

Statistische Analyse von DNA-Sequenzen

2.1 Grundidee

Viele Aspekte der Bioinformatik handeln davon, biologische Situationen in eine symbolische Sprache zu übertragen. In Kap. 1 haben wir mit der Reduktion des linearen Moleküls DNA auf die Abfolge der unterschiedlichen Basen und der Identifikation eines Proteins mit der Abfolge von Aminosäuren entlang des Polypeptidstrangs schon Beispiele für solche Übertragungen kennengelernt. Diese Reduktion der Komplexität realer biologischer Zusammenhänge bildet eine notwendige Voraussetzung für weitere Untersuchungen. Eine wichtige Fragestellung ist dabei, welche biologischen (oder chemischen) Eigenschaften durch diese Übertragung ausgeblendet werden, also einer Analyse auf der Grundlage der Symbolsequenz nicht mehr zugänglich sind, und – umgekehrt – welche biologischen Eigenschaften in der Symbolsequenz gerade einfacher diskutiert werden können. In den meisten Fällen wird die Verbindung von statistischen Eigenschaften der Symbolsequenzen und biologischen Eigenschaften des realen Objekts (z. B. des DNA-Strangs) durch *Wahrscheinlichkeitsmodelle* hergestellt. Solche auf statistischen Aussagen basierenden Funktionszuweisungen sind Gegenstand einer großen Klasse bioinformatischer Analysen. Abbildung 2.1 fasst diese Situation zusammen. Der Weg (reale Struktur → abstraktes, mathematisch zugängliches Objekt → Funktionszuweisung durch Analyse des abstrakten Objekts) erweist sich als universelles Arbeitsprinzip der Bioinformatik. Abbildung 2.2 zeigt, wie z. B. die Analyse genetischer Netzwerke mithilfe der Graphentheorie genau diesem Vorgehen entspricht. Auf Einzelheiten dieser speziellen Betrachtung werden wir in Kap. 5.2 noch ausführlich eingehen.

Ein mathematisches Modell hat in der Regel den Zweck, ein reales (beobachtetes) Phänomen numerisch zu simulieren, es also in Form mathematischer Regeln nachzustellen und es dadurch in seiner Funktion besser zu verstehen. Dabei lassen sich Einzelheiten des realen Systems so weit ausblenden, dass man versteht, welche Systemkomponenten oder -eigenschaften essenziell für bestimmte Aspekte des beobachteten Verhaltens sind (minimale Modelle). Ebenso lassen sich aus dem Mo-

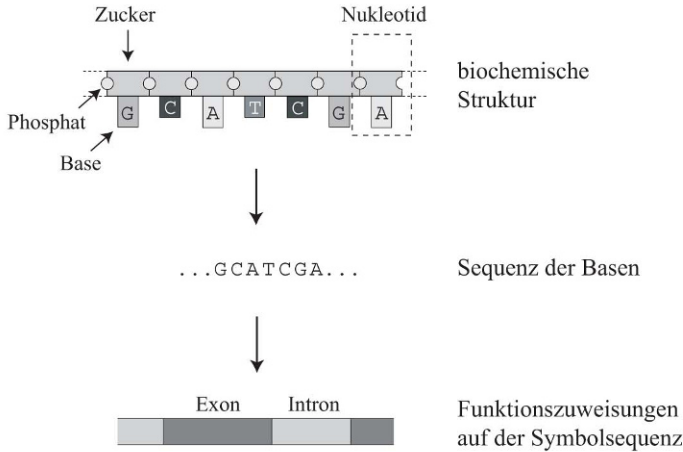


Abb. 2.1 Schematische Darstellung des durch die Begriffsabfolge Struktur, Abstraktion, Analyse gegebenen bioinformatischen Arbeitsprinzips am Beispiel einer DNA-Sequenz. Mit dem oberen Bildelement ist hier das Makromolekül selbst mit seinen vielfältigen biochemischen Eigenschaften gemeint. Damit ist klar, welche enorme Abstraktion in dem Schritt zur Symbolsequenz liegt. Der Weg von der Sequenz zur zugewiesenen Funktion führt dann oft über spezielle Modelle

