

**Einführung in die Angewandte Bioinformatik:
Nukleotidsequenz-Datenbanken
06.05.2010**

Prof. Dr. Sven Rahmann

Datenbanken am NCBI über Entrez

<http://www.ncbi.nlm.nih.gov/Entrez>

- NIH = National Institute of Health
- NLM = National Library of Medicine
- NCBI = National Center for Biotechnology Information

gemeinsamer Zugang über Entrez
(einheitliche Eingabe von Suchanfragen)

Welcome to the Entrez cross-database search page

 PubMed: biomedical literature citations and abstracts	 Books: online books
 PubMed Central: free, full text journal articles	 OMIM: online Mendelian Inheritance in Man
 Site Search: NCBI web and FTP sites	 OMIA: online Mendelian Inheritance in Animals
 Nucleotide: sequence database (includes GenBank)	 UniGene: gene-oriented clusters of transcript sequences
 Protein: sequence database	 CDD: conserved protein domain database
 Genome: whole genome sequences	 3D Domains: domains from Entrez Structure
 Structure: three-dimensional macromolecular structures	 UniSTS: markers and mapping data
 Taxonomy: organisms in GenBank	 PopSet: population study data sets
 SNP: single nucleotide polymorphism	 GEO Profiles: expression and molecular abundance profiles
 Gene: gene-centered information	 GEO DataSets: experimental sets of GEO data
 HomoloGene: eukaryotic homology groups	 Cancer Chromosomes: cytogenetic databases
 PubChem Compound: unique small molecule chemical structures	 PubChem BioAssay: bioactivity screens of chemical substances
 PubChem Substance: deposited chemical substance records	 GENSAT: gene expression atlas of mouse central nervous system
 Genome Project: genome project information	 Probe: sequence-specific reagents
 dbGaP: genotype and phenotype	 Protein Clusters: a collection of related protein sequences

Datenbanken am NCBI über Entrez <http://www.ncbi.nlm.nih.gov/Entrez>

Datenbanken-Netzwerk in Entrez
<http://www.ncbi.nlm.nih.gov/Database/>

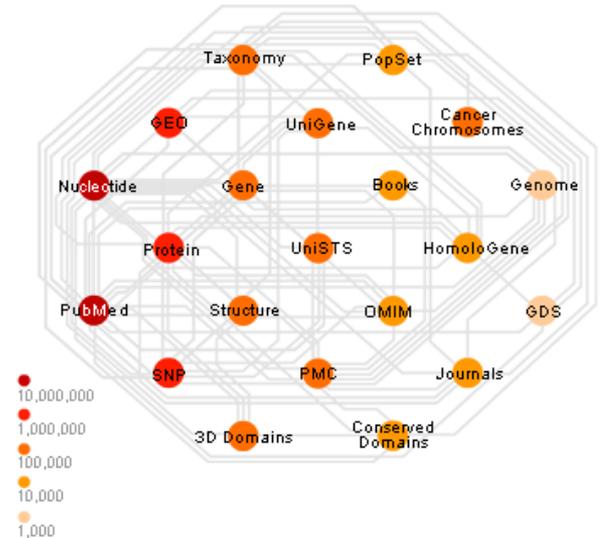
unten: Nukleotid-Datenbanken

Für uns zunächst am wichtigsten:

- NCBI GenBank (NCBI Nucleotide)
- NCBI RefSeq
- NCBI Gene
- NCBI Genome

Bemerkung:

„Nucleotide“ ist eine Obermenge von GenBank, enthält z.B. auch RefSeq u.a.



The complete list of Entrez databases can be viewed in the search pull down menu.

Nucleotide Databases

- | | |
|----------------------------|-------------------------------|
| dbEST | PopSet |
| dbGSS | Probe |
| dbSNP | RefSeq |
| dbSTS | TPA |
| Nucleotide | Trace Archive |
| GenBank | UniGene |
| HomoloGene | UniSTS |
| MGC | |

Warum viele Datenbanken, nicht eine (große)?

Historische Gründe; Spezialisierung; Expertenwissen

Verschiedene Personen an verschiedenen Instituten interessier(t)en sich für verschiedene Bereiche und sammel(te)n nur dort Daten.

(Den besten Käse gibt's im Käsefachgeschäft -- nicht im Supermarkt.)

Ausfallsicherheit

Information ist physikalisch an verschiedenen Orten gespeichert.

Nur geringer Verlust an Bequemlichkeit

durch moderne WWW-basierte Systeme spielt es kaum eine Rolle, ob im Browser angezeigte Informationen aus einer oder mehreren DBen kommen.

Achtung! Sicherheitsaspekte

Keine der öffentlichen Datenbanken unterstützt derzeit Verschlüsselung. Jeder kann „mitlesen“, für welche Sequenzen man sich interessiert. Ungünstig für Firmen, die patent-relevante Informationen suchen.

