Seminar: Mathematische Biologie

A Phase Transition for the Score in Matching Random Sequences Allowing Deletions

Julian Hüne

1 Einleitung

In diesem Seminarvortrag wird als Hauptresultat der Beweis eines Phasenübergangs zwischen linearer und logarithmischer Entwicklung von Trefferzahlen (scores) erbracht.
Hierfür benötigen wir ein endliches Alphabet $\mathcal{A}$. Auf $\mathcal{A}^2$ wird eine Trefferfunktion $s(a,b)$ und eine subadditive Lückenfunktion (gap weight function) $g(k)$ mit $k \geq 1$, die den positiven Wert für die Zuordnung eines Buchstaben des Alphabets zu einer Lücke angibt, definiert.

Definition 1. *Sei* $(\mu, \delta) \in [0, \infty]^2$

$$s : \mathcal{A}^2 \to \mathbb{R}, \qquad s(a,b) = \begin{cases} +1, & a = b \\ -\mu, & a \neq b \end{cases} \tag{1}$$

$$g : \mathcal{A} \times \{-\} \to \mathbb{R}, \qquad g(k) = \delta k \tag{2}$$

Die folgenden Sätze sind auch für allgemeinere s und g gültig. Der Einfachheit halber sollen jedoch diese studiert werden.

In diesem Modell definieren wir die Zufallsvariablen für die globale und lokale Trefferabgleichung (global alignment score/ local alignment score) wie folgt:
Die globale Trefferabgleichung S_n soll das Maximum über alle möglichen Abgleichungen von zwei Buchstabenfolgen $A = A_1 ... A_n$ und $B = B_1 ... B_n$ sein, wobei alle Buchstaben A_i und B_j unabhängig und identisch verteilte Zufallsvariablen eines endlichen Alphabets sind.

Definition 2. *Für die globale Trefferabgleichung (global alignment score) gilt:*

$$\begin{aligned} S_n = S(A,B) = S(A_1 ... A_n, B_1 ..., B_n) \\ = \max_{l \in \{0,...,n\}} \{-\delta(n - l + n - l) + \sum_{k=1}^{l} s(A_{a(k)}, B_{b(k)})\}, \end{aligned} \tag{3}$$

wobei das Maximum über alle Abgleichungen ist, die durch die aufsteigenden Sequenzen:

$$\begin{aligned} 0 = a(0) < a(1) < a(2) < ... < a(l) < a(l+1) = n+1 \\ 0 = b(0) < b(1) < b(2) < ... < b(l) < b(l+1) = n+1 \end{aligned}$$

gegeben sind.

Die lokale Trefferabgleichung H_n ist nun die optimale Trefferzahl $S(I,J)$ mit $I \subset A$ und $J \subset B$ über alle möglichen Abgleichungen von zwei zusammenhängenden Buchstabenfolgen, d.h.

Definition 3. *Für die lokale Trefferabgleichung (local alignment score) gilt:*

$$\begin{aligned} H(A_1...A_n, B_1..., B_n) &\equiv H(A, B) \\ &\equiv H_n = max\{S(I, J) : I \subset A, J \subset B\}, \end{aligned} \tag{4}$$

wobei $I = A_{g+1}...A_{g+i}$ *und* $J = B_{h+1}...B_{h+j}$ *mit* $1 \leq g+1 \leq g+i \leq n$ *und* $1 \leq h+1 \leq h+j \leq n$. *Für* $S(A_{g+1}...A_{g+i}, B_{h+1}...B_{h+j})$ *gilt:*

$$\begin{aligned} &S(A_{g+1}...A_{g+i}, B_{h+1}...B_{h+j}) \\ &= \max_{\substack{g=a(0)<a(1)<...<a(l+1)=g+i+1 \\ h=b(0)<b(1)<...<b(l+1)=h+j+1}} \{-\delta(i-l+j-l) + \sum_{k=1}^{l} s(A_{a(k)}, B_{b(k)})\} \end{aligned}$$

Das Hauptresultat, dass bewiesen werden soll, ist das folgende Theorem:

Theorem 1. *Seien* $A_1, A_2, ...$ *und* $B_1, B_2, ...$ *zwei Buchstabenfolgen, wobei alle Buchstaben* A_i *und* B_j *unabhängig und identisch verteilte Zufallsvariablen eines endlichen Alphabets sind.*
Bei der lokalen Trefferabgleichung (local alignment score) $H_n = H(A_1A_2...A_n, B_1B_2...B_n)$ *mit den Parametern* μ *und* δ *tritt ein Phasenübergang zwischen linearer Entwicklung in n für kleine* μ *und* δ *und logarithmischer Entwicklung in n für große* μ *und* δ *auf.*

Beweis. Die Behauptung folgt unmittelbar durch Kombination von Satz 4, 5 und 6. □

An die folgenden beiden Sätze sei erinnert:

Satz 1. *Seien* $A = A_1...A_n$ *und* $B = B_1...B_n$ *mit* A_i *und* B_j *iid. Es existiert eine Konstante* $\rho \geq \mathbb{E}(s(A, B))$, *so dass:*

$$\begin{gathered} \lim_{n\to\infty} \frac{S_n}{n} = \rho \textit{ fast sicher} \\ \rho = \rho(\mu, \delta) = \lim_{n\to\infty} \frac{\mathbb{E}(S_n)}{n} = \sup_{n\geq 1} \frac{\mathbb{E}(S_n)}{n} \end{gathered} \tag{5}$$

Beweisidee: Überprüfe Kingman's Theorem und Subadditivität □

Satz 2 (Korrolar des Azuma-Hoeffding Lemma)**.** *Seien* $A_1, ..., A_n$ *und* $B_1, ..., B_n$ *zwei Buchstabenfolgen, wobei alle Buchstaben* A_i *und* B_j *unabhängig und identisch verteilte Zufallsvariablen eines endlichen Alphabets sind, und sei* γ>0. *Dann gilt mit* $c := \min\{2+4\delta, 2+2\mu\}$

$$\mathbb{P}(\frac{S_n}{n} - \rho \geq \gamma) \leq \exp(-\frac{\gamma^2 n}{2c^2}) \tag{6}$$

2 subadditive Folgen

Definition 4. *Eine Folge* $(a_n)_{n\in\mathbb{N}}$ *heißt subadditiv, wenn*

$$a_{n+m} \leq a_n + a_m \qquad \forall n, m \in \mathbb{N} \tag{7}$$

Definition 5. *Eine Folge* $(a_n)_{n\in\mathbb{N}}$ *heißt superadditiv, wenn*

$$a_{n+m} \geq a_n + a_m \qquad \forall n, m \in \mathbb{N} \tag{8}$$

Bemerkung 1. *Die globale Trefferabgleichung (global alignment score)* S_n *ist eine superadditive Folge*

$$S_{n+m} \geq S_n + S_m \tag{9}$$

Satz 3 (Fekete's Lemma)**.** *Für jede superadditive Folge* $(a_n)_{n\in\mathbb{N}}$ *existiert der Grenzwert* $\lim_{n\to\infty} \frac{a_n}{n}$ *und es gilt:*

$$\lim_{n\to\infty} \frac{a_n}{n} = \sup_{m\geq 1} \frac{a_m}{m} \tag{10}$$

Ganz analog, existiert der Grenzwert $\lim_{n\to\infty} \frac{a_n}{n}$ *für jede subadditive Folge* $(a_n)_{n\in\mathbb{N}}$ *und es gilt:*

$$\lim_{n\to\infty} \frac{a_n}{n} = \inf_{m\geq 1} \frac{a_m}{m} \tag{11}$$

Beweis. Zunächst einmal gilt:

$$\lim_{j\to\infty} \sup_{n\geq j} \frac{a_n}{n} \leq \sup_{m\geq 1} \frac{a_m}{m}$$

