DOGMA CENTRAL DA BIOLOGIA MOLECULAR E INTRODUÇÃO À BIOINFORMÁTICA

w.Educacional

Brasília, 2011.

Elaboração

Luis Fernando Reys

Joci Neuby Alves Macedo

Julio Cesar Pissuti Damalio

Produção

Equipe Técnica de Avaliação, Revisão Linguística e Editoração

W Educacional Editora e Cursos Ltda.
Av. L2 Sul Quadra 603 Conjunto C
CEP 70200-630
Brasília-DF
Tel.: (61) 3218-8314 – Fax: (61) 3218-8320
www.ceteb.com.br
equipe@ceteb.com.br | editora@WEducacional.com.br

SUMÁRIO

APRESENTAÇÃO

Caro aluno

A proposta editorial deste Caderno de Estudos e Pesquisa reúne elementos que se entendem necessários para o desenvolvimento do estudo com segurança e qualidade. Caracteriza-se pela atualidade, dinâmica e pertinência de seu conteúdo, bem como pela interatividade e modernidade de sua estrutura formal, adequadas à metodologia da Educação a Distância – EaD.

Pretende-se, com este material, levá-lo à reflexão e à compreensão da pluralidade dos conhecimentos a serem oferecidos, possibilitando-lhe ampliar conceitos específicos da área e atuar de forma competente e conscienciosa, como convém ao profissional que busca a formação continuada para vencer os desafios que a evolução científico-tecnológica impõe ao mundo contemporâneo.

Elaborou-se a presente publicação com a intenção de torná-la subsídio valioso, de modo a facilitar sua caminhada na trajetória a ser percorrida tanto na vida pessoal quanto na profissional. Utilize-a como instrumento para seu sucesso na carreira.

Conselho Editorial

ORGANIZAÇÃO DO CADERNO DE ESTUDOS E PESQUISA

Para facilitar seu estudo, os conteúdos são organizados em unidades, subdivididas em capítulos, de forma didática, objetiva e coerente. Eles serão abordados por meio de textos básicos, com questões para reflexão, entre outros recursos editoriais que visam a tornar sua leitura mais agradável. Ao final, serão indicadas, também, fontes de consulta, para aprofundar os estudos com leituras e pesquisas complementares.

A seguir, uma breve descrição dos ícones utilizados na organização dos Cadernos de Estudos e Pesquisa.

Provocação

Pensamentos inseridos no Caderno, para provocar a reflexão sobre a prática da disciplina.

Para refletir

Questões inseridas para estimulá-lo a pensar a respeito do assunto proposto. Registre sua visão sem se preocupar com o conteúdo do texto. O importante é verificar seus conhecimentos, suas experiências e seus sentimentos. É fundamental que você reflita sobre as questões propostas. Elas são o ponto de partida de nosso trabalho.

Textos para leitura complementar

Novos textos, trechos de textos referenciais, conceitos de dicionários, exemplos e sugestões, para lhe apresentar novas visões sobre o tema abordado no texto básico.

Sintetizando e enriquecendo nossas informações

Espaço para você fazer uma síntese dos textos e enriquecê-los com sua contribuição pessoal.

Sugestão de leituras, filmes, *sites* e pesquisas

Aprofundamento das discussões.

Praticando

Atividades sugeridas, no decorrer das leituras, com o objetivo pedagógico de fortalecer o processo de aprendizagem.

Para (não) finalizar

Texto, ao final do Caderno, com a intenção de instigá-lo a prosseguir com a reflexão.

Referências

Bibliografia consultada na elaboração do Caderno.

INTRODUÇÃO

Até meados do século passado as Ciências Biológicas se questionavam sobre qual era a molécula fundamental da vida que determina as características hereditárias e sobre quais as características químicas que ela possuía. Com as pesquisas desenvolvidas principalmente por Watson e Crick em 1953, determinou-se a estrutura molecular do DNA, molécula responsável pelo armazenamento da informação genética. Graças a esta descoberta, somada aos estudos posteriores, surgiu o postulado do **Dogma Central da Biologia Molecular**, o qual sumariza que a informação genética contida no DNA das células e dos vírus é preservada, transmitida e traduzida.

No final do século XX e início do século XXI, a Biologia Molecular desenvolveu-se rapidamente e novos métodos e técnicas de manipulação do DNA foram criados. O salto tecnológico, associado a esta área da ciência, permitiu determinar a composição do genoma de várias espécies, incluindo o do código genético humano, assim como de espécies de interesse comercial/econômico e até de microrganismos (fungos, bactérias e vírus). Quando falamos em decifrar genomas (e/ou seus "omas" ou "ômicos" derivados – transcriptomas, proteomas, metabolomas), falamos em "ler" bioquimicamente as milhões de "letras" (pares de bases) contidas numa molécula de DNA particular de uma espécie de interesse. Isto gera, inevitavelmente, uma enorme quantidade de informação, a qual precisa ser processada, organizada e armazenada para depois ser interpretada e analisada. Em resposta a esta necessidade, surge o que hoje conhecemos como **Bioinformática**, uma das áreas das Ciências Biológicas mais desenvolvidas nas últimas décadas. Hoje em dia, existem bancos de dados biológicos contendo informações genéticas valiosas de quase 1000 espécies disponíveis através de sites na internet. Todas as informações biológicas relacionadas com o DNA, RNA e Proteínas estão disponibilizadas para quem quiser pesquisar sobre um gene em particular, determinar a sua sequência, a sua função, compará-lo com outros genes de outras espécies, ou mesmo para comparar genomas completos entre espécies relacionadas evolutivamente, dentre outras utilidades.

Esta apostila esta dividida em duas unidades. A primeira contém o conhecimento teórico básico sobre os processos bioquímicos que definem o Dogma Central da Biologia Molecular. São elas: a replicação do DNA, a transcrição e a tradução. Já a segunda unidade contém informações básicas sobre Bioinformática e, também, mostra, de maneira prática e com um breve tutorial, como acessar as diferentes informações contidas nos principais bancos de dados disponíveis na internet.

Boa leitura e bons estudos!

Objetivos

» Estudar o conceito de fluxo de informação genética por meio do postulado pelo Dogma Central da Biologia Molecular.

» Entender os experimentos realizados pelo postulado em estudo.

» Conhecer os processos moleculares envolvidos na transmissão da informação e das biomoléculas envolvidas no processo.

» Compreender o que é o Gene e o Código Genético.

» Entender os principais eventos envolvidos no processo de replicação, transcrição do DNA e na tradução ou síntese de proteínas.

» Entender os processos moleculares que mudaram, na atualidade, o que foi postulado pelo Dogma Central da Biologia Molecular.

» Reconhecer a importância da Bioinformática para a ciência.

» Apreender os conceitos básicos sobre Genômica e seus derivados "ômicos".

» Reconhecer o grau de envolvimento do Brasil neste tipo de pesquisa.

» Definir os Bancos de Dados Biológicos entendendo suas funções.

» Acessar os diferentes Bancos de Dados através de um Tutorial, usando como modelo o NCBI.

UNIDADE I

DOGMA CENTRAL DA BIOLOGIA MOLECULAR

CAPÍTULO 1

O dogma central de Francis Crick

O Dogma central foi postulado por Francis Crick anos depois de ter participado, junto com James Watson, da elucidação da estrutura do DNA. Estudando a relação entre a informação contida no DNA (na forma de alfabeto com 4 letras – bases nucleotídicas A, T, C e G), as proteínas e os ácidos ribonucléicos (RNAs), surgiu a interrogação sobre como a informação contida na molécula de DNA fluía para gerar uma proteína, em outras palavras, sobre como ocorre o fluxo da informação genética. Na época, Crick contava apenas com dados bioquímicos e estruturais do DNA e com alguns dados bioquímicos sobre os outros biopolímeros (proteínas e ácidos ribonucléicos: tRNA, rRNA, mRNA). Não se conheciam, na época, todos os detalhes dos processos de replicação, transcrição e síntese de proteínas. Portanto, baseando-se majoritariamente em suposições teóricas, Crick postulou a "**Hipótese da sequência**" e que é mais conhecida como o **Dogma Central da Biologia Molecular** (Figura 1).

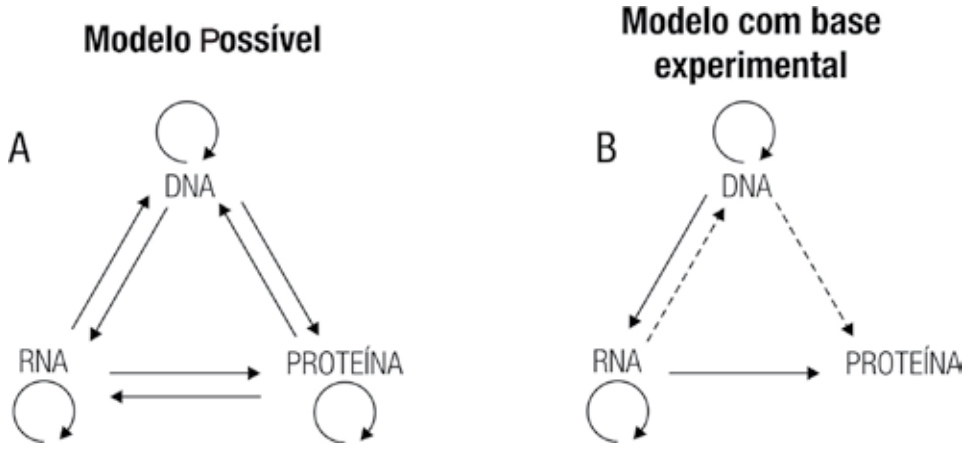

Figura 1. Dois esquemas representando as principais ideias propostas pelo Dogma Central da Biologia Molecular. no Esquema A aparece o modelo possível de transferência do código genético, enquanto em B está retratado o modelo comprovado experimentalmente e que até hoje é aceito.

Figura extraída e modificada do artigo publicado por Crick em 1958. ("Central Dogma of Molecular Biology", Nature, 227: 561–563/ http://www.nature.com/nature/journal/v227/n5258/ pdf/227561a0.pdf).

A **Hipótese da Sequência**, na sua forma mais simples, assume que a especificidade de uma porção de ácido nucleico é expressa somente pela sequência de seus pares de bases e, por sua vez, estas contém o código necessário para gerar uma proteína. Já o **Dogma Central** versa sobre o fluxo da informação contida no código genético e postula que a informação de uma sequência de ácido nucleico, uma vez transformada em proteína, não pode retornar para o ácido nucleico. Em maiores detalhes, a transferência de informação de ácido nucleico para outro ácido nucleico e deste para proteína pode ser possível, mas de proteína para proteína ou desta para o ácido nucleico seria impossível. É importante salientar que aqui o termo informação

significa a precisa determinação da sequência, seja na forma das 4 bases nucleotídicas ou da sequência de aminoácidos.

No modelo possível (Figura 1 A) de fluxo de informação genética podemos observar que todos os três biopolímeros, DNA, RNA e proteínas, têm a capacidade de se replicar, por um lado, e, por outro lado, de transmitir a sua informação para gerar outra biomolécula. Isso significa que o fluxo de informação flui em todos os sentidos. Por exemplo, a sequência de aminoácidos de uma proteína poderia não só originar cópias de si mesma como também de DNA e RNA (o mesmo acontecendo com estas últimas). Já no modelo com base experimental (Figura 1 B) postulado por Crick, observamos que este caminho não é possível.

Na mesma época, Watson postulou algo parecido sobre o fluxo de informação genética e que, até hoje, causa confusão até nos livros de texto de Biologia Molecular. Portanto, o postulado verdadeiro, mesmo que não completamente correto (como veremos mais adiante) ou fatível de ser modificado, é o originalmente postulado por Crick em 1958 (Figura 1 B).

Toda hipótese e teoria, em geral, são passíveis de crítica, modificações e até mesmo rejeição pela comunidade científica geral.O **Dogma Central**, por sua vez, não escapou de todos esses embates. Até hoje, ele tem sido considerado o pontapé inicial, uma espécie de arcabouço, onde os biólogos moleculares tentam acomodar as peças do quebra-cabeça para se entender o fluxo de informação genética dentro de uma célula qualquer.

CAPÍTULO 2

A replicação do DNA

A replicação do DNA consiste na síntese completa e idêntica do DNA de uma célula e só acontece a cada ciclo de divisão celular (Mitose e Meiose). Da replicação do DNA depende a subsistência e perpetuação de uma célula (organismos). A capacidade que cada célula tem de preservar o seu material genético e transmiti-lo para a geração seguinte dependerá deste processo. A transmissão da informação genética implica que haja uma cópia do material genético de uma célula, em outras palavras, gera uma cópia fiel do DNA de maneira tal que este passe para a geração seguinte.

Como já visto na disciplina anterior, de acordo com o ponto de vista químico, o DNA é um longo polímero de unidades simples (monômeros) de nucleotídeos, cuja cadeia principal é formada por moléculas de açúcares e fosfato intercalados e unidos por ligação fosfodiéster. Ligada à molécula de açúcar está uma das quatro bases nitrogenadas, A, T, C ou G.

A composição química pouco variável (apenas quatro bases combinadas sequencialmente), em princípio, fazia o DNA pouco atraente, já que, até então, pensava-se que as proteínas fossem as únicas moléculas responsáveis pela variação de formas de vida. Esta suposição não foi mais sustentada depois que o DNA tomou seu devido lugar como a molécula da vida e do material hereditário. De fato, para responder às necessidades de sobrevivência e replicação de uma célula é necessário que a informação genética seja mantida ou conservada de geração em geração. Portanto, a suposta monotonia do DNA seria uma consequência dessa necessidade.

Analisando-se a estrutura da dupla hélice de DNA foi possível prever o mecanismo pelo qual esta molécula seria copiada em cada ciclo de divisão celular. Segundo o modelo de Watson, Crick e Franklin, o processo de replicação estaria sustentado pela complementaridade de bases entre as duas fitas constituintes (Figura 2) de modo que uma fita sirva como molde para gerar a complementar.

Figura 2. Modelo da estrutura de DNA revelada por Watson, Crick e Franklin em 1954.

Figura obtida e modificada a partir de: http://pt.wikipedia.org/wiki/Ficheiro:DNA_Overview.png, em 17/05/2011.

Três hipóteses ajudavam a explicar o processo da replicação.

» **Hipótese Semiconservativa**: as fitas do DNA seriam separadas, atuando, cada uma, como molde para a síntese de uma fita nova (modelo previsto por Watson e colaboradores).

» **Hipótese Conservativa**: ambas as fitas atuariam como molde da reação de replicação sem que houvesse separação das fitas.

» **Hipótese Dispersiva**: o DNA seria replicado a partir de fragmentos dele próprio previamente gerados por clivagem da molécula.

Estas três hipóteses geraram três modelos diferentes para serem testados no laboratório. Cada modelo predizia um resultado diferente na constituição tanto do DNA "progenitor" – o molde –, quanto do DNA "filho" ou cópia, como mostra a Figura 3.

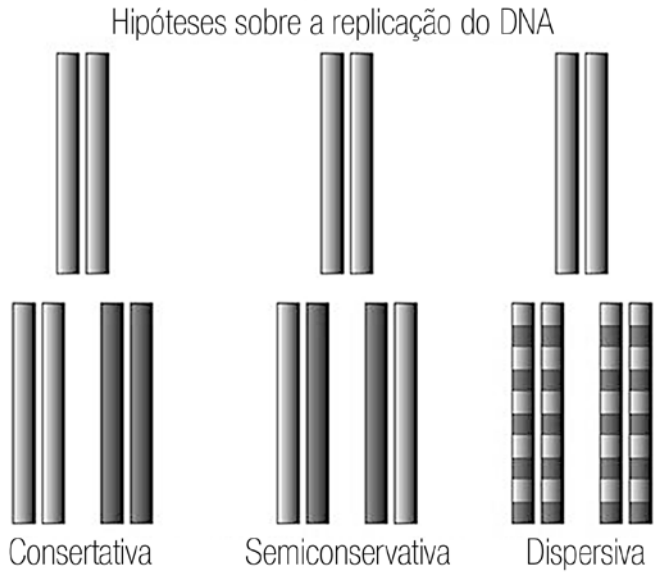

Figura 3. Resultados esperados para a replicação do DNA em função das hipóteses possíveis.

Figura obtida e modificada a partir de: http://is.wikipedia.org/wiki/Mynd:DNA-Replikationsmechanismen.svg, em 26/05/11.

Meselson e Stahl, por meio de um elegante experimento, demonstraram que a replicação do DNA é **semiconservativa**, ou seja, cada fita do DNA serve como um molde para a síntese de uma nova fita, produzindo duas novas moléculas de DNA (Figura 4). Assim, a molécula replicada é constituída de duas fitas: uma "parental" e outra complementar, a "filha".

Em um experimento, Meselson e Stahl fizeram crescer bactérias da espécie *Escherichia coli* em meio de cultura e com suplementação de uma fonte de nitrogênio radioativo (isótopo N^{15}). A primeira parte do experimento consistiu em determinar a incorporação do isótopo ao DNA. Após várias gerações de cultura (lembrando que uma geração de bactérias surge a cada 20-30 minutos), o isótopo radiativo se incorporou em toda a molécula do DNA bacteriano. Este DNA foi extraído e purificado da cultura bacteriana. Ainda, sua massa for determinada através da metodologia de ultracentrifugação em gradiente de densidade e comparada com a massa do DNA bacteriano controle (contendo o isótopo mais comum e mais leve – N^{14}).

Como resultado, observou-se que geração após geração o DNA ficava com massa maior quando comparada como o controle e, no final do experimento, 100% das moléculas correspondiam à forma denotada como DNA^{N15}. Em seguida, uma amostra de bactérias contendo DNA^{N15} foi transferida para um meio de cultura suplementado com N14 de maneira que, a cada geração, as bactérias o incorporassem na sua molécula de DNA. Se a replicação fosse semiconservativa, seria esperado observar, logo após uma geração (ou ciclo de replicação), moléculas de DNA constituídas por 50% de fitas pesadas e 50% de fitas leves, ou seja, DNA com massa intermediária (DNA^{N14-15}), entre pesada (DNA^{N15}) e leve (DNA^{N14}). Os resultados obtidos pela ultracentrifugação do DNA coincidiram com os resultados esperados, comprovando a hipótese da replicação semiconservativa do DNA.

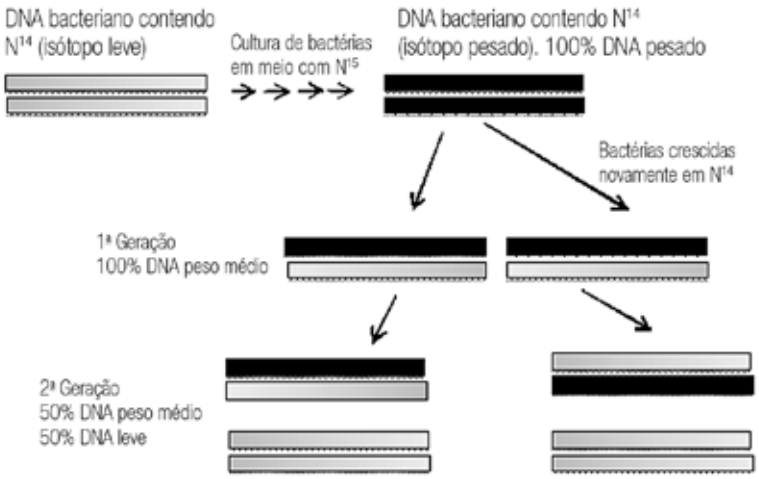

Figura 4. Esquema representativo do experimento realizado por Meselson e Stahl em 1958 que demonstra que a replicação do DNA é semiconservativa.

Pensando nas características dos átomos que constituem o DNA, por que neste experimento foi utilizado apenas nitrogênio radiativo e não fósforo ou carbono radioativos?

As grandes descobertas da Biologia Molecular têm sido realizadas graças à utilização da bactéria *Escherichia coli* como modelo de estudo. Sem ela, não teria sido possível avançar tão rapidamente no conhecimento dos processos bioquímicos fundamentais que acontecem em todos os seres vivos.

Animações dos experimentos de Meselson e Stahl.

Disponíveis em: <http://www.dnalc.org/view/15879-Semi-conservative-replication.html>

<http://www.youtube.com/watch?v=mfnDVV518es>

<http://www.youtube.com/watch?v=OzSIGxKWqoo&feature=related>

Método de gradiente de densidade de CsCl (em inglês):

Disponível em: <http://www.youtube.com/watch?v=UvGCXtX5MAM&feature=related>

Origem de replicação

O genoma de uma bactéria, como a *Ɛ.coli*, é constituído por mais de nove milhões de pares de bases. A cada divisão celular, a bactéria tem que copiar esses nove milhões de letras eficientemente, de maneira que a bactéria filha receba uma cópia do genoma. Mas onde começa a replicação do DNA? A replicação começa em qualquer ponto da molécula ou em vários pontos ao mesmo tempo?

Estas questões começaram a ser elucidadas a partir dos estudos feitos utilizando como modelo de replicação a bactéria *Ɛ.coli* e radioisótopos para marcar o seu DNA. Quando as bactérias eram cultivadas em meio contendo o radioisótopo de timidina marcada com trítio (3H, timidina tritiada) e o seu DNA isolado e revelado em papel fotográfico, era possível evidenciar o DNA circular (radioativo) da bactéria (Figura 5). À medida que o mesmo se replicava, formava-se uma alça extra, que representa a formação de duas fitas radioativas filhas, sendo cada uma complementar a uma fita parental. Uma ou ambas as extremidades da alça são pontos dinâmicos chamados de **forquilhas de replicação**. Esta forquilha é o sítio onde o DNA parental está sendo desenovelado e as fitas separadas, rapidamente, para serem replicadas.

