

Datenbanken in der Bioinformatik

Alexander Hinneburg

hinneburg@informatik.uni-halle.de

www.informatik.uni-halle.de/~hinneburg

Vorlesungsplan

1. Übersicht
2. Datenmodelle
 - Textdateien, Entry-Modell
 - Relationale DB
 - XML
3. Genom-DB
4. Genexpressions-DB
5. Protein-DB
6. Pathway-DB
7. Publikations-DB
8. Zugriff und Anfragesysteme
9. Datenintegration von Bio-DB
10. Webservices

Veranstaltungsmodalitäten

- Vorlesung, 2 SWS
 - Basiert auf Vorlesungen
 - Prof. Rahm, Prof. Müller, Uni Leipzig
 - Prof. Leser, HU Berlin
 - Mischung aus Folien und Tafel
 - Links und Forschungsartikel auf der Vorlesungsseite
- Übung, 1 SWS
 - Praktische Aufgaben, Textdateien, XML, Oracle
 - Präsentation der Ergebnisse, 3-5 Folien
 - Erfahrungsaustausch
- Schein und Prüfung
 - Klausur Ende des Semesters
 - Durchführung eines Abschlußprojektes: +1 SWS
 - Prüfung umfaßt Vorlesung und Übung
- Blockseminar
 - 2-3 Tage, Anfang Oktober 2006 (vor Vorlesungsbeginn)

Biodatenbanken Überblick

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

Motivation

- n Abspeicherung von Genom-, Protein- und Stoffwechsellinformationen in konsistenten 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	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	Grazyme A precursor (EC 3.4.21.78) (Cytotoxic T-lymphocyte p. 1) (Hanukkah factor) (H factor) (HF) (Grazyme 1) (CTL tryptase) (Fragmentin 1) OR HFSP.			
GN	GZMA OR CTLA3 OR HFSP.			
OS	Homo sapiens (Human).			
OC	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.			
RY	NCBI TaxID: 9606.			
RN	[1]			
RP	SEQUENCE FROM N.A. TISSUE-T-cell;			
RC	MEDLINE=88125000; PubMed=3257574;			
RX	Gershenfeld H.K., Hershberger R.J., Shows T.B., Weissman I.L.;			
RT	"Cloning and chromosomal assignment of a human cDNA encoding a T cell- and natural killer cell-specific trypsin-like serine protease.";			
RL	Proc. Natl. Acad. Sci. U.S.A. 85:1184-			
RN	[2]			
RP	SEQUENCE OF 29-53.			
RX	MEDLINE=88330824; PubMed=3019119;			
RA	Poe M., Bennett C.L., Fridlson W.E., Blake J.T., Norton G.P.,			
RD	Rodkey J.A., Sigal I.H., Turner R.V., Wu J.K., Zweerink H.J.;			
RT	"Human cytotoxic lymphocyte tryptase. Its purification from granules and the characterization of inhibitor and substrate specificity.";			
RL	J. Biol. Chem. 263:13215-13222(1988).			
RN	[3]			
...				

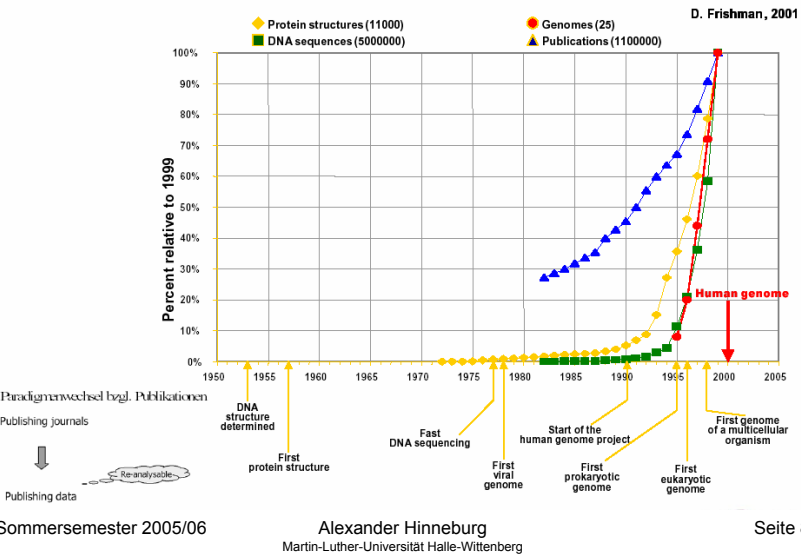
- n Zum Entry-Modell mehr in Kapitel 3 (Datenmodelle von Bio-Datenbanken)

