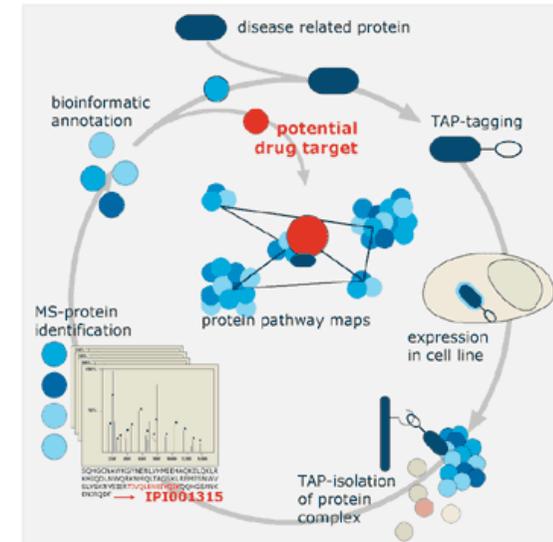


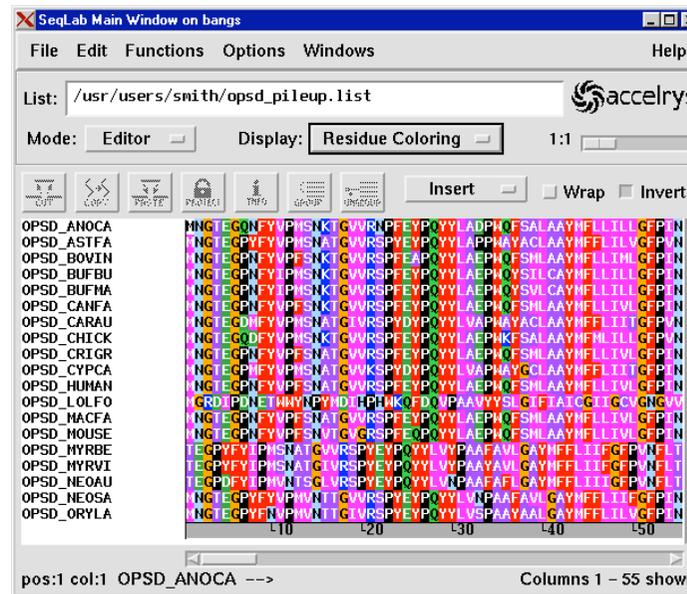
Softwarewerkzeuge der Bioinformatik

Inhalt dieser Veranstaltung: Softwarewerkzeuge kennenlernen für

- I Sequenzanalyse
- II Analyse von Proteinstruktur und Ligandenbindung
- III 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
Seminarraum 007, Geb. E 2 1

Donnerstag 10.15 – 12.00 Uhr
Dozent: Prof. Helms

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

Übungen „**hands-on**“ Beginn heute am 22.10:

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.

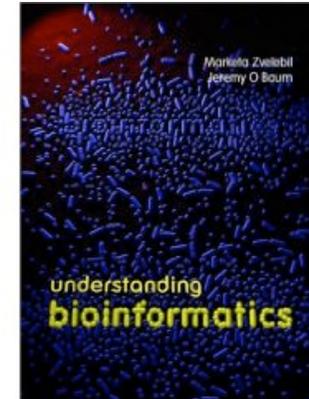
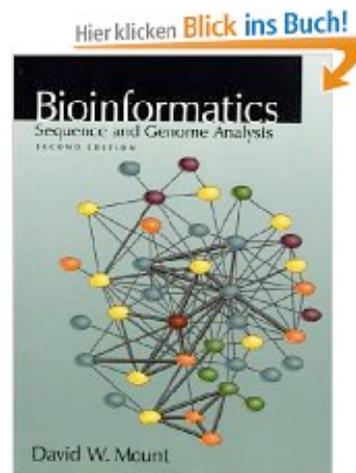
Die Veranstaltung gilt als bestanden, wenn in der abschließenden 120-minütigen Klausur über die Inhalte der Vorlesung, der Übungen und der Minipraktika mindestens die Note 4 erreicht wurde.

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

- 2 Einführung, Datenbanken
- 3 Paarweises Sequenzalignment
- 4 Multiples Sequenzalignments;
Phylogenie
- 5 Genvorhersage, Motivsuche

II Proteinstruktur

- 5. Proteinstruktur; Sekundärstruktur
- 6. Homologie-Modellierung

III Zellsimulationen/Netzwerke

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

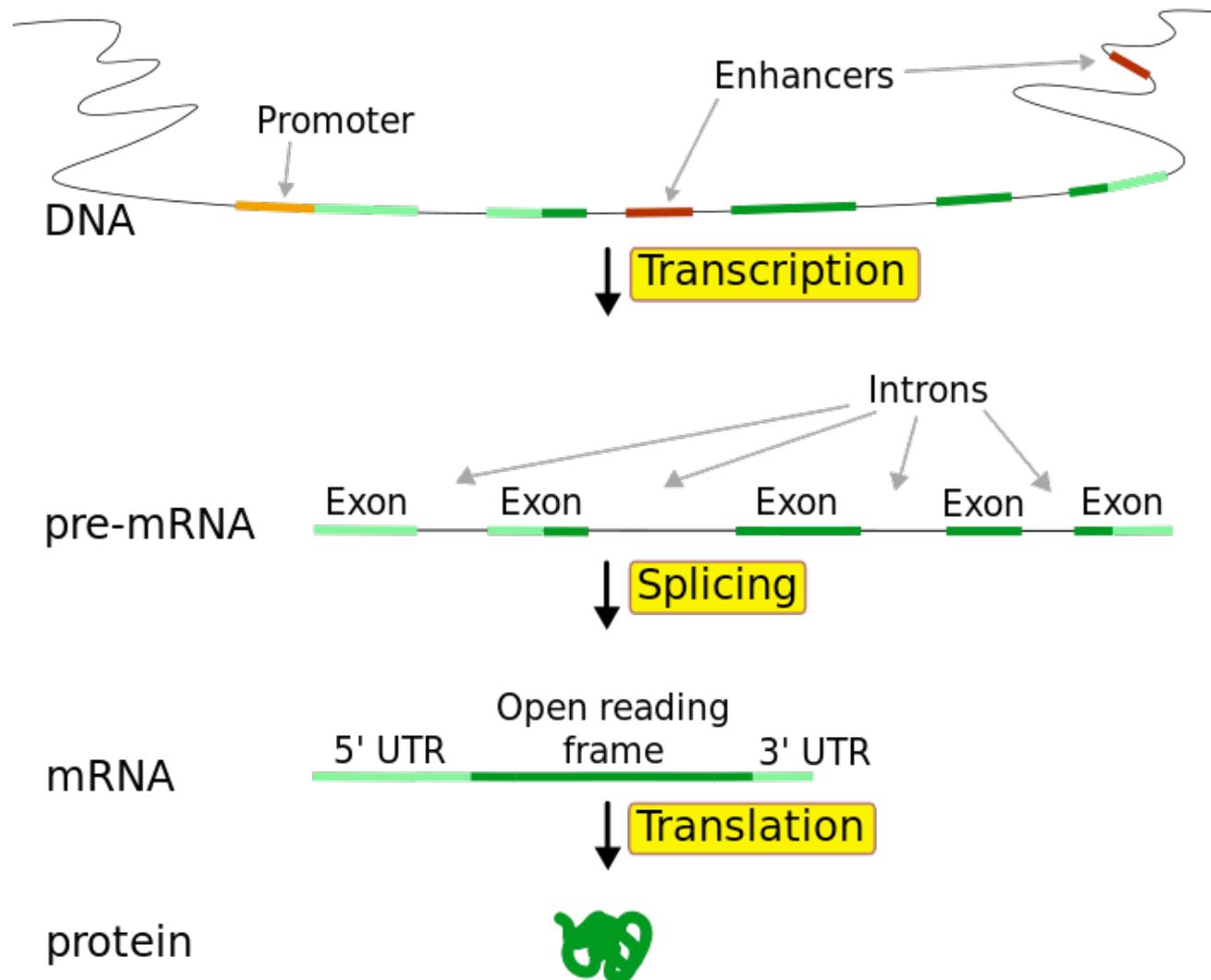
1992: Sekundärstrukturvorhersage mit Neuronalen Netzwerken (PHD)
machine learning

