**Übungen Bioinformatik im Rahmen des Modul 4 Methodenpraktikum Bachelor Biologie KIT**

**Sommersemester 2015**

|  |  |
| --- | --- |
| **Name:****Matrikel-Nr.** | **Name:****Matrikel-Nr.** |
| **Datum:** |  |

Dateien mit Ergebnissen auf dem eigenen Computer ablegen, Inhalte der Dateien hier einfügen. Bitte das Skript per Email zurückschicken.

**Sequenzanalysen 1**

Aufgaben können unter Umständen durch eine Abfolge von Klicks und *copy* / *paste* gelöst werden. Es ist ratsam, sich etwas Zeit zu nehmen und sich mit dem, was Sie machen, vertraut zu machen. Voraussichtlich werden Sie später Anwendungen, die hier zum Einsatz kommen, oft benutzen.

**Übung S1, Zugang zur PDB Protein Data Bank**

Koordinaten von Protein 3D Strukturen werden in der Protein Data Bank (PDB) abgelegt. Die Zuordnung erfolgt über einen "PDB-Code" (auch "PDB-ID"), der aus 4 Zeichen (Buchstaben oder Zahlen) besteht. Über diesen Code kann man die Struktur eines Proteins schnell und eindeutig finden. Jedem Teilnehmer wird ein Code zugeteilt, jeder arbeitet mit einem anderen Code. Gehen Sie zu der Seite <http://www.rcsb.org/>, tippen sie den Code unter “PDB ID or text” ein und klicken auf “Search”. So kommen Sie zu einer Seite, über die der Sie Information zu der entsprechenden Protein Struktur erhalten können. Sehr wahrscheinlich wird das Protein nicht in seiner gesamten Länge dargestellt. Außerdem werden oft mehrere Untereinheiten gezeigt.

Hier soll zunächst die Aminosäure-Sequenz einer Untereinheit ausgewählt werden. Die Aminosäure Sequenz im FASTA Format soll auf Ihrem Computer abgelegt werden. Die erste Zeile der Datei muss mit >beginnen, gefolgt von einer Beschreibung. Die nächsten Zeilen stellen die Protein Sequenz dar. Die Datei wird unter dem Namen **S1\_protein\_pdb.txt** abgelegt. Begleitende Information soll unten eingetragen werden.

Inhalt der FASTA Datei:

PDB ID:

Name des Proteins:

Artname:

Länge des Proteins (AS):

Wenige Worte zur Funktion:

Bild hier einfügen:

Öffnen Sie dann die pdb Datei mit Pymol. Dieses Programm muss auf Ihrem Rechner installiert sein. Es gibt freie Lizenzen von der Version für Studierende (http://pymol.org/educational/). Mit Pymol können Sie die 3D Struktur darstellen. Klicken Sie auf "all" "preset" "publication" und speichern Sie das Bild. Details werden in den Übungen besprochen. Das Bild soll mit den anderen Aufgaben abgegeben werden, Dateiname **S1\_pymol.png.** Bild bitte oben reinkopieren.

**Übung S2, Swissprot**

Die PDB enthält 107000 Einträge (März 2015). Insgesamt sind aber viel mehr Proteinsequenzen bekannt. Es gibt im Internet mehrere Datenbanken, über die man auf frei zugängliche Proteinsequenzen zugreifen kann. Das gleiche gilt für DNA Sequenzen.

Auf dem Swissprot Server <http://www.expasy.ch/> finden Sie unter UniprotKB die Swissprot Datenbank mit von Hand annotierten Sequenzen (544 996 Einträge, April 2014) und TrEMBL mit allen Protein Sequenzen, die automatisch von DNA Daten generiert werden (54 958 551 Einträge, April 2014).

Über eine Swissprot BLAST Suche bei <http://www.expasy.ch/tools/blast/> sollen Sie die vollständige Sequenz Ihres Proteins finden (es war ja "nur" die Sequenz des Proteins im Kristall). Dazu die Sequenz aus Übung 1 kopieren und in das Feld einfügen, dann “Run BLAST” klicken. Sie erhalten eine lange Liste von homologen Proteinen. Sie müssen in dieser Liste das richtige finden. Vermutlich ist es das erste, aber es gibt auch Ausnahmen von dieser Regel. Bitte überprüfen Sie die Spezies.

Wenn Sie auf das erste markierte Feld in der Zeile klicken, öffnet sich eine neue Seite, in der Sie Information zum Protein erhalten. Sie müssen die Sequenz des ausgewählten Proteins im FASTA Format darstellen. Gibt es Übereinstimmung mit S1\_protein1\_pdb.txt? Gibt es Unterschiede? Das hier gefundene Protein ist Protein1. Die Datei soll als **S2\_protein1.txt** gespeichert werden. Bitte nun wieder Inhalt der FASTA Datei sowie Begleitinformation unten eintragen.

