Bio Data Management

Kapitel 2

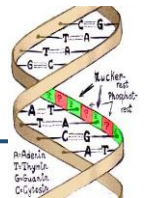
Bio-Datenbanken

Wintersemester 2014/15
Dr. Anika Groß
Universität Leipzig, Institut für Informatik, Abteilung Datenbanken
http://dbs.uni-leipzig.de

UNIVERSITÄT LEIPZIG

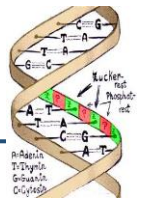
Vorläufiges Inhaltsverzeichnis

1. Motivation und Grundlagen
2. Bio-Datenbanken
3. Datenmodelle und Anfragesprachen
4. Modellierung von Bio-Datenbanken
5. Sequenzierung und Alignments
6. Genexpressionsanalyse
7. Annotationen
8. Matching
9. Datenintegration: Ansätze und Systeme
10. Versionierung von Datenbeständen
11. Neue Ansätze



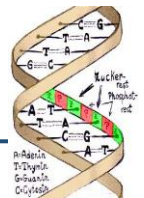
Lernziele

- Wiedergabe von
  - Allgemeinen Kenntnissen zu Eigenschaften von biologischen Daten und typischen Problemfeldern
  - Anforderungen an Bio-Datenbanken



Gliederung

- Motivation und historische Entwicklung
- Biologische Daten
- Klassifikation nach Inhalt
- Anforderungen/Eigenschaften

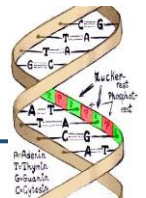


Ziele

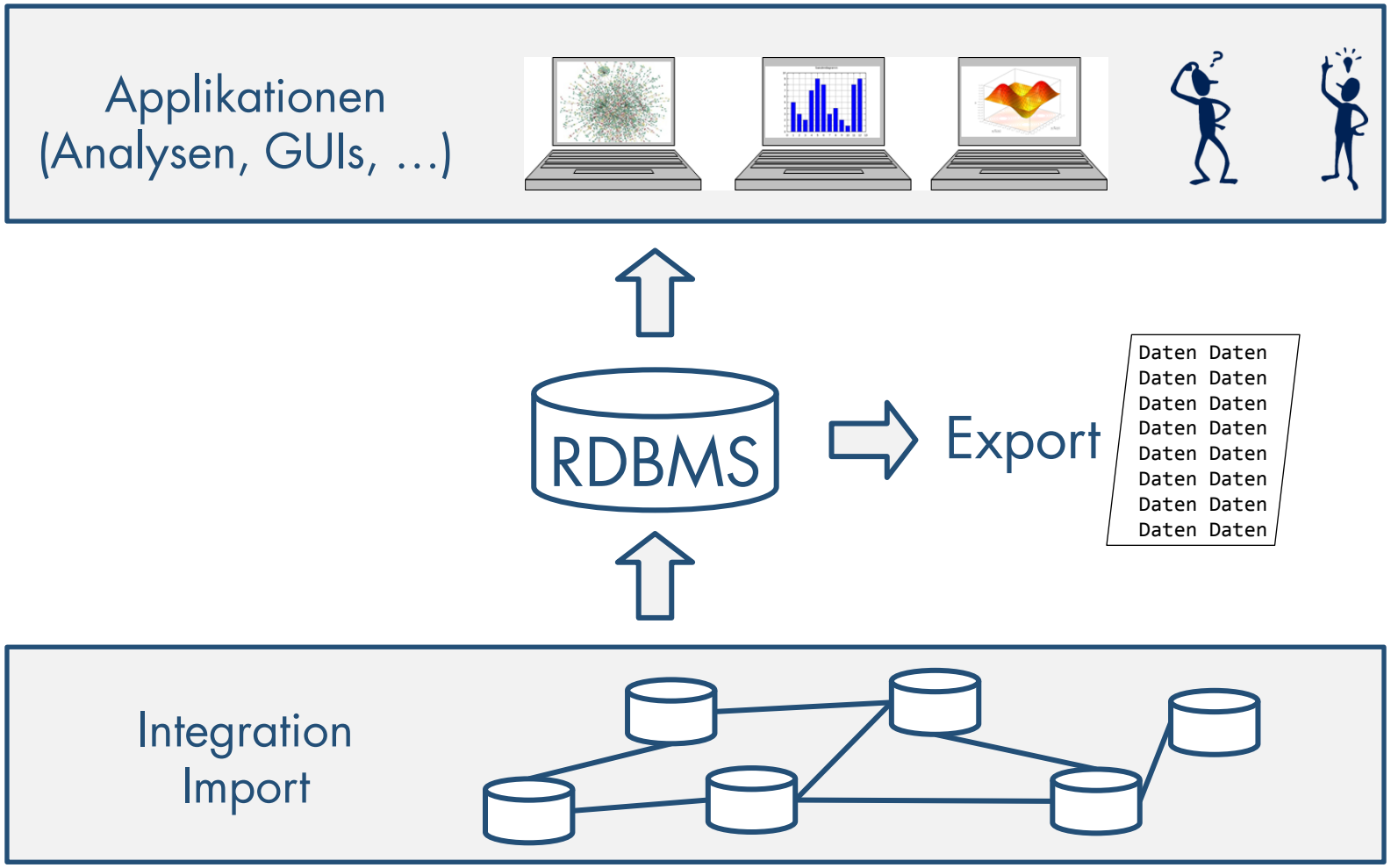
- Abspeicherung von Genom-, Protein-, Stoffwechsel-, Experiment-, … Informationen in konsistenter und effizienter Art und Weise
- Unterstützung von biowissenschaftlichen Anfragen und Analysen

  Beispiel: **Insulin**
  - Identifizieren Sie die **Insulin** mRNA und Proteinsequenz für Mensch, Huhn und Schwein!
  - In welche Stoffwechselwege ist **Insulin** eingebunden?
  - Auf welchem Chromosom liegt (das Gen für) **Insulin** im Menschen?
  - Gibt es eine Krankheit, die auf einer Mutation in **Insulin** beruht?

