Poisson-Approximationen für genetische Fingerabdrücke

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Mathematik
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Dirk Metzler
aus Bad Homburg

Frankfurt 1999
(D F 1)

vom Fachbereich Mathematik
der Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr. L. Führer
Gutachter: Prof. Dr. A. Wakolbinger
Prof. Dr. G. Kersting
Datum der Disputation: 4. Juni 1999

Inhaltsverzeichnis

Einleitung

0.1 Fragestellung

Genetische Fingerabdrücke spielen außer in der Forensik und der medizinischen Diagnostik in vielen Fachrichtungen der Biologie eine wichtige Rolle. In der Ökologie kann man zum Beispiel durch den Vergleich der genetischen Fingerabdrücke von Angehörigen einer Population Rückschlüsse auf die Entwicklung der Population ziehen, und in der Evolutionsforschung wird die Abstammungsgeschichte von Arten rekonstruiert, indem deren genetische Fingerabdrücke verglichen werden. Die RAPD-PCR ist eine sehr schnelle und kostengünstige Methode zur Herstellung eines genetischen Fingerabdrucks, der dann als Strichmuster auf einem sogenannten Gel vorliegt. Eine ihrer wichtigsten Anwendungen ist es, den Stammbaum von Individuen aufgrund von Ähnlichkeiten ihrer RAPD-Fingerabdrücke zu schätzen (vgl. Abb. 1). Zur Klärung der Frage, für welchen Stammbaum RAPD-Daten am ehesten sprechen und wie zuverlässig eine derartige Aussage ist, bedarf es eines mathematischen Modells für die gemeinsame Verteilung von RAPD-Fingerabdrücken verwandter Individuen. Da der RAPD-Fingerabdruck jedes Individuums vom Vorkommen bestimmter Muster auf seiner DNA-Sequenz abhängt (vgl. Abb. 2, Abb. 3 sowie Anhang A), ergibt sich ein solches Modell aus einem stochastischen Modell für die DNA-Sequenzevolution entlang der Abstammungslinie des Stammbaums. Das einfachste derartige Evolutionsmodell ist das Jukes-Cantor-Modell (vgl. Abb. 4). Dieses Modell führt aber

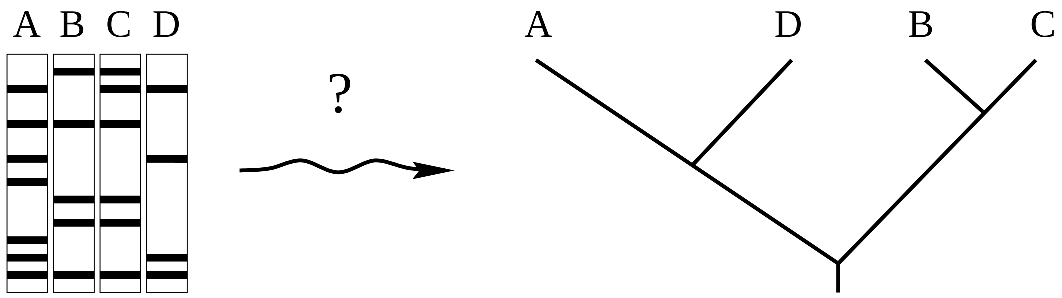


Abbildung 1: Aus den Ähnlichkeiten zwischen den RAPD-Fingerabdrücken können Rückschlüsse auf die Verwandtschaft gezogen werden. Es stellt sich jedoch die Frage, wie zuverlässig solche Schätzungen sind.

bereits zu komplizierten Abhängigkeiten innerhalb der RAPD-Fingerabdrücke (vgl. Abschnitt 0.2). Statistische Analysen von RAPD-Fingerabdrücken, die auf dem Jukes-Cantor-Modell in seiner vollen Feinheit beruhen, sind deshalb praktisch nicht durchführbar.

Gesucht ist daher:

- Ein Modell für die gemeinsame Verteilung der RAPD-Fingerabdrücke verwandter DNA-Stränge, innerhalb dessen statistische Analysen möglich sind und für das sich beweisen läßt, daß der Totalvariationsabstand zur Verteilung, die sich aus dem Jukes-Cantor-Modell ergibt, hinreichend klein ist.

Erst auf der Basis eines solchen Modells läßt sich dann die folgende Frage behandeln:

- Welche Irrtumswahrscheinlichkeiten treten bei gängigen Verfahren zur Schätzung von Stammbaumtopologien auf der Basis von RAPD-Daten auf?

0.2 Abhängigkeiten zwischen den RAPD-Banden im Jukes-Cantor-Modell

Technische Details zur RAPD-PCR können in Anhang A nachgelesen werden. Uns genügt folgende modellhafte Beschreibung: Unter einer *DNA-Sequenz* verstehen wir eine Folge der endlichen Länge n_{dna} über dem Alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Die Elemente dieses Alphabets heißen *Basen.* $\mathcal{P}$ und $\mathcal{K}$ seien Folgen der Länge l über $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Typische Werte sind $n_{\mathrm{dna}} = 3 \cdot 10^9$ und $l = 10$. Wir bezeichnen $\mathcal{P}$ und $\mathcal{K}$ im folgenden auch als *Primer* und *(Primer-)Komplement.* Außer $\mathcal{P}$ und $\mathcal{K}$ sei eine natürliche Zahl n_{amp}, der *Amplifikationsbereich* vorgegeben. (Ein typischer Wert wäre etwa $n_{\mathrm{amp}} = 3000$.) Eine *(RAPD-)Bande* ist ein Paar natürlicher Zahlen (s, k), deren Differenz $k - s$ im Intervall $[l, n_{\mathrm{amp}}]$ liegt. Die Differenz $k - s$ nennen wir *Länge* der Bande (s, k). Wir sagen von einer Bande (s, k), sie *liege auf einer DNA-Sequenz d vor*, wenn

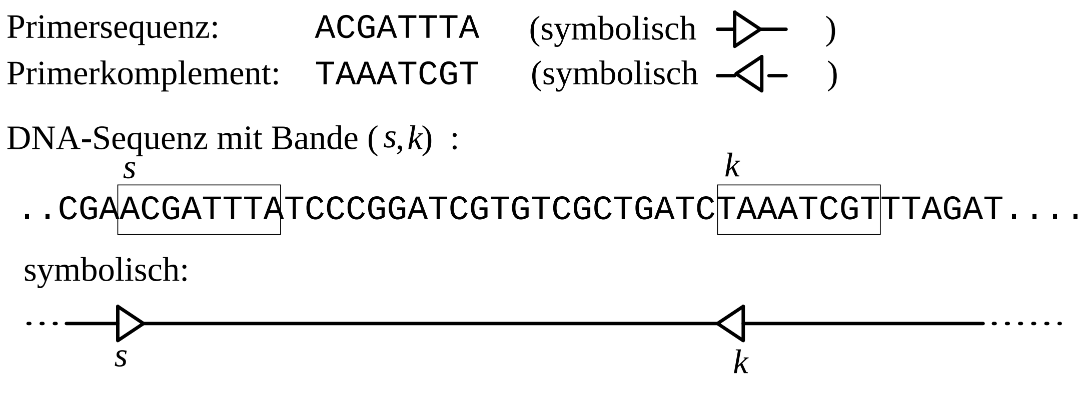


Abbildung 2: Wenn auf der DNA-Sequenz an Position s eine Kopie und an Position k ein Komplement des Primers beginnt, so liegt die Bande (s, k) vor. (Dabei muß $l \leq k - s \leq n_{\mathrm{amp}}$ gelten.)

auf d an der Stelle s das Muster $\mathcal{P}$ und an der Stelle k das Muster $\mathcal{K}$ beginnt (vgl. Abbildung 2). Falls auf d außerdem zwischen s und k weder ein Muster $\mathcal{P}$ noch ein Muster $\mathcal{K}$ beginnt, so heißt die Bande (s, k) *sichtbar auf* d. Oft werden im Labor sogenannte elektrophoretische Methoden verwendet, bei denen jede sichtbare Bande auf einem Gel an einer Position, die von der Länge der Bande abhängt, ein *Signal* hervorruft. Es ist dann nur eine bestimmte Anzahl n_{gp} von Gelpositionen unterscheidbar (z. B. $n_{gp} = 100$) (vgl. Abbildung 3). Wir gehen davon aus, daß für $i \in \{l, ..., n_{gp}\}$ jeweils alle Banden, deren Längen im Intervall $[l + (i-1)\frac{n_{\text{amp}}-l}{n_{gp}-1}, l + i\frac{n_{\text{amp}}-l}{n_{gp}-1}[$ liegen, auf dieselbe Gelposition abgebildet werden. Der *RAPD-Fingerabdruck* einer DNA-Sequenz (bzgl. $\mathcal{P}$ und $\mathcal{K}$) ist dann die Gesamtheit aller Gelpositionen, an denen ein Bandensignal vorliegt.

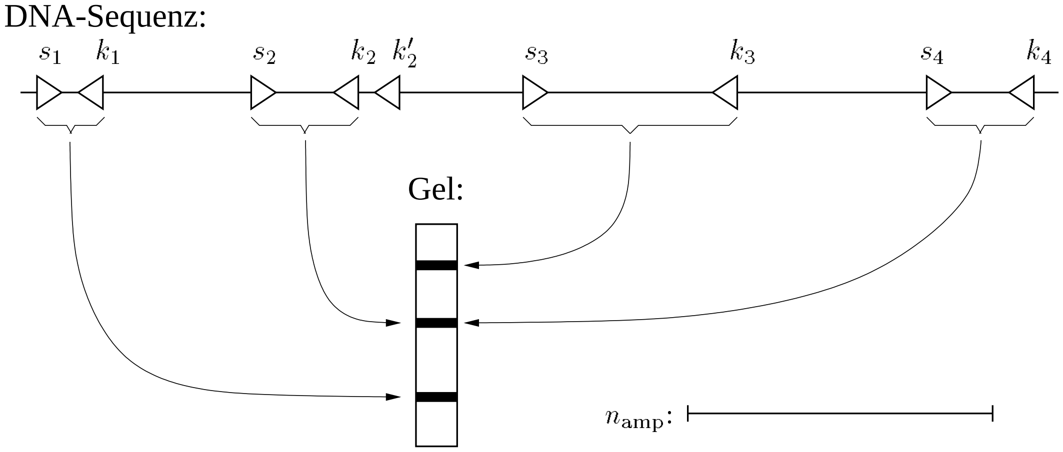


Abbildung 3: Schematische Darstellung einer DNA-Sequenz mit Primerkopien ($\triangleright$) und Primerkomplementen ($\triangleleft$), sowie dem sich daraus ergebenden RAPD-Fingerabdruck auf dem Gel. Die Banden (s_1, k_1), (s_2, k_2), (s_3, k_3) und (s_4, k_4) sind sichtbar, (s_2, k_2) und (s_4, k_4) sind aber auf dem Gel nicht zu unterscheiden, da sie dieselbe Länge haben. Die Bande (s_2, k_2') ist nicht sichtbar, da zwischen den Positionen s_2 und k_2' ein Primerkomplement liegt.

Wir betrachten nun das Jukes-Cantor-Modell (vgl. Jukes und Cantor (1969)) für die DNA-Sequenzevolution längs eines Stammbaumes. Wir ergänzen dabei den Baum durch eine Kante, die von der Wurzel unendlich weit in die Vergangenheit verläuft. Wir betrachten den Stammbaum als geometrisches Gebilde mit Längenmaß λ. Für jede Position $s < n_{\text{dna}}$ gibt es einen Poisson'schen Punkt-Prozeß auf dem Baum mit Intensität λ. Die Punkt-Prozesse zu verschiedenen *Sites* (d. h. Positionen) sind voneinander unabhängig. Wir nennen die Poisson'schen Punkte im folgenden *Mutationen*. Jeder Mutation wird unabhängig gleichverteilt eine Base aus $\{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$ zugeordnet. Um die Base zu ermitteln, die sich bei einem Blatt des Baumes an einer bestimmten Position befindet, gehe man in negativer Zeitrichtung bis zur nächsten Mutation zurück, die es an der betreffenden Position gegeben hat, und übernehme deren Base (siehe Abbildung 4).

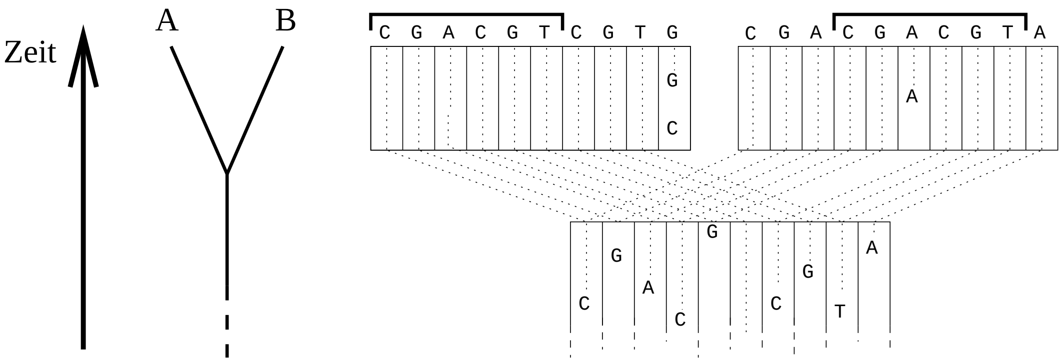


Abbildung 4: Die DNA von A ist `CGACGTCGTG`*, die von B ist* `CGACGACGTA`*. Die Mutationen sind im Jukes-Cantor-Modell Poisson'sch in die Kanten des Baumes eingestreut. Das Muster* `CGACGT` *beginnt in der DNA-Sequenz von A an erster Position und in der DNA-Sequenz von B an vierter Position. Der Unterschied kommt durch eine einzige Mutation zustande. Dies ist möglich, da die Teilfolge* `CG` *in dem Muster doppelt vorkommt.*

Wie schon im vorigen Abschnitt erwähnt, ergeben sich auf der Basis des Jukes-Cantor-Modells komplizierte stochastische Abhängigkeiten zwischen den RAPD-Banden verwandter DNA-Sequenzen. Schon bei einer einzelnen rein zufälligen Sequenz sind die Ereignisse des Auftretens von vorgegebenen Mustern an einer bestimmten Position nicht unabhängig. Die Wahrscheinlichkeit, daß etwa an einer Position s das Muster $M = \texttt{ATAT}$ auftritt, ist $1/256$. Gegeben, daß M an der Stelle s beginnt, ist die bedingte Wahrscheinlichkeit, daß M auch an der Stelle $s+1$ bzw. $s+2$ beginnt, 0 bzw. $1/16$. Dabei kommt es offensichtlich auf die möglichen Überlappungsweiten der Muster an. Wie das Beispiel in Abbildung 4 zeigt, sind für die gemeinsame Verteilung von Mustern auf verwandten DNA-Sequenzen auch solche Überlappungsweiten von Bedeutung, die erst durch zusätzliche Mutationen möglich werden. Wie stark diese Abhängigkeiten sind, hängt also in komplizierter Weise von Ähnlichkeiten der relevanten Muster ab, im Falle der RAPD-Banden also von den Primer-Sequenzen und -Komplementen.

Eine weitere Klasse stochastischer Abhängigkeiten in den RAPD-Banden-Konfigurationen verwandter DNA-Sequenzen kommt dadurch zustande, daß Primerkopien oder Primerkomplemente bei zwei verschiedenen Banden beteiligt sein können. Diese können dann bei verschiedenen Blättern des Baumes sichtbar werden (vgl. Abbildung 5).

0.3 Vorgehensweise

In **Kapitel 1** befassen wir uns mit Überlappungseffekten zwischen Mustern auf verwandten DNA-Strängen. Wir stellen dem Jukes-Cantor-Modell ein Modell für die Evolution der Banden gegenüber, in dem Muster-Überlappungseffekte vernachlässigt werden. Dazu betrachten wir den Raum aller zusammenhängenden l-Tupel von Basenpositionen, die einem Punkt auf dem Stammbaum zugeordnet sind. Die Mutationen des Jukes-Cantor-Modells bilden einen Poisson-

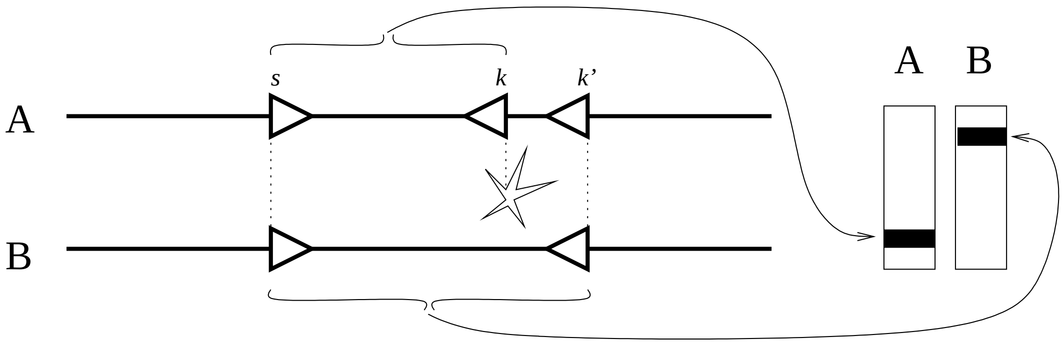


Abbildung 5: Bei A ist nur die Bande (s, k) sichtbar, bei B nur die Bande (s, k'). Die RAPD-Fingerabdrücke von A und B unterscheiden sich an zwei Gelpositionen, obwohl nur eine Mutation (durch den Stern symbolisiert) zwischen den Sequenzen liegt.

Cluster-Prozeß auf diesem Raum. Die Überlappungseffekte zu vernachlässigen bedeutet, diesen Poisson-Cluster-Prozeß durch einen Poisson-Prozeß zu approximieren. Wir konstruieren einen Markoff-Prozeß, der in dem Poisson-Cluster-Prozeß startet und für den die Verteilung des Poisson-Prozesses eine Gleichgewichtsverteilung ist. Verteilungseigenschaften dieses Markoff-Prozesses liefern eine obere Schranke für den Totalvariationsabstand der Verteilungen von Bandenkonfigurationen, die diese beiden Modelle implizieren (**Satz 1.2**, Seite 16). Mit diesem Resultat kann man für anwendungsrelevante Szenarien zeigen, daß die Überlappungseffekte in Hinblick auf die Verteilung der RAPD-Fingerabdrücke vernachlässigbar sind.

In **Kapitel 2** geht es um stochastische Abhängigkeiten der RAPD-Banden verwandter Individuen, die zum Beispiel dadurch ins Spiel kommen, daß eine Primerkopie, die an einem bestimmten Site vorkommt, bei zwei verschiedenen Individuen an unterschiedlichen Banden beteiligt sein kann. Wir vergleichen dazu zunächst zwei Modelle: In einem Feinmodell evolvieren alle Muster gemäß dem Modell ohne Muster-Überlappungseffekte aus Kapitel 1. Wenn dann bei einem Individuum eine Primerkopie und ein Primerkomplement in einem bestimmten Abstand aufeinanderfolgen, liegt eine Bande vor. Dem stellen wir zunächst ein sehr einfaches Modell gegenüber, bei dem alle Banden unabhängig evolvieren.

Wie sich zeigen wird, ist der Totalvariationsabstand zwischen den beiden resultierenden Verteilungen der Bandenkonfigurationen *nicht* hinreichend klein. Die Abhängigkeiten zwischen den Banden lassen sich also nicht generell vernachlässigen. Wir werden daher ein drittes Modell für die Konfiguration der Banden konstruieren, bei dem wesentliche Abhängigkeiten berücksichtigt werden. Komplizierte Abhängigkeiten höherer Ordnung werden jedoch vermieden. Dem Modell liegt die Idee der Poisson-Clumping-Heuristik zugrunde: Die Banden treten in Klumpen auf, wobei zwei Banden, die eine Primerkopie oder ein Primerkomplement gemeinsam haben, im selben Klumpen liegen, und die Konfiguration der Klumpen läßt sich durch einen Poisson-Prozeß approximieren. Mit Hilfe einer Variante der Chen-Stein-Methode wird es uns gelingen, den Totalvariationsabstand zwischen den Verteilungen der Bandenkonfigurationen, die aus dem Feinmodell und dem Klumpen-Modell resultieren, zu kontrollieren (**Satz 2.1**, Seite 45). Zusammen mit den Ergebnissen aus Kapitel 1 und der Dreiecksungleichung erhalten wir dann auch eine obere Abschätzung für den Totalvariationsabstand zwischen den Verteilungen der Bandenkonfigurationen, die aus dem Klumpen-Modell und dem Jukes-Cantor-Modell resultieren. Das Klumpen-Modell bietet einerseits eine gute Approximation der Verteilung der Bandenkonfigurationen des Jukes-Cantor-Modells und eignet sich andererseits für sehr effiziente Monte-Carlo-Simulationen. Erst dadurch werden Simulationsstudien zur Beurteilung von Auswertungsverfahren für RAPD-Daten möglich.

In **Kapitel 3** werden wir diese Simulationsmöglichkeiten einsetzen, um Stammbaum-Topologie-Schätzer auf der Basis von RAPD-Fingerabdrücken zu untersuchen. Speziell betrachten wir dabei den Fall von vier Individuen. (Diese bilden die Bausteine für die Stammbaumschätzung auch einer größeren Anzahl von Individuen, vgl. Strimmer und von Haeseler (1996).)

Für die Stammbaumrekonstruktion aus RAPD-Daten kommen verschiedene Verfahren in Betracht. Wir vergleichen einen parsimonischen Ansatz mit solchen, die auf der Maximum-Likelihood-Idee beruhen. Die Berechnung der Likelihood einer Baumtopologie für gegebene RAPD-Daten ist zwar auf der Basis des Klumpen-Modells möglich, aber wegen der zweibandigen Klumpen noch immer sehr rechenintensiv, insbesondere wenn auch die Verwechslung von Banden zu berücksichtigen ist. Wir konstruieren daher Maximum-Likelihood-Schätzer für die Baumtopologien auf der Basis weiter vereinfachter Ersatzmodelle und untersuchen die Irrtumswahrscheinlichkeiten dieser Schätzer bei Anwendung auf RAPD-Daten, die wir nach dem Klumpen-Modell simulieren. Insbesondere werden wir diskutieren, wie weit Ansätze tragen, die von einem gedächtnislosen Entstehen und Vergehen der Banden längs der Abstammungslinien ausgehen.

Danksagung

An erster Stelle möchte ich Professor Anton Wakolbinger für die äußerst engagierte Betreuung herzlich danken. Er war jederzeit zu fachlichen Diskussionen bereit und hat mir viele wichtige Anregungen und Hinweise gegeben.

Auch Herrn Dr. Brooks Ferebee verdanke ich viel. Seine kritischen Fragen waren oft richtungsweisend.

Der Graduiertenförderung des Landes Hessen danke ich für die finanzielle Unterstützung von August 1995 bis Juli 1997.

Für inspirierende Gespräche danke ich den Professoren Götz Kersting, Hermann Dinges, Arndt von Haeseler und Bernd Schierwater, sowie meinen Kolleginnen und Kollegen Andrea Ender, Johannes Lenhard, Jochen Geiger, Steffen Grossmann, Matthias Birkner und – auch für die seelisch-moralische Unterstützung – Gudrun Back.

Kapitel 1

Poisson-Approximation für die Evolution der Muster

Wir befassen uns in diesem Kapitel mit solchen Abhängigkeiten zwischen den Konfigurationen von Mustern auf DNA-Sequenzen, die durch Überlappungseffekte zustande kommen (vgl. Abschnitt 0.2). Bedingt man etwa im Kontext der RAPD-PCR darauf, daß auf einer der DNA-Sequenzen an einer Position s eine Primerkopie der Länge l beginnt, so ändern sich im allgemeinen die Wahrscheinlichkeiten für das Auftreten von Primerkopien oder -komplementen an den Positionen $s-l+1$ bis $s+l-1$ auf allen DNA-Sequenzen (vgl. Abbildung 4 der Einleitung). Dabei kommt es darauf an, wie die Primerkopien und -komplemente zusammengesetzt sind und wie die DNA-Sequenzen miteinander verwandt sind.

Nach Belieben kann zum Einstieg zunächst Anhang C gelesen werden, wo wir uns mit Überlappungseffekten von Mustern auf einer einzelnen Sequenz befassen.

1.1 Zwei Modelle für die Evolution der Muster

1.1.1 Allgemeines

Es sei ein Stammbaum $\mathbf{T}$ von endlich vielen Individuen gegeben, also ein verwurzelter Binärbaum, dessen Blätter mit den DNA-Sequenzen der Individuen beschriftet sind. Wir betrachten den Baum als ein geometrisches Gebilde mit Längenmaß λ. (Die Länge $\lambda([v_1, v_2])$ einer Kante K mit Endknoten v_1 und v_2 gibt im biologischen Kontext die zeitliche Differenz zwischen v_1 und v_2 an.) Wir erweitern den Stammbaum um eine unendlich lange Kante, die zur Wurzel führt und deren Ahnen beheimatet. Die Menge aller Punkte, die auf einer der Kanten des Baumes liegen (einschließlich aller Knoten und insbesondere aller Blätter von $\mathbf{T}$), bezeichnen wir im Folgenden mit $V_{\mathbf{T}}$, die Menge der Blätter von $\mathbf{T}$ mit $B_{\mathbf{T}}$.

1.1.2 Die Evolution der Muster im Jukes-Cantor-Modell

Im Jukes-Cantor-Modell (vgl. Abschnitt 0.2) bilden die Mutationen an jeder Position $i \in \{1, \dots, n_{\text{dna}}\}$ einen Poisson'schen Punkt-Prozeß auf $V_{\mathbf{T}}$ zur Intensität λ. Die Poisson-Prozesse sind voneinander unabhängig, und jeder Mutation wird unabhängig gleichverteilt eine Base aus $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ zugeordnet.

Für alle $i \in \{1, \dots, n_{\text{dna}}\}$ sei θ_i ein Poisson-Prozeß auf $V_{\mathbf{T}} \times \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ mit Intensität $\mu(U \times \{\mathtt{B}\}) := \frac{1}{4}\lambda(U)$ und für $j \in \{1, \dots, l\}$ mit $i+j-1 \leq n_{\text{dna}}$ sei $\Theta_{ij} := \theta_{i+j-1}$. Wir definieren außerdem einen Poisson-Cluster-Prozeß auf $\Gamma := V_{\mathbf{T}} \times \{1, \dots, n_{\text{dna}} - l + 1\} \times \{1, \dots, l\} \times \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ durch $\Theta := \sum_i \sum_{j=1}^{l} \delta_{(i,j)} \otimes \Theta_{ij}$ (mit der offensichtlichen Permutation der vier Komponenten von Γ).

Wir definieren dann $W_{ij}(v)$, indem wir jeweils zum jüngsten Vorfahren v' von v, für den es ein $\mathtt{B} \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ mit $\Theta_{ij}(v', \mathtt{B}) = 1$ gibt, zurückgehen und $W_{ij}(v) := \mathtt{B}$ setzen (siehe Abbildung 1.1). Für $i \in \{1, \dots, n_{\text{dna}} - l + 1\}$ ist $W_i(v) := (W_{i1}(v), \dots, W_{il}(v))$ dann das Basenwort, das in der DNA-Sequenz von v an Position i beginnt.

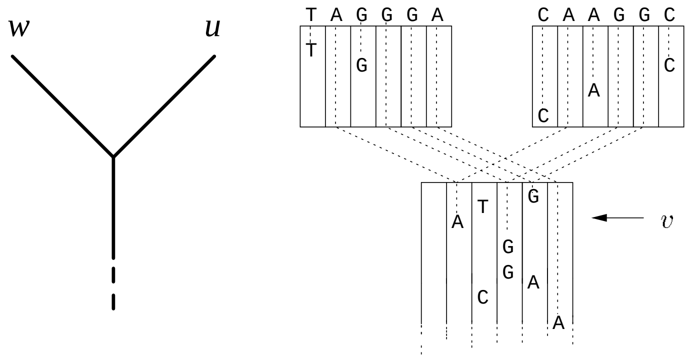


Abbildung 1.1: Die DNA von w ist `TAGGGA`*, die von u ist* `CAAGGC`*. Bei $l = 3$ gilt z. B. $W_2(v) = \mathtt{ACG}$ und $W_3(v) = \mathtt{CGA}$.*

Für jedes $i \in \{1, \dots, n - l + 1\}$ ist W_i eine mit dem Baum $\mathbf{T}$ indizierte Markoff-Kette auf $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ mit sehr einfachen Eigenschaften: Sie folgt der Jukes-Cantor-Dynamik. Der zu W_i gehörige „Mutationsprozeß" $\Theta_i := \sum_{j=1}^{l} \delta_j \otimes \Theta_{ij}$ ist ein Poisson-Prozeß auf $\{1, \dots, l\} \times V_{\mathbf{T}} \times \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Die Verteilung von $W = (W_i)_i$ ist aber recht kompliziert, da Θ_i und Θ_{i+k}, und damit auch W_i und W_{i+k}, für $|k| < l$ stochastisch abhängig sind. Es gilt nämlich $\Theta_{(i+k),j} = \Theta_{i,(j+k)}$ und $W_{(i+k),j} = W_{i,(j+k)}$.

1.1.3 Ein Modell mit unabhängig evolvierenden Mustern

Wir definieren nun Analoga $\widetilde{\Theta}$ und $\widetilde{W}$ zu Θ und W, bei denen die am Ende von 1.1.2 beschriebenen Abhängigkeiten nicht vorhanden sind. $\widetilde{\Theta}$ sei also ein Poisson-Prozeß auf Γ, dessen

Intensität μ gegeben ist durch:

$$\mu(U \times \{i\} \times \{j\} \times \{\mathtt{B}\}) := \frac{1}{4}\lambda(U) \qquad \text{für alle meßbaren } U \subset V_{\mathbf{T}} \text{ und alle } i, j, \mathtt{B}.$$

$(\widetilde{\Theta}_i)_i$ (für alle in Abschnitt 1.1.2 erlaubten i) ist demnach eine Folge von unabhängigen Poisson-Prozessen, so daß $\mathcal{L}(\Theta_i) = \mathcal{L}(\widetilde{\Theta}_i)$ für jedes i gilt.

Die Konstruktion von $\widetilde{W}$ aus $\widetilde{\Theta}$ erfolgt analog zur Konstruktion von W aus Θ. Ebenso wie W_i können wir $\widetilde{W}_i$ als baumindizierte Markoff-Kette mit Werten in $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}^l$ auffassen. Im Unterschied zu $W := (W_i)_{1\le i\le n-l+1}$ ist $\widetilde{W} := (\widetilde{W}_i)_{1\le i\le n-l+1}$ allerdings eine Folge stochastisch unabhängiger Prozesse. Daher läßt sich $\widetilde{W}$ mathematisch wesentlich leichter handhaben als W.

1.2 Exkurs über Poisson-Approximationen und den Totalvariationsabstand

Wir stellen uns in diesem Kapitel die Frage, wie gut sich der Poisson-Cluster-Prozeß Θ auf der Menge Γ durch den Poisson-Prozeß $\widetilde{\Theta}$ in Hinblick auf die genetischen Fingerabdrücke der Blätter des Stammbaums approximieren läßt. Auch in Kapitel 2 werden wir die Vernachlässigbarkeit bestimmter Abhängigkeiten nachweisen, indem wir eine Verteilung durch die Verteilung eines Poisson-Prozesses approximieren. Wir geben uns dabei nicht mit asymptotischen Aussagen zufrieden, sondern suchen nach oberen Abschätzungen für den Totalvariationsabstand zwischen den jeweiligen Verteilungen.

Der Totalvariationsabstand zweier Maße μ_1 und μ_2 auf einem meßbaren Raum $(\mathbf{X}, \mathcal{A})$ ist definiert durch:

$$d_{TV}(\mu_1, \mu_2) := \sup_{A\in\mathcal{A}} |\mu_1(A) - \mu_2(A)| = \sup_{f:\mathbf{X}\to[0,1]} \left| \int f d\mu_1 - \int f d\mu_2 \right|$$

Handelt es sich bei $\mathbf{X}$ um einen separablen metrischen Raum und sind μ_1 und μ_2 Wahrscheinlichkeitsverteilungen, so gilt

$$d_{TV}(\mu_1, \mu_2) = \min \Pr(X_1 \neq X_2),$$

wobei über Paare von Zufallsvariablen (X_1, X_2) minimiert wird, für die X_1 und X_2 die Wahrscheinlichkeitsmaße μ_1 und μ_2 haben (siehe Barbour *et al.* (1992)). Ein derartiges Paar (X_1, X_2) heißt auch *Kopplung* von μ_1 und μ_2; eine Kopplung, für die der Totalvarationsabstand minimal wird, heißt *maximale Kopplung*.

Mit der Chen-Stein-Methode (vgl. Anhang C) und den zu ihr verwandten Techniken, die in diesem Kapitel zum Einsatz kommen, erhalten wir Abschätzungen des Totalvariationsabstandes zwischen gewissen Verteilungen und ihren Poisson-Approximationen. Nutzt uns das etwas bei der Analyse genetischer Fingerabdrücke?

Wir stellen uns folgende Situation vor: Es soll anhand der genetischen Fingerabdrücke einiger Blätter entschieden werden, ob eine bestimmte Hypothese über ihre Verwandtschaft zu verwerfen ist. Dies soll auf der Basis eines bestimmten Modells für die Entstehung der genetischen

Fingerabdrücke geschehen. Das Modell enthalte aber gewisse Abhängigkeiten, die beim Berechnen der Likelihood der Hypothese vernachlässigt werden, indem eine Poisson-Approximation verwendet werde. Angenommen, es gelingt zu zeigen, daß der Totalvariationsabstand zwischen der Verteilung, die sich aus dem Modell ergibt, und der Poisson-Approximation sehr gering ist; Kann man dann eine Aussage darüber machen, wie wahrscheinlich es ist, daß man sich bei der Entscheidung über die Hypothese aufgrund der vernachlässigten Abhängigkeiten irrt? – Man kann:

Wir gehen z. B. davon aus, daß die Hypothese zutrifft, und betrachten die Wahrscheinlichkeit, daß man die Hypothese fälschlicherweise verwirft. Sei μ_1 das Wahrscheinlichkeitsmaß für die entstehende Gesamtheit an genetischen Fingerabdrücken auf der Basis des Modells mit Abhängigkeiten und μ_2 das Analogon für das Modell ohne Abhängigkeiten. Sei A die Menge der genetischen Fingerabdrücke, die auf der Basis des Modells ohne Abhängigkeiten zum Verwerfen der Hypothese führen. Nach der Definition des Totalvariationsabstandes gilt $|\mu_1(A) - \mu_2(A)| \leq d_{TV}(\mu_1, \mu_2)$. Die Tatsache, daß sich in den Daten die betreffenden Abhängigkeiten befinden, kann also gegenüber dem idealisierten Fall, in dem die Daten ohne die betreffenden Abhängigkeiten entstanden sind, höchstens zu einer Verschlechterung der Irrtumswahrscheinlichkeit um den Summanden $d_{TV}(\mu_1, \mu_2)$ führen.

Allerdings wäre in der Fragestellung dieses Kapitels $d_{TV}(\mathcal{L}(\Theta), \mathcal{L}(\widetilde{\Theta}))$ sicherlich eine zu feine Maßeinheit für den Unterschied zwischen den beiden Verteilungen. Man betrachte nämlich die Funktion f, die einer Konfiguration ξ von Elementen von Γ eine 1 zuordnet, falls ξ zwei Elemente $(v, i, j, \mathtt{B})$ und $(v', i', j', \mathtt{B}')$ mit $v = v'$ enthält, und sonst 0. Dann gilt fast sicher $f(\Theta) = 1$ und $f(\widetilde{\Theta}) = 0$. Also gilt $d_{TV}(\mathcal{L}(\Theta), \mathcal{L}(\widetilde{\Theta})) \geq \left|\mathbb{E}f(\Theta) - \mathbb{E}f(\widetilde{\Theta})\right| = 1$. In der Tat sind nur solche Unterschiede zwischen $\mathcal{L}(\Theta)$ und $\mathcal{L}(\widetilde{\Theta})$ relevant, die sich auf die genetischen Fingerabdrücke der Blätter des Stammbaumes auswirken. Also besteht unser Ziel darin, den Totalvariationsabstand der von $\mathcal{L}(\Theta)$ und $\mathcal{L}(\widetilde{\Theta})$ induzierten Verteilungen auf den genetischen Fingerabdrücken der Blätter abzuschätzen.

In Kapitel 2 werden wir es mit Familien von Ereignissen zu tun haben, bei denen lokale stochastische Abhängigkeiten vorliegen, die so stark sind, daß eine Poisson-Approximation in der oben skizzierten Art und Weise nicht sinnvoll ist. Um dennoch einige Abhängigkeiten vernachlässigen zu können, gehen wir dann folgendermaßen vor: Wir bezeichnen Unterfamilien von Ereignissen, zwischen denen starke Abhängigkeiten vorliegen und die alle eintreten, als *Klumpen*[1] und approximieren die Konfiguration der auftretenden Klumpen durch einen Poisson-Prozeß auf dem Raum der möglichen Klumpen. Derartige Prozesse heißen bei Aldous (1989) *Mosaik-Prozesse* und bei einigen anderen Autoren *Poisson-Cluster-Prozesse* (vgl. z. B. Karr (1986)). Aldous zeigt viele Möglichkeiten auf, Mosaik-Prozesse in verschiedenen stochastischen Kontexten zumindest heuristisch einzusetzen.

Unter einer *Poisson-Approximation* verstehen wir eine Approximation durch einen Poisson-

[1] Der Begriff „Klumpen" orientiert sich am Titel des Buches „Probability Approximations via the Poisson Clumping Heuristic" von D. Aldous (1989).

Prozeß oder einen Poisson-Cluster-Prozeß.

1.3 Abstand der beiden Modelle in Hinblick auf die genetischen Fingerabdrücke der Blätter

Nun gehen wir der Frage nach, inwieweit die Unterschiede zwischen Θ und $\widetilde{\Theta}$ hinsichtlich der genetischen Fingerabdrücke der Blätter vernachlässigbar sind.

Es sei $\mathcal{Z}$ die Familie der lokal endlichen, einfachen Zählmaße auf Γ. Θ und $\widetilde{\Theta}$ sind also $\mathcal{Z}$-wertige Zufallsvariablen.

Sei $\mathbf{W}$ der Raum der Abbildungen von $B_{\mathbf{T}} \times \{1, \ldots, n-l+1\}$ nach $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \mathtt{O}\}^l$. Wir definieren eine Abbildung $\Psi : \mathcal{Z} \to \mathbf{W}$: Es sei $\Psi(\xi)(v,i)$ das Wort $(w_1, \ldots, w_l)$, das wir erhalten, indem wir w_j folgendermaßen setzen: Falls unter allen Vorfahren u von v, für die es ein $\mathtt{B}_u \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ mit $\xi(v,i,j,\mathtt{B}_u) = 1$ gibt, ein jüngster u' existiert und das dazugehörige $\mathtt{B}_{u'}$ eindeutig bestimmt ist, so sei $w_j := \mathtt{B}_{u'}$. Sonst sei $w_j := \mathtt{O}$.

Damit gilt (fast sicher) $\Psi(\Theta) = W\rfloor_{B_{\mathbf{T}}}$ und $\Psi(\widetilde{\Theta}) = \widetilde{W}\rfloor_{B_{\mathbf{T}}}$, wobei $\rfloor$ für „eingeschränkt auf“ steht.

Allerdings gehen wir davon aus, daß nicht die Elemente von $\mathbf{W}$ beobachtbar sind, sondern nur ihr Bild unter einer „Vergröberung“ $F : \mathbf{W} \to F(\mathbf{W})$, und setzen $\Phi := F \circ \Psi : \mathcal{Z} \to F(\mathbf{W})$.

Beispiel: Im Falle der RAPD-PCR können nur solche Positionen auf den DNA-Sequenzen der Blätter einen Einfluß haben, bei denen die Primersequenz $\mathcal{P}$ oder ihr Komplement $\mathcal{K}$ steht und denen innerhalb von n_{amp} Positionen ein $\mathcal{K}$ folgt bzw. ein $\mathcal{P}$ vorangeht. Für Φ setzen wir also in diesem Kontext $\overline{\Psi}$ ein, welches für $\xi \in \mathcal{Z}$, $i \leq n$ und $b \in B_{\mathbf{T}}$ folgendermaßen definiert sei:

$$\overline{\Psi}(\xi)(b,i) := \begin{cases} \mathcal{P} & \text{falls} \quad \Psi(\xi)(b,i) = \mathcal{P} \\ & \text{und } \exists k \in \{l+1, \ldots, n_{\mathrm{amp}}\} : \Psi(\xi)(b, i+k) = \mathcal{K} \\ \mathcal{K} & \text{falls} \quad \Psi(\xi)(b,i) = \mathcal{K} \\ & \text{und } \exists k \in \{l+1, \ldots, n_{\mathrm{amp}}\} : \Psi(\xi)(b, i-k) = \mathcal{P} \\ \mathtt{OO}\ldots\mathtt{O} & \text{sonst} \end{cases}$$

Wir werden später auf $\overline{\Psi}$ zurückkommen (vgl. Satz 1.2).

Es ist in diesem Kapitel unser Ziel, für gewisse Φ den Totalvariationsabstand $d_{TV}(\mathcal{L}(\Phi(\Theta)), \mathcal{L}(\Phi(\widetilde{\Theta})))$ abzuschätzen. $\mathcal{F}_\Phi$ sei die Familie der meßbaren Abbildungen $f : \mathcal{Z} \to [0,1]$ der Form $f = g \circ \Phi$. Offenbar gilt:

$$d_\Phi(\mathcal{L}(\Theta), \mathcal{L}(\widetilde{\Theta})) := d_{TV}(\mathcal{L}(\Phi(\Theta)), \mathcal{L}(\Phi(\widetilde{\Theta}))) = \sup_{f \in \mathcal{F}_\Phi} \left| \mathbb{E}(f(\Theta)) - \mathbb{E}(f(\widetilde{\Theta})) \right|$$

Um eine obere Abschätzung dafür zu finden, verwenden wir eine ähnliche Kopplungsmethode wie Barbour, Holst und Janson (1992) zum Beweis von Theorem 10.B ihres Buches „Poisson

Approximation“. Wir definieren dazu einen zeitkontinuierlichen Markoff-Prozeß $(Z_t)_{t\geq 0}$ auf $\mathcal{Z}$ mit $Z_0 = \Theta$, für dessen Übergangsdynamik $\mathcal{L}(\widetilde{\Theta})$ eine Gleichgewichtsverteilung ist.

Wir konstruieren Z als Immigrations-Todesprozeß über Γ mit Immigrationsintensität μ und pro-Kopf-Sterberate 1. Der Generator von Z ist von der Form

$$(\mathcal{A}h)(\xi) = \int_\Gamma [h(\xi + \delta_\alpha) - h(\xi)] \; d\mu(\alpha) + \int_\Gamma [h(\xi - \delta_\alpha) - h(\xi)] \; d\xi(\alpha)$$

Anschaulich bedeutet das: *Jede Mutation stirbt mit Rate 1, und auf einem Kantenabschnitt der Länge dx kommen unabhängig bei jedem Muster an jeder Stelle des Musters Mutationen mit der Rate von jeweils dx hinzu.*

Z starte in $Z_0 := \Theta$. Für $\alpha = (v, i, j, \mathtt{B})$ ist $\Gamma_\alpha := \{\beta : \; \beta = (v, i+k, j-k, \mathtt{B})$ für geeignetes $k\}$ der Abhängigkeitsbereich von α. Wir setzen außerdem $\Gamma'_\alpha := \Gamma_\alpha \setminus \{\alpha\}$.

Für $\alpha \in \Gamma$, $\beta \in \Gamma'_\alpha$ und ein zufälliges $X \in \mathcal{Z}$ mit $\mathbb{E}X(\alpha) = \mathbb{E}X(\beta) = 0$ sei

$$\varphi_{\alpha\beta}(\mathcal{L}(X)) := \Pr(\Phi(X + \delta_\alpha) \neq \Phi(X) \neq \Phi(X + \delta_\beta)),$$

also die Wahrscheinlichkeit, daß sich sowohl durch die Hinzunahme von δ_α als auch durch die Hinzunahme von δ_β zur Mutationenkonfiguration X beobachtbare Veränderungen in den Blättern ergeben.

Ein $\alpha \in \Gamma$ heißt Φ-*unwirksam gegenüber* ξ (mit $\xi \in \mathcal{Z}$), falls $\Phi(\xi + (1 - 2\xi(\alpha)) \cdot \delta_\alpha) = \Phi(\xi)$ gilt. Φ heißt *Cluster-neutral*, falls für jedes Paar $(\alpha, \beta) \in \Gamma \times \Gamma'_\alpha$ gilt, daß α genau dann Φ-unwirksam gegenüber ξ ist, falls es Φ-unwirksam gegenüber $\xi + (1 - 2\xi(\beta)) \cdot \delta_\beta$ ist.

Satz 1.1 *Ist Φ Cluster-neutral, so gilt:*

$$d_\Phi(\mathcal{L}(\Theta), \mathcal{L}(\widetilde{\Theta})) \;\leq\; 2 \cdot \int_\Gamma \int_0^\infty \sum_{\beta \in \Gamma'_\alpha} \varphi_{\alpha\beta}(\mathcal{L}(Z_t)) \cdot e^{-2t} \; dt \; d\mu(\alpha)$$

Für unser Leitthema, die Frage nach der Vernachlässigbarkeit von Abhängigkeiten zwischen RAPD-Banden, ziehen wir aus Satz 1.1 folgendes Korollar, für das wir in Abschnitt 1.3.1 ein Anwendungsbeispiel diskutieren:

Satz 1.2 *Für $\alpha \in \Gamma$, $\beta \in \Gamma'_\alpha$ und eine $\mathcal{Z}$-wertige Zufallsvariable X mit $\mathbb{E}X(\alpha) = \mathbb{E}X(\beta) = 0$ sei*

$$\overline{\varphi}_{\alpha\beta}(\mathcal{L}(X)) := \Pr(\overline{\Psi}(X + \delta_\alpha) \neq \overline{\Psi}(X) \neq \; \overline{\Psi}(X + \delta_\beta))$$

Dann gilt:

$$d_{TV}\left(\mathcal{L}(\overline{\Psi}(\Theta)), \mathcal{L}(\overline{\Psi}(\widetilde{\Theta}))\right) \;\leq\; 2 \cdot \int_\Gamma \int_0^\infty \sum_{\beta \in \Gamma'_\alpha} \overline{\varphi}_{\alpha\beta}(\mathcal{L}(Z_t)) \cdot e^{-2t} \; dt \; d\mu(\alpha)$$

Beweis von Satz 1.2: Um Satz 1.1 anwenden zu können, müssen wir die Cluster-Neutralität von $\overline{\Psi}$ nachweisen. Sei also $\alpha = (v, i, j, \mathtt{B}) \in \Gamma$, $\beta = (v, i', j', \mathtt{B}) \in \Gamma'_\alpha$ und $\xi \in \mathcal{Z}$. O.B.d.A. sei $\xi(\alpha) = \xi(\beta) = 0$.

Wir nehmen nun an, es gelte $\overline{\Psi}(\xi + \delta_\alpha) \neq \overline{\Psi}(\xi)$. O.B.d.A. gebe es ein $b \in B_{\mathbf{T}}$ mit $\overline{\Psi}(\xi + \delta_\alpha)(b, i) = \mathcal{P} \neq \overline{\Psi}(\xi)(b, i)$. Dann gilt $\Psi(\xi + \delta_\alpha)(b, i) = \mathcal{P} \neq \Psi(\xi)(b, i)$ und es existiert ein $k \in \{l, \ldots, n_{\text{amp}}\}$ mit $\Psi(\xi + \delta_\alpha)(b, i + k) = \mathcal{K} = \Psi(\xi)(b, i + k)$.

Wegen $|i - i'| \in \{1, \cdots, l - 1\}$ kann man durch Hinzunehmen von δ_β keine Veränderungen an den Positionen i und $i + k$ im Bild bzgl. Ψ bewirken. Also gilt $\overline{\Psi}(\xi + \delta_\alpha + \delta_\beta)(b, i) = \mathcal{P} \neq \overline{\Psi}(\xi + \delta_\beta)(b, i)$.

Die umgekehrte Beweisrichtung folgt analog.

Also ist $\overline{\Psi}$ Cluster-neutral und die Aussage folgt aus Satz 1.1.

□

Für den Beweis von Satz 1.1 benötigen wir einige Lemmata. Wir verwenden folgende Schreibweisen:

Für $\xi, \zeta \in \mathcal{Z}$ und $v \in V_{\mathbf{T}}$ sei $\mathbf{T}_v$ die Menge der von v abstammenden $w \in V_{\mathbf{T}}$, $\Gamma^v := \{(w, i, j, \mathtt{B}) \in \Gamma \ : \ w \in \mathbf{T}_v\}$ und

$$\xi \overset{v}{\wedge} \zeta(\alpha) := \begin{cases} \xi(\alpha) & \text{falls } \alpha \in \Gamma^v \\ \zeta(\alpha) & \text{sonst} \end{cases}$$

Für $\xi \in \mathcal{Z}$ sei Z^ξ ein Prozeß auf $\mathcal{Z}$ mit $Z_0^\xi = \xi$ und derselben Übergangsdynamik wie Z.