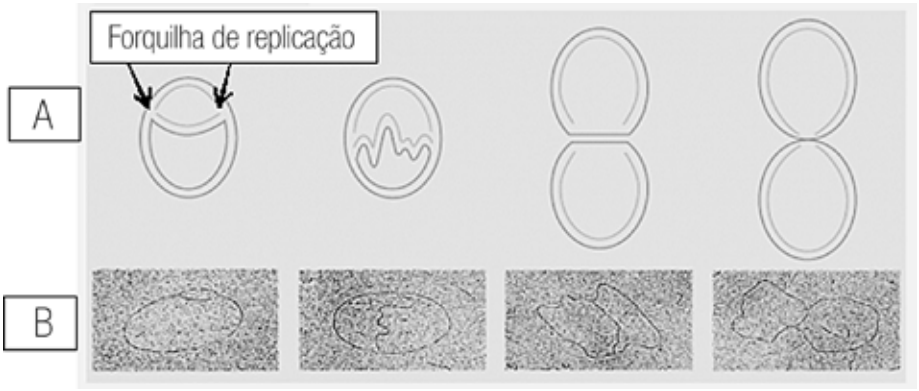

Figura 5. Esquema representativo da replicação do DNA circular bacteriano observado em *E.coli* na presença de timidina tritiada. Evidencia-se que, durante a replicação, forma-se uma alça extra de DNA com a formação da chamada Forquilha de Replicação

Extraído de http://www2.iq.usp.br/docente/goldberg/veterinaria2006/Aula_2_vet_2006.pdf, em 18/05/2011.

O fato de que o DNA circular, durante a replicação, adota a forma da letra grega teta (θ), permitiu inferir que, nas bactérias, existe apenas um único sítio de início ou de origem de replicação. Além do mais, o processo é bidirecional, isto é, quando a dupla fita se separa uma da outra, formam-se duas forquilhas de replicação, cada uma sintetizando DNA novo em ambos os sentidos. Como o genoma bacteriano contém apenas um "cromossomo" e uma única origem de replicação, ele é denominado um *replicon*. Assim, as unidades de replicação e segregação são coincidentes. Quando se inicia a replicação, ela continua até copiar o cromossomo inteiro, uma vez por cada ciclo de divisão celular bacteriana.

Uma nova fita de DNA é sintetizada sempre no sentido 5′ → 3′, de tal maneira que, na extremidade 3′, sempre fique um grupo hidroxila (OH) disponível para o alongamento do DNA. Mas, como o molde da reação de síntese é a fita complementar e esta é antiparalela, a leitura é feita no sentido 3′ → 5′, como esquematizado na Figura 6.

Figura 6. Sentido da replicação do DNA a partir das fitas molde. Cada fita, uma vez separada da complementar, serve como molde para a síntese da nova fita sempre seguindo o sentido 5′ → 3′.

Características da origem de replicação

Sabe-se mais sobre o sítio de origem de replicação em procariotos do que em eucariotos. O sítio de origem de replicação – ou oriC bacteriano – consiste em uma sequência de DNA de aproximadamente 245 pb (pares de bases) presente em um sítio especifico do cromossomo de *Ɛ.coli* e, também, em outros elementos extracromossômicos, como, por exemplo, os denominados plasmídeos. A função do oriC é a de controlar o início da replicação do DNA. A comparação entre as sequências de oriCs de diversas espécies de bactérias revelou que, em todas elas, existe um grau de conservação alto com presença de repetições de 9 pb e 13 pb ricas em AT (Figura 7). Como veremos mais adiante, estas repetições, conservadas em todas as espécies de bactérias, servem como sítio de reconhecimento e ligação da enzima encarregada de iniciar a replicação, a enzima DnaA. Outra característica importante é a presença de um alto conteúdo de nucleotídeos A+T na vizinhança do oriC. Esta região rica em AT facilita a separação das fitas do DNA.

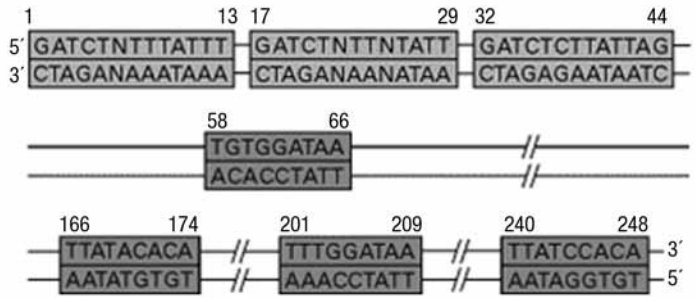

Figura 7. Sequência consenso do oriC determinada por meio da comparação das sequências de DNA de várias espécies de bactérias, dentre elas a E.coli.

Extraído de http://www.ncbi.nlm.nih.gov/books/NBK21650/figure/A3175/?report=objectonly, em 02/06/11.

Na levedura *Saccharomyces cerevisiae*, representante de eucarioto, o sítio de origem de replicação é chamado de ARS (*autonomously replicating sequences*) ou sequências autônomas replicantes. As ARS são mais curtas que as oriCs e contêm, também, subdomínios que desempenham funções distintas durante o início da replicação. A principal diferença entre procariontes e eucariontes está no número de ARS presentes. Como sabemos, o genoma de uma levedura, por exemplo, é mais complexo que o da bactéria. O genoma de leveduras está dividido em 17 cromossomos, cada um contendo várias destas sequências – ARS. Estima-se que existam mais de 400 ARS espalhadas pelo seu genoma (Figura 8).

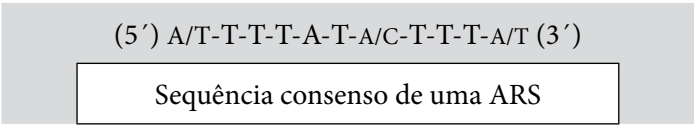

Figura 8: Sequência típica de uma ARS encontrada em leveduras.

Modificado de Molecular Cell Biology. 4th edition. Lodish H, Berk A, Zipursky SL, et al. New York: W. H. Freeman; 2000. http://www.ncbi.nlm.nih.gov/books/NBK21650/#3179, 02/09/2011.

A sequência consenso da origem de replicação em eucariontes superiores, como o *Homo sapiens*, é mais complexa de ser determinada, mas alguns estudos indicam que esta seria análoga à das leveduras. Levando em conta as distâncias evolutivas entre as espécies mencionadas, as respectivas sequências responsáveis pelo início da replicação compartilham algumas características.

- As origens de replicação são constituídas de sequências curtas e repetitivas.
- As sequências repetitivas são sítios específicos para a ligação de proteínas específicas envolvidas na iniciação da replicação do DNA.
- São ricas em pares de bases AT. Esta característica está relacionada à facilidade que este tipo de sequência tem de se separar durante a formação da bolha de replicação (lembrando que a ligação entre pares A-T é mais fraca do que entre C-G).
- As proteínas que se ligam nas origens de replicação atuam diretamente com a maquinaria que inicia o processo de replicação de DNA.

Uma das principais diferenças entre os genomas de procariontes e eucariontes é a presença de uma única origem de replicação nos primeiros e múltiplos sítios nos últimos. Qual a explicação plausível para o grande número destes sítios espalhados pelos genomas dos eucariontes?

A replicação do DNA é semidescontínua

O fato da dupla hélice de DNA ser constituída por duas fitas antiparalelas representa um problema para a replicação. À medida que a forquilha de replicação vai avançando, duas fitas complementares têm que ser sintetizadas ao mesmo tempo nas duas fitas simples expostas. Como visto, a forquilha move-se no sentido 5′ → 3′em uma das fitas e no sentido oposto na outra. Também, como visto, a síntese ocorre só no sentido 5′ → 3′. Este problema é resolvido mediante a síntese de pequenos fragmentos curtos sobre a fita que vai na direção oposta à forquilha de replicação. Portanto, ao se tratar apenas de uma das fitas, a replicação é dita semidescontínua. Esta afirmação surgiu graças às pesquisas feitas por Reiji Okazaki e colaboradores na década de 1960.

Okazaki observou que, durante a replicação do DNA, a dupla hélice não era copiada de maneira contínua e sim de forma descontínua. Enquanto uma das fitas era copiada de maneira contínua, na complementar,

a cópia ocorria em trechos curtos. A fita copiada continuamente é chamada de fita líder e a síntese da fita filha ocorre na mesma direção da forquilha de replicação. A fita descontínua é, também, chamada de atrasada e a fita filha é sintetizada em pequenos fragmentos de DNA de aproximadamente 200-1000 nucleotídeos (ou pares de bases – pb) chamados de Fragmentos de Okazaki (Figura 9).

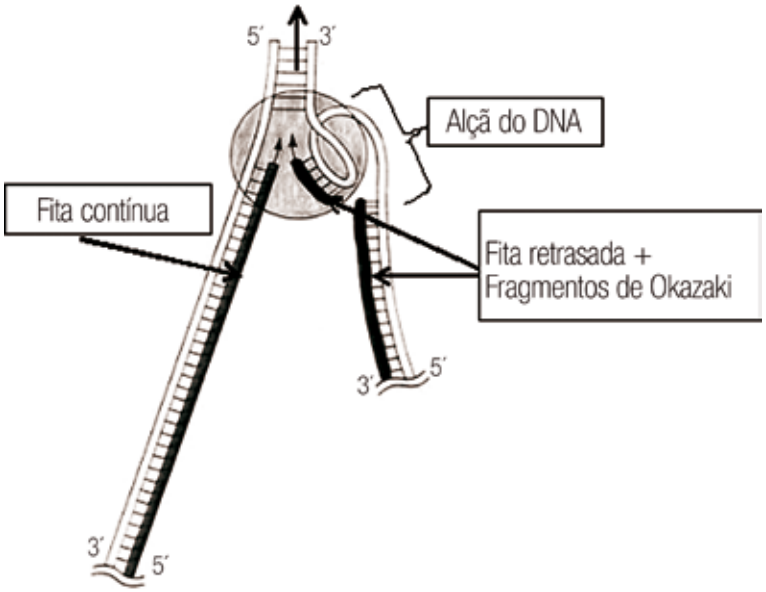

Figura 9. Forquilha de replicação mostrando ambas as fitas de DNA sendo copiadas. O sentido antiparalelo de ambas cria um problema à medida que a replicação/síntese das novas fitas acontece. Em uma das fitas, a síntese é contínua, enquanto na outra é descontínua. Por isso, a replicação é considerada semidescontínua e os fragmentos de DNA sintetizados são conhecidos como Fragmentos de Okazaki.

Modificado de http://media.wiley.com/Lux/56/24356.nfg012.gif.

Etapas da replicação e enzimas envolvidas

Os mecanismos celulares responsáveis pela replicação do DNA foram desvendados primeiro em bactérias, mas, recentemente, estes também têm sido desvendados em leveduras e eucariontes superiores com menos detalhes. Considerando que as informações sobre o processo de replicação são bem conhecidas em bactérias, tomaremos estas como modelo Tendo sempre em consideração que nos eucariontes os mecanismos e etapas envolvidas são análogos.

Do ponto de vista bioquímico, a replicação do DNA começa com a descompactação e abertura da dupla hélice a ser copiada. A abertura envolve a quebra das ligações de hidrogênio entre as bases. Uma vez expostas, as bases de cada uma das fitas de DNA separadas servem como molde para orientar a inserção das bases nucleotídicas (entrando na forma de nucleotídeo trifosfato) complementares à fita molde. Cada novo nucleotídeo é covalentemente ligado à fita que é sintetizada. A ligação resultante é do tipo fosfodiéster entre a extremidade hidroxila 3´ livre da base predecessora e o fosfato do nucleotídeo a ser adicionado (Figura 10).

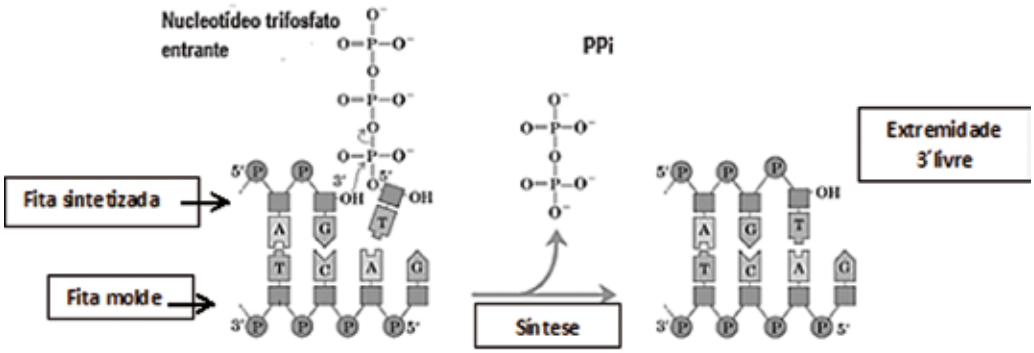

Figura 10. Esquema mostrando a síntese de DNA durante a replicação. A fita molde recebe um novo nucleotídeo (trifosfato). A fita complementar é sintetizada seguindo a regra de complementaridade de bases, A-T e C-G. Cada nucleotídeo é incorporado e ligado na extremidade 3´ do nucleotídeo predecessor. Um fosfato inorgânico (PPi) é liberado durante a reação.

Extraído e modificado de wikipedia http://en.wikipedia.org/wiki/Primer_%28molecular_biology%29,01/09/11

A replicação é um evento enzimático complexo e envolve, nas *E. coli,* mais de 20 enzimas. Entre as principais, encontramos as denominadas DNAs e RNAs polimerases. Cada uma tem uma função específica durante o processo, mas, de modo geral, todas contribuem para que, durante a replicação, não haja erros de cópia e, assim, a informação genética seja preservada a cada ciclo de divisão celular. O processo de replicação é um evento complexo onde todas as enzimas envolvidas atuam de maneira concatenada e ordenada. Elas são denominadas de SISTEMA DE REPLICAÇÃO DE DNA ou simplesmente de REPLISSOMO. O processo de replicação é altamente dinâmico, mas, para poder ser entendido em detalhes, pode ser dividido em três etapas distintas.

- **Iniciação**: o DNA é separado nas duas fitas constituintes para que o replissomo tenha acesso e inicie a síntese de novas fitas de DNA. Esta é a única etapa fortemente regulada pela maquinaria do ciclo celular.

- **Alongamento**: os nucleotídeos, na forma de nucleotídeo trifosfato, são incorporados à medida que o replissomo avança ao longo da respectiva fita molde. A incorporação das bases segue a regra de complementaridade. Nesta etapa, há correção de possíveis erros de incorporação.

- **Terminação**: uma vez replicado todo o genoma da célula, o replissomo se solta da molécula de DNA quando atinge uma região terminadora específica. Como resultado, a célula contem 2X o material genético.

Iniciação

Uma propriedade crucial da dupla fita de DNA é a habilidade de se separar em duas fitas sem quebrar as ligações covalentes, possibilitando que ambas voltem a se emparelhar. Novamente, a especificidade de tal emparelhamento é determinada pela complementaridade de bases do DNA. Esta propriedade é crucial para iniciar a replicação, assim como para terminar a mesma, garantindo, por um lado, que o replissomo tenha acesso a cada uma das fitas e, por outro, que, finalizada a replicação, o DNA volte a adotar a estrutura de dupla fita original. Talvez, a decisão mais importante que toda célula tem que tomar

é quando e como iniciar a replicação. Considerando que a replicação do genoma é um processo custoso do ponto de vista energético, este processo deve ser controlado de alguma maneira.

A iniciação da replicação do DNA bacteriano fisicamente se localiza no oriC. Este sítio é reconhecido pela enzima DnaA (Figura 11). Especificamente, a enzima reconhece a região de 250 pb, que contém 5 repetições de 9pb e uma região rica em AT. Ao menos 10 enzimas diferentes participam da iniciação da replicação (Quadro 1).

Quadro1. Componentes requeridos para iniciar a replicação do DNA em *E.coli.* Os dados extraídos o livro princípios da Bioquímica, LEHNINGER, A. L.; NELSON, D. L.; COX, s/d.

PROTEÍNA	FUNÇÃO
Proteína DnaA	Reconhece o oriC e abre a dupla fita de DNA.
Helicase (DnaB)	Desenovela o DNA.
Helicase (DnaC)	Auxilia a DnaB a se ligar no oriC.
Iniciase (Primase)	Síntese de *Primers ou iniciadores.*
SSB	Proteína que liga DNA de fita simples.
DNA girasse (topoisomerase II)	Libera a torção da dupla fita provocada pelo desenrolamento do DNA.
Outras: Metilases, HU, FIS IHF	Auxiliam a iniciação.

Todas estas enzimas têm como função principal o reconhecimento do sítio de replicação e a abertura do DNA nesse local para, assim, permitir a formação do complexo de pré-iniciação, processo esse que requer a presença de ATP (adenosina trifosfato) (Figura 9).

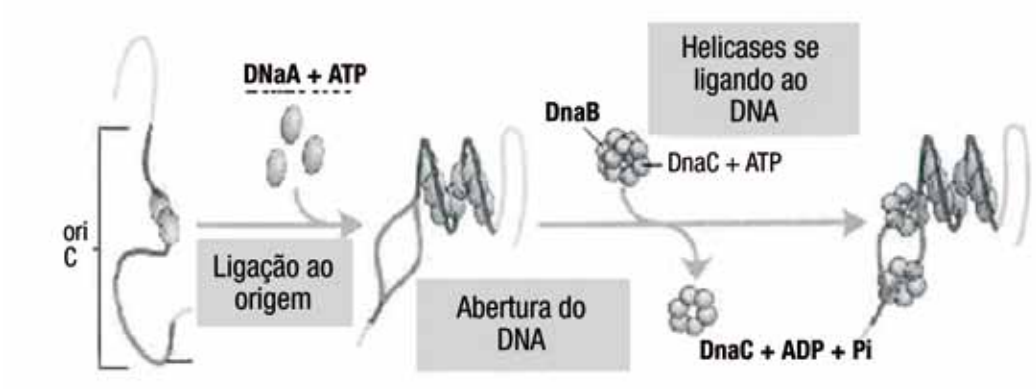

Figura 11: Modelo da iniciação da replicação na origem de replicação – oriC bacteriana – mostrando as primeira enzimas que catalisam esta etapa.

Modificado de Lehninger, 1995.

Uma vez separadas as fitas do DNA, proteínas estabilizadoras chamadas de enzimas **SSBP** (*single strand binding protein*) ou **proteínas de ligação à fita simples ligam-se a cada uma das fitas**. A sua ação estabilizadora evita que as fitas de DNA voltem a se associar novamente durante o processo de replicação.

Com o desenrolamento das fitas acontecendo em um ponto determinado, as regiões adjacentes sofrem um superenrolamento, o que dificulta a continuação do processo de replicação. As **topoisomerases**

resolvem esse problema fazendo cortes em uma das fitas de DNA para liberar a tensão causada. A **DNA polimerase (III)** é a encarregada de sintetizar a nova fita. Para que a DNA polimerase inicie a síntese é necessária a presença de uma sequência curta da RNA (10-12 bases) chamada de *primer* ou iniciador, a qual é sintetizada pela **primase**. Este iniciador será eliminado e substituído, posteriormente, por DNA, uma vez que a replicação tenha começado.

Alongamento

A fase de alongamento da replicação é a fase de síntese de DNA propriamente dita e inclui a síntese da cadeia contínua e da cadeia atrasada. O DNA é aberto e é formada a forquilha de replicação (Figura 12) onde são reunidas todas as enzimas do replissomo. A síntese da cadeia contínua e da atrasada é marcadamente diferente. A síntese da cadeia contínua é iniciada logo após a ação da primase (DnaG), a qual sintetiza uma curta sequência de nucleotídeos de RNA (10-60 bases) na origem de replicação. Uma vez ligada no DNA, a DNA polimerase III, com ajuda de proteínas auxiliares e energia em forma de ATP, inicia o processo de alongamento e cópia de toda a cadeia de DNA. Já a fita atrasada, como mencionado anteriormente, é sintetizada primeiramente como fragmentos de Okazaki. Primeiro, os *primers* de RNA são sintetizados pela primase (como na fita contínua). Logo, a DNA polimerase III liga-se a estes *primers* para, assim, adicionar os nucleotídeos necessários. Esta reação de síntese é uma das mais complexas já vistas na Biologia Molecular. Envolve a coordenação do processo de síntese em ambas as fitas do DNA por uma polimerase que possue uma taxa de incorporação de nucleotídeos de aproximadamente 1000 nt/s (nucleotídeos por segundo).

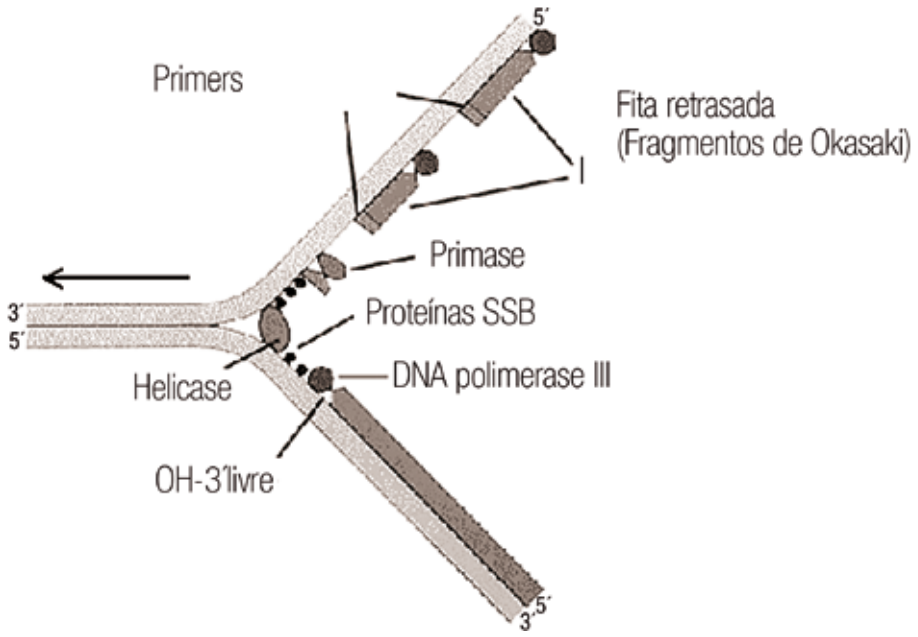

Figura 12. Esquema simplificado da fase de alongamento da replicação. A forquilha de replicação representada inclui as principais enzimas envolvidas neste passo. Todas elas fazem parte do replissomo.

