Andrew Torda
Björn Hansen

Zentrum für Bioinformatik

Übung zur Vorlesung
Grundlagen der Strukturanalyse

Wintersemester 2016/2017

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

12.12.2016

Übung 6: Structure Comparison[1]

1. Einführung

In der vorliegenden Übung versuchen wir Proteinstrukturen zu vergleichen und zu analysieren. Für die Berechnung von Sequenzalignment und struktureller Überlagerung von Proteinen werden wir die Programme *DaliLite* und *SalamiLite* verwenden. Zur Auswertung der Ergebnisse benötigen wir erneut *Chimera*.
Um weitere Informationen über die Proteine in Erfahrung zu bringen, werden wir die Proteinstruktur-Klassifikation *CATH* verwenden. Neben *CATH* gibt es auch andere Klassifikationen wie SCOP, FSSP und Pfam.

Wir werden mit drei Paaren von Proteinen arbeiten.
Bitte legen Sie für diese Übung ein neues Verzeichnis an:

mkdir structure_comparison

[1] Die Ursprungsfassung dieser Übung wurde von Irina Bondarenko erstellt.

2. *DaliLite*

Dali ist ein Netzwerkdienst für Proteinstruktur-Vergleiche. Unter der Adresse

http://ekhidna.biocenter.helsinki.fi/dali_server/start

können Sie Anfragen an den Server schicken. *Dali* vergleicht die eingegebene Proteinstruktur paarweise mit allen Proteinen in der PDB und listet anschließend die Proteine mit der größten strukturellen Übereinstimmung auf. Das Programm *DaliLite*, welches auf unseren Rechnern lokal installiert ist, werden wir für paarweise Strukturalignments von jeweils zwei Proteinen verwenden.

Die ersten beiden Proteine, die wir überlagern wollen, sind *1bop* und *2cpf*.

Bitte wechseln Sie in das soeben erstellte Verzeichnis

cd structure_comparison

und kopieren Sie alle für diese Übung erforderlichen Dateien in dieses Verzeichnis:

cp –r ~hansen/teaching/structure_comparison/* .

Wechseln Sie jetzt Unterordner 1/ und starten Sie anschließend *DaliLite*:

cd 1
perl ../dali_start.pl --name1=2BOP.pdb --name2=2CPF.pdb

Nach erfolgreicher Ausführung des Programms wird die Datei *alignment.msf* im aktuellen Verzeichnis erstellt. Anhand dieses Alignments können Sie die beiden Proteine in *Chimera* strukturell überlagern. Starten Sie hierfür bitte *Chimera* und öffnen Sie in *Chimera* die PDB-Dateien der beiden zu überlagernden Proteine (*2BOP.pdb* und *2CPF.pdb*). Die Datei *2CPF.pdb* enthält ein Ensemble von 20 strukturell leicht unterschiedlichen Varianten dieses Proteins. Öffnen Sie jetzt die Datei *alignment.msf* im **Multialign Viewer**:

Tools -> Sequence -> Multialign Viewer

Chimera wird das Alignment daraufhin einlesen und die Sequenzen automatisch den geöffneten Proteinen zuordnen. In der linken unteren Ecke des „Multialign Viewer“-Fensters können Sie nach dem Öffnen für einen kurzen Augenblick sehen, welche Sequenzen welcher Proteinstruktur zugeordnet wurden. Innerhalb des Alignment-Fensters wählen Sie jetzt **Structure** -> **Match.** Basierend auf dem Alignment wird *Chimera* die beiden Proteine überlagern, sodass Sie die strukturelle Ähnlichkeit erkennen können.

DaliLite erstellt außerdem die Datei *index.html,* welche Sie im Browser öffnen können. Diese Datei enthält zusätzliche Informationen über das strukturbasierte Alignment der überlagerten Proteine: Sequenzähnlichkeit, RMSD-Wert, Z-Score, …

3. *SalamiLite*

Salami ist unter

http://flensburg.zbh.uni-hamburg.de/~wurst/salami/

zu finden. Der *Salami*-Server führt für ein Anfrageprotein eine strukturbasierte Proteinsuche in der PDB durch und gibt eine Liste mit Proteinen ähnlicher Struktur zurück. In Rahmen dieser Übung benutzen wir lediglich die Lite-Version des *Salami*-Servers (*SalamiLite*), welcher zwei einzelne Proteine überlagert.
Führen Sie über die Eingabeaufforderung den folgenden Befehl aus:

../salamiLite 2BOP.pdb 2CPF.pdb 2bop_out.pdb -a 0 -r 3.0

Das Programm erwartet zwei Proteine für die Überlagerung. Dies sind die ersten beiden Parameter. Der dritte Parameter ist der Name einer zu erstellende pdb-Datei. Diese enthält die Koordinaten des ersten Proteins in einer Raumorientierung, in der das erste mit dem zweiten Protein überlagert ist.

