Revisitando o Dogma Central: a relação entre genes e proteínas

Felipe Tadeu Galante Rocha de Vasconcelos[1]*, Igor Neves Barbosa[1]*, Laura Machado Lara Carvalho[1]*, Lucas Santos e Souza[1]*, Ana Cristina Victorino Krepischi[2]

[1] Pós-graduando do Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, SP

* Felipe, Igor, Laura e Lucas contribuíram igualmente na produção do texto.

[2] Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, SP

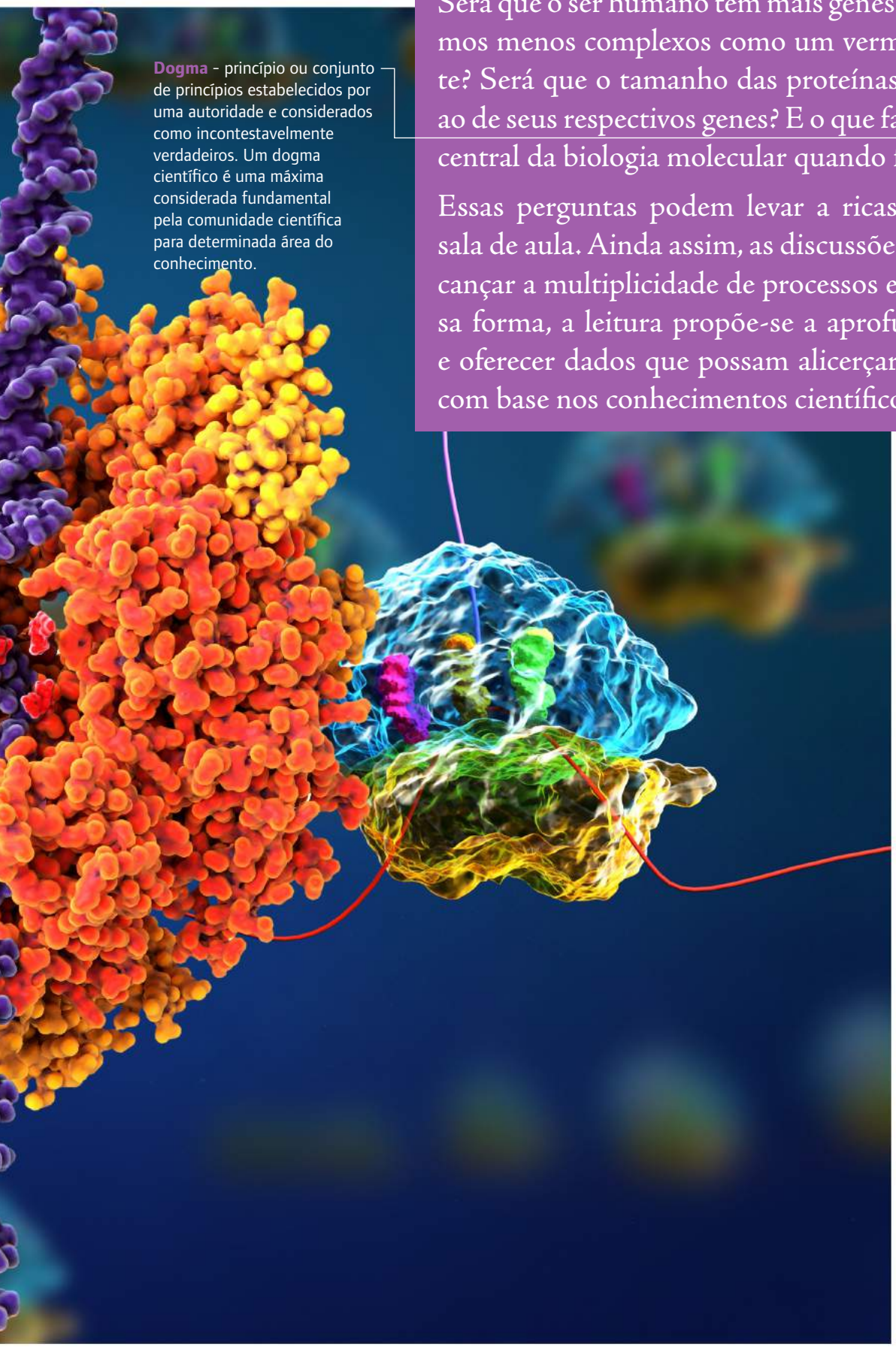
Autor para correspondência - ana.krepischi@ib.usp.br

Palavras-chave: dogma central, genes, proteínas, transcrição, tradução, *splicing*

Será que o ser humano tem mais genes do que organismos menos complexos como um verme ou um tomate? Será que o tamanho das proteínas é proporcional ao de seus respectivos genes? E o que faltava ao **dogma** central da biologia molecular quando foi proposto?

Essas perguntas podem levar a ricas discussões em sala de aula. Ainda assim, as discussões podem não alcançar a multiplicidade de processos envolvidos. Dessa forma, a leitura propõe-se a aprofundar conceitos e oferecer dados que possam alicerçar tais discussões com base nos conhecimentos científicos atuais.

Dogma - princípio ou conjunto de princípios estabelecidos por uma autoridade e considerados como incontestavelmente verdadeiros. Um dogma científico é uma máxima considerada fundamental pela comunidade científica para determinada área do conhecimento.



A complexidade de processos de transmissão da informação genética e o dogma central da biologia molecular

Em 1957, em uma palestra na Universidade de College London intitulada *síntese proteica*, Francis Crick, o mesmo cientista que propôs o modelo da dupla hélice de DNA, juntamente com James Watson, fez uma famosa suposição baseada nas informações disponíveis à época de que, uma vez que a informação genética é traduzida no processo de formação de proteína, não pode retornar aos níveis de DNA, RNA ou uma nova proteína.

Anos mais tarde (1965), na primeira edição do livro Biologia Molecular do Gene, James Watson propôs que a síntese proteica poderia ser representada pelo esquema DNA→RNA→proteína (DNA determina a síntese de RNA que determina a síntese de proteína) (Figura 1). Esse processo passou a ser tratado então pela comunidade científica como o dogma central da biologia molecular, termo proposto por Crick. Cabe frisar que esse modelo era limitado ao conhecimento científico disponível à época.

Figura 1.
Dogma central da biologia molecular, amplamente adotado pela comunidade científica há muitos anos. Nessa concepção, o DNA serve como molde para sua replicação e também para a síntese de RNA (transcrição) que, por sua vez, contém as informações necessárias para direcionar a síntese de proteína (tradução).

Naquela época, as proteínas eram as moléculas mais conhecidas da biologia molecular por algumas de suas funções **estruturais**, enzimáticas e **reguladoras** nos organismos, enquanto o DNA era visto como uma molécula que contém somente informações para a produção das proteínas. Também se suspeitava que o RNA poderia ser um intermediário entre DNA e proteína. Hoje sabemos que o RNA mensageiro é o intermediário responsável pela transmissão da informação genética à síntese proteica. Embora este seja, de fato, o fluxo geral do processo, é apenas uma parte do todo.

Atualmente, sabe-se, por exemplo, que certos tipos de vírus são capazes de sintetizar DNA a partir de RNA, em um processo conhecido como transcrição reversa, o que quebra o **paradigma** da unidirecionalidade proposta como dogma central. A transcrição reversa ocorre por ação da enzima transcriptase reversa dos **retrovírus**, como o **HIV**. O DNA sintetizado a partir da ação dessa enzima sobre o genoma viral contém os genes virais, que serão integrados ao genoma de DNA da célula hospedeira, permitindo a multiplicação viral (Figura 2).

