

Bianca Büttner
Bioinformatik 6. Semester
bbuettne@hotmail.com



Seminar: Aktuelle Themen der Bioinformatik SS 2005

Rekonstruktion und orthologische Analyse eines Genbaums über Gen-/Artenbaum-Reconcilierung

Papers von L.Arvestad, A.Berglund, J.Lagergren & B.Sennblad
(Stockholm Bioinformatics Center):

- 1) Bayesian gene/species tree reconciliation and orthology analysis using MCMC (2003)
- 2) Gene tree reconstruction and orthology analysis based on an integrated model and sequence evolution (2004)

GLIEDERUNGSÜBERSICHT

1. EINFÜHRUNG	3
2. DIE WICHTIGSTEN DEFINITIONEN.....	5
3. EIN KURZER ÜBERBLICK ÜBER DEN ALGORITHMUS.....	5
4. DAS GEN-(SEQUENZ)-EVOLUTIONS-MODELL.....	6
5. DER MCMC-ALGORITHMUS FÜR GENBÄUME.....	9
6. DIE ORTHOLOGISCHE ANALYSE.....	10
7. DER MCMC-ALGORITHMUS FÜR ARTENBÄUME.....	11
8. DIE BERECHNUNG DER LIKELIHOOD FÜR EINE RECONCILIERUNG.....	11
9. DIE BERECHNUNG DER MAXIMUM-LIKELIHOOD.....	14
10. DIE BERECHNUNG DER WAHRSCHEINLICHKEIT EINES GENBAUMS.....	15
11. DISKUSSION.....	15
12. LITERATURREFERENZEN.....	16

1. EINFÜHRUNG

Die phylogenetische Analyse von Genomdaten ist in der vergleichenden Genetik von fundamentaler Bedeutung. Dabei sind die Genbaum-Rekonstruktion und die orthologische Analyse wichtige Methoden, um Genominformationen von unterschiedlichen Organismen zu gewinnen und Genfamilien zu studieren. Zum Beispiel ist es somit möglich, die Funktion eines Genes vorauszusagen, da Gene einer Genfamilie häufig für ähnliche Proteine mit gleicher Funktion kodieren. Die hier von Lars Arvestad et al. entwickelten Algorithmen und Berechnungsmethoden sollen für dieses Teilgebiet wichtige Lösungsansätze liefern.

Mit der orthologischen Analyse untersucht man bei Genfamilien, aus welchem gemeinsamen Gen-Vorfahren sich ihre Mitglieder entwickelt haben. Sie basiert auf Walter Fitch's original Definition der Orthologie (1970, Systematische Zoologie). Fitch's Definition unterscheidet zwischen paralogen und orthologen Genen. Beide Genarten sind homolog, aber paraloge Gene entstehen durch Genduplikation und gehören zum Genom einer einzigen Spezies, während orthologe Gene das Resultat einer Speziation (Artenbildung) sind und durch sie der letzte gemeinsame Vorfahre zweier verschiedener Spezies festgestellt werden kann. In Abb. 1 wird ein Beispiel dazu gezeigt [7].

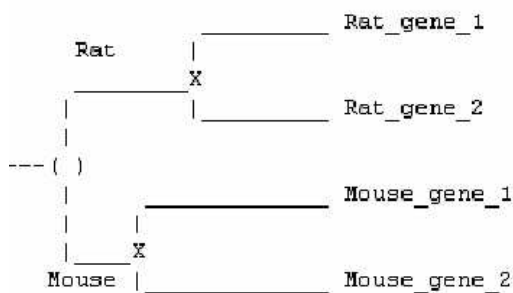


Abb. 1: () = Speziation, X = Duplikation
Die Rattengene sind zueinander paralog, genauso wie die beiden Mausgene zueinander paralog sind. Vergleicht man jedoch die Maus- und die Rattengene miteinander, so stellt man fest, dass sie durch eine Speziation getrennt wurden und daher ortholog zueinander sind.

Aufgrund der hohen Sequenzähnlichkeit innerhalb einer Genfamilie, können also durch die entstandenen Genduplikationen Rückschlüsse auf die Entwicklung einer solchen Familie zugelassen werden. Bereits 1962 entdeckten Zuckerkandl und Pauling, dass es unterschiedliche Arten von Globinen innerhalb einer Spezies gab und führten Vergleiche zwischen den Mitgliedern dieser Genfamilie durch. So entstanden die ersten Genbäume. Genduplikationen und andere biologische Mechanismen, wie z. B. Genverluste (Duplikate, die nach einiger Zeit wieder aus dem Genom verschwinden), lateraler Gentransfer (Genübertragung außerhalb der geschlechtlichen Fortpflanzung) und konvergente Entwicklung, sind allerdings auch der Grund, warum ein Genbaum häufig nicht exakt mit einem Artenbaum übereinstimmt. Goodman et al. suchten 1979 als Erste nach einem Algorithmus, der Gen- und Artenbäume in Übereinstimmung miteinander bringen sollte. Ihr Ziel war es, einen kombinierten Gen-/Artenbaum – in Form einer Reconcilierung – zu finden, der die geringste Anzahl an Substitutionen, Genduplikationen und Genverlusten aufwies. Die folgende Abbildung zeigt an einem Beispiel, wie die Reconcilierung von Goodman et al. aussah [3].

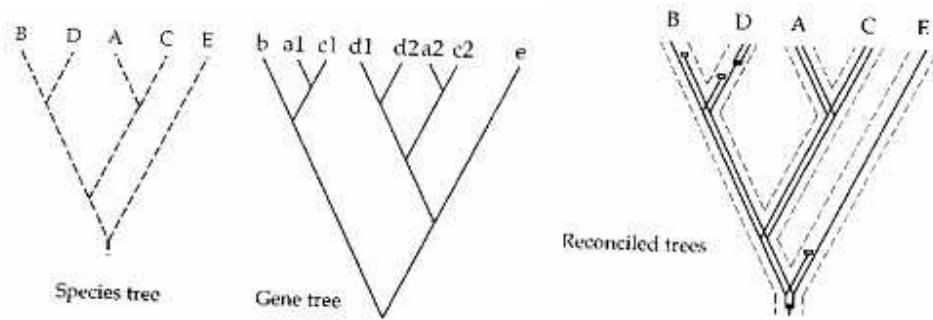


Abb. 2: Arten- und Genbaum wurden für die Reconcilierung übereinandergelegt und abgeglichen. Die gefüllten Rechtecke stellen Genduplikationen dar, die Genverluste werden durch die leeren Rechtecke symbolisiert. Diese Reconcilierung zeigt zwei Duplikationen und drei Genverluste.

