Seminararbeit: Modeling DNA Sequences

Lisa Klein-Schmeink

29. März 2012

Inhaltsverzeichnis

1 Einleitung

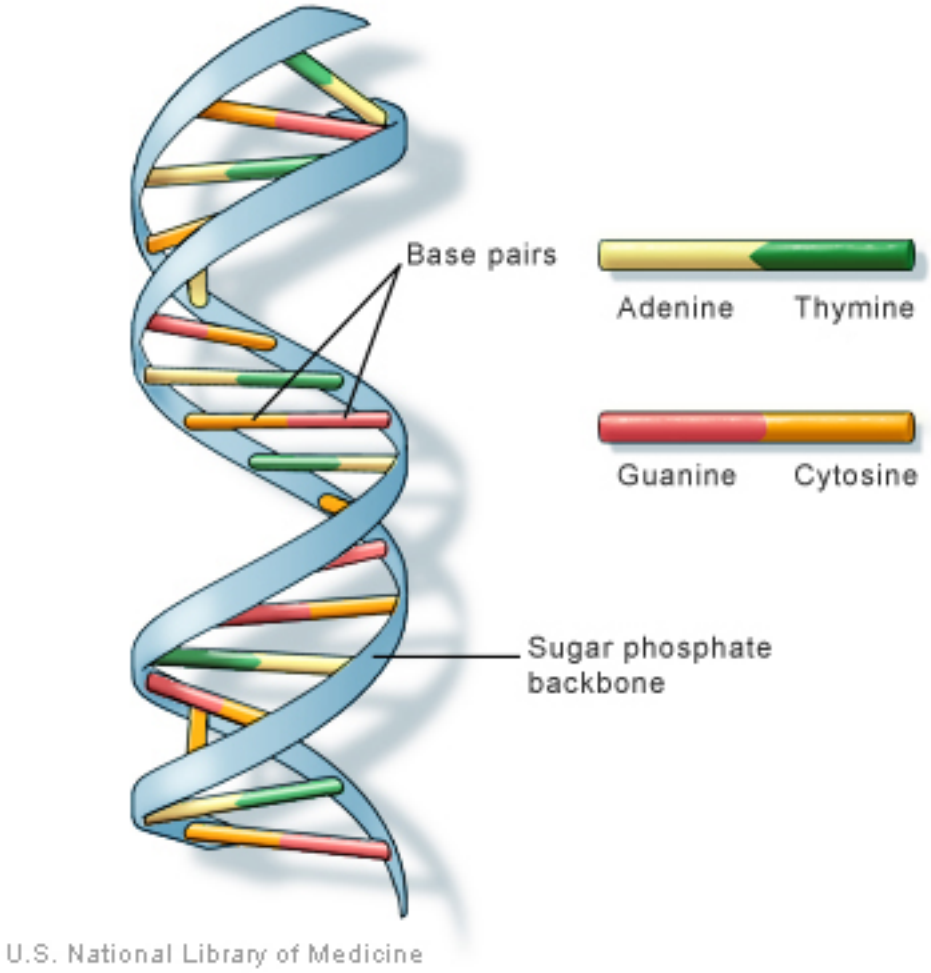


Abbildung 1: DNA-Struktur

Dieser Seminarvortrag gibt eine Einführung in die Modellierung von DNA-Sequenzen auf Basis von „Statistical Methods in Bioinformatics: An Introduction“(Ewans, Grant, 2005). Bevor man diese analysieren kann, muss man zunächst natürlich verstehen, was sich hinter einer DNA-Sequenz tatsächlich verbirgt.

DNA-Sequenzen (*Desoxyribonukleinsäure*) sind Moleküle, die einen Hauptbestandteil der Zellkerne aller Lebewesen (der *Eukaryoten*) darstellen und als Träger von Erbinformationen dienen. DNA-Sequenzen sind Nukleinsäuren, die aus Nukleotiden aufgebaut sind. Diese bestehen aus einer Phosphatgruppe, einem Zucker (Desoxyribose bei DNA) und einer stickstoff-

haltigen Base – Adenin (**A**), Cytosin (**C**), Guanin (**G**) oder Thymin (**T**). Diese bilden das DNA-Alphabet $\{A, C, G, T\}$. Dabei liegt eine komplementäre Basenpaarung vor: Adenin ist verbunden mit Thymin (A-T) und Cytosin mit Guanin (C-G). Ein DNA-Molekül hat die Form einer Doppelhelix. Die Längsstränge dieser verdrehten Strickleiter bestehen aus sich abwechselnden Desoxribose- und Phosphatmolekülen. Die beiden Stränge verlaufen in entgegengesetzter Richtung, so dass ein 3'-Ende eines Stranges einem 5'-Ende gegenüber liegt. Daher wird auch nur ein Strang einer DNA-Sequenz angegeben, um diese zu charakterisieren.

Meist betrachtet man die *proteinkodierenden Gene*, das sind die Teilstücke eines DNA-Strangs, durch die Proteine erstellt werden können. Proteine sind Makromoleküle, die aus den 20 verschiedenen Aminosäuren bestehen. Der Hauptbestandteil dieser proteinkodierenden Gene sind Nukleobasen-Tripletts *(Codons)*, die jeweils eine Aminosäure kodieren. Der *genetische Code* ist die Art, wie die Codons in Aminosäuren übersetzt werden. Ein solches Gen beginnt mit einem *Start-Codon* und endet mit einem *Stop-* bzw. *Nonsense-Codon*.

