

Kapitel 7: Zugriffsmethoden in Bio-Datenbanken

n Navigation

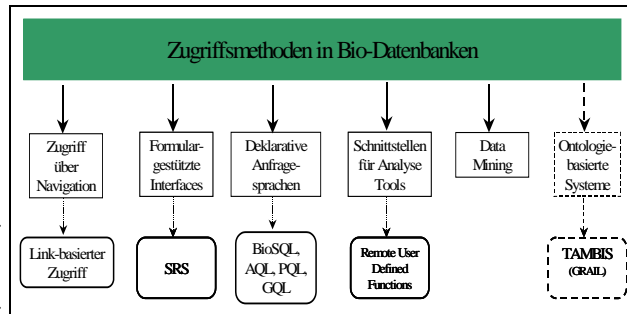
n Stichwortsuche

- Formulargestützte Interfaces
- SRS

n Deklarative Anfragesprachen

n Schnittstellen für Analyse-Tools

n Data Mining



Navigation

n Ansatz

- Browsen in den Datenbeständen; Zugriff auf benötigtes (Zusatz)-Wissen über html-Links

n Vorteile

- Einfach zu realisieren
- Für Standardfälle effizient

n Problematik

- Wenig flexibel, kein Muster-basierter Zugriff
- Verlinkung unterliegt der "Willkür" der DB-Anbieter
- "Lost in Hyperspace"-Phänomen
- Referentielle Integrität schwierig zu wahren

n Wesentliche größere Flexibilität erst durch Bio-Ontologien

- Ontologie: Explizite begriffliche Formalisierung eines Anwendungsbereiches (Fachsprache)
- Dazu mehr in Kapitel 8 (Integration von Bio-Datenbanken)

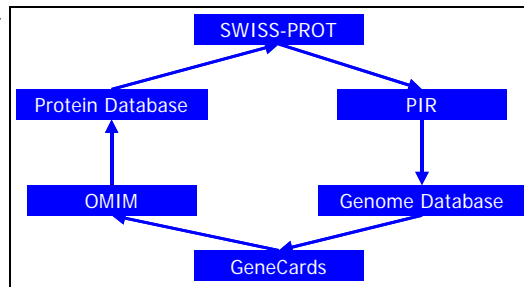


Navigation: Beispiel

n Stichwort: Duchenne Muskeldystrophie

n Startpunkt:

- Swiss-Prot (EBI) mit Stichwort "Duchenne" (DMD_HUMAN; menschliche Duchenne Muskeldystrophie)



Swiss-Prot: Eingabe



Swiss-Prot: Ausgabe (Auszug)

- kurze Beschreibung (mit Lit.-Referenzen etc.)
- Link u.a. zu Datenbank PIR (Protein Information Resource): A27605

General Information
 Entry name: DMD_HUMAN
 Accession number: P11532, Q14169, Q14170
 Created: Feb 12, 1-OCT-1999
 Sequence update: Ref. 12, 1-OCT-1999
 Annotation update: Ref. 42, 15-SEP-2003

Description and origin of the Protein
 Description: Dystrophin.
 Gene name(s): DMD.
 Organism source: Homo sapiens (human).
 Taxonomy: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia; Homo.
 NCBI TaxID: 9606

DISEASE
 [10] Piluso, G., Mirabella, M., Ricci, E., Belsto, A., Abbondanza, G., et al. (2003) J. Biol. Chem. 278: 15581-15586. Position: INTERACTION WITH SNTG1 AND SNTG2. Medline: 20283612. PubMed: 10747910

MISCELLANEOUS
 DEFECTS IN DMD ARE THE CAUSE OF DUCHENNE MUSCULAR DYSTROPHY (DMD). DMD IS A SEX-LINKED RECESSIVE DISORDER. IT IS PROXIMAL MUSCLE WEAKNESS CAUSING FALLS AND DIFFICULTY IN STANDING. UN-AFFECTED FIRST, THEN THE SHOULDER CONFINED TO A WHEELCHAIR BY AGE OF ULTIMATELY OCCUR. ABOUT 50% OF 94 EXPECTATIONS WOULD SUGGEST. THE HEREDITARY AND CLINICAL FEATURES BE DEFECTS IN DMD ARE A CAUSE OF X-LINKED THE DMD GENE IS THE LARGEST KNOWN AND COMPRISES 79 EXONS.

Database Links: Protein Database, PIR (A27605), OMIM, Genome Database.



