

Übungen zur Vorlesung
Einführung in die Bioinformatik
Sommersemester 2007

1 Einführung

Sehen Sie sich die in der Vorlesung vorgestellten Webseiten (insbesondere die Übersichtsseiten) an, um einen Eindruck der dort verfügbaren Informationen zu bekommen.

2 Biologische und Literatur-Datenbanken

1. Machen Sie sich mit den Datenbanken des National Center for Biotechnology Information (NCBI) vertraut.
2. Finden Sie in PubMed den Artikel über das in der Vorlesung vorgestellte DNA Computing Verfahren von Adleman. Wo ist der Artikel erschienen? Machen Sie sich mit MeSH Database vertraut (auf der Startseite finden Sie auch kurze Einleitungsvideos, die Ihnen weitere Hilfestellungen geben). Sehen Sie sich den Unterschied zur allgemeinen Keywordsuche in PubMed an: Suchen Sie in PubMed ohne und mit MeSH nach outbreaks von AIDS, um Informationen über HIV-Outbreaks zu erhalten. Wie viele Treffer erhalten Sie jeweils? Warum? Suchen Sie auch im Web of Science nach passenden Artikeln.
3. Machen Sie sich mit den Datenbanken bei NAR (<http://www3.oup.co.uk/nar/database/c/>) vertraut. Wie viele metabolic pathway databases sind dort verlinkt? Schauen Sie auch zum Vergleich unter <http://www.pathguide.org/> nach. Folgen Sie auch einigen Links auf Pathway DBs, um einen ersten Eindruck dieser Datenbanken zu erhalten.
4. Finden Sie mittels Entrez die mRNA- und Proteinsequenz für das Gen Pax6 in Mensch und Maus. Finden Sie heraus, wie man in OMIM eine Liste aller Einträge über Gene auf Chromosome 3 erhält. Wie viele Einträge gibt es?
5. Öffnen Sie das Interface zu SRS und machen Sie sich mit dem Layout der Seite vertraut. Suche in SRS (Quick Search vs. Standard Query Form): Suchen Sie nach der Proteinsequenz von Barnase, schränken Sie dann unter Verwendung der Standard Query Form die Suche auf den *Bacillus intermedius* ein. Beachten Sie: vor der Suche muss zunächst einmal eine (oder mehrere) Datenbank(en) ausgewählt werden. Machen Sie sich mit der Ausgabe vertraut. Suchen Sie die Proteinsequenz von Cytochrome c in dem Hefe Organismus. Suchen Sie in SRS nach dem Protein zum Gen APX1 aus dem Organismus *Arabidopsis thaliana*. Wie lautet die Accession Number?
6. Versuchen Sie, mittels SRS das Gen zum Protein ABFA aus dem Organismus *Bacillus subtilis* zu finden. Finden Sie dann einen Eintrag aus der Nukleotidsequenz-Datenbank EMBL zu diesem Gen. Suchen Sie in SRS die Proteinsequenz mit der Accession Number P00042.

3 Sequenzanalyse und BLAST

1. Machen Sie sich mit den in der Vorlesung vorgestellten BLAST Versionen vertraut. Wählen Sie beispielsweise eine beliebige Proteinsequenz und blasten Sie gegen eine DNA-Sequenzdatenbank.
2. Suchen den besten Match in *C. elegans* des Proteins NTF1 aus *D. melanogaster*. (Das Protein NTF1 ist ein Transkriptionsfaktor, verwenden Sie die Isoform A.) Wie gut sind score und e-Value?
3. Suchen Sie mit einer Sequenz im FASTA Format. Um eine Sequenz einzusetzen, finden Sie in SRS die Proteinsequenz mit der Acc-Nummer P00042. Kopieren Sie die Sequenz und fügen Sie diese in das Feld Search ein. Wählen Sie nun eine Datenbank aus und starten Sie die Suche. Limitieren Sie die vorherige Suche auf Sequenzen im Menschen. Wie gut sind hier der/die gefundene(n) Treffer (score, e-Value)?
4. Verändern Sie einige Werte (Parameter) für die Suche aus der vorherigen Aufgabe, z.B. Matrix und Expect, um die Funktion dieser Optionen genau zu verstehen. Sehen Sie sich auch andere Optionen für die Ergebnisausgabe an.
5. Finden Sie die 10 homologsten Sequenzen zu dem Protein MJ0577 aus dem Organismus *Methanococcus jannaschii*. Wenn ihr e-Value besser als e^{-20} sein soll, wie viele Sequenzen bleiben als Treffer übrig?
6. Versuchen Sie die homologsten Sequenzen zu dem Protein mit SRS Access-Nr. P00042 in der PRINTS Datenbank zu finden. Welche Einträge finden Sie? Was enthält die PRINTS Datenbank?

