

**Einführung in die Angewandte Bioinformatik:
Datenanalyse mit R
Proteinsequenz-Datenbanken
20.05.2010**

Prof. Dr. Sven Rahmann

Fortsetzung

Statistik mit R
Datenanalyse mit R

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
```

Beschreibende Statistik mit R

Sei v ein Vektor von reellen Zahlen.

Wir definieren und berechnen einige Kennzahlen.

Mittelwert	<code>mean(v)</code>	Durchschnitt
Varianz	<code>var(v)</code>	Mittlere quadrat. Abw. zum MW
Standardabweichung	<code>sd(v)</code>	Wurzel aus Varianz
Median	<code>median(v)</code>	50% der Werte kleiner, 50% größer
p -Quantil	<code>quantile(v,p)</code>	Anteil p der Werte kleiner; $1-p$ größer
p -Quantile	<code>quantile(v,p)</code>	p kann Vektor sein!
Interquartilabstand	<code>IQR(v)</code>	75%-Quantil – 25%-Quantil

Beschreibende Statistik mit R

5 Kennzahlen `fivenum(v)` # dazu ?in R `fivenum` lesen

- Minimum (oder mindestens $\text{Median} - 1.5 \cdot \text{IQR}$)
- 25%-Quantil
- Median
- 75%-Quantil
- Maximum (oder höchstens $\text{Median} + 1.5 \cdot \text{IQR}$)

Boxplot `boxplot(v)` # Visualisierung von `fivenum`

Histogramm `hist(v)` # Anzahl der Zellen sinnvoll gewählt
 `hist(v,n)` # Histogramm mit n Zellen

Ein Histogramm zeigt zu jedem Wert(ebereich) die Anzahl der Elemente in v an, die in diesen Bereich fallen.

Tabellen (data frames) einlesen

Wie kommen die Daten in den R-Workspace ?

- Manuelle Eingabe: langsam, umständlich
- Einlesen aus Datei: schnell, aber Format muss beachtet werden

Aktuelles Verzeichnis anzeigen / wechseln (wie `cd` in der Shell):

`getwd()` / `setwd("neuesVerzeichnis")`

Tabelle aus Datei in Variable `x` einlesen:

`x = read.table(dateiname, ...)`

Optionen dabei z.B.:

`sep = ""` oder `sep=","` oder `sep=";"` (Trennungszeichen der Elemente)

`header = FALSE` oder `header=TRUE` (erste Zeile enthält Namen?)

`col.names = vektor` (Spaltennamen explizit angeben)

Mit Tabellen (data frames) arbeiten

Eingelesen Daten `x` stehen in sog. **data frames** (Datenrahmen = Tabellen).
Jede Zeile ist ein Datensatz; jede Spalte repräsentiert ein Attribut der Daten.

Größe anzeigen lassen: `dim(x)`

Spalten haben Namen (aus Datei oder automatisch): `colnames(x)`

Spalten umbenennen: z.B. `colnames(x) = c("Anna", "Bert")`

Kurzübersicht: `str(x)`

Zugriff auf Zeile 17: `x[17,]` (Komma beachten!)

Zugriff auf mehrere Zeilen 17 bis 23: `x[17:23,]` (Komma beachten!)

Zugriff auf Spalte "Anna" als Vektor: `x$Anna` oder `x[, "Anna"]`

Auf Spalten direkt (ohne Präfix `x$`) zugreifen: `attach(x)`

Tabelle wieder speichern: `write.table(x, Dateiname, ...)`

Erste statistische Analysen mit R

Vergleich von zwei Vektoren x, y gleicher Länge

Scatterplot	<code>plot(x,y)</code>	Punkte $(x[i], y[i])$
Korrelationskoeffizient	<code>cor(x,y)</code>	+ -1 bei linearer Abhängigkeit 0 bei Unabhängigkeit

Komplexes Beispiel einer beschreibenden Datenanalyse

Mit einer neuen Sequenzieretechnologie (ABI SOLiD) wurden kurze nicht-codierende RNA-Stücke (ncRNA reads) sequenziert.

Gegeben

- FASTA-Datei mit ca. 670000 Sequenzen der Länge 35.
- FASTA-ähnliche Datei mit Qualitätswerten ($-10 \cdot \log_{10}(\text{Fehlerwahrscheinlichkeit})$) ebenfalls ca. 670000 Sequenzen à 35 Werte (.qual-Datei)

Fragestellung

Betrachte nur die Qualitätswerte, nicht die Sequenzen.

Nimmt die Qualität „hinten“ ab?

Sind Qualitätswerte an Positionen 30+ niedriger als an den ersten Positionen?

Erste Schritte – Daten anschauen

Gegeben

FASTA-ähnliche Datei mit ca 670000 x 35 Qualitätswerten (.qual-Datei)

In der Shell

```
% ls -l reads.qual           # wie groß ist die Datei?
% head -n 30 reads.qual      # erste 30 Zeilen
% tail -n 30 reads.qual      # letzte 30 Zeilen
% cat reads.qual             # ganze Datei (Unsinn!)
% more reads.qual            # durchblättern
% less reads.qual            # wie more
% wc less.qual               # Wie viele Zeilen?
```

FASTA-ähnliche Datei mit Qualitätswerten

```

...
# Title: s0329_20090331_552to561_613to614_2_552_561
>854_648_594_F3
25 27 27 2 29 30 25 3 2 27 26 27 4 5 25 27 24 2 2 28 28 31 3 3 21 26 30 3 2 27 21 26 5 5 6
>854_824_731_F3
14 20 20 23 23 14 26 20 28 26 18 26 22 26 30 14 25 24 26 28 17 28 18 28 29 7 25 11 26 27 18 21
>854_1300_825_F3
32 28 27 26 24 28 30 27 26 21 24 29 26 30 29 27 30 23 28 29 26 31 26 31 29 21 25 28 19 27 23 24
>855_103_1176_F3
31 25 27 28 32 29 28 28 19 32 30 25 23 20 19 20 21 27 29 28 19 21 26 27 26 28 28 29 19 25 5 30
>855_133_1168_F3
26 32 28 25 15 31 27 29 19 28 27 27 24 24 29 29 24 25 23 28 25 25 8 27 29 25 24 25 13 25 31 8 1
...

```

Es gibt

- Kommentarzeilen (mit #)
- Kopfzeilen (mit >)
- Datenzeilen (Zahlen)

Wir wollen nur die Qualitätswerte (670000 x 35 – Matrix) extrahieren.

Erzeugung einer Datei mit ausschließlich Qualitätswerten

```

% grep -v '#>' reads.qual > q.txt
% head q.txt
% wc q.txt                                     # 676773 Zeilen
  
```

Die Option `-v` bei `grep` invertiert die Logik und gibt die Zeilen aus, die die Bedingung, `#` oder `>` zu enthalten, nicht erfüllen:

```

25 27 27 2 29 30 25 3 2 27 26 27 4 5 25 27 24 2 2 28 28 31 3 3 21 26 30 3 2 27 21 26 5 5 6
14 20 20 23 23 14 26 20 28 26 18 26 22 26 30 14 25 24 26 28 17 28 18 28 29 7 25 11 26 27 18 21
32 28 27 26 24 28 30 27 26 21 24 29 26 30 29 27 30 23 28 29 26 31 26 31 29 21 25 28 19 27 23 24
31 25 27 28 32 29 28 28 19 32 30 25 23 20 19 20 21 27 29 28 19 21 26 27 26 28 28 29 19 25 5 30
26 32 28 25 15 31 27 29 19 28 27 27 24 24 29 29 24 25 23 28 25 25 8 27 29 25 24 25 13 25 31 8 1
29 24 23 27 24 5 8 25 24 32 26 18 27 21 28 23 26 27 22 20 24 21 19 15 25 27 18 24 8 26 24 18 24
28 30 17 32 29 24 28 28 23 24 30 30 26 24 22 27 23 26 21 22 23 25 27 11 18 29 23 27 4 16 28 24
  
```

Einlesen in R

```
% R                                     # Shell: Starten von R

> q = read.table('q.txt')               # R: Einlesen der Datei

> dim(q)                                 # Orientierung: Größe?
[1] 676773      35

> colnames(q)                            # Namen der 35 Spalten: automatisch