PIR-Eintrag

- Link u.a. zu Datenbank GDB (Genome Database): GDB:119850

General Information
 ID: A27605
 Accession: A27605, S07710, A27162, S05291, A40134, S06051, S10346, S02243, S02242, S02244, S022109, S23736, S09071, S5186, S68509, S168510, S51175, S51666, S03902.
 Date: 19-Nov-1998 #sequence_revision 27-Jun-1994 #text_change 16-Jun-2000
 Description: dystrophin, muscle - human
 Superfamily: dystrophin, alpha-actinin-actin-binding domain homology; spectrin/dystrophin repeat homology; WW repeat homology;
 Species: Homo sapiens; man;
 Sequence Length: 3685
 Keywords: actin binding; alternative splicing; calmodulin binding; cytoskeleton; leucine zipper; membrane-associated protein; dystrophin; structural protein; tandem repeat; triple helix;
 Comment: Dystrophin is proposed to play a role in anchoring the cytoskeleton to the plasma membrane.
 Alternate Names: Duchenne muscular dystrophy protein
 Map Position: Xp21.2-Xp21.2

Genetics
 Gene: DMD; DMD
 Cross-Reference: GDB: 119850; OMIM: 300370
 Intron: 11/1; 31/3; 62/3; 66/3; 119/3; 177/2; 217/1; 277/3; 320/3; 383/3; 444/2; 494/3; 534/3; 568/3; 604/3; 664/3; 723/2; 764/3; 849/3; 897/1; 1816/3; 1862/3; 1913/3; 1974/3; 2890/1; 3028/3; 3055/1; 3075/2; 3096/1; 3121/1; 3188/2; 3217/1; 3269/3; 3325/2; 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
 >P1:A27605
 MHNREVEEDC VRSVQVQKTF PFMVNAQFS RFSQKREND FSLQDQKAL IQLLQKLVQ
 KLEKREKTR VHALNRYRA LRVGKQVWV UNVQKQVIV DSKRDLGSL ENMLIDWV
 FQKQKDKAG LQKREKSL LKRVQKQVNF TRERDQKAL ALTRERDQ
 FQNRVYVQQK RARQLERAF NIKARVQLQIE KILQREKVDV TYPKRFSLM YTEFLPQVLE
 QQVYKALQCE YMLRFRFRVY TRERFQKRIH QMVGQVQTV ELAQKERTS RERERFQKA

Database Links: Protein Database, PIR, OMIM, Genome Database.



GDB-Eintrag

- Genome Database*
- Aliase, Clone, Loci
- Karten
- Link to GeneCard DMD

General Information
 Entry name: DMD_HUMAN
 Accession number: GDB:119850
 Created: Feb 12, 1-OCT-1999
 Sequence update: Ref. 12, 1-OCT-1999
 Annotation update: Ref. 42, 15-SEP-2003

Description and origin of the Protein
 Description: Dystrophin.
 Gene name(s): DMD.
 Organism source: Homo sapiens (human).
 Taxonomy: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia; Homo.
 NCBI TaxID: 9606

DISEASE
 [10] Piluso, G., Mirabella, M., Ricci, E., Belsto, A., Abbondanza, G., et al. (2003) J. Biol. Chem. 278: 15581-15586. Position: INTERACTION WITH SNTG1 AND SNTG2. Medline: 20283612. PubMed: 10747910

MISCELLANEOUS
 DEFECTS IN DMD ARE THE CAUSE OF DUCHENNE MUSCULAR DYSTROPHY (DMD). DMD IS A SEX-LINKED RECESSIVE DISORDER. IT IS PROXIMAL MUSCLE WEAKNESS CAUSING FALLS AND DIFFICULTY IN STANDING. UN-AFFECTED FIRST, THEN THE SHOULDER CONFINED TO A WHEELCHAIR BY AGE OF ULTIMATELY OCCUR. ABOUT 50% OF 94 EXPECTATIONS WOULD SUGGEST. THE HEREDITARY AND CLINICAL FEATURES BE DEFECTS IN DMD ARE A CAUSE OF X-LINKED THE DMD GENE IS THE LARGEST KNOWN AND COMPRISES 79 EXONS.

Database Links: Protein Database, PIR, OMIM, Genome Database.



Gene-Card DMD

- SNPs, Mutationen, ...
- Link u.a. zu OMIM ID: 300377

GeneCard for gene DMD
 GCOXM029822 Approved UCL/HGN/CUGO Human Gene Nomenclature database symbol
 DMD (dystrophin (muscular dystrophy, Duchenne and Becker types))

Aliases and Additional Descriptions
 (According to GDB, HUGO, LocusLink, SWISS-PROT, and/or GeneLog)
 • BMD
 • DXS142
 • DXS164
 • DXS206
 • DXS230
 • DXS239
 • DXS268
 • DXS269
 • DXS270
 • DXS272
 • dystrophin (muscular dystrophy, Duchenne and Becker types)
 • dystrophin (muscular dystrophy, Duchenne and Becker types), includes DXS142, DXS164, DXS206, DXS230, DXS239, DXS268, DXS269, and Dystrophin.

