3

O Genoma Humano: Estrutura e Função dos Genes e Cromossomos

Nos últimos 20 anos, um grande progresso foi feito em nossa compreensão sobre a estrutura e a função dos genes e cromossomos no nível molecular. Mais recentemente, isto foi suplementado por uma compreensão aprofundada da organização do genoma humano no nível de sua seqüência de DNA. Estes avanços deveram-se em grande parte às aplicações da genética molecular e da genômica a muitas situações clínicas, fornecendo assim os instrumentos para um enfoque novo e diferente à genética médica. Neste capítulo apresentamos uma visão geral da organização do genoma humano e os aspectos da genética molecular que são necessários para uma compreensão do enfoque genético à medicina. Este capítulo não pretende fornecer uma ampla descrição da abrangência das novas informações sobre a estrutura e a regulação gênica. Para suplementar a informação discutida aqui, o Cap. 4 descreve muitos enfoques experimentais da genética molecular moderna que estão se tornando cruciais para a prática e a compreensão da genética humana e médica.

O aumento do conhecimento dos genes, bem como de sua organização no genoma, teve um enorme impacto na medicina e na nossa percepção da fisiologia humana. Como disse Paul Berg, ganhador do Prêmio Nobel, no início desta nova era:

> Assim como nosso atual conhecimento e nossa prática da medicina baseiam-se em um sofisticado conhecimento da anatomia humana, da fisiologia e da bioquímica, lidar com a doença no futuro demandará uma compreensão detalhada da anatomia molecular, fisiologia e bioquímica do genoma humano... Precisaremos de um conhecimento mais detalhado sobre como os genes humanos são organizados e como eles funcionam e são regulados. Precisaremos também de médicos que conheçam a anatomia molecular e a fisiologia dos cromossomos e genes, como o cirurgião cardíaco conhece o funcionamento do coração.

ESTRUTURA DO DNA: UMA BREVE REVISÃO

O DNA é uma macromolécula polimérica de ácido nucleico composta de três tipos de unidades: um açúcar de cinco carbonos, a desoxirribose, uma base nitrogenada e um grupo fosfato (Fig. 3.1). As bases são de dois tipos, purinas e pirimidinas. No DNA, existem duas bases purinas, adenina (A) e guanina (G), e duas bases pirimidinas, timina (T) e citosina (C). Os nucleotídeos, cada um composto de uma base, um fosfato e um açúcar, polimerizam-se em longas cadeias polinucleotídicas por ligações fosfodiéster 5'-3' formadas entre unidades adjacentes de desoxirribose (Fig. 3.2). No genoma humano, estas cadeias polinucleotídicas (em sua forma de dupla hélice) têm centenas de milhões de nucleotídeos de tamanho longo, variando de cerca de 50 milhões de pares de bases (para o menor cromossomo, o 21) até 250 milhões de pares de bases (para o maior cromossomo, o 1).

A estrutura anatômica do DNA leva a informação química que permite a transmissão exata da informação genética de uma célula para suas células filhas e de uma geração para a seguinte. Ao mesmo tempo, a estrutura primária do DNA especifica as seqüências de aminoácidos das cadeias polipeptídicas das proteínas, como será descrito mais adiante neste capítulo. O DNA tem características especiais que lhe dão estas propriedades. O estado nativo do DNA, como esclarecido por James Watson e Francis Crick em 1953, é uma dupla hélice (Fig. 3.3). A estrutura helicoidal assemelha-se a uma escada em caracol com giro para a direita, na qual suas duas cadeias polinucleotídicas correm em sentidos opostos, mantidas juntas por pontes de hidrogênio entre os pares de bases: A em uma cadeia pareada com T da outra e G com C (ver Fig. 3.3). Conseqüentemente, o conhecimento da seqüência de nucleotídeos em um filamento automaticamente nos permite determinar a seqüência de bases do outro filamento. A estrutura em dupla hélice das moléculas de DNA permite que elas se repliquem precisamente pela separação dos dois filamentos, seguida pela síntese de dois novos filamentos complementares, de acordo com a seqüência dos filamentos-molde originais (Fig. 3.4). Similarmente, quando necessário, a complementariedade permite um reparo eficiente e correto das moléculas danificadas de DNA.

O DOGMA CENTRAL: DNA → RNA → PROTEÍNA

A informação genética está contida no DNA dos cromossomos dentro do núcleo celular, mas a síntese de proteínas, durante a qual a informação codificada no DNA é usada, ocorre no citoplasma. Esta compartimentalização reflete o fato de que o organismo humano é um **eucarionte**. Isto significa que as células humanas têm um núcleo genuíno que contém o DNA, o qual é separado do citoplasma por uma membrana nuclear. Em contraste, nos procariontes, como na bactéria intestinal *Escherichia coli*, o DNA não está encer-

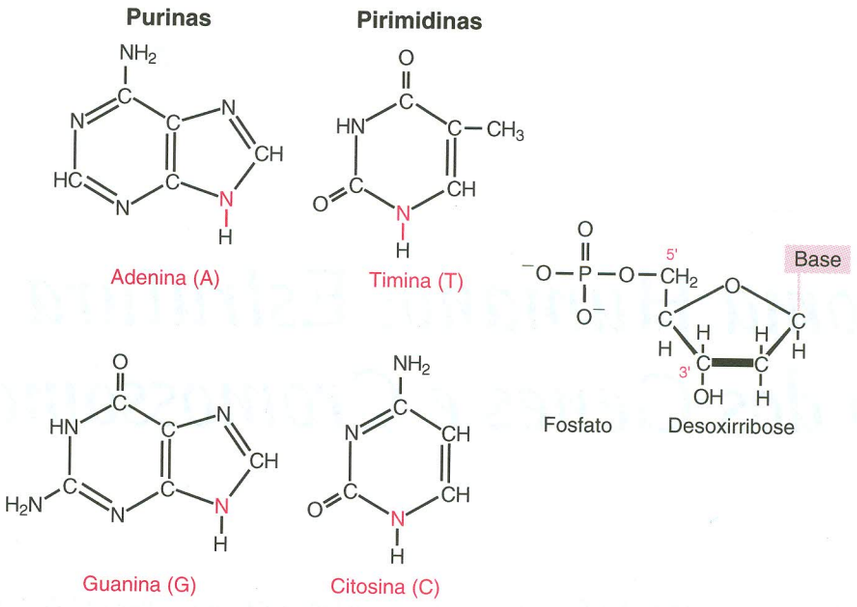

Fig. 3.1 As quatro bases do DNA e a estrutura geral de um nucleotídeo no DNA. Cada uma das quatro bases liga-se à desoxirribose (pelo nitrogênio mostrado em vermelho) e a um grupo fosfato para formar os nucleotídeos correspondentes.

rado dentro de um núcleo. Devido à compartimentalização das células eucarióticas, a transferência de informação do núcleo para o citoplasma é um processo muito complexo, que tem sido foco da atenção dos biólogos moleculares e celulares.