Zusammen mit einem Genbaum und in Abhängigkeit des entsprechenden Artenbaums erklärt eine Reconcilierung die Evolution einer Genfamilie und ist damit ein fundamentales Konzept in der orthologischen Analyse. Durch Reconcilierung ist es möglich die Anzahl und die Zeitpunkte von Duplikationen und Genverlusten in einem Genbaum festzustellen, wodurch man herausfinden kann, welche Gene paralog oder ortholog sind. Die parsimonischen Reconcilierungsmethoden von Goodman et al. und seinen Nachfolgern sind bis heute üblich. Zuerst wird ein Genbaum konstruiert und dann mit dem dazugehörigen Artenbaum abgeglichen. Bei dieser Methode findet allerdings Einiges keine Beachtung. Zum Beispiel müssen die Zeitpunkte der Duplikationen und Speziationen im Genbaum mit dem Zeitrahmen des Artenbaums übereinstimmen. Auch muss berücksichtigt werden, wie wahrscheinlich oder unwahrscheinlich der konstruierte Genbaum und die entsprechende Reconcilierung ist. Des weiteren führen die parsimonischen Methoden durch die Minimierung von Genduplikationen häufig zu Fehlinterpretationen oder lassen plausible Lösungen außer acht. Ein Beispiel für die falsche Interpretation einer Reconcilierung durch parsimonische Methoden soll Abbildung 3 zeigen [1].

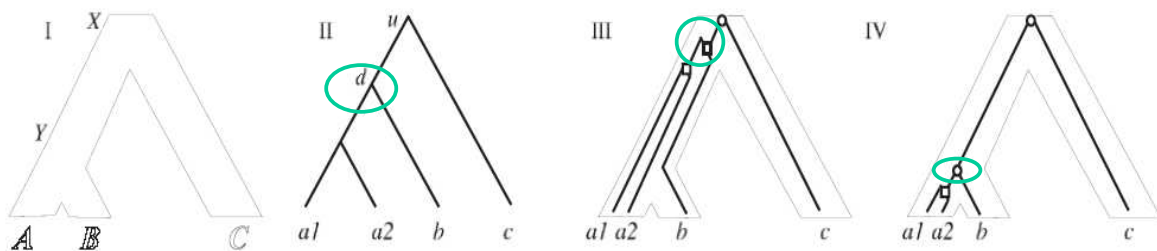


Abb.3: (I) Der Artenbaum. (II) Der Genbaum mit Gen u als Wurzel und dem Gen d als Vater von Gen b. (III) Ein mögliche Reconcilierung des Arten- und Genbaums durch den Arvestad-Algorithmus. Die Kreise repräsentieren die Speziationen und die Quadrate die Genduplikationen. Wie zu sehen ist, wird u als Speziation mit der Wurzel X des Artenbaums assoziiert. Spezies Y wird in dieser Reconcilierung kein Gen zugeordnet., während Gen a1 und a2 zum Genom der Spezies A, Gen b zu Spezies B und Gen c zu Spezies C gehören. Gen d entstand durch Genduplikation und gehört ebenfalls zum Genom der Spezies X. (IV) Die Reconcilierung eines Gen- und eines Artenbaums durch parsimonische Methoden benutzt eine minimale Anzahl an Genduplikationen. Dadurch wird hier Gen d fälschlicherweise als Speziation angegeben und dem Genom von Spezies Y zugeordnet.

All diese Probleme versuchen Arvestad et al. mit ihrem Algorithmus zu lösen.

Um zu gewährleisten, dass die zeitlichen Abstände in Gen- und Artenbaum gleich sind, lassen sie zum Beispiel mit Hilfe eines Gen-Evolutions-Modells, den Genbaum innerhalb des Artenbaums wachsen. Für die Abschätzung der Wahrscheinlichkeit einer Reconcilierung und des entsprechenden Genbaums nutzen sie die abgewandelte Form einer Markov-Chain-Monte-Carlo-Methode (MCMC).

2. KURZER ÜBERBLICK ÜBER DEN ALGORITHMUS

Der MCMC-Algorithmus von Arvestad et. al. hat im Wesentlichen drei Funktionen: Er kann zur Rekonstruktion eines Genbaums oder eines Artenbaums dienen und man kann mit ihm eine orthologische Analyse durchführen. Wenn nun der Genbaum einer Genfamilie, in Abhängigkeit eines Artenbaums, rekonstruiert werden soll, beginnt man zuerst damit, aus den gegebenen Gensequenzen - über das Gen-Sequenz-Evolutions-Modell - mögliche Reconcilierungen zu erstellen. Der nächste Schritt ist die Berechnung der Likelihood einer jeden Reconcilierung und ihres entsprechenden Genbaums. Diese Likelihood, die über rekursive Gleichungen errechnet wird, ist ein wichtiger Bestandteil des MCMC-Algorithmus, mit dem eine Wahrscheinlichkeitsverteilung der Genbäume und ihrer Reconcilierungen geschätzt werden kann.

Auch in der orthologischen Analyse spielt die Likelihood eine wichtige Rolle. Nach dem Sampeln der wahrscheinlicheren Genbäume aus der MCMC-Verteilung kann mit der Maximum-Likelihood-Berechnung gemäß der a posteriori-Verteilung die eine Reconcilierung und ihr zugehöriger Genbaum ermittelt werden. Zum Schluss kann man noch die Wahrscheinlichkeit dieses Genbaums berechnen, indem man über alle Reconcilierungen summiert.

3. DIE WICHTIGSTEN DEFINITIONEN

Zum besseren Verständnis werden nun erst mal die wichtigsten Definitionen und Schreibweisen vorgestellt. Jeder Baum T – sei es nun ein Genbaum G oder ein Artenbaum S – besitzt eine Menge von Knoten $V(T)$ und eine Menge von Kanten $A(T)$. Dabei ist $V(G)$ eine Menge von Genen, die mit Kleinbuchstaben wie z.B. u, v, w bezeichnet werden und $V(S)$ ist eine Menge von Arten, die durch Großbuchstaben wie X, Y, Z gekennzeichnet sind. Das Gleiche gilt für $L(T)$ – der Menge der Blätter eines Baumes T . Auch hier werden die Blätter des Genbaumes G durch Gene und die Blätter des Artenbaumes durch Arten repräsentiert. Da T ein gerichteter Baum ist, hat er auch eine Wurzel $r(T)$. Außerdem ist T ein binärer Baum, das heißt jeder innere Knoten hat zwei ausgehende Kanten. Der allgemeine Ausdruck für die Kinder eines inneren Knoten $u \in V(T)$ ist $c_1(u)$ und $c_2(u)$ – wobei mit $c_1(u)$ immer das linke Kind und mit $c_2(u)$ immer das rechte Kind gemeint ist. Der Ausdruck T_u bezeichnet einen Teilbaum von T mit Wurzel u , wobei $u \in V(T)$ ist. Ein Kantenteilbaum wird mit $T^{u,v}$ ausgedrückt, wobei $\langle u,v \rangle$ eine Kante der Menge $A(T)$ ist. Damit wird ein Teilbaum bezeichnet, der aus der Kante $\langle u,v \rangle$ und dem Teilbaum T_v besteht. Als Beispiel sei hier der Baum in Abbildung 4 gezeigt.

