Ihre Namen:_____ **Gruppe:**_____

Evolutionsbiologie 2, WS2016/2017: Bioinformatik - Übung 1

Erstellen Sie vor Beginn der Übung einen Ordner auf dem Desktop, in dem Sie alle benötigten Dateien speichern können (z.B. 'Uebung1').

Öffnen Sie die Fasta-Dateien nur mit einem Texteditor, z.B. Wordpad oder Notepad, nicht mit Microsoft Word oder Libre Office.

Teil 1: Dotplots

1.1 Erinnern Sie sich an die Vorlesung zum Thema "Dotplots". Betrachten Sie folgende Sequenz und die dazugehörige Genstruktur (Introns/Exons). Führen Sie nun manuell einen Dotplot der kompletten Gensequenz mit seiner CDS (also der rein exonischer Sequenz) durch und ***skizzieren Sie den Dotplot dazu in das linke Fenster.***

CDS

Gen

1.2 Wofür steht "CDS"?

WEBSITE DOTLET: http://alggen.lsi.upc.es/softpublic/dotlet/Dotlet.html

1.3 Im nächsten Schritt werden Sie Ihre obige Annahme überprüfen und einen Dotplot aus den beiden Sequenzen der Datei „Gen.fasta" erstellen. Öffnen Sie die Internetseite zu Dotlet. Wählen Sie „Input" und geben Sie die Sequenz „Gen" ein (copy-paste der obersten Sequenz ohne den FASTA header). Nennen Sie diese Sequenz „Gen". Dann wählen Sie noch einmal „Input" und wiederholen Sie den Vorgang für die Sequenz „CDS". Wenn Sie nun die beiden Sequenzen vergleichen wollen, müssen beide ausgewählt sein. Stellen Sie die Auflösung (Menüfeld neben „Compute") auf 1:4 und drücken Sie „Compute". ***Entspricht der Dotplot Ihren Erwartungen aus 1.1? Wie unterscheidet er sich?***

1.4 Im Menüfeld neben der Auflösung können Sie die Größe des „sliding window“ einstellen. Verringern Sie die Größe auf 1 und betrachten Sie die Ausgabe. Wiederholen Sie den Prozess mit größer werdenden „sliding windows“ (z.B. 3, 7, 15). ***Können Sie erklären, warum Sie ein klares Signal nur für größere „sliding windows“ sehen?***

1.5 Erstellen Sie als nächstes einen Dotplot mit dem Sequenzset der Datei „TAS14.fasta“ (Auflösung 1:1). Es handelt sich hierbei um zwei nah verwandte Proteinsequenzen. ***Wie würden Sie die zahlreichen parallelen Diagonalen in Ihrem Dotplot interpretieren?***

1.6 Die Datei „paralogs.fasta“ enthält Protein- und Nukleotidsequenzen der paralogen Gene LTP und TSW12. Erstellen Sie zunächst einen Dotplot für die Nukleotidsequenzen und danach für die Proteinsequenzen. Das Signal scheint bei den Proteinsequenzen deutlicher zu sein. ***Können Sie erklären warum? Hinweis: Überlegen Sie was ein „deutliches Signal“ bei Dotplots bedeutet.***

1.7 Beim Dotplot der Nukleotidsequenzen scheint das Signal im hinteren Bereich der Gene zu verschwinden. Dieser Bereich ist in den Sequenzen der 3‘ nicht-kodierende Bereich der Gene. ***Würden Sie ein verschwinden des Signals in diesem Bereich erwarten? Wenn ja, warum?***

1.8 Die letzte Einstellungsmöglichkeit in Dotlet in die Wahl der Substitutionsmatrix. Bei DNA Sequenzen ist diese standardmäßig auf „identity“ gestellt, bei Proteinen auf „BLOSUM62“. ***Erklären Sie wozu diese Matrix bei Dotplots benutzt wird. Hinweis: Vorlesung!***