- Integration verschiedener Datenarten
  - Experimentelle Rohdaten (z.B. Bitmaps bei Genexpressionsdaten)
  - Aufbereitete Experimentdaten (z.B. Gen- oder Proteinsequenz)
  - Textuelle Kommentare (Annotationen)
  - …



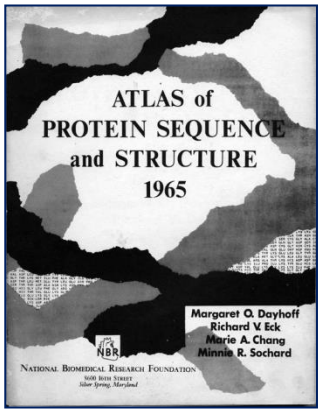
Beispiel


Applikationen
(Analysen, GUIs, …)
RDBMS
Export
Daten Daten
Daten Daten
Daten Daten
Daten Daten
Daten Daten
Daten Daten
Daten Daten
Daten Daten
Integration
Import

Historische Entwicklung

- Viele öffentliche Bio-Datenbanken entstanden aus Büchern
- Sammlungen bekannter Daten einer Art
  - DNA, Proteinsequenz, Proteinstruktur, …
  - Buch→ Band→ CD→ FTP→ WWW
  - Jährliches/quartalsweises Erscheinen

- Anfangs Verwendung von flachen, textorientierten Datenmodellen
  - Viele Beschreibungen in freier Textform
  - Für Menschen konzipiert
  - Datenbank = Menge ähnlich strukturierter "Entries"



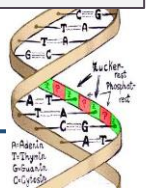
*„In **1984** I received the first available computer tape copy of the Atlas of Protein Sequence and Structure.“* [1]

*“In the summer of **1986** I decided to distribute the protein database [..] so that it would be **available free of charge** to anyone who needed it; I called it **SWISS-PROT**.*

*The **first release** [..] contained just less than **3900 sequences** (the exact number is not known as I have yet to find a copy of the **first floppy disks**!). [..]*

*In Geneva we were wondering what we should do to **make SWISS-PROT available on the Internet** [..] The software was Mosaic and the protocol was the **World Wide Web protocol**. […] The **Expasy Server** (www.expasy.ch) was born on **1 August 1993**.”* [1]

1 Amos Bairoch: Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times!, *Bioinformatics (2000)*



Entry-Modell: „Eine Seite – Ein Objekt“

Swiss-Prot Datenblatt

```
ID   INS_HUMAN               Reviewed;         110 AA.
AC   P01308; Q5EEX2;
DT   21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   21-JUL-1986, sequence version 1.
DT   21-MAR-2012, entry version 168.
DE   RecName: Full=Insulin;
DE   Contains:
DE     RecName: Full=Insulin B chain;
DE   Contains:
DE     RecName: Full=Insulin A chain;
DE   Flags: Precursor;
GN   Name=INS;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RX   MEDLINE=80120725; PubMed=6243748; DOI=10.1038/284026a0
RA   Bell G.I., Pictet R.L., Rutter W.J., Cordell B., Tischer E.,
RA   Goodman H.M.;
RT   "Sequence of the human insulin gene.";
RL
```

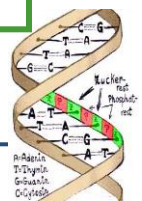
Microsyntax,
Feldabhängige Formate, NF²

Unkontrollierte
Vokabulare

Eingebettete Objekte
(keine Verweise)

Line codes (pre-XML): Referenz auf (Record-)Struktur einer Zeile
AC=Accession Code, DE = Description, DT = Date ,OS = Organism, OC =Taxonomy

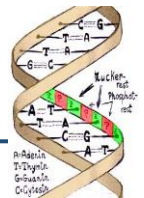
- mehr in Kapitel 3



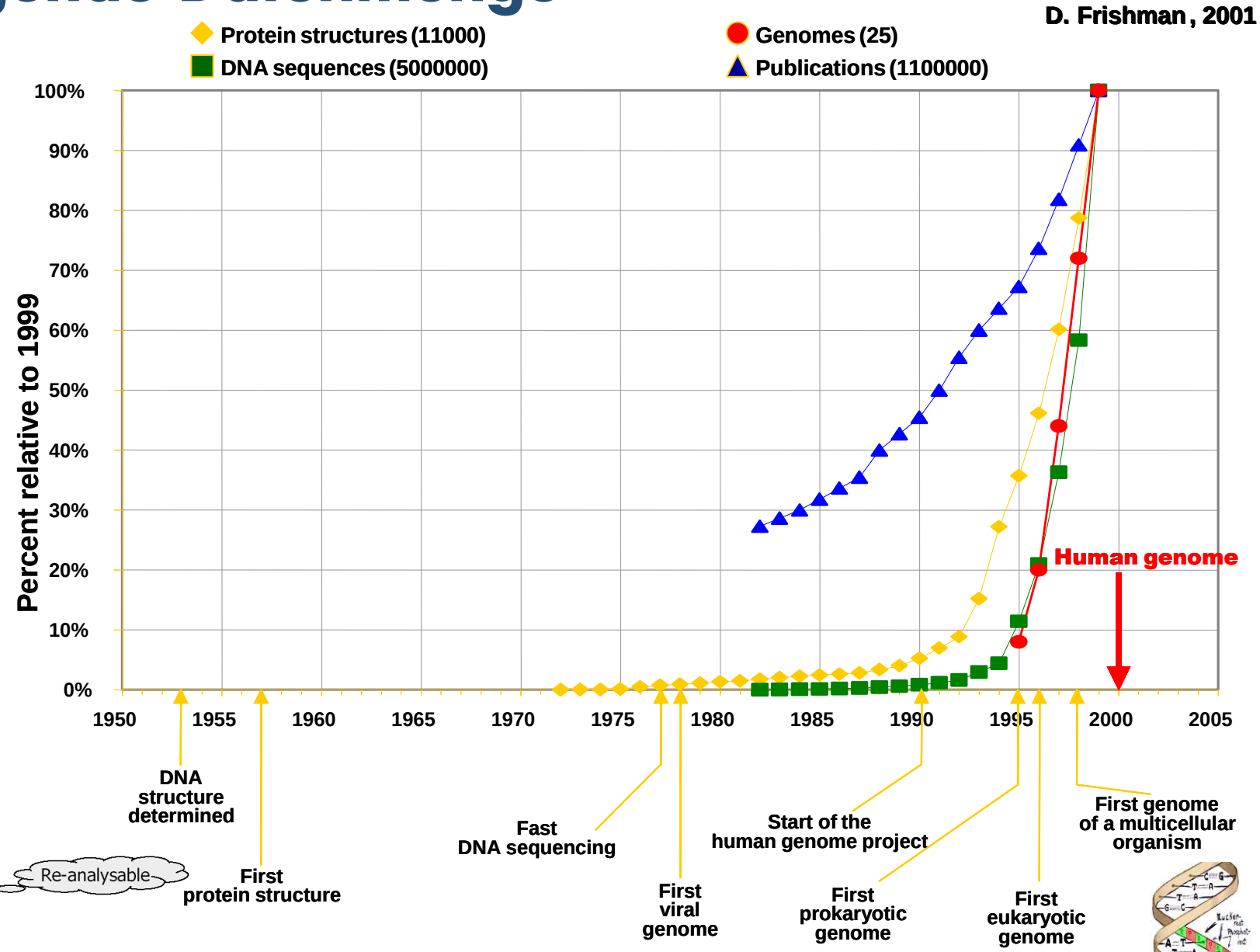
Modeltechnische Entwicklung

Aspekt	Entwicklung
Format / Struktur	Frei → definierte Felder / Entries → XML → strukturierte Daten
Vokabular / Syntax	Frei → "Controlled Vocabularies" → Ontologien
Modellierung	Ad-hoc → ER → OO/UML
Technologie	Proprietär → RDBMS / OO / ORDBMS

Sukzessive Übernahme von DB-Techniken



Steigende Datenmenge



Paradigmen-wechsel:

