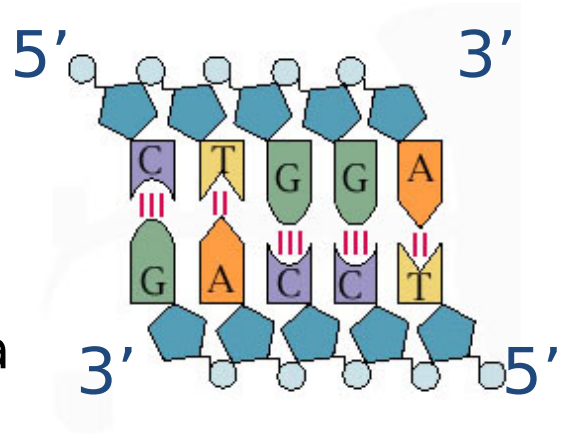
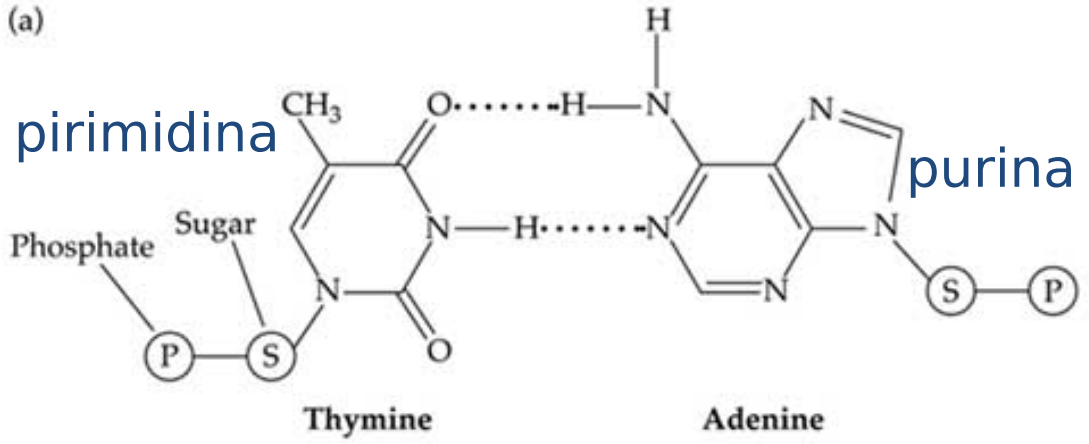
Lezione 1: Le molecole alla base di un genoma

Le molecole dell'ereditarietà

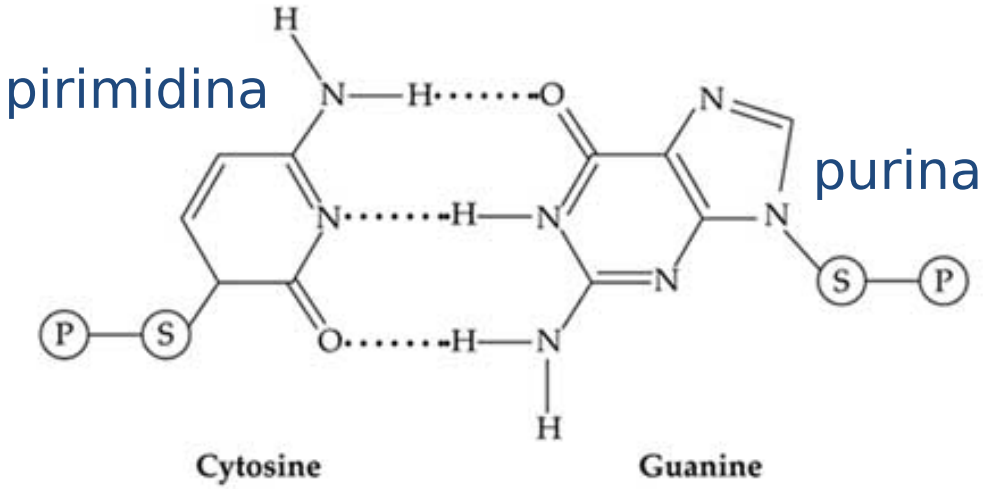


L’informazione ereditaria di tutti gli organismi viventi, con l’eccezione di alcuni virus, è a carico della molecola dell’**acido desossiribonucleico** (**DNA**).

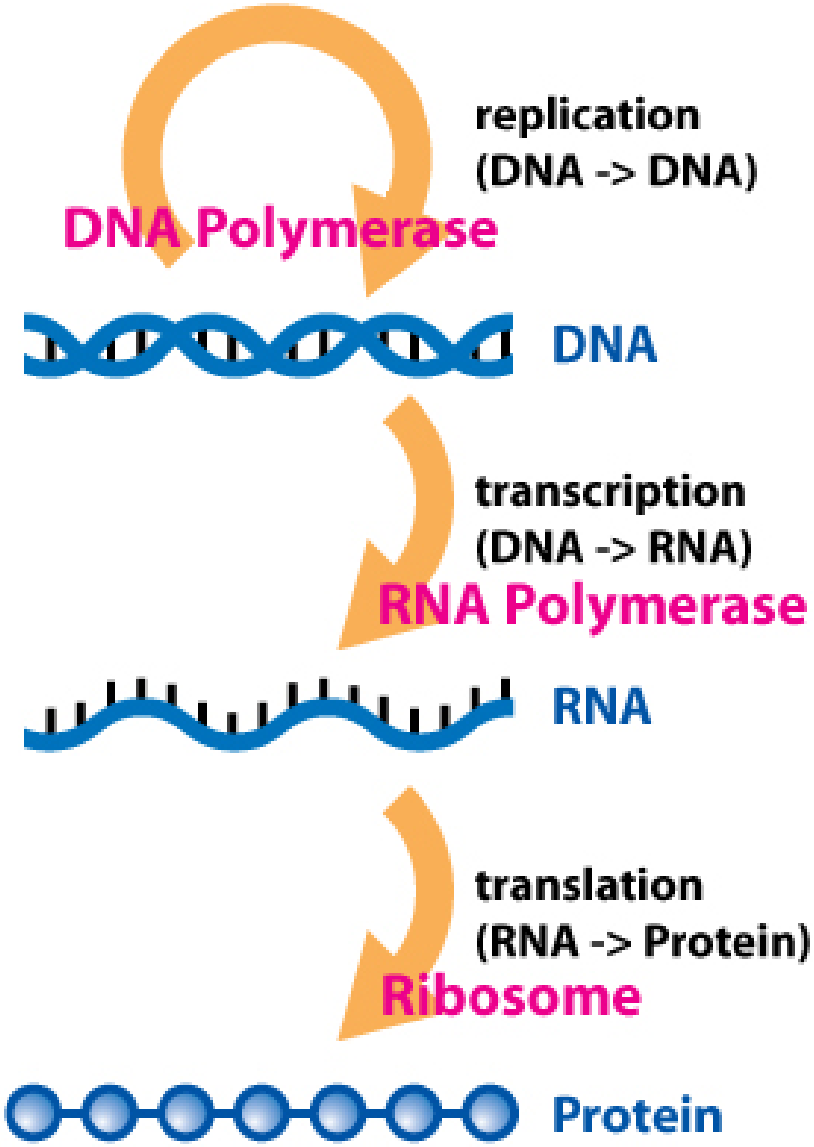
Legame debole



Legame forte



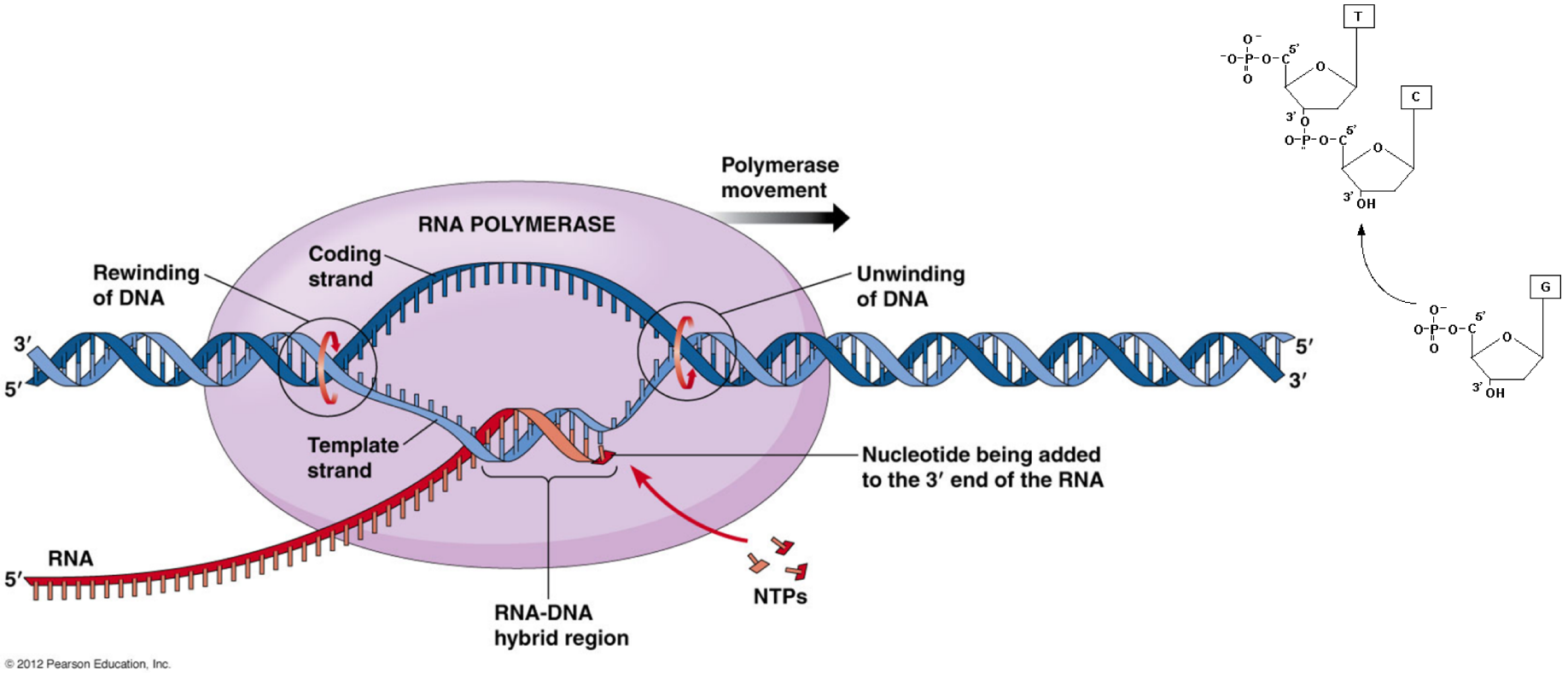
Il dogma centrale della biologia molecolare: il flusso dell'informazione