Für $v \in V_{\mathbf{T}}$, $f \in \mathcal{F}_\Phi$, $t > 0$ und $\xi \in \mathcal{Z}$ definieren wir außerdem

$$h_{f,v}^t(\xi) := -\int_0^t \mathbb{E}\Big(f(Z_s^\xi \overset{v}{\wedge} \widetilde{\Theta}) - f(\widetilde{\Theta})\Big)\, ds$$

und setzten $h_{f,v}(\xi) := h_{f,v}^\infty(\xi)$ für alle ξ, für die dies definiert ist. Nach Lemma 1.3(a) sind dies $\mathcal{L}(\Theta)$-fast alle. (Für die übrigen setzten wir $h_{f,v}(\xi) := 0$.) Für $f \in \mathcal{F}_\Phi$ und $\xi \in \mathcal{Z}_n$ sei

$$h_f(\xi) = -\int_0^\infty \mathbb{E}\Big(f(Z_t^\xi) - f(\widetilde{\Theta})\Big)\, dt.$$

Lemma 1.3 *Für $\mathcal{L}(\Theta)$-fast alle ξ, $v \in V_{\mathbf{T}}$ und $f \in \mathcal{F}_\Phi$ gilt:*

$$\begin{array}{ll} (a) & \displaystyle\int_0^\infty \mathbb{E}\Big|f(Z_s^\xi \overset{v}{\wedge} \widetilde{\Theta}) - f(\widetilde{\Theta})\Big|\, ds \ < \ \infty \\ (b) & \displaystyle\sup_{t>0} \big|h_{f,v}^t(\xi)\big| \ < \ \infty \end{array}$$

Lemma 1.4 *Sei $w_0 \in V_{\mathbf{T}}$ die Wurzel von $\mathbf{T}$, also der jüngste gemeinsame Vorfahr aller Blätter. Für $g > 0$ bezeichne v_g dasjenige Element von $V_{\mathbf{T}}$ mit $w_0 \in \mathbf{T}_{v_g}$ und $\lambda([v_g, w_0]) = g$. Dann gelten für $f \in \mathcal{F}_\Phi$ und $\mathcal{L}(\Theta)$-fast alle ξ folgende Aussagen:*

(i)

$$\int_\Gamma |h_f(\xi + \delta_\alpha) - h_f(\xi)| \ d\mu(\alpha) \ < \infty$$

(ii)

$$\int_\Gamma \Big| [h_{f,v_g}(\xi + \delta_\alpha) - h_{f,v_g}(\xi)] - [h_f(\xi + \delta_\alpha) - h_f(\xi)] \Big| \ d\mu(\alpha) \ \to 0 \qquad \textit{für } g \to \infty$$

(iii)

$$\int_\Gamma \Big| [h_{f,v_g}(\xi - \delta_\alpha) - h_{f,v_g}(\xi)] - [h_f(\xi - \delta_\alpha) - h_f(\xi)] \Big| \ d\xi(\alpha) \ \to 0 \qquad \textit{für } g \to \infty$$

Lemma 1.5 *Für $f \in \mathcal{F}_\Phi$ und $\mathcal{L}(\Theta)$-fast alle $\xi \in \mathcal{Z}$ gilt:*

$$(\mathcal{A}h_f)(\xi) \ = \ f(\xi) - \mathbb{E}f(\widetilde{\Theta})$$

Lemma 1.5 geht über folgendes Korollar zu Lemma 1.5 in den Beweis von Satz 1.1 ein:

Korollar 1.6 *Für $f \in \mathcal{F}_\Phi$ ist*

$$d_\Phi\left(\mathcal{L}(\Theta), \mathcal{L}(\widetilde{\Theta})\right) = \sup_{f \in \mathcal{F}_\Phi} |\mathbb{E}\,(\mathcal{A}h_f)\,(\Theta)|.$$

□

Beweis von Lemma 1.3: Wir konstruieren eine Kopplung zwischen $\mathcal{L}(Z_t^\xi \overset{v}{\wedge} \widetilde{\Theta})$ und $\mathcal{L}(\widetilde{\Theta})$. Dazu seien $(\widetilde{D}_t)_{t\geq 0}$ und $(D_t^\xi)_{t\geq 0}$ reine Todesprozesse über Γ mit pro-Kopf-Sterberate 1 und $\widetilde{D}_0 = \widetilde{\Theta}$ und $D_0^\xi = \xi \overset{v}{\wedge} \widetilde{\Theta}$. $(Z_t^0)_{t\geq 0}$ sei ein Prozeß mit derselben Übergangsdynamik wie $(Z_t)_{t\geq 0}$ und $Z_0^0 \equiv 0$. Dann gilt wegen $\mathcal{L}(Z_t^\xi \overset{v}{\wedge} \widetilde{\Theta}) = \mathcal{L}(D_t^\xi + Z_t^0)$ und $\mathcal{L}(\widetilde{\Theta}) = \mathcal{L}(\widetilde{D}_t + Z_t^0)$:

$$\mathbb{E}\Big| f(Z_s^\xi \overset{v}{\wedge} \widetilde{\Theta}) - f(\widetilde{\Theta})\Big| = \mathbb{E}\Big| f(D_s^\xi + Z_s^0) - f(\widetilde{D}_s + Z_s^0)\Big|$$

und

$$h_{f,v}^t(\xi) = -\int_0^t \mathbb{E}\left(f(D_s^\xi + Z_s^0) - f(\widetilde{D}_s + Z_s^0)\right) ds$$

Wegen $f \in \mathcal{F}_\Phi$ gilt $\Big| f(D_s^\xi + Z_s^0) - f(\widetilde{D}_s + Z_s^0)\Big| \leq 1$ für alle $s \geq 0$. Allerdings kann $\Big| f(D_s^\xi + Z_s^0) - f(\widetilde{D}_s + Z_s^0)\Big| \neq 0$ für ein festes s nur dann gelten, wenn es ein $\alpha \in \Gamma^v$ mit $\left(D_s^\xi - \widetilde{D}_s\right)(\alpha) \neq 0$ gibt, also nur vor dem zufälligen Zeitpunkt $\tau := \inf\{s : D_s^\xi(\alpha) = \widetilde{D}_s(\alpha) = 0 \,\forall \alpha \in \Gamma^v\}$. Für alle $t > 0$ gilt also $|h_{f,v}^t(\xi)| \leq \mathbb{E}\tau$. Außerdem gilt $\int_0^\infty \Big| f(D_s^\xi + Z_s^0) - f(\widetilde{D}_s + Z_s^0)\Big| ds \leq \tau$ und damit $\int_0^\infty \mathbb{E}\Big| f(Z_s^\xi \overset{v}{\wedge} \widetilde{\Theta}) - f(\widetilde{\Theta})\Big| ds \leq \mathbb{E}\tau$. Zu zeigen bleibt also, daß $\mathbb{E}\tau < \infty$ gilt.

Die bedingte Erwartung von τ, gegeben daß $D_0^\xi(\alpha) = 1$ für k_1 viele α und $\widetilde{D}_0(\beta) = 1$ für k_2 viele β gilt, ist $\sum_{r=1}^{k_1+k_2} 1/r \leq k_1 + k_2$. Also ist die Erwartung von τ durch die Summe aus $\mathbb{E}\int_{\Gamma^v} d\widetilde{\Theta}(\alpha) = \int_{\Gamma^v} d\mu(\alpha)$ und der Anzahl der $\alpha \in \Gamma^v$ mit $\xi(\alpha) = 1$ beschränkt und damit für $\mathcal{L}(\Theta)$-fast alle ξ endlich.

□

Beweis von Lemma 1.4: Seien D^ξ und Z^0 wie im Beweis von Lemma 1.3 definiert. Es gilt

$$h_f(\xi+\delta_\alpha)-h_f(\xi)=-\int_0^\infty \mathbb{E}(f(D_t^\xi+D_t^{\delta_\alpha}+Z_t^0)-f(D_t^\xi+Z_t^0)dt$$

und für alle t:

$$|f(D_t^\xi+D_t^{\delta_\alpha}+Z_t^0)-f(D_t^\xi+Z_t^0)|\le 1$$

Für den Beweis von *(i)* sei $g>1$ so gewählt, daß für alle i und j ein $w\in\mathbf{T}_{v_g}$ und ein $\mathtt{B}$ existieren, so daß $\xi(w,i,j,\mathtt{B})=1$ gilt. Dies ist für $\mathcal{L}(\Theta)$-fast alle ξ möglich.

Für jedes t und jedes α ist $\{D_t^{\delta_\alpha}(\alpha)=1\}$ eine notwendige Bedingung für $|f(D_t^\xi+D_t^{\delta_\alpha}+Z_t^0)-f(D_t^\xi+Z_t^0)|>0$. Die Bedingung ist mit Wahrscheinlichkeit e^{-t} erfüllt. Es gilt also:

$$|h_f(\xi+\delta_\alpha)-h_f(\xi)|\le\int_0^\infty e^{-t}dt=1$$

Wegen $\mu(\Gamma^{v_g})<\infty$ genügt es zu zeigen, daß

$$\int_{\Gamma\setminus\Gamma^{v_g}}\Big|h_f(\xi+\delta_\alpha)-h_f(\xi)\Big|\,d\mu(\alpha)\ <\infty$$

gilt. Sei also nun $\alpha=(v_l,i,j,\mathtt{B})\in\Gamma\setminus\Gamma^{v_g}$. Dann existiert ein $\beta=(w,i,j,\mathtt{B}')\in\Gamma^{v_l}$ mit $\xi(\beta)=1$. Außer $\{D_t^{\delta_\alpha}(\alpha)=1\}$ müssen dann für $|f(D_t^\xi+D_t^{\delta_\alpha}+Z_t^0)-f(D_t^\xi+Z_t^0)|\ge 0$ auch noch die Ereignisse $\{D_t^\xi(\beta)=0\}$ und $\{\forall\mathtt{B}'',u\in[v_l,w_0]:Z_t^0(u,i,j,\mathtt{B}'')=0\}$ eintreten. Diese haben die Wahrscheinlichkeiten $1-e^{-t}$ und $\exp(-l\cdot(1-e^{-t}))$. Für letzteres überlege man sich, daß Z_t^0 für festes t ein Poisson-Prozeß auf Γ mit Intensität $(1-e^{-t})\cdot\mu$ ist. Also gilt:

$$|h_f(\xi+\delta_\alpha)-h_f(\xi)|\le\int_0^\infty e^{-t}\cdot(1-e^{-t})\cdot e^{-l\cdot(1-e^{-t})}dt=\frac{1-e^{-l}-l\cdot e^{-l}}{l^2}\le\text{const.}\cdot l^{-2}$$

Damit erhalten wir:

$$\int_\Gamma\Big|h_f(\xi+\delta_\alpha)-h_f(\xi)\Big|\,d\mu(\alpha)\le\mu(\Gamma^{v_g})+\text{const.}\cdot\int_g^\infty l^{-2}\,dl<\infty$$

Für den Beweis von *(ii)* sei g hinreichend groß gewählt, so daß für jedes Paar (i,j) ein w_{ij} und ein $\mathtt{B}_{ij}$ mit $\xi(w_{ij},i,j,\mathtt{B}_{ij})>0$ existieren. Dies ist für $\mathcal{L}(\Theta)$-fast alle ξ möglich.

Für $\alpha\in\Gamma^{v_g}$ gilt

$$[h_{f,v_g}(\xi+\delta_\alpha)-h_{f,v_g}(\xi)]-[h_f(\xi+\delta_\alpha)-h_f(\xi)]=\int_0^\infty\mathbb{E}F_t(g)dt$$

mit

$$F_t(g)=[f(D_t^{\xi\overset{v_g}{\wedge}\tilde\Theta}+D_t^{\delta_\alpha}+Z_t^0)-f(D_t^{\xi\overset{v_g}{\wedge}\tilde\Theta}+Z_t^0)]-[f(D_t^\xi+D_t^{\delta_\alpha}+Z_t^0)-f(D_t^\xi+Z_t^0)].$$

Notwendig für $|F_t(g)|>0$ ist die Existenz eines Paares (i,j), so daß $Z_t^0(v,i,j,\mathtt{B})=0$ für alle $v\in\mathbf{T}_{v_g}$ und alle $\mathtt{B}$ gilt, und außerdem $D_t^\xi(w_{ij},i,j,\mathtt{B}_{ij})=0$ ist. Zusätzlich muß $\{D_t^{\delta_\alpha}(\alpha)=1\}$ eintreten. Ähnlich wie im Beweis von *(i)* folgt daraus $\int_0^\infty\mathbb{E}F_t(g)dt\le$const.$\cdot g^{-2}$.

Im Fall $\alpha \in \Gamma \setminus \Gamma^{v_g}$ gilt $h_{f,v_g}(\xi + \delta_\alpha) = h_{f,v_g}(\xi)$ und es bleibt $\int_{\Gamma\setminus\Gamma^{v_g}} |h_f(\xi+\delta_\alpha) - h_f(\xi)|$ abzuschätzen. Nach den Überlegungen aus dem Beweis von *(i)* ist dies ebenfalls $\leq$ const. $\cdot \int_g^\infty l^{-2} dl = \text{const.} \cdot g^{-1}$.

Wir erhalten also:

$$
\begin{aligned}
&\int_\Gamma \Big| [h_{f,v_g}(\xi+\delta_\alpha) - h_{f,v_g}(\xi)] - [h_{f,v_g}(\xi+\delta_\alpha) - h_{f,v_g}(\xi)] \Big| \, d\mu(\alpha) \\
&\quad \leq \quad \text{const.} \cdot g^{-2} \cdot \mu(\Gamma^{v_g}) + \text{const.} \cdot g^{-1} \\
&\quad \leq \quad \text{const.} \cdot g^{-1} \qquad \to 0 \qquad \text{für } g \to \infty
\end{aligned}
$$

Der Beweis von *(iii)* verläuft analog zum Beweis von *(ii)*. Man beachte dabei, daß $\xi(\Gamma^{v_g})$ für $\mathcal{L}(\Theta)$-fast alle ξ asymptotisch linear in g ist.

$\square$

Beweis von Lemma 1.5: Sei $v \in V_{\mathbf{T}}$ so gewählt, daß für jedes Paar $(i,j) \in \{1,\ldots,n-l+1\} \times \{1,\ldots,l\}$ mindestens ein Paar $(w,\mathtt{B}) \in \mathbf{T}_v \times \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$ mit $\xi(w,i,j,\mathtt{B}) > 0$ existiert. Das ist für $\mathcal{L}(\Theta)$-fast alle ξ möglich. Dann gilt $f(\xi \overset{v}{\wedge} \widetilde{\Theta}) = f(\xi)$. Sei außerdem $S := \min\{u : Z_0^\xi\rfloor_{\Gamma^v} \neq Z_u^\xi\rfloor_{\Gamma^v}\}$ und es sei Γ^v die Menge der $(w,i,j,\mathtt{B}) \in \Gamma$ mit $w \in \mathbf{T}_v$. Außerdem sei $q_0 := \mu(\Gamma^v)$ und $q_1 := \sum_{\alpha\in\Gamma^v} \xi(\alpha)$.

Wir unterscheiden drei Fälle, von denen fast sicher genau einer eintritt:

(A) $S \geq t$

(B) $S < t$ und $\exists \alpha = (w,i,j,\mathtt{B}) \in \Gamma^v : Z_S^\xi\rfloor_{\Gamma^v} = \xi\rfloor_{\Gamma^v} - \delta_\alpha$

(C) $S < t$ und $\exists \alpha = (w,i,j,\mathtt{B}) \in \Gamma^v : Z_S^\xi\rfloor_{\Gamma^v} = \xi\rfloor_{\Gamma^v} + \delta_\alpha$

Fall (A) tritt mit Wahrscheinlichkeit $\exp(-(q_0+q_1)t)$ ein und zieht nach sich, daß $f(Z_u^\xi \overset{v}{\wedge} \widetilde{\Theta}) = f(\xi)$ für alle $u < t$ gilt.

Daß Fall (B) eintritt und für ein $s < t$ die Stoppzeit S in $[s, s+ds]$ fällt, hat die Wahrscheinlichkeit $\exp(-(q_0+q_1)s) \cdot q_1 \cdot ds$. Die darauf bedingte Erwartung von $\int_0^t [f(Z_u^\xi \overset{v}{\wedge} \widetilde{\Theta}) - \mathbb{E}f(\widetilde{\Theta})]\, du$ ist

$$
\begin{aligned}
& - \quad (f(\xi) - \mathbb{E}f(\widetilde{\Theta})) \cdot s \\
& + \quad \frac{1}{q_1} \cdot \int_{\Gamma^v} \mathbb{E}\left(- \int_0^{t-s} \left[f(Z_0^{\xi-\delta_\alpha} \overset{v}{\wedge} \widetilde{\Theta}) - f(\widetilde{\Theta}) \right] \, du \right) \, d\xi(\alpha) \\
= & - \quad (f(\xi) - \mathbb{E}f(\widetilde{\Theta})) \cdot s \\
& + \quad \frac{1}{q_1} \cdot \int_\Gamma \cdot h_{f,v}^{t-s}(\xi - \delta_\alpha) \, d\xi(\alpha)
\end{aligned}
$$

Analog erhält man für das Eintreten von Fall (C) mit $\{S \in [s, s+ds]\}$ die Wahrscheinlichkeit $\exp(-(q_0+q_1)s) \cdot q_0 \cdot ds$ und die darauf bedingte Erwartung von $\int_0^t \left[f(Z_u^\xi \overset{v}{\wedge} \widetilde{\Theta}) - \mathbb{E}f(\widetilde{\Theta}) \right] \, du$ ist

$$
-(f(\xi) - \mathbb{E}f(\widetilde{\Theta})) \cdot s + \frac{1}{q_0} \cdot \int_{\Gamma^v} h_{f,v}^{t-s}(\xi + \delta_\alpha) d\mu(\alpha).
$$

Wir erhalten also:

$$\begin{aligned}
h^t_{f,v}(\xi) &= \mathbb{E}\Big(-\int_0^t f(Z^\xi_u \overset{v}{\wedge} \widetilde{\Theta}) - \mathbb{E}f(\widetilde{\Theta}))du\Big) \\
&= \exp(-(q_0+q_1)t)\cdot(\mathbb{E}f(\widetilde{\Theta}) - f(\xi))\cdot t \\
&\quad + \int_0^t \exp(-(q_0+q_1)s)\cdot q_1 \cdot \Big[(\mathbb{E}f(\widetilde{\Theta}) - f(\xi))\cdot s + \frac{1}{q_1}\int_{\Gamma^v} h^{t-s}_{f,v}(\xi-\delta_\alpha)d\xi(\alpha)\Big]ds \\
&\quad + \int_0^t \exp(-(q_0+q_1)s)\cdot q_0 \cdot \Big[(\mathbb{E}f(\widetilde{\Theta}) - f(\xi))\cdot s + \frac{1}{q_0}\int_{\Gamma^v} h^{t-s}_{f,v}(\xi+\delta_\alpha)d\mu(\alpha)\Big]ds
\end{aligned}$$

Unter Beachtung von Lemma 1.3 wenden wir nun den Satz von der dominierten Konvergenz an und erhalten:

$$\begin{aligned}
h_{f,v}(\xi) &= \int_0^\infty \exp(-(q_0+q_1)s)\cdot\Big[(q_1+q_0)\cdot s\cdot(\mathbb{E}f(\widetilde{\Theta}) - f(\xi)) \\
&\qquad + \int_{\Gamma^v} h_{f,v}(\xi-\delta_\alpha)d\xi(\alpha) + \int_{\Gamma^v} h_{f,v}(\xi+\delta_\alpha)d\mu(\alpha)\Big]ds \\
&= \frac{1}{q_0+q_1}\cdot\Big[\mathbb{E}f(\widetilde{\Theta}) - f(\xi) + \int_{\Gamma^v} h_{f,v}(\xi+\delta_\alpha)d\mu(\alpha) + \int_{\Gamma^v} h_{f,v}(\xi-\delta_\alpha)d\xi(\alpha)\Big]
\end{aligned}$$

Daraus folgt:

$$\begin{aligned}
f(\xi) - \mathbb{E}f(\widetilde{\Theta}) &= \int_{\Gamma^v}(h_{f,v}(\xi+\delta_\alpha) - h_{f,v}(\xi))\,d\mu(\alpha) + \int_{\Gamma^v}(h_{f,v}(\xi-\delta_\alpha) - h_{f,v}(\xi))\,d\xi(\alpha) \\
&= \int_{\Gamma}(h_{f,v}(\xi+\delta_\alpha) - h_{f,v}(\xi))\,d\mu(\alpha) + \int_{\Gamma}(h_{f,v}(\xi-\delta_\alpha) - h_{f,v}(\xi))\,d\xi(\alpha)
\end{aligned}$$

Mit Lemma 1.4 folgt:

$$f(\xi) - \mathbb{E}f(\widetilde{\Theta}) = \int_{\alpha\in\Gamma}(h_f(\xi+\delta_\alpha) - h_f(\xi))\,d\mu(\alpha) + \int_{\alpha\in\Gamma}(h_f(\xi-\delta_\alpha) - h_f(\xi))\,d\xi(\alpha)$$

□

Beweis von Satz 1.1: Wir wollen Korollar 1.6 anwenden, also schätzen wir $|\mathbb{E}(\mathcal{A}h_f)(\Theta)|$ für alle $f \in \mathcal{F}_\Phi$ gleichmäßig nach oben ab. $\Theta_\alpha := \Theta + \delta_\alpha + \sum_{\beta\in\Gamma'_\alpha}\delta_\beta$ ist für $\alpha \in \Gamma$ ein Palmscher Prozeß des Poisson-Cluster-Prozesses Θ (vgl. Lemma 10.6 in Kallenberg (1986)). Wegen $\Pr(\Theta(\alpha) \leq 1 \forall\alpha \in \Gamma) = 1$ läßt sich $\mathcal{L}(\Theta_\alpha)$ als $\mathcal{L}(\Theta|\Theta(\alpha) = 1)$ interpretieren. Außerdem gilt die Campbell Formel

$$\mathbb{E}\int_\Gamma H(\alpha,\Theta)d\Theta(\alpha) = \int_\Gamma \mathbb{E}[H(\alpha,\Theta_\alpha)]\,d\mu(\alpha)$$

für alle integrierbaren $H : \Gamma\times\mathcal{Z} \to \mathbb{R}$ (vgl. Lemma 10.1 in Kallenberg (1986)).

Speziell mit $H(\alpha,\xi) := \xi(\alpha)\cdot(h_f(\xi-\delta_\alpha) - h_f(\xi))$ folgt daraus:

$$\mathbb{E}\int_\Gamma [h_f(\Theta-\delta_\alpha) - h_f(\Theta)]\;d\Theta(\alpha) = \int_\Gamma \mathbb{E}\,(h_f(\Theta_\alpha-\delta_\alpha) - h_f(\Theta_\alpha))\,d\mu(\alpha)$$

Damit ergibt sich:

$$\mathbb{E}(\mathcal{A}h_f)(\Theta) = \int_\Gamma \mathbb{E}\Big[\ h_f(\Theta+\delta_\alpha) - h_f(\Theta) + h_f\Big(\Theta + \sum_{\beta\in\Gamma'_\alpha}\delta_\beta\Big) - h_f\Big(\Theta + \delta_\alpha + \sum_{\beta\in\Gamma'_\alpha}\delta_\beta\Big)\Big]\,d\mu(\alpha)$$

Nun ist aber

$$h_f\left(\Theta+\delta_\alpha\right) - h_f\left(\Theta\right) + h_f\left(\Theta + \sum_{\beta\in\Gamma'_\alpha}\delta_\beta\right) - h_f\left(\Theta + \delta_\alpha + \sum_{\beta\in\Gamma'_\alpha}\delta_\beta\right)$$

$$= \int_0^\infty \mathbb{E}\left[f\left(Z_t + \delta_\alpha\cdot I_{\epsilon_\alpha>t} + \sum_{\beta\in\Gamma'_\alpha}\delta_\beta I_{\epsilon_\beta>t}\right) - f\left(Z_t + \sum_{\beta\in\Gamma'_\alpha}\delta_\beta I_{\epsilon_\beta>t}\right)\right.$$
$$\left. - f\left(Z_t + \delta_\alpha\cdot I_{\epsilon_\alpha>t}\right) + f(Z_t)\right] dt$$

Dabei ist $\{\epsilon_\gamma : \gamma\in\Gamma'_\alpha\cup\{\alpha\}\}$ eine Familie untereinander und von Θ unabhängiger $\exp(1)$-verteilter Zufallsvariablen und $I_{\epsilon_\gamma>t}$ bezeichnet die Indikatorfunktion von $\{\epsilon_\gamma > t\}$.

$$= \int_0^\infty \mathbb{E}\left[f\left(Z_t + \delta_\alpha + \sum_{\beta\in\Gamma'_\alpha}\delta_\beta I_{\epsilon_\beta>t}\right) - f\left(Z_t + \sum_{\beta\in\Gamma'_\alpha}\delta_\beta I_{\epsilon_\beta>t}\right)\right.$$
$$\left. - f\left(Z_t + \delta_\alpha\right) + f(Z_t)\right]\cdot e^{-t}dt$$

$$= \int_0^\infty \mathbb{E}\left[g_t\left(\sum_{\beta\in\Gamma'_\alpha}\delta_\beta I_{\epsilon_\beta>t}\right) - g_t(0)\right]\cdot e^{-t}dt$$

mit $-1 \le g_t(\eta) := f(Z_t+\delta_\alpha+\eta) - f(Z_t+\eta) \le 1$

$$= -\sum_{k=1}^h \int_0^\infty \mathbb{E}\left[g_t\left(\sum_{j=1}^k \delta_{\beta_j} I_{\epsilon_j>t}\right) - g_t\left(\sum_{j=1}^{k-1}\delta_{\beta_j} I_{\epsilon_j>t}\right)\right]\cdot e^{-t}dt$$

Mit $\{\beta_1,\ldots,\beta_h\} := \Gamma'_\alpha$ erhält man dies durch Teleskopieren.

$$= -\sum_{k=1}^h \int_0^\infty \mathbb{E}\left[g_t\left(\delta_{\beta_k} + \sum_{j=1}^{k-1}\delta_{\beta_j} I_{\epsilon_j>t}\right) - g_t\left(\sum_{j=1}^{k-1}\delta_{\beta_j} I_{\epsilon_j>t}\right)\right]\cdot e^{-2t}dt$$

Wegen der Cluster-Neutralität von Φ ist die Gleichung

$$\Phi\left(Z_t + \delta_\alpha + \sum_{j=1}^k \delta_{\beta_j}\right) = \Phi\left(Z_t + \sum_{j=1}^k \delta_{\beta_j}\right)$$

äquivalent zu

$$\Phi\left(Z_t+\delta_\alpha\right) = \Phi\left(Z_t\right).$$

Da $\Gamma_\alpha = \Gamma_\beta$ für $\beta\in\Gamma'_\alpha$ ist, folgt auch die Äquivalenz von

$$\Phi\left(Z_t + \delta_{\beta_k} + \sum_{j=1}^{k-1}\delta_{\beta_j}\right) = \Phi\left(Z_t + \sum_{j=1}^{k-1}\delta_{\beta_j}\right)$$

und

$$\Phi\left(Z_t+\delta_{\beta_k}\right)=\Phi\left(Z_t\right).$$

Daher und wegen $f \in \mathcal{F}_\Phi$ ist $g_t(\sum_{j=1}^k \delta_{\beta_j}) = f(Z_t+\delta_\alpha+\sum_{j=1}^k \delta_{\beta_j}) - f(Z_t+\sum_{j=1}^k \delta_{\beta_j}) = 0$, falls $\Phi\left(Z_t+\delta_\alpha\right)=\Phi\left(Z_t\right)$ gilt, und es ist $g_t(\delta_{\beta_k}+\sum_{j=1}^{k-1}\delta_{\beta_j}) = g_t(\sum_{j=1}^{k-1}\delta_{\beta_j})$, falls $\Phi\left(Z_t+\delta_{\beta_k}\right)=\Phi\left(Z_t\right)$ gilt. Wegen $|g_t(x)| \leq 1$ folgt:

$$\begin{aligned}
&\left|\mathbb{E}\left[g_t\left(\delta_{\beta_k}+\sum_{j=1}^{k-1}\delta_{\beta_j} I_{\epsilon_j>t}\right)-g_t\left(\sum_{j=1}^{k-1}\delta_{\beta_j} I_{\epsilon_j>t}\right)\right]\right| \\
&\quad\leq\; 2\cdot\Pr\left(g_t\left(\delta_{\beta_k}+\sum_{j=1}^{k-1}\delta_{\beta_j} I_{\epsilon_j>t}\right)\neq g_t\left(\sum_{j=1}^{k-1}\delta_{\beta_j} I_{\epsilon_j>t}\right)\right) \\
&\quad\leq\; 2\cdot\Pr\left(\Phi\left(Z_t+\delta_{\beta_k}\right)\neq\Phi\left(Z_t\right)\neq\Phi\left(Z_t+\delta_\alpha\right)\right) \\
&\quad=\; 2\cdot\varphi_{\alpha\beta_k}\left(\mathcal{L}(Z_t)\right)
\end{aligned}$$

Also gilt für alle $f \in \mathcal{F}_\Phi$:

$$\left|\mathbb{E}\left(\mathcal{A}h_f\right)(\Theta)\right| \leq 2\cdot\int_\Gamma \sum_{\beta\in\Gamma'_\alpha}\int_0^\infty \varphi_{\alpha\beta}\left(\mathcal{L}(Z_t)\right)\cdot e^{-2t}\ dt\ d\mu(\alpha)$$

Die Behauptung des Satzes folgt nun aus Korollar 1.6.

□

1.3.1 Eine Abschätzung für $\overline{\varphi}_{\alpha\beta}(\mathcal{L}(Z_t))$

Für einen gegebenen Baum $\mathbf{T}$ und eine gegebene Menge $\mathcal{M}$ kann man $\int_\Gamma \sum_{\beta\in\Gamma'_\alpha}\overline{\varphi}_{\alpha\beta}(\mathcal{L}(Z_t))e^{-2t}d\mu(\alpha)$ mit einem ähnlichen Verfahren wie dem, welches in Abschnitt 2.2.2 für die Berechnung von e_1 und e_2 beschrieben wird, berechnen. Die nötige Rechenzeit wächst dann linear in der Anzahl der Blätter des Baumes. Ein solches Verfahren ist aber programmtechnisch relativ aufwendig, zumal auch noch über $t\ \in]0,\infty[$ numerisch integriert werden muß.

Wie sich in Abschnitt 1.3.2 zeigen wird, liefert für einen dreiblättrigen Baum bereits eine recht grobe Abschätzung für $\overline{\varphi}_{\alpha\beta}(\mathcal{L}(Z_t))$, die wir gleich herleiten werden, zusammen mit Satz 1.2 eine akzeptable Abschätzung für $d_{\overline{\Psi}}(\mathcal{L}(\Theta),\mathcal{L}(\widetilde{\Theta}))$.

Machen wir uns erst einmal klar, daß für alle $\alpha=(v_\alpha,i_\alpha,j_\alpha,\mathtt{B}_\alpha)\in\Gamma$, $\beta\in\Gamma'_\alpha$ und $t\geq 0$ gilt:

$$\begin{aligned}
\overline{\varphi}_{\alpha\beta}(\mathcal{L}(Z_t)) &= \Pr\left(\overline{\Psi}(Z_t+\delta_\alpha)\neq\overline{\Psi}(Z_t)\right)\cdot\Pr\left(\overline{\Psi}(Z_t+\delta_\beta)\neq\overline{\Psi}(Z_t)\mid\overline{\Psi}(Z_t+\delta_\alpha)\neq\overline{\Psi}(Z_t)\right) \\
&\leq \Pr\left(\overline{\Psi}(Z_t+\delta_\alpha)\neq\overline{\Psi}(Z_t)\right)\cdot \\
&\qquad\cdot\sum_{b\in B_{\mathbf{T}}}\Pr\left(\overline{\Psi}(Z_t+\delta_\beta)\neq\overline{\Psi}(Z_t)\mid\overline{\Psi}(Z_t+\delta_\alpha)(b,i_\alpha)\neq\overline{\Psi}(Z_t)(b,i_\alpha)\right).
\end{aligned}$$

In der Tat ist ja zum einen das Ereignis $\{\overline{\Psi}(Z_t+\delta_\alpha)\neq\overline{\Psi}(Z_t)\}$ die Vereinigung der Ereignisse $\{\overline{\Psi}(Z_t+\delta_\alpha)(b,i_\alpha)\neq\overline{\Psi}(Z_t)(b,i_\alpha)\}$, $b\in B_{\mathbf{T}}$, zum anderen gilt allgemein für alle Ereignisse $H,G_1,\ldots,G_k$ und $G:=\bigcup_{j=1}^k G_j$ die Ungleichung $\Pr(H|G)\leq\sum_j\Pr(H|G_j)$

Um $\Pr\left(\overline{\Psi}(Z_t + \delta_\alpha) \neq \overline{\Psi}(Z_t)\right)$ von oben abzuschätzen, numerieren wir die Blätter des Teilbaumes $\mathbf{T}_v$, der dem v_α abstammt, von links nach rechts mit $b_1, \ldots, b_{|B_{\mathbf{T}_v}|}$ durch (vgl. Abb. 1.2).

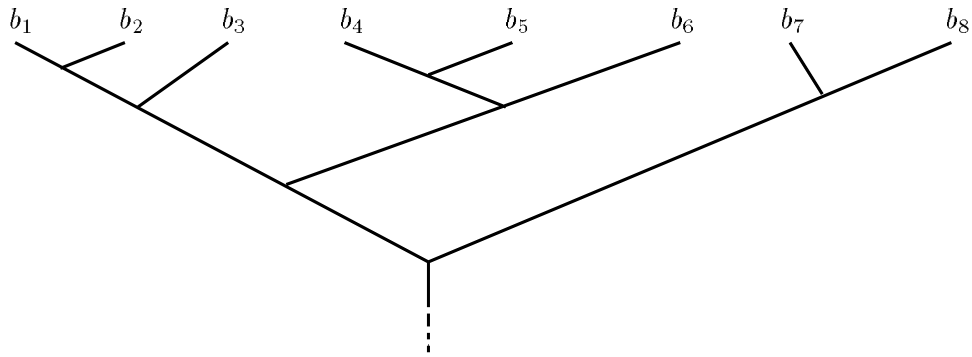


Abbildung 1.2: Ein Baum, dessen Blätter von links nach rechts durchnumeriert wurden

Es ist

$$
\begin{aligned}
&\Pr\left(\overline{\Psi}(Z_t + \delta_\alpha) \neq \overline{\Psi}(Z_t)\right) \\
&= \sum_{j=1}^{|B_\mathbf{T}|} \Pr\left(\overline{\Psi}(Z_t + \delta_\alpha)(b_j, i_\alpha) \neq \overline{\Psi}\left(Z_t\right)\left(b_j, i_\alpha\right)\right) \cdot \\
&\qquad \cdot \Pr\Bigg(\forall_{k<j}\, \overline{\Psi}\left(Z_t + \delta_\alpha\right)\left(b_k, i_\alpha\right) = \overline{\Psi}\left(Z_t\right)\left(b_k, i_\alpha\right) \;\Bigg|\; \\
&\qquad\qquad \overline{\Psi}\left(Z_t + \delta_\alpha\right)\left(b_j, i_\alpha\right) \neq \overline{\Psi}\left(Z_t\right)\left(b_j, i_\alpha\right)\Bigg) \\
&\leq \sum_{j=1}^{|B_\mathbf{T}|} \Pr\left(\overline{\Psi}(Z_t + \delta_\alpha)(b_j, i_\alpha) \neq \overline{\Psi}\left(Z_t\right)\left(b_j, i_\alpha\right)\right) \cdot \kappa(b_1, \ldots, b_{j-1} \mid b_j)
\end{aligned}
$$

Dabei sei (für $j > 1$) $\kappa(b_1, \ldots, b_{j-1}|b_j)$ die Wahrscheinlichkeit, daß es in der Konfiguration Z_t zwischen jedem Blatt $b \in \{b_1, \ldots, b_{j-1}\}$ und dem jeweiligen jüngsten mit b_j gemeinsamen Vorfahren mindestens eine Mutation gibt, welche das an einer festen Stelle i beginnende Muster betrifft (in unserem Fall geht es um $i = i_\alpha$). Für $j = 1$ setzen wir $\kappa(|b) = 1$ für alle b.

Um $\Pr\left(\Psi(Z_t + \delta_\beta) \neq \overline{\Psi}(Z_t) \mid \overline{\Psi}(Z_t + \delta_\beta)(a, i_\alpha) \neq \overline{\Psi}(Z_t)(a, i_\alpha)\right)$ nach oben abzuschätzen, sortiere man den Baum zunächst so um, daß das Blatt a ganz rechts steht, und bezeichne die Blätter dann von links nach rechts mit $b'_1, \ldots, b'_{|B_\mathbf{T}|}$ (vgl. Abb. 1.3).
Es gilt dann mit denselben Überlegungen wie oben:

$$
\begin{aligned}
&\Pr\left(\overline{\Psi}(Z_t + \delta_\beta) \neq \overline{\Psi}(Z_t) \mid \overline{\Psi}(Z_t + \delta_\alpha)(a, i_\alpha) \neq \overline{\Psi}(Z_t)(a, i_\alpha)\right) \\
&\leq \sum_{j=1}^{|B_\mathbf{T}|} \Pr\left(\overline{\Psi}(Z_t + \delta_\beta)(b'_j, i_\beta) \neq \overline{\Psi}(Z_t)(b'_j, i_\beta) \mid \overline{\Psi}(Z_t + \delta_\alpha)(a, i_\alpha) \neq \overline{\Psi}(Z_t)(a, i_\alpha)\right) \cdot \\
&\qquad \cdot \kappa(b'_1, \ldots, b'_{j-1}|b'_j)
\end{aligned}
$$

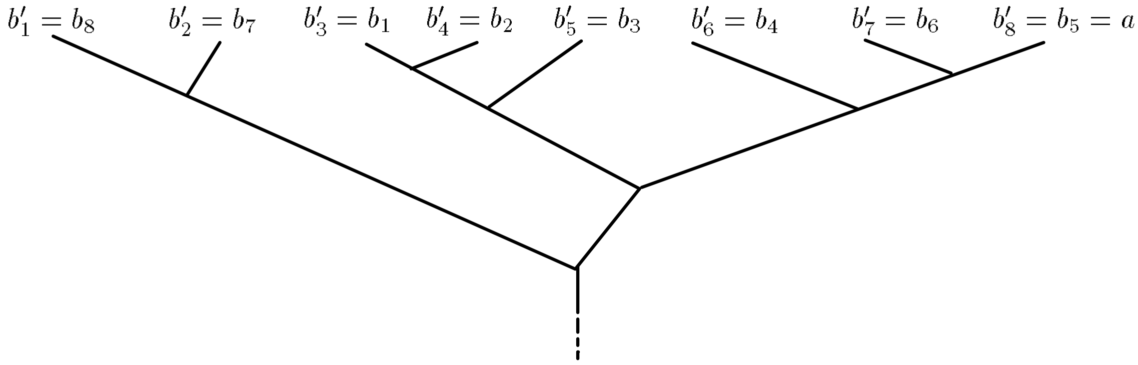


Abbildung 1.3: Eine Möglichkeit, den Baum aus Abb. 1.2 so umzusortieren, daß das Blatt a ganz rechts steht, falls $a = b_5$ ist.

1.3.2 Ein Beispiel

Wir wenden die Überlegung aus Abschnitt 1.3.1 nun auf den in Abb. 1.4 dargestellten dreiblättrigen Baum an.

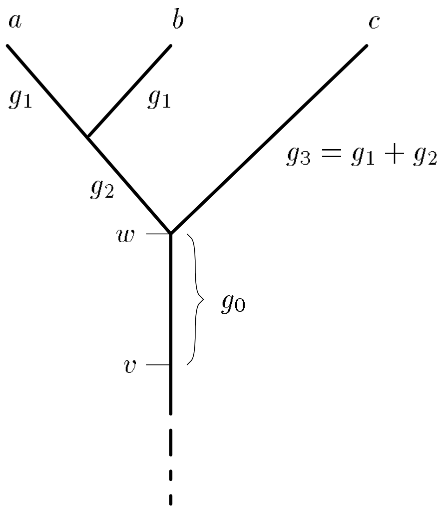


Abbildung 1.4: Ein dreiblättriger Baum als Beispiel. Es bezeichnen g_0, g_1, g_2 und g_3 die Längen (d. h. λ-Gewichte) der jeweiligen (Teil-) Kanten.

Wir gehen in diesem Beispiel von der Primersequenz $M_1 := \mathcal{P} =$ `ACGTATTA` und dem Primerkomplement $M_2 := \mathcal{K} =$ `TAATACGT` aus. Man beachte, daß es zwischen M_1 und M_2 eine Überlappungsmöglichkeit von vier Basen gibt, was wegen der geringen Länge von M_1 und M_2 relativ viel ist.

Wir gehen davon aus, daß die Länge $l = 8$ des Primers, der Amplifikationsbereich n_{amp} und die Länge n der DNA so aufeinander abgestimmt sind, daß wir auf eine DNA-Sequenz etwa 5 Banden erwarten, d.h. $n \cdot n_{\text{amp}} \cdot (1/4)^{2l} \approx 5$. Bei $n_{\text{amp}} = 3000$ ergibt sich damit eine DNA-

Länge von ungefähr 7 Millionen Basenpaaren. Für die verwandtschaftliche Nähe der Individuen nehmen wir an, daß zwischen a und b etwa jedes fünfzigste und zwischen a und c etwa jedes dreißigste Site mutiert ist, es sei also $\exp(-2g_3) = 29/30$ und $\exp(-2g_1) = 49/50$.

Die nachfolgenden Überlegungen lassen sich in vier aufeinander aufbauende Schritte gliedern:

(A) Wir betrachten die Wahrscheinlichkeit, daß ein bestimmtes Paar von Mutationen $(\alpha, \beta) \in \Gamma \times \Gamma'_\alpha$ zum Zeitpunkt t Einfluß auf das $\overline{\Psi}$-Bild von Z_t in einem gewählten Paar von Blättern ausübt.

(B) Wir übertragen die Überlegungen von **(A)** auf andere Blätter und wenden die Überlegungen aus Abschnitt 1.3.1 an, um $\overline{\varphi}_{\alpha\beta}(\mathcal{L}(Z_t))$ nach oben abzuschätzen.

(C) Wir integrieren die in **(B)** gefundene obere Schranke für $\overline{\varphi}_{\alpha\beta}(\mathcal{L}(Z_t))$ über alle Mutationen-Paare $(\alpha, \beta) \in \Gamma \times \Gamma'_\alpha$, die auf einer gewählten Kante liegen.

(D) Wir übertragen das Ergebnis von **(C)** auf alle Kanten und integrieren die Summe über $t \in]0, \infty[$.

(A) Wir betrachten zunächst ein Paar $(\alpha, \beta) \in \Gamma \times \Gamma'_\alpha$ mit $v_\alpha = v_\beta = v$ (vgl. Abbildung 1.4) und konzentrieren uns auf die Blätter a und c. Wir verwenden folgende Abkürzungen:

$$\begin{array}{rcl} M_a &:=& \overline{\Psi}(Z_t)(a, i_\alpha) \\ M'_a &:=& \overline{\Psi}(Z_t + \delta_\alpha)(a, i_\alpha) \\ M_c &:=& \overline{\Psi}(Z_t)(c, i_\beta) \\ M'_c &:=& \overline{\Psi}(Z_t + \delta_\beta)(c, i_\beta) \end{array}$$

Es ist $\Pr(M_a \neq M'_a) = \Pr(M_c \neq M'_c) < 2 \cdot (1/4)^{2l-1} \cdot n_{\text{amp}} \cdot \exp(-g_0 - g_3)$, denn für $M_a \neq M'_a$ ist z. B. notwendig, daß M'_a oder M_a mit $\mathcal{P}$ oder $\mathcal{K}$ übereinstimmt, daß innerhalb des Amplifikationsbereichs das jeweilige Komplement vorhanden ist und daß die Mutation α nicht durch eine zwischen v und a liegende Mutation überdeckt wird.

Abzuschätzen ist insbesondere:

$$\Pr(M_c \neq M'_c \mid M_a \neq M'_a) \leq \max_{k \in \{1,2\}} \Pr(M_c \neq M'_c \mid M_a \neq M'_a, M_k \in \{M_a, M'_a\})$$

(Zur Erinnerung: M_a, M'_a, M_c und M'_c sind zufällig, nicht aber M_1 und M_2.)

Betrachten wir erst einmal

$$\Pr(M_c \neq M'_c \mid M_a \neq M'_a,\ M_1 \in \{M_a, M'_a\}).$$

Eine notwendige Bedingung für $M_c \neq M'_c$ ist, daß zwischen v und c keine Mutation liegt, die β überdeckt. Die Wahrscheinlichkeit, daß zwischen w und c keine solche Mutation liegt, ist $\exp(-g_3)$. Die auf $M_a \neq M'_a$ bedingte Wahrscheinlichkeit, daß es in Z_t zwischen v und

w keine β überdeckende Mutation gibt, ist die Wahrscheinlichkeit, daß es in Z_t zwischen v und w keine solche Mutation gibt, die erst nach dem Zeitpunkt 0 hinzugekommen war, und ist demnach $\exp(-(1-e^{-t})\cdot g_0)$. (Dazu überlege man sich, daß für jede zum Zeitpunkt t vorhandene Mutation die Wahrscheinlichkeit, daß es sich um eine handelt, die erst nach dem Zeitpunkt 0 hinzugekommen ist, $1 - e^{-t}$ beträgt.)

Eine weitere notwendige Bedingung für $M_c \neq M'_c$ ist, daß $M_1 \in \{M_c, M'_c\}$ oder $M_2 \in \{M_c, M'_c\}$ gilt. Letzteres setzt voraus, daß die Primersequenz M_1 auf der DNA von c innerhalb des Amplifikationsbereichs vor i_β vorkommt. Die Wahrscheinlichkeit dafür ist von der Größenordnung von $n_{\mathrm{amp}} \cdot (1/4)^l$.

Zur Abschätzung von $\Pr(M_k \in \{M_c, M'_c\} \mid M_a \neq M'_a,\ M_1 \in \{M_a, M'_a\})$, für $k \in \{1,2\}$, sei

x_k : die Anzahl der Sites, die zwischen M_1 und M_k übereinstimmen, wenn man M_k gegenüber M_1 um $i_\beta - i_\alpha$ Positionen nach rechts verschiebt und die Position j_α bei M_1 bzw. j_β bei M_k nicht mitzählt.

y_k : die Anzahl der Sites, die zwischen M_1 und M_k **nicht** übereinstimmen, wenn man M_k gegenüber M_1 um $i_\beta - i_\alpha$ Positionen nach rechts verschiebt und die Position j_α bei M_1 bzw. j_β bei M_k nicht mitzählt.

Zur Verdeutlichung betrachten wir die Situation $i_\beta = i_\alpha + 3$, $j_\alpha = 7$ und $k = 2$. Dann ist $j_\beta = 4$ und man erhält:

```
1 2 3 4 5 6 7 8
A C G T A T T A
      + + - * +
      T A A T A C G T
      1 2 3 4 5 6 7 8
```

Die mit + gekennzeichneten Site-Paare stimmen überein, das mit - gekennzeichnete Paar stimmt nicht überein, das mit * gekennzeichnete wird nicht mitgezählt. Also ergibt sich $x_2 = 3$ und $y_2 = 1$.

