7 Markovketten und DNA-Sequenzen

Hendrik Flasche
24.Mai 2012

7.1 Diskrete Markovketten

Unter einer (zeitlich-)diskreten Markovkette versteht man folgende

Definition. *(Diskrete Markovkette)*
$X = (X_1, X_2, \dots)$ sei eine Folge von Zufallsvariablen, die Werte einer endlichen Menge $\mathcal{A} = (a_1, \dots, a_n)$ annehmen können. X heißt **Markovkette**, wenn X die elementare Markoveigenschaft besitzt, d.h. für alle $i = 1, \dots, n$ und alle Zeitpunkte $t \in \mathbb{N}$ gilt

$$\mathbb{P}(X_t = a_i \mid X_0 = x_0, \dots, X_{t-1} = x_{t-1}) = \mathbb{P}(X_t = a_i \mid X_{t-1} = x_{t-1}).$$

Die hier stets endliche Menge $\mathcal{A}$ wollen wir **Zustandsraum** nennen.

Die Folge X besteht also nicht aus stochastisch unabhängigen Zufallsvariablen, sondern die Verteilung von X_t hängt stets von der vorherigen Realisierung $X_{t-1} = x_{t-1}$ ab und zwar ausschließlich und für alle Zeitpunkte $t = 1, 2, \dots$. Schreibe kurz

$$p_{ij}(t) := \mathbb{P}(X_t = a_i \mid X_{t-1} = a_j). \tag{1}$$

Für die Modellierung von DNA-Sequenzen sind nur Markovketten mit bestimmten Eigenschaften zweckmäßig. Zur Beschreibung dieser Eigenschaften benötigt man etwas Vokabular:

Definition. *(homogen)*
Eine Markovkette X heißt **homogen**, falls die Übergangswahrscheinlichkeiten in (1) nicht von der Zeit t abhängen. Man kann dann setzen

$$p_{ij}(t) = p_{ij} \quad \text{für alle } t \in \mathbb{N} \text{ und } i, j = 1, \dots, n.$$

Für eine homogene Markovkette lässt sich eine $n \times n$-Matrix angeben, die die konstanten Übergangswahrscheinlichkeiten von einem Zustand a_i in einen Zustand a_j enthält.

$$\mathbf{P} = \begin{pmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & & \vdots \\ p_{n1} & \cdots & p_{nn} \end{pmatrix}$$

Da nach Definition die i-te Zeile der Matrix $\mathbf{P}$ die jeweiligen Übergangswahrscheinlichkeiten *von* einem Zustand a_i beschreibt, gilt $\sum_j p_{ij} = 1$ für alle $i = 1, \dots, n$. Für eine vollständige Beschreibung einer diskreten Markovkette fehlt nun noch eine Anfangsverteilung $\mu^{(0)} = (\mu_1^{(0)}, \dots, \mu_n^{(0)})$ mit $\sum_{i=1}^n \mu_i^{(0)} = 1$. Hier wird festgelegt, mit welchen Wahrscheinlichkeiten die Markovkette jeweils startet. Für einen Zeitpunkt $t \in \mathbb{N}$ wollen wir die aktuelle Verteilung auf $\mathcal{A}$ mit $\mu^{(t)}$ be-

zeichnen. Damit ist gemeint

$$\mu_i^{(t)} = \mathbb{P}(X_t = a_i) \quad \text{für alle } t \in \mathbb{N}.$$

Satz 7.1 *Für eine Markovkette* $(X_1, X_2, \dots)$ *mit Übergangsmatrix* $\mathbf{P}$ *und Anfangsverteilung* $\mu^{(0)}$ *gilt*

$$\mu^{(t)} = \mu^{(0)} \cdot \mathbf{P}^t.$$

Beweis. Übung. ∎

Bemerkung. Eine homogene, diskrete Markovkette ist vollständig festgelegt durch die Angabe einer Übergangsmatrix $\mathbf{P}$ und einer Anfangsverteilung $\mu^{(0)}$.

Für die Wahrscheinlichkeit, dass auf einen Zustand a_i in m-Schritten der Zustand a_j folgt schreibt man kurz

$$p_{ij}^{(m)} := \mathbb{P}(X_{t+m} = a_j \mid X_t = a_i).$$

Man kann sich schnell überlegen, dass sich diese Wahrscheinlichkeiten mit der m-ten Potenz der Matrix $\mathbf{P}$ angeben lassen (ohne Beweis):

$$p_{ij}^{(m)} = (\mathbf{P}^m)_{ij}. \tag{2}$$

Definition. *(Periode eines Zustandes)*
Sei $a_i \in \mathcal{A}$ ein Zustand. Unter der **Periode** ω eines Zustandes versteht man

$$\omega(i) := \text{ggT}\,\{k \in \mathbb{N} \mid p_{ii}^{(k)} > 0\}.$$

Definition. *(aperiodisch, irreduzibel und ergodisch)*
Sei X eine homogene Markovkette mit Zustandsraum $\mathcal{A} = \{a_1, \dots, a_n\}$.