1996: Vergleich von Proteinstrukturen mit DALI

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

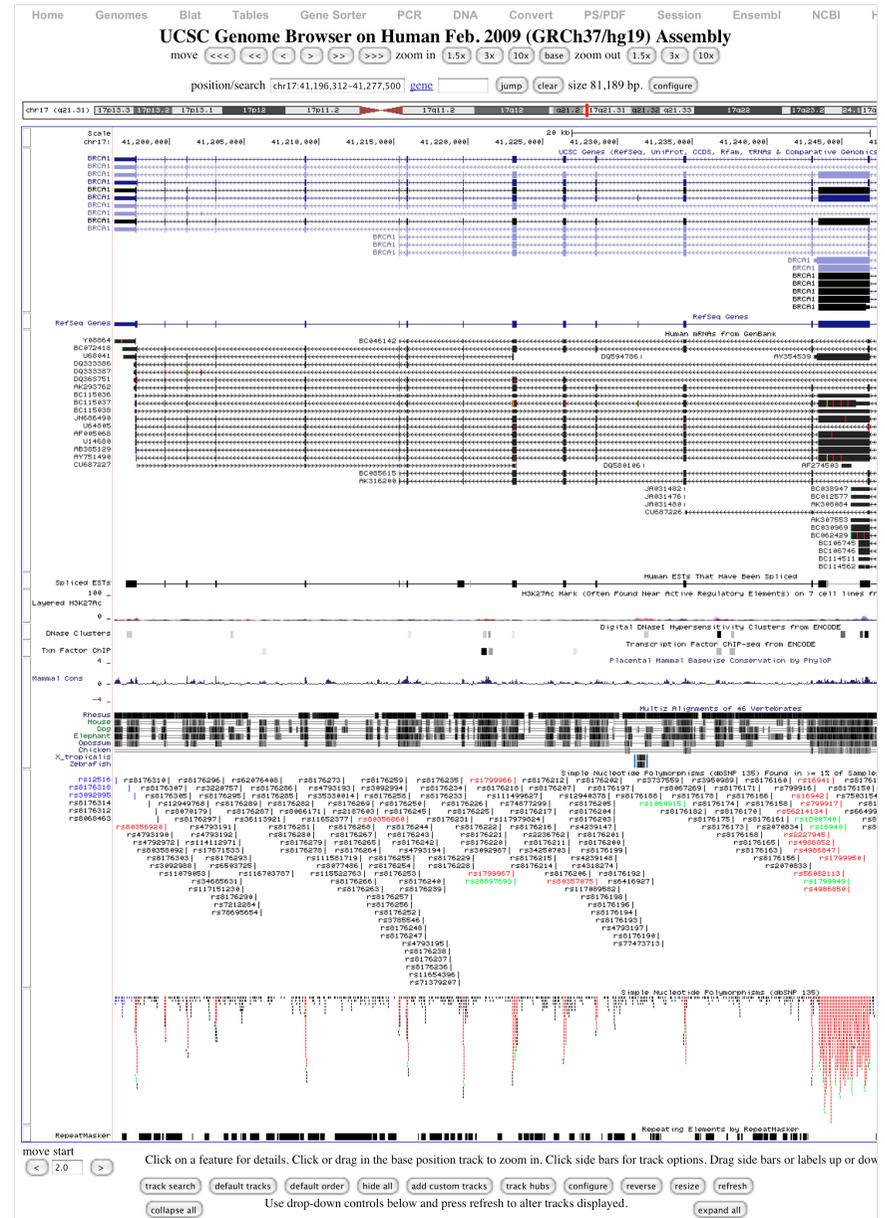
2012: ENCODE-Projekt

Die Struktur von Genen



www.wikipedia.org

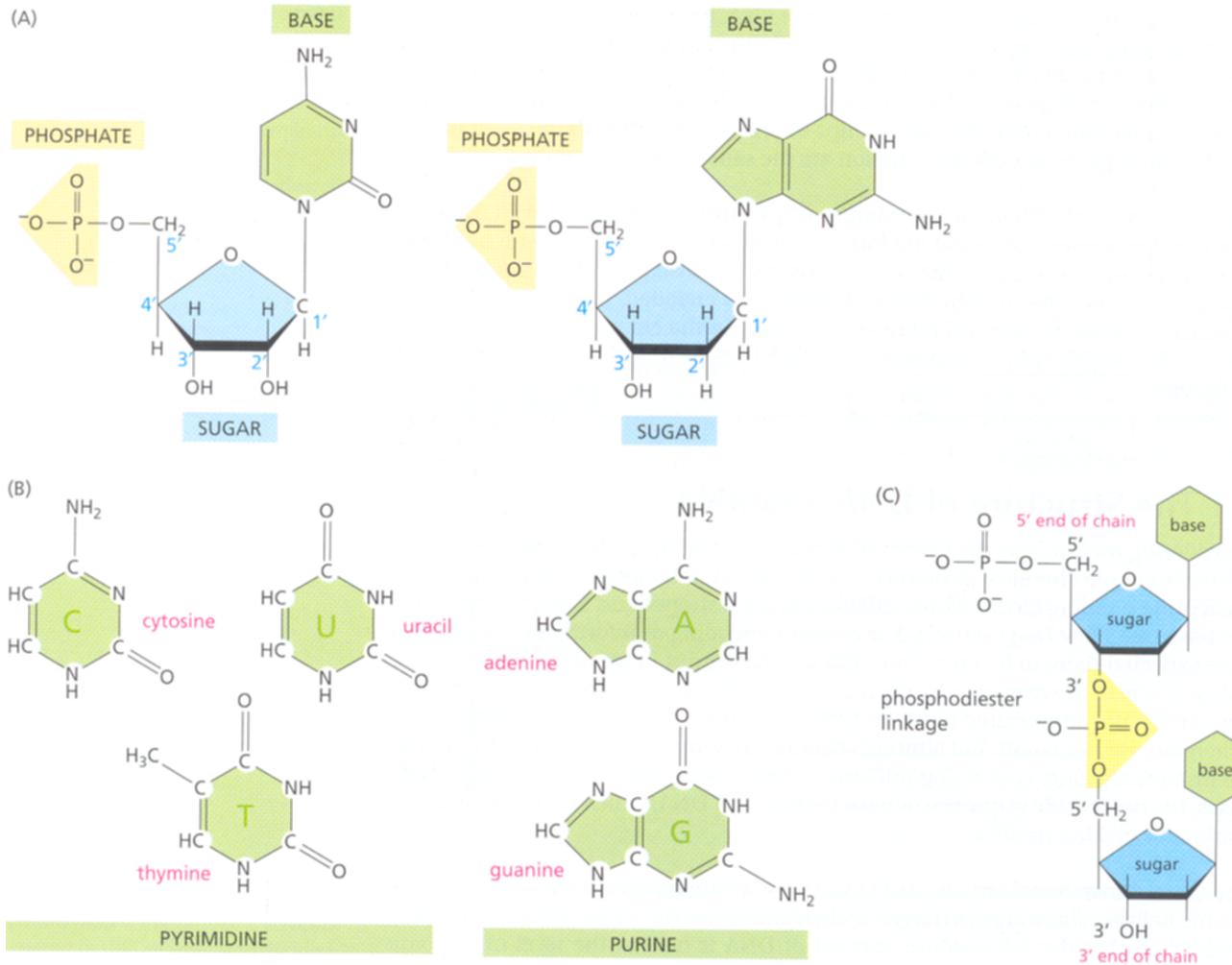
UCSC Genome Browser



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

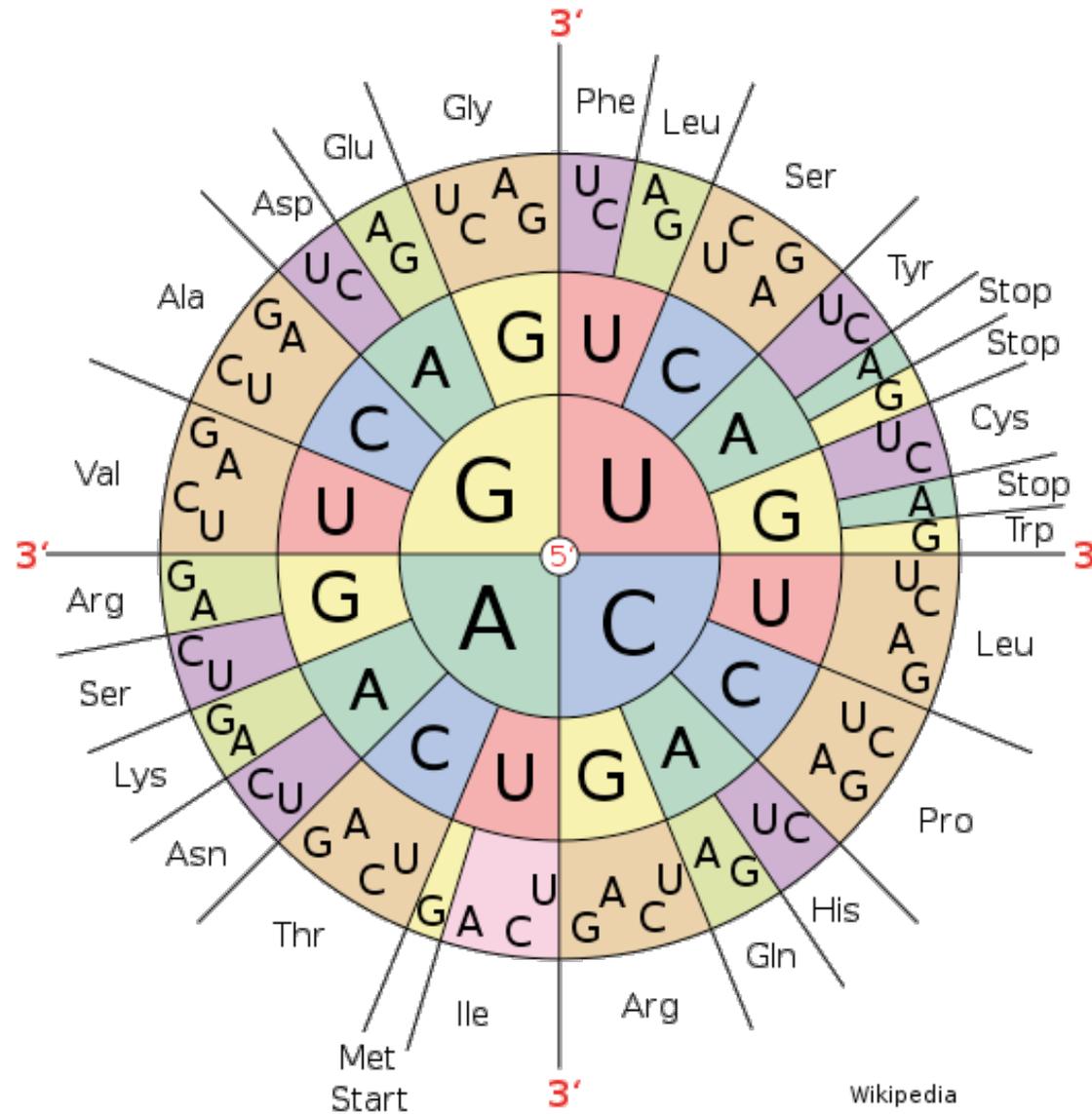
<http://genome.ucsc.edu/cgi-bin/hgGateway>

