Übung 6: Die Proteindatenbank (PDB) und Vergleiche von Proteinstrukturen

GST, Wintersemester 2019/2020

1 Einleitung

Der erste Teil dieser Übung ist eine Einführung zur Proteindatenbank (PDB), der wohl wichtigsten Quelle von Information in der Strukturbioinformatik. Der zweite Teil beschäftigt sich mit dem Vergleich von Proteinstrukturen. Anders als in den letzten Übungen befinden sich viele Fragen bereits im Text in gekennzeichneten Boxen.

2 Die PDB

Die PDB beherbergt eine große Sammlung dreidimensionaler Strukturen von Proteinen und Nukleinsäuren. Die meisten dieser Strukturen wurden mit NMR, Roentgen-Kristallographie oder Cryo-Elektronenmikroskopie (cryo-EM) bestimmt. Neben den Dateien mit Atomkoordinaten stellt die PDB eine Sammlung von Werkzeugen zur Suche und Analyse der Makromoleküle zur Verfügung. Über die Webseite kann man zum Beispiel Informationen zu Sekundärstrukturen, Proteindomänen, Liganden und modifizierten Seitenketten sowie Klassifizierungen von Strukturen und Sequenzen bekommen.

2.1 Ein Protein finden

Besuchen Sie die Seite `www.rcsb.org`. Jetzt befinden Sie sich auf der Startseite der PDB. Oben rechts befindet sich eine Suchleiste. Suchen Sie nach *topoisomerase*. Auf der Ergebnisseite befindet sich rechts eine Liste mit den Ergebnissen der Suche. Zu jeder Struktur gibt es eine eindeutige ID unter welcher sie auch zu finden ist. Dazu sieht man zu jedem Ergebnis eine Abbildung, den Namen des Moleküls, die Autoren und etwas Information zum Experiment.

Links davon befindet sich unter *Refinements* eine Liste mit Begriffen, mit welchen sich die gesuchte Struktur weiter spezifizieren lässt. Zu jeder Eigenschaft ist die Anzahl der verbleibenden Strukturen in Klammern angegeben. Schauen Sie sich die Liste mit Eigenschaften an und wählen Sie ein paar aus bis es nur noch eine Struktur gibt.

Aufgabe 2.1
Suchen Sie nach Insulin. Wie viele Strukturen wurden für menschliches Insulin zwischen 2000 und 2005 in der PDB hinterlegt, die mittels Roentgen-Kristallographie mit einer Auflösung zwischen 1.5 und 2 Ångström bestimmt wurden?

2.2 Die Übersicht

Im Folgenden sollen beispielhaft einige der Funkionen der PDB erkundet werden. Suchen Sie die Struktur mit der ID `1QCF` und gehen Sie auf die entsprechende Seite. Sie sollten nun auf der Übersichtsseite sein. Oben befinden sich Tabs mit Begriffen wie Structure Summary, 3D View, Annotations, Sequence. Bleiben Sie zunächst auf der Übersichtsseite und schauen Sie sich die Seite gut an. Weiter unten auf der Seite befinden sich Boxen mit Informationen zu den in der Struktur enthaltenen Makromolekülen, kleinen Molekülen wie Liganden oder der Validierung der Experimentellen Daten.

Aufgabe 2.2

1. Um was für ein Protein handelt es sich bei `1QCF`?
2. Aus welchem Organismus kommt das Protein?
3. Wann wurde das Molekül veröffentlicht?
4. Wie viele Ketten hat das Molekül?
5. Wie viele und welche Liganden sind vorhanden? Gibt es modifizierte Reste?
6. Mit welcher Methode wurde diese Struktur bestimmt?
7. Was zeigt die Grafik unter *Structure Validation*?

3 Klassifizierungen und Domänen in der PDB

Proteine bestehen oft aus mehreren Domänen. Häufig haben diese Domänen unterschiedliche Funktionen. Über die PDB-Webseite sind Domänenannotationen aus mehreren Quellen zugänglich. Mit dem Programm *DomainParser* werden zum Beispiel komplett automatisch Domänen in 3D-Strukturen bestimmt.

