Como genes codificam proteínas

QBQ102

Prof. João Carlos Setubal

Universidade de São Paulo
Instituto de Química

Como DNA permite…

- A atividade da vida?
- A reprodução da vida?
- Hoje vamos ver a parte da atividade da vida

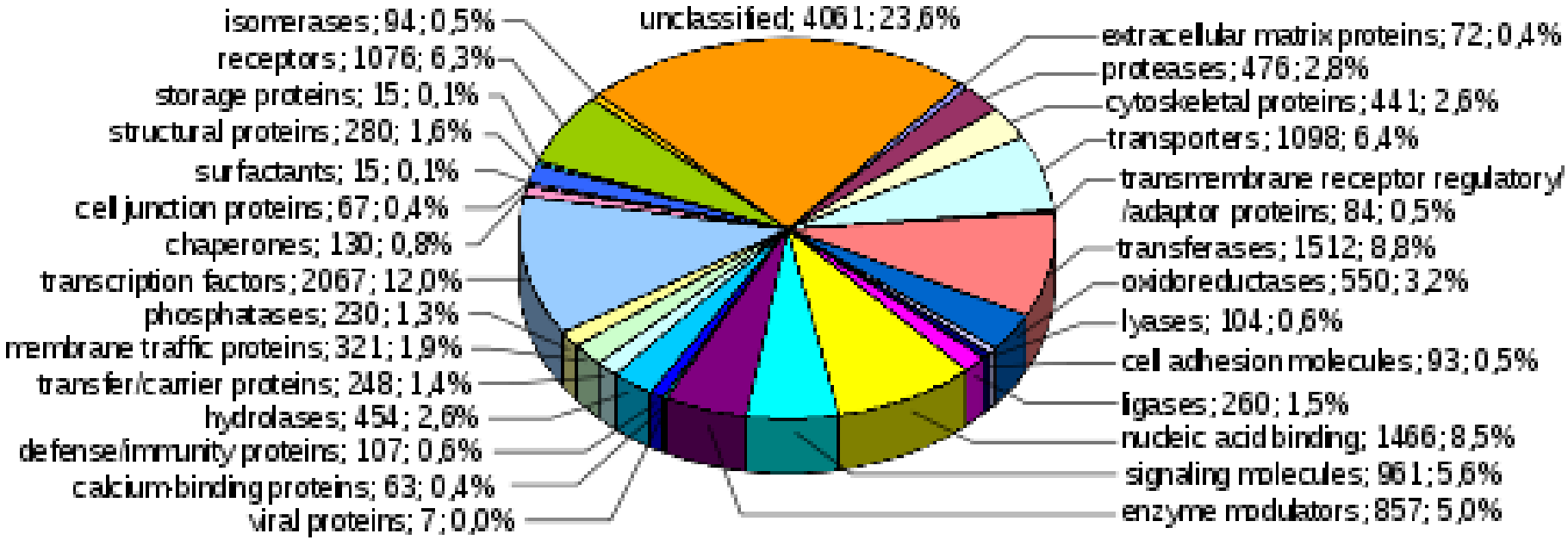
Atividade da vida significa...

- …basicamente…
- Montar (ou sintetizar) proteínas

Proteínas são as moléculas trabalhadoras dos organismos

- “nós somos nossas proteínas”
- Enzimas: catalizam reações químicas, essenciais para o metabolismo celular
- Funções estruturais ou mecânicas (actina e miosina em músculos)
- Sinalização celular
- Síntese de DNA
  - A DNA polimerase é uma proteína

Variedade das proteínas humanas



Mas o que tem DNA a ver com proteínas?

- É no DNA que está a "receita" para a fabricação das proteínas
- Hélice dupla + complementaridade

  ⇨Replicação
- A cadeia de nucleotídeos

  ⇨ Informação

    ⇨ "receita" para sintetizar ou montar proteínas

Como assim?

- Da mesma forma como letras podem formar palavras em português que nós entendemos:
  - ANTICONSTITUCIONALISSIMAMENTE

- Em “celulês” ou “genomês” é possível formar palavras que a maquinaria da célula é capaz de entender:
  - ATGCCGGTCGTCGCGGACGACGACGG

Mas como são só 4 letras...

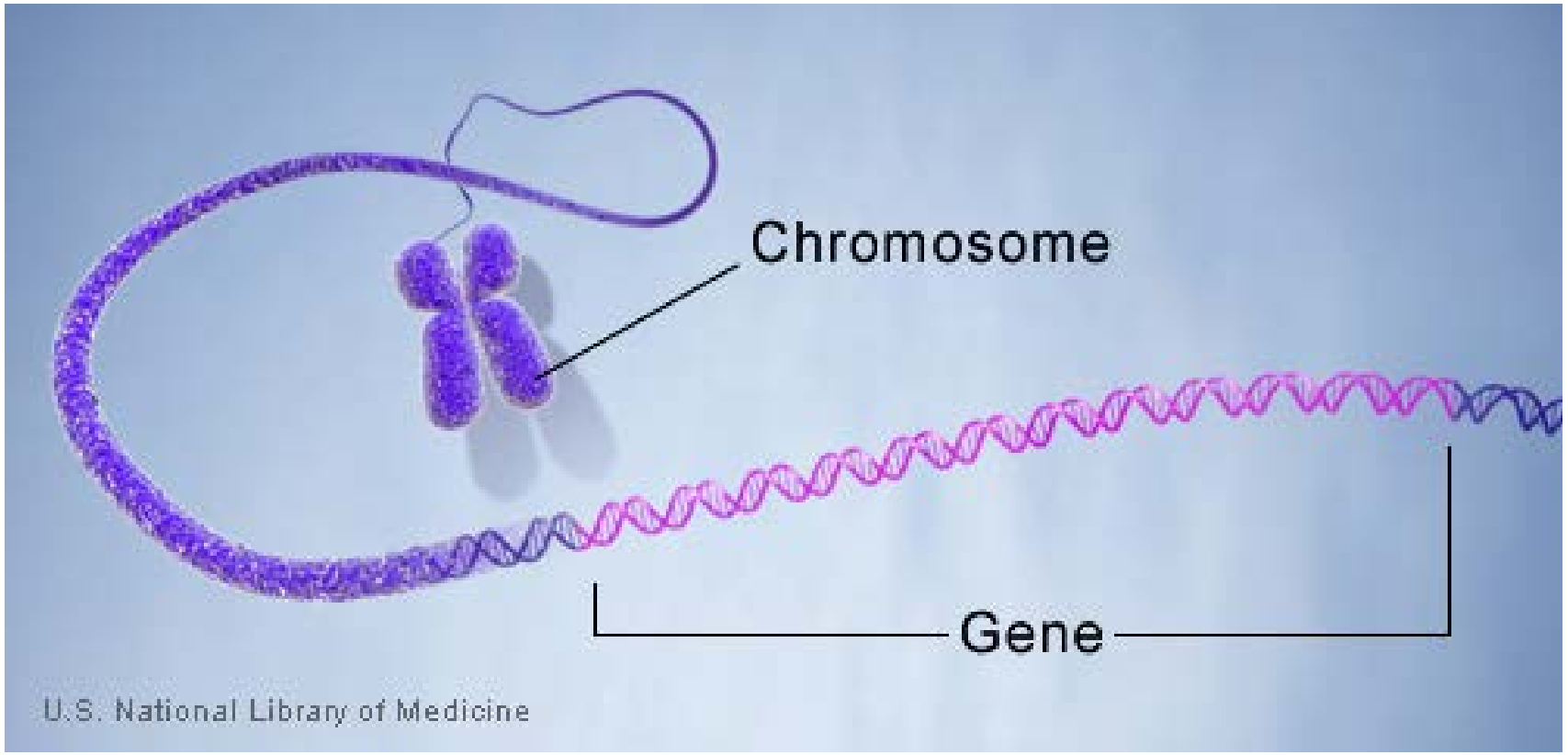
- Essas “palavras” são muito mais longas
- Tipicamente (numa bactéria) elas tem comprimento ~1000 pb
- Em outras palavras, um trecho de ~1000 pb é o tamanho da “receita” para montar uma proteína

Mas não é *qualquer* trecho de ~1000 pb…

- Assim como não é qualquer trecho de 10 letras que corresponde a uma palavra num texto em português
- Por exemplo:
  - quertre

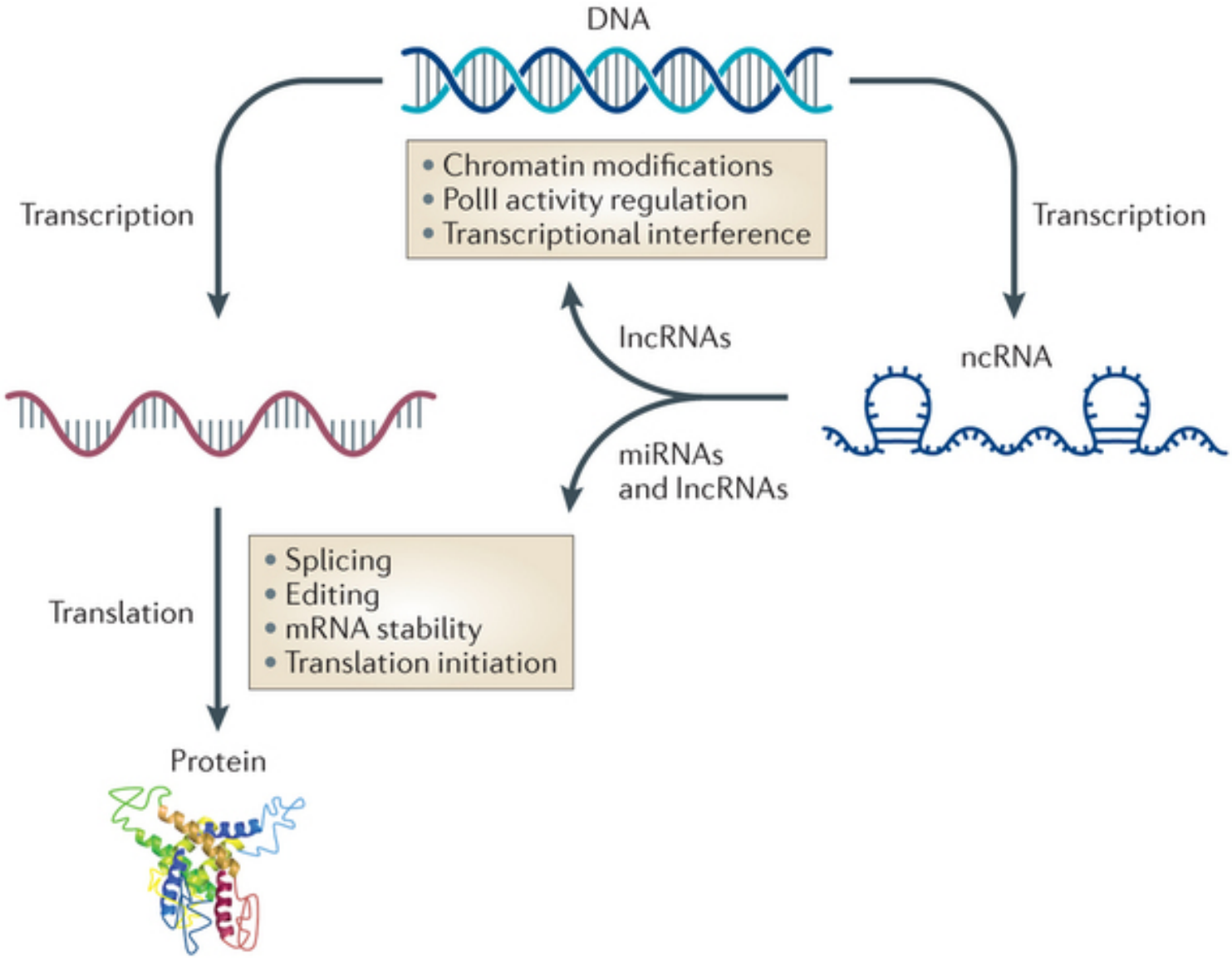
Como se chamam os trechos com as receitas?

- Ou seja, as “palavras” em “genomês”?
- Genes!

