ASE Übung 2 - Datenbanken / Alignments

Gen-Datenbank

Nutzen Sie das *NCBI*-Webinterface zu Gendatenbanken (`https://www.ncbi.nlm.nih.gov/gene/`) um nach Sequenzdaten des `SOD1` Gens zu suchen. Finden und öffnen Sie die jeweiligen Einträge der *Homo sapiens* und *Mus musculus* Taxa. Falls notwendig, lassen sich die Suchergebnisse in der rechten Spalte nach Taxon und anderen Kriterien filtern.

Verschaffen Sie sich einen Überblick über die in den Einträgen präsentierten Daten und lesen Sie die „Summary“ und „Genomic context“ Abschnitte, um mehr über die Art, Funktion und Locus des Gens zu erfahren.

Abschnitt `Genomic regions, transcripts, and products` gibt neben einer annotierten Visualisierung der Sequenz die Möglichkeit den Eintrag im *GenBank* oder *FASTA* Format zu betrachten. Hierbei handelt es sich um zwei der gängigsten textbasierten Sequenz-Dateiformate, welche sowohl vom Menschen als auch automatisiert verarbeitet werden können. Betrachten Sie beide Dateiformate. Welche Informationen werden mit welcher Syntax jeweils gespeichert? Was sind die Unterschiede (und daraus möglicherweise folgende Einsatzzwecke) der Formate?

Betrachten Sie nun die erwähnte Visualisierung der Sequenz. Um die zahlreichen Informationstypen auf das Wesentliche zu beschränken, klicken Sie in der Menuleiste auf:

`Tracks -> NCBI Recommended Track Sets -> Genes`

Sie sehen nun das `SOD1` Gen eingebettet in der Chromosomsequenz mit zwei darunter liegenden Tracks. Versuchen Sie mit Hilfe der Legende (in Menuleiste: `? -> Legends`) die Visualisierung zu interpretieren. Aus wie vielen Nukleotiden besteht das Gen der jeweiligen Taxa? Stimmen die Anzahl und grobe Position der Exons miteinander überein?

Sequenzalignment

Behalten Sie die Datenbankeinträge weiterhin offen. Öffnen Sie das *NCBI* Webinterface zu verschiedenen Alignment- und Sequenzsuche-Programmen: `https://blast.ncbi.nlm.nih.gov/Blast.cgi`

Wählen Sie hier das `Global Align` Programm, welches ein globales Sequenzalignment zweier Nukleotid- bzw. Proteinsequenzen mittels des *Needleman-Wunsch* Algorithmus durchführt. Kopieren Sie die Sequenzen des `SOD1` Gens der beiden Taxa aus den jeweiligen *GenBank* oder *FASTA* Dateiformaten in die `Query` bzw. `Subject` Sequenzfelder. Schauen Sie sich die Parametermöglichkeiten des Alignments unter `Algorithm parameters` an, lassen Sie jedoch die voreingestellten Werte. Klicken sie auf `Align`, um das Alignment durchzuführen.

Betrachten Sie das Ergebnis. Wie groß sind die Anteile von Matches bzw. Gaps und daraus folgend der angenommenen Substitutionen? Was beobachten Sie hinsichtlich der unterschiedlichen Längen der Genvarianten (die `Dot Matrix View` gibt hier einen guten Überblick)?

Betrachten Sie die Textdarstellung des Alignments. Überprüfen Sie grob, ob die Exonbereiche auf einander abgebildet sind. Nutzen Sie dazu die Indexspalte und die Visualisierungen der Sequenzen in ihren jeweiligen Datenbankeinträgen. Dort ist es sinnvoll mit dem `Zoom` Tool zu

arbeiten und einen neuen Indexheader mittels `Tools -> Set Origin` zu erzeugen, wobei der neue Ursprung die Startposition des Gens in der Chromosomsequenz ist.

Protein-Datenbank

Nutzen Sie die *RCSB Protein Data Bank* (`rcsb.org`) um die Struktureinträge des von dem Gen codierten Proteins zu finden. Suchen sie dabei nach dem Gennamen oder dem gängigen Namen des codierten Proteins (welchen Sie in den *NCBI* Einträgen unter `General protein information` finden).

Die Suchergebnisse werden nicht eindeutig sein, da die Strukturen vieler Proteine unter unterschiedlichen Umständen mehrfach ermittelt wurden. Filtern Sie auch hier die Ergebnisse nach Taxon (linke Spalte) und öffnen Sie wieder jeweils einen Eintrag für *homo sapiens* und *mus musculus*.

Verschaffen Sie sich auch hier einen kurzen Überblick über die präsentierten Daten. Öffnen Sie den Tab `Sequence` und betrachten Sie die `Sequence Chain View`. Welche Information wird hier vermittelt?

Protein Sequenzalignment

Führen Sie nun ein Sequenzalignment der jeweiligen Proteinsequenzen durch. Öffnen Sie dazu ein neues `Global Align` Fenster und wählen Sie diesmal oben links `Protein` anstatt von `Nucleotide` aus.

Anstatt die Proteinsequenzen aus den Datenbanken zu kopieren, ist es möglich die ID der Einträge (auch *accession number*) direkt zu übergeben. Hier muss jedoch zusätzlich die ID der Proteinkette mit dem „_“ Zeichen angehängt werden, da viele Struktureinträge aus mehreren Proteinketten bestehen - diese sind mit Großbuchstaben indiziert. So würde z.B. für die Sequenz der ersten Kette des *PDB* Eintrags `3GTT` die ID `3GTT_A` in das `Query` bzw. `Subject` Feld eingetragen werden.

Betrachten Sie das Alignment. Wie unterscheidet es sich vom Alignment der jeweiligen Gene hinsichtlich der Qualität?

Fragen

1. Sequenzdatenbanken speichern neben den reinen Sequenzen i.d.R. weitere Daten. Geben Sie Beispiele, um welche Information es sich dabei handeln kann und wofür diese benötigt wird.
2. Welche Parameter verwendet der *Needleman-Wunsch* Algorithmus zum Vergleich von Nukleotidsequenzen? Könnten die Parameter für den Vergleich von Proteinsequenzen sinnvoll erweitert werden?
3. Es besteht die Möglichkeit die Nukleotid- bzw. Aminosäuresequenzen der proteincodierenden Gene bzw. Proteine mittels Sequenzalignments zu vergleichen. Wann ist welcher Vergleich sinnvoll?