

# Bioinformática

Licenciaturas em Biologia, Bioquímica e  
Biotecnologia

Paulo Martel / João Varela  
[pmartle@ualg.pt](mailto:pmartle@ualg.pt)

# Docentes

- João Varela (bioinformática: conceitos, bases de dados, aplicações, pesquisa de ORFs e suas funções, anotação de sequências, uso da ferramenta Annotathon, 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 Annotathon (JV)

# Avaliação - Bioinformática

- **Exame teórico** – JV 50%, PM 50%
- **Componente Prática** - Anotação de 1 sequência metagenômica (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.”

# 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

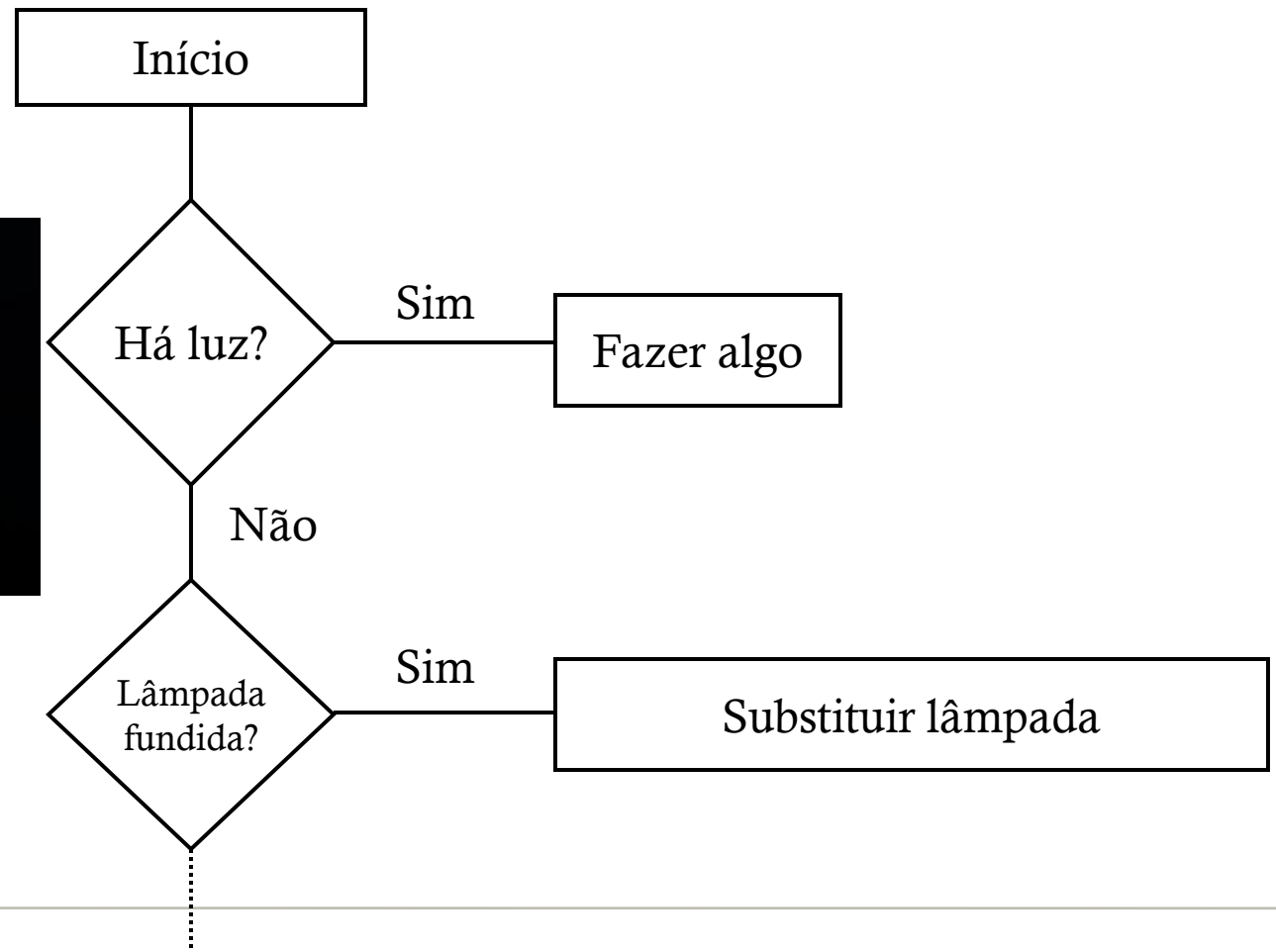
- **Bioinformática** – aplicação de ferramentas computacionais à análise e processamento de dados biológicos, geralmente de grande volume.
- **Biologia Computacional** – desenvolvimento mais genérico de **algoritmos computacionais** capazes de resolver uma variedade de problema biológicos incluindo (mas 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?

