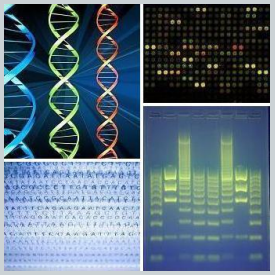


Bioinformática

Licenciaturas em Biologia, Bioquímica e Biotecnologia

Paulo Martel / João Varela
pmartle@ualg.pt

Docentes

- João Varela (bioinformática: conceitos, bases de dados, aplicações, pesquisa de ORFs e suas funções, anotação de sequências, uso da ferramenta Annothaton, incluindo filogenia, construção de árvores e identificação de OTUs).
- Paulo Martel (alinhamentos, pesquisas de sequências homólogas em base de dados, bioinformática estrutural)

Competências

- Conceito de ORF, pesquisa de ORFs em sequências nucleotídicas (JV)
- Previsão da localização de genes em genomas e metagenomas (JV)
- Bioinformática aplicada à taxonomia molecular (JV)
- Estimação de massa molecular de proteínas e ácidos nucleicos (JV)
- Alinhamentos de sequências: ferramentas e aplicações (PM)
- Construção de árvores filogenéticas: modelos, ferramentas e aplicações (JV)
- Previsão da função de genes e/ou proteínas por pesquisa de domínios funcionais (JV)
- Previsão da estrutura de proteínas; previsão da função e compreensão dos mecanismos de acção de proteínas através da bioinformática estrutural (JV e PM)
- Anotação de sequências de bases de dados via Annothaton (JV)

Avaliação - Bioinformática

- **Exame teórico** – JV 50%, PM 50%
- **Componente Prática** - Anotação de 1 sequência metagenómica (Annotathon) – não há grupos; exame prático na Época de Recorrência caso haja reprovação
- NOTA FINAL = 70% Teórica + 30% Prática

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."

Hesper B, Hogeweg P. Bioinformatica: een werkconcept. *Kameleon.* 1970;1(6):28–29

Bioinformática: Conceito

Bioinformática - campo interdisciplinar da Biologia, Ciências Informáticas, Matemática e Estatística para analisar dados de sequências biológicas e genomas (genes e disposição de genes em cromossomas) e prever a estrutura e função de macromoléculas

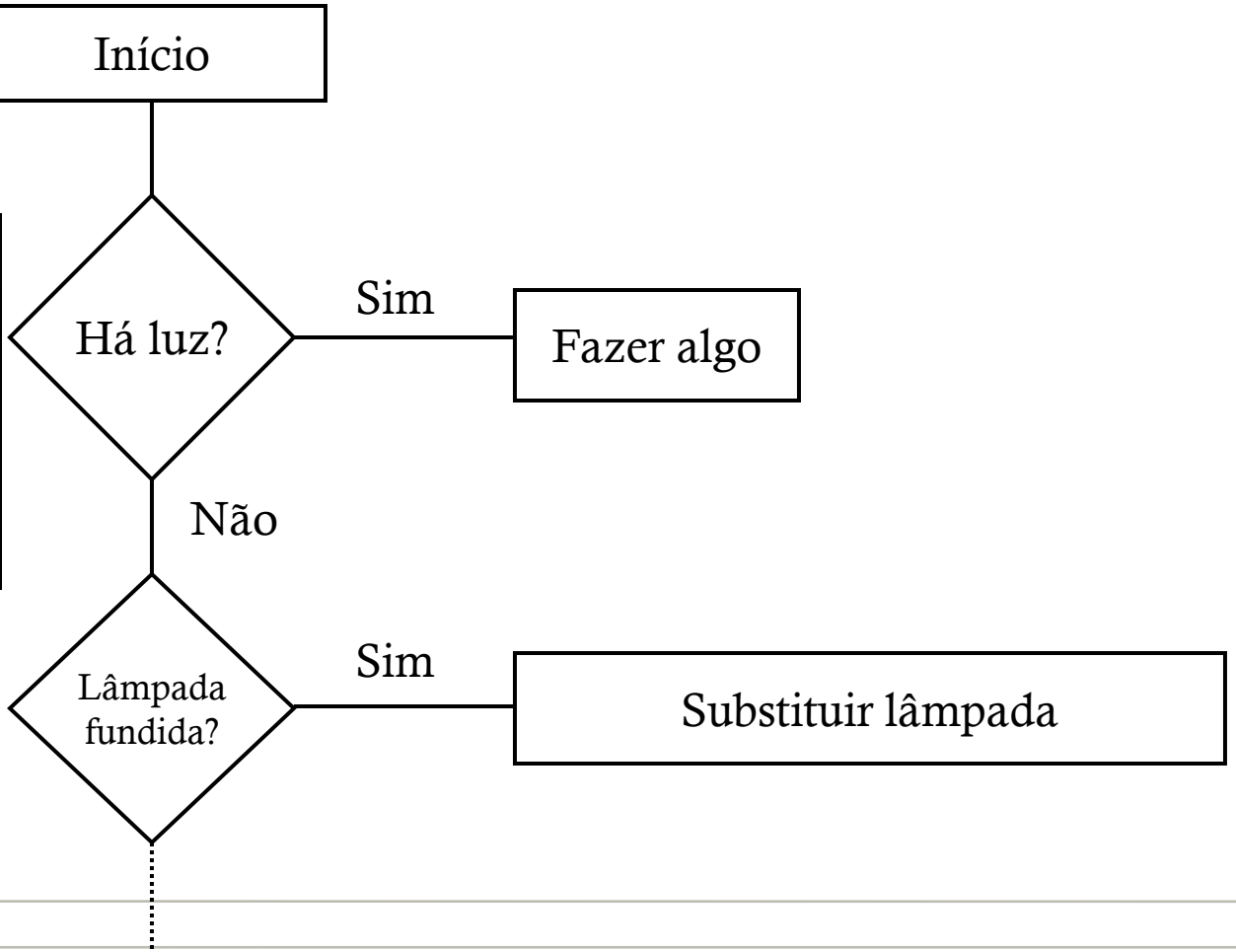
Bioinformática vs. Biologia Computacional

- Bioinformática – aplicação de ferramentas computacionais à análise e processamento de dados biológicos, geralmente de grande volume.
- Biologia Computacional – desenvolvimento mais genérico de algoritmos computacionais capazes de resolver uma variedade de problema biológicos incluindo (mas não só), algoritmos bioinformáticos como o alinhamento, pesquisa de sequências, comparação de estruturas, identificação de motivos, filogenia molecular, etc.