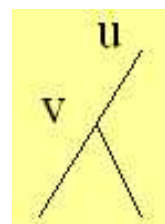


Abb.4: Der Kantenteilbaum $T^{u,v}$

Formal wird die Reconcilierung eines Artenbaums S und eines Genbaums G durch ein Paar von (γ, G') beschrieben, wobei G' ein Unterbaum von G ist und γ als Reconcilierungsfunktion wie folgt definiert ist:

$$\gamma : V(S) \rightarrow 2^{V(G')}$$

Diese Funktion bedeutet, dass jeder Art eine Menge von Genen zugeordnet wird. Zum Beispiel: $\gamma(X) = \{u\}$ und $\gamma(A) = \{a_1, a_2\}$ (siehe Abb.3). Eine andere Schreibweise, die hier ebenfalls verwendet wird, ist $u \in \gamma(X)$ - also: Gen u gehört zum Genom von X . Zu beachten ist dabei, dass G' nur ein Unterbaum von G ist. Das heißt, nach Anwendung des Gen-Evolutions-Modells, das später noch genauer erklärt wird, entsteht ein Genbaum G' , der noch Zwischenknoten mit nur einer Ausgangskante besitzt. Da G aber als binärer Baum definiert ist, werden diese Knoten entfernt und es entsteht der fertige Genbaum G (s. Abb.5).

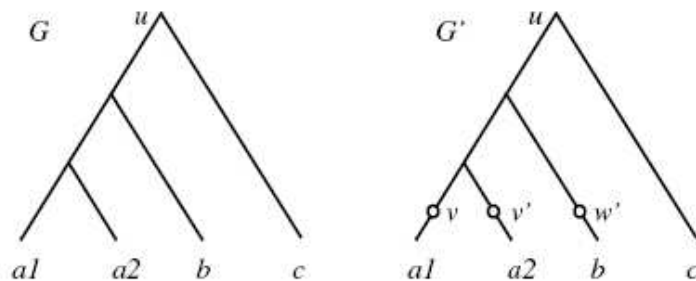


Abb.5: G' ist ein Unterbaum von G mit den Zwischenknoten v, v' und w' , die in der Reconcilierung zum Genom von Y gehören: $\gamma(Y) = \{v, v', w'\}$ (s. Abb.6).

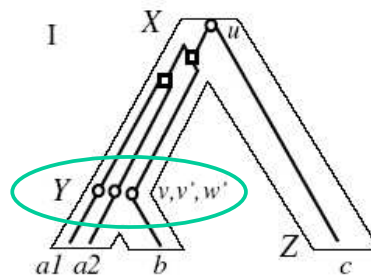


Abb.6: Reconcilierung eines Artenbaums mit einem Genbaum.

4. DAS GEN-(SEQUENZ)-EVOLUTIONS-MODELL

Eine Reconcilierung beschreibt, wie sich ein Genbaum in Bezug auf einen Artenbaum entwickelt. Das von Arvestad et. al. entwickelte probabilistische Gen-Evolutions-Modell lässt einen Genbaum innerhalb eines gegebenen Artenbaums S wachsen und induziert damit eine Reconcilierung. Dieses Modell beschreibt also, wie sich ein Genbaum G durch Speziationen, Genduplikationen und Genverluste über die Zeit entwickelt. Seine Entwicklung wird nur durch den gegebenen Artenbaum S beschränkt. Speziationen und Duplikationen führen dabei Knoten in G ein, während durch Genverluste Kanten und Knoten wieder entfernt werden. Der eigentliche Entwicklungsprozess von Genbaum G wird durch einen, in der phylogenetischen Analyse oft

verwendeten, Geburts-Todes-Prozess über den Kanten von S modelliert. Dabei entsprechen die Geburten den Genduplikationen mit einer konstanten Geburtsrate λ und die Tode entsprechen den Genverlusten mit einer konstanten Todesrate μ .

Entwickelt wurde dieser Prozess bereits 1948 von dem Mathematiker D.G. Kendall. 1992 betrachteten Nei et. al. molekulare Daten der Genfamilien von MHC- und Immunglobulinen [4]. Ihre phylogenetischen Analysen zeigten, dass das Entwicklungsmuster solcher Genfamilien mit dem Geburts-Todes-Modell übereinstimmt, indem neue Gene durch wiederholte Genduplikationen entstehen. Manche dieser Gene bleiben für lange Zeit im Genom erhalten, während andere nach kurzer Zeit wieder verschwinden oder durch Mutationen ihre Funktionsfähigkeit verlieren (Pseudogene). Nei et. al. stellten während ihrer Analysen ebenfalls fest, dass die Evolution von Genfamilien, deren Mitglieder nicht nur in einer Spezies, sondern in mehreren vorkommen, besonders durch den Geburts-Todes-Prozess charakterisiert wird, da die Mitglieder einer Genfamilie innerhalb einer Spezies nicht notwendigerweise auch näher miteinander verwandt sein müssen, als die Mitglieder einer Genfamilie, die sich über mehrere Arten erstreckt. Dies wäre bei dem Konzept der „gerichteten Evolution“ jedoch der Fall (Abb.7).

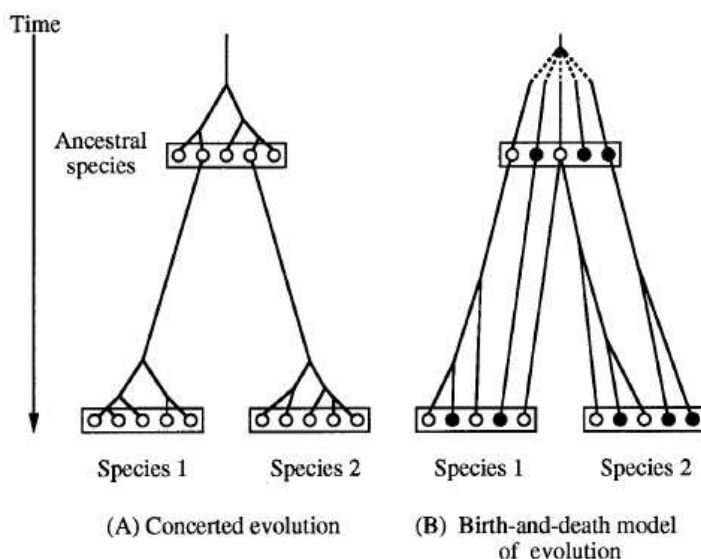
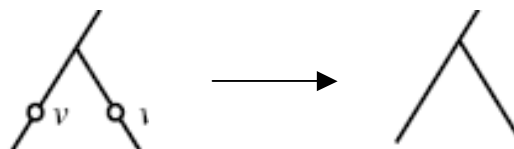


Abb.7: Zwei verschiedene Modelle der Evolution einer Genfamilie.
 (A) Nach dem Konzept der gerichteten Evolution.
 (B) Nach dem Konzept des Geburts-Todes-Modells. Die dunklen Kreise stellen die Genverluste dar.