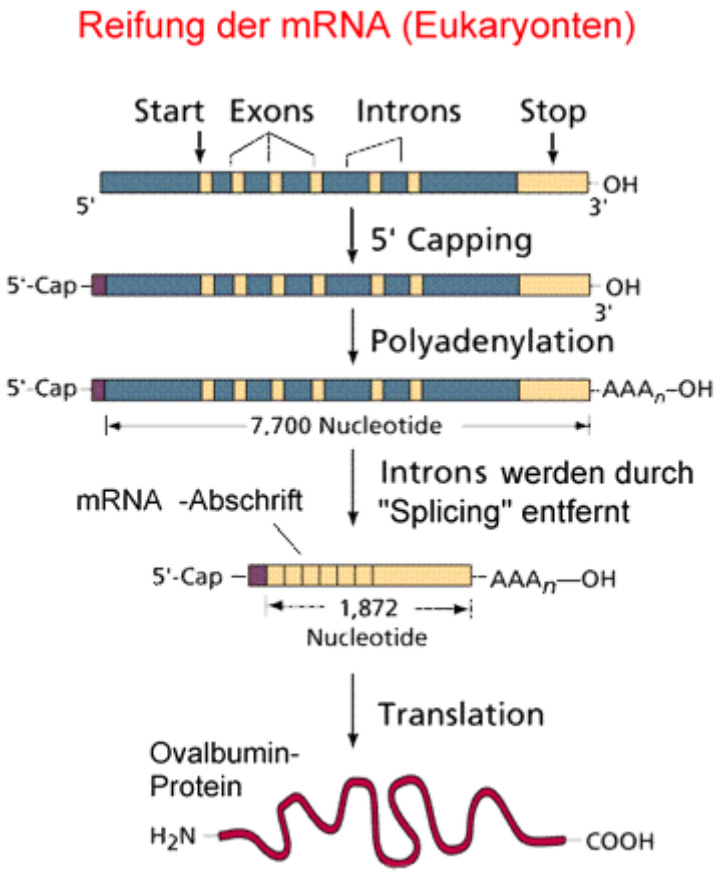


Abbildung 2: Schema der Proteinsynthese

Proteine werden aus Genen gewonnen, indem sie zunächst durch *Transkription* in RNA umgewandelt wird. Dabei werden nur die kodierenden Bereiche, die *Exons* (für *expressed region*),

benutzt während die nicht-kodierenden Regionen, die *Introns* (für *intervening region*), durch die sie unterbrochen werden, aus der Sequenz herausgespleißt werden. RNA *(Ribonukleinsäure)* sind einzelsträngige Ketten von Nukleotiden ähnlich der DNA, die als Zucker Ribose verwenden und bei dem die Base Thymin durch Uracil **(U)** ersetzt sind. Die durch die Synthese anhand von DNA gewonnene RNA wird auch Boten-RNA oder *mRNA* (für *messenger RNA*) genannt. Im zweiten Schritt, der *Translation*, wird die mRNA benutzt, um ein Protein zu generieren.

2 Methoden zur Analyse einer DNA-Sequenz?

Bei der Analyse von uncharakterisierten DNA-Sequenzen interessiert man sich häufig dafür, ob diese in einer kodierenden Region liegen. Introns und intergenische Bereiche weisen andere Eigenschaften als Exons auf, so dass z.B. unter der simpelsten Annahme, der unabhängigen und identischen Verteilung *(i.i.d.)* der Nukleotiden, die Auftrittshäufigkeiten der einzelnen Nukleotiden in Exons bzw. Introns betrachtet werden. Dafür ermittelt man aus schon charakterisierten Sequenzen, sogenannten Trainingsdaten, die Verteilung der Nukleotiden. Man testet die Unabhängigkeits-Hypothese mit Hilfe von Markov-Ketten gegen die Alternative, dass die Nukleotiden Markov-abhängig sind, d.h. dass die Auftrittswahrscheinlichkeit eines Nukleotides nur von der vorhergehenden Nukleotiden abhängt, bzw. gegen Markov-Abhängigkeit höherer Ordnung bzw. nichtlinearer.

Definition 1. *Eine stochastische Folge $M = (M_n)_{n\geq 0}$, also eine Folge von Zufallsexperimenten, mit Zustandsraum $(S, \mathfrak{S})$ heißt Markov-Kette, wenn sie die Markov-Eigenschaft besitzt, d.h.*

$$\mathbb{P}(M_{n+1} \in A | M_0, \ldots, M_n) = \mathbb{P}(M_{n+1} \in A | M_n) \quad \textit{für alle } n \geq 0 \textit{ und alle } A \in \mathfrak{S}. \tag{1}$$

Die stochastische Folge heißt Markov-Kette k-ter Ordnung, wenn sie folgender Eigenschaft genügt

$$\mathbb{P}(M_{n+1} \in A | M_0, \cdots, M_n) = \mathbb{P}(M_{n+1} \in A | M_{n-t+1}, \ldots, M_n) \tag{2}$$

für alle $n \geq 0$ und alle $A \in \mathfrak{S}$. Die Wahrscheinlichkeit für M_{n+1} hängt hier also von den k vorherigen M_i ab.

In den meisten Fällen ergibt ein solcher Assoziationstest, dass die Markov-Modelle die Realität weit aus besser widerspiegeln, wobei auch sowohl Hypothese und Alternative falsch seien können.