Primärdatenbanken und Sekundärdatenbanken

Primärdatenbanken

enthalten „Rohdaten“, z.B. biologische Sequenzdaten,
dienen zur Ablage von aus Experimenten gewonnenen Daten,
können Annotationen enthalten (z.B. Herkunft, Literaturverweis)
z.B. GenBank am NCBI

Sekundärdatenbanken

enthalten aus Primärdatenbanken gewonnenes „Wissen“,
d.h. die Daten in einer Primärdatenbank werden gefiltert, überprüft, annotiert, ...
z.B. RefSeq am NCBI

Beispiel

Die Firma Celera Genomics hat um 2000 das Humangenom sequenziert,
dabei pro Tag 175000 DNA-Sequenzen (reads) à 500bp erzeugt.
Die reads speichert man in einer Primärdatenbank.
Das daraus rekonstruierte Humangenom in einer Sekundärdatenbank.

Datenbanken für Nukleotidsequenzen - Primärdatenbanken

International Nucleotide Sequence Database Collaboration
(INSDC, <http://www.insdc.org>)

- NCBI Nucleotide (GenBank)
<http://www.ncbi.nlm.nih.gov/Genbank/>
- EMBL/EBI Nucleotide Sequence Database
<http://www.ebi.ac.uk/embl/>
- DDBJ (DNA DataBase of Japan)
<http://www.ddbj.nig.ac.jp/>

Gleicher Inhalt, automatischer Abgleich

Zugriff über „Metasuchmaschinen“

Entrez (NCBI), **SRS** (sequence retrieval system; EMBL),

ARSA (all-round retrieval of sequence and annotation; DDBJ)

DNA Data Bank of Japan (DDBJ), Mishima, Japan



- [Home page](#)
- [Sequence retrieval](#)
- [DNA sequence Submissions](#)

EMBL Nucleotide Sequence Database (EBI), Hinxton, UK



- [Home page](#)
- [Sequence retrieval](#)
- [DNA sequence Submissions](#)

GenBank(NCBI), Bethesda, MD, USA



- [Home page](#)
- [Sequence retrieval](#)
- [DNA sequence Submissions](#)

Datenbanken für Nukleotidsequenzen - Sekundärdatenbanken

NCBI RefSeq (Referenz-Sequenz)

- genomische DNA, RNA-Transkripte
(es gibt auch RefSeq für Proteine, siehe später)
- stabile Basis für weitere Annotationen
 - wo liegen Gene, Transkripte, ... im Genom,
 - wo gibt es Mutationen,
 - wo konservierte und variable Segmente verglichen mit verwandten Organismen?
- breite taxonomische Abdeckung: es gibt RefSeqs für
 - Eukaryoten, Prokaryoten, Viren
- seltene Änderungen

Datenbanken für Nukleotidsequenzen - Sekundärdatenbanken

NCBI Gene (Informationen zu Genen)

- Verweise auf zugehörige Nukleotid-Sequenzen
- Funktion eines Gens
- verschiedene Isoformen eines Gens
- Proteine dazu

NCBI Genome (komplette Genome)

- Liste aller öffentlich bisher sequenzierten Genome, viele Bakterien, verschiedene Strains verfügbar

Wichtige Unterschiede Primärdatenbank GenBank vs. Sekundärdatenbank RefSeq

GenBank

Not curated
Author submits
Only author can revise
Multiple records for same loci common

Records can contradict each other
No limit to species included
Data exchanged among INSDC members
Akin to primary literature
Proteins identified and linked

Access via NCBI Nucleotide databases

RefSeq

Curated
NCBI creates from existing data
NCBI revises as new data emerge
Single records for each molecule of major organisms

Limited to model organisms
Exclusive NCBI database
Akin to review articles
Proteins and transcripts identified and linked

Access via Nucleotide & Protein databases

NCBI Entrez GenBank / Nucleotide

Primärdatenbank,
nicht kuriert, daher viele doppelte Einträge, „Müll“

Publikation neuer Nukleotidsequenzen in GenBank
bevor man in einer relevanten molekularbiologischen Zeitschrift veröffentlicht.

Exponentielles Wachstum seit den 1990er Jahren.

Feb. 2008: 85,759,586,764 nt in 82,853,685 Sequenz-Einträgen

Apr. 2009: 102,980,268,709 nt in 103,335,421 Sequenz-Einträgen

Apr. 2010: 114,348,888,771 nt in 119,112,251 Sequenz-Einträgen

Quelle: Release Notes der aktuellen GenBank-Version unter
<ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>

Zugriffsnummern (Accession Numbers; ACs)

Nummer, die einen Datenbankeintrag eindeutig identifiziert („Schlüssel“)

Beispiel: PubMed ID; Feld [PMID] bei der PubMed-Datenbanksuche.