Chromosomal Location
 (According to GeneLoc and/or HUGO, and/or LocusLink/NCBI build 31)
 Chromosome: X
 LocusLink cytogenetic band: xp21.2
 Ensembl cytogenetic band: Xp21.2
 Gene in genomic location: bands according to Ensembl, locations according to GeneLoc (and/or LocusLink and/or Ensembl if different)
 Start: 29,822,399 bp from pter
 End: 32,042,786 bp from pter
 Size: 2,220,387 bases
 Orientation: minus strand

Genomic View
 Locus Golden Path
 UCSC Golden Path with GeneCards custom track

Disorders & Mutations
 search databases for MIM named disorders
 • Duchenne muscular dystrophy
 • Becker muscular dystrophy
 • SWISS-PROT: DMD_HUMAN
 • Disease: Defects in DMD are the cause of Duchenne muscular dystrophy (DMD) and Becker muscular dystrophy (BMD).

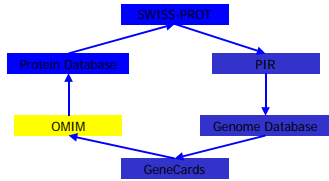
Database Links: Protein Database, PIR, OMIM, Genome Database.



OMIM

n OMIM
300377
DMD

- Beschreibung
- Klinischer Verlauf
- Molekulare Grundlagen
- Populationsgenetik
- Mutationen und Häufigkeiten
- Links u.a. zu PDB



PDB

PDB
PROTEIN DATA BANK

Structure Explorer - 1DXX



Stichwortsuche / Suchformulare

n Typische Zugriffsmöglichkeit im Web (Google, Altavista, Internet-Shopping etc.)
- Einfach, schnell, verständlich, bekannt

n Vorstrukturierte Suchformulare

n Verwendung von Methoden des IR

- Ranking der Ergebnisse
- Operatoren: AND, OR, NOT, + / -

n Probleme

- Ergebnis nicht zwingend Treffer
- Wortformen: Zeiten, Sing. / Plural, Casus, ...
- Synonym / Homonymprobleme
- Treffer sind Dokumente, nicht Attribute

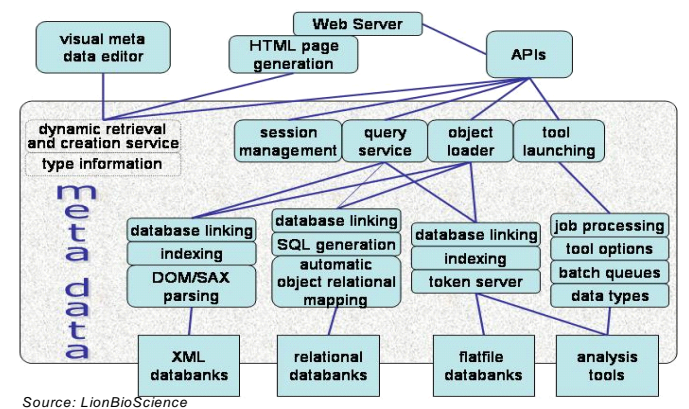
n Nachteil: Starke Einschränkung der Expressivität, keine Unterstützung vom komplexen Anfragen



SRS (Sequence Retrieval System)

n Ursprünglich als Zugriffstool für Sequenzdatenbank EMBL entwickelt

- Mehr als 400 wissenschaftliche Datenbanken ansprechbar
- Einheitliche graphische Nutzerschnittstelle
- Formularbasiert
- Fokus auf Flatfile-Datenbanken, Weiterentwicklung zur Unterstützung relationaler Datenquellen
- Übernahme und kommerzieller Vertrieb durch LION BIOSCIENCE*



Source: LionBioScience



SRS: Konzept (1)

- Ein "Konzept" wird über eine Anzahl von Feldern (Entries, Attributen) beschrieben
- Beispiel: Konzept *Gen*

The screenshot shows the SRS search interface. At the top, there are navigation tabs: TOP PAGE, QUERY, RESULTS, SESSIONS, VIEWS, DATABANKS, and HELP. Below this is a search bar with a 'Reset' button and a dropdown menu for 'Info' (set to 'about field'). The search criteria are 'AllText' and 'homeobox'. On the left, there are options to 'append wildcards to words' (checked), 'combine searches with AND', and 'Number of entries to display per page' (set to 30). There is also an 'Extended query form' button. A list of fields to search in is shown, including AllText, ID, AccNumber, Description, GeneName, Keywords, Date, Organism, Organelle, SeqLength, ProteinID, and others. A 'retrieve entries of type' dropdown is set to 'Entry', and a 'SeqSimpleView' dropdown is visible. A 'to display' list and a 'sequence format' dropdown (set to 'swiss') are also present.

SRS: Konzept (2)

- Ein Konzept kann als html-Seite visualisiert werden

The screenshot shows the SRS entry page for SWISSPROT:CD22_HUMAN. The page is titled 'Text Entry | SwissEntry' and has a 'Reset' button. It displays 'General Information about the Entry' with fields for Entry name (SWISSPROT:CD22_HUMAN), Prim. accession # (P20273), Sec. accession # (Q01665, Q92872, Q95699, Q95701, Q95702, Q95703), Created (Release 17, 1-FEB-1991), Last sequence update (Release 38, 15-JUL-1999), and Last annotation update (Release 38, 15-JUL-1999). Below this is 'Description and Origin of the Protein' with Keywords (Glycoprotein, Cell adhesion, Transmembrane, Signal, B-cell, Immunoglobulin domain, Alternative splicing, Phosphorylation, Polymorphism), Description (b-cell receptor cd22 precursor (Iu-14) (b-lymphocyte cell adhesion molecule) (bi-cam)), Gene name(s) (cd22), and Organism source (homo sapiens (human)). At the bottom, there is a protein structure visualization with the label 'VARSPIC: 241 417, MISSING (IN CD22-ALPHA)'.

