

Praktikumsteil Molekulare Systematik (Christian Printzen)

Molekulare Methoden sind nicht nur Grundlage der Genetik oder Mikrobiologie, sie haben auch die sogenannte „organismische Biologie“ revolutioniert. Die gesamte Geschichte eines Organismus, von seinen evolutionären Wurzeln bis zu seiner familiären Vergangenheit, hinterlässt Spuren im Genom. Sind Transposons die phylogenetischen Vorgänger von Retroviren? Wie oft hat sich Carnivorie in der Evolution der Pflanzen entwickelt? Kann die Plattentektonik das Auftreten flugunfähiger Vögel in Australien, Afrika und Südamerika erklären? Wie soll man die letzten Bestände vom Aussterben bedrohter Pflanzen- und Tierpopulationen managen, um ihre genetische Variabilität zu erhalten? Wer molekulare Marker – DNA oder Proteine – verwendet, kann biologische Fragestellungen angehen, die weit über das bloße Verständnis zellulärer Vorgänge hinausreichen. Dieser Praktikumsteil versucht, Ihnen die Arbeitsschritte der Datengewinnung und -analyse in der modernen Systematik zu zeigen. Die im Kurs durchgenommenen Arbeitsschritte sind genau die gleichen, die bei der wissenschaftlichen Arbeit anfallen. Sie können dieses Praktikum darum als einen „Kochkurs“ für molekulare Systematik betrachten.

Der Kurs besteht aus zwei Teilen: Datengewinnung (hier von DNA-Sequenzdaten) und Datenauswertung am Computer. Nach dem ersten Teil wird eine kurze Pause eingelegt, da wir wie viele andere Labors keinen Sequenzierer besitzen und die Sequenz-Chromatogramme von einem externen Labor erstellen lassen. Auch dies entspricht dem normalen Arbeitsablauf in den meisten Forschungslabors. Dieses Skript soll Sie durch den Kurs führen, indem es in Kurzfassung die Hintergründe der molekularen Systematik darstellt. Die theoretischen Teile werden in der Vorbesprechung genauer durchgenommen. Zusätzlich finden Sie (gelb hinterlegt) genaue Anweisungen zu den einzelnen Arbeitsschritten, die Sie im Labor oder am Computer nacheinander „abhaken“ können.

Die Labormethoden zur Gewinnung von DNA-Sequenzen sind technisch nicht anspruchsvoll, auch wenn die wenigsten Protokolle gleich von Beginn an funktionieren. Statistische Verfahren zur Berechnung von phylogenetischen Stammbäumen sind dagegen nicht immer leicht zu verstehen. Das Verhalten dieser Methoden unter bestimmten evolutionären Bedingungen ist vielfach noch schlecht erforscht. An der Verbesserung wird laufend gearbeitet, regelmäßig werden auch neue Verfahren vorgeschlagen. Dieser Kurs kann deshalb nur einen ersten Einblick vermitteln. Für die vielen spannenden Fragen (und Anfälle von Verzweiflung) gibt es später ggf. Betreuer und erfahrene Mitglieder der Arbeitsgruppe.

DNA-Sequenzierung, eine kurze Übersicht

DNA zu „sequenzieren“ bedeutet, die Reihenfolge der Nukleotid-Basen – die Basensequenz – eines DNA-Abschnitts zu ermitteln. Die für den Organismus wichtigen Informationen zur Synthese von Proteinen sind in Form von Basentriplets (Dreiergruppen von Nukleotidbasen) auf der DNA gespeichert. Daneben gibt es bei Eukaryoten eine große Menge nicht-codierender DNA, die scheinbar keinerlei Informationen trägt. DNA-Sequenzen werden in der Regel unverändert von einer Generation an die nächste weiter gegeben. Durch gelegentliche Fehler bei der DNA-Replikation (Mutationen) summieren sich aber doch im Laufe der Zeit Sequenz-Unterschiede zwischen den Organismen. Die für Systematiker wichtigen Informationen sind genau diese erblich fixierten Veränderungen (Substitutionen), die Sequenzen im Laufe der Evolutions- oder Populationsgeschichte durchmachen. Vereinfacht kann man sagen: Je näher zwei Organismen miteinander verwandt sind, desto ähnlicher sind sich ihre DNA-Sequenzen. Es geht in der molekularen Systematik also darum, anhand von DNA- (oder Protein-) Sequenzunterschieden die Evolutionsgeschichte der Organismen nachzuvollziehen. Dies ist ein Prozess, der viele Arbeitsschritte umfasst. Die folgende Zusammenfassung soll Ihnen die Übersicht erleichtern.

- 1) Der erste Schritt der DNA-Sequenzierung ist die **Extraktion der DNA** aus Gewebeproben. Die DNA muss aus den Zellen freigesetzt und alle anderen Zellbestandteile beseitigt werden.
- 2) Ein oder mehrere vorher bestimmte DNA-Abschnitte müssen in so großer Konzentration vorliegen, dass man ihre Basensequenz bestimmen kann. Die **Polymerase-Kettenreaktion (PCR)** dient dazu, die ausgewählten DNA-Abschnitte zu vermehren.
- 3) Das Endprodukt der PCR-Reaktion dient als Ausgangsprodukt für die eigentliche Sequenzierreaktion. Bevor diese Reaktion gestartet werden kann, müssen durch **Reinigung der PCR-Produkte** alle im nächsten Schritt störenden Substanzen entfernt werden.
- 4) Das gereinigt PCR-Produkt wird in der **Sequenzierreaktion** einer weiteren PCR unterzogen. Diesmal verwendet man nur einen Primer, so dass es zu keiner Verdoppelung der DNA mehr kommt. Meist möchte man zur Sicherheit und zum Datenabgleich beide DNA-Stränge sequenzieren und setzt deshalb zwei oder mehr Reaktionen mit jeweils unterschiedlichen Primern an.
- 5) Im letzten Schritt wird das Produkt der Sequenzierreaktion in einem automatischen **DNA-Sequencer** elektrophoretisch aufgetrennt, wobei mit Hilfe von Fluoreszenzmarkern die DNA-Sequenz bestimmt wird. Diesen Schritt lassen wir von einem kommerziellen Labor ausführen.

Kursziele

- 1) Nach dem ersten Kursteil sollen Sie in der Lage sein, ohne weitere Anleitung:
 - a. Aus Pflanzen- oder Pilzmaterial DNA zu extrahieren,
 - b. PCR-Reaktionen mit vorher ausgewählten Primern anzusetzen,
 - c. Den Erfolg Ihrer PCR-Reaktionen auf Agarose-Gelen zu dokumentieren,
 - d. PCR-Produkte zu reinigen,
 - e. Sequenzierreaktionen anzusetzen und
 - f. Das Produkt der Sequenzierreaktion zu reinigen und zu trocknen.
- 2) Nach dem zweiten Kursteil sollten Sie in Grundzügen mit folgenden Begriffen und Arbeitsverfahren vertraut sein.
 - a. Editieren und alignen von Sequenzen
 - b. Phylogenetische Bäume und Methoden zu ihrer Berechnung
 - c. Berechnungsmethoden für phylogenetische Unsicherheit
 - d. Phylogenetische Hypothesen und einfache Testmethoden
- 3) Nach den Vorbesprechungen und eigener Lektüre sollten Sie in der Lage sein, einfachere molekularsystematische Publikationen zu verstehen und Ihre eigenen Daten mit kritischen Augen zu betrachten.

Aufgabe vor Beginn des Kurses: Lesen Sie Kapitel 1 und 2 aus: „von Haeseler, A. & Liebers, D. (2003) *Molekulare Evolution*. 128 pp. S. Fischer Verlag, Frankfurt.“ (8,90 €).

Datengewinnung: DNA-Sequenzierung, 1. Arbeitstag

Aufarbeitung des Pflanzenmaterials

Der erste Arbeitsgang ist der einfachste: Zerstören Sie das Pflanzenmaterial bis zur Unkenntlichkeit! Dieser Schritt ist notwendig, weil im nächsten Schritt die DNA aus den Zellen freigesetzt werden soll. Je größer die Oberfläche, desto schneller wirken die dazu notwendigen Enzyme. Trotzdem heisst es vorsichtig vorgehen: Handschuhe tragen und die DNA nicht unnötigen mechanischen Belastungen aussetzen!

1. Mischen Sie Puffer C1 aus dem NucleoSpin Plant Extraktionskit gut und stellen Sie sicher, dass er warm ist.
2. Stellen Sie ein Becherglas bereit.
3. Bereiten Sie ihre Pflanzenproben vor: geben Sie je ca. 100 mg Pflanzenmaterial und eine Spatelspitze Sand in ein Eppendorf-Gefäß (1,5 ml).
4. Beschriften Sie die Gefäße unverkennbar.
5. Erhitzen Sie den Puffer C1 im Wasserbad oder auf dem Heizblock auf 45 °C.
6. Bereiten Sie ein kleines Styroporgefäß mit flüssigem Stickstoff vor.
7. Stecken Sie ein Plastikpistill in das Gefäß.
8. Tauchen Sie das Gefäß bis zur Hälfte in den Stickstoff und warten Sie einige Sekunden (Vorsicht: Finger nicht mit einfrieren!).
9. Zermahlen Sie das Material mit dem Pistill, ohne das Gefäß zu zerbrechen, und ohne dass das Material auftaut (ggf. erneut in den Stickstoff!).
10. Arbeiten Sie zügig weiter mit Schritt 11.

DNA-Extraktion

(Während der Extraktion wieder mit Handschuhen arbeiten!)

11. Mischen Sie Puffer C1 und RNase A mit der Pipette gut durch. Pipettieren Sie 400 µl C1 und 10 µl RNase A auf das zerkleinerte Pflanzenmaterial. Rühren Sie mit dem Pistill gut um.
12. Prüfen Sie, ob alles Pflanzengewebe homogenisiert ist. Zerquetschen Sie ggf. größere Stückchen an der Wand des Gefäßes.
13. Legen Sie das Pistill in das Becherglas und schließen Sie das Eppendorf-Gefäß.
14. Rühren Sie auf dem Schüttler kurz durch (nicht zu lange, genomische DNA ist empfindlich gegen mechanische Belastungen).
15. Stellen Sie das Gefäß bei 70 °C in den Heizblock.
16. Gehen Sie zurück auf 1. und fahren unmittelbar mit der nächsten Probe fort. Benutzen Sie **frische Pipettenspitzen** für jede einzelne Extraktion!
17. Wenn alle Proben versorgt sind und im Wasserbad (Heizblock) stehen, warten Sie 30 Minuten. In dieser Zeit drehen Sie die Gefäße 3-4 mal auf den Kopf und stellen sie sie zurück, um den Inhalt neu zu vermischen. Währenddessen die folgenden Arbeitsschritte vorbereiten.
18. Mischen Sie 4 Teile Puffer C2 und 1 Teil Puffer C3 (ergibt Puffer C4), und dann 3 Teile Puffer C4 mit 2 Teilen Ethanol. Mischen Sie mit der Pipette gut um. Sie benötigen für jede Probe 500 µl C4/ Ethanol. Berechnen Sie die benötigte Gesamtmenge (mit etwas „Pipettierverlust“).

19. Fügen Sie die erforderliche Menge Ethanol zu Puffer C5 hinzu und mischen.
20. Beschriften Sie für jede Ihrer Proben 1 Eppendorf-Gefäß (1,5 ml), 1 Eppendorf-Gefäß (2 ml) mit einer NucleoSpin Säule, 1 weiteres Eppendorf-Gefäß (1,5 ml).
21. Nach 30-45 Minuten, zentrifugieren Sie das DNA-Lysat 5 min. bei maximaler Umdrehungszahl.
22. Pipettieren Sie für jedes Extrakt 500 µl Puffer C4/ Ethanol in ein neues Eppendorf-Gefäß (1,5 ml).
23. Pipettieren Sie den Überstand aus den zentrifugierten DNA-Lysaten hinzu. Nehmen Sie möglichst nichts vom Bodensatz aber möglichst viel DNA-Lysat mit. **Pipettenspitzen wechseln!**
24. Pipettieren Sie 500 µl des Lysats auf die Trennsäule. **Pipettenspitzen wechseln!** Verschiessen Sie diese.
25. Zentrifugieren Sie 1 min. bei max. Umdrehungszahl. Verwerfen Sie den Durchlauf.
26. Wiederholen Sie Schritte 23 und 24 mit dem Rest des Lysats bis alles auf die Säule aufgetragen ist.
27. Waschen Sie die Säule mit 400 µl Puffer CW. 1 min. bei max. Umdrehungszahl zentrifugieren. Durchlauf verwerfen. (Wenn Sie die Trennsäule nicht berühren, können Sie bei den Schritten 26-28 Pipettenspitzen mehrfach verwenden. Nach jeder versehentlichen Berührung Pipettenspitzen wechseln.)
28. Waschen Sie die Säule mit 700 µl Puffer C5. 1 min. bei max. Umdrehungszahl zentrifugieren. Durchlauf verwerfen.
29. Waschen Sie die Säule mit 200 µl Puffer C5. 2 min. bei max. Umdrehungszahl zentrifugieren, um die Membran völlig zu trocknen. Den Durchlauf mitsamt dem 2-ml-Gefäß (aber ohne die grüne Säule!) verwerfen.
30. Grüne Säule in das letzte 1,5-ml-Gefäß stellen. 100 µl Puffer CE auf die Mitte der Membran pipettieren, ohne diese zu berühren und 5 min. bei Zimmertemperatur inkubieren. **Pipettenspitzen wechseln!**
31. 1 min. bei max. Umdrehungszahl zentrifugieren. Trennsäule verwerfen. Gefäß verschliessen und nochmals sauber beschriften.

Nach dem letzten Schritt halten Sie nun die extrahierte DNA in einer Pufferlösung in Ihren Händen. Im einzelnen haben Sie in Schritt 1-7 das Material mechanisch aufgeschlossen und in Schritt 9-16 die Zellmembranen lysiert und die DNA freigesetzt. In Schritt 20-21 haben Sie feste Bestandteile und die meisten Polysaccharide entfernt. In Schritt 22-25 die DNA an eine Trägermembran gebunden. Das so gebundenen Material haben Sie in 26-28 von Proteinresten, weiteren Polysacchariden, Nukleotiden und anorganischen Zellbestandteilen gereinigt. Die Lösung ist im Kühlschrank nun mehrere Jahre lang haltbar. Frieren Sie DNA-Lösungen nur ein, wenn Sie sicher sind, dass sie nur noch sehr selten gebraucht werden. Häufiges Auftauen und Einfrieren zerstört die DNA.

Sollte es Probleme bei der PCR geben (s.u.), kann man versuchen, statt mit Elutionspuffer mit Wasser zu eluieren, dem danach 10% TE-Puffer beigelegt wird. Das in vielen Elutionspuffern in höherer Konzentration vorhandene EDTA bindet Mg-Ionen, was in der PCR oft zu schlechter Produktausbeute führt.

Datengewinnung: DNA-Sequenzierung, 2. Arbeitstag

PCR-Reaktion

Die Polymerase-Kettenreaktion (polymerase chain reaction = PCR) ist eine Abfolge biochemischer Reaktionen, die bei verschiedenen Temperaturen in einem einzigen Reaktionsgefäß ablaufen. Im Grunde simuliert die PCR-Reaktion die in der Natur vor der Zellteilung erfolgende Replikation der DNA im Reagenzglas. Die isolierte genomische DNA wird in einer Pufferlösung mit einer DNA-Polymerase, Mg^{2+} -Ionen und Desoxyribonukleotiden (dNTPs) zusammengebracht. Damit die Polymerase die DNA als Vorlage verwenden kann und mit den in der Lösung vorhandenen Nukleotiden einen Komplementärstrang synthetisieren kann, müssen die beiden Stränge der DNA zuerst getrennt (denaturiert) werden. Dies geschieht bei 90-95 °C. Zusätzlich benötigt das Enzym einen Anfangsstrang, an den es weitere Nukleotide anfügen kann. Diesen kurzen Strang fügt man in Form von zwei „Primern“ hinzu. Diese kurzen Oligonukleotide binden bei 40-60 °C an passende, komplementäre DNA-Abschnitte (annealing). Damit hat man die Möglichkeit zu bestimmen, welche Teile der DNA man amplifizieren will. Die am häufigsten verwendete Polymerase stammt aus dem in heißen Quellen lebenden Bakterium *Thermophilus aquaticus* und ist deshalb sehr hitzestabil (*taq*-Polymerase). Die optimale Reaktionstemperatur liegt bei 72° C. Der Zyklus von DNA-Denaturierung, Primer-annealing und DNA-Synthese wird in einem automatischen Thermocycler 30 bis 40 mal durchlaufen, wobei sich die Menge des ausgewählten DNA-Abschnitts im (nie erreichten) Idealfall jedesmal verdoppelt.

