BANCO DE DADOS BIOLÓGICOS

Aula 11

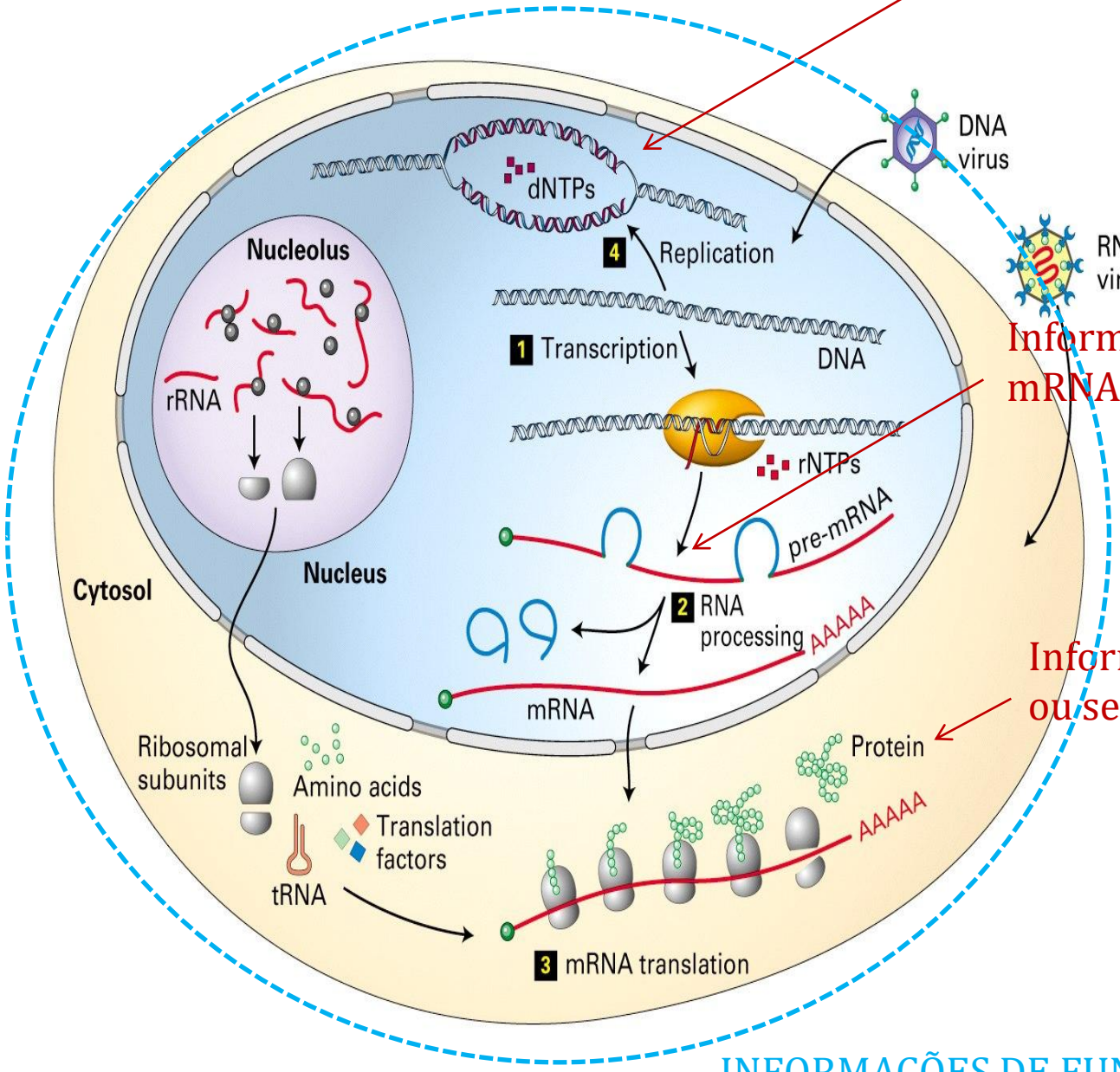
Estudo dirigido

1. O que fazer com uma sequência de DNA?
2. Bancos de dados públicos e internacionais: GenBank, ENA, DDBJ;
3. NCBI; EMBL; DDBJ;
4. Sequências completas de genomas de organismos dos três domínios;
5. Definição de Bioinformática;
6. Análise da sequência no GenBank;
7. Busca de sequências por similaridade;
8. BLAST e Banco de dados de sequências.

DOGMA CENTRAL

Informação da sequência do genôma

GENOMA



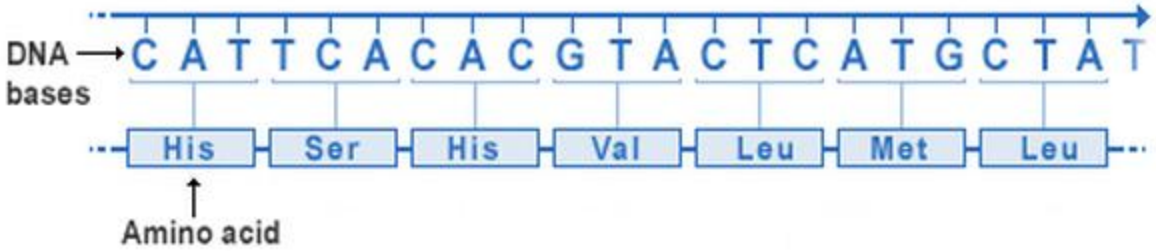
Informação da expressão gênica
mRNA; cDNA; EST. RNAseq