Proteínas estruturais - proteínas estruturais são as que sustentam a estrutura dos tecidos. Exemplos: queratina (presente na pele, unhas e cabelos) e colágeno (responsável pela integridade das cartilagens, da pele e vasos sanguíneos).

Paradigma - modelo, exemplo típico ou conceito bem consolidado para uma comunidade (no caso a científica). Diz-se que houve a quebra de um paradigma científico quando há, a partir de novas evidências, o rompimento de um modelo ou conceito amplamente aceito.

Retrovírus - grupo de vírus que tem genoma composto por RNA que, por sua vez, serve como molde para produção de DNA pela enzima transcriptase reversa.

Reguladoras - proteínas reguladoras ajudam a regular atividades no organismo. Entre elas estão, por exemplo, hormônios e fatores de transcrição (estes atuam ativando ou reprimindo a transcrição de genes).

HIV - Vírus da imunodeficiência humana. Infecta células do sistema imunológico (de defesa), o que faz com que os indivíduos não tratados tornem-se mais suscetíveis a outras infecções.

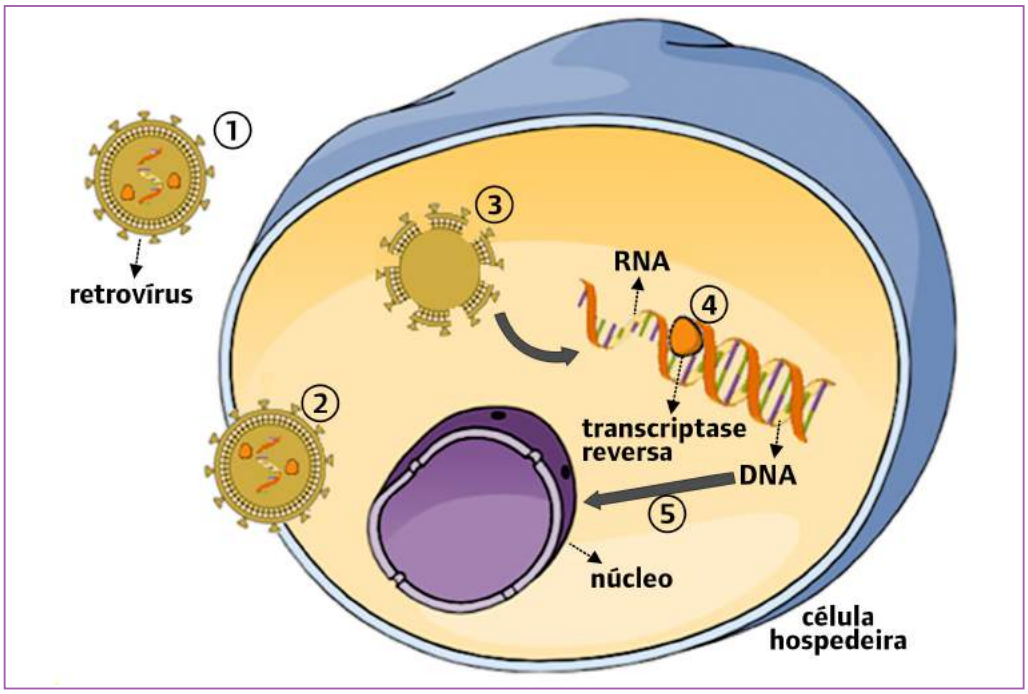


Figura 2.
Representação esquemática da transcrição reversa. A transcriptase reversa - enzima presente em retrovírus - é capaz de utilizar o RNA do genoma viral como molde para a síntese de DNA dentro da célula hospedeira. Nesta figura, são representadas as etapas: (**1** e **2**) entrada do retrovírus na célula hospedeira; (**3**) liberação do material genético e da enzima transcriptase reversa no citoplasma; (**4**) síntese de DNA a partir do RNA viral; (**5**) integração do DNA produzido ao genoma da célula hospedeira.

Alguns outros vírus, como os **rotavírus**, são capazes de realizar a síntese de RNAm a partir de dsRNA (do inglês *double-stranded RNA*; RNA dupla fita, que é o genoma deste tipo de vírus), isto é, produz-se RNA sem que um DNA seja utilizado como molde, pois eles têm uma enzima especial, a **polimerase do RNA** dependente de RNA (nos rotavírus chamada de VP1 - proteína viral 1) (Figura 3). Já os **coronavírus**, como Sars-CoV-2, que causa Covid-19, têm genoma baseado em RNA de uma fita (**+ssRNA** - fita simples de sentido positivo), porém durante a interação com o hospedeiro há a produção de intermediários de replicação viral baseados em dsRNA.

+ssRNA - material genético constituído por RNA de cadeia simples e sentido positivo. Sentido positivo significa que o RNA genômico atua diretamente como RNA mensageiro e é traduzido pelos ribossomos da célula hospedeira. Já os vírus de RNA de sentido negativo (-ssRNA) não podem ter seu genoma diretamente traduzido, pois o RNA genômico é complementar ao RNA que é traduzido. Os vírus +ssRNA são os mais abundantes do planeta (exemplos: vírus da hepatite C, da dengue, MERS , SARS-CoV-2 e os rinovírus - que causam o resfriado comum).

Rotavírus - vírus de transmissão fecal-oral que causa gastroenterite, sendo umas das principais causas de diarreia grave em crianças. Uma especificidade desse tipo de vírus é seu material genético constituído por RNA de dupla fita (dsRNA).

Polimerase do RNA - é uma enzima que atua na síntese de RNA a partir de um molde de DNA.

Coronavírus - família diversa de vírus com genoma baseado em RNA fita simples da qual faz parte o SARS-CoV-2 – vírus que causou a pandemia de COVID-19 a partir de 2019.

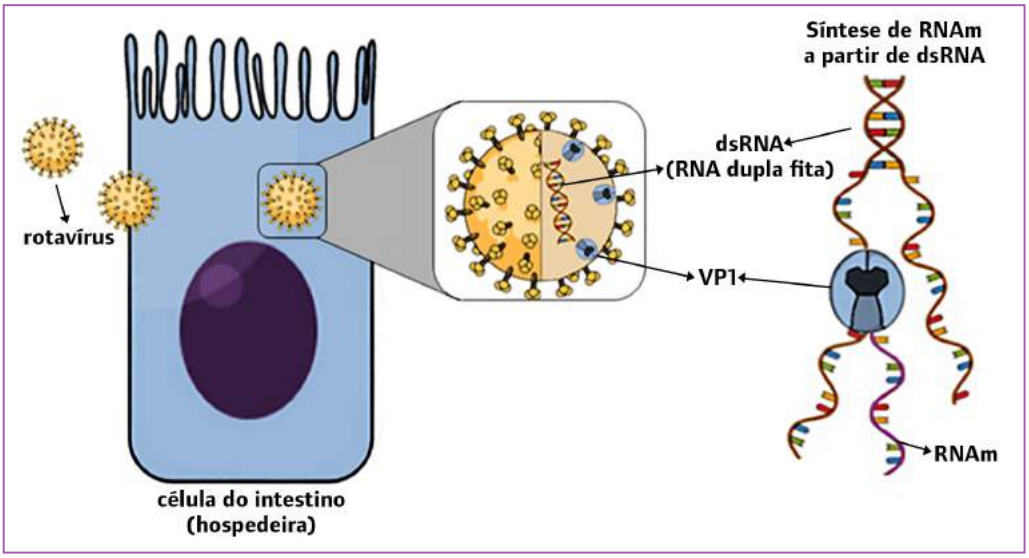


Figura 3.
Transcrição a partir de dsRNA em rotavírus. Os rotavírus utilizam uma polimerase de RNA dependente de RNA (VP1 – proteína viral 1) para produzir RNA sem que DNA seja utilizado como molde. Perceber que o molde utilizado é o dsRNA presente no genoma viral.