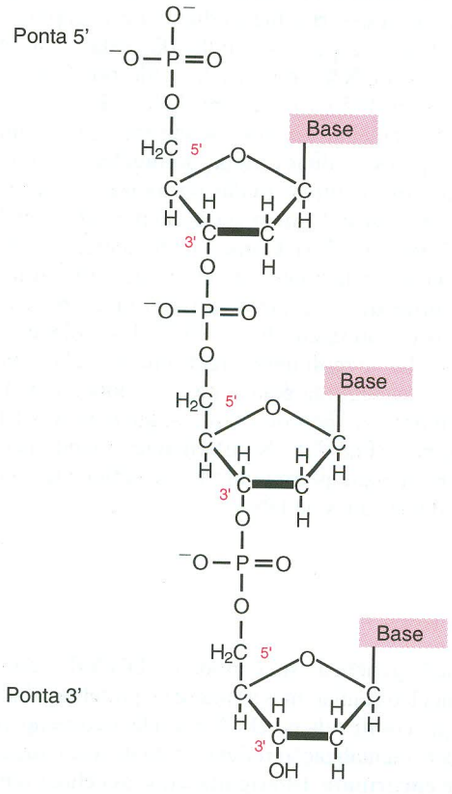

Fig. 3.2 Um trecho de uma cadeia polinucleotídica de DNA, mostrando as ligações fosfodiéster 3'-5' que unem nucleotídeos adjacentes.

A ligação molecular entre estes dois tipos relacionados de informação (o código de DNA dos genes e o código de aminoácidos das proteínas) é o **ácido ribonucleico (RNA)**. A estrutura química do RNA é similar à do DNA, exceto pelo fato de que cada nucleotídeo no RNA tem um açúcar ribose em vez de desoxirribose. Além disso, uracil (U) substitui timina como uma das pirimidinas do RNA (Fig. 3.5). Uma diferença adicional entre o RNA e o DNA é que na maioria dos organismos o RNA existe como uma molécula unifilamentar, enquanto o DNA existe como uma dupla hélice.

A correlação informacional entre DNA, RNA e proteína está interligada: o DNA dirige a síntese e a seqüência do RNA, o RNA dirige a síntese e a seqüência dos polipeptídeos e especifica as proteínas que estão envolvidas na síntese e no metabolismo do DNA e do RNA. Este fluxo de informação é chamado de "dogma central" da biologia molecular.

A informação genética é estocada no DNA por meio de um código (o **código genético**, discutido mais adiante), no qual a seqüência de bases adjacentes determina a seqüência de aminoácidos no polipeptídeo codificado. Primeiro, o RNA é sintetizado a partir de um molde de DNA por um processo conhecido como **transcrição**. O RNA, levando a informação codificada em uma forma chamada de **RNA mensageiro (mRNA)**, é então transportado do núcleo para o citoplasma, onde a seqüência de RNA é decodificada, ou traduzida, para determinar a seqüência de aminoácidos na proteína que está sendo sintetizada. O processo de **tradução** ocorre nos **ribossomos**, que são organelas citoplasmáticas com sítios de ligação para todas as moléculas que interagem, incluindo o mRNA, envolvidas na síntese de proteínas. Os ribossomos são feitos de muitas proteínas estruturais diferentes em associação com um tipo específico de RNA, conhecido como **RNA ribossômico (rRNA)**. A tradução envolve ainda um terceiro tipo de RNA, o **RNA transportador (tRNA)**, que fornece a ligação molecular entre a seqüência de bases do mRNA e a seqüência de bases da proteína.

Devido ao fluxo interdependente de informação representado pelo dogma central, podemos começar a discussão da genéti-

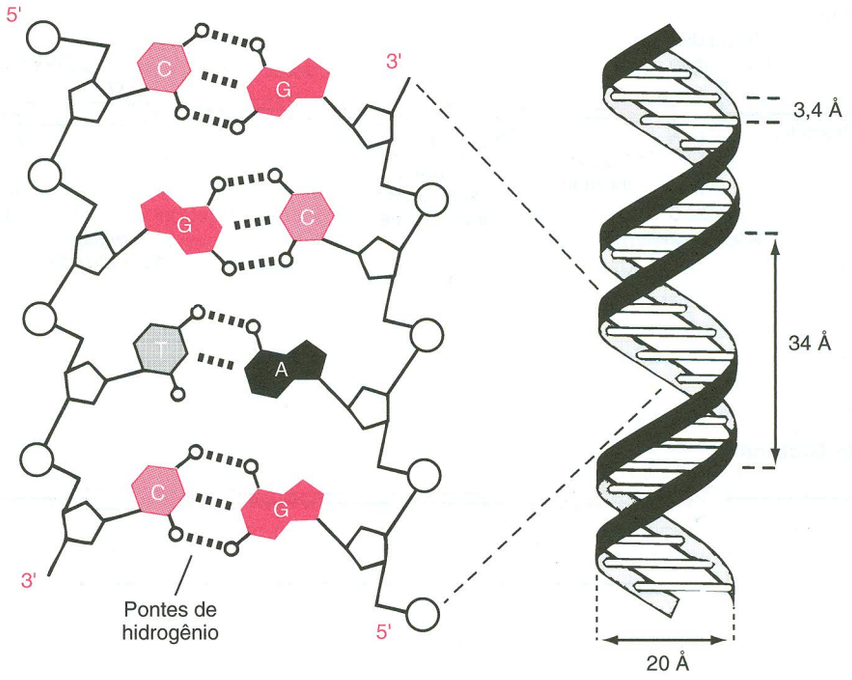

Fig. 3.3 A estrutura do DNA. *Esquerda*: Uma representação bidimensional dos dois filamentos complementares de DNA, mostrando os pares de bases AT e GC. Notar que a orientação dos dois filamentos é invertida. *Direita*: O modelo da dupla hélice de DNA, como proposto por Watson e Crick. Os "degraus" horizontais representam as bases pareadas. A hélice é dita dextrógira porque o filamento que vai do canto inferior esquerdo para o superior direito cruza o filamento oposto. (Baseado em Watson J. D., Crick F. H. C. (1953) Molecular structure of nucleic acids — A structure for deoxyribose nucleic acid. Nature 171:737-738.)

ca molecular da expressão gênica em qualquer um dos três níveis informacionais: DNA, RNA ou proteína. Começaremos a examinar a estrutura dos genes como uma base para nossa discussão do código genético, transcrição e tradução.

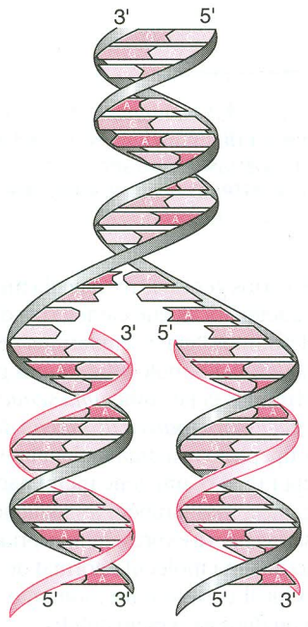

Fig. 3.4 Replicação de uma dupla hélice de DNA que resulta em duas moléculas filhas idênticas, cada uma composta de um filamento parental (*preto*) e um recém-sintetizado (*vermelho*).

