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

Beschreibende Statistik mit R

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

- Mimimum (oder mindestens Median – 1.5*IQR)
- 25%-Quantil
- Median
- 75%-Quantil
- Maximum (oder höchstens Median + 1.5* 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	plot(x,y)	Punkte (x[i],y[i])
Korrelationskoeffizient	cor(x,y)	+-1 bei linearer Abhängigkeit 0 bei Unabhängigkeit

Komplexes Beispiel einer beschreibenden Datenanalyse

Mit einer neuen Sequenziertechnologie (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*log10(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")
```

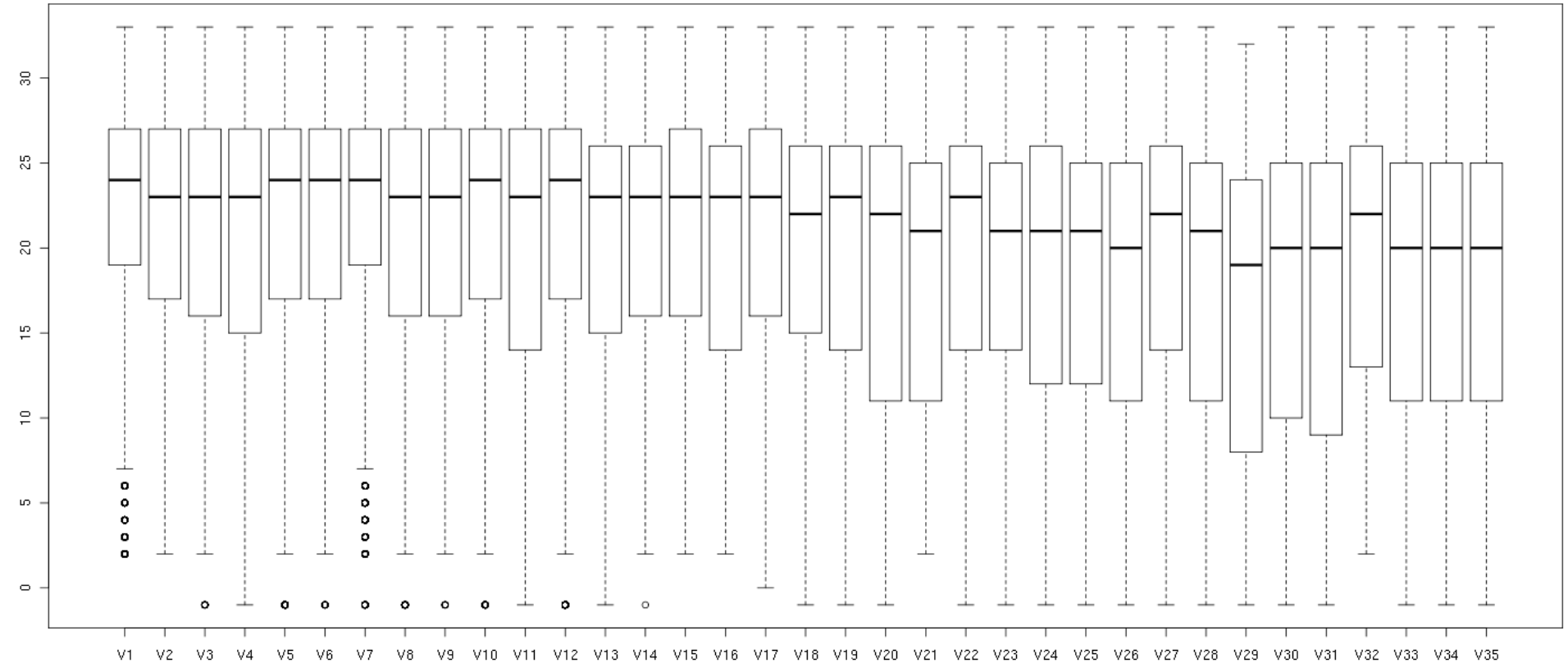
Merke

- 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

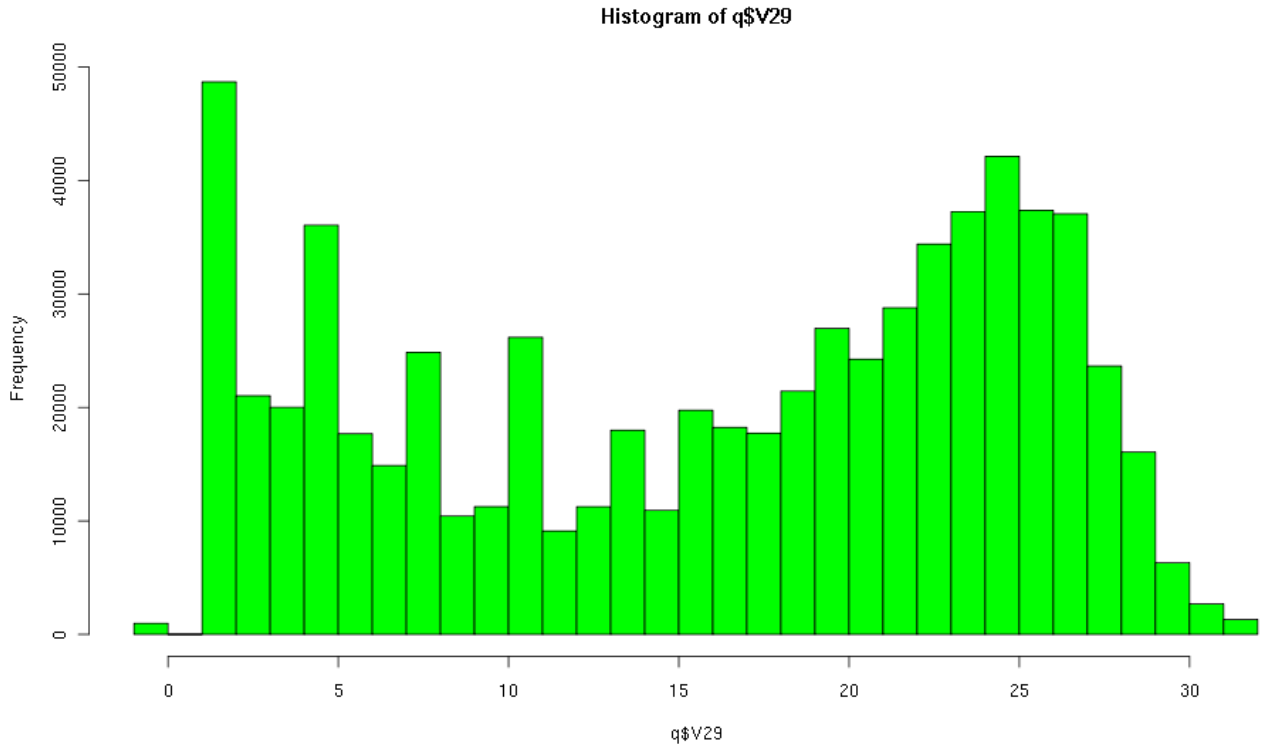
```
> boxplot(q)          # dauert...
```

Position 29 fällt auf.



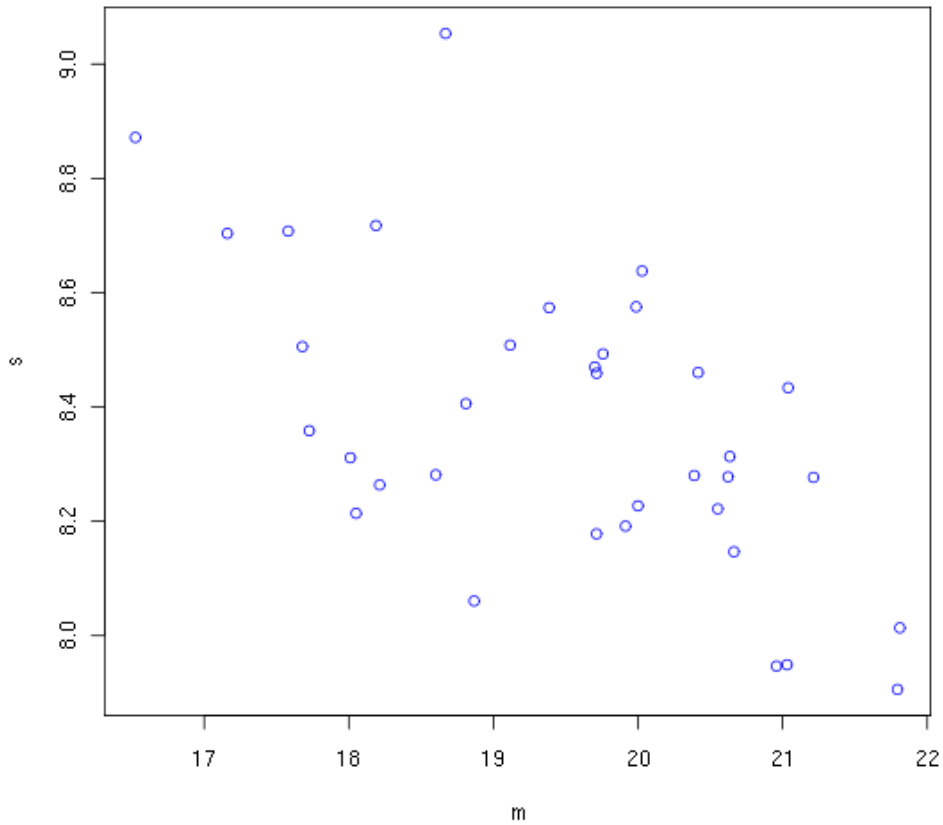
Qualitätshistogram (Position 29)

