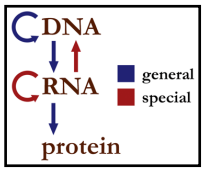
Instituto de Ciências Biomédicas / Universidade de São Paulo
Elisabete José Vicente



Introdução à Bioinformática - do DNA à proteína

CAPÍTULO 2: INTRODUÇÃO AOS BANCOS DE DADOS PRIMÁRIOS ("PRIMARY DATA BANKS")

Roteiro Teórico-prático

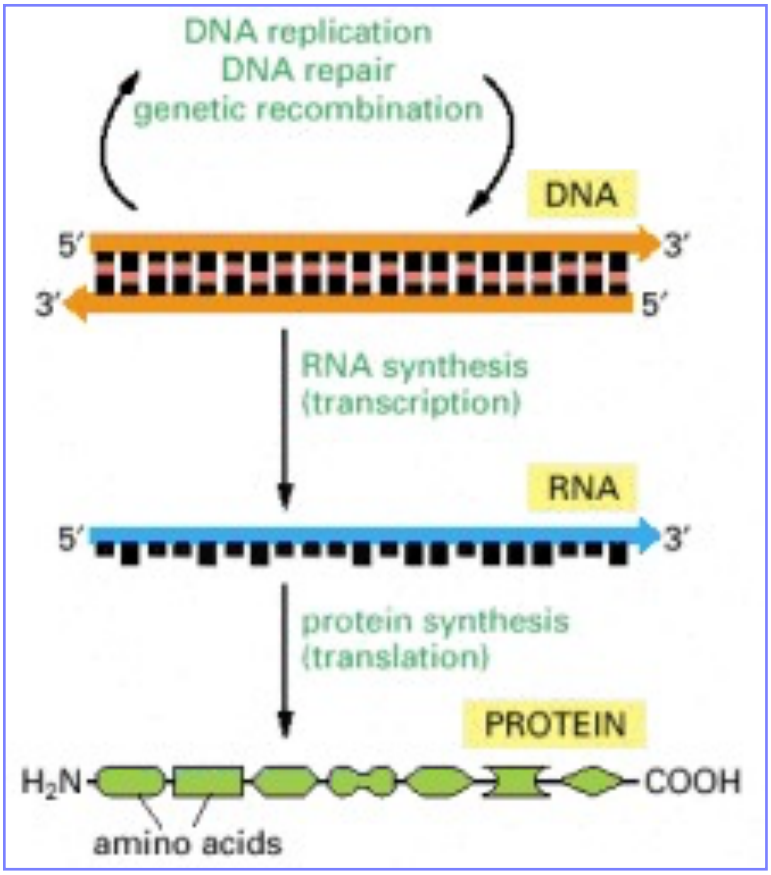


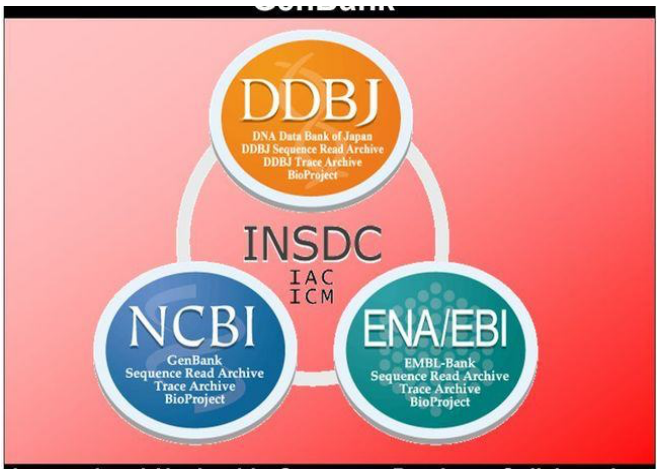
Fig 1: **Dogma central da Biologia:** O fluxo de informação genética do **DNA** para o RNA (**transcrição**) e do RNA para a **proteína** (**tradução**) ocorre em todas as células vivas. Copyright © 2002, Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter; Copyright © 1983, 1989, 1994, Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, and James D. Watson
(Fonte: https://www.ncbi.nlm.nih.gov/books/NBK21050/figure/A974/)

2020

2020.07.0

INTRODUÇÃO À BIOINFORMÁTICA – DO DNA À PROTEÍNA
CAPÍTULO 2: INTRODUÇÃO AOS BANCOS DE DADOS PRIMÁRIOS ("PRIMARY DATA BANKS")

1. Bancos de dados primários e sua importância em armazenar e disponibilizar dados de sequências disponíveis



Os bancos de dados primários (também conhecidos como repositórios ou depósitos de dados) são portais ("gateways") altamente organizados e amigáveis para os usuários consultarem uma enorme quantidade de dados biológicos produzidos por pesquisadores.

Nos anos 1980–1990, foram desenvolvidos os primeiros **bancos de dados primários de DNA e de proteínas**. Estes foram inicialmente desenvolvidos para o **armazenamento de sequências** que haviam sido experimentalmente determinadas. Naqueles tempos, as proteínas haviam sido sequenciadas um aminoácido por vez, e o sequenciamento de DNA estava em sua infância. Então, os bancos de dados continham um número limitado de sequências.