Also bleibt zu zeigen: $\forall m \in \mathbb{N}$ gilt: $\lim\limits_{j\to\infty} \inf\limits_{n\geq j} \frac{a_n}{n} \geq \frac{a_m}{m}$
Sei $m \in \mathbb{N}$
und $n = km + l$ mit $0 \leq l < m$.
Mit mehrfachem Anwenden von (8) folgt:

$$a_n \geq ka_m + a_l$$

Division durch n ergibt

$$\begin{aligned}
\frac{a_n}{n} \geq \frac{ka_m + a_l}{n} &= \frac{km}{km+l}\frac{a_m}{m} + \frac{s_l}{n} \\
\lim_{j\to\infty} \inf_{n\geq j} \frac{a_n}{n} &\geq \lim_{j\to\infty} \inf_{n\geq j} \left(\frac{km}{km+l}\frac{a_m}{m} + \frac{s_l}{n}\right) \\
&= \frac{a_m}{m}
\end{aligned}$$

und so folgt die Behauptung, denn:

$$\sup_{m\geq 1} \frac{a_m}{m} \leq \lim_{j\to\infty} \inf_{n\geq j} \frac{a_n}{n} \leq \lim_{j\to\infty} \sup_{n\geq j} \frac{a_n}{n} \leq \sup_{m\geq 1} \frac{a_m}{m}$$

Die 2. Behauptung folgt ganz analog. □

3 Der Phasenübergang

Zuerst behandeln wir die Menge mit $\rho = 0$

Satz 4. *Die Menge* $\{(\mu, \delta) \in [0, \infty]^2 : \rho(\mu, \delta) = 0\}$ *definiert eine Linie in dem Parameterraum* $[0, \infty]^2$*, die positive und negative Regionen* $\{\rho < 0\}$ *und* $\{\rho > 0\}$ *separiert.*

Beweis. ρ ist monoton fallend in beiden Parametern. Dies folgt aus der Definition der globalen Trefferabgleichung (global alignment score).
Behauptung: Es gilt die globale Ungleichung $\rho(\mu + \epsilon, \delta + \frac{\epsilon}{2}) \geq \rho(\mu, \delta) - \epsilon$.

Es gilt

$$\begin{aligned} S_k &= l * 1 - j\mu - i\delta \text{ mit } l, j, i, k \in \mathbb{N} \text{ s.d. } l + j + \frac{i}{2} = k \\ S'_k &:= l * 1 - j(\mu + \epsilon) - i(\delta + \frac{\epsilon}{2}) \end{aligned}$$

Somit folgt:

$$\begin{aligned} \frac{S'_k}{k} &= \frac{l * 1 - j(\mu + \epsilon) - i(\delta + \frac{\epsilon}{2})}{k} \\ &= \frac{l - j\mu - i\delta}{k} + \frac{-j\epsilon - i\frac{\epsilon}{2}}{k} = \frac{S_k}{k} - \frac{\epsilon(j + \frac{i}{2})}{k} \\ &\geq \frac{S_k}{k} - \frac{k\epsilon}{k} && (j + \frac{i}{2} \leq k) \\ &= \frac{S_k}{k} - \epsilon \end{aligned}$$

Dies gilt für alle $k \in \mathbb{N}$ und wir wissen, dass:

$$\frac{S_k}{k} \to \rho(\mu, \delta) \text{ fast sicher und } \frac{S'_k}{k} \to \rho(\mu + \epsilon, \delta + \frac{\epsilon}{2}) \text{ fast sicher}$$

So folgt:

$$\rho(\mu + \epsilon, \delta + \frac{\epsilon}{2}) \geq \rho(\mu, \delta) - \epsilon$$

Behauptung: ρ ist stetig.
sei $x_0 = (\mu_0, \delta_0) \in [0, \infty]^2$
Zu zeigen:
$\forall \epsilon > 0 \; \exists \nu > 0 : \forall x$ mit $||x, x_0||_1 \leq \nu$ gilt $|\rho(x) - \rho(x_0)| \leq \epsilon$
Setze $\nu := \frac{\epsilon}{2}$
Da ρ in beiden Parametern monoton fallend ist gilt:
$\exists$ Q, Q' $\in \{x \in [0, \infty]^2 : ||x, x_0||_1 \leq \frac{\epsilon}{2}\}$ mit

$Q = (\mu_0 + \frac{\epsilon}{4}, \delta_0 + \frac{\epsilon}{4})$ und $Q' = (\mu_0 - \frac{\epsilon}{4}, \delta_0 - \frac{\epsilon}{4})$, so dass gilt:

$$\begin{aligned}
|\rho(x), \rho(x_0)| &\leq |\rho(Q), \rho(Q')| \\
&= |\rho(\mu_0 + \frac{\epsilon}{4}, \delta_0 + \frac{\epsilon}{4}), \rho(\mu_0 - \frac{\epsilon}{4}, \delta_0 - \frac{\epsilon}{4})| \\
&= \rho(\mu_0 - \frac{\epsilon}{4}, \delta_0 - \frac{\epsilon}{4}) - \rho(\mu_0 + \frac{\epsilon}{4}, \delta_0 + \frac{\epsilon}{4}) \\
&\leq \rho(\mu_0 - \frac{\epsilon}{2}, \delta_0 - \frac{\epsilon}{4}) - \rho(\mu_0 + \frac{\epsilon}{2}, \delta_0 + \frac{\epsilon}{4}) && \text{(Monotonie von } \rho) \\
&\leq \rho(\mu_0 - \frac{\epsilon}{2}, \delta_0 - \frac{\epsilon}{4}) - \rho(\mu_0, \delta_0) + \frac{\epsilon}{2} && \text{(globale Ungleichung)} \\
&= \rho(\mu_0', \delta_0') - \rho(\mu_0' + \frac{\epsilon}{2}, \delta_0' + \frac{\epsilon}{4}) && (\mu_0' := \mu_0 - \frac{\epsilon}{2} \\
& && \delta_0' := \delta_0 - \frac{\epsilon}{4}) \\
&\leq \rho(\mu_0', \delta_0') - \rho(\mu_0', \delta_0') + \frac{\epsilon}{2} + \frac{\epsilon}{2} && \text{(globale Ungleichung)} \\
&= \epsilon
\end{aligned}$$

