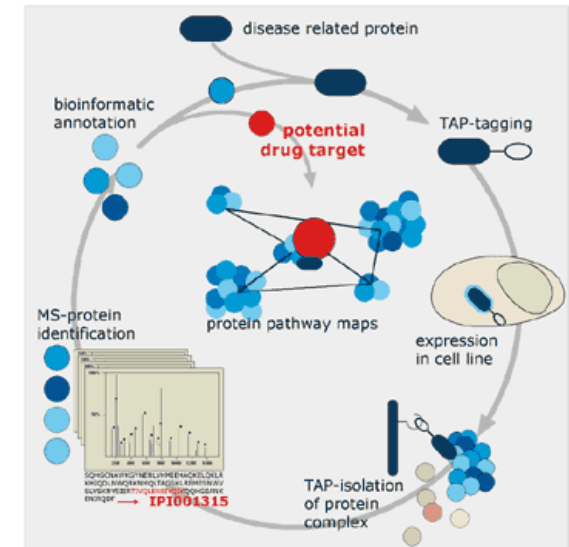


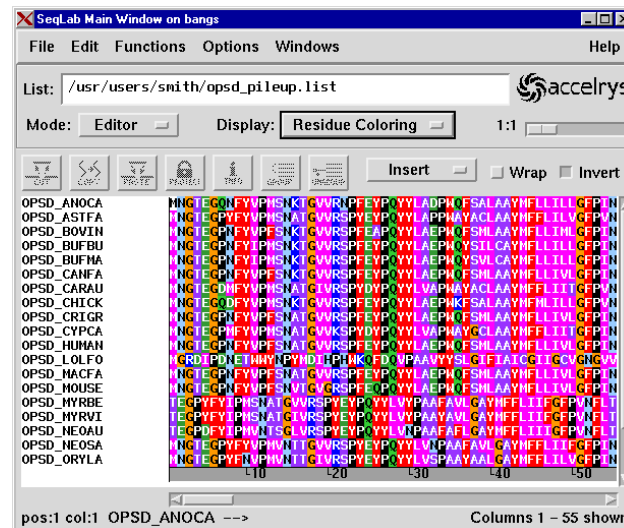
Softwarewerkzeuge der Bioinformatik

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



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 17.10:

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

Verantwortliche Betreuer der Übungen

Sequenz-Analyse

Markus Hollander

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
- (Wahl-) 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 2 der 3 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.

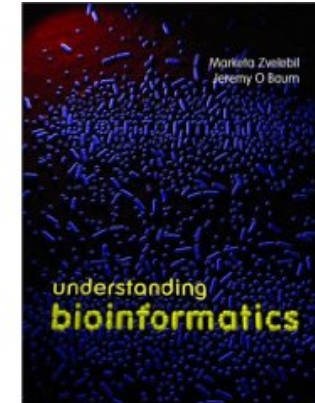
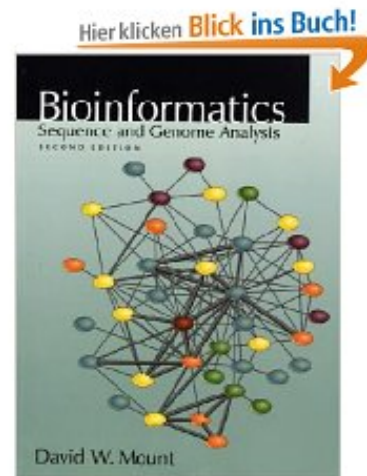
Die Veranstaltung gilt als bestanden, wenn in der abschließenden 120-minütigen Klausur über die Inhalte der Vorlesung, der Übungen und der Minipraktika mindestens die Note 4 erreicht wurde.

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€

Vorlesungsfolien ebenfalls auf
<https://www-cbi.cs.uni-saarland.de/teaching/ws-1920/sww-bioinformatik-ws-1920/>

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. Diffusionsprozesse - 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

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

1992: Sekundärstrukturvorhersage mit Neuronalen Netzwerken (PHD)
machine learning

1996: Vergleich von Proteinstrukturen mit DALI

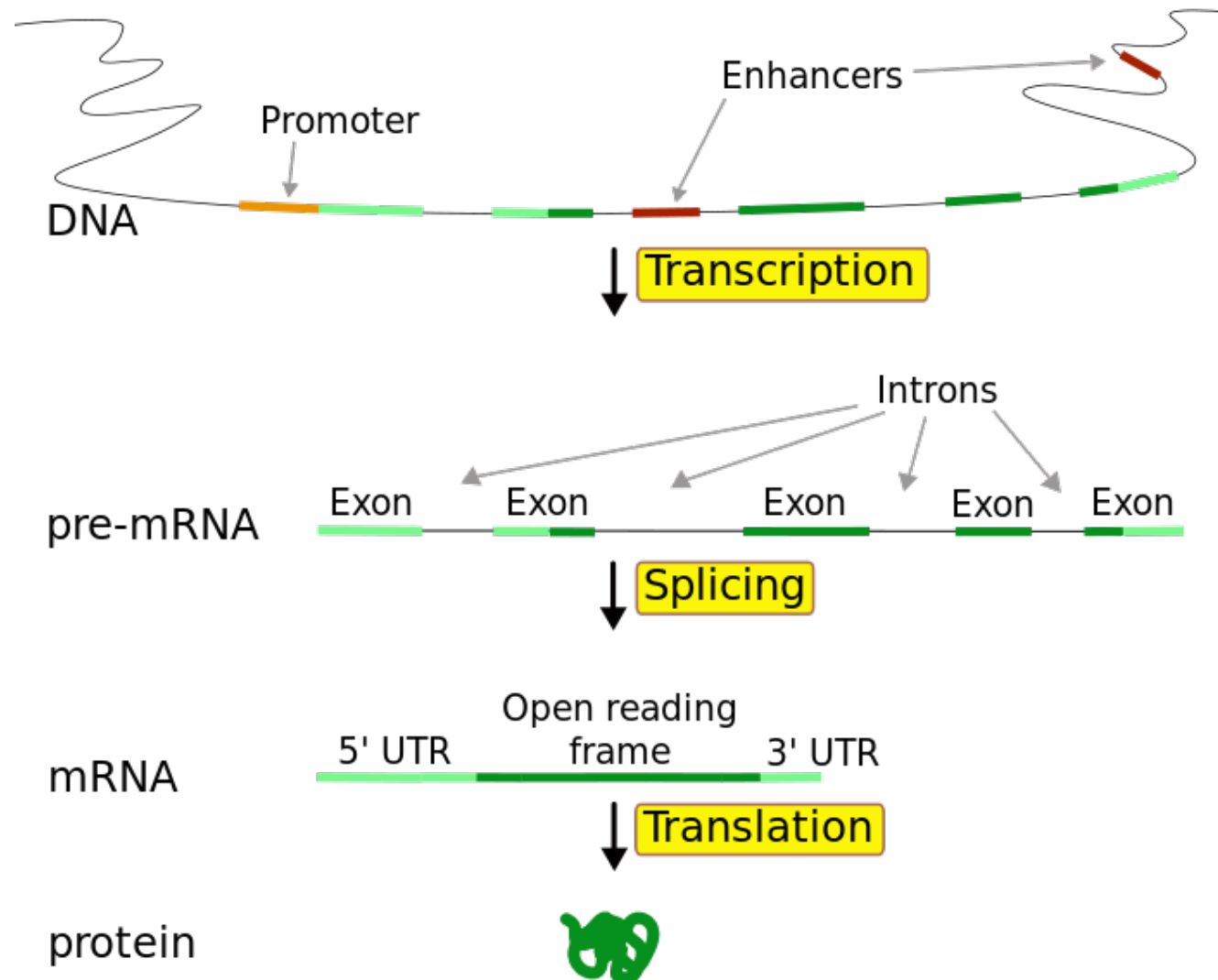
1990er: Detektion von Genexpression mit Microarrays
-> differentielle Expressionsanalyse

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

2012: ENCODE-Projekt (<https://www.encodeproject.org/>)