(a) X heißt **aperiodisch**, falls gilt

$$\omega(i) = 1 \quad \text{für alle } i = 1, \dots, n.$$

(b) X heißt **irreduzibel**, falls für alle $i, j \in \{1, \dots, n\}$ ein $k \in \mathbb{N}$ existiert mit

$$p_{ij}^{(k)} > 0.$$

(c) Eine aperiodische und irreduzible Markovkette heißt **ergodisch**.

Irreduzibilität bedeutet in Worten nichts anderes, als dass jeder Zustand von jedem Zustand aus erreicht werden kann, unerheblich, wieviele Schritte dafür benötigt werden.

Für das DNA-Alphabet $\mathcal{A} = \{A, C, G, T\}$ zeigt Abb. 1 eine graphische Darstellung einer dazugehörigen Markovkette. Die Pfeile repräsentieren den Übergang zwischen zwei Zuständen und sind mit den dementsprechenden Wahrscheinlichkeiten für einen solchen Übergang beschriftet (exemplarisch). Ein Pfeil zwischen zwei Zuständen symboliert ebenfalls, dass die entsprechende

Übergangswahrscheinlichkeit nicht verschwindet. Gilt $p_{ij} > 0$ für alle $i, j = 1, \ldots, n$, so verfügt die Markovkette über vollständige Konnektivität. Aus jedem Zustand kann also jeder Zustand direkt in einem Schritt folgen. Abb. 1 zeigt eine Markovkette mit vollständiger Konnektivität. Eine solche ist trivialerweise stets aperiodisch und irreduzibel.

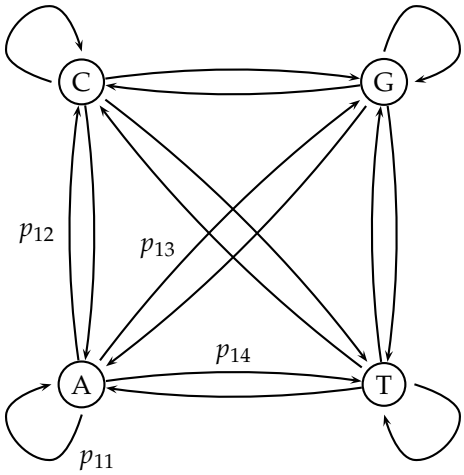

Abbildung 1: Graph einer Markovkette mit vollständiger Konnektivität

Im Folgenden gehen wir immer von einer homogenen, aperiodischen und irreduziblen Markovkette aus. Das besondere einer solchen Markovkette ist, dass für diese stets eine eindeutige Grenzverteilung existiert, was nun gezeigt werden soll. Der Beweis benutzt die Methode des **„coupling"**, welche eine wichtige Beweistechnik in der Wahrscheinlichkeitstheorie darstellt.

Definition. *(Stationare Verteilung)*
Ein Zeilenvektor $\pi = (\pi_1, \ldots, \pi_n)$ heißt **stationäre Verteilung** einer Markovkette mit Übergangsmatrix $\mathbf{P}$ und Zustandsraum $\mathcal{A} = \{a_1, \ldots, a_n\}$, wenn gilt

(i) π beschreibt eine Verteilung auf $\mathcal{A}$, also $\pi_i \geq 0$ für alle $i = 1, \ldots, n$ und $\sum_{i=1}^{n} \pi_i = 1$.

(ii) $\pi = \pi \cdot \mathbf{P}$.

Für den Beweis benötigen wir einige Lemmata, die aus Zeitgründen nicht bewiesen werden, sich aber intuitiv einsehen lassen:

Lemma 7.2 *Sei X eine irreduzible und aperiodische Markovkette mit Zustandsraum $\mathcal{A} = \{a_1, \ldots, a_n\}$ und Übergangsmatrix $\mathbf{P}$. Dann existiert ein $N < \infty$ derart, dass*

$$p_{ij}^{(n)} > 0 \quad \textit{für alle } i, j \in \{1, \ldots, m\} \textit{ und alle } n \geq N.$$

Definition. *(Eintrittszeit)*
Sei X eine Markovkette mit Zustandsraum $\mathcal{A} = \{a_1, \ldots, a_n\}$ und Übergangsmatrix $\mathbf{P}$. Startet diese in a_i, so nennen wir die Zufallsvariable

$$T_{i,j} := \min\{k \geq 1 : X_k = a_j\}$$

die **Eintrittszeit (hitting time)** für a_j. Weiterhin definieren wir

$$\tau_{i,j} = \mathbb{E}[T_{i,j}]$$

als die **durchschnittliche Eintrittszeit (mean hitting time)**. $\tau_{i,i}$ heißt die **durchschnittliche Rückkehrzeit (mean return time)** für den Zustand a_i. Trifft die Markovkette nach einem Start bei a_i niemals auf a_j, so setze $T_{i,j} = \infty$.