 [1] "v1"  "v2"  "v3"  "v4"  "v5"  "v6"  "v7"  "v8"  "v9"  "v10" "v11"
[13] "v13" "v14" "v15" "v16" "v17" "v18" "v19" "v20" "v21" "v22" "v23"
[25] "v25" "v26" "v27" "v28" "v29" "v30" "v31" "v32" "v33" "v34" "v35"
```

Mittlere Qualität und Variabilität jeder Position

```
> m = mean(q)
> s = sd(q)

> plot(m, col="red", type="o")      # Plot der Mittelwerte

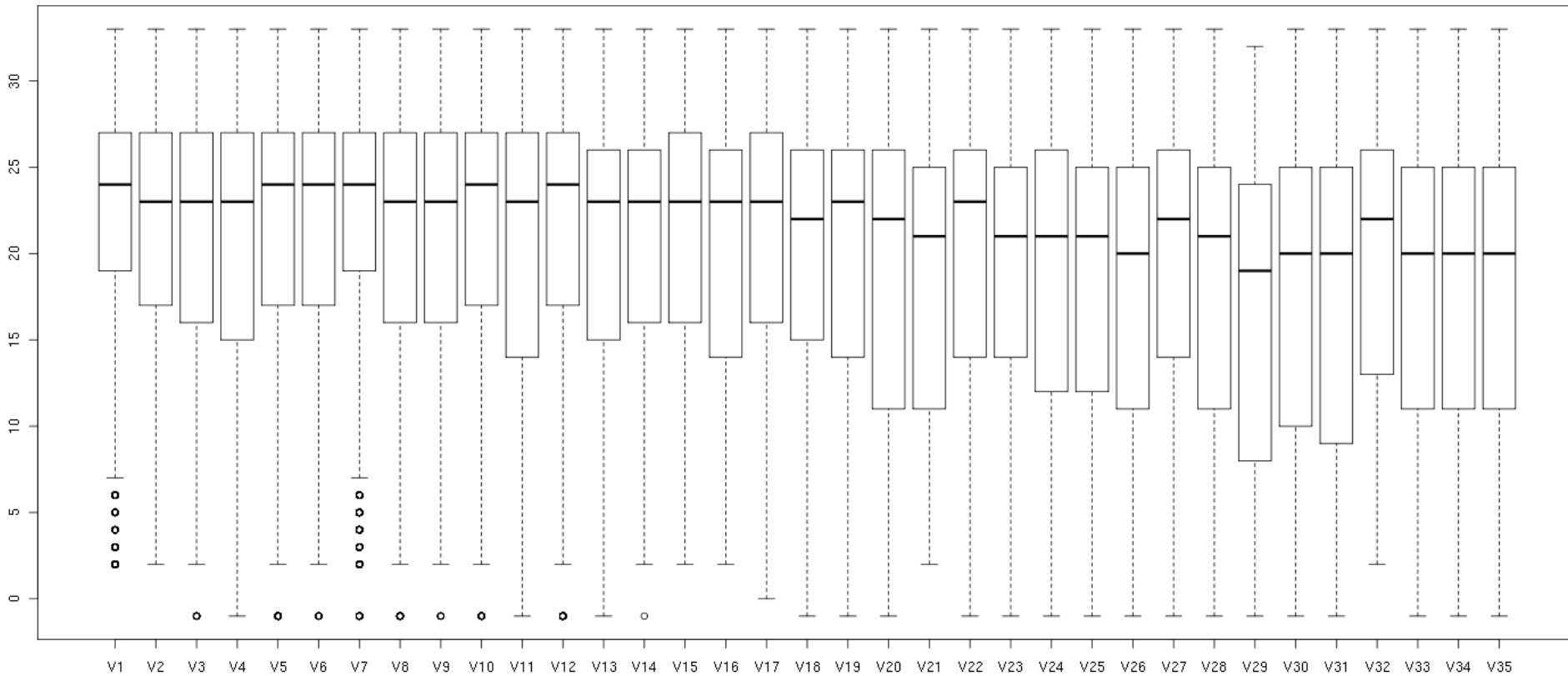
> x11()
> plot(s, pch="+", col="blue", type="o")
```

Merke

- Mit `X11()` kann man ein weiteres Plot-Fenster öffnen.
- Die einfachen beschreibenden Statistiken (wie `mean`, `sd`)
arbeiten spaltenweise auf Tabellen (dataframes)!
- Das funktioniert leider nicht mit `median`, `IQR`.
- Aber mit `boxplot`...

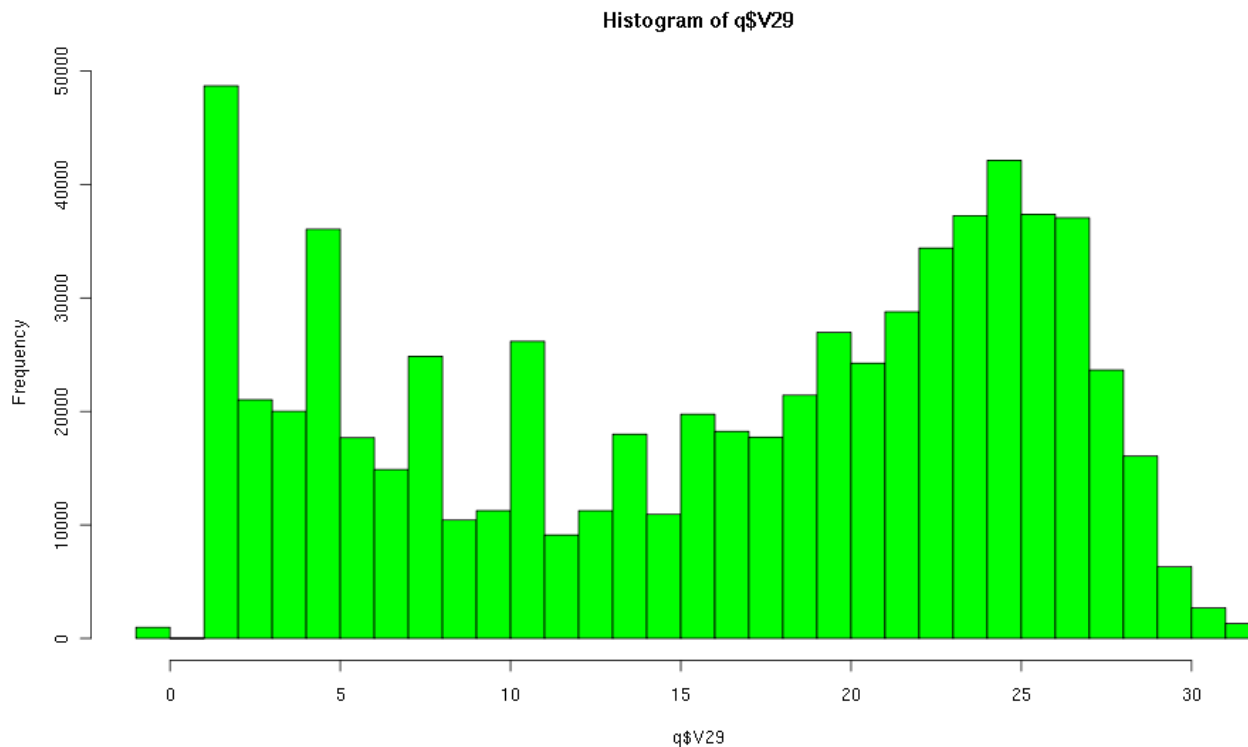
Boxplot der Qualität jeder Position

> boxplot(q) # dauert... Position 29 fällt auf.



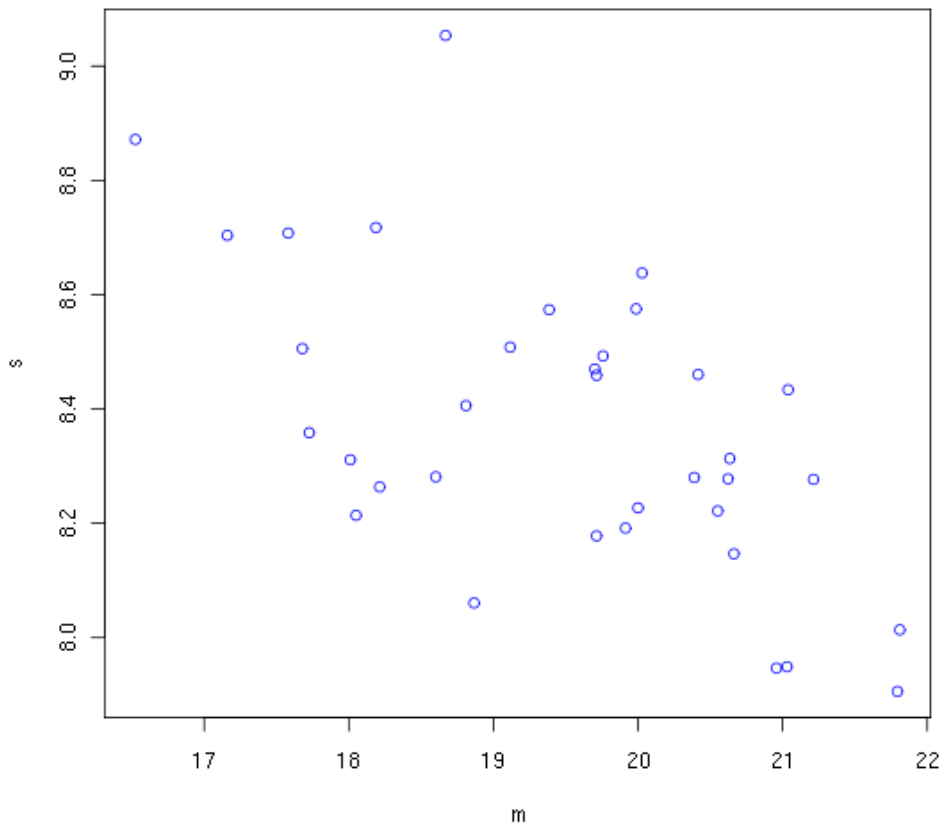
Qualitätshistogramm (Position 29)

