Inhaltsverzeichnis

1 Aminosäure-Sequenzvergleiche und Proteinstrukturen....2
1.1 Aminsäure-Sequenzvergleich vier verschiedener Arten am Beispiel von Cytochrom b561....2
1.1.1 Suche in der UniProt-Proteindatenbank....2
1.1.2 Aminosäure-Sequenzvergleich direkt im Webbrowser durchführen....3
1.1.3 Herunterladen der Sequenzinformationen im FASTA-Format....4
1.1.4 Druckausgabe für wissenschaftlichen Artikel mit dem LaTeX Paket Texshade erzeugen....4
1.1.5 Baumansicht online und mit lokalem Programm erzeugen....5
1.2 Aminsäure-Sequenzvergleich am Beispiel Fasciculin (Schlangengift)....6
1.2.1 Suche in der UniProt-Proteindatenbank....6
1.2.2 Alignment direkt im Webbrowser durchführen....6
1.2.3 Abbildung der 3D-Ansicht eines bestimmten Proteins mit bestimmten Einstellungen....7
1.2.4 Membran-Diagramme mit TexTopo am Beispiel Cytochrome b561 domain-containing protein 1 des Menschen....8
1.3 Übersicht zu Aminosäuresequenz- und Proteindatenbanken....9
1.4 Übersicht zu Dateiformaten für Proteine....9
1.5 Übungen....10
1.5.1 Pre-B-cell Colony- Enhancing Factor (PBEF) 10
1.5.2 Hämoglobin....13
2 Nucleotid-Sequenzvergleiche....15
2.1 dUTP pyrophosphatase von E. coli....15
2.1.1 Suche in der UniProt-Proteindatenbank....15
2.1.2 Nukleotid-Sequenzen im FASTA-Format herunterladen....15
2.1.3 Nukleotid-Sequenzvergleich mit BLAST....16
2.2 Dateiformate für Nukleotidsequenzen....18
2.3 Übersicht zu Nukleotidsequenz- und Gendatenbanken....18
2.4 Übungen....18
2.4.1 Zu welchem Enzym gehört ein bestimmtes DNA-Fragment?....18
2.4.2 Homologe Gene, die in Eukaryoten konserviert sind....19
3 Fachbegriffe und Abkürzungen....20
4 Software....23
4.1 Texmaker, Texshade und Textopo (Druckausgabe von Alignments)....23
4.2 Clustalx (globale und lokale Alignments erzeugen)....23
4.3 NJplot (Bäume anzeigen und Druckausgabe erzeugen)....23
4.4 Emboss (globale und lokale Alignments erzeugen)....23
4.5 PHYLIP (PHYLogeny Inference Package)....23
4.6 MAFFT (Multiple alignment program for amino acid or nucleotide sequences)....23
4.7 Archaeopteryx (Bäume anzeigen und bearbeiten)....24
5 Workflow....25
5.1 Aminosäure-Sequenzen....25
5.2 Nukleotid-Sequenzen....26
6 Literaturverzeichnis....27

1 Aminosäure-Sequenzvergleiche und Proteinstrukturen

In vielen Standard-Lehrbüchern findet man im Evolutionskapitel zu Homologien auf molekularer Ebene Abbildungen von Ausschnitten der Aminosäure-Sequenzen verschiedener Arten [1], [2].

Aus dem Vergleich der Aminosäure-Sequenzen wird dann häufig ein daraus abgeleiteter molekularer Stammbaum angegeben. Daneben zeigen manche Bücher die Ähnlichkeit des gleichen Proteins bei verschiedenen Arten zusätzlich auch auf Ebene der Tertiärstruktur anhand einer räumlichen Darstellung der gesamten Polypetidkette [3].

Als Schüler kann man durch Nachvollziehen dieses Einführungslehrgangs ein tieferes Verständnis sowohl für die Realisierung der genetischen Information gewinnen, als auch für die Bewertung von Homologien auf molekularer Ebene.

Als Lehrer kann man die hier vorgestellten Software-Werkzeuge und Methoden für die eigene Unterrichtsvorbereitung und nicht zuletzt auch für die grafische Gestaltung von Leistungserhebungen mit selbst gewählten Beispielproteinen nutzen.

Die Kapitel 1.1.4 und 1.2.4 beschäftigen sich ergänzend mit der Anfertigung diesbezüglicher Abbildungen unter Zuhilfenahme TEX basierter Textsatzprogramme für wissenschaftliche Arbeiten, etwa im Rahmen einer W-Seminararbeit. Für die Bewertung der Aussagekraft von Abbildungen in Fachzeitschriften oder Büchern ist es sicher hilfreich, sich auch der Einschränkungen bewusst zu sein, die mit der technischen Umsetzung einhergehen.

1.1 Aminsäure-Sequenzvergleich vier verschiedener Arten am Beispiel von Cytochrom b561

1.1.1 Suche in der UniProt-Proteindatenbank

www.uniprot.org > Tab: Search > Button: Advanced Search > Field > Protein Name [DE]: Cytochrome b561

	Entry	Entry name	Status	Protein names	Gene names	Organism	Length
☑	P0ABE5	C561_ECOLI	★	**Cytochrome b561**	**cybB** b1418 JW5224	Escherichia coli (strain K12)	176
☑	P49447	CY561_HUMAN	★	**Cytochrome b561**	**CYB561**	Homo sapiens (Human)	251
☑	Q60720	CY561_MOUSE	★	**Cytochrome b561**	**Cyb561** Mcyt	Mus musculus (Mouse)	250
☑	P0ABE6	C561_SHIFL	★	**Cytochrome b561**	**cybB** SF1794 S1478	Shigella flexneri	176

Tipp: Um die englische Bezeichnung z.B. eines Proteins oder einer Art zu erhalten, kann man einfach zunächst den deutschsprachigen Wikipedia-Artikel (de.wikipedia.org/wiki/Cytochrome) aufrufen und dann im linken Seitenmenü unter In anderen Sprachen > English auswählen.
Für einzelne Begriffe oder kurze Texte kann auch unter translate.google.com der zu übersetzende Text einfach direkt übersetzt werden, wobei beim Überfahren der Wörter im Zielfeld ein Auswahlmenü für mögliche Synonyme erscheint.

1.1.2 Aminosäure-Sequenzvergleich direkt im Webbrowser durchführen

Oben wechselt man nun innerhalb der UniProt-Seite in den Tab: Align und kopiert die Einträge aus der Spalte Entry name aus der unten auf der Webseite weiterhin angezeigten Ergebnisliste manuell ganz oben in das Feld Sequences (in FASTA format) or UniProt identifiers jeweils in eine neue Zeile und klickt anschließend rechts daneben den Button: Align an:

C561_ECOLI
C561_SHIFL
CY561_HUMAN
CY561_MOUSE

Der Vorteil dieser Methode besteht darin, dass die spätere Reihenfolge der Zeilen in der Sequenzvergleichs-Darstellung einfach über die Abfolge der eingegebenen UniProt identifier selbst festgelegt werden kann.

Nach einer kurzen Wartezeit erscheinen dann unten u.a. die beiden Abschnitte Alignment (Sequenzvergleich) und Tree.
Rechts neben dem Alignment kann die Ansicht im Webbrowser dann noch nach verschiedenen Kriterien (wie z.B. Ähnlichkeit, Polarität, Ladungen) eingefärbt werden, z.B. unter Amino acid properties > Similarity.
Für die Darstellung wird eine nichtproprotionale Schriftart mit fester Breite (Monospace) wie `Courier New` gewählt:

```
1    ------------------------------------------------------------    0  P0ABE5   C561_ECOLI
1    ------------------------------------------------------------    0  P0ABE6   C561_SHIFL
1    MEGGAAAATPTALPYYVAFSQLLGLTLVAMTGAWLGLYRGGIAWESDLQFNAHPLCMVIG   60  P49447   CY561_HUMAN
1    -MEHSSASVPAALPYYVAFSQLLGLTVVAVTGAWLGLYRGGIAWESSLQFNVHPLCMVIG   59  Q60720   CY561_MOUSE


1    --------------MENKYSRLQISIHW--LVFLLVIAAYCAMEFRGFFPRSDRPLINMI   44  P0ABE5   C561_ECOLI
1    --------------MENKYSRLQISIHW--LVFLLVIAAYCAMEFRGFFPRSDRPLINMI   44  P0ABE6   C561_SHIFL
61   LIFLQGNALLVYRVFRNEAKRTTKVLHGLLHIFALVIALVGLVAVFDYHRKKGYADLYSL  120  P49447   CY561_HUMAN
60   MIFLQGDALLVYRVFRREAKRTTKILHGLLHVFAFIIALVGLVAVFDYHKKKGYADLYSL  119  Q60720   CY561_MOUSE
                  :..: .*    :*    :* ::**    : .  :. :.    :  :

45   HVSCGISILVLMVVRLLLRLKYPTPPIIP-------KPKPMMTGLAHLGHLVIYLLFIAL   97  P0ABE5   C561_ECOLI
45   HVSCGISILVLMVVRLLLRLKYPTPPIIP-------KPKPMMTGLAHLGHLVIYLLFIAL   97  P0ABE6   C561_SHIFL
121  HSWCGILVFVLYFVQWLVGFSFFLFPGASFSLRSRYRPQHIFFG------ATIFLLSVGT  174  P49447   CY561_HUMAN
120  HSWCGILVFVLYFVQWLVGFSFFLFPGASFSLRSRYRPQHIFFG------ATIFLFSVGT  173  Q60720   CY561_MOUSE
     *  *** ::** .*: *: :.:   *          :*: :: *       .*:*: :.

98   PVIGL--VMMYNRGNPWFAFGLTMPYASEANFERVDSLKSWHETLANLGYFVIGLHAA-A  154  P0ABE5   C561_ECOLI
98   PVIGL--VMMYNRGNPWFAFGLTMPYASEANFERVDSLKSWHETLANLGYFVIGLHAA-A  154  P0ABE6   C561_SHIFL
175  ALLGLKEALLFNLGGKYSAFEPE------GVLAN------------VLGLLLACFGGAVL  216  P49447   CY561_HUMAN
174  ALLGLKEALLFKLGSKYSTFEPE------GVLAN------------VLGLLLVCFGVVVL  215  Q60720   CY561_MOUSE
      ::**  .:::: *  : :*         . : .             ** ::  :  .

155  ALAHHYFWKDNTLLR-----M----MPRKRS----  176  P0ABE5   C561_ECOLI
155  ALAHHYFWKDNTLLR-----M----MPRKRS----  176  P0ABE6   C561_SHIFL
217  YILTRADWKRPSQAEEQALSMDFKTLTEGDSPGSQ  251  P49447   CY561_HUMAN
216  YILAQADWKRPSQAEEQALSMDFKTLTEGDSPSPQ  250  Q60720   CY561_MOUSE
      :  :  **  :  .    *    : .  *
```

