Sommersemester 2003
Dipl.-Inf. A. Thum, Dr. A. Gerisch

9. Übung „Algorithmen der Bioinformatik I“

Abgabe am 3. Juli 2003 in der Übung

Institut für Informatik
Martin-Luther-Universität Halle-Wittenberg

Sei Σ ein gegegebenes Alphabet und Σ' das Alphabet erweitert um '-' für den Freiraum. Eine Ähnlichkeitsmatrix (auch *score matrix*) für Σ' ist dann eine symmetrische Matrix $C = (C(i,j)) \in \mathbb{R}^{|\Sigma'|,|\Sigma'|}$. Der Wert von $C(i,j)$ gibt die Ähnlichkeit des i-ten Zeichens von Σ' mit dem j-ten Zeichen von Σ' (also inklusive der Zeichen '-') an.
Für ein gegebenes Alignment A der Strings S_1 und S_2 über Σ bezeichnen S_1' und S_2' die Strings der Länge l nach der Einfügung der benötigten Freiräume '-'. Der Wert des Alignment A bezogen auf die Ähnlichkeitsmatrix C ist dann definiert als die Ähnlichkeit von S_1' und S_2'

$$\sum_{k=1}^{l} C(S_1'(k), S_2'(k)) \,.$$

Die Ähnlichkeit der Strings S_1 und S_2 bezogen auf die Ähnlichkeitsmatrix C ist definiert als der Wert desjenigen Alignment A von S_1 und S_2, welches maximalen Wert bezogen auf C hat. Ein solches Alignment heißt dann auch optimales Alignment von S_1 und S_2.

Aufgabe 1. Das Problem der Bestimmung der Ähnlichkeit (bezogen auf die Ähnlichkeitsmatrix C) zweier Strings S_1 und S_2 über Σ kann mit einem Dynamisches Programmieren Ansatz auf ähnliche Weise gelöst werden wie er auch für die Bestimmung der Edit-Distanz genutzt wird. Dazu bezeichne $V(i,j)$ den Wert des optimalen Alignments der Präfixe $S_1[1..i]$ und $S_2[1..j]$ für $i = 0, 1, 2, ..., |S_1|$ und $j = 0, 1, 2, ..., |S_2|$. Geben Sie diesen Algorithmus in Pseudo-Code an unter der Annahme, dass sie C kennen. Insbesondere müssen die Basiswerte $V(0,j)$ und $V(i,0)$ sowie eine Rekursionsformel für $V(i,j)$ angegeben und kurz begründet werden. (Der Wert von $V(0,0)$ kann hierbei frei gewählt werden, ohne das Ergebnis zu beeinflussen.)

Beim Vergleich von Proteinsequenzen werden sogenannte k-PAM Matrizen $C^{(k)}$ als Ähnlichkeitsmatrix verwendet. Σ sei also das 20-elementige Aminosäure-Alphabet. Wir wollen die Elemente $C_{ab}^{(k)} := C^{(k)}(a,b)$ bestimmen, in denen a und b für Proteinsymbole und nicht für '-' stehen. Im folgenden sei dies die Matrix $C^{(k)}$ (wir entfernen also die Zeile und Spalte aus dem ursprünglichen $C^{(k)}$, die für das Zeichen '-' standen).
PAM steht für *point accepted mutations* oder *percent accepted mutations* in Anbetracht der Tatsache, dass die 1-PAM Matrix gerade die Menge an Evolution wiederspiegelt, die im Durchschnitt eine Mutation pro 100 Aminosäuren verursacht. Wir wollen im folgenden erarbeiten, wie k-PAM Matrizen berechnet werden und was sie bedeuten. Die Zahl k gibt dabei den evolutionären Abstand an, auf dessen Grundlage wir Sequenzen vergleichen wollen. Die 250-PAM Matrix ist gut für den Vergleich von Sequenzen, die 250 Einheiten an Evolution auseinander sind (also vor 250 Einheiten identisch waren). Zuerst definieren wir 1-PAM Matrizen und leiten daraus die k-PAM Matrizen ab.

Für jeden evolutionären Abstand k haben wir eine Übergangsmatrix $M^{(k)}$, aus der sich die Ähnlichkeitsmatrix $C^{(k)}$ ergibt. Deshalb beginnen wir mit der Beschreibung Matrix der $M := M^{(1)}$. Um die 1-PAM Übergangsmatrix M zu konstruieren benötigt man (1) eine Liste von akzeptierten Mutationen und (2) die Wahrscheinlichkeiten p_a des Auftretens von Aminosäure a.
Akzeptierte Mutationen sind solche, die nicht zum Tod des Organismus führen und sie können aus dem Alignment homologer Proteine verschiedener Spezies gewonnen werden: jede Position des Aligments an der die Sequenzen sich unterscheiden gibt eine akzeptierte Mutation $a \leftrightarrow b$. Wir machen keinen Unterschied zwischen den Mutationen $a \leftrightarrow b$ und $b \leftrightarrow a$. Sei $f_{ab} = f_{ba} > 0$ die Anzahl des Auftretens der akzeptierten Mutation $a \leftrightarrow b$ und somit $f_{ab} = 0$ für nicht akzeptierte Mutationen. Weiter ist $f_a := \sum_{b \neq a} f_{ab}$ die Gesamtzahl der Mutationen in die a involviert war und $f := \sum_{a \in \Sigma} f_a$ die Gesamtzahl von Aminosäurevorkommen in Mutationen.
Die Auftrittswahrscheinlichkeiten p_a können über die relativen Häufigkeiten der einzelnen Proteine in einer großen, hinreichend variablen Menge von Proteinsequenzen bestimmt werden. Es gelten

