Proseminarvortrag

Markov-Ketten in der Biologie (Anwendungen)

von

Peter Drössler

20.01.2010

Inhalt

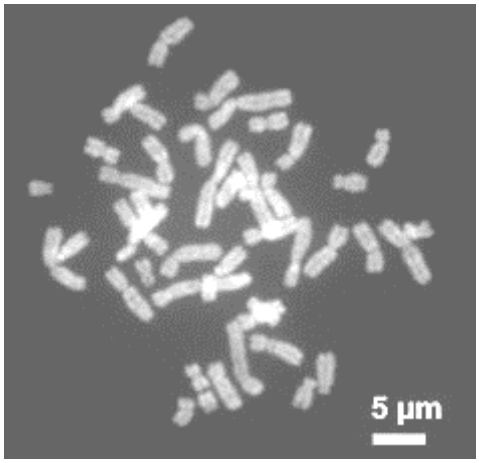
1. Das Wright-Fisher Modell

1.1. Notwendige Definition aus der Biologie (Allel):

Ein **Allel** bezeichnet eine mögliche Ausprägung eines Gens, das sich an einem bestimmten Ort auf einem Chromosom befindet.

1.2. Notwendige Definition aus der Biologie (Chromosom):

Chromosomen sind Strukturen, die Gene und damit Erbinformationen enthalten. Sie bestehen aus DNA, die mit vielen Proteinen verpackt ist. Diese Mischung aus DNA und Proteinen wird auch als Chromatin bezeichnet.



Diese Allelen aus Definition 1.1 und deren Ausprägung in verschiedenen Generationen werden wir uns mithilfe von Markov-Ketten genauer betrachten.

Problemstellung: Gene ändern sich im Verlauf der Zeit, werden von Generation zu Generation neu vermischt.

Ebenso interessiert uns die Fragestellung, ob sich der Gen-Pool irgendwann nicht mehr ändert oder Genausprägungen ganz verschwinden (Absorption).

1.3. Definition: Wright-Fisher-Model

Es sei eine diskrete Markov-Kette $(X_n)_{n\geq 0}$ mit $I = \{0, 1, \dots, \mathrm{m}\}$ gegeben mit Übergangsmatrix P und und der Übergangswahrscheinlichkeit:

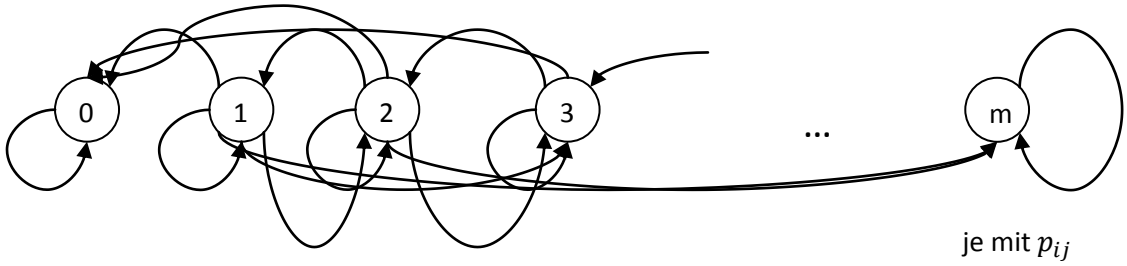
$$p_{ij} = \binom{m}{j} * \left(\frac{i}{m}\right)^{j} * \left(\frac{m-i}{m}\right)^{m-j}$$

Es gilt:

- Es ist möglich von jedem **nicht absorbierenden** Zustand in jeden anderen überzugehen: $p_{ij} > 0 \; \forall ij \epsilon (1 \dots m-1)$
- Die Zustände m und 0 sind absorbierend:

$$p_{00} = \binom{m}{0} \cdot \left(\frac{0}{m}\right)^{0} \cdot \left(\frac{m-0}{m}\right)^{m-0} = 1 \text{ (Absorptionszustand)}$$

$$p_{mm} = \binom{m}{m} \cdot \left(\frac{m}{m}\right)^{m} \cdot \left(\frac{m-m}{m}\right)^{m-m} = 1 \cdot \; 1^{m} \cdot 0^{0} = 1 \text{ (Absorptionszustand)}$$



1.4. Beschreibung der Markov-Kette:

In jeder Generation befinden sich m Allelen, welche vom Typ **A** und **a**. Die Gattung der Allelen in Generation $n+1$ werden bestimmt durch zufälliges Ziehen **mit Zurücklegen** (!) der Allelen in Generation *n*.

