Hidden-Markov-Modelle zur Bestimmung wahrscheinlichster Ereignisse

Hans-Joachim Böckenhauer
Dennis Komm

Volkshochschule Zürich

07. Mai 2014

Eine Fragestellung aus der Biologie

Beobachtung einer Bakterienkultur

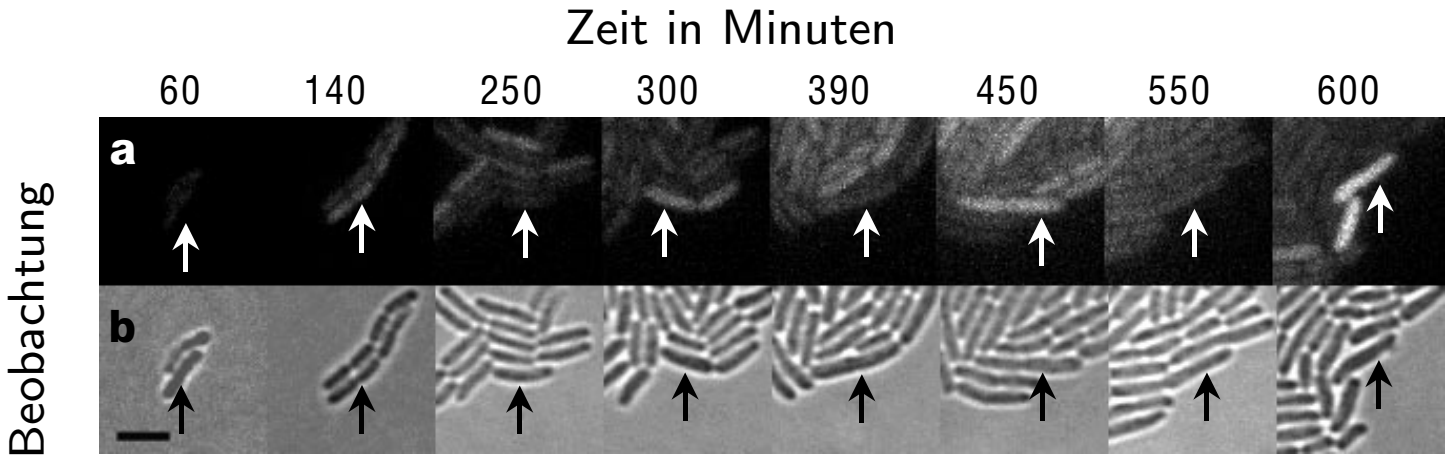
Wie verändert sich eine Bakterienkultur über die Zeit?

Konkreter Versuchsaufbau:

- Nährlösung unter sich verändernden Einflüssen (Temperaturanstieg, Hinzugabe von Chemikalien, . . .)
- Wir wollen untersuchen:
  Wann sind Bakterien krank, wann gesund?
- Wenn gesund, wird Protein von Bakterium produziert
- Veränderte Bakterien-DNA (Fluoreszenz-Marker)
- Beobachte Fluoreszenz-Level über die Zeit

Beobachtung einer Bakterienkultur

Wir machen folgende Beobachtungen:



Elowitz, Leibler, 2000, Nature 403

Für das Fluoreszenz-Level beobachten wir
Niedrig, **Mittel**, **Mittel**, **Hoch**, **Mittel**, **Hoch**, **Mittel**, **Hoch**

Was folgt nun für den Gesundheitszustand des Bakteriums?

Hidden-Markov-Modelle

„Das unfaire Kasino“

- Ein fairer Würfel (Alle Resultate mit Wahrscheinlichkeit $\frac{1}{6}$)
- Ein unfairer Würfel (Resultat „6“ mit Wahrscheinlichkeit $\frac{1}{2}$)
- Zu Beginn wird ein Würfel gewählt (Wahrscheinlichkeit $\frac{1}{2}$)
- Nach Wurf kann gewechselt werden (Wahrscheinlichkeit $\frac{1}{20}$)

Markov-Modell

- **Zustand:** Benutzter Würfel
- **Übergangs-Wahrscheinlichkeit:** Wahrscheinlichkeit, den Würfel zu wechseln
- **Emissions-Wahrscheinlichkeit:** Wahrscheinlichkeit, eine konkrete Augenzahl zu beobachten; abhängig vom Zustand

Wahrscheinlichkeit einer konkreten Ausgabe

- Nehmen wir an, wir beobachten die Zahlen **6**, **6**, **6**, **6**
- Ausserdem vermuten wir, dass zunächst zweimal der unfaire Würfel verwendet wurde, dann zweimal der faire
- Dies entspricht einem festen Pfad durch unser Markov-Modell

Wie gross ist die Wahrscheinlichkeit?