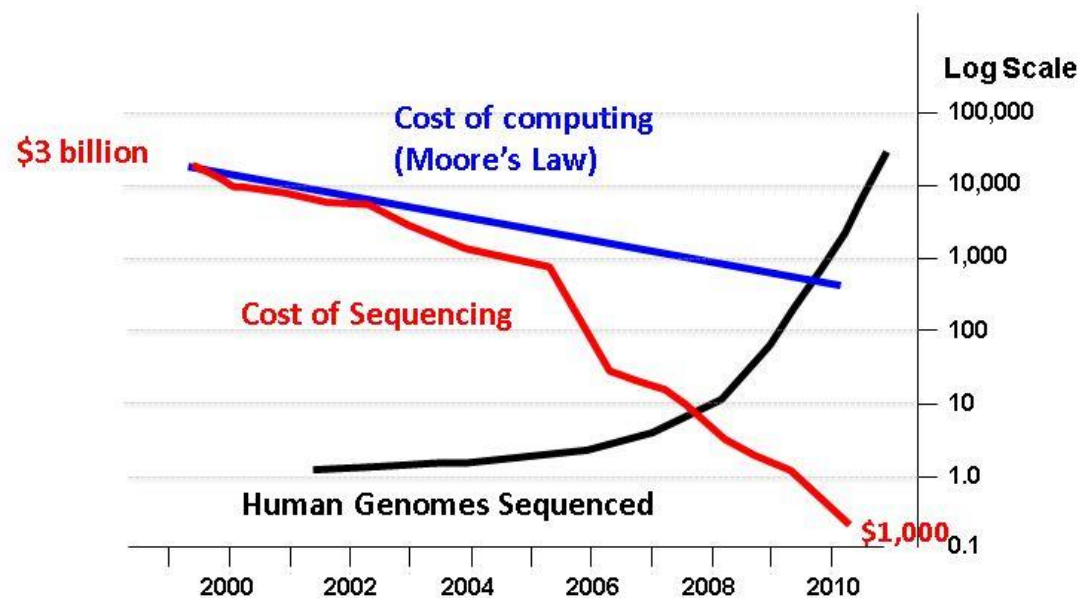
Limitação da mente humana em [armazenar](#) e [processar](#) informação devido a:

- Complexidade dos genomas
- N° crescente de genomas sequenciados
- N° crescente de anotações estruturais, funcionais e bibliográficas
- N° crescente de estruturas de proteínas

