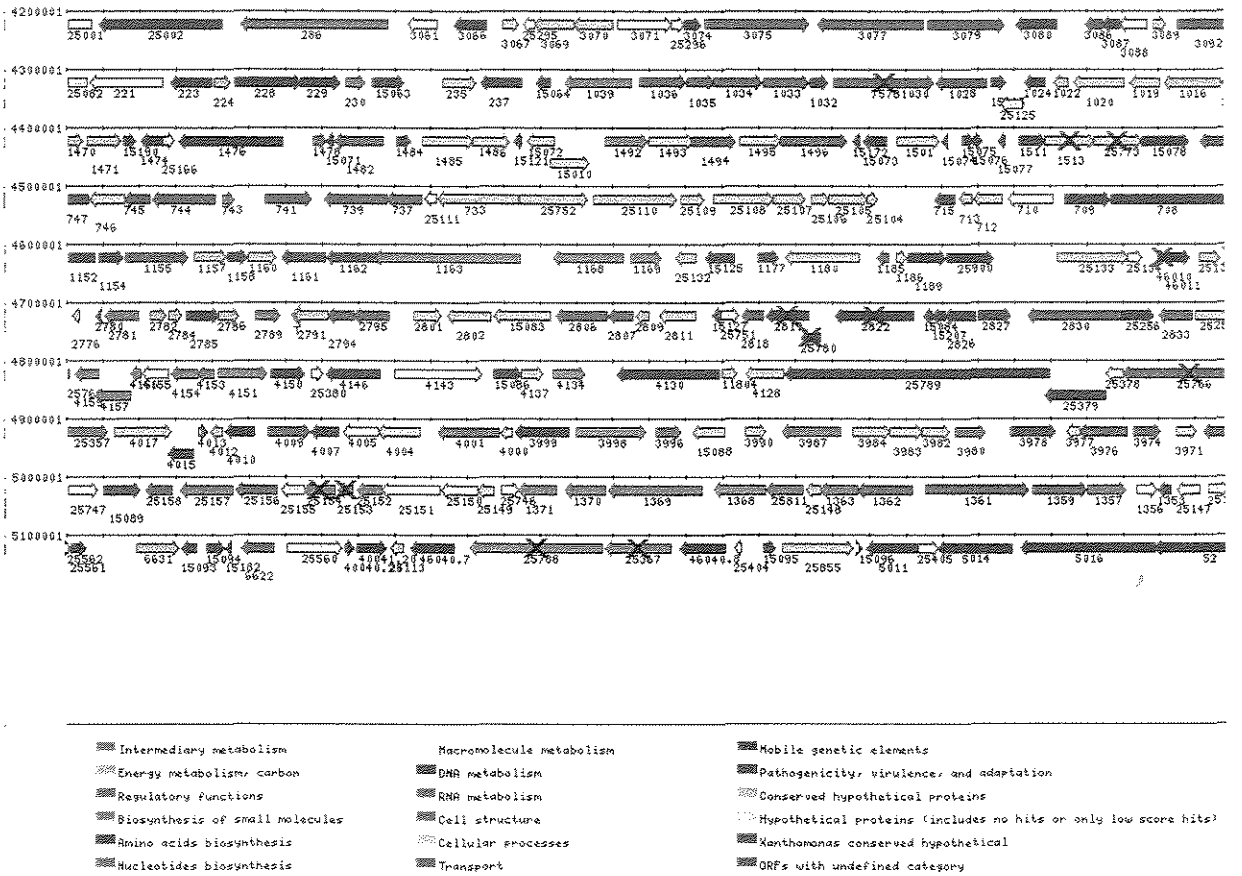


Figura 5.7: Mapa de genes mostrado parcialmente.

Capítulo 6

Outras Ferramentas

6.1 Digestão Eletrônica

Como visto na Seção 2.6.1, digestão é o processo biológico pelo qual uma molécula de DNA é cortada por enzimas de restrição. Além de ser utilizada no processo de fragmentação de moléculas, a digestão pode ser usada também na construção de mapas de restrição. Um mapa de restrição mostra os lugares ou pontos onde uma uma molécula foi digerida com uma ou mais enzimas de restrição.

O programa de digestão eletrônica[1] foi desenvolvido com propósito de comparar versões eletrônicas de mapas restrição com os mapas de restrição reais. O propósito dessa comparação é confirmar a colinearidade entre uma molécula e sua respectiva seqüência. Dadas uma seqüência de DNA S e a seqüência de uma enzima E, o programa de digestão eletrônica produz uma figura mostrando os pontos (sítios de restrição) onde E corta S, identificando quais os fragmentos gerados e seus respectivos comprimentos. Além disso, é mostrado também como seria a distribuição desses fragmentos em um gel de eletroforese, segundo o comprimento desses fragmentos. Um exemplo do mapa de restrição produzido para uma seqüência de cosmídeo é apresentado na Figura 6.1.

Achar a seqüência de uma enzima em uma seqüência de DNA corresponde ao problema de procurar uma determinada sub-cadeia em uma cadeia de caracteres maior (*string matching*) [19]. No programa de digestão foi usado a função `index` da linguagem Perl, que dada uma cadeia S e uma sub-cadeia r, retorna a posição da primeira ocorrência de r em S.

