

Informationsmaterial

Resistenz gegen HIV – Recherche und Analyse molekularer Daten

(inkl. Anleitungen zur Recherche von Sequenzen mit GenBank® und zur Analyse mit GeneDoc)

In der Computer-basierten Version der Aufgabe „Resistenz gegen HIV“ werden eine Recherche von DNA-Sequenzen im Internet und ein Sequenzvergleich mit dem Programm GeneDoc durchgeführt. Um Ihnen den Einstieg in die Sequenz-Recherche und den Sequenz-Vergleich zu erleichtern, finden Sie im Folgenden eine kurze Einführung und Anleitung zur Nutzung der Medien.

Die Recherche nach Sequenzen mit GenBank®

Um authentische Sequenzen zu vergleichen, kann man sich die Gendatenbanken im Internet zu nutze machen. Die Sequenzdaten aller Gen-Datenbanken werden öffentlich verwaltet und stehen jedem zur Nutzung und Analyse über die www-Seiten der Organisationen zur Verfügung. Eine dieser Datenbanken, die DNA-Sequenzen sammelt, verwaltet und nutzbar macht, ist die GenBank® des NIH (National Institute of Health) in den USA, die vom NCBI, dem National Center for Biotechnology Information verwaltet wird (www.ncbi.nlm.nih.gov). Die Gendatenbanken sind das Resultat wissenschaftlicher Publikationen, denn die meisten Zeitschriften fordern, dass zeitgleich mit der schriftlichen Veröffentlichung von Forschungsarbeiten auch die Sequenzen in zentralen öffentlichen Datenbanken gespeichert werden müssen. Im traditionellen Zweig von GenBank® werden zurzeit über 106 Milliarden Basen in über 108 Millionen Sequenzen verwaltet (Stand August 2009).

Die Startseite des NCBI bietet einen einfachen Einstieg in die Suche nach Sequenzdaten. Ein Ausklappmenü im Kopf der Seite erlaubt es, die Suche gezielt auf bestimmte Daten einzuschränken, so kann man z.B. gezielt nach DNA-Sequenzen suchen, indem man bei *Search:* „Nucleotide“ anwählt. Da die Suche textbasiert abläuft, ist es möglich, als Suchbegriffe Artnamen, Gen-Bezeichnungen, Autoren, Akzessionsnummern o.ä. anzugeben, z.B. *Search „Nucleotide“ for „CCR5 mutant chemokine receptor“*.

Die Suche liefert dann eine Auflistung aller Einträge, die als Hyperlinks direkt zum vollständigen Datenbankeintrag weiterleiten. Unterhalb der Bezeichnung findet man die Akzessionsnummer, die der eindeutigen Identifizierung einer Sequenz dient und immer als Referenz angegeben werden sollte, da insbesondere Gen-Bezeichnungen keinem universellen

Benennungsmuster unterliegen. Die Auflistung ändert sich ständig, da die Datenbank stetig aktualisiert und erweitert wird. Aus dieser Liste wählt man einen für die jeweilige Fragestellung geeigneten Eintrag aus. Im aktuellen Beispiel ‚Resistenz gegen HIV‘ wird nach Sequenzen im menschlichen Genom gesucht, daher könnte man z.B. ‚*Homo sapiens mutant CCR5 chemokine receptor (CCR5) gene, complete cds*‘ (Akzessionsnummer AY744141.1) anwählen. Ein solcher Datenbankeintrag enthält dann detaillierte Angaben zu der gesuchten Sequenz (z.B. Anzahl der Basen, Vollständigkeit der Sequenz, etc.), der biologischen Art des betreffenden Organismus, der zugehörigen Publikation, weiteren Literaturangaben sowie die abgeleitete Aminosäureübersetzung der DNA-Sequenz. Einige dieser Angaben sind zusätzlich mit Verlinkungen versehen, die zu weiteren relevanten Datenbanken (z.B. mit Informationen zur Systematik etc.) führen. Am Ende des Eintrags findet man die gesuchte DNA-Sequenz. Bei dieser Darstellung handelt es sich um das GenBank-Format, weitere Darstellungs-Optionen findet man unterhalb der fettgedruckten Gen-Bezeichnung: *FASTA* oder *Graphics*. Das GenBank-Dateiformat ist jedoch für den Transfer zwischen den Analyseprogrammen wenig geeignet. Besser gelingt dieses mit dem FASTA-Format (.aa), da dieses kompatibel mit zahlreichen Analyse-Programmen ist. Dies gilt für Nukleotid- wie auch Aminosäuresequenzen gleichermaßen. Rechts oben auf der Seite wird die Option *Send* angezeigt. Durch Anwählen dieses Links öffnet sich ein Menü, in dem empfohlen wird ‚*Complete Record*‘ und unter ‚*Choose Destination*‘: ‚*File*‘ zu markieren. Daraufhin öffnet sich ein kleines Untermenü, in dem man das Datei-Format anwählen kann. Durch Auswahl des *Formats FASTA* kann man die gewünschte Sequenz für die weitere Bearbeitung in einem Analyseprogramm im FASTA-Format an einem geeigneten Ort abspeichern.