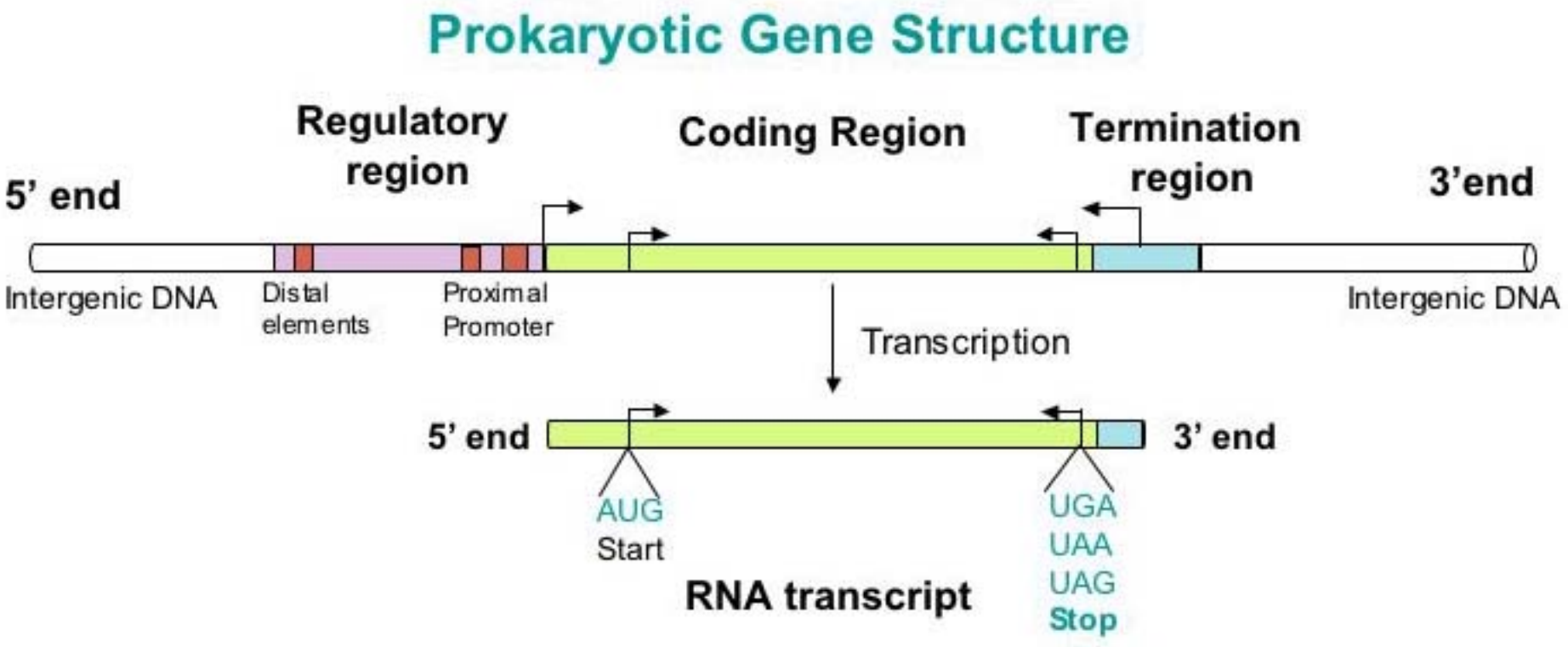

Chromosome
Gene
U.S. National Library of Medicine

Informações armazenadas num genoma

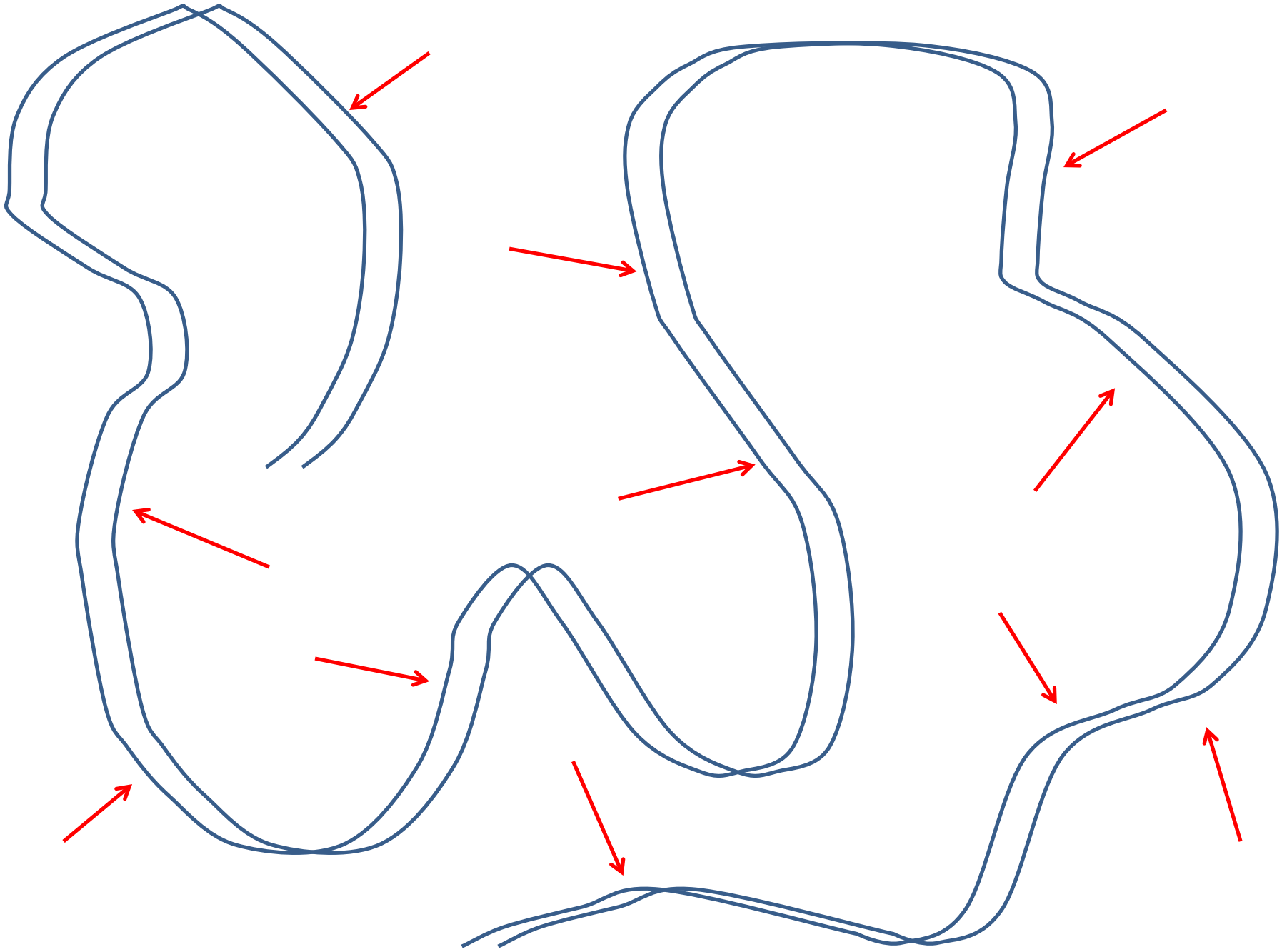
- Genes codificadores de proteína
- Genes de RNA
  - tRNA
  - RNA ribossomal
  - Outros pequenos RNAs, a maioria reguladores



http://www.nature.com/nrd/journal/v12/n6/full/nrd4018.html

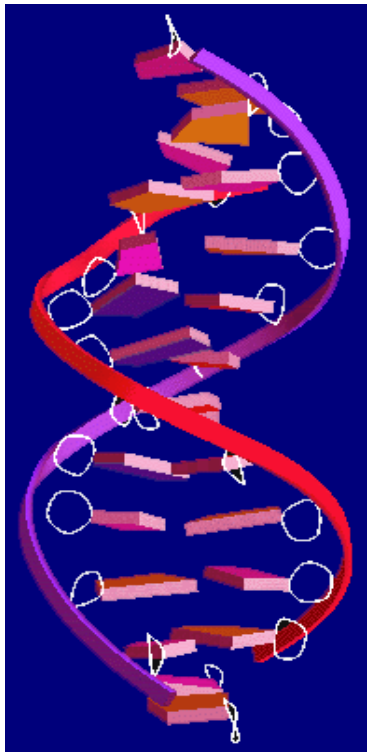


http://iweb.langara.bc.ca/biology/mario/Biol2315notes/biol2315chap21.html

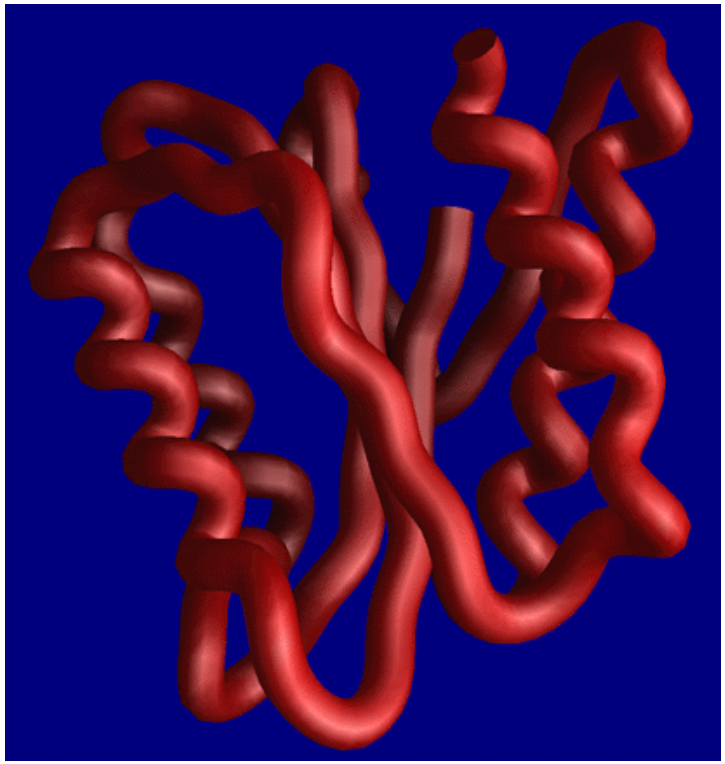


Genes e proteínas

DNA



Proteína



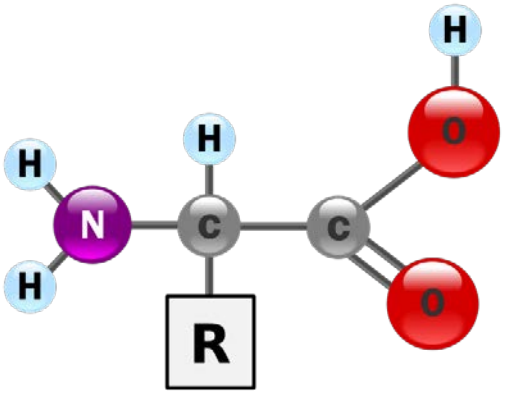
Hugues Sicotte, NCBI

Proteínas são macromoléculas

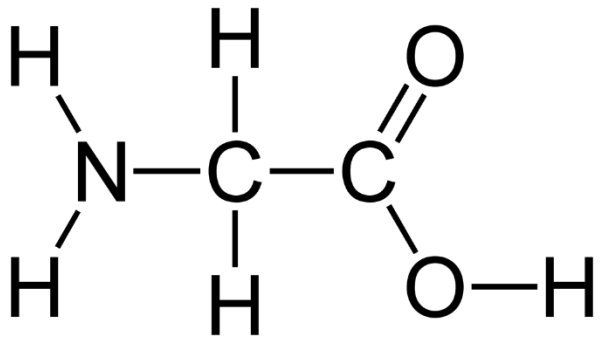
São cadeias de aminoácidos

> Protein sequence
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNI
DELLNEDILILGCSAMGDEVLEESEFEPFIEEISTKISGK
KVALFGSYGWGDGKWMRDFEERMNGYGCVVVETPLIVQNE
PDEAEQDCIEFGKKIANI