[1] Desenvolvido por Marcos Renato R. Araujo

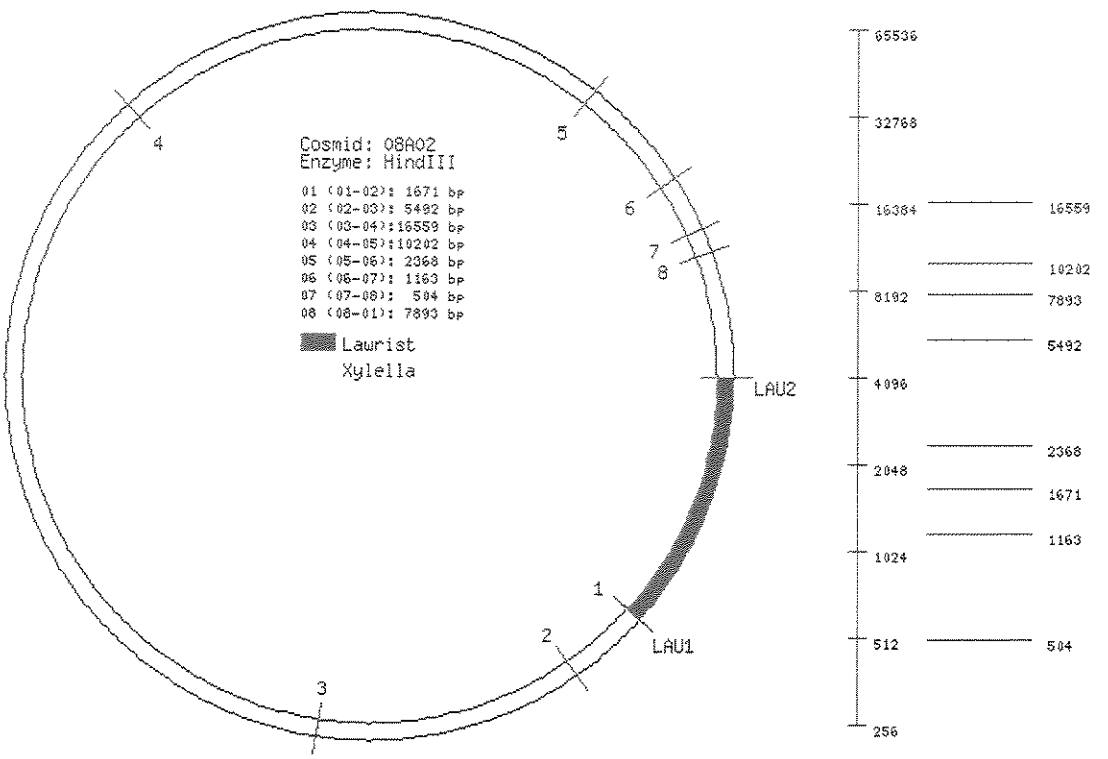


Figura 6.1: Mapa de restrição produzido pela digestão eletrônica.

6.2 PCR Eletrônico

O PCR, como visto na Seção 2.6.2.3, é uma das alternativas para fazer cópias de um DNA. Entre os elementos usados para realização desse experimento temos o primer, a partir do qual se inicia a cópia do DNA. Uma das exigências para a eficiente execução do processo é que a seqüência do primer usado não esteja repetida no DNA molde. Uma boa escolha de uma sequência de primer a ser construído, em casos como o problema de fechamento de buracos, é extremamente importante. A presença do sítio do primer em mais de um lugar da molécula original pode acarretar em cópias de trechos distintos, o que não é desejável.

O objetivo do programa de PCR eletrônico[2] é mostrar onde nos contigs a seqüência (candidata à primer) passada como entrada se liga, ou seja, onde a seqüência de entrada ocorre na seqüência dos contigs. Se a seqüência do primer for encontrada em algum contig, o programa mostra graficamente o contig e quais os pontos onde ele se liga. O programa permite um nível de tolerância, ou seja, um número aceitável de diferenças nas bases entre a seqüência do primer e a seqüência do contig. Isso é feito com a utilização do programa *psearch* [28], que implementa um algoritmo de casamento de cadeias (*string matching*) que permite erros no padrão ou no texto. O programa de PCR eletrônico exibe como resultado uma figura contendo para cada contig, duas linhas espessas (uma linha para cada orientação do contig) com as posições (representada por uma pequena linha vertical) onde se localizam seqüência de entrada (Figura 6.2). Colocando o mouse sobre cada posição, é mostrado na janela abaixo das linhas o respectivo alinhamento entre a seqüência do primer e a seqüência do contig. No exemplo da Figura 6.2, é exibido o alinhamento corresponde ao da primeira posição da linha inferior, indicando que a seqüência de entrada é complementar ao segmento do contig.

6.3 Miscelânea

Query Reads

O programa tem a finalidade de deteminar em qual contig um determinado read foi montado. Pode ser especificado mais de um read por consulta. Se o read for encontrado em algum contig, é retornado o número e nome do contig, a posição inicial e a posição final do read no contig, e a orientação do read no contig. A procura é feita usando a saída da montagem realizada pelo programa phrap.

[2]Desenvolvido por Marcos Renato R. Araujo

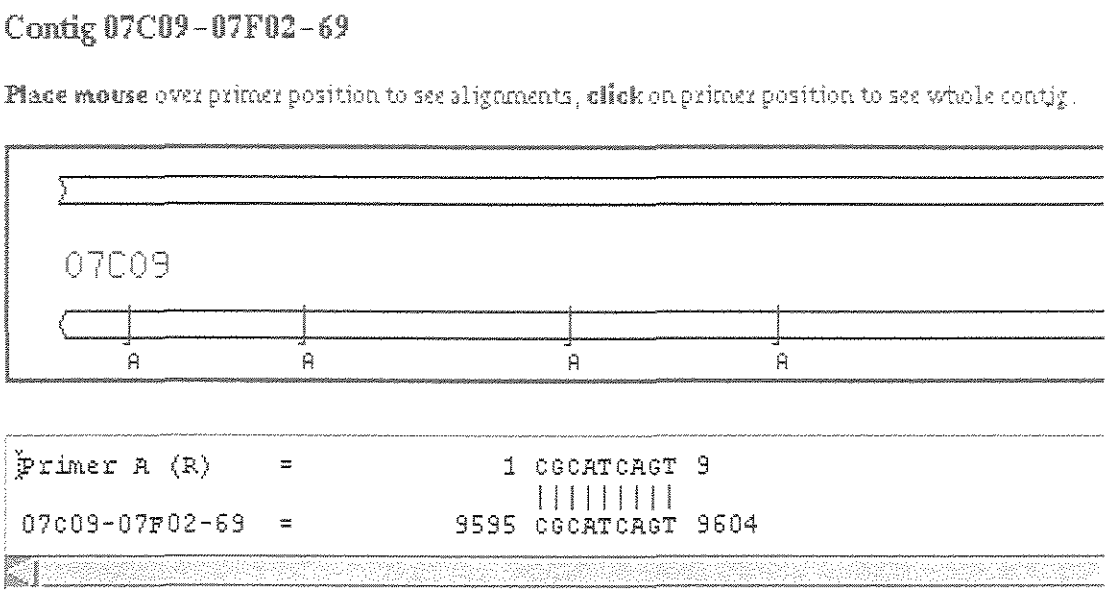


Figura 6.2: Resultado de uma pesquisa usando o PCR eletrônico.

