

Übungen zur Vorlesung
**Einführung in die angewandte
Bioinformatik**
Sommersemester 2009

Übungsblatt 6
Bearbeitungszeit:
28.05.2009

Aufgabe 6.1 – Dotlet

Wenn Sie diese Aufgabe in den Poolräumen bearbeiten, benutzen Sie bitte den Firefox.

Sie sollen in dieser Aufgabe einen Dotplot der Aminosäuresequenz des Proteins *Zinc finger protein 492* mit sich selbst erstellen, um herauszufinden, welche interessante Eigenschaft dieses Protein hat.

Nutzen Sie auf der UniProt-Webseite die Zugriffsnummer Q9P255, um sich die Aminosäuresequenz des Proteins im FASTA-Format anzuzeigen. Kopieren Sie den Inhalt in die Zwischenablage (lassen Sie dabei die Kommentarzeile weg). Gehen Sie auf die Dotlet-Webseite, klicken Sie erst auf „input“ und fügen Sie dann die Aminosäuresequenz ein. Wenn Sie es schön haben wollen, geben Sie der Sequenz noch einen passenden Namen. Klicken Sie auf „compute“ und stellen Sie die Regler über und unter dem Histogramm so ein, dass die interessanten Stellen des Plots gut sichtbar sind. Wie interpretieren Sie das Bild, das sich Ihnen bietet?

Analysieren Sie jetzt auch noch das Protein mit der Zugriffsnummer P21997. Hierbei bietet es sich an, die „window size“ (drittletzter Menüpunkt) von 15 auf 1 zu ändern. Wie interpretieren Sie diesen Dotplot?

Aufgabe 6.2 – Globales Alignment

Suchen Sie in Ihrer Lieblingsdatenbank nach dem Pax6-Gen im Menschen. Laden Sie sich die Proteinsequenzen der beiden Genproduktvarianten (Isoform a und Isoform b) im FASTA-Format herunter. (Zur Kontrolle: Die RefSeq-Einträge haben die Zugriffsnummern NP_000271.1 und NP_001595.2.) Schauen Sie sich zunächst die Sequenzen einfach nur an und versuchen Sie herauszufinden, welche Aminosäuresequenz zum Exon gehört, das in einer Isoform fehlt. Nachdem Sie festgestellt haben, dass dies etwas mühselig ist, erinnern Sie sich daran, dass Sie diese Information viel einfacher bekommen können, indem Sie ein globales Alignment zwischen den beiden Sequenzen durchführen lassen. Gehen Sie daher auf die EMBOSS-Align-Webseite. Wählen Sie die Needleman-Wunsch-Methode (für globales Alignment). Falls nicht schon voreingestellt, wählen Sie als Strafen für *Gap Open* und *Gap Extend* 10.0 bzw. 0.5, als Score-Matrix BLOSUM62, und natürlich die richtige Molekülart aus. Laden Sie nun die beiden Sequenzen hoch und starten den Vergleich.

Wie lauten also die ersten und letzten drei Aminosäuren des fehlenden Exons? _____

Wie hoch ist der Score des Alignments? _____

Aufgabe 6.3 – Lokales Alignment

Blieben Sie auf der EMBOSS-Align-Webseite, aber lassen Sie nun ein lokales Alignment mit Smith-Waterman berechnen.

Nehmen Sie an, Sie hätten ein DNA-Fragment sequenziert, das Sie als FASTA-Datei `fragment.fasta` auf der Übungs-Webseite finden. Sie vermuten, dass es aus einem Gen stammt, dessen mRNA-Sequenz Sie mit der Zugriffsnummer NM_007808.4 aus einer Datenbank bekommen können. Versuchen Sie, das Fragment mit lokalem Alignment in diesem Gen zu lokalisieren. Denken Sie daran, den *Molecule*-Parameter geeignet zu ändern, aber lassen Sie die restlichen Parameter wie voreingestellt. Wie hoch ist der Score dieses Alignments? _____ An welcher Position der Gensequenz aus der Datenbank beginnt

das Alignment? _____ Wurde Ihre Vermutung bzgl. der Herkunft des Fragments bestätigt? _____

Aufgabe 6.4 – BLAST bei NCBI

Tipp: Aktivieren Sie auf der BLAST-Seite das Häkchen bei „Show results in a new window“.

