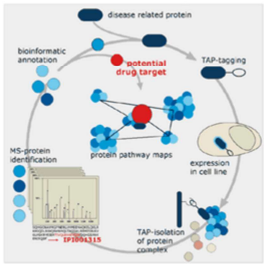
Softwarewerkzeuge der Bioinformatik

Inhalt dieser Veranstaltung: Softwarewerkzeuge kennenlernen für

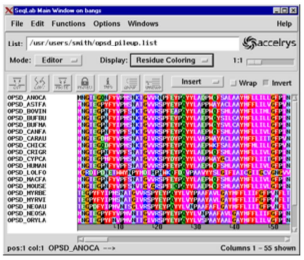
I Sequenzanalyse

II Analyse von Proteinstruktur und Ligandenbindung

III Analyse von Omics-Daten, Zell- bzw. Netzwerksimulationen



www.cellzome.com



www.accelrys.com

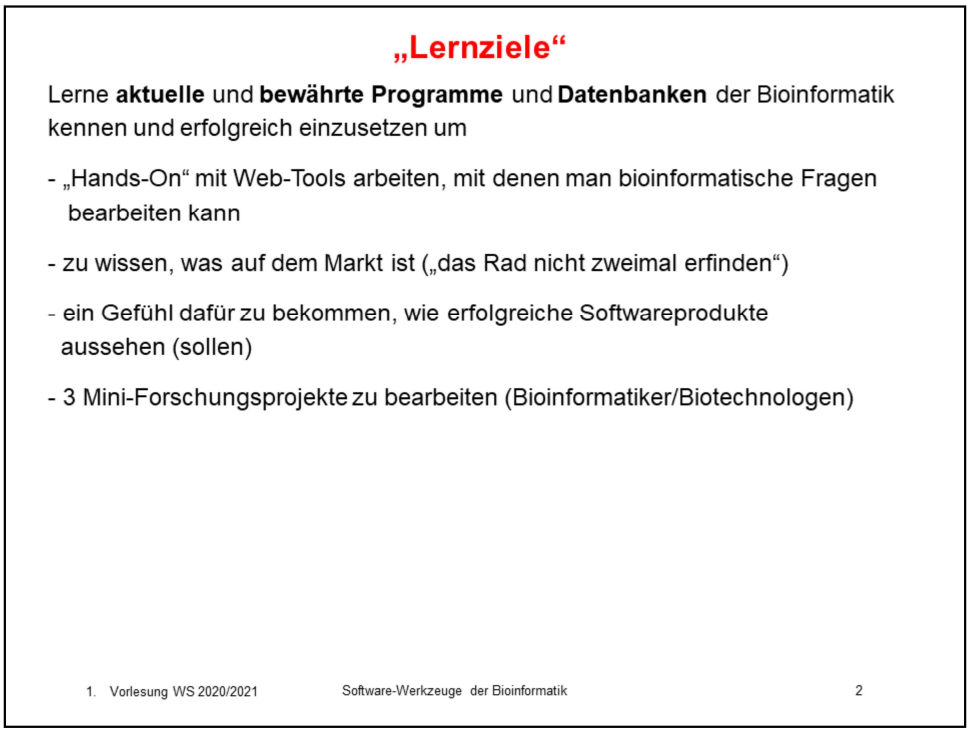
1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 1

Willkommen zu der Vorlesung „Softwarewerkzeuge der Bioinformatik“. Diese Vorlesung bietet eine **Einführung** in die Verwendung von Bioinformatik-Tools in 3 verschiedenen Bereichen.

Der erste Bereich ist das Gebiet der **Sequenzanalyse**. Dies ist das klassische Gebiet der sogenannten algorithmischen Bioinformatik. Die Sequenzen von Genen und Proteinen waren die ersten Datensätze in der Genetik und Molekularbiologie, die mit Bioinformatik-Methoden analysiert werden konnten. Obwohl dieses Gebiet daher fast schon ein altmodischen „Touch“ hat, ist die Sequenz-Bioinformatik heute wieder sehr „angesagt“, seit man durch die Entwicklung von modernen Next Generation Sequencing-Methoden routinemäßig ganze Genome in großer Anzahl sequenzieren kann. Einen weiteren Aufschwung nahm das Gebiet durch den modernen Trend, einzelne Zellen eines Organismus zu sequenzieren.

Der zweite Bereich ist das Gebiet der **Proteinstrukturanalyse** sowie der Bindung von kleinen und großen Liganden. Dieses Gebiet weist eine wichtige Schnittstelle zur medizinischen Chemie auf und ist insbesondere für die Wirkstoffentwicklung von großem Interesse.

Der dritte Bereich ist das große und sehr heterogene Gebiet der **Analyse von Hochdurchsatzdaten** wie Transkriptom und Metabolom.



Dies sind die **Lernziele** dieser Veranstaltung. Als Vorkenntnisse wird lediglich vorausgesetzt, dass die Teilnehmenden schon einmal etwas von DNA, Proteinen und Zellen gehört haben. Im Bereich Bioinformatik werden keine Vorkenntnissen vorausgesetzt. Im Gegensatz zu anderen einführenden Veranstaltungen wie Bioinformatik 1 und 2, die intensiv (mit einem Informatik-geschulten Blick) auf die algorithmischen Konzepte in den einzelnen Bereichen der Bioinformatik eingehen, werden wir nicht tief in die Theorie eindringen. Dennoch wird erwartet, dass die AbsolventInnen dieser Veranstaltung im Anschluss erklären können, wie z.B. der BLAST-Algorithmus, der Sankoff-Algorithmus, der Neighbor-Joining-Algorithmus, Homologie-Modellierung, differentielle Genexpression etc. funktionieren. Also zumindest in der Abschlussklausur kommen diese Themen dran.

Ein weiteres wichtiges Lernziel ist, dass die Teilnehmenden den **Umgang** mit modernen, populären Bioinformatik-Tools erlernen und üben. Die meisten dieser Tools sind webbasiert und können auch von zuhause aus ausprobiert werden. In diesem Corona-Semester werden wir tatsächlich nur mit Web-Tools arbeiten. Mit einigen dieser Tools werden Sie dann in Kleingruppen kleine Forschungsprojekte bearbeiten. Darauf kommen wir in ein paar Folien zurück.

Organisatorisches

Jede Woche Vorlesung	Donnerstag 10.15 – 12.00 Uhr (evtl. 15 minütige Pause)
online via MS Teams	Dozent: Prof. Helms

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

Übungen „**hands-on**“ Beginn **heute** am 5.11:

Donnerstag, 14:00 Uhr – 16:00 Uhr, online via Teams/Moodle.

Verantwortliche Betreuer der Übungen

Sequenz-Analyse	Markus Hollander
Proteinstruktur	Dr. Michael Hutter
Zellsimulationen	Andreas Denger

1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 3

Die Veranstaltung besteht aus einer Vorlesung und einem Tutorial bzw. Übung.

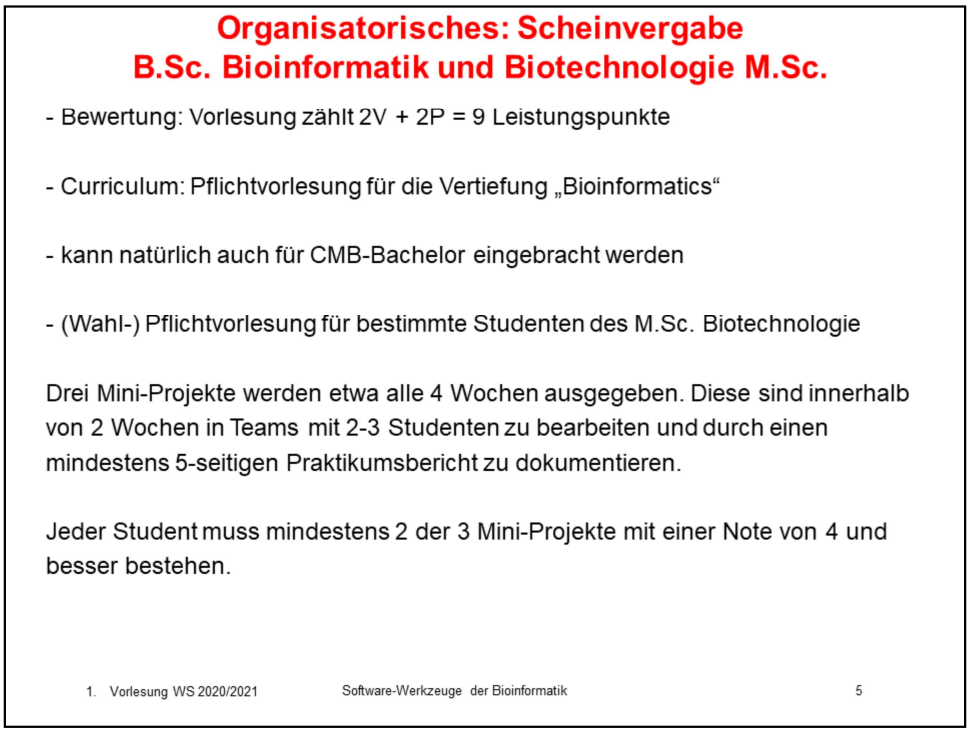
Die **Vorlesung** wird online mittels der Software MS Teams angeboten.

Es wird vorausgesetzt, dass sich die Teilnehmenden im Vorfeld jeder Vorlesung mit den Vorlesungsfolien vertraut gemacht haben.

Die Vorlesung dient der Vertiefung des Stoffs und dem Eingehen auf Fragen der Studierenden.

Das **Tutorial** wird ebenfalls online angeboten. Die **Teilnahme** am Tutorial ist **verpflichtend**.

Die Teilnehmenden werden in Kleingruppen aufgeteilt, die jeweils von einem/r Betreuer/in betreut werden.

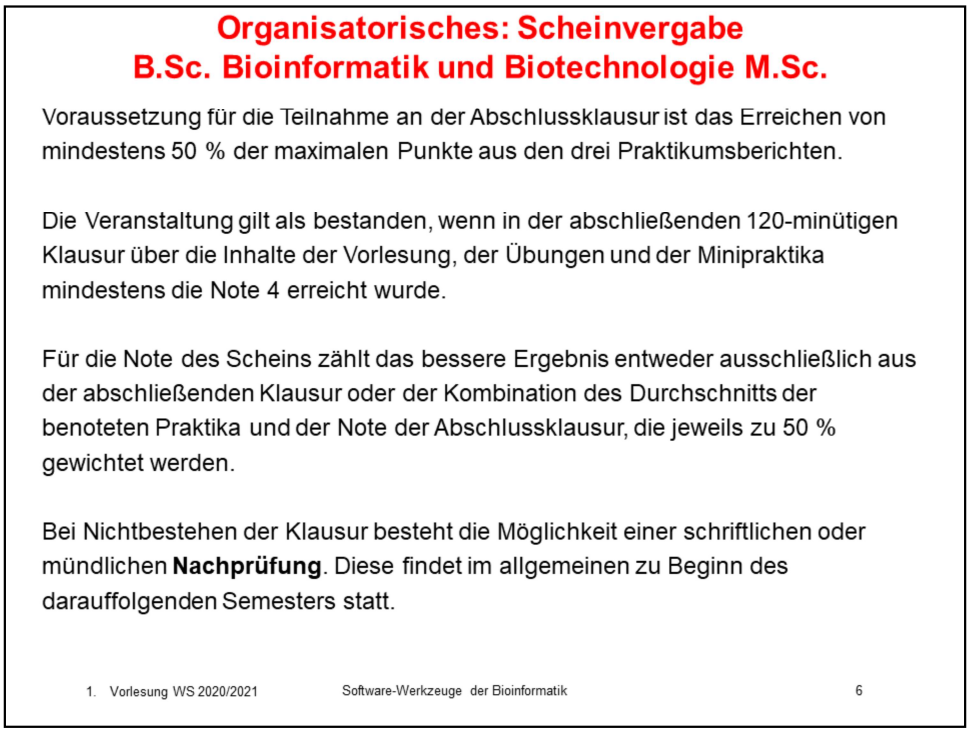


Meist kommt die Mehrzahl an Teilnehmenden aus dem **Masterstudiengang Biotechnologie**, typischerweise im 1. Semester. Diese Studierenden haben natürlich einen guten biologischen Hintergrund und manchmal bereits ein wenig Kontakt mit Bioinformatik-Tools gehabt (z.B. für Sequenzalignment oder Genexpressionsanalyse). Der andere Teil der Teilnehmenden kommt aus dem **Bachelorstudiengang Bioinformatik**, meist im 5. Semester. Diese Studierenden haben meist schon Bioinformatik 1 und 2 gehört und kennen sich daher mit dem Stoff der ersten 7 Vorlesungswochen (Sequenzalignments, Proteinstruktur) gut aus.