Weiterhin benötigen wir

Lemma 7.3 *Sei X eine aperiodische und irreduzible Markovkette mit Werten in $\mathcal{A} = \{a_1, \ldots, a_n\}$ und Übergangsmatrix $\mathbf{P}$. Dann gilt für alle $i, j = 1, \ldots, n$, dass*

$$\mathbb{P}(T_{i,j} < \infty) = 1,$$

und weiter auch, dass der Erwartungswert endlich ist

$$\tau_{i,j} = \mathbb{E}[T_{i,j}] < \infty.$$

Mit diesen Werkzeugen können wir beweisen:

Satz 7.4 ***(Existenz einer stationären Verteilung)***
Für jede irreduzible und aperiodische Markovkette existiert mindestens eine stationäre Verteilung.

Beweis. Sei $\mathcal{A} = \{a_1, \ldots, a_n\}$ Zustandsraum und $\mathbf{P}$ Übergangsmatrix. Angenommen die Markovkette beginnt in a_1 dann definiere

$$\rho_i := \sum_{t=0}^{\infty} \mathbb{P}(X_t = a_i, T_{1,1} > t).$$

ρ_i gibt für alle $i = 1, \ldots, n$ an, wie oft zu erwarten ist, dass die Markovkette den Zustand i in der Zeit $T_{1,1} - 1$ einnimmt, also bevor sie zum Anfangszustand a_1 zurückkehrt. Wir wissen mit Lemma 7.3, dass $\tau_{1,1} < \infty$ und man sieht direkt, dass $\rho_i < \tau_{1,1}$, also ist ρ_i für alle i ebenfalls endlich. Wir definieren

$$\pi = (\pi_1, \ldots, \pi_n) = \left(\frac{\rho_1}{\tau_{1,1}}, \ldots, \frac{\rho_n}{\tau_{1,1}} \right)$$

und behaupten, dass dies die gesuchte stationäre Verteilung ist. Warum gerade diese Definiton?

Angenommen, es gibt eine stationäre Verteilung π und wir starten mit dieser. Dann gilt $\mathbb{P}(X_t = a_i) = \pi_i$ für alle $t \in \mathbb{N}$ und alle $i = 1, \ldots, n$, d.h. zwischen zwei a_1 kommt ein Zustand a_i mit $i \neq 1$ im Schnitt mit dieser Häufigkeit vor.

Also sind die erwarteten relativen Häufigkeiten im Vektor π ein guter Kandidat für die stationäre Verteilung.

Natürlich muss dies noch verifiziert werden. Nach (ii) der Definition muss gelten, dass

$$\sum_{i=1}^{n} \pi_i p_{ij} = \pi_j$$

und nach (i) muss π auch wirklich eine Wahrscheinlichkeitsverteilung auf $\mathcal{A}$ sein. Auf beide Verifikationen (vgl. Häggström, Kap. 5) soll hier aus Zeitgründen verzichtet werden. ∎

Für die Formulierung des tatsächlichen Konvergenzsatzes benötigen wir noch eine Metrik auf dem Raum aller Verteilungen auf $\mathcal{A}$, die hier sehr nützlich ist.

Definition. *(total variation distance)*
Seien $\varphi = (\varphi_1, \dots, \varphi_n)$ und $\lambda = (\lambda_1, \dots, \lambda_n)$ zwei Wahrscheinlichkeitsverteilungen auf $\mathcal{A} = (a_1, \dots, a_n)$, dann definiere eine Metrik durch

$$\mathrm{d_{TV}}(\varphi, \lambda) := \frac{1}{2} \sum_{i=1}^{n} |\varphi_i - \lambda_i|.$$