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



Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 8

Biodatenbanken: Übersicht

n Weltweit derzeit über 500 Bio-Datenbanken



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

Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 9

Bio-Datenbanken: Vernetzungsproblematik



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

Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 10

Anforderungen

- Verwaltung biologischer Daten
- Flexibilität und Offenheit
- Datenqualität
- Integration und Datenaustausch
- DB-Anfragen und Analyse

Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 11

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 Man
† 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‡

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

Sommersemester 2005/06

```
<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">
MPTTTPVINGPIGDGDVIDTNRPTNPSTQNPDPQVFTIDIKIARENRPLS
VDSDREFNSVREHREVIDNARTPTIISSTALVIGCIASPLQVPTTL
VNWTSQCENWLLNKSFCNCGVNEVSTSNLLERQPPSCAQLWIA
WGHVSPFIIFMLLSVLPAIISTLLVKKVAPMATGSGISIXVVSQFVY
NEPFGPLTLVIGSVALPAISSGLSVGKRGPSVHYATCCGVLLTKLLR
DTLTYSQTEVITAASGAQVAFAPGAPIGOVLPGLREIASANRPNSTLW
KSYVALVATITLKYIDPFRNGRVILPNVTYDHDWVQRIPIALGIFG
GLYKYISEWNIIFIRFMYLSSMPVQRVLPALTLTALISYPNEFLKLD
NTESMGILFHCVENKDTSTPSHRLCQDSENTHAEPFLKIPFSLCFATVI
KALLVTVSTGARYFAGIPVPSHAYGATPGRVSLVVERFISQPSVITPGA
YAPUGAATUSGITNGLTVVIMFELTGAWIIPLMITVALTITLST
SGISGGLADQIMVNOPTLDEQDEERESTLKYTAGQMGSKLITIE
TITLSELESLLYDSASETSVHGPFITKDBDKFKEKRCIGVYKRLASK
IMQSVNSTEAQTTLVYFWSNNEELGHRNCIGPKDINNSPISVKAVP
VTLFRNPFEELGCKTIIVERSGILKGLVTAKDILKPKRIKRYRVRGAKFT
YNEALDRRCWSVIRPIIKRPTTNRNGNVI
<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

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 16

Anfragen 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

Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 17

Klassifizierungsmerkmale

• Klassifizierung von Bio-DB nach

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

Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 18

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

Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 19

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

Klassifizierung nach Inhalt (3)

