

Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2016/17

Martin Vingron · Annalisa Marsico · Alena van Bömmel · Edgar Steiger · Thimo Wellner

Blatt 15 · Ausgabe am 6.2.2017

Abgabe am 13.2.2017 vor Beginn der Vorlesung

Name:

Matrikelnummer:

Übungsgruppe:

Aufgabe 1 (*40 Punkte; Programmieren*). Auf dem letzten Übungsblatt haben Sie ein Programm geschrieben, das zwei PDB-Dateien einliest, translatiert und die RMSD berechnet. Nun möchten wir die zwei Proteine so rotieren, dass sie optimal übereinander liegen.

1. Die Rotation eines Punktes um einen bestimmten Winkel kann mit Hilfe einer Rotationsmatrix erreicht werden. In 3D kann man diese Rotationsmatrix durch drei aufeinanderfolgende Rotationen (also 3 Rotationsmatrizen) um jeweils eine der drei Achsen beschreiben. Geben Sie diese 3 Rotationsmatrizen an.
2. Erweitern Sie Ihr Programm um eine 3D-Rotation. Implementieren Sie dazu eine Funktion, die ein Protein um alle 3 Achsen um jeweils denselben Winkel rotiert.
3. Wenden Sie nun Ihre Funktion auf ein bereits mithilfe des Massezentrums translatiertes Protein vom letzten Übungsblatt an und rotieren Sie die Struktur schrittweise um 20 Grad. Berechnen Sie nach jeder Rotation die RMSD zur anderen translatierten (aber nicht rotierten) Struktur und geben Sie sie aus. Speichern Sie am Ende die Positionen der Atome mit der kleinsten RMSD.
4. Visualisieren Sie die Originalpositionen und die translatierten und rotierten Positionen der C α -Atome in Pymol¹.

Aufgabe 2 (*Bonus, 60 Punkte; Praxis*). In der folgenden Aufgabe sollen Sie ein ChIP-Seq Experiment auswerten und die Bindestellen des GATA4-Transkriptionsfaktors im Mausgenom bestimmen. Nutzen Sie dazu den Galaxy-Server². Sie müssen sich zum Bearbeiten der Aufgabe wieder als Nutzern anmelden.

Achtung: Es handelt sich um echte Experimente. Planen Sie also entsprechend Bearbeitungszeit ein. Dokumentieren Sie Ihr Vorgehen.

1. Laden Sie die FastQ-Dateien des Experiments mit der Accession-Nummer *SRX547823* und des Kontrolllexperiments mit der Accession-Nummer *SRX547825* über *Get Data - EBI SRA* in den Galaxyserver. Setzen Sie den Datentyp der Dateien unter *Edit attributes* auf *fastqsanger.gz*. Vorsicht, nicht mit *fastqcssanger.gz* verwechseln!
2. Mappen Sie die Reads in den FastQ-Dateien mittels *bowtie* für Illumina jeweils auf das Mausgenom (*mm9*). Konvertieren Sie für das Peakcalling die entstandenen SAM-Dateien in BAM-Dateien (*SAM tools - SAM-to-BAM*).
3. Bestimmen Sie die *tag/read-size* im ChIP-Experiment. Die tag-size bekommen Sie zum Beispiel mittels des FastQC-Qualitätsreports.

¹Benutzen Sie die Pymol Version für Ausbildungszwecke <http://http://pymol.org/edu/>

²<https://usegalaxy.org/>

4. Finden Sie nun die Bindestellen von GATA4 mit MACs (*NGS Peak calling - MACS*). Setzen Sie dazu die Option *MFOLD* auf 10 und passen Sie die *tag-length* auf die von Ihnen bestimmte Länge an. Finden Sie die Peaks mit folgenden unterschiedlichen Optionen:

- Geben Sie einmal das Kontrollexperiment an.
- Geben Sie kein Kontrollexperiment an.
- Vertauschen Sie das Kontrollexperiment mit dem ChIP-Experiment.

Wie viele Peaks erhalten Sie jeweils?

5. Generieren Sie ein Fasta-file mit den Sequenzen in den Bindestellen (*Fetch Sequences*), die Sie im ChIP-Experiment unter Angabe der Kontrolle gefunden haben.

6. Nutzen Sie RSAT³ um Transkriptionsfaktor-Motive in den Bindestellen zu finden (*NGS ChIP-Seq - peak-motifs*). Welche bekannten Motive finden Sie?

³<http://pedagogix-tagc.univ-mrs.fr/rsat/>