

# Soluzioni del Capitolo 10

**10.1** a. 9; b. 11; c. 10; d. 2; e. 6; f. 7; g. 12; h. 4; i. 5; j. 8; k. 3; l. 1.

**10.2** Alcuni metodi possibili per identificare le sequenze geniche nel genoma umano sono: (1) Mettere a confronto le sequenze genomiche complete di una specie di mammiferi correlata come il topo. Le regioni non funzionali dei genomi si sono significativamente divergenti, mentre le regioni funzionali sono relativamente conservate. (2) Utilizzare l'analisi computazionale per cercare nella sequenza del DNA registri di lettura aperti e potenziali confini introne/esone. (3) Individuare le regioni trascritte confrontando la sequenza genomica con le sequenze di cloni di cDNA copiati da mRNA in vari tessuti.

**10.3** a. L'uomo e gli scimpanzé hanno condiviso per l'ultima volta un antenato comune 6-7 Mya (milioni di anni fa). L'ultimo antenato comune di esseri umani e topi è stato 75 Mya; di esseri umani e cani 92 Mya; di esseri umani e polli, 310 Mya; di esseri umani e rane, 360 Mya.

**b.** L'orologio più lento sono le mutazioni di senso (I). Il motivo è che queste mutazioni modificano gli amminoacidi nelle proteine e un tale cambiamento potrebbe avere conseguenze deleterie sul fenotipo perché sfavorite dalla selezione naturale. Molte mutazioni di senso che si verificano nel corso della storia andrebbero così perse, e quindi le mutazioni di senso si accumulano lentamente nei genomi, il che significa che l'orologio cammina più lentamente. Le mutazioni silenziose non influenzano il fenotipo perché nessun amminoacido risulta modificato; quindi, le mutazioni silenziose sono, fra le tre, quelle che segnano il tempo più velocemente. La maggior parte delle mutazioni negli introni non influenza la funzione genica, sebbene alcune potrebbero influenzare l'efficienza o l'accuratezza dello splicing, quindi una minoranza di esse potrebbe influenzare la fun-

zione genica in modi che sarebbero bersaglio dalla selezione naturale.

**c.** Si noti che 400 Mya è un tempo molto lungo, più indietro dell'ultimo antenato comune di umani e rane. Per le specie che hanno condiviso per l'ultima volta un antenato comune così tanto tempo fa, è necessario usare un orologio lento come le mutazioni di senso (I). Se si utilizzasse un orologio più veloce (II o III), potrebbero essersi accumulate così tante mutazioni da non poter fare un confronto molto accurato, perché spesso potrebbe non trovarsi alcuna relazione fra i genomi delle due specie. Se si fa riferimento alla Figura 10.3, si vede che le uniche sequenze che hanno chiare omologie tra uomo e il pesce zebra (l'ultimo antenato comune risale a ~416 Mya) sono le sequenze codificanti per i geni più altamente conservati.

È interessante notare che se le specie sono più strettamente correlate in termini evolutivi, come esseri umani e scimpanzé, si deve usare un orologio più veloce, come confrontare i cambiamenti negli introni. Se si usano mutazioni di senso, tra questi genomi potrebbero essersi accumulate così poche mutazioni da non poter ottenere una stima molto accurata del tempo di divergenza.

**d.** L'orologio che ha meno probabilità di variare nel tasso di accumulo di mutazioni in geni diversi è quello delle mutazioni chiaramente silenziose (sinonime). L'orologio che ha maggiori probabilità di variare nel tasso di accumulo di mutazioni in geni diversi è quello delle mutazioni di senso (I), perché alcuni geni sono più sensibili di altri alle mutazioni che cambiano gli amminoacidi. (Alcune proteine sono altamente conservate perché la maggior parte delle sostituzioni di amminoacidi inibisce la loro funzione, mentre altre proteine sono scarsamente conservate perché molti cambiamenti nella composizione degli ammi-

noacidi sono ancora coerenti con la funzione delle proteine.) Anche una risposta (II) sarebbe stata ragionevole, perché alcuni cambiamenti negli introni potrebbero influenzare la funzione del gene; tuttavia, la maggior parte dei cambiamenti di introni non ha alcun effetto; infatti, come si osserva nella Figura 10.3, che le sequenze di introni sono generalmente scarsamente conservate.

**10.4** Il primo enzima di cui si ha bisogno è la trascrittasi inversa (g) per copiare l'mRNA nel primo filamento di cDNA. Probabilmente sarebbe necessaria anche la DNA polimerasi (a) per copiare il primo filamento di cDNA per produrre cDNA a doppio filamento. La DNA polimerasi non è stata inclusa in Figura 10.4, perché non è strettamente necessaria dato che la trascrittasi inversa non è solo una DNA polimerasi RNA-dipendente, ma può anche fungere da DNA polimerasi DNA-dipendente (cioè, ha il potenziale per copiare il primo filamento di cDNA per creare cDNA a doppio filamento). Successivamente è necessario inserire i cDNA a doppio filamento nel vettore. Anche questo non è stato mostrato in Figura 10.4, ma il modo più semplice è usare un enzima di restrizione (c). Quindi, si aggiungono oligonucleotidi contenenti il sito di riconoscimento per l'enzima di restrizione alle estremità del cDNA a doppio filamento, e si creano molecole di DNA ricombinante come in Figura 9.5. (Esistono altri metodi più complicati che permetterebbero di inserire il cDNA a doppio filamento nel vettore senza un enzima di restrizione.) Dopo aver mescolato il vettore, tagliato con l'enzima di restrizione, ai frammenti di cDNA a doppio filamento che hanno le stesse estremità coesive è necessario aggiungere la DNA ligasi (d) per saldare la molecola di DNA ricombinante.

**10.5 a.** La sequenza 1 è il frammento genomico; La sequenza 2 è la sequenza di cDNA. Sul frammento genomico scritto sotto, gli esoni sono in rosso e l'introne è in blu:

5' TAGGTGAAAGAGTAGCCTAGAAATCAGTTA 3'  
3' ATCCACTTCTCATCGGATCTAGTCAAT 5'

Questa operazione è possibile perché la sequenza complementare inversa della sequenza 2 è uguale alla sequenza 1 (il filamento superiore), tranne per il fatto che la sequenza 2 manca della sequenza dell'introne (blu).

