



I Curso de Aperfeiçoamento em Bioinformática da UFMG

Bancos de Dados em Biologia

Alessandra C. Faria-Campos, D. Sc.

Sistema de Bancos de Dados



“Um sistema de banco de dados é basicamente **um sistema computadorizado de manutenção de registros**. Pode ser considerado o equivalente eletrônico de um armário de arquivamento; ou seja um repositório ou recipiente para uma coleção de arquivos de dados”

Banco de Dados



- **Finalidade geral:** Armazenar informações e permitir que os usuários busquem e atualizem estas informações de forma eficiente e rápida quando necessário
- Os usuários podem:
 - Inserir,
 - Buscar,
 - Atualizar e
 - Remover dados.

Vantagens

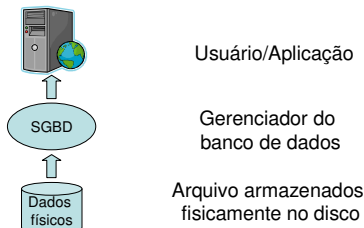


- **Organização dos dados:** a estrutura dos dados é muito complicada para ser armazenada em tabelas
- **Volume de dados:** o volume dos dados é muito grande para ser armazenado em uma tabela
- **Velocidade/Eficiência:** Os dados são obtidos e atualizados rapidamente no BD
- **Concorrência:** diferentes aplicações/usuários podem compartilhar os mesmos dados
- **Redundância controlada:** dados compartilhados não precisam ser replicados
- **Aplicabilidade:** Diminuição no tempo de desenvolvimento da aplicação.

Sistema Gerenciador de Banco de Dados (SGBD)



- Camada de Software entre o banco de dados físico – dados fisicamente armazenados – e os usuários do sistema.



Características do SGBD



- **Controle de Concorrência** – Limita alterações e leituras simultâneas do mesmo dado por diferentes usuários;
- **Backup** – mecanismo de reconstrução que permite que a base de dados retorne ao estado consistente;
- **Segurança** – pode estabelecer restrições ao acesso de cada item de informação;
- **Integridade** – manutenção da consistência da base de dados através da validação de restrições;
- **Atomicidade** – ou a transação ocorre ou não.

Modelos de Bancos de Dados

Modelo conceitual : define o modo pelo qual os dados se organizarão no BD

- Modelo Hierárquico;
- Modelo em Rede;
- Modelo Orientado a Objetos;
- Modelo Relacional: mais usado!

Projeto Lógico do BD (tabelas)

Modelo Relacional

Conceitos centrais

- Entidade: Conjunto de objetos da realidade sobre os quais deseja-se manter informações no banco de dados. Ex: proteína, categoria funcional biológica – instância: ocorrência específica da entidade ex: hemoglobina
- Relacionamento: Conjunto de associações entre entidades sobre as quais deseja-se manter informações na base de dados. Ex: pertencer a uma categoria funcional
- Atributo: Informação associada a entidade. Ex: proteína tem sequência primária, secundária
- Generalização/especialização:
 - Globina => hemoglobina

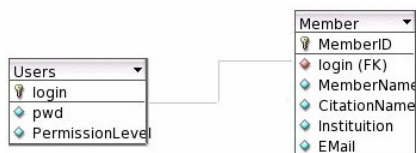
Diagrama entidade-relacionamento



Modelo Relacional

- Os dados são percebidos como tabelas (relações);
- Atributos (campos da tabela);
- Chaves:
 - primárias: atributo que identifica a entidade. Deve ser único para cada registro;
 - estrangeiras: atributo para referenciar entidades em outras tabelas.

Modelo Relacional



Modelo Relacional

login	PermissionLevel
hrausch	5
alesa	5
heron	4
gloria	4

Tabela de usuários:

MemberID	login	MemberName	CitationName	Institution	E-Mail
16	hrausch	Herbert Rausch	Rausch-Fernandes, H	UFMG	hrausch@gmail.com
25	alesa	Alessandra Faria Campos	Faria-Campos, A	UFMG	alesa@icb.ufmg.br
26	heron	Heron Oliveira Hilario	Hilario, H.O.	UFMG	heron_oh@hotmail.com
27	gloria	Gloria Regina Franco	Franco, G.R.	UFMG	gfranco@cb.ufmg.br

Tabela de membros:

Exemplos



Bancos de Dados Biológicos

“BD biológicos se tornaram uma importante ferramenta no entendimento da vasta quantidade de fenômenos biológicos existentes, desde a estrutura das biomoléculas e sua interação ao metabolismo como um todo e a evolução das espécies. Este entendimento contribui para facilitar a luta contra doenças, auxilia no desenvolvimento de novos fármacos e na descoberta de relações entre espécies.” (Wikipedia, 2006)

Requisitos para um bom banco de dados biológico

- Qualidade dos dados
- Anotações consistentes
- Integração
- Fácil acesso às informações disponíveis
- Mecanismos para extrair do conjunto de dados apenas aqueles de interesse do pesquisador, isto é que respondem a uma pergunta biológica específica

Bancos de dados Biológicos

- BD são uma parte significativa do trabalho desenvolvido em Bioinformática
- Os BD de biologia são tanto bancos públicos (ex: Genbank) como privados
- Mais de 1000 bancos de dados biológicos comerciais e públicos disponíveis atualmente
- O acesso a esses bancos de dados através de padrões abertos (*open standards*) como a *web* é importante dada as características dos usuários destes bancos – Servidores UNIX x Machintosh
- A revista Nucleic Acids Research é um importante recurso com informações sobre estes BD (<http://www3.oup.co.uk/nar/database/c/>)

Conteúdo

- Contém dados de genômica, transcriptômica, proteômica, taxonomia, ecologia, doenças, fármacos, etc.
- Informações:
 - Sequências de nucleotídeos, aminoácidos,
 - Função, estrutura, localização no cromossomo
 - Mapas metabólicos
 - Efeitos clínicos de mutações
 - Características genéticas de populações específicas
 - Catálogo de espécies ou recursos naturais
 - Etc.

