Kapitel 2: Bio-Datenbanken Überblick

- Inhalt
  - Motivation
  - Historische Entwicklung
  - Anforderungen
  - Klassifizierungsmerkmale
  - Zusammenfassung

Motivation

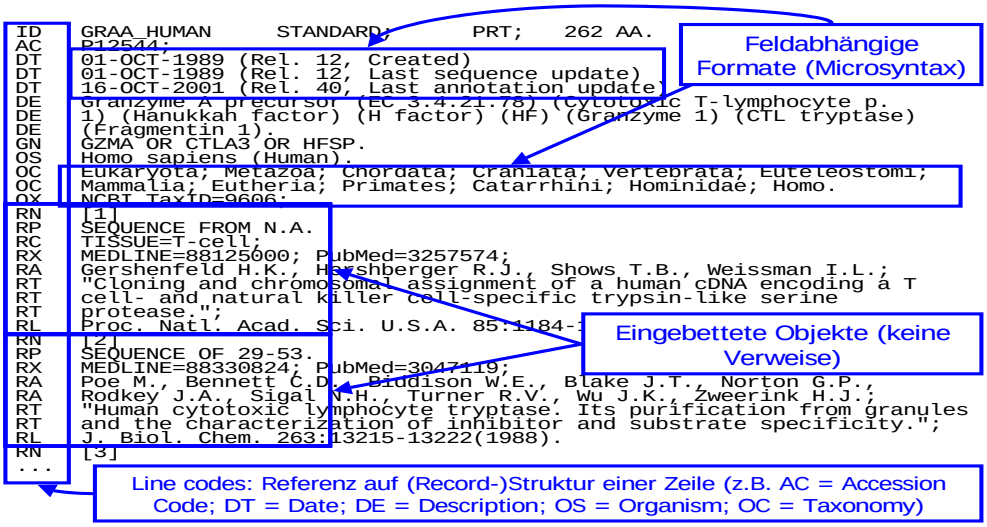
- Abspeicherung von Genom-, Protein- und Stoffwechselinformationen in konsistenter und effizienter Art und Weise
- Unterstützung von biowissenschaftlichen Anfragen und Analysen
  - Beispiel: Hypoxanthin-Guanin Phosphoribosyltransferase (HPRT)
  - Identifizieren Sie die HPRT mRNA und Proteinsequenz für 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?
- 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

- Alle (großen) öffentlichen Bio-Datenbanken entstanden aus Büchern
- Sammlungen bekannter Daten einer Art: DNA, Proteinsequenz, Proteinstruktur
  - Jährliches / quartalsweises Erscheinen
  - Buch → Band → CD → FTP → WWW
- Anfangs Verwendung von flachen, textorientierten Datenmodellen
  - Viele Beschreibungen in freier Textform
  - Für Menschen konzipiert, nicht für Weiterverarbeitung durch Computer
  - Datenbank = Menge ähnlich strukturierter "Entries"
- 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

- Beispiel Swiss-Prot*; Hanukkah-Faktor (Zytotoxische T-Lymphozyten Proteinase)