Das Alignment der Sequenzen

Das Alignment (englisch für Abgleich, Anordnung, Ausrichtung, Alignierung) dient dem Vergleich von 2 oder mehr Sequenzen (Nukleotid- oder Aminosäuresequenzen). Ermittelt werden Muster in DNA- oder Aminosäuresequenzen; Unterschiede zwischen den Sequenzen können sich im Laufe der Evolution beispielsweise durch Substitutionen, Deletionen oder Insertionen ergeben haben.

Aus den Nukleotid- oder Aminosäuresequenzen wird ein sogenanntes Alignment erstellt, eine Matrix, in der die untersuchten Taxa untereinander aufgelistet werden und die Sequenzen in horizontaler Richtung angeordnet sind. Dabei werden die Sequenzen einander so zugeordnet, dass die Reihenfolge erhalten bleibt und jedes Element (Nukleotid) einem anderen Element

oder einer Lücke (gap) in jeder Sequenz zugeordnet ist. Die besondere Schwierigkeit bei der Konstruktion eines guten Alignments sind die sog. Lücken (gaps), die eingefügt werden müssen, um dem Umstand Rechnung zu tragen, dass sich Nukleotidsequenzen nicht nur durch den Austausch von Nukleotiden (Substitution) verändern, sondern auch durch Verlust (Deletion) oder Einfügung (Insertion).

Die Alignierung von proteincodierenden Nukleotidsequenzen ist i.d.R. einfacher als die von anderen Sequenzabschnitten, da sie anhand der Aminosäureübersetzung kontrolliert werden können. Bei solchen Nukleotidsequenzen sollten Deletionen und Insertionen immer als ein Vielfaches von 3 auftreten, da ansonsten das Leseraster des codierten Proteins zerstört werden würde. Die meisten Substitutionen sollten in solchen Alignments auf die 3. Codonposition entfallen, da hier durch Veränderungen der Sinngehalt des Codons oft nicht verändert wird.

Was in kleinen Beispielaufgaben (wie z.B. ‚Resistenz gegen HIV‘) noch von Hand lösbar ist, muss in einem Alignment von vielen Taxa mit zahlreichen Mutationen zumindest teilweise von Computerprogrammen übernommen werden. Ein computerbasiertes Alignment sollte aber immer noch einmal manuell überarbeitet werden, bevor weitere Analysen erfolgen.

Eine sehr attraktive Darstellung alignierter Sequenzen gelingt mit dem Programm GeneDoc, wobei das Alignment hier nicht automatisch von dem Programm generiert wird, sondern manuell bearbeitet werden muss (Programmbeschreibung und Tutorial siehe unter ‚GeneDoc‘, S. 4). Ein automatisches Alignment und anschließende Analysen werden beispielsweise im Programm Mega4 durchgeführt, auf das an dieser Stelle aber nicht weiter eingegangen werden soll.

Ein Sequenz-Alignment dient in vielen Fällen als Grundlage für die Analyse von Verwandtschaftsbeziehungen zwischen Organismen, z.B. für die Berechnung molekularer Stammbäume. Daher muss bereits beim Alignment unbedingt sorgfältig gearbeitet werden. Die Alignierung ist der wichtigste erste Schritt zur Ableitung einer molekularen Phylogenie!

Tutorial GeneDoc

GeneDoc

Multiple Sequence Alignment Editor & Shading Utility Version 2.6.002

Copyright © 2000 by Karl Nicholas

Dieses Tutorial stellt erste grundlegende Schritte eines Sequenzalignments mit dem Programm GeneDoc dar. Es soll einen Einblick in die wissenschaftliche Arbeit bei Sequenzanalysen geben und erhebt keinen Anspruch auf Vollständigkeit.

Das Programm wird von den Herstellern zum kostenlosen Download angeboten (www.nrbsc.org/gfx/genedoc/).