In der Praxis testet man ob die betrachtete DNA-Sequenz zu einem *Signal* gehört.

Definition 2. *Ein **Signal** ist eine kurze Sequenz, die einem spezifischen Zweck, wie eine Grenze zwischen Exon und Intron zu definieren, dient. Dies ermöglicht, dass Gene erkannt und in Proteine umgewandelt werden können. Dabei existieren oft viele **Mitglieder** eines Signals, also DNA-Sequenzen, die die gleiche Funktion haben.*

Meist sind nicht alle Mitglieder bekannt, so dass diejenigen, die bekannt sind, benutzt werden, um festzustellen, ob die betrachtete Sequenz auch zu dem Signal gehört. Außerdem kommen sie manchmal auch zufällig in nichtkodierenden Regionen vor. Man schränkt das Modell auf die empirisch meist begründete Annahme ein, dass Mitglieder eines Signals durchgängig die gleiche Länge *n* haben.

3 Lange Ketten

Eine weitere Möglichkeit besteht darin, die Hypothese, dass ein Nukleotid zufällig auftritt gegen die Alternative, dass eine Tendenz für ein wiederholtes Auftreten dieses Nukleotiden in Folge besteht, zu testen.

Definition 3. *Definiere unter der Zufälligkeits-Hypothese die Zufallsvariable* $\mathbf{Y}$ *als die Länge einer Kette der Nukleotiden* a *in einer DNA-Sequenz der groß genügenden Länge* N *nach dem letzten Auftreten eines anderen Nukleotiden.*

Da dies der Anzahl der „Erfolge“ nach einem „Misserfolg“entspricht, hat sie eine geometrische Verteilung mit einem bekannten Parameter p. Die Wahrscheinlichkeit, dass a an einem beliebigen Ort in der Sequenz auftritt ist also:

$$P_Y(y) = \mathbb{P}(Y = y) = (1-p)p^y \tag{3}$$

Definition 4. *Unter einer Ordnungsstatistik* $X_{(1)}, \ldots, X_{(n)}$ *versteht man die Ordnung der Realisationen* $x_1, \ldots, x_n$ *einer Zufallsvariablen* X *der Größe nach.* $X_{(i)}$ *bezeichnet die* i*-te Ordnungsstatistik. Insbesondere sind damit* X_{max} *und* X_{min} *die* n*-te bzw. die* 1*-te Ordnungsstatistik.*

Satz 5. *Für* n *solche Sequenzen folgt dann für die längste unter diesen Sequenzen, der Ordnungsstatistik* Y_{max}*:*

$$\mathbb{P}(Y_{max} \geq y) = 1 - (1-p^y)^n \tag{4}$$

Beweis. Sei $F_{Y_{max}}(y)$ die Verteilungsfunktion von Y_{max}, dann gilt

$$\mathbb{P}(Y_{max} \geq y) = 1 - \mathbb{P}(Y_{max} \leq y-1) = 1 - (F_{Y_{max}}(y))^n = 1 - (1-p^y)^n$$

□

Um den Wert in (4) zu bestimmen, muss die Anzahl dieser Sequenzen n approximiert werden. Im folgenden sei unter $\approx$ ein gerundetes Ergebnis und unter $\simeq$ die asymptotische Annäherung (d.h. für $n \to \infty$) gemeint.

Lemma 6. *Man geht unter der Hypothese davon aus, dass die Misserfolge binomialverteilt sind, also lässt sich die Wahrscheinlichkeit, dass* Y_{max} *größer als ein beobachteter Wert* y_{max} *ist, wie folgt abschätzen:*

$$\mathbb{P}(Y_{max} \geq y_{max}) \approx 1 - (1-p^{y_{max}})^{(1-p)N} \tag{5}$$

Beweis. Da man von einer Binomialverteilung ausgeht, wird erwartet, dass $(1-p)N$ Misserfolge und, da Ketten von Erfolgen (wobei diese auch die Länge 0 haben dürfen) sich einem Misserfolg anschließen, ebenso auch $(1-p)N$ Erfolge auftreten. Damit erhält man das gewünschte Ergebnis für eine maximale beobachtete Länge y_{max}. □

Korolar 7. *Wenn* $(1-p)Np^{y_{max}} \leq 1$ *ist, gilt sogar*

$$\mathbb{P}(Y_{max} \geq y_{max}) \approx 1 - \mathrm{e}^{-(1-p)Np^{y_{max}}} \tag{6}$$

Beweis. Da für große n und $|t| \leq 1$ gilt, dass $\left(1+\frac{t}{n}\right)^n \simeq \mathrm{e}^n$, folgt für $(1-p)Np^{y_{max}} \leq 1$:

$$\mathbb{P}(Y_{max} \geq y_{max}) \approx 1-(1-p^{y_{max}})^{(1-p)N} \simeq 1-\mathrm{e}^{-(1-p)Np^{y_{max}}}$$

□

Im folgenden Beispiel erkennt man, dass diese Abschätzungen gute Ergebnisse liefern:

Beispiel 8. *Sei $N = 100,000$, $p = \frac{1}{4}$, dann ergibt Gleichung* (5) *ungefähr* 0.0690272 *und* (6) 0.0690275. *Auch die Annahme, dass die Misserfolge binomialverteilt sind, lässt sich hiermit nachvollziehen. Für $1-p = \frac{3}{4}$ und $Y_{max} \geq 10$ folgt mit dem binomischen Lehrsatz $(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}$:*

$$\begin{aligned}
\mathbb{P}(Y_{max} \geq 10) &= \sum_{j=0}^{100,000}\binom{100,000}{j}\left(\frac{3}{4}\right)^j\left(\frac{1}{4}\right)^{100,000-j}\left(1-\left(1-\left(\frac{1}{4}\right)^{10}\right)^j\right)\\
&= \sum_{j=0}^{100,000}\binom{100,000}{j}\left(\frac{3}{4}\right)^j\left(\frac{1}{4}\right)^{100,000-j}\\
&\quad -\sum_{j=0}^{100,000}\binom{100,000}{j}\left(\frac{3}{4}\right)^j\left(\frac{1}{4}\right)^{100,000-j}\left(\left(\frac{1}{4}\right)^{10}\right)^j \qquad (7)\\
&= \left(\frac{3}{4}+\frac{1}{4}\right)^{100,000} - \left(\frac{3}{4}\left(1-\left(\frac{1}{4}\right)^{10}\right)+\frac{1}{4}\right)^{100,000}\\
&= 1-\left(1-\frac{3}{4}\left(\frac{1}{4}\right)^{10}\right)^{100,000} \approx 0.069276
\end{aligned}$$

Somit stimmen die Wahrscheinlichkeiten bei (4) - (6) *bis zur siebten Nachkommastelle überein. Für diese Werte könnte man allerdings für keinen der gängigen Fehler 1. Art, also das Zurückweisen der Hypothese, obwohl diese wahr ist, die Hypothese ablehnen. Als Fehler 1. Art wird meist* 0.05 *oder* 0.01 *gewählt.*

Satz 9. *Der Erwartungswert und die Varianz von Y_{max} seien definiert durch:*

$$\mu_{max} \approx \frac{\gamma+\log(n)}{-\log(p)}-\frac{1}{2}, \quad \sigma^2_{max} \approx \frac{\pi^2}{6(\log p)^2}+\frac{1}{12} \qquad (8)$$

Dabei sei $\gamma \approx 0.5772$ die Eulerkonstante.

Beweis. Definiere zunächst eine $Z_{max} \sim Exp(p)$, so dass $Y_{max} = \lfloor Z_{max} \rfloor$ mit $p = \mathrm{e}^{-\lambda}$. Dies gilt, da der ganzzahlige Anteil einer exponentialverteilten Zufallsvariablen geometrisch verteilt ist:

$$\mathbb{P}(\lfloor Z_{max} \rfloor \geq y) = \mathbb{P}(y \geq Z_{max} < y+1) \overset{Exp(\mathrm{e}^{-\lambda})}{=} (1-\mathrm{e}^{-\lambda(y+1)})^n - (1-\mathrm{e}^{-\lambda y})^n \overset{(4)}{=} \mathbb{P}(Y_{max} \geq y)$$

Definiere nun die Zufallsvariable $D = Z_{max} - \lfloor Z_{max} \rfloor$. Da $D \in (0,1)$ und für große n die Wahrscheinlichkeiten aller möglichen Werte annähernd identisch ist, gilt approximativ $D \sim \mathcal{R}(0,1)$. Damit folgt $\mathbb{E}(Z_{max}) \approx \frac{1}{2}$ und $\mathbb{V}(Z_{max}) \approx \frac{1}{12}$.

Weiter ist $\mathbb{E}(Z_{max}) = \frac{1}{\lambda} + \ldots + \frac{1}{\lambda n}$ (wegen der Eigenschaft der Gedächtnislosigkeit der Exponentialverteilung) und auf Grund ihrer Unabhängigkeit somit ebenso $\mathbb{V}(Z_{max}) = \frac{1}{\lambda^2} + \ldots + \frac{1}{(\lambda n)^2}$. Wegen der Eigenschaften der harmonischen Reihen (*) $\sum_{k=1}^{n} \frac{1}{k} \simeq \gamma + \log(n)$ und (**) $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$ folgt

$$\begin{aligned}\mathbb{E}(Y_{max}) &= \mathbb{E}(Z_{max}) - \mathbb{E}(D) = \frac{1}{\lambda} \cdots \frac{1}{\lambda n} - \frac{1}{2} = \frac{1}{\lambda}\left(\sum_{k=1}^{n} \frac{1}{k}\right) - \frac{1}{2} \overset{(*)}{\simeq} \frac{\gamma + \log(n)}{\lambda} - \frac{1}{2} \\ &= \frac{\gamma + \log(n)}{-\log(p)} - \frac{1}{2}.\end{aligned}$$

Und da die Kovarianz $\mathbb{K}(Z_{max}, D) = 0$ für $n \to \infty$

$$\begin{aligned}\mathbb{V}(Y_{max}) &= \mathbb{V}(Z_{max}) + \mathbb{V}(D) = \frac{1}{\lambda^2} + \ldots + \frac{1}{(\lambda n)^2} + \frac{1}{12} = \frac{1}{\lambda^2}\left(\sum_{k=1}^{n} \frac{1}{k^2}\right) + \frac{1}{12} \\ &\overset{(**)}{\simeq} \frac{\gamma + \log(n)}{\lambda} - \frac{1}{2} = \frac{\pi^2}{6} + \frac{1}{12}.\end{aligned}$$

□

Korolar 10. *Seien* $\mu^* = \mu_{max} + \frac{1}{2}$, $(\sigma^*)^2 = \sigma_{max}^2 - \frac{1}{12}$. *Dann lässt sich Gleichung* (4) *durch*

$$\mathbb{P}(Y_{max} \geq y_{max}) \approx 1 - \mathrm{e}^{-\mathrm{e}^{-\left(\pi(y_{max} - \mu^*)/(\sigma^*\sqrt{6}) + \gamma\right)}} \tag{9}$$

sehr gut abschätzen.