Percebe-se, portanto, que apesar de os vírus serem estruturalmente menos complexos que outros organismos, há uma multiplicidade de processos possíveis para que ocorra a replicação de seu genoma e síntese de suas proteínas, a depender do tipo de vírus.

Também já foi observada a tradução direta do DNA para proteína, sem um RNA intermediário, porém em ambiente artificial (*in vitro*), usando ribossomos de extratos de bactérias. Vale ressaltar que esse fenômeno nunca foi observado espontaneamente na natureza. Contudo, o armazenamento e a transmissão da informação para produção de proteínas não são as únicas funções das moléculas de DNA e RNA. O DNA também pode servir de molde para a transcrição de moléculas funcionais de RNA, que não são traduzidas em proteínas. Atualmente, divide-se os RNAs em RNAm (RNAs mensageiros) e RNAs não codificadores (RNAnc - do inglês *non codingRNAs*) (Figura 4). Os RNAm são os que contêm a sequência codificadora para síntese de proteínas e já tinham sido previstos no modelo unidirecional do dogma central da biologia molecular. Os RNAnc são divididos em várias classes, descritas na Tabela 1.

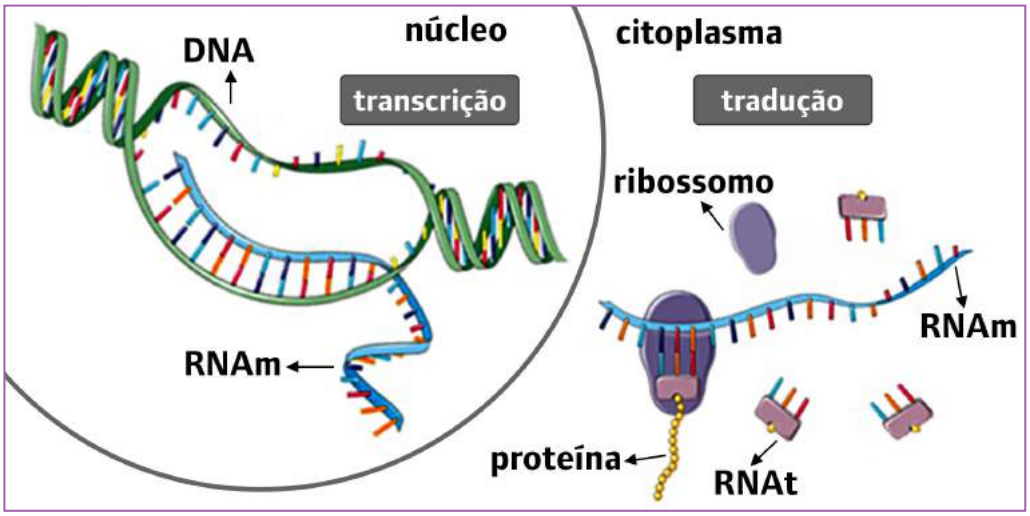


Figura 4.
Transcrição e tradução em eucariotos e algumas classes de RNA. Na transcrição, é produzido o RNAm a partir de informações contidas no DNA. Na tradução, o RNAm orienta a síntese de proteínas. Note a presença do RNAt (RNA transportador) e dos ribossomos, que são formados por RNAr (RNA ribossômico). O RNAt e o RNAr são RNAs funcionais, não codificadores de proteínas, isto é, não contêm a sequência de códons para produção da cadeia polipeptídica/proteína, mas têm papéis importantes no processo de tradução. Cabe dizer também que, em procariotos, ambos os processos (transcrição e tradução) ocorrem no citoplasma, pois não há membrana nuclear.

CLASSE	FUNÇÃO
RNAs transportadores (RNAt) → ~ 70-80 **nt**	Transportam os aminoácidos no processo de tradução
RNAs ribossômicos (RNAr) → ~ 120-5000 nt	Compõem os ribossomos
Pequenos RNAs nucleares (RNAsn) → ~ 60-360 nt	Exclusivos de eucariotos, atuam no ***splicing*** de RNAm, na manutenção dos telômeros e interagem com **fatores de transcrição**
MicroRNAs (RNAmi) e pequenos RNAs de interferência (RNAsi) → ~ 21-22 nt	Interagem com RNAm, regulando negativamente a expressão gênica
RNAs de interação piwi (RNApi) → ~ 24-31 nt	Impedem a dispersão de **elementos transponíveis** para outros *loci*
RNAs não codificadores longos (RNAlnc) → > 200 nt	Transcritos com mais de 200 nucleotídeos envolvidos na regulação da expressão gênica

Tabela 1.
Classes de RNAnc e suas respectivas funções.

nt - nucleotídeos.

Splicing - processo de maturação do pré-RNAm em RNAm por meio do qual ocorre a retirada de íntrons e junção de éxons.

Fatores de transcrição - proteínas importantes para o início e para o controle (reprimindo ou impulsionando) da taxa de transcrição.

Elementos transponíveis - trechos de DNA que se movem dentro do genoma, podendo por exemplo levar consigo genes, promover rearranjos cromossômicos e alterar a expressão de genes vizinhos, além de propiciar aumento da variabilidade genética.

Um exemplo clássico de RNA não codificador é o Xist (X-*inactive specific transcript*) que faz parte da classe dos RNAlnc. Este RNA está envolvido no silenciamento do cromossomo X em fêmeas de mamíferos placentários (Figura 5). Em fêmeas, o processo de inativação silencia transcricionalmente um dos cromossomos X do par, exceto por alguns genes que permanecem ativos, especialmente aqueles mapeados nas chamadas regiões **pseudoautossômicas**. O mecanismo de inativação permite uma equivalência de dosagem entre machos e fêmeas para a expressão da maioria dos genes do cromossomo X.

Pseudoautossômicas - são sequências homólogas entre os cromossomos sexuais (X e Y), localizadas nas extremidades de seus braços curtos e longos - chamadas de PAR1 e PAR2. Tais regiões estão presentes unicamente nos cromossomos X e Y, que pareiam entre si durante a meiose, permitindo a ocorrência de recombinação.

