

# Kapitel 7: Zugriffsmethoden in Bio-Datenbanken

n Navigation

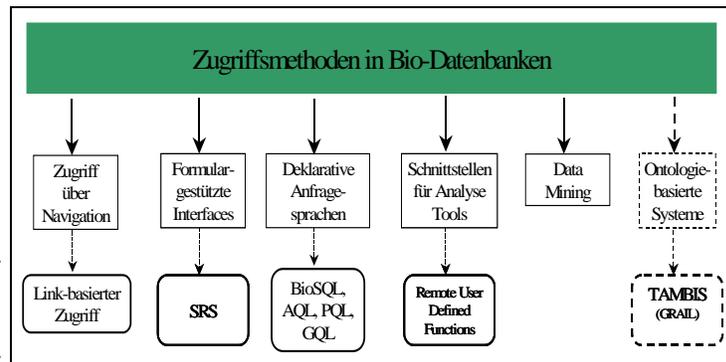
n Stichwortsuche

- 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)

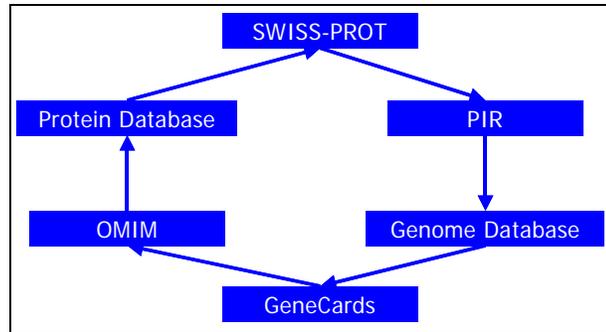


# Navigation: Beispiel

n Stichwort: Duchenne Muskeldystrophie

n Startpunkt:

- Swiss-Prot (EBI) mit Stichwort "Duchenne" (DMD\_HUMAN; menschliche Duchenne Muskeldystrophie)



# Swiss-Prot: Eingabe

 [Quick Search](#) [Library Page](#) [Query Form](#) [Tools](#) [Results](#) [Projects](#) [Views](#) [Databanks](#) [HELP](#)

[Reset](#)  [Quick Search](#)

### Search Options

- Select the **databanks** you want to search
- Enter your **search terms** in the **Quick Search** box, or choose a **query form** from below  
[Standard Query Form](#)  
[Extended Query Form](#)

You can **browse** through all the **entries** in any **databanks**. First, **select** the **databanks** you want to browse, then click:  
[Browse Entries](#)

### Available Databanks

[Expand all](#)  [Collapse all](#) Show databanks tooltips:

- Literature, Bibliography and Reference Databases**
  - [MEDLINE](#)  [MEDLINE \(Main Release\)](#)  [MEDLINE \(Updates\)](#)  [MIM](#)
  - [TAXONOMY](#)  [GENETICCODE](#)
- Nucleotide sequence databases**
  - [EMBL](#)  [EMBL \(Release\)](#)  [EMBL \(Updates\)](#)  [EMBL \(WGS\)](#)
  - [EMBL \(TPA\)](#)  [EMBL \(Contig\)](#)  [REFSEQ](#)  [ENSEMBL HUMAN](#)
  - [ENSEMBL MOUSE](#)  [ENSEMBL FLY](#)  [ENSEMBL FISH](#)  [INGTHLA](#)
  - [INGT/LIGH-DB](#)  [PATENT\\_DNA](#)
- Protein sequence databases**
  - [SWALL \(SPTR\)](#)  [Swiss-Prot](#)  [SpTrEMBL](#)  [TrEMBL \(Updates\)](#)
  - [IPI](#)  [RemTrEMBL](#)  [PIR](#)  [REFSEQP](#)
  - [PATENT\\_PRT](#)  [JPO\\_PRT](#)  [USPTO Proteins](#)  [MHCBN](#)
  - [SWISSCHANGE](#)
- Nucleotide related databases**
- Protein function databases**
- Protein structure databases**
- Enzymes, reactions and metabolic pathway databases**
- Mutation and SNP databases**
- Gene ontology resources**
- Mapping databases**
- Other databases**
- User owned databases**
- Application result databases**
- EMBOSS result databases**

### Tips

▶ bookmark this [link](#) to return to your project



# Swiss-Prot: Ausgabe (Auszug)

n Kurze Beschreibung (mit Lit.-Referenzen etc.)

