Westfälische Wilhelms Universität Münster

Fachbereich 10 - Mathematik und Informatik

Signifikanz von Alignment Scores und BLAST

Seminarvortrag von Leonie Zeune

10. Mai 2012

Veranstaltung: Seminar zur mathematischen Biologie

Betreuer: Prof. Löwe

Inhaltsverzeichnis

1 Einleitung

In meinem Vortrag beschäftige ich mich mit der Fragestellung, wie signikant ein gefundenes optimalee Alignment zweier DNA-Sequenzen ist.
Angenommen, wir haben ein optimales Alignment zweier Sequenzen x und y gefunden mit einem Score von s_0.
Frage: Ist dieser Zusammenhang biologisch bedeutsam oder einfach nur der beste Zusammenhang zwischen zwei unabhängigen Sequenzen?
Um diese Frage zu beantworten, wollen wir die Signifikanz berechnen, d.h. wenn ich zwei beliebig gewählte Sequenzen betrachte und ihren optimalen Score berechne, wie groß ist dann die Wahrscheinlichkeit, dass dieser größer oder gleich s_0 ist? Hierbei legen wir fest:

$$\text{Falls } \begin{Bmatrix} \mathbb{P}(s(x,y)) \geq s_0) \leq 0.05 \\ \text{sonst} \end{Bmatrix} \text{ gilt, dann ist } s_0 \begin{Bmatrix} \text{signifikant} \\ \text{nicht signifikant} \end{Bmatrix}.$$

Vereinfacht werden wir dies nur für den Fall von zwei lückenlosen lokalen Alignments betrachten.
Sei x^0 eine Sequenz der Länge N_1 und y^0 eine Sequenz der Länge N_2, für die wir das optimale Alignment bereits gefunden haben mit Score s_0. Seien $a, b \in \mathcal{Q}$ mit Score $s(a,b)$, wobei $\mathcal{Q}$ das DNA-Alphabet ist. Es ist $s(a,b) \in \mathbb{Z}$ und somit ist s_0 eine große Ganzzahl.
Mit p_a benennen wir die Häufigkeit von $a \in \mathcal{Q}$ in x^0 und analog mit p'_b die Häufigkeit von $b \in \mathcal{Q}$ in y^0 für alle $a, b \in \mathcal{Q}$. Sei

$$S = \{(x,y) \mid x, y \text{ Sequenzen der Länge } N_1 \text{ bzw. } N_2 \text{ mit } x_i, y_j \in \mathcal{Q}\}$$

und $x = x_1 ... x_{N_1}$ sowie $y = y_1 ... y_{N_2}$. Dann definieren wir ein Wahrscheinlichkeitsmaß auf S durch

$$\mathbb{P}\left(\{(x,y)\}\right) = \left(\prod_{i=1}^{N_1} p_{x_i}\right) \times \left(\prod_{j=1}^{N_2} p'_{y_j}\right)$$

Da S endlich ist und $\mathbb{P}$ für jedes Elementarereignis in S definiert ist, können wir das Wahrscheinlichkeitsmaß $\mathbb{P}$ auf die σ-Algebra aller Ereignisse erweitern und erhalten so einen Wahrscheinlichkeitsraum $(S, \mathcal{B}, P)$.
Auf $(S, \mathcal{B}, \mathbb{P})$ können wir nun wie folgt eine Zufallsvariable definieren:
Für $(x,y) \in S$ sei

$$s((x,y)) := \text{Score jedes optimalen Alignments zwischen x und y.}$$

Gesucht ist somit

$$\begin{aligned} \mathbb{P}\left(\{(x,y) \in S : s((x,y)) \geq s_0\}\right) &= \mathbb{P}\left(\{(x,y) \in S : s((x,y)) > s_0 - 1\}\right) \\ &= 1 - F_s(s_0 - 1) = F_s^*(s_0 - 1), \end{aligned}$$

wobei F_s die Verteilungsfunktion der Zufallsvariablen s bezeichnet.
Wie bereits zuvor festgelegt, sagen wir, dass die Ähnlichkeit zwischen zwei Sequenzen x und y signifikant ist, wenn $F_s^*(s_0 - 1) \leq 0.05$ ist. Unser Ziel ist es daher, die Verteilungsfunktion $F_s(\alpha)$ zu bestimmen, wobei α eine große Ganzzahl ist.

Um eine grundsätzliche Idee davon zu bekommen, wie diese Verteilungsfunktion aussehen könnte, stellen wir zunächst ein paar stark vereinfachte Vorbetrachtungen an:
Sei $(x, y) \in S$ mit $x = x_1...x_{N_1}$ und $y = y_1...y_{N_2}$. Wir betrachten nun zwei Teilsequenzen

$$A_{i,j,N}(x,y) := \left\{ \begin{matrix} x_i...x_{i+N} \text{ mit } 1 \leq i \leq N_1 \\ y_i...y_{i+N} \text{ mit } 1 \leq i \leq N_2 \end{matrix} \right\} \text{ mit } N \leq \min\{N_1 - i, N_2 - j\}.$$

Der Score dieses lokalen Alignments sei definiert als

$$s(A_{i,j,N}(x,y)) = s(x_i, y_j) + s(x_{i+1}, y_{j+1}) + ... + s(x_{i+N}, y_{j+N}),$$

wobei die Zufallsvariablen $s(x_{i+k}, y_{j+k})$ i.i.d. auf $(S, \mathcal{B}, \mathbb{P})$ sind. Dann folgt mit dem zentralen Grenzwertsatz, dass für großes N $s(A_{i,j,N})$ annähernd normalverteilt ist. Es gilt

$$s((x,y)) = \max_{i,j,N} s(A_{i,j,N}(x,y)),$$

wobei wir hier vereinfacht annehmen, dass die $s(A_{i,j,N}(x,y))$ auch i.i.d. sind. Somit haben wir m i.i.d. normalverteilte Zufallsvariablen $f_1, ..., f_m$ auf einem Wahrscheinlichkeitsraum $(S', \mathcal{B}', \mathbb{P}')$ und suchen die Verteilung von $f_{max} = \max\{f_1, ..., f_m\}$. Es gilt:

$$\begin{aligned} F_{f_{max}}(\alpha) &= \mathbb{P}\left(\{e \in S' : f_{max}(e) \leq \alpha\}\right) \\ &= \mathbb{P}\left(\{e \in S' : f_1(e) \leq \alpha, f_2(e) \leq \alpha, ..., f_m(e) \leq \alpha\}\right) \\ &= \prod_{k=1}^{m} \mathbb{P}\left(\{e \in S' : f_k(e) \leq \alpha\}\right) \text{ da die ZV i.i.d. sind} \\ &= (F(\alpha))^m, \end{aligned}$$

wobei F die Verteilungsfunktion einer normalverteilten Zufallsvariable ist. Sind m und α groß, so lässt sich zeigen, dass

$$F_{f_{max}}(\alpha) \approx \exp(-Km \cdot \exp(-\lambda\alpha))$$

ist mit Konstanten $K > 0$ und $\lambda > 0$. Somit können wir bereits ahnen, dass unser Ergebnis in etwa diese Form haben sollte. Um dorthin zu gelangen, brauchen wir aber zunächst einige Ergebnisse aus der Theorie der Random Walks.

2 Random Walks

Im folgenden Abschnitt werden wir einige Ergebnisse aus der Theorie der Random Walks betrachten, auf Beweise und die genaue Herleitung werden wir jedoch verzichten.

Definition 1 (Random Walk) *Ein Random Walk ist ein diskreter Zeitprozess, der in 0 startet und sich mit endlichen, festgelegten Schrittweiten und Wahrscheinlichkeiten auf und ab bewegen kann, wobei jede Bewegung unabhängig von allen vorangegangenen Bewegungen ist.*

Sei $T = \{-c, -c+1, ..., 0, 1, ..., d-1, d\}$ die Menge aller möglichen Schrittweiten mit $c, d \in \mathbb{N}$ und den zugehörigen Wahrscheinlichkeiten $p_{-c}, ..., p_0, ..., p_d$ wobei $\sum_{j=-c}^{d} p_{-j} = 1$ ist.
Auf dem Wahrscheinlichkeitsraum $(T, \mathcal{B}_T, \mathbb{P}_T)$ mit $\mathbb{P}_T(\{j\}) = p_j \ \forall j \in T$ definieren wir eine Zufallsvariable der Schrittweite als Identität von T, die mit einem Random Walk indentifiziert werden kann. In ähnlicher Weise werden wir letzendlich die Substitutionsmatrix mit einem Random Walk identifizieren und können dann die nachfolgende Theorie dort anwenden.
Zunächst brauchen wir noch drei Bedingungen:

1. $p_{-c} > 0$, $p_d > 0$.
2. Das arithmetische Mittel der Schrittweite ist negativ, d.h. $\sum_{j=-c}^{d} j \cdot p_j < 0$.
3. Der ggT aller positiven $j \in T$ mit $p_j > 0$ ist 1.

Sei S_ω die Menge aller möglichen Trajektorien eines Random Walks, d.h. in unserem Fall die Menge aller unendlichen Sequenzen $t = t_1, t_2, ..$ mit $t_j \in T$.
Als Leiterpunkt bezeichnen wir alle Punkte einer Trajektorie, die tiefer sind als alle zuvor erreichten Punkte. Den Teil einer Trajektorie von einem Leiterpunkt bis zum höchsten Punkt, der vor dem nächsten Leiterpunkt erreicht wird, bezeichnen wir als Exkursion.

Für $t \in S_\omega$ definieren wir

$$Y(t) = \begin{cases} \text{maximaler Wert, den } t \text{ erreicht, bevor sie -1 erreicht,} & \text{falls } t \in E_0^c \\ \text{beliebig,} & \text{falls } t \in E_0. \end{cases}$$

Ziel des Abschnitts ist es, das Verhalten von $F_Y(\alpha)$ für großes α zu bestimmen. Genauer gesagt, wollen wir zeigen, dass Y eine geometrisch-ähnliche Verteilung hat, d.h.

Definition 2 (geometrisch-ähnlich) *Eine diskrete Zufallsvariable Y hat eine geometrisch-ähnliche Verteilung, wenn für $y \to \infty$ gilt:*

$$\mathbb{P}(Y \geq y) \approx C \cdot e^{-\lambda y}$$

mit $C < 1$.

Sei nun α eine positive Ganzzahl. Setze

$$E_{-j} = \{t \in S_\omega : \text{ für ein } n \in \mathbb{N} \text{ gilt } \sum_{i=1}^{m} t_i \geq 0 \; \forall m < n \text{ und } \sum_{i=1}^{n} t_i = -j\}$$

und $R_{-j} = \mathbb{P}(E_{-j}) =$ Wahrscheinlichkeit, dass die erste negative Zahl, die eine Trajektorie erreicht $-j$ ist.

Für die Herleitung der Verteilung von Y verwendet man desweiteren folgende Definitionen und Theoreme:

Definition 3 (momenterzeugende Funktion) *Sei f eine diskrete Zufallsvariable die endlich viele Werte $r_1, ..., r_n$ mit $r_i \neq r_j$ für $i \neq j$ annimmt, und sei $\{p_f(r_j),\ j = 1, ...,\}$ die Wahrscheinlichkeitsverteilung von f. Dann heißt*

$$M_f(\theta) = \mathbb{E}(\exp(\theta f)) = \sum_{j=1}^{n} \exp(\theta r_j)\, p_f(r_j),\ \theta \in \mathbb{R}$$

die momenterzeugende Funktion von f.

Theorem 1 *Sei f eine diskrete Zufallsvariable, die endlich viele Werte annimmt, sodass $\mathbb{E}(f) \neq 0$. Angenommen f nimmt einen positiven Wert a und einen negativen Wert b mit jeweils positiven Wahrscheinlichkeiten $p_f(a)$ und $p_f(b)$ an. Dann existiert genau ein $\theta^* \in \mathbb{R} \setminus \{0\}$, sodass*

$$M_f(\theta^*) = 1. \tag{1}$$

Hierbei gilt: Ist $\mathbb{E}(f) = M_f'(0) < 0$, so ist $\theta^* > 0$ und umgekehrt.