1.1.3 Herunterladen der Sequenzinformationen im FASTA-Format

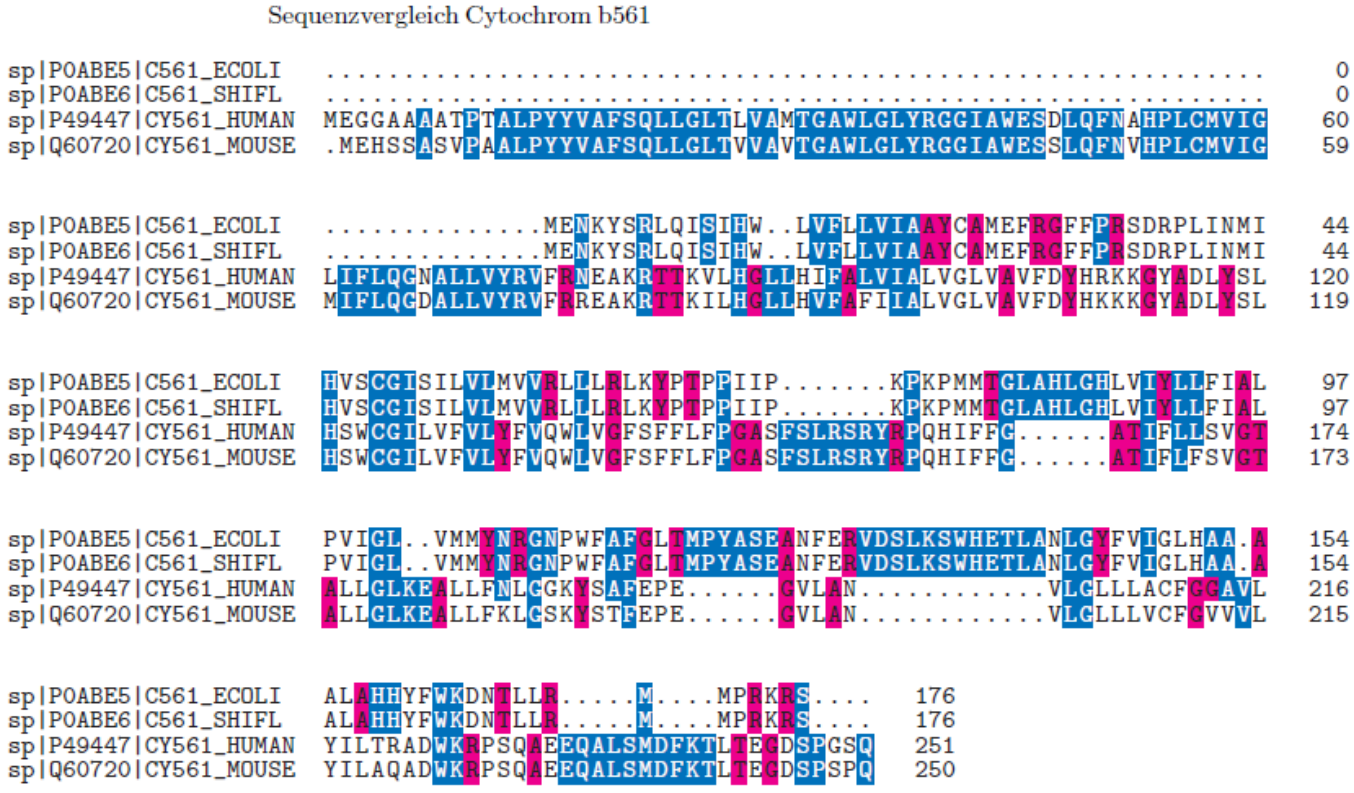
Man wählt nun die orangefarbene Schaltfläche fasta rechts oberhalb des Alignments und speichert diese Datei für die weitere Bearbeitung unter dem Dateinamen `cytochromvergleich.fasta` lokal auf dem eigenen Rechner ab. Im Gegensatz zu dem über die grüne Bearbeitungsleiste erreichbaren Button Retrieve liefert diese Methode in den Fasta-Dateien auch die mit Minuszeichen entwerteten Leerstellen mit, falls sich die Kettenlängen unterscheiden und es daher an manchen Stellen in der anderen Kette keine entsprechenden Aminosäuren gibt. Diese „Gaps“ (= Lücken) sollten aber nicht zu viele Bausteine umfassen.

1.1.4 Druckausgabe für wissenschaftlichen Artikel mit dem LaTeX Paket Texshade erzeugen

Nun legt man im Programm Texmaker die folgende Datei `alignment_cytochrome.tex` an:

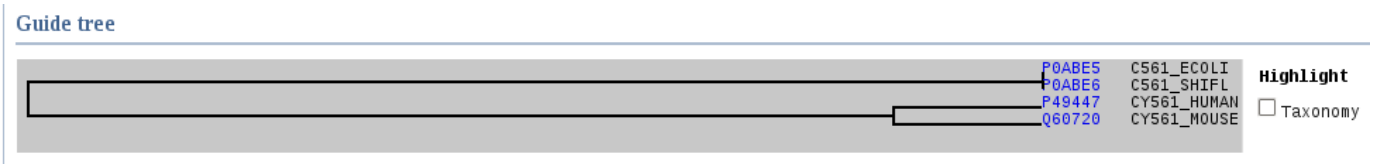
```
\documentclass[10pt,a4paper]{article}
\usepackage[utf8]{inputenc}
\usepackage{amsmath}
\usepackage{amsfonts}
\usepackage{amssymb}
%TexShade Paket benutzen
\usepackage{texshade}
\begin{document}
%vorher unter 1.1.3 gespeicherte Datei cytochromvergleich.fasta benutzen
\begin{texshade}{cytochromvergleich.fasta}
%Zeilenumbruch nach 60 Aminsäuren pro Zeile
\residuesperline*{60}
%Einfärbemodus nach Ähnlichkeit/Similarity
\shadingmode[allmatchspecial]{similar}
%Bereichsauswahl von der 1. bis maximal 251. Aminosäure
\setends{1}{1..251}
\hideconsensus
\end{texshade}
\end{document}
```

Es wird eine PDF-Datei mit dem folgenden Inhalt erzeugt:



1.1.5 Baumansicht online und mit lokalem Programm erzeugen

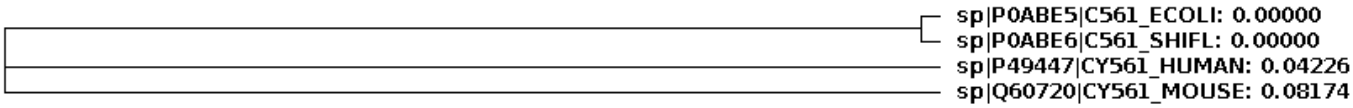
Befindet man wie in Kapitel 1.1.2 im Tab: Align , so wird unten im Abschnitt Tree ein einfacher abgeleiteter Stammbaum für das zuvor durchgeführte Alignment angezeigt:



Weg 1 mit online-ClustalW-Werkzeug:
Die unter 1.1.3 gespeicherte Datei `cytochromvergleich.fasta` kann über den Button Durchsuchen auf der Webseite www.ebi.ac.uk/Tools/phylogeny/clustalw2_phylogeny hochgeladen werden. Diese bietet die online-Nutzung des im Weg 2 lokal installierten Programms Clustalw/x. Die Berechnung wird mit Submit gestartet. Auf der Ergebnisseite wird im Absatz Phylogenetic Tree der Quelltext des Baumes angezeigt und darunter im Absatz Phylogram/Cladogram im Wechsel die beiden Ansichtsoptionen Cladogram Tree und Phylogram Tree.

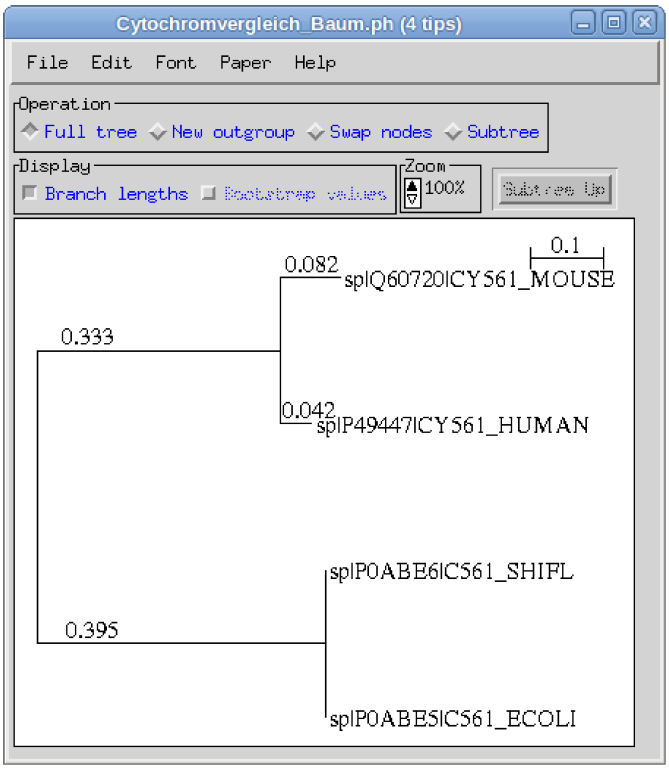
Phylogramm:

Cladogramm:



Möchte man nur die Daten des Baums speichern, wählt man die Optione View Phylogentic Tree File und speichert die dann im Browser angezeigte PHYLIP FORMAT TREE-Datei als `cytochromvergleich_baum.ph` lokal ab.

