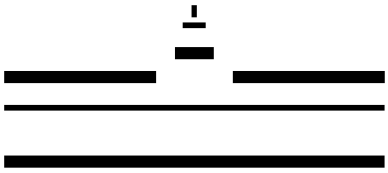

Westfälische
Wilhelms-Universität
Münster

Global Alignment of Random DNA Sequences

Seminararbeit
im Rahmen des Seminars Mathematische Biologie

Westfälische Wilhelms-Universität Münster
Fachbereich Mathematik und Informatik
Institut für Mathematische Statistik

Eingereicht von:
Karolina Weber

19.04.2012

Inhaltsverzeichnis

Global Alignment

1 Einleitung

Bei der Betrachtung von zwei gegebenen Sequenzen ist man an der Ähnlichkeit der Sequenzen interessiert. Wenn diese Sequenzen aber sehr lang sind, dann ist es nicht mehr so einfach zu entscheiden, ob sie sich ähnlich sind. Um dies zu sehen, müssen die Sequenzen zuerst anders angeordnet werden und Lücken (engl.: gaps) zugelassen werden. Das Vorhandensein eines Gaps deutet dabei auf das Hinzufügen (engl.: insertion) bzw. Wegfallen (engl.: deletion) von Residuen in der Evolution der Sequenz hin. Gegeben seien beispielsweise

$$\begin{aligned}\mathbf{A} &= TACCAGT,\\ \mathbf{B} &= CCCGTAA.\end{aligned}$$

Beide Sequenzen haben die selbe Länge und es gibt nur eine Möglichkeit sie global anzuordnen, wenn keine Gaps zugelassen sind. Wenn diese aber erlaubt sind, dann gibt es mehrere mögliche Anordnungen, z.B.

$$\begin{aligned}\mathbf{A} &= T\ \ A\ \ C\,C\ \ A\ \ G\,T\ -\ -,\\ \mathbf{B} &= C\ -\ C\,C\ -\ G\,T\ \ A\ \ A\\ &\quad \textit{oder}\\ \mathbf{A} &= T\ \ A\ \ C\,C\,A\,G\,T\ -\ -,\\ \mathbf{B} &= -\ -\ C\,C\,C\,G\,T\ \ A\ \ A.\end{aligned}$$

Um dabei entscheiden zu können welches Alignment das bessere ist, muss man den Score der Sequenzen berechnen können. Dabei ist zu beachten, dass die optimalen Alignments von der Scoring-Funktion abhängen und diese Funktion biologisch relevant sein muss.

In dieser Ausarbeitung werden zwei verschiedene Arten des Global Alignment und die Verteilungen der jeweiligen Scores betrachtet. In dem einen Fall ist die Anordnung der Sequenzen schon vornherein festgelegt und in dem anderen Fall ist diese nicht gegeben und muss optimal bestimmt werden.
Gegeben seien zwei Sequenzen $\mathbf{A} = A_1A_2 \dots A_n$ und $\mathbf{B} = B_1B_2 \dots B_m$ der Länge n und m. Dabei sind $A_i, B_j \in \mathcal{A}$ i.i.d. für $i = 1, \dots, n$, $j = 1, \dots, m$ und $\mathcal{A}$ ist ein gegebenes Alphabet.

Wenn die Anordnung der Sequenzen noch nicht festgelegt ist, kann man Lücken " – "hinzufügen, sodass man zwei neu angeordnete Sequenzen der gleichen Länge L erhält. Diese Alignment - Transformation sieht aus wie folgt:

$$\begin{aligned} A_1 ... A_n &\to A_1^* ... A_L^*, \\ B_1 ... B_m &\to B_1^* ... B_L^*. \end{aligned} \tag{1.1}$$

Hierbei ist zu beachten, dass alle Zufallsvariablen einer Sequenzen bei dem Alignment berücksichtigt werden und deren Reihenfolge auch nach der Transformation erhalten bleibt. D. h. die Teilsequenz für alle $A_i^* \neq " - "$ entspricht $A_1 ... A_n$.