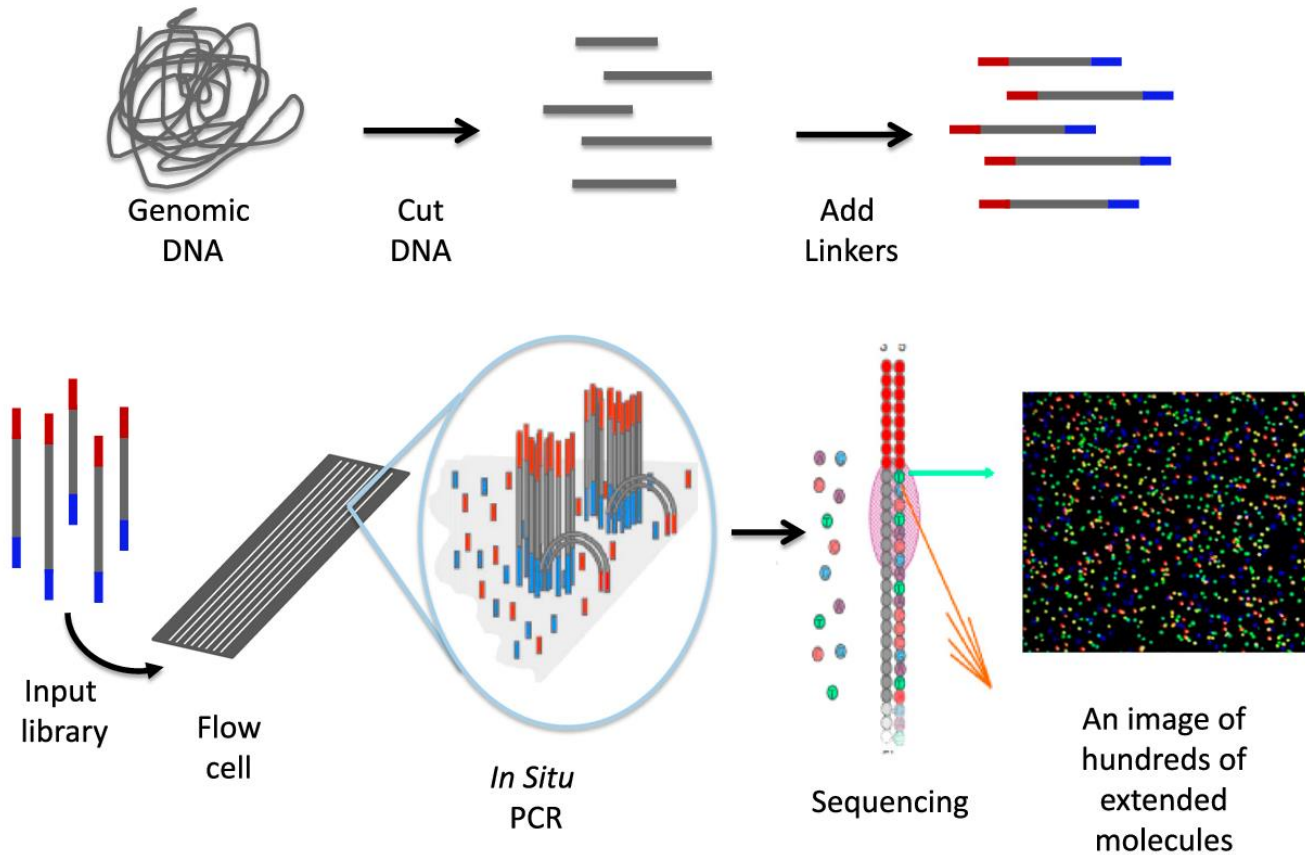
# Explosão de Sequenciação

Adapted from  
**The Economist**

## The Sequencing Explosion



# Sequencição de Nova Geração



# Como resolver esta limitação humana?

- Aplicações computadorizadas de **armazenamento** de dados: **bases de dados**
- Aplicações computadorizadas 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



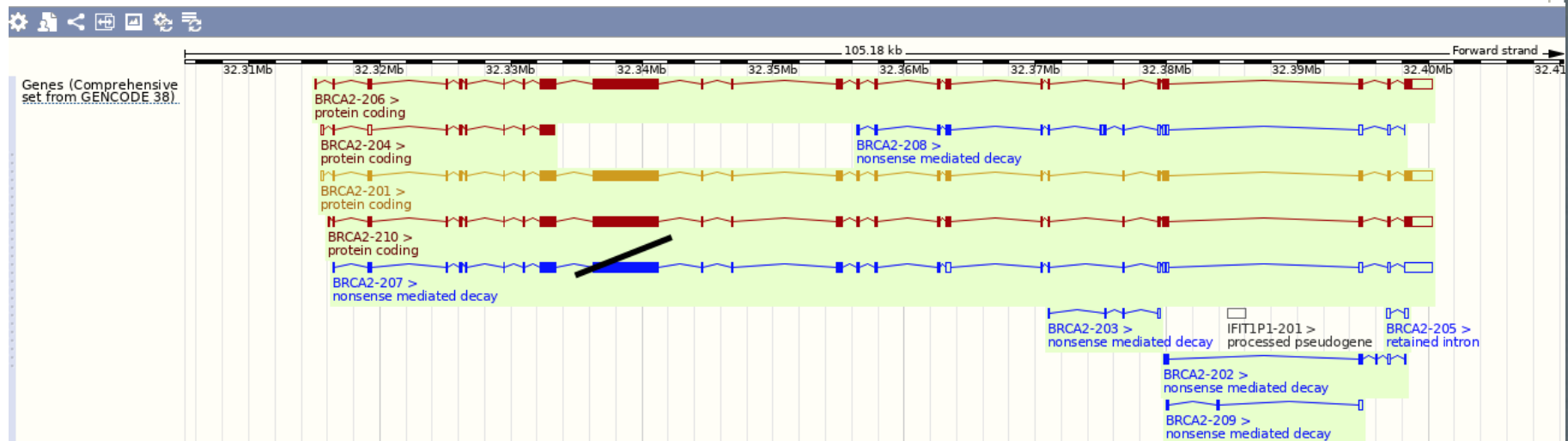
# Exemplo de Bases de Dados

(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)