Wir stellen noch eine Vorüberlegung an: Wie groß ist die Wahrscheinlichkeit, daß bei Z_t in Blatt a im i-ten Muster an der j-ten Position dieselbe Base steht wie in Blatt b im i'-ten Muster an der j'-ten Position, wobei $i + j = i' + j'$ und $i \neq i'$ ist? (Man beachte, daß die beiden Basen in Z_0 übereinstimmen.) Es kann sein, daß die Basen in beiden Positionen durch dieselbe Mutation bestimmt werden. Dazu darf zwischen w und a sowie zwischen w und b an den betreffenden Positionen keine Mutation sein. Die Wahrscheinlichkeit dafür ist $\exp(-2g_3)$. Wir gehen nun davon aus, daß sich die Mutation, die die Base an der betreffenden Position in a festlegt, bereits bei einem Vorfahren von w ereignet hat. Die Wahrscheinlichkeit, daß die Länge des Kantenstücks zwischen den Mutanten und w „in das infinitesimal kleine Intervall $[g, g + dg]$ fällt", ist $\exp(-g)dg$. Die Mutation kann nur dann auch Einfluß auf die betreffende Position in b haben, wenn sie bereits in Z_0 vorhanden war. Die Wahrscheinlichkeit für letzteres hängt nicht von g ab und ist e^{-t}. Eine letzte Bedingung muß noch erfüllt sein, damit die beiden Positionen „abstammungsgleich“ sind: Zwischen dem Urmutanten der betreffenden Position in a und dem Individuum w darf bei den Positionen, die Vorfahren der betreffenden Position in b sind, keine Mutation nach dem Zeitpunkt 0 hinzugekommen sein und bis t überlebt haben. Die Wahrscheinlichkeit dafür ist $\exp(-(1-e^{-t})g)$. Insgesamt ist also die Wahrscheinlichkeit, daß die beiden Positionen durch dieselbe Mutation bestimmt werden:

$$\exp(-2g_3) \cdot e^{-t} \cdot \int_0^\infty \exp(-g) \cdot \exp(-(1-e^{-t})g)dg = \frac{\exp(-2g_3)}{2\exp(t)-1}$$

Wenn die beiden Positionen nicht abstammungsgleich sind, so ist die Wahrscheinlichkeit, daß sie dennoch dieselbe Base haben, $1/4$. Insgesamt hat also die Wahrscheinlichkeit, daß die beiden Positionen dieselbe Base haben, den Wert

$$\frac{1}{4} + \frac{3\exp(-2g_3)}{8\exp(t)-4}$$

Demnach gilt

$$\begin{aligned} &\Pr(M_1 \in \{M_c, M_c'\} \mid M_1 \in \{M_a, M_a'\},\ M_a \neq M_a') \\ \leq\ & \left(\frac{1}{4} + \frac{3 \cdot e^{-2g_3}}{8e^t - 4}\right)^{x_1} \cdot \left(\frac{1}{4} - \frac{e^{-2g_3}}{8e^t-4}\right)^{y_1} \cdot \left(\frac{1}{4}\right)^{l-x_1-y_1-1} \end{aligned}$$

und

$$\begin{aligned} &\Pr(M_2 \in \{M_c, M_c'\} \mid M_1 \in \{M_a, M_a'\},\ M_a \neq M_a') \\ \leq\ & \left(\frac{1}{4} + \frac{3 \cdot e^{-2g_3}}{8e^t - 4}\right)^{x_2} \cdot \left(\frac{1}{4} - \frac{e^{-2g_3}}{8e^t-4}\right)^{y_2} \cdot \left(\frac{1}{4}\right)^{2 \cdot l-x_1-y_1-1} \cdot n_{\mathrm{amp}}. \end{aligned}$$

Wir erhalten also:

$$\begin{aligned} &\Pr(M_c \neq M_c' \mid M_1 \in \{M_a, M_a'\},\ M_a \neq M_a') \\ \leq\ & \exp(1-(1-e^{-t})g_0) \cdot \left[\left(\frac{1}{4} + \frac{3 \cdot e^{-2g_3}}{8e^t - 4}\right)^{x_1} \cdot \left(\frac{1}{4} - \frac{e^{-2g_3}}{8e^t-4}\right)^{y_1} \cdot \left(\frac{1}{4}\right)^{l-x_1-y_1-1} \right. \\ & \left. + \left(\frac{1}{4} + \frac{3 \cdot e^{-2g_3}}{8e^t - 4}\right)^{x_2} \cdot \left(\frac{1}{4} - \frac{e^{-2g_3}}{8e^t-4}\right)^{y_2} \cdot \left(\frac{1}{4}\right)^{2 \cdot l-x_1-y_1-1} \cdot n_{\mathrm{amp}} \right] \end{aligned}$$

(B) Führt man die Überlegungen aus **(A)** mit dem Blattpaar (b, c) anstelle von (a, c) aus, so erhält man dasselbe Ergebnis (aus Symmetriegründen). Wählt man das Blattpaar (a, b), so erhält man einen Ausdruck derselben Gestalt, aber mit $e^{-2g_1} = 49/50$ statt $e^{-2g_3} = 29/30$ und $g_0 + g_2$ statt g_0. Bei den Blattpaaren (a, a), (b, b) und (c, c) wird $e^{-2g_3} = 29/30$ durch 1 und g_0 durch $g_0 + g_3$ ersetzt. Es folgt jedenfalls für beliebige Blätter x und y:

$$\begin{aligned} &\Pr(M_x \neq M_x' \mid M_1 \in \{M_y, M_y'\},\ M_y \neq M_y') \\ \leq\ & \exp(-(1-e^{-t})g_0) \cdot \left[\left(\frac{1}{4} + \frac{3}{8e^t - 4}\right)^{x_1} \cdot \left(\frac{1}{4} - \frac{29/30}{8e^t-4}\right)^{y_1} \cdot \left(\frac{1}{4}\right)^{l-x_1-y_1-1} \right. \end{aligned}$$

$$+\left(\frac{1}{4}+\frac{3}{8e^t-4}\right)^{x_2}\cdot\left(\frac{1}{4}-\frac{29/30}{8e^t-4}\right)^{y_2}\cdot\left(\frac{1}{4}\right)^{2\cdot l-x_1-y_1-1}\cdot n_{\mathrm{amp}}\Bigg].$$

Wir verwenden im folgenden die Abkürzung

$$R_t(x,y):=\left(\frac{1}{4}+\frac{3}{8e^t-4}\right)^{x}\cdot\left(\frac{1}{4}-\frac{29/30}{8e^t-4}\right)^{y}\cdot\left(\frac{1}{4}\right)^{l-x-y-1}.$$

In unserem Beispiel gilt mit den in Abschnitt 1.3.1 eingeführten Bezeichnungen:

$$\begin{aligned}
&\kappa(a|b)=\kappa(b|a)=1-e^{-lg_1}\approx 0.13\\
&\kappa(c|a)=\kappa(c|b)=1-e^{-lg_3}\approx 0.08\\
&\kappa(a,b|c)=1-e^{-lg_2}+e^{-lg_2}\cdot(1-e^{-lg_1})^2\approx 0.07\\
&\kappa(a,c|b)=\kappa(b,c|a)=(1-e^{-lg_1})\cdot(1-e^{-lg_3})\approx 0.01
\end{aligned}$$

Mit den Überlegungen aus Abschnitt 1.3.1 und wegen $\kappa(|a)+\kappa(a|b)+\kappa(a,b|c)+\kappa(|b)+\kappa(b|a)+\kappa(a,b|c)+\kappa(|c)+\kappa(c|a)+\kappa(a,c|b)\approx 3{,}49$ erhalten wir

$$\begin{aligned}
&\Pr(\overline{\Psi}(Z_t+\delta_\beta)\neq\overline{\Psi}(Z_t)\mid\overline{\Psi}(Z_t+\delta_\alpha)\neq\overline{\Psi}(Z_t))\\
&\quad\lesssim\ 3{,}49\cdot\exp(-(1-e^{-t})g_0)\cdot\max\left\{\left[R_t(x_1,y_1)+R_t(x_2,y_2)\cdot\left(\frac{1}{4}\right)^l\cdot n_{\mathrm{amp}}\right],\right.\\
&\qquad\qquad\left.\left[R_t(x_1',y_1')\cdot\left(\frac{1}{4}\right)^l\cdot n_{\mathrm{amp}}+R_t(x_2',y_2')\right]\right\}
\end{aligned}$$

und

$$\Pr(\overline{\Psi}(Z_t+\delta_\alpha)\neq\overline{\Psi}(Z_t))\lesssim 3{,}49\cdot\left(\frac{1}{4}\right)^{2l-1}\cdot n_{\mathrm{amp}}\cdot e^{-g_0}\cdot(z_{\mathtt{B}}(M_1,j_\alpha)+z_{\mathtt{B}}(M_2,j_\alpha)).$$

Dabei sind x_1', y_1', x_2' und y_2' genauso wie x_1, y_1, x_2 und y_2 definiert, aber mit M_2 anstelle von M_1, und $z_{\mathtt{B}}(M,j)$ ist die Wahrscheinlichkeit, daß an der Stelle j durch Ersetzen einer vorhandenen, rein zufälligen Base durch B die für das Muster M an dieser Stelle passende Base neu hinzukommt oder wegfällt. Dies ist 3/4, falls B mit der an dieser Stelle für M passenden Base übereinstimmt und 1/4 sonst. (Beim Abschätzen von $\Pr(\overline{\Psi}(Z_t+\delta_\beta)\neq\overline{\Psi}(Z_t)\mid\overline{\Psi}(Z_t+\delta_\alpha)\neq\overline{\Psi}(Z_t))$ haben wir auf die notwendige Bedingung verzichtet, daß sich an der Stelle j_β durch die Substitution der zufällig vorhandenen Base überhaupt etwas tut. Dadurch haben wir einiges an Schärfe der Abschätzung verschenkt, aber einige allzu aufwendige Fallunterscheidungen gespart.)

(C) Wir lösen uns nun von der Betrachtung eines einzelnen Paares von Mutationen und integrieren über alle solche Paare, die auf der „unendlichlangen Wurzelkante“ liegen.

Es sei Γ_0 die Menge aller $(v,i,j,\mathtt{B})\in\Gamma$, für die v ein Vorfahr der Wurzel w ist. Um $\int_{\Gamma_0}\sum_\beta\overline{\varphi}_{\alpha\beta}(\mathcal{L}(Z_t))d\mu(\alpha)$ abzuschätzen, müssen wir das Produkt obiger beider Ausdrücke über $g_0\in]0,\infty[$ und verschiedene x_1, y_1 $x_2,\dots$ integrieren. Dazu sei $x_1(j,k)$ dasjenige x_1, welches wir für $\alpha = (v, i, j, \mathtt{B})$ und $\beta = (v, i+k, j-k, \mathtt{B})$ erhalten (mit $j \in \{1, \dots, l\}$ und $k \in \{j-l, \dots, j-1\} \setminus \{0\}$). Genauso verfahren wir mit x_2, y_1, etc.. Außerdem sei v_g der Vorfahr von w mit $\lambda([v_g, w]) = g$. Es folgt dann:

$$\begin{aligned}
&\int_{\Gamma_0} \sum_\beta \overline{\varphi}_{\alpha\beta}(\mathcal{L}(Z_t))d\mu(\alpha) \\
&\leq \frac{1}{4} \sum_{i,j,\mathtt{B}} \sum_{0 \neq k = j-l}^{j-1} \int_0^\infty \Pr(\overline{\Psi}(Z_t + \delta_{(v_g, i+k, j-k, \mathtt{B})}) \neq \overline{\Psi}(Z_t) \mid \overline{\Psi}(Z_t + \delta_{(v_g, i, j, \mathtt{B})}) \neq \overline{\Psi}(Z_t)) \cdot \\
&\qquad \cdot \Pr(\overline{\Psi}(Z_t + \delta_{(v_g, i, j, \mathtt{B})}) \neq \overline{\Psi}(Z_t))dg \qquad (*) \\
&\lesssim \frac{0{,}785 \cdot n \cdot (1/4)^{2l} \cdot n_{amp}}{(2 - e^{-t})} \cdot \\
&\quad \cdot \sum_{j=1}^{l} \sum_{0 \neq k = j-l}^{j-1} \max \left\{ \left[R_t(x_1(j,k), y_1(j,k)) + R_t(x_2(j,k), y_2(j,k)) \cdot \left(\frac{1}{4}\right)^l \cdot n_{\text{amp}} \right], \right. \\
&\qquad \left. \left[R_t(x_1'(j,k), y_1'(j,k)) \cdot \left(\frac{1}{4}\right)^l \cdot n_{\text{amp}} + R_t(x_2'(j,k), y_2'(j,k)) \right] \right\} \\
&\lesssim \frac{3{,}93}{(2 - e^{-t})} \cdot \sum_{j=1}^{l} \sum_{0 \neq k = j-l}^{j-1} \max \left\{ [R_t(x_1(j,k), y_1(j,k)) + R_t(x_2(j,k), y_2(j,k)) \cdot 0{,}05], \right. \\
&\qquad \left. [R_t(x_1'(j,k), y_1'(j,k)) \cdot 0{,}05 + R_t(x_2'(j,k), y_2'(j,k))] \right\}
\end{aligned}$$

(D) Nun fassen wir auch die Mutationspaare auf anderen Kanten ins Auge.

Überträgt man die Rechnung aus **(C)** auf $\int_{\Gamma_2} \sum_\beta \overline{\varphi}_{\alpha\beta}(\mathcal{L}(Z_t))d\mu(\alpha)$, wobei Γ_2 die Menge aller $(v, i, j, \mathtt{B})$ bezeichnet, für die v auf der Kante zwischen w und dem jüngsten gemeinsamen Vorfahren der Blätter a und b liegt, so muß man bei $(*)$ nur über $g \in [0, g_2]$ integrieren. Außerdem benötigt man statt $\kappa(|a) + \ldots + \kappa(a, c|b)$ nur $\kappa(|a) + \kappa(a|b) + \kappa(|b) + \kappa(b|a)$ und man kann generell g_3 durch g_1 ersetzen. Inbesondere steht dann in $R_t(x, y)$ der Bruch 49/50 an Stelle von 29/30. Ähnliches gilt für die übrigen Kanten.

Um schließlich $2 \cdot \int_0^\infty \int_\Gamma \sum_{\beta \in \Gamma_\alpha'} \overline{\varphi}_{\alpha\beta}(\mathcal{L}(Z_t))e^{-2t}d\mu(\alpha)dt$ abzuschätzen, multiplizieren wir die Summe der Abschätzungen für alle Kanten des Baumes mit $2e^{-2t}$ und integrieren das Ergebnis numerisch über $t \in]0, \infty[$. (Das ist sehr leicht machbar, da der Integrand monoton fällt und ab etwa $t = 10$ kaum von einem konstanten Vielfachen von e^{-2t} abweicht. Von beidem kann man sich am einfachsten dadurch überzeugen, daß man den Graphen des Integranden und der Funktionen $R_t(x, y)$ für $x, y \leq 8$ betrachtet.) Ein C-Programm, in welchem die numerische Berechnung durchgefüht wird, findet sich unter `http://stoch.math.uni-frankfurt.de/Leute/metzler/dissprgs.html`. Als Ergebnis erhalten wir:

$$d_{TV}(\mathcal{L}(\overline{\Psi}(\Theta)), \mathcal{L}(\overline{\Psi}(\widetilde{\Theta}))) \leq 2 \cdot \int_0^\infty \int_\Gamma \sum_{\beta \in \Gamma_\alpha'} \overline{\varphi}_{\alpha\beta}(\mathcal{L}(Z_t))e^{-2t}d\mu(\alpha)dt \lesssim 0{,}009$$

Wie schon zu Beginn des Unterabschnitts angemerkt, ist dies eine eher konservative Schranke, die aber in Hinblick auf die in Abschnitt 0.1 beschriebene Anwendung durchaus brauchbar ist.

1.4 Fazit

Sind ein Stammbaum, eine Sequenzlänge und eine Menge von Mustern einer festen Länge gegeben, so sind die Ereignisse des Auftretens von Mustern in den DNA-Sequenzen der Blätter stochastisch abhängig. Satz 1.2 liefert eine Methode, die Vernachlässigbarkeit dieser Abhängigkeiten in Hinblick auf die RAPD-Bandenkonfigurationen der Blätter nachzuweisen. Ein Anwendungsbeispiel legt den Schluß nahe, daß die Abhängigkeiten in typischen RAPD-PCR-Versuchsszenarien nur von sehr geringer Bedeutung sind, falls die Primersequenzen und ihre Komplemente nicht zu viele Überlappungsmöglichkeiten haben.

Kapitel 2

Poisson-Approximation für die Bandenkonfiguration

Gegeben sei ein verwurzelter Binärbaum $\mathbf{T}$, der die Verwandtschaft seiner Blätter $b \in B_{\mathbf{T}}$ beschreibt. Die Topologie und die Astlängen von $\mathbf{T}$ sollen aus den RAPD-Fingerabdrücken der Blätter geschätzt werden. Sollte man dabei Abhängigkeiten berücksichtigen, die dadurch auf den Plan treten, daß verschiedene Banden eine Primerkopie oder ein Primerkomplement gemeinsam haben können? (vgl. Einleitung, Abb. 5)

Um dieser Frage nachzugehen, vergleichen wir zunächst wieder zwei Modelle: eines, welches solche Abhängigikeiten berücksichtigt, und eines, in dem die Banden unabhängig sind.

Im folgenden sei $n := n_{\text{dna}}$ die Länge der DNA, l die Länge des Primers und n_{amp} die Größe des Amplifikationsbereichs, jeweils in der Anzahl der Basen gemessen. Wir bezeichnen die Primersequenz mit $\mathcal{P}$ und ihr Komplement mit $\mathcal{K}$.

2.1 Ein Beispiel zur Problematik

Wir stellen uns folgende Situation vor: Ein Biologe will zwischen zwei Hypothesen über die Verwandtschaft von vier Individuen A, B, C und D entscheiden und führt zu diesem Zweck eine RAPD-PCR-Analyse durch. Hypothese (a) lautet, daß A und B eng verwandt sind, sowie C und D, während nach Hypothese (b) A und C sowie B und D jeweils miteinander eng verwandt sind (siehe Abbildung 2.1).

Nehmen wir an, daß eine der RAPD-Banden nur bei A und B sichtbar wird und eine andere nur bei C und D. Alle anderen Banden seien jeweils bei allen Arten vorhanden. Offensichtlich stützt ein solches Versuchsergebnis Hypothese (a), denn es erscheint auf der Basis von Hypothese (a) plausibel, daß das beobachtete Ergebnis durch nur zwei Mutationen entstanden ist. So könnten zum Beispiel in der Wurzel beide Banden vorhanden sein und dann durch jeweils eine Mutation in der Kante zwischen der Wurzel und dem jüngsten gemeinsamen Vorfahren von C und D zerstört werden. In der Situation von Hypothese (b) müßte hingegen jede der

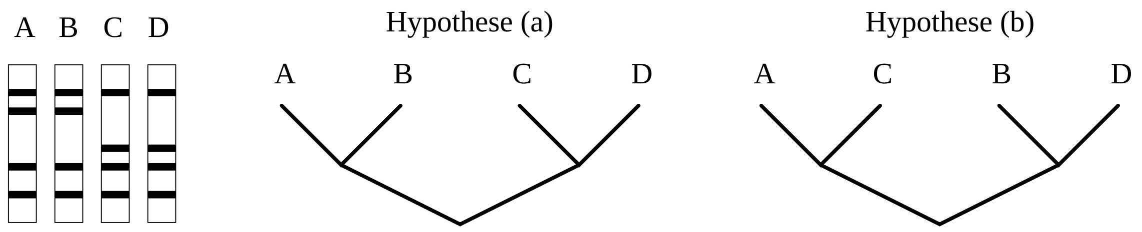


Abbildung 2.1: RAPD-PCR-Fingerabdrücke der Arten A, B, C und D und zwei Hypothesen über ihre Verwandtschaft.

beiden Banden von mindestens zwei Mutationen getroffen werden, um das Versuchsergebnis hervorzubringen.

Man könnte jedoch einwenden, daß dieselben beiden Mutationen, die dafür sorgen, daß die eine Bande nur bei A und B sichtbar ist, auch dafür verantwortlich sein könnten, daß die andere Bande nur bei C und D sichtbar ist, wenn man die in Abbildung 5 der Einleitung dargestellte Möglichkeit einer gekoppelten Evolution der beiden Banden in Betracht zieht. (Mit dieser Möglichkeit könnte man natürlich das Versuchsergebnis auf der Basis von Hypothese (a) mit nur einer Mutation erklären. Eine Differenz von zwei Mutationen, wie sie auftritt, wenn man die die in Abbildung 5 der Einleitung dargestellte Möglichkeit außer Acht läßt, wäre aber sicherlich ein wesentlich überzeugenderes Argument für Hypothese (a).)

Dieses Beispiel ist in mehrfacher Hinsicht recht einfach: Es geht um nur vier Individuen, es werden nur zwei Hypothesen gegenübergestellt, es gibt nur zwei relevante Banden, die nicht bei allen Individuen vorkommen, und die Interpretation der Versuchsergebnisse beruht nur auf einfachen ad-hoc-Argumenten. Tatsächlich geht es oft um eine größere Anzahl an Taxa, über die möglichen Stammbäume werden keine einschränkenden Annahmen gemacht, es sind in der Regel mehr Banden im Spiel und zur Auswertung werden raffinierte, computergestützte Verfahren verwendet. Es ist hinlänglich bekannt, daß statistische Verfahren oft nicht sehr robust gegenüber der Vernachlässigung stochastischer Abhängigkeiten sind. Die Berechnung der Likelihood von Bäumen für gegebene RAPD-Daten ist aber in der Regel zu aufwendig, wenn die Abhängigkeiten zwischen den Banden, die sich durch gemeinsame Primerkopien oder Primerkomplemente ergeben, berücksichtigt werden. Wir werden daher in diesem Kapitel untersuchen, welche Abhängigkeiten allgemein vernachlässigt werden können, und ein Modell entwickeln, das die wesentlichen Abhängigkeiten berücksichtigt und für weitere Untersuchungen – wie etwa die Simulationsstudien in Kapitel 3 – hinreichend praktikabel ist.

2.2 Ein Poisson-Modell für die Bandenkonfigurationen

Ähnlich wie in Kapitel 1 werden wir nun von einem Modell ausgehen, in dem gewisse stochastische Abhängigkeiten vorhanden sind, und diesem eine Poisson-Approximation gegenüberstellen. Abhängigkeiten, die durch Musterüberlappungseffekte zustande kommen, vernachlässigen wir von vornherein: Wir gehen davon aus, daß wir uns in einem Szenario befinden, in dem sich mit den Methoden aus Kapitel 1 zeigen läßt, daß solche Effekte vernachlässigbar sind.

Das Feinmodell für die zufälligen Banden: Wir erinnern an das in Kapitel 1 hergeleitete Poisson-Modell $\widetilde{\Theta}$, das in Hinblick auf die Bandenkonfigurationen zu einer brauchbaren Approximation des Jukes-Cantor-Modells Θ geführt hat. $\widetilde{\Theta}$ erzeugt für jedes $i = 1, \ldots, n$ eine mit dem Baum $\mathbf{T}$ indizierte (d.h. von der Wurzel des Baumes in die Blätter laufende) Markoff-Kette $M_i(.) := \widetilde{W}_i(.)$ auf dem Zustandsraum $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^l$, die der Jukes-Cantor-Übergangsdynamik genügt. Die M_i sind unabhängig, und jedes M_i startet in seiner Gleichgewichtsverteilung, nämlich in der uniformen Verteilung auf $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^l$.

Sei $\mathcal{B}(\widetilde{\Theta})$ die Menge der Banden (s, k), die durch $\widetilde{\Theta}$ irgendwo in den Blättern von $\mathbf{T}$ erzeugt werden. Für jedes $(s, k) \in \mathcal{B}(\widetilde{\Theta})$ sei $U_{(s,k)}$ die Menge derjenigen Blätter, auf denen die Bande (s, k) auftritt. Für $U \subseteq B_{\mathbf{T}}$ sei dann $X_{(U,s,k)}$ die Indikatorvariable für $\{U_{(s,k)} = U\}$. Wir bezeichnen die Gesamtheit X dieser Indikatorvaiablen im Folgenden auch als das Feinmodell.

Unser Ziel ist es, eine Poisson-Approximation für die zufällige Konfiguration $X = (X_\alpha)_{\alpha \in \Gamma}$ zu finden. *Wir bezeichnen dabei den Grundraum* $\{(U, s, k) : \emptyset \neq U \subseteq B_{\mathbf{T}}, (s, k)$ *ist Bande* $\}$, *um den es hier geht, wegen der konzeptionellen Ähnlichkeit dieses Problems zu dem aus Kapitel 1 wieder mit* Γ.

Ein Poisson-Modell: Folgender Ansatz erscheint nun naheliegend: Es sei $Y = (Y_\alpha)_{\alpha \in \Gamma}$ eine Familie unabhängiger Zufallsvariablen, wobei jeweils Y_α zum Parameter $\mathbb{E}X_\alpha$ Poisson-verteilt sei. Y ist also ein Poisson-Prozeß auf der endlichen Menge Γ.

Wir sagen, ein Blatt v habe eine Bande der Länge $k - s \leq n_{\text{amp}}$, falls eine natürliche Zahl $s \leq n$ und eine Teilmenge $U \subseteq B_{\mathbf{T}}$ mit $v \in U$ und $Y_{(U,s,k)} \geq 1$ existieren.

2.2.1 Zur Approximationsgüte des Poisson-Modells

2.2.1.1 Eine obere Schranke für $d_{TV}(X, Y)$

Für $(U, s, k) \in \Gamma$ sei $\Gamma_{(U,s,k)}$ die Menge aller $(V, r, h) \in \Gamma$ mit $\{s, k\} \cap \{r, h\} \neq \emptyset$. $\beta \in \Gamma_\alpha$ bedeutet also, daß die von α und β beschriebenen Banden entweder die Primerkopie oder das Primerkomplement an derselben DNA-Position haben, oder daß eine der beiden Banden die Primerkopie an derselben DNA-Position hat wie die andere das Primerkomplement. Damit ist X_α von $(X_\gamma)_{\gamma \in \Gamma \setminus \Gamma_\alpha}$ stochastisch unabhängig. Aus Theorem 10.A in Barbour *et al.* (1992) (vgl. auch Theorem 1 in Arratia *et al.* (1996)) ergibt sich daher:

$$d_{TV}(X, Y) \leq \sum_{\alpha \in \Gamma,\ \beta \in \Gamma_\alpha} \mathbb{E}X_\alpha \mathbb{E}X_\beta + \sum_{\substack{\alpha,\ \beta \\ \Gamma \ni \alpha \neq \beta \in \Gamma_\alpha}} \mathbb{E}X_\alpha X_\beta \tag{2.1}$$

Es bezeiche im folgenden $g_{s,k}$ die Wahrscheinlichkeit im Modell mit Abhängigkeiten, daß ein Blatt mit einer Bande (s, k) existiert. Diese hängt nicht von s und k ab. Wir setzen $g := g_{(s,k)}$.

Da für (s,k) höchstens ein U mit $X_{(U,s,k)} = 1$ existiert, gilt $\sum_U \mathbb{E}X_{(U,s,k)} = \mathbb{E}\sum_U X_{(U,s,k)} = g_{(s,k)} = g$. Für $\alpha = (U,s,k)$ mit $n_{\mathrm{amp}} < s < n - 2n_{\mathrm{amp}}$ folgt also:

$$\begin{aligned}
\sum_{\beta\in\Gamma_\alpha} \mathbb{E}X_\beta &= \sum_{h=s+l}^{s+n_{\mathrm{amp}}} g_{s,h} + \sum_{r=s-n_{\mathrm{amp}}}^{s-l} g_{r,s} + \sum_{\substack{r=k-n_{\mathrm{amp}}\\ r\neq s}}^{k-l} g_{r,k} + \sum_{h=k+l}^{k+n_{\mathrm{amp}}} g_{k,h} \\
&\approx 4\cdot n_{\mathrm{amp}}\cdot g
\end{aligned}$$

Für sonstige α wird über weniger summiert, was jedoch wegen $n_{\mathrm{amp}} << n$ vernachlässigbar ist. Mit $\sum_{\alpha\in\Gamma} \mathbb{E}X_\alpha \approx n\cdot n_{\mathrm{amp}}\cdot g$ erhalten wir also:

$$\sum_\alpha \sum_{\beta\in\Gamma_\alpha} \mathbb{E}X_\alpha \cdot \mathbb{E}X_\beta \approx \sum_\alpha \mathbb{E}X_\alpha \cdot 4\cdot n_{\mathrm{amp}}\cdot g \approx 4\cdot n\cdot (n_{\mathrm{amp}}\cdot g)^2 \tag{2.2}$$

In ähnlicher Weise berechnen wir die zweite Summe in (2.1). Zunächst einmal gilt:

$$\begin{aligned}
\sum_{\substack{\alpha,\,\beta\\ \Gamma\ni\alpha\neq\beta\in\Gamma_\alpha}} \mathbb{E}X_\alpha X_\beta &= \sum_{\substack{s,k\leq n\\ k\leq s+n_{\mathrm{amp}}}} \sum_{\substack{(r,h)\neq(s,k)\\ \{s,k\}\cap\{r,h\}\neq\emptyset}} \underbrace{\mathbb{E}\sum_{\substack{U,V\in\mathbf{L}\\ U\neq\emptyset\neq V}} X_{(U,s,k)}X_{(V,r,h)}}_{=:\,E_{s,k,r,h}} \\
&= \sum_{\substack{s,k\leq n\\ k\leq s+n_{\mathrm{amp}}}} \left(\sum_{\substack{h=s+l\\ h\neq k}}^{s+n_{\mathrm{amp}}} E_{s,k,s,h} + \sum_{r=s-n_{\mathrm{amp}}}^{s-l} E_{s,k,r,s} + \right. \\
&\qquad \left. + \sum_{\substack{r=k-n_{\mathrm{amp}}\\ r\neq s}}^{k-l} E_{s,k,r,k} + \sum_{h=k+l}^{k+n_{\mathrm{amp}}} E_{s,k,k,h} \right)
\end{aligned}$$

Dabei sei $E_{s,k,r,h} = 0$, falls einer der vier Indizes außerhalb von $\{1,\ldots,n\}$ liegt. Da $\sum_{\substack{U,V\subseteq B_{\mathbf{T}}\\ U\neq\emptyset\neq V}} X_{(U,s,k)}X_{(V,r,h)}$ nur die Werte 0 und 1 annehmen kann, ist $E_{s,k,r,h}$ gerade die Wahrscheinlichkeit, daß es mindestens ein Blatt mit der Bande (s,k) und mindestens ein Blatt mit der Bande (r,h) gibt. Die Summanden $E_{s,k,s,h}$, $E_{s,k,r,k}$, $E_{s,k,r,s}$ und $E_{s,k,k,h}$ hängen nicht von der speziellen Wahl von (s,k,r,h) ab, solange letzteres im jeweiligen Summationsbereich liegt. Außerdem gilt aus Symmetriegründen $E_{s,k,s,h} = E_{s,k,r,k} =: e_1$ und $E_{s,k,r,s} = E_{s,k,k,h} =: e_2$. Analog zu (2.2) folgt:

$$\sum_{\substack{\alpha,\,\beta\\ \Gamma\ni\alpha\neq\beta\in\Gamma_\alpha}} \mathbb{E}X_\alpha X_\beta \approx 2\cdot n\cdot n_{\mathrm{amp}}^2 \cdot (e_1+e_2) \tag{2.3}$$

Dabei bedeutet „$\approx$“ ebenso wie in (2.2) „gleich bis auf einen Faktor, der höchstens um $\frac{n_{\mathrm{amp}}}{n} + \frac{1}{n_{\mathrm{amp}}}$ von 1 abweicht“.

Damit ist die Suche nach einer oberen Abschätzung von $d_{TV}(X,Y)$ auf die Berechnung von g, e_1 und e_2 zurückgeführt.

2.2.2 Die Berechnung von g, e_1 und e_2.

Die in Abschnitt 2.2.1.1 definierten Größen g, e_1 und e_2 sind für den Rest dieses Kapitels sehr wichtig. Daher untersuchen wir nun, wie man g, e_1 und e_2 für einen gegebenen Baum $\mathbf{T}$ effizient berechnen kann.

Für die Berechnung von $g = g_{(s,k)}$ sei

$$Z_v := \sharp\{j : M_{sj}(v) = \mathcal{P}_j\} + \sharp\{j : M_{kj}(v) = \mathcal{K}_j\},$$

wobei $M_{ij}(v)$, $\mathcal{P}_j$ und $\mathcal{K}_j$ jeweils die j-te Base im jeweiligen Muster bezeichnet. Damit ist Z eine mit dem Baum $\mathbf{T}$ indizierte Markoff-Kette auf $\{0, 1, \ldots, 2l\}$. Ist v' der Vaterknoten eines Knotens v und $a, b \in \{0, \ldots, 2l\}$, so sei $p_v(a, b) := \Pr(Z_v = b \mid Z_{v'} = a)$.

Für einen beliebigen Knoten v sei $L_v = \{v\}$, falls v ein Blatt ist. Sonst sei L_v die Menge der Blätter, die von v abstammen. $B_v \subset \{1, \ldots, n\}^2$ sei die Menge aller in L_v vorkommenden Banden. Für $a \in \{0, \ldots, 2l\}$ sei

$$w_1^v(a) := \Pr((s, k) \in B_v \mid Z_v = a) = \Pr(\exists u \in L_v : Z_u = 2l \mid Z_v = a).$$

Ist v ein Blatt, so gilt offensichtlich $w_1^v(a) = \delta_{a,2l}$.

Sei λ die Länge der Kante zwischen den beiden Knoten. Wir können $p_v(a, b)$ berechnen, indem wir über die Anzahl c der Basen, die in beiden Knoten stimmen, partitionieren. Wir erhalten dann:

$$\begin{aligned} p_v(a,b) \quad = \quad & \sum_{c=\max\{0,a+b-2l\}}^{\min\{a,b\}} \binom{a}{c}\binom{2l-a}{b-c} \cdot \left(\frac{1}{4}\right)^{2l} \cdot \left(1+3e^{-\lambda}\right)^{c} \cdot \\ & \cdot \left(3-3e^{-\lambda}\right)^{a-c} \cdot \left(1-e^{-\lambda}\right)^{b-c} \cdot \left(3+e^{-\lambda}\right)^{2l-a-b+c} \end{aligned}$$

Es sei v ein Knoten mit zwei Nachkommen x und y. Es folgt dann mit der Markoff-Eigenschaft von Z:

$$\begin{aligned} w_1^v(a) \quad &= \quad \Pr((s,k) \in B_x \mid Z_v = a) + \Pr((s,k) \in B_y \mid Z_v = a) \\ &\qquad - \Pr((s,k) \in B_x \cap B_y \mid Z_v = a) \\ &= \quad \sum_{b=0}^{2l} \Big(p_x(a,b) \cdot w_1^x(b) + p_y(a,b) \cdot w_1^y(b) \Big) \\ &\qquad - \left(\sum_{c=0}^{2l} p_x(a,c) \cdot w_1^x(c) \right) \cdot \left(\sum_{d=0}^{2l} p_y(a,d) \cdot w_1^y(d) \right) \end{aligned}$$

Für die Wurzel ρ gilt:

$$g_{(s,k)} = \sum_{a=0}^{2l} \binom{2l}{a} \cdot \left(\frac{1}{4}\right)^{a} \cdot \left(\frac{3}{4}\right)^{2l-a} \cdot w_1^{\rho}(a)$$

Mit diesen Formeln können wir $g = g_{(s,k)}$ berechnen, indem wir $w_1^\cdot(.)$ von den Blättern ausgehend bis zur Wurzel, den Baum in negativer Zeitrichtung durchlaufend, berechnen. Die Laufzeit dieses Verfahrens ist offensichtlich linear in der Anzahl der Blätter und polynomial in der Länge des Primers.

e_1 und e_2 lassen sich ähnlich wie g berechnen. Auch hier konstruieren wir jeweils eine mit dem Baum $\mathbf{T}$ indizierte Markoff-Kette Z' bzw. Z'', so daß wir an $(Z'_u)_{u \in B_\mathbf{T}}$ bzw. $(Z''_u)_{u \in B_\mathbf{T}}$ ablesen können, ob die Ereignisse eintreten, deren Wahrscheinlichkeiten e_1 bzw. e_2 angeben. Wir können dann wieder die Markoff-Eigenschaft von Z' bzw. Z'' benutzen, um e_1 bzw. e_2 zu berechnen.

Zur Berechnung von $e_1 = E_{s,k,s,h}$ verwenden wir folgende mit $\mathbf{T}$ indizierte Markoff-Kette auf $\{0, 1, \ldots, l\}^3$:

$$Z'_v := (\,\sharp\{j : M_{sj}(v) = \mathcal{P}_j\}\,,\,\sharp\{j : M_{kj}(v) = \mathcal{K}_j\}\,,\,\sharp\{j : M_{hj}(v) = \mathcal{K}_j\}\,)$$

Ist v' der Vaterknoten des Knotens v, so gilt

$$\Pr(Z'_v = (a, b, c) \,|\, Z'_{v'} = (a', b', c')) = q_v(a', a) \cdot q_v(b', b) \cdot q_v(c', c),$$

wobei $q_v(.,.)$ genauso definiert ist wie $p_v(.,.)$, aber mit l statt $2l$.

Außerdem sei

$$w_2^v(a, b, c) := \Pr((s, k) \in B_v \wedge (s, h) \in B_v \,|\, Z'_v = (a, b, c)).$$

Ist v ein Blatt, so folgt $w_2^v(a, b, c) = \delta_{(a,b,c),(l,l,l)}$.

Sei nun wieder v der Vaterknoten der Knoten x und y. Für die Formulierung der Rekursionsgleichung zur Berechnung von $w_2^\cdot(.,.,.)$ verwenden wir die Abkürzung $\Pr_{abc}(A) := \Pr(A \,|\, Z'_v = (a, b, c))$. Mit der Markoff-Eigenschaft von Z' und der Einschluß-Ausschluß-Formel erhalten wir:

$$\begin{aligned}
w_2^v(a, b, c) \;=\; & \Pr_{abc}[(s, k) \in B_x \wedge (s, h) \in B_x] \;+\; \Pr_{abc}[(s, k) \in B_x \wedge (s, h) \in B_y] \\
& + \Pr_{abc}[(s, k) \in B_y \wedge (s, h) \in B_x] \;+\; \Pr_{abc}[(s, k) \in B_y \wedge (s, h) \in B_y] \\
& - \Pr_{abc}[(s, k) \in B_x \cap B_y \wedge (s, h) \in B_y] - \Pr_{abc}[(s, k) \in B_x \wedge (s, h) \in B_x \cap B_y] \\
& - \Pr_{abc}[(s, k) \in B_x \cap B_y \wedge (s, h) \in B_x] - \Pr_{abc}[(s, k) \in B_y \wedge (s, h) \in B_x \cap B_y] \\
& + \Pr_{abc}[(s, k) \in B_x \cap B_y \wedge (s, h) \in B_x \cap B_y] \\
\;=\; & \Pr_{abc}[(s, k) \in B_x \wedge (s, h) \in B_x] \;+\; \Pr_{abc}[(s, k) \in B_x] \;\cdot\; \Pr_{abc}[(s, h) \in B_y] \\
& + \Pr_{abc}[(s, k) \in B_y] \;\cdot\; \Pr_{abc}[(s, h) \in B_x] \;+\; \Pr_{abc}[(s, k) \in B_y \wedge (s, h) \in B_y] \\
& - \Pr_{abc}[(s, k) \in B_x] \;\cdot\; \Pr_{abc}[(s, k) \in B_y \wedge (s, h) \in B_y] \\
& - \Pr_{abc}[(s, k) \in B_x \wedge (s, h) \in B_x] \;\cdot\; \Pr_{abc}[(s, h) \in B_y] \\
& - \Pr_{abc}[(s, k) \in B_x \wedge (s, h) \in B_x] \;\cdot\; \Pr_{abc}[(s, k) \in B_y] \\
& - \Pr_{abc}[(s, h) \in B_x] \;\cdot\; \Pr_{abc}[(s, k) \in B_y \wedge (s, h) \in B_y] \\
& + \Pr_{abc}[(s, k) \in B_x \wedge (s, h) \in B_x] \;\cdot\; \Pr_{abc}[(s, k) \in B_y \wedge (s, h) \in B_y]
\end{aligned}$$

Für $z \in \{x, y\}$ gilt:

$$\begin{aligned}
\Pr_{abc}((s,k) \in B_z) &= \sum_{d=0}^{2l} p_z(a+b, d) \cdot w_1^z(d) \\
\Pr_{abc}((s,h) \in B_z) &= \sum_{d=0}^{2l} p_z(a+c, d) \cdot w_1^z(d) \\
\Pr_{abc}((s,k) \in B_z \wedge (s,h) \in B_z) &= \sum_{a',b',c'=0}^{l} q_z(a,a') \cdot q_z(b,b') \cdot q_z(c,c') \cdot w_2^z(a',b',c')
\end{aligned}$$

Wir berechnen dann

$$\begin{aligned}
e_1 = E_{(s,k,s,h)} &= \sum_{a,b,c=0}^{l} \Pr(Z'_\rho = (a,b,c)) \cdot w_2^\rho(a,b,c) \\
&= \sum_{a,b,c=0}^{l} \binom{l}{a}\binom{l}{b}\binom{l}{c} \cdot \left(\frac{1}{4}\right)^{a+b+c} \cdot \left(\frac{3}{4}\right)^{3l-a-b-c} \cdot w_2^\rho(a,b,c),
\end{aligned}$$

indem wir, ausgehend von den Blättern, iterativ $w_2^v(a,b,c)$ für jeden Knoten v und jedes Zahlentripel $(a,b,c) \in \{0,1,\ldots,l\}^3$ berechnen.

Für die Berechnung von $e_2 = E_{s,k,k,h}$ setzen wir

$$\begin{aligned}
Z''_v \quad := \quad (& \sharp\{j : M_{sj}(v) = \mathcal{P}_j\}, \\
& \sharp\{j : M_{hj}(v) = \mathcal{K}_j\}, \\
& \sharp\{j : M_{kj}(v) = \mathcal{K}_j \neq \mathcal{P}_j\}, \\
& \sharp\{j : M_{kj}(v) = \mathcal{P}_j \neq \mathcal{K}_j\}, \\
& \sharp\{j : M_{kj}(v) = \mathcal{P}_j = \mathcal{K}_j\})
\end{aligned}$$

und

$$w_2^v(a,b,c,d,e) := \Pr(\{(s,k),(k,h)\} \subset B_v \mid Z''_v = (a,b,c,d,e)).$$

Wir gehen dann genauso vor wie bei der Berechnung von g und e_1, wenn auch mit mehr Rechenaufwand. (Dabei ist zu beachten, daß die Koordinaten von Z'' im Gegensatz zu den Koordinaten von Z' stochastisch abhängig sind.)

Wir erhalten damit also auch für e_1 und e_2 Berechnungsverfahren, deren Laufzeit linear in der Anzahl der Blätter und polynomial in l ist. In biologisch sinnvollen Szenarien liegt l etwa bei 10, und g, e_1 und e_2 sind in akzeptabler Zeit berechenbar (vgl. Abschnitt 2.2.2.1).

2.2.2.1 Ein Beispiel

Um uns einen Eindruck von g, e_1 und e_2 zu verschaffen, betrachten wir den in Abbildung 2.2 dargestellten Stammbaum. Wir gehen von einem Primer der Länge 10, einer DNA der Länge

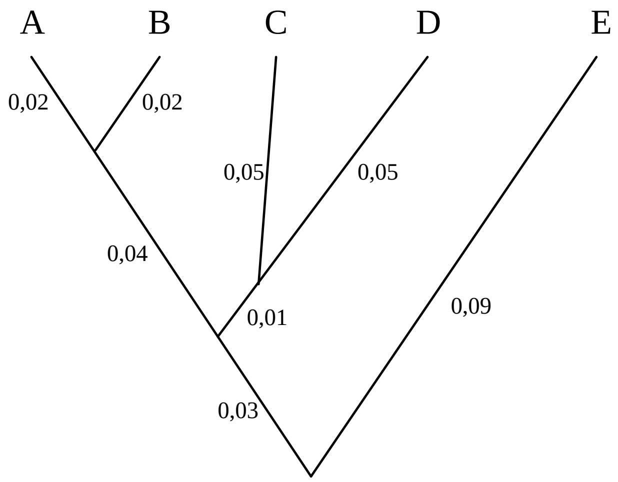


Abbildung 2.2: Ein Stammbaum von fünf Taxa. Ein Label λ an einer Kante bedeutet, daß auf der Kante pro Basenplatz eine zum Parameter λ Poisson-verteilte Anzahl an Mutationen einwirken.

10^9 und von einem Amplifikationsbereich von 3000 Basenpaaren aus. Die erwartete Anzahl an Banden eines DNA-Strangs liegt also bei $10^9 \cdot (1/4)^{20} \cdot 3000 \approx 2,7$.

Zunächst einmal schätzen wir die Größenordnungen von g, e_1 und e_2 mit einer grob heuristischen Überschlagsrechnung, um mehr Gefühl dafür zu bekommen, wovon g, e_1 und e_2 abhängen. Die Aussagen der nächsten Absätze sind in diesem Sinne zu verstehen. Wer sich nur für die Ergebnisse interessiert, kann die nächsten Absätze überspringen.

Greift man rein zufällig zwei Blätter G und H heraus, so unterscheiden sich bei den beiden die Basen an ungefähr jeder zehnten Position, da die mittlere Distanz zwischen zwei Blättern in der Größenordnung von 1/10 liegt. Die Wahrscheinlichkeit, daß ein in G vorhandenes 10er-Muster in H noch intakt ist, ist also ungefähr $(9/10)^{10} \approx 0{,}35$. Die erwartete Anzahl an verschiedenen in den Blättern vorhandenen 10-er-Mustern an einem Site ist also ungefähr $1 + 4 \cdot 0{,}35 \approx 2{,}5$.

Die Wahrscheinlichkeit, daß an 20 ausgewählten Positionen die Basen bei G und H übereinstimmen, ist ungefähr $(9/10)^{20} \approx 0{,}12$. Eine Bande, die in einem Blatt vorliegt, ist also in ungefähr $4 \cdot 0{,}12 \approx 0{,}5$ weiteren Blättern vorhanden. Für $(s,k) \in \{l,\dots,n\}^2$ mit $k - s \in \{1,\dots,n_{\mathrm{amp}}\}$ gibt es also ungefähr $5/1{,}5 \approx 3,3$ verschiedene Musterpaare $(M_s(v), M_k(v))$ mit $v \in B_{\mathbf{T}}$. Die Wahrscheinlichkeit, daß eines dieser Musterpaare eine Bande ist, ist also ungefähr $3{,}3 \cdot (1/4)^{20} \approx 3 \cdot 10^{-12} =: \widetilde{g}$.

In einem Blatt gebe es eine Bande (s,k). Die erwartete Anzahl an Blättern, in denen an der Stelle s die Primersequenz $\mathcal{P}$ vorliegt, ist dann ungefähr $5/2{,}5 = 2$. Die erwartete Anzahl an verschiedenen Sequenzen an einer ausgewählten Stelle $h \neq k$ mit $h - s \in \{l,\dots,n_{\mathrm{amp}}\}$ in den beiden Blättern ist dann ungefähr $1 + (1 - 0{,}35) = 1{,}65$. Die Wahrscheinlichkeit, daß es in den Blättern eine Bande (s,h) gibt, bedingt darauf, daß es eine Bande (s,k) gibt, ist also ungefähr $1{,}65 \cdot (1/4)^{10}$. Wir schätzen also e_1 durch $\widetilde{e}_1 := \widetilde{g} \cdot 1{,}65 \cdot (1/4)^{10} \approx 0{,}5 \cdot 10^{-7}$.

Sei a die Anzahl der Banden, die zwischen $\mathcal{P}$ und $\mathcal{K}$ übereinstimmen. Wir gehen davon aus, daß dies nicht extrem viele sind, sagen wir $a \leq 5$. Angenommen, in einem Blatt G existiert eine Bande (s, k). Damit ein Blatt mit einer Bande (r, s) existieren kann (für gewähltes r), muß an der Stelle s das Muster $\mathcal{P}$ nach $\mathcal{K}$ mutieren und in einem der Blätter, in denen an der Stelle s dann $\mathcal{K}$ vorliegt, muß es an Stelle r ein $\mathcal{P}$ geben. Die Wahrscheinlichkeit dafür, e_2, ist also etwa von der Größenordnung $(1/30)^{(10-a)} \cdot e_1 \leq (1/30)^5 \cdot e_1$, aber jedenfalls wesentlich kleiner als e_1. Da e_2 in der uns interessierenden Abschätzung (2.1) nur in Summe mit e_1 auftritt, wollen wir es mit dieser Feststellung bewenden lassen.

Wir schätzen also

$$\sum_{\alpha\in\Gamma,\ \beta\in\Gamma_\alpha} \mathbb{E}X_\alpha \mathbb{E}X_\beta \approx 4 \cdot n \cdot (n_{\text{amp}} \cdot g)^2 \overset{(?)}{\approx} 4 \cdot n \cdot (n_{\text{amp}} \cdot \widetilde{g})^2 \approx 3 \cdot 10^{-7}$$

$$\sum_{\substack{\alpha,\,\beta \\ \Gamma\ni\alpha\neq\beta\in\Gamma_\alpha}} \mathbb{E}X_\alpha X_\beta \approx 2 \cdot n \cdot n_{\text{amp}}^2 \cdot e_1 \overset{(?)}{\approx} 2 \cdot n \cdot n_{\text{amp}}^2 \cdot \widetilde{e}_1 \approx 0{,}09.$$

In der Tat gilt:

$$\sum_{\alpha\in\Gamma,\ \beta\in\Gamma_\alpha} \mathbb{E}X_\alpha \mathbb{E}X_\beta \approx 4 \cdot n \cdot (n_{\text{amp}} \cdot g)^2 \approx 4{,}18 \cdot 10^{-7}$$

$$\sum_{\substack{\alpha,\,\beta \\ \Gamma\ni\alpha\neq\beta\in\Gamma_\alpha}} \mathbb{E}X_\alpha X_\beta \approx 2 \cdot n \cdot n_{\text{amp}}^2 \cdot e_1 \approx 0{,}0916$$

Diese Werte habe ich mit dem weiter oben beschriebenen Verfahren berechnet. Den kommentierten C++-Quellcode des dazu geschriebenen Computerprogramms habe ich im Internet unter `http://stoch.math.uni-frankfurt.de/Leute/metzler/dissprgs.html` abgelegt. Das Programm benötigte zur Berechnung der beiden Werte auf einer HP-Workstation ebenso wie auf einem PC mit einem Intel Pentium 100 Prozessor ungefähr eine Minute. Mit dem Programm lassen sich die Werte g und e_1 für beliebige verwurzelte, mit Kantenlängen versehene Binärbäume berechnen.

Verlängert man die Kanten des Baumes um den Faktor 10, so erhält man die Ergebnisse $7{,}743 \cdot 10^{-7}$ und $0{,}0798$, verkürzt man hingegen die Kanten mit dem Faktor $1/10$ so erhält man $6{,}17 \cdot 10^{-8}$ und $0{,}0265$.

Da man in der Biologie meistens auf einem Signifikanzniveau von mindestens 95% arbeitet, ist die Information, daß der Totalvariationsabstand zwischen dem Feinmodell und dem handhabbaren Modell kleiner gleich 9% ist, nicht besonders befriedigend. Das Beispiel zeigt also, daß mindestens eine der folgenden Fragen zu bejahen ist:

- Ist der Unterschied zwischen X und Y in biologisch relevanten Szenarien zu groß, als daß man ihn vernachlässigen dürfte?