Pagamentos

Esse programa contabiliza os valores a serem pagos para cada laboratório pelo seu esforço de seqüenciamento. Os valores são apresentados em uma tabela, que inicialmente contabilizava os valores para as seqüências de cosmídeos e de plasmídeos. Para os cosmídeos eram contados o número de bases seqüenciadas e para os plasmídeos o número de reads seqüênciados. Com o progresso do projeto *Xylella* novos seqüenciamentos foram feitos, sendo então acrescidas à tabela mais três colunas:

- GFS (gap filling sequences): seqüências utilizadas no processo de fechamento de buracos do genoma. Valor pago pelo número de bases da seqüência.
- Novos reads de shotgun: Valor pago pelo número de reads.
- end reads: reads de pontas de cosmídeos e de pontas de fagos lambda. Valor pago pelo número de reads.

Os valores das colunas dessa tabela correspondem aos valores totais e são links que apontam para outra tabela apresentando os números que deram origem ao valor total contabilizado. No caso dos cosmídeos e das GFSs são listados para cada seqüência, o nome, o número de bases e o valor a ser pago. No caso dos plasmídeos, dos reads de shotgun e dos reads de ponta, são listados para cada ano/mês o número de reads considerados

pagáveis e o número de reads considerados não pagáveis. Um read é considerado pagável se possui um número mínimo de bases com qualidade boa.

Procura de Reads Quiméricos

O programa lê a saída da montagem produzida pelo phrap e procura quais reads são identificados como quiméricos. Para cada read indentificado como quimérico, o programa mostra o nome do read, o nome do cosmídeo que o read tem sobreposição, e a posição relativa do read com relação ao cosmídeo.

Auto Corretor da Montagem

A finalidade desse programa é identificar e remover os reads de shotgun considerados problemáticos ou que não influenciam na montagem. Em uma primeira análise, são removidos os reads quiméricos e os reads identificados como "probable deletion reads". Entre os reads quiméricos, são removidos somente aqueles que fazem parte dos contigs de shotgun. Contigs de shotgun são aqueles que são formados apenas por reads de shotgun. O programa envia uma mensagem a algumas pessoas que estão diretamente envolvidas com a montagem, reportando a lista de reads removidos.

Visualizador de Genes

Esse programa é usado no site oficial (público) do Projeto *Xylella* para acessar os genes. No site há uma caixa de entrada onde podem ser digitadas palavras chave para buscar genes no banco de dados. As consultas podem feitas usando os operadores lógicos *and*, *or* ou *not*. Como resultado, o programa mostra a lista de genes em cujo registro foi encontrada a palavra chave, ou no caso da lista resultante conter apenas um gene, o registro desse gene é exibido, onde são apresentadas informações como a sua identificação, o nome do produto, o nome do gene, a categoria, coordenadas e comprimento da seqüência.

Visualizador de Categorias

Esse programa é usado também no site oficial do Projeto *Xylella* e visa acessar os genes segundo a sua categoria. Ao ser invocado, o programa mostra as categorias usadas na classificação dos genes e o número de genes em cada categoria. Ao clicar na descrição de uma categoria, é exibida a lista de genes que fazem parte da categoria escolhida.

Fastaqual2Phd

O pacote phred/phrap/consed traz consigo um programa chamado fasta2phd.perl, que recebe como entrada um arquivo fasta e produz um arquivo phd, ou seja, um arquivo contendo as qualidades das bases do arquivo fasta. No entanto, a qualidade é a mesma para todas as bases e tem um valor padrão. Esse tipo de coisa é geralmente útil quando se quer incorporar seqüências que não tem qualidade a uma montagem que usa seqüências que tem qualidade. Porém, nos projetos *Xylella* e *Xanthomonas* houve a necessidade de incorporar à montagem seqüências com qualidade que não tinham um arquivo phd. Um exemplo, são as seqüências montadas. O montador gera automaticamente um arquivo com a seqüência de bases no formato fasta. Assim, através de uma adaptação, foi criado o programa fastaqual2Phd, que recebe como entrada o arquivo fasta com as bases e o arquivo fasta com as qualidades e dá como saída o arquivo phd.

Extrator de Regiões Upstream de Genes

Esse programa[3] visa identificar o segmento de comprimento NB imediatamente anterior ao início de cada gene. A informação extraída por esse programa é útil no estudo para encontrar regiões de RBS (*ribosome binding site*) dos genes. O programa leva em conta se o gene está na fita direta ou inversa, e se há sobreposição do segmento com o gene anterior na mesma fita. Para esse último caso, o programa imprime a parte do segmento que não está se sobrepondo com o gene anterior, desde que tenha um comprimento mínimo U. Além disso, o programa aceita uma bandeira (*flag* -A) que ignora se há ou não sobreposição.

[3] Desenvolvido em parcecia com João Carlos Setubal

Capítulo 7

Conclusão

Objetivamos neste trabalho, descrever sistematicamente como foram realizadas as tarefas de bioinformática baseando-se nos pipelines de montagem e anotação usados, respectivamente, no projeto genoma *Xylella* e no projeto genoma *Xanthomonas*. Além disso, esses pipelines foram também utilizados, com algumas modificações, no projeto genoma *Xylella fastidiosa Pierce's Disease* e no projeto genoma *Agrobacterium tumefaciens*. Nesse último, foi usado apenas o sistema de anotação.