Ein wichtiges Element dieser Veranstaltung sind die **3 Mini-Forschungsprojekte**. Diese werden in Gruppen von 2-3 Studierenden bearbeitet, die sich selbständig zusammenfinden (also nicht vom Dozenten oder von den Übungsleitern zusammengestellt werden). Wie in jedem anderen Praktikum ist über den Hintergrund der Projekte, die darin verwendeten Methoden, die Ergebnisse der Aufgabe und ggf. eine kurze Diskussion ein **Bericht** zu verfassen. Als formale Mindestbedingung gilt eine Mindestlänge von 5 Seiten.

Die Mini-Forschungsprojekte werden benotet. Alle Mitglieder einer Gruppe erhalten dieselbe Note. Jede/r Studierende muss mindestens für 2 der 3 Mini-Projekte mindestens die Note „ausreichend“ erhalten. Mini-Forschungsprojekte können nicht wiederholt werden. Die **Bearbeitungszeit**

wird bei Ausgabe der Projekte angegeben (mindestens 2 Wochen) und ist **verpflichtend einzuhalten**. Verspätete Abgaben werden mit „mangelhaft“ bewertet. Falls 1 oder 2 Mitglieder einer Gruppe aus von ihnen nicht zu vertretenden Gründen (z.B. Krankheit) das Projekt teilweise nicht bearbeiten können, müssen eben die übrigen Gruppenmitglieder das Projekt bearbeiten und rechtzeitig fertigstellen.



Die Teilnahme an der Abschlussklausur ist nur möglich, wenn mindestens 50% der maximalen Punkte in den Mini-Forschungsprojekten erreicht wurden und höchstens ein Projekt mit „mangelhaft“ bewertet wurde.

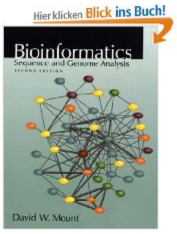
Die **Abschlussklausur** findet voraussichtlich als Präsenzveranstaltung statt. Falls dies aufgrund der Entwicklung der Corona-Epidemie nicht möglich sein sollte, finden stattdessen mündliche Prüfungen über den Klausurstoff statt.

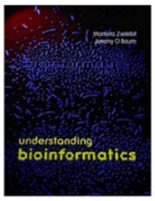
Die Note auf dem **Schein** ist das 50%:50% Mittel aus Klausurnote und der gemittelten und (ab)gerundeten Projektnote. Falls die Klausurnote besser als die gemittelte Projektnote ist, erhält der/die Studierende auf dem Schein die Klausurnote. Dadurch gleichen wir das „Pech“ aus, falls ein/e Studierende/r 1-2 schwache oder arbeitsunwillige ProjektpartnerInnen hatte und deswegen schwache Noten in den Projekten erhielt.

Für alle TeilnehmerInnen der Klausur, deren Leistungen mit „mangelhaft“ bewertet wurden, oder die ihre Note verbessern möchten, wird zu Beginn des SS 2021 eine **Nachprüfung** angeboten. Die bessere Note zählt dann, man kann sich also nicht verschlechtern. Es gibt immer mal wieder die Frage, ob man in einem späteren Semester den Kurs zur Notenverbesserung wiederholen kann. Dies ist grundsätzlich möglich. Allerdings gehören alle Leistungen eines Semester zusammen. In solch einem Fall muss der/die Studierende deshalb noch einmal drei neue Mini-Forschungsprojekte bearbeiten. Dadurch ist der Aufwand relativ hoch.

Literatur

David Mount
Bioinformatics
72€





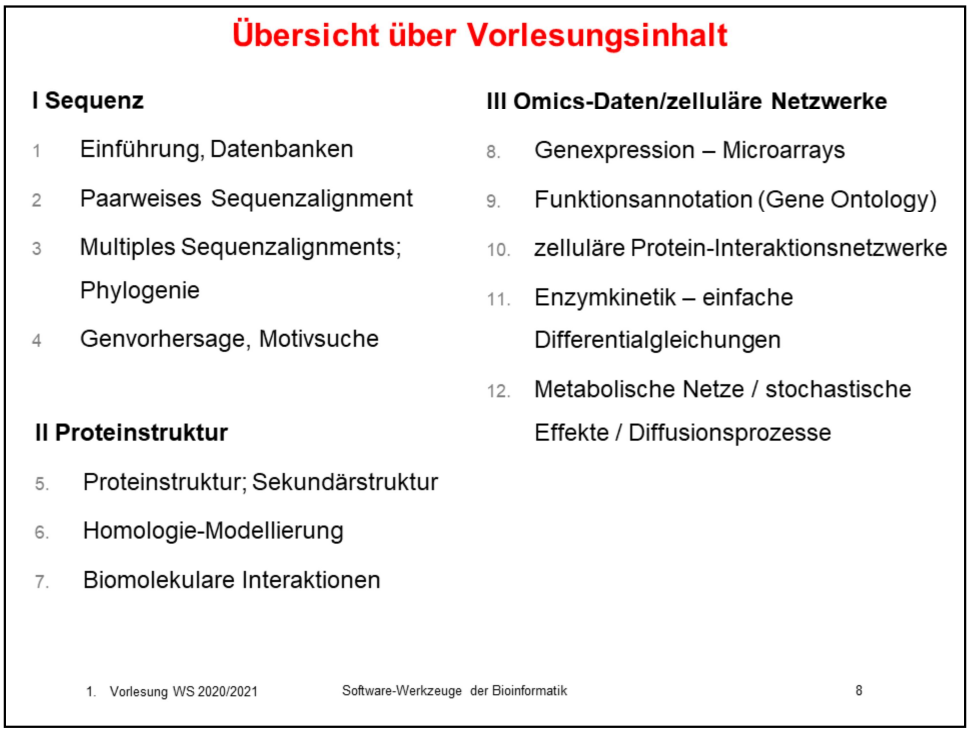
Marketa Zvelebil & Jeremil O. Baum
Understanding bioinformatics, 63€

Vorlesungsfolien auf
https://www-cbi.cs.uni-saarland.de/teaching/ws-2020-21/softwarewerkzeuge-der-bioinformatik-ws-20-21/

1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 7

Dies sind zwei einführende **Lehrbücher** in die Bioinformatik. Sie sind in der Campus-Bibliothek Informatik vorhanden.

Die kommentierten **Vorlesungsfolien** werden auf unserer Lehrstuhlseite zur Verfügung gestellt.



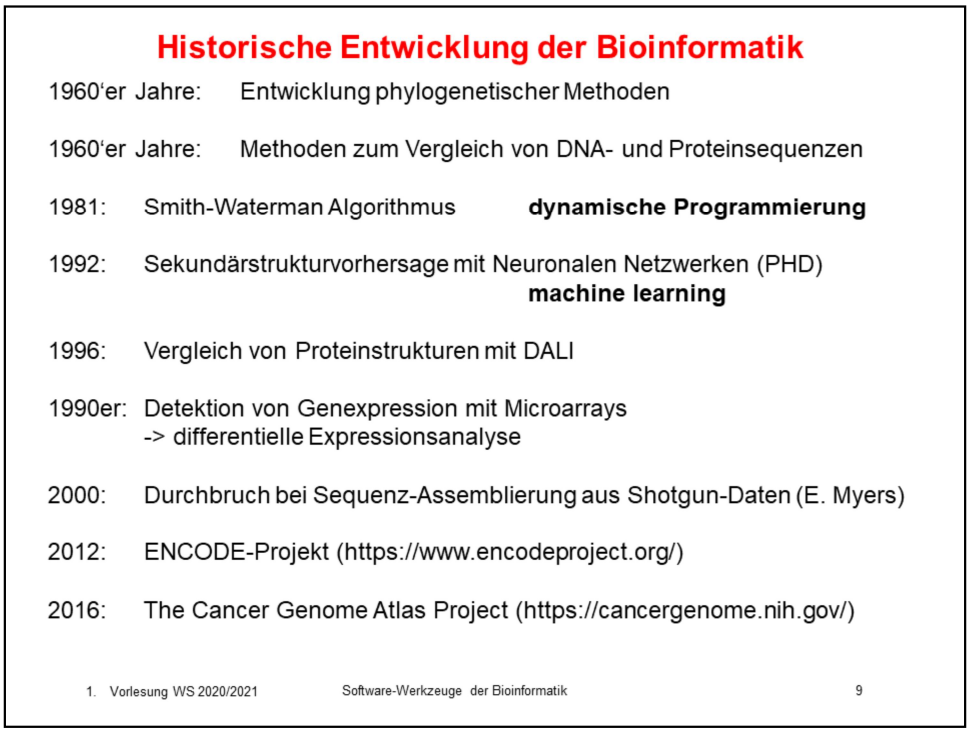
Dies ist eine Übersicht über den **Inhalt** der 12 Vorlesungswochen im WS 2020/21.

Die Vorlesung besteht aus 3 Blöcken: Sequenzanalyse, Proteinstrukturanalyse und die Analyse von Daten aus Hochdurchsatzexperimenten.

Die ersten beiden Gebiete sind klassische Gebiete der Bioinformatik, in denen seit den 1960er-Jahren geforscht wird.

Das letzte Gebiet ist erst durch die technische Entwicklungen seit den 1990er-Jahren entstanden. Hier spielen jedoch auch klassische Konzepte aus der Biochemie (Enzymkinetik) und der Physik (Diffusion, Stochastik) hinein.

Obwohl die Vorlesung einen einführenden Charakter hat und methodisch nicht besonders schwierig ist, wird von den Studierenden meist als Herausforderung angesehen, dass die Vorlesung solch einen großen thematischen Bogen spannt. Jedes einzelne Thema kommt nur in einer einzigen Vorlesung dran.



Dies sind ein paar Notizen zur **Geschichte** des Faches Bioinformatik.

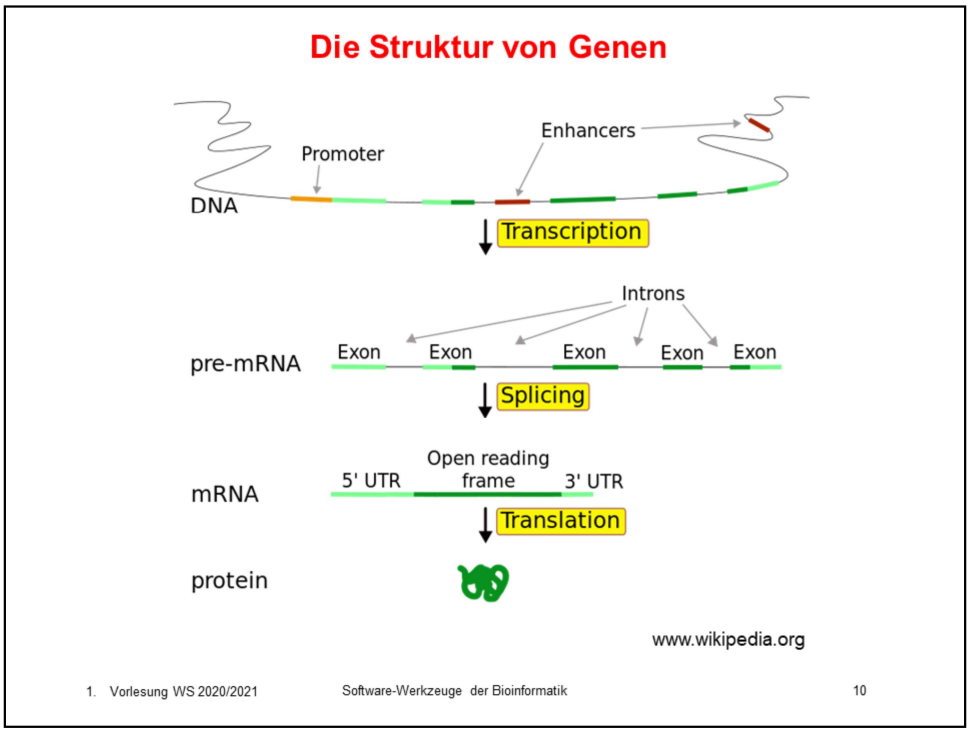
Auf der Webseite der „Fachgruppe Bioinformatik“ gibt es weitere Hinweise: https://www.bioinformatik.de/de/bioinformatik-3/was-ist-bioinformatik.html

Die englische Version von Wikipedia enthält deutlich mehr Infos als die deutsche Seite: https://en.wikipedia.org/wiki/Bioinformatics

Mehr zur Historie des Faches ausserdem: S. 6-8 in https://chagall.med.cornell.edu/BioinfoCourse/presentations2010/Lecture1_2010.pdf

In einer späteren Version dieser Vorlesung https://chagall.med.cornell.edu/BioinfoCourse/presentations2013/Lecture1_Intro_2013_slides.pdf fehlt die historische Zeitskala.

