Local Alignment of Random DNA-Sequences

Nora Miriam Steinle

17. April 2012

Matrikelnummer: 349620
Veranstaltung: Seminar zur mathematischen Biologie im SS 2012
Betreuer: Prof. Löwe

Einleitung

Beim Vergleich zweier DNA-Sequenzen miteinander ist es von Interesse die best zusammenpassenden Teilsequenzen herauszufinden. Wir haben uns bereits damit beschäftigt, welche Ansätze bei nicht-zufälligen Sequenzen angewendet werden. Für die Bewertung der Ergebnisse ist es hilfreich zu wissen, wie sich beispielsweise die Länge der längsten Übereinstimmungsfolge verhält, wenn die Sequenzen zufällig sind. Das ist Thema dieser Ausarbeitung.
In einem ersten Teil betrachten wir zwei Sequenzen

$$\begin{aligned} A &: A_1 A_2 \dots A_n \\ B &: B_1 B_2 \dots B_n \end{aligned}$$

mit unabhängig identisch verteilten Buchstaben $A_1, A_2, \dots, B_1, B_2, \dots$ und vergleichen die Länge der längsten Übereinstimmungsfolge, wenn

- Fall 1: keine Verschiebung der Sequenzen
- Fall 2: Verschiebung der Sequenzen

erlaubt ist.
In einem zweiten Teil betrachten wir die Länge der längsten Übereinstimmungsfolge, wenn $A_1, A_2, \dots$ nach ξ und $B_1, B_2, \dots$ nach ν verteilt sind und alle Buchstaben unabhängig sind.

1 Part I

Als Erstes betrachten wir die Länge R_n der längsten Übereinstimmungsfolge zweier Sequenzen der Länge n von iid Buchstaben. Die Buchstaben seien aus einem endlichen Alphabet $\mathcal{A}$. Mit A_i bzw. B_i bezeichne die Zufallsvariable, die die Ziehung des i-ten Buchstaben der Sequenz A bzw. B beschreibe. Die Wahrscheinlichkeit für eine Übereinstimmung beträgt dann

$$p = \sum_{i \in \mathcal{A}} \xi_i^2,$$

wobei ξ_i die Wahrscheinlichkeit für den Buchstaben i sei. Daher lässt sich unseres Modell auf jenes eines n-fachen Münzwurf mit Erfolgswahrscheinlichkeit p reduzieren.
Der erste Satz, den wir beweisen werden, besagt, dass sich R_n asymptotisch wie $\log_{\frac{1}{p}}(n)$ verhält. Die zugrund liegende Idee ist die folgende:
Eine Übereinstimmungsfolge der Länge m hat die Wahrscheinlichkeit p^m. Für große n und verhältnismäßig kleine m gibt es ungefähr n verschiedene Startpositionen einer Übereinstimmungsfolge der Länge m. Daher ergibt sich hieraus

$$\mathbb{E}\left(\text{Anzahl der Übereinstimmungsfolgen der Länge } m\right) \cong np^m.$$

Gehen wir nun davon aus, dass die längste Übereinstimmungsfolge eindeutig ist, so gilt $1 = np^{R_n}$, was äquivalent zu $R_n = \log_{\frac{1}{p}}(n)$ ist.

Wir werden zeigen, dass für großes n mit Wahrscheinlichkeit 1 tatsächlich $R_n = \log_{\frac{1}{p}}(n)$ gilt. Hierfür erinnern wir an:

Bemerkung 1.1 *(Borel-Cantelli) Gilt* $\sum_{n=1}^{\infty} \mathbb{P}(C_n) < \infty$ *für eine Folge von Ereignissen* $(C_n)_n$, *so folgt* $\mathbb{P}(C_n \text{ für unendlich viele } n) = 0$.

