Ihre Namen: _____

Übung 9: Molekulare Evolution II

Themen: *GenBank – BLAST – Orthologie – Paralogie – Einfluss von Selektion auf Sequenzvariation*

Teil 1: Identifizierung einer unbekannten DNA Sequenz

Im ersten Teil der heutigen Übung werden Sie sich näher mit GenBank und dem Programm BLAST beschäftigen und einen ersten Einblick in multiple Sequenzalignments erhalten.

Als Grundlage dazu verwenden Sie eine Studie von 2004:
Im Juli 2003 wurde an einem Strand in Los Muermos, Chile, eine große Gewebemasse gefunden ("Chilean Blob"), die zunächst nicht identifiziert werden konnte. Die Masse wog über 13 Tonnen und hatte einen Durchmesser von 12 Metern. Forscher weltweit mutmaßten lange darüber, was diese Masse sein könnte. Überreste einer unbekannten Spezies? Ein riesiger Oktopus beispielsweise?
Es gelang, DNA aus dieser Masse zu isolieren und zu sequenzieren. Diese DNA-Sequenz werden Sie verwenden, um dem Ursprung des "Chilean Blob" auf den Grund zu gehen.

Abbildung 1: Der "Chilean Blob" wie er am Strand von Pinuno in Chile gefunden wurde. Photo by Elsa Cabrera (© E. Cabrera, 2003).

Grundsätzlich lässt sich diese Vorgehensweise natürlich auch auf andere DNA-Sequenzen unbekannten Ursprungs anwenden, beispielsweise in der Aufklärung von Kriminalfällen.

Laden Sie die Datei "***MonsterBlob.fas***" von der Kursseite.

Hinweis: Öffnen Sie Dateien mit der Endung ***.fas*** immer mit einem Text-Editor (in Windows z. B.: Notepad/Wordpad), nicht mit Word!

Schauen Sie sich das Format dieser Datei an. Es handelt sich hierbei um das **Fasta-Format**. Sequenzen sind normalerweise in Textdateien gespeichert, welche mit einem Text-Editor geöffnet und bearbeitet werden können. Das Fasta-Format ist eines der am häufigsten genutzten Formate in der Biologie. Fasta-Dateien sind einfache Textdateien und können eine oder mehrere Sequenzen enthalten. Jeder Sequenz ist eine Kopfzeile (der sogenannte "Header") vorangestellt, welche mit ">" beginnt. Die Kopfzeile enthält in der Regel eine ID (Identifizierungsnummer) und evtl. weitere Informationen zur Sequenz. In der nächsten Zeile folgt die eigentliche Sequenz (Nukleotid- oder Aminosäuresequenz) – nichts anderes!

Die erste Aufgabe besteht darin herauszufinden, von welcher Art diese Sequenz stammt. Dazu werden Sie das Programm BLAST verwenden um diese Sequenz mithilfe der Internetdatenbank GenBank zu identifizieren.

a) Warum kann man mit dieser Vorgehensweise den Ursprung unserer unbekannten DNA-Sequenz untersuchen?

Gehen Sie auf die NCBI-BLAST-Webseite: http://blast.ncbi.nlm.nih.gov/Blast.cgi. Sie arbeiten heute ausschließlich im Bereich "***Basic BLAST***". Wählen Sie "***nucleotide blast***" und kopieren Sie die Sequenz in das Eingabefenster. Halten Sie alle Standardeinstellungen bei und klicken Sie auf "***BLAST***". Die Suche kann einige Zeit dauern.

b) Klicken Sie auf "***Search Summary***". Hier können Sie sehen wieviele Sequenzen und wieviele Nukleotide es in der Datenbank gibt, in der Sie gerade gesucht haben. Tragen Sie die Werte in die Tabelle mit ein.