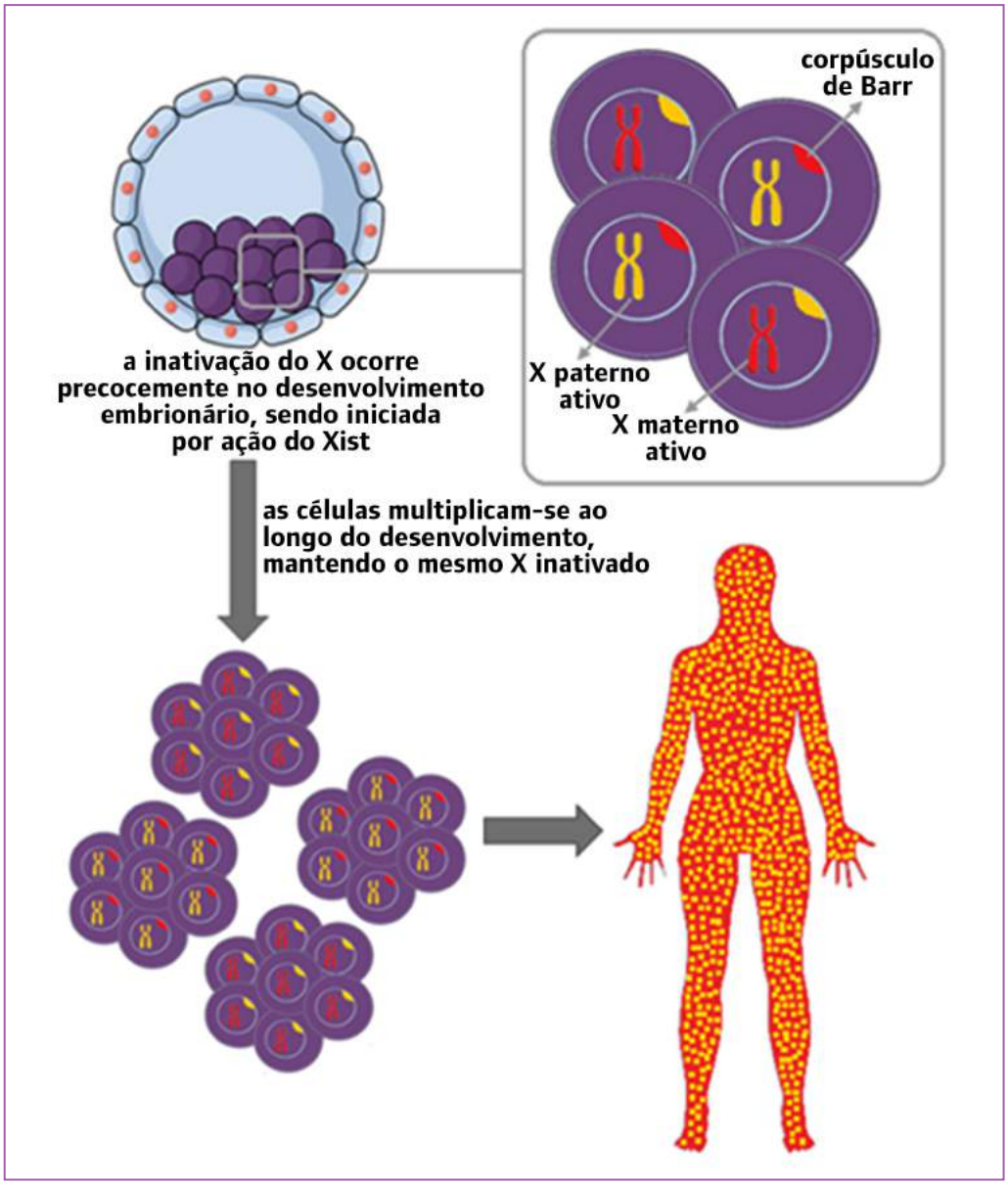


Figura 5.
Esquema da inativação do cromossomo X. O RNA não codificador Xist atua no silenciamento do cromossomo X em fêmeas. Ao observar as células, microscopicamente, é possível visualizar corpúsculo de Barr na periferia do núcleo, que corresponde ao X inativo. O processo de inativação é aleatório, de maneira que algumas células têm o X materno inativo e, as demais, o paterno.

A escolha do cromossomo X a ser inativado é aleatória, de forma que algumas células terão o X paterno ativo e, outras, o X materno. Ao observar as células ao microscópio, é possível visualizar o X inativo na periferia do núcleo, como uma estrutura condensada à qual se dá o nome de corpúsculo de Barr. Após o processo de inativação no início do desenvolvimento embrionário, há a manutenção da inativação do mesmo cromossomo X nas células-filhas. Dessa forma, uma fêmea de mamíferos é um mosaico quanto à expressão de genes do cromossomo X (algumas células do organismo têm X materno ativo e, outras, o X paterno).

Tendo claro que as instruções contidas no DNA não se limitam à função de produção proteica, atualmente o conceito de gene não abrange apenas sequências codificadoras, mas também as que orientam a síntese de RNAs funcionais (RNAnc).

Em eucariotos, o tamanho de um gene não é rigidamente proporcional ao tamanho da proteína

A distrofina e a titina são proteínas presentes principalmente nos músculos. O gene da distrofina (*DMD*) é um dos maiores da nossa espécie, com aproximadamente 2,2 Mb, isto é, 2.200.000 pb (dois milhões e duzentos mil pares de bases), o que corresponde a cerca de 0,07% do genoma humano. Por ser tão grande, estima-se que o processo de transcrição requer 16 horas. Por outro lado, o gene da titina (*TTN*), embora seja menor (~281,4 Kb), produz uma proteína maior (35.991 aminoácidos) do que a proteína distrofina (de 3.685 aminoácidos).

Kb - quilobases (mil pares de bases).

Por que essa discrepância na relação entre o tamanho do gene e o tamanho da proteína? Para entender isso, é preciso comparar as regiões codificadoras (CDS do inglês *coding sequence*) que dão origem a cada uma dessas proteínas. Mas o que é região codificadora? Para compreender esse conceito, vamos precisar entender o *splicing* do RNAm (Figura 6).

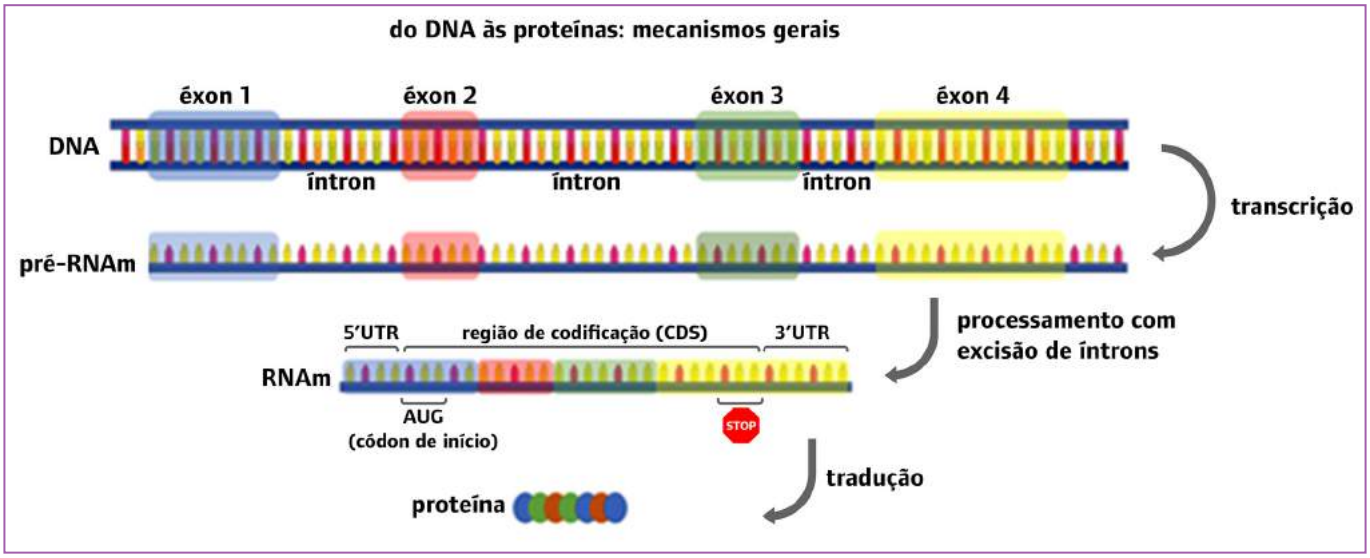


Figura 6.
Transcrição, processamento de RNAm e tradução. Após a retirada dos íntrons, em um processo denominado *splicing*, o RNAm passa a ser formado pelas regiões exônicas. As extremidades do RNAm maduro (denominadas regiões UTR, do inglês *unstranslated regions*) não são codificadoras de proteína. A sequência que orienta a construção da cadeia peptídica no processo de tradução está contida na região de codificação (CDS).