Kennt man die AC, braucht man keine komplexen Suchanfragen.

ACs werden benutzt, um Datenbanken zu verbinden.

Beispiel: Bei einer (codierenden) DNA-Sequenz in einer Nukleotiddatenbank findet man die AC der zugehörigen Proteinsequenz(en) für eine Proteindatenbank.

Verwendung:

- Manuell – notiere ACs auf einem Zettel, suche damit in anderen Datenbanken
- Per Link – direkter WWW-Link zu den referenzierten Datenbanken
- Datenintegrationssysteme – fragen automatisch mehrere Datenbanken ab, stellen ggf. Ergebnisse übersichtlich zusammen

Suche in NCBI GenBank mit Accession Number

The screenshot shows the NCBI search interface. At the top, there is a navigation bar with 'All Databases', 'PubMed', 'Nucleotide', 'Protein', and 'Genome'. The search bar contains 'Nucleotide' in the dropdown menu and 'for K03160' in the text input field. Below the search bar are buttons for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The search results section shows 'Found 1 nucleotide sequence. Nucleotide [1]'. Below this, there are controls for 'Display' (set to 'Summary'), 'Show' (set to '20'), 'Sort by', and 'Send to'. At the bottom, there are buttons for 'All: 1', 'Bacteria: 0', 'RefSeq: 0', and 'mRNA: 0'. The search results list shows one entry: '1: [K03160](#) Reports Auricula auricula-judae 5S ribosomal RNA gi|173593|gb|K03160.1|AAURRA[173593]'. In the background, there is a colorful diagram of a nucleotide structure and a DNA sequence.

Eingabe einer Accession Number führt sofort zum gewünschten Eintrag. In diesem Fall (da „K03160“ niemals sonst als Stichwort auftritt) muss man nicht explizit angeben, dass es sich um [AC] handelt.

Ergebnis im GenBank-Format

```

1: K03160. Reports Auricularia auricula...[gi:173593]
Features Sequence

LOCUS      AAURRA                      118 bp    rRNA     linear   PLN 13-DEC-1995
DEFINITION Auricularia auricula-judae 5S ribosomal RNA.
ACCESSION K03160
VERSION   K03160 GI:173593
KEYWORDS   5S ribosomal RNA; ribosomal RNA.
SOURCE     Auricularia auricula-judae (ear fungus)
ORGANISM   Auricularia auricula-judae
           Eukaryota; Fungi; Dikarya; Basidiomycota; Agaricomycotina;
           Agaricomycetes; Auriculariales; Auriculariaceae; Auricularia.
REFERENCE  1 (bases 1 to 118)
AUTHORS    Huysmans,E., Dams,E., Vandenberghe,A. and De Wachter,R.
TITLE      The nucleotide sequences of the 5S rRNAs of four mushrooms and
           their use in studying the phylogenetic position of basidiomycetes
           among the eukaryotes
JOURNAL    Nucleic Acids Res. 11 (9), 2871-2880 (1983)
PUBMED     6856478
COMMENT    Original source text: Auricularia auricula-judae rRNA.
FEATURES   Location/Qualifiers
           source             1..118
                               /organism="Auricularia auricula-judae"
                               /mol_type="rRNA"
                               /db_xref="taxon:29892"
           rRNA             1..>118
                               /product="5S ribosomal RNA"
ORIGIN
1 atccacggcc ataggactct gaaagcactg catcccgtcc gatctgcaaa gtaaccaga
61 gtaccgcccc gttagtacca cgggtggggga ccacgcggga atcctgggtg ctgtggtt
//

```

Ein GenBank-Eintrag

- beginnt mit LOCUS
- endet mit //
- Sequenz hinter ORIGIN
- Annotation: FEATURES

GeneBank-Format (flat file) – kann man herunterladen

```

LOCUS      AAURRA          118 bp ss-rRNA          RNA          16-JUN-1986
DEFINITION A.auricula-judae (mushroom) 5S ribosomal RNA.
ACCESSION  K03160
VERSION    K03160.1  GI:173593
KEYWORDS   5S ribosomal RNA; ribosomal RNA.
SOURCE     A.auricula-judae (mushroom) ribosomal RNA.
  ORGANISM Auricularia auricula-judae
            Eukaryota; Fungi; Eumycota; Basidiomycotina; Phragmobasidiomycetes;
            Heterobasidiomycetidae; Auriculariales; Auriculariaceae.
REFERENCE  1 (bases 1 to 118)
  AUTHORS  Huysmans,E., Dams,E., Vandenberghe,A. and De Wachter,R.
  TITLE    The nucleotide sequences of the 5S rRNAs of four mushrooms and
            their use in studying the phylogenetic position of basidiomycetes
            among the eukaryotes
  JOURNAL  Nucleic Acids Res. 11, 2871-2880 (1983)
FEATURES   Location/Qualifiers
  rRNA     1..118
            /note="5S ribosomal RNA"
BASE COUNT 27 a      34 c      34 g      23 t
ORIGIN     5' end of mature rRNA.
            1 atccacggcc ataggactct gaaagcactg catcccgtcc gatctgcaaa gttaaccaga
            61 gtaccgcca gttagtagca cgggtgggga ccacgcggga atcctgggtg ctgtggtt
//
LOCUS      ABCRRAA          118 bp ss-rRNA          RNA          15-SEP-1990
...

