



Vorlesung: Einführung in die Bioinformatik Sommersemester 2014

Übungsblatt Nr. 1

Elektronische Abgabe an: Joachim.Fritscher, per Email an joachim.fritscher@student.uni-tuebingen.de

1. Dotplots

- (a) Erläutern Sie kurz, was ein Dotplot im Zusammenhang von Sequenzvergleichen ist.
- (b) Beispiel per Hand. Gegeben seien die folgenden Sequenzen:

```
SEQ_A   G H R Q S G G
SEQ_B   S G G R G
```

Berechnen Sie mit Papier und Bleistift einen Dotplot D . Es gelte folgende Bedingung für das Füllen der Matrix D :

$$D[i, j] = \begin{cases} 1 & \text{falls } SEQ_A[i] = SEQ_B[j] \\ 0 & \text{sonst.} \end{cases}$$

Zeichnen Sie D für die obigen 2 Sequenzen und verwenden Sie statt einer 1 einen ‘•’ und statt der 0 einen ‘white space’. Was bedeuten Diagonalen in der Matrix?

- (c) Für folgende Übung verwenden Sie das Java Wbestart-basierte Tool JDotter (<http://athena.bioc.uvic.ca/virology-ca-tools/jdotter/>, Achtung erst registrieren!, benötigt Java!):

Übergeben Sie die folgende Sequenz per copy and paste an das Applet:

```
PrP-Human (P04156)
MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPHGGGGWGQPHG
GGWGQPHGGGGWGQPHGGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGAVVGGLGGYMLGSAMSRPIIH
FGSDYEDRYRENHRYPNQVYRPMDEYSNQNNFVHDCVNITIKQHTVTTTTKGENFTETDVKMMERVV
EQMCITQYERESQAYYQRGSSMVLFSPPVILLISFLIFLIVG
```

Vergleichen Sie die Sequenz PrP-Human mit sich selbst. Wie interpretieren Sie das Muster? Verändern Sie in JDotter Window size und Zoom Factor, derart, dass identische Regionen gut sichtbar werden. Stellen Sie durch Navigieren mit der Maus fest, an welchen Positionen innerhalb der Sequenz ähnliche Teilsequenzen liegen. Hier finden Sie den zur Sequenz zugehörigen Datenbankeintrag: <http://www.ncbi.nlm.nih.gov/protein/P04156> Lesen Sie nach, ob bzw. wie die von Ihnen identifizierten Regionen im Abschnitt ‘Features’ aufgeführt und annotiert sind. Reichen Sie einen exportierten Plot des Dotplots mit ein.

2. BLAST

In dieser Aufgabe wollen wir BLAST einsetzen, um homologe Sequenzen für das folgende gegebene humane Gen in der Datenbank GenBank zu finden. BLAST ist unter der folgenden URL als Webservice zu verwenden: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>

Das Referenzgen besitzt die GI-Nummer 194353959 (GI=gene identifier) bzw. Refseq ID NP_004561. Finden Sie zunächst unter <http://http://www.ncbi.nlm.nih.gov/protein> heraus, um welches Gen es sich handelt.

Wenden Sie dann ‘protein blast’ (blastp) mit der obigen ID als Eingabe (sogenannte ‘query’ im BLAST-Jargon) an. Wählen Sie als Datenbank „Reference proteins (refseq_protein)“ aus. Wählen Sie als Organismen die folgenden aus (m.a.W. beschränken Sie die Suche auf die folgenden Organismen):

- *Mus musculus*
- *Bos taurus*
- *Rattus norvegicus*
- *Macaca mulatta*
- *Xenopus tropicalis*
- *Equus caballus*
- *Canis familiaris*
- *Monodelphis domestica*
- *Ornithorhynchus anatinus*

Stellen Sie eine Tabelle auf mit den oben aufgeführten lateinischen Organismennamen und zugehörigen gebräuchlichen Namen, sowie dem Ergebnis Ihrer Blastsuche. Versuchen Sie herauszubekommen, wie Sie anhand der von BLAST ausgegebenen Parameter, u.a.

- Bit-Score
- Query Coverage
- *E*-Value

entscheiden, welcher der Treffer in der Datenbank Ihrer eingegebenen Sequenz am Ähnlichsten ist. Versuchen Sie das Ergebnis kurz zu interpretieren.