**b.** Poiché non si conosce la direzione della trascrizione, il trascritto primario potrebbe corrispondere a entrambi i filamenti riportati nella ri-

sposta alla parte (a) sopra. Per decidere tra queste due possibilità, è necessario cercare le sequenze consenso di splicing alle estremità 5' e 3' dell'introne (esone|GU.....AG|esone; vedi Figura 8.15). Queste sequenze si trovano nelle posizioni corrette solo se il filamento superiore del DNA è il filamento simile all'RNA:

5' TAGGTGAAAGA GTAGCCTAG AATCAGTTA 3'

**c.** Poiché l'introne è irrealisticamente breve, manca una sequenza di ramificazione a lazo. Inoltre, l'introne è privo di alcuni dei nucleotidi ricchi di pirimidina che si trovano solitamente appena a monte del sito accettore di giunzione (vedi Figura 8.15).

**d.** Il filamento di DNA simile all'RNA mostrato nella parte (b) produce il seguente trascritto primario:

5' UAGGUGAAAGA GUAGCCUAG AAUCAGUUA 3'

Dopo lo splicing, l'mRNA prodotto è il seguente, con i codoni di stop indicati in rosso:

5' UAGGUGAAAGAAUCAGUUA 3'

Poiché il problema afferma che entrambi gli esoni sono costituiti solo da codoni [e nessun codone di inizio (AUG) è presente in nessun registro di lettura], possiamo presumere che questa parte dell'mRNA contenga sequenze che si trovano nel mezzo del registro di lettura aperto. Solo uno dei tre possibili registri di lettura di questo mRNA è aperto:

5' XUA GGU GAA AGA AAU CAG UUA 3'

(X è un nucleotide indefinito che deve trovarsi nell'mRNA.) La traduzione di questo mRNA dà:

**N ...Gly Glu Arg Asn Gln Leu... C**

**10.6 a.** Poiché i cDNA sono costituiti da mRNA, mancano di introni e inoltre non contengono informazioni sulle regioni regolative come promotori ed enhancer.

**b.** I cloni in un cDNA possono sicuramente includere sequenze 5' UTR e 3' UTR. Il motivo è che le sequenze 5' UTR e 3' UTR si trovano negli esoni e i cloni di cDNA contengono questi esoni (vedi Figura 8.14).

**c.** Sarebbe più probabile trovare ORF più lunghi in media nei cloni di cDNA che nei cloni genomici. Nel DNA di sequenza casuale, ci si aspetta di trovare che il registro di lettura aperto medio codifichi solo circa 20 amminoacidi (perché ci so-

no 64 possibili codoni di cui 3 significano stop), ma la maggior parte delle proteine ha molti più amminoacidi. Se si ottenessero sequenze geniche da un clone di DNA genomico, gli esoni verrebbero interrotti da introni; se gli introni fossero di sequenza casuale, un frame di lettura iniziato in un esone si fermerebbe entro circa 60 nucleotidi (20 amminoacidi) nell'introne. I cloni di cDNA non hanno introni; contengono l'intera sequenza codificante. Quindi, se la proteina fosse, per esempio, lunga 1000 amminoacidi, un clone di cDNA avrebbe un registro di lettura aperto che codificherebbe per tutti questi amminoacidi.

- 10.7** Il motivo principale per cui i genetisti che studiano le cellule eucariotiche creano librerie di cDNA è determinare come il trascritto primario di un gene viene processato in un determinato tipo di cellula. Quasi tutti i geni eucariotici sono interrotti da introni. In effetti, solo una piccola parte (~1-2% negli esseri umani) della maggior parte dei genomi eucariotici sono costituiti da esoni codificanti proteine. Al contrario, i genomi dei batteri non hanno introni (con pochissime rare eccezioni) e la grande maggioranza delle copie di basi nei genomi batterici sono regioni codificanti proteine. Le genoteche a cDNA corrispondenti agli mRNA eucariotici sono quindi una fonte arricchita di esoni codificanti proteine. Inoltre, privi di introni, i cloni di cDNA hanno registri di lettura aperti completi, mentre i cloni genomici non avranno tali frame di lettura aperti (vedi Problema 10.6).

Sebbene i genetisti dei microrganismi di solito non abbiano bisogno di creare genoteche a cDNA, tali librerie sono preziose per determinati usi specifici, come per determinare quali geni vengono trascritti in condizioni ambientali diverse. La principale difficoltà per i genetisti dei microrganismi nella costruzione di genoteche a cDNA è che gli mRNA batterici non hanno code poli-A. Come si è visto in Figura 10.4, il primo passo nella creazione di librerie a cDNA eucariotiche è quello di far rinaturare l'oligo-dT alle code poli-A in modo che l'oligo-dT possa fungere da primer per la trascrittasi inversa. Ci sono modi in cui i genetisti dei microrganismi possono aggirare questo problema. Questi metodi sono per lo più al di fuori dello scopo di questo libro, ma il Problema 10.12 suggerisce uno dei diversi possibili approcci.

- 10.8 a.** La libreria genomica conterrebbe l'intero genoma umano, mentre le genoteche a cDNA del cervello e del fegato conterrebbero solo gli esoni

di alcuni geni. Le librerie genomiche contengono frammenti del genoma, inclusi sia i geni sia le sequenze intergeniche. Al contrario, le sequenze genomiche nelle librerie di cDNA sono limitate in due modi. Primo, poiché i cDNA sono costituiti da DNA a doppio filamento, copia di mRNA, i cDNA contengono solo esoni, non introni, né sequenze regolatrici di geni, né sequenze intergeniche. In secondo luogo, poiché non tutti i geni sono trascritti in tutte le cellule, le genoteche a cDNA del cervello o del fegato conterrebbero cDNA corrispondenti solo al sottoinsieme di geni trascritti in quel tessuto.

**b.** Queste tre genoteche dovrebbero condividere alcune sequenze. La libreria genomica includerà gli esoni trovati nelle librerie di cDNA. Alcuni geni sono espressi (trascritti) in tutti i tipi di cellule, quindi i cDNA per questi geni si troverebbero nelle genoteche a cDNA sia del cervello sia del fegato. Tuttavia, alcuni geni sono espressi nel cervello ma non nel fegato o viceversa, quindi le librerie di cDNA del cervello e del fegato avranno anche cloni di cDNA tessuto-specifici.

**c.** La libreria genomica parte dal DNA genomico umano; la libreria del cDNA del fegato parte dal pool totale di mRNA estratto dal fegato; il materiale di partenza per la libreria di cDNA del cervello è il pool di mRNA totale estratto dai cervelli.