Wikipedia

Il dogma centrale della biologia molecolare: il flusso dell'informazione

La trascrizione:
il DNA antisenso (strand -) 3' -5' viene trascritto in un RNA 5'-3' (copia esatta del filamento "senso" cioè di quello codificante)



Codice genetico

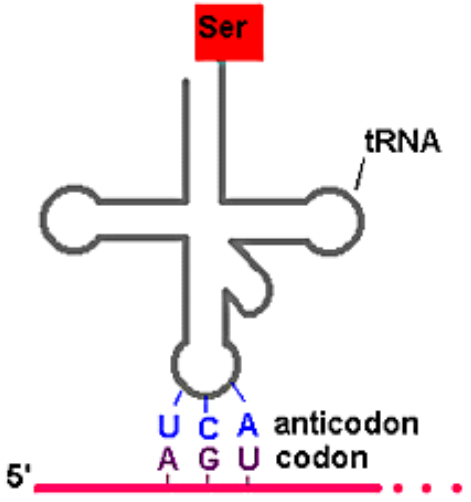
La trascrizione

```
5'  T G A C C T T C G A A C G G G A T G G A A A G G  3'  DNA strand+
3'  A C T G G A A G C T T G C C C T A C C T T T C C  5'  DNA strand-

5'  U G A C C U U C G A A C G G G A U G G A A A G G  3'  mRNA
```

DNA strand +: la stessa tripletta dell'mRNA con T al posto di U

La traduzione



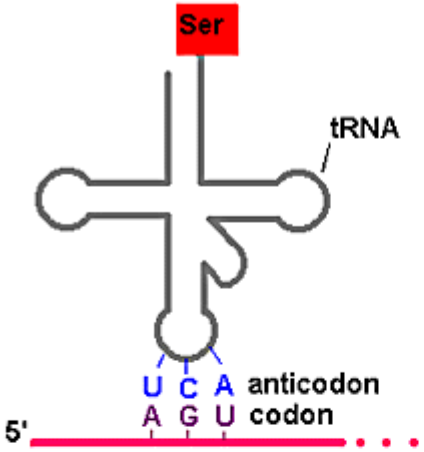
Perchè triplette?

4 basi disponibili, 20 AA da codificare.

Scopriamo quante lettere mettere in un codone (n)
Combinazioni possibili: 4^n
$4^1 = 4$
$4^2 = 16$ ancora troppo piccolo
$4^3 =$ **64** prima potenza di 4 più grande del numero di AA

Codice genetico

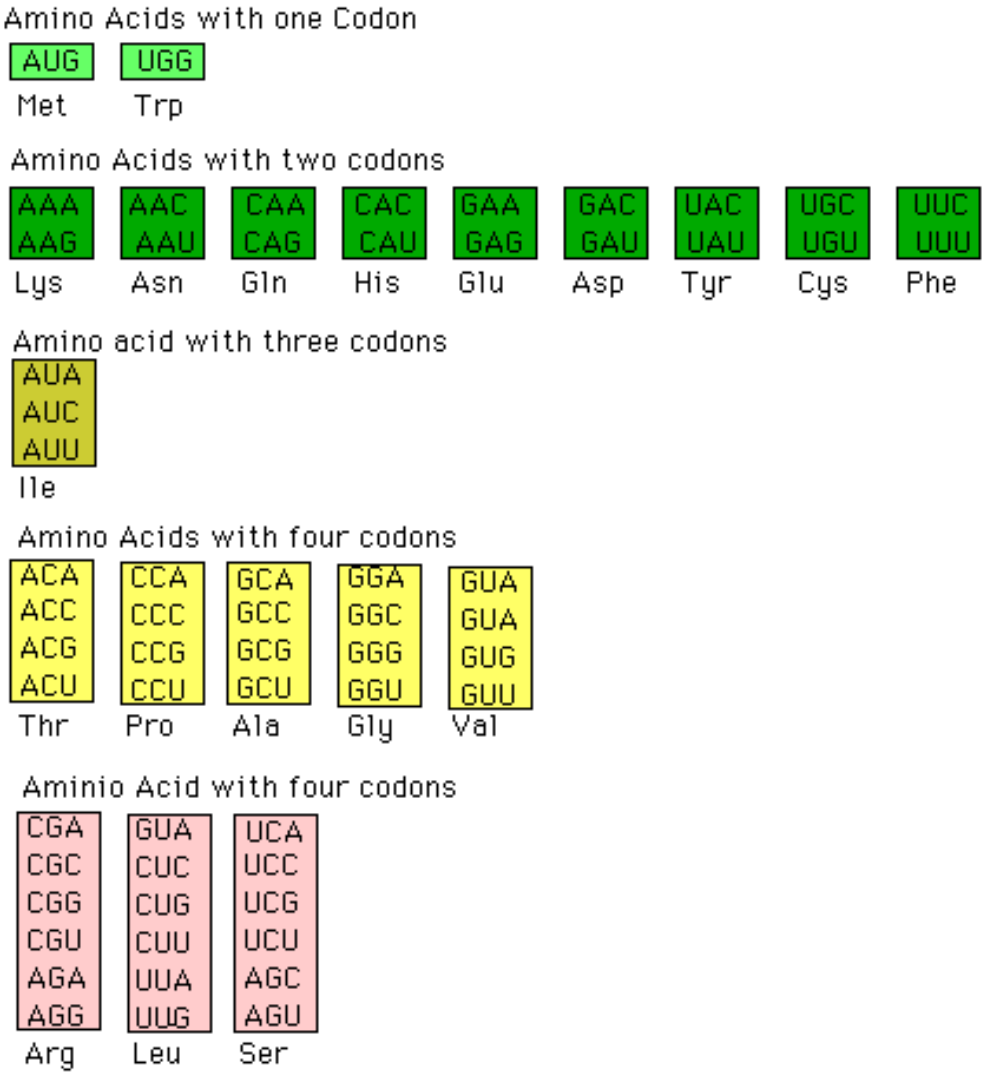
Il codice genetico universale è ridondante



1st letter	Second Letter: U		Second Letter: C		Second Letter: A		Second Letter: G		3rd letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Codice genetico

Il codice genetico organizzato secondo un criterio di degenerazione



Perchè non 20 triplette codificanti e 44 stop codon?