dell oft Eigenschaften analytisch herleiten, so dass man eine Information darüber erhält, welche Aspekte des realen Phänomens zwangsläufig aus den (oft wenigen) im Modell festgelegten Grundeigenschaften folgen. Während die beiden letztgenannten Punkte üblicherweise Gegenstand der *theoretischen Biologie* sind, findet sich der erste Aspekt, die numerische Simulation, in einer Vielzahl bioinformatischer Untersuchungen. Die dabei verwendeten Modelle sind Wahrscheinlichkeitsmodelle. Ein Wahrscheinlichkeitsmodell erzeugt z. B. Sequenzen. Die (in dem Modell möglichen) Sequenzen werden im Allgemeinen mit unterschiedlichen Wahrscheinlichkeiten hervorgebracht, die von den Parametern des Modells abhängen. Dieses elementare Szenario werden wir in den Kapiteln 2.2 und 2.3 diskutieren. Wir werden im Folgenden anhand einfacher Beispiele sehen, wie man Wahrscheinlichkeitsmodelle formuliert, und vor allem, wie sich dann bei gegebener Modellstruktur die Modellparameter aus empirischen Befunden schätzen lassen. Die derzeit beste Darstellung, um von dieser pragmatischen und qualitativen Ebene zu einer mathematisch exakteren Ebene zu gelangen, ist das Buch von Durbin et al. (1998).

2.2 Wahrscheinlichkeitsmodelle

Einen Zugang zu dem abstrakten Begriff der Wahrscheinlichkeitsmodelle gewinnt man über Beispiele. Es ist klar, dass die für uns relevanten Objekte Sequenzen sind

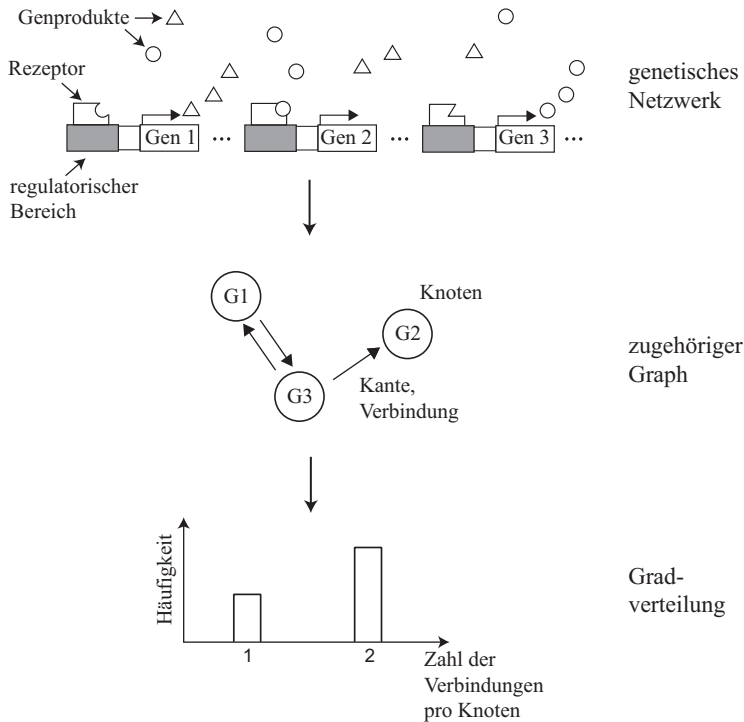


Abb. 2.2 Dasselbe durch Struktur, Abstraktion und Analyse gegebene Arbeitsprinzip wie in Abb. 2.1 hier für genetische Netzwerke. In diesem einfachen Schema reguliert Gen 3 über seine Produkte (Kreise) die Aktivität der Gene 1 und 2 und wird selbst von Gen 1 reguliert (Dreiecke). Damit gibt es in diesem formalen Graphen zwei Knoten (Gene) mit zwei Verbindungen und einen Knoten mit einer Verbindung. Diese Häufigkeitsverteilung ist im unteren Teil der Abbildung aufgetragen. Mit der Zahl der Verbindungen ist hier die *Summe* aus einlaufenden und herausgehenden Verbindungen für jeden Knoten gemeint

(für die dann z. B. funktionelle Bestandteile oder Aspekte der Molekülstruktur bestimmt werden sollen). Im Fall der DNA oder der Proteine sind dies dann Symbolsequenzen. In vielen anderen Beispielen stellen diese Sequenzen eine Abfolge von Zahlen dar.

