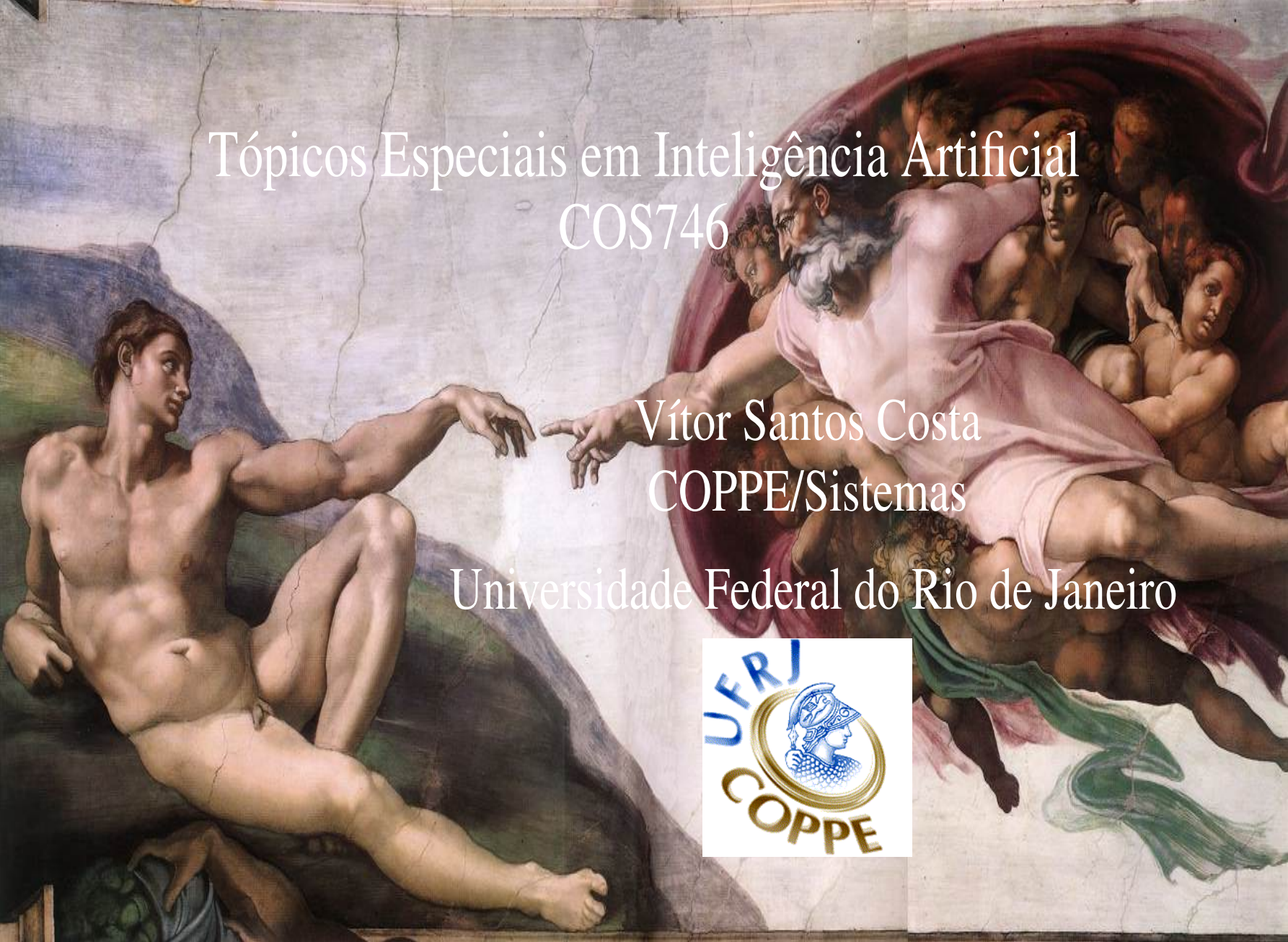


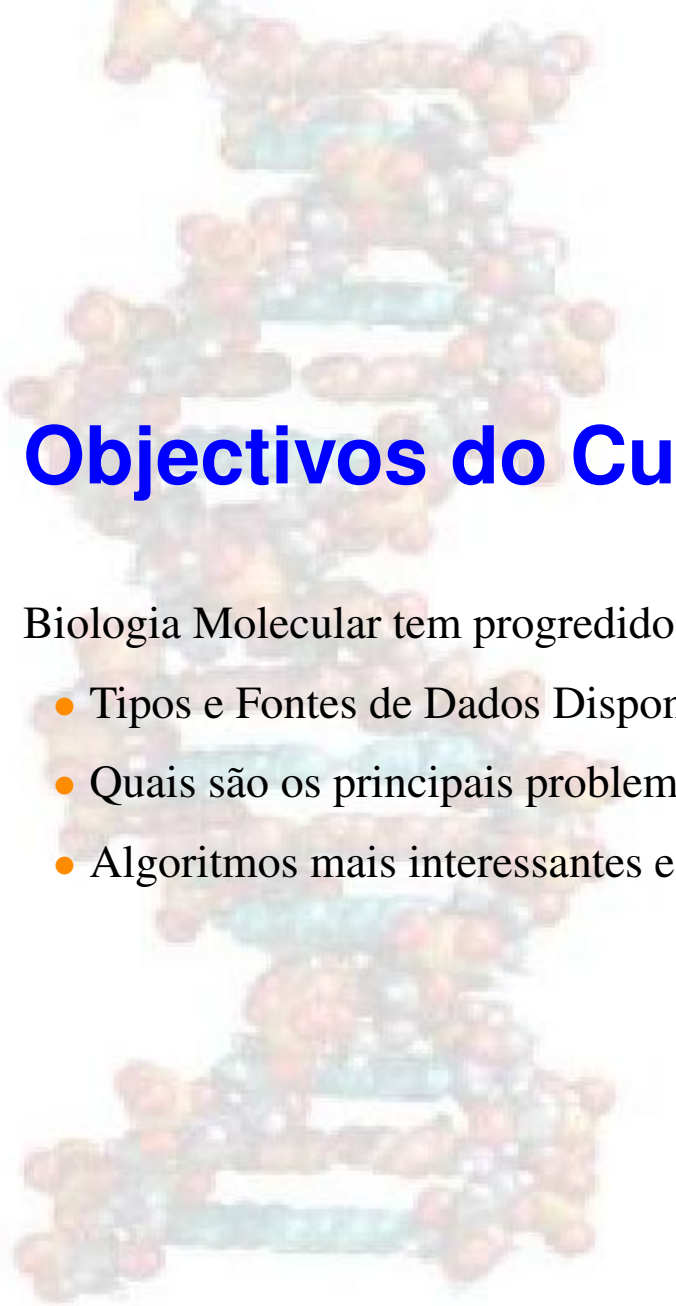
Tópicos Especiais em Inteligência Artificial
COS746

Vítor Santos Costa
COPPE/Sistemas
Universidade Federal do Rio de Janeiro

Introdução

Fundamentos:

1. Algoritmos e Estruturas de Dados (COS)
2. Estatística: conveniente.
3. Biologia Molecular: não se espera bk, mas interesse na área é requerido.



Objectivos do Curso

Biologia Molecular tem progredido rapidamente nos últimos anos:

- Tipos e Fontes de Dados Disponíveis em Biologia Molecular;
- Quais são os principais problemas computacionais;
- Algoritmos mais interessantes e relevantes.

Bibliografia

1. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Cambridge University Press, 1998.
2. Introduction to Computational Molecular Biology. J. Setubal and J. Meidanis. PWS Publishing, 1997.
3. Papers, etc.

Cursos Relacionados

- Introduction to BioInformatics (Mark Craven, UW-Madison): http://www.biostat.wisc.edu/bmi576/
- MO640/MC931 Biologia Computacional (João Carlos Setúbal, Unicamp): http://onsona.lbi.ic.unicamp.br/biocomp/
- Computational Molecular Biology (Sean Eddy, Washington University): http://bio5495.wustl.edu/
- Algorithms for Molecular Biology (Ron Shamir, Tel Aviv University): http://www.math.tau.ac.il/~rshamir/algmb.html
- Computational Molecular Biology (Doug Brutlag & Lee Kozar, Stanford): http://biochem.stanford.edu/biochem218/
- Representations and Algorithms for Computational Molecular Biology (Russ Altman, Stanford): http://www.smi.stanford.edu/projects/helix/bmi214/
- Introduction to Computational Molecular Biology (Peter Clote, MIT): http://theory.lcs.mit.edu/~bab/class/01-18.417-home.html

BioInformática

Processamento/armazenamento/apresentação/pesquisa de dados biológicos:

1. *sequências*;
2. *estruturas*;
3. *funções*;
4. *níveis de actividade*;
5. *redes de interacção*;

de/entre biomoléculas.

Também conhecida como *Biologia Computacional* ou *Biologia Molecular Computacional*

Problemas Computacionais em Biologia Molecular

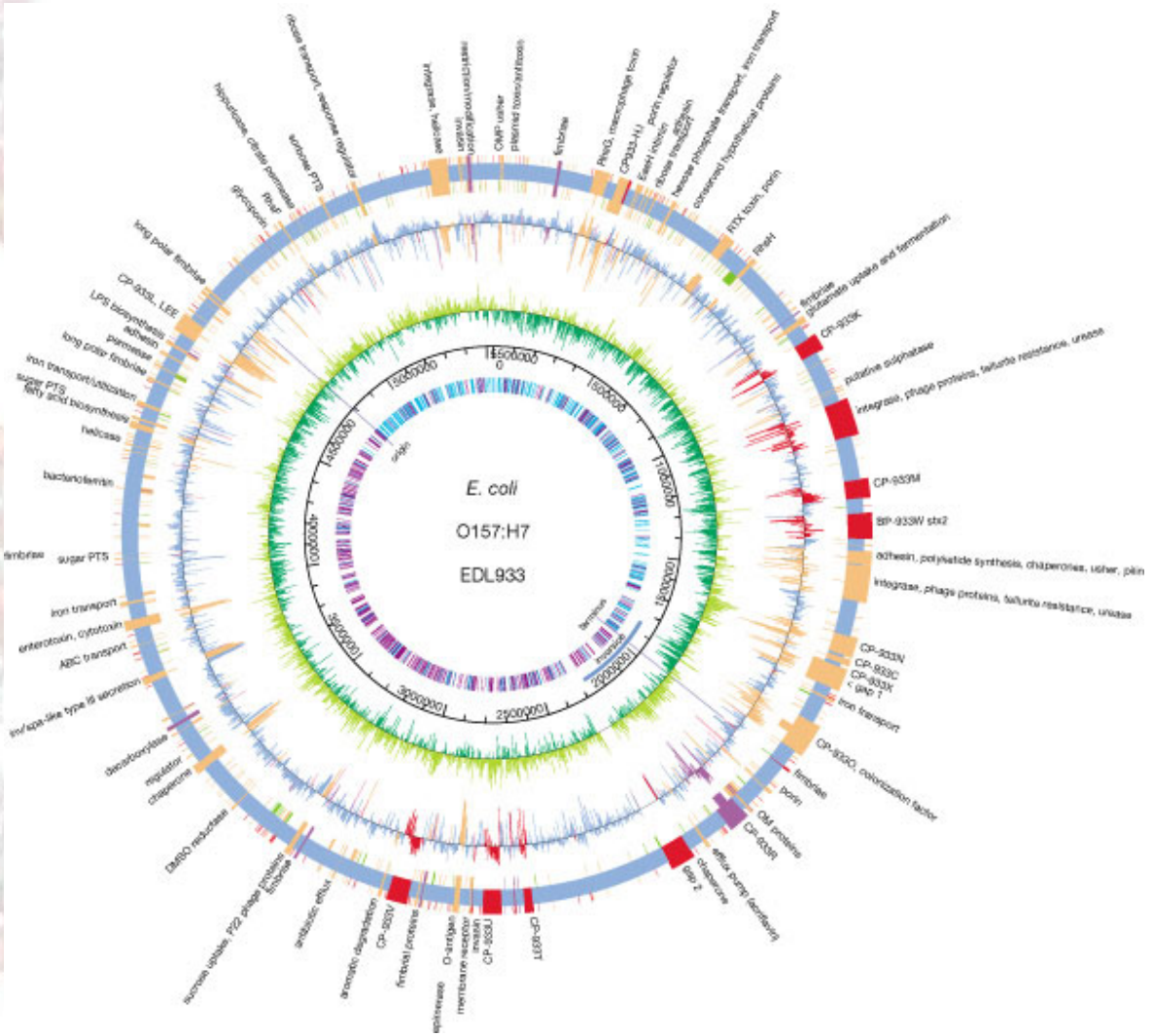
- Alinhamento de pares de sequências;
- Procura em bancos de dados de sequências;
- Alinhamento múltiplo de sequências;
- Modelagem e reconhecimento de genes;
- Modelagem e reconhecimento de “sinais”;
- Estrutura e funções de proteínas;
- Análise da Expressão de genes;
- Construção de árvores filogenéticas.

Técnicas de CS

- Algoritmos sobre Sequências
- Programação Dinâmica
- Aprendizagem por Computador
- Modelos baseados em cadeias de Markov
- Cadeias de Markov escondidas (HMM)
- Algoritmos EM
- “Clustering”
- Algoritmos sobre Árvores
- . . .