ρ ist nicht streng monoton fallend in beiden Parametern überall im Parameterraum $[0, \infty]^2$. Aber es gilt:
Behauptung: ρ ist streng monoton fallend in der (1,1)-Richtung in einer Umgebung der Linie $\rho = 0$.
Sei zunächst $(\mu, \delta) \in [0, \infty)^2$. Dafür setze $\gamma := max\{\mu, 2\delta\}$.
Es gilt:

$$\begin{aligned}
S_k &= l - j\mu - \frac{i}{2}2\delta \text{ (mit l, j, i, k} \in \mathbb{N}\text{, s.d. } l + j + \frac{i}{2} = k) \\
&\geq l - (j + \frac{i}{2})\gamma
\end{aligned}$$

Sei x der Anteil von nicht übereinstimmend abgeglichenen Buchstabenpaaren:

$$S_k \geq k(1 - x) - kx\gamma$$

Und somit:

$$\frac{S_k}{k} \geq (1 - x) - x\gamma \Leftrightarrow x \geq \frac{1 - \frac{S_k}{k}}{\gamma + 1} \tag{12}$$

Erhöht man bei solchen Abgleichungen die Parameter μ, δ um jeweils ϵ>0, so muss die Trefferzahl (score) um mindestens ϵx abnehmen, d.h. mit
$S_k' = l - j(\mu + \epsilon) - \frac{i}{2}2(\delta + \epsilon)$ folgt für alle $k \in \mathbb{N}$:

$$\begin{aligned}
\frac{S_k'}{k} &\leq \frac{S_k}{k} - \epsilon x \overset{(12)}{\leq} \frac{S_k}{k} - \epsilon \frac{1 - \frac{S_k}{k}}{\gamma + 1} \\
&= \frac{S_k}{k} + \epsilon \frac{\frac{S_k}{k}}{\gamma + 1} - \epsilon \frac{1}{\gamma + 1}
\end{aligned}$$

Und da wir außerdem wissen, dass gilt:

$$\frac{S_k}{k} \to \rho(\mu, \delta) \text{ fast sicher}$$
$$\frac{S'_k}{k} \to \rho(\mu + \epsilon, \delta + \epsilon) \text{ fast sicher,}$$

folgt schließlich:

$$\begin{aligned}
\rho(\mu + \epsilon, \delta + \epsilon) &\leq \rho(\mu, \delta) + \epsilon \frac{\rho(\mu, \delta)}{\gamma + 1} - \frac{\epsilon}{\gamma + 1} \\
&= \rho(\mu, \delta) - \epsilon \frac{1 - \rho(\mu, \delta)}{\gamma + 1} \\
&\leq \rho(\mu, \delta) - \epsilon \frac{1 - \rho(\mu, \delta)}{1 + \mu + 2\delta}
\end{aligned}$$

Da wir uns in einer kleinen Umgebung von $\rho = 0$ befinden, gilt:

$$\rho(\mu + \epsilon, \delta + \epsilon) \leq \rho(\mu, \delta) - \underbrace{\epsilon \frac{1 - \rho(\mu, \delta)}{1 + \mu + 2\delta}}_{>0} < \rho(\mu, \delta)$$

Sei nun $\mu = \infty$ und $\delta \in [0, \infty)$. Auch dann gilt, dass $\rho(\infty, \delta)$ streng monoton fallend in der (1,1)-Richtung ist.
Da $\mu = \infty$ ist, werden nicht übereinstimmende Buchstaben jeweils zu einer Lücke abgeglichen. Das bedeutet: $S_k = l * 1 - \frac{i}{2} 2\delta$. Somit folgt:

$$S_k \geq k(1 - x) - kx2\delta \Leftrightarrow x \geq \frac{1 - \frac{S_k}{k}}{2\delta + 1}$$

Und so folgt genauso auch hier:

$$\rho(\infty + \epsilon, \delta + \epsilon) \leq \rho(\infty, \delta) - \epsilon \frac{1 - \rho(\mu, \delta)}{1 + 2\delta} < \rho(\infty, \delta)$$

Sei als letztes $\delta = \infty$ und $\mu \in [0, \infty)$
Dieser Fall ist analog zu dem vorherigen. Die optimalen Abgleichungen von Sequenzen in diesem Fall beinhalten keine Lücken. □

Bemerkung 2. *Der Satz sagt nicht, dass für alle $\epsilon > 0$:*
$\rho(\mu + \epsilon, \delta) < \rho(\mu, \delta)$ *und* $\rho(\mu, \delta + \epsilon) < \rho(\mu, \delta)$. *Insbesondere das 1. ist falsch.*

Definition 6. *Für alle $q \in [0, \infty)$ definieren wir die Funktion*
$r : [0, \infty) \to [0, \infty]$ *durch:*

$$r(q) := \lim_{n \to \infty} \frac{-\log \mathbb{P}(S_n \geq qn)}{n} = \inf_{n \geq 1} \frac{-\log \mathbb{P}(S_n \geq qn)}{n} \tag{13}$$

Bemerkung 3. *Wir setzen* $\log(0) = -\infty$

Lemma 1. *Der Grenzwert existiert und ist gleich dem Infimum.*

Beweis. Dies beweisen wir mit Hilfe von Fekete's Lemma. Es gilt:

$$\begin{aligned}
\mathbb{P}(S_{n+l} \geq q(n+l)) &\geq \mathbb{P}(S_n + S(A_{n+1}...A_{n+l}, B_{n+1}...B_{n+l}) \geq qn + ql)) \\
&= \mathbb{P}(S_n + S_l \geq qn + ql) \qquad \text{(identisch verteilt)} \\
&\geq \mathbb{P}(S_n \geq qn \wedge S_l \geq ql) \\
&= \mathbb{P}(S_n \geq qn)\mathbb{P}(S_l \geq ql) \qquad \text{(Unabhängigkeit)}
\end{aligned}$$

Somit folgt:

$$-\log(\mathbb{P}(S_m \geq qm)) \leq -\log(\mathbb{P}(S_n \geq qn)) + (-\log(\mathbb{P}(S_l \geq ql)))$$

Das zeigt, dass $-\log \mathbb{P}(S_n \geq qn)$ eine subadditive Folge ist. Aufgrund Fekete's Lemma wissen wir:

$$\lim_{n\to\infty} -\frac{\log \mathbb{P}(S_n \geq qn)}{n} = \inf_{n\geq 1} -\frac{\log \mathbb{P}(S_n \geq qn)}{n}$$

und der Grenzwert existiert. □

Als nächstes werden einige Eigenschaften der Funktion r aufgeführt, die der Anschauung dienen und in den späteren Beweisen benötigt werden.

Bemerkung 4. *r ist monoton steigend*

Beweis. $\mathbb{P}(S_n \geq qn)$ ist für alle $n \in \mathbb{N}$ monoton fallend in q. Die Monotonie überträgt sich auch auf den Grenzwert. Somit ist r monoton steigend. □

Bemerkung 5. *Für* $q \in [0, 1]$ *gilt:* $0 \leq r(q) < \infty$ *und für* $q > 1$*:* $r(q) = \infty$.