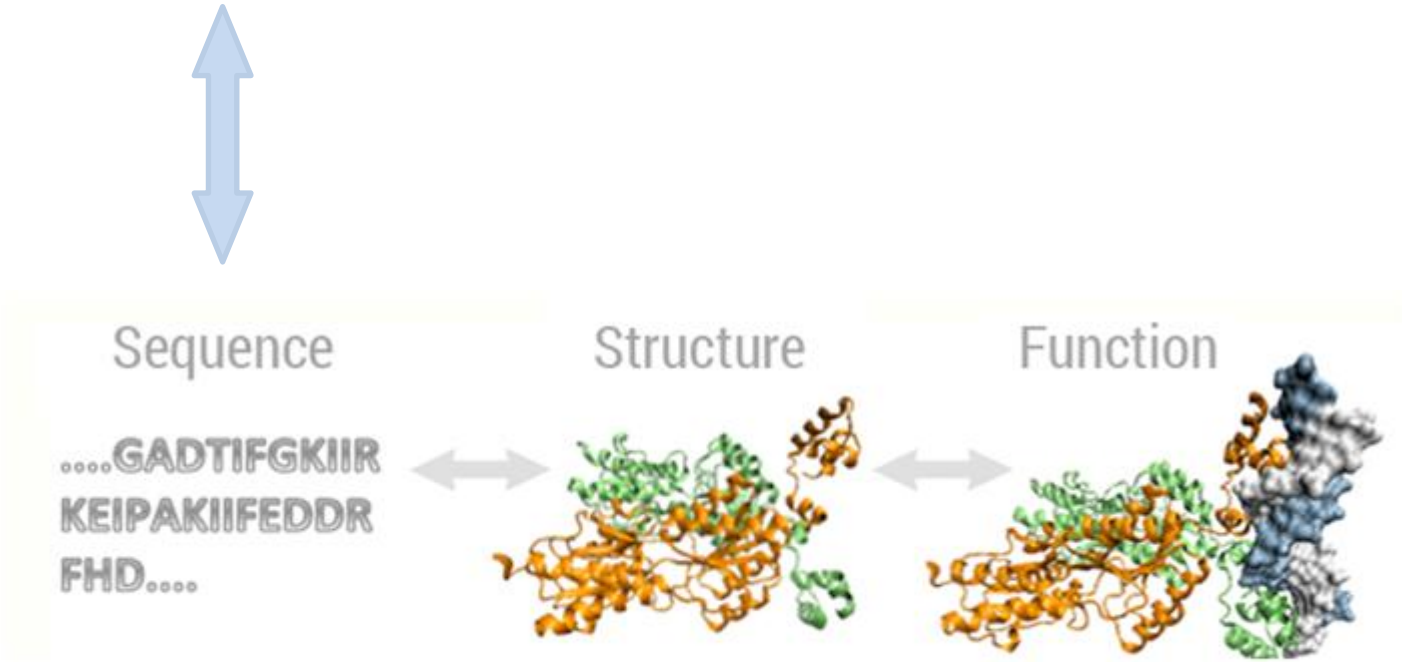
TRANSCRIPTOMA

Informação proteína: inferência
ou sequenciamento direto

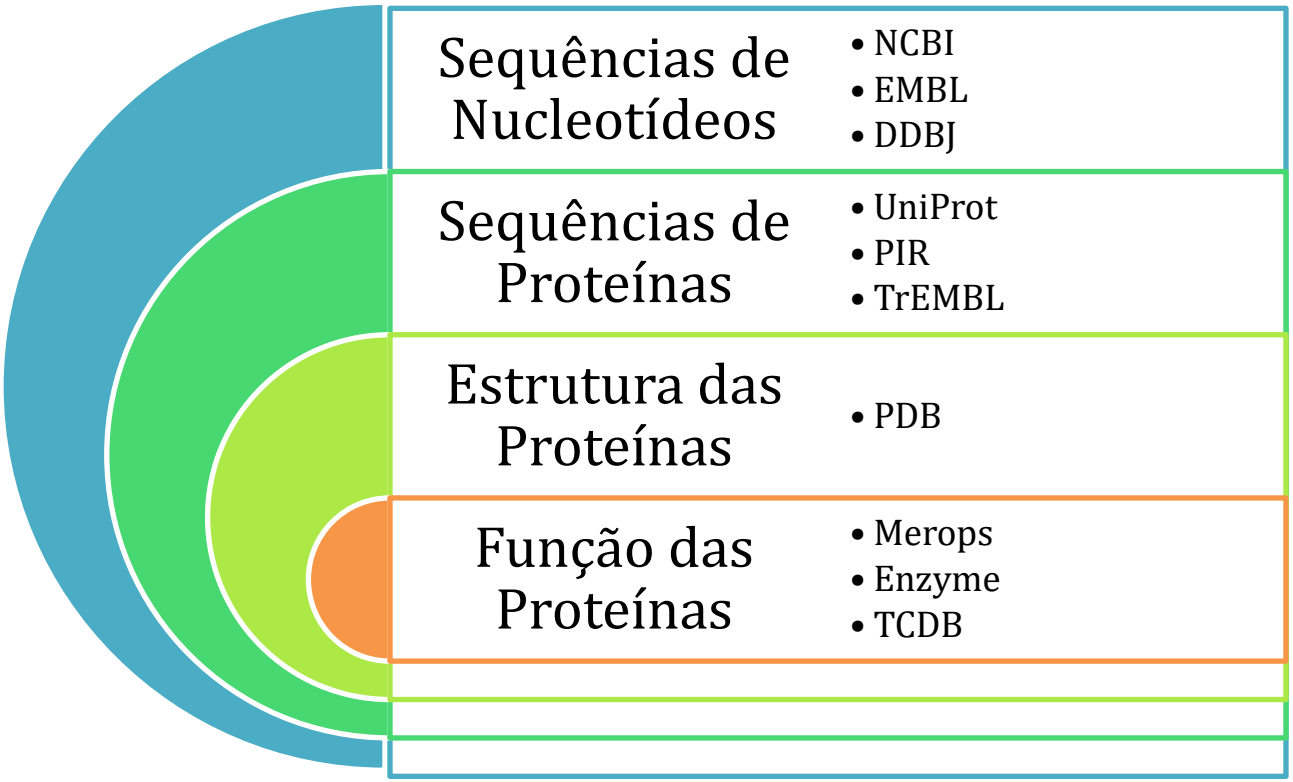
PROTEOMA

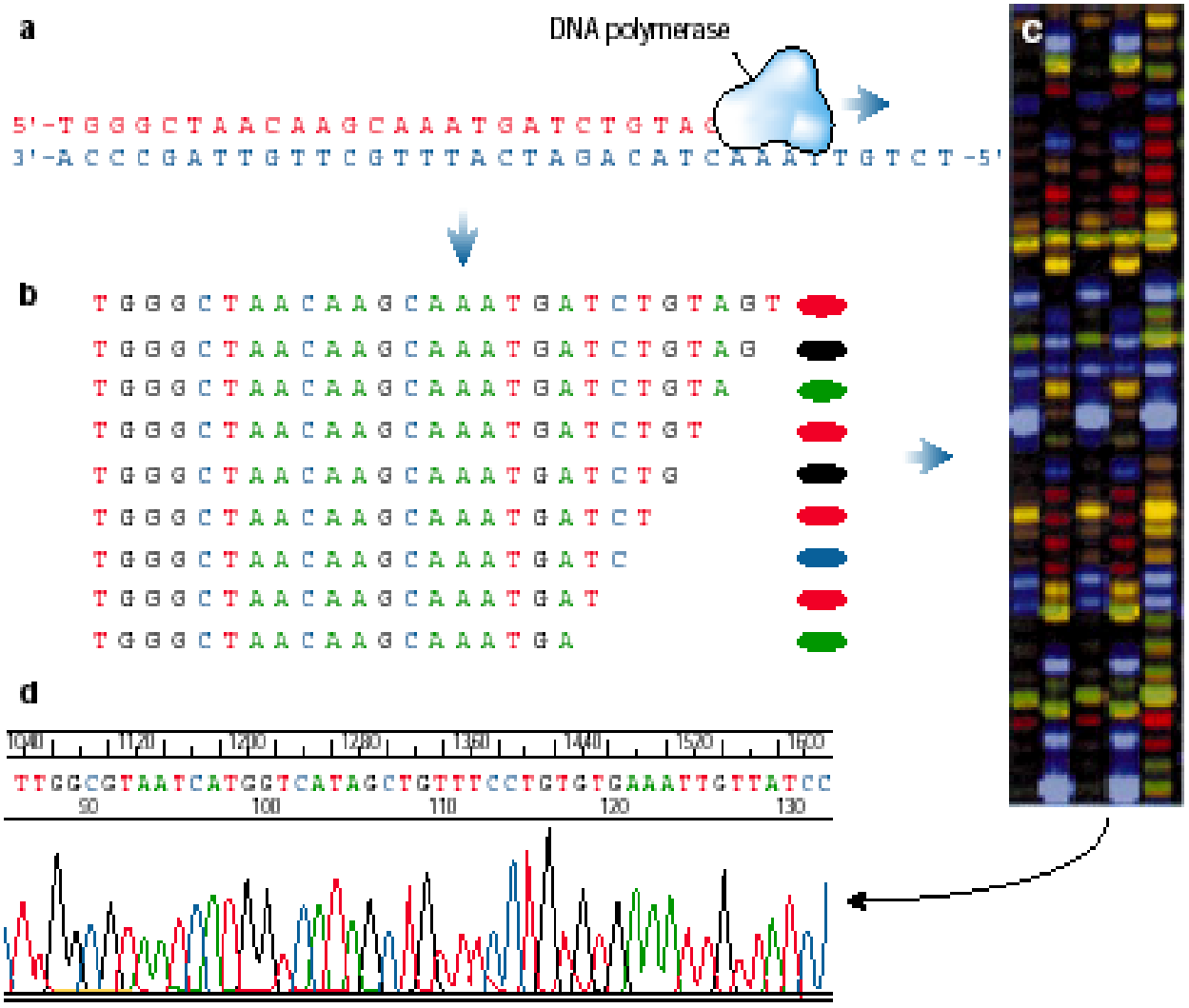
INFORMAÇÕES DE FUNÇÃO


DNA bases
C A T T C A C A C G T A C T C A T G C T A T
His
Ser
His
Val
Leu
Met
Leu
Amino acid


Sequence
Structure
Function
....GADTIFGKIIR
KEIPAKIIFEDDR
FHD....