Beweis. Parametrisiere p als $p = \mathrm{e}^{-\lambda}$ um und setze $\lambda = \frac{\pi}{\sigma^*\sqrt{6}}$, damit ist

$$\begin{aligned}\mathbb{P}(Y_{max} \geq y_{max}) &= 1 - (1 - p^{y_{max}})^n = 1 - (1 - \mathrm{e}^{\lambda y_{max}})^n \\ &\simeq 1 - (1 - \mathrm{e}^{-n\,\mathrm{e}^{-\lambda y_{max}}}), \text{ wegen } \lim_{n\to\infty}\left(1 + \frac{t}{n}\right)^n = \mathrm{e}^t \text{ für beliebige } t \\ &= 1 - (1 - \mathrm{e}^{-\mathrm{e}^{-\lambda(y_{max} - \log(n)/\lambda}})\end{aligned}$$

$$\begin{aligned}\text{Da } -\lambda\left(y_{max} - \frac{\log(n)}{\lambda}\right) &= -\lambda\left(y_{max} - \frac{\log(n)}{\log(p)}\right) = -\left(\lambda\left(y_{max} - \frac{\log(n)}{\log(p)} - \frac{\gamma}{\lambda}\right) + \gamma\right) \\ &= -\left(\frac{\pi(y_{max} - \mu^*)}{\sigma^*\sqrt{6}} + \gamma\right) \text{ folgt somit die Abschätzung.}\end{aligned}$$

□

Ähnlich kann man auch auf wiederholtes Auftreten beliebiger Nukleotiden testen. Dies wird besonders für Basentripletts bei der Untersuchung auf Krankheiten benutzt.

Korolar 11. *Für den einfachen Fall, dass die Wahrscheinlichkeiten für alle Nukleotiden gleich sind, also $p = \frac{1}{4}$, folgt für den Erwartungswert und die Varianz von Y_{max} damit:*

$$\mu_{max} \approx \frac{\gamma + \log(n)}{\log(4)} + \frac{1}{2}, \quad \sigma^2_{max} \approx \frac{\pi^2}{6(\log 4)^2} + \frac{1}{12} \tag{10}$$

Beweis. Die gleichen Varianzen und nur um 1 verschobenen Erwartungswerte ergeben sich bei Betrachtung des Auftretens eines Nukleotiden an einem beliebigen Ort, definiert als Erfolg. Es sei Y_1 die Länge einer Folge von Erfolgen. Bei Betrachtung beliebiger Wiederholungen liegt ein Erfolg vor, wenn eine Nukleotide mit einer benachbarten übereinstimmt. Die Länge einer solchen Abfolge beläuft sich dann auf $1 + Y_2$, die erste Nukleotide plus die Länge der sich anschließenden Folge. Da für identische Nukleotidenwahrscheinlichkeiten Y_1 und Y_2 die gleiche Verteilung haben, ist $\mathbb{E}(1 + Y_2) = 1 + \mathbb{E}(Y_2) = 1 \cdot \mathbb{E}(Y_1)$ und $\mathbb{V}(1 + Y_2) = \mathbb{V}(Y_1)$. □

4 r-Scans

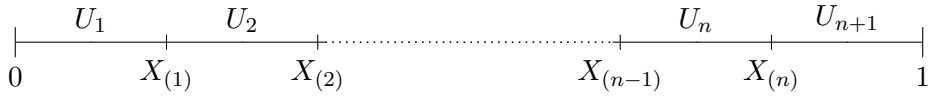
Als nächstes teste man, ob bestimmte genomische Eigenschaften, wie Gene, an unabhängigen und gleichverteilten Orten innerhalb eines Teils des Genoms auftreten. Diese Tests für die räumliche Ungleichmäßigkeit der Verteilung der Eigenschaft werden **r-Scans** genannt. Man nimmt an, dass die Gene relativ zur betrachteten Sequenz so kurz sind, dass ihre Lage als Punkt verstanden werden kann. Im Folgenden werden Tests unter der Hypothese der Unabhängigkeit und Gleichverteilung (hier normiert auf die Gleichverteilung auf dem Intervall $(0, 1)$) gegen die Alternativen

1. des geballten Auftretens der Gene → Teststatistik: U_{max}
2. des Auftretens in regelmäßigen Intervallen → Teststatistik: U_{min}

vorgestellt.

