

Kapitel 2: Bio-Datenbanken Überblick

n Inhalt

- Motivation
- Historische Entwicklung
- Anforderungen
- Klassifizierungsmerkmale
- Zusammenfassung



Motivation

- n Abspeicherung von Genom-, Protein- und Stoffwechsellinformationen in konsistenter und effizienter Art und Weise

- n Unterstützung von biowissenschaftlichen Anfragen und Analysen
 - Beispiel: Hypoxanthin-Guanin Phosphoribosyltransferase (HPRT)
 - Identifizieren Sie die HPRT mRNA und Proteinsequenz für Mensch, Maus und Ratte
 - In welchem Stoffwechselweg ist HPRT eingebunden?
 - Auf welchem Chromosom liegt (das Gen für) HPRT im Menschen?
 - Gibt es eine Krankheit, die auf einer Mutation in HPRT beruht?
 - In welchen Kontexten (z.B. bei welchen Krankheiten) wird HPRT verstärkt gebildet?

- n Integration verschiedenster Datenarten
 - Experimentelle Rohdaten (subsymbolisches Level, z.B. Bitmaps bei Genexpressionsdaten)
 - Aufbereitete Experimentdaten (symbolisches Level, z.B. Gen- oder Proteinsequenz)
 - Textuelle Kommentare (Annotationen)



Historische Entwicklung

- n Alle (großen) öffentlichen Bio-Datenbanken entstanden aus Büchern
- n Sammlungen bekannter Daten einer Art: DNA, Proteinsequenz, Proteinstruktur
 - Jährliches / quartalsweises Erscheinen
 - Buch → Band → CD → FTP → WWW
- n Anfangs Verwendung von flachen, textorientierten Datenmodellen
 - Viele Beschreibungen in freier Textform
 - Für Menschen konzipiert, nicht für Weiterverarbeitung durch Computer
 - Datenbank = Menge ähnlich strukturierter "Entries"
- n Entry-"Modell"
 - Entry: Menge von Feldern (Attribute, Lines) zu einem Bio-Objekt (z.B. zu einem Protein)
 - Von nahezu allen Bio-Datenbanken verwendet
 - Kein Datenmodell im engeren Sinn (wie z.B. RM, OO)
 - Keine deklarativen Konsistenzbedingungen, kein Klassen- oder Objektbegriff



Entry-Modell

n Beispiel Swiss-Prot^{*}; Hanukkah-Faktor (Zytotoxische T-Lymphozyten Proteinase)

```

ID      GRAA_HUMAN          STANDARD;          PRT;   262 AA.
AC      P12544;
DT      01-OCT-1989 (Rel. 12, Created)
DT      01-OCT-1989 (Rel. 12, Last sequence update)
DT      16-OCT-2001 (Rel. 40, Last annotation update)
DE      Granzyme A precursor (EC 3.4.21.78) (Cytotoxic T-lymphocyte p.
DE      1) (Hanukkah factor) (H factor) (HF) (Granzyme 1) (CTL tryptase)
DE      (Fragmentin 1)
GN      GZMA OR CTLA3 OR HFSP.
OS      Homo sapiens (Human).
OC      Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC      Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OC      NCBI_TaxID=9606;
RN      [1]
RP      SEQUENCE FROM N.A.
RC      TISSUE=T-cell;
RX      MEDLINE=88125000; PubMed=3257574;
RA      Gershenfeld H.K., Hershberger R.J., Shows T.B., Weissman I.L.;
RT      "Cloning and chromosomal assignment of a human cDNA encoding a T
RT      cell- and natural killer cell-specific trypsin-like serine
RT      protease.";
RL      Proc. Natl. Acad. Sci. U.S.A. 85:1184-
RN      [2]
RP      SEQUENCE OF 29-53.
RX      MEDLINE=88330824; PubMed=3047119;
RA      Poe M., Bennett C.D., Biddison W.E., Blake J.T., Norton G.P.,
RA      Rodkey J.A., Sigal N.H., Turner R.V., Wu J.K., Zweerink H.J.;
RT      "Human cytotoxic lymphocyte tryptase. Its purification from granules
RT      and the characterization of inhibitor and substrate specificity.";
RL      J. Biol. Chem. 263:13215-13222(1988).
RN      [3]
...
  
```

Feldabhängige Formate (Microsyntax)

Eingebettete Objekte (keine Verweise)

Line codes: Referenz auf (Record-)Struktur einer Zeile (z.B. AC = Accession Code; DT = Date; DE = Description; OS = Organism; OC = Taxonomy)

n Zum Entry-Modell mehr in Kapitel 3 (Datenmodelle von Bio-Datenbanken)