2 Alignment Given

In diesem Abschnitt betrachten wir das Global Alignment für den Fall, dass die Anordnung schon vorab gegeben ist und es keine Gaps gibt (d. h. keine Insertions und Deletions). Beide Sequenzen haben die gleiche Länge, also $\mathbf{A} = A_1 ... A_n$ und $\mathbf{B} = B_1 ... B_n$ und es wird angenommen, dass das Alphabet $\mathcal{A}$ endlich ist.
Sei $s(A, B)$ eine reellwertige Funktion. Dann gilt für den Score S

$$S = \sum_{i=1}^{n} s(A_i, B_i). \tag{2.1}$$

Das folgende Theorem liefert Eigenschaften über den Erwartungswert,die Varianz und das Verhalten des Scores bei großem n.

Theorem 2.1. *Seien $\boldsymbol{A} = A_1 ... A_n$, $\boldsymbol{B} = B_1 ... B_n$ gegeben, wobei A_i und B_j i.i.d. und sei der Score wie in (2.1) definiert. Dann gilt*
(a) $\mathbb{E}(S) = n\mathbb{E}(s(A, B)) = n\mu$,
(b) $\mathrm{Var}(S) = n\,\mathrm{Var}(s(A, B)) = n\sigma^2$,
(c) $\lim_{n\to\infty} \mathbb{P}(\frac{S-n\mu}{\sqrt{n}\sigma} \leq x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\,dt$, *d. h. der Score konvergiert für große n gegen eine Standardnormalverteilung.*

3 Große Abweichungen für binomialverteilte Zufallsvariablen

In diesem Kapitel sei die reellwertige Funktion s definiert durch

$$s(a, b) = \begin{cases} 1 & \text{, a=b,} \\ 0 & \text{, a} \neq \text{b,} \end{cases} \tag{3.1}$$

und die Indel penalty Funktion sei $g(k) = \infty$ für Gaps der Länge k. Diese Funktion bewertet vorhandene Gaps mit $-g(k)$. Dabei ist die Funktion g nichtnegativ und subadditiv. Der Alignment Score $S(\mathbf{A}, \mathbf{B})$ für $\mathbf{A} = A_1 ... A_n$ und $\mathbf{B} = B_1 ... B_n$ mit fester Anordnung ist dann eine $\mathcal{B}(n, p)$ - verteilte Zufallsvariable, wobei $p = \mathbb{P}(A_i = B_j) = \mathbb{P}(s(A, B) = 1)$.
Das Ziel ist hierbei einen p - Wert für beobachtete Werte der binomialverteilten Zufallsvariablen zu schätzen, welcher weit vom Mittelwert entfernt ist. Große Abweichungen liefern dabei Abschätzungen, welche viel akkurater sind als die Abschätzungen durch den Zentralen Grenzwertsatz.

Sei $Y_n \sim \mathcal{B}(n, p)$ und die Erfolgsrate α für k von n Erfolgen gegeben durch

$$\alpha = k/n, \tag{3.2}$$

wobei $\alpha \in (p, 1)$.

Definition 3.1. *Gegeben sei die relative Entropie* $\mathcal{H} \equiv \mathcal{H}(\alpha, p) \equiv (\alpha)\, log\left(\frac{\alpha}{p}\right) + (1-\alpha)\, log\left(\frac{1-\alpha}{1-p}\right)$. *Der Wert* $\mathcal{H}$ *heißt Kullback-Leibler Distanz.*

Man kann beobachten, dass $\mathcal{H}(\alpha, p)$ von 0 bis $log(1/p)$ wächst mit wachsendem $\alpha \in (p, 1)$.
$\mathcal{H}$ misst die Distanz zwischen der $\mathcal{B}(n, p)$ - Verteilung (unter der die Daten generiert wurden) und der alternativen $\mathcal{B}(n, \alpha)$ - Verteilung. Das zentrale Konzept bei großen Abweichungen besteht darin, simultan mit zwei Wahrscheinlichkeitsmaßen auf dem gleichen Raum von möglichen Ereignissen zu arbeiten. Das folgende Theorem gibt eine obere Schranke, die für alle n, p und α gültig ist.