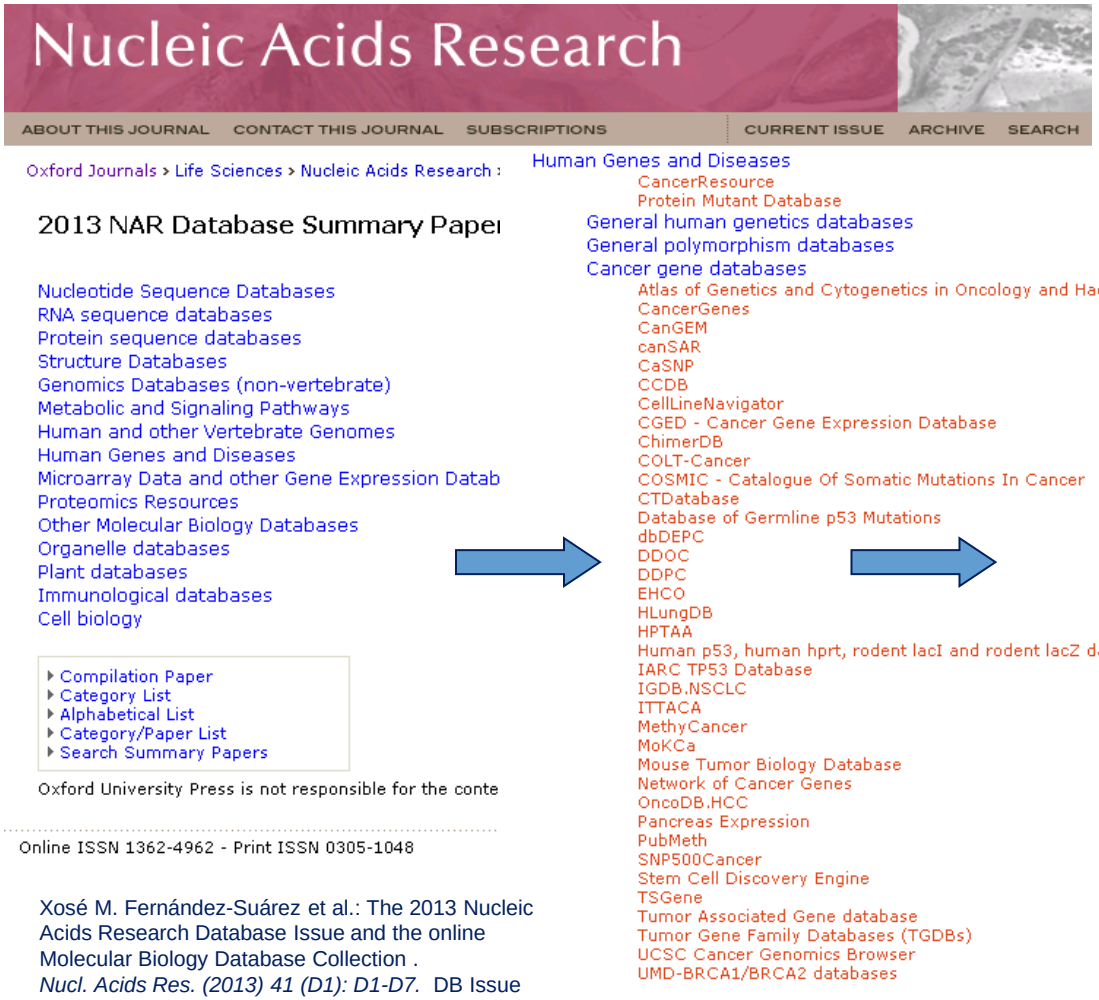
Publishing journals

Re-analysable

Publishing data

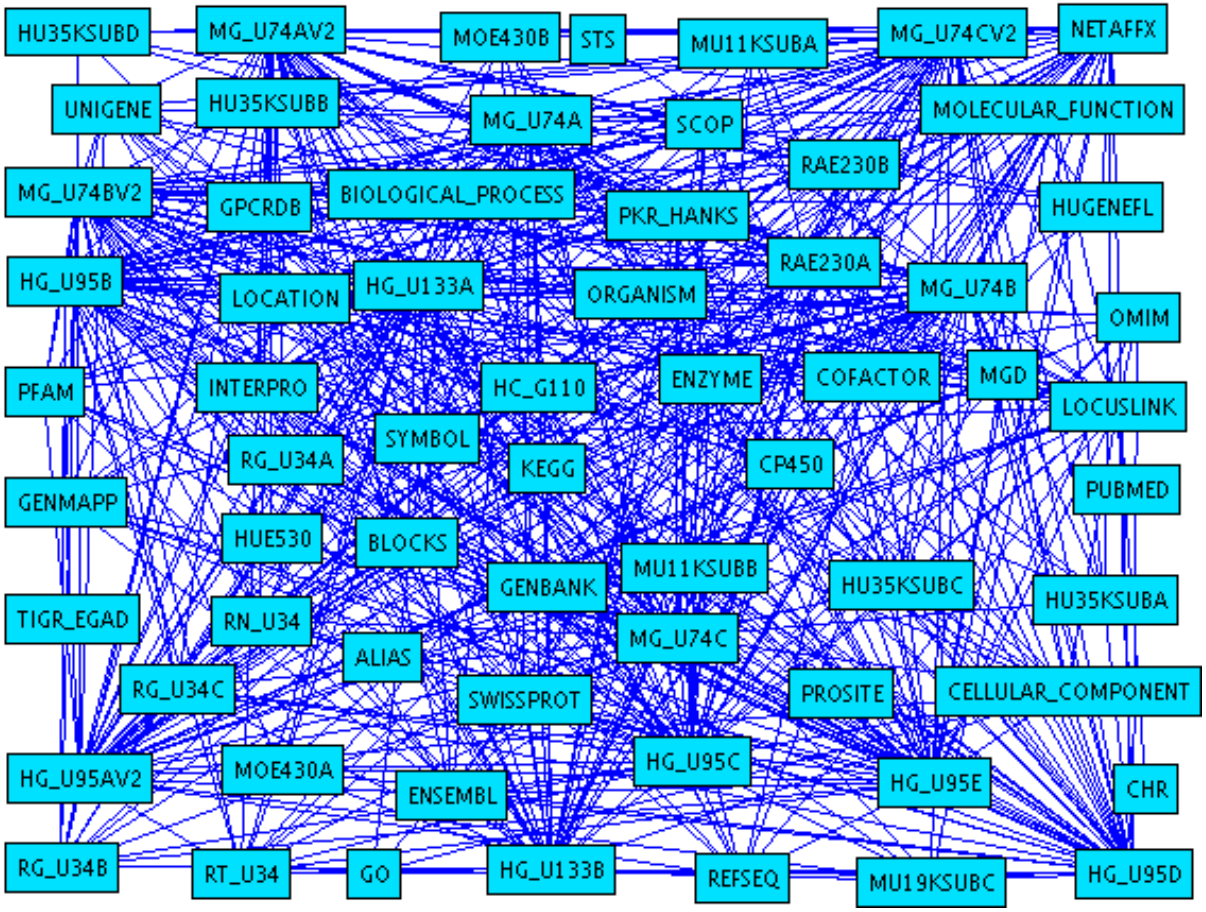
Viele verschiedene Bio-Datenbanken ...

