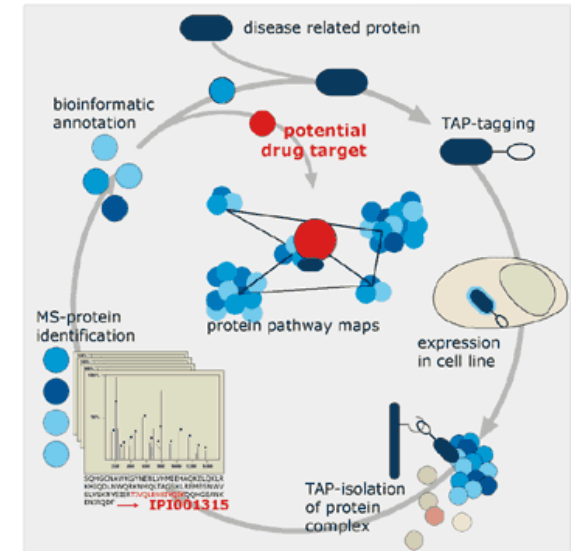


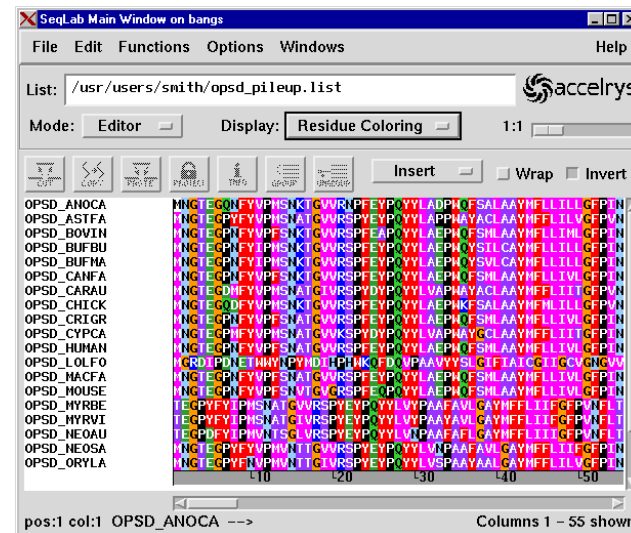
# Softwarewerkzeuge der Bioinformatik

Inhalt dieser Veranstaltung: Softwarewerkzeuge kennenlernen für

- I Sequenzanalyse
- II Analyse von Proteinstruktur und Ligandenbindung
- III Analyse von Omics-Daten, Zell- bzw. Netzwerksimulationen



[www.cellzome.com](http://www.cellzome.com)



[www.accelrys.com](http://www.accelrys.com)

# „Lernziele“

Lerne **aktuelle** und **bewährte Programme** und **Datenbanken** der Bioinformatik kennen und erfolgreich einzusetzen um

- „Hands-On“ mit Web-Tools arbeiten, mit denen man bioinformatische Fragen bearbeiten kann
- zu wissen, was auf dem Markt ist („das Rad nicht zweimal erfinden“)
- ein Gefühl dafür zu bekommen, wie erfolgreiche Softwareprodukte aussehen (sollen)
- 3 Mini-Forschungsprojekte zu bearbeiten (Bioinformatiker/Biotechnologen)



## Organisatorisches

Jede Woche Vorlesung

Donnerstag 10.15 – 12.00 Uhr

(15 minütige Pause)

Seminarraum 007, Geb. E 2 1

Dozent: Prof. Helms

Die Teilnahme an der Vorlesung ist nicht obligatorisch,  
jedoch die Teilnahme an der Übung.

Übungen „**hands-on**“ Beginn **heute** am 19.10:

Donnerstag, 14:00 Uhr – 16:00 Uhr, CIP-Pool E 2 1 CIP.

### Verantwortliche Betreuer der Übungen

**Sequenz-Analyse**

Kerstin Reuter

**Proteinstruktur**

Dr. Michael Hutter

**Zellsimulationen**

Daria Gaidar

# Organisatorisches

Jeder Teilnehmer an den Übungen benötigt einen Rechneraccount für den CIP-Pool.

Biotechnologen: bitte in Liste eintragen



## 4. Pflichten der Benutzer

Der Benutzer verpflichtet sich,

- a) die bereitgestellten Betriebsmittel sorgfältig zu benutzen;
- b) das Passwort des ihm zugeteilten Benutzerkennzeichens geheim zu halten ...;
- ...
- d) alles zu unterlassen, was den ordnungsgemäßen Ablauf der Anlage stört;
- e) in den Arbeitsräumen sich so zu verhalten, dass andere Benutzer nicht gestört werden;
- f) Störungen ... zu melden und diese nicht auszunutzen;
- g) in den Räumen ... sowie bei Inanspruchnahme seiner Geräte ... den Weisungen des Personals des Anlagenbetreibers Folge zu leisten;
- ...
- l) lizenzierte Software nur nach Absprache mit dem jeweiligen BfR einzuspielen und zu verwenden;
- m) von der Fak6 oder der Universität des Saarlandes bereitgestellte Software, Dokumentationen oder Daten weder zu kopieren noch an Dritte weiterzugeben, sofern dies nicht ausdrücklich erlaubt ist, noch zu anderen als den erlaubten Zwecken zu verwenden,

**Zugang zum CIP-Pool während der Übungsstunden.**

# Organisatorisches: Scheinvergabe B.Sc. Bioinformatik und Biotechnologie M.Sc.

- Bewertung: Vorlesung zählt 2V + 2P = 9 Leistungspunkte
- Curriculum: Pflichtvorlesung für die Vertiefung „Bioinformatics“
- kann natürlich auch für CMB-Bachelor eingebracht werden
- Wahlfach Pharmazie/Diplom
- Pflichtvorlesung für bestimmte Studenten des M.Sc. Biotechnologie

Drei Mini-Projekte werden etwa alle 4 Wochen ausgegeben. Diese sind innerhalb von 2 Wochen in Teams mit 2-3 Studenten zu bearbeiten und durch einen mindestens 5-seitigen Praktikumsbericht zu dokumentieren.

Jeder Student muss mindestens zwei der drei Mini-Projekte mit einer Note von 4 und besser bestehen.

# Organisatorisches: Scheinvergabe

## B.Sc. Bioinformatik und Biotechnologie M.Sc.

Voraussetzung für die Teilnahme an der Abschlussklausur ist das Erreichen von mindestens 50 % der maximalen Punkte aus den drei Praktikumsberichten.

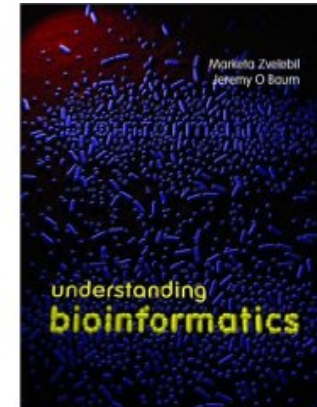
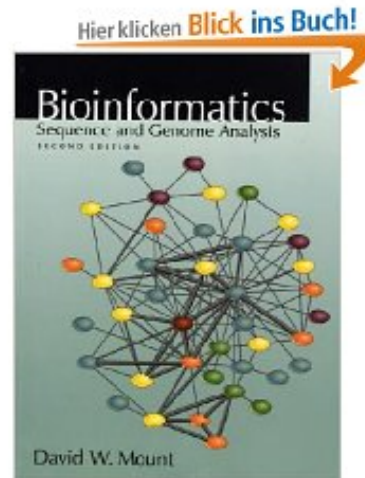
Die Veranstaltung gilt als bestanden, wenn in der abschließenden 120-minütigen Klausur über die Inhalte der Vorlesung, der Übungen und der Minipraktika mindestens die Note 4 erreicht wurde.

Für die Note des Scheins zählt das bessere Ergebnis entweder ausschließlich aus der abschließenden Klausur oder der Kombination des Durchschnitts der benoteten Praktika und der Note der Abschlussklausur, die jeweils zu 50 % gewichtet werden.

Bei Nichtbestehen der Klausur besteht die Möglichkeit einer schriftlichen oder mündlichen **Nachprüfung**. Diese findet im allgemeinen zu Beginn des darauffolgenden Semesters statt.

# Literatur

David Mount  
Bioinformatics  
70€



Marketa Zvelebil & Jeremil O. Baum  
Understanding bioinformatics, 96€

Zu empfehlen ist ebenfalls:

**Vorlesungsskript aus 2010** (176 Seiten)

kann von <http://gepard.bioinformatik.uni-saarland.de/teaching/ss-2011/sww-bioinformatik/script/SW10-Skript.pdf>

heruntergeladen werden.

Vorlesungsfolien ebenfalls auf

<http://gepard.bioinformatik.uni-saarland.de/teaching/ws-14-15/sww-bioinformatik-ws1415>

# Übersicht über Vorlesungsinhalt

## I Sequenz

1. Einführung, Datenbanken
2. Paarweises Sequenzalignment
3. Multiples Sequenzalignments;  
Phylogenie
4. Genvorhersage, Motivsuche

## II Proteinstruktur

5. Proteinstruktur; Sekundärstruktur
6. Homologie-Modellierung
7. Biomolekulare Interaktionen

## III Zellsimulationen/Netzwerke

8. Genexpression – Microarrays
9. Funktionsannotation (Gene Ontology)
10. Systembiologie: metabolische Pfade;  
Protein-Interaktion,  
Genregulationsnetzwerke
11. Enzymkinetik – einfache  
Differentialgleichungen
12. Diffusionssysteme - Virtual Cell
13. Stochastische Effekte