Weg 2 mit lokal installiertem ClustalX-Programm:
Man kann die Baumdatei alternativ auch mit dem lokal installierten Programm Clustalx erzeugen, indem man unter File > Load Sequences die unter 1.1.3 gespeicherte Datei `cytochromvergleich.fasta` lädt und unter Trees > Draw Tree den Baum als `cytochromvergleich_baum.ph` lokal abspeichert.



Egal auf welchem Weg sie erzeugt wurde, öffnet man anschließend diese Datei mit mit dem lokal installierten Programm NJPlot und erzeugt den folgende Baum (rooted = mit Wurzel):

Unter Display > Branch lengths werden die Abstände auf den Ästen angezeigt.
Unter File > Save as Postscript kann dann eine Postscript Datei mit der Endung `ps` erzeugt werden, die dann entweder direkt ausgedruckt oder nach einem Import mit einem Bildbearbeitungsprogramm wie Gimp weiterverarbeitet werden kann.

Alternativen bieten die Programme Treeview (taxonomy.zoology.gla.ac.uk/rod/treeview.html), TreeViewX (code.google.com/p/treeviewx) und Tree-Puzzle (www.tree-puzzle.de)

Mit dem Programm CDTree (nur Windows und Mac, www.ncbi.nlm.nih.gov/Structure/cdtree/cdtree.shtml) können auch weiterführende Analysen durchgeführt werden.

Tipp: Einen interaktiven Überblick über aktuelle Verwandtschaftsbeziehungen aufgrund molekularer Abstände bietet die Webseite *Interactive Tree of Life* (itol.embl.de).

1.2 Aminsäure-Sequenzvergleich am Beispiel Fasciculin (Schlangengift)

1.2.1 Suche in der UniProt-Proteindatenbank

www.uniprot.org > Search > Fasciculin

	Entry	Entry name	Status	Protein names	Gene names	Organism	Length
☑	P0C1Y9	TXFA1_DENAN	★	**Fasciculin-1**		Dendroaspis angusticeps (Eastern green mamba)	61
☑	P01391	NXL1_NAJKA	★	**Long neurotoxin 1**		Naja kaouthia (Monocled cobra) (Naja siamensis)	71
☑	P60615	NXL1A_BUNMU	★	**Alpha-bungarotoxin isoform A31**		Bungarus multicinctus (Many-banded krait)	95

Folgt man einzelnen Einträgen auf ihre Seiten, z.B. www.uniprot.org/uniprot/P0C1Y9
so findet man dort im Absatz General annotation (Comments) im Abschnitt Sequence similarities den Hinweis:
Belongs to the snake three-finger toxin family. Acn-esterase inhibitor subfamily.

Hinweis: Bei der Suche zu Fasciculin finden wir in der Ergebnistabelle neben den drei oben verglichenen Schlangengiften auch die menschliche Acetylcholinesterase:

	Entry	Entry name	Status	Protein names	Gene names	Organism	Length
☐	P22303	ACES_HUMAN	★	**Acetylcholin-esterase**	**ACHE**	Homo sapiens (Human)	614

1.2.2 Alignment direkt im Webbrowser durchführen

Anschließend führt man ein Align aus, indem man die Einträge der drei Entry name Felder wie folgt im Tab Align in das entsprechende Eingabefeld Sequences (in FASTA format) or UniProt identifiers einträgt:

TXFA1_DENAN
NXL1_NAJKA
NXL1A_BUNMU

Nach der Auswahl von Align, erhält man die folgende Anordnung:

```
1    --------------------TMCYSHTTT-SRAILTNCGENSCYRKSRRH------PPK   32  P0C1Y9   TXFA1_DENAN
1    --------------------IRCF--ITPDITSKDCP-NGHVCYTKTWCDAFCSIRGKR   36  P01391   NXL1_NAJKA
1    MKTLLLTLVVVTIVCLDLGYTIVCHTTATSPISAVTCPPGENLCYRKMWCDAFCSSRGKV   60  P60615   NXL1A_BUNMU
                             *.   *    :       . ** *   .

33   MVLGRGCGCPPGDDYLEVKCCTSPDKCNY-------   61  P0C1Y9   TXFA1_DENAN
37   VDLGCAATCPTVKTGVDIQCCST-DNCNPFPTRKRP   71  P01391   NXL1_NAJKA
61   VELGCAATCPSKKPYEEVTCCST-DKCNPHPKQRPG   95  P60615   NXL1A_BUNMU
     : ** .. **  .   :: **:: *:**
```

Ein Sternchen „*“ bedeutet völlige Übereinstimmung, ein Doppelpunkt „:“ sehr ähnlich, ein einfacher Punkt „.“ ähnlich. Wählt man in der rechten Spalte unter Amino acid properties > Similarity aus, so werden bei allen verglichenen Proteinen konservierte Bereiche hellgrau unterlegt.
Wählt man in der rechten Spalte unter Annotation > Disulfide bond aus, wird die Aminosäure Cystein hellblau unterlegt.

Beobachtung:

- Man erkennt hier eindeutig, dass die Aminosäuresequenz v.a. an den Cysteinresten sehr konservativ ist.

Interpetation:

- Die Disulfidbrücken sind entscheidend für den übergeordneten räumlichen Bau der Tertiärstruktur und somit für die Funktionsweise dieser Toxine.

1.2.3 Abbildung der 3D-Ansicht eines bestimmten Proteins mit bestimmten Einstellungen

Nun folgt man auf der Suchergebnisseite um Abschnitt Results dem Link in der Spalte Entry für das Protein mit dem Namen Fasciculin-1, hier also www.uniprot.org/uniprot/P0C1Y9
Im Absatz Cross-references wählt man im Abschnitt 3D structure databases unter PDBe den Entry 1FAS aus, der auf die Seite www.ebi.ac.uk/pdbe-srv/view/entry/1FAS führt.

Dort kann man im linken Seitenmenü unter Visualisation die Ansicht in Jmol auswählen. Auf der neuen Seite www.ebi.ac.uk/pdbe-srv/view/entry/1fas/jmol wählt man rechts unter Rendering > Rockets aus. Nun kann man durch einen Rechtsklick auf das Jmol-Applet das Kontextmenü aufrufen und z.B. unter Render > Disulfide bonds > On die Disulfidbrücken sichtbar machen.

Kennt man die Structure-ID für die RCSB Protein Data Bank (hier also „1FAS“), so kann man alternativ auch unter www.rcsb.org/pdb/ > Search: 1FAS suchen.
Auf der Ergebnisseite www.rcsb.org/pdb/explore/explore.do?structureId=1fas kann man unter dem Eintrag View in Jmol wieder eine 3D-Ansicht erhalten. Hier kann zusätzlich unter Datei > Exportiere PNG Abbild die aktuelle Ansicht als Grafikdatei lokal gespeichert werden.

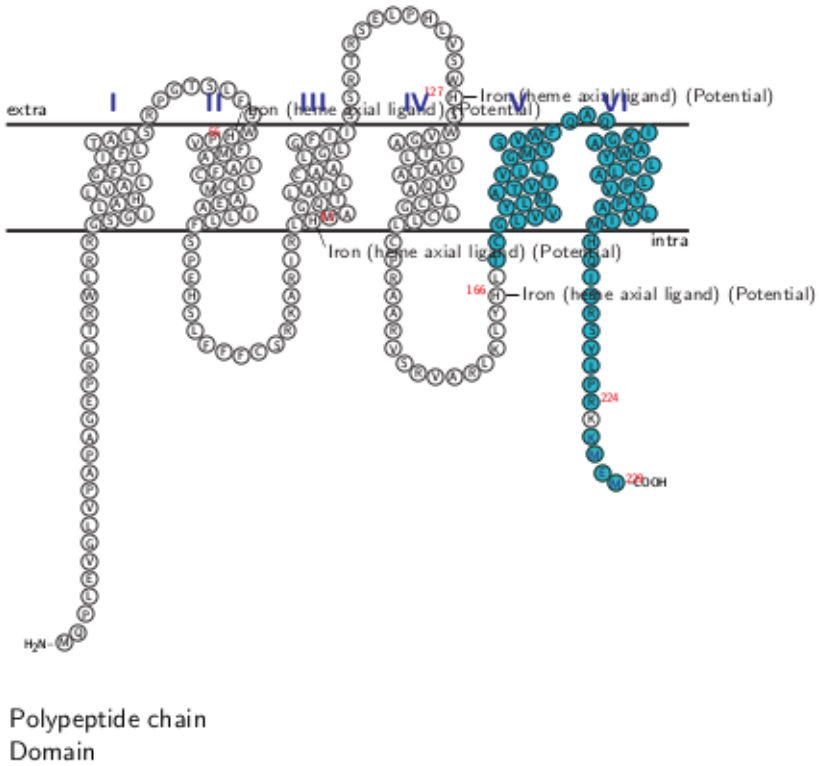
1.2.4 Membran-Diagramme mit TexTopo am Beispiel Cytochrome b561 domain-containing protein 1 des Menschen

Auf der Seite www.uniprot.org/uniprot/Q8N8Q1 wählt man rechts oben den orangefarbenen Knopf **text** und speichert die Datei als `Q8N8Q1.SP`

Nun legt man im Programm Texmaker die folgende Datei `diagramm_Q8N8Q1.tex` an:

```
\documentclass[10pt,a4paper]{article}
\usepackage[latin1]{inputenc}
\usepackage{amsmath}
\usepackage{amsfonts}
\usepackage{amssymb}
%TexTopo Paket benutzen
\usepackage{textopo}
\begin{document}
\begin{textopo}
%die unter 1.3.4 gespeicherte Datei Q8N8Q1.SP benutzen
\getsequence[make new]{SwissProt}{Q8N8Q1.SP}
\end{textopo}
\end{document}
```

Im Erfolgsfall wird die folgende Grafik in einer PDF-Datei erzeugt:



Listen geeigneter Membranproteine bieten die Seiten MPtopo (blanco.biomol.uci.edu/mptopo) und Stephen White's Database (blanco.biomol.uci.edu/mpstruc/listAll/list).