- Weltweit große Anzahl an Quellen online verfügbar
- z.B. sind derzeit 1512 online Datenbanken bei NAR aufgelistet



Xosé M. Fernández-Suárez et al.: The 2013 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection . *Nucl. Acids Res. (2013) 41 (D1): D1-D7.* DB Issue

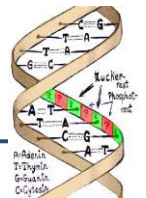
Hohe Vernetzung der Datenquellen



- BioDBs enthalten Links zu anderen BioDBs
- Externe Referenzen durch Verwendung von IDs (Accession Number)
- Keine zentrale Verwaltung / Kontrolle (→ oft keine Konsistenz)

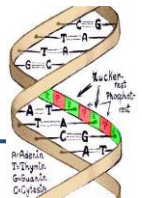
Biologische Daten

- Unterschiedliche Datenarten
  - unstruktruiert, z.B. das Bild eines Genexpressionschips
  - strukturiert, z.B. Nukleotidsequenz, Proteinsequenz
  - semistrukturiert, z.B. Annotationen
- Sehr große Datenmengen
  - Bio-Datenbanken **ohne Experimentdaten** im Bereich 1–200 GB
    - Informationen zu Proteinen, Genen, Pathways etc.
    - z.B. Uniprot: Swiss-Prot + TrEMBL: **>15 GB**
  - **Mit Experimentdaten** deutlich größere Datenmengen
    - TIFF eines Genexpressionschips: ca. 50 MB
    - Rohspektrum eines MS-Experimentes (MS = Massenspektrometrie)
    - Tracefiles von Sequenziermaschinen
    - Bilder von 2D-Gel-Elektrophorese-Experimenten
    - z.B. ArrayExpress: functional genomics data, > 30.000 Experimente, >1 Million Assays, **> 13 TB** archivierte Daten



Art der Datengewinnung

- "Passiv"
  - Alle Daten werden von externen Forschungsgruppen und Institutionen eingebracht ("submittet")
  - Sinn: Archivierung, ID-Vergabe und "roher" Zugriff
  - Auf freiwilliger Basis, oder Verpflichtung durch Geldgeber, Journale ("Publikation nur, wenn Daten eingebracht werden") etc.
  - Beispiele: Genbank/EMBL, PDB, …
- "Aktiv"
  - Relevante (öffentlich zugängliche) Datenquellen werden regelmäßig abgegriffen (z.B. Online-Abstracts bei Bio-Journalen)
  - Sinn: Integration, Veredlung, Vollständigkeit
  - Ermöglicht zentralen Zugriff ohne Verpflichtung
  - Beispiele: Swiss-Prot, Protein Information Resource (PIR), …
- Mischformen

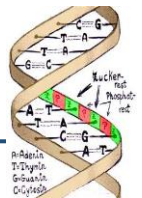


Klassifikation nach Inhalt

- Organismus, Gewebe, Chromosome, …
- Typen der abgespeicherten Bio-Objekte: Sequenzen, Strukturen, Motifs, …

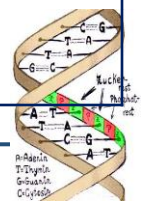
Tertiär-Datenbanken	Ontologie-basiert, strukturierte Annotationen Verteter: GeneOntology, PFAM, PRINTs, InterPro, CATH, ….
Sekundär-Datenbanken	Aufbereitete Daten mit Annotationen (meist nur semi-strukturiert) und Verlinkungen Vertreter: Swiss-Prot, MGD, OMIM, …
Primär-Datenbanken	Enthalten die unmittelbaren Experiment-Daten ("Nah am Experiment") Wenig Verarbeitung, kurze Annotationspipelines Vertreter: Genbank/EMBL, PDB, UniGene

- Grenze vor allem zw. Sekundär- und Tertiärdatenbanken oft fließend



Klassifikation nach Inhalt

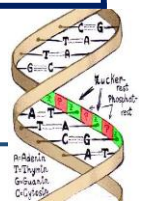
Thema / Daten	Datenbanktyp	Schwerpunkte	Unterstützte Fragestellungen	Vertreter
Genom	Kartierungs-DB	Genlokalisation	Verwandschafts-beziehungen, phylogenetische Stammbäume	Ensembl / UCSC Genome Browser
	Sequenz-DB	Basensequenzen, Nukleinsäuresequenzen		Genbank / EMBL / DDBJ (DNA Data Bank of Japan)
	Krankheits-/ Mutations-DB	Genveränderungen Punktmutationen	Welche Krankheiten sind durch welche Genveränderungen bedingt?	OMIM, dbSNP (Single Nucleotide Polymorphism)
	Genexpressions-Datenbanken	Expressionsniveaus Genexpressionsmuster	Unter welchen Bedingungen exprimiert eine Zelle welche Gene?	GEO, ArrayExpress
Proteine	Proteinsequenz-Datenbanken	Primärstruktur von Proteinen	Proteindesign (z.B. für neue Medikamente)	UniProt/Swiss-Prot
	Proteinstruktur-Datenbanken	Sekundär-, Tertiär- und Quartärstruktur von Proteinen		PDB



Klassifikation nach Inhalt

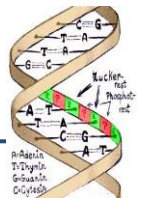
Thema / Daten	Datenbanktyp	Schwerpunkte	Unterstützte Fragestellungen	Vertreter
Protein Interaktionen/ Stoffwechsel	Protein-Domain/ family	Protein-Gruppen	Welche Proteingruppe ist für bestimmte Stoffwechselprozesse (z.B. Blutgerinnnung) zuständig?	PFAM (Protein families database of alignments and HMMs)
	Protein Interaktion	Biomolekulare Interaktionen	Welche Interaktionen bestehen zwischen verschiedenen Proteinen?	PPI, BIND
	Pathway-Datenbanken	Metabolische Pfade Regulatorische Pfade	Welche Stoffwechselprozesse werden von welchen Proteinen (Enzymen) gesteuert?	KEGG (Kyoto Encyclopedia of Genes and Genomes), Reactome
Publikationen	Publikations-datenbank			MedLine PubMed

Keine strikte Klassifikation!
Übergang teilweise fließend, mehrere Bereiche können abgedeckt werden



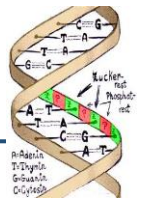
Anforderungen

- Datenqualität
- Offenheit, Verfügbarkeit
- Querying
- Flexibilität
- Performance
- Integration und Datenaustausch
- Analyse
- …



