Conceitos Básicos de Biologia Molecular e de Seqüenciamento de DNA

Vagner Katsumi Okura*

1 Introdução

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: [3, 6, 7, 5, 2, 4].

2 Células

As células são consideradas as unidades vitais de vida. 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.

*Este documento constitui o Capítulo 2 de Okura, V. K. - Bioinformática de Projetos Genoma de Bactérias - Dissertação de Mestrado submetida ao Instituto de Computação - Unicamp - Fevereiro de 2002

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.

3 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.

4 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 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 (`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`.

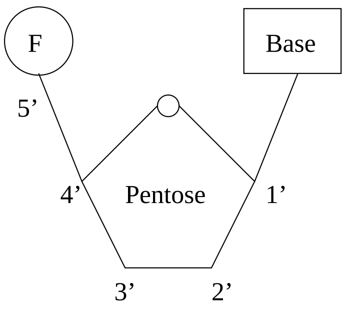


Figura 1: Esquema do nucleotídeo. O nucleotídeo possui três partes: fosfato (F), ribose e uma base. O fosfato liga-se ao carbono número 5 da pentose, enquanto que a base liga-se ao carbono número 1.

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). Por convenção, as seqüências são representadas na orientação $5' \to 3'$.

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). 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 $5' \to 3'$ 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 (**mRNA**), RNA transportador (**tRNA**) e RNA ribossomal (**rRNA**). A descrição das moléculas de RNA está na Seção 6.1.

4.1 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 $5' \to 3'$.

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.

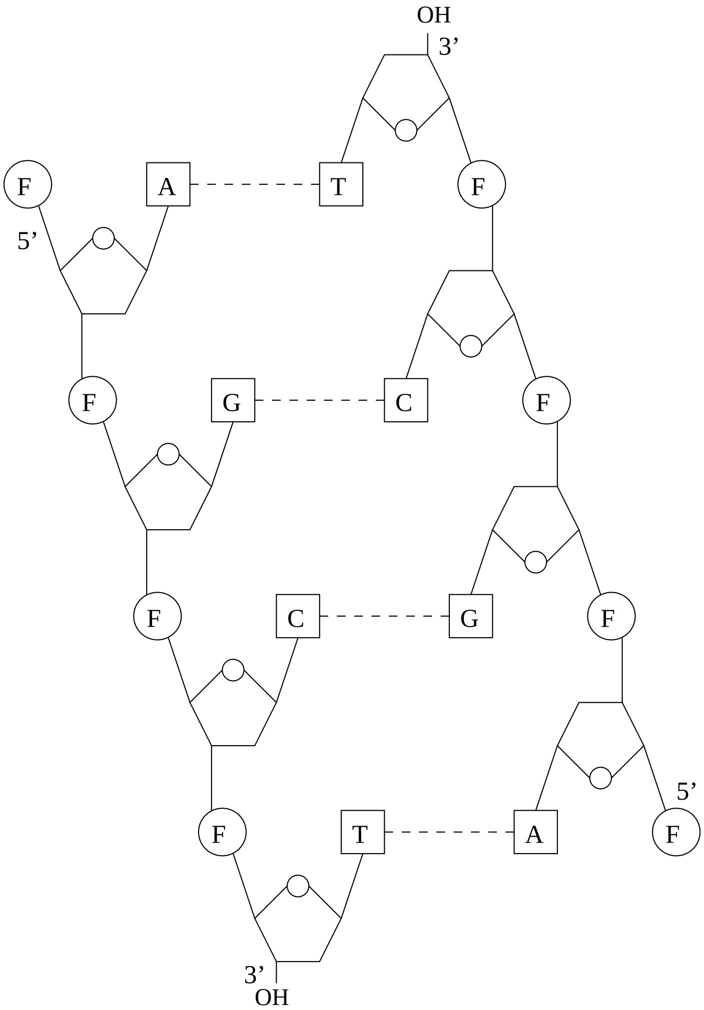


Figura 2: Representação linear da hélice dupla. As fitas possuem orientação inversa e ligam-se através de suas bases complementares: A com T e C com G.

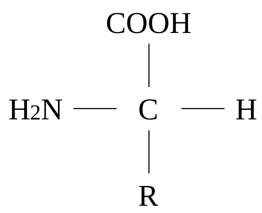


Figura 3: Estrutura química geral dos aminoácidos.

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 6.1.

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.

5 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 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 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.

6 Expressão de Genes

Como visto na Seção 3, 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 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.

6.1 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 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 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.