**d.** Per annotare un genoma, si devono sequenziare molti cloni da molte genoteche a cDNA, perché geni diversi vengono trascritti in tessuti diversi e perché alcuni geni vengono trascritti solo raramente in qualsiasi tessuto.

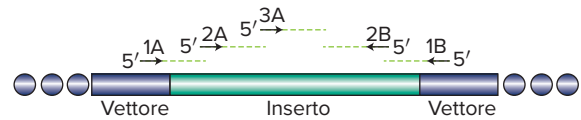
- 10.9** Nelle genoteche a cDNA sono rappresentati solo gli mRNA (almeno quelle realizzate con le tecniche descritte in Figura 10.4) perché gli RNA da copiare vengono purificati dal citoplasma in virtù delle loro code poli-A. Gli RNA che non sono tradotti (cioè gli RNA non codificanti [ncRNA]) non hanno code poli-A. Questi ncRNA includono tRNA, rRNA, vari RNA che regolano la trascrizione e la traduzione ed enzimi a RNA.

- 10.10** Diversi tipi di geni sono difficili da scovare, quindi i conteggi attuali del numero di geni sono quasi certamente incompleti. Questi geni difficili da trovare includono: (1) geni che vengono trascritti solo raramente (quindi i cDNA corrispondenti all'mRNA rappresenterebbero solo una percentuale molto piccola di tutti i cDNA sequenziati); (2) geni che codificano per proteine molto piccole (quindi il registro di lettura aperto sarebbe troppo breve per essere identificato dall'analisi al computer); (3) geni che codificano per proteine che

sono scarsamente conservate attraverso l'evoluzione (quindi i confronti tra specie non sarebbero informativi); (4) geni che sono trascritti in RNA non codificanti (ncRNA) che non sono tradotti in proteine. Diverse classi di ncRNA sono state trovate solo negli ultimi anni, quindi è molto probabile che rimangano da identificare ulteriori ncRNA (e i geni che li codificano).

**10.11 a.** Sì, se si sequenziassero molti cloni da una genoteca a cDNA, si sarebbero sequenziate molte copie indipendenti delle stesse molecole di mRNA. Per esempio, se un particolare mRNA fosse molto abbondante tra tutti gli mRNA in un tessuto, allora molti cloni avrebbero in effetti gli stessi inserti di cDNA corrispondenti a questo mRNA. Tuttavia, queste informazioni ridondanti hanno ancora valore. Ovviamente, la frequenza con cui un particolare tipo di clone di cDNA viene trovato nella libreria riflette l'abbondanza dell'mRNA corrispondente, quindi è possibile stimare i livelli relativi di espressione genica dal sequenziamento di molti cloni provenienti dallo stesso mRNA. Altri usi di queste informazioni potrebbero essere l'identificazione di mRNA prodotti dallo splicing alternativo dallo stesso trascritto primario o, più speculativamente, la ricerca di alterazioni post-trascrizionali inaspettate apportate ad alcuni mRNA.

**b.** Segue un diagramma della procedura di avanzamento progressivo del primer. Si devono utilizzare le informazioni sulla sequenza ottenute in ogni passaggio della procedura per progettare i primer necessari per il passaggio successivo. Questi primer si troverebbero vicino all'estremità 3' delle sequenze ottenute e l'orientamento da 5' a 3' dei primer "punterebbe" nella direzione della sequenza di DNA dell'inserto che vorresti ottenere nel passaggio successivo. Nella figura seguente, le frecce nere rappresentano i primer da usare. I primer 1A e 1B rappresentano i due primer del primo passaggio (in reazioni di sequenziamento separate provenienti da giunzioni vettore/inserto opposte), 2A e 2B i due primer nel passaggio successivo, ecc. Le **linee verdi tratteggiate** indicano le nuove sequenze di DNA che si ottengono con ogni primer. (I primer sono disegnati su scala più grande di quanto non sarebbero nella realtà: i primer sono solo circa 20-25 basi, mentre puoi sequenziare circa 1000 basi di DNA da ciascun primer.) I **cerchietti blu** alle estremità del vettore indicano che il vettore è in realtà circolare.



**10.12 a.** 5' CCCCCG 3'  
3' GGGGGCTTAA 5'

**b.** 5' AATTCGGGGG — CCCCCG 3'  
3' GGGGGCTTAA — GGGGGCTTAA 5'  
cDNA

**c.** Questi adattatori forniscono ai cDNA estremità compatibili con i vettori digeriti con *EcoRI*.

**d.** Il successo della reazione di ligasi delle estremità nette dipende dalla concentrazione delle estremità nette nella miscela di ligatura. Nel caso delle ligature con le estremità coesive, l'ibridazione delle estremità coesive tiene insieme le due molecole di DNA, in modo che la ligasi abbia la possibilità di unirle in modo covalente. Le molecole di DNA con estremità nette che si agitano nella soluzione non si legano insieme come fanno le molecole con estremità coesive. Una concentrazione molto elevata di estremità nette nella miscela può superare questo impedimento: più le estremità nette si trovano vicine l'una all'altra, più è probabile che la ligasi sarà in grado di creare un legame fosfodiesterico tra due di esse. Poiché le molecole adattatrici sono molto più piccole di un tipico vettore plasmidico (~3000 bp), è possibile creare una soluzione con una concentrazione di molecole adattatrici molto più elevata di quella che potresti raggiungere con le molecole del vettore. Per esempio, una soluzione contenente 100 ng degli adattatori della parte (b) avrà 750 volte più estremità smussate rispetto a una soluzione contenente 100 ng di molecole di un vettore da 3000 bp.  $[(3000 \text{ bp/vettore} \times 2 \text{ estremità}) / (8 \text{ bp/adattatore} \times 1 \text{ estremità})] = 750 \text{ estremità adattatore/vettore estremità.}]$

**10.13** (1) I geni nel genoma umano sono interrotti frequentemente da introni, mentre i genomi batterici mancano di introni. (2) Le regioni intergeniche (regioni tra i geni) possono essere piuttosto lunghe nel genoma umano, mentre sono quasi tutte molto corte nei genomi batterici. (3) Il genoma umano ha lunghi tratti di sequenze di DNA ripetitivo, per esempio ai centromeri e ai telomeri, che mancano nei genomi batterici.

**10.14** I due diversi cloni di cDNA rappresentano prodotti dallo splicing alternativo del trascritto primario dello stesso gene. Questi mRNA alternativi derivano dall'unione di esoni diversi. Per esempio, supponiamo che il gene abbia 5 esoni. Un

prodotto di splicing potrebbe mantenere gli esoni 1, 2, 3 e 5, mentre l'altro potrebbe avere gli esoni 1, 2, 4 e 5. Le estremità dei due mRNA (esone 1 ed esone 5) sono le stesse, ma gli esoni intermedi sono diversi.