```
> attach(q) # spart Tipparbeit  
> min(V29) # min. Qualität an Position 29  
> max(V29) # max. Qualität an Position 29  
> hist(V29, 35, col="green") # Histogramm
```



Korrelieren Mittelwert und Standardabweichung?

```
> plot(m, s, col="blue")      # Scatterplot m gegen s  
> cor(m, s)                  # Tendenz?
```



Neues Thema

Protein-Datenbanken

3 Proteinsequenz-Datenbanksysteme

- NCBI Entrez Proteins
- EBI SRS Proteins
- UniProt (empfohlen)

UniProt

Search in: Protein Knowledgebase (UniProtKB) [dropdown]

Query: [input field]

Core Data

- Protein Knowledgebase (UniProtKB)
- Sequence Clusters (UniRef)
- Sequence Archive (UniParc)

Supporting Data

- Literature citations
- Taxonomy
- Keywords
- Subcellular locations

Information

- News
- Documents
- User manual
- FAQ
- Help

EMBL-EBI EB-eye Search All Databases [dropdown] Enter Text Here [input] Go [button] Reset [button] Give us feedback [button]

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index [RSS icon] [PDF icon]

Quick Search | Library Page | Query Form | Tools | Results | Projects | Views | Databanks | HELP

SRS

[Start a Permanent Project](#)

Tips

★ Want to know more about using SRS?
- go to the [Help Center](#) for online searchable help.
- look in our [SRS@EBI FAQ](#) for answers to commonly

Quick Text Search Search Tips

Find: Proteins [dropdown] matching: Enter Text Here [input]

Search: [input] [button] Search [button]

News [input] Search Tips

Import [input]

Proteins

- Nucleotides
- Proteins
- Structures
- Protein Families
- Literature
- Genome
- Mutations
- Metabolic Pathways

NCBI Entrez Proteins

„Entrez“-Sicht auf verschiedene Datenbanken,
auch Nicht-NCBI-Datenbanken.

Vorteil: in gewohnter Weise mit Entrez durchsuchbar

Sequenz-Inhalt (wie bei UniProt):

- Swiss-Prot
- PIR
- Übersetzungen der kodierenden Nukleotidsequenzen in GenBank und RefSeq.

Weiterer Inhalt aus:

- Literatur: PRF (Protein Research Foundation, Japan)
- Proteinstruktur: wwPDB (world-wide Protein Data Bank, Proteinstrukturen)

Suche in NCBI Entrez Proteins

- Anzeige weiterer Treffer in der Gene-Datenbank
- Anzeige der Organismen mit vielen Treffern

The screenshot shows the NCBI Entrez Protein search interface. The search term 'DtxR' has been entered, and 799 results are displayed. The top results are for Corynebacterium glutamicum ATCC 13032, Corynebacterium diphtheriae NCTC 13129, and Frankia alni ACN14a. The page also features a 'Top Organisms' sidebar and a 'Recent Activity' section.

NCBI Entrez Protein

Search: Protein for DtxR [Go] [Clear] [Save Search]

Display: Summary Show 20 Sort By Send to

All: 799 Bacteria: 698 RefSeq: 275 Related Structures: 634

Items 1 - 20 of 799 Page 1 of 40 Next

This search in Gene shows [283 results](#), including:

- [dtxR](#) (*Corynebacterium glutamicum* ATCC 13032): IRON dependent regulatory protein-DTXR-like protein
- [dtxR](#) (*Corynebacterium diphtheriae* NCTC 13129): diphtheria toxin repressor
- [dtxR](#) (*Frankia alni* ACN14a): Iron-dependent repressor

1: [YP_250879](#) Reports Conserved Domains, BLink, Links
iron-dependent repressor DtxR [*Corynebacterium jeikeium* K411]
gi|68536174|ref|YP_250879.1|[68536174]

2: [ZP_03933519](#) Reports Conserved Domains, BLink, Links
iron-dependent repressor DtxR [*Corynebacterium accolens* ATCC 49725]
gi|227503470|ref|ZP_03933519.1|[227503470]

3: [EEI13936](#) Reports Conserved Domains, BLink, Links
iron-dependent repressor DtxR [*Corynebacterium accolens* ATCC 49725]
gi|227075973|gb|EEI13936.1|[227075973]

Top Organisms [Tree]

- [Corynebacterium diphtheriae](#) (63)
- [Mycobacterium tuberculosis](#) (27)
- [Bacteroides capillosus](#) ATCC 29799 (20)
- [Mycobacterium bovis](#) (9)
- [Clostridium botetoe](#) ATCC BAA-613 (8)
- All other taxa (688)

Recent Activity

Turn Off Clear

🔍 DtxR (799) Protein

Suchergebnisse in NCBI Entrez Proteins

- Anzeige im GenPept-Format (wie GenBank-Format), andere möglich.
- Viele Tools und Links auf der rechten Seite (BLAST, CD || Struktur, Literatur...).

Format: **GenPept** [FASTA](#) [Graphics](#) [More Formats](#) ▼

[Download](#) ▼

[Save](#) ▼

[Links](#) ▼

NCBI Reference Sequence: YP_250879.1

iron-dependent repressor DtxR [*Corynebacterium jeikeium* K411]

Change Region Shown 

[BLAST Sequence](#)

Find regions of similarity between this sequence and other sequences using BLAST.

[Conserved Domains](#)

View conserved domains detected in this protein sequence using CD-search.

[Articles about dtxR](#)

- ▶ Complete genome sequence and analysis of the multire: [J Bacteriol. 2005]

» See all...

[More about the gene dtxR](#)

Also Known As: jk1097

[Comment](#) [Features](#) [Sequence](#)

LOCUS YP_250879 240 aa linear BCT 26-APR-2009