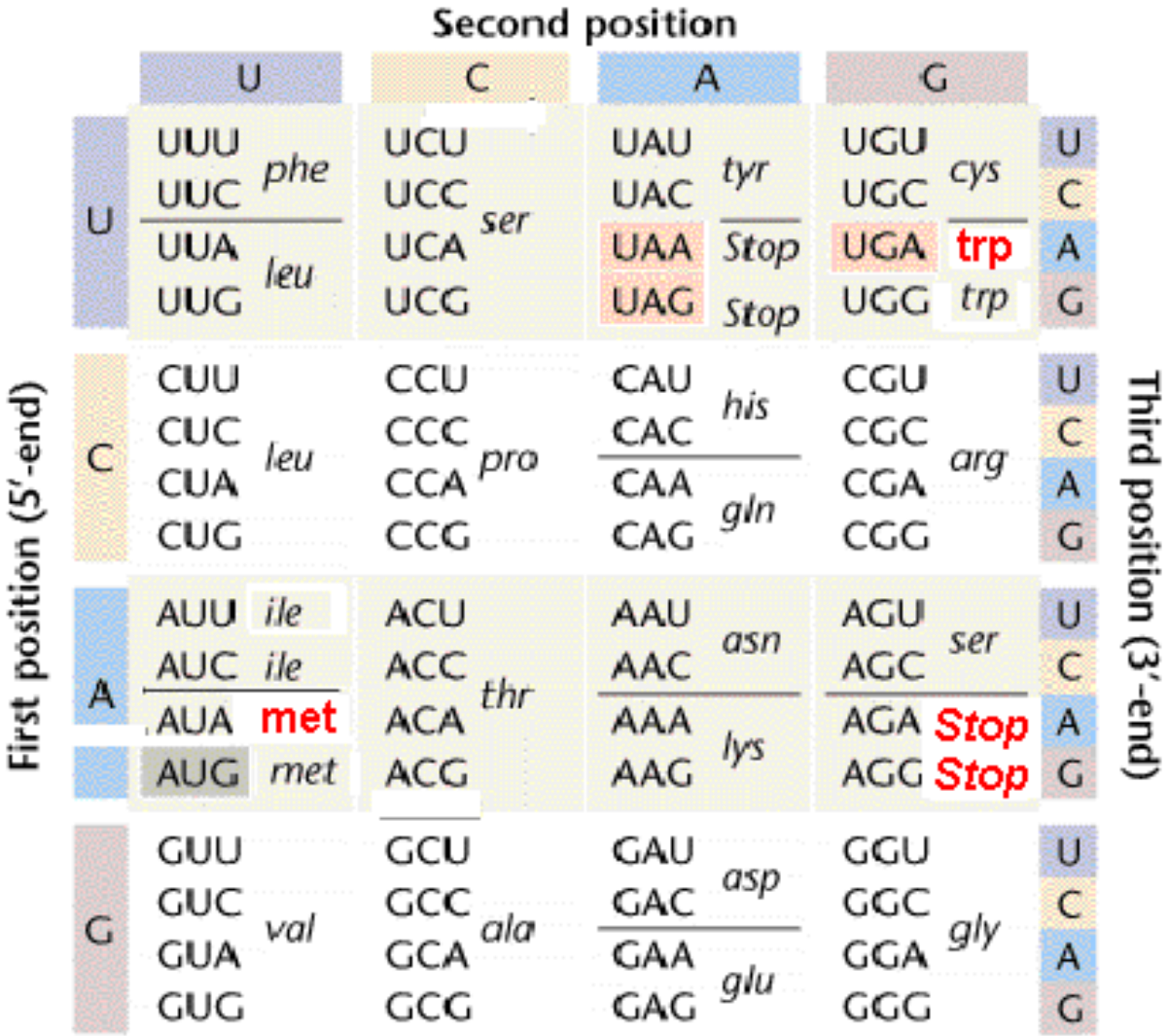
Alta probabilità che una mutazione produca uno stop codon (pericoloso!)

Perchè alcuni aminoacidi sono codificati da pochi codoni e altri da molti?

Ad esempio, il numero di codoni che codificano un particolare aminoacido correla con la sua frequenza nelle proteine (“importanza” dell’AA, necessità di assicurarne la sintesi)

Codice genetico

Il codice genetico dei mitocondri dei vertebrati



The Genetic Codes

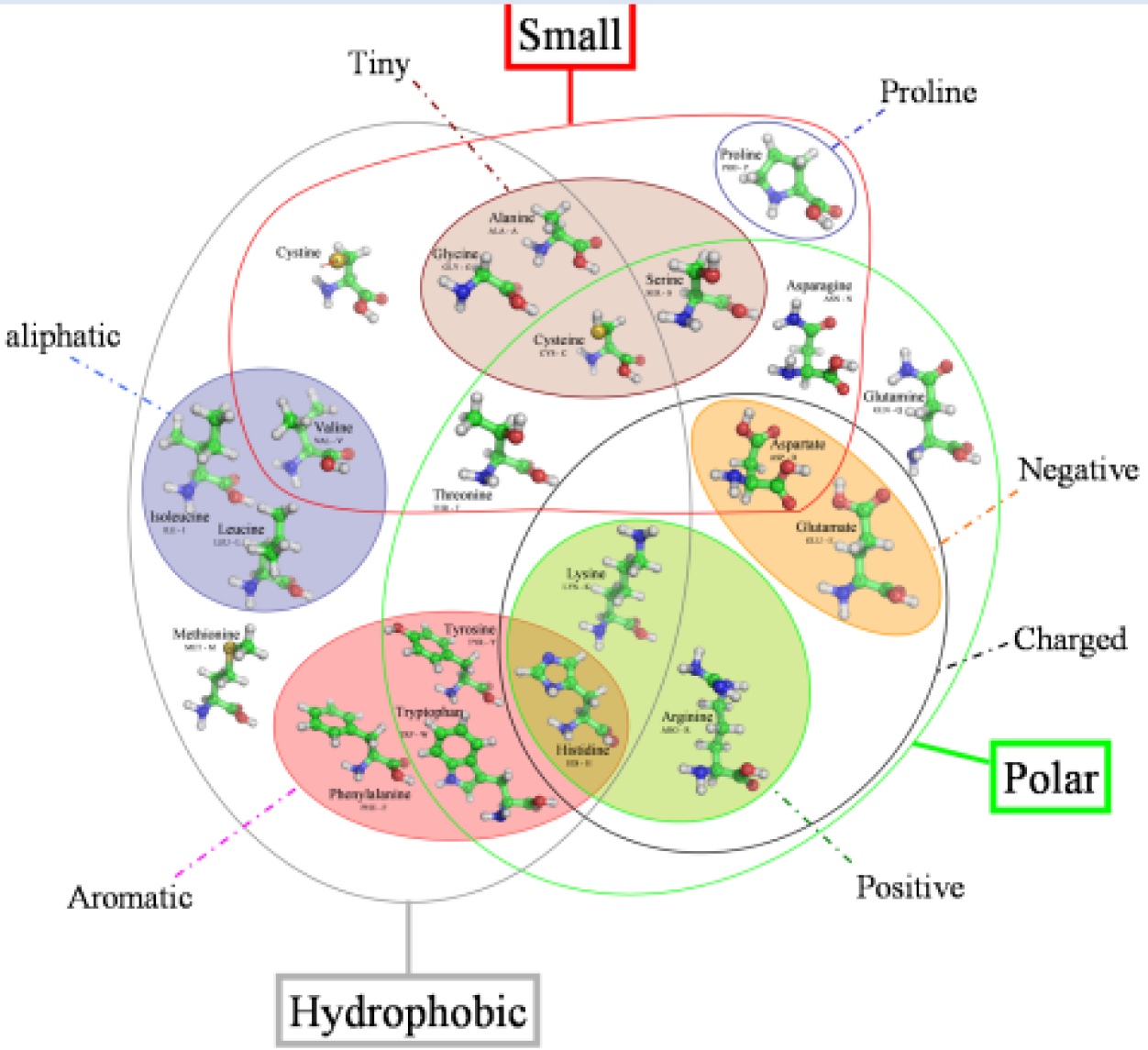
Compiled by Andrzej (Anjay) Elzanowski and Jim Ostell
National Center for Biotechnology Information (NCBI), Bethesda, Maryland, U.S.A.

Last update of the Genetic Codes: April 30, 2013

The following genetic codes are described here:

http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

La traduzione: aminoacidi



In bioinformatica spesso si deve valutare il “peso” del cambiamento AA in una proteina o nel confronto tra due proteine per poi poter proseguire con il resto delle analisi > **matrici di sostituzione**

Twenty-One Amino Acids

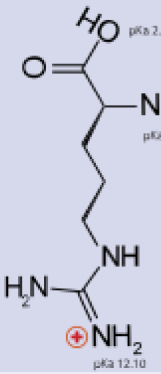
⊕ Positive ⊖ Negative

• Side chain charge at physiological pH 7.4

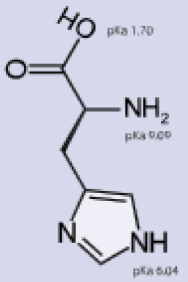
A. Amino Acids with Electrically Charged Side Chains