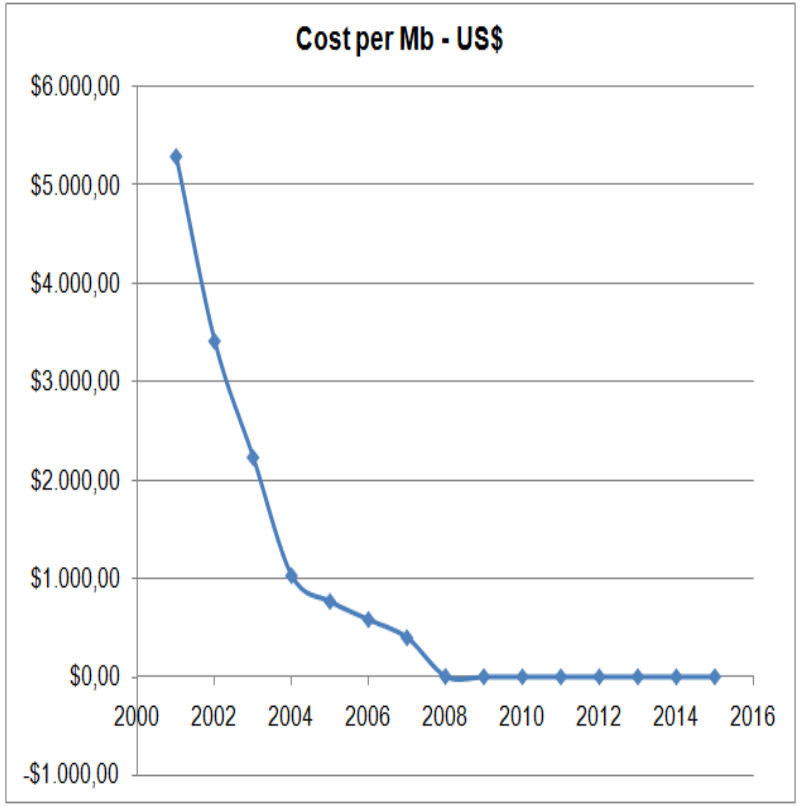
Bancos de dados biológicos




a
DNA polymerase
5'-TGGGCTAACAAGCAAATGATCTGTAG
3'-ACCCGATTGTTCGTTTACTAGACATCAAATTGTCT-5'
b
TGGGCTAACAAGCAAATGATCTGTAGT
TGGGCTAACAAGCAAATGATCTGTAG
TGGGCTAACAAGCAAATGATCTGTA
TGGGCTAACAAGCAAATGATCTGT
TGGGCTAACAAGCAAATGATCTG
TGGGCTAACAAGCAAATGATCT
TGGGCTAACAAGCAAATGATC
TGGGCTAACAAGCAAATGAT
TGGGCTAACAAGCAAATGA
c
d
1040
1120
1200
1280
1360
1440
1520
1600
TTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCC
50
100
110
120
130

Sequencing costs

Date	Cost per Mb	Cost per Genome
2001	$5.292,39	$95.263.072
2002	$3.413,80	$61.448.422
2003	$2.230,98	$40.157.554
2004	$1.028,85	$18.519.312
2005	$766,73	$13.801.124
2006	$581,92	$10.474.556
2007	$397,09	$7.147.571
2008	$3,81	$342.502
2009	$0,78	$70.333
2010	$0,32	$29.092
2011	$0,09	$7.743
2012	$0,07	$6.618
2013	$0,06	$5.096
2014	$0,06	$5.731
2015	$0,014	$1.245



NIH: https://www.genome.gov/sequencingcostsdata/

2.365,5 trilhões de bases
18 trilhões de sequências

Sanger Sequencing 1977

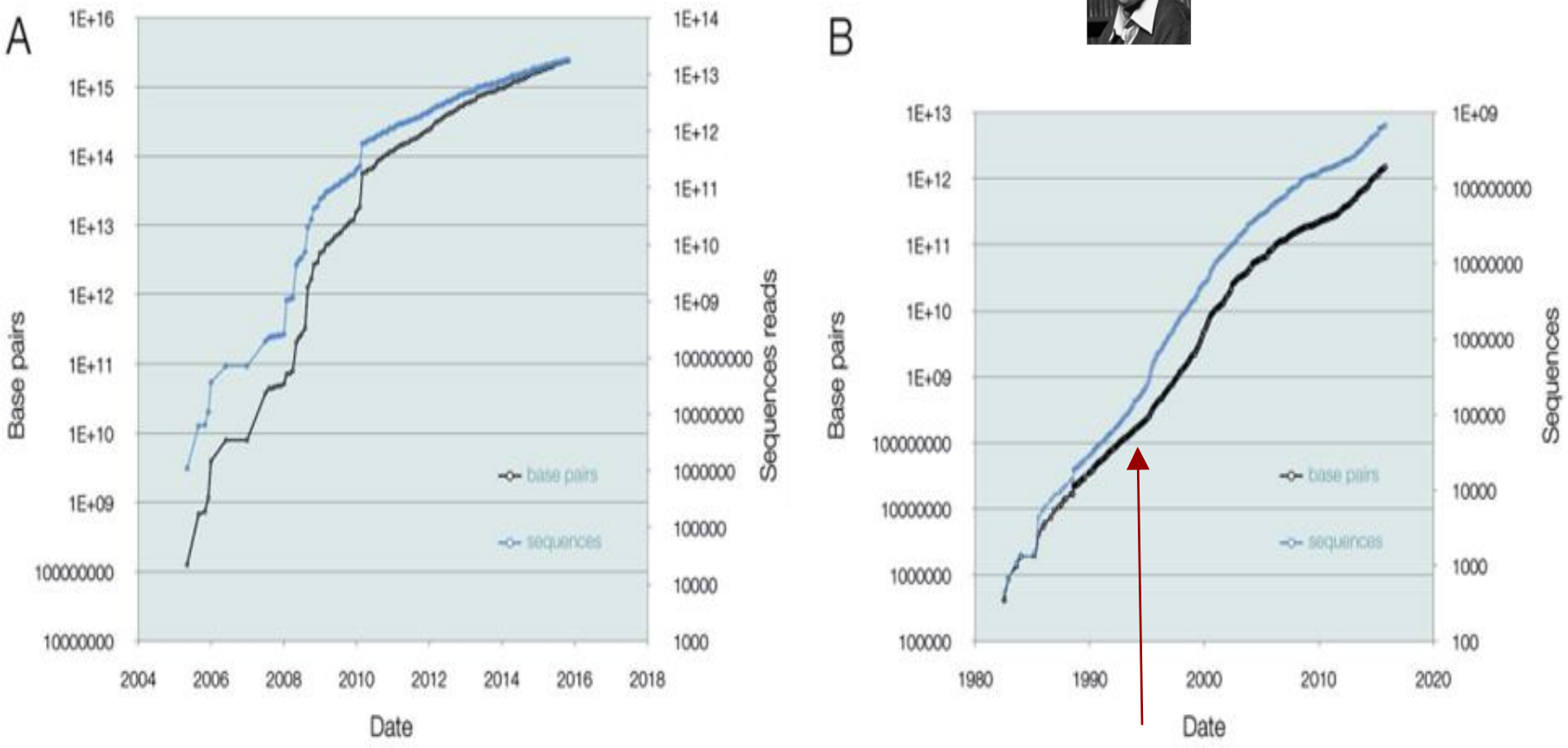
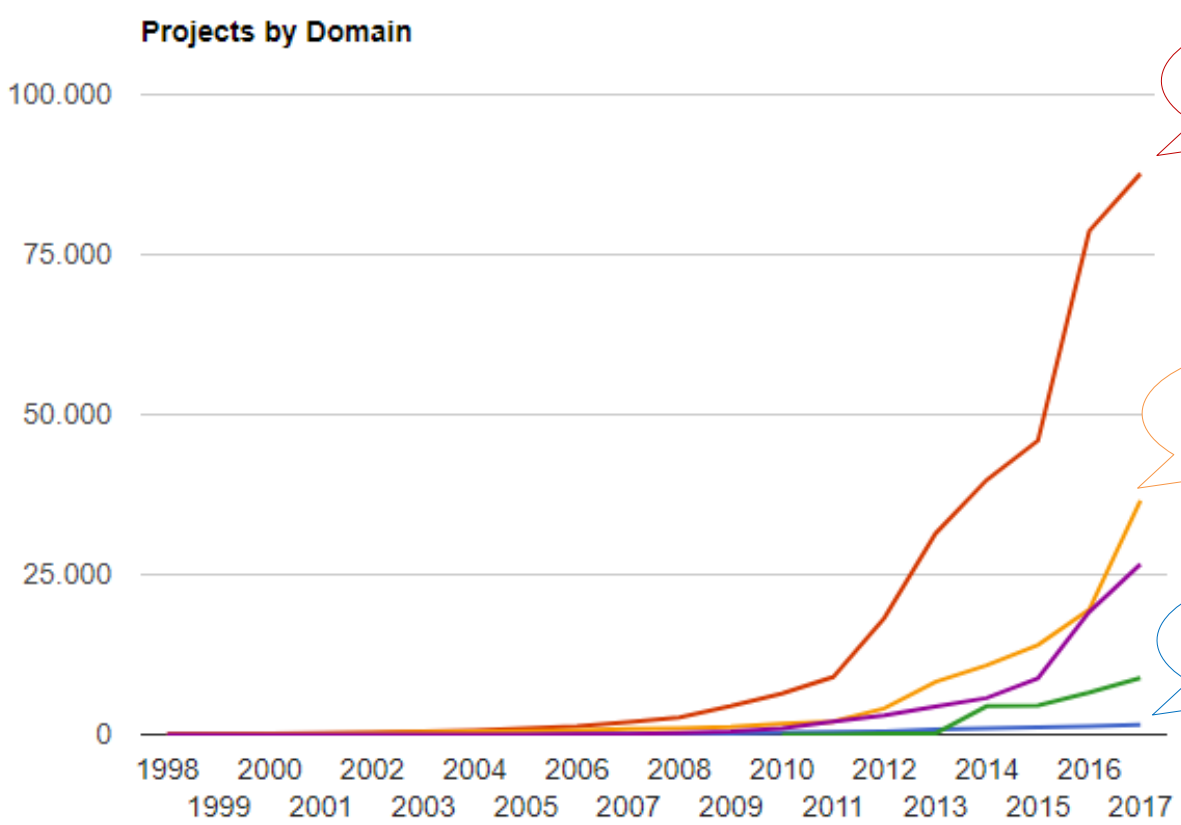