Wahrscheinlichkeit einer konkreten Ausgabe

Betrachten wir nur das Werfen der ersten **6**

1. Ereignis $A =$ „Der unfaire Würfel wird zu Beginn gewählt“
2. Ereignis $B =$ „Es wird **6** gewürfelt“

Jetzt können wir einfach die Wahrscheinlichkeiten ausrechnen

$$
\begin{aligned}
&\mathit{Wahr}(\text{Der unfaire Würfel wird zu Beginn gewählt und } \mathbf{6} \text{ gewürfelt}) \\
&\quad = \mathit{Wahr}(A \text{ und } B) \\
&\quad = \mathit{Wahr}(A) \cdot \mathit{Wahr}(B \text{ unter der Voraussetzung } A) \\
&\quad = \mathit{Wahr}(A) \cdot \mathit{Wahr}(B \mid A) \\
&\quad = \frac{1}{2} \cdot \frac{1}{2}
\end{aligned}
$$

Wahrscheinlichkeit einer konkreten Ausgabe

Nun führen wir diese Anwendung für **6**, **6**, **6**, **6** fort

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{19}{20} \cdot \frac{1}{2} \cdot \frac{1}{20} \cdot \frac{1}{6} \cdot \frac{19}{20} \cdot \frac{1}{6}$$

Wahrscheinlichkeit, dass schliesslich wieder eine **6** geworfen wird

Wahrscheinlichkeit einer konkreten Ausgabe

***Hidden*-Markov-Modell**

Aber was passiert, wenn wir nur die Ausgaben beobachten?

- Nehmen wir an, wir beobachten die Zahlen **6**, **5**, **2**, **6**, **6**
- Was ist der wahrscheinlichste Pfad?
- Für drei Zahlen ist der unfaire Würfel „besser“
- Für zwei Zahlen der faire
- Es ist nicht klar, ob es sich „lohnt“ den Würfel zu wechseln

Wahrscheinlichkeit einer konkreten Ausgabe

- Anzahl möglicher Pfade für n Würfe ist 2^n
- Für 300 Würfe ist dies mehr als

$$10^{90}$$

- Aber müssen wir wirklich alle ausprobieren?
- Nein
- **Dynamische Programmierung**

Dynamische Programmierung

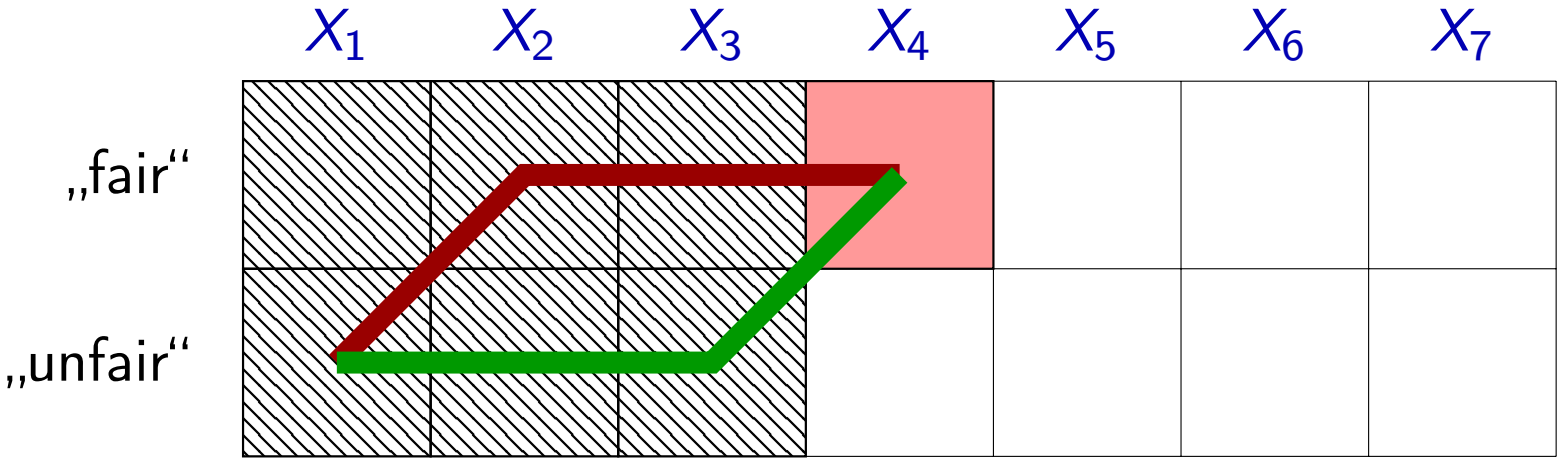
Prinzip der dynamischen Programmierung
Lösung für die gesamte Eingabe zusammensetzen aus Teillösungen für Teilprobleme, beginnend mit den kleinsten Teilproblemen