Algoritmo

Conjunto de instruções bem definidas para resolver um problema; quando aplicado em informática, um algoritmo pode ser transformado em um programa e este numa aplicação informática.

Algoritmo representado por um diagrama de fluxo



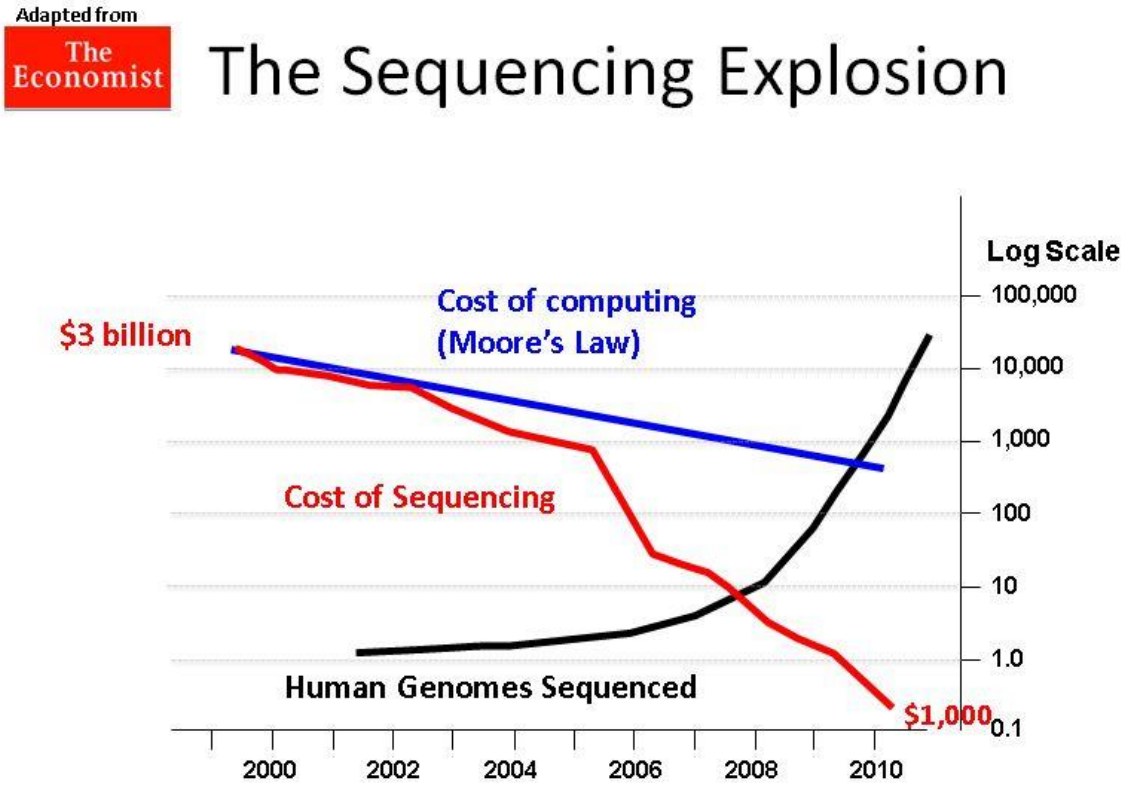
Bioinformática: para que serve?

Limitação da mente humana em armazenar e processar informação devido a:

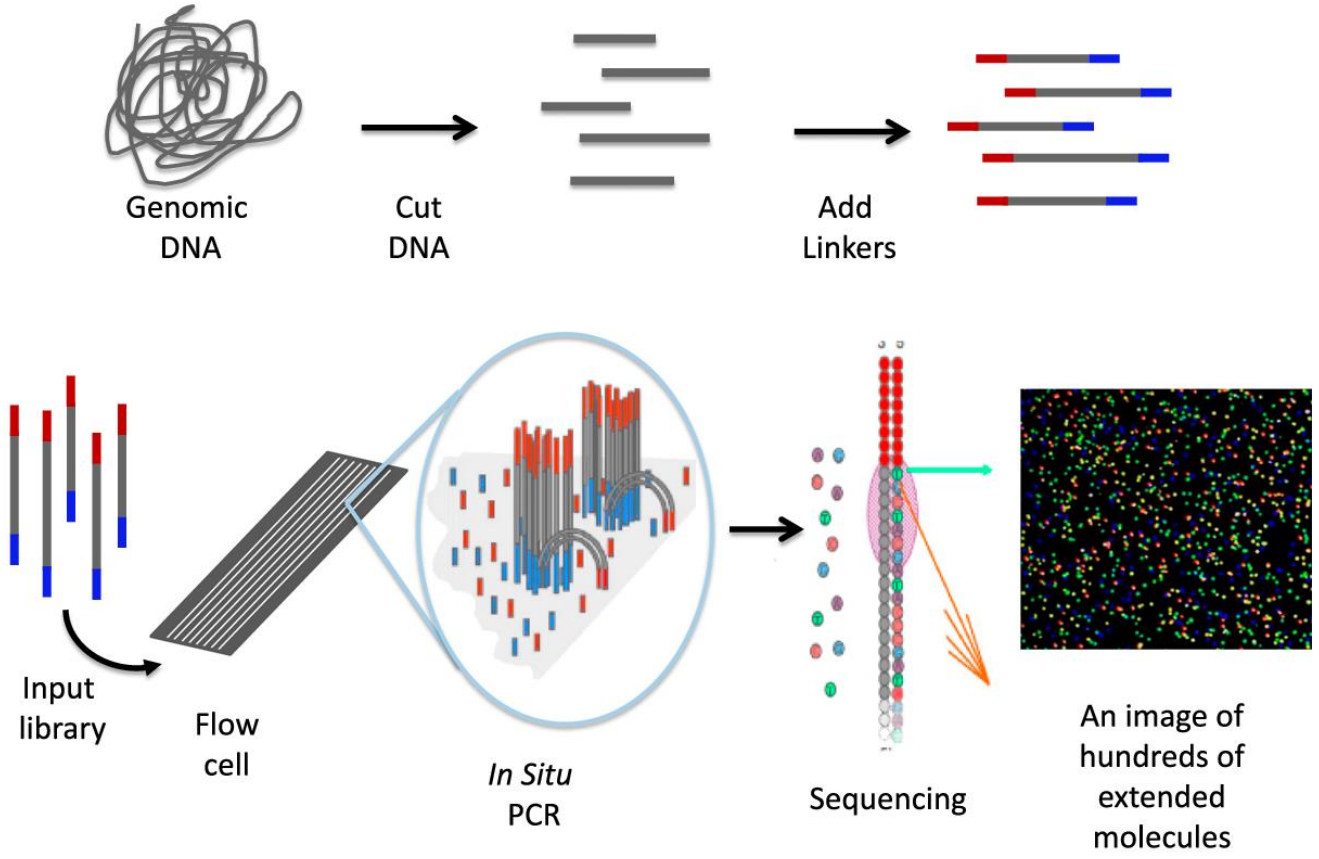
- Complexidade dos genomas
- Nº crescente de genomas sequenciados
- Nº crescente de anotações estruturais, funcionais e bibliográficas
- Nº crescente de estruturas de proteínas

Explosão de Sequenciação

Adapted from The Economist



Sequenciação de Nova Geração



Como resolver esta limitação humana?