# Historische Entwicklung der Bioinformatik

1960'er Jahre: Entwicklung phylogenetischer Methoden

1960'er Jahre: Methoden zum Vergleich von DNA- und Proteinsequenzen

1976: erste MD-Simulation eines Proteins

1981: Smith-Waterman Algorithmus **dynamische Programmierung**

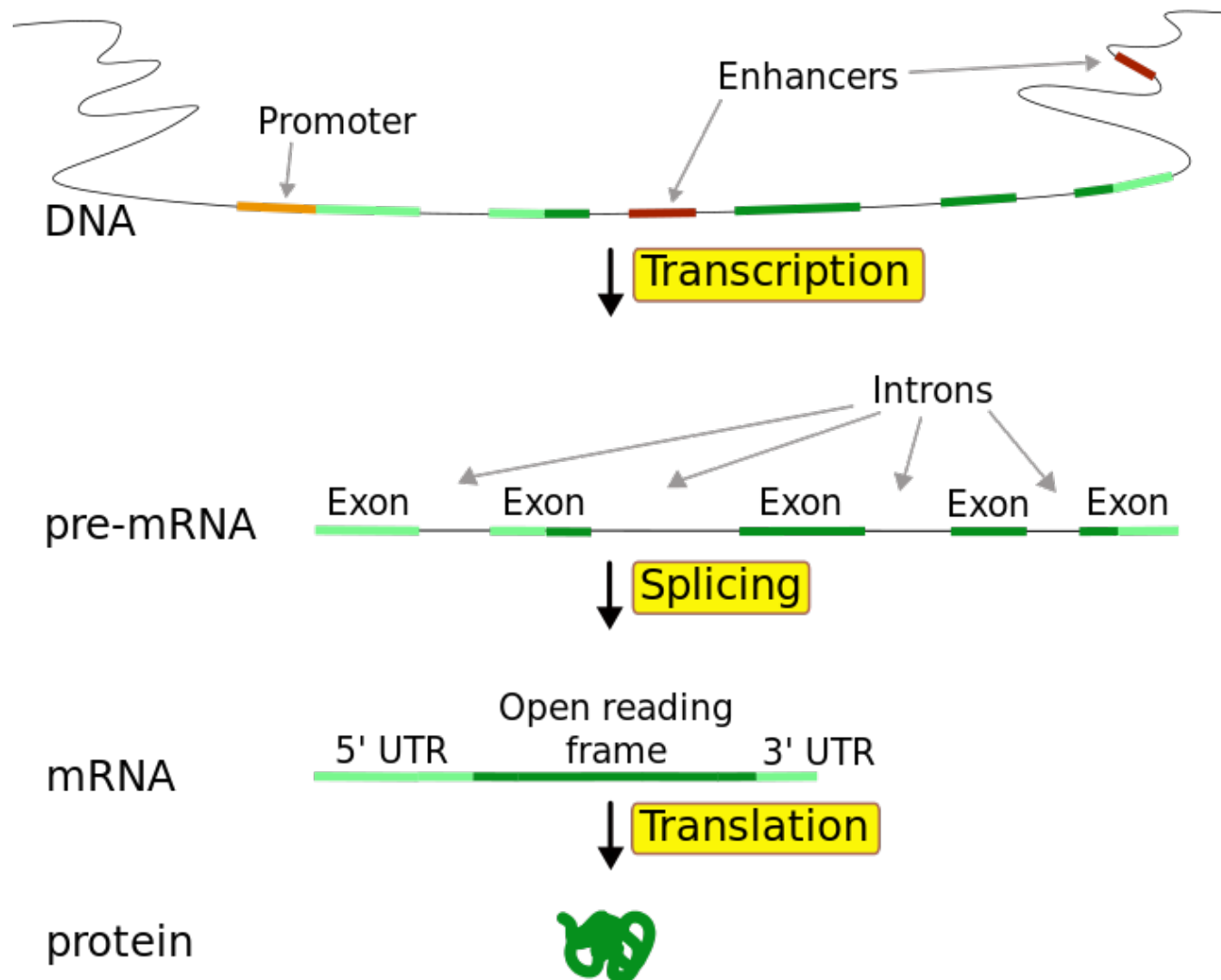
1992: Sekundärstrukturvorhersage mit Neuronalen Netzwerken (PHD)  
**machine learning**

1996: Vergleich von Proteinstrukturen mit DALI

2000: Durchbruch bei Sequenz-Assemblierung aus Shotgun-Daten (E. Myers)

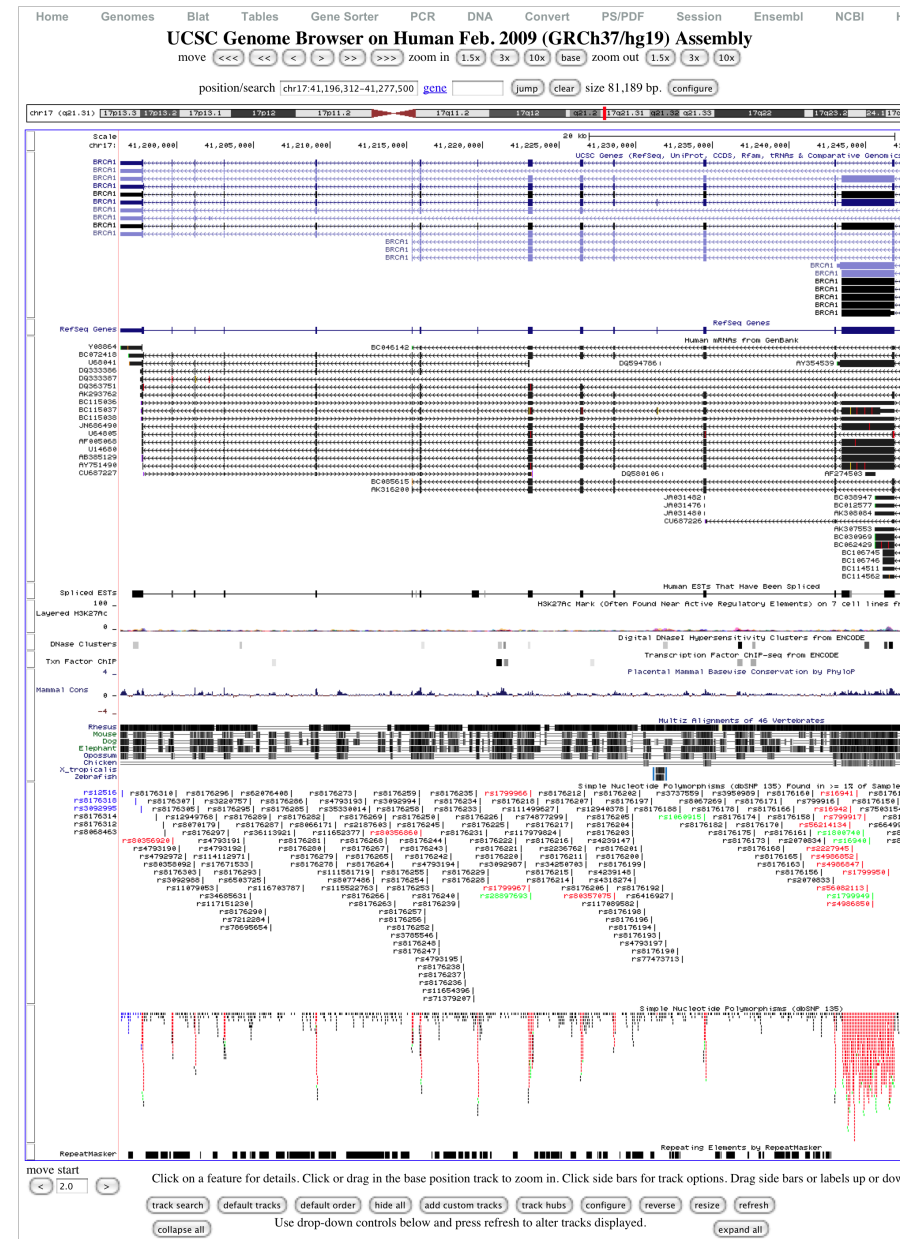
2012: ENCODE-Projekt

# Die Struktur von Genen



[www.wikipedia.org](http://www.wikipedia.org)

