Database Systems
in Bioinformatics

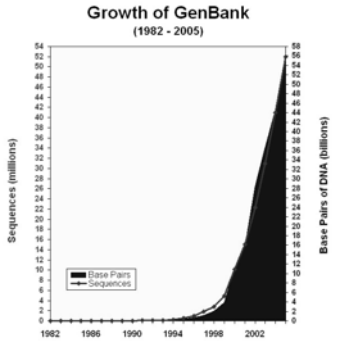
1

Bioinformatics

Application of mathematical and Computer Science methods to solve problems in Molecular Biology that require

large scale data

2 / 17



Quelle: http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html, 6.9.07

3 / 17

Databases in Bioinformatics (DBB)

There are several data

- Of one (closed) project, only
- Which consist of the world wide and ongoing cooperation of teams of research
- Of a single organism
- About the existence of a protein in all possible organisms

4 / 17

Questions to DBB

DNA Analysis and Sequencing	sequence DB
Determination of Phylogenetic Trees	sequence DB
Gene Expression Data Analysis	→ special DB
Determination of metabolical paths	sequence DB → special DB
Protein Structure Prediction	Protein sequence and protein structure DB

5 / 17

Examples for DBB

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

6 / 17

Examples for DBB

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

7 / 17

Examples for DBB

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

8 / 17

Characteristics of biological data

- Very **complex**
  → you cannot describe all aspects of data with traditional DBMS
- **Amount** and **variability** of data are **very large**
  → The types and values of data must be **very flexible**

9 / 17

Characteristics of biological data

- **Schemes** in biological databases **change very quickly**
  → Systems of today at least once a year create a new database scheme
- **The descriptions** of the same data by several biologists often **differ**
- Most of the biologists **don't know** the **inner structure** of the data base

10 / 17

Characteristics of biological data

- Definition and description of **complex queries** are important
  → Tools for the formulation of queries must be provided
- User often need the **Access to "old" values of data**
  → changes of values must be archived

11 / 17

Conclusions

- Conventional DBMS don't meet all requirements of complex biological data
  → further developments of DBMS are necessary
- **GENOME** (Georgia Tech Emory Network Object Management Environment) is one of such developments (Emory is a university in Georgia)

12 / 17

Models for DBB

- Have common characteristics in relation to their data models and their management
- DBB use for example the following models (the percents refer to the DB that have been analyzed by Bry and Kröger)

13 / 17

Models for DBB

Model	Percent	Remarks
Flat files	40 %	ASCII and GIF files, unstructured
Relational	30 %	Common Model, for molecular biological data not so suitable
Object	9 %	structured data, suitable for data of Molecular Biology
ACEDB (A C. elegans Database)	4 %	Special model, representation of genetic data

14 / 17

Queries

- Most DBB offer web forms to create queries, often there are only special types of queries
- Examples
  SWISS-PROT – largest protein database
  BLAST – Basic Local Alignment Search Tool
  Entrez – Life Sciences Search Engine
  Tree of Life

15 / 17

Literature

- 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
- Elmasri, R., Navathe, S. B.: Fundamentals of Database Systems. Pearson Education 2000
- http://www.sbc.su.se/~pjk/strbio2001/databases/index.html

16 / 17

Literature

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

contains among other things:

- 124 references
- 111 analyzed DBB
- Table with URLs for methods of data analysis (Sequence Alignment, Gene Finding, Gene Expression)

17 / 17