Estrutura e Organização do Gene

Em sua forma simples, um gene pode ser visualizado como um segmento de uma molécula de DNA contendo o código para a seqüência de aminoácidos de uma cadeia polipeptídica e as seqüências reguladoras necessárias para a expressão. Esta descrição, entretanto, é inadequada para os genes do genoma humano (e, na verdade, para a maioria dos genomas eucarióticos), pois alguns genes existem como seqüências codificantes contínuas. A grande maioria dos genes é interrompida por uma ou mais regiões não-codificantes. Estas seqüências intercalares, chamadas **íntrons**, são inicialmente transcritas em RNA no núcleo, mas não estão presentes no mRNA final no citoplasma. Assim, a informação das seqüências intrônicas normalmente não é representada no produto proteico final. Os íntrons alternam-se com seqüências codificantes, ou **éxons**, que finalmente codificam a seqüência de aminoácidos da proteína (Fig. 3.6). Embora alguns genes do ge-

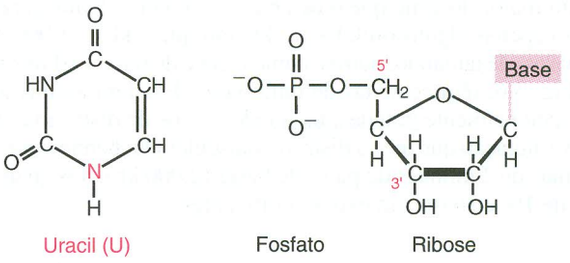

Fig. 3.5 A pirimidina uracil e a estrutura de um nucleotídeo no RNA. Notar que o açúcar ribose substitui o açúcar desoxirribose do DNA. Comparar com a Fig. 3.1.

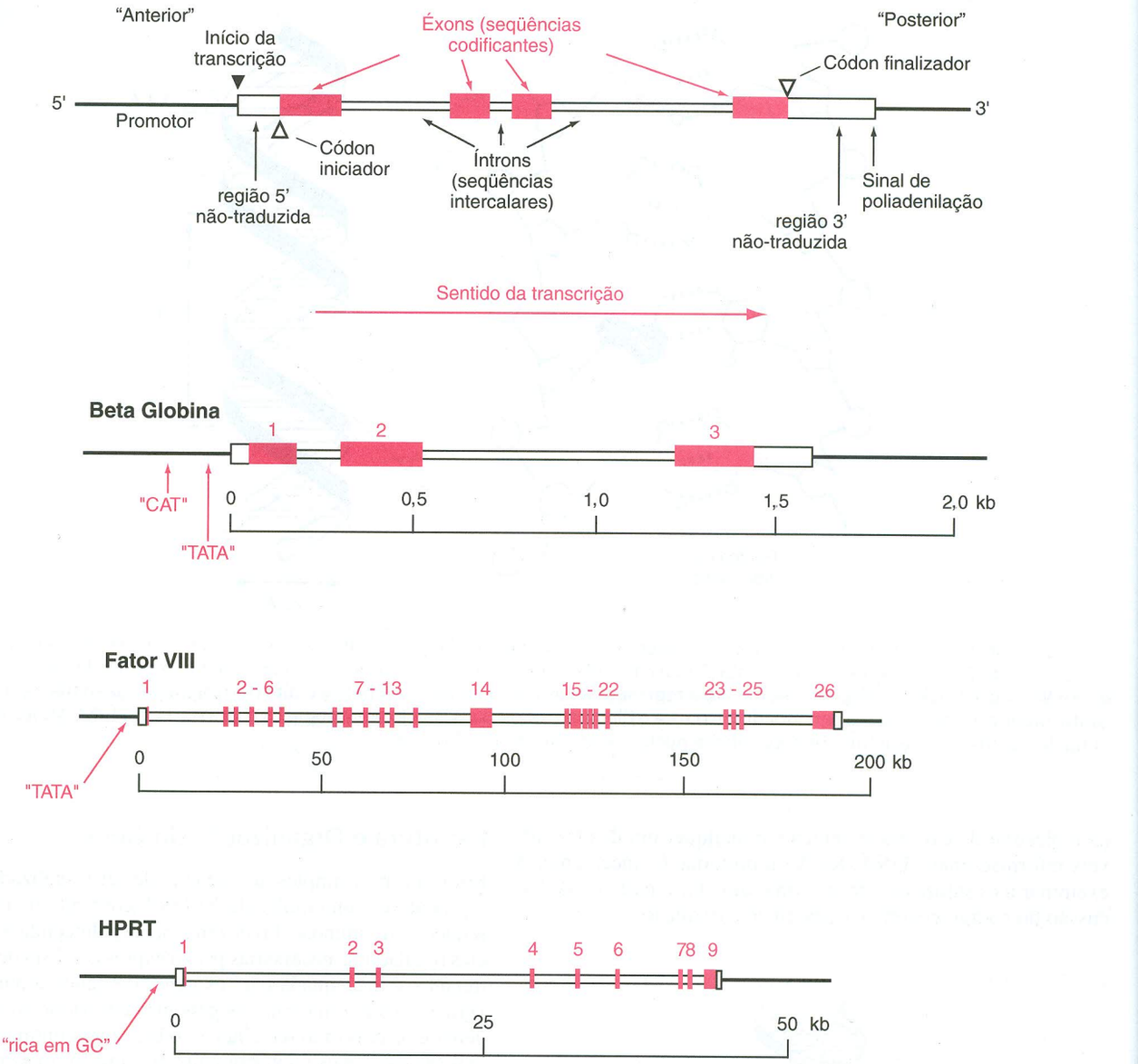

Fig. 3.6 Estrutura geral de um gene humano típico. As características individuais estão marcadas na figura e são discutidas no texto. Os exemplos de três genes humanos de importância médica são apresentados na parte inferior da figura. Os éxons individuais são numerados. Mutações diferentes no gene de β-globina causam uma variedade de hemoglobinopatias importantes. As mutações no gene de fator VIII causam hemofilia A. As mutações no gene de hipoxantina fosforibosiltransferase (*HPRT*) levam à síndrome de Lesch-Nyhan.

noma humano não tenham íntrons, a maioria dos genes contém pelo menos um e em geral vários íntrons. Surpreendentemente, em muitos genes, o tamanho cumulativo dos íntrons é uma proporção muito maior do gene que o dos éxons. Embora alguns genes tenham apenas alguns quilobases (kb, em que 1 kb = 1.000 pares de bases) de tamanho, outros, como o gene de fator VIII mostrado na Fig. 3.6, têm centenas de quilobases. Existem alguns genes excepcionalmente grandes, incluindo o gene de distrofina ligado ao X (mutação que leva à distrofia muscular Duchenne), que ocupa mais de 2 milhões de pares de bases (2.000 kb), dos quais menos de 1% consiste em éxons codificantes.

Características Estruturais de um Gene Humano Típico

Uma representação esquemática de uma parte do DNA cromossômico contendo um gene típico é mostrada na Fig. 3.6, juntamente com a estrutura de vários genes de relevância médica. Juntos, ilustram a gama de características que caracteriza os genes humanos. Nos Caps. 1 e 2, definimos "gene" em termos gerais. Neste ponto, podemos dar uma definição molecular de um gene. Em circunstâncias típicas, definimos gene como *uma seqüência de DNA cromossômico que é necessária para a produção de um produto funcional*, seja um polipeptídeo ou uma molécula funcional de RNA. Como está claro na Fig. 3.6, um gene inclui não só as seqüências realmente codificantes, mas também as seqüências nucleotídicas adjacentes necessárias para a expressão apropriada do gene — isto é, para a produção de uma molécula normal de mRNA, na quantidade correta, no local correto e no momento correto durante o desenvolvimento ou durante o ciclo celular.