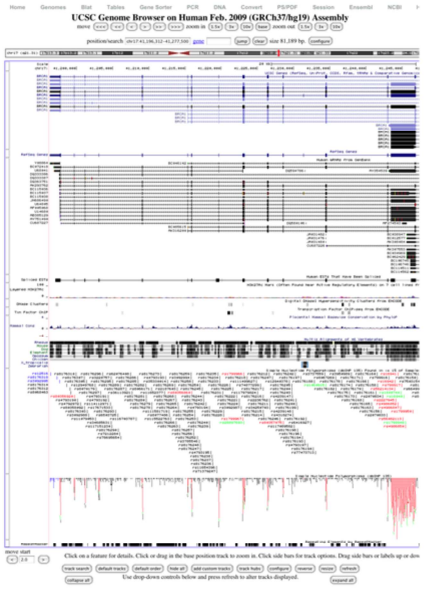
Eigentlich beschäftigt sich in der Bioinformatik aber kaum jemand mit der Vergangenheit des Feldes, sondern viel mehr mit der Zukunft.



Diese **grundlegenden Dinge** über DNA, Transkription, das Spleissen von mRNA, und die Translation in Aminosäureketten sollten Ihnen allen vertraut sein.

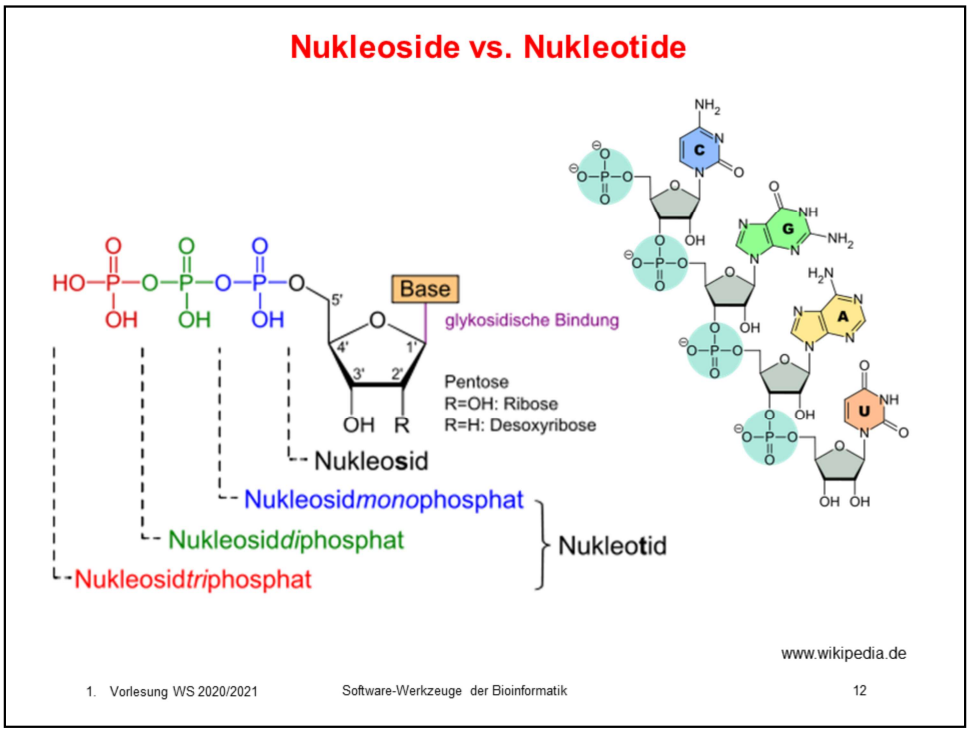
In Eukaryoten wird die Transkription von protein-kodierenden Abschnitten auf der DNA sowie von nicht-kodierenden Abschnitten (diese können z.B. für microRNAs kodieren) durch Bindung von regulatorischen Faktoren an die Promoter- und Enhancerabschnitte auf der DNA kontrolliert.

UCSC Genome Browser



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

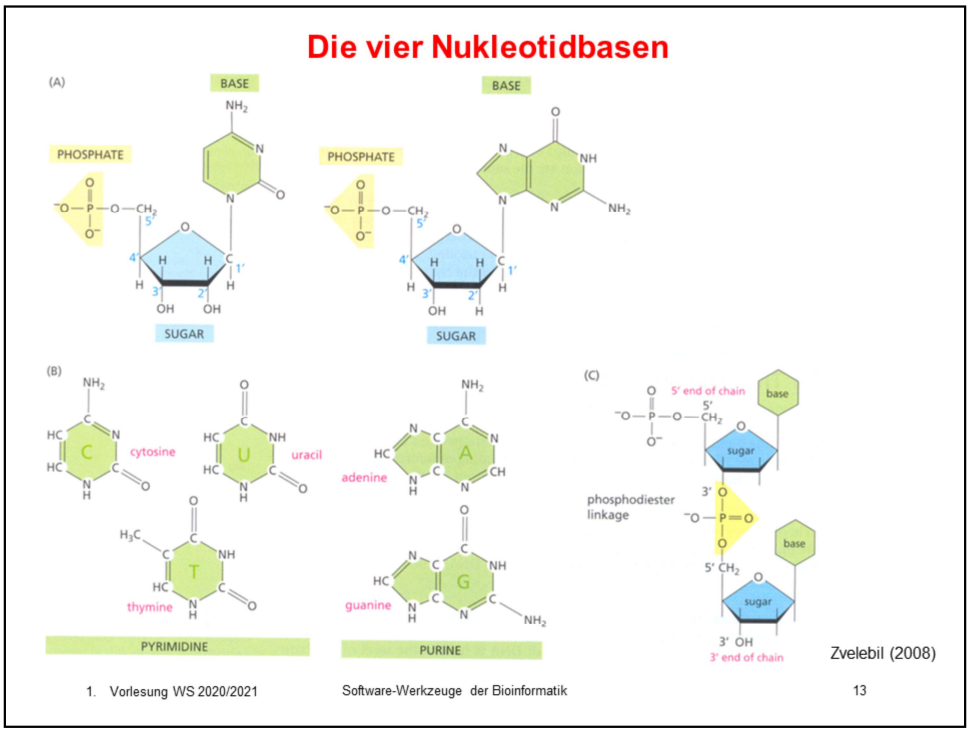
Ein Standardtool, um sich genomische Sequenzen anzuschauen, sowie um Zusatzinformationen darzustellen, ist der UCSC Genome Browser.



(links) Nukleotide sind Nukleoside mit Phosphatgruppen.

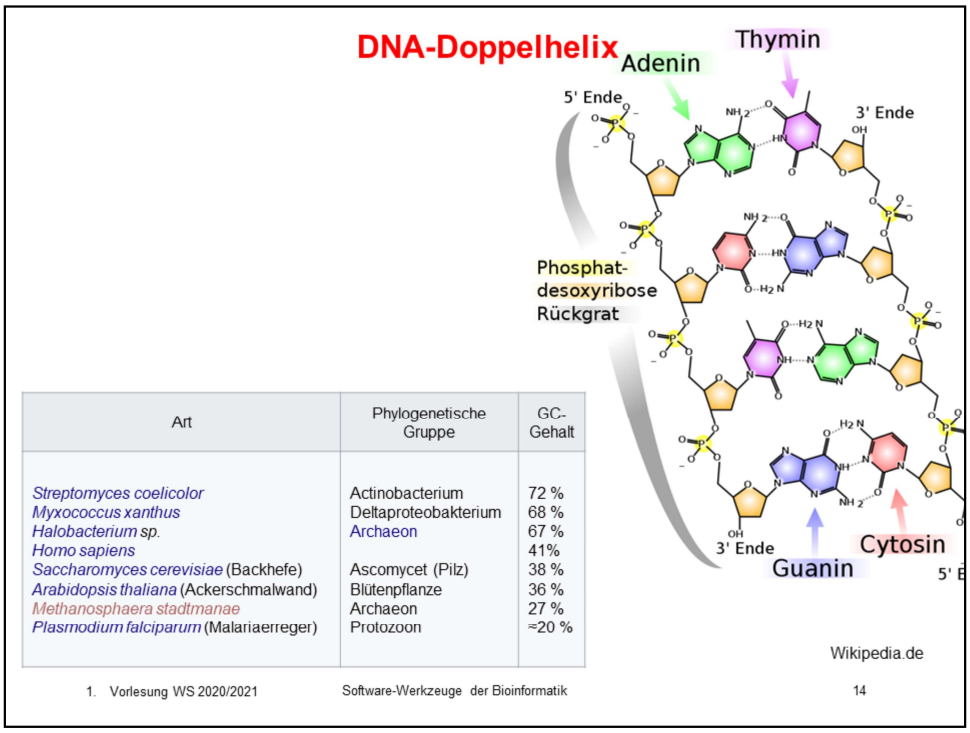
Ein **Nukleosid** ist die Verbindung einer Nukleinbase (Base) mit einem Einfachzucker, einer Pentose. Deren 2'-Rest (R) ist im Falle der Ribose eine Hydroxygruppe (OH-), im Falle der Desoxyribose hingegen Wasserstoff (H-). Bei einem Nukleotid ist die 5'-OH-Gruppe der Pentose eines Nukleosids mit einem Phosphatrest verestert. Ein Nukleosidtriphosphat (NTP) weist drei Phosphatgruppen auf, die untereinander Säureanhydridbindungen ausbilden. Mit Adenin als Base und Ribose als Saccharid liegt das **Adenosintriphosphat** (ATP) vor.

(rechts) Das Makromolekül einer DNA oder einer RNA ist jeweils aus vier verschiedenen Sorten von Nukleotiden zusammengesetzt, die durch kovalente Bindungen zum Strang des polymeren Biomoleküls verknüpft werden, einem Polynukleotid. Die hierbei ablaufende Reaktion ist eine Kondensationsreaktion. Von den monomeren Nukleosidtriphosphaten wird dabei ein Pyrophosphatrest abgespalten, sodass die Monosaccharide der Nukleotide über je eine Phosphatgruppe aneinander gekoppelt werden, die das C5'-Atom der nächsten mit dem C3'-Atom der vorangehenden Pentose verbindet.



Dies sind die 4 Nukleotidbasen Adenin, Cytosin, Guanin und Thymin, aus denen DNA-Stränge aufgebaut sind.

Die **Pyrimidine** C und T enthalten einen einfachen aromatischen Ring, die **Purine** A und G einen aromatischen Doppelring.