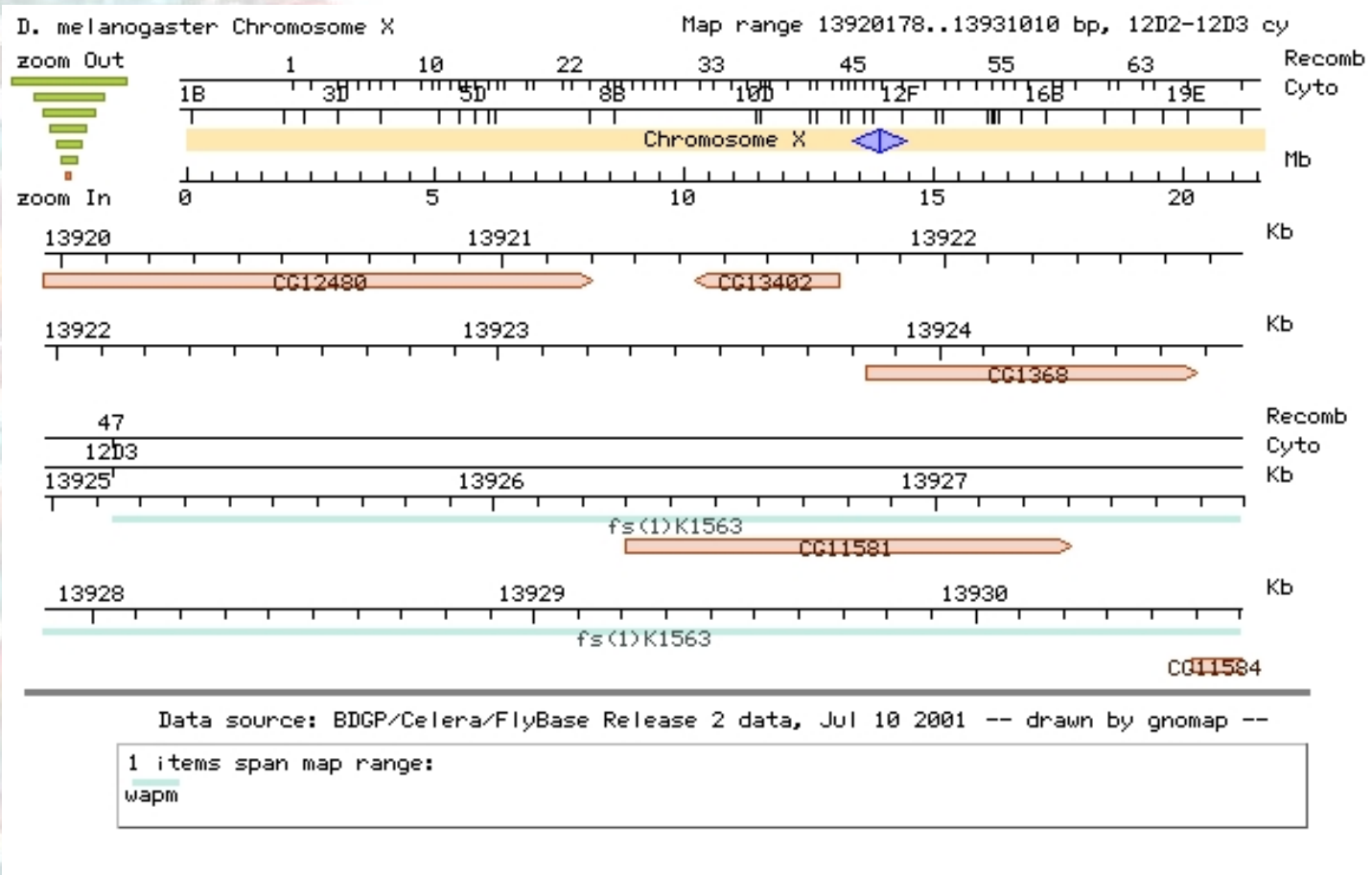
Comparar duas sequências de Genes

- Sequências
- Programação Dinâmica



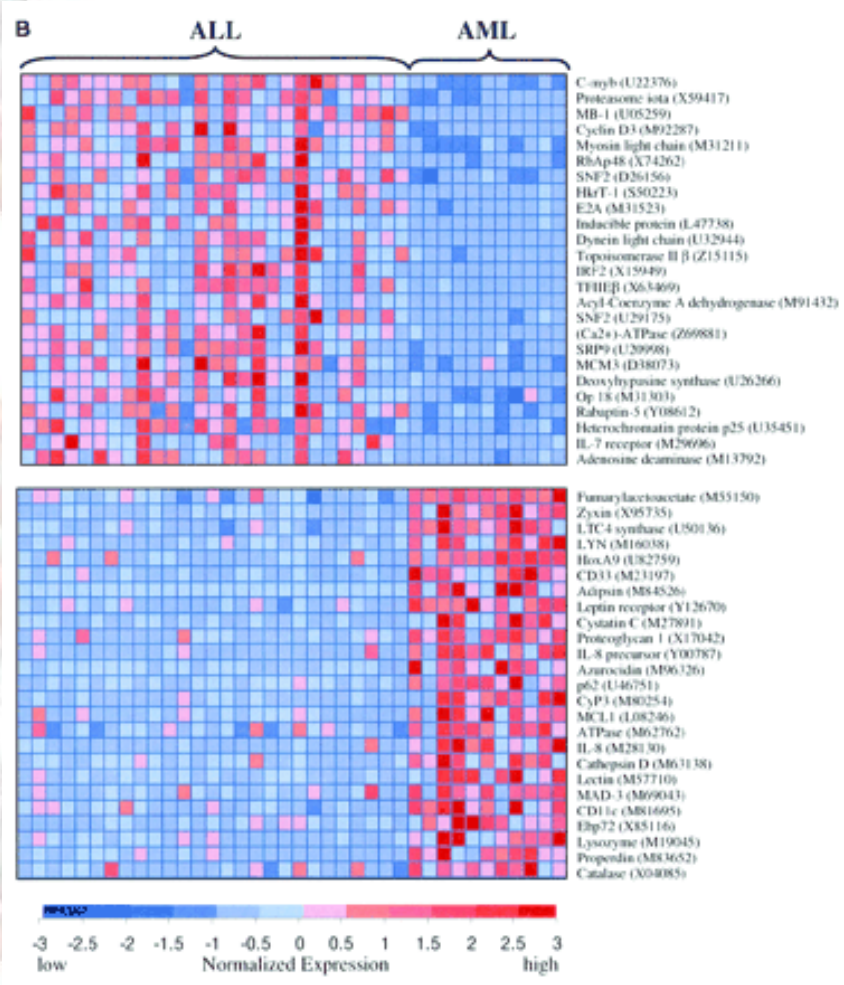
Encontrar Genes no Genoma

- Markov Models
- Hidden Markov Models



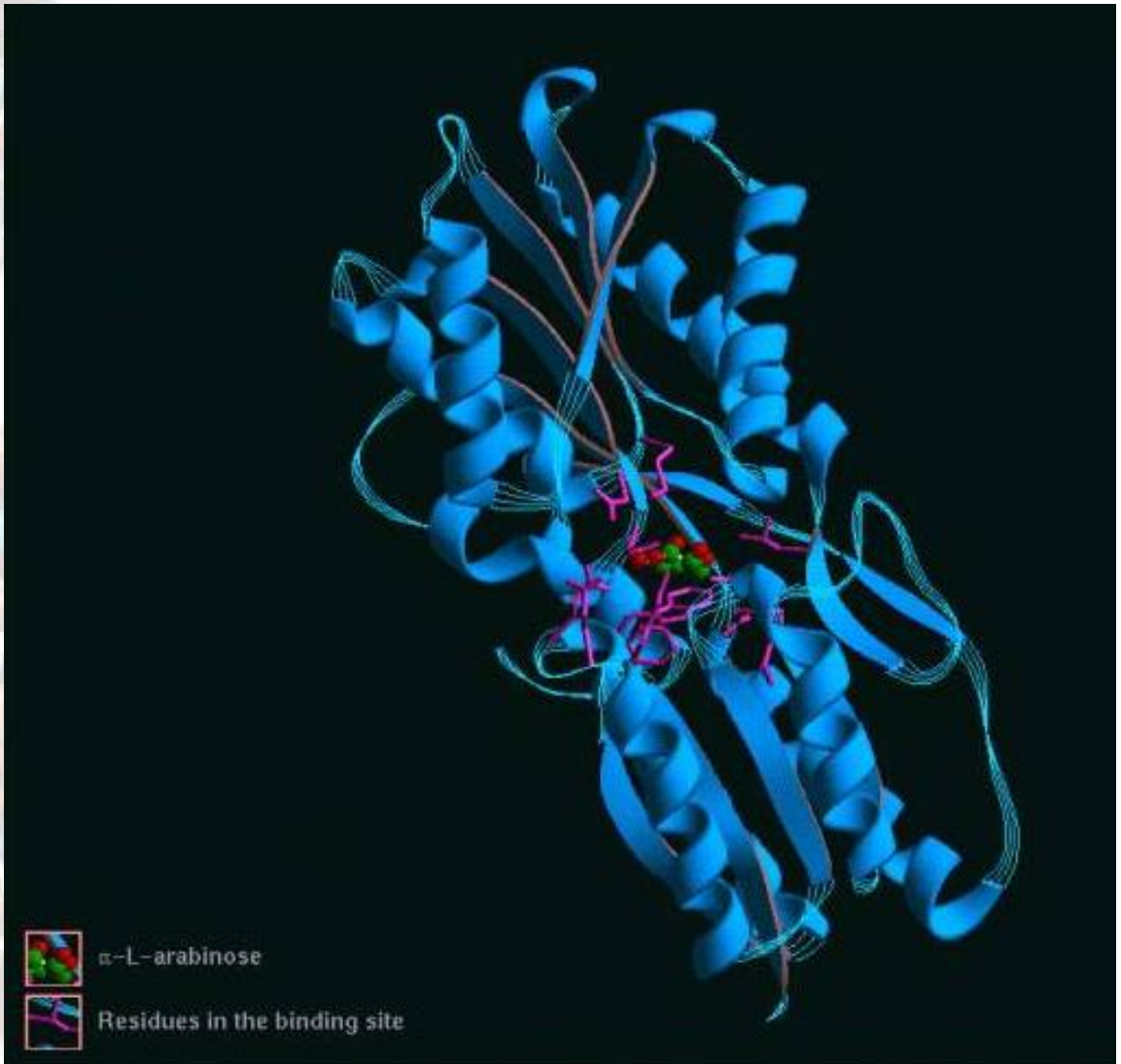
Actividade de Genes

- Clusters
- Machine Learning



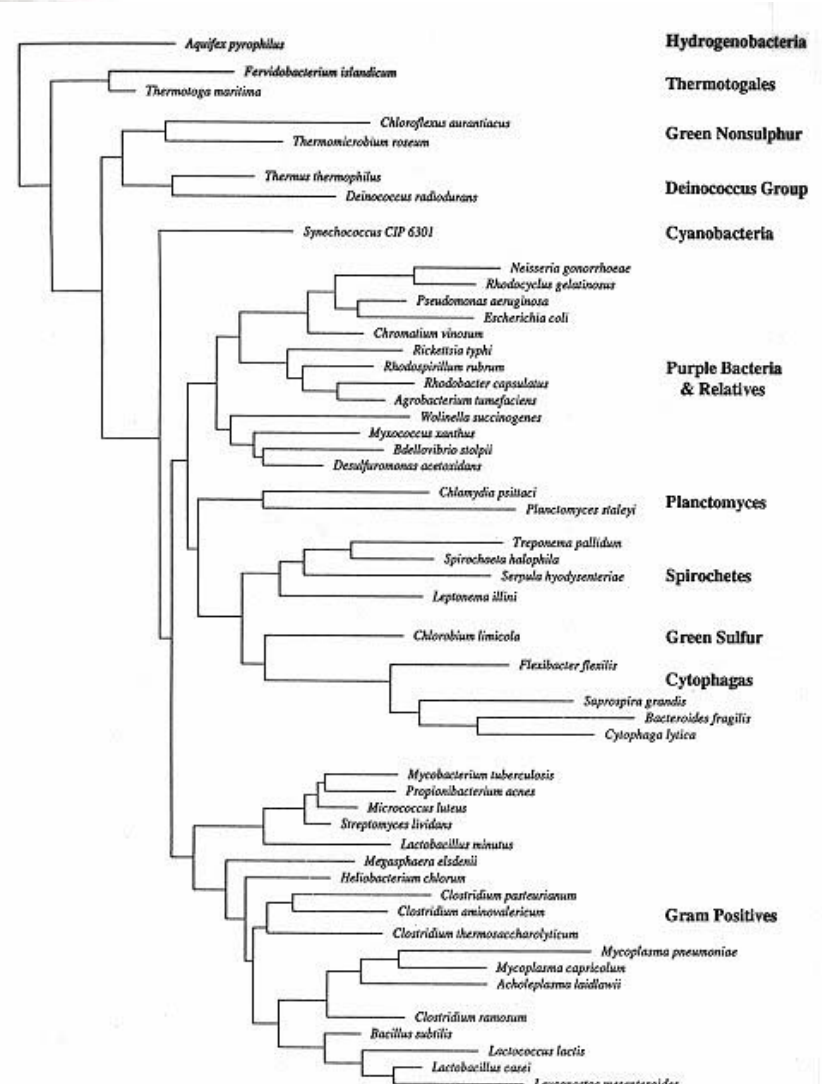
Estrutura de Proteínas?