Problem: Finde geeignete Teilprobleme

Idee (Viterbi, 1967):

- Alle wahrscheinlichsten Pfade für Anfangsstücke (Präfixe) der gegebenen Folge von Würfen, die in gegebenem Zustand enden, als Teilprobleme
- Berechne wahrscheinlichste Pfade für längere Präfixe aus den wahrscheinlichsten Pfaden für kürzere Präfixe

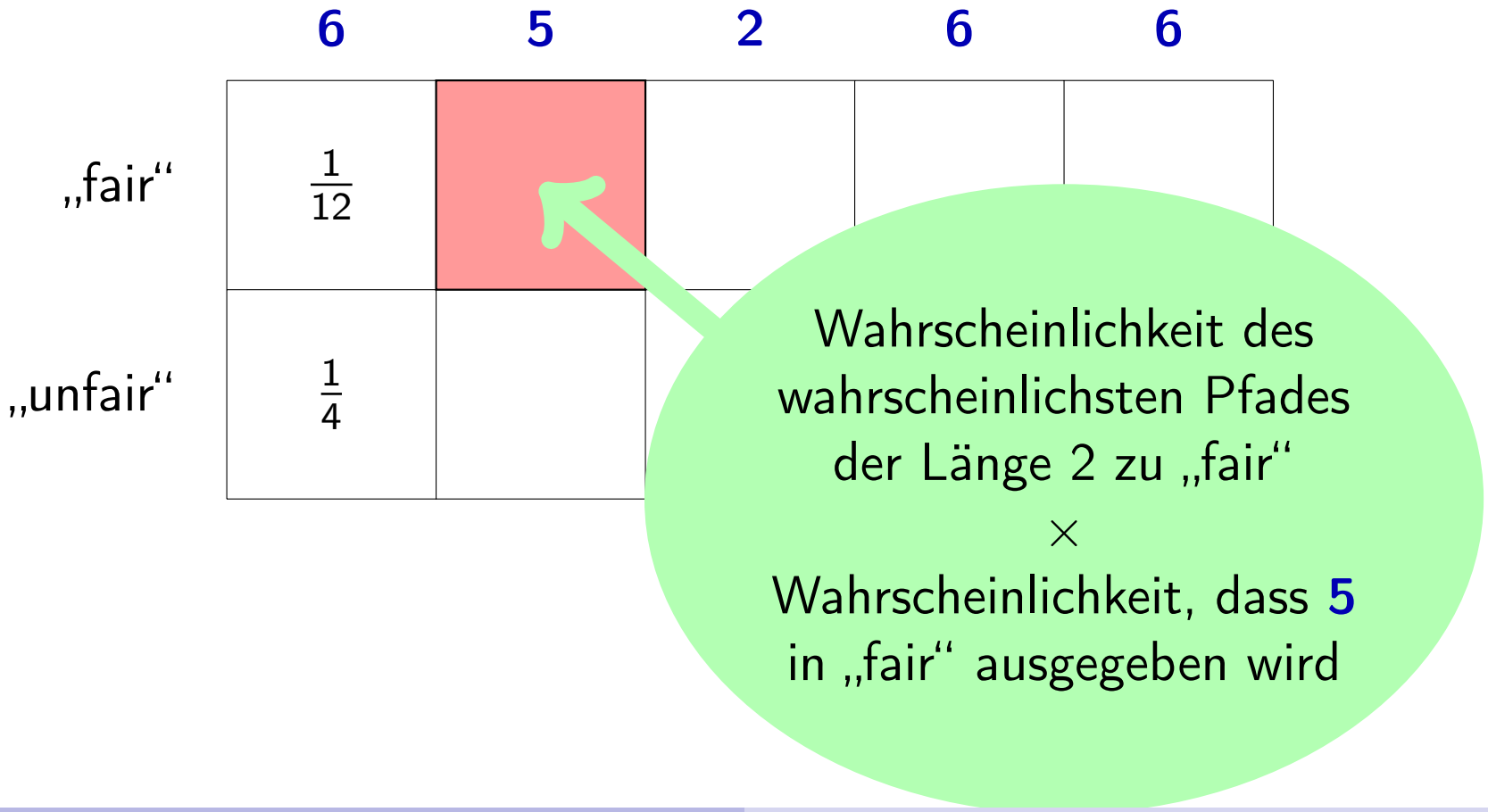
Dynamische Programmierung



- Nehmen wir an, die schraffierten Felder sind bereits ausgefüllt
- Wahrscheinlichste Pfade der Länge 3 sind bekannt
  (z.B. Pfad, der in X_3 endet: unfair, fair, fair)
- Berechne jetzt solchen Pfad der Länge 4, der in „fair“ endet
- Verlängere bekannte Pfade und nimm wahrscheinlicheren