Inhalt der FASTA Datei von Protein 1:

Name des Proteins:

Länge (Aminosäuren):

Unterschiede zu S1\_protein.txt:

Jedem Protein entspricht eine DNA Sequenz. Im nächsten Schritt soll die Sequenz der codierenden DNA ermittelt werden (CDS). In der Regel kommt man über “Cross-references” und Sequence Databases weiter. Unter "Translation" kann man die Sequenz anklicken. Erneut soll die Sequenz im FASTA Format gespeichert werden. Die Datei heißt **S2\_DNA.txt**. Den Inhalt der FASTA Datei sowie die Länge der DNA in Basenpaaren unten eintragen:

Inhalt der CDS FASTA Datei:

Länge der Sequenz in Nukleotiden:

Es ist nicht immer so einfach, die DNA Sequenz zu finden. Man kann nicht alle Eventualitäten vorhersehen, zumal sich die Programme der Server ändern. Bitte helfen Sie sich gegenseitig und fragen Sie Tutoren.

**Übung S3, Protein und DNA Sequenzen auf NCBI Server**

Sie sollen nun die gleiche Datenbankrecherche auf dem NCBI Server durchführen. Unter NCBI sind viele Datenbanken zusammengefasst, man kann unter anderem nach Genomen, Proteinen, DNA, Literatur oder Strukturen suchen. Wie in Übung S2 soll das Protein1 mit BLAST gesucht werden.

Dazu unter <http://www.ncbi.nlm.nih.gov/> unter “popular resources” auf “BLAST” klicken. Unter “Basic BLAST”, auf “protein blast” klicken, die Sequenz ins Fenster einfügen und auf den “BLAST” Knopf klicken. (Sie können den Prozess beschleunigen, wenn Sie den Organismus eingeben.)

Die FASTA Datei sollte nun unter **S3\_protein.txt** abgespeichert werden. Die codierende DNA Sequenz soll ebenfalls ermittelt werden. Diese im FASTA Format unter **S3\_DNA.txt** ablegen. Sind es die gleiche Sequenzen wie unter S2? Bitte hier aufschreiben:

Handelt es sich um die gleichen Sequenzen wie in S2?

FASTA Sequenz des Proteins:

FASTA Sequenz der DNA:

**Übung S4, Suche nach einem homologen Protein**

Proteine mit gleicher Funktion haben meist eine homologe Sequenz, umgekehrt kann man bei Homologie von der gleichen Funktion ausgehen. Wie ähnlich Protein-Sequenzen sind, hängt unter anderem von ihrem evolutionären Abstand ab. Hier sollten Sie ein Protein finden, welches 50%-55% Identität mit Protein1 besitzt. Wegen der sehr großen Zahl von bekannten Proteinsequenzen findet man bei einer Standard BLAST Suche nur Proteine mit höherer Homologie. Geben Sie die Sequenz von Protein1 in NCBI BLAST ein, und schließen Sie die Organismengruppe, zu der das Protein gehört, aus. Wenn in der Liste ein Protein mit 50%-55% Identität erscheint, ist das Ziel fast erreicht. Wenn die Homologie zu hoch ist, müssen Sie entweder die Zahl der gelisteten Proteine auf 1000 erhöhen oder eine Taxonomie Gruppe wählen, die deutlich von der des Protein1 abweicht. Sie sollten außerdem unter "formatting options" einen Identitäts-Bereich (z.B. 45-55%) angeben. Das so identifizierte Protein wird im folgenden als **Protein2** bezeichnet. Unter **S4\_protein.txt** bitte die Sequenz im FASTA Format ablegen. Außerdem zusätzliche Information angeben:

Inhalt der FASTA Datei von Protein 2:

Name des Protein2:

Organismus (Artname) und Gruppe (z.B. Grünalge):

Länge des Protein2 (AS):

Übereinstimmung mit Protein1 (AS):

Identität Protein1 und Protein2 (%):

Kurze Beschreibung, wie Protein2 gefunden wurde:

**Übung S5, Needleman Wunsch Analyse**

Diese Übung wird nur im Kurs gemacht. Der Needleman Wunsch Algorithmus resultiert in einem *alignment* von zwei Sequenzen. Details werden in der Vorlesung besprochen. In dieser Übung sollen Sie nachvollziehen, wie der Algorithmus im Prinzip funktioniert. Sie sollten dazu zwei kurze DNA Sequenzen mit je 8 bp „erfinden“, die ähnlich, aber nicht identisch sind. Ein Beispiel wäre: ATGTCTAT und ATTCTAGT

Folgende Parameter werden vorgegeben:

Gap cost: -1; match: 2; mismatch: 1

In der unteren Tabelle ist das Ergebnis dargestellt.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|   |   | A | T | G | T | C | T | A | T |
|   | 0 | -1 | -2 | -3 | -4 | -5 | -6 | -7 | -7 |
| A | -1 | 2 | 1 | 0 | -1 | -2 | -3 | -4 | -5 |
| T | -2 | 1 | 4 | 3 | 2 | 1 | 0 | -1 | -2 |
| T | -3 | 0 | 3 | 5 | 5 | 4 | 3 | 2 | 1 |
| C | -4 | -1 | 2 | 4 | 6 | 7 | 6 | 5 | 4 |
| T | -5 | -2 | 1 | 3 | 6 | 7 | 9 | 8 | 7 |
| A | -6 | -3 | 0 | 2 | 5 | 7 | 8 | 11 | 10 |
| G | -7 | -4 | -1 | 2 | 4 | 6 | 8 | 10 | 12 |
| T | -7 | -5 | -2 | 1 | 4 | 5 | 8 | 9 | 12 |