- Programação Dinâmica
- Branch & Bound
- HMMs

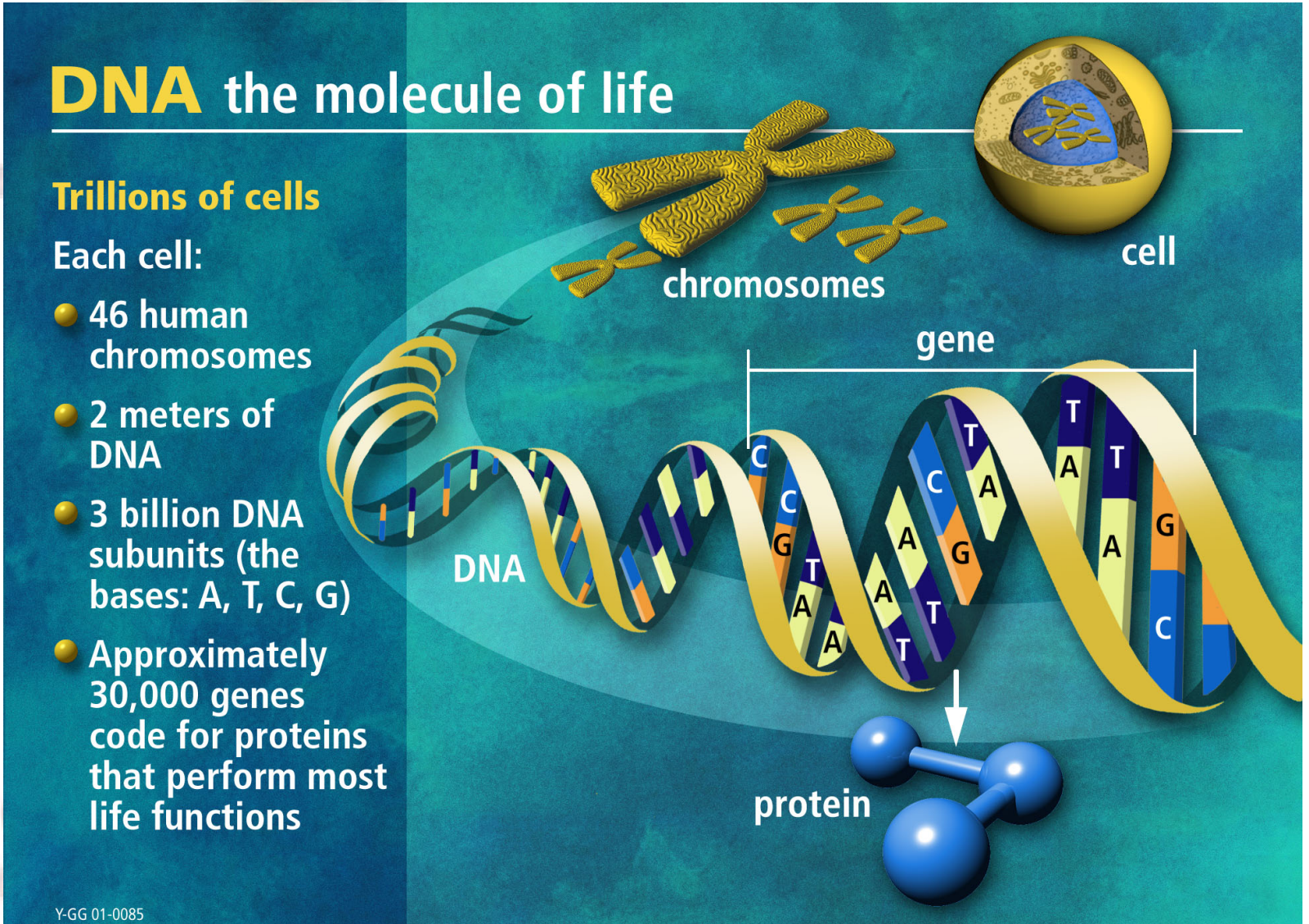


Árvores Filogenéticas

- Inferência em Árvores
- Métodos de Procura?

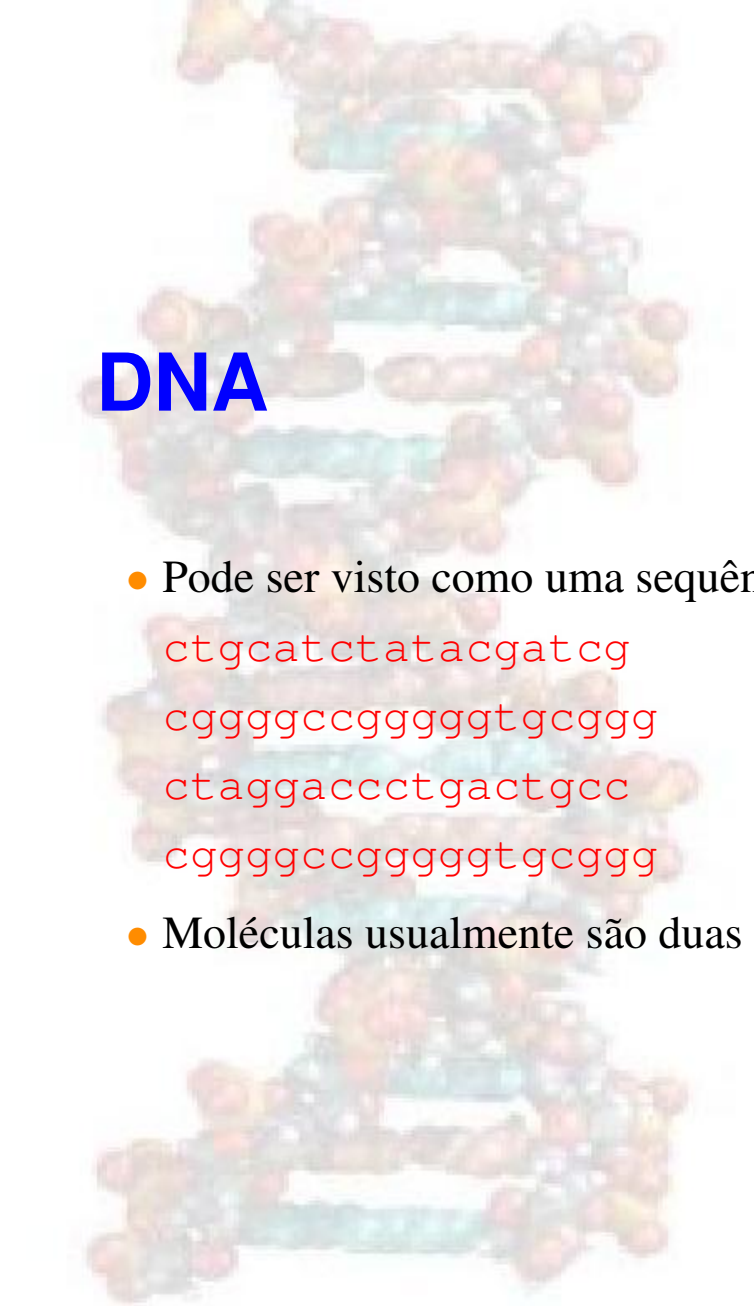


DNA


DNA the molecule of life
Trillions of cells
Each cell:
46 human chromosomes
2 meters of DNA
3 billion DNA subunits (the bases: A, T, C, G)
Approximately 30,000 genes code for proteins that perform most life functions
chromosomes
cell
gene
DNA
protein
Y-GG 01-0085

DNA

- Vista como sendo a Matriz que codifica o organismo
- Composta de pequenas moléculas chamados *nucleotídeos*
- Distinguidos por uma base:
  - **A**: adenina
  - **C**: citosina
  - **G**: guanina
  - **T**: timina
- *Polímero*: molécula enorme composta de moléculas similares



DNA

- Pode ser visto como uma sequência de 4 letras:

  ctgcatctatacgatcg

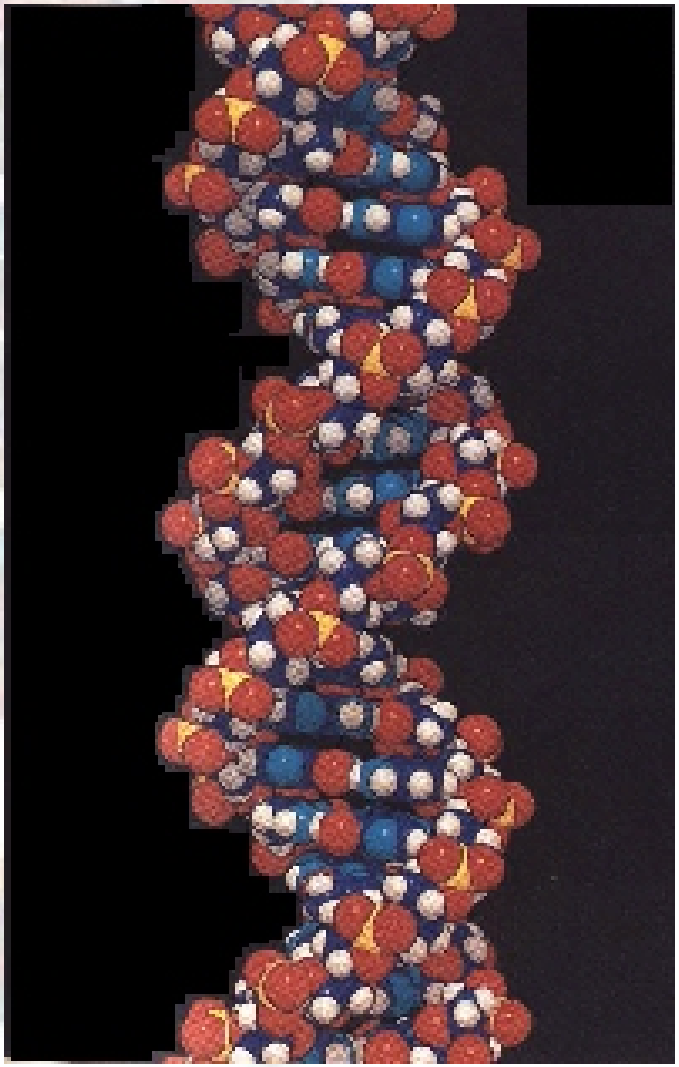
  cggggccgggggtgcggg

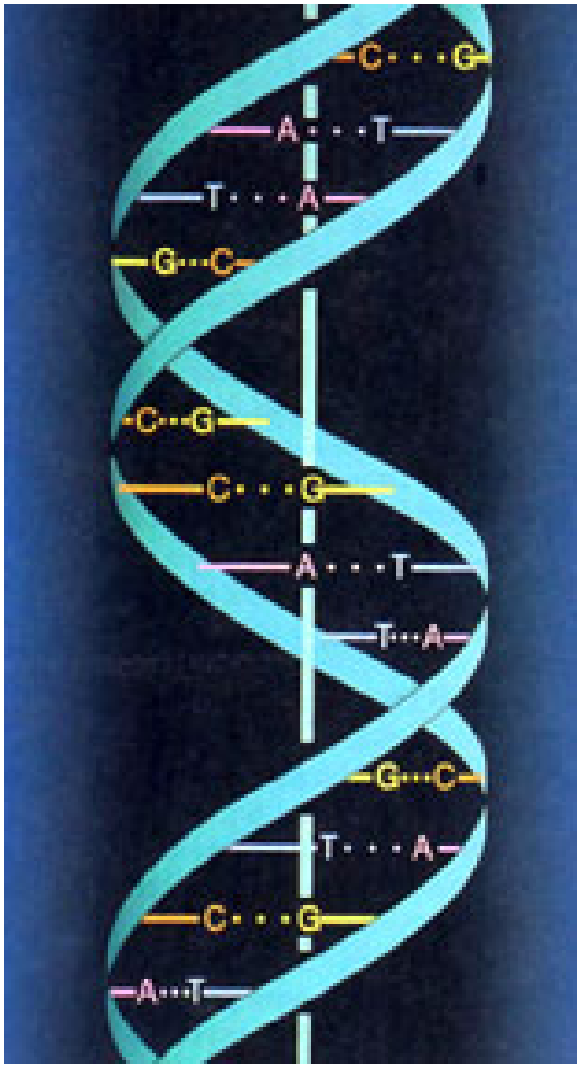
  ctaggaccctgactgcc

  cggggccgggggtgcggg

- Moléculas usualmente são duas fitas formando a famosa hélice dupla.