No entanto, com a chegada do **sequenciamento automático de DNA**, esses bancos de dados começaram a crescer exponencialmente. Atualmente, as submissões de sequências são feitas por laboratórios individuais, bem como "a granel" por centros de sequenciamento em todo o mundo, e as submissões de DNA agora excedem em muito o número de submissões de sequências de proteínas. Assim, atualmente, a **maioria das sequências de proteínas encontradas em bancos de dados é o produto da tradução conceitual dos genes** e genomas determinados usando o sequenciamento de DNA.

Há **três bancos de nucleotídeos** ou **bancos de dados primários** para a submissão de sequências de **nucleotídeos** e **genoma**:

- GenBank hospedado pelo **National Center for Biotechnology Information (NCBI)** da "**National Library of Medicine" (nim)** dos "**National Institutes of Health (nih),** nos EUA;
- The "**European Nucleotide Archive"** ou ENA hospedado pelo "**European Molecular Biology Laboratories"** (**EMBL**, do "European Bioinformatics Institute" (EBI);
- The **DNA Data Bank of Japan** ou DDBJ hospedado pelo **National Centre for Genetics**.

Estes **três bancos dedados juntos** formam o **Internacional Nucleotide Sequence Database Collaboration** e, felizmente para todos, todos eles "espelham" uns aos outros. Ou seja, independentemente de onde uma sequência é depositada, a entrada aparecerá em todos os três bancos de dados. Uma vez que os dados são depositados em bancos de dados primários, eles podem ser acessados livremente por qualquer pessoa em todo o mundo.

Exemplo: Um grupo de pesquisadores identificam uma cepa de *Staphylococcus aureus* isolada de um paciente e, depois de algumas investigações, suspeitam que essa cepa possa ser geneticamente diferente daquelas previamente identificadas. Eles decidem sequenciá-la

e, depois de comparar as sequências de DNA já disponíveis nos bancos públicos (das cepas até então "conhecidas"), concluem que, de fato, este novo isolado é diferente. A comunidade de pesquisa se beneficiará de ter essa nova sequência no repositório público, de modo que na próxima vez que um pesquisador encontrar a mesma cepa, ele/ela será capaz de reconhecer se seu isolado é novo, ou está de alguma forma relacionado com cepas previamente sequenciadas.

O acúmulo de **conhecimento coletivo** em **bancos de dados públicos** permite o acesso rápido e eficiente aos dados por indivíduos e instituições. A **rápida identificação** de uma **cepa virulenta de patógeno microbiano com base em sua sequência** e o **compartilhamento de resultados e experiências entre pesquisadores e médicos** podem ajudar a colocar restrições no local para evitar a disseminação de um patógeno na comunidade. Em outras situações, a **identificação** correta do patógeno causador da doença pode auxiliar no tratamento e escolha dos antibióticos, possibilitando uma melhor e mais rápida solução para o tratamento da doença.

2. O formato dos arquivos GenBank ("the GenBank file format")

O formato de um arquivo **Genbank,** além d**e** permitir o armazenamento da **sequência** de DNA/proteína, também permite o **armazenamento de sua informação**.

Assim, o formato **Genbank** contém muito mais informações do que o formato **FASTA**. Formatos semelhantes ao **Genbank** foram desenvolvidos pelo **ENA** (**formato EMBL**) e pelo **DDBJ** (**formato DDBJ**).

Já discutimos como as sequências de proteínas e DNA são representadas de uma maneira que nos permite salvar essa sequência em um arquivo de computador para posterior referência ou manipulação. Vamos agora investigar como informações adicionais sobre uma sequência podem ser armazenadas de maneira sistemática e controlada. O objetivo dessa maneira altamente organizada de fornecer dados adicionais é tornar possível que essas informações sejam padronizadas para a interpretação humana e possam ser tratadas por muitos diferentes programas de computador.

O formato **FASTA** discutido anteriormente é provavelmente o mais simples de todos os formatos de arquivo de dados de sequências.

Embora possamos adicionar algum texto em seu cabeçalho (primeira linha indicada por ">"), em alguns casos pode ser necessário adicionar mais informações. Além disso, essa informação pode precisar ser categorizada para permitir sua diferenciação. Por exemplo, para discriminar um número que indica uma coordenada do genoma de um número que indica o comprimento de um gene.