n Link u.a. zu Datenbank PIR (Protein Information Resource): A27605

General information	
Entry name	DMD_HUMAN
Accession number	<b>P11532</b> , Q14169, Q14170
Created	Rel. 12, 1-OCT-1989
Sequence update	Rel. 12, 1-OCT-1989
Annotation update	Rel. 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; Eumammalia; Primates; Hominoidea; Hominidae; Homo.
NCBI TaxID	9606
...	
[10]	<p> <a href="#">Piluso,G., Mirabella,M., Rico,E., Belisio,A., Abbondanz Gammal1- and gammal2-syntrophins, two novel (2000) J. Biol. Chem. 275: 15851-15860</a>            Position INTERACTION WITH SNTG1 AND SNTG2.            Medline <a href="#">20283612</a>            PubMed <a href="#">10747910</a> </p>
...	
DISEASE	<p>           DEFECTS IN DMD ARE THE CAUSE OF DUCHENNE MUSCULAR DYSTROPHY (BMD). DMD IS A SEX-LINKED RECESSIVE DISORDER. IT IS CHARACTERIZED BY PROXIMAL MUSCLE WEAKNESS CAUSING WALKING DIFFICULTY, FALLS, AND DIFFICULTY IN STANDING UP. THE DISEASE IS AFFECTED FIRST, THEN THE SHOULDER AND UPPER LIMBS. CHILDREN CONFINED TO A WHEELCHAIR BY AGE OF 10 YEARS. ULTIMATELY OCCUR. ABOUT 50% OF PATIENTS WHOSE EXPECTATIONS WOULD SUGGEST. THE DISEASE IS A HEREDITARY AND CLINICAL FEATURES BY...         </p>
DISEASE	<p>           DEFECTS IN DMD ARE A CAUSE OF X-LINKED...         </p>
MISCELLANEOUS	<p>           THE DMD GENE IS THE LARGEST KNOWN GENE IN HUMANS. IT IS 2.4 MILLION BASE-PAIRS IN SIZE AND COMPRISES 79 EXONS.         </p>

EMBL	<a href="#">U006179,CAA29545</a> 1,ALT_SEQ.
	<a href="#">U006178,CAA29544</a> 1,-
	<a href="#">X14298,CAA32479</a> 1,ALT_SEQ.
	<a href="#">X15495,CAA33518</a> 1,-
	<a href="#">X54820,CAA38589</a> 1,-
	<a href="#">X13045,CAA31451</a> 1,-
	<a href="#">X13046,CAA31452</a> 1,-
	<a href="#">X13047,CAA31453</a> 1,-
	<a href="#">X13048,CAA31454</a> 1,-
	<a href="#">U27203,AAA86115</a> 1,-
	<a href="#">U27203,AAA86116</a> 1,-
	<a href="#">X15148,CAA33545</a> 1,-
PIR	<a href="#">A27605,A27605</a> .
	<a href="#">A27162,A27162</a> .
	<a href="#">S05291,S05291</a> .
ISSP	<a href="#">P46939,IQAG</a> .
Gene	<a href="#">HGNC:2928,DMD</a> .
	<a href="#">300377</a> ,-
	<a href="#">310200</a> ,-
MIM	<a href="#">300376</a> ,-



# PIR-Eintrag

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

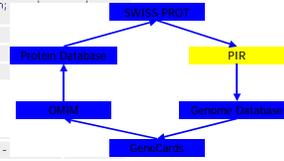
General Information	
ID	A27605
Accession	A27605; S07710; A27162; S05291; A40134; S06051; S10346; S02243; S02242; S02244; S02109; S23736; S09071; I54186; I68509; I68510; I54175; I54166; S03902;
Date	19-Nov-1988 #sequence_revision 27-Jun-1994 #text_change 16-Jun-2000
Description	dystrophin, muscle - human
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; dystrophy; 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

...

Region	leucine zipper motif	3572 -
--------	----------------------	--------

Genetics	
Gene	GDB:DMD
Cross References	GDB:119850; OMIM:310200
Intron	11/1; 31/3; 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; 723/2; 764/3; 774/3; 815/1; 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	
MLMWEVEVC YEREDVQKKT PFGWNAQPS RFGKQIENE FSLDQERAL LDDLEGLGQ	
KLPERKQSPR VHALNNVNA LAVLQNNVD LQNIQSTDIY DQHRRLLEGL IHWILLRQV	
EDWENIDAG LQQNREKIL LSWRQSTEN PFGWYIHTP FWRDGLAN ALRHRRL	
FDWNSVVOQ RATORLEHAF NIARVQLGIE KLDDREVDY TYEDKRSILM VITSLFQVLE	
QQVIEAIQE VMKLEPFKVV FKEHPQLHH QHYSQQTIV SLAQSVETS SPKRPFKYA	



# GDB-Eintrag

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

The screenshot displays the GDB interface for the DMD gene. The main window shows a genomic map with various features labeled. On the right, a diagram illustrates the database connections: SWISS-PROT, Protein Database, PIR, QMAM, Genome Database, and GeneCards.