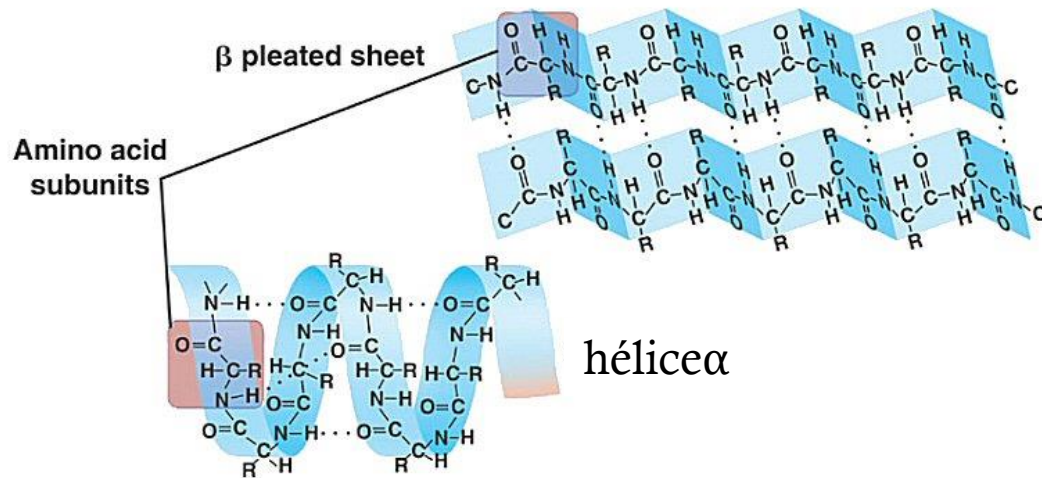
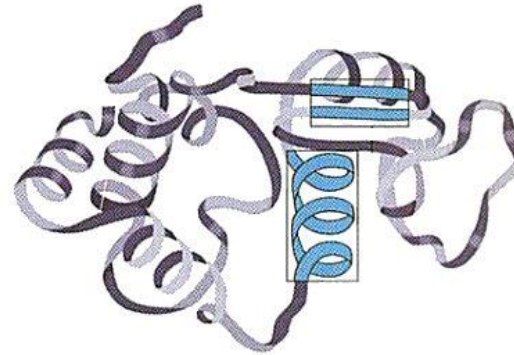
- InterPro

<https://www.ebi.ac.uk/interpro/>

- ScanProsite

<https://prosite.expasy.org/scanprosite/>

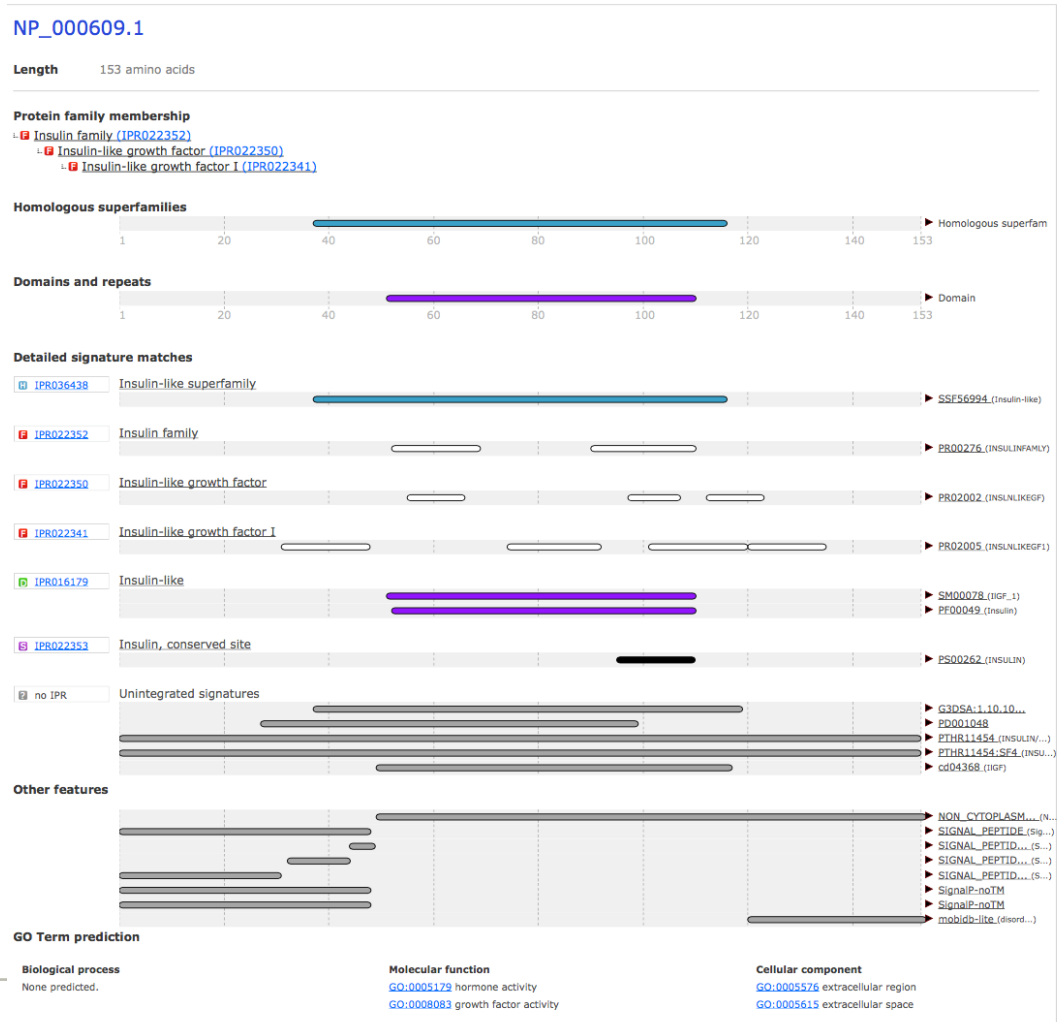
# Estrutura Secundária e Domínios Funcionais