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

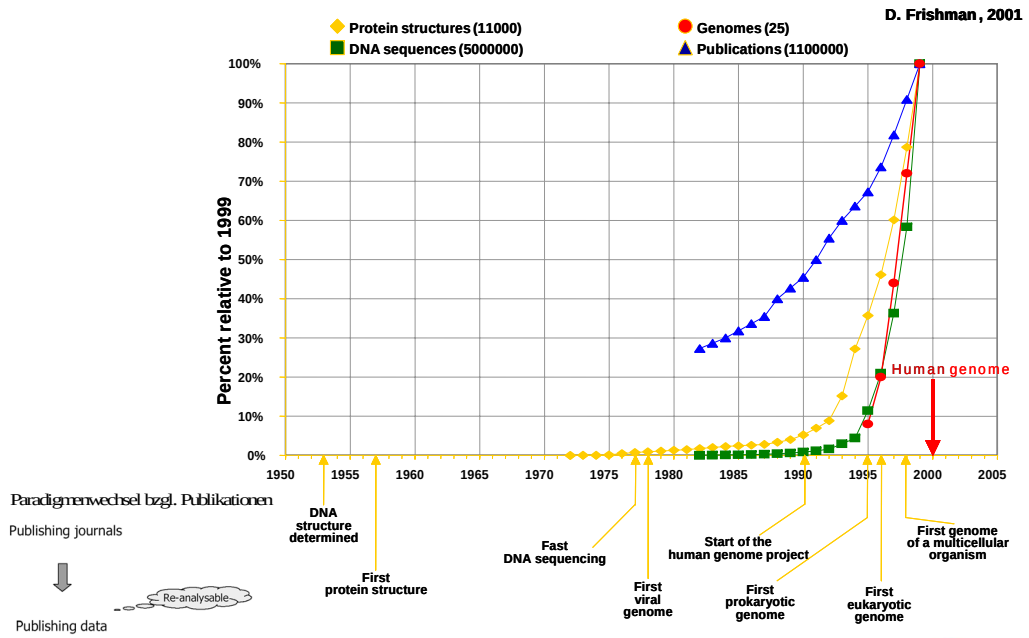
* Swiss-Prot = Protein knowledgebase

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

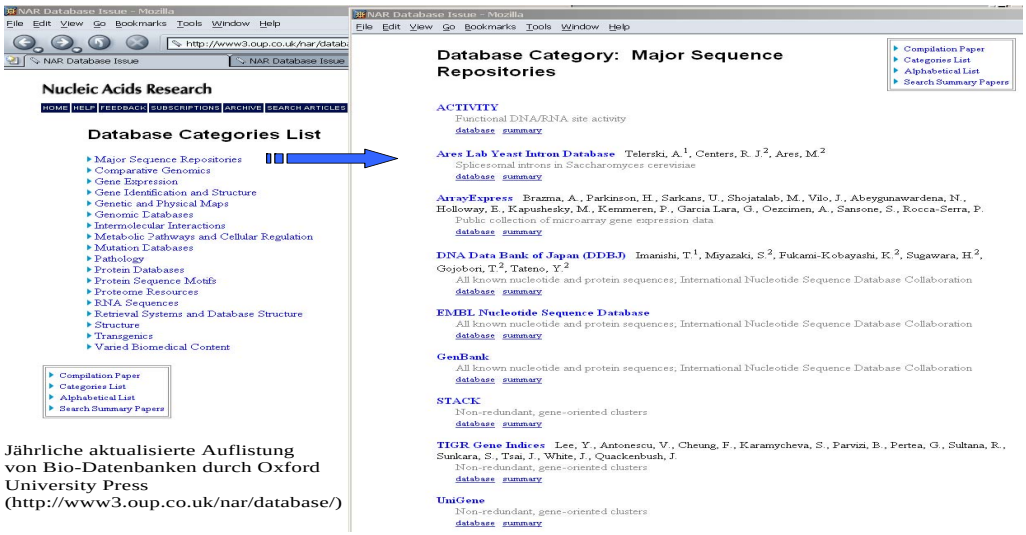
Bio-Daten: Historische Entwicklung



Paradigmenwechsel bzgl. Publikationen

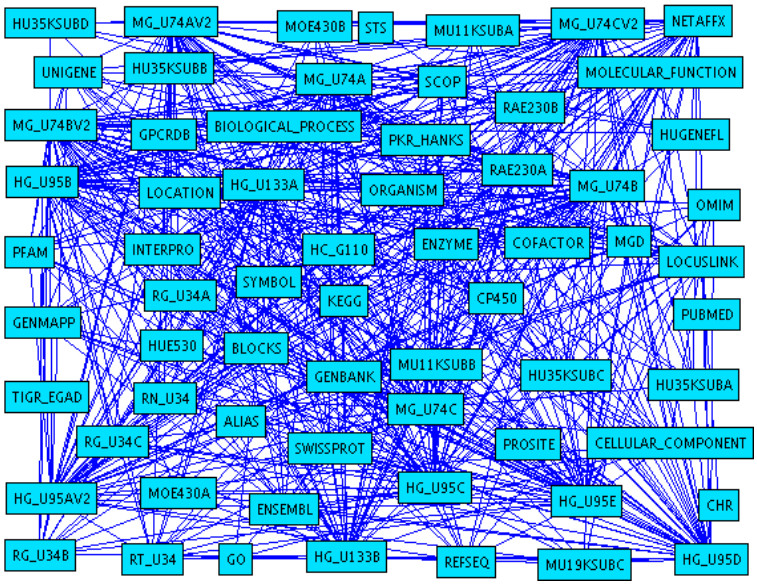
Bio-Datenbanken: Übersicht

- Weltweit derzeit über 500 Bio-Datenbanken



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

Bio-Datenbanken: Vernetzungsproblematik



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

Anforderungen

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

Verwaltung biologischer Daten

- Unterschiedliche Datenarten
  - unstruktruiert, z.B. TIFF eines Genexpressionschips
  - strukturiert, z.B. Nucleotidsequenz, Proteinsequenz
  - semistrukturiert, z.B. Annotationen
- 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)
- 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

- Forschungsfragen ändern sich ständig
  - 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
- Zugriff von verschiedensten Clients aus erforderlich (Java, CGI, Perl, PHP, ...)
- 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
- Integration von z.B. Blast in den DB2 Information Integrator

* Laboratory Information Mangagement System

Datenqualität

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

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 nach wie vor offenes Problem