SRS: Konzept (3)

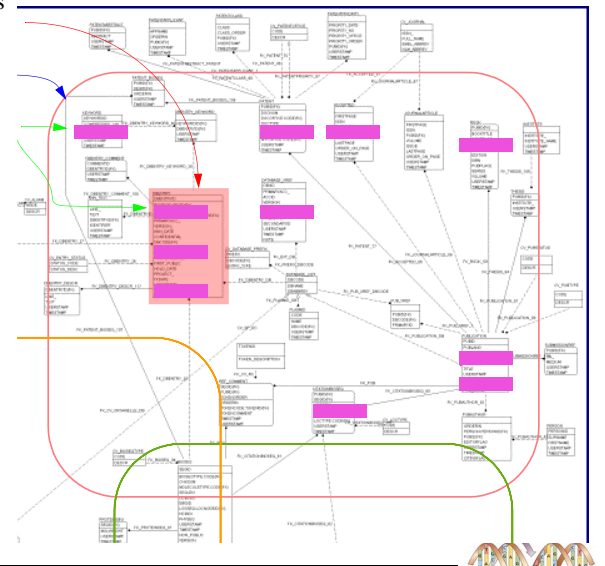
- Export als SRS-"Objekt" (C++, Java, Perl, CORBA) oder Speicherung als XML

The screenshot shows the SRS interface with a table of 'Data-fields in SRS' and their corresponding XML export code. The table has columns for Name, Short Name, Type, No of Keys, and No of Refers. The XML code is shown on the right, with red circles highlighting specific fields and their corresponding XML tags. The XML code includes tags for <name>, <abstract>, <db_xref dbkey>, <db_xref dbkey>, <db_xref dbkey>, <parent_list>, <contains>, <member_list>, <db_xref dbkey>, <member_list>, <sec_ac acc>, and </interpro>.

Name	Short Name	Type	No of Keys	No of Refers
AllText	all	group	0	
Accession	acc	id	3915	6
SecAcc	oac	index	3915	0
Short Name	snm	index	3911	0
Full Name	fnm	index	7726	0
Type	ty	index	4	0
GC-terms	gc	index	2049	0
Abstracts	abs	index	123444	0
ProteinName	pnm	index	35396	0
ProteinRef	prf	index	333	0
ChildRef	chr	index	913	0
ContainsRef	has	index	127	0
FoundInRef	in	index	207	0
Taxon	taxon	index	2686	0
PubId	pubid	index	1064	0
Authors	aut	index	15626	0
Title	tit	index	34060	0
BookTitle	bkttl	index	102	0
Journal	jnl	index	338	0
VolumeNo	vol	num	500	0
FirstPage	fp	num	536	0
LastPage	lp	num	269	0
Year	yr	num	0	0
URL	url	index	0	0
MedlineID	mid	index	4776	0
DbName	dbn	index	11	0
Dclass	dr	index	9449	0
ref_Name	nm	index	5888	0

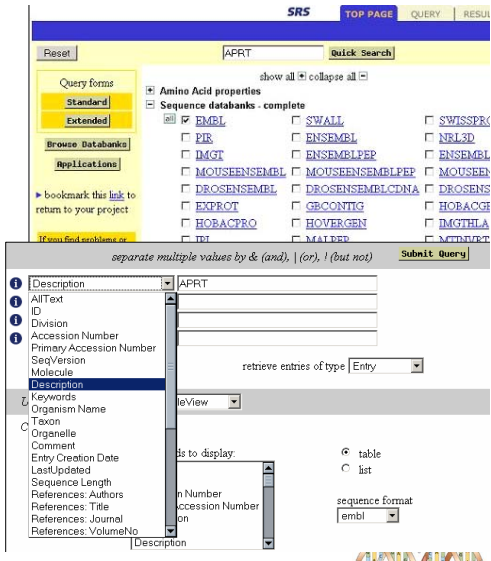
SRS: Zugriff auf relationale Datenbanken

- Auswahl einer *hub*-Tabelle (als "Aufhänger"-Konzept) (→)
- Angabe der Tabellen die zum Konzept "dazu gehören" (→)
- Angabe der abfragbaren Attribute (→)
- Interne Umsetzung via Views/Joins