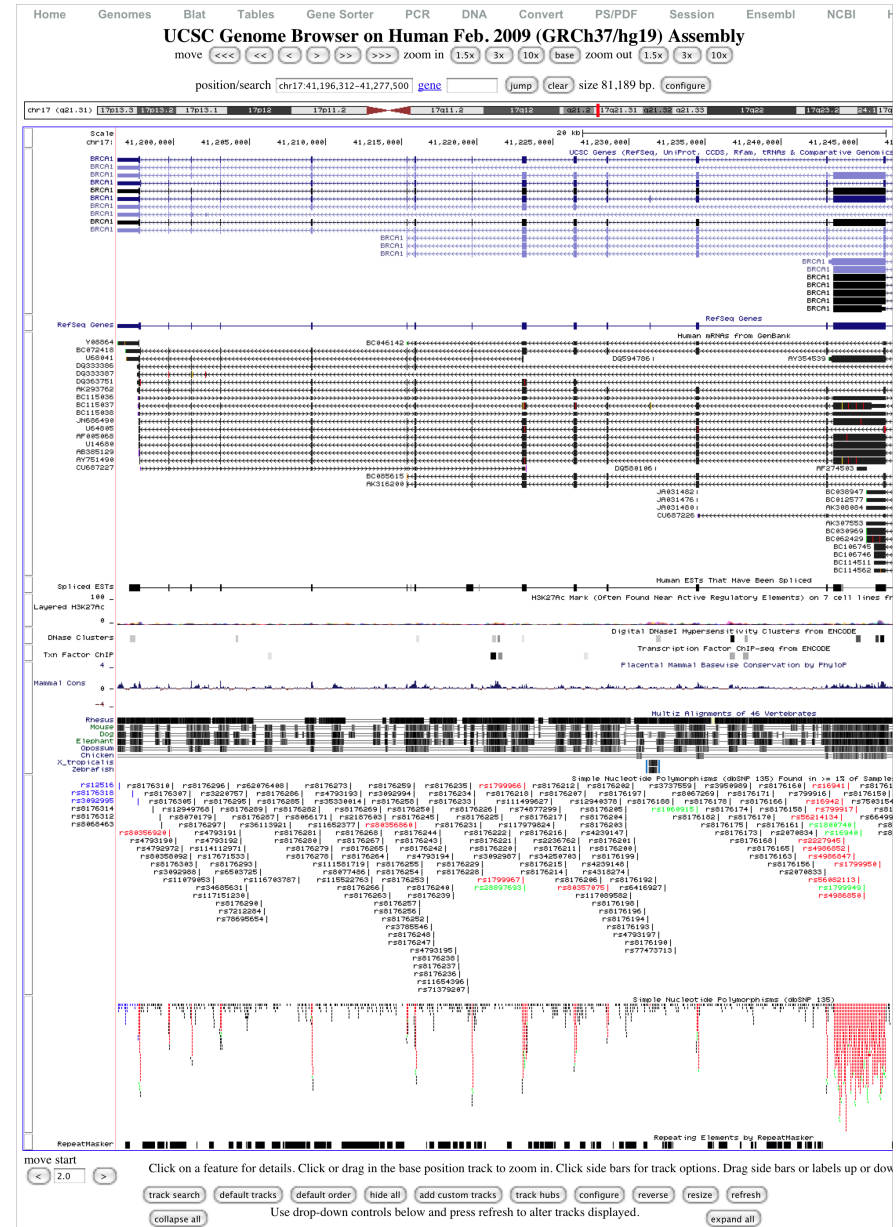
2016: The Cancer Genome Atlas Project (<https://cancergenome.nih.gov/>)