- Die meisten Datenbanken sind "nur" integriert im Sinne einer Verlinkung
  - z.B. Verlinkung Swiss-Prot $\leftrightarrow$ OMIM* $\leftrightarrow$ GDB†
- 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)
- Integration im Sinne eines globalen Schemas oft nicht vorhanden (und überhaupt nötig?)

* Online Mendelian Inheritance in Men
† Genom Datebase

Datenaustausch

- Verschiedene Austauschformate
  - EMBL Format (Sequenzen)
  - ASN.1 (Sequenzen)
  - MIAME (Genexpressionsdaten)
- Export üblicherweise in Flat Files
- XML zunehmend von Bedeutung
- 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">
MPTTYVPINQPIGDGEDVIDTNRFTNIPETQNFDQFVTIDKIAEENRPLS
VDSDREFLNSKYRHYREVIWDRAKTFITLSSTAIVIGCIAGFLQVFTETL
VNWKTGHCQRNWLLNKSFCCNGVVNEVTSTSNLLLKRQEFECEAQGLWIA
WKGHVSPFIIFMLLSVLFALISTLLVKYVAPMATGSGISEIKVWVSGFEY
NKEFLGFLTLVIKSVALPLAISSGLSVGKEGPSVHYATCCGYLLTKWLLR
DTLTYSSQYEYITAASGAGVAVAFGAPIGGVLFGLEEIASANRFNSSTLW
KSYYVALVAITTLKYIDPFRNGRVILFNVTYDRDWKVQEIPIFIALGIFG
GLYGKYISKWNINFIHFRKMYLSSWPVQEVLFLATLTALISYFNEFLKLD
MTESMGILFHECVKNDNTSTFSHRLCQLDENTHAFEFLKIFTSLCFATVI
RALLVVVSYGARVPAGIFVPSMAVGATFGRAVSLLVERFISGPSVITPGA
YAFLGAAATLSGITNLTLTVVVIMFELTGAFMYIIPLMIVVAITRIILST
SGISGGIADQMIMVNGFPYLEDEQDEEEEETLEKYTAEQLMSSKLITINE
TIYLSELESLLYDSASEYSVHGFPITKDEDKFEKEKRCIGYVLKRHLASK
IMMQSVNSTKAQTTLVYFNKSNEELGHRENCIGFKDIMNESPISVKKAVP
VTLLFRMFKELGCKTIIVEESGILKGLVTAKDILRFKRIKYREVHGAKFT
YNEALDRRCWSVIHFIIKRFTTNRNGNVI
<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