 DEFINITION iron-dependent repressor DtxR [*Corynebacterium jeikeium* K411].
 ACCESSION YP_250879
 VERSION YP_250879.1 GI:68536174
 DBSOURCE REFSEQ: accession [NC 007164.1](#)
 KEYWORDS .
 SOURCE *Corynebacterium jeikeium* K411
 ORGANISM [Corynebacterium jeikeium K411](#)
 Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
 Corynebacterineae; Corynebacteriaceae; Corynebacterium.
 REFERENCE 1 (residues 1 to 240)
 AUTHORS Tauch,A., Kaiser,O., Hain,T., Goesmann,A., Weisshaar,B.,
 Albersmeier,A., Bekel,T., Bischoff,N., Brune,I., Chakraborty,T.,
 Kalinowski,J., Meyer,F., Rupp,O., Schneiker,S., Viehoyer,P. and
 Puhler,A.
 TITLE Complete genome sequence and analysis of the multiresistant
 nosocomial pathogen *Corynebacterium jeikeium* K411, a

Wichtige Suchfelder für NCBI Proteins

[accession]

[gene name]

[protein name]

[EC/RN number]

[organism]

[molecular weight] (Bereichssuche, z.B. 1000:1500)

[sequence length] (Bereichssuche)

...

Universal Protein Resource (UniProt)

<http://www.uniprot.org>

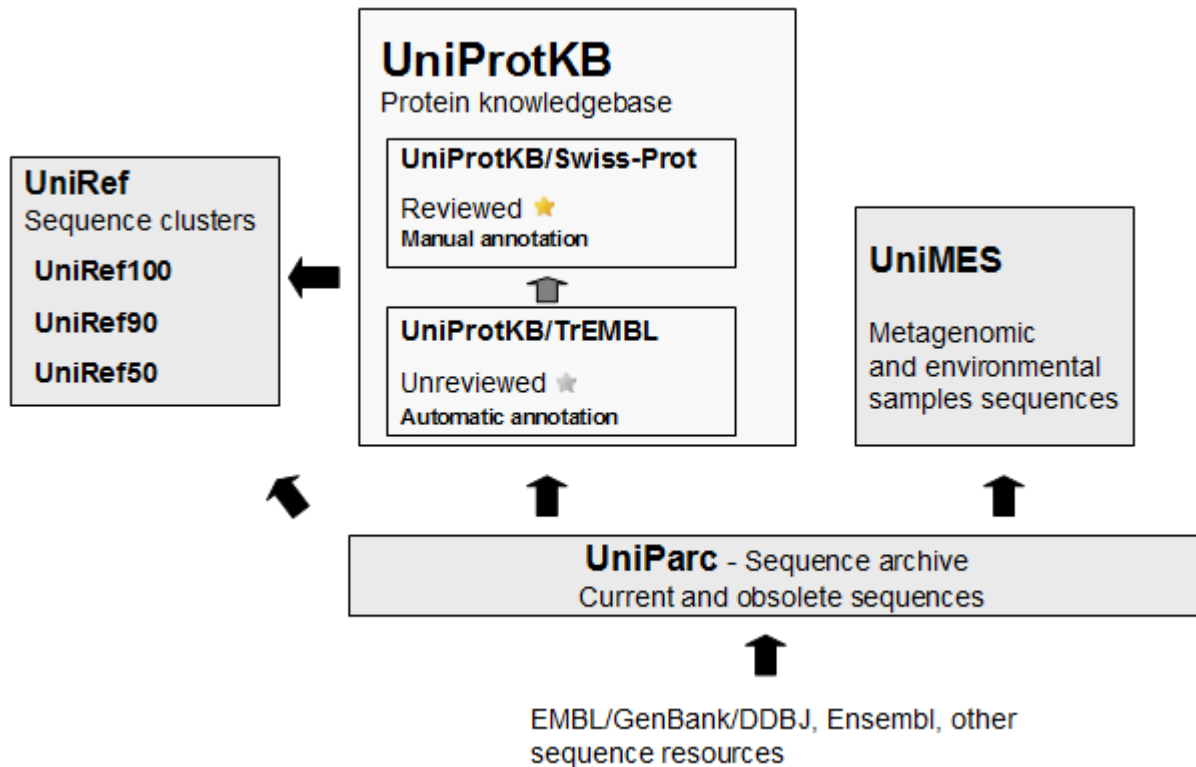
UniProt besteht aus vier Datenbanken

- Protein Knowledgebase (UniProtKB) – bestehend aus
 - TrEMBL – Translated EMBL (übersetzte EMBL-Nukleotideinträge)
 - Swiss-Prot (von Hand annotierte Protein-Datenbank)
- Sequenz-Cluster (UniRef) – repräsentative Sequenzen für Proteinfamilien
- Sequenz-Archiv (UniParc) – Historie der Proteinsequenzen
- Proteinsequenzen aus Metagenomprojekten (UniMES) (neu)

Universal Protein Resource (UniProt)

<http://www.uniprot.org>

UniProt besteht aus vier Datenbanken



Universal Protein Resource (UniProt)

<http://www.uniprot.org>

Beteiligte Institute und verfügbare Dienste

- European Bioinformatics Institute (EBI) mit
 - TrEMBL (Teildatenbank von UniProtKB)
- Swiss Institute of Bioinformatics (SIB) mit
 - Swiss-Prot (Teildatenbank von UniProtKB)
 - ExPASy Server (Expert Protein Analysis System)
- Georgetown University mit
 - PIR (Protein Information Resource)

UniProtKB-Suche

Achtung:
Die Syntax für die Suche ist anders als bei Entrez:

feldname: "Suchbegriff"

Die Feldnamen rechts funktionieren nicht!

Wichtig sind:
gene: Gen-Name
name: Protein-Name
organism: Organismus

The screenshot shows the UniProt search interface. At the top, there is a 'Search in' dropdown menu set to 'Protein Knowledgebase (UniProtKB)' and a 'Query' input field. Below this, there are two columns: 'Field' and 'Term'. The 'Field' dropdown menu is open, showing a list of search fields: All, Protein name [DE], Gene name [GN], Protein family, Domain, Organism [OS], Taxonomy [OC], Virus host [OH], Organelle [OG], General annotation [CC], Sequence annotation [FT], Interacts with, Keyword [KW], Subcellular Location, Gene Ontology (GO), and Enzyme classification (EC). The 'Term' input field is empty. Below the search interface, there is a text box containing information about UniProtKB, including a star icon and the text: 'protein knowledgebase, consists of two sections: ★ Swiss-Prot, which is manually annotated and reviewed. ★ TrEMBL, which is automatically annotated and is not reviewed.'

UniProtKB-Suche: Beispiel

UniProt > UniProtKB Downloads · Contact · Documentation/Help

Search in: Protein Knowledgebase (UniProtKB) Query: `gene:dtxr AND organism:"Corynebacterium jeikeium"` Search Clear Fields »

Search
Blast
Align
Retrieve
ID Mapping *

1 result for `gene:dtxr` AND `organism:"Corynebacterium jeikeium"` in UniProtKB

Reduce sequence redundancy to 100%, 90% or 50% | Customize display

Download...

Page 1 of 1

All	Accession	Entry name	Status	Protein names	Gene names	Organism	Length
<input type="checkbox"/>	Q4JV96	DTXR_CORJK	★	Diphtheria toxin repressor (Iron-dependent diphtheria toxin regulatory element) (Tox regulatory factor)	dtxR (jk1097)	Corynebacterium jeikeium (strain K411)	240

Page 1 of 1

UniProtKB-Suche: Beispiel

Datenbankeintrag Q4JV96

<http://www.uniprot.org/uniprot/Q4JV96>

- Names and Origin (Protein-, Gen-Name, Organismus)
- Protein attributes (Länge, Zustand der Sequenz, Existenz des Proteins)
- General annotation (Funktion, Ort, Ähnlichkeiten)
- Ontologies (kontrollierte Stichwörter, GO-Terme)
- Sequence annotation / Features (bekannte Proteinketten, Domänen)
- Sequences (Proteinsequenz, Zugriff auf Tools)
- References (Literaturangaben)
- Cross-references (Verweise auf andere Datenbanken)
- Entry information (Geschichte dieses Datenbank-Eintrags)

UniProtKB-Suche: Beispiel Q4JV96 - Cross References

Beispiele für in UniProt
verlinkte Datenbanken:

- RefSeq
- PDB – 3D-Proteinstrukturen
- PDBSum – graphische Übersicht über
in PDB enthaltene Strukturen
- SWISS-2DPAGE – Lage des Proteins
in 2D-Gelen
- KEGG – Reaktionswege (Pathways)
mit Beteiligung des Proteins
[Kyoto Encyclopedia