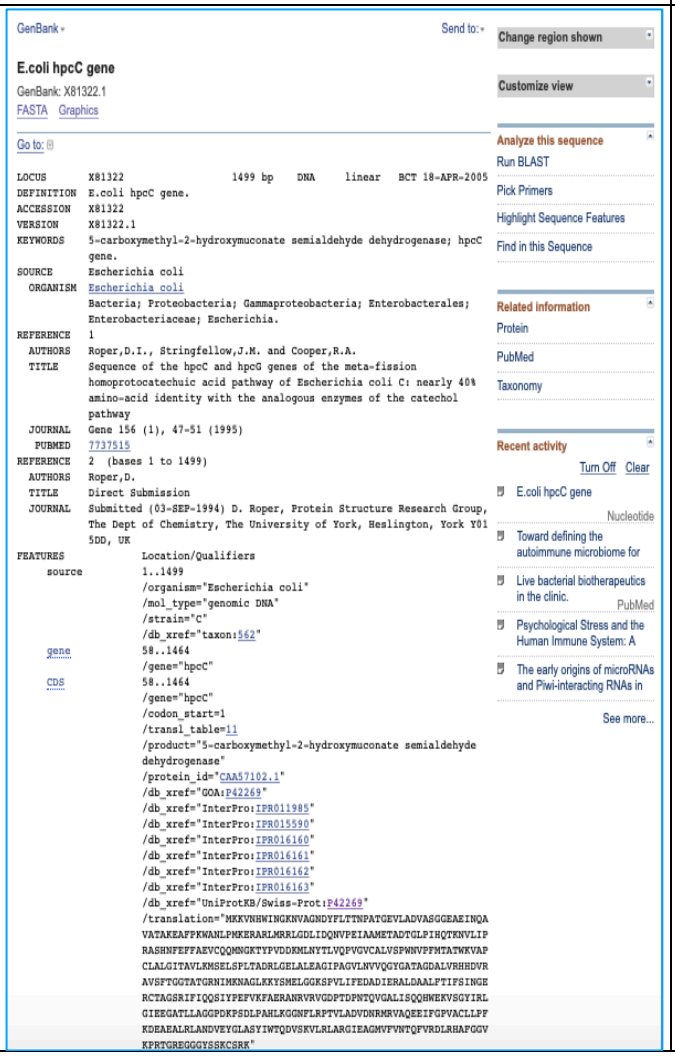
Os bancos de dados primários desenvolveram formatos de **arquivo de dados altamente estruturados** que permitem o armazenamento de todos esses dados adicionais que **acompanham a sequência de DNA**, que acompanham a sequência "nua" representada em um arquivo **FASTA**. O "estilo" ("layout") estrito é necessário para que o arquivo seja compatível com uma variedade de programas de computador. Cada um dos **três bancos de dados p**ri**mários possui seu próprio** "**layout" de formato de arquivo de sequência**. No entanto, todos eles contêm quase os mesmos campos e as mesmas informações, tornando-os intercambiáveis. Vale notificar que **existem muitos outros formatos de arquivo** que foram **personalizados para atender às finalidades específicas**.

Os formatos de arquivos que serão discutidos aqui armazenam informações adicionais relacionadas às sequências de DNA e de proteínas. Por simplicidade, será apresentado aqui penas o formato de arquivo de sequência **GenBank**. O formato **EMBL** será discutido nas seguintes atividades.

Use este link do **GenBank** para visualizar uma entrada para o **gene hpcC de *Escherichia coli*.** A primeira parte desta entrada também é fornecida abaixo. Como pode ser verificado, o formato **GenBank** apresenta cinco elementos ou partes essenciais (**Quadro 1**):
- A **primeira parte** da entrada (**A**) inclui **LOCUS**, **DEFINITION**, **ACCESSION**, **VERSION**;
- A **segunda parte** da entrada (**B**) é identificada por **ORIGIN** e representada a sequência real.

Quadro 1: **Entrada no GenBank para o gene *hpcC* de *E. coli*:** A) Na **Primeira parte** do arquivo estão**: LOCUS DEFINITION**, **ACCESSION**, **VERSION ;** e **B)** Na **Segunda Parte** do arquivo está a **SEQUENCIA** de nucleotídeos do gene do inicio (ORIGIN = 1) até o ultimo nucleotídeo.

A.



GenBank Send to: Change region shown

E.coli hpcC gene

GenBank: X81322.1

FASTA Graphics

Go to:

```
LOCUS       X81322                1499 bp    DNA     linear   BCT 18-APR-2005
DEFINITION  E.coli hpcC gene.
ACCESSION   X81322
VERSION     X81322.1
KEYWORDS    5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase; hpcC
            gene.
SOURCE      Escherichia coli
  ORGANISM  Escherichia coli
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1
  AUTHORS   Roper,D.I., Stringfellow,J.M. and Cooper,R.A.
  TITLE     Sequence of the hpcC and hpcG genes of the meta-fission
            homoprotocatechuic acid pathway of Escherichia coli C: nearly 40%
            amino-acid identity with the analogous enzymes of the catechol
            pathway
  JOURNAL   Gene 156 (1), 47-51 (1995)
    PUBMED  7737515
REFERENCE   2  (bases 1 to 1499)
  AUTHORS   Roper,D.
  TITLE     Direct Submission
  JOURNAL   Submitted (03-SEP-1994) D. Roper, Protein Structure Research Group,
            The Dept of Chemistry, The University of York, Heslington, York Y01
            5DD, UK
FEATURES             Location/Qualifiers
     source          1..1499
                     /organism="Escherichia coli"
                     /mol_type="genomic DNA"
                     /strain="C"
                     /db_xref="taxon:562"
     gene            58..1464
                     /gene="hpcC"
     CDS             58..1464
                     /gene="hpcC"
                     /codon_start=1
                     /transl_table=11
                     /product="5-carboxymethyl-2-hydroxymuconate semialdehyde
                     dehydrogenase"
                     /protein_id="CAA57102.1"
                     /db_xref="GOA:P42269"
                     /db_xref="InterPro:IPR011985"
                     /db_xref="InterPro:IPR015590"
                     /db_xref="InterPro:IPR016160"
                     /db_xref="InterPro:IPR016161"
                     /db_xref="InterPro:IPR016162"
                     /db_xref="InterPro:IPR016163"
                     /db_xref="UniProtKB/Swiss-Prot:P42269"
                     /translation="MKKVNHWINGKNVAGNDYFLTTNPATGEVLADVASGGEAEINQA
                     VATAKEAFPKWANLPMKERARLMRRLGDLIDQNVPEIAAMETADTGLPIHQTKNVLIP
                     RASHNFEFFAEVCQQMNGKTYPVDDKMLNYTLVQPVGVCALVSPWNVPFMTATWKVAP
                     CLALGITAVLKMSELSPLTADRLGELALEAGIPAGVLNVVQGYGATAGDALVRHHDVR
                     AVSFTGGTATGRNIMKNAGLKKYSMELGGKSPVLIFEDADIERALDAALPTIFSINGE
                     RCTAGSRIFIQQSIYPEFVKFAERANRVRVGDPTDPNTQVGALISQQHWEKVSGYIRL
                     GIEEGATLLAGGPDKPSDLPAHLKGGNFLRPTVLADVDNRMRVAQEEIFGPVACLLPF
                     KDEAEALRLANDVEYGLASYIWTQDVSKVLRLARGIEAGMVFVNTQPVRDLRHAPGGV
                     KPRTGREGGGYSSKCSRK"
```

Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence

Related information
Protein
PubMed
Taxonomy

Recent activity
Turn Off Clear
E.coli hpcC gene
Nucleotide
Toward defining the autoimmune microbiome for
Live bacterial biotherapeutics in the clinic.
PubMed
Psychological Stress and the Human Immune System: A
The early origins of microRNAs and Piwi-interacting RNAs in
See more...

B.

```
ORIGIN
        1 gaagtagaag gcgtgggccg cctggtgaac cgaattgttg agtgaggaaa cagcgaaatg
       61 aaaaaagtaa atcattggat caacggcaaa aatgttgcag gtaacgacta cttcctgacc
      121 accaatccgg caacgggtga agtgctggcg gatgtggcct ctggcggtga agcggagatc
      181 aatcaggcgg tagcgacagc gaaagaggcg ttcccgaaat gggccaatct gccgatgaaa
      241 gagcgtgcgc gcctgatgcg ccgtctgggc gatctgatcg accagaacgt gccagagatc
      301 gccgcgatgg aaaccgcgga cacgggcctg ccgatccatc agaccaaaaa tgtgttgatc
      361 ccacgcgctt ctcacaactt tgaatttttc gcggaagtct gccagcagat gaacggcaag
      421 acttatccgg tcgacgacaa gatgctcaac tacacgctgg tgcagccggt aggcgtttgt
      481 gcactggtgt caccgtggaa cgtgccgttt atgaccgcca cctggaaggt cgcgccgtgt
      541 ctggcgctgg gcattaccgc ggtgctgaag atgtccgaac tctccccgct gaccgctgac
      601 cgcctgggtg agctggcgct ggaagccggt attccggcgg gcgttctgaa cgtggtacag
      661 ggctacggcg caaccgcagg cgatgcgctg gtccgtcatc atgacgtgcg tgccgtgtcg
      721 ttcaccggcg gtacggcgac cgggcgcaat atcatgaaaa acgccgggct gaaaaaatac
      781 tccatggaac tgggcggtaa atcgccggtg ctgatttttg aagatgccga tattgagcgc
      841 gcgctggacg ccgccctgtt caccatcttc tcgatcaacg gcgagcgctg caccgccggt
      901 tcgcgcatct ttattcaaca aagcatctac ccggaattcg tgaaatttgc cgaacgcgcc
      961 aaccgtgtgc gcgtgggcga tccgaccgat ccgaataccc aggttggggc gcttatcagc
     1021 cagcaacact gggaaaaagt ctccggctat atccgtctgg gcattgaaga aggcgccacc
     1081 ctgctggcgg gcggcccgga taaaccgtct gacctgcctg cacacctgaa aggcggcaac
     1141 ttcctgcgcc caacggtgct ggcggacgta gataaccgta tgcgcgttgc ccaggaagag
     1201 attttcgggc cggtcgcctg cctgctgccg tttaaagacg aagccgaagc gttacgcctg
     1261 gcaaacgacg tggagtatgg cctcgcgtcg tacatctgga cacaggatgt cagcaaagtg
     1321 ctgcgtctgg cgcgcggcat tgaagcaggc atggtgttcg tcaacaccca gttcgtgcgt
     1381 gacctgcgcc acgcatttgg cggcgtaaaa cctcgcaccg ggcgtgaagg cggtggatac
     1441 agttcgaagt gttcgcggaa atgaagaaga acgtctgcat tccatggcgg accatccca
//
```

Obtém-se as informações:

LOCUS: x81322
DEFINITION: E.coli hpcC gene
ACCESSION: x81322
VERSION: x81322.1
ORIGIN: 1..... //

No restante das seções são adicionadas informações que, embora importantes, não são essenciais e podem estar faltando. Note que não poderíamos carecer do campo LOCUS para um arquivo do GenBank, ou ele não poderia ser reconhecido como tal arquivo.

As partes não essenciais da entrada contêm o que é comumente conhecido como **metadados**, e podem incluir informações mais detalhadas sobre o organismo, referências cruzadas a outros bancos de dados e até mesmo uma lista de publicações nas quais essa entrada é incluída.

A parte **FEATURES** da entrada descreve características importantes da entrada, como presença de sequências de codificação, proteínas, etc. Esta seção é menos "amigável" ao homem, e pode conter campos que não fazem sentido para o olho destreinado. Mas não se preocupe, essas partes são principalmente destinadas a serem lidas por um programa de computador.