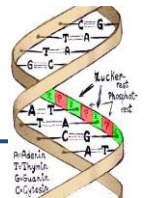
Datenqualität

- Experimente erzeugen (fast) immer unscharfe Daten
  - Arbeit mit lebenden Organismen
  - Zugrundeliegende Mechanismen größtenteils unverstanden
  - Fehleranfällige Techniken: Bilderkennung, statistische/heuristische Algorithmen, …
- Starke Redundanz
  - Ungewollt: PIR - Swiss-Prot, KEGG - Reactome
  - Gewollt: spezies-spezifische Daten
  - Manche Quellen kopieren von anderen (z.B. Ensembl)
  - Kopieren und Überarbeiten von Daten → Inkonsistenzen (c/p error)
- Eine junge Wissenschaft: viele (falsche) Daten und Veröffentlichungen
  - Viele verschiedene (konkurrierende) Arbeitsgruppen
- Herkunft der Daten sollte (auch bei Ableitungen) ermittelbar sein
- Automatisch berechnete Daten oder „curated“ (=„redaktionell betreut“)
  - Falsch-Positiv-Rate bei High-Throughput Experimenten
  - Curator: liest, fasst zusammen ..
  - Probleme: Konsistenz, Vollständigkeit, Qualitätssicherung, Objektivität,…



Datenqualität (Curation)

- Ansatz 1: (Externer) Einbringer ist "Datenherr"
  (z.B. Genbank, ArrayExpress)
  - Im Nachhinein keine (inhaltlichen) Veränderungen an einmal eingebrachten Daten
  - Vorteil: Urheber klar, hohe Datenstabilität; Nachteil: keine globale Verantwortlichkeit, übergreifende Datenqualität schwierig zu sichern
- Ansatz 2: Zentrale Nachbearbeitung/Kontrolle der Daten
  (z.B. Swiss-Prot, MIPS *(Munich Information Center for Protein Sequences)*)
  - Daten werden laufend verbessert
  - Hoher (manueller) Aufwand, da Automatisierung nur eingeschränkt möglich
  - Vorteil: Höhere Datenqualität; Nachteil: Urheber weniger klar, hohe Volatilität
- Redundanz
  - Ansatz 1: Alles aufnehmen, auch wenn teilweise redundant zu bisherigen Einträgen
  - Ansatz 2: Entfernen gleicher oder sehr "ähnlicher" Einträge
  - Beispiel Swiss-Prot: Redundanzminimierung durch (menschliche) Editoren (sicher, aber teuer)
  - Beispiel UniGene: Redundanzminimierung durch Algorithmen (ökonomisch, aber mit Unsicherheiten behaftet)

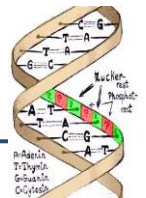


Verfügbarkeit, Zugriffsmethoden

Öffentliche Datenbanken	Nicht-öffentliche Datenbanken
Lange bestehend, international organisierte Referenzdatenbanken, öffentliche Archive (Genbank, Swiss-Prot, PIR, PDB, …)	Projektbezogene ("One-Shot")-Datenbanken von Forschungsgruppen (hochaktuell für kurze Zeit; existieren oft nur bis zur Veröffentlichung der Ergebnisse) Kommerzielle BioDBs von Firmen wie z.B. Celera

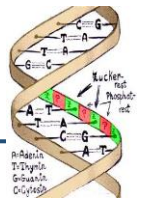
Meist freier Zugriff und Download - „share data“:

- Web-basierte Nutzerschnittstellen
  - Manuelle Suche
  - Automatischer Zugriff (via GET/POST, HTML parsen)
- Nutzen von Libraries, Anfragesprachen
  - BioPerl, BioJava, … enthalten Parser für beliebte BioDBs
  - BioSQL: transformieren typischer BioDBs in relationales Schema
- Webservices, APIs
- Manchmal direkter SQL Zugriff, oft Dumps zum Download
- Einige öffentliche Datensätze auf Amazon Elastic Compute Cloud (EC2) verfügbar (z.B. NCBI UniGene, Ensembl)



Querying

- Bio-Daten werden im Allgemeinen für komplexe Weiterverarbeitungen genutzt (Analyse Workflows)
- Querying-Anforderungen
  - Effiziente Bearbeitung
  - Benutzerfreundlichkeit
    - Vordefinierte (parametrisierbare) Masken für häufige Anfragetypen
    - Möglichkeit, Ad-hoc-Queries komfortabel zusammenstellen zu können (z.B. über grafisches Interface)
    - Interface mit voller Query-Komplexität (für sog. "Power User")
  - Unterstützung von Unschärfe bei unstrukturierten oder semi-strukturierten Daten



Web Interfaces - Browsen

- Browsen in den Datenbeständen über Links
- Unterstützung der Suche durch Ontologien
- *Ontologie: Explizite begriffliche Formalisierung eines Anwendungsbereiches., d.h. eine explizite Spezifikation von Begriffen (concepts) und deren Beziehungen in einem Bereich (domain)*

Tree Browser AmiGO

Browse a Genome

Genome assembly: GRCh37 (GCA 000001405.11)

More information and statistics

Download DNA sequence (FASTA)

Convert your data to GRCh37 coordinates

Display your data in Ensembl

Other assemblies

- NCBI36 (Ensembl release 54)

Filter tree view

Filter by ontology: Ontology — All, biological process, cellular component, molecular function

Filter Gene Product Counts: Data source — All, ASAP, AspGD, CGD; Species — All, Arabidopsis thaliana, Aspergillus fumig..., Aspergillus niger

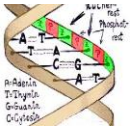
View Options: Tree view — Full / Compact