Im Gen-Evolutions-Modell von Arvestad et. al. startet der Geburts-Todes-Prozess in der Wurzel des Artenbaums S und wandert abwärts bis zu seinen Blättern. Erreicht der Prozess das Ende einer Kante des Artenbaums, spaltet er den darauffolgenden Knoten in zwei identische Kopien, läuft rekursiv rechts und links an den Kanten der beiden neuen Knoten weiter und wiederholt den Vorgang solange bis er die Blätter des Artenbaums erreicht und stoppt. Nach dem Ende des rekursiven Prozesses wird der entstandene Genbaum gekürzt, indem Knoten ohne Nachfahren in den Blättern des Artenbaums (Genverluste) samt ihren eingehenden Kanten gelöscht werden. Knoten, die nur eine ausgehende Kante (also nur ein Kind) besitzen werden ebenfalls entfernt und ihre eingehende Kante mit der Ausgehenden verbunden (s. Abb.8 oder auch Abb.5: G' wird zu G).

Abb.8: Dieses Beispiel soll zeigen, wie Knoten mit nur einer ausgehenden Kante gelöscht werden und die ausgehenden Kanten mit den eingehenden Kanten verbunden werden.



Zum Schluss werden die Blätter des resultierenden Genbaums G entsprechend der Blätter des Artenbaums benannt, in die sie während des Prozesses hineinwuchsen. Das Gen-Evolutions-Modell generiert aber nicht nur einen Genbaum G , sondern benennt auch gleichzeitig seine inneren Knoten mit den Duplikations- und Speziationszeitpunkten τ , die später noch zur Berechnung der Likelihood der Genbäume benötigt werden.

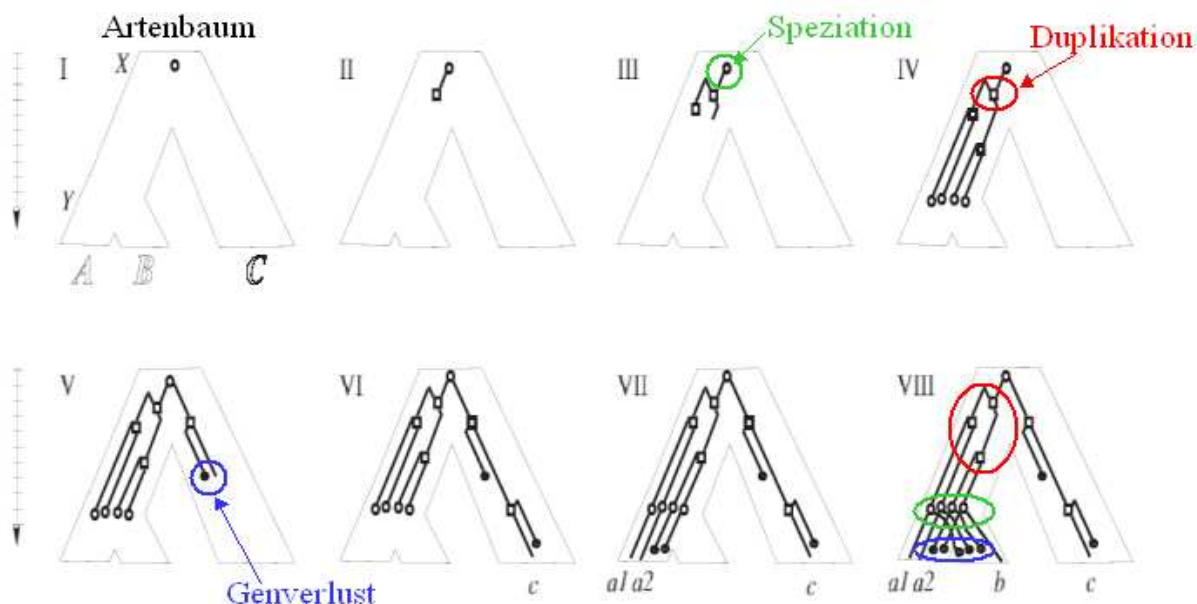


Abb.9: Ein Beispiel, wie sich ein Genbaum innerhalb eines Artenbaums entwickelt.

Wie bereits erwähnt, wurde das Gen-Evolutions-Modell zum Gen-Sequenz-Evolutions-Modell erweitert (s. Paper 2004). Dadurch wird nun auch die Evolution der Gensequenz miteinbezogen. Das heißt bei der Entstehung einer neuen Genlinie durch Duplikation oder Speziation, entwickelt sich die Gensequenz entsprechend eines Standard-Sequenz-Evolutions-Modells (hier: DNA-Substitutionsmodell von Jukes-Cantor und Molecular-Clock-Modell für die Genbaumkanten) bis diese Linie (wiederum durch Duplikation oder Speziation) in zwei neue Linien gespalten wird und zwei identische Kopien der Sequenz entstehen. Diese beiden Sequenzen entwickeln sich dann unabhängig voneinander weiter. Das Gen-Sequenz-Evolutions-Modell ist hierarchisch, d.h. es ist egal, ob zuerst der Genbaum nach dem Gen-Evolutions-Modell entwickelt und danach das Standard-Sequenz-Evolutions-Modell darauf angewendet wird, oder umgekehrt. Das Gen-Sequenz-Evolutionsmodell setzt sich somit aus dem Gen-Evolutions-Modell und dem Standard-Sequenz-Evolutions-Modell zusammen.