Theorem 3.2. *Für* $p < \alpha < 1$, $n = 1, 2, 3, ...$, *mit der relativen Entropie* $\mathcal{H} = \mathcal{H}(\alpha, p)$ *wie in Definition 3.1 und* $Y_n \sim \mathcal{B}(n, p)$ *gilt*

$$I\!P(Y_n \geq \alpha n) \leq e^{-n\mathcal{H}}. \tag{3.3}$$

Beweis. Betrachte $\mathbb{P}(Y_n \geq \alpha n)$ und wende innerhalb der Wahrscheinlichkeit die monoton wachsende Funktion e^{β} auf beiden Seiten der Ungleichung an. Nutze im Anschluss die Ungleichung von Markov, den binomischen Lehrsatz und die momenterzeugende Funktion $(e^{\beta}p + (1-p))^n$ der binomialverteilten Zufallsvariablen Y_n, um $\{e^{-\alpha\beta}(1 - p + pe^{\beta)})\}^n$ zu erhalten. Indem dieser Ausdruck nach β minimiert wird, erhält man $e^{-\mathcal{H}}$ und damit die zu beweisende Aussage. □

4 Alignment Unknown

Im Falle einer nicht vorgegebenen Anordnung von zwei Sequenzen $\mathbf{A} = A_1 ... A_n$ und $\mathbf{B} = B_1 ... B_m$ unterschiedlicher Länge wird diese durch Optimalität bestimmt. Wie in der Alignment-

Transformation (1.1) erläutert, werden Gaps für eine Anordnung benutzt. Dafür wird die reellwertige Funktion s erweitert durch $s(a,-)$ und $s(-,b)$. Dann gilt für den Alignment Score

$$S = max\left\{\sum_{i=1}^{L} s(A_i^*, B_i^*) : \text{alle Alignments}\right\}. \tag{4.1}$$

Der große Unterschied zu der Situation in Abschnitt 2 besteht darin, dass bei der Optimierung über die einzelnen Alignments die Normalverteilungseigenschaft des Scores abhanden kommt. Deshalb widmet sich dieser Abschnitt den Verteilungseigenschaften des Scores, wenn am Anfang keine Anordnung der Sequenzen gegeben ist und diese zuerst durch Optimalität bestimmt werden muss. Die Lemmata von Kingman und Azuma-Hoeffding liefern interessante Resultate zu der Verteilung des Scores, jedoch konnte die Verteilung von S bisher noch nicht vollkommen analysiert werden.

Theorem 4.1. *(Kingman)*
Seien s und t nichtnegative ganze Zahlen mit $0 \leq s \leq t$. $X_{s,t}$ sei eine Sammlung von Zufallsvariablen, die folgende Bedingungen erfüllt
(a) Wenn $s < t < u$, dann gilt $X_{s,u} \leq X_{s,t} + X_{t,u}$, d. h. $X_{s,u}$ ist subadditiv.
(b) Die gemeinsame Verteilung von $\{X_{s,t}\}$ ist die gleiche wie die von $\{X_{s+1,t+1}\}$.
(c) Der Erwartungswert $h_t = I\!E[X_{0,t}]$ existiert und es gilt $h_t \geq -Kt$ für eine Konstante K und alle $t > 1$.

Dann existiert der endliche Grenzwert $\lim_{t\to\infty} \frac{X_{0,t}}{t} = p$ fast sicher und konvergiert im Mittel, also liegt L_1 - Konvergenz vor.