**10.15 a.** In questa regione esistono due geni annotati. Un gene produce cinque trascritti alternativi rappresentati dai cloni di cDNA A-E; il secondo gene produce due trascritti alternativi F e G.

**b.** L'idiogramma nella parte superiore della figura mostra che questa regione si trova nel braccio lungo (q) del cromosoma 9. Il gene corrispondente ai trascritti A-E è trascritto in direzione dal telomero al centromero (questa è la direzione in cui si muove l'RNA polimerasi durante la trascrizione del gene). Il secondo gene corrispondente ai trascritti F e G è trascritto nella direzione opposta, dal centromero verso il telomero.

**c.** I dati suggeriscono 3 promotori. Un promotore viene utilizzato per produrre gli mRNA A e B; questo promotore si trova all'incirca in posizione 76 905 000 sul cromosoma 9. Il secondo promotore è usato per produrre gli mRNA C-E; questo promotore è all'incirca in posizione 76 695 000. Il terzo promotore per gli RNA F e G si trova approssimativamente nella posizione 76 770 000.

**d.** I trascritti corrispondenti ai cDNA A, C, D ed E potrebbero codificare ciascuno per polipeptidi unici. (Si noti che C, D ed E incorporano piccoli esoni diversi codificanti per proteine.) Tuttavia, i trascritti B, F e G apparentemente non contengono alcun registro di lettura aperto; sembra che nessuno di questi trascritti codifichi un polipeptide. Pertanto, i dati indicano che queste sequenze di DNA potrebbero codificare per 4 diverse proteine. Poiché il registro di lettura aperto del trascritto A è molto più lungo di quello dei trascritti C, D ed E, è ipotizzabile che solo la proteina codificata dall'mRNA A sia funzionale.

**e.** Due cose sembrano essere insolite. Innanzitutto, uno dei due geni (corrispondenti ai trascritti F e G) viene trascritto in RNA non codificanti che non vengono tradotti. In secondo luogo, questo gene non codificante (F e G) si trova interamente all'interno di un lungo introne dell'altro gene. Nonostante le apparenze, nessuna di queste situazioni è in realtà così insolita. Cioè, il genoma umano include migliaia di geni non codificanti e ci sono molte centinaia di esempi di geni situati all'interno degli introni di altri geni. Questi geni all'interno dei geni possono essere codificanti o non codificanti.

**10.16 a.** I trascritti F e G sono RNA non codificanti, che di solito non hanno code poli-A.

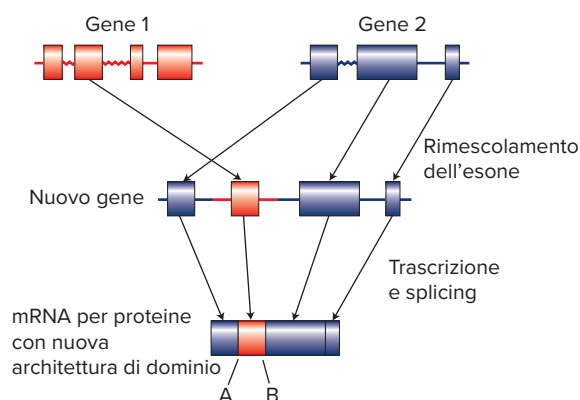
**b.** Il primo passo nella sintesi del cDNA è la purificazione degli mRNA usando le loro code poli-A.

**c.** Per produrre cDNA, si deve usare l'enzima trascrittasi inversa e questo enzima, come la DNA polimerasi, richiede un primer. Sono possibili due tipi di approcci per fornire questi primer.

In primo luogo, si potrebbero utilizzare oligonucleotidi di DNA casuali come primer; alcuni di questi sarebbero complementari all'estremità 3' dell'RNA non codificante. Tuttavia, in questo modo genererebbero un numero enorme di cDNA, che potrebbero iniziare ovunque all'interno di un trascritto, non solo all'estremità.

Una strategia alternativa consiste nell'aggiungere una sequenza all'estremità 3' degli RNA non codificanti. Questo obiettivo può essere raggiunto con l'enzima RNA ligasi o polimerizzando sequenze come la poli-C sulle estremità 3' con enzimi di funzione simile alla poli-A polimerasi. Si dovrebbe quindi innescare la trascrizione inversa con un primer complementare alle sequenze aggiunte.

**10.17** Per quanto riguarda lo scenario rappresentato in Figura 10.11 e riprodotto qui sotto, supponiamo che tutti gli esoni siano codificanti proteine in modo che un'unità costituita da un esone del Gene 1 più parte dei suoi introni fiancheggiati (rossi) venga riasssemblata all'interno di un introne del Gene 2 (blu). Il nuovo gene può produrre, dopo lo splicing, un nuovo mRNA maturo con tutti gli esoni del Gene 2 più il nuovo esone del Gene 1. Indichiamo le due giunzioni tra le parti blu e rossa di questo mRNA, con A e B (vedi figura).



Alla giunzione A, l'esone blu potrebbe essere unito all'esone rosso in uno qualsiasi dei 6 modi, solo



uno dei quali rispetta il registro di lettura. (In primo luogo, la parte riasssemblata del Gene 1 dovrebbe essere inserita nel Gene 2 con la stessa polarità [contro la polarità opposta] [probabilità =  $1/2$ ]). In secondo luogo, anche se le polarità relative sono corrette, gli introni devono trovarsi nella stessa posizione rispetto ai codoni in entrambi i geni [probabilità =  $1/3$ ]. Per esempio, se l'introne nel gene 2 si trovava esattamente tra due codoni [piuttosto che tra due nucleotidi che fanno parte dello stesso codone], per creare una nuova proteina composta da domini di entrambi i prodotti genici originali, anche l'introne nel gene 1 dovrebbe trovarsi tra due codoni.)

Se la giunzione A nell'mRNA maturo è in frame, è necessario considerare cosa succede alla giunzione B. Gli orientamenti devono già essere corretti. Tuttavia, la fine dell'esone rosso a questa giunzione potrebbe trovarsi in una qualsiasi delle tre posizioni rispetto ai codoni e solo una di queste tre possibilità sarà nella stessa cornice di lettura dell'esone blu.