Art	Phylogenetische Gruppe	GC-Gehalt
Streptomyces coelicolor	Actinobacterium	72 %
Myxococcus xanthus	Deltaproteobakterium	68 %
Halobacterium sp.	Archaeon	67 %
Homo sapiens		41%
Saccharomyces cerevisiae (Backhefe)	Ascomycet (Pilz)	38 %
Arabidopsis thaliana (Ackerschmalwand)	Blütenpflanze	36 %
Methanosphaera stadtmanae	Archaeon	27 %
Plasmodium falciparum (Malariaerreger)	Protozoon	≈20 %

1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 14

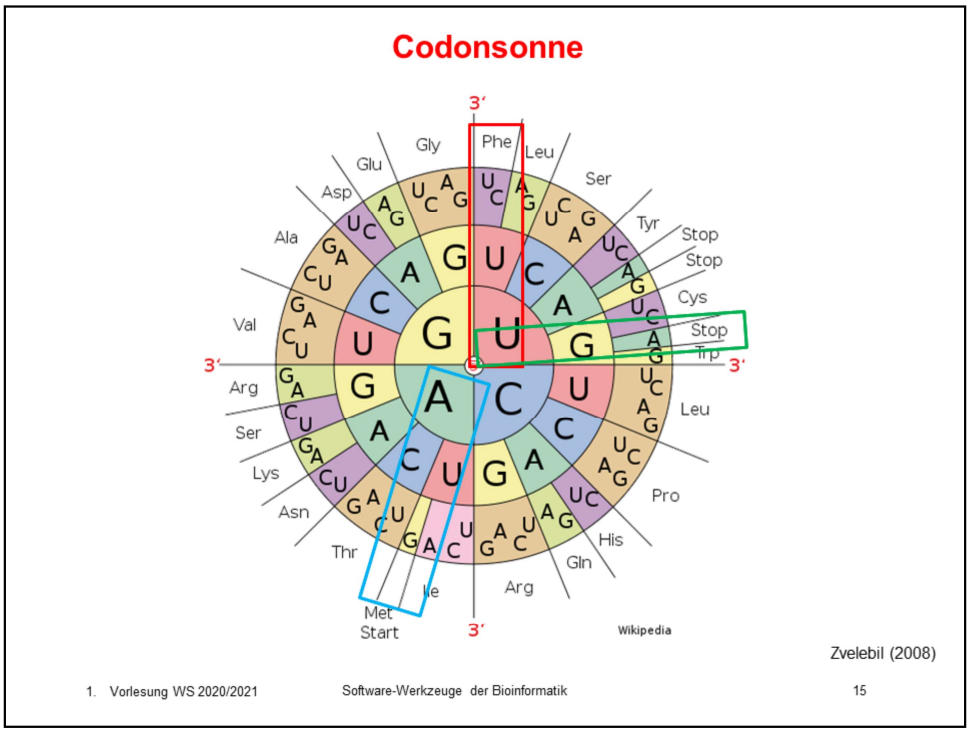
Bei einem **DNA-Doppelstrang** liegen die Basen der Nukleotide des einen DNA-Einzelstrangs den Basen der Nukleotide des anderen Einzelstrangs gegenüber; deren Phosphat-Desoxyribose-Rückgrat zeigt somit nach außen.

Typischerweise bilden dabei jeweils eine (kleinere) Pyrimidinbase (T, C) und eine (größere) Purinbase (A, G) ein Paar. Als komplementär werden die Basenpaare aus T und A sowie aus C und G bezeichnet.

Die einander gegenüberliegenden Basen der Nukleotide zweier Stränge sind in der DNA-Doppelhelix über **Wasserstoffbrückenbindungen** miteinander verbunden. Zwischen den Basen G und C bilden sich drei, zwischen A und T nur zwei. Mit Kraftspektroskopie (AFM) hat man die Bindungskraft bestimmt: „The base binding strength of single dG/dC and single dA/dT were estimated to be 20.0 ± 0.2 pN and 14.0 ± 0.3 pN, ”
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4360479/

Dies liegt offensichtlich an der unterschiedlichen Anzahl an Wasserstoffbrückenbindungen.

Der **T_m-Wert** wird definiert als die Temperatur, bei der 50 % der Doppelhelix in denaturiertem Zustand vorliegen. Der T_m-Wert eines DNA-Doppelstranges ist direkt abhängig von dessen GC-Gehalt. Je mehr GC-Bindungen ein DNA-Molekül enthält, desto höher liegt der T_m-Wert. Wie in der Tabelle gezeigt ist, haben verschiedene Organismen einen sehr unterschiedlichen GC-Gehalt.



Das ist die berühmte „**Codon-Sonne**“. Jeweils drei aufeinanderfolgende Nukleotidbasen bilden ein **Codon**. In diesem Schema enthält der innere Ring das jeweils erste Nukleotid. Der mittlere Ring enthält das mittlere Nukleotid.

Der äußere Ring erhält das dritte Nukleotid.

Außen markiert ist die jeweilige Aminosäure, die das Ribosom anhand des 3-Tripletts an Basen synthetisiert.

Beispielsweise kodieren UUU und UUC für die Aminosäure Phenylalanin („Phe“), durch einen roten Kasten markiert.

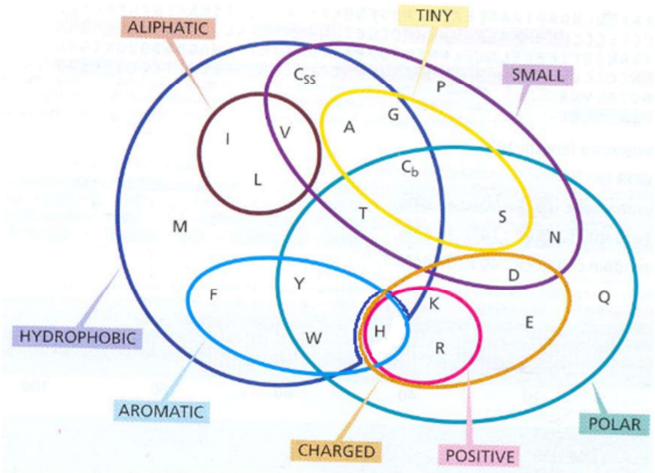
Offensichtlich besteht oft an der dritten Aminosäure eines Codons eine große Redundanz.

Dies liegt daran, dass man durch 3 Basen $4^3 = 64$ verschiedene Tripletts zusammensetzen kann. Allerdings gibt es nur 20 verschiedene natürliche Aminosäuren. 3 Tripletts (UAA, UAG und UGA) sind **Stop-Codons**, bei denen die Transkription durch RNA-Polymerese stoppt. UGA ist durch einen grünen Kasten markiert.

Die Translation in Aminosäuresequenzen beginnt üblicherweise mit einem AUG-Triplett, das durch einen hellblauen Kasten markiert ist.

Eigenschaften der Aminosäuren

Aminosäuren unterscheiden sich in ihren physikochemischen Eigenschaften.



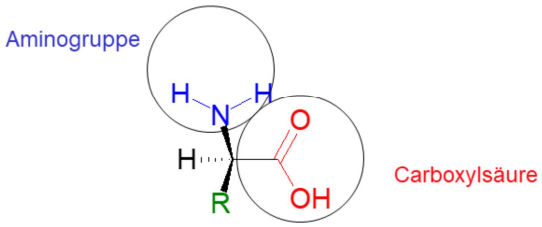
Q: müssen Bioinformatiker die Eigenschaften von Aminosäuren kennen?

Die 20 natürlich in Proteinen vorkommenden Aminosäuren kann man aufgrund ihrer **physikochemischen Eigenschaften** klassifizieren:

Aliphatische (Fett liebende), winzige, kleine, polare, elektrostatisch geladene, positiv oder negativ geladene, aromatische und hydrophobe Aminosäuren.

Einleitung: Aminosäuren

Aminosäuren sind die **Bausteine** von Proteinen:

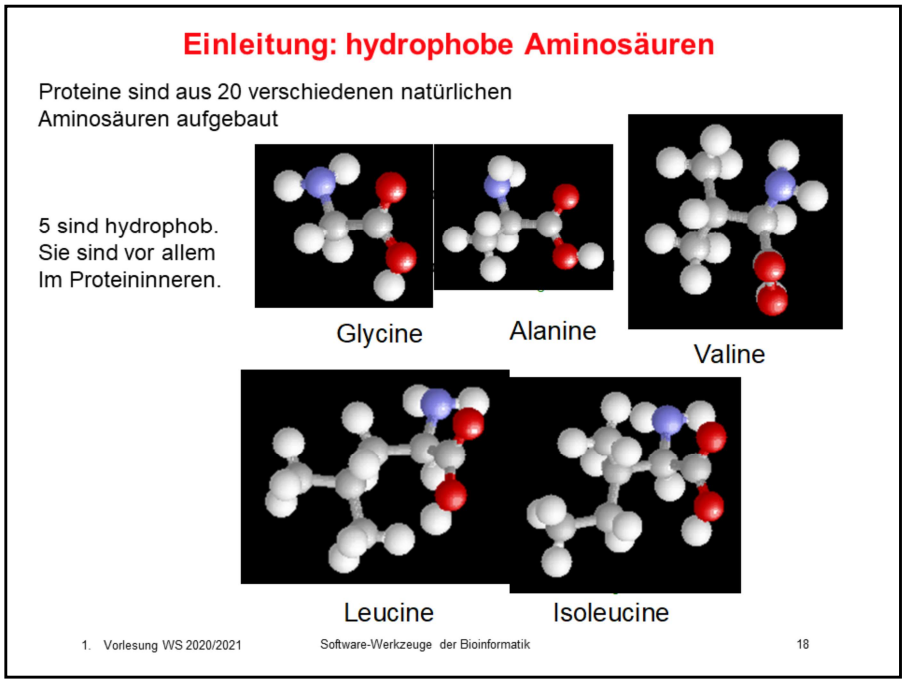


Aminosäuren unterscheiden sich hinsichtlich ihrer
- Größe
- elektrischen Ladung
- Polarität
- Form und Steifigkeit

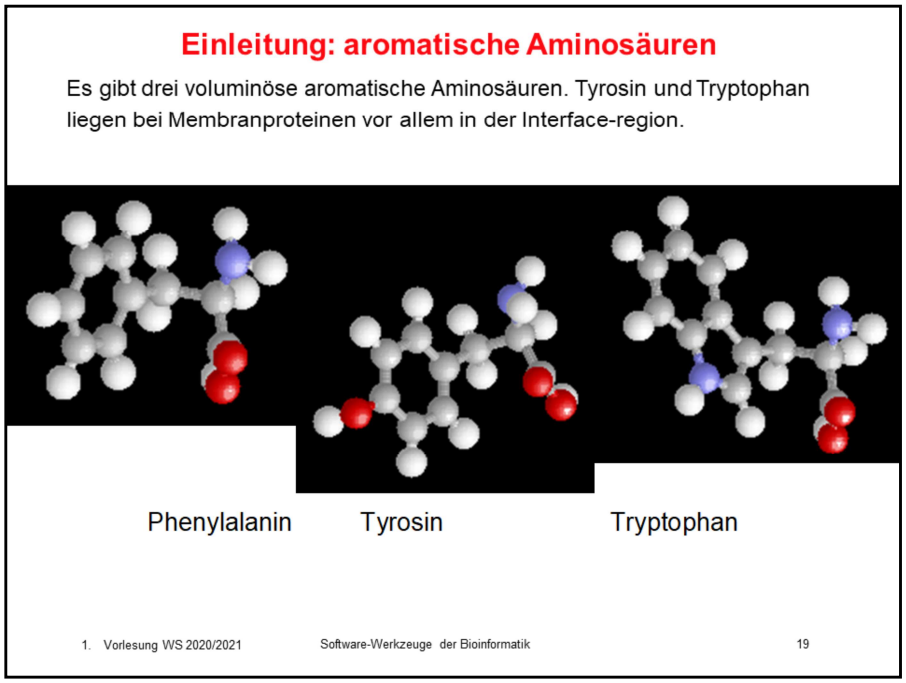
1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 17

Jede Aminosäure enthält eine **Aminogruppe**, eine **Carboxylsäuregruppe** und eine **Seitenkette** („Rest“). Aminogruppe und Carboxylsäuregruppe sind sehr polar. Allerdings gehen diese beiden Gruppen Peptidbindungen mit anderen Aminosäuren ein (bis auf die jeweils erste und letzte einer Aminosäurekette).

Daher hängen die physikochemischen Eigenschaften einer Aminosäure von ihrem Rest ab.



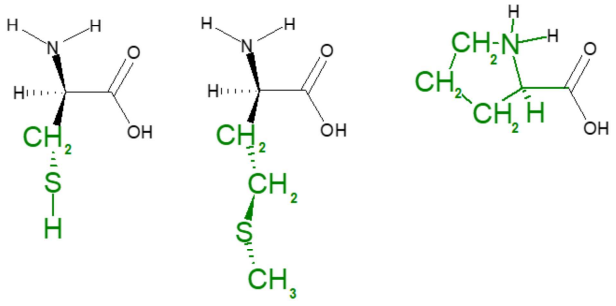
Gezeigt sind hier die 5 hydrophoben Aminosäuren, die man vor allem im **Proteininneren** antrifft.



Von den drei aromatischen Aminosäuren Phe, Tyr und Trp ist Phe die am meisten hydrophobe und Tyr die am meisten polare AS. Man findet diese AS in der Interface-Region von Lipiddoppelschicht-Membranen bzw. an Protein-Protein-Schnittstellen angereichert.