4.1 Lineares Wachstum des Alignment Scores

In diesem Abschnitt soll gezeigt werdem, dass der Alignment Score S linear wächst, wenn dieser eine Form wie in (4.1) aufweist. Dabei wirken sich Gaps negativ auf den Score aus, da Deletions der Länge k mit $-g(k)$ bewertet werden. Die Funktion g ist nichtnegativ und subadditiv. Dies führt dazu, dass Gaps der Länge $s+t$ sich weniger oder mindestens genauso negativ auf den Score auswirken wie zwei Gaps der Länge s und t.

Definition 4.1.1. *Der Score von (t-s) Zufallsvariablen ist definiert als*

$$-X_{s,t} = \textit{Score von}\, A_{s+1} \ldots A_t \,\textit{und}\, B_{s+1} \ldots B_t.$$

Dabei ist $-X_{s,t}$ das Maximum einer endlichen Anzahl von Alignment Scores.

Lemma 4.1.2. *Sei $-X_{0,t}$ = Score von $A_1 ... A_t$ und $B_1 ... B_t$ gegeben. Dann erfüllt $-X_{s,t}$ die Voraussetzungen des Theorems von Kingman (vgl. Theorem 4.1) und somit gilt*

$$\lim_{t\to\infty} \frac{-X_{0,t}}{t} = p \qquad \text{f.s. und in } L_1. \tag{4.2}$$

Beweis. (a) Wegen der Optimalitätseigenschaft des Scores und der Definition von $X_{s,t}$ gilt

$$-X_{s,u} \geq (-X_{s,t}) + (-X_{t,u}), \text{ also}$$
$$X_{s,u} \leq X_{s,t} + X_{t,u}. \quad \text{Also ist } X_{s,u} \text{ subadditiv.}$$

(b) $\{X_{s,t}\}$ und $\{X_{s+1,t+1}\}$ haben die gleiche gemeinsame Verteilung, da beide von $(t-s)$ Zufallsvariablen abhängen.
(c) $h_t = \mathbb{E}(X_{0,t})$ existiert, da der Erwartungswert einer einzelnen Anordnung existiert und $-X_{0,t}$ als das Maximum einer beschränkten Anzahl von Alignment Scores definiert ist. Also bleibt zu zeigen, dass $h_t \geq -Kt$ für eine Konstante K und alle $t > 1$.
Setze $s^* = max\{s(a,b)\ a,\ b \in \mathcal{A}\}$. Dann gilt $\mathbb{E}(-X_{0,t}) \leq max\{ts^*, -2g(t)\} = t\, max\{s^*, \frac{-2g(t)}{t}\}$.
Der maximale Score bei einer kompletten Übereinstimmung von t Zufallsvariablen wäre ts^*.
Wenn aber die Sequenzen so angeordnet werden, dass überhaupt keine Übereinstimmung vorhanden ist, dann existieren $2t$ Gaps, die mit $-2g(t)$ in den Score einfließen.
Wenn $s^* < -2g(t)/t$, dann existiert $\lim_{t\to\infty} g(t)/t$ wegen der Subadditivität von g, da man für größer werdendes t die Funktion durch die Summe von kleineren Funktionswerten abschätzen kann, also $g(t+t) \leq g(t) + g(t)$.
Als Maximum der Funktion s existiert s^*. In dem Fall $s^* > -2g(t)/t$ ist $\mathbb{E}(-X_{0,t})$ also ebenso beschränkt. Damit existiert eine Konstante K, sodass $h_t \geq -Kt$. Mit dem Theorem von Kingman folgt dann die Behauptung. □

Damit ist auch schon das folgende Theorem bewiesen, welches aussagt, dass der optimale Alignment Score linear mit der Länge der Sequenzen wächst.