Definition 12. *Seien $X_1, \ldots, X_n$ die Lagen der n Punkte. Man betrachte ihre Ordnungsstatistiken $X_{(1)}, \ldots, X_{(n)}$, die das Intervall $(0, 1)$ in $n + 1$ Teilintervalle der Längen $U_1, \ldots, U_{n+1}$ teilen. Dabei seien*

$$U_1 = X_{(1)}, \quad U_2 = X_{(2)} - X_{(1)}, \ldots, U_{n+1} = 1 - X_{(n)}$$



Lemma 13. *Die gemeinsame Dichtefunktion der Ordnungsstatistik $X_{(i)}$ ist:*

$$f_{X_{(1)},\ldots,X_{(n)}}(x_{(1)}, \ldots, x_{(n)}) = n! \tag{11}$$

Beweis. Da die $X_{(i)}$ i.i.d. verteilt sind, folgt

$$f_{X_{(1)},\ldots,X_{(n)}}(x_{(1)},\ldots,x_{(n)}) = n!\prod_{i=1}^{n} f_{X_{(i)}}(x_{(i)}) = n!\frac{1}{(1-0)^n} = n!$$

□

Um nun die gemeinsame Verteilung der $U_1,\ldots,U_{n+1}$ zu bestimmen, muss die Dichte in (11) auf die Funktionen $U_i = U_i(X_{(1)},\ldots,X_{(n)}), i = 1,\ldots,n$ transformiert werden. Dafür wird folgender Satz benötigt:

Satz 14. *Seien $X_1,\ldots,X_n$ stetige Zufallsvariablen und $V_i = V_i(X_1,\ldots,X_n)$ für $i = 1,\ldots,n$ injektive und differenzierbare Funktionen von $X_1,\ldots,X_n$ mit einer differenzierbaren Inversen. Dann ist die gemeinsame Dichtefunktion der $V_1,\ldots,V_n$ definiert durch*

$$f_{\mathbf{V}_i}(v_1,\ldots,v_n) = f_{\mathbf{X}_{(n)}}(x_{(1)},\ldots,x_{(n)})|J^{-1}|$$

mit der Jakobi-Matrix $J = \begin{pmatrix} \frac{\partial v_1}{\partial x_1} & \frac{\partial v_1}{\partial x_2} & \cdots & \frac{\partial v_1}{\partial x_4} \\ \frac{\partial v_2}{\partial x_1} & \frac{\partial v_2}{\partial x_2} & \cdots & \frac{\partial v_2}{\partial x_4} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial v_4}{\partial x_1} & \frac{\partial v_4}{\partial x_2} & \cdots & \frac{\partial v_4}{\partial x_4} \end{pmatrix}$

Beweis. siehe z.B. „Nichtparametrische statistische Methoden“(Brüning, Trenkle, 1994) □

Korolar 15. *Die gemeinsame Dichte der $\mathbf{U_i}$ ist:*

$$f_{\mathbf{U}_i}(u_1,\ldots,u_n) = n!, \quad u_j > 0, \sum_{j=1}^{n} u_j \leq 1. \tag{12}$$

Dabei sei $\mathbf{U_i} = (U_1,\ldots,U_i), i = 1,\ldots,n+1$, wobei $\mathbf{U_{n+1}}$ wegen $U_{n+1} = 1-(U_1+\cdots+U_n)$ durch $\mathbf{U_i},\ldots,\mathbf{U_n}$ bestimmt wird.

Beweis. Die Jakobi-Matrizen der U_i sind $J = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -1 & 1 \end{pmatrix} \Rightarrow J^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & 1 \end{pmatrix}$, so dass $|J^{-1}| = 1$ ist. Dadurch erhält man die gemeinsame Dichte der $\mathbf{U_i}$:

$$f_{\mathbf{U}_i}(u_1,\ldots,u_n) = f_{\mathbf{X}_{(n)}}(x_{(1)},\ldots,x_{(n)})|J^{-1}| = n!$$

□

4.1 Geballtes Auftreten

Man teste nun zunächst die Hypothese gegen die Alternative, dass die Punkte, an denen die Genomeigenschaft auftritt, geballt auftreten, also Klumpen bilden. Dafür betrachte man als die Teststatistik das Maximum U_{max} der $U_1,\ldots,U_{n+1}$ und bestimme ihre Verteilung unter der Hypothese. Wenn zufällig g der Teilintervalle $U_1,\ldots,U_{n+1}$ ausgewählt werden, suche zunächst

die Wahrscheinlichkeit, dass alle der g gewählten U_i länger als ein beliebiges $u \in (0,1)$ sind. Dabei sei ohne Einschränkung $ug < 1$, da für $ug > 1$ die Wahrscheinlichkeit 0 ist. Die gemeinsame Dichtefunktion $f_{\mathbf{U_g}}(u_1, \ldots, u_g)$ einer g-elementigen Teilmenge der $U_1, \ldots, U_{n+1}$ ist unabhängig von der Wahl der Teilmenge, da die Dichte (12) symmetrisch ist.

Lemma 16. *Die gemeinsame Dichte der $U_1, \ldots, U_g$ ist:*

$$f_{\mathbf{U_g}}(u_1, \ldots, u_g) = \frac{n!}{(n-g)!}(1 - u_1 - \cdots - u_g)^{n-g} \tag{13}$$