- Ist der Totalvariationsabstand zwischen X und Y eine zu feine Maßeinheit für die biologisch relevanten Fragestellungen?

- Ist die Abschätzung (2.1) zu grob?

In Abschnitt 2.2.2.2 werden wir sehen, daß es nicht an der Grobheit der Abschätzung liegt: Der Totalvariationsabstand zwischen X und Y ist in obigem Beispiel tatsächlich größer als es wünschenswert wäre. Natürlich ist der Totalvariationsabstand zwischen X und Y eine für viele Fragestellungen zu feine Maßeinheit. Wir wollen aber in diesem Kapitel relativ allgemein bleiben und uns nicht auf zu spezielle Anwendungen der RAPD-PCR beschränken. Wir werden daher in Abschnitt 2.3 ein Modell untersuchen, in dem bestimmte Abhängigkeiten zwischen den Banden berücksichtigt werden, deren Vernachlässigung bei Y offensichtlich einen großen Anteil zum Totalvariationsabstand zwischen X und Y beiträgt. In Kapitel 3 werden wir dann eine bestimmte Klasse von Fragestellungen ins Auge fassen, für die sich die Abhängigkeiten in vielen Fällen als vernachlässigbar herausstellen werden.

2.2.2.2 Eine untere Schranke für $d_{TV}(X,Y)$

Um der am Ende von Abschnitt 2.2.2.1 aufgeworfenen Frage, ob die Abschätzung (2.1) zu grob sei, nachzugehen, suchen wir nun nach einer unteren Abschätzung von $d_{TV}(X,Y)$. Wegen $d_{TV}(X,Y) = \sup_{f:\{0,1,\dots\}^\Gamma\to[0,1]} |\mathbb{E}f(X) - \mathbb{E}f(Y)|$ erhalten wir für jedes solche f die Abschätzung $d_{TV}(X,Y) \geq |\mathbb{E}f(X) - \mathbb{E}f(Y)|$. Wir suchen also ein f, so daß die rechte Seite dieser Gleichung möglichst groß wird. Die bisherigen Überlegungen legen nahe, f so zu definieren, daß $f(X)$ die Indikatorfunktion dafür wird, daß ein Paar $(\alpha,\beta) \in \Gamma^2$ mit $\alpha \neq \beta \in \Gamma_\alpha$ und $X_\alpha = X_\beta = 1$ existiert. Wir setzen also für $\xi \in \{0,1,2,\dots\}^\Gamma$:

$$f(\xi) := \begin{cases} 1 & \text{falls } \exists(\alpha,\beta) \in \Gamma^2,\ \alpha \neq \beta \in \Gamma_\alpha:\ \xi_\alpha \cdot \xi_\beta \neq 0 \\ 0 & \text{sonst} \end{cases}$$

Es gilt dann

$$\mathbb{E}f(Y) \leq \sum_{\substack{\{\alpha,\,\beta\} \\ \Gamma\ni\alpha\neq\beta\in\Gamma_\alpha}} Y_\alpha Y_\beta = \frac{1}{2} \sum_{\substack{\alpha,\,\beta \\ \Gamma\ni\alpha\neq\beta\in\Gamma_\alpha}} Y_\alpha Y_\beta \approx 2 \cdot n \cdot (n_{\text{amp}} \cdot g)^2.$$

$\mathbb{E}f(X)$ vernünftig abzuschätzen, ist etwas diffiziler. Sei dazu $\Lambda := \{\{\alpha,\beta\} \mid \Gamma \ni \alpha \neq \beta \in \Gamma_\alpha\}$ und für $\{\alpha,\beta\} \in \Lambda$ sei $V_{\{\alpha,\beta\}} := X_\alpha X_\beta$ und $W := \sum_{\Phi\in\Lambda} V_\Phi$. Offensichtlich gilt:

$$\mathbb{E}f(X) = \Pr(\exists_{\Phi\in\Lambda}\ V_\Phi \neq 0) = \Pr(W > 0)$$

Wenn die Abhängigkeiten zwischen den V_Φ nicht zu stark sind, dann ist W annähernd Poisson-verteilt zum Parameter $\lambda_W := \mathbb{E}W = \sum_\Phi \mathbb{E}V_\Phi = \frac{1}{2}\sum \mathbb{E}X_\alpha X_\beta \approx n \cdot n_{\text{amp}}^2 \cdot (e_1 + e_2)$ und es gilt $\mathbb{E}f(X) \approx 1 - e^{-\lambda_W} \approx 1 - \exp(-n \cdot n_{\text{amp}}^2 \cdot (e_1 + e_2))$. Für unser Beispiel in Abschnitt 2.2.2.1 würde das bedeuten:

$$d_{TV}(X,Y) \geq |\mathbb{E}f(X) - \mathbb{E}f(Y)| \approx 1 - e^{-0{,}045} - 2{,}09 \cdot 10^{-7} \approx 4{,}4\%$$

Dabei bleibt noch zu klären, wie gut die Verteilung von W durch die Poisson-Verteilung approximiert wird. Wir verwenden dazu wieder die Chen-Stein-Methode (genauer gesagt Theorem 10.A in Barbour *et al.*, 1992): Für $\{\alpha,\beta\} \in \Lambda$ sei $\Lambda_{\{\alpha,\beta\}} := \{\{\gamma,\delta\} \in \Lambda \mid \{\gamma,\delta\} \cap (\Gamma_\alpha \cup \Gamma_\beta) \neq \emptyset\}$. Dann sind V_Φ und V_Ψ offensichtlich stochastisch unabhängig, falls $\Psi \notin \Lambda_\Phi$ ist. Also gilt:

$$\left|\Pr(W>0) - (1-e^{-\lambda_W})\right| \leq d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda_W)) \leq \sum_{\Phi\in\Lambda,\,\Psi\in\Lambda_\Phi} \mathbb{E}V_\Phi \mathbb{E}V_\Psi + \sum_{\substack{\Phi,\Psi \\ \Lambda\ni\Phi\neq\Psi\in\Lambda_\Phi}} \mathbb{E}V_\Phi V_\Psi$$

Beim Abschätzen dieser beiden Summen verfahren wir analog zu Abschnitt 2.2.1.1. Die erste Summe ist kleiner gleich $4\cdot n\cdot n_{\mathrm{amp}}^4\cdot(e_1+2\cdot e_2)^2$, die zweite kleiner gleich $4\cdot n\cdot n_{\mathrm{amp}}^3\cdot(2\cdot e_3+e_4)$. Dabei ist e_3 für fest gewählte s,r,k mit $0<s<r<k<s+n_{\mathrm{amp}}<n$ die Wahrscheinlichkeit, daß jeweils mindestens ein Blatt mit einer Bande (s,k), einer Bande (r,k) und einer Bande (s,r) existieren. e_4 ist für fest gewählte s,r,k,h mit $0<s<r<k<h<s+n_{\mathrm{amp}}<n$ die Wahrscheinlichkeit, daß jeweils mindestens ein Blatt mit einer Bande (s,k), einer Bande (r,k) und einer Bande (r,h) existieren. Wir gehen wieder davon aus, daß keine besondere Ähnlichkeit zwischen $\mathcal{P}$ und $\mathcal{K}$ besteht und daß e_3 daher im Vergleich zu e_4 zu vernachlässigen ist. Schätzen wir e_4 dann grob durch g^2 ab, so erhalten wir für die zweite Summe die obere Abschätzung $8\cdot n\cdot n_{\mathrm{amp}}^3\cdot g^2$. Für unser konkretes Beispiel erhalten wir also $|\Pr(W>0)-(1-e^{-\lambda_W})| \leq$ 0,25% und damit 4,1% $\lesssim d_{TV}(X,Y) \lesssim$ 9,16%. Der Totalvariationsabstand zwischen X und Y liegt also in unserem Beispiel in einem Bereich, der für unsere Fragestellung nicht akzeptabel erscheint.

2.3 Ein Poisson-Cluster-Modell für die Bandenkonfigurationen

Wir werden jetzt ein Modell konstruieren, welches die Abhängigkeiten berücksichtigt, die für den größten Teil des Totalvariationsabstandes zwischen X und Y verantwortlich sind. Schwächere Abhängigkeiten werden vernachässigt. Wie sich zeigen wird, spielen diese in relevanten Szenarien eine so kleine Rolle, daß der Totalvariationsabstand zwischen dem hier zu konstruierenden Modell und dem Feinmodell sehr klein wird. Die übrigbleibenden Abhängigkeiten sind recht überschaubar.

Hinter diesem Modell steckt die Vorstellung, daß sich gegenseitig begünstigende Ereignisse – in unserem Fall das Erscheinen von Banden in bestimmten Blättern an gewissen Positionen – in „Klumpen" auftreten und daß diese Klumpen einigermaßen unabhängig voneinander sind[1].

Wir betrachten die von $M_.(.)$ induzierte (zufällige) *Musterkonfiguration (in den Blättern)*

$$G := \sum_{v\in B_{\mathbf{T}}} \sum_{s=1}^{n} \delta_{(v,s,M_s(v))} I_{\{\mathcal{P},\mathcal{K}\}}(M_s(v))$$

auf $B_{\mathbf{T}} \times \{1,\ldots,n\} \times \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}^l$. Gewisse Atome von G (nämlich solche der Art $(v,s,\mathcal{P})$, $(v,k,\mathcal{K})$ mit $k-s\in\{l,\ldots,n_{\mathrm{amp}}\}$) bilden zusammen eine Bande, andere (nämlich solche der

[1] Der Begriff „Klumpen" orientiert sich am Titel des Buches „Poisson Clumping Heuristic", Aldous (1989).

Art $(v,s,\mathcal{M})$, $(v',s,\mathcal{M})$) beeinflussen sich gegenseitig. Insgesamt erzeugt dies eine Abhängigkeitsstruktur innerhalb von G, die wir mit folgender Begriffsbildung beschreiben:

Zwei Tripel $(v,s,\mathcal{M})$ und $(v',s',\mathcal{M}')$ aus $B_{\mathbf{T}} \times \{1,\ldots,n\} \times \{\mathcal{P},\mathcal{K}\}$ heißen *befreundet*, falls eine der folgenden Bedingungen gilt:

(a) $s = s'$

(b) $v = v'$, $s + l \leq s' \leq s + n_{\text{amp}}$, $\mathcal{M} = \mathcal{P}$, $\mathcal{M}' = \mathcal{K}$
(D. h. es gibt eine Bande (s,s') auf $v = v'$.)

(c) $v = v'$, $s' + l \leq s \leq s' + n_{\text{amp}}$, $\mathcal{M}' = \mathcal{P}$, $\mathcal{M} = \mathcal{K}$
(D. h. es gibt eine Bande (s',s) auf $v = v'$.)

Eine Teilmenge χ von $B_{\mathbf{T}} \times \{1,\ldots,n\} \times \{\mathcal{P},\mathcal{K}\}$ heißt *Clique*, wenn jedes Paar von Elementen von χ durch eine Folge in χ verbunden werden kann, bei der aufeinanderfolgende Glieder jeweils befreundet sind, und die Projektion von χ auf $\{1,\ldots,n\}$ mehr als ein Site enthält. Eine Clique muß also mindestens eine Bande umfassen. Sei Ξ die Menge aller Cliquen.

Wir definieren den Abhängigkeitsbereich $A(\chi)$ einer Clique χ durch:

$$A(\chi) := \{(v,s,\mathcal{M}) : (v,s,\mathcal{M}) \text{ ist mit einem Element von } \chi \text{ befreundet}\}$$

Ein *Klumpen* (in G) ist eine Clique χ, die im Träger von G enthalten und in diesem maximal ist, d.h. es muß $\operatorname{supp}(G) \cap A(\chi) = \chi$ gelten. Die *Konfiguration der Klumpen (in G)* sei:

$$Q := \sum_{\chi \text{ ist Klumpen in } G} \delta_\chi$$

Q ist also ein zufälliges Zählmaß auf der Menge Ξ der Cliquen und für $\chi \in \Xi$ ist $Q(\chi)$ die Indikatorvariable für das Ereignis „*für alle $(v,s,\mathcal{M}) \in \chi$ gilt $M_s(v) = \mathcal{M}$, und für alle mit einem Element von χ befreundeten $(v',s',\mathcal{M}') \notin \chi$ gilt $M_{s'}(v') \neq \mathcal{M}'$*“. Ein Beispiel ist in Abbildung 2.3 gegeben.

2.3.1 Vernachlässigbare Abhängigkeiten zwischen den Klumpen

Offensichtlich gibt es Abhängigkeiten zwischen den Atomen von Q: Gilt $Q(\chi) = 1$ für eine Clique χ, so folgt zum Beispiel $Q(\varphi) = 0$ für alle $\varphi \neq \chi$ mit $\varphi \cap \chi \neq \emptyset$. Wir stellen diesem Szenario ein Modell ohne Abhängigkeiten gegenüber: Sei K eine Poissonsche Konfiguration auf Ξ mit Intensität $\mathbb{E}Q$, d. h. $(K(\chi))_{\chi \in \Xi}$ ist eine Familie unabhängiger $\{0,1,2,\ldots\}$-wertiger Zufallsvariabler, und $K(\chi)$ ist jeweils zum Parameter $\mathbb{E}Q(\chi)$ Poisson-verteilt.

Wir verwenden diesmal Theorem 10.B aus Barbour *et al.* (1992). Wir konstruieren dazu für jedes $\chi \in \Xi$ eine $\{0,1\}$-wertige Zufallsvariable R_χ mit $\mathcal{L}(R_\chi + \delta_\chi) = \mathcal{L}(Q \mid Q(\chi) = 1)$. Wir betrachten dazu die bedingte Veteilung der Familie baumindizierter Markoff-Ketten $M = (M_i(.))_i$, gegeben daß Q das Atom δ_χ besitzt, und konstruieren eine Familie mit $\mathbf{T}$ indizierter Markoff-Ketten $\{M_1'(.),\ldots,M_n'(.)\}$ auf $\{\mathtt{A},\mathtt{G},\mathtt{C},\mathtt{T}\}^l$, die dieser Verteilung genügt. Wegen der

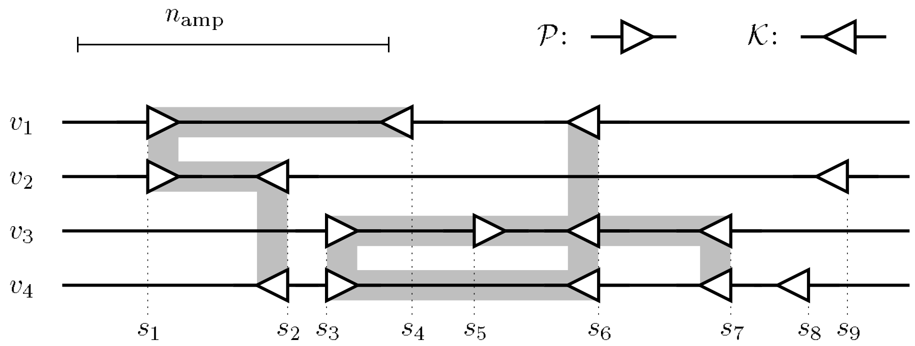


Abbildung 2.3: In dieser Musterkonfiguration gibt es die beiden Klumpen $\{(v_1, s_1, \mathcal{P}), (v_1, s_4, \mathcal{K}), (v_2, s_1, \mathcal{P}), (v_2, s_2, \mathcal{K}), (v_4, s_4, \mathcal{K})\}$ *und* $\{(v_1, s_6, \mathcal{K}), (v_3, s_3, \mathcal{P}), (v_3, s_5, \mathcal{P}), (v_3, s_6, \mathcal{K}), (v_3, s_7, \mathcal{K}), (v_4, s_3, \mathcal{P}), (v_4, s_6, \mathcal{K}), (v_4, s_7, \mathcal{K})\}$.

Unabhängigkeit der M_s $(s = 1, \ldots, n)$ faktorisiert die gesuchte bedingte Verteilung in das Produkt der bedingten Verteilungen der M_s, gegeben $\{\mathrm{supp}(G) \cap A(\chi) \cap (B_{\mathbf{T}} \times \{s\} \times \{\mathcal{P}, \mathcal{K}\}) = \chi \cap (B_{\mathbf{T}} \times \{s\} \times \{\mathcal{P}, \mathcal{K}\}\}$.

Für jedes $s \in \{1, \ldots, n\}$, welches in einem Element von χ als mittlere Koordinate vorkommt, wählen wir also $M'_s(.)$ so, daß gilt:

$$\mathcal{L}(M'_s(.)) = \mathcal{L}(M_s(.)|\forall_{v \in B_{\mathbf{T}}} : [M'_s(v) = \mathcal{P} \Longleftrightarrow (v, s, \mathcal{P}) \in \chi] \wedge [M'_s(v) = \mathcal{K} \Longleftrightarrow (v, s, \mathcal{K}) \in \chi])$$

Für jedes $s \in \{1, \ldots, n\}$, welches in keinem Element von χ als mittlere Koordinate vorkommt, für das aber ein mit einem Element von χ befreundetes Tripel $(v', s, \mathcal{M}) \notin \chi$ existiert, gelte:

$$\mathcal{L}(M'_s(.)) = \mathcal{L}(M_s(.)\,|\,\forall_{v \in B_{\mathbf{T}}} : (v, s, M_s(v)) \text{ ist mit keinem Element von } \chi \text{ befreundet})$$

Für alle sonstigen s sei $M'_s(.) := M_s(.)$. Wir konstruieren dann $(R^\chi + \delta^\chi)$ aus $M'.(.)$ analog zur Konstruktion von Q aus $M.(.)$. Nach Theorem 10.B aus Barbour *et al.* (1992) gilt:

$$d_{TV}(Q, K) \leq \sum_{\chi \in \Xi} \mathbb{E}\left(Q(\chi) \mathbb{E} \sum_{\varphi \in \Xi} |Q(\varphi) - R_\chi(\varphi)| \right)$$

Dabei kann $|Q(\varphi) - R_\chi(\varphi)|$ nur die Werte 0 und 1 annehmen, und damit es den Wert 1 annimmt, ist notwendig, daß eins der folgenden Ereignisse eintritt (vgl. Abb. 2.4):

(A) Es gilt $Q(\varphi) = 1$ und es existiert ein Site s, der sowohl in φ als auch in χ vorkommt.

(B) Es gilt $Q(\varphi) = 1$ und ein Element von χ ist mit einem Element von φ befreundet.

(C) Es gilt $R_\chi(\varphi) = 1$ und es gibt ein Site s und zwei (nicht notwendigerweise verschiedene) Blätter v und v', so daß $(v, s, M_s(v))$ mit φ und $(v', s, M_s(v'))$ mit χ befreundet ist.

Es sei $\mathcal{N}$ die Menge aller $(s, r, s', r') \in \{1, \ldots, n\}^4$, für die die Ungleichungen $s \leq r \leq s + n_{\text{amp}}$, $s' \leq r' \leq s' + n_{\text{amp}}$ und $s - n_{\text{amp}} \leq s' \leq s + n_{\text{amp}}$ gelten, und es sei $\mathcal{B}$

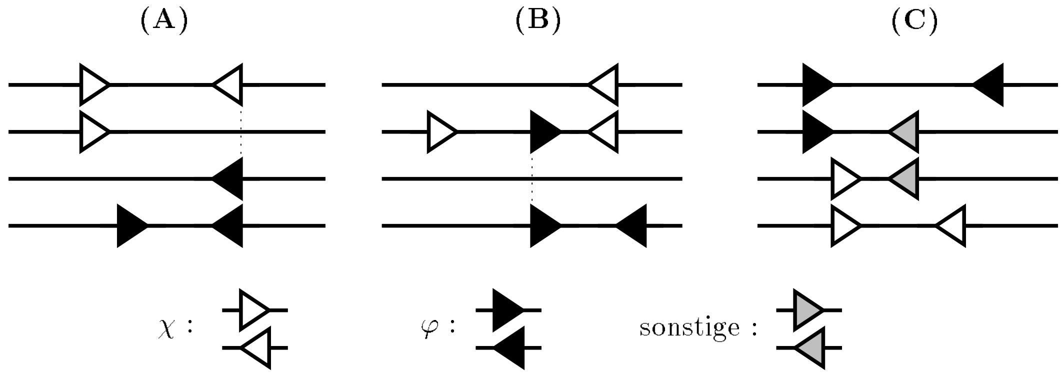


Abbildung 2.4: Beispiele für die Fälle **(A)**, **(B)** *und* **(C)**, *in denen* $|Q(\varphi) - R_\chi(\varphi)| = 1$ *gilt. Für die weiß eingezeichneten, zu* χ *gehörigen Tripel* $(v, s, \mathcal{M})$ *gelte* $M_s(v) \neq \mathcal{M}$ *und natürlich* $M'_s(v) = \mathcal{M}$, *für die schwarz eingezeichneten, zu* φ *gehörigen Tripel gelte* $M_s(v) = \mathcal{M}$ *und für die grau eingezeichneten* $M_s(v) = \mathcal{M}$ *und folgerichtig* $M'_s(v) \neq \mathcal{M}$. *Bei* **(A)** *und* **(B)** *gilt* $Q(\varphi) = 1$ *und* $R_\chi(\varphi) = Q(\chi) = 0$, *bei* **(C)** *gilt* $R_\chi(\varphi) = 1$ *und* $Q(\chi) = Q(\varphi) = 0$.

die Menge aller $(\varphi, \chi) \in \Xi^2$, für die ein $(s, r, s', r') \in \mathcal{N}$ und ein Paar $(v, v') \in B_{\mathbf{T}}$ mit $((v, s, \mathcal{P}), (v, r, \mathcal{K}), (v', s', \mathcal{P}), (v', r', \mathcal{K})) \in \varphi \times \varphi \times \chi \times \chi$ existieren.

Für die folgende Überlegung sei $(Q', M''_.(.))$ genauso verteilt wie $(Q, M_.(.))$, aber unabhängig von Q und $(R(\chi))_{\chi \in \Xi}$. In jedem der drei Fälle (A), (B) und (C) gilt $(\varphi, \chi) \in \mathcal{B}$. Also folgt:

$$\begin{aligned}\sum_{\chi,\varphi} \mathbb{E}\left(Q(\chi)\mathbb{E}\left|Q(\varphi) - R_\chi(\varphi)\right|\right) &= \mathbb{E}\sum_{\chi,\varphi} Q'(\chi)\left|Q(\varphi) - R_\chi(\varphi)\right| \\ &\leq \mathbb{E}\sum_{(\chi,\varphi)\in\mathcal{B}} Q'(\chi) \cdot Q(\varphi) + \mathbb{E}\sum_{(\chi,\varphi)\in\mathcal{B}} Q'(\chi) \cdot R_\chi(\varphi)\end{aligned}$$

Für jedes Paar von Sites (s, r) kann es nur ein $\varphi \in \Xi$ mit $Q(\varphi) = 1$ geben, für das ein Blatt v existiert, so daß $\{(v, s, \mathcal{P}), (v, r, \mathcal{K})\} \subseteq \varphi$ gilt. Dasselbe gilt für $Q'(\chi)$. Also folgt:

$$\begin{aligned}\mathbb{E}\sum_{(\varphi,\chi)\in\mathcal{B}} Q'(\chi)Q(\varphi) &\leq \sum_{(s,r,s',r')\in\mathcal{N}} \Pr(\exists v \in B_{\mathbf{T}} : \{(v, s, \mathcal{P}), (v, r, \mathcal{K})\} \subseteq \varphi) \cdot \\ &\qquad\qquad \cdot \Pr(\exists L' \in B_{\mathbf{T}} : \{(v', s', \mathcal{P}), (v', r', \mathcal{K})\} \subseteq \chi) \\ &= |\mathcal{N}| \cdot g^2 \ \leq\ 2 \cdot (n_{\text{amp}} + 1)^3 \cdot n \cdot g^2\end{aligned}$$

Mit derselben Argumentation für $\mathbb{E}\sum_{(\varphi,\chi)\in\mathcal{B}} Q'_\chi R^\chi_\varphi$ erhalten wir:

Satz 2.1

$$d_{TV}(Q, K) \leq \mathbb{E}\sum_{\chi,\varphi} Q'(\chi)\left|Q(\varphi) - R_\chi(\varphi)\right| \leq 4 \cdot n \cdot (n_{\text{amp}} + 1)^3 \cdot g^2$$

Bemerkung Die linke Seite der Ungleichung in Satz 2.1 liegt in unserem Referenzbeispiel aus Abschnitt 2.2.2.1 etwa bei 0,0013.

2.3.2 Wieviele Klumpen gibt es von welchem Typ?

Analog zu den Überlegungen am Ende des vorangehenden Abschnitts ergibt sich auch, die approximative obere Abschätzung $nn_{\text{amp}}^3(2g^2 + e_3)$ für die erwartete Anzahl der Klumpen, die Muster an mehr als drei verschiedenen Sites enthalten, wobei e_3 die Wahrscheinlichkeit ist, daß für ein beliebiges aber fest gewähltes Tupel $(s, r_1, r_2, r_3) \in 1, \ldots, n - n_{\text{amp}}$ drei Blätter v_1, v_2 und v_3 existieren, so daß für $i \in \{1, 2, 3\}$ jeweils $M_s(v_i) = \mathcal{P}$ und $M_{r_i}(v_i) = \mathcal{K}$ gilt. Mit ähnlichen Überlegungen wie in Abschnitt 2.2.2 kann man für die von uns ins Auge gefaßten Szenarien zeigen, daß nicht nur $2nn_{\text{amp}}^3 g^2$ sondern auch $nn_{\text{amp}}^3 e_3$ sehr klein wird. Daher gehen wir in den folgenden, teilweise etwas heuristischen Überlegungen dieses Unterabschnitts davon aus, daß wir es nur mit Klumpen mit zwei bis drei verschiedenen Sites zu tun haben. Wir werden außerdem wieder davon ausgehen, daß n so groß ist, daß wir Randeffekte vernachlässigen können. Wir nehmen außerdem an, daß die Muster $\mathcal{P}$ und $\mathcal{K}$ nicht zu ähnlich sind, so daß die Möglichkeit, daß an einem Site (bei verschiedenen Blättern) ein $\mathcal{P}$ und ein $\mathcal{K}$ vorkommen, vernachlässigbar ist (d. h. e_2 ist sehr klein).

Die erwartete Anzahl an Paaren $(s, k) \in \{1, \ldots, n\}^2$, die als Banden vorkommen, ist $\approx n \cdot n_{\text{amp}} \cdot g$. Die erwartete Anzahl an Tripeln (s, r, k) mit $s < r$, für die sowohl eine Bande (s, k) als auch eine Bande (r, k) vorkommt, ist ebenso wie die erwartete Anzahl an Tripeln (s, k, h) mit $k < h$, für die eine Bande (s, k) und eine Bande (s, h) vorkommen, ungefähr $\frac{1}{2} n \cdot n_{\text{amp}}^2 \cdot e_1$. Wir erwarten also $n \cdot n_{\text{amp}} \cdot (g - 2 \cdot n_{\text{amp}} \cdot e_1)$ verschiedene Banden (s, r), die in Klumpen ohne weitere Banden liegen, und wir erwarten $2n \cdot n_{\text{amp}}^2 \cdot e_1$ Banden, die sich paarweise je einen Klumpen teilen.

Das bedeutet, daß in dem durch K beschriebenen Modell die Anzahl der 2-Sites-Klumpen zu einem Parameter von ungefähr $n \cdot n_{\text{amp}} \cdot (g - 2 \cdot n_{\text{amp}} \cdot e_1)$ Poisson-verteilt ist, und die Anzahl der 3-Sites-Klumpen ungefähr zum Parameter $n \cdot n_{\text{amp}}^2 \cdot e_1$. In unserem Beispiel aus Abschnitt 2.2.2.1 liegen diese beiden Werte bei 10,11 und 0,045. Es sei daran erinnert, daß sowohl die Anzahlen an Klumpen der beiden Typen als auch die Werte (Cliquen), die diese Klumpen dann annehmen, im K-Modell untereinander unabhängig sind.

2.3.3 Schranken für den erwarteten Anteil an Klumpen mit mehr als einer Bande

Nach den Überlegungen des vorangegangenen Abschnitts enthält von allen Klumpen ein Anteil von ungefähr $\frac{n_{\text{amp}} \cdot e_1}{g - n_{\text{amp}} \cdot e_1}$ mehr als eine Bande. (Es sei daran erinnert, daß gemäß der hier verwendeten Sprechweise eine Bande auch dann **eine** Bande bleibt, wenn sie bei mehreren Taxa auftritt.) Wir gehen im Folgenden weiterhin davon aus, daß der Anteil an Klumpen mit mehr als drei Banden vernachlässigbar ist. Nach Abschnitt 2.2.2 können wir e_1 und g heuristisch schätzen oder für einen gegebenen Baum mit Hilfe eines Computers berechnen. Nun soll es darum gehen, für die Wahrscheinlichkeit, daß ein Klumpen mehr als eine Bande enthält, eine brauchbare obere Schranke zu finden, die man auch ohne ein spezielles Computerprogramm berechnen kann, und die nur überschaubare Eigenschaften des zugrundeliegenden Baumes verwendet oder gar für alle Bäume mit einer gegebenen Anzahl an Blättern konstant ist.

Als Beispiel betrachten wir zunächst die in Abbildung 2.5 dargestellte Klasse von Viererbäumen. Alle vier Blattkanten haben die Länge x und die beiden Wurzelkanten haben die

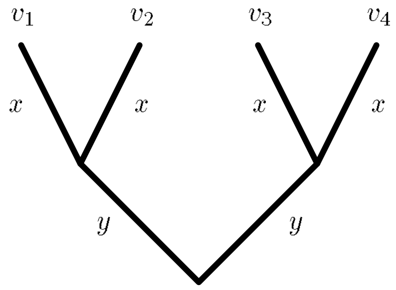


Abbildung 2.5: Die Klasse aller symmetrischen Viererbäume mit Blattkanten der Länge x und Wurzelkanten der Länge y.

Länge y. Der Graph in Abbildung 2.6 zeigt den Wert von $\frac{n_{\mathrm{amp}} \cdot e_1}{g - n_{\mathrm{amp}} \cdot e_1}$ für einen solchen Baum in Abhängigkeit von x und y, wobei von $n_{\mathrm{amp}} = 3000$ und $l = 10$ ausgegangen wird. Es fällt auf, daß das Maximum bei einem sternförmigen Baum angenommen wird. Dies kann man folgendermaßen interpretieren: Wegen der bei sternförmigen Bäumen besonders geringen Abhängigkeiten zwischen den in den Blättern angenommenen zufälligen Mustern steigt die erwartete Anzahl an verschiedenen Mustern pro Site und damit die Wahrscheinlichkeit, daß es zu einer gegebenen Bande, die auf einem der vier Blätter existiert, ein Blatt mit einer befreundeten Bande gibt.

Für $x = y = 0$, also den Fall ohne Mutationen, erreicht die in Abbildung 2.6 dargestellte Funktion ihr Minimum. Da in diesem Fall alle Taxa dieselbe DNA-Sequenz haben, ist dieses Minimum gerade die Wahrscheinlichkeit, daß ein Klumpen, der auf einem einzelnen DNA-Strang gegeben sei, mehr als eine Bande umfaßt. Da wir die Möglichkeit, daß ein Klumpen mehr als zwei Banden enthält, vernachlässigen können, entspricht diese Wahrscheinlichkeit in etwa der Hälfte der Wahrscheinlichkeit, daß es auf dem DNA-Strang zu einer existierenden, fest gewählten Bande (s, k) eine Bande (s, h) oder eine Bande (r, k) mit $r \neq s$ bzw. $h \neq k$ gibt. Dies ist wiederum ungefähr die bedingte Wahrscheinlichkeit, daß eine Bande (s, h) mit $h \neq k$ auf dem DNA-Strang existiert, gegeben eine Bande (s, k) liegt vor. Bei $n_{\mathrm{amp}} = 3000$ und $l = 10$ ist dies $2990 \cdot \left(\frac{1}{4}\right)^{10} \approx 0{,}00285$, da die Bedingung nichts anderes bedeutet, als daß der Primer für die Bande (s, h) schon mal existiert. Ganz offensichtlich ist $(n_{\mathrm{amp}} - 10) \cdot \left(\frac{1}{4}\right)^{l}$ für jeden beliebigen Baum eine untere Schranke für den Anteil der Klumpen mit mehr als einer Bande, denn für jeden Klumpen kann man unter allen Taxa, die mindestens eine zu dem Klumpen gehörige Bande besitzen, eines rein zufällig auswählen, und die Wahrscheinlichkeit, daß dieses Taxon dann mehr als eine zu dem Klumpen gehörige Bande besitzt, ist bereits größer gleich $(n_{\mathrm{amp}} - l) \cdot \left(\frac{1}{4}\right)^{l}$.

Nun suchen wir nach einer oberen Schranke. Dazu stellen wir zunächst eine Vorüberlegung

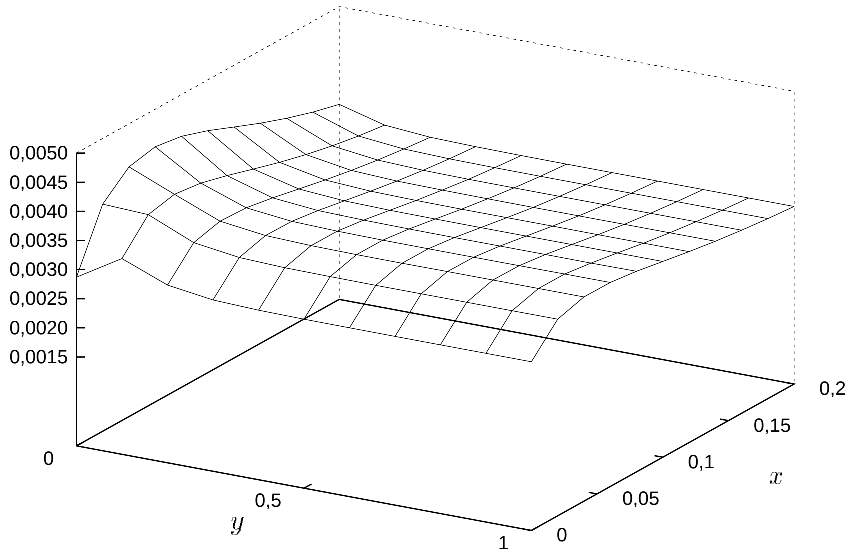


Abbildung 2.6: Der erwartete Anteil an Klumpen mit mehr als einer Bande in dem in Abbildung 2.5 dargestellten Baum in Abhängigkeit von x und y. Wir gehen dabei von $n_{\mathrm{amp}} = 3000$ und einer Primerlänge von zehn Basen aus.

an: Ein Baum habe b Blätter $v_0, v_1, \ldots, v_{b-1}$, und an einem ausgewählten Blatt v_0 gebe es am Site s einen Primer. Es sei eine Zahl $k \in \{s+l-1, \ldots, s+n_{\mathrm{amp}}\}$ gegeben. Die Wahrscheinlichkeit, daß im Baum ein Blatt mit der Bande (s,k) existiert, ist $\left(\frac{1}{4}\right)^l + \left(1 - \left(\frac{1}{4}\right)^l\right) \cdot z$, wobei z die bedingte Wahrscheinlichkeit bezeichnet, daß ein Blatt mit der Bande (s,k) existiert, gegeben daß v_0 zwar die Primerkopie $\mathcal{P}$ in s, nicht aber das Komplement $\mathcal{K}$ in k hat. Wir suchen jetzt zunächst eine obere Abschätzung für z. Sind $\lambda_1, \ldots, \lambda_{b-1}$ die Abstände der Blätter $v_1, \ldots, v_{b-1}$ zu v_0 (gemessen in erwartete Anzahl an Mutationen pro Site), so ist die erwartete Anzahl an Blättern, die in s, aber nicht in k mit v_0 übereinstimmen:

$$\sum_{i=1}^{b-1} f(\lambda_i)\ , \qquad \text{mit} \qquad f(\lambda) := \left(\frac{1+3\cdot e^{-\lambda}}{4}\right)^l \cdot \left(1 - \left(\frac{1+3\cdot e^{-\lambda}}{4}\right)^l\right)$$

Die erwartete Anzahl an Mustern, die im Site k bei Blättern auftreten, die sich im Site k, nicht aber in Site s von v_0 unterscheiden, ist also kleiner gleich $\sum_{i=1}^{b-1} f(\lambda_i)$. Jedes dieser Muster ist mit Wahrscheinlichkeit $\frac{1}{4^l-1}$ das Primerkomplement (man denke hier doppeltstochastisch). Die bedingte Wahrscheinlichkeit, daß ein Blatt mit einer Bande (s,k) existiert, gegeben daß bei b_0 in s der Primer und in k nicht dessen Komplement vorliegt, ist also kleiner gleich

$$\frac{1}{4^l-1} \cdot \sum_{i=1}^{b-1} f(\lambda_i).$$

Da die Funktion $x \mapsto x(1-x)$ ihr Maximum $\frac{1}{4}$ für $x = \frac{1}{2}$ annimmt, gilt:

$$f \leq f\left(-\ln \frac{4 \cdot \sqrt[l]{1/2} - 1}{3}\right) = \frac{1}{4}$$

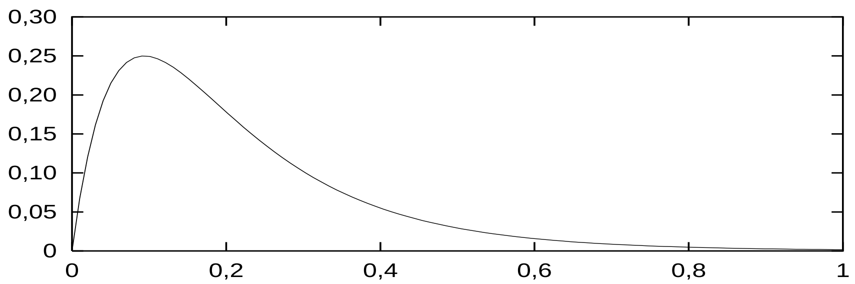


Abbildung 2.7: Die Funktion $f : \lambda \mapsto \left(\frac{1+3 \cdot e^{-\lambda}}{4}\right)^l \cdot \left(1 - \left(\frac{1+3 \cdot e^{-\lambda}}{4}\right)^l\right)$ *auf dem Intervall zwischen 0 und 1, für* $l = 10$. *Als maximalen Wert erreicht sie 0,25. Dies gilt für beliebige* $l \in \mathbb{N}$.

Für jeden Baum mit b Blättern ist der erwartete Anteil an Klumpen mit mehr als einer Bande also kleiner gleich $\left(\left(\frac{1}{4}\right)^l + \left(1 - \left(\frac{1}{4}\right)^l\right) \cdot \frac{1}{4} \cdot \frac{b-1}{4^l-1}\right) \cdot n_{\text{amp}}$, und falls μ das Maximum von $f(\lambda)$ über alle Abstände λ zwischen zwei Blättern des Baumes ist, ist der erwartete Anteil an Klumpen mit mehr als einer Bande also kleiner gleich $\left(\left(\frac{1}{4}\right)^l + \left(1 - \left(\frac{1}{4}\right)^l\right) \cdot \mu \cdot \frac{b-1}{4^l-1}\right) \cdot n_{\text{amp}}$.

2.3.4 Simulation von Klumpen

Nach den Überlegungen der letzten Abschnitte erscheint für viele RAPD-PCR-Szenarien die Modellierung durch eine Poisson-verteilte Anzahl an Klumpen mit einer Bande und eine zu einem wesentlich kleineren Parameter Poisson-verteilte Anzahl an Klumpen mit zwei Banden angemessen. Die Sites, bei denen diese Klumpen landen, sind rein zufällig (d. h. uniform verteilt und von allem unabhängig), ebenso die Längen der Banden.

Die eigentlich interessierende Eigenschaft der Klumpen ist jedoch die Menge der Blätter, bei denen die jeweilige Bande vorhanden ist, denn dies ist die Information, aus der die Phylogenien rekonstruiert werden sollen. Für einen gegebenen Baum kann man die Wahrscheinlichkeit, daß eine Bande genau bei einer bestimmten Menge von Blättern vorliegt, leicht berechnen. Man kann dabei ähnlich wie in Abschnitt 2.2.2 bei der Berechnung von g und e_1 vorgehen, also mit einem Verfahren, dessen Laufzeit linear in der Anzahl der Blätter ist. Wenn man dies aber für alle möglichen Mengen von Blättern durchführen möchte, hat man natürlich dennoch exponentiell viel zu tun. Es empfiehlt sich daher, die Verteilung der Blättermenge durch Simulationen zu schätzen. Aus diesem Grund, und auch weil man dadurch die Verteilung besser versteht, machen wir uns an dieser Stelle klar, wie man die gesuchte Verteilung effizient simulieren kann.

Zunächst zur Simulation einbandiger Klumpen: Gegeben sei also ein fester Baum und eine Primerlänge l. Wir wollen darauf bedingen, daß eine Bande (s,r) in einem einbandigen Klumpen existiert, und die zufällige Menge von Blättern simulieren, bei denen die Bande vorliegt. Es sei also $U^{(1)}_{(s,r)} := \{v \in B_{\mathbf{T}} : M_s(v) = \mathcal{P}\}$, $U^{(2)}_{(s,r)} := \{v \in B_{\mathbf{T}} : M_r(v) = \mathcal{K}\}$ und $U_{(s,r)} := U^{(1)}_{(s,r)} \cap U^{(2)}_{(s,r)}$. Wir erzeugen zunächst eine zufällige Teilmenge U von $B_{\mathbf{T}}$ mit Verteilung $\mathcal{L}(U_{(s,r)} | U_{(s,r)} \neq \emptyset)$. Als ersten Schritt simulieren wir die zufällige Partitionierung von $B_{\mathbf{T}}$ in Äquivalenzklassen bzgl. der Äquivalenzrelation $v \sim v' :\Longleftrightarrow (M_v(s), M_v(r)) = (M'_v(s), M'_v(r))$. Dazu wählen wir ein Blatt des Baumes aus und starten von dort aus eine mit $\mathbf{T}$ indizierte Markoff-Kette W auf $\{\mathtt{A},\mathtt{G},\mathtt{C},\mathtt{T}\}^{2l}$ mit der Jukes-Cantor-Übergangsdynamik (siehe Abschnitt 1.1.2), o. B. d. A. mit der Startsequenz $\mathtt{A},\mathtt{A},\ldots,\mathtt{A}$. Dadurch wird jedem Blatt ein Wort der Länge $2l$ zugeornet. Gegeben, daß an einem bestimmten $2l$-Tupel von Sites in den Blättern des Baumes w verschiedene Wörter stehen, nimmt die bedingte Wahrscheinlichkeit, daß eines davon den für eine Bande nötigen Zustand $(\mathcal{P},\mathcal{K})$ annimmt, den Wert $w \cdot \left(\frac{1}{4}\right)^{2l}$ an. Dieser Proportionalität in w werden wir durch *rejection sampling* gerecht: Ist b die Anzahl der Blätter des Baumes und w die Anzahl der Äquivalenzklassen, die wir durch die oben beschriebene Prozedur erhalten haben, so betätigen wir einen Zufallsmechanismus, der mit Wahrscheinlichkeit $1 - \frac{w}{b}$ dafür sorgt, daß die Partitionierung verworfen wird und das ganze Verfahren neu startet. Dies wiederholen wir solange, bis eine Partitionierung nicht verworfen wird. Aus der Menge der dann vorliegenden Äquivalenzklassen wählen wir eine uniform aus (also jede mit Wahrscheinlichkeit $\frac{1}{w}$, egal wieviele Blätter sie enthält). Diese nennen wir U. Seien $U^{(1)} \supseteq U$ und $U^{(2)} \supseteq U$ die Mengen der Blätter, in denen die ersten (bzw. letzten) l Komponten von W dieselben sind wie in den Blättern in U. Damit hat $(U, U^{(1)}, U^{(2)})$ die Verteilung $\mathcal{L}((U_{(s,r)}, U^{(1)}_{(s,r)}, U^{(2)}_{(s,r)}) | U_{(s,r)} \neq \emptyset)$. Wir vergessen also nun W und gehen davon aus, daß $U^{(1)}$ und $U^{(2)}$ die Mengen aller Blätter sind, bei denen eine Kopie von $\mathcal{P}$ bzw. $\mathcal{K}$ an der Stelle s bzw. r vorliegt, und daß U die Menge aller Blätter ist, bei denen die Bande (s,r) existiert.

Nun hängt es aber von $U^{(1)}_{(s,r)}$ und $U^{(2)}_{(s,r)}$ ab, wie groß die Wahrscheinlichkeit ist, daß die Bande (s,r) in ihrem Klumpen allein ist. Wir führen deshalb noch ein Zufallsexperiment durch, welches uns mit der Wahrscheinlichkeit, daß die Bande Teil eines zweibandigen Klumpens ist, dazu führt, daß $(U, U^{(1)}, U^{(2)})$ verworfen und die Simulation wiederholt wird. Dazu lassen wir ein weiteres l-Tupel von Basen längs des Baumes evolvieren. Wir zählen, wieviele l-Worte dabei auf $U^{(1)}$ und wieviele auf $U^{(2)}$ angenommen werden. Es bezeichnen S und R diese beiden Anzahlen. Offensichtlich ist $S \cdot \left(\frac{1}{4}\right)^{l}$ ein erwarungstreuer Schätzer für die Wahrscheinlichkeit, daß in $U^{(1)}$ ein Blatt existiert, bei dem an einem fest vorgegebenen Site ein Primerkomplement auftritt. Für $R \cdot \left(\frac{1}{4}\right)^{l}$ gilt die analoge Aussage in bezug auf $U^{(2)}$. Wenn wir dann also ein zum zufälligen Parameter $(n_{\mathrm{amp}} - 1) \cdot (S + R) \cdot \left(\frac{1}{4}\right)^{l}$ Bernoulli-verteiltes Experiment ausführen, so ist die Wahrscheinlichkeit, daß dieses letztlich eine 1 liefert, ungefähr die Wahrscheinlichkeit, daß die oben als Zwischenergebnis erhaltene Bande Teil eines zweibandigen Klumpens ist. Der dabei auftretende Fehler entspricht der Wahrscheinlichkeit, daß der Klumpen mehr als zwei Banden enthält, und ist daher in unserem Szenario vernachlässigbar.

Wie man zweibandige Klumpen simuliert, sollte nun klar sein: Man verwendet eine Markoff-Kette auf $\{\mathtt{A},\mathtt{G},\mathtt{C},\mathtt{T}\}^{3l}$. Jedem Blatt v ist dann ein Wort $W_1(v)$ und ein Wort $W_2(v)$ zugeordnet, und zwar besteht $W_1(v)$ aus den ersten $2l$ Basen der in v vorliegenden Sequenz und $W_2(v)$ ist aus den ersten l und den letzten l Basen zusammengesetzt. Sei Ω die Menge der Paare $(W, W') \in (\{\mathtt{A},\mathtt{G},\mathtt{C},\mathtt{T}\}^{2l})^2$, für die Blätter v und v' mit $W = W_1(v)$ und $W' = W_2(v')$ existieren, und bei denen die ersten l Basen in W und W' übereinstimmen. Wir akzeptieren das Zwischenergebnis mit Wahrscheinlichkeit $\frac{|\Omega|}{b}$, wobei b wieder die Anzahl der Blätter bezeichnet. Wir wiederholen die Prozedur bis wir ein Zwischenergebnis akzeptiert haben. Dann wählen wir aus Ω uniform ein Paar $(\mathcal{W}, \mathcal{W}')$ aus. Als Simulationsergebnis erhalten wir damit einen Klumpen mit zwei Banden, die gerade bei den Blättern vorliegen, bei denen W_1 das Wort $\mathcal{W}$ annimmt, bzw. W_2 das Wort $\mathcal{W}'$. Wir gehen o. B. d. A. davon aus, daß es sich um einen Klumpen mit zwei Banden handelt, deren Primer am selben Site sitzt. Das Site, an dem der Primer sitzt, und die Längen der Banden sind uniform aus $\{1, \ldots, n\}$ bzw. $\{l, \ldots, n_{\mathrm{amp}}\}$ zu wählen. (Um genau zu sein, muß man darauf bedingen, daß die beiden Banden nicht gleich lang sind.)

2.4 Fazit

Als adäquates Modell für die Verteilung der RAPD-Bandenkonfigurationen auf den DNA-Sequenzen der Blätter eines Stammbaums erscheint das Modell der Poisson'sch eingestreuten ein- bis zweibandigen Klumpen. Mit den zweibandigen Klumpen enthält es Korrelationen zwischen den Banden, die sich aus der Jukes-Cantor-Evolutionsdynamik ergeben. Für einen vorgegebenen Baum läßt sich der Totalvariationsabstand zwischen den Verteilungen der Bandenkonfigurationen, die sich aus dem Poisson-Klumpen-Modell und dem Modell mit unabhängig evolvierenden Mustern (siehe Abschnitt 1.1.3) ergeben, nach oben abschätzen (Satz 1.2, Seite 16). Der Aufwand dafür ist linear in der Anzahl der Blätter. Beispiele zeigen, daß das Ergebnis für typische RAPD-PCR-Szenarien hinreichend klein wird. Zusammen mit den Ergebnissen aus Kapitel 1 und der Dreiecksungleichung ergibt sich dann für viele anwendungsrelevante Situationen, daß die Unterschiede zwischen dem Poisson-Klumpen-Modell und dem Jukes-Cantor-Modell in Hinblick auf die Bandenkonfigurationen vernachlässigbar sind.