Die Markov-Kette X_n mit Übergangswahrscheinlichkeit aus Def. 1.3. zählt nun die Anzahl der Allelen des Typs A in Generation *n*.

1.5. Kommunikations- & Absorptionsklassen (in I)

Die „Klassen" unserer Markov-Ketten sind gegeben durch

$$\{0\}, \{1, \dots, m-1\}, \{m\}$$

wobei 0 und m absorbierend sind und $\{1, \dots, m-1\}$ transient (Beh 1.5.1), d.h.

$$\mathbb{P}_i(X_n = i \text{ für } \infty \text{ viele n}) = 0 \; \forall i \in \{1, \dots, m-1\}$$

Bew 1.5.1.:

Mit *Th. 1.5.5.:* „Jede endlich rekurrente Klasse ist abgeschlossen" und *Th. 1.5.6.* „Eine endlich abgeschlossene Klasse ist rekurrent" ⇨ Jede nicht abgeschlossene endliche Klasse ist transient ⇨ Die endliche Klasse $\{1, \dots, m-1\}$ ist nicht abgeschlossen ⇨ Behauptung ■

1.6. Behauptung: Die Trefferzeit für m (pur AA) ist gegeben durch

$$h_i = \sum_{j=0}^{m} p_{ij} * h_j = \sum_{j=0}^{m} \binom{m}{j} * \left(\frac{i}{m}\right)^j * \left(\frac{m-i}{m}\right)^{m-j} * h_j = \frac{i}{m}$$
$$= \mathbb{P}_i(X_n = m \text{ für } \infty \text{ viele } n)$$

Beweis:

<u>Erinnerung Beispiel 3 (Proseminarvortag am 14.01.2010)</u>

Sei $\mathcal{I} = \{0 \dots m\}$ wobei $\{1 \dots m-1\}$ nicht geschlossene endliche Klasse und $\{0, m\}$ absorbierend, dann gilt:

a) $h_i^{\{0,m\}} = 1 \; \forall i \in \mathcal{I}$

b) Ist x mit $0 \leq x \leq 1$ eine Lösung des folgenden Gleichungssystems:

$$(1.6.1): \begin{cases} x_0 = m \\ x_m = 1 \\ x_i = \sum_{j=0}^{m} p_{ij} \cdot x_j \text{ für } i = 1 \dots (m-1) \end{cases}$$

Dann gilt $h_i^{\{m\}} = x_i$ und $h_i^{\{0\}} = 1 - x_i$ für $i = 0 \dots m$

Sei $h_i = h_i^{\{m\}}$ die Trefferzeit von m bei Start in i und $(x_j) := \left(\frac{i}{m}\right)_{i=0,\dots,m}$

Zu zeigen ist also: $x = (x_j)$ löst $(1.6.1)$

Beweis:

Es gilt für i=m:

$$x_m = \sum_{j=0}^{m} \binom{m}{j} \cdot \left(\frac{m}{m}\right)^j \cdot \left(\frac{m-m}{m}\right)^{m-j} \cdot \frac{j}{m} = \sum_{j=0}^{m} \binom{m}{j} \cdot 1 \cdot 0^{m-j} \cdot \frac{j}{m} = 1 \cdot 1 \cdot 0^0 \cdot \frac{m}{m}$$
$$= 1$$

Es gilt außerdem für $i \neq m$

$$x_i = \sum_{j=0}^{m} p_{ij} \cdot x_j = \sum_{j=0}^{m} \binom{m}{j} \cdot \left(\frac{i}{m}\right)^j \cdot \left(\frac{m-i}{m}\right)^{m-j} \cdot \frac{j}{m}$$

$$= \sum_{j=0}^{m} \underbrace{\frac{(m-1)!}{(m-j)! \cdot (j-1)!}}_{\binom{m-1}{j-1}} \cdot \left(\frac{i}{m}\right)^j \cdot \left(\frac{m-i}{m}\right)^{m-j} = \{l := j-1\}$$