Teil 2: Alignments

WEBSITE EMBOSS: http://emboss.bioinformatics.nl

2.1 In den folgenden Teilaufgaben werden Sie paarweise Alignments erstellen. Sie verwenden dafür die Programme “needle” und “water” von der Emboss-Seite.
Führen Sie die Alignments mit „needle” UND „water” für das Sequenzset aus „Proteinsequenzen.fasta“ durch. Laden Sie jeweils die 1. Sequenz aus diesem Set in das obere Fenster und ALLE anderen in das untere Fenster. Kopieren Sie dabei die FASTA header mit, dadurch werden Ihre Sequenzen automatisch benannt. Die Programme erstellen dann alle möglichen paarweisen Alignments mit der 1. Sequenz. ***Tragen Sie die Werte für „Identity“, „Similarity“, „Gaps“ und „Length“ für alle drei paarweisen Vergleiche in folgende Tabelle ein.***

NEEDLE	Identity	Similarity	Gaps	Length
TPA_HUMAN - Q9BZW_HUMAN				
TPA_HUMAN - HGFA_MOUSE				
TPA_HUMAN - FAKE				
WATER	Identity	Similarity	Gaps	Length
TPA_HUMAN - Q9BZW_HUMAN				
TPA_HUMAN - HGFA_MOUSE				
TPA_HUMAN - FAKE				

2.2 Was ist der konzeptionelle Unterschied zwischen „Identity“ und „Similarity“?

2.3 Die beiden Programme (oder Algorithmen) „needle" und „water" produzieren „globale" bzw. „lokale" Alignments? ***Erkennen Sie was das bedeutet? Hinweis: Betrachten Sie die Länge der Alignments sowie die Start- und Endkoordinaten der ursprünglichen Sequenzen in den Alignments die von „needle" und „water" produziert werden.***

2.4 Können Sie herausfinden, was die vollständigen Namen für diese Algorithmen sind? (Wir behandeln sie in der nächsten Vorlesung)

2.5 Welche Sequenz ist der 1. Sequenz am ähnlichsten?

.

Teil 3: BLAST

WEBSITE (NCBI) BLAST: http://blast.ncbi.nlm.nih.gov/Blast.cgi

3.1 Wir wollen die Herkunft der DNA Sequenz „Nukleotidsequenz.fasta" mit Hilfe einer BLAST-Suche herausfinden. ***Welches der Programme auf der BLAST Homepage würden Sie dazu verwenden?***

3.2 Führen Sie eine Suche mit der Sequenz durch. Setzen Sie einen Haken bei „Show results in a new window", dadurch können mehrere BLAST-Läufe später miteinander vergleichen. ***Ist die gegebene Sequenz in der Datenbank enthalten? Woran erkennen Sie das?***

3.3 Um was für eine Sequenz handelt es sich (Genname/Spezies)?

3.4 Von welchen drei weiteren Spezies befinden sich homologe Sequenzen mit der größten Ähnlichkeit in der Datenbank? Wie groß ist hier die Sequenzidentität zu Ihrer Originalsequenz?

3.5 In der BLAST Eingabemaske befinden sich oben Reiter die mit Sie unter mehreren Algorithmen auswählen können: blastp, blastn, tblastx, tblastn, blastx
Je nachdem ob Ihre Anfrage (Query) eine Nukleotid- oder Proteinsequenzen ist und ob Sie in Nukleotid- oder Proteindatenbanken nach ähnlichen Sequenzen such wollen muss ein unterschiedlicher Algorithmus benutzt werden. ***Tragen Sie in folgende Tabelle ein, welcher Algorithmus mit welchen Datenbanken bzw. Anfragen (Queries) genutzt werden können.***

		Anfrage (Query)	
		Nukleotid	**Protein**
Datenbank	**Nukleotid**		
	Protein		

3.6 Im 2. Teil dieser Übung werden wir BLAST-Suchen mit einer Nukleotidsequenz und einer Proteinsequenz durchführen. Im Eingabefenster von BLAST können Sie nicht nur Sequenzen, sondern auch „GenBank IDs“ oder „Accession Numbers“ direkt eingeben. ***Führen Sie eine Nukleotid-Nukleotid-BLAST-Suche mit „AY843504.1“ durch. Wie viele Treffer erhalten Sie? HINWEIS: Informationen über die Anzahl finden Sie über der Ergebnis-Grafik. Sind das alle möglichen Treffer?***

