Manuela Helmer Citterich, Fabrizio Ferrè, Giulio Pavesi, Chiara Romualdi, Graziano Pesole

Fondamenti di bioinformatica

ZANICHELLI

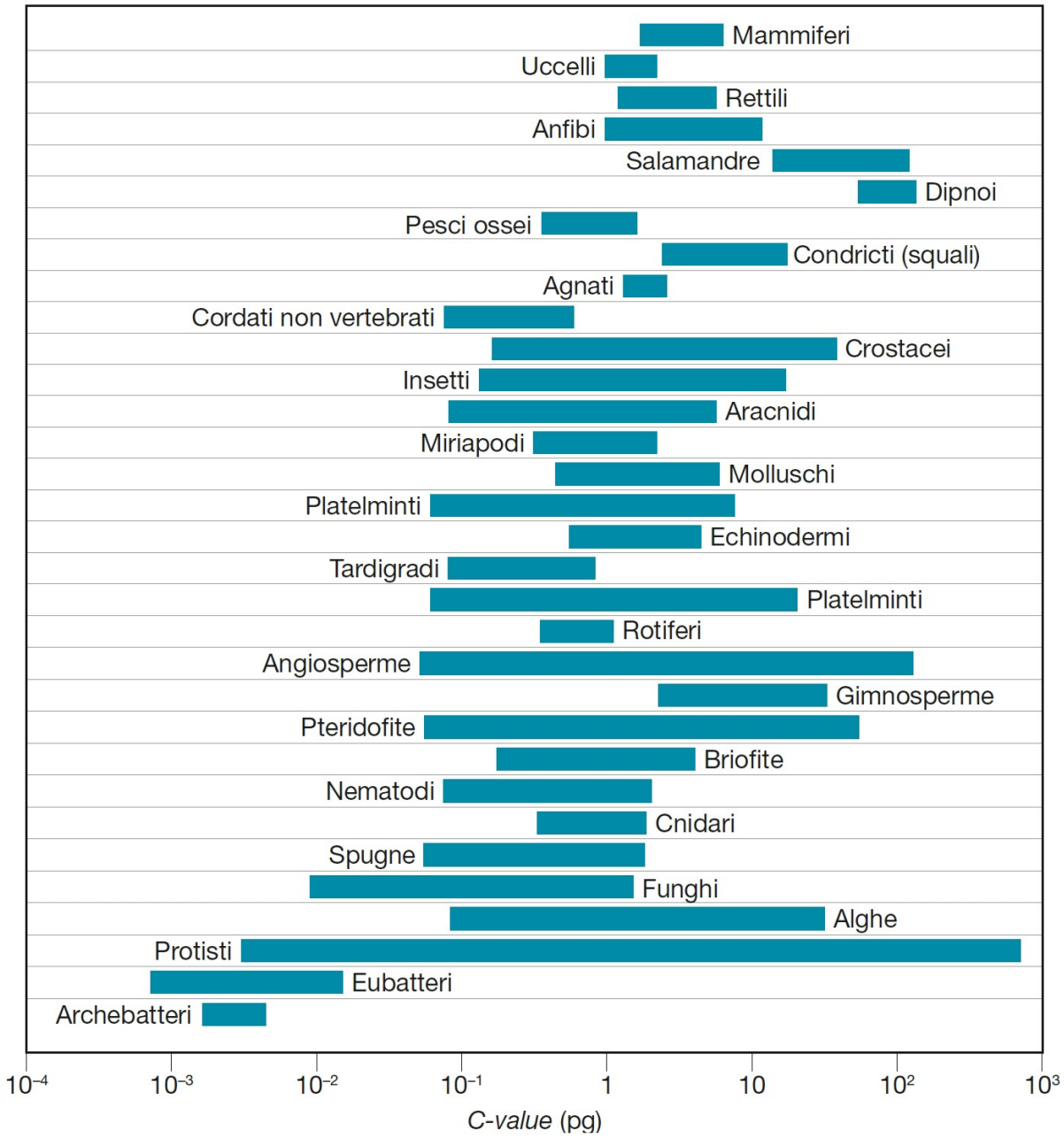
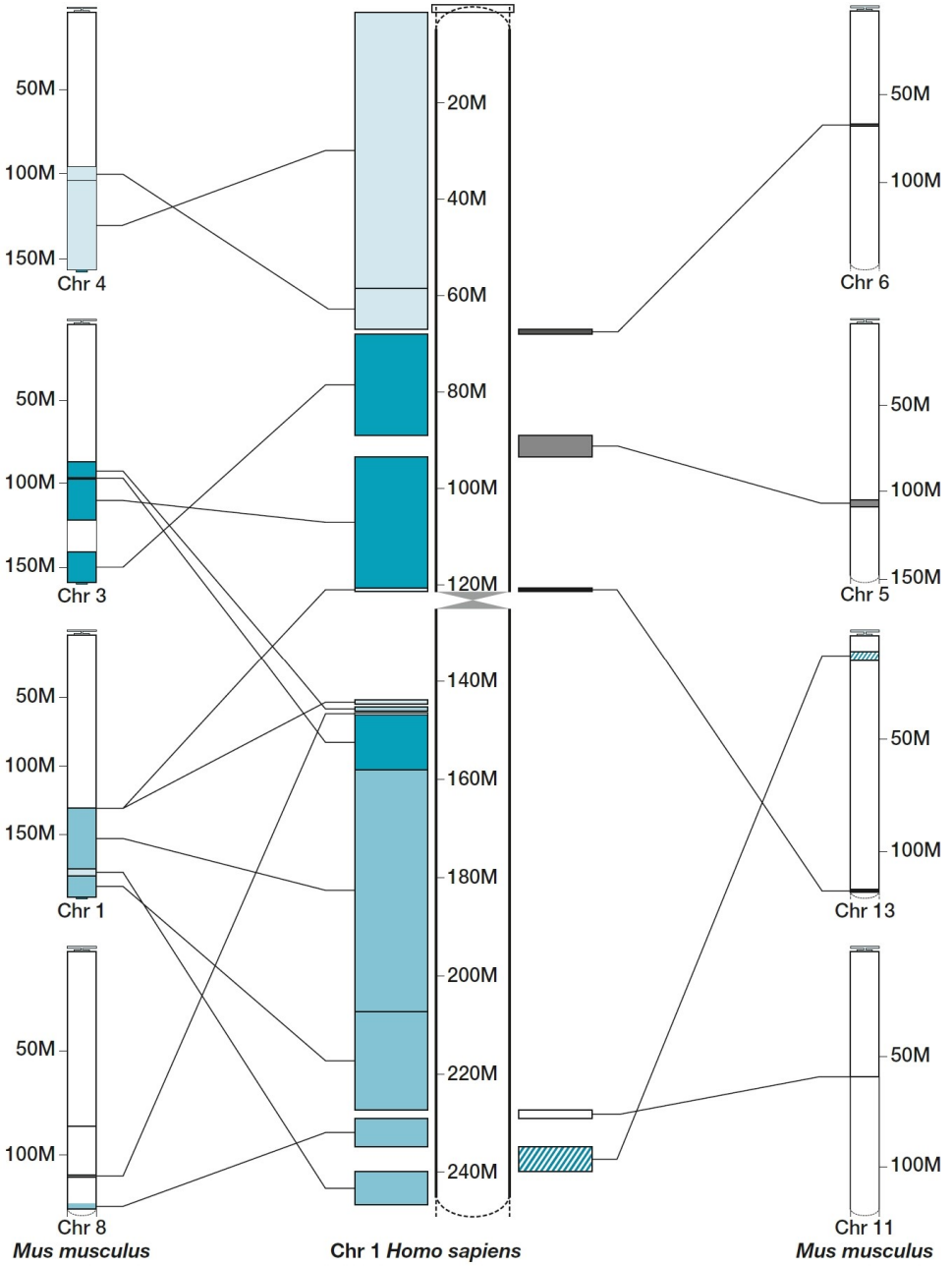
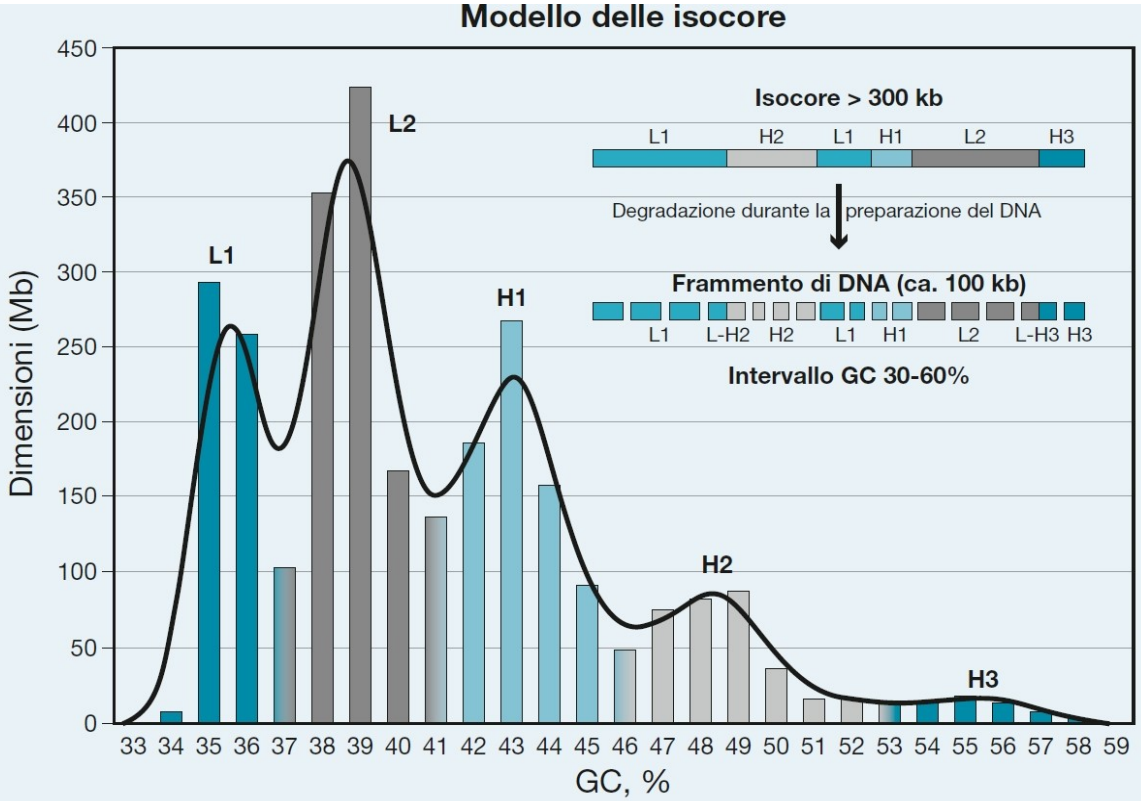


Figura 1.8
Dimensioni dei genomi procariotici ed eucariotici. (*Adattata da*: www.genomesize.com/statistics.php)

Figura 1.9
Mappa di sintenia del cromosoma 1 umano rispetto al genoma di topo che mostra come questo presenti omologia con estese regioni genomiche (> 100 kbp) di 8 diversi cromosomi di topo. Nelle regioni sinteniche si osserva una sostanziale conservazione dell'ordine genico. (*Adattata da*: Drillon G. e Fischer G., *Comptes Rendus Biologies*, 2011, 334(8-9):629-638. *Vedi anche* www.apps.webofknowledge.com/full_record.do?product=UA&search_mode=CitingArticles&qid=2&SID=N2gRjm5ADvkt5BMzr3r&page=1&doc=1)





Le isocore del genoma umano sono state identificare mediante frammentazione e successiva centrifugazione in gradiente di densità (*vedi* pannello in alto a destra). La maggior parte del genoma umano è costituita dalle isocore L1, L2 e H1 che hanno un contenuto in G+C compreso tra il 34 e il 46%, mentre le isocore più ricche in G+C (H2 e H3) sono quelle con la più alta densità genica. (*Fonte:* Maria Costantini *et al.*, *Genome Res.* 2006; 16: 536-541.)

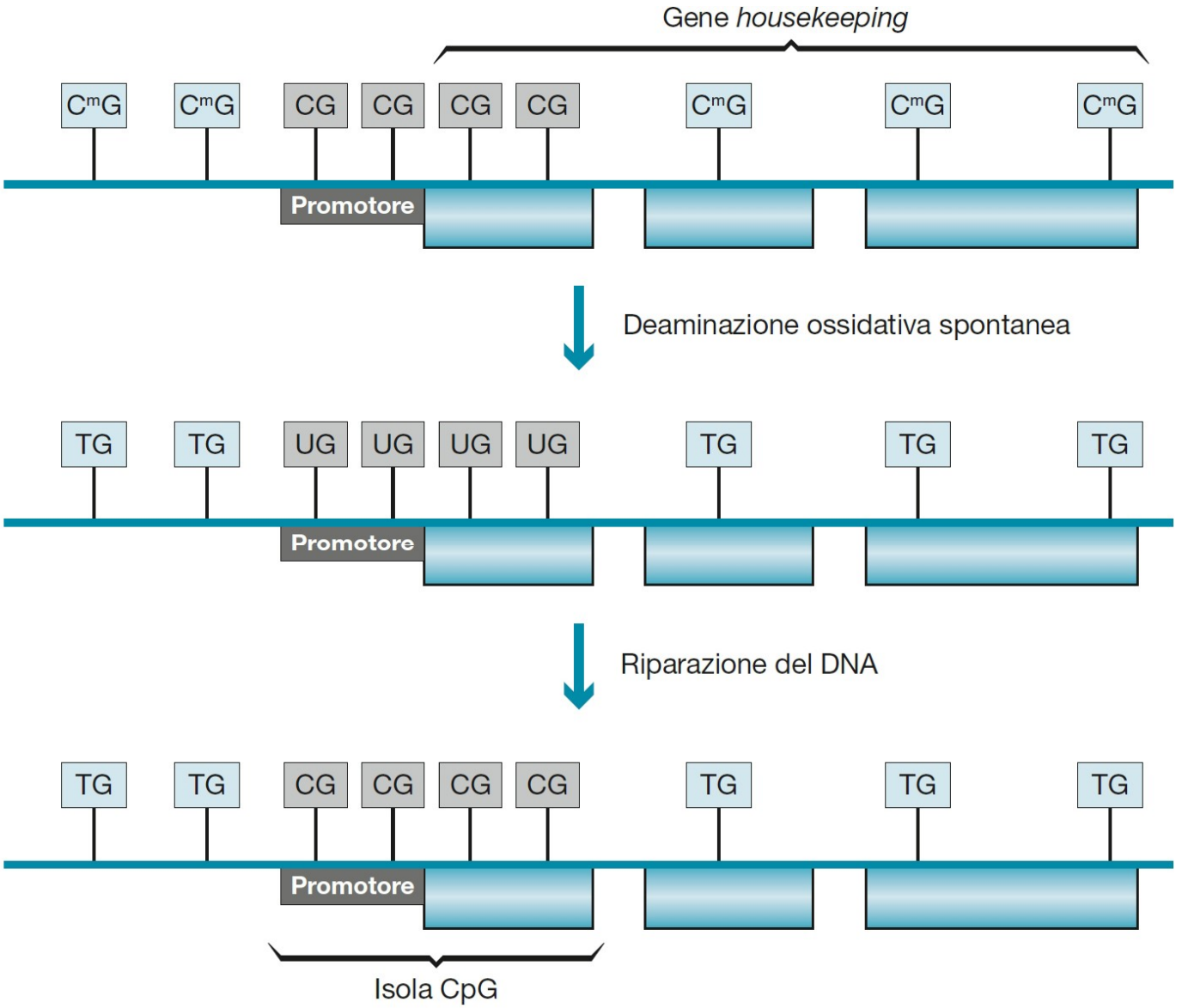


Figura 1.10
Le isole CpG, di lunghezza pari a 1-2 kpb, sono localizzate in corrispondenza del promotore dei geni *housekeeping* e corrispondono a tratti ipometilati del genoma.

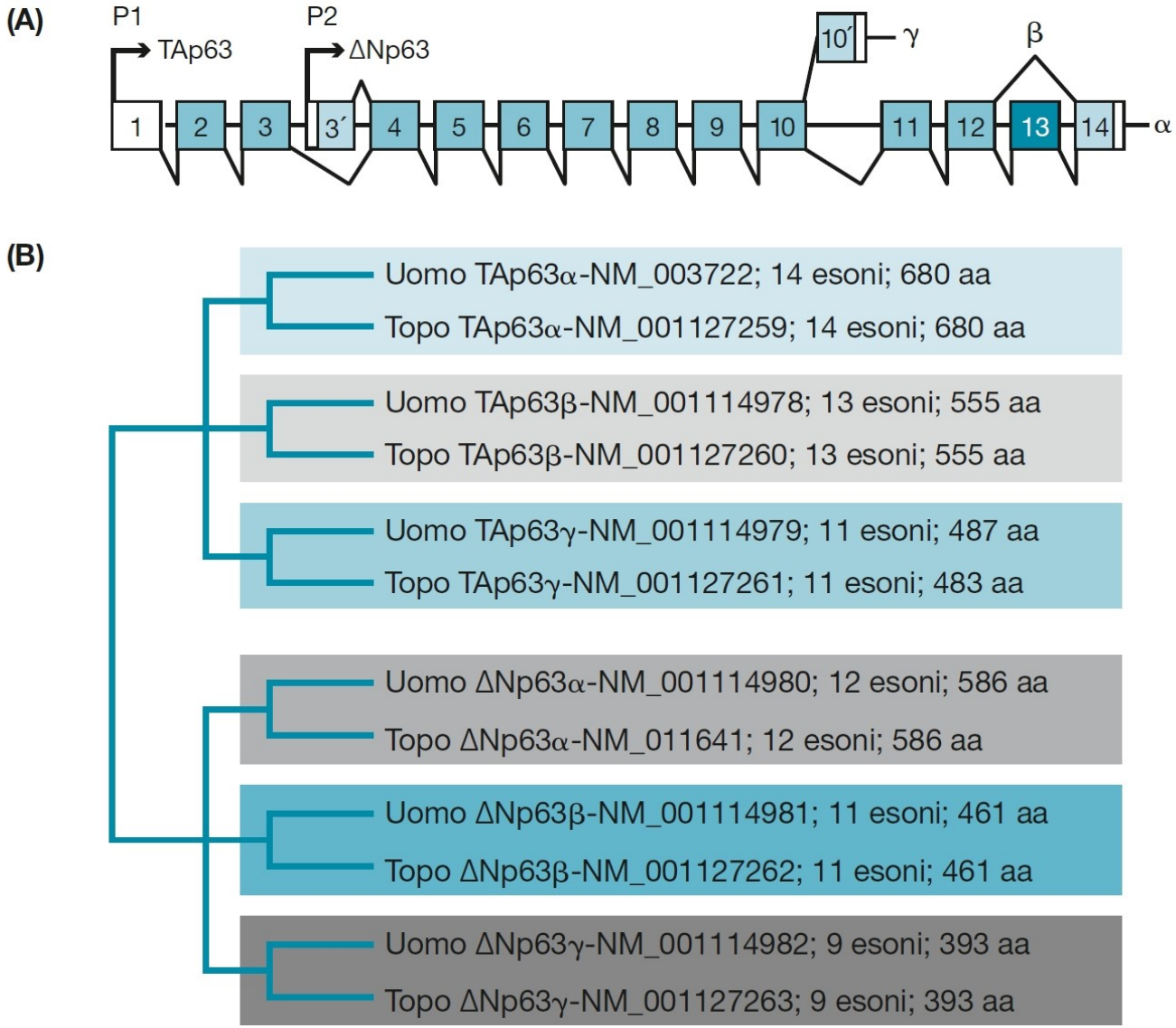


Figura 1.11
Splicing alternativo del gene *TP63* umano (**A**). Le isoforme osservate nell'uomo che danno origine a proteine diverse sono anche conservate nel topo (**B**).

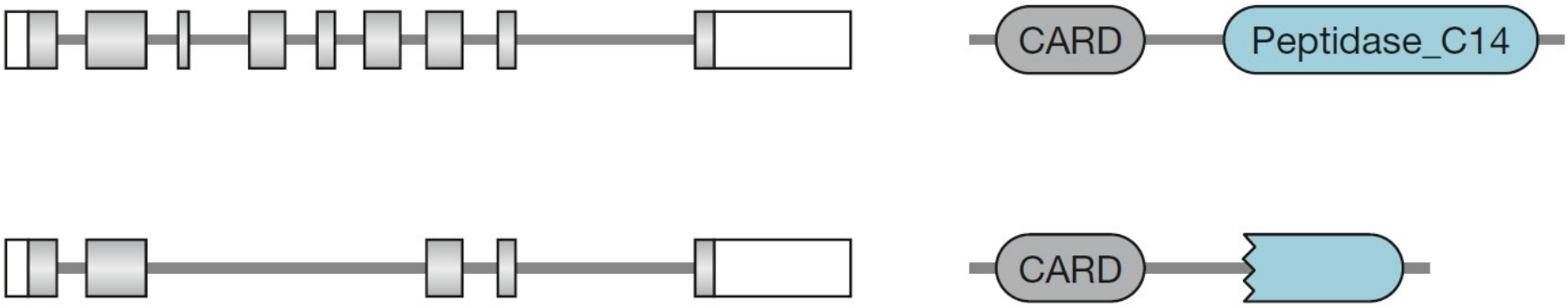


Figura 1.12
Il gene per la caspasi 9 può esprimere due proteine con funzioni antagoniste. La forma costitutiva della proteina (CASP9, 9 esoni, 416 aa) induce apoptosi. Essa contiene un Caspase recruitment domain (CARD) e un dominio caspasi Peptidase_C14. L'isoforma più corta della proteina (CASP9S, 5 esoni, 266 aa) contiene un dominio Caspase recruitment domain (CARD) e un dominio tronco della Peptidase_C14. Questa isoforma è priva dell'attività proteasica e agisce da inibitore dell'apoptosi.

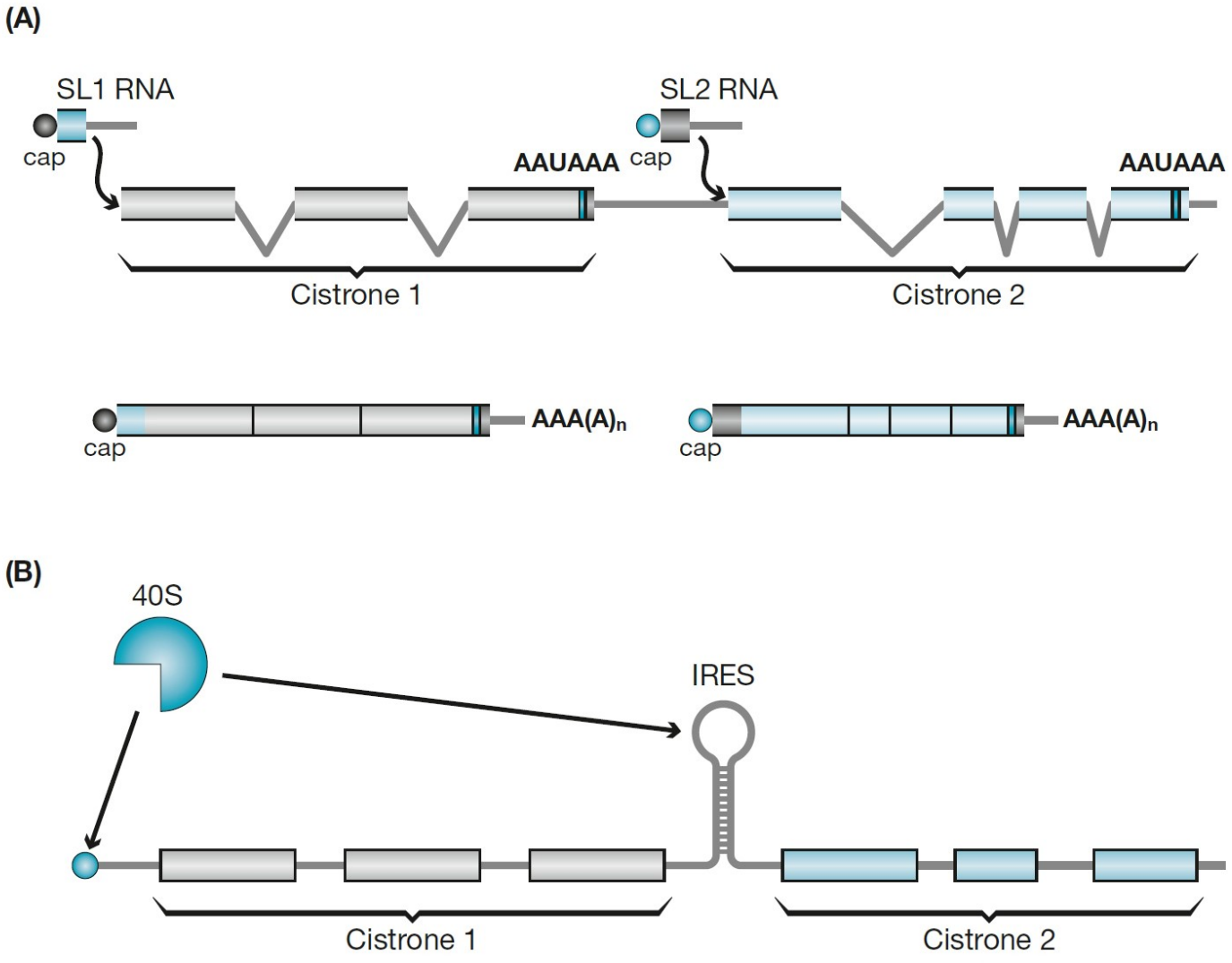


Figura 1.13
(**A**) Meccanismo del *transplicing* attraverso il quale un trascritto policistronico viene maturato in mRNA mediante l'aggiunta di un piccolo trascritto leader (SL1, SL2) dotato di cap.
(**B**) Traduzione cap-indipendente mediata dal legame della subunità minore del ribosoma (40S) all'elemento IRES.

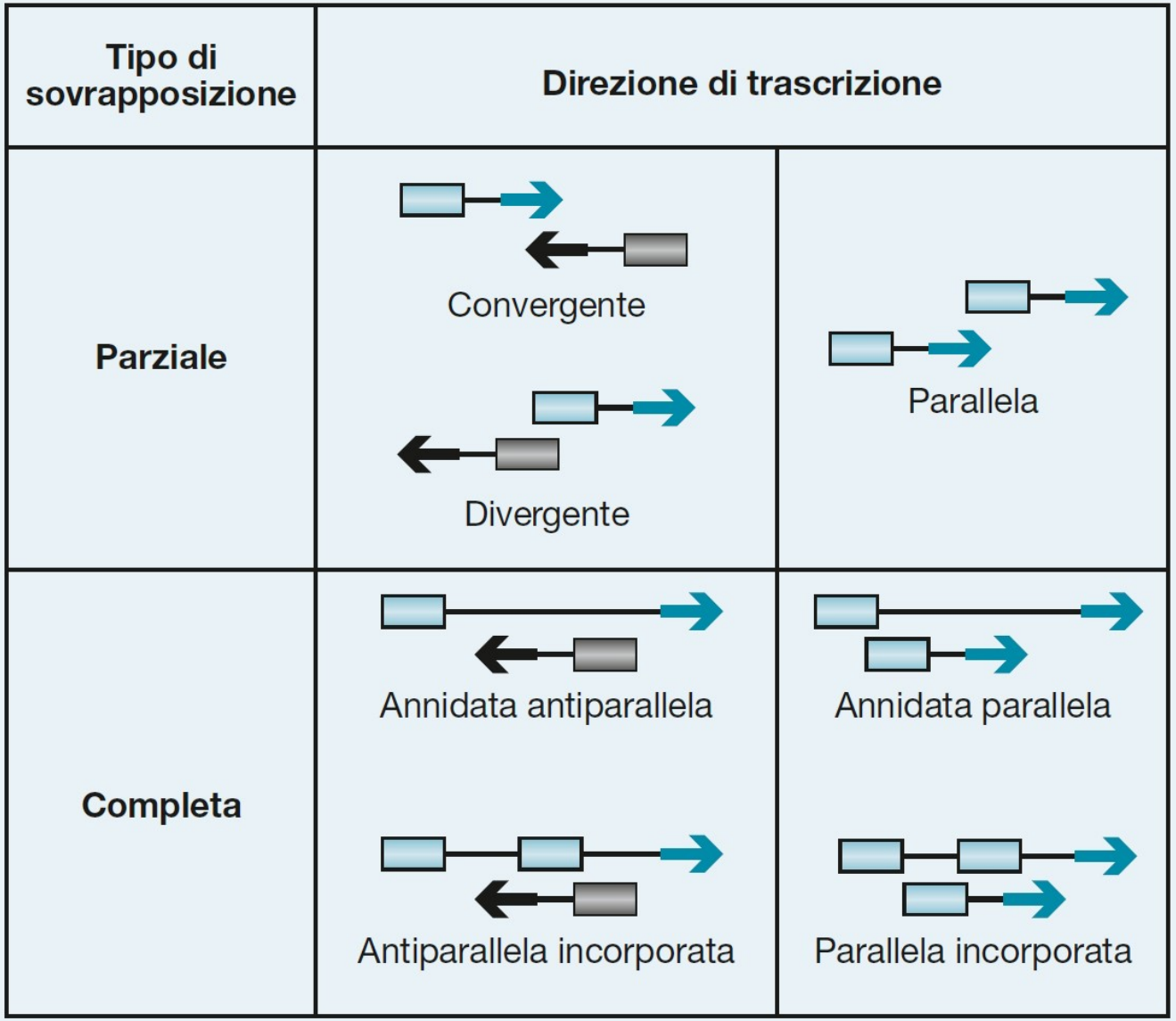


Figura 1.14
I geni eucariotici sono localizzati su entrambi i filamenti di DNA e possono essere sovrapposti in molti modi diversi.

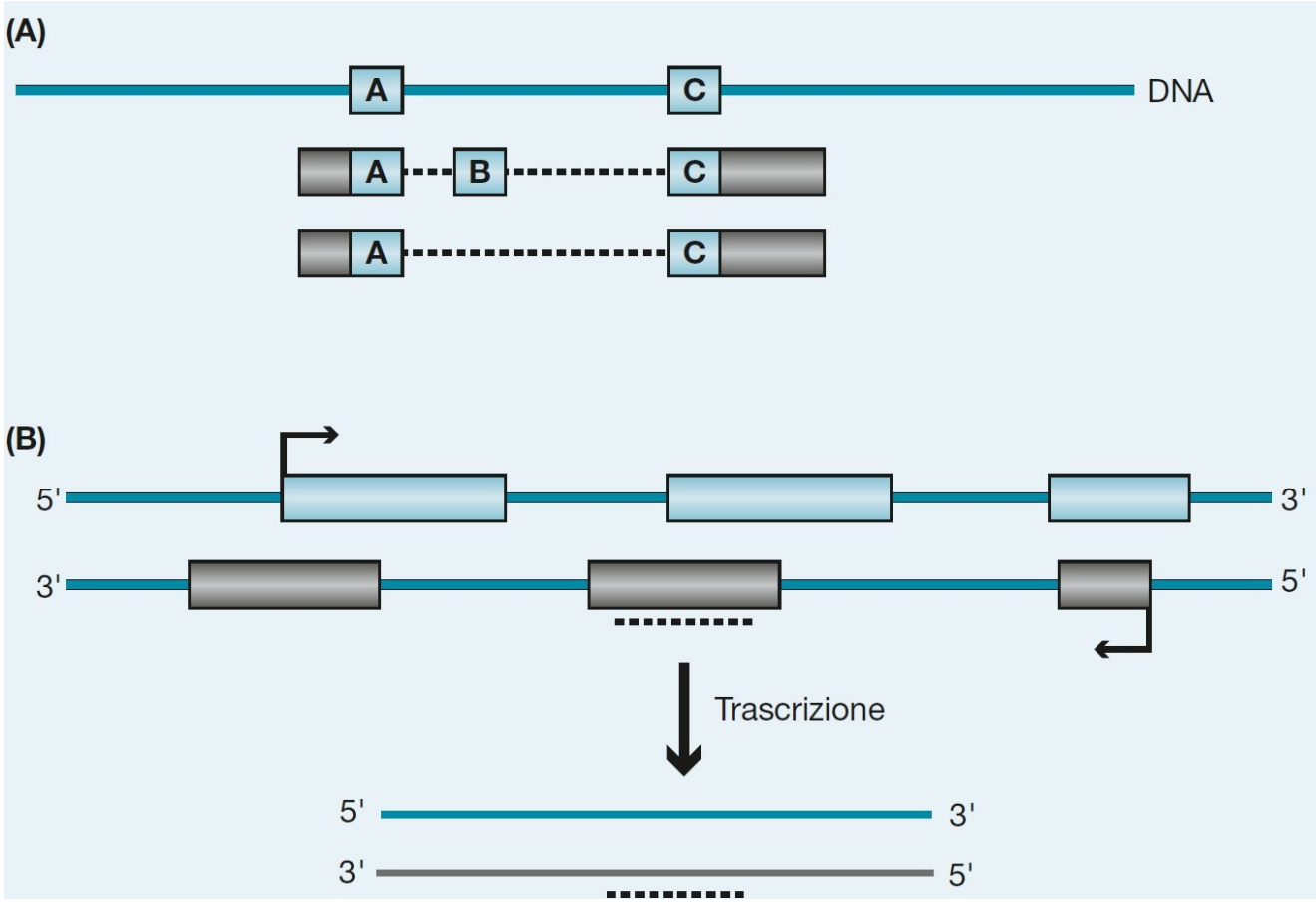


Figura 1.15
(**A**) Due trascritti alternativi possono essere assegnati a uno stesso gene se le loro proiezioni sul genoma, limitatamente alle regioni corrispondenti al prodotto funzionale, sono sovrapposte anche se parzialmente; in questo esempio le porzioni codificanti del primo e terzo esone (regioni A e C). (**B**) Una read generata in un esperimento di sequenziamento RNA-seq (segmento grigio tratteggiato) può essere assegnata a uno dei due geni sovrapposti localizzati sui filamenti complementari solo se ottenuta con un kit direzionale.

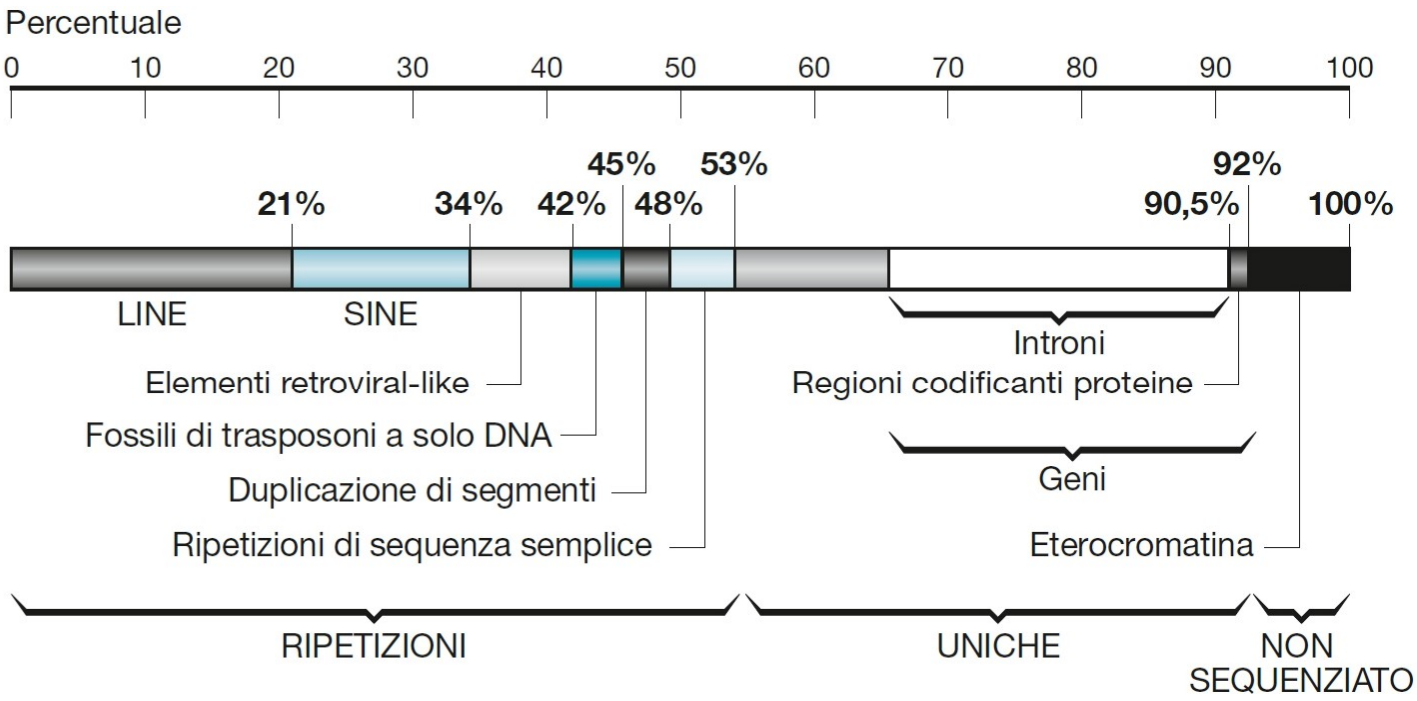


Figura 1.16
Composizione del genoma umano.