$$= \sum_{l=0}^{m-1} \binom{m-1}{l} \cdot \left(\frac{i}{m}\right)^{l+1} \cdot \left(\frac{m-i}{m}\right)^{(m-1)-l} = \frac{i}{m} \cdot \underbrace{\sum_{l=0}^{m-1} \binom{m-1}{l} \cdot \left(\frac{i}{m}\right)^l \cdot \left(\frac{m-i}{m}\right)^{(m-1)-l}}_{1}$$

$$= \frac{i}{m}$$

Es gilt $1 = \left(\frac{i}{m} + \frac{m-i}{m}\right)^{m-1} = \sum_{l=0}^{m-1} \binom{m-1}{l} \cdot \left(\frac{i}{m}\right)^l \left(\frac{m-i}{m}\right)^{(m-1)-l}$

■

1.7. Bemerkungen:

- Das Wright-Fischer Modell modelliert genetische Änderungen, die sich vorwärts in der Zeit abspielen.
- D. h.: Es modelliert die genetisches Zusammensetzung einer Generation n+1 anhand der Daten aus Generation n.
- Genetische Daten aus Generationen <n spielen hier keine Rolle → Markov Eigenschaft gilt (anschaulich). Es geschieht von Generation zu Generation alles zufällig. Es gibt kein Gedächtnis. Für einen Beweis dieser Behauptung müsste man Beobachtungen über mehrere Jahre hinweg anführen.
- Dieses Modell ist jedoch realitätsfremd, da mit Zurücklegen gezogen wird und somit möglich ist, dass sich ein Allel mit sich selbst paart.

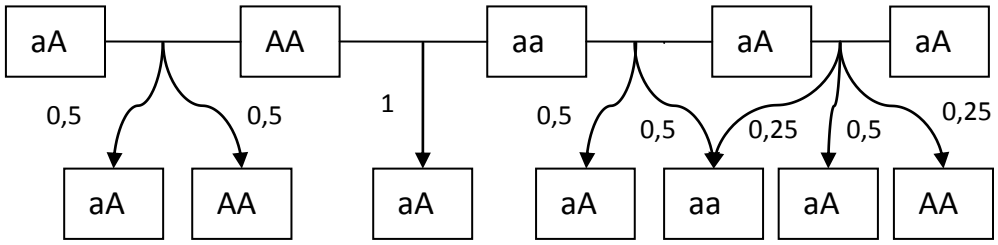
1.8. Beispiel: Betrachtung als Doppel-Allele:

Betrachten wir vereinfacht Individuen mit l Allelen, in unserem Beispiel $l = 2$. Betrachtet werden also Gene mit je 2 Allelen. Wir erwarten also, dass jedes Individuum zwei Allele besitzt, damit ergeben sich die Möglichkeiten

$$AA, Aa, aa$$

Sei nun $m = 2k$ (k Individuen, m Allele) und nehmen wir an, dass alle Individuen der aktuellen Generation mit zufälligen anderen der gleichen Generation Nachkommen zeugen und diese je ein Allel von beiden Elternteilen bekommen. Wir erlauben hierbei, dass beide Elternteile durchaus die gleichen Allel-Ausprägungen haben, wir machen auch keine Unterscheidung zwischen beiden Geschlechtern.

Anschaulich:



Sei in der aktuellen Generation n folgende Situation gegeben:

$$AA \quad aA \quad AA \quad AA \quad aa$$

Es wird nun rein zufällig und unabhängig voneinander m-mal hintereinander ein Doppelgen herausgegriffen. Dann ist jedes Gen in der Generation $n + 1$ von der Ausprägung A mit der Wahrscheinlichkeit $0{,}7$ und a mit $0{,}3$.

Begründung: Dieses Modell lässt sich auf ein einfaches zweistufiges Experiment reduzieren. Hier wählen wir statt in ersten Instanz die Eltern und in der zweiten das Gen, das vererbt wird, einfach die Geneausprägungen aus dem Pool, die Vererbt werden, unabhängig davon wie Mutter und Vater beschaffen sind. (dies gelte hier ohne Beweis).