Finalmente, no final do arquivo, encontramos a **sequência real** que **poderia ser** de **DNA** ou de **proteína**. Note que a **última linha** da entrada tem um **"//".** Esses **dois caracteres** são muito importantes e indicam o **final da entrada/arquivo**. Embora possa estar claro para você que esse é o fim do arquivo porque não há mais nada abaixo, os programas de computador precisam ser instruídos quando devem parar de ler.

Assim, dentro de um arquivo, há uma riqueza de informações, desde a sequência de nucleotídeos do genoma até as publicações relacionadas a dessa entrada do genoma e as referências cruzadas a outros bancos de dados.

Questão 1:

Faça o download do arquivo do mesmo gene analisado acima (**GenBank**, *hpcC* de *E. coli*), mas agora no banco de dados do **European Nucleotide Archive – European Bioinformatics Institute (EMBL-EBI ENA),** formato **EMBL.**

Para facilitar, use o link abaixo para fazer o download do arquivo da sequencia equivalente a entrada *E. coli* X81322 no formato **EMBL**. Compare os formatos **EMBL** e **GenBank**.

```
ID   X81322; SV 1; linear; genomic DNA; STD; PRO; 1499 BP.
XX
AC   X81322;
XX
DT   15-SEP-1994 (Rel. 41, Created)
DT   18-APR-2005 (Rel. 83, Last updated, Version 4)
XX
DE   E.coli hpcC gene
XX
KW   5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase; hpcC gene.
XX
OS   Escherichia coli
OC   Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
OC   Enterobacteriaceae; Escherichia.
XX
RN   [1]
RX   DOI; 10.1016/0378-1119(95)00082-H.
RX   PUBMED; 7737515.
RA   Roper D.I., Stringfellow J.M., Cooper R.A.;
RT   "Sequence of the hpcC and hpcG genes of the meta-fission homoprotocatechuic
RT   acid pathway of Escherichia coli C: nearly 40% amino-acid identity with the
RT   analogous enzymes of the catechol pathway";
RL   Gene 156(1):47-51(1995).
XX
RN   [2]
RP   1-1499
RA   Roper D.;
RT   ;
RL   Submitted (03-SEP-1994) to the INSDC.
RL   D. Roper, Protein Structure Research Group, The Dept of Chemistry, The
RL   University of York, Heslington, York Y01 5DD, UK
XX
DR   MD5; ff367c259d405558292d651deac8730f.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1499
FT                   /organism="Escherichia coli"
FT                   /strain="C"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:562"
FT   CDS             50..1464
FT                   /transl_table=11
FT                   /gene="hpcC"
FT                   /product="5-carboxymethyl-2-hydroxymuconate semialdehyde
FT                   dehydrogenase"
FT                   /db_xref="GOA:P42269"
FT                   /db_xref="InterPro:IPR011985"
FT                   /db_xref="InterPro:IPR015590"
FT                   /db_xref="InterPro:IPR016160"
FT                   /db_xref="InterPro:IPR016161"
FT                   /db_xref="InterPro:IPR016162"
FT                   /db_xref="InterPro:IPR016163"
FT                   /db_xref="InterPro:IPR029510"
FT                   /db_xref="UniProtKB/Swiss-Prot:P42269"
FT                   /protein_id="CAA57102.1"
FT                   /translation="MKKVNHWINGKNVAGNDYFLTTNPATGEVLADVASGGEAEINQAV
FT                   ATAKEAFPKWANLPMKERALMERLGDLIDQNVPEIAAMETADTGLPIHQTKNVLIPRA
FT                   SHNFEFPAEVCQQMNGKTYFVDDKMLNYTLVQPVGVCALVSPWNVPFMTATWKVAPCLA
FT                   LGITAVLKMSELSPLTADRLGELALEAGIPAGVLNVVQGYGATAGDALVRHHDVRAVSF
FT                   TGGTATGRNIMKNAGLKKYSMELGGKSPVLIFEDADIERALDAALFTIFSINGERCTAG
FT                   SRIFIQQSIYPEFVKFAERANRVRVGDPTDPNTQVGALISQQHWEKVSGYIRLGIEEGA
FT                   TLLAGGPDKPSDLPAHLKGGNFLRPTVLADVDNRMRVAQEEIFGPVACLLPFKDEAEAL
FT                   RLANDVEYGLASYIWTQDVSKVLRLARGIEAGMVFVNTQFVRDLRHAFGGVKPRTGREG
FT                   GGYSSKCSRK"
```

```
ORIGIN
        1 gaagtagaag gcgtgggccg cctggtgaac cgaattgttg agtgaggaaa cagcgaaatg
       61 aaaaaagtaa atcattggat caacggcaaa aatgttgcag gtaacgacta cttcctgacc
      121 accaatccgg caacgggtga agtgctggcg gatgtggcct ctggcggtga agcggagatc
      181 aatcaggcgg tagcgacagc gaaagaggcg ttcccgaaat gggccaatct gccgatgaaa
      241 gagcgtgcgc gcctgatgcg ccgtctgggc gatctgatcg accagaacgt gccagagatc
      301 gccgcgatgg aaaccgcgga cacgggcctg ccgatccatc agaccaaaaa tgtgttgatc
      361 ccacgcgctt ctcacaactt tgaatttttc gcggaagtct gccagcagat gaacggcaag
      421 acttatccgg tcgacgacaa gatgctcaac tacacgctgg tgcagccggt aggcgtttgt
      481 gcactggtgt caccgtggaa cgtgccgttt atgacccgca cctggaaggt cgcgccgtgt
      541 ctggcgctgg gcattaccgc ggtgctgaag atgtccgaac tctccccgct gaccgctgac
      601 cgcctgggtg agctggcgct ggaagccggt attccggcgg gcgttctgaa cgtggtacag
      661 ggctacggcg caaccgcagg cgatgcgctg gtccgtcatc atgacgtgcg tgccgtgtcg
      721 ttcaccggcg gtacggcgac cgggcgcaat atcatgaaaa acgccgggct gaaaaaatac
      781 tccatggaac tgggcggtaa atcgccggtg ctgatttttg aagatgccga tattgagcgc
      841 gcgctggacg ccgccctgtt caccatcttc tcgatcaacg gcgagcgctg caccgccggt
      901 tcgcgcatct ttattcaaca aagcatctac ccggaattcg tgaaatttgc cgaacgcgcc
      961 aaccgtgtgc gcgtgggcga tccgaccgat ccgaataccc aggttggggc gcttatcagc
     1021 cagcaacact gggaaaaagt ctccggctat atccgtctgg gcattgaaga aggcgccacc
     1081 ctgctggcgg gcggcccgga taaaccgtct gacctgcctg cacacctgaa aggcggcaac
     1141 ttcctgcgcc caacggtgct ggcggacgta gataaccgta tgcgcgttgc ccaggaagag
     1201 attttcgggc cggtcgcctg cctgctgccg tttaaagacg aagccgaagc gttacgcctg
     1261 gcaaacgacg tggagtatgg cctcgcgtcg tacatctgga cacaggatgt cagcaaagtg
     1321 ctgcgtctgg cgcgcggcat tgaagcaggc atggtgttcg tcaacaccca gttcgtgcgt
     1381 gacctgcgcc acgcatttgg cggcgtaaaa cctcgcaccg ggcgtgaagg cggtggatac
     1441 agttcgaagt gttcgcggaa atgaagaaga acgtctgcat tccatggcgg accatccca
//
```