- Bio-Daten werden im Allgemeinen für komplexe Weiterverarbeitungen genutzt
- 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 unstruktrukturierten oder semi-strukturierten Daten
- Analyse-Anforderungen
  - Integration von Data Warehouse-Ansätzen (multidimensionale Anfragen, Aggregation)
  - Integration von Data Mining Tools

Klassifizierungsmerkmale

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

Klassifizierung nach Inhalt

- Organismus, Gewebe, Chromosome, ...
- Typen der abgespeicherten Bio-Objekte: Sequenzen, Strukturen, Motifs*, ...
- Primärdatenbanken
  - Enthalten die unmittelbaren Experiment-Daten ("Nahe am Experiment")
  - Wenig Verarbeitung, kurze Annotationspipelines
  - Vertreter: Genbank/EMBL, PDB†, UniGene
- Sekundärdatenbanken
  - Aufbereite Daten mit Annotationen (meist nur semi-strukturiert) und Verlinkungen
  - Vertreter: Swiss-Prot, MGD‡, OMIM, ...
- Tertiärdatenbanken
  - Ontologie-basiert, strukturierte Annotationen
  - Verteter: GeneOntology, PFAM**, PRINTs††, InterPro‡‡, CATH***, ...
- 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	Genlokalisation	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 Stoffwechstelprozesse (z.B. Blutgerinnnung) zuständig	PFAM (Protein families database of alignments and HMMs)
Stoffwechsel	Pathway-Datenbanken	Metabolische Pfade Regulatorische Pfade	Welche Stoffwechstelprozesse werden von welchen Proteinen (Enzymen) gesteuert. Welche (Abfall-)Produkte entstehen dabei	KEGG (Kyoto Encyclopedia of Genes and Genomes)
Publikationen				MedLine

Klassifizierung nach Verfügbarkeit

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

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

* Sequence Retrieval System

Klassifizierung nach Art der externen 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, PIR*, ...
- Mischformen: GDB

* Protein Information Resource

Klassif. bzgl. 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*)
  - 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)

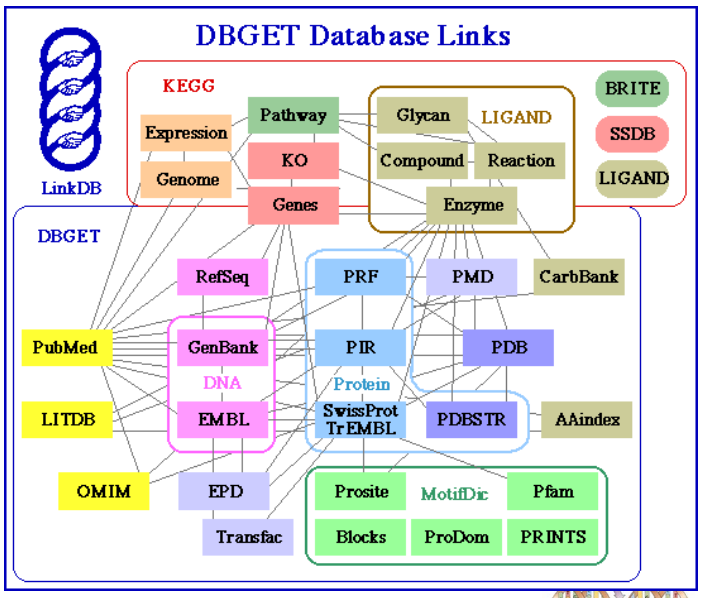
* Munich Information Center for Protein Sequences