^{*} Swiss-Prot = Protein knowledgebase



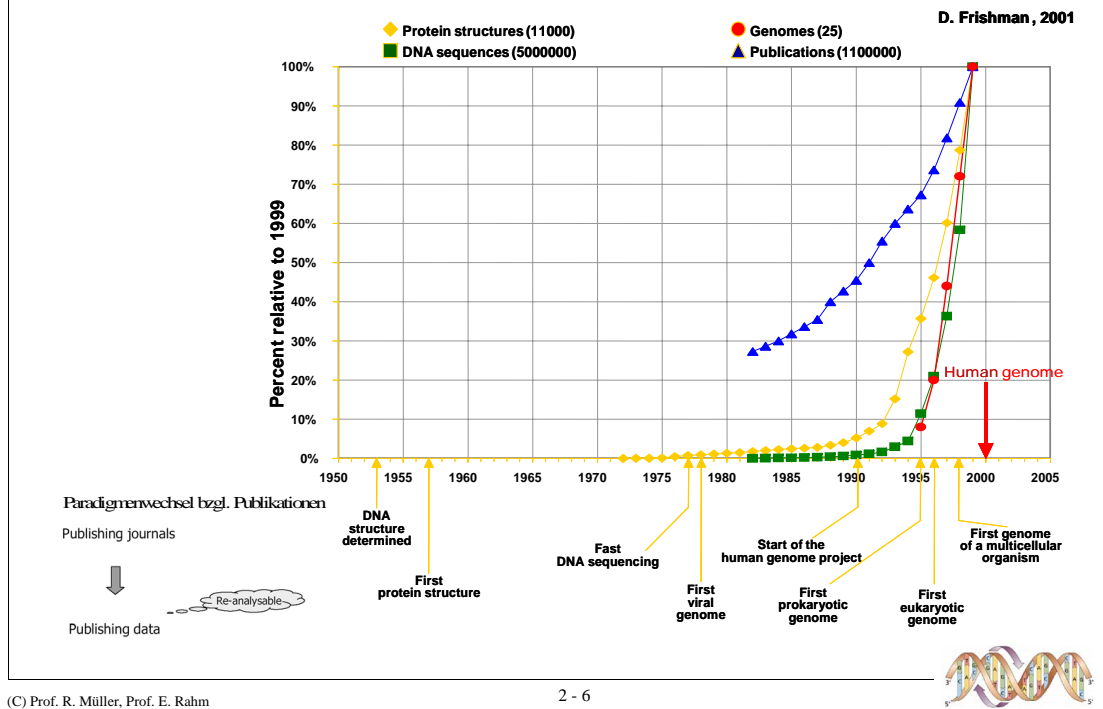
Modelltechnische Entwicklung

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

Sukzessive Übernahme von DB
Techniken



Bio-Daten: Historische Entwicklung



Bio-Datenbanken: Übersicht

n Weltweit derzeit über 500 Bio-Datenbanken

The screenshot shows two browser windows. The left window displays the 'Nucleic Acids Research' website with a 'Database Categories List' on the left sidebar and a 'Database Category: Major Sequence Repositories' section on the right. A blue arrow points from the 'Major Sequence Repositories' link in the sidebar to the corresponding section in the main content area. The right window shows a detailed view of the 'Major Sequence Repositories' section, listing various databases such as 'Ares Lab Yeast Intron Database', 'ArrayExpress', 'DNA Data Bank of Japan (DDBJ)', 'EMBL Nucleotide Sequence Database', 'GenBank', 'STACK', 'TIGR Gene Indices', and 'UniGene'. Each entry includes a brief description and links to the database, a summary, and a paper.

Nucleic Acids Research
HOME HELP FEEDBACK SUBSCRIPTIONS ARCHIVE SEARCH ARTICLES

Database Categories List

- Major Sequence Repositories
- Comparative Genomics
- Gene Expression
- Gene Identification and Structure
- Genetic and Physical Maps
- Genomic Databases
- Intermolecular Interactions
- Metabolic Pathways and Cellular Regulation
- Mutation Databases
- Pathology
- Protein Databases
- Protein Sequence Motifs
- Proteome Resources
- RNA Sequences
- Retrieval Systems and Database Structure
- Structure
- Transgenics
- Varied Biomedical Content

Compilation Paper
Categories List
Alphabetical List
Search Summary Papers

Jährliche aktualisierte Auflistung von Bio-Datenbanken durch Oxford University Press
(<http://www3.oup.co.uk/nar/database/>)

Database Category: Major Sequence Repositories

ACTIVITY
Functional DNA/RNA site activity
[database](#) [summary](#)

Ares Lab Yeast Intron Database Telerski, A.¹, Centers, R. J.², Ares, M.²
Spleosomal intron in *Saccharomyces cerevisiae*
[database](#) [summary](#)

ArrayExpress Eramo, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Garcia Lara, G., Ozdamin, A., Sansone, S., Rocca-Serra, P.
Public collection of microarray gene expression data
[database](#) [summary](#)

DNA Data Bank of Japan (DDBJ) Imanishi, T.¹, Miyazaki, S.², Fukami-Kobayashi, K.², Sugawara, H.², Gojobori, T.², Tateno, Y.²
All known nucleotide and protein sequences, International Nucleotide Sequence Database Collaboration
[database](#) [summary](#)

EMBL Nucleotide Sequence Database
All known nucleotide and protein sequences, International Nucleotide Sequence Database Collaboration
[database](#) [summary](#)

GenBank
All known nucleotide and protein sequences, International Nucleotide Sequence Database Collaboration
[database](#) [summary](#)

STACK
Non-redundant, gene-oriented clusters
[database](#) [summary](#)