Beweis. $r(q) \geq 0$, denn:

$$-\frac{\log \mathbb{P}(S_n \geq qn)}{n} \geq 0 \qquad \forall n \in \mathbb{N}$$

Und da r monoton steigend ist gilt für alle q $\in [0, 1]$:

$$\begin{aligned}
r(q) \leq r(1) &= \inf_{n\geq 1} -\frac{\log \mathbb{P}(S_n \geq n)}{n} \\
&= \inf_{n\geq 1} -\frac{\log \mathbb{P}(A_i = B_i \;\; \forall i \in \{1, ..., n\})}{n} \\
&= \inf_{n\geq 1} -\frac{\log((\mathbb{P}(A_1 = B_1))^n)}{n} \qquad \text{(iid)} \\
&= \inf_{n\geq 1} -\log(\mathbb{P}(A_1 = B_1)^{\frac{n}{n}}) \\
&= -\log \mathbb{P}(A_1 = B_1) \qquad (14) \\
&< \infty
\end{aligned}$$

Da $\frac{S_n}{n} \leq 1$ für alle $n \in \mathbb{N}$ gilt für q>1:

$$\mathbb{P}(S_n \geq qn) = 0 \;\; \forall n \in \mathbb{N}$$
$$\Rightarrow r(q) = \lim_{n\to\infty} -\log(0^{\frac{1}{n}}) = -\log(0) = \infty$$

$\square$

Bemerkung 6. *Wenn $\rho(\mu, \delta) > q$, dann folgt r(q)=0.*

Beweis. Wir wissen, dass

$$\frac{S_n}{n} \to \rho \text{ fast sicher}$$

S_n ist eine superadditive Folge, d.h. $S_{n+m} \geq S_n + S_m$. Somit folgt aus Fekete's Lemma

$$\lim_{n\to\infty} \frac{S_n}{n} = \sup_{n\geq 1} \frac{S_n}{n}$$

So gilt:

$$\rho \geq \frac{S_n}{n} \;\; \forall n \in \mathbb{N} \text{ fast sicher}$$

Sei $q < p$ und setze $q = \rho - \epsilon$ mit $\epsilon > 0$. So folgt:

$$\begin{aligned}
\mathbb{P}(S_n \geq qn) &= \mathbb{P}(\frac{S_n}{n} \geq \rho - \epsilon) = \mathbb{P}(\rho - \frac{S_n}{n} \leq \epsilon) \\
&= 1 - \mathbb{P}(\rho - \frac{S_n}{n} > \epsilon) \\
&= 1 - \mathbb{P}(\rho - \frac{S_n}{n} > \epsilon) - \underbrace{\mathbb{P}(-\rho + \frac{S_n}{n} > \epsilon)}_{=0, \text{ da } \rho \geq \frac{S_n}{n} f.s} \\
&= 1 - \mathbb{P}(\rho - \frac{S_n}{n} > \epsilon \vee -\rho + \frac{S_n}{n} > \epsilon) \\
&= 1 - \mathbb{P}(|\rho - \frac{S_n}{n}| > \epsilon) \quad \underbrace{\xrightarrow{n\to\infty} 1}_{\text{da } \frac{S_n}{n} \to \rho \text{ f.s.}} \quad ,
\end{aligned}$$

denn aus fast sicherer Konvergenz folgt Konvergenz in Wahrscheinlichkeit. Jetzt können wir zeigen, dass r(q)=0 gilt für alle $q < \rho$

$$\begin{aligned}
0 \leq r(q) &= \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}(S_n \geq qn) \\
&= \lim_{n\to\infty} -\log(\mathbb{P}(S_n \geq qn)^{\frac{1}{n}}) \\
&= -\log(\lim_{n\to\infty} \mathbb{P}(S_n \geq qn)^{\frac{1}{n}}) && \text{(Stetigkeit)} \\
&= -\log(\lim_{n\to\infty} (1 - \mathbb{P}(|\rho - \frac{S_n}{n}| \leq \epsilon))^{\frac{1}{n}})
\end{aligned}$$

$\exists \epsilon' > 0$ klein: $\exists\ N \in \mathbb{N}\ \forall\ n > N\ \mathbb{P}(|\rho - \frac{S_n}{n}| \leq \epsilon))^{\frac{1}{n}}) < \epsilon'$

$$\leq -\log(\lim_{n\to\infty}(1-\epsilon')^{\frac{1}{n}})$$
$$= -\log(1) = 0$$

□

Während S_k nur Trefferzahlen (scores) zwischen Sequenzen der Länge k angibt, hängt H_k von Sequenzen beliebiger Länge zwischen 1 und k ab.

Definition 7.

$$S_{i,j} := S(A_1...A_i, B_1...B_j) \tag{15}$$

$$r^{'}(q) := \lim_{k\to\infty} -\frac{1}{k}\log(\max_{i+j=2k} \mathbb{P}(S_{i,j} \geq qk)) \tag{16}$$

Lemma 2. *Es gilt:*

$$\begin{aligned} r(q) &= \lim_{k\to\infty} \frac{-\log \mathbb{P}(S_k \geq qk)}{k} \\ &= \lim_{k\to\infty} -\frac{1}{k}\log(\max_{i+j=2k} \mathbb{P}(S_{i,j} \geq qk)) = r^{'}(q) \end{aligned} \tag{17}$$

Beweis. Da $-\log(\max_{i+j=2k} \mathbb{P}(S_{i,j} \geq qk))$ eine subadditive Folge ist (vgl. Lemma 1), folgt mit Fekete's Lemma:

$$\lim_{n\to\infty} -\frac{\log(\max_{i+j=2k} \mathbb{P}(S_{i,j} \geq qk))}{k} = \inf_{n\geq 1} -\frac{\log(\max_{i+j=2k} \mathbb{P}(S_{i,j} \geq qk)}{k}$$

und der Grenzwert existiert.
Zu zeigen: $r^{'} \leq r$
Es gilt:

$$\{i, j : i = k \wedge j = k\} \subseteq \{i, j : i + j = 2k\}$$

Und somit:

$$\begin{aligned} \mathbb{P}(S_k \geq qk) &\leq \max_{i+j=2k} \mathbb{P}(S_{i,j} \geq qk) \\ \frac{-\log \mathbb{P}(S_k \geq qk)}{k} &\geq -\frac{1}{k}\log(\max_{i+j=2k} \mathbb{P}(S_{i,j} \geq qk)) \end{aligned}$$

Die Ungleichung bleibt im Grenzwert für k $\to \infty$ erhalten.
Also ist $r^{'} \leq r$

Zu zeigen $r^{'} \geq r$
zeige daher:

$$r^{'} \geq r - \epsilon \text{ für alle } \epsilon > 0$$

seien i, j und $k = \frac{i+j}{2}$ groß genug, s.d.:

$$r^{'} \leq -\frac{1}{k} \log \mathbb{P}(S_{i,j} \geq qk) < r^{'} + \epsilon \tag{18}$$