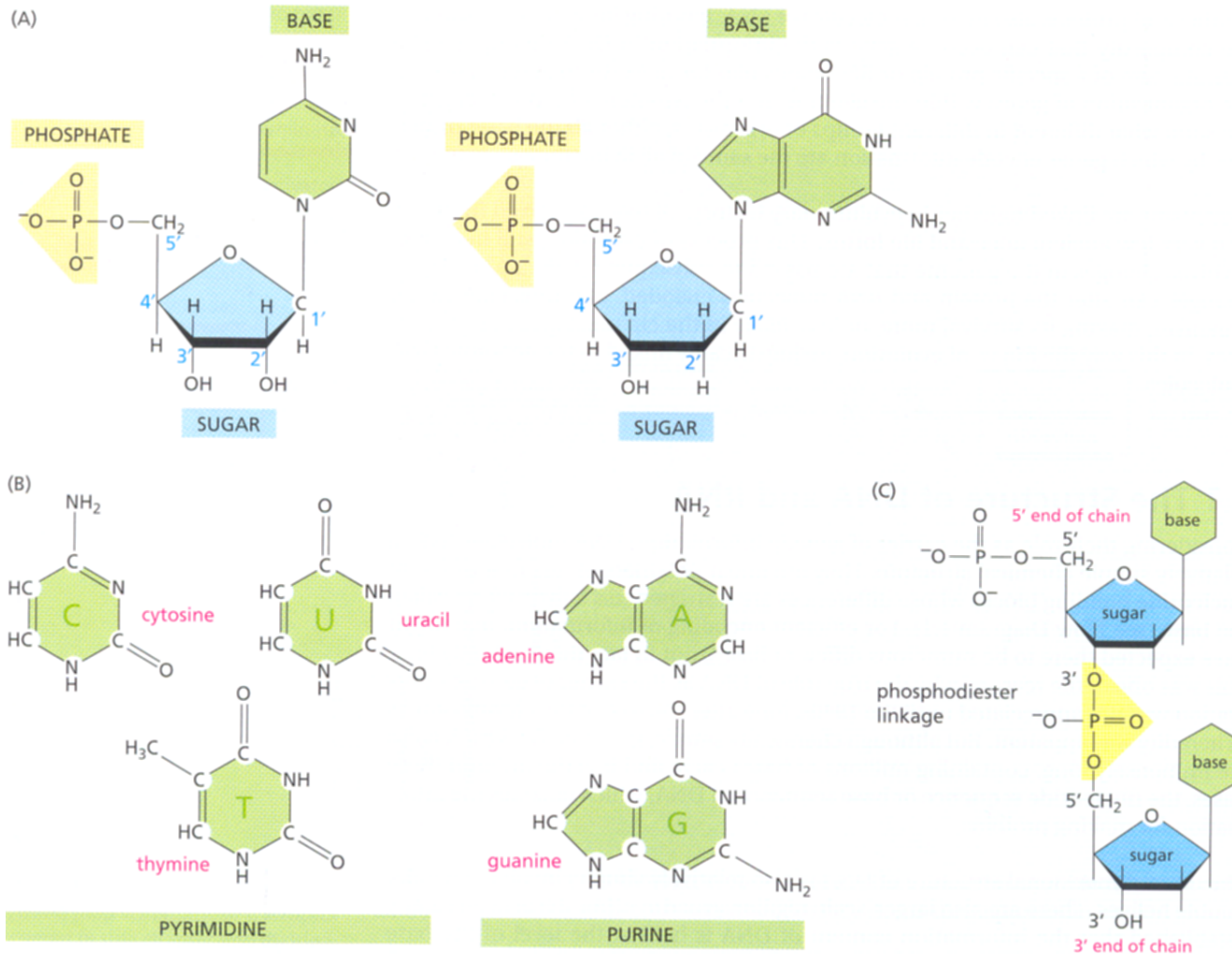
# UCSC Genome Browser



Es gibt verschiedene Assemblies  
hg17, hg18, hg19 für das *humane* Genom

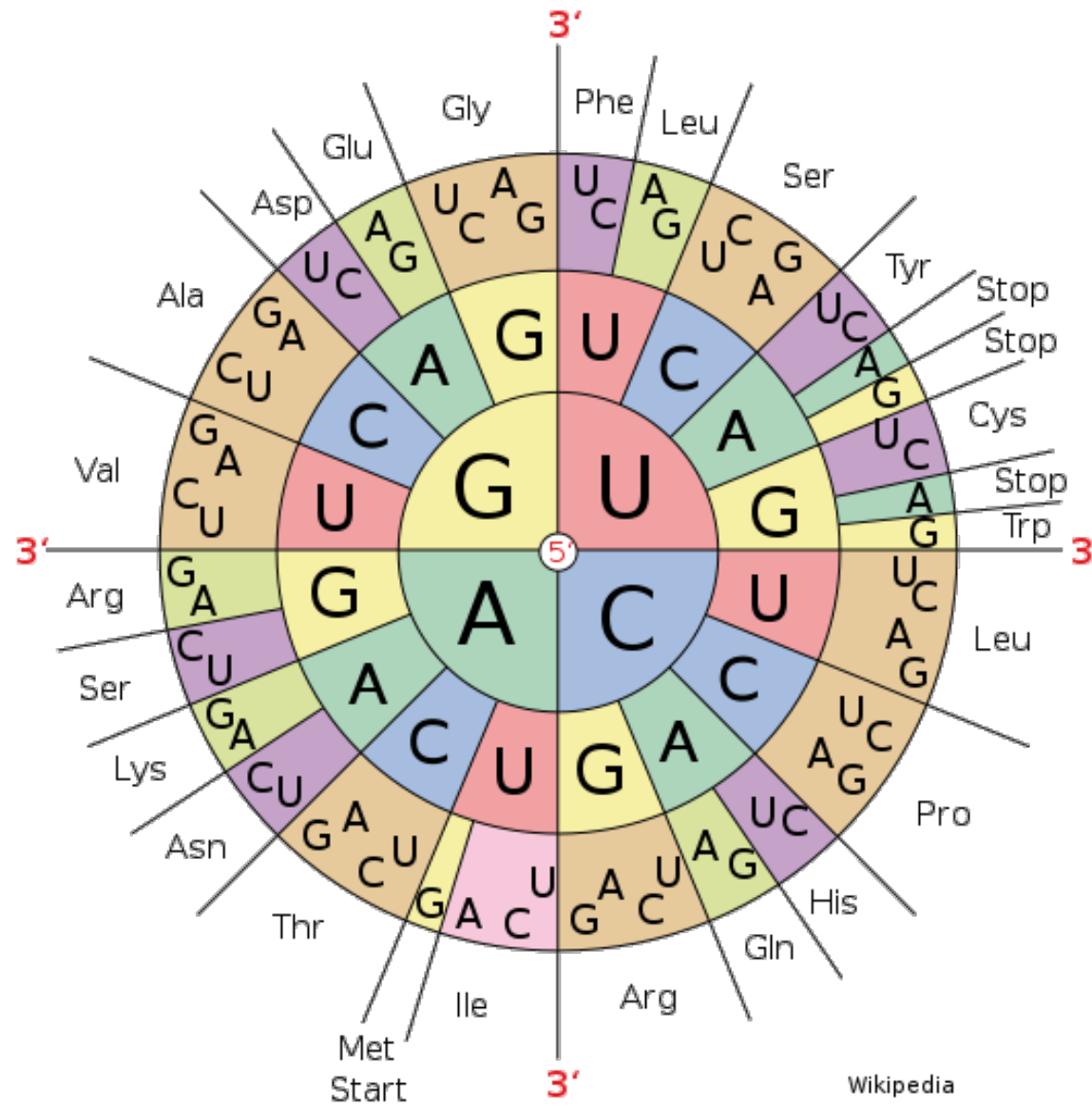
<http://genome.ucsc.edu/cgi-bin/hgGateway>

# Die vier Nukleotidbasen



Zvelebil (2008)

# Codonsonne

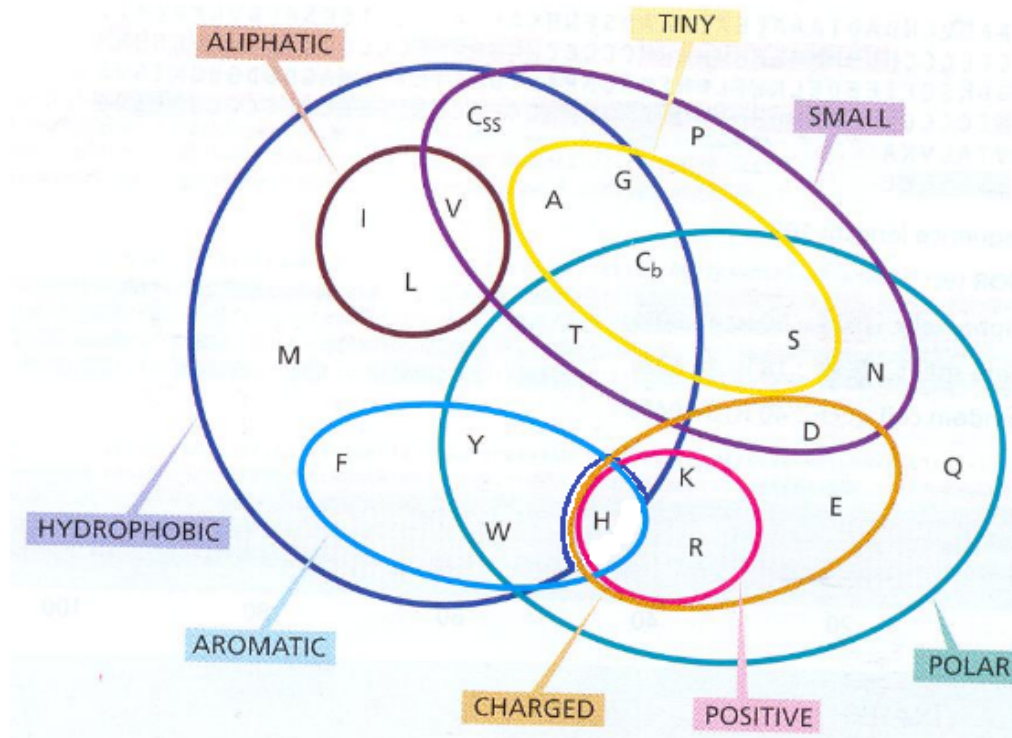


Wikipedia

Zvelebil (2008)

# Eigenschaften der Aminosäuren

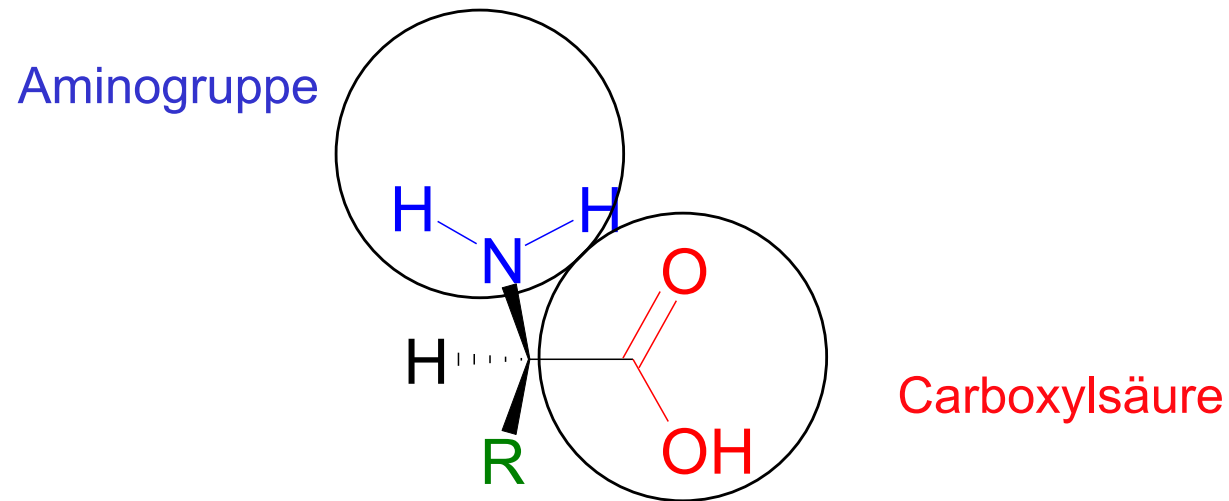
Aminosäuren unterscheiden sich in ihren physikochemischen Eigenschaften.