1.3 Übersicht zu Aminosäuresequenz- und Proteindatenbanken

UniProt

- Kontrollierte Einträge v.a. von Aminosäuresequenzen von hoher Qualität von ca. 100 000 Proteinen
- Empfohlene Einstiegsseite, da man über die Cross-references schnell zu den ausführlicheren Datenbanken gelangt
- www.uniprot.org

Protein Data Bank (RSCB, PDB ⇒ auch als Dateiformat!**)**

- Datenbank für ca. 18000 experimentell ermittelte Proteinstrukturen
- www.rcsb.org/pdb

1.4 Übersicht zu Dateiformaten für Proteine

FASTA

Enthält die Sequenzinformation für Aminosäure- oder Nukleotidsequenzen.
Wird von fast allen Programmen verstanden und erzeugt.
Kann mit jedem einfachen Texteditor geöffnet, kopiert und verändert werden.
Beginnt mit einem größer als Zeichen „>“, gefolgt von einer Kennzeichnung des Typs als Protein/Gen/Alignment in der ersten Zeile: P... ⇒ Protein; N...⇒ Nukleinsäure
Dann folgt ein Strichpunkt „;“ und anschließend ein eindeutiger Bezeichner.
Jetzt ist noch eine Kommentarzeile möglich, die auf beiden Seiten mit je einem Pipe-Symbol „|“ eingeschlossen wird.
In der folgenden Zeile folgt dann die eigentliche (möglicherweise sehr lange) Sequenz im reinen Textformat. Es werden immer Großbuchstaben empfohlen. Die Sequenz kann (muss aber nicht) mit einem Sternchen „*“ abgeschlossen werden. Zusätzliche Zeilenumbrüche oder Leerzeichen werden beim Einlesen ignoriert, so dass sie im Bereich der Sequenz für eine bessere Druckdarstellung beliebig eingefügt werden können, ohne dass sich der Informationsgehalt ändert.

Ein formal korrektes Beispiel einer FASTA-Datei für eine Aminosäuresequenz würde also lauten:

```
>P1;MEINE_KENNZEICHNUNG
|Mein Kommentartext|
AG*
```

(entsprechend hier den Aminosäuren Alanin und Glycin im Einbuchstabencode)

PDB (Format der Protein Data Bank)

Enthält die (u.a. räumlichen) Strukturinformationen von Proteinen.
Eine typische Zeile hat den folgenden Aufbau:

Anzahl	Element	Molekül	Kette	x	y	z	Koordinate
ATOM 1	N	VAL	e	16 29.582	19.112	38.968	

SP (SWISSPROT; UNIPROT)

Gekürztes Beispiel:

```
ID   C56D1_HUMAN             Reviewed;         229 AA.
AC   Q8N8Q1; B4DH97; Q52M36; Q5T6C2;
DT   10-MAY-2005, integrated into UniProtKB/Swiss-Prot.
[...]
DE   RecName: Full=Cytochrome b561 domain-containing protein 1;
[...]
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
[...]
FT   CHAIN         1    229       Cytochrome b561 domain-containing protein
FT                                1.
FT                                /FTId=PRO_0000151034.
FT   TOPO_DOM      1     24       Cytoplasmic (Potential).
FT   TRANSMEM     25     45       Helical; (Potential).
[...]
FT   DOMAIN       22    224       Cytochrome b561.
FT   METAL        55     55       Iron (heme axial ligand) (Potential).
[...]
SQ   SEQUENCE   229 AA;  25424 MW;  43978DAF7D8EC218 CRC64;
     MQPLEVGLVP APAGEPRLTR WLRRGSGILA HLVALGFTIF LTALSRPGTS LFSWHPVFMA
     LAFCLCMAEA ILLFSPEHSL FFFCSRKARI RLHWAGQTLA ILCAALGLGF IISSRTRSEL
     PHLVSWHSWV GALTLLATAV QALCGLCLLC PRAARVSRVA RLKLYHLTCG LVVYLMATVT
     VLLGMYSVWF QAQIKGAAWY LCLALPVYPA LVIMHQISRS YLPRKKMEM
//
```

Die Datenfelder werden durch einen Zweibuchstabencode eingeleitet, nach dem auch bei der Suche gefiltert wird. Über den Schlüssel DE wird z.B. der Name des Proteins gesucht.
Die Schlüssel 0S und 0C liefern die systematische Einordnung der Spezies, zu welcher das Protein gehört. Der Schlüssel FT steht für besondere Bereiche mit bestimmten Eigenschaften, z.B. Stellung innerhalb oder an einer Membran, Bindungsstellen für Metallatome/-ionen oder typische Domänen einer Proteinfamilie. Der Schlüssel SQ beinhaltete dann die eigentlichen Sequenzdaten.

1.5 Übungen

1.5.1 Pre-B-cell Colony- Enhancing Factor (PBEF)

a) Holen der Sequenzdaten

Wählen Sie auf der Seite www.ncbi.nlm.nih.gov/sites/entrez?db=unigene die Option Protein und suchen Sie nach dem Stichwort *human PBEF* . In den Suchergebnissen sollte der folgende Eintrag erscheinen:

nicotinamide phosphoribosyltransferase precursor [Homo sapiens]
491 aa protein
Accession: NP_005737.1
GI: 5031977

Speichern Sie über den Link FASTA die Aminosäuresequenz ab unter dem Namen
`NP_005737.1_mensch.fasta`

Suchen und speichern Sie anschließend die entsprechenden Sequenzen für die Ratte (Rattus norvegicus) [NP_808789.1_ratte.fasta] und die Maus (Mus musculus) [NP_067499.2_maus.fasta].

Damit in den Alignments nicht die komplizierten Accession-Numbers und Kurzbezeichnungen erscheinen, sondern einfach die Art, wird in jeder Fasta-Datei die erste Zeile z.B. von

```
>gi|5031977|ref|NP_005737.1| nicotinamide phosphoribosyltransferase precursor [Homo sapiens]
```

geändert in:

```
>gi|Mensch
```

Dies wird entsprechend auch in den Fasta-Dateien für Ratte und Maus durchgeführt.

b) Durchführen des Alignments online
Öffnen Sie anschließend die Seite www.ebi.ac.uk/Tools/msa/clustalw2 und öffnen Sie die drei Fasta-Dateien aus dem letzten Arbeitsschritt mit einem einfachen Texteditor. Kopieren Sie den Inhalt dieser Drei Dateien in das Feld STEP 1 - Enter your input sequences und wählen Sie unter STEP 3 - Set your Multiple Sequence Alignment Options unter More options... > OUTPUT Options > FORMAT: Pearson/FASTA.
Starten Sie anschließend das Alignment mit Submit.

Speichern Sie auf der Ergebnisseite das Alignment mit Download Alignment File als `mensch_ratte_maus.fasta`

c) Stammbaumanalyse online
Öffnen Sie die Seite www.ebi.ac.uk/Tools/phylogeny/clustalw2_phylogeny und laden Sie im Feld STEP 1 - Enter your input sequences unter Durchsuchen die gerade gespeicherte Datei `mensch_ratte_maus.fasta` hoch und erzeugen Sie mit Submit einen Baum. Vergleichen Sie wieder die Ansichten als Phylogram und Cladogram.

d) Stammbaumanalyse offline
Speichern Sie das in der letzten Aufgabe erzeugte Phylogenetic Tree File als `stammbaum_PBEF.ph` und öffnen Sie diese anschließend mit dem Programm NJplot. Machen Sie unter Display > Branch lenghts die ermittelten Abstände sichtbar und speichern Sie eine A4-Druckversion unter File > Save as Postscript. Die Dateiendung sollte .ps lauten. Diese Datei kann dann mit einem PDF-Betrachter angesehen oder in ein Bildbearbeitungsprogramm importiert werden.

e) Alternatives online-Alignment mit anschließender Stammbaumanalyse am NCBI [entsprechend c) & d)]
Suchen Sie auf der Seite www.ncbi.nlm.nih.gov/sites/entrez?db=unigene nach dem Stichwort dog *PBEF* .
In den Suchergebnissen sollte der folgende Eintrag erscheinen:

Nicotinamide phosphoribosyltransferase
NAMPT, Canis lupus familiaris
Cfa.39714: 9 sequences.

Folgen Sie dem Link zu den Nicotinamide phosphoribosyltransferase (NAMPT)-ähnlichen Proteinen

Folgen Sie im Kontextmenü des Eintrags
XP_540386.2 PREDICTED: similar to Nicotinamide phosphoribosyltransferase (NAmPRTase) (Nampt) (Pre-B-cell colony-enhancing factor 1 homolog) (PBEF)
der Auswahl Protein sequence und in der erscheinenden Seite dem Link FASTA und speichern Sie die Aminosäure-Sequenz ab unter `XP_540386.2_hund.fasta`

Damit in den Alignments nicht die komplizierte Accession-Number und Kurzbezeichnung erscheint, sondern einfach die Art, wird auch in dieser Fasta-Datei die erste Zeile z.B. von

```
>gi|73981945|ref|XP_540386.2| PREDICTED: similar to Nicotinamide phosphoribosyltransferase
(NAmPRTase) (Nampt) (Pre-B-cell colony-enhancing factor 1 homolog) (PBEF) [Canis familiaris]
```

geändert in:

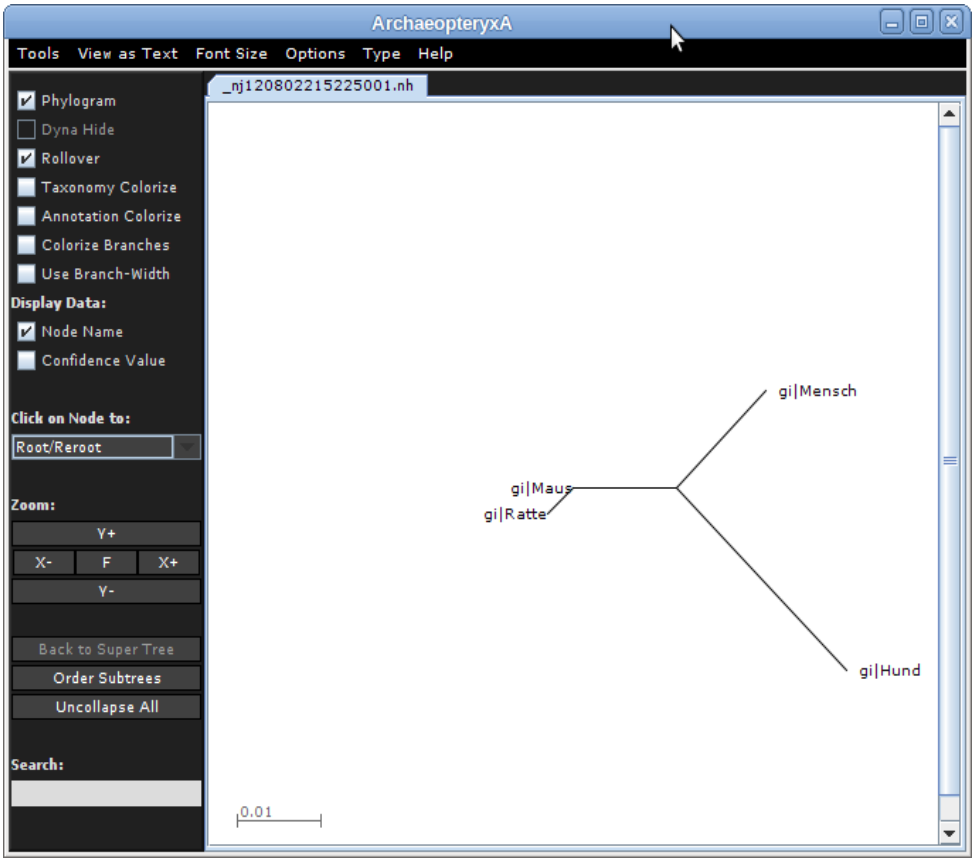
```
>gi|Hund
```

Öffnen Sie die Seite mafft.cbrc.jp/alignment/server und kopieren Sie in das Feld Input nacheinander den Inhalt der vier Dateien:

```
NP_005737.1_mensch.fasta
NP_808789.1_ratte.fasta
NP_067499.2_maus.fasta
XP_540386.2_hund.fasta
```

erzeugen Sie mit Submit zunächst das Alignment für alle vier Arten.
Wählen Sie auf der Ergebnisseite den Button Phylogenetic Tree und dann Go!
(Bei nur drei Arten im Alignment wird eine Fehlermeldung ausgegeben.)

Wählen Sie unter Type > unrooted (alpha) und betrachten Sie den folgenden (unrooted = ohne Wurzel) Stammbaum:



Falls man bereits ein Alignment mit mehr als drei verschiedenen Arten/Proteinen als Fasta-Datei hat, kann man diese auch gleich unter mafft.cbrc.jp/alignment/server/phylogeny.html zur Erzeugung des Baumes benutzen.

Eine Kurzübersicht bietet auch die Seite www.cgae.de/alignment/index.html.

f) Druckausgabe des Alignments erzeugen (nur für Fortgeschrittene!)
Versuchen Sie, wie im Kapitel 1.1.4 beschrieben, eine formatierte Ausgabe des erzeugten Alignments als PDF-Datei zu erstellen. Legen Sie dazu zunächst im Speicherort der Datei `mensch_ratte_maus.fasta` die folgende Datei `Alignment_PBEF.tex` an:

```
\documentclass[10pt,a4paper]{article}
\usepackage[utf8]{inputenc}
\usepackage{amsmath}
\usepackage{amsfonts}
\usepackage{amssymb}
%TexShade Paket benutzen
\usepackage{texshade}
\begin{document}
%vorher unter 1.1.3 gespeicherte Datei mensch_ratet_maus.fasta benutzen
\begin{texshade}{mensch_ratte_maus.fasta}
%Zeilenumbruch nach 60 Aminsäuren pro Zeile
\residuesperline*{60}
%Einfärbemodus nach Ähnlichkeit/Similarity
\shadingmode[allmatchspecial]{similar}
%Bereichsauswahl von der 1. bis maximal 251. Aminosäure
\setends{1}{1..251}
\hideconsensus
\end{texshade}
\end{document}
```

1.5.2 Hämoglobin

Erstellen Sie ein globales Alignment der Aminosäuresequenz der α- und β- Kette des menschlichen Hämoglobins und protokollieren Sie stichpunktartig die dazu notwendigen Schritte.
Speichern Sie die entsprechenden Aminosäure-Sequenz-Dateien zwischendurch im FASTA-Format lokal ab.

Tipp: Im Kapitel 5 Workflow ist ein allgemeiner Arbeitsablauf skizziert. Benutzen Sie Werkzeuge Ihrer Wahl, also je nach Systemvoraussetzungen online- oder auch lokal installierte offline-Programme.

Zusatzübungen:

- Erzeugen Sie in einem Jmol-Applet drei PNG-Dateien der 3D-Ansicht der beiden verschiedenen Ketten einzeln sowie eines komplettes Hämoglobin-Moleküls.

- Erzeugen Sie mit Hilfe einer angepassten LaTeX-Datei mit Texmaker eine PDF-Datei mit dem Alignment der beiden Ketten.

- Folgen Sie auf der Uniprot-Seite der α-Kette **www.uniprot.org/uniprot/P69905** im Abschnitt Natural variations z.B. der in Quong Sze gefundenen Punktmutation an der Position **126**, bei der die Aminosäure Leucin (L) durch die Aminosäure Prolin (P) ersetzt wird. Die veränderte Aminosäure ist farbig hervorgehoben.

- Im Absatz Cross-references wird nun unter Sequence databases GenBank ausgewählt und dann gleich der erste Eintrag **J00153**

- Über den Eintrag Related Information > Related Sequences im rechten Seitenmenü werden dann ähnliche Sequenzen gefunden. Unter dem Stichwort thalassemia (**Thalassämie**) findet man dort auch Referenzen zu Erbkrankheiten.

2 Nucleotid-Sequenzvergleiche

Aminosäure-Sequenzen bestehen meist aus etwa 100 bis 500 Bausteinen. Sie haben klar definierte Begrenzungen und spezifische Funktionseigenschaften. Die dazugehörigen Nukleotid-Sequenzen hingegen variieren viel stärker hinsichtlich ihrer Länge sowie ihrer Verteilung im Genom.

Auf der Nukleinsäure-Ebene enthält ein Gen also neben den Protein codierenden Abschnitten zusätzlich (1) Regulatorische Regionen und (2) Bereiche in denen zwar eine Transkription aber keine Translation mehr stattfindet. Bei Eukaryoten sind die Protein codierenden Abschnitte zusätzlich noch aufgeteilt in eine unterschiedliche Anzahl von (3) Exons, also Bereichen, welche zum späteren Protein beitragen sowie (4) Introns, die vorher beim Spleißen ausgeschnitten werden.

Nukleotid-Sequenzen können sich daher beziehen auf entweder
(1) primäre Transkripte vor dem Speißen, die noch alle Introns und Exons enthalten,
(2) den Abschnitt entsprechend der reifen mRNA, welche nur noch die Exons enthält,
(3) den offenen Leserahmen (OLR) zwischen einem Start- und einem Stop-Codon, der sowohl Introns als auch Bereiche enthält, die zwar transkribiert aber nicht translatiert werden, oder auf
(4) eine Vielzahl von möglichen Teilsequenzen.

Hieraus ergeben sich zwei wesentliche Folgen für den Arbeitsablauf:

1. Es wird soweit möglich zuerst das Protein gesucht und erst dann werden über Querverweise (Cross-references) die Nukleotid-Sequenzen aufgerufen.
2. Bei Proteinen ist häufig ein globales Alignment (über die gesamte Länge) sinnvoll, während bei Nukleotid-Sequenzen meist nur lokale Alignments (mit kürzeren Bruchstücken) sinnvoll sind.

2.1 dUTP pyrophosphatase von E. coli

2.1.1 Suche in der UniProt-Proteindatenbank

www.uniprot.org > Tab: Search > Suchbegriff: dUTP pyrophosphatase

	Entry	Entry name	Status	Protein names	Gene names	Organism	Length
☑	P06968	DUT_ECOLI	★	**Deoxyuridine 5'-triphosphate nucleotidohydrol..** .	**dut** dnaS sof b3640 JW3615	Escherichia coli (strain K12)	151

Nun folgt man dem Eintrag auf die Seite www.uniprot.org/uniprot/P06968
und scrollt auf der Seite nach unten bis in den Absatz Cross-references.

Im Abschnitt Sequence databases wählt man die Option GenBank aus und folgt dem GenBank Eintrag mit der Accession Number X01714 auf die Seite www.ncbi.nlm.nih.gov/nuccore/X01714

Über den Link FASTA kann wird nun die Fasta-Nukleotid-Sequenzdatei `X01714_ecoli.fasta` gespeichert.

2.1.2 Nukleotid-Sequenzen im FASTA-Format herunterladen

Nun werden die unter 2.1.1 beschriebenen Schritte immer mit dem ersten GenBank Eintrag für die folgenden Arten wiederholt, so dass man am Ende insgesamt sechs verschiedene Nukleotid-Sequenz-Fasta-Dateien erhält:

UniProt Entry name	Art	NCBI Accession Number	Fasta-Dateiname
DUT_AQUAE	Aquifex aeolicus (strain VF5) (thermophiles Bakterium)	AE000657 komplettes Genom!	`AE000657.1_aquifex.fasta`
DUT_BRAJA	Bradyrhizobium japonicum	AF042096.3 mehrere Gene	`AF042096.3_bradyrhizobium.fasta`
DUT_BPT5	Bacteriophage T5	AY543070.1 komplettes Genom!	`AY543070.1_bacteriophaget5.fasta`
DUT_CANAL	Candida albicans SC5314 (Hefepilz)	AACQ01000045.1 komplettes Genom!	`AACQ01000045.1_candida.fasta`
DUT_BUCAI	Buchnera aphidicola str. APS (Acyrthosiphon pisum)	BA000003.2 komplettes Genom!	`BA000003.2_buchnera.fasta`

Tipp: Längere Fasta-Sequenzen können unter Send > Choose Destination > File > Format: FASTA > Create File heruntergeladen werden, kürzere kann man auch einfach markieren und in ein leeres Textdokument in einem einfachen Texteditor wie Gedit oder Geany einfügen.