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/Objektrelationale DBMS
- XML Datenbanken (Tamino, XIS)

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

Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 24

Klassifizierung nach 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
Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 25

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)

Sommersemester 2005/06

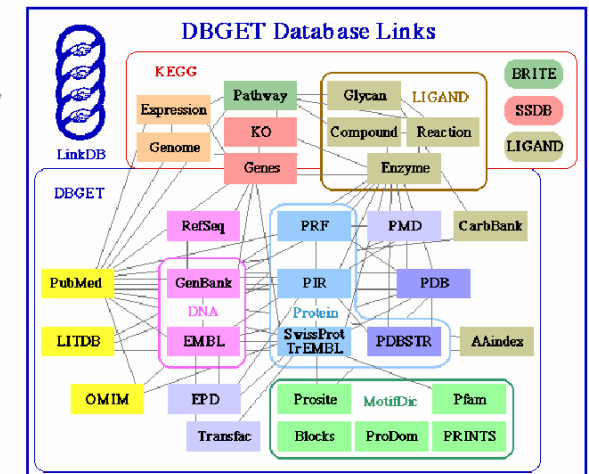
Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 26

Beispiel Datenintegration: Linkdatenbanken DBGet / DBLink

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

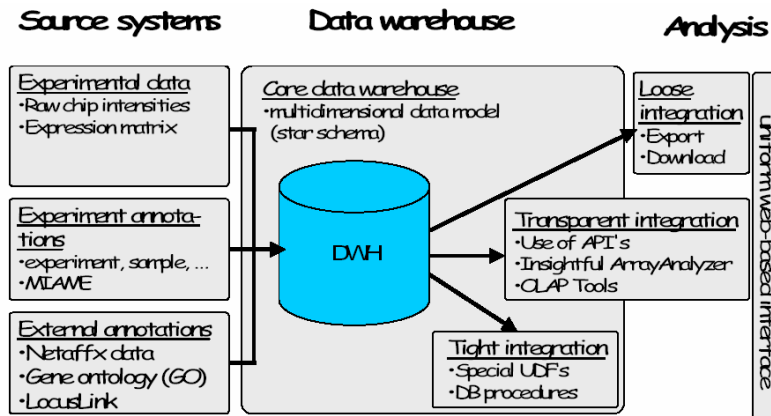
n Quelle: www.genome.ad.jp/dbget (Stand Okt. 2003)



Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 27

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>)

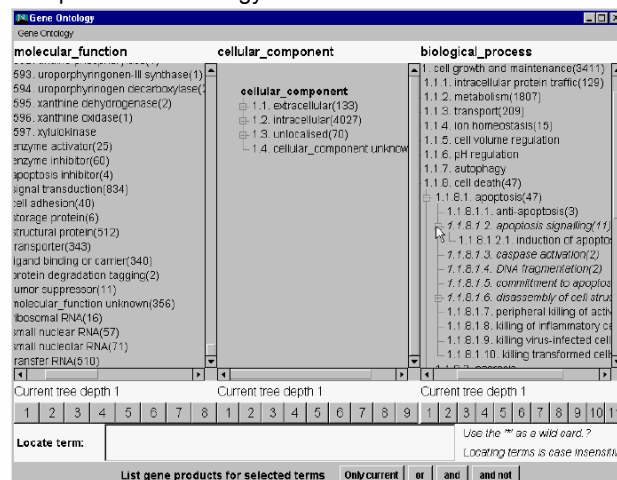
Klassifizierung nach Zugriffsmethoden

- 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)

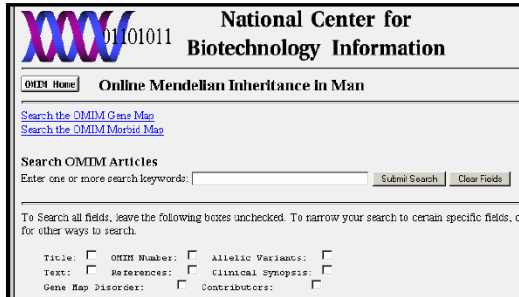
Beispiel: Gene Ontology



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



Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 32

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

Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 33

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

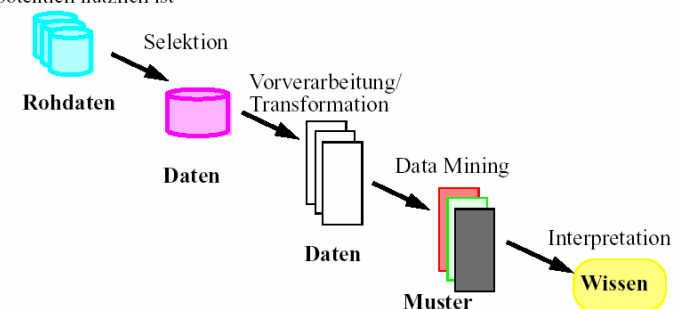
Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 34

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)

Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 35

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

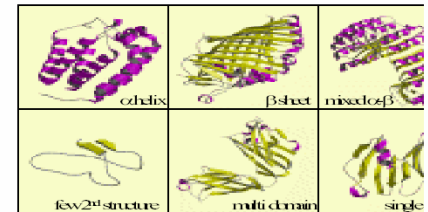
Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 36

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 => 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 . . .



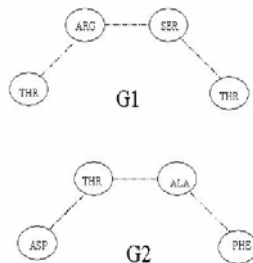
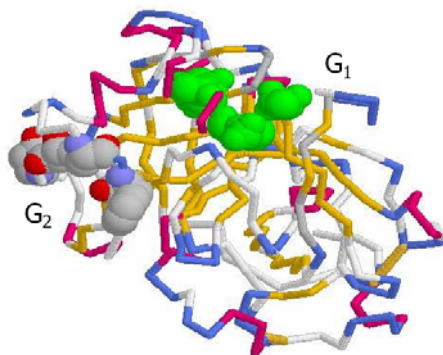
Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 37

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)



Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 38

Zusammenfassung

- n Bio-Datenbanken
 - forschungsgestrieben, 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

Sommersemester 2005/06

Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Seite 39