Bioinformatik I, WS00/01
Uebungsblatt 5

Besprechung am 7. Dezember 2000

Guide RNA

Guide RNA dirigiert z.B. in Trypanosomen das posttranslationale Editing der mRNA sogenannter Kryptogene. Man kann die durchschnittliche Länge von Guide RNA (gRNA) analytisch abschätzen, um dann gRNA fehlerfreier identifizieren zu können. Dazu teilt man die Guide RNA in Anchor und Guiding-Region auf und schätzt für beide Regionen die Länge. Die Länge der Guiding-Region lässt sich bestimmen, wenn man die Erzeugung zufälliger Alignments von Kryptogen und Guide-RNA als Markovkette modelliert. Hier soll diese Markovkette implementiert werden.

Für die Modellierung beötigen wir zunächst die Zustände der Markovkette. Dazu unterscheiden wir *edit-events* und *matches*. Die ersten fünf Zustände

$$\begin{array}{c}\underline{\mathrm{A}}\!\rightarrow\\ \mathrm{u}\\ |\\ \mathrm{A}\end{array}\quad \begin{array}{c}\underline{\mathrm{A}}\!\rightarrow\\ \mathrm{u}\\ |\\ \mathrm{G}\end{array}\quad \begin{array}{c}\underline{\mathrm{G}}\!\rightarrow\\ \mathrm{u}\\ |\\ \mathrm{A}\end{array}\quad \begin{array}{c}\underline{\mathrm{G}}\!\rightarrow\\ \mathrm{u}\\ |\\ \mathrm{G}\end{array}\quad \text{und}\quad \begin{array}{c}\underline{\mathrm{C}}\!\rightarrow\\ \mathrm{u}\\ |\\ \mathrm{A}\end{array}$$

entsprechen edit-events. Nach den Regeln für gRNA dürfen zwischen den edit-events bis zu drei matches liegen. Um die matches zu zaehlen benötigen wir also noch drei match-Zustände.

$$\begin{array}{c}*\\ \blacksquare\\ *\end{array}\!\times 1\quad \begin{array}{c}*\\ \blacksquare\\ *\end{array}\!\times 2\quad \text{und}\quad \begin{array}{c}*\\ \blacksquare\\ *\end{array}\!\times 3 .$$

Die Wahrscheinlichkeit für eine Base $\alpha \in A, C, G, T$ im Kryptogen sei gegeben als p_α und in der gRNA als q_α.

Die Kette beginnt mit einem edit-event, also ergeben sich folgende Wahrscheinlichkeiten für den Anfangszustand s_0.

$$P\left[s_0 = \begin{array}{c}\underline{\mathrm{A}}\!\rightarrow\\ \mathrm{u}\\ |\\ \mathrm{A}\end{array}\right] = p_A q_A \quad P\left[s_0 = \begin{array}{c}\underline{\mathrm{A}}\!\rightarrow\\ \mathrm{u}\\ |\\ \mathrm{G}\end{array}\right] = p_A q_G \quad P\left[s_0 = \begin{array}{c}\underline{\mathrm{G}}\!\rightarrow\\ \mathrm{u}\\ |\\ \mathrm{A}\end{array}\right] = p_G q_A$$

$$P\left[s_0 = \begin{array}{c}\underline{\mathrm{G}}\!\rightarrow\\ \mathrm{u}\\ |\\ \mathrm{G}\end{array}\right] = p_G q_G \quad P\left[s_0 = \begin{array}{c}\underline{\mathrm{C}}\!\rightarrow\\ \mathrm{u}\\ |\\ \mathrm{A}\end{array}\right] = p_C q_A .$$

$$P\left[s_0 = \begin{array}{c}*\\ \blacksquare\\ *\end{array}\times 1\right] = 0 \quad P\left[s_0 = \begin{array}{c}*\\ \blacksquare\\ *\end{array}\times 2\right] = 0 \quad P\left[s_0 = \begin{array}{c}*\\ \blacksquare\\ *\end{array}\times 3\right] = 0.$$

Die Übergangswahrscheinlichkeiten sind in Tabelle 1 gegeben. Die Tabelle entspricht einer substochastischen Matrix, d.h. die Summe der Wahrscheinlichkeiten aus einem Zustand in den naechsten zu gehen darf kleiner als 1 sein. Die Restwahrschienlichkeit ist jeweils die Wahrscheinlichkeit für das Ende des Alignments.

Aufgabe

Schreiben Sie ein Programm zur Generierung von Cryptogen-gRNA Alignments. (Die Wahrscheinlichkeiten p_α, q_α sollen eingegeben werden koennen, Ausgabe ist eine Zustandsfolge der Markovkette)

Gehen Sie z.B. wie folgt vor.

1. Definieren Sie Daten-Strukturen (Arrays) zur Aufnahme der Anfangswahrscheinlichkeiten und der Matrix, sowie der Wahrscheinlichkeiten p_α und q_α. Schreiben Sie eine Funktion zur Initialisierung der Anfangswahrscheinlichkeiten und der Übergangsmatrix aus den p und q Wahrscheinlichkeiten. Schreiben Sie ein Programm, dass bei gegebenen p und q Wahrscheinlichkeiten, die Anfangswahrscheinlichkeiten und die Matrix ausgibt.

2. Erweitern Sie das Programm, so dass es eine zufällige Zustandsfolge der Markovkette generiert und ausgibt. (für die Erzeugung/Verwendung von Zufallszahlen siehe ggf. Programm DNA-Generierung von Blatt 2).

Generieren Sie mit dem Programm mind. 100 Zustandssequenzen für gleich verteilte Nucleotid-Wahrscheinlichkeiten $p_\alpha = q_\alpha = 0.25$ und berechnen Sie die durchschnittliche Länge der Guiding-Region-Alignments (= Länge der Zustandssequenzen).

	A→ u A	A→ u G	G→ u A	G→ u G	C→ u A	* * ×1	* * ×2	* * ×3
A→ u A	q_A	q_G	0	0	0	q_U	0	0
A→ u G	q_A	q_G	0	0	0	q_U	0	0
G→ u A	0	0	q_A	q_G	0	$q_U + q_C$	0	0
G→ u G	0	0	q_A	q_G	0	$q_U + q_C$	0	0
C→ u A	0	0	0	0	q_A	q_G	0	0
* * ×1	$p_A q_A$	$p_A q_G$	$p_G q_A$	$p_G q_G$	$p_C q_A$	0	$p_1 + p_2$	0
* * ×2	$p_A q_A$	$p_A q_G$	$p_G q_A$	$p_G q_G$	$p_C q_A$	0	0	$p_1 + p_2$
* * ×3	$p_A q_A$	$p_A q_G$	$p_G q_A$	$p_G q_G$	$p_C q_A$	0	0	0

Tabelle 1: Übergangswahrscheinlichkeiten