Es gilt zunächst einmal:

$$\begin{aligned}
&\mathbb{P}(S_{2k} = S_{2k,2k} \geq q(2k)) && (19)\\
&\quad = \mathbb{P}(S_{i+j,i+j} \geq q(2k)) \\
&\quad \geq \mathbb{P}(S_{i,j} + S(A_{i+1}...A_{i+j}, B_{j+1}...B_{j+i}) \geq q(2k)) && \text{(Superadditivität)}\\
&\quad = \mathbb{P}(S_{i,j} + S_{j,i}) \geq q(2k)) && \text{(identisch verteilt)}\\
&\quad = \mathbb{P}(S_{i,j} + S_{i,j}) \geq q(2k)) && \text{(Symm. zw. A,B)}\\
&\quad \geq \mathbb{P}(S_{i,j} \geq qk \wedge S_{i,j} \geq qk) \\
&\quad = \mathbb{P}(S_{i,j} \geq qk)\mathbb{P}(S_{i,j} \geq qk) && \text{(Unabhängigkeit)}\\
&\quad = (\mathbb{P}(S_{i,j} \geq qk))^2 && (20)
\end{aligned}$$

Außerdem gilt weiter:

$$\begin{aligned}
r = r(q) &\overset{Def}{\leq} -\frac{1}{2k} \log \mathbb{P}(S_{2k} \geq 2qk) \\
&\overset{(20)}{\leq} -\frac{1}{2k} \log(\mathbb{P}(S_{i,j} \geq qk))^2 \\
&= -\frac{1}{k} \log \mathbb{P}(S_{i,j} \geq qk) \overset{(18)}{<} r^{'} + \epsilon
\end{aligned}$$

$\Rightarrow r = r^{'}$ □

Als nächstes soll gezeigt werden, dass für große μ und δ die lokale Trefferabgleichung (local alignment score) sich logarithmisch in n entwickelt.
Um dies zu beweisen, benötigen wir zunächst folgende Vorbemerkungen:

Lemma 3. *Wenn $\rho(\mu, \delta) < 0$ und $q \geq 0$, dann ist auch $r(q) > 0$.*

Beweis. Es gilt:

$$\begin{aligned}
\mathbb{P}(S_n \geq qn) &= \mathbb{P}(S_n - \rho n \geq (q - \rho)n) && (q - \rho > 0)\\
&\overset{(6)}{\leq} \exp\left(\frac{-(q-\rho)^2 n}{2c^2}\right) && \text{(Azuma-Hoeff.-Ungl.)}
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \frac{-\log \mathbb{P}(S_n \geq qn)}{n} &\geq -\frac{1}{n} \log\left(\exp\left(\frac{-(q-\rho)^2 n}{2c^2}\right)\right) \\
&= \frac{(q-\rho)^2}{2c^2} \overset{\rho<0, q\geq 0}{>} 0
\end{aligned}$$

Dies ist unabhängig von n, also gilt $r(q) > 0$. □

Bemerkung 7. *Es gilt sogar allgemeiner: wenn* $q > \rho(\mu, \delta)$ *folgt* $r(q) > 0$.

Beweis. Dies ist sofort aus dem vorherigem Beweis ersichtlich. □

Definition 8. *Sei* $\rho(\mu, \delta) < 0$, *dann definiere:*

$$b = b(\mu, \delta) = \max_{q \geq 0} \frac{q}{r(q)} = \max_{q \in [0,1]} \frac{q}{r(q)} \tag{21}$$

Bemerkung 8. *Es gilt:*
$b > 0$, *da* $r(1) \overset{(14)}{=} -\log \mathbb{P}(A_1 = B_1)$. *Außerdem wird das Maximum auf dem Kompaktum* $[0, 1]$ *angenommen und ist endlich.*

Die Idee dieser Definition ist die folgende:
Möchte man zwei Sequenzen der Länge n vergleichen und die lokale Trefferabgleichung (local alignment score) H_n bestimmen, so gibt es $(n - t + 1) * (n - t + 1)$ Möglichkeiten für eine Abgleichung der Länge t. $\mathbb{P}(S_t \geq qt)$ ist die Wahrscheinlichkeit, dass die Trefferzahl der Teilabgleichung S_t größer als qt ist. Die erwartete Anzahl von Abgleichungen der Länge t mit einer Trefferzahl von mindestens qt ist $(n - t + 1)^2 * \mathbb{P}(S_t \geq qt)$:

$$\begin{aligned}
X &:= \#\text{Abgleichungen der Länge t mit } S_t \geq qt \\
X_{i,j} &:= 1_{\{\text{Sequenzstück von i bis i+t abgeglichen mit Sequenzstück von j bis j+t, mit } S_t \geq qt\}} \\
X &= \sum_{i=1}^{n-t+1} \sum_{j=1}^{n-t+1} X_{i,j} \\
\mathbb{E}[X] &= \mathbb{E}[\sum_{i=1}^{n-t+1} \sum_{j=1}^{n-t+1} X_{i,j}] = (n - t + 1)^2 \mathbb{E}[X_{i,j}] \qquad \text{(identisch verteilt)} \\
&= (n - t + 1)^2 \mathbb{P}(S_t \geq qt)
\end{aligned}$$

Hält man t fest, wird n groß und beachtet man, dass gilt:
$\mathbb{P}(S_t \geq qt) \cong \exp(-tr(q))$, so gilt für die erwartete Anzahl an Teilsequenzen mit einer Trefferzahl von mindestens qt approximativ:

$$\mathbb{E}[X] \cong n^2 \exp(-tr(q))$$

Nehmen wir weiter an, dass die lokale Trefferabgleichung (local alignment score) eindeutig ist. So gilt:

$$\begin{aligned}
1 &= n^2 \exp(-tr(q)) \\
t &= \frac{log(n^2)}{r(q)}
\end{aligned}$$

Und somit sollte für die maximale Trefferzahl gelten:

$$\max_{q \in [0,1]} qt = \max_{q \in [0,1]} \frac{q}{r(q)} \log(n^2) = 2b \log(n)$$

Satz 5. *Für alle* $(\mu,\delta) \in [0,\infty]^2$ *mit* $\rho(\mu,\delta) < 0$ *gilt:*

$$\mathbb{P}((1-\epsilon)b < \frac{H_n}{\log(n)} < (2+\epsilon)b) \to 1 \textit{ für alle } \epsilon > 0$$

Beweis. Wir führen den Beweis in zwei Schritten, für die untere Grenze und für die obere Grenze.

