Datenbanken für Bioinformatiker
(Fortsetzung von Einf. in DBS)

H. Schweppe
FU Berlin, SS 2005

hs@inf.fu-berlin.de

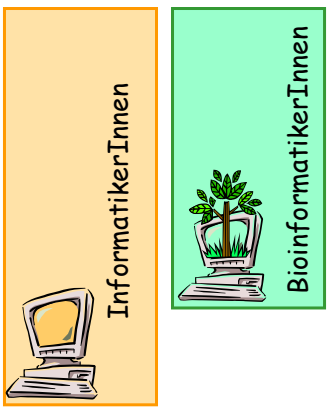
Schwerpunkte

1. Datenmodellierung:
   systematischer Entwurf von DB
   Schwerpunkt: Relationale DB

2. Datenbanknutzung:
   Zugriff auf die Daten mit SQL (Structured Query Language), interaktiv oder mittels Applikationsprogrammen

3. Datenbankimplementierung

Spezielle Aspekte von Datenbanken in der Bioinformatik



..\Search Results in the Bio Netbook [database genomic].htm

1 Introduction

Lit.: Bry, Kröger: A Computational Biology Database Digest, see reader
Course of Prof. Leser, HU, SS 2003, http://www.informatik.hu-berlin.de/wbi/
Lacroix, Critchlow (eds): Bioinformatics – Managing Scientific Data, Morgan Kaufmann, 2003
Kranss, Raymer: fundamental Concepts of Bioinformatics, Pearson Education, 2003

Introduction

- Not a course on specific molecular–biologic data bases and their usage
- … but on methods used for data management beyond relational DBS within this application domain
- Most methods also employed in completely different application domains
- Some Key words: data integration, data warehouse, mining, clustering, representation, XML, information retrieval / text retrieval, similarity, Ontologies, object oriented database (modeling), …

```
LOCUS       AL33DMOR4     314 bp ds-DNA           PHG       24-MAY-1991
DEFINITION  Bacteriophage alpha3 deletion mutant DNA for origin region (-ori)
            of replication
ACCESSION   X15716
KEYWORDS    origin of replication.
SOURCE      Bacteriophage alpha3 DNA.
  ORGANISM  Bacteriophage alpha3
            Viridae; ds-DNA nonenveloped viruses; Siphoviridae.
REFERENCE   1  (bases 1 to 314)
  AUTHORS   Kodaira,K.I.
  JOURNAL   Unpublished (1989)
  STANDARD  full automatic
REFERENCE   2  (sites)
  AUTHORS   Nakano,K., Kodaira,K.I. and Taketo,A.
  TITLE     Properties of the bacteriophage alpha3 mutants with deletion and/or
            insertion in the complementary strand origin
  JOURNAL   Biochim. Biophys. Acta 1048, 43-49 (1990)
  STANDARD  full staff_review
COMMENT     *source: clone=delo L105; See <J02444>,<M25640>,<X13332>,<X15713>
            to <X15715>,and <X15717> to <X15721> for other -ori sequences.

            Data kindly reviewed (06-JAN-1990) by Kodaiva K.-I.

            From EMBL     entry BA3DMOR4;  dated 16-JAN-1991.
BASE COUNT      68 a     67 c     82 g     97 t
ORIGIN
      1 tgaagttgag cattacccaa ttgaaatgtc tgttggttct ggtggtgttt gttctgctcg
     61 cgattgtgct actgttgata ttcatcctcg tacttctggt aataatgttt ttgttggtgt
    121 gatttgttct agcgctaaat ggacctccgg tcgtgtgatt ggtaccatcg ctacgactca
    181 ggttattcat gaataccaag tccttcagcc gcttaaataa aaggctgccg cactcccggt
    241 tagatgcctg cccagtgtag ggcagaccgg tacggagata cccgataaac taggaacgtg
```

How to represent this?

structured fields

multiple values

Natural Language (NL) text

reference to foreign data sets, "micro syntax"

long fields

1.1 DBS versus Bio Databases

- Standard DB management

  Query – Answer paradigm (Q-A)