Das bekannteste Beispiel eines Wahrscheinlichkeitsmodells ist ein Würfel. Die Modellparameter sind dann die Wahrscheinlichkeiten p_i , die Zahl i zu würfeln. Implizit sind wir damit vom klassischen Würfel, der eine Gleichverteilung aller sechs Zahlen besitzt, abgewichen und haben faire und unfaire Würfel gleichermaßen zugelassen. Ebenso könnten wir von den möglichen Zahlen $i = 1, \dots, 6$ abweichen und ein Modell eines Würfels mit k Zahlen formulieren. Um die Notation einfach zu halten, wollen wir an der Beschränkung auf sechs Zahlen jedoch festhalten. Ein Wahrscheinlichkeitsmodell (in unserem Sinne) bringt also Zahlen und Symbole hervor. Die Menge

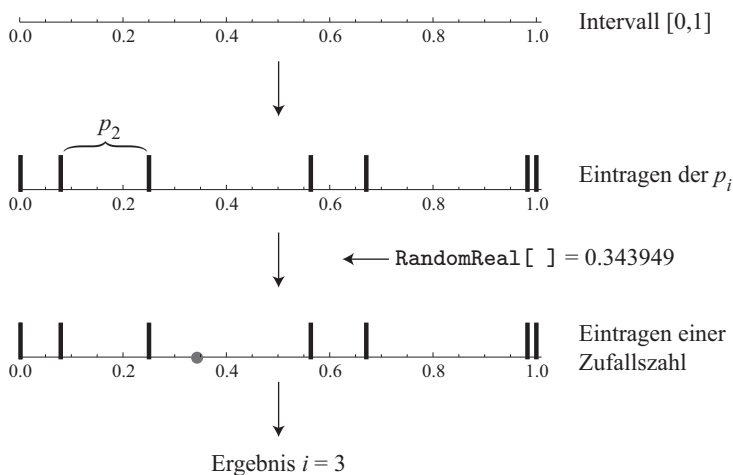


Abb. 2.3 Die Zerlegung des Einheitsintervalls als Grundgedanke diskreter Wahrscheinlichkeitsverteilungen. Die Intervallgrößen geben die Wahrscheinlichkeit für jedes (diskrete) Ereignis an. Ein kontinuierliches Zufallsereignis kann so eingeordnet und in ein entsprechendes diskretes Zufallsereignis übersetzt werden. Im dargestellten Beispiel gelangt die Zufallszahl `RandomReal [ ]` in das fünfte Intervall. Das Ergebnis ist also $i = 3$

der möglichen Zahlen oder Symbole, die ein Wahrscheinlichkeitsmodell erzeugen kann, bezeichnet man auch als *Zustandsraum* des Wahrscheinlichkeitsmodells. Bei solchen Parametern eines Modells handelt es sich um *Wahrscheinlichkeiten*, wenn die Parameter noch zwei Nebenbedingungen erfüllen: Zum einen dürfen die p_i nicht negativ sein, zum anderen muss die Summe über alle p_i gerade Eins ergeben,

$$p_i \geq 0 ; \quad \sum_{i=1}^6 p_i = 1 , \quad i = 1, \dots, 6 . \quad (2.1)$$

Damit ist gesagt, dass als sicheres Ereignis, also mit der Wahrscheinlichkeit Eins, eines der Symbole aus dem Zustandsraum des Wahrscheinlichkeitsmodells tatsächlich als nächstes Symbol auftritt. Diese Vollständigkeitsbedingung stellt die erste mathematisch interessante Eigenschaft solcher Wahrscheinlichkeitsmodelle dar, denn je nach Perspektive hat sie die Gestalt einer Normierungsbedingung oder aber die Form der Zerlegung des Einheitsintervalls $[0, 1]$ durch die Wahrscheinlichkeitsverteilung $p_1, \dots, p_6$. Der Aspekt der Normierungsbedingung ist unmittelbar zu verstehen, wenn man sich die Schätzung der Parameter p_i aus den Häufigkeiten $N_1, \dots, N_6$ der in einer Symbolsequenz tatsächlich vorliegenden Ereignisse vorstellt. Dabei gibt N_i die absolute Häufigkeit des Symbols i in einer durch den Würfel (also unser Wahrscheinlichkeitsmodell) hervorgebrachten Sequenz an. Verwandelt man die absoluten