Positive

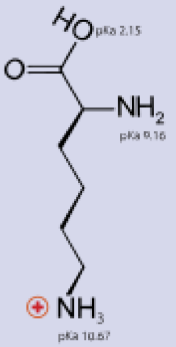
Arginine (Arg) **R**



Histidine (His) **H**

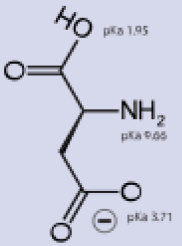


Lysine (Lys) **K**

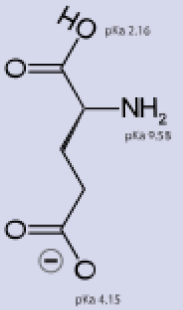


Negative

Aspartic Acid (Asp) **D**

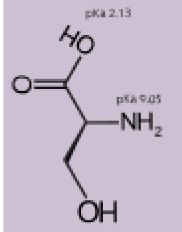


Glutamic Acid (Glu) **E**

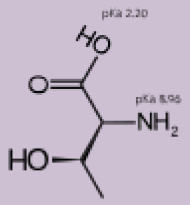


B. Amino Acids with Polar Uncharged Side Chains

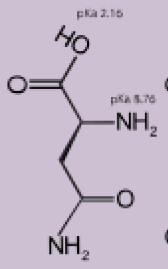
Serine (Ser) **S**



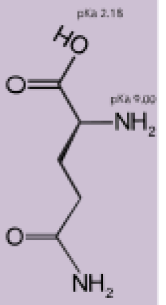
Threonine (Thr) **T**



Asparagine (Asn) **N**

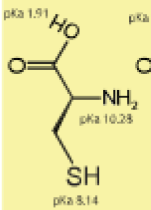


Glutamine (Gln) **Q**

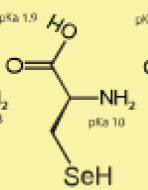


C. Special Cases

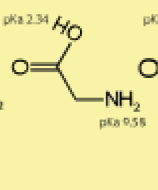
Cysteine (Cys) **C**



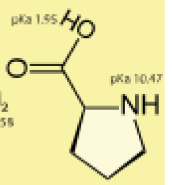
Selenocysteine (Sec) **U**



Glycine (Gly) **G**

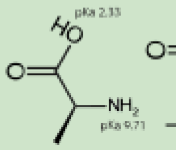


Proline (Pro) **P**

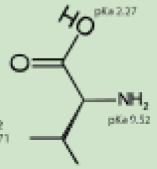


D. Amino Acids with Hydrophobic Side Chain

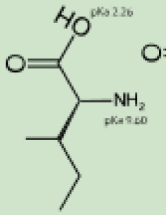
Alanine (Ala) **A**



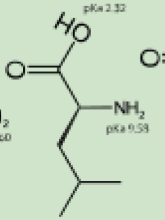
Valine (Val) **V**



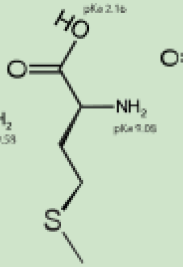
Isoleucine (Ile) **I**



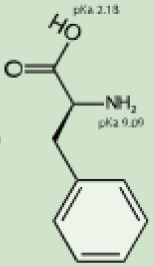
Leucine (Leu) **L**



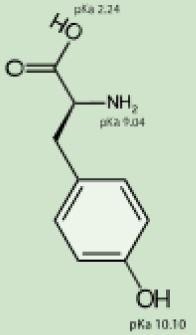
Methionine (Met) **M**



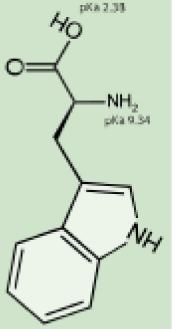
Phenylalanine (Phe) **F**



Tyrosine (Tyr) **Y**



Tryptophan (Trp) **W**



Matrici di sostituzione

	A	C	G	T
A	5	-4	-4	-4
C	-4	5	-4	-4
G	-4	-4	5	-4
T	-4	-4	-4	5

Es. Matrice di sostituzione nucleotidica
4 nucleotidi > 6 possibili sostituzioni

Nelle sequenze proteiche ci sono 20 aminoacidi con determinate dimensioni, cariche, codone di codifica, caratteristiche chimiche.

Matrici di sostituzioni AA hanno un punteggio per ognuna delle 210 possibili coppie di AA (180 = ((20*20)/2) – 20))

Queste matrici vengono calcolate dando un punteggio alla relazione tra due AA sulla base di alcune precise caratteristiche

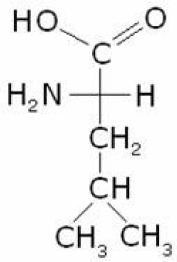
Sostituzioni aminoacidiche > matrici

Amino acids substitution matrix
Residue-features matrices

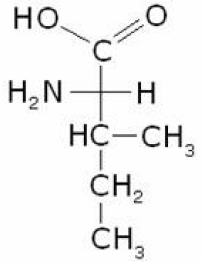
A selection of substitution matrices based on amino acids features-

- Mutation values for the interconversion of amino acid pairs (Fitch, 1966)
- Genetic code matrix (Benner et al., 1994)
- Residue replace ability matrix (Cserzo et al., 1994)
- Structure-Genetic matrix (Feng et al., 1985)
- Hydrophobicity scoring matrix (George et al., 1990)
- Chemical distance (Grantham, 1974)
- Chemical similarity scores (McLachlan, 1972)
- Base-substitution-protein-stability matrix (Miyazawa-Jernigan, 1993)
- Hydrophobicity scoring matrix (Riek et al., 1995)
- WAC matrix constructed from amino acid comparative profiles (Wei et al., 1997)

Sostituzioni aminoacidiche



leucina

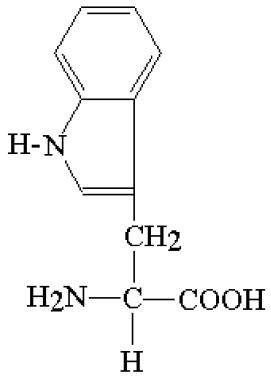


isoleucina

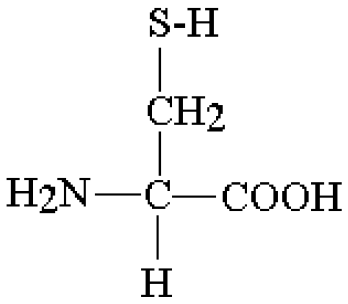
TABLE 4.7 Physicochemical distances between pairs of amino acids[a]

Arg	Leu	Pro	Thr	Ala	Val	Gly	Ile	Phe	Tyr	Cys	His	Gln	Asn	Lys	Asp	Glu	Met	Trp	
110	145	74	58	99	124	56	142	155	144	112	89	68	46	121	65	80	135	177	Ser
	102	103	71	112	96	125	97	97	77	180	29	43	86	26	96	54	91	101	Arg
		98	92	96	32	138	5	22	36	198	99	113	153	107	172	138	15	61	Leu
			38	27	68	42	95	114	110	169	77	76	91	103	108	93	87	147	Pro
				58	69	59	89	103	92	149	47	42	65	78	85	63	81	128	Thr
					64	60	94	113	112	195	86	91	111	106	126	107	84	148	Ala
						109	29	50	55	192	84	96	133	97	152	121	21	88	Val
							135	153	147	159	98	87	80	127	94	98	127	184	Gly
								21	33	198	94	109	149	102	168	134	10	61	Ile
									22	205	100	116	158	102	177	140	28	40	Phe
										194	83	99	143	85	160	122	36	37	Tyr
											174	154	139	202	154	170	196	**215**	Cys
												24	68	32	81	40	87	115	His
													46	53	61	29	101	130	Gln
														94	23	42	142	174	Asn
															101	56	95	110	Lys
																45	160	181	Asp
																	126	152	Glu
																		67	Met

n (1974).



tryptophan
is a nonpolar amino acid



cysteine

Cysteine is a uncharged polar amino acid.

FASTA format

In bioinformatics, FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes.

Amino Acid Code	Meaning
A	Alanine
B	Aspartic acid (D) or Asparagine (N)
C	Cysteine
D	Aspartic acid
E	Glutamic acid
F	Phenylalanine
G	Glycine
H	Histidine
I	Isoleucine
J	Leucine (L) or Isoleucine (I)
K	Lysine
L	Leucine
M	Methionine
N	Asparagine
O	Pyrrolysine
P	Proline
Q	Glutamine
R	Arginine
S	Serine
T	Threonine
U	Selenocysteine
V	Valine
W	Tryptophan
Y	Tyrosine
Z	Glutamic acid (E) or Glutamine (Q)
X	any
*	translation stop
-	gap of indeterminate length

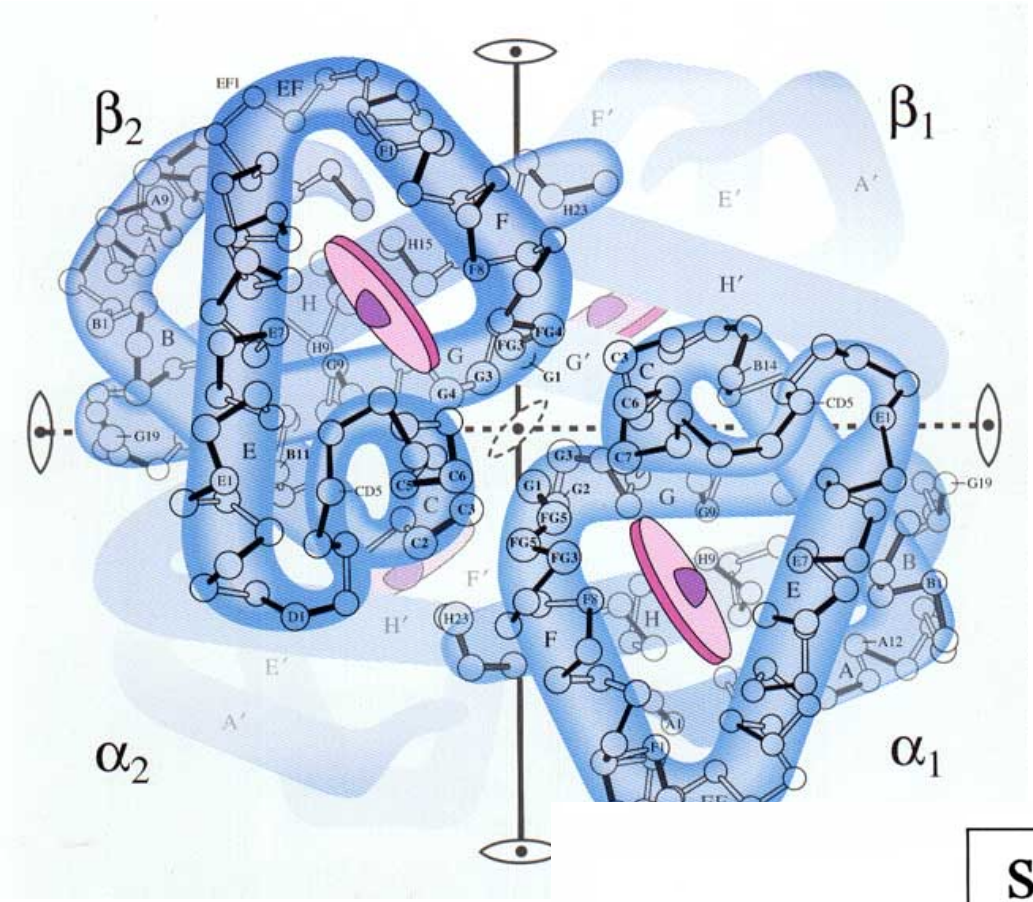
Amino acids substitution matrix
Residue-features matrices