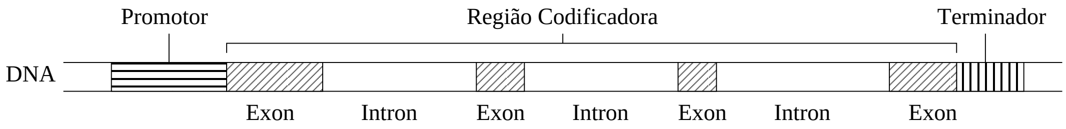


Figura 4: Estrutura de um gene de eucarioto. Genes de procarioto (em geral) não possuem *introns*.

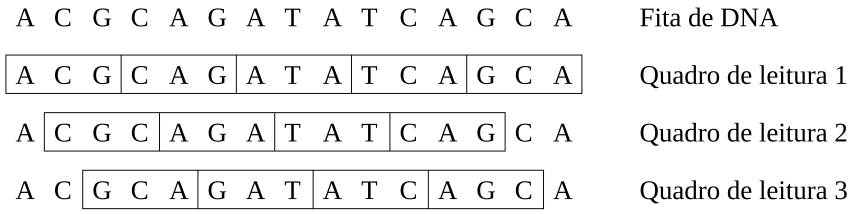
6.2 Tradução: Síntese de Proteínas

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 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 6.1) 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 seja encerrado prematuramente por um codon 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.

7 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 seqüências 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 7.2.3). 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.

7.1 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 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 de 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 5 apresenta um esboço de um gel fotografado, onde as

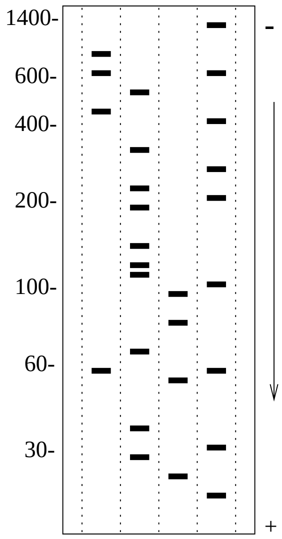


Figura 5: Esboço de autoradiograma de um gel. Os traços horizontais indicam a presença de DNA. As linhas verticais pontilhadas representam as canaletas onde são colocados os fragmentos.

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 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.

7.2 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.

7.2.1 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 chamade 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 7.4). 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*, diferente dos vetores descritos acima são incubados em células da bactéria *Escherichia coli.*

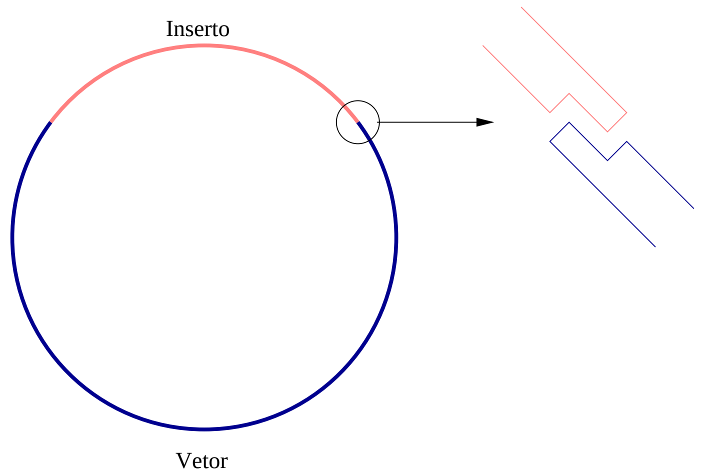


Figura 6: Inserto e vetor ligados pelo mesmo sítio de restrição.

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.

7.2.2 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 (vetores abertos), 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 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

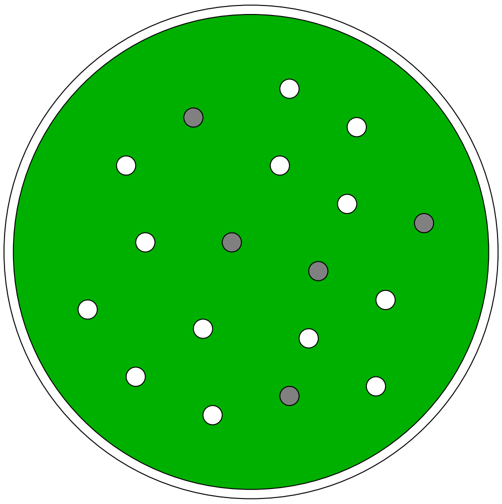


Figura 7: Ilustração de uma placa contendo colônias de bactérias. Os pontos brancos referem-se às colônias com insertos e os pontos sombreados referem-se às colônias sem insertos.

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;
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 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 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.

7.2.3 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.