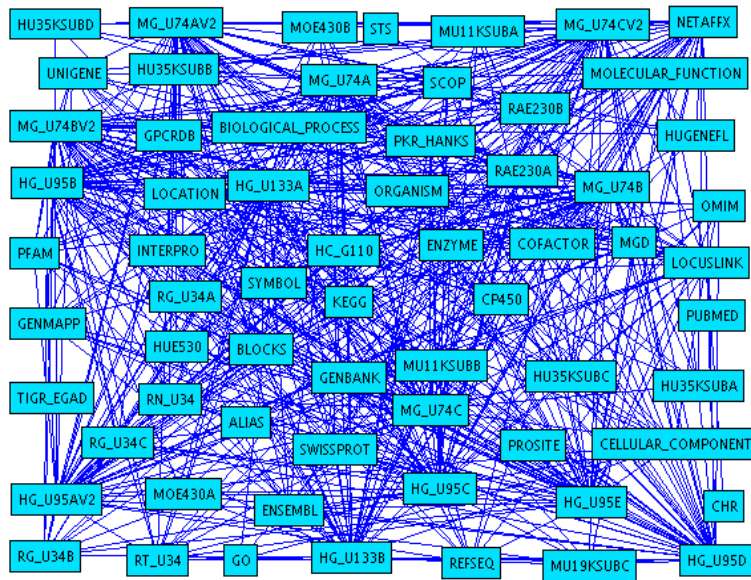
TIGR Gene Indices Lee, Y., Antonescu, V., Cheung, F., Karamycheva, S., Farvizi, B., Perte, G., Sultana, R., Sunkara, S., Tsai, J., White, J., Quackenbush, J.
Non-redundant, gene-oriented clusters
[database](#) [summary](#)

UniGene
Non-redundant, gene-oriented clusters
[database](#) [summary](#)

n A. D. Baxevanis: The Molecular Biology Database Collection: 2003 update. Nucl. Acids. Res. 2003 31 (NAR): 1-12.



Bio-Datenbanken: Vernetzungsproblematik



n Quelle: GenMapper (Do & Rahm; <http://sun1.izbi.uni-leipzig.de:8080/GenMapper/servlet/gui.MainFrame>)



Anforderungen

- n Verwaltung biologischer Daten
- n Flexibilität und Offenheit
- n Datenqualität
- n Integration und Datenaustausch
- n Querying und Analyse



Verwaltung biologischer Daten

n Unterschiedliche Datenarten

- unstrukturiert, z.B. TIFF eines Genexpressionschips
- strukturiert, z.B. Nucleotidsequenz, Proteinsequenz
- semistrukturiert, z.B. Annotationen

n Bio-Datenbanken ohne Experimentdaten im Bereich 1–200 GB

- GenBank: 110 GB (Uncompressed Flatfiles, Release No. 134, 2/2003)
- Swiss-Prot + TrEMBL^{*}: 1 GB (Oracle Export, Compressed, Stand 2/2003)

n Mit Experimentdaten deutlich größere Datenmengen

- TIFF eines Genexpressionschips: ca. 50 MB
- Rohspektrum eines MS[†]-Experimentes
- Tracefiles von Sequenziermaschinen
- Bilder von 2D-Gel-Elektrophorese-Experimenten[‡]

^{*} EMBL = European Molecular Biology Laboratory; TrEMBL = Proteinsequenz-Datenbank von EMBL (als Ergänzung zu Swiss-Prot)

[†] MS = Massenspektrometer / Massenspektrometrie

[‡] Gel-Elektrophorese: Verfahren, um Proteine in einer organischen Substanz (Gel) gemäß ihrer Ladung und ihrem Molekulargewicht zu trennen



Flexibilität und Offenheit

- n Forschungsfragen ändern sich ständig
 - Andere wissenschaftliche Fragestellungen → Andere Daten, andere Queries
- n Design muss Wartbarkeit und Flexibilität in Vordergrund stellen
 - Schemaänderungen, Einbringung neuer Datentypen, Optimierung auf neue Anforderungen
- n Bio-Datenbanken meist Teil eines Forschungsprojekts
 - Datenbeschaffung (LIMS^{*}), Datenarchivierung, Datenanalyse
- n Zugriff von verschiedensten Clients aus erforderlich (Java, CGI, Perl, PHP, ...)
- n Integration mit selbstentwickelten Analyse-Algorithmen nötig
 - Blast/Fasta (Algorithmen für Ähnlichkeitssuche/Alignments bzgl. Genen und Proteinen (<http://www.ncbi.nlm.nih.gov/BLAST/> bzw. <http://www.ebi.ac.uk/fasta33/>))
 - Strukturberechnung, Motifsuche, Gensuche
- n Integration von z.B. Blast in den DB2 Information Integrator

* Laboratory Information Management System



Datenqualität

- n Experimente erzeugen (fast) immer unscharfe Daten
 - Arbeit mit lebenden Organismen
 - Zugrundeliegende Mechanismen größtenteils unverstanden
 - Fehleranfällige Techniken: Bilderkennung, Statistische/heuristische Algorithmen, ...
- n Eine junge Wissenschaft: viele falsche Daten und Veröffentlichungen
- n Herkunft der Daten sollte (auch bei Ableitungen) ermittelbar sein