5. MCMC-SCHÄTZUNG DER POSTERIOR-VERTEILUNG VON GENBÄUMEN

Das Markov-Chain-Monte-Carlo-Verfahren ist eine Methode zur Schätzung der stationären Wahrscheinlichkeitsverteilung einer Markov-Kette. Der hier verwendete Algorithmus basiert auf dem Metropolis-Hastings-Algorithmus, der zu den MCMC-Methoden gehört und folgendermaßen funktioniert:

In einem Zustandsraum wird durch eine zufällige Übergangswahrscheinlichkeit (proposal distribution, hier: „nearest neighbor interchange branch swapping“) eine Markov-Kette aus Zuständen gebildet. Eine Akzeptanz-Wahrscheinlichkeit (der Metropolis-Hastings-Quotient, s. Abb. 10) entscheidet, ob und mit welcher Wahrscheinlichkeit der Übergang von einem Zustand in den Anderen akzeptiert wird. Dieser Quotient sorgt dafür, dass die Markov-Kette sich nach einer gewissen Zeit der stationären Zielverteilung π (hier: Die a posteriori-Verteilung der Reconcilierungen) nähert. Da man vorher nicht weiß, wie gut der Startpunkt der Markov-Kette gewählt ist, läßt man die Kette zunächst eine lange Zeit („burn in“) laufen, bis man annehmen kann, dass die Verteilung des Wertes, der ausgegeben wird, nahe genug an der stationären Zielverteilung - und damit an einem Gleichgewicht - ist. Dieser Wert wird mehrmals - jedes Mal nach einer konstanten Anzahl von Iterationen des Algorithmus - ausgegeben („gesampelt“). Die gesampelten Werte nähern die Verteilung π an, d.h. die wahrscheinlichen Zustände werden auch mit größerer Wahrscheinlichkeit gezogen.

$$\alpha(X, Y) = \min \left(1, \frac{\pi(Y)q(X|Y)}{\pi(X)q(Y|X)} \right)$$

Abb.10: Der Metropolis-Hastings-Quotient

$\alpha(X,Y)$ gibt die Wahrscheinlichkeit an, mit der der neue Zustand Y akzeptiert wird.
 $\pi(Y)$ ist die Wahrscheinlichkeit, daß Zustand Y der neue vorgeschlagene Zustand ist.
 $\pi(X)$ ist die Wahrscheinlichkeit, daß Zustand X der aktuelle Zustand ist.
 $Q(X|Y)$ ist die Wahrscheinlichkeit, mit der der Übergang von Zustand Y nach X stattfindet.
 $Q(Y|X)$ ist die Wahrscheinlichkeit, mit der der Übergang von Zustand X nach Y stattfindet.
 $\alpha(X,Y)$ kann höchstens den Wert 1 annehmen, auch wenn der Quotient einen größeren Wert hat.

Im von Arvestad et. al. adaptierten MCMC-Verfahren setzt sich ein Zustand der Markov-Kette aus einem Triple von Werten zusammen: (G, λ, μ) – also aus dem Genbaum G und den zugehörigen Geburts- und Todesraten λ und μ . Außerdem wird noch ein weiterer Parameter eingeführt: $F = \{q_1, \dots, q_n\}$ ist die Menge aller zu betrachtenden Gensequenzen einer Genfamilie, für die der Genbaum rekonstruiert werden soll. Dabei existiert für jede Art i eine Menge von Gensequenzen q_i in F, d.h. n entspricht der Anzahl der Arten, die Gene dieser Genfamilie besitzen. Die gesuchte stationäre Zielverteilung ist eine a posteriori Verteilung von Genbäumen – posterior, da F bereits gegeben ist. Sie wird über den Satz von Bayes (Prinzip der bedingten Wahrscheinlichkeit) berechnet und sieht wie folgt aus:

$$Pr[G, \lambda, \mu | F] = \frac{Pr[F, G | \lambda, \mu] Pr[\lambda, \mu]}{Pr[F]}$$

$\Pr [G, \lambda, \mu | F]$ ist also der Wert, der aus der Zielverteilung gesampelt wird. Er gibt die Wahrscheinlichkeit an, daß für einen Zustand (G, λ, μ) im Gen-Evolutions-Prozess G der Genbaum, λ die Geburtsrate und μ die Todesrate – bei gegebenen Gensequenzen F - war. Die Wahrscheinlichkeit $\Pr[F, G | \lambda, \mu]$ wird Likelihood genannt und kann mit einem kleinen Fehler approximiert werden. Ihre gegebenen λ - und μ -Werte sind angenommene a priori – also zufällige – Werte und werden erst zu a posteriori Werten, wenn sie zusammen mit dem zugehörigen Genbaum aus der Zielverteilung gesampelt wurden. Die Akzeptanz-Wahrscheinlichkeit mit der von einem aktuellen Zustand (G, λ, μ) in einen neuen, vorgeschlagenen Zustand (G', λ', μ') übergegangen wird, bildet sich aus dem Verhältnis von $\Pr [G, \lambda, \mu | F]$ zu $\Pr [G', \lambda', \mu' | F]$. Die Wahrscheinlichkeiten $\Pr[\lambda, \mu]$ (ebenfalls a priori-Werte) und $\Pr[F]$ bleiben dabei sowohl bei dem alten als auch bei dem neuen Zustand die Gleichen und lassen sich als Konstanten aus dem Term herauskürzen. Die Akzeptanz-Wahrscheinlichkeit sieht somit (entsprechend dem Metropolis-Hastings-Quotienten) folgendermaßen aus:

$$\alpha_{ij} = \min \left(1, \frac{\bar{Pr}[F, G | \bar{\lambda}, \bar{\mu}]}{Pr[F, G' | \lambda', \mu']} \right)$$

Damit ist offensichtlich, wie wichtig die Approximierung der Likelihood ist. Sie setzt sich aus drei Wahrscheinlichkeiten zusammen, die über τ und γ summiert werden:

$$\Pr[F, G | \lambda, \mu] = \sum_{\gamma, \tau} Pr[F | \tau, \gamma, G, \lambda, \mu] Pr[\tau | \gamma, G, \lambda, \mu] Pr[\gamma, G | \lambda, \mu],$$

τ ist – wie bereits unter 4. erwähnt – die Menge aller Speziations- und Duplikationszeitpunkte und wird durch den Gen-Evolutions-Prozess ermittelt, durch den wir auch die entsprechende Reconcilierung γ erhalten. Wenn dann alle Parameter bekannt sind, ist es möglich die Wahrscheinlichkeit $\Pr[F | \tau, \gamma, G, \lambda, \mu]$ zu errechnen. $\Pr[\gamma, G | \lambda, \mu]$ kann durch die, von Arvestad et. al. entwickelte, dynamische Programmierung gesampelt werden, die später in 8. noch erklärt wird.

6. ORTHOLOGISCHE ANALYSE

Um zu ermitteln, ob ein Genpaar a und b bei gegebenen Gensequenzen F ortholog ist, muß ebenfalls eine posterior Wahrscheinlichkeit berechnet werden, bei der wiederum die Likelihood eine wichtige Rolle spielt. Diese Wahrscheinlichkeit kann wie folgt ausgedrückt werden:

$$\Pr[a \text{ und } b \text{ sind ortholog} | F] = \sum_{G, \lambda, \mu} \frac{Pr[a \text{ and } b \text{ orthologs}, F, G | \lambda, \mu]}{Pr[F, G | \lambda, \mu]} Pr[G, \lambda, \mu | F]$$

Die Wahrscheinlichkeit $\Pr [G, \lambda, \mu | F]$ kennen wir noch aus 5. – sie ist der gesampelte Wert aus der MCMC-posterior Verteilung der Genbäume. Mit der Wahrscheinlichkeit,

$$\sum_{\gamma, \tau} Pr[F | \tau, \gamma, G, \lambda, \mu]$$

die ebenfalls noch aus der Likelihood-Berechnung unter 5. bekannt ist, wird dann für jede Reconcilierung separat berechnet, wo ein Genpaar ortholog und wo es paralog ist. Daraus kann dann die Wahrscheinlichkeit $\Pr[a \text{ und } b \text{ sind ortholog, } F, G | \lambda, \mu]$ für $\Pr[a \text{ und } b \text{ sind ortholog} | F]$ geschätzt werden.