Bancos de Dados Biológicos

Alguns tipos de bancos de dados biológicos:

- Bancos de dados primários de seqüência (nucleotídeos e aminoácidos) – GenBank, UniProt
- Bancos de genomas – Mouse Genome Database, NCBI Genomic Biology
- Bancos de dados especializados - Flybase, Wormbase, CGAP
- Bancos de dados de vias bioquímicas – KEGG
- Bancos de dados de estrutura de proteínas – PDB, SCOP
- Bancos de dados de *microarrays* – Array Express, SMD
- Bancos de dados de interações proteína-proteína – STRING, BioGRID
- Bancos de Cadastro de recursos naturais – AmazonLink, ENDS, National Whale and Dolphins Stranding Database

Uso dos BD Biológicos

O que se pode descobrir por meio da busca em BD biológicos?

- Informação evolutiva: genes homólogos, freqüências dos alelos, ...
- Informação genômica: localização no cromossomo, introns, UTRs, regiões reguladoras, ...
- Informação estrutural: estruturas da proteína correspondente, tipos de folds, domínios estruturais, ...
- Informação de expressão: expressão específica a um dado tecido, fenótipos, doenças, ...
- Informação funcional: função molecular/enzimática, papel em diferentes vias, papel em doenças, ...

Busca de Informação em BDB

Busca de informação sobre genes e produtos gênicos

- Gene e produtos gênicos são geralmente organizados por seqüência
 - Seqüências genômicas codificam todas as características de um organismo
 - Produtos gênicos são descritos unicamente por sua seqüência
- Seqüências similares entre biomoléculas indica tanto uma função similar quanto um relacionamento evolutivo
- Seqüências de macromoléculas proporciona chaves biologicamente significativas para busca em BD

Busca em BDB

- Comece com uma seqüência, encontre informação sobre ela
- Muitos tipos de seqüências de entrada
 - Pode ser uma seqüência de aminoácido ou de nucleotídeo
 - Genômica, cDNA/mRNA, proteína
 - Completa ou fragmentada
- *Matches* exatos são raros
 - Em geral, o objetivo é recuperar um conjunto de seqüências similares

Busca em BD de Seqüências


O que queremos saber sobre a seqüência?

- Ela é similar ao algum gene conhecido? Quão próximo é o melhor *match*? Significância?
- O que sabemos sobre este gene?
 - Genômica (localização no cromossomo, regiões reguladoras, ...)
 - Estrutural (estrutura conhecida? ...)
 - Funcional (molecular, celular e doença)
- Informação evolutiva
 - Este gene é encontrado em outros organismos?
 - Qual é sua árvore taxonômica?

ENTREZ

PDB

Gene Ontology



the Gene Ontology Search

menus

downloads

documentation

about GO

contact GO

map

Gene Ontology Home

The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. [Read more about the Gene Ontology...](#)

Search the Gene Ontology Database

Search for genes, proteins or GO terms using AmiGO:

gene or protein name GO term or ID


AmiGO is the official GO browser and search engine. Browse the Gene Ontology with AmiGO.

GO website

- [GO downloads](#), including ontology files, annotations and the GO database
- [Tools](#) for using GO, including OBO-Edit downloads and AmiGO
- Request new terms or ontology changes via the [GO curator requests tracker](#); help with new term submission is available.
- [Documentation](#) on all aspects of the GO project and the [GO FAQ](#)
- [Gene Ontology mailing lists](#) and [contact details](#)

www.geneontology.org

NCBI Taxonomy Browser



Taxonomy Home - Mozilla Firefox

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

Loading... Last modified: ...

About Entrez

Entrez Taxonomy

LinkOut Manual

Taxonomy browser

Taxonomy Common tree

Taxonomy resources

Taxonomy Statistics

Taxonomy FTP site

Taxonomy FAQs

The NCBI Entrez Taxonomy Homepage

The NCBI taxonomy database contains the names of all organisms that are represented in the genetic databases with at least one nucleotide or protein sequence. Click on the [tree](#) if you want to browse the taxonomic structure or retrieve sequence data for a particular group of organisms.

These are direct links to some of the organisms commonly used in molecular research projects:

<input type="checkbox"/> Arabidopsis thaliana	<input type="checkbox"/> Escherichia coli	<input type="checkbox"/> Panicum umbricola
<input type="checkbox"/> Bos taurus	<input type="checkbox"/> Homo sapiens	<input type="checkbox"/> Rattus norvegicus
<input type="checkbox"/> Canis familiaris	<input type="checkbox"/> Mus musculus	<input type="checkbox"/> Saccharomyces cerevisiae
<input type="checkbox"/> Gallus gallus	<input type="checkbox"/> Drosophila melanogaster	<input type="checkbox"/> Schistosoma mansoni
<input type="checkbox"/> Dactylopusia	<input type="checkbox"/> Xenopus laevis	<input type="checkbox"/> Zea mays
<input type="checkbox"/> Drosophila	<input type="checkbox"/> Plasmodium falciparum	<input type="checkbox"/> Lepidoptera

Comments and questions to info@ncbi.nlm.nih.gov