Integration

- n Viele Daten machen erst Sinn im Kontext
 - (Teil-)Sequenz: Genkontext, Regulationskontext, Homologie
 - Protein: Welcher Organismus?, Strukturkontext, Domänen
 - Expression: Regulationskontext, Phänotypen, Krankheitsverläufe, ...
- n Integration von Bio-Daten aus externen Quellen nach wie vor offenes Problem
- n Die meisten Datenbanken sind "nur" integriert im Sinne einer Verlinkung
 - z.B. Verlinkung Swiss-Prot ↔ OMIM* ↔ GDB†
- n Typische Bio-Anfragen implizieren bereits Zugriffe auf mehrere Datenbanken
 - Beispiel: Hypoxanthin-Guanin Phosphoribosyltransferase (HPRT)
 - Identifizieren Sie die HPRT mRNA und Proteinsequenz für Mensch, Maus und Ratte (DB: **GeneCards**, **NCBI-LocusLink**, **NCBI-GenBank** (für Nucleotide), **NCBI-GenBank** (für Proteine))
 - In welchem Stoffwechselweg ist HPRT eingebunden? (DB: **KEGG**)
 - Auf welchem Chromosom liegt HPRT beim Mensch? (DB: **NCBI-LocusLink**, **NCBI-OMIM**, **GeneCards**)
 - Gibt es eine Krankheit, die auf einer Mutation in HPRT beruht? (Datenbank: **NCBI-OMIM**)
- n Integration im Sinne eines globalen Schemas oft nicht vorhanden (und überhaupt nötig?)

* Online Mendelian Inheritance in Men
† Genom Database



Datenaustausch

n Verschiedene Austauschformate

- EMBL Format (Sequenzen)
- ASN.1 (Sequenzen)
- MIAME (Genexpressionsdaten)

n Export üblicherweise in Flat Files

n XML zunehmend von Bedeutung

n DTD's definiert für verschiedene Projekte, z.B.

- GAME^{*}
- BIOML[†]
- BSML[‡]

```
<db_entry id="5" label="Huang, M.E. (1995)" format="MEDLINE"
entry="95397595">
Yeast (1995) 11:775-781
</db_entry id="6" label="Embl: L36344" format="EMBL" entry="L36344"/>
</db_entry>
</reference>
<reference id="2" label="Databases">
<db_entry id="7" label="Embl: Z49540" format="EMBL" entry="Z49540"/>
</reference>
<peptide id="1" label="translated sequence" start="1" end="779">
MPTTYVPINQPIDGDEVIDTNRFTNIPETQNFDFVTIDKIAENRPLS
VDSDFRFLNSKYRHYREVINDRAKTFITLSSTAIVIGCIAGFLQVFTETL
VNWKTGHCQRNWLKNSFCCNGVVNEVTSTSNLLLRQEFCEAAQGLWIA
WKGHVSPFIIIFMLLSVLFALISTLLVKYVAPMATSGGISEIKVWVSGFEY
NKEPLGFLTLVIXSVALPLAISSGLSVGKEGPSVHYATCCGYLLTKWLLR
DLTYSSQYXEYITAASGAGVAVAFGAPIGVLPGLERIASANRPNSSTLW
KSYVALVAITTLKYIDPPFRNGRVILFNVTYDRDNKVVQEIPIFIALGIFG
GLYGYISKWNINFIHFRKMYLSSWPVQEVLFALTLALISYFNEPLKLD
MTESMGILFHECVKNDNTSTFSHRLCQLDENTHAFELKIFTSLCFATVI
RALLVVVSYGARVPAGIFVPSNAVGATPGRAVSLVERPISGSPSVITPGA
YAFLGAATLSGITNLTLTVVIMFELTGAFMYIIPLMIVVAITRIILST
SGISGGIADQIMVNGPFYLEDQDEEEETLEKYTARQLMSSKLITINE
TIYLSLESLLYDASEYSVHGFFITKDEDRKFEKRCIGYVLRRLHASK
IMQSVNSTKAQTTLVYFNKSNEELGHRNCIGFKDIMNESPISVKKAVP
VTLFRMPKELGCKTIIVBESGILKGLVAKDILRFKRIKYREVHGAFT
YNEALDRRCWSVIHFIIKRFTNRRNGNVI
<domain id="1" start="1" end="779" label="chloride channel protein CLC-1" />
<domain id="2" start="76" end="99" label="transmembrane" />
<domain id="3" start="156" end="197" label="transmembrane" />
<domain id="4" start="206" end="226" label="transmembrane" />
<domain id="5" start="262" end="288" label="transmembrane" />
<domain id="6" start="378" end="399" label="transmembrane" />
<domain id="7" start="435" end="567" label="transmembrane" />
<domain id="8" start="695" end="731" label="transmembrane" />
</peptide>
</subunit>
```

BIOML-Beispiel

* Genome Annotation Markup Elements
† BIOPolymer Markup Language
‡ Bioinformatic Sequence Markup Language



Querying und Analyse

n Bio-Daten werden im Allgemeinen für komplexe Weiterverarbeitungen genutzt

n Querying-Anforderungen

- Vordefinierte (parametrisierbare) Masken für häufige Anfragetypen
- Möglichkeit, Ad-hoc-Queries komfortabel zusammen stellen zu können (z.B. über grafisches Interface)
- Alphanumerisches Interface mit voller Query-Komplexität (für sog. "Power User")
- Unterstützung von Unschärfe bei unstrukturierten oder semi-strukturierten Daten

n Analyse-Anforderungen

- Integration von Data Warehouse-Ansätzen (multidimensionale Anfragen, Aggregation)
- Integration von Data Mining Tools