```
> 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?
```

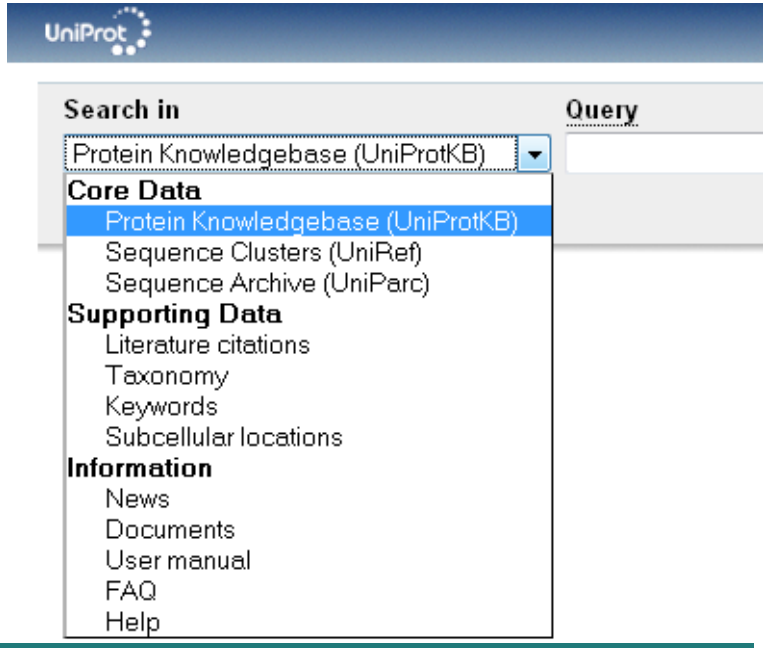


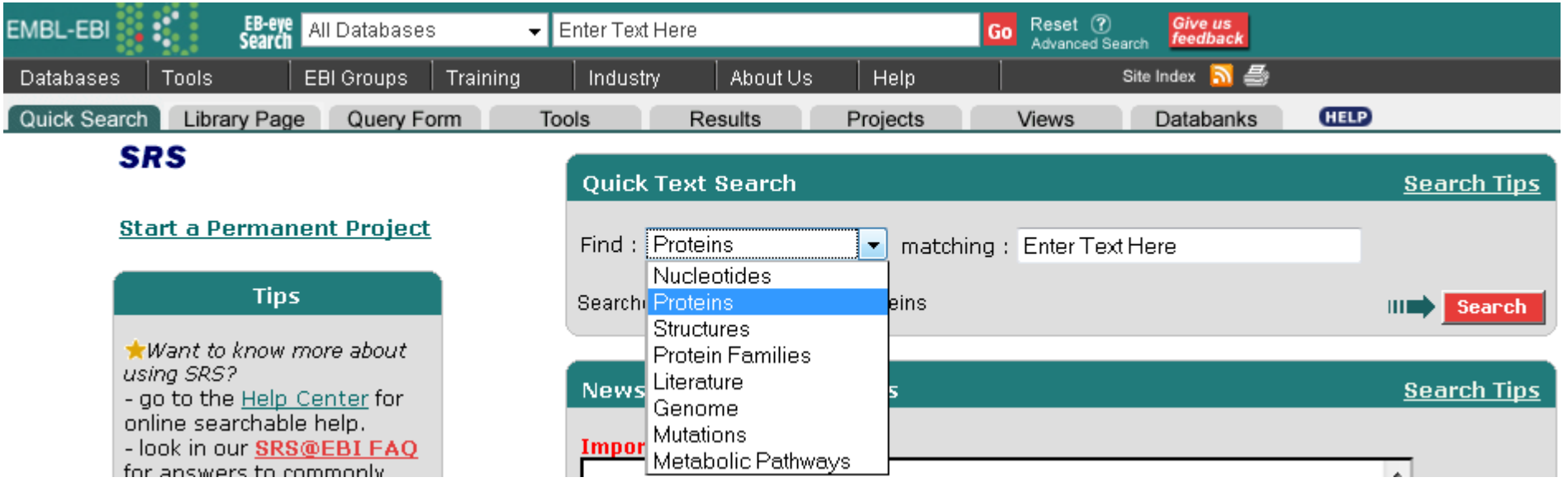
Neues Thema

Protein-Datenbanken

3 Proteinsequenz-Datenbanksysteme

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





NCBI Entrez Proteins

„Entrez“-Sicht auf verschiedene Datenbanken,
auch Nicht-NCBI-Datenbanken.
Vorteil: in gewohnter Weise mit Entrez durchsuchbar

Sequenz-Inhalt (wie bei UniProt):