Damit würde sich unser Pool aus Doppelallen vereinfachen auf:

$$A\ A\ a\ A\ A\ A\ A\ A\ a\ a$$

Nun greifen wir k-mal nacheinander zwei Allele aus dem Genpool und und bilden daraus die jeweiligen Nachkommen (auch Doppelallelen), wobei diese Verteilung nach wie vor unabhängig ist. Wir würden so beispielhaft folgendes Ergebnis bekommen:

$$aa \quad aA \quad Aa \quad AA \quad AA$$

1.9. Folgerung:

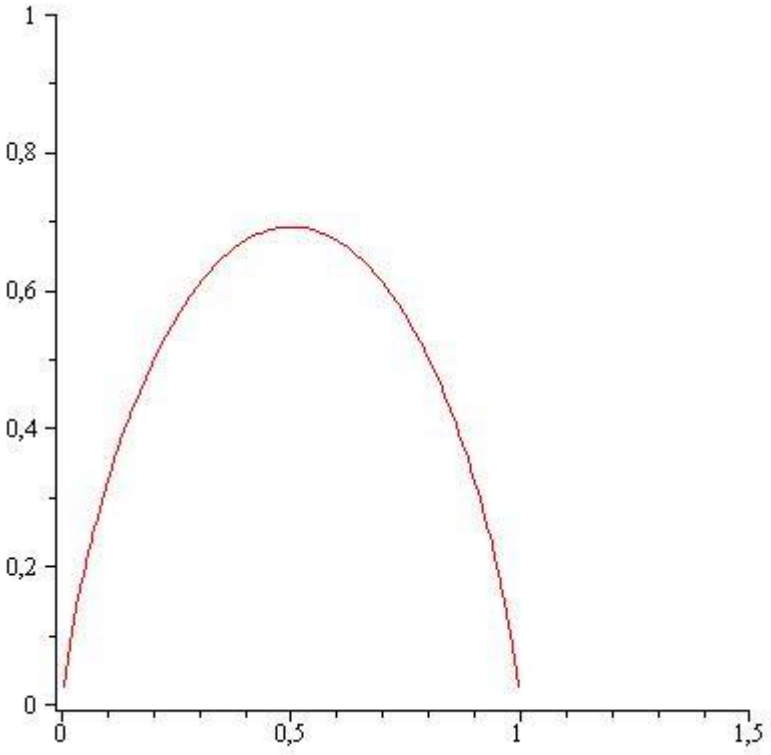
Die Struktur der Paare ist irrelevant in der Markov-Kette$(X_n)_{n\geq 0}$, die nur die Anzahl der Allelen des Typs A zählt.

1.10. Erwartungswert & Folgerung:

Dieses Modell wirft die Frage auf, ob die biologische Vielfalt der Paare irgendwann verschwindet oder sich stabilisiert. Es ist bekannt für $\frac{i}{m} = p \in (0,1)$ und $m \to \infty$

$$\mathbb{E}_{pm}(\boldsymbol{T}) \cong (-\boldsymbol{2m}) * \underbrace{\left((\boldsymbol{1}-\boldsymbol{p}) * \boldsymbol{\ln}(\boldsymbol{1}-\boldsymbol{p}) + \boldsymbol{p} * \boldsymbol{\ln} \boldsymbol{p}\right)}_{<0}$$

wobei T die Trefferzeit von $\{0, m\}$, also der absorbierenden *Zustände* ist.



m=1

1.7.1. Folgerung: In einer großen Population (für große m) geht die biologische Vielfalt also nicht verloren.

1.8. Modifikationen, Annahmen & hinreichende Findung von p

Wenn man andere Aspekte aus der Theorie der Genetik mit einbezieht lässt sich die Übergangswahrscheinlichkeit unserer Markov-Kette (X_n) genauer beschreiben:

1.8.1. Modifikation 1:

Es könnte z. B. sein, dass die drei genetischen Typen AA, aA, aa je einen gewissen relativen Vorteil gegenüber den jeweils anderen besitzen, gegeben durch die Parameter: $\alpha, \beta, \gamma > 0$. Aus diesem Aspekt lässt sich die Wahrscheinlichkeit für A im Falle $X_n = i$ angeben:

$$\psi_i = \frac{\alpha\left(\frac{i}{m}\right)^2 + 0{,}5 * \beta * i * \frac{m-i}{m^2}}{\alpha\left(\frac{i}{m}\right)^2 + \frac{\beta(m-i)}{m^2} + \gamma\left(\frac{m-i}{m}\right)^2}$$