$$p_a > 0 \quad \text{und} \quad \sum_{a \in \Sigma} p_a = 1 \,.$$

Die Häufigkeiten f_{ab} und die Wahrscheinlichkeiten p_a genügen, um die 1-PAM Übergangsmatrix M zu konstruieren. Sei M_{ab} die Wahrscheinlichkeit, dass a in b übergeht, wobei $a = b$ möglich ist. Die relative Mutabilität von a ist definiert als

$$m_a = \frac{f_a}{\alpha f p_a} \quad \text{mit einer positiven Konstante } \alpha \in \mathbb{R}$$

und ist ein Maß dafür, wie stark a sich verändert. Die Wahrscheinlichkeit, dass a sich nicht verändert, ist daher gegeben als

$$M_{aa} := 1 - m_a \,.$$

Weiterhin ergibt sich die Wahrscheinlichkeit, dass a nach $b \neq a$ verändert wird als Produkt der Wahrscheinlichkeit, dass a verändert wird, mit der bedingten Wahrscheinlichkeit, dass a nach b verändert wird unter der Bedingung das a verändert wird. Wir approximieren dieses Produkt durch

$$M_{ab} := \frac{f_{ab}}{f_a} m_a, \quad a \neq b \,.$$

Diese Definitionen nutzen ein stark vereinfachtes Model der Evolution von Aminosäuren, in dem z.B. Aminosäuren unabhängig von ihrer Evolutionsgeschichte mutieren.

Aufgabe 2.
(i) Zeigen Sie, dass die Zahlen M_{ab}, $b \in \Sigma$, für jedes $a \in \Sigma$ die Eigenschaften einer Wahrscheinlichkeitsverteilung erfüllen. Wie muss α dabei eingeschränkt werden?
(ii) Wie muss die Konstante α in der Definition von m_a gewählt werden, so dass in unserem Modell eine Einheit Evolution im Durchschnitt 1 von 100 Aminosäuren verändern wird? Betrachten Sie dazu die Summe $\sum_{a \in \Sigma} p_a M_{aa}$.

Aus der Übergangsmatrix M, die einer Evolutionseinheit entspricht, können wir nun die Übergangsmatrizen (-wahrscheinlichkeiten) für größere Mengen Evolution bestimmen.

Aufgabe 3. Zeigen Sie, dass die Wahrscheinlichkeit einer Mutation $a \leftrightarrow b$ in 2 Einheiten Evolution gegeben ist durch $(M^2)_{ab}$, also ein Eintrag im Quadrat $M \cdot M$ von M. Damit gilt also $M^{(2)} = M^2$. Zeigen Sie weiterhin, dass die Wahrscheinlichkeit einer Mutation $a \leftrightarrow b$ in k Einheiten Evolution gegeben ist durch $(M^k)_{ab}$, also $M^{(k)} = M^k = \underbrace{M \cdot M \cdots M}_{k \text{ mal}}$.

Nun können wir die Ähnlichkeitsmatrix $C^{(k)}$ definieren. Deren Einträge stehen in Beziehung zum Verhältnis der Wahrscheinlichkeiten, dass ein Paar (a, b) eine Mutation im Gegensatz zu einer Zufallsvertauschung ist (Likelihood). Die Wahrscheinlichkeit, dass in k Evolutionseinheiten ein a in ein b mutiert ist $(M^k)_{ab}$ und die Wahrscheinlichkeit, dass es zufällig zu einem b wurde ist p_b, das Verhältnis also $(M^k)_{ab}/p_b$. Wir definieren

$$C_{ab}^{(k)} := 10 \log_{10} \frac{(M^k)_{ab}}{p_b}.$$

Aufgabe 4. Zeigen Sie, dass die k-PAM Matrix $C^{(k)}$ symmetrisch ist.

Um einen Vergleich zweier Sequenzen zu ermöglichen, muss k fixiert werden obwohl wir die evolutionäre Distanz der Sequenzen nicht kennen. In diesem Fall sollte der Vergleich für verschiedene Werte k, z.B. $k = 40, 120, 250$ erfolgen, wobei kleine Werte k gut für das Finden kurzer, starker Ähnlichkeiten sind und große Werte k detektieren lange, schwache Ähnlichkeit besser.

Aufgabe 5. Wir haben gesehen, dass die 1-PAM Matrix im Durchschnit 1% geänderter Aminosäuren entspricht. Kann man sagen, dass die 2-PAM Matrix im Durchschnit 2% geänderter der Aminosäuren entspricht?

Aufgabe 6. Wenn k-PAM Matrizen in der Bestimmung optimaler Alignments benutzt werden, dann kann es passieren, dass der Wert eines identischen Alignments ($S_1 = S_2 = S_1' = S_2'$) kleiner ist als der Wert für bestimmte nicht perfekte Alignments ($S_1 = S_2, S_1 \neq S_1', S_2 \neq S_2'$). Erklären Sie, wie das passieren kann. Ist das ein unerwünschtes Verhalten der Ähnlichkeitsmatrix?