**Q:** müssen Bioinformatiker die Eigenschaften von Aminosäuren kennen?

# Einleitung: Aminosäuren

Aminosäuren sind die **Bausteine** von Proteinen:



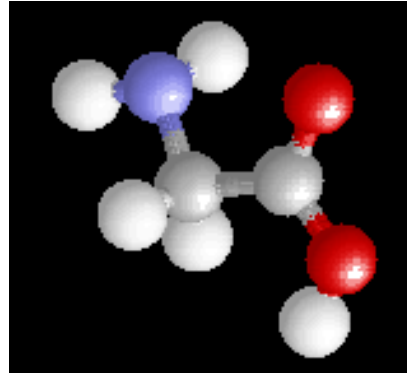
Aminosäuren unterscheiden sich hinsichtlich ihrer

- Größe
- elektrischen Ladung
- Polarität
- Form und Steifigkeit

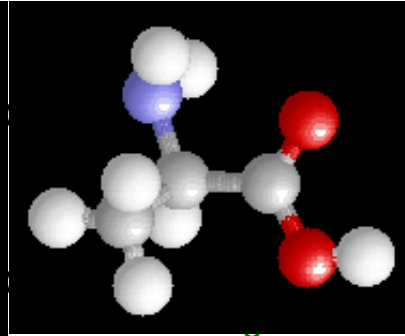
# Einleitung: hydrophobe Aminosäuren

Proteine sind aus 20 verschiedenen natürlichen Aminosäuren aufgebaut

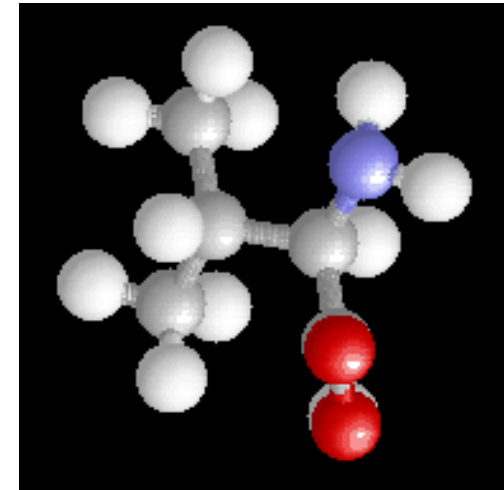
5 sind hydrophob.  
Sie sind vor allem  
Im Proteininneren.



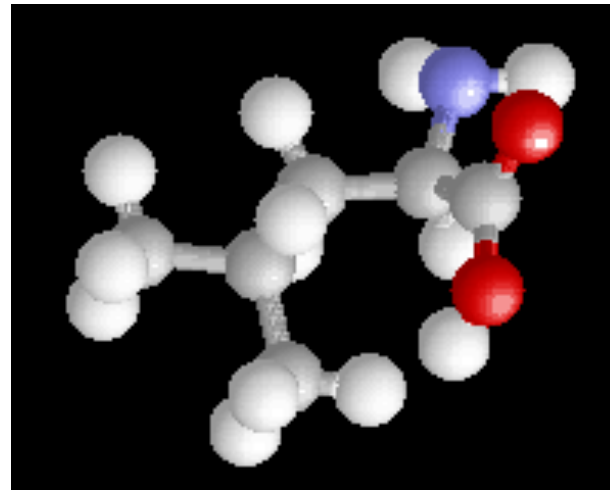
Glycine



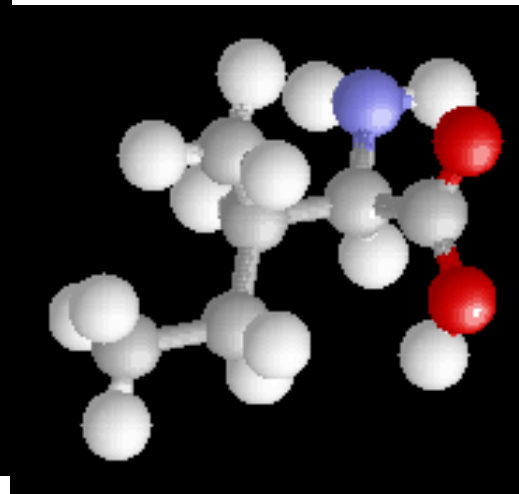
Alanine



Valine



Leucine

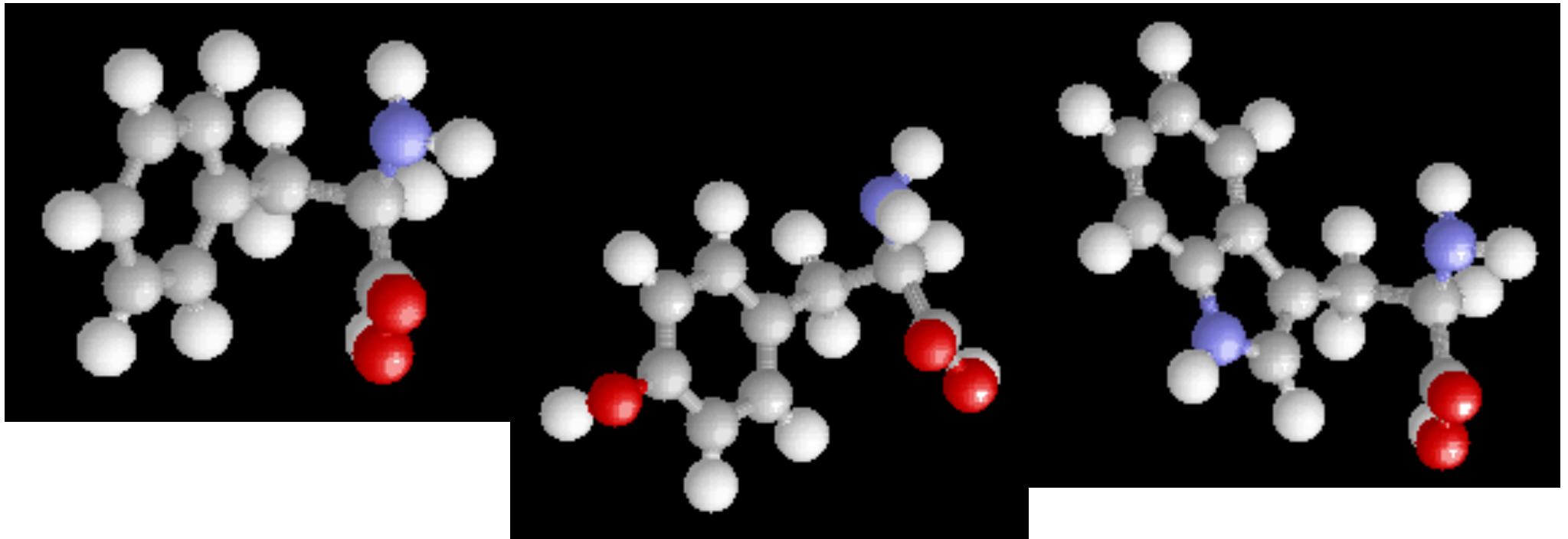


Isoleucine



## Einleitung: aromatische Aminosäuren

Es gibt drei voluminöse aromatische Aminosäuren. Tyrosin und Tryptophan liegen bei Membranproteinen vor allem in der Interface-region.



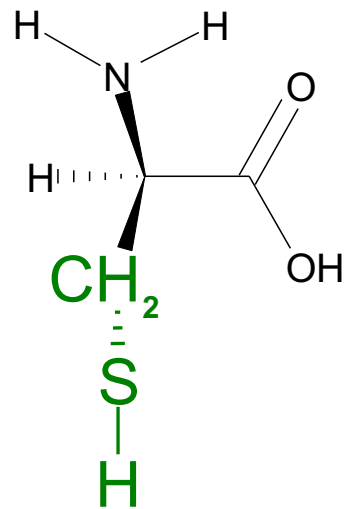
Phenylalanin

Tyrosin

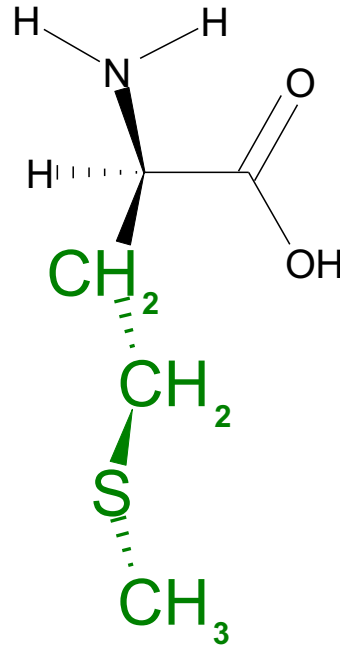
Tryptophan

# Einleitung: Aminosäuren

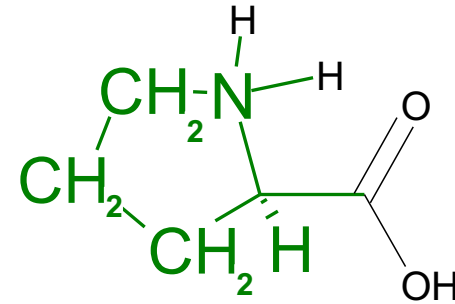
Es gibt 2 Schwefel enthaltende Aminosäuren und das ungewöhnliche Prolin.  
Cysteine können Disulfidbrücken bilden.  
Prolin ist ein "Helixbrecher".