Satz 1.2 *Seien* $A_1, A_2, \ldots, B_1, B_2, \ldots$ *unabhängig und identisch verteilt und sei* $0 < p = \mathbb{P}(A_1 = B_1) < 1$. *Definiere*

$$R_n = \max\{m : A_{i+k} = B_{i+k} \text{ für } k = 1, \ldots, m, 0 \leq i \leq n - m\}.$$

Dann gilt

$$\mathbb{P}\left(\lim_n \frac{R_n}{\log_{\frac{1}{p}}(n)} = 1\right) = 1.$$

Beweis: Im ersten Teil des Beweises ist das Ziel

$$\mathbb{P}\left(\limsup_n \frac{R_n}{\log_{\frac{1}{p}}(n)} \leq 1\right) = 1$$

zu zeigen. Dies wäre erreicht, wenn wir zeigen könnten, dass mit Wahrscheinlichkeit 1 für alle bis auf endlich viele $n \in \mathbb{N}$

$$R_n \leq \log_{\frac{1}{p}}(n)$$

gilt. Um die Rechnungen einfacher zu machen, beweisen wir Obiges zunächst für eine Teilfolge $(n_k)_k$ von $(n)_n$. Sei dazu $D_i(n) := \{A_{i+k} = B_{i+k} \text{ für } k = 1, \ldots, m\}$, wobei hier $0 \leq i \leq n - m$. Sei $\epsilon > 0$ und $m := \lfloor (1+\epsilon) \log_{\frac{1}{p}}(n) \rfloor$. Dann gilt

$$(1+\epsilon)\log_{\frac{1}{p}}(n) \leq m+1 \quad \Leftrightarrow \quad -(m+1) \leq \log_{\frac{1}{p}}\left(n^{-(1+\epsilon)}\right) \Leftrightarrow -m \leq \log_{\frac{1}{p}}\left(\frac{1}{p}n^{-(1+\epsilon)}\right)$$
$$\Leftrightarrow \quad p^m \leq \frac{1}{p}n^{-(1+\epsilon)},$$

also

$$p^m = \mathbb{P}(D_i(n)) \leq \frac{1}{p}n^{-(1+\epsilon)}.$$

Für $n_k := \lfloor \left(\frac{1}{p}\right)^k \rfloor$ und $m_k := \lfloor (1+\epsilon)\log_{\frac{1}{p}}(n_k) \rfloor$ definiere

$$E(n_k) := \bigcup_{i=0}^{n_k - m_k} \{A_{i+k} = B_{i+k} \text{ für } k = 1, \ldots, m_k\} = \bigcup_{i=0}^{n_k - m_k} D_i(n_k)$$

als das Ereignis, dass $R_{n_k} \geq m_k$ gilt. Mit obiger Abschätzung sowie $n_k \geq 1, m_k \geq 0$ ergibt dies

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbb{P}(E(n_k)) &\leq \sum_{k=1}^{\infty} \sum_{i=0}^{n_k - m_k} \mathbb{P}(D_i(n_k)) = \sum_{k=1}^{\infty} (n_k - m_k + 1) p^{m_k} \\
&\leq \sum_{k=1}^{\infty} 2 n_k p^{m_k} \leq \sum_{k=1}^{\infty} \frac{2}{p} n_k n_k^{-(1+\epsilon)} \leq \sum_{k=1}^{\infty} \frac{2}{p} \left(\frac{1}{p}\right)^{-\epsilon k} \\
&= \frac{2}{p} \sum_{k=1}^{\infty} p^{\epsilon k} < \infty,
\end{aligned}$$

da $0 < p^\epsilon < 1$. Mit Borel-Cantelli folgt $\mathbb{P}(E(n_k) \text{ für unendlich viele } k) = 0$, d.h., mit Wahrscheinlichkeit 1 existiert ein $K \in \mathbb{N}$, sodass für alle $k \geq K$ gilt $R_{n_k} < \lfloor (1+\epsilon)\log_{\frac{1}{p}}(n_k) \rfloor$. Da mit $R_{n_k} < \lfloor (1+\epsilon)\log_{\frac{1}{p}}(n) \rfloor$ auch $R_{n_k} < (1+\epsilon)\log_{\frac{1}{p}}(n_k)$ gilt und $\epsilon > 0$ beliebig war, folgt $R_{n_k} \leq \log_{\frac{1}{p}}(n_k)$. Insgesamt erhalten wir also, dass mit Wahrscheinlichkeit 1 ein $K \in \mathbb{N}$ existiert, sodass

$$R_{n_k} \leq \log_{\frac{1}{p}}(n_k) \text{ für alle } k \geq K.$$

Mit Wahrscheinlichkeit 1 ist daher auch

$$\limsup_k \frac{R_{n_k}}{\log_{\frac{1}{p}}(n_k)} \leq 1,$$

d.h.