- Aplicações computorizadas de armazenamento de dados: bases de dados
- Aplicações computorizadas de processamento de dados

Exemplo de Bases de Dados

(sequências primárias de ácidos nucleicos e proteína com anotação básica)

- GenBank (NCBI, EUA) *http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi*
- EMBL Nucleotide Sequence Database (Europa) *http://www.ebi.ac.uk/embl/*
- DNA Data Bank of Japan (DDBJ, Japão) *http://www.ddbj.nig.ac.jp/*

Tipo de dados armazenados

- Sequências de DNA (desoxinucleótidos)
- Sequências de RNA (nucleótidos)
- Sequências de Proteína (aminoácidos)
- Anotações estruturais
- Anotações funcionais
- Anotações de localização intracelular
- Anotações de localização genómica
- Anotações bibliográficas

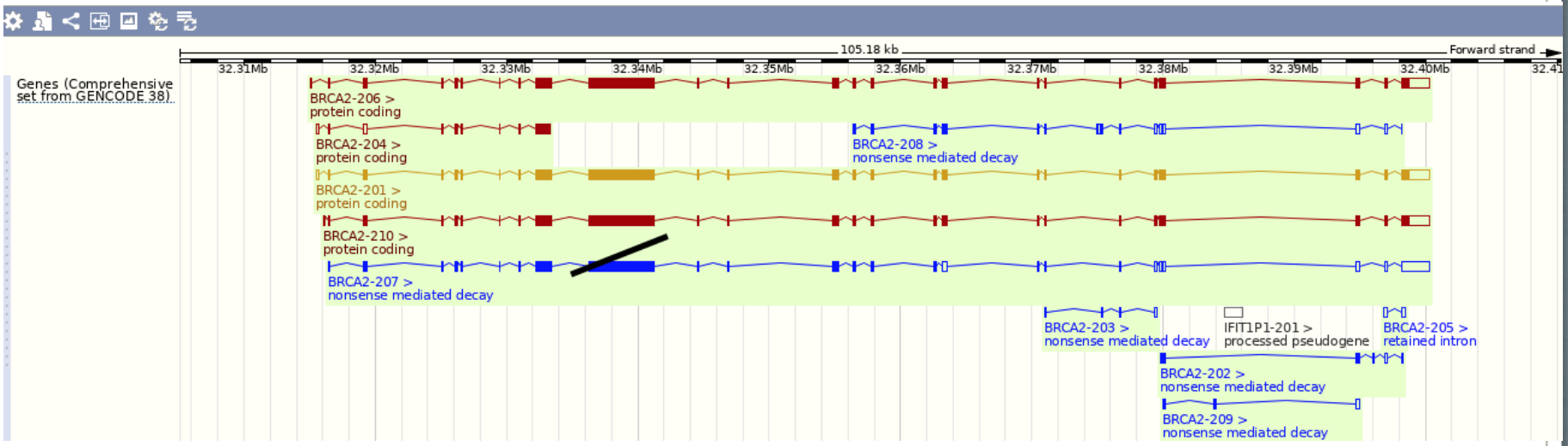
Exemplo de Bases de Dados

(sequências primárias de ácidos nucleicos e proteína com anotação mais complexa)

- Ensembl (Genomas de vertebrados)
  https://www.ensembl.org/
- Ensembl (Genomas Bacterianos)
  https://bacteria.ensembl.org/
- UniProtKB/Swiss-Prot (Proteínas)
  https://www.uniprot.org

Exemplo de um gene anotado

(anotações do gene *BRCA* humano na plataforma Ensembl)

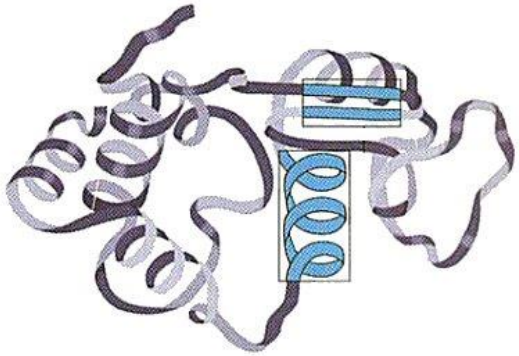


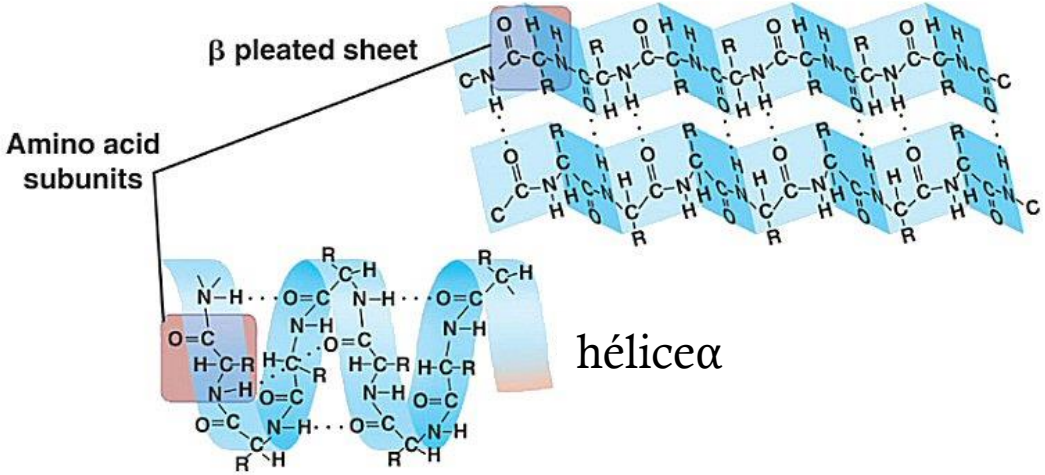
Exemplo de Bases de Dados

(estruturas secundárias de proteínas - domínios funcionais)

- InterPro
  https://www.ebi.ac.uk/interpro/
- ScanProsite
  https://prosite.expasy.org/scanprosite/