A Hélice Dupla



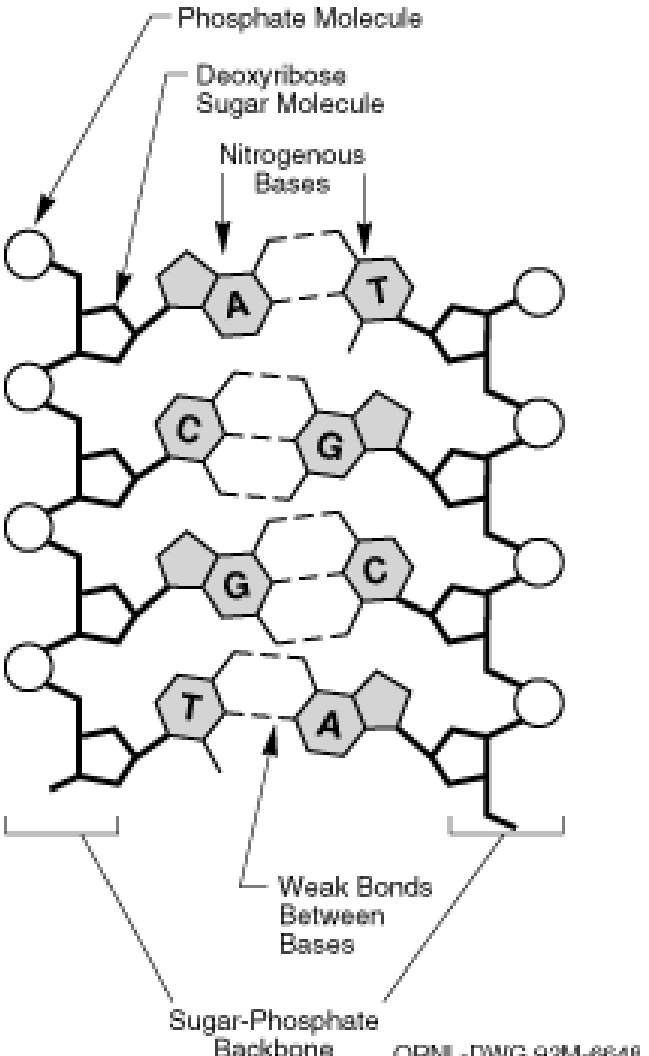


Pares de Watson-Crick

- No DNA de 2 fitas:

  ⋆ A sempre liga com T e

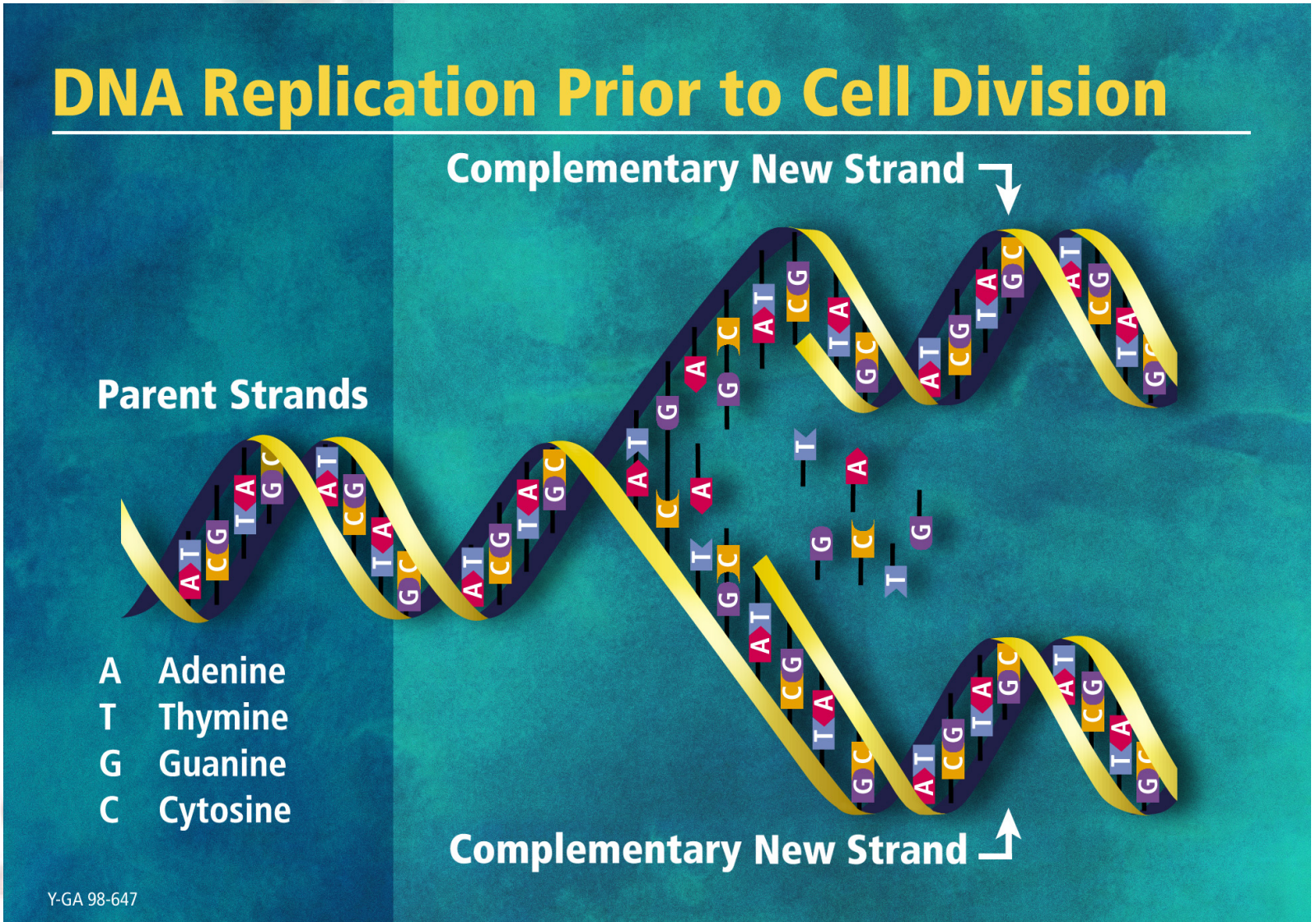
  ⋆ G sempre liga com C



A Hélice Dupla

- Cada fita de DNA tem uma "direcção":
  - ⋆ Num lado o carbono terminal da coluna está ligado ao carbono 5' do açúcar
  - ⋆ No oposto, está ligado ao carbono 3'
- Podemos portanto falar do terminal 5' e 3' de uma fita
- As fitas são *antiparalelas*

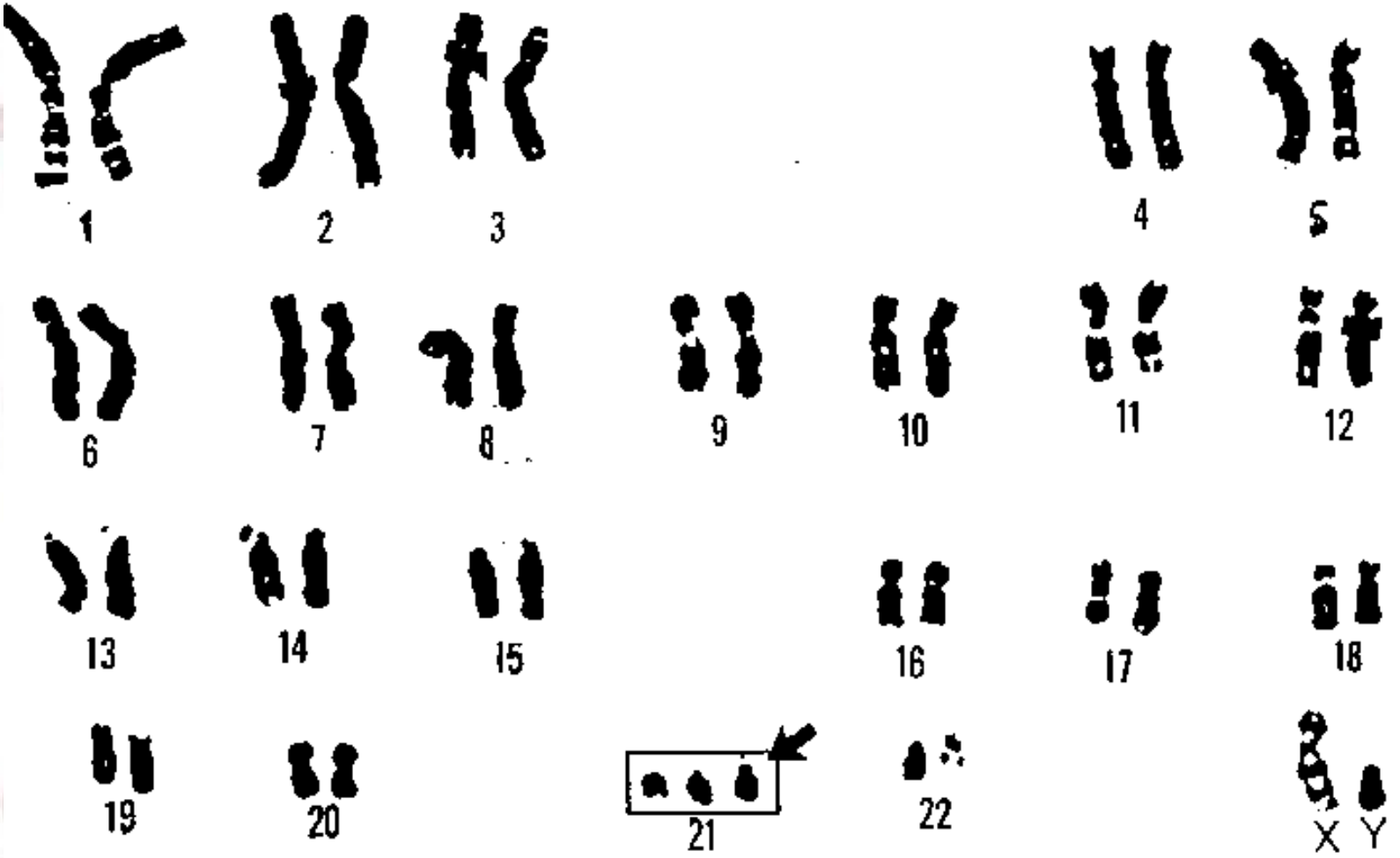
Replicação


DNA Replication Prior to Cell Division
Complementary New Strand
Parent Strands
A Adenine
T Thymine
G Guanine
C Cytosine
Complementary New Strand
Y-GA 98-647

Cromossomas

- O DNA está armazenado em cromossomas (juntamente com proteínas)
- *procariotes* são organismos uni-celulares sem núcleo e têm apenas um cromossoma circular
- *eucariotes* são organismos com núcleo e têm um número específico de cromossomas lineares.

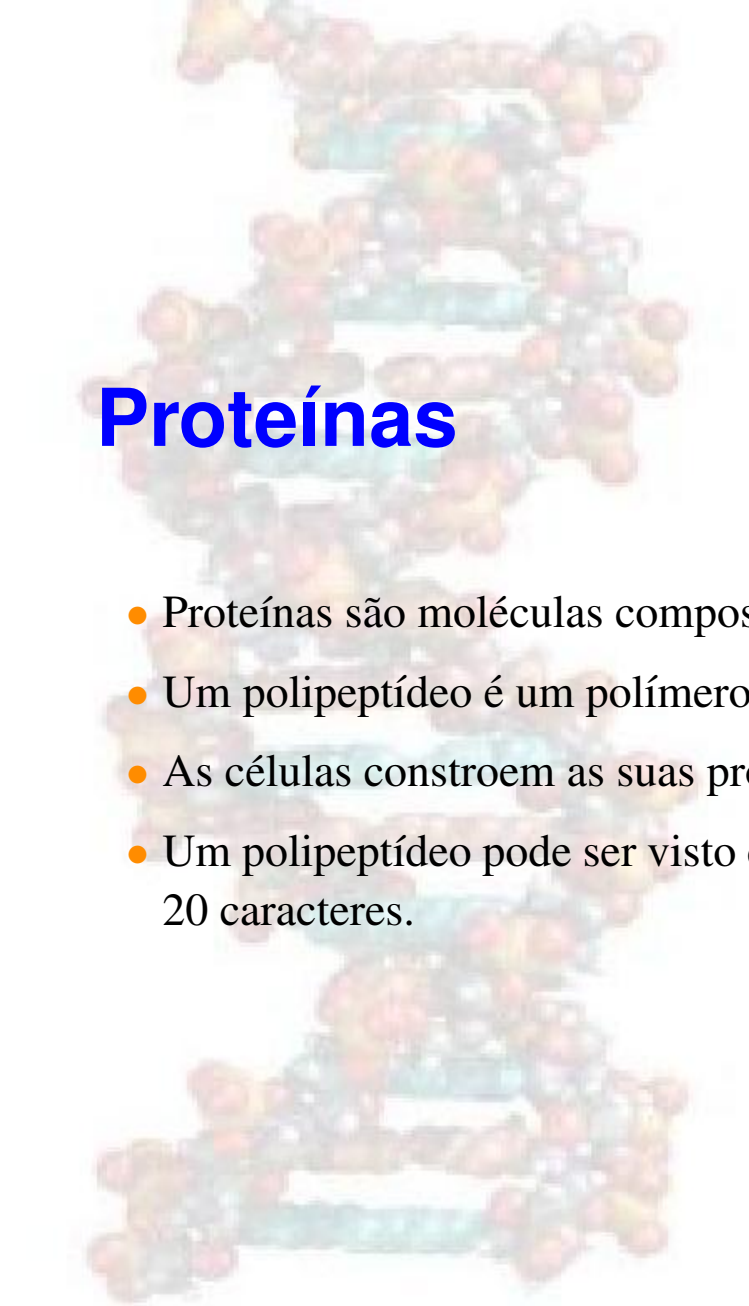
Cromossoma Humano



Genoma

O termo *genoma refere-se ao DNA completo para uma espécie*

- O ser humano tem 46 cromossomas;
- Todas as células têm o genoma completo
  - ⋆ Excepções: células sexuais e células vermelhas maturas no sangue.
- Podemos portanto falar do terminal 5’ e 3’ de uma fita
- As fitas são *antiparalelas*



Proteínas

- Proteínas são moléculas compostas de polipeptídeos;
- Um polipeptídeo é um polímero composto de amino-ácidos
- As células constroem as suas proteínas de cerca de 20 amino-ácidos diferentes
- Um polipeptídeo pode ser visto como uma sequência composta de um alfabeto com 20 caracteres.

