Introdução aos bancos de dados biológicos

By Editor-in-chief

23 de junho de 2021

Introdução aos bancos de dados biológicos

Leandro Libório , Victor Hugo Resende

Revisão: Pedro Martins , Diego Mariano

BIOINFO – Revista Brasileira de Bioinformática. Edição #01. Julho, 2021.

DOI: [10.51780/978-6-599-275326-16](https://doi.org/10.51780/978-6-599-275326-16)

O conhecimento biológico é armazenado através de bancos de dados. O fundamento mais importante para a bioinformática aplicada é a coleta de dados de sequência e suas informações biológicas associadas [6]. O armazenamento de dados torna-se necessário diante ao aumento de volume gerado nos últimos anos e da necessidade de tratamento de informações através de buscas, processamentos e análises de resultados. Portanto, os bancos de dados biológicos utilizam aplicações e sistemas gerenciadores de bancos de dados para manter acessíveis e funcionais as informações de sequências e anotações genômicas, dados de complexos proteicos e suas interações, mapas metabólicos, catálogo de espécies ou recursos naturais, dentre outros exemplos.

Os bancos de dados biológicos têm grande utilidade para a comunidade científica. Existem diversas aplicações e recursos que funcionam como fontes de pesquisa e armazenamento. Espera-se encontrar uma série de benefícios diante ao uso destas aplicações tais como: a informação da localização de um gene, perspectivas de evolução de organismos, dados de expressão gênica e características fenotípicas, avaliação do uso de fármacos, interações moleculares e funcionais de proteínas, etc. Há uma lista considerável de possibilidades e desafios no tratamento de dados biológicos. Nos próximos tópicos serão abordados maiores detalhes referentes à classificação destes bancos de dados, de sua estrutura, funcionamento e exemplos de utilização.

Introdução

Embora o conceito de banco de dados esteja intrinsecamente ligado à Ciência da Computação, boa parte das atividades do nosso dia a dia envolvem organização e armazenamento de informações com um objetivo específico. Neste capítulo, iremos discutir alguns conceitos gerais de banco de dados e sua importância no contexto da bioinformática.

O que é um banco de dados?

Um banco de dados é uma coleção de dados relacionados que possuem um significado implícito. Desta maneira, os dados devem ser coerentes e populados nos registros com o intuito de atender alguma finalidade específica ligada a algum aspecto do mundo real. Múltiplos usuários finais realizam transações que devem ser atualizadas e refletidas no banco de dados [1]. Por exemplo: um site de vendas de produtos esportivos disponibiliza diversos itens em sua *home page*. O acesso ao site é realizado por inúmeros clientes que podem realizar suas compras, e consequentemente atualizam as informações do banco de dados deste vendedor. A manutenibilidade de estoque e oferta destes produtos é automatizada através de operações computacionais por um grupo de programas escritos para esta tarefa.

Sistema de Gerenciamento de Banco de Dados (SGBD)

O Sistema de Gerenciamento de Banco de Dados (SGBD), do inglês *Data Base Management System* (DBMS) é um conjunto de programas que gerenciam a estrutura do banco de dados e controlam o acesso aos dados armazenados. O SGBD provê entre os usuários e o banco de dados controle de armazenamento, operações de acesso, compartilhamento, segurança e organização. Existem diversos *softwares* gerenciadores de bancos de dados como Oracle, SQL Server, PostgreSQL, MySQL, Sybase, DB2, entre outros. Do ponto de vista do gerenciamento real, o SGBD é apenas um dos componentes do Sistema de Banco de Dados (SBD), que por sua vez é composto de cinco partes principais: hardware, software, pessoas, procedimentos e dados [2].

Figura 1: Ambiente simplificado do SBD – Sistema de Banco de Dados. Fonte: Elmasri & Navathe, 2017 [1].