Beweis. Die gemeinsame Dichte berechnet sich dann als Randdichte von $U_1, \ldots, U_g$, wobei die oberen Intervallgrenzen wegen $\sum_{j=1}^{n} u_j \leq 1$ durch $w_j = 1 - u_1 - \cdots - u_{j-i}$ definiert sind. Dies ergibt:

$$\begin{aligned}
f_{\mathbf{U_g}}(u_1, \ldots, u_g) &= \int_0^{w_{g+1}} \cdots \int_0^{w_{n-1}} n!(1 - u_1 - \cdots - u_{n-1})\, \mathrm{d}u_{n-1} \ldots\, \mathrm{d}u_{g+1} \\
&= \int_0^{w_{g+1}} \cdots \int_0^{w_{n-2}} n! \left[(1 - u_1 - \cdots - u_{n-2})u_{n-1} - \frac{u_{n-1}^2}{2}\right]_0^{w_{n-1}} \mathrm{d}u_{n-2} \ldots\, \mathrm{d}u_{g+1} \\
&= \int_0^{w_{g+1}} \cdots \int_0^{w_{n-2}} \frac{n!}{2}(1 - u_1 - \cdots - u_{n-2})^2\, \mathrm{d}u_{n-2} \ldots\, \mathrm{d}u_{g+1} \\
&= \int_0^{w_{g+1}} \cdots \int_0^{w_{n-3}} \frac{n!}{2} \left[-\frac{(1 - u_1 - \cdots - u_{n-2})^3}{3}\right]_0^{w_{n-2}} \mathrm{d}u_{n-3} \ldots\, \mathrm{d}u_{g+1} \\
&= \int_0^{w_{g+1}} \cdots \int_0^{w_{n-3}} \frac{n!}{6} \left(0 - \left(-(1 - u_1 - \cdots - u_{n-2})^3\right)\right) \mathrm{d}u_{n-3} \ldots\, \mathrm{d}u_{g+1} \\
&= \quad \cdots \\
&= \frac{n!}{(n-g)!}(1 - u_1 - \cdots - u_g)^{n-g}
\end{aligned}$$

Beachte, dass dies auch als $f_{\mathbf{U_g}}(u_1, \ldots, u_g) = \binom{n}{g}(1 - \sum_{j=1}^{n} u_j)^{n-g} g!$ geschrieben werden kann. □

Korolar 17. *Die Wahrscheinlichkeit, dass alle Abstände $U_1, \ldots, U_g$ größer als u sind, ist damit:*

$$\mathbb{P}(U_1 > u, \ldots, U_g > u) = (1 - gu)^n \tag{14}$$

Beweis. Die gesuchte Wahrscheinlichkeit ergibt sich durch Integration über die verbleibenden $U_1, \ldots, U_g$ wobei wie oben die oberen Intervallgrenzen 0 ergeben.

$$\begin{aligned}
\mathbb{P}(U_1 > u, \ldots, U_g > u) &= \int_u^1 \int_u^{w_1} \cdots \int_u^{w_g} \frac{n!}{(n-g)!}(1 - u_1 - \cdots - u_g)^{n-g}\, \mathrm{d}u_g \cdots\, \mathrm{d}u_1 \\
&= \int_u^1 \cdots \int_u^{w_{g-1}} \frac{n!}{(n-g+1)!}(1 - u_1 - \cdots - u_{g-1} - u)^{n-g+1}\, \mathrm{d}u_{g-1} \cdots\, \mathrm{d}u_1 \\
&= \quad \cdots \\
&= (1 - gu)^n
\end{aligned}$$

□

Satz 18. *Sei h_u die größte natürliche Zahl so dass $h_u u < 1$. Die Dichte der Ordnungsstatistik U_{max} ist definiert durch*

$$f_{U_{max}}(u) = n \sum_{g=1}^{h_u} (-1)^{g+1} \binom{n+1}{g} g(1-gu)^{n-1} \tag{15}$$

Beweis. Definiere die Ereignisse $A_i = \{U_i \geq u\}, i = 1, \ldots, n+1$. Seien weiter $i_1, \ldots, i_g$ paarweise verschiedene Indizes, dann gilt, damit die Bedingung $ug < 1$ nicht verletzt wird:

$$\mathbb{P}(A_{i_1} \cdots A_{i_g}) = \begin{cases} (1-gu)^n, & g \leq h_u \\ 0, & g > h_u. \end{cases} \tag{16}$$

Mit Hilfe des Einschluss-Ausschluss-Prinzips folgt dann für die Wahrscheinlichkeit, dass mindestens eines der Ereignisse eintritt, also dass das Maximum der U_i länger als u ist:

$$\mathbb{P}(A_{i_1} \cup \cdots \cup A_{i_g}) = \sum_{j=1} \mathbb{P}(A_{i_j}) + \cdots + (-1)^{g-1} \mathbb{P}(A_{i_1} \cdots A_{i_g}) = \sum_{g=1}^{h_u} (-1)^{g+1} \binom{n+1}{g} (1-gu)^n. \tag{17}$$

Hieraus lässt sich über die Verteilungsfunktion $F_{U_{max}}(u)$ von U_{max} die Dichte $f_{U_{max}}(u)$ berechnen, weil (17) $= 1 - F_{U_{max}}(u)$:

$$\begin{aligned} f_{U_{max}}(u) &= \frac{\partial}{\partial u} \left(F_{U_{max}}(u) \right) = \frac{\partial}{\partial u} \left(1 - \sum_{g=1}^{h_u} (-1)^{g+1} \binom{n+1}{g} (1-gu)^n \right) \\ &= n \sum_{g=1}^{h_u} (-1)^{g+1} \binom{n+1}{g} g(1-gu)^{n-1} \end{aligned} \tag{18}$$