Frage: Warum werden für die PCR zwei Primer verwendet, die am 5'- und 3'-Ende des zu vermehrenden Abschnittes jeweils auf den Komplementärsträngen sitzen?

PCR-Reaktionen sind sehr launisch, geringe Veränderungen im Protokoll führen oft zu kompletten Fehlschlägen. Trotzdem werden in verschiedenen Labors oft ganz unterschiedliche Protokolle verwendet. Das am häufigsten verwendete Verfahren zur Optimierung von PCR Reaktionen ist „Versuch und Irrtum“: Verläuft die PCR schließlich wie erwünscht, verändert man das Protokoll nicht mehr (oder nur noch, wenn man testen will, ob sich das Resultat noch verbessern lässt).

Multiple Banden, unsaubere Banden oder fehlendes PCR-Produkt sind die häufigsten Probleme. Fehlendes PCR-Produkt z. B. kann die unterschiedlichsten Ursachen haben:

- Das Extraktionsprotokoll hat versagt; das Eluat enthält keine oder zu wenig DNA.
- Man hat zuviel DNA hinzugefügt (auch das verursacht Probleme).
- Die Ausgangs-DNA ist degeneriert (z. B. durch häufiges Auftauen und wieder Einfrieren).
- Der DNA-Extrakt enthält Stoffe (z. B. Polysaccharide, EDTA), die die PCR stören.
- Die Polymerase ist überaltert/ nicht tiefgefroren gelagert.
- Einer oder beide Primer sind überaltert/ nicht tiefgefroren gelagert.
- Die Primer passen nicht auf die Bindungsstellen (kann sogar bei angeblich universellen Primern hin und wieder geschehen).
- Man hat eine Zutat im Ansatz vergessen.
- Die Annealing-Temperatur ist zu hoch.
- Der Thermocycler ist defekt.

Der Lösung kann man nur auf die Spur kommen, wenn man der Reihe nach die verschiedenen Möglichkeiten ausschließt. Im Zweifelsfall empfiehlt es sich, erfahrene Kollegen um Rat zu fragen. Im Internet finden sich viele Seiten, in denen durch Fehlversuche gestählte Biologen Tips zu „PCR-troubleshooting“ geben.

(z. B. http://206.53.227.20/prod_inf/manuals/pcr_man/Chapter02/CHAP01-Seite29.htm).

Die Polymerase-Kettenreaktion erfolgt in PCR-Gefäßen (0,2 ml). Obwohl die optimale Reaktionstemperatur bei 72° C liegt, ist die Polymerase auch bei Zimmertemperatur schon aktiv und beginnt, hier und da Nukleotide an Primer anzubauen. Das stört im allgemeinen die spätere PCR-Reaktion empfindlich, weshalb das **Ansetzen der PCR auf Eis** erfolgt. Teurere sog. „hot start“ Enzyme enthal-

ten Antikörper, die die Polymerase deaktivieren. Durch mehrminütiges Vorerhitzen auf 96 °C im Thermocycler (s.u.) werden die Antikörper denaturiert und das Enzym aktiviert. Mit solchen Enzymen kann man auch bei Zimmertemperatur arbeiten.

1. Ziehen Sie Handschuhe an und tauen Sie die folgenden Zutaten auf:
 - destilliertes und autoklaviertes Wasser
 - PCR-Puffer
 - MgCl₂
 - dNTP-Mix
 - Primer C und F
2. Beschriften Sie für jeden PCR-Ansatz und für eine Nullprobe jeweils 1 PCR-Gefäß (0,2 ml) und stellen Sie sie auf Eis.
3. Die Polymerase ist trotz Aufbewahrung im Gefrierschrank flüssig und wird direkt auf Eis überführt.
4. Ihr DNA-Extrakt überführen sie direkt vom Kühlschrank auf Eis.

Abgesehen von der Ausgangs-DNA enthalten Ihre PCR-Ansätze die gleichen Chemikalien in gleicher Konzentration. Um sich unnötiges Pipettieren zu ersparen, stellt man daher zuerst einen sog. „Mastermix“ her, der alles enthält außer der zu amplifizierenden DNA. Jeder PCR-Ansatz enthält 50 µl, davon entfallen 5 µl auf das DNA-Extrakt. Reaktionen mit 25 µl sind auch üblich, 100 µl Ansätze wegen der hohen Kosten nur, wenn große Mengen DNA benötigt werden.

Aufgabe: Der Mastermix enthält die folgenden Bestandteile.

	Ausgangs- konzentration	Konzentration/ Reaktion	Volumen/ Reaktion	Volumen/ 15 Reaktionen
PCR-Puffer	10 x	1 x		
MgCl ₂	50 mM	2 mM		
dNTP-Mix	2 mM	0,2 mM		
Primer 1	5 µM	0,4 µM		
Primer 2	5 µM	0,4 µM		
<i>taq</i> -Polymerase	5 U/ µl	2,5 U		
Wasser	-	-		

Berechnen Sie die Volumina der einzelnen Bestandteile für eine 50µl-Reaktion und einen Mastermix für 15 25µl-Reaktionen.

5. Berechnen Sie die Mengen für jeden PCR-Ansatz und für den Mastermix. Berechnen Sie 1 Ansatz für jede zu untersuchende Art, sowie 1,5 Reaktionen zusätzlich (1 Nullprobe und 0,5 Mengeneinheiten als „Pipettierreserve“)
6. Stellen Sie ein Eppendorf-Gefäß (1,5 ml) auf Eis und pipettieren Sie erst Wasser, dann die anderen Zutaten in der angegebenen Reihenfolge zusammen. Benutzen Sie für jedes Reagens eine frische Pipettenspitze. Mischen Sie jedes Reagens (außer der Polymerase, s. u.!) vor dem Pipettieren durch Ein- und Auspipettieren bei gleichzeitigem Rühren!
7. Pipettieren Sie als letztes die Polymerase. Rühren Sie die Polymerase-Lösung nur sehr vorsichtig, aber sorgfältig mit der Pipettenspitze um. Die Lösung ist sehr zähflüssig; sorgen Sie dafür, dass nichts an der Außenseite der Pipettenspitze hängenbleibt. Dieses Enzym ist extrem teuer (1 ml kostet ungefähr 1000 Euro).

8. Mischen Sie den Mastermix bis sie keine Schlieren mehr sehen. Pipettieren Sie 45 µl in jeden Ansatz.
9. Pipettieren Sie zuletzt 5 µl ihrer DNA-Extrakte (5 µl H₂O für die Nullprobe) in das jeweilige Gefäß, verschließen Sie das Gefäß.
10. Zentrifugieren Sie die PCR-Ansätze **kurz** ab (wenige Umdrehungen) und stellen Sie sie sofort wieder auf Eis.
11. Schalten Sie den Thermocycler ein und starten Sie das Programm CHL1.
12. Sobald die Temperatur von 94 °C erreicht ist, unterbrechen Sie das Programm, setzen zügig die Proben ein, verschließen den Deckel und setzen das Programm fort.
13. Setzen Sie für den morgigen Arbeitstag 2 Liter 5 × TBE-Puffer an. Dazu stellen Sie zunächst 200 ml 0,5 M EDTA (pH 8,0) her: 29,225 g EDTA, Aqua dest. ad 200 ml, langsam KOH-Plätzchen dazugeben und pH auf dem Magnetrührer einstellen.
14. Füllen Sie 108 g Tris, 55 g Borsäure, 40 ml EDTA mit Aqua dest. ad 2000 ml auf. Auf dem Magnetrührer mischen.
15. Arbeitsflächen aufräumen.

ACHTUNG!

Protokollieren Sie die PCR-Reaktion, indem Sie für jedes Gefäß das DNA-Extrakt, die PCR-Nummer und die Mengen aller Bestandteile des Ansatzes notieren (am besten in Tabellenform wie folgt).

Extr.-Nr.	PCR-Nr.	DNA	H ₂ O	Puffer	dNTPs	Primer1	Primer2	Enzym
127	1345	5 µl	x µl	x µl	x µl	x µl	x µl	x µl
128	1346							
...	...							

Dieses Protokoll ist sehr wichtig. Zu Beginn einer neuen Versuchsreihe (neue Organismen, neue Genabschnitte usw.) müssen PCR-Reaktionen fast immer optimiert werden. Das ist nur möglich, wenn man die Versuchsbedingungen jedes Laufs protokolliert.

NOCH EIN HINWEIS

Sobald die zu untersuchenden Organismen nur noch in Form von DNA-Extrakten im Labor herumstehen, lassen sie sich äußerlich nicht mehr unterscheiden. Ab diesem Zeitpunkt muss man sich peinlich genau vor Verwechslungen der Reaktionsgefäße oder –nummern hüten. Man sollte deshalb alle Eppendorf-Gefäße sauber und permanent beschriften, und unbedingt auch das Datum jedes Ansatzes notieren. Nur so hat man wenigstens eine geringe Chance, Verwechslungen auch im Nachhinein noch auf die Spur zu kommen.

UND EIN DRITTER HINWEIS

PCR-Produkte enthalten DNA in millionenfach höherer Konzentration als DNA-Extrakte. Selbst kleinste Verunreinigungen durch Tröpfchen an Handschuhen, Pipettenspitzen oder achtlos weggeworfenen PCR-Gefäßen können desaströse Folgen haben. Die DNA konkurriert mit der genomischen DNA um Polymerase, Primer und dNTPs. Dabei zieht die genomische DNA fast immer den Kürzeren. Im günstigsten Fall stört das die PCR lediglich. Im ungünstigsten Fall erhalten Sie wunderschöne PCR-Produkte, die sich nach Sequenzieren (für 10 €/je DNA-Strang) immer wieder als ein und dieselbe Art erweisen. Reinigen des gesamten Labors (einschließlich Mobiliar, Geräten, Glasflaschen usw.) und Wegwerfen kontaminierten Verbrauchsmaterials ist dann oft die einzige Lösung. Arbeiten Sie in der „Post-PCR Phase“ mit pingeliger Genauigkeit, sonst zerrütten Sie das Verhältnis zu Ihren Labor-kollegen.

Nachweis von PCR-Produkten

Die spannende Frage lautet nun: Hat die PCR funktioniert oder nicht? Von außen sehen Ihre PCR-Röhrchen unverändert aus. 4 µl Ihres PCR-Produktes müssen Sie deshalb opfern, um es mittels Elektrophorese auf einem Agarose-Gel aufzutrennen. Das Gel wird mit Ethidiumbromid (EtBr) behandelt, das sich mit DNA zu einem fluoreszierenden Komplex verbindet. Bei Betrachten des Gels auf einem UV-Transilluminator finden Sie so heraus, ob (1) überhaupt PCR-Produkt entstanden ist, (2) nur ein spezifisches oder mehrere Produkte unterschiedlicher Länge entstanden sind, und (3) wie lang das amplifizierte DNA-Stück in etwa ist, d. h. ob das richtige Stück amplifiziert wurde. Auch bei diesem Arbeitsgang sind Handschuhe erforderlich.

1. Wiegen Sie 2 g Agarose in ein Becherglas ein. Vorsicht, Agarose und Kokain unterscheiden sich in Aussehen und Preis nicht sonderlich (vermutlich aber in der Wirkung).
2. Fügen Sie 200 ml 1 × TBE und ein Rührfischchen hinzu und setzen einen Deckel lose auf.
3. Erhitzen Sie die Mischung 2 min auf höchster Stufe in der Mikrowelle.
4. Rühren Sie einmal mit dem Magnetrührer durch.
5. Erhitzen Sie jetzt in kleinen Schritten (20-30 sek mit zwischenzeitlichem Rühren) weiter in der Mikrowelle bis die Lösung völlig klar ist. Agarose kocht in Sekundenschnelle über (eine Riesenschweinerei!). Beaufsichtigen Sie das Glas gut und stoppen Sie die Mikrowelle, sobald die ersten Blasen erscheinen. Auch auf dem Magnetrührer beginnt heiße Agarose-Lösung sehr leicht zu schäumen.
6. Wenn die Agarose-Lösung klar ist, lassen Sie sie auf dem Magnetrührer bis etwa 60° C abkühlen (mit Handschuhen knapp unterhalb der Schmerzgrenze).
7. Setzen Sie eine Gelwanne um 90° gedreht in die Elektrophoresekammer, so dass sie dicht mit den Wänden abschließt. Hängen Sie einen oder zwei Kämme ein.
8. Gießen Sie ca. 50 ml Agarose-Lösung (1 %) in die Gelwanne, ohne Blasen zu erzeugen. Etwaige Blasen mit einer Pipettenspitze an den Rand des Gels manövrieren. Warten Sie 20 Minuten, bevor Sie mit den folgenden Arbeitsschritten beginnen.

Es folgt die Vorbesprechung.

9. Wenn das Gel milchig aussieht, ziehen Sie die Gelwanne aus dem Tank und setzen sie richtig ein. Füllen Sie Puffer 1 × TBE in den Elektrophoresetank, bis das Gel bedeckt ist. Achten Sie auf die richtige Ausrichtung des Gels, die DNA wandert zur Anode (dem roten Anschluss). Ziehen Sie **ganz vorsichtig** die Kämme aus dem Gel (auf einer Seite beginnen und schräg ziehen, um das Gel nicht zu beschädigen).
10. Zentrifugieren Sie die PCR-Produkte kurz ab.
11. Schneiden Sie einen Streifen Parafilm ab und legen ihn flach auf den Tisch. Notieren Sie auf einen Zettel, in welcher Reihenfolge Sie die PCR-Produkte auf das Gel laden. Die folgenden Schritte erfordern zügiges Vorgehen!
12. Pipettieren Sie nebeneinander für jede PCR-Reaktion (inklusive der Nullprobe) 1 µl Ladepuffer auf den Parafilm-Streifen. Genügend Abstand halten. Pipettenspitze muss nicht gewechselt werden.
13. Mischen Sie die erste PCR-Reaktion mit einer neuen Pipettenspitze gut um (DNA setzt sich unten ab).
14. Pipettieren Sie 4 µl PCR-Produkt auf den ersten Tropfen Ladepuffer. Verschließen Sie das PCR-Gefäß wieder. **Pipettenspitze wechseln!**

15. Wiederholen Sie Schritt 13-14 für jedes PCR-Produkt und für die Nullprobe. Arbeiten Sie so zügig, dass die Tropfen nicht eindampfen.
16. Nehmen Sie die Tropfen (5 µl) der Reihe nach mit einer Pipette auf und pipettieren Sie sie vorsichtig in die Taschen des Gels. Versuchen Sie das Gel dabei nicht zu berühren und vermeiden Sie Wirbelstürme im Elektrophoresetank. Lassen Sie in jeder Reihe die letzte Tasche des Gels frei.
17. Pipettieren Sie 5 µl Größenmarker („100 bp ladder“) in die letzte Tasche des Gels.
18. Setzen Sie den Deckel auf, kontrollieren Sie noch einmal die richtige Ausrichtung des Gels und starten Sie die Elektrophorese bei 80 V.

Nach etwa 45-60 min, wenn der blaue Marker 5-7 cm weit gewandert ist, kann das Gel unter dem Transilluminator betrachtet werden.

19. Ziehen Sie blaue Nitrilhandschuhe an. Heben Sie die Gelwanne vorsichtig aus dem Elektrophoresetank und trocknen sie mit Papier ab.
20. Überführen Sie das Gel **ohne es fallen zu lassen** in die Färbewanne unter dem Abzug. Die Färbelösung enthält Ethidiumbromid (EtBr), das im Tierversuch karzinogen ist. Latex ist durchlässig für EtBr, deshalb die Nitrilhandschuhe. Entsorgen Sie alle Materialien, die mit EtBr in Berührung gekommen sind im Spezial-Abfallbehälter unter dem Abzug.
21. Lassen Sie die Färbelösung ca. 20 min. einwirken.
22. Überführen Sie das Gel erst auf das durchsichtige Tablett und dann **ohne zu tropfen** auf den Transilluminator. Schliessen Sie die Tür und schalten Sie das Gerät ein. Wechseln Sie danach die Handschuhe (Sondermüll), bevor Sie andere Gegenstände berühren.
23. Schalten Sie den Computer der Geldokumentationsanlage ein.
24. Starten Sie das Programm „EasyWin 32“ und klicken Sie im Drop-down-Menü **>Datei>Kamera aktivieren** an.
25. Speichern Sie das Bild (**>Datei>Bild speichern**), drucken Sie ein Foto des Gels aus (**>Datei>Bild drucken**) und kleben es zur Dokumentation unter ihr PCR-Protokoll.
26. Entsorgen Sie das Gel als Sondermüll. Reinigen Sie die Oberfläche des Transilluminators mit destilliertem Wasser und Papiertüchern.
27. Entsorgen Sie ihre Handschuhe im Sondermüll.