Tipos de SGBDs ou Modelo de Dados

- **Hierárquico** – consiste em registros armazenados em estruturas de dados em árvore com dois componentes principais: tipo de registro e ligação. Estes registros são conectados através de ponteiros de forma que cada registro contém um único valor e cada ligação está relacionada exatamente a dois registros nesta estrutura que se assemelha a uma árvore genealógica com subordinações do tipo pai-filho. O acesso a qualquer registro é feito a partir do registro raiz. Exemplos de bancos de dados com modelo hierárquico: IMS, Adabas.
- **Em Rede** – baseado no modelo hierárquico adiciona uma extensão das limitações do tipo anterior. Este modelo permite uma estrutura mais completa, organizada em uma rede interligada de registros. O acesso do registro pode ser feito direto ao nó ao invés da necessidade de caminhar a partir da raiz.
- **Relacional** – proposto em 1970 por Edgar Frank Codd, este modelo é o mais utilizado atualmente e apresenta conceitos de lógica e teoria de conjuntos. Os bancos de dados relacionais possibilitam o tratamento e organização dos dados a partir de relações (tabelas) e das restrições definidas nestes relacionamentos. As consultas sobre as tabelas são baseadas na álgebra relacional através da linguagem SQL (*Structured Query Language*), conforme veremos com mais detalhes na seção 2.4 deste capítulo. Cada tabela deste modelo é composta por linhas (tuplas) e colunas (atributos). Os principais SGBDs da atualidade utilizam este modelo, por exemplo, Oracle, SQL Server, MySQL, PostgreSQL, etc.
- **Orientado a Objetos** – cada informação é armazenada na forma de objetos através do paradigma de representação dos dados a partir de abstrações do mundo real. Esse modelo permite maior flexibilidade do que o modelo relacional diante de dimensionamento de dados complexos e na integração com linguagens de programação orientadas a objetos como Python, C++ e Java, por exemplo.

Figura 2: Comparativo entre as representações dos modelos de dados. Fonte: próprio autor.

O que é SQL?

SQL – *Structured Query Language* (Linguagem de Consulta Estruturada), é uma linguagem declarativa e não procedural implementada na IBM Research como a interface para um Sistema de Banco de Dados relacional experimental chamado SYSTEM R. É a linguagem padrão dos SGBDs relacionais comerciais atualmente. Sua padronização é realizada pela *American National Standards Institute* (ANSI). SQL é uma linguagem abrangente que possui instruções para definições de dados (DDL – *Data Definition Language*) – criação de esquemas e tabelas; e instruções para manipulação de dados (DML – *Data Manipulation Language*) – inserção, deleção e atualização de registros [1].

- **Comando DDL:**

```
CREATE TABLE clientes (  
  cliente_id number(10) NOT NULL,  
  nome_cliente varchar2(50) NOT NULL,  
  cidade varchar2(50)  
);
```

O exemplo do comando acima, na linguagem Oracle PL/SQL, representa a criação de uma tabela clientes que possui três colunas: cliente_id, nome_cliente e cidade. Os atributos “number” e “varchar” referem-se aos tipos de dados, valores numéricos com até 10 caracteres e valores alfanuméricos com até 50 caracteres, respectivamente. Já a expressão “NOT NULL” indica quais as colunas não podem conter valores nulos – cliente_id e nome_cliente.

- **Comando DML:**

```
mysql > select * from result_blast limit 10;
```

O exemplo do comando acima, seleciona todas as colunas (*) da tabela result_blast e limita o resultado aos 10 primeiros registros encontrados (cláusula limit 10). Exemplo adaptado de http://biodados.icb.ufmg.br/cromatina/bacharelado/Manual_aula_MySQL.pdf.