of Genes and Genomes]
- Pfam – Proteindomänen-Familien

Cross-references

Sequence databases

EMBL	CR931997 Genomic DNA. Translation: CAI37261.1 .
RefSeq	YP_250879.1 .

3D structure databases

ModBase	Search...
---------	---------------------------

Genome annotation databases

GeneID	3433720 .
GenomeReviews	Gene locus jk1097 in contig CR931997_GR .
KEGG	cjk:jk1097 .
NMPDR	figl306537.3.peg.1020 .

Organism-specific databases

CMR	Search...
-----	---------------------------

Phylogenomic databases

HOGENOM	Q4JV96 .
OMA	Q4JV96 . KVHDEAC.

Enzyme and pathway databases

BioCyc	CJEI306537:JK1097-MON .
--------	---

Family and domain databases

InterPro	IPR001367 . HTH_DbxR. IPR011991 . Wing_hlx_DNA_bd. [Graphical view]
Gene3D	G3DSA:1.10.10.10 . Wing_hlx_DNA_bd. 1 hit.
Pfam	PF02742 . Fe_dep_repr_C. 1 hit. PF01325 . Fe_dep_repress. 1 hit.

UniProtKB-Suche: Beispiel Q4JV96 - Tools

Sequences

Sequence	Length	Mass (Da)	Tools
<input type="checkbox"/> Q4JV96-1 [UniParc]. Last modified August 2, 2005. Version 1. Checksum: 74319E88EC0B1A4C	240	26,964	Blast ProtParam Compute pI/MW ProtScale PeptideMass PeptideCutter
<pre> 10 20 30 40 50 MRDLVDTTEM YLRTIYELEE EGIPPLRARI AERLDQSGPT VSQTVARMER DEL 70 80 90 100 110 120 SLKLSAQGRA LATAVMRKHR LAERLLTDVI GLPWEKVHDE ACRWEHVMGD EVEVQLVKVL 130 140 150 160 170 180 SEYATSPFGN PIPGLDELM E GIPDSERAEL QOKIDNLQVV TSQRASDIEP PEPIQVKILS 190 200 210 220 230 240 INEIIQVEHK LMAKFHALGM RPGSVVDLVA TEDGLEFSND NGAMVVPEEL GHAVRVEKVN </pre>			
« Hide			

föhren zum
BLAST-Server
des NCBI
bzw. zum
ExPASy-Server
des SIB

References

ExPASy (Expert Protein Analysis System) - ProtParam

Zahlreiche Protein-Parameter

(Auswahl der gesamten Proteinkette (chain))

- Länge
- Molekulargewicht (Masse)
- Theoretischer pI (isoelektrischer Punkt)
- Aminosäureverteilung (mit Ladungsverteilung)
- Atomare Zusammensetzung, chemische Summenformel
- Extinktionskoeffizient
- Halbwertszeit (Proteindegradationsrate)
- Instabilitätsindex

...

ExPASy Hauptseite (<http://www.expasy.org/>) - Tools

Site Map

Search ExPASy

Contact us

Search for



ExPASy Proteomics Server

The ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE ([Disclaimer](#) / [References](#) / [Linking to ExPASy](#)).

[\[Databases\]](#) [\[Tools & Software\]](#) [\[Education & Services\]](#) [\[Links\]](#)
[\[Announcements\]](#) [\[Mirror Sites\]](#) [\[Job openings\]](#)

Databases

- [UniProt Knowledgebase \(Swiss-Prot and TrEMBL\)](#) - Protein knowledgebase
- [ViralZone](#) - Portal to viral UniProtKB/Swiss-Prot entries new
- [PROSITE](#) - Protein families and domains
- [SWISS-2DPAGE](#) - Two-dimensional polyacrylamide gel electrophoresis
- [World-2DPAGE Repository](#) - A public standards-compliant repository for gel-based proteomics data published in the literature
- [MIAPEGelDB](#) - A public repository for MIAPE Gel electrophoresis documents
- [ENZYME](#) - Enzyme nomenclature
- [UniPathway](#) - Metabolic pathways
- [SWISS-MODEL Repository](#) - Automatically generated protein models

Tools and software packages

- [Proteomics and sequence analysis tools](#)
 - Identification and characterization ([Aldente](#), [FindMod](#), [Popitam](#), [Phenyx](#), [pl/Mw](#), [ProtParam](#)...)
 - DNA -> Protein
 - Similarity searches ([BLAST](#)...)
 - Pattern and profile searches ([ScanProsite](#)...)
 - Post-translational modification and topology prediction
 - Primary structure analysis
 - Secondary and tertiary structure tools ([Swiss-PdbViewer](#)...)
 - Alignment and Phylogenetic analysis
- [Melanie / ImageMaster](#) - Software for 2-D PAGE analysis
- [MSight](#) - Mass Spectrometry Imager
- [Roche Applied Science's Biochemical Pathways](#)

Einzelne Tools in der Übung!

Proteindomänen und Proteinfamilien

Domänen

sind wiederkehrende modulare Bausteine von Proteinen.
Durch verschiedene Kombinationen von Domänen
entstehen Proteine mit unterschiedlichen Eigenschaften.
Ziel: alle existierenden Domänen katalogisieren, analysieren

Proteinfamilien

Eine Domäne oder eine bestimmte Kombination von Domänen
kann eine bestimmte Familie von Proteinen charakterisieren.

Datenbanken zu Domänen

Pfam: <http://pfam.sanger.ac.uk/>

protein families

SMART: <http://smart.embl-heidelberg.de/>

simple modular architecture
research tool

Modellierung von Proteindomänen

Wie kann man eine Domäne beschreiben?

- Aminosäuresequenz (Konsensus + Variationsmöglichkeiten)
Sequenz angeben, evtl. mehrere Symbole pro Position
(nicht sehr nützlich wg. Variationen)
- multiples Alignment aus bekannten Beispiel-Sequenzen
Beispiel folgt
- statistisches Modell (Hidden-Markov Model, HMM)
Beispiel folgt

Beschreibung durch Multiples Alignment

Beispiel: Serpin-Domäne (Serin Protease Inhibitor)