Theorem 4.1.3. *Seien zwei Sequenzen $\boldsymbol{A} = A_1 ... A_n$ und $\boldsymbol{B} = B_1 ... B_n$ gegeben, wobei A_i und B_j i.i.d sind. Sei $S_n = S(\boldsymbol{A}, \boldsymbol{B}) = max\{\sum s(A_i^*, B_i^*) : \text{alle Alignments}\}$. Dann existiert eine Konstante $p \geq \mathbb{E}(s(A,B))$, sodass $\lim_{n\to\infty} \frac{S_n}{n} = p$ fast sicher und in L_1 konvergiert.*
Da $\lim_{n\to\infty} \frac{S_n}{n} = p$ fast sicher, konvergiert auch $\frac{\mathbb{E}(S_n)}{n} \to p$.

Hierbei ist aber zu beachten, dass selbst in dem einfachsten Fall, in dem der Score binomialverteilt ist, die Konstante p bis jetzt nicht bestimmt werden konnte. Über die Varianz des Scores S_n ist ebenso nicht viel bekannt. Es existieren aber Grenzen für p: $0.7615 \leq p \leq 0.8575$ und ohne Anordnung hätte man $\mathbb{E}(s(A,B)) = 0.5$ (vgl. dazu Theorem 2.1). Ohne Anordnung hat man also einen kleineren Wert für die Konstante p als bei der Anordnung, wenn ein binomialverteilter Score zugrunde gelegt wird.

4.2 Die Azuma-Hoeffding Ungleichung

Definition 4.2.1. *(Martingal)*
Sei $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq ...$ eine Folge von Sub-σ-Algebren und X_n ein $\mathcal{F}_n$- messbarer stochastischer Prozess. $\{X_n\}_{n\geq 0}$ ist ein Martingal, falls
(a) $\mathbb{E}|X_n| < \infty$ für alle n, und
(b) $\mathbb{E}(X_n|\mathcal{F}_{n-1}) = X_{n-1}, n \geq 1$.

Beispiel 4.2.2. *Der Random-Walk $S_n = \sum_{i=1}^{n} X_i$, $\mathcal{F}_n = \sigma(S_1, ..., S_n)$ mit $\{X_i\}_{i\in\mathbb{N}}$ i.i.d, $\mathbb{E}|X_i| < \infty$ und $E(X_1) = 0$ ist ein Martingal, da*

$$\mathbb{E}(S_{n+1} \mid \mathcal{F}_n) = \mathbb{E}(S_n + X_{n+1} \mid \mathcal{F}_n) = Sn + \mathbb{E}(X_{n+1} \mid \mathcal{F}_n) = S_n + \mathbb{E}(X_{n+1}) = S_n.$$

Dies gilt wegen der $\mathcal{F}_n$-Messbarkeit von S_n, der Unabhängigkeit von X_{n+1} von $\mathcal{F}_n$ und da $E(X_1) = 0$ für die i.i.d Zufallsvariablen gilt.

Lemma 4.2.3. *Sei $\{X_n\}_{n\geq 0}$ mit $X_0 = 0$ ein Martingal bzgl. der Filtration $\{F_n\}$, sodass $X_{n-1} = \mathbb{E}(Y|\mathcal{F}_{n-1}), n \geq 1$. Wenn für eine Sequenz von positiven Konstanen c_n,*

$$|X_n - X_{n-1}| \leq c_n \,, \text{für } n \geq 1,$$
$$\text{dann} \qquad \mathbb{E}(e^{\beta X_n}) \leq e^{\beta^2/2 \sum_{k=1}^{n} c_k^2}. \tag{4.3}$$

Beweis. Bei diesem Beweis werden zwei Ungleichungen benötigt. Die erste Ungleichung ist

$$e^{\beta x} \leq \frac{c-x}{2c} e^{-\beta c} + \frac{c+x}{2c} e^{\beta c} \text{ für alle } x \in [-c, c]\,.$$