  – Well defined query language, well defined data set
  – Answer A(Q) well defined, independent of individual judgment
  – Basically only one structuring principle (e. g. "relations")
  – Data are correct

    … more or less.

    e.g. address DB vs account management

Paradigms for data management

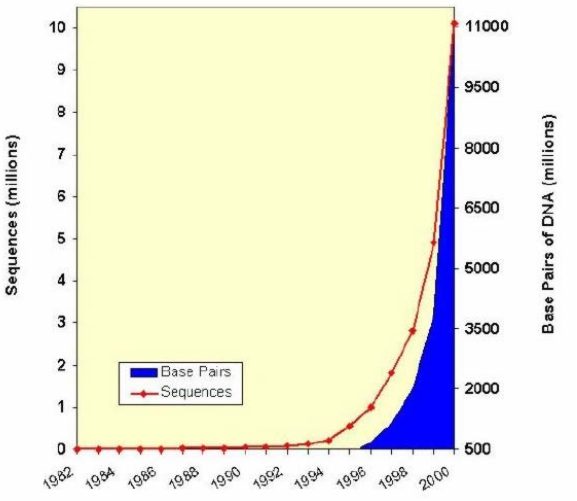
- Management of Life Science Data

  Problem solving paradigm

  " Any diseases similar to neurofibromatosis (type 2) with a known gene on chromosome 22?"

  - Boolean results infrequent,
    Answer to query has to be interpreted and put into context
  - Data structurally very different (gene bank vs. protein-protein interaction DB vs publications DB vs…)
  - Data incomplete
  - Questionable quality
  - DBs interact heavily with tools (e.g. for alignment)

Growth



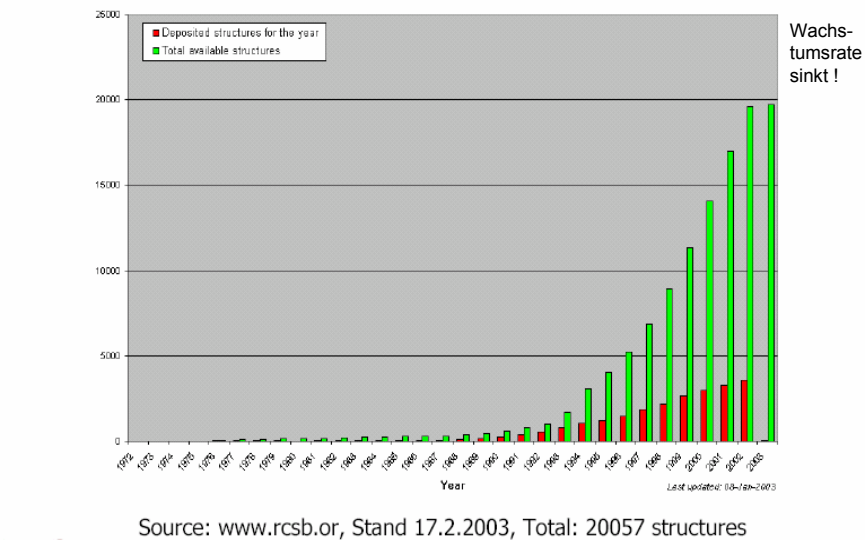
2003:
36,553,368,485 Base Pairs
30,968,418 Sequences

only 36 GB ?

Problem: there are MANY data sources for different species, different kinds of info, …

Growth of the GenBank DB (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html)

Growth of Protein Structure DBs



Wachs-
tumsrate
sinkt !

Source: www.rcsb.or, Stand 17.2.2003, Total: 20057 structures

The great challenge in biological research today is how to turn data into knowledge.

Sydney Brenner. The Scientist 16[6]:12, March 18, 2002

… but what is knowledge – in a computing machine?