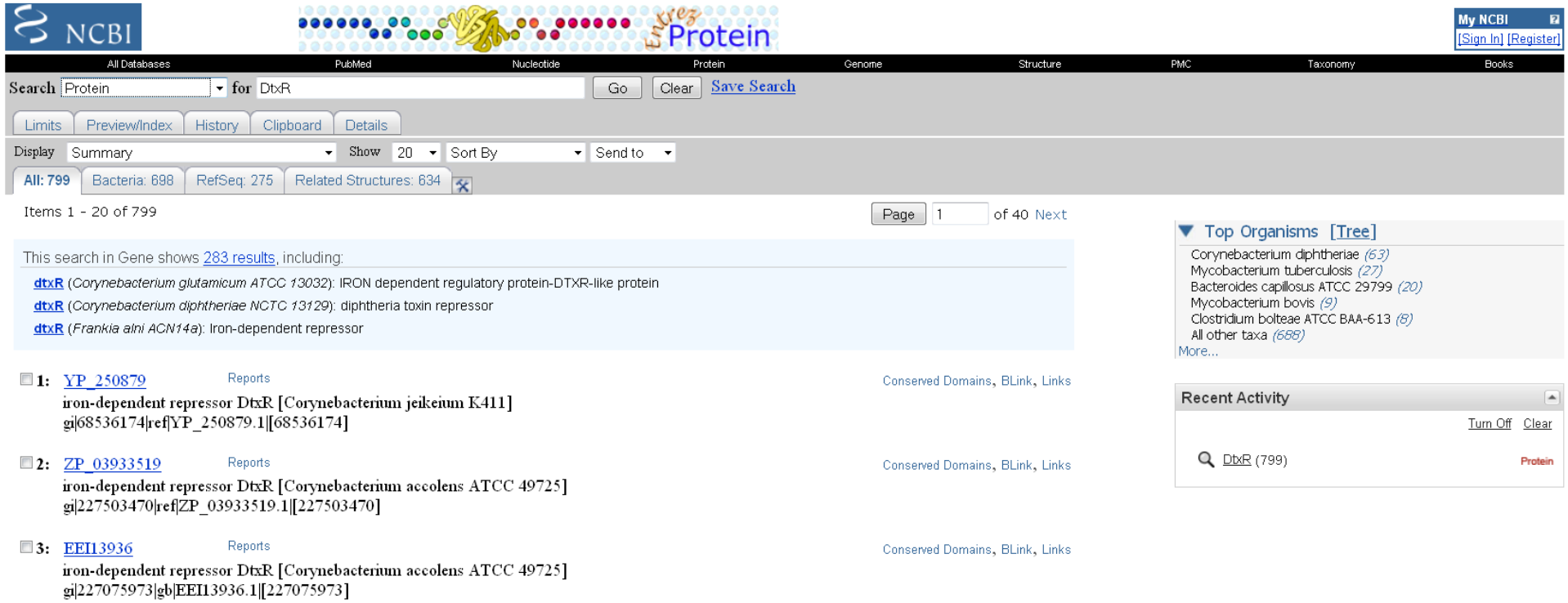
- Swiss-Prot
- PIR
- Übersetzungen der kodierenden Nukleotidsequenzen in GenBank und RefSeq.

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



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

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., Viehoever,P. and
            Puhler,A.
  TITLE     Complete genome sequence and analysis of the multiresistant
            nosocomial pathogen Corynebacterium jeikeium K411, a
```