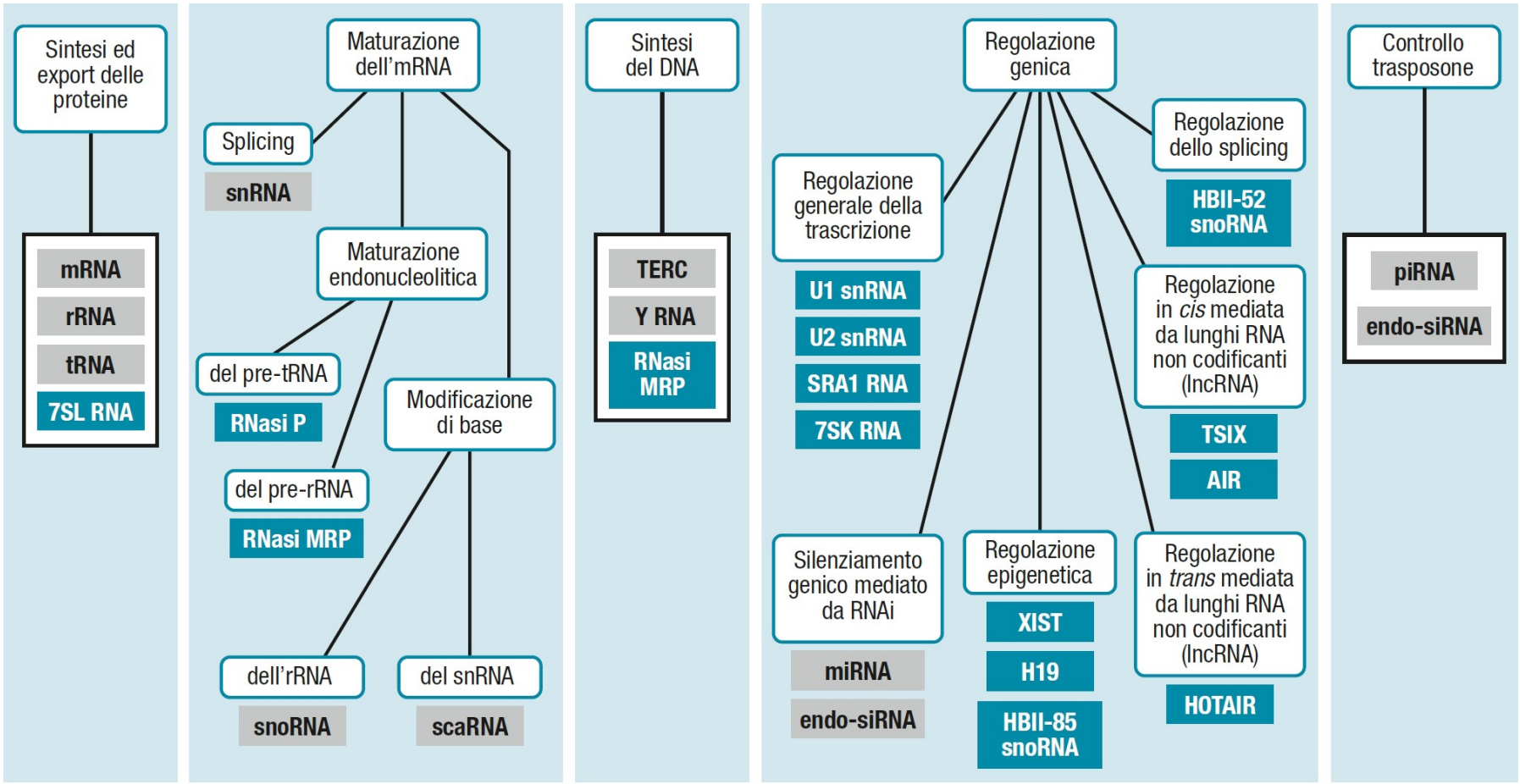


Figura 1.17
Funzioni dei sncRNA (per es. snoRNA, scaRNA, miRNA ecc.) e dei lncRNA (per es. XIST, HOTAIR ecc.).
(*Adattata da*: Strachan T. e Read A., *Genetica molecolare umana*, Zanichelli, Bologna, 2012)

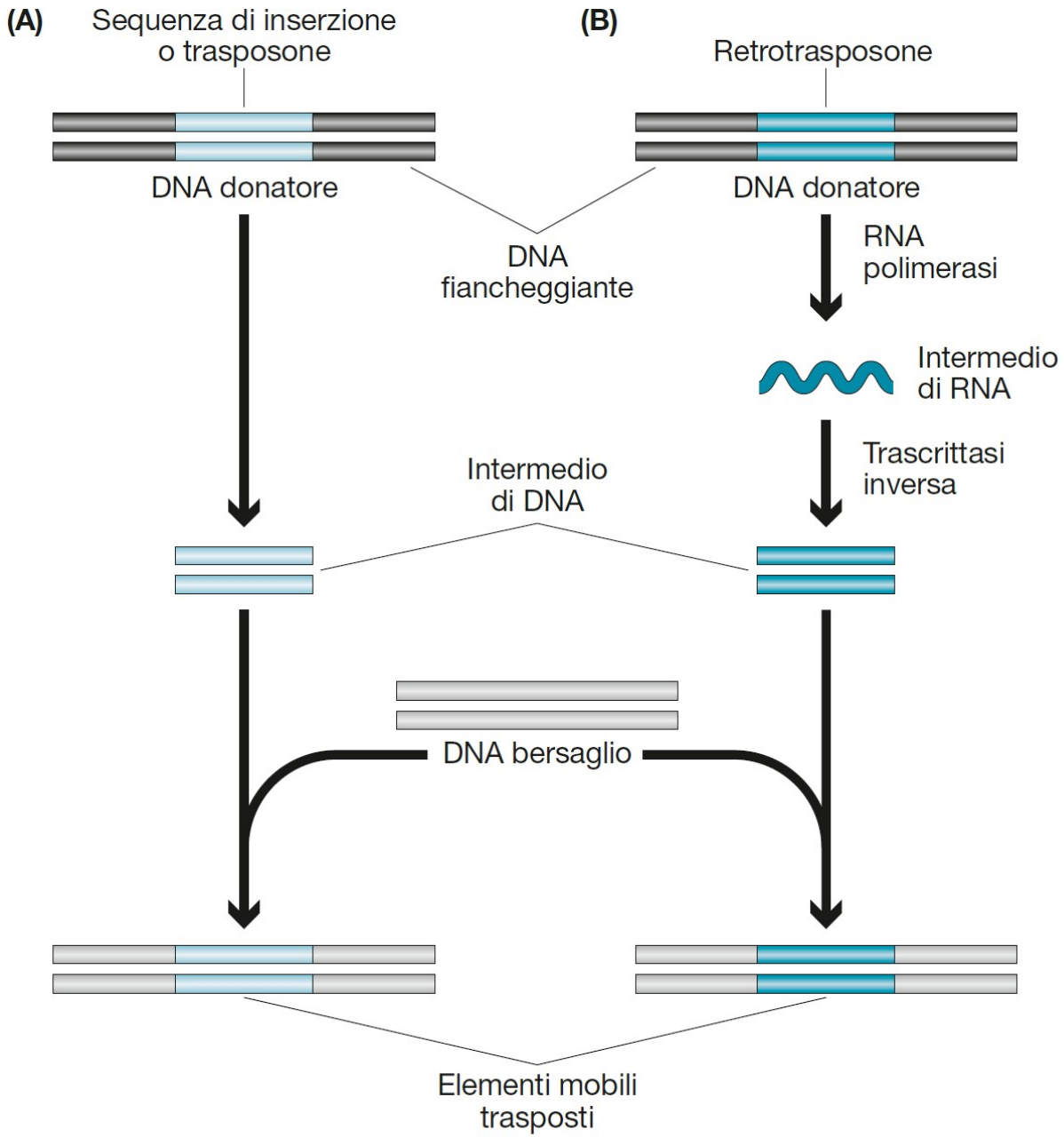


Figura 1.18
Le ripetizioni intersperse del genoma umano possono utilizzare un intermedio a DNA (**A**) o a RNA (**B**).

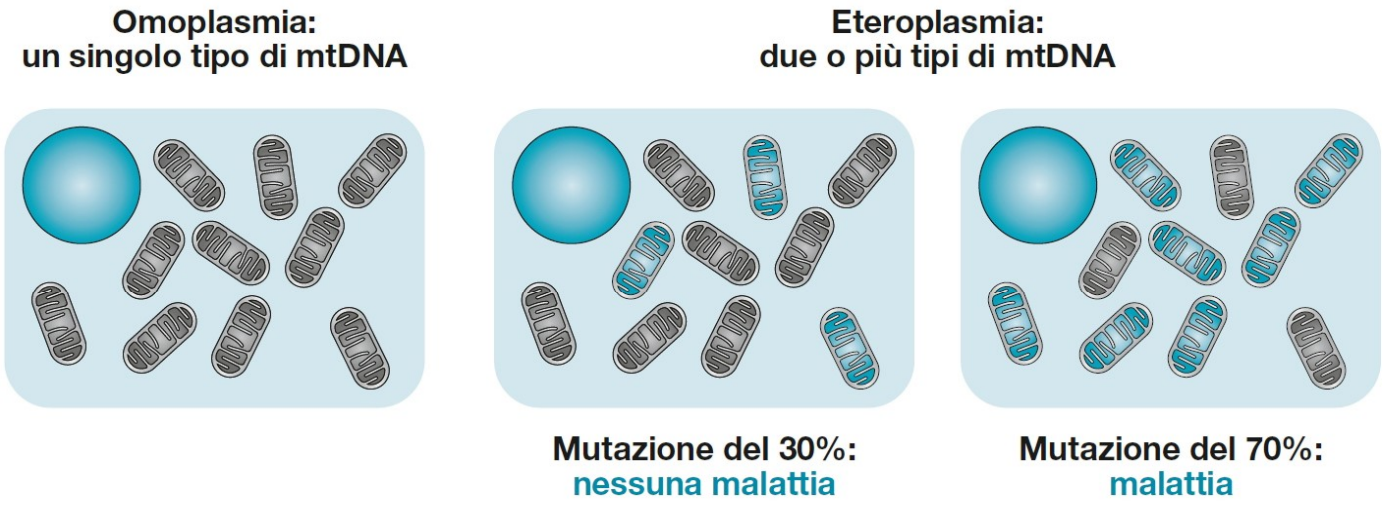


Figura 1.19
Omoplasmia, eteroplasmia ed effetto soglia nelle patologie mitocondriali.

Capitolo 3

La statistica essenziale

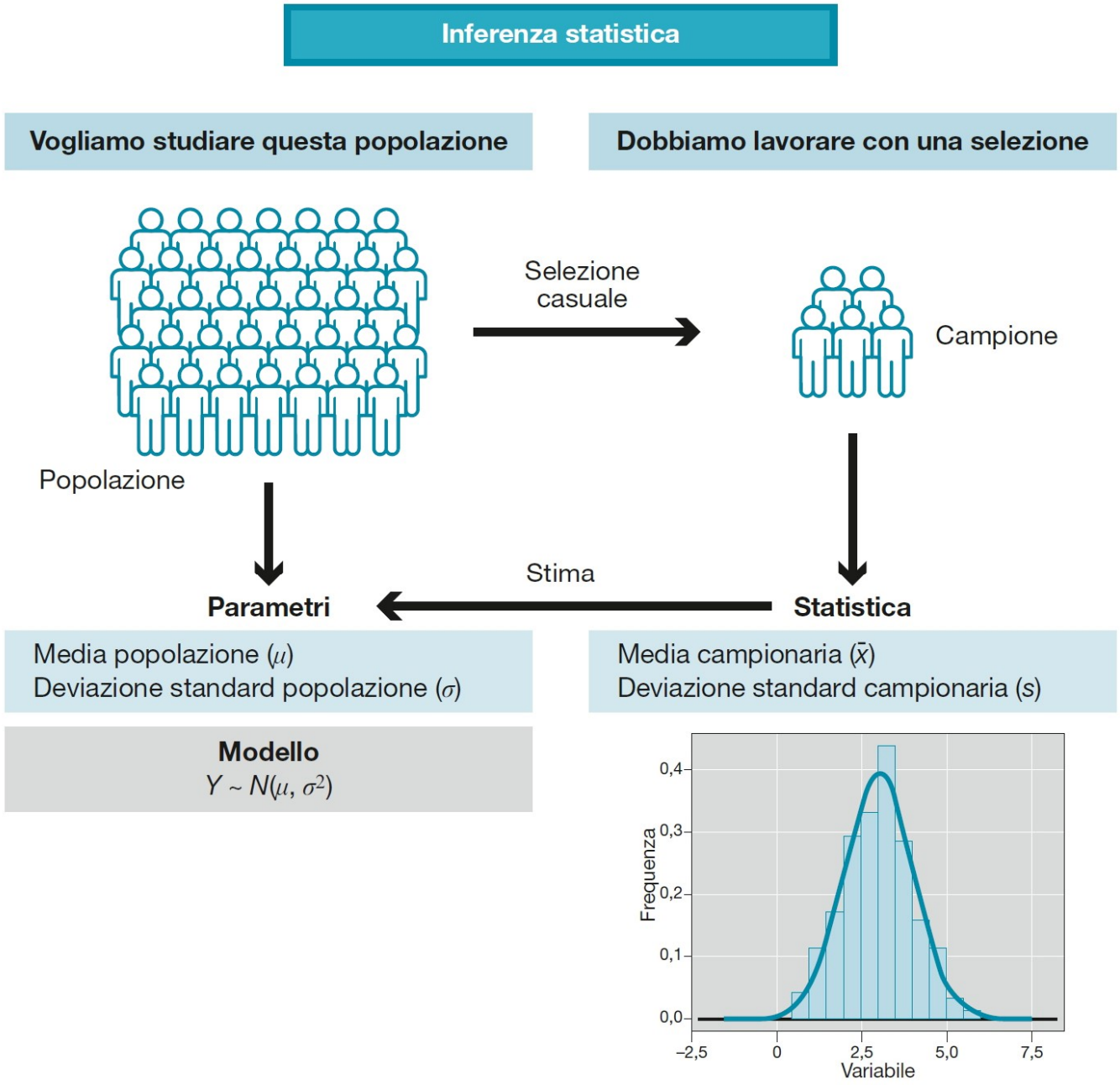


Figura 3.1
Schema del processo di inferenza e induzione. Dalla popolazione di riferimento si seleziona un campione su cui viene misurata la variabile di interesse. Sul campione si stimeranno poi la distribuzione campionaria, la media e la deviazione standard per fare verifica d'ipotesi sui parametri della popolazione che sono parametri non noti.
La variabile di interesse Y nella popolazione di riferimento si assume essere distribuita (~) come una gaussiana con media μ e varianza σ^2, $N(\mu, \sigma^2)$.

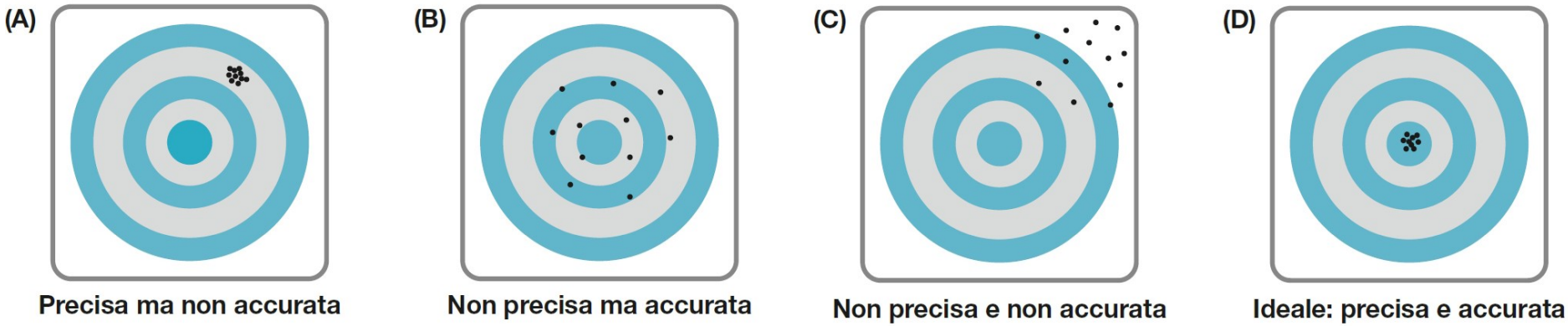


Figura 3.2
Precisione e accuratezza di una stima.

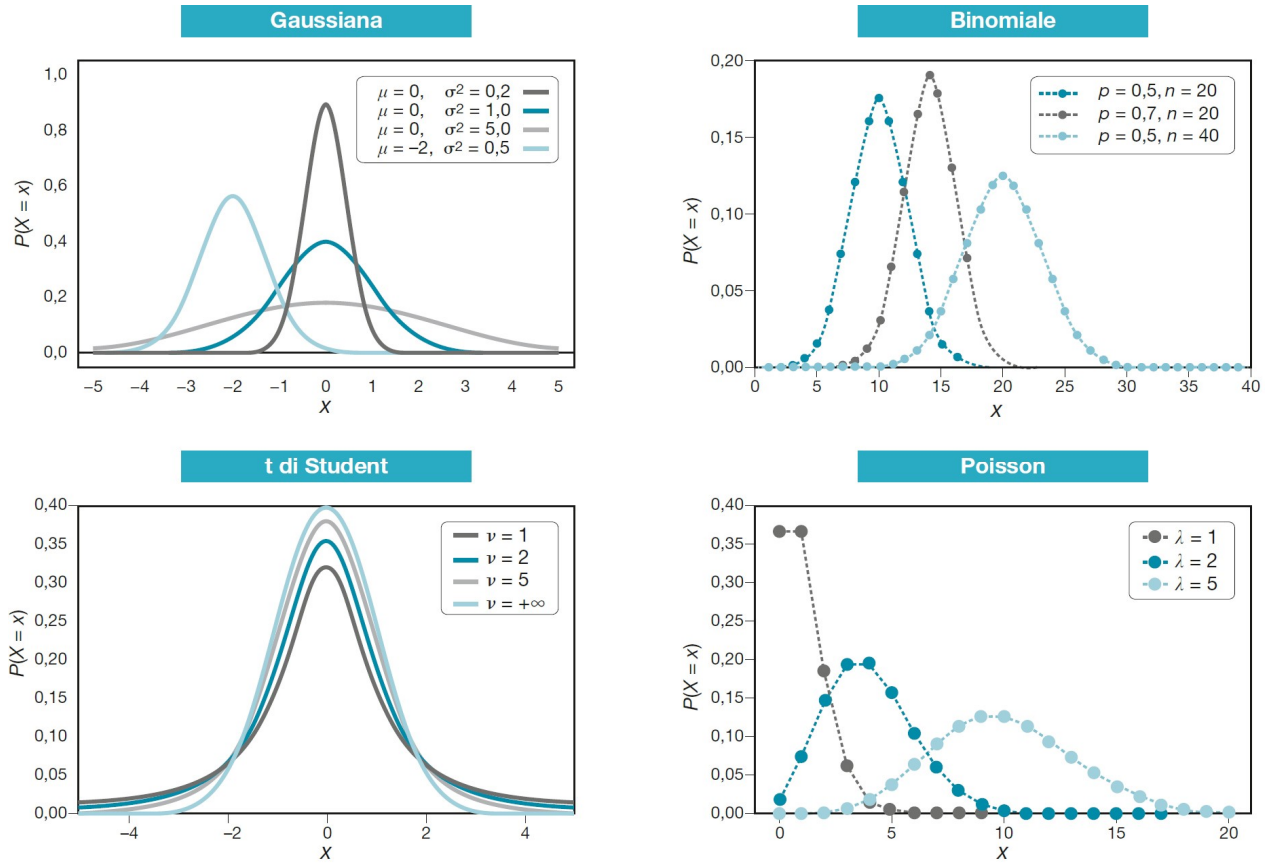


Figura 3.3
Principali distribuzioni teoriche continue Gaussiana, t di Student, e discrete Binomiale e Poisson. Ognuna di queste distribuzioni dipende da parametri che ne definiscono la posizione, la variabilità e la simmetria. Si noti come nelle distribuzioni discrete i valori assunti dalla variabile sono solo quelli relativi ai valori segnati dai pallini colorati.

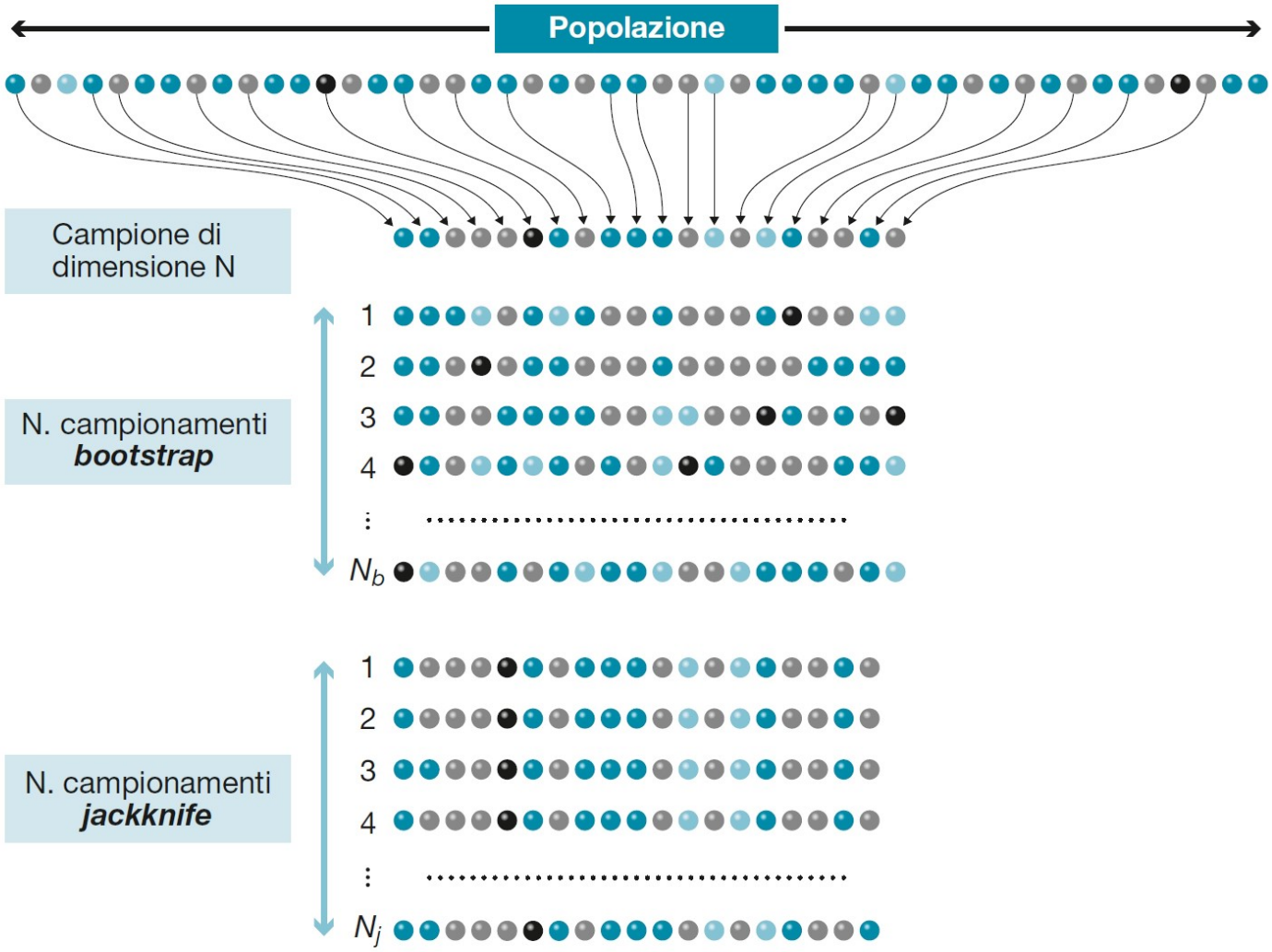


Figura 3.4
Metodi di ricampionamento *bootstrap* e *jackknife*.

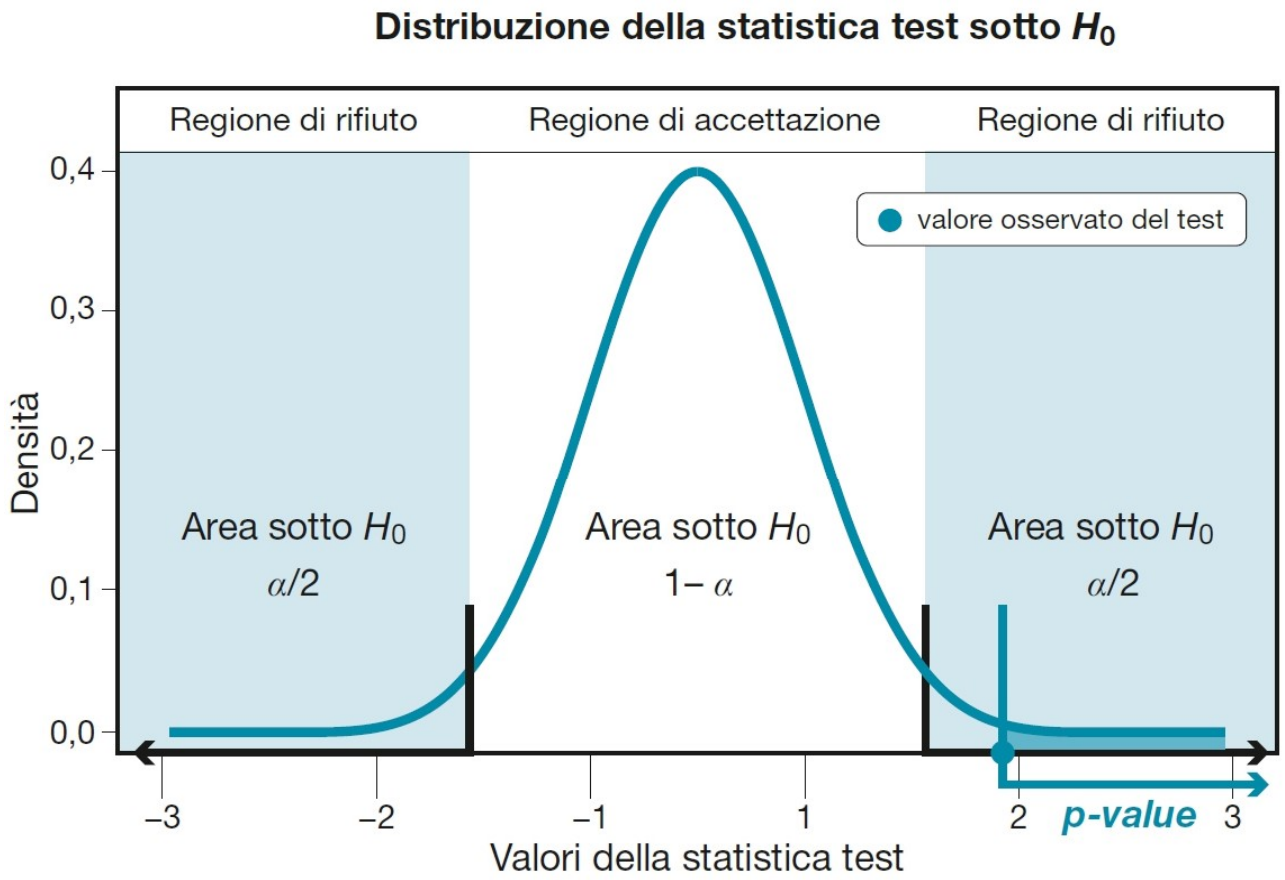


Figura 3.5
Processo di inferenza. Sulla distribuzione del test sotto H_0, si definiscono sulla base del valore fissato di α, la regione di rifiuto e la regione di accettazione. La posizione del valore del test osservato (T_{oss}) rispetto a queste regione determina il rifiuto o meno di H_0. Il valore osservato del test determina inoltre l'area della curva sotto H_0, definita come *p-value*.

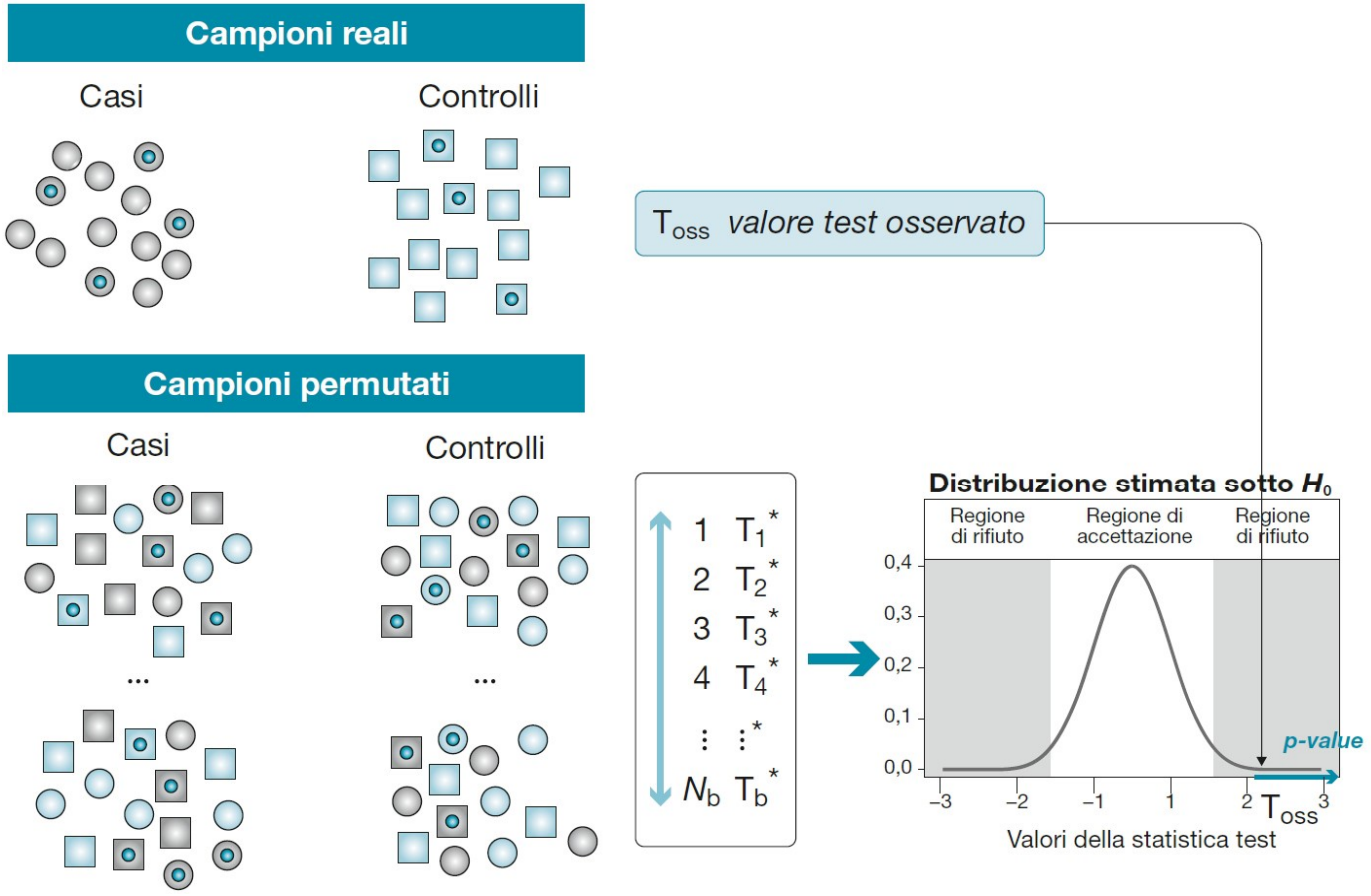


Figura 3.6
Approccio dei test di permutazione. Per ognuno dei campioni virtuali ottenuti riassegnando casualmente le classi alle unità statistiche, si calcolano le statistiche test T*. L'insieme di queste T* determina la stima della distribuzione sotto H_0. La posizione del valore osservato della statistica test su questa distribuzione determina il rifiuto o meno di H_0.

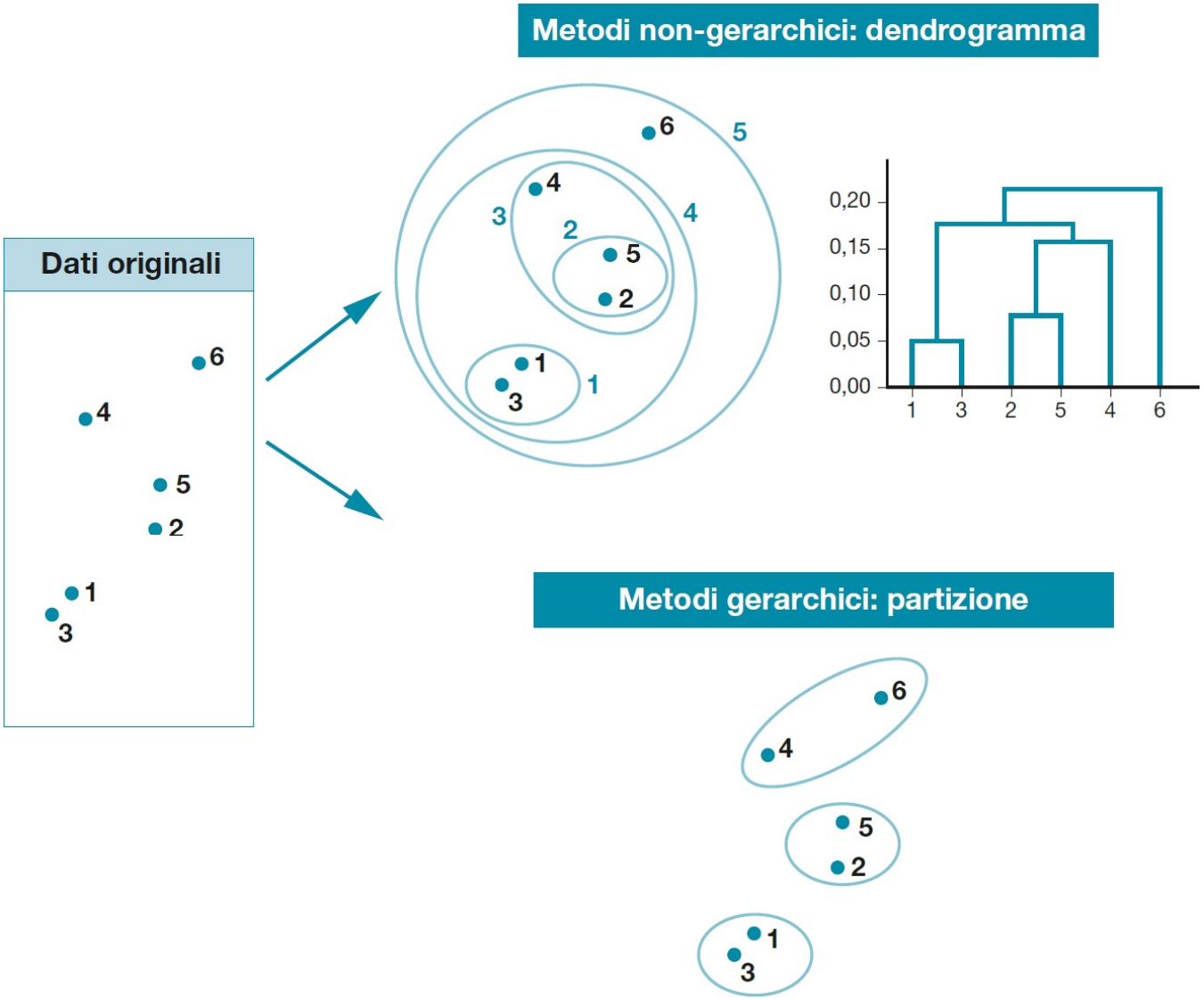


Figura 3.7
Differenza tra metodi di raggruppamento gerarchico e non-gerarchico.

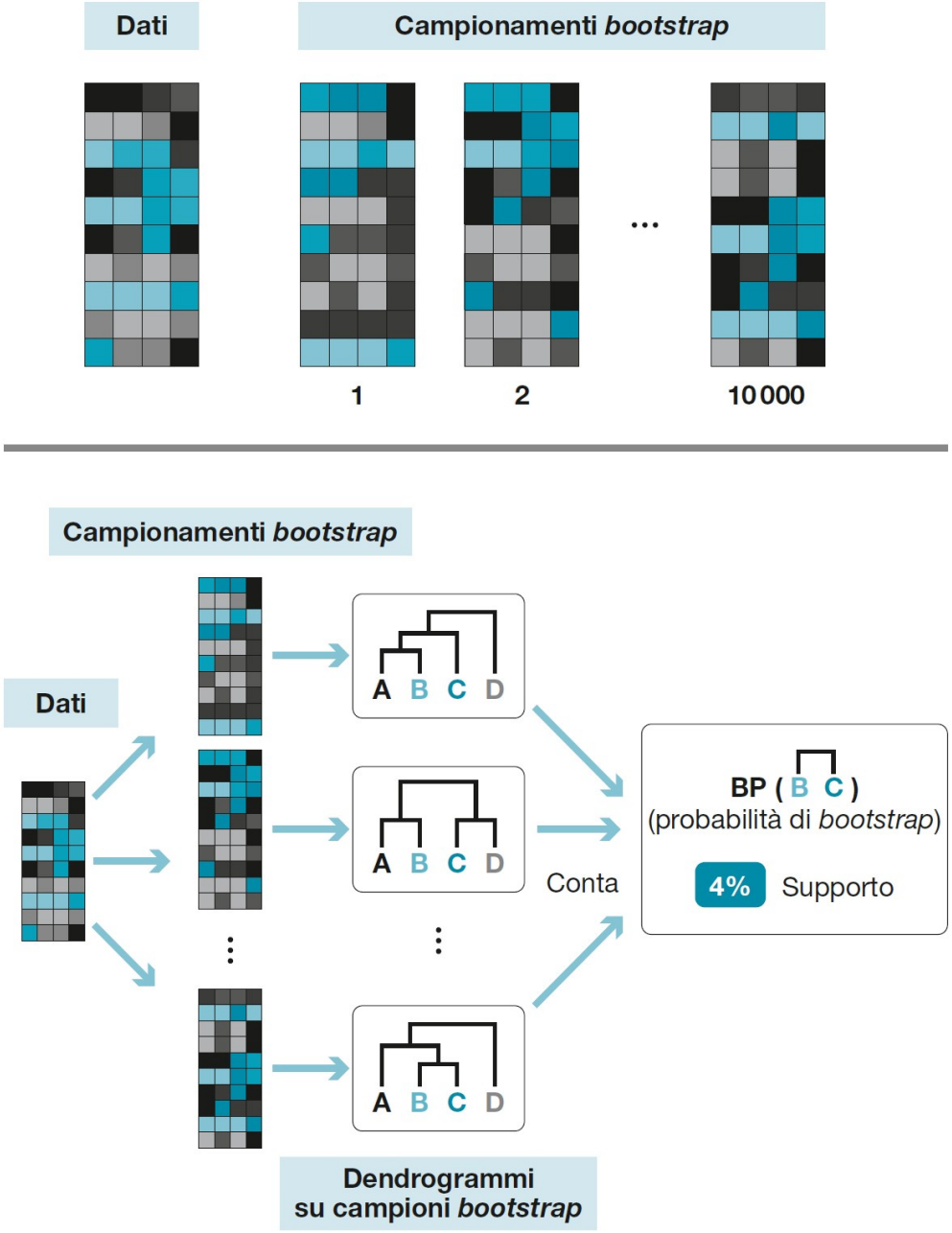


Figura 3.8
Supporto di *bootstrap* per raggruppamenti. In questo caso la matrice dei dati ha le osservazioni sulle colonne e le variabili sulle righe. Dalla matrice originale si generano 10000 matrici *bootstrap* in cui vengono ricampionate le variabili. Per ognuna delle matrici viene stimato il dendrogramma sulle unità A, B, C e D. Gli alberi vengono confrontati contando quante volte il gruppo B-C è raggruppato insieme. Questo numero diviso per il numero di campionamenti *bootstrap* definisce il supporto.

Capitolo 4

L’evoluzione biologica

Figura 4.1
Tipi di mutazione che intervengono nel corso del processo di evoluzione molecolare. A partire da una sequenza progenitore (**A**) possono verificarsi: sostituzioni puntiformi (**B**); delezioni (**C**); inserzioni (**D**); inversioni (**E**). Questo fa sì che la sequenza discendente possa risultare differente da quella ancestrale anche nella lunghezza (**F**).