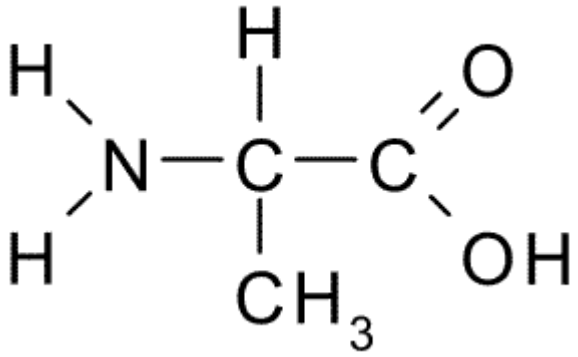
Aminoácidos



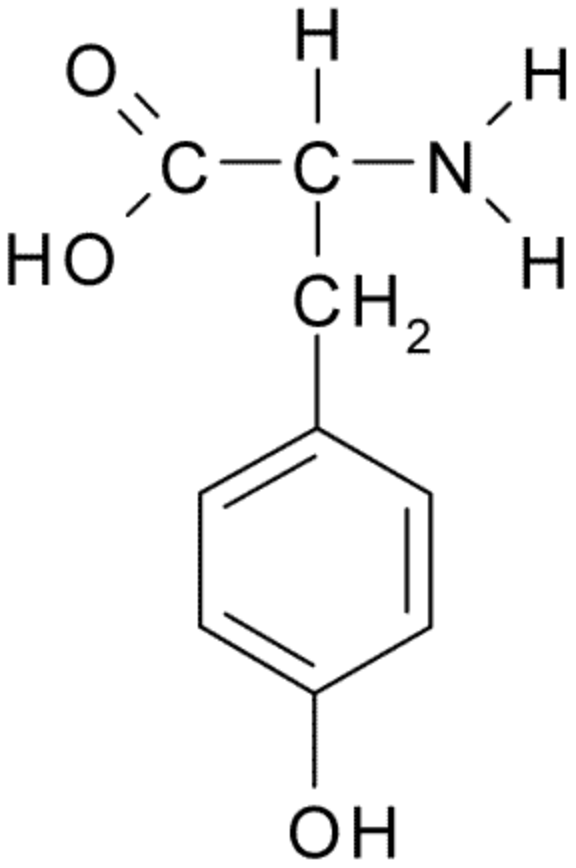
Estrutura genérica de um aminoácido



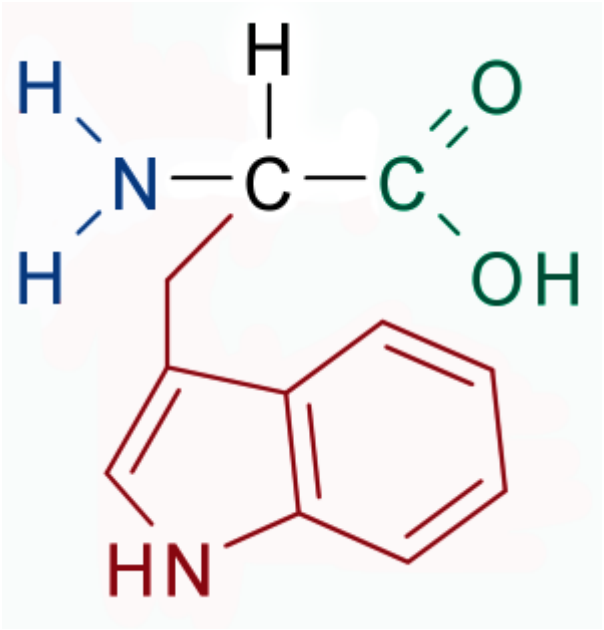
glicina



alanina



tirosina



triptofano

Twenty-One Amino Acids

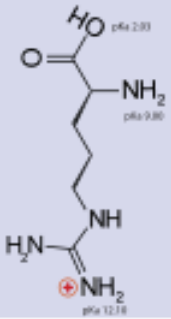
Positive Negative

• Side chain charge at physiological pH 7.4

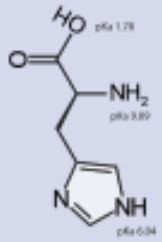
A. Amino Acids with Electrically Charged Side Chains

Positive

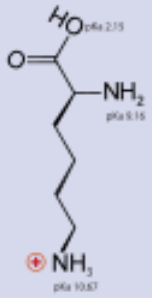
Arginine (Arg) R



Histidine (His) H

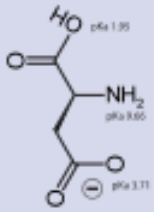


Lysine (Lys) K

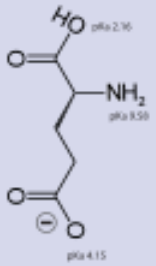


Negative

Aspartic Acid (Asp) D

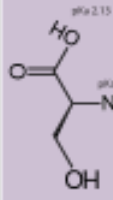


Glutamic Acid (Glu) E

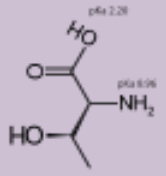


B. Amino Acids with Polar Uncharged Side Chains

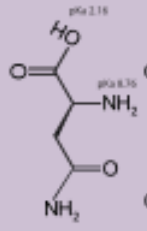
Serine (Ser) S



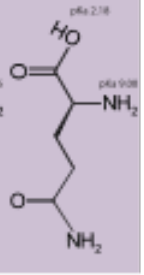
Threonine (Thr) T



Asparagine (Asn) N

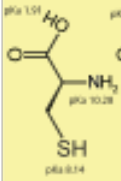


Glutamine (Gln) Q

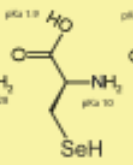


C. Special Cases

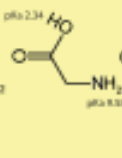
Cysteine (Cys) C



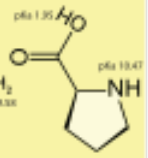
Selenocysteine (Sec) U



Glycine (Gly) G

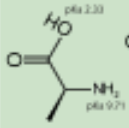


Proline (Pro) P

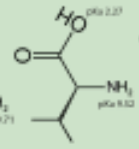


D. Amino Acids with Hydrophobic Side Chain

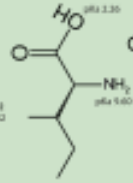
Alanine (Ala) A



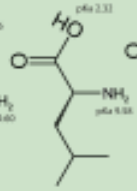
Valine (Val) V



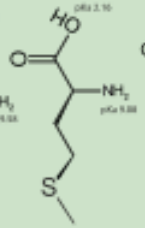
Isoleucine (Ile) I



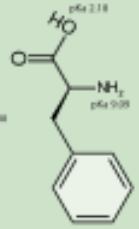
Leucine (Leu) L



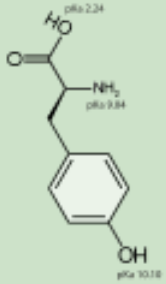
Methionine (Met) M



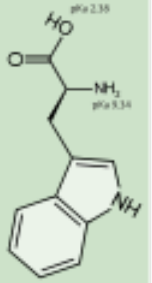
Phenylalanine (Phe) F



Tyrosine (Tyr) Y



Tryptophan (Trp) W

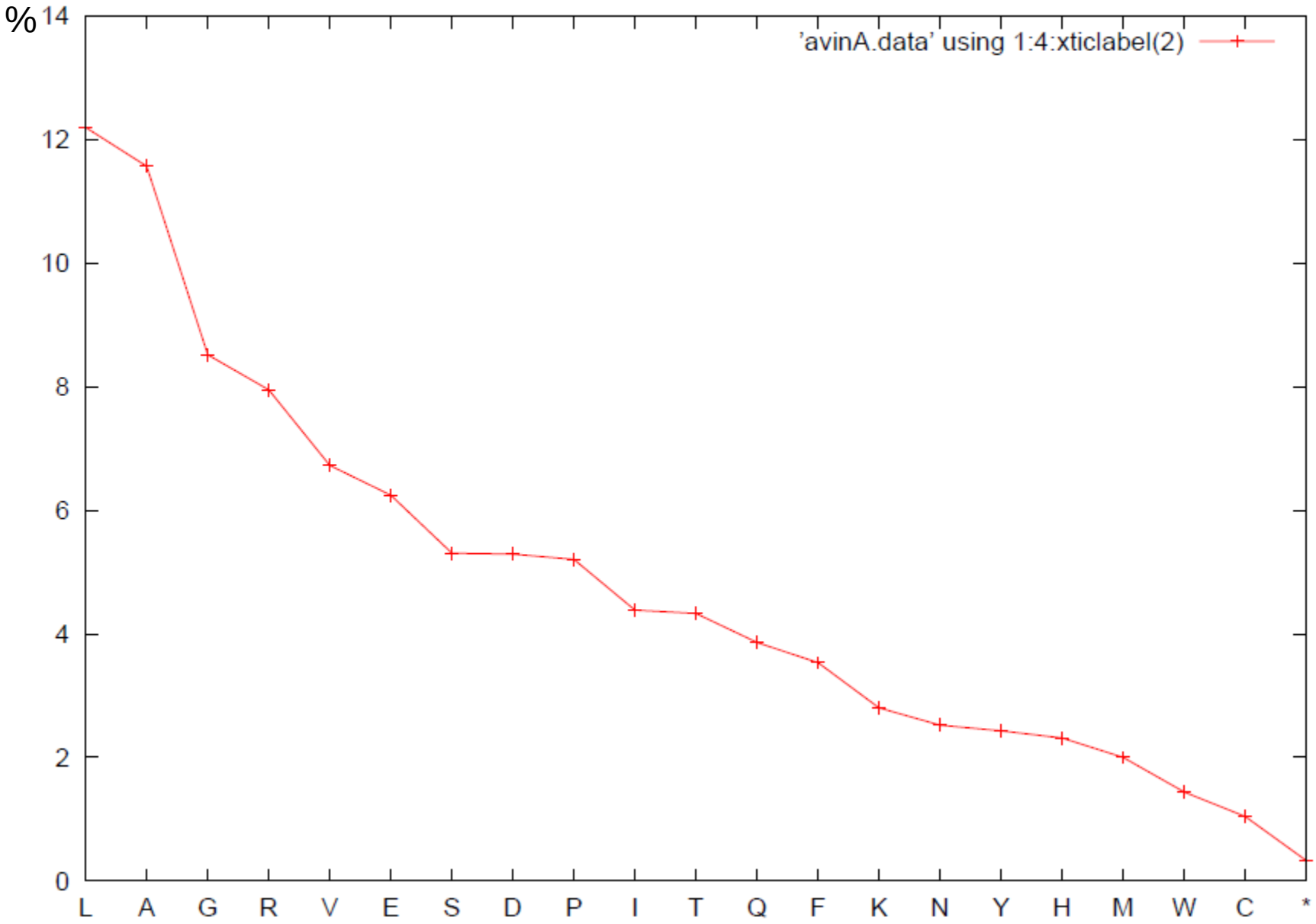


pKa Data: **CRC Handbook of Chemistry**, v. 2010

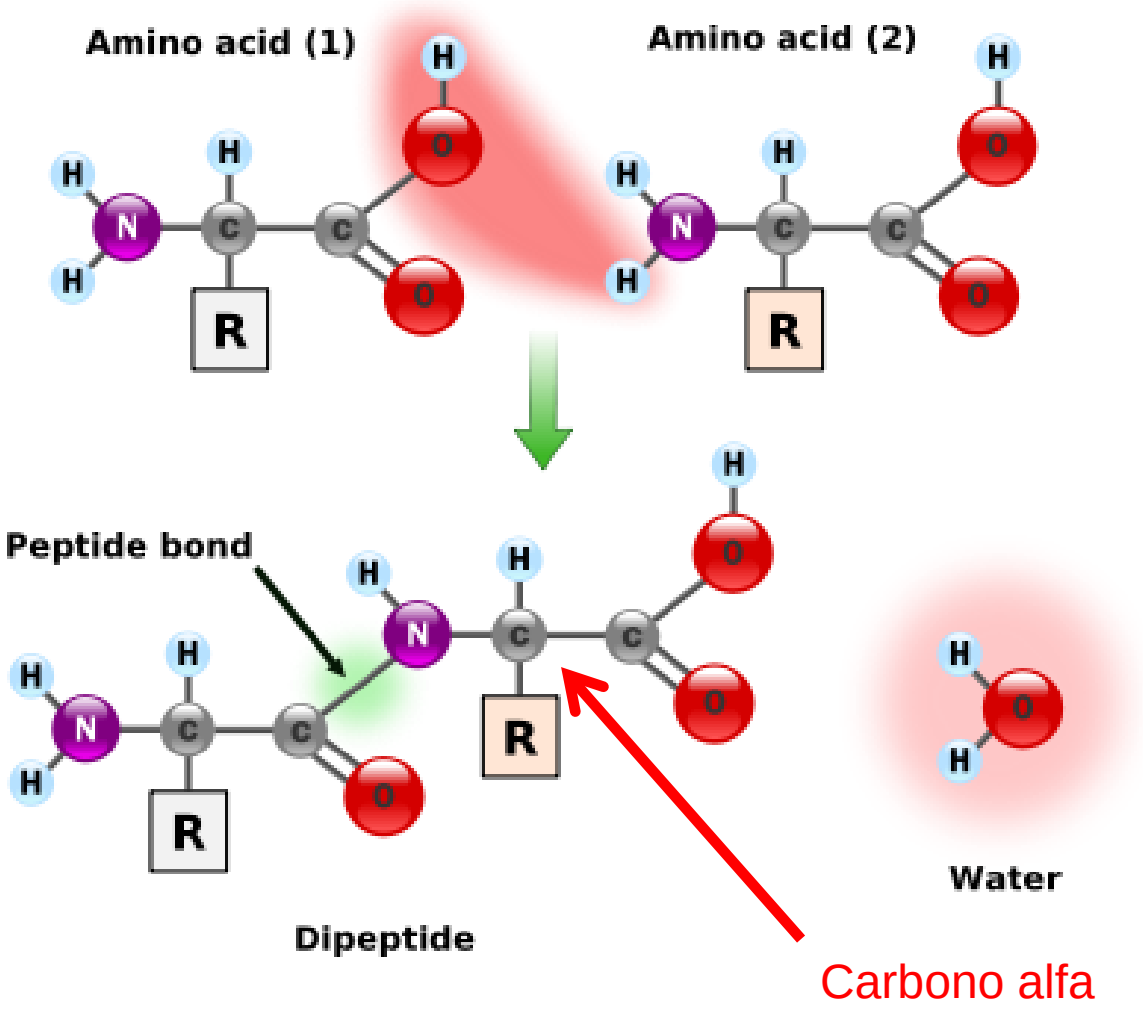
Dan Cojocari, Department of Medical Biophysics, **University of Toronto** 2011

