B. Morgenstern / M. Stanke

Übungen zur Vorlesung "Algorithmen der Bioinformatik II"
Wintersemester 2005/2006

Blatt 3

1. Löse das Netzwerk-Alignment-Problem für ein Wort S und ein Netzwerk $G = (V, E)$ mit einem etwas anderen Ansatz. Definiere für eine *Kante* $e \in E$ und eine Position $1 \leq i \leq |S|$ die Variable

   $B(e, i)$ = maximale Ähnlichkeit von $S[1..i]$ zu einem Wort, das von einem Pfad in G buchstabiert wird, der in Knoten Start beginnt und mit Kante e endet.

   Finde eine Rekursionsformel für diese Variablen und gib an, wie sich aus der Tabelle der $B(e, i)$'s die gesuchte maximale Ähnlichkeit von S zu einem Wort in G ergibt. Warum ist die Methode im Skript obiger Methode vorzuziehen?

   6 Punkte

2. Zu einer DNA-Eingabesequenz S sei eine Menge von Exonkandidaten gegeben. Jedem Exonkandidaten $e = (a, b, r, f, s)$ sei seine Anfangsposition $1 \leq a \leq |S|$ und seine Endposition $1 \leq b \leq |S|$ mit $b > a$ zugeordnet. $r \in \{\rightarrow, \leftarrow\}$ gibt an, ob er auf dem Vorwärts- bzw. Rückwärtsstrang ist. Der Leserahmen $f \in \{0, 1, 2\}$ des Exonkandidatens sei die Anzahl der Basen vor der in Leserichtung ersten Kodongrenze in e. Die Sorte $s \in \{\text{single}, \text{initial}, \text{internal}, \text{terminal}\}$ gibt an, ob e das einzige (single) Exon des Gens ist, oder ob es in Leserichtung, das erste (initial), eines der inneren (internal) oder das letzte (terminal) Exon ist. Gib an, wie man die Menge von Typen T, und die Funtionen vor(e) und nach(e) in Abschnitt 1.4.3 definieren kann, so dass die konsistenten Ketten gerade die sind, die biologisch möglichen Exon-Abfolgen entsprechen.

   6 Punkte

3. Wie muss man den Gene-Assembly-Algorithmus aus der vorhergehenden Aufgabe abändern, so dass nur vollständige Gene gefunden werden?

   4 Punkte

4. Wir wollen schätzen, wieviele Exonkandidaten im Durchschnitt eine Lange DNA-Sequenz s der Länge n enthält. Eine ungefähre Antwort reicht uns aus. Deswegen treffen wir folgende vereinfachenden Annahmen. Erstens nehmen wir an, dass jede Base der Sequenz zufällig ist, mit gleicher Wahrscheinlichkeit jede der 4 möglichen Basen ist und dass die Basen unabhängig voneinander sind. Zweitens beschränken wir uns auf Kandidaten von inneren Exons auf dem Vorwärtsstrang. Als Exonkandidat betrachten wir hier jedes Sequenzintervall von Position a bis Position b zusammen mit einem gegebenen Leserahmen f, das folgende drei Bedingungen erfüllt: Unmittelbar vor a kommt in s das Akzeptor-Spleißstellen-Dinukleotid ag. Unmittelbar nach b kommt in s das Donor-Spleißstellen-Dinukleotid gt. Und in dem angegebenen Leserahmen kommt keines der drei Stoppkodons taa, tga, tag vor.

   Die Menge der Exonkandidaten sei also

$$\begin{aligned} K(s) = \{ \quad & (a,b,f) \mid 3 \le a \le b \le n-2, f \in \{0,1,2\}, \\ & s[a-2,a-1] = ag, s[b+1,b+2] = gt, \\ & s[k,k+2] \notin \{taa, tga, tag\} \text{ für alle } k \text{ mit } a \le k \le b-2 \text{ und } k-a \equiv f (\text{mod } 3)\} \end{aligned}$$

   Hierbei bedeutet $k - a \equiv f (\text{mod } 3)$, dass $k - a$ bei Division durch 3 den Rest f ergibt. Zur Lösung der Aufgabe kannst du einen von zwei Wegen gehen.

   (a) Schreibe ein Computerprogramm, dass für einige große Werte von n die mittlere Anzahl von Exonkandidaten schätzt. Hierzu lässt du für ein gegebenes n mindestens 100 Mal eine zufällige DNA-Sequenz der Länge n erzeugen und zählst jeweils die Anzahl der Exonkandidaten. Danach bildest du den Mittelwert der Anzahlen. Berechne diesen Mittelwert für die Werte $n = 1000$, $n = 10000$, $n = 100000$.

   (b) Oder schätze die mittlere Anzahl der Exonkandidaten theoretisch für große n. Du solltest, wenn du diesen Weg gehst, eine Funktion in Abhängigkeit von n erhalten, die bis auf einen möglichst kleinen Faktor am wahren Wert dran liegt. Vorsicht: Dieser zweite Lösungsweg ist schwer.

10 Punkte

Abgabe bis Dienstag, den 22. November, bei Mario Stanke. Quelltext von Programmen an mstanke@gwdg.de. Lösungen werden am Donnerstag, den 24. November, besprochen.