Pertanto, la probabilità che un evento di mescolamento degli esoni come quello mostrato in Figura 10.11 produca una nuova proteina con i domini che erano presenti nel prodotto del Gene 2 originale più un nuovo dominio del Gene 1 originale è  $(1/6) \times (1/3) = 1/18$ . Se il nuovo mRNA nella parte inferiore della figura dovesse trovarsi in frame solo in corrispondenza di una giunzione anziché di due (come sarebbe il caso se per esempio il codone di inizio fosse nell'esone rosso, la probabilità sarebbe  $1/6$ ).

Anche se solo una frazione dei possibili eventi di rimescolamento degli esoni che potrebbero verificarsi produrrebbe una nuova proteina con una nuova conformazione dei domini, il rimescolamento degli esoni è ancora molto più efficiente della semplice unione di pezzi casuali del genoma per creare nuovi geni. Nel mescolamento degli esoni, le rotture dei cromosomi possono verificarsi praticamente ovunque all'interno degli introni. Se si unissero pezzi casuali, si avrebbero nuove architetture dei domini solo quando le rotture si trovavano fra i codoni appropriati e i prodotti uniti mantenevano il registro di lettura

**10.18** Il genoma umano contiene due tipi di pseudogeni: pseudogeni duplicati prodotti come risultato della duplicazione di geni e della successiva divergenza delle loro sequenze (vedi Figura 10.13) e pseudogeni maturati, prodotti attraverso la trascrizione inversa di mRNA (in cui le copie di cDNA degli mRNA vengono inserite nel geno-

ma). Gli pseudogeni di entrambi i tipi si modificano in tempi evolutivi, fino a non essere più riconoscibili.

**a.** Gli pseudogeni maturati dovrebbero conservare i resti della struttura dell'mRNA. Tali pseudogeni non dovrebbero contenere introni e possono anche avere una coda poli-A nella regione corrispondente all'estremità 3' del trascritto. Si noti che è improbabile che questi pseudogeni maturati vengano trascritti, poiché l'mRNA non contiene il promotore.

**b.** È molto probabile che i meccanismi che danno origine agli pseudogeni maturati (quelli copiati dagli mRNA) si traducano in pseudogeni sparsi nel genoma. Gli mRNA non rimangono associati ai geni da cui sono trascritti, perché lasciano il nucleo per essere tradotti nel citoplasma. Di conseguenza, ci sarebbero poche possibilità che un cDNA a doppio filamento venga incorporato nel genoma in una posizione adiacente a quella del gene originale. Al contrario, alcuni meccanismi di duplicazione genica, come il crossing-over ineguale favoriscono la creazione di famiglie di geni, i cui membri sono raggruppati insieme nella stessa area del genoma.

**10.19 a.** Poiché i domini a dita di zinco consentono il legame al DNA, l'ipotesi più probabile è che la proteina sia coinvolta in un qualche tipo di metabolismo dipendente dal DNA. In termini di espressione genica, tale proteina potrebbe essere un fattore di trascrizione che modula la trascrizione di geni vicini ai siti di legame sul DNA riconosciuti dalla proteina. Le proteine che legano il DNA potrebbero anche essere coinvolte in altri processi come la replicazione del DNA, l'impacchettamento del DNA nei cromosomi o la segregazione del DNA durante la divisione cellulare.

**b.** Se due geni nello stesso organismo condividono una somiglianza di base significativa, è molto probabile che i due geni siano omologhi, correlati tra loro attraverso un gene ancestrale comune che ha subito la duplicazione e i cui geni discendenti si sono poi differenziati l'uno dall'altro. Se la somiglianza è molto alta, o l'evento di duplicazione è avvenuto nel recente passato evolutivo, oppure entrambi i geni svolgono funzioni essenziali che sono sensibili alla mutazione (cioè, la selezione naturale rimuoverebbe gli alleli mutanti di entrambi i geni). Rispetto alla terminologia introdotta in Figura 10.14, due geni nello stesso organismo strettamente correlati tra loro sono geni paraloghi.

**10.20 a.** La migliore sequenza consenso è:

5' NNATATAAAANNNNNNNN 3'  
       G      T

Otto delle posizioni nel gruppo di quattro sequenze possono avere qualsiasi nucleotide; questo è indicato con una N. La posizione 3 può avere una A o una G, la posizione 8 può avere una A o una T. I restanti nucleotidi sono invarianti: hanno sempre T o A come mostrato.

**b.** Una sequenza consenso mostra i nucleotidi che probabilmente sono i più importanti per la funzione del gene o del cromosoma. Questi nucleotidi sono stati conservati durante il periodo evolutivo che parte dall'ultima volta che i quattro organismi hanno condiviso un antenato evolutivo comune. Poiché qualsiasi breve tratto di sequenze casuali potrebbe erroneamente sembrare di avere nucleotidi conservati, è necessario prima determinare se queste sequenze abbiano qualche relazione funzionale. È necessario scegliere sequenze di DNA che si trovano in geni ortologi e nella stessa posizione relativa (per esempio, all'interno dello stesso esone) nelle quattro specie.

**c.** Se si confrontano gli amminoacidi delle proteine codificate da geni omologi, si dovrebbe essere in grado di trovare particolari amminoacidi conservati da usare per definire un consenso. Anche se l'amminoacido non è identico, è possibile che possa essere sostituito in altre specie da un altro amminoacido con proprietà chimiche simili. Per esempio, la sostituzione dell'amminoacido acido aspartico con acido glutammico è definita una sostituzione conservativa. Questo tipo di confronto è più utile per le sequenze di amminoacidi che per le sequenze di acidi nucleici, perché la degenerazione del codice genetico potrebbe mascherare la conservazione. (Cioè, due codoni potrebbero codificare lo stesso amminoacido pur essendo differenti.)

**10.21** Per esaminare quali eventi combinatori accadono a livello di DNA delle immunoglobuline, si possono creare librerie di DNA genomico da cellule non del sistema immunitario (in cui non si sarebbero verificati eventi combinatori) e da cellule del sistema immunitario a vari stadi di differenziamento in cellule B, comprese le stesse cellule B mature che producono gli anticorpi. Quindi si isolano diverse copie dei geni che codificano per gli anticorpi da ciascuna di queste librerie e si determina le loro sequenze di DNA. Qualsiasi differenza rilevata mostrerebbe il tipo di eventi che si verificano per riorganizzare o al-

terare il DNA dei geni degli anticorpi in modo che possano produrre così tanti tipi di anticorpi.