```

Suchmöglichkeiten in GenBank (und anderen NCBI-DBs)

[ACCN] – sucht nach Accession Number

[SLEN] – sucht nach bestimmten Sequenzlängen

[ORGN] – sucht nach Sequenzen im angegebenen Organismus

Bereichssuche (:)

Bei numerischen Eingaben kann mit einem Doppelpunkt (:) ein ganzer Bereich durchsucht werden.

Beispiel: 200:222 [SLEN]

findet Sequenzen der Länge 200 bis 222

Wildcard-Suche (*)

Ist das Wortende nicht genau bekannt oder soll offen gelassen werden, kann es mit einem Stern (*) abgeschnitten werden.

Beispiel: HUM* [ORGN]

findet Sequenzen in *Homo sapiens* (human) oder *Humulus lupulus* (Hopfen)

Suchmöglichkeiten in EMBL Nucleotide mit SRS

EMBL Nucleotide ist inhaltlich eine Kopie von GenBank.

Statt mit Entrez durchsucht man sie mit SRS (Sequence Retrieval System)

URL: <http://srs.ebi.ac.uk/>

- schnelle Stichwortsuche über „Quick Text search“
- genaue Eingrenzung durch Klick auf Reiter „Library page“, „Query Form“

The screenshot shows the EMBL SRS website interface. At the top, there is a navigation bar with 'EMBL-EBI' and 'EB-eye Search' on the left, a search input field with 'All Databases' and 'Enter Text Here', and 'Go', 'Reset', and 'Give us feedback' buttons on the right. Below this is a menu with 'Databases', 'Tools', 'EBI Groups', 'Training', 'Industry', 'About Us', and 'Help'. A secondary menu below that includes 'Quick Search' (circled in red), 'Library Page', 'Query Form', 'Tools', 'Results', 'Projects', 'Views', 'Databanks', and 'HELP'. The main content area features a 'Quick Text Search' section with a dropdown menu set to 'Nucleotides' (circled in red) and a search input field. Below the search field is a 'Search' button. To the left of the search section is a 'Tips' box with advice on using SRS. Below the search section is a 'News and Announcements' section with a 'Search Tips' link and an 'Important announcements:' section containing a notice about a service interruption on Monday 16th July.

EMBL Nucleotide mit SRS – Library Page

Auswahl der zu durchsuchenden Datenbanken

The screenshot shows the EMBL SRS Library Page. At the top, there is a navigation bar with links for Databases, Tools, FBI Groups, Training, Industry, About Us, and Help. Below this is a search bar with a 'Quick Search' button and a 'Library Page' button, which is circled in red. The main content area is divided into two columns. The left column, titled 'Search Options', contains instructions for selecting databanks and search terms, with buttons for 'Standard Query Form', 'Extended Query Form', and 'Browse Entries'. The right column, titled 'Available Databanks', lists various databases under different categories. The 'EMBL' checkbox is circled in red, and the 'RefSeq Genome' checkbox is also circled in red. Other categories include Literature, Bibliography and Reference Databases; Gene Dictionaries and Ontologies; and Nucleotide sequence databases - subsections.

Search Options

1. Select the **databanks** you want to search
2. 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

Taxonomy OMIM OMIM Morbid Map MEDLINE
 Patent Abstracts Karyn's Genomes

Literature, Bibliography and Reference Databases - subsections

MEDLINE (Updates) MEDLINE (Main Release 2007) MED2PUB

Gene Dictionaries and Ontologies

Nucleotide sequence databases

EMBL Patent DNA IMGT/LIGM-DB IMGT/HLA
 IPD-KIR EMBL (Contig) EMBL (Contigs expanded) EMBL (Annotated Cons)
 EMBL (Coding Sequences) Genome Reviews GR Gene Sets RefSeq Genome
 LiveLists EMBL ID/Accession Mapping EMBL MGA

Nucleotide sequence databases - subsections

EMBL (Updates) EMBL (Release) EMBL (Whole Genome Shotgun)
 EMBL (Whole Genome Shotgun release) EMBL (Whole Genome Shotgun updates) EMBL (Contig release)