Dies folgt aus der Konvexität der Funktion $\varphi(t) = e^{\beta t}$. Dabei nutzt man $\varphi(\gamma x_1 + (1-\gamma)x_2) \leq \gamma\varphi(x_1) + (1-\gamma)\varphi(x_2)$, $\gamma \in [0, 1]$. Dabei sind $x_1 = -c$, $x_2 = c$,

$$\gamma = \frac{c-x}{2c},$$
$$\text{und} \qquad 1-\gamma = \frac{c+x}{2c}.$$

Die zweite Ungleichung

$$\frac{e^{-x} + e^{x}}{2} = cosh\, x \leq e^{x^2/2} \quad \text{für alle } x,$$

folgt durch Anwenden der Taylor-Formel und anschließendes Vergleichen beider Seiten. Es

gilt nun

$$\begin{aligned}\mathbb{E}(e^{\beta X_n}) &= \mathbb{E}(\mathbb{E}(e^{\beta X_n} \mid \mathcal{F}_{n-1})) \\ &= \mathbb{E}(e^{\beta X_{n-1}}\mathbb{E}(e^{\beta(X_n - X_{n-1})} \mid \mathcal{F}_{n-1}).\end{aligned}$$

Da $|X_n - X_{n-1}| \leq c_n$ kann man die erste Ungleichung auf $e^{\beta(X_n - X_{n-1})}$ anwenden. Des Weiteren nutzt man die Linearität des Erwartungswerts, die Martingaleigenschaft und die Messbarkeit von X_{n-1} bzgl. der Filtration $\mathcal{F}_{n-1}$. Damit gilt $\mathbb{E}(e^{\beta(X_n - X_{n-1})} \mid \mathcal{F}_{n-1}) \leq cosh\,\beta c_n$. Durch iteratives Anwenden der eben gezeigten Abschätzung und der Martingaleigenschaft auf $\mathbb{E}(e^{\beta X_n})$ folgt die Behauptung. □

Lemma 4.2.4. *Unter den gleichen Voraussetzungen wie in Lemma 4.2.3 und $\lambda > 0$ gilt*

$$I\!P(X_n \geq \lambda) \leq e^{-\lambda^2/(2\sum_{k=1}^{n} c_k^2)}. \tag{4.4}$$

Beweis. Benutze für die Ungleichung von Markov die monoton wachsende Funktion $g : \mathbb{R} \to [0, \infty)$ mit $g(t) = e^{\beta t}$, $\beta > 0$. Dann gilt

$$\mathbb{P}(X_n \geq \lambda) \leq \frac{\mathbb{E}(e^{\beta X_n})}{e^{\beta\lambda}}.$$

Durch Anwenden von Lemma 4.2.3 und minimieren nach β erhält man die Ungleichung. □

Die Azuma-Hoeffding Ungleichung gibt eine Beschränkung für große Abweichungen bzw. die Wahrscheinlichkeit, dass eine Zufallsvariable seinen Mittelwert um einen bestimmten Wert überschreitet.

4.3 Große Abweichungen vom Mittelwert

Lemma 4.3.1. *Sei $S = S(a_1...a_k, b_1...b_k) = S(c_1...c_k)$ der Alignment Score für k Paare von Buchstaben, $c_i = (a_i, b_i)$. Und sei $S' = S(c_1...c_{i-1}, c_i', c_{i+1}...c_k)$ der Score für k Paare von Buchstaben, wobei nur das i-te Paar verändert wird. Sei $s^* = max\{s(a,b) : a, b \in \mathcal{A}\}$ und $s_* = min\{s(a,b) : a, b \in \mathcal{A}\}$. Dann gilt*

$$S - S' \leq max\{min\{2s^* + 4g(1), 2s^* - 2s_*\}, 0\} = c \tag{4.5}$$

Beweis. Betrachte das i-te Paar $c_i = (a_i, b_i)$ in S und $c_i' = (a_i', b_i')$ in S'. Um die maximale Differenz $S - S'$ bei einer Veränderung in dem i-ten Paar zu beschränken muss man eine Fallunterscheidung machen.
1. Fall: a_i stimmt mit b_j und b_i stimmt mit a_l überein.
Dabei wird ein maximaler Score von $2s^*$ in dem Score S erzielt. Wenn a_i in a_i' und b_i in b_i' umgeändert wird, wird mindestens ein Score von $2s_*$ erzielt oder alle vier Buchstaben

