# ASE Übung 7 - BLAST

#### Einführung

Das Ziel dieser Übung ist es sich erneut mit dem Kontext der Sequenz- und Strukturkonservierung auseinander zu setzen und sich weiter mit den gängigen Tools zur Datenbanksuche, Alignments, Sequenz- und Strukturanalyse vertraut zu machen.

Zunächst wird in einer Datenbank nach Proteinsequenzen menschlicher Globine gesucht. Die Suchergebnisse werden evaluiert und mittels eines multiplen Sequenzalignments und eines phylogenetischen Baums dargestellt. Anschließend werden einzelne Strukturen analysiert und mit einander verglichen.

#### BLAST

Öffnen Sie das *NCBI* Webinterface für Proteinsequenzen (https://www.ncbi.nlm. nih.gov/protein/) und suchen Sie nach dem Eintrag der menschlichen Hämoglobin alpha Untereinheit (NP\_000508.1). Verschaffen Sie sich einen Überblick über das Molekül.

Mittels des *NCBI* Webinterfaces zum *Basic Local Alignment Search Tool (BLAST)* lassen sich Sequenzdatenbanken nach zu einer Eingabesequenz ähnlichen Sequenzen durchsuchen. Öffnen Sie ein neues *BLAST* Fenster für die gegebene Sequenz, indem sie in der rechten Spalte unter Analyze this sequenze auf Run BLAST klicken. Als Eingabesequenz ist nun die richtige *accession number* bereits eingetragen.

Das Ziel ist es nun alle bekannten ähnlichen Proteinsequenzen des Menschen zu finden. Wählen Sie hierzu unter Choose Search Set die Datenbank Reference proteins

(refseq\_protein) aus und tragen Sie als Zielorganismus Homo sapiens ein. Aktivieren Sie weiter unten die Checkbox Show results in a new window und starten Sie eine suche, indem Sie auf den blauen BLAST Button klicken. Abhängig von den Parametern, Zielsequenzen und der Serverauslastung kann eine Suche einige Sekunden bis Minuten dauern.

Betrachten Sie die Ergebnisse. Unter Graphic Summary sehen Sie die annotierte Eingabesequenz, sowie die nach Alignmentscore gefärbte Sequenzüberdeckung der gefundenen Sequenzen. Unter Descriptions sehen Sie die Liste der gefundenen Sequenzeinträge sortiert nach Alignmentscore, sowie die Sequenzüberdeckung, E-Wert (Signifikanz der Ähnlichkeit) und die Sequenzidentität. Schließlich werden unter Alignments die paarweisen Sequenzalignments zwischen Eingabesequenz und Ergebnissen dargestellt.

In der ersten Suche mit Standardparametern sollten nur einige andere Hämoglobin Untereinheiten gefunden werden. Die Suche soll nun erweitert werden. Wechseln Sie zurück zum *BLAST* Parameterfenster und klicken Sie unten auf Algorithm parameters. Setzen Sie die Ihnen bekannten Parameter der Substitutionsmatrix und der Gap Kosten auf Werte, die es wahrscheinlicher machen, weniger ähnliche Sequenzen zu finden, und starten Sie eine neue Suche.

## **PSI-BLAST**

Sollten die Ergebnisse nicht signifikant erweitert werden, so ist es notwendig den Algorithmus zu ändern. Wählen Sie unter Program Selection den aus der Vorlesung bekannten PSI-BLAST (Position-Specific Iterated BLAST) Algorithmus aus. Hier wird in jeder Suchiteration zunächst ein durchschnittliches Profil aus den bisher gefundenen Sequenzen erstellt und als neue Sucheingabe verwendet, was i.d.R. das Finden weniger verwandter Sequenzen mit weiterhin signifikanter Homologie erlaubt. Starten Sie zunächst eine Suche und anschließend jeweils eine neue Iteration, indem Sie im Ergebnisfenster unter Descriptions neben Run PSI-Blast iteration ... with max ... auf Go klicken. Neu gefundene Sequenzen werden gelb markiert. Führen Sie Suchiterationen aus, bis keine neuen Sequenzen mehr gefunden werden.