Antes de se tornarem RNAm, os transcritos de genes codificadores de proteínas são chamados transcritos primários ou pré-RNAm. Um dos principais eventos do processamento de pré-RNAm é a excisão de íntrons e união dos éxons, porém, nem todo o RNAm resultante desse processamento é traduzido em proteína. O início da tradução se dá pela presença de uma trinca de nucleotídeos com sequência AUG, que codifica o aminoácido metionina. O segmento de RNAm anterior ao primeiro AUG é chamado de 5'UTR (do inglês *untranslated region* - região não traduzida). Também há uma região não traduzida na extremidade 3', a chamada 3'UTR, que corresponde ao segmento posterior ao códon de parada (UAA, UAG ou UGA). No RNAm, a região codificadora fica entre as regiões UTR. Cabe dizer ainda que as regiões UTR podem envolver apenas o éxon inicial e o final, mas podem envolver também mais éxons, a depender da localização do primeiro AUG e do códon de parada.

Tendo entendido o conceito de região codificadora, comparando o tamanho dessa região nos transcritos da distrofina e da titina (Tabela 2), percebe-se que, apesar de o gene da distrofina ser muito maior que o da titina (2,2 Mb contra 281Kb), a região codificadora da titina é quase dez vezes maior que o da distrofina e isso significa que apenas uma pequena porcentagem da sequência do gene da distrofina é codificadora. O gene da distrofina tem muitos íntrons e esses introns são muito grandes.

Tabela 2.
Informações sobre tamanhos da sequência do gene no DNA, da região codificadora e da proteína da distrofina, titina e insulina humanas. CDS – região codificadora; Mb – megabases (um milhão de pares de bases); Kb – quilobases (mil pares de bases); nt – nucleotídeos; aa – aminoácidos.

	Gene	CDS	Proteína
DMD distrofina	2,2 Mb	11.058 nt	3.685 aa
TTN titina	281,4 Kb	107.976 nt	35.991 aa
INS insulina	1,4 Kb	333 nt	51 aa

Outro evento que pode alterar consideravelmente o tamanho de uma proteína são alterações pós-traducionais. A proteína insulina madura, por exemplo, passa por essas alterações (Figura 7). A região codificadora do gene da insulina leva à produção de uma cadeia peptídica de 110 aminoácidos, correspondente à pré-pró-insulina. Durante o processo de maturação, ela sofre duas quebras promovidas por enzimas específicas, que levam à retirada do peptídeo sinal e da cadeia C. Dessa forma, a insulina madura tem apenas 51 aminoácidos, correspondentes às cadeias A e B.

Alterações pós-traducionais - eventos de processamento que mudam as propriedades das proteínas por clivagem (quebra) ou por adição de um grupo químico a um ou mais aminoácidos.

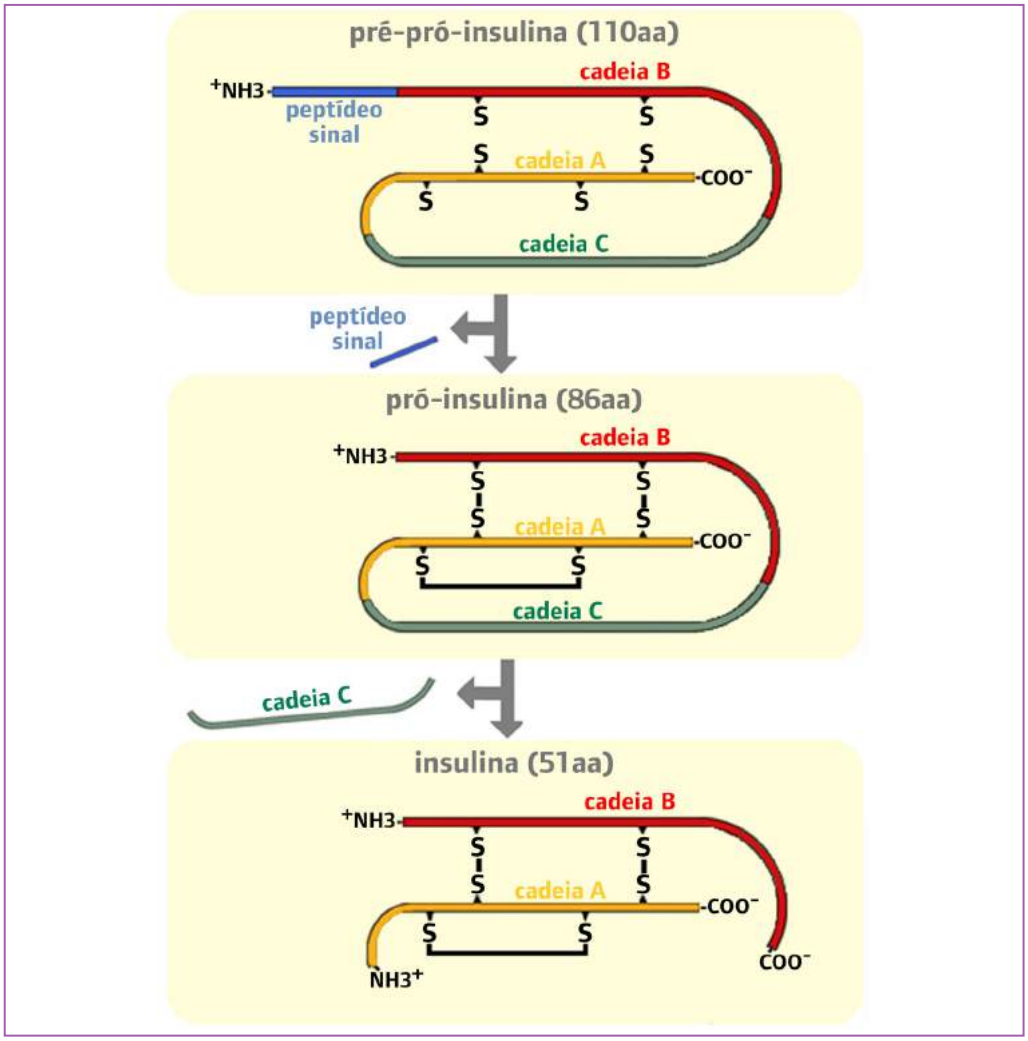


Figura 7.
Modificações pós-traducionais para produção da insulina. Perceba que é inicialmente traduzida a cadeia polipeptídica chamada de pré-pró-insulina, com 110 aminoácidos, mas ocorrem modificações que retiram trechos dessa cadeia (primeiro o peptídeo sinal e depois a cadeia C), resultando no hormônio insulina, de apenas 51 aminoácidos.