As seqüências adjacentes de nucleotídeos fornecem os sinais moleculares de "início" e "fim" para a síntese do mRNA transcrito do gene. Na ponta 5' do gene está uma região **promotora**,

que inclui seqüências responsáveis pelo início apropriado da transcrição. Dentro da região 5' estão vários elementos do DNA cuja seqüência é conservada entre muitos genes diferentes. Esta conservação, junto com os estudos funcionais da expressão gênica em muitos laboratórios, indica que estes elementos particulares de seqüência têm um papel importante na regulação. Existem vários tipos diferentes de promotor encontrados no genoma humano, com diferentes propriedades reguladoras que especificam os padrões de desenvolvimento, bem como os níveis de expressão de um determinado gene em tecidos diferentes. Os papéis dos elementos promotores individualmente conservados, identificados na Fig. 3.6, são discutidos em maiores detalhes na seção "Fundamentos da Expressão Gênica". Tanto os promotores quanto outros elementos reguladores (situados em 5' ou 3' de um gene ou em seus íntrons) podem ser sítios de mutação nas doenças genéticas que podem interferir com a expressão normal de um gene. Estes elementos reguladores, incluindo os **acentuadores**, os **silenciadores** e as **regiões controladoras de locus** (LCRs), serão discutidos mais amplamente mais adiante neste capítulo.

Na ponta 3' de um gene está uma região não-traduzida de importância que contém um sinal de adição de uma seqüência de unidades adenosina (a chamada cauda poliA) na ponta do mRNA final. Embora em geral seja aceito que tais seqüências reguladoras proximamente vizinhas são parte do que é chamado de "gene", as dimensões exatas de qualquer gene específico permanecerão um tanto incertas até que as funções potenciais das seqüências mais distantes sejam totalmente caracterizadas.

Famílias de Genes

Muitos genes pertencem a famílias de seqüências de DNA proximamente relacionadas, reconhecidas como famílias por causa da similaridade da seqüência de nucleotídeos dos próprios genes ou da seqüência de aminoácidos dos polipeptídeos codificados.

Uma família de genes pequena, mas medicamente importante, é composta de genes que codificam as cadeias de proteína encontradas nas hemoglobinas. Os grupamentos gênicos de α-globina e de β-globina, nos cromossomos 16 e 11, respectivamente, são mostrados na Fig. 3.7 e são tidos como tendo surgido por duplicação de um gene precursor primitivo há cerca de 500 milhões de anos. Estes dois grupamentos contêm genes que codificam cadeias de globina proximamente relacionadas expressas em estágios de desenvolvimento diferentes, do embrião até o adulto. Os genes individuais dentro de cada grupamento são mais similares em seqüência uns aos outros que a genes de outros grupamentos. Assim, cada grupo é tido como tendo evoluído de uma série de eventos seqüenciais de duplicação gênica nos últimos 100 milhões de anos. Os padrões éxon-íntron de genes de globina parecem ter sido bastante conservados durante a evolução. Cada um dos genes funcionais de globina mostrados na Fig. 3.7 tem dois íntrons em locais similares, embora as seqüências contidas dentro dos íntrons tenham acumulado muito mais mudanças de bases nucleotídicas com o tempo que as seqüências codificantes de cada gene. O controle da expressão dos vários genes de globina, no estado normal bem como nas muitas hemoglobinopatias herdadas, será considerado em maiores detalhes tanto mais adiante, neste capítulo, quanto no Cap. 11.

Vários genes de globina não produzem nenhum RNA ou produto proteico e, portanto, é improvável que tenham alguma função. As seqüências de DNA que se assemelham muito a genes conhecidos mas não são funcionais são chamadas de **pseudogenes**. Os pseudogenes estão dispersos no genoma e são tidos como subprodutos da evolução, representando genes que já foram funcionais mas que agora são vestigiais, tendo sido inativados por mutações em seqüências codificantes ou reguladoras. Em alguns casos, como nos genes de pseudo-α-globina e pseudo-β-globina, os pseudogenes supostamente surgiram por duplicação gênica, seguida da introdução de várias mutações nas cópias extras do gene que já foi funcional. Em outros casos, os pseudogenes foram formados por um processo, chamado de **retrotransposição**, que envolve a transcrição, a geração de uma cópia de DNA do mRNA e, finalmente, da integração de tais cópias de DNA ao genoma. Os pseudogenes criados por retrotransposição não têm íntrons e são chamados de **pseudogenes processados**. Eles não estão necessária ou usualmente no mesmo cromossomo (ou região cromossômica) que seu gene genitor.

A maior família de genes conhecida no genoma humano é a chamada **superfamília de imunoglobulina**, que inclui muitas centenas de genes envolvidos nos eventos de reconhecimento da superfície celular nos sistemas imune e nervoso, tais como os genes nos cromossomos 2, 14 e 22, que codificam as próprias cadeias pesada e leve de imunoglobulinas, os genes no cromossomo 6, que constituem o complexo principal de histocompatibilidade, os genes nos cromossomos 7 e 14, cujos produtos constituem o receptor de células T, e os genes que são expressos primariamente em tecidos neurais, tais como os genes para moléculas de adesão celular ou glicoproteínas associadas à mielina. A estrutura e a função de muitos destes genes serão examinadas em detalhes no Cap. 14.

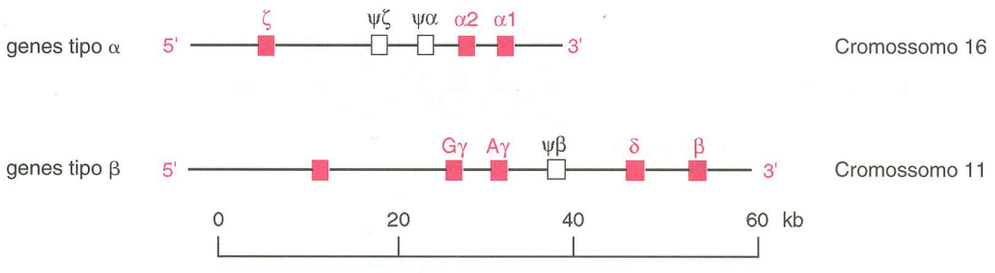

Fig. 3.7 Organização cromossômica de dois grupos de genes de globina humana. Os genes funcionais são indicados em vermelho. Os pseudogenes são indicados pelos boxes vazados. (Redesenhado de Nienhuis A. W., Maniatis T. [1987] Structure and expression of globin genes in erythroid cells. *In* Stamatoyannopoulos G., Nienhuis A. W., Leder P., Majerus P. W. [eds.] The Molecular Basis of Blood Diseases. W. B. Saunders, Philadelphia, pp. 28-65.)

FUNDAMENTOS DA EXPRESSÃO GÊNICA

O fluxo de informações do gene para o polipeptídeo envolve várias etapas (Fig. 3.8). O início da transcrição de um gene está sob a influência dos promotores e de outros elementos reguladores, bem como de proteínas específicas, conhecidas como **fatores de transcrição**, que interagem com seqüências específicas dentro destas regiões. A transcrição de um gene é iniciada no "ponto de início" transcricional no DNA cromossômico antecedente às seqüências codificantes e continua ao longo do cromossomo, indo desde várias centenas de pares de bases a mais de um milhão de pares de bases, ao longo tanto de íntrons quanto de éxons e além do final das seqüências codificantes. Após a modificação tanto das pontas 5' quanto das pontas 3' do transcrito primário de RNA, as partes correspondentes a íntrons são removidas e os segmentos correspondentes a éxons são reunidos. Após a recomposição do RNA, o mRNA resultante (agora colinear às partes codificantes do gene) é transportado do núcleo para o citoplasma, onde o mRNA é finalmente traduzido na seqüência de aminoácidos do polipeptídeo codificado. Cada uma das etapas desta via complexa está sujeita a erro, e as mutações que interferem nas etapas individuais foram implicadas em vários distúrbios genéticos (ver Caps. 5, 11 e 12).