Figure 1. Cumulative growth in INSDC. (**A**) Base pairs (black, 2365.5 trillion) and sequence reads (blue, 17.8 trillion) for INSDC raw data. (**B**) Base pairs (black 1449 billion) and sequences (blue, 651.5 million) in INSDC assembled/annotated data.

Nucleic Acids Research, 2016, Vol. 44, Database issue doi: 10.1093/nar/gkv1323

Total de Projetos no GOLD (domínios da vida)



BACTERIA :
87 651

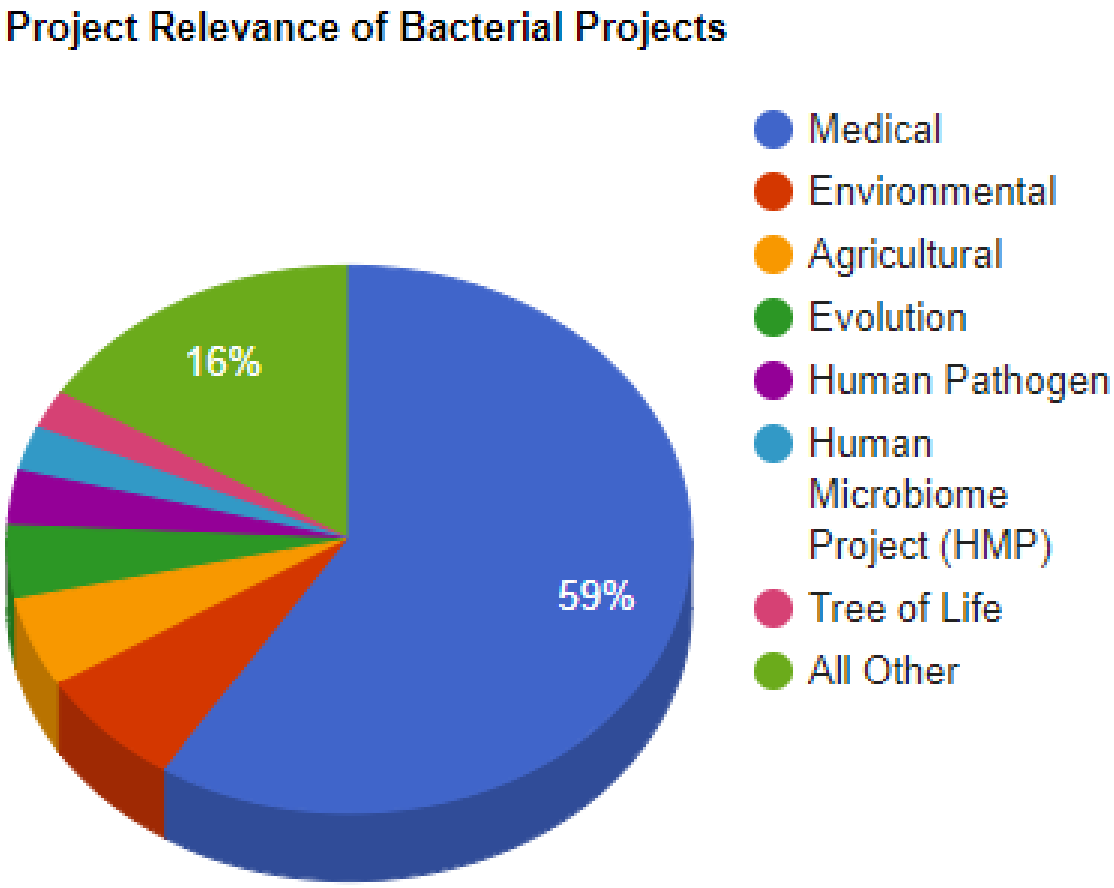
EUKARYA :
36 577

ARCHAEA:
1 560

Vírus:
8 844

Metagenômica:
26 642

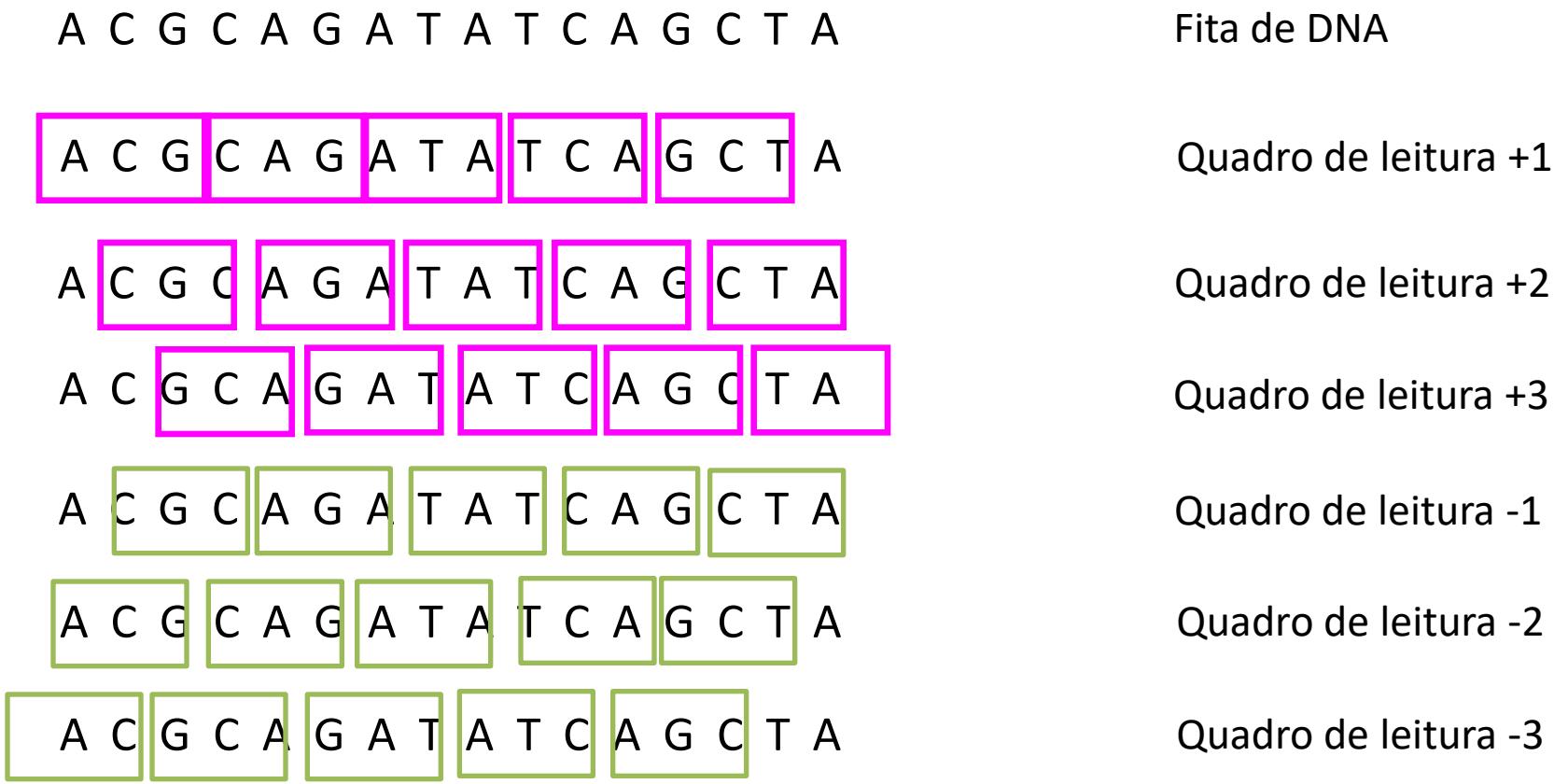
https://gold.jgi-psf.org/statistics

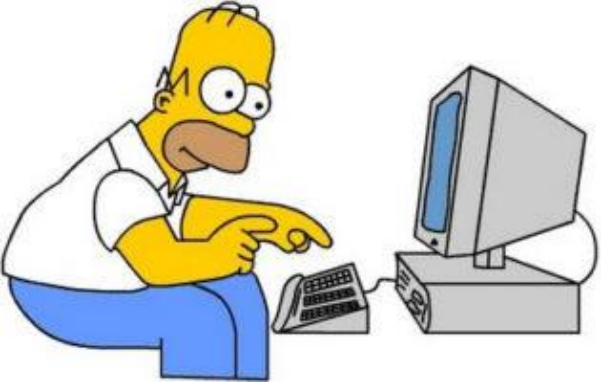

Project Relevance of Bacterial Projects
16%
59%
Medical
Environmental
Agricultural
Evolution
Human Pathogen
Human Microbiome Project (HMP)
Tree of Life
All Other

TTCATACTTGGTTAAGACCTTTACAAGCCGACCAACGTGGTGACAGTGTCGTCCTTTA
CGCACCGAATCCCTTTATCATTGAATTAGTAGAAGAGCGATACTTAGGACGTCTTCGG
ATGGAATCTTGGTCCCGTTGCCTGGAACGTCTTGAAACTGAATTCCCGCCAGAAGATG
TTCATACTTGGTTAAGACCTTTACAAGCCGACCAACGTGGTGACAGTGTCGTCCTTTA
CGCACCGAATCCCTTTATCATATTGAATTAGTAGAAGAGCGATACTTAGGACGTCTTC
GGGAATTGTTATCCTATTTCTCAGGAATACGTGAAGTAGTCCTTGCAATTGGCTCACG
ACCTAAAACAACAGAACTACCCGTACCAGTAGACACTACAGGACGTTTGTCTTCAACA
GTCCCATTTAACGGAAATCTCGACACACACTATAACTTTGATAATTTTGTTGAGGGAC
GAAGCAATCAACTCGCTCGTGCTGCAGCTTGGCAAGCGGCACAGAAACCGGGAGACCG
TACTCACAACCCTCTATTGCTCTATGGTGGGACTGGTTTGGGTAAAACCCATTTAATG
TTTGCTGCAGGTAACGTAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATCTTC
GTTCGGAACAGTTTTTCAGCGCCATGATAAGAGCGTACAAGATAAAAGTATGGATCAT
AAGGGTAAAACCCATTTAATGTTTGCTGCAGGTAACGTAATGCGGCAAGTAAACCCAA
CTTATAAAGTAATGTATCTTCGTTCGGAACAGTTTTTCAGCGCCATGATAAGAGCGTA
CAAGATAAAAGTATGGATCATAAGGGTAAAACCCATTTAATGTTTGCTGCAGGTAACG
TAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATCTTCGTTCGGAACAGTTTTT
CAGCGCCATGATAAGAGCGTACAAGATAAAAGTATGGATCATAAAACGTAATGCGGCA
AGTAAACCCAACTTATAAAGTAATGTATCTTCGTTCGGAACAGGGTAAAACCCATTTA
ATGTTTGCTGCAGGTAACGTAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATC
TTCGTTCGGAACAGTTTTTCAGCGCCATGATAAGAGCGTACAAGATAAAAGTATGGAT
CATAAAACGTAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATCTTCGTTCGGA
ACAAAAACGTAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATCTTCGTTCGGA
AG

Reverse genetics:

https://en.wikipedia.org/wiki/Reverse_genetics


A C G C A G A T A T C A G C T A
Fita de DNA
A C G C A G A T A T C A G C T A
Quadro de leitura +1
A C G C A G A T A T C A G C T A
Quadro de leitura +2
A C G C A G A T A T C A G C T A
Quadro de leitura +3
A C G C A G A T A T C A G C T A
Quadro de leitura -1
A C G C A G A T A T C A G C T A
Quadro de leitura -2
A C G C A G A T A T C A G C T A
Quadro de leitura -3



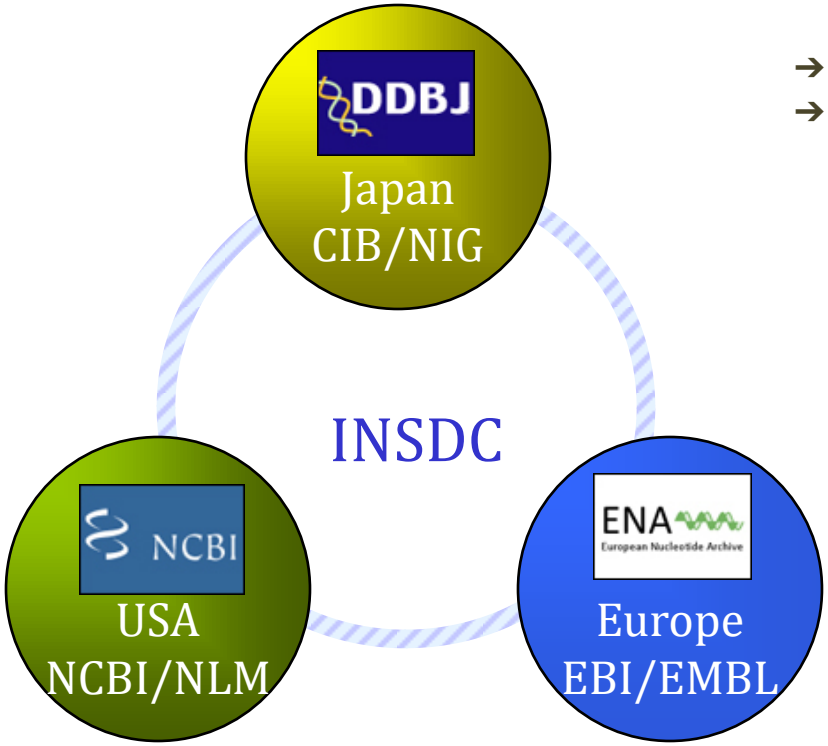
Bioinformática

*A bioinformática consiste no desenvolvimento de métodos computacionais, matemáticos e estatísticos para **organizar** e **analisar** informações biológicas em grande escala e de maneira integrada.*