Häufigkeiten N_i nun in relative Häufigkeiten r_i , indem man durch die Gesamtzahl von Symbolen in der Sequenz teilt,

$$p_i \approx r_i = \frac{N_i}{\sum_{j=1}^6 N_j}, \quad i = 1, \dots, 6, \quad (2.2)$$

so erhält man eine Approximation¹ an die Wahrscheinlichkeiten p_i . Dass die relativen Häufigkeiten r_i (unter ziemlich schwachen Modellannahmen) im Limes unendlich vieler Stichproben ($n \rightarrow \infty$ mit $n = \sum_{j=1}^6 N_j$) gegen die Größen p_i konvergieren, folgt aus dem *Gesetz der großen Zahlen*, einem Grundpfeiler der Wahrscheinlichkeitstheorie. Gerade aufgrund der Normierung beim Schritt zu den relativen Häufigkeiten erfüllen die so geschätzten p_i die oben aufgeführte Nebenbedingung *per constructionem*. In all diesen Zusammenhängen geht man typischerweise davon aus, dass die Stichproben (z. B. die hintereinander ausgeführten Würfelwürfe) unabhängig voneinander sind (also insbesondere nicht von ihrer Reihenfolge abhängen). Die lückenlose Aufteilung des Einheitsintervalls $[0, 1]$ durch die in dem Modell auftretenden Wahrscheinlichkeiten, die Zerlegung der Einheit, ist die zweite mögliche Sichtweise auf diese Nebenbedingung. Die Größe eines Unterintervalls ist dabei gerade durch den Wert der Wahrscheinlichkeit gegeben. Eine im Intervall $[0, 1]$ gleichverteilte reelle Zufallszahl befindet sich also in dieser Vorstellung in einem der Unterintervalle (z. B.: im i ten Unterintervall). Das hervorgebrachte Symbol ist somit gerade i . Zufallszahlen (und die praktischen Schwierigkeiten ihrer Erzeugung) haben wir bereits in Kap. 1.5 kurz angesprochen. Vollführt man dieses Zufallsexperiment oft hintereinander, so ist über die Intervallgröße sichergestellt, dass die relativen Häufigkeiten N_i mit einer immer größeren Sequenz von Zahlen (größere Stichprobe) immer besser die Wahrscheinlichkeit p_i approximieren. Abbildung 2.3 fasst diese Situation zusammen. Die grafische Vorstellung der Zerlegung der Einheit durch eine diskrete Wahrscheinlichkeitsverteilung $p_1, \dots, p_6$ ist ein nützliches Gedankenmodell für das Funktionieren von Wahrscheinlichkeitsverteilungen und zugleich ein geeigneter gedanklicher Ausgangspunkt für die konkrete informatische Implementierung solcher Zufallsprozesse. Sie stellt daher auch das Thema unseres ersten *Mathematica*-Exkurses dar. Zur weiteren Ausgestaltung des durch die p_i gegebenen Würfels als Wahrscheinlichkeitsmodell kann man nun *Sequenzen* von Zahlen diskutieren, also Abfolgen von Würfelergebnissen. Dies ist hier besonders einfach durch die bereits erwähnte Annahme, dass die Würfe alle unabhängig voneinander erfolgen. Man hat dann als (Gesamt-)Wahrscheinlichkeit z. B. für eine Sequenz $x = 2, 6, 5$ das *Produkt* der individuellen Wahrscheinlichkeiten:

$$P(x) = p_2 p_6 p_5. \quad (2.3)$$

Diese Annahme einer Unabhängigkeit der Ereignisse definiert ein Wahrscheinlichkeitsmodell, das wir im Folgenden häufig zum direkten Vergleich mit realen Sequenzen heranziehen werden, nämlich das Modell einer rein *zufälligen Symbolsequenz* (engl. *random sequence model*, auch Bernoulli-Sequenz).