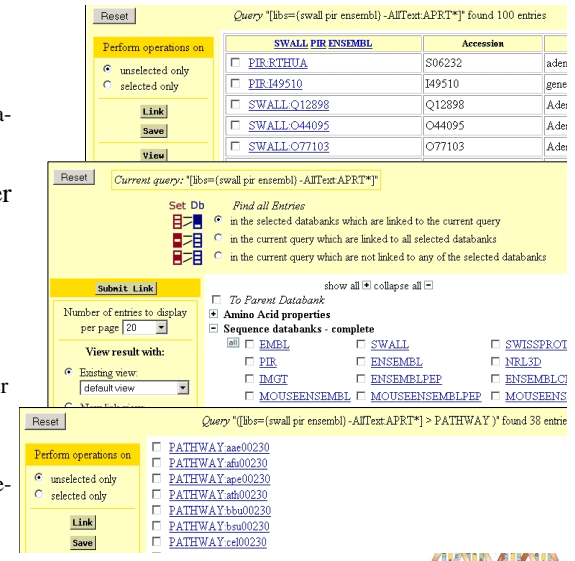
SRS: Datenzugriff (1)

- Objektsuche
 - Auswahl von Datenquellen
 - Spezifikation der Suchkriterien für abfragbaren Attribute
 - Schnittmenge der Attribute aller ausgewählten Quellen (z.B. ID)
- Unterstützte Query-Möglichkeiten
 - Textsuche, Bereichsuche für numerische/ Datumattribute
 - Regulär Ausdrücke
- Automatische Übersetzung von SRS-Queries nach SQL zum Zugriff auf relationale Datenbanken
- Ergebnis als Vereinigung der Suchergebnisse über einzelne Quellen



SRS - Datenzugriff (2)

- Querverweis-Suche
 - Für Objekte einer Ergebnismenge oder für eine ganze Quelle
 - Referenzierte Objekte in anderen Datenquellen
- Automatische Bestimmung der Pfade zwischen Quellen
 - Shortest-Path-Algorithmus
 - Gleiche Semantik der Beziehungen
- Tool-Integration
 - Anwendung auf Ergebnismengen der Abfragen
 - Anzeige direkt im Web-Browser
 - Große Anzahl von Tools bereits integriert



SRS Anfrage Sprache

- Anfragen können an vielen Stellen direkt formuliert werden, z.B. Results -> Search using a query expression
- Anfrage-Syntax
 - Stringsuche: [Menge-Mengenattribut:Suchmuster], Menge kann eine Datenbank, DB-Gruppe, Index, Index-Gruppe oder Suchausdruck sein
 - z.B. [pir-des:elastase], oder [swissprot-AllText:duchenne*]
 - Wildcards: [swissprot-aut:sanger,f*!coulson,a*]
 - Reguläre Ausdrücke: [swissprot-aut:/mue?l|er/]
 - Zahlenbereiche: [swissprot-SeqLength#400:500]
 - mehrere DBs: [{swissprot swissnew sptrembl}-des:kinase], [dbs={swissprot swissnew sptrembl}-des:kinase] &[dbs-org:human]

SRS Anfrage Sprache

- Einfache Anfragen kombinieren:
 - operand operator operand ...
 - z.B. verlinken: [swissprot-AllText:duchenne*] > pdb
- Operatoren
 - logisch: | oder, & und, ! aber nicht
 - Links:
 - > liefere die linken Einträge, < liefere die rechten Einträge
 - Komplexe Verweise: (q = [{swissprot swissnew}-des:kinase])!(q<swissnew)
 - Hierarchische Suche
 - >^ liefere Teilbaum-Einträge definiert durch linke Seite
 - >_ liefere Blatt-Einträge des Teilbaums definiert durch linke Seite

SRS Anfrage Sprache

- Mehrfache Links
 - [swissprot-AllText:duchenne*] >omim
OMIM-Einträge zu denen man direkt von SwissProt-Einträgen mit Begriff “duchenne” gelangt
 - [swissprot-AllText:duchenne*] >pdb >omim
wie oben, doch SwissProt und OMIM müssen über PDB verlinkt sein
 - [swissprot-id:acha_human] > prosite > swissprot
Suche nach Eintrag “acha_human”, Link nach Prosite (Protein Fam.) “neuronal acetylcholine receptors”, Link zu Swissprot Einträgen => Sequenzen aus einer Proteinfamilie
 - [swissprot-id:gshr_caee] > prodom > pdb
kein direkt Link von „gshr_caee” zur PDB, aber über ProDom (Protein Domänen), Ergebnis homologe Proteine zu „gshr_caee”



© Dr. A. Hinneburg, Prof. R. Müller, Prof. E. Rahm

SRS Anfrage Sprache

- Einträge und Untereinträge
 - [swissprot-keywords:transmembrane] alle Einträge mit Keyword Transmembr.
 - [swissprot-ftkey:transmem] Menge von Untereinträgen mit Typ transmem
- Einträge und Untereinträge können über Links kombiniert werden
 - [swissprot-org:human] > [swissprot-ftkey:transmem]
Menge der Transmembransegmente in menschlichen Proteinen
 - [swissprot-org:human] < [swissprot-ftkey:transmem]
Menge der menschlichen Proteine mit Transmembransegmenten
 - [swissprot-ftkey:transmem] > parent
Konvertierung der Untereinträge zu den zugehörigen Einträgen
 - [swissprot-ftkey:transmem] > parent | [swissprot-key:transmembrane]
Einträge mit Transmembransegmenten oder mit dem Schlüsselwort “transmembrane”