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

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

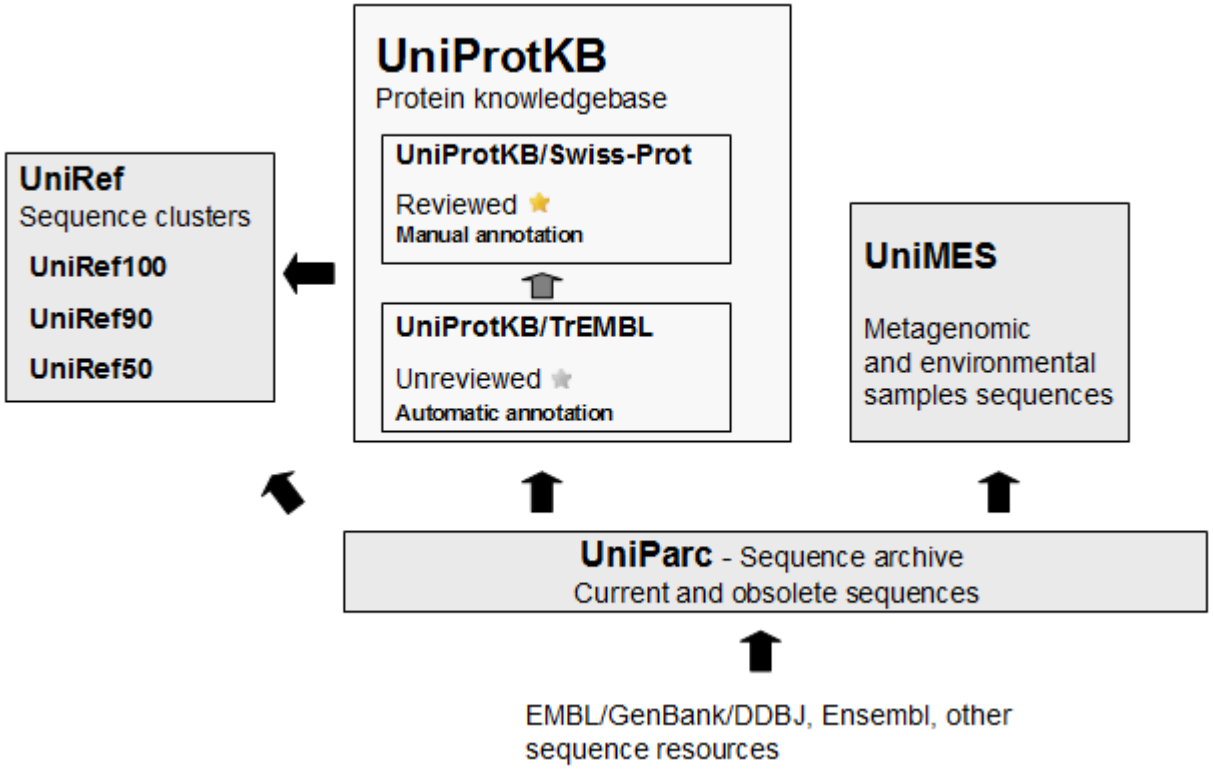
UniProt besteht aus vier Datenbanken

- 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)

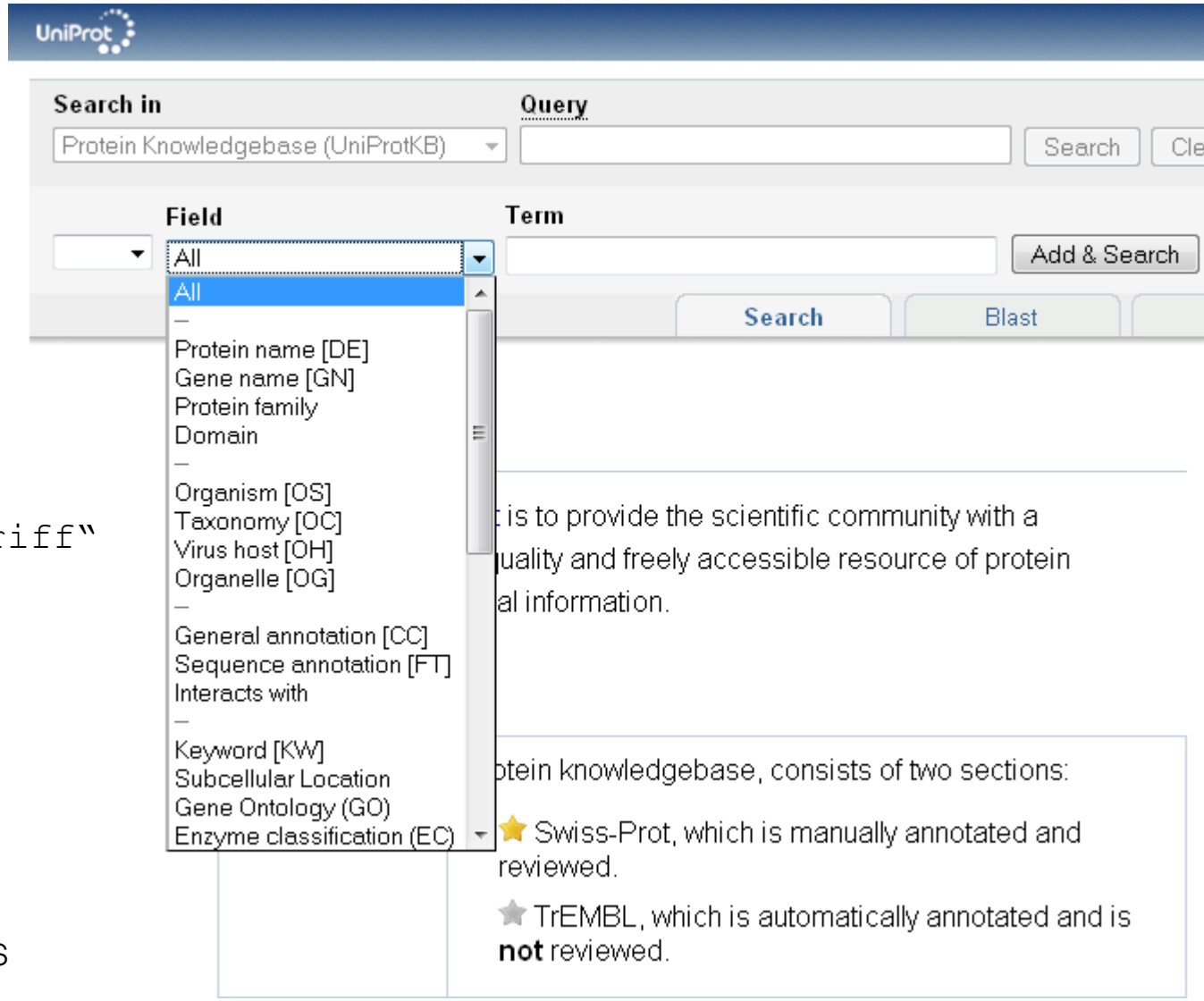
UniProtKB-Suche

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



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
☐	Q4JV96	DTXR_CORJK	★	**Diphtheria toxin repressor** (Iron-dependent diphtheria tox 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