Klassifizierungsmerkmale

n Klassifizierung nach

- Inhalt
- Verfügbarkeit
- Datenhaltungssystem
- Externer Datengewinnung
- Datenqualität
- Art der Integration
- Zugriffsmethoden



Klassifizierung nach Inhalt

- n Organismus, Gewebe, Chromosome, ...
- n Typen der abgespeicherten Bio-Objekte: Sequenzen, Strukturen, Motifs^{*}, ...
- n Primärdatenbanken
 - Enthalten die unmittelbaren Experiment-Daten ("Nahe am Experiment")
 - Wenig Verarbeitung, kurze Annotationspipelines
 - Vertreter: Genbank/EMBL, PDB[†], UniGene
- n Sekundärdatenbanken
 - Aufbereitete Daten mit Annotationen (meist nur semi-strukturiert) und Verlinkungen
 - Vertreter: Swiss-Prot, MGD[‡], OMIM, ...
- n Tertiärdatenbanken
 - Ontologie-basiert, strukturierte Annotationen
 - Vertreter: GeneOntology, PFAM^{**}, PRINTs^{††}, InterPro^{‡‡}, CATH^{***},
- n Grenze vor allem zw. Sekundär- und Tertiärdatenbanken oft fließend

^{*} (kurze) Sequenz von Sekundär-Struktur-Elementen mit im Allg. spezifischer biologischer Funktion
[†] Protein Database
[‡] Mouse Genome Database
^{**} Protein families database of alignments and HMMs
^{††} Protein fingerprints database
^{‡‡} Integrated documentation resource for protein families, domains and sites
^{***} Class(C), Architecture(A), Topology(T) and Homologous superfamily (H) of protein domain structures



Klassifizierung nach Inhalt (2)

Biologischer "Bereich"	Datenbanktyp	Schwerpunkte	Unterstützte Fragestellungen	Vertreter
Genom	Kartierungs-DB	Genlokalisierung	Verwandschaftsbeziehungen, phylogentische Stammbäume	GDB
	Sequenz-DB	Basensequenzen Nucleinsäuresequenzen		Genbank / EMBL / DDBJ (DNA Data Bank of Japan)
	Mutations-DB	Genveränderungen	Welche Krankheiten sind durch welche Genveränderungen bedingt?	dbSNP (Single Nucleotide Polymorphism Database)
	Genexpressions-Datenbanken	Expressionsniveaus Genexpressionsmuster	Unter welchen Bedingungen exprimiert eine Zelle welche Gene?	GeneX, ArrayExpress



Biologischer "Bereich"	Datenbanktyp	Schwerpunkte	Unterstützte Fragestellungen	Vertreter
Proteine	Proteinsequenz-Datenbanken	Primärstruktur von Proteinen	Proteindesign (z.B. für neue Medikamente)	Swiss-Prot
	Proteinstruktur-Datenbanken	Sekundär-, Tertiär- und Quartärstruktur von Proteinen		PDB
	Protein-Domain/family	Protein-Gruppen	Welche Proteingruppe ist für bestimmte Stoffwechselprozesse (z.B. Blutgerinnung) zuständig	PFAM (Protein families database of alignments and HMMs)
Stoffwechsel	Pathway-Datenbanken	Metabolische Pfade Regulatorische Pfade	Welche Stoffwechselprozesse werden von welchen Proteinen (Enzymen) gesteuert. Welche (Abfall-)Produkte entstehen dabei	KEGG (Kyoto Encyclopedia of Genes and Genomes)
Publikationen				MedLine



Klassifizierung nach Verfügbarkeit

n Öffentliche Datenbanken

- Lange bestehend, international organisiert
- Referenzdatenbanken, öffentliches Archive (Genbank, Swiss-Prot, PIR, PDB, ...)

n Nicht-öffentliche Datenbanken

- Projektbezogene ("One-Shot")-Datenbanken von Forschungsgruppen (hochaktuell für kurze Zeit; existieren oft nur bis zur Veröffentlichung der Ergebnisse)
- Kommerzielle Datenbanken von Bio-Firmen (z.B. Celera)



Klassifizierung nach Datenhaltungssystem

n Verwendetes Speichersystem

- Flatfiles
- Proprietäre Systeme (ACeDB, Icarus/SRS*)
- Relationale DBMS
- Objektorientierte/Objektrationale DBMS
- XML Datenbanken (Tamino, XIS)

* Sequence Retrieval System



Klassifizierung nach Art der externen Datengewinnung

n "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, ...

n "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, PIR*, ...

n Mischformen: GDB

* Protein Information Resource



Klassif. bzgl. Datenqualität (Curation)

- n 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
- n Ansatz 2: Zentrale Nachbearbeitung/Kontrolle der Daten (z.B. Swiss-Prot, MIPS*)
 - Daten werden laufend verbessert
 - Hoher (manueller) Aufwand, da Automatisierung nur eingeschränkt möglich
 - Vorteil: Höhere Datenqualität; Nachteil: Urheber weniger klar, hohe Volatilität
- n 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)