SCOP und CATH sind hierarchische Klassifizierungen von Proteindomänen, die zu großen Teilen per Hand von Experten bestimmt wurden. In SCOP ist das Schema

$$class \rightarrow fold \rightarrow superfamily \rightarrow family$$

und in CATH

$$class \rightarrow architecture \rightarrow topology \rightarrow homology.$$

Auf die Datenbanken kann man zwar über entsprechende Webseiten (http://scop.mrc-lmb.cam.ac.uk/scop/ und http://www.cathdb.info/) zugreifen, die Information lässt sich aber auch bequem durch die PDB-Webseite anzeigen. So lassen sich die Ergebnisse außerdem gut vergleichen.

PFAM Domänen wiederum wurden nur aufgrund der Aminosäuresequenzen bestimmt. In dieser Übung sollen in erster Linie SCOP und CATH Domänen betrachtet werden. Klicken sie auf den Annotations-Reiter und schauen Sie sich die SCOP und CATH Klassifizierungen an.

Aufgabe 3.1
Aus wie vielen Domänen besteht `1QCF` laut SCOP und laut CATH?

Im Folgenden betrachten wir die Domänen und die Unterschiede zwischen der SCOP und der CATH Annotation näher. Klicken Sie dafür auf den Reiter *Sequence*. Hier befindet sich eine Box, welche die Annotationen für die Kette A zu der Sequenz abbildet. Rechts befindet sich die Sequenz. Links befindet sich eine Tabelle mit Annotationen. Unter *Add an Annotation* lassen sich Annotationen hinzufügen. In dem entsprechenden Feld der Annotation kann man diese durch klicken auf *hide* wieder verschwinden lassen. Die ausgewälten Annotationen sind in Form farbiger Balken über der Sequenz abgebildet.

Lassen Sie sich die SCOP und CATH-Domänen zu der Sequenz anzeigen und dazu die mit dem Programm *DSSP* bestimmte Sekundärstruktur. Eine Legende mit den Symbolen für Sekundärstrukturelemente befindet sich unter der Sequenz. Um die Domänen gleichzeitig mit Struktur und Sequenz zu betrachten klicken Sie auf *Display Jmol* unter *Display Options*. Bei den Annotationen kann man nun in der *Details* Spalte auf die farbig unterlegte Domäne klicken um diese in dem Strukturfenster angezeigt zu bekommen. Zeit das mal auszuprobieren.

Aufgabe 3.2

1. Wie unterscheiden sich die Domänen aus `1QCF` von SCOP und CATH? Gibt es Gemeinsamkeiten?
2. Welcher Anteil der gesamten Struktur ist Helikal und welcher Beta-Faltblatt?
3. Welches Sekundärstrukturelement deckt den größten Bereich in der kleinsten SCOP-Domäne ab? Welches deckt den größten Bereich in der größten SCOP-Domäne ab?

Noch etwas detaillierter kann man sich die Struktur im Bereich *3D View* ansehen. Gehen Sie über den entsprechenden Tab dorthin. Probieren Sie die verschiedenen Anzeigeoptionen aus. Wechseln Sie anschließend den Viewer und wählen Sie JSmol aus. Nun erscheint unter dem Viewer eine Box in der *Ligands*, *Domain* und *Modification* ausgewählt werden können. Probieren Sie wieder alles aus um die folgenden Fragen zu beantworten.

Aufgabe 3.3

1. An welche SCOP-Domäne bindet der Ligand?
2. In welcher SCOP-Domäne befindet sich die modifizierte Aminosäure?

4 Strukturvergleiche

In diesem Abschnitt geht es um den Vergleich von Proteinstrukturen.

4.1 Strukturbasierte Alignments

Hier soll das Programm `salamiLite` verwendet werden. Dieses Programm nimmt zwei Proteinstrukturen und berechnet ein Strukturalignment.

Analog zum Sequenzalignment werden beim Strukturalignment Sequenzpositionen einander zugeordnet. Diesmal aber nicht aufgrund von Sequenzähnlichkeiten, sondern aufgrund struktureller Eigenschaften. `Salami` benutzt dafür die Torsionswinkel des Proteinrückrats. Aufgrund dieser Winkel wird eine Matrix mit lokalen Ähnlichkeiten zwischen den beiden Strukturen berechnet. Die Optimale Zuordnung der Reste wird wie bei Sequenzalignments mit Hilfe des `Needleman-Wunsch` bzw. `Smith-Waterman` Algorithmus gefunden.