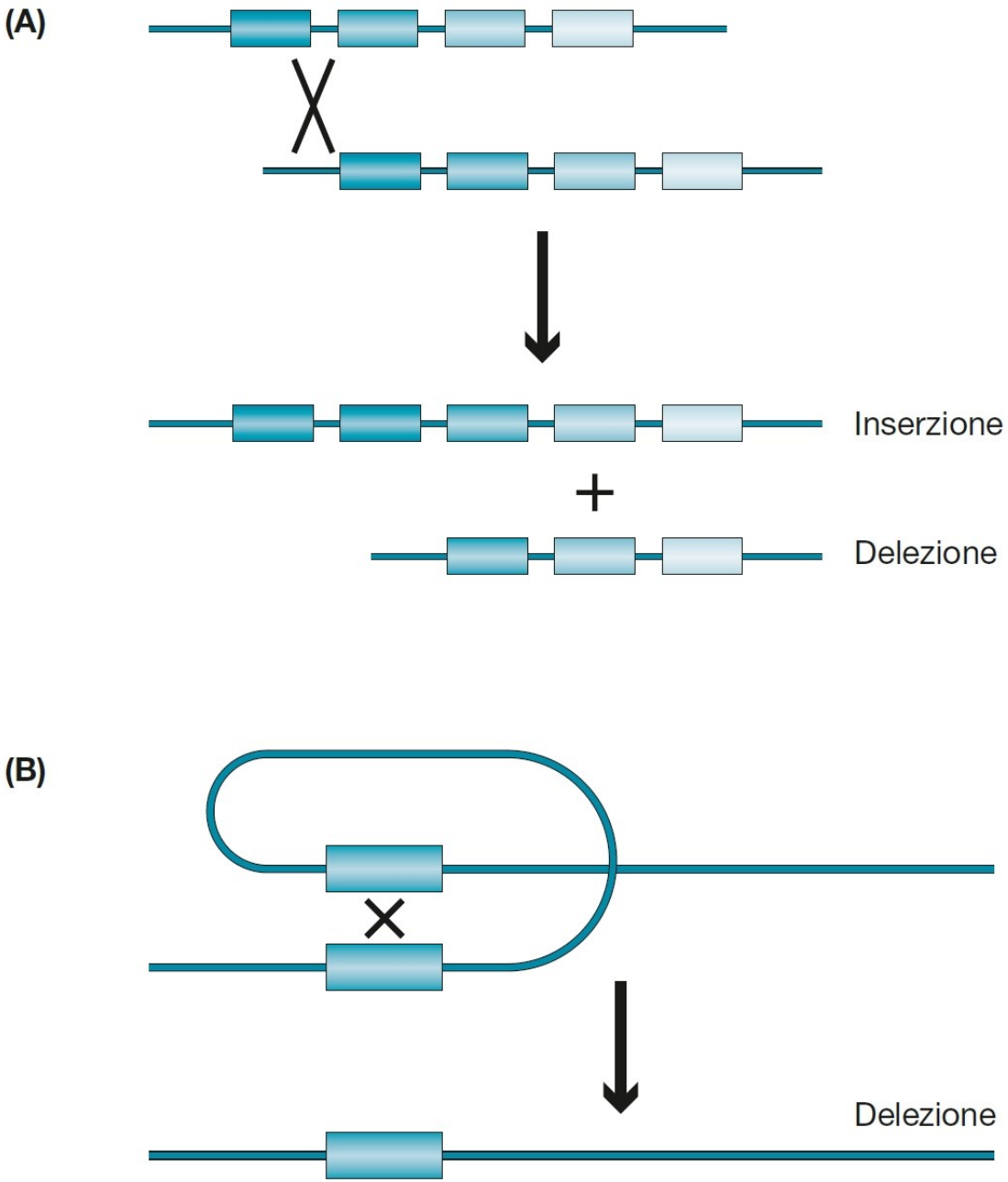


Figura 4.2
Generazione di inserzioni e delezioni attraverso il meccanismo del *crossing-over* disuguale mediato da segmenti ripetuti (rappresentati da rettangoli con colore uguale) nel caso di appaiamento intercromosomico (**A**) e intracromosomico (**B**).

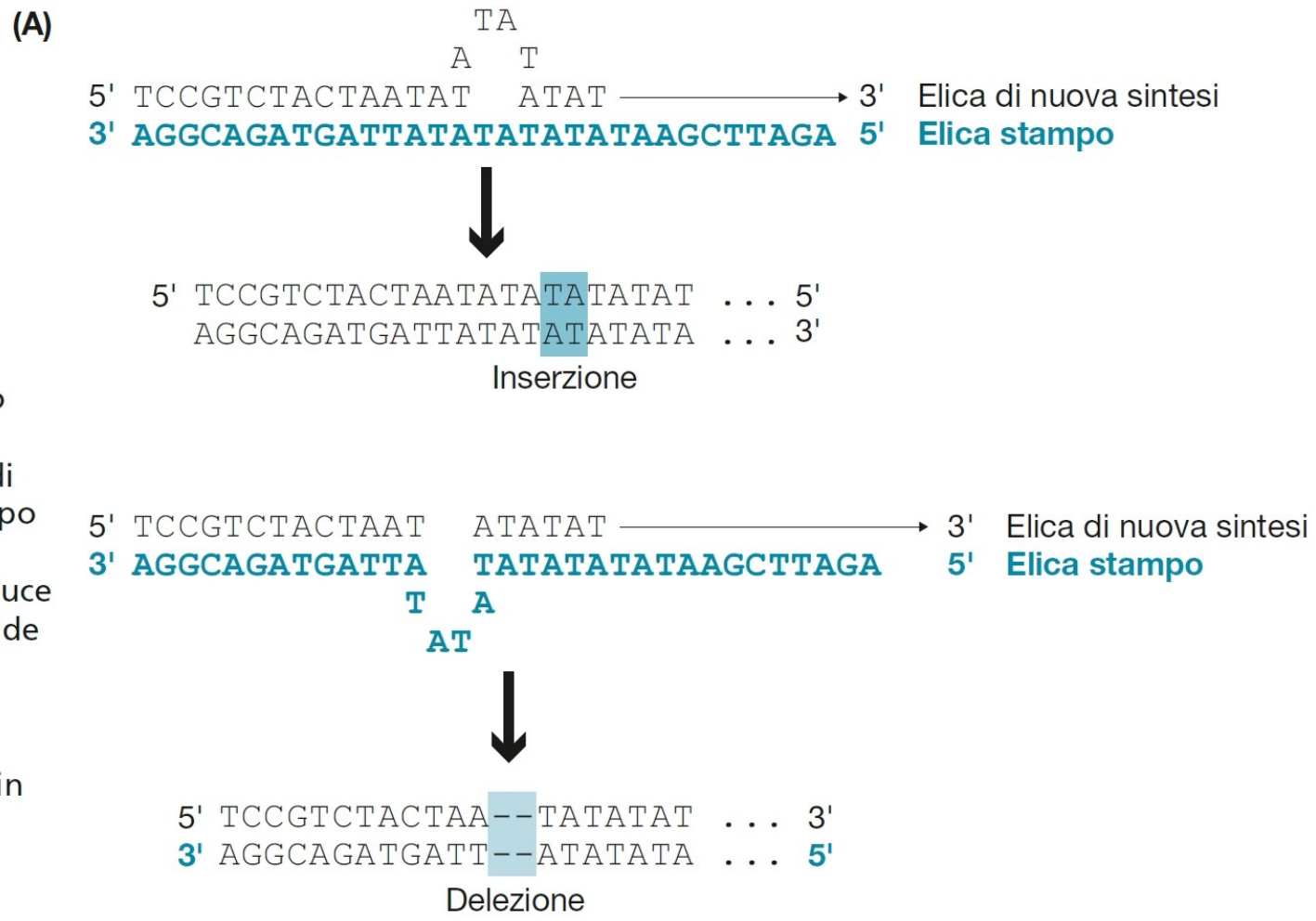


Figura 4.3
Generazione di piccole inserzioni o delezioni attraverso il meccanismo dello *slippage*. Nell'esempio qui riportato si osserva un misappaiamento dell'elica di nuova sintesi sull'elica stampo dovuto alla presenza di un microsatellite (TA) che produce l'inserzione di un dinucleotide TA in una delle due eliche figlie (**A**). Allo stesso modo il misappaiamento può riguardare l'elica stampo e in questo caso si produce una delezione in una delle due eliche figlie (**B**).

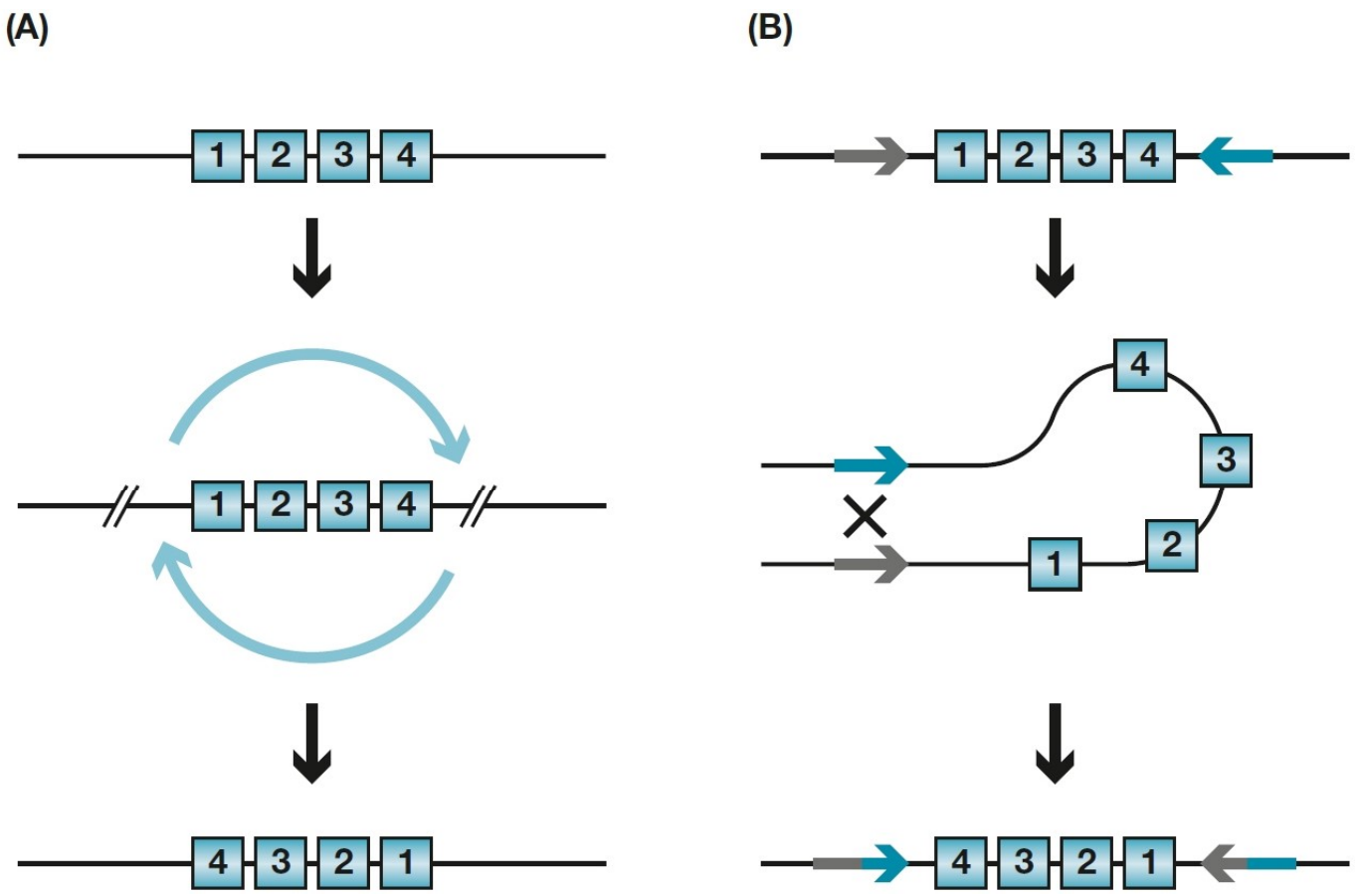


Figura 4.4
Generazione di inversioni per escissione, inversione e ricongiunzione di un segmento cromosomico (**A**) o per ricombinazione omologa intracromosomica mediata da elementi ripetuti (rappresentati da frecce) (**B**).

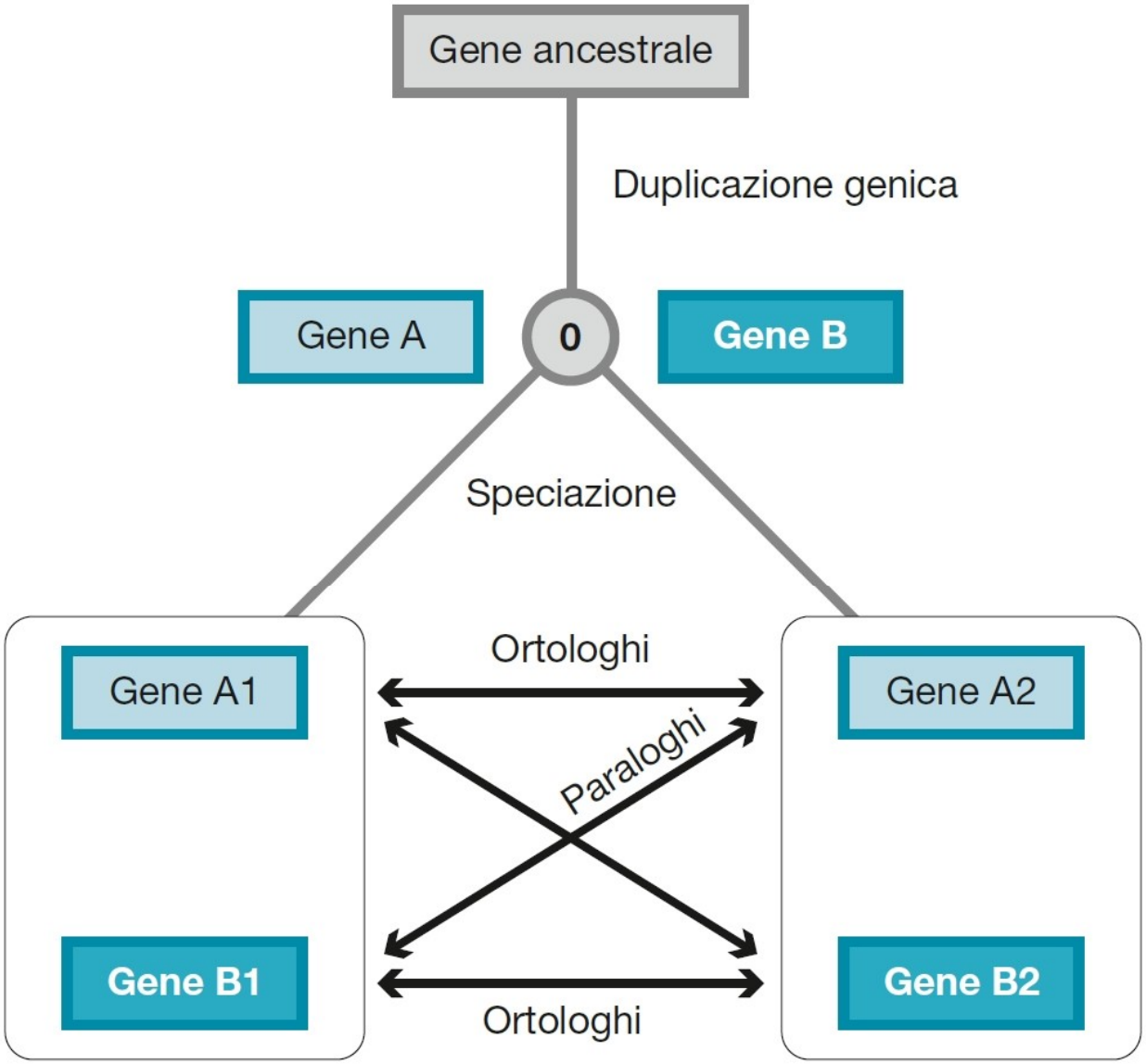


Figura 4.5
Relazioni tra geni ortologhi e paraloghi generati in un processo evolutivo a partire da un gene ancestrale, con una duplicazione genica nella specie 0, che genera i geni A e B, seguita da una speciazione che origina i geni A1 e B1 nella specie 1 e A2 e B2 nella specie 2.

Capitolo 5

Allineamenti tra sequenze

	–	A	G	A	T	T	C	C	A	T
–	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	0	1	0
G	0	1	2	2	2	2	2	2	2	2
T	0	1	2	2	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4
C	0	1	2	2	3	3	4	5	5	5
C	0	1	2	2	3	3	4	5	5	5
A	0	1	2	3	3	3	3	3	6	6
T	0	1	2	3	4	4	4	4	6	7

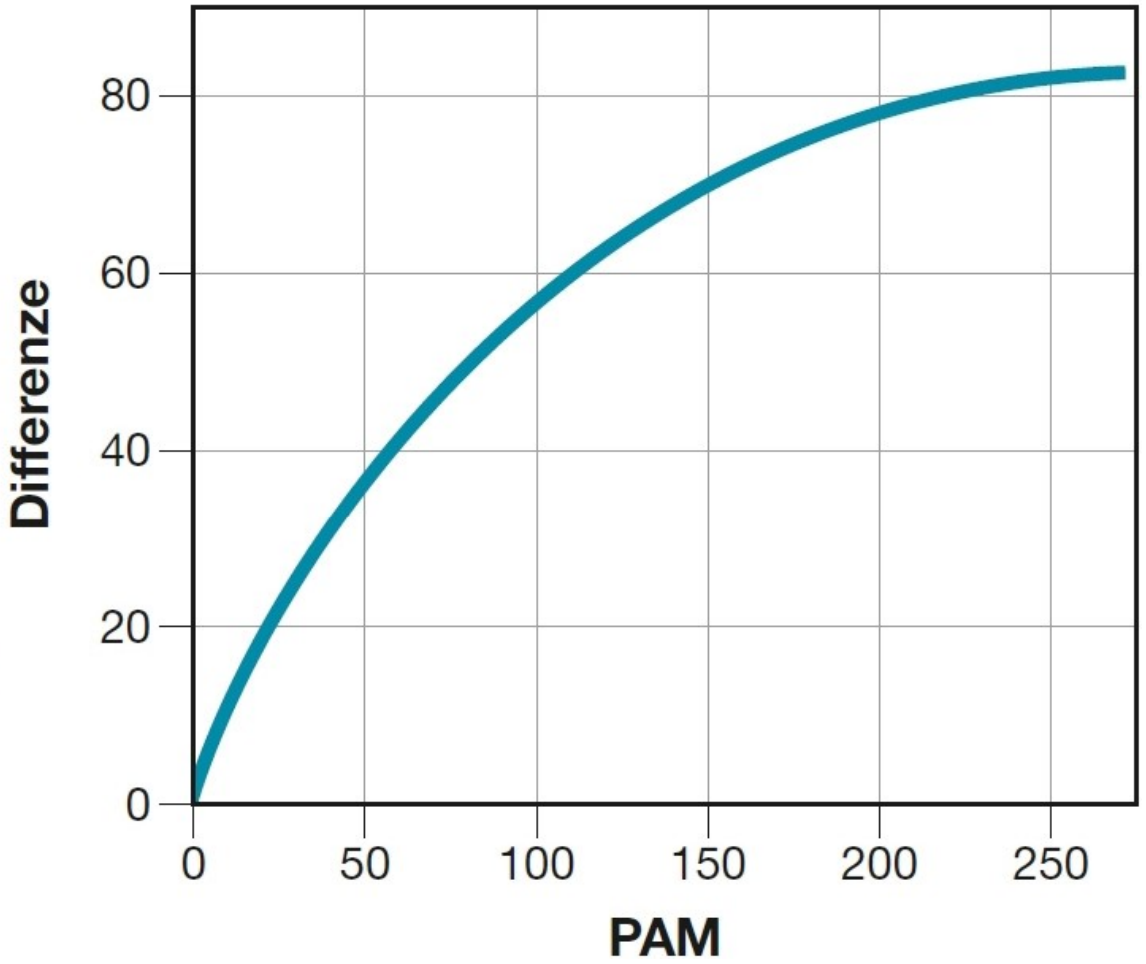
Figura 5.1
Per l'allineamento di due sequenze di lunghezza m e n si può costruire una matrice di dimensione $m + 1$, $n + 1$ e riempirla come spiegato nel testo. Le sequenze usate come esempio sono le stesse dell'allineamento (1).

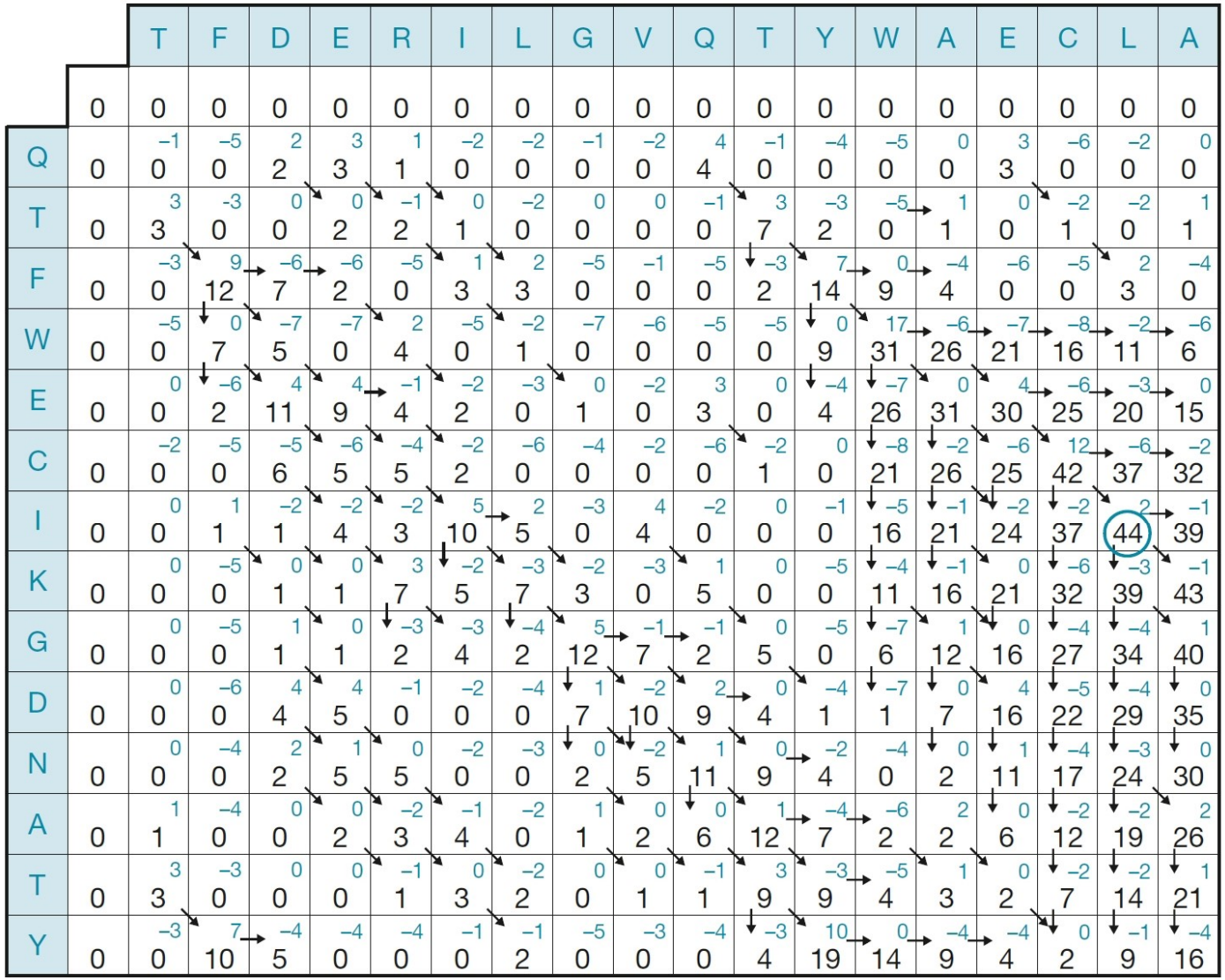
Figura 5.2
Matrice di sostituzione PAM250. Le righe e le colonne descrivono i valori di sostituzione calcolati per ogni possibile coppia di residui. I valori più alti di ogni riga e di ogni colonna sono quelli che si trovano sulla diagonale, in quanto corrispondono alla sostituzione di un residuo con se stesso. È importante notare che ci sono residui per i quali il valore sulla diagonale è particolarmente alto (per es. W e C), in quanto la loro frequenza nelle proteine è bassa e le loro particolari proprietà li rendono particolarmente difficili da sostituire: il triptofano in quanto particolarmente voluminoso e idrofobico, la cisteina a causa dei ponti disolfuro che è in grado di formare. La matrice PAM250 è idealmente usata per allineare sequenze molto divergenti (~20% di identità).

PAM250

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	2	–2	0	0	–4	1	–1	–1	–1	–2	–1	0	1	0	–2	1	1	0	–6	–2
C	–2	12	–5	–5	–4	–3	–3	–2	–5	–6	–5	–4	–3	–5	–4	0	–2	–2	–8	0
D	0	–5	4	3	–6	1	1	–2	0	–4	–3	2	–1	2	–1	0	0	–2	–7	–4
E	0	–5	3	4	–5	0	1	–2	0	–3	–2	1	–1	2	–1	0	0	–2	–7	–4
F	–4	–4	–6	–5	9	-5	–2	1	–5	2	0	–4	–5	–5	–4	–3	–3	–1	0	7
G	1	–3	1	0	–5	5	–2	–3	–2	–4	–3	0	–1	–1	–3	1	0	–1	–7	–5
H	–1	–3	1	1	–2	–2	6	–2	0	–2	–2	2	0	2	2	–1	–1	–2	–3	0
I	–1	–2	–2	–2	1	–3	–2	–5	–2	2	2	–2	–2	–2	–2	–1	0	4	–5	–1
K	–1	–5	0	0	–5	–2	0	–2	5	–3	0	1	–1	1	3	0	0	–2	–3	–4
L	–2	–6	–4	3	2	–4	–2	2	–3	6	4	–3	–3	–2	–3	–3	–2	2	–2	–1
M	–1	–5	–3	–2	0	–3	–2	2	0	4	6	–2	–2	–1	0	–2	–1	2	–4	–2
N	0	–4	2	1	–4	0	2	–2	1	–3	–2	2	–1	1	0	1	0	–2	–4	–2
O	1	–3	–1	–1	–5	–1	0	–2	–1	–3	–2	–1	6	0	0	1	0	–1	–6	–5
Q	0	–5	2	2	–5	–1	3	–2	1	–2	–1	1	0	4	1	–1	–1	–2	–5	–4
R	–2	–4	–1	–1	–4	–3	2	–2	3	–3	0	0	0	1	6	0	–1	–2	2	–4
S	1	0	0	0	–3	1	–1	–1	0	–3	–2	1	1	–1	0	2	1	–1	–2	–3
T	1	–2	0	0	–3	0	–1	0	0	–2	–1	0	0	–1	–1	1	3	0	–5	–3
V	0	–2	–2	–2	–1	–1	–2	4	–2	2	2	–2	–1	–2	–2	–1	0	4	–6	–2
W	–6	–8	–7	–7	0	–7	–3	–5	–3	–2	–4	–4	–6	–5	2	–2	–5	–6	17	0
Y	–3	0	-4	–4	7	–5	0	–1	–4	–1	–2	–2	–5	–4	–4	–3	–3	–2	0	10

Figura 5.3
Il grafico mostra la relazione tra gli indici delle matrici PAM (riportati in ascisse) e la divergenza tra le sequenze, ovvero la percentuale di residui non identici.





		T	F	D	E	R	I	L	G	V	Q	T	Y	W	A	E	C	L	A
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	2	3	1	0	0	0	0	4	0	0	0	0	3	0	0	0
T	0	3	0	0	2	2	1	0	0	0	0	7	2	0	1	0	1	0	1
F	0	0	12	7	2	0	3	3	0	0	0	2	14	9	4	0	0	3	0
W	0	0	7	5	0	4	0	1	0	0	0	0	9	31	26	21	16	11	6
E	0	0	2	11	9	4	2	0	1	0	3	0	4	26	31	30	25	20	15
C	0	0	0	6	5	5	2	0	0	0	0	1	0	21	26	25	42	37	32
I	0	0	1	1	4	3	10	5	0	4	0	0	0	16	21	24	37	44	39
K	0	0	0	1	1	7	5	7	3	0	5	0	0	11	16	21	32	39	43
G	0	0	0	1	1	2	4	2	12	7	2	5	0	6	12	16	27	34	40
D	0	0	0	4	5	0	0	0	7	10	9	4	1	1	7	16	22	29	35
N	0	0	0	2	5	5	0	0	2	5	11	9	4	0	2	11	17	24	30
A	0	1	0	0	2	3	4	0	1	2	6	12	7	2	2	6	12	19	26
T	0	3	0	0	0	1	3	2	0	1	1	9	9	4	3	2	7	14	21
Y	0	0	10	5	0	0	0	2	0	0	0	4	19	14	9	4	2	9	16

Figura 5.4
Allineamento locale ottenuto con l'algoritmo di Smith e Watermann, con una matrice di sostituzione PAM240 e utilizzando 5 come punteggio di penalità per l'apertura di gap.

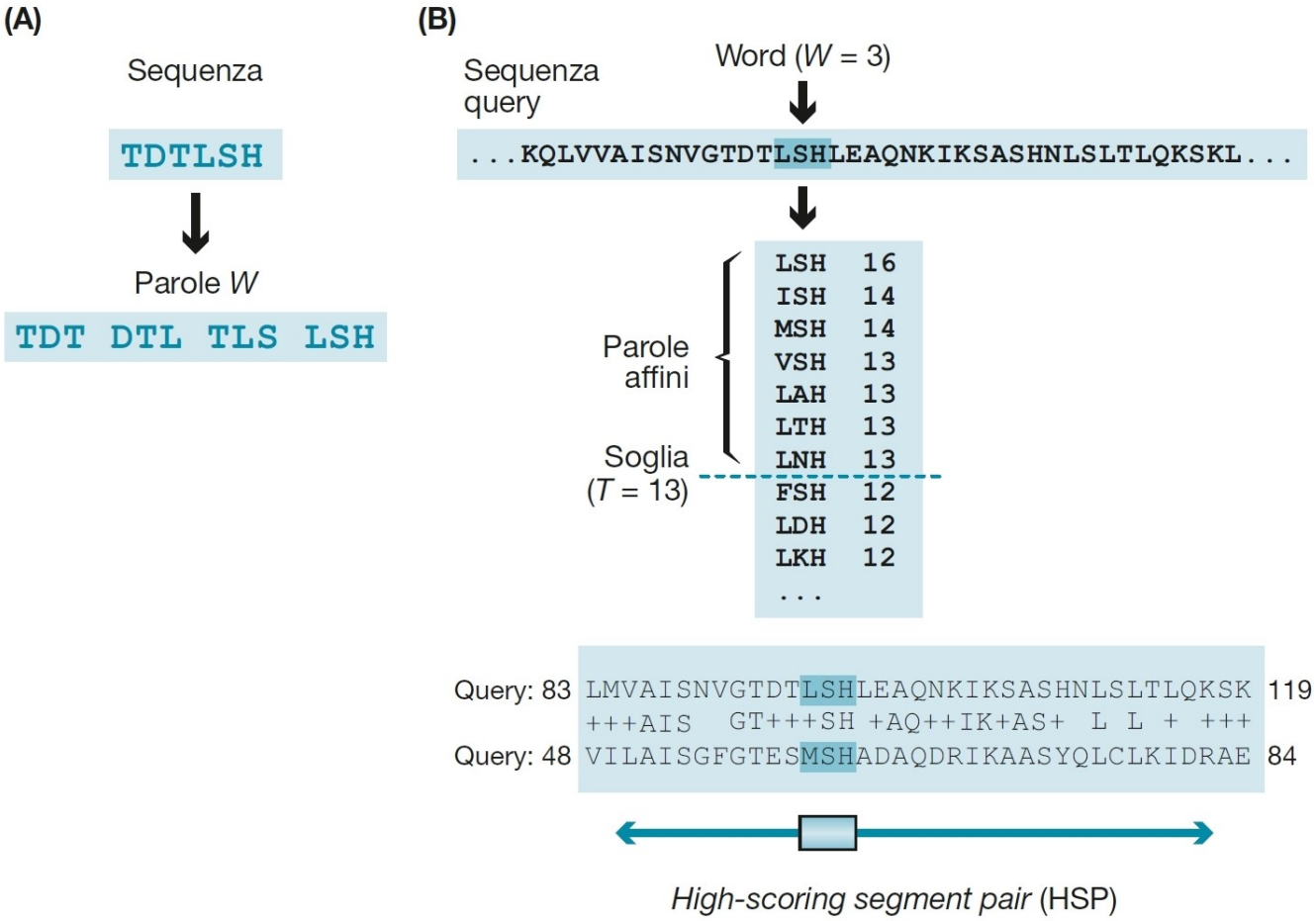
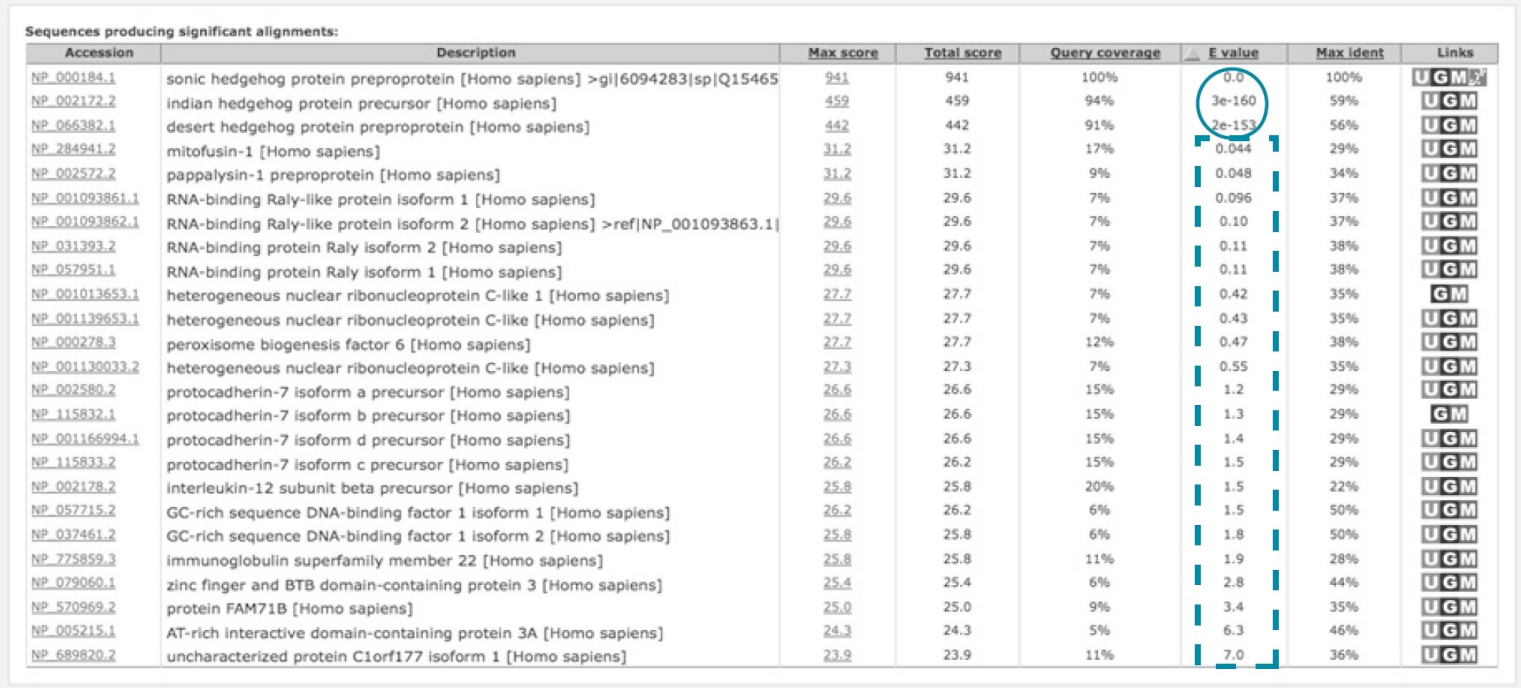


Figura 5.5
(**A**) Ogni sequenza viene suddivisa in parole di lunghezza *W* (per es. di 3 residui con *W* = 3), sovrapposte in quanto ottenute spostandosi di un residuo alla volta. (**B**) Per ognuna di tali parole, viene generato un elenco di parole affini, dette *W-mer*, che hanno un valore di allineamento superiore a *T* con la *W* di riferimento. L'algoritmo approfondisce l'analisi delle coppie di sequenze che hanno almeno un allineamento identico con un *W-mer*, in questo caso MSH. Nella parte inferiore è mostrata l'estensione a monte e a valle del *W-mer* identificato e la costruzione di un HSP.



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NP_000184.1	sonic hedgehog protein preproprotein [Homo sapiens] >gi|6094283|sp|Q15465	941	941	100%	0.0	100%	UGM
NP_002172.2	indian hedgehog protein precursor [Homo sapiens]	459	459	94%	3e-160	59%	UGM
NP_066382.1	desert hedgehog protein preproprotein [Homo sapiens]	442	442	91%	2e-153	56%	UGM
NP_284941.2	mitofusin-1 [Homo sapiens]	31.2	31.2	17%	0.044	29%	UGM
NP_002572.2	pappalysin-1 preproprotein [Homo sapiens]	31.2	31.2	9%	0.048	34%	UGM
NP_001093861.1	RNA-binding Raly-like protein isoform 1 [Homo sapiens]	29.6	29.6	7%	0.096	37%	UGM
NP_001093862.1	RNA-binding Raly-like protein isoform 2 [Homo sapiens] >ref|NP_001093863.1|	29.6	29.6	7%	0.10	37%	UGM
NP_031393.2	RNA-binding protein Raly isoform 2 [Homo sapiens]	29.6	29.6	7%	0.11	38%	UGM
NP_057951.1	RNA-binding protein Raly isoform 1 [Homo sapiens]	29.6	29.6	7%	0.11	38%	UGM
NP_001013653.1	heterogeneous nuclear ribonucleoprotein C-like 1 [Homo sapiens]	27.7	27.7	7%	0.42	35%	GM
NP_001139653.1	heterogeneous nuclear ribonucleoprotein C-like [Homo sapiens]	27.7	27.7	7%	0.43	35%	UGM
NP_000278.3	peroxisome biogenesis factor 6 [Homo sapiens]	27.7	27.7	12%	0.47	38%	UGM
NP_001130033.2	heterogeneous nuclear ribonucleoprotein C-like [Homo sapiens]	27.3	27.3	7%	0.55	35%	UGM
NP_002580.2	protocadherin-7 isoform a precursor [Homo sapiens]	26.6	26.6	15%	1.2	29%	UGM
NP_115832.1	protocadherin-7 isoform b precursor [Homo sapiens]	26.6	26.6	15%	1.3	29%	GM
NP_001166994.1	protocadherin-7 isoform d precursor [Homo sapiens]	26.6	26.6	15%	1.4	29%	UGM
NP_115833.2	protocadherin-7 isoform c precursor [Homo sapiens]	26.2	26.2	15%	1.5	29%	UGM
NP_002178.2	interleukin-12 subunit beta precursor [Homo sapiens]	25.8	25.8	20%	1.5	22%	UGM
NP_057715.2	GC-rich sequence DNA-binding factor 1 isoform 1 [Homo sapiens]	26.2	26.2	6%	1.5	50%	UGM
NP_037461.2	GC-rich sequence DNA-binding factor 1 isoform 2 [Homo sapiens]	25.8	25.8	6%	1.8	50%	UGM
NP_775859.3	immunoglobulin superfamily member 22 [Homo sapiens]	25.8	25.8	11%	1.9	28%	UGM
NP_079060.1	zinc finger and BTB domain-containing protein 3 [Homo sapiens]	25.4	25.4	6%	2.8	44%	UGM
NP_570969.2	protein FAM71B [Homo sapiens]	25.0	25.0	9%	3.4	35%	UGM
NP_005215.1	AT-rich interactive domain-containing protein 3A [Homo sapiens]	24.3	24.3	5%	6.3	46%	UGM
NP_689820.2	uncharacterized protein C1orf177 isoform 1 [Homo sapiens]	23.9	23.9	11%	7.0	36%	UGM

Figura 5.6
Nel cerchio con linea continua gli allineamenti che possiamo considerare significativi perché hanno un *E-value* sufficientemente basso, nel rettangolo con linea tratteggiata quelli non significativi.

Figura 5.7
Per ognuna delle proteine riportate nell'elenco di Figura 5.6, è disponibile l'allineamento.