Die Struktur von Genen



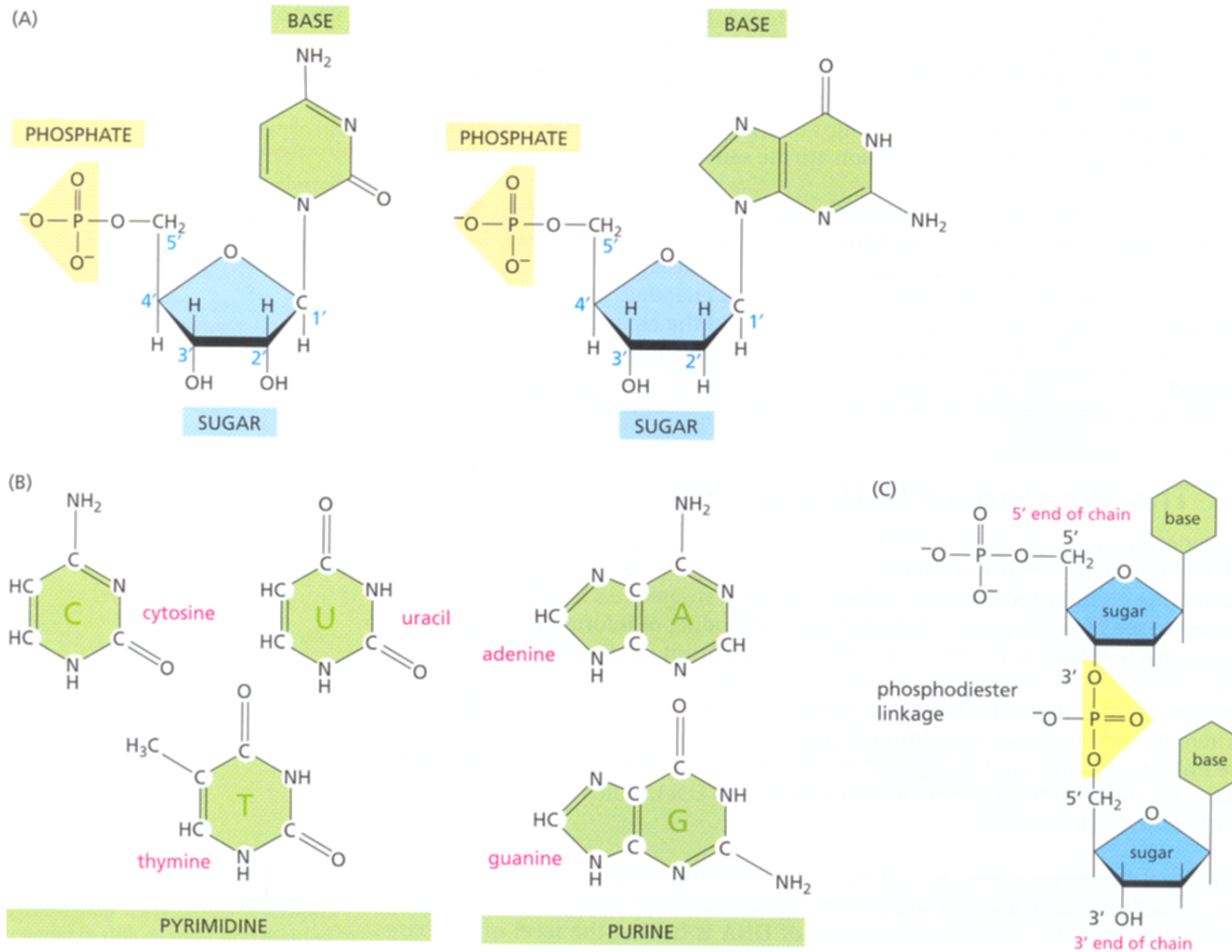
www.wikipedia.org

UCSC Genome Browser



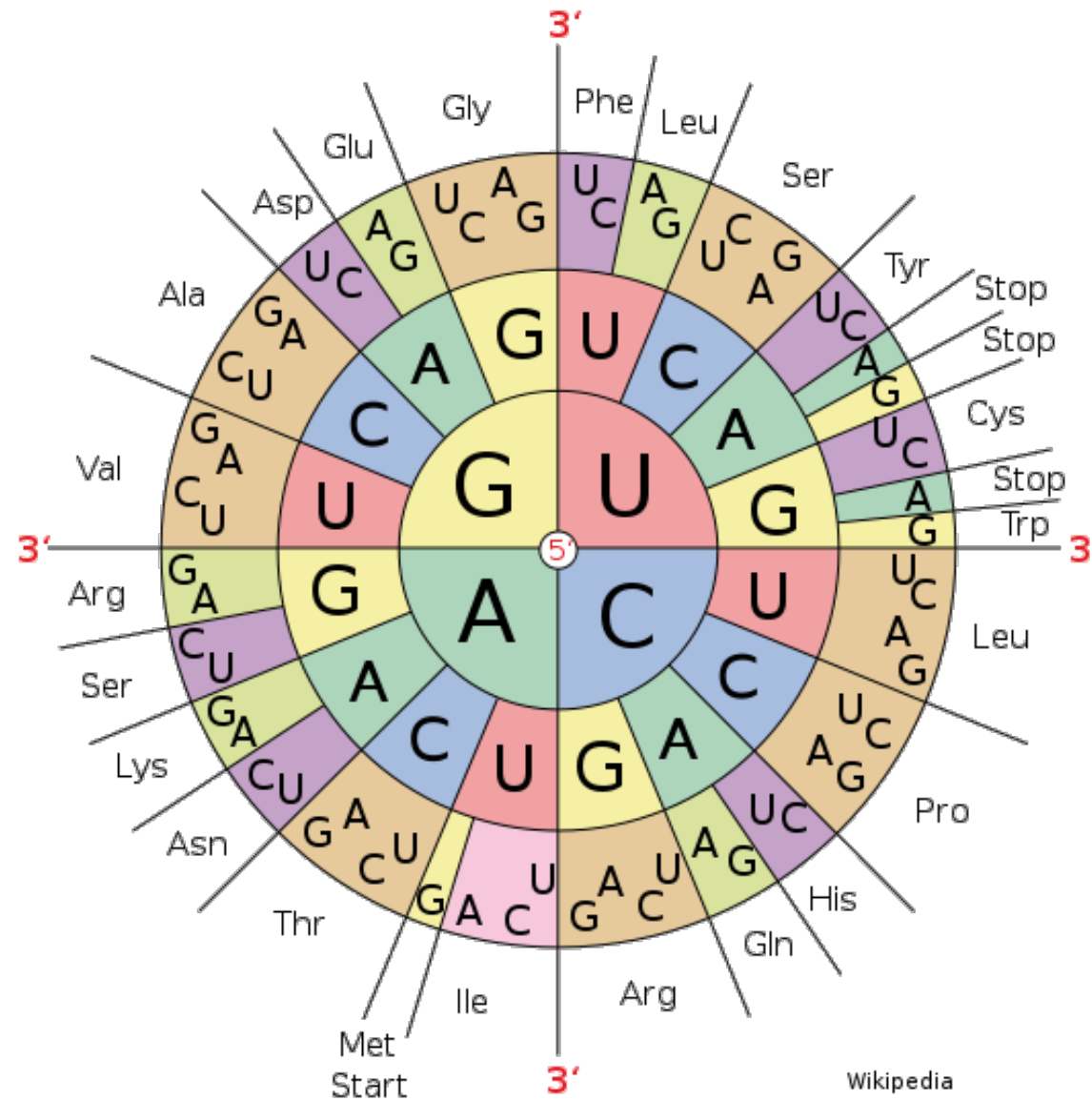
Es gibt verschiedene Assemblies
hg17, hg18, hg19 für das *humane* Genom
<http://genome.ucsc.edu/cgi-bin/hgGateway>

Die vier Nukleotidbasen



Zvelebil (2008)

Codonsonne

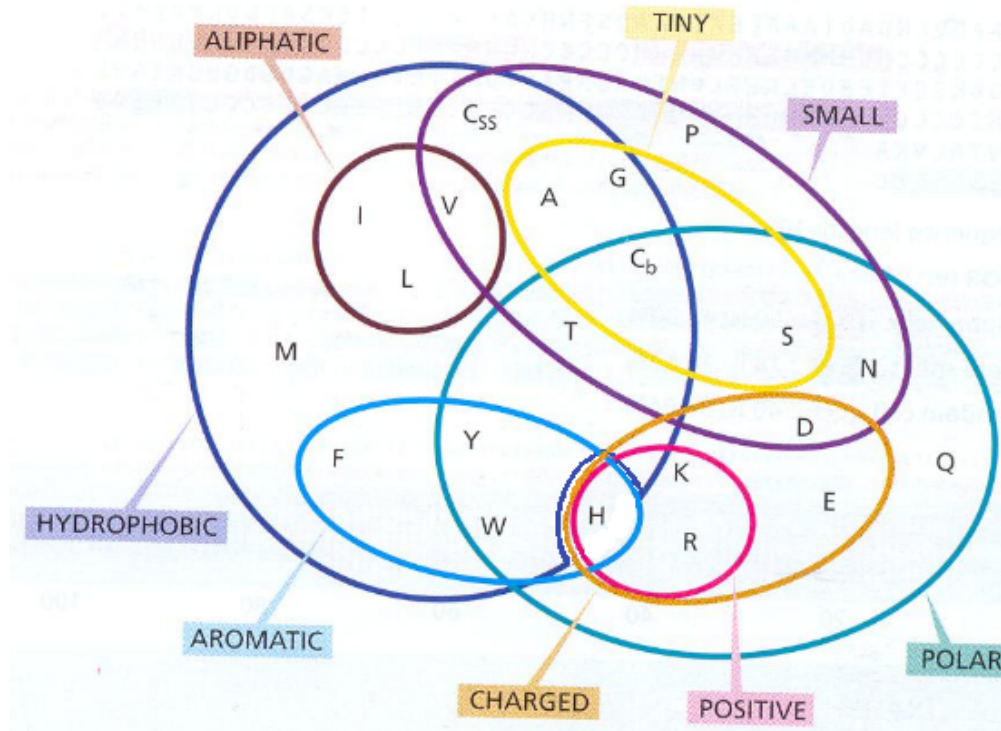


Wikipedia

Zvelebil (2008)

Eigenschaften der Aminosäuren

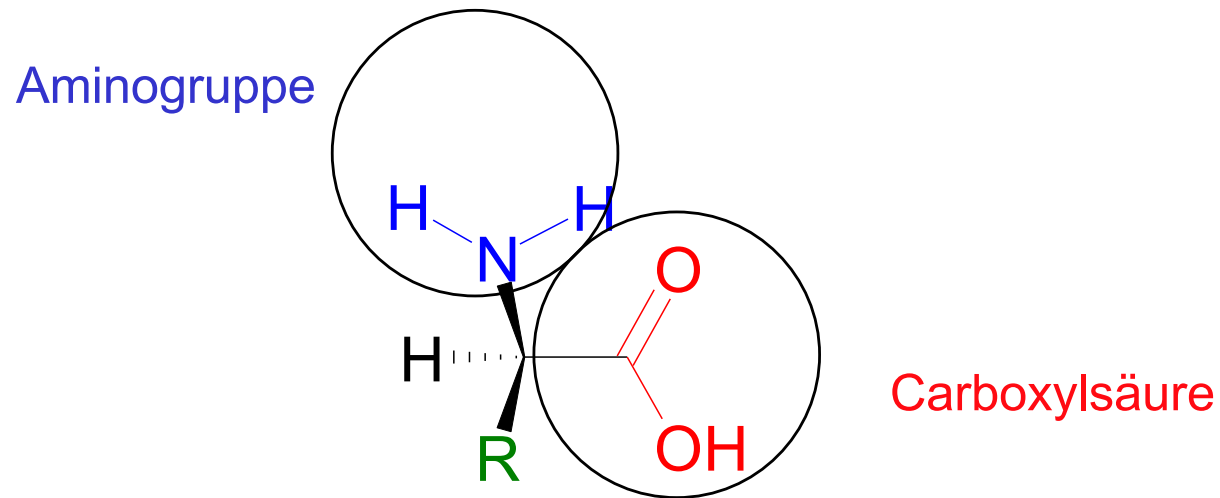
Aminosäuren unterscheiden sich in ihren physikochemischen Eigenschaften.



Q: müssen Bioinformatiker die Eigenschaften von Aminosäuren kennen?

