

Molekularbiologische Datenbanken

Einführung

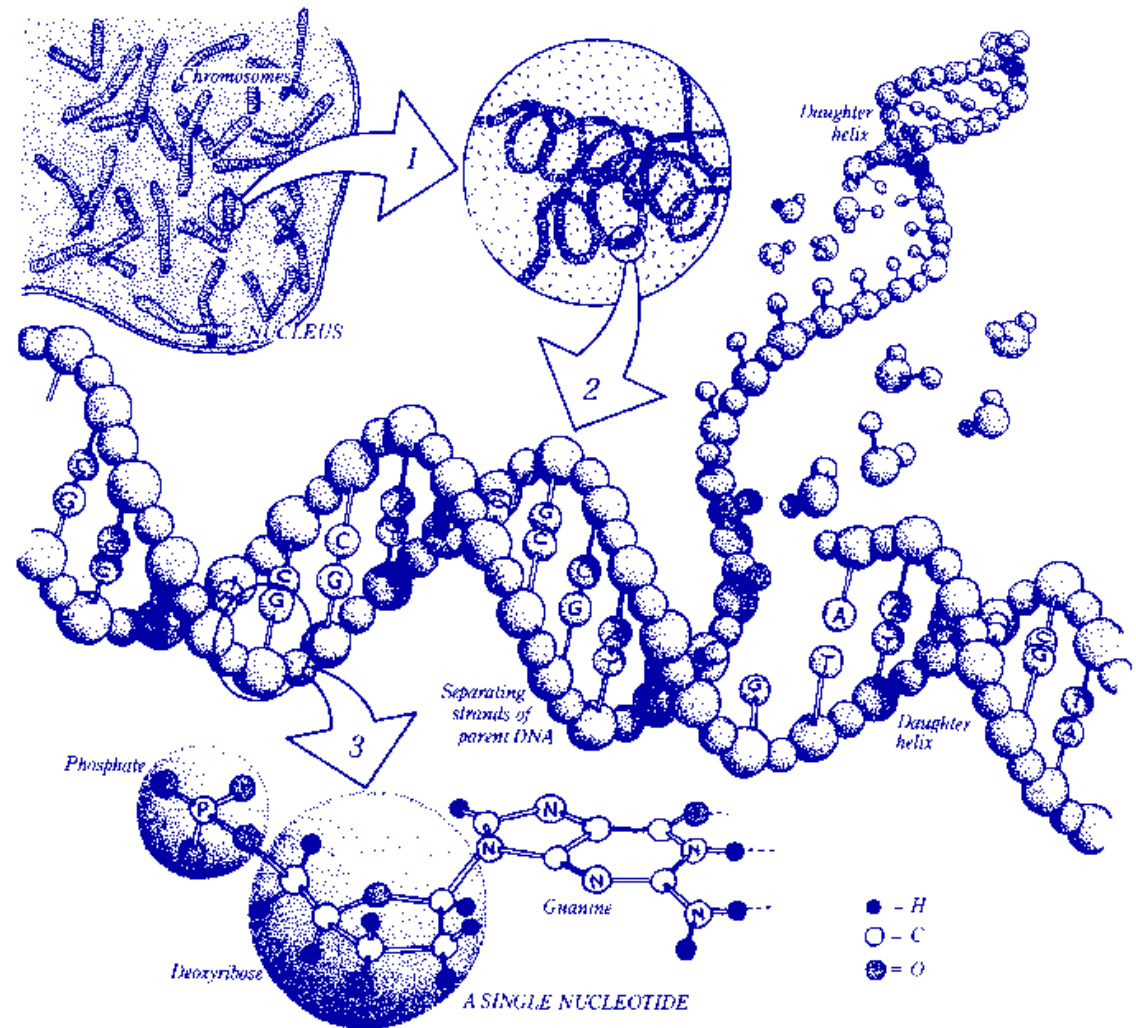
Ulf Leser

Wissensmanagement in der
Bioinformatik



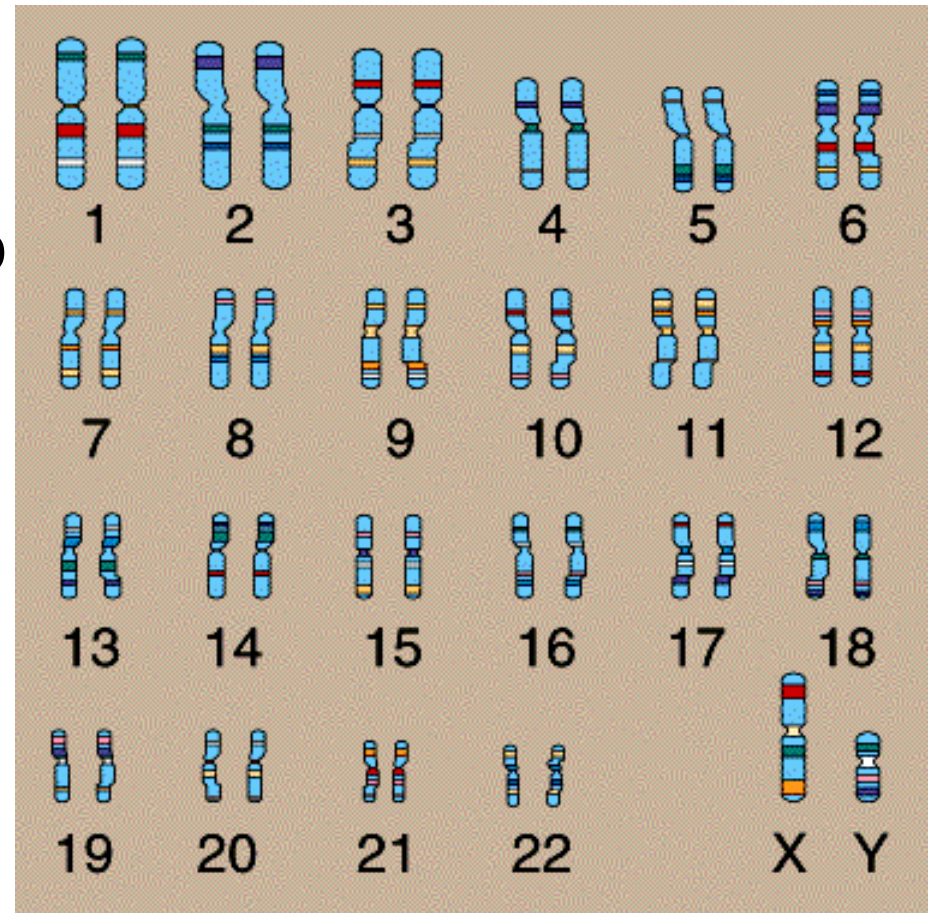
Zusammenfassung letzte Vorlesung

- Organismus
- Zelle
- Proteine
- Gene
- Sequenz

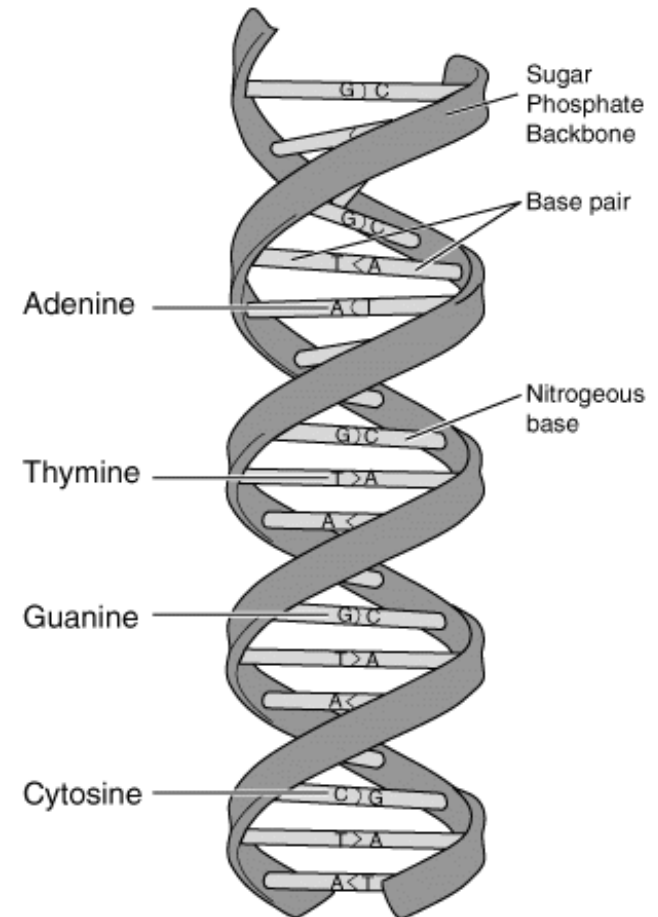
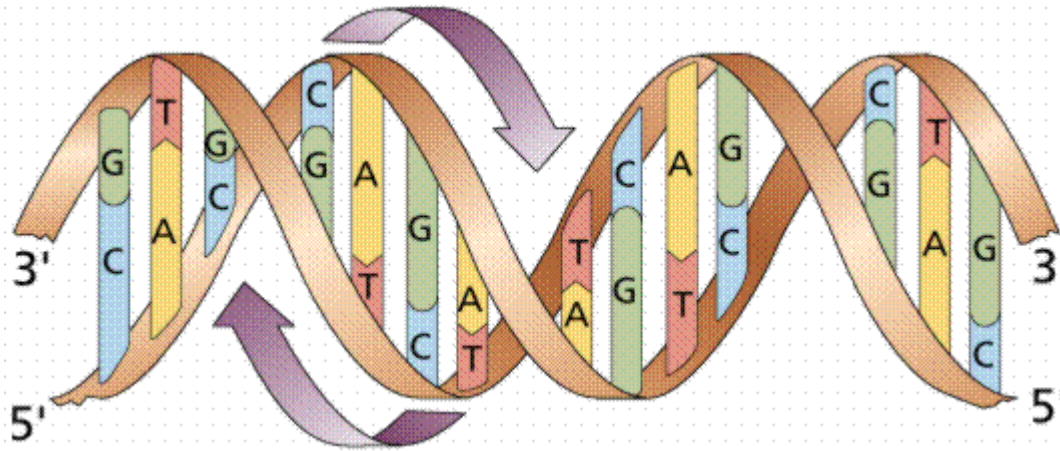