Organização e Armazenamento → - Bancos de Dados Biológicos

Visualização e Análise →
- Ferramentas computacionais
- Montagem de genomas
- Compreensão do significado biológico

Bancos de dados



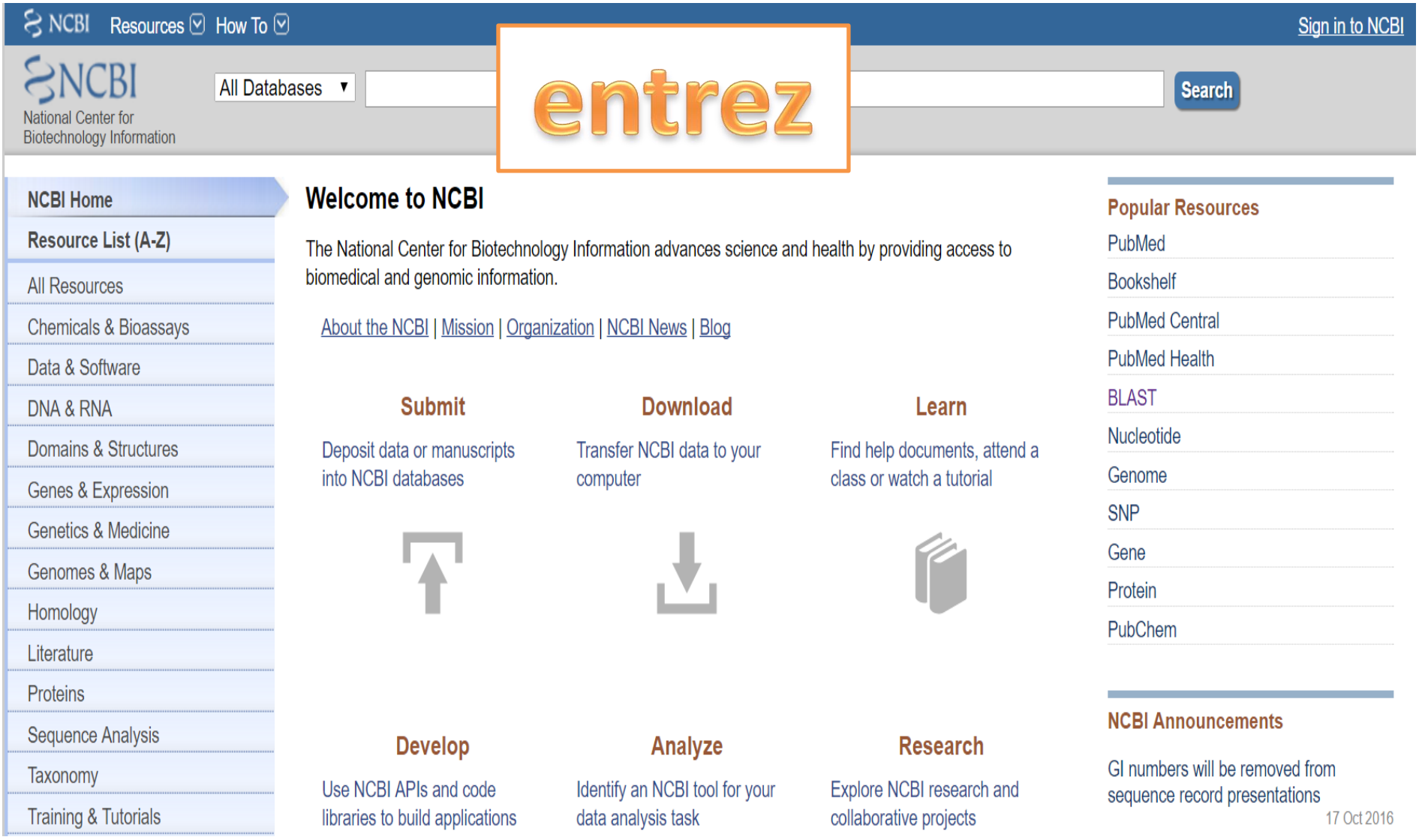
http://www.insdc.org/

- 667.903 espécies representadas (UniProtKB/TrEMBL 2016_07)
- 15.536 genomas (Eukarya, Archeae, Bacteria) (GOLD database)

- *National Center for Biotechnology Information* (NCBI)
  - *National Institutes of Health* (NIH)
  - Maryland, EUA
- *European Molecular Biology Laboratory* (EMBL)
  - *European Bioinformatics Institute* (EBI)
  - Hinxton, Inglaterra
- *DNA Data Bank of Japan* (DDBJ)
  - *Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ)*
  - Mishima, Japão

National Center for Biotechnology Information (NCBI)

http://www.ncbi.nlm.nih.gov/



HOME PAGE | MY TIMES | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

The New York Times

Dining & Wine

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

FASHION & STYLE | **DINING & WINE** | HOME & GARDEN | WEDDINGS

Creamy, Healthier Ice Cream? What's the Catch?

By JULIA MOSKIN
Published: July 26, 2006

SIGN IN TO E-MAIL THIS
PRINT
SINGLE PAGE
REPRINTS

IN its quest to create ice cream as voluptuous as butter and as virtuous as broccoli, the ice cream industry has probed the depths of the Arctic Ocean, studied the intimate structures of algae and foisted numerous failures on the American public.

Tony Cenicola/The New York Times

SCOOP OF SCIENCE Companies are using new methods to make ice cream a guiltless pleasure.

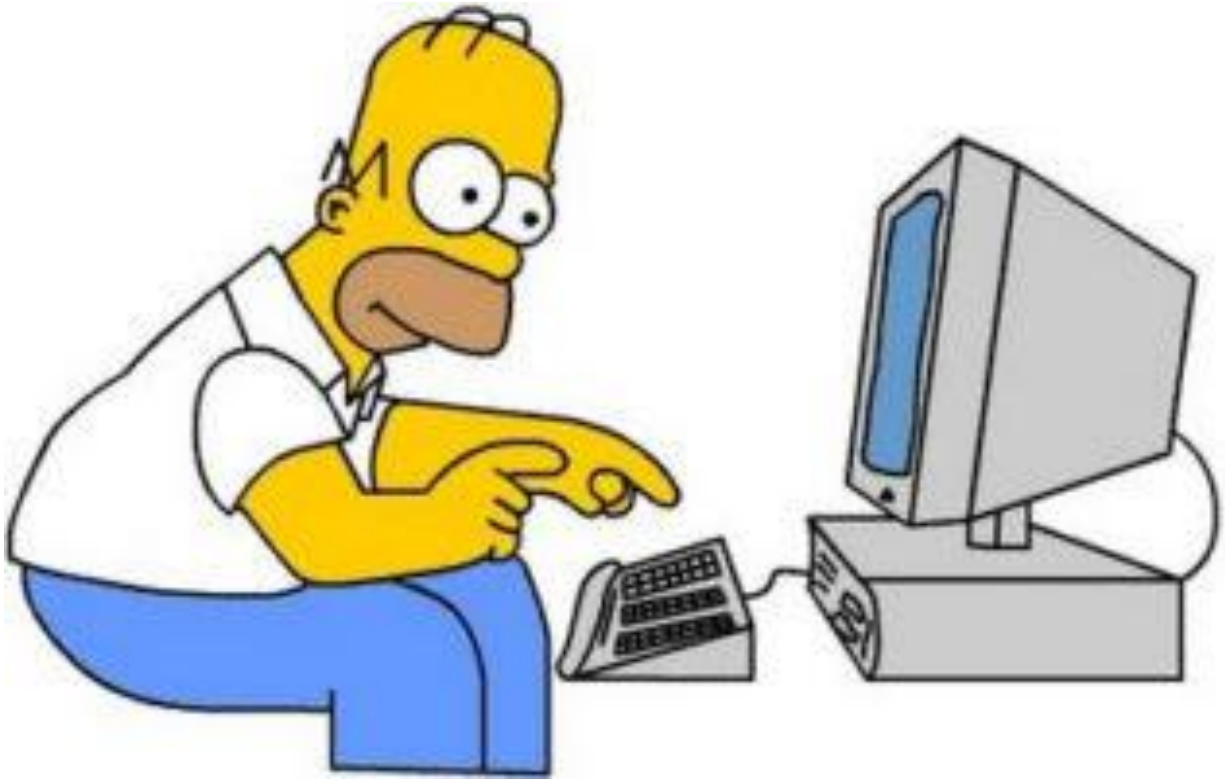
"I have tried them all as they came down the pike: dairy-free, fat-free, sugar-free; with tofu, yogurt, rice, whatever," said Linda Calhoun, a teacher who lives near Flagstaff, Ariz., cataloguing the disappointments she has tasted over the years. "They always make me sad."