Cystein



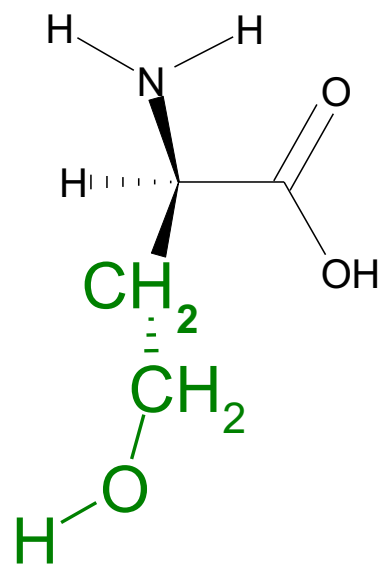
Methionin



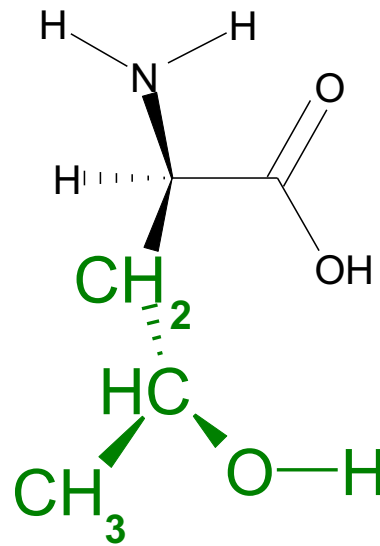
Prolin

# Einleitung: Aminosäuren

Es gibt zwei Aminosäuren mit terminalen polaren Hydroxylgruppen:



Serin

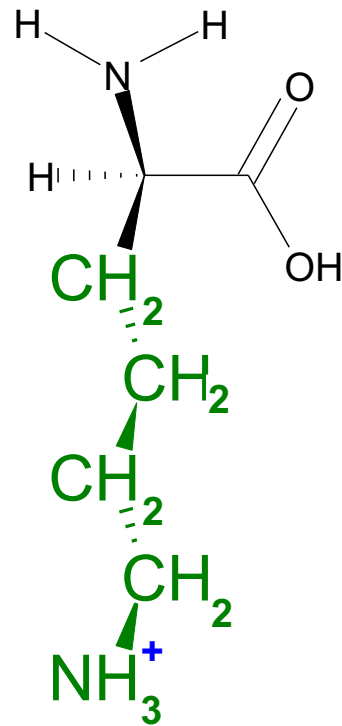


Threonin

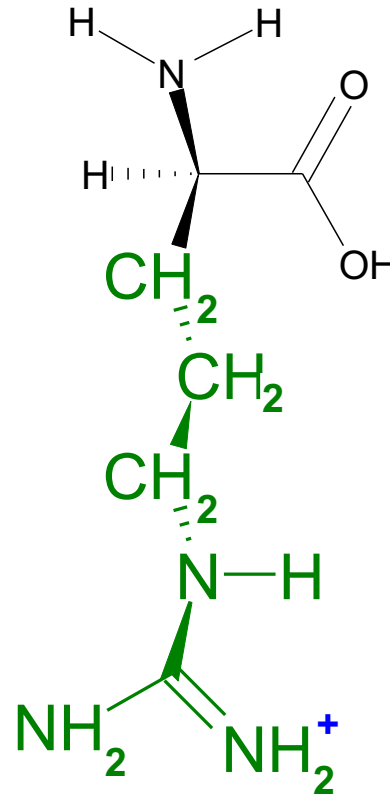
# Einleitung: Aminosäuren

Es gibt 3 positiv geladene Aminosäuren. Sie liegen vor allem auf der Proteinoberflächen und in aktiven Zentren.

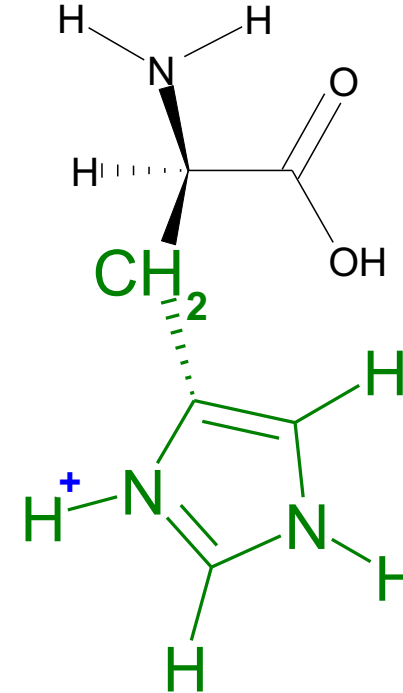
Thermophile Organismen besitzen besonders viele Ionenpaare auf den Proteinoberflächen.



Lysin



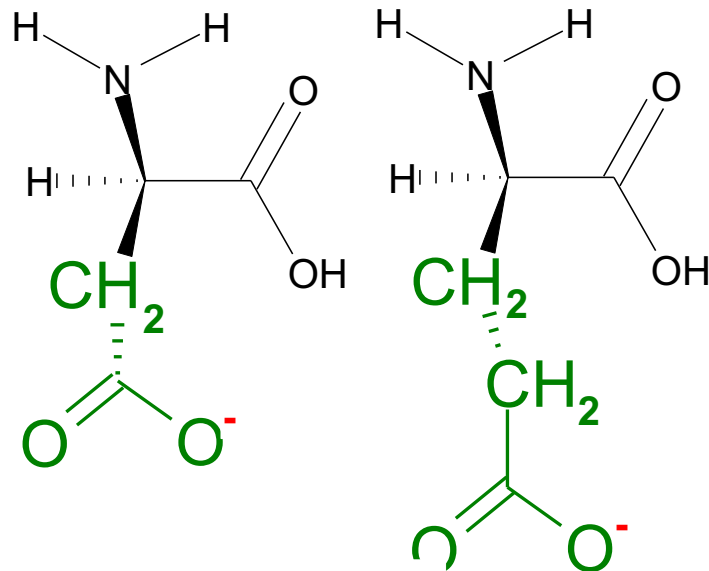
Arginin



Histidin

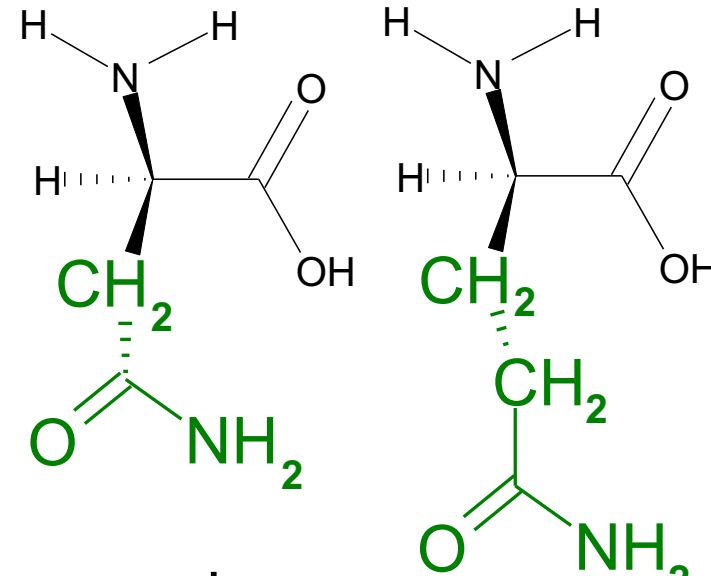
# Einleitung: Aminosäuren

Es gibt 2 negativ geladene Aminosäuren und ihre zwei neutralen Analoga. Asp und Glu haben  $pK_a$  Werte von 2.8. Das heisst, erst unterhalb von  $pH=2.8$  werden ihre Carboxylgruppe protoniert.



Asparaginsäure

Glutaminsäure



Asparagin

Glutamin

# Buchstaben-Code der Aminosäuren

- Ein- und Drei-Buchstaben-Codes der Aminosäuren

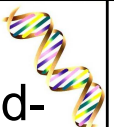
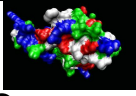

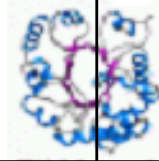
<b>G</b>	Glycin	<b>Gly</b>	<b>P</b>	Prolin	<b>Pro</b>
<b>A</b>	Alanin	<b>Ala</b>	<b>V</b>	Valin	<b>Val</b>
<b>L</b>	Leucin	<b>Leu</b>	<b>I</b>	Isoleucin	<b>Ile</b>
<b>M</b>	Methionin	<b>Met</b>	<b>C</b>	Cystein	<b>Cys</b>
<b>F</b>	Phenylalanin	<b>Phe</b>	<b>Y</b>	Tyrosin	<b>Tyr</b>
<b>W</b>	Tryptophan	<b>Trp</b>	<b>H</b>	Histidin	<b>His</b>
<b>K</b>	Lysin	<b>Lys</b>	<b>R</b>	Arginin	<b>Arg</b>
<b>Q</b>	Glutamin	<b>Gln</b>	<b>N</b>	Asparagin	<b>Asn</b>
<b>E</b>	Glutaminsäure	<b>Glu</b>	<b>D</b>	Asparaginsäure	<b>Asp</b>
<b>S</b>	Serin	<b>Ser</b>	<b>T</b>	Threonin	<b>Thr</b>

## Zusätzliche Codes

**B** Asn/Asp   **Z** Gln/Glu   **X** Irgendeine Aminosäure

**Die Kenntnis dieser Abkürzungen ist essentiell für Sequenzalignments und für Proteinstrukturanalyse!**

# Datenbanktypen

primär				sekundär				
DNA-/ Nukleotid-Sequenzen 	Protein-/ Aminosäure-Sequenzen 	Protein-, DNA-Strukturen		Protein-/ Aminosäure-Sequenzen			Protein-Strukturen	
GenBank	NCBI Protein Database	Swiss Prot (Uniprot)	PDB 	PROSITE	Prints	Pfam	SCOP 	CATH

- Sequenzinformationen
- zugehörige Annotationen
- Kreuzreferenzen zu anderen Datenbanken

- Analysen auf Basis der primären Datenbanken
- Klassifizierungen nach Ähnlichkeit

# Sequenzdaten

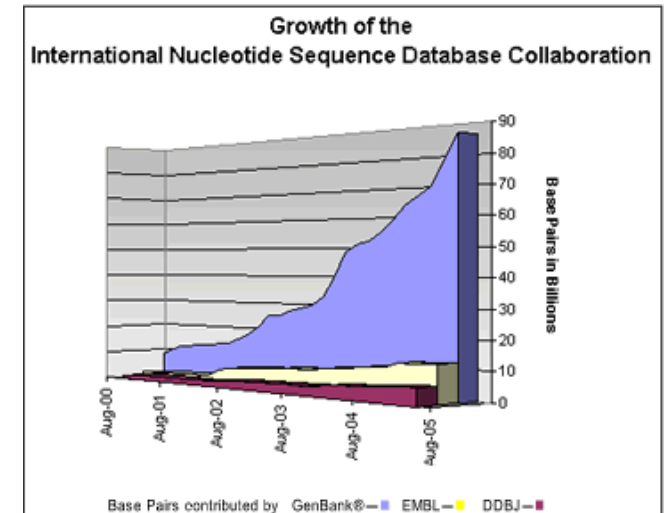
- in Okt. 2016 ~197 Mio. **Nukleotidsequenzen**  
(Quelle: GenBank <http://www.ncbi.nlm.nih.gov/genbank/index.html>)

~363 Mio. WGS-Nukleotidsequenzen

- 114.767 **3D-Strukturen** von biologischen

Makromolekülen (Proteine, DNA, RNA, ...)