Falls eine Folge $\nu_1, \nu_2, \dots$ bzgl. dieser Metrik gegen ν konvergiert, also falls gilt

$$\lim_{n \to \infty} \mathrm{d_{TV}}(\nu_n, \nu) = 0$$

so schreiben wir

$$\nu_n \xrightarrow{\mathrm{TV}} \nu.$$

Satz 7.5 ***(Konvergenzsatz für Markovketten)***
Sei X eine homogene, irreduzible und aperiodische Markovkette X mit Zustandsraum $\mathcal{A}$, festgelegt durch $(\mathbf{P}, \mu^{(0)})$. Dann existiert eine eindeutige stationäre Grenzverteilung π. Diese ist unabhängig von der Anfangsverteilung $\mu^{(0)}$. Sei $\mu^{(t)}$ für $t \in \mathbb{N}$ die Verteilung auf $\mathcal{A}$ für die t-te Stelle der Markovkette, dann gilt

$$\mu^{(t)} \xrightarrow{TV} \pi.$$

Beweis. Die Beweisidee benutzt das **coupling**-Argument. Dies funktioniert folgendermaßen: Alle im Beweis konstruierten Markovketten seien nach Annahme stets irreduzible und aperiodisch. Sei X eine mit Anfangsverteilung $\mu^{(0)}$ und Übergangsmatrix $\mathbf{P}$. Sei π eine für $X = (X_0, X_1, \dots)$ stationäre Verteilung (existiert nach Satz 7.4). Sei weiterhin $Y = (Y_0, Y_1, \dots)$ ebenfalls eine Markovkette mit gleicher Übergangsmatrix $\mathbf{P}$ und Anfangsverteilung π. Es ist möglich, X und Y unabhängig zu konstruieren (kann über den Hastings-Algorithmus im folgenden Kapitel erläutert werden). Definiere dann eine Zufallsvariable T, die angibt, wann sich X und Y das erste Mal „treffen“, also

$$T := \min\{k : X_k = Y_k\},$$

wobei $T = \infty$, falls sie sich niemals treffen. Wir möchten zeigen, dass sich X und Y mit Wahrscheinlichkeit 1 irgendwann in endlicher Zeit treffen. Lemma 7.2 liefert uns ein $N < \infty$ mit

$$p_{ij}^{(N)} > 0 \quad \text{für alle } i, j \in \{1, \dots, n\}.$$

Setze

$$\alpha := \min\{p_{i,j}^{(N)} : i \in \{1, \dots, n\}\} > 0,$$

dann folgt

$$\begin{aligned}
\mathbb{P}(T \leq N) &\geq \mathbb{P}(X_N = Y_N) \\
&\geq \mathbb{P}(X_N = a_1, Y_N = a_1) \\
&= \mathbb{P}(X_N = a_1)\mathbb{P}(Y_N = a_1) \\
&= \left(\sum_{i=1}^{n} \mathbb{P}(X_0 = a_i, X_N = a_1)\right)\left(\sum_{i=1}^{n} \mathbb{P}(Y_0 = a_i, Y_N = a_1)\right) \\
&= \left(\sum_{i=1}^{n} \mathbb{P}(X_0 = a_i)\mathbb{P}(X_N = s_1 \mid X_0 = a_i)\right) \\
&\cdot \left(\sum_{i=1}^{n} \mathbb{P}(Y_0 = a_i)\mathbb{P}(Y_N = s_1 \mid Y_0 = a_i)\right) \\
&\geq \left(\alpha \sum_{i=1}^{n} \mathbb{P}(X_0 = a_i)\right)\left(\alpha \sum_{i=1}^{n} \mathbb{P}(Y_0 = a_i)\right) = \alpha^2.
\end{aligned}$$

Damit folgt aber direkt, dass

$$\mathbb{P}(T > N) \leq 1 - \alpha^2.$$

Ebenso gilt

$$\begin{aligned}
\mathbb{P}(T > 2N) &\leq \mathbb{P}(T > N)\mathbb{P}(T > 2N \mid T > N) \\
&\leq (1 - \alpha^2)\mathbb{P}(T > 2N \mid T > N) \\
&\leq (1 - \alpha^2)\mathbb{P}(X_{2N} \neq Y_{2N} \mid T > N) \\
&= (1 - \alpha^2)(1 - \mathbb{P}(X_{2N} = Y_{2N} \mid T > N)) \\
&\leq (1 - \alpha^2)^2.
\end{aligned}$$

Damit gilt aber dann iterativ auch für alle $l \in \mathbb{N}$, dass

$$\mathbb{P}(T > lN) \leq (1 - \alpha^2)^l.$$

Also folgt

$$\lim_{k \to \infty} \mathbb{P}(T > k) = 0 \qquad (*)$$

und damit die Konvergenz einer Treffwahrscheinlichkeit beider Markovketten gegen 1.