Overview: databases in bioinformatics

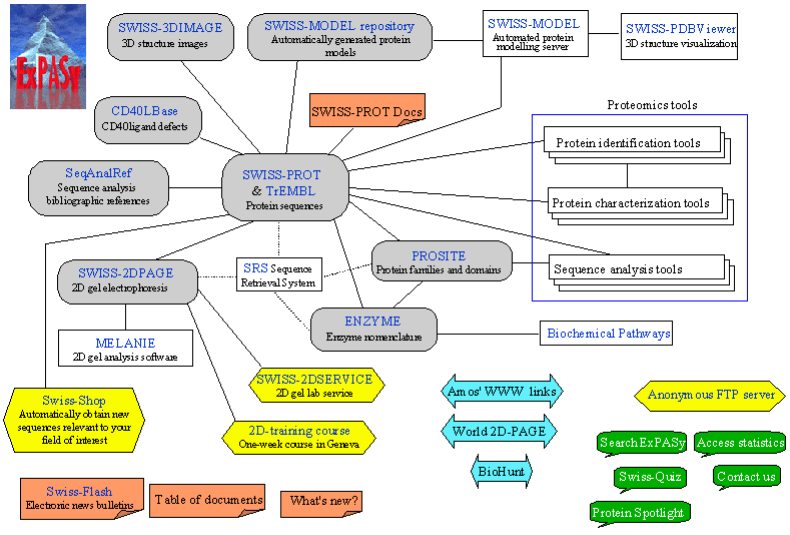
- Meta Databases, e.g. dbCat
  http://www.bioinformatics.vg/biolinks/bioinformatics/verbose/Databases.shtml
- Sequence DB, eg. EMBL or GenBank
  http://www.ebi.ac.uk/embl/
- Protein DB, e.g. Swiss-Prot

  much less entries: Swiss-Prot: $15*10^3$

  http://www.expasy.org/
- Secondary protein structures
  http://www.sander.ebi.ac.uk/dssp/
- Species specific DB, e.g. Flybase
  http://flybase.bio.indiana.edu/

Contain typically many correlated "simple" databases

Swiss-Prot portal



1.2 Types, structure and usage of databases

Content

- DNA Sequencing / Analysis
- Protein Structure Prediction
  - eg. homology based:
    given the coding sequence, find similar sequences which code a protein with known structure
    "similarity search" : very important compared to standard DBS
- Phylogenetic Trees
  - modeling of evolution of protein codes in DNA, DB play some role when search for mutations
- Metabolic pathways
  - find the "path" of metabolic processes within a cell
- Gene Expression

classification basically according to Leser / 2003

DB types: content

- Static Data
  - "Genotype data", i.e. data on bio entities like DNA sequences, genes, proteins and their relationships
- Dynamic data
  - data on phenotypes, data about the dynamics of biological processes
- Data on analysis (software) tools
- Annotations and Scientific papers
  - textual descriptions and explanations about the data above

DB types: active - passive

- Active
  - Data gathering, from journals and other sources
  - target: integration, completeness , central access to distributed data
  - Example: SWISS-PROT
- Passive
  - Data submitted
  - Archiving function
  - Identification (ID assignement)
  - Example GenBank / EMBL
- Mixed forms

DB types: curated - raw

- Curated Database
  - Error correction in data gathered
  - Improvement of data sets – error correction, cross referencing
  - manual integration and cleaning, high cost
  - example: SWISS PROT Versioning essential
- Raw data sets
  - no improvement except by owner of the data
  - submitter is owner ⇨ origin of data is clear
  - primary function is archiving, not improvement
  - Example: GenBank / EMBL

Curated vs archival DBs

- SWISS-Prot

  "The UniProt/Swiss-Prot Protein Knowledgebase is a curated protein sequence database that provides a high level of annotation, a minimal level of redundancy and high level of integration with other databases. UniProt, a "one-stop shop" that allows easy access to all publicly available information of protein sequence annotation"

- EMBL
  - All database records submitted to the INSD will remain permanently accessible as part of the scientific record. Corrections of errors and update of the records by authors are welcome and erroneous records may be removed from the next database release, but all will remain permanently accessible by accession number.
  - … the quality and accuracy of the record are the responsibility of the submitting author, not of the database.