3.7 Gehen Sie in der BLAST Eingabemaske auf „Algorithm parameters“. ***Welche Einstellung müssen Sie vornehmen, um die Anzahl an Hits zu erhöhen?***

3.8 Da Sie eine BLAST Suche mit einer Sequenz direkt aus GenBank durchgeführt haben sollte der beste Treffer die Sequenz selbst sein. ***Finden Sie Ihre Eingabesequenz in den Ergebnissen? Entspricht der gefundene BLAST Treffer zu 100% der Eingabesequenz über die komplette Länge?***

3.9 Falls Ihre Eingabesequenz nach dem BLAST nicht zu 100% mit sich selbst übereinstimmt, versuche Sie zu ergründen wieso. Sehen Sie Sich dazu die DNA Sequenz des GenBank Eintrags an. ***Was für potentielle Probleme können Sie dort erkennen?***

3.10 Führen Sie nun eine Protein-Protein-BLAST-Suche mit „AAV91975.1“ durch. Wie viele und welche Domänen hat dieses Protein?

3.11 Was hat dieses Protein mit der DNA Sequenz „AY843504.1“aus 3.6 zu tun? Hinweis: Betrachten Sie die GenBank Einträge des Proteins und der DNA.

3.12 Versuchen Sie als nächstes die zur Protein-Sequenz gehörige Publikation zu finden. ***Wie lautet der Titel und in welchem Journal wurde publiziert?***

3.13 Neben den Informationen die direkt im GenBank Eintrag zu finden sind gibt es auch noch Links zu externen Datenbanken („Related information“ rechts auf der Seite). ***Was ist hier ein PopSet? Wie viele Sequenzen sind im PopSet zur angegebenen Publikation enthalten?***

Teil 4: Proteindomänen

WEBSITE PDB: http://www.rcsb.org/pdb

4.1 Die PDB Datenbank enthält 3D-Strukturen von Proteinen und Proteindomänen, die meist durch Röntgenkristallographie erlangt wurden. Finden Sie den Eintrag 1CE9. Betrachten Sie die Übersichtsseite „Structure Summary“. ***Aus welchem Organismus stammt diese Struktur?***

4.2 Wählen Sie nun die Registerkarte „Sequence”. Falls Proteinstrukturen dadurch entstehen, dass mehrere Aminosäureketten miteinander interagieren, dann besitzt diese Struktur mehrere „chains“. Diese können in diesem Reiter einzeln betrachtet werden. ***Aus wie vielen chains besteht diese Proteinstruktur? Unterscheiden sich die chains voneinander?***

4.3 Betrachten Sie die Struktur Ihrer chains unter „Sequence chain view“. ***Was für Strukturmotive kommen in den chains vor? Erkennen Sie die chains in der 3D-Strukur unter „Structure Summary“?***

4.4 Laden Sie unter „Download Files” die Fasta-Datei mit den Einträgen zu 1CE9 herunter. Wir werden die Proteinsequenz der chains nun an die Sequenz des Proteins alignen. Verwenden Sie dazu das „needle“ Programm aus Übung 2.1: In das obere Fenster geben sie die Sequenz „unbekannt.fasta“ (komplettes Protein) ein, in das untere Fenster die gesamte Datei, die Sie gerade herunter geladen haben. ***Welcher Teil des Proteins bildet die 3D-Struktur aus?***

Teil 5: Bioinformatische Vorhersage von Proteineigenschaften

Für diese Übung werden Sie mit vier verschiedenen Proteinsequenzen arbeiten: „protein_1.fasta“, „protein_2.fasta“, „protein_3.fasta“ und „protein_4.fasta“. Ziel ist es, Informationen über die strukturellen Eigenschaften dieser Proteine zu erhalten. Hierzu benutzen wir Programme, die versuchen aus der Aminosäuresequenz des Proteins verschiedene Eigenschaften abzuleiten.