Genome

- Sequence of all chromosomes of a species is its genome
- Humans
 - App. 3.100.000.000 bp
 - 22 chromosomes + 2 sex chromosomes
 - Length 50–250 MB
 - App. 30.000 genes
 - App. 100.000 proteins

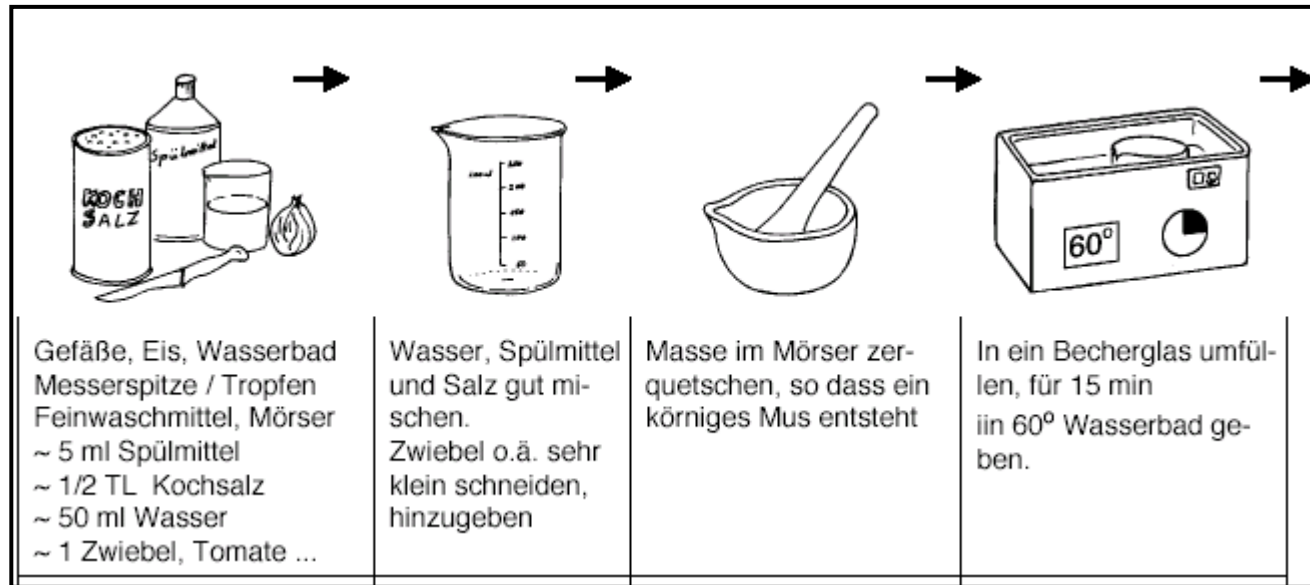


DesoxyriboNucleicAcid



- Chromosome: string of DNA
- Only 4 different nucleic acids
- Fixed pairs: A-T, G-C
- Same mechanism in all species

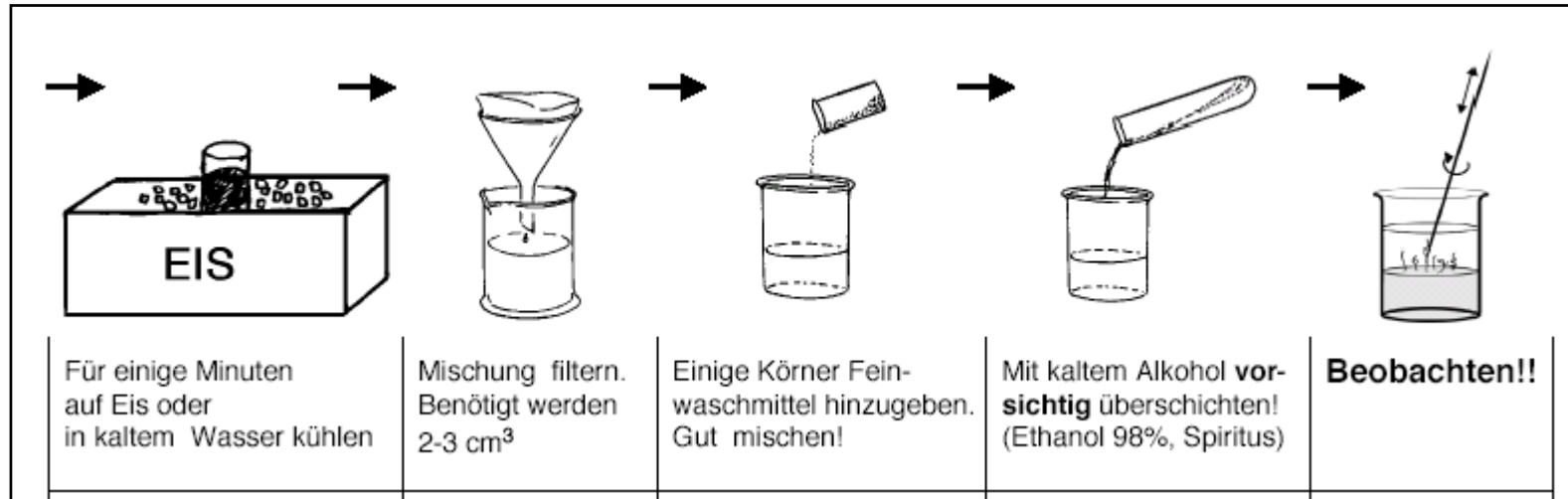
DNA sehen 1



Spülmittel (Tenside) zerstören
Zell- und Zellkernwände

Hitze unterstützt Zerstörung;
außerdem Degradation der DNAasen

DNA sehen 2



Nicht zu stark erhitzen;
auch DNA zerfällt

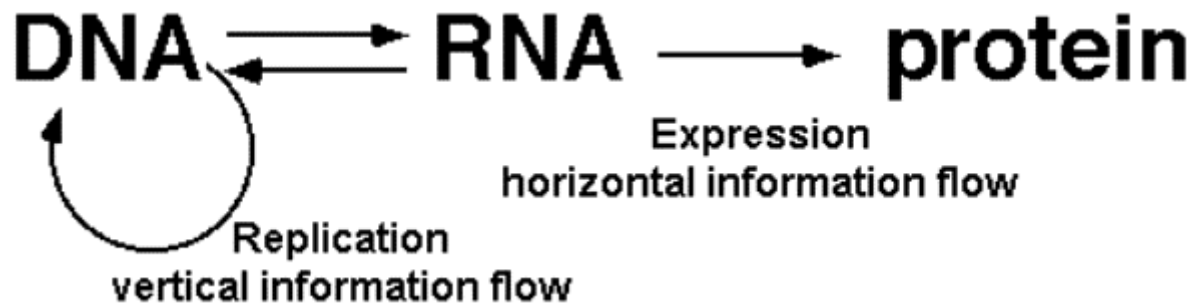
Proteasen zerstören Proteine
im Gemisch

Filtert Wandbestandteile
von DNA und Proteinen

DNA fällt in Alkohol in
Schlieren aus

Central Dogma

The Central Dogma of Molecular Genetics



Inhalt dieser Vorlesung

- Geschichte
 - Proteine & DNA
 - Human Genome Project
- Beispiele molekularbiologischer Datenbanken
 - DNA und Proteinsequenz
- Eigenschaften von MDB
 - Curation, Redundanz, Integration, ...

Teil I: Geschichte und Entwicklung

- Warum überhaupt Datenbanken öffentlich machen ?
- Anfänge molekularbiologischer Datenbanken
- Human Genome Projekt

Frühgeschichte

- 1865 Mendel'sche Vererbungslehre
 - Kreuzung von Erbsen
- 1909 Prägung des Wortes „Gen“ als Träger von vererbter Information
- 1910 Positionierung einzelner Gene auf Chromosomen der Fruchtfliege
 - Häufigkeit gemeinsamer Vererbung
- 1944 Nachweis, dass DNA die Erbinformation trägt
- 1951/53 Röntgenstrukturaufnahmen von DNA (Roslin), Aufdeckung der Doppelhelixstruktur (Watson/Crick)
- 1965 Genetischer Code: Triplets werden Aminosäuren (Mathaei/Ochoa)
- 1977 Methode zur Sequenzierung (Gilbert, Maxam, Sanger)



Motivation

- Warum gibt es öffentlich zugängliche molekularbiologische Datenbanken (MDB) ?
- Data Submission = **Veröffentlichung**
- Veröffentlichung
 - Austausch von Informationen und Daten
 - Erlangung von Reputation
 - Öffentliche Gelder – öffentliche Daten
- **MDB: Veröffentlichungen, nicht Tupel**