$$\mathbb{P}\left(\limsup_k \frac{R_{n_k}}{\log_{\frac{1}{p}}(n_k)} \leq 1\right) = 1.$$

Um die Aussage auf die Folge $(n)_{n\in\mathbb{N}}$ zu übertragen zeigen wir

- $\lim\limits_k \frac{\log_{\frac{1}{p}}(n_{k+1})}{\log_{\frac{1}{p}}(n_k)} = 1$
  um
- $\limsup\limits_n \frac{R_n}{\log_{\frac{1}{p}}(n)} \leq \limsup\limits_k \frac{R_{n_k}}{\log_{\frac{1}{p}}(n_k)}$

und damit die Gleichheit der Limiten zu erhalten. Der Schlüssel für den ersten Punkt liegt in den Abschätzungen

$$\begin{aligned}\log_{\frac{1}{p}}(n_k) &= \log_{\frac{1}{p}}\left(\lfloor\left(\frac{1}{p}\right)^k\rfloor\right) \leq \log_{\frac{1}{p}}\left(\left(\frac{1}{p}\right)^k\right) = k\\ \text{sowie}\\ \log_{\frac{1}{p}}(n_k) &= \log_{\frac{1}{p}}\left(\lfloor\left(\frac{1}{p}\right)^k\rfloor\right) \geq \log_{\frac{1}{p}}\left(\left(\frac{1}{p}\right)^k - 1\right) = \log_{\frac{1}{p}}\left(\left(\frac{1}{p}\right)^k\right) + \log_{\frac{1}{p}}\left(1-p^k\right)\\ &= k + \log_{\frac{1}{p}}\left(1-p^k\right).\end{aligned}$$

Analog gilt $\log_{\frac{1}{p}}(n_{k+1}) \geq (k+1) + \log_{\frac{1}{p}}\left(1-p^{k+1}\right)$. Definieren wir $\epsilon_k := \log_{\frac{1}{p}}\left(1-p^k\right) < 0$ sowie $\delta_k := \log_{\frac{1}{p}}\left(1-p^{k+1}\right) < 0$, so gilt $\epsilon_k \to 0$ und $\delta_k \to 0$. Nun liefern die obigen Überlegungen

$$\frac{k+1+\delta_k}{k} \leq \frac{\log_{\frac{1}{p}}(n_{k+1})}{\log_{\frac{1}{p}}(n_k)} \leq \frac{k+1}{k+\epsilon_k},$$

was $\lim\limits_k \frac{\log_{\frac{1}{p}}(n_{k+1})}{\log_{\frac{1}{p}}(n_k)} = 1$ ergibt. Für den zweiten Punkt beachte, dass zu $m \in \mathbb{N}$ beliebig ein $k_m \in \mathbb{N}$ existiert mit $n_{k_m} \leq m \leq n_{k_m+1}$ und dass mit der Monotonie von R_n die Ungleichung

$$\begin{aligned}\frac{R_m}{\log_{\frac{1}{p}}(m)} &\leq \frac{R_{n_{k_m+1}}}{\log_{\frac{1}{p}}(n_{k_m+1})}\frac{\log_{\frac{1}{p}}(n_{k_m+1})}{\log_{\frac{1}{p}}(m)}\\ &\leq \frac{R_{n_{k_m+1}}}{\log_{\frac{1}{p}}(n_{k_m+1})}\frac{\log_{\frac{1}{p}}(n_{k_m+1})}{\log_{\frac{1}{p}}(n_{k_m})}\end{aligned}$$