Transcrição

A transcrição de genes codificantes de proteínas pela RNA polimerase II (uma das várias classes de RNA polimerases) é iniciada um pouco antes da primeira seqüência codificante no ponto inicial de transcrição, o ponto que corresponde à ponta 5' do produto final de RNA (ver Figs. 3.6 e 3.8). A síntese do transcrito primário de RNA ocorre no sentido 5' para 3', enquanto o filamento do gene transcrito é de fato lido no sentido 3' para 5' com relação à direção da estrutura desoxirribose fosfodiéster (ver Fig. 3.2). Como o RNA sintetizado corresponde tanto em polaridade quanto em seqüência de bases (substituindo T por U) ao filamento do DNA 5' para 3', este filamento de DNA não-transcrito às vezes é chamado de "codificante" ou "**com sentido**". O filamento de DNA 3' para 5' transcrito é então chamado de "não-codificante" ou "**anti-sentido**".* A trancrição continua incluindo íntrons e éxons do gene, além da posição no cromossomo que eventualmente corresponde à ponta 3' do mRNA final. Não sabemos se a transcrição termina em ponto final 3' predeterminado.

O transcrito primário de RNA é processado pela adição de uma estrutura química "*cap*" na ponta 5' do RNA e uma clivagem na ponta 3' em um ponto específico ao final da informação codificante. Esta clivagem é seguida pela adição de uma cauda poliA na ponta 3' do RNA. A cauda poliA parece aumentar a estabilidade do RNA poliadenilado resultante. O ponto de poliadenilação é especificado em parte pela seqüência AAUAAA

*N.T.: Esta inversão é uma causa constante de confusão por parte dos alunos.

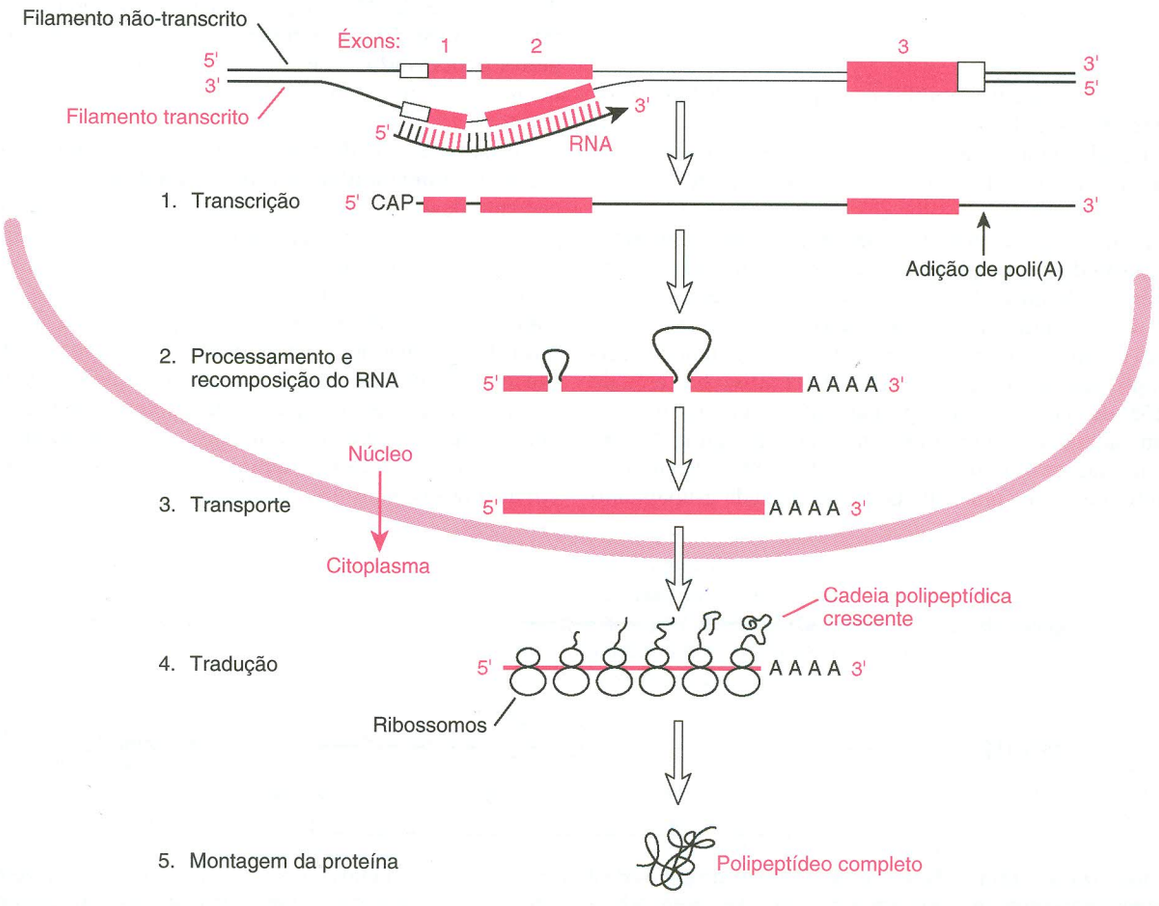

Fig. 3.8 Fluxo de informação do DNA para o RNA para a proteína de um gene hipotético com três éxons e dois íntrons. As etapas incluem transcrição, processamento e recomposição do RNA, transporte do núcleo para o citoplasma e tradução.

(ou uma variante disto), em geral encontrada na parte 3' não-traduzida do RNA transcrito. Estas modificações pós-transcricionais ocorrem no núcleo, assim como o processamento do RNA. O RNA totalmente processado agora é chamado de mRNA e é transportado para o citoplasma, onde ocorre a tradução (ver Fig. 3.8).

Tradução e o Código Genético

No citoplasma, o mRNA é traduzido em proteína pela ação de uma variedade de moléculas de tRNA, cada uma específica para um determinado aminoácido. Estas moléculas, cada uma com apenas de 70 a 100 nucleotídeos de tamanho, têm a função de transferir os aminoácidos corretos para suas posições ao longo do molde de mRNA, para que sejam adicionados à cadeia polipeptídica crescente. A síntese de proteínas ocorre nos ribossomos, complexos macromoleculares feitos de rRNA (codificados pelos genes de rRNA 18S e 28S) e várias dúzias de proteínas ribossomiais (ver Fig. 3.8).

A chave para a tradução é um código que relaciona aminoácidos específicos a combinações de três bases adjacentes ao longo do mRNA. Cada conjunto de três bases constitui um **códon** específico para um determinado aminoácido (Quadro 3.1). Em teoria, são possíveis variações quase infinitas na disposição das bases ao longo de uma cadeia polinucleotídica. Em qualquer posição existem quatro possibilidades (A, T, C ou G). Assim, existem 4^n combinações possíveis em uma seqüência de n bases. Para três bases, há 4^3, ou 64, combinações possíveis de trincas. Estes 64 códons constituem o **código genético**.

Como existem apenas 20 aminoácidos e 64 possíveis códons, a maioria dos aminoácidos é especificada por mais de um códon. Por isto o código é dito **degenerado**. Por exemplo, a base na terceira posição da trinca pode ser uma purina (A ou G), ou uma pirimidina (T ou C) ou, em alguns casos, qualquer uma das quatro bases, sem alterar a mensagem codificada (ver Quadro 3.1). A leucina e arginina são especificadas por seis códons. Apenas a metionina e o triptofano são cada um deles especificado por um único códon. Três dos códons são chamados de **códons finalizadores** (ou **sem sentido**) porque designam o término da tradução do mRNA neste ponto.