\* <http://gdbwww.dkfz-heidelberg.de/>

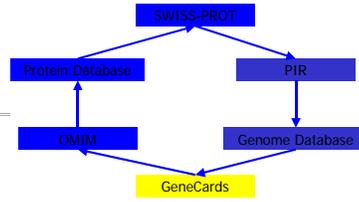


# Gene-Card DMD

n SNPs,  
Mutation-  
nen, ...

n Link u.a.  
zu OMIM  
ID:  
300377

<b>GeneCard for gene <i>DMD</i></b> <b>GC0XM029822</b>		Approved UCL/HGNC/HUGO Human Gene Nomenclature database symbol <b>DMD (dystrophin (muscular dystrophy, Duchenne and Becker types))</b>
<b>Aliases and Additional Descriptions</b> <small>(According to GDB, HUGO, LocusLink, SWISS-PROT, and/or GeneLoc)</small>	<ul style="list-style-type: none"> <li>• BMD</li> <li>• DXS142</li> <li>• DXS164</li> <li>• DXS206</li> <li>• DXS230</li> <li>• DXS239</li> <li>• DXS268</li> <li>• DXS269</li> <li>• DXS270</li> <li>• DXS272</li> <li>• dystrophin (muscular dystrophy, Duchenne and Becker types)</li> <li>• dystrophin (muscular dystrophy, Duchenne and Becker types), includes DXS142, DXS164, DXS206, DXS230, DXS239, DXS268, DXS269, and Dystrophin.</li> </ul>	
<small>Previous GC identifier: GC0XM029640</small>		
<b>Chromosomal Location</b> <small>(According to GeneLoc and/or HUGO, and/or LocusLink(NCBI build 31), Genomic Views According to UCSC and Ensembl)</small>		
<p>Chromosome: X <a href="#">GeneLoc gene densities</a></p> <p>LocusLink cytogenetic band: <b>xp21.2</b> Ensembl cytogenetic band: <b>Xp21.2</b></p> <p>Gene in genomic location: bands according to Ensembl, locations according to GeneLoc (and/or LocusLink and/or Ensembl if different)</p> <p>GeneLoc location for GC0XM029822: <small>(about GC identifiers)</small></p> <p>Start: <b>29,822,399</b> bp from pter</p> <p>End: <b>32,042,786</b> bp from pter</p> <p>Size: <b>2,220,387</b> bases</p> <p>Orientation: <b>minus</b> strand</p> <p>Genomic View:  <a href="#">UCSC Golden Path</a>  <a href="#">UCSC Golden Path with GeneCards custom track</a></p>		
<small>OMIM: HUMAN</small> <small>OMIM ID: 300377</small>		
<small>search databases for MIM named disorders:</small> <ul style="list-style-type: none"> <li>• Duchenne muscular dystrophy</li> <li>• Becker muscular dystrophy</li> </ul>		
<small>SWISS-PROT: DMD_HUMAN</small>		
<b>Disorders &amp; Mutations</b> <ul style="list-style-type: none"> <li>• <b>Disease:</b> Defects in DMD are the cause of Duchenne muscular dystrophy (DMD) and Becker muscular dystrophy (BMD). DMD is the most severe form of muscular dystrophy, characterized by progressive muscle weakness and wasting.</li> </ul>		



## n OMIM 300377 DMD

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

NCBI  
OMIM  
Online Mendelian Inheritance in Man  
Johns Hopkins University

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM

Search OMIM for 300377 Go Clear

Limits Preview Index History Clipboard Details

Entrez

OMIM  
Search OMIM  
Search Gene Map  
Search Morbid Map

Help  
OMIM Help  
How to Link

FAQ  
Numbering System  
Symbols  
How to Print  
Citing OMIM  
Download

OMIM Facts  
Statistics  
Update Log  
Restrictions on Use

Allied Resources  
Genetic Alliance  
Databases  
HGMD  
Locus-Specific  
Model Organisms  
Map Map  
Phenotype  
Davis Human/Mouse  
Homology Maps  
Coriell  
The Jackson  
Laboratory  
Human Gene  
Nomenclature