Dynamische Programmierung

- Beobachtete Würfelfolge **6**, **5**, **2**, **6**, **6**
- Erstelle Tabelle (ähnlich wie bei Alignment)
- Merke, welcher Pfad bislang der beste war



Dynamische Programmierung

- Beobachtete Würfelfolge **6**, **5**, **2**, **6**, **6**
- Erstelle Tabelle (ähnlich wie bei Alignment)
- Merke, welcher Pfad bislang der beste war

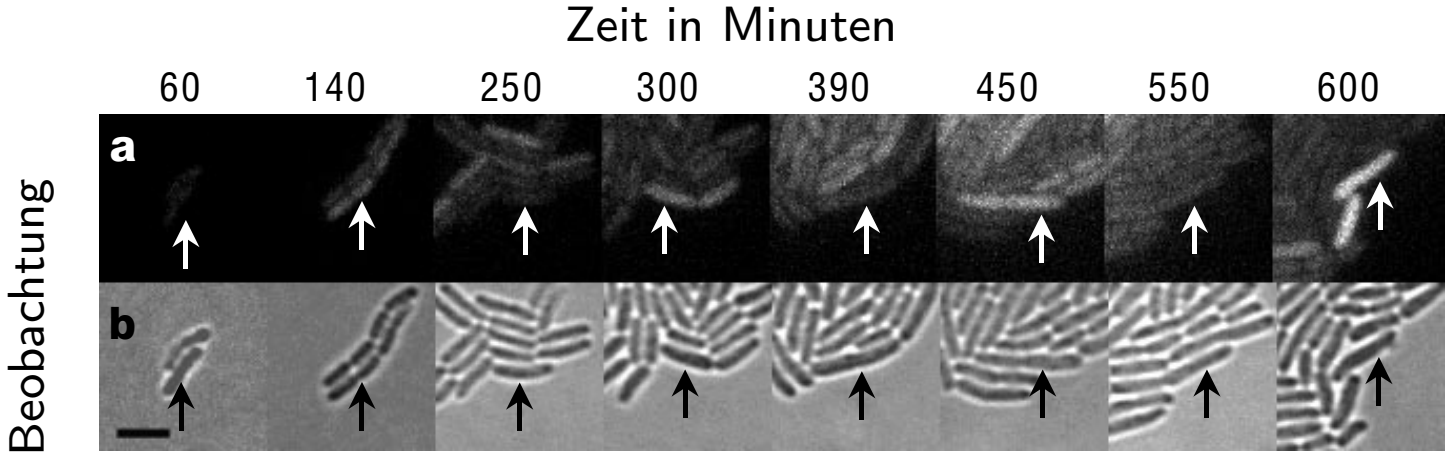
	6	**5**	**2**	**6**	**6**
„fair“	$\frac{1}{12}$	$\frac{19}{1440}$	$\frac{361}{172800}$	$\frac{6859}{20736000}$	$\frac{6859}{76800000}$
„unfair“	$\frac{1}{4}$ →	$\frac{19}{800}$ →	$\frac{361}{16000}$ →	$\frac{6859}{640000}$ →	$\frac{130321}{25600000}$

Der wahrscheinlichste Pfad ist also der, bei dem nur der unfaire Würfel verwendet wurde

Zurück zur Fragestellung aus der Biologie

Beobachtung einer Bakterienkultur

Wir machen folgende Beobachtungen:



Elowitz, Leibler, 2000, Nature 403

Beobachtung einer Bakterienkultur

Bakterie kann sich in einem von drei Zuständen befinden:

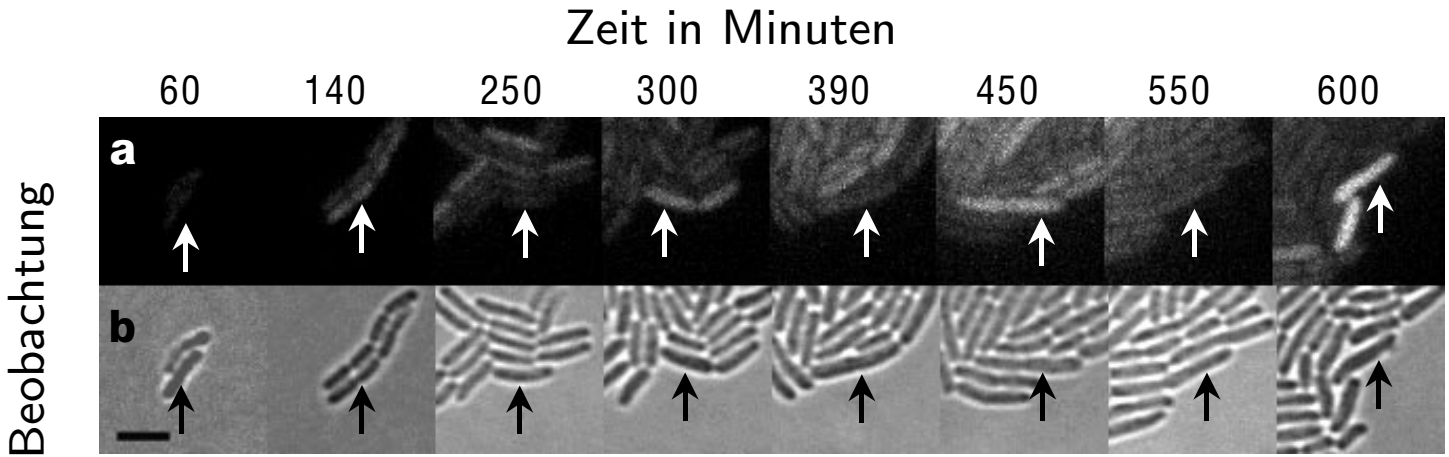
1. Gesund
2. OK
3. Krank

DNA wurde so verändert, das Bakterien fluoreszieren, und zwar je mehr desto gesünder sie sind; Fluoreszenz-Level kann beobachtet werden:

1. Hoch
2. Mittel
3. Niedrig

Beobachtung einer Bakterienkultur

Wir machen folgende Beobachtungen:



Elowitz, Leibler, 2000, Nature 403

Für das Fluoreszenz-Level beobachten wir
Niedrig, **Mittel**, **Mittel**, **Hoch**, **Mittel**, **Hoch**, **Mittel**, **Hoch**

Was folgt nun für den Gesundheitszustand des Bakteriums?