Einleitung: Aminosäuren

Aminosäuren sind die **Bausteine** von Proteinen:



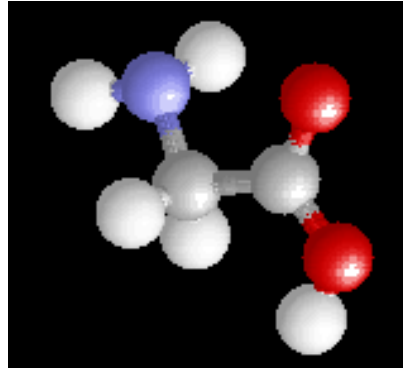
Aminosäuren unterscheiden sich hinsichtlich ihrer

- Größe
- elektrischen Ladung
- Polarität
- Form und Steifigkeit

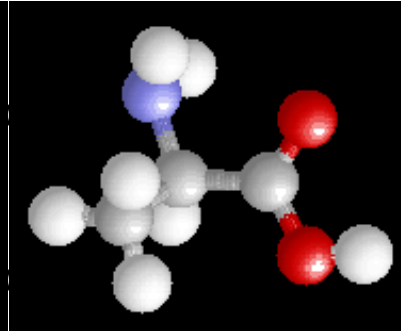
Einleitung: hydrophobe Aminosäuren

Proteine sind aus 20 verschiedenen natürlichen Aminosäuren aufgebaut

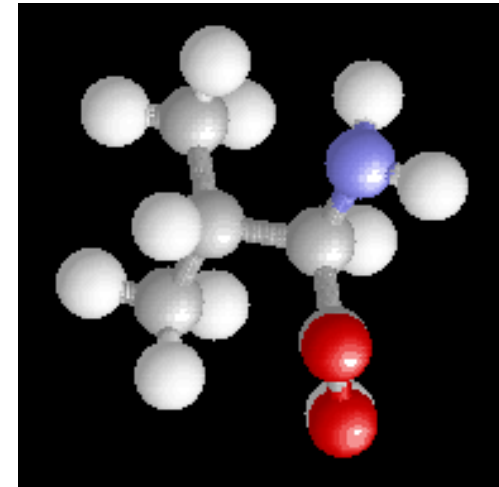
5 sind hydrophob.
Sie sind vor allem
Im Proteininneren.



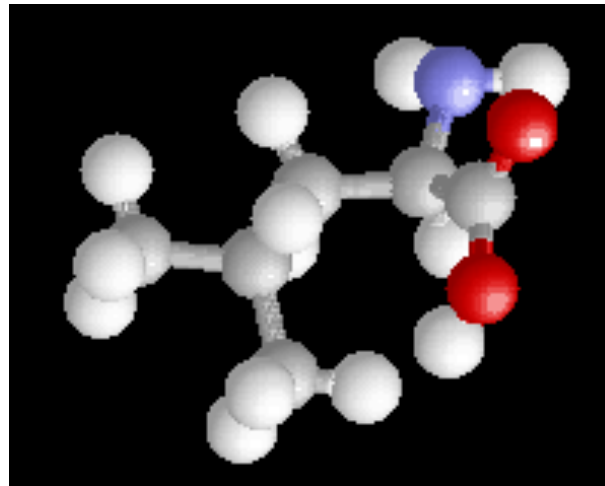
Glycine



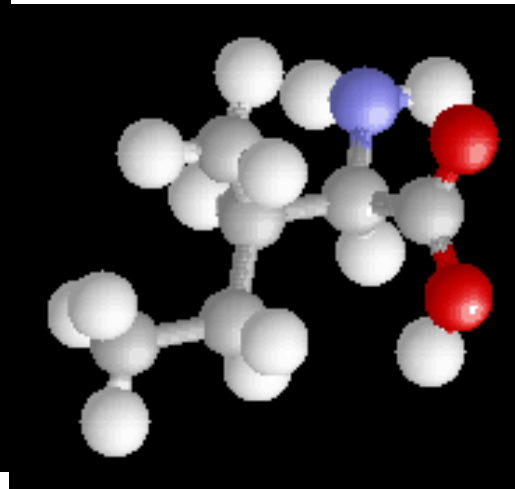
Alanine



Valine



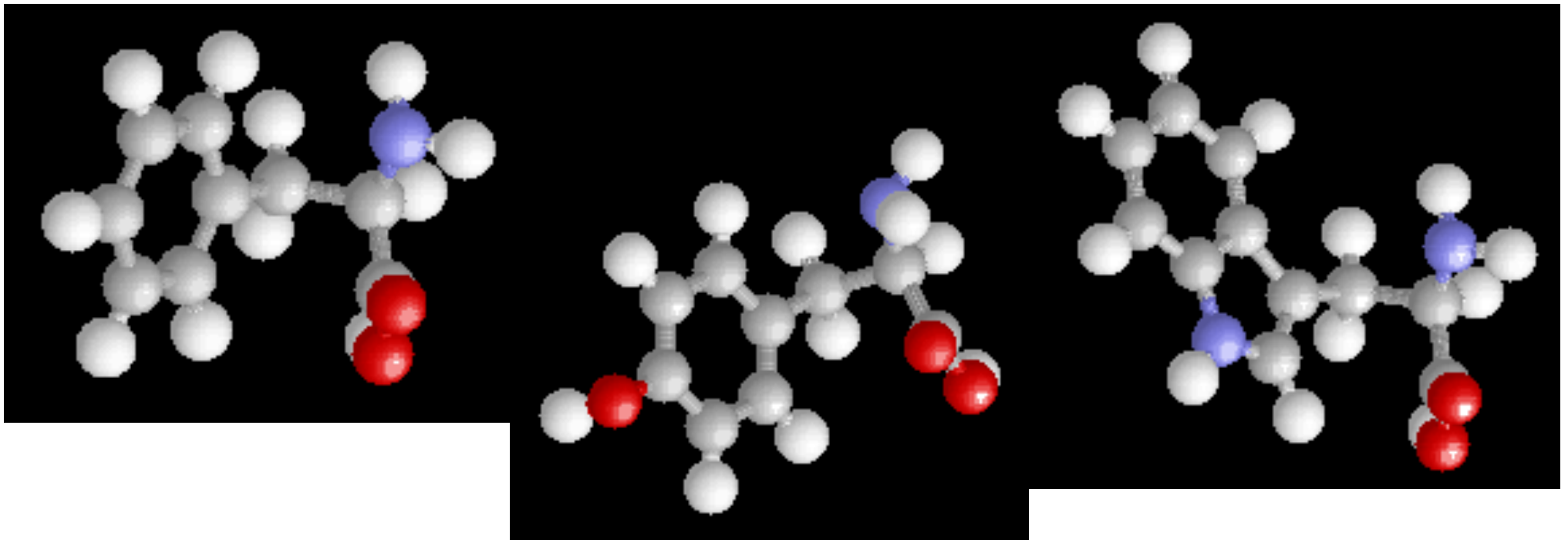
Leucine



Isoleucine

Einleitung: aromatische Aminosäuren

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



Phenylalanin

Tyrosin

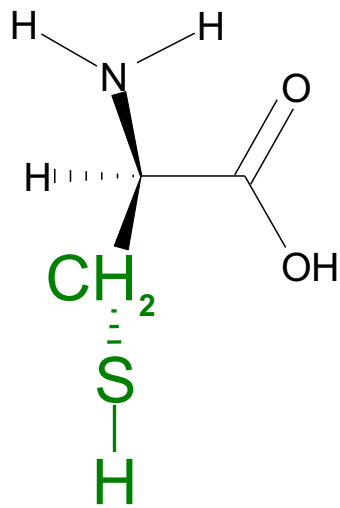
Tryptophan

Einleitung: Aminosäuren

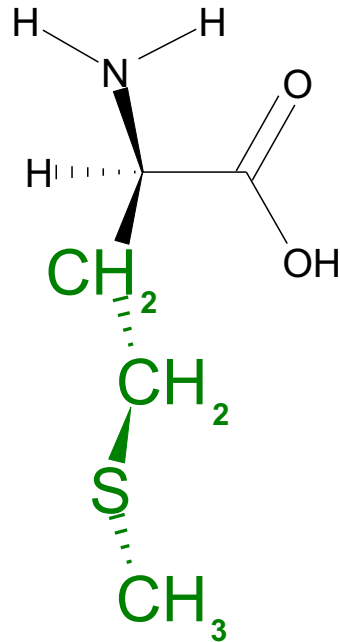
Es gibt 2 Schwefel enthaltende Aminosäuren und das ungewöhnliche Prolin.

Cysteine können Disulfidbrücken bilden.

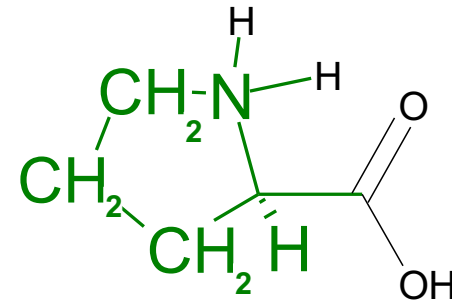
Prolin ist ein "Helixbrecher".