Tragen Sie nun in die nächsten Tabelle folgende Sequenzen ein:

Quer: AGTCCGTA / senkrecht: GTCAGTAC. Füllen Sie die Felder entsprechend von Hand aus. Dies wird vermutlich etwas Zeit in Anspruch nehmen. Sie sollen die Richtigkeit selbst überprüfen. Dazu gibt es eine Excel Datei.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |
|   |   |    |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |

Für das *alignment* ist die Tabelle erst der erste Schritt, im nächsten Schritt werden die Sequenzen übereinander dargestellt entsprechend der höchsten Werte in der Tabelle.

**Übung S6, globales Alignment nach Needleman Wunsch**

Hier machen Sie ein Needleman Wunsch *alignment* von Protein1 und Protein2 mit Hilfe eines Internet Programms.

<http://mobyle.pasteur.fr/cgi-bin/portal.py>

alignment->pairwise->global->needle

Das Ergebnis soll in der Datei **S6\_needle.txt** abgelegt werden.

|  |
| --- |
|  |
| ######################################### Program: needle# Rundate: Wed 10 Mar 2010 10:20:37# Commandline: needle….## Length: 516# Identity: 287/516 (55.6%)# Similarity: 376/516 (72.9%)….#======================================= 1 MFINRWLFSTNHKDIGTLYLLFGAWAGMVGTALSLLIRAELGQPGT--LL 48 .|||:||||||||.||::|..:.|:|||.||:|||.:|...|| |. 1 ---MRWLYSTNHKDIGMLYIIFAFFGGLVGTGLSVLIRLQLATTGTGILQ 47 :|||||.:.||||||:|||.:| ...||.:||||||||||||||||||| 196 VPVLAAALIMLLTDRSLNTAYF--VDSGDLLLYQHLFWFFGHPEVYILIL 243 |
|  |

Bitte Inhalt der Datei hier einfügen:

**Übung S7, globales vs. lokales *alignment*, Vergleich Needleman Wunsch und Smith Waterman**

Hier werden andere Sequenzen verwendet als in S6!

In einem globalen *alignment* werden die Sequenzen in ihrer gesamten Länge berücksichtigt. Aber sehr oft sind nur Teile von Sequenzen homolog zueinander. Proteine bestehen meist aus mehreren Domänen, die Domänen Anordnung kann variieren, selbst zwischen eng verwandten Proteinen. Daher ist eine globale Analyse oft nicht sinnvoll. Die beiden Sequenzen unten enthalten beide eine "GAF" Domäne, die anderen Domänen sind verschieden (siehe SMART Abbildungen). Mit einem **globalen** *alignment* wird man voraussichtlich nicht den korrekten, homologen Bereich finden. Ein **lokales** *alignment* identifiziert zumindest den Bereich mit der höchsten Homologie. Sie sollten nun die beiden unten angegebenen Sequenzen mit NEEDLE (Needleman Wunsch, globales *alignment*) und dann mit WATER (Smith Waterman, lokales *alignment*) vergleichen. Sie müssen für Parameter (und evtl Matrix) verschiedene Möglichkeiten testen. Die Sequenzen müssen sich in jedem Fall überlappen.

<http://mobyle.pasteur.fr/cgi-bin/portal.py>

alignment->pairwise->global->needle

alignment->pairwise->local->water



SMART Domänen von *Mesorhizobium* sp. BNC1 protein Q11GZ0\_MESSB (Q11GZ0)

>tr|Q11GZ0|Q11GZ0\_MESSB Sensor protein OS=Mesorhizobium sp. (strain BNC1) GN=Meso\_1942 PE=4 SV=1