folgt. Gehen wir auf den Limes superior über, so ergibt sich

$$\begin{aligned}
\limsup_m \frac{R_m}{\log_{\frac{1}{p}}(m)} &\leq \limsup_m \frac{R_{n_{k_m}+1}}{\log_{\frac{1}{p}}(n_{k_m+1})} \frac{\log_{\frac{1}{p}}(n_{k_m+1})}{\log_{\frac{1}{p}}(n_{k_m})} \\
&\leq \limsup_k \frac{R_{n_{k+1}}}{\log_{\frac{1}{p}}(n_{k+1})} \frac{\log_{\frac{1}{p}}(n_{k+1})}{\log_{\frac{1}{p}}(n_k)} \\
&\leq \limsup_k \frac{R_{n_k+1}}{\log_{\frac{1}{p}}(n_{k+1})} \limsup_k \frac{\log_{\frac{1}{p}}(n_{k+1})}{\log_{\frac{1}{p}}(n_k)} \\
&= \limsup_k \frac{R_{n_{k+1}}}{\log_{\frac{1}{p}}(n_{k+1})},
\end{aligned}$$

wobei wir beim zweiten Ungleichheitszeichen Teilfolgeneigenschaften ausgenutzt haben. Damit ist das Ziel, $\mathbb{P}\left(\limsup_n \frac{R_n}{\log_{\frac{1}{p}}(n)} \leq 1\right) = 1$, erreicht.
Zeige nun $\mathbb{P}\left(\liminf_n \frac{R_n}{\log_{\frac{1}{p}}(n)} \geq 1\right) = 1$. Dazu sei $1 > \epsilon > 0$ beliebig und $C_n := \left\{R_n \leq (1-\epsilon)\log_{\frac{1}{p}}(n)\right\}$. Wir zeigen mit Borel-Cantelli, dass

$$\mathbb{P}\left(C_n \text{ für unendlich viele n}\right) = 0.$$

Das bedeutet wieder, dass mit Wahrscheinlichkeit 1 ein $N \in \mathbb{N}$ existiert mit "'C_n tritt nicht auf"' für alle $n \geq N$. Mit anderen Worten existiert dann mit Wahrscheinlichkeit 1 ein $N \in \mathbb{N}$, sodass $R_n \geq \log_{\frac{1}{p}}(n)$ für alle $n \geq N$, d.h.

$$\mathbb{P}\left(\liminf_n \frac{R_n}{\log_{\frac{1}{p}}(n)} \geq 1\right) = 1.$$

Definiere hierzu $m := \lfloor (1-\epsilon)\log_{\frac{1}{p}}(n) \rfloor$ und

$$E_i(n) := \left\{A_{(m+1)i+k} = B_{(m+1)i+k} \text{ für } k = 1, ..., m+1\right\}$$

für $0 \leq i \leq \lfloor \frac{n}{m+1} \rfloor - 1$. Wir vergleichen die beiden Sequenzen also blockweise, wobei jeder Block die Länge $m+1$ hat. Die $E_i(n)$ sind für festes n unabhängig und jedes $E_i(n)$ hat Wahrscheinlichkeit p^{m+1}. Nun ist

$$p^{m+1} \geq pn^{-(1-\epsilon)},$$

da

$$\begin{aligned} m = \lfloor (1-\epsilon)\log_{\frac{1}{p}}(n) \rfloor &\leq (1-\epsilon)\log_{\frac{1}{p}}(n) \\ \Leftrightarrow -m-1 &\geq -1-(1-\epsilon)\log_{\frac{1}{p}}(n) \\ \Leftrightarrow p^{m+1} &\geq pn^{-(1-\epsilon)}. \end{aligned}$$

Wir wollen als nächstes sehen, dass

$$\lfloor \frac{n}{m+1} \rfloor \geq \frac{n}{\log_{\frac{1}{p}}(n)} \tag{1}$$

für n groß genug. Für $n \geq N$ und $N \in \mathbb{N}$ geeignet gilt $(1-\epsilon)\log_{\frac{1}{p}}(n) > 1$ sowie

$$\lfloor \frac{n}{m+1} \rfloor \geq \frac{n}{m+1} - 1 \geq \frac{n}{(1-\epsilon)\log_{\frac{1}{p}}(n)+1} - 1 \geq \frac{n}{2(1-\epsilon)\log_{\frac{1}{p}}(n)} - 1$$