7. MCMC-SCHÄTZUNG EINER POSTERIOR VERTEILUNG VON ARTENBÄUMEN

Der MCMC-Algorithmus von Arvestad et. al. ist auch zur Schätzung einer posterior Verteilung von Artenbäumen anwendbar. Der größte Unterschied, der hierbei beachtet werden muß, ist dass der Artenbaum hier natürlich nicht– wie bei der Rekonstruierung eines Genbaums – bereits gegeben ist, da man ja einen Artenbaum rekonstruieren will. Auch wird nicht eine einzelne Genfamilie betrachtet, sondern es werden gleich mehrere Genfamilien miteinbezogen. Die Eingabe ist demnach nicht F als eine Menge von einzelnen Gensequenzen, sondern $f = \{F_1, \dots, F_k\}$ als eine Menge von Genfamilien, wobei jedes F_i eine Menge von Sequenzdaten einer Genfamilie darstellt und aus einer Gensequenz pro Spezies besteht (wie bei 5.).

Die gesuchte Verteilung $\Pr[S | f]$ erhält man aus der posterior Wahrscheinlichkeit:

$$\Pr[S, G_1, \dots, G_k, \lambda, \mu | \mathcal{F}] = \frac{\Pr[\mathcal{F}, G_1, \dots, G_k | S, \lambda, \mu] \Pr[S, \lambda, \mu]}{\Pr[\mathcal{F}]}$$

Der Quotient wird wiederum über den Satz von Bayes gebildet. Für den Artenbaum S und die Geburts- und Todesraten λ und μ werden auch hier zur Berechnung zuerst a priori Werte angenommen. Aber, wie unter 5., muß die Likelihood extra berechnet werden. Die Likelihood sieht folgendermaßen aus:

$$\Pr[\mathcal{F}, G_1, \dots, G_k | S, \lambda, \mu] = \prod_{1 \leq i \leq k} \Pr[F_i, G_i | S, \lambda, \mu]$$

Der Index k entspricht der Anzahl der Genfamilien, d. h., für jede Genfamilie (wird über den Geburts-Todes-Prozess) bei gegebenen (a priori) S , λ - und μ -Werten ein Genbaum erstellt. Dabei kann mit jeder einzelnen Wahrscheinlichkeit $\Pr[F_i, G_i | S, \lambda, \mu]$ für $i = 1, \dots, k$ so verfahren werden, wie mit der Likelihood bei der posterior Verteilung von Genbäumen (Berechnung der Akzeptanzwahrscheinlichkeit etc.). Alle weiteren Berechnungen zur Rekonstruktion von Artenbäumen entsprechen denen, die auch für die Genbaum-Rekonstruktion verwendet werden.

8. BERECHNUNG DER LIKELIHOOD FÜR EINE RECONCILIERUNG

Zur Berechnung der Likelihood einer Reconcilierung und des entsprechenden Genbaums – bei gegebenem Artenbaum S und den λ - und μ -Parametern – ist es nötig, einen dynamisch programmierten Algorithmus zu verwenden, der ein Rechenproblem in kleine Teilprobleme zerlegt. Nach diesem Prinzip wird der Genbaum in sogenannte „sliced subtrees“ zerlegt. Ein „sliced subtree“ besteht aus dem Teil des Genbaums, dessen zugehörige Reconcilierung innerhalb einer Kante des Artenbaums liegt. Ein Beispiel für einen „sliced subtree“ wird in Abbildung 10 unter III gezeigt. $G_{u\gamma(y)}$ ist ein „sliced subtree“, für die Kante des Artenbaums

$\langle x,y \rangle \in A(S)$ und dem Gen $u \in \gamma(x)$, das zum Genom der Spezies X gehört. Er endet bei den Genen die zum Genom der Spezies Y gehören (v, v', w').

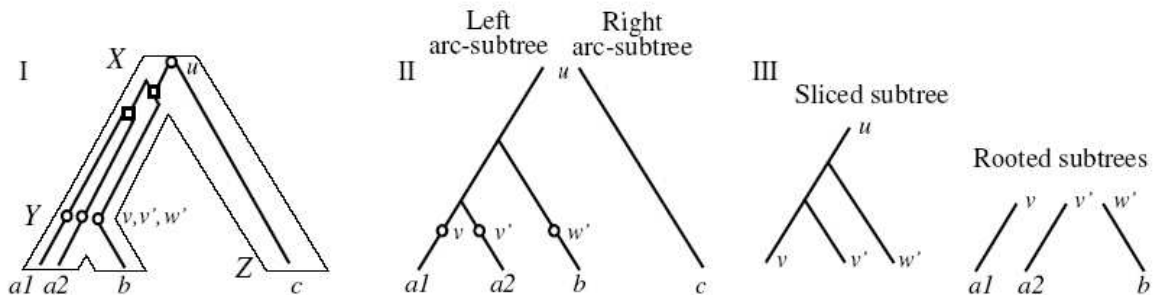


Abb.10: (I) Eine Reconciliation γ von G und S . (II) Links der Teilbaum $G_u^{x,y}$ und rechts der Teilbaum $G_u^{x,z}$ in Bezug auf (I). (III) Der „sliced subtree“ $G_{u\gamma(y)}$ und die gewurzelten Teilbäume $G_v, G_{v'}$ und $G_{w'}$, für den gleichen Gen- und Artenbaum und die gleiche Reconciliation wie in (I).

$G_u^{x,y}$ ist noch mal ein spezieller „sliced subtree“, bei dem von vorneherein festgelegt ist, dass er genau eine Kante $\langle x,y \rangle$ lang und bei u gewurzelt ist. Es gibt keine zusätzliche Spezies zwischen X und Y . Der Unterschied zu $G_{u\gamma(y)}$ ist der, dass per Definition auch die Gene, die der Spezies Y nachfolgen – also schon der nächsten Spezies angehören (hier: $a1, a2$ und b) - ebenfalls zu dem „sliced subtree“ $G_u^{x,y}$ gehören.

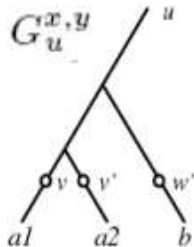


Abb.11: Beispiel eines „sliced subtrees“.