Die amplifizierte DNA einer erfolgreichen Reaktion liegt als mehr oder weniger breite Bande auf dem Gel vor. Diese Banden sollten deutlich sichtbar und sauber begrenzt sein, und die Fragmente sollten in der Länge dem erwarteten Produkt entsprechen. Die Gelspur der Nullprobe sollte schwarz sein. Sollten Sie in der Nullprobe eine Bande finden, die einer der Banden in Ihren PCR-Ansätzen entspricht, sind ihre Reaktionen sehr wahrscheinlich kontaminiert. Meist sind verunreinigte Reagenzien die Ursache solcher Kontaminationen. Diese Verunreinigungen entstehen z. B., wenn Pipettenspitzen nicht gewechselt wurden und genomische DNA übertragen wurde. In diesem Fall muss die PCR wiederholt werden. Im schlimmsten Fall hat man es mit Laborkontaminationen zu tun (s.o.).

Sie können jetzt entscheiden, welche PCR-Produkte Sie nach dem Standardprotokoll (nächster Schritt) weiterverarbeiten wollen und welche Reaktionen wiederholt werden müssen. Oft treten multiple Banden auf. Solche Reaktionen können nicht einfach gereinigt und sequenziert werden. Man muss die Reaktion aber auch nicht unbedingt wiederholen. Wenn die Banden gut getrennt auf dem Gel liegen und die erwünschte Bande deutlich und sauber begrenzt ist, kann man das gesamte PCR-Produkt auf ein möglichst großes und dickes Agarose-Gel mit größeren Taschen aufladen und auftrennen. Die gewünschte Bande wird unter UV-Licht geortet, mit einer Skalpellklinge ausgeschnitten und mit einem speziellen Geextraktionskit gereinigt. Das Erkennen und Ausschneiden der richtigen Bande soll-

te nicht auf dem Transilluminator erfolgen. Die hohe Intensität des UV-Lichtes zerstört in kürzester Zeit die DNA. Stattdessen eignen sich gewöhnliche UV-Lampen, wie sie z. B. für die Dünnschichtchromatographie verwendet werden. Zusätzlich zu Handschuhen muss ein Augenschutz getragen werden.

Datengewinnung: DNA-Sequenzierung, 3. Arbeitstag

Reinigung der PCR-Produkte

Rückstände der PCR-Reaktion, besonders Polymerase und nicht verbrauchte Primer müssen entfernt werden, bevor die eigentliche Sequenzierreaktion angesetzt werden kann. Dieser Reinigungsschritt wird ähnlich wie die DNA-Extraktion am besten mit einem fertigen Kit durchgeführt. Wir verwenden das NucleoSpin Reinigungskit.

1. Fügen Sie 28 ml Ethanol zu jeder Flasche Puffer NT3.
2. Beschriften Sie für jeden gelungenen PCR-Ansatz ein Eppendorf-Gefäß (1,5 ml), eine Extraktionssäule in einem Eppendorf-Gefäß (2 ml), und ein weiteres Eppendorf-Gefäß (1,5 ml).
3. Berechnen Sie für jede zu reinigende PCR 30 µl Puffer NE, pipettieren Sie diese Menge in ein 1,5 ml Eppendorf-Gefäß und erhitzen Sie sie im Heizblock auf 45° C.
4. Pipettieren Sie 184 µl Puffer NT2 in die erste Serie Eppendorf-Gefäße. Pipettenspitze braucht nicht gewechselt zu werden.
5. Pipettieren Sie nun das PCR-Produkt der ersten zu untersuchenden Art vollständig in das erste Eppendorf-Gefäß und mischen Sie Puffer und Produkte mit der Pipette. **Pipettenspitzen wechseln und Vorsicht vor Verwechslungen!**
6. Wiederholen Sie Schritt 4 für die PCR-Produkte aller Arten.
7. Pipettieren Sie die Mischungen in die jeweiligen Extraktionssäulen. Dabei für jede Probe **Pipettenspitze wechseln.**
8. Zentrifugieren Sie 1 min bei 13000 rpm. Verwerfen Sie den Durchlauf.
9. Pipettieren Sie 600 µl Puffer NT3 auf die Trennsäule. (Wenn Sie die Trennsäule nicht berühren, können Sie bei den Schritten 8-10 Pipettenspitzen mehrfach verwenden. Nach jeder versehentlichen Berührung Pipettenspitzen wechseln.)
10. Zentrifugieren Sie 1 min bei 13000 rpm. Verwerfen Sie den Durchlauf.
11. Pipettieren Sie 200 µl Puffer NT3 auf die Trennsäule.
12. Zentrifugieren Sie 2 min bei 13000 rpm. Verwerfen Sie den Durchlauf mitsamt dem 2-ml-Gefäß.
13. Stellen Sie die Trennsäule in das zweite Eppendorf-Gefäß (1,5 ml) und lassen Sie sie 5 min mit geöffnetem Deckel stehen, um die Membran vollständig zu trocknen.
14. Pipettieren Sie 25 µl Puffer NE (vom Heizblock) genau ins Zentrum der Membran, ohne diese zu berühren. **Pipettenspitzen wechseln!** Schliessen Sie den Deckel der Trennsäule und lassen Sie den Ansatz 5 min bei Raumtemperatur stehen.
15. Zentrifugieren Sie 1 min bei 13000 rpm. Verwerfen Sie die Trennsäule, behalten Sie das Eluat (Ihr gereinigtes PCR-Produkt). Beschriften Sie das Gefäß deutlich lesbar.

DNA-Quantifizierung

Nach dem vorigen Arbeitsschritt haben Sie nun ein gereinigtes PCR-Produkt, das darauf wartet sequenziert zu werden. Die Qualität der Sequenzen kann aber sehr empfindlich auf schwankende DNA-Mengen reagieren. Unterhalb eines gewissen Schwellenwertes versagt die Sequenzreaktion, oberhalb werden die Sequenzen oft unsauber. Solche schlechten Sequenzchromatogramme muss man in nervenaufreibender Arbeit editieren. Selbst dann bleiben oft viele Positionen der Sequenz unsicher, was die Datenanalyse erschwert. Um sich diesen Ärger zu ersparen, ist es besser, die Menge der DNA vorher zu bestimmen und 200 ng DNA pro Reaktion (10 μ l) zu verwenden.

Die Mengenbestimmung erfolgt auf einem 1 %igen Agarose-Gel mit EtBr. Arbeiten Sie nach dem Protokoll für Schritt 4 und verwenden Sie den Rest der vorher angesetzten Agarose-Lösung und 4 μ l Ihres gereinigten Eluats. Füllen Sie die ersten beiden Taschen des Gels mit je 1 μ l und 2 μ l λ -DNA-Lösung (entspricht 25 und 50 ng DNA).

1. Die ersten Schritte sind dieselben wie bei der Überprüfung der PCR-Produkte. Verwenden Sie nur 1 μ l des gereinigten PCR-Produktes zur Quantifizierung. Es ist nicht notwendig, ein Bild des Gels auszudrucken.
2. Wenn Sie ein Bild des Gels auf dem Bildschirm haben, klicken Sie **>Auswertung>Molekular-MW und Volumen>Einzelne Banden** an.
3. Es erscheint ein Fadenkreuz. Setzen Sie das Fadenkreuz auf die linke obere Ecke der ersten Bande. Mit einem linken Mausklick verschwindet das Fadenkreuz, ziehen Sie das rote Rechteck auf, so dass die DNA-Bande vollständig umfasst wird. Mit einem weiteren linken Mausklick wird das Rechteck fixiert und das Fadenkreuz erscheint wieder. Fahren Sie fort, bis alle Banden markiert sind, dann löschen sie das Fadenkreuz mit einem rechten Mausklick.
4. Zum Quantifizieren der einzelnen DNA-Banden verwenden Sie die beiden Banden mit bekannter Konzentration. Ein Doppelklick in die obere Hälfte der 1. Bande öffnet eine Dialogbox mit einer Zeile verschiedener Zahlenwerte.
5. Mit einem rechten Mausklick auf die Zahlenwerte öffnet sich ein Menü, wählen Sie **>Volumen zuweisen**.
6. In der sich öffnenden Dialogbox wählen Sie **>Verwenden** und ersetzen den angezeigten Wert durch „25“.
7. Verfahren Sie genauso mit der 2. Bande. Hier tragen Sie „50“ ein.
8. Bei Doppelklick auf die übrigen Banden können Sie nun die ermittelte DNA-Menge ablesen. Tragen Sie die Werte in Ihr Protokoll ein und berechnen Sie, wieviel DNA in 1 μ l PCR-Produkt enthalten ist.
9. Wieviel PCR-Produkt müssen Sie in der Sequenzreaktion einsetzen, um auf 200-500 ng DNA zu kommen?

Datengewinnung: DNA-Sequenzierung, 4. Arbeitstag

Sequenzierreaktion

In der ersten PCR-Reaktion ging es darum, einen spezifischen Abschnitt der genomischen DNA exponentiell zu vermehren, um ihn später sequenzieren zu können. In der Sequenzierreaktion wird diese hoch konzentrierte Templat-DNA nicht mehr exponentiell vermehrt. Man fügt deshalb nur einen Primer hinzu, so dass die Polymerase nur einen der beiden DNA-Stränge synthetisieren kann. Die Sequenzierreaktion enthält neben gewöhnlichen dNTPs einen kleinen Anteil Didesoxiribonukleotide (ddNTPs). Gewöhnlich verwendet man fertige Mischungen (sog. Terminator-Kits mit meist blumigen Namen). Wir verwenden im Kurs das BigDye®-Kit, weil es von unserem Sequenzierlabor verwendet wird. An jedes der vier unterschiedlichen ddNTPs ist ein anderer Fluoreszenzfarbstoff gebunden. Wird ein solches dd-Nukleotid zufällig eingebaut, kann der DNA-Strang nicht weiter verlängert werden und die Reaktion bricht ab. Am Ende vieler Reaktionszyklen erhält man so ein Gemisch von fluoreszenzmarkierten DNA-Strängen aller unterschiedlichen Längen, bei denen das letzte Nukleotid sich durch seine spezifische Fluoreszenz verrät.

Diese vielen hundert verschiedenen DNA-Moleküle werden im letzten Schritt auf einem Polyacrylamid-Gel elektrophoretisch nach ihrer Größe aufgetrennt. In automatischen Sequencern, die hierfür verwendet werden, befinden sich am Ende des Gels meist vier Infrarot-Laser, die die vorbeiwandernden DNA-Moleküle zur Fluoreszenz anregen. Diese Fluoreszenz wird von einem Detektor aufgezeichnet, dessen Signale in Form eines Sequenzchromatogramms als Computerdatei gespeichert werden.

Frage: Terminator-Kits sind sehr teuer. Eine alternative, wesentlich billigere Methode besteht darin, nicht die einzelnen ddNTPs zu markieren, sondern einen Fluoreszenzfarbstoff an die eingesetzten primer zu binden. Wieviele Reaktionen muss man bei dieser Methode je DNA-Strang ansetzen? Lässt sich die Anzahl der Reaktionen verringern, wenn man beide Stränge sequenziert?

Sequenz-Chromatogramme sind nur über eine begrenzte Länge hinweg lesbar. Die Qualität oder Leselänge variiert sehr stark, überschreitet aber selbst im Idealfall selten 800 bp. In vielen Fällen sind die sequenzierten Genabschnitte länger (in unserem Fall ca. 1050-1200 bp), so dass man schon aus diesem Grund beide Stränge sequenzieren muss. Um sicher zu gehen, kann man zusätzlich interne Primer einsetzen. Wir verwenden zusätzlich zu den Primern C und F die internen Primer D und E.

Achtung: Jeder DNA-Strang wird separat sequenziert. Sie brauchen vier Reaktionsansätze mit je einem Primer für jedes PCR-Produkt.

1. Schreiben Sie eine Tabelle. Tragen Sie die ermittelte Menge jedes PCR-Produkts (200 ng DNA) ein und füllen Sie jeweils mit Wasser auf 10 µl auf.

PCR-Produkt	Sequenz-Nr.	DNA	H ₂ O	Primer 1	BigDye	Total
1235	145	xx µl	xx µl	1 µl	4 µl	10 µl
1237	146	xx µl	xx µl	1 µl	4 µl	10 µl
...						
PCR-Produkt	Sequenz-Nr.	DNA	H ₂ O	Primer 2	BigDye	Total
1235	155	xx µl	xx µl	1 µl	4 µl	10 µl
1237	156	xx µl	xx µl	1 µl	4 µl	10 µl
...						

2. Die folgenden Arbeitsschritte erfolgen auf Eis. Auch Terminator Kits leiden unter häufigem Einfrieren und Auftauen. Sollten Sie einmal größere Mengen bekommen, aliquotieren Sie diese und frieren die Aliquots separat ein.
3. Beschriften Sie für jeden Ansatz ein PCR-Gefäß (0,2 ml) anhand der Tabelle.
4. Pipettieren Sie nacheinander Big Dye (auf vollständiges Auftauen achten und gut mischen!), Wasser und Primer in die PCR-Gefäße. Die Pipettenspitze nur beim Wechsel zwischen den Reagentien erneuern.
5. Pipettieren Sie zuletzt die jeweilige Menge DNA hinzu und verschließen Sie die Gefäße. **Pipettenspitze nach jeder Probe wechseln!**
6. Die Big Dye Lösung ist lichtempfindlich. Frieren Sie sie schnell wieder ein.
7. Schalten Sie den Thermocycler ein und starten Sie das Programm SQ1.
8. Sobald die Temperatur von 94 °C erreicht ist, unterbrechen Sie das Programm, setzen zügig die Proben ein, verschließen den Deckel und setzen das Programm fort.

Reinigen und Trocknen der Sequenzierreaktion

Im letzten Arbeitsgang muss die Sequenzierreaktion gereinigt und das Produkt getrocknet werden. Das getrocknete Produkt wird von uns an ein kommerzielles Sequenzierlabor weitergegeben. Die Ethanol-fällung erfolgt bei -20 °C. Arbeiten Sie auf Eis.

1. Zentrifugieren Sie die Sequenzierreaktion kurz in den Boden des PCR-Gefäßes.
2. Bereiten Sie für jede Reaktion ein Eppendorf-Gefäß (1,5 ml) vor und beschriften Sie es.
3. Pipettieren Sie erst 1 µl 3 M Natriumacetat (pH 5,2) in jedes Gefäß, pipettieren dann die jeweilige Reaktion vollständig hinzu und füllen mit 10 µl Wasser auf.
4. Geben Sie je 50 µl eiskaltes Ethanol (100 %) hinzu, und vortexen die Mischung.
5. Stellen Sie die Mischung für 20 min in den Gefrierschrank.
6. Zentrifugieren Sie die Ansätze 20 min bei 13000 rpm.
7. Giessen Sie den Überstand ab.
8. Geben Sie 400 µl Ethanol (70 %) hinzu und zentrifugieren Sie 3 min bei 13000 rpm.
9. Giessen Sie den Überstand ab.
10. Lassen Sie die PCR-Gefäße 20-25 min offen an der Luft trocknen.
11. Drücken Sie die Daumen, dass alles geklappt hat und wir in ein paar Tagen saubere Sequenzen geliefert bekommen.

Datenanalyse: Erstellen des Datensatzes

Das Erzeugen von Daten ist ein wichtiger, vorläufiger Schritt jeder wissenschaftlichen Arbeit. Die eigentliche Wissenschaft beginnt aber erst mit der Auswertung der Daten. Seit mehr als 30 Jahren haben Systematiker neue Labormethoden aus der Genetik und Biochemie übernommen, um bis dahin unbekannte molekulare Merkmale für ihre Arbeit zu nutzen. Seither ist so etwas wie eine industrielle Produktion neuer Auswertungsmethoden entstanden, mit denen sich immer mehr Informationen aus diesen Daten extrahieren lassen. Mittlerweile nutzen nun Genetiker diese neuen Auswertungsmethoden für ihre Zwecke. Statistische Verfahren, die ursprünglich für systematische Fragestellungen entwickelt wurden, sind heutzutage z. B. aus der Aids-Forschung oder Genomkartierungsprojekten nicht mehr wegzudenken.