können gelöscht werden. Wegen der Subadditivität von g ist die zusätzliche deletion penalty höchstens gleich $4g(1)$. In diesem Fall kann die Differenz von S und $S^{'}$ beschränkt werden durch $S - S^{'} \leq min\{2s^* + 4g(1),\ 2s^* - 2s_*\}$. Für den Fall, dass a_i oder b_i gelöscht werden, gilt auch die obige Abschätzung.
2. Fall: a_i stimmt mit b_i überein.
Eine hohe Scoring-Übereinstimmung wird durch eine niedrige Scoring-Übereinstimmung ersetzt oder beide Buchstaben können gelöscht werden. Also $S - S^{'} \leq min\{s^* + 2g(1),\ s^* - s_*\}$. Wenn dabei $s^* + 2g(1) \leq 0$, dann wäre $S^{'} \geq S$, was der Optimalität des Alignment Scores S widersprechen würde. In diesem Fall ist $S_k = 2g(k)$ unabhängig von den Sequenzen und es gilt $S - S^{'} = 0$. Damit gilt die Behauptung. □

Theorem 4.3.2. *Gegeben seien zwei Sequenzen $\boldsymbol{A} = A_1 ... A_n$ und $\boldsymbol{B} = B_1 ... B_n$, wobei A_i und B_j i.i.d sind. Sei $S = S(\boldsymbol{A}, \boldsymbol{B})$ der globale Alignment Score. Wenn c die Konstante aus Lemma 4.3.1 ist, dann*

$$I\!P(S - I\!E(S) \geq \gamma n) \leq e^{-\gamma^2 n/2c^2}. \tag{4.6}$$

Beweis. Wir wollen hier Lemma 4.2.4 anwenden. Dafür benutzen wir, dass $X_i = \mathbb{E}(Y \mid \mathcal{F}_i)$ mit $Y = S(C_1...C_n) - \mathbb{E}(S(C_1...C_n))$ ein Martingal ist, wobei $\mathcal{F}_i = \sigma(C_1...C_i)$ die von den ersten i Paaren von Zufallsvariablen $C_i = (A_i, B_i)$ erzeugte σ- Algebra ist. Dies gilt, da

$$\mathbb{E}(X_n \mid \mathcal{F}_{n-1}) = \mathbb{E}(\mathbb{E}(Y \mid \mathcal{F}_n) \mid \mathcal{F}_{n-1}) = \mathbb{E}(Y \mid \mathcal{F}_{n-1}) = X_{n-1}.$$

Da Y $\mathcal{F}_n$-messbar ist, gilt $X_n = \mathbb{E}(Y \mid \mathcal{F}_n) = Y = S - \mathbb{E}(S)$ mit den gegebenen Definitionen und der Martingaleigenschaft.
Der entscheidende Punkt des Martingals ist

$$\mathbb{E}(S \mid \mathcal{F}_i) = \sum_{c_{i+1},...,c_n} S(C_1, ..., C_i, c_{i+1}, ..., c_n)\mathbb{P}(C_{i+1} = c_{i+1}, ...C_n = c_n).$$

Dies gilt, da man die Werte von $C_{i+1}, ..., C_n$ mitteln muss, um $S(C_1...C_n)$ auf $\mathcal{F}_i = \sigma(C_1, ..., C_i)$ messbar zu machen. Im Anschluss betrachtet man die Differenz $\mid X_i - X_{i-1} \mid$ der zuvor definierten X_i und bekommt mit Lemma 4.3.1 eine obere Grenze für diese Differenz. Dann kann man Lemma 4.2.4 anwenden und bekommt damit die Behauptung. □