Klassifizierung nach Art der Integration

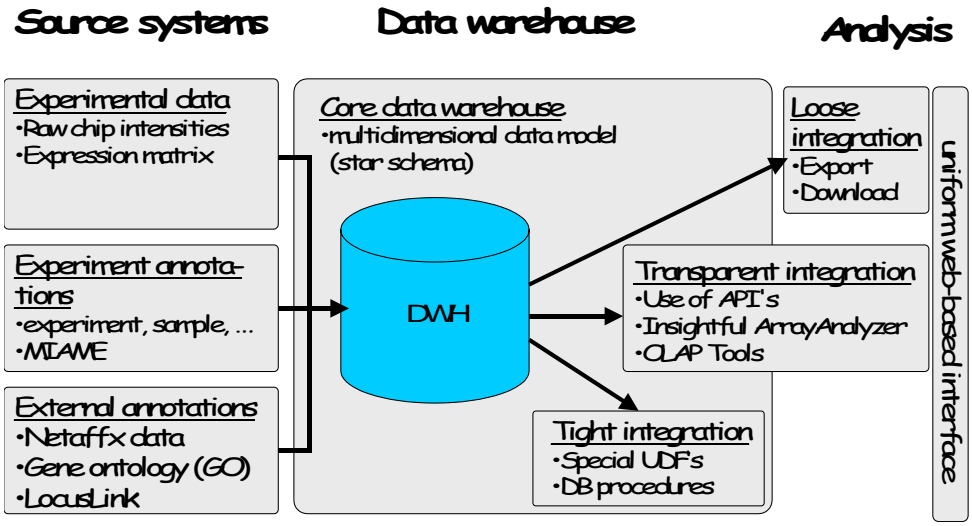
- Ansatz 1: Virtuelle Integration (über Links)
  - "lockerer Verbund" zwischen Datenquellen, deren Objekte durch Verweise miteinander verbunden sind
  - Häufigste Integrationsart in Bio-Datenbanken
- Ansatz 2: Materialisierte Integration
  - Daten werden kopiert und zentral aufbereitet
  - Data Warehouse-Ansatz
- 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)