Einleitung: Aminosäuren

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



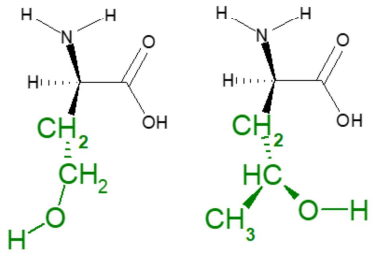
1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 20

Zwei gegenüberliegende Cystein-AS können unter Abspaltung von H_2 **Disulfidbrücken** ausbilden. Dabei wird die Oxidationszahl ihres Schwefelatoms um 1 erhöht, also oxidiert. Meist wird H_2 nicht als Gas abgespalten (bzw. aufgenommen), sondern damit eine andere Disulfidbrücke auf der Oberfläche eines anderen Proteins, oder mit einer anderen Redox-aktiven Sustanz wie GSSH aufgebrochen (bzw. gebildet). Das Schwefelatom von Methionin ist durch seine Methylgruppe nicht frei zugänglich.

Prolin ist eine besondere Aminosäure, da der Aminosäurerest unter Abgabe von H kovalent an die NH-Gruppe des Aminosäurerückgrats gebunden ist. Dadurch ist die Aminosäureseitenkette nicht mehr frei beweglich, sondern kann bzgl. der Atome des Rückgrats nur 2 Orientierungen annehmen, bei denen der AS-Ring entweder oberhalb oder unterhalb der Ebene der Nachbaratome liegt („cis“ bzw. „trans“).

Einleitung: Aminosäuren

Es gibt zwei Aminosäuren mit terminalen polaren Hydroxlgruppen:



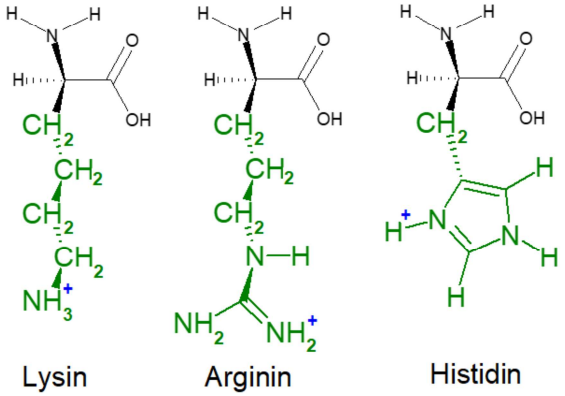
Serin Threonin

1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 21

Die beiden kurzen AS Serin und Threonin haben beide eine relativ frei-bewegliche **Hydroxylgruppe**, die sowohl mit dem H-Atom als auch mit dem O-Atom Wasserstoffbrücken ausbilden kann. Die zusätzliche Methylgruppe von Threonin macht die AS etwas hydrophober als Serin.

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 Protein-oberflächen.



1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 22

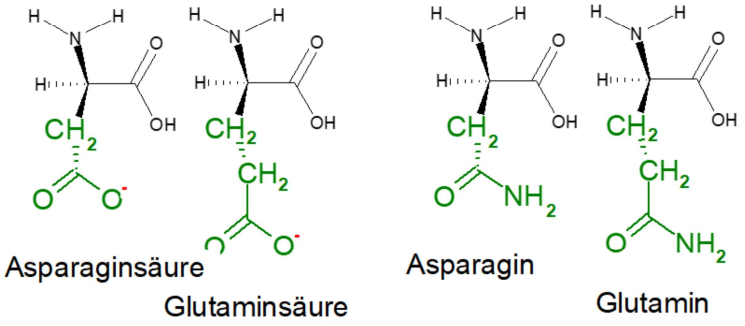
Lysin und Arginin haben relativ **lange Seitenketten**, jeweils mit einer (bei neutralem pH) positiv geladenen, terminalen Aminogruppe. Beim Arginin ist die positive Ladung über beide Aminogruppen delokalisiert. Außerdem besitzt Arginin eine weitere N-H-Bindung. Alle diese N-N-Bindungen sind mögliche Donatoren für Wasserstoffbrückenbindungen.

Histidin kann bei neutralem pH entweder in neutraler Form oder in positiv geladener Form vorliegen. Der pKa-Wert fürs Umklappen zwischen den beiden Formen ist 6.5.

Vorkommen: Arginin hat einen gemischt geladenen / aromatischen Charakter. Es ist an Protein-Protein-Bindungsschnittstellen angereichert. Histidin findet man oft in den aktiven Zentren von Enzymen, da es leicht Protonen aufnehmen und abgeben kann.

Einleitung: Aminosäuren

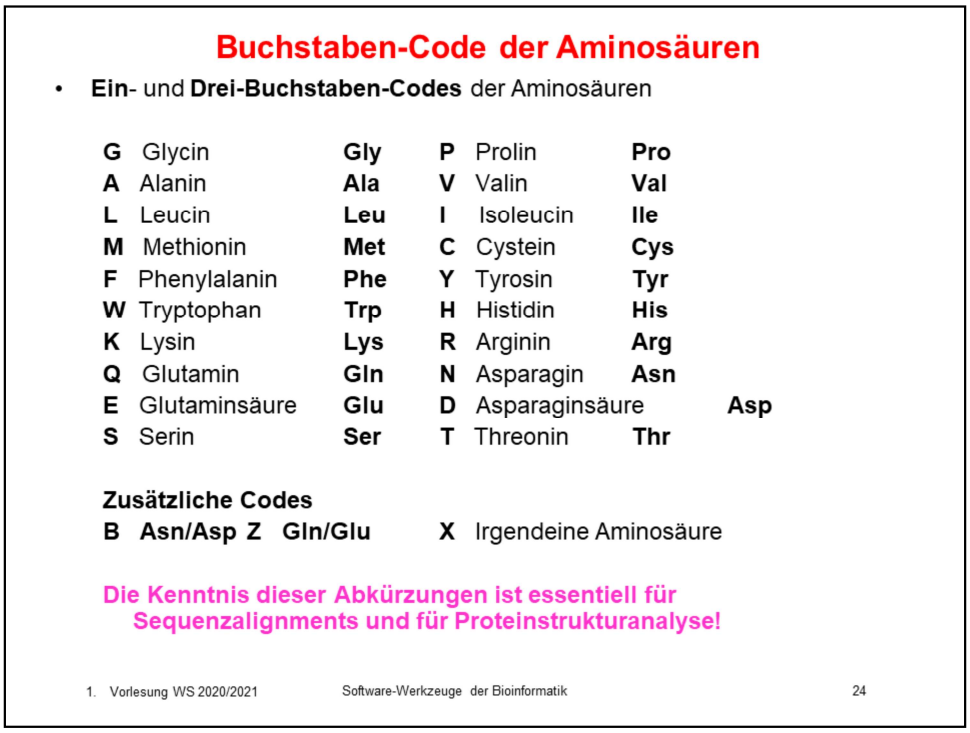
Es gibt 2 negativ geladene Aminosäuren und ihre zwei neutralen Analoga.
Asp und Glu haben pK_a Werte von 2.8. Das heisst, erst unterhalb von pH=2.8 werden ihre Carboxylgruppe protoniert.



1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 23

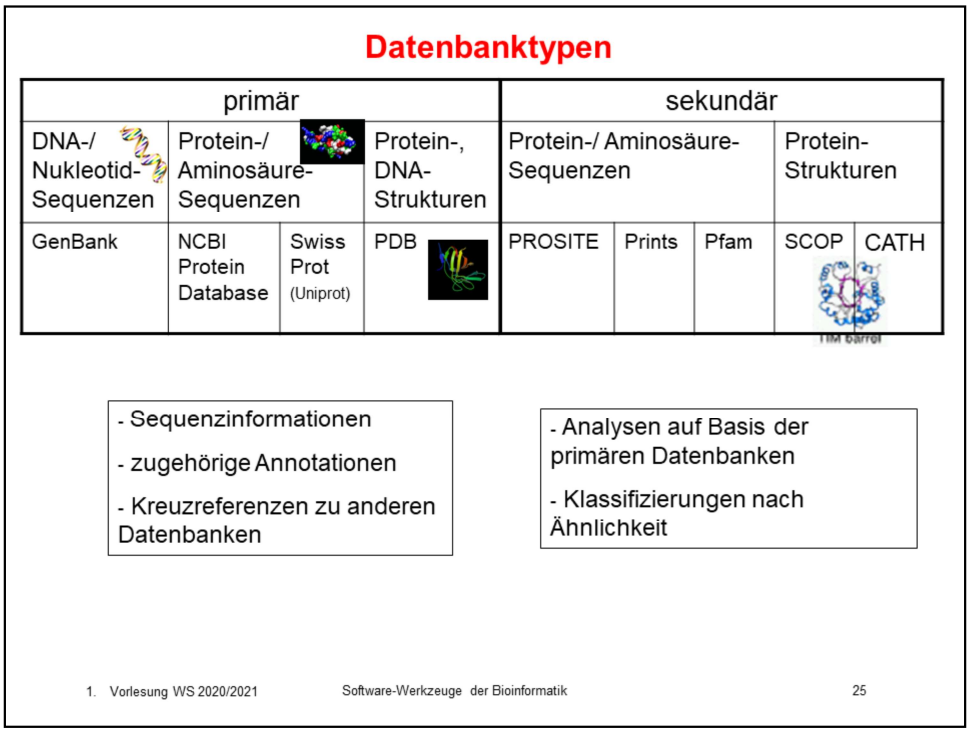
Analog zu den positiv geladenen AS gibt es zwei negativ geladene, saure AS namens Asparaginsäure und Glutaminsäure. Sie sind sehr ähnlich zueinander und unterscheiden sich lediglich in der Länge ihrer Seitenketten.

Zu beiden AS gibt es neutrale Analoge, bei denen die Carboxylgruppe durch eine Amidgruppe ersetzt ist. In Elektronendichtekarten sind NH2 und O kaum voneinander unterscheidbar. Deswegen muss die Orientierung von Asn / Gln-Seitenketten durch Betrachtung ihrer Umgebung bestimmt werden. Die Seitenketten werden so positioniert, dass sich optimale H-Bindungen zu benachbarten AS ergeben.



Um den täglichen Umgang mit Aminosäuren zu vereinfachen und abzukürzen, dient entweder ein 3-Buchstaben-Code oder ein 1-Buchstaben-Code.

Letzterer ist insbesondere für den Umgang mit Proteinsequenzen essentiell. Nur durch das physikochemische Verständnis über die Eigenschaften der 20 AS erschließen sich wichtige strukturelle und mechanistische Eigenschaften von Proteinen. **Sie sollten sich diese beiden Codes daher gut einprägen.**



Ein wichtiges Arbeitswerkzeug für Bioinformatiker sind **Datenbanken**. Datenbanken haben die Hauptaufgabe, Daten effizient und sicher abzuspeichern und zur Verfügung zu stellen. Der Zugriff auf Datenbanken geschieht über eine Frontend-Software, mit der man den gesamten Datenbestand nach Schlüsselwörtern oder Eigenschaften **durchsuchen** kann.

Weiterhin werden die einzelnen, von Forschern eingereichten Datensätze (z.B. Proteinsequenzen oder Proteinstrukturen) üblicherweise vor dem Abspeichern auf die **Qualität** hin überprüft. Wenn man Daten aus einer gut gepflegten Datenbank (z.B. am NCBI oder EBI) verwendet, kann man daher zuversichtlich sein, dass die Daten sorgfältig überprüft wurden.

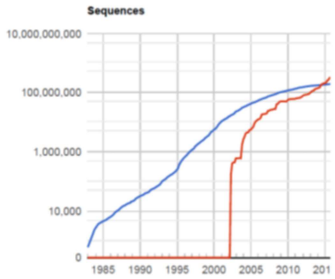
Man unterscheidet **primäre** und **sekundäre Datenbanken**. Primäre Datenbanken enthalten die „Rohdaten“, die z.B. experimentell bestimmt wurden und von den Forschern eingereicht / hochgeladen wurden.

Sekundäre Datenbanken verwenden Daten der primären Datenbanken und bereiten diese in geeigneter Form, z.B. zusammen mit Zusatzinformationen auf.