Als erstes Produkt einer phylogenetischen Analyse will man fast immer einen oder mehrere Stammbäume erstellen. Wenn man nur die Verwandtschaft der untersuchten Organismen klären will, reicht das aus. Mit Hilfe solcher Stammbäume kann man aber auch weitergehende Fragen, z. B. nach der Evolution bestimmter Merkmale oder der Biogeografie von Arten klären. Die umfangreichsten Erkenntnisse z. B. zur Evolution und Ausbreitung des Menschen basieren mittlerweile nicht mehr auf Fossilfunden, sondern auf der Auswertung genetischer Daten.

Wie im ersten Teil werden wir eine phylogenetische Analyse in Einzelschritten durchlaufen. Die Analyse erfolgt fast ausschließlich am Computer und erfordert den Einsatz vieler Programme. Manche lassen sich mit Hilfe der Maus oder von drop-down Menüs bedienen, andere erfordern die manuelle Eingabe von Befehlen in Befehlszeilen. Die Anleitung hält sich an folgende Konventionen: **Fettschrift** bezeichnet Befehle aus drop-down-Menüs oder Dialogboxen. Schrift in *Courier* bedeutet Eingabe in Befehlszeilen.

„>**Datei**>**Öffnen**“ bedeutet „Klicken Sie im Menü **Datei** auf den Befehl **Öffnen**“.

„hsearch addseq=random nreps=10000“ bedeutet „Tippen Sie diese Befehle über die Tastatur in eine Befehlszeile ein.“

Editieren von Sequenzen

Bisher haben wir lediglich überprüfen können, ob überhaupt PCR-Produkt der erwarteten Länge gebildet wurde. Wir haben angenommen, dass es sich um die gewünschte DNA handelt, und haben gehofft, dass wir lesbare Sequenzen erhalten. Die vom automatischen Sequenzierer vom Fluoreszenzdetektor aufgezeichnete Abfolge der Genfragmente erscheint in der Computerdatei („Trace File“) als eine Abfolge von „Peaks“. Abb.1 zeigt einen Ausschnitt aus einem sehr gut lesbaren Trace File.

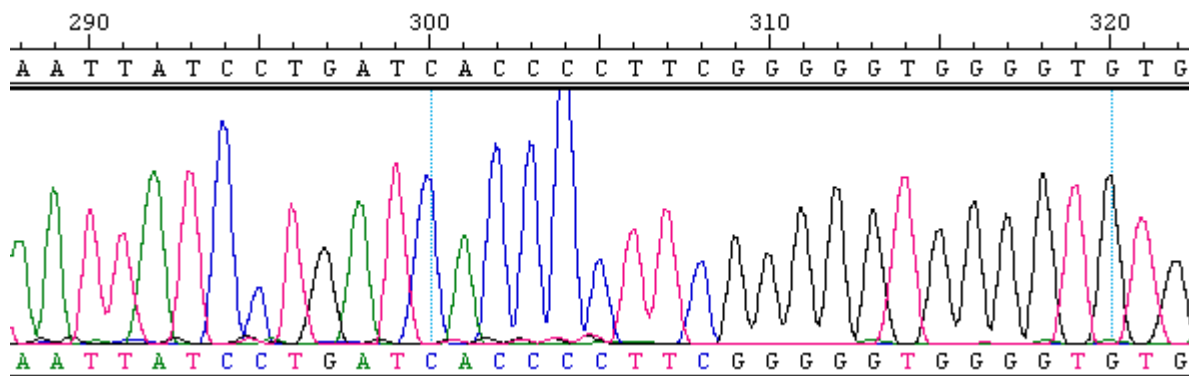


Abb. 1: Perfektes Sequenzchromatogramm: deutliche Peaks, fast kein Rauschen.

Bei nahezu jedem bisher durchgeführten Arbeitsschritt kann es aber zu Fehlern kommen, die zu unleserlichen Chromatogrammen führen (Abb. 2). Die Qualität der Chromatogramme liegt oft zwischen diesen Extremen. Einzelne Positionen sind nicht deutlich zu lesen, die meisten sind gut erkennbar. Sequenziert man beide Stränge eines Genfragments, kann man unsichere Positionen des einen mit Hilfe des anderen Strangs ergänzen. Diesen Vorgang bezeichnet man als „Editieren“ der Sequenzen. Bei langen PCR-Produkten verwendet man wegen der begrenzten Leselänge zusätzlich interne Primer zum Sequenzieren. Fehler in den Trace files treten auch dadurch auf, dass die Wellenlängen der Infrarotlaser im Sequencer relativ nahe beieinander liegen. Dadurch kommt es, besonders bei schwachem Signal am Anfang oder Ende einer Sequenz zu Fehlablesungen des Detektors.

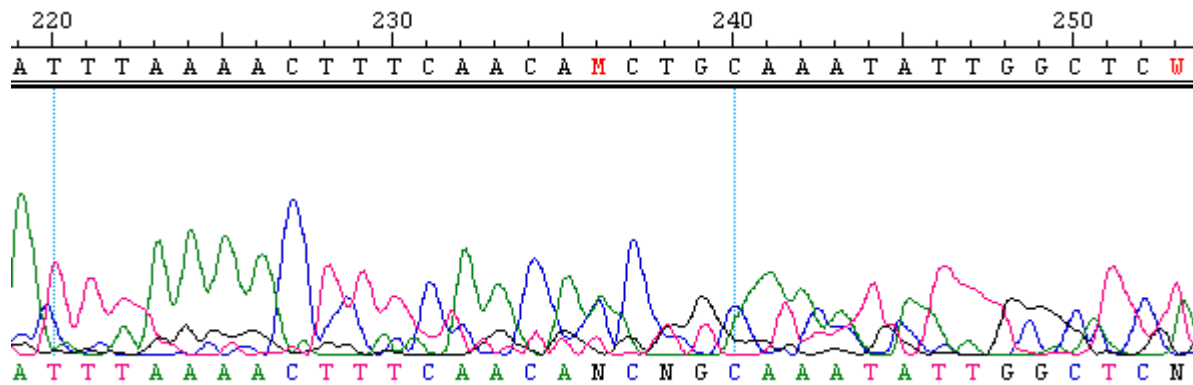


Abb. 2: Nahezu unleserliches Sequenzchromatogramm: nur wenige deutliche Peaks, viel Rauschen.

Frage: Die DNA-Polymerase arbeitet nicht fehlerfrei. Mit einer geringen Wahrscheinlichkeit werden auch falsche Nukleotide in die synthetisierten DNA-Stränge eingebaut. Warum führt das normalerweise nicht zu Sequenzierfehlern?

1. Öffnen Sie das Programm „Seqman II“. Wir werden zunächst die verschiedenen trace files eines PCR-Produktes zu einem sog. „Contig“ zusammenführen.
2. **>File>New.** Es öffnen sich 2 Fenster namens „Unassembled Sequences“ und „Untitled“.
3. Unter “Unassembled Sequences” **>Add Sequences ...** Es öffnet sich eine Dialogbox, in der sie nach den trace files suchen können.
4. Gehen Sie in den Ordner, der Ihre gewünschten trace files enthält.
5. Klicken Sie die trace files an, die zu der ersten zu untersuchenden Art gehören. Klicken Sie auf **>Add**, um die files in das rechte Teilfenster zu kopieren. Falsche files können sie durch **>Remove** wieder entfernen.
6. Wenn Sie die gewünschten files kopiert haben, klicken Sie **>Done**. Die Dialogbox schließt sich wieder. Eine Liste der gewählten files erscheint im linken Fenster.
7. Klicken Sie **>Assemble**. Der Sequenzeditor fügt die Sequenzen zu einem Contig im rechten Fenster zusammen. Erscheinen mehrere Contigs, hat die Sequenzierung in einem oder mehreren Fällen nicht geklappt oder Sie haben versehentlich sehr unterschiedliche Sequenzen verschiedener Arten zusammengefügt.
8. Doppelklicken Sie auf „Contig 1“ und vergrößern Sie das erscheinende Fenster auf Bildschirmgröße. Sie sehen eine Liste der einzelnen Sequenzen im Contig und den Beginn der Sequenz.
9. Um die Chromatogramme zu sehen, klicken Sie auf die grauen Pfeile links neben den Namen der Sequenzen.

10. Am linken Bildschirmrand sehen sie verschiedenen blaue Felder. Mit der Lupe können Sie die Chromatogramme horizontal, mit dem Regler der Skala vertikal vergrößern. Sie sehen unterhalb jedes Chromatogramms die Sequenz als Buchstabenfolge. Über dem obersten Chromatogramm erscheint die vom Editor ermittelte Konsensussequenz.
11. Am unteren Bildschirmrand finden Sie einen Scrollbalken. Scrollen Sie durch das Chromatogramm und suchen Sie in der Konsensussequenz nach unsicheren Positionen. Diese Positionen sind rot markiert. Treffen Sie anhand der vorhandenen Daten eine Entscheidung, welches Nukleotid hier stehen muss. Mit der Maus können Sie in die Chromatogramme klicken und manuell einzelne Basen verändern. Um zu kennzeichnen, dass es sich um manuelle Veränderungen handelt, schreiben Sie mit Kleinbuchstaben.
12. Sichern Sie das Contig durch **>File>Save as**. Wählen Sie den Artnamen als Dateinamen in Form einer Abkürzung von max. 10 Zeichen (z. B. dem ersten Buchstaben der Gattung und max. neun Buchstaben des Artnamens). Verwenden Sie keine Leerzeichen, Punkte oder Striche, sondern nur Buchstaben.
13. Sichern Sie die Konsensussequenz durch **>Contig>Save Consensus>Single File**. Wählen Sie als Dateinamen wieder den Artnamen und als Dateityp „FastA (*.fas)“.
14. Schliessen Sie die Datei mit **>File>Close** und beginnen Sie für die nächste Art bei Schritt 1.

Wenn Sie auf diese Weise alle Sequenzen editiert haben, schließen Sie Seqman. Als nächstes muss aus den Einzelsequenzen ein alignment erstellt werden. Wir müssen uns aber vorher noch vergewissern, ob die Sequenzen wirklich zu den von uns zu untersuchenden Arten gehören. Hierfür wird ein „BLAST-search“ in Genbank durchgeführt.

BLAST-search in Genbank

Durch unzählige Sequenzierungsprojekte, am spektakulärsten das „Human Genome Project“ und „Assembling the Tree of Life“ (http://www.nsf.gov/bio/pubs/awards/ato1_02.htm), sind seit den 70er Jahren ungeheure Mengen an DNA- und Proteinsequenzen erzeugt worden. Um diese Datenflut zu speichern und der Forschung zugänglich zu machen, unterhalten das National Institute of Health, das European Molecular Biology Laboratory und die DNA Databank of Japan untereinander vernetzte Sequenz-Datenbanken. Alle renommierten wissenschaftlichen Zeitschriften verlangen heute, dass Sequenz-Daten von publizierten Artikeln in einer dieser Datenbanken hinterlegt werden. Diese Daten können dann von jedem anderen für wissenschaftliche Arbeiten verwendet werden.

Systematiker nutzen diese Datenbank auf zweierlei Weise. Sie verwenden bereits publizierte Sequenzdaten in eigenen Datensätzen. Das spart Zeit und Kosten. Mit Hilfe eines Suchalgorithmus (namens BLAST) kann man in Genbank aber auch nach Sequenzen suchen, die einer selbst erstellten Sequenz ähneln. Auf diese Weise findet man heraus, ob die erstellten Sequenzen wirklich vom untersuchten Organismus stammen (und nicht von kontaminierenden Bakterien, Pilzen oder einem selbst). Wir testen zunächst eine der im Praktikum erstellten Sequenzen.

1. Öffnen Sie die Datei mit der Konsensussequenz einer Art in Word.
2. Öffnen Sie den Internet Explorer und gehen Sie auf folgende Seite:
<http://www.ncbi.nlm.nih.gov/BLAST/>
3. Wählen Sie [Nucleotide-nucleotide BLAST \(blastn\)](#)
4. Kopieren Sie die vollständige Konsensussequenz in das „Search“-Fenster, entfernen Sie eventuelle Zeilenumbrüche und klicken Sie auf „Blast!“.
5. Genbank teilt Ihnen mit, wie lange die Suche ungefähr dauern wird. Die angegebene „Request ID“ kopieren Sie mit einer kleinen Erklärung für späteren Gebrauch in einen

Textfile. Es kann vorkommen, dass die Datenbank (oder das Netzwerk) versagt. Mit dieser Nummer können sie ihre Ergebnisse auch später noch abfragen.

6. Klicken Sie nach Ablauf der Zeit auf „Format!“ In einem separaten Fenster erhalten Sie das Resultat Ihrer Anfrage.
7. Neben einigen Informationen zu BLAST und der von Ihnen eingegebenen Sequenz erhalten Sie als Resultat eine graphische Darstellung mit ihrer „query“ Sequenz und ähnlichen Sequenzen aus Genbank. Die Farbe der Balken gibt den Grad der Ähnlichkeit an. Weiter unten finden Sie eine Liste mit Genbank-Nummern, Arten und Kurzbeschreibungen der Gene. Die Angaben sind nach Ähnlichkeit sortiert. Noch weiter unten sehen Sie paarweise „alignments“ ihrer „query“ Sequenz mit den ähnlichsten Sequenzen.
8. Eine gute Übereinstimmung ihrer Sequenz mit den ähnlichsten Sequenzen in Genbank bedeutet in der Regel eine enge Verwandtschaft. Wenn also die ähnlichsten Sequenzen in Genbank von Menschen und Mäusen stammen, haben Sie mit Sicherheit keine Bromeliaceen oder Flechten sequenziert.
9. Die Familienzugehörigkeit der ähnlichsten Sequenzen können Sie herausfinden, indem Sie auf die Genbank-accession-number links von den Artnamen klicken. Unter „Organism“ ist die Klassifizierung angegeben.
10. Wenn Sie feststellen, dass die ähnlichsten Organismen in Genbank zu den Flechten gehören, können Sie den BLAST-search abschließen. Wenn nicht, testen Sie andere Sequenzen, bis sie sicher sind, eine Flechten-Sequenz erwischt zu haben.
11. Die nicht zu den Flechten gehörenden Sequenzen können Sie bei allen folgenden Analysen nicht verwenden.

Im nächsten Schritt werden die Einzelsequenzen zu einem Gesamtdatensatz zusammengefügt, aus dem dann ein sogenanntes „Alignment“ erstellt wird.

Alignen der Sequenzen

Bei der phylogenetischen Auswertung von Sequenzdaten spielt im Grunde die Ähnlichkeit der Sequenzen die Hauptrolle. Diese Ähnlichkeit lässt sich auf ganz verschiedenen Weise messen. Man kann zum Beispiel zählen, an wievielen Stellen in jeweils zwei Sequenzen unterschiedliche Nukleotide eingebaut sind. Natürlich kann man die Sequenzen dafür nicht willkürlich nebeneinander legen und die Unterschiede notieren, sondern muss „gleiche“, sogenannte homologe Positionen der beiden Sequenzen miteinander vergleichen. Mit Hilfe des Programms „Bioedit“

(<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>) fassen wir die Sequenzen zunächst in einem Rohdatensatz zusammen. Bioedit bietet eine Vielzahl weiterer Funktionen, die wir nicht alle besprechen können.

1. Starten Sie „Bioedit“.
2. **>File>New Alignment**
3. **>File>Import**
4. Suchen Sie den Ordner mit Ihren Sequenzfiles. Markieren Sie alle von Ihnen erstellten Fasta-Dateien (*.fas), indem sie gleichzeitig die Taste <Strg> gedrückt halten, und klicken Sie zum Abschluss **>Öffnen**.
5. Speichern Sie den Datensatz durch **>File>Save As** unter einem selbst gewählten Namen.

6. Am unteren Rand des Fensters finden Sie wieder einen Scrollbalken. Scrollen Sie an das Ende des Datensatzes und überzeugen Sie sich, dass ihre Sequenzen verschieden lang sind.

Beim Scrollen durch den Datensatz werden Sie feststellen, dass Sie im Gewimmel der Sequenzen anfangs noch eine Struktur erkennen. Weiter hinten verschwindet diese Struktur zunehmend. In fast allen Genen kommt es nämlich gelegentlich zu Insertionen oder Deletionen („Indels“) von einzelnen Nukleotiden oder längeren Genabschnitten. Durch die zunehmende Zahl von Indels werden die homologen Positionen der Sequenzen immer weiter voneinander weggerückt, je weiter man durch den Datensatz scrollt. Durch Einsetzen von „Leerstellen“, sog. „Gaps“ müssen die Sequenzen wieder auf gleiche Länge gebracht und homologe Nukleotide verschiedener Sequenzen an die gleiche Position gerückt werden. Diesen Vorgang nennt man „alignen“, das Resultat ist das „Alignment“, das als Datengrundlage der phylogenetischen Rekonstruktion dient. Die Schwierigkeit beim Alignen besteht darin, die genauen Positionen der Indels ausfindig zu machen, und Gaps an der richtigen Stelle einzufügen. Die beiden folgenden Sequenzen lassen sich z. B. auf zwei verschiedene Arten alignen.