Estrutura Secundária e Domínios Funcionais



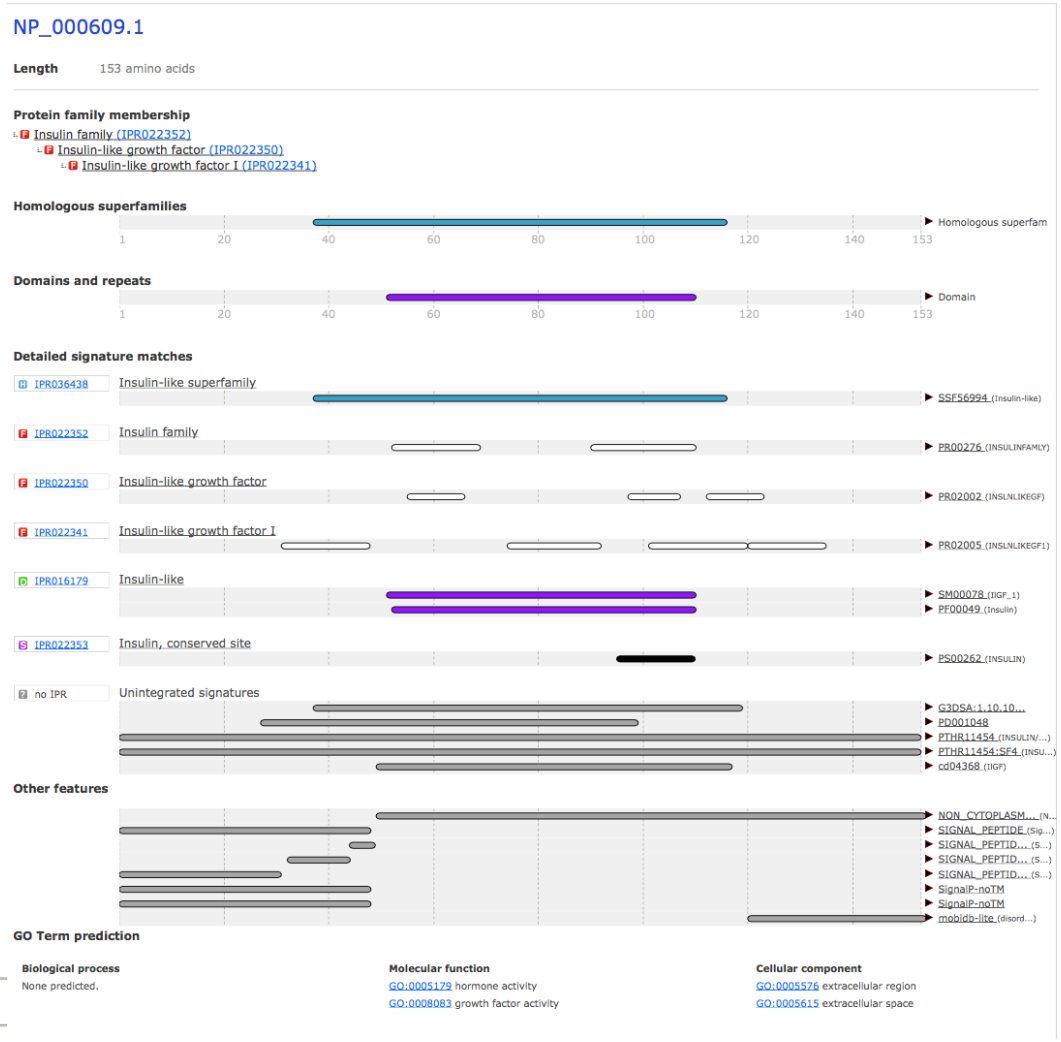


Folhas β pregueadas

hélice α

Domínios Funcionais Anotados

(Ex: InterProScan – Anotações estruturais e funcionais)