Ausserdem definieren wir analog zu E_{-j} und R_{-j}

$$E_k = \{t \in S_\omega : \text{ für ein } n \in \mathbb{N} \text{ gilt } \sum_{i=1}^{m} t_i \leq 0 \; \forall m < n \text{ und } \sum_{i=^1}^{n} t_i = k\}$$

und $Q_k = \mathbb{P}(E_k) =$ Wahrscheinlichkeit, dass die erste positive Zahl, die eine Trajektorie erreicht, k ist. Dann gilt $Q_k = 0\ \forall k > d$ und $Q_0 := 0$. Ausserdem gilt, dass $\sum_{k=1}^{d} Q_k < 1$ ist, daher definieren wir $\bar{Q} = 1 - \sum_{k=1}^{d} Q_k$ als die Wahrscheinlichkeit, dass eine Trajektorie nie einen positiven Wert annimmt.

Theorem 2 (Erneuerungstheorem) *Angenommen die drei Sequenzen $\{a_0, a_1, ...\}$, $\{b_0, b_1, ...\}$ und $\{c_0, c_1, ...\}$ aus nichtnegativen Zahlen erfüllen die Gleichung*

$$c_j = a_j + (c_j b_0 + c_{j-1} b_1 + ... + c_1 b_{j-1} + c_0 b_j) \quad \forall j \geq 0 \tag{2}$$

und $\{c_j\}$ sei beschränkt, $\sum_{j=0}^{\infty} b_j = 1$ und $\sum_{j=0}^{\infty} a_j$ sowie $\sum_{j=0}^{\infty} j \cdot b_j$ konvergieren. Sei $A = \sum_{j=0}^{\infty} a_j$ und $\mu = \sum_{j=0}^{\infty} b_j$ und ausserdem gelte,

dass der ggT aller Ganzzahlen j mit $b_j > 0$ 1 ist.
Dann existiert der Grenzwert $\lim_{j\to\infty} c_j$ *und es gilt*

$$\lim_{j\to\infty} c_j = \frac{A}{\mu}.$$

Theorem 3 (Wald's Identität) *Sei N eine Zufallsvariable auf $(S_\omega, \mathcal{B}_\omega, \mathbb{P}_\omega)$, deren Wert $N(t)$ für alle $t \in E_0^c$ gleich der Anzahl der Schritte ist, die die Trajektorie t braucht, um entweder einen positiven Wert oder $-L$ das erste Mal zu erreichen. Sei außerdem T_N die Zufallsvariable auf $(S_\omega, \mathcal{B}_\omega, \mathbb{P}_\omega)$, deren Wert für alle $t \in E_0^c$ gleich dem Wert ist, den t nach $N(t)$ Schritten annimmt. Dann gilt für alle $\theta \in \mathbb{R}$*

$$\mathbb{E}\left(M(\theta)^{-N} \cdot \exp(\theta T_N)\right) = 1.$$

Wichtig hierbei ist, dass die Trajektorie nicht genau $-L$ erreichen muss, sie kann auch $-L$ überschreiten. Mit $\theta = \theta^*$ erhalten wir $\mathbb{E}(\exp(\theta^* T_N)) = 1$.

Letzendlich erhalten wir folgende Verteilungsfunktion der Zufallsvariablen Y:

$$F_Y^*(\alpha) \approx C \cdot \exp(-\theta^*(\alpha + 1)),$$

wobei

$$C = \frac{\bar{Q}\left(1 - \sum_{j=1}^{c} R_{-j} \exp(-\theta^* j)\right)}{(1 - \exp(-\theta^*)) \sum_{k=1}^{d} k \cdot Q_k \exp(\theta^* k)} \tag{3}$$

ist. Somit gilt, dass Y geometrisch-ähnlich verteilt ist, was wir in diesem Abschnitt zeigen wollten.

3 Signifikanz eines Alignment Scores

In diesem Abschnitt wollen wir nun die allgemeine Theorie der Random Walks aus Abschnitt 2 auf unseres Problem anwenden, indem wir die Substitutionsmatrix mit einem Random Walk identifizieren. Betrachte dazu die Substitutionsmatrix $(s(a,b))$ mit $a,b \in \mathcal{Q}$ und die Wahrscheinlichkeiten $\{p_a\}$ und $\{p'_b\}$, die wir zuvor in Abschnitt 1 definiert haben. Wir definieren

$$T = \{x \in \mathbb{Z} \mid x \in [\tilde{s}, \hat{s}]\},$$

wobei $\tilde{s}$ der minimale Eintrag und $\hat{s}$ der maximale Eintrag in $(s(a,b))$ ist. Für $j \in (s(a,b))$ definiere

$$p_j = \sum_{(a,b):s(a,b)=j} p_a \cdot p'_b \quad \text{beziehungsweise} \quad p_j = 0,$$

falls keine $a,b \in \mathcal{Q}$ existieren mit $s(a,b) = j$. Hierbei nehmen wir an, dass

1. $\hat{s} > 0$, $p_{\hat{s}} > 0$ und $p_{\tilde{s}} > 0$

2. $\sum_{a,b\in\mathcal{Q}} s(a,b) p_a p'_b < 0$

3. der ggT aller positiven Einträge in $(s(a,b))$ ist 1.

Um eine Idee davon zu bekommen, wie genau die Substitutionsmatrix und der Random Walk zusammenhängen betrachten wir folgendes Beispiel:

Beispiel 1 *Gegeben seien zwei Sequenzen*

$$\begin{aligned} x:&\ AGCTAGCAATGGCT... \\ y:&\ AGATCGATCAGTAC... \end{aligned}$$

,

sowie eine Scoring-Funktion

$$s((a,b)) = \begin{cases} 1 & a = b \\ -1 & a \neq b \end{cases}$$

Dann gilt

x:	*A*	*G*	*C*	*T*	*A*	*G*	*C*	*A*	*A*	*T*	*G*	*G*	*C*	*T*	*...*
y:	*A*	*G*	*A*	*T*	*C*	*G*	*A*	*T*	*C*	*A*	*G*	*T*	*A*	*C*	*...*
Score:	*1*	*1*	*-1*	*1*	*-1*	*1*	*-1*	*-1*	*-1*	*-1*	*1*	*-1*	*-1*	*-1*	*...*

und die jeweiligen Scoring-Werte entsprechen den Schrittweiten der Trajektorie des Random Walks. Das heißt, die Trajektorie des zugehörigen Random Walks sieht wie folgt aus.