MKFAESRTELLLYSAIAGLMALTFIADLYFPLGTAIWVIYLFPIALCYLVSRPVVPVAAA

TVVSFFVVIGFFYAPAGVDPDIARTNRVLGALTSWIIALTGYFFIRNKLAIQKQEWLQAG

QVHLARSVSGNQKIEELSGNILAFLAEYLDAQAGALYVRRGSTFGRSGVYGVPAGAPIPD

RFKPGDGLLGQAVADRRSFVLRDVPENYLYFGSGLGRAKPRHLLIAPASVDDEVNAVLEL

GFLKEAGEEVGELLERISEPMAIALRSARYRAQLQELLEETQRQAEELQTQSEELRSANE

ELENQSRSLQDSQVRLEEQQAELEQSNAQLEEQTQLLEVQRDELSRAQGALQAKARDLEE

ASRYKSEFLANMSHELRTPLNSLLIMARLLADNREGNLTSDQVKFAETIETAGNDLLNLI

NDILDISKIEAGHIELQIERVRVEPMLEKLRATFQSSASEKGLNFRIETVPGTPDRIESD

PQRIEQVLKNFLSNAIKFTEKGEVVLSVAEEEDGRIAFAVRDTGIGIPEDQQQAIFEAFR

QADGSINRKYGGTGLGLSISSELTRVLGGEIRVESTVGQGSSFILVVPKIYDAASVQKSA

GAGRRTPVVSRAAPAAAESEAEKIDTVPARPRIIEDDRDKLTGGRVILVVEDDPAFARIL

MDLAHELGFECLVAETADEGVLLARRYLPAGVILDIGLPDHTGLSVLDRIKHDVRTRHIP

VHVVSVSDYTRTARSFGAAGYMLKPVKREELERALTGLAKQLAQTVRRVLIVEDDATQLE

ALQLLLSSQDVETIAAASAAECFGKLKAETFDCMVLDLSLPDASGFEVLEQLSAEDAYSF

PPTIVYTGRDLTPDEELRLRKFSKSIIIKGAKSPERLLDEVTLFLHQMVAELPERKREML

VRSLNRDAALEGRRILVVEDDIRNVYALMSVFEGHGAAVQIARNGREALQALEKMADEGN

PAIDLVLMDVMMPEMDGLTATREIRKRERWKNLPIIVLTAKAMPGDQEQCLAAGANDYLA

KPLDVDKLLSLARVWMPR



SMART Domänen von *Sinorhizobium medicae* Protein A6UJ84\_9RHIZ (A6UJ84)

>tr|A6UJ84|A6UJ84\_SINMW Signal transduction histidine kinase OS=Sinorhizobium medicae (strain WSM419) GN=Smed\_4932 PE=4 SV=1

MHGAEILPKGVFDDETSLHEPDFRAVLDALTVAVYATDAEGVVTYCNAAAAAIAGREPEL

GKDRWCVSWQLRRPDGSPLPHEQCPMARALKERRPIRGEELIAVRPNGAAVPLLPHPTPM

FDAAGALTGAVNVLIDLSNTKDMERSSRYLAAIVESSDDAIVAKDLNGIITSWNRGAERL

FGYTAEDVIGKSITILIPQDRLSEEPGILERIRRGDKVDHFETKRRRKDGSLIDISLTVS

PVRDATGRIVGASKIARDVTETKRAAEALSDRFREQAALYRFTERVQRTKRIEDVFEAAL

DAILEALSCDRASILLFDPSGTMRFVASRGLSEAYQSAVDGHSPWASGASEPEPIFVENV

DSTDMSRALKEAIAQEGIGALGFFPLVLDGKLIGKFMTYYDEPHRFAESEIGVAVTIARQ

LGFSLQRMETENARRQAEERLRRNEANERARAGELAAIMEAVPALIWIARTPDCHVISGN

KSSYEMLRLPHDSNLSLSAPPGERPTNFRVFSGGRALSPDELPVQRAARGEEVRNFEEEV

WFEDGSSRYIFGNATPLRDAAGEVIGAVAASVDITERRQAEEALQESERRLQLALDAGHM

GAWEWNLGTDEVIWSPGLEALHRLERGSFGRTLADFKRNVHPDDLPLIDREIAKAVETKG

DYHVVYRARLLDGTVRWMEAFAQFSPSNGPSRRLIGVCMDITERREAEAQRNLLVAELSH

RVKNTLAIVGSIARQTFSTSPDARTAHRSFDARIRALAQTHTRLAEASWSGVSLQTVLYD

ELAPYQDDGRTNVKLQGPTAMLPPKHALTLGMAAHELATNAAKHGALSVKSGKVDIEWVV

DVDTDRLRICWRESGGPPVVTPAHNGFGRLLLERVLASDLGGEVHLEFKPQGLVCTIDVP

YPRGAPG

|  |
| --- |
| ######################################### Program: needle…# Aligned\_sequences: 2# 1: A6UJ84\_SINMW# 2: Q11GZ0\_MESSB# Matrix: EPAM250# Gap\_penalty: 10.0# Extend\_penalty: 10.0…#=======================================A6UJ84\_SINMW 1 MHGAEILPKGVFDDETSLHEPDFRAVLDALTVAVYATDAEGVVTYCNAAA 50 Q11GZ0\_MESSB 0 -------------------------------------------------- 0A6UJ84\_SINMW 51 AAIAGREPELGKDRWCVSWQLRRPDGSPLPHEQCPMARALKERRPIRGEE 100  |
|  |
| ######################################### Program: water# Rundate: Wed 10 Mar 2010 17:08:43…Q11GZ0\_MESSB 241 GFLKEAGEEVGELLERISEPMAIALRSARYRAQLQE-LLEETQRQAEELQ 289 :|.::...:...|::| ..:.|:..::. |:...:. ||:.|.|..|.: A6UJ84\_SINMW 635 DFKRNVHPDDLPLIDR-EIAKAVETKGD-YHVVYRARLLDGTVRWMEAF- 681… |

Ergebnis der Needleman Wunsch Analyse

Ergebnis der Smith Waterman Analyse

Die Übungen knüpfen an die "paarweisen Sequenzvergleiche" an. Die ersten Aufgaben dienen der Vorbereitung, damit Sie schließlich durch selbständige Arbeit eine vernünftige Zusammenstellung von homologen Sequenzen erhalten. Es sollte klar werden, wozu BLAST angewandt wird, wozu SMART bzw. PFAM wichtig sind, wozu multiple *alignments* sinnvoll sind, und wie ein *alignment* durchgeführt werden kann

**Übung M1, BLAST Suche nach homologen Proteinen aus unterschiedlichen Organismen**

Für diese Aufgabe finden Sie im Anhang (Ende dieses Skripts) eine Zusammenstellung von Organismengruppen und Organismen, die Sie auswählen können. Beachten Sie bitte, dass die Sequenz von Protein1 immer dabei sein muss, auch wenn Protein1 nicht in einem der Organismen vorkommt. Protein2 sollte ebenfalls dabei sein.