Damit wären die Übergangswahrscheinlichkeiten unserer Markov-Ketten gegeben durch:

$$p_{ij} = \binom{m}{j} * \psi_i^j (1-\psi_i)^{m-j}$$

1.8.2. Modifikation 2: Mutation

Gleichzeitig können wir auch annehmen, dass gewisse Gene **mutieren**. Sie die Wahrscheinlichkeit, dass A zu a mutiert gegeben durch $\boldsymbol{u}$ und im umgekehrten Fall $\boldsymbol{v}$, dann ist die Wahrscheinlichkeit von A für $X_n = i$ gegeben durch

$$\phi_i = \frac{i(1-u) + (m-i)v}{m}$$

und

$$p_{ij} = \binom{m}{j} * \phi_i^j * (1-\phi_i)^{m-j}$$

1.8.3. Folgerungen aus 1.8.2.:

Mit $u, v > 0$ sind die Zustände 0 und m nicht mehr absorbierend. Damit wird die Markov-Kette irreduzibel und die Übergangswahrscheinlichkeiten gehen in eine invariante Verteilung π über.

Dann gilt:

a) $X_1 \sim \mathrm{Bin}(m, \phi_i)$
b) $\mathbb{E}_i(X_1) = m \cdot \phi_i$
c) $\mu := \mathbb{E}_\pi(X_1) = \frac{m \cdot v}{u+v}$

Beweis c):

$$\mu = \sum_{i=0}^{m} \underbrace{P(X_0 = i)}_{\pi_i} \cdot \underbrace{\mathbb{E}_\pi(X_1 | X_0 = i)}_{\mathbb{E}_i(X_1)} = \sum_{i=0}^{m} \pi_i \mathbb{E}_i(X_1)$$

$$= \sum_{i=0}^{m} m\pi_i\phi_i = \sum_{i=0}^{m} (i(1-u) + (m-i)v) \cdot \pi_i = (1-u)\mu + mv - v\mu$$

$\Rightarrow$ Behauptung ■

2. Moran-Modell

2.1. Einführung:

Das Moran-Modell ist eine Variante des Wright-Fisher-Modells. Zu diesem Ergebnis werden wir aber erst am Ende kommen. Das Moran Modell spiegelt eine Entwicklung einer Bevölkerung in Form einer Geburten-Sterbe-Kette (en: „Birth-and-death-chain“) wieder.

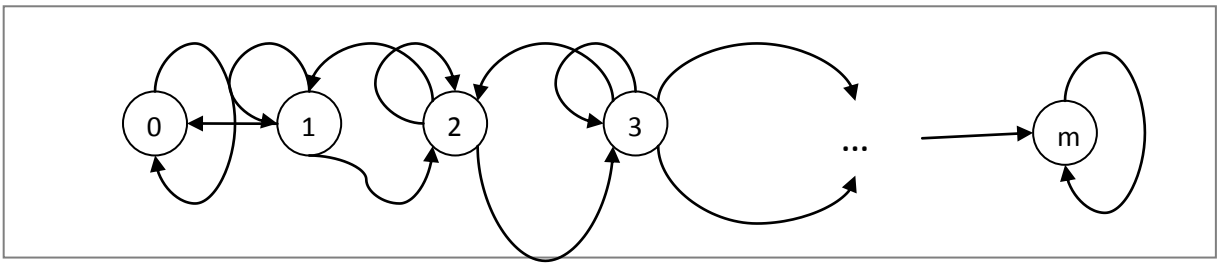
2.2. Definition:

Sei $(X_n)_{n\geq 0}$ eine Markov-Kette mit $I = \{0,1,\dots,m\}$ mit der Übergangsmartix P und der Übergangswahrscheinlichkeiten:

$$p_{i,i-1} = \frac{i(m-i)}{m^2}$$

$$p_{ii} = \frac{i^2 + (m-i)^2}{m^2}$$

$$p_{i,i+1} = \frac{i(m-i)}{m^2}$$



2.3. Genetische Interpretation:

Eine Population besteht aus Individuen zweier Typen A und a. Wir ziehen zufällig zweimal ein Individuum aus dieser Population zur Zeit n und fügen ein Gen des Typs der ersten Ziehung hinzu, legen dafür das zweite gezogene nicht mehr zurück: Damit bekommen wir die Population in Generation $n+1$.

X_n zähle hierbei wieder die Anzahl der Allelen des Typs A.