DB types

- Redundant
  - take everything or eliminate "similar" entries?
  - "similar" ?
    - homologous protein in different species?
    - homologous gene in different position?
- Integrated
  - complex net of interrelated objects: difficult to add a new one

  or compilation of data
  - different depth of integration: Links, automatic according to specified criteria, manual

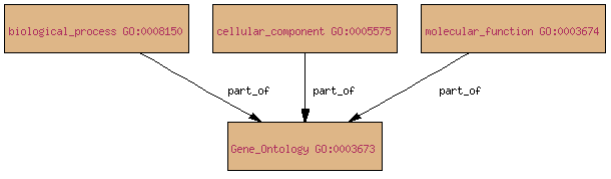
Integration: THE added value of molecular DB (MDB)

Integration by links

- Similarity links between different databases:
  - starting from some entity, find similar entities in different DBs
  - Uses Similarity search tools – e.g. Blasta, Fasta
  - Usually dynamically calculated according to user's metrics
- Biology links
  - relates different kind of entities, e.g. a DNA sequence s, the publications about this sequence and the metabolic processes, s is involved in.

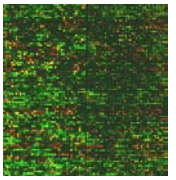
DB types

- Primary / secondary / tertiary
  - data from experiments, few processing steps, no data cleansing
  - integration, curation, data improvement
  - combining processes, functions, components
    e.g. GeneOntology



DB types

- Intention
  - Archive , reference data sets
    - from the 1960ies: Protein Sequence atlas (68-78)
    - published as a book, SWISS-PROT since 1986#
  - Project database
    - short lived, up-to-date, migrate into archival type DB
  - Lab databases
    - integrated with lab information management systems (LIMS: http://www.limsource.com/intro.html), stores raw experimental data, e.g. images of gene expression data (50 MB / image)
    - base for data analysis



1.3 Technical aspects of MDB

Data management

- Standard DBMS (relational, object oriented)
- Management system designed for molecular data (ACEDB)
- Object based system for scientific data (Object Protocol Model OPM)
- Flat files
  using specific representation syntax
  (GenBank: ASN(1), obsolete)
- Native XML databases

Many MDB evolved from ad hoc tools – no systematic devlopment

DBS software

111 randomly selected molecular biology DB
(study by Bry et al., 2001)

- 43 flat file systems
- 42 relational DBMS
- 7 object oriented DBS / tools
- 3 object relational
- 16 other special purpose (like ACEDB?)

Nearly all have hypertext references to foreign DB

DB size

- MDB 1 – 20 GB
  without raw data
- GenBank more than 100 GB
- SWISS PROT 1...2 GB (Oracle export)

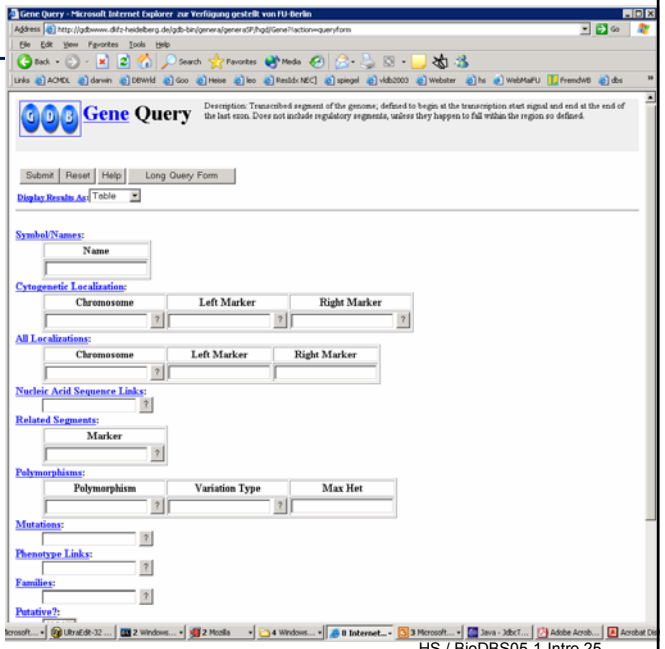
- Much larger volumes of raw data
  - images,
  - sequence traces of sequencing equipment
  - 2D gel images …