Cystein



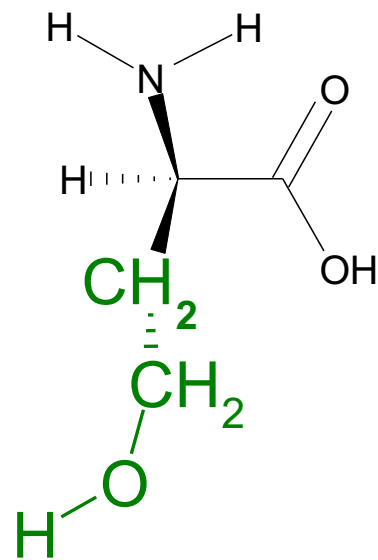
Methionin



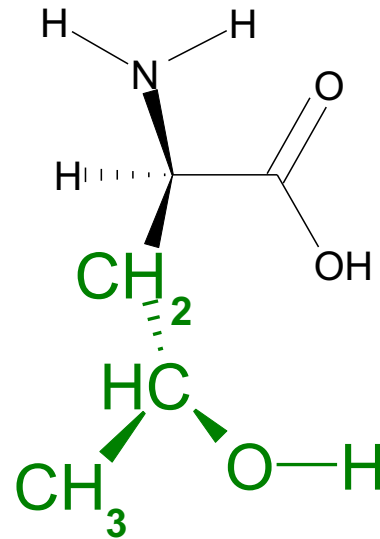
Prolin

Einleitung: Aminosäuren

Es gibt zwei Aminosäuren mit terminalen polaren Hydroxylgruppen:



Serin

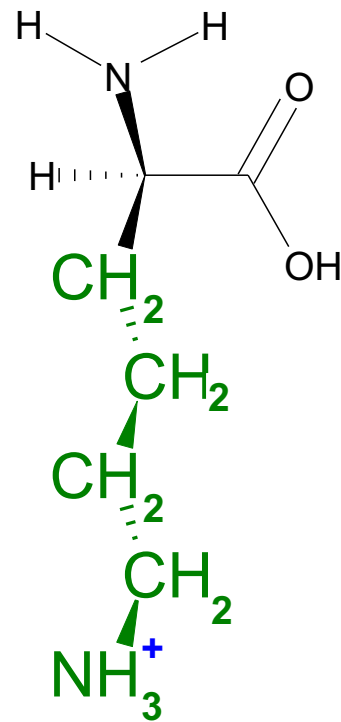


Threonin

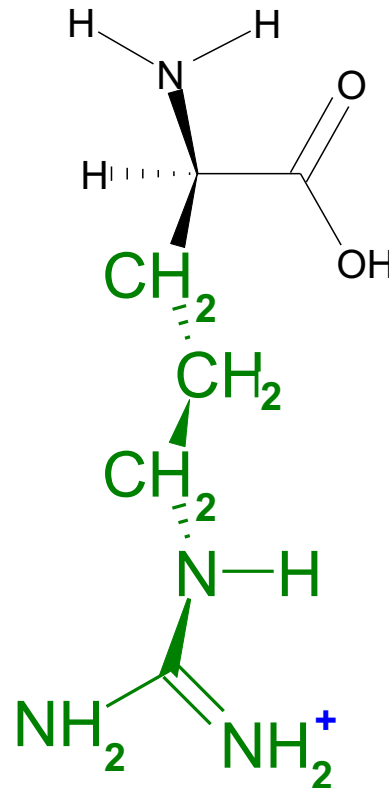
Einleitung: Aminosäuren

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

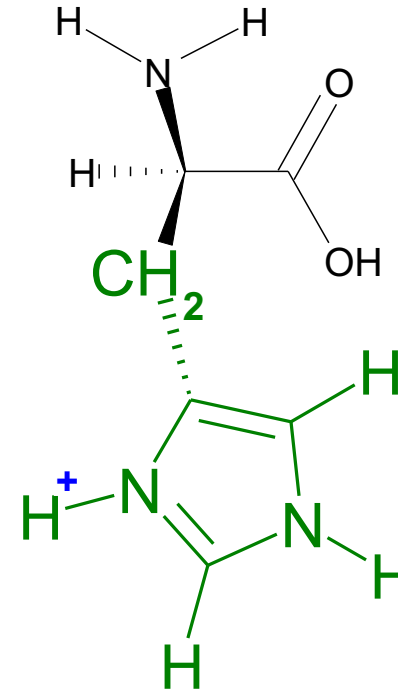
Thermophile Organismen besitzen besonders viele Ionenpaare auf den Proteinoberflächen.



Lysin



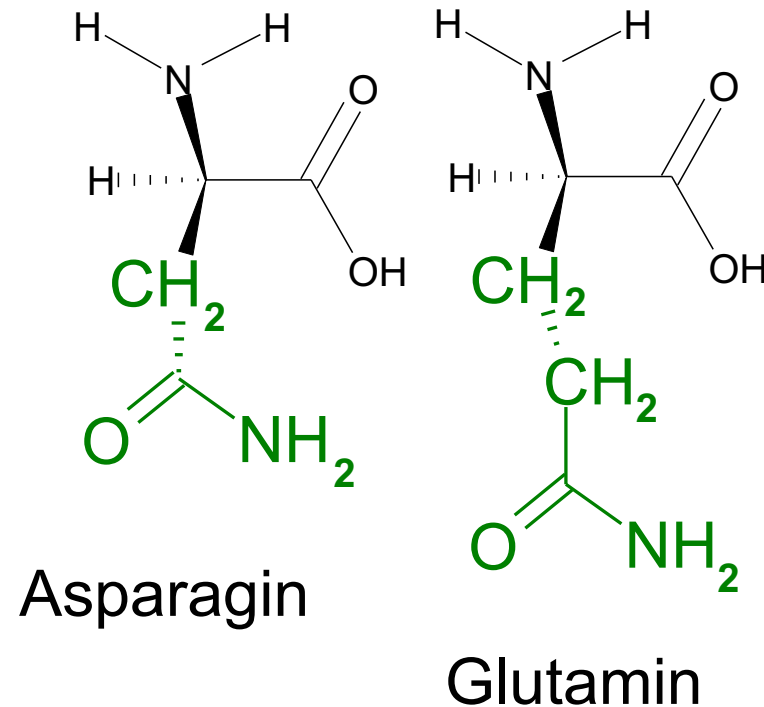
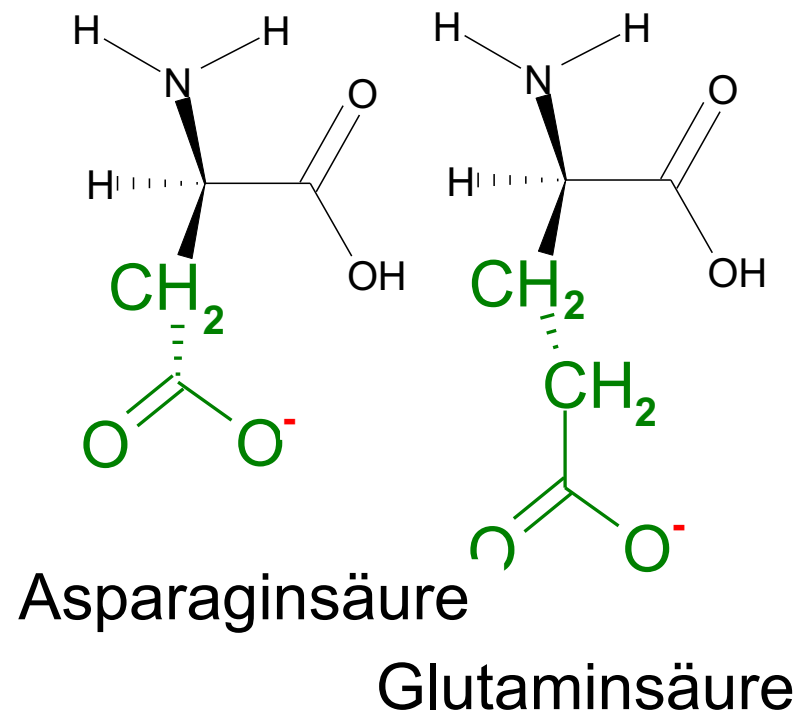
Arginin



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.