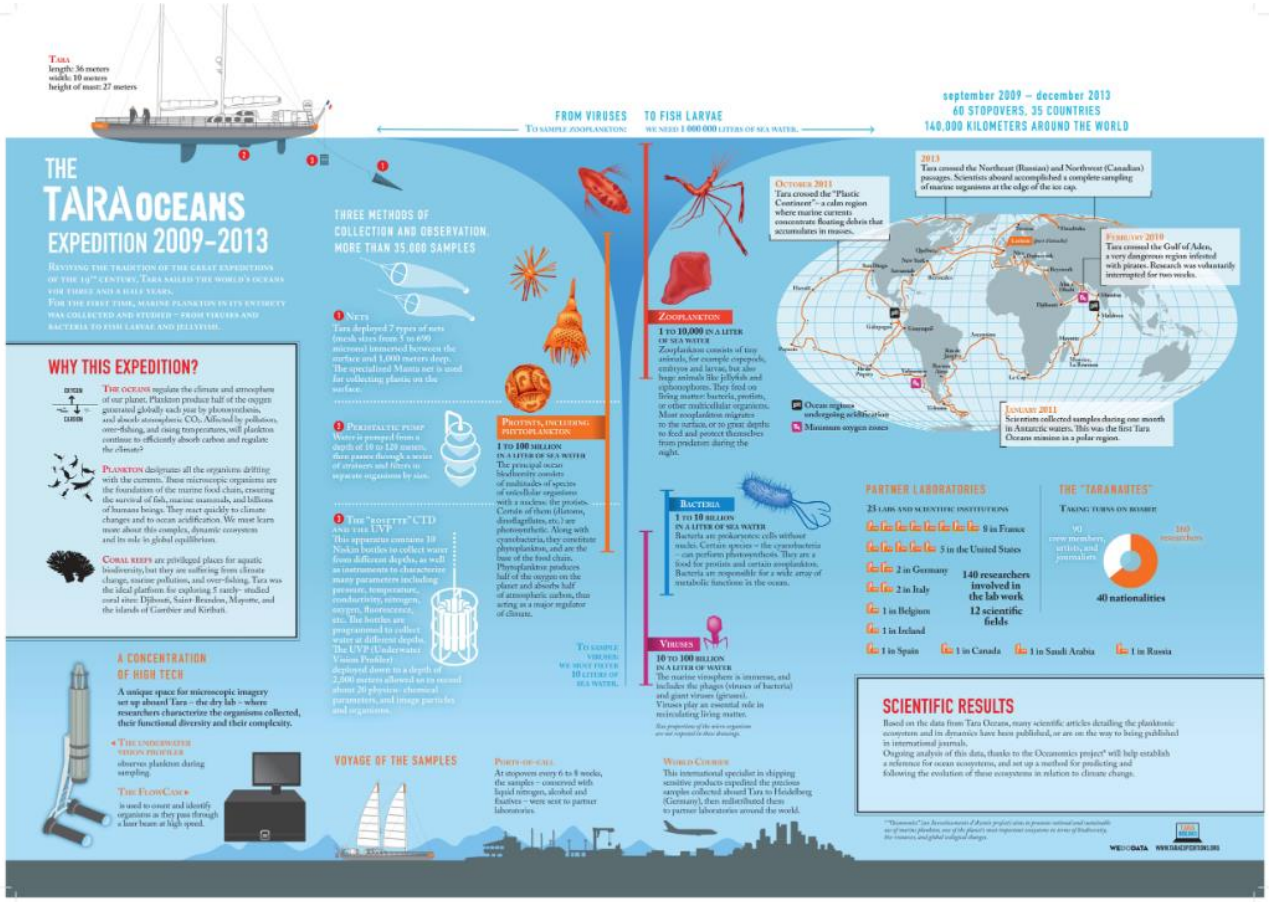


Como funciona a bioinformática em termos práticos?

1. Recolha de amostras (laboratoriais, ambientais)
2. Isolamento de DNA (ou RNA) [por vezes opcional]
3. Amplificação de sequências de DNA (ou cDNA) por PCR [opcional]
4. Sequenciação de DNA
5. Armazenamento da sequência em ficheiros informáticos
6. Tratamento bioinformático das sequências

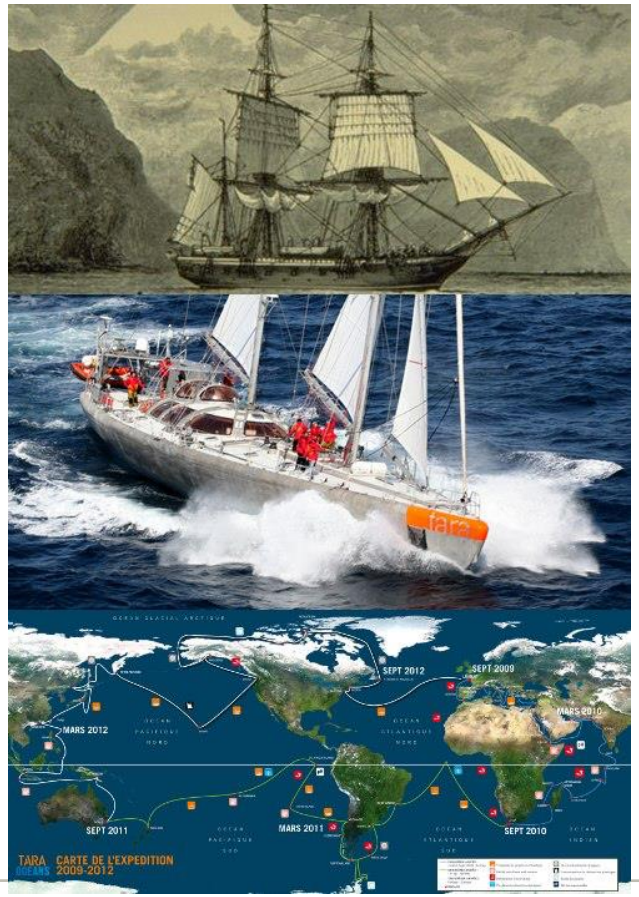
Recolha de Amostras Ambientais

Annothaton: Amostras de água do mar



Recolha de Amostras Ambientais

Annothaton: Amostras de água do mar (http://annotathon.org/)



Preparação para a aula teórico-prática

- Ir ao site ANNOTHATON
  http://annotathon.org/
- Leiam a página "Rule Book" para ver o que tem que fazer nas aulas teórico-práticas

Como funciona a bioinformática em termos práticos?

1. Recolha de amostras (laboratoriais, ambientais)
2. Isolamento de DNA (ou RNA) [por vezes opcional]
3. Amplificação de sequências de DNA (ou cDNA) por PCR [opcional]

Genética Molecular

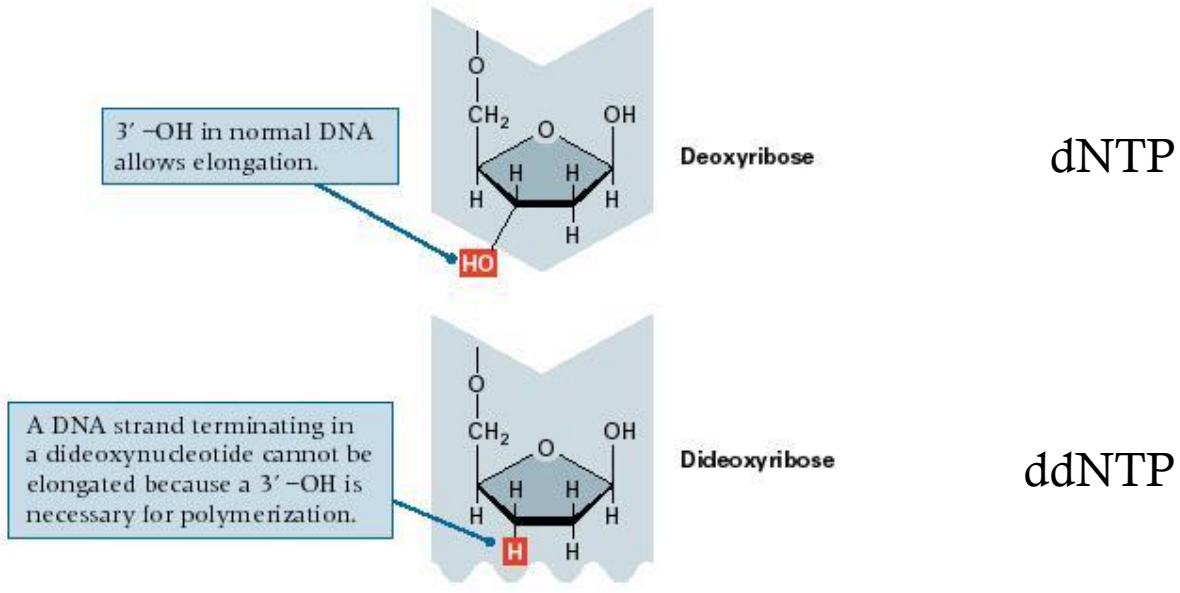
4. Sequenciação
5. Armazenamento da sequência em ficheiros informáticos
6. Tratamento bioinformático das sequências

Como funciona a bioinformática em termos práticos?