Extraído e modificado de Bruce Alberts, et al , Biologia Molecular da Célula/ http://www.ncbi.nlm.nih.gov/books/NBK21054/.

Terminação da replicação

No final do processo de cópia, as forquilhas de replicação do DNA encontram-se numa região de terminação, a qual é formada por uma sequência de 20 pb chamada de Ter (Figura 14). As sequências Ter estão arranjadas no DNA bacteriano de maneira que a forquilha de replicação passe por elas, mas não siga em frente.

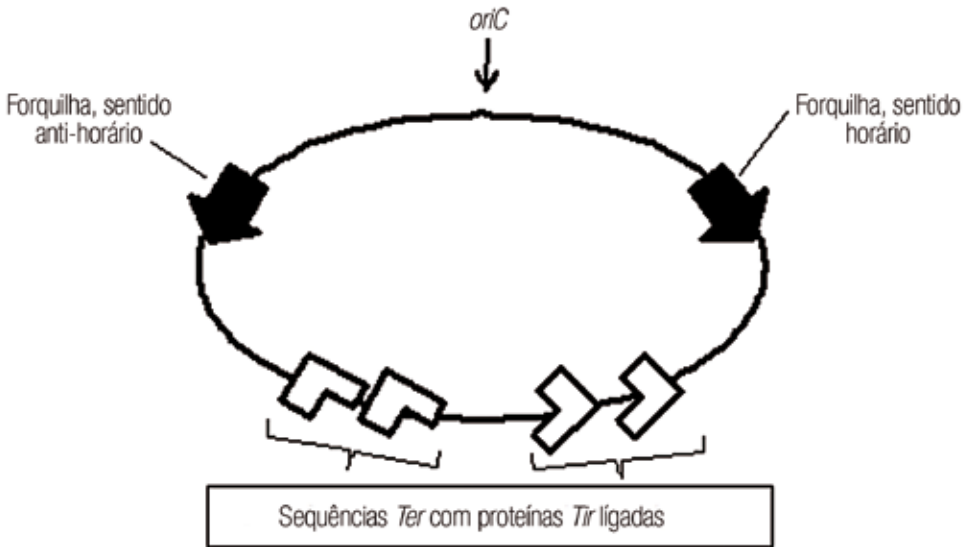

Figura 13. DNA circular bacteriano mostrando a direção das duas forquilhas de replicação (setas pretas) e as sequências TER onde termina a síntese de DNA.

Modificado a partir de Lenhinger, 2007.

A função principal destas sequências é servir como sítio de ancoragem das proteínas Tus (*terminus utilization substance*) de tal maneira que, ao se ligarem ao DNA, elas impessam que a forquilha de replicação avance. Como resultado da replicação bacteriana originam-se dois cromossomos circulares. Por um mecanismo ainda não bem esclarecido, estes dois cromossomos são separados um do outro pela ação da enzima topoisomerase IV que ocorre juntamente com o processo de divisão celular bacteriano.

DNA polimerases bacterianas

Nas bactérias existem diferentes tipos de DNA polimerases que exercem funções diferentes no processo de replicação do DNA.

A DNA **polimerase I**, primeira a ser descoberta, é uma proteína constituída de uma subunidade catalítica capaz de sintetizar DNA com baixa processividade. Antigamente, pensava-se que esta enzima era a responsável pela replicação do DNA bacteriano. Hoje em dia, sabe-se que esta enzima cumpre funções de correção e manutenção do DNA durante todo o ciclo celular bacteriano. Destaca-se pela sua atividade exonucleásica em ambos os sentidos. – 5′ → 3′ (atividade raramente encontrada em outras DNA polimerases) e 3′ → 5′ – da fita dupla de DNA. A velocidade de incorporação de nucleotídeos é de aproximadamente 15 por segundo e com processividade de 200.

Definição/significado de Processividade: característica de enzimas que participam do processo de replicação. Relaciona-se à quantidade de nucleotídeos que elas conseguem adicionar antes de se desligarem do molde.

A **DNA polimerase II** assim como as IV e V recentemente descobertas, tem atividade exonucleásica 3´ → 5´, maior processividade e maior velocidade de incorporação de nucleotídeo, quando comparada à DNA polimerase I. Também, está envolvida na reparação do dano causado ao DNA durante o ciclo celular bacteriano quando este é, por exemplo, danificado por radiação ultravioleta (UV).

A **DNA polimerase III** é a principal enzima responsável pela replicação do DNA bacteriano (Figura 13). A estrutura desta enzima é mais complexa do que das outras mencionadas. Suas atividades de polimerização e revisão residem nas subunidades α e ε, respectivamente. A subunidade θ associa-se a estas duas para formar o núcleo básico da holoenzima. A processividade da DNA polimerase aumenta significativamente quando a subunidade chamada de braçadeira β está presente.

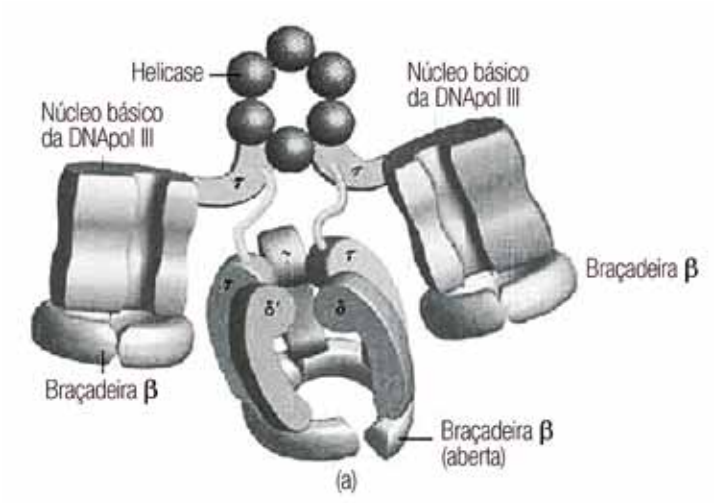

Figura 14. Arquitetura da DNA polimerase III (holoenzima). Ela é composta por 2 núcleos básicos, cada um contendo 3 subunidades (α, ε e θ) ligadas entre si e também ligadas à helicase através de um complexo posicionamento da braçadeira composta por várias subunidades (τ_2, δ, δ´e γ). Existem outras duas subunidades importantes que se ligam a esta enzima e que não estão mostradas na Figura. Em cada núcleo também existem duas braçadeiras ligadas.

Modificado a partir de Lenhinger, 2007.

A fidelidade da replicação do DNA

A replicação do DNA deve ocorrer com o menor erro possível. A informação nele contida deve ser transferida de geração a geração de maneira correta com a maior fidelidade factível. No entanto, como acontece em todo sistema vivo, os erros durante a replicação podem acontecer. Pode ocorrer de um ou mais nucleotídeos serem incorporados em posições incorretas, ou seja, sem serem complementares à base presente na fita molde. Isto é observado quando, por exemplo, ocorre a incorporação de uma Adenina (A) pareando-se com uma Citosina (C) (Figura 15).

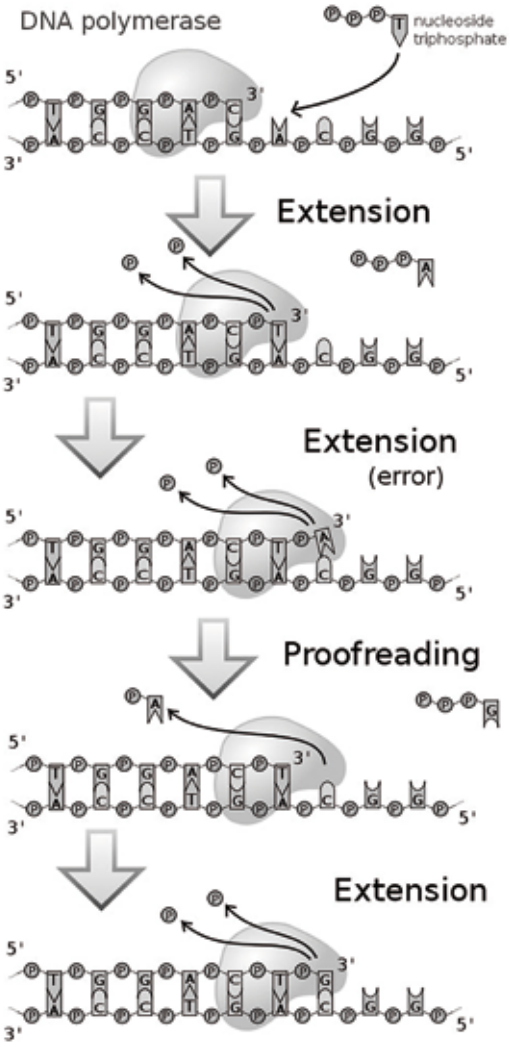

Figura 15. Prova de leitura feita por uma DNA polimerase. Quando a enzima reconhece a base erroneamente incorporada fita de DNA que está sendo estendida (neste exemplo A—C em vez de G—C), ela para a síntese momentaneamente e retrocede no sentido contrário removendo a base erroneamente incorporada para, em seguida, adicionar a base correta segundo a complementaridade de bases.

Extraído e modificado de Wikipedia ou de Allison, Lizabeth A. Fundamental Molecular Biology. Blackwell Publishing. 2007. p.112.

Esses tipos de erros (ou mutações pontuais) podem ser corrigidos imediatamente ao mesmo tempo em que a replicação está acontecendo. De fato, há algumas DNA polimerases encarregadas de corrigir estes tipos de erros. Quando um par de bases errôneas é reconhecido durante a replicação, a DNA polimerase reverte a direção na qual se movimenta e corrige o erro mediante a remoção da base e troca por uma base correta no local.

Assim, a atividade corretiva da proteína é chamada de **prova de leitura com atividade exonucleotídica 3′ → 5′**, já que acontece no sentido contrário ao da síntese de DNA. Estas enzimas têm uma taxa de erro calculada em 1 erro (incorporação de nucleotídeo errôneo) a cada 109 nucleotídeos adicionados.

Considerando um DNA com 9 milhões de pares de bases e só uma DNA polimerase, quanto tempo demoraria a célula para replicar seu genoma completo?

A mutação, do ponto de vista bioquímico, é sinônimo de erro e, dependendo de onde aconteça, pode ser fatal para um indivíduo. Por outro lado, do ponto de vista da evolução das espécies é considerada a principal fonte de geração de variabilidade genética nas espécies.

A replicação em eucariotos

Tudo o que vimos até agora e o que virá toma como modelo o sistema bacteriano. Quando tentamos comparar com um sistema modelo eucarioto, tomando como exemplo a levedura (unicelular e menos complexa que um mamífero, por exemplo), o panorama torna-se mais complexo e difícil de estudar. Primeiro, devido ao fato de que o DNA eucarioto é mais extenso que o bacteriano, a organização do mesmo no espaço físico celular é distinto e, assim, poderíamos citar várias razões para explicar este fato. No entanto, os processos evolutivos, ao menos em termos moleculares, nos permitem generalizar sobre este ponto e dizer que muitas das enzimas envolvidas no processo de replicação do DNA estão conservadas em todos os organismos vivos.

O que determina o grau de complexidade são os mecanismos que regulam o processo em si, tanto espacial quanto temporalmente. Nos eucariontes existem DNA polimerases homólogas às das bactérias e com as mesmas funções. Neste sentido, o processo de replicação em si, com a presença de origens e forquilhas de replicação, e com síntese contínua e descontínua, dentre outras características, também foi demonstrado em eucariontes.

Como todo processo biológico, a replicação do DNA precisa ser regulada. Nos eucariontes a replicação é regulada dentro do contexto do ciclo celular. Assim que a célula cresce e se divide, ela passa por diferentes etapas do ciclo celular. A replicação, especificamente, acontece durante a fase S (Síntese). Tanto o início como o término desta fase é controlada e envolve uma série de enzimas (quinases, fosfatases) chamadas, em geral, de ciclinas. Antes de iniciar a replicação do DNA, a célula deve passar pelo primeiro ponto de controle: o G1/S ou ponto de restrição. Se a célula não estiver preparada para replicar seu DNA, ela não atravessará este ponto. Por exemplo, isto acontece quando o DNA sofre algum dano e precisar ser reparado ou frente a condições fisiológicas extremas. Este tipo de controle evita que erros durante a replicação ocorram e que o DNA seja copiado de forma incompleta. Mas, uma vez que a célula transpõe este ponto de controle, inevitavelmente iniciará a replicação do DNA. As células que não progridem do ponto G1/S ficam no ponto denominado G0 Isto ocorre com os neurônios que, raramente, dividem-se e, portanto, não replicam seu DNA. Nos procariontes o controle da replicação é diferente, já que não possuem um ciclo celular bem definido. O rápido crescimento bacteriano exige que o DNA seja quase constantemente replicado. Nas *E.coli* sabe-se que a replicação é controlada por hemimetilação e sequestro do oriC. Também, pode ser regulado através dos níveis de ATP/ADP da célula que, por sua vez, influenciam as enzimas (dependentes destes nucleotídeos) envolvidas na replicação, como a DnaA.

Esta é apenas a imagem instantânea de um processo altamente dinâmico e complexo que acontece a cada ciclo de divisão celular, quando o nosso material hereditário precisa ser transferido com fidelidade para a geração seguinte.

Informações sobre a replicação

Disponíveis em: <http://www.wiley.com/college/pratt/0471393878/student/animations/dna_replication/index.>

<http://www.johnkyrk.com/DNAreplication.html>

<http://www.youtube.com/watch?v=teV62zrm2P0>

<http://www.youtube.com/watch?v=gL3aigv7w4A&feature=related>

<http://media.pearsoncmg.com/bc/bc_campbell_biology_6/cipl/ins/16/16-13-LeadingStndNarrAnim_S.mov>

<http://media.pearsoncmg.com/bc/bc_campbell_biology_6/cipl/ins/16/16-13-LaggingStrandAnim_B.mov>

CAPÍTULO 3

A transcrição do DNA

O processo de transcrição é o evento que envolve a leitura do código genético contido no DNA e a sua cópia para RNA. Até agora, vimos como se copia a informação de DNA para DNA. Agora, passaremos a ver como esta informação é TRANSCRITA para outro formato semelhante – de DNA para RNA (Figura 16).

A evolução determinou que o DNA armazenasse e transferisse o código da vida ao longo dos tempos. Já a molécula de RNA, mesmo tendo quase as mesmas características químicas, atua transportando esta informação na célula. Para entender qual o significado deste processo, podemos realizar uma analogia com a escrita de um resumo bibliográfico. A informação contida em um livro texto é lida e transcrita para logo ser interpretada.

A expressão da informação contida nos genes geralmente envolve a produção de moléculas de RNA transcritas a partir do DNA molde. Apesar de serem moléculas semelhantes, o RNA contém características distintas do DNA, o que lhe permite realizar funções variadas dentro da célula. Portanto, o RNA seria mais complexo que o DNA em termos funcionais e estruturais, chegando a ser considerado como o material genético que primeiramente surgiu na terra.

A transcrição do DNA gera três tipos de RNAs: o mensageiro (mRNA), o de transferência (tRNA) e o ribossomal (rRNA). O conjunto destes RNAs transcritos em um determinado momento do ciclo celular é denominado transcriptoma. Em outros capítulos, veremos quais as abordagens utilizadas pela Biologia Molecular para caracterizar o transcriptoma de uma célula qualquer e a importância que isto tem para entender diferentes alterações ocorridas nela, seja durante o desenvolvimento e diferenciação celular ou na transformação em uma célula tumoral.

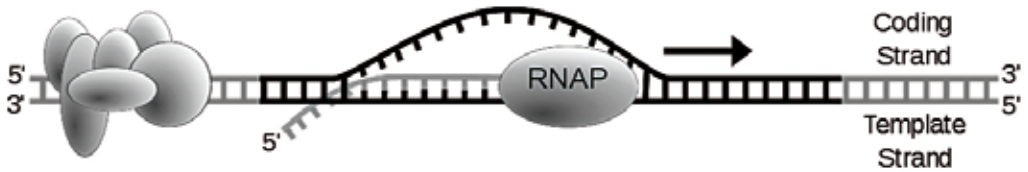

Figura 16. Transcrição do DNA. Uma única fita de RNA é gerada a partir do DNA (gene) sempre seguindo o sentido 5´ → 3´.

(Extraído e modificado de wikipedia http://en.wikipedia.org/wiki/File:Simple_transcription_elongation1.svg 01/09/11).

A transcrição se assemelha, em vários aspectos, com a replicação tanto em procariontes como em eucariontes. Os motivos são os seguintes:

» o RNA é transcrito (sintetizado) a partir de um molde, o DNA;

- a sequência do RNA transcrito é determinada pela complementaridade de bases – A=U(T) e C=G;
- a direção de síntese é, também, no sentido 5′ → 3′;
- a transcrição pode ser dividida em iniciação, alongamento e terminação.

Mas existem diferencias também:

- o RNA, após a sua síntese, não se mantém ligado ao DNA;
- o RNA é constituído de uma única fita de ácido nucléico em vez de uma dupla fita;
- a transcrição envolve a cópia de alguns trechos (genes) do DNA e não do genoma completo;
- a iniciação da transcrição não precisa de sequências iniciadoras (*primers)*, pois o RNA é sintetizado diretamente.

A RNA polimerase

A RNA polimerase é a enzima que catalisa o processo de transcrição. Utiliza o DNA como molde da reação de síntese, sendo, portanto, chamada de RNA polimerase dependente de DNA (Figura 17). Esta enzima, para ser ativa, precisa, além do DNA, da presença dos 4 nucleotídeos – ATP, GTP, UTP e CTP, como precursores do RNA. Também, precisa de íons de Magnésio e Zinco (Mg +2 e Zn +2). Tanto a química como o mecanismo da síntese de RNA se assemelham à replicação. A RNA polimerase alonga o transcrito de RNA pela adição de ribonucleotídeos à extremidade 3′, crescendo no sentido 5′ → 3′. A RNA polimerase requer de DNA para ser ativa e a sua atividade é maior quando se liga à fita dupla. Apesar disso, apenas uma das fitas de DNA é utilizada como molde da reação. Esta fita é lida no sentido 3′ → 5′, assim como acontece com a replicação. E, como nesta, a adição de cada nucleotídeo segue a regra de complementaridade de bases. A única diferença está na base uracila "U", que pareia com A.

Lembre-se que o nucleotídeo uracila está presente somente no RNA. No DNA, a uracila é substituída pela timina "T". Tanto a uracila quanto a timina se pareiam com a adenina "A".

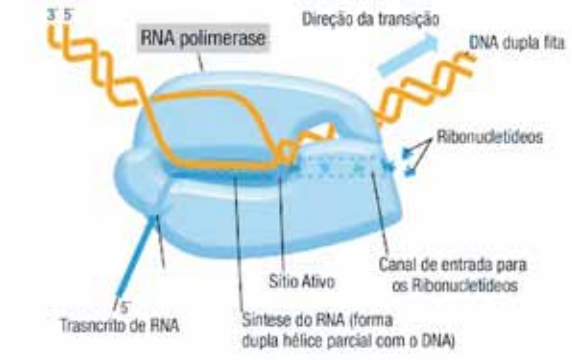

Figura 17. RNA polimerase sintetizando um transcrito de RNA a partir do molde de DNA. A síntese do transcrito é sempre na direção 5′ → 3′ (ou lendo o DNA no sentido 3′ → 5′), adicionando ribonucleotídeos (no sítio ativo) conforme avança sobre a dupla fita de DNA.

Extraído e modificado a partir de Lenhinger, 2007.

Unidade de transcrição

Podemos definir a unidade de transcrição como todos os elementos que determinam a expressão de um gene. Para iniciar a transcrição de um determinado gene, a RNA polimerase tem que localizar uma região chamada de PROMOTORA, a qual é constituída por uma sequência de nucleotídeos específicos e conservados que sinalizam e direcionam a transcrição do gene adjacente. No caso das bactérias, esta unidade de transcrição é conhecida como OPERON, o qual possibilita a produção de várias proteínas envolvidas em uma mesma via metabólica.

A unidade de transcrição mínima para um gene é determinada pela presença da própria sequência do gene e por mais duas regiões importantes que sinalizam o início e o fim da unidade de transcrição são elas: a região PROMOTORA e a região TERMINADORA da transcrição (Figura 18)

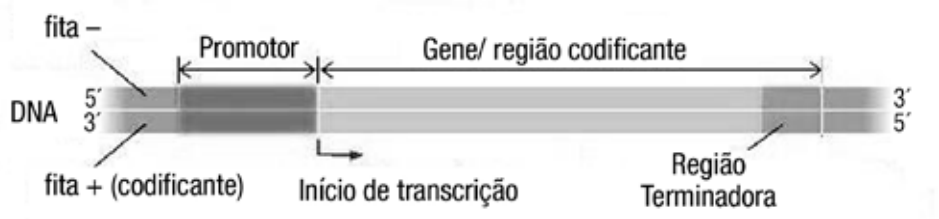

Figura 18. Organização da unidade de transcrição básica de um gene bacteriano qualquer

Extraído e modificado a partir de Lenhinger, 2007.

O início da transcrição

A iniciação é um processo complexo que envolve vários fatores reguladores, além da enzima RNA polimerase, e, como mencionado anteriormente, envolve também o promotor ou região promotora. A iniciação da transcrição é um ponto importante de regulação da expressão gênica, já que é ponto onde se determina qual gene, em qual momento e em que quantidade será expresso pela célula.

Nas bactérias, um promotor típico tem em média 70 pb e se localiza adjacente à sequência gênica que será transcrita. Por convenção, as posições referentes aos sítios de ligação da enzima são numeradas com sinais (+) e (–) dependendo da distância que se encontre um determinado nucleotídeo do início da transcrição gênica. O ponto inicial do gene é a posição 0 (Figura 19).