Korollar 4.3.3. *Unter den Voraussetzungen von Theorem 4.3.2 gilt*

$$I\!P(S_n/n - p \geq \gamma) \leq e^{-\gamma^2 n/2c^2}. \tag{4.7}$$

Beweis. Wegen der Subadditivität des Scores folgt mit Theorem 4.1.3 $p = \lim_{n\to\infty} \frac{\mathbb{E}(S_n}{n} =$

$sup_n \frac{\mathbb{E}(S_n}{n}$. Also hat man mit $\mathbb{E}(S_n) \leq np$

$$\mathbb{P}(S_n \geq (\gamma + p)n) \leq \mathbb{P}(S_n - \mathbb{E}(S_n) \geq \gamma n). \tag{4.8}$$

Durch Anwenden von Theorem 4.3.2 erhält man die Gleichung. □

Korollar 4.3.3 sagt dabei aus, dass eine exponentielle Konvergenzgeschwindigkeit der gegebenen Wahrscheinlichkeit vorliegt.

Theorem 4.3.4. *(Steele)*
Sei $f(x_1, ..., x_n)$ eine Funktion und X_i, X_i', $1 \leq i \leq n$, seien $2n$ Zufallsvariablen, dann

$$\mathrm{Var}(f) \leq \frac{1}{2}\{\mathbb{E}\sum_{i=1}^{n}(f - f_{(i)})^2\}, \tag{4.9}$$

wobei $f = f(X_1, ..., X_n)$ und $f_{(i)} = f(X_1, ..., X_i', ..., X_n)$ durch Ersetzen von X_i durch X_i' entsteht.

Theorem 4.3.5. *Gegeben seien zwei Sequenzen $\boldsymbol{A} = A_1...A_n$ und $\boldsymbol{B} = B_1...B_n$, wobei A_i und B_j i.i.d sind. Sei $S_n = S(\boldsymbol{A}, \boldsymbol{B})$ der globale Alignment Score. Wenn $c^* = max\{0, min\{s^* + 2g(1), s^* - s_*\}\}$ und $p = \mathcal{P}(A_1 = B_1)$, dann gilt*

$$\mathrm{Var}(S_n) \leq n(1-p)c^{*2}. \tag{4.10}$$

Beweis. Es wird nur eine der $2n$ Zufallsvariablen zur gleichen Zeit geändert. Damit gilt $|S - S_{(i)}| \leq max\{0, min\{s^* + 2g(1), s^* - s_*\}\} = c^*$, wobei $c^* = \frac{1}{2}c$ aus Lemma 4.3.1. Mit Wahrscheinlichkeit $\mathbb{P}(A_1 = B_1)$ hat sich i-te Zufallsvariable nicht verändert, also gilt hier $S - S_{(i)} = 0$. Zudem

$$(S - S_{(i)})^2 = \begin{cases} 0 & , \text{mit } \mathbb{P}(A_1 = B_1), \\ (S - S_{(i)})^2 & , \text{mit } (1 - \mathbb{P}(A_1 = B_1)) \end{cases}.$$

Also $\mathbb{E}(S - S_{(i)})^2 = 0 \cdot \mathbb{P}(A_1 = B_1) + (S - S_{(i)})^2 \cdot (1 - \mathbb{P}(A_1 = B_1)) \leq c^{*2} \cdot (1 - \mathbb{P}(A_1 = B_1))$. Da es $2n$ Terme gibt gilt mit dem Theorem von Steele (vgl. 4.3.5)

$$\begin{aligned} Var(S_n) &\leq \frac{1}{2}\{\mathbb{E}\sum_{j=1}^{2n}(S_n - S_{n(j)})^2\} \\ &= \frac{1}{2}\{\sum_{j=1}^{2n}\mathbb{E}(S_n - S_{n(j)})^2\} \\ &\leq \frac{1}{2}\{\sum_{j=1}^{2n} c^{*2} \cdot (1-p)\} \\ &= \frac{1}{2} \cdot 2n(1-p)c^{*2} \end{aligned}$$

□

Damit haben wir eine Beschränkung von $Var(S_n)$, die linear von n abhängt.