Das coupling-Prinzip geht nun so vor, dass wir eine dritte Markovkette $Z = (Z_0, Z_1, \dots)$ konstruieren, die so beschaffen ist, dass sie bis zum Aufeinander von X und Y identisch ist mit X und

danach identisch ist mit Y:

$$Z_0 := X_0$$
$$Z_{t+1} := \begin{cases} X_{t+1} & \text{falls } Z_t \neq Y_t \\ Y_{t+1} & \text{falls } Z_t = Y_t \end{cases}$$

Dann lässt sich abschätzen, dass

$$\begin{aligned} \mu_i^{(t)} - \pi_i &= \mathbb{P}(Z_t = a_i) - \mathbb{P}(Y_t = a_i) \\ &\leq \mathbb{P}(Z_t = a_i, Y_t \neq a_i) \\ &\leq \mathbb{P}(Z_t \neq Y_t) \\ &= \mathbb{P}(T > n). \end{aligned}$$

Analog lässt sich $\pi_i - \mu_i^{(t)}$ abschätzen. Und mit (*) folgt dann

$$\lim_{t\to\infty} |\mu_i^{(t)} - \pi_i| = 0$$

und damit die Behauptung. ∎

7.2 Modellierung von DNA-Sequenzen

Eine durch $(\mathbf{P}, \mu^{(0)})$ vorgegebene Markovkette nimmt nun zu jedem Zeitpunkt $t = 1, 2, \ldots$ einen Zustand aus $\mathcal{A}$ an, sie erzeugt also, anders gesprochen, unendliche Sequenzen der Form

$$x : x_1 x_2 x_3 \ldots \quad \text{mit } x_i \in \mathcal{A} \text{ für alle } i \in \mathbb{N}$$

Für die Modellierung von endlichen DNA-Sequenzen betrachtet man immer die ersten L Folgenglieder für ein festes $L \in \mathbb{N}, L \geq 2$. Dazu führt man einen Anfangszustand $\mathcal{B}$ und einen Endzustand $\mathcal{E}$ ein. Die Wahrscheinlichkeiten vom Anfangszustand $\mathcal{B}$ in einen Zustand aus $\mathcal{A}$ überzugehen sind dabei durch den Vektor $\mu^{(0)}$ der Anfangsverteilung gegeben. Unter dem Endzustand $\mathcal{E}$ versteht man einen absorbierenden Zustand, also einen Zustand, der mit Wahrscheinlichkeit 1 in sich selbst zurückführt. Jede vorgegebene endliche Sequenz $x = x_1 \ldots x_L$ wird nun von der Markovkette mit einer Wahrscheinlichkeit realisiert, die sich mittels

$$\mathbb{P}(x) = p_{\mathcal{B}x_1} \cdot p_{x_1x_2} \cdot \cdots \cdot p_{x_{L-1}x_L} \cdot p_{x_L\mathcal{E}} \tag{3}$$

berechnen lässt. Die Summe aller möglichen Sequenzen mit fester Länge L sollte natürlich gleich 1 sein:

Satz 7.6 *Sei X eine Markovkette, die Werten in einer endlichen Menge $\mathcal{A}$ annimmt und durch die Übergangsmatrix $\mathbf{P}$ und die Anfangsverteilung v festgelegt ist. Betrachte Sequenzen $x = x_1x_2 \ldots x_L$ mit fester Länge L. Dann gilt*

$$\sum_x \mathbb{P}(x) = \sum_{x_1,\ldots,x_L \in \mathcal{A}} p_{\mathcal{B}x_1} \cdot \prod_{i=1}^{L-1} p_{x_ix_{i+1}} \cdot p_{x_L\mathcal{E}} = 1.$$

Beweis. Da die Länge der Sequenz mit L fixiert worden ist, gilt für alle $x_L \in \mathcal{A}$, dass $p_{x_L\mathcal{E}} = 1$. Man weiß ebenfalls, dass für eine Übergangsmatrix gilt $\sum_{a_i \in \mathcal{A}} p_{x_{i-1}a_i} = 1$. Damit folgt

$$\begin{aligned}
&\sum_{x_1,\ldots,x_L \in \mathcal{A}} p_{\mathcal{B}x_1} \cdot \prod_{i=1}^{L-1} p_{x_i x_{i+1}} \cdot p_{x_L\mathcal{E}} \\
&= \sum_{x_1,\ldots,x_{L-1} \in \mathcal{A}} p_{\mathcal{B}x_1} \cdots p_{x_{L-2}x_{L-1}} \sum_{x_L \in \mathcal{A}} p_{x_{L-1}x_L} \cdot \underbrace{p_{x_L\mathcal{E}}}_{=1} \\
&= \sum_{x_1,\ldots,x_{L-1} \in \mathcal{A}} p_{\mathcal{B}x_1} \cdots p_{x_{L-2}x_{L-1}} \underbrace{\sum_{x_L \in \mathcal{A}} p_{x_{L-1}x_L}}_{=1} \\
&= \cdots = 1. \qquad \blacksquare
\end{aligned}$$