1. Die untere Grenze
zu zeigen:

$$\mathbb{P}((1-\epsilon)b\log(n) < H_n) \to 1 \text{ für alle } \epsilon > 0$$

Sei $\epsilon > 0$, sei $\gamma > 0$ klein genug und $q > 0$ mit $\frac{q}{r(q)} \approx b$, so dass

$$(1-\epsilon)b(\frac{r(q)+\gamma}{q}) < 1 - \frac{\epsilon}{2} \tag{22}$$

Das ist möglich, denn:

$$\frac{r(q)+\gamma}{q} \approx \frac{1}{b} + \frac{\gamma}{q}$$
$$\Rightarrow (1-\epsilon)(1+\frac{b\gamma}{q}) < 1 - \frac{\epsilon}{2} \text{ für } \gamma \text{ klein genug}$$

Sei $t = (1-\epsilon)b\log(n)$.
Weiter setze $k = \lfloor \frac{t}{q} \rfloor$.
Damit existiert ein $c \in \mathbb{R}^+ : k \approx c\log(n)$
Ist n groß genug (je größer n ist, desto größer ist k), so wird k die folgende Ungleichung erfüllen:

$$r(q) \leq -\frac{1}{k}\log(\mathbb{P}(S_k \geq qk)) \leq r(q) + \gamma \tag{23}$$

(Infimum Eigenschaft)

Außerdem gilt:

$$\begin{aligned}
\mathbb{P}(S_k \geq qk) &= \exp(-k(-\frac{1}{k})\log(\mathbb{P}(S_k \geq qk))) \\
&\overset{(23)}{\geq} \exp(-k(r(q)+\gamma)) \\
&\geq \exp(-t(\frac{r(q)+\gamma}{q})) && k = \lfloor \frac{t}{q} \rfloor \leq \frac{t}{q} \\
&\overset{(22)}{\geq} \exp(-(1-\frac{\epsilon}{2})\log(n)) && t = (1-\epsilon)b\log(n) \\
&= n^{-1+\frac{\epsilon}{2}}
\end{aligned}$$

Nun zerteile $A_1...A_n$ und $B_1...B_n$ in sich nicht überlappende Blöcke der Länge k+1. So haben wir ungefähr $\frac{n}{k} \approx \frac{n}{c\log(n)}$ unabhängige Möglichkeiten für eine hohe Trefferzahl.
Jeder Block hat mindestens die Wahrscheinlichkeit $n^{-1+\frac{\epsilon}{2}}$ eine hohe Trefferzahl (score) zu erreichen.
Schließlich gilt dann mit $j = k+1$, so dass $t < qj$:

$$\begin{aligned}
&\mathbb{P}(H_n < (1-\epsilon)b\log(n)) \\
&\qquad = \mathbb{P}(H_n < t) \leq \mathbb{P}(H_n < qj) \\
&\qquad = \mathbb{P}(\max\{S(I,J) : I \subset A, B \subset J\} < qj) \\
&\qquad \leq \mathbb{P}(\bigcap_{0 \leq i \leq \lfloor \frac{n}{j} \rfloor - 1} \{S(A_{ij+1}...A_{ij+j}, B_{ij+1}...B_{ij+j}) < qj\}) \quad \text{(Teilmenge)} \\
&\qquad = \mathbb{P}(S_j < qj)^{\lfloor \frac{n}{j} \rfloor} \quad \text{(iid)} \\
&\qquad < (1 - n^{-1+\frac{\epsilon}{2}})^{\lfloor (\frac{n}{j}) \rfloor} \quad \text{(für großes n)} \\
&\qquad \longrightarrow 0 \text{ für } n \to \infty
\end{aligned}$$

Denn es gilt für kleine $\epsilon > 0$:

$$\begin{aligned}
a &:= 1 - \frac{\epsilon}{2} \\
1 - n^{-1+\frac{\epsilon}{2}} &= 1 - \frac{1}{n^a} = \frac{1}{1 + \frac{1}{n^a - 1}} =: \frac{1}{1 + \frac{1}{m}} \\
(1 - n^{-1+\frac{\epsilon}{2}})^{\lfloor \frac{n}{j} \rfloor} &= \frac{1}{(1 + \frac{1}{m})^{\lfloor \frac{n}{j} \rfloor}}
\end{aligned}$$

Zu zeigen:

$$(1 + \frac{1}{m})^{\lfloor \frac{n}{j} \rfloor} \to \infty$$

Weiter gilt:

$$(1 + \frac{1}{m})^{\lfloor \frac{n}{j} \rfloor} = (1 + \frac{1}{m})^{m \frac{1}{m} \lfloor \frac{n}{j} \rfloor}$$

Zu zeigen:

$$\frac{1}{m} \lfloor \frac{n}{j} \rfloor \to \infty$$

und das gilt, da für n groß genug $n^{1-a} > j$

2. die obere Grenze:
zu zeigen: $\mathbb{P}(H_n \geq (2+\epsilon)b\log(n)) \to 0$ für $n \to \infty$
sei $t = (2+\epsilon)b\log(n)$
$\mathbb{P}(H_n \geq (2+\epsilon)b\log(n)) = \mathbb{P}(H_n \geq t)$
Nun schauen wir uns die Menge $\{H_n \geq t\}$ genauer an:
$\{H_n \geq t\}$ besteht aus höchstens n^4 verschiedenen Elementen (beide Start und Endpunkte frei auswählbar zwischen 1 und n).
Wir teilen diese Menge in zwei Teilmengen, wobei die eine ein Vielfaches von $(n\log(n))^2$ Elemente beinhalten soll und die andere alle Restlichen.

$$\begin{aligned}
&\{H_n \geq t\} \\
&\quad = \{\max\{S(I,J) : I \subset A, J \subset B\} \geq t\} \\
&\quad \subseteq [\bigcup_{\substack{i_0,j_0\in[1,n] \\ i,j\leq n, i+j\leq 2C\log(n)}} \{S(A_{i_0+1}...A_{i_0+i}, B_{j_0+1}...B_{j_0+j}) \geq t\}] \\
&\qquad\qquad \cup [\bigcup_{\substack{i_0,j_0\in[1,n] \\ i,j\leq n, i+j > 2C\log(n)}} \{S(A_{i_0+1}...A_{i_0+i}, B_{j_0+1}...B_{j_0+j}) \geq 0\}]
\end{aligned} \tag{24}$$

1. Vereinigungsmenge
Sei $k = \frac{i+j}{2}$ und $t = qk$
Jedes Element hat höchstens die Wahrscheinlichkeit:

$$\begin{aligned}
\mathbb{P}(S_{i,j} \geq t) &= \mathbb{P}(S_{i,j} \geq qk) \\
&\leq \exp(-kr(q)) \\
&\qquad\qquad\qquad r'(q) \leq -\frac{1}{k}\log\mathbb{P}(S_{i,j} \geq qk) \\
&= \exp(-t\frac{r(q)}{q}) \\
&= \exp(-(2+\epsilon)(\max_c \frac{c}{r(c)})(\log(n))\frac{r(q)}{q}) \\
&\leq \exp(-(2+\epsilon)\log(n)) \\
&\qquad\qquad\qquad \text{denn } \max_c \frac{c}{r(c)} \geq \frac{q}{r(q)} \\
&\qquad\qquad\qquad \Leftrightarrow \max_c \frac{c}{r(c)}\frac{r(q)}{q} \geq 1 \\
&= n^{-(2+\epsilon)}
\end{aligned}$$

Da die 1. Vereinigung höchstens $n^2(2C\log(n))^2$ Elemente beinhaltet, gilt:

$$\begin{aligned}
\mathbb{P}[\bigcup_{\substack{i_0,j_0\in[1,n]\\ i,j\leq n, i+j\leq 2C\log(n)}} &\{S(A_{i_0+1}...A_{i_0+i}, B_{j_0+1}...B_{j_0+j})\geq t\}] \\
&\leq n^2(2C\log(n))^2 n^{-(2+\epsilon)} \\
&= \frac{4C^2(\log(n))^2}{n^\epsilon} = \frac{8C^2\log(n)}{n\epsilon n^{\epsilon-1}} \qquad \text{(l'Hopital)} \\
&= \frac{8C^2\log(n)}{\epsilon n^\epsilon} = \frac{8C^2}{\epsilon^2 n^{\epsilon-1} n} \\
&= \frac{8C^2}{\epsilon^2 n^\epsilon} \\
&\to 0 \text{ für } n\to\infty
\end{aligned}$$