Gilt also

$$\left(\frac{1}{2-2\epsilon}\right)\frac{n}{\log_{\frac{1}{p}}(n)} - 1 \geq \frac{n}{\log_{\frac{1}{p}}(n)}$$

für n groß genug, so auch Ungleichung **??**. Die Ungleichung ist äquivalent zu

$$\left(\frac{1}{2-2\epsilon} - 1\right)\frac{n}{\log_{\frac{1}{p}}(n)} = \left(\frac{2\epsilon-1}{2-2\epsilon}\right)\frac{n}{\log_{\frac{1}{p}}(n)} \geq 1.$$

Letzteres ist erfüllt für $n \geq N \in \mathbb{N}$, da $\lim_n \frac{n}{\log_{\frac{1}{p}}(n)} = \infty$, d.h., es gilt Ungleichung **??** für $n \geq N$. Mit der Unabhängigkeit der $(E_i(n))_i$ sind auch deren Komplemente unabhängig und es folgt für große n wegen $1 + x \leq e^x$

$$\begin{aligned} \mathbb{P}(C_n) &= \mathbb{P}\left(R_n \leq (1-\epsilon)\log_{\frac{1}{p}}(n)\right) \leq \mathbb{P}\left(\bigcap_{i=0}^{\lfloor \frac{n}{m+1} \rfloor - 1} E_i(n)^C\right) \\ &= \left(1-p^{m+1}\right)^{\lfloor \frac{n}{m+1} \rfloor} \leq \left(1-pn^{-(1-\epsilon)}\right)^{\lfloor \frac{n}{m+1} \rfloor} \leq \left(1-pn^{-(1-\epsilon)}\right)^{\frac{n}{\log_{\frac{1}{p}}(n)}} \\ &\leq exp\left(-\frac{pn^{\epsilon}}{\log_{\frac{1}{p}}(n)}\right). \end{aligned}$$

Wir wollen sehen, dass $\sum_{n\geq 2} exp\left(-\frac{pn^{\epsilon}}{\log_{\frac{1}{p}}(n)}\right) < \infty$ und damit auch

$\sum_{n\geq 2} \mathbb{P}(C_n) < \infty$. Sei dazu $\lambda \in (0,1)$. Nun ist

$$\begin{aligned} exp\left(\frac{-pn^{\epsilon}}{\log_{\frac{1}{p}}(n)}\right) &\leq \lambda^n \\ \Leftrightarrow \frac{-pn^{\epsilon}}{\log_{\frac{1}{p}}(n)} &\leq n\ln(\lambda) \\ \Leftrightarrow \frac{1}{n^{1-\epsilon}\log_{\frac{1}{p}}(n)} &\leq -\frac{\ln(\lambda)}{p} = c. \end{aligned}$$

Da die linke Seite gegen Null konvergiert und c eine positive Konstante ist, gilt $exp\left(\frac{-pn^{\epsilon}}{\log_{\frac{1}{p}}(n)}\right) \leq \lambda^n$ für alle n groß genug. Also konvergieren alle betrachteten Reihen und es folgt

$$\mathbb{P}\left(R_n \geq (1-\epsilon)\log_{\frac{1}{p}}(n) \text{ für unendlich viele n}\right) = 0.$$

Dies impliziert

$$\mathbb{P}\left(\liminf_n \frac{R_n}{\log_{\frac{1}{p}}(n)} \geq 1\right) = 1$$

und der Satz ist bewiesen. □

Bemerkung 1.3 *Der Satz benutzt an keiner Stelle, dass $A_1, A_2, \ldots, B_1, B_2, \ldots$ identisch verteilt sind. Daher kann er wie folgt verallgemeinert werden: Seien $A_1, A_2, \ldots$ nach ξ und $B_1, B_2, \ldots$ nach ν verteilt. $A_1, A_2, \ldots, B_1, B_2, \ldots$ seien unabhängig und $p = \mathbb{P}(A_i = B_i) \in (0,1)$. Ist*

$$R_n = \max\{m : A_{i+k} = B_{i+k} \text{ für } k = 1, \ldots, m, 0 \leq i \leq n-m\}$$

wie oben, so gilt

$$\mathbb{P}\left(\lim_n \frac{R_n}{\log_{\frac{1}{p}}(n)} = 1\right) = 1.$$

Als nächstes untersuchen wir wie sich die Länge der längsten Übereinstimmungsfolge H_n verhält, wenn die zu vergleichenden Sequenzen gegeneinander verschoben werden dürfen. Dieser Fall ist von größerem biologischem Interesse. Für eine Übereinstimmungsfolge der Länge m gibt es hier n^2 verschiedene