1. Recolha de amostras (laboratoriais, ambientais)
2. Isolamento de DNA (ou RNA) [por vezes opcional]
3. Amplificação de sequências de DNA (ou cDNA) por PCR [opcional]
4. → Sequenciação de DNA
5. Armazenamento da sequência em ficheiros informáticos
6. Tratamento bioinformático das sequências

Sequenciação de DNA

O Método de Sanger baseia-se no uso de desoxinucleótidos artificiais em reacções de síntese de DNA (polimerização).



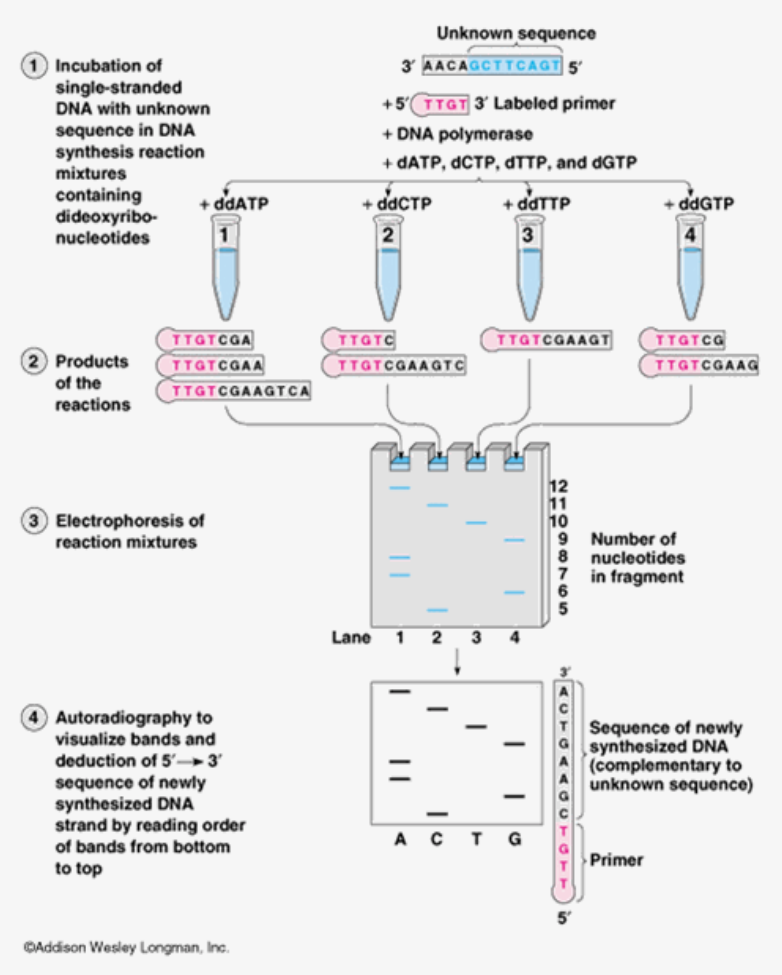
Desoxinucleótidos vs. Didesoxinucleótidos

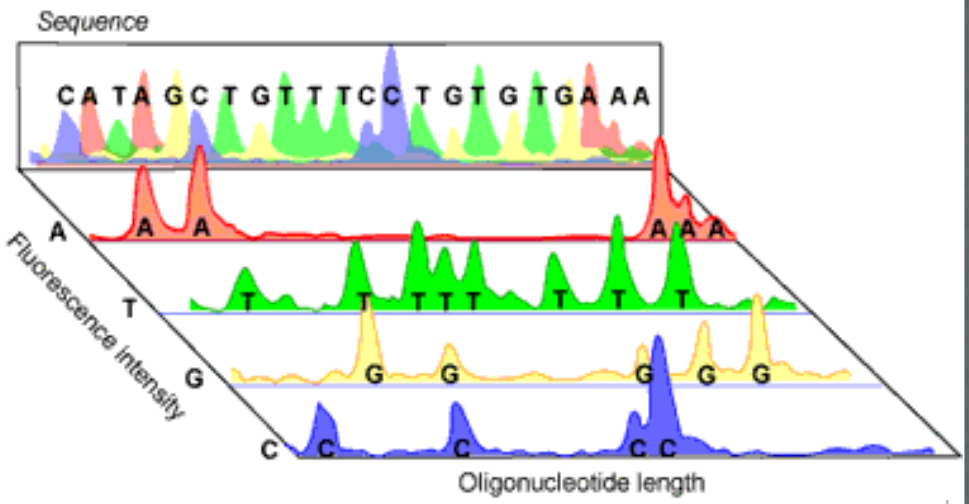
Sequenciação de DNA

Os didesoxinucleótidos bloqueiam a polimerização

Ex: Uso de ddTTP para gerar fragmentos de diferente tamanho terminando na posição das T

Sequenciação de DNA





Como funciona a bioinformática em termos práticos?

1. Recolha de amostras (laboratoriais, ambientais)
2. Isolamento de DNA (ou RNA) [por vezes opcional]
3. Amplificação de sequências de DNA (ou cDNA) por PCR [opcional]
4. Sequenciação de DNA
5. → Armazenamento da sequência em ficheiros informáticos
6. Tratamento bioinformático das sequências

Ficheiros Bioinformáticos

- Ficheiros de texto
- Ficheiros de DNA contêm letras que representam desoxinucleótidos
- Ficheiros de Proteína contêm letras que correspondem a aminoácidos
- Podem conter também notas explicativas sobre a sequência que se denominam: ANOTAÇÕES

Códigos de representação de sequências nucleotídicas em ficheiros bioinformáticos

Table 1: List of Nucleotides

Symbol	Meaning	Origin of designation
a	a	adenine
g	g	guanine
c	c	cytosine
t	t	thymine
u	u	uracil
r	g or a	purine
y	t/u or c	pyrimidine
m	a or c	amino
k	g or t/u	keto
s	g or c	strong interactions 3H-bonds
w	a or t/u	weak interactions 2H-bonds
b	g or c or t/u	not a
d	a or g or t/u	not c
h	a or c or t/u	not g
v	a or g or c	not t, not u
n	a or g or c or t/u, unknown, or other	any

Códigos de representação de sequências proteicas em ficheiros bioinformáticos

One-letter symbol	Three-letter symbol	Amino acid
A	Ala	alanine
B	Asx	aspartic acid or asparagine
C	Cys	cysteine
D	Asp	aspartic acid
E	Glu	glutamic acid
F	Phe	phenylalanine
G	Gly	glycine
H	His	histidine
I	Ile	isoleucine
K	Lys	lysine
L	Leu	leucine
M	Met	methionine
N	Asn	asparagine
P	Pro	proline
Q	Gln	glutamine
R	Arg	arginine
S	Ser	serine
T	Thr	threonine
U*	Sec	selenocysteine
V	Val	valine
W	Trp	tryptophan
X**	Xaa	unknown or 'other' amino acid
Y	Tyr	tyrosine
Z	Glx	glutamic acid or glutamine (or substances such as 4-carboxyglutamic acid and 5-oxoproline that yield glutamic acid on acid hydrolysis of peptides)

Formato de ficheiros de sequências nucleotídicas e a.a.