Beispiel virtuelle Integration: Linkdaten-banken DBGet / DBLink

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

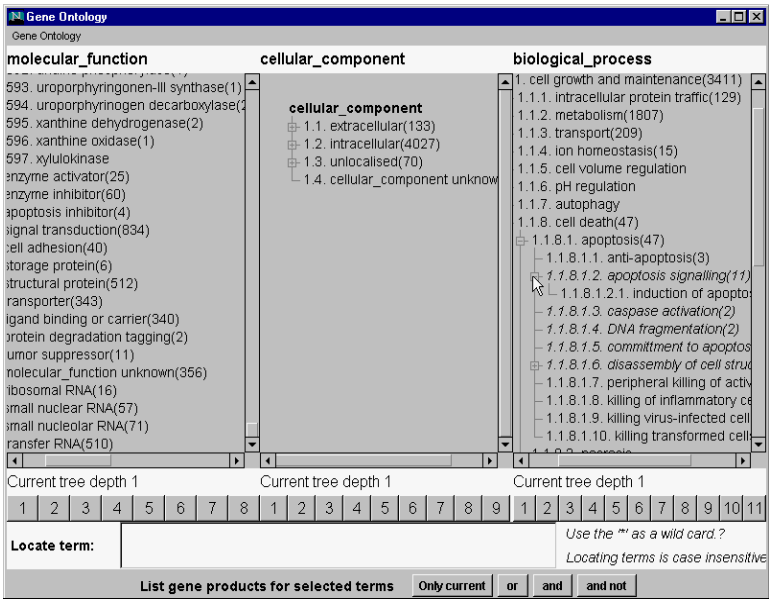
Klassifikation nach Zugriffsmethoden

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

Navigation

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

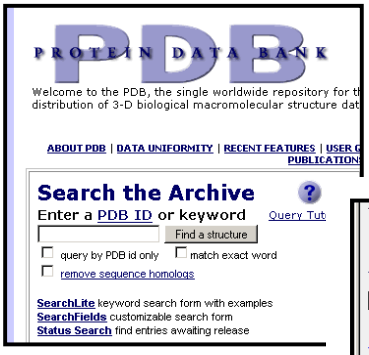
Gen Ontology (http://www.geneontology.org/)

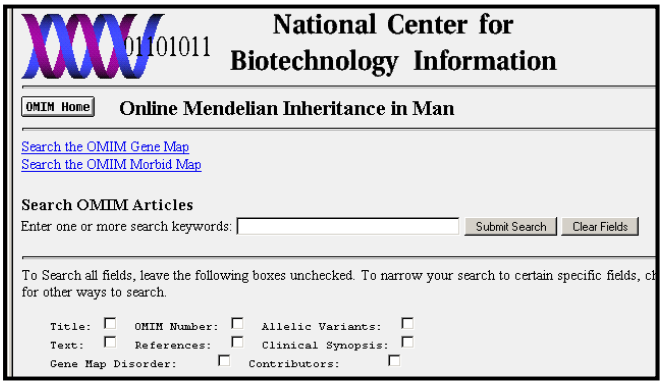


Stichwortsuche

- Typische Zugriffsmöglichkeit im Web (Google, Altavista 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 geeignet zur automatischen Weiterverarbeitung (Joins, ...)

Stichwortsuche: Beispiele



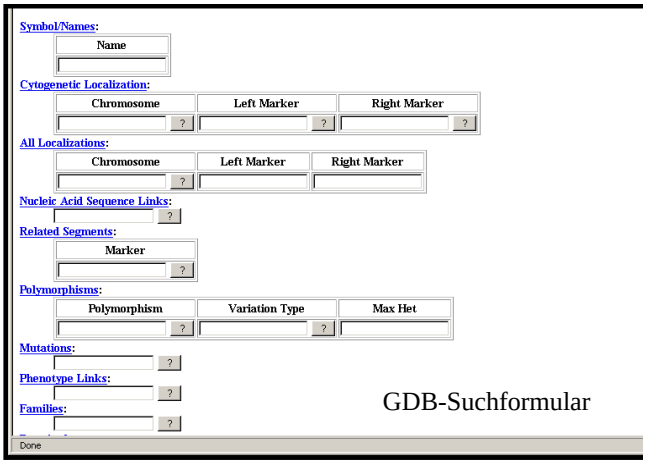


Anfragesprachen

- 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
- Spezialsprachen (z.B. GQL: Genom Query Language) → Kapitel 7 (Zugriffsmethoden in Bio-Datenbanken)
- Kaum Verwendung als "öffentliches" Interface, da zu komplex

Suchformulare

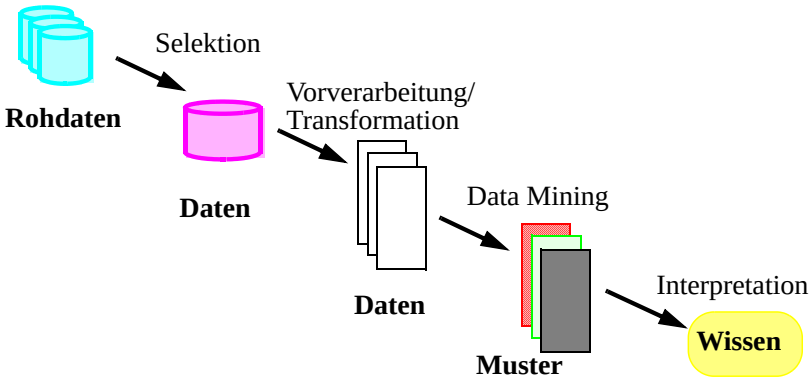
- Anfragen erfolgen oft in vorstrukturierten Suchformularen ("Canned Queries")
  - Select / From: Meist Fest innerhalb des Suchformulars
  - Where: Wahlmöglichkeiten bzgl. der Werte, der Vergleichsoperatoren und der Verknüpfungen zwischen einzelnen Bedingungen
- Benutzerunterstützung durch Drop-Downlisten, Checkboxes etc.
- Übersetzung in strukturierte Anfragesprache relativ trivial
- Vorteil: Einfach zu realisieren, benutzerfreundlich, effizient
- Nachteil: Starke Einschränkung der Expressivität, keine Unterstützung vom komplexen Anfragen