Wahlen der Startposition (i,j). Mit derselben Heuristik wie oben würden wir also nun erwarten, dass sich H_n asymptotisch wie $\log_{\frac{1}{p}}(n^2)$ verhält. Dies ist tatsächlich der Fall:

Satz 1.4 *Seien $A_1, A_2, \ldots, B_1, B_2, \ldots$ unabhängig und identisch verteilt mit $0 < p = \mathbb{P}(A_1 = B_1) < 1$. Definiere*

$$H_n = max\left\{m : \ A_{i+k} = B_{j+k} \text{ für } k = 1, \ldots, m \wedge 0 \leq i, j \leq n - m\right\}.$$

Dann gilt $\mathbb{P}\left(\lim_n \frac{H_n}{\log_{\frac{1}{p}}(n)} = 2\right) = 1.$

Beweis: Die obere Schranke für den Limes superior erhält man wie im vorhergehenden Beweis. Für die untere Schranke braucht man noch andere Hilfsmittel. Der Beweis ist zu lang, um ihn hier ordentlich aufzuführen.

Lassen wir also Verschiebung der Sequenzen zu, so verdoppelt sich die Länge der längsten Übereinstimmungsfolge.

2 Part II

Satz 2.1 *Seien $A_1, A_2, \ldots$ nach ξ verteilt, $B_1, B_2, \ldots$ nach ν verteilt, wobei alle Buchstaben unabhängig seien und $p = \mathbb{P}(A_1 = B_1) \in (0,1)$. Dann gibt es eine Konstante $\mathcal{C}(\xi, \nu) \in [1,2]$, sodass*

$$\mathbb{P}\left(\lim_n \frac{H_n}{\log_{\frac{1}{p}}(n)} = \mathcal{C}(\xi, \nu)\right) = 1.$$

Außerdem gilt

$$\mathcal{C}(\xi, \nu) = \sup_{\gamma \in Pr(\mathcal{A})} min\left\{\frac{\log\left(\frac{1}{p}\right)}{\mathcal{H}(\gamma, \xi)}, \frac{\log\left(\frac{1}{p}\right)}{\mathcal{H}(\gamma, \nu)}, \frac{2\log\left(\frac{1}{p}\right)}{\log\left(\frac{1}{p}\right) + \mathcal{H}(\gamma, \beta)}\right\},$$

wobei $\beta_a := \frac{\xi_a \nu_a}{p}, \mathcal{H}(\eta, \psi) := \sum_{a \in \mathcal{A}} \eta_a \log\left(\frac{\eta_a}{\psi_a}\right)$ und γ alle Wahrscheinlichkeitsverteilungen auf dem Alphabet $\mathcal{A}$ durchläuft. Die Basis des Logarithmus ist

frei wählbar.
Desweiteren ist $\mathcal{C}(\xi,\nu)=2$ *genau dann, wenn*

$$max\left\{\mathcal{H}(\beta,\nu),\mathcal{H}(\beta,\xi)\right\}\leq\frac{1}{2}\log\left(\frac{1}{p}\right).$$

Bemerkung 2.2 *Das oben definierte* β *mit* $\beta(a)=\beta_a$ *für* $a\in\mathcal{A}$ *ist ein Wahrscheinlichkeitsmaß auf dem Alphabet* $\mathcal{A}$*, denn mit der Unabhängigkeit der* A_i *und* B_i *folgt*

$$\begin{aligned}\sum_{a\in\mathcal{A}}\beta_a &= \frac{1}{p}\sum_{a\in\mathcal{A}}\xi_a\nu_a=\frac{1}{p}\sum_{a\in\mathcal{A}}\mathbb{P}(A_1=a)\,\mathbb{P}(B_1=a)=\frac{1}{p}\sum_{a\in\mathcal{A}}\mathbb{P}(A_1=a=B_1)\\ &= \frac{1}{p}\mathbb{P}(A_1=B_1)=1.\end{aligned}$$