* Munich Information Center for Protein Sequences



Klassifizierung nach Art der Integration

n Ansatz 1: Virtuelle Integration (über Links)

- "lockerer Verbund" zwischen Datenquellen, deren Objekte durch Verweise miteinander verbunden sind
- Häufigste Integrationsart in Bio-Datenbanken

n Ansatz 2: Materialisierte Integration

- Daten werden kopiert und zentral aufbereitet
- Data Warehouse-Ansatz

n Beide Ansätze mit oder ohne globales Schema

n Manuelle versus automatische Integration

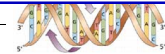
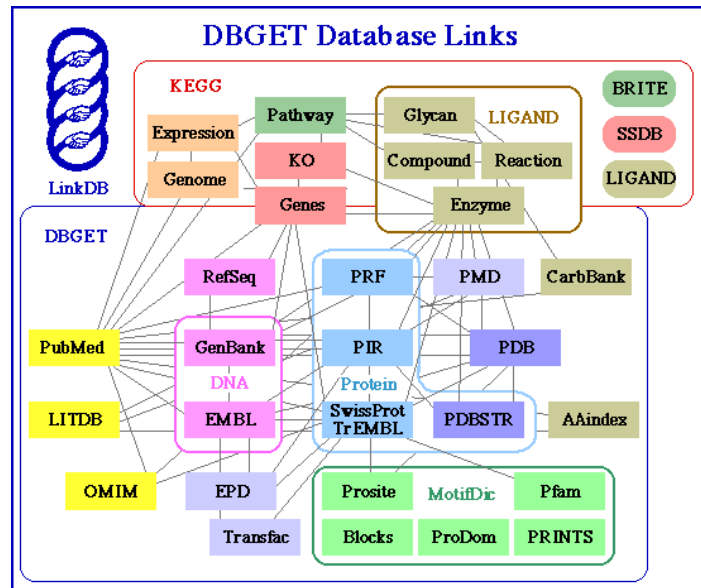
- Automatische Integration anhand def. Kriterien (Ensembl)
- Manuelle Integration anhand Wissen des Editors (Swiss-Prot)



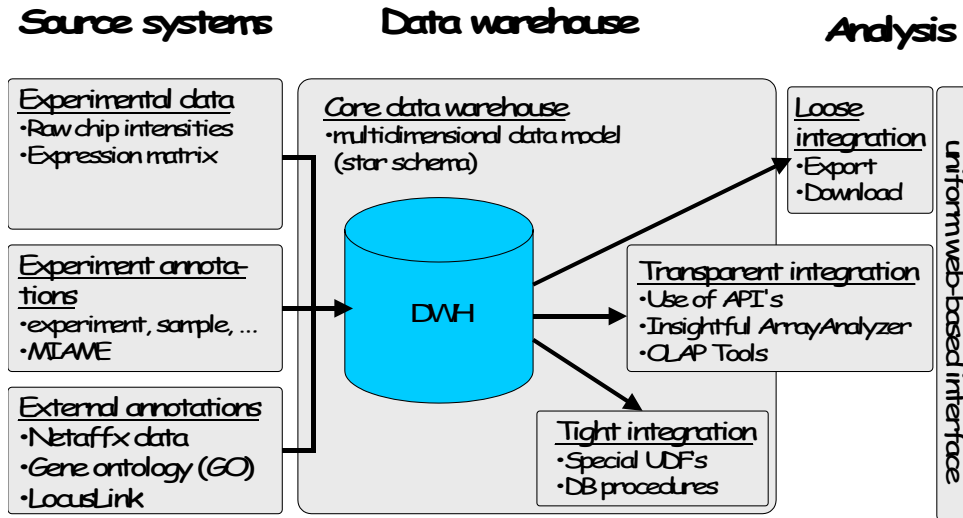
Beispiel virtuelle Integration: Linkdatenbanken DBGet / DBLink

n DBGET: Retrieval System für breite Palette von Bio-Datenbanken

n Quelle: www.genome.ad.jp/dbget (Stand Okt. 2003)



Beispiel materielle Integration: GeWare



- Quelle: Do, H.H., Kirsten, T., Rahm, E.: Comparative Evaluation of Microarray-based Gene Expression Databases. Proc. 10. Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW 2003), Leipzig, Feb. 2003 (siehe auch: <http://dbs.uni-leipzig.de/de/projekte/BIOINF/bioinformatics.html>)



Klassifikation nach Zugriffsmethoden

- n Navigation (über Links)
- n Stichwortsuche
- n Anfragesprachen
- n Data Mining



Navigation

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

Gene Ontology (<http://www.geneontology.org/>)

The screenshot displays the Gene Ontology web interface. It features three main columns: **molecular_function**, **cellular_component**, and **biological_process**. The **cellular_component** column is expanded to show a hierarchical tree structure. The tree starts with 'cellular_component' and branches into '1.1. extracellular(133)', '1.2. intracellular(4027)', '1.3. unlocalised(70)', and '1.4. cellular_component unknown'. The '1.2. intracellular(4027)' branch is further expanded to show sub-terms like '1.2.1. cytoplasm(100)', '1.2.2. nucleus(100)', and '1.2.3. organelle(100)'. Below the columns are three 'Current tree depth 1' indicators and a search bar with the text 'Locate term:'. To the right of the search bar are the instructions 'Use the "*" as a wild card?' and 'Locating terms is case insensitive'. At the bottom of the interface are buttons for 'List gene products for selected terms', 'Only current', 'or', 'and', and 'and not'.