Amino acid	3-letter code	1-letter code	MW (Da)	Structure
Alanine	Ala	A	89.1	CH3-CH(NH2)-COOH
Arginine	Arg	R	174.2	HN=C(NH2)-NH-(CH2)3-CH(NH2)-COOH
Asparagine	Asn	N	132.1	H2N-CO-CH2-CH(NH2)-COOH
Aspartic Acid	Asp	D	133.1	HOOC-CH2-CH(NH2)-COOH
Cysteine	Cys	C	121.2	HS-CH2-CH(NH2)-COOH
Glutamic Acid	Glu	E	147.1	HOOC-(CH2)2-CH(NH2)-COOH
Glutamine	Gln	Q	146.1	H2N-CO-(CH2)2-CH(NH2)-COOH
Glycine	Gly	G	75.1	NH2-CH2-COOH
Histidine	His	H	155.2	NH-CH=N-CH=C-CH2-CH(NH2)-COOH |_____|
Isoleucine	Ile	I	131.2	CH3-CH2-CH(CH3)-CH(NH2)-COOH
Leucine	Leu	L	131.2	(CH3)2-CH-CH2-CH(NH2)-COOH
Lysine	Lys	K	146.2	H2N-(CH2)4-CH(NH2)-COOH
Methionine	Met	M	149.2	CH3-S-(CH2)2-CH(NH2)-COOH
Phenylalanine	Phe	F	165.2	Ph-CH2-CH(NH2)-COOH
Proline	Pro	P	115.1	NH-(CH2)3-CH-COOH |_____|
Serine	Ser	S	105.1	HO-CH2-CH(NH2)-COOH
Threonine	Thr	T	119.1	CH3-CH(OH)-CH(NH2)-COOH
Tryptophan	Trp	W	204.2	Ph-NH-CH=C-CH2-CH(NH2)-COOH |_____|
Tyrosine	Tyr	Y	181.2	HO-p-Ph-CH2-CH(NH2)-COOH
Valine	Val	V	117.1	(CH3)2-CH-CH(NH2)-COOH

Frequência de aminoácidos em proteínas da bactéria *Azotobacter vinelandii*

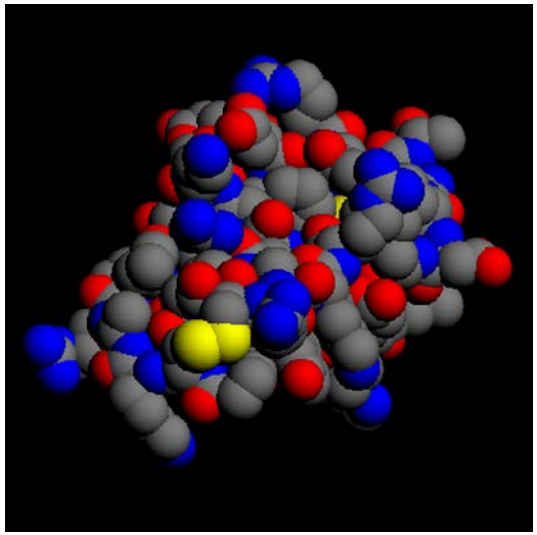


Os aminoácidos se ligam entre si por ligações peptídicas



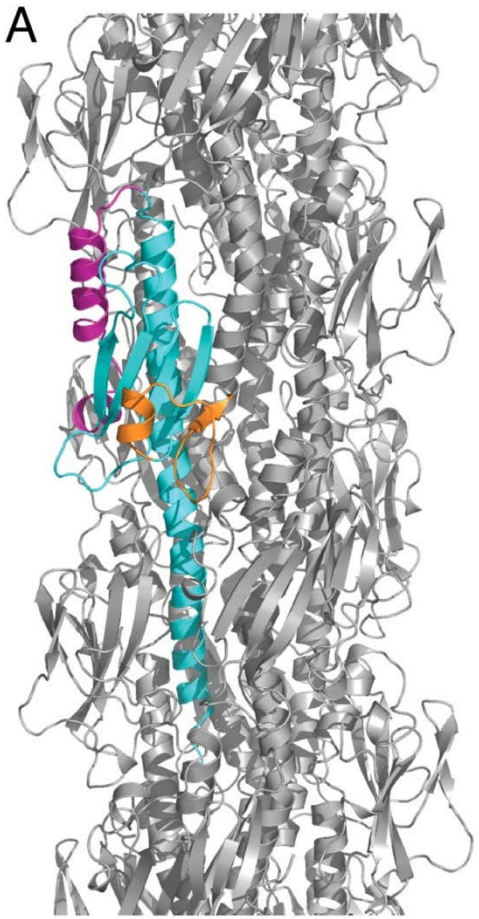
Proteínas formam estruturas tridimensionais

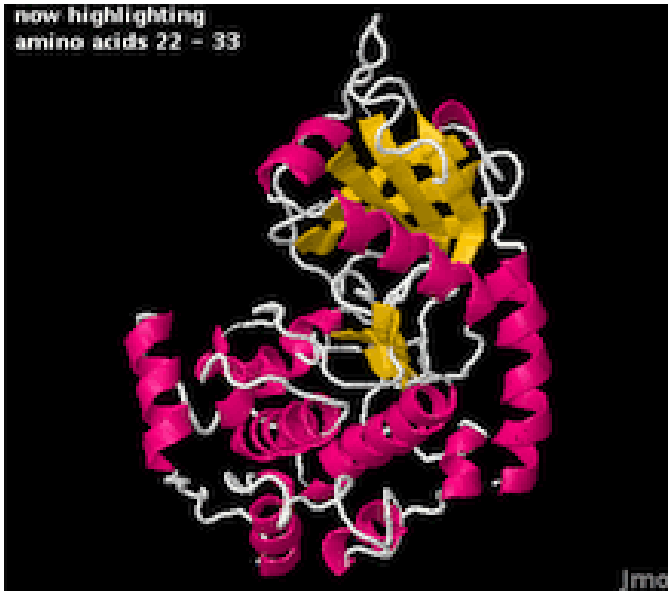
- São complexas e variadas
  - Diferentes proteínas tem diferentes estruturas



- (DNA é sempre hélice dupla)

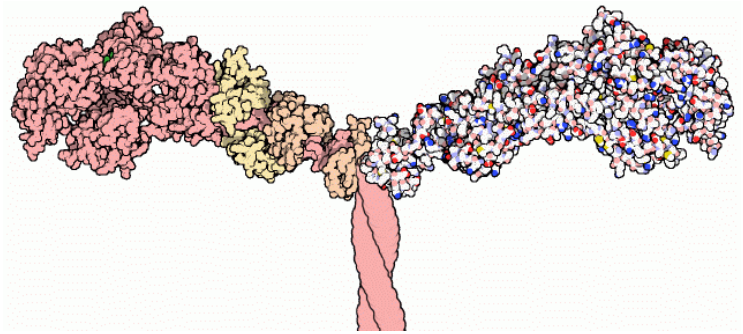
Exemplos de proteínas

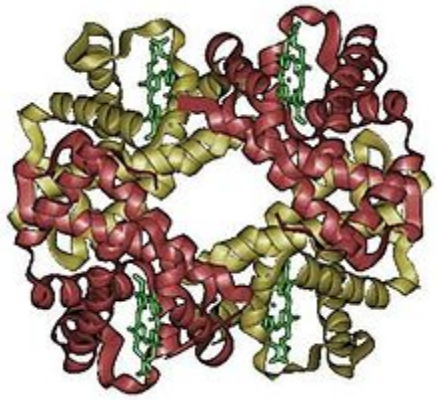




Diferentes estruturas conferem diferentes funções

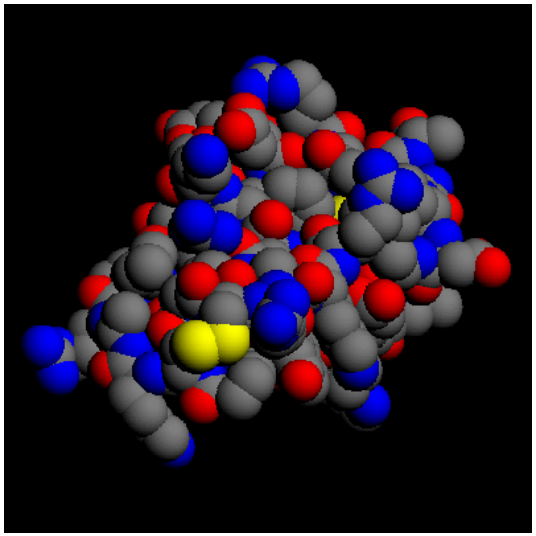
Hemoglobina: transporta oxigênio no sangue

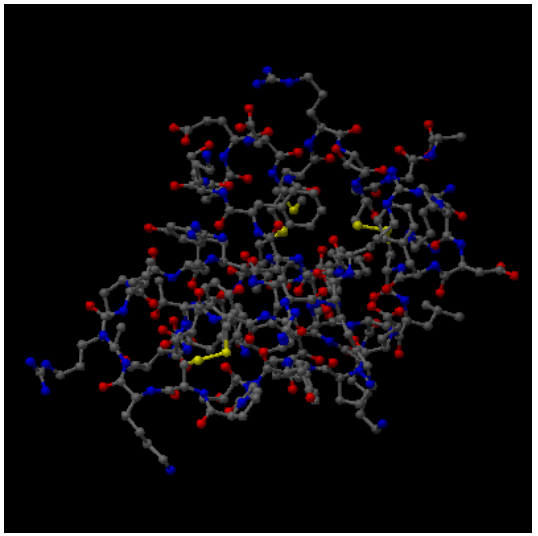




Miosina: contração do músculo

Proteínas podem ser visualizadas de diferentes formas

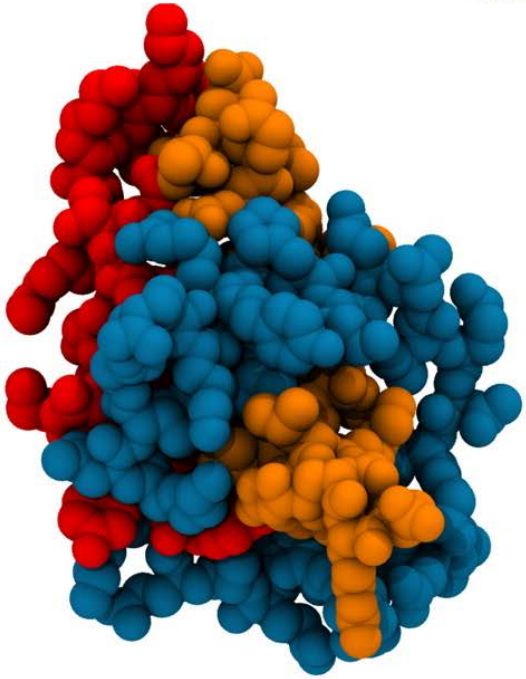


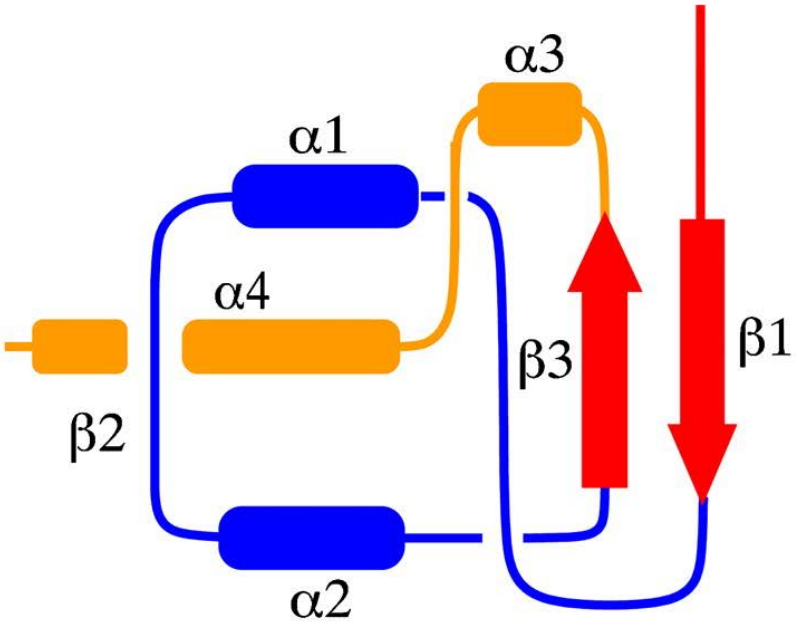


```
> Protein sequence
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNI
DELLNEDILILGCSAMGDEVLEESEFEPFIEEISTKISGK
KVALFGSYGWGDGKWMRDFEERMNGYGCVVVETPLIVQNE
PDEAEQDCIEFGKKIANI
```