Die Größe der „sliced subtrees“ wird durch eine obere und eine untere Schranke begrenzt. Es sind komplizierte rekursive Gleichungen notwendig, um einen Genbaum in diese Teilbäume zu zerlegen, die hier aber nicht gezeigt werden. Da es aber exponentiell viele „sliced subtrees“ gibt – aufgrund der hohen Anzahl von mögliche Reconciliationen und der Größe eines Genbaums – genügt es für die Berechnung, eine kleinere Konfiguration der „sliced subtrees“ zu betrachten: den Vaterknoten, seine zwei Kinder und eine kleine Anzahl von gespeicherten Werten. Die folgenden rekursiven Gleichungen, auf denen die dynamische Programmierung basiert, verdeutlichen dieses Prinzip. Diese rekursiven Likelihoodberechnungen werden für jeden „sliced subtree“ einzeln angestellt. Zusammengesetzt ergeben sie dann die Likelihood einer einzelnen Reconciliation bzw. Genbaums. Die wichtigste rekursive Gleichung ist $e_v(\gamma, x, u)$, auf der alle anderen Gleichungen basieren. Sie ist als die Wahrscheinlichkeit definiert, dass G_u und γ_u sich

aus der Wurzel u des Genbaum-Subtrees und der entsprechenden Startspezies x in S_x entwickelt haben. Ähnlich ist auch $e_A(\gamma, y, u)$ definiert, nämlich als die Wahrscheinlichkeit mit der sich $G_u^{x,y}$ und $\gamma_u^{x,y}$ aus u und x in $S^{x,y}$ entwickelt haben – mit y endet die Kante in der Zielspezies. Daher kann man $e_v(\gamma, x, u)$ auch als Knotenwahrscheinlichkeit und $e_A(\gamma, y, u)$ als Kantenwahrscheinlichkeit bezeichnen. Hier nun die Rekursionsgleichungen:

$$e_v(\gamma, x, u) = \begin{cases} 1, & x \in L(S), u \in L(G) \\ e_A(\gamma_u^{x,y}, y, u) e_A(\gamma_u^{x,z}, z, u), & \text{otherwise} \end{cases}$$

$e_v(\gamma, x, u)$ bekommt den Wert 1 zugewiesen, wenn x und u die Blättern ihrer jeweiligen Bäume sind. Sind sie es nicht, wird e_A für die linke und rechte Kante von x berechnet (y und z sind Kinder von x).

Für die Berechnung von $e_A(\gamma, y, u)$ sind zwei Fälle zu beachten. Der 1. Fall trifft ein, wenn $\gamma_u(y)$ nicht leer ist, d.h. Gene der Spezies Y existieren und u daher kein Blatt sein kann. Der 2. Fall geht vom Gegenteil aus: $\gamma_u(y) = 0$, d.h. das Genom von Y ist leer und u ein Blatt. Für den 1. Fall sieht die rekursive Gleichung von $e_A(\gamma, y, u)$ folgendermaßen aus:

$$e_A(\gamma, y, u) = p_y(|\gamma(y)|) h(\gamma, y, u) \phi(\gamma, y, u) \prod_{v \in \gamma_u(y)} e_v(\gamma_v, y, v)$$

Die Rekursionsgleichung für den 2. Fall ist deutlich kürzer: $e_A(\gamma, y, u) = p_y(0)$

e_A setzt sich aus vier Wahrscheinlichkeiten zusammen, von denen drei wiederum rekursiv berechnet werden müssen. $p_y(l)$ ist lediglich ein diskreter Parameter, nämlich die Wahrscheinlichkeit, dass der Genteilbaum der Kante $\langle x, y \rangle$, der vom Geburts-Todes-Prozess generiert wurde, genau l viele Blätter hat. Ist y kein Blatt, sondern ein Knoten mit Kindern wie bei x , muß für $v \in \gamma(y)$ wiederum rekursiv die Knotenwahrscheinlichkeit $e_v(\gamma_v, y, v)$ berechnet werden.

Der rekursive Ausdruck $h(\gamma, y, u)$ berechnet die Wahrscheinlichkeit der Baumstruktur (Topologie). Während des Geburts-Todes-Prozesses werden sogenannte „fully labeled trees“, also Bäume mit Bezeichnungen an Knoten und Blättern, generiert. Diese „fully labeled trees“ sind alle gleichwahrscheinlich, da es von jedem Baum nur einen gibt, d.h. jeder Baum besitzt eine bestimmte „history“ (eine zeitliche Abfolge der inneren Knoten) und eine bestimmte Baumform, die durch die Blattbenennung festgelegt wird. Durch das Entfernen der Labels an den inneren Knoten (während des Zerlegens in die „sliced subtrees“) geht die „history“ der Bäume verloren. Danach werden auch die Labels der Blätter gelöscht und übrig bleiben die sogenannten „unlabeled trees“, die nur noch aus einer der beiden möglichen Baumstrukturen (s. Abb.12) bestehen. Dabei ist die eine Topologie wahrscheinlicher als die Andere und diese Wahrscheinlichkeit wird durch $h(\gamma, y, u)$ ausgedrückt.

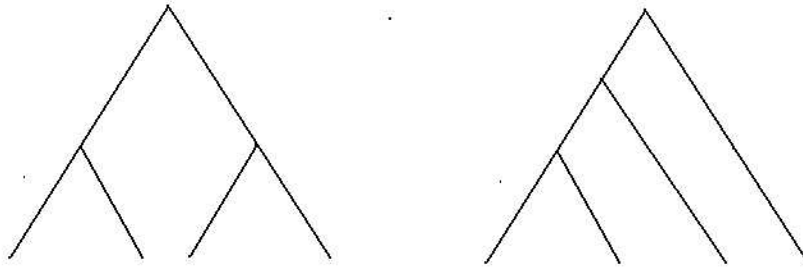


Abb.12: Die rechte Baumstruktur hat mehr Permutationsmöglichkeiten und ist daher wahrscheinlicher als die linke Baumstruktur.

Die Rekursionsgleichung für $h(\gamma, y, u)$ ist wie folgt aufgebaut:

$$h(\gamma, y, u) = \begin{cases} 1 & \text{if } u \in \gamma(y) \\ \kappa h(\gamma, y, c_1(u))h(\gamma, y, c_2(u)) & \text{otherwise,} \end{cases}$$

$$\text{where } \kappa = \frac{2^{\delta(G_{u, \gamma(y)})}}{|L(G_{u, \gamma(y)})| - 1}.$$

Auch hier wird $h(\gamma, y, u)$ der Wert 1 zugewiesen, wenn u bereits ein Blatt ist. Wenn dies nicht der Fall ist, wird für das linke und rechte Kind von u die rekursive Gleichung $h(\gamma, y, u)$ erneut aufgerufen und die Werte miteinander multipliziert. Wichtig ist dabei der Parameter κ , der sich aus einem δ und der Anzahl der Blätter des „sliced subtrees“ - 1 zusammensetzt. Das δ ist der eigentliche Parameter, der die Topologie eines „sliced subtrees“ in die Gleichung miteinfließen lässt. Ihm wird der Wert 1 zugewiesen, wenn die Struktur des Teilbaums des zu betrachtenden Knotens der rechten Struktur in Abb.12 entspricht. Ansonsten nimmt δ den Wert 0 an, für die linke, ausbalanciertere Struktur.