Betrachten Sie die Ergebnisse. Sind auch nicht-Globin Sequenzen vorhanden? Wenn ja, wie schätzen Sie deren Ähnlichkeit zu der Eingabe ein, wenn Sie Alignmentscore, Sequenzidentität und E-Wert berücksichtigen? Schauen Sie sich an, wie der Alignmentscore und die Sequenzidentitäten sich entwickeln. Was fällt hier auf und wie ist dies zu erklären?

## Phylogenetischer Baum

Wählen Sie unter den *PSI-BLAST* Ergebnissen alle Sequenzen der Globin-Familie aus, indem Sie die Checkboxen in der linken Spalte anklicken. Erzeugen Sie einen phylogenetischen Baum der ausgewählten Sequenzen, indem Sie über der Headerzeile auf Distance tree of results klicken, und betrachten Sie das Ergebnis. Scheint Ihnen die Einteilung der Sequenzen konsistent zu sein?

Überprüfen Sie, wie robust der Baum ist, indem Sie andere Verfahren bzw. Metriken unter Tree method und Distance wählen, und schauen, ob der Baum sich signifikant ändert. Experimentierten Sie auch mit anderen Darstelltungsoptionen unter Tools -> Layout.

### Strukturanalyse

Im Folgenden sollen die Strukturen einiger gefundener Proteine mit einander verglichen werden. Starten Sie das Programm *Chimera* mit folgendem Kommandozeilenbefehl:

/usr/local/zbhtools/chimera/1.13/bin/chimera &

Öffnen Sie die Struktur eines menschlichen Hämoglobins (File -> Fetch by ID...) mit der PDB ID 6HAL. Öffnen Sie das selbe Modell auch über das PDB Webinterface (www.rcsb.org) und finden Sie unter Macromolecules heraus, welche Ketten die alpha Unterheinheiten bilden. Löschen Sie alle Ketten bis auf eine alpha Untereinheit in Chimera, indem Sie zunächst eine Kette auswählen (Select -> Chain) und dann die Auswahl löschen (Actions -> Atoms/Bonds -> delete).

Wiederholen Sie die Schritte für das die zeta Untereinheit beinhaltende Modell 3W4U und das Neuroglobin 4MPM, so dass am Ende nur je eine Hb-alpha, -zeta und eine Neuroglobin Kette übrig bleiben.

Starten Sie das MatchMaker Tool (unter Tools -> Structure Comparison). Hiermit ist es möglich Strukturüberlagerungen durchzuführen. Wählen Sie unter Reference structure das Modell der alpha Untereinheit und unter Structure(s)to match die beiden anderen Modelle aus. Wenn Sie auf Apply klicken, werden letztere auf das Referenzmodell überlagert, indem zunächst ein Sequenzalignment durchgeführt wird und anschließend versucht wird das Sequenzalignment geometrisch möglichst gut anzunähern.

Die Strukturen sollten nun überlagert sein. Nutzen Sie das ModelPanel (unter Favorites ), um die Sichtbarkeit der einzelnen Strukturen ein- und auszuschalten (Shown Checkbox), und evaluieren Sie visuell die Ähnlichkeit der zeta und der Neuroglobin Untereinheiten jeweils zur alpha Untereinheit. Im Reply Log (unter Favories) werden die nach dem Überlagern berechneten RMSD (*root mean square deviation*) Werte in Ångström (0.1 nm) angegeben. Hierbei handelt es sich um die durchschnittliche Abweichung der Atomkoordinaten der Überlagerten Strukturen.

Prüfen Sie, ob die beobachteten strukturellen Abweichungen grob den Sequenzunterschieden folgen. Schauen Sie hierzu nochmal auf den phylogenetischen Baum.

Aus den *BLAST* Ergebnissen lässt sich auch ein multiples Sequenzalignment erstellen. Klicken Sie hierzu in der Descriptions Headerzeile auf Multiple alignment. Zoomen Sie weiter hinein um die Sequenzelemente zu sehen. Experimentierten Sie mit Färbungen nach unterschiedlichen Eigenschaften unter Tools -> Coloring. Setzen Sie den Parameter z.B. auf BLOSUM45 um die Residues nach mittels der Substitutionsmatrix ermittelten Ähnlichkeit innerhalb der jeweiligen Spalte einzufärben.

In der Mitte des Alignments sollten Sie zwei stark konservierte Histidin Spalten beobachten. Lassen sich diese Residues und ihre Funktion in den Strukturen sofort erkennen?