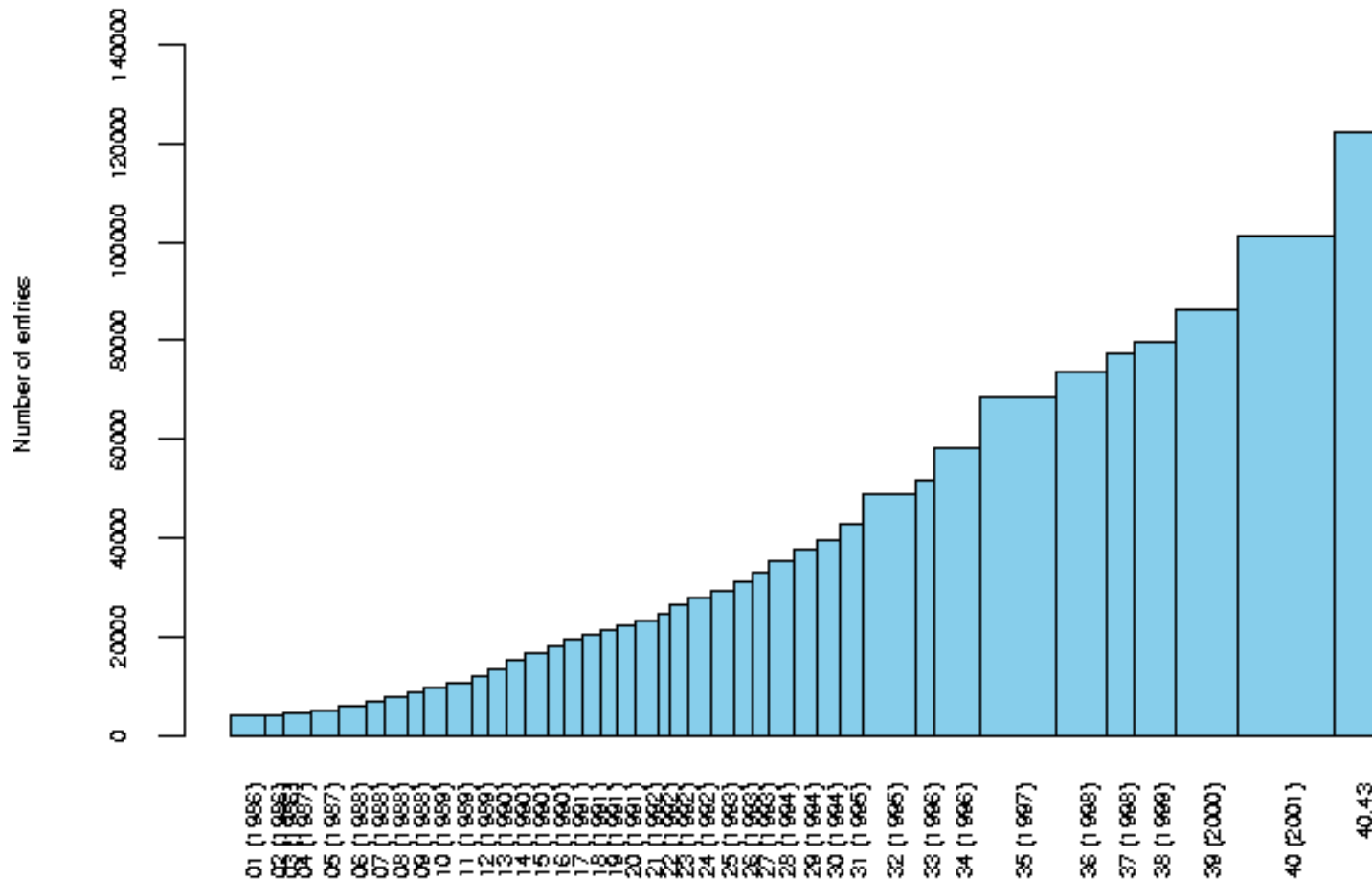
Veröffentlichungen

- Paper, Bücher, Reports
- 1907 / 1927: *Chemical/Biological Abstracts*
 - Sekundärliteratur: Abstracts, bibliographische Daten
- 1972 *Medline*
 - Publication Index (einer von vielen)
 - *PubMed*: <http://www.ncbi.nlm.nih.gov/entrez>
 - Heute: > 14.000.000 Artikel, > 5000 Journals
 - Online Links zu Volltextartikeln

Proteinsequenzen

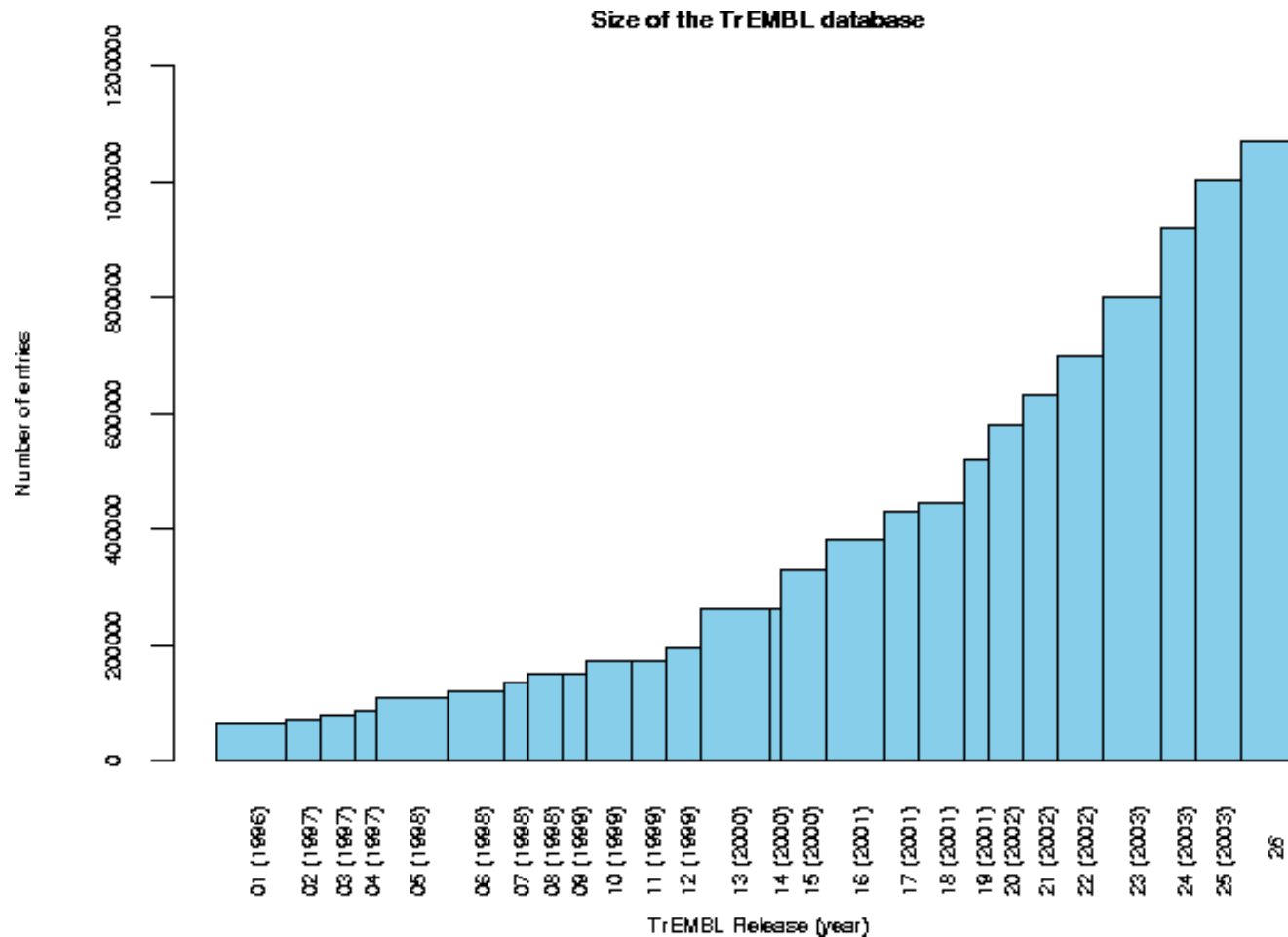
- Erste Sequenzen: Proteine
 - Länger bekannt und untersucht, einfacher zugänglich
 - DNA: Zellkern, kompakte Packung, große Länge, Instabilität
- Erste Proteinsequenz 1951 (Sanger & Tuppy):
Seitenkette von Insulin
- Sammlung ab Anfang der 1960er (Dayhoff et al. 1965)
 - Protein Sequence Atlas: [Buchform](#), 1968-1978
 - Motivation: Evolutionäre Untersuchungen
 - 1980: Protein Information Resource (seit 1988: PIR-Int.)
 - 1986: SWISS-PROT: Genf (Amos Bairoch) plus EBI

Wachstum SWISS-PROT



Source: ExPASy, Release 40.43 of 12-Feb-2003; 122.088 sequence entries

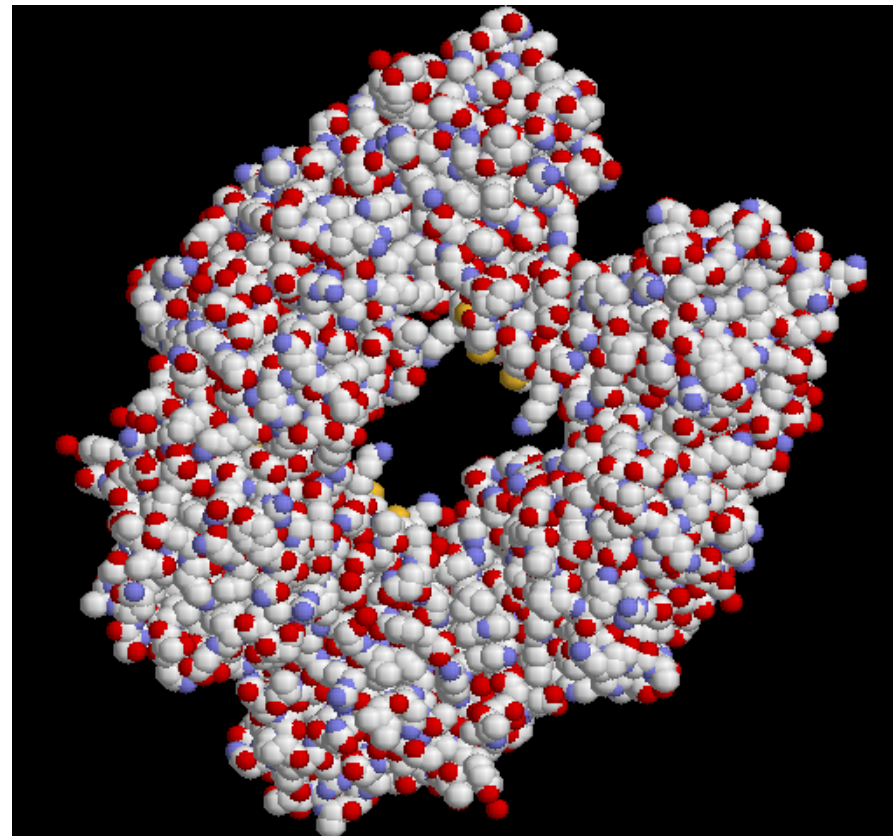
Zu viel Arbeit



Source: EBI, TrEMBL Release 26.0 04-Apr-2004; 1.069.649 sequence entries

Proteinstrukturen

- Proteine falten sich in komplexe Strukturen, die entscheidend für die Funktion ist
- Strukturaufklärung
 - Röntgenkristallographie (seit 50`er Jahren)
 - Nuclear Magnetic Resonance (NMR)
- Protein Data Bank (PDB)
 - Seit 1971 in Brookhaven
 - Seit 1999: Rutgers University
- Cambridge Structural Database: 1965