Folhas  $\beta$   
pregueadas

# Domínios Funcionais Anotados

(Ex: InterProScan – Anotações estruturais e funcionais)

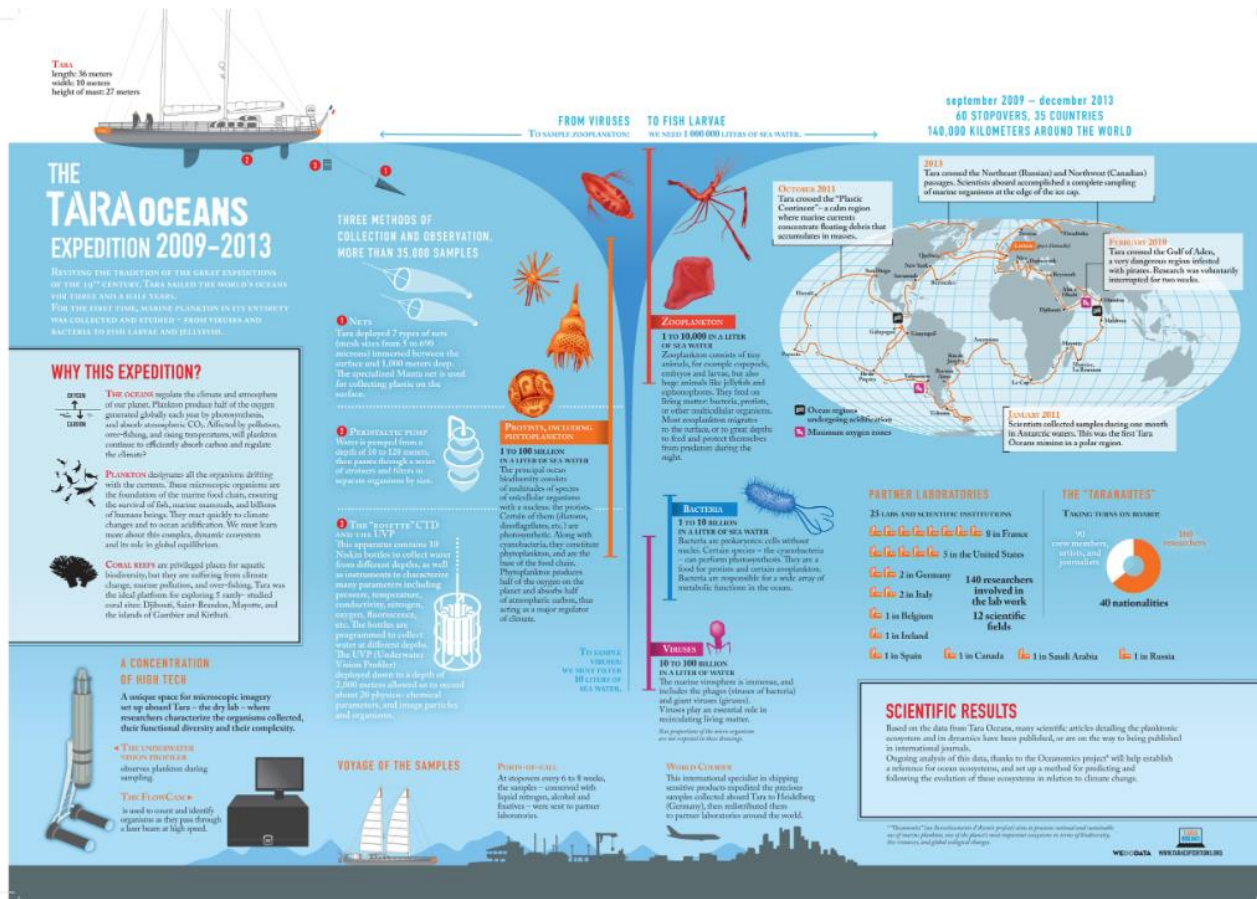


# Como funciona a bioinformática em termos práticos?

- 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

## Annotaton: Amostras de água do mar





# Recolha de Amostras Ambientais

Annotathon: 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?