Auf der letzten Seite des Übungsblattes finden Sie eine grafische Repräsentation dieser vier Proteine. Für jedes dieser Proteine, sollen Sie Informationen zu Proteindomänen, Sekundärstruktur, Hydrophobizität und coiled-coil-Strukturen bestimmen und in der Grafik einzeichnen. Sie müssen bei Ihren Zeichnungen nicht exakt sein. Ziel dieser Aufgabe ist es, einen Überblick über Gemeinsamkeiten und Unterschiede dieser Proteine zu erhalten.

5.1 Vorhersage von Proteindomänen

WEBSITE HMMER: http://www.ebi.ac.uk/Tools/hmmer/search/hmmscan

Für jedes dieser Proteine führen Sie nun eine Vorhersage von Proteindomänen durch. Gehen Sie dafür auf die HMMER Internetseite und geben die Sequenz in das Suchfenster ein (der FASTA header kann mitkopiert werden). Betrachten Sie die grafische Ausgabe und tragen Sie die Domänen in das Arbeitsblatt ein. Verwenden Sie dazu die Pfam-Annotation.

5.2 Vorhersage der Sekundärstruktur

WEBSITE NPS@:
https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_server.html

Als nächstes werden Sie eine Strukturvorhersage für alle Proteine durchführen. Nutzen Sie die NPS@ Internetseite. Gehen Sie unter „Secondary structure prediction” auf „Secondary structure consensus prediction” und kopieren Sie die Sequenzen einzeln in das Eingabefeld. Für diesen Schritt dürfen Sie nur die reinen Sequenzen verwenden, kopieren Sie die Sequenzen NICHT mit der Fasta-ID („>...”). Zeichnen Sie in das Arbeitsblatt ein einfaches Schema, das zeigt, wo α-Helices (h) und erweiterte β-Stränge (Faltblätter) (e) in den Proteinen vorhanden sind. Benutzen Sie dazu die Konsensus-Vorhersage („Sec.cons.“).

5.3 Vorhersage von Hydrophobizität

WEBSITE WHAT 2: http://www.tcdb.org/progs/?tool=hydro

Nun betrachten wir die Hydrophobizität (auch Hydropathie) der vier Proteine. Diese Eigenschaft ist besonders wichtig, da längere hydrophobe Abschnitte auf einen Transmembranbereich hinweisen können. Benutzen Sie hierzu die Website What 2 und kopieren Sie die Proteinsequenzen. Potentielle hydrophobe Bereiche werden im Ausgabefenster mit orangen Balken farbig markiert. Gibt es Proteine, die möglicherweise einen Transmembranbereich aufweisen? Markieren Sie derartige hydrophobe Bereiche im Arbeitsblatt.

5.4 Vorhersage von coiled-coil-Strukturen

WEBSITE COILS: http://www.ch.embnet.org/software/COILS_form.html

Coiled-coils sind wichtige strukturelle Motive die u.a. bei der Protein-Dimerisierung eine Rolle spielen. Meist bestehen diese aus ineinander verwobenen α-Helices (siehe auch 5.2). Zur Vorhersage benutzen Sie das Programm COILS. Kopieren Sie die Proteinsequenzen und lassen Sie das Programm für alle vier Proteine laufen. Der Output gibt Ihnen die Wahrscheinlichkeit, dass in einem Bereich eine coiled-coil Struktur vorhanden ist. Markieren Sie auf dem Arbeitsblatt die Bereiche, in denen coiled-coils mit hoher Wahrscheinlichkeit (>50%) vorkommen.