Entrada da sequência *E. coli* X81322 no formato **EMBL** obtida na base de dados **EMBL-EBI ENA**.

Questão 2:

Identifique as principais diferenças dos **formatos** das informações obtidas quando se faz a entrada da sequência de *E. coli*_ X81322 na base de dados **GenBank** (do NCBI) e na Base de dados do **EMBL**.

- Pense: por que existem essas diferenças?
- Compartilhe suas descobertas e discuta comparando com as de os seus colegas.

Confira se você concorda com as Respostas:

As diferenças entre **GenBank** e **EMBL** são:

1- **GenBank** é mais específico e simples.

2- EMBL é mais detalhado que o **GenBank**.

3- A ordem das informações é diferente. Ex. no **GenBank DEFINITION** vem antes de **ACCESSION**, ou no **EMBL DEFINITION** vem depois que **ACCESSION.**

4- No **EMBL** são usadas abreviaturas, mas no **GenBank** abreviaturas não são usadas.

5- O **GenBank** e o **EMBL** nomeiam as sequências de DNA ORIGEM, mas no **EMBL** chama-se “SEQUENCE” e os números escrevem-se no lado direito e no **GenBank** no lado esquerdo.

3. Comparando os formatos das sequências nos bancos GenBank, EMBL, DDJ

Cada um dos três repositórios de sequência de DNA (ou seja, cada um dos três bancos de **dados de nucleotídeos primários**) tem formato de arquivo ligeiramente diferente. No geral, eles são muito semelhantes e contêm as mesmas informações, mas são organizados de maneiras diferentes.

Veja como os formatos diferem comparando a mesma entrada apresentada nos três formatos diferentes:

Escherichia coli Ent plasmid P307 in GenBank format

Escherichia coli Ent plasmid P307 in EMBL format

Escherichia coli Ent plasmid P307 in DDBJ format

Suas Observações:

O alto nível de estrutura ajuda os biólogos e bioinformáticos na organização dos dados. Além disso, a formatação rigorosa desses arquivos é essencial para tarefas computacionais.

.

Questão 3:
Identifique **duas vantagens** e **duas desvantagens** de usar um nível tão alto de estruturação dos bancos de dados.

Questão 4:
Imagine o que aconteceria se cada um dos campos abaixo estivesse ausente de um dos arquivos: Explique as consequências

a) LOCUS:
b) DEFINITION:
c) ACCESSION:
d) VERSION:
e) ORIGIN:

Respostas:

3-

4- a)

b)

c)

D

e)

4. Como recuperar sequências de DNA / proteína de repositórios públicos

No vídeo a seguir é demonstrado como baixar sequências de DNA e / ou de proteínas de dois repositórios públicos muito populares: o **NCBI ("National Center for Biotechnology Information")** e o **ENA** ("European Nucleotide Archive").

Por que precisamos saber disso?
Procurar por um dado gene ou proteína em um banco de dados é onde muitas pesquisas se iniciam.
Exemplo: Imagine que você está lendo um artigo de pesquisa ou um livro texto sobre **resistência a antibióticos** em uma cepa de ***E. coli*** específica. O texto informa que o gene ou proteína responsável pela resistência é chamado ***hpcC***. Você quer saber a sequência desse gene ou proteína. Como você faz para achar isso? De onde você tira essas informações?