2. Vereinigungsmenge
Diese setzt sich höchstens aus n^4 Elementen der Form $\{S_{i,j}\geq 0\}$ zusammen.
Jedes dieser Elemente lässt sich wie folgt abschätzen:
Sei $k=\frac{i+j}{2} > C\log(n)$ und $C=\frac{5}{r(0)}$. Da $\rho<0$, ist $r(0)>0$.
Dann folgt:

$$\begin{aligned}
\mathbb{P}(S_{i,j}\geq 0) &\leq \exp(-kr(0)) \qquad\qquad r'(0)\leq -\frac{1}{k}\log\mathbb{P}(S_{i,j}\geq 0) \\
&\leq \exp(-(C\log(n))r(0)) \\
&= \exp(-5\log(n)) = \frac{1}{n^5}
\end{aligned}$$

Und weiter gilt:

$$\begin{aligned}
\mathbb{P}[\bigcup_{\substack{i_0,j_0\in[1,n]\\ i,j\leq n, i+j\geq 2C\log(n)}} &\{S(A_{i_0+1}...A_{i_0+i}, B_{j_0+1}...B_{j_0+j})\geq 0\}] \\
&\leq n^4\frac{1}{n^5} = \frac{1}{n} \\
&\longrightarrow 0 \text{ für } n\to\infty
\end{aligned}$$

Schließlich setzen wir alles zusammen, so folgt:

$$\begin{aligned}
&\mathbb{P}(\{H_n\geq(2+\epsilon)b\log(n)\}) \\
&\quad\leq \mathbb{P}[\bigcup_{\substack{i_0,j_0\in[1,n]\\ i,j\leq n, i+j\leq 2C\log(n)}} \{S(A_{i_0+1}...A_{i_0+i}, B_{j_0+1}...B_{j_0+j})\geq t\}] \\
&\qquad + \mathbb{P}[\bigcup_{\substack{i_0,j_0\in[1,n]\\ i,j\leq n, i+j\geq 2C\log(n)}} \{S(A_{i_0+1}...A_{i_0+i}, B_{j_0+1}...B_{j_0+j})\geq 0\}] \\
&\quad\leq \frac{8C^2}{\epsilon^2 n^\epsilon}+\frac{1}{n} \\
&\quad\longrightarrow 0 \text{ für } n\to\infty
\end{aligned}$$

$\square$

Als letztes kommen wir zu dem Fall, dass $\rho(\mu, \delta) > 0$

Satz 6. *Wenn $\rho = \rho(\mu, \delta) > 0$, dann gilt:*

$$\frac{S_n}{n} \to \rho \textit{ in Wahrscheinlichkeit}$$
$$\textit{und}$$
$$\frac{H_n}{n} \to \rho \textit{ in Wahrscheinlichkeit}$$

Beweis. Es reicht zu zeigen:

$$\mathbb{P}(H_n > (1+\epsilon)n\rho) \to 0 \text{ für } n \to \infty \tag{25}$$
und
$$\mathbb{P}(S_n < (1-\epsilon)n\rho) \to 0 \text{ für } n \to \infty, \tag{26}$$

denn es gilt $H_n \geq S_n$ nach Definition und somit:

$$\mathbb{P}(S_n > (1+\epsilon)n\rho) \leq \mathbb{P}(H_n > (1+\epsilon)n\rho)$$
$$\mathbb{P}(H_n < (1-\epsilon)n\rho) \leq \mathbb{P}(S_n < (1-\epsilon)n\rho)$$

Sind die beiden Gleichungen (siehe oben) gezeigt, so folgt:

$$\mathbb{P}(H_n < (1-\epsilon)n\rho \vee (1+\epsilon)n\rho < H_n) \to 0 \text{ für } n \to \infty$$
$$\mathbb{P}(S_n < (1-\epsilon)n\rho \vee (1+\epsilon)n\rho < S_n) \to 0 \text{ für } n \to \infty$$

und somit die Beh., denn: sei $\epsilon > 0$

$$\mathbb{P}(H_n < (1-\epsilon)n\rho \vee (1+\epsilon)n\rho < H_n) \overset{n\to\infty}{\longrightarrow} 0$$
$$\Leftrightarrow \mathbb{P}(|\frac{H_n}{n} - \rho| > \epsilon\rho) \overset{\epsilon' := \rho\epsilon}{=} \mathbb{P}(|\frac{H_n}{n} - \rho| > \epsilon') \overset{n\to\infty}{\longrightarrow} 0$$
$$\Leftrightarrow \frac{H_n}{n} \to \rho \text{ in Wahrscheinlichkeit}$$

S_n ganz analog.

Wir zeigen zunächst die erste Gleichung: (26).
Laut Satz 1 gilt:

$$\frac{S_n}{n} \to \rho \text{ fast sicher}$$

Außerdem ist S_n eine superadditive Folge und somit folgt mit Fekete's Lemma:

$$\lim_{n\to\infty} \frac{S_n}{n} = \sup_{m\geq 1} \frac{S_m}{m} \Rightarrow \rho \geq \frac{S_n}{n} \;\; \forall n \text{ fast sicher}$$

So gilt weiter:

$$\begin{aligned}
\mathbb{P}(S_n < (1-\epsilon)n\rho) &= \mathbb{P}(\frac{S_n}{n} - \rho < -\epsilon\rho) \\
&\overset{\epsilon' := \epsilon\rho}{=} \mathbb{P}(\rho - \frac{S_n}{n} > \epsilon') \\
&= \mathbb{P}(\rho - \frac{S_n}{n} > \epsilon') + \underbrace{\mathbb{P}(-\rho + \frac{S_n}{n} > \epsilon')}_{=0, \text{ da } \rho \geq \frac{S_n}{n} f.s} \\
&= \mathbb{P}(\rho - \frac{S_n}{n} > \epsilon' \vee -\rho + \frac{S_n}{n} > \epsilon') \\
&= \mathbb{P}(|\rho - \frac{S_n}{n}| > \epsilon') \\
&= \mathbb{P}(|\frac{S_n}{n} - \rho| > \epsilon') \quad \underbrace{\xrightarrow{n\to\infty} 0}_{\text{da } \frac{S_n}{n} \to \rho \text{ f.s.}} \qquad \forall \epsilon' > 0
\end{aligned}$$