1	ATGCGTCGTT	1	AT--GCGTCGTT
2	ATCCG-CGTC	2	ATCCGCGTC

Besonders bei großen Datensätzen mit vielen Arten und variablen Sequenzen lässt sich ein Alignment deshalb nicht einfach nach Augenmaß durchführen. Erstens schädigt das auf Dauer die Augen. Zweitens ist die Methode nicht objektiv. Man neigt im Zweifelsfall dazu, Gaps so zu verteilen, dass ähnliche Arten auch ähnliche Sequenzen haben. So erzeugt man einen Datensatz, der a priori die eigenen Hypothesen unterstützt.

Durch Einfügen beliebig vieler Gaps lässt sich ein perfektes Alignment ohne Substitutionen erzielen, bei dem an jeder homologen Position nur gleiche Nukleotide oder Gaps vorkommen. Ein solches Alignment würde nicht den natürlichen Verhältnissen entsprechen, unter denen Substitutionen erwiesenermaßen vorkommen. Wenn ein Computer ein optimales Alignment erstellen soll, muss man ihm ein Optimierungskriterium vorgeben, an dem er sich bei seinen Berechnungen orientieren kann. Bei dem von uns verwendeten Programm ClustalX werden Indels genau wie Substitutionen als Evolutionsergebnisse gewertet und sozusagen mit Strafpunkten („Gap Penalties“) belegt. Das Programm findet dann die Lösung, die mit den wenigsten Strafpunkten zu erreichen ist. Die Frage bleibt, wie man Indels im Vergleich zu Substitutionen wichten soll. Bei verschiedenen Gap Penalties erhält man verschiedene Alignments, eine vertrackte Situation, da die Analyse auf der Annahme aufbaut, dass die Positionen eines Alignments homolog sind und nicht mehrere Positionen derselben Sequenz homolog sein können. Mehrere Auswege sind dafür vorgeschlagen worden: (1) Man verwirft Positionen mit unsicherem Alignment (Positionen mit vielen Gaps); (2) Man wichtet bei der späteren Analyse solche Positionen geringer als andere Stellen; (3) Man erstellt mehrere Alignments und kombiniert die Daten zu einem Superdatensatz („elision method“). Hierbei wichten sich die unsichereren Positionen sozusagen von selbst herunter, da sie in den unterschiedlichen Alignments zu verschiedenen phylogenetischen Resultate führen. (4) Man verwendet Methoden, die phylogenetische Bäume ohne Alignment errechnen können (z. B. POY: <http://www.csc.fi/molbio/progs/poy/>).

ClustalX vergleicht zunächst alle Sequenzen paarweise miteinander, erstellt paarweise Alignments und eine Ähnlichkeitsmatrix. Aufgrund der Ähnlichkeit wird dann ein Dendrogramm, eine Art vorläufiger Stammbaum, der Sequenzen errechnet. Anhand dieses Dendrogramms alignt ClustalX als erstes die zwei ähnlichsten Sequenzen und arbeitet sich langsam zu den unähnlicheren vor, bis es den ganzen Datensatz alignt hat.

1. Öffnen Sie „ClustalX“.
2. **>File>Load Sequences**
3. Laden Sie den in Bioedit erstellten Datensatz.
4. **>Alignment>Do Complete Alignment**. Das Programm alignt die Sequenzen mit seinen eigenen Voreinstellungen.
5. **>File>Save Sequences As ...** Wählen Sie einen neuen Namen, unter dem Sie das alignment speichern. Klicken Sie unter „Format“ **>Clustal** an.
6. Spaßeshalber wollen wir jetzt noch ein alignment erstellen, bei dem das Einfügen von gaps kostenfrei ist, bei dem also das Programm nur nach Ähnlichkeit der Nukleotide entscheidet. Markieren Sie alle Sequenzen mit der Maus und drücken Sie **>Edit>Remove All Gaps**
7. **>Alignment>Alignment Parameters>Multiple Alignment Parameters**. Die für uns interessanten Parameter sind die „Gap Opening Penalty“ und die „Gap Extension Penalty“. Der erste Wert bezeichnet die Kosten für das Einfügen einer neuen gap, der zweite den für die Verlängerung einer solchen gap. Die Verlängerung eines indels ist „kostengünstiger“, weil längere indels oft in einem einzigen Schritt als Ganzes entstehen und nicht als mehrere Ereignisse gerechnet werden müssen.
8. Stellen Sie beide Parameter auf „0.00“ und drücken Sie **>CLOSE**.
9. **>Alignment>Do Complete Alignment**
10. Speichern Sie den file wieder unter einem neuen Namen und vergleichen Sie die beiden alignments.
11. Zur weiteren Bearbeitung der alignments, schließen Sie ClustalX und kehren Sie zu Bioedit zurück.
12. Laden Sie das erste alignment in Bioedit: **>File>Open**. Zusätzlich zu den Sequenzen sehen Sie eine weitere Zeile, die invariable Positionen des alignments anzeigt.
13. Unter **>View** finden Sie verschiedene Ansichtsmodi. Wählen Sie den, der Ihnen für die Kontrolle des alignments am besten erscheint.
14. Scrollen Sie durch das alignment und verbessern Sie es, wenn Sie deutliche Fehler erkennen.
15. Wenn Sie mit dem alignment zufrieden sind, löschen Sie die Zeile mit der Clustal Consensus Information (Titel anklicken **>Edit>Delete Sequence(s)**) und exportieren Sie das endgültige alignment in PAUP/NEXUS-Format (**>File>Export>Sequence Alignment**, dann Namen wählen und unter Dateityp **>PAUP/NEXUS (*.pau, *.nex** wählen). Dieses Format benötigen wir bei den folgenden Analysen.
16. Wenn es Sie interessiert, können Sie sich die Unterschiede zwischen den Formaten in einem Texteditor oder in Word ansehen, indem Sie die entsprechenden Dateien laden.

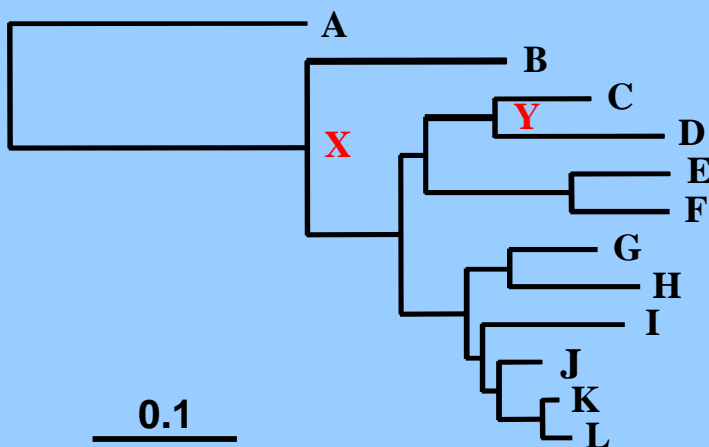
Datenanalyse: Errechnen phylogenetischer Stammbäume

Aus molekularen Daten kann man eine Vielzahl von Informationen extrahieren. Wir beschränken uns in diesem Kurs aber auf einige Fragen, die üblicherweise in der Systematik auftauchen. Die am häufigsten gestellte Frage ist die nach der Verwandtschaft der Organismen. Nach der Evolutionstheorie gehen alle Organismen auf einen Uroorganismus zurück und sind demnach miteinander verwandt. Da die DNA von Generation zu Generation repliziert und weitergegeben wird, sind auch alle homologen DNA-Sequenzen und die daraus synthetisierten Proteine miteinander verwandt. Deshalb kann man Protein- oder DNA-Sequenzen dazu verwenden, die Evolutionsgeschichte zu rekonstruieren und die Verwandtschaftsverhältnisse der Organismen aufzudecken. Je länger zwei Sequenzen unabhängig

voneinander existieren, desto mehr Substitutionen haben sie in der Regel akkumuliert. Man kann also genetische Unterschiede (die genetische Divergenz) zwischen Arten als Maß für ihre Verwandtschaft heranziehen und aus dieser Divergenz einen phylogenetischen Stammbaum errechnen. Ein solcher Stammbaum ist ein **Modell** oder eine **Hypothese** der verwandtschaftlichen Beziehungen zwischen den untersuchten Arten.

Box 1: Phylogenetische Bäume

Ein phylogenetischer Stammbaum verdeutlicht modellhaft die evolutionären Beziehungen zwischen Organismen. Er besteht aus Ästen, die an Knoten miteinander verbunden sind. Im folgenden Beispiel sind zwei interne Knoten mit X und Y bezeichnet. Die Organismen stellen die Endpunkte der Äste dar, und werden auch als terminale Knoten bezeichnet. In unserem Fall sind dies die Endpunkte A-L. Die Arten K und L sind in diesem Beispiel näher miteinander verwandt als H und L. Die Länge der Äste ist proportional zur Anzahl der Substitutionen von einem Knoten zum nächsten. Der Maßstab unten links zeigt die Astlänge für 0.1 Substitution je Position des Alignments an.



Die meisten phylogenetischen Stammbäume sind wie dieser dichotom verzweigt. Das heißt, es haben sich an jedem Knoten aus einer Ursprungsart zwei Tochterarten gebildet. Es können aber auch mehr als zwei Tochteräste von einem Knoten ausgehen (polytome Verzweigung). Ohne dass sich der Informationsgehalt eines phylogenetischen Stammbaums ändert, kann man die Äste wie bei einem Mobile um ihre Knoten drehen. E und F sind z. B. in diesem Baum nicht näher mit der Gruppe G-L verwandt als C und D. Phylogenetische Bäume können eine Wurzel haben oder nicht. Sind sie gewurzelt, weiss man, von wo die Evolution ihren Ausgang genommen hat. Im anderen Falle könnte jeder interne Knoten den ersten Artbildungsschritt darstellen.

Die mögliche Anzahl verschiedener phylogenetischer Bäume für eine bestimmte Anzahl von Arten ist immens. Für n Arten beträgt sie:

$$U_n = (2n-5)(2n-7) \dots (3)(1) \text{ für ungewurzelte Bäume und}$$

$$R_n = (2n-3)(2n-5) \dots (3)(1) \text{ für gewurzelte.}$$

Für $n = 20$ ist $R_n = 8\,200\,794\,532\,637\,891\,559\,000$, für $n = 100$ gibt es mehr Bäume als Atome im Universum. Aus dieser Menge gilt es, den optimalen Baum herauszufinden. Schon bei relativ kleinen Datensätzen lassen sich nicht mehr alle möglichen Bäume testen. Man muss auf sogenannte heuristische Suchverfahren zurückgreifen, um zu einem Ergebnis zu kommen (s. Box 3).

Für phylogenetische Bäume hat sich eine Kurzschreibweise eingebürgert, die auch von vielen Computerprogrammen gelesen wird: das Newick Format, das die Verwandtschaftsverhältnisse in Form von Klammern ausdrückt. Die Gruppe I-L sieht im Newick-Format so aus: $(I(J(K,L)))$

Aufgabe: Schreiben Sie den vollständigen Baum als ungewurzelten Baum im Newick-Format auf.

Nach welchen Kriterien soll man vorgehen, wenn man aus allen möglichen phylogenetischen Stammbäumen den herausfinden will, der die Evolution der Gruppe am besten wiedergibt? Um den „besten“ Baum unter allen möglichen herauszufinden, braucht man ein Optimierungskriterium, nach dem entschieden werden kann, wie „gut“ ein Baum im Vergleich zu anderen ist. Drei solcher Kriterien wollen wir kurz besprechen: (1) Distanzmethoden, (2) Maximum-Parsimonie-Methoden (MP) und (3) Maximum-Likelihood-Methoden (ML).

(1) Distanzmethoden berechnen die genetische Distanz zwischen den Sequenzen des Datensatzes und benutzen zum Berechnen des Baums die Distanzmatrix. Die Sequenzen selber spielen bei der Berechnung des Baums keine Rolle.

Taxon	1	2	3	4
2	d_{12}			
3	d_{13}	d_{23}		
4	d_{14}	d_{24}	d_{34}	
5	d_{15}	d_{25}	d_{35}	d_{45}

Beim einfachsten Verfahren, der UPGMA-Methode (unweighted pair-group method using arithmetic averages) werden die beiden Arten mit der geringsten Distanz zu einer Gruppe zusammengefasst. Im nächsten Schritt wird eine vereinfachte Distanzmatrix errechnet, bei der die zusammengefassten Taxa als Gruppe vorkommen. Nehmen wir an, Taxa 1 und 2 weisen die geringste genetische Distanz auf. Dann sieht die neue Matrix wie folgt aus:

Taxon	i = (1-2)	3	4
3	d_{i3}		
4	d_{i4}	d_{34}	
5	d_{i5}	d_{35}	d_{45}

Auf diese Weise werden schrittweise die Arten und Gruppen mit den jeweils kleinsten genetischen Distanzen miteinander verbunden bis alle Taxa zu einem einzigen Baum verknüpft sind.

(2) MP-Methoden greifen direkt auf die Sequenzen zurück, um den optimalen Baum zu berechnen. Sie gehen davon aus, dass die Topologie die Verhältnisse am besten wiedergibt, die mit den wenigsten Substitutionsereignissen die heute zu beobachtenden Sequenzunterschiede erklärt. Praktisch geht man hierbei von den Sequenzen des Alignments aus, ermittelt für jede Nukleotid-Position das Nukleotid der Vorläufersequenzen an den internen Knoten des Baumes, und zählt für jede Topologie die minimale Anzahl an Substitutionen, die die Evolution der Gruppe erklärt. Von allen untersuchten Topologien ist dann diejenige mit den wenigsten Substitutionen die optimale. MP verwendet nicht alle Positionen eines Alignments. Positionen, die in allen Sequenzen gleich sind, sind generell uninformativ. Bei Positionen, an denen nur eine Sequenz eine bestimmte Substitution aufweist, muss diese Substitution auf dem terminalen Ast liegen. Diese Positionen lassen unter dem Parsimonie-Kriterium keine Schlüsse auf die verwandtschaftliche Beziehung der Art zu. Als Parsimonie-informativ bezeichnet man alle Positionen, an denen mehr als eine Art eine bestimmte Substitution aufweist. MP-Methoden tun sich generell schwer mit Datensätzen, die viele Homoplasien aufweisen (s. Box 2).

Box 2: Homologie und Homoplasie

Identische Nukleotide an einer Position des Alignments die zwei oder mehr Sequenzen von einem gemeinsamen Vorfahren ererbt haben, werden als **Homologie** bezeichnet. Wenn die Sequenzen das Nukleotid unabhängig voneinander durch Mutation erworben haben, spricht man von einer **Homoplasie**. In den meisten Fällen man davon ausgehen, dass identische Nukleotide an einer Position homolog sind. Bei den vier Sequenzen im folgenden Beispiel führt das allerdings zu Schwierigkeiten.

Position	1	2	3	4	5	6	7	8	9
Sequenz A	A	A	A	C	T	T	C	G	G
Sequenz B	A	A	A	C	G	T	T	C	G
Sequenz C	A	A	A	C	G	T	T	C	G
Sequenz D	A	A	A	C	T	T	C	G	G

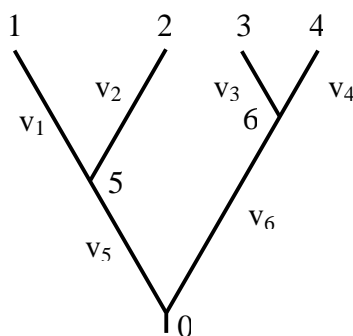
Angenommen der wahre Baum sieht so aus ((A,D)(B,C)). Dann haben entweder Sequenz A und B das G oder Sequenz C und D das C an Position 8 unabhängig voneinander erworben. Wenn die wahre Topologie ((A,B)(C,D)) ist, gilt analog dasselbe für das G und C in Position 5. Homoplastische Positionen einer DNA-Sequenz enthalten also Informationen, die der optimalen Baumtopologie widersprechen. Je höher der Anteil an Homoplasien, desto schwieriger wird die Berechnung eines optimalen Baumes. Als ein Maß für die Verlässlichkeit eines MP-Baumes werden daher meist Homoplasie-Indices angegeben:

Zur Berechnung des Consistency Index CI wird für jede Position des Alignments der Quotient $c_i = m_i / s_i$ (mit m_i = minimale Anzahl von Substitutionen in allen möglichen Topologien und s_i = beobachtete Anzahl von Substitutionen bei der betrachteten Topologie) ermittelt. CI ist dann das Mittel aus den Werten für alle Positionen:

$$CI = \frac{\sum_i m_i}{\sum_i s_i}$$

Dieser Wert ist nicht für alle Topologien identisch. Der Retention Index $RI = (\sum_i g_i - \sum_i s_i) / (\sum_i g_i - \sum_i m_i)$ (mit g_i = maximale Anzahl Subst. in allen möglichen Topologien) und der Rescaled Consistency Index $RC = CI \times RI$ sind dagegen unabhängig von der Topologie.