User interface

- Retrieval / Querying

  - No SQL :(
    … on the surface

  - restricted by form interfaces in most case
  - easy generation of SQL



http://gdbwww.dkfz-heidelberg.de/gdb-bin/genera/generaSF/hgd/Gene?!action=queryform

User Interface

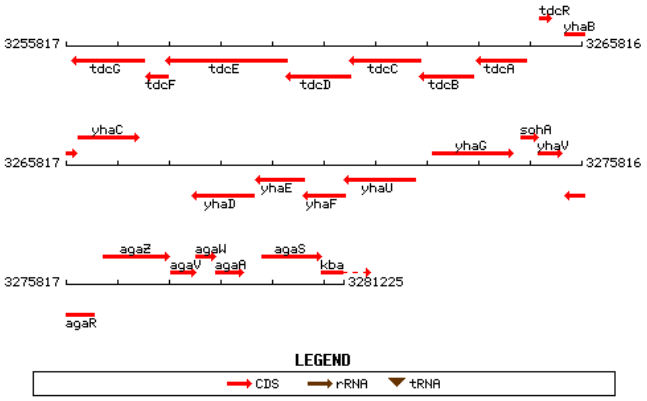
– graphical browsing interface
http://genolist.pasteur.fr/Colibri/

Web interface is the standard

– "Download interface"

– similarity search
using information retrieval methods

Most important: Information Retrieval methods



Synonym *b1999*
Type CDS
Mnemonic Systematic nomenclature
Accession number EG13384
Cross-references

SWISS-PROT	Blattner	GenBank
P76359	b1999	g1788308

Institute Pasteur:
http://genolist.pasteur.fr/Colibri/

User interface

- Retrieval by keyword
  - Boolean retrieval with pattern matching semantics on constituents (words, strings…) using boolean operators
  - vector space model or probabilistic retrieval model
  - ranking of results according to system defined similarity function

Appropriate for character based data

but…

- terminology, abbreviations,…
- which similarity functions are appropriate

Summary: Structure and use of Bio-DBs

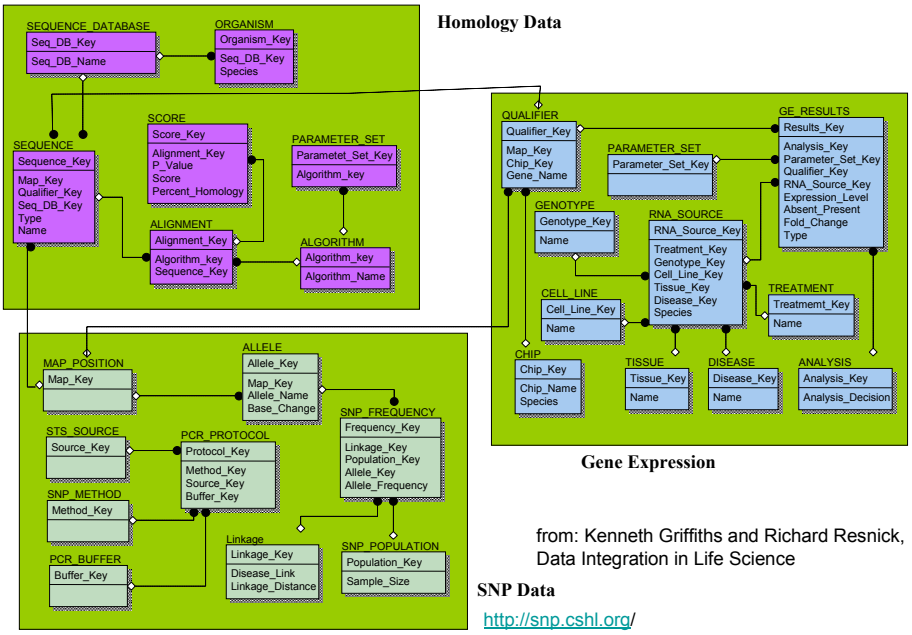
- NO common structure
  - no common schema for the same type of content, sometimes not even related
  - Relational DB used as reliable container
- Important goal: "Interoperability"
  how to use several DB within the one context? Important: the WEB!
- … and Data Integration
  how to virtually or physically integrate many databases into one?
- Usage
  - no transactions -- concurrent update of no importance
  - information retrieval, browsing
  - Interface to DB: analyzing software (e.g. for alignment)