A insulina é responsável pela redução da **glicemia**, ao promover a entrada de glicose nas células. Hoje se sabe também que o peptídeo C cumpre algumas funções fisiológicas, tais quais: é um anti-inflamatório, **antioxidante**, **antiapoptótico**, tem atuação na melhora do fluxo sanguíneo e na **bomba de sódio e potássio**. Assim sendo, alterações pós-traducionais na pré-pró-insulina levam à formação de duas proteínas importantes menores que a cadeia polipeptídica precursora: a insulina e o peptídeo C.

Glicemia - concentração de glicose no sangue.

Antioxidante - um antioxidante é uma molécula capaz de inibir a oxidação de outras moléculas. A oxidação é um tipo de reação na qual ocorre a perda de elétrons de uma determinada molécula. Embora as reações de oxidação sejam importantes em algumas vias biológicas, elas têm potencial de produzir radicais livres que, por sua vez, são moléculas muito reativas que em excesso podem ser danosas ao organismo.

Antiapoptótico - a apoptose é um tipo de morte celular programada, um "suicídio celular". Como consequência evita que células com problemas comprometam o funcionamento adequado do organismo. Mecanismos apoptóticos favorecem a entrada em apoptose, enquanto mecanismos antiapoptóticos a evitam.

O número de genes não é proporcional à complexidade do organismo

O Projeto do Genoma Humano foi um empreendimento internacional que aconteceu entre 1990 e 2003 e contou com pesquisadores de 18 países. Tinha como objetivo determinar a sequência nucleotídica completa do genoma humano, além do mapeamento e identificação de seus genes. No início do projeto, estimava-se que o genoma humano teria cerca de 100.000 (cem mil) genes. O raciocínio era baseado em uma ideia simples: cada gene produz uma proteína e quanto mais genes, maior a complexidade do organismo, por isso esperavam que o ser humano tivesse mais genes que outros organismos para os quais o genoma era razoavelmente conhecido, no entanto, era equivocado esse raciocínio.

Bomba de sódio e potássio - proteína localizada na membrana plasmática cuja atividade utiliza a energia proveniente da degradação do ATP (adenosina trifosfato) em ADP (adenosina difosfato) para transportar íons de sódio e potássio entre os ambientes intracelular e extracelular, sempre contra o gradiente de concentração, isto é, do ambiente de menor para o de maior concentração.

Surpreendentemente, chegou-se a um número total de genes do genoma humano muito inferior a 100.000, o que estava abaixo das expectativas. De acordo com dados atuais do ***Ensembl***, o genoma humano tem cerca de 20.500 genes que codificam proteínas e outros 24.000 genes não codificadores, ou seja, que contêm instruções para síntese de RNAs funcionais.

Ensembl - é um *Genome Browser* - isto é, um navegador - que permite o acesso a inúmeros bancos de dados, disponíveis *online* com informações genômicas de diversas espécies. O *Ensembl* original (lançado em 1999 em resposta à iminente conclusão do Projeto do Genoma Humano) concentrava-se apenas em genomas de vertebrados. Desde 2009, há também portais específicos *online* do *Ensembl* para *metazoa*, plantas, fungos, bactérias e protistas.

De acordo com dados atuais do ***UniProt***, nossa espécie tem pouco mais de 75.500 proteínas. Mas por que temos mais proteínas do que genes codificadores de proteínas em humanos? Há mais de um motivo, na verdade. O primeiro deles são as alterações pós-traducionais, das quais já tratamos aqui. Voltando ao exemplo da insulina, perceba que após a tradução ocorrem modificações na cadeia polipeptídica inicial e o resultado é a produção de duas proteínas derivadas: a insulina e o peptídeo C.

UniProt - banco de dados disponível *online* com informações de sequências de aminoácidos nas proteínas e suas funções.

Calcitonina - produzida pela tireoide, é um hormônio peptídico que diminui a concentração de cálcio no sangue e aumenta sua fixação nos ossos.

Catacalcina - peptídeo produzido na tireoide que atua na redução de cálcio no sangue.

Outro mecanismo que contribui para termos mais proteínas que genes codificadores de proteínas em humanos é um evento de nome *splicing* alternativo, que já era conhecido desde o início da década de 1980. Para entender o mecanismo de *splicing* alternativo, é necessária a explicação de como é a expressão da **calcitonina**, da **catacalcina** e do **CGRP** a partir do gene *CALCA* (Figura 8).

CGRP - peptídeo produzido por neurônios e tem diversas funções: é vasodilatador (dilatador de vasos sanguíneos), atua na transmissão de sinais de dor ao cérebro e na regeneração do tecido nervoso após lesão. A sigla significa peptídeo relacionado ao gene da calcitonina (do inglês, *calcitonin gene related peptide*).

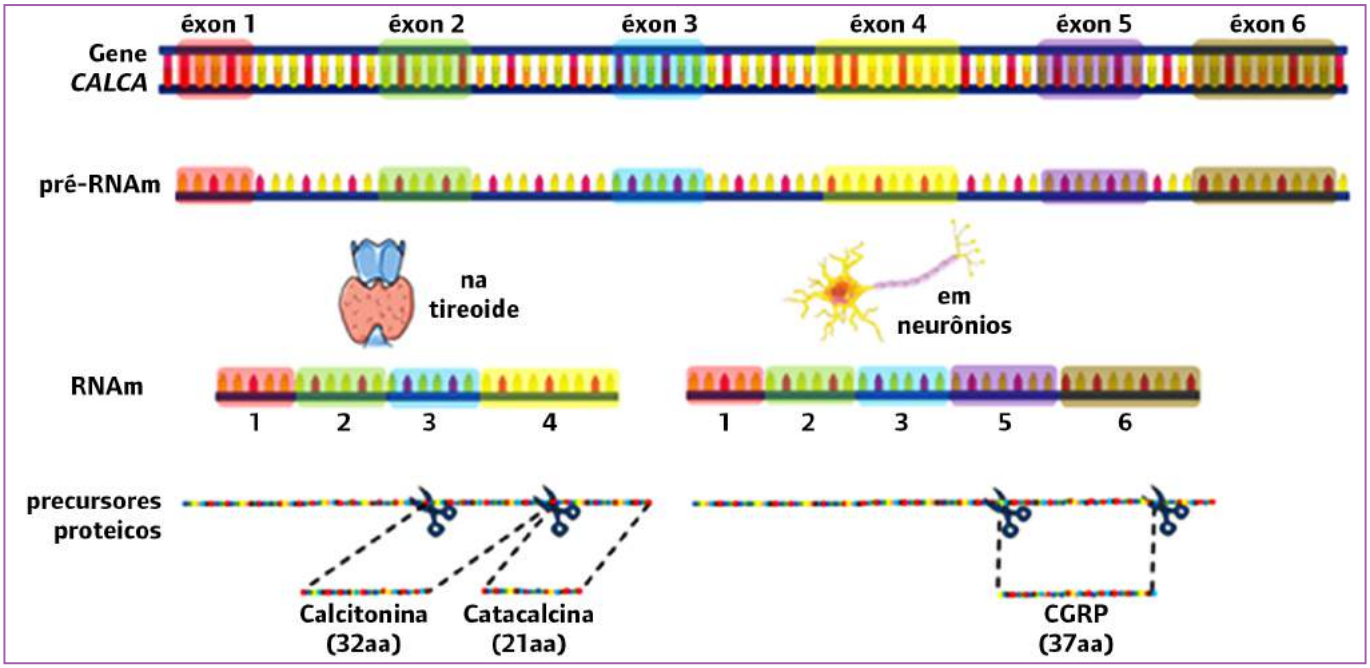


Figura 8.
A expressão do gene *CALCA* envolve eventos de *splicing* alternativo tecido-específico e modificações pós-traducionais, resultando em três produtos proteicos diferentes. O *splicing* na tireoide não ocorre da mesma maneira que em neurônios. Perceba as diferenças no conteúdo exônico do mRNA da tireoide e de neurônios. Além disso, o precursor proteico traduzido na tireoide passa por alterações pós-traducionais – representadas pelas tesouras - e resulta em dois produtos (calcitonina e catacalcina), enquanto o de neurônios forma apenas o CGRP após alterações pós-traducionais.