¹ In der Statistik bezeichnet man eine solche Approximation einer Wahrscheinlichkeit (also eines Modellparameters) durch empirische Größen als *Schätzer*.

Im Fall von DNA-Sequenzen hat man den aus den möglichen Nukleotiden gebildeten Zustandsraum

$$\Sigma = \{A, G, C, T\}. \quad (2.4)$$

Die Parameter des Modells sind dann die Wahrscheinlichkeiten p_a für das Vorliegen eines beliebigen Symbols $a \in \Sigma$ in der Sequenz.² Erneut wird das Modell ergänzt durch die Annahme, dass die Zufallsexperimente unabhängig sind. Betrachten wir eine Sequenz $x = x_1, x_2, \dots, x_n$. Die Wahrscheinlichkeit $P(x)$ der Sequenz x in diesem Modell ist dann

$$P(x) = p_{x_1} p_{x_2} \cdots p_{x_n} = \prod_{i=1}^n p_{x_i}. \quad (2.5)$$

Die Bernoulli-Sequenz ist eine Art Minimal- oder Grundmodell, das man verwendet, um Aussagen aus anderen, fortgeschrittenen Modellen mit den Ergebnissen dieses Grundmodells zu vergleichen. Es dient als eine *Nullhypothese* der Sequenzmodelle.

Mathematica-Exkurs: Zerlegung der Einheit

Im Rahmen dieses Exkurses soll das prinzipielle Phänomen der Zerlegung des Einheitsintervalls durch eine diskrete Wahrscheinlichkeitsverteilung und die in Abb. 2.3 dargestellte Abfolge von Schritten zur praktischen Verwendung dieser Vorstellung in *Mathematica* rekonstruiert werden.

Das erste Element auf diesem Weg ist eine geeignete Wahrscheinlichkeitsverteilung. Dazu wird mit der Funktion **RandomReal**[], die eine im Intervall [0, 1] gleichverteilte reellwertige Zufallszahl erzeugt, der Ausgangspunkt gelegt:

```
In[1] := rand = Table[RandomReal[], {6}]
Out[1] = {0.817389, 0.11142, 0.789526, 0.187803, 0.241361, 0.0657388}
```

Diese Tabelle aus sechs Zufallszahlen lässt sich nun so normieren, dass die Summe gerade Eins ist:

```
In[2] := probDist = rand/Apply[Plus, rand]
Out[2] = {0.369318, 0.0503424, 0.356729,
          0.0848545, 0.109053, 0.0297025}
```

Die Funktion **Plus** gibt die Summe ihrer Argumente aus, **Plus[x1, x2, x3]** = $x_1 + x_2 + x_3$. Um die Funktion **Plus** auf eine Liste von Elementen anzuwenden, ist der Befehl **Apply** nötig. **Plus[{x1, x2, x3}]** würde das Argument unausgewertet

² Aus unserer Diskussion der relativen Häufigkeiten ist klar, dass diese Größen von der Länge der betrachteten Sequenz abhängen. Diesen Aspekt betrachten wir hier nicht weiter. Im praktischen Umgang mit Sequenzmodellen benötigt man zur Parameterschätzung eine geeignete Menge von *Trainingsdaten*.

lassen, erst **Apply[Plus, {x1, x2, x3}]** führt auf $x_1 + x_2 + x_3$. Mit dem Anwenden der Funktion **Plus** auf die resultierende Liste,

```
In[3] := Apply[Plus, probDist]
Out[3] = 1.
```

lässt sich diese Normierung überprüfen. Eine Abkürzung dieses **Apply**-Befehls ist durch **@@** gegeben:

```
In[4] := Plus@@probDist
Out[4] = 1.
```