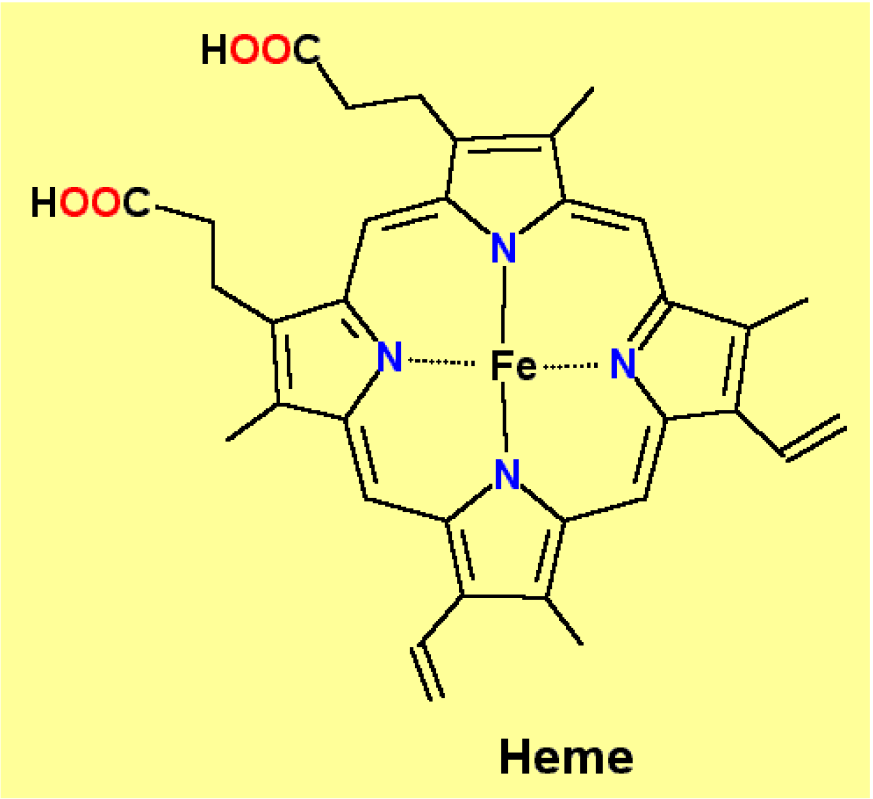
Hydrophobicity aa substitution matrix

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

Il significato evolutivo dei cambiamenti aminoacidici

Proteine





Tasso di sostituzione per sito per 10^9 anni

	surface	heme pocket
α-globin	1.35	0.17
β-globin	2.73	0.24

Mutazioni

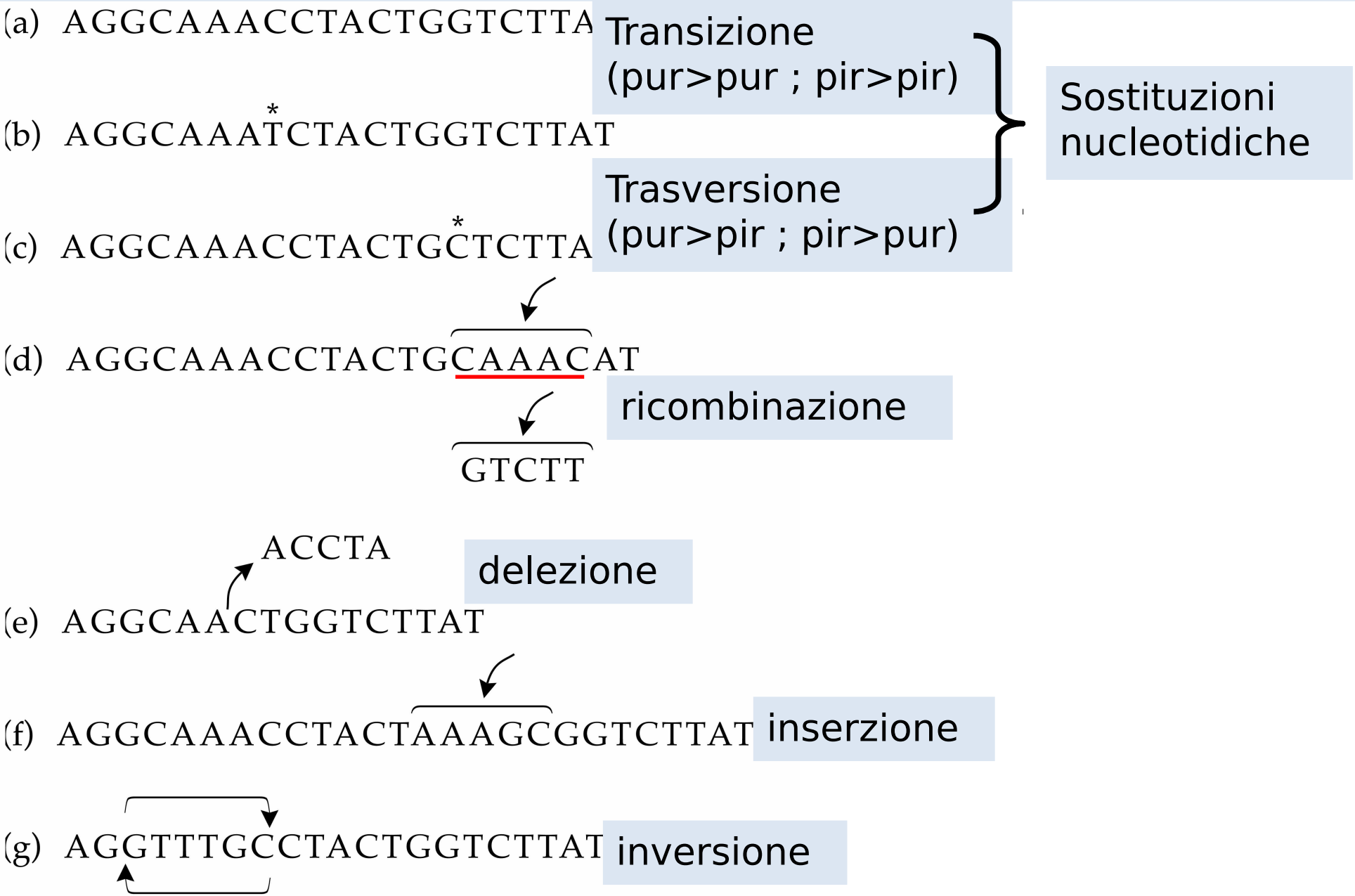
Le sequenze di DNA sono solitamente copiate in modo preciso durante la replicazione.

Raramente tuttavia possono avvenire degli errori che originano nuove sequenze. Questi errori si chiamano **mutazioni**.

Da un punto di vista evolutivo una mutazione è una sequenza nella linea germinale che differisce dalla sua controparte nelle cellule somatiche, che viene ereditata dalla progenie la quale sarà dunque caratterizzata da una **“novità” genetica**.