(Quelle: RCSB-PDB <http://www.rcsb.org>, Okt. 2016)



Einträge sind teilweise **redundant**,

d.h. es gibt mehrere Versionen derselben Sequenz/Struktur



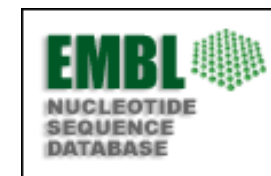
# NCBI DNA-Datenbank

National Center for Biotechnology Information  
National Library of Medicine      National Institutes of Health



GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>)

- öffentliche Nukleotid-Sequenzdatenbank
- ~197 Mio. Sequenzeinträge
- fast jeder kann Sequenzen einreichen
- Mindestlänge der eingereichten Sequenzen: 50 bp
- jeder Eintrag bekommt eine eindeutige *Accession Number*
- wird alle 24h gegen EMBL-Bank (EMBL Nucleotide Sequence Database, <http://www.ebi.ac.uk/>) und DDBJ (DNA DataBank of Japan, <http://www.ddbj.nig.ac.jp>) synchronisiert
- redundant



# NCBI Protein-Datenbank



NCBI Protein Database (<http://www.ncbi.nlm.nih.gov/>)

- öffentliche, primäre Protein-Sequenzdatenbank
- Zusammenstellung aus den folgenden Protein-Sequenzdatenbanken:
  - UniProtKB
  - PIR (Protein Identification Resources)
  - PDB (Protein Data Bank, Strukturen)
  - Proteintranslationen der GenBank-Datenbank
  - und weiteren
- redundant
- Vorteil: Links zu Original-Datenbanken

# UniProtKB/Swiss-Prot



(<http://www.expasy.org/sprot/>)

- Universal Protein Resource Knowledge Base
- öffentliche, primäre Proteinsequenz-Datenbank
- “nur” 552.000 Einträge (Okt 2016)
- wichtigste Sammlung von Proteinsequenzen:
  - Daten stammen aus der Datenbank TrEMBL (*translated* EMBL)
  - manuell überprüft; manuelle Annotationen von Experten
  - nicht redundant
  - Querverweise zu Funktionsbeschreibung, Domänenstruktur, posttranslationalen Modifikationen und ~60 anderen Datenbanken
- UniProtKB/TrEMBL enthält Einträge, die noch nicht in UniProtKB/Swiss-Prot aufgenommen wurden



# Webinterface: Entrez



Datenbank wählen

The screenshot shows the Entrez Protein search results for the query 'melibiase'. The search was performed in the 'Protein' database. The results list 866 items, with the first seven displayed. Each result includes a protein ID, a title, and a description. The first result is NP\_822252, melibiase [Streptomyces avermitilis MA-4680]. The second is CAA69852, alpha-galactosidase; melibiase [Thermoanaerobacter ethanolicus ATCC 33223]. The third is NP\_189269, glycosyl hydrolase family protein 27 / alpha-galactosidase family protein / melibiase family protein [Arabidopsis thaliana]. The fourth is AAA34770, pre-alpha galactosidase (melibiase). The fifth is AAO78237, alpha-galactosidase (melibiase) [Bacteroides thetaiotaomicron VPI-5482]. The sixth is AAO77957, alpha-galactosidase (melibiase) [Bacteroides thetaiotaomicron VPI-5482]. The seventh is XP\_001315888, Melibiase family protein [Trichomonas vaginalis G3].

Stichwort, hier Name des Proteins

# Detaillierte Suche bei Entrez

Protein Result - Mozilla Firefox  
http://www.ncbi.nlm.nih.gov/sites/entrez

NCBI  
Entrez Protein

Search Protein for melibiase "bacillus subtilis" [ORGN] Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Relevance Send to

All: 1 Bacteria: 1 RefSeq: 0 Related Structures: 1

1: [O34645](#) Reports BLink, Conserved Domains, Links  
Alpha-galactosidase (Melibiase)  
gi|3912990|sp|O34645|AGAL\_BACSU[3912990]

Suche nach dem Protein Melibiase in genau diesem Organismus

weitere nützliche Beschränkungen:

- [ACCN]: Accession Number
- [KYWD]: Stichwort zur Funktion etc.
- X:Y [SLEN]: Sequenzlänge zwischen X und Y
- [TITL]: Wort muß im Titel des Eintrags stehen
- [AUTH]: Name des Autors bei Suche nach einer Publikation in PubMed (elektronische Zeitschriftenbibliothek)
- logische Verknüpfungen mit NOT, OR
  - AND als automatische Voreinstellung

# Eintrag bei NCBI Protein Database

```
NCBI Sequence Viewer v2.0 - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=3912990
Google

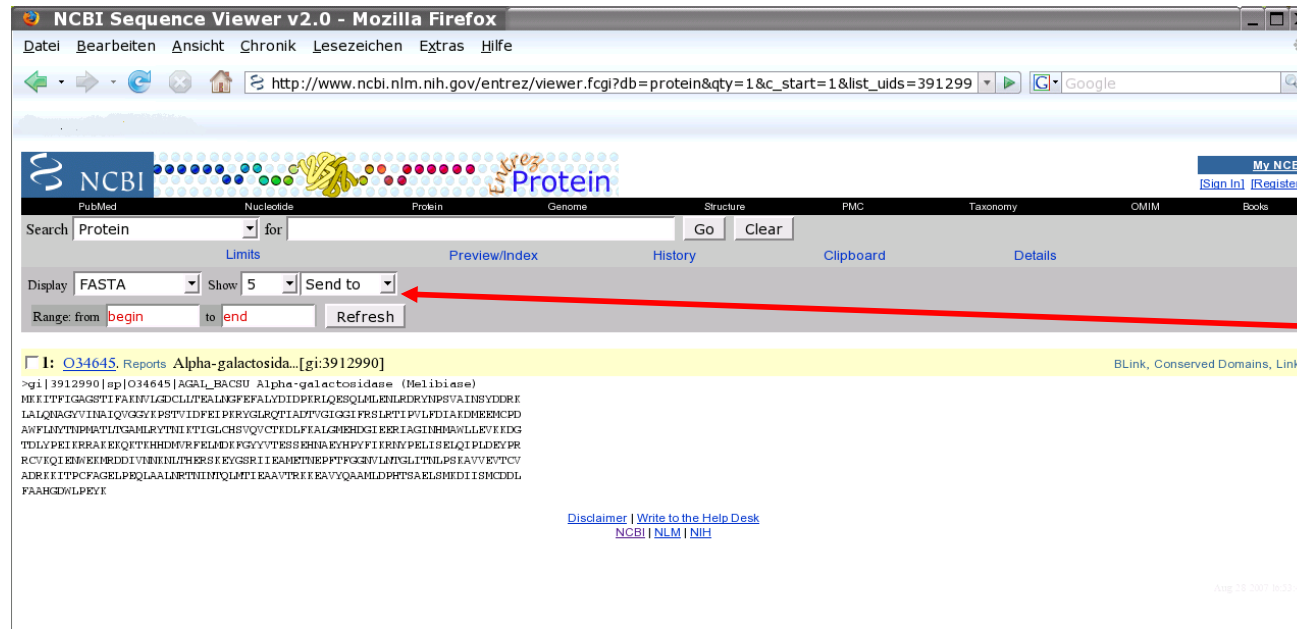
Region
3..162
/gene="melA"
/locus_tag="BSU30300"
/region_name="Ldh_L_N"
/notes="Lactate/malate dehydrogenase, NAD binding domain.
L-lactate dehydrogenases are metabolic enzymes which
catalyse the conversion of L-lactate to pyruvate, the last
step in anaerobic glycolysis; pfam02056"
/db_xref="CDD:65809"

Site
148
/gene="melA"
/locus_tag="BSU30300"
/site_type="binding"
/inference="non-experimental evidence, no additional
details recorded"
/notes="Substrate (By similarity)."
```

ORIGIN

```
1  mdkitfigag stfaknvlq delltealg fefalydidp krlqesqlal enldrnyps
61  vainsyddrk lalgnagyvi naigvgykp stvidfeipk ryrlqtias tvyiggifrs
121 lrtip'lfdi akdnsemcpd awflnytnpm atltgamry tnikitqlch svqvctkdlf
181 kalqmehdgi eeriaqinhn awlle'okdq tdlpeikrr akekqtbhhd dmvrfelndk
241 fgyyvtesse hnaeyhyfi knypelise lqipideyr rcvkiemwe kurddivnkk
301 nltherskey qsrrieamet nepftfgm' lntglitnlp skavvvtcv adrkkitpcf
361 agelpeqlaa lnrntintql mtieavtrk keavyqaam dphsaalam kdiismcddl
421 faahgdwlpe yk
//
Fertig
```

# Fasta-Format



Umstellung der Anzeige, Beschränkung auf bestimmten Abschnitt der Sequenz, ...

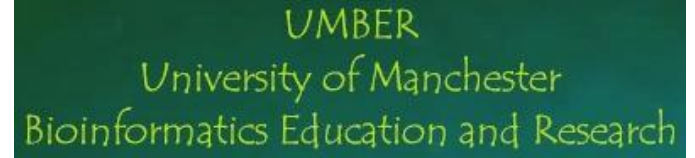
>DNA-Sequenz-Bezeichnung  
ACGT

....

>Protein-Sequenz-Bezeichnung  
ACDEFGHIKLMNPQRSTVWY

....

