Bioinformática

2023-2024

Docente

- **Paulo Martel**

Gabinete: FCT, Edifício C8, 3.12

Email: pmartel@ualg.pt

Homepage: http://pmartel.github.io/teaching

Funcionamento da Disciplina

- Aulas teóricas (14T)
- Aulas teórico-práticas (14TP)
  - Annothaton usado em algumas aulas, mas não conta para a avaliação
- Duas frequências (avaliação on-line no Moodle)
  - 1ª Frequência: semana 6-10 Novembro (7/11 ?)
  - 2ª Frequência: semana 18-22 Dezembro (18/12 ?)
- Exame Final (avaliação on-line no Moodle)
- Frequência das aulas:
  - Teóricas: livre
  - Práticas: livre

Página da disciplina: http://pjmartel.github.io/teaching/bioinfo

Programa

2023-09-18	Aula T1	Introdução à Bioinformática
2023-09-25	Aula T2	Portais, bases de dados e formatos de representação de objetos biológicos
2023-10-02	Aula T3	Alinhamento de Sequências
2023-10-09	Aula T4	Matrizes de Score
2023-10-16	Aula T5	Significância de Alinhamentos
2023-10-23	Aula T6	Pesquisa de sequências em bases de dados
2023-10-30	Aula T7	Alinhamento múltiplo de sequências
2023-11-06	Aula T8	Motivos e Perfis
2023-11-13	Aula T9	Análise Genómica I
2023-11-20	Aula T10	Análise Genómica II
2023-11-27	Aula T11	Filogenia Molecular
2023-12-04	Aula T12	Bioinformática Estrutural I
2023-12-11	Aula T13	Bioinformática Estrutural II
2023-12-18	Aula T14	Bioinformática Estrutural III

- **Paulo Martel**

Gabinete: FCT, Edifício C8, 3.12
Email: pmartel@ualg.pt
Homepage: http://pmartel.github.io/teaching

Bibliografia

- Choudhuri, Supratim, and Michael Kotewicz. *Bioinformatics for Beginners: Genes, Genomes, Molecular Evolution, Databases, and Analytical Tools*. Elsevier/AP, 2014.
- Claverie, Jean-Michel, and Cedric Notredame. *Bioinformatics for Dummies*. 2nd ed, Wiley Pub, 2007.
- Lesk, Arthur M. *Introduction to Bioinformatics*. Fifth edition, Oxford University Press, 2019.
- Mount, David W. *Bioinformatics: Sequence and Genome Analysis*. 2nd ed, Cold Spring Harbor Laboratory Press, 2004.

Bioinformática: o que é ?

O termo “bioinformática” foi utilizado pela primeira vez em 1970 por Pauline Hog e Ben Hesper, mas o seu significado alterou-se um pouco ao longo do tempo, não existindo uma definição universalmente aceite.

“Bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical-chemistry) and then applying “informatics” techniques (derived from disciplines such as applied math, CS, and statistics) to understand and organize the information associated with these molecules, on a large-scale.”

“**(1)** Bioinformatics is the development of computational methods for studying the structure, function, and evolution of genes, proteins and whole genomes.
(2) bioinformatics is the development of methods for the management and analysis of biological information arising from genomics and high throughput experiments.”

Bioinformática: o que é ?

- Processamento e análise de dados **biológicos**, na forma de **sequências** e **estruturas** moleculares
- **Armazenamento** e organização dos dados biológicos em **bancos de dados**
- **Algoritmos** e **ferramentas** de análise de dados biológicos em computadores digitais
- **Plataformas Web** de acesso a ferramentas e dados
- **Interconexão** de **bancos de dados** e **serviços** na rede digital

Origens

- Os primórdios da bioinformática ocorreram há mais de 50 anos, quando os computadores pessoais ainda eram uma hipótese e o DNA ainda não podia ser sequenciado.

