Bioinformatik I, WS00/01
Übungsblatt 8

Besprechung am 18.01.2001

1 Evolution als Markovprozess

Die Substitution von Basen in DNA im Verlauf der Zeit kann durch kontinuierliche Markovprozesse modelliert werden. $q(t)$ bezeichne die Wahscheinlichkeit, daß zum Zeitpunkt t zwei von der gleichen Base abstammende Basen gleich sind. Für das Jukes/Cantor Evolutionsmodell soll der kontinuierliche Markovprozess durch eine diskrete Markovkette angenähert werden und so die Wahrscheinlichkeit $q(t)$ abhängig von der Zeit t approximiert werden.

Im Jukes/Cantor Evolutionsmodell wird von einer für alle Nukleotide gleichen Substitutionsrate α ausgegangen. Man erhält die folgende Ratenmatrix

$$\Lambda = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}.$$

Analytisch kann die Wahrscheinlichkeit $q(t)$ bestimmt werden als

$$q(t) = \frac{1}{4} \cdot (1 + 3 \cdot e^{-8 \cdot \lambda \cdot t}).$$

Ein kontinuierlicher Markovprozess kann als Grenzwert diskreter Markovketten interpretiert werden. Eine diskrete Markovkette die einen Markovprozess annähert erhält man, indem man die kontinuierliche Zeit in kleine diskrete Schritte mit Zeit Δt zerlegt. Für kleine Δt kann man die Übergangsmatrix $P(\Delta t)$ aus den Substitutionsraten annähern, nämlich

$$P(\Delta t) \approx I + \Delta t \cdot \Lambda.$$

Die Aufgabe ist nun ein Programm zu schreiben, das zu einem gegebenen kontinuierlichen Markovprozess für das Jukes/Cantor Modell, d.h. gegebenem α, die entsprechende diskrete Markovkette für ein gegebenes Zeitintervall Δt erzeugt und Approximationen der Wahrscheinlichkeiten $q(t)$ berechnet. Aus einer Transitionsmatrix $P(2t)$, die man durch Potenzierung von $P(\Delta t)$ bekommt, erhält man eine Approximation $q(t) \approx p_{i,i}(2 \cdot t)$ für beliebiges $i \in \{1, 2, 3, 4\}$.[1]

Die approximierten Werte für $q(t)$ können dann mit der analytisch bestimmten Funktion $q(t)$ verglichen werden. Wie klein muss das Zeitintervall Δt

[1] Diese Beziehung $q(t) = p_{i,i}(2 \cdot t)$ ergibt sich aus folgender Überlegung. Eine Base X

gewählt werden, um eine gute Annäherung zu erreichen? Nach welcher Zeit ist $q(t)$ praktisch konstant?

Testen Sie Mutationsraten $\alpha = 10^{-6}$ und $\alpha = 10^{-9}$ und vergleichen Sie unterschiedliche Zeitintervalle Δt.

Hinweis zum Plotten von Funktionen Unter Linux können Funktionen und Datenreihen z.B. mit dem Programm gnuplot gezeichnet werden. Um die vom Programm erzeugte Funktion q zu zeichnen sollte das Programm nur die t und $q(t)$ Werte in je einer Zeile ausgeben, d.h. z.B.

```
1 0.999
2 0.998
3 0.996
usw.
```

Zur Anzeige mit gnuplot erzeugt man eine Datei mit dieser Ausgabe, startet dann gnuplot und gibt ein

```
plot "daten" with lines
```

Zum Vergleich mit der analytisch bestimmten Funktion q gebe z.B. ein

```
plot "daten" with lines, 0.25*(1+3*exp(-8*0.000001*x))
```

für Rate $\alpha = 0.000001$.

2 Felsensteins ML Methode

a) Bestimme nach Felsensteins Maximum Likelihood Methode die Distanz zweier Sequenzen (d.h. die Topologie des Baums ist gegeben), z.B.

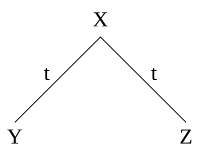
```
ATGCTA
ACGGTT
```

für gegebene stationäre Verteilung $\pi_1, \ldots, \pi_4$ z.B. mit Hilfe einer Tabellenkalkulation.

b) Schreibe ein Programm für die Berechnung einer solchen Distanz für zwei gegebene Sequenzen und gegebene π_i.

---

evolviere in Zeit t zu den Basen Y und Z.



Die Wahrscheinlichkeit, daß die Base Y gleich Z ist, ist gleich der Wahrscheinlichkeit, daß die Base Y in Zeit $2 \cdot t$ zu einer gleichen Base evolviert (Umkehrbarkeit der Evolution). Über alle (gleichverteilten) möglichen Basen Y summiert erhält man

$$q(t) = \sum_{i=1}^{4} \frac{1}{4} \cdot p_{i,i}(2 \cdot t).$$

Da alle $p_{i,i}(2 \cdot t)$ in unserem Modell für alle i gleich sind, ergibt sich die obige Beziehung.