Zahlreiche Proteine, die verwandt sind mit **Protein1**, werden mit BLAST gesucht. Damit soll ein Stammbaum erstellt werden, aus dem Verwandtschaftsverhältnisse abgelesen werden können. Sie sollen 30 oder mehr verschiedene Proteine zusammenstellen, die Identität soll zwischen 30% und 90% liegen. Geben Sie bei NCBI Blast **Protein1** als Suchsequenz ein und schränken die Suche auf einzelne taxonomische Gruppen ein, indem Sie den taxonomischen Namen in das entsprechende Feld eintippen. Vorschläge für taxonomische Gruppen finden Sie im Anhang.

Nun gibt es jedesmal mehrere Möglichkeiten: (i) es gibt gar kein homologes Protein. In diesem Fall machen Sie sich eine Notiz und suchen bei der nächsten Gruppe weiter. (ii) Es gibt eines oder mehrere homologe Proteine, die Homologie erstreckt sich über >50% der Länge von **Protein1**. In dem Fall wählen Sie das Protein mit der höchsten Homologie aus (niedrigster *E-value* oder höchste *identity* oder *score*). Bitte notieren Sie die Art und den Stamm. Die Proteinsequenz bitte im FASTA Format abspeichern. Speichern Sie weitere homologe Sequenzen aus anderen Arten ebenso ab. Diese Sequenzen sollten sich einerseits unterscheiden, d.h. die Identität mit Protein1 soll deutlich anders sein, andererseits muss sich aber die Homologie über einen genügend langen Bereich erstrecken. (iii) Es gibt homologe Bereiche, aber die Homologie erstreckt sich nur über einen kleinen Teil der Sequenz. Speichern Sie die Sequenz wie unter (ii). Wenn später genügend "bessere" Sequenzen zur Verfügung stehen, können Sie diese vernachlässigen. Wenn nicht, sollten Sie die spätere Analyse auf den homologen Bereich beschränken. Bitte Rücksprache mit Betreuer halten (z.B. Email).

Suchen Sie auf jeden Fall so lange nach homologen Sequenzen, bis es 30 oder mehr sind. Wenn alle Sequenzen abgespeichert vorliegen, muss aus Gründen der Einheitlichkeit und Verständlichkeit die Bezeichnung der Proteine vereinfacht werden. Dazu gehen Sie bitte wie folgt vor:

Wählen Sie eine Abkürzung von 4-5 Buchstaben für Ihr Protein (zum Beispiel CytoC für Cytochrom C). Diese Abkürzung wird im Folgenden mit YYYYY gekennzeichnet. Dann müssen Sie eine Abkürzung aus genau 5 Zeichen für jeden Artnamen verwenden. Zum Beispiel steht Arath für *Arabidopsis thaliana* und Homsa für *Homo sapiens*. (3 Buchstaben für den Gattungsnamen und 2 für die Bezeichnung der Art). Diese Abkürzung wird durch XXXXX gekennzeichnet.

Editieren Sie die erste Zeile in jeder FASTA Datei so, dass nach dem ">” Zeichen direkt XXXXXYYYYY steht, danach folgt ein Leerzeichen (z.B.. >ArathCytoC ……). Verändern Sie ansonsten die Zeile nicht. Bitte benennen Sie nun auch den Dateinamen der FASTA Dateien wie folgt um: "XXXXXYYYYY.fasta” (z.B. "ArathCytoC.fasta”). Die Dateien von **Protein1** und **Protein2** sollen auch in der Weise bearbeitet werden. Bitte tragen Sie die Namen hier ein:

Namen von Protein1 und Protein1-homologe:

**Übung M2, BLAST Fragen**

BLAST ist in der biologischen Forschung ein sehr wichtiges und viel benutztes Programm. Sie sollten sich selbständig mit BLAST etwas vertraut machen, siehe auch <http://www.ncbi.nlm.nih.gov/books/NBK21097/>.

Bitte beantworten Sie folgende Fragen zu BLAST:

Was bedeutet Bit score?

Was bedeutet Expect oder E-value?

Wofür steht die Abkürzung BLAST?

Die Antworten bitte hier eintragen.

**Übung M3, PFAM und SMART**

Das multiple *alignment*, welches später durchgeführt werden soll, setzt voraus, dass Proteine homolog zueinander sind. Es liegt in der Natur der Dinge, dass man bei der BLAST Suche auch Proteine finden kann, die nur über einen kleinen Bereich homolog sind. Homologe Proteine besitzen homologe Domänen, und diese Domänen können Sie mit speziellen Programmen darstellen. Wenn dabei herauskommt, dass "Ausreiser" dabei sind, müssen Sie ggf. zu M2 zurück gehen und neue Sequenzen suchen.