Bekanntlich gilt für $\xi=\nu$, dass $\mathcal{C}(\xi,\nu)=2$ ist. Der Satz zeigt, dass es viele Verteilungen ξ auf $\mathcal{A}$ gibt, sodass $\mathcal{C}(\xi,\nu)=2$ ist, was die Stärke des $2\log_{\frac{1}{p}}(n)$-Gesetz belegt. Es folgt ein Beispiel, bei dem wir explizit bestimmen können, wann Verschiebung die Länge der längsten Übereinstimmungsfolge verdoppelt.

Beispiel: Sei $\mathcal{A}=\{H,T\}$ und seien $A_1,A_2,\ldots$ mit $\mathbb{P}(A_i=H)=\mathbb{P}(A_i=T)=\frac{1}{2}$ sowie $B_1,B_2,\ldots$ mit $\mathbb{P}(B_i=H)=\theta=1-\mathbb{P}(B_i=T)$ und $\theta\in[0,1]$. Dann ist mit der Unabhängigkeit der A_i und B_i

$$\begin{aligned}p &= \mathbb{P}(A_i=B_i)=\mathbb{P}(A_i=B_i=H)+\mathbb{P}(A_i=B_i=T)\\ &= \mathbb{P}(A_i=H)\,\mathbb{P}(B_i=H)+\mathbb{P}(A_i=T)\,\mathbb{P}(B_i=T)\\ &= \frac{1}{2}\theta+\frac{1}{2}(1-\theta)=\frac{1}{2}.\end{aligned}$$

Nach Satz **??** hat R_n ein Wachstum von $R_n\sim\log_2(n)$. Ist $\theta=1$, also $\mathbb{P}(B_i=H)=1$, so entspricht H_n der Länge der längsten Übereinstimmungsfolge des Buchstaben H in $A_1A_2\ldots A_n$. Daher gilt wieder nach Satz **??** $H_n=R_n\sim\log_2(n)$. Analog gilt für $\theta=0$, dass H_n der Länge der längsten Übereinstimmungsfolge des Buchstaben T in $A_1A_2\ldots A_n$ entspricht und wieder $H_n=R_n\sim\log_2(n)$. Für $\theta=\frac{1}{2}$ haben die A_i und B_i dieselbe Verteilung und es gilt damit $H_n\sim2\log_2(n)$ nach Satz **??**.

Satz **??** besagt, dass $H_n \sim C \log_2(n)$ für ein $C \in [1,2]$ und $C = 2$ genau dann, wenn

$$max\left\{\mathcal{H}(\beta,\nu), \mathcal{H}(\beta,\xi)\right\} \leq \frac{1}{2}\log(2).$$

In diesem Fall ist

$$\begin{aligned}
\xi &= (\xi_H, \xi_T) = \left(\frac{1}{2}, \frac{1}{2}\right) \\
\nu &= (\nu_H, \nu_t) = (\theta, 1-\theta) \\
\beta &= (\beta_H, \beta_T) = \left(\frac{\xi_H \nu_H}{p}, \frac{\xi_T \nu_T}{p}\right) = \nu
\end{aligned}$$

und daher

$$\mathcal{H}(\beta,\nu) = \beta_H \log\left(\frac{\beta_H}{\nu_H}\right) + \beta_T \log\left(\frac{\beta_T}{\nu_T}\right) = 0.$$

Somit ist $C = 2$ genau dann, wenn $\mathcal{H}(\beta,\xi) \leq \frac{1}{2}\log(2)$. Hierbei ist

$$\mathcal{H}(\beta,\xi) = \beta_H \log\left(\frac{\beta_H}{\xi_H}\right) + \beta_T \log\left(\frac{\beta_T}{\xi_T}\right) = \theta\log(2\theta) + (1-\theta)\log(2(1-\theta)).$$

Löst man die Ungleichung nach θ, so erhält man, dass $C = 2$ genau dann, wenn $\theta \in [0.11002786\ldots, 0.88997214\ldots]$ gilt.