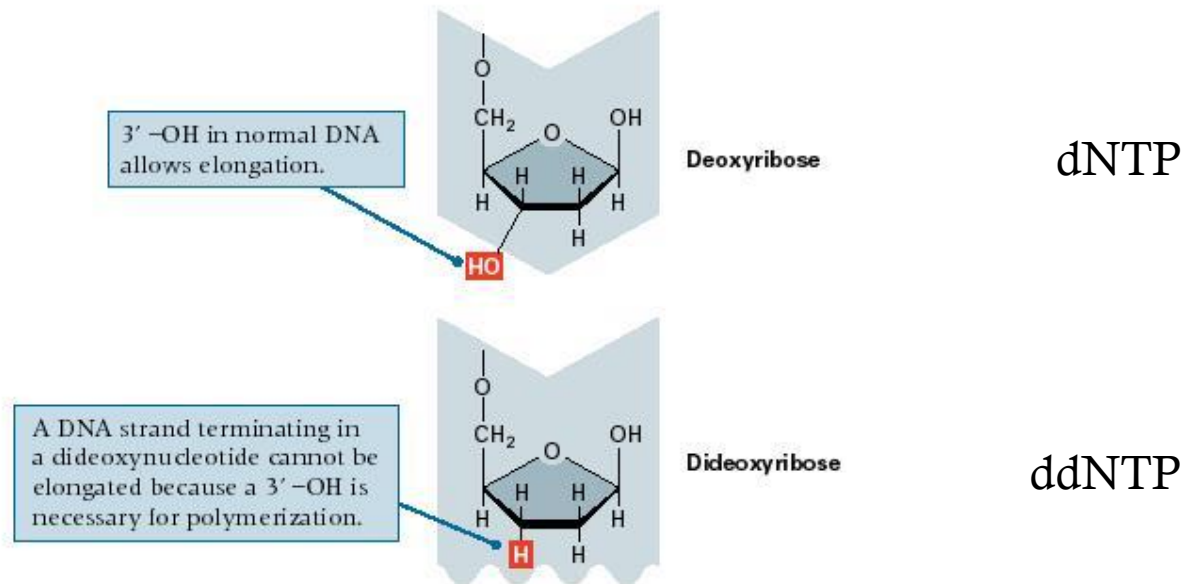
1. Recolha de amostras (laboratoriais, ambientais)
2. Isolamento de DNA (ou RNA) [por vezes opcional]
3. Amplificação de sequências de DNA (ou cDNA) por PCR [opcional]  
**Genética Molecular**
4. Sequenciação
5. Armazenamento da sequência em ficheiros informáticos
6. Tratamento bioinformático das sequências

# 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

# Sequenciación de DNA

Os didesoxinucleótidos bloqueiam a polimerização

DNA Polymerase reads the template strand and synthesizes a new second strand to match:



IF 5% of the T nucleotides are actually dideoxy T, then each strand will terminate when it gets a ddT on its growing end:



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



# 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	<u>a</u> denine
g	g	<u>g</u> uanine
c	c	<u>c</u> ytosine
t	t	<u>t</u> hymine
u	u	<u>u</u> racil
r	g or a	<u>p</u> urine
y	t/u or c	<u>p</u> irimidine
m	a or c	<u>a</u> mino
k	g or t/u	<u>k</u> eto
s	g or c	<u>s</u> trong interactions 3H-bonds
w	a or t/u	<u>w</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	<u>a</u> ny

# 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

One-letter symbol	Three-letter symbol	Amino acid
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

Header

**RefSeq Id**  
↓

LOCUS **NC\_000854** 1669695 bp DNA circular BCT 03-DEC-2005

DEFINITION Aeropyrum pernix K1, complete genome.  
ACCESSION NC\_000854  
VERSION NC\_000854.1 GI:14600379  
KEYWORDS .  
SOURCE Aeropyrum pernix K1  
ORGANISM Aeropyrum pernix  
Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales;  
Desulfurococcaceae; Aeropyrum.