1: \*300377 GeneTests, Links

**DYSTROPHIN; DMD**  
Alternative titles: symbols

APO-DYSTROPHIN 1, INCLUDED

**TABLE OF CONTENTS**

- [TEXT](#)
- [CLONING](#)
- [MOLECULAR GENETICS](#)
- [GENOTYPE/PHENOTYPE CORRELATIONS](#)
- [ANIMAL MODEL](#)
- [ALLELIC VARIANTS](#)
  - [View List](#)
- [SEE ALSO](#)
- [REFERENCES](#)
- [CONTRIBUTORS](#)
- [CREATION DATE](#)
- [EDIT HISTORY](#)

Gene map locus [Xp21.2](#)

```

    graph TD
      SWISS-PROT --> Protein_Database
      SWISS-PROT --> PIR
      Protein_Database --> OMIM
      PIR --> Genome_Database
      OMIM --> GeneCards
      Genome_Database --> GeneCards
  
```





### Summary Information



#### Summary Information

[View Structure](#)

[Download/Display File](#)

[Structural Neighbors](#)

[Geometry](#)

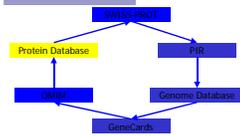
[Other Sources](#)

[Sequence Details](#)

[Structure Factors](#)  
(compressed)

Explore

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



*Title:* N-Terminal Actin-Binding Domain Of Human Dystrophin

*Compound:* Mol\_Id: 1; *Molecule:* Dystrophin; *Chain:* A, B, C, D; *Fragment:* Actin-Binding; *Engineered:* Yes; *Mutation:* Yes

*Authors:* F. L. Norwood, A. J. Sutherland-Smith, N. H. Keep, J. Kendrick-Jones

*Exp. Method:* X-ray Diffraction  
*Classification:* Structural Protein

*Source:* Homo sapiens

*Primary Citation:* [Norwood, F. L., Sutherland-Smith, A. J., Keep, N. H., Kendrick-Jones, J.](#): The Structure of the N-Terminal Actin-Binding Domain of Human Dystrophin and How Mutations in This Domain May Cause Duchenne or Becker Muscular Dystrophy *Structure (London)* 8 pp. 481 (2000)  
[\[Medline\]](#)

*Deposition Date:* 19-Jan-2000

*Resolution [Å]:* 2.60

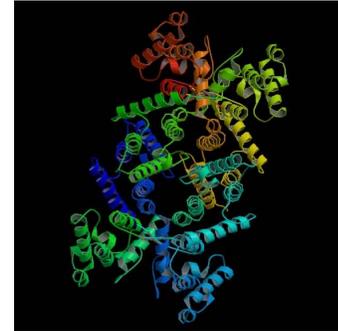
*Space Group:* P 1

*Unit Cell:* dim [Å]: a 59.69 b 79.33 c 81.95  
angles [°] alpha 61.08 beta 78.22 gamma 70.54

*Polymer Chains:* A, B, C, D

*Atoms:* 7622

*HET groups:* HOH



# 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

Symbol/Names:

Name
<input type="text"/>

Cytogenetic Localization:

Chromosome	Left Marker	Right Marker
<input type="text"/>	<input type="text"/>	<input type="text"/>

All Localizations:

Chromosome	Left Marker	Right Marker
<input type="text"/>	<input type="text"/>	<input type="text"/>

Nucleic Acid Sequence Links:

<input type="text"/>
----------------------

Related Segments:

Marker
<input type="text"/>

Polymorphisms:

Polymorphism	Variation Type	Max Het
<input type="text"/>	<input type="text"/>	<input type="text"/>

Mutations:

<input type="text"/>
----------------------

Phenotype Links:

<input type="text"/>
----------------------

Families:

<input type="text"/>
----------------------

GDB-Suchformular

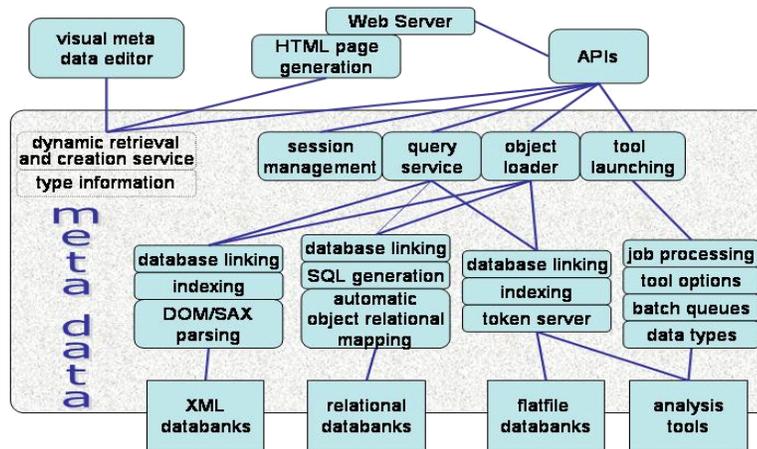
n Nachteil: Starke Einschränkung der Expressivität, keine Unterstützung vom komplexen Anfragen



# 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

\* <http://www.lionbioscience.com>



# SRS: Konzept (1)

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

The screenshot shows the SRS query interface. At the top, there is a navigation bar with links: TOP PAGE, QUERY, RESULTS, SESSIONS, VIEWS, DATABANKS, and HELP. Below this is a search bar with a 'Reset' button, a search scope dropdown (SWISSPROT SPTREMBL REMTREMBL TREMBLNEW), and an 'Info about field' dropdown (AllText). The main area is divided into a left sidebar and a central query builder. The sidebar contains options: 'append wildcards to words' (checked), 'combine searches with' (AND), and 'Number of entries to display per page' (30). The central query builder has a dropdown menu for 'AllText' and a list of fields: ID, AccNumber, Description, GeneName, Keywords, Date, Organism, Organelle, SeqLength, and ProteinID. To the right of the list, there is a 'retrieve entries of type' dropdown (Entry) and a 'SeqSimpleView' dropdown. At the bottom right, there is a 'sequence format' dropdown (swiss).



# SRS: Konzept (2)

n Ein Konzept kann als html-Seite visualisiert werden

The screenshot displays the Swiss-Prot entry for CD22\_HUMAN (P20273). The interface includes a navigation bar with 'Text Entry' and 'SwissEntry', a 'Reset' button, and a 'This entry is from:' section with a 'SWISS-PROT' logo and a 'Save' button. A 'Link' button is also present. A 'Launch' button is next to a 'BlastP' dropdown menu, and a 'Printer Friendly' button is at the bottom left of the main content area.

**General Information about the Entry**

Entry name	SWISSPROT:CD22_HUMAN
Prim. accession #	P20273
Sec. accession #	Q01665, Q92872, O95699, O95701, O95702, O95703;
Created	Release 17, 1-FEB-1991
Last sequence update	Release 38, 15-JUL-1999
Last annotation update	Release 38, 15-JUL-1999

**Description and Origin of the Protein**

Keywords	Glycoprotein, Cell adhesion, Transmembrane, Signal, B-cell, Immunoglobulin domain, Alternative splicing, Phosphorylation, Polymorphism,
Description	b-cell receptor cd22 precursor (leu-14) (b-lymphocyte cell adhesion molecule) (bl-cam).
Gene name(s)	cd22.
Organism source	homo sapiens (human).

VARSPPLIC: 241 417, MISSING (IN CD22-ALPHA).



# SRS: Konzept (3)

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

Name	Short Name	Type	No of Keys	No of Refers
AllText	all	group		0
AccessionNumber	acc	id	3915	
SecAcc	oac	index	67	
ShortName	sm	index	3911	
FullName	fm	index	7726	
Type	tp	index	4	
GO terms	so	index	2049	
Abstract	abs	index	123444	
ProtExample	sem	index	33396	
ProteinRef	pr	index	333	
ChildRef	chr	index	913	
ContainsRef	has	index	127	
FoundInRef	in	index	207	
Taxon	taxon	index	2686	
PubId	pubid	index	1064	
Authors	aut	index	15626	
Title	tit	index	34060	
BookTitle	bkttl	index	102	
Journal	jnl	index	338	
VolumeNo	vol	num	500	
FirstPage	fp	num	1586	
LastPage	lp	num	2669	
Year	yr	num	30	
URL	url	index	4	
MedlineID	mid	index	4776	
DbName	dbn	index	11	
DBacc	dr	index	9449	
Ref Name	nm	index	5888	