Bancos de dados biológicos

Classificação dos bancos de dados biológicos

Os bancos de dados biológicos são desenvolvidos para diversos propósitos, abrangem vários tipos de dados com cobertura heterogênea e são curados em diferentes níveis e com métodos diferentes, de modo que existem vários critérios aplicáveis à classificação dos bancos de dados biológicos, como observamos abaixo:

- Escopo da cobertura de dados: abrangentes ou especializados
- Nível de curadoria: primários ou secundários
- Método de biocuração: curadoria por especialistas ou pela comunidade científica
- Tipo de dados gerenciados: DNA, RNA, proteína, expressão, vias metabólicas, doenças, entre outros. [4]

Alguns exemplos de bancos de dados biológicos

São exemplos de bancos de dados biológicos [3]:

- **Bancos de dados de sequências e estruturas:** contém dados de sequências de nucleotídeos, aminoácidos, e de estruturas de proteínas. Exemplos: [GenBank](#); [DDBJ \(DNA Data Bank of Japan\)](#); [PDB \(Protein Data Bank\)](#).
- **Bancos de dados de genomas:** embora as sequências de genomas constituam entradas nos arquivos-padrão de sequência de ácidos nucleicos, muitas espécies têm bancos de dados especiais que juntam a sequência do genoma e sua anotação com outros dados relacionados às espécies. O [Genome](#) é o banco de dados do NCBI que armazena estruturas completas.
- **Bancos de dados de padrões de expressão e de proteômica:** armazenam as medições dos níveis de mRNA, geralmente por meio das ESTs (*Express Sequence Tag* ou Etiquetas de Sequências Expressas – sequências terminais curtas de cDNA sintetizadas a partir de mRNA) descrevendo padrões de transcrição de genes. Bancos de dados de proteômica armazenam medidas realizadas com proteínas, descrevendo padrões de tradução de genes.
- **Bancos de dados de vias metabólicas:** contém vias metabólicas de uma variedade de espécies com integração das informações bioquímicas. Um dos mais importantes e completo banco de dados é o [KEGG \(Kyoto Encyclopedia of Genes and Genomes\)](#).
- **Bancos de dados de referências da literatura científica:** abrange as informações de citações, resumos e artigos relacionados às pesquisas biológicas. Por exemplo, o [PubMed](#), plataforma popular de pesquisa mantida pelo NCBI.

Outros bancos de dados biológicos

- [NAR](#) – a revista internacional *Nucleic Acids Research* publica anualmente uma lista extensa de bancos de dados de biologia molecular desde o ano de 1996.
- **Bancos de dados NCBI** – todos os bancos de dados moleculares mantidos pelo NCBI. Esses bancos de dados estão divididos em 6 grupos principais: nucleotídeo, proteína, estrutura, taxonomia, genoma e expressão.
- **Banco de dados EBI** – o portal principal de todos os bancos de dados do EBI dividido em vários grupos, como literatura, microarranjo, nucleotídeo, proteína, estrutura, caminho e ontologia [9].
- **Protein Data Bank** – principal repositório mundial de estruturas de proteínas, DNA e RNA.

Formatos de arquivos de dados comuns em Bioinformática

- **Formato FASTA:** um arquivo no formato FASTA começa com uma linha de descrição seguida por linhas de dados da sequência. A linha de descrição é diferenciada dos dados da sequência por um símbolo maior que (">") no início. É recomendável que todas as linhas de texto tenham menos de 80 caracteres. Uma sequência de exemplo no formato FASTA é:

```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDITLVNVAIFYKGMWKTAFNAEDTREMPPHVTQESKPVQ
MMCMNNSFNATLPAEKMKILELPFASGDLMLVLLPDEVSDLERIEKTINFEK
LTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTSVLMALGMTDLFIPSANLTGISS
AESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESQFRADHPFLFLIK
HNPTNTIVYFGRYWSP
```