Aus fast sicherer Konvergenz folgt Konvergenz in Wahrscheinlichkeit.
Nun bleibt noch (25):

$$\mathbb{P}(H_n > (1+\epsilon)n\rho) \to 0 \text{ für } n \to \infty$$

zu zeigen.
Seien $i, j, n \in \mathbb{N}$ und $k = \frac{i+j}{2} \leq n$
Mit Hilfe von Lemma 2 gilt:

$$\begin{aligned}
r &:= r((1+\epsilon)\rho) \\
&= \lim_{k\to\infty} (-\frac{1}{k} \log \mathbb{P}(S_k \geq (1+\epsilon)\rho k)) \\
&\overset{(17)}{=} \inf_{k\geq 1} (-\frac{1}{k} \log \max_{i+j=2k} \mathbb{P}(S_{i,j} \geq (1+\epsilon)\rho k)) \\
&= r'
\end{aligned}$$

weiter gilt r>0, denn es gilt $\rho > 0$ und somit:

$$\begin{aligned}
\mathbb{P}(S_k \geq (1+\epsilon)\rho k) &= \mathbb{P}(\frac{S_k}{k} - \rho \geq \epsilon\rho) \\
&\overset{(6)}{\leq} \exp(-\frac{\epsilon^2\rho^2 k}{2c^2}) \qquad \text{(Azuma-Hoeff.-Ungl.)}
\end{aligned}$$

Und deshalb:

$$\begin{aligned}
-\frac{1}{k}\log \mathbb{P}(S_k \geq (1+\epsilon)\rho k) &\geq -\frac{1}{k}\log(\exp(-\frac{\epsilon^2\rho^2 k}{2c^2})) \\
&= \frac{\epsilon^2\rho^2}{2c^2} > 0 \text{ ,da } \rho \neq 0 \text{ und } \epsilon > 0
\end{aligned}$$

Dies ist unabhängig von k und damit gilt $r > 0$, so folgt weiter:

$$\begin{aligned}
\mathbb{P}(S_{i,j} \geq (1+\epsilon)\rho k) &= \exp(-k(-\frac{1}{k})\log(\mathbb{P}(S_{i,j} \geq (1+\epsilon)\rho k))) \\
&\leq \exp(-k(\inf_{k\geq 1}(-\frac{1}{k})\log(\mathbb{P}(S_{i,j} \geq (1+\epsilon)\rho k)))) \\
&\leq \exp(-k(\inf_{k\geq 1}(-\frac{1}{k})\log(\max_{i+j=k}\mathbb{P}(S_{i,j} \geq (1+\epsilon)\rho k)))) \\
&= \exp(-kr') \overset{(17)}{=} \exp(-kr)
\end{aligned}$$

Weiter gilt:

$$\begin{aligned}
\mathbb{P}(S_{i,j} \geq (1+\epsilon)n\rho) &\leq \mathbb{P}(S_{i,j} \geq (1+\epsilon)k\rho) && (n \geq k) \\
&\leq \exp(-rk)
\end{aligned}$$

Nun ergibt sich aufgrund der Trefferfunktion:

$$\begin{aligned}
S_{i,j} &= S(A_1 ... A_i, B_1 ... B_j) \\
&\leq k
\end{aligned}$$

So bedingt $S_{i,j} \geq (1+\epsilon)n\rho$:

$$k \geq (1+\epsilon)n\rho$$

Somit:

$$\begin{aligned}
\mathbb{P}(S_{i,j} \geq (1+\epsilon)n\rho) &\leq \exp(-rk) \\
&\leq \exp(-r(1+\epsilon)n\rho)
\end{aligned}$$

Zu guter Letzt gilt:

$$\begin{aligned}
\mathbb{P}(H_n \geq (1+\epsilon)n\rho) &= \mathbb{P}(\max\{S(I,J) : I \subset A, J \subset B\} \geq (1+\epsilon)n\rho) \\
&= \mathbb{P}(\bigcup_{\substack{i,j \\ k,l}}\{S(A_{i+1} ... A_{i+k}, B_{j+1} ... B_{j+l}) \geq (1+\epsilon)n\rho\}) \\
&\leq n^4\mathbb{P}(S_{i,j} \geq (1+\epsilon)n\rho) \\
&\qquad \text{(beide Start- und Endpunkte frei auswählbar)} \\
&\leq n^4\exp(-r(1+\epsilon)n\rho) \\
&\longrightarrow 0 \text{ für } n \to \infty
\end{aligned}$$

□

Damit ist das Theorem bewiesen.
Vergleicht man zwei Sequenzen der Länge n und vorausgesetzt, dass $n \to \infty$, so gibt es einen Phasenübergang zwischen linearer und logarithmischer Entwicklung. Das bedeutet:
Wenn $\rho(\mu, \delta) > 0$ wächst H_n mindestens so schnell wie $\rho(\mu, \delta)n$ und wenn $\rho(\mu, \delta) < 0$ ist, wächst H_n höchstens so schnell wie $b(\mu, \delta) \log(n^2)$ und mindestens so schnell wie $b(\mu, \delta) \log(n)$.
Auf der Linie $\{(\mu, \delta) : \rho(\mu, \delta) = 0\}$ findet dieser Phasenübergang statt.

4 Fazit

In dieser Arbeit wurde der Phasenübergang bewiesen für unabhängig und identisch verteilte Zufallsvariablen eines endlichen Alphabets. Außerdem haben wir eine spezielle Trefferfunktion und eine bestimmte Lückenfunktion vorgegeben:

$$s(a,b) = \begin{cases} +1, & a = b \\ -\mu, & a \neq b \end{cases}$$
$$g(k) = \delta k \text{ mit } (\mu,\delta) \in [0,\infty]^2$$

Der vorgestellte Beweis lässt sich auf Sequenzen unterschiedlicher Länge erweitern. Außerdem ist nicht zwingend erforderlich, dass $A_1, A_2, ..., B_1, B_2, ...$ unabhängig und identisch verteilt sind. Es genügt auch, dass $A_1, A_2, ...$ unabhängig und identisch nach μ verteilt sind und $B_1, B_2, ...$ nach ν.
Zur genaueren Interpretation von biologischen Sequenzabgleichungen benötigt man häufig ein komplizierteres Trefferschema $s(a,b)$, sowie eine komplexe Lückenfunktion $g(k)$.
Eine Verallgemeinerung besteht darin, eine allgemeine Lückenfunktion zuzulassen. Mit einer beliebigen Lückenfunktion $g : \mathbb{N} \to \mathbb{R}^+$ wird die subadditive Kostenfunktion definiert:

$$w(i) = \min\{\sum_{j=1}^{l} g(i_j) | l, i_1, ..., i_l \in \mathbb{N}, \sum_{j=1}^{l} i_j = i\}$$
$$w(0) := 0$$

Diese bewertet das Löschen von i zusammenhängenden Buchstaben. Damit ergibt sich für die globale Trefferabgleichung (global alignment score) die folgende Formel:

$$S_{n,m} = \max_{\substack{l \in \{0,...,\min\{n,m\}\} \\ 0=a(0)<...<a(l+1)=m+1 \\ 0=b(0)<...<b(l+1)=n+1}} \{-\sum_{k=1}^{l+1} w(a(k)-a(k-1)-1)$$
$$+w(b(k)-b(k-1)-1) + \sum_{k=1}^{l} s(A_{a(k)}, B_{b(k)})\}$$

Auch in diesem Fall erhält man den Phasenübergang in etwas allgemeinerer Form.
Außerdem lassen sich die meisten Ergebnisse auch auf zwei irreduzible aperiodische Markovketten verallgemeinern.