Download GenPept Graphics

indian hedgehog protein preproprotein [Homo sapiens]

Sequence ID: NP_002172.2 **Length:** 411 **Number of Matches:** 1

Range 1: 28 to 407 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
459 bits(1182)	8e-160	Compositional matrix adjust.	258/436(59%)	298/436(68%)	57/436(13%)

```
Query  24   CGPGRGFG-KRRHPKKLTPLAYKQFIPNVAEKTLGASGRYEGKISRNSERFKELTPNYNP  82
            CGPGR  G +RR P+KL PLAYKQF PNV EKTLGASGRYEGKI+R+SERFKELTPNYNP
Sbjct  28   CGPGRVVGSRRRPPRKLVPLAYKQFSPNVPEKTLGASGRYEGKIARSSERFKELTPNYNP  87

Query  83   DIIFKDEENTGADRLMTQRCKDKLNALAISVMNQWPGVKLRVTEGWDEDGHHSEESLHYE  142
            DIIFKDEENTGADRLMTQRCKD+LN+LAISVMNQWPGVKLRVTEGWDEDGHHSEESLHYE
Sbjct  88   DIIFKDEENTGADRLMTQRCKDRLNSLAISVMNQWPGVKLRVTEGWDEDGHHSEESLHYE  147

Query  143  GRAVDITTSDRDRSKYGMLARLAVEAGFDWVYYESKAHIHCSVKAENSVAAKSGGCFPGS  202
            GRAVDITTSDRDR+KYG+LARLAVEAGFDWVYYESKAH+HCSVK+E+S AAK+GGCFP
Sbjct  148  GRAVDITTSDRDRNKYGLLARLAVEAGFDWVYYESKAHVHCSVKSEHSAAAKTGGCFPAG  207

Query  203  ATVHLEQGGTKLVKDLSPGDRVLAADDQGRLLYSDFLTFLDRDDGAKKVFYVIETREPRE  262
            A V LE G    +  + PGDRVLA  + G   +SD L FLDR+    + F VIET++P
Sbjct  208  AQVRLESGARVALSAVRPGDRVLAMGEDGSPTFSDVLIFLDREPHRLRAFQVIETQDPPR  267

Query  263  RLLLTAAHLLFVAPHNDSATGEPEASSGSGPPSGGALGPRALFASRVRPGQRVYVVAERD  322
            RL LT AHLLF A   D+ T EP A              RA FAS V+PGQ V V
Sbjct  268  RLALTPAHLLFTA---DNHT-EPAARF------------RATFASHVQPGQYVLVA----  307

Query  323  GDRRLLPAAVHSVTLSEEAAGAYAPLTAQGTILINRVLASCYAVIEEHSWAHRAFAPFRL  382
            G   L PA V +V+ +  A GAYAPLT  GT+++  V+ASC+A + +H  A  AF P RL
Sbjct  308  GVPGLQPARVAAVS-THVALGAYAPLTKHGTLVVEDVVASCFAAVADHHLAQLAFWPLRL  366

Query  383  AHALLAALAPARTDRGGDSGGGDRGGGGGRVALTAPGAADAPGAGATAGIHWYSQLLYQI  442
             H+L                                 +  PG     G+HWY QLLY++
Sbjct  367  FHSLA-------------------------------WGSWTPG----EGVHWYPQLLYRL  391

Query  443  GTWLLDSEALHPLGMA  458
            G  LL+  + HPLGM+
Sbjct  392  GRLLLEEGSFHPLGMS  407
```

Una sola sequenza non contiene informazioni sull'importanza relativa dei vari residui

Molte sequenze possono dare **MOLTE** informazioni

Figura 5.8
Se analizziamo una sola sequenza, ogni residuo ha lo stesso peso degli altri. I due triptofani (W nel codice a una lettera) presenti nella sequenza in alto non possono essere associati a una maggiore o minore importanza. Se invece abbiamo un allineamento multiplo di sequenze omologhe, ogni residuo viene immediatamente caratterizzato dalla sua maggiore o minore conservazione nelle altre sequenze omologhe. Per esempio, uno dei due W può essere molto conservato mentre l'altro può non essere conservato per nulla.

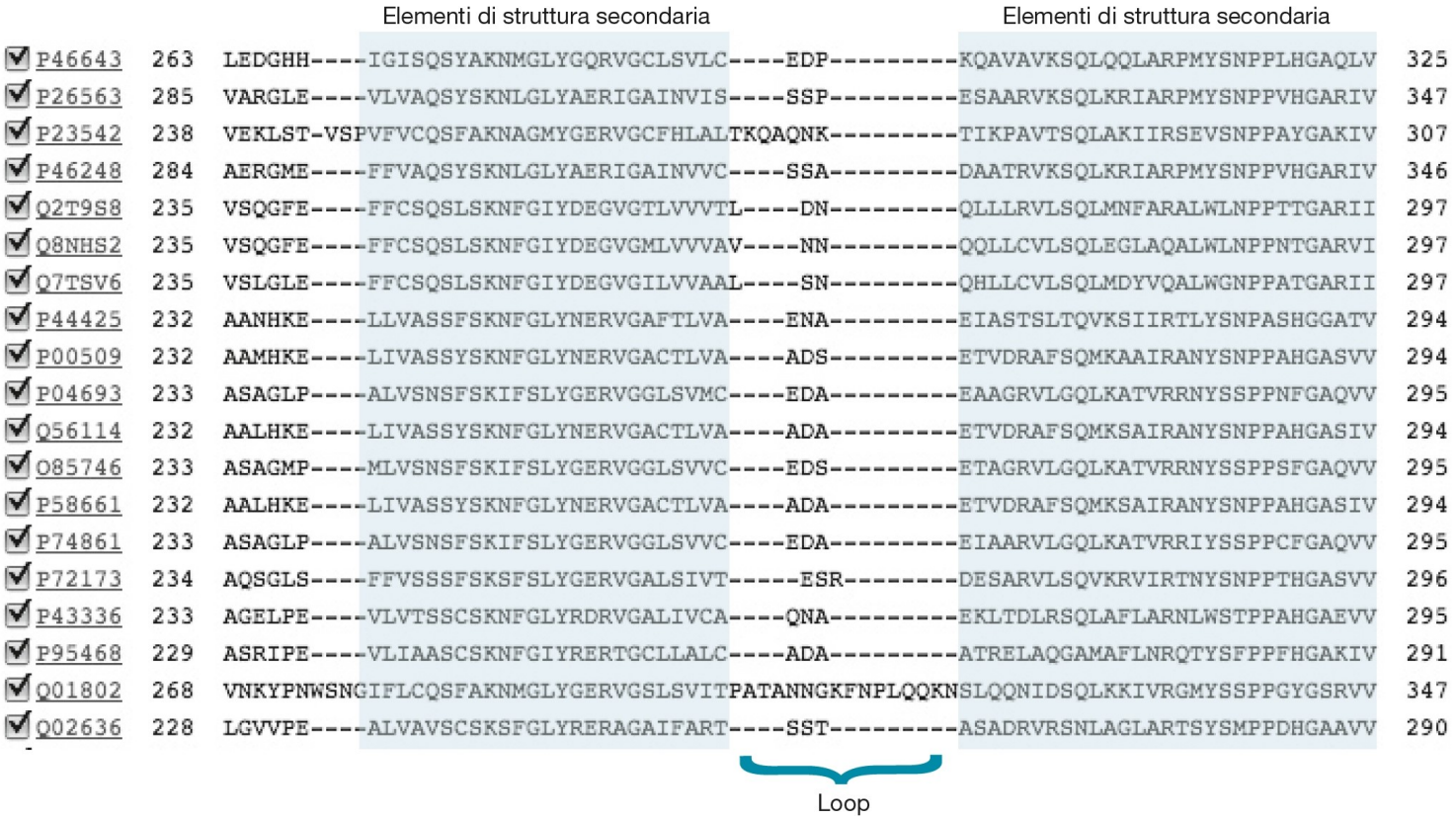
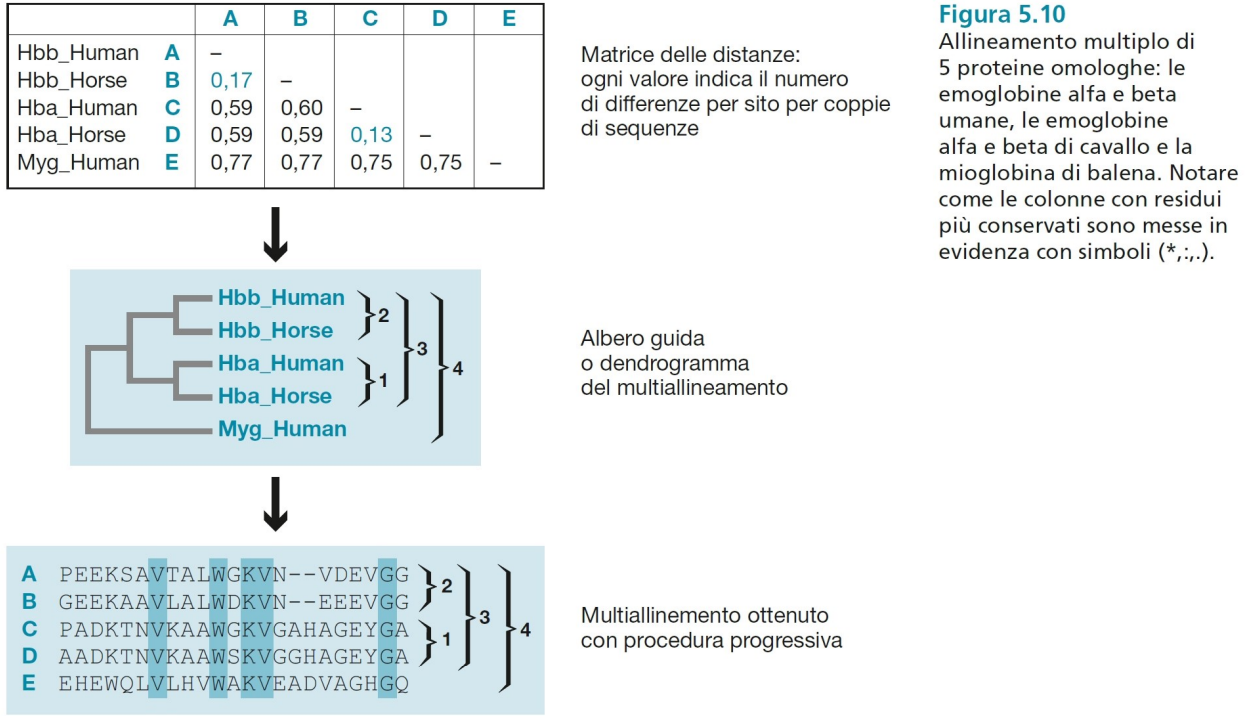


Figura 5.9
Un allineamento multiplo di proteine omologhe contiene importanti informazioni sulla struttura secondaria dei residui che lo compongono. Risulta evidente la possibile identificazione di elementi di struttura secondaria separati da un numero di residui variabile nelle diverse sequenze, e quindi probabilmente appartenenti a un loop. Ulteriori analisi a carico dei residui nei putativi elementi di struttura secondaria possono rendere facilmente identificabile il tipo di struttura secondaria (α-elica o filamento β).



		A	B	C	D	E
Hbb_Human	A	–				
Hbb_Horse	B	0,17	–			
Hba_Human	C	0,59	0,60	–		
Hba_Horse	D	0,59	0,59	0,13	–	
Myg_Human	E	0,77	0,77	0,75	0,75	–

Matrice delle distanze: ogni valore indica il numero di differenze per sito per coppie di sequenze

Albero guida o dendrogramma del multiallineamento

Multiallinemento ottenuto con procedura progressiva

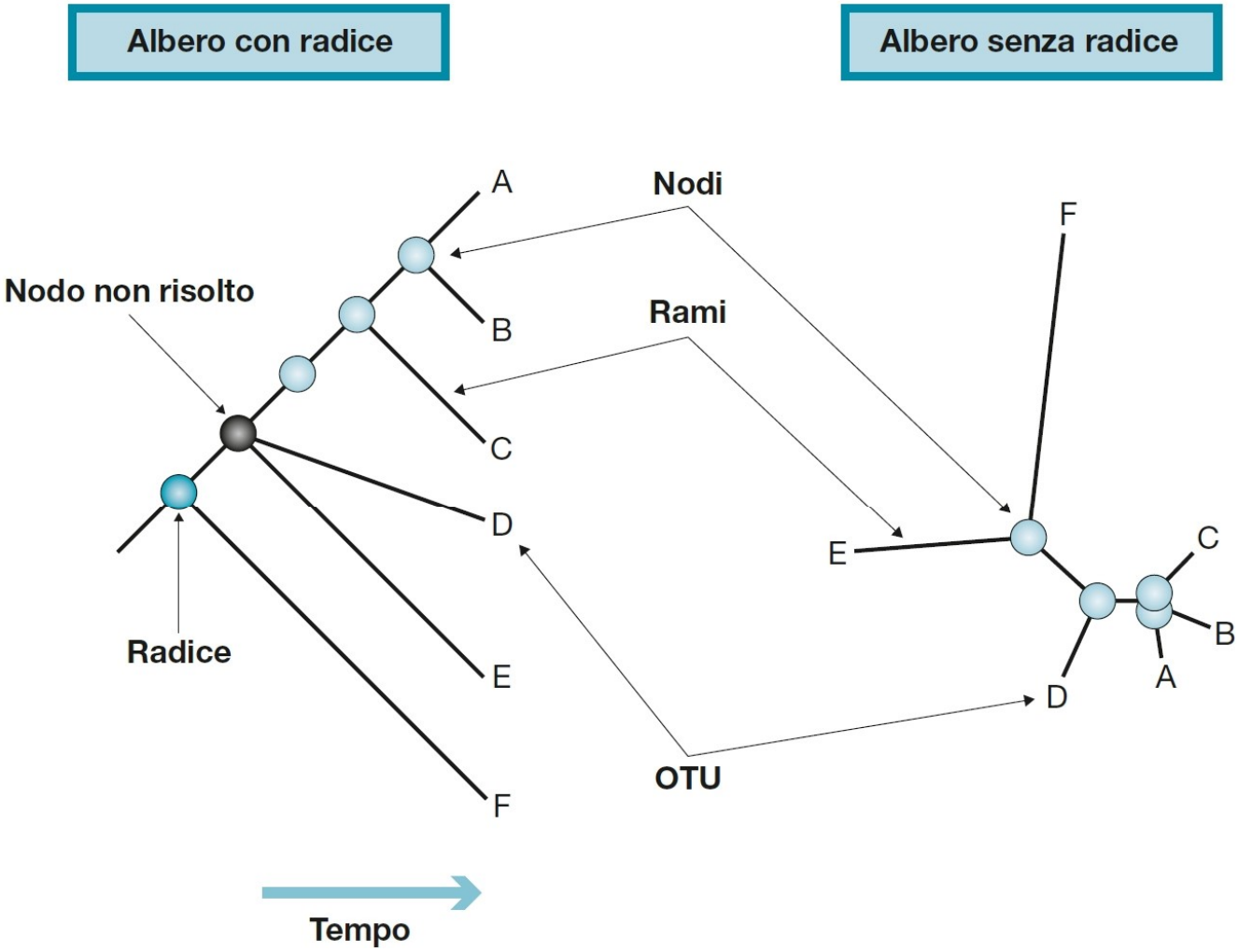
Figura 5.10
Allineamento multiplo di 5 proteine omologhe: le emoglobine alfa e beta umane, le emoglobine alfa e beta di cavallo e la mioglobina di balena. Notare come le colonne con residui più conservati sono messe in evidenza con simboli (*,:,.).

```
sp|P02144|MYG_HUMAN   -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKAS   59
sp|P69905|HBA_HUMAN   ------------------------------------------------------------   0
sp|P01958|HBA_HORSE   ------------------------------------------------------------   0
sp|P68871|HBB_HUMAN   MVHLTPEEKSAVTALWGKVNVDEV--GGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN   58
sp|P02062|HBB_HORSE   ------------------------------------------------------------   0

sp|P02144|MYG_HUMAN   EDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSK   119
sp|P69905|HBA_HUMAN   -------KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAH   53
sp|P01958|HBA_HORSE   -------KKVGDALTLAVGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVH   53
sp|P68871|HBB_HUMAN   PKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHH   118
sp|P02062|HBB_HORSE   ---KAHGKKVLHSFGEGVHHLDNLKGTFAALSELHCDKLHVDPENFRLLGNVLVVVLARH   57
                             .*  ::   : : ..    :  *:: *. * ::    :.::.. :: .*  :

sp|P02144|MYG_HUMAN   HPGDFGADAQGAMNKALELFRKDMASNYKELGFQG    154
sp|P69905|HBA_HUMAN   LPAEFTPAVHASLDKFLASVSTVLTSKYR------    82
sp|P01958|HBA_HORSE   LPNDFTPAVHASLDKFLSSVSTVLTSKYR------    82
sp|P68871|HBB_HUMAN   FGKEFTPPVQAAYQKVVAGVANALAHKYH------    147
sp|P02062|HBB_HORSE   FGKDFTPELQASYQKVVAGVANALAHKYH------    86
                         :*    :.: :* :  . . :: :*:
```

Capitolo 6

Alberi filogenetici

Figura 6.1
Illustrazione della topologia degli alberi con radice (*rooted*) e senza radice (*unrooted*). In un albero con radice esiste un nodo particolare, definito radice, che rappresenta il comune progenitore di tutti i nodi rappresentati nell'albero. In questo caso, tutti i rami dell'albero possono essere orientati in funzione del tempo. Un albero senza radice, invece, descrive esclusivamente le relazioni evolutive tra le OTU senza fornire alcuna informazione circa il processo evolutivo in funzione del tempo. In altre parole, non si può stabilire la collocazione temporale dei nodi interni in termini di quali siano più antichi e quali più recenti.



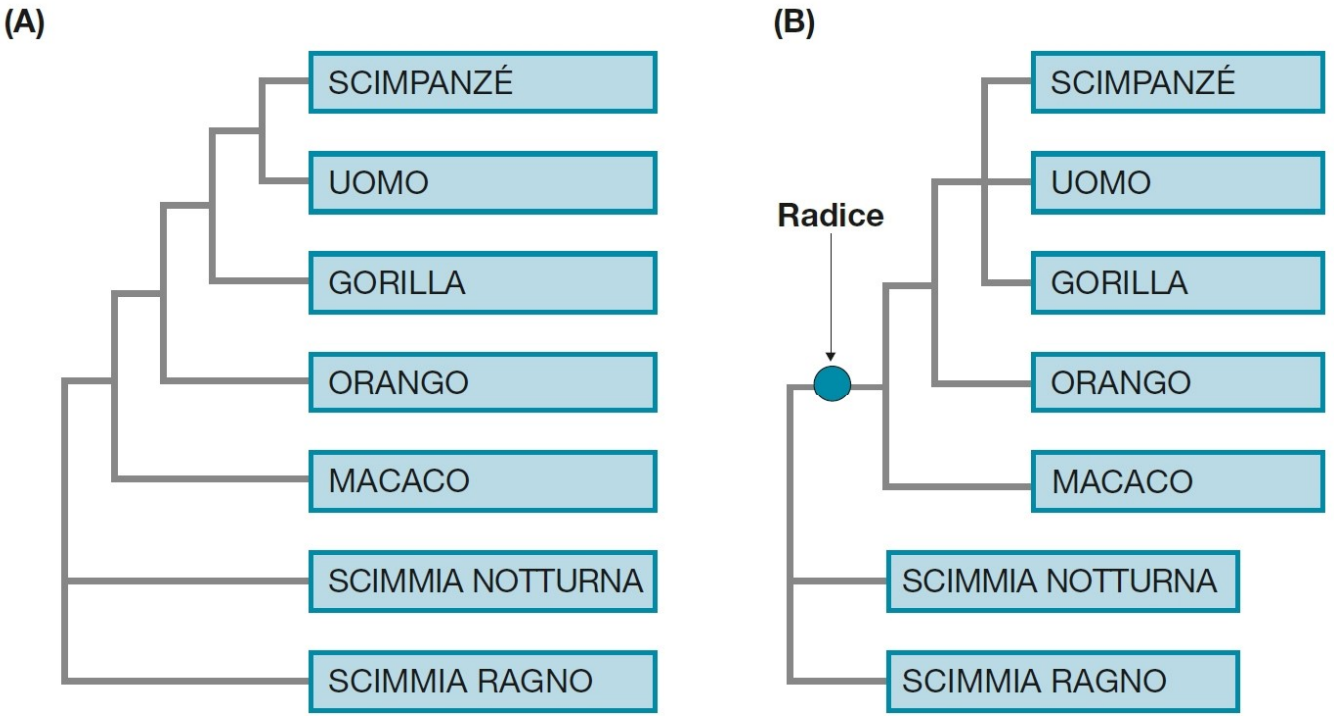


Figura 6.2
Esempio di cladogramma (**A**) e di filogramma (**B**) che descrive le relazioni filogenetiche tra sette sequenze di ψη-globine di primati. Nel filogramma la presenza di uno o più *outgroup* consente di collocare la radice all'interno del ramo che congiunge gli *outgroups* con tutte le altre OTU (*ingroups*).

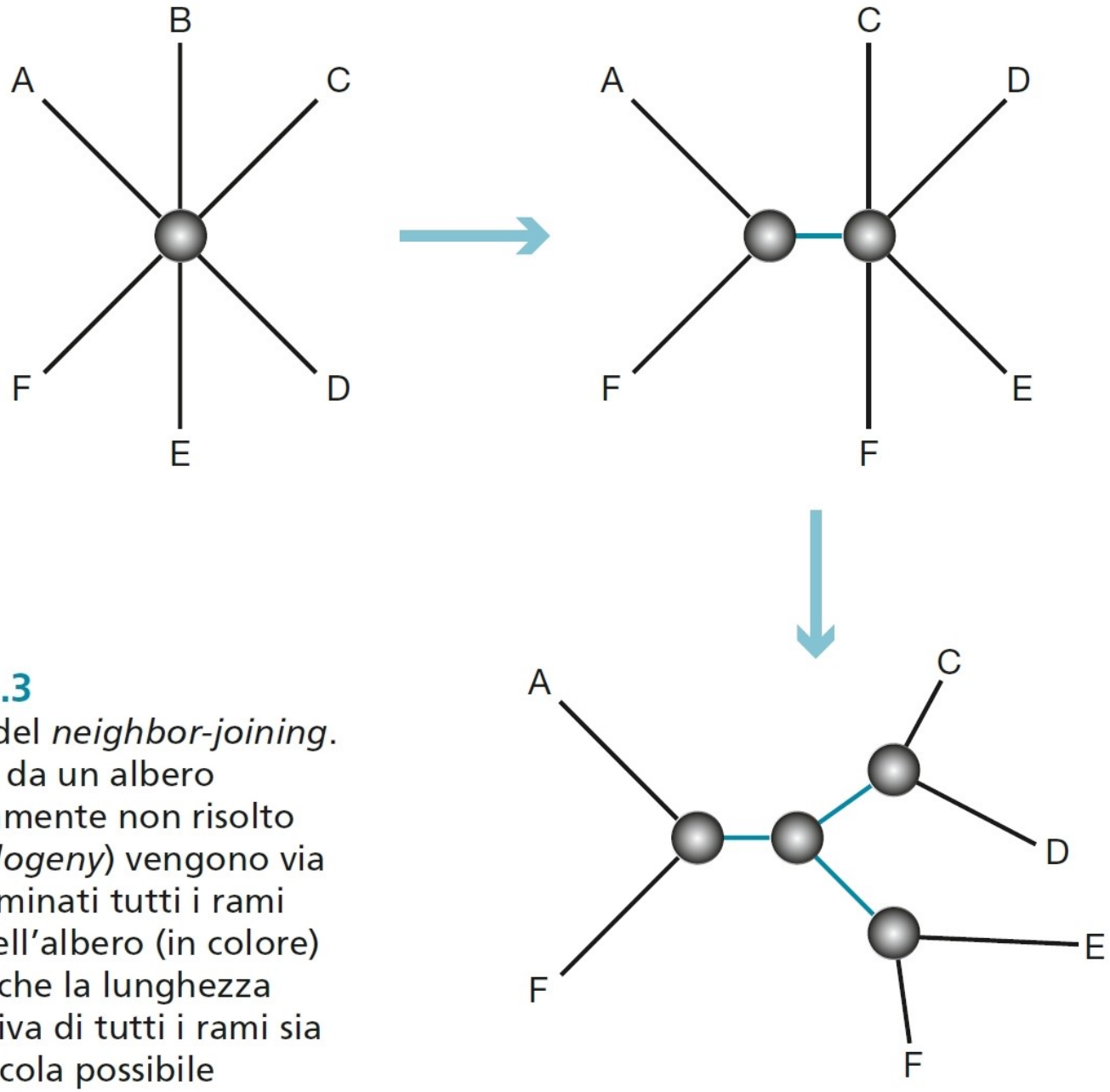


Figura 6.3
Metodo del *neighbor-joining*. A partire da un albero completamente non risolto (*star phylogeny*) vengono via via determinati tutti i rami interni dell'albero (in colore) in modo che la lunghezza complessiva di tutti i rami sia la più piccola possibile

(A)

	Sito									
Sequenza	1	2	3	4	5	6	7	8	9	10
1	G	G	C	A	G	T	C	C	A	C
2	G	A	G	C	G	T	G	C	G	C
3	G	A	T	G	A	T	T	C	A	C
4	G	A	T	T	A	T	T	C	G	C

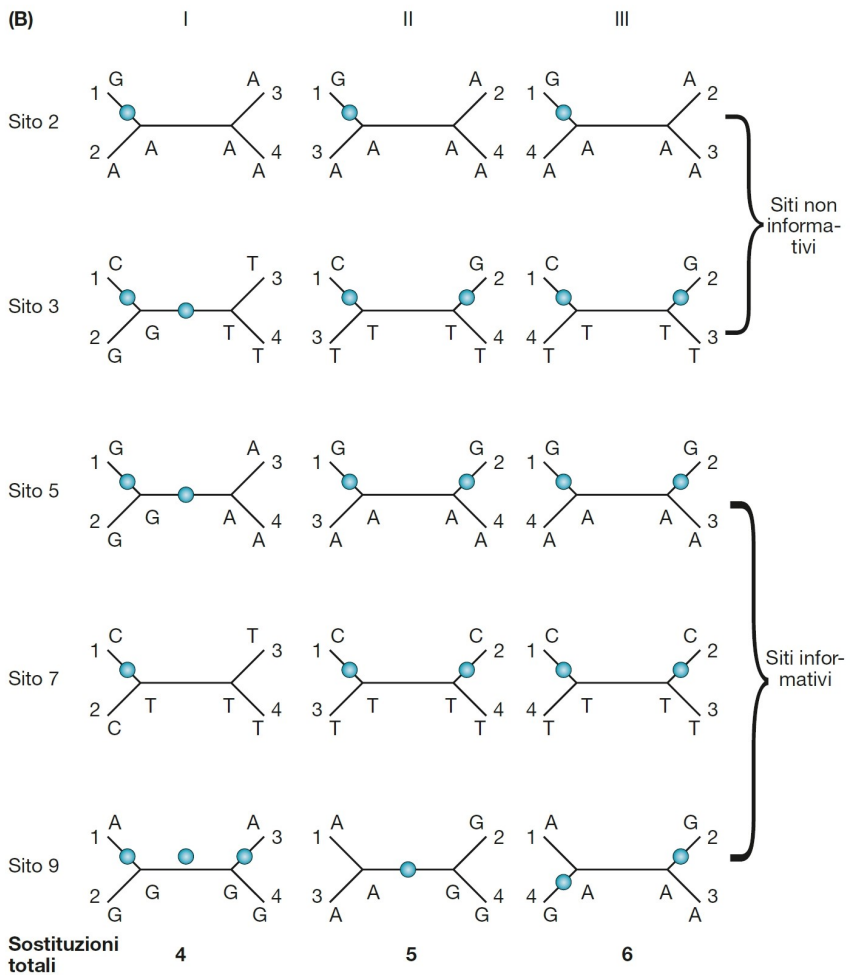


Figura 6.4
(A) Allineamento multiplo di 4 ipotetiche sequenze ciascuna costituita da 10 siti. I siti informativi sono riportati in blu. **(B)** I tre alberi senza radice (I, II e III) che descrivono le possibili relazioni filogenetiche tra le sequenze del multiallineamento. I nodi terminali indicano i nucleotidi presenti in posizioni omologhe delle sequenze appartenenti alle specie considerate. I cerchietti blu sui rami rappresentano altrettante sostituzioni nucleotidiche. I nucleotidi presenti nei due nodi ancestrali rappresentano una delle possibili ricostruzioni tra diverse alternative ugualmente parsimoniose. Si noti come i siti "non informativi" non favoriscono alcuno dei tre alberi alternativi, mentre i tre siti informativi determinano l'albero I come quello di "massima parsimonia", richiedendo solo 4 sostituzioni nucleotidiche complessive.

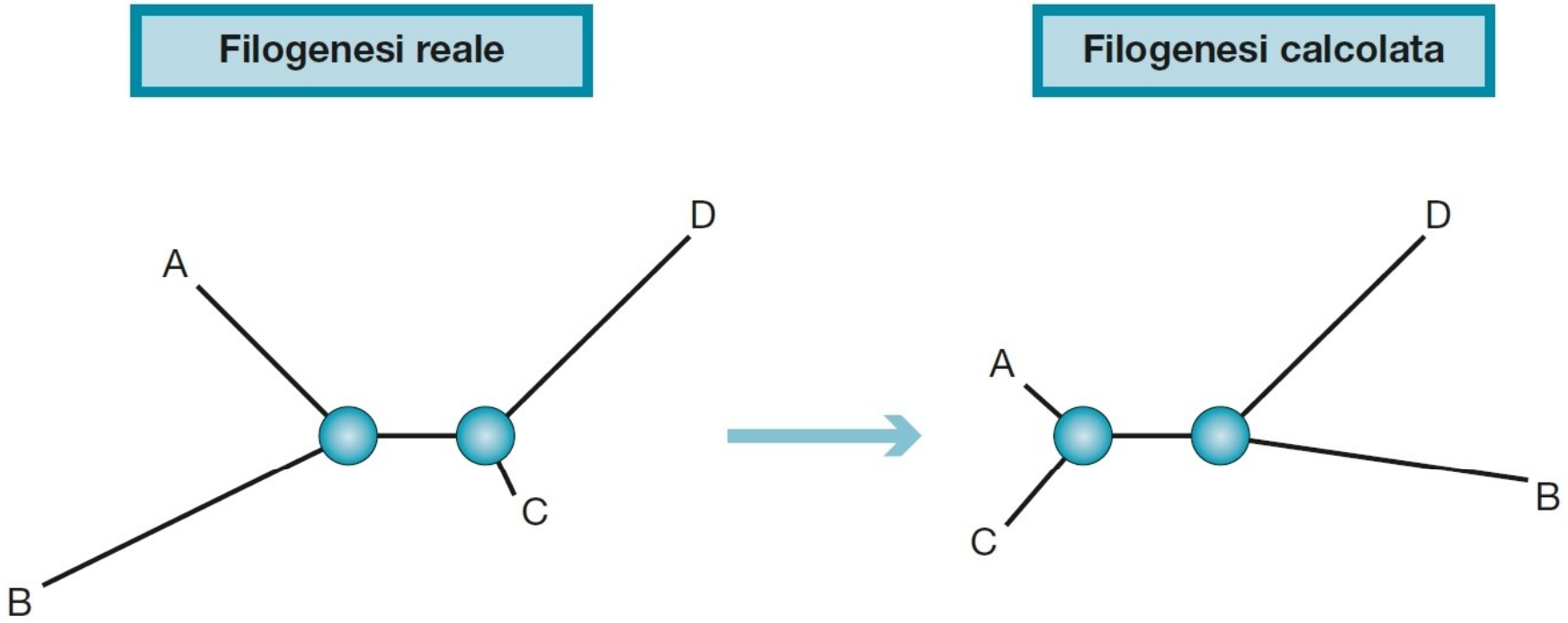
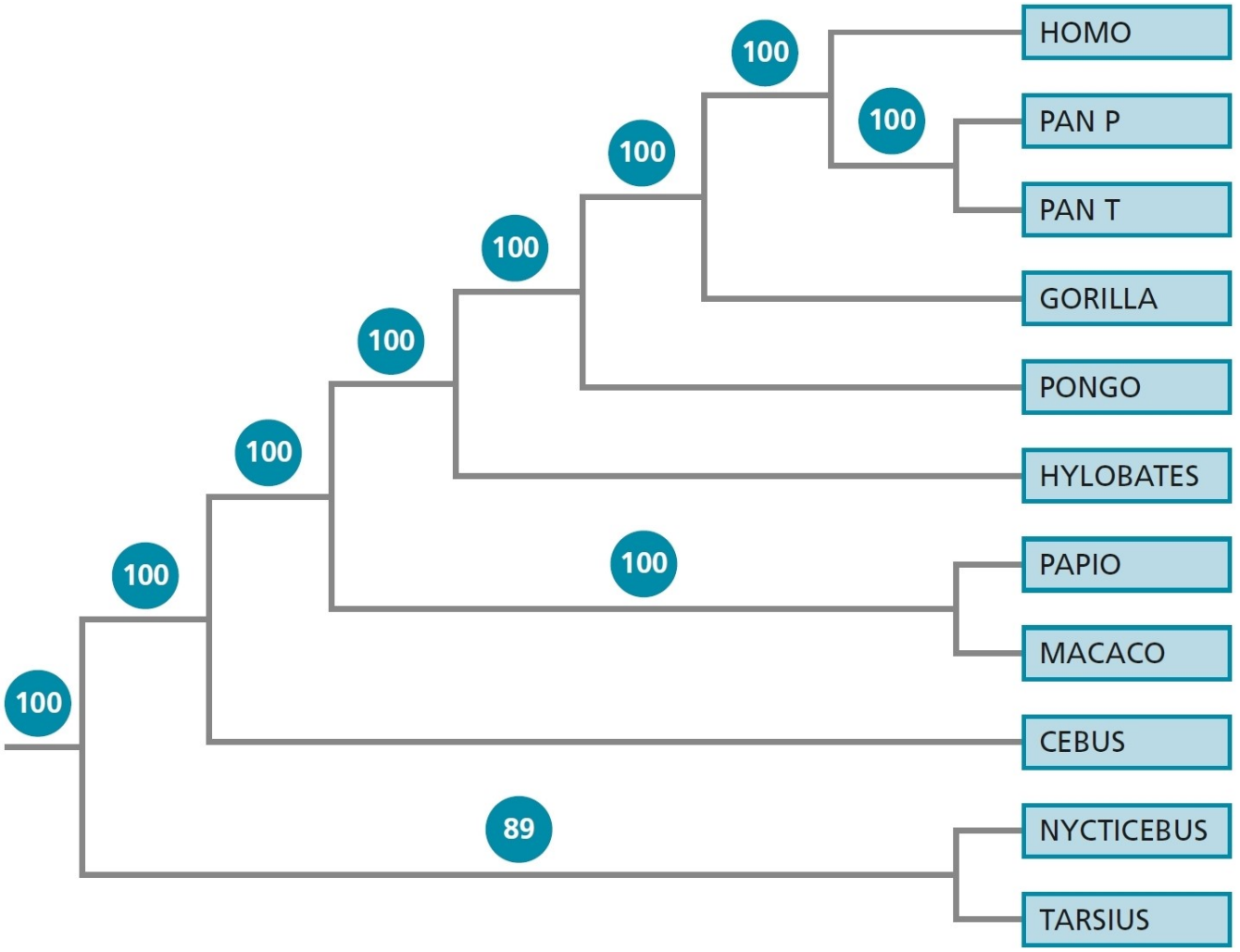


Figura 6.5
Effetto della *long branch attraction* nella ricostruzione filogenetica. Se l'albero che descrive la reale filogenia tra le specie contiene sia rami brevi che lunghi, la ricostruzione filogenetica produrrà un albero nel quale le OTU alle estremità dei rami più lunghi tenderanno ad "attrarsi".

Figura 6.6
Albero filogenetico di alcune specie di primati determinato dal programma MrBayes sulla base dell'analisi evolutiva dei geni mitocondriali con l'indicazione dei valori di *bootstrap* a supporto dei diversi nodi.