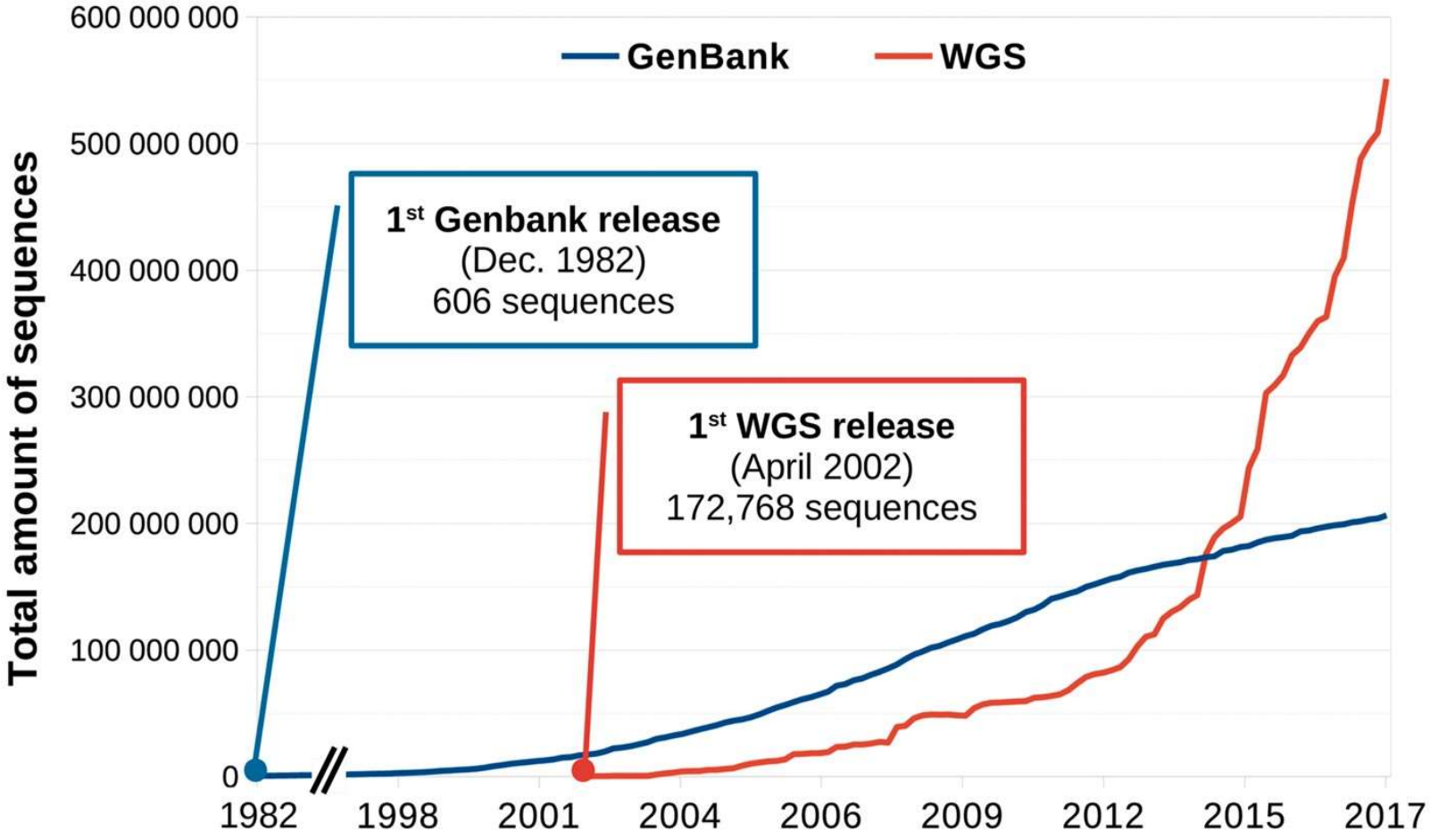
- Na década de 1960, foi desenvolvido o primeiro montador de sequência de peptídeo de novo, o primeiro banco de dados de sequências de proteínas e o primeiro modelo de substituição de aminoácidos para filogenética.

- Ao longo das décadas de 1970 e 1980, avanços paralelos na biologia molecular e na ciência da computação traçaram o caminho para empreendimentos cada vez mais complexos, como a análise de genomas completos.

- Nas décadas de 1990 a 2000, o uso da Internet, aliado à sequenciação de próxima geração, levou a um influxo exponencial de dados e a uma rápida proliferação de ferramentas de bioinformática.

- Hoje, a bioinformática enfrenta múltiplos desafios, como lidar com Big Data, garantir a reprodutibilidade dos resultados e uma integração adequada nos currículos acadêmicos.