1.4 Key problems of MDB

Diversity and heterogeneity of data

- hard facts
  - sequences
  - what else?
- experimental data
- interpretations (e.g. in publications)
- wrong data?
- Semistructured data
  - relational appropriate?
    - ⇨ Object oriented data modeling, XML
- Independent, related data sets

Schemas for three independent Genomic systems



http://snp.cshl.org/

from: Kenneth Griffiths and Richard Resnick, Data Integration in Life Science

Data Integration

Biggest problem: semantic heterogeneity

Data

- are from a variety of incompatible sources
- have no standard naming convention
- are inconsistent among species

	Protease	Inhibitor	signal
Fruit fly	tolloid	Sog	dpp
frog	Xolloid	Chordin	BMP2/4
Zebrafish	Minifin	Chordino	swirl

c.f. Schütze, Course Text Retrieval, Data Mining

- abbreviations (1990 -2000; ~30000 -> 50000)
- heavily overloaded: PCA has more than 60 expansions (!)

`para-chloramphetamine ,... , post conceptional age`

Data Integration

- Semantic conflicts due to different scientific viewpoints (!)
  - Gene:
    "DNA fragment which can be transcribed and translated into a protein" (Def. of GDA – the Human Genome DB)
    " DNA fragment carrying a genetic trait of phenotype" including introns(!) (Def. of Genbank)
- Methodologies
  - "Ontologies": conceptual description of words and their relationships
  - Statistical techniques, data mining
  - using standardized descriptive information (metadata)
    ⇨ XML and its application in biosciences

Data Modeling

Today: Many ad hoc techniques

- GenBank using ASN(1)
- Relational DB: questionable
- Object orientation: sufficient?
- Modeling of processes ?
- lack of standards (at least in the early phase)

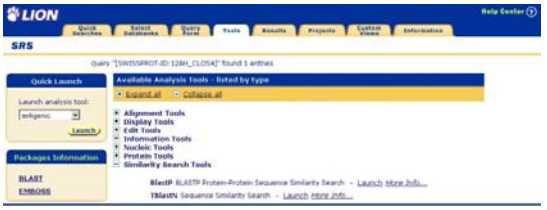
To come:

- widely accepted exchange formats ⇨ XML
- Data management tools for these formats (XML DB !?)
- information extraction from non-structured data (texts)

Query and exploration language

- Should keep the goodies of SQL
  - fine granular search predicates
  - powerful set operations
  - easy to use
- … integrate domain-specific analysis tools
- …be able to navigate (e.g. web references)
- … integrate retrieval and processing of data by analysis tools

  Example: SRS –



Data correctness

- Empirical data always fuzzy
  - measurements imprecise in principle -> statistical techniques to calculate errors
  - unclear mechanisms underlying data in molecular biology
  - heuristic techniques, e.g. image analysis
- Fuzzy data: model immanent, to be quantified
- Incorrect data:
  - in most cases not intentionally
  - difficult to assess ⇨ data cleansing techniques for data bases
  - Danger: propagation of errors
    protein functions: "30 % of annotations are incorrect"

[Science] – a procedure whose rationality consists in the fact that we learn from our mistakes
(Karl Popper: The Growth of Scientific Knowledge)

Versioning

Improvement of data quality, growth of data sets

⇨ new interpretations

Analytical investigations use latest known data sets

⇨ not reproducible if data sets have not been conserved

Versioning is essential!

What does versioning mean?

store differences between objects / attributes?
keep data accessible?

make differences transparent?

Versioning: another dimension of complexity in management of molecular data

Course plan

- Computer Science techniques
  - used in current systems (except relational DB)
  - useful for solving the problem / improving solutions
- Information Retrieval techniques
- Object oriented data management
- XML
- Data Integration
- Data Mining