Eine weitere für uns wichtige Eigenschaft ist die folgende

Definition. *(nichttrivial verbunden)*
Eine Markovkette heißt **nichttrivial verbunden** falls für alle Zustände $a_i \in \mathcal{A}$ gilt: Existiert ein Pfad von $\mathcal{B}$ nach a_i, so existiert auch ein Pfad von a_i nach $\mathcal{E}$. Formal gilt also

$$p_{\mathcal{B}a_i} > 0 \quad \Rightarrow p_{a_i\mathcal{E}} > 0 \quad \text{für alle } i = 1, \ldots, n.$$

Fordert man diese Eigenschaft für eine Markovkette, so gilt Satz 7.6 ohne eine Länge L zu fixieren:

Satz 7.7 *Sei X wie in Satz 7.3 und nichttrivial verbunden. Dann gilt*

$$\sum_x \mathbb{P}(x) = \sum_{L=1}^{\infty} \sum_{x_1,\ldots,x_L \in \mathcal{A}} p_{\mathcal{B}x_1} \cdot \prod_{i=1}^{L-1} p_{x_i x_{i+1}} \cdot p_{x_L\mathcal{E}} = 1 \tag{4}$$

Beweis. Übung.

Zur Berechnung der Übergangswahrscheinlichkeiten benötigt man Trainingsdaten. Man fixiert die Konnektivität und wählt die Trainingsdaten so aus, dass sie zu dieser passen (Gibt es keine Pfad von a_i nach a_j, so darf auch in den Trainingsdaten auf a_i kein a_j folgen). Nun berechnet man die Matrix $\mathbf{P}$ durch

$$p_{ij} = \frac{H_{ij}}{\sum_{k=1}^{n} H_{ik}}, \tag{5}$$

wobei H_{ij} die Anzahl der Fälle darstellt, in denen auf den Zustand Z_i der Zustand Z_j folgt. Die Anfangsverteilung wird ebenso bestimmt. Bei der Auswahl der Trainingsdaten ist darauf zu achten, dass nur Markovketten entstehen, die nichttrivial verbunden sind.

Beispiel. Gegeben seien die folgenden DNA-Sequenzen (jeweils in Fett gedruckt sind die Stop-

und Startkodons, diese spielen keine Rolle).

$$x_1 : \mathbf{ATG}AACGTGCAGTTAGTG\mathbf{TAG}$$
$$x_2 : \mathbf{ATG}AAACTGACGCTGTTAACGTGA\mathbf{TAA}$$
$$x_3 : \mathbf{ATG}AGCAAGCTAGCTAGCTAA\mathbf{TAG}$$
$$x_4 : \mathbf{GTG}GCGACAAGA\mathbf{TAA}$$

Daraus ergeben sich mit (5) die folgenden relativen Häufigkeiten:

$$p_{\mathcal{B}A} = \frac{3}{4} \qquad p_{\mathcal{B}C} = \frac{1}{4} \qquad p_{\mathcal{B}G} = \frac{1}{4} \qquad p_{\mathcal{B}T} = \frac{0}{4}$$
$$p_{A\mathcal{E}} = \frac{3}{4} \qquad p_{C\mathcal{E}} = \frac{0}{4} \qquad p_{G\mathcal{E}} = \frac{1}{4} \qquad p_{T\mathcal{E}} = \frac{0}{4}$$
$$p_{AA} = \frac{6}{17} \qquad p_{AC} = \frac{5}{17} \qquad p_{AG} = \frac{6}{17} \qquad p_{AT} = \frac{0}{17}$$
$$p_{CA} = \frac{4}{13} \qquad p_{CC} = \frac{0}{13} \qquad p_{CG} = \frac{4}{13} \qquad p_{CT} = \frac{5}{13}$$
$$p_{GA} = \frac{4}{16} \qquad p_{GC} = \frac{7}{16} \qquad p_{GG} = \frac{0}{16} \qquad p_{GT} = \frac{5}{16}$$
$$p_{TA} = \frac{5}{12} \qquad p_{TC} = \frac{0}{12} \qquad p_{TG} = \frac{5}{12} \qquad p_{TT} = \frac{2}{12}$$