Buchstaben-Code der Aminosäuren

- **Ein- und Drei-Buchstaben-Codes** der Aminosäuren

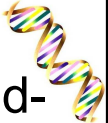
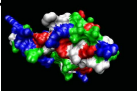
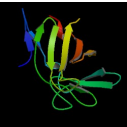
G	Glycin	Gly	P	Prolin	Pro
A	Alanin	Ala	V	Valin	Val
L	Leucin	Leu	I	Isoleucin	Ile
M	Methionin	Met	C	Cystein	Cys
F	Phenylalanin	Phe	Y	Tyrosin	Tyr
W	Tryptophan	Trp	H	Histidin	His
K	Lysin	Lys	R	Arginin	Arg
Q	Glutamin	Gln	N	Asparagin	Asn
E	Glutaminsäure	Glu	D	Asparaginsäure	Asp
S	Serin	Ser	T	Threonin	Thr

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

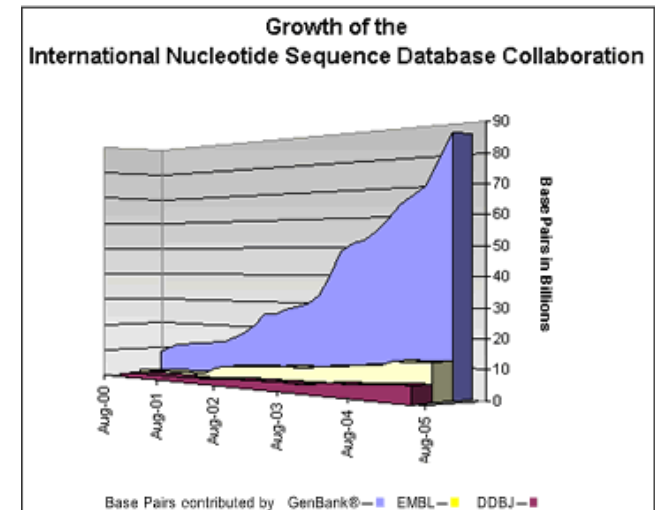
- in Okt. 2016 ~197 Mio. **Nukleotidsequenzen**
(Quelle: GenBank <http://www.ncbi.nlm.nih.gov/genbank/index.html>)

~363 Mio. WGS-Nukleotidsequenzen

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

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

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



RCSB **PDB**
PROTEIN DATA BANK

Einträge sind teilweise **redundant**,

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

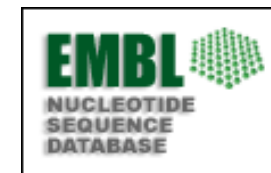
NCBI DNA-Datenbank

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

UniProt
the universal protein resource



20 Years
SIB
swissprot
EBI

- 



Webinterface: Entrez



Datenbank
wählen

melibiase - Protein Results - Mozilla Firefox

http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein&cmd=search&term=melibiase

NCBI

Search Protein for melibiase

Display Summary Show 20 Sort by Relevance Send to

All: 866 Bacteria: 366 RefSeq: 306 Related Structures: 654

Items 1 - 20 of 866

1: NP_822252 Reports
melibiase [Streptomyces avermitilis MA-4680]
gi|29827618|ref|NP_822252.1|[[29827618]]

2: CAA69852 Reports
alpha-galactosidase; melibiase [Thermoanaerobacter ethanolicus ATCC 33223]
gi|2920686|emb|CAA69852.1|[[2920686]]

3: NP_189269 Reports
glycosyl hydrolase family protein 27 / alpha-galactosidase family protein / melibiase family protein [Arabidopsis thaliana]
gi|30688284|ref|NP_189269.2|[[30688284]]

4: AAA34770 Reports
pre-alpha galactosidase (melibiase)
gi|171926|gb|AAA34770.1|[[171926]]

5: AAO78237 Reports
alpha-galactosidase (melibiase) [Bacteroides thetaiotaomicron VP1-5482]
gi|29340445|gb|AAO78237.1|[[29340445]]

6: AAO77957 Reports
alpha-galactosidase (melibiase) [Bacteroides thetaiotaomicron VP1-5482]
gi|29340163|gb|AAO77957.1|[[29340163]]

7: XP_001315888 Reports
Melibiase family protein [Trichomonas vaginalis G3]
gi|123456302|ref|XP_001315888.1|[[123456302]]

Fertig

Stichwort,
hier Name
des Proteins

Detaillierte Suche bei Entrez

Protein Result - Mozilla Firefox

http://www.ncbi.nlm.nih.gov/sites/entrez

NCBI

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

Limits PreviewIndex History Clipboard Details

Display Summary Show 20 Sort by Relevance Send to

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

1: [O34645](#) Reports

Alpha-galactosidase (Melibiase)

gi|3912990|sp|O34645|AGAL_BACSU[3912990]

BLINK, Conserved Domains, Links

Fertig

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

http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=3912990

Annotations:

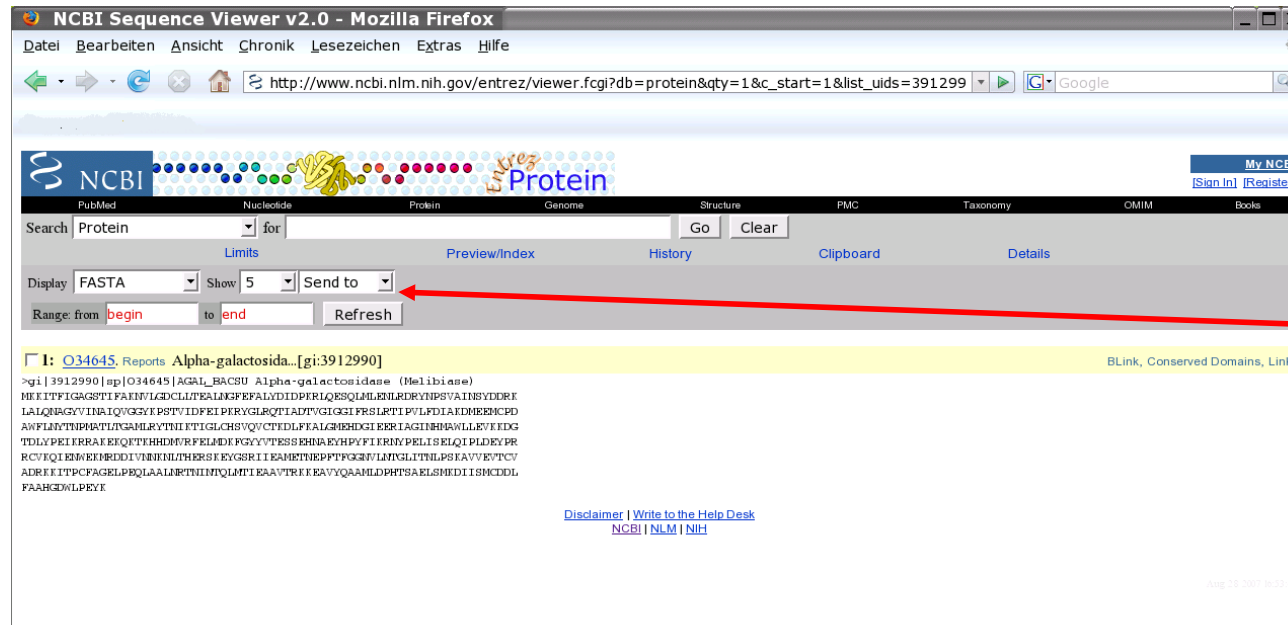
- Gene:** melA
- Locus Tag:** BSU30300
- Region Name:** Glyco_hydro_4
- Note:** Family 4 glycosyl hydrolase; pfam02056
- db_xref:** CDD:65809
- Region:** 3..162
- Gene:** melA
- Locus Tag:** BSU30300
- Region Name:** Ldh_1_N
- Note:** lactate/malate dehydrogenase, NAD binding domain. L-lactate dehydrogenases are metabolic enzymes which catalyze the conversion of L-lactate to pyruvate, the last step in anaerobic glycolysis; pfam00056
- db_xref:** CDD:79437
- Site:** 148
- Gene:** melA
- Locus Tag:** BSU30300
- Site Type:** binding
- Inference:** non-experimental evidence, no additional details recorded
- Note:** Substrate (By similarity)."
- Site:** 169
- Gene:** melA
- Locus Tag:** BSU30300
- Site Type:** metal-binding
- Inference:** non-experimental evidence, no additional details recorded
- Note:** Manganese (By similarity)."
- Site:** 170
- Gene:** melA
- Locus Tag:** BSU30300
- Site Type:** active
- Inference:** non-experimental evidence, no additional details recorded
- Note:** Proton donor (By similarity)."
- Site:** 199
- Gene:** melA
- Locus Tag:** BSU30300
- Site Type:** metal-binding
- Inference:** non-experimental evidence, no additional details recorded
- Note:** Manganese (By similarity)."