5.5 Die Strukturen die Sie bis jetzt erhalten haben waren nur Vorhersagen von Algorithmen. Wir wollen diese nun mit den Tatsächlichen Strukturen in der PDB Datenbank vergleichen. Der in der Fasta-Datei angegebene Name (z. B. 1HGA) ist eine sogenannte PDB-ID. Sollte nach dem Namen noch ein „:“ stehen (z. B. 1HGA:A), so bezeichnet dies den chain der Struktur aus der die Sequenz stammt. Gehen Sie auf die PDB Internetseite und suchen Sie den passenden Eintrag für jedes Protein. ***Welche Proteine haben Sie betrachtet? Vergleichen Sie die tatsächlichen Sekundärstruktur-Motive (unter „Sequence“ in PDB) mit den Vorhersagen. Waren die Vorhersagen zuverlässig? Hinweis: die Struktur in der Datenbank zu 1CE9 deckt nur einen Teil der benutzten Proteinsequenz ab. Es ist dieselbe Sequenz wie in Übung 4.4.***

1HGA	Proteinname: Vorhersage:
1PQ7	Proteinname: Vorhersage:
1BE3	Proteinname: Vorhersage:
1CE9	Proteinname: Vorhersage:

Teil 6: Genomdatenbanken

WEBSITE FLYBASE: http://www.flybase.org

Organismen für die die Sequenz des kompletten Genoms bekannt ist haben oftmals eigene Datenbanken in denen die Sequenz hinterlegt ist, inklusive eines „Genome Browsers“, in dem das Genom und sein Inhalt graphisch dargestellt wird. Für *Drosophila melanogaster* ist dies FlyBase. Öffnen Sie die Website. Wir wollen Sie benutzen um Informationen über ein paar Gene zu sammeln.

6.1 Viele Gene in *Drosophila* wurden nach Phänotypen benannt, die eine Mutation im jeweiligen Gen auslöst. Suchen Sie unter „Jump to Gene“ das Gen das den Phänotyp der “rosy” (ry) Mutante bei *Drosophila* ausmacht. Sie erhalten eine Übersicht über alle bekannten Informationen zu diesem Gen. ***Zu was für einer Familie von Enzymen gehört dieses Gen?***

6.2 Wie lautet die Annotationsnummer (annotation symbol) des rosy Gens?

6.3 Auf welchem Chromosomenarm befindet sich das rosy Gen? Hinweis: Wenn Sie Sich unter „Genomic Maps -> GBrowse“ die chromosomale Lage des Gens darstellen lassen.

6.4 Mit der „Get FastA“ Funktion können Sie Sich die Sequenzabschnitte von bestimmten Bereichen des Gens herunterlade***. Benutzen Sie dies um herauszufinden wie viele Introns das rosy Gen hat.***

6.4 Wie viele Aminosäuren ist das von rosy kodierte Protein lang? Sie können wieder die „Get FastA“ Funktion benutzen, versuchen Sie die Information aber auch im Unterbereich „Gene model and products“ zu finden.

6.5 Ein weiteres Gen im Drosophila Genom ist *disco*. ***Wie lautet der vollständige Name des Gens und auf welchem Chromosom liegt es?***

6.6 Betrachten Sie den Unterbereich Gene Ontology (GO). ***Was ist die molekulare Funktion des disco Proteins? In welchem Teil der Zelle ist es aktiv? Passt das mit der molekularen Funktion zusammen? Warum oder warum nicht?***

6.7 Betrachten sie nun das Gen *Frost*. Suchen den Unterbereich in dem Ortologe dieses Gens in anderen Arten aufgelistet werden. ***Können Sie Orthologe in anderen Drosophila Spezies finden? Wenn ja, in welchen?***

Löschen Sie bitte nach Beendigung der Übung alle heruntergeladenen und erstellten Dateien bevor Sie den Computer herunterfahren.

Sequenz 1

	N-terminus	*C-terminus*
Domänen		
Sekundärstruktur		
Hydrophobizität		
Coiled coils		

Sequenz 2

	N-terminus	*C-terminus*
Domänen		
Sekundärstruktur		
Hydrophobizität		
Coiled coils		

Sequenz 3

	N-terminus	*C-terminus*
Domänen		
Sekundärstruktur		
Hydrophobizität		
Coiled coils		

Sequenz 4

	N-terminus	*C-terminus*
Domänen		
Sekundärstruktur		
Hydrophobizität		
Coiled coils		