Der letzte rekursive Ausdruck $\phi(\gamma, y, u)$ errechnet die Anzahl der isomorphen (äquivalenten) Reconcilierungen und entsprechenden Genbäume. Die Rekursionsformel ist ähnlich aufgebaut wie bei der Berechnung von $h(\gamma, y, u)$. Auch hier existiert ein δ , dass die Werte 1 oder 0 annehmen kann. Allerdings bekommt δ nur dann den Wert 1 zugewiesen, wenn die Reconcilierungen einer isomorphen Klasse am Anfang und Ende gleich sind, jedoch nicht komplett übereinstimmen. Die Berechnung des $\phi(\gamma, y, u)$ selbst, erfolgt dann wieder rekursiv von den Blättern bis zur Wurzel wie bei $h(\gamma, y, u)$.

9. BERECHNUNG DER MAXIMUM-LIKELIHOOD

Die Berechnung der Maximum-Likelihood soll die Likelihood der wahrscheinlichsten Reconcilierung liefern und ist als das Maximum von $e_V(\gamma, r(S), r(G))$ definiert. Sie basiert auf den gleichen Rekursionsgleichungen, die bereits in 8. für die Likelihood-Berechnung einer einzelnen Reconcilierung vorgestellt wurden. In der dynamischen Programmierung der Maximum-Likelihood werden für jeden Knoten $y \in V(S)$ und $u \in V(G)$ die drei höchsten Werte von e_A und e_V (von insgesamt drei verschiedenen Reconcilierungen aus der MCMC-Verteilung) gespeichert.

Die Rekursionsgleichungen von e_A und e_V sind mit denen aus 8. fast identisch, allerdings wurden sie ein wenig umgeformt. Zum Einen wurden die beiden δ 's aus den Gleichungen $h(\gamma, y, u)$ und $\phi(\gamma, y, u)$ zu einer Variablen ξ zusammengefasst. Dieses ξ findet in der Rekursionsgleichung f_A

Verwendung, die bis auf den Wert $p_y(|\gamma(y)|)$ aus den gleichen Variablen besteht wie e_A . Um e_A selbst zu berechnen, wird f_A separat mit $p_y(|\gamma(y)|)$ multipliziert, da $p_y(|\gamma(y)|)$ ein diskreter Parameter ist und keiner rekursiven Berechnung bedarf. So werden also rekursiv für jeden Knoten in einer Reconcilierung die e_A - und e_V -Werte berechnet (von den Blättern aufwärts bis zur Wurzel) und jeweils die drei höchsten Werte der wahrscheinlichsten Reconcilierungen gespeichert. Die Speicherung der Werte erfolgt in aufsteigender Reihenfolge, d.h. der größte Wert wird unter 1. gespeichert, der zweitgrößte Wert unter 2. und der Drittgrößte unter 3. So gelangt das Maximum von e_V schließlich bis in die Wurzel des Genbaums, dessen zugehörige Reconcilierung die Wahrscheinlichste ist.

10. BERECHNUNG DER WAHRSCHEINLICHKEIT EINES GENBAUMS

Nun, da man die wahrscheinlichste Reconcilierung herausgefunden hat, kann man die exakte Wahrscheinlichkeit für den entsprechenden Genbaum - bei gegebenem Artenbaum - berechnen. Dies geschieht, indem man über Γ - der Menge aller Reconcilierungen - summiert:

$$\Pr(G|S) = \sum_{\gamma \in \Gamma} \Pr(G, \gamma|S)$$

Natürlich erfolgt die eigentliche Berechnung auch hier wieder über die einzelnen „sliced subtrees“ - mit Rekursionsgleichungen die analog zu den e_A - und e_V -Gleichungen sind.

$$\Pr(G_u^{x,y}|S^{x,y}) = \sum_{\gamma \in \Gamma_u^{x,y}} e_B(\gamma, y, u)$$

Die einzige Veränderung ist die, dass e_A zu e_B umbenannt wurde, da der ϕ -Faktor aus e_A entfernt wurde. Der ϕ -Wert - also die Anzahl der äquivalenten Reconcilierungen - ist nicht mehr nötig, da ja bereits über alle Reconcilierungen summiert wird. e_V heißt nun e_W und ruft entsprechend e_B rekursiv auf.

11. DISKUSSION

Arvestad et. al. implementierten ihren dynamisch programmierten Algorithmus und testeten ihn bereits erfolgreich auf sowohl künstlichen wie auch realen biologischen Daten. Zum Beispiel konnten sie bei einem biogeographischen Problem eine Hypothese bestätigen, die von parsimonischen Methoden verworfen worden war. Auch bei dem Genbaum der MHC-Genfamilie konnte durch ihren Algorithmus bewiesen werden, dass parsimonische Methoden an einer Stelle des Genbaums ein Genpaar fälschlicherweise als ortholog erkannt hatten. Der von Arvestad et. al. entwickelte Algorithmus hat daher großes Potential in Zukunft zum Standardwerkzeug der phylogenetischen Analyse zu werden und die parsimonischen Methoden zu verdrängen. Allerdings sind die parsimonischen Methoden weniger komplex und ihre Durchführung ist deutlich schneller möglich als es bei der neuen, hier vorgestellten, Methode der Fall ist.

12. LITERATURREFERENZEN

- [1] L.Arvestad et al. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics/ISMB`03*, 19, i7-i15, 2003.
- [2] L.Arvestad et al. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *RECOMB`04*, March 27-31, 326-335, 2004.
- [3] J.Felsenstein. *Inferring Phylogenies*. Sinauer, 2004, Chapter 29, pp 509-515.
- [4] M.Nei et al. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl. Acad. Sci. USA* Vol. 94, pp. 7799–7806, July 1997, Colloquium Paper.
- [5] Arthur M.Lesk. *Bioinformatik – Eine Einführung*. Spektrum, 2002, Kapitel 4, pp198-209.
- [6] D.Metzler. *Algorithmen und Modelle der Bioinformatik*. Skript zur Vorlesung vom WS 2003/2004.
- [7] Fred R. Opperdoes. *Methods for the inference of protein phylogeny*. <http://www.icp.ucl.ac.be/%7Eopperd/private/phylogeny.html> [Letztes Update: 25.09.1997].
- [8] Wikipedia – Die freie Enzyklopädie. <http://de.wikipedia.org/wiki/Bayes-Theorem> [14.07.05].