Damit sind die einzelnen Intervallgrößen festgelegt. Durch Bildung der Partialsummen

$$q_i = \sum_{j=1}^i p_j$$

erhält man daraus die Intervallgrenzen im Einheitsintervall $[0, 1]$. In *Mathematica* sind Partialsummen effizient zu formulieren, indem man die Funktion **Plus** iterativ auf eine Liste anwendet. Dazu verwenden wir die Funktion **FoldList**. Diese Funktion besitzt drei Argumente, nämlich eine zweiaargumentige Funktion f , das Anfangselement A und eine Liste $L = \{L_1, \dots, L_n\}$. Dann hat man

$$\mathbf{FoldList}[f, A, L] = \{A, f[A, L_1], f[f[A, L_1], L_2], \dots\}.$$

Die Iteration endet, wenn das letzte Listenelement erreicht ist. In unserem Fall bildet die untere Intervallgrenze, $A = 0$, den Anfangspunkt. Man hat also

```
In[5] := cumDist = FoldList[Plus, 0, probDist]
Out[5] = {0, 0.369318, 0.419661, 0.77639, 0.861244, 0.970297, 1.}
```

In eine grafische Darstellung des Einheitsintervalls,

```
In[6] := pl1 = Plot[0, {x, 0, 1}, Axes → {True, False},
PlotRange → {{0, 1.01}, {-1, 1}}]
```

lassen sich nun diese Intervallgrenzen einzeichnen:

```
In[7] := Do[
  plX[k] = ListLinePlot[
    {{cumDist[[k]], 0}, {cumDist[[k]], 0.2}},
    PlotStyle → {AbsoluteThickness[3], Black}],
  {k, 1, Length[cumDist]}]
```

Die **Do**-Schleife ist ein zentrales Element der Programmierung in *Mathematica*. Die erste geschweifte Klammer enthält die Abfolge von Befehlen, die in jedem Schleifenschritt ausgeführt werden. In unserem Beispiel enthält sie nur einen einzigen Befehl. Die zweite Klammer legt die Schleifenvariable (hier: **k**) und ihren Anfangs-

und Endwert in der Schleife fest. Es ist deutlich zu sehen, dass in den Befehlen (erste Klammer) diese Variable verwendet werden kann.

```
In[8] := p100 =
      Show[
        Flatten[{p11, Table[p1X[k], {k, 1, Length[cumDist]}]}]]
```

Die Ausführung der beiden letztgenannten Befehle ergibt die ersten Teile der Abb. 2.3. Die hier verwendeten Grafikoptionen eignen sich besonders für eigene kleine Experimente mit den *Mathematica*-Befehlen. Eine Zufallszahl

```
In[9] := x = RandomReal[]
Out[9] = 0.542247
```

kann nun in diese Abbildung eingetragen werden:

```
In[10] := p12 = ListPlot[{{x, 0}},
      PlotStyle → {AbsolutePointSize[8], Red}];
      Show[p100, p12]
```

Das zugehörige Ergebnis, also die Intervallnummer oder – in dem Gedankenexperiment dieses Kapitels – die von dem Würfel, der dieser Wahrscheinlichkeitsverteilung folgt, ausgegebene Zahl, kann man nun über einen kleinen Umweg ermitteln. Dazu fügt man die Zahl **x** der Liste von Intervallgrenzen hinzu (mit **Append**), sortiert die Einträge dieser Liste nach ihrer Größe (mit **Sort**) und lässt sich die Position von **x** in der sortierten Liste ausgeben (mit **Position**). Subtrahiert man 1 (sonst würden die möglichen Ergebnisse zwischen 2 und 7 liegen), so erhält man die Intervallnummer:

```
In[11] := Position[Sort[Append[cumDist, x]], x][[1, 1]] - 1
Out[11] = 3
```