Die vier Nukleotidbasen



Zvelebil (2008)

Codonsonne

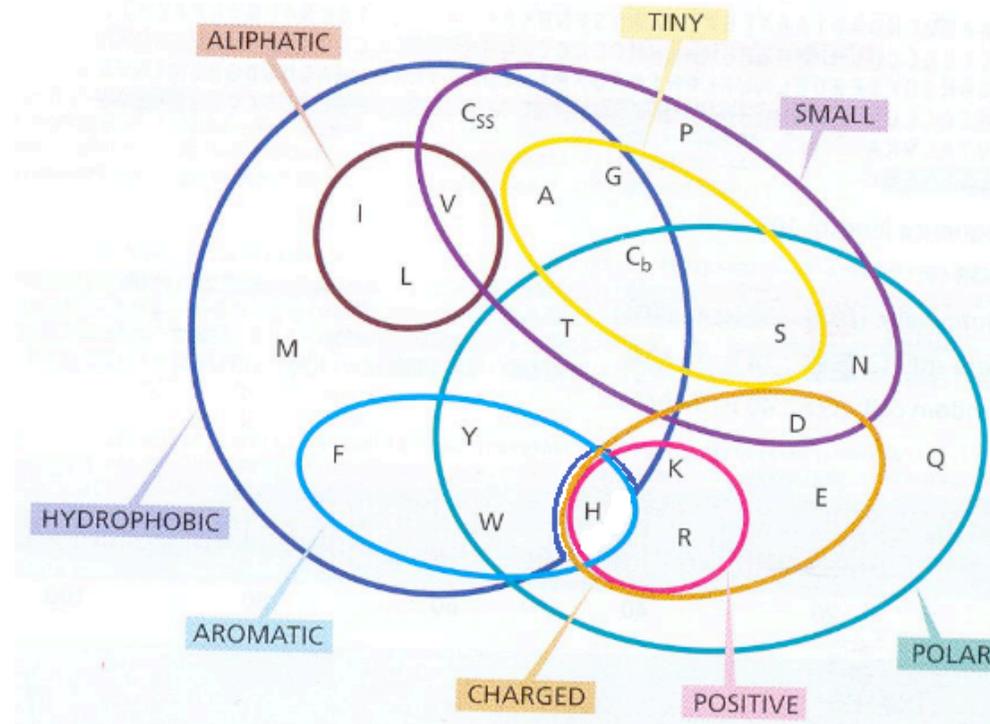


Wikipedia

Zvelebil (2008)

Eigenschaften der Aminosäuren

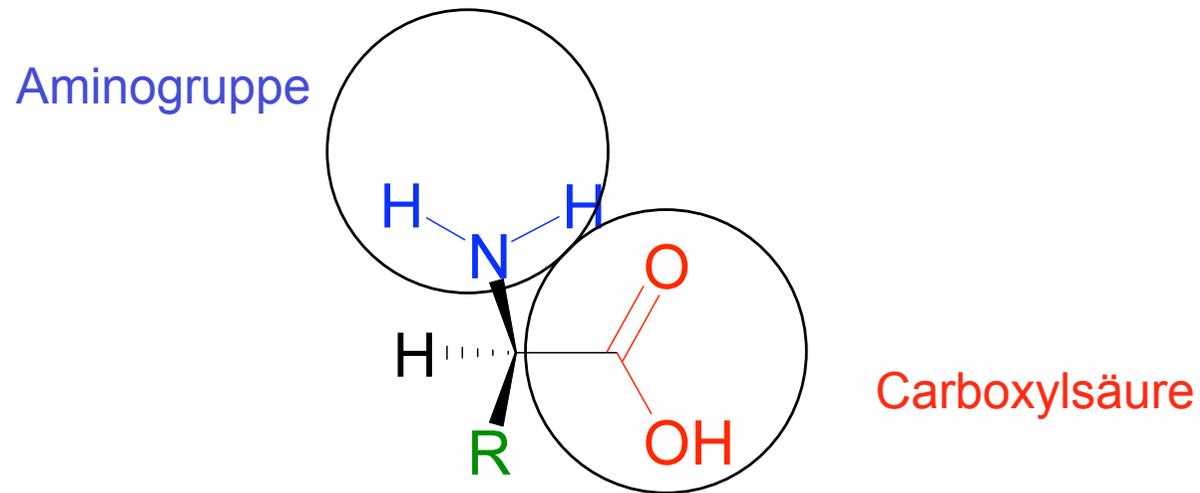
Aminosäuren unterscheiden sich in ihren physikochemischen Eigenschaften.



Q: müssen Bioinformatiker die Eigenschaften von Aminosäuren kennen?

Einleitung: Aminosäuren

Aminosäuren sind die **Bausteine** von Proteinen:



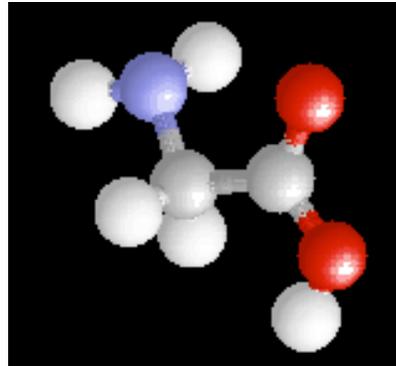
Aminosäuren unterscheiden sich hinsichtlich ihrer

- Größe
- elektrischen Ladung
- Polarität
- Form und Steifigkeit

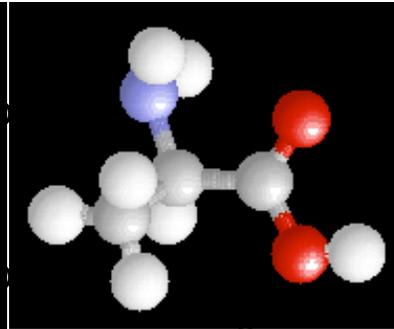
Einleitung: hydrophobe Aminosäuren

Proteine sind aus 20 verschiedenen natürlichen Aminosäuren aufgebaut

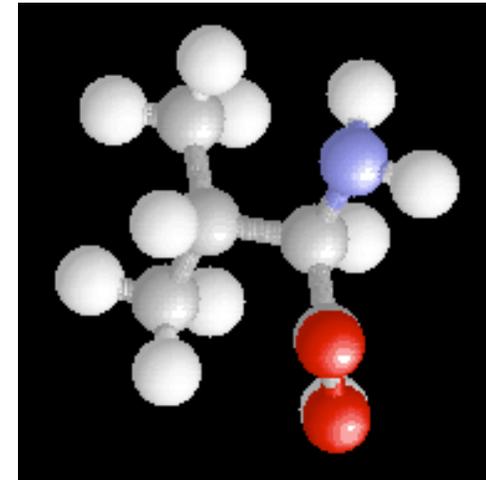
5 sind hydrophob.
Sie sind vor allem
Im Proteininneren.