Posted date	Jul 21, 2013 6:14 AM	Jun 11, 2014 3:13 AM	
Number of letters	48,071,108,788	59,495,734,569	
Number of sequences	18,983,527	23,396,504	

Im Vergleich zu den Zahlen aus den letzten Jahren können Sie sehen, dass sich diese Werte ändern. Die Datenbank GenBank ist nicht statisch und die Anzahl der Sequenzen/Buchstaben wächst ständig.

c) Gehen Sie in den Ergebnissen bis zu "***Sequences producing significant alignments***". Wie sind diese Sequenzen sortiert? Welche Sequenzen sind "***MonsterBlob.fas***" am ähnlichsten?

d) Schauen Sie sich den GenBank Eintrag an. Klicken Sie dazu rechts unter "***Accession***" auf die Akzessionsnummer. Klassifizieren Sie die Art, die Sie gerade identifiziert haben. Um welches Tier handelt es sich?

Teil 2: Ortholog vs. Paralog

Im zweiten Teil der heutigen Übung werden Sie sich die Verwendung von BLAST etwas genauer ansehen. Ziel ist es auch zu verstehen, wie wichtig es ist, die richtigen Treffer für eine gegebene Fragestellung in der Datenbank zu wählen. Dies wird am Konzept orthologer und paraloger Sequenzen verdeutlicht.

2.1 Schauen Sie sich zunächst die aus Teil 1 bereits bekannte NCBI-BLAST-Webseite http://blast.ncbi.nlm.nih.gov/Blast.cgi etwas genauer an. Neben dem in Teil 1 verwendetem "***nucleotide blast***", gibt es weitere BLAST Programme im Bereich "***Basic BLAST***". Finden Sie heraus, welche Möglichkeiten es dort gibt.

Tipp: Sie können es mit der Nukleotidsequenz vom „MonsterBlob" aus Teil 1 ausprobieren...

Laden Sie nun die Datei "***unknown.fas***" von der Kursseite. Wählen Sie "***protein blast***" und verwenden Sie "***unknown.fas***" für eine Blast-Suche mit ***blastp***. Halten Sie zunächst alle Standardeinstellungen bei. Die Suche kann eine Weile dauern. Die Ergebnisseite ist ähnlich wie beim "***nucleotide blast***". Zusätzlich werden hier auch noch konservierte Domänen innerhalb der Proteinsequenz graphisch angezeigt.

a) Schauen Sie sich Ihre Ergebnisse an. Ist die Anfrage-Sequenz in der Datenbank enthalten? Können Sie die unbekannte Sequenz identifizieren? Aus welcher Art kommt diese Sequenz?

2.2 Die gefundene Sequenz wird nun verwendet, um das Prinzip der Orthologie und Paralogie etwas genauer zu untersuchen.

a) Wie könnten Sie bei Ihrer BLAST-Suche geschickt vorgehen, wenn Sie herausfinden wollten, ob es zur der unbekannten Sequenz paraloge Sequenzen gibt und welche dies sind? Tipp: Schauen Sie sich die Einstellungsoptionen in der BLAST-Suchmaske an.

b) Wie würden Sie geschickt in der Datenbank nach orthologen Sequenzen suchen?

c) Nehmen Sie an, dass Sie eine Gensequenz nutzen wollen um die Beziehungen zwischen Arten zu untersuchen. Nehmen Sie auch an, dass dieses Gen in der Vergangenheit dupliziert wurde und dass daher Orthologe und Paraloge vorliegen können. Was müssen Sie bei der Suche von geeigneten Sequenzen bei der Erstellung des Baumes auf jeden Fall beachten?

d) Welche Gene verwenden Sie? Orthologe oder Paraloge?

e) Was kann passieren, wenn Sie das nicht beachten?

Im Folgenden werden sie **zwei Dateien** mit mehreren Sequenzen im Fasta-Format erstellen. Achten Sie auf eine eindeutige Benennung der Sequenzen und darauf, dass jeder Header (>...) nur **einmal** vorkommt! **Notieren Sie sich bitte auch die Akzessionsnummern aller Sequenzen die Sie auswählen.**

f) Erstellen Sie nun eine Text-Datei mit einem Editor Ihrer Wahl (nicht Word!) und speichern Sie darin die zuvor unbekannte Sequenz "***unknown***" (als >h2) im Fasta-Format (Kopieren/Einfügen). Nennen Sie diese Datei "***h2Sequenzen.fas***". Im Folgenden werden noch weitere Sequenzen dazu kommen.