# PRINTS



(<http://bioinf.manchester.ac.uk/dbbrowser/PRINTS/>)

- sekundäre Protein-Datenbank
- 2.156 Einträge und 12.444 Motive (in 2012)
- Fingerabdruck (*fingerprint*): Gruppe von konservierten Motiven
- mehrere funktionelle Bereiche (Faltung, Ligandenbindung, Komplexbildung, ...) -> mehrere Sequenzmotive für ein Protein
- Motive aus kurzen lokalen Alignments
  - Abstände zwischen Motiven und Reihenfolge spielen keine Rolle
    - spezifisch für individuelle Proteine
    - keine Zusammenfassung zu gemeinsamem Motiv



# Finger-PRINTS

==SPRINT==> Query Results - Mozilla Firefox

http://www.bioinf.manchester.ac.uk/cgi-bin/dbbrowser/sprint/searchprintss.cgi?display\_opts=Pi

Final Motifs

Motif	width	Element	Seqn Id	St	Int	Rpt		
Motif 1	width=16	SVVVAGGGSTFTPGIV	<a href="#">GLVG_RCOLI</a>	5	5	-		
		SIVIAGGGSTFTPGIV	<a href="#">GLVG_BACSU</a>	7	7	-		
		KITFIGAGSTIFVKHI	<a href="#">AGAL_RCOLI</a>	6	6	-		
		KITFIGAGSTIFAKIV	<a href="#">AGAL_BACSU</a>	3	3	-		
		KVVTIGGGSSYTPRELL	<a href="#">CRLF_RCOLI</a>	6	6	-		
		KIVTIGGGSSYTPRELV	<a href="#">CRLF_BACSU</a>	6	6	-		
		SILLAGGGSTFTPGII	<a href="#">HALH_FUSHM</a>	5	5	-		
		KIAYIGGGSGWARSLS	<a href="#">LPLD_BACSU</a>	11	11	-		
		Motif 2	width=17	ALSAADIVIISILPGSL	<a href="#">LPLD_BACSU</a>	76	49	-
				ALBDADFVVAQIGGY	<a href="#">AGAL_RCOLI</a>	75	53	-
AFPTIDFVNAHIRVGKY	<a href="#">HALH_FUSHM</a>			74	53	-		
ALKDADFVTTQLRVGQL	<a href="#">CRLF_RCOLI</a>			77	55	-		
AFSDVDFVNAHIRVGKY	<a href="#">GLVG_RCOLI</a>			74	53	-		
AFPTVDFVNAHIRVGKY	<a href="#">GLVG_BACSU</a>			76	53	-		
ALKDADFVTTQFRVGLL	<a href="#">CRLF_BACSU</a>			77	55	-		
ALQHAGYVINAIQVGGY	<a href="#">AGAL_BACSU</a>			72	53	-		
Motif 3	width=14			LDRQIPLKYGVVGG	<a href="#">GLVG_BACSU</a>	97	4	-
				LDEKIPLRHGVVGG	<a href="#">HALH_FUSHM</a>	95	4	-
		LDEKIPLRHGVVGG	<a href="#">GLVG_RCOLI</a>	95	4	-		
		TDFEVCKRHGLKQT	<a href="#">AGAL_RCOLI</a>	97	5	-		
		LDRRIPLSHGVLGQ	<a href="#">CRLF_RCOLI</a>	98	4	-		
		KDRRIPLYGVIGQ	<a href="#">CRLF_BACSU</a>	98	4	-		
		IDFKIPKRYGLRQT	<a href="#">AGAL_BACSU</a>	94	5	-		
		VDVHLPRRCGIYQS	<a href="#">LPLD_BACSU</a>	97	4	-		
		Motif 4	width=21	DTVGGGIIRGLRAVPFAEI	<a href="#">LPLD_BACSU</a>	113	2	-
				ETCGPGGIAYGHRSIGGVIGL	<a href="#">HALH_FUSHM</a>	109	0	-
ETCGPGGIAYGHRSIGGVLEL	<a href="#">GLVG_RCOLI</a>			109	0	-		
ETCGPGGIAYGHRSIGGVLEI	<a href="#">GLVG_BACSU</a>			111	0	-		
DTLGGGIIHRLRTIPHLWQI	<a href="#">AGAL_RCOLI</a>			113	2	-		
ETHGPGGLFKGLRTIPVILEI	<a href="#">CRLF_BACSU</a>			112	0	-		
DTVGGGIFRSLRTIPVLFDI	<a href="#">AGAL_BACSU</a>			110	2	-		
ETHGAGGLFKGLRTIPVIFDI	<a href="#">CRLF_RCOLI</a>			112	0	-		
Motif 5	width=18			PDAWHLHYSNPAAIVAEA	<a href="#">GLVG_BACSU</a>	140	8	-
				PNAWHLHYSNPAAIVAEA	<a href="#">GLVG_RCOLI</a>	138	8	-
		PDATHLHYVHPNMTWA	<a href="#">AGAL_RCOLI</a>	142	8	-		
		PDAWHLHYVHPNMTLGA	<a href="#">AGAL_BACSU</a>	139	8	-		
		PNAWHLHYSNPAAIVAEA	<a href="#">HALH_FUSHM</a>	138	8	-		
		PNAWHLHYVHPNMTLGA	<a href="#">CRLF_RCOLI</a>	141	8	-		
		PNAWHLHYVHPNMTLGA	<a href="#">CRLF_BACSU</a>	141	8	-		
		PNAWHLHYVHPNMTLGA	<a href="#">LPLD_BACSU</a>	142	8	-		
		Suchen: tetL	Abwärts	Aufwärts	Hervorheben	Groß-/Kleinschreibung	Das Seitenende wurde erreicht, Suche vom Seitenanfang fortg	

# PRINTS - Example

Illustration of a hierarchical PRINTS diagnosis. The UniProtKB/TrEMBL entry Q9NSV5\_HUMAN was annotated as putative uncharacterized protein DKFZp434D2030; the family- and domain-database cross-references suggested membership of the major intrinsic protein (MIP) superfamily, but provided no specific family affiliation. The FingerPRINTScan result (inset) diagnoses the sequence both as a member of the MIP superfamily and as an aquaporin 6 subtype.

The screenshot shows the UniProtKB entry for Q9NSV5\_HUMAN. The entry is annotated as 'Putative uncharacterized protein DKFZp434D2030'. The taxonomic lineage is Eukaryota · Metazoa · Chordata · Craniata · Vertebrata · Euteleostomi · Mammalia · Eutheria · Euarchontoglires · Primates · Haplorhini · Catarrhini · Hominidae · Homo.

The 'Family and domain databases' section includes:

InterPro	IPR000425. MIP. IPR022357. MIP_CS. [Graphical view]
PANTHER	PTHR19139. MIP. 1 hit.
Pfam	PF00230. MIP. 1 hit. [Graphical view]
PRINTS	PR00783. MINTRINSICP.
PROSITE	PS00221. MIP. 1 hit. [Graphical view]

The FingerPRINTScan result (inset) shows the following information:

PRINTS42\_0 and matrix blo62

Scan of sequence: Q9NSV5\_HUMAN  
SubName: Full=Putative uncharacterized protein DKFZp434D2030;

Highest scoring fingerprints for Q9NSV5\_HUMAN

Fingerprint	E-value	GRAPHScan	Motif3D
MINTRINSICP (relations)	6.590555e-24	Graphic	
AQUAPORIN6 (relations)	9.409030e-23	Graphic	

Attwood et al. Database (2012) 2012 : bas019 doi: 10.1093/database/bas019

# Pfam – Protein-Familien-Datenbank

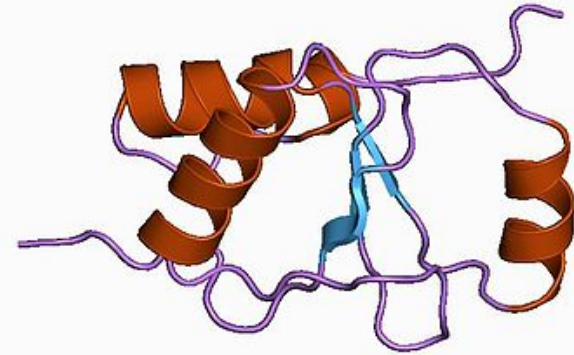


*(<http://pfam.xfam.org>)*

- sekundäre Protein-Datenbank
- 74% aller Proteinsequenzen haben mindestens einen Pfam-Eintrag
- Profile = funktionell interessante Domänen
- **Profil**: Auftrittswahrscheinlichkeiten bestimmter Aminosäuren an bestimmten Positionen in Form einer Matrix
- **Pfam**: genau untersuchte Profile aus multiplen Alignments und Hidden Markov Modellen (HMM), teilweise manuelle Alignments, >16.3000 Protein-Familien (Okt 2016, Pfam v30.0)

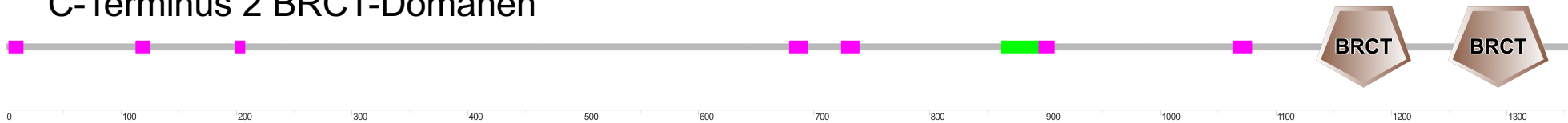
# Pfam – Profil für BRCA1

Hollywoodstar Angelina Jolie hat sich aus Angst vor Krebs vorsorglich beide Brüste abnehmen lassen. Sie habe sich für den Eingriff entschieden, weil sie ein defektes Gen namens BRCA1 in sich trage, das ihr Risiko für Brust- und Eierstockkrebs erheblich erhöhe, schrieb die 37-Jährige in der „New York Times“.