From Databank	Entries Linked	To Databank
INTERPRO	1061	PROSITE
INTERPRO	1403	PRINTS
INTERPRO	2664	PFAM
INTERPRO	3435	SWISS
INTERPRO	913	INTERPF
IPRMATCHES	376770	INTERPF

```

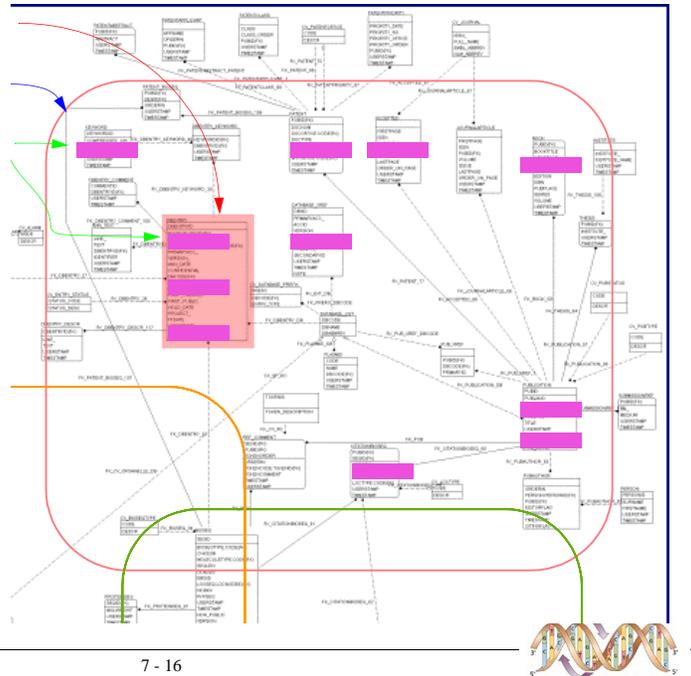
<interpro id="IPR003881" type="Family" short_name="Isochorsmtase_sub">
<name>Isochorsmtase</name>
<abstract><p>Iron is essential for growth in both bacteria and mammals. Controlling the amount of free iron in solution is often used as a tactic by hosts to limit invasion of pathogenic microbes; binding iron tightly within protein molecules can accomplish this. Such iron-protein complexes include haem in blood, lactoferrin in tears/saliva and transferrin in blood plasma. Some bacteria express surface receptors to capture eukaryotic iron-binding compounds, while others have evolved siderophores to scavenge iron from iron-binding host proteins [PMID: 8057905].</p><p>The absence of free iron molecules in the surrounding environment triggers transcription of gene clusters that encode both siderophore-synthesis enzymes, and receptors that recognise iron-bound siderophores [PMID: 2521621]. Classic examples are the enterobactin/enterochelin clusters found in Escherichia coli and Salmonella spp., although similar moieties in other pathogens have been identified. The enzymic machinery that produces vibrionectin in Vibrio cholera is such a homologue [PMID: 9371453].</p><p>EntB, an isochorsmate enzyme, is involved in the second stage of enterobactin biosynthesis. It has a molecular weight of 35kDa, and is believed to possess bifunctional activity. Deletion studies involving EntB- mutants have shown that it is essential for virulence [PMID: 2521622].</p></abstract>
- <example_list>
- <example>
- <db_xref dbkey="007900" db="SWISS" />
- </example>
- <example>
- <db_xref dbkey="P45743" db="SWISS" />
- </example>
- <example>
- <db_xref dbkey="P15048" db="SWISS" />
- </example>
- </example_list>
- <parent_list>
- <ref_ipr ipr_ref="IPR000868" />
- </parent_list>
- <contains>
- <ref_ipr ipr_ref="IPR003880" />
- </contains>
- <member_list>
- <db_xref db="PRINTS" dbkey="PRO1396" name="ISCHRISMTASE" />
- </member_list>
- <acc_acc>
- <db_xref db="IPR000255" />
- </acc_acc>
</interpro>

```



# SRS: Zugriff auf relationale Datenbanken

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



# SRS: Datenzugriff (1)

## n Objektsuche

- Auswahl von Datenquellen
- Spezifikation der Suchkriterien für abfragbaren Attribute
- Schnittmenge der Attribute aller ausgewählten Quellen (z.B. ID)