Capitolo 7

Piattaforme di sequenziamento degli acidi nucleici

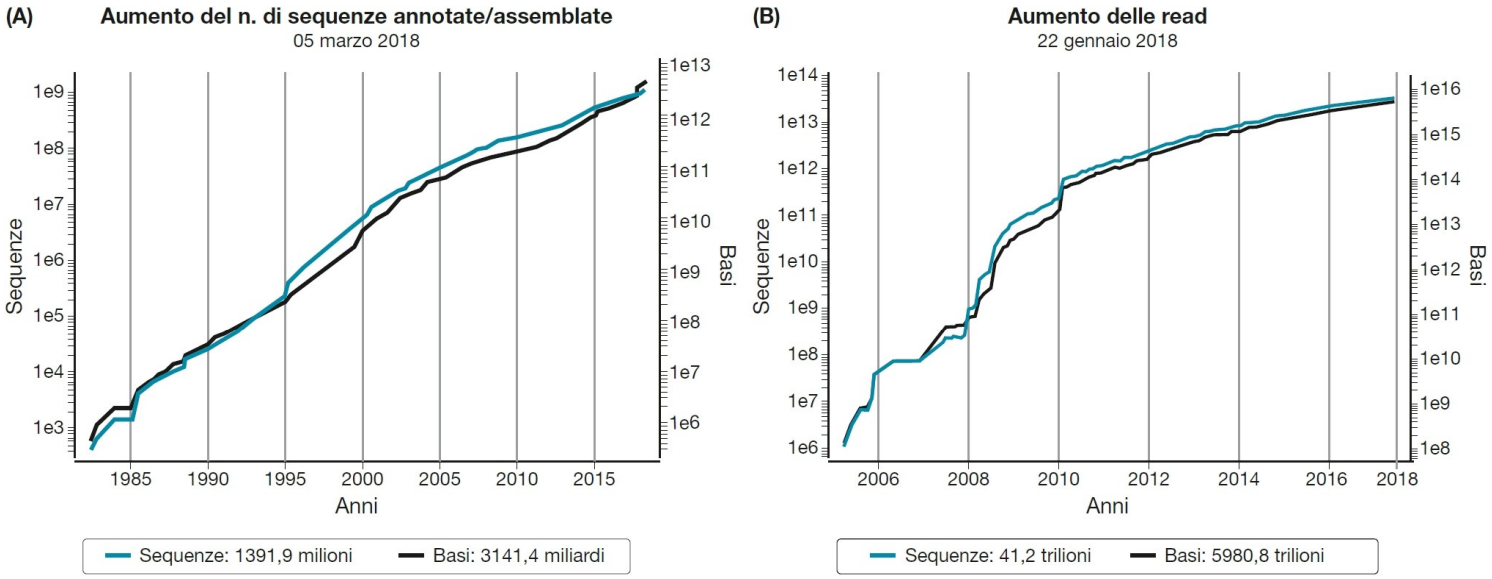


Figura 7.1
Incremento esponenziale del numero di sequenze nucleotidiche (**A**) e del numero di read NGS (**B**) nelle banche dati negli anni. (*Fonte*: www.ebi.ac.uk/ena/about/statistics)

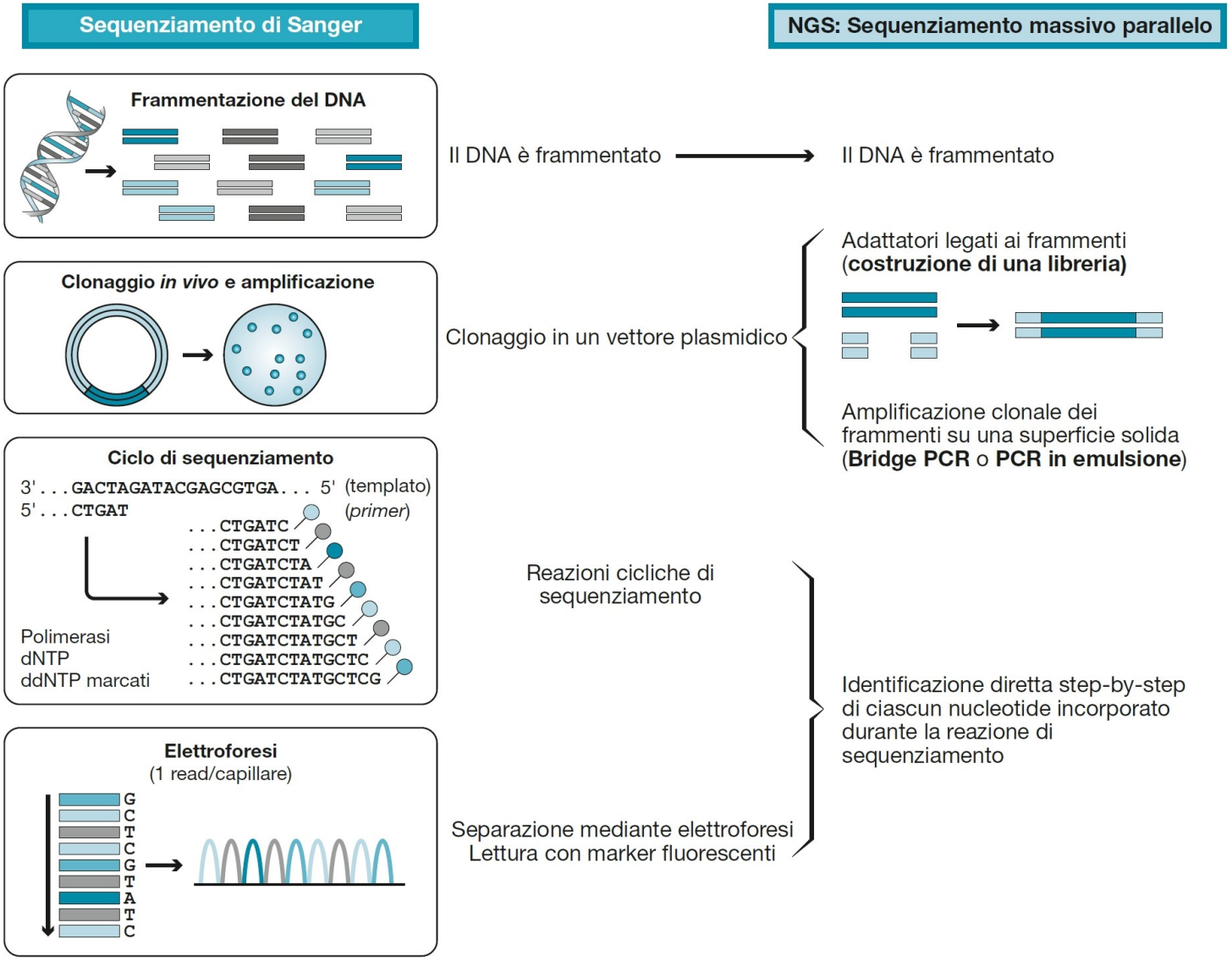
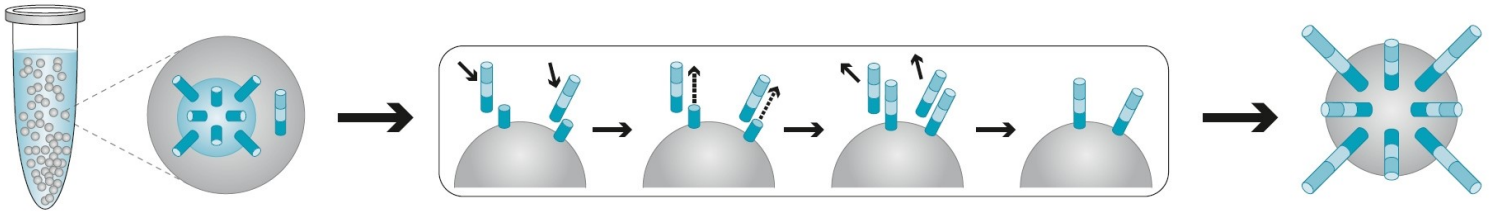


Figura 7.2
Confronto tra il sequenziamento tradizionale basato sul metodo di Sanger (*a sinistra*) e il sequenziamento di nuova generazione (*a destra*).

PCR a emulsione

454 (Roche)®, SOLiD® (ThermoFisher), GeneReader® (Qiagen), Ion Torrent® (ThermoFisher)



Emulsione
Gocce di micelle sono caricate con *primer*, templato, dNTP e polimerasi

Amplificazione clonale su biglia
I frammenti di DNA (templati) ibridizzano con i *primer* legati alla biglia e vengono amplificati: dopo l'amplificazione, i filamenti complementari si dissociano, lasciando molte copie del frammento di DNA a singolo filamento legate alle biglie

Prodotto finale
100-200 millioni di biglie con migliaia di templati legati

Biglie con templati di DNA amplificati clonalmente ed enzimi di sequenziamento

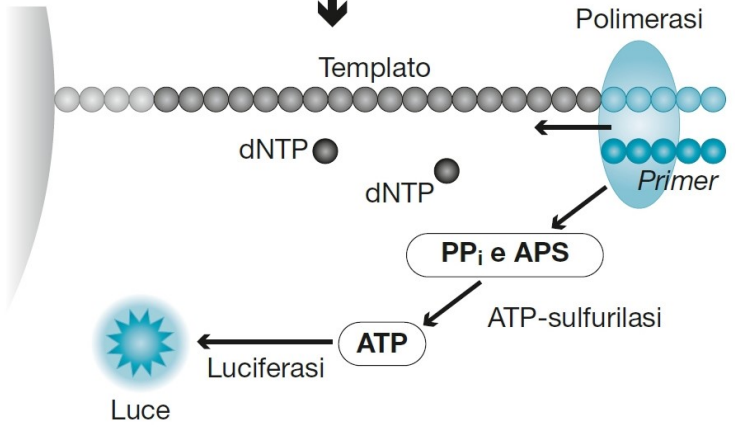


Figura 7.3
Schema del sequenziamento Roche 454.

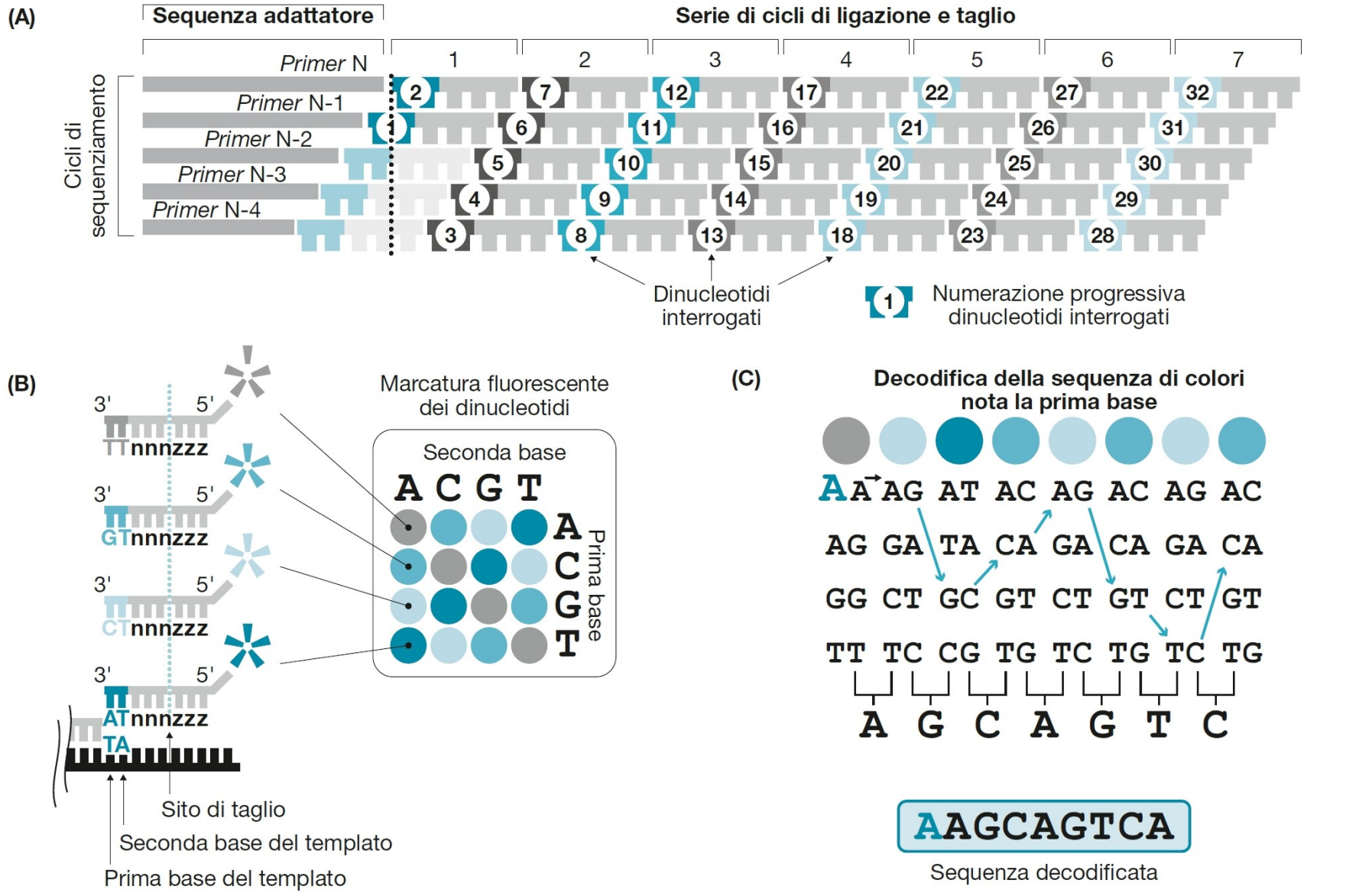


Figura 7.4
Schema del sequenziamento SOLiD.

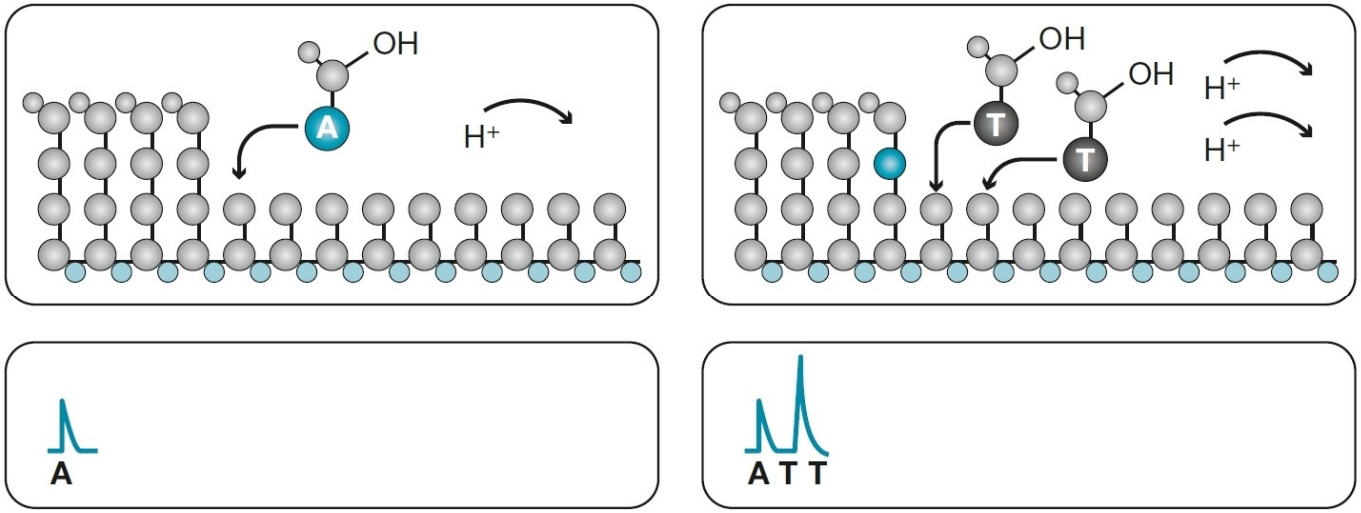


Figura 7.5
Schema del sequenziamento Ion Torrent.

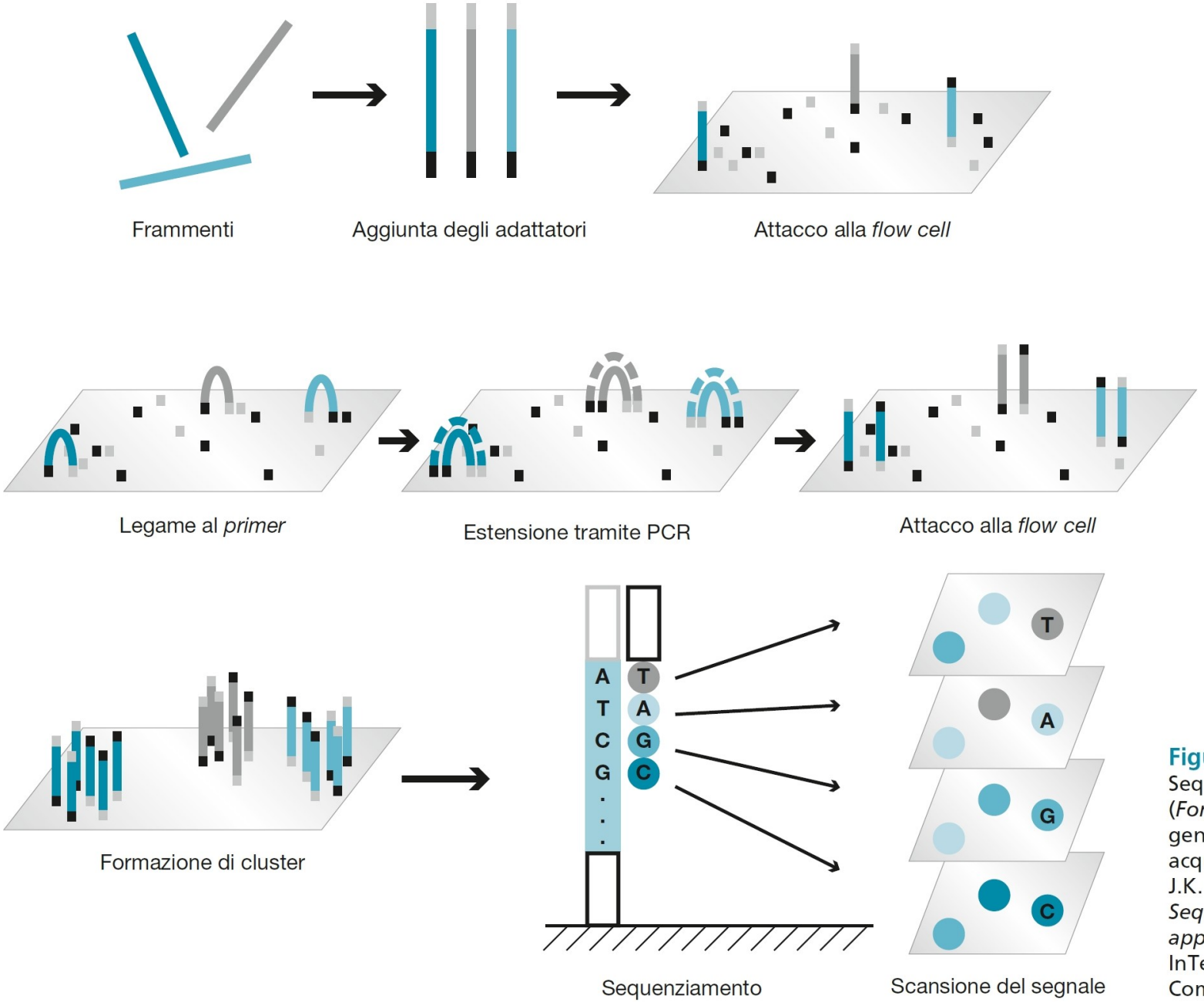


Figura 7.6
Sequenziamento Illumina. (*Fonte*: Lu B.Y. *et al.*, Next generation sequencing in acquatic models, in Kulski J.K. (ed.), *Next Generation Sequencing - Advances, applications and challanges*, InTech (2016), Creative Commons 3.0)

Figura 7.7
Il MinIon della Oxford Nanopore. (*Fonte:* http://labiotech.eu/interested-minion-first-pocket-dna-sequencer/)

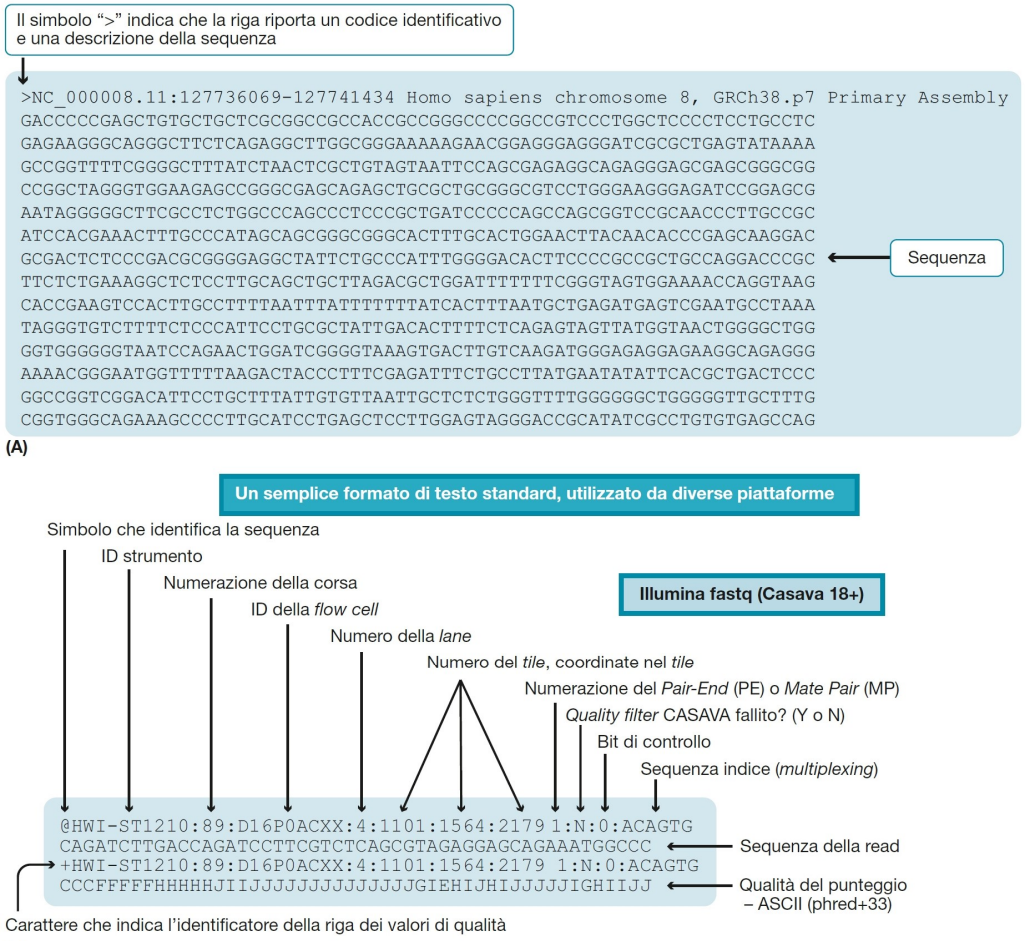


Figura 7.8
(**A**) Il formato *FASTA*.
(**B**) Il formato *FASTQ* (gentile concessione di David Horner).
(**C**) Corrispondenza tra i valori di Phred score (phred+33) e i codici ASCII.

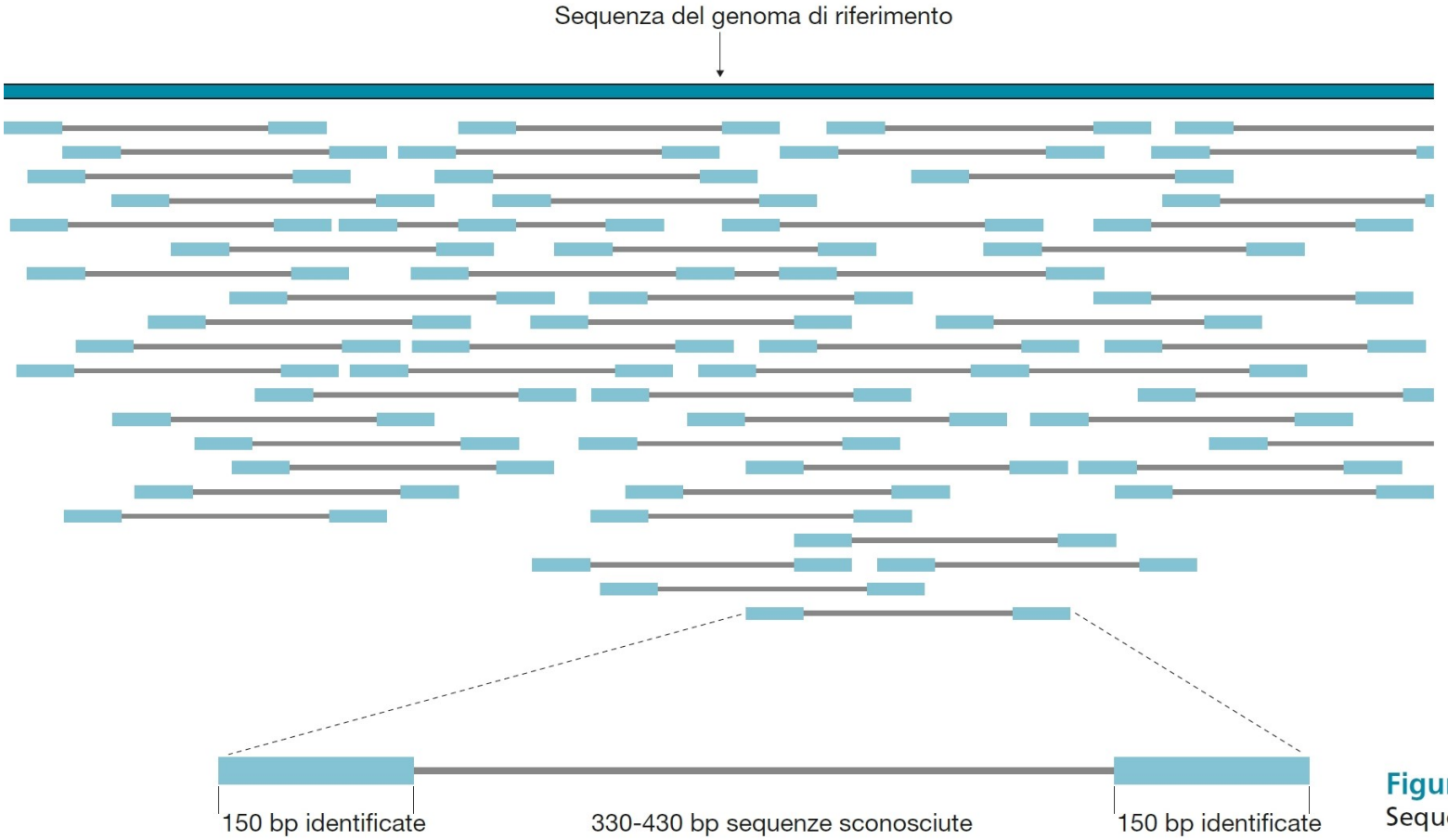


Figura 7.9
Sequenziamento *paired-end*.

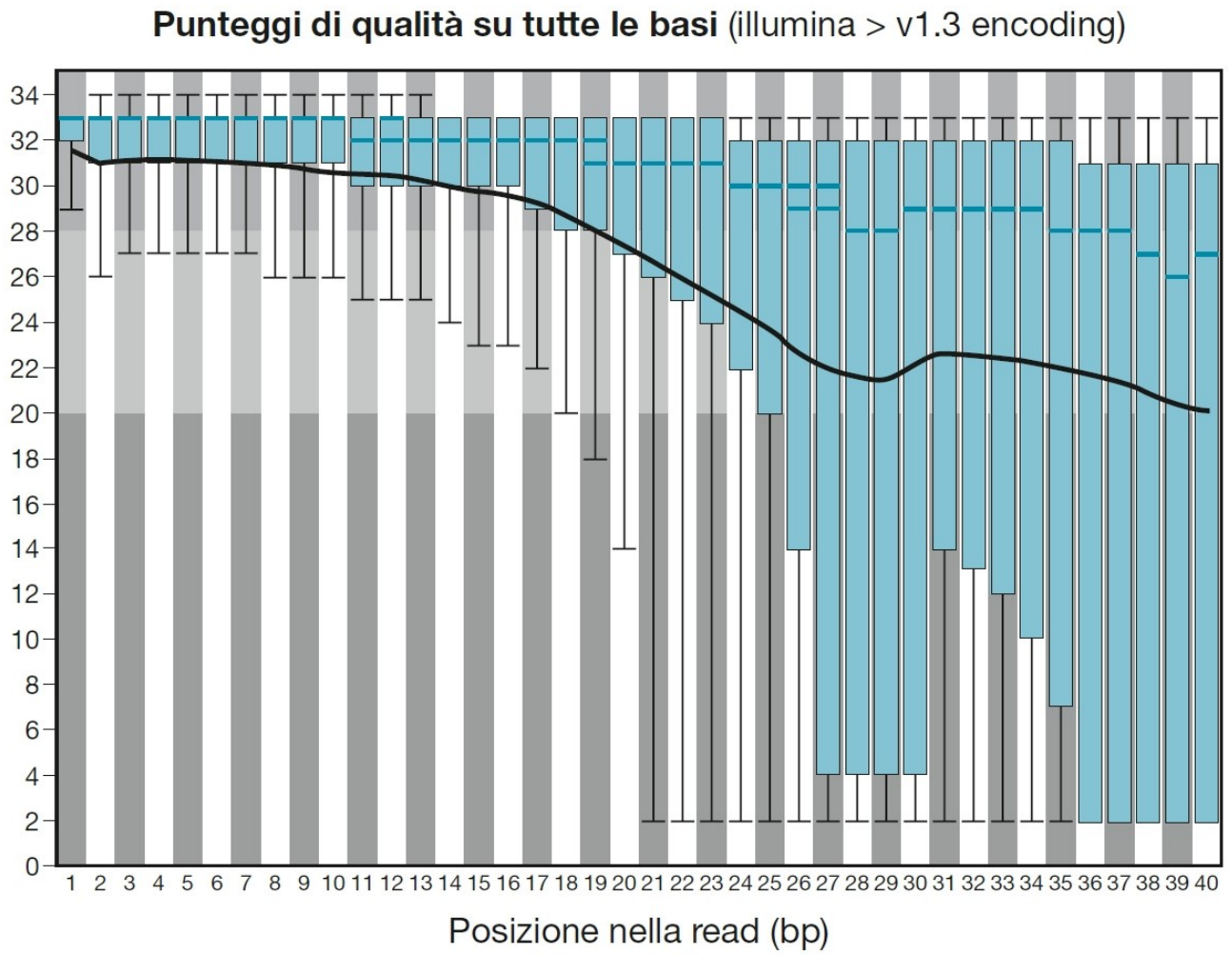


Figura 7.10
Esempio di output di FastQC. Il grafico mostra la distribuzione dei punteggi di qualità (sulle ordinate) in ogni posizione delle read (riportata sulle ascisse dal 5' al 3'), mettendo in evidenza in questo caso un calo medio di qualità procedendo verso l'estremità 3', e suggerendo che sia necessario un *trimming* di queste porzioni poco affidabili. (*Fonte:* www.bioinformatics.babraham.ac.uk/projects/fastqc/)

Capitolo 8

Ricostruzione e annotazione di genomi

Figura 8.1
Assemblaggio di sequenze nucleotidiche. Da un genoma (**A**), presente identico in tutte le cellule di un campione, si ottengono frammenti diversi a partire da ogni cellula (**B**), dei quali solo una parte sarà sequenziata, ottenendo delle read (**C**). La sovrapposizione delle read in base alla loro somiglianza permette la ricostruzione della sequenza originale (**D**) e di correggere eventuali errori di sequenziamento, mostrati nell'esempio dal nucleotide "t" nella read *f4*, che può essere corretto in A in base alla sequenza delle altre due read (*f2* e *f6*) che sono allineate nella stessa posizione.

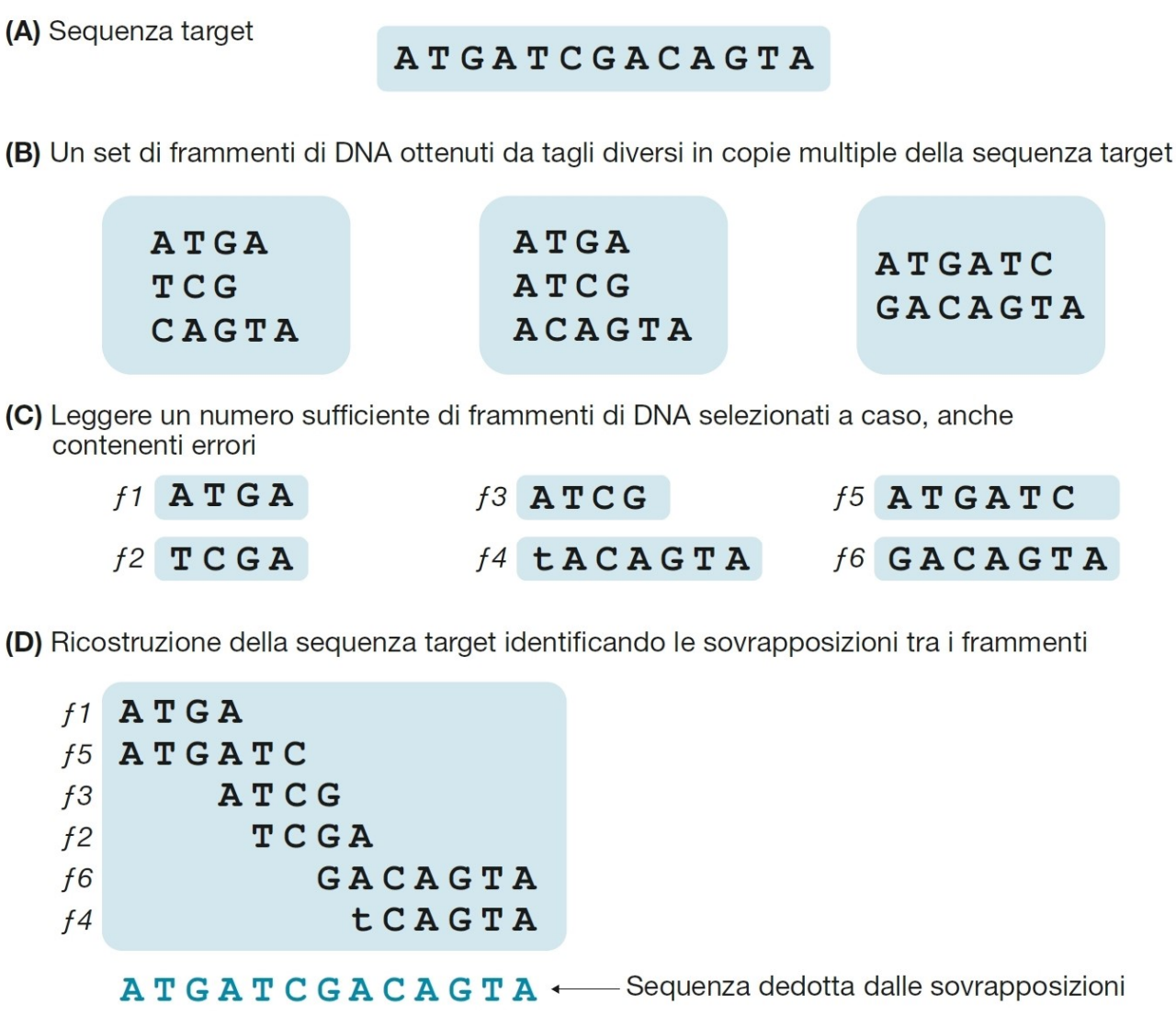
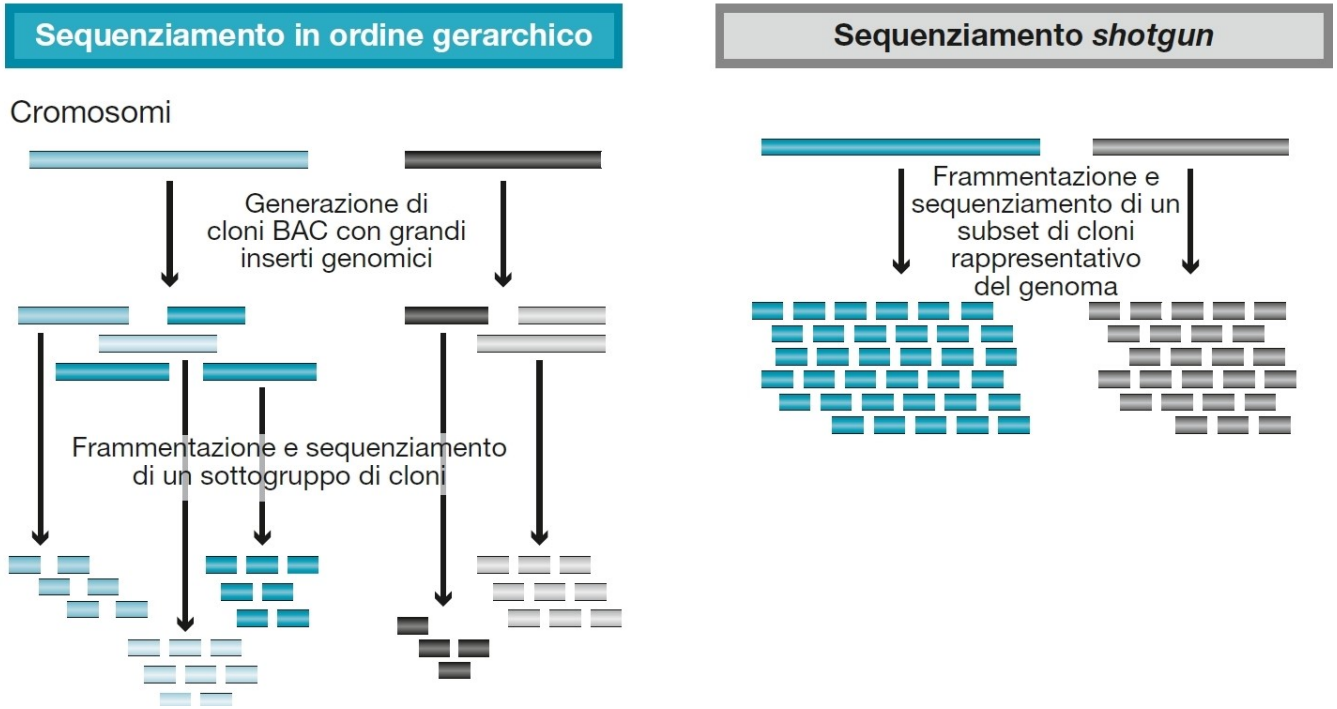


Figura 8.2
Sequenziamento gerarchico e *shotgun*. Nel sequenziamento gerarchico il sottogruppo di cloni da sequenziare è determinato sulla base di una mappa fisica del genoma, ottenuta precedentemente, che consente di ordinare i cloni BAC e di selezionarne il più piccolo sottoinsieme (mostrato in tonalità più scura) che copra l'intero genoma.



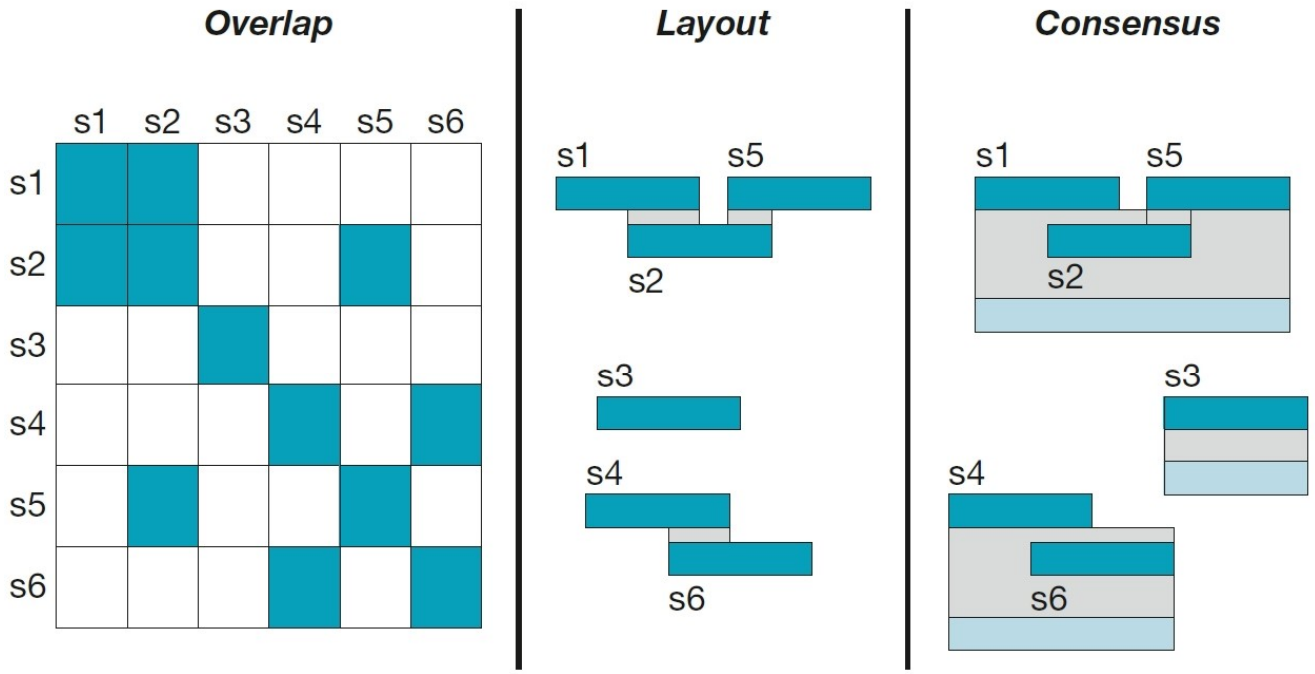
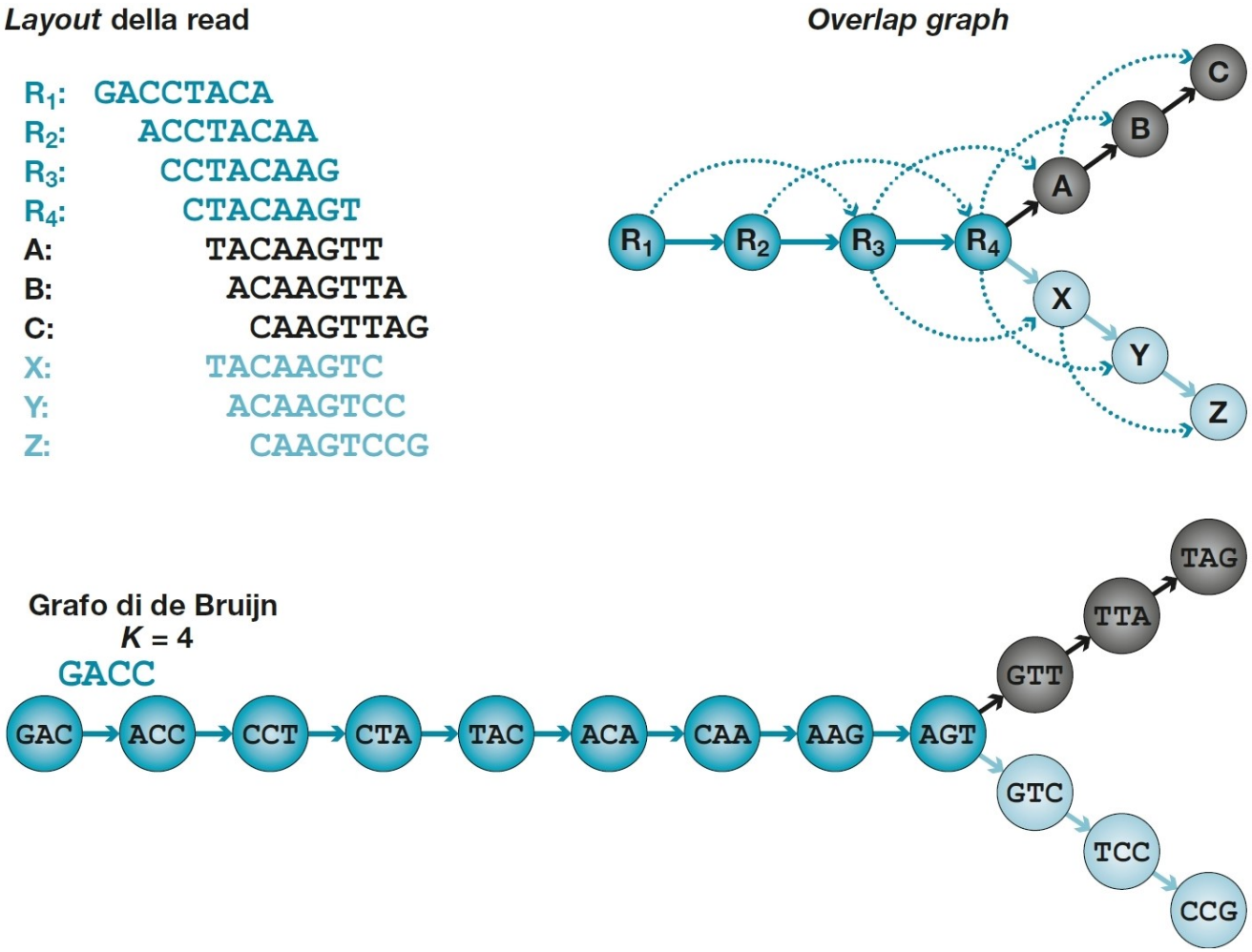


Figura 8.3
Algoritmi di assemblaggio *Overlap-Layout-Consensus*. Nella fase di *overlap*, la similarità fra ogni coppia di read è calcolata e riportata in una matrice. Da questa matrice, si cerca la combinazione migliore fra read che si estendono l'un l'altra (fase di *layout*). Infine, dalla sovrapposizione della sequenza delle read si ricava la sequenza rappresentativa di ogni porzione del genoma (*consensus*).