```

THBG_RAT/38-415      QNATLYKMP SINADFAFRLYRK LSV . ENPDLNIFFPVSI SAALAMLSFGSGSSTQTQILEVLGFNLT DTPVKE . . . .
THBG_HUMAN/35-412   PNATLYKMSSINADFAFNLYRR FTV . ETPDKNIFFPVSI SAALVMLSFGACCSTQT EIVETLGFNLT DTPMVE . . . .
A1AT_RAT/37-409     QSPTYRKISSNLADFAFSLYRE LVH . QSNTSNIFFPVSI ITTAFAMLSLGSKGDTRKQILEGLEFNLTQIPEAD . . . .
A1AT2_MOUSE/37-410 QSPASHEIATNLGDFATSLYRE LVH . QSNTSNIFFPVSI IATAFAMLSLGSKGDHTHTQILEGLQFNLTQTSEAD . . . .
A1AT_BOVIN/41-413   QEAACHKIAPNLANFAFSIYHH LAH . QSNTSNIFFPVSI ASAFAMLSLGAKNTHTEILKGLGFNLT ELAEAE . . . .
A1AT_HUMAN/43-415   DHPFTFNKIPNLAFAFSLYRQ LAH . QSNTSNIFFPVSI IATAFAMLSLGTKADTHDEILEGLNFNLT EIPFAQ . . . .
A1AF_RABIT/38-410   DHPACHRIAPSLAEFALSYRE VAH . ESNTTNIFFPVSI ALAFAMLSLGAKGDHTHTQVLEGLKFNLT ETAEAE . . . .
A1AF_CAVPO/28-400   AQQPSQIIPRSLAHFAHMYRV LTQ . QSNTSNIFFPVSI IATALAMVSLGAKGDHTHTQILWGLEFNLT EIAEAD . . . .
A1AT_DIDMA/36-407   EYSSTRRI SPYMTDFSIDFYRL LVS . KSNTTNIFFPVSI IYTAFTLLALGAKSATRDQILTGLRFNRTE ISEEH . . . .
A1ATR_HUMAN/46-417 EDLACQKISYNVTDLAFDLYKSWLIY . . . . HNQHVLVPTPSVAMAFRMLSLGTKADTRTEILEGLNFNLT ETPEAK . . . .
AACT_HUMAN/45-420   VD . . . . LGLASANVDFAFSLYKQ LVL . KAPDKNVVIFSPVLSI STALAFSLGAHNNTLLEILKGLKFNLT ETSEAE . . . .
CPI6_RAT/42-417     LDS . . . . LTLASINTDFAFSLYKK LAL . RNPDKNVVIFSPVLSI SAALAVVSLGAKGNSMEEILEGLKFNLT ETPETE . . . .
SPA3C_MOUSE/42-414 LDS . . . . LTLASINTDFAFSLYKK LAL . KNPDTNIVFSPVLSI SAALAVVSLGAKGNTLEEILEGLNFNLT ETPEAD . . . .
SPA3K_MOUSE/43-417 DDS . . . . LTLASVNTDFAFSLYKK LAL . KNPDTNIVFSPVLSI SAALAVVSLGAKGKTMEEILEGLKFNLT ETPPAD . . . .
CPI1_RAT/40-415     LHS . . . . LTLASINTDFATLSLYKK LAL . RNPDKNVVIFSPVLSI SAALAILSLGAKDSTMEEILEVLKFNLT EITEEE . . . .
IPSP_HUMAN/34-406   LHV GATVAPSSRRDFTFDLYRA LAS . AAPSQNIFFPVSI SMALAMLSLGA SSKMQLLEGLGLNLTQKSSEKE . . . .
CBG_MOUSE/27-396    DSSSHRDLAPTNDFAFNLYKR LVA . LNSDKNTLISPVSI SMALAMLSLSTRGST . QYLENLGFNM SKMSEAE . . . .
CBG_RAT/27-395      SSNSHRGLAPTNDFAFNLYQR LVA . LNPDKNTLISPVSI SMALAMVSLGS . . . . AQTQSLSLGFNLT ETSEAE . . . .
CBG_HUMAN/32-404    MSNHHRGLASANVDFAFSLYKH LVA . LSPKKNIFISPVSI SMALAMLSLGTGCHTRAQLLQGLGFNLT ERSETE . . . .
CBG_RABIT/10-382    TRSPPRGLAPANVDFAFSLYRQ LVS . SAPDRNICISPVSVSMALAMLSLGA SGTHTRTQLLQGLGFNLT EMPEAE . . . .
EP45_XENLA/61-432   LTKEEKILSEENSDFSVNLFNQL STESKRSPRKNIFFPVSI SAAFYMLALGAKSETHQQLKGLSFNKKKLSSEQ . . . .
HEP2_HUMAN/119-496 GKSRIQRINILNAKFAFNLYRV LKDQ . VNTFDNIFIAPVGI STAMGMISLGLKGETHEQVHSILHF KDFVNASSKYEIT . . . .
OVALY_CHICK/1-388   MDS . . . . ISVTNAKFCFDVENE MKV . HHVNEINILYCPVLSI TALAMVYLGARGNTESQMKKVLHFD SITGAGSTTDSQ . . . .
OVAL_CHICK/2-386    GS . . . . IGAASMEFCFDVFEK LKV . HHANENIFYCPIAIMSALAMVYLGAKDSTRTQINKVVRFDKLPFGFGDSIEAQ . . . .
SPB6_HUMAN/1-376    MDV . . . . LAEANGTFAINLLKT LG . . . . KDNSKNVFFSPVSM SCALAMVYMGAKGNTAAQMAQILSFNKS GGGGD . . . .
ILEU_HORSE/1-379    MEQ . . . . I STANTHFAVDLFRA LNE . SDPTGNIFISPLISSALAMIFLGT RGNNTAAQVSKALYFDTVED . . . .
SPB5_HUMAN/1-375    MDA . . . . IQLANSFAVDLFKQ LCE . KEPLGNVLFSPICLSTSLSLAQV GAKGDTANEIGQVLFHFNVD . . . .
ANT3_HUMAN/76-461   TNRRVWELSKANSRFATTFYQH LADS . KNDNDNIFLSPVLSI STAFAMTKLGA CNDTLQQLMEVFKFDTISEKTS DQ . . . .
SERPH_CHICK/23-396 LSDKATTLADRSTTLAFNLYHA MAK . DKNMENILSPVWVASSLGLVSLGG KATTASQAKAVLSADKLNDY . . . .
PRTZ_HORVU/6-395    ATDVRLSIAHQ . TRFALRLRSA ISSNPERAA GNVAFSPVLSLHV ALSLITAGA . AATRDQLVA I LGGGADKELNA . . . .
PRTZ_BOVIN/27-482   PRTZ_BOVIN/27-482 PRTZ_BOVIN/27-482 PRTZ_BOVIN/27-482 PRTZ_BOVIN/27-482 PRTZ_BOVIN/27-482 PRTZ_BOVIN/27-482

```

Beschreibung durch HMM