© Dr. A. Hinneburg, Prof. R. Müller, Prof. E. Rahm

SRS Anpassung

- Komplizierte Anfragen können vordefiniert werden
- Perl-ähnlich Sprache ICARUS
- Beispiel: Suche einen Swissprot-Eintrag nach Zugriffsnummer oder Beschreibung

```
$CannedQuery:[sampleQuery
prompt:|Sample canned query
options:{
  $AppOpt:[ac prompt:'Access number' defStr:'Q1']
  $AppOpt:[des prompt:'Description' defStr:'cancer']
}
queryStr:
@"|[swissprot-des:(Sdes)*]|/[swissprot-acc:(Sac)*]"
]
```
- Das @ zeigt, das die Anfrage zur Laufzeit bearbeitet wird.
- Datei mit Anfragedefinition wird ähnlich wie ein HTML Link in SRS-Seite eingebunden.



© Dr. A. Hinneburg, Prof. R. Müller, Prof. E. Rahm

Aufruf von SRS aus HTML Seiten

- Wegtz kann als CGI Programm direkt als HTML Link aufgerufen werden
- Einfaches Anzeigen von Einträgen
 - wgetz?-e+[embl-id:melas]
 - wgetz?-e+[{embl%20emblnew}-acc:X012345]
 - Die Option -e lässt wgetz volle Einträge anzeigen
- Anzeigen von Mengen
 - wgetz?[embl-all:elastase]
 - wgetz?swissprot
 - wgetz?swissprot+-lv+30+-bv+31 (-lv Anzahl Einträge pro Seite, -bv erster Eintrag)
 - wgetz?swissprot+-lv+30+-bv+31+-view+SequenceSimple
zeigt ZusatzInfos mit Hilfe eines vordef. Views
 - wgetz?swissprot+-lv+30+-bv+31+-view+SequenceSimple+-ascii+-rs+||+-cs+@@
Wie oben aber als HTML Tabelle



© Dr. A. Hinneburg, Prof. R. Müller, Prof. E. Rahm

Kapitel 7 (Forts.): Schnittstellen für Analyse-Tools

- n Vielzahl an Analyseprogrammen für Verarbeitung von Bio-Daten, z.B.
 - CLUSTALX (Graphisches Tool für ClustalW multiple sequence alignment program)
 - FASTA
 - BLAST (Basic Local Alignment Search Tool)
 - Sacc3D (Structural Information for Yeast Proteins)
- n Installation auf lokalem Rechner vs. Remote-Zugang über Web-Interface
- n Installation auf lokalem Rechner: Effizient, aber Update-Problematik
- n Zugang über Web-Interface
 - Via Formulare und ftp/e-mail: Einfach zu realisieren, aber oft umständlich und wenig flexibel
 - Via API von Remote User Defined Functions (RUDFs): Flexibler, aber komplexer
 - Realisierung von RUDFs als Internet Functions



Eigenschaften von Internet Functions (IFs)

- n Kein Teil der Datenbank
- n Werden von einem externen System gestartet
- n Erreichbar über das Internet
- n Beinhalten Kommunikationsprotokoll zum Datenaustausch mit der Datenbank
- n Verschachtelungen mehrerer IFs beim Aufruf möglich



Beispiel-Szenario

Database: Program:

Enter an accession, gi, or a sequence in FASTA format:

```
>Ab000001
1 aattt...aatg aagagtgttg ttgttagctgg cccattaatt taggcatgtg
cacaccttc
61 tcttttccc catacacacc tgtgaacttg tgagacagat ggggaatta
tttattgttt
121 ttttttttaa tataaagatg ataagtcatg gaacctctct gtctactcaa
```

Ablauf einer Berechnung des Sequenz-Vergleichs mittels online zugänglichen Analyse Tools

The request ID is

or

The results are estimated to be ready in 4 seconds but may be done sooner.