g) Suchen Sie nun zu ihrer Sequenz "***unknown***" (>h2) eine paraloge Sequenz (Tipp: siehe 2.2a). Verwenden Sie dazu wieder ***blastp***. Für welches Paralog haben Sie sich entschieden? (Hinweis: Paraloge werden gewöhnlich durchnummeriert)

Akzessionsnummer:

Erstellen Sie eine weitere Text-Datei und nennen Sie die Datei "***h1Sequenzen.fas***". Dann speichern Sie darin das Paralog (>h1) indem Sie in der Liste rechts auf die Akzessionsnummer "***Accession***" klicken. Sie werden auf eine Website weitergeleitet, die alle Informationen zu dieser Sequenz beinhaltet. Klicken Sie oben links auf "***FASTA***". Um die Sequenz abzuspeichern, kopieren Sie diese einfach in ihre Text-Datei (fasta-Format beibehalten).

h) Suchen Sie als nächstes Orthologe zu ihrer Sequenz h2 in folgenden Arten:

- Maus (*Mus musculus*) [>h2_maus], Akzessionsnummer:
- Zebrafisch (*Danio rerio*) [>h2_fisch], Akzessionsnummer:
- Huhn (*Gallus gallus*) [>h2_huhn], Akzessionsnummer:

Denken Sie auch hier wieder daran, passende Einstellungen im Suchfenster vorzunehmen, um die Sequenzen in den jeweiligen Arten leicht finden zu können. Wählen Sie jeweils den besten Treffer und speichern Sie die Sequenz in "***h2Sequenzen.fas***" (über Kopieren/Einfügen wie zuvor beschrieben).

Die Fruchtfliege *Drosophila melanogaster* [>fliege] wird als Außengruppe verwendet. Nehmen Sie den besten Treffer und speichern Sie die Sequenz in Ihrer Datei "***h2Sequenzen.fas***" (Kopieren/Einfügen).

Akzessionsnummer:

i) Als nächstes bestimmen Sie die Orthologe zu Ihrer Sequenz h1 aus den gleichen Arten und speichern den jeweils besten Treffer in "***h1Sequenzen.fas***" (Kopieren/Einfügen):

- Maus (*Mus musculus*) [>h1_maus], Akzessionsnummer:
- Zebrafisch (*Danio rerio*) [>h1_fisch], Akzessionsnummer:
- Huhn (*Gallus gallus*) [>h1_huhn], Akzessionsnummer:

Jetzt haben Sie eine Text-Datei pro Paralog und zu jedem Paralog orthologe Sequenzen aus weiteren Vertebraten. Außerdem haben Sie eine Außengruppe (*D. melanogaster*) zu beiden Paralogen. Die Dateien sollten folgendermaßen aussehen:
"***h2Sequenzen.fas***" mit h2, h2_maus, h2_fisch, h2_huhn, fliege
"***h1Sequenzen.fas***" mit h1, h1_maus, h1_fisch, h1_huhn
Bevor Sie weitermachen, zeigen Sie die beiden Text-Dateien einem Tutor oder Dozenten.

j) Bevor Sie nun anfangen Bäume zu erstellen, überlegen Sie sich A) wie der Baum aussieht, wenn Sie nur die Sequenzen aus "***h2Sequenzen.fas***" verwenden und B) wie der Baum aussieht, wenn Sie die Sequenzen aus beiden Text-Dateien ("***h2Sequenzen.fas***" und "***h1Sequenzen.fas***") verwenden.
Skizzieren Sie beide Bäume:
A) B)