Hidden Markov Model (HMM)

Stochastisches generatives Modell:

- wird aus gegebenen Beispiel-Sequenzen (Alignment) erstellt
- kann weitere ähnliche Sequenzen generieren
- kann benutzt werden, um zu prüfen, ob eine neue Sequenz zum Modell passt
(Berechne Wahrscheinlichkeit, dass HMM diese Sequenz generiert)

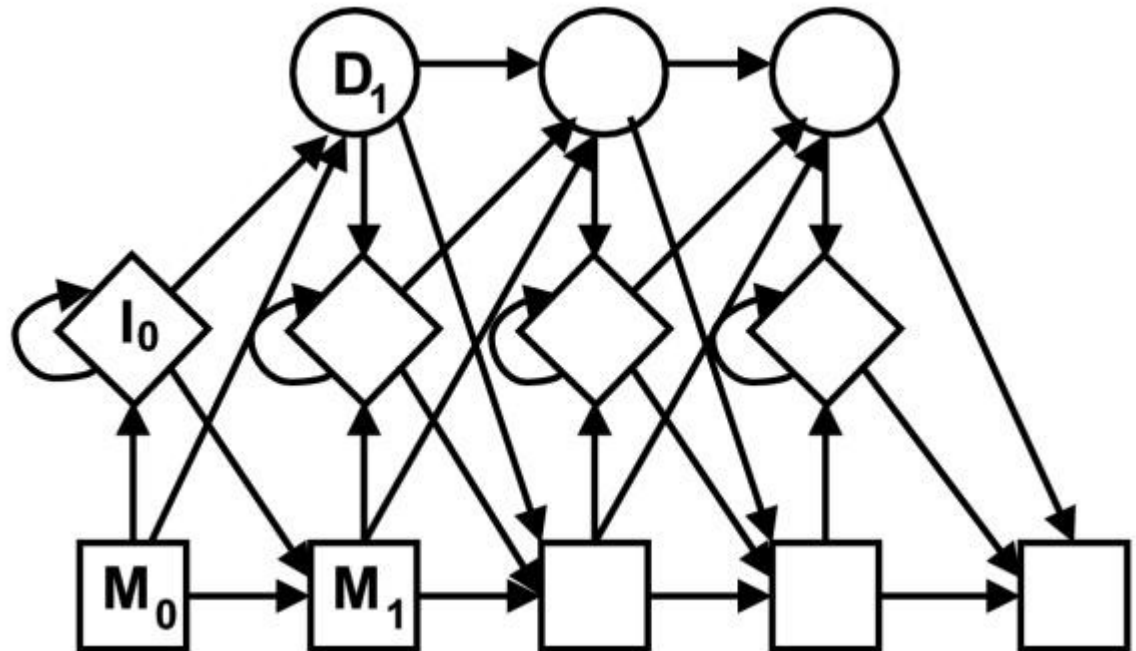
Modellparameter

Für jede Position wird angegeben:

- Wahrscheinlichkeit, Position auszulassen
- Aminosäure-Verteilung (Wahrscheinlichkeiten)
- Wahrscheinlichkeit, dahinter zusätzliche AS einzufügen
- Aminosäure-Verteilung der eingefügten AS

HMM (Profil-HMM)

Visuelle Vorstellung nach
Durbin, Eddy, Krogh,
Mitchison:
Biological Sequence Analysis
Cambridge University Press



Modellparameter

Für jede Position wird angegeben:

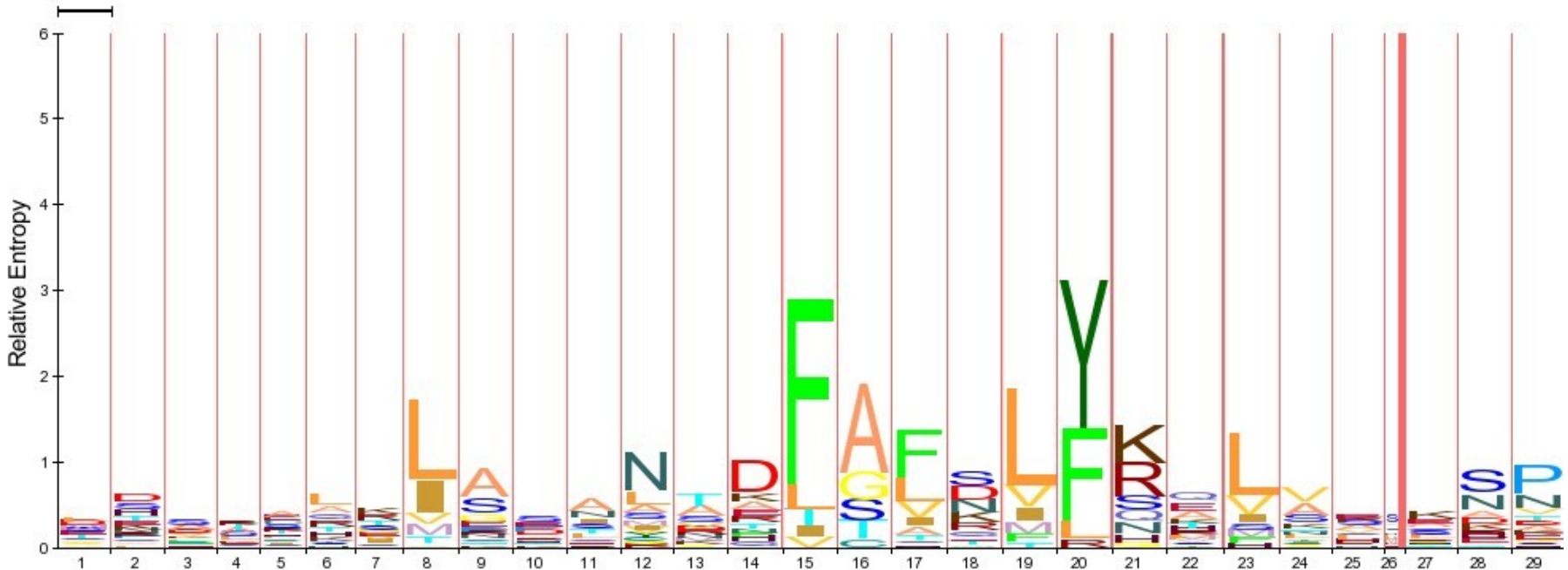
- Wahrscheinlichkeit, Position auszulassen („deletion“ - D)
- Aminosäure-Verteilung (Wahrscheinlichkeiten in M_0, M_1, \dots)
- Wahrscheinlichkeit, dahinter zusätzliche AS einzufügen („insertion“ - I)
- Aminosäure-Verteilung der eingefügten AS (Wahrscheinlichkeiten in I_0, \dots)

Visualisierung durch HMM-Logos

Höhe der Türme: Grad der Konserviertheit

Breite der Türme: Wahrscheinlichkeit, nicht ausgelassen zu werden

Breite der roten Balken: Insertionswahrscheinlichkeit zwischen zwei Positionen



If you use HMM-Logos in your publication, please cite

"Schuster-Boeckler B, Schultz J, Rahmann S: HMM Logos for visualization of protein families. BMC Bioinformatics 2004, 5:7"

The paper is "open access": <http://www.biomedcentral.com/1471-2105/5/7>

Spezielle Protein-Datenbanken

Molecular Class-Specific Information System (MCSIS) project



Available MCSIS

- **The GPCRDB:** a Molecular-Specific Information System for G Protein-Coupled Receptors (created in 1994)
 - [The GPCRDB](#) at the CMBI, the Netherlands
- **The NucleaRDB:** a Molecular-Specific Information System for Nuclear Receptors (created in April 2000)
 - [The NucleaRDB](#) at the CMBI, the Netherlands
 - Mirror site at UCSF, USA (no longer available)
- **The PrionDB:** a Molecular-Specific Information System for Prion proteins (created July, 21 2003)
 - [The PrionDB](#) at the CMBI, the Netherlands
 - Mirror site at UCSF, USA (no longer available)
- **The KChannelDB:** a Molecular-Specific Information System for potassium channels (created July, 25 2003)
 - [The KChannelDB](#) at the CMBI, the Netherlands
 - Mirror site at UCSF, USA (no longer available)
- **The GPCRIPDB:** a Molecular-Specific Information System for GPCR Interacting Partners (G proteins & GPCRs)
 - [The GPCRIPDB](#) at the CMBI, the Netherlands

jeweils auf eine Proteinfamilie zugeschnittene Datenbanken, enthalten viel Expertenwissen, meist von Hand gepflegt:

wenige Einträge, aber qualitativ hochwertig

<http://www.gpcrdb.org>

Zusammenfassung

Protein-Sequenzdatenbanken

- UniProt (<http://www.uniprot.org>)
- Zugriff auf Analyse-Werkzeuge
- Links zu anderen Datenbanken (z.B. Gene, Strukturen)

Protein-Sequenzanalyse

- ExPASy (Expert Protein Analysis System), via UniProt
- Vielzahl von nützlichen Werkzeugen

Modellierung von Proteinen / Domänen

- Hidden Markov Modelle (HMMs)

Ausblick

Ähnlichkeit, Alignments, Homologie

- Sequenzähnlichkeit
- Alignment von homologen Sequenzen
- Homologiesuche: BLAST; Statistik dazu
- BLAST-Statistik
- Multiple Alignments (Clustal)
- Alignment – Algorithmen und Komplexität

Proteinstruktur

- Sekundärstruktur
- Tertiärstruktur (3D)
- Quartärstruktur