Figura 8.4
L'*overlap graph* e il grafo di de Bruijn (Schatz *et al.*, 2010). Il grafo di *overlap* mostra le similarità (riportate come archi) fra le read in esame (riportate come nodi), da cui si può ricostruire il *layout*. Gli archi tratteggiati indicano un *overlap* con *shift* di 2 nucleotidi, mentre gli archi pieni indicano un *overlap* con *shift* di 1 nucleotide. Nell'esempio, due set di read (A, B, e C; X, Y e Z) mostrano due strade alternative, che potrebbero corrispondere a una sequenza ripetuta. Il grafo di de Bruijn è un modo di rappresentazione alternativo più adatto a read provenienti dal sequenziamento di nuova generazione.



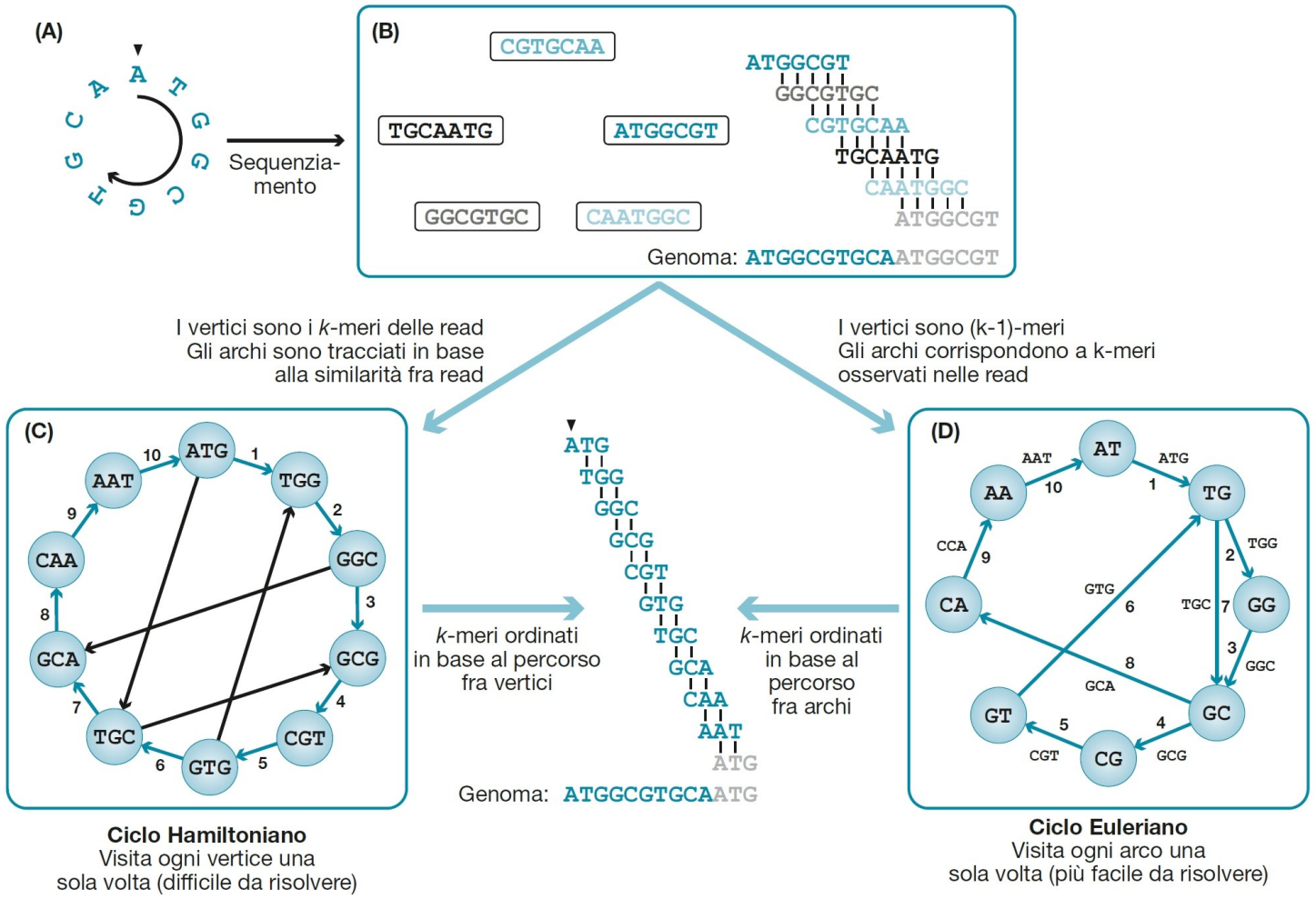


Figura 8.5
Percorsi Hamiltoniani ed Euleriani. Partendo da una sequenza genomica (**A**) e dalle read ottenute dal suo sequenziamento (**B**), il *layout* può essere ricostruito da un grafo di *overlap* cercando un percorso Hamiltoniano (**C**) che includa ogni nodo (cioè ogni frammento di read lungo *k* nt) una sola volta, che in questo caso è reso difficoltoso da similarità spurie (mostrate da frecce nere), o da un grafo di de Bruijn, in cui la risoluzione è più semplice e accurata. (*Adattata da*: Compeau P.E. *et al.*, *Nat Biotechnol*, 2011, 29(11):987-991)

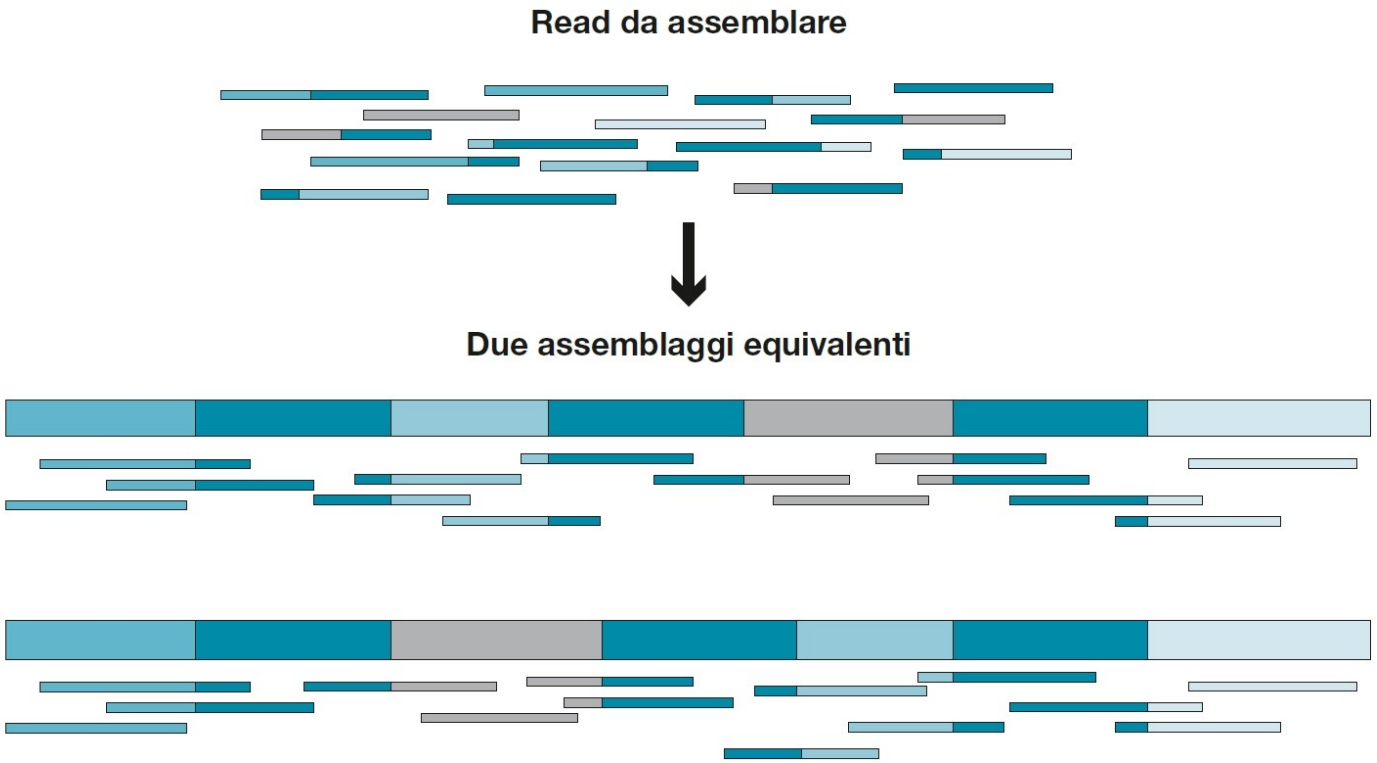
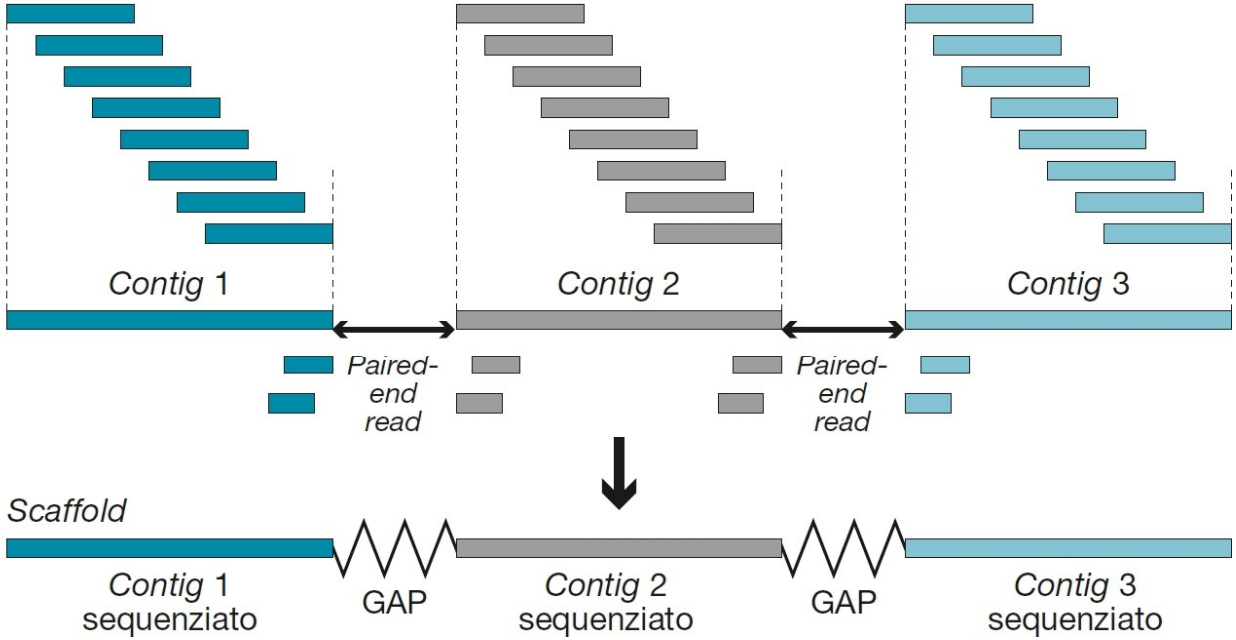


Figura 8.6
Problemi dovuti a sequenze ripetute. Nell'esempio è mostrata una regione genomica contenente una ripetizione presente in tre copie, che causa ambiguità di ricostruzione, dando luogo a due assemblaggi diversi ma altrettanto buoni, di cui uno solo è quello giusto (ma non si può stabilire quale).

Figura 8.7
Dai *contig* agli *scaffold*. Le *paired read* possono essere usate per formare ponti fra *contig*, permettendo di stimarne posizione reciproca, distanza e orientamento, ottenendo gli *scaffold*. Non permettono però di colmare le parti non sequenziate o non ricostruite fra di essi (gap).



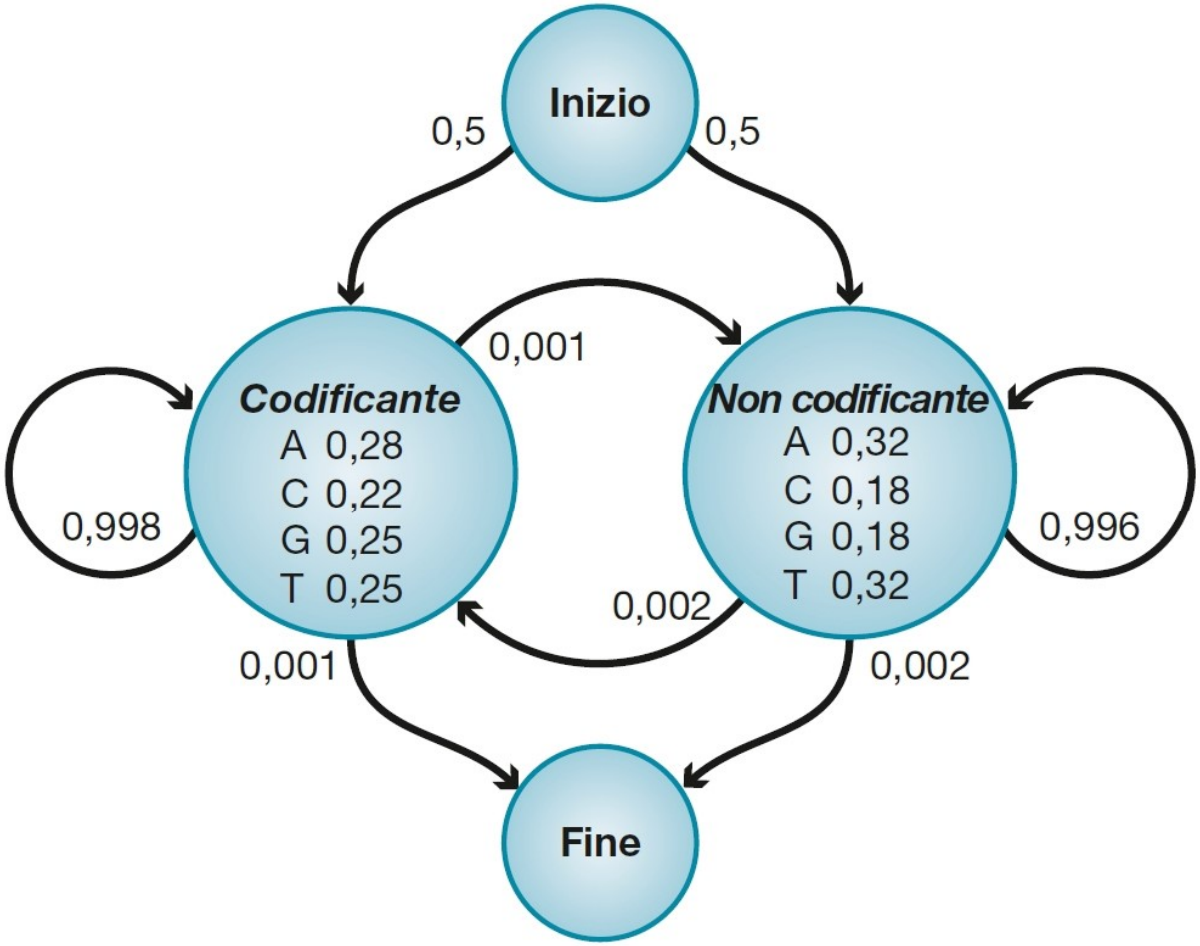


Figura 8.8
Un esempio di HMM, composto da due stati, uno che rappresenta la parte codificante di un genoma, l'altro la parte non codificante. Le frequenze dei 4 nucleotidi corrispondenti a questi due stati sono stimate da esempi noti, così come le probabilità di permanere in un dato stato o di passare da uno all'altro.

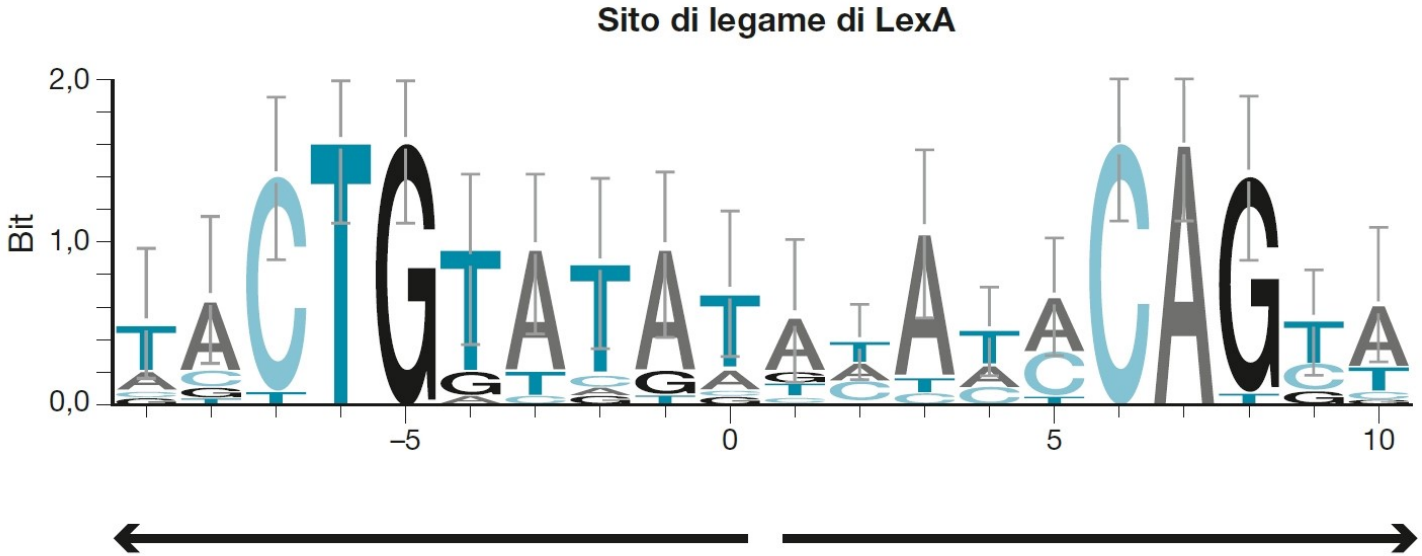
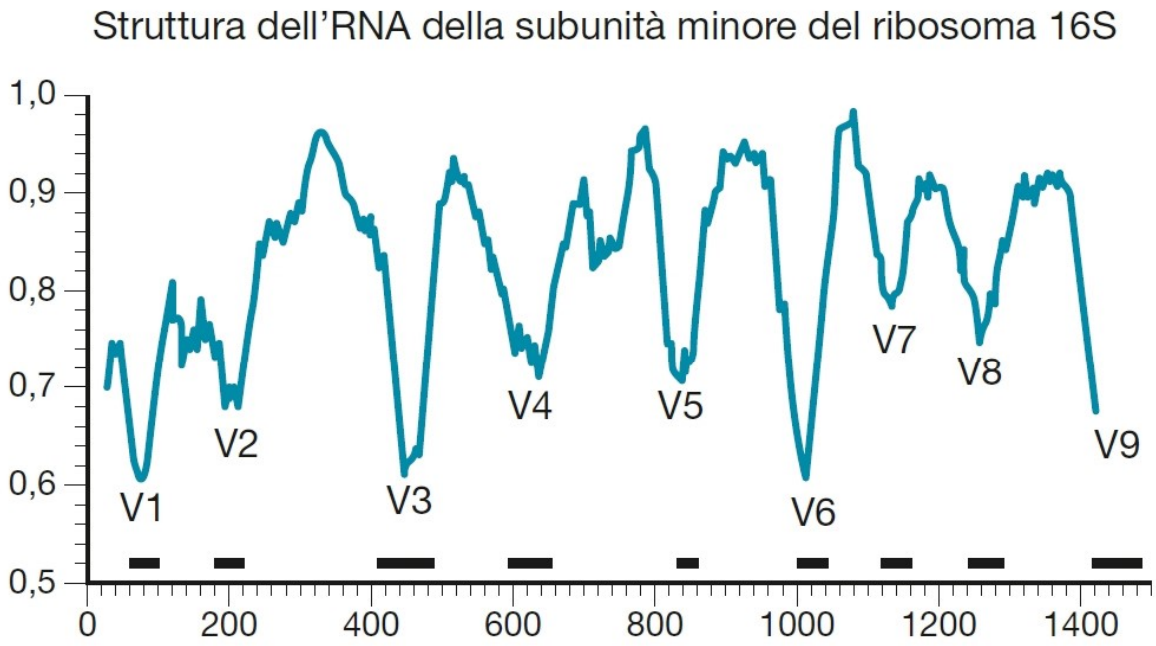


Figura 8.9
L'altezza di ciascuna base *a* nella posizione *i* è proporzionale alla sua frequenza relativa in quella posizione. L'altezza complessiva di ogni colonna *i* è proporzionale alla conservazione dei residui in quella posizione e inversamente proporzionale alla sua entropia, cioè a quanta variabilità si osserva in ogni posizione del sito di legame.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	**1**	**0**	**1**	**5**	**32**	**5**	**35**	**23**	**34**	**14**	**43**	**13**	**34**	**4**	**52**	**3**
C	**50**	**1**	**0**	**1**	**5**	**6**	**0**	**4**	**4**	**13**	**3**	**8**	**17**	**51**	**2**	**0**
G	**0**	**0**	**54**	**15**	**5**	**5**	**12**	**2**	**7**	**1**	**1**	**3**	**1**	**0**	**1**	**52**
T	**5**	**55**	**1**	**35**	**14**	**40**	**9**	**27**	**11**	**28**	**9**	**32**	**4**	**1**	**1**	**1**
Somma	**56**	**56**	**56**	**56**	**56**	**56**	**56**	**56**	**56**	**56**	**56**	**56**	**56**	**56**	**56**	**56**

Matrice posizionale di frequenze di residui per il sito di legame del repressore della trascrizione LexA, calcolate da 56 siti di legame noti nella banca dati Prodoric.
Le frequenze relative sono ottenute dividendo le conte di ciascun nucleotide in ogni posizione per il totale (cioè 56).

Figura 8.10
Profilo di conservazione dell'RNA ribosomiale 16S. Si osservano regioni molto conservate, che possono essere usate per disegnare *primer* per amplificazione per PCR, e regioni ipervariabili, indicate da V1 a V9, possibilmente diverse in specie differenti, che possono essere usate per rilevare la presenza di una specie in un campione metagenomico.



Capitolo 10

L’analisi del trascrittoma

Figura 10.1
La complessità del trascrittoma è principalmente dovuta al fatto che uno stesso *locus* genico può dare origine a numerosi trascritti alternativi a causa dell'utilizzo di siti alternativi di inizio e terminazione della trascrizione, dello *splicing* alternativo e delle modificazioni post-trascrizionali quali l'RNA editing o altre modificazioni delle basi [per es. N(6)-metiladenosina, m6A]. Tutti questi fenomeni possono generare trascritti differenti sia nella porzione codificante (e quindi proteine differenti) che in quella non codificante (5'e 3'UTR).
(*Fonte*: Bangru S. e Kalsotra A., *F1000Res*, 2016, 14(5):2668.)

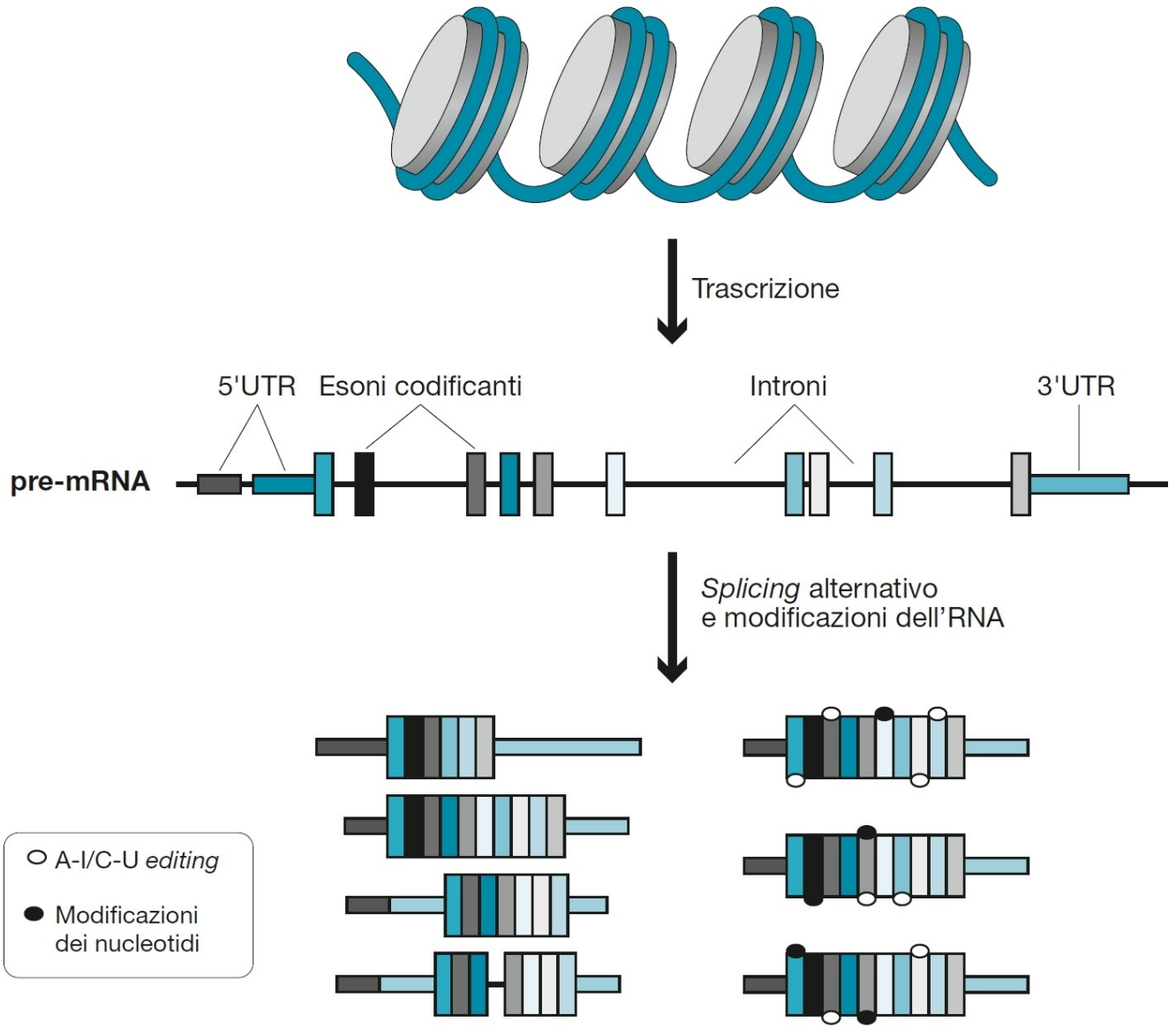


Figura 10.2
Output del software SPLIGN in grado di allineare trascritti multiesonici a sequenze genomiche. Il pannello (**A**) mostra gli ID delle sequenze allineate con i relativi parametri dell'allineamento. Il pannello (**B**) illustra i dettagli dell'allineamento, incluso il modello esoni-introni risultante. È anche possibile visualizzare l'allineamento di ciascuno degli esoni, come nel pannello (**C**).

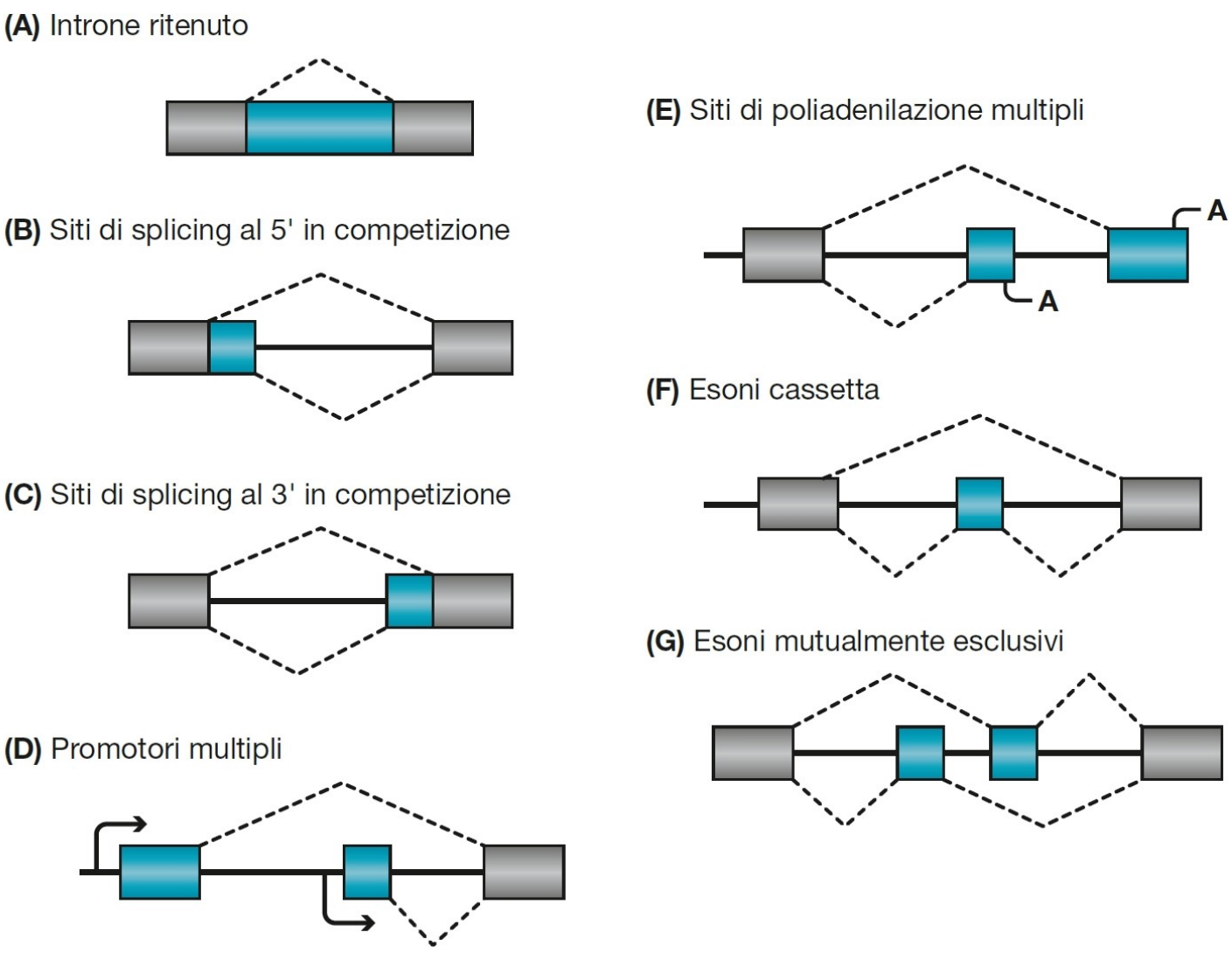


Figura 10.3
Diverse tipologie di eventi di trascrizione e *splicing* alternativi.

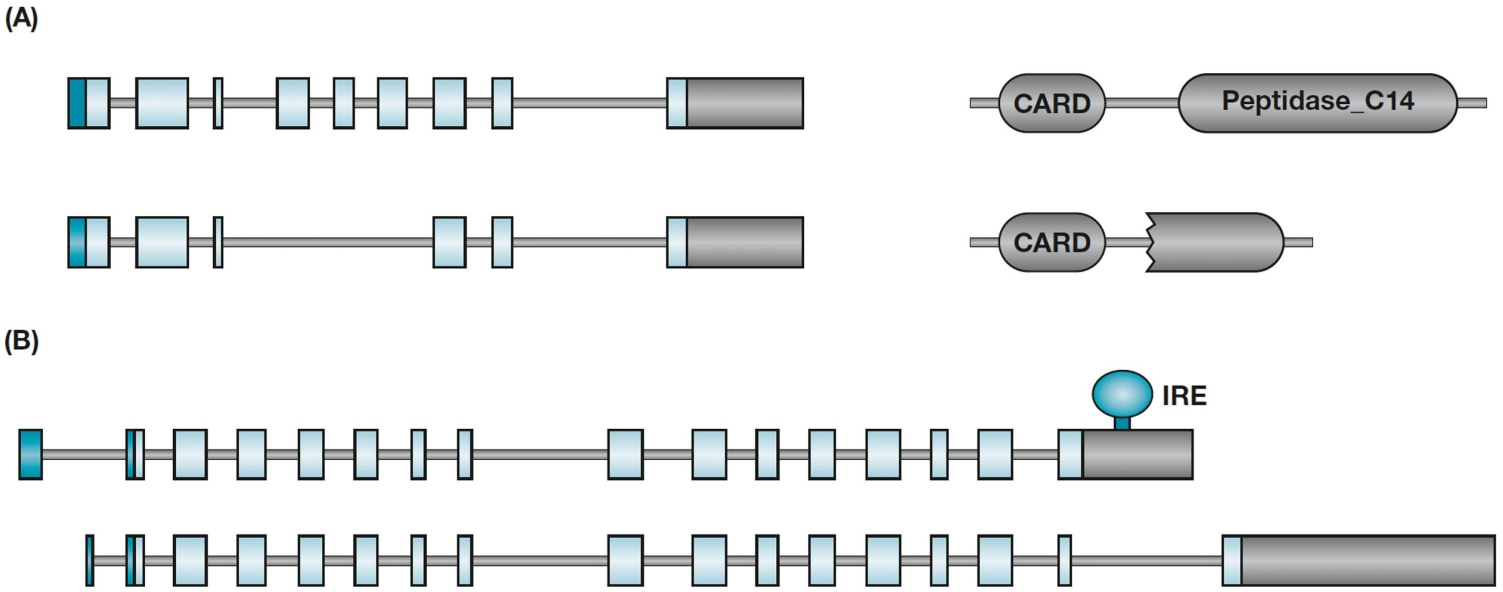


Figura 10.4
(**A**) L'isoforma costitutiva della proteina CASP9 (9 esoni, 416 aa) induce apoptosi. Essa contiene un dominio CARD (*CAspase Recruitment Domain*) e un dominio peptidasico (*Peptidase_C14 caspase domain*). L'isoforma più corta CASP9S (6 esoni, 266 aa) contiene un dominio CARD integro e un dominio peptidasico tronco, per cui non avendo attività proteolitica agisce da inibitore dell'apoptosi. (**B**) Il gene *SLC11A2* (*divalent cation transporter*) codifica per almeno due isoforme, solo una delle quali sensibile alla concentrazione di ferro (i livelli di proteina aumentano sensibilmente in assenza di ferro). Questo meccanismo di regolazione è mediato da un *Iron Responsive Element* (IRE) localizzato nel 3'UTR di una delle due isoforme. Nell'uomo, il trascritto dotato di IRE (16 esoni) codifica per una proteina di 561 aa (NM_000617). Il trascritto privo di IRE (17 esoni) non è presente nel database RefSeq e codifica una proteina di 568 aa.