Sequenzdaten (Genbank) – Strukturdaten (PDB)

• in Aug. 2020 ~218 Mio. **Nukleotidsequenzen**
(Quelle: https://www.ncbi.nlm.nih.gov/genbank/statistics/)

~1408 Mio. WGS-Nukleotidsequenzen



• 168.095 **3D-Strukturen** von biologischen

Makromolekülen (Proteine, DNA, RNA, …)
(Quelle: RCSB-PDB *http://www.rcsb.org*, Aug. 2020)

RCSB PDB
PROTEIN DATA BANK

Einträge sind teilweise **redundant**,

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

1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 26

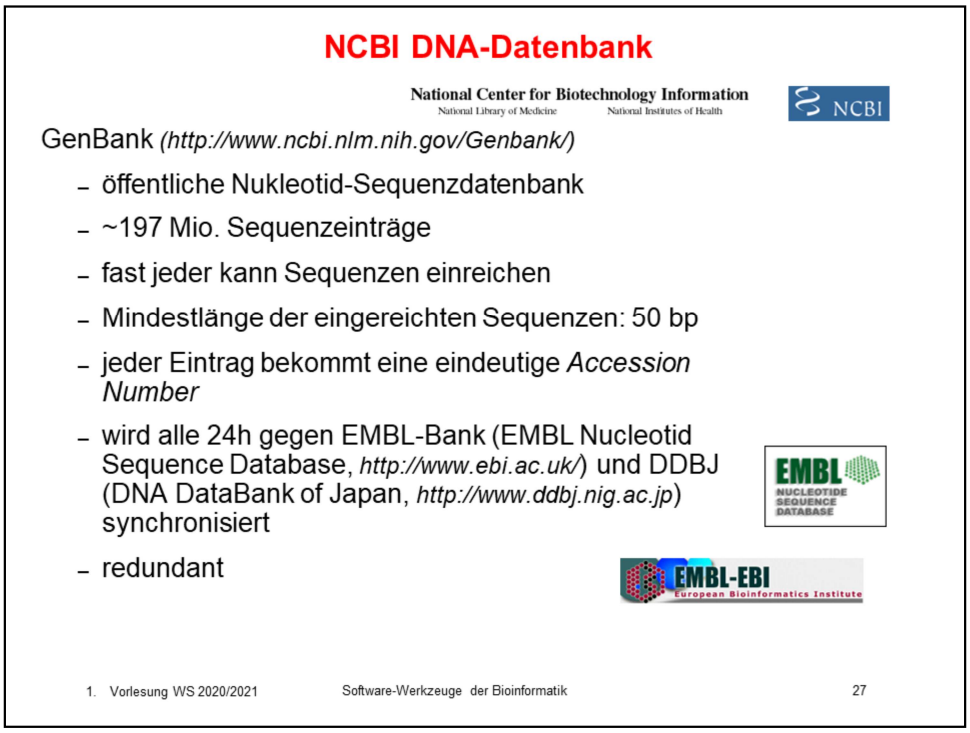
Die beiden wichtigsten Informationsquellen sind NCBI (National Center for Biotechnology Information) in den USA und das EBI (European Bioinformatics Institute) in Hinxton/Cambridge in Grossbritannien.

Die Proteinsequenzen stammen zumeist aus translatierten DNA-Sequenzen. In der logarithmischen Abbildung sind die Anzahl an einzeln bestimmten Nukleotid-sequenzen von Genen (blau) sowie diejenigen aus Ganzgenomsequenzierung (WGS = whole genome sequencing in rot) gezeigt. Die rote Kurve begann erst in 2002, im Wesentlichen mit der Sequenzierung des menschlichen Genoms, und steigt seither steil an. Im Vergleich dazu gibt es etwa 1000 mal weniger hochaufgelöste dreidimensionale Strukturen von Proteinen. Strukturdaten werden in der Protein-Datenbank (PDB) gesammelt. Zudem ist der Anstieg an Strukturdaten wesentlich langsamer als der Anstieg an Sequenzdaten.

Primäre Datenbanken enthalten oft mehrfach dieselbe Information (Proteinsequenz). Diese **Redundanz** resultiert z.B. daraus, dass mehrere Forscher die Sequenz desselben Proteins unabhängig voneinander bestimmten und an die Datenbank schickten. Da man die Qualität des einzelnen Experiments nicht im Detail überprüfen kann, legt die Datenbank dann einfach mehrere unabhängige Einträge an.

Von den meisten Datenbanken gibt es jedoch auch nicht-redundante Versionen, bei denen absolut identische Einträge miteinander verschmolzen werden bzw. nur einer der Einträge behalten wird. Dies ist z.B. nützlich für Homologiesuchen mit dem Programm BLAST (man erhält nicht mehrere Treffer zu identischen Einträgen) und

essentiell für manche Arten von statistischen Auswertungen. Sonst könnten solche Einträge, die mehrfach oder vielfach in den Datenbanken enthalten sind, die statistischen Auswertungen verfälschen.



Bei der Datenbank GenBank kann fast jeder Daten einreichen!

Es gibt mehrere Möglichkeiten dafür, siehe
https://www.ncbi.nlm.nih.gov/genbank/submit/

BankIt ist ein Webtool, **tbl2asn** ist ein command-line Programm (falls man eine grössere Anzahl an Sequenzen oder ganze Genome einreichen möchte).

Genbank, EMBL-Bank und DDBJ enthalten aufgrund der täglichen Synchronisierung im Prinzip dieselben Sequenzen.

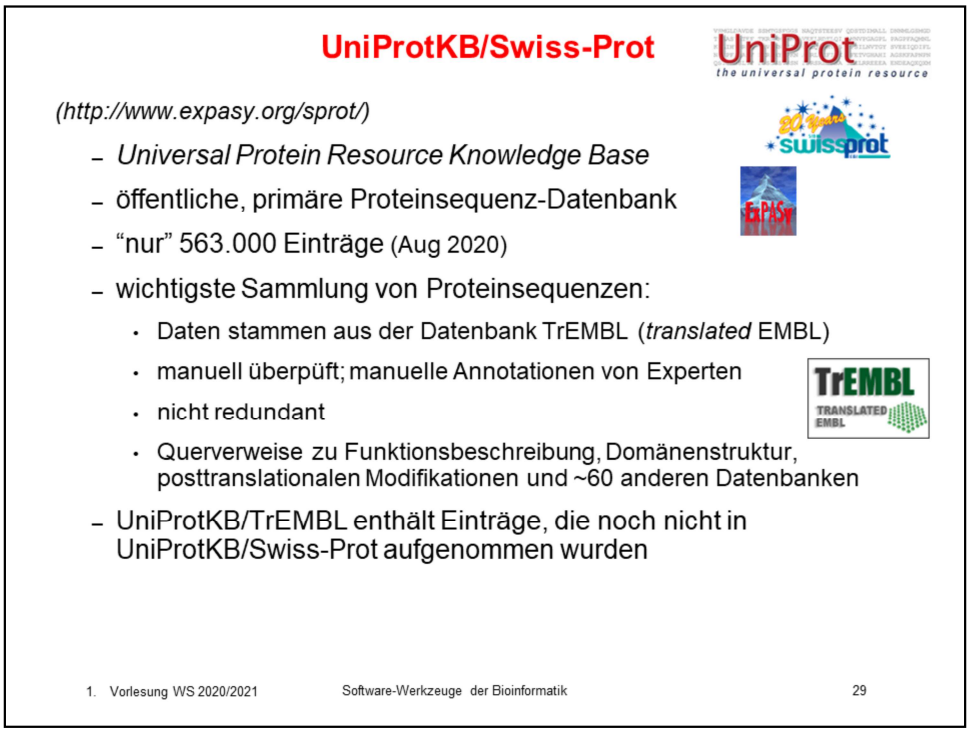
NCBI Protein-Datenbank

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

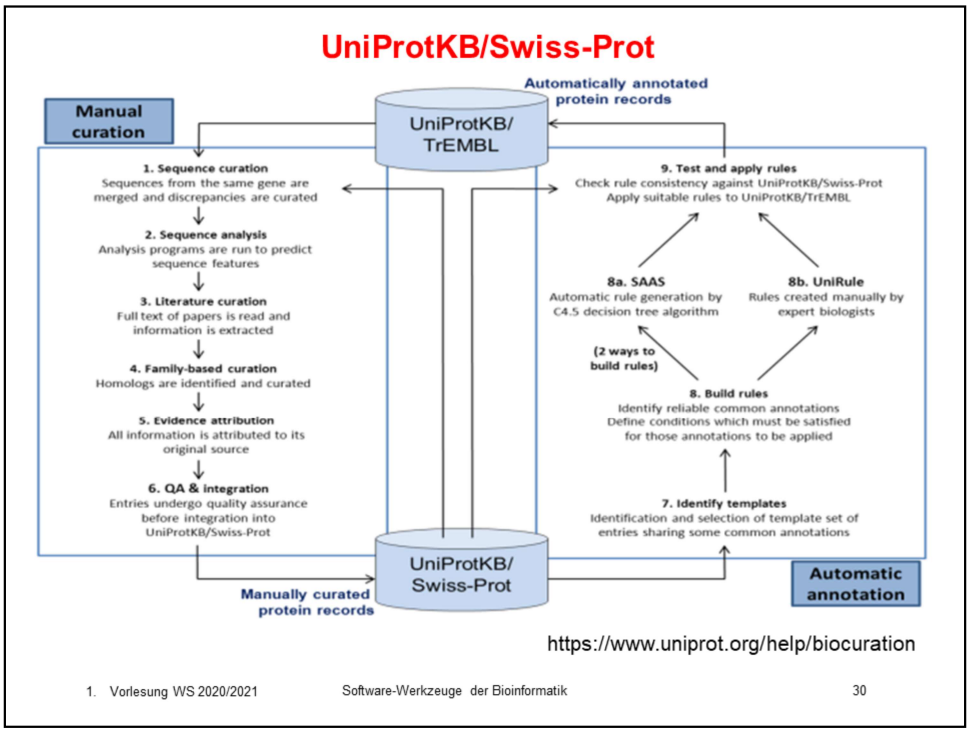
- öffentliche, primäre Protein-Sequenzdatenbank
- Zusammenstellung aus den folgenden Protein-Sequenzdatenbanken:
  - UniProtKB
  - PIR (Protein Identification Resources)
  - PDB (Protein Data Bank, Strukturen)
  - Proteintranslationen der GenBank-Datenbank
  - und weiteren
- redundant
- Vorteil: Links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

Für Proteinsequenzen gibt es wesentlich mehr primäre Datenbanken als für Nukleotidsequenzen.

Die am sorgfältigsten gepflegte darunter ist UniProtKB/SwissProt

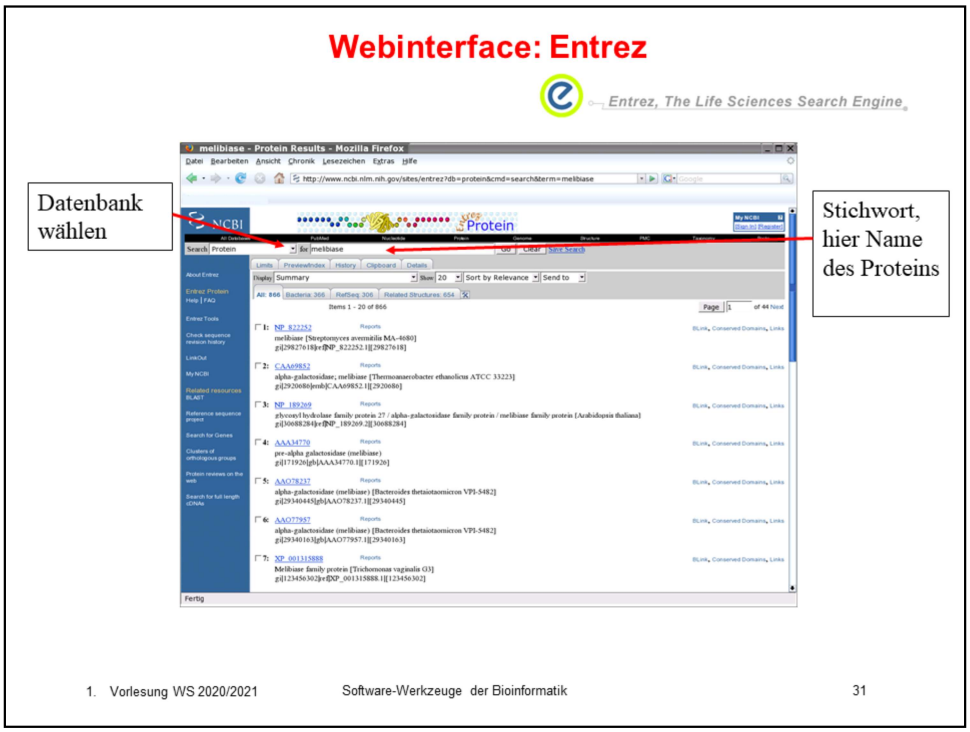


UniProt/SwissProt ist unsere "Lieblingsdatenbank". Die Daten stammen aus dem TrEMBL-Pool.