Remove all filters

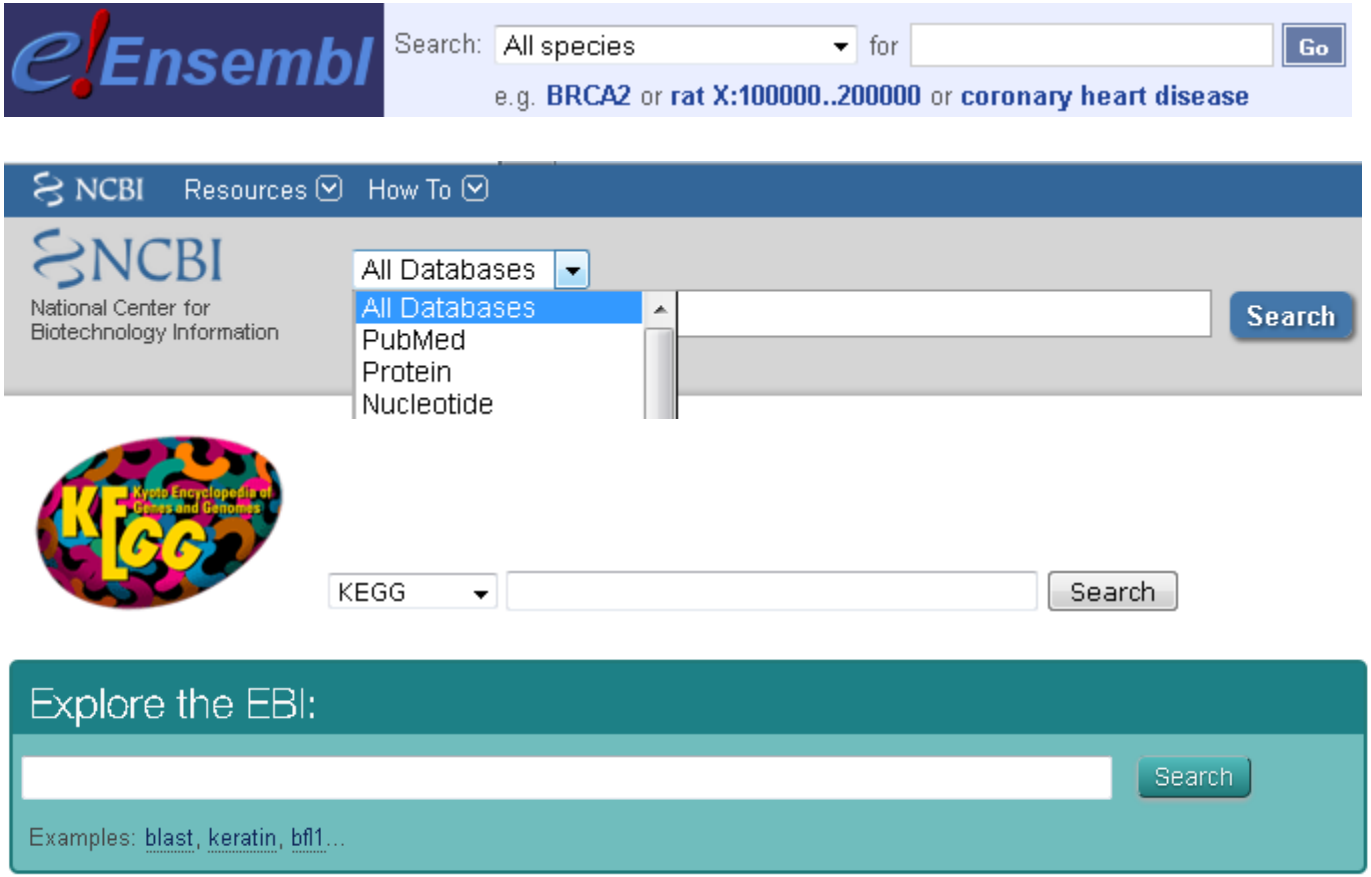
all : all [644210 gene products]

- GO:0008150 : biological_process [500364 gene products]
- GO:0005575 : cellular_component [459348 gene products]
- **GO:0003674 : molecular_function [525747 gene products]**
  - GO:0016209 : antioxidant activity [3273 gene products]
  - GO:0005488 : binding [229340 gene products]
  - GO:0003824 : catalytic activity [207957 gene products]
  - GO:0016247 : channel regulator activity [805 gene products]

Actions...
Last action
GO:00036
Graphical
Permalink
Download.
OBO
RDF-XML
GraphViz

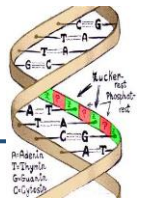


Web Interfaces - Stichwortsuche



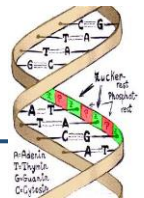
Stichwortsuche

- Typische Zugriffsmöglichkeit im Web (Google, etc.)
  - Einfach, Schnell, Verständlich, Bekannt
- "Google-Effekt": geringe Akzeptanz nicht-stichwortbasierter Interfaces
- Verwendung von Methoden des Information Retrieval
  - Ranking der Ergebnisse (fehlt oft)
  - Operatoren zur Verknüpfung: AND, OR, NOT, + / -
- Probleme
  - Suchergebnis nicht zwingend Treffer
  - Wortformen: Zeiten, Singular / Plural, Casus, …
  - Synonym / Homonymprobleme
  - Treffer sind Dokumente, nicht Attribute
- Geeignet für Menschen, aber nicht gut geeignet zur automatischen Weiterverarbeitung (Joins, …)