Über die Option -a wird festgelegt, welcher Algorithmus für das Sequenzalignment verwendet werden soll. -a 0 (Null) bedeutet, dass der Needleman-Wunsch-Algorithmus verwendet wird (globales Alignment). Mit -a 1 wird dagegen festgelegt, dass der Smith-Waterman-Algorithmus zum Einsatz kommt (lokales Alignment). Mit der Option -r 3.0 wird eine obere RMSD-Wert-Grenze für die Überlagerung festgelegt, in diesem Fall 3Å. *SalamiLite* wird anschließend versuchen, die Proteine räumlich so gut zu überlagern, dass diese Grenze erreicht wird. Falls dies zunächst nicht gelingt, wird das Programm versuchen, die Paare von Aminosäuren im Alignment, welche die Überlagerung stören, möglichst nicht zu berücksichtigen, sondern stattdessen nur die wichtigsten Paare von Aminosäuren.

Schauen Sie sich die Ausgabe des Alignments genau an: Die großen Buchstaben im Alignment geben an, dass diese Stellen bei der Überlagerung berücksichtigt wurden. Die mit kleinen Buchstaben gekennzeichneten Positionen wurden dagegen nicht berücksichtigt. Neben der Anzahl der berücksichtigten Aminosäure-Paare des Sequenzalignments finden Sie vor dem Alignment auch Informationen zum RMSD-Wert der Überlagerung sowie zur Sequenz-Identität.
Öffnen Sie jetzt die Kooardinaten-Dateien *2CPF.pdb* und *2bop_out.pdb* in *Chimera*, um sich die Überlagerung anzuschauen.

Wählen Sie nun eine besonders große Grenze für den RMSD-Wert aus (z.B. 10.0) und untersuchen Sie die neuen Dateien in *Chimera*. Ist es nun schwerer, die räumliche Überlagerung der Strukturen zu erkennen?

4. *CATH*

CATH (eine Abkürzung von engl. **C**lass, **A**rchitecture, **T**opology und **H**omologous superfamily) ist eine halbautomatische Klassifizierung von Proteinen:

http://www.cathdb.info/

Bei dieser Klassifizierung werden Proteinsequenzen zunächst automatisch auf Grundlage ihrer Aminosäuresequenz klassifiziert und sogenannten homologen Überfamilien (Homologous superfamily) zugeordnet. Danach werden diese Überfamilien strukturbasiert in verschiedene topologische Gruppen einsortiert. Die Zuordnung zu einem Architektur-Level erfolgt manuell. Hierbei werden alle Topologien, welche die gleiche 3D-Strukturen bilden, zum selben Architektur-Level zusammengefasst. Diejenigen Architekturen, welche zu ähnlichen Anteilen aus den verschiedenen Sekundärstrukturelementen bestehen, werden wiederum derselben Klasse zugeordnet (alpha domains only, beta domains only, alpha and beta, …).

Wir wollen nun versuchen, unsere Proteine in der *CATH*-Klassifikation zu finden. Besuchen Sie dafür mit dem Browser Ihrer Wahl die Website

http://www.cathdb.info/search

und geben Sie dort die PDB-Namen Ihrer Proteine ins Suchfeld ein. Auf der Ergebnisseite klicken Sie bitte auf den Eintrag unter "Matching *CATH* Domains" und notieren sich die wichtigsten Informationen über diese Familie.

Sind die Proteine in der gleichen Klasse? Falls ja, warum ist das so? Welche Proteine gibt es in dieser Klasse?

In Ihr Arbeitsverzeichnis *structure_comparison* haben Sie auch die Unterverzeichnisse *2/* und *3/* kopiert, welche jeweils ein weiteres Paar von Proteinen enthalten. Wiederholen Sie Ihre Analyse (*DaliLight*, *SalamiLight*, *CATH*) für diese zwei Proteinpaare.

Sind *1tul* und *2brq* in der gleichen Klasse? Falls ja, warum? Hätten Sie dieses Proteinpaar gefunden, wenn Sie nur mit Hilfe der *CATH*-Klassifikation gearbeitet hätten? Wie würden Sie die strukturelle Ähnlichkeit bei einer derartig niedrigen Sequenzähnlichkeit erklären?

5. Aufgaben

Bitte beantworten Sie die folgenden Fragen und bringen Sie Ihre Antworten zur Übung am **9. Januar 2017** mit. Der Zufall bestimmt, welche Studenten Ihre Lösungen präsentieren werden.

1. Welche Programme haben Sie für die strukturbasierte Suche von Proteinen verwendet?

2. Wofür benötigen wir strukturbasierte Alignments?
   Sind sequenzbasierte allein nicht ausreichend?

3. Wie kann man hohe strukturelle Ähnlichkeit bei sehr geringer bis nicht nachweisbarer Sequenzähnlichkeit erklären?

4. Welche Anforderungen sollten wir an ein Programm stellen, dass die strukturelle Überlagerung zweier Proteine ermittelt?

5. Was ist der RMSD-Wert? Ist der RMSD-Wert immer ausreichend, um die Qualität eines strukturbasierten Alignments zu beurteilen?

6. Welche Funktion haben f-dme- und Z-Score?

7. Informieren Sie sich über CATH, SCOP, FSSP und Pfam.
   Diskutieren Sie, welche Probleme bei solchen Klassifikationen auftauchen können.
   Ist eine dieser Klassifikationen ausreichend? Was spricht für ihre Verwendung?
   Fällt Ihnen auch ein Argument gegen die Verwendung einer Klassifikation von Proteindomänen ein?

8. Würden Sie eine andere Klassifikation vorschlagen und falls ja, wie würde diese aussehen?