Die Protein Domänen sollen mit SMART und PFAM überprüft werden. Bevor Sie die Domänenstruktur Ihrer "eigenen" Proteine anschauen, sollten Sie zum Einstieg und zur Übung die Domänenstrukturen der beiden Proteine "ArathPhyB" und "XanaxUncha" ermitteln. Sie sollten jedes Mal sowohl PFAM (<http://pfam.sanger.ac.uk/>) als auch SMART (<http://smart.embl-heidelberg.de/>) benutzen.

>ArathPhyB

MVSGVGGSGGGRGGGRGGEEEPSSSHTPNNRRGGEQAQSSGTKSLRPRSNTESMSKAIQQYTVDARLHAVFEQSGESGKSFDYSQSLKTTTYGSSVPEQQITAYLSRIQRGGYIQPFGCMIAVDESSFRIIGYSENAREMLGIMPQSVPTLEKPEILAMGTDVRSLFTSSSSILLERAFVAREITLLNPVWIHSKNTGKPFYAILHRIDVGVVIDLEPARTEDPALSIAGAVQSQKLAVRAISQLQALPGGDIKLLCDTVVESVRDLTGYDRVMVYKFHEDEHGEVVAESKRDDLEPYIGLHYPATDIPQASRFLFKQNRVRMIVDCNATPVLVVQDDRLTQSMCLVGSTLRAPHGCHSQYMANMGSIASLAMAVIINGNEDDGSNVASGRSSMRLWGLVVCHHTSSRCIPFPLRYACEFLMQAFGLQLNMELQLALQMSEKRVLRTQTLLCDMLLRDSPAGIVTQSPSIMDLVKCDGAAFLYHGKYYPLGVAPSEVQIKDVVEWLLANHADSTGLSTDSLGDAGYPGAAALGDAVCGMAVAYITKRDFLFWFRSHTAKEIKWGGAKHHPEDKDDGQRMHPRSSFQAFLEVVKSRSQPWETAEMDAIHSLQLILRDSFKESEAAMNSKVVDGVVQPCRDMAGEQGIDELGAVAREMVRLIETATVPIFAVDAGGCINGWNAKIAELTGLSVEEAMGKSLVSDLIYKENEATVNKLLSRALRGDEEKNVEVKLKTFSPELQGKAVFVVVNACSSKDYLNNIVGVCFVGQDVTSQKIVMDKFINIQGDYKAIVHSPNPLIPPIFAADENTCCLEWNMAMEKLTGWSRSEVIGKMIVGEVFGSCCMLKGPDALTKFMIVLHNAIGGQDTDKFPFPFFDRNGKFVQALLTANKRVSLEGKVIGAFCFLQIPSPELQQALAVQRRQDTECFTKAKELAYICQVIKNPLSGMRFANSLLEATDLNEDQKQLLETSVSCEKQISRIVGDMDLESIEDGSFVLKREEFFLGSVINAIVSQAMFLLRDRGLQLIRDIPEEIKSIEVFGDQIRIQQLLAEFLLSIIRYAPSQEWVEIHLSQLSKQMADGFAAIRTEFRMACPGEGLPPELVRDMFHSSRWTSPEGLGLSVCRKILKLMNGEVQYIRESERSYFLIILELPVPRKRPLSTASGSGDMMLMMPY

>XanaxUncha tr|Q8PLK2|Q8PLK2\_XANAC Putative uncharacterized protein OS=Xanthomonas axonopodis pv. citri GN=XAC1795 PE=4 SV=1

MPSLPASAESLLPPTLDESTRLEVLRGLCLLDSPPDPVFDTVAAMAARSLDAEIAVVSLV

DEHRQWFKARIGLEARETPRSQAFCAHAIRSDEVMVVPDAQLDPRFCDNPLVLGPPFIRF

YAGAPLKLRDGHRIGTLCVIGTSPRPGLDQAAITQLEGLRDLAVLRVENLRSTTYRDGPT

GLPNRSRFSEDLDTWLSQRDTAPATTAVAIDVCGSDYFRDMVKALGWEYADGYIALAQRR

LAAYLPGGTLLYRLDPTTFGFLAQAEGQRLATLCTKVSKAFTEPLEHQGIPHTAVASIGA

VSLQSSYGAADTIRSLTTAVDIARERCLPWSMYERKHDVAQRNTFRLLAALPAALDSSSQ

LRLHFQPRVDLHDHRCVGVEALLRWQHPMIGPVMPSDFIPMAEKTALINRITAWVIDNGI

AQAARWQQQGLDFNLALNVSAADLDRPGFAGLLRRGLDRHKLDPRRLEIEFTESAMIRHP

DHLAEQLAAIAALGVHIAIDDFGTGYSNFSYLKQLPASSLKIDQSFIRSLPDSRTDRTLV

PAMIQLGHSLGQRVVAEGIESAEAYAQLRAWGCDEGQGYWIAKPMPAAALETWLETPWHE

QHAAPVNLLAASLAAARV

Beispiel für SMART Bild



Beispiel für PFAM Bild.