(3) Das Optimierungskriterium bei ML-Methoden ist die Likelihood (nicht Wahrscheinlichkeit) des Datensatzes unter Vorgabe einer bestimmten Topologie und eines bestimmten Substitutionsmodells. Die Topologie, bei der die Likelihood maximal ist, ist die optimale.



ML-Methoden sind rechnerisch extrem aufwändig. In diesem Beispielbaum mit vier Taxa 1, 2, 3, 4 und zwei internen Knoten 5, 6 bezeichnen $v_1, v_2 \dots v_6$ die Astlängen (richtiger: die erwartete Anzahl von Substitutionen für den i -ten Ast). Diese erwartete Anzahl entspricht dem Produkt aus der Substitutionsrate r_i und der verstrichenen Zeit t_i : $v_i = r_i t_i$. Als $x_1, x_2 \dots x_6$ bezeichnen wir die Nukleotide an einer bestimmten Position k des Alignments, mit g_{x0} die Wahrscheinlichkeit, dass der Knoten 0 an Position k das Nukleotid x hat, und mit P_{ij} die Wahrscheinlichkeit, dass ein Nukleotid x_i durch ein

anderes Nukleotid x_j ersetzt wird. Dann entspricht die Likelihood-Funktion für die Position k des Alignments:

$$l_k = g_{x_0} P_{x_0x_5}(v_5) P_{x_5x_1}(v_1) \dots P_{x_6x_4}(v_4)$$

Die Wahrscheinlichkeiten des Übergangs von x_i nach x_j kann man durch ein Substitutionsmodell errechnen. Da die Nukleotide an den inneren Knoten in der Praxis nicht bekannt sind, muss diese Funktion für alle möglichen Nukleotide berechnet werden (L_k). Zuletzt muss man die Berechnung für alle Positionen des Alignments durchführen. Die Likelihood für die gesamte Sequenz entspricht dem Produkt der L_k Werte aller Positionen

$$L = \prod_{k=1 \text{ bis } n} L_k$$

Die gängigen Computerprogramme geben immer den Logarithmus dieses Wertes, die sogenannte log-likelihood, an:

$$\ln L = \sum_{k=1 \text{ bis } n} L_k$$

Wie man sich denken kann, ist die Ermittlung eines ML-Baums mit einer Wahnsinnsrechnung verbunden. Schon bei relativ wenigen Taxa können auch leistungsfähige Computer nicht mehr alle möglichen Baumtopologien durchtesten. Stattdessen wendet man sogenannte heuristische Suchalgorithmen an, die den besten Baum aus einer begrenzten Anzahl untersuchter Bäume zu ermitteln versuchen.

Box 3: Heuristische Suche (heuristic search)

Eine heuristische Suche nach dem optimalen Baum erfolgt in zwei Schritten. Im ersten Schritt wird ein Ausgangsbaum erzeugt. Eine häufig verwendete Methode ist der Stepwise Addition Algorithm. Einem Ausgangsbaum von 3 Arten werden schrittweise weitere Arten angefügt, wobei man diese immer an der Stelle einfügt, an der man einen optimalen (Teil-)Baum erhält.

Ausgehend vom vollständigen Ausgangsbaum, beginnt das Programm, Äste zu vertauschen und berechnet, ob sich so ein besserer Baum finden lässt. Es gibt wiederum verschiedene Methoden für dieses sogenannte Branch Swapping. Beim Nearest Neighbour Interchange (NNI) werden alle Bäume untersucht, die sich vom Ausgangsbaum in der Position zweier Arten unterscheiden. Beim Subtree Pruning and Regrafting (SPR) wird ein Ast abgeschnitten und an allen möglichen Stellen des Ausgangsbaumes wieder angesetzt. Dies wird für alle Äste des Baumes wiederholt. Beim Tree Bisection and Reconnection Algorithmus (TBR), wird der Ausgangsbaum in zwei Teile geschnitten und in allen möglichen Positionen wieder zusammengefügt. Dies wird ebenfalls für alle möglichen „Schnittstellen“ wiederholt. Die Zahl der untersuchten Bäume ist bei TBR größer als bei SPR und NNI. Trotzdem untersucht man immer nur eine kleine Zahl aller möglichen Bäume.

Um die Chance zu verkleinern, dass man auf diese Weise den optimalen Baum verfehlt, führt man meist mehrere heuristische Suchläufe durch, wobei man im ersten Schritt die Arten jeweils in willkürlicher Reihenfolge einfügt.

Die am häufigsten verwendeten Computerprogramme zur phylogenetischen Analyse sind PAUP, Phylip und MEGA. Phylip (<http://evolution.genetics.washington.edu/phylip.html>) und MEGA (<http://www.megasoftware.net/>) sind kostenlos im Internet erhältlich. PHYLIP implementiert eine Vielzahl verschiedener Methoden und ist das wahrscheinlich am weitesten verbreitete phylogenetische Programmpaket. PAUP ist ähnlich weit verbreitet und in der Benutzung einfacher, allerdings nicht kostenlos. Es existieren Versionen für Mac und Windows. Die drei Programme benötigen unterschiedliche Datenformate als input files. Mit Hilfe von Bioedit können Sie alignments auch in Phylip-Format (*.phy) **speichern**. PAUP verwendet das sog. NEXUS-Format (*.nex) und kann selber Phylip-files konvertieren. Sie können alignments aus Bioedit in Nexus-Format **exportieren**. MEGA benutzt ein Format, das von Bioedit nicht erstellt werden kann, konvertiert aber selber Phylip- und NEXUS-files. Die Windows-Version von PAUP wird im wesentlichen über eine Befehlszeile gesteuert. PAUP ist ein enorm vielseitiges Programm (mit einer ungeheuren Menge von Befehlen). Wir können nicht mehr als ein paar Grundfunktionen kennenlernen und beginnen mit einer einfachen Analyse. PAUP benutzt in der Grundeinstellung Maximum Parsimony (MP) als Optimierungskriterium.

1. Öffnen Sie das Programm „PAUP“ und laden Sie das endgültige alignment aus der sich automatisch öffnenden Dialogbox. Wenn alles gut gegangen ist, sollte der file ohne Fehlermeldung eingelesen werden. Häufige Fehlerquelle bei diesem Schritt sind Leerzeichen, Punkte oder Striche in den Artnamen.
2. Sie erhalten im Fenster oberhalb der Befehlszeile einige Informationen zu ihrem Datenfile. Wenn Sie auf **>Window>[Name Ihres Files]** klicken, öffnet sich ein zweites Fenster mit dem Datenfile. Sie können nun zwischen diesen Ansichten wechseln.
3. Schon bei Datensätzen von 10 Arten ist die Zahl der möglichen Bäume so groß, dass die vollständige Suche nach dem besten Baum sehr lange dauert. Sie haben in der Vorbesprechung von heuristischen Suchmethoden gehört. Im ersten Versuche lassen wir PAUP mit einer heuristischen Suchmethode nach dem optimalen Baum suchen.
4. Eingabe: `hsearch addseq=asis; <Enter>`.
Das Programm berechnet den Baum, indem es die Sequenzen in der Reihenfolge zusammenfügt, wie sie im Datensatz erscheinen. Im output wird u. a. angezeigt, wieviele Merkmale der Datensatz enthält, wie viele davon parsimonie-informativ sind, und welcher „branch-swapping algorithm“ verwendet wurde. Die default-Einstellung verwendet TBR (tree-bisection-reconnection). Außerdem erfahren Sie in einer Tabelle unter „score“ wie viele Substitutionsschritte der beste gefundene Baum hat.
5. Um sich den Baum anzusehen, geben Sie ein:
`describetrees 1/ brlens=yes; <Enter>`
Jetzt zeigt das Programm den Baum Nr. 1 an und gibt eine Tabelle mit den Astlängen (brlens) aus. Ausserdem werden verschiedene Werte angegeben, die alle das Ausmaß an Homoplasie in ihrem Datensatz beschreiben. Der Consistency Index (CI) variiert mit der Topologie des Baums; seine untere Grenze ist nicht gleich 0. Informativer sind deshalb der Retention Index (RI) und der Rescaled Index (RC), die zwischen 0 und 1 variieren. Diese Indices sollten nur aufgrund der informativen Positionen berechnet werden. Der Homoplasy Index ($HI = 1 - CI$) bezeichnet den proportionalen Anteil an Positionen mit parallelen oder Rückmutationen.
6. Bei der heuristischen Suche wird nur ein kleiner Teil der möglichen Bäume wirklich untersucht. Um die Chance, den optimalen Baum zu finden, zu vergrößern, kann man mehrere Zyklen heuristischer Suchen durchlaufen. Dabei wird die Reihenfolge der Sequenzen am besten mit jedem Zyklus variiert.
Eingabe: `hsearch addseq=random swap=NNI nreps=100; <Enter>`
Es werden 100 Zyklen (nreps) durchlaufen, wobei die Sequenzen jedesmal in einer anderen Reihenfolge (random) hinzugefügt werden. Als branch swapping Algorithmus wird diesmal nearest neighbor interchange (NNI) verwendet.
7. Um den errechneten Baum anzuzeigen, klicken Sie diesmal auf den Pfeil rechts neben der Befehlszeile. Sie sehen eine Liste der zuletzt verwendeten Befehlszeilen. Wählen Sie die richtige aus und drücken Sie `<Enter>`.
8. Die Bäume, die von PAUP ausgegeben werden, sind ungewurzelt. Jetzt wollen wir den Baum mit Hilfe einer Außengruppe (outgroup) verwurzeln. Oft bestimmt man solche outgroups schon vor der Analyse und nimmt sie genau deswegen in die Analyse auf. Wählen Sie nach Rücksprache mit dem Betreuer eine outgroup. Der Befehl lautet:
`outgroup [Artname(n)]; <Enter>`
9. Um den Baum mit einer outgroup zu wurzeln, geben Sie ein:
`describetrees 1/ root=outgroup outroot=monophyl; <Enter>`
Die outgroup erscheint als monophyletischer Clade neben den Arten der ingroup.

10. Um den Baum/ die Bäume mitsamt Astlängen zu speichern, geben Sie ein:

```
savetrees file=[Dateiname] brlens=yes; <Enter>
```

Sie haben den ersten errechneten Baum nun als vorläufiges Ergebnis gespeichert. Die Ansicht von Bäumen in PAUP ist allerdings recht unbequem. Um den Baum manipulieren zu können, die Ansicht zu verändern und den Baum ausdrucken zu können, starten Sie jetzt das Programm Treeview.

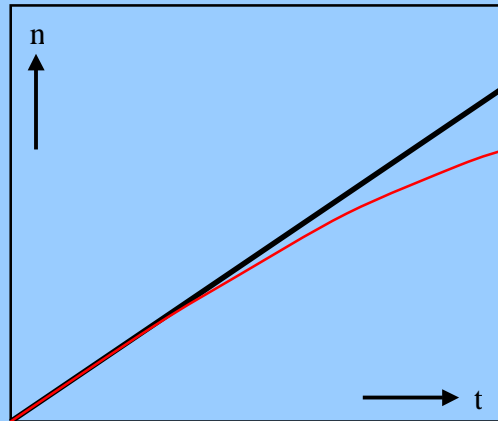
1. Laden Sie den treefile durch **>File>Open**
2. In der Menüzeile über dem Fenster sehen Sie verschiedene Symbole für die Darstellung von Bäumen. Probieren Sie diese verschiedenen Ansichten aus. Worin unterscheiden sie sich?
3. Sie können auch in Treeview eine outgroup für ihren Baum bestimmen und den Baum danach rooten. Drücken Sie **>Tree>Define outgroup** Es erscheint eine Dialogbox mit zwei Fenstern.
4. Markieren Sie die Arten ihrer outgroup im linken Fenster (Ingroup) und schieben Sie sie durch Druck auf die Pfeiltaste in das rechte Fenster (Outgroup). Schließen Sie mit **>OK** ab.
5. **>Tree>Root with outgroup**. Die outgroup steht jetzt zuoberst, der restliche Baum „hängt“ sozusagen an der outgroup.
6. Wenn Sie die outgroup lieber am unteren Ende des Baumes anhängen wollen, drücken Sie **>Tree>Order>Ladderize left**.
7. Sie können den Baum in Treeview durch verschiedene andere Manipulationen für einen Ausdruck vorbereiten. **>Edit>Edit tree**.
8. In der Menüzeile oberhalb des neu geöffneten Fensters sehen Sie 9 Symbole und ein erklärendes Textfenster. Mit der linken Gruppe von 4 Symbolen verändern Sie die Topologie des Baumes. Dies kann für spezielle Fragestellungen nötig sein, stellt aber eine Manipulation ihrer Ergebnisse dar. Mit der mittleren Gruppe von 4 Symbolen verändern Sie dagegen nur das Aussehen des Baumes, nicht die Topologie. Diese Funktionen können Sie jederzeit frei verwenden, um Ihren Baum besser darzustellen. Mit dem rechten Symbol können Sie interne Knoten beschriften. Verändern Sie den Baum in ihrem Sinne und speichern Sie ihn unter einem neuen Namen ab.
9. Sie können mit den Befehlen **>Edit>Preferences** oder den Befehlen unter **>Style** die Beschriftung des Baums verändern. Mit **>File>Print preview** können Sie eine Druckansicht des Baumes öffnen (schließen mit **>Close**). Zuletzt können Sie die Ansicht mit **>File>Save as graphic** als Graphik speichern, um sie in anderen Programmen (z. B. Powerpoint) zu verwenden.

Auswahl des besten Substitutionsmodells

Wenn man Maximum Likelihood (ML) als Optimierungskriterium verwendet, muss man sich im voraus zwischen verschiedenen Substitutionsmodellen entscheiden. Im Prinzip ist dabei jedes Modell falsch, da es die wahren Verhältnisse nur grob vereinfacht wiedergibt. Je komplizierter ein Modell ist, desto genauer spiegelt es normalerweise die Wirklichkeit wieder, aber desto größer ist auch die Gefahr, dass die Annahmen des Modells in der Wirklichkeit nicht zutreffen. ML bietet die Möglichkeit, sich aufgrund eines statistischen Kriteriums zwischen Modellen zu entscheiden. Man maximiert, wie bei der Suche nach dem optimalen Baum, die Likelihood-Funktion über alle Parameter eines vorher ausgewählten Modells. Dann wählt man ein komplizierteres Modell und wiederholt die Berechnung. So sucht man diesmal nicht nach dem Baum, sondern nach dem Substitutionsmodell, für das die Likelihood-Funktion maximal wird. Im Prinzip lassen sich Modelle unbegrenzt verkomplizieren, um damit die Likelihood-Funktion zu „verbessern“. Oberhalb einer gewissen Grenze sind diese Verbesserungen aber nur noch minimal.

Box 4: Substitutionsmodelle

Zwei Sequenzen, die von einer Ursprungssequenz abstammen, akkumulieren im Laufe der Zeit Substitutionen. Ihre genetische Divergenz wächst an. Die Zahl der beobachteten Substitutionen sollte also eigentlich linear von der Zeit abhängen, die verstrichen ist, seit zwei Sequenzen sich voneinander getrennt haben (schwarze Linie). Mit geringer Wahrscheinlichkeit kommt es aber auch zu Rückmutationen oder mehrfachen Mutationen an derselben Position. Deshalb ist die Zahl der beobachteten Substitutionen geringer als die Zahl der Mutationsereignisse (rote Linie). Das einfachste Maß: $p = n_d / n$ gibt also die Zeitverhältnisse nicht korrekt wieder.



Das einfachste Substitutionsmodell (Jukes-Cantor Modell) nimmt an, dass die Substitutionsraten zwischen allen Nukleotiden gleich sind. In Wirklichkeit unterscheiden sich Transitions- und Transversionsraten in der Regel (Kimura 2-Parameter Modell). Auch kommen die verschiedenen Nukleotide in unterschiedlichen Mengen vor, was die Substitutionsraten ebenfalls beeinflusst (HKY Modell). Im GTR Modell nimmt man unterschiedliche Raten für alle Arten von Substitutionen an.