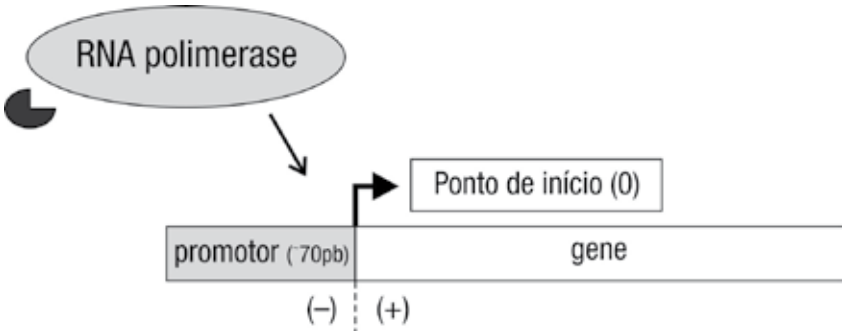

Figura 19. Promotor bacteriano típico formado por aproximadamente 70 pb localizado antes (-) do ponto de início da transcrição gênica (0). O promotor é o sítio de reconhecimento e ligação da RNA polimerase além de fatores que regulam a transcrição.

Extraído e modificado a partir de Lenhinger, 2007.

A análise comparativa de vários promotores bacterianos permitiu determinar a existência de duas sequências consenso em todos eles. Uma localizada a -10pb (5´-TATAAT-3´) bases do ponto de início da transcrição e outra a -35 pb (5´-TTGAC-3´) (Figura 20). As regiões -35 e -10 determinam a afinidade da RNA polimerase pelo promotor. De fato, mutações provocadas nas bases do consenso provocam mudanças drásticas no processo de transcrição, podendo causar desde um aumento na taxa de transcrição gênica até a inibição do processo, dependendo da região mutada. Fora desta região existem outras sequências importantes para regular a transcrição gênica. Em certos genes altamente expressos em bactérias existem sequências reguladoras em posições chamadas *UPSTREAM* (acima da região promotora), a -40 ou -60 pb do início.

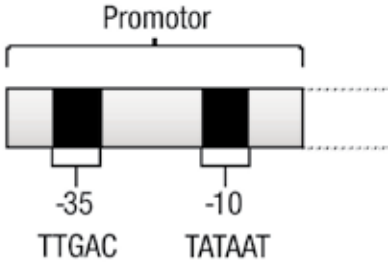

Figura 20: Promotor típico bacteriano mostrando as duas regiões consenso determinadas através da análise das sequências de vários promotores bacterianos.

Extraído e modificado a partir de Lenhinger, 2007.

O início da transcrição envolve múltiplos passos e a ação de várias proteínas (RNA polimerase e fatores de transcrição). Em geral, a transcrição começa com a ligação da RNA polimerase no promotor do(s) gene(s) e a posterior abertura do DNA. Esta abertura da dupla hélice é conhecida como complexo aberto da transcrição e acontece perto da região -10. Antes de começar a transcrição, há uma alternância de abertura e fechamento do complexo (RNA polimerase-DNA) e síntese de um pequeno fragmento de RNA (8-10pb). Neste ponto, a RNA polimerase liga-se com inespecificidade ao promotor. Outro fator protéico (componente da RNA polimerase), o fator σ (Sigma), deve estar, momentaneamente, presente junto à enzima para iniciar a transcrição. É ele que determina a especificidade da RNA polimerase no promotor. Apesar dessa atuação, o fator O tem que estar ausente para que esta avance na transcrição (Figura 21).

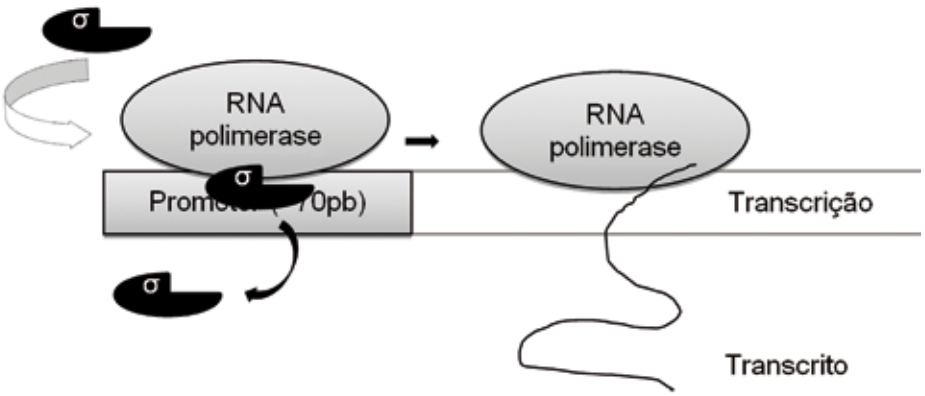

Figura 21. Reconhecimento do promotor pela RNA polimerase. A entrada do fator sigma e sua posterior liberação são os passos necessários para iniciar a transcrição do DNA. À medida que a RNA polimerase avança, uma molécula de RNA (transcrito) é gerada.

Extraído e modificado a partir de Lenhinger, 2007.

Os genes que estão sob controle de fatores de transcrição, como o fator sigma, estão, geralmente, envolvidos na produção de proteínas/enzimas necessárias para a bactéria durante todo ciclo de vida, ou seja, são genes constitutivos. No entanto, existem outros genes menos expressos ou expressos em determinadas circunstâncias, como, por exemplo, em condições de estresse ambiental para as quais existem outros fatores de transcrição específicos. Estes fatores têm afinidades distintas pelos promotores e, por sua vez, os genes que os produzem são fortemente regulados pela célula. Os fatores que se associam à RNA polimerase são determinantes para o reconhecimento de distintos tipos de promotores bacterianos.

Terminação da transcrição

Uma vez iniciada a transcrição, a RNA polimerase literalmente anda ao longo do DNA conforme o transcrito de RNA é sintetizando. No entanto, como esta enzima sabe onde termina um gene?

Como vimos na Figura 18, o final de um determinado gene é determinado por uma sequência terminadora especifica. Quando a RNA polimerase atinge esta sequência, ocorrem mudanças conformacionais no transcrito que fazem a enzima parar de transcrever e se soltar do DNA. Esse tipo de terminação é denominado "terminação independente de Rho" para diferenciá-la do segundo tipo de terminação, a "terminação dependente de Rho". A terminação dependente de Rho necessita da presença da proteína Rho e não de uma sequência terminadora.

Regulação da transcrição

O processo de transcrição gênica é regulado em vários pontos, já que nem todos os genes precisam ser expressos continuamente durante o ciclo celular. Por exemplo, durante a divisão celular, genes específicos envolvidos neste processo celular são altamente expressos, enquanto outros genes envolvidos no metabolismo – ou com a resposta a estímulos químicos e ambientais – encontram-se reprimidos e não são expressos.

Como vimos, o promotor é o sitio onde a RNA polimerase se liga para iniciar a transcrição e é nesse mesmo sítio que proteínas reguladoras da transcrição atuam para ativar ou desativar os diferentes genes. Estas proteínas reguladoras são denominadas repressoras e ativadoras da transcrição e elas competem para se associarem ao promotor. Também, existem genes que são expressos em quase todo o ciclo celular. São genes constitutivos essenciais para a vida de todas as células. Nesse caso, a regulação da transcrição é menos controlada.

Como sabemos, as bactérias não possuem núcleo celular. Nelas, o cromossomo bacteriano está em contato com o citoplasma da célula. Portanto, o processo de transcrição acontece no mesmo espaço físico onde acontecem outros processos celulares, como é o caso da síntese de proteínas. Essa disposição espacial também afeta e regula o processo de transcrição gênica. A expressão ou síntese proteica, nesse caso, acontece concomitantemente ao processo de transcrição. À medida que o transcrito de um determinado gene (que codifica uma proteína e não tRNAs ou rRNAs) é sintetizado, os ribossomos já estão se ligando a ele e começando a produzir a cadeia polipeptídica (como veremos à frente).

Tanto o processo de replicação do DNA, quanto a transcrição do DNA em RNA ocorrem no núcleo celular (caso o organismo em questão seja eucarionte). Já o processo de tradução, que veremos adiante, ocorre no citoplasma.

Atualização do Dogma Central

Como vimos anteriormente, o fluxo de informação genética ocorre em um sentido só: DNA→RNA→PROTEÍNAS. A replicação e a transcrição utilizam como molde o DNA. Entretanto, existem algumas enzimas que utilizam o RNA como molde para produzir DNA e para replicar o próprio RNA. Por isso, o que postulado inicialmente pelo Dogma Central da Biologia tem sido modificado (Figura 22). Essa modificação se deve à existência de alguns vírus que (retrovírus que infectam células eucariontes) possuem a enzima específica chamada de **Retrotranscriptase Reversa**. Esta enzima utiliza o RNA e não o DNA como molde para a reação de síntese. Quando esses vírus infectam uma célula, liberam no citoplasma desta o seu genoma, uma molécula de RNA de fita única de ~10.000 nucleotídeos, para gerar mais cópias da partícula viral. Primeiramente, o genoma de RNA é transcrito para DNA pela retrotranscriptase reveresa, gerando um híbrido RNA/DNA. Depois, essa enzima degrada a fita de RNA e a substitui por DNA novamente. O DNA *duplex* frequentemente se integra ao genoma da célula hospedeira, ficando disponível para ser transcrito, posteriormente, pela maquinaria da célula. Quando o genoma viral é transcrito para RNA, novas partículas virais são geradas, permitindo que o vírus continue o ciclo de infecção e replicação nas células hospedeiras.

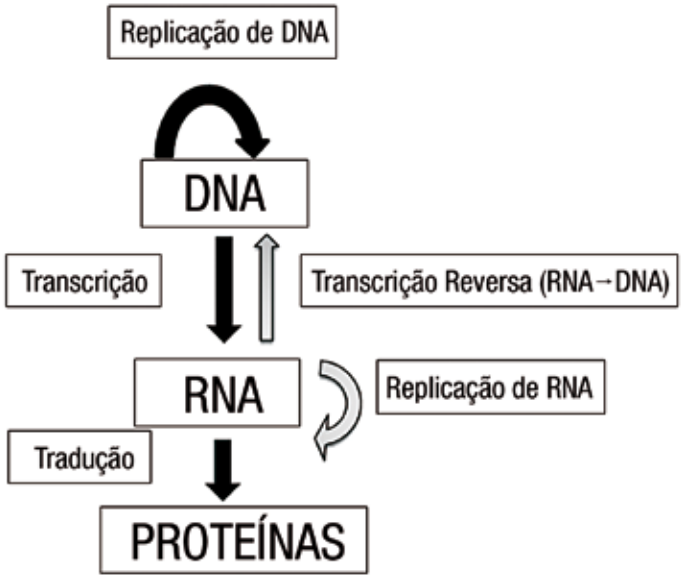

Figura 22. Fluxo da informação genética segundo o Dogma Central da Biologia Molecular atualizado, de acordo com as evidências científicas conhecidas recentemente. As setas em preto mostram o sentido do fluxo da informação mencionado por Crick correspondente aos processos de replicação do DNA, transcrição e tradução. Já as setas em cinza indicam o fluxo da informação de RNA para DNA (transcrição reversa) e replicação do RNA.

Extraído e modificado a partir de Lenhinger, 2007.

A replicação do RNA foi descoberta em alguns vírus, conhecidos como bacteriófagos, que infectam bactérias – os vírus F2, MS2, R17 e o QB – e, também, naqueles que infectam células de eucariontes. Em

todos eles, o RNA não só é replicado como também funciona como mRNA. De fato, para replicar o seu genoma de RNA, o vírus precisa expressar a enzima que cumpre esta função (já que a célula hospedeira carece de tal enzima). A enzima em questão é uma RNA polimerase dependente de RNA ou, simplesmente, RNA replicase. Todos os retrovírus (exceto os retrovírus como o HIV) precisam desta enzima. O curioso é que esta enzima utiliza, também, fatores de tradução do hospedeiro para funcionar corretamente. Em bacteriófagos, a enzima possui uma subunidade de 65 KDa (Mr= 65.000) codificada pelo próprio genoma viral. Existem outras 3 subunidades que se juntam a esta para ativar a RNA replicase. As subunidades são os fatores de elongamento Tu e Ts e a proteína ribossomal S1 da bactéria (*E.coli*). Tem sido demonstrado que esta enzima não reconhece o DNA como molde, nem o RNA da célula hospedeira, apenas reconhece o RNA viral.

A complexidade e organização que distinguem os organismos vivos dos sistemas inanimados são as principais manifestações do processo fundamental da vida. Poderíamos considerar um vírus como organismo vivo?

Transcrição: procariontes versus eucariontes

Apesar dos genomas de eucariontes serem mais complexos em relação à estruturação e distribuição dos genes do que os genomas de procariontes, podemos compará-los entre si no que tange a transcrição, como é demonstrado no Quadro 2.

Quadro 2. Comparação da transcrição em procariontes e eucariontes.

TRANSCRIÇÃO EM PROCARIONTES	TRANSCRIÇÃO EM EUCARIONTES
1. Catalizada por uma Enzima RNA polimerase.	1.Catalizada por 3 RNAs polimerases.
2. Promotor reconhecido diretamente pela RNA polimerase.	2. Precisa de vários fatores de transcrição para o reconhecimento do promotor.
3. Unidade de Transcrição simples.	3. Unidade de transcrição complexa.
4. Genes transcritos sem íntrons.	4. Genes transcritos majoritariamente com íntrons e éxons.
5. Transcrição e tradução acopladas.	5. Transcrição (núcleo) separada temporal e especialmente da tradução (citoplasma).

(Dados extraídos de Lenhinger, 2007)

Íntrons são regiões não-codificantes do DNA, isto é, não produzem proteínas ou tRNA e rRNA. Os Éxons, por sua vez, são regiões codificantes do DNA. Em um processo denominado *splicing,* regiões específicas do RNA mensageiro (os íntrons) são recortadas e eliminadas. Esses íntrons eliminados são segmentos não-codificantes, pois não levam nenhuma mensagem para produção de proteínas. Depois que eles são eliminados, os segmentos resultantes (os éxons) unem-se entre si formando a molécula de RNA mensageiro funcional com a mensagem madura (mensagem propriamente dita). Acredita-se que a principal vantagem da

ocorrência desse processo nos eucariontes é que seus transcritos primários podem ser processados de vários modos para a produção de diferentes RNAs mensageiros maduros. Dependendo do organismo ou do estágio de desenvolvimento em que ele se encontra, pode ocorrer a produção de diferentes proteínas a partir de um mesmo segmento de DNA.

Processamento do RNA

Depois de gerado o transcrito de RNA, tanto nas bactérias como nos eucariotos, este precisa ser modificado de alguma maneira para assegurar que a informação seja, posteriormente, interpretada corretamente. Alguns dos eventos moleculares mais interessantes no metabolismo do RNA ocorrem depois da sua síntese. O mais chamativo é que, em muitos dos casos, as enzimas que estão envolvidas nesses eventos são os próprios RNA e por isso, eles também são conhecidos como ribozimas. Dentre as ribozimas, temos os íntrons autoprocessantes (grupo I), as RNAses P e a ribozima cabeça de (Figura 23). As principais atividades catalíticas que as ribozimas desenvolvem são duas transesterificação e hidrólise da ligação fosfodiéster (clivagem) do próprio RNA. Novamente, o reconhecimento do substrato é feito graças à complementaridade de bases dos ácidos nucléicos.

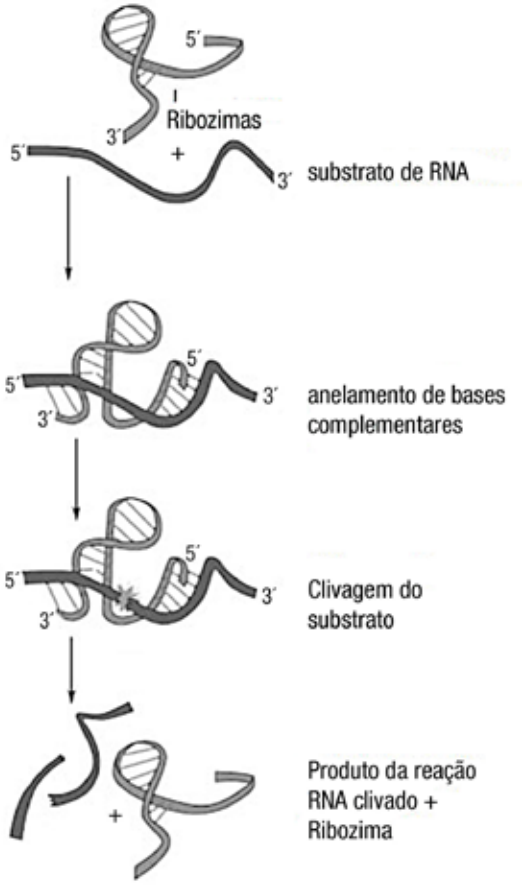

Figura 23. Ribozima: moléculas de RNA com atividade catalítica capaz de degradar a si mesma ou outras moléculas de RNA.

Modificado a partir de: http://caibco.ucv.ve/caibco/vitae/VitaeDieciseis/Articulos/Bacteriologia/ArchivosHTML/degradacionmarn.htm, retirado em 20/06/11.

Vídeos recomendados sobre a descoberta das Ribozimas.

Disponível em:

<http://www.youtube.com/watch?v=eGKg5i4FHHw&feature=related>

<http://www.youtube.com/watch?v=WAChisSiW3o&feature=related>

<http://www.youtube.com/watch?v=4Jdc7qW6hNM&feature=related>

<http://www.scripps.edu/news/press/010809.html>

Os mRNAs e os tRNAs são os transcritos que sofrem mais modificações pós-transcricionais. De fato, os mRNAs eucariotos são transcritos contendo sequências não codificantes (íntrons) que precisam ser removidas enzimaticamente para que o mRNA seja traduzido. Também, estes mRNAs passam por modificações na extremidade 5′, onde é adicionado um nucleotídeo modificado (5-metil guanosina) para formar um capacete 5′. Enquanto isso a região 3′ é clivada e modificada com a adição de 80 a 250 nucleotídeos de Adenina para formar, assim, uma cauda poliadenilada ou, simplesmente cauda poli-A.

Qualquer que seja o destino e função do RNA dentro da célula, inevitavelmente, esta molécula será posteriormente degradada. A taxa de renovação dos RNAs desempenha um papel importante na regulação da expressão gênica e, portanto, no balanço total de proteínas que estão, ou não, sendo expressas em um determinado momento do ciclo celular. Alguns RNAs têm um tempo de vida muito curto, com duração de minutos, de horas ou mais, dependendo das necessidades da célula.

CAPÍTULO 4

Tradução ou síntese proteica

As proteínas constituem mais da metade da massa seca total de uma célula e sua síntese tem uma importância fundamental para a manutenção e o crescimento celular. A síntese de proteínas ocorre nos ribossomos e envolve vários tipos de moléculas de RNA que atuam nas diversas etapas do processo.

Inicialmente, uma molécula de RNA mensageiro deve ser sintetizada a partir de uma das cadeias do DNA que codifica a proteína. No citoplasma, moléculas de cada um dos 20 aminoácidos que entram na composição das proteínas devem se unir a seus respectivos RNAs transportadores. As subunidades ribossômicas que promoverão a síntese devem se associar com proteínas auxiliadoras no processo. A síntese de proteínas tem início quando todos os componentes mencionados acima, ou seja, um RNA mensageiro, um dos RNAs transportadores e as subunidades de um ribossomo se reúnem para formar um ribossomo funcional. Cada ribossomo percorre, então, a molécula de RNAm traduzindo a sequência de códons em uma sequência de aminoácidos. Aqui, o formato de ácido nucléico tem que ser traduzido para um formato novo, o das proteínas, isto é, RNA→Aminoácidos. Ou seja, a informação passa a ser escrita em um formato de 4 letras (os nucleotídeos A/T ou U/C/G) para um de 20 (os aminoácidos essenciais).

Assim, a sequência de nucleotídeos do mRNA é lida seguindo a regra do código genético, ou seja, uma trinca de nucleotídeos codifica um aminoácido. Segundo o cálculo matemático de combinação, temos 43=64 possibilidades de trincas diferentes, ATG, ATC, ATT etc., mas existem apenas 20 aminoácidos. Isto indica que o código genético é degenerado, ou seja, duas ou mais trincas (códons) de nucleotídeos podem codificar o mesmo aminoácido (Figura 24)

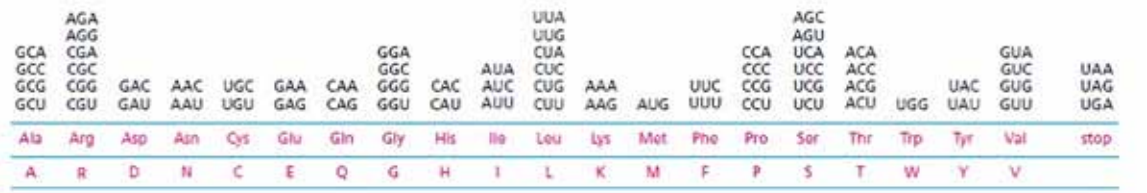

Figura 24. O código genético é degenerado. Um aminoácido é codificado por mais de um códon. Existem três códons que não codificam aminoácidos e servem como sinal de parada da tradução. O início da tradução é codificado, sempre, pelo códon da metionina (AUG).

Extraído de Lodish H, Berk A, Zipursky SL, et al.New York: W. H. Freeman; 2000.

Apesar de o código genético ser degenerado, isto é, trincas diferentes de nucleotídeos codificam o mesmo aminoácido, ele não é dúbio. Isso porque aminoácidos diferentes não são produzidos por um mesmo códon.