Außerdem kann man Bandenkonfigurationen nach dem Modell der ein- bis zweibandigen Klumpen effizient simulieren.

Kapitel 3

Stammbaumrekonstruktion mit RAPD-Daten

> Of course, for a model to be good, you must show it leads somewhere: This may be done by mathematical 'experiment', i.e. by computations or by the first step in its analysis.
>
> *David Mumford, 1998*

Nach den Überlegungen aus Kapitel 2 entstehen in relevanten Szenarien keine wesentlichen Ungenauigkeiten, wenn man die Verteilung der Banden, die sich aus dem Jukes-Cantor-Modell ergibt, durch das Modell der unabhängig eingestreuten Bandenklumpen beschreibt und sich dabei auf Klumpen mit ein oder zwei Banden beschränkt.

Nun soll der Nutzen dieser Vereinfachung aufgezeigt werden. Dazu wird am Beispiel der Schätzung von Viererbaumtopologien demonstriert, wie sich aus dem Klumpen-Modell Auswertungsansätze ergeben und wie man diese dann mit Hilfe der in Abschnitt 2.3.4 dargestellten Simulationsmethoden in Verbindung mit explorativer Datenanalyse bewerten kann. Derartige Bewertungen sollen hier nicht in erschöpfender Weise durchgeführt werden. Es geht eher darum, Möglichkeiten aufzuzeigen und dabei einige Schlaglichter auf Sachverhalte zu werfen, die nicht zuletzt auch für Anwender der RAPD-PCR interessant sein dürften.

Wir beschränken uns in diesem Kapitel auf vierblättrige Bäume. Diese bilden eine einigermaßen überschaubare Beispielklasse, und außerdem gibt es sehr effiziente Baumrekonstruktionsverfahren, bei denen zunächst die vierblättrigen Teilbäume eines Stammbaums mit Maximum-Likelihood-Methoden geschätzt und die Ergebnisse dieser Schätzung zu einem Baum zusammengesetzt werden (vgl. Strimmer und von Haeseler (1996, 1997)).

3.1 Mögliche Probleme der Stammbaumrekonstruktion mit RAPD-Daten

Die Problematik der Stammbaumrekonstruktion mit RAPD-PCR-Daten wurde teilweise bereits in Abschnitt 2.1 dargestellt. Leider erhält man bei der Durchführung der RAPD-RCR nicht die volle Information über die Klumpen. Bei der Betrachtung der Signale auf dem Gel erhält man nur Informationen der Art „Banden der Länge x liegen bei diesen und jenen Blättern vor“. Banden, die im Rahmen der Meßgenauigkeit dieselbe Länge haben, sind auf dem Gel nicht unterscheidbar. Es ist außerdem nicht feststellbar, ob eine Bande zu einem einbandigen oder zu einem zweibandigen Klumpen gehört. Bei zweibandigen Klumpen tritt das zusätzliche Problem auf, daß in Blättern, in denen beide Banden vorliegen, in der Regel nur die kürzere sichtbar ist (vgl. Anhang B und Abschnitt 2.1). Wie wir in Kapitel 2 gesehen haben, wird der Anteil an zweibandigen Klumpen in vielen Fällen recht gering sein, aber es ist theoretisch durchaus denkbar, daß Likelihood-Quotienten stark verzerrt werden, wenn man auch nur einen zweibandigen Klumpen als zwei einbandige fehlinterpretiert.

3.2 Szenario und Art der simulierten Daten

Es gibt drei mögliche Topologien für den unverwurzelten Stammbaum von vier Individuen. Wir betrachten die Problemstellung, die Topologie des Stammbaumes aus den RAPD-Bandenmustern der vier Blätter zu schätzen.

Wie wahrscheinlich ein bestimmtes Bandenmuster ist, hängt nicht nur von der Topologie sondern auch von den Kantenlängen ab. Die Likelihood einer Topologie bei gegebenen Daten ist die maximale Likelihood aller Bäume mit Kantenlängen, die diese Topologie haben. Um die zur Diskussion stehenden Schätzverfahren in Hinblick auf den Rechenaufwand zu vereinfachen, lassen wir nur endlich viele Kantenlängen zu. Für das gesamte Kapitel wählen wir die sechs Werte 0,00625, 0,0125, 0,025, 0,05, 0,1 und 0,2. Wenn eine Kante die Länge λ hat, so ist $\left(\frac{1}{4}+\frac{3}{4}\cdot e^{-\lambda}\right)^{2l}$ die Wahrscheinlichkeit, daß eine in einem der beiden zugehörigen Knoten vorhandene Bande die Kante überlebt, wobei l die Länge des Primers bezeichnet. Für $l=10$ erhalten wir zu den sechs genannten Kantenlängen also die Überlebenswahrscheinlichkeiten 0,91, 0,83, 0,69, 0,47, 0,23 und 0,054. Bei 6 möglichen Kantenlängen gibt es für jede der drei Baumtopologien $6^5 = 7776$ verschiedene Bäume. Wir vereinfachen die Problemstellung insofern, als daß bekannt sei, daß sich unter diesen $3 \cdot 6^5 = 23328$ Bäumen der richtige befindet. Dessen Topologie soll anhand der RAPD-Daten geschätzt werden. (Natürlich liegt in der Auswahl der Testbäume eine gewisse Willkür und es ist durchaus denkbar, daß eine andere Population von Testbäumen das eine oder andere Verfahren in einem anderen Licht erscheinen läßt.)

Für den „wahren“ Baum simulieren wir zunächst zu den erwarteten Häufigkeiten Poisson-verteilt viele ein- und zweibandige Klumpen (siehe 2.3.2 und 2.3.4). Da wir untersuchen wollen, wie sich die im Klumpenmodell übriggebliebenen Abhängigkeiten zwischen den Banden und die

Möglichkeit, daß Banden ähnlicher Länge verwechselt werden können, auf die Schätzbarkeit der Baumtopologie auswirken, erzeugen wir aus der Ausgabe des in Abschnitt 2.3.4 beschriebenen Verfahrens drei verschiedene „Datensätze", bei denen es sich jeweils um eine zufällige Abbildung von der Familie der nichtleeren Teilmengen von $B_{\mathbf{T}}$ nach $\mathbb{N}_0$ handelt.

Datensatz D^I: Hier liegt das „Modell ohne Abhängigkeiten" aus Kapitel 2 zugrunde. Es handelt sich also um Daten, die in zweierlei Hinsicht idealisiert sind: Alle Banden evolvieren voneinander unabhängig und sind voneinander zu unterscheiden, sogar wenn sie exakt dieselbe Länge haben. Wir generieren D^I gekoppelt mit D^{II} und D^{III} (s.u.) mit Hilfe des in Abschnitt 2.3.4 beschriebenen Verfahrens. Dabei machen wir aus den zweibandigen Klumpen einbandige, indem wir von den beiden Banden eine rein zufällig auswählen und die andere ignorieren. Für eine nichtleere Menge U von Blättern sei dann D^I_U die Anzahl an Banden, die bei allen Blättern in U und bei sonst keinen vorkommen.

Datensatz D^{II}: Bei diesem Datensatz werden die Abhängigkeiten berücksichtigt, die dadurch zustandekommen, daß es Klumpen mit zwei Banden gibt. Verwechslungen von Banden finden ebenso wie bei D^I nicht statt. D^{II}_U ist die Anzahl an Banden, die bei allen Blättern in U und bei sonst keinen sichtbar sind. Dabei ist eine Bande bei einem Blatt genau dann sichtbar, wenn sie bei dem Blatt existiert und es im selben Klumpen keine kürzere Bande gibt, die ebenfalls bei dem Blatt existiert (vgl. Abschnitt 0.2).

Datensatz D^{III}: Hier berücksichtigen wir die Abhängigkeiten wie bei D^{II} und gehen zusätzlich davon aus, daß Banden ähnlicher Länge nicht unterschieden werden können bzw. daß nur eine bestimmte Anzahl n_{gp} an Gelpositionen zu unterscheiden sind. Für jede nichtleere Menge U von Blättern und $i \in \{l, \ldots, n_{gp}\}$ sei $S_U(i)$ die Indikatorfunktion dafür, daß U die Menge aller Blätter ist, bei denen eine Bande sichtbar ist, deren Länge im Intervall $G_i := [l + (i-1) \cdot \frac{n_{\mathrm{amp}}-l}{n_{gp}-1}\ ,\ l + i \cdot \frac{n_{\mathrm{amp}}-l}{n_{gp}-1}[$ liegt (und die damit auf dem Gel in der i-ten Position zu sehen ist). Es sei dann D^{III}_U die Anzahl an Gelpositionen i, für die $S_U(i) = 1$ gilt.

Mit dem Datensatz D^{III} simulieren wir die Bandensignale, die man bei der RAPD-PCR beobachten kann. Die idealisierten Datensätze D^I und D^{II} sollen lediglich Vergleiche ermöglichen, ob man beim Schätzen der Baumtopologie bedeutend weniger Fehler machen würde, wenn es keine Abhängigkeiten bzw. Verwechslungen zwischen den Banden gäbe. Diese Vergleiche sollen Aufschluß darüber geben, inwieweit die genannten Effekte im Kontext der Baumtopologieschätzung vernachlässigbar sind.

3.3 Methoden der Topologieschätzung

Die folgenden Methoden der Topologieschätzung sollen in Abschnitt 3.4 auf die Datensätze D^I, D^{II} und D^{III} angewandt werden, die aus simulierten Bandenklumpen für Viererbäume erzeugt wurden. Die Schätzmethoden sind prinzipiell auch in Szenarien mit einer größeren Anzahl an Individuen anwendbar. Das wesentliche Kriterium für die Beurteilung der Verfahren ist die Fehleranfälligkeit beim Schätzen der Baumtopologie auf der Basis des Datensatzes D^{III}. Verfahren, die Abhängigkeiten und Verwechslungen zwischen den Banden vernachlässigen, werden wir zum Vergleich auf die Datensätze D^I und D^{II} anwenden, um uns einen Eindruck zu verschaffen, wie stark sich die Vernachlässigung bei dem jeweiligen Verfahren auswirkt.

Der Schätzer MP: Bei diesem Schätzer lassen wir uns vom Prinzip der maximalen Parsimonie leiten, d. h. wir „glauben" an die Baumtopologie, bei der so wenige Rückmutationen wie möglich erfolgen. Da wir es mit vierblättrigen Bäumen zu tun haben, wird demnach eine Bande, die bei genau zwei Blättern auftritt, als Hinweis für die Baumtopologie gewertet, bei der diese beiden Blätter von den beiden anderen durch den Zentralast getrennt werden. Banden, die bei nur einem, bei drei oder bei allen vier Taxa vorkommen, werden als nicht informativ betrachtet. Der Schätzer MP liefert die Baumtopologie, auf die die meisten Banden hinweisen. (Algorithmen für parsimonische Analysen von Sequenzdaten findet man bei Swofford und Olsen (1990).)

Der Schätzer ML$_1$: Hier gehen wir davon aus, daß uns ein Datensatz vorliegt, der wie D^I ohne Abhängigkeiten zwischen den Banden und ohne Verwechslungen zustande gekommen ist, und wir schätzen die Topologie des Stammbaumes nach dem Maximum-Likelihood-Prinzip. Wir orientieren uns also am „Modell ohne Abhängigkeiten" aus Abschnitt 2.2.

Nach demselben Prinzip wie bei der Berechnung von g in Abschnitt 2.2.2 können wir für jede Bande die Wahrscheinlichkeit ausrechnen, daß sie in einer gegebenen Menge von Blättern vorliegt, sofern der Baum einschließlich seiner Kantenlängen bekannt ist. Der dazu benötigte Zeitaufwand ist linear in der Anzahl der Blätter. Für vierblättrige Bäume ist allerdings folgender Rechenweg günstiger: Wir betrachten zunächst eine der $n \cdot n_{\text{amp}}$ möglichen Banden und berechnen für jede Menge von Blättern $U \subseteq B_{\mathbf{T}}$ die Wahrscheinlichkeit p_U, daß bei allen $v \in U$ die Bande vorkommt. Dazu betrachten wir eine mit dem Baum indizierte Markoff-Kette W, die der Jukes-Cantor-Dynamik auf $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^{2l}$ folgt (vgl. Abschnitt 1.1.2). Die zufällige Menge aller Blätter v, in denen $W(v)$ den Zustand $(\mathtt{A}, \mathtt{A}, \ldots, \mathtt{A})$ annimmt, ist offensichtlich genauso verteilt wie die Menge aller Blätter, in denen die Bande vorliegt. p_U ist also die $2l$-te Potenz der Wahrscheinlichkeit, daß an einem vorgegebenen Site bei allen $v \in U$ ein $\mathtt{A}$ steht, und läßt sich daher leicht berechnen. Wir berechnen dann für jedes $U \subset B_{\mathbf{T}}$ die Wahrscheinlichkeit q_U, daß die Bande *genau* bei den Blättern in U vorkommt, mit der Rekursion $q_U = p_U - \sum_{V \supsetneq U} q_V$. Da $\{D^I_U : U \text{ ist nichtleere Menge von Taxa}\}$ eine Familie unabhängiger, Poisson-verteilter Zufallsvariabler mit $\mathbb{E}D^I_U = n \cdot n_{\text{amp}} \cdot q_U$ ist, erhalten wir damit die Verteilung von D^I für jeden gegebenen Baum – oder anders gesagt: Wir können für eine gegebene Realisierung von D^I die Likelihood jedes Baumes berechnen. Der Schätzer ML$_1$ liefert die Topologie des Baumes mit der maximalen Likelihood.

Der Schätzer ML$_2$: Wir orientieren uns auch hier am Maximum-Likelihood-Prinzip und berücksichtigen dabei, daß bei D^{III} Banden allzu ähnlicher Länge nicht zu unterscheiden sind. (Abhängigkeiten zwischen den Banden werden hingegen vernachlässigt.) Wir nähern den Evolutionsprozeß der Bandensignale durch einen Markoffprozeß an, bei dem es für jede Gelposition in jedem Knoten zwei Zustände gibt.

Wir bezeichnen wieder mit n_{dna} die Länge der DNA, mit n_{amp} die Länge des Amplifikationsbereichs, mit n_{gp} die Anzahl der unterscheidbaren Gelpositionen und mit l die Länge des Primers. Für $i \in \{1, \ldots, n_{gp}\}$ und einen Knoten v sei $I_i(v)$ die Indikatorfunktion dafür, daß in v mindestens eine Bande existiert, deren Länge im Intervall G_i liegt, μ sei die erwartete Anzahl solcher Banden in einem beliebigen, aber fest gewählten Knoten. Für zwei Knoten v_1 und v_2, die durch eine Kante der Länge λ verbunden sind, sei ν_λ die erwartete Anzahl an Banden, die sowohl bei v_1 als auch bei v_2 vorkommen und deren Länge in G_i liegt. Die Anzahl solcher Banden ist approximativ Poisson-verteilt zum Parameter ν_λ. Die Anzahl an Banden, deren Länge in dem genannten Intervall liegt und die nur in v_1, nicht aber in v_2 vorkommen, ist ungefähr Poisson-verteilt zum Parameter $\mu - \nu_\lambda$. Dasselbe gilt, wenn man v_1 und v_2 vertauscht. Wir erhalten also:

$$\begin{aligned}
\mu &\approx \frac{n_{\text{dna}} \cdot n_{\text{amp}}}{n_{gp}} \cdot \left(\frac{1}{4}\right)^{2l} \\
\mathbb{E} I_i(v_1) &\approx 1 - e^{-\mu} \\
\nu_\lambda &\approx \mu \cdot \left(\frac{1}{4} + \frac{3}{4} \cdot e^{-\lambda}\right)^{2l} \\
\mathbb{E}\left(I_i(v_1) \cdot I_i(v_2)\right) &\approx \left(1 - e^{-\nu_\lambda}\right) + e^{-\nu_\lambda} \cdot \left(1 - e^{-\mu+\nu_\lambda}\right)^2
\end{aligned}$$

Wir haben also die gemeinsame Verteilung von $I_i(v_1)$ und $I_i(v_2)$ zumindest approximativ im Griff. Es ist aber sehr aufwendig, die gemeiname Verteilung von $\{I_i(v)$: U ist Knoten im Baum$\}$ zu berechnen. Wir vernachlässigen daher gewisse Abhängigkeiten und nehmen für die Konstruktion des Schätzers ML$_2$ an, daß die Gelpositionen in folgendem Sinne von Knoten zu Knoten gedächtnislos evolvieren: Wenn V eine Menge von Knoten ist, so daß v_1 topologisch zwischen V und v_2 liegt, so gilt:

$$\begin{aligned}
&\mathbb{E}(I_i(v_2) \mid \{I_i(v) : v \in V \cup \{v_1\}\}) \approx \mathbb{E}(I_i(v_2) \mid I_i(v_1)) \\
&\quad = \frac{\mathbb{E}(I_i(v_1) \cdot I_i(v_2))}{\mathbb{E}(I_i(v_1))} \cdot I_i(v_1) + (1 - I_i(v_1)) \cdot \frac{\mathbb{E}(I_i(v_2)) - \mathbb{E}(I_i(v_1) \cdot I_i(v_2))}{1 - \mathbb{E}(I_i(v_1))}
\end{aligned}$$

Wir erhalten damit bis auf eine Umskalierung der Kantenlängen eine Modellierung des Entstehens und Vergehens der Bandensignale durch eine Markoffkette längs der Kanten des Baumes mit den Zuständen {sichtbar, nicht sichtbar}. Mit dem Schätzer ML$_2$ ahmen wir also nach, was passiert, wenn man Software, die wie etwa PHYLIP DNAML (siehe Felsenstein, 1993) zur Maximum-Likelihood-Phylogenieschätzung auf der Basis von DNA-Sequenzdaten entwickelt wurde, in naiver Weise auf RAPD-Daten anwendet; mehr dazu in Abschnitt 3.4.2.5.

Der Schätzer ML_3: Ähnlich wie ML_2 ist auch ML_3 ein Maximum-Likelihood-Schätzer, der auf einer Modellierung der Bandensignaldynamik durch eine baumindizierte Markoffkette beruht. Nun soll aber berücksichtigt werden, daß eine Bande, die von einem zum nächsten Knoten verschwunden ist, beim übernächsten mit erhöhter Wahrscheinlichkeit wieder auftreten kann, da unter Umständen nur ein einziges Site rückmutieren muß. Wir unterscheiden daher für jede Gelposition die folgenden drei Zustände:

1. Es gibt mindestens eine Bande, die an der betreffenden Gelposition ein Signal hervorruft.

2. Es gibt kein Signal an der betreffenden Gelposition, aber man könnte eines erhalten, indem man eine einzige Base ändern würde.

3. Man müßte mindestens zwei Basen ändern, um an der Gelposition ein Signal zu erhalten.

Die Übergangswahrscheinlichkeiten zwischen diesen Zuständen lassen sich für zwei benachbarte Knoten in Abhängigkeit von der Länge ihrer Verbindungskante ähnlich wie bei ML_2 leicht berechnen. Ob sich eine Gelposition bei einem Taxon im zweiten oder im dritten Zustand befindet, ist vom Beobachter nicht zu unterscheiden. Analog zu ML_2 wird auch bei ML_3 eine Gedächtnislosigkeit von Knoten zu Knoten angenommen.

3.4 Simulationsergebnisse und erste Interpretationen

Wir rechnen in diesem Abschnitt mit einer DNA-Länge von 3 Milliarden Basenpaaren. (Dies entspricht etwa der Länge eines haploiden menschlichen Chromosomensatzes.) Für den Amplifikationsbereich nehmen wir $n_{\mathrm{amp}} = 3000$ an und für die Länge der Primer meistens $l = 10$. Für die Anzahl der unterscheidbaren Gelpositionen werden wir verschiedene Werte einsetzen. Den C++-Quellcode des Simulationsprogramms habe ich im Internet unter `http://stoch.math.uni-frankfurt.de/Leute/metzler/dissprgs.html` abgelegt.

Wir gehen von einer festen Baumtopologie aus. Da für jede der fünf Kanten sechs mögliche Längen angenommen werden, ergeben sich 6^5 verschiedene Bäume. Bei einem Simulationslauf werden für jeden dieser Bäume die Datensätze D^I, D^{II} und D^{III} simuliert und mit den verschiedenen Verfahren ausgewertet. Anschließend wird verglichen, welches Verfahren wie häufig ein falsches Ergebnis lieferte. Es ist zu erwarten, daß es umso leichter ist, die Baumtopologie zu schätzen, je länger der Zentralast des Baumes ist. Daher betrachten wir die Fehlerhäufigkeiten in Abhängigkeit von der Zentralastlänge. Bei der graphischen Darstellung der Simulationsergebnisse wählen wir als Maßeinheit für die Astlänge die in Prozent angegebene Wahrscheinlichkeit, daß eine im einen Endknoten der Kante existierende Bande (die sich gemäß Jukes-Cantor entwickelt) im anderen Endknoten verschwunden ist. Da für jede Zentralastlänge die übrigen 4 Kantenlängen jeweils die 6 Werte 8,94%, 17,1%, 31,2%, 52,5%, 77,3% und 94,6% annehmen konnten, wurden für jede Zentralastlänge Datensätze zu $6^4 = 1296$ verschiedenen Bäumen simuliert und ausgewertet. Von diesen Bäumen sollte jedes halbwegs sinnvolle Verfahren zumindest $6^4/3 = 432$ Bäumen die richtige Topologie zuordnen, denn diesen Wert könnte man in Erwartung erreichen, indem man jedem Baum eine rein zufällige Topologie zuordnet.

In einigen Fällen ordnen die Verfahren zwei oder gar allen drei Topologien dieselbe Likelihood bzw. Parsimonie zu. Wir werden diesen Fällen durch die Verwendung des folgenden Bewertungsschemas Rechnung tragen: Für jede richtig geschätzte Topologie erhält ein Verfahren einen Punkt. Wenn es hingegen zwei Topologien favorisiert und die richtige dabei ist, erhält es dafür einen halben Punkt, und wenn es allen drei Topologien dieselbe Likelihood bzw. Parsimonie zuordnet, dann erhält es einen drittel Punkt. Die Vorstellung dabei ist, daß ein Verfahren bei mehreren Favoriten rein zufällig einen davon auswählen müßte, und daß es die Wahrscheinlichkeit, dann den richtigen zu treffen, als Punktanteil erhält. In den nachfolgenden Graphiken ist die Anzahl an Punkten zu sehen.

3.4.1 Zur Robustheit von ML_1

Zunächst betrachten wir Simulationsergebnisse in Hinblick auf die Frage, ob Abhängigkeiten zwischen den Banden und die Möglichkeit der Verwechslungen von Banden vernachlässigbar sind. Wir untersuchen also die Robustheit von ML_1 gegenüber diesen Effekten, indem wir erkunden, ob ML_1 auf der Basis von D^I wesentlich zuverlässiger die Topologie schätzen kann als auf der Basis von D^{II} oder D^{III}.

3.4.1.1 Nur ein Primer

Wir betrachten zunächst ein Szenario, in dem die Daten mit einer einzigen RAPD-PCR-Primersequenz erzeugt wurden und 100 Gelpositionen unterscheidbar sind. In Abbildung 3.1 ist zu sehen, wieviele Bewertungspunkte ML_1 mit D^I, D^{II} und D^{III} bei verschiedenen Zentralastlängen erhielt. (Die Kurven für ML_2 und MP werden wir in Abschnitt 3.4.2 diskutieren.)

Offensichtlich spielen Abhängigkeiten zwischen den Banden hier keine bedeutende Rolle, denn das Verfahren ML_1 konnte mit dem Datensatz D^{II}, der diese Abhängigkeiten enthält, genauso viele Bäume richtig schätzen wie mit dem Datensatz D^I, bei dem solche Abhängigkeiten ausgeschlossen sind. Es macht allerdings schon einen Unterschied, wenn man die Möglichkeit von Bandenverwechslungen einbezieht: Mit zunehmender Zentralastlänge schneidet der Datensatz D^{III} schlechter ab als die anderen beiden.

3.4.1.2 Mehrere Primer

Oft werden mehrere RAPD-PCR-Versuche, jeder mit einem anderen Primer, durchgeführt und die Daten werden dann gepoolt. Da klar ist, welches Bandensignal aus welchem Versuch stammt, wächst die Anzahl der beobachteten Bandensignale linear mit der Anzahl der Primer (bis auf Randeffekte, die wir vernachlässigen). Wir können für diesen Fall dasselbe Simulati-

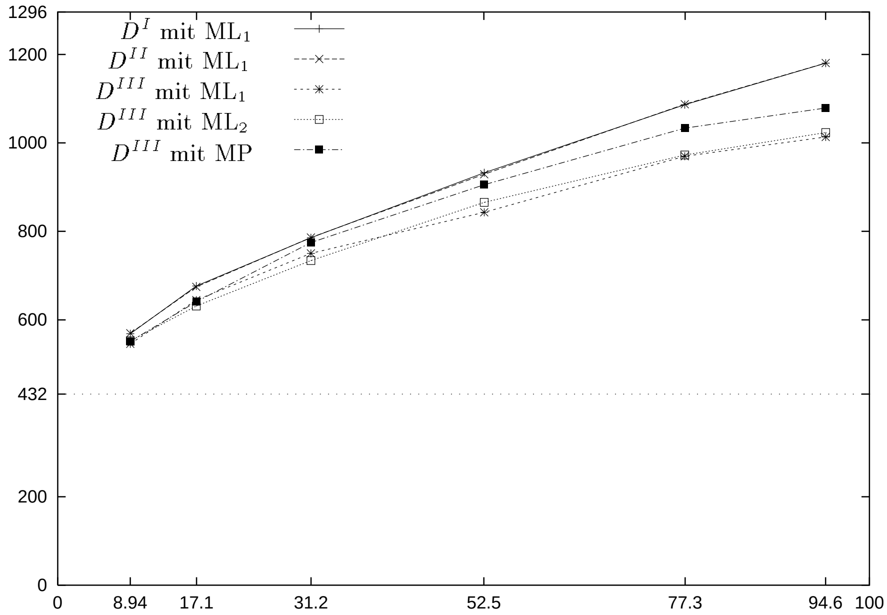


Abbildung 3.1: Bei $n_{\mathrm{dna}} = 3 \cdot 10^9$, $n_{\mathrm{amp}} = 3000$, $n_{gp} = 100$ und $l = 10$ spielen die Abhängigkeiten zwischen den Banden offenbar keine große Rolle: Die Topologieschätzug gelingt auf der Basis von D^{II} ebenso gut wie mit D^I. Hingegen führen die in D^{III} enthaltenen Verwechslungen von Banden zu einer deutlichen Erhöhung der Fehlerwahrscheinlichkeit. Das parsimonische Verfahren MP schneidet bei der Topologieschätzung auf der Basis von D^{III} etwas besser ab als die ML-Verfahren.

onsprogramm wie im vorangegangenen Abschnitt verwenden, indem wir n_{dna} und n_{gp} mit der Anzahl der verwendeten Primer multiplizieren.

Wir stellen uns vor, daß die Ergebnisse von 20 RAPD-Versuchen mit verschiedenen Primern mit einer 3 Milliarden Basenpaare langen DNA gepoolt wurden und daß auf dem Gel 100 Positionen unterscheidbar sind. Abbildung 3.2 zeigt Ergebnisse einer entsprechenden Simulation. Offensichtlich kann ML_1 auch in diesem Szenario mit D^{II} die Topologie praktisch genauso gut schätzen wie mit D^I. Abhängigkeiten zwischen den Banden scheinen also auch hier vernachlässigbar zu sein. Dasselbe Bild ergibt sich auch aus den in Abbildung 3.7 und 3.8 dargestellten Ergebnissen. Wir werden in Abschnitt 3.4.2.2 in Zusammenhang mit Abbildung 3.8 noch einmal darauf zu sprechen kommen, weshalb bei der Verwendung von D^{III} im allgemeinen mehr Fehler gemacht werden als etwa mit D^I.

3.4.1.3 Kurze Primer

Wir werfen noch einen Blick auf ein Szenario mit Primern, die nur fünf Basen lang sind. Bei Verwendung kurzer DNA-Sequenzen wären kurze Primer wünschenswert, da sonst kaum Banden auftreten. Mit den bislang labortechnisch möglichen Versuchsbedingungen scheint allerdings

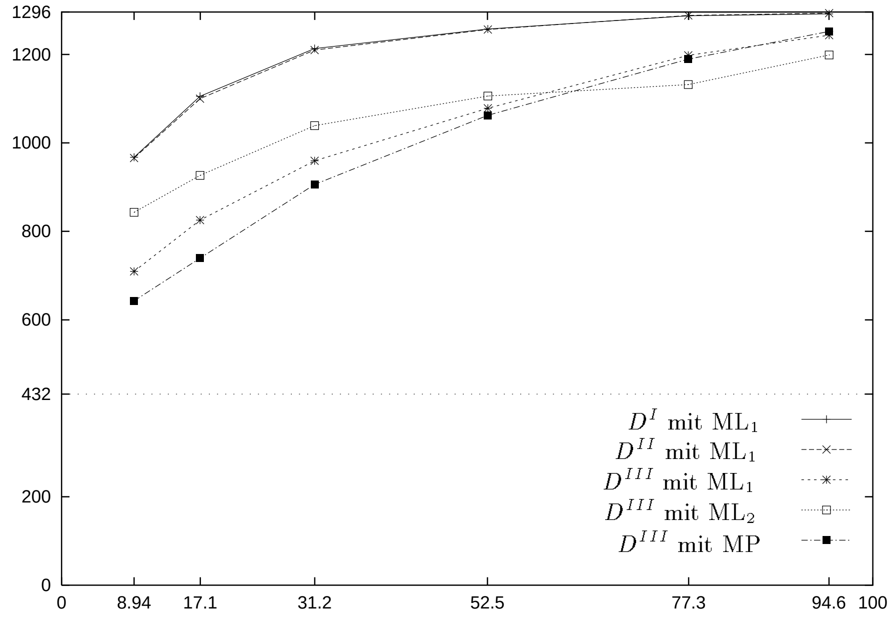


Abbildung 3.2: 20 Primer mit jeweils $l = 10$, $n_{gp} = 100$, $n_{\text{dna}} = 3 \cdot 10^9$ und $n_{\text{amp}} = 3000$. Die Verwechslung von Banden wirkt sich sehr deutlich aus, während Abhängigkeiten zwischen den Banden keine wesentliche Rolle spielen. Auf der Basis von D^{III} liefert ML_1 bessere Resultate als MP. Die besten Ergebnisse liefert dann allerdings ML_2, zumindest wenn der Zentralast des tatsächlichen Baumes nicht sehr lang ist.

ein Einsatz derartig kurzer Primer nicht sinnvoll zu sein, und ich wage nicht zu beurteilen, ob sich daran – etwa durch qualitative Verbesserungen der Labormaterialien – in absehbarer Zeit etwas ändern wird. Wir betrachten das Szenario hier als ein Extrembeispiel. Seine theoretische Bedeutung wird auch bei Clark und Lanigan (1993) deutlich.

Wir denken an eine DNA der Länge 100000 Basenpaare, einen Amplifikationsbereich von 3000 Basenpaaren und 1000 unterscheidbare Gelpositionen. Wir arbeiten wieder mit dem Klumpen-Modell, bei dem nur Klumpen mit ein bis zwei Banden auftreten, obgleich dieses hier nicht unbedingt als Approximation der Bandenverteilung, die sich aus dem Jukes-Cantor-Modell ergeben würde, aufgefaßt werden kann. Letztere würde nämlich einen nicht unerheblichen Anteil an Klumpen mit mehr als zwei Banden beinhalten, und die stochastischen Abhängigkeiten zwischen den Klumpen wären wohl nicht ohne weiteres zu vernachlässigen.

Wie bisher verwenden wir als mögliche Kantenlängen jene, die pro Site die erwarteten Mutationshäufigkeiten 0,00625, 0,0125, 0,025, 0,05, 0,1 und 0,2 hervorbringen. Übersetzt man dies in die Wahrscheinlichkeiten, daß eine Bande auf einer Kante der jeweiligen Länge verschwindet, so erhalten wir bei einer Primerlänge von $l = 5$ die Werte 0,58, 8,94, 17,0, 31,1, 52,3 und 76,8%.

Wir erwarten bei dem gegebenen Szenario bei jedem Individuum ungefähr 300 Banden. Wie wir in Abbildung 3.3 sehen, zeigen die Abhängigkeiten zwischen den Banden durchaus Wirkung.

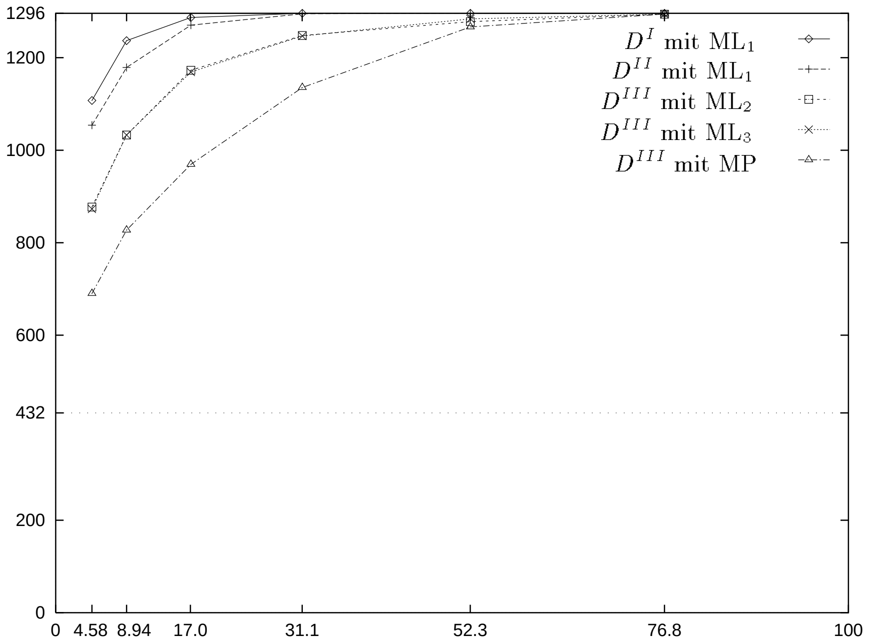


Abbildung 3.3: Bei $l = 5$, $n_{\text{dna}} = 100000$, $n_{\text{amp}} = 3000$ und $n_{gp} = 1000$ wirken sich neben den Verwechslungen von Banden auch Abhängigkeiten zwischen den Banden aus. Das parsimonische Verfahren MP schneidet außerdem deutlich schlechter ab als die ML-Verfahren.

Wenn man bedenkt, daß hier die meisten Klumpen (über 95%) zwei Banden enthalten, ist es allerdings eher erstaunlich, daß die Unterschiede zwischen den Fehlerkurven für die Auswertung von D^I und D^{II} mit ML$_1$ derartig gering sind. Allerdings macht sich auch hier das Verwechseln von Banden bemerkbar: Auf der Basis von D^{III} werden wesentlich mehr Fehler gemacht.

3.4.2 Vergleich der Methoden zur Topologieschätzung auf der Basis von D^{III}

Wir vergleichen nun die in Abschnitt 3.3 dargestellten Methoden. Wir gehen davon aus, daß in der Labor-Realität Banden nur im Rahmen einer gewissen Meßgenauigkeit vergleichbar sind, und verwenden daher für die Vergleiche simulierte Daten vom Typ D^{III}.

3.4.2.1 Nur ein Primer

Wir kommen zunächst wieder auf die Situation zurück, daß nur Daten vorliegen, die mit einer einzeigen RAPD-PCR-Primersequenz erzeugt wurden, und betrachten Abbildung 3.1. Das parsimonische Verfahren MP konnte bei dieser Parameterkombination mehr Punkte gewinnen als die beiden Maximum-Likelihood-Verfahren ML$_1$ und ML$_2$. Außerdem ist zu bemerken, daß ML$_1$, welches von einer Verteilung ausgeht, die offensichtlich nicht auf D^{III} zutrifft, ähnliche Zensuren erhält wie ML$_2$, obwohl letzteres zumindest ansatzweise die Besonderheiten von D^{III} berücksichtigt. Dies könnte dafür sprechen, daß das ML$_2$ zugrundeliegende Modell die Verteilung von D^{III} nicht hinreichend gut approximiert. Könnte es sein, daß hier

ML_3 eher angebracht ist? Für jede Gelposition erwartet man bei den gegebenen Parametern $\frac{n_{\mathrm{dna}} \cdot n_{\mathrm{amp}}}{n_{gp}} \cdot 2l \cdot \left(\frac{1}{4}\right)^{2l-1} = \frac{3 \cdot 10^9 \cdot 3000 \cdot 20}{100 \cdot 4^{19}} \approx 6{,}5$ Zwanzigtupel von Basen, von denen jedes ein Signal an der betreffenden Gelposition ergeben würde, wenn man nur eine einzige seiner Basen ändern würde. Daher würde sich so gut wie keine Gelposition im Zustand 3 des ML_3 zugrundliegenden Modells befinden. Da sich ML_3 somit im aktuellen Szenario praktisch auf zwei Zustände beschränkt, ist bei den gewählten Parametern kein großer Unterschied zwischen ML_2 und ML_3 zu erwarten.

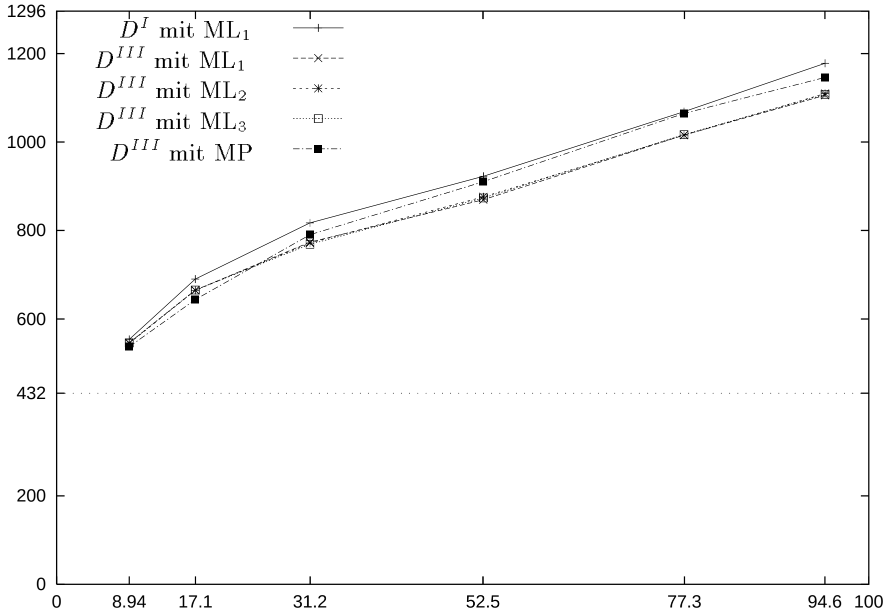


Abbildung 3.4: Bei $n_{\mathrm{dna}} = 3 \cdot 10^9$, $n_{\mathrm{amp}} = 3000$, $n_{gp} = 300$ und $l = 10$ schneidet MP bei der Topologieschätzung auf der Basis von D^{III} etwas besser ab als die ML-Verfahren. ML_3 bringt offensichtlich keine nennenswerten Vorteile gegenüber ML_2.

Wir machen die Probe auf's Exempel: Wir geben dabei ML_3 sogar noch eine etwas bessere Chance, indem wir die Anzahl der unterscheidbaren Gelpostionen auf $n_{gp} = 300$ erhöhen. Damit verringert sich die pro Gelposition erwartete Anzahl an Zwanzigtupeln, von denen eines genügen würde, damit sich die Gelposition nicht in Zustand 3 des ML_3 zugrundeliegenden Modells befindet, auf ungefähr 2,2.

Das Simulationsergebnis ist in Abbildung 3.4 zu sehen. Die drei Kurven für die Auswertung von D^{III} mit den ML-Verfahren verlaufen fast gleich und sind kaum zu unterscheiden. Anscheinend ist in diesem Szenario der ML-Ansatz relativ robust in bezug auf Veränderungen der zugrundegelegten Bandendynamik. Insgesamt schneidet bei den in Abbildung 3.4 gewählten Parametern das parsimonische Verfahren MP bei der Auswertung von D^{III} etwas besser ab als die ML-Verfahren. Allerdings sehen die ML-Verfahren bei Bäumen mit sehr kurzen Zentralästen etwas besser aus. Wir werden auf die Frage, wieso sich MP so gut gegenüber den

ML-Verfahren behaupten kann in Abschnitt 3.4.2.4 zurückkommen.

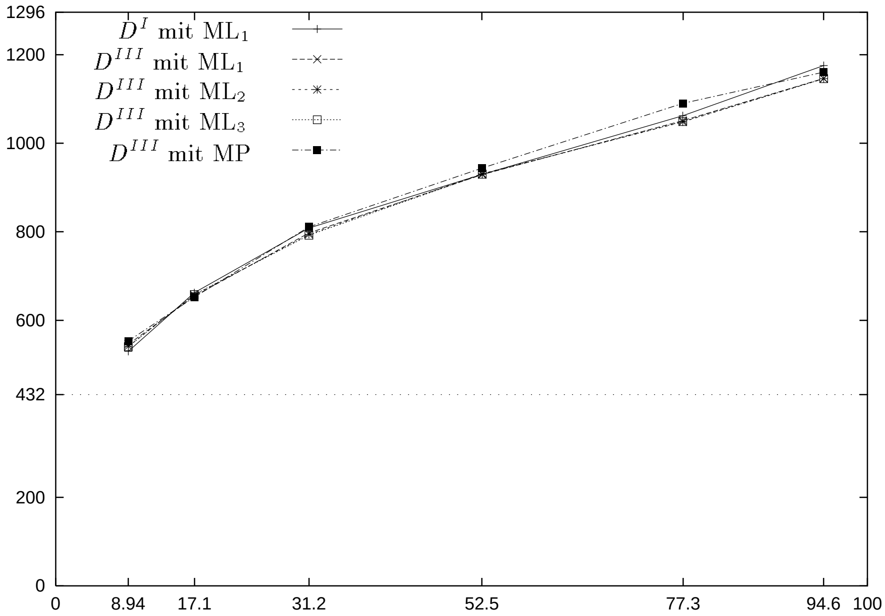


Abbildung 3.5: Bei $n_{gp} = 1000$, $n_{\text{dna}} = 3 \cdot 10^9$, $n_{\text{amp}} = 3000$ und $l = 10$ funktionieren alle Verfahren ungefähr gleich gut. Verwechslungen und Abhängigkeiten zwischen den Banden spielen hier keine große Rolle.

In Abbildung 3.5 ist zu sehen, daß die Verfahren in der Situation von 1000 unterscheidbaren Gelpositionen alle ungefähr gleich gut funktionieren. Es werden dann jeweils nur 3 Bandenlängen zu einer Gelposition zusammengefaßt, so daß nur wenige Verwechslungen von Bandenlängen auftreten, die dann offensichtlich nur wenig Einfluß auf die Qualität der Daten haben.

3.4.2.2 Mehrere Primer

Wir betrachten nun wieder Szenarien, in denen durch die Verwendung mehrerer Primer mehr Daten zur Verfügung stehen. Bekanntlich funktionieren Maximum-Likelihood-Verfahren besonders gut, wenn hinreichend viele Daten vorliegen. Dies läßt erwarten, daß sich in diesem Abschnitt die ML-Methoden gegenüber MP etwas besser behaupten können.

Abbildung 3.6 zeigt, wieviele Punkte die einzelnen Methoden bei Anwendung auf D^{III} sowie ML_1 bei D^I erhielten, wobei im Simulationsprogramm n_{gp} auf 3000 und n_{dna} auf 9 Milliarden gesetzt wurde. Man stelle sich dazu etwa vor, daß die DNA 3 Milliarden Basenpaare lang ist und daß 3 Primer verwendet werden, für die jeweils 1000 Gelpositionen unterscheidbar sind.

Offensichtlich gibt es auch hier keine nennenswerten Unterschiede zwischen den ML-Verfahren in ihrer Fähigkeit, die Baumtopologie aus dem Datensatz D^{III} zu schätzen. Der idealisierte Datensatz D^I liefert nur geringfügig bessere Ergebnisse. Das parsimonische Verfah-

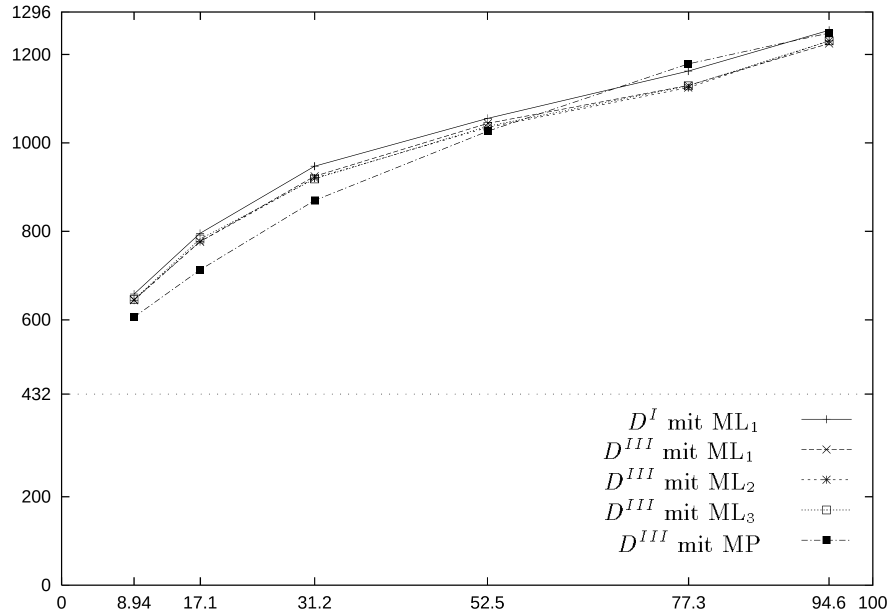


Abbildung 3.6: Auch bei 3 Primern mit jeweils $l = 10$, $n_{\text{dna}} = 3 \cdot 10^9$, $n_{\text{amp}} = 3000$ *und* $n_{gp} = 1000$ *funktionieren alle Verfahren ungefähr gleich gut.*

ren MP schneidet diesmal bei geringeren Zentralastlängen schlechter ab als die ML-Verfahren, dafür ist es bei längern Zentralästen noch etwas besser.

Wir vergrößern nun die Datenmenge weiter und stellen uns folgendes Szenario vor: Die DNA ist 3 Milliarden Basenpaare lang, auf dem Gel sind 100 Positionen unterscheidbar und es werden 20 verschiedene Primer verwendet. (Wir setzen also $n_{gp} = 2000$ und $n_{\text{dna}} = 60$ Milliarden.) Die in Abbildung 3.2 dargestellten Ergebnisse einer solchen Simulation lassen erstmals einen deutlichen Unterschied zwischen den Fehlerkurven von ML_2 und ML_1 bei Anwendung auf D^{III} erkennen. Bei Bäumen mit langen Zentralästen macht ML_1 etwas weniger Fehler, ansonsten ist ML_2 besser. Das parsimonische Verfahren MP ist, außer bei langen Zentralästen, bei der Auswertung von D^{III} etwas schlechter als ML_1.

In Abbildung 3.7 geht es um ein Szenario mit 30 Primern, für die jeweils 100 Gelpositionen unterscheidbar sind. Wie man sieht, sind dabei zwischen ML_2 und ML_3 keine Qualitätsunterschiede feststellbar. Die beiden ML-Verfahren konnten aber bei der größeren Datenmenge ihren Vorsprung gegenüber dem parsimonischen Verfahren MP weiter ausbauen. Nur bei der maximalen Zentralastlänge steht MP etwas besser da. Offenbar spielen Abhängigkeiten zwischen den Banden auch in diesem Szenario keine Rolle, denn die beiden Kurven für die Auswertung von D^I und D^{II} mit ML_1 sind fast gleich.