Bitte beschreiben Sie kurz alle 4 Ergebnisse in Worten: welche Domänen kommen vor? Was ist der Unterschied zwischen den Proteinen? Was ist der Unterschied zwischen SMART und PFAM Ergebnissen?

Ergebnis der SMART und PFAM

Jetzt sollen Sie alle Proteine, die Sie in M1 zusammengestellt haben (die Protein1-Homologe), in SMART und PFAM analysieren. Bitte kopieren Sie alle Bilder in eine Datei, die Sie für sich selbst abspeichern. Bitte immer eine kleine Bildunterschrift dazu fügen. Wenn alle Bilder fertig sind, sollte die Datei im pdf Format gespeichert werden (nicht abgeben).

Bitte beschreiben Sie das Ergebnis mit eigenen Worten: welche Domänen kommen vor? Sind alle Proteine gleich oder kommen unterschiedliche Domänen vor? Was kann man über die Länge der Proteine sagen?

Ergebnis der Domänenanalyse der Protein1-homologen Proteine

Anmerkung: auch bei NCBI können Sie die Domänen-Anordnung darstellen lassen (Interpro). Diese Ergebnisse können Sie auch gerne mit verwerten.

**Übung M4, Wenn die Domänen Ihrer Proteine nicht übereinstimmen**

Wenn die Proteine die gleichen Domänen in der gleichen Anordnung besitzen, können Sie einfach weiter machen. Wenn nicht, müssen später die nicht homologen Bereiche mit *jalview* entfernt werden. Sonst wird die phylogenetische Analyse „falsch“.

**Übung M5, multiples *alignment* mit ClustalX2**

In dieser Übung führen Sie ein multiples *alignment*, das heißt ein *alignment* mit mehr als 2 Sequenzen, mit dem "klassischen" Programm ClustalX2 durch. Dieses Programm arbeitet mit einer einfachen graphischen Oberfläche. Hier geht es um ein *alignment* der in M2 ausgewählten Sequenzen.

Ihre Sequenzen liegen einzeln vor. Sie sollten alle Sequenzen in einer Datei zusammenfassen. Im Folgenden sind vier Möglichkeiten dafür beschrieben: (i) Öffnen Sie Windows "Powershell" (unter "Ausführen" oder "Durchsuchen") und wechseln Sie auf Ihr Verzeichnis durch (mit "cd.." und "cd xy" und "ls"). Dann tippen Sie "cat \*.fasta > alle". In der Datei "alle" stehen nun alle FASTA Sequenzen nacheinander. Anschließend benennen Sie die Datei "alle" zu "alle.fasta" ("mv alle alle.fasta”). (ii) Statt Powershell verwenden Sie "cmd" und den Befehl "copy \*.fasta alle" (iii) Kopieren Sie mit einem Texteditor alle Datei-Inhalte zusammen und speichern den Inhalt als Textdatei. (In Word oder OO gespeicherte Dateien sind nicht brauchbar). Dies ist einfacher, dauert aber länger. (iv) Sie lesen eine Datei nach der anderen in ClustalX ein.

Weitere Programme: MAFFT (http://mafft.cbrc.jp/alignment/software/), ClustalΩ (http://www.clustal.org/omega/), Tcoffee

Die Datei(en) mit den Sequenzen sollten sich auf dem ClustalX2 Verzeichnis befinden. Nach dem Öffnen von ClustalX2 müssen Sie die FASTA Sequenzen einlesen. Dazu "File"->"Load Sequences" klicken. Die Datei "alle.fasta" klicken. Einzelne Sequenzen können über "Append Sequences" nachgeladen werden. (Das heißt, Sie können sowohl alle Sequenzen einzeln einlesen, oder fehlende Sequenzen ergänzen.) Unter "Alignment" -> "Alignment parameters" -> "multiple alignment parameters” sollten Sie die Matrix in "BLOSUM series” ändern. Danach können Sie gleich mit "Alignment” -> "Do complete alignment” starten. Die Berechnung dauert eine Weile. Sie werden sehen, dass ClustalX2 einfach Leerzeichen in die Sequenzen einfügt wurden. Sie sollten nun das *alignment* in drei Formaten speichern: fasta (\*.fasta), Clustal (\*.aln), PHYLIP (\*.phylip). Daraus die Dateien **M5\_clustal.fasta** **M5\_clustal.aln** und **M5\_clustal.phylip** machen. Sie sollten dann unter "File" -> "Write alignment as postscript" eine *postscript* Datei erstellen, die dann in eine pdf Datei umgewandelt werden muss, zum Beispiel hier: http://online2pdf.com/. So entsteht die Datei **M5\_alignment.pdf**. Zur Dokumentation, dass die Übung richtig gemacht wurde, kopieren Sie bitte grafisch den vorderen Teil des Clustal Bildes.