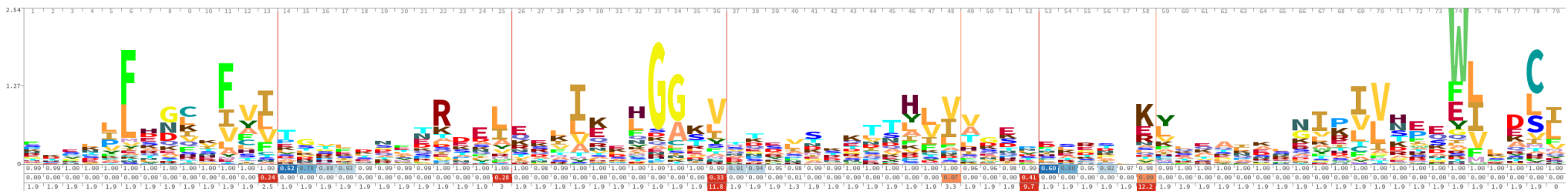


Kristallstruktur der BRCT-Domäne

<http://smart.embl-heidelberg.de> -> Domänenstruktur, BRCA1 enthält mehrere low complexity Regionen (lila), 1 coiled coil Region (grün) und am C-Terminus 2 BRCT-Domänen



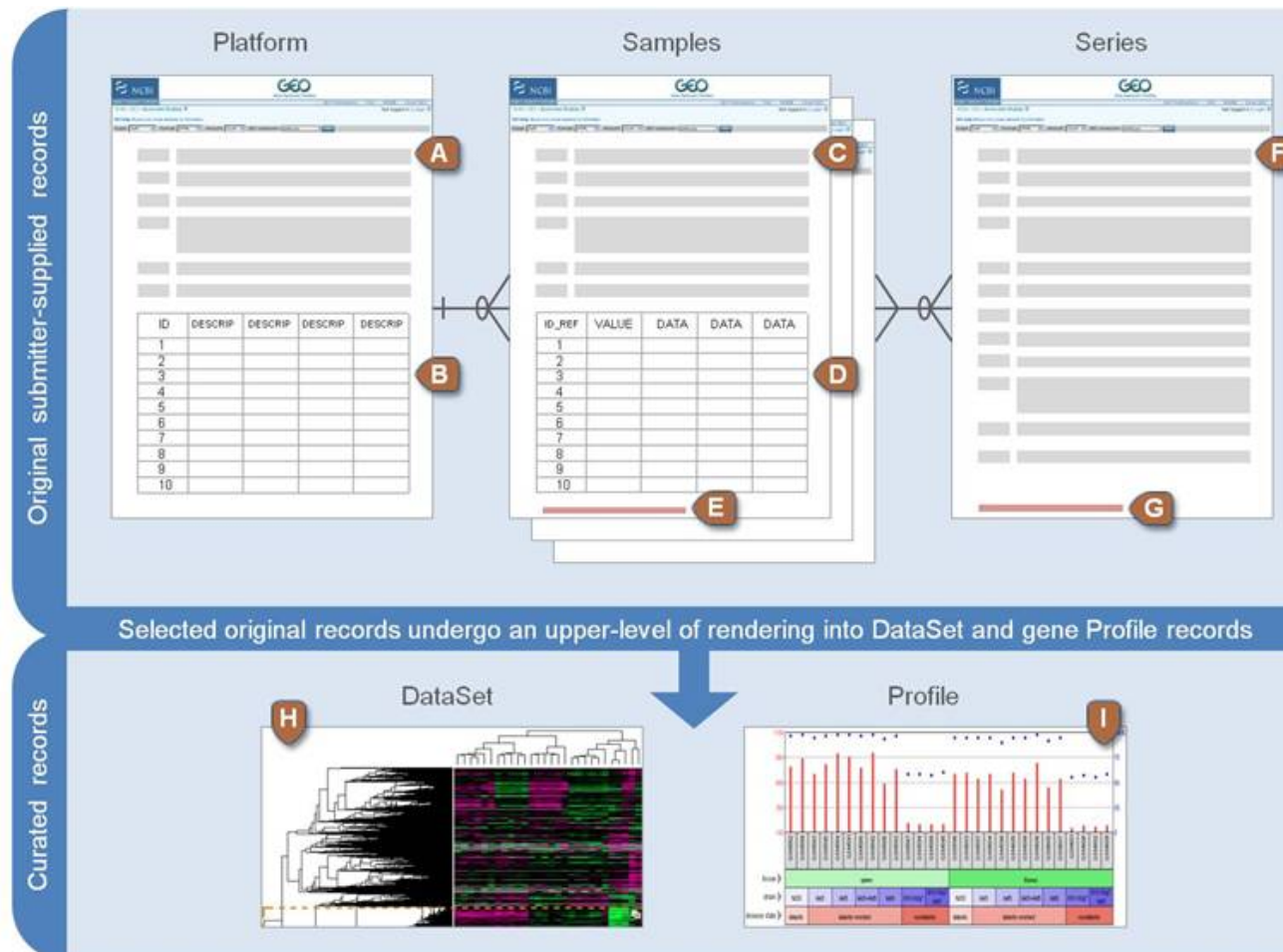
## HMM-Logo von PFAM für die BRCT-Domäne



# GEO – Gene Expression Omnibus

(<http://www.ncbi.nlm.nih.gov/geo/>)

- Genexpressions-Datensätze
- entweder mit Microarrays oder NGS gemessen



# GEO – Gene Expression Omnibus

## Cancer Research

ACR

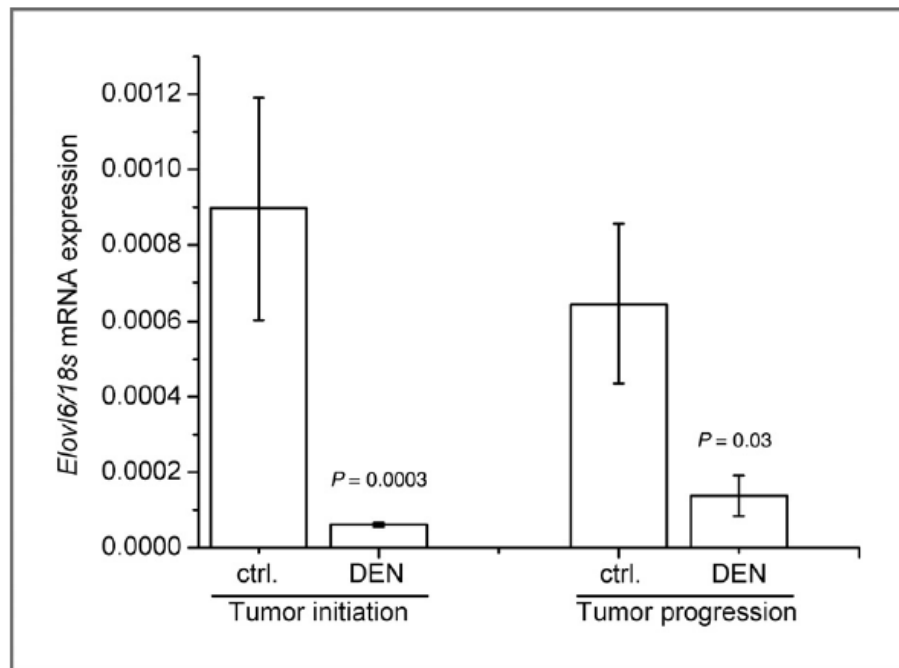
Ist die im Arbeitskreis Kiemer in Mäusen mit Leberkrebs (HCC) beobachtete Runterregulation von *Elovl6* auch im Mensch relevant?

Ja, dies konnten wir anhand von öffentlich zugänglichen GEO-Daten zeigen.

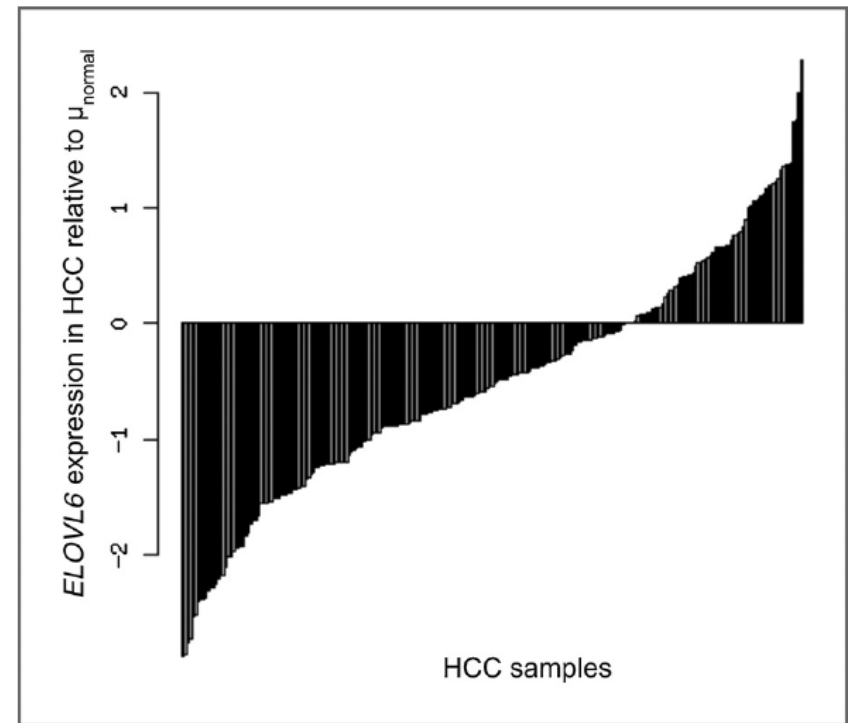
### Lipid Metabolism Signatures in NASH-Associated HCC—Letter

Sonja M. Kessler, Stephan Laggai, Ahmad Barghash, et al.

*Cancer Res* Published OnlineFirst April 28, 2014.



**Figure 2.** Wild-type mice were treated with the carcinogen DEN at the age of 2 weeks. Livers were analyzed after 24 weeks to assess the tumor initiation state. Analyses in the tumor progression stadium were done after 36 weeks. *Elov6* mRNA expression as determined by real-time reverse transcriptase PCR with  $n = 8-18$  per group. Data were normalized to 18S. Statistical differences compared with untreated animals of the same age (ctrl.) were calculated by Mann-Whitney  $U$  test.



**Figure 1.** mRNA levels of *ELOVL6* in 247 human HCC samples relative to the mean of 239 nontumor liver tissue ( $\mu_{normal}$ ). Samples of dataset GSE14520 [ $\log_2$  (expression) values from GEO after Robust Multi-array Average normalization] were mapped to hgu133a.db using bioconductor. Significance values:  $P = 3.8E-11$ , Kolmogorov-Smirnov test;  $P = 6.7E-11$ ,  $t$  test;  $5.1E-11$ , Mann-Whitney  $U$  test.

**Übungen** heute Nachmittag

## **Ausblick**

Bioinformatik-Software muss man **hands-on** kennenlernen.

Im **Tutorial** zeigen wir Ihnen den Umgang mit weit verbreiteter Bioinformatik-Software.

Das Tutorial ist genauso wichtig wie die Vorlesung!

In wenigen Wochen sollen Sie mit diesen Tools in einer kleinen Gruppe ein **Mini-Forschungsprojekt** bearbeiten. Also passen Sie bitte gut auf ... 😊

Gute **Statistik-Kenntnisse** sind essentiell für das Design von Experimenten, für das Aufstellen von Arbeitshypothesen und für die Arbeit mit Datenmengen.

Wichtig ist zudem das Verständnis, wie die **Daten gewonnen** wurden und welche **Fehlerquellen** auftreten können.