Fase de leitura do código genético

Como vimos anteriormente, a informação genética está contida nos genes. A produção de uma proteína depende da sequência contida nesse gene e no seu respectivo mRNA. Essa informação do código genético é chamada de fase de leitura ou sequência de códons e é lida durante a tradução. Em princípio, todo mRNA (e, portanto, o gene que o codifica) contém 3 fases de leitura potencialmente traduzíveis (Figura 25). Dada a sequência hipotética 5´-CUCAGCGUUACCAU- 3´ do mRNA, esta pode ser lida começando pelo códon CUC(1ª fase), UCA (2ª fase) ou CAG (3ª fase), podendo originar aminoácidos/proteínas diferentes.Para que a proteína resultante tenha a sequência correta e, consequentemente, desempenhe sua função, ela deve ser produzida na fase correta.

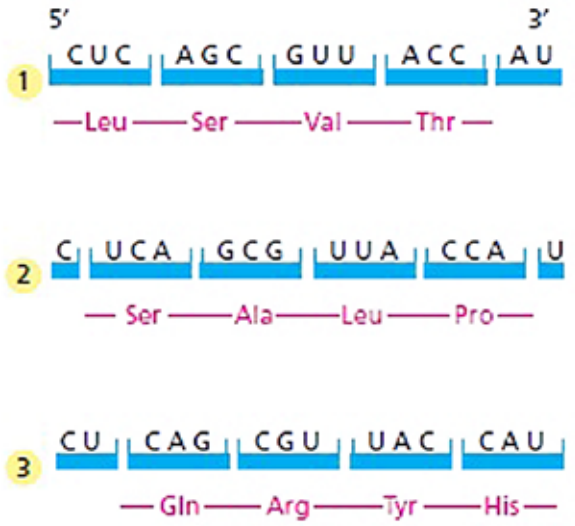

Figura 25. Três fases de leitura possíveis para um mRNA hipotético considerando o código genético de trincas.

Extraído de Lodish H, Berk A, Zipursky SL, et al.New York: W. H. Freeman; 2000 Molecular Cell Biology. 4th edição.

A função do RNA de transferência (tRNA)

Durante a síntese de proteínas, os códons do mRNA não são diretamente reconhecidos pelos aminoácidos que eles codificam. Precisam de um adaptador entre eles e a maquinaria de síntese protéica (os ribossomos). Esta função adaptadora é levada à cabo pelo RNA de transferência ou tRNA. Cada molécula de tRNA (com uma média de 80 nucleotídeos) carrega em uma extremidade um aminoácido e, na outra, uma sequência (trinca) complementar ao códon no mRNA. Esta sequência complementar ao códon do mRNA é chamada de anticódon. A primeira base do códon no mRNA (lido na direção 5´→3´) pareia com a terceira base do anticódon (Figura 26 A).

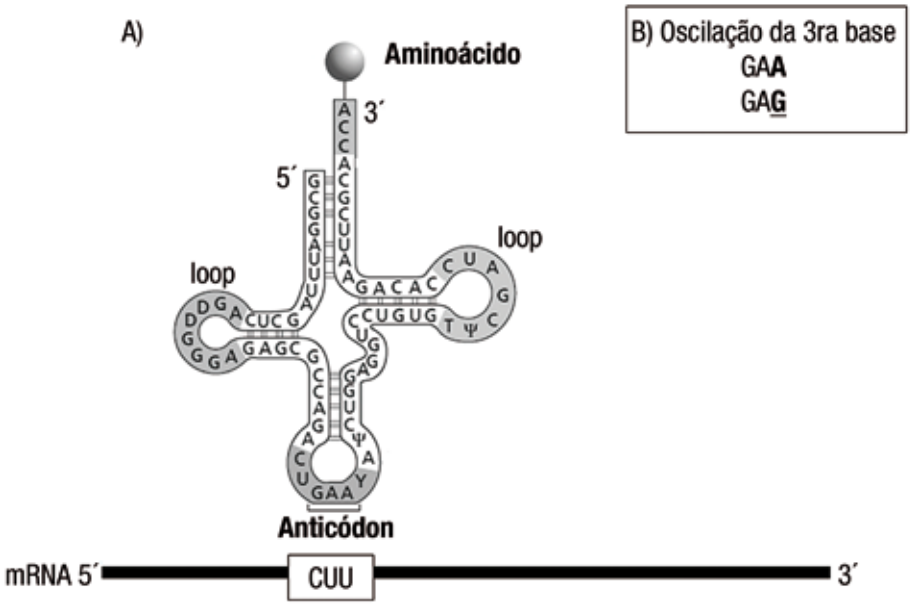

Figura 26. A) Molécula de tRNA carregada com um aminoácido e ligada ao mRNA através do pareamento de bases entre o códon e o anticódon. Note que o tRNA adota uma forma de folha de trevo (*loops*). B) Oscilação na terceira base do codón que permite que dois códons para o mesmo aminoácido (GAA e GAG) sejam reconhecidos e ligados pelo mesmo tRNA.

Modificado a partir de Alberts, 2006.

Como visto anteriormente, os aminoácidos são codificados por mais de um códon (exceto a metionina (M) e o triptofano (W)). Este fato não implica na existência do mesmo número de anticódons, isto é, de tRNAs diferentes para um mesmo aminoácido. O pareamento entre as bases complementares do códon e anticódon não ocorre, necessariamente, com as três bases. Se olharmos em detalhes as sequências dos códons de um aminoácido, poderemos observar que, na maioria, há apenas mudança na terceira base do códon (Figura 26 B). Isso permite que dois ou mais códons para um mesmo aminoácido sejam reconhecidos pelo mesmo tRNA ocorre o pareamento entre as bases complementares na primeira e na segunda posição do códon, sendo que a terceira base fica "flutuando". Desta forma, as duas primeiras bases são as determinantes da especificidade. Na célula existem, no mínimo, 32 tRNAs diferentes para os 61 códons do código genético, isto é, 31 para os aminoácidos e 1 para a iniciação (AUG=Metionina). No entanto, o número exato de tRNA dependerá da cada espécie. Nos humanos, por exemplo, há 500 genes codificando para tRNAs, mas apenas 48 deles são produzidos (transcritos).

Os ribossomos

Os ribossomos são as organelas citoplasmáticas responsáveis pela síntese proteica de toda célula. Eles são constituídas de 2/3 de moléculas de rRNA (RNA ribossômico) e 1/3 de proteínas (ribossomais) distribuídas em duas subunidades independentes. Cada subunidade é chamada de "maior" e "menor" (Figura 27), devido a diferenças na massa de cada uma (determinada pelo coeficiente de sedimentação S). Tanto as subunidades como seus constituintes podem ser separados, isolados e depois reconstituídos *in vitro* sem que o Ribossomo perca a sua atividade. Até não muito tempo atrás, pensava-se que eram as proteínas as principais responsáveis pela atividade catalítica desta organela. Mas, com a elucidação da sua

estrutura tridimensional (há uma década), somada a alguns dados bioquímicos, foi possível demonstrar que é o RNA, e não as proteínas, o principal responsável pela síntese proteica. De fato, as proteínas têm um papel secundário e não se encontram dentro do sítio catalítico do ribossomo, mas sim na superfície associadas ao rRNA.

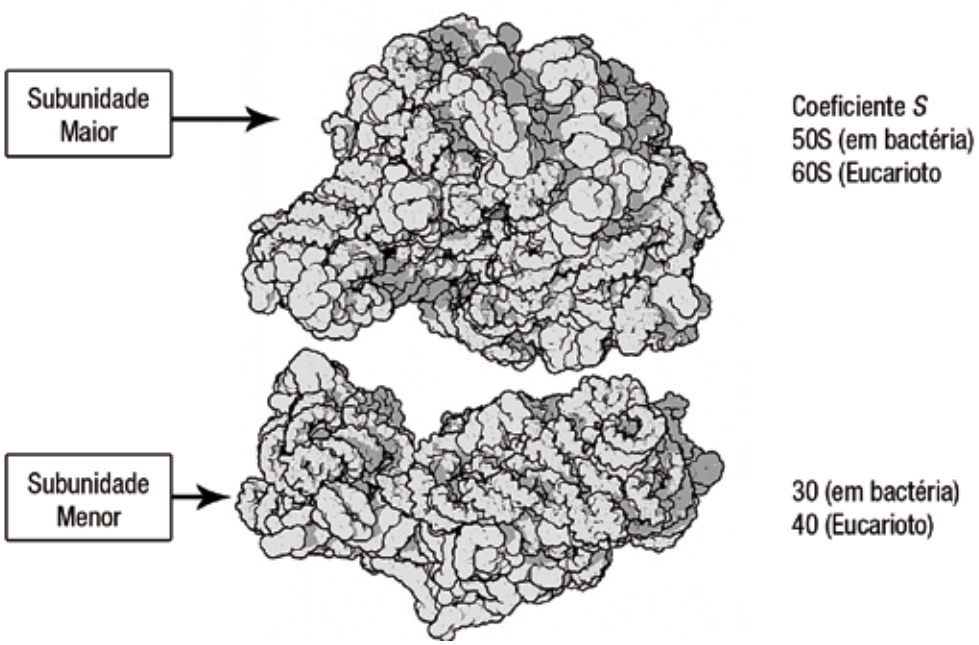

Figura 27. Subunidade maior e menor do ribossomo típico de bactérias.

Extraído do protein data bank , http://www.rcsb.org/pdb/home/home.docódigo 1ffk e 1fka, respectivamente, em10/07/2011.

A subunidade maior contém o sítio ativo do ribossomo. Nas bactérias, esta subunidade é constituída por dois rRNAs de massas diferentes, 5S e 23S. Já em eucariotos, esta subunidade possui 3 rRNAs com massas de 5S, 5,28S e 28S. A subunidade menor contém apenas um rRNA tanto nas bactérias como nos eucariotos. As massas são o 16S e 18S, respectivamente. Durante a síntese proteica, essas duas subunidade se juntam para formar o ribossomo ativo, criando uma fenda estreita por onde o mRNA se desliza à medida que ocorre a sua tradução.

Localização e tipos de ribossomos

Ribossomos são encontrados nas células sob duas formas: livres e associados ao retículo endoplasmático.

Ribossomos livres

Encontrados no citoplasma. Podem ocorrer como um único ribossomo ou em grupos conhecidos como polirribossomos ou polissomos. Ocorrem em maior número que os ribossomos associados ao retículo e em células que retém a maioria das proteínas produzidas.

Ribossomos associados ao retículo

São encontrados associados à membrana do retículo endoplasmático (RE) constituindo o RE rugoso. Estes ribossomos sintetizam proteínas de secreção, as quais são exportadas para outras organelas e para o espaço extracelular via RE. Portanto, ocorrem em maior número que os ribossomos livres e em células que expressam grande número de proteínas para secreção (por exemplo, nas células pancreáticas produtoras de enzimas digestivas) (Figura 28).

Encontramos ribossomos, também, dentro das mitocôndrias e nos cloroplastos de células eucarióticas. Eles são sempre menores que os ribossomos citoplasmáticos e são comparáveis aos ribossomos procarióticos em tamanho e sensibilidade a antibióticos. Entretanto, os valores de sedimentação "S" variam entre as espécies. As células aplicam considerável esforço para a produção destas organelas essenciais. Por exemplo, uma célula de *E. coli* contém, mais ou menos, 15000 ribossomos cada uma com um peso molecular de, aproximadamente, 3 x 10^6 Dalton constituindo 25% da massa total dessas células.

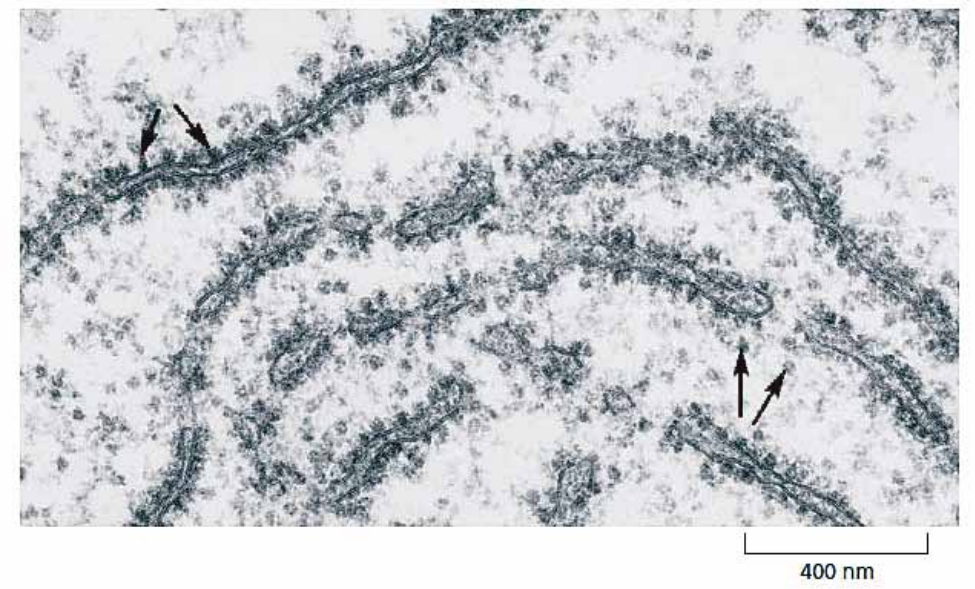

Figura 28. Micrografia eletrônica de transmissão (MET) mostrando o RE rugoso com as partículas ribossomais aderidas na membrana.

Extraído de Leningher, 2007.

Existem RNAs que têm atividade catalítica e, por isso, são chamados de ribozimas. Seria adequado denominar os ribossomos da mesma maneira?

O processo de síntese proteica

A reação fundamental da síntese de proteínas é a formação de uma ligação peptídica entre o grupo carboxila (C-COOH) de um aminoácido (aa) e o grupo amino (C-NH_2) livre de outro aminoácido (aa_{n+1}) (Figura 29). Consequentemente, uma cadeia polipeptídica (proteína) é sintetizada com uma

orientação definida, isto é, no sentido N-terminal até o C-terminal da cadeia polipeptídica. Como visto anteriormente, a informação que determina a sequência de aminoácidos está codificada no mRNA.

A síntese proteica pode ser dividida em 5 estágios:

- ativação dos aminoácidos;
- início da síntese proteica;
- alongamento da cadeia polipeptídica;
- terminação da síntese protéica;
- enovelamento tridimensional da cadeia polipeptídica.

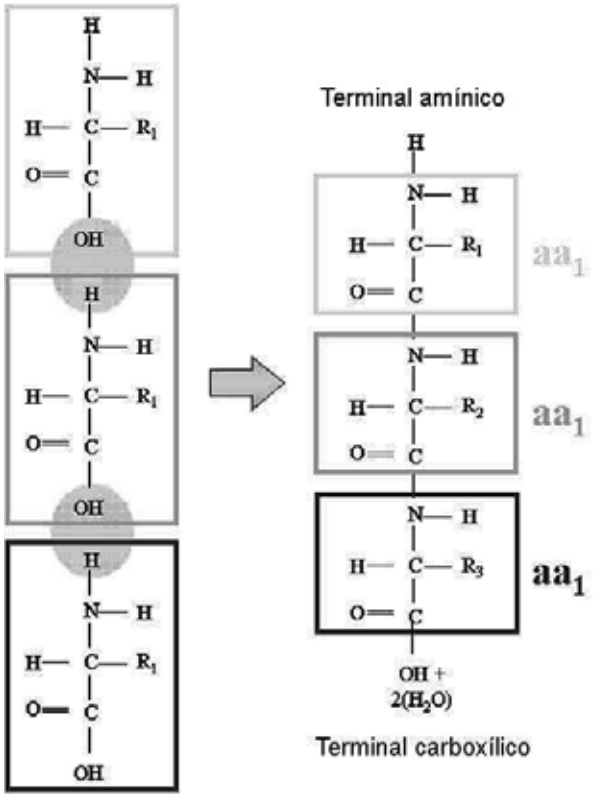

Figura 29: Esquema da formação da ligação peptídica entre dois aminoácidos, durante a síntese proteica.

Imagem extraída de http://www.biorede.pt/zoom_image.asp?ImageID=2159, em 24/06/11.

Estágio I: ativação dos aminoácidos

Ativar um aminoácido é deixá-lo num estado com energia necessária para realizar uma ligação peptídica para a polimerização de uma proteína. Em outras palavras, consiste em ligar cada um dos 20 aminoácidos livres existentes no citoplasma celular a seus respectivo tRNAs. Esta reação é realizada por enzimas específicas – a aminoacil tRNA sintase – e sem a participação do ribossomo (Figura 30).

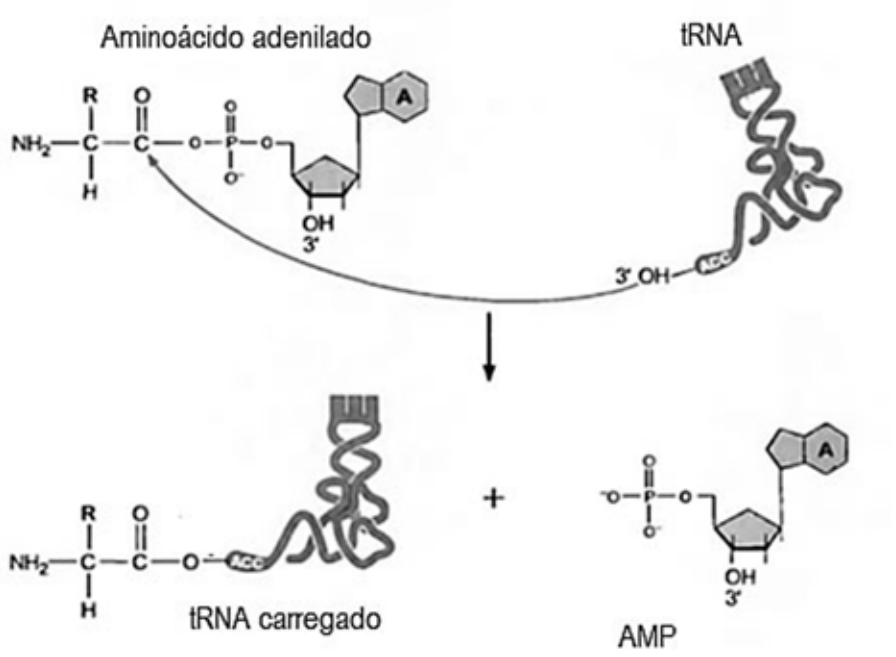

Figura 30. Ativação de aminoácidos. Os aminoácidos livres na célula são carregados nas respectivas moléculas de RNA transportador (tRNA) antes de serem utilizados na síntese de proteínas. Esta ativação envolve a ligação covalente do aminoácido com a extremidade 3´ livre do tRNA.

Retirado e modificado a partir de Alberts, 2006.

Estágio II: a iniciação da síntese protéica.

Nesta etapa, o mRNA codificante para uma cadeia polipeptídica se liga à subunidade (livre) menor do ribossomo através de uma sequência sinalizadora (em bactérias chamada de Shine-Delgarno) e, também, a um tRNA carregado com o primeiro aminoácido da cadeia polipeptídica. Em todos os mRNA bacterianos e eucariotos, o primeiro códon de iniciação codifica para uma metionina. Portanto, neste estágio, a molécula que estará presente será o ^{MET}-tRNA. Uma vez presente o ^{MET}-tRNA ligado ao códon respectivo de iniciação – AUG – do mRNA, a subunidade maior do Ribossomo liga-se ao complexo ativando, assim, a síntese protéica.

No ribossomo existem três sítios específicos de ligação: os sítios A, P e E. O sítio A é onde o tRNA carregado entra e se liga ao mRNA. No sítio P, ocorre a formação da ligação peptídica e é pelo sítio E que o tRNA sem o aminoácido sai do ribossomo. Estes passos requerem energia e proteínas auxiliares associadas ao complexo (Figura 31).

Estágio III: alongamento da cadeia polipeptídica.

Neste estágio, a cadeia polipeptídica é sintetizada pela ligação covalente de cada aminoácido que, sucessivamente, é transportado pelo respectivo ^{aa}-tRNA (o qual se liga ao códon respectivo do mRNA). À medida que uma ligação peptídica é formada, o ribossomo se desloca para que um ^{aa}-tRNA possa reconhecer o próximo códon no mRNA (Figura 31)

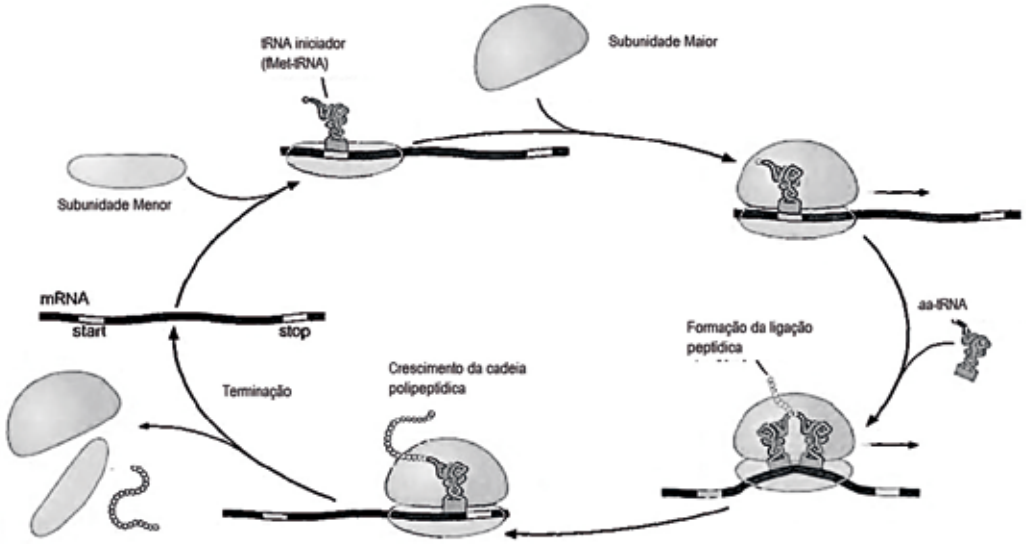

Figura 31: Esquema simplificado da síntese proteica mostrando as etapas de iniciação, alongamento da cadeia polipeptídica e de terminação da síntese proteica.