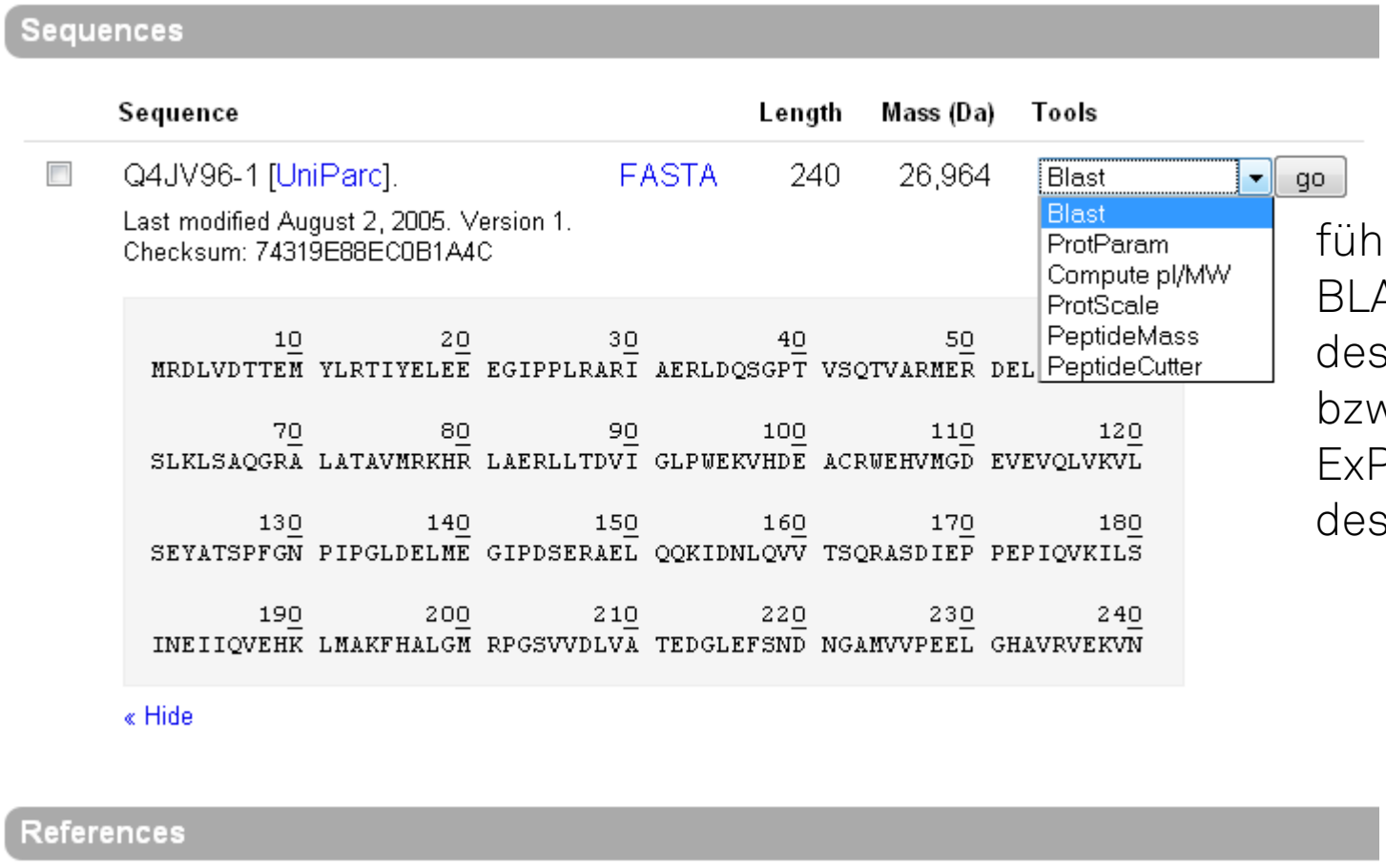
Beispiele für in UniProt verlinkte Datenbanken:

- 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	fig|306537.3.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_DtxR. 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



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

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
- Halbwertzeit (Proteindegradationsrate)
- Instabilitätsindex

...

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

Site Map | Search ExPASy | Contact us

Search Swiss-Prot/TrEMBL for Go Clear

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, pI/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?