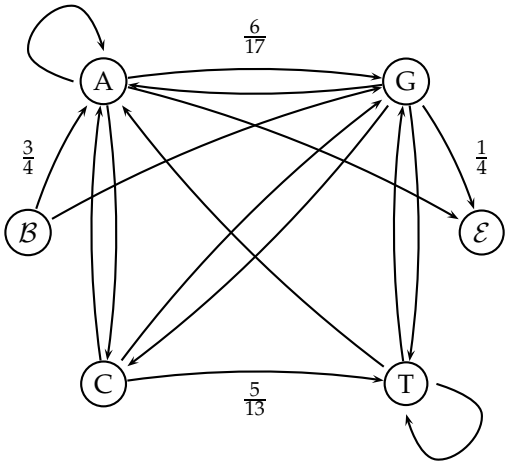

Abbildung 2: Beispiel einer aus Trainingsdaten gewonnenen Markovkette

7.3 Der „Hasting-Metropolis"-Algorithmus

Das Ziel in diesem Abschnitt ist die Konstruktion einer aperiodischen, minimalen Markovkette mit vorgeschriebener stationärer Verteilung $\pi = (\pi_1, \ldots, \pi_n)$. Unter den Voraussetzungen der Aperiodizität und der Irreduzibilität konvergiert die Markovkette gegen diese stationäre Verteilung. Doch wie simuliert man mit dem Computer eine Markovkette?

7.3.1 Simulation einer Markovkette

Wir setzen voraus, dass wir beliebig viele Zufallsvariablen $U_1, U_2, \ldots$ mit dem Computer erzeugen können, die alle gleichmäßig auf $[0,1]$ verteilt sind. Dies geschieht entweder mithilfe physikalischer Phänomene (Rauschen von elektronischen Bauteilen, Geigerzähler, quantenphysikalische Effekte...) oder mithilfe von Pseudozufallszahlen. Letztere sind durch ein festes vorgegebenes Verfahren erzeugte „Zufallszahlen", die streng genommen keinen echten Zufall repräsentieren, sondern nur von diesem nicht unterschieden werden können. Für viele Zwecke (wie auch unserem) sind diese ausreichend.

Um nun eine Markovkette, gegeben durch $(\mathbf{P}, \mu^{(0)})$, simulieren zu können, teilen wir das Intervall $[0,1]$ in disjunkte Teilintervalle $(A_i : i \in \{1,\ldots,n\})$, die jeweils die Länge

$$|A_i| = \mu_i^{(0)}$$

besitzen. Weiterhin unterteilen wir $[0,1]$ für alle $i \in \{1,\ldots,n\}$ in disjunkte Intervalle $(A_{ij} : j \in \{1,\ldots,n\})$ der Längen

$$|A_{ij}| = p_{ij}.$$

Dann definiere Funktionen

$$\begin{aligned} G_0 : [0,1] &\to \{1,\ldots,n\} \\ G : \{1,\ldots,n\} \times [0,1] &\to \{1,\ldots,n\} \end{aligned}$$

durch

$$\begin{aligned} G_0(u) &= i \text{ falls } u \in A_i, \\ G(i,u) &= j \text{ falls } u \in A_{ij}. \end{aligned}$$

Definiere weiterhin für $t \geq 0$.

$$\begin{aligned} X_0 &:= G_0(U_0) \\ X_{t+1} &:= G(X_t, U_{t+1}). \end{aligned}$$

Dann ist durch $(X_t)_{t\in\mathbb{N}}$ eine diskrete Markovette mit den Parametern $(\mathbf{P}, \mu^{(0)})$ gegeben, denn

$$\begin{aligned} &\mathbb{P}(X_0 = i) = \mathbb{P}(U_0 \in A_i) = \mu_i^{(0)}, \\ &\mathbb{P}(X_{t+1} = i_{t+1} \mid X_0 = i_0, \ldots, X_n = i_n) = \mathbb{P}(U_{t+1} \in A_{i_t, i_{t+1}}) = p_{i_t i_{t+1}} \end{aligned}$$

7.3.2 Der Hasting-Metropolis-Algorithmus

Wir möchten eine gegebene Verteilung

$$\pi = (\pi_1, \ldots, \pi_n) \quad \text{mit} \quad \sum_{i=1}^{n} \pi_i = 1$$

durch eine Markovkette simulieren. Die Markovkette ist so zu konstruieren, dass π stationär Verteilung ist. Eine dafür hinreichende Bedingung ist durch

$$\pi_i p_{ij} = \pi_j p_{ij} \quad \text{für alle} \quad i,j \in \{1,\dots,n\} \tag{6}$$

gegeben („detailed balance equation"), denn sei $\mathbf{P} = (p_{ij})$, dann gilt

$$\begin{aligned}
\pi \cdot \mathbf{P} &= (\pi_1,\dots,\pi_m) \cdot \begin{pmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & & \vdots \\ p_{n1} & \cdots & p_{nn} \end{pmatrix} \\
&= \begin{pmatrix} \pi_1 p_{11} + \cdots + \pi_n p_{n1} \\ \vdots \\ \pi_1 p_{1n} + \cdots + \pi_n p_{nn} \end{pmatrix}^T \overset{(6)}{=} \begin{pmatrix} \pi_1 p_{11} + \cdots + \pi_1 p_{1n} \\ \vdots \\ \pi_n p_{n1} + \cdots + \pi_m p_{nn} \end{pmatrix}^T \\
&= \begin{pmatrix} \pi_1 \cdot \sum_{i=1}^n p_{1i} \\ \vdots \\ \pi_n \sum_{i=1}^n p_{n1} \end{pmatrix}^T = (\pi_1,\dots,\pi_n) = \pi.
\end{aligned}$$