- **Formato EMBL:** armazena a sequência e sua anotação juntas. O início da seção de anotação é marcado por uma linha que começa com a palavra "ID". A seção de início da sequência é marcada por uma linha que começa com a palavra "SQ". A linha "//" (terminador) também não contém dados ou comentários e designa o final de uma entrada. O arquivo EMBL pode terminar com a extensão .embl ou .txt.
- **PDB/mmCIF:** é o formato padrão de arquivos do [wwPDB](#) (*World Wide Protein Data Bank*). Armazenam informações de estruturas de biomoléculas.
- **GFF:** O formato GFF (*General Feature Format*) é um tipo de arquivo delimitado por tabulação que descreve genes e outras características de sequência de DNA, RNA e proteínas.
- **GTF:** *General Transfer Format* – idêntico à versão GFF2.
- **SAM:** *Sequence Alignment Map* – armazenam sequências biológicas alinhadas a uma sequência de referência. O formato SAM consiste em uma seção de cabeçalho e uma seção de alinhamento. As linhas na seção de cabeçalho começam com o caractere '@' e as linhas na seção de alinhamento não. Todas as linhas são delimitadas por TAB. [13]
- **PIR:** uma sequência no formato PIR consiste em: uma linha começando com um sinal ">" (maior que), seguida por um código de duas letras que descreve o tipo de sequência (P1, F1, DL, DC, RL, RC ou XX) , seguido por um ponto e vírgula, seguido pelo código de identificação de sequência (o código de identificação do banco de dados). Uma linha contendo uma descrição textual da sequência. Uma ou mais linhas contendo a própria sequência. O final da sequência é marcado com um caractere "*" (asterisco). O formato PIR também é conhecido como formato NBRF [14]. Exemplo de Formato PIR:

```
>P1;CRAB_ANAPL
ALPHA CRYSTALLIN B CHAIN (ALPHA(B)-CRYSTALLIN)
MDITIHNP LI RRPLFSWLAP SRIFDQIFGE HLQESSELLPA SPSLSPFLMR
SPIFRMPSWL ETGLSEMRLE KDKFSVNLDV KHFSPEELKV KVLGDMVEIH
GKHEERQDEH GFIAREFNRK YRIPADVDP L TITSSLSLDG VLTVSAPRKQ
SDVPERSIPI TREEKPAIAG AQRK*
```

Tutorial de exemplo de busca no GenBank

Agora, para ilustrar o uso de banco de dados biológicos vamos apresentar um breve tutorial de acesso ao banco de dados GenBank. Inicialmente, faça acesso à página inicial de pesquisa: <https://www.ncbi.nlm.nih.gov/genbank/>.

Figura 3: Página inicial do GenBank. Fonte: próprio autor.

Existem várias maneiras de pesquisar e recuperar dados do GenBank. Neste exemplo, iremos buscar as informações de sequências para a proteína P53 em *Homo Sapiens*.

Figura 4: Campo de busca no GenBank. Fonte: próprio autor.

Abaixo estão os resultados da pesquisa. A lista ao lado direito “Results by taxon” mostra o número de resultados retornados por espécie. A busca também pode ser realizada informando os identificadores *Accession* e *GI number*. Nesta consulta, o resultado recuperado foi de 22783 registros.

Figura 5: Resultados da consulta realizada. Fonte: próprio autor.

Também podemos analisar os registros recuperados detalhando-os no formato do GenBank, FASTA ou Graphics.

Homo sapiens mRNA for P53, complete cds
GenBank: AB082923.1
[FASTA](#) [Graphics](#)

[Go to:](#) [⌵](#)