Per verificare la possibilità di splicing alternativo, è necessario prima creare una libreria di cDNA dalle cellule B che producono anticorpi. (Non è necessario creare librerie di cDNA da altri tipi di cellule perché non producono mRNA di anticorpi.) Si sequenziano quindi molti cDNA corrispondenti agli mRNA che codificano per anticorpi. Se si è verificato uno splicing alternativo, è possibile rilevarlo confrontando le sequenze di cloni diversi. Si potrebbe anche caratterizzare lo splicing coinvolto confrontando queste sequenze di cDNA con le sequenze genomiche dei geni degli anticorpi nei linfociti B.

**10.22 a.** Tutti questi geni, sia negli esseri umani, sia negli scimpanzé, sarebbero considerati omologi poiché derivano tutti da un antenato comune. Tuttavia, il tipo di omologia differirebbe per le diverse coppie di geni [vedi parte (c)]. I geni  $\alpha 1$ ,  $\alpha 2$ ,  $\beta$ , G,  $A\gamma$ ,  $\delta$ ,  $\epsilon$  e  $\zeta$  nell'uomo costituirebbero un insieme di geni paraloghi; anche i geni  $\alpha 1$ ,  $\alpha 2$ ,  $\beta$ , G,  $A\gamma$ ,  $\delta$ ,  $\epsilon$  e  $\zeta$  negli scimpanzé costituirebbero un insieme di geni paraloghi. Il gene  $\alpha 1$  nell'uomo sarebbe ortologo al gene  $\alpha 1$  negli scimpanzé, il gene  $\alpha 2$  nell'uomo sarebbe ortologo al gene  $\alpha 2$  negli scimpanzé, ecc.

**b.** È più probabile che i geni ortologi abbiano funzioni strettamente correlate, poiché sono stati derivati più recentemente da un antenato comune rispetto ai geni paraloghi.

**c.** I geni ortologi (il gene  $\beta$  umano e il gene  $\beta$  dello scimpanzé) dovrebbero avere una maggiore somiglianza nucleotidica, poiché sono stati derivati più recentemente da un antenato comune rispetto ai geni  $\alpha$  e  $\beta$  umani paraloghi.

**d.** La maggior parte degli eventi di duplicazione genica illustrati nella Figura 10.13 che hanno dato origine ai diversi tipi di geni dell'emoglobina ( $\alpha 1$ ,  $\alpha 2$ ,  $\beta$ , G $\gamma$  A,  $\delta$ ,  $\epsilon$  e  $\zeta$ ) deve essersi verificata prima che si differenziasse la linea evolutiva degli umani e degli scimpanzé. Cioè, l'ultimo antenato comune di queste specie deve aver già avuto tutti questi geni. In effetti, la maggior parte di questi geni deve essere stata presente in un'antica specie ancestrale di tutti i mammiferi poiché la maggior parte di questi geni si trova in tutti i mammiferi odierni.

**10.23** Il proteoma umano è molto più complesso di quello del nematode, anche se il genoma umano ha solo circa 5000 geni in più. (La maggior parte dei geni codifica per proteine in entrambe le specie). Il motivo per cui i numeri dei geni non ri-

specchiano la complessità dell'organismo non è completamente noto, ma è probabile che siano coinvolti diversi meccanismi. Un fattore che contribuisce è che le proteine umane potrebbero essere soggette a più tipi diversi di modificazioni chimiche rispetto alle proteine nei nematodi. Più di 400 diversi tipi di reazioni chimiche possono influenzare le proteine nelle cellule umane, quindi il nostro proteoma è molto più grande di quanto sia il nostro genoma. La maggior parte di queste reazioni di modificazione si verifica anche nei nematodi, ma è possibile che alcune non abbiano luogo. Un altro fattore importante è che l'amplificazione combinatoria sia a livello di DNA (per esempio, unione V-D-J di geni anticorpali), sia di RNA (splicing alternativo) si verifica più frequentemente negli esseri umani che nei vermi.

**10.24 a.** Molto verosimilmente, gli axolotl non hanno un numero di geni 10 volte superiore a quello degli esseri umani. La dimensione molto più grande del genoma dell'axolotl rispetto al genoma umano è probabilmente dovuta al fatto che il genoma dell'axolotl ha più DNA cosiddetto "spazzatura" e introni più grandi.

**b.** È probabile che i geni coinvolti nella rigenerazione siano espressi a livelli elevati nei siti di rigenerazione. Gli scienziati potrebbero creare librerie di cDNA da arti rigeneranti e da arti normali e confrontarne il contenuto.

**10.25 a.** Il gene CFTR nell'uomo ha 27 esoni (e 26 introni). (Nel browser, gli esoni sono più spessi degli introni. Gli introni sono indicati da una linea sottile con punte di freccia che mostrano la direzione della trascrizione. Non ci si deve preoccupare se si è perso il conteggio di uno dei piccoli esoni.)

**b.** Sull'idiogramma nella parte superiore della finestra del browser, si può vedere che CFTR si trova circa a metà del braccio lungo (q) del cromosoma 7 (nella banda 7q31.3).

**c.** La direzione della trascrizione (indicata dalle punte di freccia lungo gli introni) è dal centromero, verso il telomero. Cioè, l'RNA polimerasi si muove verso il telomero mentre trascrive il gene CFTR.

**d.** Il gene a sinistra di CFTR (cioè il gene successivo più vicino al centromero) è chiamato ASZ1. Viene trascritto nella direzione opposta a quella per CFTR, quindi viene trascritto dall'altro filamento. Il gene appena a destra di CFTR (più vicino al telomero) è chiamato CTTNBP2. Viene anch'esso trascritto sull'altro filamento (ovvero, il suo stampo per la trascrizione è il filamento opposto al filamento stampo per CFTR).

**e.** Il cromosoma 7 è lungo circa 160 Mb (milioni di paia di basi). Si può vedere questo numero nella parte superiore della finestra del browser dopo aver rimpicciolito abbastanza da evidenziare l'intero cromosoma in rosso sull'idiogramma. Questo numero non è del tutto accurato, perché le regioni cromosomiche intorno ai centromeri hanno grandi quantità di DNA ripetitivo le cui sequenze sono difficili da determinare.

**f.** Il centromero si trova a circa 60 milioni di paia di basi, con la numerazione che inizia all'estremità sinistra [cioè al telomero del braccio corto (p)].

**g.** I geni sembrano sovrapporsi perché non c'è abbastanza spazio per visualizzarli tutti sul cromosoma uno dopo l'altro sulla stessa riga, quando si visualizza l'intero cromosoma in un'unica finestra.