	A	T	C	G
A	-	α	α	α
T	α	-	α	α
C	α	α	-	α
G	α	α	α	-

Jukes-Cantor

	A	T	C	G
A	-	β	β	α
T	β	-	α	β
C	β	α	-	β
G	α	β	β	-

Kimura

	A	T	C	G
A	-	βg_T	βg_C	αg_G
T	βg_A	-	αg_C	βg_G
C	βg_A	αg_T	-	βg_G
G	αg_A	βg_T	βg_C	-

HKY

	A	T	C	G
A	-	$a g_T$	$b g_C$	$c g_G$
T	$a g_A$	-	$d g_C$	$e g_G$
C	$b g_A$	$d g_T$	-	$f g_G$
G	$c g_A$	$e g_T$	$f g_C$	-

GTR

Alle Substitutionsmodelle sind nur Annäherungen an die Verhältnisse in der Natur. Im Prinzip lassen sich beliebig komplizierte Modelle formulieren. Wenn die Substitutionen in der Natur einem komplizierten Modell folgen, dann unterschätzen einfachere Modelle die Zahl der Substitutionen. Allerdings sind komplizierte Modelle unnötig, wenn die genetische Divergenz $p \leq 0,1$ ist. Einfache Modelle sind dann ebenso genau und die erwartete Zahl von Substitutionen hat eine kleinere Varianz, ist also präziser zu errechnen.

Alle diese Modelle nehmen an, dass die Substitutionsrate für alle Positionen einer Sequenz die gleiche ist. Das stimmt in der Regel nicht. Proteine haben aktive Zentren, die sehr konservativ sind, während in anderen Bereichen Substitutionen häufiger sind. 1., 2. und 3. Codonpositionen haben ebenfalls unterschiedliche Substitutionsraten, weil viele Mutationen an der 3. Position „still“ sind, d. h. nicht zu Veränderungen der Aminosäuresequenz führen. Die Substitutionsrate variiert also von Position zu Position. Auch diese Variation kann man in Substitutionsmodellen berücksichtigen, indem man verschiedenen Positionen des Alignments unterschiedliche Substitutionsraten aus einer Häufigkeitsverteilung (der sogenannten Gamma-Verteilung) zu weist. Substitutionsmodelle mit variabler Rate sind meist realistischer als solche mit konstanter Rate und sind z. B. für die genaue Berechnung von Astlängen besser geeignet.

Wenn das einfache Modell jeweils ein Spezialfall des komplizierteren ist, kann man testen, ob die Unterschiede zwischen beiden Modellen noch signifikant sind. Ein nicht signifikanter Unterschied bedeutet dabei, dass das kompliziertere Modell den Datensatz nicht wesentlich besser erklären kann als das einfachere (obwohl die Likelihood vielleicht noch ein wenig größer ist). In diesem Falle entscheidet man sich für das einfachere Modell und bricht die Suche ab. In der Praxis führt man zu diesem Zweck eine ganze Reihe von Likelihood-Verhältnistests (Likelihood Ratio Tests, s. unter „Testen phylogenetischer Hypothesen“) durch. Das Programm „Modeltest“ kann 56 verschiedene Modelle testen und erledigt diese Aufgabe in Zusammenarbeit mit PAUP.

1. Öffnen Sie PAUP und laden Sie Ihren Nexus-file durch **>File>Open**.
2. Eingabe: `execute [Dateiname]; <Enter>`
3. Eingabe: `default lscores longfmt=yes; <Enter>`
4. Öffnen Sie jetzt die Datei C:\Programme\Modeltest\Modeltest 3.06 folder\example files\modelblock3.nex. durch **>File>Open**. PAUP berechnet für jedes Modell die Likelihood-Werte und speichert sie unter C:\Programme\Modeltest\Modeltest 3.06 folder\example files\model.scores.
5. Schließen Sie PAUP nach Beendigung des Rechengvorgangs.
6. Verschieben Sie die Datei model.scores in den Ordner C:\Programme\Modeltest\ Modeltest 3.06 folder und geben Sie ihr einen neuen Namen.
7. Öffnen Sie die DOS-Eingabeaufforderung im Windows-Startmenü durch **>Start >Programme>Zubehör>Eingabeaufforderung**
8. Durch die Eingabe `cd` gefolgt vom jeweils nächsten Ordnernamen können Sie sich langsam aus dem Verzeichnis C: in den Ordner „C:\Programme\Modeltest\Modeltest 3.06 folder“ vorarbeiten.
9. Wenn Sie dort angekommen sind, starten Sie Modeltest durch die Eingabe: `modeltest3-06.exe <[Dateiname]`. Dateiname ist der soeben verschobene output-file von PAUP.
10. Modeltest führt nun die Likelihood-Verhältnistests durch, bis sich keine signifikante Verbesserung mehr ergibt.
11. Im Ergebnis sehen Sie das optimale Substitutionsmodell und einige Zeilen, die mit „BEGIN PAUP;“ anfangen. Dieser Teil lässt sich in PAUP übertragen, um mit den richtigen Parametern eine Analyse durchführen zu können.
12. Markieren Sie den Text von „BEGIN PAUP;“ bis „END;“ mit der Maus, indem Sie auf das Symbol links in der Kopfzeile des Eingabefensters klicken und **>Bearbeiten>Markieren** anwählen.
13. **>Bearbeiten>Kopieren**
14. Wechseln Sie in PAUP, öffnen Sie die Ansicht des NEXUS-files und kopieren Sie den Textblock an das Ende des Datei. Im nächsten Teil werden wir hieraus einen Befehlsblock erstellen, der es ermöglicht, PAUP auch ohne Eingabe in die Befehlszeile zu steuern.

Erstellen eines PAUP-Blocks

Im NEXUS-Format lässt sich nicht nur ein Alignment zur Bearbeitung speichern. In verschiedenen zusätzlichen Textblöcken kann man Annahmen zum Datensatz (ASSUMPTIONS Block), vorher errechnete Bäume (TREES Block) oder Details zur Analyse (PAUP Block) eingeben. Dies ist z. B. sehr praktisch, wenn man Analysen schrittweise durchführt, wobei ein Schritt auf den Ergebnissen des vorherigen aufbaut. Ein solches iteratives Verfahren kann z. B. notwendig sein, wenn man sehr große

Datensätze unter ML mit komplizierten Substitutionsmodellen untersucht. Statt selber stunden- oder tagelang auf die Ergebnisse zu warten und dann die nächsten Befehle per Hand einzugeben, gibt man PAUP am Freitag nachmittag alle notwendigen Befehle in einem PAUP Block mit, startet die Analyse und nimmt am Montag die Ergebnisse in Empfang. Wir beginnen mit einem einfachen Beispiel, indem wir als Ausgangspunkt die Substitutionsparameter aus Modeltest verwenden.

1. Entfernen Sie den Zeilenumbruch aus der „lset“-Zeile. Mit `lset` werden die Parameter für eine ML Analyse festgelegt. Dabei bedeutet:
`base=(x y z)` Frequenz der Basen A, C und G (T ergibt sich dann von alleine),
`nst=x` Zahl der verschiedenen Substitutionsraten (je nach Modell 1, 2 oder 6)
`rmat=(x y z ...)` Rate der verschiedenen Substitutionsraten,
`rates=[equal, gamma]` Substitutionsraten sind entweder konstant oder variabel und folgen der Gamma-Verteilung,
`shape=x` Form-Parameter der Gamma-Funktion,
`pinvar=x` Anteil invariabler sites
2. Um ML als Optimierungskriterium zu verwenden und dafür zu sorgen, dass das Programm die Analyse ohne Ihre Bestätigung abschließt, fügen Sie Nach „Begin PAUP“ die folgende Zeile ein:
`set autoclose=yes criterion=likelihood;`
3. Um die Suche nach dem besten Baum zu starten fügen Sie nach der „lset“ Zeile die folgende bekannte Zeile ein:
`hsearch addseq=random swap=SPR nreps=10;`
4. In der nächsten Zeile geben Sie an, dass der beste Baum mit Astlängen angezeigt werden soll. Zur Übung stellen Sie die Zeile selbst zusammen.
5. Zuletzt lassen Sie PAUP den Baum mit Astlängen unter einem selbstgewählten neuen Dateinamen speichern.
6. Speichern und schliessen Sie den NEXUS-file.
7. Laden Sie den file jetzt erneut. Die Analyse wird durchgeführt, ohne dass Sie weitere Befehle eingeben müssen.

Phylogenetische Unsicherheit, nicht-parametrischer Bootstrap

Die Ergebnisse verschiedener phylogenetischer Analysen (z. B. MP- und ML-Bäume) desselben Datensatzes liefern oft nicht vollkommen gleiche Ergebnisse. Speziell unter MP findet man in einer Analyse häufig mehrere, gleich lange Bäume. Die Unterschiede sind in der Regel klein, können aber wichtige Details betreffen, etwa die Frage, ob eine Gattung monophyletisch ist oder nicht. Wieviel Vertrauen kann man also in das Ergebnis einer phylogenetischen Untersuchung haben? Die Frage berührt grundsätzlich alle wissenschaftlichen Ergebnisse. Während aber ein Physiker seine Messungen wiederholen kann und Durchschnitt und Standardabweichung der Messwerte errechnen kann, hat sich die Evolution nur einmal abgespielt und kann nicht experimentell wiederholt werden.

Eine Lösung dieses Problems ist die Pseudoreplikation des Datensatzes. Man sammelt nicht neue Daten in der Natur, sondern „besammelt“ den bereits vorhandenen Datensatz mehrere 100 bis 1000 mal. Das am häufigsten verwendete Verfahren hierfür ist der sog. nicht-parametrische Bootstrap. Dabei erstellt man neue Datensätze von gleicher Größe wie der ursprüngliche Datensatz, indem man willkürlich einzelne Positionen des Alignments auswählt. Durch die zufällige Auswahl werden einige Positionen mehrmals „gesammelt“, andere fallen weg. Jede Pseudoreplikation des Datensatzes unterscheidet sich von den anderen. Man errechnet für jeden Datensatz einen separaten Stammbaum, erstellt aus den Bäumen einen Konsensusbaum, und kann für jeden Ast dieses Baumes ermitteln, in wie vielen der

Bootstrap-Bäume er vorhanden war. Diesen „bootstrap-support“ (in Prozent ausgedrückt) sieht man an fast allen publizierten phylogenetischen Bäumen. Wegen der langen Rechenzeit unter ML führen wir die Bootstrap-Analyse unter MP durch.

1. Öffnen Sie PAUP und laden Sie den NEXUS-file.
2. Sollte das Programm ohne Aufforderung zu rechnen beginnen, stoppen sie den Vorgang und wechseln Sie in die Ansicht des NEXUS-files. Sie können den PAUP-Block am Ende deaktivieren, indem Sie jede Zeile in eckige Klammern setzen. Speichern und schliessen Sie die Datei und laden Sie sie dann neu.
3. Eingabe: `bootstrap nreps=500 search=heuristic conlevel=50 treefile=[Dateiname]; <Enter>`
Durch diese Eingabe starten Sie eine bootstrap-Analyse mit 500 Pseudoreplikationen (der kleinsten Menge, die verlässliche Resultate liefert) und einer heuristischen Suche nach dem besten Baum jedes Replikats. Das Programm soll als Ergebnis einen Konsensus-Baum mit allen Äste anzeigen, die in mindestens 50% der Bäume vorkommen. Die Bäume werden in einer Datei gespeichert, deren Namen Sie selber festlegen müssen.
4. Als Anzeige sehen Sie den gewünschten Konsensus-Baum und eine Tabelle mit allen Art-Gruppierungen, die sich in mindestens einem Baum fanden. Die Arten sind als Zahlen oberhalb der Spalten abgekürzt. Sternchen in den Spalten markieren eine Gruppierung dieser Arten. Die rechten Spalten zeigen, in wievielen Bäumen und in wieviel Prozent der Bäume diese Gruppe angetroffen wurde. Gruppierungen, die in weniger als 5% der Bäume auftraten, sind nicht gelistet. Vergleichen Sie den Baum mit der Tabelle.
5. Um den Konsensus-Baum zu speichern, müssen Sie erst den Treefile laden. Die Warnung “The limit of 100 trees (= “MaxTrees”) has been reached” beantworten Sie mit **>Reset Maxtrees>Automatically increase by 100**. Hierdurch erhöhen Sie die von PAUP vorgegebene maximale Anzahl von Bäumen im Arbeitsspeicher. In der Anzeige lesen Sie, wieviele Bäume PAUP errechnet hat (oft mehr als 500, weil bei einigen Bootstrap-Datensätzen mehrere gleich gute Bäume errechnet wurden).
6. Eingabe: `contree 1-[Anzahl Bäume] / strict=no majrule=yes percent=50 treefile=[Dateiname]; <Enter>`
PAUP errechnet aus allen Bäumen nochmals einen „majority rule“ Konsensus-Baum mit allen Äste, die in mindestens 50% der Bäume vorkommen. Ein strikter Konsensus-Baum, der nur Äste anzeigt, die in allen Bäumen vorkommen, wird nicht errechnet. Der Konsensus-Baum wird in einer Datei mit selbst gewähltem Namen gespeichert.
7. In der Anzeige sehen Sie nun neben dem neuen Konsensus-Baum auch eine vollständige Tabelle mit allen beobachteten Gruppierungen. Der neue Treefile kann wieder in Treeview geöffnet und bearbeitet werden.

Die Interpretation von Bootstrap-Werten ist nicht ganz einfach. Bootstrap-Werte werden intuitiv meist als Wahrscheinlichkeitswerte missdeutet. Ein Wert von 85 % bedeutet aber nicht, dass der entsprechende Ast mit einer „Wahrscheinlichkeit“ von 0,85 auch im (unbekannten) wahren Stammbaum der Organismen vorkommt, sondern dass die verwendete Methode auf der Grundlage unserer Daten in 85 % der Fälle diesen Ast rekonstruiert. Die Werte zeigen eher Präzision als Genauigkeit an. Ein Ast mit hohem Bootstrap-support kann aus den Daten mit größerer Präzision ermittelt werden. Niedrige Bootstrap-Werte zeigen an, in welche Gruppierungen eines Baumes wir nicht allzu viel Vertrauen setzen sollten. In wissenschaftlichen Publikationen werden meist nur Werte oberhalb von 50 % angegeben. Strenggenommen sollte man nur Werte ≥ 95 % als signifikante Unterstützung für einen Ast werten. Der Bootstrap-Test ist aber konservativ, d. h. die Zuverlässigkeit einer Rekonstruktion wird

regelmäßig eher unterschätzt. Auf Ästen mit Bootstrap-Werten unterhalb von 70 % sollte man aber auf keinen Fall weitergehende Hypothesen aufbauen oder brisante Schlüsse ziehen.

Testen phylogenetischer Hypothesen; die molekulare Uhr

Oft ist man nicht nur an einem phylogenetischen Stammbaum interessiert, sondern hat weitergehende Fragen, die man aufgrund des Stammbaumes beantworten möchte. Eine typische Frage bei systematischen Arbeiten lautet z. B.: Hat sich das Merkmal X in der Evolution einmal oder mehrmals entwickelt? Häufig will man auch wissen, wann sich verschiedene Arten einer Gattung oder Gattungen einer Familie voneinander abgespalten haben. Mit solchen Daten kann man oft biogeografische Fragen, etwa zur Besiedlung von Inselgruppen, beantworten oder Aussagen zu Mechanismen der Evolution treffen.

Die Astlängen eines Baumes sind das Produkt aus Substitutionsrate und Zeit. Wenn die Substitutionsrate in allen Ästen des Baumes gleich ist, geben die Astlängen deshalb direkt das relative Alter unterschiedlicher Linien an. Um die Knoten eines Baums datieren zu können, müsste also die Substitutionsrate in allen Ästen des Baumes homogen sein. In den meisten Fällen variieren aber die Substitutionsraten des optimalen Baumes in verschiedenen Ästen erheblich, was sich an sehr unterschiedlichen Astlängen zeigt. Ein Baum mit homogener Substitutionsrate gibt in diesem Fall den Datensatz offenbar verzerrt wieder. Die Frage ist, wie stark diese Verzerrung ist: Ob der Baum den Datensatz fast genau so gut erklären kann wie der optimale Baum oder ob er signifikant schlechter ist. Nur wenn der Baum nicht signifikant schlechter ist, kann man ihn zur Datierung verwenden.

Unter ML kann man die Substitutionsrate wie einen Parameter des Substitutionsmodells betrachten. Im Normalfall maximiert das Programm die Likelihood-Funktion, wobei die Substitutionsrate frei über den Baum variieren kann. Genauso wie alle anderen Parameter des Substitutionsmodells (etwa das Verhältnis von Transitionen zu Transversionen; s. oben „Auswahl des besten Substitutionsmodells“) kann man aber auch die Substitutionsrate zwischen verschiedenen Evolutionslinien homogen halten. Ein Modell mit homogener Substitutionsrate stellt lediglich einen Spezialfall des Modells mit heterogener Substitutionsrate dar.