LOCUS AB082923 2451 bp mRNA linear PRI 01-APR-2003
DEFINITION Homo sapiens mRNA for P53, complete cds.
ACCESSION AB082923
VERSION AB082923.1
KEYWORDS .
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
REFERENCE 1
AUTHORS Azuma,K., Shichijo,S., Maeda,Y., Nakatsura,T., Nonaka,Y., Fujii,T.,
Koike,K. and Itoh,K.
TITLE Mutated p53 gene encodes a nonmutated epitope recognized by
HLA-B*4601-restricted and tumor cell-reactive CTLs at tumor site
JOURNAL Cancer Res. 63 (4), 854-858 (2003)
PUBMED 12591737
REFERENCE 2 (bases 1 to 2451)
AUTHORS Shichijo,S. and Itoh,K.
TITLE Direct Submission
JOURNAL Submitted (26-MAR-2002) Shigeki Shichijo, Kurume Univ. School of
Med., Dep. Immunol.; 67-Asahi-machi, Kurume, Fukuoka 830-0011,
Japan (E-mail:shichijo@med.kurume-u.ac.jp, Tel:81-942-31-7551,
Fax:81-942-31-7699)
FEATURES
source Location/Qualifiers
1..2451
/organism="Homo sapiens"
/mol_type="mRNA"
/db_xref="taxon:9606"
/clone="79"
/cell_line="SW620"
/cell_type="colon cancer"
1..2451

Customize view

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

Articles about the TP53 gene
[Cerebrospinal fluid TP53 gene mutation in patients with lung (Zhonghua Yi Xue Za Zhi. 2020)
Evaluation of p53 and Its Target Gene Expression as P [Asian Pac J Cancer Prev. 2020]
COUP-TFII Overexpression Inhibits Cell Proliferation and Invasion [Anticancer Res. 2020]
See all...

Pathways for the TP53 gene
Ferroptosis
Mitophagy - animal
Fluid shear stress and atherosclerosis
See all...

Related sequence information

Figura 6: Exemplo de um resultado detalhado – formato de arquivo do GenBank. Temos a descrição, informações das publicações de artigos relacionados, a sequência de nucleotídeos, regiões codificadoras de proteínas – CDS (Coding Sequence), etc. Fonte: próprio autor.

É possível realizar a *download* do arquivo conforme o formato desejado selecionando a opção “*Send to*”, marcando o botão de opção “*Choose Destination: File*” e em seguida basta clicar em “*Create File*”.

(A)

Homo sapiens mRNA for P53, complete cds
GenBank: AB082923.1
[FASTA](#) [Graphics](#)

[Go to:](#) [⌵](#)

LOCUS AB082923 1182 bp mRNA linear PRI 01-APR-2003
DEFINITION Homo sapiens mRNA for P53, complete cds.
ACCESSION AB082923 REGION: 64..1245
VERSION AB082923.1
KEYWORDS .
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
REFERENCE 1
AUTHORS Azuma,K., Shichijo,S., Maeda,Y., Nakatsura,T., Nonaka,Y., Fujii,T.,
Koike,K. and Itoh,K.
TITLE Mutated p53 gene encodes a nonmutated epitope recognized by
HLA-B*4601-restricted and tumor cell-reactive CTLs at tumor site
JOURNAL Cancer Res. 63 (4), 854-858 (2003)
PUBMED 12591737
REFERENCE 2 (bases 1 to 1182)

Send to: **Complete Record**
Coding Sequences
Gene Features
Choose Destination
File **Clipboard**
Collections **Analysis Tool**
Download 1 item.
Format
GenBank
Show GI
Create File

Change region shown

Customize view

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

Articles about the TP53 gene
[Cerebrospinal fluid TP53 gene mutation in patients with lung (Zhonghua Yi Xue Za Zhi. 2020)
Evaluation of p53 and Its Target Gene Expression as P [Asian Pac J Cancer Prev. 2020]
COUP-TFII Overexpression Inhibits Cell Proliferation and Invasion [Anticancer Res. 2020]
See all...

Pathways for the TP53 gene
Ferroptosis
Mitophagy - animal
Fluid shear stress and atherosclerosis
See all...

(B)

Homo sapiens mRNA for P53, complete cds
GenBank: AB082923.1
[GenBank](#) [Graphics](#)

FASTA **Send to:** **Change region shown**

Customize view

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

Articles about the TP53 gene
[Cerebrospinal fluid TP53 gene mutation in patients with lung (Zhonghua Yi Xue Za Zhi. 2020)
Evaluation of p53 and Its Target Gene Expression as P [Asian Pac J Cancer Prev. 2020]
COUP-TFII Overexpression Inhibits Cell Proliferation and Invasion [Anticancer Res. 2020]
See all...