ORIGIN

```
1 mdkitfigag stfaknvg delltealg fefalydidp krlgesqlnl enldrnyps
61 vainsyddrk lalgnagvvi naigvgykp stvidfeipk ryqlrtiad tvigqifrs
121 lrtipvlfdi akdmeemcpd awflnytnpm atttgamry tnktgtich svqvctkdlf
181 kalgmebdgi eeriaginhm awllekkdg tdlpeikrr akakgtkdh dmrvfelndk
241 fgyvtesse hnaeghyfi knypelise lqipldeypr rcvkiemse knrddivnwk
301 nlthrskey garieamet nepftfgnv lntglitnlp skavvevtev adrkkitpcf
361 agelpqlea ltrntintql mtieavtrk keavyqaam dphtsaalam kdliismcdl
421 faahgdvlpe yk
//
```

Fertig

Fasta-Format



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

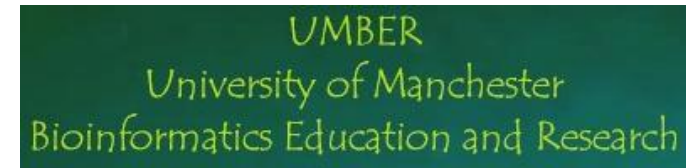
>DNA-Sequenz-Bezeichnung
ACGT

....

>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

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

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	16	SVVVAGGGSTFTPGIV	GLVG_RCOLI	5	5	-		
		SIVVAGGGSTFTPGIV	GLVG_BACSU	7	7	-		
		KITFIGAGSTIFVKNI	AGAL_RCOLI	6	6	-		
		KITFIGAGSTIFAKNV	AGAL_BACSU	3	3	-		
		KVVTIGGGSSYTPELL	CRIF_RCOLI	6	6	-		
		KIVTIGGGSSYTPELV	CRIF_BACSU	6	6	-		
		SILVAGGGSTFTPGII	NALM_FUSHR	5	5	-		
		KIAYIGGGSQGWARS	LPID_BACSU	11	11	-		
		Motif 2	17	ALSAADIVIISILPGSL	LPID_BACSU	76	49	-
				ALRDADFVVAFQIGGY	AGAL_RCOLI	75	53	-
AFTDIDFVNAHVRGKY	NALM_FUSHR			74	53	-		
ALKDADFVTTQLRVGQL	CRIF_RCOLI			77	55	-		
AFSDVDVFNHVRGKY	GLVG_RCOLI			74	53	-		
AFTDIDFVNAHVRGKY	GLVG_BACSU			76	53	-		
ALKDADFVTTQFRVGLL	CRIF_BACSU			77	55	-		
ALQNGYVIAHAIQVGGY	AGAL_BACSU			72	53	-		
Motif 3	14			LDEQIPLKYGVVGG	GLVG_BACSU	97	4	-
				LDERIPLRHGVVGQ	NALM_FUSHR	95	4	-
		LDERIPLRHGVVGQ	GLVG_RCOLI	95	4	-		
		TDFEVCKRHGLEQT	AGAL_RCOLI	97	5	-		
		LDERIPLSHGYLGQ	CRIF_RCOLI	98	4	-		
		KDERIPLKYGIVGQ	CRIF_BACSU	98	4	-		
		IDFERIPKRYGLRQT	AGAL_BACSU	94	5	-		
		VDVHLPRRCGIYQS	LPID_BACSU	97	4	-		
		Motif 4	21	DTVGGGIIIRGLRAVPIFAKI	LPID_BACSU	113	2	-
				ETCGPGGIAYGHRISIGGVIGL	NALM_FUSHR	109	0	-
ETCGPGGIAYGHRISIGGVLEL	GLVG_RCOLI			109	0	-		
ETCGPGGIAYGHRISIGGVLEI	GLVG_BACSU			111	0	-		
DTLGGGIIHRALETIPHLWQI	AGAL_RCOLI			113	2	-		
ETNGGGGLFKGLRTIPVILKEI	CRIF_BACSU			112	0	-		
DTVGIGGIIFRSLRTIPVLFDI	AGAL_BACSU			110	2	-		
ETNGAGGLFKGLRTIPVIFDI	CRIF_RCOLI			112	0	-		
Motif 5	18			PDAWMLNYSHPAAIVAKA	GLVG_BACSU	140	8	-
				PHAWMLNYSHPAAIVAKA	GLVG_RCOLI	138	8	-
		PDATHLNYSHPAHITWA	AGAL_RCOLI	142	8	-		
		PDATF LNYTHPATLTGA	AGAL_BACSU	139	8	-		
		PHAWMLNYSHPAAIVAKA	NALM_FUSHR	138	8	-		
		PHAWVINFTHPAGHVTBA	CRIF_RCOLI	141	8	-		
		PHAWLINFTHPAGHVTBA	CRIF_BACSU	141	8	-		
		PHAWLINFTHPAGHVTBA	CRIF_BACSU	142	8	-		
		Suchen:	tetL	Abwärts	Aufwärts	Hervorheben	☐ Groß-/Kleinschreibung	Das Seitenende wurde erreicht, Suche vom Seitenanfang fortge

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 entry is part of the 'MIP' superfamily. The FingerPRINTScan result (inset) diagnoses the sequence both as a member of the MIP superfamily and as an aquaporin 6 subtype.

Names and origin

Protein names	Submitted name: Putative uncharacterized protein DKFZp434D2030 (EMBL CAB70889.1)
Gene names	Name:DKFZp434D2030 (EMBL CAB70889.1)
Organism	Homo sapiens (Human) (EMBL CAB70889.1)
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota > Metazoa > Chordata > Craniata > Vertebrata > Euteleostomi > Mammalia > Eutheria > Euarchontoglires > Primates > Haplorhini > Catarrhini > Hominidae > Homo

Cross-references

Family and domain databases

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]

PRINTS42_0 and matrix blos62

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

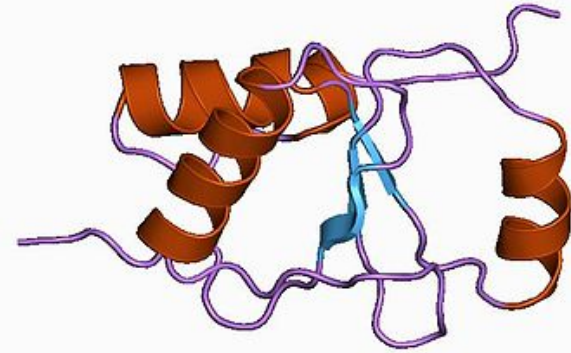


(<http://pfam.xfam.org>)

- 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)

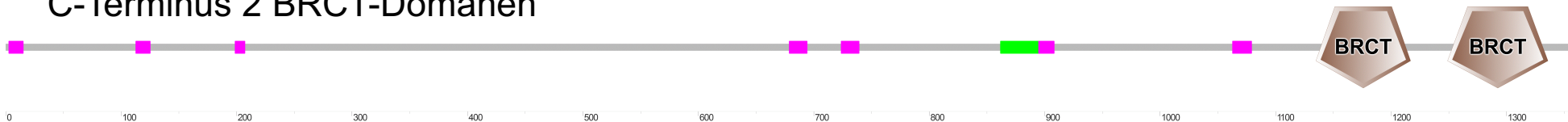
Pfam – Profil für BRCA1

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“.