Führen Sie ein Protein-BLAST auf der NCBI-Seite durch. Geben Sie die Zugriffsnummer P00393 ein. Suchen Sie in der non-redundant-protein-Datenbank. Sehen Sie sich zunächst in Ruhe das Ergebnis an (es kann ein paar Sekunden dauern, bis es erscheint). Sehen Sie sich dann den ersten Treffer an. Wie hoch ist die

Sequenzidentität? _____ Wiederholen Sie die Suche, schränken sie dabei aber nun auf den Organismus *Gallus gallus* ein. Betrachten Sie das Alignment für den ersten Treffer. Wie lang ist die Lücke, die in der Query-Sequenz ca. an Position 215 beginnt? _____

Wiederholen Sie auch diese eingeschränkte Suche, aber mit geänderten Parametern. Klicken Sie dazu unten auf *Algorithm parameters*. Nachdem Sie freudig festgestellt haben, dass Sie viele dieser Parameter aus der Vorlesung kennen, ändern Sie die Kosten für Lücken auf **Existence: 7 Extension: 2**. Sehen Sie sich wieder den ersten Treffer an und vergewissern Sie sich, dass es sich um dasselbe Protein wie bei der letzten

Suche handelt. Finden Sie die Lücke aus der letzten Suche wieder? _____
Wie erklären Sie sich das?

Aufgabe 6.5 – BLAST mit beschränkter Ausgabe

BLASTen Sie das Protein MJ0577 (das ist *nicht* die Zugriffsnr., sondern der Proteinname) aus dem Organismus *Methanocaldococcus jannaschii* gegen die non-redundant-Datenbank. Schränken Sie die Ausgabe auf die

10 ähnlichsten Sequenzen ein. Wie hoch ist der E-Value des schlechtesten Treffers? _____
Wiederholen Sie die Suche, schränken Sie aber die Ausgabe auf Treffer mit einem E-Value besser als 10^{-20} ein. Wie viele Treffer bleiben übrig? _____

Aufgabe 6.6 – Gensuche

Laden Sie sich von der Übungswebseite die FASTA-Datei *unbekannt.fasta* herunter, die z.B. wieder ein sequenziertes Fragment sein könnte. Finden Sie mittels Nukleotid-BLAST gegen die Genomes-Datenbank den Organismus, aus dem diese Sequenz wahrscheinlich stammt. Schauen Sie sich den ersten Treffer an:

Wie groß ist die Zahl der Nukleotide, die übereinstimmen? _____

Aus welchem Organismus stammt die Sequenz? _____

Aufgabe 6.7 – Verschiedene Matrizen

In dieser Aufgabe geht es um die Auswirkungen verschiedener BLOSUM-Matrizen auf die Ergebnisse einer Protein-BLAST-Suche. BLASTen Sie dazu nach der Sequenz mit der RefSeq-Zugriffsnummer NP_727165 (Isoform A des *sex-lethal*-Gens aus *Drosophila melanogaster*). Stellen Sie folgende Parameter ein: Suchen Sie nur in der SwissProt-Datenbank, schränken Sie die Suche auf *Escherichia coli* ein, stellen Sie den maximalen E-Value auf 15 ein, stellen Sie die *compositional adjustments* auf *no adjustment* und aktivieren Sie den *low-complexity*-Filter. Lassen Sie den Score-Matrix-Parameter zunächst unverändert (die BLOSUM62-Matrix sollte voreingestellt sein) und wiederholen Sie die Suche dann mit der BLOSUM45-Matrix.

Vergleichen Sie die beiden Suchergebnisse. Wie lauten die Zugriffsnummern und E-Values der beiden jeweils ersten Treffer?

BLOSUM62: _____

BLOSUM45: _____

Welche gemeinsamen Treffer finden Sie in beiden Ergebnissen?