For Americans who spend each summer wrestling with temptation, there is fresh hope in the freezer case. New industrial processes, including one that involves a protein

Ocean pout vive em regiões polares

http://www.unilever.com/innovation/productinnovations/coolicecreaminnovations/

http://academicsreview.org/reviewed-content/genetic-roulette/section-7/part-7-3/



http://www.ncbi.nlm.nih.gov/nuccore/X07506?

```
LOCUS       X07506                  1095 bp    DNA     linear   VRT 14-NOV-2006
DEFINITION  Winter flounder antifreeze protein gene (AFP).
ACCESSION   X07506
VERSION     X07506.1  GI:64211
KEYWORDS    antifreeze protein.
SOURCE      Pseudopleuronectes americanus (winter flounder)
  ORGANISM  Pseudopleuronectes americanus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Euteleostei; Neoteleostei;
            Acanthomorpha; Acanthopterygii; Percomorpha; Pleuronectiformes;
            Pleuronectoidei; Pleuronectidae; Pleuronectinae;
            Pseudopleuronectes.
```

```
REFERENCE   1  (bases 1 to 1095)
  AUTHORS   Scott,G.K., Davies,P.L., Kao,M.H. and Fletcher,G.L.
  TITLE     Differential amplification of antifreeze protein genes in the
            pleuronectinae
  JOURNAL   J. Mol. Evol. 27 (1), 29-35 (1988)
   PUBMED   3133486
FEATURES             Location/Qualifiers
     source          1..1095
                     /organism="Pseudopleuronectes americanus"
                     /mol_type="genomic DNA"
```

```
     CAAT signal     26..29
     TATA signal     77..83
     misc feature    109
                     /note="pot. transcription initiation region"
     gene            order(158..213,710..902)
                     /gene="AFP"
     CDS             join(158..213,710..902)
                     /gene="AFP"
                     /codon_start=1
                     /product="antifreeze protein"
                     /protein_id="CAA30389.1"
                     /db_xref="GI:64212"
                     /db_xref="GOA:P04002"
                     /db_xref="InterPro:IPR000104"
                     /db_xref="PDB:1ATF"
                     /db_xref="PDB:1J5B"
                     /db_xref="PDB:1WFA"
                     /db_xref="PDB:1WFB"
                     /db_xref="UniProtKB/Swiss-Prot:P04002"
                     /translation="MALSLFTVGQLIFLFWTMRITEARPDPAAKAAPAAAAAPAAAAP
                     DTASDAAAAAALTAANAKAAAELTAANAAAAAAATARG"
     intron          214..709
                     /gene="AFP"
                     /number=1
     polyA signal    969..974
```

polyA signal 969..974

ORIGIN

```
       1 gcacaacact gggggagtgt tgtaccaatc tgctcagatt ggtcgacagt caagcgatga
      61 cccaggctcc agttactata aaacagattc acattgacct ggatattcac cacatcttca
     121 ttttgtagtg aaccagtgct ccctacaagt tctcaaaatg gctctctcac ttttcactgt
     181 cggacaattg attttcttat tttggacaat gaggtacgtg aacactcact ttgtttcttc
     241 tatgaatctg gttttactgt aaatatcttg gaaggaagga aggatatctg cattatcccc
     301 gaggggccat ttgttttaca gccagcggtg aaagatgaag atcttcatcc gtgttcatct
     361 gtttgaccct gattaacaca agatggtcac atggaccatc tttatttaca taatgtttca
     421 tcagcacttc ctgttttcag cccgaaactt aaagaggcct catggaaact tcctgatgat
     481 ctggtgacac ctgctggttg aaggaaacag agtttgagag gcggcagaaa aaattatttt
     541 agtttgaatg aagaagctgt catttgattt catgttgggg gggggggggg tcatcacaca
     601 cagatattga taactgtcat cactgagttt ggtgaaagtg acggaccagt aaatgttgtg
     661 atatataata ttatcataat aattataata ataccattaa tctctgcaga atcactgaag
     721 ccagacccga ccccgcagcc aaagccgccc cagcagcagc tgccgccccc gccgcagccg
     781 ccccagacac cgcctctgac gccgccgctg cagccgccct taccgccgcc aacgccaaac
     841 ccgctgccga actcactgcc gccaacgccg ccgccgccgc agcagccacc gccagaggtt
     901 aaggatcgtg gtcgtcttga tgtgggatca tgtgaacatc tgagcagcga gatgttacca
     961 atctgcttaa taaacctgag aagctgattg ttaaaaacca agtgtcctgt tcatttcatc
    1021 tctgaaagtc cgtcacagtt tctgtagatc atgtagactc caggaagtga tgccattgtg
    1081 ctgttgaacc tgcag
//
```

FORMATO FASTA??

NUCLEOTÍDEOS

```
>gi|47933333|gb|AY262820.1| Pinus radiata cellulose synthase (CesA10) mRNA,
complete cds
GCACGAGGATTTAATCGAACTCGGTAATTGTTATCATCGTGGTGAGGACTAGTGCTTGATATTTTAGTTT
TATTCTCGAAATTTCATAATAGCTTGGGCTTTCTAAAAAGGGGAATGGTGGAATGGGTGTGAGAGTGAAG
AGGAATGGTATCGAACCACTAAGAAAAGTAGTCGTGCAAGTATTAGATGGTTGGCTGTGATAGTTGGAAA
```