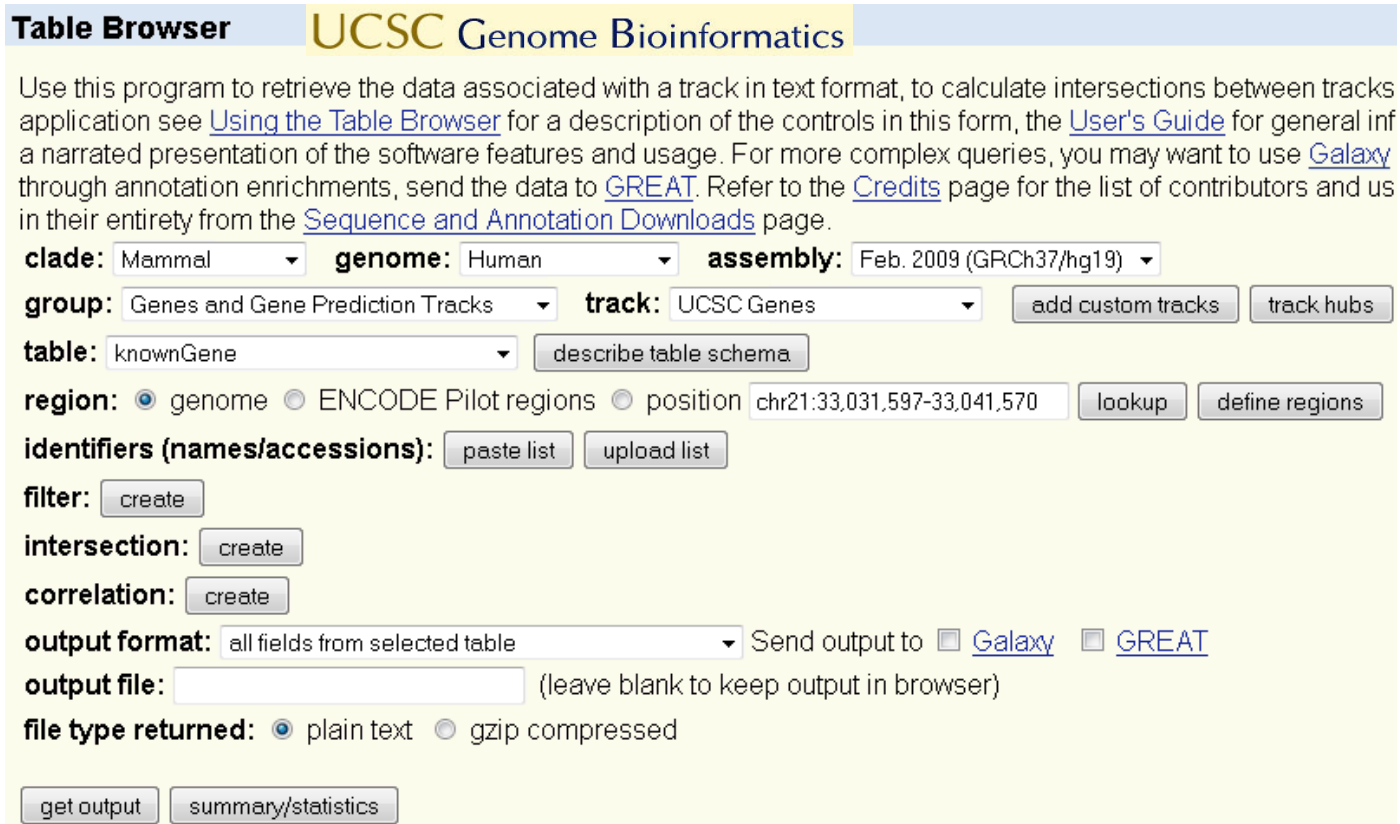
Anfragesprachen

- Unterstützen (semi-)strukturierte Anfragen
- Basieren üblicherweise auf Mengenoperationen
- Vertreter
  - SQL92 (relational), SQL:1999 (objekt-relational)
  - OQL (objektorientiert)
  - XPath / XQuery (XML-basiert)
- Typische Elemente (SQL92)
  - Select: Auswahl dessen, was Ergebnis ausmacht
  - From: Auswahl der Datenherkunft
  - Where: Auswahl der Bedingungen, die Ergebnisse erfüllen müssen (Filterkriterien)
- Spezialsprachen (z.B. GQL: Genom Query Language)
- Kaum Verwendung als "öffentliches" Interface, da zu komplex



Suchformulare

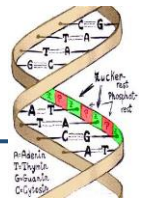
- Anfragen erfolgen oft in vorstrukturierten Suchformularen ("Canned Queries")
  - Select / From: Meist fest innerhalb des Suchformulars
  - Where: Wahlmöglichkeiten bzgl. der Werte, der Vergleichsoperatoren und der Verknüpfungen zwischen einzelnen Bedingungen
- Benutzerunterstützung durch Drop-Downlisten, Checkboxes, …
- Einfache Übersetzung in strukturierte Anfragesprache



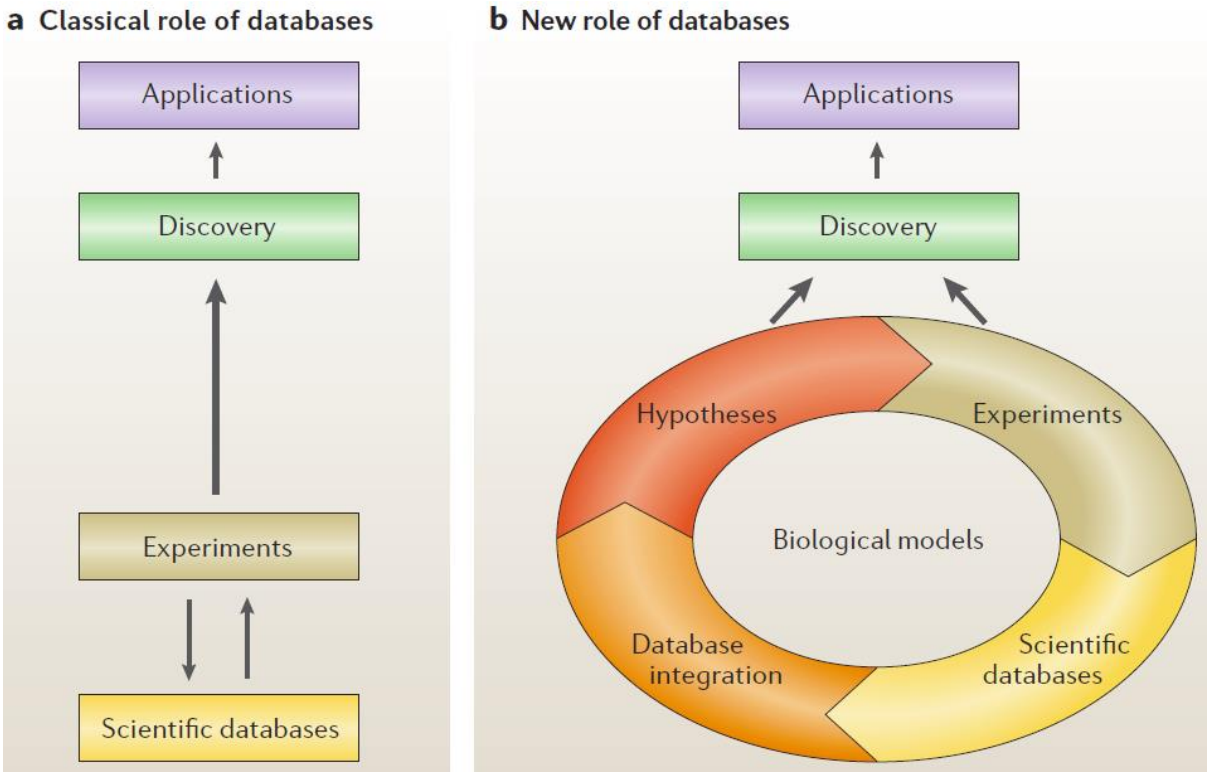
- Vorteil: Einfach zu realisieren, benutzerfreundlich, effizient
- Nachteil: Starke Einschränkung der Expressivität, keine (flexible) Unterstützung von komplexen Anfragen

Flexibilität

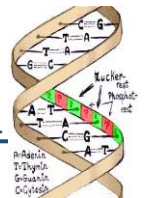
- Forschungsfragen ändern sich ständig (regelmäßig neue Versionen)
  - Andere wissenschaftliche Fragestellungen
    → Andere Daten, andere Queries
- Design muss Wartbarkeit und Flexibilität in Vordergrund stellen
  - Schemaänderungen, Einbringung neuer Datentypen, Optimierung auf neue Anforderungen
- Bio-Datenbanken meist Teil eines Forschungsprojekts
  - Datenbeschaffung (LIMS), Datenarchivierung, Datenanalyse
- Flexible Analysemöglichkeiten notwendig



Klassische vs. Neue Rolle von BioDBs



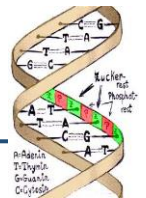
- Klassisch: einfacher Zugriff auf experimentelle Ergebnisse, Langzeitspeicherung der Ergebnisse; zentrales Datenmanagement
- Aktuellere, systembiologische Ansätze generieren automatisch Hypothesen aus den Informationen in DB, welche experimentell verifiziert werden und wiederum in andere DB einfließen → DATENINTEGRATON!



Philippi, Köhler: Addressing the problems with life-science databases for traditional uses and systems biology. *Nature Reviews Genetics*, 2006.