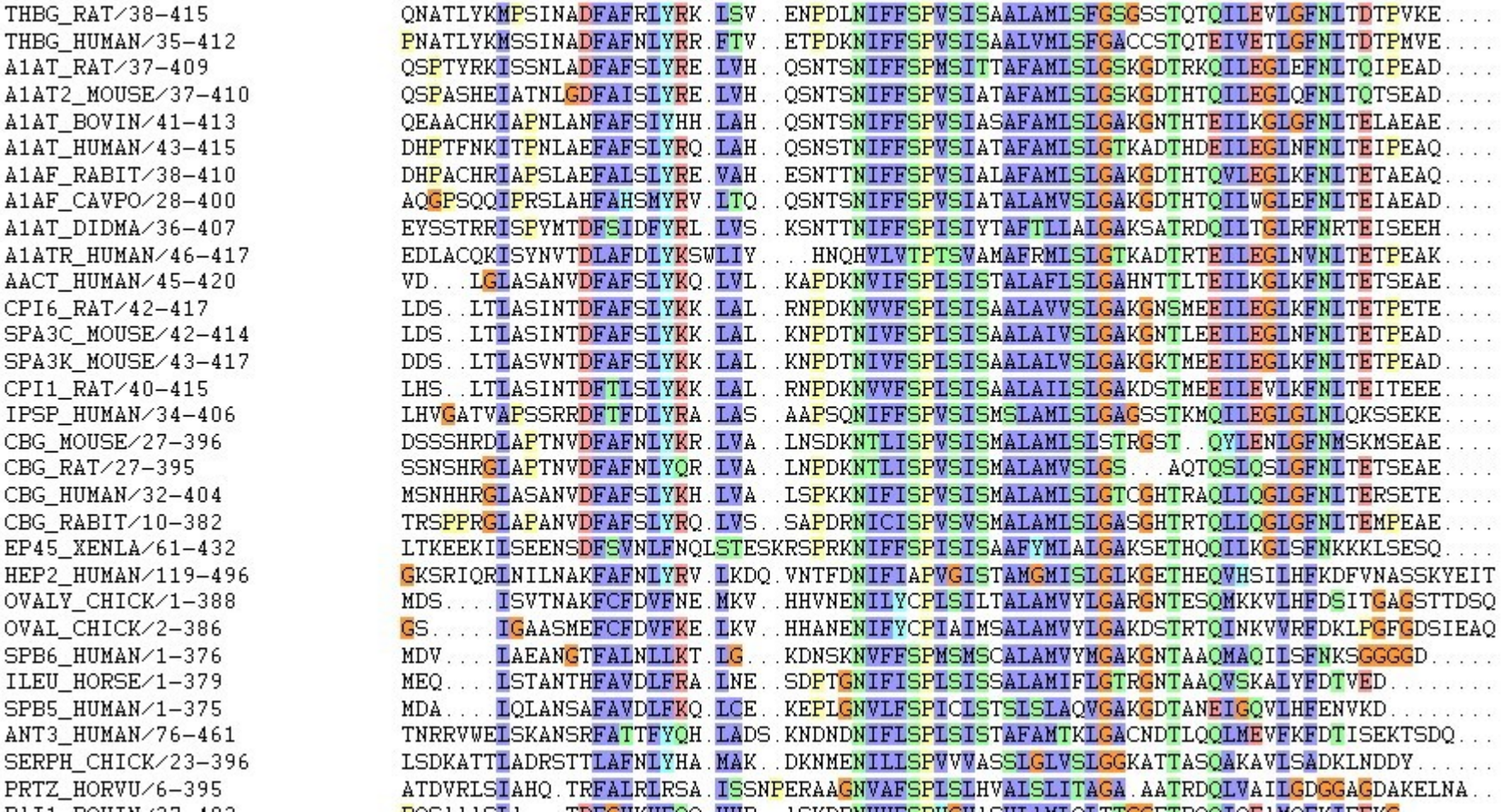
- Aminosäuresequenz (Konsensus + Variationsmöglichkeiten)
  Sequenz angeben, evtl. mehrere Symbole pro Position
  (nicht sehr nützlich wg. Variationen)

- 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)



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)