Para isso, apresentamos uma breve introdução sobre alguns conceitos de biologia molecular e de seqüenciamento de DNA, para um melhor um entendimento do trabalho e também para servir como um pequeno guia para iniciantes na área de bioinformática. É bom ressaltar a importância do profissional dessa área ter bons conhecimentos em ambas as disciplinas de computação e biologia, seja um profissional de computação ou um biólogo, para assim conhecer com mais precisão os problemas e dessa forma oferecer soluções ou ainda melhorar as que já existem.

Apresentamos também uma descrição sobre projetos genoma, suas abordagens de seqüenciamento e análise, suas etapas principais e como a bioinformática têm um papel fundamental na sua realização.

Apresentamos detalhes dos dois pipelines desenvolvidos pelo LBI, juntamente com a descrição dos conceitos relativos ao problema e das principais ferramentas utilizadas. O primeiro sistema serve para recepção, análise de qualidade e montagem de seqüências, com a característica de ser simples, portável e flexível. O segundo sistema, voltado para a anotação de genes, permite que tal análise sobre o genoma seja realizada paralelamente com a montagem, independente se a seqüência completa esteja fechada ou não. E para que tal análise seja realizada com segurança, é necessário que sejam fornecidas junto com as seqüências as qualidades de suas bases. A principal característica desses sistemas é a sua flexibilidade, ou seja, podem ser facilmente customizáveis. Apesar disso, essa característica pode levar por outro lado a dificuldade de manter atualizado entre os diferentes

projetos em andamento, as melhorias gerais ou novas funcionalidades que são realizadas no sistema de um projeto específico.

E finalmente apresentamos diversos programas com aplicações específicas, que podem ser usados junto aos dois sistemas descritos, como ferramentas para análises adicionais, visando a confirmação e a melhoria de dados gerados.

Os dois pipelines descritos podem ser considerados como dois sistemas de software, aplicados respectivamente ao problema da montagem de seqüências e ao problema de anotação de um genoma. Há algumas melhorias que devem ser realizadas em ambos para que sejam sistemas mais compactos e mais estruturados. No caso do sistema de montagem, as melhorias que podem ser realizadas são:

1. Centralizar as configurações dependentes do projeto e da plataforma, para facilitar a customização e portabilidade do sistema. Exemplos dessas configurações são o padrão do nome das seqüências e localização das ferramentas usadas pelo sistema.

2. Organizar os diversos programas desenvolvidos, estabelecendo uma divisão modular mais eficiente;

3. Criar um pacote de instalação para o sistema.

Diferente do sistema de montagem, o sistema de anotação foi um pouco mais elaborado, e teve um planejamento (embora informal) antes de sua implementação. As melhorias dos itens 1 e 3 para o sistema de montagem servem também para o sistema de anotação, embora para esse último o primeiro item tem bem menos peso. Outras melhorias gerais sobre o sistema de anotação estão relacionadas a total implementação e incorporação de outros módulos projetados. Um exemplo é o módulo de análise e disponibilização de informações sobre regiões intergênicas, que podem conter genes não achados pelas ferramentas de identificação. Outro exemplo é o módulo de edição de *features* do genoma. Algumas partes desses módulos já foram implementadas, mas fazem parte da customização do sistema em projetos distintos.

As minhas principais contribuições são:

- sistematização e documentação dos pipelines de montagem e anotação, e que portanto deve ser útil para outros projetos genoma;

- introdução básica sobre Biologia Molecular e seqüenciamento de DNA, servindo como um guia para iniciantes na área de bioinformática;

- desenvolvimento de programas e módulos específicos;

- participação (co-autoria) nos trabalhos científicos do projeto *Xanthomonas* [10] e do projeto *Agrobacterium* [40].

O estudo de genomas é uma área de pesquisa que tem ainda muitos desafios. O escopo deste trabalho mostra apenas uma parcela disso, restringindo-se ao que tem sido feito a nível de bioinformática para os genomas de bactérias, que mesmo sendo considerados organismos vivos simples, apresentam certo grau de dificuldade quando estudados. E ainda há todo o restante dos organismos do domínio dos procariotos, sem falar nos eucariotos, que são organismos muito mais complexos de serem estudados. Além disso, em paralelo ao estudo dos genomas, surgem os desafios pós-genoma relacionados ao estudo da expressão gênica baseado na análise de RNAs expressos (trascriptoma) e ao estudo das proteínas em um organismo (proteoma).

Apêndice A

Tabelas Adicionais

Nesta seção são colocadas algumas tabelas referenciadas nos capítulos da dissertação.

O mapeamento genético apresentado abaixo é feito usando-se bases de RNA, pois são as moléculas que fazem a ligação entre DNA e a síntese de proteínas. As quatro bases são arranjadas em triplas e resultam em 64 combinações. Apesar desse número, somente 20 aminoácidos são utilizados na composição de proteínas, resultando em triplas distintas levando a um mesmo aminoácido. O codon `AUG` que codifica a metionina é usado com um codon especial, denominado *start codon*, que marca o ínicio dos genes. Há outras três triplas especiais, denominadas de *stop codons*. Elas são identificadas pela palavra STOP e determinam o fim da composição de uma cadeia de aminoácidos. Esse código genético é usado por grande parte dos organismos vivos, sendo tratado como código universal. Outros organismos usam um código ligeiramente modificado.

Tabela A.1: Código genético: mapeamento de codons para aminoácidos.

Primeira posição	**Segunda posição**				**Terceira posição**
	G	A	C	U	
G	Gly	Glu	Ala	Val	G
	Gly	Glu	Ala	Val	A
	Gly	Asp	Ala	Val	C
	Gly	Asp	Ala	Val	U
A	Arg	Lys	Thr	Met	G
	Arg	Lys	Thr	Ile	A
	Ser	Asn	Thr	Ile	C
	Ser	Asn	Thr	Ile	U
C	Arg	Gln	Pro	Leu	G
	Arg	Gln	Pro	Leu	A
	Arg	His	Pro	Leu	C
	Arg	His	Pro	Leu	U
U	Trp	STOP	Ser	Leu	G
	STOP	STOP	Ser	Leu	A
	Cys	Tyr	Ser	Phe	C
	Cys	Tyr	Ser	Phe	U

Apêndice B

Relatórios e Sumários

Nesta seção são colocados exemplos de relatórios gerados pelo programas de submissão de seqüências e sumários gerais sobre análise de cosmídeos e contigs, referenciados nos capítulos da dissertação.