Das Erstellen des Datensatzes

Import von Sequenzen:

Wählt man in der Menüleiste ‚File‘ (Datei) > ‚Import‘ (Importieren) an, öffnet sich ein kleines separates Fenster ‚Import Type‘ (Art des Imports), in dem das Dateiformat der Sequenz erfragt wird. I.d.R. passt die Voreinstellung ‚FASTA (Pearson)‘ und der Import der Sequenz kann durch Anklicken von ‚Import‘ veranlasst werden. [Das Dateiformat FASTA (Pearson) ist mit vielen Programmen kompatibel und zu Beginn wurde bereits empfohlen, Sequenzen aus GenBank® im FASTA-Format abzuspeichern (Dateiname.aa)]. Die gewünschte Datei kann nun angewählt und geöffnet werden. Nach jedem Öffnen einer Sequenzdatei erscheint das ‚Import Type‘ Fenster erneut und es können weitere Sequenzen ergänzt werden. Ist der Datensatz vollständig, wird im Fenster ‚Import Type‘ die Option ‚Done‘ (Erledigt) gewählt.

Das Programm zeigt nun eine Auflistung der Sequenzen. Dabei erscheint links die Bezeichnung der jeweiligen Sequenz und nach einem Doppelpunkt wird die vollständige Sequenz in der Zeile angezeigt.


Möchte man ein bereits bestehendes Alignment zur weiteren Bearbeitung öffnen, wählt man ebenfalls unter ‚File‘ > ‚Import‘ das entsprechende Dateiformat an. Dieses kann z.B. ebenfalls bereits als FASTA (Pearson), also Dateiname.aa, vorliegen oder in anderen Formaten, wie z.B. im Clustal-Format als Dateiname.aln.

Bearbeitung der Sequenzen / des Alignments

Umbenennen der Sequenzen

Die etwas unpraktische Bezeichnung der Gene kann vom Nutzer geändert werden. Zu diesem Zweck wählt man in der Menü-Leiste ‚Project‘ > ‚Edit Sequence List‘ (Bearbeiten der Sequenz-Liste). Es erscheint ein Fenster, in dem alle Sequenzen des Datensatzes aufgelistet sind. Durch einen Doppelklick auf die gewünschte Sequenz erscheint ein neues Fenster mit dem aktuellen Namen der Sequenz. Dieser kann durch eine andere Bezeichnung ersetzt und mit ‚Ok‘ bestätigt werden [im Alignment angezeigt werden die ersten 10 Zeichen (Voreinstellung)].

Grad der Übereinstimmung der Sequenzen

Die Nukleotide sind mit unterschiedlichen Graustufen hinterlegt, die den Grad der Übereinstimmung zwischen den verschiedenen Sequenzen deutlich machen. Dabei ist es möglich, über Knöpfe (Buttons) in der Menü-Leiste unterschiedliche Level der Schattierung anzuwählen:  0, 2, 3 oder 4 Graustufen (Level 4 im Beispiel unten). Nach jeder Verschiebung der Sequenzen sollte die Schattierung durch Anklicken des Buttons aktualisiert werden.

```

540          *          560          *          580
TCTCATTTTCCATACAGTCAGTATCAATTCTGGAAGAATTTCCAGAC
TCTCATTTTCCATACATTAAAGATAGTCATCTTGGGGCTGGTCCTGC
TCTCATTTTCCATACA T A AT          G T C C


```

Verschieben der Sequenzen

In der Menü-Leiste sind zwei Knöpfe (Buttons) zu finden, die die Verschiebung einzelner Nukleotide (Grab and Slide →1) bzw. ganzer Sequenzabschnitte (Grab and Drag →2) ermöglichen.



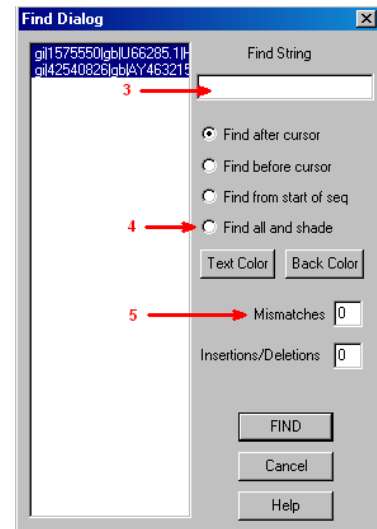
Suche nach Übereinstimmungen

Um die Sequenzen miteinander abzugleichen, ist es notwendig, nach Übereinstimmungen zu suchen. Dies kann manuell erfolgen oder aber mit Hilfe der Suchfunktion des Programms. Dazu wählt man in der Menü-Leiste ‚Edit‘ (Bearbeiten) > ‚Find‘ (Finden). Eine „Abkürzung“ besteht in dem Knopf (Button): . Es öffnet sich ein kleines Fenster ‚Find Dialog‘, in dem man in das Feld ‚Find String‘ (Finde Sequenz) (→3) eine Abfolge von Nukleotiden eingibt, die in einer der vorliegenden Sequenzen vorhanden ist. Es sollte die Option ‚Find all and Shade‘ (Finden und Schattieren) (→4) aktiviert werden. Die Schattierungs-Farbe kann bei