A explosão de dados biológicos

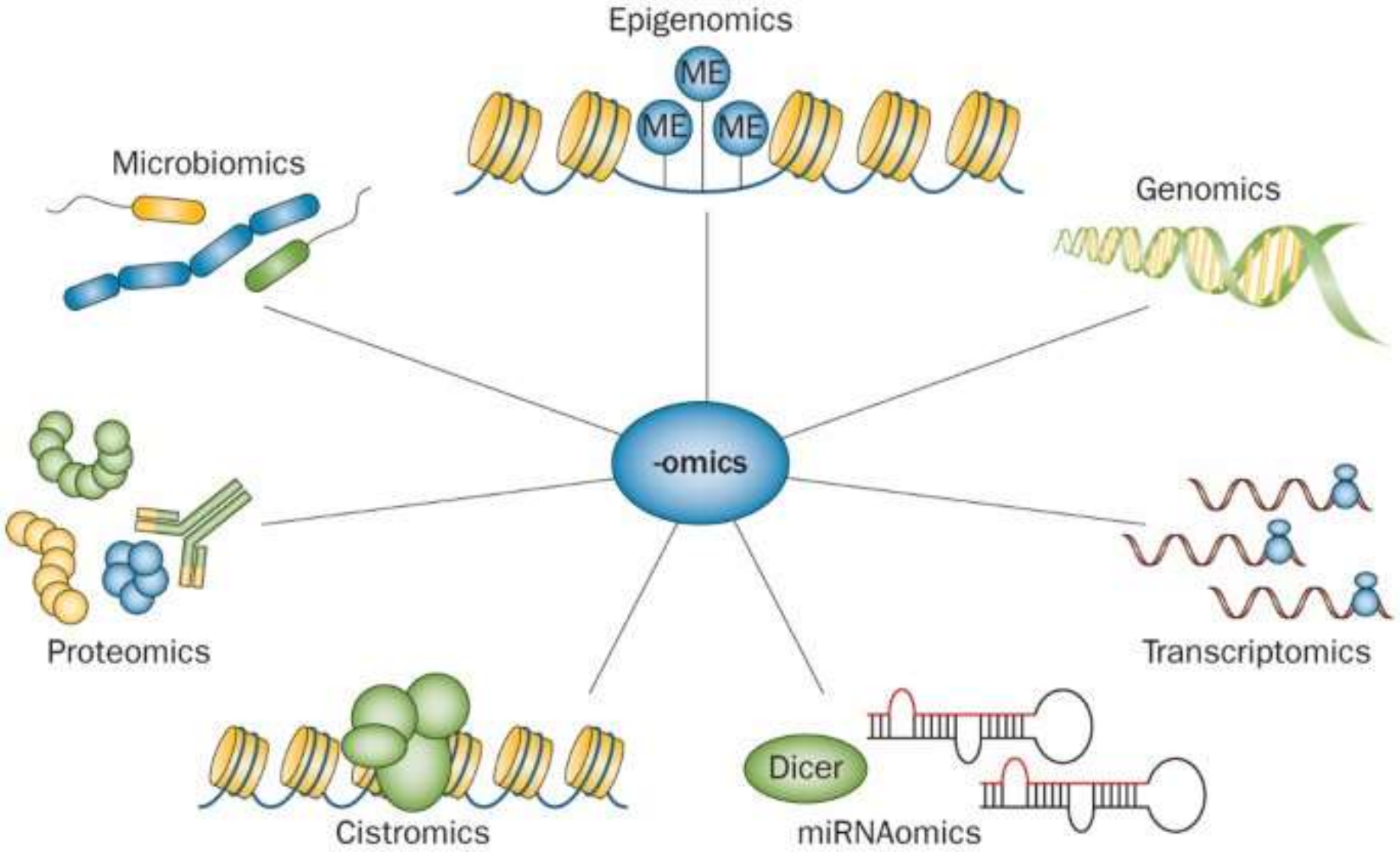


Total amount of sequences on the NCBI GenBank and WGS (Whole Genome Shotgun) databases over time. The number ...

Informação biológica

- Genomas (~1000)
  - Humanos: ~250 000
- Sequências genéticas (~$6.5x10^{7}$)
- Sequências de proteína (~$2.5x10^{6)}$
- Estruturas de proteínas (~$2.0x10^{5}$)
- Expressão genética
- População proteica (proteoma)
- Vias metabólicas (metaboloma)
- Interações proteína-proteína (interactoma)
- Bibliografia

ómicas


Epigenomics
ME
ME
ME
Microbiomics
Genomics
-omics
Proteomics
Transcriptomics
Cistromics
Dicer
miRNAomics