Agora vamos aprender como recuperar uma **entrada de um gene de um repositório (banco de dados)**. Aqui, vamos usar dois repositórios:
- o **NCBI (**"National Center for Biotechnology Information") hospedado no **NIH** ("National Institute of Health") nos EUA, e
- o **ENA** ("European Nucleotide Archive"), na Europa.....

Neste vídeo, você aprenderá **onde** e **como** obter as sequências de genes e proteínas de um gene de interesse.

Se **não** desejar baixar o VIDEO,
siga as INTRUÇÕES
TRANSCRITAS DO VIDEO
no final deste Capítulo

VIDEO 1: Como obter uma sequência de DNA/proteína dos bancos de dados GenBank, EMBL
Clique no link ao lado para assistir ao vídeo.

Olá a todos, sou professora Elisabete Vicente, do Instituto de Ciências Biomédicas da USP (ICB/USP). ....

Questão 5:

Compare os arquivos **EMBL**, **GenBank** e **FASTA** - investigue e discuta.
Você terá agora a chance de testar suas habilidades bioinformáticas recém-adquiridas.

1. Use o número de acesso **CDS63663** para recuperar entradas gene/proteína do **NCBI** e do **ENA**. Faça o download do arquivo **GenBank** (ou GenPept) do **NCBI** e do arquivo **EMBL** (do **ENA)**.

2. Compare o conteúdo dos arquivos **GenBank/EMBL**. Eles são iguais ou diferentes? Se forem diferentes, o que é diferente entre neles?

3. Faça o download dos arquivos FASTA correspondentes para o mesmo número de acesso.

4. Compare o conteúdo dos arquivos FASTA. Eles são iguais ou diferentes? Se este último, o que é diferente entre eles?

- Link: Para entrada nos arquivos **Genbank** e **ENA**

- Dica: Observe que no **ENA** após a busca surge mais de uma correspondência, escolha aquela que represente um único gene/proteína).

- o que correu bem e o que não correu tão bem em suas buscas?

Questão 6

Para um determinado organismo, selecione todas as respostas que você acha corretas

a) Cada aminoácido é codificado por apenas um códon
b) Cada aminoácido é codificado por até dois códons diferentes
c) Cada códon pode codificar apenas um aminoácido
d) Em diferentes organismos, aminoácidos podem ser codificados por diferentes códons preferenciais

Questão 7

Selecione todas as afirmações que são verdadeiras. Selecione todas as respostas que você acha corretas:

a) O formato FASTA é adequado para sequências de DNA e proteínas
b) O formato FASTA é exclusivo para sequências de DNA
c) O formato FASTA é exclusivo para sequências de proteínas
d) o formato FASTA é adequado para sequências de RNA

Questão 8

Selecione todas as declarações verdadeiras sobre uma entrada do **GenBank**. Selecione todas as respostas que você acha corretas.

a) Uma entrada do GenBank possui uma seção dedicada aos dados da sequência.
b) Uma entrada típica do GenBank possui uma seção dedicada a dados sequenciais no formato FASTA.
c) Uma entrada do GenBank inclui a data de envio.
d) Um GenBank pode ou não incluir referências a um ou mais trabalhos publicados.

Respostas:
6-c, d 7-a, d; 8-a, c, d.

ANEXO 1: Transcrição do VIDEO 1

"Olá a todos, sou professora Elisabete Vicente, do Instituto de Ciências Biomédicas da USP (ICB/USP). ....
Hoje, vou demonstrar como recuperar dados de uma **entrada de um gene** desejado **de um** repositório **(banco de dados de bioinformática)**.

Neste vídeo, vamos demonstrar como usar dois repositórios:
- **NCBI (**"National Center for Biotechnology Information") hospedado no **NIH** ("National Institute of Health)" nos EUA, e
- **ENA** ("European Nucleotide Archive"), na Europa.

E,
o gene que vamos usar nessa demonstração é ***hpcC*** da bactéria ***E. coli.***

Então, vamos lá!
(baixando a sequência e nucleotídeos no site do **GenBank**), no formato GenBank)
Inicialmente, vamos começar entrando no site do **NCBI** digitando: **www.ncbi.nlm.nih.gov**. Na parte superior esquerda da página, há um ícone seletor. Neste momento, vamos usar o banco de dados **Nucleotide,** clique nesta opção**.** Na caixa de pesquisa, digite **E.coli hpcC**.
Agora, clique **Search**.

Surgem os resultados. Observe que em algumas das entradas surgem informações indesejadas, como: genomas completos, dados de outras bactérias, etc. Não estamos interessados nessas entradas, pois estamos interessados **apenas na entrada de um único gene**, que está disponibilizada em terceiro lugar; assim, clique nessa entrada. Agora, é mostrada a entrada do GenBank para o nosso gene. Observe que, do lado esquerdo, você tem **tags** como: **LOCUS**, **DEFINIÇÃO**, **ACESSION, VERSION**, **KEYWORDS, SOURCE**. Estes são preenchidos por códigos. **Então, aqui temos: o número de acesso**, a **definição do gene**, **algumas palavras-chave**. Também, temos links importantes para, por exemplo, o organismo *Escherichia coli*.

Também, temos links do **PUBMED** que permite obter mais informações sobre essa entrada. Não vamos fazer isso agora, mas você pode fazer isto em seu tempo. Mais abaixo, na parte inferior da página, está a sequência de DNA do gene. Note que a sequência da proteína, que é uma **tradução conceitual** deste gene, está logo acima da sequência do gene.