Bitte hier den vorderen Teil der ClustalX Ergebnis mit den Namen der Proteine und dem Anfangsbereich der Sequenzen z.B. über *snipping tool* reinkopieren:

**Übung M6 multiples *alignment* mit MUSCLE**

Clustal wird nach und nach von besseren Programmen verdrängt. Diese Programme laufen entweder auf einem Server im Internet, und Sequenzen werden über eine Maske eingegeben, oder kommandozeilenorientiert. Hier soll das leistungsfähige Programm MUSCLE (<http://www.drive5.com/muscle>) als Beispiel für Eingabe über Kommandozeile angewandt werden. Das Programm muscle.exe in ein extra Verzeichnis installieren, z.B. unter C:/muscle. Die "alle.fasta" Datei in das gleiche Verzeichnis kopieren. Dann unter powershell oder cmd tippen: "muscle.exe -in alle.fasta -out alle\_muscle.fasta". (Dateien können auch andere Namen haben, aber die erste muss eine Datei sein, die alle Sequenzen im fasta Format enthält.) In der Datei alle\_muscle.fasta steht nun das *alignment*. Die Ausgabedatei soll wiederum in ClustalX2 eingelesen und in verschiedenen Formaten abgespeichert werden. Sie erhalten die Dateien **M6\_muscle.fasta**, **M6\_muscle.aln**, **M6\_muscle.phylip,** **M6\_muscle.pdf**. Bitte erneut wie in Aufgabe M5 den Anfang hierher kopieren:

Bitte hier den vorderen Teil des Muscle alignments, dargestellt mit ClustalX, mit den Namen der Proteine und dem Anfangsbereich der Sequenzen z.B. über *snipping tool* reinkopieren:

**Übung M7 Entfernen von *gaps***

Hier wird mit Jalview <http://www.jalview.org/Web_Installers/install.htm> gearbeitet. Es lohnt sich, dieses Programm besser kennen zu lernen. Es wird oft verwendet und bietet viele Möglichkeiten.

Beliebige Bereiche können blockweise editiert werden. So können Sie Bereiche im N-terminus oder C-terminus ausschneiden, die nicht die gleichen Domänen enthalten. Bei *alignments* werden, wie gesagt, Leerzeichen (*gaps*) eingefügt. Für phylogenetische Studien ist es erforderlich, Bereiche mit zu vielen *gaps* zu entfernen. Das Entfernen von *gaps* muss spaltenweise erfolgen, dabei gehen Aminosäuren verloren! Öffnen Sie mit Jalview (nacheinander) die beiden *alignment* Dateien aus M5 und M6, die Sie im aln oder fasta Format abgespeichert haben. Dokumentieren Sie durch *screenshots* (snipping tool) wie das alignment vor und nach dem Editieren aussieht (Datei **M7\_jalview.pdf**). Das Markieren der Spalten und anschließende Löschen sollte ohne weitere Anleitung gemacht werden können. Abschließend sollten die Dateien unter **M7\_clustal\_wo\_gaps.fasta M7\_muscle\_wo\_gaps.fasta** abgespeichert werden In späteren Aufgaben können Sie mit oder ohne *gaps* weiter arbeiten.

Bitte zeigen Sie einen Bereich vor und nach dem Entfernen von gaps (snipping tools).

Bild eines Bereich vor Entfernen von Spalten mit gaps

Bild gleicher Bereich nach Entfernen von Spalten mit gaps

**PDB Codes**

1J77 1JFF 2VVH 3JW0 3O18 1ATN 1B4K 1B7T 1B7Y 1BL8 1BNC 1BOU 1C17 1DPI 1E0U 1FBB 1FSZ 1FXA 1HOX 1HTB 1IQR 1J77 1JBO 1JFF 1M1N 1M1Y 1MBN 1QCQ 1RCX 2C2A 2C7L 2HYX 2IX5 2ONE 2QWM 2SOD 2WVH 2YPI 2ZT9 3ABL 3AEK 3C2W 3EYX 3FKY 3FWY 3JW0 3KZI 3L0F 3LDH 3LGB 3LGB 3LW5 3MYR 3NUL 3PGK 3PGM 4ALD 4PFK 4A16 3B13 4DJ9

**Liste der taxonomischen Gruppen**:

Archaea

Bacteria:

Actinobacteria

Chlorobi

Chloroflexi

Cyanobacteria

Deinococcus

Firmicutes

Planctomycetes

Proteobacteria alpha beta gamma

Thermotoga

Eukaryota

 Alveolata

 Amoebozoa

 Cryptophyta

 Euglenozoa

 Glaucocystophyceae

 Choanoflagellida

 Fungi

 Metazoa

 Porifera

 Placozoa

 Bilateria

 Rhizaria

 Rhodophyta

 Stramenopyles

 Viridiplantae

**Liste der Organismen**

**Bakterien**

*Synechocystis PCC6803*

*Nostoc PCC 7120*

*Trichodesmium erythraeum IMS101*

*Prochlorococcus marinus*

*Escherichia coli*

*Thermotoga maritima*

*Agrobacterium tumefaciens*

*Rhizobium etli*

*Rhodobacter sphaeroides*

*Xanthomonas citri*

*Deinococcus radiodurans*

*Bacillus thurengiensis*

*Bacillus anthracis*

*Lactobacillus johnsonii*

*Streptococcus pneumonia*