2.1.3 Nukleotid-Sequenzvergleich mit BLAST

Für den Sequenzvergleich wird hier eine online-Version von BLAST genutzt.
Auf der BLAST-Startseite des NCBI http://blast.ncbi.nlm.nih.gov/Blast.cgi findet man die folgende Option: nucleotide_blast, welche die Nukleotid-Datenbank nach ähnlichen Stellen zu einer gegebenen Nukleotid-Sequenzen durchsucht.
Dabei treten zwei Probleme auf:
(1) Werden homologe Sequenzen übersehen? Dies kann passieren, wenn man zu spezifisch sucht;
(2) Werden falsche Sequenzen mitgefunden? Dies kann passieren, wenn man zu unspezifisch sucht.
⇒ Es muss immer ein Kompromiss aus (1) Spezifität und (2) Sensitivität gefunden werden!

Fragestellung 1:
In welchen Organismen ist allgemein eine zum dUTP pyrophosphatase-Gen von E. coli ähnliche Nukleotid-Sequenz enthalten?

- In das obere Feld Enter Query Sequence wird nun zunächst die Accession Number des E. coli Gens kopiert:
  X01714
- Die Option Align two or more sequences wird NICHT ausgewählt.
- Im Abschnitt Choose Search Set wird unter Database Others (nr etc.): Nucleotide colelction (nr/nt) ausgewählt. Hinweis: nr steht für nonredundant, die Suche läuft also über alle bekannten Nukleotid-Sequenzen.
- Im Abschnitt Program Selection wird unter Optimize for ausgewählt Somewhar similar sequences (blastn)
- Nach Drücken des großen BLAST-Knopfes ganz unten links, kann es mehrere Minuten dauern, bis eine Ergebnisseite erscheint.

Die Interpretation ist hier leider nicht mehr so einfach wie bei den Aminosäure-Seuqenzvergleichen. Wichtig für eine gute Übereinstimmung sind allgemein ein hoher Wert in der Spalte Total Score sowie ein möglichst niedriger Wert in der Spalte E value (E = Expect-Wert, E ≦ 0,02 ⇒ vermutlich homolog, E ≧ 1 ⇒ die gefundene Übereinstimmung beruht vermutlich auf Zufall).

Identische Sequenzen sind natürlich zu 100% identisch und die Wahrscheinlichkeit beim Abgleich mit ganzen Genomen ist ebenfalls viel höher als mit einzelnen Gensequenzen.

Fragestellung 2:
Ist in einer der fünf anderen Gensequenzen aus Kapitel 2.1.2 eine dem dUTP pyrophosphatase-Gen von E. coli ähnliche Nukleotid-Sequenz enthalten?

- In das obere Feld Enter Query Sequence wird nun zunächst die Accession Number des E. coli Gens kopiert:
  X01714
- Die Option Align two or more sequences wird jetzt ausgewählt.
- Im erscheinenden Abschnitt Enter Subject Sequence wird bei jedem Testdurchlauf EINE der folgenden Accession-Numbers eingegeben:
  AE000657 *ODER* AF042096.3 *ODER* AY543070.1 *ODER* AACQ01000045.1 *ODER* BA000003.2
- Im Abschnitt Program Selection wird unter Optimize for ausgewählt More dissimilar sequences (discontiguous megablast)
- Nach Drücken des großen BLAST-Knopfes ganz unten links, kann es mehrere Minuten dauern, bis eine Ergebnisseite erscheint.

Im Erfolgsfall erscheint auf der Ergebnisseite im Abschnitt Graphic Summary unterhalb des oberen durchgezogenen roten Balkens, der die vorgegebene Abfragesequenz darstellt, eine oder viele Linien für gefundene Alignments.
Die Länge der Balken zeigt an, wie lange die gefundene Sequenz im Vergleich zur vorgegebenen Abfragesequenz ist. Die Farbe des Balkens zeigt den Grad der Ähnlichkeit an (Bei Nukleotidsequenzen in Prozent Identität, bei Aminosäuresequenzen in Prozent Ähnlichkeit/Similarity).

Im Abschnitt Alignments wird eine genauere Ansicht für jedes dieser gefundenen Alignments gegeben, etwa in der Form:

```
>dbj|BA000003.2| Download subject sequence BA000003 spanning the HSP Buchnera aphidicola
str. APS (Acyrthosiphon pisum) genomic DNA,
complete sequence
Length=640681

 Score = 68.0 bits (74),  Expect = 3e-12
 Identities = 81/110 (74%), Gaps = 0/110 (0%)
 Strand=Plus/Minus

Query  545     TGCCGCGCTCCGGATTGGGACATAAGCACGGTATCGTGCTTGGTAACCTGGTAGGATTGA  604
               ||||  | || ||| | || |||||  | ||||| || || |||||  | || || || |
Sbjct  593108  TGCCTAGGTCTGGACTAGGTCATAAAAAAGGTATTGTACTAGGTAATTTAGTTGGTTTAA  593049

Query  605     TCGATTCTGACTATCAGGGCCAGTTGATGATTTCCGTGTGGAACCGTGGT  654
               | ||||||||||||| || || |||||||| ||  | ||||| ||| ||
Sbjct  593048  TTGATTCTGACTATCAAGGTCAATTGATGATATCTCTCTGGAATCGTAGT  592999
```

2.2 Dateiformate für Nukleotidsequenzen

FASTA

Vgl. Kapitel 1.4
Enthält die Sequenziformation für Nukleotid- oder Aminosäuresequenzen.
Ein formal korrektes Beispiel einer FASTA-Datei für eine Nukleotidsequenz würde also lauten:

```
>N1;MEINE_KENNZEICHNUNG
|Mein Kommentartext|
AUG GCG UAG*
```

(entsprechend der mRNA für: Met/Start-Ala-Stop)

2.3 Übersicht zu Nukleotidsequenz- und Gendatenbanken

European Molecular Biology Laboratory (EMBL)

- Umfassende Datenbank mit mehr als $20x10^6$ Einträgen
- Die Inhalte werden mit der **DNA Database of Japan DDBJ** (www.ddbj.nig.ac.jp) und der **GenBank** des **National Center for Biotechnology Information NCBI** (www.ncbi.nlm.nih.gov/gene) abgeglichen
- www.ebi.ac.uk/embl

Nucleotide Database am National Center for Biotechnology Information

- Für kürzere Nukleotid-Sequenzen
- www.ncbi.nlm.nih.gov/nuccore

2.4 Übungen

Tipp: Alternativ kann neben der NCBI-Seite blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn auch das BLAST-online Werkzeug der Seite www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html genutzt werden.

2.4.1 Zu welchem Enzym gehört ein bestimmtes DNA-Fragment?

Gegeben ist das folgende Fragement einer DNA-Sequenz aus einem wichtigen Enzym des Menschen

```
>N_Fragment1
TGTCGCACAG ATGACCACGT GGTTAGTGGC AACCTGGTGA CCCCCCTTCC TGTGATTTTA
```

```
>N_Fragment2
GGTGATAAAG TCATCCCGCT CTTTACTCCT CAGTGTGGAA AATGCAGAGT TTGTAAAAAC
CCGGAGAGCA ACTACTGCTT GAAAAATGAT CTAGGCAATC CTCGGGGGAC CCTGCAGGAT
```

Ermitteln Sie für beide DNA-Fragmente jeweils getrennt voneinander und jeweils mit Hilfe der beiden angegebenen BLAST-Seiten, zu welchem Enzym des Menschen dieses DNA-Fragement vermutlich gehört.

2.4.2 Homologe Gene, die in Eukaryoten konserviert sind

Vergleichen Sie auf der Seite www.ncbi.nlm.nih.gov/sites/homologene/37552 die im Absatz Protein-Alignments im Kasten Pairwise alignments generated uning BLAST das Protein-Alignment zwischen NP_001017916.1 (H. sapiens) und NP_031831.2 (M. musculus).

```
>ref|NP_031831.2| cytochrome b561 [Mus musculus]
 sp|Q60720.2|CY561_MOUSE RecName: Full=Cytochrome b561; AltName: Full=Cytochrome b-561
 dbj|BAB29321.1| unnamed protein product [Mus musculus]
 6 more sequence titles
 Length=250

 Score =  387 bits (993),  Expect = 1e-140, Method: Compositional matrix adjust.
 Identities = 210/233 (90%), Positives = 219/233 (94%), Gaps = 0/233 (0%)

Query  19   FSQLLGLTLVAMTGAWLGLYRGGIAWESDLQFNAHPLCMVIGLIFLQGNALLVYRVFRNE  78
            FSQLLGLT+VA+TGAWLGLYRGGIAWES LQFN HPLCMVIG+IFLQG+ALLVYRVFR E
Sbjct  18   FSQLLGLTVVAVTGAWLGLYRGGIAWESSLQFNVHPLCMVIGMIFLQGDALLVYRVFRRE  77

Query  79   AKRTTKVLHGLLHIFalvialvglvavFDYHRKKGYADLYSLHSWCGILVFVLYFVQWLV  138
            AKRTTK+LHGLLH+FA +IALVGLVAVFDYH+KKGYADLYSLHSWCGILVFVLYFVQWLV
Sbjct  78   AKRTTKILHGLLHVFAFIIALVGLVAVFDYHKKKGYADLYSLHSWCGILVFVLYFVQWLV  137

Query  139  GFSFFLFPGASFSLRSRYRPQHIFFGATIFLLSVGTALLGLKEALLFNLGGKYSAFEPEg  198
            GFSFFLFPGASFSLRSRYRPQHIFFGATIFL SVGTALLGLKEALLF LG KYS FEPEG
Sbjct  138  GFSFFLFPGASFSLRSRYRPQHIFFGATIFLFSVGTALLGLKEALLFKLGSKYSTFEPEG  197

Query  199  vlanvlgllaCFGGAVLYILTRADWKRPSQAEEQALSMDFKTLTEGDSPGSQ  251
            VLANVLGLLL CFG  VLYIL +ADWKRPSQAEEQALSMDFKTLTEGDSP  Q
Sbjct  198  VLANVLGLLLVCFGVVVLYILAQADWKRPSQAEEQALSMDFKTLTEGDSPSPQ  250
```

Begründen Sie, woran man erkennen kann, dass in diesem Fall auch ein globales Alignment gut geeignet gewesen wäre!