Agora, vamos "**baixar" essa entrada em um arquivo** em nosso computador. Para isso, vamos para o menu, no canto superior direito, "**Send To"** (Enviar para). Escolha o formato "**Complete Record"**; no item "**Choose Destination**", escolha "**File"**; e, escolha o formato "**GenBank"**. Então, clique em "**Create File**" (Criar Arquivo). Isso vai baixar automaticamente uma sequência na nossa pasta Download.

Para acessar essa sequência, precisamos abri-la em um editor de texto. Neste exemplo, estou trabalhando em um Mac, e o arquivo já é disponibilizado automaticamente. Mas você também terá softwares semelhantes em seus computadores **PC** ou **Linux**, basta encontrar o arquivo. Observe sua extensão "**.gb**".
Caso a extensão ".gb" tenha sido perdida, será preciso abrir o arquivo no **WordPad** do Windows. Para isto, abra no **WordPad** ou similar. Pode-se abrir o arquivo diretamente da pasta Downloads. Observe que o arquivo não aparece na lista. Isso ocorre porque a extensão não é necessariamente compatível com o WordPad, mas podemos optar por mostrar todos os documentos. Pronto, surge a nossa sequência. Então, clique na nossa sequência, e clique em "Open" (Abrir). E aqui está a nossa entrada. Observe que esse arquivo tem o mesmo formato que você observou no navegador da Web, mas alguns dos links foram removidos. Isso ocorre porque agora é um arquivo simples e contém apenas texto. Mas esta é uma boa maneira de manter um registro de sua sequência de interesse.

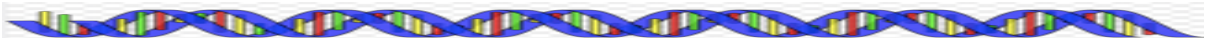
(Baixando a sequência e nucleotídeos no site do **GenBank**), no formato FASTA)
Agora vamos fechar esse arquivo, e vou mostrar como fazer o download de uma **sequência FASTA desse gene**. Para isto vamos usar o mesmo "menu download "**Send to**". Mas em vez de selecionar "**Complete Record**", escolha "**Coding Sequences**". No **download de formato**, há duas opções: podemos baixar o nucleotídeo, ou podemos baixar a proteína. Agora, vamos baixar a opção **nucleotídeo**. Clique em "**Create File**", e outro arquivo será baixado. Faça o mesmo procedimento anterior. Abra o arquivo no **WordPad** ou, caso estiva já estiver aberto, pode ser usado diretamente.

E daqui, indo para a pasta Downloads, repetindo este procedimento, pode ser obtido o documento. Esta é uma sequência **FASTA**. A primeira linha tem um símbolo seguido pelo nome da entrada, que é bastante longa neste caso. O texto desta linha pode ser bem pequeno conter somente o número de acesso e, em seguida, segue toda sequência.

(Baixando a sequência e AMINOACIDOS no site do **GenBank**), no formato FASTA)
Para baixar uma **sequência de proteínas**, podem ser repetidos os mesmos passos. Vá para o menu "**Send To**", "**coding sequence**", e escolha "**Protein file**". Este download, traz uma sequência de aminoácidos. Este é o nosso **terceiro arquivo** baixado. Como os dois anteriores, este também pode ser aberto no **WordPad** ou equivalente. E aqui está a nossa sequência da proteína de **E.coli hpcC**. Observe que a sequência é diferente desta vez. É de aminoácidos em vez de nucleotídeos.

(Baixando as mesmas sequências no site Europeu - **ENA**)
Vamos agora baixar a mesma sequência no site do **European Nucleotide Archive** (**ENA**). Então, vamos recuperar **a mesma entrada** genética de um **banco de dados diferente**. Para isso, vamos navegar para o site **www.ebi.ac.uk/ena**. Na pesquisa de texto, vamos digitar exatamente a mesma coisa - **E.coli hpcC**, e vamos clicar em "**Search**".
6:44
E este procedimento retorna dois resultados. Clique no link desejado. E aqui, temos uma entrada semelhante àquela que tínhamos no GenBank. Não fique confuso sobre o layout diferente dos bancos, pois a mesma informação está aqui. Clique em "**Text**" (Texto). Isto irá automaticamente baixar um arquivo. Podemos encontrar este arquivo no "**Finder"**. Este é o nosso arquivo. Da mesma forma que anteriormente, ele pode ser aberto com o WordPad.
7:15
Como pode ser visto, esta entrada é equivalente à entrada do GenBank que foi baixada anteriormente. Todas as mesmas informações estão aqui. É apenas apresentada ligeiramente diferente: em vez de ter todas as palavras, como "**ACCESSION**" ou "**PUBMED"**, há apenas dois códigos de letras, mas todas informações estão aqui. É apenas uma questão de costumar-se com a apresentação. no final, também está a sequência de nucleotídeos e a sequência de aminoácidos.
7:43
Para baixar apenas a sequência FASTA, use o link FASTA no final. O mesmo procedimento pode ser repetido, mas agora a extensão pode apresentar alguns problemas, e para abri-lo pode ser usado o WordPad para obter a sequência.
8:08
É possível alterar a extensão dos arquivos usando apenas a opção renomear no seu software.

Em resumo: Aqui demonstramos como obter a sequência de um gene e de uma proteína usando os bancos **NCBI** e **ENA**, e de como fazer o download destas sequências em seu computador.



Parabéns, você concluiu o Capítulo 2