Função das Proteínas

- Suporte Estrutural
- Armazenamento de Amino Ácidos
- Transporte de outras substâncias
- Coordenação das actividades do organismo
- Resposta ao estímulos químicos
- Movimento
- Protecção contra doenças
- Aceleração selectiva de reacções químicas

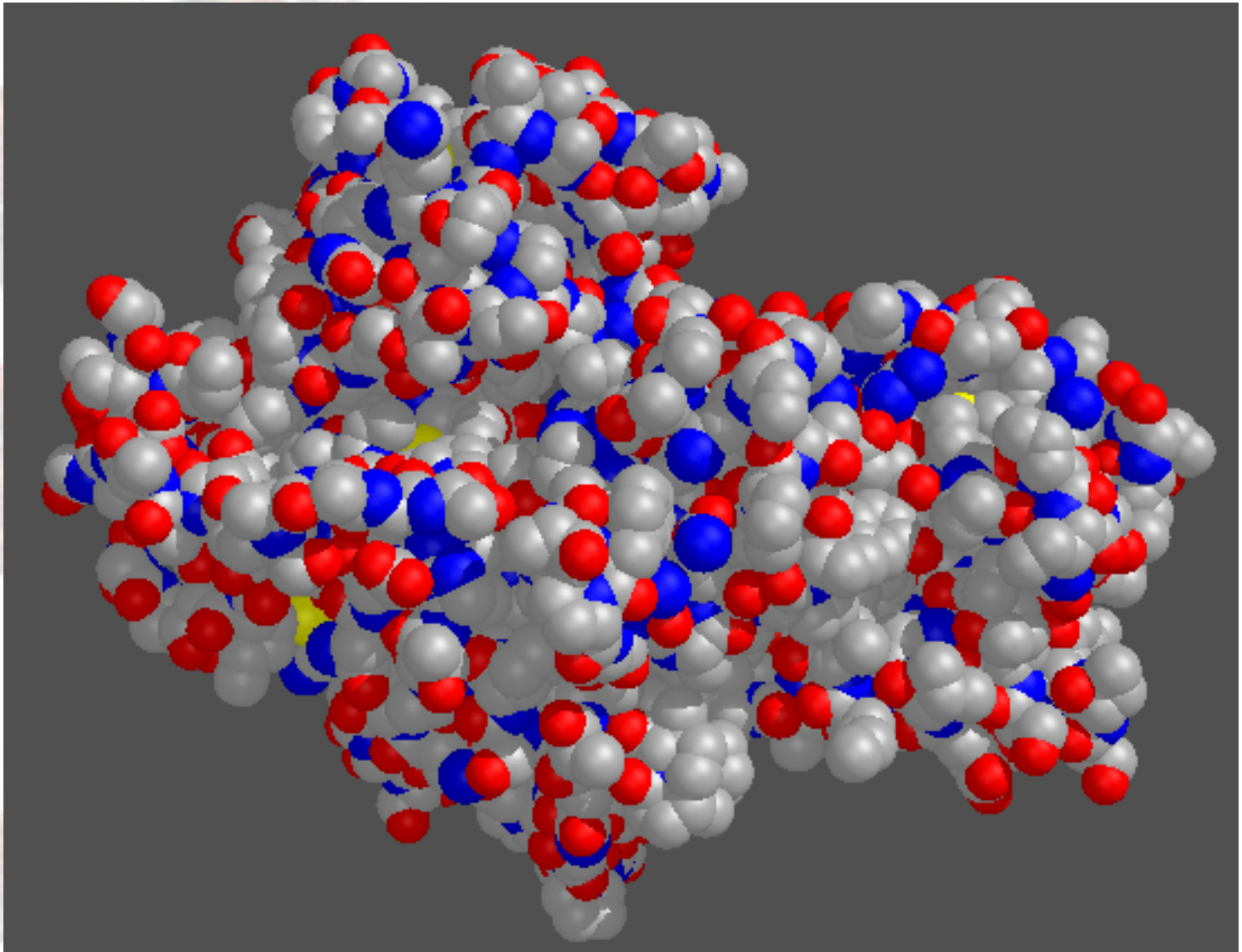
Amino-Ácidos

Alanina	Ala	**A**	Isoleucina	Ile	**I**
Arginina	Arg	**R**	Leucina	Leu	**L**
Ácido Aspártico	Asp	**D**	Licina	Lys	**K**
Asparagina	Asn	**N**	Metionina	Met	**M**
Cisteína	Cys	**C**	Prolina	Pro	**P**
Ácido Glutâmico	Clu	**E**	Serina	Ser	**S**
Fenilalanina	Phe	**F**	Treonina	Thr	**T**
Glutamina	Gln	**Q**	Triptofan	Trp	**W**
Glicina	Cly	**G**	Tirosina	Tyr	**Y**
Histina	His	**H**	Valina	Val	**V**

Hexokinase

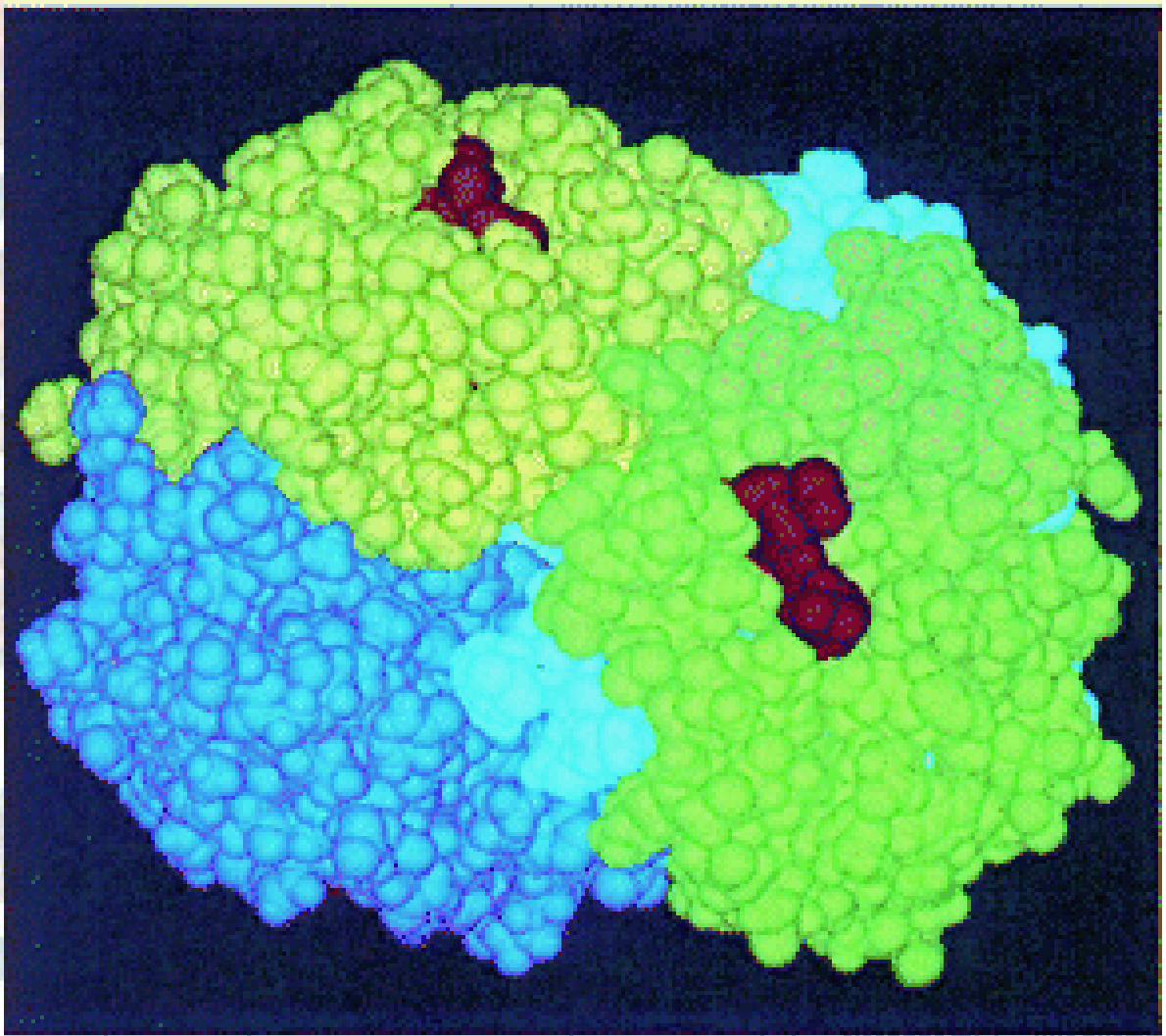
```
          5         10        15        20        25        30
  1 A A S X D X S L V E V H X X V F I V P P X I L Q A V V S I A
 31 T T R X D D X D S A A A S I P M V P G W V L K Q V X G S Q A
 61 G S F L A I V M G G G D L E V I L I X L A G Y Q E S S I X A
 91 S R S L A A S M X T T A I P S D L W G N X A X S N A A F S S
121 X E F S S X A G S V P L G F T F X E A G A K E X V I K G Q I
151 T X Q A X A F S L A X L X K L I S A M X N A X F P A G D X X
181 X X V A D I X D S H G I L X X V N Y T D A X I K M G I I F G
211 S G V N A A Y W C D S T X I A D A A D A G X X G G A G X M X
241 V C C X Q D S F R K A F P S L P Q I X Y X X T L N X X S P X
271 A X K T F E K N S X A K N X G Q S L R D V L M X Y K X X G Q
301 X H X X X A X D F X A A N V E N S S Y P A K I Q K L P H F D
331 L R X X X D L F X G D Q G I A X K T X M K X V V R R X L F L
361 I A A Y A F R L V V C X I X A I C Q K K G Y S S G H I A A X
391 G S X R D Y S G F S X N S A T X N X N I Y G W P Q S A X X S
421 K P I X I T P A I D G E G A A X X V I X S I A S S Q X X X A
451 X X S A X X A
```

Hexokinase: Modelo Espacial



Hemoglobina

Construída com 4 Polipeptídeos:



Genes

Genes são a unidade básica de hereditariedade:

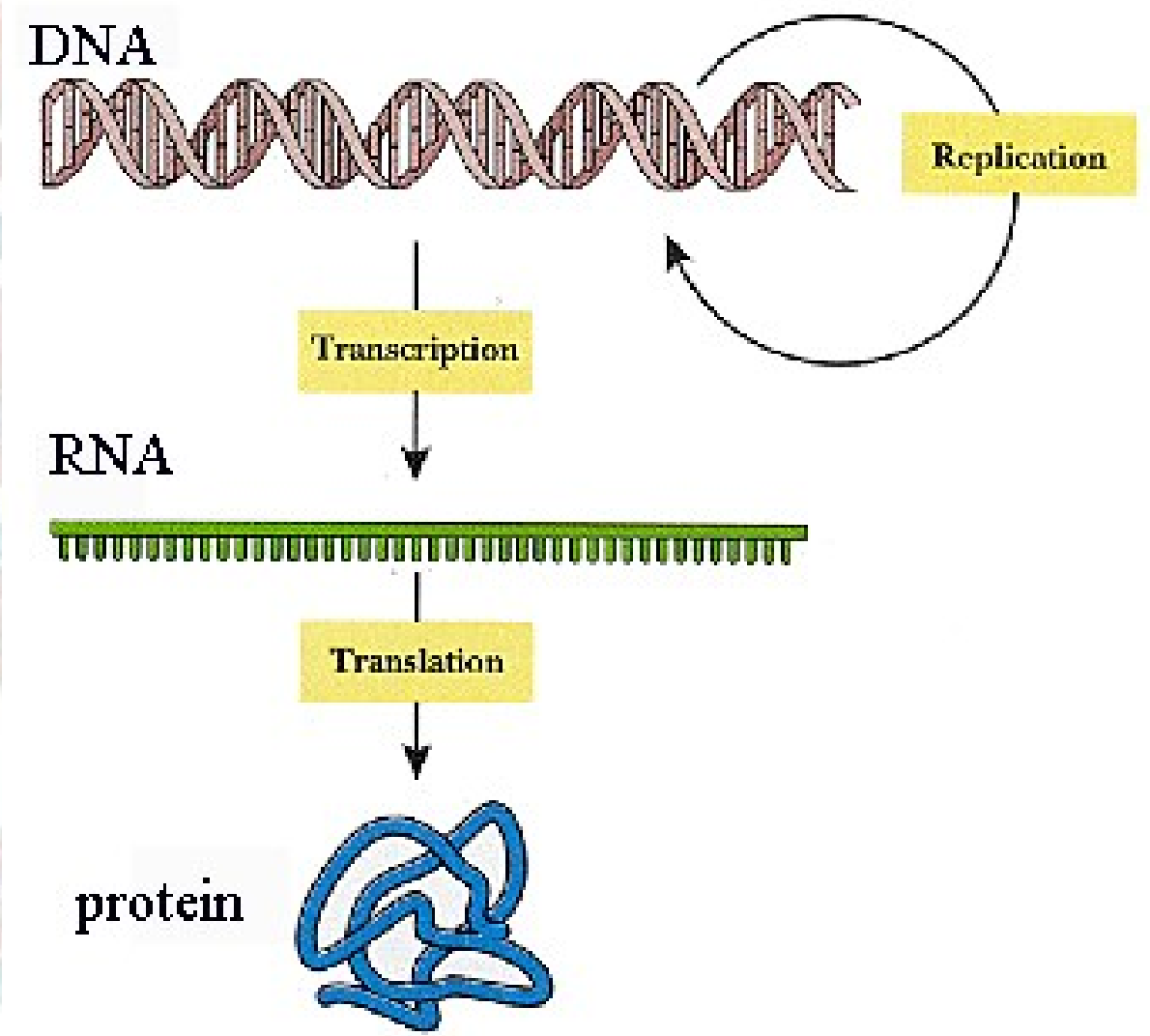
- sequência de base que carrega a informação necessária para construir uma certa proteína (polipeptídeo)
- diz-se que genes *codificam* proteínas
- estimativa: o nosso genoma tem cerca de 4000 genes

Densidade de Genes

Nem todo o DNA no genoma codifica proteínas:

```
micróbios 90% codificação   gene/kb
humanos    3% codificação gene/35kb
```

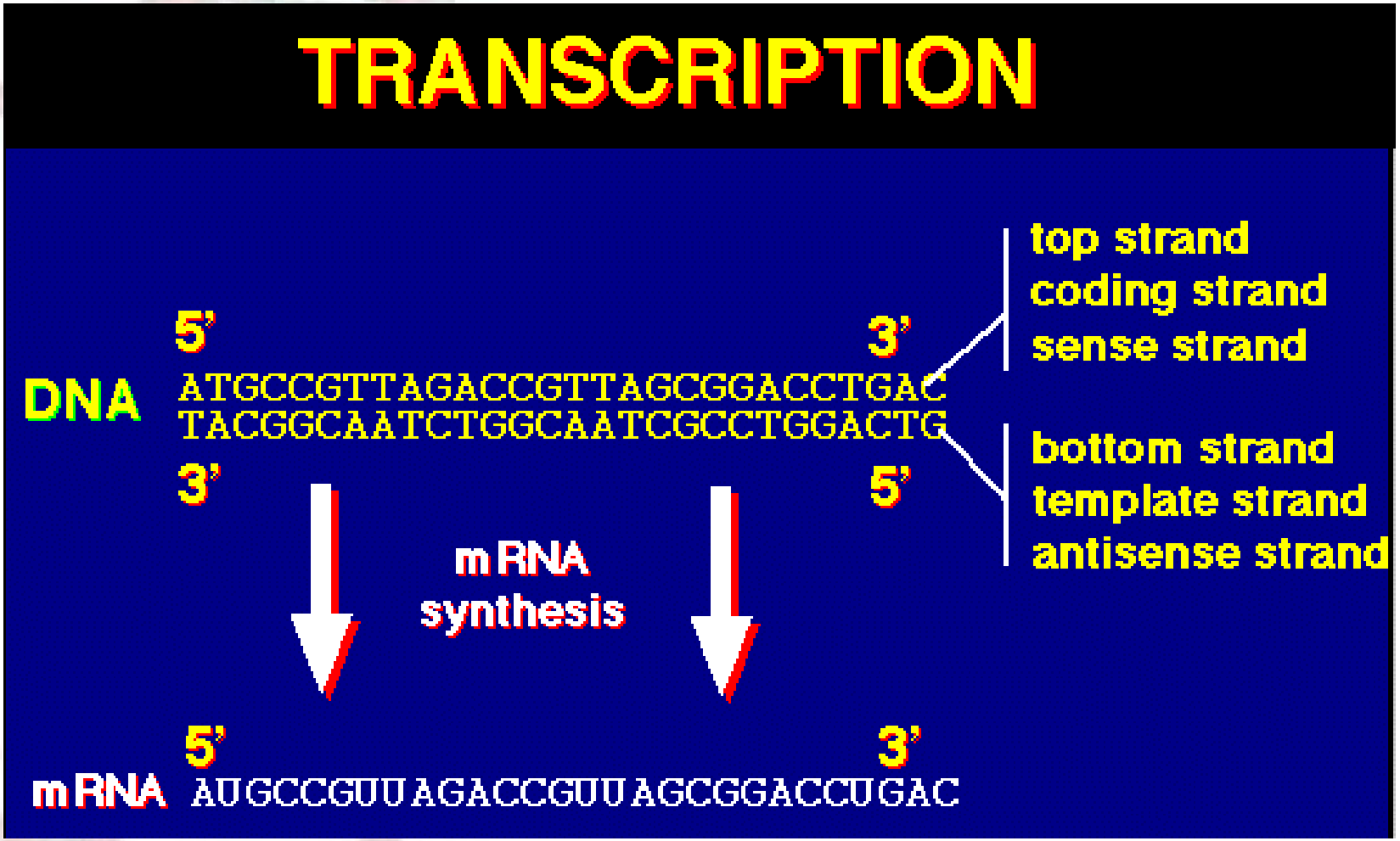
O Dogma Central



RNA

- RNA é como DNA excepto que:
  - ⋆ coluna um pouco diferente;
  - ⋆ habitualmente apenas uma fita
  - ⋆ usa uracilo (U) em vez de timina (T)
- Uma fita de RNA pode ser vista como uma sequência formada com 4 letras: A, C, G, U.

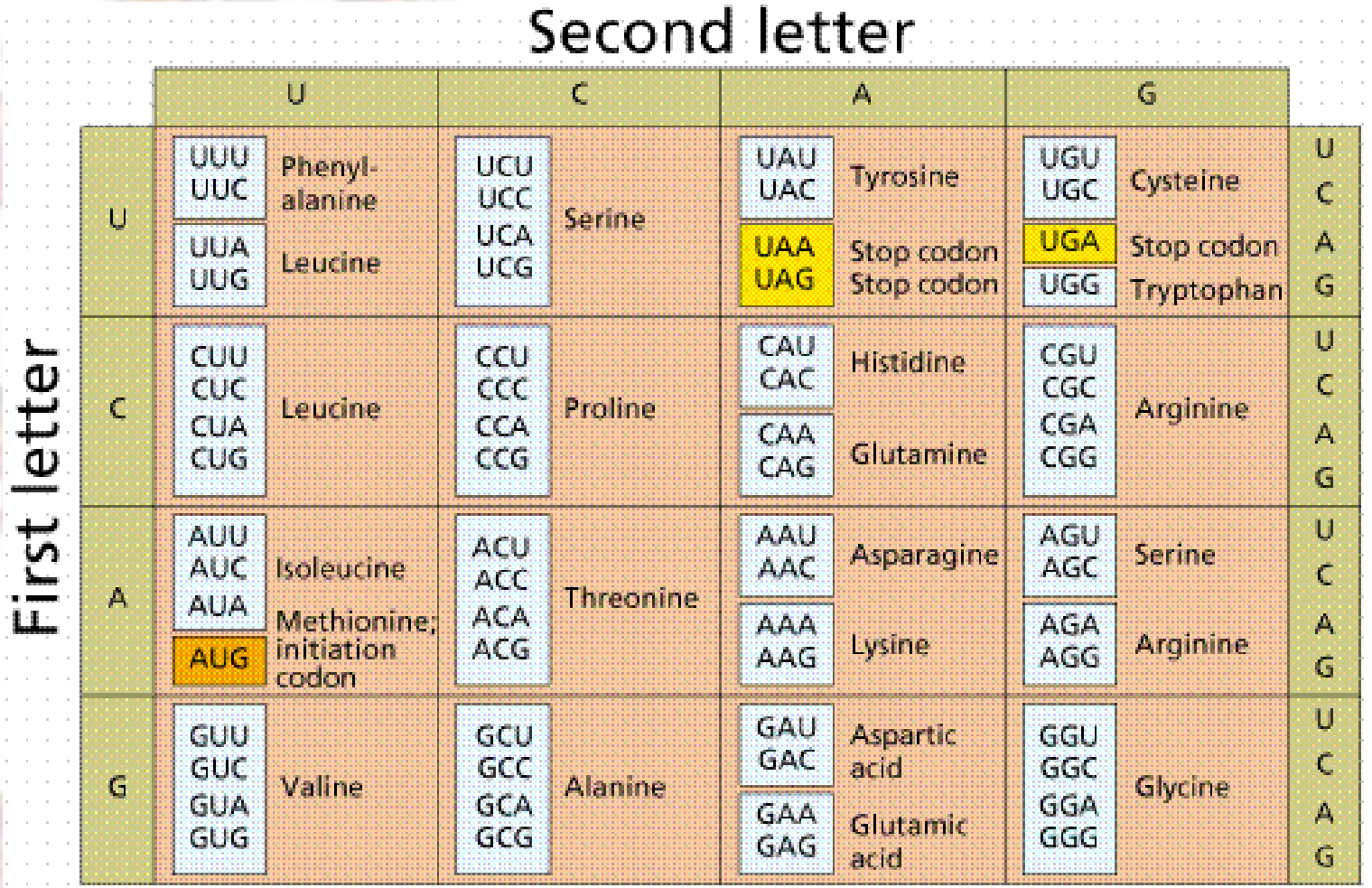
Translado


TRANSCRIPTION
5'
3'
DNA
ATGCCGTTAGACCGTTAGCGGACCTGAC
TACGGCAATCTGGCAATCGCCTGGACTG
3'
5'
top strand
coding strand
sense strand
bottom strand
template strand
antisense strand
mRNA
synthesis
5'
3'
mRNA
AUGCCGUUAGACCGUUAGCGGACCUGAC

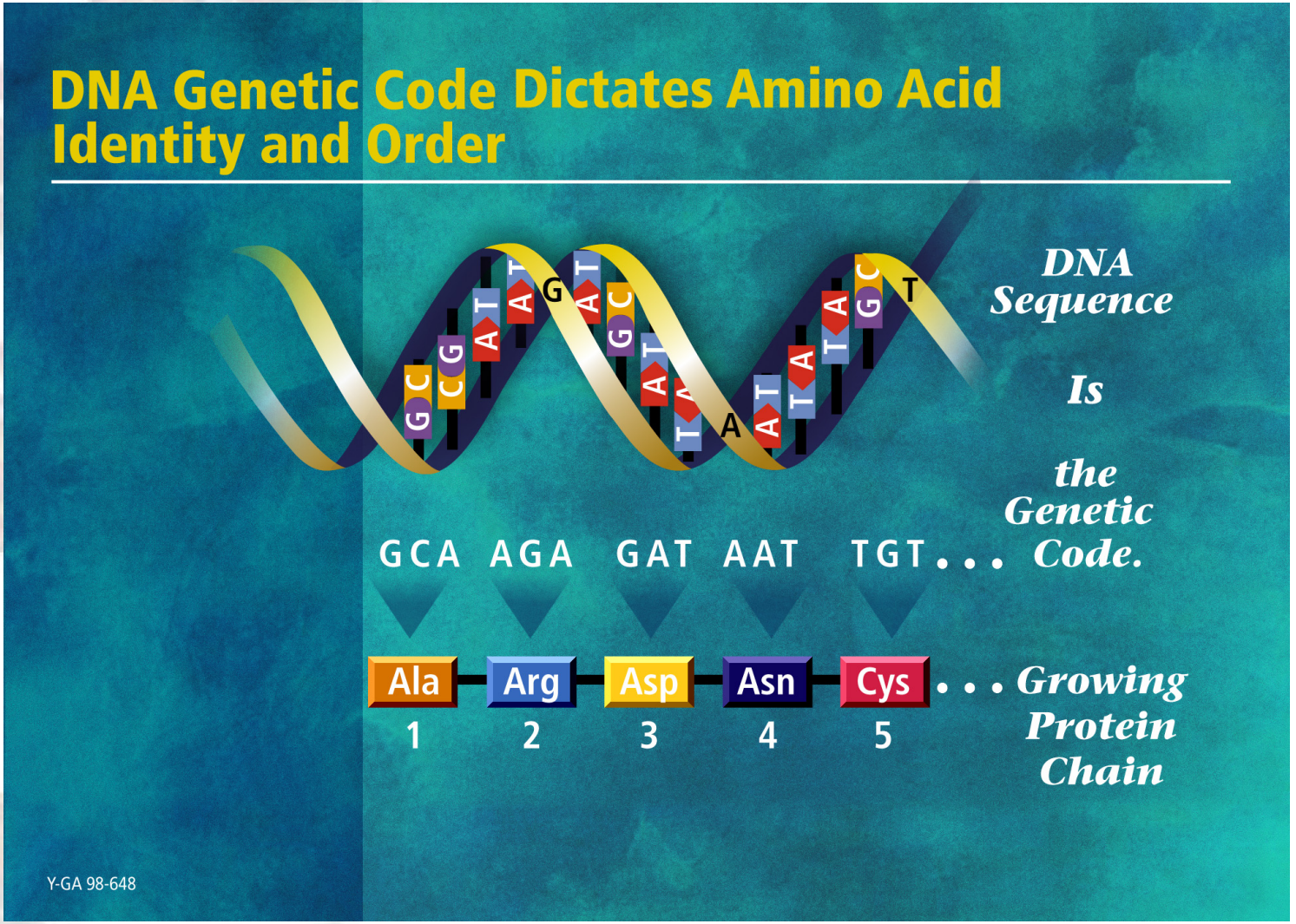
Translado

- RNA Polimerase é o enzima que constrói uma fita de RNA a partir de um gene.
- O RNA que é transcrito é chamado de RNA mensageiro: RNA-m.
- Existem mais variedades de RNA.