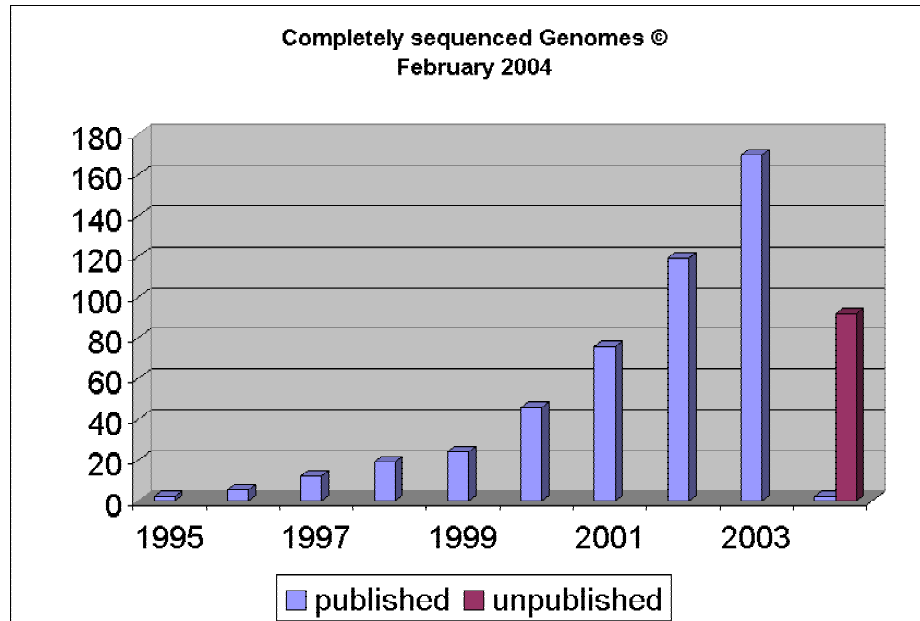


DNA Sequenzierung

- Doppelhelix der DNA: 1953 (Watson, Crick)
- Entwicklung DNA Sequenzierung Beginn der 1970er (Sanger, 1972): „radioactive dideoxy sequencing“
- Sammlungen von DNA Sequenzen
 - Los Alamos National Laboratory seit 1979
[GenBank](#) am NCBI (National Center f Biotech. Information)
 - European Molecular Biology Laboratory 1980
[EMBL](#) am EBI (European Bioinformatics Institute)
 - DNA Data Bank of Japan: 1986
[DDBJ](#) am NIG (National Inst. of Genetics)
- Int. Nuc. Sequence DB Collaboration seit ca. 1986
 - Täglicher Austausch
 - Jeweils verantwortlich für submittete Sequenzen

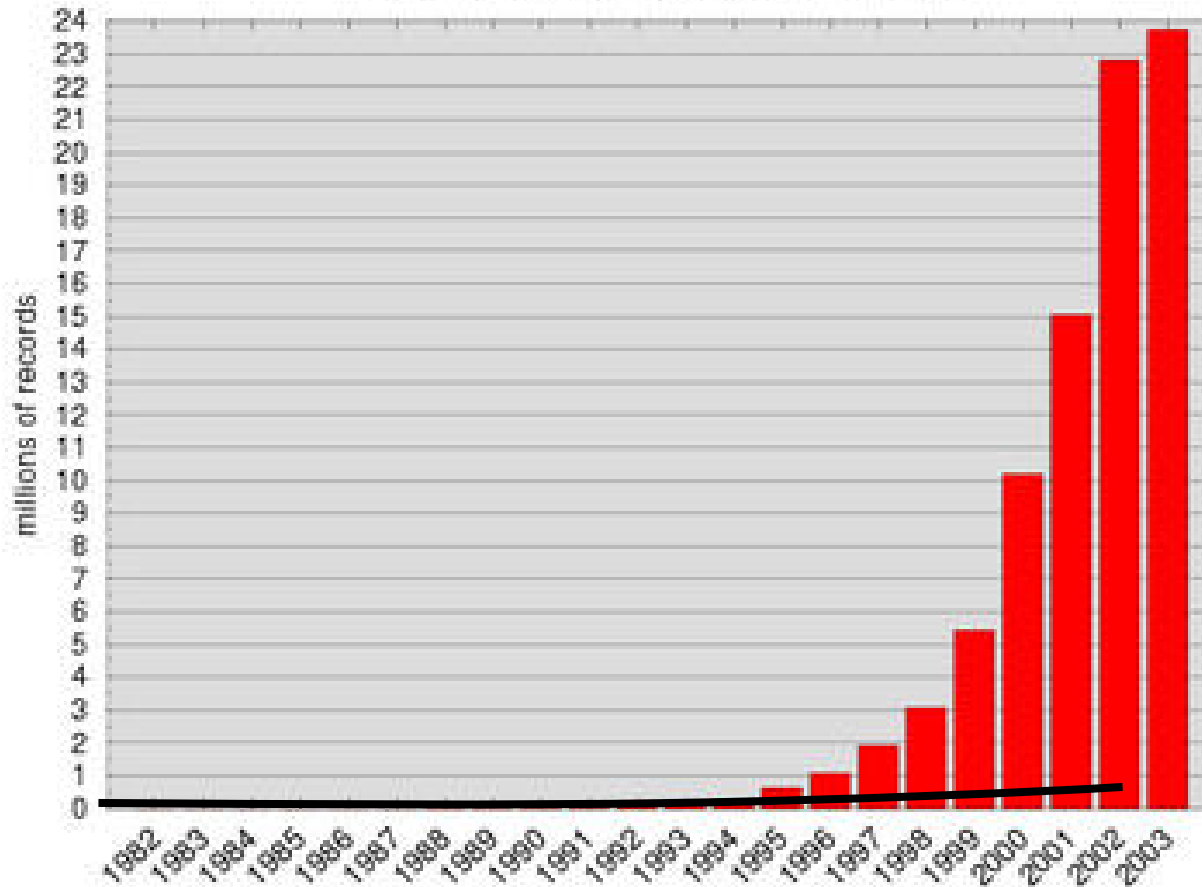
DNA Sequenzierung

- *Bacteriophage fX174*, 1979 (Sanger): 5.38 KB
- *Bäckerhefe*, 1996 (Internat.): 4.000.000 KB
- *Homo sapiens* (draft), 2001: 3.000.000.000 KB



Quelle: <http://wit.integratedgenomics.com/GOLD>, Stand 5.4.2004

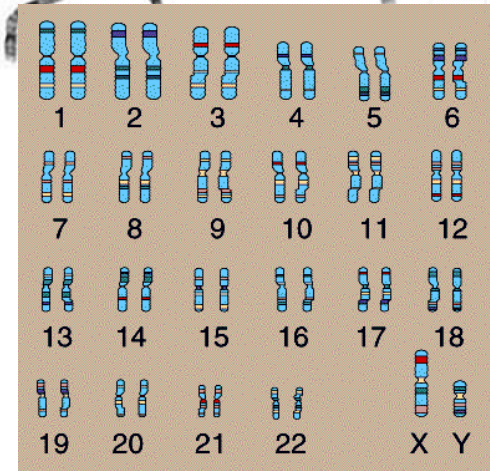
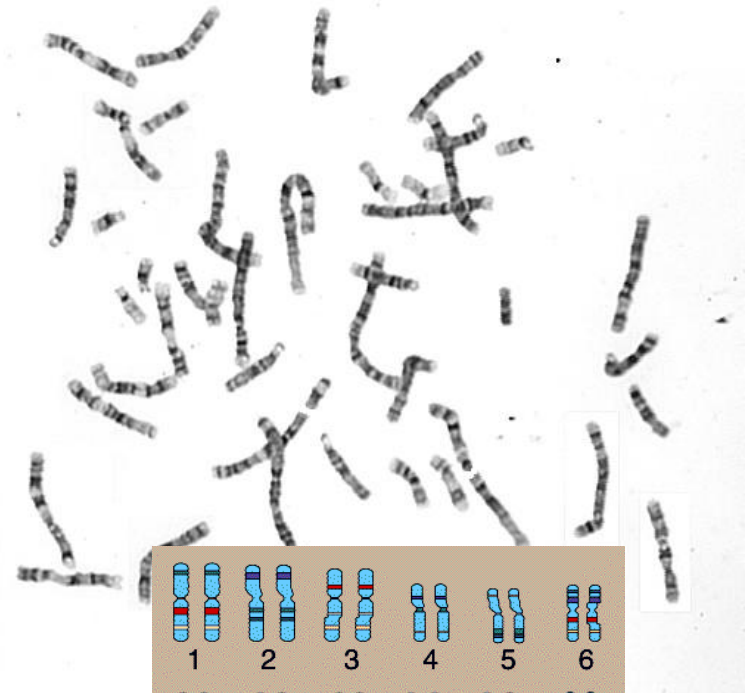
Vergleich EMBL / PDB



Quelle: EMBL, Genome Monitoring Tables, Stand 12.2.2003

Human Genome Projekt

- Start: ca. 1990
- Ziel
 - Sequenzierung des humanen Genoms
 - Identifikation aller Gene
 - ~3.000.000.000 Basenpaare
 - 25.000 – 33.000 Gene
- Ursprünglicher Plan: bis 2005
- Effekt wie Mondlandung:
Vision erzeugt Beschleunigung



Human Genome Projekt –2-

- Im Verlauf des Projektes
 - Neue Techniken, Strategien und Geräte
 - Sequenzierung immer billiger (5-8 € -> ~20 cents / base pair)
- Ergebnisse
 - Genom im Draft: 26.6.2000
 - Publikation Draft: 15.2.2001
 - Abschluss: 14.4.2003
- Wettlauf
 - öffentlich finanziert: Sanger, Whitehead, ...
 - privat finanziert: Celera

Commercial Sequencing

Project sequencing (0.9 to 30 kb)

Primer walking on both DNA strands

(99.5 %, per base pair) ds-pw 0.23

Publication quality

(99.9 %, per base pair) 0.46

Patent quality

(100 %, per base pair) 1.43

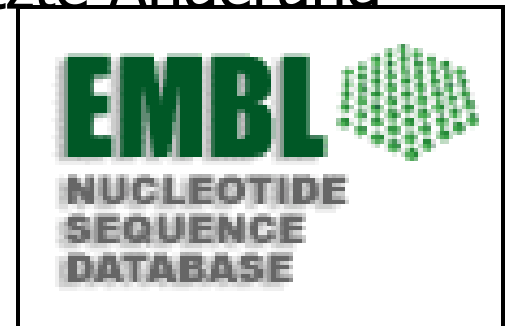
Quelle: www.genotype.de, Stand Feb 2003

Teil II: Zwei Beispiele

- EMBL
- SWISS-PROT

EMBL / Genbank / DDBJ

- Repository aller bekannten DNA Sequenzen
- Basiert auf Submission
- Archivcharakter, keine Verbesserung der Daten
- **Flaches Datenmodell**
 - Sequenz
 - Autor, Datum, Länge, Methode, letzte Änderung
 - Sequenzart, Organismus
 - „Feature“ auf der Sequenz
 - Literaturreferenzen
 - ...