GDB-Suchformular

Data Mining

- 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



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

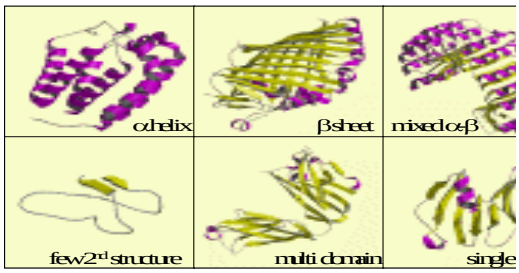
Data Mining

- Data Mining: Anwendung eff. Algorithmen, die in DB enthaltene Muster liefern
- bisher meist Mining auf speziell aufgebauten Dateien
- 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
- 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

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

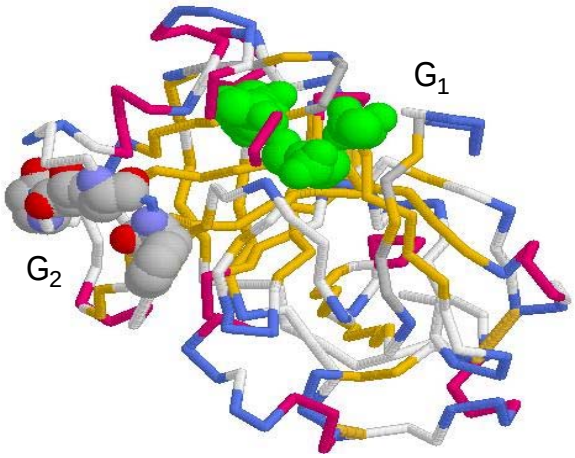
- Assoziationsregeln
  - z.B. Bei Genexpression vom Grad x bei Gen y => Hinweis auf Erkrankung z)
  - Sonderformen zur Berücksichtigung von Dimensionshierarchien (z.B. Gentgruppen), quantitativen Attributen, zeitlichen Beziehungen (sequence mining)
- Klassifikation
  - Zuordnung von Objekten (z.B. Proteinen) 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

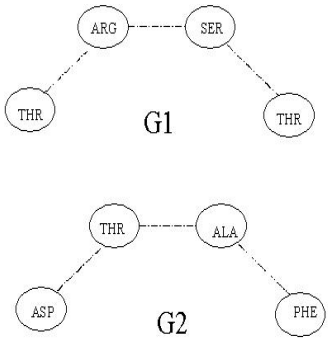


- Weitere Ansätze:
  - Genetische Algorithmen (multivariate Optimierungsprobleme, z.B. beim Proteindesign)
  - Regressionsanalyse zur Vorhersage numerischer Attribute . . .

Beispiel: Fingerprint-Suche von Proteinen

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





Zusammenfassung

- Bio-Datenbanken
  - forschungsgetrieben, forschungsbegleitend, Forschungsgegenstand: Ständige Veränderung
- Anforderungen
  - Verwaltung biologischer Daten
  - Flexibilität und Offenheit
  - Datenqualität
  - Integration und Datenaustausch
  - Querying und Analyse
- 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