O gene *CALCA* tem seis éxons. A transcrição deste gene na tireoide, seguida de processamento com excisão de íntrons (*splicing*), resulta em um RNAm maduro, contendo apenas os éxons 1, 2, 3 e 4. Já nos neurônios, o RNAm maduro contém os éxons 1, 2, 3, 5 e 6. A esse fenômeno dá-se o nome de *splicing* alternativo, em que RNAs com conteúdos exônicos diferentes são formados (transcritos diferentes). O mecanismo ocorre durante o processamento de pré-RNAm em RNAm maduro.

No caso específico do gene *CALCA*, o RNAm formado na tireoide dá origem a uma cadeia polipeptídica precursora que sofre uma série de alterações pós-traducionais e origina duas proteínas: a calcitonina e a catacalcina. Já o RNAm formado em neurônios dá origem a uma cadeia polipeptídica que também sofre alterações pós-traducionais, produzindo apenas o CGRP, ou seja, a partir do gene *CALCA* há a síntese de três produtos proteicos graças a dois mecanismos aqui tratados: *splicing* alternativo e alterações pós-traducionais.

O *splicing* alternativo permite que um único pré-RNAm tenha diversas possibilidades de *splicing*, aumentando consideravelmente o número possível de produtos proteicos.

Entretanto, esse mecanismo de *splicing* e as alterações pós-traducionais não são os únicos a contribuir para que haja uma enorme diversidade de proteínas sintetizadas por cada genoma eucarioto, com número que pode ser inclusive superior ao de genes, como em humanos. Outro mecanismo importante é a existência de sítios alternativos de início da transcrição. Nos genomas de humanos, por exemplo, mais de 50% dos genes têm sítios alternativos de início da transcrição, sendo em média quatro por gene.

Mas, como esse mecanismo funciona? Para entendermos, é preciso saber que a transcrição de um gene em organismos eucariotos é altamente controlada. Cada gene possui regiões regulatórias, nas quais estão presentes sequências específicas que controlam a transcrição. Uma dessas sequências é a região promotora que, como o nome já diz, ajuda a promover a transcrição do gene, funcionando como um local de montagem de um complexo proteico (formado por polimerase do RNA e fatores de transcrição). O sítio de início da transcrição corresponde ao primeiro nucleotídeo transcrito (comumente chamado de +1) (Figura 9).

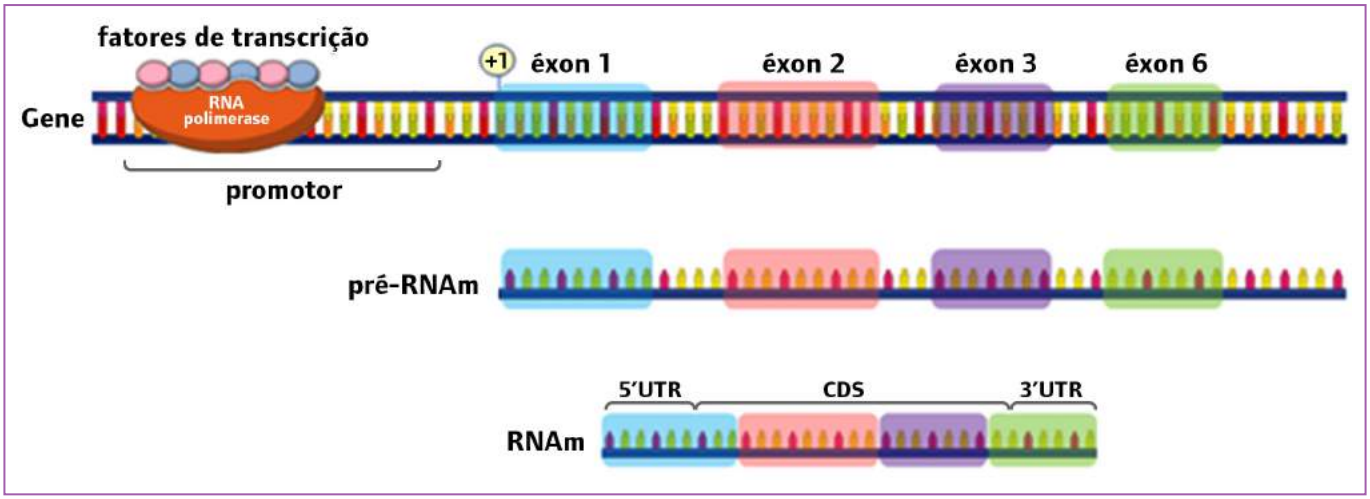


Figura 9.
Estrutura gênica com sítio de início da transcrição (+1), que é o primeiro nucleotídeo transcrito. A polimerase do RNA reconhece a região promotora do gene, mas só inicia sua transcrição a partir do sítio +1.

A depender da localização do promotor (um mesmo gene pode ter mais de um promotor), a posição do sítio +1 pode variar, originando diferentes transcritos por gene (Figura 10).

Assim sendo, a chave da complexidade humana provavelmente não está no número de genes, mas em como as sequências gênicas são usadas como módulos para construir diferentes produtos. Mecanismos como alterações pós-traducionais, processos alternativos de *splicing* e múltiplos sítios de início de transcrição contribuem para a ampla gama de possibilidades na produção de proteínas.