O Código Genético



O Código Genético e Proteínas



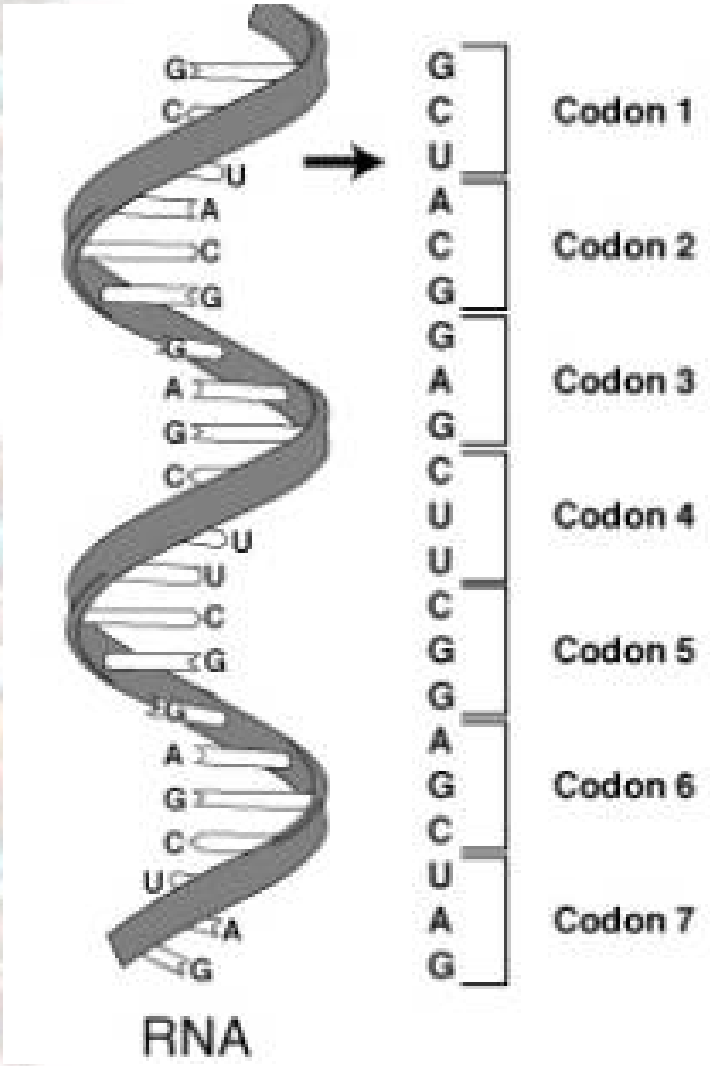
Tradução

- Ribossomas são as máquinas que sintetizam proteínas a partir do mRNA;
- Um grupo de codões é chamado de quadro de leitura (“reading frame”):

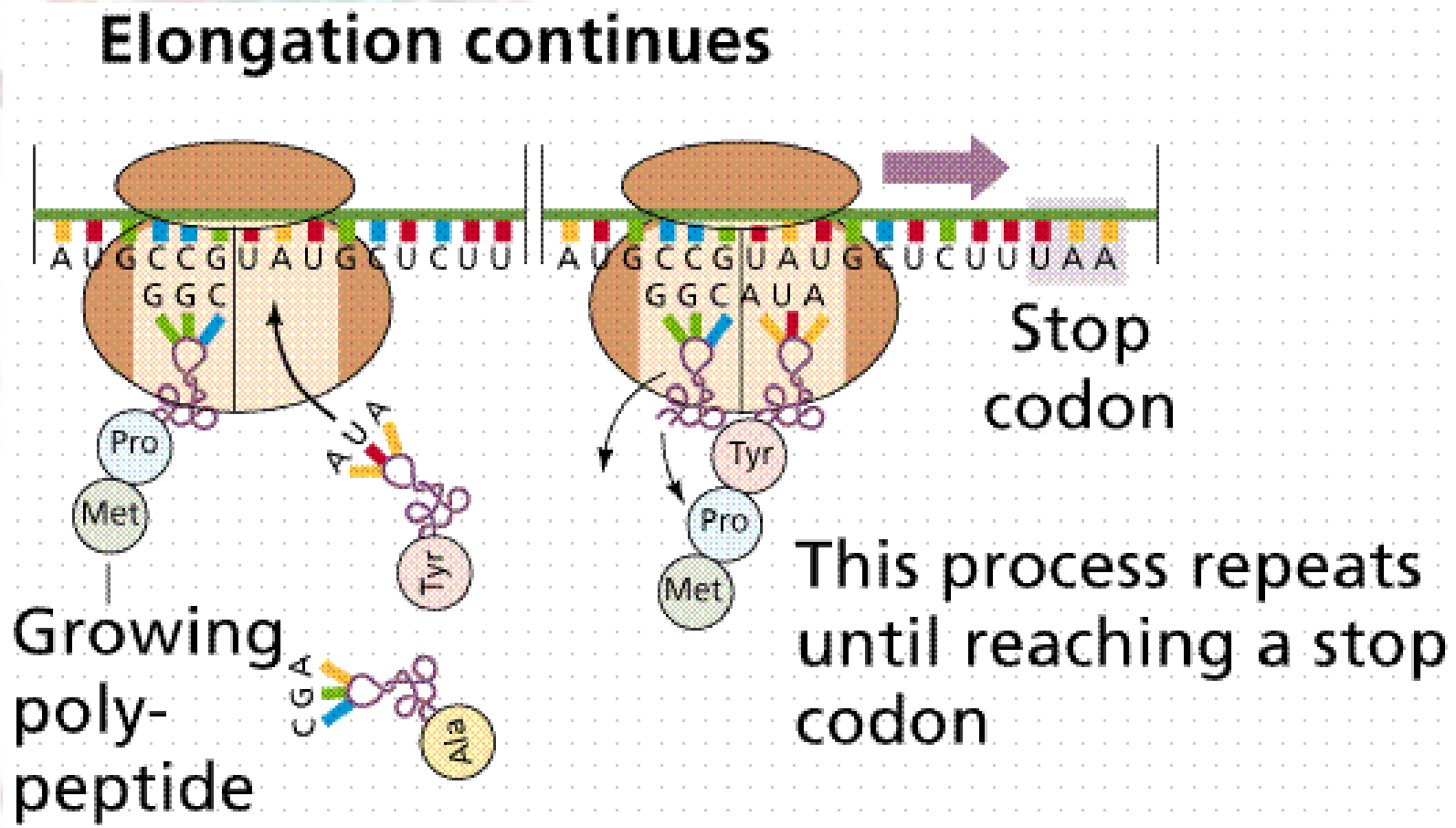
Fita de DNA	A	C	G	C	A	G	A	T	A	T	C	A	T	G	A
	A	C	G	C	A	G	A	T	A	T	C	A	T	G	A
	A	C	G	C	A	G	A	T	A	T	C	A	T	G	A
	A	C	G	C	A	G	A	T	A	T	C	A	T	G	A

- a tradução começa com o “start codon”
- a tradução termina com o “stop codon”

Codons e Quadros de Leitura



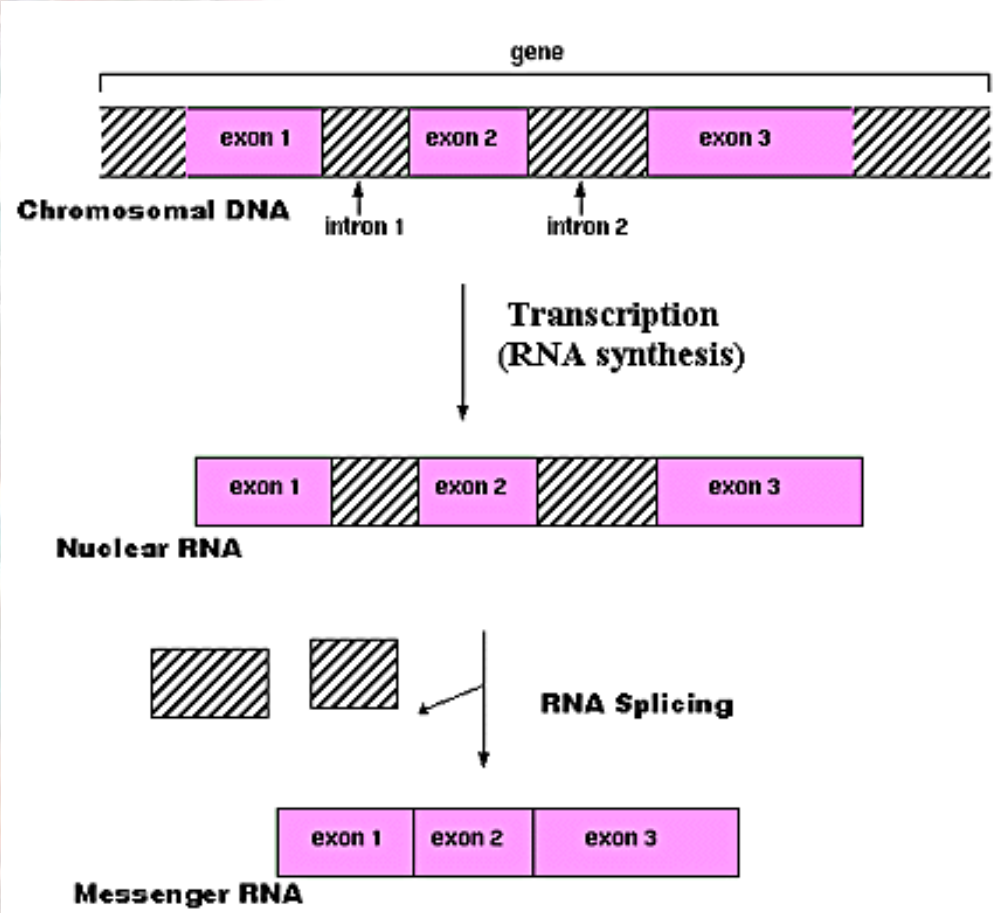
Tradução


Elongation continues
A U G C C G U A U G C U C U U
G G C
Pro
Met
A U A
Tyr
C G A
Ala
Growing poly-peptide
A U G C C G U A U G C U C U U U A A
G G C A U A
Tyr
Pro
Met
Stop codon
This process repeats until reaching a stop codon