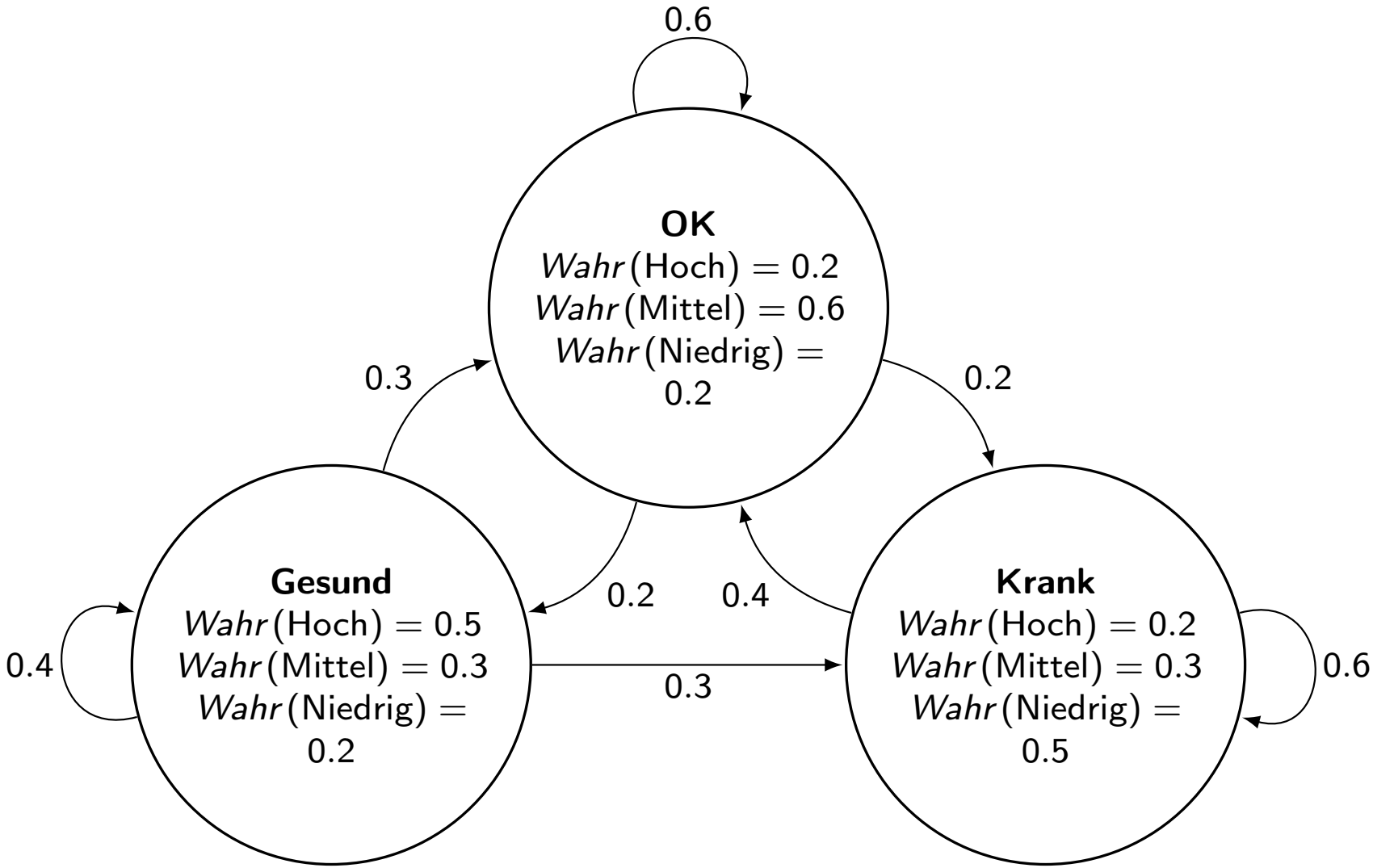
Beobachtung einer Bakterienkultur

Emissions-Wahrscheinlichkeiten
(erworben durch empirische Untersuchungen)

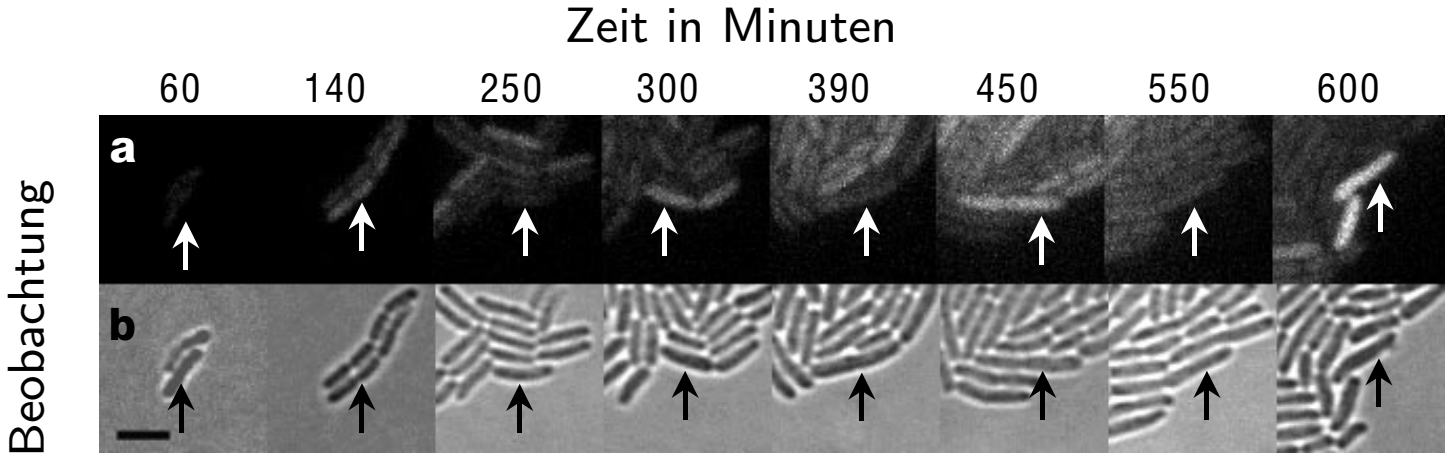
1. Gesund
   - $Wahr(\text{Hoch}) = 0.5$
   - $Wahr(\text{Mittel}) = 0.3$
   - $Wahr(\text{Niedrig}) = 0.2$
2. OK
   - $Wahr(\text{Hoch}) = 0.2$
   - $Wahr(\text{Mittel}) = 0.6$
   - $Wahr(\text{Niedrig}) = 0.2$
3. Krank
   - $Wahr(\text{Hoch}) = 0.2$
   - $Wahr(\text{Mittel}) = 0.3$
   - $Wahr(\text{Niedrig}) = 0.5$

Beobachtung einer Bakterienkultur

Übergangs-Wahrscheinlichkeiten
(ebenfalls erworben durch empirische Untersuchungen)



Beobachtung einer Bakterienkultur



Elowitz, Leibler, 2000, Nature 403

Wir machen über die Zeit also folgende Beobachtungen:
Niedrig, **Mittel**, **Mittel**, **Hoch**, **Mittel**, **Hoch**, **Mittel**, **Hoch**