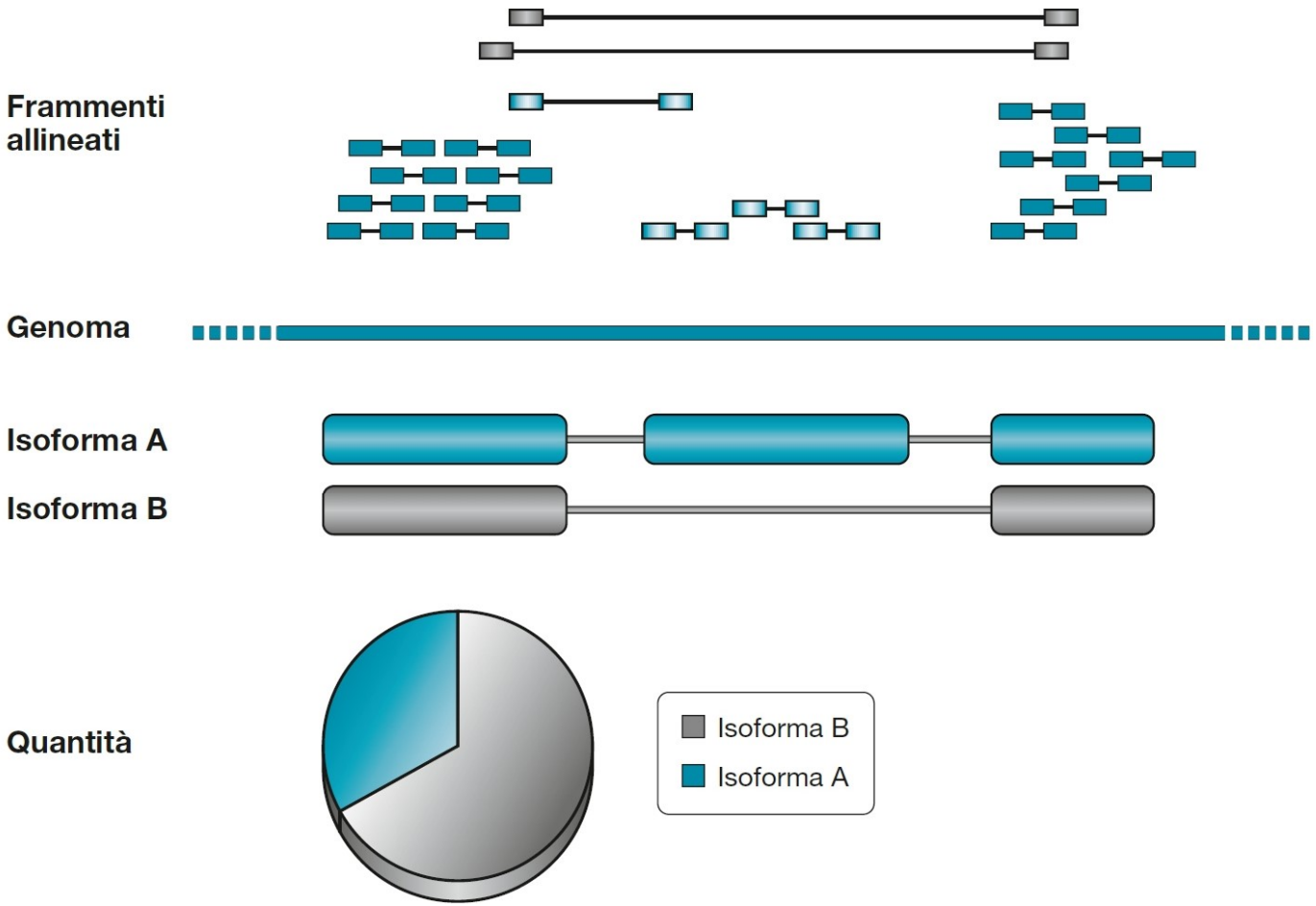


Figura 10.5
L'abbondanza relativa dei trascritti alternativi espressi da uno stesso gene viene determinata con metodi statistici che considerano le read univocamente attribuibili a una specifica isoforma (in blu chiaro e in grigio), considerando anche le read mappate su entrambe le isoforme (in blu scuro). (*Fonte*: https://cgrlucb.wikispaces.com/Isoform+Deconvolution+and+Unannotated+Species)

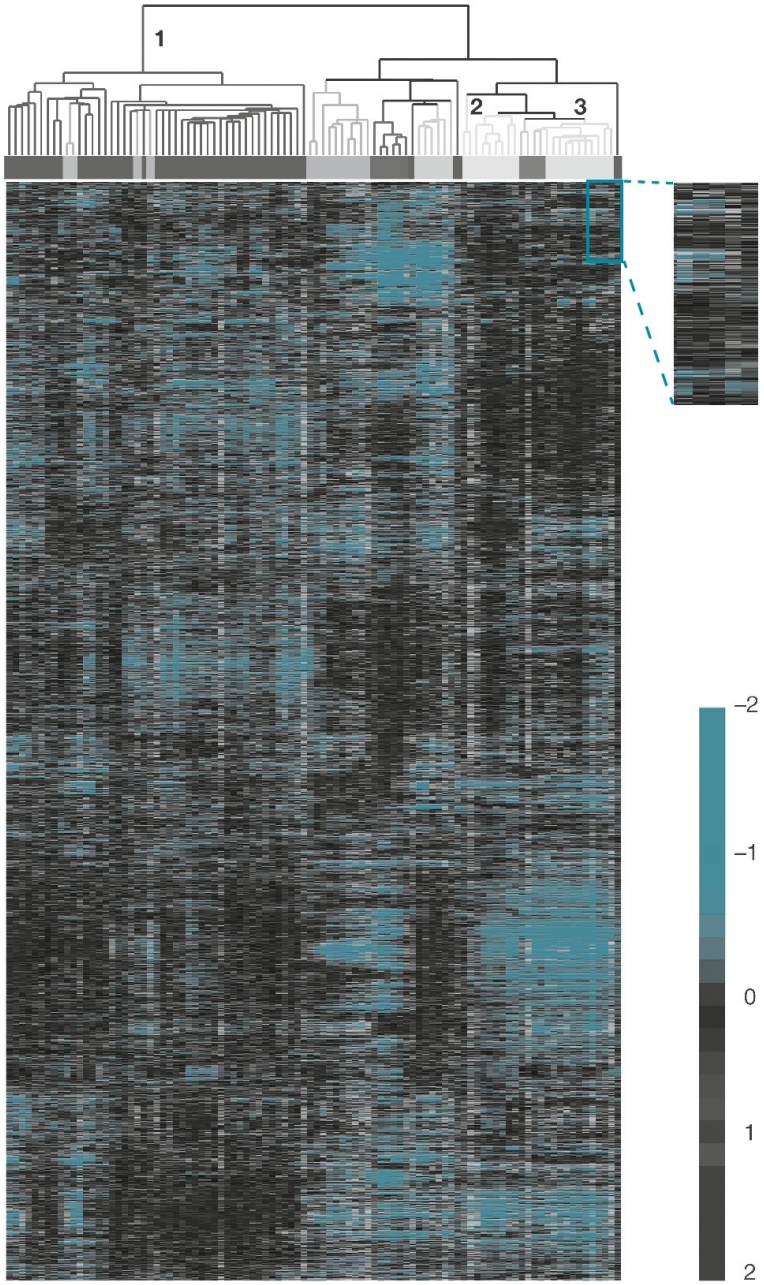


Figura 10.6
Tipica *heatmap* dei trascritti differenzialmente espressi, dove sulle righe abbiamo i diversi geni e sulle colonne i diversi campioni.

Capitolo 11

La struttura dell'RNA

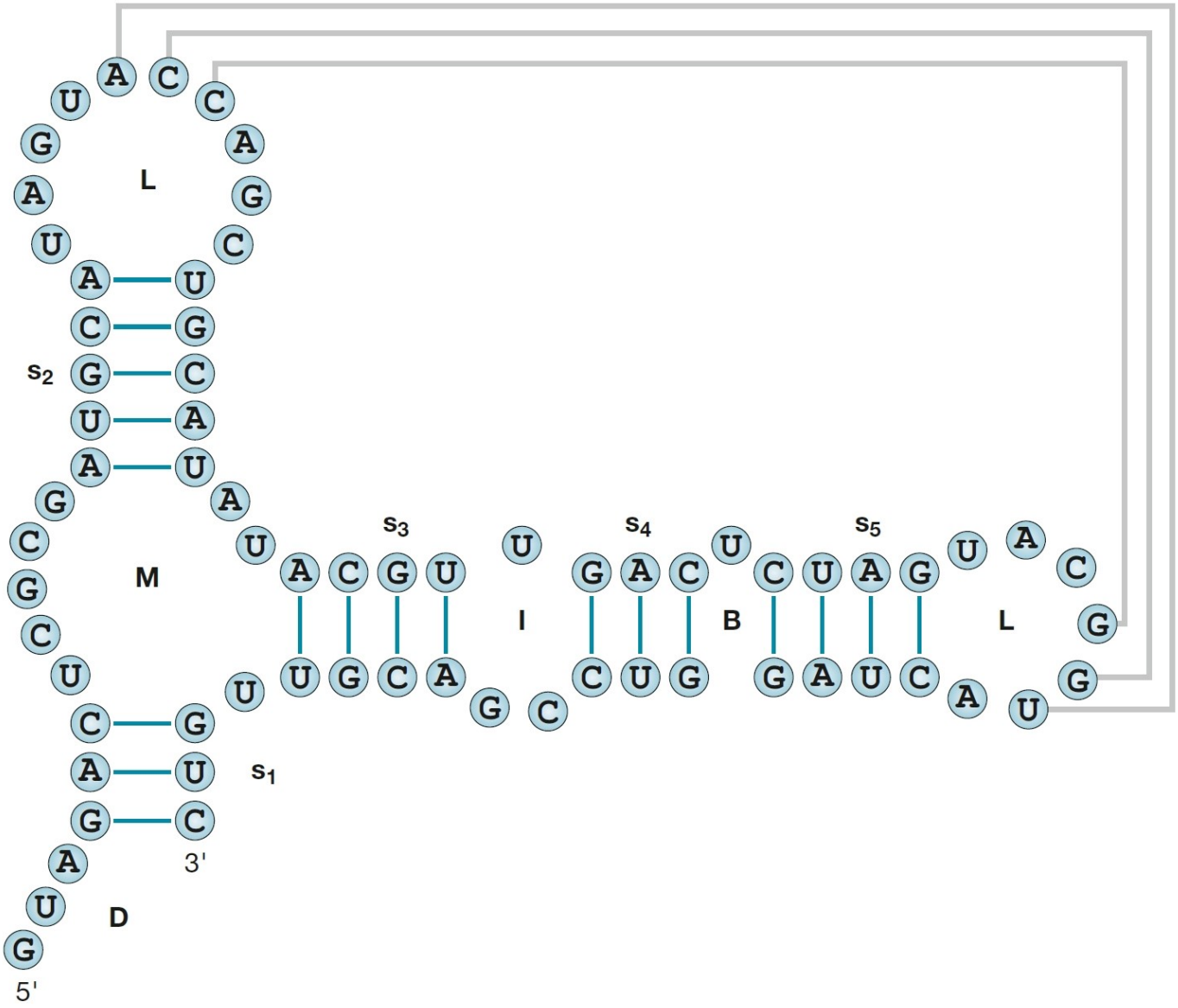


Figura 11.1
Motivi strutturali elementari delle strutture secondarie di RNA: *stem* (s1-s5); *hairpin loop* (L); *multibranched loop* (M); *internal loop* (I); *bulge* (B); *dangling end* o *junction* (D). Le linee in grigio chiaro mostrano anche un possibile *pseudoknot*.

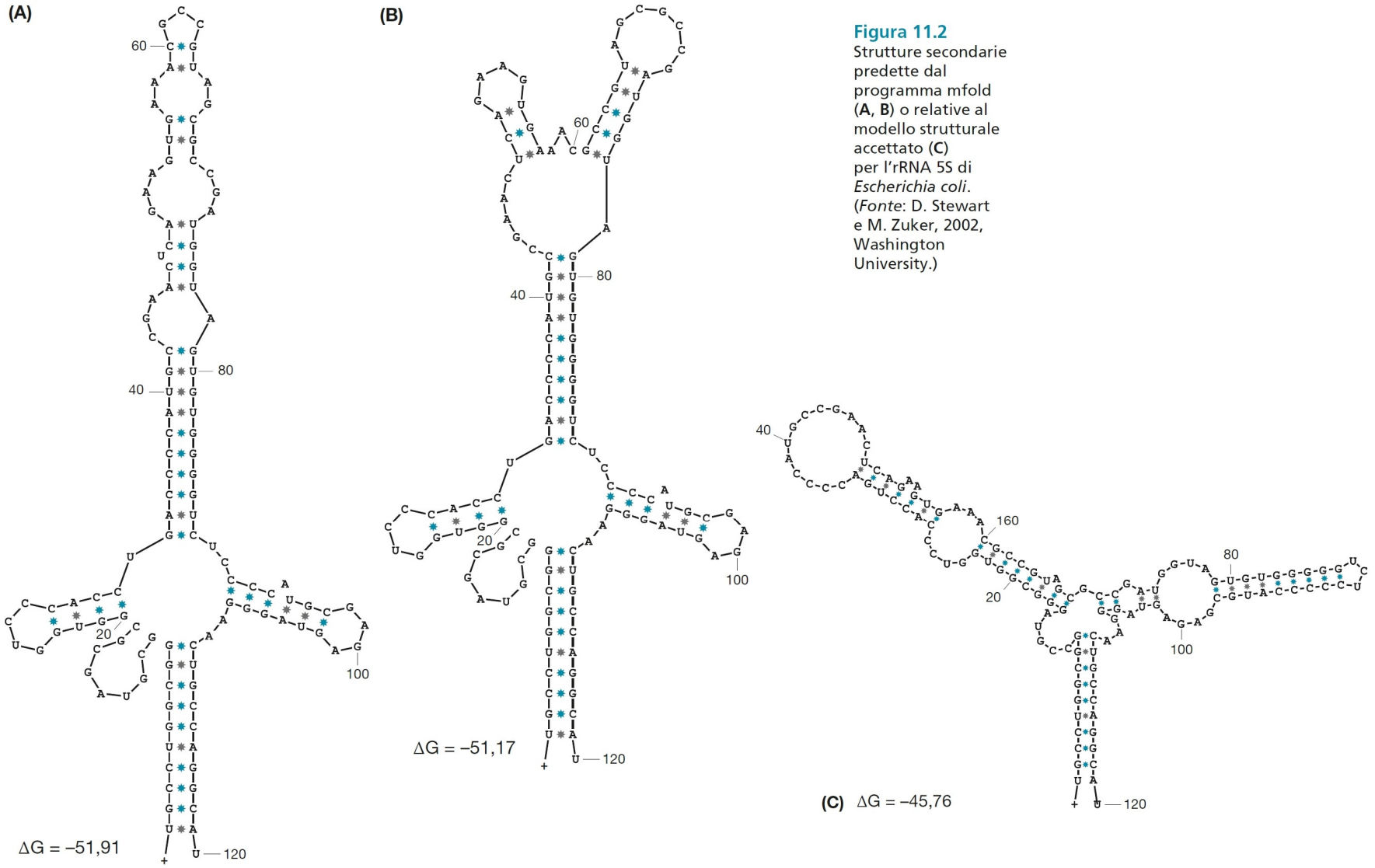


Figura 11.2
Strutture secondarie predette dal programma mfold (**A**, **B**) o relative al modello strutturale accettato (**C**) per l'rRNA 5S di *Escherichia coli*. (*Fonte*: D. Stewart e M. Zuker, 2002, Washington University.)

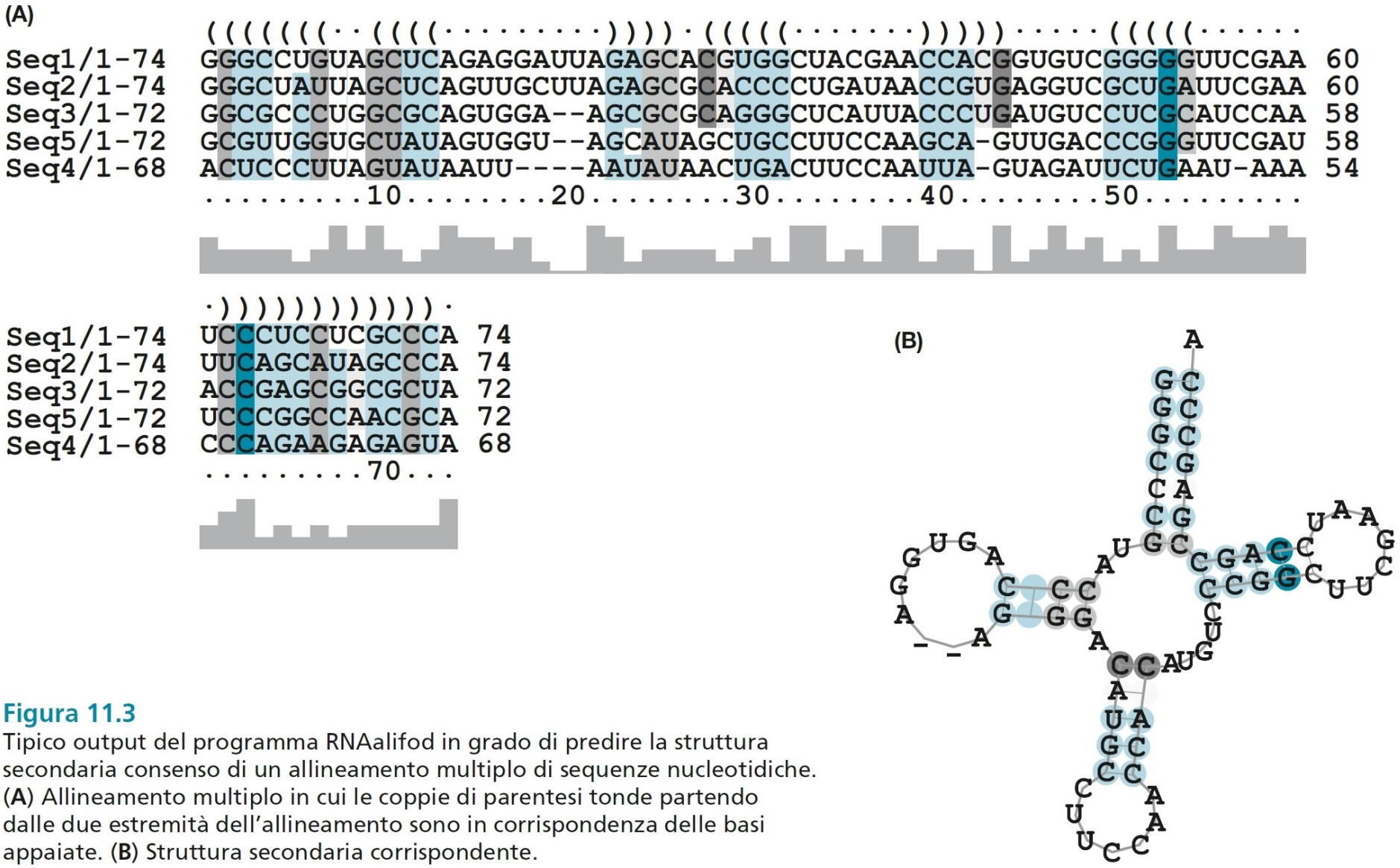
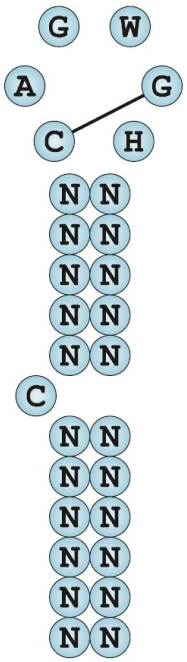


Figura 11.3
Tipico output del programma RNAalifod in grado di predire la struttura secondaria consenso di un allineamento multiplo di sequenze nucleotidiche. (**A**) Allineamento multiplo in cui le coppie di parentesi tonde partendo dalle due estremità dell'allineamento sono in corrispondenza delle basi appaiate. (**B**) Struttura secondaria corrispondente.

(A)



(B)

```
r1 = {au,ua,gc,cg,gu,ug}
p1 = 2...8 c p2 = 5...5 CAGWGH r1~ p2 r1~ p1
```

(C)

Nome dell'entry | Posizione del pattern | Pattern

```
R000001C10 : [37,59] :     TTG  C TTCAA CAGTGT TTGAA  CGG
R000001M11 : [35,57] :     TTG  C TTCAA CAGTGT TTGAA  CGG
R000013E09 : [339,361] :   GCA  C TGGGC CAGTGT GTCTA  TGT
ZX00044F24 : [1216,1240] : AGGC C TGGGT CAGAGC ACCTG GCTT
ZX00045P16 : [625,647] :   GCA  C TGGGC CAGTGT GTCTA  TGT
ZX00047P10 : [36,58] :     TTG  C TTCAA CAGTGT TTGAA  CGG
ZX00048A04 : [37,59] :     TTG  C TTCAA CAGTGT TTGAA  CGG
PX00006J03 : [146,168] :   TGG  C TTCGG CAGAGC CCGGG  CTA
PX00020018 : [229,251] :   ACT  C TGATT CAGAGA AGTTG  AGT
R000002J21 : [35,57] :     TTG  C TTCAA CAGTGT TTGAA  CGG
```

Figura 11.4
Struttura consensus dell'*Iron Responsive Element* (IRE) (**A**) e relativa sintassi adottata dal programma PatSearch (**B**). È mostrato anche un esempio di output prodotto da PatSearch nella ricerca dell'elemento IRE in una collezione di mRNA di topo (**C**).

Capitolo 15

•Interazioni proteiche

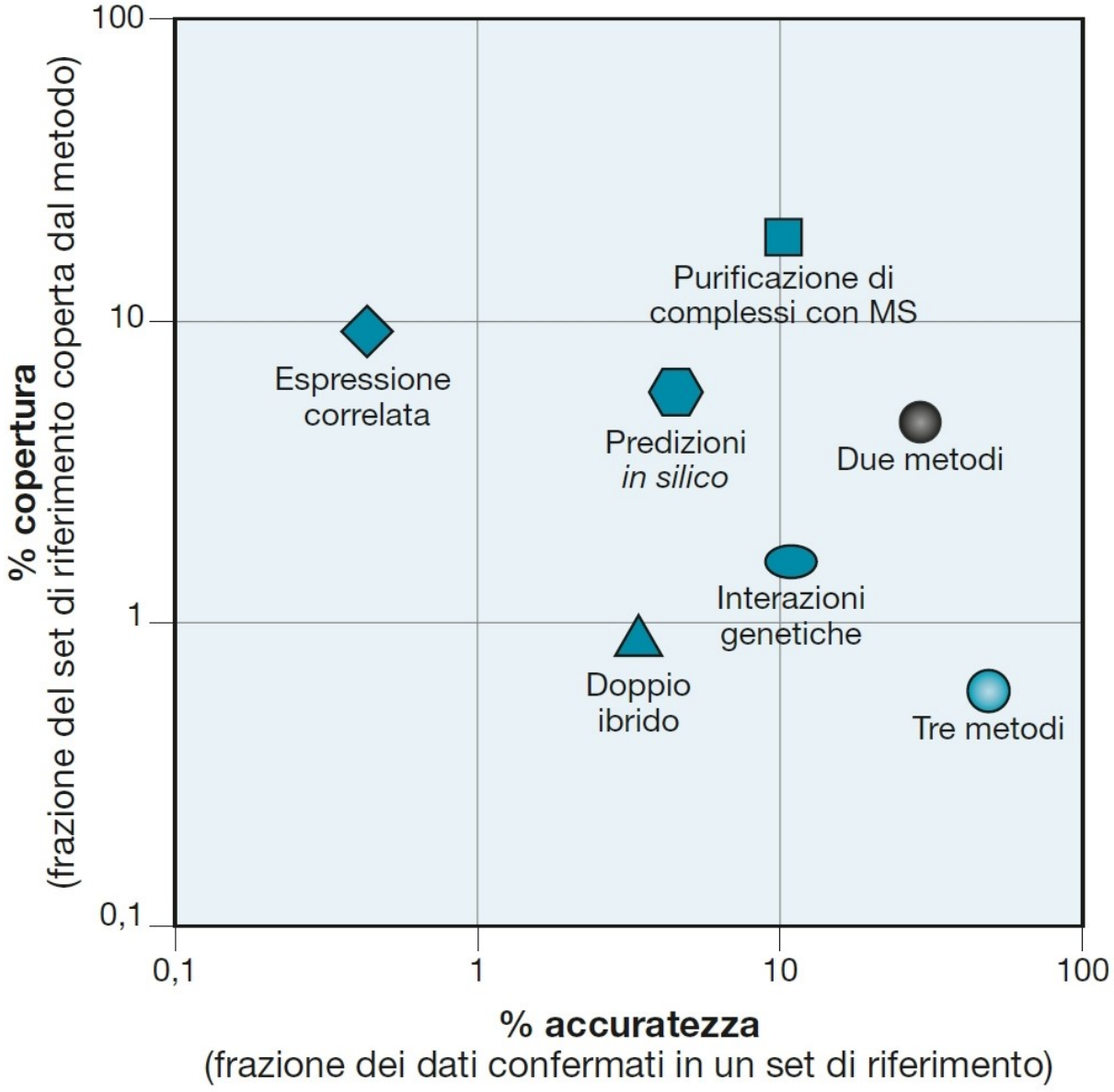
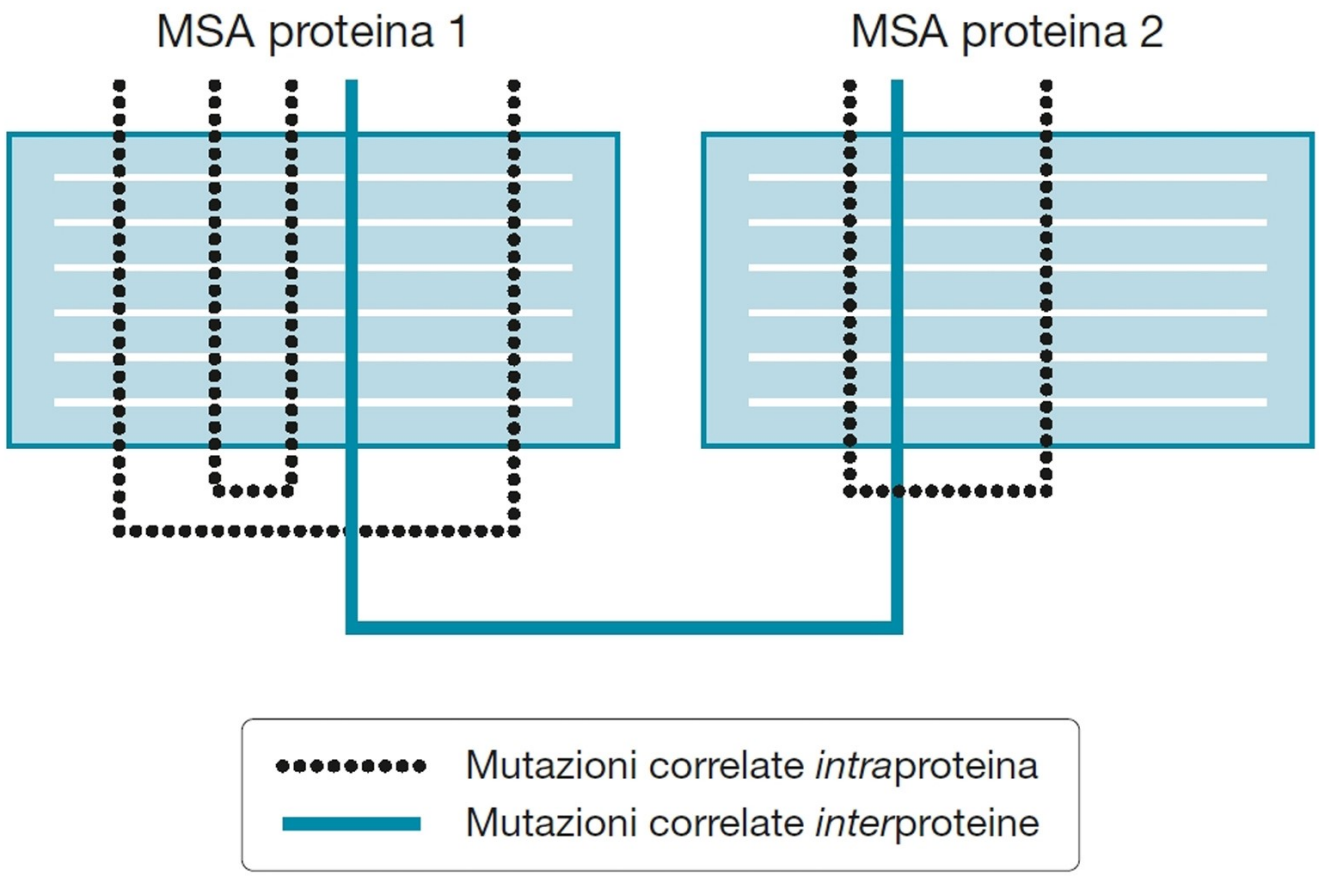


Figura 15.1
Ogni metodo sperimentale o computazionale è associato a una certa accuratezza ed è in grado di dare informazioni su una proporzione più o meno grande dell'interattoma. Considerare l'intersezione dei risultati di più metodi porta ad aumentare l'affidabilità del dato, ma in generale diminuisce il numero di interazioni. (MS = *Mass Spectrometry*, spettrometria di massa.)

Figura 15.2
Mutazioni correlate: dato un allineamento multiplo di proteine omologhe, spesso si osservano posizioni dell'allineamento che variano in maniera coordinata, in modo tale che l'osservazione di un certo aminoacido in una colonna va a specificare quale aminoacido si troverà in un'altra colonna del MSA (*Multiple Sequence Alignment*). Questi casi sono mostrati come righe tratteggiate. Lo stesso principio si può applicare a coppie di proteine interagenti, in cui la presenza di un particolare aminoacido in una specifica posizione di una proteina si rispecchia nella presenza di un particolare aminoacido in una specifica posizione del suo partner di interazione (riga continua). Nei due allineamenti multipli che si confrontano, è essenziale che siano incluse le stesse specie, che devono essere riportate nello stesso ordine in entrambi gli MSA.



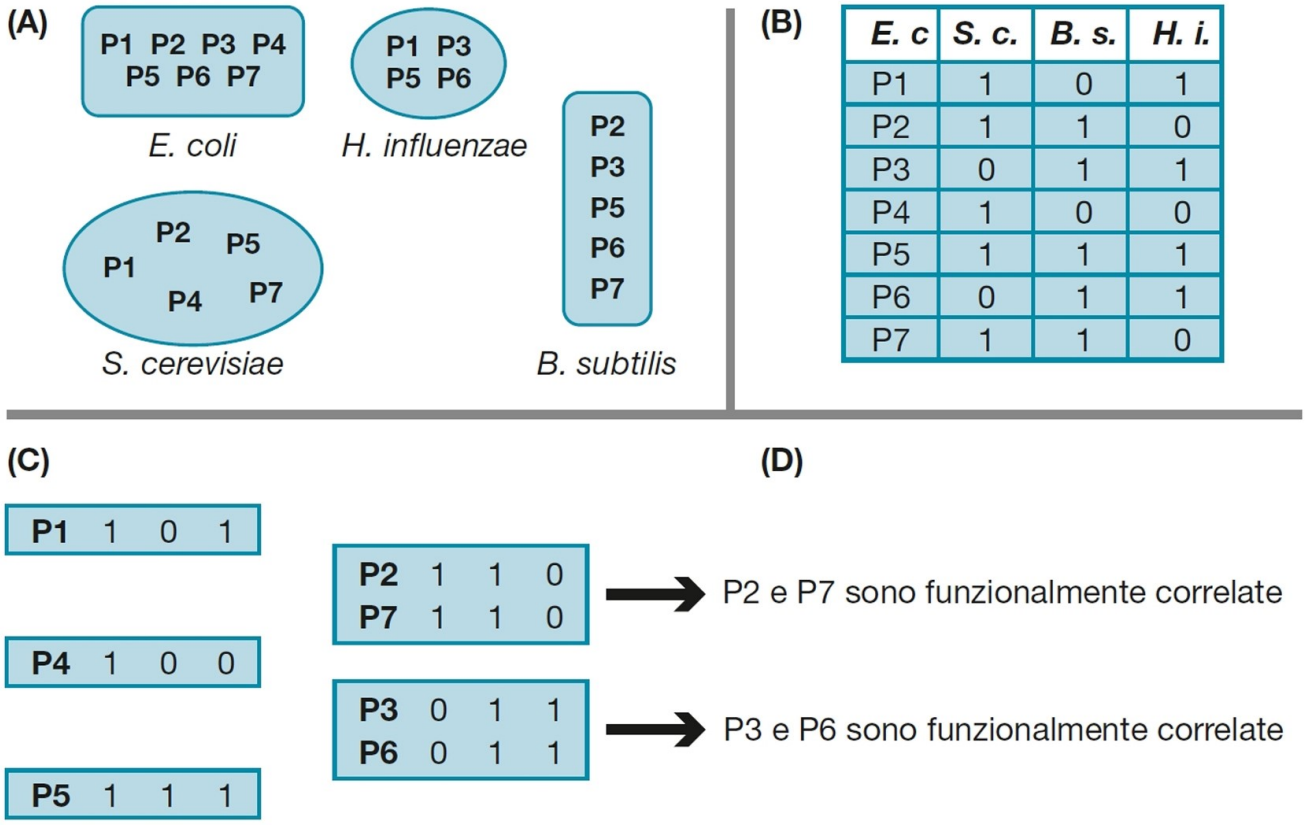


Figura 15.3
Metodo dei profili filogenetici. Supponiamo di voler analizzare le proteine del batterio *Escherichia coli* e di avere a disposizione, per semplicità, solo i genomi di altri tre organismi (*Saccharomyces cerevisiae*, *Bacillus subtilis* e *Haemophilus influenzae*). Per poter costruire il profilo filogenetico di alcune proteine di *E. coli* si dovrà prima di tutto stabilire se esistano le corrispondenti proteine ortologhe negli altri genomi considerati (**A**). Il profilo filogenetico di una proteina sarà costituito dalle informazioni sulla presenza (1) o assenza (0) di un ortologo in ognuno dei genomi considerati (**B**). Proteine con lo stesso profilo filogenetico vengono raggruppate insieme e si ipotizza che siano funzionalmente correlate (**C** e **D**).

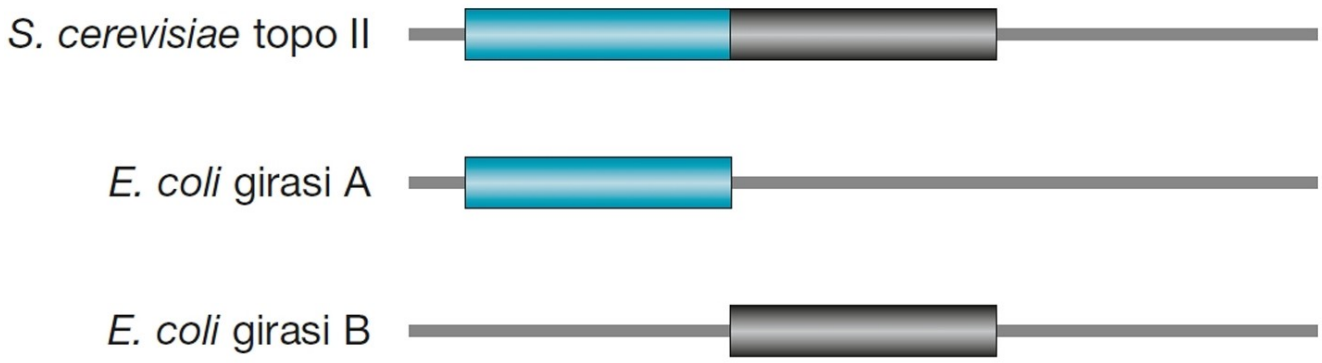


Figura 15.4
Metodo della stele di Rosetta. Se per una proteina, come per esempio la topoisomerasi II di lievito, si trovano due diversi ortologhi nel genoma di un'altra specie corrispondenti a due porzioni diverse della proteina, come per esempio le girasi A e B nel genoma di *E. coli*, si può allora assumere che questi due ortologhi possano interagire fra di loro per ripristinare l'unità funzionale.

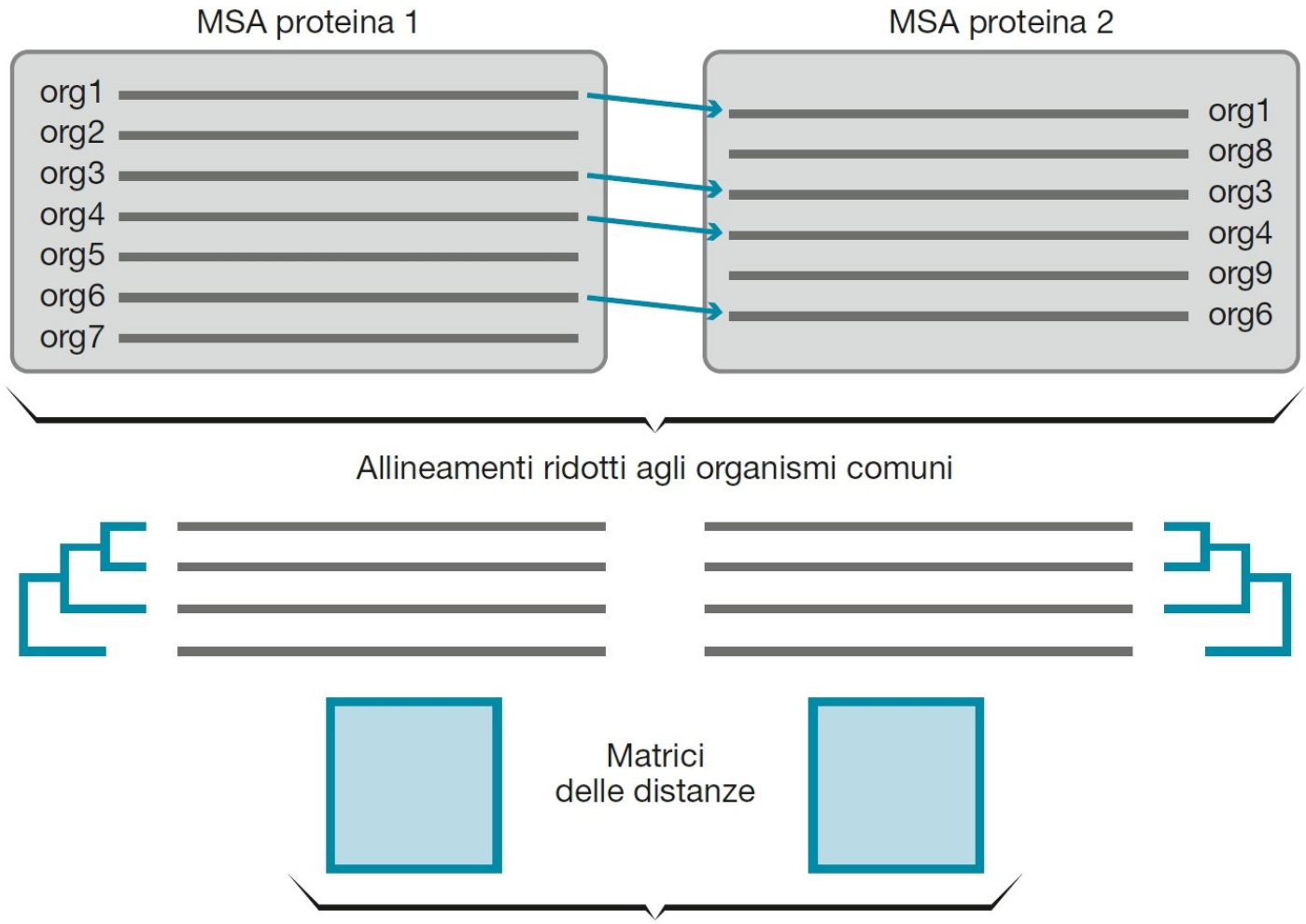


Figura 15.5
Metodo della similarità degli alberi filogenetici.