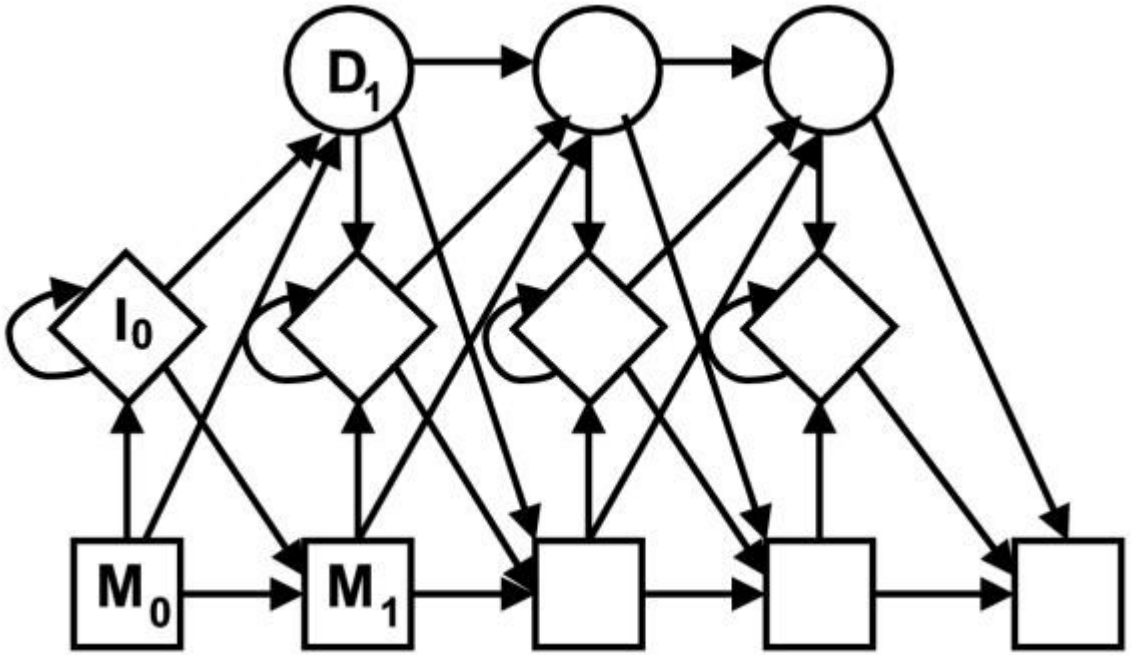
Modellparameter

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



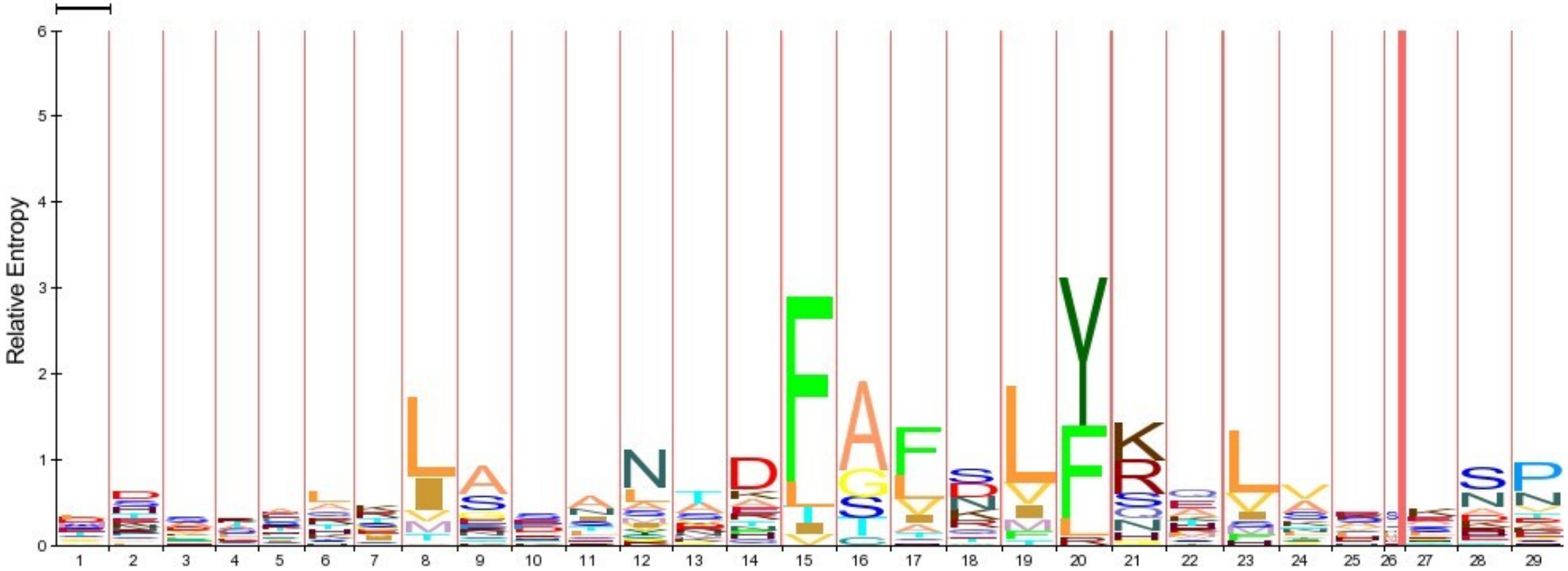
Modellparameter

Für jede Position wird angegeben:

- Wahrscheinlichkeit, Position auszulassen („deletion“ - D)
- Aminosäure-Verteilung (Wahrscheinlichkeiten in M0, M1, ...)
- Wahrscheinlichkeit, dahinter zusätzliche AS einzufügen („insertion“ - I)
- Aminosäure-Verteilung der eingefügten AS (Wahrscheinlichkeiten in I0, ...)

Visualisierung durch HMM-Logos

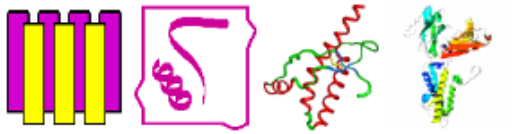
Höhe der Türme: Grad der Konserviertheit
Breite der Türme: Wahrscheinlichkeit, nicht ausgelassen zu werden
Breite der roten Balken: Insertionswahrscheinlichkeit zwischen zwei Positionen



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, 25 200
  - 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 &
  - 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 Datenbaken (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