Anschließend werden die Proteine noch überlagert. Das bedeutet, dass ein Protein im Raum so verschoben und gedreht wird, dass der RMSD-Wert (engl. root mean square difference) minimiert wird. Da weniger konservierte und flexiblere Bereiche der Proteine sich oft stark unterscheiden, kann es sinnvoll sein, diese bei der Überlagerung zu ignorieren. `salamiLite` bekommt deswegen einen Grenzwert mitgegeben. Aminosäuren, die nach der Überlagerung weit voneinander entfernt sind, werden entfernt und die Überlagerung wird wiederholt. Das wird solange wiederholt bis alle paarweisen Distanzen zwischen den einander zugeordneten Resten unter dem Grenzwert liegen.

Legen Sie für diese Übung nun ein lokales Verzeichnis an.

Kopieren Sie `/home/petersen/teaching/structure_comparison/salamiLite` in dieses Verzeichnis. Suchen sie nun in der PDB nach den Strukturen mit den IDs 1BDG und 1J6Z. Laden sie die Dateien mit den Strukturen herunter, indem Sie auf der Übersichtsseite unter *Download Files* die Option *PDB Format* auswählen. Speichern Sie beide Dateien in dem eben angelegten Verzeichnis. Starten Sie nun salamiLite mittels

`./salamiLite 1bdg.pdb 1j6z.pdb 1bdg_out.pdb -a 0 -r 3.0`

Die ersten beiden Parameter sind hier die Eingabedateien. `1bdg_out.pdb` ist der Name der Datei, in welche die Struktur `1BDG` geschrieben wird, nachdem sie mittels Verschieben und

Rotation über die Struktur `1J6Z` gelegt wurde. Mit `-r` wird der Grenzwert für die Überlagerung übergeben. `-a 0` bedeutet, dass der Needleman-Wunsch Algorithmus für globale Alignments verwendet wird (mit 1 wird ein lokales Smith-Waterman Alignment berechnet).

Das Programm gibt das Alignment direkt in der Kommandozeile aus. Positionen, die für die Überlagerung verwendet wurden, lassen sich daran erkennen, dass sie im Alignment groß geschrieben werden. Außerdem wird die Sequenzübereinstimmung hinter `Seq ID` angegeben. Betrachten Sie das Ergebnis der Überlagerung in Chimera indem Sie

`/usr/local/zbhtools/chimera/1.11/bin/chimera 1bdg_out.pdb 1J6Z.pdb &`

in der Konsole eingeben.

Aufgabe 4.1

1. Beschreiben Sie kurz inwiefern sich die beiden Strukturen ähnlich sind und wo Sie sich unterscheiden.
2. Wie hoch ist die Sequenzübereinstimmung?
3. Um welche Moleküle handelt es sich?
4. Betrachten Sie die SCOP Annotationen der beiden Moleküle. Bis zu welcher Ebene sind sie gleich?
5. Probieren Sie verschiedene Grenzwerte für die Überlagerung aus. Machen Sie ein Bild von ihrer schönsten Überlagerung in chimera ($File \rightarrow SaveImage$).

4.2 Datenbanksuchen mit Strukturen und Sequenzen

Ein klassischer Anwendungsfall von Alignments ist die Suche nach ähnlichen Sequenzen bzw. Strukturen in einer Datenbank. Hier soll die PDB nach ähnlichen Strukturen und Sequenzen durchsucht und die Ergebnisse miteinander verglichen werden.

Für die Strukturalignments werden wir den Webservice *DALI* in Anspruch nehmen. *DALI* ist wie *salamiLite* eine Methode zur Berechnung von Strukturalignments. Über den Webservice kann man eine Struktur mit allen Strukturen der PDB vergleichen und so ähnliche Strukturen finden.