Mit einem „Likelihood Ratio Test“ (LRT) kann man feststellen, ob die Likelihood des „Uhr“-Baumes signifikant schlechter ist als die Likelihood des optimalen Baumes. Die Werte von:

$$2 \left([-\ln \text{Spezialfall}] - [-\ln \text{komplexes Modell}] \right)$$

folgen einer χ^2 -Verteilung mit $s-2$ Freiheitsgraden (s = Anzahl der Arten im Datensatz). Statistische Tafeln enthalten Signifikanzwerte für die χ^2 -Verteilung. Alternativ kann man p-Werte unter: <http://ergo.ucsd.edu/unixstats/probcalc/index.shtml> errechnen lassen. Ein p-Wert $> 0,05$ bedeutet dabei, dass die Likelihood-Werte nicht signifikant verschieden sind. Die Annahme einer molekularen Uhr ist gültig. Bei p-Werten $\leq 0,05$ ist der „Uhr“-Baum signifikant schlechter als der optimale Baum und die Annahme homogener Substitutionsraten wird statistisch zurückgewiesen. Dieser Test ist nur zulässig, wenn die Topologie und das Substitutionsmodell zur Errechnung beider Bäume identisch ist. Der einzige variable Parameter darf die Substitutionsrate zwischen verschiedenen Evolutionslinien sein. Man errechnet die Likelihood des optimalen Baumes mit variabler Substitutionsrate und vergleicht die Likelihood desselben Baumes mit homogener Substitutionsrate.

1. Öffnen Sie PAUP und laden Sie den NEXUS-file mit anhängendem PAUP-Block, den Sie vorher erstellt haben.
2. Brechen Sie die Berechnung ab.
3. Laden Sie nun den optimalen Baum, den Sie vorher mit Hilfe des PAUP-Blocks errechnet haben.
4. Berechnen Sie noch einmal den Likelihood-Wert dieses Baumes unter den mit Modeltest bestimmten optimalen Parametern.
Eingabe: `lscores 1;` <Enter>
5. Die Berechnung des Likelihood-Wertes unter Annahme einer molekularen Uhr erfordert einen gewurzelten Baum. Bestimmen Sie die outgroup mit dem bekannten Befehl.

6. Wurzeln Sie den Baum mit dem Befehl:
`roottrees outroot=monophyl; <Enter>`
7. Um den Likelihood-Wert unter Annahme einer molekularen Uhr zu erhalten, geben Sie ein:
`lscores 1/ clock=yes; <Enter>`
8. Berechnen Sie die Likelihood Ratio und prüfen Sie in der Tabelle oder auf der angegebenen Internetseite, ob die molekulare Uhr zurückgewiesen wird oder nicht.

Bayessche Analyse phylogenetischer Datensätze

ML berechnet die Likelihood des Datensatzes unter einem bestimmten Modell (Baumtopologie, Astlängen, Substitutionsparameter). Der optimale Baum ist der, bei dem die Likelihood des Datensatzes am höchsten ist. Dabei wird der Gaul in gewisser Weise von hinten aufgepäht. Intuitiv wäre es einleuchtender, die Wahrscheinlichkeit eines Baumes auf der Grundlage des Datensatzes zu berechnen. Der presbyterianische Reverend Thomas Bayes fand als erster eine Formel, mit der man Likelihood-Werte in A-posteriori-Wahrscheinlichkeiten („Posterior Probabilities“) konvertieren kann. 1763 wurde dieses „Bayes’sche Theorem“ posthum publiziert. Der statistische Hintergrund ist relativ kompliziert. Bei einfachen Experimenten mit wenigen möglichen Resultaten (Kopf oder Zahl beim Werfen einer Münze) lassen sich solche Wahrscheinlichkeiten analytisch berechnen. Die A-posteriori-Wahrscheinlichkeit phylogenetischer Bäume mit vielen Parametern und einer quasi unendlichen Anzahl möglicher Ergebnisse lässt sich leider nicht exakt berechnen. Man kann aber zu einer Näherungslösung kommen, indem man eine Zufallsstichprobe der wahrscheinlichsten Bäume „sammelt“. Die Wahrscheinlichkeit eines Parameters der Phylogenie (z. B. einer bestimmten Evolutionslinie oder Astlänge) entspricht dann der Häufigkeit seines Auftretens in dieser Stichprobe.

Erst 1997-1999 wurden Methoden publiziert, mit denen man die A-posteriori-Wahrscheinlichkeit aus Zufallsstichproben von Bäumen berechnen kann, in denen die Bäume proportional zu ihrer Likelihood vertreten sind. Die dabei verwendete, statistische Methode wird als „Markov Chain Monte Carlo“, MCMC, bezeichnet. Stellen Sie sich einen Roboter vor, der darauf programmiert ist, in einer Hügellandschaft herumzulaufen. Immer wenn der nächste Schritt bergauf geht, tut der Roboter den Schritt. Wenn dieser Schritt bergab führt, dann berechnet der Roboter das Höhenverhältnis zwischen der alten und der neuen Position, zieht aus einem Zufallsgenerator eine Zahl zwischen 0 und 1 und geht nur dann weiter, wenn diese Zahl kleiner ist als das Höhenverhältnis. Dieser Roboter würde die Punkte der Landschaft proportional zu ihrer Höhe abschreiten. Meist läuft er auf den Hügeln herum, seltener an den Hängen, noch seltener in den Tälern. Ganz ähnlich lässt sich die „Landschaft“ der Bäume proportional zu ihrer Likelihood begehen. Ausgehend von einem Zufallsbaum verändert das Computerprogramm bei jedem Schritt der „Markov-Kette“ einen Parameter der Phylogenie. Verbessert sich die Likelihood, „geht“ das Computerprogramm auf den neuen Baum, verschlechtert sie sich, wird die Veränderung nur mit einer geringen Wahrscheinlichkeit akzeptiert. Eine geringe Anzahl der so „besuchten“ Bäumen (z. B. jeder 10. oder jeder 100.) wird gespeichert. Auf diese Weise bewegt sich das Programm durch den „Wahrscheinlichkeitsraum“ aller möglichen Phylogenien, wobei es Bäume proportional zu ihrer Likelihood sammelt. Dass man nicht jeden Baum speichert, hängt damit zusammen, dass man eigentlich eine Zufallsstichprobe sammeln möchte. Innerhalb der Markov Kette ist aber jeder Baum von seinen unmittelbaren Vorgängern abhängig. Diese Abhängigkeit wird umso geringer, je mehr Schritte zwischen den gesammelten Bäumen liegen. Die Markov-Kette sammelt also nicht wirklich sondern nur näherungsweise eine Zufallsstichprobe.

Die ersten Bäume, die das Programm sammelt, liegen wahrscheinlich weit von den Regionen hoher Likelihood entfernt. Deshalb verwirft man bei Bayesschen Analysen die ersten 10.000 bis 100.000 Schritte als sogenannte „Burn-in“ Phase. Außerdem kann die Likelihood-Landschaft mehrere Hügel aufweisen. Eine einzige Markov-Kette könnte sich auf einem suboptimalen „Nebenhügel“ festsetzen, von dem sie auch in vielen Schritten nicht mehr herunterkommt. Deshalb lässt man meist mehrere Markov-Ketten gleichzeitig laufen, die von verschiedenen Zufallsbäumen ausgehen. Das erhöht die Chance, dass die Analyse im optimalen Bereich der Likelihood-Oberfläche konvergiert.

Die Bayessche Methode ist aus mehreren Gründen sehr populär geworden.

- Obwohl hunderttausende oder Millionen von Bäumen berechnet werden, ist die Methode extrem schnell.
- Alle Methoden, die nur nach einem Baum suchen, versagen unter bestimmten Bedingungen. Ob diese Bedingungen gegeben sind, kann man nicht immer wissen. Deshalb bleibt eine gewisse Unsicherheit, ob der gefundene optimale Baum auch wirklich der wahre Baum ist.
- Bootstrap-Analysen unter ML sind extrem zeitaufwändig. Die Bayessche Analyse berechnet den Baum und die Wahrscheinlichkeit einzelner Parameter in einem Arbeitsgang.
- Damit lässt sich auch die Wahrscheinlichkeit bestimmter Baumtopologien leicht vergleichen.

1. Kopieren Sie den NEXUS-file in den Ordner, in dem sich MrBayes befindet.
2. Öffnen Sie den NEXUS-file mit einem Text-Editor oder in PAUP (stoppen Sie dann die Berechnung und gehen Sie in die Ansicht des NEXUS-files).
3. Löschen Sie den Text zwischen „#NEXUS“ und „begin data;“. Die ersten 3 Zeilen der Datei lauten jetzt:

```
#NEXUS
```

```
begin data;
```

4. Ergänzen Sie in der 2. Zeile des Datenblocks „interleaved“ mit „=yes“.
5. Gehen Sie zum Ende der Datei und ersetzen Sie die Zeile „begin paup;“ durch „begin mrbayes;“
6. Im MrBayes-Block werden genau wie im PAUP-Block die Parameter der Analyse festgelegt. Das Substitutionsmodell muss ebenfalls festgelegt werden, allerdings bleiben die einzelnen Parameter variabel. Wir übernehmen das mit Modeltest errechnete Modell, lassen aber Angaben zu Nukleotidfrequenz, Gamma-shape Parameter α oder dem Anteil invariabler Sites weg. Diese Parameter des Modells werden im Laufe der Markov-Kette variiert.

```
set autoclose=yes;
```

```
lset nst=[Wert] rates=[equal/gamma/invgamma];
```

```
mcmc ngen=100000 printfreq=100 samplefreq=100 nchains=4
```

```
savebrlens=yes;
```

```
END;
```

Die erste Zeile bestimmt, dass das Programm die Analyse nach dem letzten Baum abschließt. Unter „lset“ wird wieder das Modell festgelegt (ohne, dass die Parameter fixiert werden). „Nst“ ist die Zahl der verschiedenen Substitutionsraten [Werte 1, 2 oder 6] „rates“ legt fest, ob

die Substitutionsratekonstant sein soll oder einer Gamma-Verteilung mit oder ohne invariable Sites folgt. Unter „mcmc“ wird der Suchlauf näher bestimmt. „Ngen“ = Zahl der Generationen (= Schritte) der Markov-Kette, „printfreq=100“ = die Likelihood-Werte jedes 100. Schritts der MARKOV-Ketten werden auf dem Bildschirm angezeigt; „samplefreq=100“ = jeder 100. Baum wird gesammelt; „nchains=4“ = vier parallel laufenden Markov-Ketten.

7. Speichern Sie den NEXUS-file im Ordner „MrBayes“ auf dem Desktop.
8. Öffnen Sie den Ordner „MrBayes“ und starten Sie „MrBayes3_0b3.exe“ durch Doppelklick.
9. Eingabe: `execute [Dateiname]`
Das Programm beginnt mit der Analyse. Auf dem Bildschirm wird jeder 100. Schritt angezeigt. In der rechten Spalte sehen Sie, wie viele Sekunden das Programm voraussichtlich noch rechnen wird.
10. Nachdem das Programm die Berechnung beendet hat, kann man die Ergebnisse anzeigen lassen. Als Ergebnis würden wir uns jetzt gerne den Konsensus-Baum mit Astlängen und den A-posteriori-Wahrscheinlichkeiten der einzelnen Clades ansehen. Wir müssen bei der Anzeige der Ergebnisse aber die „Burnin“-Phase berücksichtigen und

die ersten gesammelten Bäume verwerfen. Dazu lassen wir uns zuerst die Wahrscheinlichkeiten aller Parameter des Modells (Mittelwerte, Varianzen und 95 %-Vertrauensintervalle) und einen Graph der Likelihood-Werte anzeigen.

Eingabe: `sumt filename=[Dateiname.p]` Anhand des Graphen können wir abschätzen, nach wievielen Generationen der Markov-Kette die Likelihood-Werte ein stabiles Plateau erreicht haben. Die entsprechende Anzahl gesammelter Bäume (nicht Generationen!) ziehen wir als Burnin-Phase nicht in Betracht. Je nach Substitutionsmodell gibt das Programm die wahrscheinlichsten Werte für folgenden Parameter an: Baumlänge TL, verschiedene Substitutionsraten (r...), Nukleotidfrequenzen (pi...), Gamma Shape-Parameter (alpha), Anteil invariabler Positionen (pinvar).

11. Zur Auswertung des Baumes geben wir jetzt ein:

```
sumt filename=[Dateiname.t] burnin=[Zahl zu verwerfender Bäume]
```

Als output erhalten wir zuletzt eine Liste der Arten und der Häufigkeit verschiedener Gruppierungen, einen Baum mit A-posteriori-Wahrscheinlichkeiten, einen Baum mit Astlängen und eine Tabelle mit „Credible sets of trees“. Alle diese Daten werden in neuen Dateien namens [Dateiname].parts (Gruppierungen), [Dateiname].con (Bäume) und [Dateiname].trprobs gespeichert.

12. Die Datei [Dateiname].con lässt sich in Treeview zur besseren Ansicht öffnen.

13. Falls Sie nach der Wahrscheinlichkeit eines bestimmten, nicht im Konsensus-Baum enthaltenen Clades suchen, können Sie [Dateiname].parts mit einem Texteditor öffnen.

14. Zuletzt sehen wir uns an, was unter einem „Credible set of trees“ verstanden wird. Kein einzelner Baum wird normalerweise eine statistisch signifikante Wahrscheinlichkeit besitzen. In der Datei [Dateiname].trprobs finden Sie die Einzelwahrscheinlichkeiten (p) und kumulativen Wahrscheinlichkeiten (P) aller errechneten Bäume. Anhand der kumulativen Wahrscheinlichkeit können Sie die Bäume herausuchen, die gemeinsam eine signifikante Wahrscheinlichkeit besitzen. Den Schwellenwert der Signifikanz können Sie selber festlegen; in der Statistik üblich sind 95 % 99 % oder 99,9 %. Statt aufgrund eines Konsensus-Baums können Sie Ihre Daten auch auf der Grundlage eines solchen Sets möglicher Bäume durchführen.

Schlussbemerkungen

Schon mit dieser kleinen Auswahl an Methoden zur Rekonstruktion von Phylogenieen stößt man bei der Auswertung auf gewisse Probleme:

1. Wir wissen in der Regel nur sehr wenig über die molekulare Evolution der Gruppen, an denen wir arbeiten.
2. Über die statistischen Eigenschaften der verschiedenen Rekonstruktionsmethoden weiß man auch noch viel zu wenig.

Bei der Auswahl der Methode spielen praktische Aspekte eine große Rolle. Mit ML kann man neben dem Baum auch zahlreiche Parameter des Substitutionsmodells errechnen, aber die Berechnungen lassen sich bei großen Datensätzen nur noch auf Supercomputern durchführen. Hier schaffen Bayesische Methoden vielleicht Abhilfe. MP liefert schnelle Ergebnisse, kommt aber bei Datensätzen mit vielen Homoplasien (also stark divergenten Sequenzen) ins Straucheln. Eine scheinbar salomonische Lösung wäre es, mehrere Methoden anzuwenden und die Ergebnisse miteinander zu vergleichen. Das darf aber nicht dazu führen, dass man sich am Ende die Ergebnisse aussucht, die einem am besten in den Kram passen. Außerdem darf man sich nicht verleiten lassen, Ergebnisse nur deswegen für zuverlässiger zu halten, weil sie sich mit verschiedenen Methoden ermitteln lassen. Alle Methoden arbeiten mit dem gleichen Datensatz und es wäre bedenklich, wenn sie dabei zu grob unterschiedlichen Ergebnissen kämen.

Simulationsstudien haben gezeigt, dass es sich nicht lohnt, mit exzessivem Aufwand nach dem besten Baum zu suchen, da dieser Baum oft „besser“ ist als die Wirklichkeit. Man sollte die Zeit besser nutzen, um die statistische Unterstützung der errechneten, vielleicht suboptimalen Bäume zu bestimmen.

Literatur

Goldman N, Anderson LP, Rodrigo G (2000) Likelihood-based tests of topologies in phylogenetics. *Systematic Biology* 49: 652-670.

Nei M, Kumar S, Takahashi K (1998) The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. *Proceedings of the National Academy of Sciences USA* 95: 12390-12397.

Suzuki Y, Glazko GV, Nei M (2002) Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proceedings of the National Academy of Sciences USA* 99:16138-16143.

von Haeseler, A. & Liebers, D. (2003) Molekulare Evolution. 128 pp. S. Fischer Verlag, Frankfurt.