Abbildung 3.8 bezieht sich auf ein Szenario mit einer DNA-Länge von 3 Milliarden Basenpaaren und 10 Primern, bei dem die Banden als perfekt unterscheidbar angenommen werden. Alle Verfahren werden am Datensatz D^I (und ML_1 zusätzlich an D^{II}) erprobt. Eigentlich

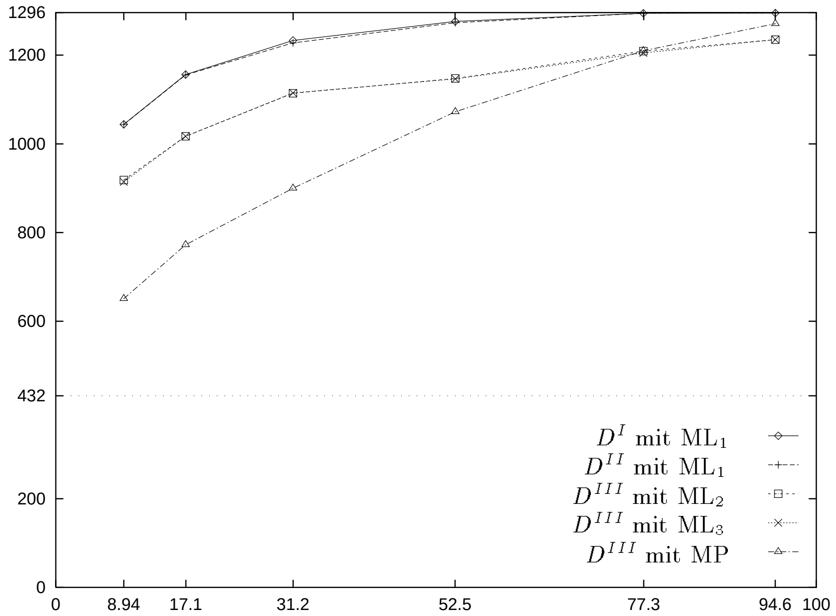


Abbildung 3.7: Bei 30 Primern mit jeweils $l = 10$, $n_{gp} = 100$, $n_{\text{dna}} = 3 \cdot 10^9$ und $n_{\text{amp}} = 3000$ macht MP bei der Topologieschätzung auf der Basis von D^{III} mehr Fehler als die ML-Verfahren – außer bei sehr langen Zentralästen. Qualitätsunterschiede zwischen ML_2 und ML_3 sind auch hier nicht zu erkennen.

müßte sich dabei der Umstand, daß von einem zum nächsten Knoten verschwundene Banden beim übernächsten Knoten mit erhöhter Wahrscheinlichkeit wieder entstehen, am ehesten bemerkbar machen. Daher könnte man annehmen, daß die Methode ML_2 bei diesen Parametern ihre Schwächen zeigt. Es stellt sich jedoch heraus, daß die Fehlerhäufigkeiten der ML-Verfahren kaum voneinander abweichen. Dies deutet darauf hin, daß sowohl die Abhängigkeiten zwischen den Banden als auch das „Gedächtnis" der Bandendynamik zu vernachlässigen sind, und legt die Vermutung nahe, daß bei der Verwendung von D^{III} im allgemeinen deswegen mehr Fehler gemacht werden, weil beim Verwechseln von Banden Information verlorengeht, und daß es keine große Rolle spielt, daß die ML_2 und ML_3 zugrundeliegenden Modelle die Verteilung von D^{III} nur ungenau beschreiben. Das parsimonische Verfahren ist – außer bei der maximalen Zentralastlänge – deutlich schlechter als die ML-Verfahren, die angesichts der Fülle an „guten Daten" offensichtlich „ganz in ihrem Element" sind.

3.4.2.3 Kurze Primer

Wir kommen noch einmal auf das Szenario mit kurzen Primern der Länge 5 Basen zurück und betrachten die in Abbildung 3.3 dargestellten Simulationsergebnisse. Die Verwendung von ML_3 bringt offenbar auch hier keine Vorteile gegenüber ML_2. Die ML-Verfahren schneiden wesentlich besser als das parsimonische Verfahren MP ab, was wieder daran liegen mag, daß recht viel Information zur Verfügung steht.

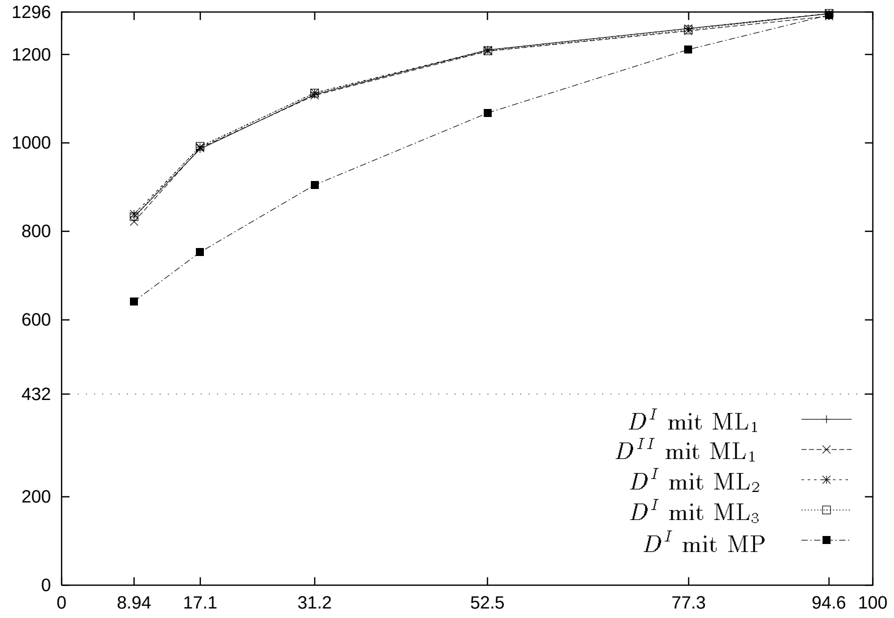


Abbildung 3.8: Bei 10 Primern mit jeweils $l = 10$, $n_{\text{dna}} = 3 \cdot 10^9$, $n_{\text{amp}} = 3000$ *und perfekt unterscheidbaren Banden schneiden alle drei ML-Verfahren gleich gut ab, und deutlich besser als MP.*

3.4.2.4 Wieso funktioniert MP in einigen Fällen besser als ML$_2$?

Wie zum Beispiel in Abbidung 3.1 zu sehen ist, gibt es Fälle, in denen das Verfahren ML$_2$ schlechter abschneidet als das sehr einfache Verfahren MP, das weniger Informationen aus den Daten benutzt. Die Abbildung 3.7 zeigt, daß auch bei großen Datenmengen das Verfahren ML$_1$ mehr Fehler als MP macht, wenn der Zentralast des wahren Baumes sehr lang ist. Wir möchten in diesem Abschnitt diskutieren, woran das liegen könnte.

Wir betrachten die Bäume aus dem Simulationslauf, dessen Ergebnisse in Abbildung 3.7 zu sehen sind, die maximale Zentralastlänge haben und bei denen ML$_1$ die falsche Topologie geschätzt hat. Da die Zentralastlänge fest ist, ist jeder Baum durch vier Zahlen, nämlich die Längen seiner Außenäste, gegeben. Zur graphischen Darstellung ordnen wir jedem Baum eine Linie im $\mathbb{R}^2$ zu, indem wir als Anfangspunkt den Punkt nehmen, der die beiden Längen der Kanten auf der einen Seite des Zentralastes als Koordinaten hat, und als Koordinaten des Endpunktes nehmen wir die Längen der beiden Kanten auf der anderen Seite des Zentralastes. Da nur 6 verschiedene Kantenlängen auftreten, würden einige Bäume zu übereinanderliegenden Linien führen. Um dies zu verhindern, verschieben wir in der Graphik die Anfangs- und Endpunkte jeder Linie jeweils um einen kurzen, zufälligen Vektor.

In Abbildung 3.9 ist auf der linken Seite zu sehen, daß alle Bäume mit maximaler Zentralastlänge aus dem Simulationslauf zu Abbildung 3.7, bei denen sich ML$_2$ geirrt hatte, mindestens einen Außenast maximaler Länge hatten. Die rechte Hälfte von Abbildung 3.9 stellt

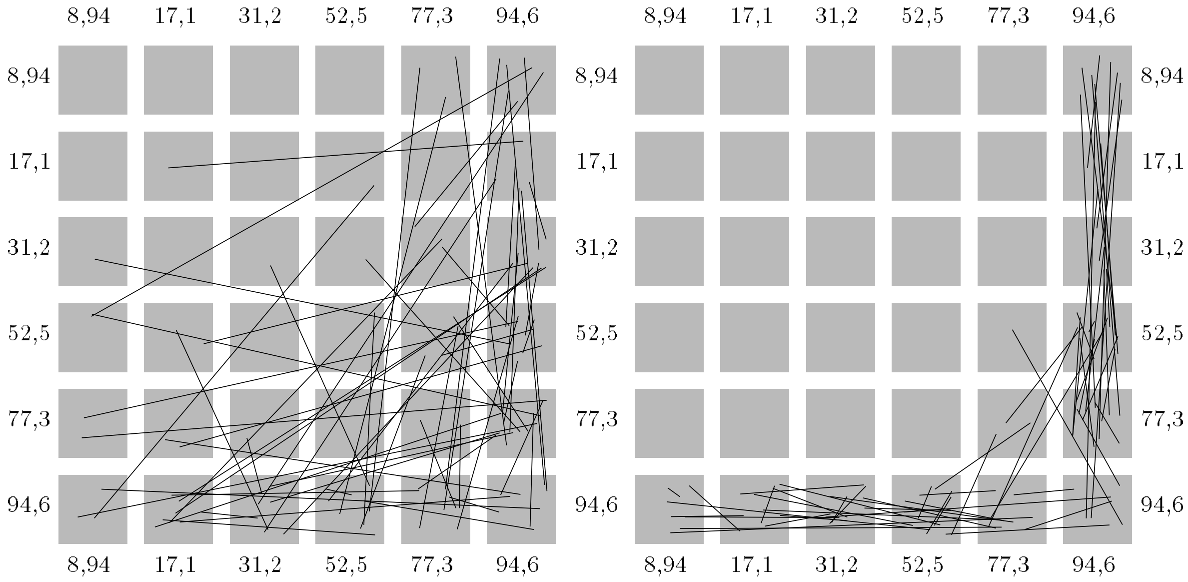


Abbildung 3.9: Darstellung derjenigen Bäume mit maximaler Zentralastlänge aus dem Szenario von Abbidung 3.7, bei denen sich ML_2 geirrt hat (links) und derjenigen Bäume, die in diesen Fällen die maximale Likelihood hatten (rechts). Die Endpunkte der Linien haben als Koordinaten jeweils die Längen benachbarter Außenkanten des zugehörigen Baumes.

die Außenastlängen der Bäume dar, die in diesen Fällen von ML_1 als Bäume mit maximaler Likelihood ausgegeben wurden. (Diese können verschiedene Zentralastlängen haben.) Wie man sieht, haben fast alle diese Bäume auf beiden Seiten des Zentralasts je eine Außenkante maximaler Länge. Außerdem laufen die Linien vorzugsweise entweder vertikal oder horizontal und nur wenige Ausnahmen sind schräg. Die beiden ersten Koordinaten des Anfangs- und des Endpunktes einer Linie sind die Längen der Kanten, die zu den Knoten führen, die im *wahren* Baum auf der einen Seite des Zentralastes liegen, und dementsprechend gehören die zweiten Koordinaten des Anfangs- und des Endpunktes zu Knoten, die im wahren Baum auf der anderen Seite des Zentralastes liegen. Typisch scheinen also solche Fehler zu sein, wie sie in Abbildung 3.10 dargestellt werden.

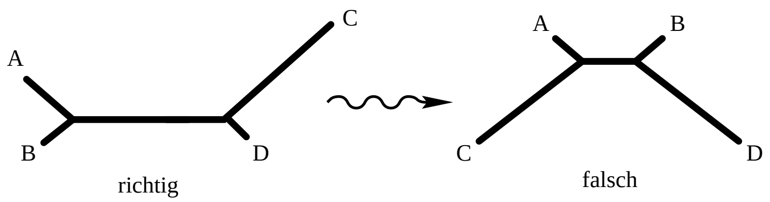


Abbildung 3.10: Wenn der links dargestellte Baum der wahre ist, so ähnelt der Baum mit der maximalen Likelihood nach Abbildung 3.9 offenbar in vielen Fällen dem rechts dargestellten.

Da der zu C führende Außenast des linken (richtigen) Baumes maximale Länge hat, ist es denkbar, daß C keine gemeinsame Bande mit den anderen Knoten hat. Die einzige Information, die wir über C erhalten, ist dann, daß es im wahren Baum weit von allen anderen Blättern entfernt ist, und dies ist auch beim rechten (falschen) Baum der Fall. Die Verwandtschaft von A, B und C untereinader ist in beiden Bäumen einschließlich der ungefähren Entfernungen zueinander dieselbe. Das Verfahren MP macht den in Abbildung 3.10 dargestellten Fehler nur selten. Umgekehrt besteht bei MP die Gefahr, daß die Topologie des linken Baumes ausgegeben wird, wenn der tatsächliche Baum der rechte ist. Dies liegt daran, daß auch im rechten Baum A und B am engsten miteinander verwandt sind und daß parsimonische Verfahren die Tendenz haben, die engsten Verwandten zusammenzufassen (vgl. Huelsenbeck und Hillis (1993)). Dies trägt wohl wesentlich dazu bei, daß MP bei den Simulationsläufen mit größeren Datenmengen (siehe Abbildungen 3.7 und 3.8) bei kurzen Zentralästen im Vergleich zu den ML-Verfahren besonders schlecht abschneidet.

3.4.2.5 Vergleich von ML_2 mit DNAML

Die Methode ML_2 ist ein Analogon der klassischen Maximum-Likelihood-Stammbaumrekonstruktionsmethoden für den Fall einer diskreten Menge möglicher Astlängen. Wir vergleichen nun die Fehlerhäufigkeiten von ML_2 mit denen des Programms DNAML (Felsenstein, 1993) in den Test-Szenarien, die den Abbildungen 3.1 und 3.7 zugrundliegen. Die simulierten RAPD-Daten wurden zunächst in Pseudo-DNA-Sequenzdaten übersetzt. Für jede zweite Gelposition wurde dazu bei jedem Taxon, bei dem an dieser Gelposition keine Bande zu sehen war, an dem entsprechenden Site ein A und bei allen Taxa, bei denen an dieser Position ein Bandensignal auftrat, ein G eingetragen. Für die übrigen Gelpositionen wurde an den entsprechenden Sites das Vorhandensein einer Bande mit T und die Abwesenheit der Bande mit C gekennzeichnet. Das Entstehen eines Bandensignals übersetzt sich also in eine Punktmutation von A nach G oder von C nach T, und das Verschwinden eines Bandensignals übersetzt sich in eine Punktmutation von G nach A oder von T nach C. Die genannten Mutationen heißen Transitionen, die übrigen heißen Transversionen. Da DNAML auch auf der Basis des Kimura-2-Modells (vgl. Kimura (1983)) schätzen kann, nach dem Transversionen und Transitionen in unterschiedlichen Häufigkeiten auftreten, können wir berücksichtigen, daß in unserem Pseudo-DNA-Datensatz keine Transversionen auftreten, indem wir den DNAML-Programmparameter für das Ratenverhältnis zwischen Transversionen und Transitionen auf einen sehr großen Wert (z. B. 1000) setzen. Die erwarteten Häufigkeiten von A, G, C und T lassen wir DNAML aus dem Datensatz schätzen. Außerdem wird die Eingabereihenfolge der zu den Taxa gehörigen Pseudo-DNA-Sequenzen zufällig permutiert.

Die Ergebnisse sind in den Abbildungen 3.11 und 3.12 zu sehen. DNAML schneidet jeweils insgesamt geringfügig schlechter als ML_2 ab. Daß sich die Abweichungen zwischen den beiden Verfahren in Grenzen halten, deutet darauf hin, daß wir unsere bisherigen Erkenntnisse über ML_2 wohl auch großteils auf den Fall kontinuierlicher Kantenlängen übertragen können. Daß DNAML etwas schlechter als ML_2 abschneidet, ist insofern nicht überraschend, als daß wir dem

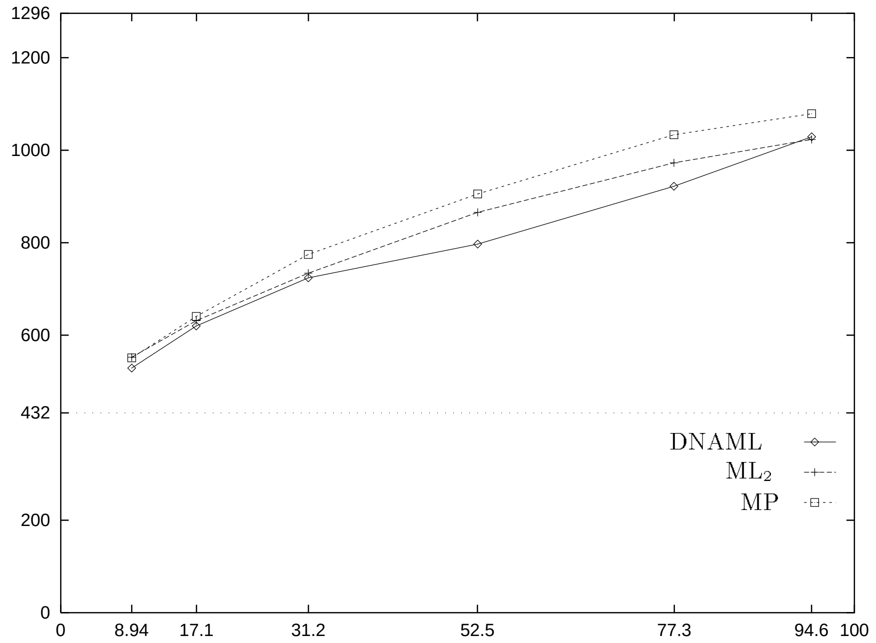


Abbildung 3.11: Vergleich von DNAML mit ML_2 und MP für $l = 10$, $n_{\mathrm{dna}} = 3 \cdot 10^9$, $n_{\mathrm{amp}} = 3000$ *und* $n_{gp} = 100$ *(vgl. Abbildung 3.1).*

Programm weniger Information zur Verfügung gestellt haben. Im Gegensatz zu ML_2 verfügt nämlich DNAML insbesondere nicht über die Information, daß bei unseren Testbäumen für jede Kante nur sechs verschiedene Längen möglich sind.

Es fällt auf, daß bei beiden Test-Szenarien DNAML bei Bäumen mit maximaler Zentralastlänge geringfügig weniger Fehler macht als ML_2. Eine mögliche Erklärung für diese Beobachtung ist die folgende: Für ML_2 sind die beiden in Abbildung 3.10 gezeigten Bäume nicht nur wegen der Ähnlichkeit ihrer Bandenverteilungen schwer zu unterscheiden, der rechte Baum wird womöglich auch noch tendenziell bevorzugt, wodurch die Fehlerrate für Bäume mit langen Zentralästen steigt (und die für Bäume mit kurzen Zentralästen sinkt). Eine derartige Bevorzugung der Bäume mit kürzeren Zentralästen könnte bei ML_2 folgendermaßen entstehen: Da die Taxa C und D in Abbildung 3.10 jeweils sehr weit von allen anderen entfernt sind, ist die meiste Information darüber, wie sie mit den jeweils anderen verwandt sind, durch viele Mutationen verrauscht, so daß nur klar ist, daß sie mit keinem anderen eng verwandt sind. Die Information, die aus der Sicht von ML_2 für die Topologieschätzung am besten verwertbar erscheint, ist daher der Abstand zwischen A und B. Der von ML_2 implizit verwendete Schätzer für diesen Abstand unterliegt allerdings zufälligen Schwankungen, die dazu führen können, daß der geschätzte Abstand zwischen A und B auf einen anderen Baum hindeutet. Dabei ist es wahrscheinlicher, daß ein Baum wie der linke in Abbildung 3.10 für einen dem rechten ähnelnden Baum gehalten wird, als umgekehrt. Dies ergibt sich daraus, daß nur sechs verschiedene Längen pro Kante möglich sind, und in der Menge der Bäume mit der Topologie

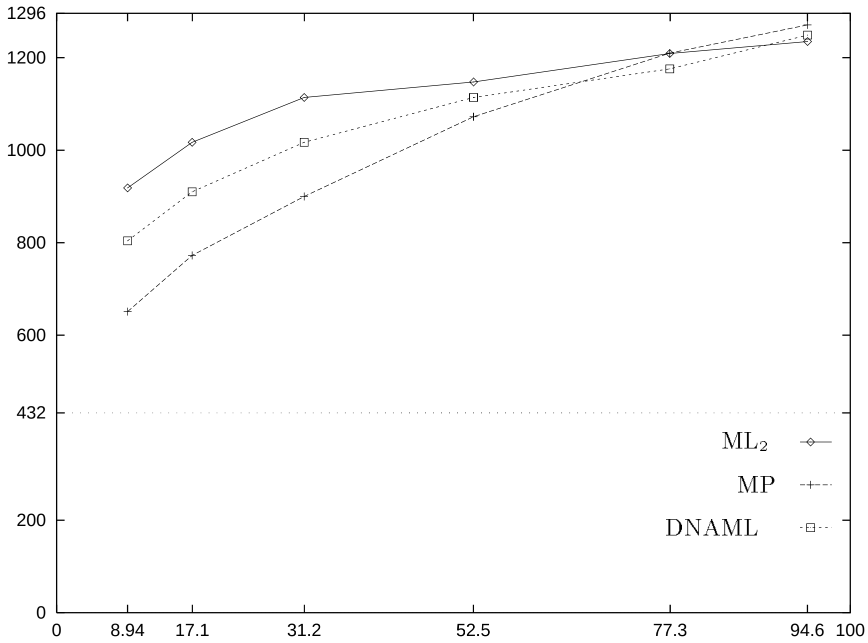


Abbildung 3.12: Vergleich von DNAML mit ML_2 und MP für ein Szenario mit 30 Primern mit jeweils $l = 10$, $n_{gp} = 100$, $n_{\mathrm{dna}} = 3 \cdot 10^9$ und $n_{\mathrm{amp}} = 3000$ (vgl. Abbildung 3.7).

des rechten Baumes mehr verschiedene Abstände zwischen A und B auftreten, da bei dieser Topologie zwischen A und B drei Kanten liegen. Dieser Effekt tritt bei DNAML nicht auf, da dieses Verfahren auf einem Modell basiert, bei dem ohnehin für jede Kante beliebige Längen möglich sind.

3.5 Fazit

Am Beispiel der Topologieschätzung vierblättriger Stammbäume haben wir gesehen, wie man das in Kapitel 2 entwickelte Poisson-Cluster-Modell für RAPD-Banden benutzen kann, um Auswertungsverfahren für RAPD-Daten zu beurteilen. Die Simulationsergebnisse legen den Schluß nahe, daß zumindest für die exemplarisch betrachtete Anwendung folgende Faustregeln angegeben werden können:

- Abhängigkeiten zwischen den Banden im Sinne der Kapitel 1 und 2 können vernachlässigt werden.

- Die Likelihood eines Baumes für gegebene RAPD-Daten läßt sich mit dem für ML$_2$ angegebenen Algorithmus hinreichend genau approximieren. Die Tatsache, daß die Evolution der Bandensignale auf den Zuständen {„sichtbar“, „nicht sichtbar“} nicht Markoff'sch ist, kann also vernachlässigt werden.

- Bei der Berechnung der Likelihood von Bäumen auf der Basis von RAPD-Daten sollte (wie bei ML_2) die Möglichkeit berücksichtigt werden, daß Banden, die von unterschiedlichen Abschnitten auf der DNA kommen, auf dem Gel ununterscheidbar sein könnten.

- Liegen nur wenige RAPD-Daten vor, so sind Punktschätzungen von Baumtopologien sehr fehleranfällig, insbesondere auch wenn Maximum-Likelihood-Schätzer verwendet werden.

- Bei größeren Datenmengen liefern Maximum-Likelihood-Ansätze bessere Resultate als parsimonische Verfahren, insbesondere wenn der Zentralast des wahren Baumes sehr kurz ist. Außerdem erhält man Informationen über die Sicherheit einer Maximum-Likelihood-Schätzung, wenn man für jede der drei Baumtopologien vierblättriger Bäume die Likelihood berechnet und die Werte dann vergleicht. Dies spielt auch eine Rolle bei Verfahren, bei denen Topologieschätzer größerer Bäume aus geschätzten Topologien vierblättriger Bäume zusammengesetzt werden, vgl. Strimmer und von Haeseler (1996, 1997).

Es sei daran erinnert, daß wir vom Jukes-Cantor-Modell ausgegangen sind. Obige Aussagen können sich also nur auf Situationen beziehen, in denen dieses Modell für die Sequenzevolution angemessen ist. Völlig andere Effekte sind denkbar, wenn außer den Basen-Substitutionen noch andere Arten von Mutationen – etwa Insertionen und Deletionen – eine bedeutende Rolle spielen.

Anhang A

Molekulargenetische Grundlagen

A.1 Die Desoxyribonukleinsäure (DNA)

In der belebten Welt wird die Erbinformation in der molekularen Struktur der Nukleinsäuren gespeichert. Man unterscheidet Desoxyribonukleinsäure (DNA, engl. *deoxyribo nucleic acid*) und Ribonukleinsäure (RNA, engl. *ribo nucleic acid*). Nukleinsäuren sind lange Ketten von Nukleotiden. Diese bestehen jeweils aus einem Zuckermolekül, einem Phosphorsäuremolekül und einer Base. Für die Basen kommen bei der DNA, auf die wir uns beschränken, Adenin (A), Guanin (G), Cytosin (C) und Thymin (T) in Frage. Die Erbinformation ist durch die Folge der Basen kodiert, also sozusagen in einem vierbuchstabigen Alphabet niedergeschrieben. Die Struktur der DNA wurde in den fünfziger Jahren von J. Watson und F. Crick (1953) entdeckt. DNA-Moleküle bestehen im allgemeinen aus zwei gegenüberliegenden Nukleotid-Strängen. Durch die molekulare Zusammensetzung der Nukleotide ist für jeden Strang eine Richtung vorgegeben, die sogenannte 5'-3'-Richtung. Die beiden Einzelstränge sind zueinander gegenläufig ausgerichtet und werden durch Wasserstoffbrückenbindungen zwischen den Basen zusammengehalten. Dabei liegen sich jeweils A und T, sowie C und G gegenüber (siehe Abbildung A.1). Die beiden

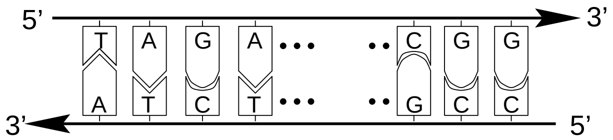


Abbildung A.1: Schematische Darstellung eines DNA-Doppelstrangs

Einzelstränge sind also komplementär zueinander und enthalten jeweils die gesamte Information. Dies ist für die Vervielfältigung der Information – wie etwa zum Zwecke der Fortpflanzung – wesentlich. Dazu werden nämlich die beiden Einzelstränge voneinander getrennt, und jeder der beiden wird durch Anlagerung von Nukleotiden, die die jeweils komplementären Basen enthalten, zu einem Doppelstrang ergänzt. Die Synthese der neu entstehenden Einzelstränge geschieht dabei in 5'-3'-Richtung. Die Vervielfältigung der Information funktioniert nicht immer ganz fehlerfrei: Es treten verschiedene Arten von Mutationen auf. Dabei können einzelne oder ganze Gruppen von Basen vertauscht werden, es können Basen wegfallen oder zusätzliche eingefügt werden. Wir beschränken uns auf Mutationen, bei denen die Base an einer einzelnen Position durch eine andere Base ersetzt wird.

A.2 Die Polymerasekettenreaktion (PCR)

Die Polymerasekettenreaktion (PCR, engl. *polymerase chain reaction*) ist ein Laborverfahren zur Amplifikation (Vervielfältigung) von Abschnitten der DNA, die einige tausend Basenpaare lang sein können (vgl. Hadrys *et al.* (1992)). Soll ein bestimmter Abschnitt amplifiziert werden, so müssen von seinem Anfang und seinem Ende (in 5'-3'-Richtung auf einem der beiden Einzelstränge betrachtet) jeweils circa 10 bis 20 Basen bekannt sein. Die DNA-Vorlage wird dann als Doppelstrang zusammen mit jeweils einer großen Anzahl an Duplikaten der Anfangssequenz und Komplementen der Endsequenz sowie mit einzelnen Nukleotiden der verschiedenen Typen in eine Ionenlösung gegeben. Hinzu kommt Polymerase, das Enzym, welches die Anbindung der passenden Nukleotide an DNA-Einzelstränge katalysiert. Durch Erhöhen der Temperatur werden die beiden Doppelstränge voneinander getrennt. Durch anschließendes leichtes Absenken der Temperatur wird ermöglicht, daß sich die Komplemente der Endsequenz am Ende des zu amplifizierenden Abschnitts anlagern. Außerdem können sich die Duplikate der Anfangssequenz an dem zum Anfang der Sequenz komplementären Abschnitt des zweiten Einzelstrangs anlagern. Nach nochmaligem Absenken der Temperatur können einzelne Nukleotide an die beiden Einzelstränge anbinden. Dies kann allerdings nur dort geschehen, wo sich bereits Nukleotide angelagert haben, also nur dort, wo sich bereits ein Komplement der Endsequenz oder ein Duplikat der Anfangssequenz des zu amplifizierenden Abschnitts befindet. Daher werden diese Moleküle als *Primer* bezeichnen, ihre Komplemente als *Inverted Repeats*. Von einem Primer ausgehend, werden die Nukleotide also in 5'-3'-Richtung zusammengesetzt, und im Bereich des zu amplifizierenden Abschnitts entsteht ein Doppelstrang, siehe Abbildung A.2. Die entstehenden Teilstränge beginnen am 5'-Ende mit den Primersequenzen. Je nach der Qualität der Polymerase beträgt die Länge des möglichen Amplifikationsbereichs einige tausend Basenpaare. Wichtig ist dabei, daß der entstehende Strang auch das zweite Inverted Repeat enthält, so daß er auch selbst wieder als Amplifikationsvorlage dienen kann. Durch wiederholtes Erhitzen und Abkühlen werden die entstehenden Doppelstränge nämlich immer wieder getrennt und somit werden weitere Kopien des betreffenden Abschnitts hergestellt. Diese beginnen dann mit einer Primersequenz und enden mit dem Komplement das jeweils anderen Primers, siehe Abbildung A.3. Es werden üblicherweise ungefähr 40 Zyklen des Erhitzens und Abkühlens durchgeführt. Würde wirklich in jedem Zyklus an jedem Inverted Repeat ein Primer anbinden und eine Amplifikation in Gang setzen, so würde die Anzahl der DNA-Fragmente in jedem Zyklus verdoppelt. Dies scheint aber nicht der Fall zu sein; mehr dazu in Anhang B.

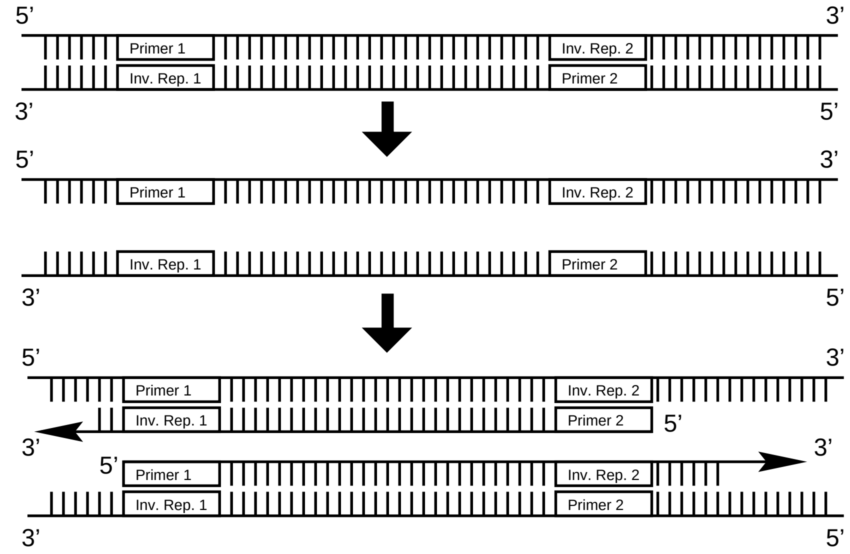


Abbildung A.2: Der erste Zyklus der PCR

A.3 Genetische Fingerabrücke mit der RAPD-PCR

Die PCR eröffnet eine kostengünstige und schnelle Möglichkeit, genetische Fingerabdrücke herzustellen. Dazu werden nicht spezifische Primer mit bekannten Anbindungsstellen zu der DNA-Vorlage gegeben, sondern es wird nur ein relativ kurzer Primer (etwa 10 Nukleotide lang) verwendet. Wenn die Primersequenz – mehr oder weniger zufällig – irgendwo auf der DNA vorkommt und innerhalb weniger tausend Basenpaare ein Inverted Repeat der Primersequenz folgt, so wird der betreffende Abschnitt vervielfältigt. Dieses Verfahren heißt RAPD-PCR (RAPD=Abk. f. engl. *random amplified polymorphic DNA*). Am Ende überprüft man, welche Fragmentlängen vorkommen. Dazu unterzieht man die Amplifikationsprodukte einer Elektrophorese. Man erhält dabei ein Gel, in welchem die Fragmente je nach Länge unterschiedlich weit „gewandert“ sind. Ab einer gewissen Konzentration kann man Anhäufungen von DNA-Fragmenten auf dem Gel sichtbar machen. Man erhält dann auf dem Gel einen Streifen (eine sogenannte *Bande*), an dessen Position die Länge der zugehörigen DNA-Fragmente abgelesen werden kann. Allerdings ist es durchaus möglich, daß PCR-Produkte, die von verschiedenen Abschnitten der DNA-Vorlage kommen, dieselbe oder fast dieselbe Länge haben und daher auf dem Gel nicht unterscheidbar sind (vgl. Kapitel 3). Wir verstehen dann allerdings unter der „Bande“ das Paar der Positionen, an denen sich die Kopie und das Komplement des Primers befinden (vgl. Abschnitt 1.1.2).

Genetische Fingerabdrücke haben in der Biologie, der Medizin und natürlich auch in der Forensik ein weites Anwendungsfeld. Wir haben in erster Linie die biologischen Anwendungen im Blick. Durch Vergleiche der genetischen Fingerabdrücke verschiedener Individuen einer Population sollen Erkenntnisse über deren Verwandtschaft erzielt werden. Daher spielt die RAPD-PCR in der Ökologie eine gewisse Rolle (siehe Hadrys, 1992). In der Evolutionsforschung werden ge-

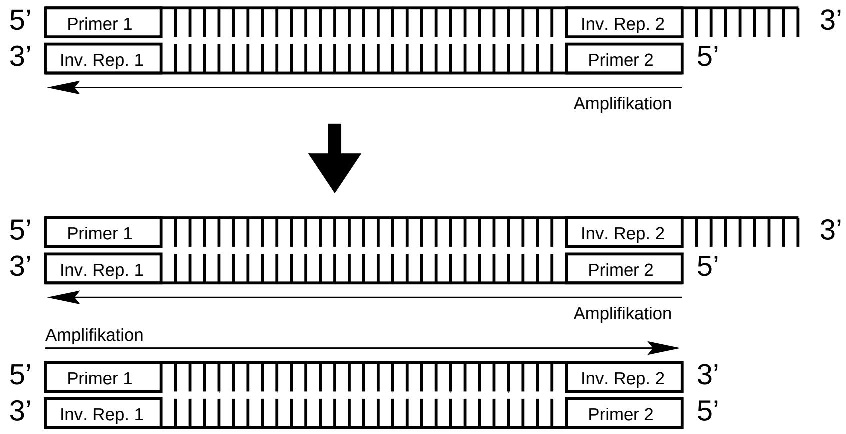


Abbildung A.3: Nach dem zweiten Zyklus der PCR gibt es Amplifikationsprodukte, die – in 5'-3'-Richtung betrachtet – mit einer der Primersequenzen beginnen und mit dem Komplement der anderen Primersequenz enden.

netische Fingerabdrücke verschiedener Arten benutzt, um Aussagen über deren gemeinsamen Stammbaum zu machen.

Einige Biologen stehen der RAPD-PCR kritisch gegenüber. Eine Befürchtung ist etwa, daß die Ergebnisse der RAPD-PCR in gewissen Situationen nicht reproduzierbar sind, da es vom Zufall abhängt, ob an den Inverted Repeats in den einzelnen PCR-Zyklen Primer anbinden. Dies wird in Anhang B ausführlich diskutiert. Eine weitere Frage ist, ob davon ausgegeangen werden kann, daß die Banden unterschiedlicher Länge voneinander stochastisch unabhängig sind (vgl. Abschnitt 2.1). Die meisten Untersuchungen in dieser Arbeit haben mit dieser Frage zu tun.

Anhang B

Spielen konkurrierende RAPD-Banden eine Rolle?

B.1 Modell einer PCR mit überlappenden Banden

Wir befassen uns nun mit einer Situation, bei der von einer DNA-Vorlage in jedem PCR-Zyklus nur jeweils eine von zwei verschiedenen Banden amplifiziert werden kann. Wir gehen davon aus, daß es vom Zufall abhängt, welche Bande jeweils amplifiziert wird, und untersuchen, ob diese Art von Zufall zu nicht reproduzierbaren Versuchsergebnissen führen kann. Einige Ergebnisse dieses Abschnitts sind bereits in eine Veröffentlichung eingeflossen (siehe Schierwater *et al.* (1996)). Dort findet man auch weitere Details zur RAPD-PCR.

Wir stellen uns nun vor, daß in der DNA-Vorlage einer RAPD-PCR auf eine Kopie der Primersequenz innerhalb des Amplifikationsbereichs zwei Primerkomplemente folgen (siehe Abbildung B.1). Wenn ein Fragment eine Primerkopie und zwei darauf (in 5'-3'-Richtung) folgende

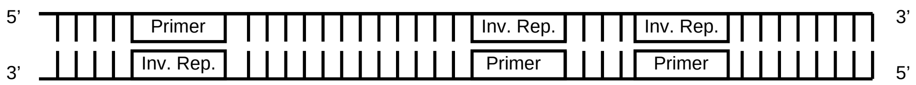


Abbildung B.1: Wenn auf einem der beiden Einzelstränge auf eine Primersequenz innerhalb des Amplifikationsbereichs zwei Primerkomplemente folgen, können Banden mit zwei verschiedenen Längen auftreten: Es kann der Abschnitt zwischen dem Primer und dem ersten Komplement sowie der Abschnitt zwischen dem Primer und dem zweiten Komplement amplifiziert werden.

Primerkomplemente enthält, so gibt es in einem PCR-Zyklus vier Möglichkeiten:

1. Es bindet kein Primer an und es wird nichts amplifiziert.

2. Nur am ersten Komplement bindet ein Primer an und es wird das Stück zwischen der Primerkopie und dem ersten Komplement amplifiziert.

3. Nur am zweiten Komplement bindet ein Primer an und es wird das gesamte Fragment amplifiziert.

4. Sowohl am ersten als auch am zweiten Komplement bindet ein Primer an. Was nun passiert, hängt von der Polymerase ab. Wir gehen davon aus, daß eine Polymerase verwendet wird, die wie die meisten folgendes bewirkt (vgl. Schierwater *et al.* (1996)): Das Stück zwischen der Primerkopie und dem ersten Komplement wird amplifiziert, und ausgehend von dem Primer, der am zweiten Komplement angebunden ist, beginnt eine Amplifikation, die abbricht, sobald sie auf den Primer trifft, der am ersten Komplement angebunden ist. Es entsteht also ein Fragment, welches der kürzeren Bande entspricht und ein kurzes Reststück, aus dem in späteren Zyklen nichts amplifiziert werden kann, da es kein Primerkomplement enthält (siehe Abbildung B.2).

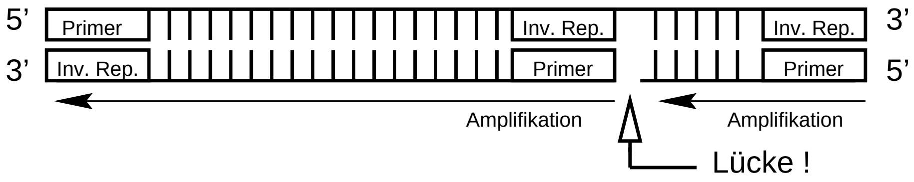


Abbildung B.2: Wenn eine Amplifikation auf einen bereits angebundenen Primer trifft, wird sie dort unterbrochen und es entstehen zwei neue Fragmente.

Je nachdem, wo Primer anbinden, kann es also zur Synthese der längeren oder der kürzeren Bande kommen. In diesem Sinne konkurrieren die Banden miteinander. Kann es sein, daß bei einem PCR-Versuch durch Zufall nur von der längeren Bande und in einem anderen PCR-Versuch nur von der kürzeren Bande eine hinreichend große Menge an Kopien entsteht?

Wir nehmen nun an, daß jede Primeranbindung in jedem Zyklus und an jede mögliche Anbindungsstelle mit derselben Wahrscheinlichkeit p erfolgt und daß Primeranbindungen an verschiedenen Anbindungsstellen und/oder in verschiedenen PCR-Zyklen stochastisch bedingt unabhängige Ereignisse sind, gegeben daß die Fragmente, auf denen die betreffenden Anbindungsstellen liegen, im jeweiligen Zyklus existieren. (Ähnliche Modellannahmen werden auch in Krawczack *et al.* (1989), Nedelman *et al.* (1992) sowie Weiss und von Haeseler (1995) zur Untersuchung verschiedener Fragen zur PCR verwendet.)

Wir teilen nun die beteiligten Fragmente in vier verschiedene Typen auf: In Typ A fassen wir alle Fragmente zusammen, die jeweils genau eine Primerkopie und ein Inverted Repeat des Primers enthalten; Typ B umfasst alle Fragmente, die zwei Primerkopien und ein Komplement enthalten. Ein Fragment ist vom Typ C, wenn es eine Primerkopie und zwei Komplemente enthält und vom Typ D, wenn es keine Primerkopie enthält. In Abbildung B.3 sind die Typen zusammen mit den Wahrscheinlichkeiten, daß sie in einem fest gewählten PCR-Zyklus die Synthese von Molekülen der anderen Typen bewirken, dargestellt.

Die Typklassen A, B und C enthalten auch Fragmente, bei denen die in Abbildung B.3

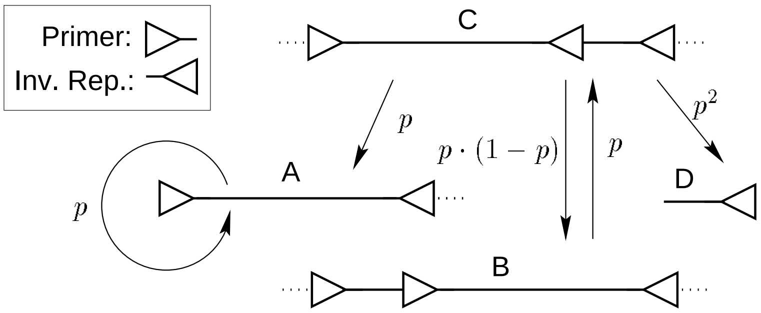


Abbildung B.3: Die im Szenario der konkurrierenden Banden an der PCR beteiligten Fragment-Typen. Der Pfeil von einem Typ X zu einem Typ Y ist mit der Wahrscheinlichkeit beschriftet, daß in einem fest gewählten PCR-Zyklus an einem gegebenen Fragment vom Typ X ein Fragment vom Typ Y entsteht.

durch gepunktete Linien angedeuteten Abschnitte vorkommen. Solche Fragmente können aber nur die Originalstränge sein oder solche Stränge, die in einem der Zyklen direkt an einem Originalstrang amplifiziert wurden (vgl. Abbildung A.3). Nach 40 Zyklen können also höchstens 82 mal so viele Fragmente mit Reststücken existieren wie am Anfang und damit zu wenige, um irgendein Bandensignal hervorzurufen. Wir können also die gepunkteten Abschnitte in Abbildung B.3 vernachlässigen.

B.2 Erwartungswerte und asymptotische Betrachtungen

Es sei A_i, B_i, C_i und D_i für $i \in \{1, 2, \ldots\}$ jeweils die Anzahl der Moleküle vom Typ A, B, C bzw. D nach dem i-ten Zyklus und für $i = 0$ die jeweilige (nicht zufällige) Anzahl zu Beginn der PCR. Es sei $X_i = (A_i, B_i, C_i, D_i)$. Wir gehen von $X_0 = (0, m, m, 0)$ aus. Wie wir aus B.3 ablesen können, gilt $\mathbb{E}(X_{i+1}|X_i) = X_i \cdot M$ mit

$$M = \begin{pmatrix} p+1 & 0 & 0 & 0 \\ 0 & 1 & p & 0 \\ p & p(1-p) & 1 & p^2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Aus $\mathbb{E}X_i = \mathbb{E}[\mathbb{E}(X_i|X_{i-1})] = \mathbb{E}[X_{i-1}M] = (\mathbb{E}X_{i-1})M$ (für $i \geq 1$) folgt $\mathbb{E}X_n = (0, m, m, 0) \cdot M^n$ für alle $n \geq 1$.

Die Eigenwerte mit zugehörigen linken Eigenvektoren von M sind:

$$\begin{array}{ll}
\lambda_1 := 1 + p & \left(1 \;,\; 0 \;,\; 0 \;,\; 0 \right) \\
\lambda_2 := 1 + p\sqrt{1-p} & \left(-\frac{1-p+\sqrt{1-p}}{p^2} \;,\; \frac{1-p}{p} \;,\; \frac{\sqrt{1-p}}{p} \;,\; 1 \right) \\
\lambda_3 := 1 & \left(0 \;,\; 0 \;,\; 0 \;,\; 1 \right) \\
\lambda_4 := 1 - p\sqrt{1-p} & \left(-\frac{1-p-\sqrt{1-p}}{p^2} \;,\; \frac{1-p}{p} \;,\; -\frac{\sqrt{1-p}}{p} \;,\; 1 \right)
\end{array}$$

Sei T die Matrix, die man erhält, wenn man die linken Eigenvektoren untereinander schreibt, und D sei die Diagonalmatrix mit den Eigenwerten als Diagonaleinträgen (beides in der oben gegebenen Reihenfolge). Wegen $M = T^{-1}DT$ gilt dann für $n \geq 0$:

$$\mathbb{E}X_n = (0, m, m, 0)T^{-1}D^nT = \begin{pmatrix} 2m\left(\frac{1}{p} + f_1(n,p)\right) \cdot {\lambda_1}^n \\ \frac{1}{2}m\left(1 + \sqrt{1-p} + f_2(n,p)\right) \cdot {\lambda_2}^n \\ \frac{1}{2}m\left(1 + \frac{1}{\sqrt{1-p}} + f_3(n,p)\right) \cdot {\lambda_2}^n \\ \frac{1}{2}m\frac{p}{1-p}\left(1 + \sqrt{1-p} + f_4(n,p)\right) \cdot {\lambda_2}^n \end{pmatrix}$$

Dabei konvergieren die Funktionen f_1, f_2, f_3 und f_4 exponentiell schnell gegen 0, wenn ihr erstes Argument gegen ∞ strebt.

Das Teilsystem $(B_i, C_i)_{i \geq 0}$ bildet einen irreduziblen, superkritischen 2-Typ-Galton-Watson-Prozeß, und $\lambda_2 = 1 + p \cdot \sqrt{1-p}$ ist der größte Eigenwert seiner Mittelwertmatrix

$$\begin{pmatrix} 1 & p \\ p(1-p) & 1 \end{pmatrix}.$$

Nach einem klassischen Satz aus der Theorie der Verzweigungsprozesse (siehe Athreya und Ney, 1972) konvergiert die Folge $(B_i/{\lambda_2}^i, C_i/{\lambda_2}^i)_{i \geq 0}$ für $i \to \infty$ fast sicher gegen einen Zufallsvektor $(V, V/\sqrt{1-p})$, wobei die $\mathbb{R}_+$-wertige Zufallsvariable V mit positiver Wahrscheinlichkeit größer als 0 ist. Die zweiten Momente von B_i und C_i verhalten sich asymptotisch wie const.$\cdot {\lambda_2}^{2i}$.

Um die Kovarianzmatrix für X_i iterativ zu berechnen, können wir folgendes einfache Lemma verwenden:

Lemma B.1 *Seien N und M Zufallsvariablen mit Werten in einer beschränkten Teilmenge von $\mathbb{N}_0$ und sei S bedingt binomialverteilt zu (N, p) und R bedingt binomialverteilt zu (M, q). Außerdem seien R und S bedingt unabhängig bezüglich (N, M). Dann gelten folgende Gleichungen:*

$$\begin{aligned} \mathrm{Var}(S) &= p(1-p) \cdot \mathbb{E}N + p^2 \cdot \mathrm{Var}(N) \\ \mathrm{Cov}(S, R) &= p \cdot q \cdot \mathrm{Cov}(N, M) \end{aligned}$$

Beweis: Die erste Gleichung wird z. B. im Buch von Dinges und Rost (1982) auf Seite 253 bewiesen. Die zweite Gleichung folgt analog.

□

Nun sei C_i^B die Anzahl an Molekülen vom Typ C, die im Zyklus i an einem Molekül vom Typ B entstehen. A_i^A, A_i^C, B_i^C und D_i^C seien analog definiert. Wir kürzen im Folgenden „binomialverteilt“ mit bin(. , .) ab. Es gilt für $i \geq 1$:

$$\begin{aligned} \mathcal{L}(A_i^A | X_{i-1}) &= \mathrm{bin}(A_{i-1}, p) & \mathcal{L}(C_i^B | X_{i-1}) &= \mathrm{bin}(B_{i-1}, p) \\ \mathcal{L}(A_i^C | X_{i-1}) &= \mathrm{bin}(C_{i-1}, p) & \mathcal{L}(D_i^C | A_i^C) &= \mathrm{bin}(A_i^C, p) \\ \mathcal{L}(B_i^C | X_{i-1}, A_i^C) &= \mathrm{bin}(C_{i-1} - A_i^C, p) \end{aligned}$$

Mit Lemma B.1 erhalten wir daraus ein Gleichungssystem, mit dem wir die Kovarianzmatrix von X_i aus $\mathbb{E}X_{i-1}$ und der Kovarianzmatrix von X_{i-1} berechnen können. Wegen der Größe des Gleichungssystems empfiehlt sich die Verwendung eines Computeralgebrasystems. Es läßt sich dann u. a. leicht die folgende Gleichung herleiten:

$$\mathrm{Var}(A_i) = {\lambda_1}^2 \mathrm{Var}(A_{i-1}) + p(1-p)(\mathbb{E}C_{i-1} + \mathbb{E}A_{i-1}) + p^2 \mathrm{Var}(C_{i-1}) + 2p\lambda_1 \mathrm{Cov}(A_{i-1}, C_{i-1})$$

(Weitere Gleichungen entnehme man dem MAPLE-Worksheet `pcrcov.mws` unter `http://stoch.math.uni-frankfurt.de/Leute/metzler/dissprgs.html`.) Da $\mathbb{E}C_{i-1}$, $\mathbb{E}A_{i-1}$, $\mathrm{Var}(C_{i-1})$ und $\mathrm{Cov}(A_{i-1}, C_{i-1})$ durch const.$\cdot{\lambda_1}^{2i}$ beschränkt sind, folgt, daß sich $\mathrm{Var}(A_i)$ asymptotisch wie const.$\cdot{\lambda_1}^{2i}$ verhält. Aus dem Gleichungssystem läßt sich außerdem ablesen, daß sich die Varianz von D_i asymptotisch wie const.$\cdot{\lambda_2}^{2i}$ verhält. Dies ist nicht gerade erstaunlich, da der Typ D vom Typ C gespeist wird.