Tips

▶ bookmark this [link](#) to

EMBL Nucleotide mit SRS – Query Form

Einstellung der Suchparameter; dann Klick auf „Search“

The screenshot shows the EMBL SRS Query Form interface. The 'Query Form' tab is highlighted with a red circle. Below the navigation bar, there is a search input field containing 'EMBL' and a 'Reset' button. The main content area is divided into several sections:

- Search Options:** Includes options for combining search terms (set to '& (AND)'), using wildcards (checked), and getting results of type 'Entry'.
- Fields you can search:** A table with two columns: 'Fields you can search' and 'Your search terms'. The 'Accession Number' field is highlighted with a red circle. The text above the table states: 'In a single field, you can separate multiple values by: &, | or !'. A 'Search' button is also highlighted with a red circle.
- Result Display Options:** Includes options for viewing results using 'EMBLSeqSimpleView' or 'Create a view'.
- Create a view:** Includes options for choosing fields (ID, Topology, Molecule, Data Class) and displaying as 'Table' or 'List'. The 'Sequence Format' is set to 'embl'.

EMBL Nucleotide mit SRS – Ergebnisseite

General Information			
Primary Accession #	K03160		
Accession #	K03160		
SRS Entry ID	EMBL:K03160 (formerly EMBL:AARRA)		
Molecule Type	linear rRNA		
Sequence Length	118		
Entry Division	FUN (<i>Fungi</i>)		
Entry Data Class	STD (<i>Standard</i>)		
Sequence Version	K03160.1		
Creation Date	17-JAN-1991		
Modification Date	04-MAR-2000		
EMBL-SVA	K03160		
Description			
Description	Auricularia auricula-judae 5S ribosomal RNA.		
Keywords	5S ribosomal RNA; ribosomal RNA.;		
Organism	Auricularia auricula-judae (ear fungus)		
Organism Classification	Eukaryota; Fungi; Dikarya; Basidiomycota; Agaricomycotina; Agaricomycetes; Auriculariales; Auriculariaceae; Auricularia.		
References			
1.	Huysmans,E.; Dams,E.; Vandenberghe,A.; De Wachter,R.; The nucleotide sequences of the 5S rRNAs of four mushrooms and their use in studying the phylogenetic position of Nucleic Acids Res. 11(9):2871-2880 (1983)		
	DOI	10.1093/nar/11.9.2871	
	PubMed	6856478        CiteXplore	
	Position	1-118	
Database Cross-references			
	RFAM	RF00001 .	
Features			
Key	Location	Qualifier	Value
source	1..118	organism	Auricularia auricula-judae
		mol_type	rRNA
		db_xref	taxon:29892
rrna	1..>118	product	5S ribosomal RNA
Sequence			
Characteristics	Length: 118 BP, A Count: 27, C Count: 34, G Count: 34, T Count: 23, Others Count: 0		
Sequence	<pre>>emb1 K03160 K03160 Auricularia auricula-judae 5S ribosomal RNA. atcgagggcctggcctctggaagcctggatccgctggcctctgcaagcttaagcaga</pre>		

Ergebnis der
Suche nach
Accession Number
K03160:

Inhalt wie
bei NCBI-Entrez,
nur andere Darstellung

EMBL-Format (flat file) mit SRS

2-Buchstaben-
Kürzel
identifizieren
die Elemente
des Eintrags

```

ID   K03160; SV 1; linear; rRNA; STD; FUN; 118 BP.
AC   K03160;
DT   17-JAN-1991 (Rel. 26, Created)
DT   04-MAR-2000 (Rel. 63, Last updated, Version 7)
DE   Auricula auricula-judae 5S ribosomal RNA.
KW   5S ribosomal RNA; ribosomal RNA.
OS   Auricularia auricula-judae (ear fungus)
OC   Eukaryota; Fungi; Dikarya; Basidiomycota; Agaricomycotina; Agaricomycetes;
OC   Auriculariales; Auriculariaceae; Auricularia.
RN   [1]
RP   1-118
RX   DOI; 10.1093/nar/11.9.2871
RX   PUBMED; 6856478.
RA   Huysmans E., Dams E., Vandenberghe A., De Wachter R.;
RT   "The nucleotide sequences of the 5S rRNAs of four mushrooms and their use
RT   in studying the phylogenetic position of basidiomycetes among the
RT   eukaryotes";
RL   Nucleic Acids Res. 11(9):2871-2880(1983).
XX
DR   RFAM; RF00001.
FH   Key          Location/Qualifiers
FH
FT   source          1..118
FT                   /organism="Auricularia auricula-judae"
FT                   /mol_type="rRNA"
FT                   /db_xref="taxon:29892"
FT   rRNA           1..>118
FT                   /product="5S ribosomal RNA"
XX
SQ   Sequence 118 BP; 27 A; 34 C; 34 G; 23 T; 0 other;
      atccacggcc ataggactct gaaagcactg catcccgtcc gatctgcaaa gttaaccaga      60
      gtaccgcca gttagtagca cgggtggggga ccacgcggga atcctgggtg ctgtgggt      118
//