B.1 Relatório de Submissão de Reads

```
========== RELATORIO DE PROCESSAMENTO DO LOTE 000011 ==========

seq_id: X0QR0002F03F
vetor(pbs-bluescribe): 47 -> 127
n_bases= 815 vetor= 81 (9.9%) >=20= 486 (59.6%) >=30= 428 (52.5%)
seq_id: X0QR0002F04F
vetor(pbs-bluescribe): 34 -> 102
n_bases= 814 vetor= 69 (8.5%) >=20= 189 (23.2%) >=30= 44 (5.4%)
seq_id: X0QR0002F06F
vetor(pbs-bluescribe): 37 -> 106
vetor(pbs-bluescribe): 783 -> 820
n_bases= 822 vetor= 108 (13.1%) >=20= 451 (54.9%) >=30= 332 (40.4%)
seq_id: X0QR0002F07F
vetor(pbs-bluescribe): 66 -> 119
n_bases= 818 vetor= 54 (6.6%) >=20= 171 (20.9%) >=30= 24 (2.9%)
seq_id: X0QR0002F09F
vetor(pbs-bluescribe): 67 -> 136
n_bases= 835 vetor= 70 (8.4%) >=20= 526 (63.0%) >=30= 428 (51.3%)
seq_id: X0QR0002F10F
vetor(pbs-bluescribe): 37 -> 93
n_bases= 824 vetor= 57 (6.9%) >=20= 369 (44.8%) >=30= 222 (26.9%)
seq_id: X0QR0002F11F
vetor(pbs-bluescribe): 47 -> 124
n_bases= 826 vetor= 78 (9.4%) >=20= 436 (52.8%) >=30= 315 (38.1%)
seq_id: X0QR0002F12F
vetor(pbs-bluescribe): 17 -> 97
n_bases= 814 vetor= 81 (10.0%) >=20= 515 (63.3%) >=30= 482 (59.2%)
seq_id: X0QR0002G01F
vetor(pbs-bluescribe): 47 -> 127
n_bases= 812 vetor= 81 (10.0%) >=20= 523 (64.4%) >=30= 467 (57.5%)
```

```
seq_id: X0QR0002G02F
vetor(pbs-bluescribe): 31 -> 111
vetor(pbs-bluescribe): 630 -> 803
n_bases= 822 vetor= 255 (31.0%) >=20= 501 (60.9%) >=30= 452 (55.0%)
seq_id: X0QR0002G04F
vetor(pbs-bluescribe): 35 -> 103
n_bases= 814 vetor= 69 (8.5%) >=20= 495 (60.8%) >=30= 450 (55.3%)
seq_id: X0QR0002G10F
vetor(pbs-bluescribe): 24 -> 100
n_bases= 807 vetor= 77 (9.5%) >=20= 457 (56.6%) >=30= 361 (44.7%)
seq_id: X0QR0002G11F
vetor(pbs-bluescribe): 47 -> 129
n_bases= 813 vetor= 83 (10.2%) >=20= 518 (63.7%) >=30= 476 (58.5%)
seq_id: X0QR0002H02F
vetor(pbs-bluescribe): 20 -> 95
n_bases= 810 vetor= 76 (9.4%) >=20= 245 (30.2%) >=30= 131 (16.2%)
seq_id: X0QR0002H03F
vetor(pbs-bluescribe): 6 -> 97
vetor(pbs-bluescribe): 550 -> 784
n_bases= 784 vetor= 327 (41.7%) >=20= 452 (57.7%) >=30= 452 (57.7%)
seq_id: X0QR0002H04F
vetor(pbs-bluescribe): 16 -> 97
vetor(pbs-bluescribe): 714 -> 810
n_bases= 816 vetor= 179 (21.9%) >=20= 492 (60.3%) >=30= 428 (52.5%)
seq_id: X0QR0002H05F
vetor(pbs-bluescribe): 49 -> 129
n_bases= 822 vetor= 81 (9.9%) >=20= 503 (61.2%) >=30= 455 (55.4%)
seq_id: X0QR0002H06F
vetor(pbs-bluescribe): 29 -> 112
n_bases= 824 vetor= 84 (10.2%) >=20= 499 (60.6%) >=30= 418 (50.7%)
seq_id: X0QR0002H07F
vetor(pbs-bluescribe): 26 -> 107
n_bases= 797 vetor= 82 (10.3%) >=20= 539 (67.6%) >=30= 490 (61.5%)
seq_id: X0QR0002H09F
vetor(pbs-bluescribe): 47 -> 129
n_bases= 814 vetor= 83 (10.2%) >=20= 529 (65.0%) >=30= 492 (60.4%)
seq_id: X0QR0002H11F
vetor(pbs-bluescribe): 53 -> 130
n_bases= 827 vetor= 78 (9.4%) >=20= 71 (8.6%) >=30= 4 (0.5%)
seq_id: X0QR0002H12F
vetor(pbs-bluescribe): 17 -> 97
vetor(pbs-bluescribe): 374 -> 812
n_bases= 813 vetor= 520 (64.0%) >=20= 276 (33.9%) >=30= 276 (33.9%)
seq_id: X0QR0003A01F
vetor(pbs-bluescribe): 44 -> 124
n_bases= 823 vetor= 81 (9.8%) >=20= 530 (64.4%) >=30= 473 (57.5%)
seq_id: X0QR0003A02F
vetor(pbs-bluescribe): 16 -> 96
n_bases= 809 vetor= 81 (10.0%) >=20= 504 (62.3%) >=30= 439 (54.3%)

>>>>>>>>>>> Totais para lote 000011 <<<<<<<<<<<

n_reads= 24 n_bases= 19575 vetor= 2835 (14.5%) >=20= 10277 (52.5%) >=30= 8539 (43.6%)

----------------  final do relatorio ------------------------
```