**10.26** Quando si esegue la ricerca BLAST, la finestra di risposta dovrebbe avere tre componenti principali come mostrato nel diagramma allegato. Nelle edizioni più recenti di BLAST, questi componenti sono mostrati come schede. (Nella discussione di seguito, ignoreremo la scheda chiamata "Tassonomia", che non è pertinente a questo problema.)

La prima scheda della finestra di risposta si chiama "Descrizioni" e presenta un elenco degli hit, ovvero le sequenze di DNA nel database cercato, che sono più strettamente correlate alla query (la domanda con cui si è interrogato il database). I primi due risultati sono mostrati nella immagine seguente. Si noti che i valori E di questi risultati sono entrambi  $3e-07$ , il che significa più o meno che la possibilità che una sequenza casuale di DNA abbia lo stesso grado di somiglianza con la Query è solo di circa  $3 \times 10^{-7}$ , ovvero inferiore a uno su un milione.

#### Sequences producing significant alignments:

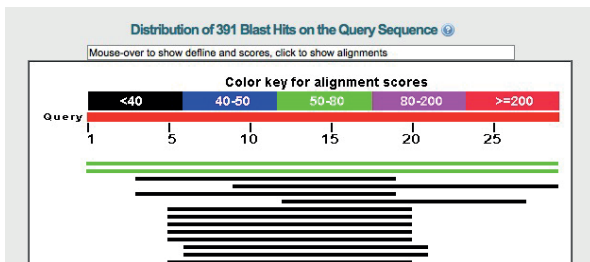
Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
Transcripts							
<input type="checkbox"/>	<a href="#">Homo sapiens protein phosphatase 1, regulatory (inhibitor) subunit 14A (PPP1R14A), transcript variant 2, mRNA</a>	58.0	58.0	100%	3e-07	100%	<a href="#">NM_001243947.1</a>
<input type="checkbox"/>	<a href="#">Homo sapiens protein phosphatase 1, regulatory (inhibitor) subunit 14A (PPP1R14A), transcript variant 1, mRNA</a>	58.0	58.0	100%	3e-07	100%	<a href="#">NM_033256.2</a>



La seconda scheda denominata “Riepilogo grafico” porta a una rappresentazione grafica dei risultati. (le cui prime voci appaiono come segue). Più alti sono i punteggi di allineamento, più alti sono i valori E per la regione indicata. Pertanto, nello screenshot del Riepilogo grafico di seguito, le linee verdi indicano che due sequenze nel database hanno forti omologie con la sequenza Query di Figura 9.7f. (Corrispondono ai primi due risultati nella Descrizioni mostrate sopra.) Le linee nere sono altre sequenze con gradi di somiglianza inferiori; le ignoreremo.



La terza scheda nella finestra di risposta chiamata “Allineamenti” mostra le corrispondenze effettive tra la sequenza di query e i vari risultati. L’output dovrebbe essere il seguente per la hit più elevata:

**a.** Come visto nella finestra precedente, il gene umano (il Soggetto) più simile alla sequenza Query codifica per la proteina fosfatasi 1, subunità regolatoria (inibitrice) 14A (nota anche come PPP1R14A); questa proteina inibisce un tipo specifico di enzima che rimuove i gruppi fosfato dalle proteine.

**b.** La corrispondenza è esatta. Tutti i nucleotidi della Query combaciano perfettamente con la sequenza Soggetto, quindi questo è quasi certamente il gene la cui sequenza appare in Figura 9.7f. La sequenza Soggetto è in realtà una sequenza di cDNA corrispondente a una variante di splicing alternativo dell’mRNA per questo gene. (L’altro colpo verde corrispondente è una variante di splicing dello stesso gene che include anche la sequenza di query.)

**c.** La maggior parte di questa parte del gene PPP1R14A è abbastanza ben conservata nel genoma del topo; l’allineamento della sequenza è mostrato nella schermata seguente. Si noti tuttavia che l’identità della sequenza inizia solo al nucleotide 9 della query, quindi i primi 8 nucleotidi all’estremità 5’ della query non sono ben conservati nella copia murina di questo gene.

Download ▾ GenBank Graphics
Next Previous Descriptions

Homo sapiens protein phosphatase 1, regulatory (inhibitor) subunit 14A (PPP1R14A), transcript variant 2, mRNA  
Sequence ID: [ref|NM\\_001243947.1|](#) Length: 701 Number of Matches: 1

Range 1: 254 to 282 GenBank Graphics
Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
58.0 bits(29)	3e-07	29/29(100%)	0/29(0%)	Plus/Plus

Query 1 TGGCAGCTCAGCGGCTGGGCAAGCGCGTG 29  
Sbjct 254 TGGCAGCTCAGCGGCTGGGCAAGCGCGTG 282

Related Information  
[Gene](#) - associated gene details  
[UniGene](#) - clustered expressed sequence tags  
[Map Viewer](#) - aligned genomic context

---

Download ▾ GenBank Graphics
Next Previous Descriptions

Mus musculus protein phosphatase 1, regulatory (inhibitor) subunit 14A (Ppp1r14a), mRNA  
Sequence ID: [ref|NM\\_026731.3|](#) Length: 531 Number of Matches: 1

Range 1: 25 to 45 GenBank Graphics
Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
42.1 bits(21)	0.014	21/21(100%)	0/21(0%)	Plus/Plus

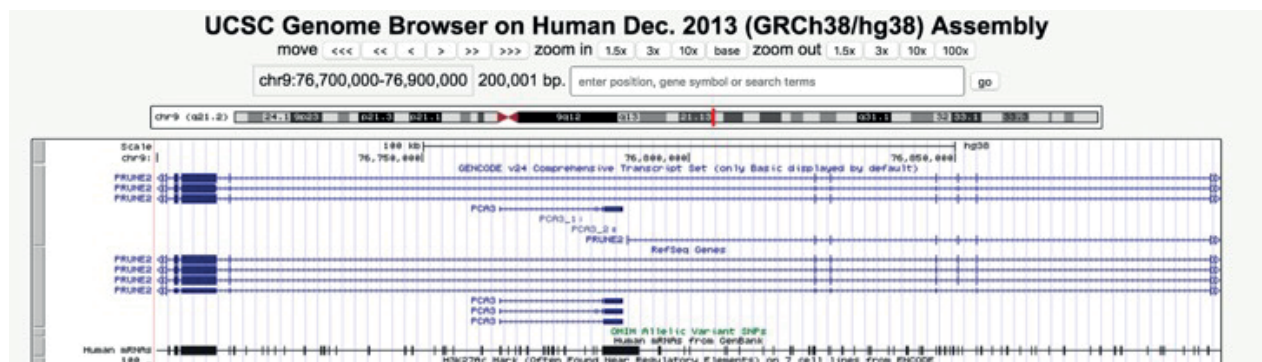
Query 9 CAGCGGCTGGGCAAGCGCGTG 29  
Sbjct 25 CAGCGGCTGGGCAAGCGCGTG 45