4 Phylogenetik

1. Sehen Sie sich die in der Vorlesung vorgestellten Webseiten (insbesondere die Übersichtsseite und das Tree of Life Project) an, um einen Eindruck der dort verfügbaren Informationen zu bekommen.
2. Verwenden Sie die online-Version von Phylip, um einen phylogenetischen Baum der in der Datei `five_species.fasta.txt` (siehe Webseite) gegebenen Sequenzen zu berechnen. Verwenden Sie beispielsweise `protdist` (distanzbasierte Methode) und zeichnen Sie den Baum anschließend mit `drawtree` und `drawgram`. Sehen Sie sich den Effekt von verschiedenen Methoden zur Berechnung eines Baums aus der Distanzmatrix an, vergleichen Sie die Bäume.

5 Proteinsequenzen und Proteinstruktur

1. Sehen Sie sich die in der Vorlesung vorgestellten Webseiten und Visualisierungswerkzeuge an.
2. Suchen Sie in SRS die UniProt Accession Number des Enzymes 1.1.1.25 (EC Nummer) in *Arabidopsis*. Verwenden Sie ExPASy, um einige Eigenschaften des Proteins vorherzusagen. Was ist das vorhergesagte Molekulargewicht?
3. Finden und visualisieren Sie das oben gefundene Protein mit drei verschiedenen Tools der PDB Datenbank. Probieren Sie die in die Webseite eingebetteten Werkzeuge Jmol, WebMol und King.

4. Verwenden Sie die CATH Datenbank, um einen Eindruck vom Aufbau des Proteins mit der PDB ID 1FUP zu erhalten. Wählen Sie eine Domain aus und schauen Sie sich die zugehörige Hierarchie an.

6 Statistische Analyse mit R

1. Sehen Sie sich die in der Vorlesung vorgestellten Webseiten an, um einen Eindruck der dort verfügbaren Informationen zu bekommen.
2. Machen Sie sich mit R vertraut. Gehen Sie entsprechend des Beispiels am Ende des Übungszettels vor, probieren Sie aber auch ein paar eigene Sachen aus, um ein Gefühl für die Programmierung in R zu erhalten.

7 Biologische Netzwerke, Netzwerk-Datenbanken, Modellierung und Simulation

1. Sehen Sie sich die in der Vorlesung vorgestellten Webseiten (insbesondere KEGG und EcoCyc) an, um einen Eindruck der dort verfügbaren Informationen zu bekommen. Wie ist der Pathway für die Synthese von Alanin in den verschiedenen Datenbanken repräsentiert, welche Informationen liefern die einzelnen DBs? Welche Pathways finden sich hauptsächlich in TransPATH?
2. Informieren Sie sich auf der SBML Seite (<http://sbml.org>), welche Werkzeuge zur Simulation von biologischen Prozessen/Netzwerken das Austauschformat SBML unterstützen. Finden Sie wenigstens drei.

Anhang: Projekt für R

Starten Sie R mit den folgenden Anweisungen:

```
% module add r  
% R
```

R enthält standardmäßig den Datensatz "Iris". Teilen Sie R mit, dass Sie ihn verwenden wollen, und lassen Sie sich diesen Datensatz ausgeben.

```
> data(iris)  
> iris
```

Wieviele Attribute gibt es? Versuchen Sie im Internet herauszufinden, woher dieser Datensatz stammt.

Lassen Sie sich im nächsten Schritt einen Boxplot für jedes Attribut ausgeben. Beispiel: Einen Boxplot für Petal.Length erzeugen Sie so:

```
> boxplot(iris$Petal.Length)
```

Was sehen Sie?

Visualisieren Sie nun jedes der Attribute mit einem Histogramm. Beispiel: Ein Histogramm für `Petal.Length` erzeugen Sie so:

```
> hist(iris$Petal.Length,main="Histogram")
```

Lassen Sie sich jetzt einen Scatterplot anzeigen, der jeweils ein Paar von Attributen gegeneinander abbildet. Beispiel:

```
> plot(iris$Petal.Length, iris$Petal.Width)
```

Erkennen Sie Abhängigkeiten zwischen den Attributen?

Mit

```
> plot(iris)
```

können Sie sich Scatterplots von allen Paaren anzeigen lassen.

Zwischen welchen der Attributen herrscht der stärkste Zusammenhang? Prüfen Sie dies zusätzlich, indem Sie die Korrelation zwischen zwei Attributen berechnen. Beispielsweise:

```
> cor(iris$Petal.Length, iris$Sepal.Width)
```

In einem nächsten Schritt geht es darum, den Zusammenhang der Attribute `Petal.Length`, `Petal.Width`, `Sepal.Width` und `Sepal.Length` zur Spezies herauszufinden.

Mit dem Befehl

```
> plot(iris[,1:4], col = as.numeric(iris$Species))
```

können Sie sich einen Scatterplot zwischen diesen Attributen anzeigen lassen, wobei die Spezies eingefärbt ist. Welche Attributpaare eignen sich gut, um die Spezies voneinander zu trennen, welche eignen sich nicht so gut?