REFERENCE 1  
AUTHORS Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y.,  
Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankaï, A., Kosugi, H.,  
Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Nakazawa, H.,  
Takamiya, M., Masuda, S., Funahashi, T., Tanaka, T., Kudoh, Y.,  
Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Kubota, K.,  
Nakamura, Y., Nomura, N., Sako, Y. and Kikuchi, H.  
TITLE Complete genome sequence of an aerobic hyper-thermophilic  
crenarchaeon, Aeropyrum pernix K1  
JOURNAL DNA Res. 6 (2), 83-101 (1999) PUBMED 10382966  
REFERENCE 2 (bases 1 to 1669695)  
AUTHORS  
TITLE Direct Submission  
JOURNAL Submitted (05-JUL-2001) National Center for Biotechnology  
Information, NIH, Bethesda, MD 20894, USA  
REFERENCE 3 (bases 1 to 1669695)  
AUTHORS Tanaka, T., Hino, Y., Kawarabayasi, Y. and Kikuchi, H.  
TITLE Direct Submission  
JOURNAL Submitted (14-DEC-1998) National Institute of Technology and  
Evaluation, Biotechnology Center, 2-49-10 Nishihara, Shibuyaku,  
Tokyo 151-0066, Japan  
COMMENT PROVISIONAL REFSEQ: This record has not yet been subject to final  
NCBI review. The reference sequence was derived from BA000002.  
COMPLETENESS: full length.

# 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="MVDILSSLLLSLPPGVIGFLLVLSPGSIVTPVKESIGYVYVSRR
                    VTKASKLLGSLTLLASLISFVVGAAAYGITIQASTLALLLALITVVTVEYSMRLAEIE
                    SLNQPVLEGFEPVGSIKLKYLTIILLVYLSIVFSIEGSLKLYSIGAYGTLASHLSIE
                    ILAGYTVFLSVKRPEAYVIPGLSRETIELLQFFMPTSLSLIAIGVYMLAGFHMWWII
                    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
                    WTPPETSLYDALKRLEKLGLIKRWVRSGRGPLRKYIEITDAGRETYEVVVKDFSKMV
                    GWLICRKGRE"
misc_feature        complement(1001..>1180)
                    /locus_tag="APE0002"
                    /db_xref="CDD:43477"
                    /note="Transcriptional regulator PadR-like family;
                    Region: PadR"
```



# Formato GenBank: Sequência

Sequence