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

7.3 Leitura de DNA

Os procedimentos que envolvem a leitura de bases de DNA são baseados na técnica de **eletroforese de gel** (Seção 7.1). 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):

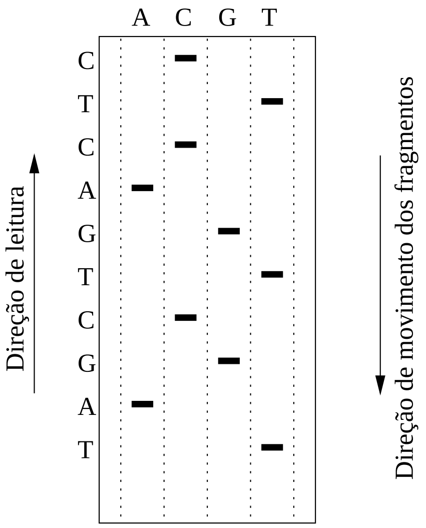


Figura 8: Esboço de uma imagem de gel de seqüenciamento de DNA, mostrando a seqüência de bases obtida, que no caso é lida de baixo para cima, como mostrado à esquerda da figura.

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

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 7.2.3), 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 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 [1], 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 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 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 9), e deslocamento de canaletas vizinhas. Os problemas no gel podem ser causados por uma série de fatores [1], 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.

7.4 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 chegar na seqüência de seus fragmentos, como ilustrado no diagrama da Figura 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 7.1) 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 7.2), onde a técnica de clonagem (Seção 7.2.1) é predominantemente usada. Nesta etapa, os fragmentos são inseridos em vetores, e em seguida colocados em bactérias pelo processo de transformação (Seção 7.2.2). 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 subclonagem (sub-fragmentação e clonagem) em fragmentos de DNA menores, como ilustrado pela linha tracejada na Figura 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 7.3. 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