Retirado e modificado a partir de Alberts, 2006.

Estágio IV: terminação e reciclagem do ribossomo

A terminação da síntese proteica é determinada uma vez que o ribossomo reconhece um códon de terminação (*STOP*) no mRNA. Com o auxílio de fatores proteicos de terminação, a cadeia polipeptídica recém-sintetizada é liberada do ribossomo. O ribossomo se desmonta e suas subunidades ficam disponíveis para outro ciclo de síntese.

Estágio V: enovelamento protéico e modificações pós-traducionais

Nesta etapa, já na ausência do ribossomo, a cadeia polipeptídica adota a estrutura tridimensional característica da proteína. Este passo é fundamental para que a proteína seja funcional. Caso ocorra algum erro durante esse passo, a proteína incorretamente enovelada é direcionada para a degradação proteolítica. O processo de enovelamento pode acontecer tanto no citoplasma celular como no lúmen do RE e ocorre com a presença de proteínas auxiliares chamadas de chaperonas moleculares que catalisam o enovelamento proteico. No entanto, o enovelamento correto não garante completamente que a proteína seja funcional. São necessários algumas modificações, conhecidas como modificações pós-traducionais, para garantir sua funcionalidade. Uma modificação comum é a eliminação de alguns resíduos da cadeia polipeptídica (como exemplo, a metionina inicial, a sequência peptídica de direcionamento para diferentes compartimentos celulares e clivagem proteolítica de sequências internas). Além disso, há a adição de resíduos de açúcares (glicosilação) e alterações químicas que permitem a proteína se ligar à membrana plasmática (por exemplo, a miristilação), dentre outras modificações importantes.

Visualização do processo de síntese proteica.

Disponível em:

<http://www.youtube.com/watch?v=ZmkEPuYQE8k>

<http://www.youtube.com/watch?v=DcCnmPeutP4&NR=1&feature=fvwp>

UNIDADE II

INTRODUÇÃO À BIOINFORMÁTICA

UNIDADE II

INTRODUÇÃO À BIOINFORMÁTICA

Perante a carência de livros de textos básicos e em português sobre esta importante área da ciência, apresentamos nesta unidade noções gerais sobre Bioinformática e um breve tutorial sobre como acessar as fontes de informação sobre bancos de dados biológicos em geral. Para tal, tomamos como referência o sítio do Centro Nacional de Informação de Biotecnologia – NCBI (**National Center of Biotechnology *Information***). Esta unidade tem como propósito apresentar os conhecimentos atualizados da área e, ao mesmo tempo, auxiliar o acesso do(a) aluno(a) a este tipo de sítio, ajudando na compreensão e interpretação das informações encontradas, de forma a servir como guia e roteiro para consulta durante e após o curso.

CAPÍTULO 5

Definição de Bioinformática e sua relação com as áreas Genômicas e derivados "ômicos" (Transcriptômica, Proteômica, Metabolômica)

Em poucas palavras, podemos definir a Bioinformática como a área da ciência que envolve a Biologia, a Ciência da Computação e a Tecnologia da Informação. A Bioinformática tem como principal objetivo organizar, armazenar e analisar a informação biológica contida nas principais biomoléculas da vida, que são o DNA, o RNA e as Proteínas. Após a criação de bancos de dados, a Bioinformática se volta à interpretação destas informações, gerando conhecimentos que podem ajudar na identificação de novos fenômenos biológicos e de novos genes, por meio de análises comparativas de genomas de diversas espécies, dentre inúmeras outras abordagens científicas.

Análoga a uma biblioteca que armazena o conhecimento da humanidade em formato de livros, a Bioinformática teve que criar bancos de dados para armazenar a informação na forma de letras originadas do "decifrамento" do código genético. Os projetos de genomas de diversas espécies, dentre eles o do genoma humano, constituem a principal fonte desses bancos de dados.

Antes de a Bioinformática ser consolidada, havia nas Ciências Biológicas, duas maneiras de realizar os experimentos biológicos.

- *In vivo*: com organismos vivos e condições experimentais que sejam o mais próximas da realidade.

- *In vitro*: com ou sem organismos vivos sob condições ambientais restringidas.

Hoje em dia, graças ao desenvolvimento da Bioinformática, contamos com uma terceira forma de realizar pesquisa. Esta é chamada de pesquisa biológica *in silico* ou pesquisa biológica computacional, utilizando apenas um computador. A Bioinformática começou como uma tendência e, hoje em dia, é uma área da ciência completamente consolidada, de fato, grande parte das descobertas feitas nos últimos anos dentro da grande área da Biologia Molecular envolveu procedimentos bioinformáticos. Como exemplo, temos o projeto de sequenciamento do genoma humano.

A Bioinformática foi se desenvolvendo à medida que outras espécies de interesse tiveram seus genomas decodificados. Sem a Bioinformática não haveria genomas estudados, os quais geraram uma enorme quantidade de informação biológica. No começo, a Bioinformática funcionava apenas como uma ferramenta para organizar, armazenar e analisar as milhões de sequências de DNA obtidas de cada organismo, tecido ou célula. Hoje em dia, esta se tornou uma área de pesquisa dentro da grande área das Ciências Biológicas e continua se expandindo cada dia mais.

Era da Genômica e seus derivados "omicos"

Quando falamos em Bioinformática, rapidamente associamos com a Genômica. A Genômica, como o nome indica, é a área dedicada a decifrar o código genético de uma espécie, ou seja, caracterizar o seu genoma. Apesar da Bioinformática existir desde 1960, ela só se tornou protagonista décadas depois com a Genômica. Isso ocorreu, principalmente, após ter iniciado o projeto de sequenciamento do genoma humano, o qual visaria compreender a constituição gênica do ser humano, assim como relacioná-la com diversas doenças.

Desde 1980 até 2009, mais de 900 organismos tiveram seus genomas sequenciados (Figura 32 e Quadro 3).

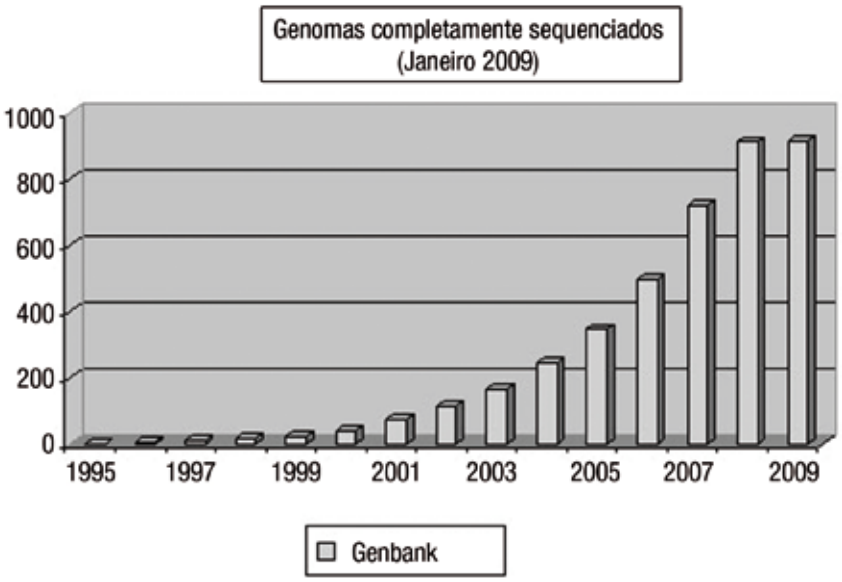

Figura 32: Gráfico mostrando o número de espécies que tiveram seus genomas completamente sequenciados e depositados nas ultimas duas décadas (1995-2009).

Extraído de http://www.genomesonline.org/images/gold_s1.gif, retirado em 20/07/2011.

De forma geral, um projeto genoma consiste, basicamente, em 3 grandes fases.

- **Sequenciamento**: análise do conteúdo de sequências nucleotídicas através da técnica de sequenciamento automático (método de Sanger) ou sequenciamento de nova geração.
- **Anotação:** depósito das sequências nucleotídicas, *in silico*, nos bancos de dados.
- **Processamento e análise:** organização das sequências de DNA e montagem do mapa genômico.

Uma vez anotadas e processadas, as informações derivadas dos genomas ficam disponíveis para serem acessadas para várias análises *in silico,* como, por exemplo, para:

- determinar similaridades e diferenças entre genomas completos de espécies relacionadas (desta forma, estabeleceu-se que o genoma humano é muito mais parecido com o genoma do camundongo);

- determinar o número de genes por genoma, classificar estes genes em famílias gênicas etc;
- identificar genes não descritos;
- estabelecer relações filogenéticas através de sequências nucleotídicas, dentre outras possíveis análises.

Quadro 3: Lista de alguns organismos que tiveram seus genomas sequênciados mostrando o tamanho (em pares de bases) e o número estimado de genes em cada um.

ORGANISMO	TAMANHO DE GENOMA (PARES DE BASES)	Nº ESTIMADO DE GENES
Homem (*Homo sapiens*)	3 bilhões	30.000
Rato (*M. musculus*)	2,6 bilhões	30.000
Mostarda (*A. thaliana*)	100 milhões	25.000
Roundworm (*C. elegans*)	97 milhões	19.000
Mosca das frutas (*D. melanogaster*)	137 milhões	13.000
Levedura (*S. cerevisiae*)	12,1 milhões	6.000
Bactéria (*E. coli*)	4,6 milhões	3.200
Vírus da AIDS (HIV)	9700	9

Tabela de Binneck, Eliseu. As ômicas: integrando a bioinformação. Biotec Ci & Des 32: 28-37. http://www.biotecnologia.com.br/revista/bio32/ômicas_32.pdf.

A Genômica decifra e organiza o código genético de uma ou mais espécies, mas tem limitações quanto aos tipos de informações que podem ser inferidas com seus dados. Não permite, por exemplo, analisar quais genes e em que intensidade são expressos numa célula ou organismo em um dado momento do ciclo de vida ou sob condições específicas. Podemos saber se uma célula ou organismo possui determinado gene se este tem variação em relação a genes similares, mas não poderemos saber se este gene é ativo, o quanto ele é expresso e quando é expresso na célula.

Como visto na primeira unidade, o RNA e as proteínas são os produtos de expressão gênica. Portanto, para analisarmos a expressão gênica de um ou mais genes de uma célula, temos que analisar estas moléculas especificamente. Assim, a Genômica originou três áreas importantes.

- Transcriptômica: área que estuda o conjunto de transcritos de RNA produzidos por uma célula ou organismo em um determinado momento do ciclo de vida.
- Proteômica: área que estuda de forma quantitativa e qualitativa o conjunto de proteínas presente na célula em determinado momento.
- Metabolômica: a área mais recente das três que estuda moléculas orgânicas importantes para a célula.

Essas três áreas são muito mais dinâmicas e apresentam mais informações sobre a expressão de genes e proteínas do que a Genômica. A maioria das células de um organismo contém o genoma idêntico independente do tipo celular, estágio celular ou condições ambientais. Células sanguíneas normais terão um conjunto de transcritos de RNA (e proteínas) típico para cada tipo celular, por exemplo, hemácias e leucócitos. No entanto, poderão variar em função das variações ambientais, como em resposta à presença de algum agente patogênico. Ao analisarmos os níveis de expressão global de genes e proteínas para cada tipo celular, podemos observar significativas diferenças se comparamos os níveis de expressão (transcriptoma e proteoma) entre células ou mesmo de um tipo celular sob condições ambientais diferentes. Seria possível observar, em princípio, quais são os genes que estão sendo expressos em cada situação.

Assim, a Biologia Molecular tem expandido suas formas de análises, passando dos estudos de um único gene, RNA, e da proteína, para uma visão mais global destas moléculas (tanto qualitativa quanto quantitativamente). Atualmente, espera-se que esta nova era "ômica" permita analisar doenças, como o câncer, de um ponto de vista mais real e holístico, o que ajudará a achar as causas, a nível molecular, de como uma célula se transforma em cancerosa e, assim, atacar o problema de maneira específica e o mais eficaz possível.

Em 26 de junho de 2000, Francis Collins e J. Craig Venter anunciaram, em cerimônia na Casa Branca com a presença do presidente Bill Clinton, que haviam completado o primeiro rascunho do genoma humano. O detalhe é que ambos obtiveram de maneira independente. Collins e Venter não eram colegas. Eram rivais que travaram uma corrida contra o tempo para ver quem publicaria os primeiro resultados. Por que haveria tanto interesse envolvido em saber quantos DNAs contém o ser humano?

A Genômica tem evoluído tanto que, hoje, o genoma de uma pessoa pode ser sequenciado em questão de dias. Seria possível, no futuro próximo, incluir a Genômica entre os estudos clínicos de rotina de uma pessoa?

Informações adicionais interessantes sobre o tópico.

Disponíveis em:

<http://www.youtube.com/watch?v=xPtOmviT7zs&feature=related>

<http://www.youtube.com/watch?v=taENQMitua8&feature=related>

<http://genome.wellcome.ac.uk/doc_WTX056439.html>

<http://genome.wellcome.ac.uk/doc_WTD020758.html>

<http://revistapesquisa.fapesp.br/?art=888&bd=1&pg=1&lg=>

<http://www.youtube.com/watch?v=Bixu2gQVvuA&feature=related>

<http://cienciahoje.uol.com.br/colunas/deriva-genetica/dez-anos-de-genoma-humano>

O Brasil no contexto das "ômicas"

A participação do Brasil nas "ômicas" da Biologia Molecular tem sido significativa. Através dos distintos projetos de pesquisas (envolvendo sequenciamento genômico) financiados, principalmente, pela agência de fomento Fapesp (Fundação de Amparo à Pesquisa do Estado de São Paulo), tem sido possível gerar muita informação sobre o patrimônio genética de espécies de interesse comercial ou mesmo do próprio homem. Nesse sentido, a Bioinformática tem sido altamente importante na hora de analisar os resultados obtidos. Para citar alguns exemplos, o Brasil tem participado do sequenciamento do genoma da *xylella fastidiosa*, da Cana-de-açúcar, do genoma humano e clínico do câncer, do genoma da *xanthomonas*, do eucalipto, do *schistossoma* mansoni, da bactéria *leifsonia xyli,* dentre outros.

O uso da Bioinformática é imprescindível para atender aos desafios científicos oriundos do gigantesco volume de dados produzidos pelos atuais projetos na área biológica. Uma vez que diversas espécies têm tido seus genomas decifrados, resta saber agora a função e o papel biológico de cada uma das sequências de proteínas que foram determinadas, mostrando o valor inestimável que a Bioinformática tem para as Ciências Biológicas. A sua importância se viu refletida logos após a publicação dos resultados do projeto genoma humano, onde milhares de sequências de DNA foram decifradas, permitindo a descoberta de mais de 30.000 genes e seus respectivos produtos proteicos.

Para trabalharmos com Bioinformática, basta apenas termos um computador conectado a internet e boas perguntas para começarmos a pesquisar nas diversas bases ou bancos de dados que contém todas as informações biológicas relevantes dos seres humanos e de muitas outras espécies. O mais importante é que todas essas informações são, na maioria dos casos, disponibilizadas de maneira gratuita para o público em geral, seja ele um simples pesquisador, um ganhador do prêmio Nobel, um estudante ou apenas uma pessoa curiosa.

Reportagens e leituras interessantes sobre o assunto:

Disponíveis em:

<http://revistapesquisa.fapesp.br/?art=4156&bd=1&pg=1&lg=>

<http://revistapesquisa.fapesp.br/?art=4142&bd=1&pg=1&lg=>

<http://revistapesquisa.fapesp.br/?art=4028&bd=1&pg=1&lg=>

<http://revistapesquisa.fapesp.br/?art=4004&bd=1&pg=1&lg=>

<http://revistapesquisa.fapesp.br/?art=3921&bd=1&pg=1&lg=>

<http://revistapesquisa.fapesp.br/?art=237&bd=2&pg=1&lg=>

<http://www.comciencia.br/reportagens/genoma/genoma1.htm>

<http://www.comciencia.br/reportagens/genoma/genoma3.htm>

Projetos genomas desenvolvidos no Brasil.

Projeto genoma *xylella fastidiosa*.

Disponível em: <http://www.lbi.ic.unicamp.br/xf/>

Projeto genoma da cana-de-açúcar.

Disponível em: <http://revistapesquisa.fapesp.br/?art=540&bd=1&pg=1&lg=>

CAPÍTULO 6

Os bancos de dados biológicos

O sequenciamento de qualquer molécula de DNA/RNA gera, inevitavelmente, informação e esta tem que ser armazenada de alguma maneira. Quando se trata de genomas inteiros, com bilhões de pares de bases, o volume de informação gerado é enorme. Por esse motivo, precisa-se criar um banco de sequências biológicas.

Por definição, um banco de dados biológico constitui um grande conjunto de sequências de DNA/RNA/proteínas armazenados e processados com *softwares* específicos. As informações mais relevantes geralmente organizam-se em forma de tabelas.

Classificação: bancos primários e secundários

Os bancos de dados primários são criados a partir de resultados de dados do sequenciamento gênico publicados com alguma interpretação adicional, mas sem uma análise exaustiva desses dados. Como exemplo deste tipo de banco, temos o *GenBank*/NCBI, o EMBL, dentre outros.

» *GenBank*/NCBI (disponível em: <http://www.ncbi.nlm.nih.gov/Entrez>).

» EMBL-EBI (disponível em: <http://www.ebi.ac.uk>).

Os bancos primários geralmente estão interconectados de maneira que a informação genética contida neles seja a mesma. Cada sequência de nucleotídeos é depositada e identificada com um **número de acesso** (*Accesion Number,* AN) e pode variar dependendo do tipo de sequência depositada. Para melhorar a identificação de sequências antigas e evitar sobreposição de sequências ou redundância, o AN vem acrescentado do número da sua versão. Dessa forma, pode-se ver o número de acesso, um ponto

e o número de atualizações feitas em uma determinada sequência Por exemplo, o número de acesso A21645.3 é a terceira atualização da sequência A21645 e as versões anteriores permanecem armazenas e acessíveis através dos números de submissão A21645.1 e A21645.2.

Um código similar ao *AN.version* é dado, também, para sequências de proteínas. Para criar um índice mais robusto para suas entradas, o NCBI, em 1992, criou um novo identificador, o *GenInfo Identifier* (gi), um número inteiro simples. Esse é um identificador único para cada sequência. Se houver duas sequências com 99% de identidade, mesmo assim, elas terão um **gi** diferente.

Como mencionado anteriormente, estes bancos de dados são públicos e qualquer pessoa pode ter acesso a eles de qualquer computador que esteja conectado a internet. Além disto, muitas outras informações podem ser obtidas através destas organizações públicas nos respectivos sítios na *web*. Por exemplo, através do NCBI (*National Center of Biotechnological Information*) podemos ter acesso a diversas bases de dados derivadas do *Genbank*. Ao acessar pela primeira vez um banco de dados de sequências, seja de DNA ou proteínas, nos deparamos com uma grande quantidade de nomes, siglas e números de identificação que podem nos confundir e atrapalhar durante a busca de informações relevantes. Para entendermos a terminologia da Bioinformática, primeiramente, começaremos por definir e compreender o que é um banco de dados primário e um banco de dados secundário e os números de acesso atribuídos a cada um.

O ENTREZ é a plataforma (NCBI) que nos permite ter acesso às sequências de DNA e proteínas de diversos organismos. Cada caso nos oferece informações sobre a localização do gene dentro do cromossomo (mapeamento gênico), assim como a homologia com genes de outras espécies. Além disso, os dados da literatura científica que sustentam essas informações estão disponíveis. Cada uma das sequências depositadas, assim como as informações relevantes, são constantemente atualizadas. Novas ferramentas sempre são incorporadas para melhor compreendermos as informações fornecidas.

Os **bancos de dados secundários**, por definição, são derivados dos bancos primários. Neles há uma compilação e interpretação dos dados de entrada (sequências) por um ou mais grupos de pesquisadores. Desta forma, ele é atualizado à medida que novas descobertas surgem. Portanto, os bancos de dados secundários são mais acurados que os primários.

» **COGs –** *Clusters of Orthologous Groups of proteins* (disponível em: <http://www.ncbi.nlm.nih.gov/COG/>).

» Uniprot/SWISS-PROT (disponível em: <http://www.uniprot.org/>)

» **PDB** – Protein Databank (disponível em: <http://www.pdb.org/pdb/home/home.do>)

O valor e/a importância dos bancos de dados secundários são resultado de algumas caracterísitcas relevantes:

» As informações são mais acuradas, evitando a redundância e a sobreposição de dados.

» Os dados relativos a uma espécie são, sempre que possível, correlacionados com os dados de outras espécies.

» Disposição dos dados em tabelas e/ou gráficos interativos.

» Disponibilidade de acesso a um vasto número de dados pré-processados.

» Ampla interconectividade com outros bancos de dados e *links* com informações relevantes e adicionais para cada sequência.

» Todos os dados contidos nesses bancos são suportados e identificados pelos autores responsáveis, assim como apurados por outros grupos.

No entanto, os bancos de dados secundários têm alguns inconvenientes que devem ser considerados:

» Esses bancos de dados demoram para ser atualizados, principalmente quando se trata de um volume grande de informação (genomas) proveniente dos bancos primários.

» Cada um possui uma interface de trabalho própria. O usuário deve estar familiarizado com cada uma e estar atento às mudanças e às atualizações.

» Muitas vezes, os *links* para outros sítios podem ser um problema devido ao excesso de informações.

» Alguns bancos de dados, como o SwissProt, mantém um único número identificador para cada sequência depositada. Já outros bancos de dados como o UniGene, o EST *clusters* ou o Ensembl podem mudar este identificador causando confusão.