Führen Sie wie in Kapitel 1 beschrieben auf der Seite www.uniprot.org ein solches globales Alignment mit www.ncbi.nlm.nih.gov/protein/63054830?report=fasta und www.ncbi.nlm.nih.gov/protein/31542436?report=fasta durch und vergleichen Sie das dortige Ergebnis mit dem oben angegebenen Ergebnis der BLAST-Abfrage.

3 Fachbegriffe und Abkürzungen

Globales Alignment
Es wird die Sequenz der Bausteine eines polymeren Biomoleküls über seine volle Länge hinweg verglichen.
Bei unterschiedlicher Kettenlänge werden in der kürzeren Kette die jeweils fehlenden Glieder als übersprungen gekennzeichet (z.B. durch einen Punkt oder ein Minuszeichen).
z.B. eine komplette Polypeptidkette eines Proteins einer Art mit dem entsprechenden Protein einer anderen Art order zwei verwandte Polypeptidketten nur einer Art.
z.B. der Exon-Bereich eines Gens
z.B. eine mRNA

Beispiel für ein globales (paarweises) Alignment:

```
KIEGKNVF**YWFQNHKAR**ERQKKR--
|:.||  |||:..|||...|..:
KLAGK--FYWYNKHKA**YWFQNKHAR**
```

Ein Pipe-Symbol | (oder ein Sternchen *) bedeutet völlige Übereinstimmung, ein Doppelpunkt : sehr ähnlich, ein einfacher Punkt . ähnlich. Aminosäuren, die nur in einer Kette vorkommen, werden mit einem Bindestrich - gekennzeichnet.
Die fettgedruckten Aminosäuren werden unten im Beispiel für ein lokales Alignment verwendet.

Beispiel-Programm: "GAP" (Needleman & Wunsch)

Lokales Alignment
Es werden nur kurze Abschnitte einer bestimmten Länge mit einer bestimmten Sequenz gesucht. Möglicherweise können diese auch mehrfach vorkommen. Durch Insertion oder andere Genommutationen können sie auch unterschiedlichen Stellen innerhalb eines Gens vorkommen.

Beispiel für ein lokales (paarweises) Alignment:

```
YWFQNHKAR
|||||||||
YWFQNHKAR
```

Paarweises Alignment
Es werden nur zwei Sequenzen miteinander verglichen (z.B. gleiches Protein zweier Arten vgl. Kapitel 1.1 und 1.2 oder unterschiedliches Protein einer Art, vgl. Kapitel 1.5 Übung 1.5.1).

Multiples Alignment
Es werden mehrere Sequenzen miteinander verglichen (z.B. entsprechende Proteine bei mehreren Arten).

FASTA
Dateiformat für Aminosäure- oder Nukleotid-Sequenzdaten einzelner Polypeptidketten oder DNA- oder mRNA-Sequenzen oder auch paarweiser oder multipler Alignments; bezeichnet gleichzeitig auch den Algorithmus für globale Alignments

FastA (Fast Alignment, Pearson u. Lipmann 1998)

GDE format
genau wie FASTA, aber beginnend mit einem Prozentzeichen "%" an Stelle eines größer als Zeichens ">"

BLAST (Basic Local Alignment Search Tool, Altschul et al. 1990)

Weit verbreiteter Algorithmus (und Software) für schnelle und trotzdem zuverlässige lokale Alignments

Multiple Sequence Format (MSF)
Ausgabeformat für Alignments, die mit Hilfe des Programms PHYLIP (phylogenetic inteference package, evolution.genetics.washington.edu/phylip.html) erzeugt wurden.

ALN format
Ausgabeformat für Alingments, die mit Hilfe des Programms CLUSTALW/X erzeugt wurden.

PHYLIP FORMAT TREE (Phylogenetic Tree File, Dateiendung: .ph)
z.B. die Datei `cytochromvergleich_baum.ph` aus Kapitel 1.1.5

```
(
(
sp|P0ABE5|C561_ECOLI:0.00000,
sp|P0ABE6|C561_SHIFL:0.00000)
:0.72747,
sp|P49447|CY561_HUMAN:0.04226,
sp|Q60720|CY561_MOUSE:0.08174);
```

Ähnlichkeit (Similarity)
Maß für die Übereinstimmung verschiedender Aminosäure-Sequenzen. Hierbei werden neben den Zahlenverhältnissen v.a. die Eigenschaften der verschiedenen Aminosäuren (z.B. Hydrophilie, Polarität, pKS-Wert und Ladung bei bestimmten pH-Werten) mit berücksichtigt. Sie wird in Prozent angegeben. Hinweis: Eine Homologie ist nur eine von mehreren möglichen Ursachen für eine solche Ähnlichkeit.

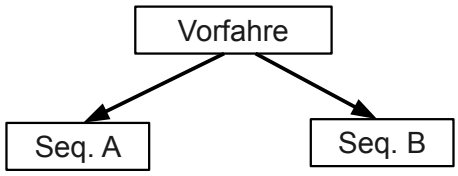
```
ILVCAGMFYWHKREQDNSTP  Aminosäure (Einbuchstabencode)
XXXXXXXXXXX·······X·  Hydrophober Rest
········XXXXXXXXXX·X  Polarer Rest
··XXXX·········XXXXX  Kleiner Rest
···················X  Prolin
····XX··········X··  Sehr kleiner Rest
XXX·················  Kettenförmiger Kohlenwasserstoffrest
·······XXXX·········  Aromatischer Rest
··········XXX·······  Positive Ladung
············X·X····  Negative Ladung
··········XXXX·X····  Geladen
```

Homologie (Homology)
Die betrachteten Sequenzen besitzen einen gemeinsamen Voräufersequenz (vgl. Orthologie). In Ausnahmefällen können sie sich sogar stark unterscheiden, etwa wenn auf Proteinebene inzwischen andere Funktionen erfüllt werden. Im Gegensatz zu nicht identischen Nukleotiden, können auch nicht identische Aminosäuren können einen evolutionären Bezug haben. Dies ist dann von ihren Eigenschaften abhängig (s.o. Ähnlichkeit/Similarity).

Identität (Identity)
Gib den Anteil der identischen Bausteine (Aminosäuren oder Nukleotide) in einer Sequenz einer bestimmten Länge in Prozent an.

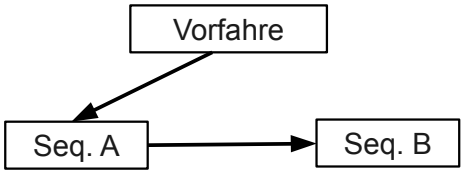
Orthologie

Sequenzähnlichkeit aufgrund eines evolutionären Zusammenhangs
(z.B. ein bestimmtes Enzym beim Menschen und bei der Maus)



Paralogie

Sequenzähnlichkeit aufgrund einer Gen-Duplikation
(z.B. zwei Kinasen des Menschen in verschiedenen Signalübertragungswegen)

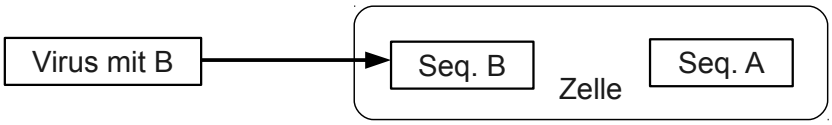


Analogie

Sequenzähnlichkeit aufgrund evolutionärer Konvergenz/ähnlicher Funktion

Xenologie

Sequenzähnlichkeit aufgrund der Übertragung/Aufnahme artfremder Gene z.B. durch ein Virus (z.B. Resistenzgene auf Plasmiden)



Vereinfachungen:

- Aminosäuren haben verschiedene chemische Eigenschaften, welche ihre relative Vertauschbarkeit beeinflussen (z.B. Hydrophilie, Polarität, pKS-Wert und Ladung bei bestimmten pH-Werten).
- Mehrfachsubstitutionen, die bewirken, dass z.B. nach einer Änderung an einer Position nach einer erneuten Änderung wieder der ursprüngliche Baustein an der entsprechenden Position steht, werden hier nicht berücksichtigt.
- Da sehr viele lokale Alignments möglich sind, ist ein aufwändiges Scoring notwendig, um diese Möglichkeiten zu bewerten. Dessen zugrundeliegenden Ähnlichkeitsmatrizen (Score Matrices) werden hier nicht thematisiert.
- Für die Bewertung alternativer Stammbaumvarianten ist das Parsimonieprinzip nützlich, das besagt, dass man mit möglichst wenigen Sequenzänderungen (und Stammbaumverzweigungen) ans Ziel kommen sollte.
- Der Unterschied zwischen phylogenetischen Stammbäumen und statistischen Abstandsbäumen wird hier nicht thematisiert.