Folha beta

Hélice alfa





Os aminoácidos estão para as proteínas assim como os núcleotídeos estão para o DNA

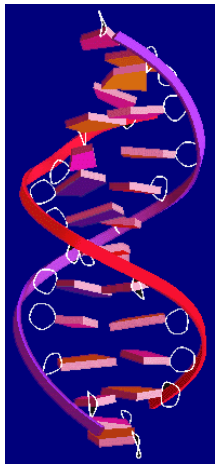
- Proteínas: 20 aminoácidos
- DNA: 4 nucleotídeos
- Ambas as cadeias têm direcionalidade
- DNA: 5’ → 3 ’
- Proteína: Amino terminal (N) → Carboxi terminal (C)
- Proteínas são bem mais curtas do que DNA
- Protéina típica tem 300 aa

Se 1 angstrom = 1 mm

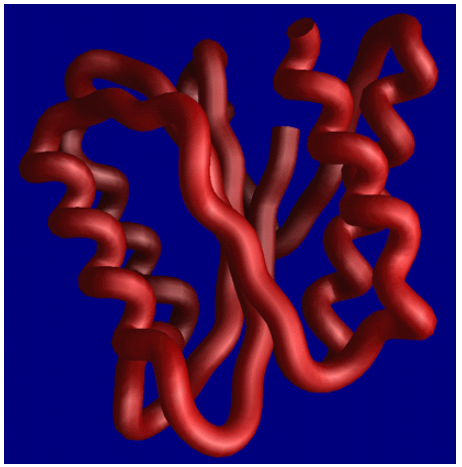
- Um cromossomo de 100.000.000 bp teria 340 km de comprimento
- Uma proteína de 300 aa teria 3 metros

Genes e proteínas

DNA



Proteína



```
> DNA sequence
ATGTCATGAAAATCGTATACTGGTCTGGTACCGGCAACAC
TGAGAAAATGGCAGAGCTCATCGCTAAAGGTATCATCGAA
TCTGGTAAAGACGTCAACACCATCAACGTGTCTGACGTTA
ACATCGATGAACTGCTGAACGAAGATATCCTGATCCTGGG
TTGCTCTGCCATGGGCGATGAAGTTCTCGAGGAAAGCGAA
TTTGAACCGTTCATCGAAGAGATCTCTACCAAAATCTCTG
GTAAGAAGGTTGCGCTGTTCGGTTCTTACGGTTGGGGCGA
CGGTAAGTGGATGCGTGACTTCGAAGAACGTATGAACGGC
TACGGTTGCGTTGTTGTTGAGACCCCGCTGATCGTTCAGA
ACGAGCCGGACGAAGCTGAGCAGGACTGCATCGAATTTGG
TAAGAAGATCGCGAACATCTAGTAGA
```

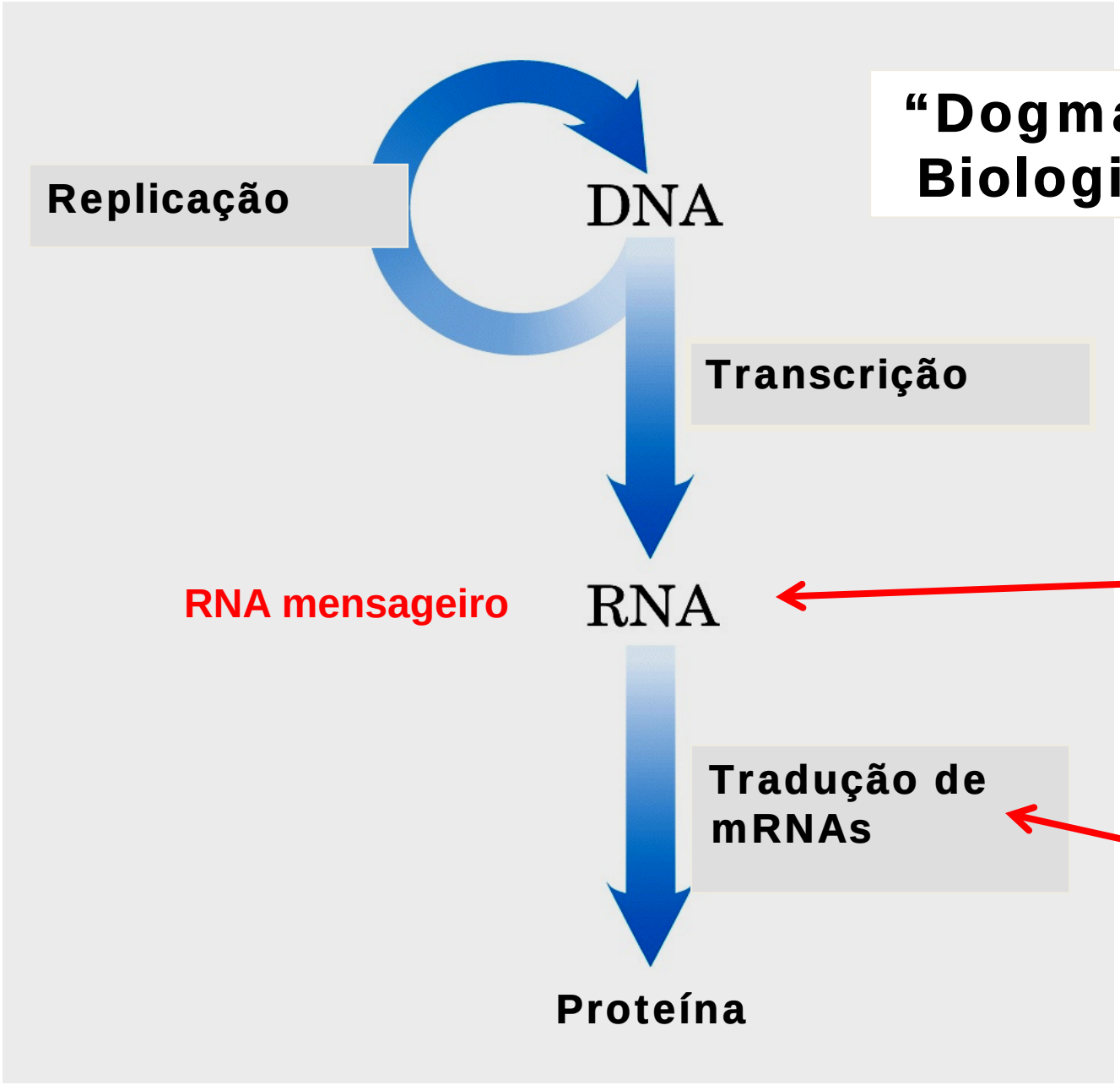
```
> Protein sequence
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNI
DELLNEDILILGCSAMGDEVLEESEFEPFIEEISTKISGK
KVALFGSYGWGDGKWMRDFEERMNGYGCVVVETPLIVQNE
PDEAEQDCIEFGKKIANI
```

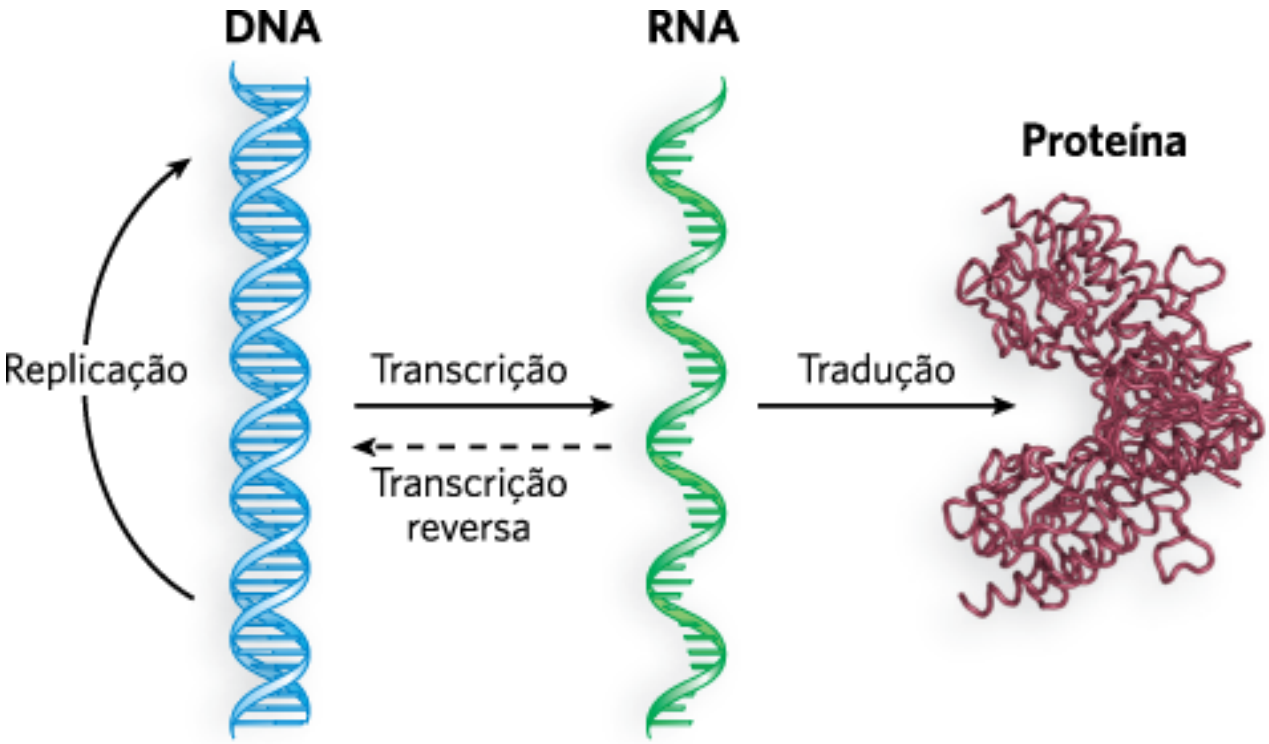

Hugues Sicotte, NCBI

Como um pedaço de DNA (gene) pode gerar uma proteína?

- Informacionalmente por meio de um código (o famoso código genético)
- Mecanicamente por meio de processos celulares chamados de
  - Transcrição
  - Tradução
- Esta aula: o processo informacional

“Dogma Central”da Biologia Molecular

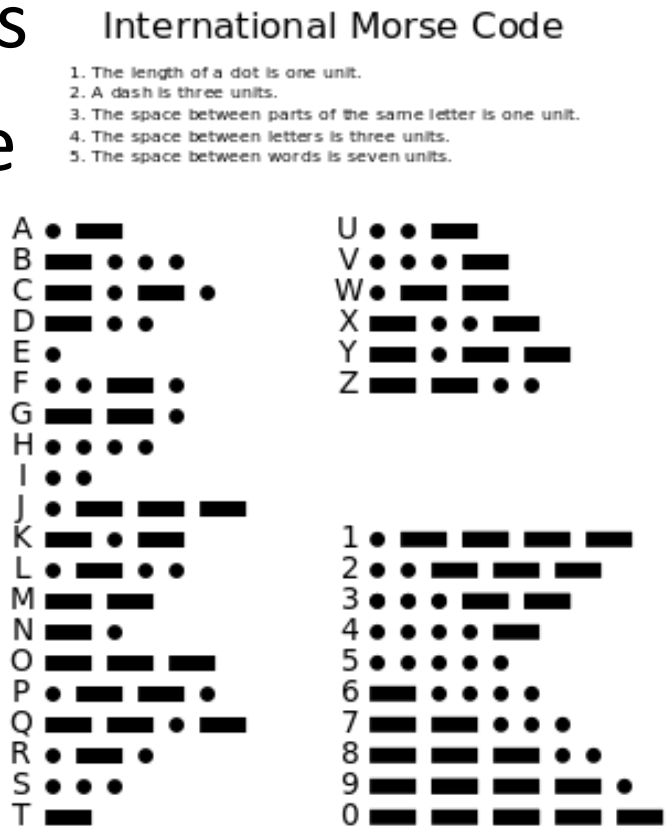




O dogma central do fluxo da informação: DNA→RNA→proteína. A informação para replicar o DNA está inerente na sua estrutura (seta curvada). A informação flui do DNA para o RNA por transcrição. A informação flui do RNA para a proteína por tradução. Em alguns casos, a informação também pode fluir de volta, do RNA para o DNA (transcrição reversa). Não existe nenhuma evidência de informação que flui da proteína para o ácido nucleico.