Processamento do RNA nos eucariotes

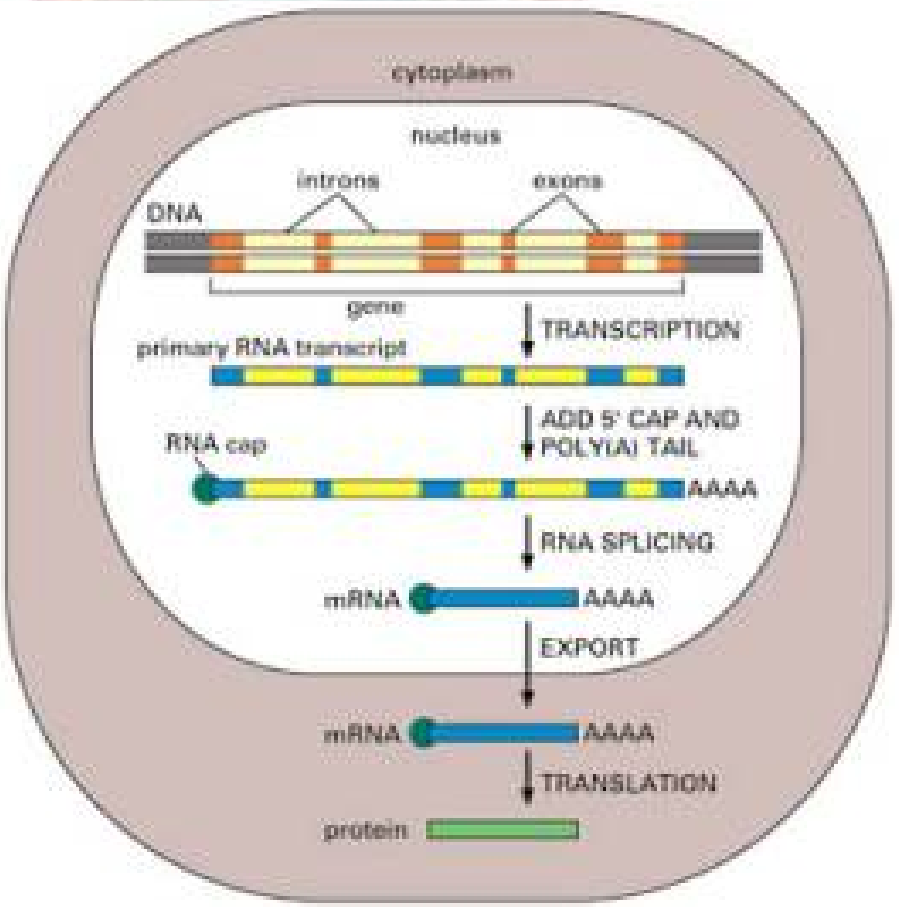
- Eucariotes são organismos que têm núcleos fechados nas suas células
- Nos eucariotes, o mRNA consiste de segmentos alternados de exons e introns:
  - ⋆ os exons são as componentes responsáveis por codificação
  - ⋆ os introns são removidos antes da tradução

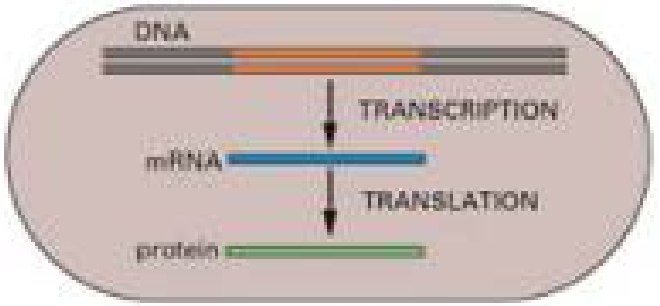
Remoção do DNA



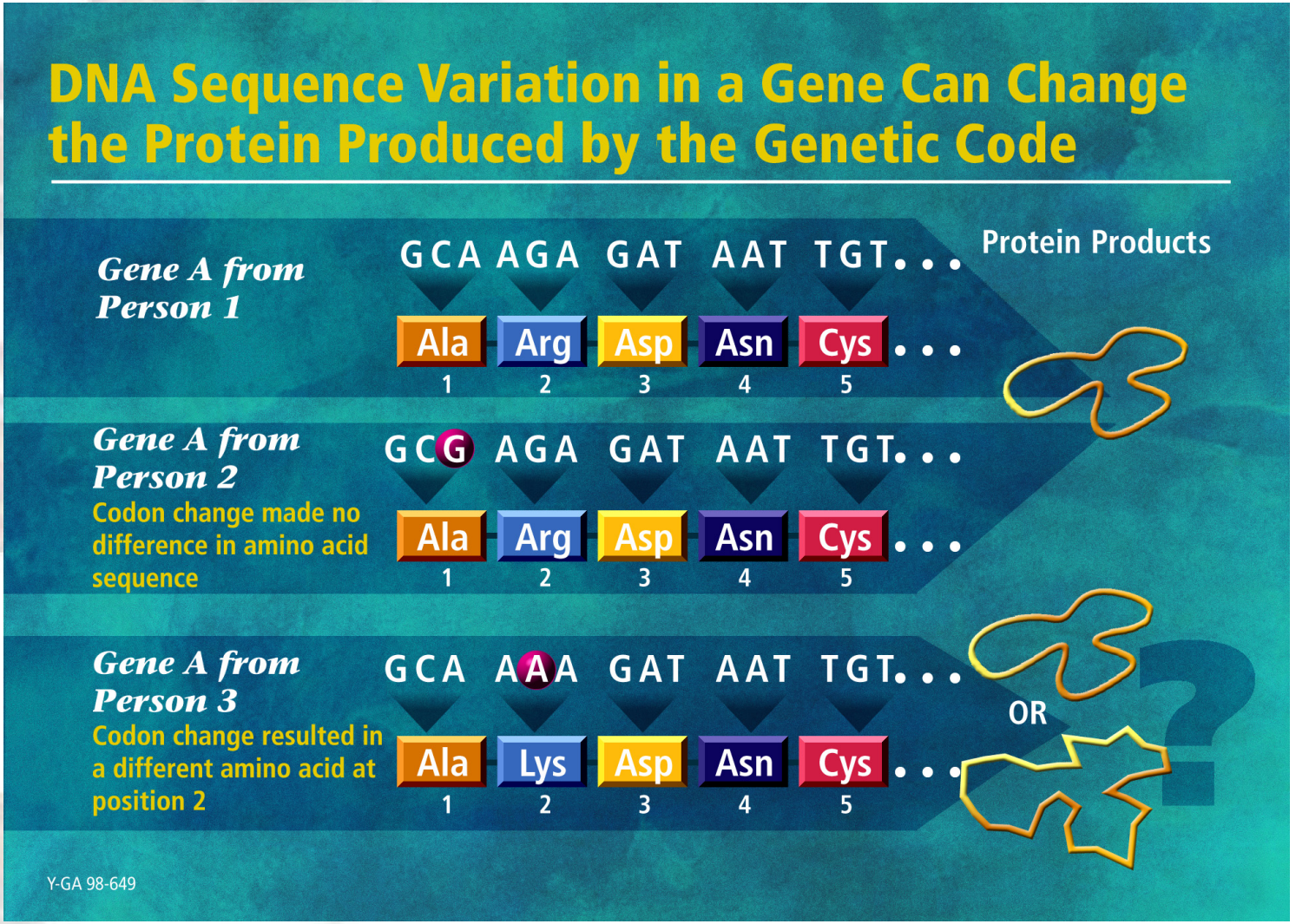
RNA synthesis and processing

Síntese de Proteínas





Variação do DNA



Genomas Completamente Publicados

Tipo	Número Aproximado
Archaea	16
Bacteria	96
Eucariotes	17

- dados de http://wit.integratedgenomics.com/GOLD/
- Não conta vírus, fagos, etc.
- Um encontrado no Brasil: bactéria Xanthomonas axonopodis pv. citri
- Organismos multicelulares: rato, fungos, e homo sapiens.
- Em progresso:
  - ⋆ 345 procariotes
  - ⋆ 235 eucariotes

Os Grandes Sucessos do Genoma

Genoma	One	Ano
H. Influenza	TIGR	1995
E. Coli K-12	Wisconsin	1997
S. Cerevisiae (fermento)	colab. interna.	1997
C. Elegans (verme)	Washington U./Sanger	1998
Drosophila M. (mosca da fruta)	multiple groups	2000
E. Coli 0157:H7 (patogeneo)	Wisconsin	2000
H. Sapiens	colaboração internacional/Celera	2001

Tamanhos de Alguns Genomas

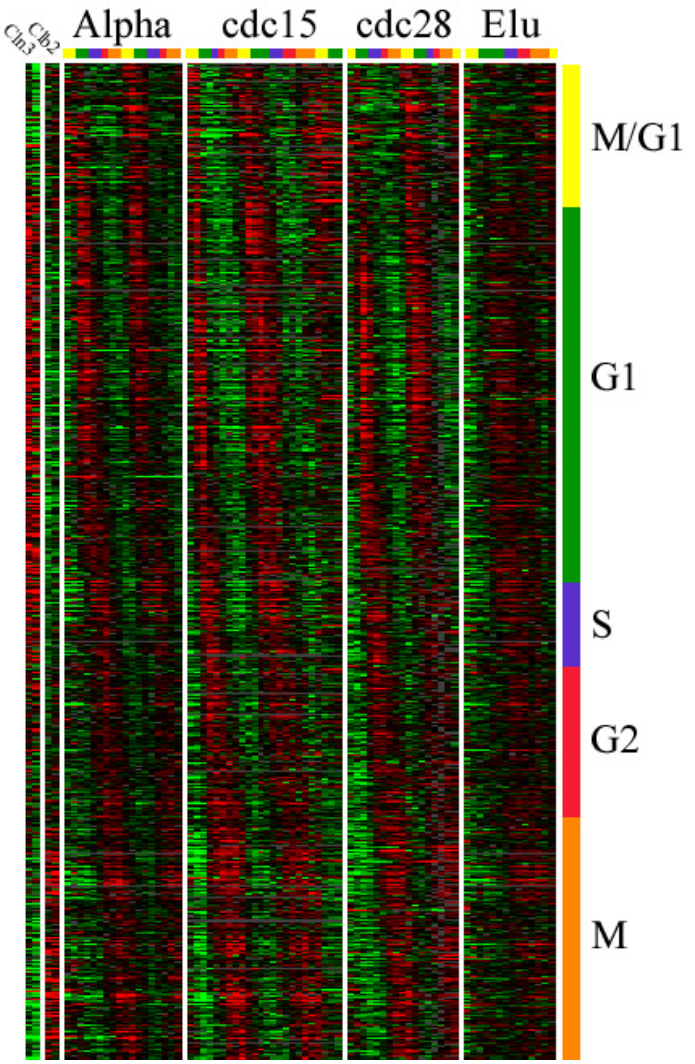
Genoma	#bps
HIV	9750
E. coli	4.6 milhões
S. cerevisiae	12 milhões
C. elegans	97 milhões
Drosophila M.	137 milhões
human	3.1 billion

Há Mais

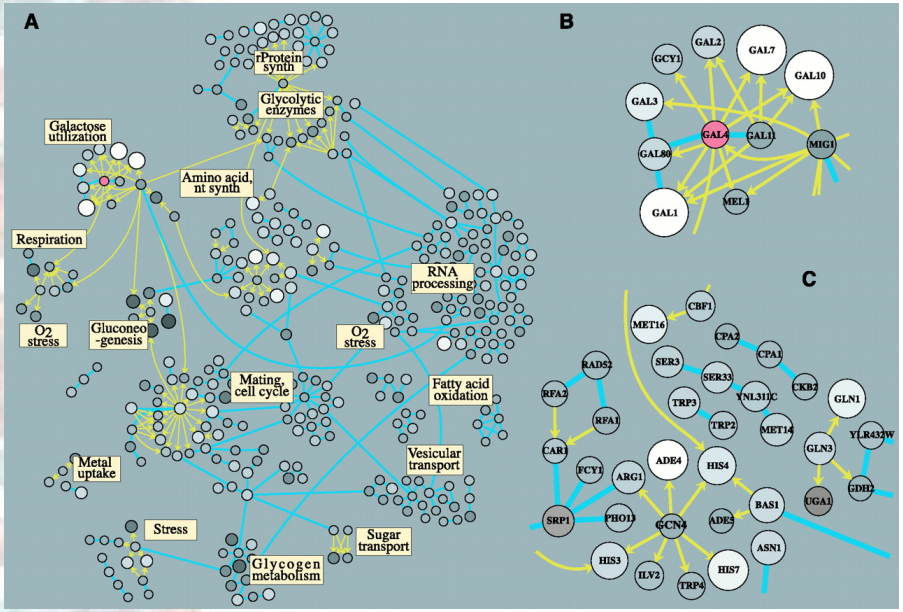
- > 300 outros bancos de dados sobre biologia nuclear.
- GenBank (Feb 2006):
  - 12.465.546 sequências
  - 59.750.386.305 bases
- SWISS-PROT (7.4):
  - 215.741 entradas com sequências de proteínas
  - 79.098.200 amino-ácidos
- Protein Data Bank (Abril 06):
  - 35.917 proteinas e estruturas relacionadas.

Dados Sobre a Expressão de Genes

- a figure mostra a expressão de um gene de fermento:
  - ⋆ cada linha é um gene
  - ⋆ coluna representa medida da expressão de genes em certa altura
- vermelho: acima de um certo nível
- azul: abaixo de um certo nível



Interacções



- cada nó representa o produto de um gene (proteína)
- linhas azuis representam interacções directas entre proteínas
- linhas amarelas mostram interacções em que uma proteína associa a DNA e altera a expressão de outra.

Significado da Revolução Genómica

- Biologia baseada em dados:
  - ⋆ genómicas funcional
  - ⋆ genómicas comparadas
  - ⋆ biologia de sistemas
- Medicina Molecular:
  - ⋆ Identificação de componentes genéticos de várias doenças
  - ⋆ diagnose/prognose a partir de sequências/expressões
  - ⋆ terapia com genes
- Farmacogenómicas:
  - ⋆ Desenvolver drogas altamente especializada
- Toxicogenómicas:
  - ⋆ Que genes são afectadas por que agentes químicos.

Bioinformática Revisitada

Representação/Armazenamento/Recuperação/Análise de dados biológicos sobre sequências (DNA, protein)

- estruturas (proteínas)
- funções (proteínas, sinais de sequências)
- níveis de actividade (mRNA, proteínas)
- redes de interacções (caminhos metabólicos, caminhos regulatórios, caminhos de sinalização)

de/entre biomoléculas