## n Unterstützte Query-Möglichkeiten

- Textsuche, Bereichsuche für numerische/Datumattribute
- Regulär Ausdrücke

## n Automatische Übersetzung von SRS-Queries nach SQL zum Zugriff auf relationale Datenbanken

## n Ergebnis als Vereinigung der Suchergebnisse über einzelne Quellen

The screenshot displays the SRS web interface. At the top, there are navigation links: SRS, TOP PAGE, QUERY, and RESULT. Below this is a search bar with the text 'APRT' and a 'Quick Search' button. To the left of the search bar are 'Query forms' buttons: Standard, Extended, Browse Databanks, and Applications. To the right, there are sections for 'Amino Acid properties' and 'Sequence databanks - complete' with a grid of checkboxes for various databases like EMBL, SWALL, PIR, ENSEMBL, etc. A dropdown menu is open, showing a list of attributes such as Description, AllText, ID, Division, Accession Number, etc. The 'Description' attribute is selected. Below the dropdown, there are options for 'retrieve entries of type' (Entry), 'View' (table/list), and 'sequence format' (embl). A 'Submit Query' button is visible at the bottom right of the search area.



# SRS - Datenzugriff (2)

## n Querverweis-Suche

- Für Objekte einer Ergebnismenge oder für eine ganze Quelle
- Referenzierte Objekte in anderen Datenquellen

## n Automatische Bestimmung der Pfade zwischen Quellen

- Shortest-Path-Algorithmus
- Gleiche Semantik der Beziehungen

## n Tool-Integration

- Anwendung auf Ergebnismengen der Abfragen
- Anzeige direkt im Web-Browser
- Große Anzahl von Tools bereits integriert

The screenshot displays the SRS web interface with three panels. The top panel shows search results for the query "[libs=(swall pir ensembl) -AllText:APRT\*]" with 100 entries. It includes a table with columns for database name, accession number, and source. The middle panel shows options for "Set Db" and "Find all Entries", along with a "Submit Link" button and a "View result with:" dropdown. The bottom panel shows search results for the query "[libs=(swall pir ensembl) -AllText:APRT\*] > PATHWAY" with 38 entries, listing various pathway identifiers.

SWALL_PIR_ENSEMBL	Accession	
<input type="checkbox"/> PIR_RTHUA	S06232	ader
<input type="checkbox"/> PIR_I49510	I49510	gene
<input type="checkbox"/> SWALL_Q12898	Q12898	Ade
<input type="checkbox"/> SWALL_O44095	O44095	Ade
<input type="checkbox"/> SWALL_O77103	O77103	Ade

PATHWAY
<input type="checkbox"/> PATHWAY_aae00230
<input type="checkbox"/> PATHWAY_ah00230
<input type="checkbox"/> PATHWAY_ape00230
<input type="checkbox"/> PATHWAY_ath00230
<input type="checkbox"/> PATHWAY_bbu00230
<input type="checkbox"/> PATHWAY_bsu00230
<input type="checkbox"/> PATHWAY_cel00230



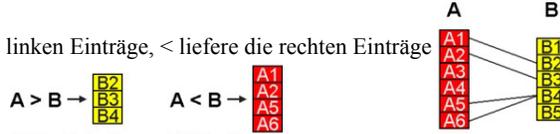
# SRS Anfrage Sprache

- 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:elestase], oder [swissprot-AllText:duchenne\*]
  - Wildcards: [swissprot-aut:sanger,f\*!coulson,a\*]
  - Reguläre Ausdrücke: [swissprot-aut:/mue?ller/]
  - 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\_caeel] > prodom > pdb  
kein direkt Link von „gshr\_caeel” zur PDB, aber über ProDom (Protein Domainen), Ergebnis homologe Proteine zu „gshr\_caeel”



# 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”



# 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']  
    $AppOpt: [des prompt: 'Description' defStr: 'cancer']  
}  
queryStr:  
@ "[swissprot-des:($des)*] | [swissprot-acc:($ac)*]"  
]
```

- Das @ zeigt, dass die Anfrage zur Laufzeit bearbeitet wird.
- Datei mit Anfragedefinition wird ähnlich wie ein HTML Link in SRS-Seite eingebunden.



# 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:rnelas]`
  - `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



# 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 aatttcaatg aagagtgttg ttgtagctgg cccattaatt taggcatgtg
cacaccttc
61 tttttcctcc catacacacc tgtgaacttg tgagacagat ggggaattta
tttattgttt
121 tttttttaa tataaagatg ataagtcatt gaacccttct gtctactcaa
```

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

**Eingabe der Sequenz in das Formular**

The request ID is

or

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

**Bearbeitungszeit...**

**request ID für die Einsendung wird zurückgegeben**

```
>ref|NT_025938.2|Hs22_26094 Homo sapiens chromosome 22 contig12211.1
Length = 65461

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

Query: 261 tcgatgaagaacgcagcgaatgcgataagtaatgtgaattgcag 305
|||||
Sbjct: 16868 tcgatgaagaacgcagctagctgcgagattaatgtgaattgcag 16824
```