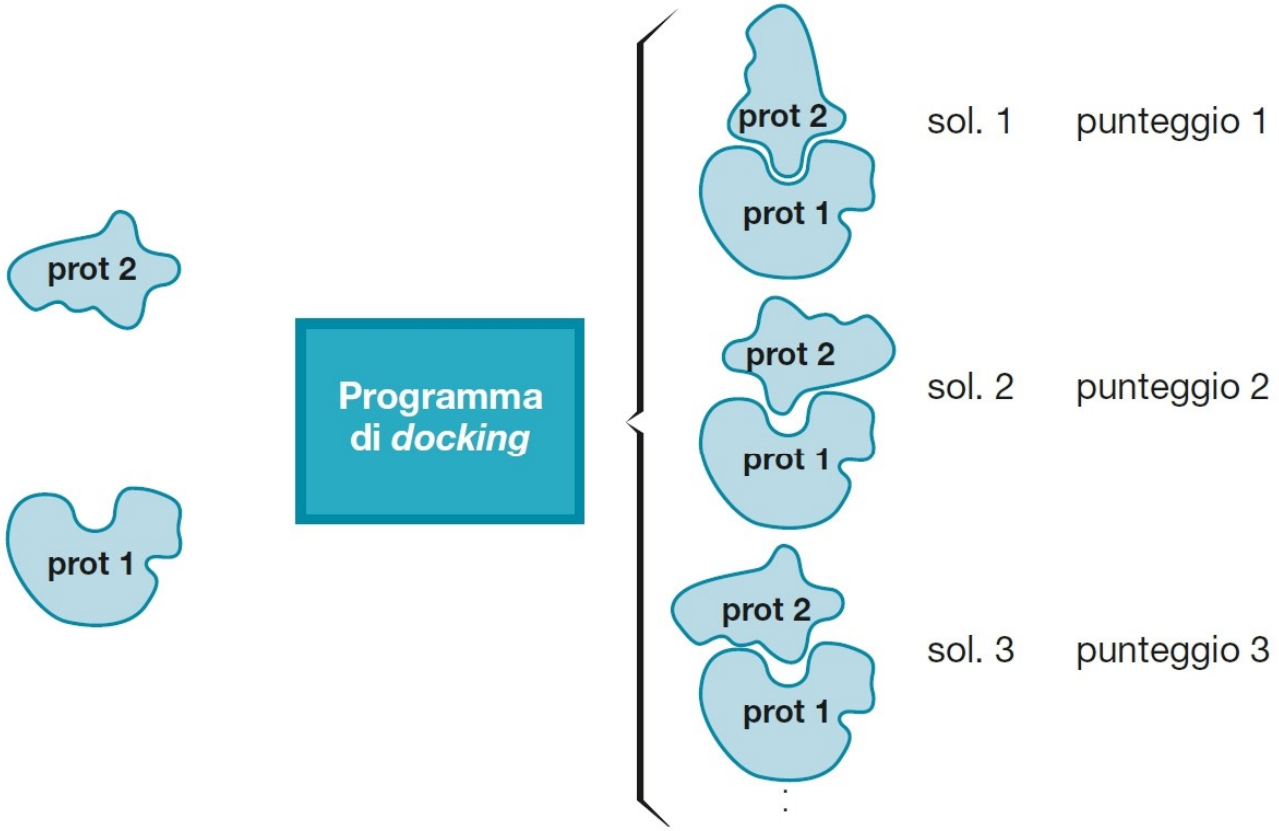


Figura 15.6
Docking: date due strutture di proteine potenzialmente interagenti, qui rappresentate come forme bidimensionali, un algoritmo di *docking* può proporre vari possibili complessi alternativi, ognuno associato a un punteggio.

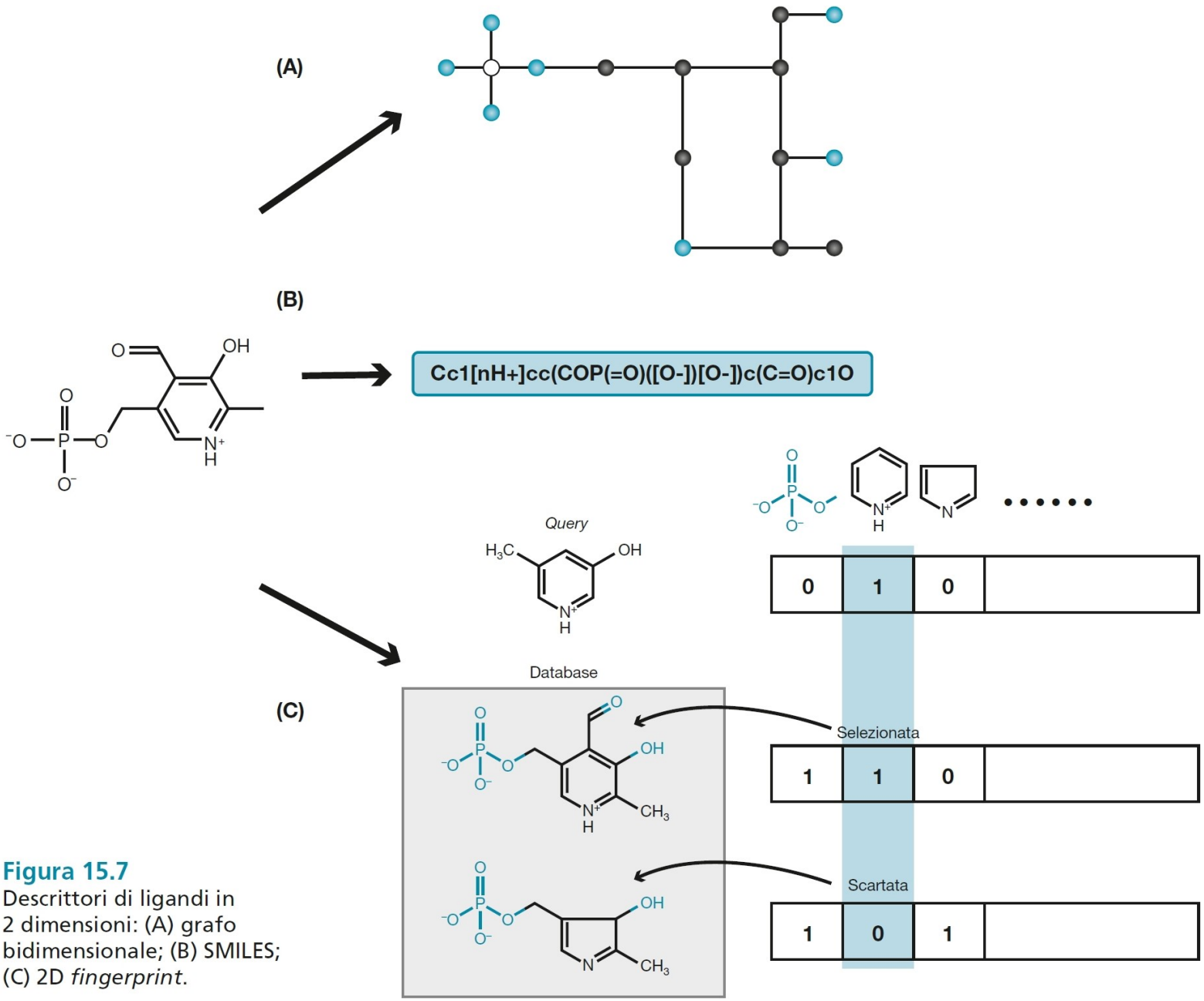


Figura 15.7
Descrittori di ligandi in 2 dimensioni: (A) grafo bidimensionale; (B) SMILES; (C) 2D *fingerprint*.

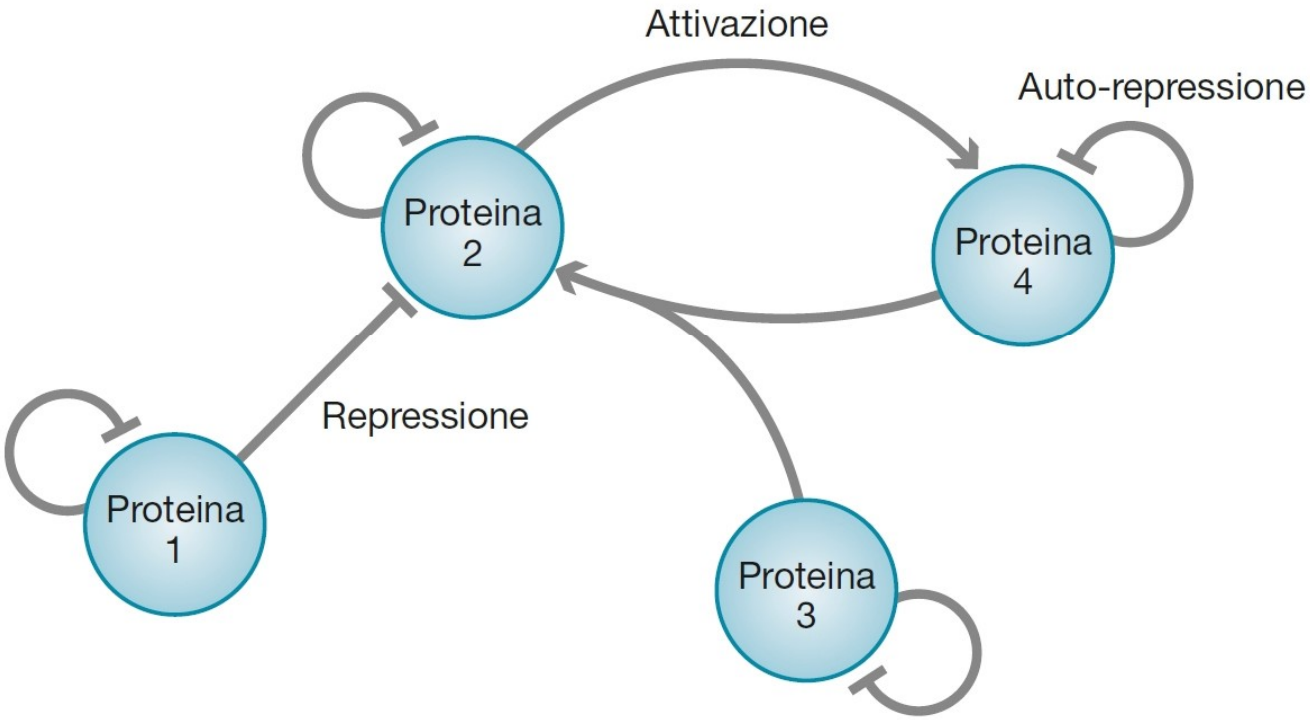
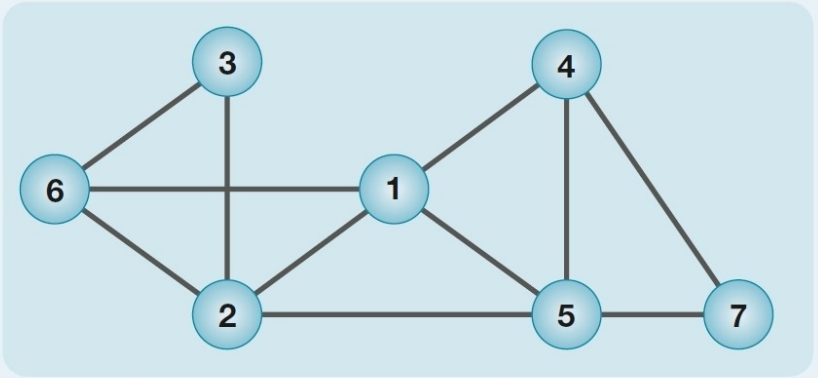


Figura 15.8
Grafo di interazione fra quattro proteine. Nel grafo, gli archi descrivono l'interazione fra proteine e l'effetto di tale interazione. Archi che terminano in una freccia indicano che il nodo (cioè la proteina) da cui l'arco fuoriesce ha un'attività attivatoria sul nodo in cui l'arco arriva. Archi che terminano in un trattino indicano attività repressoria. Archi che escono ed entrano dallo stesso nodo indicano attività che la proteina ha su se stessa, per esempio auto-repressione.

(A)



V = {1,2,3,4,5,6,7}
E = {3-6,3-2,6-1,6-2,1-2,1-5,2-5,1-4,4-5,4-7,5-7}

G = (V,E)

Grado V1: 4
Shortest-path tra V1-V3: 2
Diametro: 4
Coefficiente di *clustering* V6: 2/3

(B) *Scale-free* network

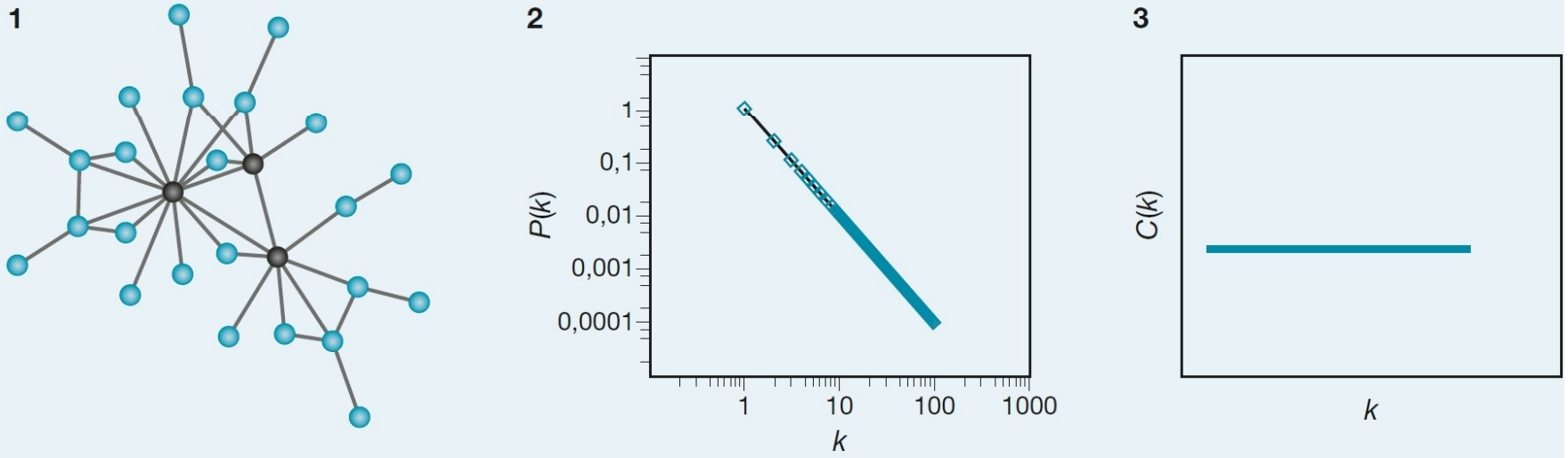


Figura 15.9
Teoria dei Grafi. (**A**) È riportato un esempio di network non-orientato composto di 7 nodi e 11 archi. Per alcuni nodi vengono riportate le misure topologiche descritte nel testo. (**B**) Definizione di network *scale free*, grafico della relazione grado vs frequenze in scala logaritmica e della relazione coefficiente di *clustering* e grado. La relazione grado frequenza è esponenziale negativa che, trasformata in logaritmo, diventa lineare inversa.

Capitolo 16

La bioinformatica: tra presente e futuro

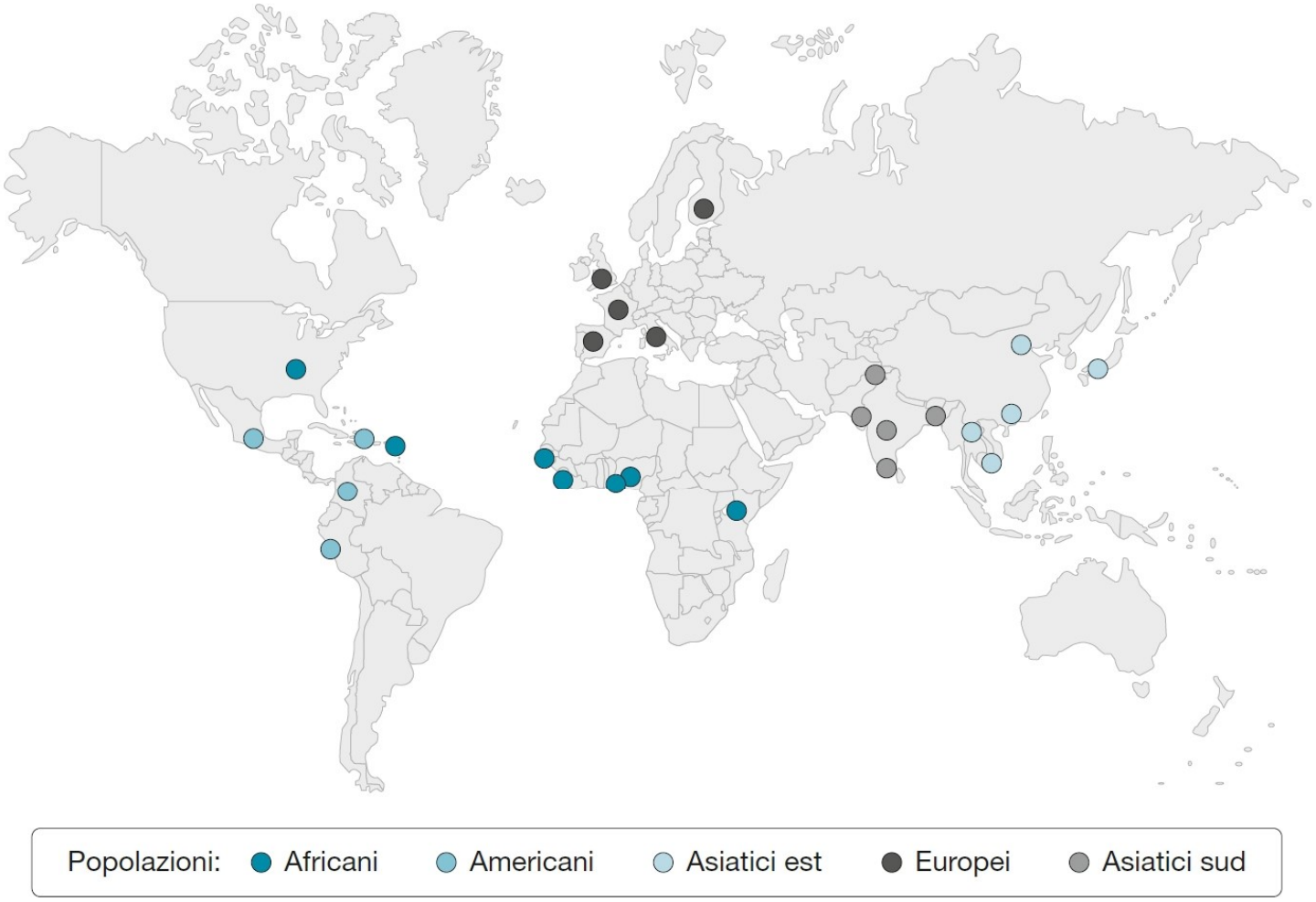


Figura 16.1
Localizzazione geografica e divisione per etnia dei campioni utilizzati nel Progetto 1000 Genomi.

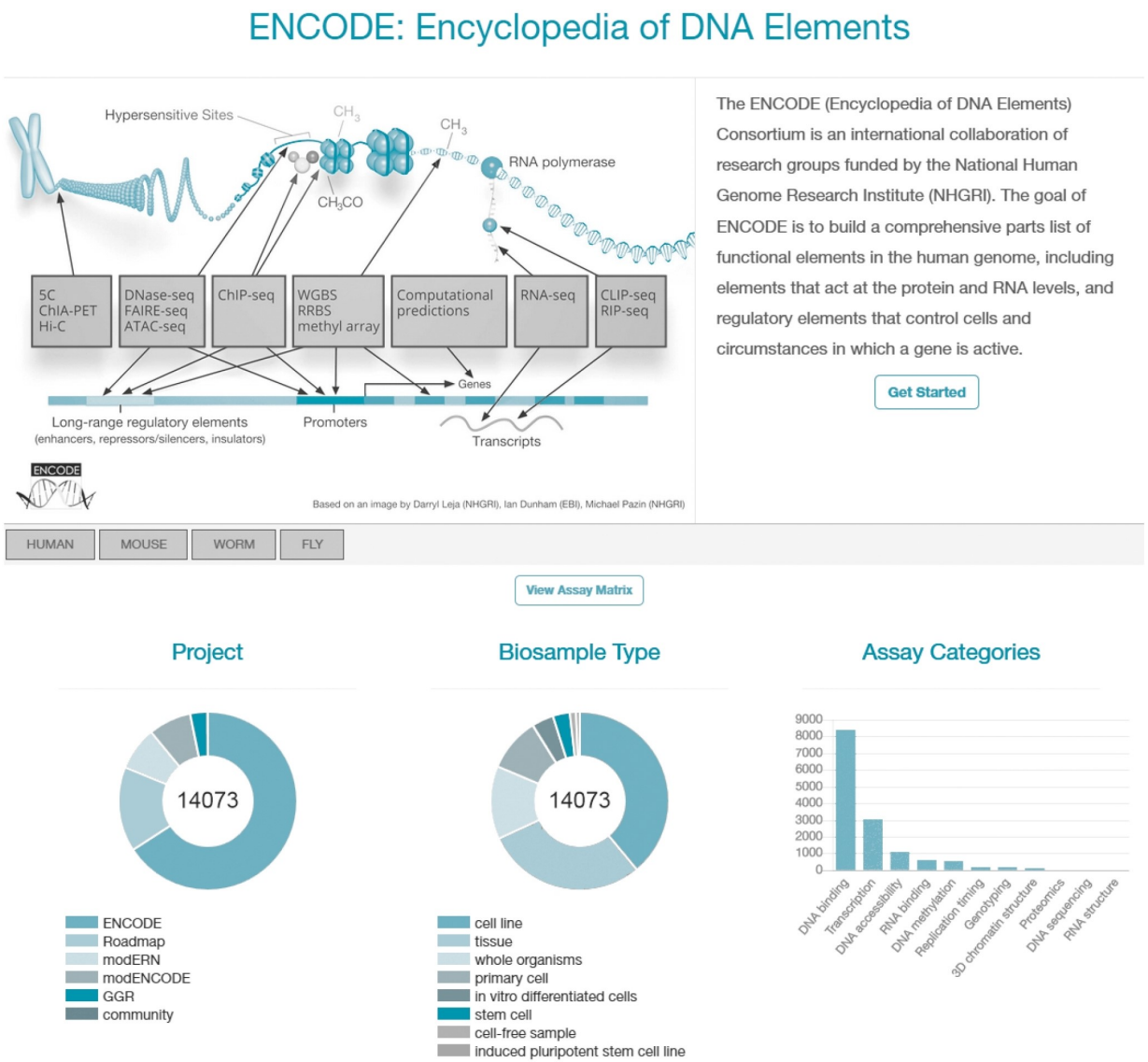


Figura 16.2
Pagina iniziale del portale ENCODE che racchiude tutti i dati dei progetti ENCODE, modENCODE e Roadmap. Nella prima parte vengono riportati in uno schema gli innumerevoli tipi di esperimenti utilizzati per studiare i diversi tipi di elementi regolativi. Nella seconda parte della pagina web vengono riportate in modo dinamico (cioè aggiornati al momento in cui la pagina viene caricata) per i quattro organismi modello disponibili le statistiche riguardanti il numero di esperimenti divisi per progetto, per tipo di campione e per tipo di esperimento. (*Fonte*: www.encodeproject.org/)

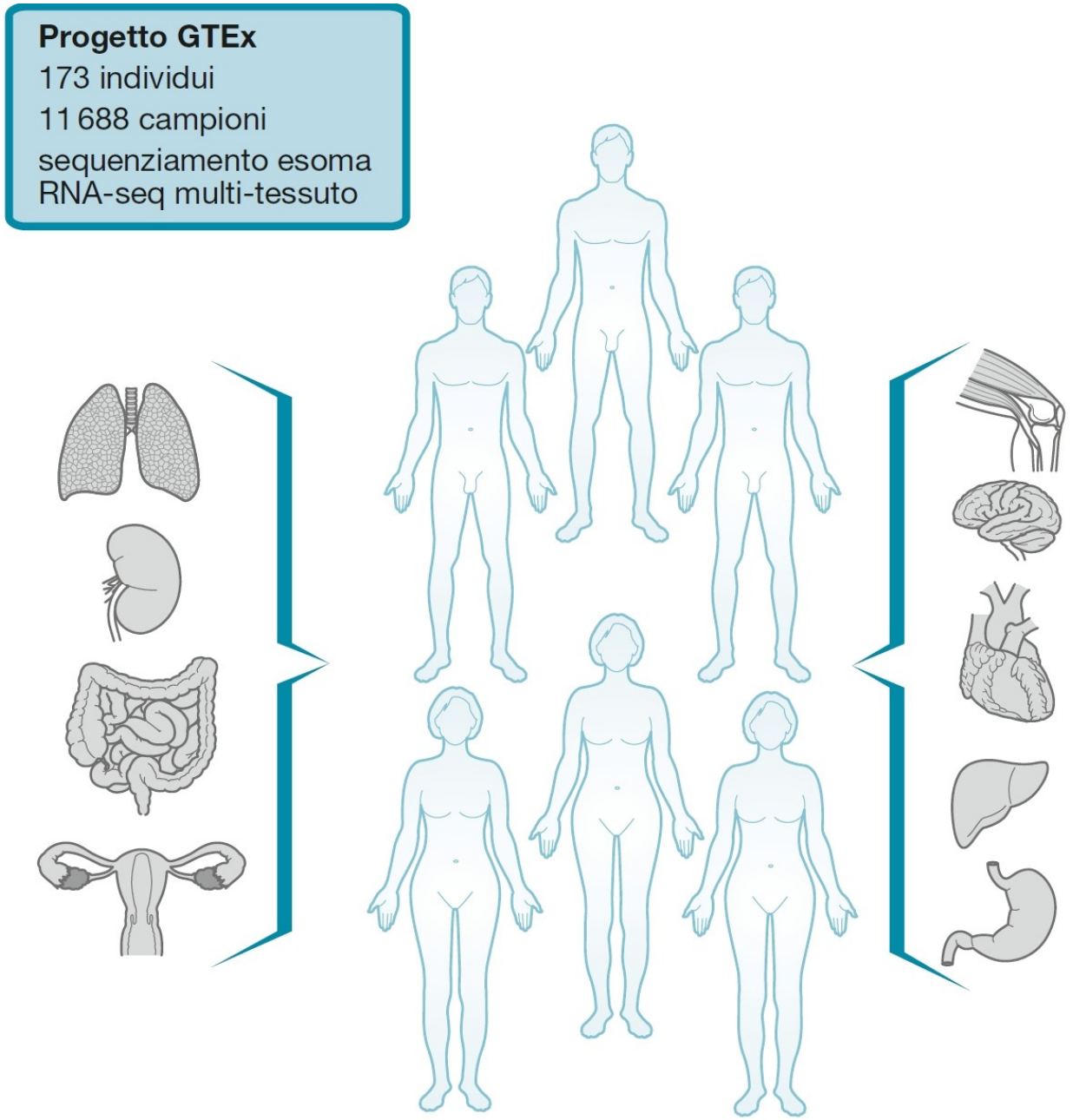


Figura 16.3
Molteplicità dei dati e tessuti considerati nel progetto GTEx.

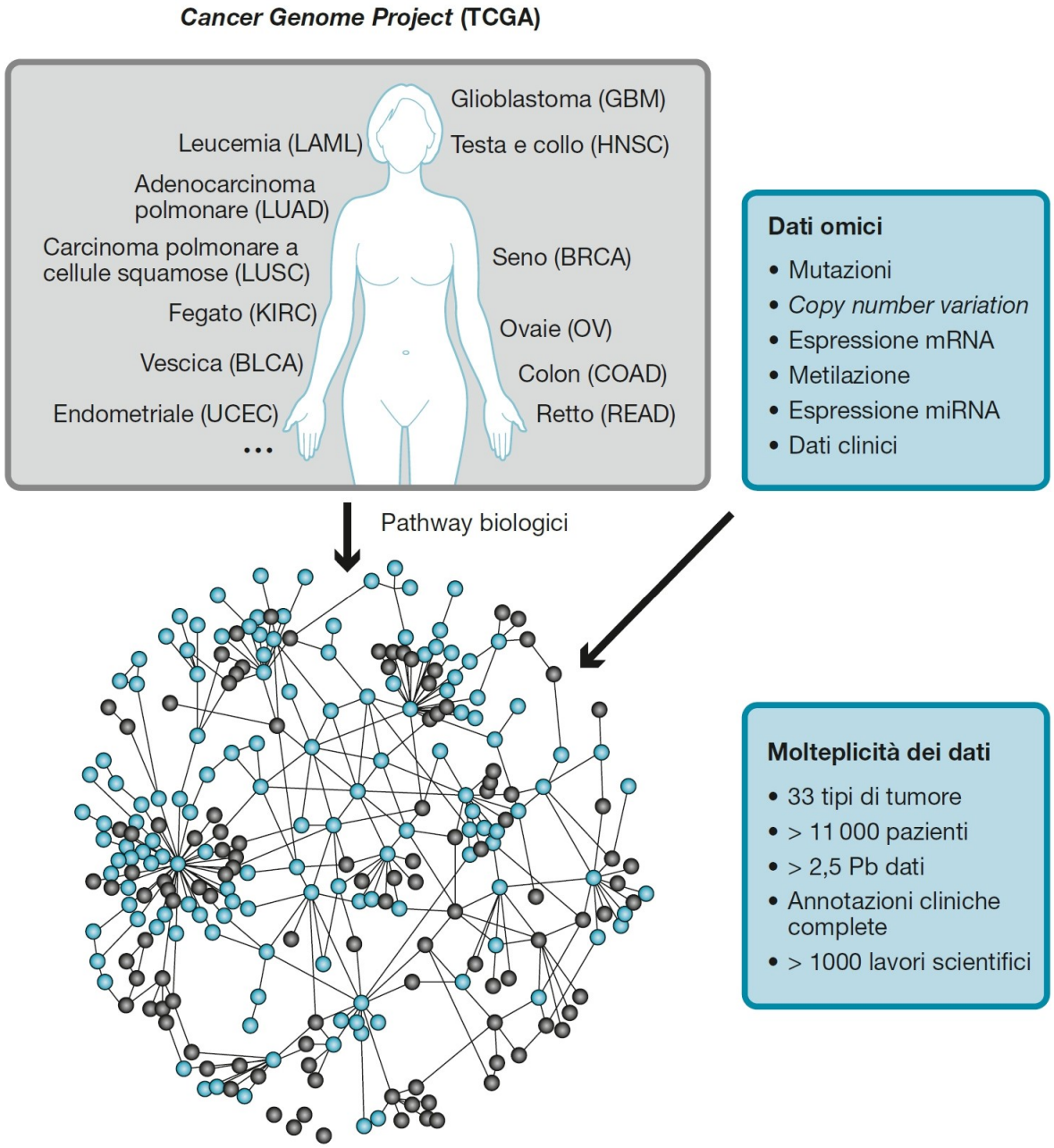


Figura 16.4
Schema del progetto TCGA in cui 33 tessuti tumorali sono stati prelevati da pazienti per generare i profili di mutazione, *copy number variation*, espressione genica (di mRNA e miRNA) e metilazione. Lo scopo del progetto è l'integrazione dei dati omici con le caratteristiche cliniche dei pazienti per definire i marcatori di malattia e identificare i circuiti di regolazione alterati a causa dell'insorgenza del tumore. In evidenza sono riportati alcuni numeri per descrivere la grande massa di dati raccolta.

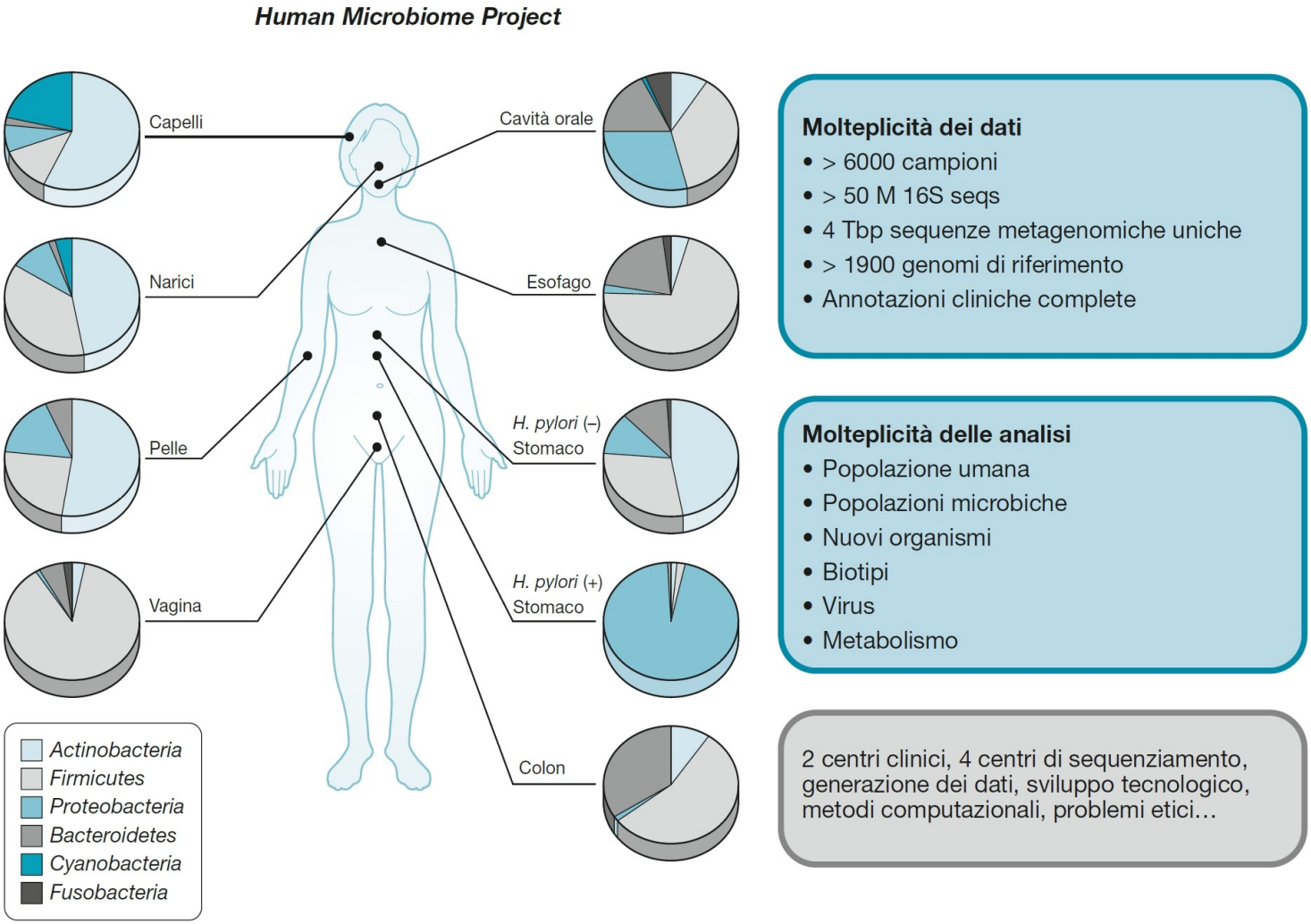


Figura 16.5
Schema riassuntivo dell'organizzazione, dei tipi di campione e dei tipi di dati raccolti dal progetto *Human Microbiome*.

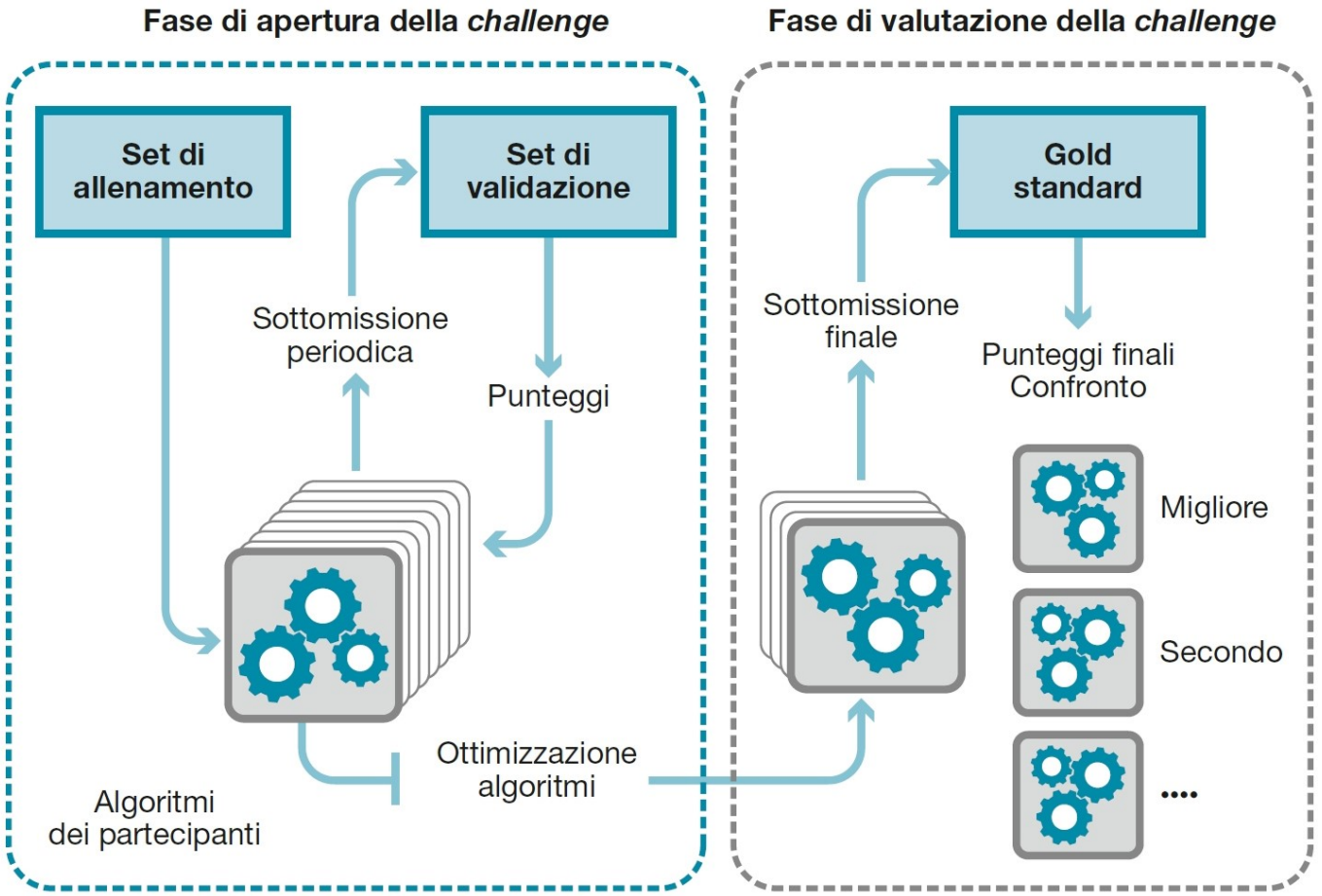


Figura 16.6
Schema tipico di una competizione scientifica. Un set di dati viene suddiviso in un set di allenamento, in un set di validazione e nel gold standard. Dopo l'ottimizzazione e la sottomissione finale, gli algoritmi vengono testati sul gold standard e classificati sulla base di diverse misure di bontà del metodo, quali sensibilità, specificità, accuratezza ecc. (*Adattata da*: Butros P.C. *et al.*, *Genome Biology*, 2014, 15(9):462.)

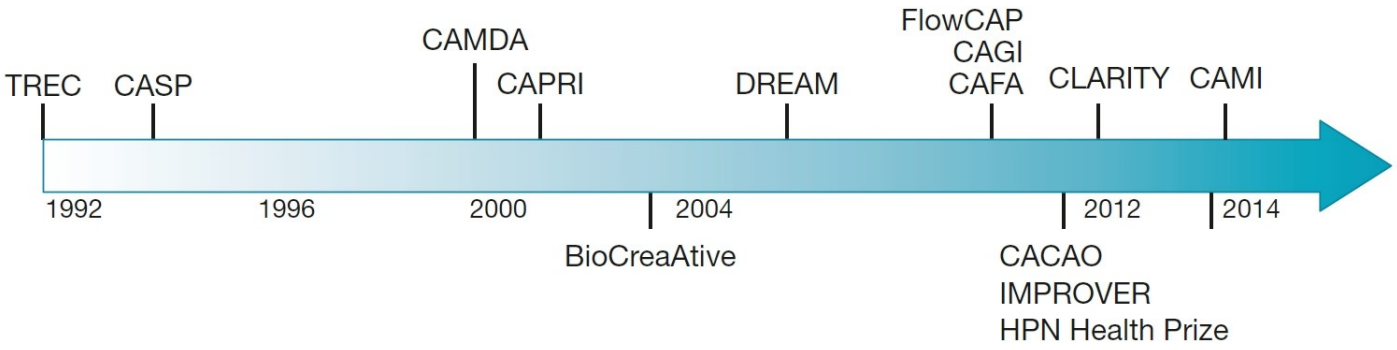


Figura 16.7
Lista delle maggiori competizioni scientifiche con il loro nome, acronimo, URL e anno di inizio.