B.2 Relatório de Submissão de Seqüências Montadas

```
PRELIMINARY CERTIFICATION REPORT FOR COSMID 03C03
DATE: Mon 29-Mar-1999

READ COMPOSITION: 1385 reads;

BF-03C03 :  432 reads
IL-03C03 :  180 reads
PM-03C03 :    2 reads
RC-07E05 :  433 reads
UV-07B07 :  280 reads
plasmid  :   58 reads

Contig size: 41769 bp
Insert size: 39387 bp (excludes lawrist and/or cosmid extensions)

LAWRIST LOCATIONS:

Left  : first base after  LAU2 = 940
Right : last  base before LAU1 = 40326

EXPECTED ERROR RATE: 0.28 / 10000 bp

LOW CONSENSUS QUALITY (LCQ) REGIONS:

  19278 - 19280 (3)
  19333 - 19335 (3)
  19398 - 19403 (6)
  19405 - 19408 (4)
  19458 - 19460 (3)

TOTAL: 19

HIGH QUALITY DISCREPANCIES (HQD):

XOIL-03C03-A035F   26504 - 26506 (3)
XOIL-03C03-A035F   26508 - 26509 (2)
XOBF-03C03-Z261F   34654 - 34654 (1)
XOIL-03C03-A037F   16317 - 16317 (1)
XOIL-03C03-A037F   16318 - 16319 (2)
XOIL-03C03-A037F   16321 - 16321 (1)
XOIL-03C03-A037F   16324 - 16324 (1)
XOIL-03C03-A037F   16326 - 16326 (1)
XOIL-03C03-A032F   27092 - 27093 (2)
XOIL-03C03-A032F   27096 - 27100 (5)
XOIL-03C03-A092F   13947 - 13947 (1)
XOIL-03C03-A096F   13751 - 13751 (1)
XOBF-03C03-Z259F   36212 - 36213 (2)
XOIL-03C03-A005F   12765 - 12766 (2)
XOIL-03C03-A005F   12769 - 12769 (1)
XOIL-03C03-A005F   12772 - 12772 (1)

TOTAL: 27

POSITIONS NOT CONFIRMED ON BOTH STRANDS (NCBS):

   19437-19437 (1)
```

```
   19514-19514 (1)
   19517-19517 (1)
   19520-19520 (1)
   19525-19525 (1)
   19539-19539 (1)
   19551-19551 (1)
   19553-19564 (12)
   19566-19568 (3)
   19571-19576 (6)
   19578-19581 (4)
   19583-19585 (3)
   19587-19589 (3)
   19593-19593 (1)
   19596-19596 (1)
   19598-19600 (3)
   19602-19612 (11)
   19614-19615 (2)
   19618-19618 (1)

TOTAL: 57


CONCLUSION: Cosmid still not close to finished;
            please check problems pointed out above

END OF ANALYSIS
```

B.3 Sumário de Cosmídeos

A figura a seguir apresenta o sumário criado para os cosmídeos. Conforme seu estado, a linha do cosmídeo na tabela é pintada com uma cor. Alguns dos campos da tabela são apontadores Web para outras informações relativas a coluna. Por exemplo, ao clicar no item `Sequence` (mostra seu tamanho) de uma dada linha leva a seqüência do cosmídeo, ou no item `Report` leva ao relatório gerado pela certificação (ver Apêndice B.2).

Cosmid summary

The number in parenthesis besides the color code indicates the number of cosmids in that state.
Finished (117) Close to Finished (0) Shotgunned (0) Submitted (0) Not Submitted (0)

Note that some column headers are links: they point to important explanations about that column's content; please read them.
Last updated: Fri Apr 12 16:00:01 EDT 2002

Code	Lab	# reads	Status	Report	Sequence	Overlaps	Annotation
01A01	PM	360	Finished		37082	1 piece	Done
01A03	UV	983	Finished	report.txt	39112	1 piece	Done
01E01	AG	394	Finished	report.txt	43547	1 piece	Automatic only
01E01	IU	156	Finished	report.txt	N/A	1 piece	Automatic only
01G04	BG	553	Finished	report.txt	37493	1 piece	Done
01G04	IU	796	Finished	report.txt	N/A	1 piece	Done
01G06	CE	438	Finished	report.txt	40185	1 piece	Automatic only
01G06	UT	506	Finished	report.txt	N/A	1 piece	Automatic only
01H09	QR	825	Finished	report.txt	37445	1 piece	Done
02A11	BZ	775	Finished	report.txt	38839	1 piece	Done
02D03	IU	1220	Finished	report.txt	44196	1 piece	Done
02D09	QR	1034	Finished	report.txt	42240	1 piece	Done
02F06	QH	1012	Finished	report.txt	43113	1 piece	Done
02F10	IU	925	Finished	report.txt	35283	1 piece	Done
02G04	BZ	1039	Finished *	report.txt	40751	1 piece	Done
02G12	QH	473	Finished	report.txt	41105	1 piece	Done
03G12	QS	456	Finished	report.txt	N/A	1 piece	Done
02H01	QR	839	Finished	report.txt	40885	1 piece	Automatic only
03a12	IC	756	Finished	report.txt	18095	1 piece	Automatic only
03C03	BF	657	finished	report.txt	N/A	1 piece	Done
03C03	IU	497	finished	report.txt	39387	1 piece	Done
03C11	UV	725	Finished	report.txt	39444	1 piece	Done
03C12	MC	453	Finished	report.txt	39990	1 piece	Done
03D03	QV	598	Finished	report.txt	39433	1 piece	Done
03E01	IU	849	Finished	report.txt	40502	1 piece	Done

B.4 Sumário de Contigs

Para acompanhar o andamento da montagem geral, foi elaborado um sumário para os contigs, como mostrado na figura a seguir. Nesse sumário são apresentados dados sobre cada contig obtido pela montagem mais recente, incluindo o número de identificação, o nome dado ao contig (formado pelo nome do cosmídeo mais a esquerda e nome do cosmídeo mais a direita do contig), o comprimento da sua seqüência, o número de reads de shotgun, e quais os possíveis contigs adjacentes (à direita e à esquerda), indicados pelo programa phrap. A tabela traz também apontadores Web para outras informações sobre

os contigs. O sumário de contigs mostrado na figura é do genoma já fechado, e por isso está cheio de "n/a"s na coluna `Contig links`.