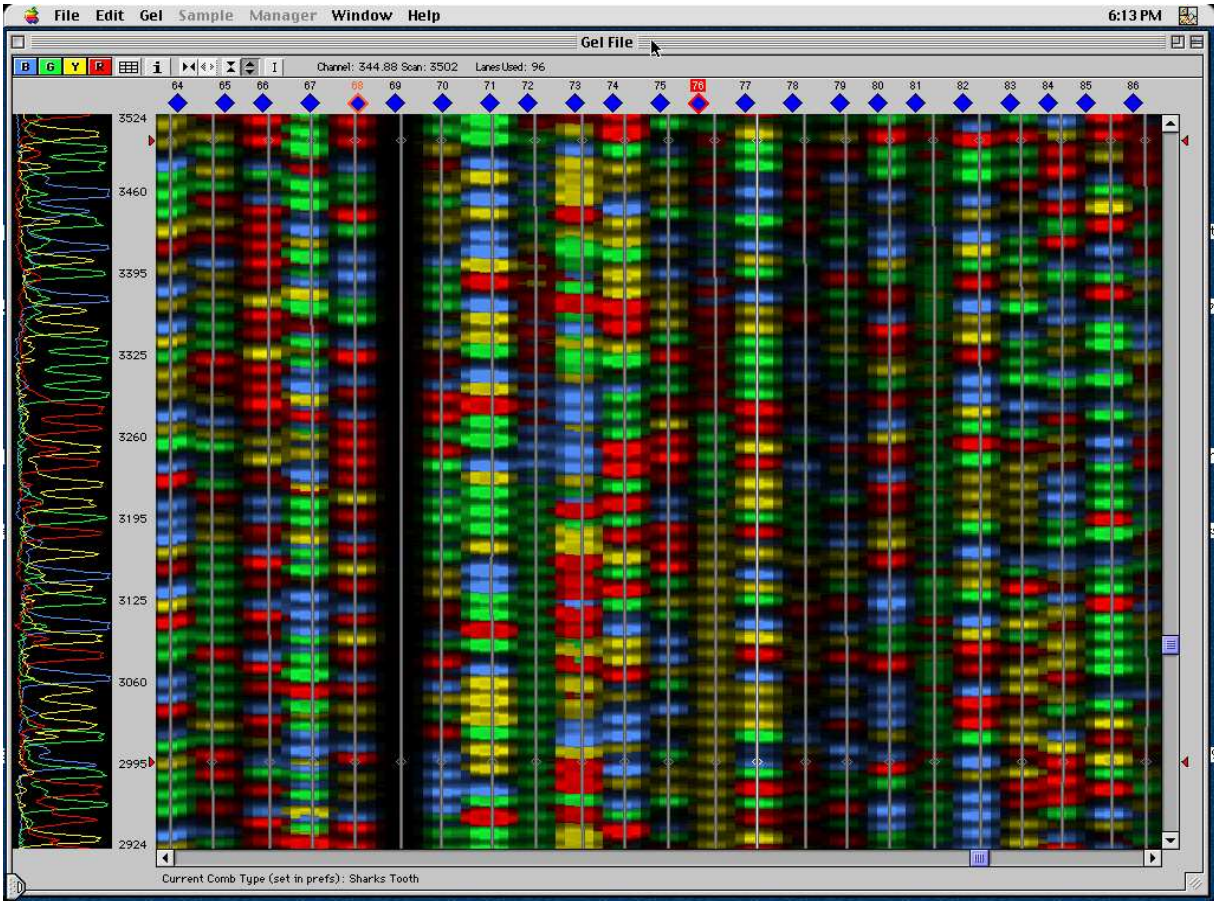


Figura 9: Figura de um arquivo de gel obtido com seqüenciamento automático de DNA. Na vertical estão as canaletas. As cores representam cada uma das bases: verde como `A`, azul como `C`, amarelo como `G`, vermelho como `T`. À esquerda, está a visualização do cromatograma da seqüência correspondente a canaleta 77.

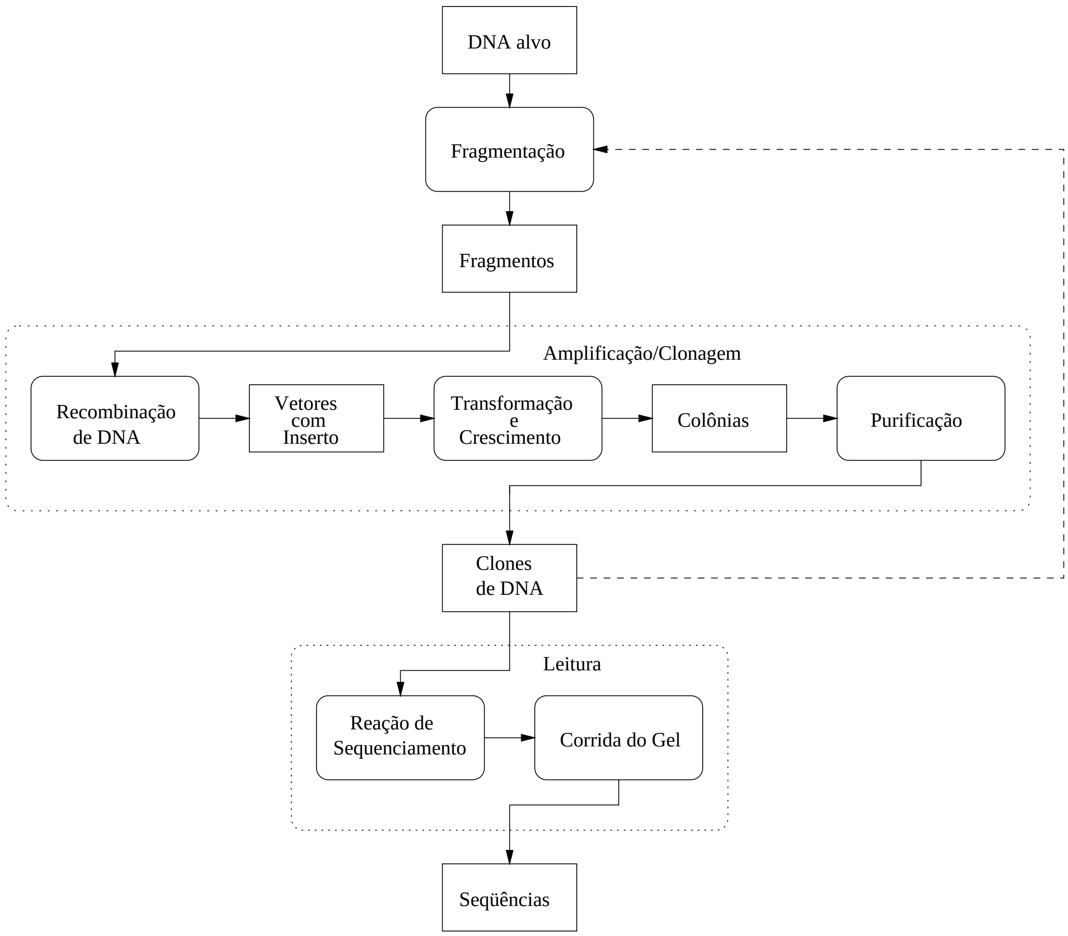


Figura 10: Diagrama ilustrando o processo global de seqüenciamento.

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 1: Código genético: mapeamento de codons para aminoácidos.

Primeira posição	**Segunda posição** G	A	C	U	**Terceira posição**
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

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 7.3, 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 7.2.3, 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.

Referências

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

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

[3] 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.

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

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

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

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