```

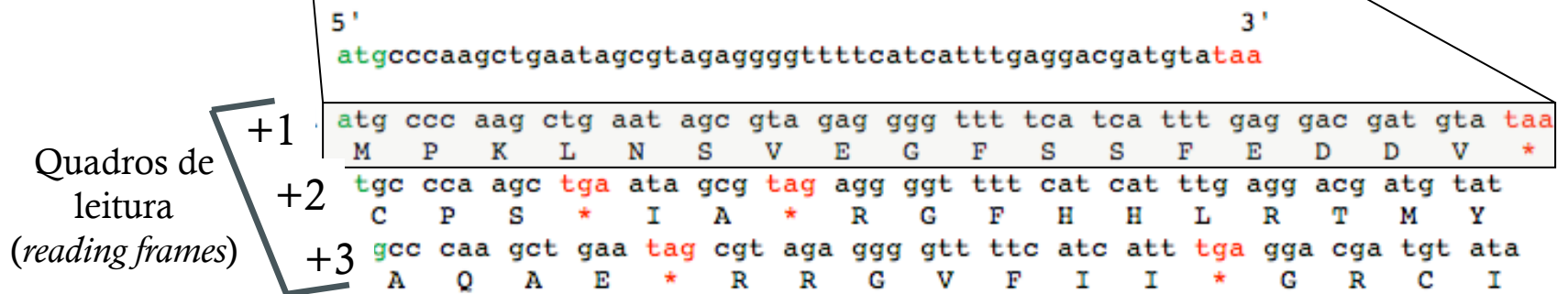
                                     origin: 1448
BASE COUNT   360022 a 473378 c 466849 g 369446 t
ORIGIN
1   aaataaat  aaaattaag  tgactcatgc  attatcctac  gaggtaaaa  tatgttataa
61  attgtcccg  actaccatca  atttagggac  aatagtgttt  aagggatggc  cttcggagct
121 ggagctcgc  gggttcaaac  tcgctagggg  cccgagttct  agttatagtt  gcgtggattt
181 agataaattg  agtatgatct  ctcaagttta  tatcaatact  tacctctttt  attaatcata
241 attaacattg  ttacaacgaa  tagagtggtc  actcccgcga  acaggattat  ccaccacata
301 tggaaatcctg  ctaaaatcat  atatacacct  atagctatga  gagataagga  ggttggcatg
361 aaaaattgta  atagctcgat  cgtttccgga  cttagctctg  gtattacata  tgctccggcg
421 ctttttacag  atagaaaaac  ggtatatcct  gctaataatt  caatagataa  atgtgaagct
481 aacgttccgt  atgcaccaat  actatatagt  ttaagagAAC  cttcaattga  gaatacaatc
541 gagattagat  aaactagtaa  tataatagtc  aaatatttta  atttaataga  acctactggc
601 tcgaatcctt  caaggactgg  ttggtttaaa  gactctattt  cggcgagcct  catagaatat
661 tcacaggtaa  ctacggttat  caatgcgaag  aatagggcta  gtgtagatgc  ctgaatagta
721 ataccatattg  ctgcaccaac  cacaaaagat  attaaactcg  ctacaaaagt  aagcgaacc
781 aatagcttgc  tcgctttaac  cgtgacgcgc  ctagaaacat  aaacatagcc  tatgctctcc
841 tttacaggcg  tccatattga  cccaggagac  aataccaaga  gaaatccaat  tacaccaaat
901 ggaagcgaca  gcaggagtga  agatagtata  tctaccatta  ctctctcccc  tttctgcaaa
961 taagccagcc  aacctctttt  gagaatcct  ttactacaac  ctcatatgtc  tctctaccag
1021 catcggttat  ttcatagtat  ttctttaaag  gccccctacc  gctcctaacc  catcggccct
1081 ttattagccc  cagcttttct  aacctcttca  aagcatcata  aagactcgtc  tctggaggcg
1141 tccatcccat  tagattgcca  agaattttcc  tcaaccaata  cccatgtaga  gctccatgat
1201 tgacaagtac  gtgtaafact  gccaatgcaa  gcacaccagt  ccttgcactc  ttcacagct
1261 tatctgctgg  atccacgtga  caccacacat  tttattagga  agcctactat  tagcatggag
1321 accacgacag  agataccggc  tggaggggca  acaagcctgt  taccgatagt  tagggctgca
1381 aaaactcctc  caataccatt  aaccgttcca  tgcgctattg  ctggagtaat  gatggagttt
1441 gaatgtctcc  taagaggtaa  aaggatgctt  gtaaatgcta  tgggtgataa  tgtgaagact
1501 actatagcgg  gccaccctcg  ggaatagctt  ccacaactc  cttagcataga  tacgttgtaa
1561 ttataaccag  cataaattaa  gggagcatgc  cagacactcc  agataagacc  tataataatg
1621 acctaccga  gatcgttaac  tttcttatcg  agtatggtga  agagatatcc  tctccagccg
1681 agttcttctc  caagtgcaac  aagtgcgttc  atggtaaact  ctgctataag  accaagtat
1741 attagtatta  taaccgttgt  aattagcaga  gtagtattag  atacctctt  gaagtatccg
1801 catgggtcaa  tactaacgcc  taagcctta  gcgattggtg  atgacatcac  atatgaggtc
1861 aatggcgcta  ccgctgataa  tatggtccat  ttcaaggatg  gaattataat  tctcaagatt
1921 tcttttattt  tctccatggt  acgatctct  tcgaccata  aggcctgcgat  aacgcctgta
1981 gcaggtagcc  acatctaaa  taggaggacg  atcgtgagga  ggagtttatt  tcggggttag
2041 gttgtggtgg  gctcttgcac  tgacgttagt  aactttatgg  ctattgtata  gtctaggagg
2101 tatgctggga  cgaatgatac  tghtaggaag  actgctaaac  ctatgtaatg  gcgtttatcg
2161 atttctatc  gcatactacg  tccaccgggg  tatttatcat  gttatagatg  tatttaagac
2221 caaagctgat  ttaagaacct  aacattgtat  atatagtttg  gtgttacctg  tggcggtaga
2281 gcaattaacg  attgctggaa  cggagctact  aaaaatgag  ctacaaagca  agctagttat
2341 cggcgtatta  ttgtccgttt  ggatagtcac  agttgcagtc  atcaagctca  ggaagactc
2401 gaggatagg  cagatagtcg  gtctaattgt  agcggctgta  gccacagctg  tggcttagg
2461 tacaatagcc  tatatatta  acccgctcca  aacctatggc  gcttatctcg  agagtagaac
2521 attcaaaatt  agattctaca  tgaatgatga  agtggctgtt  gacttatgta  acgctcagtt
2581 gtcattgcta  tcgaggagca  acgcaataaa  cttactctac  attagaacta  acggtattgc
2641 tgatcctttc  tcaggtatta  ctgcccggata  ttacaaaact  gtggatgaac  gggaaagcct
2701 tgcattatc  gctggttaagg  acattgatga  tgccttgca  atcgaattcg  acagtaaaat
2761 tatcctccta  ggacttaaa  gagcgaatga  cttttacc  aaactcata  tttacaaaag
```

# 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

quadro de leitura aberto (*open reading frame* [ORF])



# Genômica vs. Metagenômica