EMBL – Beispieleintrag

Global identifier →

Accession id →

Local identifier & version →

Description: free →

Keyword: free →

Taxonomy: ctrl ! →

References: redundant →

X-Ref: free →

Comment: free →

Feature: partly ctrl →

Sequence →

```
ID  HSIGHAF  standard; RNA; HUM; 1089 BP.
XX
AC  J00231;
XX
NI  g185041
XX
DT  17-DEC-1994 (Rel. 42, Last updated, Version 6)
XX
DE  Human Ig gamma3 heavy chain disease OMM protein mRNA.
XX
KW  C-region; gamma heavy chain disease protein;
XX
OC  Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Primates;
XX
RN  [1]
RP  1-1089
RX  MEDLINE; 82247835.
...
DR  GDB; 119339; IGHG3.
DR  GDB; G00-119-339.
...
CC  The protein isolated from patient OMM is a gamma heavy chain
FH
FT  CDS                23. .964
FT                    /codon_start=1
FT                    567112"
XX
SQ  Sequence 1089 BP; 240 A; 358 C; 271 G; 176 T; 44 other;
    CCTGGACCTC CTGTGCAAGA ACATGAAACA NCTGTGGTTC TTCCTTCTCC TGGTGGCAGC      60
    TCCCAGATGG GTCCTGTCCC AGGTGCACCT GCAGGAGTCG GGCCCAGGAC TGGGGAAGCC      120
    ...
```

Ulf Leser:

SWISS-PROT

- Repository aller bekannten Proteinsequenzen
- Basiert auf Submission, DNA-Übersetzung und aktiver Suche
- Intensive (manuelle) Verbesserung und Ergänzungen
- **Flaches Datenmodell**
 - Proteinsequenz, Version, Vorgänger
 - Autor, Datum, Länge, Methode, letzte Änderung
 - Organismus
 - Feature auf der Sequenz
 - Links zu anderen Datenquellen
 - Literaturreferenzen
 - ...



SWISS-PROT Beispieleintrag

ID	GUMB_CLOTM	STANDARD;	PRT;	563 AA.		
AC	P04956;					
DT	13-AUG-1987	(Rel. 05, Created)				
DT	13-AUG-1987	(Rel. 05, Last sequence update)				
DT	01-FEB-1995	(Rel. 31, Last annotation update)				
DE	ENDOGLUCANASE B PRECURSOR	(EC 3.2.1.4) (EGB) (ENDO				
DE	(CELLULASE B).					
GN	CELB.					
OS	Clostridium thermocellum.					
OC	Bacteria; Firmicutes; Bacillus/Clostridium group;					
OC	Clostridium.					
OX	NCBI_TaxID= 1515 ;					
RN	[1]					
RP	SEQUENCE FROM N.A.					
RC	STRAIN=NCIB 10682;					
RX	MEDLINE= 86148508 ; PubMed= 3453102 ;					
RA	Grepinet O., Beguin P.;					
RT	"Sequence of the cellulase gene of Clostridium the					
RT	endoglucanase B.";					
RL	Nucleic Acids Res. 14:1791-1799(1986).					
CC	-!- FUNCTION: THIS ENZYME CATALYZES THE ENDOHYDROL					
CC	GLUCOSIDIC LINKAGES IN CELLULOSE, LICHENIN AND					
CC	GLUCANS.					
CC	-!- CATALYTIC ACTIVITY: ENDOHYDROLYSIS OF 1,4-BETA					
CC	LINKAGES IN CELLULOSE.					
CC	-!- DOMAIN: A 24 RESIDUES DOMAIN IS REPEATED TWICE					
CC	WELL AS IN OTHER C.THERMOCELLUM CELLULOSOME EN					
CC	MAY FUNCTION AS THE BINDING LIGAND FOR THE SL					
CC	-!- SIMILARITY: BELONGS TO CELLULASE FAMILY A (FAM					
CC	HYDROLASES).					
DR	EMBL; X03592 ; CAA27266.1 ; -.					
DR	PIR; A23512; CZCLBM.					
DR	HSSP; P54583; IECE .					
DR	InterPro; IPR002105 ; Dockerin_1.					
DR	InterPro; IPR002048 ; EF-hand.					
DR	InterPro; IPR001547 ; Glyco_hydro_F5.					
DR	Pfam; PF00150 ; cellulase; 1.					
DR	Pfam; PF00404 ; Dockerin_1; 2.					
DR	PROSITE; PS00018 ; EF_HAND; UNKNOWN_1.					
DR	PROSITE; PS00448 ; CLOS_CELLULOSOME_RPT; 2.					
DR	PROSITE; PS00659 ; GLYCOSYL_HYDROL_F5; 1.					
KW	Cellulose degradation; Hydrolase; Glycosidase; Repeat; Signal.					
FT	SIGNAL	1	27	OR 31.		
FT	CHAIN	28	563	ENDOGLUCANASE B.		
FT	ACT_SITE	204	204	PROTON DONOR (BY SIMILARITY).		
FT	ACT_SITE	363	363	NUCLEOPHILE (BY SIMILARITY).		
FT	DOMAIN	502	557	2 X 24 AA APPROXIMATE REPEATS.		
FT	REPEAT	502	526	1.		
FT	REPEAT	534	557	2.		
SQ	SEQUENCE	563 AA;	63929 MW;	866FE55704A1DE4B CRC64;		
	MKKFLVLLIA	LIMIATLLVV	PGVQTSAEGS	YADLAEPDDD	WLHVEGTNIV	DKYGNKVVIT
	GANWFGFNCR	ERMLLD SYHS	DIIADIELVA	DKGINVVRMP	IATDLLYAWS	QGIYPPSTD
	SYNMPALAGL	NSYELFNFML	ENFKRVGIVK	ILDVHSPETD	NQGHNYPLWY	NTTITEEIFK
	KAWVWVAERY	KNDDTIIGFD	LKNEPHTMTG	TMKIKAQSAI	WDDSNHPNHW	KRVAEETALA
	ILEVHPNVL I	FVEGVEMYPK	DGIWDETTFD	TSPWTGNNDY	YGNWWGGNLR	GVKDYPINLG
	KYQSQLVYSP	HDYGPVVEQ	DWFKGDFITA	NDEQAKRILY	EQCWRDNWAY	IMEEGISPLL
	LGEWGGMTEG	GHPLLDLNLK	YLRCMRDFIL	ENKYKLHHTF	WCINIDSADT	GGLFTRDEGT
	PFPGGRD LKW	MDNKYDNYLY	PVLWKTEDGK	FIGLDHKIPL	GRNGISISQL	SNYTPSVTPS
	PSATPSPTTI	TAPPTDTVTY	GDVNGDGRVN	SSDVALLKRY	LLGLVENINK	EADVNVSGT
	VNSTDLAIMK	RYVLRSEL	PYK			
	//					

SWISS-PROT Web Interface

NiceProt View of SWISS-PROT: P29358

P29358

Printer-friendly view Quick BlastP search

[\[General\]](#) [\[Name and origin\]](#) [\[References\]](#) [\[Comments\]](#) [\[Cross-references\]](#) [\[Keywords\]](#) [\[Features\]](#) [\[Sequence\]](#) [\[Tools\]](#)

General information about the entry

Entry name	143B_BOVIN
Primary accession number	P29358
Secondary accession numbers	None
Entered in SWISS-PROT in	Release 24, December 1992
Sequence was last modified in	Release 33, February 1996
Annotations were last modified in	Release 41, June 2002

Name and origin of the protein

Protein name	14-3-3 protein beta/alpha
Synonyms	Protein kinase C inhibitor protein-1 KCIP-1
Gene name	YWHAB
From	Bos taurus (Bovine) [TaxID: 9913] Ovis aries (Sheep) [TaxID: 9940]
Taxonomy	Eukaryota ; Metazoa ; Chordata ; Craniata ; Vertebrata ; Euteleostomi ; Mammalia ; Eutheria ; Cetartiodactyla ; Ruminantia ; Pecora ; Bovoidea ; Bovidae ; Bovinae ; Bos .

References

[1] SEQUENCE.
SPECIES=Bovine;
MEDLINE=91108808; PubMed=1671102; [NCBI, ExPASy, EBI, Israel, Japan]

Dokument: Done (2.273 Sek.)

Text -> Daten

- Alle (großen) „klassischen“ und öffentlichen Datenbanken entstanden aus Büchern
 - Sammlungen bekannter Daten einer Art: DNA, Proteinsequenz, Proteinstruktur
 - Jährliche / quartalsweises Erscheinen
 - Buch – Band – CD – FTP – WWW
- Flaches, textorientiertes Datenmodell
 - Aufbau nach „Entries“
 - Viele Beschreibungen in freier Textform
 - Für Menschen, nicht Computer

Entwicklung

Thema	Entwicklung
Austauschformat	Frei → definierte Felder → XML
Vokabular	Frei → Controlled Vocabularies → Ontologien
Struktur	Entryliste → Featuretables (einfache Schachtelung) → Subentries (hierarchische Schachtelung)
Datenmodell	Ad-Hoc → ER/Relational → OO/UML
Technologie	Proprietär → RDBMS / OR

Sukzessive Übernahme von DB
Techniken

Teil III: Eigenschaften von MDB

- Weltweit über 200 ? 400 ? 1000 ? MDB
- NAR 1/2002: 335 „Key Databases“
- Wichtige Unterscheidungskriterien
 - I.d.R. unscharf
 - Nützlich zur Charakterisierung
 - Nicht dogmatisch interpretieren

E1: Passiv oder aktiv

- Model 1:
 - Alle Daten werden von Erzeugern eingeschickt
 - Submission-based
 - Sinn: **Archivierung**, ID Vergabe, globaler Zugriff
 - Verpflichtung durch Journals, Geldgeber und Gewissen
 - Beispiele: Genbank/EMBL/DDBJ, PDB, ArrayExpress, ...
- Model 2:
 - Daten werden von den Datenbankbetreibern gesammelt
 - Journals, Experten, andere Datenbanken, ...
 - Sinn: Integration, **Veredlung, Vollständigkeit**
 - Beispiele: SWISS-PROT, PIR, ...
- Mischformen existieren: GDB

Passiv oder aktiv

„Public money – public data“
Patentierung und Lizenzierung
Wem gehört eine Sequenz ?
Biodiversity (Nutzpflanzen)

E2: Database Curation

- Curation: Verbesserung, „Heilung“
- Modell 1:
 - Submitter/Quelle ist Datenherr
 - Keine (inhaltlichen) Veränderungen
 - Beispiele: PDB, Genbank, ArrayExpress, ...
- Modell 2:
 - Daten werden laufend verbessert
 - Hoher (manueller) Aufwand
 - Beispiele: SWISS-PROT, MGD, MIPS, ...