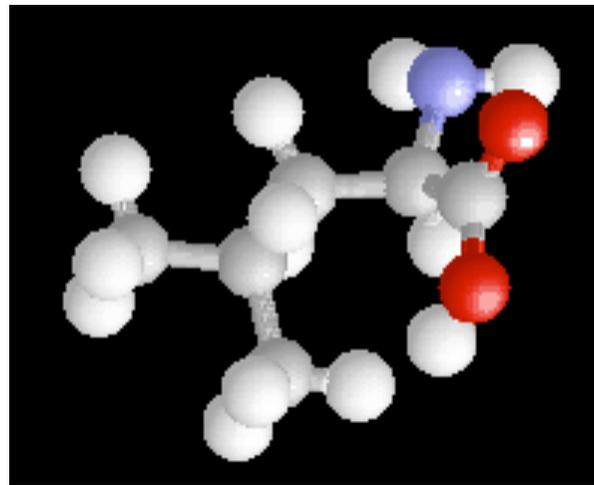
Glycine



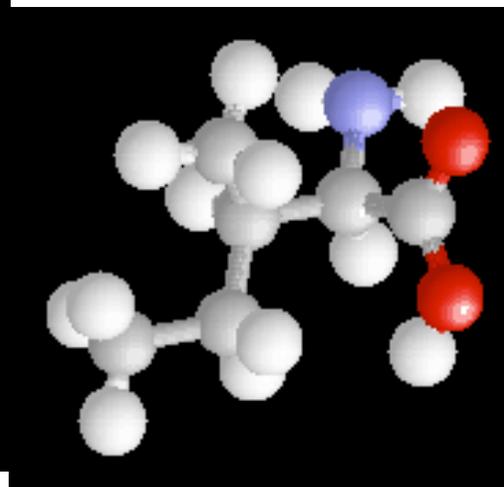
Alanine



Valine



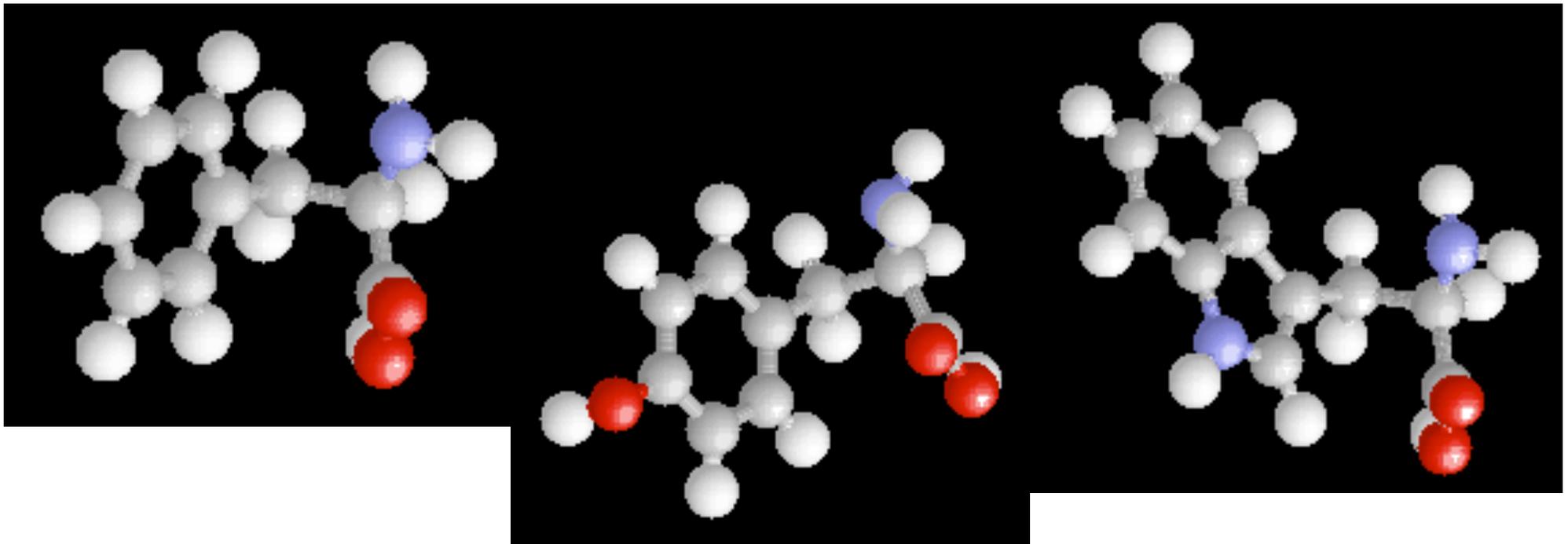
Leucine



Isoleucine

Einleitung: aromatische Aminosäuren

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



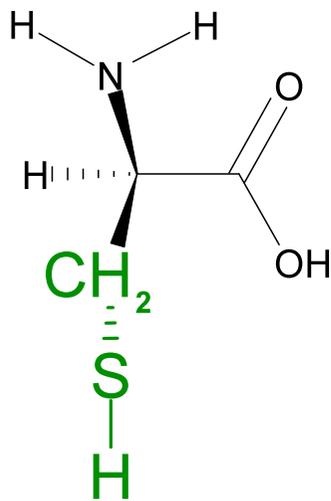
Phenylalanin

Tyrosin

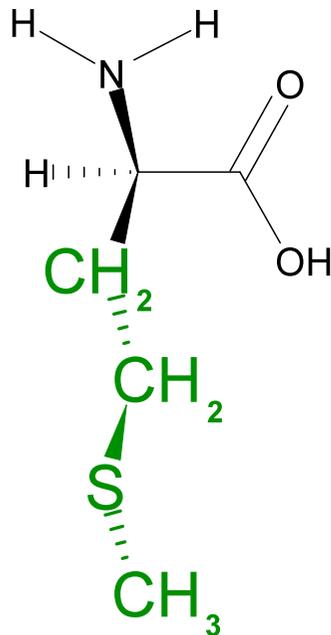
Tryptophan

Einleitung: Aminosäuren

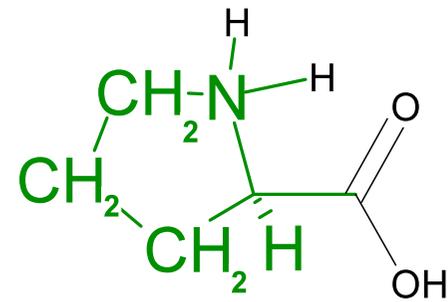
Es gibt 2 Schwefel enthaltende Aminosäuren und das ungewöhnliche Prolin.
Cysteine können Disulfidbrücken bilden.
Prolin ist ein "Helixbrecher".



Cystein



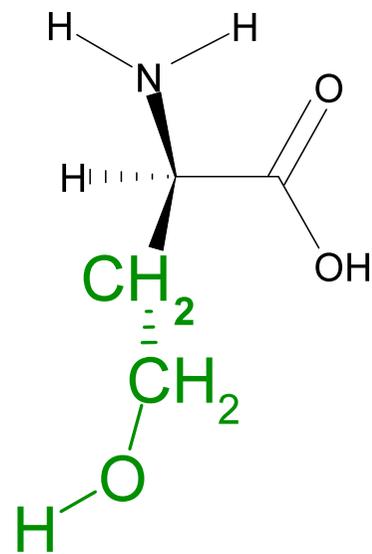
Methionin



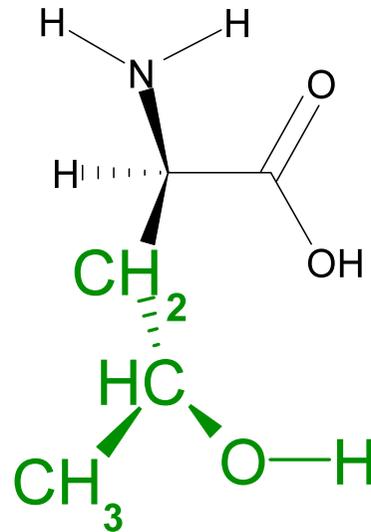
Prolin

Einleitung: Aminosäuren

Es gibt zwei Aminosäuren mit terminalen polaren Hydroxylgruppen:



Serin

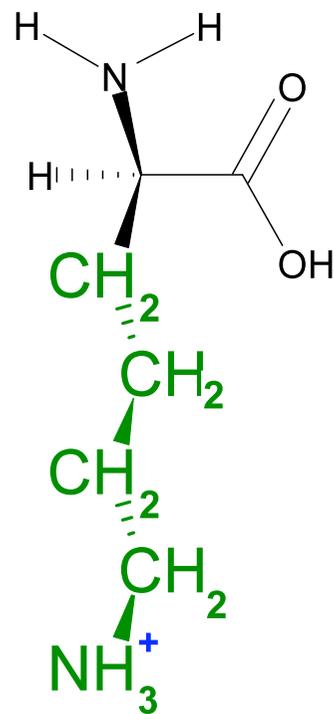


Threonin

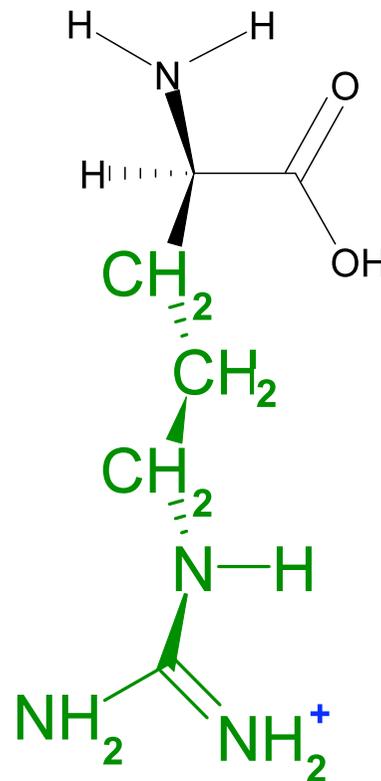
Einleitung: Aminosäuren

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

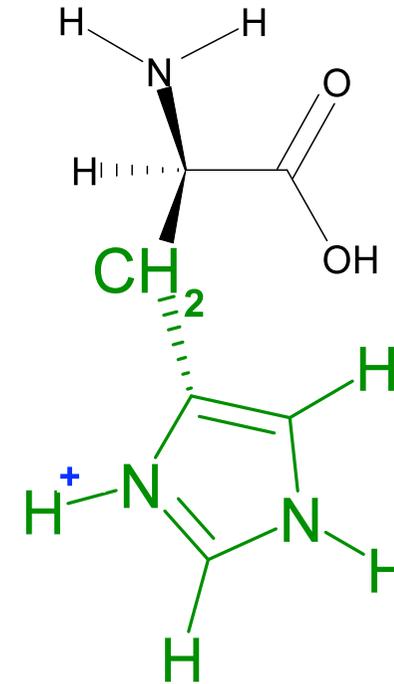
Thermophile Organismen besitzen besonders viele Ionenpaare auf den Proteinoberflächen.



Lysin



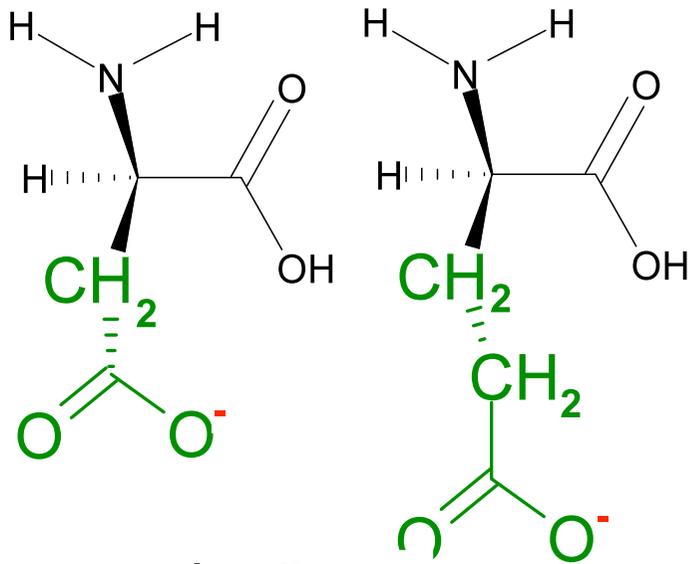
Arginin



Histidin

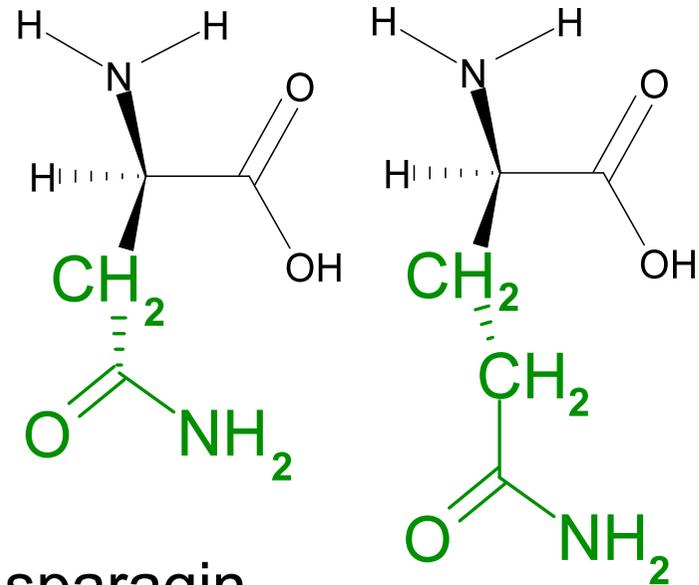
Einleitung: Aminosäuren

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

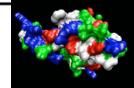
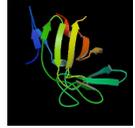
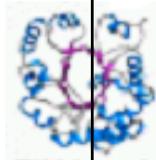
G Glycin	Gly	P Prolin	Pro
A Alanin	Ala	V Valin	Val
L Leucin	Leu	I Isoleucin	Ile
M Methionin	Met	C Cystein	Cys
F Phenylalanin	Phe	Y Tyrosin	Tyr
W Tryptophan	Trp	H Histidin	His
K Lysin	Lys	R Arginin	Arg
Q Glutamin	Gln	N Asparagin	Asn
E Glutaminsäure	Glu	D Asparaginsäure	Asp
S Serin	Ser	T Threonin	Thr

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

- ~174 Mio. **Nukleotidsequenzen**

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

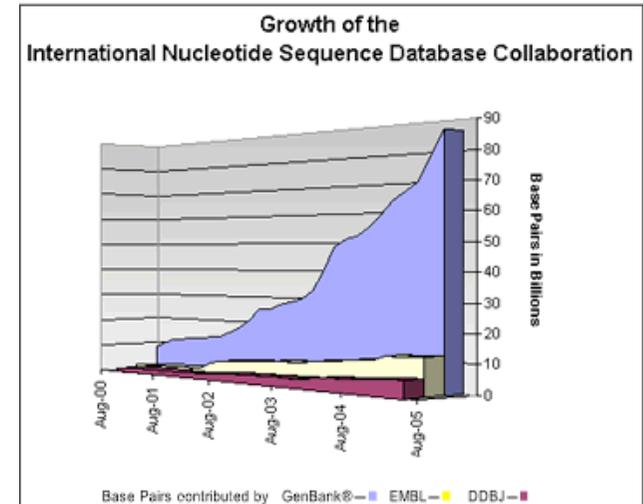
[GenBank/index.html](http://www.ncbi.nlm.nih.gov/GenBank/index.html))

~189 Mio. WGS-Nukleotidsequenzen

- 102.720 **3D-Strukturen** von biologischen

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

(Quelle: RCSB-PDB <http://www.rcsb.org>, 25.08.2014)



RCSB **PDB**
PROTEIN DATA BANK

Einträge sind teilweise **redundant**,

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

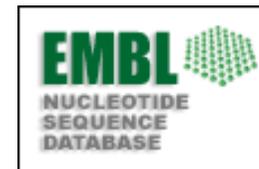
NCBI DNA-Datenbank

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



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

- öffentliche Nukleotid-Sequenzdatenbank
- ~174 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” 546.000 Einträge (Juli 2014)
- 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].

Item	Protein ID	Protein Name	Description
1	NP_822252	melibiase	[Streptomyces avermitilis MA-4680]
2	CAA69852	alpha-galactosidase; melibiase	[Thermoanaerobacter ethanolicus ATCC 33223]
3	NP_189269	glycosyl hydrolase family protein 27 / alpha-galactosidase family protein / melibiase family protein	[Arabidopsis thaliana]
4	AAA34770	pre-alpha galactosidase (melibiase)	
5	AAO78237	alpha-galactosidase (melibiase)	[Bacteroides thetaiotaomicron VPI-5482]
6	AAO77957	alpha-galactosidase (melibiase)	[Bacteroides thetaiotaomicron VPI-5482]
7	XP_001315888	Melibiase family protein	[Trichomonas vaginalis G3]