Genome Contig Summary

Last update: Mon Jun 5 16:16:03 EST 2000

Number *(Link to analysis)*	Name *(Link to Map)*	Sequence Pick subsequence	# Shotgun	Contig links		Annotation
				Left	Right	
69	07C08-07F02	2678561	36550	n/a	n/a	n/a
68	megaplasmid	54132	2369	n/a	n/a	n/a
67	shotgun contig	2949	48	n/a	n/a	n/a
66	miniplasmid	3185	46	n/a	n/a	n/a
65	shotgun contig	2147	27	n/a	n/a	n/a
64	shotgun contig	2236	16	n/a	n/a	n/a
63	shotgun contig	762	16	n/a	C52	n/a
62	shotgun contig	1253	16	n/a	n/a	n/a
61	shotgun contig	1146	14	n/a	n/a	n/a
60	shotgun contig	1364	14	n/a	n/a	n/a
59	shotgun contig	1528	14	n/a	n/a	n/a
58	shotgun contig	2611	11	n/a	n/a	n/a
57	shotgun contig	303	12	n/a	n/a	n/a
55	shotgun contig	60	12	n/a	n/a	n/a
54	shotgun contig	698	10	n/a	n/a	n/a
53	shotgun contig	820	8	n/a	n/a	n/a
52	shotgun contig	574	6	n/a	C63	n/a
51	shotgun contig	738	6	n/a	n/a	n/a
49	shotgun contig	725	5	n/a	C37	n/a
48	shotgun contig	1247	5	n/a	n/a	n/a

Apêndice C

Endereços WEB

Nesta seção são colocados os endereços WEB dos projetos mencionados na dissertação e alguns outros relacionados ao trabalho.

- Projeto Genoma *Xylella fastidiosa*:
  http://aeg.lbi.ic.unicamp.br/xf
- Projeto Genoma *Xylella fastidiosa Pierce's Disease*:
  http://www.lbi.ic.unicamp.br/world/xf-grape
- Projeto Genoma *Xanthomonas*:
  Xanthomonas axonopodis pv citri
  Endereço geral: http://genoma4.iq.usp.br/xanthomonas
  Endereço com resultados de anotação:
  http://cancer.lbi.ic.unicamp.br/xantho
  Xanthomonas campestris pv campestris
  Endereço geral: http://genoma.fcav.unesp.br/xc-campestris
  Endereço com resultados de anotação:
  http://cancer.lbi.ic.unicamp.br/campestris
- Projeto Genoma *Agrobacterium tumefaciens*:
  Endereço geral: http://www.agrobacterium.org
  Endereço com resultados de anotação:
  http://cancer.lbi.ic.unicamp.br/agroC58
- Laboratório de Bioinformática - IC - Unicamp:
  http://www.lbi.ic.unicamp.br
- Rede ONSA:
  http://watson.fapesp.br/genoma3.htm

Bibliografia

[1] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[2] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[3] Applied Biosystems. *Automated DNA Sequencing*, 1998.

[4] Alex Bateman, Ewan Birney, Richard Durbin, Sean R. Eddy, Robert D. Finn, and Erik L. L. Sonnhammer. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Research*, 27(1):260–262, 1999.

[5] Alex Bateman, Ewan Birney, Richard Durbin, Sean R. Eddy, Kevin L. Howe, and Erik L. L. Sonnhammer. The Pfam Protein Families Database. *Nucleic Acids Research*, 28(1):263–266, 2000.

[6] Andreas D. Baxevanis and B. F. Francis Ouellette. *Bioinformatics - A Practical Guide to the Analysis of Genes and Proteins*. John Wiley & Sons, 1998.

[7] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, Barbara A. Rapp, and David L. Wheeler. Genbank. *Nucleic Acids Research*, 28(1):15–18, 2000.

[8] Mark Borodovsky and James McIninch. GenMark: Parallel Gene Recognition for both DNA Strands. *Computers & Chemistry*, 17(2):123–133, 1993.

[9] Adams M. D., Kelley J. M., Gocayne J. D., Dubnick M., Polymeropoulos M. H., Xiao H., Merril C. R., Wu A., Olde B., and Moreno R. F. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–1656, Jun 1991.

[10] A.C.R. da Silva, J.A. Ferro, F.C. Reinach, C.S. Farah, L.R. Furlan, R.B. Quaggio, C.B. Monteiro-Vitorello, M.A. Van Sluys, N.F. Almeida Jr., L.M.C. Alves, A.M. do Amaral, M.C. Bertolini, L.E.A. Camargo, G. Camarotte, F. Cannavan, J. Cardozo, F. Chambergo, L.P. Ciapina, R.M.B. Cicarelli, L.L. Coutinho, J.R. Cursino-Santos, H. El-Dorry, J.B. Faria, A.J.S. Ferreira, R.C.C. Ferreira, M.I.T. Ferro, E.F. Formighieri, M.C. Franco, C.C. Greggio, A. Gruber, A.M. Katsuyama, L.T. Kishi, R.P. Leite Jr., E.G. M. Lemos, M.V.F. Lemos, E.C. Locali, M.A. Machado, A.M.B.N. Madeira, N.M. Martinez-Rossi, E.C. Martins, J. Meidanis, C.F.M. Menck, C.Y. Miyaki, D.H. Moon, L.M. Moreira, M.T.M. Novo, V.K. Okura, M.C. Oliveira, V.R. Oliveira, H.A. Pereira Jr., A. Rossi, J.A.D. Sena, C. Silva, R.F. de Souza, L.A.F. Spinola, M.A. Takita, R.E. Tamura, E.C. Teixeira, R.I.D. Tezza, M.T. dos Santos, D. Truffi, S.M. Tsai, F.F. White, J.C. Setubal, and J.P. Kitajima. Complete genome sequences of two *Xanthomonas* pathogens with similar genomes but different host specificities. *Aceito para publicação - Nature*, 2002.