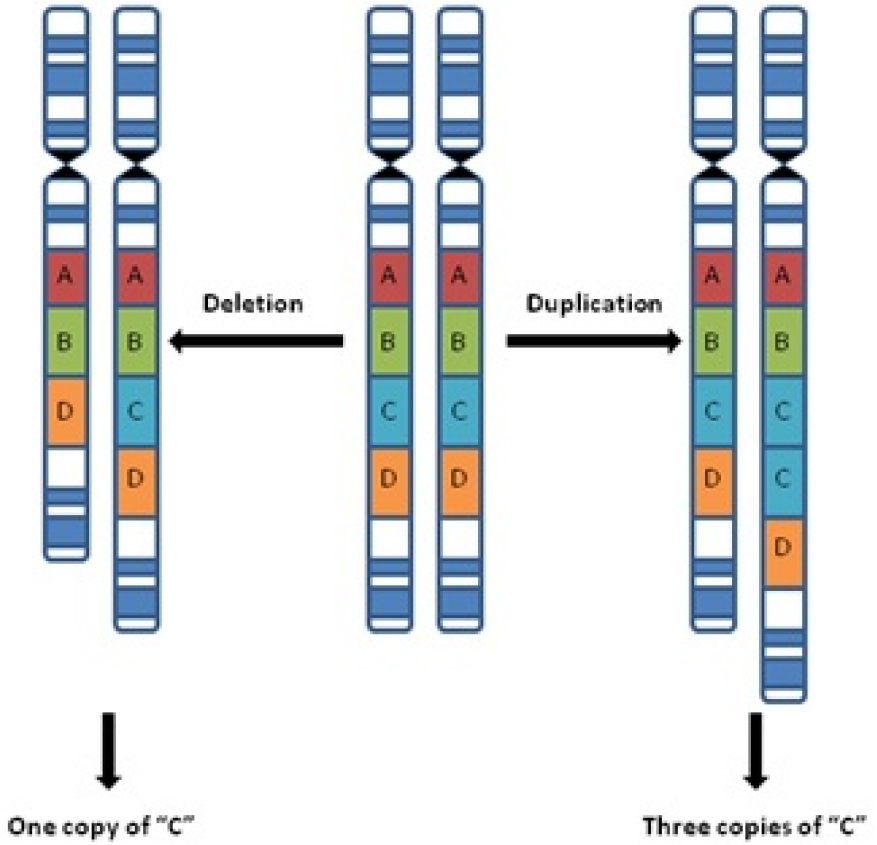
Le mutazioni sono quindi la **fonte di variabilità e di novità evolutiva**

Mutazioni nucleotidiche

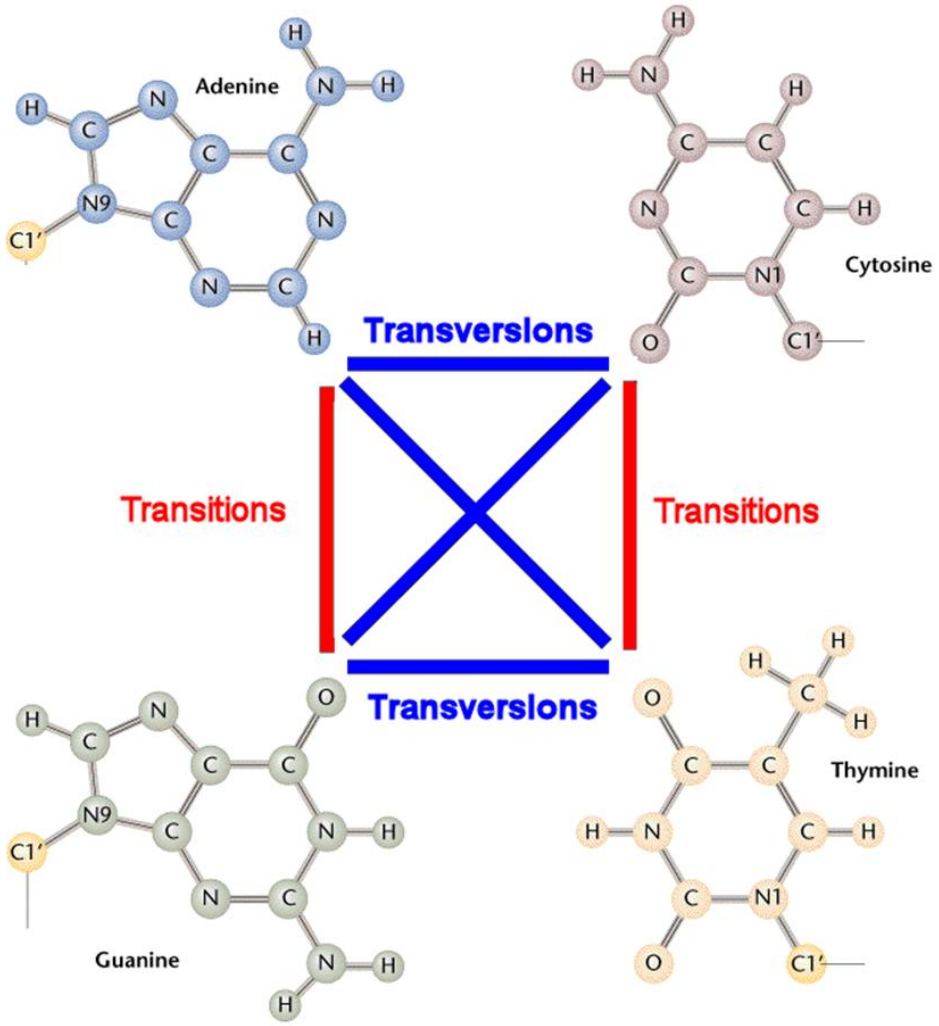


Mutazioni nucleotidiche

Copy number variant



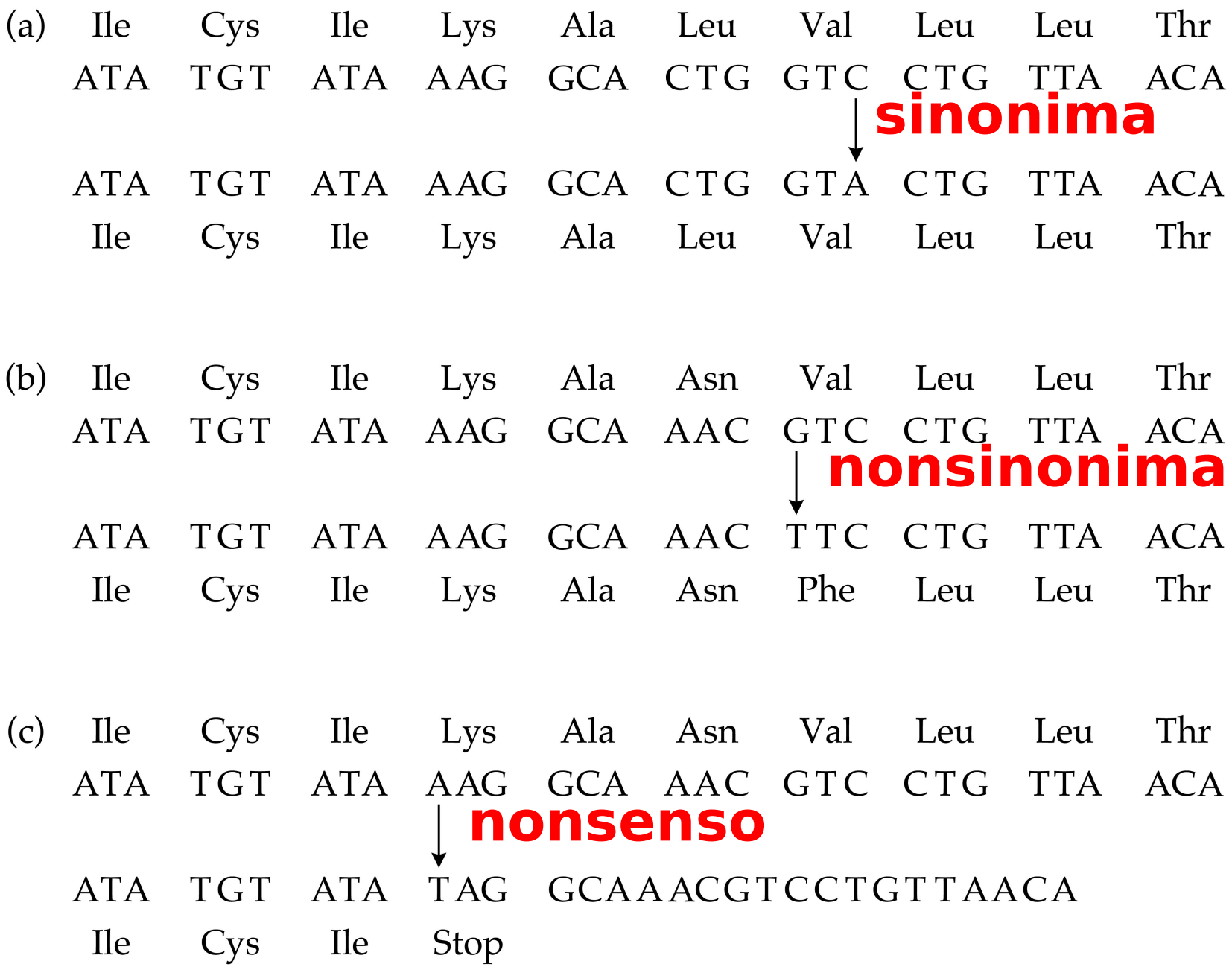
Mutazioni nucleotidiche



Nucleotide substitutions

Transitions:	A→G
	G→A
	C→T
	T→C
Transversions:	A→C
	A→T
	C→A
	C→G
	T→A
	T→G
	G→C
	G→T

Mutazioni nucleotidiche: effetto sulla traduzione



Mutazioni nucleotidiche: effetto sulla traduzione

Ogni codone codificante un AA può mutare in altri 9 codoni attraverso sostituzioni di un singolo nucleotide.

Esempio:

CCU (Pro)

6 possibili sostituzioni nonsinonime
UCU (Ser)
ACU (Thr)
GCU (Ala)
C**U**U (Leu)
C**A**U (His)
C**G**U (Arg)

3 possibili sostituzioni sinonime
CC**C**
CC**A**
CC**G**

Ogni codone codificante un AA può mutare in altri 9 codoni attraverso sostituzioni di un singolo nucleotide.