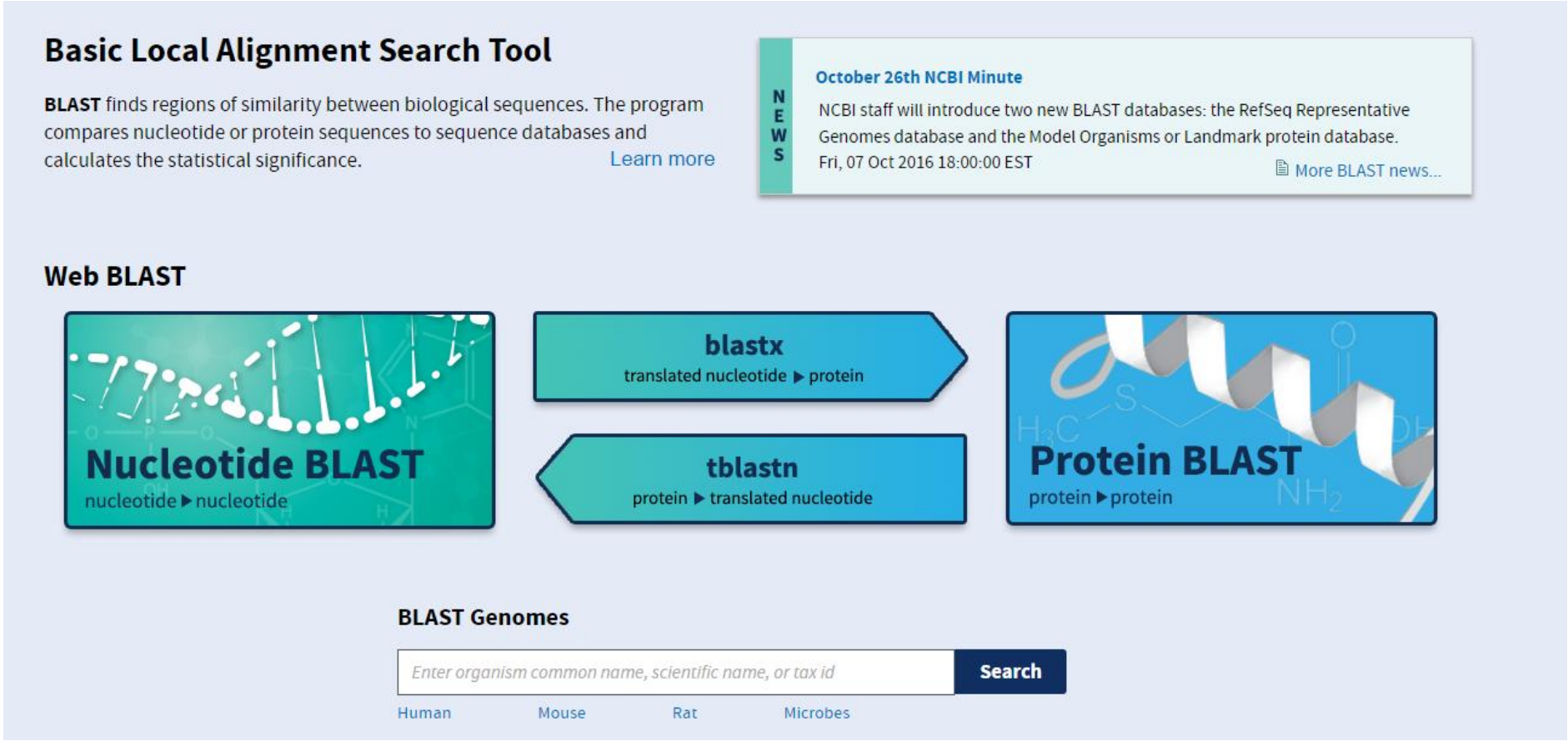
PROTEÍNAS

```
>gi|47933334|gb|AAQ63935.1| cellulose synthase [Pinus radiata]
MEARTNTAAGSNKRNVRVSVRDDGELGPKPPQHINSHICQICGEDVGLAADGEFFVACNECAFPVCRPCY
EYEWKDGNQSCPQCKTRYKWHKGSPQVDGDKEDECADDLDHDFNSTQGNRNEKQQIAEAMLHWQMAYGRG
EDVGPSRSESQELPQLQVPLITNGQAISGELPAGSSEYRRIAAPPTGGGSGKRVHPLPFPDSTQTGQVRA
```

>LINHA DO NOME

MÁXIMA DE 80 CARACTERES POR LINHA

https://blast.ncbi.nlm.nih.gov/Blast.cgi



BLAST (NCBI)

BLASTP: compara uma sequência de entrada (**proteína**) com um banco de dados (**proteínas**).

BLASTN: compara uma sequência de entrada (**DNA**) a um banco de dados (**DNA**).

BLASTX:compara uma sequência de nucleotídeos, traduzida em todas os 6 frames, em um banco de dados de proteínas.

TBLASTN: compara uma sequência de proteína com um banco de dados de nucleotídeos traduzido em 6 frames.

BLAST (NCBI)

QUIZ

Formato FASTA: formato universalmente aceito para se processado

```
>sequência 1
PQLVALGLALLCAVAGPAAAQNCGCQPNVCCSKFGYCGTTDEYCGDGCQSGPCRSGRGGGGSGGGGANVA
SVVTSSFFNGIKNQAGSGCEGKNFYTRSAFLSAVKGYPGFAHGGSQVQGKREIAAFFAHATHETGHFCYI
SEINKSNAYCDPTKRQWPCAAGQKYYGRGPLQISWNYNYGPAGRAIGFDGLGDPGRVARDAVVAFKAALW
FWMNSVHGVVPQGFGATTRAMQRALECGGNNPAQMNARVGYYRQYCRQLGVDPGPNLTC


>sequência 2
ATGTTAGATACTAATAAAGTTTATGAAATAAGCAATCTTGCTAATGGATTATATACATCAACTTATTTAA
GTCTTGATGATTCAGGTGTTAGTTTAATGAGTAAAAAGGATGAAGATATTGATGATTACAATTTAAAATG
GTTTTTATTTCCTATTGATAATAATCAATATATTATTACAAGCTATGGAGCTAATAATTGTAAAGTTTGG
AATGTTAAAAATGATAAAATAAATGTTTCAACTTATTCTTCAACAAACTCTGTACAAAAATGGCAAATAA
AAGCTAAAGATTCTTCATATATAATACAAAGTGATAATGGAAAGGTCTTAACAGCAGGAGTAGGTGAATC
TCTTGGAATAGTACGCCTAACTGATGAATTTCCAGAGAATTCTAACCAACAATGGAATTTAACTCCTGTA
CAAACAATTCAACTCCCACAAAAACCTAAAATAGATGAAAAATTAAAAGATCATCCTGAATATTCAGAAA
CCGGAAATATAAATCCTAAAACAACTCCTCAATTAATGGGATGGACATTAGTACCTTGTATTATGGTAAA
TGATTCAGGAATAGATAAAAACACTCAAATTAAAACTACTCCATATTATATTTTTAAAAAATATAAATAC
TGGAATCTAGCAAAAGGAAGTAATGTATCTTTACTTCCACATCAAAAAAGATCATATGATTATGAATGGG
GTACAGAAAAAAATCAAAAAACATCTATTATTAATACAGTAGGATTGCAAATTAATATAGATTCAGGAAT
GAAATTTGAAGTACCAGAAGTAGGAGGAGGTACAGAAGACATAAAAACACAATTAACTGAAGAATTAAAA
GTTGAATATAGCACTGAAACCAAAATAATGACGAAATATCAAGAACACTCAGAGATAGATAATCCAACTA
ATCAACCAATGAATTCTATAGGACTTCTTATTTATACTTCTTTAGAATTATATCGATATAACGGTACAGA
AATTAAGATAATGGACATAGAAACTTCAGATCATGATACTTACACTCTTACTTCTTATCCAAATCATAAA
GAAGCATTATTACTTCTCACAAACCATTCGTATGAAGAAGTAGAAGAAATAACAAAAATACCTAAGCATA
CACTTATAAAATTGAAAAAACATTATTTTAAAAAATAA
```

Exercício:

http://www.ncbi.nlm.nih.gov/nuccore/M63845.1

http://www.uniprot.org/uniprot/P0A370

1) Quantos nucleotídeos tem o gene que codifica a proteina?
2) Quantos nucleodídeos tem a região codante?
3) Qual a localização das regiões regulatórias?
4) Quantos exons e introns tem o gene?

Estudo dirigido

1. O que fazer com uma sequência de DNA?
2. Bancos de dados públicos e internacionais: GenBank, ENA, DDBJ;
3. NCBI; EMBL; DDBJ;
4. Sequências completas de genomas de organismos dos três domínios;
5. Definição de Bioinformática;
6. Análise da sequência no GenBank;
7. Busca de sequências por similaridade;
8. BLAST e Banco de dados de sequências.