Beobachtung einer Bakterienkultur

Wir können nun folgende Tabelle ausfüllen

Werte nach 6 Nachkommastellen abgeschnitten

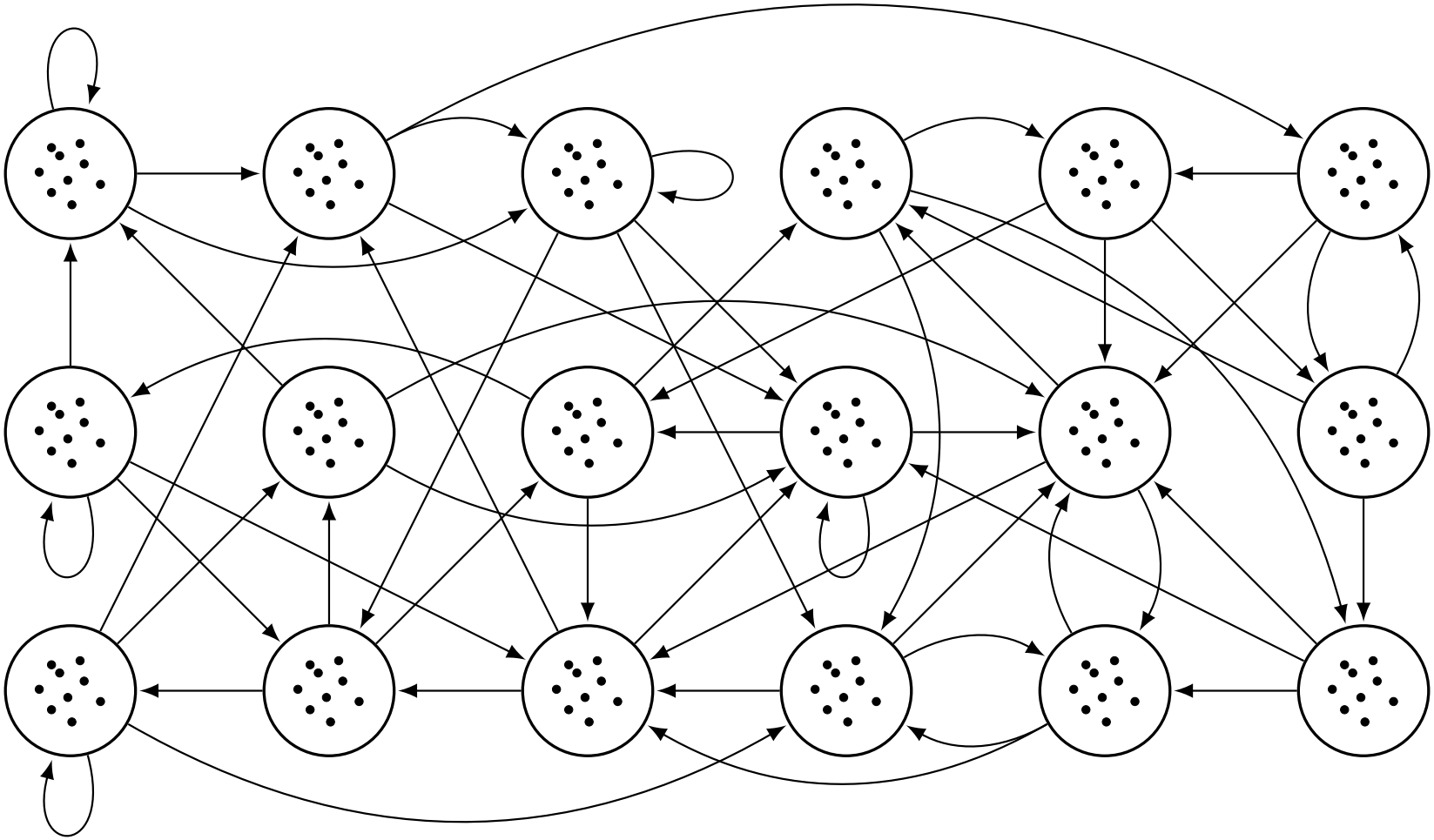
	Niedrig	**Mittel**	**Mittel**	**Hoch**	**Mittel**	**Hoch**	**Mittel**	**Hoch**
Gesund	$0.0\overline{6}$	0.008	0.0024	0.00144	0.000172	0.000062	0.000005	0.000001
OK	$0.0\overline{6}$	0.04	0.0144	0.001728	0.000622	0.000074	0.000002	0.000003
Krank	$0.1\overline{6}$	0.03	0.0054	0.000648	0.000129	0.000024	0.000007	0.000002

Laufzeit-Analyse

Dieser Algorithmus ist als **Viterbi**-Algorithmus bekannt.