Urheber klar
Hohe Datenstabilität

Urheber weniger klar
Hohe Volatilität
Versionierung essentiell

Beispiel

- SWISS-PROT

- *The Swiss-Prot Protein Knowledgebase is a **curated protein sequence database that provides a high level of annotation, a minimal level of redundancy and high level of integration** with other databases.*

- EMBL

- *All database records submitted to the INSD will remain permanently accessible as part of the scientific record. **Corrections of errors and update of the records by authors are welcome** and erroneous records may be removed from the next database release, but all will remain permanently accessible by accession number.*
- *Beyond limited editorial control and some internal integrity checks (for example, proper use of INSD formats and translation of coding regions specified in CDS entries are verified), the **quality and accuracy of the record are the responsibility of the submitting author**, not of the database.*

E3: Redundanz

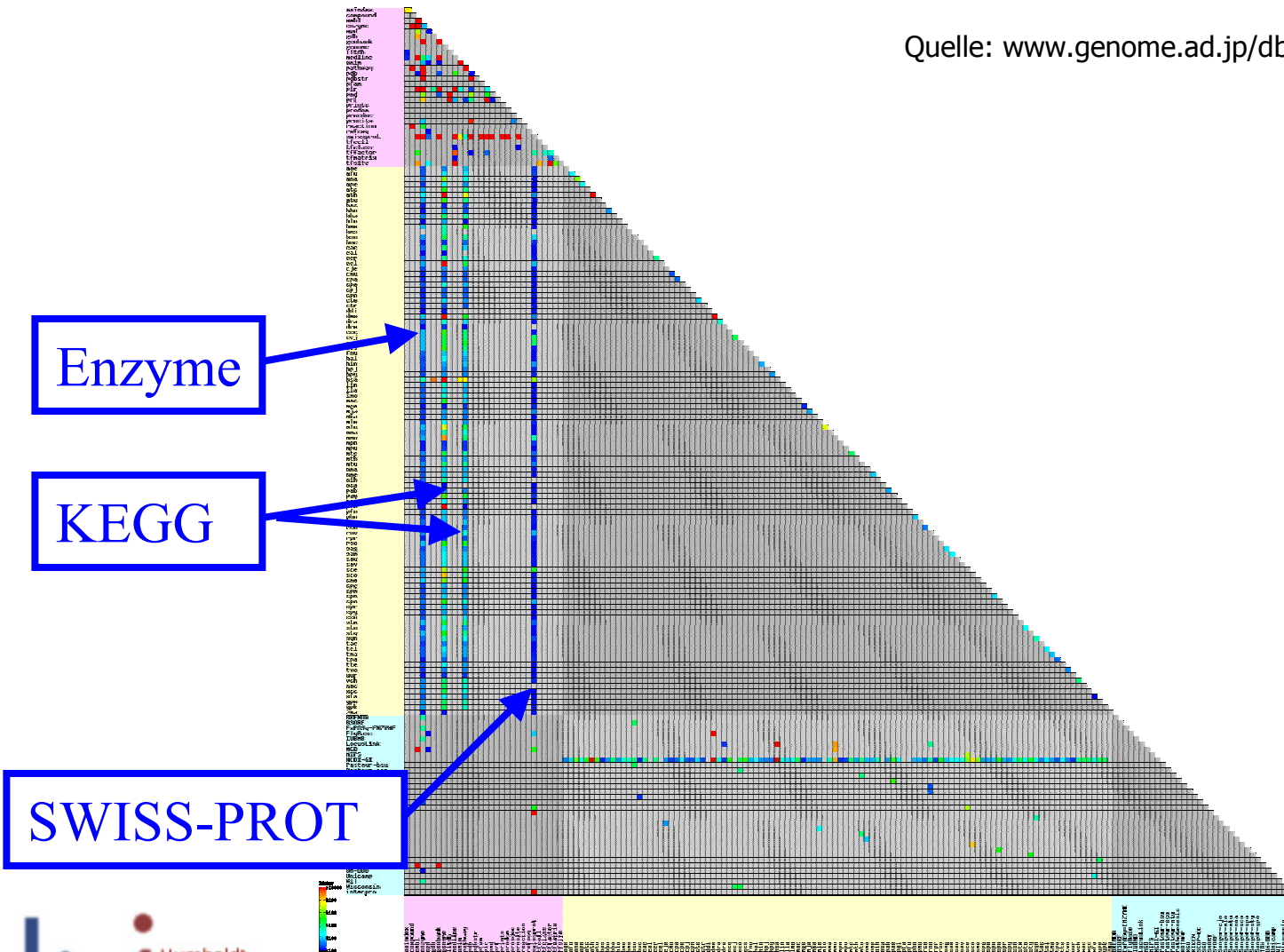
- Model 1:
 - Alles aufnehmen
- Modell 2:
 - Entfernen gleicher oder sehr ähnlicher Einträge
 - Kann nur bei „kurierten“ Datenbanken funktionieren
- Was ist sehr ähnlich?
 - Gleiches Protein in anderer Spezies?
 - Gleiches Gen an anderer Position?
- Ähnlichkeit muss individuell definiert werden
 - Beispiel Swiss-Prot: Redundanzminimierung durch Editoren
 - Beispiel UniGene: Redundanzminimierung durch Algorithmen

E4: Integration

- Modell 1
 - DB als Archiv unverbundener Objekte
 - Problemloses Löschen / Hinzufügen von Objekten
- Modell 2
 - Objekte als Knoten in einem komplexen Geflecht
 - Kanten ([Cross-references](#)) zwischen Knoten müssen laufend erzeugt, erneuert, ergänzt werden
 - Neue „Knoten“ können Änderungskaskaden auslösen
 - Kanten im Interface: WWW Hyperlinks
- Verschiedene Integrationstiefen
 - „Reine“ Links zu anderen Quellen (DBGET, Integr8)
 - Inhaltliche Integration anhand def. Kriterien (Ensembl)
 - Manuelle Integration anhand Wissen des Editors (Swiss-Prot)

Linkdatenbanken DBGet / DBLink

Quelle: www.genome.ad.jp/dbget Stand 18.2.2003



Enzyme

KEGG

SWISS-PROT

Enge Integration

SWISS-PROT Entry
Manuelle Erstellung
XRefs zu ca. 50 anderen DB

Comments

FUNCTION

MAY PLAY A FUNCTION.

SUBCELLULAR LOCATION

CYTOPLASM

TISSUE SPECIFICITY

WIDELY EXPRESSED IN THE BRAIN, INCLUDING THE CEREBELLAR CORTEX, THE NEOCORTEX, THE STRIATUM, AND THE HIPPOCAMPAL FORMATION.

PTM

CLEAVED BY APOPAIN DOWNSTREAM OF THE POLYGLUTAMINE STRETCH. THE RESULTING AMINO-TERMINAL FRAGMENT IS CYTOTOXIC AND PROVOKES APOPTOSIS.

POLYMORPHISM

THE POLY-GLN REGION OF HD IS HIGHLY POLYMORPHIC (10 TO 35 REPEATS) IN THE NORMAL POPULATION AND IS EXPANDED TO ABOUT 36-120 REPEATS IN HD PATIENTS. THE REPEAT LENGTH USUALLY INCREASES IN SUCCESSIVE GENERATIONS, BUT CONTRACTS ALSO ON OCCASION. THE LONGER EXPANSIONS RESULT IN EARLIER ONSET AND MORE SEVERE CLINICAL MANIFESTATIONS OF THE DISEASE. THE ADJACENT POLY-PRO REGION IS ALSO POLYMORPHIC AND VARIES BETWEEN 7-12 RESIDUES. POLYGLUTAMINE EXPANSION LEADS TO ELEVATED SUSCEPTIBILITY TO APOPAIN CLEAVAGE AND LIKELY RESULT IN ACCELERATED NEURONAL APOPTOSIS.

DISEASE

DEFECTS IN HD ARE THE CAUSE OF HUNTINGTON'S DISEASE, AN AUTOSOMAL DOMINANT NEURODEGENERATIVE DISORDER CHARACTERIZED BY INVOLUNTARY MOVEMENTS (CHOREA), GENERAL MOTOR IMPAIRMENT, PSYCHIATRIC DISORDERS AND DEMENTIA. ONSET OF THE DISEASE OCCURS USUALLY IN THE THIRD OR FOURTH DECADE OF LIFE AND SYMPTOMS PROGRESSIVELY WORSEN LEADING TO DEATH IN 10 TO 20 YEARS. IT AFFECTS 1 IN 10,000 INDIVIDUALS OF EUROPEAN

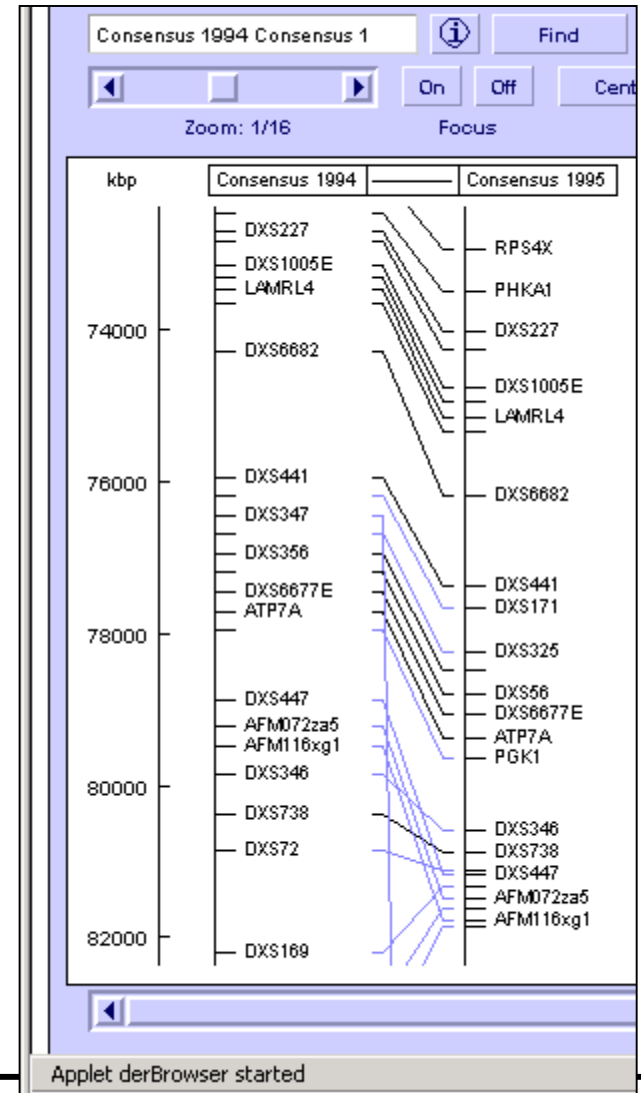


Enge Integration –2-

- IXDB
- Vergleich von Genomkarten nach Namen
- Aufwendige und manuelle Synonymsuche