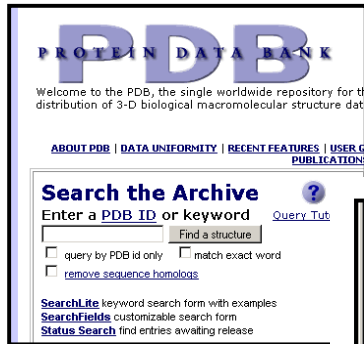


Stichwortsuche

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



Stichwortsuche: Beispiele



PDB
PROTEIN DATA BANK

Welcome to the PDB, the single worldwide repository for the distribution of 3-D biological macromolecular structure data.

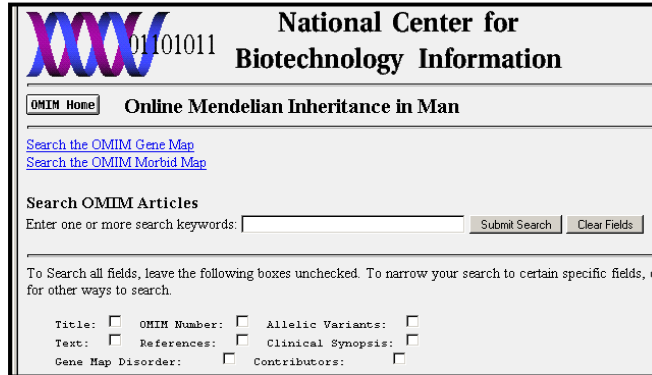
[ABOUT PDB](#) | [DATA UNIFORMITY](#) | [RECENT FEATURES](#) | [USER GUIDANCE](#) | [PUBLICATION](#)

Search the Archive

Enter a **PDB ID** or **keyword** [Query Tutorial](#)

query by PDB id only match exact word
 [remove sequence homologs](#)

[Search by Life](#) keyword search form with examples
[Search by Fields](#) customizable search form
[Status Search](#) find entries awaiting release



OMIM 01101011

National Center for Biotechnology Information

[OMIM Home](#) **Online Mendelian Inheritance in Man**

[Search the OMIM Gene Map](#)
[Search the OMIM Morbid Map](#)

Search OMIM Articles

Enter one or more search keywords:

To Search all fields, leave the following boxes unchecked. To narrow your search to certain specific fields, check for other ways to search.

Title: OMIM Number: Allelic Variants:
Text: References: Clinical Synopsis:
Gene Map Disorder: Contributors:



Anfragesprachen

- n Unterstützen (semi-)strukturierte Anfragen
- n Basieren üblicherweise auf Mengenoperationen
- n Vertreter
 - SQL92 (relational), SQL:1999 (objekt-relational)
 - OQL (objektorientiert)
 - XPath / XQuery (XML-basiert)
- n Typische Elemente (SQL92)
 - Select: Auswahl dessen, was Ergebnis ausmacht
 - From: Auswahl der Datenherkunft
 - Where: Auswahl der Bedingungen, die Ergebnisse erfüllen müssen
- n Spezialsprachen (z.B. GQL: Genom Query Language) → Kapitel 7 (Zugriffsmethoden in Bio-Datenbanken)
- n Kaum Verwendung als "öffentliches" Interface, da zu komplex



Suchformulare

n 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

n Benutzerunterstützung durch Drop-Downlisten, Checkboxes etc.

n Übersetzung in strukturierte Anfragesprache relativ trivial

n Vorteil: Einfach zu realisieren, benutzerfreundlich, effizient

n Nachteil: Starke Einschränkung der Expressivität, keine Unterstützung vom komplexen Anfragen

The image shows a screenshot of the GDB search form. It is a structured query interface with several sections, each containing input fields and a search button (indicated by a question mark icon):

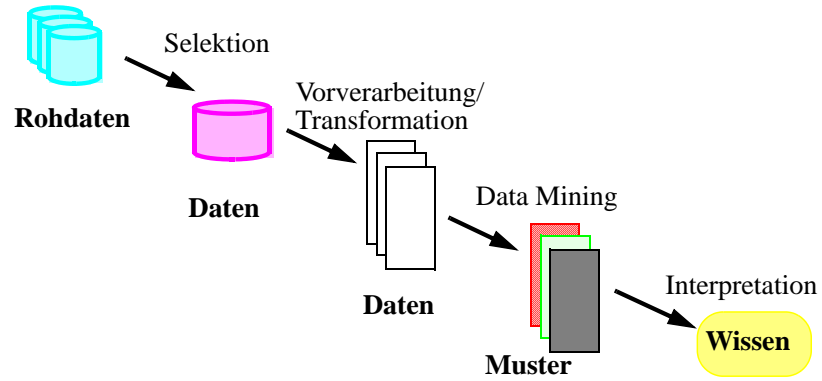
- Symbol/Names:** A single input field for 'Name'.
- Cytogenetic Localization:** Three input fields for 'Chromosome', 'Left Marker', and 'Right Marker'.
- All Localizations:** Three input fields for 'Chromosome', 'Left Marker', and 'Right Marker'.
- Nucleic Acid Sequence Links:** A single input field.
- Related Segments:** A single input field for 'Marker'.
- Polymorphisms:** Three input fields for 'Polymorphism', 'Variation Type', and 'Max Het'.
- Mutations:** A single input field.
- Phenotype Links:** A single input field.
- Families:** A single input field.