Diese Übersicht fasst die Analyseschritte bei Swissprot (linke Hälfte des Bildes) zusammen. Der dritte Schritt (wissenschaftliche Literatur lesen) ist aufwändig und wird typischerweise von promovierten Wissenschaftlern durchgeführt.

Auf der rechten Hälfte wird beschrieben, wie die Regeln aufgestellt werden, mit denen die Einträge von TrEMBL automatisch annotiert werden.



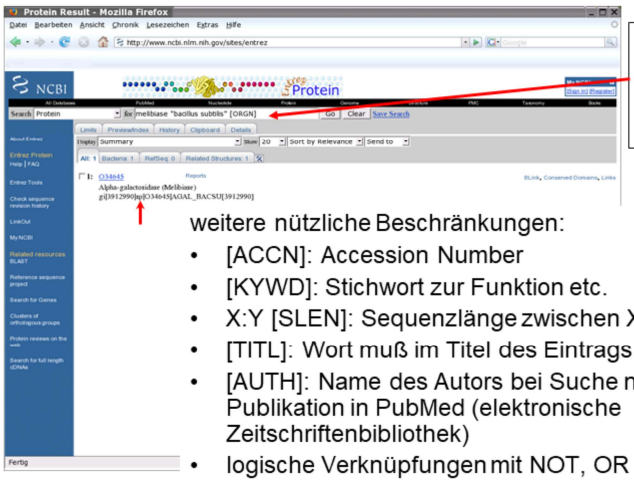
Entrez integriert viele Datenbanken.

Man kann entweder in allen Organismen oder nur in einzelnen nach Treffern suchen.

Es gibt simple und detaillierte Suchfunktionen sowie Links zu Einträgen der Datenbanken.

Wenn man z.B. nach “Melibiase” sucht, bekommt man eine große Anzahl an Treffern, da durch Textsuche alles angezeigt wird, wo irgendwo im Eintrag "Melibiase" steht.

Detaillierte Suche bei Entrez



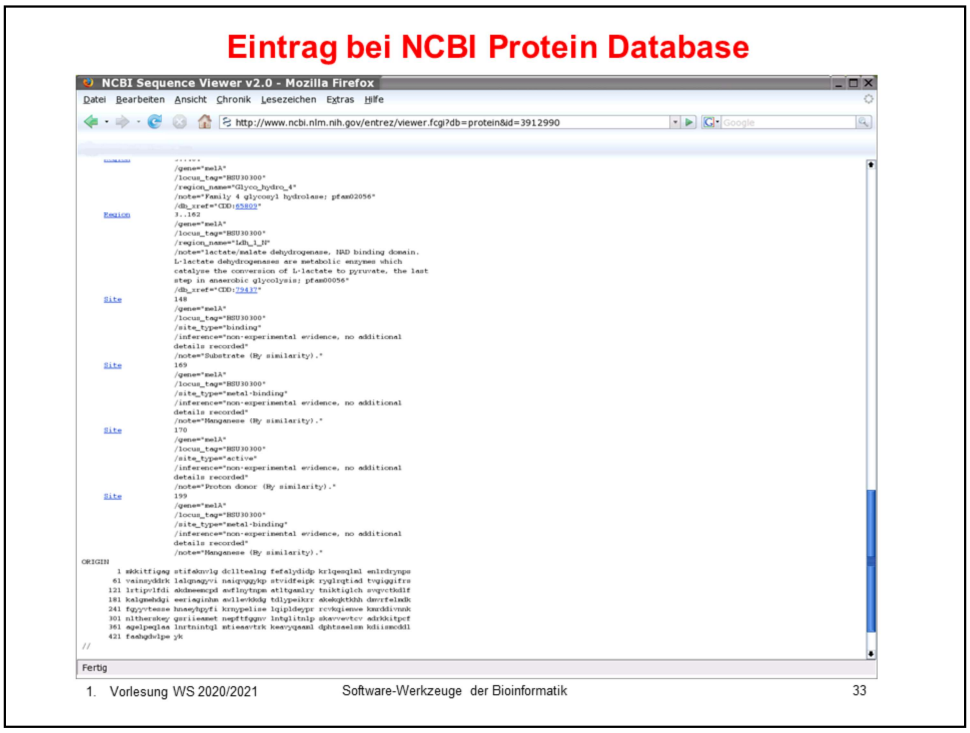
Suche nach dem Protein Melibiase in genau diesem Organismus

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

1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 32

Verschiedene Organismen enthalten Melibiasen. Damit der gewünschte Organismus gefunden wird, ist eine Beschränkung nötig! -> Nur noch ein Eintrag, der gewünschte, den wollen wir uns jetzt mal anschauen.



Der Eintrag enhält verschiedene interessante Details: Accession Number, Beschreibung, Stichworte, die Publikation dazu, wie das Gen heißt, wo die funktionalen Regionen sind, am Ende die Sequenz. Die kann man sich auch "schöner" anzeigen lassen.

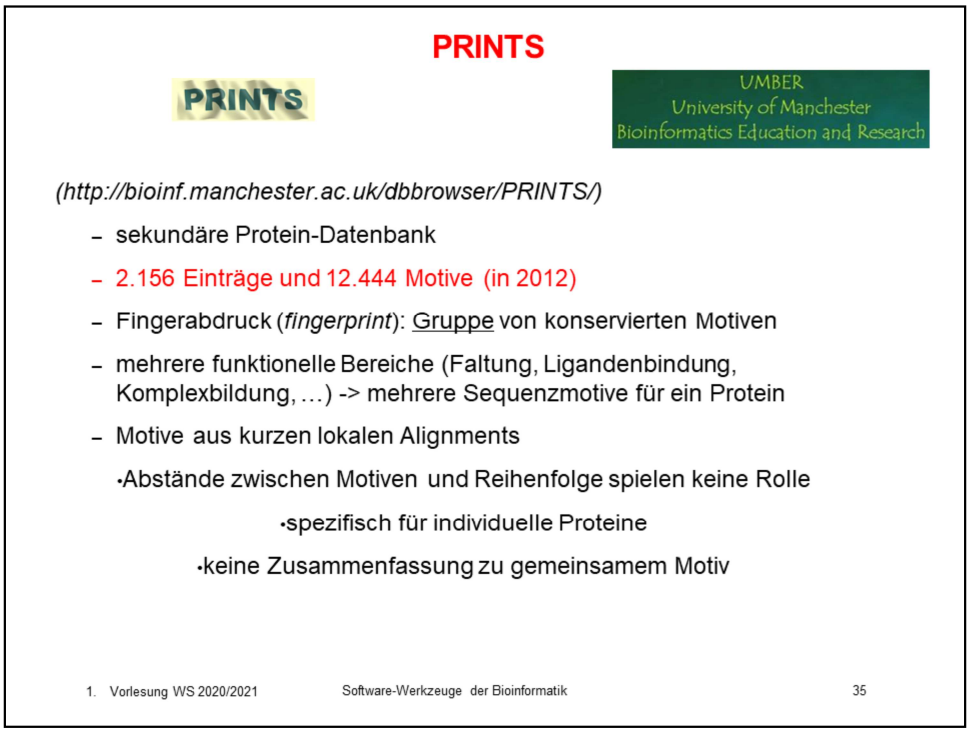
Fasta-Format

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

>DNA-Sequenz-Bezeichnung
ACGT
....
>Protein-Sequenz-Bezeichnung
ACDEFGHIKLMNPQRSTVWY
....

1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 34

und jetzt schauen wir uns mal an, mit welchen interessanten Datenbanken der Eintrag verlinkt ist



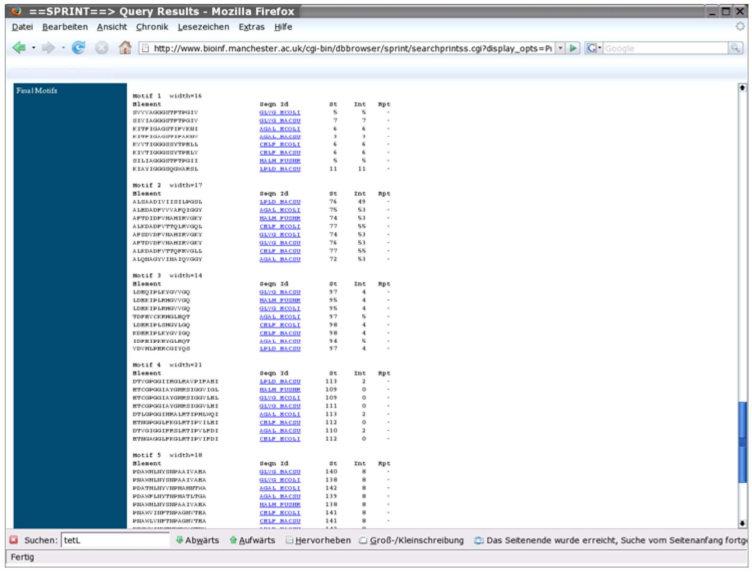
Prints ist eine sekundäre Datenbank, enthält also Zusatzinformationen zu Proteinsequenzen.

Für ein Protein liefert Prints einen **Fingerabdruck** aus mehreren konservierten Sequenzmotiven. Obwohl diese Motive in der Sequenz weit auseinander liegen können, können diese Abschnitte im 3-dimensionalen Raum, d.h. in der gefalteten Proteinstruktur evtl. eng beieinander liegen. Diese Motive können entweder für die Funktion des Proteins wichtig sein (z.B. in dem aktiven Zentrum eines Proteins) oder auch einfach für dessen stabile Struktur.

Anhand eines Fingerabdrucks kann man ein einzelnes Protein sicher bestimmen.

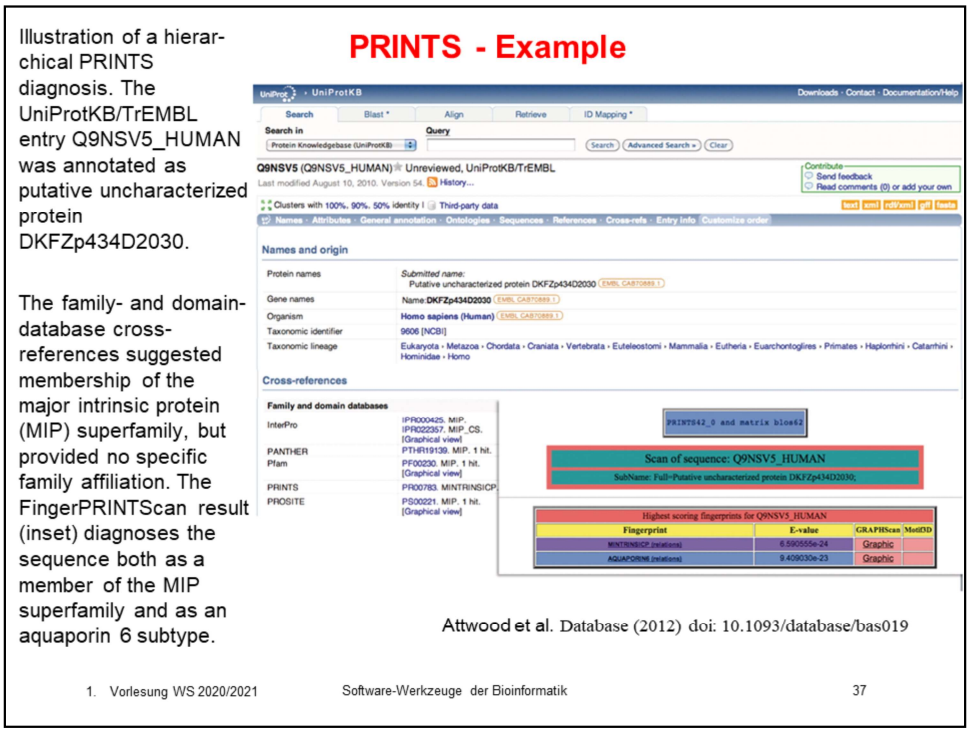
Es ist aber mühsam, mit solchen Gruppen von Sequenzmotiven zu suchen …