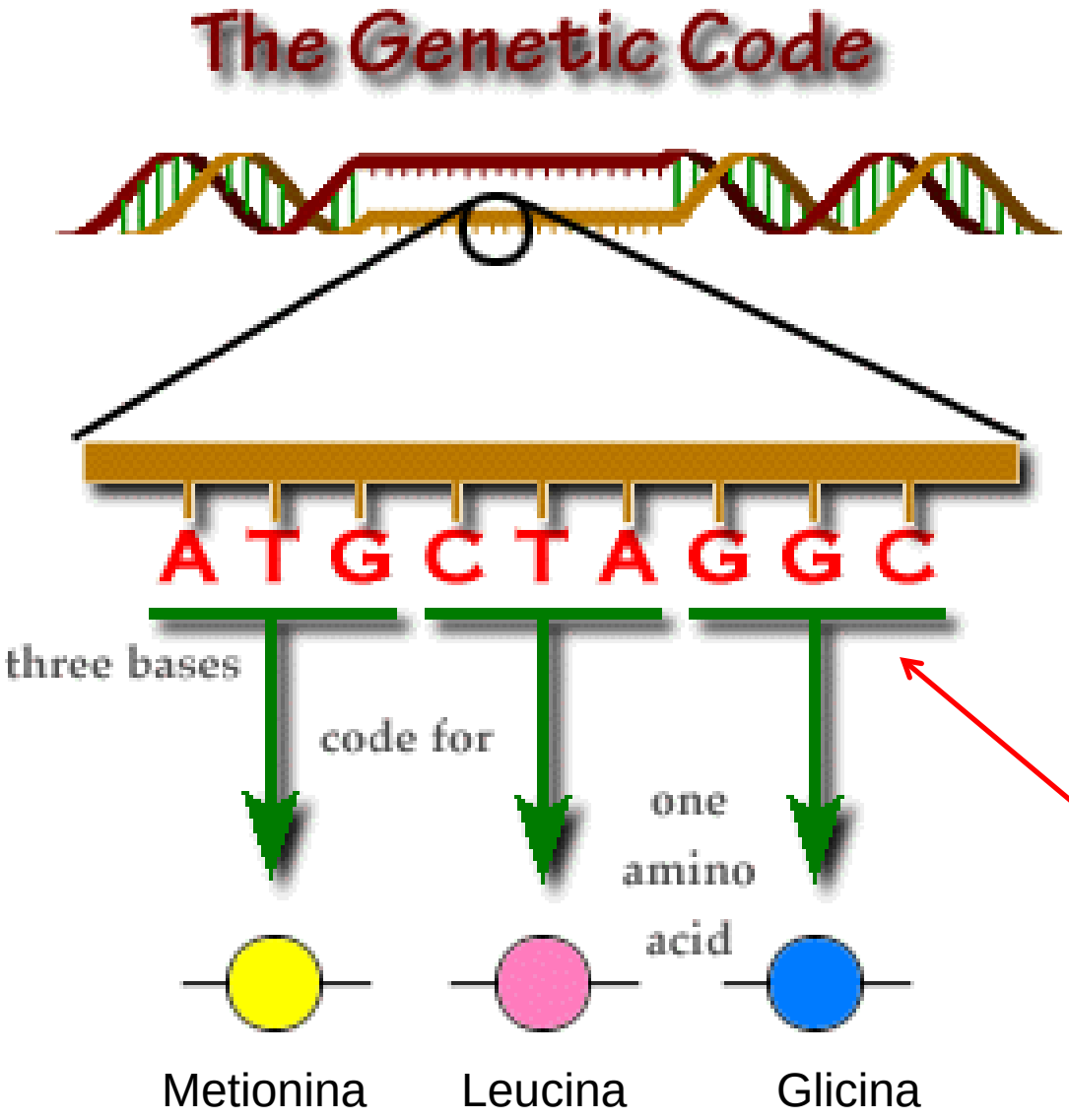
Código Genético

- Funciona como uma tabela
- Nucleotídeos → Aminoácidos
- Semelhante ao código Morse



Código genético

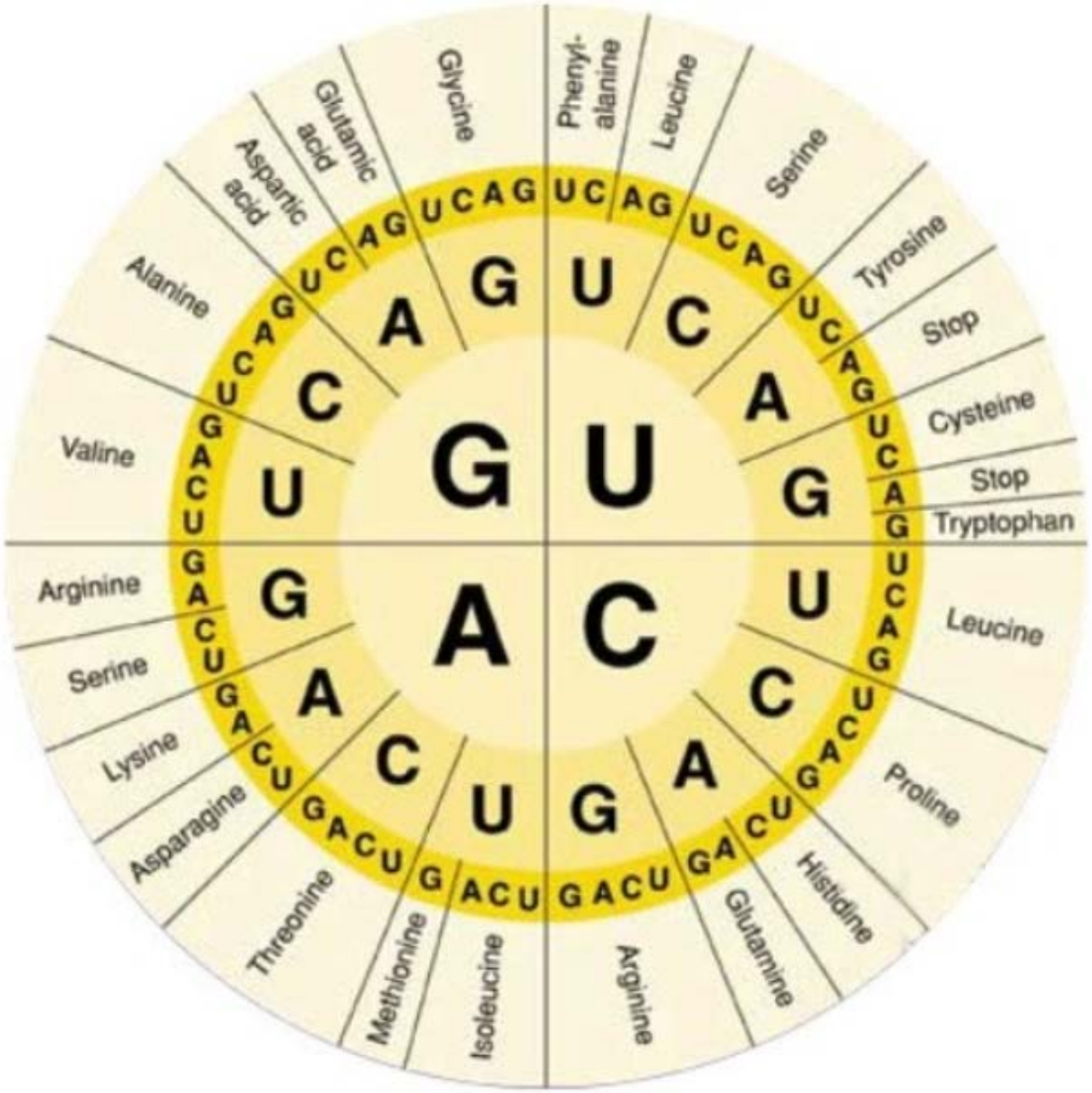
- 4 bases e 20 aminoácidos
- Um código 1:1 não dá
- Um código 2:1 também não dá
  - Apenas 16 possíveis pares
- Um código 3:1 dá (e sobra)
  - 64 possíveis trincas
- Lembrar que é preciso um sistema de pontuação (início e fim da região codificadora)


The Genetic Code
A T G C T A G G C
three bases
code for
one
amino
acid
Metionina
Leucina
Glicina

códon

A tabela com o código genético

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine	UCA		UAA	Stop (*Ochre*)	UGA	Stop (*Opal*)	A
	UUG		UCG		UAG	Stop (*Amber*)	UGG	(Trp/W) Tryptophan	G
C	CUU		CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	AUG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G


G U C A
G U C A
G U C A
G U C A
Glycine
Glutamic acid
Aspartic acid
Alanine
Valine
Phenyl-alanine
Leucine
Serine
Tyrosine
Stop
Cysteine
Stop
Tryptophan
Leucine
Proline
Histidine
Glutamine
Arginine
Isoleucine
Methionine
Threonine
Asparagine
Lysine
Serine
Arginine

Exercício de tradução

- Dada uma sequencia em DNA, mostrar sua tradução em aminoácidos

- Informacionalmente, podemos “pular” o passo de mRNA; ou seja, do DNA ir direto para proteína

- Na célula, nunca ocorre esse “pulo”

Exercícios

- Que aminoácidos são codificados por TCGTCTGATATTCTA?
- E por

CGGCCCTGGCCTCCGACATCGGCGCC?

O código genético é quase universal (o mesmo para todas as formas de vida)

- Em *Mycoplasma* (bactéria), UGA é Trp
- Em *Candida* (fungo), CUG é Ser
- …e outras pequenas variações

O código genético é degenerado

# codons	Aminoácidos	# aa	Total codons
6	Leu, Ser, Arg	3	18
4	Ala, Thr, Pro, Gly, Val	5	20
3	Ile, Stop	2	6
2	Phe, Tyr, His, Gln, Asn, Lys, Asp, Glu, Cys	9	18
1	Met, Trp	2	2
Totais		20+1	64

A degeneração ocorre principalmente por meio da terceira base

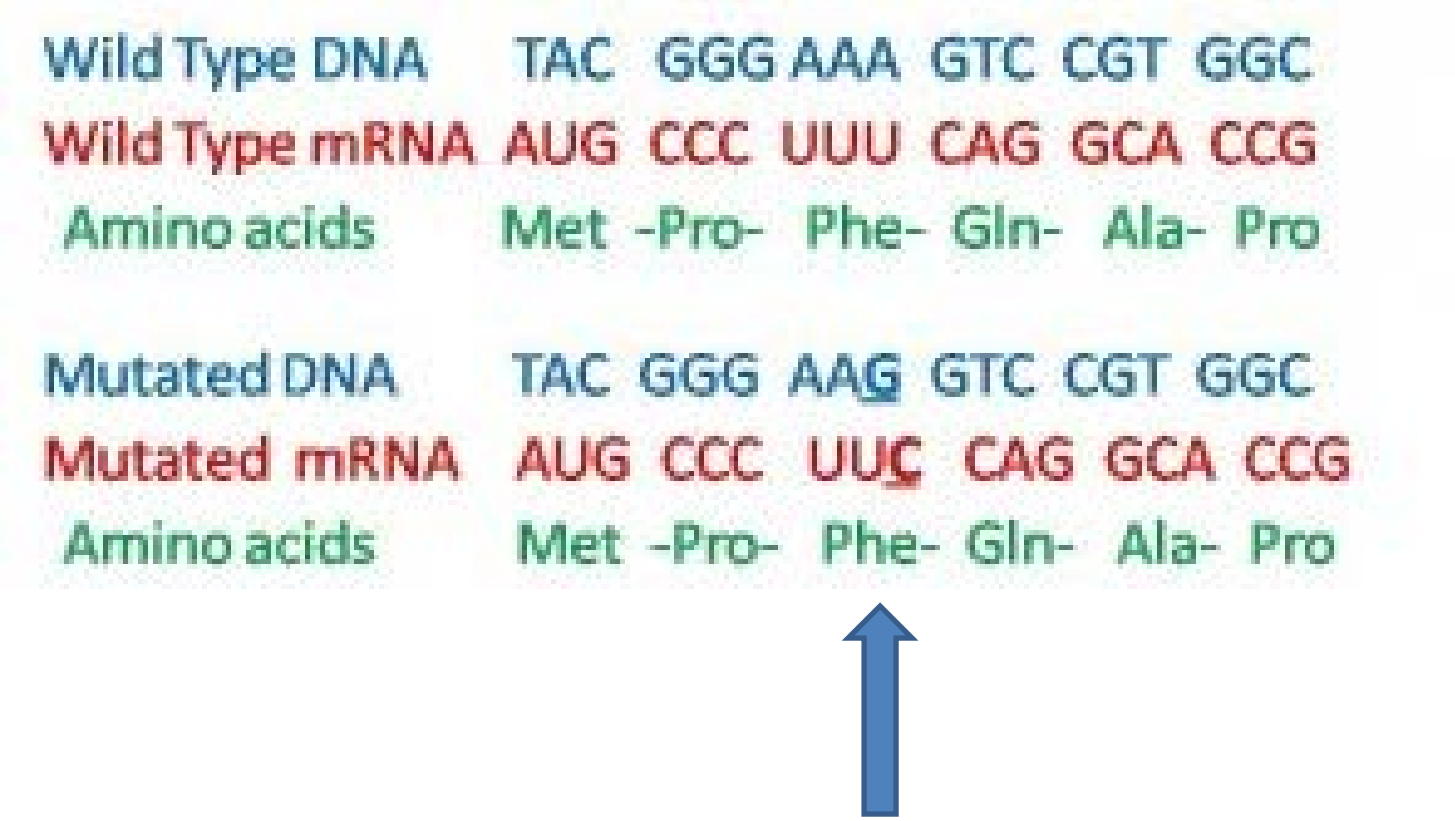
codon	AA
GG	Gly
CC	Pro
GC	Ala
CG	Arg (+2)
GU	Val
CU	Leu (+2)
UC	Ser (+2)
AC	Thr

AU: 3 possibilidades correspondem a Isoleucina; a outra é Met
UA: 2 possibilidades correspondem a STOP e 2 a Tirosina

Degeneração significa redundância

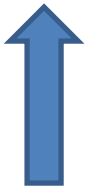
- Robustez em relação a erros
- Mutações sinônimas (ou silenciosas)
- Mutações não-sinônimas

Mutação sinônima ou silenciosa



Mutação não-sinônima

Wild Type DNA	TAC	CAC	CCC	GCC	ATC
Wild Type mRNA	AUG	GUG	GGG	CGG	UAG
Amino acids	Met	-Val-	Gly-	Arg-	Stop
Mutated DNA	TAC	**G**AC	CCC	GCC	ATC
Mutated mRNA	AUG	**C**UG	GGG	CGG	UAG
Amino acids	Met	-Leu-	Gly-	Arg-	Stop



Mutações são boas ou más?