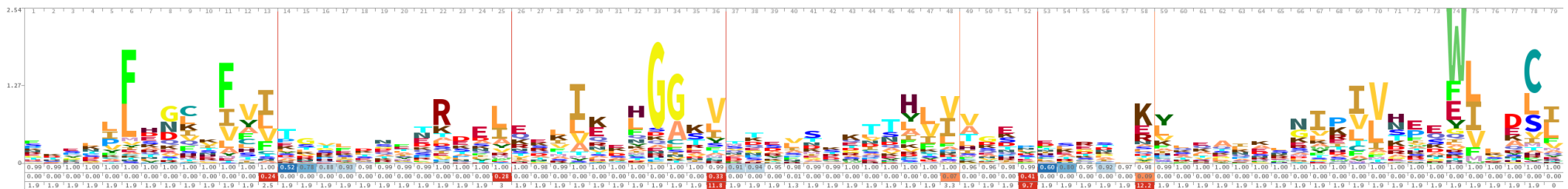


Kristallstruktur der BRCT-Domäne

<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



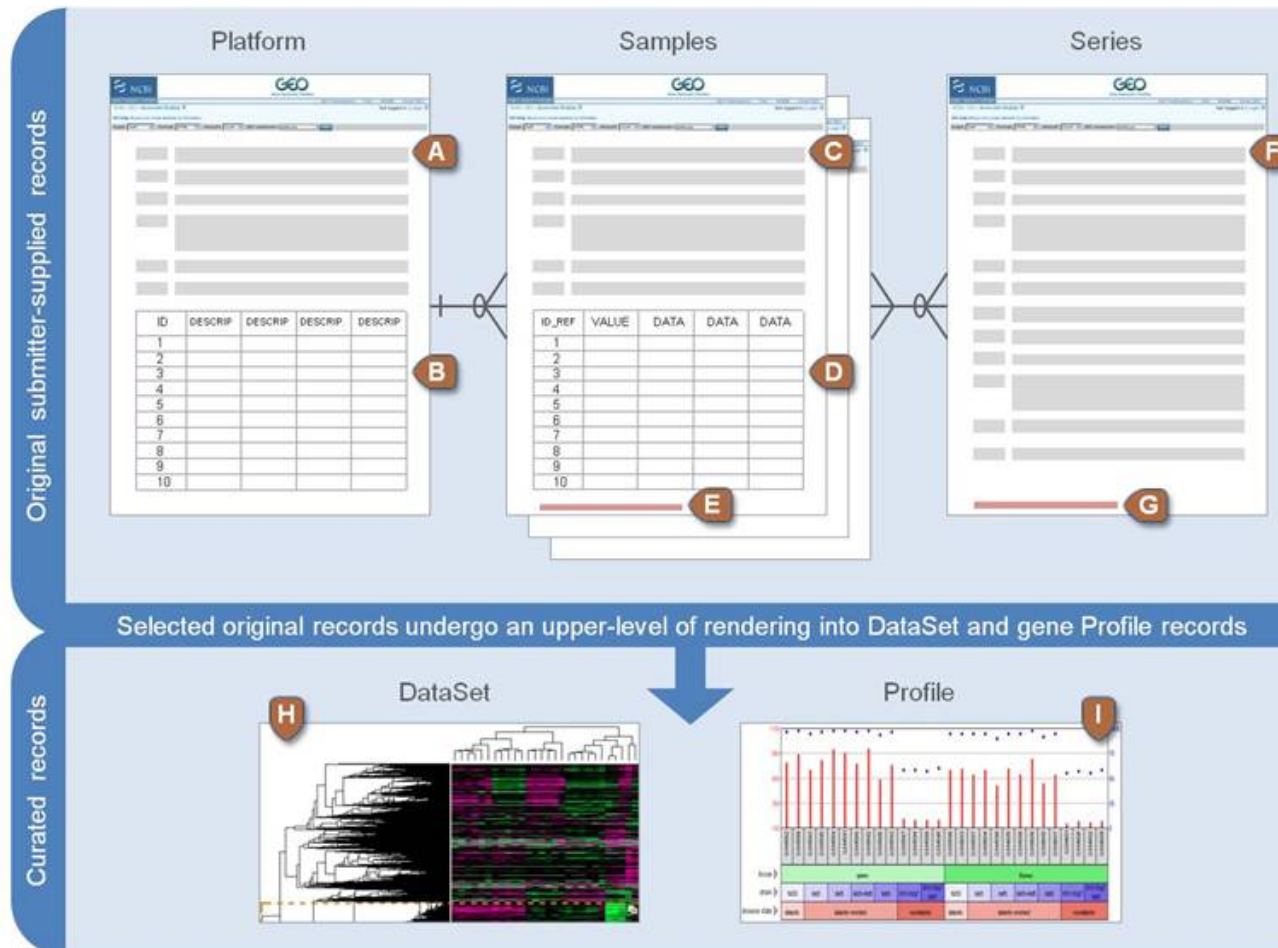
HMM-Logo von PFAM für die BRCT-Domäne



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



Lipid Metabolism Signatures in NASH-Associated HCC— Letter

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

Cancer Res Published OnlineFirst April 28, 2014.

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.

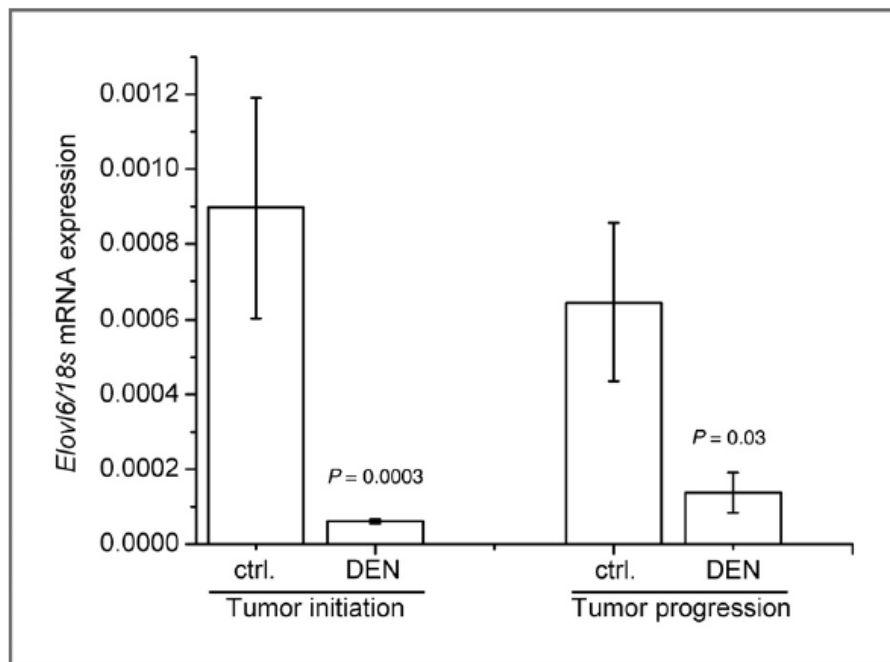


Figure 2. Wild-type mice were treated with the carcinogen DEN at the age of 2 weeks. Livers were analyzed after 24 weeks to assess the tumor initiation state. Analyses in the tumor progression stadium were done after 36 weeks. *Elovl6* mRNA expression as determined by real-time reverse transcriptase 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.

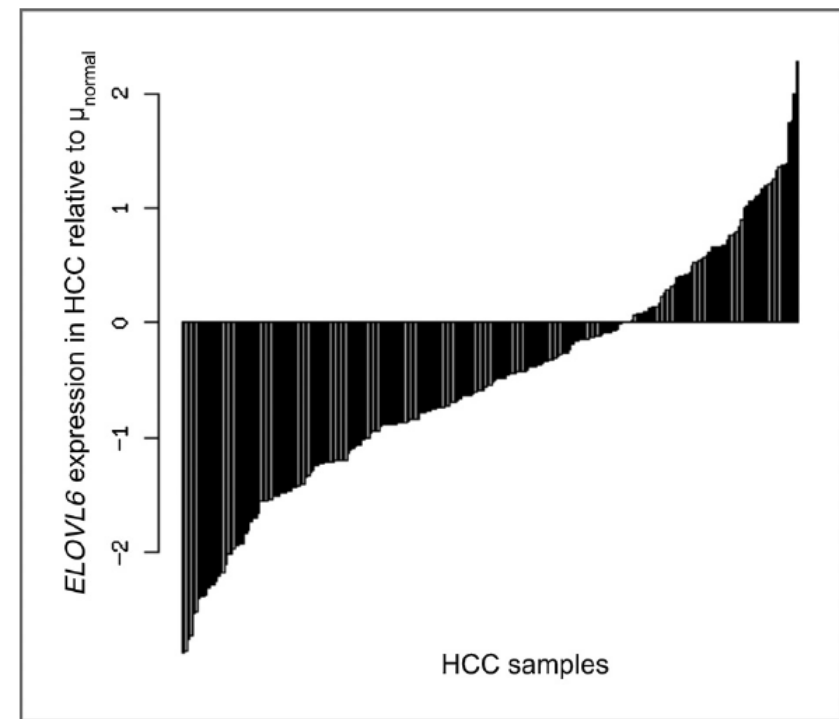


Figure 1. mRNA levels of *ELOVL6* in 247 human HCC samples relative to the mean of 239 nontumor liver tissue (μ_{normal}). Samples of dataset GSE14520 [$\log_2$ (expression) values from GEO after Robust Multi-array Average normalization] were mapped to hgu133a.db using bioconductor. Significance values: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