Finger-PRINTS

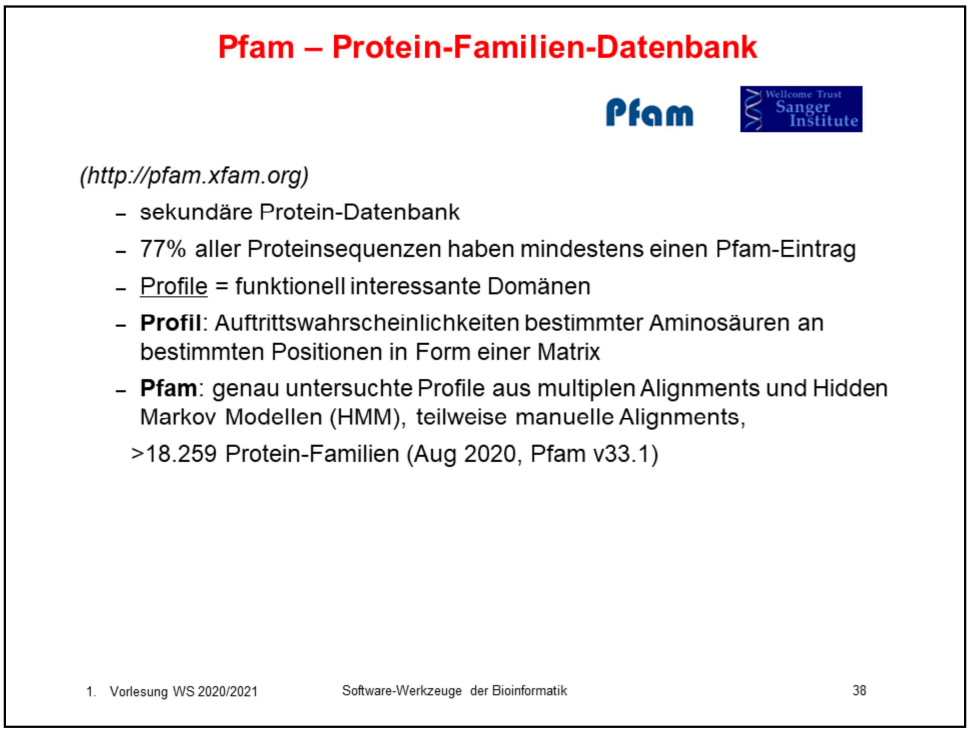


Diese Übersicht zeigt, aus welchen Sequenzen der Fingerabdruck abgeleitet wurde (die Sequenzen sind auch alle verlinkt), wieviele Sequenzen wieviele der Prints enthalten, und eine Liste der ganzen Motive.

Die Darstellung ist ein wenig unübersichtlich. Darum gibt es die Möglichkeit, die einzelnen Motive zu einer Matrix zusammenzufassen. DAS macht PFAM.

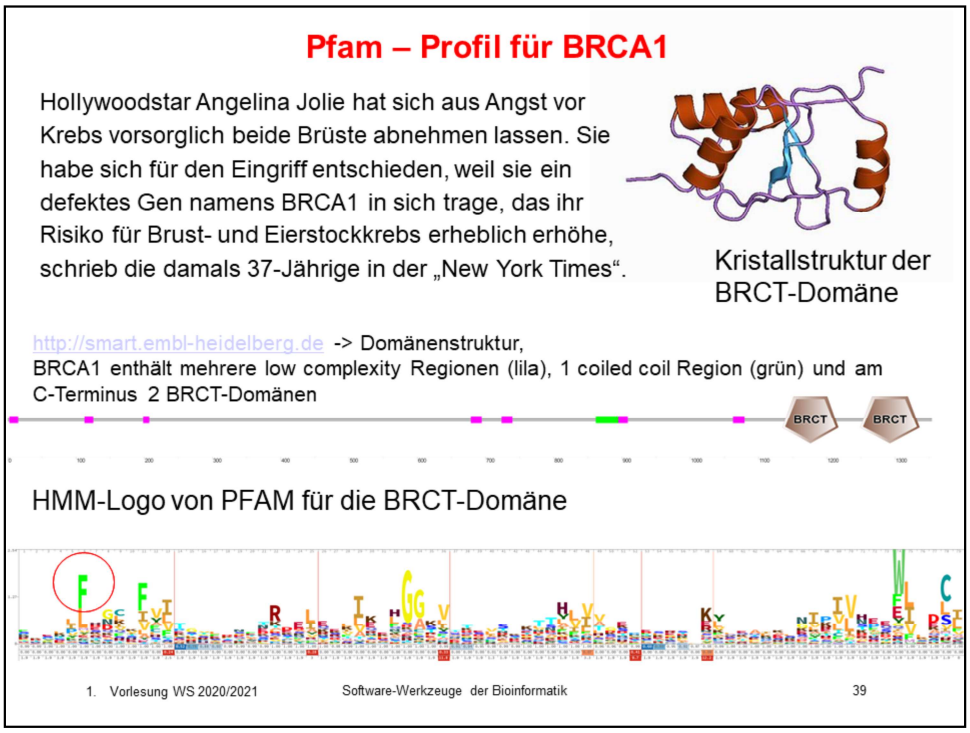


Dieses Beispiel stammt aus der Publikation, die PRINTS beschreibt.



Wenn wir wissen wollen, zu welcher Familie ein Protein gehört, benutzen wir normalerweise Pfam (Pi-Fäm). Dies ist ebenfalls eine sekundäre Datenbank, wird aber regelmässig gepflegt (d.h. bekommt updates).

Der Unterschied zwischen Motiven und Profilen ist: **Motive** enthalten an jeder Position einen Buchstaben, **Profile** enthalten üblicherweise 20 Zahlen für die Häufigkeit der 20 natürlichen Aminosäuren an jeder Position der Sequenz. Ein solches Profil ist dann eine Matrix und enthält wesentlich mehr Informationen als ein Motiv. Anstelle von Profilen erstellt Pfam HMM-Profile, also Übergangswahrscheinlichkeiten in einem Hidden Markov-Modell. HMMs behandeln wir in der vierten Vorlesung.



BRCA1 ist ein sehr kompliziert aufgebautes Gen. Nur für kurze Abschnitte haben wir eine Idee über die dreidimensionale Struktur des davon kodierten Proteins.

Am C-Terminus liegen 2 sogenannte BRCT-Domänen. Oben rechts ist die Kristallstruktur einer solchen Domäne gezeigt.

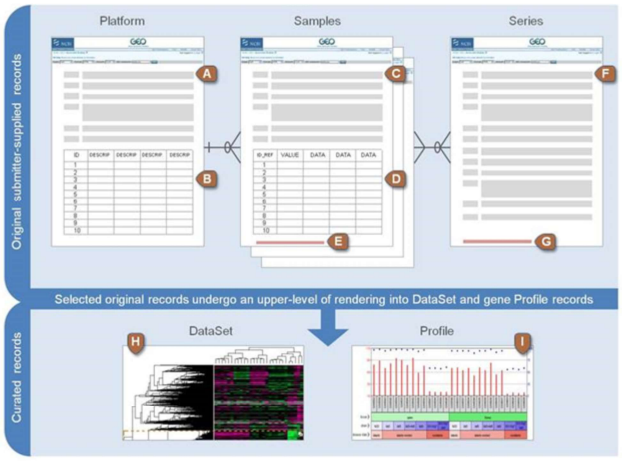
In der Mitte sehen wir aneinandergereiht die vorhergesagte Domänenstruktur von BRCA1.

Unten sehen wir das Pfam-Logo für die BRCT-Domäne. In solch einem Sequenzlogo ist an jeder Position gezeigt, wie häufig die einzelnen Aminosäuren an jeder Position sind. Von links gesehen, ist die erste Position mit einer klaren Präferenz die Position 6, an der sich am häufigsten die Aminosäure F (für Phenylalanin) befindet (rot umkreist).

GEO – Gene Expression Omnibus

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

- Genexpressions-Datensätze
- entweder mit Microarrays oder NGS gemessen



1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 40

Eine andere wichtige Datenquelle ist GEO, wo Genexpressionsdatensätze und andere Omics-Daten gespeichert sind.

Wir kommen in Vorlesung 8 darauf zurück.

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.

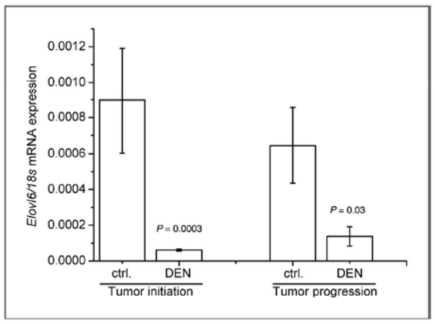


Figure 2. Wild-type mice were treated with the carcinogen DEN at the age of 2 weeks. Livers were analyzed after 24 weeks to assess the tumor initiation state. Analyses in the tumor progression stadium were done after 36 weeks. *Elovl6* mRNA expression as determined by real-time reverse transcriptase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

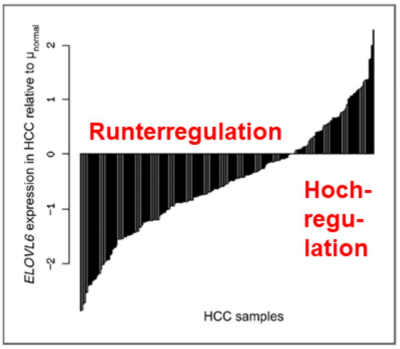


Figure 1. mRNA levels of *ELOVL6* in 247 human HCC samples relative to the mean of 239 nontumor liver tissue (μ_{normal}). Samples of dataset GSE14520 [$\log_2$ (expression) values from GEO after Robust Multi-array Average normalization] were mapped to hgu133a.db using bioconductor. Significance values: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

1. Vorlesung WS 2020/2021 Software-Werkzeuge der Bioinformatik 41

Dies ist ein Beispiel aus unserer gemeinsamen Forschung mit der Arbeitsgruppe von Prof. Alexandra Kiemer in der Pharmazie. Die Kiemer-Gruppe interessiert sich für die Entstehung und Charakterisierung von **Leberkrebs** (hepatozelluläres Karzinom, HCC). In Mäusen kann man durch Zugabe der krebserregenden Substanz Diethylnitrosamin die Entstehung von HCC auslösen.

Wie im linken Plot gezeigt wird, ist nach 2 Wochen („tumor initiation“) z.B. die mRNA für das Protein C18 Fettsäure-produzierende Elongase (**Elovl6**) stark runterreguliert. Nach 24 Wochen („tumor progression“) ist der Unterschied immer noch signifikant (p-Wert < 0.05). Allerdings interessiert sich kaum jemand in der normalen Gesellschaft für die Entstehung von Leberkrebs in Maus.

Die Frage war also, ob im Menschen eine ähnliche Runterregulation stattfindet. Dazu ist in der rechten Abbildung die ELOVL6-Expression in HCC-Patienten im GEO-Datensatz GSE14520 im Vergleich zu gesundem Lebergewebe gezeigt.

Der Großteil der Patienten zeigte ebenfalls eine Runterregulation (linker Teil), aber nicht alle. Die Signifikanz der Ergebnisse ist im rechten Beispiel dramatisch höher (p-Werte 10^{-11}). Woran liegt dies? Dies werden wir in der Vorlesung diskutieren.

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.

Dies war bereits die erste Vorlesung.

Zu Beginn wurden einige formale Punkte zu Vorlesung und Tutorial vorgestellt.

In der vierten Woche erhalten Sie bereits das Thema für Ihr erstes Mini-Forschungsprojekt.

Entsprechend den bis dahin behandelten Themen wird es sich mit Proteinsequenzen beschäftigen.

Bis dahin sollten Sie sich mit den Eigenschaften und Abkürzungen der Aminosäuren vertraut gemacht haben.

In der nächsten Woche behandeln wir das Thema “paarweises Alignment von Proteinsequenzen”.

Bitte denken Sie daran, sich die Folien VOR der Vorlesung schon einmal anzuschauen.