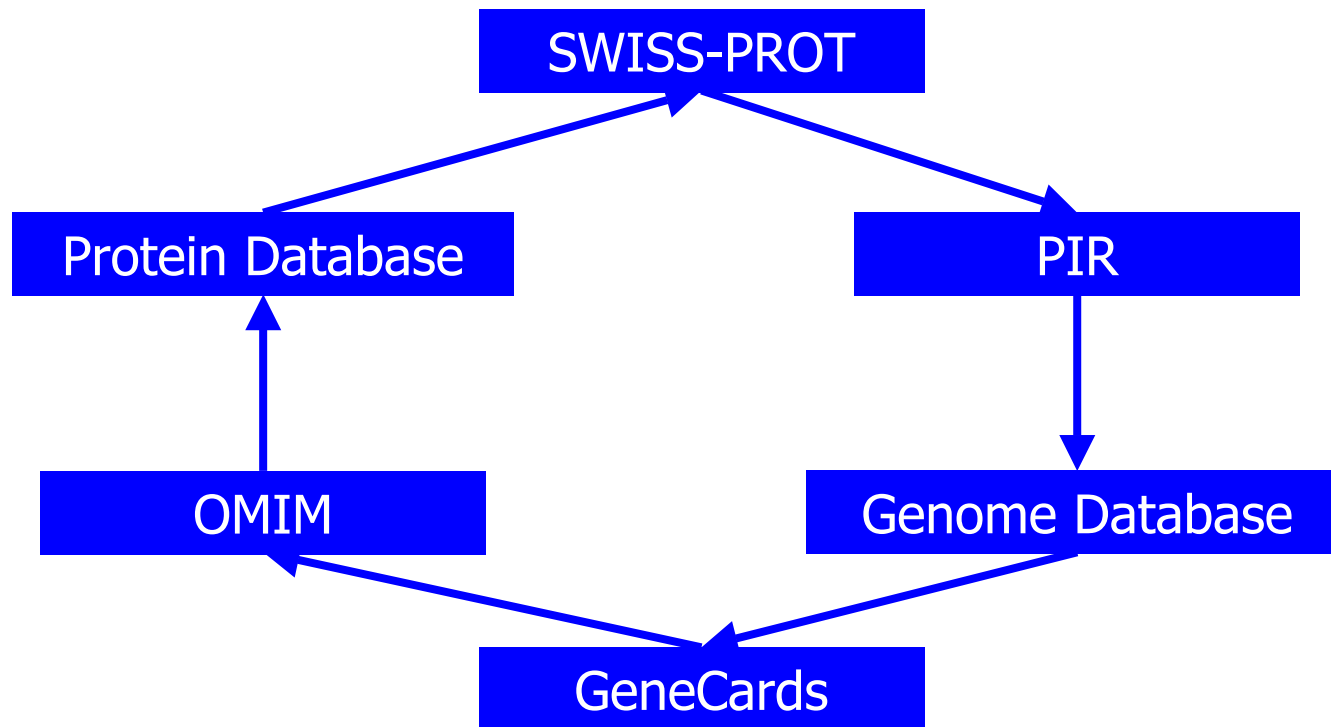
Inhaltliche Integration ist der Mehrwert !

Quelle: IXDB, Stand 18.2.2003



Einschub: Linktour

Duchenne Muskel Dystrophie



Links in MDB –1–

The screenshot shows the MDB search interface. At the top, there is a search bar with the text 'duchenne' and a 'Quick Search' label. Below the search bar, there are two blue circles: one around the search bar and another around the 'SWALL (SPTR)' link in the 'Sequence libraries - complete' section. The interface includes a 'Reset' button, 'All Entries' link, and a sidebar with 'Query forms' (Standard and Extended) and 'bookmark this link' information. The main content area lists various database resources with checkboxes for selection.

Reset Quick Search All Entries

show all + collapse all -

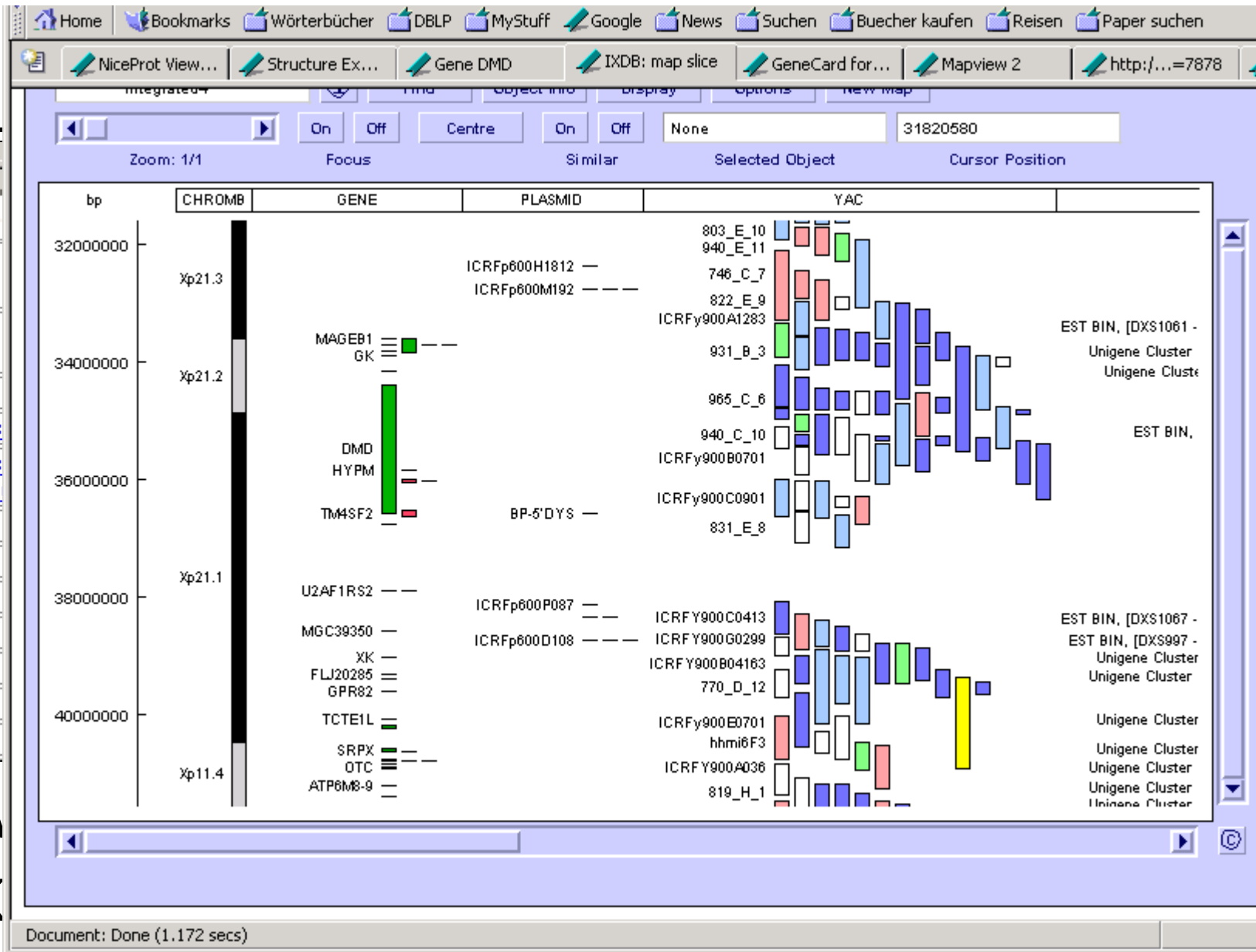
- [-] **References**
 - all [MEDLINE](#)
 - References - subsections
 - all [MEDLINE \(Updates\)](#) [MEDLINE \(Main Release\)](#)
- [-] **Gene Ontology resources**
 - all [GOA](#) [GO](#)
- [-] **Sequence libraries - complete**
 - all [SWALL \(SPTR\)](#) [PIR](#) [IPI](#)
 - [RemTrEMBL](#) [ENSEMBL HUMAN](#) [ENSEMBL MOUSE](#) [ENSEMBL FLY](#)
 - [PATENT PRT](#) [JPO PRT](#) [PATENT DNA](#) [USPO PRT](#)
 - [ENSEMBL FISH](#) [IMGTLIGM-DB](#) [IMGTHLA](#) [MHCBN](#)
- [-] **Sequence libraries - subsections**