A tradução de um mRNA processado é sempre iniciada em um códon que especifica metionina. A metionina é, portanto, o primeiro aminoácido codificado (amino-terminal) de cada cadeia polipeptídica, embora em geral seja removido antes que a síntese da proteína esteja completa. O códon para metionina (o códon

QUADRO 3–1

O Código Genético

Primeira Base	Segunda Base: *U*		*C*		*A*		*G*		Terceira Base
U	UUU	fen	UCU	ser	UAU	tir	UGU	cis	U
	UUC	fen	UCC	ser	UAC	tir	UGC	cis	C
	UUA	leu	UCA	ser	UAA	fim	UGA	fim	A
	UUG	leu	UCG	ser	UAG	fim	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC	leu	CCC	pro	CAC	his	CGC	arg	C
	CUA	leu	CCA	pro	CAA	gln	CGA	arg	A
	CUG	leu	CCG	pro	CAG	gln	CGG	arg	G
A	AUU	ile	ACU	tre	AAU	asn	AGU	ser	U
	AUC	ile	ACC	tre	AAC	asn	AGC	ser	C
	AUA	ile	ACA	tre	AAA	lis	AGA	arg	A
	AUG	met	ACG	tre	AAG	lis	AGG	arg	G
G	GUU	val	GCU	ala	GAU	asp	GGU	gli	U
	GUC	val	GCC	ala	GAC	asp	GGC	gli	C
	GUA	val	GCA	ala	GAA	glu	GGA	gli	A
	GUG	val	GCG	ala	GAG	glu	GGG	gli	G

Abreviações dos aminoácidos:

ala (A)	alanina	leu (L)	leucina
arg (R)	arginina	lis (K)	lisina
asn (N)	asparagina	met (M)	metionina
asp (D)	ácido aspártico	fem (F)	fenilalanina
cis (C)	cisteína	pro (P)	prolina
gln (Q)	glutamina	ser (S)	serina
glu (E)	ácido glutâmico	tre (T)	treonina
gli (G)	glicina	trp (W)	triptofano
his (H)	histidina	tir (Y)	tirosina
ile (I)	isoleucina	val (V)	valina

Outra abreviação:

fim	códon finalizador

Os códons são mostrados em termos do mRNA, que são complementares aos códons correspondentes no DNA.

iniciador, AUG) estabelece a **matriz de leitura** do mRNA. Cada códon subseqüente é lido, por sua vez, para prever a seqüência de aminoácidos da proteína.

As ligações moleculares entre os códons e os aminoácidos são as moléculas específicas de tRNA. Um determinado sítio de cada tRNA forma um **anticódon** com três bases que é complementar a um códon específico no mRNA. A ligação entre o códon e o anticódon coloca o aminoácido apropriado na posição seguinte no ribossomo para a ligação pela formação de uma ligação peptídica com a ponta carboxila e a cadeia polipeptídica crescente. O ribossomo então desliza ao longo do mRNA a cada três bases, alinhando o códon seguinte para o reconhecimento por outro tRNA com o aminoácido seguinte. Assim, as proteínas são sintetizadas a partir de seu terminal amino até o terminal carboxila, que corresponde à tradução do mRNA no sentido 5' para 3'.

Como mencionado antes, a tradução termina quando um códon finalizador (UGA, UAA ou UAG) é encontrado na mesma matriz de leitura que o códon iniciador. (Os códons finalizadores em uma das matrizes de leitura não são lidos e, portanto, não têm efeito na tradução.) O polipeptídeo completo é então liberado do ribossomo, que se torna disponível para começar a síntese de outra proteína.

Processamento Pós-traducional

Muitas proteínas sofrem amplas modificações pós-traducionais. A cadeia polipeptídica que é o produto primário da tradução é dobrada e associada em uma estrutura tridimensional específica que é determinada pela própria seqüência de aminoácidos. Duas ou mais cadeias polipeptídicas, produtos do mesmo gene ou de genes diferentes, podem se combinar para formar um único complexo proteico final. Por exemplo, duas cadeias de α-globina e duas cadeias de β-globina associam-se de modo não convalente para formar a molécula tetramérica de hemoglobina $\alpha_2\beta_2$. Os produtos proteicos também podem ser modificados quimicamente por, por exemplo, adição de fosfato ou carboidratos em sítios específicos. Outras modificações podem envolver a clivagem da proteína, seja para remover seqüências amino-terminais específicas após terem direcionado uma proteína para o seu local correto dentro da célula (p. ex., proteínas que funcionam dentro do núcleo ou mitocôndrias) ou para dividir a molécula em cadeias polipeptídicas menores. Por exemplo, as duas cadeias que constituem a molécula de insulina, uma com 21 e a outra com 30 aminoácidos, originalmente são parte de uma tradução primária com 82 aminoácidos, chamada de pró-insulina.

A Expressão Gênica em Ação: O Gene de Beta-Globina

O fluxo de informações destacado nas seções anteriores pode ser mais bem apreciado com relação a um gene particularmente bem estudado, o gene de β-globina. A cadeia de β-globina é um polipeptídeo com 146 aminoácidos, codificada por um gene que ocupa aproximadamente 1,6 kb no braço curto do cromossomo 11. O gene tem três éxons e dois íntrons (ver Fig. 3.6). O gene de β-globina, bem como outros genes no grupo de β-globina (ver Fig. 3.7), é transcrito em uma direção do centrômero para o telômero. Esta orientação, entretanto, é diferente para outros genes no genoma e depende de qual filamento da dupla hélice cromossômica é o filamento codificante de um determinado gene.

As seqüências de DNA necessárias para uma iniciação precisa da transcrição do gene de β-globina estão situadas no promotor em cerca de 200 pares de bases antecedentes ao ponto de início da transcrição. A seqüência de dupla hélice de DNA desta região do gene de β-globina, a seqüência correspondente de RNA e a seqüência traduzida dos 10 primeiros aminoácidos são mostradas na Fig. 3.9 para ilustrar as relações entre estes três níveis de informação. Como mencionado antes, é o filamento 3' para 5' do DNA que serve como molde e é transcrito, mas é o filamento 5' para 3' do DNA que corresponde mais diretamente à seqüência 5' para 3' do mRNA (e, de fato, é idêntica a ele, exceto por U substituindo T). Devido à sua correspondência, o filamento 5' para 3' do DNA de um gene (o filamento que *não é* transcrito) é o filamento em geral relatado na literatura científica ou nos bancos de dados.

De acordo com esta convenção, a seqüência completa de aproximadamente 2,0 kb do cromossomo 11 que inclui o gene de β-globina é mostrada na Fig. 3.10. (Vale a pena refletir que esta página de nucleotídeos representa apenas 0,000067% da seqüência de todo o genoma humano). Dentro destes 2,0 kb está contida a maioria dos elementos de seqüência necessários para codificar e regular a expressão deste gene, mas não todos. Na Fig. 3.10 são indicadas muitas das características importantes do gene de β-globina, incluindo os elementos de seqüência promotora conservados, os limites de íntrons e éxons, os sítios de corte do RNA, os códons iniciadores e finalizadores e o sinal de poliadenilação, todos os quais são conhecidos como estando mutados em vários defeitos herdados do gene de β-globina (ver Cap. 11).

