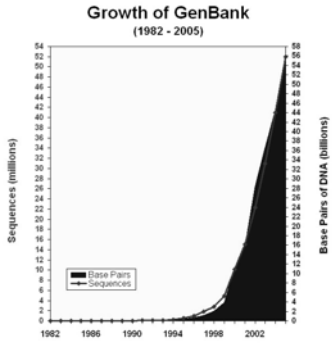
Datenbanken in der Bioinformatik

Bioinformatik

Anwendung von Informatik-methoden zur Untersuchung von Problemen der Molekularbiologie, die auf **sehr großen Datenmengen** beruhen



Quelle: http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html, 15.11.07

Bioinformatikdatenbanken (BIDB)

Es geht um Daten

- eines einzigen (abgeschlossenen) Projektes
- die auf einer weltweiten und andauernden Zusammenarbeit zwischen Forschungsteams beruhen
- eines einzigen Organismus
- über das Vorkommen eines Proteins in allen möglichen Organismen

Fragestellungen an BIDB

DNS-Analyse und Sequenzierung	Sequenz-DB
Ermittlung phylogenetischer Bäume	Sequenz-DB
Genexpressionsanalyse	→ Spezielle DB
Ermittlung biochemischer Pfade	Sequenz-DB → Spezielle DB
Strukturvorhersage	Proteinsequenz- und Proteinstruktur-DB

Beispiel-DB*

Gen bank NCBI	▪www.ncbi.nlm.nih.gov ▪www.ebi.ac.uk ▪www.ddbj.nig.ac.jp
German Human Genom Project	www.dhgp.de (2004 eingestellt) www.genome.gov

* von Prof. Cieslik zusammengestellt

Beispiel-DB

Protein bank  	▪ www.expasy.ch ▪ www.embl-heidelberg.de ▪ www.rcsb.org/pdb/

Beispiel-DB

Phylogeny 	▪www.ucmp.berkeley.edu/exhibit/phylogeny.html ▪http://evolution.genetics.washington.edu ▪http://tolweb.org/tree ▪http://awcmee.massey.ac.nz

Merkmale biologischer Daten

- Sehr **komplex**
  → mit traditionellen DBMS kann man i. a. nicht alle Datenaspekte erfassen
- **Menge** und **Variabilität** der Daten sind **sehr hoch**
  → Handhabung von Datentypen und -werten muss sehr flexibel sein

Merkmale biologischer Daten

- **Schemata** in biologischen Datenbanken **ändern sich sehr schnell**
  → heutige Systeme erstellen wenigstens einmal pro Jahr das gesamte Datenbankschema neu
- **Darstellungen** der gleichen Daten durch verschiedene Biologen **unterscheiden sich** meist
- Die meisten Biologen haben **keine Kenntnis über** die **interne Struktur** der Datenbank

Merkmale biologischer Daten

- Definition und Darstellung **komplexer Abfragen** sind wichtig
  → Werkzeuge für die Formulierung von Abfragen müssen bereitgestellt werden
- Benutzer benötigen oft auch den **Zugriff auf „alte" Datenwerte**
  → Änderungen der Datenwerte müssen archiviert werden

Schlussfolgerungen

- herkömmliche DBMS erfüllen nicht alle Anforderungen komplexer biologischer Daten
  → Weiterentwicklungen der DBMS erforderlich
- **GENOME** (Georgia Tech Emory Network Object Management Environment) ist eine solche Weiterentwicklung (Emory ist eine Universität in Georgia)

BIDB-Modelle

- weisen Gemeinsamkeiten bezüglich ihrer Datenmodellierung und ihrem Management auf
- BIDB verwenden z. B. folgende Modelle (Prozentangaben beziehen sich auf die von Bry und Kröger untersuchten DB)

BIDB-Modelle

Modell	Prozent	Bemerkungen
Flat-Files	40 %	ASCII-Texte, unstrukturiert
Relationales	30 %	Herkömmliches Modell, für molekularbiol. Daten wenig geeignet
Objektorien-tiertes	9 %	Strukturierte Daten, passend für Daten der Molekularbiologie
ACEDB (A C. elegans Database)	4 %	Eigenes, spezielles Modell, Repräsentation von genetischen Daten

Abfragen

- Die meisten BIDB bieten Webformulare an, mit denen man Abfragen zu den DB erstellen kann. Es sind meist nur begrenzte Abfragetypen.
- Beispiele
  SWISS-PROT – derzeit größte Proteindatenb.
  BLAST – Basic Local Alignment Search Tool
  Entrez – Suchmaschine
  Tree of Life

Literatur

- Bry, F, Kröger, P.: A computational biology database digest: data, data analysis, and data management. Research Report PMS-FB-2002-8
- http://www.pms.informatik.uni-muenchen.de/publikationen/#PMS-FB-2002-8
- Bry, F, Kröger, P.: Datenbanken in der Bioinformatik. Informatik Spektrum 17, Okt. 2002
- Elmasri, R., Navathe, S. B.: Grundlagen von Datenbanksystemen. Pearson Studium 2002
- Gibas, C., Jambeck, P.: Praktische Bioinformatik. O'Reilly Verlag 2002

Literatur

A computational biology database digest: data, data analysis, and data management

enthält u. a.:

- 124 Literaturhinweise
- 111 untersuchte BIDB
- Tabelle mit URLs zu Methoden der Datenanalyse (Sequence Alignment, Gene Finding, Gene Expression)