```

NCBI Entrez Gene (Sekundärdatenbank)

Motivation für Entrez Gene

GenBank enthält alle möglichen Arten von Nukleotid-Sequenzen. Nicht jedes „Stück DNA“ im Genom gehört zu einem Gen. Ein Gen kann durch viele Einträge in GenBank repräsentiert sein. Man benötigt eine Gen-zentrierte „Sicht“ auf diese Sequenzen; Gene sind ein viel häufigerer Startpunkt für eine Suche als einzelne Sequenzen.

Beispiel:

In einem Artikel wird das *dtxR*-Gen in *Corynebacterium glutamicum* erwähnt.

The screenshot shows the NCBI Entrez Gene search interface. The search term 'dtxR' is entered in the search box, and the results are displayed in a summary view. The search results show 195 items, with the first 20 items displayed. The text 'All: 195' is circled in red, indicating the total number of results. The interface includes navigation buttons like 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The search results are sorted by relevance, and the display format is set to 'Summary'.

(195 Treffer sind zu viele!)

Suche mit Feldnamen in Entrez Gene

[gene name] – Name des Gens

[organism] = [orgn] – Organismus, in dem man nach Genen sucht

Beispiel: dtxr[gene name] AND Corynebacterium glutamicum[organism]

All Databases PubMed Nucleotide Protein Genome Structure PMC

Search Gene for dtxr[gene name] AND corynebacterium glutamicu Go Clear [Save Search](#)

Limits Preview/Index History Clipboard Details

Display Full Report Show 20 Sort by Relevance Send to

All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 0

1: dtxR IRON DEPENDENT REGULATORY PROTEIN-DTXR HOMOLOG [*Corynebacterium glutamicum* ATCC 13032]

GeneID: 3343964

updated 05-Apr-2008

Summary

Gene name dtxR

Locus tag cg2103

Gene type protein coding

RefSeq status Provisional

Organism [Corynebacterium glutamicum ATCC 13032 \(strain: DSM 20300; ATCC 13032\)](#)

Lineage Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Corynebacteriaceae; Corynebacterium

Genomic regions, transcripts, and products

Go to [reference sequence details](#)

[Try our new Sequence Viewer](#)



Suche mit Feldnamen – Hinweise

Feldnamen sind hilfreich zur Präzisierung der Suche.

Liste aller Feldnamen: Hilfe-Seiten am NCBI

Beispiele auch in den Übungsaufgaben.

Erinnerung: Boole'sche Operatoren AND, OR, NOT

Entrez wertet ungewöhnlich aus: von links nach rechts

Daher Klammern setzen, um Gruppen von Bedingungen abzugrenzen.

Beispiel (in der NCBI Gene-Datenbank):

Gene der Maus auf Chromosom 1 oder 2, die mit Krebs in Verbindung stehen?

Falsch: cancer [dis] AND mouse [orgn] AND 1 [chr] OR 2 [chr] – 128588 Ergebnisse

Richtig: cancer [dis] AND mouse [orgn] AND (1 [chr] OR 2 [chr]) – 7 Ergebnisse

Richtig: 1 [chr] OR 2 [chr] AND cancer [dis] AND mouse [orgn] – 7 Ergebnisse

Richtig: (1 [chr] OR 2 [chr]) AND cancer [dis] AND mouse [orgn] – 7 Ergebnisse

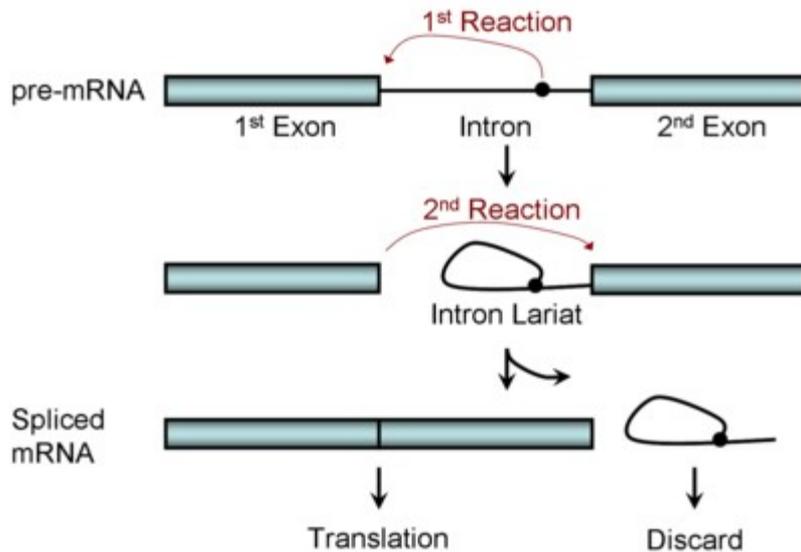
Isoformen von Genen (alternatives Splicing)