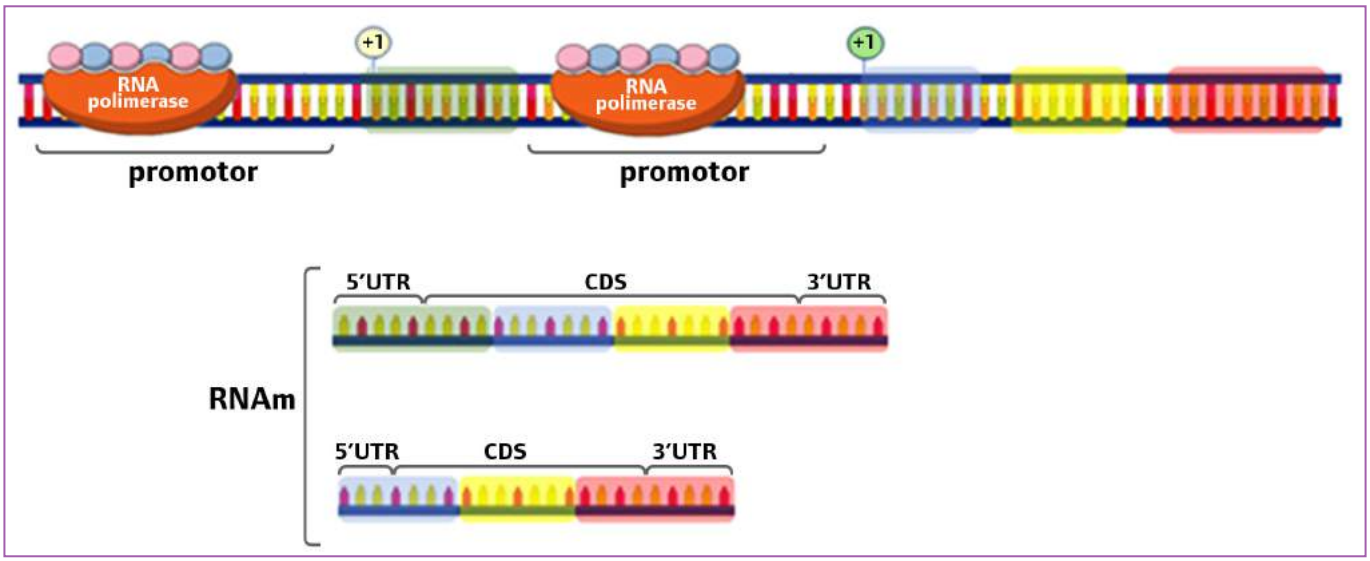


Figura 10.
Sítios alternativos de início da transcrição. Perceber que o início da transcrição pode variar em um mesmo gene, gerando RNAs mensageiros maduros diferentes e, consequentemente, proteínas diferentes.

Dessa forma, não é tão surpreendente que, de acordo com dados atuais do *Ensembl*, o *zebrafish* (*Danio rerio* – peixe paulistinha) tenha aproximadamente 5.000 (cinco mil) genes codificadores de proteínas a mais que os humanos, ou que o tomate (*Solanum lycopersicum*) tenha cerca de 14.000 (quatorze mil) genes codificadores de proteína a mais que os humanos. E, também, que o milho apresente cerca de 19.000 (dezenove mil) a mais. Já o verme *Caenorhabditis elegans* tem números próximos de genes codificadores e não codificadores de proteínas em relação aos humanos (Figura 11).

Os dados do gráfico (Figura 11) evidenciam que a complexidade de um organismo, tal qual era entendida, não pode ser diretamente avaliada a partir do número de genes em seu genoma, uma vez que há um amplo repertório de mecanismos reguladores intermediários entre o gene e o produto de sua expressão.

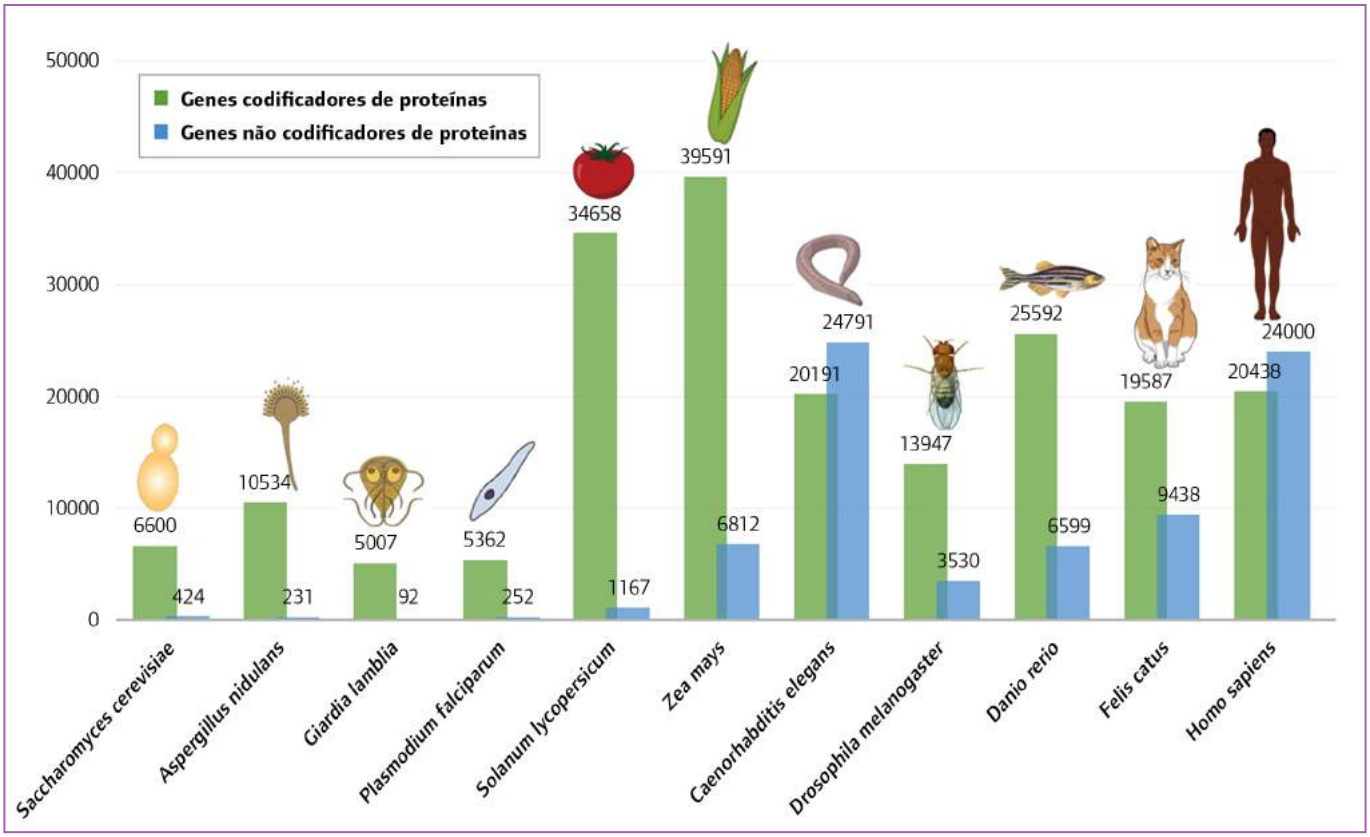


Figura 11.
Número de genes codificadores e não codificadores de proteínas em diferentes espécies de eucariotos. Obs.: gráfico com informações dos bancos de dados *Ensembl*, baseadas no conhecimento científico atual. Consulta em 08/06/2020, que está sujeito a atualizações.

Financiamento: os autores são gratos às agências de fomento CAPES, CNPq e FAPESP.

Ilustrações: As figuras foram produzidas pelos autores utilizando recursos *Mind the Graph* (https://www.mindthegraph.com/) e *Servier Medical Art* (https://smart.servier.com/), respeitando as condições de seus termos de uso.

Para saber mais

STRACHAN, T.; READ, A. *Genética molecular humana*. 4ª edição. Porto Alegre. Artmed, 2013.

MCCARTHY, B. J. HOLLAND, J. J. Denatured DNA as a direct template for in vitro protein synthesis. *Proc. Natl. Acad. Sci* v. 54, n. 3, p. 880-886, 1965.

NCBI - Gene. Disponível em: https://www.ncbi.nlm.nih.gov/gene/

Ensembl genome browser. Disponível em: https://www.ensembl.org/

Ensembl plants. Disponível em: https://plants.ensembl.org/

Ensembl protists. Disponível em: https://protists.ensembl.org/

Ensembl fungi. Disponível em: https://fungi.ensembl.org/

Ensembl metazoa. Disponível em: https://metazoa.ensembl.org/

UniProt. Disponível em: https://www.uniprot.org/