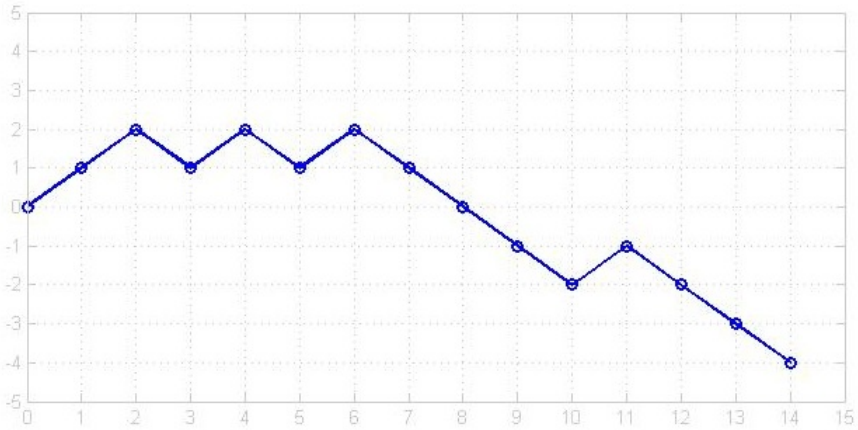

Abbildung 1: Random Walk mit Schrittweiten 1 und -1.

Um nun die Signifikanz eines Scores zu bestimmen, wollen wir das Verhalten von F_s bestimmen, wobei s die in Abschnitt 1 eingeführte Zufallsvariable auf $(S, \mathcal{B}, \mathbb{P})$ ist mit

$$s((x, y)) := \text{Score jedes optimalen Alignments zwischen x und y.}$$

Jedes Alignment zwischen x und y kann in beide Richtungen jeweils soweit wie möglich verschoben werden, dass sich die Sequenzen zumindest in einer Variablen noch überschneiden, und definiert dadurch eine Überdeckung zwischen x und y. Das heißt, für zwei Sequenzen $AGCTAG$ und $AGATCG$ sind

$$\begin{array}{ll} AGCTAG & \quad AGCTAG \\ \quad AGATCG & \qquad AGATCG \end{array}$$

mögliche Überdeckungen der Länge 4 beziehungsweise 2. Insgesamt gibt es für zwei Sequenzen der Länge N_1 bzw. N_2 gerade $N_1 + N_2 - 1$ solcher Überdeckungen. Wir wählen nun für alle $(x, y) \in S$ eine feste Überdeckung $\mathcal{O}$ der Länge $N \leq \min\{N_1, N_2\}$. Betrachte dazu die Zufallsvariable $s_{\mathcal{O}}$ auf $(S, \mathcal{B}, \mathbb{P})$ definiert durch

$$\begin{aligned} s_{\mathcal{O}}(x, y) := \ & \text{Score eines lückenlosen lokalen Alignments mit dem} \\ & \text{höchsten Score unter allen lückenlosen lokalen Align-} \\ & \text{ments zwischen } x \text{ und } y, \text{ die in die Überdeckung } \mathcal{O} \text{ passen.} \end{aligned}$$

Betrachten wir erneut unser Beispiel 1, dann ist

$$\begin{array}{ccc|ccccccccccc|l} & & & & & & & & & & & & & & \\ A & G & C & T & A & G & C & A & A & T & G & G & C & T & ... \\ & & & A & G & A & T & C & G & A & T & C & A & G & T\ A\ C\ ... \end{array}$$

eine Überdeckung der Länge 11. Die Zufallsvariable $s_{\mathcal{O}}$ sucht jetzt den maximalen Score unter alle lokalen Alignments, die in diese Überdeckung passen. Dies kann ein lokales Alignment der Länge 1, wie z.B. $\left|\begin{array}{c} T \\ A \end{array}\right|$ sein, aber auch ein lokales Alignment der Länge 5, wie z.B. $\left|\begin{array}{ccccc} A & G & C & A & A \\ G & A & T & C & G \end{array}\right|$ sein.

Wir wollen nun das Verhalten von $F_{s_{\mathcal{O}}}$ bestimmen. Sei dafür

$$S_N := \{(x, y) \mid x, y \text{ Sequenzen der Länge } N \text{ mit } x_1, ..., x_N, y_1, ..., y_N \in \mathcal{Q}\}$$

und für $(x, y) \in S_N$ definieren wir

$$\mathbb{P}_N\left(\{(x, y)\}\right) = \prod_{i=1}^{N} p_{x_i} p_{y_i'}.$$

Da S_N endlich ist und $\mathbb{P}_N$ für alle Elementarereignisse in S_N definiert ist, können wir $\mathbb{P}_N$ auf $\mathcal{B}_N = \mathcal{B}\left(S_N\right)$ erweitern und erhalten so einen Wahrscheinlichkeitsraum $(S_N, \mathcal{B}_N, \mathbb{P}_N)$.
Betrachten wir nun erneut den Random Walk, der mit der Substitutionsmatrix assoziiert wird. Dieser ist zunächst auf dem Raum $(S_\omega, \mathcal{B}_\omega, \mathbb{P}_\omega)$ definiert, doch wollen wir diesen nun einschränken. Dazu sei $S_{\omega,N}$ die Menge aller möglichen Trajektorien der Länge N. Für $t = t_1, ..., t_N \in S_{\omega,N}$ mit $t_j \in T$ für alle $j = 1, ..., n$ definieren wir das Wahrscheinlichkeitsmaß

$$\mathbb{P}_{\omega,N}\left(\{t\}\right) = \prod_{j=1}^{N} p_{t_j}.$$

Dann gilt wieder, dass $S_{\omega,N}$ endlich ist und $\mathbb{P}_{\omega,N}$ auf allen Elementarereignissen in $S_{\omega,N}$ definiert ist, sodass wir $\mathbb{P}_{\omega,N}$ auf $\mathcal{B}_{\omega,N} = \mathcal{B}\left(S_{\omega,N}\right)$ erweitern können und einen Wahrscheinlichkeitsraum $(S_{\omega,N}, \mathcal{B}_{\omega,N}, \mathbb{P}_{\omega,N})$ erhalten. Für $N \to \infty$ würde

$$(S_{\omega,N}, \mathcal{B}_{\omega,N}, \mathbb{P}_{\omega,N}) \to (S_\omega, \mathcal{B}_\omega \mathbb{P}_\omega)$$

gelten, denn betrachtet man nur die ersten N Elemente einer Trajektorie $t \in E_{t_{i_1}^0,...,t_{i_m}^0}$, wobei $E_{t_{i_1}^0,...,t_{i_m}^0}$ ein Elementarereignis in S_ω ist, dann gilt

$$\mathbb{P}_{\omega,N}\left(E_{t_{i_1}^0,...,t_{i_m}^0,N}\right) \to \mathbb{P}_\omega\left(E_{t_{i_1}^0,...,t_{i_m}^0}\right).$$