‚Text Color‘ verändert werden. Durch Anklicken von ‚Find‘ durchsucht das Programm alle Sequenzen auf eine Übereinstimmung mit der angegebenen Nukleotid-Sequenz und markiert diese farbig.

620 * 640 * 660
 GCTGCTTGTTCATGCTCATCTGCTACTCGGGAATCCTAAAAACT
 TCCTAAAAACTCTGCTTCGGTGTGCGAAATGAGAAGAAGAGGCA
 C C G T GA A A C

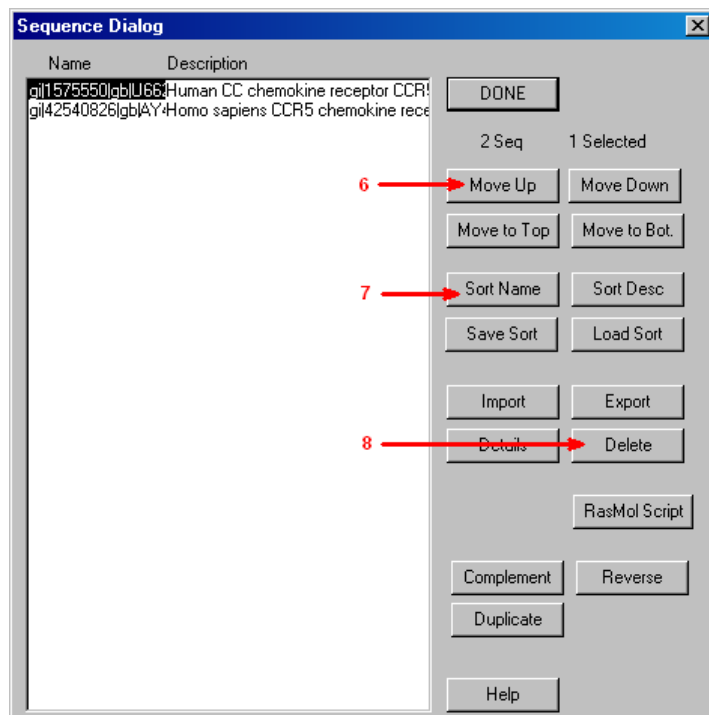
Dadurch wird die Alignierung der Sequenzen erheblich erleichtert. Zur Überprüfung der Übereinstimmungen kann immer wieder die Graustufen-Schattierung aller Sequenzen aktualisiert werden. Farbige Markierungen gehen dann verloren. Werden keine Treffer angezeigt, kann man entweder nach neuen Abschnitten suchen oder im Fenster ‚Find Dialog‘ auch eine Anzahl von Fehltreffern (Mismatches oder Insertion/Deletion) (→5) zulassen.



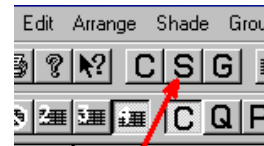
Veränderung der Reihenfolge der Sequenzliste, Entfernen von Sequenzen

Um eine Übereinstimmung verschiedener Sequenzen zu verdeutlichen, kann es sinnvoll sein, die Reihenfolge der Sequenzen zu verändern. Dazu wählt man in der Menü-Leiste ‚Project‘ > ‚Edit Sequence List‘ und hat nun die Möglichkeit, eine Sequenz durch Anklicken zu markieren und ihre Position durch den Knopf (Button) ‚Move Up‘ (hoch bewegen), ‚Move Down‘ (runter bewegen), ‚Move to Top‘ (an den Anfang setzen) oder ‚Move to Bottom‘ (ans Ende setzen) zu verändern (→6). Zudem besteht die Möglichkeit, Sequenzen nach Namen zu sortieren (→7).

Soll eine Sequenz aus dem Alignment entfernt werden, besteht ebenfalls unter ‚Project‘ > ‚Edit Sequence List‘ die Möglichkeit dazu. Durch Anklicken der Option ‚Delete‘ (Löschen) (→8) werden ausgewählte Sequenzen gelöscht.