[11] Rick Danell. *HTML 4 - Professional Reference Edition - Unleashed.* Sams.net Publishing, 1998.

[12] Arthur L. Delcher, Douglas Harmon, Simon Kasif, Owen White, and Steven L. Salzberg. Improved microbial gene identification with Glimmer. *Nucleic Acids Research*, 27(23):4636–4641, 1999.

[13] Luiz Carlos Donadio and Celio Soares Moreira. *Citrus Variegated Chlorosis.* Fundecitrus and Fapesp, 1998.

[14] Jeffry Dwight, Michael Erwin, and Robert Nile. *Using CGI.* Que Corporation, second edition, 1997.

[15] Brent Ewing and Phil Green. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research*, 8:186–194, 1998.

[16] Brent Ewing, LaDeana Hillier, Michael C. Wendl, and Phil Green. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Research*, 8:175–185, 1998.

[17] David Gordon, Chris Abajian, and Phil Green. Consed: A Graphical Tool for Sequence Finishing. *Genome Research*, 8:195–202, 1998.

[18] Phil Green. Phrap Documentation. *www.phrap.org.*

[19] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology.* Cambridge University Press, 1997.

[20] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[21] Peter D. Karp, Monica Riley, Suzanne M. Paley, and Alida Pellegrini-Toole. EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Research*, 24(1):32–39, 1996.

[22] Peter D. Karp, Monica Riley, Milton Saier, Ian T. Paulsen, Suzanne M. Paley, and Alida Pellegrini-Toole. The EcoCyc and MetaCyc databases. *Nucleic Acids Research*, 28(1):56–59, 2000.

[23] Benjamin Lewin. *Genes VI.* Oxford University Press, New York, 1997.

[24] Harvey Lodish, David Baltimore, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira, and James Darnell. *Molecular Cell Biology.* Scientific American Books, New York, third edition, 1995.

[25] Todd M. Lowe and Sean R. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5):955–964, 1997.

[26] Marcos Machado. *Citrus Variegated Chlorosis (CVC), A New Destructive Citrus Disease In Brazil, And The Xylem-Limited Bacteria, Xylella fastidiosa.* Instituto Agronômico de Campinas, Centro de Citricultura Sylvio Moreira, Aug 1997.

[27] Christopher K. Matheus and K. E. van Holde. *Biochemistry.* The Benjamin/Cummings Publishing Company, 1990.

[28] G. Navarro and R. Baeza-Yates. Improving an Algorithm for Approximate String Matching. *Algorithmica*, 2000.

[29] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.

[30] ONSA/AEG. Em preparação. 2002.

[31] Steven L. Salzberg, Arthur L. Delcher, Simon Kasif, and Owen White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2):544–548, 1998.

[32] João C. Setubal and João Meidanis. *Introduction to Computational Molecular Biology.* PWS, Boston, 1997.

[33] A.J.G. Simpson et al. The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature*, 406:151–157, Jul 2000.

[34] T. F. Smith and M. S. Waterman. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[35] Erik L. L. Sonnhammer, Sean R. Eddy, Ewan Birney, Alex Bateman, and Richard Durbin. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research*, 26(1):320–322, 1998.

[36] Robert H. Tamarin. *Principles of Genetics*. Wm. C. Brown Publishers, fifth edition, 1996.

[37] Roman L. Tatusov, Michael Y. Galperin, Darren A. Natale, and Eugene V. Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):33–36, 2000.

[38] Larry Wall, Tom Christiansen, and Randal L. Schwartz. *Programming Perl*. O'Reilly & Associates, second edition, 1996.

[39] James D. Watson, John Tooze, and David T. Kurtz. *Recombinant DNA*. Labor, New York, second edition, 1992.

[40] Derek W. Wood, João C. Setubal, Rajinder Kaul, Dave E. Monks, João P. Kitajima, Vagner K. Okura, Yang Zhou, Lishan Chen, Gwendolyn E. Wood, Nalvo F. Almeida Jr., Lisa Woo, Yuching Chen, Ian T. Paulsen, Jonathan A. Eisen, Peter D. Karp, Donald Bovee Sr., Peter Chapman, James Clendenning, Glenda Deatherage, Will Gillet, Charles Grant, Tatyana Kutyavin, Ruth Levy, Meng-Jin Li, Erin McClelland, Anthony Palmieri, Christopher Raymond, Gregory Rouse, Channakhone Saenphimmachak, Zaining Wu, Pedro Romero, David Gordon, Shiping Zhang, Heayun Yoo, Yumin Tao, Phyllis Biddle, Mark Jung, William Krespan, Michael Perry, Bill Gordon-Kamm, Li Liao, Sun Kim, Carol Hendrick, Zuo-Yu Zhao, Maureen Dolan, Forrest Chumley, Scott V. Tingey, Jean-Francois Tomb, Milton P. Gordon, Maynard V. Olson, and Eugene W. Nester. The Genome of the Natural Genetic Engineer *Agrobacterium tumefaciens* C58. *Science*, 294:2317–2323, Dec 2001.

[41] E. J. Wood, C. A. Smith, and W. R. Pickering. *Life Chemistry and Molecular Biology*. Portland Press, London, 1997.

[42] Randy Jay Yager, George Reese, and Tim King. *MySQL and mSQL*. O'Reilly & Associates, first edition, 1999.

[43] Zheng Zhang, Alejandro A. Schäffer, Webb Miller, Thomas L. Madden, David J. Lipman, Eugene V. Koonin, and S.F. Altschul. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Research*, 26(17):3986–3990, 1998.