61 codoni
"senso"
↓
61x9 =549
possibili
sostituzioni
nucleotidiche

1st letter	Second Letter U	Second Letter C	Second Letter A	Second Letter G	3rd letter
U	UUU Phe UUC UUA Leu UUG	UCU UCC UCA UCG Ser	UAU Tyr UAC UAA Stop UAG Stop	UGU Cys UGC UGA Stop UGG Trp	U C A G
C	CUU CUC CUA CUG Leu	CCU CCC CCA CCG Pro	CAU His CAC CAA Gln CAG	CGU CGC CGA CGG Arg	U C A G
A	AUU AUC AUA Ile AUG Met	ACU ACC ACA ACG Thr	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G
G	GUU GUC GUA GUG Val	GCU GCC GCA GCG Ala	GAU Asp GAC GAA Glu GAG	GGU GGC GGA GGG Gly	U C A G

Se assumiamo che

1. Tutti i codoni siano ugualmente presenti nelle regioni codificanti
2. Ogni sito abbia la stessa probabilità di mutare

In un gene codificante qualunque ci aspettiamo una frequenza relativa dei diversi tipi di sostituzioni come in tabella

Alcune caratteristiche importanti:

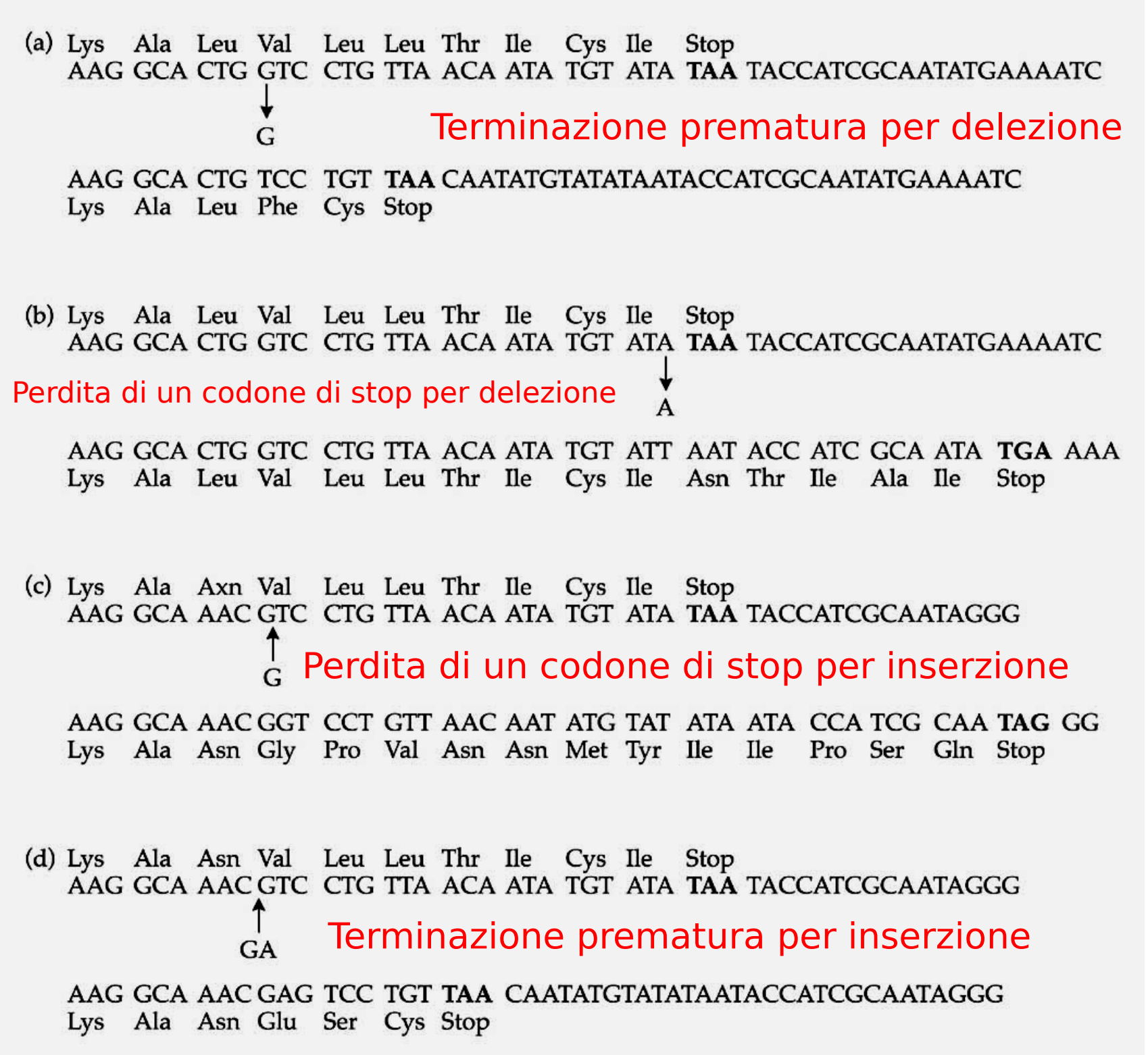
- Circa il 70% dei cambiamenti in 3° base sono sinonimi
- Il 100% dei cambiamenti in 2° base sono nonsinonimi
- Il 96% dei cambiamenti in 1° base sono nonsinonimi

sostituzioni	numero	Frequenza
Totali (1,2,3 base)	549	100
Sinonime	134	25
Nonsinonime	415	75
Missenso (non senso)	392 (23)	71 (4)
Totali (1 base)	183	100
Sinonime	8	4
Nonsinonime	175	96
Missenso (non senso)	166 (9)	91 (5)
Totali (2 base)	183	100
Sinonime	0	0
Nonsinonime	183	100
Missenso (non senso)	176 (7)	96 (4)
Totali (3 base)	183	100
Sinonime	126	69
Nonsinonime	57	31
Missenso (non senso)	50 (7)	27 (4)

Inserzioni e delezioni

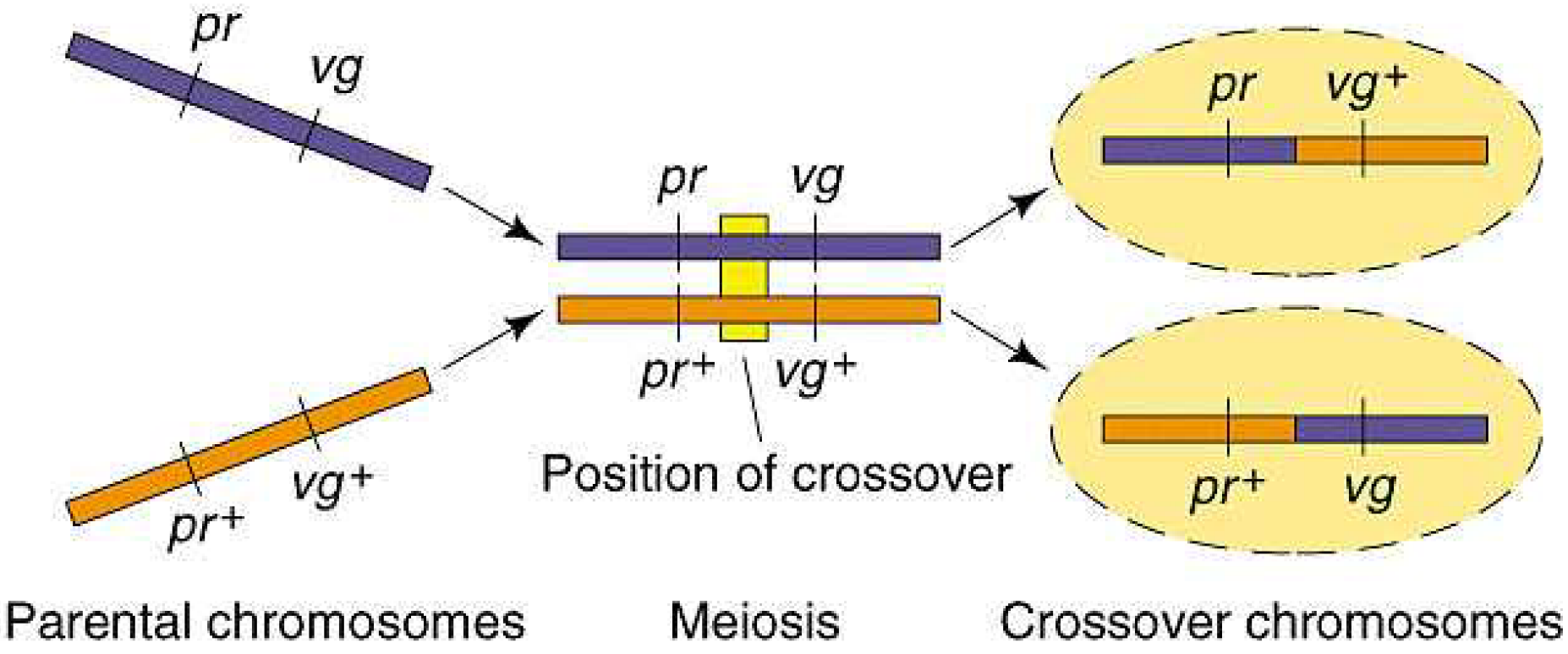
Nel confronto tra due sequenze è impossibile capire se ci sia stata una delezione in una delle due o una inserzione nell'altra