Início da Transcrição

O promotor de β-globina, como muitos outros promotores gênicos, consiste em uma série de elementos funcionais relati-

Fig. 3.9 Estrutura da seqüência de nucleotídeos da ponta 5' do gene humano de β-globina no braço curto do cromossomo 11. A transcrição do filamento 3' para 5' (*inferior*) começa no ponto indicado para produzir o mRNA de β-globina. A matriz de leitura traducional é determinada pelo códon iniciador AUG (***); os códons subseqüentes que especificam aminoácidos são indicados em vermelho. As outras duas matrizes potenciais não são usadas.

Fig. 3.10 Seqüência de nucleotídeos no gene humano completo de β-globina. É mostrada a seqüência do filamento 5' para 3' no gene. As letras maiúsculas representam seqüências que correspondem ao mRNA final. As letras minúsculas indicam íntrons e seqüências flanqueadoras. As seqüências CAT e TATA box na região flanqueadora 5' são indicadas. O códon iniciador ATG (AUG no mRNA) e o códon finalizador TAA (UAA no mRNA) são mostrados em vermelho. A seqüência de aminoácidos de β-globina é mostrada acima da seqüência codificante. As abreviações com três letras do Quadro 3.1 são usadas aqui. Os dinucleotídeos GT e AG, importantes para a recomposição do RNA e nas junções íntron/éxon, são mostrados em boxes. (De Lawn R. M., Efstratiadis A., O'Connell C., *et al* [1980] The nucleotide sequence of the human β-globin gene. Cell 21:647-651.)

vamente curtos que são tidos como interagindo com proteínas específicas (em geral chamadas de **fatores de transcrição**) que regulam a transcrição, incluindo, no caso dos genes de globina, as proteínas que restringem a expressão destes genes a células eritróides, as células nas quais é produzida a hemoglobina. Uma seqüência promotora importante é o "TATA boxe", uma região conservada rica em adeninas e timinas que está cerca de 25 a 30 pares de bases antecedendo o ponto de início da transcrição (ver Figs. 3.6 e 3.10). O TATA boxe parece ser importante para a determinação da posição do início da transcrição, que no gene de β-globina está aproximadamente a 50 pares de bases antecedentes ao ponto de início da tradução (ver Fig. 3.9). Assim, neste gene existem cerca de 50 pares de bases de seqüência que são transcritos mas não são traduzidos. Em outros genes, esta região 5' transcrita mas não traduzida (chamada de 5' UTR) pode ser muito maior e, de fato, pode ser interrompida por um ou mais íntrons. Uma segunda região conservada, o chamado CAT boxe (na verdade, CCAAT), está antecedente há algumas dezenas de pares de bases (ver Fig. 3.10). Tanto as mutações experimentalmente induzidas quanto aquelas de ocorrência natural em um destes elementos de seqüência, bem como em outras seqüências reguladoras mais antecedentes, levam a uma intensa redução no nível de transcrição, demonstrando assim a importância destes elementos para a expressão normal do gene. Muitas mutações nestes elementos reguladores foram identificadas em pacientes com o distúrbio β-talassemia (ver Cap. 11).

Nem todos os promotores gênicos contêm os dois elementos específicos descritos. Em particular, os genes que são expressos constitutivamente na maioria ou em todos os tecidos (chamados de "genes de manutenção" [*housekeeping genes*]) em geral não têm o CAT e TATA boxes, que são mais típicos de genes histoespecíficos. Os promotores de muitos genes de manutenção em geral contêm uma alta proporção de citosinas e guaninas em relação ao DNA circundante (ver o promotor do gene de hipoxantina fosforibosiltransferase na Fig. 3.6). Tais promotores ricos em CG em geral estão situados nas regiões do genoma chamadas de **ilhas de CG** (ou de **CpG**), assim chamadas em função da alta concentração do dinucleotídeo 5'-CG-3' que fica fora da região cromossômica mais geral rica em AT. Algumas das seqüências ricas em CG encontradas nestes promotores são tidas como servindo como pontos de ligação para fatores específicos de transcrição.

Em adição às seqüências que constituem o próprio promotor, existem outros elementos de seqüência que podem alterar muito a eficiência da transcrição. As mais bem caracterizadas destas seqüências "ativadoras" são chamadas de **acentuadores** (*enhancers*). Os acentuadores são elementos de seqüência que podem agir à distância (em geral vários kb) de um gene para estimular a transcrição. Ao contrário dos promotores, os acentuadores são independentes de posição e orientação e podem estar localizados a 5' ou a 3' do ponto de início da transcrição. Os elementos acentuadores funcionam apenas em alguns tipos de células e, portanto, parecem estar envolvidos no estabelecimento da especificidade tissular e/ou o nível de expressão de muitos genes, em conjunto com um ou mais fatores de transcrição. No caso do gene de β-globina, vários acentuadores histoespecíficos estão presentes dentro do próprio gene e em suas regiões flanqueadoras. A interação de acentuadores com determinadas proteínas leva a níveis aumentados de transcrição.

A expressão normal do gene de β-globina durante o desenvolvimento também requer uma seqüência mais distante, chamada de **região controladora de locus (LCR)** e situada antecedendo o gene de ε-globina (ver Fig. 3.7), que é necessária para a expressão apropriada em alto nível. Como esperado, as mutações que perturbam ou deletam as seqüências acentuadoras ou LCR interferem ou impedem a expressão do gene de β-globina (ver Cap. 11).

Recomposição do RNA

O transcrito primário de RNA do gene de β-globina contém dois éxons, cerca de 100 e 850 pares de base de tamanho, que precisam ser reunidos. O processo é exato e muitíssimo eficiente. Noventa e cinco por cento dos transcritos de β-globina são tidos como sendo recompostos precisamente para gerar um mRNA funcional de globina. As reações de recomposição são guiadas por seqüências específicas tanto nas pontas 5' quanto nas pontas 3' dos íntrons. A seqüência 5' consiste em nove nucleotídeos, dos quais dois (o dinucleotídeo GT, situado no íntron imediatamente adjacente ao ponto de corte) são virtualmente invariantes entre os pontos de corte em genes diferentes (ver Fig. 3.10). A seqüência 3' consiste em cerca de uma dúzia de nucleotídeos, dos quais, mais uma vez, dois, os AG situados imediatamente a 5' do limite íntron/éxon, são obrigatórios para a recomposição normal. Os próprios pontos de corte não estão relacionados com a matriz de leitura do mRNA em particular. Em alguns casos, como no do íntron 1 do gene de β-globina, o íntron corta um códon específico (ver Fig. 3.10).

A importância médica da recomposição do RNA é ilustrada pelo fato de que as mutações dentro das seqüências conservadas nos limites íntron/éxon comumente prejudicam a recomposição do RNA, com a redução concomitante da quantidade de mRNA final normal de β-globina; as mutações nos dinucleotídeos GT ou AG mencionadas antes invariavelmente eliminam a remoção normal do íntron contendo a mutação. Várias mutações de sítio de corte, identificadas em pacientes com β-talassemia, serão discutidas em detalhes no Cap. 11.

Poliadenilação

O mRNA final de β-globina contém aproximadamente 130 pares de bases do material de 3' não-traduzido (3' UTR) entre o códon finalizador e o local da cauda poliA (ver Fig. 3.10). Como em outros genes, a clivagem da ponta 3' do mRNA e a adição da cauda poliA é controlada, pelo menos em parte, por uma seqüência AAUAAA de cerca de 20 pares de bases antes do sítio de poliadenilação. As mutações neste sinal de poliadenilação nos pacientes com β-talassemia (bem como mutações no sinal de poliadenilação correspondente no gene de α-globina nos pacientes com α-talassemia) documentam a importância deste sinal para a clivagem apropriada em 3' e a poliadenilação (ver Cap. 11). A região 3' não-traduzida de alguns genes pode ser bem longa, de até vários kb. Outros genes têm vários sítios de poliadenilação alternativa, e a seleção entre eles pode influenciar a estabilidade do mRNA resultante e, portanto, o estado de equilíbrio de cada mRNA.