- Startpunkt: SWISS-PROT (SRS, EBI) DMD_HUMAN

Links in MDB –3–

Domain	WW repeat homology	3055 - 309
Region	cysteine-rich	3080 - 330
Region	leucine zipper motif	3506 - 350
Region	leucine zipper motif	3572 - 359
Genetics		
Gene	GDB:DMD	
Cross-References	GDB: 119850 ; OMIM:310200	
Intron	11/1; 51/3; 62/3; 88/3; 119/3; 177/2; 217/1; 277/3; 320/3; 383/3; 444/2; 494/3; 534/3; 568/3; 604/3; 664/3; 72/3; 874/3; 935/1; 1816/3; 1862/3; 1913/3; 1974/3; 2890/1; 3028/3; 3055/1; 3075/2; 3096/1; 3121/1; 3188/2; 3217/3; 3362/3; 3408/2; 3421/2; 3443/2; 3465/2; 3518/2; 3599/3; 3641/1; 3672/1; 3682/3	
Note	the list of introns is incomplete	
Sequence		
	MLWEEVEDC YEREDVQKKT FTKWVNAQFS KFGKQHIENL FSDLQDGRRL LDLLBGLTCQ	60
	KLPKEKGSTR VHALNNVKA LRVLQNNVVD LVNIGSTDIV DGNHRLTLGL IWNIIILHWQV	120
	KNVMKNIMAG LQQTNSEKIL LSWVRQSTRN YPQVNVINFT TSWSDGLALN ALIHSHRPDL	180
	FDWNSVVCQQ SATQRLEHAF NIARYQLGIE KLLDPEDVDT TYPDKSILM YITSLFQVLP	240
	QQVSIEAIQE VEMLPRPPKV TREEHFQLHH QMHYSQITV SLAQGYERTS SPKPRFKSYA	300
	YTQAAAYVTT DPTRSPFPSQ HLEAPEDKSF GSSLMSEVNV LDRYQTALEE VLSWLLSAED	360
	TLQAQGEISN DVEVVKDQFH THEGYMMDLT AHQGRVGNIL QLGSKLIGTC KLSDEDETEV	420
	QEQMNLNLSR WECLRVASME QQSNLHRVLM DLQNQKLKEL NDWLTKEER TRKMEEEPLC	480

- Domains, Sequence
- Link zu GDB [119850](#)



- NiceProt view...
- View Maps of Region
- w...
- bol/Names:**
- [DMD](#)
 - [dystrophin \(mus](#)
 - [dystrophin \(mus](#)
 - [DXS239, DXS2](#)
 - [DXS272](#)
 - [DXS142](#)
 - [DXS164](#)
 - [DXS206](#)
 - [DXS230](#)
 - [DXS239](#)
 - [DXS268](#)

- A
- K

Links in MDR _7_

Structure Explorer - 1DXX - Mozilla {Build ID: 2002082606}

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop

http://www.rcsb.org/pdb/cgi/explore.cgi?pid=11801045588836&pdbId=1DXX

Print

Home Bookmarks Wörter

NiceProt View of Swiss-Prot: P

PDB
PROTEIN DATA BANK

Summary Information

[View Structure](#)

[Download/Display File](#)

[Structural Neighbors](#)

[Geometry](#)

[Other Sources](#)

[Sequence Details](#)

[Structure Factors \(compressed\)](#)

Explore

[SearchLite](#) [SearchFields](#)

Document: Done (3.075 secs)

View Structure

[Summary Information](#)

[View Structure](#)

[Download/Display File](#)

[Structural Neighbors](#)

[Geometry](#)

[Other Sources](#)

[Sequence Details](#)

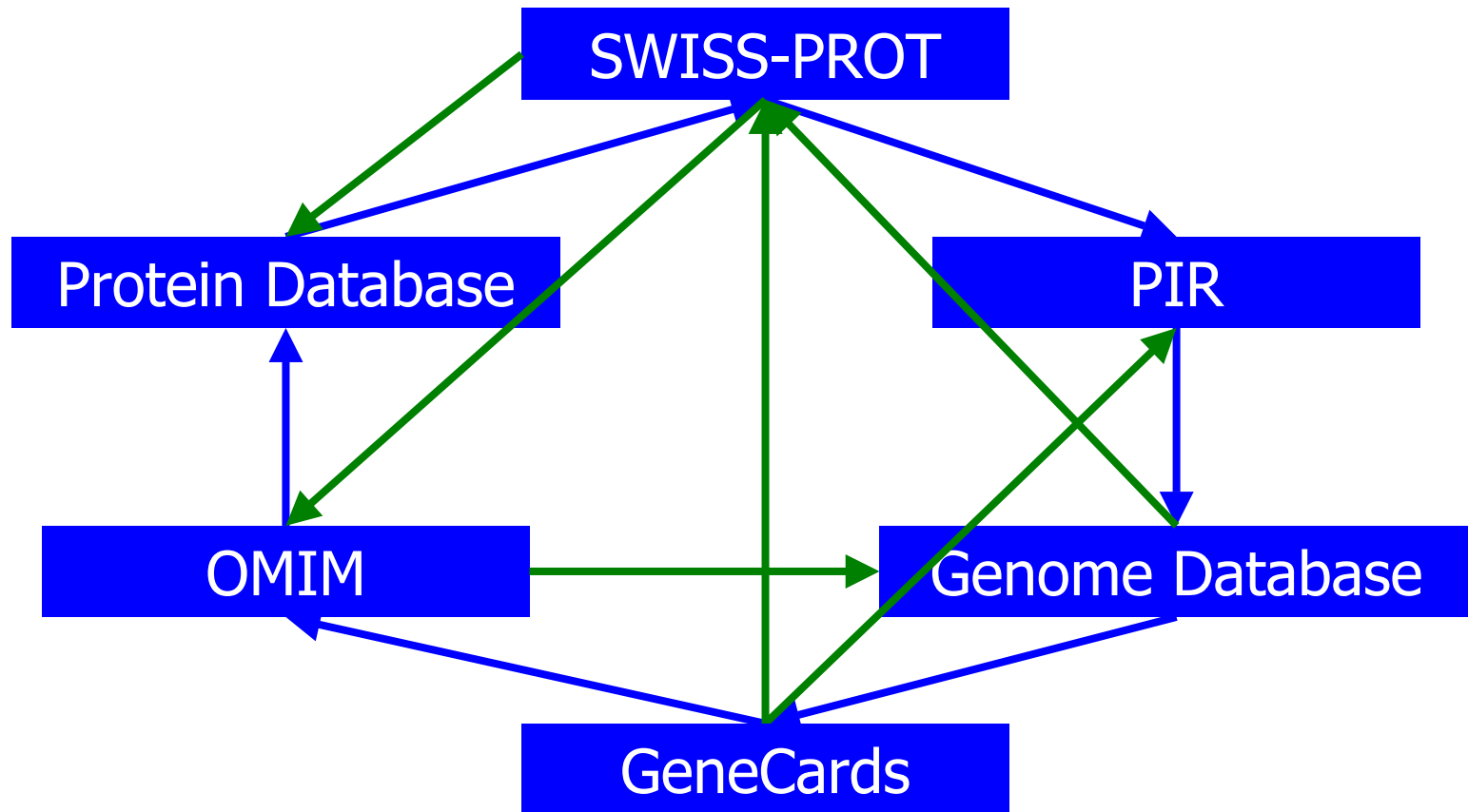
[Structure Factors \(compressed\)](#)

Explore

[SearchLite](#) [SearchFields](#)

- PDB: 1

Linktour



E5: Fokus

- Abdeckung
 - Organismus, Gewebe, Chromosom, Krankheit, Mutation, ...
- Datentyp
 - Sequenzen, Strukturen, Motive, Expressionsintensitäten, ...
- „Breite Datenbanken“
 - Wenig biologische Typen, viele Objekte
 - EMBL, ArrayExpress, GDB, 2D Page, ...
- „Tiefe Datenbanken“
 - Viele biologische Objekte, wenig Objekte
 - MGD, FlyBase, Gencards, ...

E6: Position in „Wissensschöpfungskette“

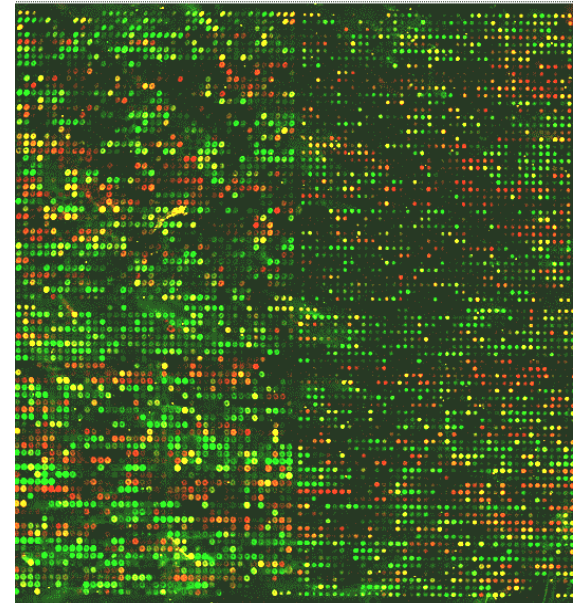
- Primärdatenbanken
 - Nahe am Experiment
 - Wenig Verarbeitung (kurze Annotationspipelines)
 - Genbank/EMBL, PDB, UniGene
- Sekundärdatenbanken
 - Intensive Arbeit zur Datenverbesserung
 - Reichhaltige Annotation und Verlinkung
 - Intensive Integration und Curation über Datentypen hinweg
 - SWISS-PROT, MGD, OMIM, ...
- Tertiärdatenbanken
 - Datenbanken von „Annotationstypen“ oder „Featuretypen“
 - GeneOntology, PFAM, PRINTs, InterPro, CATH,

E7: Zweck

- Lange bestehend, international organisiert
 - Referenzdatenbanken, öffentliches Archiv
 - Genbank, SWISS-PROT, PIR, PDB, ...
- Projektbezogen, One-Shot
 - Existieren ca. bis zur Veröffentlichung
 - Ergebnis einer Datenanalyse
 - Hochaktuell für kurze Zeit
- Labor-Datenbanken
 - Verzahnung mit LIMS (Laboratory Info. Mang. S.)
 - Grundlage der Datenanalyse
 - Speichert Rohdaten (Traces, Bilder, MS-Spektren)

Beispiel: Genexpressionsdaten

- Rohdaten
 - TIFF Images
 - Ca. 50MB pro Bild
- Intensitätsliste
 - Eine Zeile pro Gen
 - Ca. 300 KB
 - Reduktionsfaktor 100
 - Weitere Reduktion durch Matrixdarstellung einfach



Zusammenfassung

- Jede Datenbank ist (etwas) anders
 - Aktiv/passiv, Roh/curated, Redundant, Integriert/verlinkt, Fokus, Primär/abgeleitet, Projektbezogen/langfristig, ...
- **Form follows function**: Anforderungen bestimmen die Eigenschaften
- Eindeutige Einordnungen i.d.R. nicht möglich
- Forschungsgetrieben, forschungsbegleitend, Forschungsgegenstand: **Ständige Veränderung**
- MDB häufig von Nicht-Informatikern entwickelt und betrieben
- Zunehmende Benutzung etablierter Datenbanktechniken