Damit haben wir Abb. 2.3 vollständig in *Mathematica* reproduziert. Zugleich haben wir ein mathematisch nicht triviales Objekt implementiert, nämlich eine Routine, die eine zwischen Null und Eins gleichverteilte kontinuierliche Zufallszahl in eine diskrete Zufallszahl überführt, die einer vorgegebenen Wahrscheinlichkeitsverteilung folgt. Diese Aussage können wir nun überprüfen. Abbildung 2.4 gibt eine Abfolge von zehn Zufallsexperimenten mit jeweils zehn Einträgen an. Die erheblichen Unterschiede zwischen den Punktverteilungen zeigen deutlich, dass zehn Wiederholungen keinesfalls ausreichen, um die zugrunde liegende Wahrscheinlichkeitsverteilung angemessen abzubilden. Abbildung 2.5 stellt drei Häufigkeitsverteilungen für unterschiedliche Stichprobengrößen sowie die tatsächliche Verteilung dar. Der optische Eindruck aus Abb. 2.4 bestätigt sich: Mit nur zehn Wiederholungen des Einzelereignisses lassen sich die Intervallgrößen nur unzureichend nachzeichnen. Mit wachsender Zahl der Wiederholungen (Stichprobengröße) approximieren die relativen Häufigkeiten jedoch immer besser die Intervallgrenzen (Abb. 2.5d).

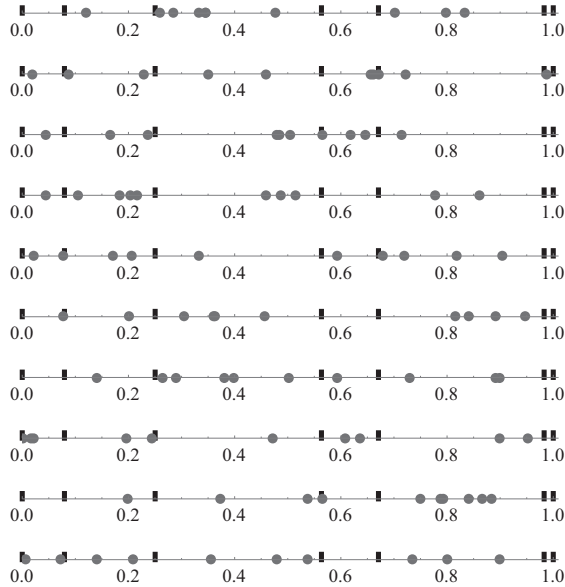


Abb. 2.4 Zehn Zufallsexperimente gemäß dem in Abb. 2.3 dargestellten Schema. Jedes Zufallsexperiment (also jede Zeile) besteht aus zehn Einzelereignissen. Man erkennt die bei dieser geringen Stichprobengröße starken Schwankungen in der Ausnutzung des (diskreten) Zustandsraums

2.3 Bedingte Wahrscheinlichkeiten

Wir haben nun an einfachen Beispielen gesehen, wie man Wahrscheinlichkeitsmodelle prinzipiell formuliert. In diesem Kapitel werden wir diskutieren, wie man solche Modelle praktisch verwendet und – vor allem – wie man sich zwischen konkurrierenden Modellen³ entscheidet. Dazu müssen Datensätze gegeben sein (also Sequenzen; das können DNA-Sequenzen in einer Datenbank sein oder Würfelerggebnisse, die in einer Datei abgelegt sind), um die Parameter zu bestimmen. Das wahrscheinlichkeitstheoretische Werkzeug, um solche Fragestellungen zu beantworten, ist die *bedingte Wahrscheinlichkeit*. Es wird sich im Folgenden auch als zentral bei der Formulierung fortgeschrittenerer Wahrscheinlichkeitsmodelle (z. B. für eine DNA-Sequenz) als eine einfache Bernoulli-Sequenz herausstellen. Wir betrachten zwei unterschiedliche Würfel $\mathcal{W}_1$ und $\mathcal{W}_2$. Die Wahrscheinlichkeit, mit $\mathcal{W}_1$ die Zahl i ,

³ An dieser Stelle meinen wir mit ‘Modelle’ zwei Versionen desselben Wahrscheinlichkeitsmodells, die durch unterschiedliche Parameter ausgestaltet werden. Das ist für die vorliegende Diskussion ein angemessenes Vorgehen, später jedoch nicht mehr. Ein praktikablerer (und interessanterer) Modellbegriff ergibt sich, wenn Varianten, die sich nur in ihren Parametern unterscheiden, immer noch als *ein Modell* bezeichnet werden. Beide Vorstellungen findet man in bioinformatischen Untersuchungen. Die entsprechende Untersuchungsstrategie werden wir im Verlauf dieses Kapitels noch ausführlich diskutieren.