CAPÍTULO 7

Como acessar um banco de dados primário: o NCBI

Como já dito, existem várias formas de acessar e obter as informações contidas em um banco de dados biológico. Para entendermos melhor a maneira de entrar nesses bancos, utilizaremos o sítio do NCBI (http://www.ncbi.nlm.nih.gov/Entrez), o qual é um dos principais bancos de informações biológicas do mundo (senão o melhor). Veremos, de uma forma geral, e por meio de exemplos práticos, quais informações que podemos encontrar num banco de dados biológicos e como utilizá-las.

Neste momento, é importante salientar alguns pontos importantes que o estudante tem que levar em consideração a partir de agora.

- Para realizar pesquisas relacionadas à Bioinformática, precisamos ter, obviamente, um computador conectado à internet. Não é necessário ter uma conexão muito rápida, mas, quanto mais rápida, melhor.

- Não é necessário instalar nenhum programa para acessar e utilizar os bancos de dados. A configuração básica de qualquer computador permite, em princípio, trabalhar com estes bancos.

As informações presentes nos bancos de dados que veremos à frente estão em inglês, por ser esta a língua universal da ciência.

A página principal do sítio da NCBI utilizada para a busca nos distintos bancos de dados chama-se ENTREZ (Figura 33). Nela, podemos acessar os vários bancos de dados disponibilizados pelo sítio. Os mesmos estão organizados em forma de lista e indicados com *links* e ícones respectivos. Também, existe uma caixa de busca para uma pesquisa global em todos os bancos de dados. A interface desta página é simples e de fácil compreensão.

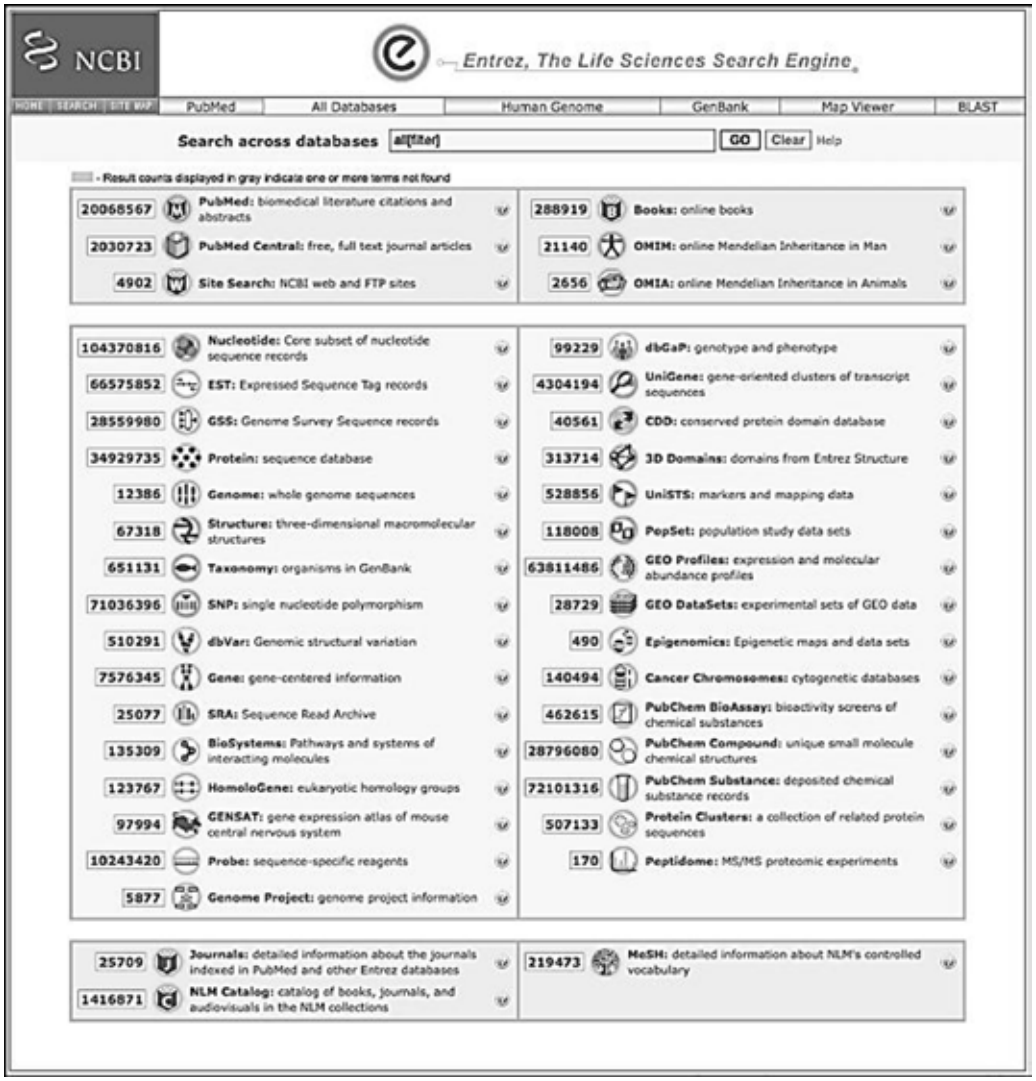

Figura 33. Página do *Entrez Globa*l (NCBI) mostrando os vários bancos de dados disponíveis. As pesquisas podem ser realizadas diretamente em cada um dos *links* ou utilizando um buscador geral *(search across databases).*

http://www.ncbi.nlm.nih.gov/Entrez

CAPÍTULO 8

Como buscar por literatura científica: PubMed

O PubMed é uma fonte de informação científica pública e gratuita que pertence e é mantido pelo *National Center for Biotechnology Information* – NCBI –, da biblioteca nacional americana de medicina que entegra o *National Institute of Health* (NIH). Ele contém mais de 20 milhões de citações de literatura sobre Biologia Médica (MEDLINE), jornais sobre as ciências da vida, assim como livros de texto *online*. Portanto, é o site ideal para pesquisas bibliográficas sobre qualquer assunto referente à Biologia Molecular e áreas afins (por exemplo, os tópicos abordados nesta apostila).

Também, é possível acessar, através do PubMed, resumos de artigos científicos *(papers)* de diversas áreas. Em alguns poucos casos, é possível acessar gratuitamente o artigo completo *online* ou em formato pdf. Na maioria dos casos, o acesso a esse tipo de informação é pago. Abaixo, apresentamos duas maneiras de buscar na base de dados do PubMed (http://www.nlm.nih.gov/bsd/disted/pubmed.html).

- Busca por termos gerais: Quando queremos saber sobre um assunto sem conhecer de antemão o que há publicado.
- Busca avançada: Quando precisamos encontrar um ou mais artigos específicos e contamos com informações mais detalhadas do assunto, como título do artigo, nome do autor, ano de publicação, jornal onde foi publicado etc.

Busca por termos gerais:

Primeiramente, acessamos o sítios do PubMed. Uma vez lá, clicamos o *link* PubMed e digitamos no buscar o termo *bioinformatics*, por exemplo. Por fim, clicamos *search* e obtemos os resultados da busca. Na tela do computador veremos algo semelhante à imagem apresentada na Figura 34.

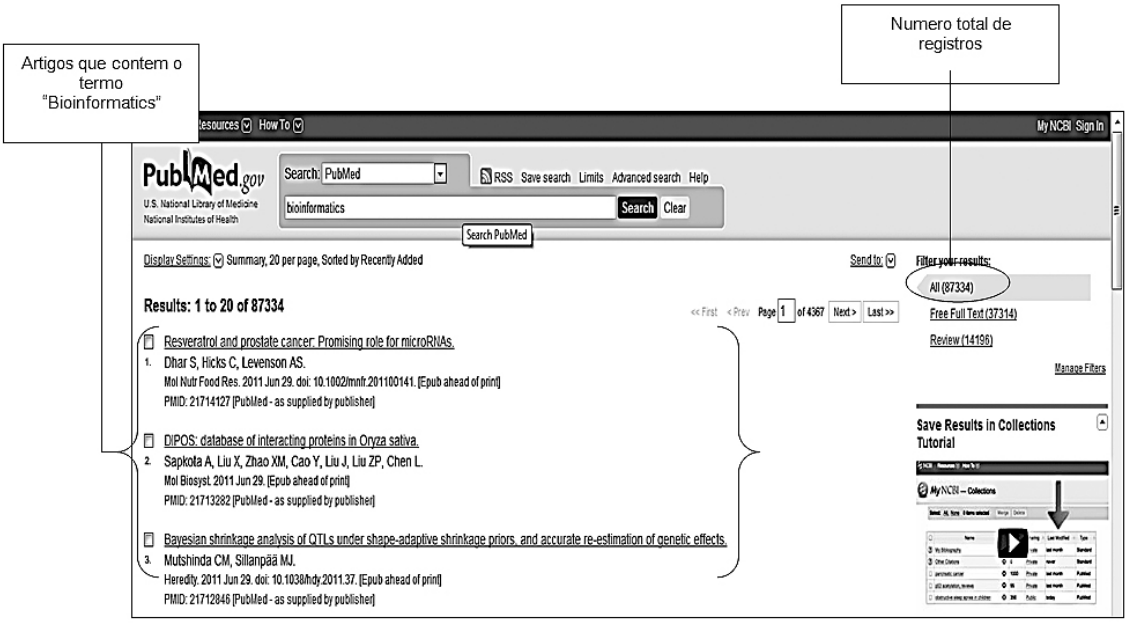

Figura 34. Imagem mostrando os resultados gerados pela pesquisa inicial do PubMed.

http://www.ncbi.nlm.nih.gov/pubmed

Clicando no primeiro resultado mostrado na Figura 34, visualizaremos mais detalhes do artigo escolhido (Figura 35).

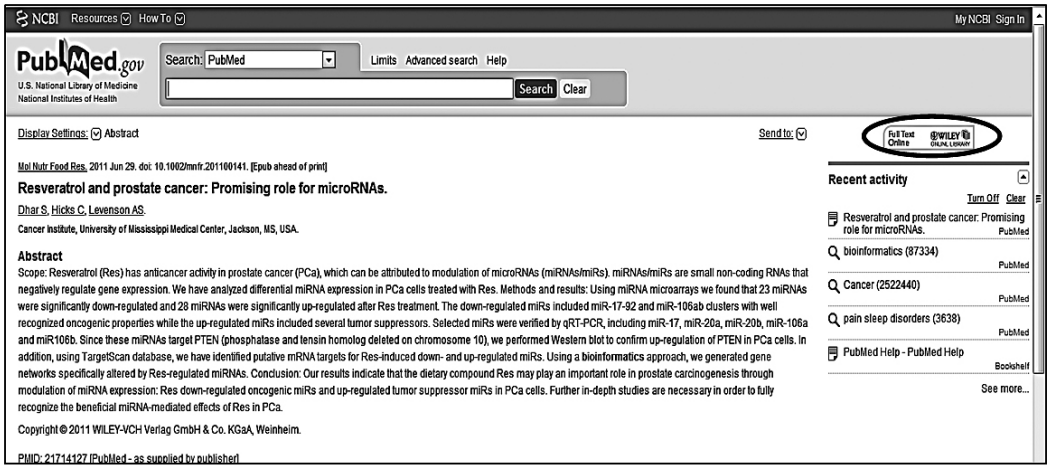

Figura 35. Imagem mostrando os detalhes do artigo selecionado. Vemos, entre outras informações, o título do artigo e os autores, além do resumo do trabalho *(abstract),* onde podemos ler as principais descobertas realizadas sem ter que acessar o texto completo.

http://www.ncbi.nlm.nih.gov/pubmed.

No resumo ou *abstract* podemos obter as principais informações oferecidas pelo artigo. Se este for de nosso interesse e quisermos ter acesso ao texto publicado na íntegra *(full text),* basta clicar no *link* que aparece, geralmente, na parte superior à direita (Figura 35) como um ícone da revista ou editora que o distribui. No nosso exemplo, trata-se de um artigo publicado pela editorial *Wiley*®.

Esse tipo de busca pode ser trabalhosa dependendo da maneira e dos termos usados para a pesquisa. Termos gerais como *CANCER*, DNA, RNA, *DOWN*, ou *VIRUS*, por exemplo, sempre terão como resposta uma quantidade enorme de informações (muitas vezes desnecessárias) que não se ajusta ao nosso foco de pesquisa. O mesmo acontece se usamos apenas os sobrenomes de autores, como Watson, Crick ou Da Silva. Por isso, recomenda-se utilizar uma combinação de termos específicos para, que, assim, possamos restringir o volume de informação adquirida em cada procura. Como exemplo, podemos acrescentar à palavra *Bioinformatics* os termos *genomics* e *Brazil* (*bioinformatics genomics brazil,* sem vírgulas) e assim, reduziremos o número de 87 mil para 212 registros. Por isso, é importante nos perguntarmos qual o objetivo da nossa pesquisa bibliográfica e determinamos (o que requer prática) as palavras-chave que nos permitam reduzir o tempo necessário para a obtenção das informações que esperamos encontrar.

Busca avançada

A outra forma de procurar referências bibliográficas é entrando diretamente no *link* PubMed *Single Citation Matcher* (http://www.ncbi.nlm.nih.gov/pubmed/citmatch), o qual permite realizar uma busca mais avançada, especificando o nome do autor, o jornal, ano de publicação etc. (Figura 36).

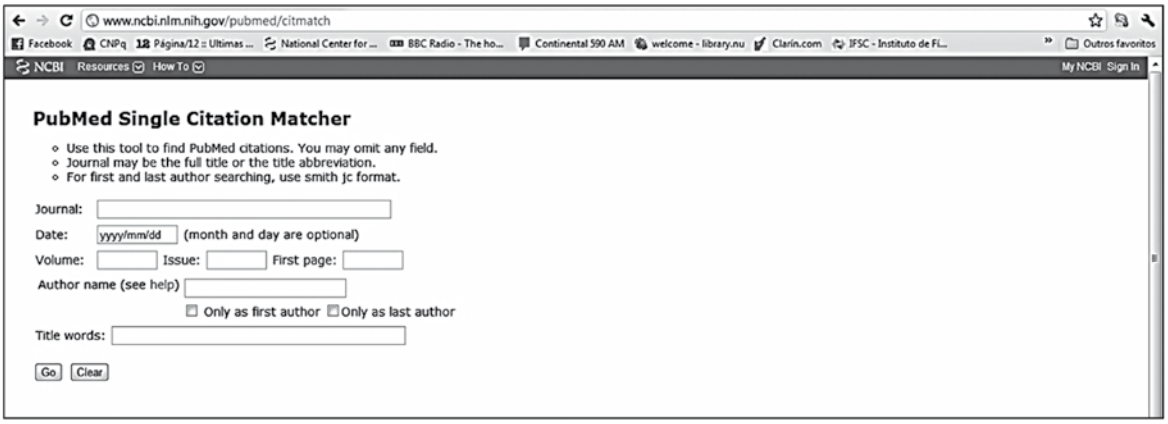

Figura 36. *Pubmed Single Citation Matcher*. Interface do PubMed que permite realizar uma busca por artigos científicos mais especifica. Essa ferramenta é útil quando conhecemos algumas das informações específicas de um artigo, como nome do autor, nome da revista, ano de publicação etc.

http://www.ncbi.nlm.nih.gov/pubmed.

Mais informações sobre o PubMed Disponível em: <http://www.nlm.nih.gov/bsd/disted/pubmed.html>

Buscando na base de dados de livros

Os conceitos teóricos tratados aqui e no resto do curso podem ser consultados no acervo de livros digitalizados e disponibilizados no NCBI através do link *bookshelf* (http://www.ncbi.nlm.nih.gov/books/). Esta base de dados contém 700 textos sobre ciências da vida e saúde humana. Todos estão escritos na língua inglesa.

CAPÍTULO 9

Buscando na base de dados de sequências nucleotídicas

Dentro do NCBI podemos acessar três tipos de bancos de dados de sequências nucleotídicas.

- *Nucleotide*: contém sequências nucleotídicas de várias fontes, incluindo *Genbank*.
- ESTs *(Expressed sequence tags)*: contém um conjunto de transcritos curtos provenientes de sequências de RNAm obtidos do *Genbank*. Neste banco podem ser analisados os níveis de expressão de um gene em particular, em um organismo ou tecido específico.
- GSS: contém um conjunto de sequências curtas de DNA genômico.

Os três estão interconectados através de *links*. Se o objetivo é encontrar a sequência de nucleotídeos que codifica uma proteína, tomemos, como exemplo, a enzima quinase. Devemos escrever o termo kinase na caixa de busca e, assim, teremos como resposta todas as sequências de nucleotídeos depositadas no banco que correspondem ao respectivo termo (Figura 37).

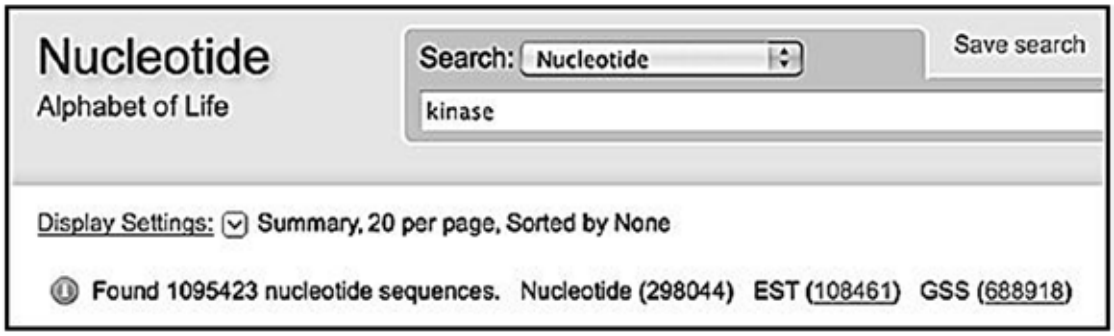

Figura 37: Banco de dados de nucleotídeos do NCBI. Resultado da procura pela palavra chave *Kinase*.

http://www.ncbi.nlm.nih.gov/nucleotide.

Como resultado, teremos o número de sequências nucleotídicas ESTs e GSS que respondem à palavra *Kinase*. Além do mais, podemos saber a distribuição destas sequências em distintos organismos no *link top organism* (Figura 38).

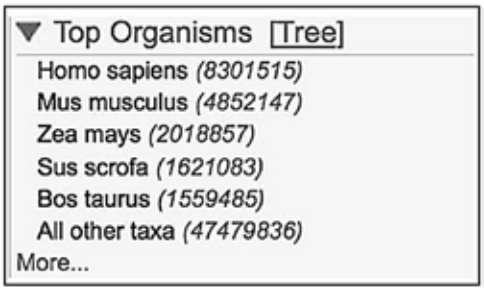

Figura 38. Lista de espécies que respondem ao termo KINASE

http://www.ncbi.nlm.nih.gov/nucleotide.

Selecionando uma das 8301515 sequências depositadas de humanos – por exemplo, *Homo sapiens* ROR2 (ROR2) *gene*, *partial cds* (Figura 39) –, teremos uma série de informações referentes à sequência. Assim, encontraremos uma lista de palavras-chave típicas, como *locus, definition, accession*, *version, keywords, source,* dentre outras que serão definidas à frente.

Figura 39. Descrição do gene ROR2.

http://www.ncbi.nlm.nih.gov/nucleotide.

» *locus*: atribui um nome ao *loco* do gene (SEG_AF), o comprimento da sequência em pares de bases (pb), a natureza da molécula (DNA ou RNA), a topologia (linear ou circular) e a data de depósito da sequência.

» *definition*: define de forma breve o gene.

» *accession*: mostra todos os números de acessos atribuídos à sequência.

» *keywords*: mostra as palavras-chave ou termos que caracterizam a sequência gênica/protéica depositada.

» *source*: divulga o nome do organismo ao qual pertence a sequência.

» *organism*: divulga a identificação taxonômica do organismo, isto é, o reino, filo, gênero e espécie.

» *reference*: apresenta todas as referências dos autores que descreveram a sequência (diferentes partes de uma mesma sequência pode ser atribuída a diferentes autores). Cada referencia tem um *link* que direciona para o PubMed.

Para obter apenas as informações básicas, como a sequência de interesse, clicamos no ícone FASTA (formato muito comum em bases de dados) e obtemos algo semelhante ao mostrado na Figura 40.

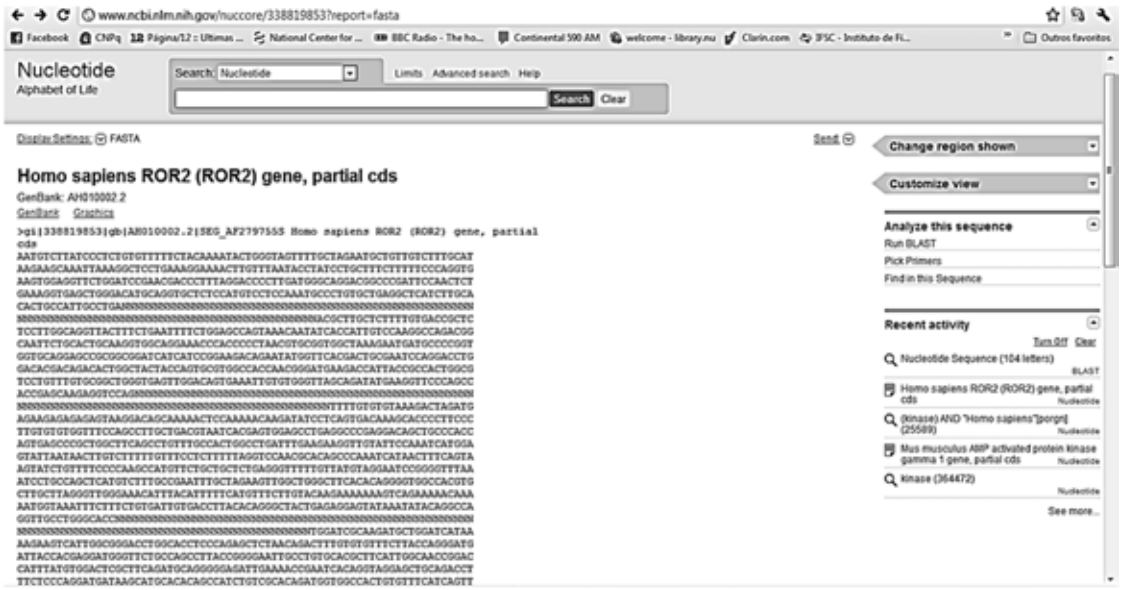

Figura 40. Obtenção da sequência de DNA de um gene qualquer.

http://www.ncbi.nlm.nih.gov/nucleotide.