Eingabe der Sequenz in das Formular

Bearbeitungszeit...

request ID für die Einsendung wird zurückgegeben

Ergebnisse der Berechnung (e-mail oder online)

```
>ref|NT_025938.2|Hs22_26094 Homo sapiens chromosome 22
Length = 65461

Score = 50.1
bits (25), Expect = 0.002
Identities = 40/45 (88%)
Strand = Plus / Minus

Query: 261 tcgatgaagaacgcagcgaatgcgataagtaagtgtgaattgcag 305
          ||| | | | | | | | | | | | | | | | | | | | | | | | | | |
Sbjct: 16868 tcgatgaagaacgcagctagctgcgagaatataatgtgaattgcag 16824
```



Internet Function Definition Language (IFDL)

- n Erweiterung der SQL DDL
 - *Kein* Teil des SQL-Standards
 - Nicht verwechseln mit der *Independent Form Definition Language* (Abk. auch IFDL)
- n Definition für die Einbindung von IFs in SQL-Queries
- n IF-Definition: 5-Tupel
 - Funktionsname der aufzurufenden IF
 - URL der Funktion
 - Liste von Eingabeparametern mit Typen
 - Typ des Rückgabewertes
 - HTQL*-Wert

* Hyper Text Query Language



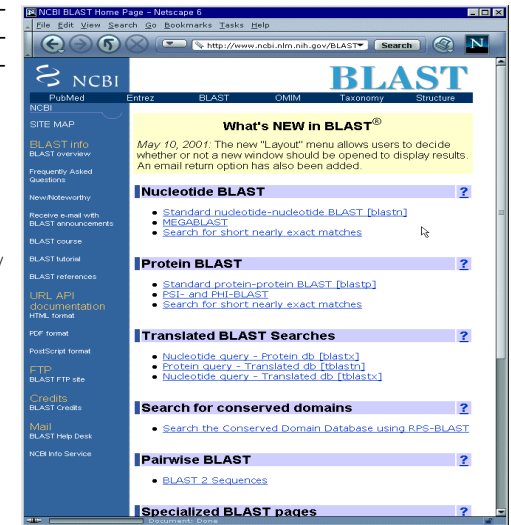
Hyper Text Query Language (HTQL)

- n Unterstützt Suche in Web-Dokumenten
- n Nutzung der Tags in HTML/XML-Dokumenten zur Navigation und Ausgabestrukturierung (ähnlich wie bei XPath, XQuery)
- n Dient im Zusammenhang mit IFs als Schnittstelle zw. DB und IFs
 - HTQL-Wert in IFDL Function Definition informiert über Position des Rückgabewertes im (generierten) Web-Rückgabedokument
 - IF extrahiert für weitere Verarbeitung Rückgabewert aus dem (generierten) Web-Dokument



Beispiel für IF-Ausführung (GenBank)

- n Vergleich von neuen, noch uncharakterisierten Gensequenzen (in Tabelle *local*) mit bekannten Sequenzen der Kenianischen Fruchtfliege
 - Nur solche Sequenzen von Interesse, die Ähnlichkeit von mind. 98 % zu geg. Sequenz aufweisen
 - Für Berechnung der Editierdistanz Verwendung des über Internet Schnittstelle zugänglichen Programms BLAST (<http://ncbi.nlm.nih.gov/blast/blast.cgi>)



Beispiel für IF-Ausführung (2)

- n Verwendung von zwei IFs: *blast*, *get_seq*
 - *blast* verwendet als Eingabe eine Sequenz der Tabelle *local* und liefert "requestID"-Wert (rein "technischer" Parameter für weitere Referenz auf Anfrage, vor allem wg. oft langer Laufzeit von *blast*)
 - Ergebnis von *blast* befindet sich nach Ausführung in von *blast.cgi* erzeugter HTML-Datei im Eingabefeld des 1. form-Tags
 - Dieser Wert wird extrahiert und dient als Parameter für *get_seq*, dessen Ergebnis nach dem 2. pre-Tag der erzeugten HTML-Datei steht
 - Ergebnis von *get_seq* ist Tabelle, welche die ermittelten Ähnlichkeitsscores, Herkunft und Art enthält

```
define function blast
href "http://www.ncbi.nlm.nih.gov/blast/Blast.cgi"
parameters query varchar(10000)
results request_id varchar(40)
htql value: <form>.<input>;
```

```
define function get_seq
href "http://www.ncbi.nlm.nih.gov/blast/Blast.cgi"
parameters rid varchar(40)
results sequence varchar(10000)
htql value: <pre>.<pre>;
```



Beispiel für IF-Ausführung (3)

- n SQL Query mit IF-Aufrufen

```
SELECT b.sequence
FROM (SELECT get_seq( blast( a.sequence ))
      FROM local as a) as b
WHERE b.organism = "Drosophila" AND b.source(country) = "Kenia"
      AND b.e-value >= 0.98
```
- n Teile der Anfrage
 - Führe für alle Sequenzen der DB-Tabelle *local* (hier *a*) die BLAST-Suche in GenBank durch
 - Speichere die Ergebnisse in die Tabelle *b* ab und gib alle Sequenzen aus, die zur Kenianischen Fruchtfliege gehören und einen Ähnlichkeits-Score von mehr als 0.98 haben



Anwendung: LifeDB

n Biologisches Datenbanksystem (Prototyp der Mississippi State University)

n Unterstützt das Konzept der IFs

n Dreiteiliges Interface

- Textfeld links oben für Eingabe der IFs und SQL-Anfragen
- Query-Ergebnisseite rechts
- Feld links unten zur Auflistung bereits getätigter Anfragen