Transkription von Protein-codierenden Genen:

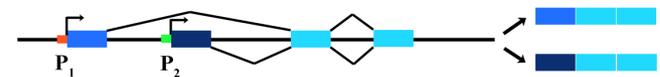
Genomische DNA wird durch RNA-Polymerase abgelesen und in mRNA kopiert.

Spleißen (splicing): Introns werden aus mRNA entfernt.

Durch alternatives Splicing können aus einem Transkript verschiedene mRNAs (Isoformen desselben Gens) und damit Proteine entstehen.



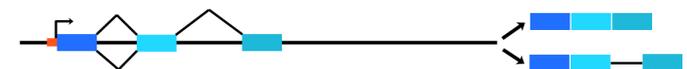
(a) Alternative selection of promoters (e.g., *myosin* primary transcript)



(b) Alternative selection of cleavage/polyadenylation sites (e.g., *tropomyosin* transcript)



(c) Intron retaining mode (e.g., *transposase* primary transcript)



(d) Exon cassette mode (e.g., *troponin* primary transcript)



Alternatives Splicing in NCBI Entrez Gene

Ergebnis-Seiten zeigen alle bekannten Isoformen eines Gens.

Beispiel: Homo sapiens, Myosin heavy chain 14

1: MYH14 myosin, heavy chain 14 [*Homo sapiens*]

GeneID: 79784

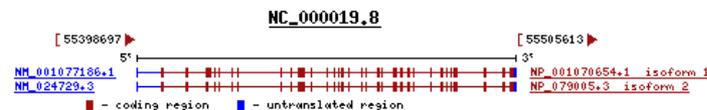
updated 25-Apr-2008

Summary	
Official Symbol	MYH14 provided by HGNC
Official Full Name	myosin, heavy chain 14 provided by HGNC
Primary source	HGNC:23212
See related	Ensembl:ENSG00000105357 ; HPRD:10543 ; MIM:608568
Gene type	protein coding
RefSeq status	Reviewed
Organism	Homo sapiens
Lineage	<i>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo</i>
Also known as	DFNA4; MHC16; myosin; FP17425; FLJ13881; FLJ43092; KIAA2034; NMHC-II-C; DKFZp667A1311
Summary	This gene encodes a member of the myosin superfamily. Myosins are actin-dependent motor proteins with diverse functions including regulation of cytokinesis, cell motility, and cell polarity. Mutations in this gene result in one form of autosomal dominant hearing impairment. Multiple transcript variants encoding different isoforms have been found for this gene.

Genomic regions, transcripts, and products

Go to [reference sequence details](#)

[Try our new Sequence Viewer](#)



NCBI Entrez Genome (Sekundärdatenbank)

Motivation für Entrez Genome

Es gibt mittlerweile viele komplett sequenzierte Genome;
d.h. die gesamte Erbinformation vieler Spezies ist mittlerweile bekannt.
Bei Bakterien sind sogar oft mehrere Stämme (strains) sequenziert.

Beispiel:

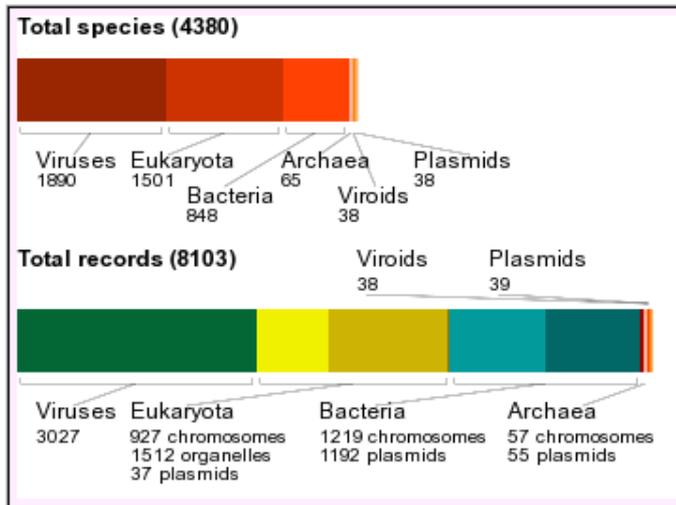
Wie viele Stämme des Tuberkulose-Erregers sind bereits bekannt?