Der Spaltenvektor $(1, 1+p, 1+p, 0)^{\mathrm{transp}}$ ist ein rechter Eigenvektor von M zum Eigenwert λ_1. Demnach gilt $\mathbb{E}\left(X_i \cdot (1, 1+p, 1+p, 0)^{\mathrm{transp}}\right) = X_{i-1} \cdot M \cdot (1, 1+p, 1+p, 0)^{\mathrm{transp}} = \lambda_1 \cdot X_{i-1} \cdot (1, 1+p, 1+p, 0)^{\mathrm{transp}}$. Also ist die durch $W_i := (A_i + p^{-1}B_i + p^{-1}C_i)/\lambda_1^i$ definierte Folge $(W_i)_{i\geq 0}$ ein nichtnegatives Martingal (bzgl. der naheliegenden Filtration). Aus den obigen Betrachtungen zu den zweiten Momenten von A_i, B_i und C_i folgt, daß die zweiten Momente von W_i für $i \to \infty$ konvergieren. Aus Standardargumenten der Martingaltheorie (siehe Doob 1953, S.319) folgt, daß $(W_i)_i$ fast sicher gegen eine Zufallsvariable W mit den Eigenschaften $\mathbb{E}W = W_0$, $\mathrm{Var}(W) < \infty$ und $\Pr(0 < W < \infty) = 1$ konvergiert. Da $(p^{-1}B_i + p^{-1}C_i)/\lambda_1^i$ für $\mathrm{Var}(W) < \infty$ fast sicher gegen 0 konvergiert, folgt insbesondere:

$$\mathbb{E}\left(\lim_{i\to\infty} A_i/\lambda_1^i\right) = A_0 + p^{-1}B_0 + p^{-1}C_0$$

Verwendet man in dieser Argumentation weitere rechte Eigenvektoren von M, so erhält man die Martingale $\left((\sqrt{1-p}\cdot C_i + B_i)/\lambda_2^i\right)_i$, $\left(\frac{p}{1-p}B_i - D_i\right)_i$ und $\left((\sqrt{1-p}\cdot C_i - B_i)/\lambda_4^i\right)_i$. Die Konvergenzaussage läßt sich jedoch nur auf das erste übertragen, da für die letzten beiden die Folgen der zweiten Momente nicht beschränkt sind. Wir können aber folgern, daß die Supermartingale $\left((\frac{p}{1-p}B_i - D_i)/\lambda_2^i\right)_i$ und $\left((\sqrt{1-p}\cdot B_i - C_i)/\lambda_2^i\right)_i$ fast sicher gegen 0 konvergieren.

Die fast sichere Konvergenz der (Linearkombinationen von) skalierten Typhäufigkeiten gegen Zufallsvariablen kann man sich auch mit elementar-stochastischen Überlegungen plausibel machen: Innerhalb der ersten Zyklen, wenn nur wenige Moleküle vorhanden sind, spielt der Zufall noch eine gewisse Rolle. Wenn nach einigen Zyklen die Anzahlen an Molekülen der einzelnen Typen sehr groß geworden sind, wirken die Gesetze der großen Zahlen und alle Typen wachsen fast deterministisch gemäß ihren erwarteten Wachstumsraten.

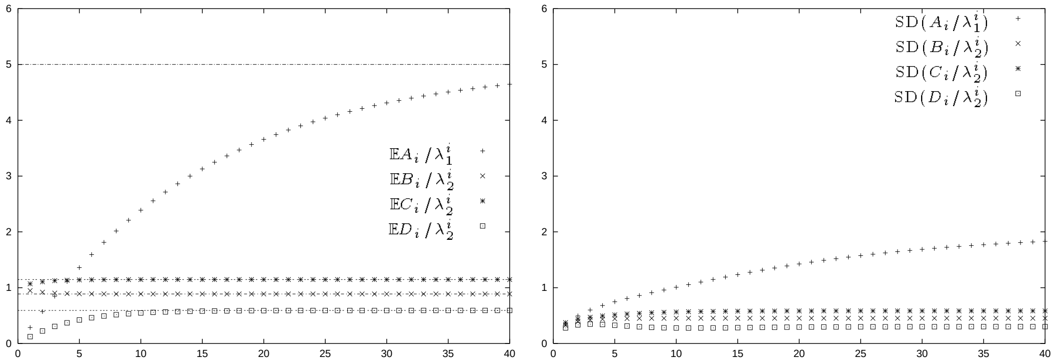
B.3 Erwartungswerte und Varianzen nach 40 Zyklen

Wir haben im vorangegangenen Abschnitt das asymptotische Verhalten von X_i untersucht. Insbesondere haben wir Approximationen für die erwartete Anzahl an Molekülen nach vielen

Zyklen hergeleitet. Wir fragen uns nun zunächst, ob sich das System nach 40 Zyklen schon in einem Zustand befindet, der durch die Asymptotik beschrieben werden kann.

Wir beziehen uns in diesem Abschnitt stets auf eine Familie von Fragmenten, die auf einen einzelnen Original-Doppelstrang zurückgehen. Wir nehmen also $X_0 = (0, 1, 1, 0)$ an.

$p = 0{,}4$:



$p = 0{,}7$:

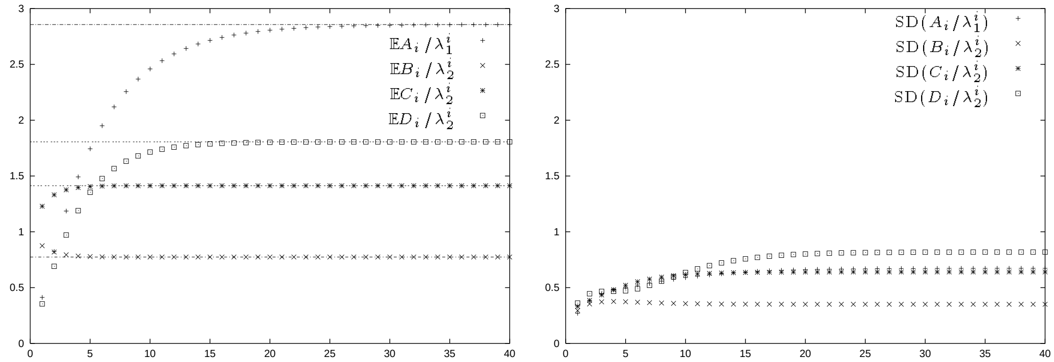


Abbildung B.4: Die Entwicklung des Erwartungswerts und der Standardabweichung (SD) der Häufigkeit jedes Typs, skaliert mit der jeweiligen asymptotischen Wachstumsrate, über die 40 Zyklen der PCR. Die beiden oberen Grafiken beziehen sich auf eine Primerbindungswahrscheinlichkeit von $p = 0{,}4$, die unteren beiden auf $p = 0{,}7$. Zu den Erwartungswerten von A_i/λ_1^i, B_i/λ_2^i, C_i/λ_2^i und D_i/λ_2^i sind auch deren asymptotische Werte $2/p$, $\frac{1}{2}(1+\sqrt{1-p})$, $\frac{1}{2}(1+\frac{1}{\sqrt{1-p}})$ und $\frac{1}{2}\frac{p}{1-p}(1+\sqrt{1-p})$ als Konstanten eingezeichnet.

Wie wir in Abbildung B.4 sehen, stimmen für $p = 0{,}7$ die erwarteten Typhäufigkeiten mit den approximativ erwarteten schon ab dem 25. Zyklus sehr gut überein, und an den Standardabweichungen der skalierten Typhäufigkeiten ändert sich auch nicht mehr viel. Bei geringeren Primerbindungswahrscheinlichkeiten ($p = 0{,}4$) ist der Erwartungswert der Anzahl an Molekülen vom Typ A nach 40 Zyklen noch einige Prozentpunkte von seiner Asymptote entfernt, befindet sich aber „in der richtigen Größenordnung“, was für die meisten praktischen Fragen ausreicht.

Wir betrachten nun die erwarteten Typhäufigkeiten nach 40 Zyklen in Abhängigkeit von der Primerbindungswahrscheinlichkeit p. Wie in Abbildung B.5 zu sehen ist, entstehen für

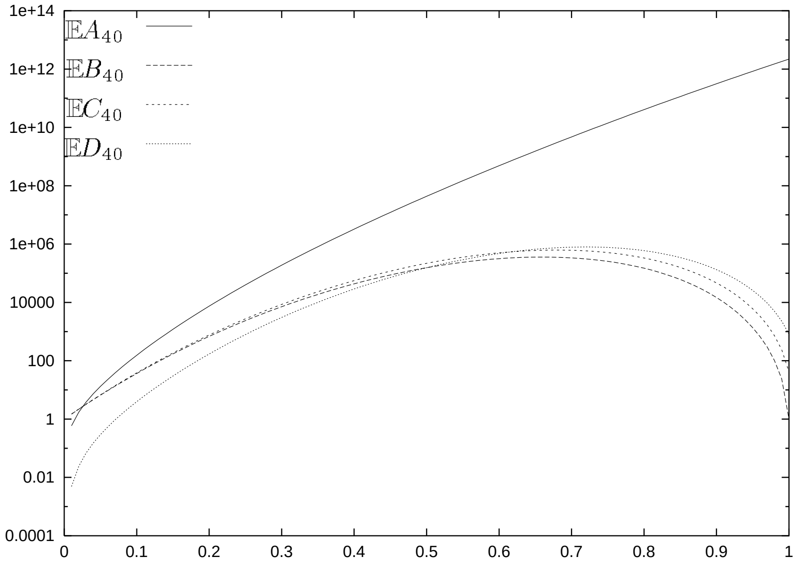


Abbildung B.5: Erwartete Häufigkeiten der einzelnen Typen nach 40 Zyklen, in Abhängigkeit von der Primer-Bindungs-Wahrscheinlichkeit.

$p \leq 1/4$ nur wenige zehntausend Fragmente pro Original-Doppelstrang, was wohl in der Regel zu keinem deutlichen Bandensignal führen wird. Für $p > 1/4$ gibt es in Erwartung deutlich mehr Fragmente vom Typ A als von allen anderen Typen. Welche Häufigkeitsverhältnisse wir zwischen den Typen B, C und D erwarten, hängt von p ab. Je größer p ist, desto mehr gewinnt D und verliert B im Vergleich zu C. Das ist leicht zu erklären: Zum Beispiel ist es ab $p > 1/2$ wahrscheinlicher, daß an beiden Inverted Repeats eines Typ-C-Moleküls je ein Primer anbindet, als daß nur an der zweiten ein Primer anbindet (siehe Abbildung B.3). Je größer p wird, desto wahrscheinlicher wird die Amplifikation eines Typ-D-Moleküls an einem gegebenen Typ-C-Molekül. Für große p wird damit aber die Amplifikation von Typ-B-Molekülen immer unwahrscheinlicher. Da Typ-C-Moleküle aber nur an Typ-B-Molekülen amplifiziert werden können, bedeutet das, daß für sehr große p auch die erwartete Anzahl an Typ-B-Molekülen und damit auch die erwartete Anzahl an Typ-D-Molekülen geringer ist als z. B. für $p = 0{,}7$.

Wieviele Fragmente einer bestimmten Länge nötig sind, um ein Bandensignal zu erzeugen, hängt stark von den jeweils vorhandenen Laborbedingungen ab und läßt sich wohl nicht pauschal sagen. Anhand der in B.5 gezeigten erwarteten Häufigkeiten würde man aber annehmen, daß Bandensignale, die auf Moleküle der Typen B, C und D zurückzuführen sind, in der Regel nur sehr schwach sind, zumindest im Vergleich zu den Typ-A-Bandensignalen.

Kann es sein, daß aufgrund zufälliger Schwankungen doch einmal ein deutliches Bandensignal von einem der Typen B, C oder D kommt? Kann es sogar sein, daß durch Zufall unter exakt denselben Versuchsbedingungen einmal ein Typ-D-Bandensignal etwas deutlicher ist und ein anderes Mal ein Typ-B-und-C-Bandensignal? (Man beachte, daß die Fragmente der Typen

B und C ohne die in Abbildung B.3 gepunktet dargestellten Stücke dieselbe Länge haben und daher – wenn überhaupt – ein gemeinsames Bandensignal hervorbringen.)

Aus Abbildung B.4 konnte man auch etwas über die Variabilität der Typhäufigkeiten erfahren. Um ein ausführlicheres Bild zu erhalten, habe ich Monte-Carlo-Simulationen für X_{40} durchgeführt. Wir wenden uns zunächst dem aus den Typen B, C und D bestehenden Teilsystem zu. Abbildung B.6 zeigt die Wertepaare $(B_{40} + C_{40} + D_{40}, B_{40})$, $(B_{40} + C_{40} + D_{40}, C_{40})$ und $(B_{40} + C_{40} + D_{40}, D_{40})$, die sich bei 1000 Monte-Carlo-Simulationen für $p = 0{,}6$ ergeben haben.

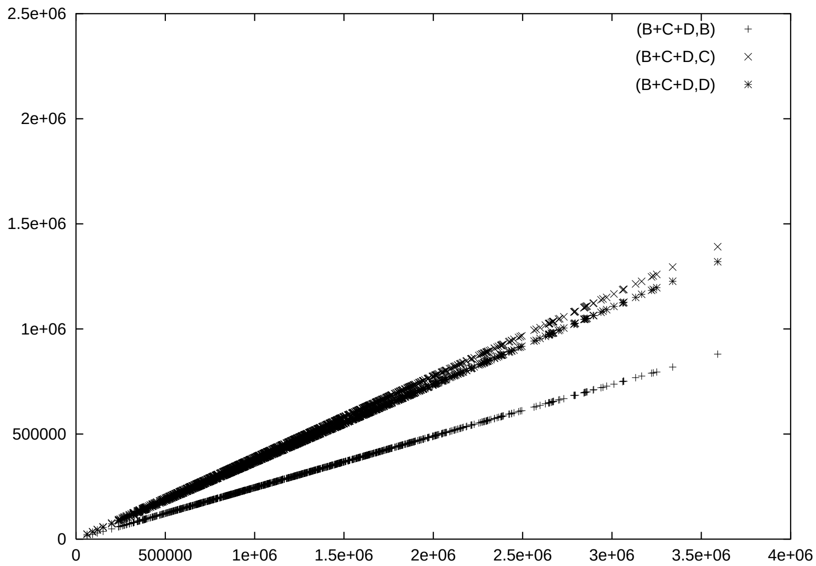


Abbildung B.6: Obgleich die absoluten Typhäufigkeiten Schwankungen unterliegen, ist das Verhältnis der Häufigkeiten nach 40 Zyklen ziemlich konstant (hier für $p = 0{,}6$).

Wie wir sehen, schwanken B_{40}, C_{40} und D_{40} von Versuch zu Versuch durchaus recht stark. Sie stehen jedoch immer ziemlich genau im selben Verhältnis zu $B_{40}+C_{40}+D_{40}$ und damit auch zueinander. (Dies gilt übrigens auch bei anderen Werten für p.) Diese Beobachtung zeigt, daß die Resultate aus Abschnitt B.2, nach denen $\left((\frac{p}{1-p}B_i - D_i)/\lambda_2^i\right)_i$ und $((\sqrt{1-p} \cdot B_i - C_i)/\lambda_2^i)_i$ fast sicher gegen 0 konvergieren, bereits nach 40 Schritten eine Bedeutung haben.

Wir können das gesamte in Abbildung B.3 dargestellte System als aus zwei zyklischen Teilsystemen bestehend auffassen: Es gibt Typ A, der sich selbst begünstigt, und es gibt die Typen B und C, die sich gegenseitig begünstigen und dabei noch Typ D fördern. Die einzige Abhängigkeit zwischen den beiden Teilsystemen ist dadurch gegeben, daß Typ C auch Typ A speist. Wie stark wirkt sich das auf die Abhängigkeiten zwischen A_{40} und $B_{40} + C_{40} + D_{40}$ aus? Wir betrachten dazu Abbildung B.7, in der das Paar $(A_{40}, B_{40} + C_{40} + D_{40})$ für jeweils 1000 Monte-Carlo-Simulationen mit verschiedenen Primerbindungswahrscheinlichkeiten eingezeichnet ist.

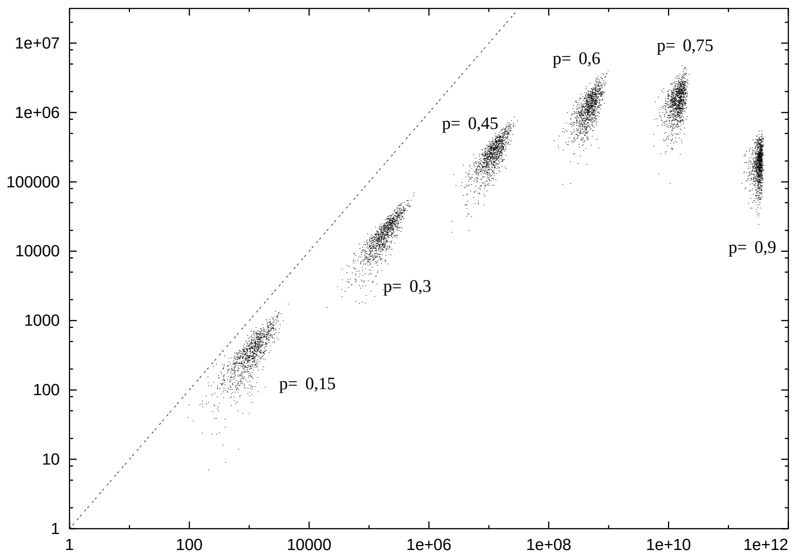


Abbildung B.7: Die Häufigkeit von Typ A, aufgetragen gegen die Gesamthäufigkeit aller anderen Typen nach 40 Zyklen bei jeweils 1000 Monte-Carlo-Simulationen mit verschiedenen Primer-Bindungs-Wahrscheinlichkeiten.

Wie man sieht, sind A_{40} und $(A_{40}, B_{40}+C_{40}+D_{40})$ (zumindest für $p < 0{,}9$) positiv korreliert. Die Abhängigkeit ist jedoch wesentlich schwächer als innerhalb der Typen B, C und D. Wie man außerdem klar erkennen kann, besteht (zumindest für $p \geq 0{,}3$) keine ernstzunehmende Gefahr, daß A einmal nicht der nach 40 Zyklen bei weitem dominierende Typ sein könnte.

Alles, was in diesem Abschnitt über Variabilität gesagt wurde, bezieht sich auf ein Szenario, bei dem die PCR mit nur einem Original-DNA-Doppelstrang gestartet wird. Wenn man mit mehr Molekülen startet, ist die Variabilität natürlich geringer. Startet man etwa mit 1000 Doppelsträngen, so kann man den Zufall wohl völlig vernachlässigen.

Man kann sagen, daß die Fragmente A und B innerhalb eines Zyklus in gewissem Sinne um die Amplifikation an Molekülen vom Typ C in Konkurrenz stehen. Wie wir gesehen haben, sind nach 40 Zyklen alle Typhäufigkeiten positiv korreliert und in diesem Sinne nicht als Konkurrenten zu bezeichnen. Die Ergebnisse legen den Schluß nahe, daß sich aus der Situation, daß auf einem DNA-Strang zwei Inverted Repeats auf eine Primerkopie folgen, keine stochastische Dynamik ergibt, die zu zufälligen „Nichtreproduzierbarkeiten“ führen kann, sofern man ausschließlich deutliche Bandensignale beachtet.

B.4 Fazit

Wir gehen davon aus, daß die RAPD-PCR ein Verfahren ist, welches für einen DNA-Strang und eine Primersequenz genau dann ein Bandensignal für die Fragmentlänge k liefert, wenn es auf der DNA eine Primerkopie gibt, auf die im Abstand von genau k Basen ein Primerkomplement folgt, so daß zwischen den beiden weder eine weitere Primerkopie noch ein weiteres Primerkomplement liegt. Das Ganze gelte nur für solche k, die kleiner als der Amplifikationsbereich sind. Für letzteren werden wir in der Regel 3000 Basenpaare annehmen.

Anhang C

Muster und Banden auf einem DNA-Strang

Hier betrachten wir die Problematik der Überlappung von Mustern in zufälligen Folgen und Lösungsansätze mit der Chen-Stein-Methode und der Poisson Clumping Heuristik.

Wenn ein Vektor $X = (X_1, \dots, X_m)$ von $\{0,1\}$-wertigen Zufallsvariablen und eine Folge $(\Gamma_1, \dots, \Gamma_m)$ von Teilmengen von $\Gamma := \{1, \dots, m\}$ gegeben ist, so daß (für $i \leq m$) X_i unabhängig von $(X_j)_{j \in \Gamma \setminus \Gamma_i}$ ist, so läßt sich der Totalvariationsabstand $d_{TV}(\mathcal{L}(X), \mathrm{Po}(\mathbb{E}X))$ zwischen der Verteilung von X und der eines Poisson-Prozesses zum Intensitätsvektor $\mathbb{E}(X) = (\mathbb{E}X_1, \dots, \mathbb{E}X_m)$ durch $\sum_{i \leq m, j \in \Gamma_i} (\mathbb{E}X_i \mathbb{E}X_j + \mathbb{E}(X_i X_j))$ abschätzen. Theorem 10.A aus Barbour *et al.* (1992) liefert sogar eine etwas allgemeinere Abschätzung, die auch dann noch sinnvoll einsetzbar ist, wenn es (für $i \leq n$) schwache, kontrollierbare Abhängigkeiten zwischen X_i und $(X_j)_{j \in \Gamma \setminus \Gamma_i}$ gibt (siehe auch Arratia *et al.* (1996)). Das Finden geeigneter Γ_i und das Anwenden dieses Theorems wird oft als „die Chen-Stein-Methode für Poisson-Prozesse“ bezeichnet, da es sich bei dem Theorem um eine Verallgemeinerung von Resultaten von Chen (1975) handelt, die er mit Techniken erzielt hat, die auf Stein (1971) zurückgehen.

C.1 Poisson-Approximation für Muster-Konfigurationen auf der DNA

Von einem DNA-Strang sei nur seine Länge bekannt. Was kann man über Anzahl und Länge der Banden sagen, die man erhält, wenn man eine RAPD-PCR mit dem Strang durchführt? Als Grundlage für die Beantwortung dieser und ähnlicher Fragen betrachten wir nun eine Menge $\mathcal{M} = \{M_1, \dots, M_k\}$ von Mustern der Länge l und die Verteilung der Positionen, an denen die Muster auf einem zufälligen DNA-Strang $D = (D_1, \dots, D_{n_{\mathrm{dna}}})$ vorkommen. Wir betrachten also die zufälligen Mengen $\mathcal{V}_{M_i}(D) := \{s : (D_s, D_{s+1}, \dots, D_{s+l-1}) = M_i\}$. Dabei seien die D_s unabhängig identisch verteilt mit $\Pr(D_s = \mathtt{A}) =: P(\mathtt{A}), \dots, \Pr(D_s = \mathtt{T}) =: P(\mathtt{T})$.

Wie sieht die gemeinsame Verteilung von $\mathcal{V}_{M_1}(D), \dots, \mathcal{V}_{M_k}(D)$ aus? Sei X_{sj} die Indikatorfunktion für $\{s \in \mathcal{V}_{M_j}(D)\}$. Mit $M_j =: (m_{j1}, \ldots, m_{jl})$ gilt $\Pr(s \in \mathcal{V}_{M_j}(D)) = \mathbb{E}X_{sj} = P(m_{j1}) \cdots P(m_{jl})$ für alle $s \in \{1, \ldots, n_{\text{dna}} - l + 1\}$ und alle $j \in \{1, \ldots, k\}$. Es treten aber viele stochastische Abhängigkeiten auf. So gelten z. B. für $M_1 = \mathtt{AA}$ und $M_2 = \mathtt{GG}$ die Gleichungen $\mathbb{E}(X_{s,1} | X_{s-1,1} = 1) = P(\mathtt{A})$ und $\mathbb{E}(X_{s,1} | X_{s-1,2} = 1) = \mathbb{E}(X_{s,1} | X_{s,2} = 1) = 0$.

Kann man die Abhängigkeiten zwischen den Komponenten von X vernachlässigen? Wir werden dieser Frage nachgehen, indem wir den Totalvariationsabstand der Verteilung von $X = (X_{sj})_{sj}$ zu einem Poisson-Prozeß $\widetilde{X} = (\widetilde{X}_{sj})_{sj}$ mit $\mathcal{L}(\widetilde{X}) = \text{Po}((\mathbb{E}X_{sj})_{sj})$ abschätzen.

Wir verwenden die Chen-Stein-Methode für Poisson-Prozesse. Dazu sei $\Gamma := \{1, \ldots, n_{\text{dna}} - l + 1\} \times \{1, \ldots, k\}$ und für $(s, j) \in \Gamma$ sei $\Gamma_{(s,j)} := \{(s', j') \in \Gamma \ : \ s' \in \{s - l + 1, \ldots, s + l - 1\}\}$. Dann ist für jedes $\alpha \in \Gamma$ die Zufallsvariable X_α von der Familie von Zufallsvariablen $(X_\gamma)_{\gamma \in \Gamma \setminus \Gamma_\alpha}$ stochastisch unabhängig.

Aus Theorem 10.A in Barbour *et al.* (1992) folgt also:

$$d_{TV}(\mathcal{L}(X), \mathcal{L}(\widetilde{X})) \leq \sum_{\substack{\alpha,\beta \in \Gamma \\ \beta \in \Gamma_\alpha}} \mathbb{E}X_\alpha \cdot \mathbb{E}X_\beta + \sum_{\substack{\alpha,\beta \in \Gamma \\ \alpha \neq \beta \in \Gamma_\alpha}} \mathbb{E}(X_\alpha \cdot X_\beta)$$

Die erste Summe ist offensichtlich $2 \cdot (n_{\text{dna}} - l + 1) \cdot (l - 1) \cdot \left(\sum_{j=1}^k \prod_{h=1}^l P(m_{jh})\right)^2$.

Zur Berechnung der zweiten Summe sei $\mathcal{U}_{AB}$ für jedes Paar von Mustern $A = (a_1, \ldots, a_l)$, $B = (b_1, \ldots, b_l)$ die Menge der möglichen Überlappungsweiten $< l$ der Muster:

$$\mathcal{U}_{AB} := \{x \in \{1, \ldots, l - 1\} \ : \ a_{l-x+1} = b_1, a_{l-x+2} = b_2, \ldots, a_l = b_x\}$$

Für $\mathcal{U}_{M_i,M_j}$ schreiben wir auch $\mathcal{U}_{ij}$. Offensichtlich gilt:

$$\begin{aligned}\sum_{\substack{\alpha,\beta \in \Gamma \\ \alpha \neq \beta \in \Gamma_\alpha}} \mathbb{E}(X_\alpha X_\beta) \quad &\leq \quad (n_{\text{dna}} - l + 1) \sum_{(i,j) \in \{1,\ldots,k\}^2} P(m_{i1}) \cdots P(m_{il}) \cdot \\ &\cdot \left(\sum_{x \in \mathcal{U}_{ij}} P(m_{j,x+1}) \cdots P(m_{jl}) + \sum_{y \in \mathcal{U}_{ji}} P(m_{j,1}) \cdots P(m_{j,l-y}) \right)\end{aligned}$$

(Das $\leq$ ergibt sich dabei ausschließlich daher, daß bei Sites, die am Anfang oder Ende von D liegen, einige Überlappungsmöglichkeiten wegfallen.)

Beispiele Es sei $P(\mathtt{A}) = P(\mathtt{G}) = P(\mathtt{C}) = P(\mathtt{T}) = 1/4$, $\mathcal{M}$ bestehe nur aus dem Muster $\mathtt{AAAAAAAG}$. Es sei $\lambda = (n_{\text{dna}} - 7)(\frac{1}{4})^8$ die erwartete Anzahl an Mustern auf D. Da es keine Überlappungsmöglichkeiten gibt, gilt:

$$d_{TV}(\mathcal{L}(X), \mathcal{L}(\widetilde{X})) \leq \sum_{\substack{\alpha,\beta \in \Gamma \\ \beta \in \Gamma_\alpha}} \mathbb{E}X_\alpha \cdot \mathbb{E}X_\beta \leq 2 \cdot (n_{\text{dna}} - 7) \cdot 7 \cdot \left(\frac{1}{4}\right)^{16} \approx 0{,}0002 \cdot \lambda$$

Wählt man hingegen als Menge von Mustern diejenige, die nur aus $M = \mathtt{AGCTAGCT}$ besteht, so gibt es eine Überlappungsmöglichkeit: Es gilt $\mathcal{U}_{MM} = \{4\}$. Wir erhalten dann

$$\sum_{\substack{\alpha,\beta \in \Gamma \\ \alpha \neq \beta \in \Gamma_\alpha}} \mathbb{E}(X_\alpha X_\beta) \leq (n_{\text{dna}} - 7) \cdot \left(\frac{1}{4}\right)^8 \cdot 2 \cdot \left(\frac{1}{4}\right)^4 \approx \lambda \cdot 0{,}0078$$

und damit

$$d_{TV}(\mathcal{L}(X), \mathcal{L}(\widetilde{X})) \lesssim 0{,}008 \cdot \lambda.$$

Das Beispiel zeigt also, daß wir für stark überlappungsfähige Muster mit diesem Ansatz keine besonders guten Resultate bekommen, falls λ nicht gerade sehr klein ist. Für solche Szenarien ist der Ansatz, der im nächsten Abschnitt diskutiert wird, offensichtlich eher angemessen.

C.2 Poisson-Approximationen für Klumpen von Mustern

Das Problem beim letzten Beispiel des vorangegangenen Abschnitts war offensichtlich die Tatsache, daß ein Muster vom Typ M, welches in D auftritt, häufig noch ein um 4 Sites verschobenes Muster vom selben Typ nach sich zieht. Es tauchen also häufig für Poisson-Prozesse „untypische“ Paare von Ereignissen auf. Wir verfolgen daher nun den Ansatz, von mehreren Mustern, die in überlappender Weise auf D vorkommen, jeweils nur das erste zu beachten und die Konfigurationen, die wir so erhalten, durch Poisson-Prozesse zu approximieren. Man kann es auch noch etwas prägnanter formulieren: Da Muster in Klumpen auftreten, approximieren wir nicht die Konfiguration der Muster sondern die der Klumpen durch einen Poisson-Prozeß. Hierzu ein Beispiel: Seien A = `TCAT` und B = `ATGT` die uns interessierenden Muster. Dann ist eine Kopie der Sequenz `TCATCATGT` ein Klumpen. Die ersten vier Basen bilden eine Kopie von A. Diese überlappt mit einer mit der vierten Base beginnenden Kopie von A, welche mit einer Kopie von B überlappt, die mit der sechsten Base beginnt. Die Sequenz `TCATTCAT` ist hingegen kein Klumpen, sondern besteht aus zwei Klumpen, da die beiden Kopien von A nicht überlappen.

Es sei daher in diesem Abschnitt $Y_{(s,j)}$ für $(s, j) \in \Gamma = \{1, \ldots, n_{\text{dna}}-l+1\} \times \{1, \ldots, k\}$ die Indikatorvariable des Ereignisses $\{s \in \mathcal{V}_{M_j}(D)$ und $\{s-l+1, \ldots, s-1\} \cap (\mathcal{V}_{M_1}(D) \cup \ldots \cup \mathcal{V}_{M_k}(D)) = \emptyset\}$; in Worten: an Site s beginnt M_j und wird von links von keinem Muster überlappt. Wir registrieren also nur, an welchem Site der Klumpen beginnt und mit welchem Muster er anfängt. Für $A, B \in \{M_1, \ldots, M_k\}$ und $x \in \{1, \ldots, l-1\}$ sei $Z_{AB,x}$ die Indikatorvariable für das Ereignis $\{x \in \mathcal{U}_{AB}\} \cap \{\nexists\ (C, y) \in \{M_1, \ldots, M_k\} \times \{1, \ldots, x-1\}\ :\ y \in \mathcal{U}_{AC}, (x-y) \in \mathcal{U}_{CB}\}$. Damit gilt für $l < s < n_{\text{dna}} - l$:

$$\mathbb{E}Y_{(s,j)} = P(m_{j1}) \cdots P(m_{jl}) \cdot \left(1 - \sum_{i=1}^{k} \sum_{x=1}^{l-1} Z_{M_i M_j,x} \cdot P(m_{i1}) \cdots P(m_{i,l-x})\right)$$

Die Indikatorvariablen $Z_{AB,x}$ sind gerade so definiert, daß wir in der letzten Summe nur Wahrscheinlichkeiten disjunkter Ereignisse aufaddieren.

Wir ordnen jedem Klumpen eine Folge auf $\{M_1, \ldots, M_k\} \cup \{\mathtt{X}\}$ zu, wobei `X` lediglich ein formales Symbol ist. Die Folge heißt *Profil* des Klumpens. Ihr i-tes Glied ist das i-te Muster, das in dem Klumpen auftritt, oder `X`, falls es weniger als i Muster in dem Klumpen gibt. Ist {`TCAT`, `ATGT`} die Menge der Muster, so hat der Klumpen `TCATCATGT` das Profil (`TCAT`, `TCAT`, `ATGT`, `X`, `X`, ...). Gegeben, daß auf D an der Stelle s ein Klumpen beginnt, so ist sein Profil $(K_i)_{i\in\mathbb{N}}$ offensichtlich eine Markoff-Kette mit absorbierendem Zustand $\mathtt{X}$. Die Übergangswahrscheinlichkeiten sind gegeben durch:

$$\begin{aligned}
\Pr(K_n = M_i | K_{n-1} = M_j) &= \sum_{x=1}^{l} Z_{M_i M_j, x} \cdot \Pr(m_{j,x+1}) \cdots \Pr(m_{jl}) \\
\Pr(K_n = \mathtt{X} | K_{n-1} = M_j) &= 1 - \sum_{i=1}^{k} \Pr(K_n = M_i | K_{n-1} = M_j)
\end{aligned}$$

Damit läßt sich leicht die Wahrscheinlichkeit berechnen, daß ein Klumpen, der mit einem Muster M_i beginnt, ein Muster M_j enthält.

Ist gegeben, daß sowohl bei r als auch bei $s > r$ je ein Klumpen beginnt, so zieht dies für den ersten nach sich, daß er höchstens $s - r$ Basen enthält. Die Auswirkungen dieser Bedingung auf die Wahrscheinlichkeitsverteilung seines Profils sind aber offensichtlich vernachlässigbar, wenn $s - r$ hinreichend groß ist. Die Wahrscheinlichkeit, daß ein Klumpen mehr als n Basen enthält, fällt exponentiell in n, und die Abstände zwischen zwei Klumpen sind typischerweise sehr viel größer als die erwartete Länge eines Klumpen.

Wir wollen nun $Y = (Y_\alpha)_{\alpha\in\Gamma}$ durch eine Familie unabhängiger Poisson-verteilter Zufallsvariablen $\widetilde{Y} = (\widetilde{Y}_\alpha)_{\alpha\in\Gamma}$ mit $\mathbb{E}Y_\alpha = \mathbb{E}\widetilde{Y}_\alpha$ approximieren. Man beachte, daß anders, als es bei $(X_\alpha)_{\alpha\in\Gamma}$ der Fall ist, $Y_{(s,j)}$ und $Y_{(r,k)}$ i. a. auch dann stochastisch abhängig sind, wenn $l \le |s - r| \le 2 \cdot (l - 1)$ gilt. Für die Anwendung der Chen-Stein-Methode setzen wir daher

$$\Gamma^Y_{(s,j)} := \{s - 2(l - 1), \ldots, s + 2(l - 1)\} \times \{1, \ldots, k\}.$$

Jedes Y_α ist von $(Y_\gamma)_{\gamma\in\Gamma\setminus\Gamma^Y_\alpha}$ unabhängig. Aus Theorem 10.A in Barbour *et al.* (1992) folgt also:

$$d_{TV}(\mathcal{L}(Y), \mathcal{L}(\widetilde{Y})) \le \sum_{\substack{\alpha,\beta\in\Gamma \\ \beta\in\Gamma^Y_\alpha}} \mathbb{E}Y_\alpha \cdot \mathbb{E}Y_\beta + \sum_{\substack{\alpha,\beta\in\Gamma \\ \alpha\ne\beta\in\Gamma^Y_\alpha}} \mathbb{E}(Y_\alpha \cdot Y_\beta)$$

Für die erste Summe erhalten wir:

$$\begin{aligned}
&\sum_{\substack{\alpha,\beta\in\Gamma \\ \beta\in\Gamma^Y_\alpha}} \mathbb{E}Y_\alpha \cdot \mathbb{E}Y_\beta \\
&\quad\approx\ (n_{\mathrm{dna}} - l + 1) \cdot (4(l - 1) + 1) \cdot \\
&\qquad \cdot \left[\sum_{j=1}^{k} P(m_{j1}) \cdots P(m_{jl}) \cdot \left(1 - \sum_{i=1}^{k} \sum_{x=1}^{l-1} Z_{M_i M_j, x} \cdot P(m_{j1}) \cdots P(m_{j,l-x})\right)\right]^2
\end{aligned}$$

Die Ungenauigkeit entsteht dabei dadurch, daß es einige wenige Sites am Anfang und Ende von D gibt, bei denen Randeffekte zu berücksichtigen sind. Wir gehen aber davon aus, daß n_{dna} sehr groß ist, und daß die Randeffekte folglich zu vernachlässigen sind.

Für $(s, i) \ne (r, j)$ mit $|s - r| \le l - 1$ gilt $\mathbb{E}(Y_{(s,i)} Y_{(r,j)}) = 0$. Für die Berechnung von $\sum_{\Gamma\ni\alpha\ne\beta\in\Gamma^Z_\alpha} \mathbb{E}(Y_\alpha Y_\beta)$ braucht uns also nur der Fall $|s - r| \in \{l, \ldots, 2(l - 1)\}$ zu interessieren.

Sei o. B. d. A. $r \in \{s+l, \dots, s+2(l-1)\}$. Dann gilt (falls die Sites s und r nicht zu nahe am Anfang oder Ende von D liegen):

$$\begin{aligned}\mathbb{E}(Y_{(s,j)} \cdot Y_{(r,i)}) &= \mathbb{E}Y_{(s,j)} \cdot P(m_{i1}) \cdots P(m_{il}) \cdot \\ &\cdot \Bigg(1 - \sum_{h=1}^{k} \Big(\sum_{x \in \mathcal{U}_{jh}} Z_{M_h M_i, s+2l-r-x} \cdot P(m_{h,x+1}) \cdots P(m_{h,x+r-s-l}) \\ &\qquad + \sum_{y=s+2l-r}^{l-1} Z_{M_h M_i, y} \cdot P(m_{h,1}) \cdots P(m_{h,l-y}) \Big) \Bigg)\end{aligned}$$

Ist $t = l$, so habe das „leere Produkt" $P(m_{h,x+1}) \cdots P(m_{h,x+t-l})$ in dieser Gleichung den Wert 1. Es folgt also:

$$\begin{aligned}\sum_{\substack{\alpha,\beta \in \Gamma \\ \alpha \neq \beta \in \Gamma_\alpha^Y}} \mathbb{E}(Y_\alpha \cdot Y_\beta) &\approx 2 \cdot (n_{\text{dna}} - l + 1) \sum_{j=1}^{k} \mathbb{E}Y_{(s,j)} \cdot \sum_{i=1}^{k} \sum_{t=l}^{2(l-1)} P(m_{i1}) \cdots P(m_{il}) \cdot \\ &\cdot \Bigg(1 - \sum_{h=1}^{k} \Big(\sum_{x \in \mathcal{U}_{jh}} Z_{M_h M_i, 2l-t-x} \cdot P(m_{h,x+1}) \cdots P(m_{h,x+t-l}) \\ &\qquad + \sum_{y=2l-t}^{l-1} Z_{M_h M_i, y} \cdot P(m_{h,1}) \cdots P(m_{h,l-y}) \Big) \Bigg)\end{aligned}$$

(Dabei ist s eine beliebige ganze Zahl zwischen $4l$ und $n_{\text{dna}} - 4l$.) Wir müssen hier wieder $\approx$ statt $=$ schreiben, da wir Randeffekte vernachlässigen.

Beispiele Wir betrachten nun wieder die beiden Beispiele aus Abschnitt C.1. Für `AAAAAAAG` erhalten wir

$$d_{TV}(\mathcal{L}(Y), \mathcal{L}(\widetilde{Y})) \lesssim (n_{\text{dna}} - 7) \cdot 29 \cdot \left(\frac{1}{4}\right)^{16} + 2 \cdot (n_{\text{dna}} - 7) \cdot \left(\frac{1}{4}\right)^{8} \cdot 7 \cdot \left(\frac{1}{4}\right)^{8} \approx 0{,}00066 \cdot \lambda.$$

Wir erhalten für dieses Muster also ein schlechteres Ergebnis als in Abschnitt C.1. Dies liegt nicht daran, daß der Totalvariationsabstand zwischen $\mathcal{L}(Y)$ und $\mathcal{L}(\widetilde{Y})$ größer wäre als zwischen $\mathcal{L}(X)$ und $\mathcal{L}(\widetilde{X})$. Da das Muster keine Überlappungsmöglichkeiten hat, gilt nämlich $X = Y$ und $\mathcal{L}(\widetilde{X}) = \mathcal{L}(\widetilde{Y})$. Wir haben den Totalvariationsabstand einfach gröber abgeschätzt, da Γ_α^Y größer ist als Γ_α. Dies bringt im Falle von nicht überlappenden Mustern nur Nachteile.

Wie sieht es also aus, wenn wir das überlappende Muster $M =$ `AGCTAGCT` betrachten? Es gilt dann $Z_{MM,x} = 0$ für $x \neq 4$ und $Z_{MM,4} = 1$. Daraus folgt:

$$\begin{aligned}d_{TV}(\mathcal{L}(Y), \mathcal{L}(\widetilde{Y})) &\lesssim (n_{\text{dna}} - 7) \cdot 29 \cdot \left[\left(\frac{1}{4}\right)^{8} \cdot \left(1 - \left(\frac{1}{4}\right)^{4}\right)\right]^2 \\ &\quad + 2(n_{\text{dna}} - 7) \cdot \left(\frac{1}{4}\right)^{8} \cdot \left(1 - \left(\frac{1}{4}\right)^{4}\right) \cdot\end{aligned}$$

$$
\begin{aligned}
&\cdot \left[3\cdot\left(\frac{1}{4}\right)^8 + 3\cdot\left(\frac{1}{4}\right)^8\cdot\left(1-\left(\frac{1}{4}\right)^4\right)\right]\\
\approx\ & 0{,}00062\cdot\lambda
\end{aligned}
$$

Wir erhalten also ein deutlich besseres Resultat als mit dem Ansatz in Abschnitt C.1. Wenn man sich allerdings für die Stellen auf der DNA interessiert, wo Muster beginnen, so muß man hier je nach Szenario noch weiter untersuchen, mit welcher Wahrscheinlichkeit welche Muster in Klumpen auftreten, die mit einem bestimmten anderen Muster beginnen, und ob die stochastische Abhängigkeit zwischen Y und den Profilen der Klumpen für die jeweilige Fragestellung vernachlässigbar ist.

C.3 Abstände zwischen Mustern auf der DNA

Als erstem Anwendungsbeispiel für die in den Abschnitten C.1 und C.2 dargestellten Approximationen von X durch $\widetilde{X}$ bzw. von Y durch $\widetilde{Y}$ wenden wir uns nun einem klassischen Problem aus dem Bereich der zufälligen Sequenz-Muster zu und vergleichen in diesem Kontext die Poisson-(Clumping-)Approximation mit einer klassischen Approximation. Chen-Stein-Abschätzungen spielen hierbei keine explizite Rolle. Wir gehen davon aus, daß die Totalvariationsabstände hinreichend klein sind.

Wir bedingen in diesem Abschnitt darauf, daß der DNA-Strang D mit einem Muster $A = (a_1, \ldots, a_l)$ beginnt. Nach wievielen Basen tritt zum ersten Mal das Muster $B = (b_1, \ldots, b_l) \neq A$ auf? Ist A ein RAPD-Primer und B das dazugehörige Komplement, so könnten wir auch fragen: „Wie lang ist die kürzeste Bande, an der eine gegebene Kopie von A beteiligt ist?“ Wir können diese Frage auch als Vorüberlegung für die Frage nach der Verteilung der Länge von *sichtbaren* Banden auffassen.

Elegante Methoden zur Berechnung des Erwartungswertes von $N_{AB} := \min\{s \in \mathbb{N} : D_{s-l+1} = b_1, \ldots, D_s = b_l\}$ findet man bei S.-Y. R. Li (1980). Die erzeugende Funktion für die Verteilung von N_{AB} haben H. U. Gerber und S.-Y. R. Li (1981) hergeleitet. Weitere Resultate zu überlappenden Mustern in zufälligen Folgen findet man bei L. J. Guibas und A. M. Odlyzko (1981).(Die Untersuchungen in diesen drei Arbeiten beziehen sich nicht auf molekulargenetische Anwendungen, insbesondere wird nicht vorausgesetzt, daß die Glieder der Zufallsfolge genau vier Zustände haben. Auch die nun folgenden Untersuchungen sind direkt auf allgemeinere Szenarien übertragbar.)

Wenn wir die Verteilung der Muster- oder der Klumpen-Konfiguration auf D in der in C.1 bzw. C.2 dargestellten Weise approximieren, so erhalten wir auch Approximationen für die Verteilung von N_{AB}. Wir werden diese Approximationen mit asymptotischen Betrachtungen zu $p_n := \Pr(N_{AB} = n)$ vergleichen.

Die Verteilung von N_{AB} hängt davon ab, ob und wie weit Kopien des Musters B untereinander und mit A überlappen können. Dazu ein zunächst etwas paradox erscheinendes Beispiel:

Beispiel Wenn man eine Münze solange wirft, bis das Muster `Kopf,Kopf,Zahl,Zahl` auftritt, so beträgt der Erwartungswert für die benötigte Anzahl an Würfen 16. Ist man hingegen auf das Muster `Kopf,Zahl,Kopf,Zahl` aus, so muß man die Münze in Erwartung 20 mal werfen. Eine Formel zur Berechnung solcher Erwartungswerte wurde von S.-Y. R. Li (1980) besonders elegant bewiesen.

Wie kann man sich die relativ große Differenz zwischen den beiden Erwartungswerten erklären? Die Wahrscheinlichkeit, daß das Muster sofort bei den ersten vier Würfen entsteht, ist natürlich für beide Muster gleich. Wenn man die Münze schon $n\times$ geworfen hat (für $n \geq 5$), und das betreffende Muster bisher nicht aufgetreten ist, so ist die Wahrscheinlichkeit, daß das Muster im nächsten Wurf vervollständigt wird, im Falle des zweiten Musters kleiner. Die Bedingung, daß das Muster bisher nicht aufgetreten ist, bedeutet nämlich insbesondere, daß die Würfe $n-4$, $n-3$, $n-2$ und $n-1$ zusammen *nicht* das Muster bilden. Diese Bedingung verkleinert die Wahrscheinlichkeit dafür, daß der Wurf $n-2$ `Kopf` und der Wurf $n-1$ `Zahl` war, was jedoch eine notwendige Bedingung dafür ist, daß das zweite Muster im $(n+1)$-ten Wurf vervollständigt werden kann.

Wenn man jeweils die letzten vier Würfe der Münzwurffolge betrachtet, so erhält man eine Markoff-Kette auf $\{\texttt{Kopf,Zahl}\}^4$, und es bietet sich folgende Interpretation an: So wie sich die Muster der Länge 4, die mit dem Muster `Kopf,Zahl,Kopf,Zahl` überlappen können, gegenseitig begünstigen, begünstigen sich auch die Muster, die keine Überlappungsmöglichkeiten mit diesem Muster haben: Wenn wir darauf bedingen, daß wir gerade nicht im Zustand `Kopf,Zahl,Kopf,Zahl` sind, vergrößern wir die Wahrscheinlichkeit, daß wir auch zwei Sites später nicht im Zustand `Kopf,Zahl,Kopf,Zahl` sind.

Eine andere Betrachtungsweise ist die folgende: Wenn das Muster `Kopf,Zahl,Kopf,Zahl` in einer Münzwurffolge vollendet wird, so ist die Wahrscheinlichkeit, daß es zwei Würfe später wieder vollendet wird, relativ hoch. Wenn hingegen das Muster `Kopf,Kopf, Zahl,Zahl` auftritt, so kann dieses Muster frühestens vier Würfe später wieder vervollständigt werden. Der Erwartungswert dafür, wie oft das Muster in einer Folge von n Münzwürfen vollendet wird, ist jedoch für beide Muster $(n-3)/16$ (wie für alle Muster der Länge 4). Daher müssen die Klumpen, die aus einer Kopie oder mehreren überlappenden Kopien von `Kopf,Zahl,Kopf,Zahl` bestehen, in Erwartung seltener vorkommen, als Kopien von `Kopf,Kopf,Zahl,Zahl`.