Integration

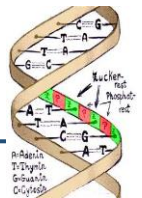
- Viele Daten machen erst Sinn im Kontext
  - (Teil-)Sequenz: Genkontext, Regulationskontext, Homologie
  - Protein: Welcher Organismus?, Strukturkontext, Domänen
  - Expression: Regulationskontext, Phänotypen, Krankheitsverläufe, …
- Integration von Bio-Daten aus externen Quellen
- Die meisten Datenbanken sind "nur" integriert im Sinne einer Verlinkung z.B. Verlinkung Ensembl ↔ Swiss-Prot ↔ OMIM
- Integration im Sinne eines globalen Schemas oft nicht vorhanden
- Typische Bio-Anfragen implizieren bereits Zugriffe auf mehrere Quellen
- Beispiel: **Insulin**
  - Identifizieren Sie die **Insulin** mRNA und Proteinsequenz für Mensch, Huhn und Schwein *(DB: NCBI-Entrez, GeneCards, UCSC Genome Browser, NCBI-GenBank (für Nucleotide), NCBI-GenBank (für Proteine)*
  - In welchem Stoffwechselweg ist **Insulin** eingebunden? *(DB: KEGG)*
  - Auf welchem Chromosom liegt **Insulin** beim Mensch? *(DB: NCBI-Entrez, GeneCards)*
  - Gibt es eine Krankheit, die auf einer Mutation in **Insulin** beruht? *(DB: NCBI-OMIM)*



Art der Integration

- Ansatz 1: Virtuelle (logische) Integration
  - "lockerer Verbund" zwischen Datenquellen, deren Objekte durch Links/Verweise miteinander verbunden sind
  - Mediator-Wrapper-System
  - Häufigste Integrationsart in Bio-Datenbanken
- Ansatz 2: Materialisierte (physische) Integration
  - Daten werden kopiert und zentral aufbereitet
  - Data Warehouse
- Beide Ansätze mit oder ohne globales Schema

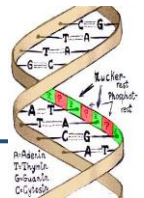
- Manuelle versus automatische Integration
  - Automatische Integration anhand def. Kriterien (Ensembl)
  - Manuelle Integration anhand Wissen des Editors (Swiss-Prot)



Analyse

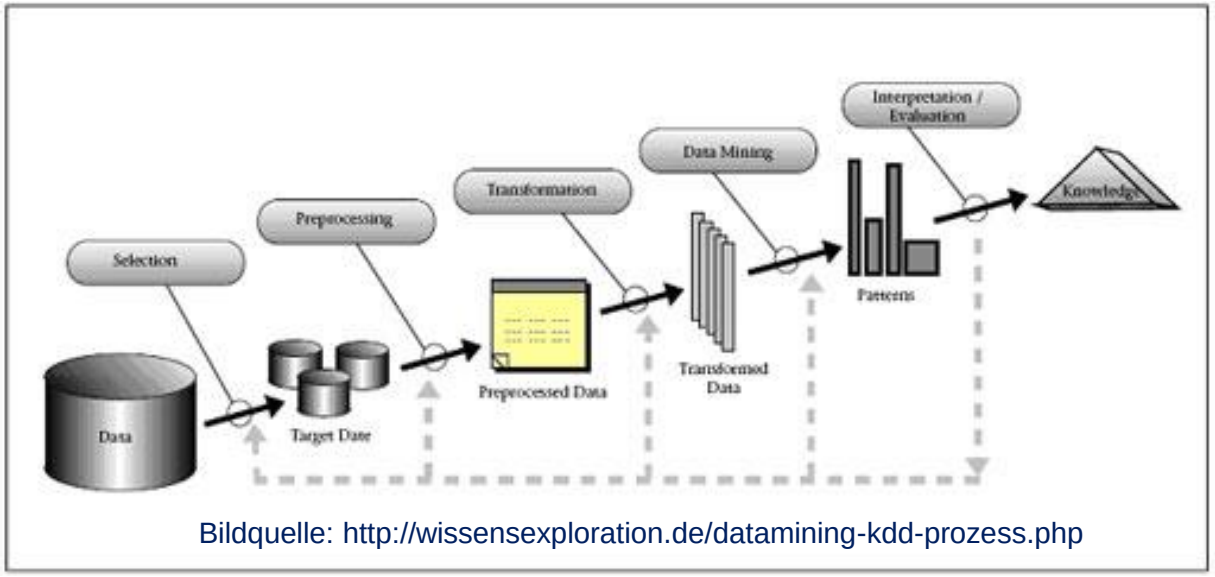
- Neben manueller Suche → automatische Auswertung der Daten

- Integration mittels Data Warehouse-Ansätzen (multidimensionale Anfragen, Aggregation)
- Integration mit (selbstentwickelten) Analyse-Algorithmen nötig
  - z.B. Algorithmen für Ähnlichkeitssuche bzgl. Genen und Proteinen (Blast/Fasta), Motif-Suche, Gensuche …
- Integration von Statistik und Data Mining Tools



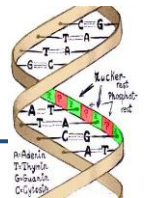
Data Mining

- Knowledge Discovery in Databases (KDD): Prozess der (semi-) automatischen Extraktion von Wissen aus Datenquellen, das
  - gültig (im statistischen Sinn)
  - bisher unbekannt
  - und potentiell nützlich ist

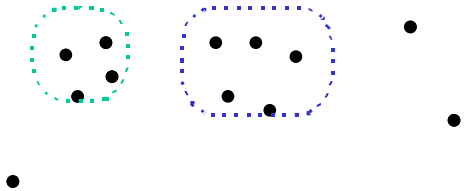


Bildquelle: http://wissensexploration.de/datamining-kdd-prozess.php

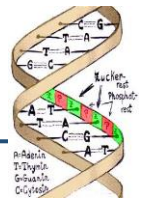
- Kombination von Verfahren zu Datenbanken, Statistik und KI (maschinelles Lernen)



Data Mining

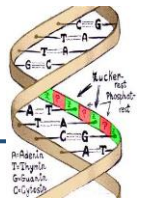


- Data Mining: Anwendung eff. Algorithmen, die in DB enthaltene Muster liefern
- Clusteranalyse, Klassifikation, Assoziationsregeln …
- Oft Mining auf speziell aufgebauten Dateien
- Notwendig: Data Mining auf Datenbanken bzw. Data Warehouses
  - Skalierbarkeit auf großen Datenmengen
  - Nutzung von Performance-Techniken (Indizes, materialisierte Sichten, Parallelverarbeitung)
  - Vermeidung von Redundanz und Inkonsistenzen
  - Integration mehrerer Datenquellen, Portabilität
- Datenaufbereitung für Data Mining
  - Datenintegration und Datenbereinigung (data cleaning)
  - Diskretisierung numerischer Attribute (Aufteilung von Wertebereichen in Intervalle, z.B. Genexpressionsgruppen)
  - Erzeugen abgeleiteter Attribute (z.B. Aggregationen für bestimmte Dimensionen, Genexpressionsänderungen)
  - Einschränkung der auszuwertenden Attribute



Zusammenfassung

- Bio-Datenbanken
  - Große Datenmengen, viele Datenquellen, forschungsgetrieben, forschungsbegleitend, ständige Veränderung, …
- Anforderungen
  - Datenqualität
  - Offenheit, Verfügbarkeit
  - Querying
  - Flexibilität
  - Performance
  - Integration und Datenaustausch
  - Analyse



Fragen ?