Es gibt viele Möglichkeiten für (p_{ij}), die Bedingung (6) erfüllen. Eine davon kann folgendermaßen konstruiert werden: Wir wählen beliebige Konstanten $\{q_{ij}\}$ mit $q_{ij} > 0$ für alle $i,j = 1,\dots,m$ und $\sum_j q_{i,j} = 1$. Dann definieren wir

$$a_{ij} := \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}, \qquad p_{ij} := q_{ij} a_{ij} \quad \text{für } i \neq j, \qquad p_{ii} := 1 - \sum_{i \neq j} p_{ij}. \tag{7}$$

Satz 7.8 *Die durch (7) definierte Markovkette ist aperiodisch und minimal. Weiterhin ist ihre stationäre Verteilung gegeben durch π.*

Beweis. Es ist $q_{ij} > 0$ für alle $i,j \in \{1,\dots,m\}$ nach Definition. Also sind auch $p_{ij} > 0$. Für $i \neq j$ ist dies klar, da $a_{ij} > 0$. Weiter ist nach Definition $a_{ij} \leq 1$ und $\sum_{i=1}^m p_{ji} = 1$ und damit gilt für alle $j \in \{1,\dots,m\}$

$$\begin{aligned}
\sum_{j \neq i} p_{ij} &= \sum_{i=1, j \neq i}^{n} a_{ij} q_{ij} \leq \sum_{i=1, j \neq i} q_{ij} < 1 \\
\Rightarrow \quad p_{ii} &= 1 - \sum_{j \neq i}^{n} p_{ij} > 0.
\end{aligned}$$

Also gilt $p_{ij} > 0$ auch für $i = j$ und die Markovkette ist damit irreduzibel und aperiodisch. Wir zeigen, dass Gleichung (6) erfüllt ist. Aus dieser folgt direkt, dass π stationäre Verteilung der gegebenen Markovkette sein muss.

1.Fall : $\frac{\pi_j q_{ij}}{\pi_i q_{ij}} < 1$.

$$\frac{\pi_j q_{ji}}{\pi_i q_{ij}} < 1 \quad \Rightarrow \quad \frac{\pi_i q_{ij}}{\pi_j q_{ji}} > 1$$

und damit folgt

$$a_{ij} = \frac{\pi_j q_{ij}}{\pi_i q_{ij}}, \qquad p_{ij} = \frac{\pi_j q_{ji}}{\pi_i}, \qquad a_{ji} = 1, \qquad p_{ji} = q_{ji}$$

und daraus folgt direkt (6).

2.Fall : $\frac{\pi_j q_{ij}}{\pi_i q_{ij}} = 1$. Dann folgt

$$\pi_j q_{ji} = \pi_i q_{ij}, \qquad a_{ij} = a_{ji} = 1 \qquad \Rightarrow \quad q_{ij} = p_{ij} \quad \text{und} \quad q_{ji} = p_{ji}$$

und damit direkt (6). ∎

Die obigen Definition in (14) haben folgende Interpretation: Falls $X_t = i$, dann wird eine neue Zufallsvariable simuliert, fur die $\mathbb{P}(Y_t = a_j) = r_{ij}$ gilt. Ist dann $Y_t = j$, dann setze

$$X_{t+1} = \begin{cases} Y_t & \text{mit Wahrscheinlichkeit } \min\{\frac{\pi_j r_{ji}}{\pi_i r_{ij}}, 1\} \\ X_t & \text{sonst} \end{cases}$$

Dieses Verfahren nennt man einen **Hastings-Algorithmus**. Die Konstanten $\{q_{ij}\}$ sind willkürlich gewählt. Wählt man sie so, dass

$$q_{ij} = q_{ji} \quad \text{für alle } i, j \in \{1, \ldots, n\},$$

so erhält man

$$p_{ij} = \left(\min \left\{ \frac{\pi_j}{\pi_i}, 1 \right\} \right) \cdot q_{ij} \quad \text{für } i \neq j.$$

Dieser Speziallfall trägt den Namen **Metropolis-Algorithmus**.