Genauer gesagt gilt sogar für $N \geq i_m$

$$\mathbb{P}_{\omega,N}\left(E_{t_{i_1}^0,...,t_{i_m}^0,N}\right) = \mathbb{P}_\omega\left(E_{t_{i_1}^0,...,t_{i_m}^0}\right).$$

Wir definieren die Zufallsvariable $Y_N(t)$ auf $(S_{\omega,N}, \mathcal{B}_{\omega,N}, \mathbb{P}_{\omega,N})$ analog zur Zufallsvariablen $Y(t)$ auf $(S_\omega, \mathcal{B}_\omega \mathbb{P}_\omega)$, das heißt

$Y_N(t) =$ maximaler Wert, den $t = t_1, ..., t_N$ erreicht, bevor sie -1 erreicht.

Das entspricht gerade der Höhe der Exkursion zwischen dem ersten und zweiten Leiterpunkt von t. Besucht t nie einen negativen Wert, so definiere $Y_N(t)$ als den höchsten Wert, den t annimmt. Dann gilt für $N \to \infty$

$$F_{Y_N} \to F_Y.$$

Sind also N und α groß, so gilt

$$F^*_{Y_N}(\alpha) \approx C \cdot \exp\left(-\theta^*(\alpha + 1)\right),$$

wobei C und θ^* über die Formeln (3) bzw. (1) gefunden werden können.

Diese Verteilung von Y_N wollen wir nun in Bezug zu der Verteilung von $s_{\mathcal{O}}$ setzen. Es gilt, dass jedes $(x, y) \in S_N$ eine Trajektorie $t((x, y)) \in S_{\omega,N}$ erzeugt, auf die Art und Weise wie schon zu Beginn des Kapitels gezeigt. Gleichzeitig wird jede Trajektorie $t' \in S_{\omega,N}$ von einem $(x', y') \in S_N$ erzeugt, wobei die Abbildung

$$(x, y) \mapsto t(x, y)$$

nicht bijektiv ist. (Es gilt z.B. $t(y, x) = t(x, y)\ \forall (x, y) \in S_N$.)
Für $(x, y) \in S_N$ definiere eine Zufallsvariable auf $(S_N, \mathcal{B}_N, \mathbb{P}_N)$ durch

$$Y'_N((x, y)) := Y_N(t(x, y)).$$

Dann gilt, dass $F_{Y'_N}$ auf $(S_N, \mathcal{B}_N, \mathbb{P}_N)$ mit F_{Y_N} auf $(S_\omega, \mathcal{B}_\omega, \mathbb{P}_\omega)$ übereinstimmt. Damit folgt: für großes N und großes α gilt

$$F^*_{Y'_N}(\alpha) \approx C \cdot \exp\left(-\theta^*(\alpha + 1)\right).$$

Nun definieren wir einen Pfad $\bar{t} \in S_{\omega,N}$ wie folgt: Angenommen die Trajektorie $t \in S_{\omega,N}$ nimmt nach m_1 Schritten den ersten negativen Wert an, dann übernehmen wir die ersten $m_1 - 1$ Schritte von t und im m_1-ten Schritt setzen wir $\bar{t} = 0$. Dann fahren wir fort, indem wir wieder alle Schritte von t übernehmen, bis $\bar{t}$ das nächste Mal einen negativen Wert annehmen würde und setzen dort $\bar{t}$ wieder gleich 0 und so weiter. Somit nimmt $\bar{t}$ nie negative Werte an.

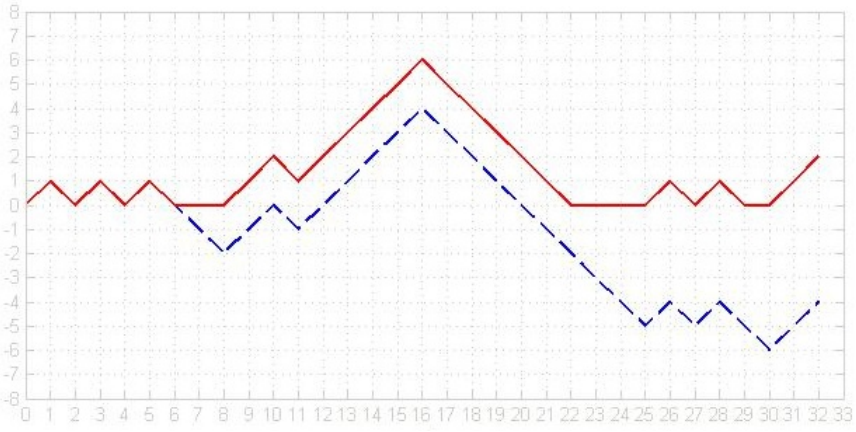

Abbildung 2: Beispiel einer Trajektorie t und ihres Pfades $\bar{t}$

Für diesen Pfad $\bar{t}$ definieren wir eine Exkursion als einen Teil von $\bar{t}$, der in 0 startet und bis zur nächsten 0 geht. Für $(x, y) \in S_N$ entspricht dann $s_{\mathcal{O}}((x, y))$ (definiert als Score eines lückenlosen, lokalen Alignments mit dem höchsten Score unter allen lückenlosen Alignments zwischen x und y, die zur Überdeckung $\mathcal{O}$ passen) gerade dem Maximum aller Höhen von Exkursionen des Pfads $\bar{t}((x, y))$. Das gilt, da jeder Abschnitt des Pfads als lokales Alignment in der Überdeckung gesehen werden kann. Dann wird die maximale Höhe aller Exkursionen genau an der Stelle angenommen, an der auch der maximale Score aller lokalen Alignments in der Überdeckung erreicht wird. Das Maximum der Höhe von Exkursionen des Pfads $\bar{t}(x, y)$ zwischen 0 und dem nächsten Leiterpunkt der Trajektorie $t(x, y)$ ist der Definition nach $Y'_N(x, y)$. Nehmen wir nun an, dass für alle $(x, y) \in S_N$ die Trajektorien $t((x, y))$ die gleiche Anzahl an Leiterpunkten besitzen und dass der letzte Schritt immer auch ein Leiterpunkt ist, so gilt:

$$s_{\mathcal{O}}((x, y)) = \max\{Y_1((x, y)), Y_2((x, y)), ..., Y_n((x, y))\} = Y_{\max}((x, y)),$$

wobei die Y_i's i.i.d. Zufallsvariablen auf $(S_N, \mathcal{B}_N, \mathbb{P}_N)$ sind, die alle die Verteilung von Y'_N haben. Dies sind natürlich starke Einschränkungen, die in der Realität nicht unbedingt gegeben sind, da wir aber „nur“ eine Approximation suchen, werden wir mit diesen Annahmen weiter arbeiten. Hierfür setzen wir

$$n = \frac{N}{\varepsilon\left(N'\right)} \text{ mit } \varepsilon\left(N'\right) = -\frac{\sum_{j=1}^{c} j \cdot R_{-j}}{\sum_{j=-c}^{d} j \cdot p_j}. \tag{4}$$

Dann gilt, dass sich $Y_{\max}$ und $s_{\mathcal{O}}$ für $N \to \infty$ beliebig nah annähern. Sind N und α groß, dann ist

$$F^*_{Y_j}(\alpha) = F^*_{Y'_N}(\alpha) \approx C \cdot \exp\left(-\theta^*\left(\alpha + 1\right)\right)$$

und damit folgt

$$F^*_{Y_{\max}}(\alpha) = 1 - F_{Y_{\max}}(\alpha) = 1 - \prod_{j=1}^{n} F_{Y_j}(\alpha) = 1 - \prod_{j=1}^{n} \left(1 - F^*_{Y_j}(\alpha)\right)$$
$$= 1 - \left(1 - C \cdot \exp\left(-\theta^*(\alpha+1)\right)\right)^n. \tag{5}$$

Definieren wir $\beta := C \cdot \exp\left(-\theta^*(\alpha+1)\right)$ so gilt für großes α und daher kleines β

$$1 - \beta \approx \exp(-\beta)$$

und Gleichung (5) ist äquivalent zu

$$F^*_{s_{\mathcal{O}}}(\alpha) \approx F^*_{Y_{\max}}(\alpha) \approx 1 - \exp\left(-nC \cdot \exp\left(-\theta^*(\alpha+1)\right)\right). \tag{6}$$

Somit haben wir bereits die Verteilung von $s_{\mathcal{O}}$ bestimmt, unser Ziel ist es aber $F^*_s(s_0 - 1)$ zu berechnen. Es gilt für alle $(x, y) \in S$

$$s((x,y)) = \max_{\mathcal{O}} s_{\mathcal{O}}((x,y)),$$

wobei das Maximum über alle möglichen Überdeckungen gebildet wird. Weiter nehmen wir an, dass alle $s_{\mathcal{O}}$ unabhängig sind und dass (6) für $n = \frac{N_{\mathcal{O}}}{\varepsilon(N')}$ und großes α eine gute Approximation ist, wobei wir mit $N_{\mathcal{O}}$ die Länge der Überdeckung $\mathcal{O}$ bezeichnen. Dann gilt

$$F_s(\alpha) = \prod_{\mathcal{O}} F_{s_{\mathcal{O}}}(\alpha) \approx \exp\left(-\frac{C}{\varepsilon(N')} \cdot \exp\left(-\theta^*(\alpha+1)\right) \sum_{\mathcal{O}} N_{\mathcal{O}}\right).$$

Sei o.B.d.A. $N_1 \geq N_2$, dann gilt

$$\sum_{\mathcal{O}} N_{\mathcal{O}} = 2 \sum_{i=1}^{N_2 - 1} i + N_2 (N_1 - N_2 + 1)$$
$$= 2 \cdot \frac{N_2(N_2-1)}{2} + N_2(N_1 - N_2 + 1) = N_1 N_2$$

und damit

$$F_s(\alpha) \approx \exp\left(-\frac{C}{\varepsilon(N')} N_1 N_2 \exp\left(-\theta^*(\alpha+1)\right)\right)$$

beziehungsweise

$$F^*_s(\alpha) \approx 1 - \exp\left(-\frac{C}{\varepsilon(N')} N_1 N_2 \exp\left(-\theta^*(\alpha+1)\right)\right) \tag{7}$$

Wir haben also gezeigt, dass s tatsächlich eine Verteilung der Form $\exp(-Km \cdot \exp(-\lambda\alpha))$ hat, wie wir bereits in der Einleitung anhand der einfachen Betrachtung vermutet haben. Wichtig ist aber hierbei zu beachten, dass wir in

unserer Herleitung einige sehr starke Annahmen gemacht haben, die in der Praxis so nicht zu finden sind, und dass die Approximation (6) nur für sehr große Überdeckungen $\mathcal{O}$ gilt, obwohl es in der Praxis durchaus auch sehr kleine Überdeckungen gibt. Dennoch wollen wir unser Ergebnis nutzen, um noch einige Überlegungen anzustellen.
Da der maximale Score s_0 eine große Ganzzahl war, können wir Formel (7) nutzen, um die Signifikanz des optimalen Alignments zwischen zwei Sequenzen x^0 und y^0 zu berechnen. Hierzu definieren wir zunächst

$$\theta^* := \lambda \text{ und } K = \frac{C}{\varepsilon\left(N'\right)} \cdot \exp\left(-\lambda\right).$$