Stichwort, hier Name des Proteins

Detaillierte Suche bei Entrez

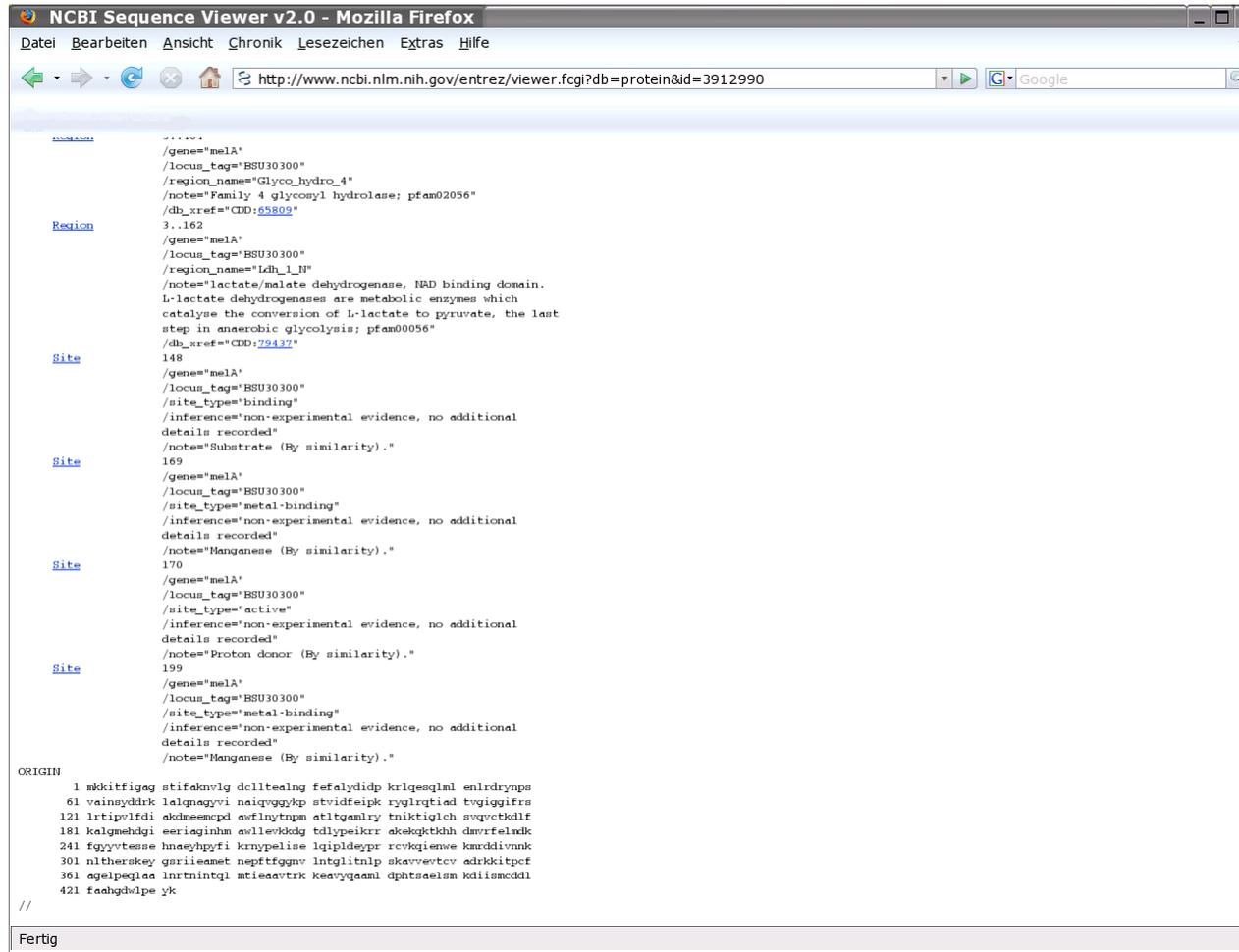
Protein Result - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://www.ncbi.nlm.nih.gov/sites/entrez
NCBI
All Databases PubMed Nucleotide Protein Genome structure PDB Taxonomy Books
Search Protein for melibiase "bacillus subtilis" [ORGN] Go Clear Save Search
Limits Preview/Index 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



```
.....  
/gene="mela"  
/locus_tag="BSU30300"  
/region_name="Glyco_hydro_4"  
/note="Family 4 glycoyl hydrolase; pfam02056"  
/db_xref="CDD:68809"  
Region  
3..162  
/gene="mela"  
/locus_tag="BSU30300"  
/region_name="ldh_I_N"  
/note="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; 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 mskitfigag stifaovlg dcltealng fefalydidp krqesqlal enldrnyps  
61 vainsyddrk lalqagyvi naigvgykp stvidfeipk ryqlrqtied tvqigqifrs  
121 lrtipvfdi ekdmeecpd avflnytnpm atltqanry tniktiglih svqvtkdlf  
181 kalgmehdgi eeriaginhn avllevdkgd tdlypeikrr akakqtkdhh dmrvfelndk  
241 fgyvtesse hnaeqhpyfi kmypelise lqipldeypr rcvkqiemwe kmrddivnkk  
301 nltherskey gsrileeant nepftfgnv lntglitnlp skavvetcv adrkkitpcf  
361 agelpeqlaa lnrntintql ntieaevtrk keavyqaanl dphtsaelm kdiiacddl  
421 fahdgdvlpe yk  
//  
Fertig
```

Fasta-Format

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

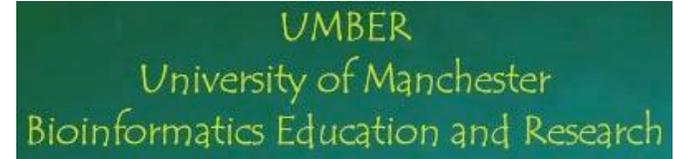
>DNA-Sequenz-Bezeichnung
ACGT

....

>Protein-Sequenz-Bezeichnung
ACDEFGHIKLMNPQRSTVWY

....

PRINTS



(<http://bioinf.man.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/searchprints.cgi?display_opts=Pi

Final Motifs

Motif	width	Element	Seqn Id	st	Int	Rpt
Motif 1	width=16	Element				
		SVVVAGGGSSTFTPGIV	GLVG_KCOLI	5	5	-
		SIVVAGGGSSTFTPGIV	GLVG_BACSU	7	7	-
		KITFIGAGSTIFVKHI	AGAL_KCOLI	6	6	-
		KITFIGAGSTIFAKIV	AGAL_BACSU	3	3	-
		KVVTIGGGSSTVTKLL	CRLE_KCOLI	6	6	-
		KIVTIGGGSSTVTKLV	CRLE_BACSU	6	6	-
		SILVAGGGSSTFTPGII	HALH_FUSHR	5	5	-
KIAYIGGGSQGWARS	LPLD_BACSU	11	11	-		
Motif 2	width=17	Element				
		ALSAADIVVVISLPGSL	LPLD_BACSU	76	49	-
		ALKDADEVVVAFIGGY	AGAL_KCOLI	75	53	-
		AFTDIDFVNAHIVGKY	HALH_FUSHR	74	53	-
		ALKDADEVVTTQLRVGQL	CRLE_KCOLI	77	55	-
		AFSDVDFVNAHIVGKY	GLVG_KCOLI	74	53	-
		AFTDIDFVNAHIVGKY	GLVG_BACSU	76	53	-
		ALKDADEVVTTQFRVGLL	CRLE_BACSU	77	55	-
		ALQAGYVINAIVGGY	AGAL_BACSU	72	53	-
		Motif 3	width=14	Element		
LDRQIPLKYGVVGG	GLVG_BACSU			97	4	-
LDRKIPLRHGVVGG	HALH_FUSHR			95	4	-
LDRKIPLRHGVVGG	GLVG_KCOLI			95	4	-
TDFRVCKRHGVLGQT	AGAL_KCOLI			97	5	-
LDRKIPLSHGVLGQ	CRLE_KCOLI			98	4	-
KDRKIPKYGVLGQ	CRLE_BACSU			98	4	-
IDFRKIPKYGVLGQT	AGAL_BACSU			94	5	-
VDVHLPERCGIYQS	LPLD_BACSU			97	4	-
Motif 4	width=21			Element		
		DTVGGGGIIRGLRAVPIFAKI	LPLD_BACSU	113	2	-
		ETCGGGIAYGHRISIGGVIGL	HALH_FUSHR	109	0	-
		ETCGGGIAYGHRISIGGVLEL	GLVG_KCOLI	109	0	-
		ETCGGGIAYGHRISIGGVLEI	GLVG_BACSU	111	0	-
		DTLGGGIRALETIPHLWQI	AGAL_KCOLI	113	2	-
		ETNGGGLEFGLLETIPVILEI	CRLE_BACSU	112	0	-
		DTVGGGGIIRGLRAVPIFAKI	AGAL_BACSU	110	2	-
		ETNGGGLEFGLLETIPVIFDI	CRLE_KCOLI	112	0	-
Motif 5	width=18	Element				
		PDAWHLNYSNPAAIVAEA	GLVG_BACSU	140	8	-
		PHAWHLNYSNPAAIVAEA	GLVG_KCOLI	138	8	-
		PDATHLNYVNPAAHTWA	AGAL_KCOLI	142	8	-
		PDAWHLNYSNPAAIVAEA	AGAL_BACSU	139	8	-
		PHAWHLNYSNPAAIVAEA	HALH_FUSHR	138	8	-
		PHAWHLNYSNPAAIVAEA	CRLE_KCOLI	141	8	-
		PHAWHLNYSNPAAIVAEA	CRLE_BACSU	141	8	-
		PHAWHLNYSNPAAIVAEA	LPLD_BACSU	142	8	-

Suchen: tetL

Abwärts Aufwärts Hervorheben Groß-/Kleinschreibung Das Seitenende wurde erreicht, Suche vom Seitenanfang fortg