Moléculas biológicas

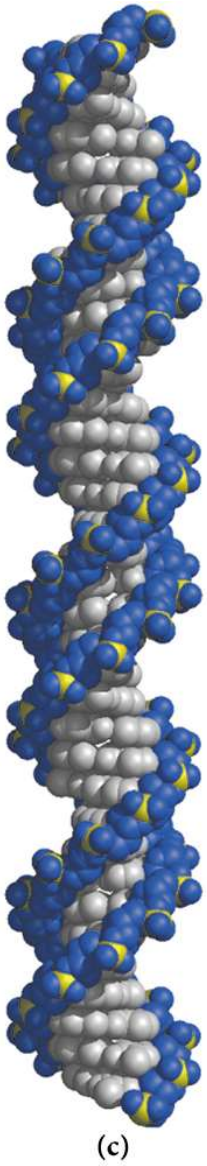
- **DNA:** repositório da informação genética na maioria dos organismos vivos
- **RNA:** transferência de informação genética, matriz para a síntese proteica, funções estruturais, etc...
- **Proteínas:** componentes estruturais (pele, ossos, músculo, cabelo, etc...), catálise de reacções bioquímicas (enzimas), transmissão de sinais, regulação, transdução de energia, etc., etc., etc.!...

Códigos biológicos

Table 1.2. Nucleic acid and protein

Macromolecule		Backbone	Repeating unit	Length	Role
Nucleic acid	DNA	Phosphodiester bonds	Deoxyribonucleotides (A, C, G, T)	10^3–10^8	Genome
	RNA	Phosphodiester bonds	Ribonucleotides (A, C, G, U)	10^3–10^5 10^3–10^4 10^2–10^3	Genome Messenger Gene product
Protein		Peptide bonds	Amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y)	10^2 – 10^3	Gene product

DNA



(c)

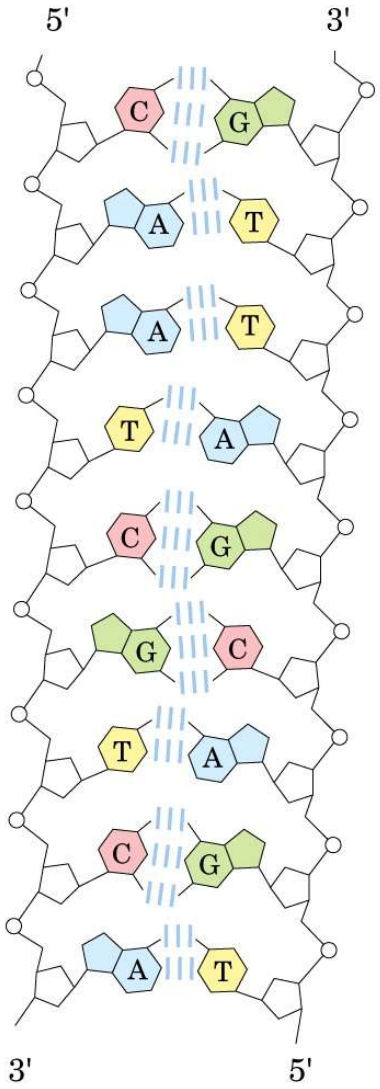
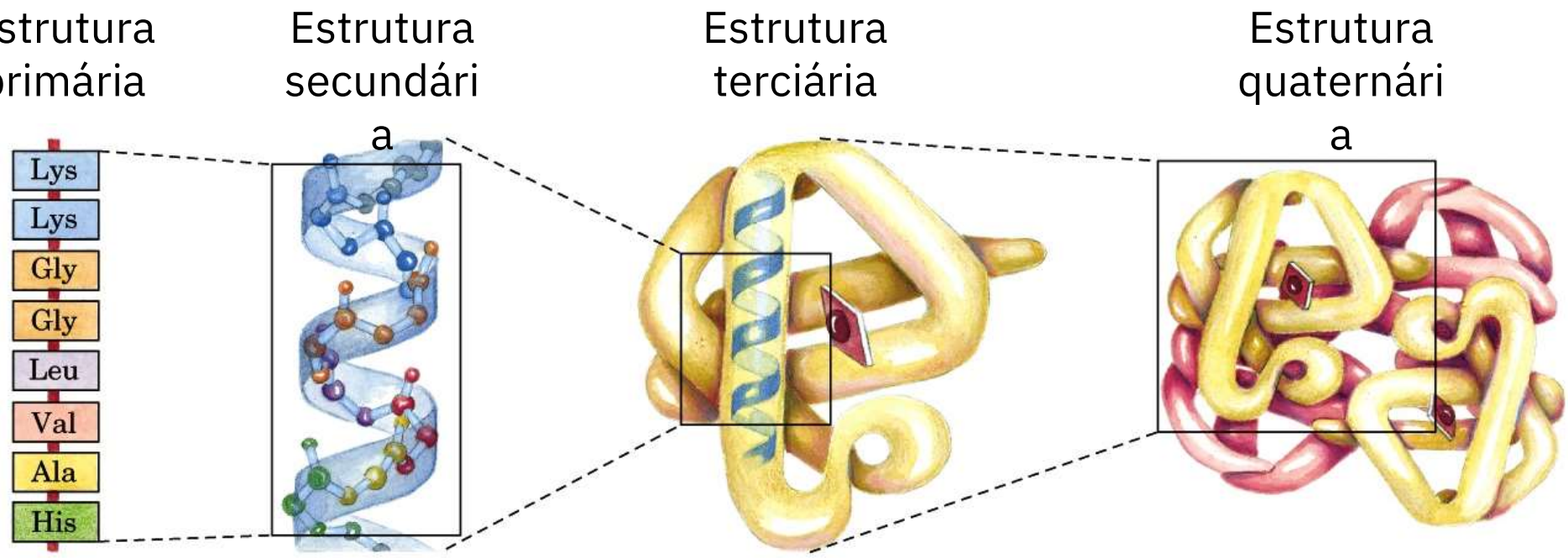