Damit folgt aus (7)

$$F_s^*\left(s_0 - 1\right) = 1 - \exp\big[- \underbrace{KN_1N_2 \exp(-\lambda(s_0 - 1))}_{*} \big] \tag{8}$$

beziehungsweise, falls s_0 sehr groß und damit $*$ sehr klein ist

$$F_s^*\left(s_0 - 1\right) = KN_1N_2 \exp\left(-\lambda\left(s_0 - 1\right)\right). \tag{9}$$

Den Wert $F_s^*\left(s_0 - 1\right)$ nennt man P-Wert und bezeichnet ihn oft mit $P(s_0-1)$ und den Wert $KN_1N_2 \exp(-\lambda(s_0 - 1))$ nennt man E-Wert und bezeichnet ihn mit $E(s_0-1)$. Hierbei entspricht der E-Wert circa der durchschnittlichen Anzahl an lokalen Alignments zwischen x und y mit einem Score größer oder gleich s_0. Es gilt also

$$P(s_0 - 1) = \begin{cases} E(s_0 - 1) & \text{falls } s_0 \text{ groß} \\ 1 - \exp\left(E(s_0 - 1)\right) & \text{sonst.} \end{cases}$$

Um also den P- bzw. den E-Wert zu berechnen und damit zu entscheiden, ob ein gefundenes lokales Alignment signifikant ist, muss man also zunächst K und λ berechnen, wobei λ durch die Beziehung

$$\sum_{a,b \in \mathcal{Q}} p_a p'_b \exp\left[\lambda s((a,b))\right] = 1 \tag{10}$$

bestimmt werden kann. (Vergleiche hierzu Theorem 3.) In der Praxis werden beide Werte meist numerisch approximiert.

Zum Abschluss dieses Abschnitts wollen wir noch ein einfaches Beispiel betrachten.

Beispiel 2 *Sei $\mathcal{Q}$ das DNA-Alphabet und $p_a = \frac{1}{4}$ sowie $p'_b = \frac{1}{4}$ für alle $a, b \in \mathcal{Q}$. Die Substitutionsmatrix sei gegeben durch*

$$s((a,b)) = \begin{pmatrix} +1 & -1 & -1 & -1 \\ -1 & +1 & -1 & -1 \\ -1 & -1 & +1 & -1 \\ -1 & -1 & -1 & +1 \end{pmatrix}.$$

Dann erfüllt $s((a,b))$ die Bedigungen

1. $\hat{s} = 1 > 0$, $p_{\hat{s}} = \frac{1}{4} > 0$ *sowie* $p_{\tilde{s}} = \frac{1}{4} > 0$

2. *der ggT aller positiven Einträge in* $s((a,b))$ *ist 1*

3. $\sum_{a,b\in\mathcal{Q}} s((a,b))p_a p'_b = -\frac{1}{2} < 0$

Mit Gleichung (10) folgt:

$$\begin{aligned}
&\sum_{a,b\in\mathcal{Q}} p_a p'_b \exp(\lambda s((a,b))) = 1 && \Leftrightarrow \ \frac{1}{4}\cdot\exp(\lambda) + \frac{3}{4}\cdot\exp(-\lambda) = 1 \\
\Leftrightarrow\ & \frac{1}{4}\cdot\exp(2\lambda) + \frac{3}{4} = \exp(\lambda) && \Leftrightarrow \ (\exp(\lambda))^2 - 4\exp(\lambda) + 3 = 0 \\
\Rightarrow\ & \exp(\lambda) = 2 \pm \sqrt{4-3} && \Rightarrow \ \lambda = \ln(3),
\end{aligned}$$

da $\lambda \in \mathbb{R}\setminus\{0\}$ sein muss. Um K zu berechnen braucht man zunächst Q_1 und R_{-1}. Es ist

$$T = \{-1;0;1\} \quad \textit{mit} \quad p_{-1} = \frac{3}{4},\ p_0 = 0,\ p_1 = \frac{1}{4}$$

und damit $c = d = 1$. Außerdem gilt, dass R_{-1} der Wahrscheinlichkeit entspricht, dass die erste negative Zahl, die eine Trajektorie annimmt, -1 ist. Da in unserem Beispiel nur Sprünge mit Schrittweite 1 möglich sind, gilt $R_{-1} = 1$. Q_1 entspricht der Wahrscheinlichkeit, dass der erste positive Wert, den die Trajektorie annimmt, 1 ist. Damit dies erfüllt ist, muss die Trajektorie eine ungerade Anzahl an Schritten zurücklegen, genau einen mehr in positive Richtung als in negative. Daher gilt

$$Q_1 = \sum_{n=1}^{\infty} B_{2n-1} \cdot \left(\frac{1}{4}\right)^n \cdot \left(\frac{3}{4}\right)^{n-1},$$

wobei B_{2n-1} gerade der Anzahl der Trajektorien in $S_{\omega,2n-1}$ entspricht, für die $\sum_{j=1}^{2n-1} t_j = 1$ und $\sum_{j=1}^{m} t_j \leq 0$ für alle $m = 1,2,...,2n-2$ ist. Dementsprechend gilt auch

$$R_{-1} = \sum_{n=1}^{\infty} B_{2n-1} \cdot \left(\frac{1}{4}\right)^{n-1} \cdot \left(\frac{3}{4}\right)^n$$

und somit folgt

$$\begin{aligned} Q_1 &= \sum_{n=1}^{\infty} B_{2n-1} \cdot \left(\frac{1}{4}\right)^n \cdot \left(\frac{3}{4}\right)^{n-1} \\ &= \sum_{n=1}^{\infty} B_{2n-1} \cdot \left(\frac{1}{4}\right)^{n-1} \cdot \frac{3}{4} \cdot \frac{1}{3} \cdot \left(\frac{3}{4}\right)^{n-1} \\ &= \frac{1}{3} \sum_{n=1}^{\infty} B_{2n-1} \cdot \left(\frac{1}{4}\right)^{n-1} \cdot \left(\frac{3}{4}\right)^{n} \\ &= \frac{1}{3} \cdot R_{-1}. \end{aligned}$$