CAPÍTULO 10

Buscando em bancos de dados gênicos

Recentemente, além do *Genbank*, o NCBI tem desenvolvido outras bases de dados (ou interfaces de bases de dados) específicas para genes bem identificados. São caracterizados, pela forma resumida, sem a presença de todas as informações que normalmente obtemos do *GenBank*. Em outras palavras, podemos obter a história completa de um gene em uma única busca. Além disso, oferecem informações quanto à localização do gene (*locus*) no genoma da espécie de origem e, também, permitem analisar as regiões do genoma que o circunda.

Esta ferramenta pode ser acessada no seguinte endereço eletrônico: http://www.ncbi.nlm.nih.gov/gene. Colocamos o nome do gene ou a sigla que o identifica – por exemplo, DUT – e teremos como resposta algo semelhante ao que visualizamos na Figura 41. Nesta interface, aparece uma breve resenha funcional do gene, isto é, a função do seu produto proteico. Podemos ver o mapeamento do mesmo no genoma e toda a bibliografia que suporta estas informações, assim como o mapeamento do mesmo no genoma da espécie e toda a bibliografia correspondente.

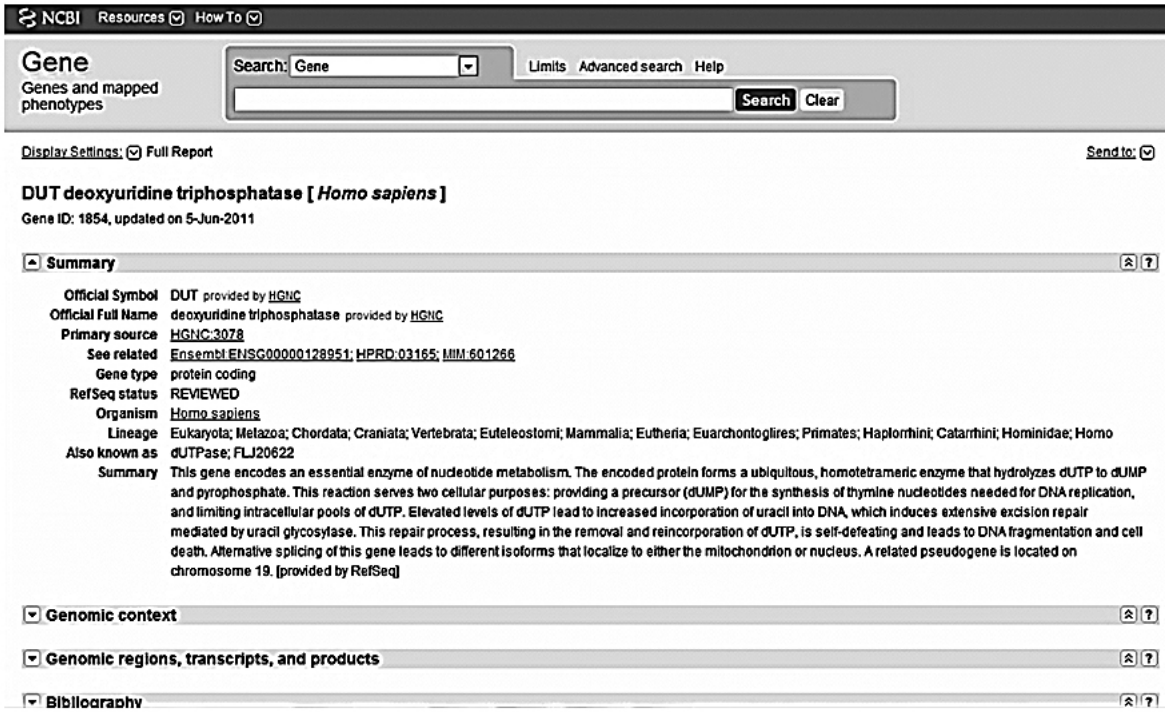

Figura 41. Descrição geral do gene DUT.

http://www.ncbi.nlm.nih.gov/gene.

Os bancos de dados genômicos

Os bancos de dados genômicos contêm sequências e mapas genéticos dos genomas de mais de 1000 espécies. Neles, encontramos tanto genomas completamente sequenciados, quanto aqueles que ainda estão em processo de sequenciamento. Todos os tipos de organismos estão representados nesse banco de dados. Encontramos desde bactérias e arqueobactérias, até plantas, fungos, animais, o ser humano e, também, vírus, bacteriófagos e plasmídeos.

Como mencionado anteriormente, o banco genômico é constituído, de forma organizada, por todas as informações referentes ao patrimônio genético de uma espécie. Dessa maneira, pode ser feita uma análise global dos genes de uma única espécie sem que apareçam dados provenientes de espécies relacionadas. No entanto, existem ferramentas para realizar análises comparativas entre genomas, como, por exemplo, entre o do ser humano e o do rato (esse tipo de análise é muito comum hoje em dia e recebe o nome de genômica comparativa).

Sítio contendo mais de 180 espécies que tiveram seus genomas sequenciados desde 1995.

Disponível em: <http://www.genomenewsnetwork.org/resources/sequenced_genomes/genome_guide_p1.shtml>

Buscando em bancos de dados secundários

A Uniprot (disponível em: http://www.ebi.ac.uk/uniprot/index.html) é a principal página consultada na *web* pela comunidade científica quando informações de sequências proteicas (acuradas e de alta qualidade) são o objeto da pesquisa. Existem várias formas de se acessar as informações contidas nela. Por exemplo, através da base de dados UniProt *Knowledgebase* (UniProtKB). Esta contém informações acuradas sobre as proteínas, incluindo a função, classificação e as referências que suportam as informações (Figura 42). Também, contém ferramentas Bioinformáticas para análise das sequências protéicas, como, por exemplo, a comparação de sequências proteicas por similaridade através da plataforma BLAST (*Basic Local Alignment Search Tool*, ClustalW (Mulitple sequence alignment). O aluno poderá notar que estas ferramentas também estão disponíveis na pagina do NCBI.

A busca nos bancos de dados secundários, em geral, costuma ser menos trabalhosa e mais compreensível por várias razões.

- as sequências proteicas são mais curtas (~350 aminoácidos) que as sequências nucleotídicas;
- possuem início e fim bem definidos, diferente do observado em várias sequências gênicas;
- as proteínas são constituídas por uma sequência única.

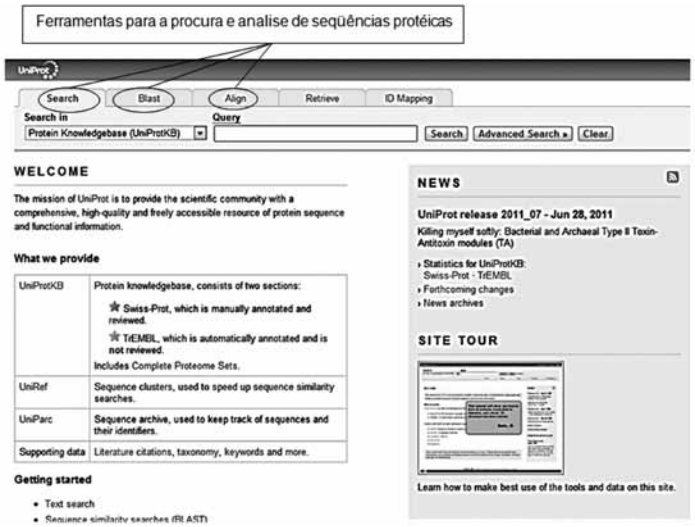

Figura 42. Ferramentas para a busca e análise de sequências proteicas. Para uma maior compreensão das ferramentas oferecidas pelo *site*, recomendamos realizar um *tour* virtual clicando no respectivo sítio *web* (*site tour* disponível em: http://www.uniprot.org/demos/diabetes).

As informações fornecidas pela ferramenta serão listadas abaixo.

- *Name and origin*: informações essenciais da sequência depositada indicando o nome comum da mesma, os sinônimos, os número de identificação etc.
- *Protein attributes*: informações básicas dos atributos da proteína, como o número de aminoácidos, processamento da cadeia polipeptídica etc.
- *General annotation*: descreve, resumidamente, a função proposta e/ou demonstrada para a proteína. No caso das enzimas, descreve a atividade catalítica. Também, apresenta algumas informações estruturais como a localização dentro da célula/tecido. Por fim, podemos encontrar informações sobre modificações pós-traducionais da proteína.
- *Ontologies*: apresentam palavras-chave *(keywords),* entre outras informações, que nos permite associar a função e localização da proteína no contexto biológico celular.
- *Sequence annotation*: contém todas as anotações feitas sobre a sequência, modificações, atualizações etc.
- *Sequence*: mostra a sequência proteica propriamente dita, contendo as informações básicas da mesma, como massa molecular e número de aminoácidos. Se a proteína tiver mais de uma forma (isoforma), ela é indicada e caracterizada junto com a sequência encontrada.
- *References*: contêm as citações de todos os trabalhos que levaram à identificação da proteína (nem sempre bem atualizada).

CAPÍTULO 11

O banco de dados estrutural: *Protein Data Bank* (RCSB PDB)

A função de uma proteína esta associada a sua estrutura e vice-versa. Portanto, a análise da sequência de aminoácidos (estrutura primária), por si só, não é suficiente para compreendermos completamente o que ela faz dentro das células.

Os dados de sequência, características físico-químicas, estrutura e função das proteínas, entre outros, fornecem uma multiplicidade de informações cruciais para entender os processos biológicos. A associação das áreas da Biologia Molecular (descobrindo genes e proteínas) junto com a bioquímica das proteínas (caracterizando a função e atividade enzimática) e com a física aplicada às biomoléculas (estudos de estruturas 3D por Raios X e RMN, por exemplo) deram origem à grande área da Biologia Molecular Estrutural. Esta associação rendeu uma grande quantidade de informações referentes à estrutura 3D das proteínas e ácidos nucleicos, tornando necessária a criação de um banco de dados para tais informações. Surgiu, assim, o PDB ou, em inglês, *Protein Data Bank* (acesso disponível em: http://www.pdb.org) (Figura 43).

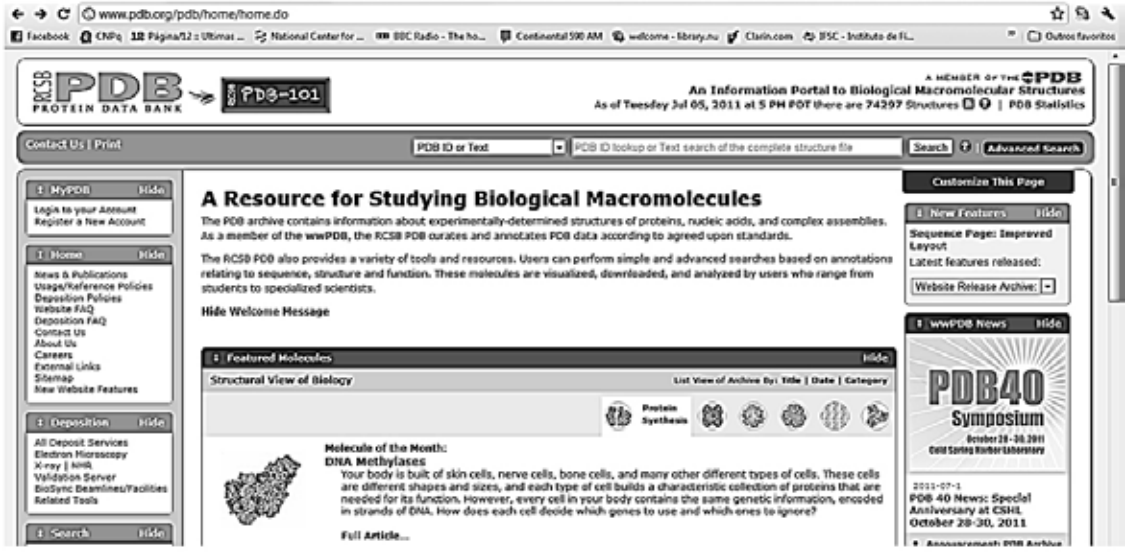

Figura 43. Página do PDB. Utilizada como fonte de estudo de macromoléculas biológicas.

http:// www.pdb.org.

Como os outros bancos de dados, o PDB é de acesso livre e gratuito. Todas as informações podem ser acessadas de qualquer computador. Na atualidade, há mais de 68.000 estruturas de proteínas disponíveis (de um total de 74.000 macromoléculas depositadas) no PDB e este número aumenta a cada semana.

Também, é possível encontrar as estruturas tridimensionais de ácidos nucléicos, como tRNAs, e até subunidades constituintes dos ribossomos de várias espécies.

O constante crescimento do PDB é um reflexo da pesquisa que está acontecendo nos laboratórios de todo o mundo. Isso pode torná-lo emocionante e desafiador tanto para quem o utiliza com fins de pesquisa, como para quem o utiliza como material educativo. Devido, ao grande volume de informações nele contidas, pode ser desafiante selecionar as mais relevantes para um estudo em particular. Para uma determinada proteína é comum identificar estruturas completas e parciais, estruturas obtidas a partir da proteína com mutações geradas *in vitro* ou naturais, estruturas de proteínas associadas a ligantes naturais ou artificiais, entre outras.

É importante destacar que, neste banco de dados, o número de sequências depositados não corresponde ao numero total de proteínas existentes nos bancos de dados primários ou secundários (NCBI e Uniprot, respectivamente). Encontramos apenas aquelas cuja estrutura tridimensional foi determinada experimentalmente com algumas das técnicas mencionadas anteriormente.

A estrutura 3D de uma proteína pode ser visualizada de duas maneiras diferentes. A mais conveniente para quem não esta familiarizado com esse banco de dados é através do próprio *site* com a utilização das ferramentas disponibilizadas. Também, podem ser obtidas e visualizadas (no formato de imagem) no editor de texto (como mostrado na Figura 41). Alternativamente, podem ser baixadas e visualizadas através de *softwares* específicos (como Rasmol – disponível em: http://www.rasmol.org/, por exemplo) que permitem manipular a estrutura protéica de diversas formas.

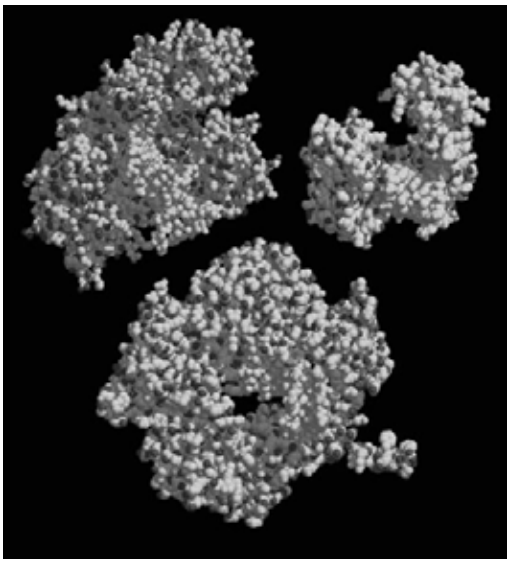

Figura 44. Três enzimas DNA polimerases obtidas do PDB e manipuladas com programa Rasmol.

http://www.pdb.org, código de acesso 1kln e 1zqa.

PARA (NÃO) FINALIZAR

Cada vez mais, tem sido intensa a necessidade da utilização da informática para a análise de dados gerados na pesquisa biomédica. Em alguns setores de pesquisa na Bioquímica e na Biologia Molecular, o uso da informática é imprescindível para que se possa chegar a resultados definitivos a partir dos dados experimentais. A análise de genomas e proteomas, bem como os estudos relativos à relação entre estrutura e função de proteínas e de outras macromoléculas de interesse biológico, são os setores no quais a importância da informática fica mais evidente. Com a explosão de informações relativas às sequências e estruturas disponíveis aos pesquisadores, o campo da Bioinformática ou Biologia Computacional está se destacando na elucidação de aspectos desconhecidos da estrutura e função de genes e proteínas (Figura 45). Assim, por exemplo, podemos determinar a sequência de um gene humano, determinar a sua posição dentro de um cromossomo, obter a sequência proteica da possível proteína codificada por ele e comparar esta sequência com os produtos de genes homólogos descobertos em outras espécies para, finalmente, determinar a estrutura e função da proteína. Ainda, se for o caso de uma proteína envolvida em algum processo patológico, podemos desenhar uma droga que inative sua função.

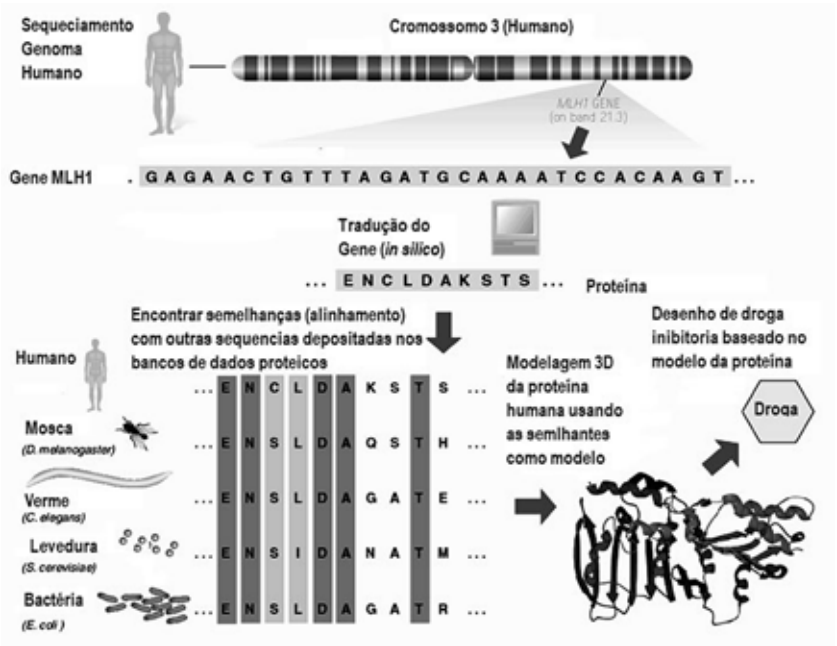

Figura 45. Esquema mostrando como a Bioinformática e a Biologia Molecular atuam em conjunto. Por exemplo, determinam e caracterizam um gene humano e seu produto proteico, resolvem a estrutura proteica e finalizam com o desenho de uma droga que inative a ação desta proteína na célula.

Retirado e modificado de http://www.webartigos.com/articles/17028/1/Tecnologia-microarray-biochips-de-DNA/pagina1.html, em 10/08/2011.

Sem a Bioinformática esse tipo de análise seria altamente complexa e somente realizada por um grupo pequeno de pesquisadores no mundo. Hoje, praticamente qualquer pessoa que tiver acesso a esses bancos de dados tem a possibilidade e as ferramentas para realizar todo tipo de pesquisa relacionada com área da Biologia Molecular que possa envolver complexos processos celulares e as inúmeras patologias que acometem os organismos e até a evolução molecular. Para não finalizar, a Bioinformática e Biologia Molecular convivem em uma espécie de simbiose virtual, onde uma se beneficia da outra e a ciência, como um todo, tira proveito dessa união. As informações estão agora disponíveis para todos nós. Resta apenas irmos atrás das respostas para as inúmeras perguntas que a Biologia Molecular ainda tem que responder.

REFERÊNCIAS

ALBERTS, B. et al. **Fundamentos da biologia celular.** 2. Ed. Porto Alegre: Artmed, 2006.

ALBERTS, B et al. **Molecular biology of the cell.** 4. ed. New York: Garland Science, 2002.

BENJAMIN, L. **Genes VII.** New York: Oxford University Press, 2000.

BENSON, D.A. et al. Genbank. **Nucleic acid res.** v. 24, p. 1-5, 1996.

BENSON, D.A. et al. Genbank. **Nucleic acid res.** v. 30, n. 1, p. 17-20, 2002.

BINNECK, E. As ômicas: integrando a bionformação. **Bioecnologia, ciência e desenvolvimento.** Uberlândia, v. 1, n. 32, p. 28-37, 2004.

CLAVERIE, J-M; NOTREDAME, C. B**ionformatics for dommies.** 2.ed. Indianapolis: Wiley Publishing, 2006.

LEHNINGER, A.L; NELSON, D. L; COX, M.M. **Princípios de bioquímica.** 4. ed. S.I. Sarvier, 2007.

LODISH, H. et al. **Molecular cell biology.** 4. Ed. New York: W. H. Freeman, 2000.

PERUSKI. L. F; PERUSKI, A. H. **The internet and the new biology:** tools dor genomic and molecular research. Washington: ASM Press, 1997.

STOESSER, G. et al. The EMBL nucleotide sequence database. **Nucleic acid res.** v. 30, n. 1, p. 21-26, 2002.

VOET, D. VOET, J.G. **Biochemistry**. 2. ed. New York: John Wiley & Sons, 1997.

WELLER, D. L. et al. Database resorcer of the national center for biotechnology information: 2002 update. **Nucleics acid res.** v. 30, n. 1, p. 13-16, 2002.

WESTBROOK, J. et al. The protein data bank: uniying the archive. **Nucleic acid res.** v. 30, n. 1, p. 245-248, 2002.

Páginas consultadas da internet

http://www.ncbi.nlm.nih.gov/l

http://www.comciencia.br

http://www.dnalc.org/

http://dnaftb.org/

http://pt.wikipedia.org/wiki/Wikip%C3%A9dia:P%C3%A1gina_principal