k) Als nächstes werden Sie diese beiden Bäume erstellen. Dafür wird das Programm MEGA verwendet. Öffnen Sie MEGA: "***Alle Programme > MEGA5/MEGA***". Gehen Sie zunächst auf "***Open A File/Session***" und öffnen Sie die Datei "***h2Sequenzen.fas***" um Baum A zu erstellen. Bevor Sie einen Baum erstellen können, müssen die Sequenzen aligniert werden (Das Alignment wird in einem neuen Fenster angezeigt werden). Wählen Sie "***Align***", dann gehen Sie auf "***Alignment***" und wählen Sie "***Align by ClustalW***". Sie werden gefragt ob Sie alle Sequenzen verwenden wollen, klicken Sie auf "***OK***". Behalten Sie die Standardeinstellungen bei, klicken Sie dazu wieder auf "***OK***". Ihre Sequenzen werden nun aligniert.
Um aus diesem Alignment einen Baum erstellen zu können, müssen Sie es erst aktivieren. Gehen Sie dazu auf "***Data > Phylogenetic Analysis > Yes***". Ihr Alignment ist nun im ersten MEGA-Fenster aktiv und Sie können den Baum erstellen. Wählen Sie "***Analysis > Phylogeny > Construct/Test Neighbor-Joining Tree***", klicken Sie auf "***Yes***" (um das aktive Alignment zu bestätigen) und "***Compute***". Baum A wird nun erstellt. Speichern Sie ihn ab: "***Image > Save as PDF***". Schließen Sie alle MEGA-Fenster.
Erstellen Sie als nächstes Baum B. Dazu müssen Sie zunächst eine FASTA-Datei, die alle Sequenzen enthält erstellen:"***h1undh2Sequenzen.fas***". Öffnen Sie nun diese Datei in MEGA und erstellen Sie Baum B.

Skizzieren Sie die entstandenen Bäume. Entspricht das Ergebnis Ihren Erwartungen?
A) B)

l) Nehmen Sie an, dass Sie nichts über die analysierten Arten wüssten und herausfinden wollten, in welchem Verwandtschaftsverhältnis die Arten stehen. Was würde passieren, wenn Sie nur folgende Sequenzen ausgewählt hätten: ***h1***, ***h1_fisch***, ***h2_maus***, ***h2_huhn***, und ***fliege***? Nehmen Sie dazu Baum B aus 2.2k und skizzieren Sie einen Teilbaum aus diesen fünf Sequenzen. Was würde dies für die Abschätzung der Verwandtschaften der untersuchten Arten bedeuten (Tipp: siehe 2.2c-e)?

Teil 3: Angewandte Populationsgenetik

In diesem Teil werden Sie sich den Einfluss von Selektion auf die Sequenzvariation innerhalb von Populationen ansehen. Dafür verwenden Sie zwei Beispiele, die aus Studien aus unserer Arbeitsgruppe stammen. Desweiteren lernen Sie ein Standard-Programm zur populationsgenetischen Analyse von Datensätzen kennen.

3.1 Evolution von Genfamilien

Genduplikationen kommen in Pflanzen häufig vor. Die daraus entstehenden Paraloge können unterschiedliche evolutionäre Wege einschlagen. Dazu gehören Konservierung, Differenzierung, Subfunktionalisierung und Pseudogenisierung. In dieser Übung verwenden Sie einen Datensatz von *Solanum peruvianum*, einer Wildtomatenart aus Südamerika. Dieser Datensatz umfasst zwei Gene der *CBF* Genfamilie. Die *CBF* Gene sind Transkriptionsfaktoren und Teil der molekularen Antwort auf Kältestress. Nach neuesten Erkenntnissen ist *CBF2* ein negativer Regulator von *CBF3*.
a) Welchen evolutionären Weg haben *CBF2* und *CBF3* möglicherweise genommen?

Als erstes werden Sie die genetische Variation dieser beiden Paraloge innerhalb einer Population miteinander vergleichen. Dazu laden Sie zunächst die Dateien ***CBF2.nex*** und ***CBF3.nex*** von der Kursseite herunter. Starten Sie die Software DnaSPv5. Im Menü dieser Software gehen Sie nun auf "***File/Open Data File***" und öffnen zunächst **"*CBF2.nex*"**. Unter "***Display/View Data***" können Sie sich den Datensatz genauer ansehen. Die Sequenzen beginnen mit den Buchstaben "TAR". TAR steht für Tarapaca und bezeichnet die *S. peruvianum* Population aus der diese Sequenzen stammen.

b) Zunächst bestimmen Sie die Polymorphismus-Werte (= genetische Variation) von *CBF2*. Gehen Sie auf "***Analysis/DNA polymorphism***". Klicken Sie auf "***Ok***". Die Ergebnisse der Analyse erscheinen nun in einem türkisen Fenster auf dem Bildschirm. DnaSP produziert eine ganze Reihe von Statistiken. Uns interessieren aber nur *S* (Anzahl der segregierenden Stellen), Theta (basierend auf *S*) und Pi (Nukleotiddiversität). Tragen Sie diese Werte für *CBF2* in die Tabelle ein. Falls Sie sich nicht sicher sind welches die richtigen Werte sind, fragen Sie die Dozenten/Tutoren.