Eine „Abkürzung“ zu diesen Optionen gibt es über den Knopf (Button) S:

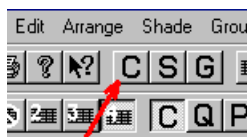


Speichern des Alignments

Ein Alignment kann für die weitere Bearbeitung in GeneDoc oder weitere Analysen in anderen Programmen gespeichert werden. In diesem Fall kann man entweder über ‚File’ > ‚Export’ die Datei im FASTA-Format oder beispielsweise im Clustal-Format speichern. Zu dem gleichen Ergebnis kommt man durch Anwahl von ‚Project’ > ‚Edit Sequence List’ > ‚Export’.

Zeichengröße und Drucken des Alignments

Um eine größere Übersicht über das Alignment zu erhalten, kann man auch die Zeichengröße ändern, indem man unter ‚Project’ > ‚Edit Sequence List’ die Option ‚Configure’ anwählt. In diesem Fenster kann man dann bei ‚Project, Font Settings’ die Zeichengröße von 12pt (Standard-Einstellung) verändern (Verkleinern der Zeichengröße wird zum Ausdrucken empfohlen). Eine „Abkürzung“ zu dieser Option gibt es über den Knopf (Button) C:



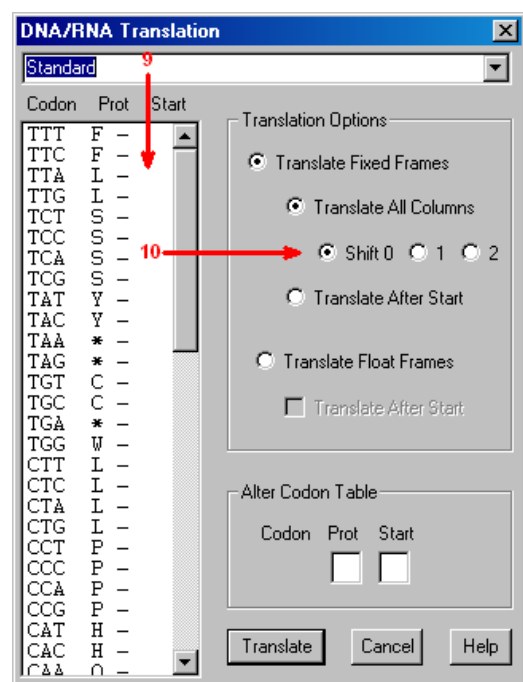
Ist das Alignment erfolgreich erstellt, kann man unter ‚File’ > ‚Print’ (Drucken) das Ergebnis ausdrucken. Es wird empfohlen, das Ergebnis im Querformat zu ausdrucken.

Translation der Sequenzen (DNA-Sequenz in Aminosäure-Sequenz)

In einem Alignment können nicht nur die Nukleotide einer DNA-Sequenz miteinander verglichen werden, sondern auch die daraus resultierenden Aminosäure-Sequenzen. Diese Translation übernimmt das Programm. Zunächst muss man jedoch alle Lücken (gaps), die zuvor für ein erfolgreiches Alignment eingefügt wurden, wieder entfernt werden (sie dienen lediglich der Veranschaulichung von Insertionen / Deletionen).

Um nun die Translation durchzuführen wählt man ‚Project’ > ‚Trans DNA to Protein’. Es öffnet sich ein kleines Fenster ‚DNA/RNA Translation’, in dem die Codierung der Aminosäuren aufgelistet ist (→9). An dieser Stelle hat man die Möglichkeit, eine Verschiebung des Leserasters („Shift“) um 1 oder 2 Nukleotide festzulegen, falls z.B. die Sequenz nicht mit einem vollständigen Triplet beginnt (→10). In einigen Fällen müssen mehrere Versionen ausgetestet werden. Dies eignet sich auch dafür, Schülerinnen und Schülern die Auswirkungen einer Leserasterverschiebung durch eine Mutation (Deletion oder Insertion) deutlich zu machen.

Die Aminosäure-Sequenz erscheint in einem neuen Fenster, so dass durch Schließen oder Verkleinern dieses Fensters wieder die DNA-Sequenz sichtbar wird.



Literatur und links:

GenBank® (National Center for Biotechnology Information: www.ncbi.nlm.nih.gov).

GeneDoc - Multiple Sequence Alignment Editor & Shading Utility Version 2.6.002, © 2000 by Karl Nicholas (www.nrbsc.org/gfx/genedoc/index.html)

Knoop, V. and K. Müller (2009). Gene und Stammbäume. Heidelberg, Spektrum Akademischer Verlag