INserzioni de **DEL**ezioni vengono in generale chiamate **INDELS**


(a) Lys Ala Leu Val Leu Leu Thr Ile Cys Ile Stop
AAG GCA CTG GTC CTG TTA ACA ATA TGT ATA TAA TACCATCGCAATATGAAAATC
G
Terminazione prematura per delezione
AAG GCA CTG TCC TGT TAA CAATATGTATATAATACCATCGCAATATGAAAATC
Lys Ala Leu Phe Cys Stop
(b) Lys Ala Leu Val Leu Leu Thr Ile Cys Ile Stop
AAG GCA CTG GTC CTG TTA ACA ATA TGT ATA TAA TACCATCGCAATATGAAAATC
Perdita di un codone di stop per delezione
A
AAG GCA CTG GTC CTG TTA ACA ATA TGT ATT AAT ACC ATC GCA ATA TGA AAA
Lys Ala Leu Val Leu Leu Thr Ile Cys Ile Asn Thr Ile Ala Ile Stop
(c) Lys Ala Axn Val Leu Leu Thr Ile Cys Ile Stop
AAG GCA AAC GTC CTG TTA ACA ATA TGT ATA TAA TACCATCGCAATAGGG
G
Perdita di un codone di stop per inserzione
AAG GCA AAC GGT CCT GTT AAC AAT ATG TAT ATA ATA CCA TCG CAA TAG GG
Lys Ala Asn Gly Pro Val Asn Asn Met Tyr Ile Ile Pro Ser Gln Stop
(d) Lys Ala Asn Val Leu Leu Thr Ile Cys Ile Stop
AAG GCA AAC GTC CTG TTA ACA ATA TGT ATA TAA TACCATCGCAATAGGG
GA
Terminazione prematura per inserzione
AAG GCA AAC GAG TCC TGT TAA CAATATGTATATAATACCATCGCAATAGGG
Lys Ala Asn Glu Ser Cys Stop

Frameshift

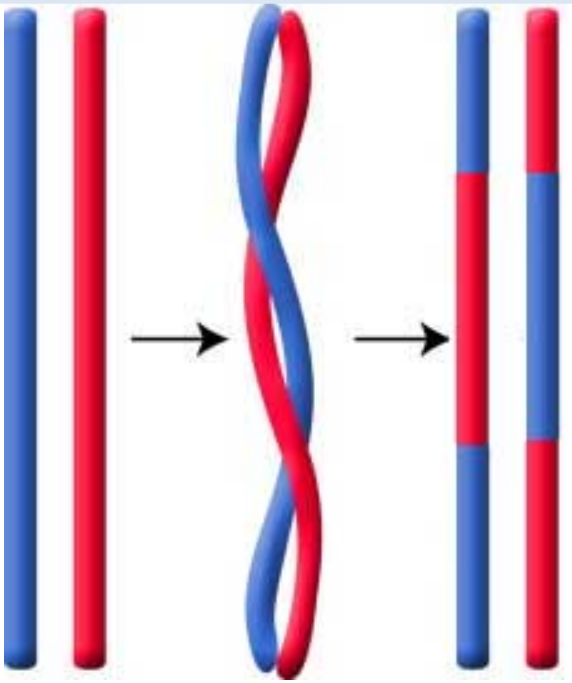
Altre fonti di variabilità: la ricombinazione

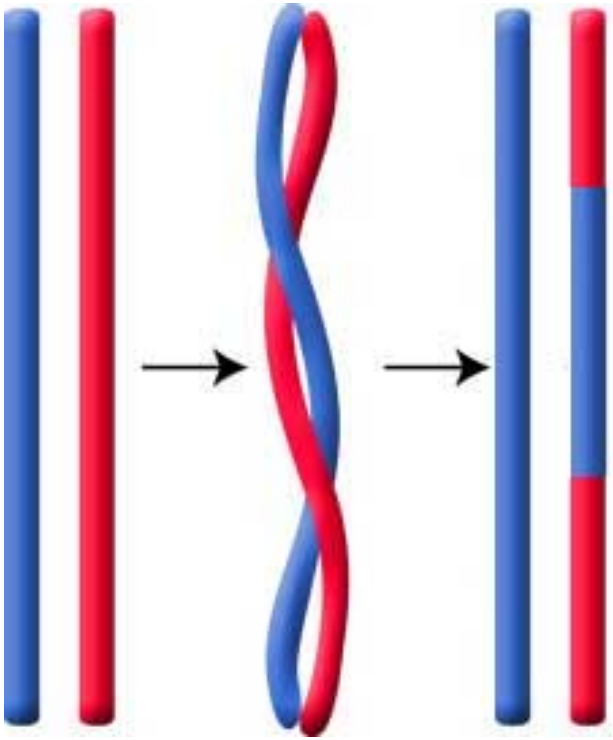


Altre fonti di variabilità: la ricombinazione

Recombination

1. Crossing-over or Reciprocal recombination.





2. Gene conversion or nonreciprocal recombination.

Altre fonti di variabilità: la ricombinazione

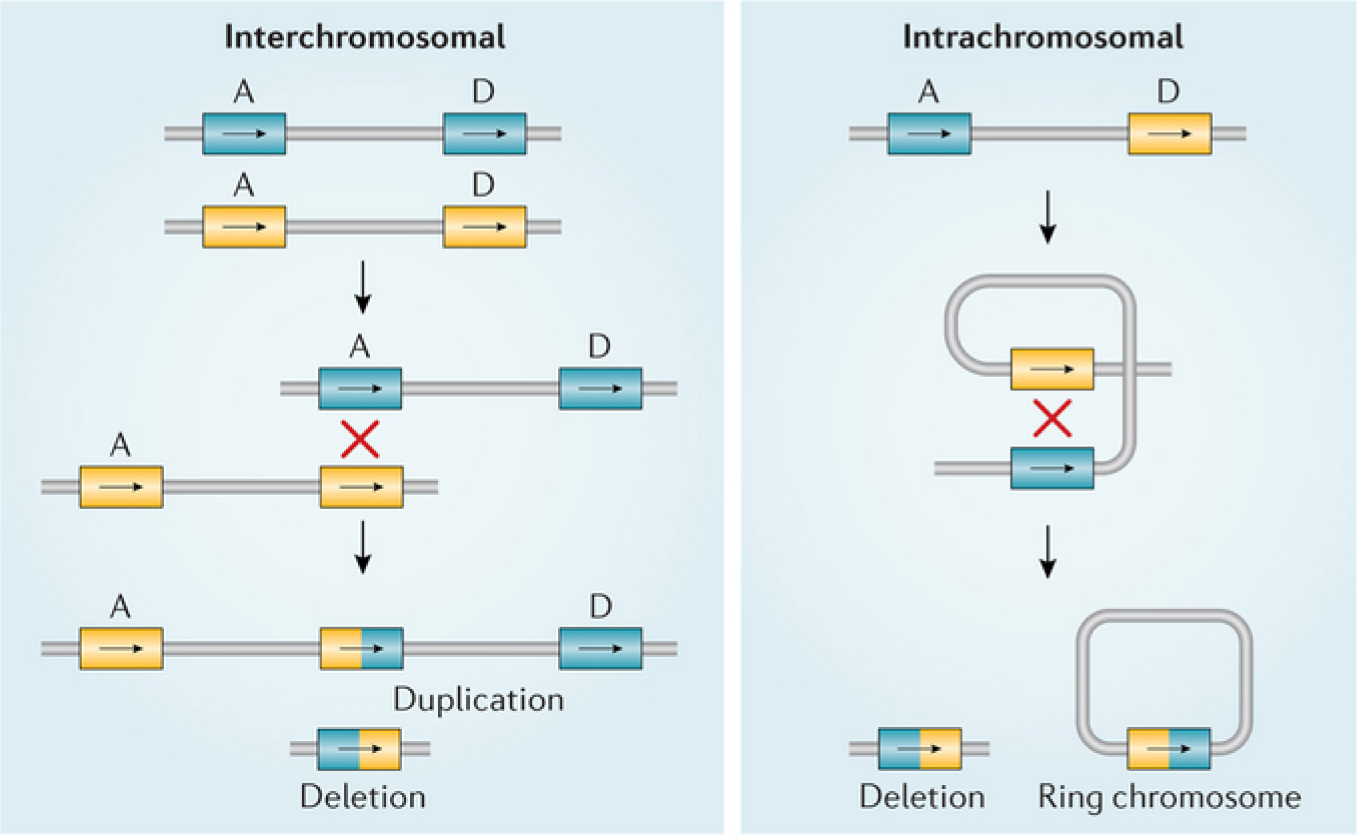
La ricombinazione reciproca è un potente mezzo di generazione della variabilità

5’—AACT—3’ and 5’—CTTG—3’ -> 6 possibili nuove sequenze**:**

5’—ATTG—3’
5’—CACT—3’
5’—AATG—3’
5’—CTCT—3’
5’—AACG—3’
5’—CTTT—3’

Altre fonti di variabilità: la ricombinazione

Inserzioni e delezioni



Crossing over ineguale

Delezione “intra strand”