The screenshot shows the NCBI Entrez Genome search interface. The search term "mycobacterium tuberculosis" is entered in the search box. The results show "All: 220" and "Item 1 - 20 of 220". The first result is "1: [NC 010612](#) Mycobacterium marinum M, complete genome".

220 ist falsch!
[orgn] verwenden!
Lösung: 21

NCBI Entrez Genome – Liste aller Genome

Alle Genome in 'NCBI Genome' lassen sich auflisten:
unter <http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome>



Inhalt der Genome – Datenbank
(Quelle: NCBI; Mai '08)

Durch Klick auf 'Bacteria' kommt man zu einer Liste, in der man auch) alle sequenzierten (öffentlich zugänglichen) Stämme von *M. tuberculosis* findet:

Mycobacterium tuberculosis CDC1551	chromosome	NC 002755	4403837 nt
Mycobacterium tuberculosis F11	chromosome	NC 009565	4424435 nt
Mycobacterium tuberculosis H37Ra	chromosome	NC 009525	4419977 nt
Mycobacterium tuberculosis H37Rv	chromosome	NC 000962	4411532 nt

Zusammenfassung

Nukleotid-Sequenzdatenbanken

- Primär: NCBI Entrez Nucleotide (GenBank) = EMBL Nucleotide = DDBJ
- Sekundär: Entrez RefSeq, Entrez Gene, Entrez Genome

Suche in Sequenzdatenbanken

- verschiedene Systeme (Entrez, SRS), gleiche Möglichkeiten
- Feldnamen verwenden zur Eingrenzung
- Bereichs-Operator „:“ und Wildcard-Operator „*“
- Boole'sche Operatoren AND, OR, NOT: Klammern setzen!

Fortsetzung

Statistik mit R

Zugriff auf Elemente von Vektoren

Vektoren werden mit eckigen Klammern [] indiziert.

Indizierung beginnt bei 1, nicht bei 0 !

Man kann auch mit Vektoren indizieren!

```
x = seq(0, 10, by=0.5)      # erzeugt 0, 0.5, 1, ..., 9.5, 10
x[0]                       # gibt es nicht
x[1]                       # 0
x[length(x)]              # 10
x[seq(1, length(x), 3)]   # jedes dritte Element von x
```

Tests von Vektoren, logische Indizierung

Elemente von Vektoren können auf Eigenschaften getestet werden.
Wie viele von 20 gleichverteilten Zufallszahlen auf $[0,1]$ sind >0.5 ?

```
x = runif(20)           # 20 Zufallszahlen zwischen 0 und 1
gr = x>0.5             # Vektor aus TRUE und FALSE
sum(gr)                # Anzahl der TRUE-Werte
y = x[gr]              # x-Werte >0.5 nach y extrahieren
mean(y)                # Durchschnitt dieser Werte
mean(x[x>0.5])        # dasselbe! Lies:
                       # Mittelwert der x-Werte, für die x>0.5
```

Merke:

Vektor-Indizierung kann auf zwei Arten erfolgen:

- numerisch (mittels Zahlenvektor: `x[c(1,2,3)]`)
- logisch (mittels T/F-Vektor: `x[c(T,T,T,F,F,F)]`)

Runde und Eckige Klammern

- **Runde Klammern: Funktionsaufruf**
 - mean ist eine Funktion; wird angewendet auf einen Vektor
 - mean(x) liefert Mittelwert des Vektors x
 - runif ist eine Funktion; wird angewendet auf eine natürliche Zahl n
 - runif(n) liefert n zufällige Zahlen im Intervall $[0,1[$.
- **Eckige Klammern: Indizierung**
 - z.B. sei x ein Vektor; das k-te Element bekommt man mit $x[k]$
 - Wir wissen: Auch der Index kann ein Vektor sein;
 - dieser kann durch eine Funktion wie seq oder c erzeugt werden:
 - $x[\text{seq}(1,10,2)]$ oder $x[\text{c}(1,3,5,7,9)]$

Funktionsaufruf mit benannten Parametern

Es ist Konvention, einer Funktion erst die nötigen Daten zu übergeben; danach benannte Parameter, die das Verhalten der Funktion verändern.

Beispiel: Plot mit Linien und Punkten in Blau mit Titel:

```
plot(x,y, type="o", col="blue", main="Ein Funktionsplot")
```

Beispiel: Logarithmische Achsen

Achsen können mit logarithmischer Skala darzustellen.

Dies geschieht (für die y-Achse) durch Angabe des Parameters `log="y"`.

Monome der Form $y=cx^n$ werden bei logarithmischen Achsen zu Geraden.

```
x = seq(1, 5, by=1/16)
y = 5 * x**3
plot(x,y, type="o")
plot(x,y, type="o", log="xy") # beide Achsen logarithmisch
```