Related Information  
[Gene](#) - associated gene details  
[UniGene](#) - clustered expressed sequence tags  
[GEO Profiles](#) - microarray expression data  
[Map Viewer](#) - aligned genomic context

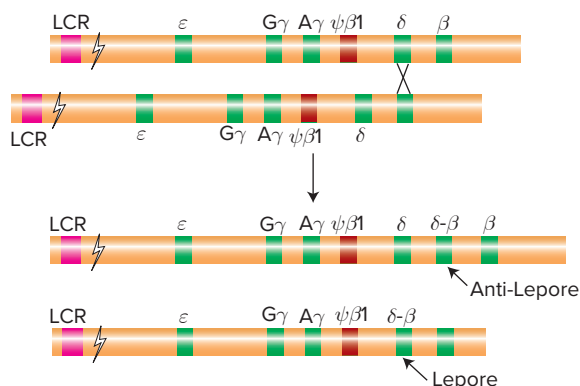
Fonte delle cifre per il problema 26: National Institutes of Health.

**10.27** Immettere chr9: 76,700,000-76,900,000 nella casella Posizione/Termine di ricerca. Si trova immediatamente che i due geni sono PRUNE2 (che

codifica per proteine) e PCA3 (che viene trascritto in un RNA non codificante):



**10.28 a.** Il crossing-over ineguale tra le due copie omologhe del cluster di geni  $\beta$ , tra il gene  $\delta$  su un cromosoma e il gene  $\beta$  sul cromosoma omologo, produrrebbe nuovi geni che codificano per catene polipeptidiche che potrebbero formare l'emoglobina Lepore  $\alpha(\delta-\beta) 2$  o l'emoglobina anti-Lepore  $\alpha 2(\beta-\delta) 2$  (vedi la figura che segue). Il crossing-over ineguale può aver luogo perché i geni  $\delta$  e  $\beta$  sono omologhi, con sequenze di DNA simili (sebbene non identiche).



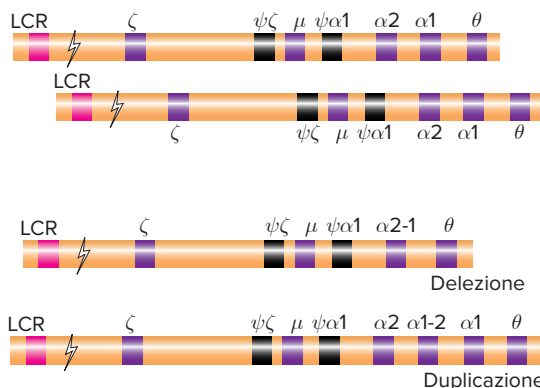
**b.** Gli individui lievemente talassemici sono eterozigoti per l'insolito allele Lepore, il che significa che meno emoglobina totale di quella che si troverebbe in un individuo completamente normale (meno del 100% ma più del 50%).

**c.** Il motivo per cui questi eterozigoti avrebbero meno emoglobina Lepore rispetto alla normale emoglobina adulta risiede nel fatto che il gene  $\delta$  viene trascritto molto meno frequentemente del gene  $\beta$ , quindi gli adulti normali hanno molta più emoglobina  $\alpha 2 \beta 2$  di  $\alpha 2 \delta 2$  (vedi Figura 10.20). Il promotore di ciascun gene si trova all'estremità sinistra (perché i geni sono trascritti da sinistra a destra come si osserva in Figura

10.21). Pertanto, la trascrizione dell'emoglobina Lepore dipende dal promotore molto debole per il gene  $\delta$ , quindi viene prodotta relativamente poca emoglobina Lepore.

Per questo motivo, i rari individui omozigoti per l'allele Lepore producono pochissima emoglobina (e questa emoglobina sarebbe tutta  $\alpha 2(\delta-\beta) 2$ ). Di conseguenza, questi omozigoti presentano una grave  $\beta$ -talassemia; con grave anemia e anomalie scheletriche dovute a difetti di crescita durante l'infanzia.

**10.29** Proprio come nel Problema 28, il crossing-over ineguale tra i geni  $\alpha 1$  e  $\alpha 2$  potrebbe spiegare come alcune persone arrivino a ereditare il cluster di geni  $\alpha$  con un solo gene  $\alpha$  (vedi il diagramma che segue). Si noti in Figura 10.22 che gli individui omozigoti per tale delezione avranno una lieve anemia, perché producono solo circa la metà della globina  $\alpha$  rispetto alle persone normali.



**10.30 a.** Si ricordi che l'emoglobina è costituita da due subunità di tipo  $\alpha$  e due subunità di tipo  $\beta$ . L'emoglobina X è  $\alpha 2 \beta 2$  (cioè due subunità  $\alpha$  e due subunità  $\beta$ ); questa è la normale emoglobina adulta mostrata nella Figura 10.20c. L'emoglobina Y è

$\alpha_2\gamma_2$ ; cioè, è l'emoglobina fetale. L'emoglobina Z è  $\alpha_2\delta_2$ ; cioè, è la forma minore di emoglobina negli adulti.

**b.** In un bambino di un anno, la  $\beta$ -talassemia dovrebbe comportare un aumento della proporzione di emoglobina HbZ, semplicemente perché un tale individuo mancherebbe dell'emoglobina X normalmente predominante (emoglobina adulta normale). Se non trattata, la  $\beta$ -talassemia porterebbe alla morte del bambino entro pochi anni, perché la malattia provoca la morte della maggior

parte dei globuli rossi. Fortunatamente, questi bambini possono essere trattati efficacemente con le trasfusioni di sangue.

**c.** Gli individui con persistenza ereditaria dell'emoglobina fetale (vedi Figura 10.21b) continuano a produrre emoglobina fetale (emoglobina Y nella figura) anche da adulti. È interessante notare che, nonostante le diverse caratteristiche di trasporto dell'ossigeno delle emoglobine fetali e adulte, nelle persone con questa condizione i disturbi clinici sono pochi o persino assenti.