ESTRUTURA DOS CROMOSSOMOS HUMANOS

A composição de genes no genoma humano, bem como os determinantes de sua expressão, é especificada no DNA dos 46

cromossomos humanos. Como vimos na seção anterior, *cada cromossomo é tido como consistindo em uma única dupla hélice contínua de DNA*. Isto é, cada cromossomo no núcleo é uma longa molécula bifilamentar e linear de DNA. Os cromossomos não são duplas hélices nuas de DNA. A molécula de DNA de um cromossomo existe como um complexo, com uma família de proteínas cromossômicas básicas chamadas histonas e com um grupo heterogêneo de proteínas não-histônicas ácidas que não são tão bem caracterizadas, mas que parecem ser cruciais para o estabelecimento de um ambiente apropriado para garantir o comportamento normal dos cromossomos e a expressão apropriada dos genes. Em conjunto, este complexo de DNA e proteínas, é chamado de **cromatina.**

Existem cinco tipos principais de histonas que têm um papel crucial no acondicionamento da fibra de cromatina. Duas cópias de cada uma das quatro histonas cerne H2A, H2B, H3 e H4 constituem um octâmero, ao redor do qual se enrola um segmento da dupla hélice de DNA (Fig. 3.11). Cerca de 140 pares de base do DNA estão associados a cada cerne de histona, dando quase duas voltas ao redor do octâmero. Depois de um curto segmento (de 20 a 60 pares de bases) "espaçador" de DNA, forma-se um novo complexo de DNA, e assim por diante, dando à cromatina um aspecto de contas em um colar. Cada complexo de DNA com histonas cerne é chamado de um **nucleossomo**, que é a unidade estrutural básica da cromatina. A quinta histona, a H1, parece ligar-se ao DNA em seguida a cada nucleossomo, na região espaçadora internucleossômica. A quantidade de DNA associada ao cerne de nucleossomo, juntamente com a região espaçadora, tem cerca de 200 pares de bases.

Durante o ciclo celular, como vimos no Cap. 2, os cromossomos passam por etapas ordenadas de condensação e descondensação (ver Fig. 2.5). No núcleo interfásico, os cromossomos e a cromatina estão bem descondensados em relação ao estado altamente condensado da cromatina na metáfase. Entretanto, mesmo nos cromossomos interfásicos, o DNA na cromatina está substancialmente mais condensado do que estaria como uma dupla hélice nativa, livre de proteínas.

Os longos filamentos de nucleossomos estão eles próprios mais compactados em uma estrutura secundária helicoidal de cromatina do que parece ao microscópio eletrônico, como uma fibra espessa de 30 nm (aproximadamente três vezes mais grossa que a fibra nucleossômica) (ver Fig. 3.11). Esta fibra "solenóide" (do grego *solenoeides*, "em forma de tubo") cilíndrica parece ser a unidade fundamental da organização da cromatina. Os solenóides são compactados em **alças** ou domínios ligados a intervalos de cerca de 100 kb a um **arcabouço** não-histônico de proteína ou matriz. Tem-se especulado que as alças são, de fato, unidades funcionais de replicação do DNA ou de transcrição gênica, ou ambos, e que os pontos de ligação de cada alça são fixados ao longo do DNA cromossômico. Assim, um nível de controle da expressão gênica pode depender de como o DNA e os genes são acondicionados nos cromossomos e de sua associação a proteínas da cromatina no processo de acondicionamento.

Os vários níveis hierárquicos de acondicionamento vistos em um cromossomo interfásico são ilustrados esquematicamente na Fig. 3.11. A enorme quantidade de DNA embalada em um cromossomo pode ser apreciada quando os cromossomos são tratados para remover a maioria das proteínas da cromatina a fim de se observar o arcabouço proteico (Fig. 3.12). Quando o DNA é liberado pelos cromossomos tratados deste modo, longas alças de DNA podem ser vistas, e o arcabouço residual pode ser visto reproduzindo uma estrutura de um típico cromossomo metafásico.

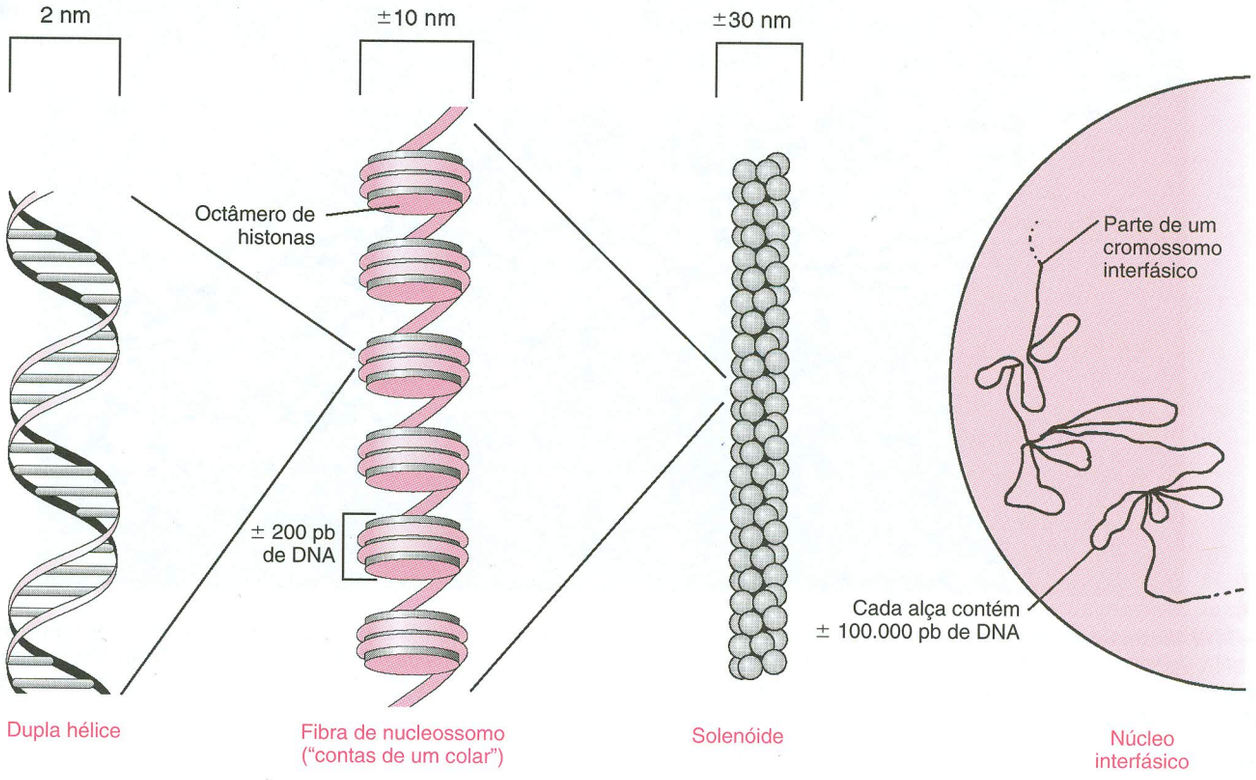

Fig. 3.11 Níveis hierárquicos de acondicionamento da cromatina em um cromossomo humano.