Da Strukturen im Allgemeinen schwieriger zu vergleichen sind als Sequenzen, brauchen die meisten Programme deutlich mehr Ressourcen zum Rechnen als Methoden zum Vergleich von Sequenzen. Dies fällt insbesondere ins Gewicht, wenn man eine Struktur mit vielen anderen Strukturen in einer Datenbank vergleicht. Deshalb muss man auf *DALI*-Ergebnisse etwas länger warten. Um die Übungszeit gut zu nutzen und keinen Stau auf dem *DALI*-server zu verursachen, wird in dieser Übung eine Seite mit bereits berechneten Ergebnissen aufgerufen.

Geben Sie

`cat /home/petersen/teaching/structure_comparison/1bdg_dalilink.txt`

ein um den Link zu lesen und kopieren sie diesen in ihren Browser um zur Ergebnisseite zu kommen.

DALI-Ergebnisseiten werden nach einer Woche gelöscht. Sollte der link nicht mehr funktionieren können Sie die Liste mit den gefundenen Proteinketten auch in der Datei

`/home/petersen/teaching/structure_comparison/1bdg_dalisearch.txt`

finden.

Aufgabe 4.2
Können sie 1J6Z auf der *DALI*-Ergebnisseite finden?

Anschließend soll eine zweite Suche durchgeführt werden. Diesmal soll allerdings nur Sequenzinformation genutzt werden. Dafür wollen wir das weit verbreitete Programm `BLAST` verwenden. Besuchen Sie dafür die Seite `https://blast.ncbi.nlm.nih.gov/Blast.cgi` und wählen Sie *Protein BLAST*.

Nun benötigen Sie die Sequenz des Proteins 1BDG. Besuchen Sie dafür wieder die PDB. Auf der Übersichtseite von 1BDG wählen Sie diesmal unter *Display Files* die Option *FASTA Sequence*. Kopieren Sie den gesamten Eintrag in das entsprechende Feld bei der *BLAST*-Suche. Wählen Sie außerdem für die *BLAST*-Suche unter *Choose Search Set* bei *Database* die Option *Protein Data Bank proteins (pdb)*. Benutzen Sie ansonsten die Voreinstellungen und klicken Sie auf BLAST.

`BLAST` verwendet einen sehr schnellen Algorithmus zum Vergleich von Sequenzen. Sie müssen trotzdem einen Moment warten.

Aufgabe 4.3
Können sie 1J6Z finden? Warum könnte das hier schwieriger sein?

Die PDB-Website bietet bereits vorberechnete Strukturalignments an. Gehen Sie dafür wieder auf die PDB-Seiten von 1BDG und 1J6Z. Klicken sie auf den beiden Seiten auf den Tab *Structure Similarity*.

Aufgabe 4.4
Können Sie in den Ergebnissen unter *Structure Similarity* in der PDB Hinweise auf die Ähnlichkeit zwischen den beiden Strukturen finden?

5 Noch ein Strukturvergleich

Aufgabe 5.1
Vergleichen Sie die Strukturen mit den IDs 1BAJ und 4X3X. Nutzen Sie

- die Information auf den PDB-Seiten
- *salamiLite*
- die DALI-Suche (ein Link zu den Ergebnissen befindet sich in `/home/petersen/teaching/structure_comparison/1baj_dalilink.txt` und die Ergebnisliste befindet sich in `/home/petersen/teaching/structure_comparison/1baj_dalisearch.txt`)
- BLAST
- die *structural similarity* Einträge in der PDB.

und beantworten Sie folgendes:

1. Um was für Moleküle handelt es sich? Aus welchem Organismen kommen Sie?
2. Beschreiben Sie sehr kurz die Strukturen.
3. Wie ähnlich sind sich die Strukturen und wie ähnlich sind die Sequenzen?
4. Haben Sie eine Hypothese um ihre Beobachtungen zu erklären?

6 Weitere Aufgaben

Aufgabe 6.1

1. Erklären sie jeweils in einem Satz, worum es sich bei folgenden Begriffen handelt:
   - SCOP, salamiLite, DALI, BLAST
2. Warum sind Strukturalignments nützlich? Reichen Sequenzaligments allein nicht aus?
3. Wie kann man hohe strukturelle Ähnlichkeit bei sehr geringer bis nicht nachweisbarer Sequenzähnlichkeit erklären?

7 Abgabe

Beantworten Sie die Fragen in den Boxen. Wir besprechen die Lösungen dazu in der Übung am 16.12.2019.