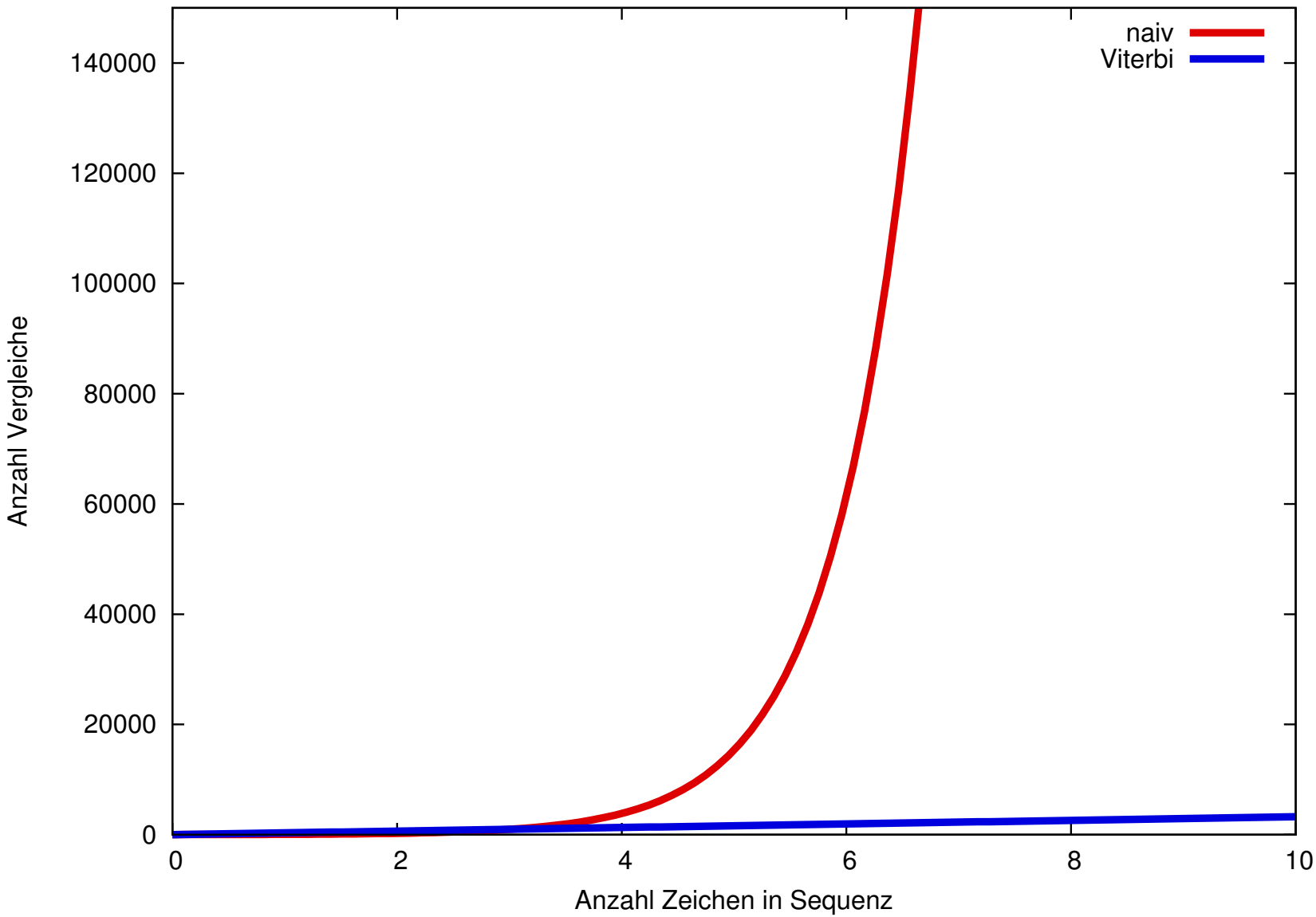
- Anzahl der Zustände: q
- Länge der beobachteten Sequenz: n
- Tabelle hat Grösse $q \times n$
- Für jede Zelle müssen q Werte verglichen werden
- Wir brauchen ca. $q \cdot q \cdot n$ Vergleiche
- „Die Laufzeit wächst proportional mit $q^2 \cdot n$“

Laufzeit-Analyse



- Hier ist $q = 18$; sei $n = 300$: Weniger als $\mathbf{100\,000}$ Vergleiche
- Naiver Ansatz: Mehr als $\mathbf{10^{143}}$ Vergleiche

Laufzeit-Analyse



CG-Inseln

Kodierende Bereiche der DNA

- DNA wird aufgefasst als String über den Buchstaben `A`, `C`, `G`, `T`
- Teile sind „kodierende Bereiche“ andere „steuernde Bereiche“
- *CG*-Inseln sind Bereiche im String, in denen die Folge `CG` häufig vorkommt
- Sie tauchen in der Nähe von Genen häufig auf, sonst selten
- Finde *CG*-Inseln in gegebenen String

Analogie zum vorher vorgestellten unfairen Kasino:

- Die jeweiligen Würfel entsprechen den Situationen, ob man in einer *CG*-Insel ist
- Es gibt aber für beide Situationen 4 verschiedene Beobachtungen

Zusammenfassung

1. **Wahrscheinlichkeiten**
   modellieren Situationen, wenn nicht alle Grössen bekannt sind
2. **Markov-Modelle**
   verknüpfen Zustände miteinander
3. **Beispiele**
   Beobachtungen im Labor, *CG*-Inseln
4. **Dynamische Programmierung**
   kann die Berechnungszeit verringern