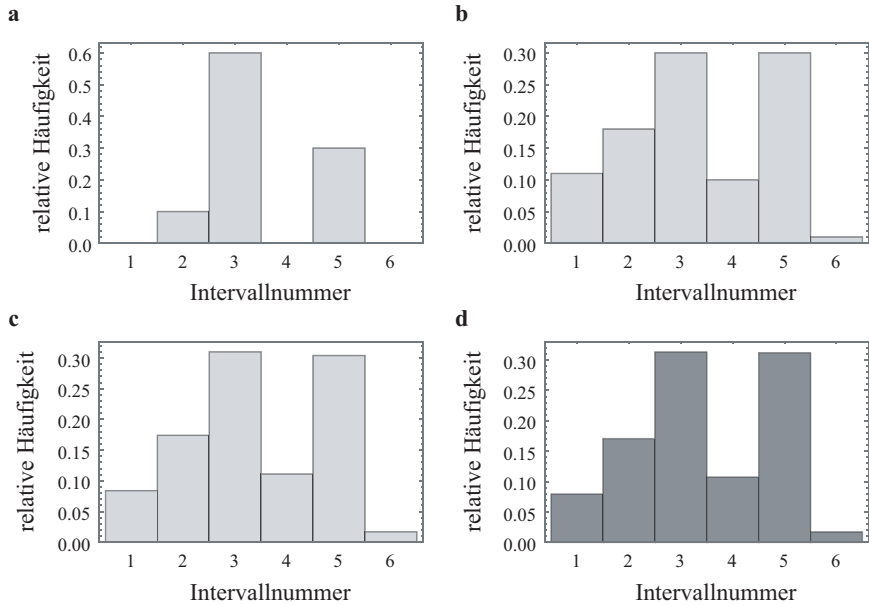


Abb. 2.5 Approximation der tatsächlichen (durch die Intervallbreiten in der Versuchsanordnung aus Abb. 2.3 gegebenen) Wahrscheinlichkeiten durch die relativen Häufigkeiten für verschiedene Stichprobengrößen der Zufallsexperimente: **a** $n = 10$, **b** $n = 100$, **c** $n = 1000$, **d** Darstellung der tatsächlichen Wahrscheinlichkeiten

mit $i \in \{1, \dots, 6\}$, zu würfeln, ist dann die bedingte Wahrscheinlichkeit $P(i | W_1)$, also die Wahrscheinlichkeit für die Zahl i unter der Bedingung des Würfels W_1 .

Zu einer Beziehung zu den üblichen Wahrscheinlichkeiten gelangt man, wenn man eine zufällige Auswahl des Würfels zulässt. Dann gibt es weitere Parameter $P(W_j)$, $j = 1, 2$, als Wahrscheinlichkeit für die Verwendung des Würfels W_j . Damit ist die gemeinsame (bzw. verknüpfte) Wahrscheinlichkeit $P(i, W_j)$ dafür, den Würfel W_j zu wählen und mit ihm die Zahl i zu würfeln:

$$P(i, W_j) = P(i | W_j) P(W_j). \quad (2.6)$$

Gleichung (2.6) ist die übliche mathematische Präzisierung unserer Vorstellung einer bedingten Wahrscheinlichkeit. Für zwei Zufallsereignisse X und Y wird so ein Zusammenhang zwischen der bedingten und der gemeinsamen Wahrscheinlichkeit hergestellt:

$$P(X, Y) = P(X | Y) P(Y). \quad (2.7)$$

Dabei ist $P(X, Y)$ die Wahrscheinlichkeit für das gemeinsame Auftreten von X und Y , $P(X | Y)$ die Wahrscheinlichkeit für X unter der Bedingung, dass Y vorliegt, und $P(Y)$ die Wahrscheinlichkeit für Y . Aus der gemeinsamen Wahrscheinlichkeit kann

Methoden der Bioinformatik

Eine Einführung zur Anwendung in Biologie und Medizin

Hütt, M.-T.; Dehnert, M.

2016, XVII, 406 S. 217 Abb., Softcover

ISBN: 978-3-662-46149-5