Da $R_{-1} = 1$ *ist, ist* $Q_1 = \frac{1}{3}$ *und* $\bar{Q} = 1 - Q_1 = \frac{2}{3}$*. Setzen wir nun alles in Gleichung (3) ein, gilt*

$$C = \frac{\frac{2}{3} \cdot [1 - 1 \cdot \exp(-\ln(3))]}{[1 - \exp(-\ln(3))] \cdot \frac{1}{3} \cdot \exp(\ln(3))} = \frac{\frac{2}{3} \cdot \frac{2}{3}}{\frac{2}{3}} = \frac{2}{3}.$$

Aus Gleichung (4) erhalten wir

$$\varepsilon\left(N'\right) = -\frac{R_{-1}}{\sum_{j=-1}^{1} j \cdot p_j} = -\frac{1}{-1 \cdot \frac{3}{4} + 1 \cdot \frac{1}{4}} = 2$$

und damit

$$K = \frac{C}{\varepsilon(N')} \cdot \exp(-\lambda) = \frac{\frac{2}{3}}{2} \cdot \exp(-\ln(3)) = \frac{1}{9}.$$

Mit diesen Ergebnissen wollen wir nun die Signifikanz des Scores zweier lokalen Alignments berechnen. Betrachte dazu zwei Sequenzen

$$x = ACATGCTG \quad \textit{und} \quad y = CATTGCGA$$

Da die Häufigkeiten aller Buchstaben in beiden Sequenzen $\frac{1}{4}$ *ist, gilt* $p_a = p'_b = \frac{1}{4}$ *und somit können wir die zuvor berechneten Werte für* K *und* λ *nutzen, wenn wir die gleiche Substitutionsmatrix verwenden. Berechnen wir dann zum Beispiel mit dem Smith-Waterman-Algorithmus die optimalen, lokalen lückenlosen Alignments, erhalten wir*

$$\begin{array}{lll} x^0 : CAT & \textit{und} & x^0 : TGC \\ y^0 : CAT & & y^0 : TGC \end{array}$$

beide mit Score 3. Dann gilt

$$s_0 = 3, \;\; N_1 = N_2 = 8, \;\; \lambda = \ln(3), \;\; K = \frac{1}{9}$$

und damit

$$\begin{aligned} P(s_0 - 1) &\approx 1 - \exp\left(-E(s_0 - 1)\right) \\ &= 1 - \exp\left[- KN_1N_2 \exp(-\lambda(s_0 - 1))\right] \\ &= 1 - \exp\left[-\frac{1}{9} \cdot 8 \cdot 8 \cdot \exp\left(-\ln(3) \cdot 2\right)\right] \\ &\approx 0.546211 > 0.05. \end{aligned}$$

Das heißt, bei diesen Annahmen sind die beiden optimalen Alignments nicht signifikant. Wobei man hierbei wieder bedenken muss, dass wir unsere Rechnungen unter der Annahme gemacht hatten, dass N_1, N_2 und s_0 groß sind.

4 Anwendung in der Praxis und BLAST

In der Praxis wird eine Sequenz x aber nicht nur mit einer anderen Sequenz y verglichen, wie wir es bisher betrachtet haben, sondern mit allen anderen Sequenzen z aus einer Datenbank $\mathcal{D}$. Wir suchen somit das optimale Alignment zwischen einem Abschnit aus x und einem Abschnitt aus allen möglichen Sequenzen z, die sich in der Datenbank befinden. Eine Möglichkeit, dieses optimale Alignment zu finden, ist die Anwendung von BLAST. BLAST steht für Basic Local Alignment Search Tool und ist ein Programm, welches eine DNA-Sequenz mit allen anderen DNA Sequenzen einer Datenbank vergleicht und einem anschließend die besten Alignments liefert. Hinzu kommt noch, dass BLAST auch angibt, wie signifikant die gefundenen Alignments sind, sodass sich die Frage stellt, wie in diesem Fall entschieden wird, ob ein Fund signifikant ist oder nicht. Eine Möglichkeit wäre, die gleichen Betrachtungen wie in Abschnitt 3 anzustellen und somit zu ignorieren, dass die optimale Lösung gefunden wurde, indem man x mit einer großen Menge an anderen Sequenzen $z \in \mathcal{D}$ verglichen hat und nicht wie zuvor mit nur einer einzigen anderen Sequenz. In der Praxis geht man aber wie folgt vor: Sei $|\mathcal{D}|$ die Summe aller Längen von Sequenzen in $\mathcal{D}$. Dann betrachten wir das lokale Alignment zwischen x und y wobei man y erhält, indem man alle Sequenzen aus $\mathcal{D}$ hintereinander hängt. Dann hat y die Länge $|\mathcal{D}|$ und den zugehörige E-Wert des Alignments erhalten wir durch

$$E_D\left(s_0 - 1\right) = KN_1|\mathcal{D}|\exp\left(s_0 - 1\right).$$

Der entsprechenden P-Wert des Alignments ist somit

$$P_D\left(s_0 - 1\right) \approx 1 - \exp\left[KN_1|\mathcal{D}|\exp\left(s_0 - 1\right)\right].$$

Es gilt, dass $P_D\left(s_0 - 1\right)$ größer ist als der P-Wert $P\left(s_0 - 1\right)$, den wir erhalten, wenn wir x nur mit einer Sequenz y vergleichen. Das heißt, wenn wir unsere ursprünglichen Ergebnisse einfach auf den Fall mehrerer Sequenzen übertragen, kann es dazu führen, dass wir einen Treffer als signifikant bezeichnen, obwohl er es bei genauerer Betrachtung gar nicht ist. Somit sollte die Entscheidung über die Signifikanz in diesem Fall anhand des Wertes $P_D\left(s_0 - 1\right)$ getroffen werden.
Das Programm BLAST geht hierbei wie folgt vor: Zuerst berechnet es den Wert $P\left(s_0 - 1\right)$ und approximiert anschließend

$$E_D\left(s_0 - 1\right) \approx E'_D\left(s_0 - 1\right) = \frac{P\left(s_0 - 1\right)|\mathcal{D}|}{N_2}.$$

Der P-Wert, anhand dessen letzendlich die Entscheidung getroffen wird, ob ein Fund signifikant ist oder nicht, wird dann durch

$$P_D\left(s_0 - 1\right) \approx 1 - \exp\left(-E'_D\left(s_0 - 1\right)\right)$$

approximiert.