C.3.1 Approximation von N_{AB} mit Hilfe von $\widetilde{X}$ und $\widetilde{Y}$

Wir gehen in diesem Abschnitt davon aus, daß wir uns in einem Szenario befinden, bei dem $d_{TV}(\mathcal{L}X, \mathcal{L}\widetilde{X})$ und $d_{TV}(\mathcal{L}Y, \mathcal{L}\widetilde{Y})$ hinreichend klein sind.

Wir wenden nun zunächst die Approximation von X durch $\widetilde{X}$ an. Dabei gehen wir von der Mustermenge $\{M_1, M_2\} = \{A, B\}$ aus. Wir erhalten als Analogon zu N_{AB} die Zufallsvariable $N_{AB}^{\widetilde{X}} := \min\{s \in \mathbb{N} \ : \ \widetilde{X}_{(s,2)} > 0\}$. Wir gehen davon aus, daß D hinreichend lang ist, so daß der Fall $\forall_s : \widetilde{X}_{(s,2)} = 0$ vernachlässigbar ist. $N_{AB}^{\widetilde{X}}$ ist daher geometrisch verteilt zum Parameter $\Pr(\widetilde{X}_{(s,2)} > 0) = 1 - e^{-P(b_1)\cdots P(b_l)} \approx P(b_1)\cdots P(b_l)$. Da bei der Approximation von X durch $\widetilde{X}$ die Überlappungsmöglichkeiten zwischen den Mustern nicht berücksichtigt werden, spielen diese bei der Verteilung von $N_{AB}^{\widetilde{X}}$ ebenfalls keine Rolle. Insbesondere geht auch nicht die Bedingung ein, daß D mit dem Muster A beginnt.

Nun argumentieren wir mit Y. Wir approximieren Y durch $\widetilde{Y}$. Die Bedingung, daß D mit einem Muster A beginnt, übersetzt sich in $\{Y_{(1,1)} = 1\}$ bzw. $\{\widetilde{Y}_{(1,1)} = 1\}$ und $\{\widetilde{Y}_{(1,2)} = 0\}$. Es genügt nicht, die Sites s zu betrachten, für die $\{\widetilde{Y}_{(s,2)} > 0\}$ gilt, da auch Klumpen, die mit A beginnen, eine Kopie von B enthalten können. Die Wahrscheinlichkeit, daß ein mit A beginnender Klumpen ein Muster B enthält, entspricht dem Ereignis, daß in seinem Profil $(K_i)_{i\in\mathbb{N}}$ bei Verlassen des Zustands A der Übergang $A \to B$, und nicht $A \to \mathtt{X}$, stattfindet. Die Wahrscheinlichkeit dafür ist:

$$\gamma := \frac{\Pr(K_N = B | K_{N-1} = A)}{\Pr(K_N \neq A | K_{N-1} = A)} = \frac{\sum_{x=1}^{l-1} Z_{AB,x} \Pr(b_{x+1})\cdots\Pr(b_l)}{1 - \sum_{y=1}^{l-1} Z_{AA,y} \Pr(a_{y+1})\cdots\Pr(a_l)}$$

Dies ist insbesondere auch die Wahrscheinlichkeit, daß gleich in dem im ersten Site beginnenden Klumpen ein Muster B enthalten ist. Wir gehen von einem Szenario aus, in dem wir für jeden an einem Site s mit A beginnenden Klumpen die stochastischen Abhängigkeiten zwischen seinem Profil und $(Y_\alpha)_{\alpha\in\Gamma\setminus\{(s,1)\}}$ vernachlässigen können. Dann können wir die Konfiguration der Sites, an denen Klumpen beginnen, die das Muster B enthalten, durch einen Poisson-Prozeß $Q = (Q_s)_{s<n_{\mathrm{dna}}}$ modellieren, wobei jedes Q_s die Summe von $\widetilde{Y}_{(s,2)}$ und einer (davon unabhängigen) zum zufälligen Parameter $(\widetilde{Y}_{(s,1)}, \gamma)$ binomialverteilten Zufallsvariable ist. Q_s ist also Poisson-verteilt zum Parameter

$$\begin{aligned}\lambda_Q \;=\;& \left(P(b_1)\cdots P(b_l)\cdot\left(1 - \sum_{x=1}^{l-1} Z_{AB,x}P(a_1)\cdots P(a_{l-x}) - \sum_{y=1}^{l-1} Z_{BB,y}P(b_1)\cdots P(b_{l-y})\right)\right) + \\ &+ \left(P(a_1)\cdots P(a_l)\cdot\left(1 - \sum_{x=1}^{l-1} Z_{AA,x}P(a_1)\cdots P(a_{l-x}) - \sum_{y=1}^{l-1} Z_{BA,y}P(b_1)\cdots P(b_{l-y})\right)\right)\cdot \\ &\cdot\frac{\sum_{x=1}^{l-1} Z_{AB,x} \Pr(b_{x+1})\cdots\Pr(b_l)}{1 - \sum_{y=1}^{l-1} Z_{AA,y} \Pr(a_{y+1})\cdots\Pr(a_l)}.\end{aligned}$$

Bezieht man die zusätzlichen Bedingungen $\{\widetilde{Y}_{(1,1)} = 1\}$ und $\{\widetilde{Y}_{(1,2)} = 0\}$ ein, so erhält man $\Pr(Q_1 > 0) = \Pr(Q_1 = 1) = \gamma$.

Wir approximieren N_{AB} durch $N_{AB}^{\widetilde{Y}} := \min\{s : Q_s > 0\}$. Für $s > 1$ gilt also:

$$p_s \approx \Pr(N_{AB}^{\widetilde{Y}} = s) = (1-\gamma)\cdot(1 - \Pr(Q_s > 0))^{s-2}\cdot\Pr(Q_s > 0)$$

Dabei ist $\Pr(Q_s > 0) = 1 - \exp(-\lambda_Q) \approx \lambda_Q$.

C.3.2 Die Verteilungsgewichte p_n von N_{AB}

Wir betrachten nun die Verteilungsgewichte von N_{AB}, um die dabei gewonnenen Erkenntnisse mit den Ergebnissen von Abschnitt C.3.1 zu vergleichen. Wenn die Konfiguration von Mustern oder Klumpen auf der DNA durch einen Poisson-Prozeß approximierbar ist, so sind die Abstände zwischen den Mustern bzw. Klumpen „ungefähr“ geometrisch verteilt. Legt man ein Klumpenmodell zugrunde, so kann es natürlich auch sein, daß B bereits im ersten Klumpen auftritt. Für größere n sollte sich also p_n wie das Produkt aus der Wahrscheinlichkeit, daß B nicht im ersten Klumpen enthalten ist, und dem Gewicht einer geometrischen Verteilung verhalten.

Wir untersuchen in diesem Abschnitt, ob N_{AB} bis auf einen Vorfaktor asymptotisch geometrisch verteilt ist, ob es also ein $p \in [0,1]$ gibt, so daß $\frac{p_n}{(1-p)^{n-1} \cdot p}$ für $n \to \infty$ gegen eine positive Konstante konvergiert.

Zunächst leiten wir dazu eine rekursive Formel zur Berechnung von $(p_n)_n$ her. Dazu benötigen wir die Mengen der möglichen Überlappungsweiten $\mathcal{U}_{AB}$ und $\mathcal{U}_{BB}$; vgl. Abschnitt C.1.

C.3.2.1 Die rekursive Berechnung von p_n

Für $n \leq l$ gilt $p_n = 0$ und für $n \in \{l+1, \dots, 2l-1\}$ gilt $p_n = Z_{AB,2l-n} \cdot P(b_{2l-n+1}) \cdots P(b_l)$. Falls eine Zahl $k \in \{1, \dots, l-1\}$ mit $k \in \mathcal{U}_{AB}$ und $l-k \in \mathcal{U}_{BB}$ existiert, gilt $p_{2l} = 0$, sonst gilt $p_{2l} = P(b_1) \cdots P(b_l)$.

Für $n > 2l$ gilt:

$$\begin{aligned}
p_n &= P(b_1) \cdot \ldots \cdot P(b_l) \\
&\quad - \sum_{n-l<k<n} \Pr(\{(Z_{n-l+1}, \dots, Z_n) = B\} \cap \{N_{AB} = k\}) \\
&\quad - \sum_{k \leq n-l} \Pr(\{(Z_{n-l+1}, \dots, Z_n) = B\} \cap \{N_{AB} = k\}) \\
&= P(b_1) \cdot \ldots \cdot P(b_l) - \sum_{\{k \mid \, n-l<k<n, \; l-(n-k) \in \mathcal{U}_{BB}\}} p_k \cdot P(b_{l-(n-k)+1}) \cdot \ldots \cdot P(b_l) \\
&\quad - \sum_{k \leq n-l} p_k \cdot P(b_1) \cdot \ldots \cdot P(b_l) \\
&= P(b_1) \cdot \ldots \cdot P(b_l) - \sum_{\{k \in \mathcal{U}_{BB} \mid \; 1 \leq k \leq l-1\}} p_{n-l+k} \cdot P(b_{k+1}) \cdot \ldots \cdot P(b_l) \\
&\quad - \sum_{k \leq n-l} p_k \cdot P(b_1) \cdot \ldots \cdot P(b_l)
\end{aligned}$$

Dementsprechend gilt für $n \geq 2l+2$:

$$\begin{aligned}
p_{n-1} &= P(b_1) \cdot \ldots \cdot P(b_l) - \sum_{\{k \in \mathcal{U}_{BB} \mid \; 1 \leq k \leq l-1\}} p_{n-l+k-1} \cdot P(b_{k+1}) \cdot \ldots \cdot P(b_l) \\
&\quad - \sum_{k < n-l} p_k \cdot P(b_1) \cdot \ldots \cdot P(b_l)
\end{aligned}$$

Daraus ergibt sich:

$$\begin{aligned} p_n &= p_{n-1} - p_{n-l} \cdot P(b_1) \cdots P(b_l) \\ &\quad + \sum_{\{k \in \mathcal{U}_{BB} \mid 1 \leq k \leq l-1\}} (p_{n-l+k-1} - p_{n-l+k}) \cdot P(b_{k+1}) \cdots P(b_l) \end{aligned}$$

Wir verwenden von nun an folgende Bezeichnung:

$$\rho_k := \begin{cases} P(b_1) \cdots P(b_l) & \text{falls } k = 0 \\ P(b_{k+1}) \cdots P(b_l) & \text{falls } 1 \leq k \leq l-1 \text{ und } k \in \mathcal{U}_{BB} \\ 1 & \text{falls } k = l \\ 0 & \text{sonst} \end{cases}$$

Demnach bezeichnet ρ_k für $k \in \{0, \ldots, l\}$ die bedingte Wahrscheinlichkeit, daß das Muster B an der Stelle $n + l - k$ vollendet wird, gegeben, daß es an der Stelle n vollendet wurde. Wir erhalten also für $n > 2l$:

$$\begin{aligned} p_n &= p_{n-1} - \rho_0 \cdot p_{n-l} + \sum_{k=1}^{l-1} \rho_k \cdot (p_{n-l+k-1} - p_{n-l+k}) && \text{(C.1)} \\ &= \sum_{k=0}^{l-1} (\rho_{k+1} - \rho_k) \cdot p_{n-l+k} && \text{(C.2)} \end{aligned}$$

C.3.2.2 Asymptotische Betrachtungen für p_n

Konvergiert der Quotient $(a_n/b_n)_n$ zweier Zahlenfolgen $(a_n)_n$ und $(b_n)_n$ gegen 1, so schreiben wir $a_n \sim b_n$.

Wir werden nun den Bereich eingrenzen, in dem wir den Parameter p einer geometrischen Verteilung, die sich asymptotsich wie p_n verhält, zu suchen haben. Wir verwenden dabei folgende Schreibweise von S.-Y. R. Li (1981):

$$B * B := \sum_{k=1}^{l} \frac{\rho_k}{\rho_0}$$

Li zeigt mit einem besonders plastischen Martingalargument, daß $B * B$ die erwartete Anzahl an Sites ist, die man warten muß, bis das Muster B zum ersten Mal auftritt, wenn auf keine Startsequenz bedingt wird.

Lemma C.1 *Wenn es ein $p \in]0, 1[$ mit $p_n \sim \text{const.} \cdot (1-p)^n \cdot p$ gibt, so gilt*

$$p = \frac{1}{B * B - (l-1) + R}$$

mit

$$\frac{-(l-1)^2}{B * B - 4(l-1)} \leq R \leq (l-1) \frac{B * B - \rho_0^{-1}}{B * B - 2(l-1)}.$$

Es ist zu bemerken, daß R betraglich sehr klein ausfällt, da $B * B$ in der Regel l um einige Größenordnungen übertrifft und ρ_0^{-1} meistens einen großen Anteil von $B * B$ ausmacht.

Beweis Wenn die Voraussetzung des Lemmas erfüllt ist, so müssen die Gewichte der geometrischen Verteilung zum Parameter p die Rekursionsgleichung (C.1) bzw. (C.2) erfüllen. Gesucht ist also eine Zahl $p \in]0,1[$, so daß gilt:

$$p(1-p)^n = \sum_{k=0}^{l-1} (\rho_{k+1} - \rho_k) \cdot p(1-p)^{n-l+k}$$

Das ist gleichbedeutend mit

$$(1-p)^l = \sum_{k=0}^{l-1} (\rho_{k+1} - \rho_k) \cdot (1-p)^k. \tag{C.3}$$

Diese Gleichung ist äquivalent zu

$$\rho_0 = \sum_{k=1}^{l} \rho_k p (1-p)^{k-1}.$$

Wir betrachten nun die Differenzenfolge

$$\Delta_k := (1-p)^{k-1} - (1-p)^k$$

und setzen $\quad \Delta := \Delta_l$ und $\delta_k := \Delta_k - \Delta$.

Aus

$$(1-p)^l = (1-p)^{l-1} + \sum_{k=1}^{l-1} \rho_k \cdot \Delta_k - \rho_0$$

folgt:

$$\begin{aligned} 0 &= \sum_{k=1}^{l} \rho_k \cdot \Delta_k - \rho_0 \cdot \left((1-p)^l + \sum_{k=1}^{l} \Delta_k \right) \\ &= \sum_{k=1}^{l} \rho_k \Delta + \sum_{k=1}^{l} \rho_k \delta_k - \rho_0 \sum_{k=1}^{l} \Delta - \rho_0 \sum_{k=1}^{l} \delta_k - \rho_0 (1-p)^l \end{aligned}$$

$\Rightarrow$

$$\begin{aligned} (1-p)^l &= \Delta \left(\sum_{k=1}^{l} \frac{\rho_k}{\rho_0} - l \right) + \left(\sum_{k=1}^{l} \left(\frac{\rho_k}{\rho_0} - 1 \right) \delta_k \right) \\ &= \left((1-p)^{l-1} - (1-p)^l \right) (B * B - l + R) \\ & \quad \text{mit } R := \Delta^{-1} \sum_{k=1}^{l} \left(\frac{\rho_k}{\rho_0} - 1 \right) \delta_k \end{aligned}$$

Wir erhalten:

$$p = (1 + B * B - l + R)^{-1}$$

Abzuschätzen bleibt also

$$R = \sum_{k=1}^{l} \frac{\delta_k}{\Delta} \left(\frac{\rho_k}{\rho_0} - 1 \right) = \sum_{k=1}^{l-1} \frac{\delta_k}{p(1-p)^{l-1}} \left(\frac{\rho_k}{\rho_0} - 1 \right) \qquad \text{(beachte } \delta_l = 0\text{).}$$

Zunächst einmal gilt

$$- \sum_{k=1}^{l-1} \frac{\delta_k}{p(1-p)^{l-1}} \leq R \leq \sum_{k=1}^{l-1} \left(\frac{\delta_k}{p(1-p)^{l-1}} \cdot \frac{\rho_k}{\rho_0} \right) \leq \frac{\delta_1}{p(1-p)^{l-1}} \cdot \sum_{k-1}^{l-1} \frac{\rho_k}{\rho_0}$$

und wegen $\delta_1 \leq \delta_k$ und

$$\begin{aligned} \frac{\delta_1}{p(1-p)^{l-1}} &= (1-p)^{1-l} - 1 \\ &\leq (1-(l-1)p)^{-1} - 1 = \frac{1}{1+p-pl} - 1 \end{aligned}$$

folgt:

$$-\frac{l-1}{1+p-pl} + l - 1 \leq - \sum_{k=1}^{l-1} \frac{\delta_k}{p(1-p)^{l-1}} \leq R \leq \frac{B*B-\rho_0^{-1}}{1+p-pl} - B*B + \rho_0^{-1}$$

Aus $p = (B*B - l + 1 + R)^{-1}$ folgt:

$$\frac{-(l-1)^2}{B*B-2l+2+R} \leq R \leq \frac{(B*B-\rho_0^{-1})(l-1)}{B*B-2l+2+R}$$

Da sonst die linke Seite größer wäre als die rechte, sind die Nenner positiv, und es folgt:

$$-(l-1)^2 \overset{(i)}{\leq} R^2 + R(B*B-2l+2) \overset{(ii)}{\leq} (B*B-\rho_0^{-1})(l-1)$$

Aus der Ungleichung (ii) ergibt sich:

$$0 \geq R^2 + R(B*B-2l+2) - (B*B-\rho_0^{-1})(l-1)$$

Nach Anwendung der „p-q-Formel" und unter Berücksichtigung der für $g > k > 0$ geltenden Ungleichung $-g + \sqrt{g^2+k} \leq k/(2g)$ erhält man:

$$R \;\leq\; (l-1)\frac{B*B-\rho_0^{-1}}{B*B-2l+2}$$

Analog dazu folgt aus (i) mit $-g + \sqrt{g^2-k} \geq -\sqrt{k}/(2\sqrt{g^2-k})$ (für $g > k > 0$):

$$R \;\geq\; \frac{-(l-1)^2}{2 \cdot \sqrt{\left(\frac{B*B-2l+2}{2}\right)^2 - (l-1)^2}} \;\geq\; \frac{-(l-1)^2}{B*B-4(l-1)}$$

□

C.3.2.3 Die Konvergenz der Gewichte

Wir zeigen nun, daß $p_n \sim \text{const.} \cdot (1-p)^{n-1} \cdot p$ tatsächlich gilt. Da es uns in erster Linie nur auf den Vergleich mit den Ergebnissen aus den Abschnitten C.1 und C.2 ankommt, schließen wir dabei die technisch aufwendigeren Fälle aus, daß B sehr kurz oder sehr stark selbstüberlappend ist oder daß eine einzelne Base ein sehr großes Gewicht hat. Wir nehmen also für den Rest des Abschnitts an, daß l größer als 4 ist, $\mathcal{U}_{BB}$ keine der Zahlen $l-1$, $l-2$, $l-3$ und $l-4$ enthält und daß $P(\mathtt{A})$, $P(\mathtt{G})$, $P(\mathtt{C})$ und $P(\mathtt{T})$ jeweils größer als 0 und kleiner als $\frac{1}{2}$ ist. Für $l > 4$ gilt dann insbesondere $\rho_{l-1} = \rho_{l-2} = \rho_{l-3} = \rho_{l-4} = 0$. Außerdem ist dann $B * B \geq 2^l$.

Wir setzen

$$\zeta_k := \begin{cases} P(b_1) \cdots P(b_l) & \text{falls } k = 0 \\ P(b_{k+1}) \cdots P(b_l) & \text{falls } 1 \leq k \leq l-1 \text{ und } k \in \mathcal{U}_{AB} \\ 0 & \text{sonst} \end{cases}$$

Mit diesen Voraussetzungen und Schreibweisen beweisen wir nun folgenden Satz:

Satz C.2 *Es existiert ein $p \in]0,1[$, das die Voraussetzung $p_n \sim \text{const.} \cdot p \cdot (1-p)^{n-1}$ von Lemma C.1 erfüllt. Genauer gilt:*

$$p_n \sim \frac{\rho_0 \cdot \left(\sum_{j=0}^{l-1} (\zeta_j - \zeta_{j+1}) \cdot (1-p)^{j-l} \right)}{p \cdot \zeta_0 \cdot \sum_{k=1}^{l} k \cdot (\rho_{l-k+1} - \rho_{l-k}) \cdot (1-p)^{3-k}} \cdot p \cdot (1-p)^{n-1}$$

Für den Beweis verwenden wir folgenden Satz (vgl. Feller (1968), Kapitel XI.4):

Satz C.3 *Die erzeugende Funktion der Verteilung einer $\mathbb{N}$-wertigen Zufallsvariablen mit den Gewichten p_n sei von der Form $G(s) = U(s)/V(s)$, wobei $U(s)$ und $V(s)$ Polynome seien. $V(s)$ habe eine einfache Nullstelle s_1, die betraglich kleiner als alle anderen Nullstellen von $V(s)$ sei. Dann gilt:*

$$p_n \sim \frac{-U(s_1)}{V'(s_1) s_1^{n+1}}$$

Beweis von Satz C.2: Die erzeugende Funktion von N_{AB} haben Gerber und Li (1981) hergeleitet:

$$\mathrm{E}[z^{N_{AB}}] = \frac{1 + (1-z) \sum_{j=1}^{l-1} \frac{\zeta_j}{\zeta_0} z^{-j}}{1 + (1-z) \sum_{j=1}^{l} \frac{\rho_j}{\rho_0} z^{-j}} = \frac{\rho_0 \left(\sum_{j=0}^{l-1} (\zeta_j - \zeta_{j+1}) \, z^{l-j} \right)}{\zeta_0 \left(1 + \sum_{j=0}^{l-1} (\rho_j - \rho_{j+1}) \, z^{l-j} \right)}$$

Um Satz C.3 anwenden zu können, müssen wir also die betraglich kleinste (möglicherweise komplexe) Nullstelle von

$$f(z) := \sum_{k=0}^{l-1} (\rho_k - \rho_{k+1}) \cdot z^{l-k} + 1 = \rho_0 z^l + \sum_{k=1}^{l} \rho_k \cdot (1-z) z^{l-k}$$

finden. Wegen $f(0) > 0$ und $f(2) = \rho_0 2^l - \sum_{k=1}^{l} \rho_k 2^{l-k} \leq \rho_0 2^l - \rho_l < \left(\frac{1}{2}\right)^l \cdot 2^l - 1 = 0$ hat f eine Nullstelle in $]0, 2[$. Also existiert auch ein $p \in]0, \frac{1}{2}[$, welches die Gleichung $f((1-p)^{-1}) = 0$ erfüllt und damit die Rekursionsgleichung (C.3) löst. Mit Lemma C.1 können wir dieses p ziemlich genau bestimmen. Zu zeigen bleibt also, daß f keine betraglich kleinere Nullstelle als

$$z_0 := \frac{1}{1 - \frac{1}{B*B-l+1+R}} = 1 + \frac{1}{B*B-l+R} \qquad \text{(mit } R \text{ wie in Lemma C.1) besitzt.}$$

Wir betrachten dazu die Ableitung

$$f'(z) = \sum_{k=1}^{l} k \cdot (\rho_{l-k} - \rho_{l-k+1}) \cdot z^{k-1} = \rho_{l-1} - 1 + \sum_{k=2}^{l} k \cdot (\rho_{l-k} - \rho_{l-k+1}) \cdot z^{k-1}.$$

Da nach Voraussetzung das Muster B keine Periode ≤ 4 besitzt und damit $\rho_{l-k} = 0$ für $1 \leq k \leq 4$ gilt, folgt für $|z| \leq z_0$:

$$\begin{aligned} \left|f'(z) + 1\right| &= \left|\sum_{k=5}^{l} k \cdot (\rho_{l-k} - \rho_{l-k+1}) \cdot z^{k-1}\right| \leq z_0^{l-1} \cdot \left(5 \cdot \left(\frac{1}{2}\right)^5 + \sum_{k=6}^{l} k \cdot \left(\frac{1}{2}\right)^{k-1}\right) \\ &\leq z_0 \cdot \frac{19}{32} < 1 \qquad \left(\text{wegen } \sum_{k=n}^{\infty} (k+1) \left(\frac{1}{2}\right)^k = \frac{n+2}{2^{n-1}}\right). \end{aligned}$$

Daraus folgt, daß der Realteil von $f'(z)$ negativ ist:

$$\Re(f'(z)) \leq \left|f'(z) + 1\right| - 1 < 0$$

Damit ist zunächst einmal sichergestellt, daß z_0 eine einfache Nullstelle ist. Außerdem gilt $f(r) > 0$ für alle $r \in \mathbb{R}$ mit $|r| < z_0$. Nach den Cauchy-Riemann'schen Differentialgleichungen beschreibt $\Re(f')$ auch das Wachstum des Imaginärteils $\Im(f)$ von f in Richtung der imaginären Achse. Da $\Im(f)$ auf der reellen Achse verschwindet, folgt:

$$\Im(f(z)) \begin{cases} > 0 & \text{für} \quad |z| < z_0 \text{ und } \Im(z) < 0 \\ < 0 & \text{für} \quad |z| < z_0 \text{ und } \Im(z) > 0 \end{cases}$$

Damit ist gezeigt, daß z_0 die Bedingungen von Satz C.3 erfüllt, und es folgt:

$$p_n \sim -\frac{\rho_0 \left(\sum_{j=0}^{l-1} (\zeta_j - \zeta_{j+1}) \cdot z_0^{l-j}\right)}{\zeta_0 \sum_{k=1}^{l} k \cdot (\rho_{l-k} - \rho_{l-k+1}) z_0^{k-1}} \cdot z_0^{-n-1}$$

Mit $z_0 = (1-p)^{-1}$ folgt die Behauptung.

□

Etwas kryptisch erscheint in Satz C.2 zunächst der Vorfaktor

$$\begin{aligned} c &:= \frac{\rho_0 \cdot \left(\sum_{j=0}^{l-1} (\zeta_j - \zeta_{j+1}) \cdot (1-p)^{j-l}\right)}{p \cdot \zeta_0 \cdot \sum_{k=1}^{l} k \cdot (\rho_{l-k+1} - \rho_{l-k}) \cdot (1-p)^{3-k}} \\ &= \frac{\frac{1}{p} - \sum_{j=1}^{l} \frac{\zeta_j}{\zeta_0} (1-p)^{j-1}}{p \cdot \sum_{k=1}^{l} k \frac{\rho_{l-k}}{\rho_0} (1-p)^{l-k+2} + \sum_{k=1}^{l} \frac{\rho_k}{\rho_0} (1-p)^{k+2} - l \cdot (1-p)^2}. \end{aligned}$$

Falls $p \approx (B*B-l+1)^{-1} \approx B*B^{-1}$ sehr klein und damit $(1-p)^x$ für Exponenten x der Größenordnung von l ungefähr 1 ist, so folgt mit $A*B := \sum_{k=1}^{l} \frac{\zeta_k}{\zeta_0}$:

$$c \approx \frac{B*B-l+1-A*B}{B*B-l} \approx 1-\frac{A*B}{B*B}$$

Die Approximation $p_n \approx \left(1-\frac{A*B}{B*B}\right)p(1-p)^{n-1}$ legt folgende heuristische Sichtweise nahe: Mit einer Wahrscheinlichkeit von ungefähr $A*B/B*B$ tritt B im „Einflußbereich" der Startsequenz A auf. Geschieht dies nicht, so ist die Anzahl der Basen, die man noch auf B warten muß, approximativ geometrisch verteilt, und zwar zum Parameter $p \approx (B*B-l+1)^{-1}$.

C.3.3 Beispiel

Wir vergleichen nun die in den Abschnitten C.3.1 und C.3.2 entwickelten Approximationen der Verteilung von N_{AB} für ein konkretes Musterpaar (A, B). Wir wählen $A = \texttt{CGTATACG}$ und $B = \texttt{ACGACGAC}$, und es sei $P(\texttt{A}) = P(\texttt{G}) = P(\texttt{C}) = P(\texttt{T}) = \frac{1}{4}$.

Zunächst zu den Approximationen aus Abschnitt C.3.1: $N_{AB}^{\widetilde{X}}$ ist einfach geometrisch verteilt zum Parameter $1-\exp\left(-\left(\frac{1}{4}\right)^8\right) \approx 1{,}526 \cdot 10^{-5}$.

Für $N_{AB}^{\widetilde{Y}}$ stellen wir zunächst fest, daß $Z_{AA,2} = Z_{AB,3} = Z_{BA,1} = Z_{BB,5} = 1$ ist, und daß ansonsten $Z_{\cdot\cdot,\cdot} = 0$ gilt. Damit ergibt sich

$$\gamma = \frac{\left(\frac{1}{4}\right)^5}{1-\left(\frac{1}{4}\right)^7} \approx 9{,}766 \cdot 10^{-4}$$

und

$$\begin{aligned}\lambda_Q &= \left(\frac{1}{4}\right)^8 \cdot \left(1-\left(\frac{1}{4}\right)^5-\left(\frac{1}{4}\right)^3\right)+\left(\frac{1}{4}\right)^8 \cdot \left(1-\left(\frac{1}{4}\right)^6-\left(\frac{1}{4}\right)^7\right) \cdot \frac{\left(\frac{1}{4}\right)^5}{1-\left(\frac{1}{4}\right)^6} \\ &\approx 1{,}5020 \cdot 10^{-5}.\end{aligned}$$

Es gilt also:

$$\Pr(N_{AB}^{\widetilde{Y}} = s) \approx (1-9{,}766 \cdot 10^{-4}) \cdot (1-1{,}5020 \cdot 10^{-5})^{s-2} \cdot 1{,}5020 \cdot 10^{-5}$$

Nun zur Asymptotik im Sinne der Abschnitte C.3.2.2 und C.3.2.3. Streng genommen erfüllt die Seqeunz B nicht die Bedingungen, die wir am Beginn von Abschnitt C.3.2.3 vorausgesetzt haben. Wie wir sehen werden, ist die Konvergenz dennoch gegeben. Für das gegebene Paar (A, B) gilt $\zeta_0 = \rho_0 = \left(\frac{1}{4}\right)^8$, $\zeta_3 = \left(\frac{1}{4}\right)^5$, $\rho_2 = \left(\frac{1}{4}\right)^6$, $\rho_5 = \left(\frac{1}{4}\right)^3$ und $\rho_8 = 1$. Damit ergibt sich $A*B = 4^3 = 64$, $B*B = 4^8+4^5+4^2 = 66576$ und

$$p \approx \frac{1}{B*B-7} \approx 1{,}5022 \cdot 10^{-5}.$$

Für den Vorfaktor erhalten wir $c \approx 1-7{,}97 \cdot 10^{-4}$. Offensichtlich ist in unserem Beispielszenario $(1-p)^l$ nicht unbedingt vernachlässigbar. Es gibt nämlich doch eine gewisse Differenz zwischen

$1-c$ und

$$1-\frac{B*B-7-A*B}{B*B-8}\approx 9{,}464\cdot 10^{-4}\quad\text{bzw.}\quad \frac{A*B}{B*B}\approx 9{,}613\cdot 10^{-4}.$$

Wir vergleichen nun die Approximationen mit der tatsächlichen Verteilung von N_{AB}, deren Gewichte p_n wir mit C.1 berechnen können. In Abbildung C.1 ist zu sehen, daß alle drei Approximationen sehr nahe an die tatsächlichen Verteilungsgewichte herankommen. Bei genauem Hinsehen sind nur die Gewichte von $N_{AB}^{\widetilde{X}}$ von denen von N_{AB} zu unterscheiden.

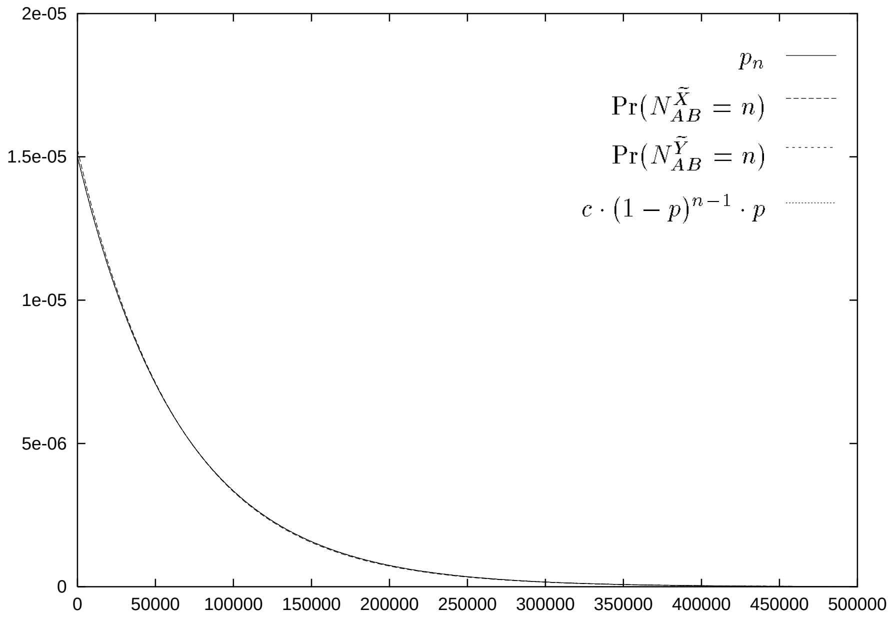


Abbildung C.1:

In Abbildung C.2 treten die Unterschiede in der Präzision der drei Approximationsverfahren weitaus deutlicher zutage. Die drei Kurven zeigen das Verhältnis zwischen den Gewichten von N_{AB} und den drei Approximationen. Wie wir sehen, ist $p_n/(c\cdot(p-1)^n\cdot p)$ bereits für kleine n sehr nahe bei 1. Die Gewichte von N_{AB} sind für kleine n um etwas mehr als 1,5% kleiner als die von $N_{AB}^{\widetilde{X}}$, und der Quotient der beiden wächst pro 100 000 um circa 2%. Die Unterschiede zwischen den Verteilungsgewichten von N_{AB} und $N_{AB}^{\widetilde{Y}}$ sind demgegenüber sehr gering. Offensichtlich läßt sich $\mathcal{L}(N_{AB})$ wegen der starken Selbstüberlappungsfähigkeit von B wesentlich besser durch $\mathcal{L}(N_{AB}^{\widetilde{Y}})$ als durch $\mathcal{L}(N_{AB}^{\widetilde{X}})$ approximieren, obgleich für viele Anwendungen auch der Unterschied zwischen $\mathcal{L}(N_{AB})$ und $\mathcal{L}(N_{AB}^{\widetilde{X}})$ bereits hinreichend klein sein mag.

C.4 Banden auf einem DNA-Strang

Wie groß ist die Wahrscheinlichkeit, daß es auf einem DNA-Strang eine RAPD-Bande (s,r) gibt? Als Muster betrachten wir hier die Menge $\{M_1,M_2\}=\{\mathcal{P},\mathcal{K}\}$, wobei $\mathcal{P}=(a_1,\ldots,a_l)$

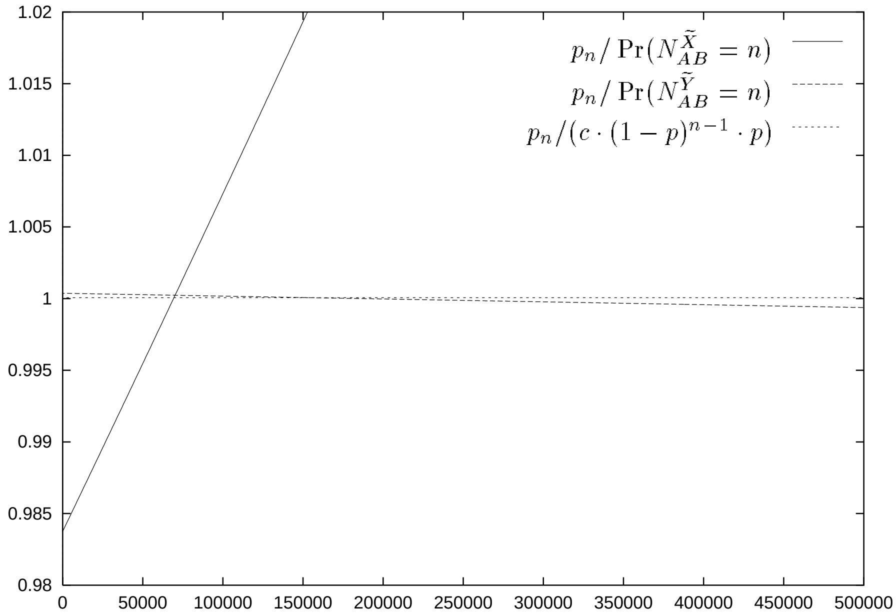


Abbildung C.2:

die verwendete Primersequenz und $\mathcal{K} = (b_1, \ldots, b_l)$ ihr Komplement bezeichnet. Wenn wir die in Abschnitt C.1 dargestellte Approximation von X durch $\widetilde{X}$ verwenden können, so ist die gesuchte Wahrscheinlichkeit gegeben durch:

$$
\begin{aligned}
&\Pr(\{X_{(s,1)} = X_{(r,2)} = 1\} \cap \{\forall_{i \in \{s+1,\ldots,r-1\}} : X_{(i,1)} = X_{(i,2)} = 0\}) \\
&\quad\approx \Pr(\{\widetilde{X}_{(s,1)} = \widetilde{X}_{(r,2)} = 1\} \cap \{\forall_{i \in \{s+1,\ldots,r-1\}} : \widetilde{X}_{(i,1)} = \widetilde{X}_{(i,2)} = 0\}) \\
&\quad= P(a_1) \cdots P(a_l) \cdot e^{-P(a_1)\cdots P(a_l)} \cdot P(b_1) \cdots P(b_l) \cdot e^{-P(b_1)\cdots P(b_l)} \cdot \\
&\qquad \cdot e^{-(r-s-1)\cdot(P(a_1)\cdots P(a_l) + P(b_1)\cdots P(b_l))} \\
&\quad= P(a_1) \cdots P(a_l) \cdot P(b_1) \cdots P(b_l) \cdot e^{-(r-s)\cdot(P(a_1)\cdots P(a_l) + P(b_1)\cdots P(b_l))}
\end{aligned}
$$

Dabei ist vorausgesetzt, daß $(r-s)$ kleiner gleich dem Amplifikationsbereich n_{amp} ist.

Wir nehmen nun für die Primer-Länge und den Amplifikationsbereich die für RAPD-PCR-Versuche plausiblen Werte $l = 10$ und $n_{\text{amp}} = 3000$ an (siehe Anhang A und B). Wenn wir zusätzlich von $P(\mathtt{A}) = P(\mathtt{G}) = P(\mathtt{C}) = P(\mathtt{T}) = \frac{1}{4}$ ausgehen, erhalten wir für die Wahrscheinlichkeit, daß eine sichtbare Bande (s, r) existiert, den Wert $\approx 9{,}1 \cdot 10^{-13} \cdot (1 - 1{,}91 \cdot 10^{-6})^{r-s}$. Wegen $0 < r - s \leq 3000$ liegt der Faktor $(1 - 1{,}91 \cdot 10^{-6})^{r-s}$ zwischen 0,9943 und 1 und ist damit wohl für die meisten Fragen bedeutungslos. Wenn wir also unter allen sichtbaren Banden, die ein zufälliger DNA-Strang hervorbringt, eine rein zufällig auswählen, so ist ihre Länge ungefähr gleichverteilt auf $\{1, \ldots, n_{\text{amp}}\}$.

Diese Überlegungen sind natürlich nur dann sinnvoll, wenn der Unterschied zwischen $\mathcal{L}(X)$ und $\mathcal{L}(\widetilde{X})$ vernachlässigbar ist. Betrachten wir also nun die in Abschnitt C.1 hergeleitete obere Abschätzung des Totalvariationsabstandes zwischen den beiden Verteilungen. Wir gehen von dem optimalen Fall aus, daß es in $\{\mathcal{P}, \mathcal{K}\}$ keine Überlappungsmöglichkeiten gibt. Dann bringt es keine Vorteile, $\mathcal{L}(Y)$ durch $\mathcal{L}(\widetilde{Y})$ statt $\mathcal{L}(X)$ durch $\mathcal{L}(\widetilde{X})$ zu approximieren, und wir erhalten die Abschätzung

$$d_{TV}(\mathcal{L}(X), \mathcal{L}(\widetilde{X})) \leq \sum_{\substack{\alpha,\beta\in\Gamma \\ \alpha\neq\beta\in\Gamma_\alpha}} \mathbb{E}X_\alpha \cdot \mathbb{E}X_\beta = 2\cdot(n_{\mathrm{dna}}-9)\cdot 9\cdot\left(2\cdot\left(\frac{1}{4}\right)^{10}\right)^2 \approx n_{\mathrm{dna}}\cdot 6{,}548\cdot 10^{-11}.$$

Was sind vernünftige Werte für n_{dna}? Bei einem DNA-Strang der Länge n_{dna} beträgt die erwartete Anzahl an sichtbaren Banden ungefähr $n_{\mathrm{dna}} \cdot \left(\frac{1}{4}\right)^{20} \cdot 3000$. Wenn die erwartete Anzahl an sichtbaren Banden sehr gering ist (z. B. 0,1), liefern die meisten Versuche kein verwertbares Ergebnis. Dies anzunehmen ergibt also wenig Sinn. Wenn wir zum Beispiel davon ausgehen, daß die erwartete Anzahl an Banden etwa 3 beträgt, so muß n_{dna} ungefähr $3\cdot 4^{20}/3000 \approx 1{,}01\cdot 10^9$, also rund eine Milliarde betragen. (Zum Vergleich: Ein einfacher menschlicher Chromosomensatz enthält ungefähr 3 Milliarden Basenpaare.) Als obere Abschätzung für $d_{TV}(\mathcal{L}(X), \mathcal{L}(\widetilde{X}))$ erhalten wir also ungefähr 0,07. Dieser Wert ist nicht hinreichend klein, als daß man ihn heranziehen könnte, wenn man bei der statistischen Analyse von RAPD-Daten die Approximation von $\mathcal{L}(X)$ durch $\mathcal{L}(\widetilde{X})$ rechtfertigen möchte. Außerdem werden bei den meisten Anwendungen die Bandensignale, die man von mehreren, stochastisch abhängigen DNA-Strängen erhält, miteinander verglichen. Andererseits ist der Totalvariationsabstand auch für solche Unterschiede zwischen Verteilungen von Musterkonfigurationen sensitiv, die sich nicht auf die Bandenkonfigurationen auswirken. In Kapitel 1 beschäftigen wir uns mit Poisson-Approximationen für gemeinsame Musterkonfigurationen auf verwandten DNA-Strängen. Insbesondere lernen wir in Kapitel 1 eine Möglichkeit kennen, daß Musterüberlappungseffekte bei RAPD-PCR-Untersuchungen verwandter DNA-Stränge vernachlässigbar sind – zumindest in praxisrelevanten Szenarien.

C.5 Fazit

Mit Hilfe der Chen-Stein-Methode läßt sich der Einfluß von Musterüberlappungseffekten auf die Verteilung von Musterkonfigurationen auf Folgen von Zufallsvariablen quantifizieren. Sind die lokalen Abhängigkeiten zu stark, so empfiehlt es sich, nicht die Stellen, an denen Muster beginnen, durch einen Poisson-Prozeß zu approximieren, sondern die Stellen, an denen Klumpen von überlappenden Mustern beginnen.

Literatur

D. Aldous (1989) *Probability Approximations via the Poisson Clumping Heuristic.* Springer-Verlag, Berlin.

R. Arratia, L. Goldstein, L. Gordon (1989) Two Moments Suffice for Poisson Approximations: The Chen-Stein Method. *The Annals of Probability* **17(1)**, 9-25.

R. Arratia, D. Martin, G. Reinert, M. Waterman (1996) Poisson Process Approximation for Sequence Repeats, and Sequencing by Hybridization. *Journal of Computational Biology* **3.3**, 425-463.

K. B. Athreya, P. E. Ney (1972) *Branching Processes.* Springer-Verlag, Berlin.

A. D. Barbour, L. Holst, S. Janson (1992) *Poisson Approximation.* Oxford University Press, Oxford.

L. H. Y. Chen (1975) Poisson approximation of dependent trials. *Annals of Probability* **3**, 534-45.

A. G. Clark, C. M. S. Lanigan (1993) Prospects for Estimating Nucleotide Divergence with RAPD's. *Mol. Biol. Evol.* **10(5)**, 1096-1111.

H. Dinges, H. Rost (1982) *Prinzipien der Stochastik.* Teubner, Stuttgart.

J. L. Doob (1953) *Stochastic Processes.* Wiley, New York.

W. Feller (1968) *An Introduction to Probability Theory and Its Applications.* Volume 1, Wiley, New York.

J. Felsenstein (1993) *PHYLIP: phylogenetic inference package.* Version 3.5c. Department of Genetics, University of Washington, Seattle.

H. U. Gerber, S.-Y. R. Li (1981) The occurence of sequence patterns in repeated experiments and hitting times in a markov chain.
Stochastic Processes and their Applications **11**, 101-108.

L. J. Guibas, A. M. Odlyzko (1981) String Overlaps, Pattern Matching, and Nontransitive Games.
Journal of Combinatorial Theory **A 30**, 183-208.

H. Hadrys, M. Siva-Jothy, B. Schierwater (1992) Applications of random amplified polymorphic DNA (RAPD) in molecular ecology.
Mol. Ecol. **1**, 55-63.

J. P. Huelsenbeck, D. M. Hillis (1993) Success of phylogenetic methods in the four-taxon case.
Syst. Biol. **42**, 247-264.

T. H. Jukes, C. R. Cantor (1969) Evolution of protein molecules.
In: H. N. Munro (Ed.): *Mammalian protein metabolism.*
Academic Press, New York, 21-132.

O. Kallenberg (1986) *Random Measures.*
(vierte Auflage), Akademie-Verlag, Berlin.

A. F. Karr (1986) *Point Processes an their Statistical Inference.*
Marcel Dekker, New York.

M. Kimura (1983) *The Neutral Theory of Molecular Evolution.*
Cambridge University Press, Cambridge.

M. Krawczack, J. Reiss, J. Schmidtke, U. Rösler (1989) Polymerase chain reaction: Replication errors and reliability of gene diagnosis.
Nucl. Acids Res. **18**, 5153-5156.

S.-Y. R. Li (1981) A martingale approach to the study of occurence of sequence patterns in repeated experiments.
The Annals of Probability **8.6**, 1171-1176.

D. Mumford (1998) Trends in the Profession of Mathematics.
DMV-Mitteilungen 2/98, Sonderbeilage zum International Congress of Mathematicians 1998 in Berlin: Zukunft der Mathematik, 25-29.

J. Nedelman, P. Heagerty, C. Lawrence (1992) Quantitative PCR: Procedures and precisions.
Bull. Mathem. Biol. **54**, 477-502.

B. Schierwater, D. Metzler, K. Krüger, B. Streit (1996) The Effects of Nested Primer Binding Sites on the Reproducibility of PCR: Mathematical Modeling and Computer Simulation Studies.
J. Comp. Biol. **3.2**, 235-251.

C. Stein (1971) A bound for the error in the normal approximation to the distribution of a sum of dependent random variables.
Proc. Sixth Berkeley Symp. Math. Statist. Prabab. **3**, 583-602, Univ. California Press.

K. Strimmer, A. von Haeseler (1996) Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies.
Mol. Biol. Evol. **13(7)**, 964-969.

K. Strimmer, A. von Haeseler (1997) Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment.
Proc. Natl. Acad. Sci. USA **94**, 6815-6819.

D. L. Swofford, G. J. Olsen (1990) Phylogeny reconstruction.
In: *Molecular Systematics.* (ed. D. M. Hill, C. Moritz), Sinauer, Sunderland, Massachusetts.

J. D. Watson, F. H. C. Crick (1953) Genetical implications of the structure of deoxyribonucleic acid.
Nature **171**, 964-967.

G. Weiss, A. von Haeseler (1995) Modeling the Polymerase Chain Reaction.
J. Comp. Biol. **2**, 49-61.

Lebenslauf

Dirk Metzler
geboren am 19. Februar 1969 in Bad Homburg v. d. H.

1974–79	Grundschule Weißkirchen (Oberursel)
1979–85	Integrierte Gesamtschule Stierstadt
1985–88	gymnasiale Oberstufe der Gesamtschule Oberursel Leistungskurse: Mathematik und Biologie Abitur
Herbst 1988	Immatrikulation für Mathematik und Informatik an der Johann Wolfgang Goethe-Universität
1988–90	Grundstudium bei den Professoren R. Bieri, O. Drobnik H. F. de Groote, G. Kersting und W. Metzler
Herbst 1990	Vordiplom in Mathematik mit Nebenfach Informatik
1990–94	Hauptstudium bei Prof. Bieri, Prof. de Groote, Dr. Ferebee, Prof. Schnorr und Prof. Wakolbinger; Diplomarbeit bei Prof. de Groote zum Thema „Garbentheoretische Methoden in der Differentialgeometrie“ (Zweitgutachter: Prof. Constantinescu)
Februar 1994	Diplom in Mathematik mit Nebenfach Informatik
ab Herbst 1994	Bearbeitung des Dissertationsthemas bei Prof. Wakolbinger
1.8.95–31.7.97	Stipendium der Graduiertenförderung des Landes Hessen
16.8.97–15.8.98	wissenschaftlicher Mitarbeiter der AG 7.1 am FB Mathematik der Universität Frankfurt (bei Prof. H. Dinges)
seit 16.10.98	wissenschaftlicher Mitarbeiter der AG 1.1 am FB Mathematik der Universität Frankfurt (bei Prof. P. Kloeden)