The text 'GDB-Suchformular' is printed to the right of the form. At the bottom left of the form area, the word 'Done' is visible.



Data Mining

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



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



Data Mining

n Data Mining: Anwendung eff. Algorithmen, die in DB enthaltene Muster liefern

n bisher meist Mining auf speziell aufgebauten Dateien

n notwendig: Data Mining auf Datenbanken bzw. Data Warehouses

- Skalierbarkeit auf große Datenmengen
- Nutzung von Performance-Techniken (Indexe, materialisierte Sichten, Parallelverarbeitung)
- Vermeidung von Redundanz und Inkonsistenzen
- Integration mehrerer Datenquellen, Portabilität

n 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



Data Mining: Techniken

n Clusteranalyse

- Objekte (z.B. Proteine) werden aufgrund von Ähnlichkeiten in Klassen eingeteilt (Segmentierung)

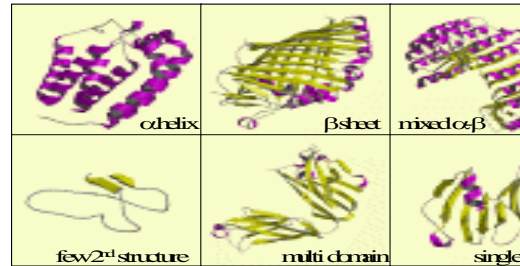


n Assoziationsregeln

- z.B. Bei Genexpression vom Grad x bei Gen $y \Rightarrow$ Hinweis auf Erkrankung z)
- Sonderformen zur Berücksichtigung von Dimensionshierarchien (z.B. Gentgruppen), quantitativen Attributen, zeitlichen Beziehungen (sequence mining)

n Klassifikation

- Zuordnung von Objekten (z.B. Proteinen) zu Gruppen/Klassen mit gemeinsamen Eigenschaften bzw. Vorhersage von Attributwerten
- explizite Erstellung von Klassifikationsregeln (z.B. "wenn Teilsequenz T dann Proteingruppe P")
- Verwendung von Stichproben (Trainingsdaten)
- Ansätze: Entscheidungsbaum-Verfahren, statistische Auswertungen (z.B. Maximum Likelihood-Schätzung / Bayes-Schätzer), neuronale Netze



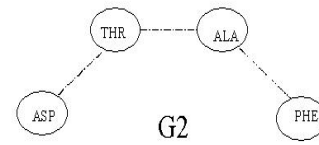
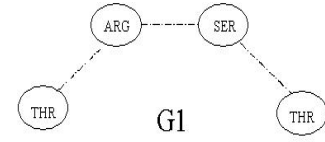
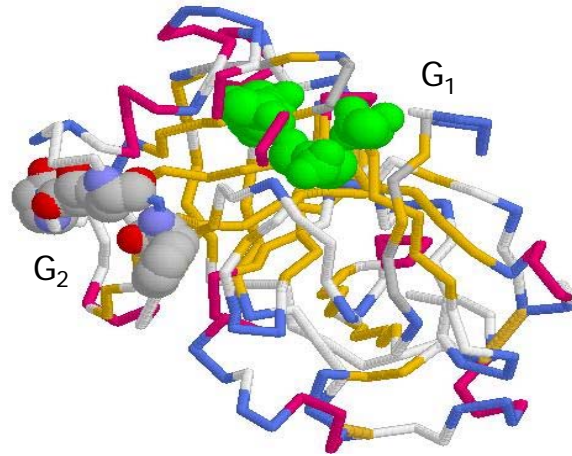
n Weitere Ansätze:

- Genetische Algorithmen (multivariate Optimierungsprobleme, z.B. beim Proteindesign)
- Regressionsanalyse zur Vorhersage numerischer Attribute . . .



Beispiel: Fingerprint-Suche von Proteinen

- n Fingerprint: Substruktur, die Protein oder Proteingruppe eindeutig identifiziert
- n Fingerprint in Prokaryotic Serine Protease (Achromobacter lyticus protease I; PDB ID 1ARB)



Zusammenfassung

n Bio-Datenbanken

- forschungsgetrieben, forschungsbegleitend, Forschungsgegenstand: Ständige Veränderung

n Anforderungen

- Verwaltung biologischer Daten
- Flexibilität und Offenheit
- Datenqualität
- Integration und Datenaustausch
- Querying und Analyse

n Klassifizierungsmerkmale

- Inhalt, Verfügbarkeit
- Datenhaltungssystem, Externe Datengewinnung, Datenqualität
- Art der Integration, Zugriffsmethoden
- Fast jede Bio-Datenbank spezifische Kombinationen der o.g. Merkmalsausprägungen bzgl. Inhalt, etc.; eindeutige Einordnungen i.d.R. nicht möglich