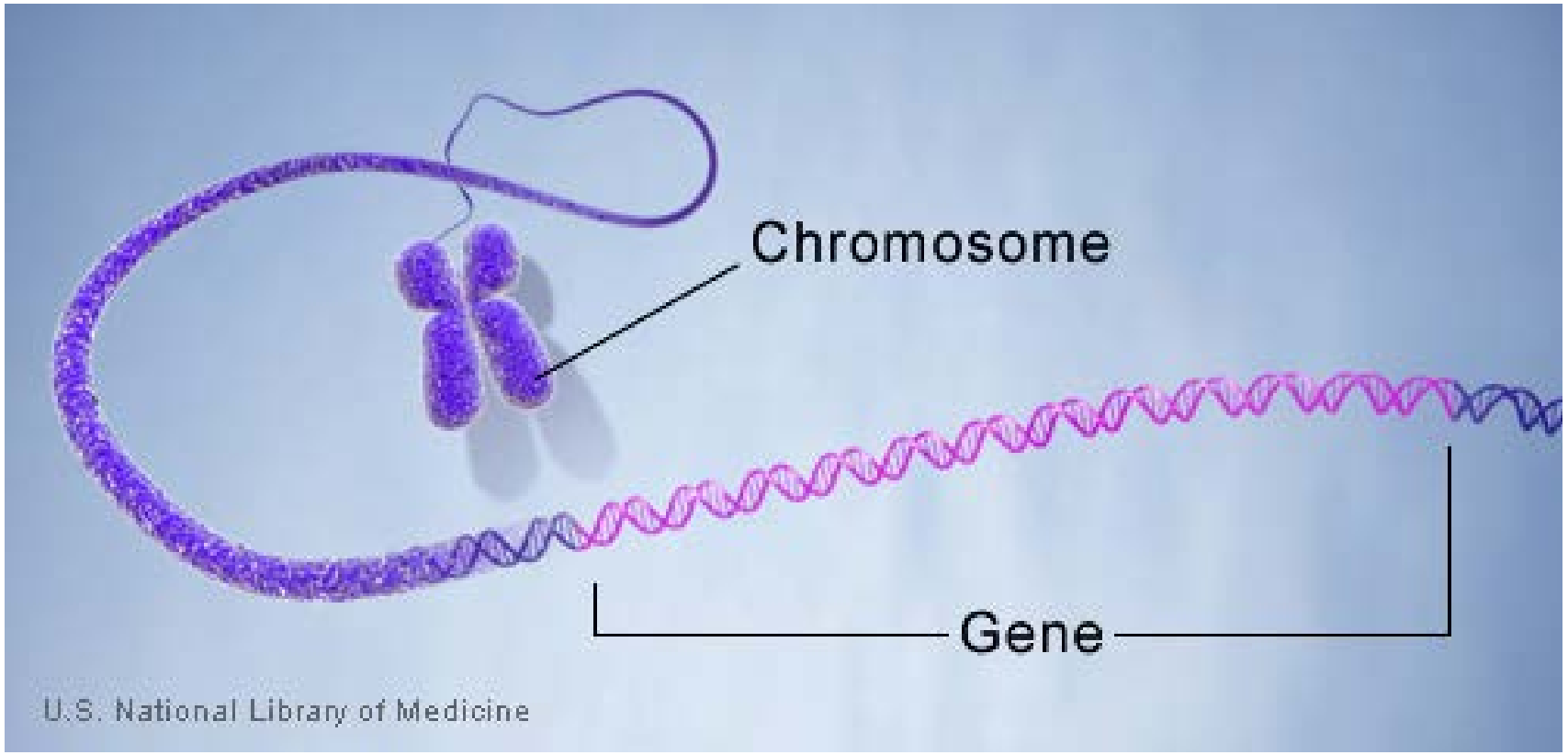
- Neutras
  - exemplo: mutação sinônima
- Más
  - Quando causam doença
- Boas
  - Quando dão uma vantagem competitiva ao indivíduo
  - Combustível da evolução!
- Algumas mutações são boas e más ao mesmo tempo!

Anemia falciforme

- Mutação não sinônima numa única posição de hemoglobina (amino ácido num. 7)
- GAA (glu) → GUA (val)
- GAG (glu) → GUG (val)
- Valina é hidrofóbica e Ácido glutâmico não é

A mutação de Af é boa e má ao mesmo tempo

- Má: Anemia falciforme é uma doença
- Boa: Indivíduos com essa mutação tem proteção contra malária!
- Anemia falciforme é prevalente nas regiões da África que historicamente foram (e são) afetadas pela malária

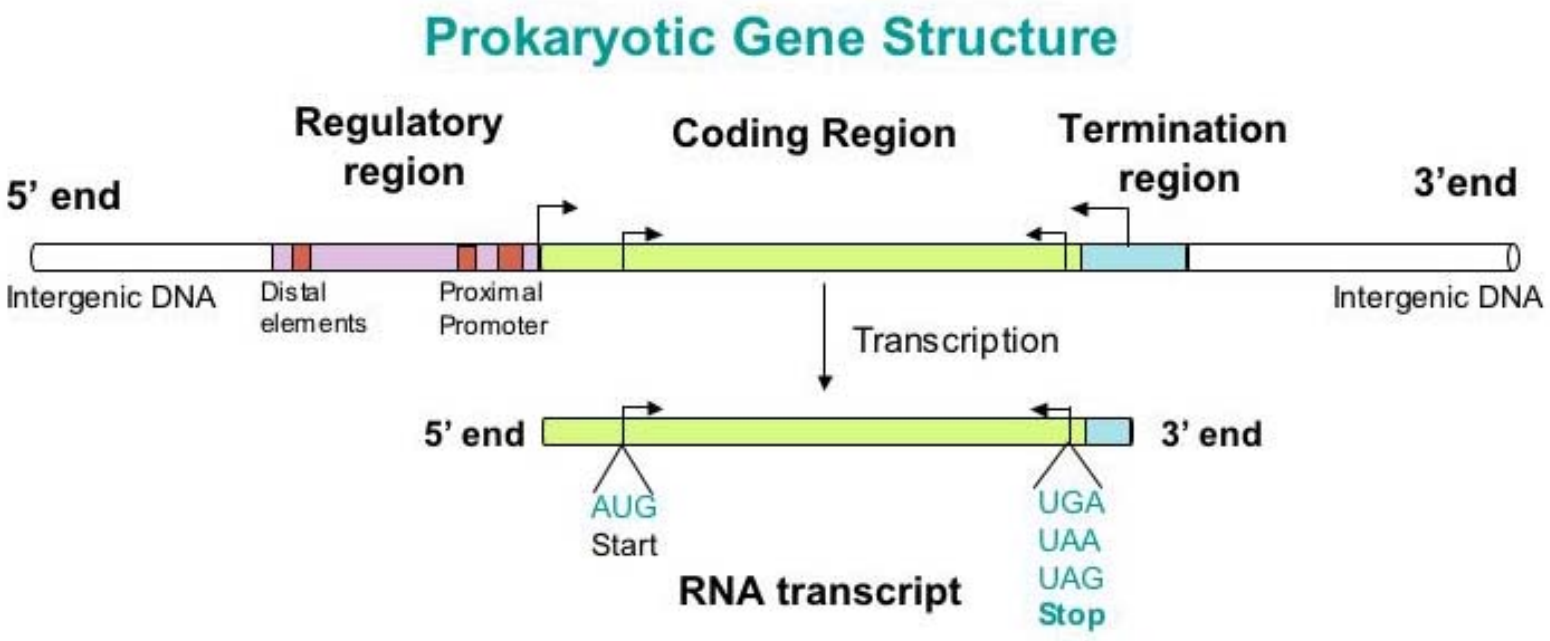

Chromosome
Gene
U.S. National Library of Medicine

Um gene numa sequência

AGCTCGCGCTCCGCATCCATCCAGTAGGGTTCGGTGTCGACGAGCGTGCC
GTCCATATCCCAGAAGACGGCGGCCGGCATCGCGTGCGGAGTCAGTTCGG
TCACGGCTGACAAGTCTATCCCGGCGGCCCCGGGCCTATTCTTGAGGGAC
GGCGTCCTGACCGGTCGCCGGATGAAAGGACCAGAACGCCCCGTGACTGA
CGCGAACAGCATCCTCGGAGGGCGCATCCTCATGGTGGCCTTCGAAGGGT
GGAACGACGCTGGCGAGGCCGCCAGCGGGGCCGTCAAGACGCTCAAGGAC
CAGCTGGATGTCGTCCCGGTCGCCGAGGTCGATCCCGAGCTGTACTTCGA
CTTCCAGTTCAACCGGCCGGTCGTCGCGGACGACGACGGCCGCCGGCGCC
TCATCTGGCCGTCCGCGGAGATCCTGGGCCCAGCTCGCCCCGGCGACACC
GGCGATGCGCGCCTGGACGCCACCGGCGCCAACGCGGGCAATATCTTCCT
TCTCCTCGGCACCGAGCCGTCGCGCAGCTGGCGCAGCTTCACCGCGGAGA
TCATGGATGCGGCCCTGGCCTCCGACATCGGCGCCATCGTCTTCCTCGGT
GCGATGCTGGCGGACGTACCGCACACCCGCCCCATCTCCATCTTCGCTTC
GAGCGAGAACGCGGCCGTCCGTGCGGAGCTCGGCATCGAACGCTCTTCGT
ACGAGGGGCCGGTCGGTATCCTGAGCGCGCTCGCCGAAGGGGCGGAGGAC
GTGGGCATTCCGACCATCTCCATCTGGGCGTCGGTTCCGCACTATGTCCA
CAATGCGCCCAGCCCGAAGGCGGTGCTCGCACTGATCGACAAGCTCGAAG
AGCTGGTGAATGTCACCATCCCGCGTGGCTCGCTGGTGGAGGAGGCCACG
GCCTGGGAAGCCGGGATCGACGCGCTGGCTCTGGACGACGACGAGATGGC
TACGTACATCCAGCAGCTGGAGCAGGCACGCGACACCGTGGACTCCCCTG
AGGCCAGCGGCGAGGCGATCGCCCAGGAGTTCGAGCGCTACCTCCGCCGC
CGCGACGGCCGCGCCGGCGATGACCCCCGCCGTGGCTGACGTCACCCCCT
CTCTGCGTCCGCCGTCCTCTGTTCCCCCCGCTCGGCCTCCCCTGAGGCCG
AGGAGTCGCGCCCACATGCCGGAAACTCCTCCTTTCCTGACTTTCTGGAG

Início e fim da porção codificadora de um gene de procarioto

- O início é quase sempre um ATG = metionina
- O final é sempre um codon de parada