Pathways for the TP53 gene
Ferroptosis
Mitophagy - animal
Fluid shear stress and atherosclerosis
See all...

(C)

Figura 7: (A) Download do arquivo no formato GenBank. (B) Seleção da visualização do registro localizado no formato FASTA. (C) Visualização da sequência – Graphics. Fonte: próprio autor.

Um pouco mais sobre o PDB

O PDB “*Protein Data Bank*” é um banco de dados de estruturas obtidas através de experimentos de difração de raios-X e de nêutrons em cristais e de Ressonância Magnética Nuclear (RMN) de milhares de proteínas [7], tornando-se o maior repositório mundial de informações sobre a estrutura 3D de grandes moléculas biológicas, incluindo os ácido nucleicos. Esse banco apresentava em 20 de abril de 2020 o total de 162.816 arquivos com estruturas de macromoléculas.

Um arquivo de estrutura ou arquivos PDB são representados em um formato específico e padronizado, chamado formato PDB. Esses arquivos são utilizados em programas de modelagem e visualização molecular. Um arquivo PDB pode ser dividido em duas partes principais, a primeira contém informações e detalhes sobre a estrutura e a segunda contém as coordenadas dos átomos. Na primeira seção, informações tais como o método utilizado para resolver a estrutura da proteína, o nome dos autores que resolveram a estrutura, nome da macromolécula e a lista de cadeias disponíveis na estrutura, informações sobre as estruturas secundárias, entre tantas outras informações estão disponíveis. A segunda seção contém informações relativas às coordenadas dos átomos da proteína e dos ligantes (registros chamados de ATOM e HETATM), e informações sobre ligações covalentes e conexões entre átomos (registros chamados de CONECT). A partir destas coordenadas, é possível calcular os tipos atômicos e interações envolvidas entre proteínas e seus ligantes [8].

Referências do capítulo

[1] ELMASRI, Ramez; NAVATHE, Sham. **Fundamentals of database systems**. Pearson, 2017.

[2] ROB, Peter; CORONEL, Carlos. **Sistemas de banco de dados: Projeto, implementação e gerenciamento**. Cengage Learning, 2014.

- [3] LESK, Arthur M. **Introdução à Bioinformática**. 2.ed. Artmed, 2008.
- [4] Zou, Dong et al. "Biological databases for human research." *Genomics, proteomics & bioinformatics* vol. 13,1 (2015): 55-63. doi:10.1016/j.gpb.2015.01.006
- [5] HARISHA, Sharma. **Fundamentals of Bioinformatics**. IK International Pvt Ltd, 2010.
- [6] SELZER, Paul M.; MARHÖFER, Richard J.; ROHWER, Andreas. *Applied bioinformatics. An introduction*—Springer, Verlag, Berlin, Heidelberg, Germany, v. 260, 2008.
- [7] <https://www.rcsb.org/>
- [8] <https://www.wwpdb.org/documentation/file-format>
- [9] JIANG, Rui; ZHANG, Xuegong; ZHANG, Michael Q. *Basics of Bioinformatics*. Springer-Verlag Berlin An, 2016.
- [10] https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp
- [11] <http://scikit-bio.org/docs/0.5.2/generated/skbio.io.format.embl.html>
- [12] <https://www.ensembl.org/info/website/upload/gff.html>
- [13] Li, Heng et al. "The Sequence Alignment/Map format and SAMtools." *Bioinformatics (Oxford, England)* vol. 25,16 (2009): 2078-9. doi:10.1093/bioinformatics/btp352
- [14] http://www.bioinformatics.nl/tools/crab_pir.html
- [15] <https://br.ccm.net/contents/66-o-que-e-o-sistema-de-gerenciamento-de-bancos-de-dados>