□

Korolar 19. *Die Wahrscheinlichkeit in* (17) *lässt sich durch folgende Abschätzungen vereinfachen:*

$$\mathbb{P}(U_{max} \geq u) \simeq 1 - \mathrm{e}^{-(n+1)\,\mathrm{e}^{-(n+1)u}} \tag{19}$$

Wenn $(n+1)\,\mathrm{e}^{-(n+1)u}$ klein genug ist, gilt sogar:

$$\mathbb{P}(U_{max} \geq u) \simeq (n+1)\,\mathrm{e}^{-(n+1)u} \tag{20}$$

Beweis. Benutze den Erwartungswert $\mu_{max} = \frac{1}{n+1} \left(\frac{1}{n+1} + \frac{1}{n} + \cdots + 1 \right)$ und die Varianz $\sigma^2_{max} = \frac{\pi^2}{6(n+1)^2}$ (ohne Beweis, siehe Karlin and Macken (1991a,b)) von U_{max}, dann ergibt sich

Abschätzung (19):

$$\begin{aligned}
\mathbb{P}(U_{max} \geq u) &\simeq 1 - \mathrm{e}^{-\mathrm{e}^{-(\pi(u-\mu_{max})/(\sigma_{max}\sqrt{6})+\gamma)}} \\
&\simeq 1 - \mathrm{e}^{-\mathrm{e}^{-\left(\pi\left(u-\frac{1}{n+1}\sum_{k=1}^{n+1}\frac{1}{k}\right)/\left(\frac{\pi}{(n+1)}\right)+\gamma\right)}} \\
&\simeq 1 - \mathrm{e}^{-\mathrm{e}^{-\left((n+1)u-\sum_{k=1}^{n+1}\frac{1}{k}+\gamma\right)}} \\
&\simeq 1 - \mathrm{e}^{-\mathrm{e}^{-((n+1)u-\ln(n+1))}} \\
&\simeq 1 - \mathrm{e}^{-(n+1)\,\mathrm{e}^{-(n+1)u}}
\end{aligned}$$

Dies gilt wieder, da die asymptotische Entwicklung der harmonischen Reihe $\sum_{k=1}^{n} \frac{1}{k} = \gamma + \ln n + \mathcal{O}(\frac{1}{n})$ ist. Ist außerdem $(n+1)\,\mathrm{e}^{-(n+1)u}$ klein genug, benutze weiter die Approximation: $\mathrm{e}^x \simeq x + 1$ für $|x|$ klein genug, um (20) zu erhalten. □

Das folgende Beispiel zeigt, dass die obigen Approximationen immerhin für große n gute Ergebnisse liefern.

Beispiel 20. *Wähle $u = 0.01$, $n + 1 = 1000$, dann liefert die exakte Gleichung folgendes Ergebnis:*
(17)*:* $\sum_{g=1}^{99} (-1)^{g+1} \binom{1000}{g} (1 - 0.01g)^{999} \approx 0.0428$ *verglichen mit*
(19)*:* $1 - \mathrm{e}^{-1000\,\mathrm{e}^{-10}} \approx 0.0444$ *und*
(20)*:* $1000\,\mathrm{e}^{-10} \approx 0.0454$

4.2 Auftreten in regelmäßigen Intervallen

Teste nun mit der Teststatistik U_{min} die Hypothese gegen die Alternative, dass die Punkte regelmäßige Abstände haben, und lehne somit die Gleichverteilungs– und Unabhängigkeits–Hypothese ab, wenn die Realisation u von U_{min} zu groß ist. Mit der Wahrscheinlichkeit aus (16) folgt für $g = n + 1$, dass für alle $u \in (0, \frac{1}{n+1})$ (damit wieder $ug < 1$ bleibt) für die Wahrscheinlichkeit, dass alle $n + 1$ Intervalllängen länger als u sind:

$$\mathbb{P}(U_{min} \geq u) = (1 - (n+1)u)^n\,. \tag{21}$$

Satz 21. *Die Dichte von U_{min} ist definiert durch:*

$$f_{U_{min}}(u) = n\,(n+1)\,(1-(n+1)u)^{n-1}\,, \quad 0 < u < \frac{1}{n+1} \tag{22}$$

Beweis. Da Gleichung (21) wieder $1 - F_{U_{min}}(u)$ entspricht, folgt analog zu (18):

$$f_{U_{min}}(u) = \frac{\partial}{\partial u} F_{U_{min}}(u) = \frac{\partial}{\partial u}\left(1 - (1-(n+1)u)^n\right) = n\,(n+1)\,(1-(n+1)u)^{n-1}$$

□

5 Fazit

Dieser Vortrag hat nur einen kleinen Einblick in die Analyse von DNA-Sequenzen geben können. Es wurden zum einen die grundsätzlichen Annahmen - Unabhängigkeit gegen Markov-abhängigkeiten - eingeführt. Zum anderen wurden die Tests, ob eine Tendenz zur Wiederholung von Nukleotiden vorliegt und welche Art der räumlichen Verteilung vorliegt, vorgestellt. Ein weiterer Analyseansatz, der hier nicht mehr vorgestellt werden kann, ist die Musteranalyse, die letztendlich auf einer Poisson-Verteilung basiert. Hierbei testet man zum einen, wie häufig bestimmte *Wörter* bzw. ganze Gruppen von Nukleotiden-Wörtern, sogenannte *Motive*, in einer DNA-Sequenz auftreten und zum anderen, welcher Abstand zwischen zwischen einzelnen Auftritten besteht.