Sequenciamento de DNA

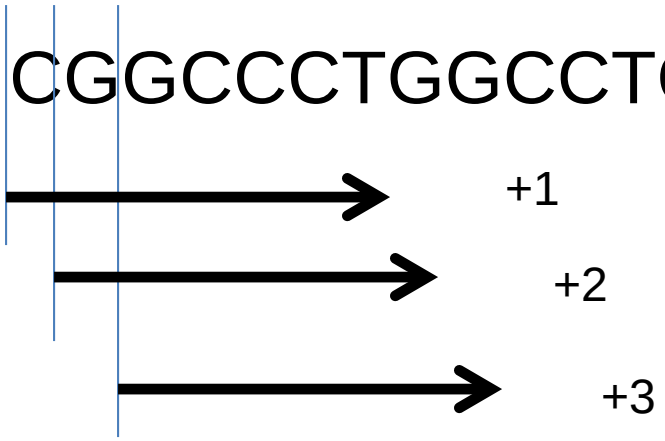
- Gera imensas quantidades de sequências de DNA, armazenadas em arquivos de computadores
- Em geral são fragmentos onde a informação da fita e da localização no cromossomo foi perdida
- Para achar onde estão os genes de forma computacional, precisamos do conceito de Quadro de Leitura

Quadros de leitura

Uma fita dupla de DNA admite

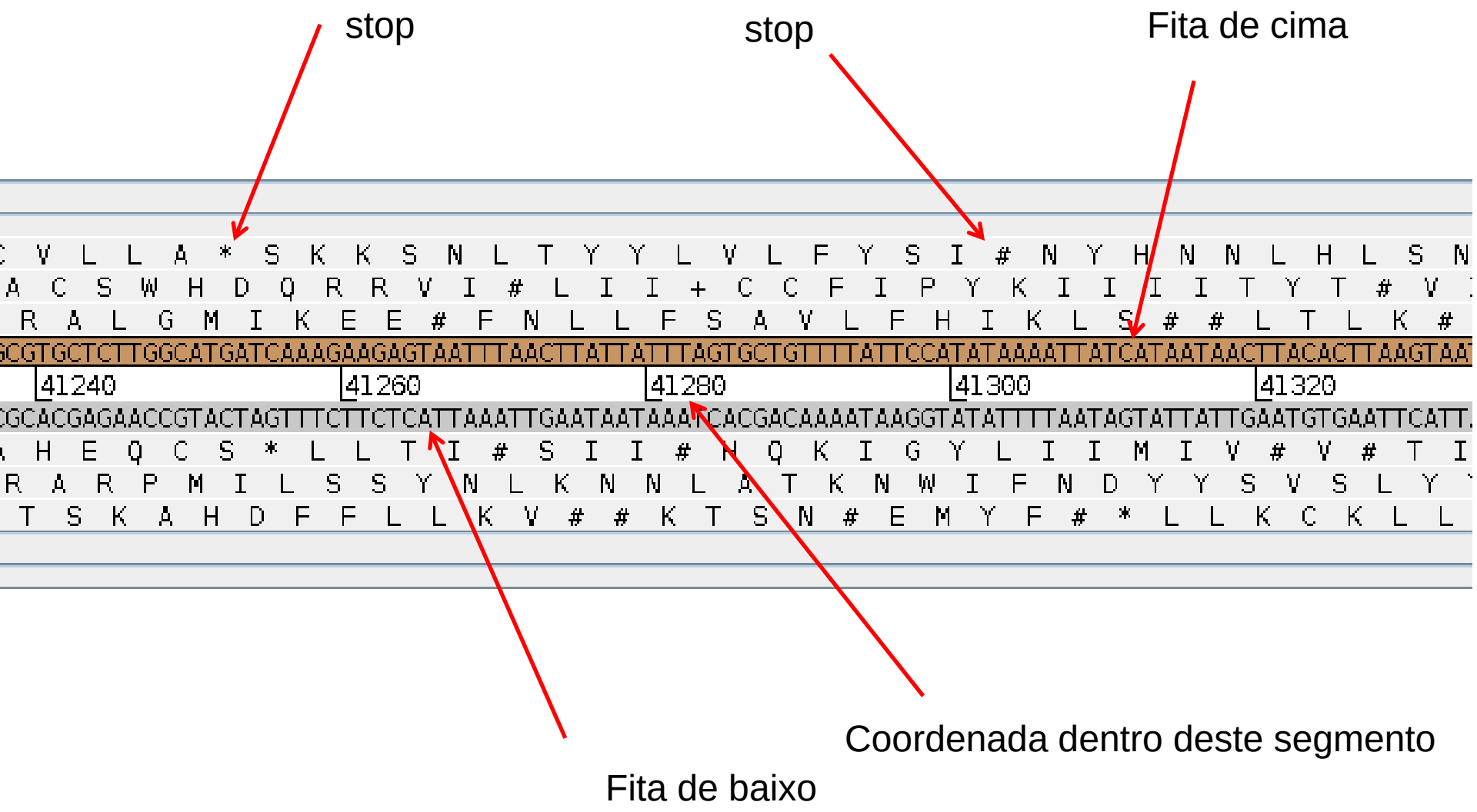
6 quadros de leitura

(3 em cada fita)



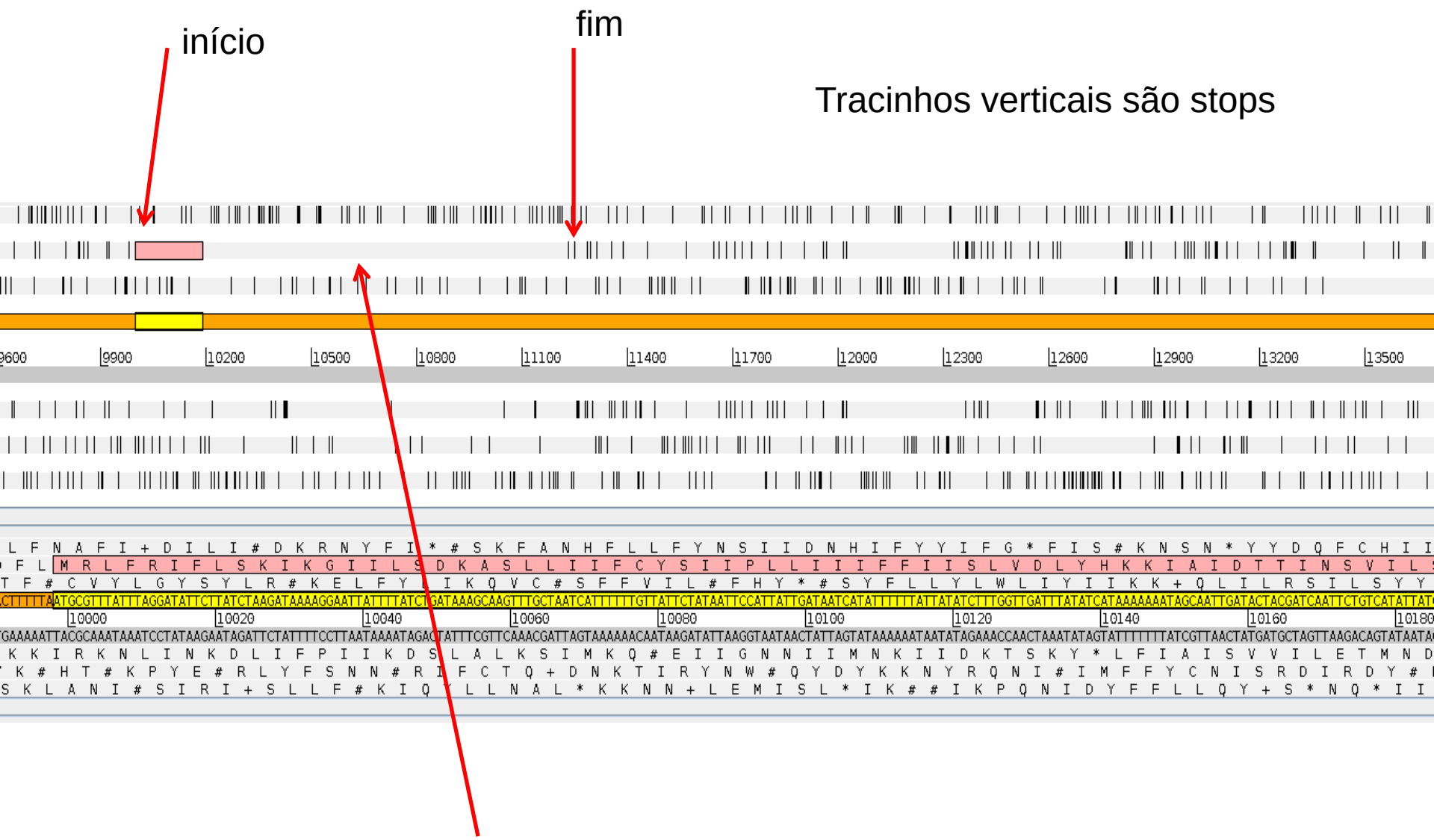
Na outra fita:
– 1, – 2, – 3

Cada quadro tem sua própria tradução


stop
stop
Fita de cima
41240
41260
41280
41300
41320
Fita de baixo
Coordenada dentro deste segmento

Quadro aberto de leitura

- *Open reading frame (ORF)*
- É um quadro de leitura
  - Com número de bases múltiplo de 3
  - Terminando em STOP
  - Sem outros STOPs no meio
- A porção codificadora (de proteína) de um gene bacteriano é um quadro aberto de leitura iniciado por ATG (muito mais raramente por GTG ou CTG)



Exemplo de uma ORF bacteriana

Estas imagens vieram do navegador de genomas ARTEMIS

Disponível (de graça) em

https://www.sanger.ac.uk/resources/software/artemis/

Exercício

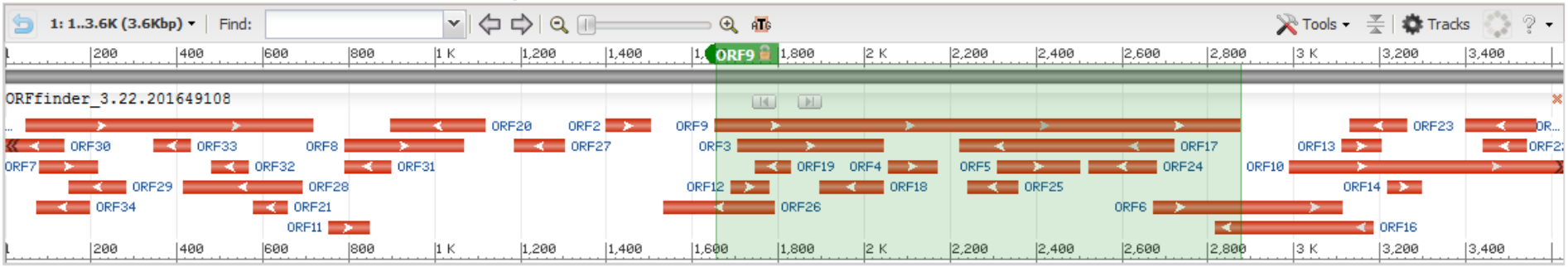
- Dada uma sequência de DNA, achar uma ORF fazendo a tradução nos 6 quadros de leitura
- Verifique o resultado usando

https://www.ncbi.nlm.nih.gov/orffinder/

Open Reading Frame Viewer

Sequence

ORFs found: 34 Genetic code: 1 Start codon: 'ATG' only



Add six-frame translation track

ORF9 (407 aa) Display ORF as... Mark

```
>lcl|ORF9
MLSPACPVIPGGHVCTVSCRSRILRTDRHAGLQPRAHRSM
HVVRLSIHRLRRFQTVELHPSSALNLLTGDNGAGKTSVLE
ALHLMAYGRSFRGRVRDGLIQQGANDLEVFVEWKEGGGAA
VERTRRAGLRHSGQEWTGRLDGEDVAQLGSLCAALAVVTF
EPGSHVLISGGGEPRRRFLDWGLFHVEPDFLTLWRRYARA
LKQRNALLKQGAQPRMLDAWDNELAESGETLTSRRMRYLE
RLQDRLVPVADAIAPALGLSALTFAPGWKRHEVSLADALL
LARERDRQNGYTSQGPHRADWMPSFHALPGKDALSRGQAK
LTALACLLAQAEDFAFERGEWPVIALDDLGSELDRHHQGR
VLQRLASAPAQVLITATETPPGLADAAALLQQFHVEHGQI
ARQATVN
```

Mark subset... Marked: 0 Download marked set as Protein FASTA

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF9	+	2	1652	2875	1224 | 407
ORF1	+	1	49	717	669 | 222
ORF10	+	2	2990	>3640	651 | 216
ORF17	-	1	2722	2222	501 | 166
ORF6	+	1	2671	3114	444 | 147
ORF16	-	1	3184	2816	369 | 122
ORF3	+	1	1705	2046	342 | 113
ORF8	+	2	791	1069	279 | 92
ORF28	-	2	693	415	279 | 92
ORF26		2	1701	1534	258 | 85

SmartBLAST ORF9

BLAST ORF9 BLAST marked set

BLAST Database:

UniProtKB/Swiss-Prot (swissprot)

Go back to the submitting page...