Table 1.3. Nucleotide codes

A	Adenine	W	Weak (A or T)
G	Guanine	S	Strong (G or C)
C	Cytosine	M	Amino (A or C)
T	Thymine	K	Keto (G or T)
U	Uracil	B	Not A (G or C or T)
R	Purine (A or G)	H	Not G (A or C or T)
Y	Pyrimidine (C or T)	D	Not C (A or G or T)
N	Any nucleotide	V	Not T (A or G or C)

Proteínas


Estrutura primária
Estrutura secundária
Estrutura terciária
Estrutura quaternária
Lys
Lys
Gly
Gly
Leu
Val
Ala
His
Sequência de aminoácidos
α-hélice
Cadeia polipetídica
Organização das subunidades

Proteínas

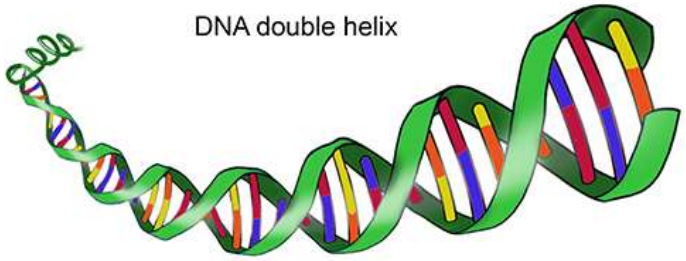
Table 1.4. Amino acid codes

Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic acid
Cys	C	Cysteine
Gln	Q	Glutamine
Glu	E	Glutamic acid
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Ser	S	Serine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine
Asx	B	Asn or Asp
Glx	Z	Gln or Glu
Sec	U	Selenocysteine
Unk	X	Unknown

Fluxo da informação biológica

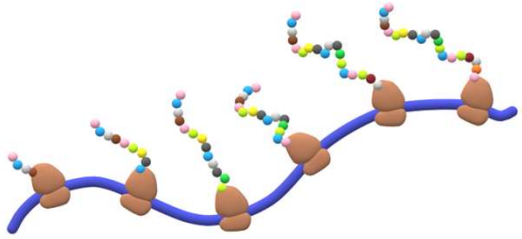
Dogma central da biologia molecular

Gene ...TTAATAA



⇩ transcrição

m-RNA ...UUAAUAAGU...

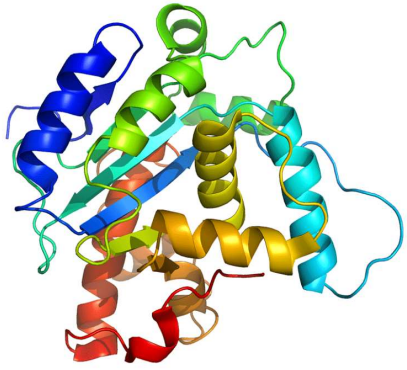


⇩ splicing, tradução

cadeia polipeptídica ...LISVHDN...

⇩ modificações pós-translacionais

proteína



Excepções: vírus de RNA, priões, ribozimas (?)

Problemas Bioinformáticos

• Dado um conjunto de sequências, alinhá-las de forma obter correspondência entre regiões homólogas

• Dada uma sequência de um gene ou proteína , encontrar sequências *homólogos* numa base de dados apropriada

• Dada a sequência de uma proteína, classificá-la na classe funcional e/ou estrutural apropriada

• Reconstruir as relações evolutivas (filogenia) entre um conjunto de sequências e/ou estruturas

• Dada a sequência de uma proteína, prever a sua estrutura tridimensional (*folding problem)*

• Dada a estrutura de uma proteína e de uma molécula pequena, prever o modo de interação e a estrutura do complexo por elas formado. (*docking problem)*