The screenshot shows a web browser window with the following content:

- Query Input:** A text area containing the SQL query: `select sts_code, GenbankID from genes where sts_GenbankID(sts_code) is not null`. Below it are buttons for "do" and "back".
- Results:** A table with columns "sts_code" and "GenbankID". It displays 5 records:

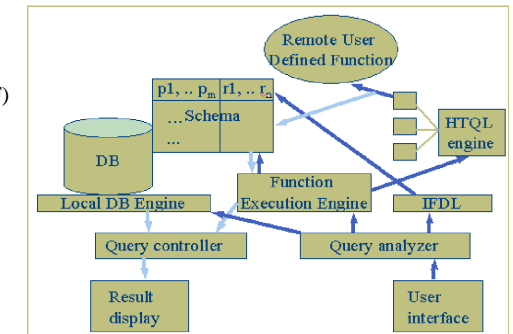
sts_code	GenbankID
WT-5270	G04845
WT-5275	G04849
WT-5276	G04850
WT-5278	G04851
WT-5279	G04852
- Queries:** A list of queries, including "345452 sts_code, GenbankID" and "1 queries. refresh".
- Footer:** "5 records found. refresh" and "Return Home" link.

(C) Prof. R. Müller, Prof. E. Rahm

LifeDB Anfrageverarbeitung: Architektur

n Query Analyzer

- Parsen von Eingaben im Query-Eingabefeld (stand-alone IF-Aufrufe sowie SQL-Anfragen mit oder ohne eingebundene IF)
- Weiterleitung von SQL-Anfragen ohne eingebundene IF direkt an die lokale DB Engine
- Weiterleitung von IF-Aufrufen an IFDL Modul



n IFDL Modul

- Weiterleitung von eingebetteten IF an Function Execution Engine
- Verfügt über Metadaten-Tabelle in lokaler Datenbank mit Strukturinformationen über die definierten IF (diese wird von Function Execution Engine benötigt, um IF über HTQL-Engine einzubinden und auszuführen)

```
create table EDLUDFS (
  function_name varchar(30),
  store_table_name varchar(30),
  address varchar(100),
  para_num number(2),
  para_name_list varchar(200),
  result_name_list varchar(200),
  result_html varchar(200)
```

IF-Metadaten-Tabelle

(C) Prof. R. Müller, Prof. E. Rahm

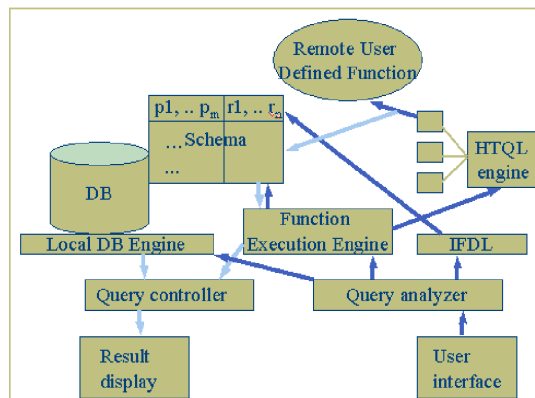
LifeDB Anfrageverarbeit.: Architektur (2)

n HTQL Engine

- Nutzt http zur Übermittlung der Eingabedaten an IF
- IF wird extern ausgeführt und legt Ergebnis in HTML-Seite ab
- HTQL Engine extrahiert Ergebnis anhand des HTQL-Wertes der IF-Definition
- Ergebnis wird in einer weiteren Tabelle in lokaler Datenbank abgelegt

n Query Controller führt SQL-Anfrage aus und verwendet dabei Ergebnistabelle aus der Ausführung der IF

n Schließlich wird Ergebnis dem Nutzer präsentiert



(C) Prof. R. Müller, Prof. E. Rahm

RUDF/IF: Zusammenfassung

n API / CLI Ansatz zur Einbindung externer Funktionalität bei der Auswertung von Bio-Daten

n Einbindung von RUDF / IFs innerhalb von SQL-Queries

n Rückgabe der Ergebnisse in html-Format mit Extraktionsinformation (wo ist gewünschte Information im Rückgabe-Dokument?)

n Vorteile: Umgehung lokaler Installation von externer Funktionalität

n Nachteile: Proprietäre Formate und Ausführungsarchitektur

n Weiter Informationen unter:

- <http://www.cse.msstate.edu/~cly/EDI/index.html>

- Chen, L.; Jamil, H. M.: Supporting Remote User Defined Functions in Heterogeneous Biological Databases. Proceedings IEEE International Conference on BIBE 2001: 144-152, 2001.

- Chen, L.; Jamil, H. M.: On Using Remote User Defined Functions as Wrappers for Biological Database Interoperability; International Journal of Cooperative Information Systems (IJCIS), Special Issue on Data Management and Modeling Support in Bioinformatics, 12(2):161-195, 2003.

(C) Prof. R. Müller, Prof. E. Rahm