(Baum)

2.4. Beispiel:

A	**A**	**a**	**a**	**A**		A gezogen
A	**A**	**a**	**a**	**A**	**A**	**a gezogen**
A	**A**	**a**		**A**	**A**	**Neue Vert.**

2.5. Bemerkungen

- Es kann dabei passieren, dass zweimal hintereinander das gleiche Element gezogen wird, was keine Änderung in der Zusammensetzung der Bevölkerung zur Folge hätte
- Unterschied zum Wright-Fischer-Modell: Hier können keine Paare von Genen wie in 1.4. betrachtet werden.
- Im Moran-Modell wird immer nur ein Individuum einer aktuellen Population zum Zeitpunkt n geändert, keinesfalls die ganze Bevölkerung.

2.6. Struktur der zugrundeliegenden Markov-Kette:

Die Klassen-Struktur der Markov-Kette ist die gleiche wie in 1.5.: Wir haben Kommunikations- und Absorptionsklassen

$$\{0\}, \{1, \dots, m-1\}, \{m\}$$

wobei wir wieder die transiente Klasse $\{1, \dots, m-1\}$ (Beweis analog wie in 1.6)haben, sowie absorbierende Klassen $\{0\}$ und $\{m\}$. Das Moran-Modell ist reversibel und ganz ähnlich aufgebaut wie das Wright-Fischer-Modell.

- Die **Trefferwahrscheinlichkeit** ist gegeben durch: $\mathbb{P}_i(X_n = m \text{ für } \infty \text{ viele } n) = \frac{i}{m}$
- Ebenfalls angeben können wir die **Absorptionszeit**: $k_i = \mathbb{E}_i(T)$ mit Trefferzeit T von $\{0, m\}$

2.7. Absorptionszeit

Die einfachste Methode, um die Absorptionszeit zu bestimmen, ist j festzuhalten und die die bekannten Gesetze für die Zeit k_i^j (Haltezeit in j, Start in i) anzuwenden:

$k_i^j = \delta_{ij} + \left(p_{i,i-1}k_{i-1}^j + p_{ii}k_i^j + p_{i,i+1}k_{i+1}^j\right)$ für $i = 1, \dots, m-1$

$k_0^j = k_m^j = 0$

Betrachte $i = 1, \dots, m-1$

$$k_{i+1}^j - 2k_i^j + k_{i-1}^j = -\frac{\delta_{ij}m^2}{j(m-j)}$$

Damit

$$k_i^j = \begin{cases} \left(\frac{i}{j}\right)k_j^j \; f\ddot{u}r\; i \le j \\ \left(\frac{m-i}{m-j}\right)k_j^j \; f\ddot{u}r\; i \ge j \end{cases}$$

wobei k_j^j bestimmt ist durch

$$\left(\frac{m-j-1}{m-j} - 2 + \frac{j-1}{j}\right)k_j^j = -\frac{m^2}{j(m-j)}$$

woraus folgt, dass $k_j^j = m$

Daraus folgt

$$k_i = \sum_{j=1}^{m-1} k_i^j = m\left\{\sum_{j=1}^{i}\left(\frac{m-i}{m-j}\right) + \sum_{j=i+1}^{m-1}\frac{i}{j}\right\}$$

2.8. **Vergleich:**

Betrachte nun die Kette für großes m und $i = mp$ für $p\epsilon(0,1)$

Dann gilt:

$$\frac{1}{m^2}\cdot k_{pm} = (1-p)\cdot\sum_{j=1}^{mp}\frac{1}{m-j} + p\cdot\sum_{j=mp+1}^{m-1}\frac{1}{j} \xrightarrow{m\to\infty} -(1-p)\ln(1-p) - p\cdot\ln p$$

Der Erwartungswert ist dann gegeben durch

$$\mathbb{E}_{pm}(T) \cong (-m^2)\cdot\left((1-p)\cdot\ln(1-p) + p\cdot\ln p\right)$$

im Vergleich zum Wright-Fisher-Modell

$$\mathbb{E}_{pm}(T) \cong (-2m)\cdot\left((1-p)\cdot\ln(1-p) + p\cdot\ln p\right)$$

Der Unterschied liegt hier gerade bei einem Faktor m/2. Diesen kann man dadurch erklären, dass das Moran-Modell ausschließlich ein Individuum pro Zeitperiode ändert, während das Wright-Fisher Modell alle m ändert.