**Ergebnisse der Berechnung (e-mail oder online)**



# 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



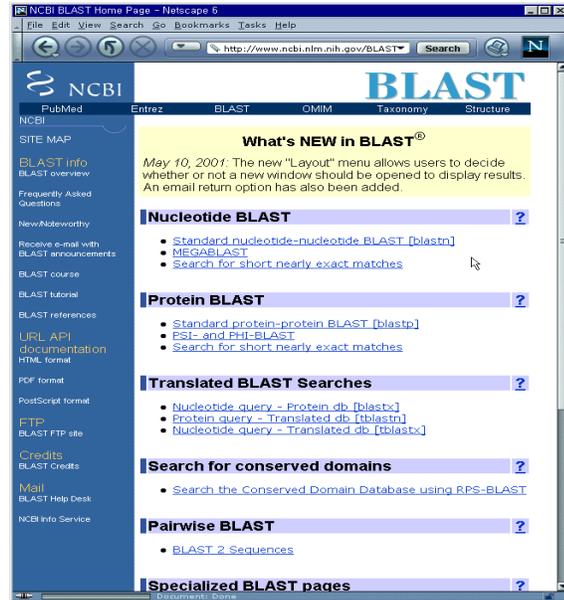
# Hyper Text Query Language (HTQL)

- 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 der 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)
html 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)
html 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 displays the LifeDB web interface. On the left, a text area contains the SQL query: `select sts_code, GenbankID from genes where sts_GenbankID(sts_code) is not null`. Below the text area is a 'do' button and a 'back' link. A scrollable area below shows a list of queries, with the first one being `345452 sts_code, GenbankID` and a 'refresh' link. On the right, the 'Results' section shows the query ID `345452` and a table with 5 records. The table has two columns: 'sts\_code' and 'GenbankID'. Below the table, it says '5 records found.' with a 'refresh' link and a 'Return Home' link.

```
select sts_code, GenbankID from genes
where sts_GenbankID(sts_code) is not
null
```

do back

Queries:

[345452](#) sts\_code, GenbankID

1 queries. [refresh](#)

## Results

query: 345452

sts_code	GenbankID
WI-5270	G04845
WI-5275	G04849
WI-5276	G04850
WI-5278	G04851
WI-5279	G04852

5 records found. [refresh](#)

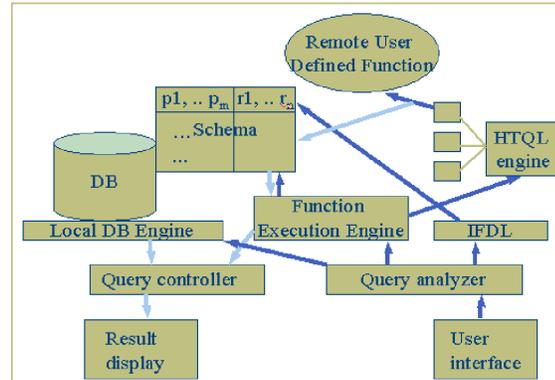
[Return Home](#)



# 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 lokalen 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_htql varchar(200)
```

IF-Metadaten-Tabelle



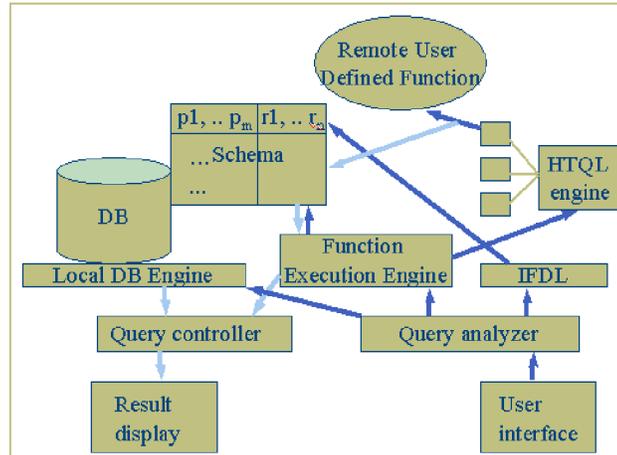
## LifeDB Anfrageverarbeitung.: Architektur (2)

### 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



# 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.