Fertig

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 a 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 a cluster with 100%, 90%, and 50% identity. The entry is annotated with various databases, including InterPro, PANTHER, Pfam, PRINTS, and PROSITE. 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

Pfam



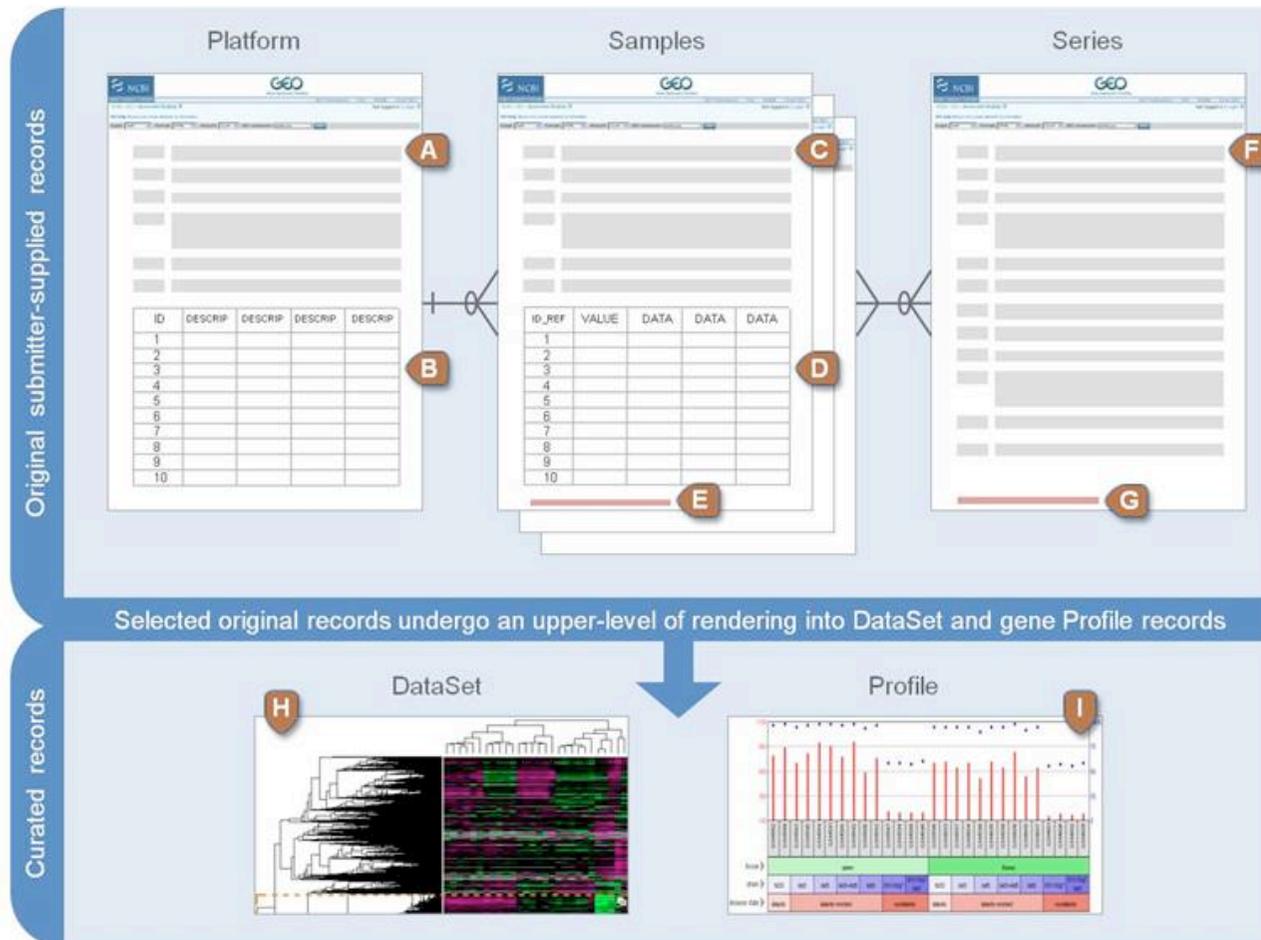
(<http://pfam.sanger.ac.uk/>)

- 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-A: genau untersuchte Profile aus multiplen Alignments, teilweise manuelle Alignments, >8000 Familien
- Pfam-B: automatisch generierte Profile: mehr Sequenzen, aber weniger präzise

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



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.

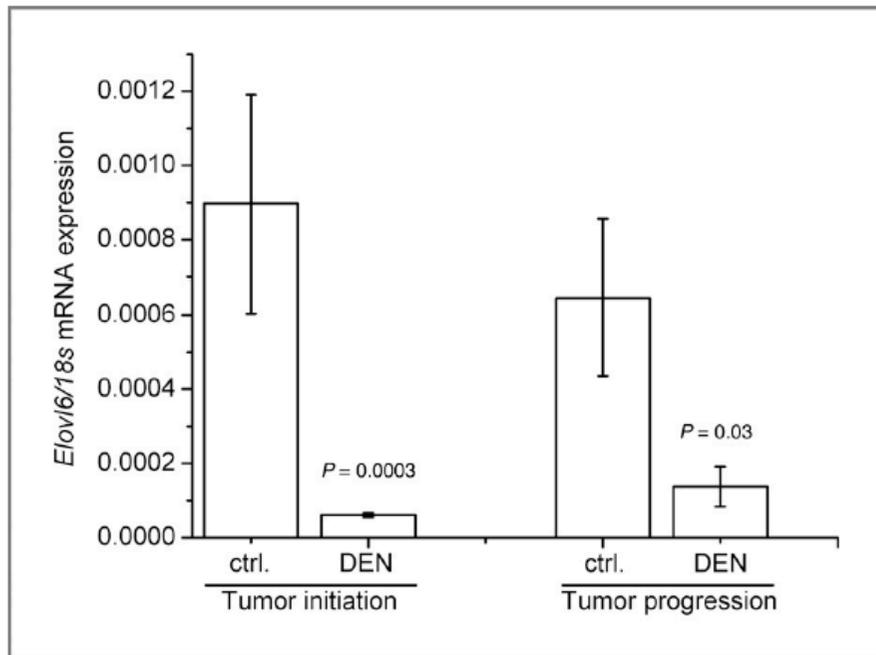


Figure 2. Wild-type mice were treated with the carcinogen DEN at the age of 2 weeks. Livers were analyzed after 24 weeks to assess the tumor initiation state. Analyses in the tumor progression stadium were done after 36 weeks. *Elovl6* mRNA expression as determined by real-time reverse transcriptase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

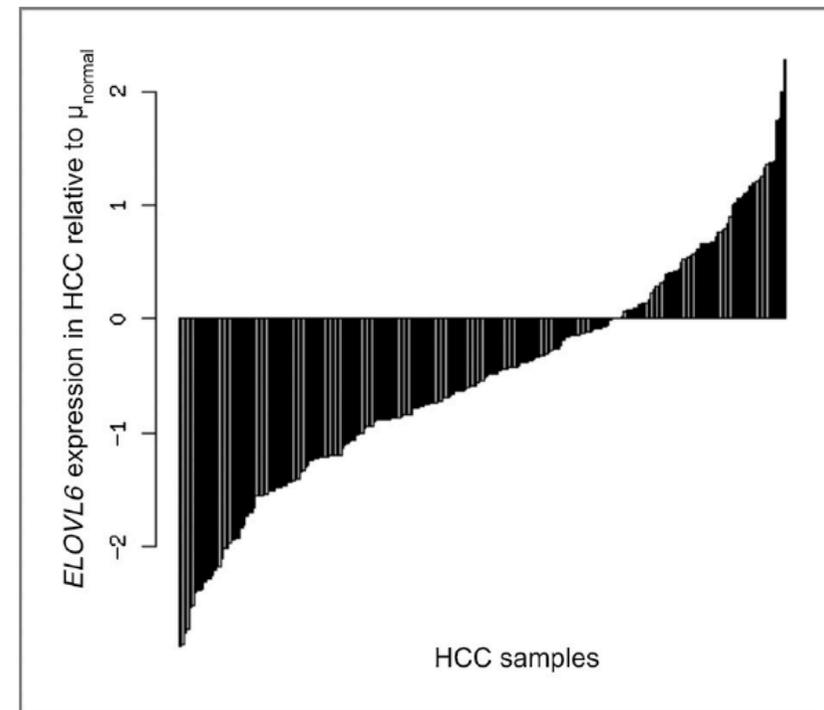


Figure 1. mRNA levels of *ELOVL6* in 247 human HCC samples relative to the mean of 239 nontumor liver tissue (μ_{normal}). Samples of dataset GSE14520 [\log_2 (expression) values from GEO after Robust Multi-array Average normalization] were mapped to hgu133a.db using bioconductor. Significance values: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

