



Vorlesung: Einführung in die Bioinformatik Sommersemester 2013

Übungsblatt Nr. 1

Elektronische Abgabe an: Benjamin Schubert: schubert@informatik.uni-tuebingen.de

1. Dotplots

- Erläutern Sie kurz, was ein Dotplot ist.
- 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 ‘•’ und für eine 0 ‘white space’. Was bedeuten Diagonalen in der Matrix?

- Die folgende Übung ist dem Buch ”Bioinformatik Interaktiv” (Merkel und Waack) entnommen. Für diese verwenden Sie das Tool Dotlet (von M. Pagni und T. Junier), welches Sie hier direkt als Java Applet verwenden können:

<http://www.ikp.unibe.ch/dotlet/dotlet.html>

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

SLIT_DROME (P24014)

```
MAAPSRTTLMPPPFRLQLRLLILPILLLLRLHDAVHAEPYSGGFGSSAVSSGGLGSVGIHIPGGGVGVITEARCPVCSCT
GLNVDCSHRGLTSVPRKISADVERLELQGNNTLVIYETDFQRLTKLRLQLTDNQIHTIERNFSQDLVSLERLDISNNVI
TTVGRRVFKGAQSLRSLQLDNNQITCLDEHAFKGLVELEILTNNNNLTSLPHNIFGGLGRLRALRLSDNPFACDCHLSW
LSRFLRSATRLAPYTRCQSPSQLKGQNVADLHDQEFKCSGLTEHAPMECGAENSCPHPCRCADGIVDCREKSLTSVPVTL
PDDTTDVRLEQNFITELPPKSFSSFRRLRRIDLSNNNISRIAHDALSGLKQLTTLVLVYGNKIKDLP SGVFKGLGSLRLLL
LNANEISCIRKDAFRDLHSLSLSLYDNNIQLSLANGTFDAMKSMKTVHLAKNPFI CDCNLRWLADYLHKNPIETSGARCE
SPKRMHRRRIESLREEKFKCSWGELRMKLSGECRMDSDCPAMHCEGTTVDCTGRRLKEIPRDIPLHTTELLLDNELGR
ISSDGLFGRPLHLVKLELKRNLGTGIEPNAFEGASHIQELQLGENKIKEISNKMFLGLHQLKTLNLYDNQISCVMPGSFE
HLNSLTSLNLASNPFCNCHLAWFAECVRKKS LNNGAARCGAPSKVRDVQIKDLPHSEFKCSSENSEGLGDGYCPPSCT
CTGTVVACSRNQLKEIPRGIPAETSELYLESNEIEQIHYERIRHLRSLTRLDLSNNQITILSNYTFANLTKLSTLIISYN
KLQCLQRHALSGLNNLRVSLHGNRISMLPEGSFEDLKSLTHIALGSNPLYCDCGLKWFSDWIKLDYVEPGIARCAEPEQ
MKDKLILSTPSSSFVCRGRVRNDILAKCNACFEQPCQNQAQCVALPQREYQCLCQPGYHGKHCFEMIDACYGNPCRNNAT
CTVLEEGRFSCQCAPGYTGARCETNIDDCLEIKCQNNATCIDGVESYKCECQPGFSGEFCDTKIQFCSPEFNPCANGAK
CMDHFTHYSCDCQAGFHGTNCTDNIDDCQNHMCQNGGTCVDGINDYQCRCPDDYTGKYCEGHNMISMYPQTSPCQNEHC
KHGVCFPQNAQGS DYLCRCHPGYTGKWCEYLTSISFVHNNSFVELEPLRTRPEANVTIVFSSAEQNGILMYDGQDAHLAV
ELFNIRVSVDGNHPVSTMYSFEMVADGKYHAVELLA IKKNFTLRVDRGLARSIINEGSNDYKLKLTTPMFLGGLPVDP
AQQAYKNWQIRNLTSFKGCMKEVWINHKLVD FGNARQQKITPGCALLEGEEQEEEDDEQDFMDETPHIKEEPVDP CLEN
KRRGRSVCVPNSNARDGYQCKCKHGQRGRYCDQEGESTEPPTVTAASCTRKEQVREYYTENDCRSRQPLKYAKCVGGCGN
QCCAIVRRRKVRMVCSNNRKYIKNLDIVRKC GCTKKCY
```

Vergleichen Sie die Sequenz SLIT_DROME mit sich selbst. Wie interpretieren Sie das Muster? Stellen Sie den Zoom-Faktor auf 1:5. Justieren Sie die Regler 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/P24014> Lesen Sie nach, ob bzw. wie die von Ihnen identifizierten Regionen im Abschnitt Features aufgeführt und annotiert sind.

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 4557818 (GI=gene identifier). Finden Sie zunächst unter <http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide> heraus, um welches Gen es sich handelt.

Wenden Sie dann ‘nucleotide blast’ (blastn) mit der obigen GI-Nummer als query-Sequenz an. Wählen Sie als Datenbank „Reference mRNA sequences (refseq_rna)“ 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

Versuchen Sie herauszubekommen, wie Sie anhand der von BLAST ausgegebenen Parameter, u.a.

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

entscheiden, welcher der Treffer der eingegebenen Sequenz am Ähnlichsten ist.