



Übung 7: Structure Comparison

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 hierfür ein neues Verzeichnis an:

```
mkdir pdbs
```

2. *DaliLite*

Dali ist ein Netzwerkserver 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 untersuchen wollen, sind *1bop* und *2cpf*.

Bitte legen Sie hierfür einen neuen Ordner an und wechseln Sie anschließend in diesen Ordner:

```
mkdir struct_analyse
cd struct_analyse
```

Kopieren Sie jetzt die beiden Proteine in den erstellten Ordner:

```
scp -r ~bondarenko/Public/teaching/gst/struct_compar/1 ./
cd 1
```

Starten Sie nun *DaliLite*.

```
perl ~bondarenko/Public/teaching/gst/struct_compar/dali_start.pl
--name1=2BOP.pdb --name2=2CPF.pdb
```

Nach erfolgreicher Ausführung des Programms, wird eine *alignment.msf* Datei 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 überlagerten Proteine (*2BOP.pdb* und *2CPF.pdb*) sowie die Datei *alignment.msf*. *Chimera* wird das Alignment jetzt einlesen und die Sequenzen automatisch den geöffneten Proteinen zuordnen. In der linken unteren Ecke können Sie 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/>

oder

http://flensburg.zbh.uni-hamburg.de/~wurst/salami/version_2

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:

```
~bondarenko/Public/wurstl/tools/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 Nedelman-Wunsch-Algorithmus verwendet wird (globales Alignment). Mit `-a 1` wird dagegen festgelegt, dass der Smith-Watherman-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 Kooordinaten-Dateien *2CPF.pdb* und *2bop_out.pdb* in *Chimera*, um sich die Überlagerung anzuschauen.

Wählen Sie jetzt eine besonders große Grenze für den RMSD-Wert (z.B. 10.0) und untersuchen Sie die neuen Dateien in *Chimera*. Sie werden sehen, dass die Ähnlichkeit der Strukturen nicht mehr so klar anhand der räumlichen Überlagerung zu erkennen ist. Bei kleinen Proteinen wie denen in diesem Beispiel kann die Ähnlichkeit noch erkannt werden, bei größeren Proteinen ist das dagegen nicht mehr so einfach.

4. CATH

CATH (eine Abkürzung von engl. Class, Architecture, Topology und Homologous 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 an. Auf der Ergebnisseite klicken Sie bitte auf "Family" und notieren sich die wichtigsten Information über diese Familie.

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

Wiederholen Sie Ihre Analyse für weitere zwei Paare von Proteinen:

```
cd ..
```

```
scp ~/bondarenko/Public/teaching/gst/struct_compar/2 ./
```

```
scp ~/bondarenko/Public/teaching/gst/struct_compar/3 ./
```

Sie haben jetzt zwei weitere Order angelegt, die jeweils ein weiteres Proteinpaar enthalten.

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?

Betrachten Sie das Paar *1jjo* und *1kct*: Ist es immer sinnvoll, Proteine domänenbasiert zu klassifizieren?

5. Aufgaben.

1. Fassen Sie kurz zusammen, welche Programme Sie für die strukturbasierte Suche von Proteinen benutzt haben. Welche grundlegende Idee steckt hinter diesen Programmen.
2. Was ist der RMSD-Wert? Ist dieser Wert ausreichend, um die Qualität eines strukturbasierten Alignments zu beurteilen?
3. Welche Funktion haben Z- und f-dme-Score?
4. Diskutieren Sie über CATH: Ist diese Klassifikation ausreichend?
5. Informieren Sie sich über SCOP, FFSP und Pfam. Diskutieren Sie, welche Probleme bei solchen Klassifikationen auftauchen können.
6. Würden Sie eine andere Klassifikation vorschlagen und falls ja, wie würde diese aussehen?