4 Software

4.1 Texmaker, Texshade und Textopo (Druckausgabe von Alignments)

Installation unter Ubuntu- oder Debian-Linux:
```
sudo apt-get install texmaker texlive-science
```
Die Dokumentation liegt dann in den Verzeichnissen:
```
/usr/share/doc/texlive-doc/latex/texshade
/usr/share/doc/texlive-doc/latex/textopo
```

Installation unter Windows:
www.xm1math.net/texmaker
www.miktex.org
www.uni-kiel.de/pharmazie/chem/Prof_Beitz/dtse.htm
www.uni-kiel.de/pharmazie/chem/Prof_Beitz/textopo.htm

4.2 Clustalx (globale und lokale Alignments erzeugen)

Installation unter Ubuntu- oder Debian-Linux:
```
sudo apt-get install clustalx
```
Installation unter Windows:
www.clustal.org/clustal2

4.3 NJplot (Bäume anzeigen und Druckausgabe erzeugen)

Installation unter Ubuntu- oder Debian-Linux:
```
sudo apt-get install njplot
```

Installation unter Windows:
pbil.univ-lyon1.fr/software/njplot

Bäume können alternativ auch über den folgenden Webservice angezeigt werden:
www.ebi.ac.uk/Tools/phylogeny/clustalw2_phylogeny

4.4 Emboss (globale und lokale Alignments erzeugen)

Installation unter Ubuntu- oder Debian-Linux:
```
sudo apt-get install emboss jemboss
```

An Stelle einer Installation wird unter Windows die Benutzung des entsprechenden Webservices empfohlen:
www.ebi.ac.uk/Tools/webservices

4.5 PHYLIP (PHYLogeny Inference Package)

Installation unter Ubuntu- oder Debian-Linux:
```
sudo apt-get install phylip embassy-phylip
```
Installation unter Windows:
evolution.genetics.washington.edu/phylip.html

4.6 MAFFT (Multiple alignment program for amino acid or nucleotide sequences)

Installation unter Ubuntu- oder Debian-Linux:
```
sudo apt-get install mafft
```
Installation unter Windows oder online-Nutzung:
mafft.cbrc.jp/alignment/software

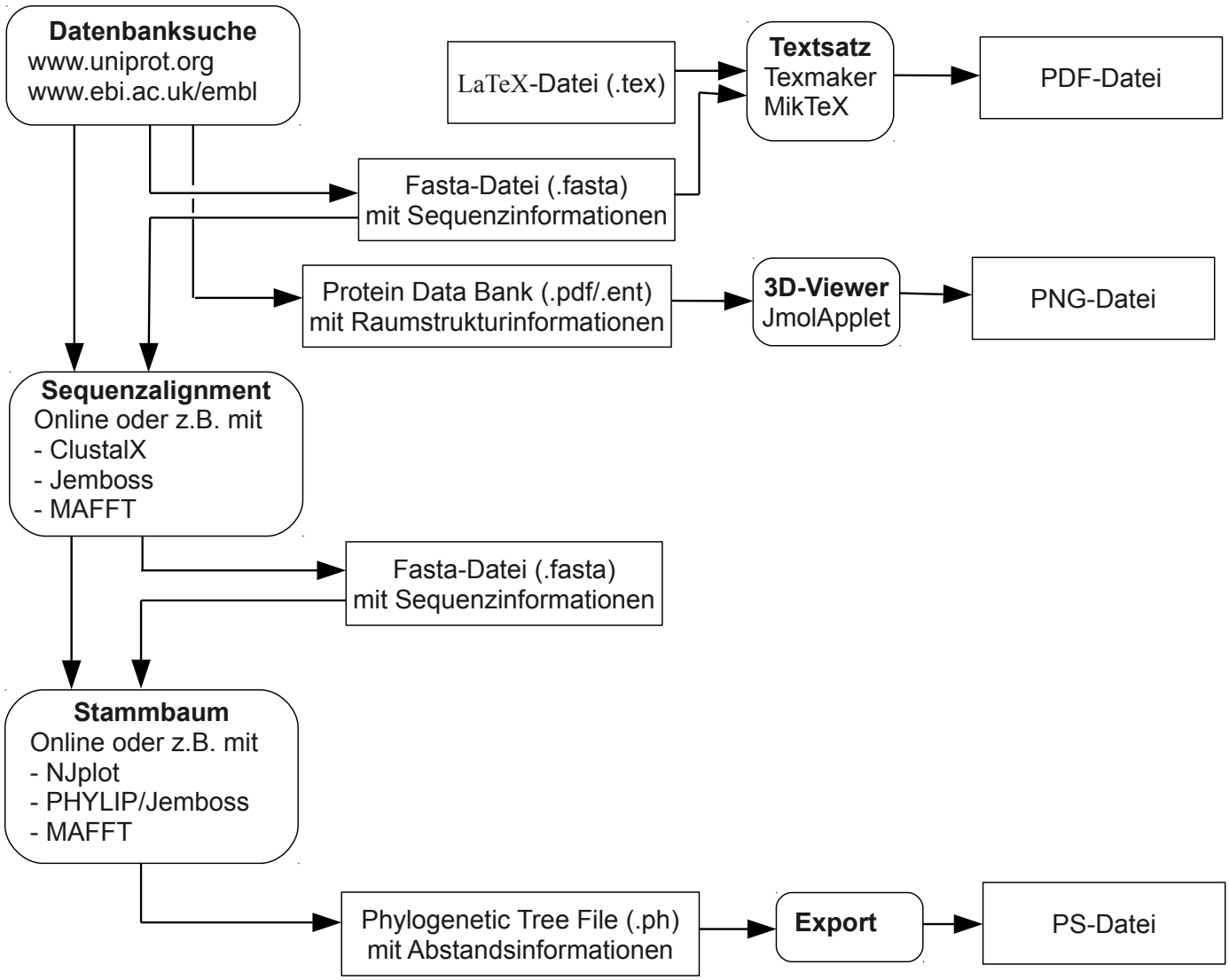
4.7 Archaeopteryx (Bäume anzeigen und bearbeiten)

Plattform übergreifendes Java basiertes Standalone-Programm oder Browser-Applet:
www.phylosoft.org/archaeopteryx

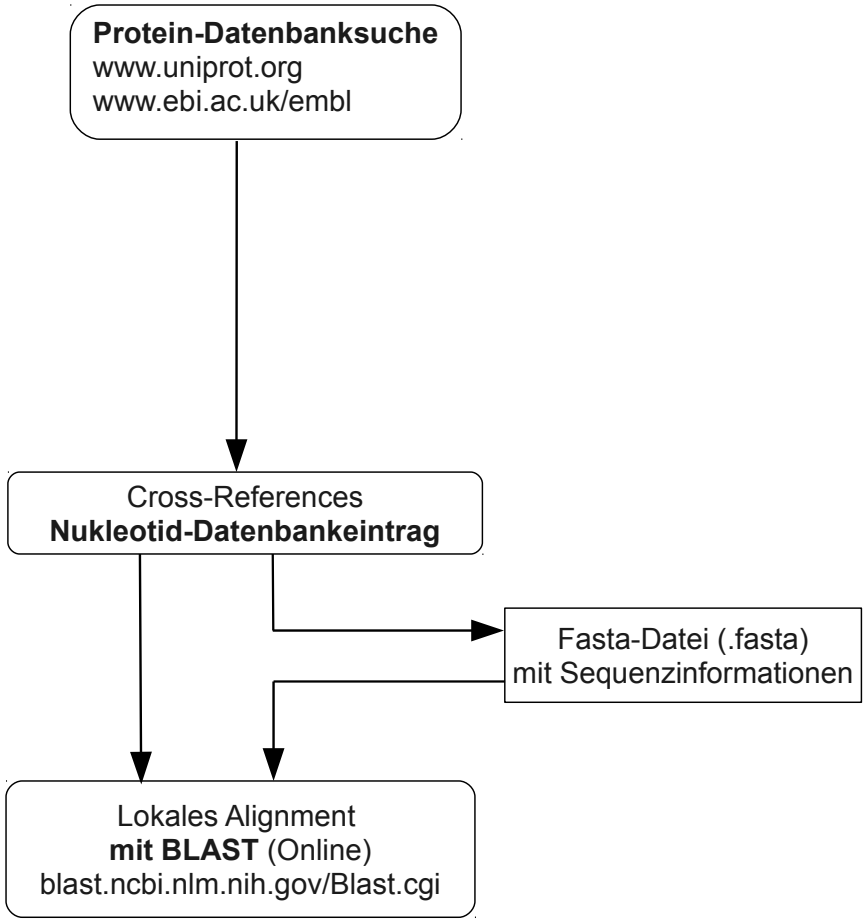
Tipp: Das Metapaket `science-biology` enthält noch viele weitere Programme zur Bioinformatik!

5 Workflow

5.1 Aminosäure-Sequenzen



5.2 Nukleotid-Sequenzen



6 Literaturverzeichnis

Einfache Beispiele für fertige Darstellungen, die mit Hilfe von Methoden aus der Bioinformatik gewonnen wurden:

[1] *Biologie: Ein Lehrbuch*, G. Czihak , H. Langer, H. Ziegler, Springer; Auflage: 6., unveränd. Aufl. (18. Oktober 1996), S. 874, 875, Abb. 10.33 Aminosäuresequenz des Proteins Cytochrom c von 20 verschiedenen Arten im Vergleich

[2] *Fokus Biologie* - Oberstufe - Gymnasium Bayern: 12. Jahrgangsstufe - Schülerbuch, Dr. S. Esders, G. Gräbe, Dr. W. Kleesattel, T. Linzmaier, Dr. F. Scholz, Prof. U. Weber, Dr. K. Wilhelm, Cornelsen Verlag (Mai 2010), S. 17, Abb. 1 Ausschnitte des Cytochroms c verschiedender Arten

[3] *Nautilus Biologie* Ausgabe B 12, H. Schauer, Bayerischer Schulbuch-Verlag (28. April 2010), S. 25, Abb. 9 Homologe Moleküle des Proteins Cytochrom c von Thunfisch und Reispflanze, Abb. 10 Ausschnitt aus einem Stammbaum

Weiterführende Literatur:

Lehrbuch der Genetik, W. Seyffert, Spektrum Akademischer Verlag; Auflage: 2. Aufl. (11. September 2003), Kapitel 33 Bioinformatik; S. 901-918

Angewandte Bioinformatik: Eine Einführung. Mit Übungen und Lösungen (Springer-Lehrbuch), P. M. Selzer, R. J.Marhöfer, A. Rohwer, Springer Berlin Heidelberg; Auflage: 1 (17. September 2003)

Bioinformatics for Dummies, J.-M. Claverie, C. Notredame, Wiley & Sons; Auflage: 1 (17. Januar 2003)

Instant Notes in Bioinformatics (Instant Notes Series), D. H. Weshead, J.H. Parish, R. M. Twyman , Taylor & Francis Ltd. (15. Juni 2002)

Bioinformatics: A Practical Approach (Chapman & Hall/CRC Mathematical and Computational Biology), Shui Qing Ye, Taylor & Francis; Auflage: 1 (9. August 2007)