	CBF2	*CBF3*
Länge ["total number of sites"]		
S		
Theta (θ_W)		
Pi (π)		

Wenn Sie mit *CBF2* fertig sind gehen Sie auf "***File/Close Data File***" und wiederholen Sie die einzelnen Schritte für *CBF3*.
c) Vergleichen Sie die Polymorphismus-Werte der beiden Gene. Was stellen Sie fest? Würden Sie sagen, dass diese beiden Gene neutral evolvieren? Begründen Sie ihre Antwort.

Eine Möglichkeit herauszufinden unter welchem Selektionstyp ein Gen evolvieren könnte, besteht darin, die Polymorphismus-Werte dieses Genes mit den durchschnittlichen Polymorphismus-Werten der Population zu vergleichen.
d) Nehmen Sie nun an, dass die durchschnittlichen Polymorphismus-Werte der Population bei $\theta_W = 0{,}0238$ und $\pi = 0{,}0226$ liegen. Unter welchem Selektionstyp könnte *CBF2* evolvieren? Unter welchem *CBF3*?

3.2 Selektion in *Drosophila melanogaster*

In dieser Übung verwenden Sie einen Datensatz von *Drosophila melanogaster* und betrachten wieder die genetische Variation innerhalb einer Population.
Eine beliebte Methode Selektionsereignisse im Genom zu finden besteht darin das Genom nach Regionen zu durchsuchen, die vom Durchschnitt abweichen. Dies können Regionen mit erhöhter oder verringerter Variation sein. Eine solche Region werden Sie sich in dieser Übung anschauen. Es handelt sich um eine Region auf dem X Chromosom.

Laden Sie die Datei ***phpRegion_africa_Dmel.fas*** von der Kursseite herunter, speichern Sie die Datei und öffnen Sie diese mit DnaSPv5. Unter "***Display/View Data***" können Sie sich den Datensatz genauer ansehen. Die Buchstaben "RG" (z. B., RG10, RG11 usw.) stehen für die *D. melanogaster* Population aus der diese Sequenzen stammen: Ruanda, Afrika.

Um die Polymorphismen (= genetische Variation) in dieser genomischen Region untersuchen, gehen Sie auf "***Analysis/DNA polymorphism***". Klicken Sie auf "***Sliding Window Analysis***" und setzen Sie die Fenstergröße ("***window length***") auf 1000 und die Anzahl der Schritte ("***steps***") auf 100. Wählen Sie "***Ok***". Die Ergebnisse der Analyse erscheinen nun in einer Tabelle auf dem Bildschirm.

a) Wählen Sie nun "***DnaSP Graph***" und sehen Sie sich die "Sliding Windows" der verschiedenen Statistiken (Pi, Theta, *S*) an. Beschreiben Sie zunächst den Graphen. Was ist auf der x- was auf der y-Achse? Können Sie erkennen was mithilfe einer "***Sliding Window Analyse***" dargestellt werden kann?

b) Beschreiben Sie was Sie in den "Sliding Windows" sehen. Können Sie eine evolutionäre Erklärung für diese Beobachtung geben?

c) Vergleichen Sie die “Sliding Windows” für die drei Statistiken. Fällt Ihnen etwas auf? Wie können Sie es erklären?

Bitte alle heute heruntergeladenen und/oder erstellten Dateien vom Computer löschen!

Zusätzliche Literatur:

Monster Blob:

Pierce, S., S. Massey, N. Curtis, G. Smith, C. Olavarría & T. Maugel 2004. Microscopic, Biochemical, and Molecular Characteristics of the Chilean Blob and a Comparison With the Remains of Other Sea Monsters: Nothing but Whales. Biological Bulletin 206: 125–133.

CBF Genfamilie:

Mboup, M., I. Fischer, H. Lainer & W. Stephan 2012. Trans-Species Polymorphism and Allele-Specific Expression in the *CBF* Gene Family of Wild Tomatoes. Mol Biol Evol 29: 3641-3652.