TCGA – The Cancer Genome Atlas

*Human Molecular Genetics, 2013
doi:10.1093/hmg/ddt158*

DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples

Ruslan Akulenko and Volkhard Helms*

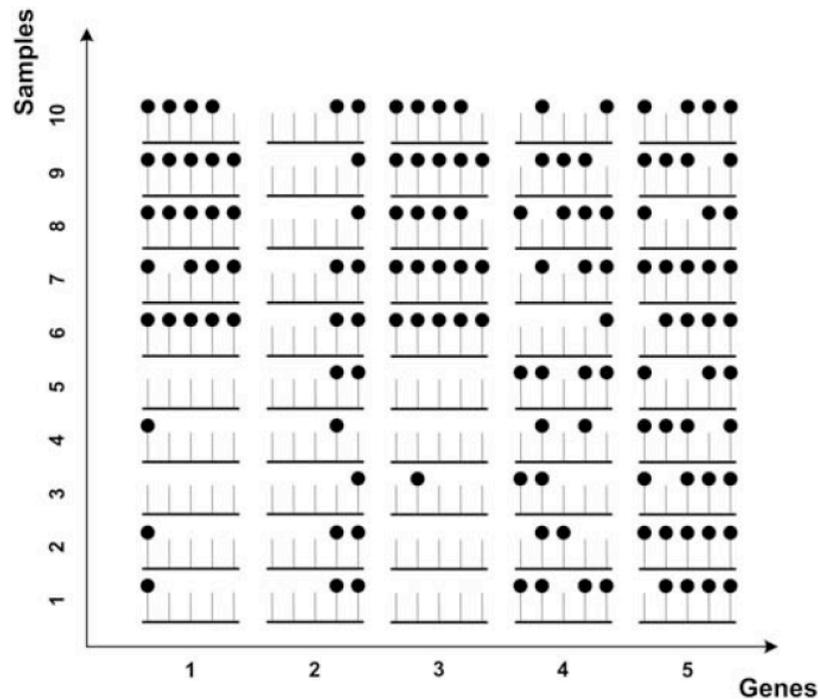


Figure 1. Schematic example of CpG methylation in five genes. The sticks indicate CpG sites. Filled circles indicate CpG methylation. The first and the third genes show highly correlated methylation levels across the 10 samples. Here, we term this behavior ‘co-methylation’. The second gene is mostly unmethylated. Even though genes five and four are mostly methylated, they are not co-methylated.

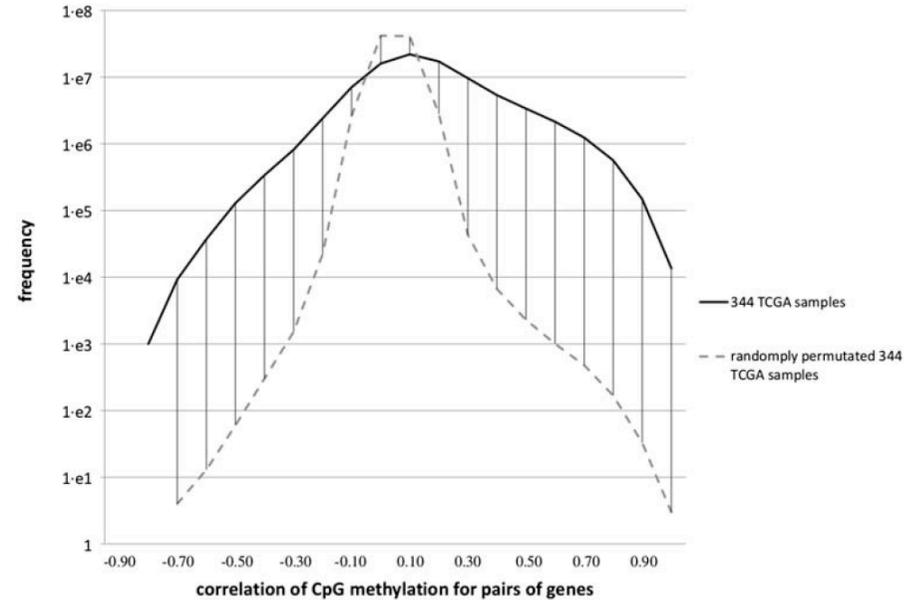


Table 1. The 10 strongest correlations for pairs of genes with respect to their β -values, obtained after three-stage filtering

First gene	Second gene	Pearson's correlation
SPRR1B	SPRR1A	0.872
FCN2	FCN1	0.870
CD244	CD48	0.866
SPRR1B	SPRR4	0.862
TAS2R13	PRB4	0.859
F7	TFF1	0.856
SH3TC2	SPARCL1	0.853
ABCE1	SC4MOL	0.849
REG1B	REG1P	0.846
SPRR3	SPRR4	0.843

TCGA – The Cancer Genome Atlas

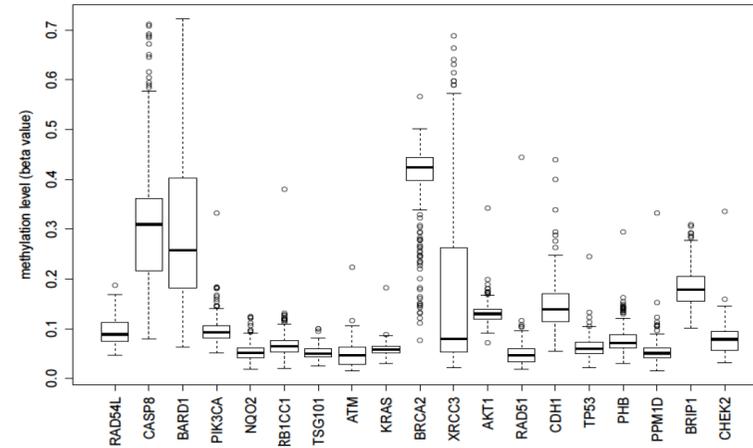
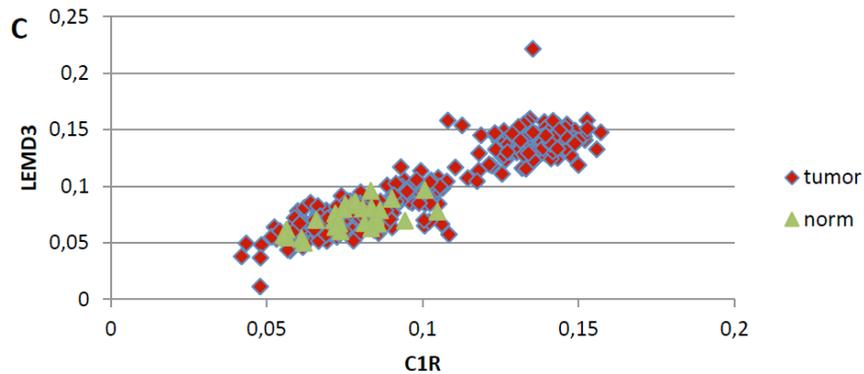
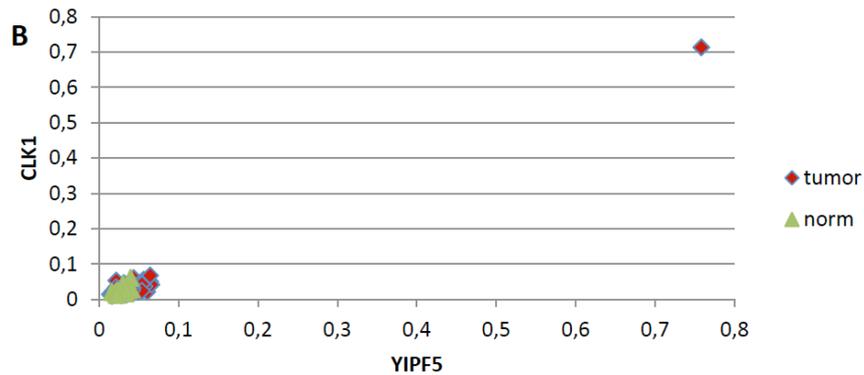


Fig S2. Methylation levels of 19 OMIM breast cancer genes found among the 344 TCGA data samples analyzed here.

Oben – bekannte Brustkrebsgene sind wenig methyliert.

Unten – statistisch signifikant angereicherte biochemische Pfade, die ko-methylierte Gene enthalten

Daten müssen bereinigt/gefiltert werden:
Oben – hohe Korrelation durch einen (zufälligen oder fehlerhaften) Ausreißer.

Unten – hohe Korrelation, aber insgesamt geringe Variabilität der Werte

Cluster ID	KEGG pathways	p-value	Genes involved in pathways	FDR
8	hsa04950:Maturity onset diabetes of the young	0.003	HNF1B, FOXA2, NEUROD1	2.622
9	hsa04640:Hematopoietic cell lineage	0.009	CD1A, CD1E, CD1D	6.229
15	hsa04730:Long-term depression	0.004	GRM5, C7ORF16, PRKG2	2.952
27	hsa04512:ECM-receptor interaction	0.005	COL5A2, COL11A1, SPP1	3.500

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.