- FASTA
- FASTA-PEARSON
- NBRF
- GCG
- PIR
- GenBank
- PHYLIP
- ASN.1
- PAUP

Formato FASTA

Ficheiro de texto (ASCII) com:

- Linha de comentário iniciada por “>”
- Sequência (nucleotídica ou proteica)
- Terminação da sequência por “*” (opcional)

Exemplo:

```
> YCZ2_Yeast protein in HMR 3’ region
MKAVVIEDGKAVVKEVGP*
```

Formato de ficheiros de sequências

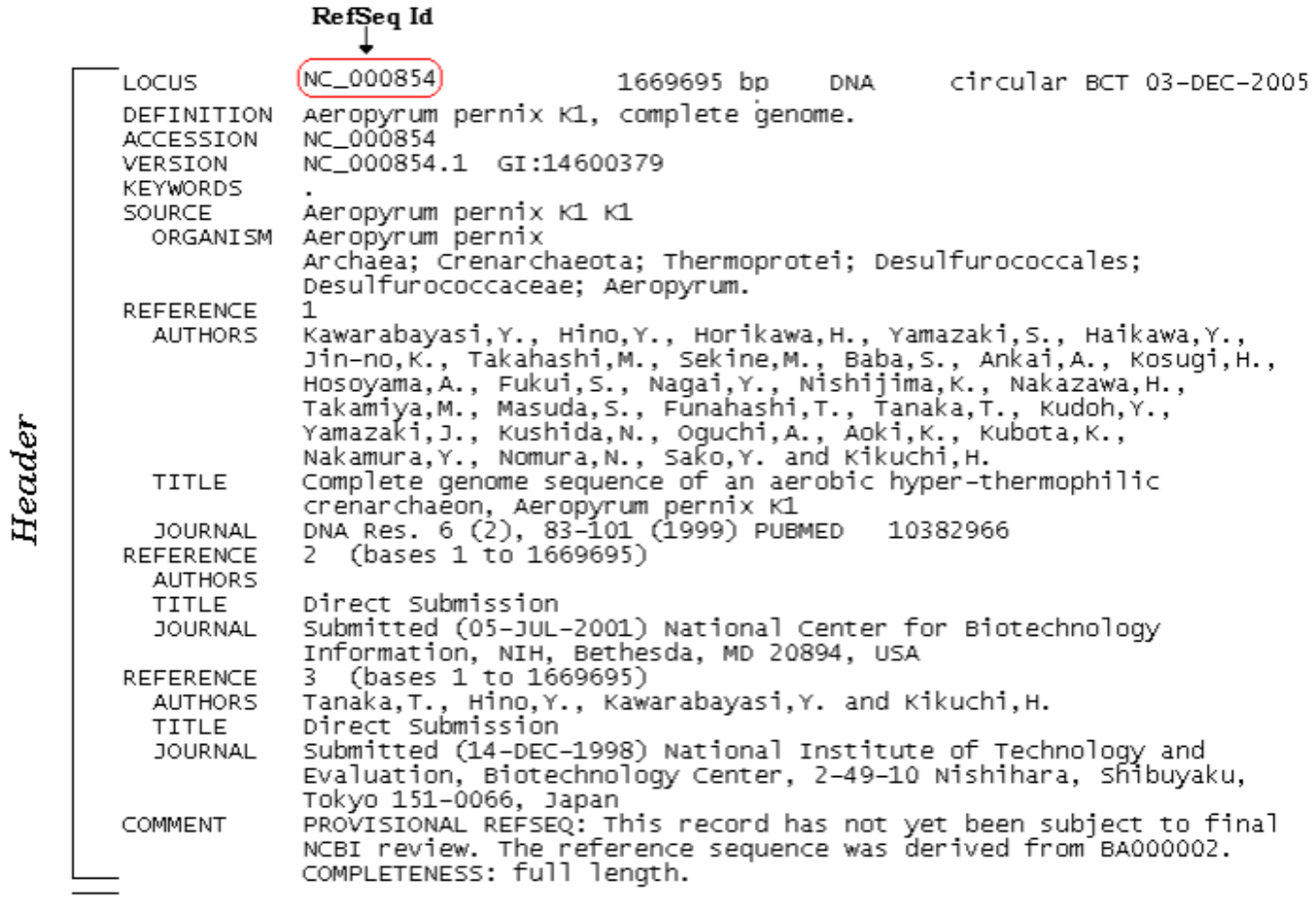
- FASTA
- FASTA-PEARSON
- NBRF
- GCG
- PIR
- GenBank
- PHYLIP
- ASN.1
- PAUP

Formato GenBank

Ficheiro de texto com:

- Cabeçalho (*header*)
- Anotações (*features*)
- Sequência

Formato GenBank: Cabeçalho

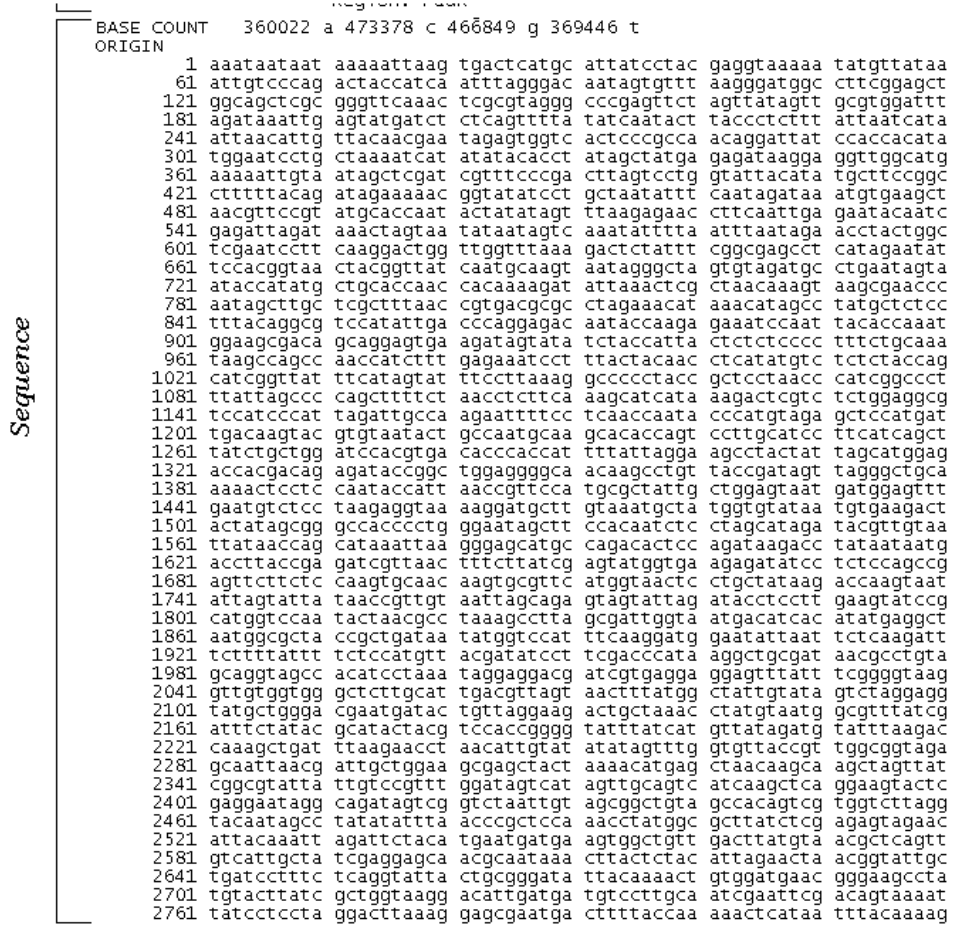


Formato GenBank: Anotações

Features

```
FEATURES             Location/Qualifiers
     source          1..1669695
                     /mol_type="genomic DNA"
                     /db_xref="taxon:272557"
                     /organism="Aeropyrum pernix K1"
     gene            complement(213..938)
                     /locus_tag="APE0001"   <-- Locus tag
                     /db_xref="GeneID:1445602"
     CDS             complement(213..938)
                     /locus_tag="APE0001"
                     /protein_id="NP_146894.1"
                     /transl_table=11
                     /db_xref="GI:14600380"
                     /db_xref="GeneID:1445602"
                     /codon_start=1
                     /product="hypothetical protein"
                     /translation="MVDILSSLLLSLPFGVIGFLLVLSPGSIWTPVKESIGYVYVSRR
                     VTVKASKLLGSLTLLASLISFVVGAAYGITIQASTLALLLALITVVTVEYSMRLAEIE
                     SLNQPVLEGFEPVGSIKLKYLTIILLVYLISIVFSIEGSLKLYSIGAYGTLASHLSIE
                     ILAGYTVFLSVKRPEAYVIPGLSRETIELLQFFMPTSLSLIAIGVYMILAGFHMWWII
                     LLAGVTTLFVVTMLIMINKEGKY"
     gene            complement(938..1276)
                     /locus_tag="APE0002"
                     /db_xref="GeneID:1445577"
     CDS             complement(938..1276)
                     /locus_tag="APE0002"   <-- Locus tag
                     /protein_id="NP_146895.1"
                     /transl_table=11
                     /note="similar to PIR:C69525 percent identity:39.583 in
                     96aa."
                     /db_xref="GI:14600381"
                     /db_xref="GeneID:1445577"
                     /codon_start=1
                     /product="hypothetical protein"
                     /translation="MDPADKLMKDARTGVLALAVLHVLVNHGALHGYWLRKILGNLMG
                     WTPPETSLYDALKRLEKLGLIKGRWVRSGRGPLRKYYEITDAGRETYEVVVKDFSKMV
                     GWLICRKGRE"
     misc_feature    complement(1001..>1180)
                     /locus_tag="APE0002"
                     /db_xref="CDD:43477"
                     /note="Transcriptional regulator PadR-like family;
                     Region: PadR"
```

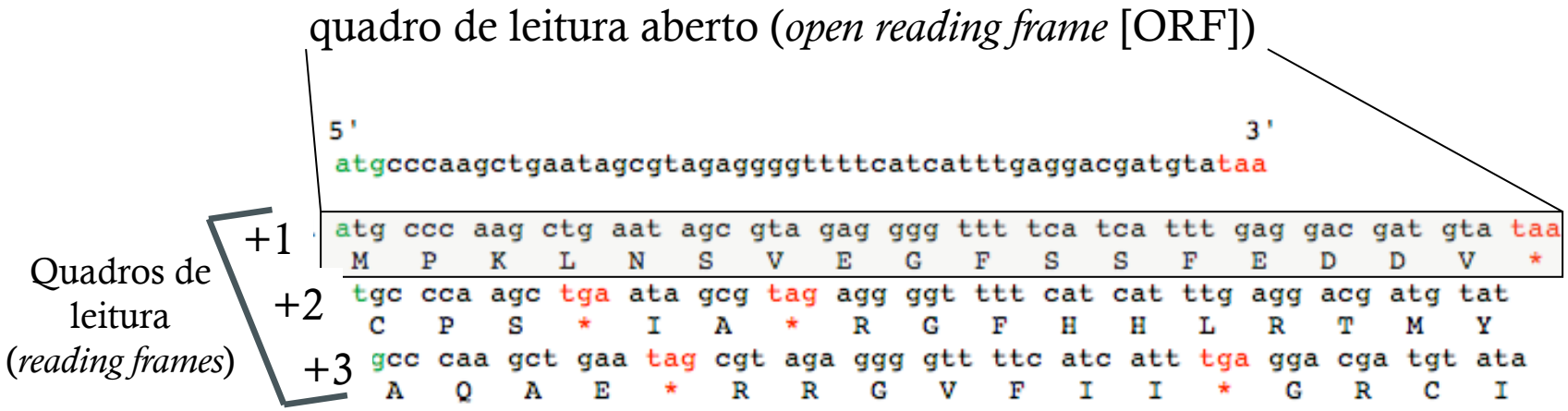
Formato GenBank: Sequência



Como funciona a bioinformática em termos práticos?

1. Recolha de amostras (laboratoriais, ambientais)
2. Isolamento de DNA (ou RNA) [por vezes opcional]
3. Amplificação de sequências de DNA (ou cDNA) por PCR [opcional]
4. Sequenciação de DNA
5. Armazenamento da sequência em ficheiros informáticos
6. → Tratamento bioinformático das sequências (anotações)

Tradução virtual de sequências nucleotídicas em proteicas



Genómica vs. Metagenómica