Fachbereich Informatik

Lehrgebiet Technische Informatik II

Kurs 1738 "Einführung in die Bioinformatik"

Lösungsvorschläge zur

Hauptklausur im WS 2003/04

am 07.02.2004

Aufgabe 1 Sequenzvergleich (20 Punkte)

a) Begründen Sie, welche Schlüsse aus hinreichender Ähnlichkeit zwischen Proteinsequenzen gezogen werden können. Aus welchen biochemischen Gesetzmäßigkeiten leiten sich diese Folgerungen ab? (2 Punkte)

Sind sich zwei Proteinsequenzen hinreichend ähnlich, d.h. weisen sie z. B. mehr als 30% identische Residuen auf, so kann mit großer Sicherheit auf ähnliche Funktion bzw. Struktur geschlossen werden. Die physikalisch/chemischen Eigenschaften der Aminosäuren determinieren die Faltung und damit die Struktur der Proteine. Zueinander ähnliche Sequenzen haben in der Regel ähnliche Struktur bzw. Funktion.

b) Welche Elemente gehören zu einem Verfahren, mit dem eine Querysequenz gegen das bisher gesammelte Wissen zu Proteinsequenzen verglichen werden kann? Nennen Sie zu jedem Element dessen Funktion und einen repräsentativen Vertreter. (3 Punkte)

Ein derartiges Verfahren besteht aus drei Elementen. 1) Einem Algorithmus, der es erlaubt, die Eingabe paarweise mit allen Sequenzen der Datenbank zu vergleichen und hierbei jeweils eine Distanz/einen Score zu bestimmen. Vertreter sind z. B. der Needleman-Wunsch-Algorithmus oder BLAST. 2) Einem Scoring-System mit dem die einzelnen Symbole und Lücken bewertet werden. Beispiel: BLOSUM-Matrix, affine Kostenfunktion. 3) Eine annotierte Datenbank, in der zu den Sequenzen Information z.B. zu deren Funktion gesammelt ist. Beispiel: SWISSPROT.

c) Der Needleman-Wunsch (NW)-Algorithmus kann abgeleitet werden aus einer speziellen Distanz zum Vergleich von Zeichenketten. Beschreiben Sie diese, insbesondere die erlaubten Operationen. (3 Punkte)

Mit der Levenshtein-Distanz wird der Aufwand bewertet, der notwendig ist, um eine Zeichenkette A in eine Zeichenkette B zu überführen. Für das Überführen sind drei Operationen erlaubt: Das Ersetzen eines Symbols durch ein anderes, das Löschen von Symbolen und das Einführen von Lücken. Jede dieser Operationen ist mit "Kosten" bewertet. Diejenige Kombination von Operationen, die für die Überführung die geringsten Kosten verursacht, ist die Levenshtein-Distanz D(A, B).

d) Geben Sie den NW-Algorithmus in Pseudocode an, kommentieren Sie kurz und analysieren Sie die Laufzeit. Beziehen Sie sich auf die scorebasierte Version. (3 Punkte)

Seien $A = a_1 \ldots a_n$ und $B = b_1 \ldots b_m$ zwei Zeichenketten der Längen n bzw. m. Seien $s(\varepsilon)$ und $s(a, b)$ Scores für das Einführen von Lücken bzw. Scores für das Alignieren der Symbole a und b. Dann kann formuliert werden:

For $i = 0$ to n do $S_{i,0} = i \times s(\varepsilon)$
For $j = 0$ to m do $S_{0,j} = j \times s(\varepsilon)$

For $i = 1$ to n do
  For $j = 1$ to m do
    $S_{i,j} = max(S_{i-1,j-1} + s(a_{i-1}, b_{j-1}), S_{i-1,j} + s(\varepsilon), S_{i,j-1} + s(\varepsilon))$

Die Laufzeit ist von $O(n^2)$, da zwei ineinandergeschachtelte Schleifen ausgeführt werden.

e) Begründen Sie, weshalb beim Vergleich von Proteinsequenzen der Smith-Waterman (SW)-Algorithmus eine wichtige Rolle spielt. Welche Änderungen müssen für die Umstellung im NW-Algorithmus eingeführt werden und wie begründen Sie diese? (3 Punkte)

Der Smith-Waterman-Algorithmus bestimmt lokale Alignments. Häufig wird in Proteinsequenzen nach Proteindomänen gesucht und weniger nach genereller Übereinstimmung zweier Sequenzen. Diese Aufgabe erfüllt der SW-Algorithmus. In den NW-Algorithmus werden die folgenden Änderungen eingeführt: 1) Die Bedingung zur Berechnung des lokalen Scores $S_{i,j}$ wird erweitert auf $max\ (\ 0,\ S_{i-1,j-1} + s(a_{i-1},b_{j-1}),\ S_{i-1,j}+s(\epsilon),\ S_{i,j-1}+s(\epsilon))$. 2) Die erste Zeile und die erste Spalte werden mit 0 initialisiert, 3) Zur Identifikation des lokalen Alignments wird in der Matrix der größte Wert gesucht. Durch 1) wird erreicht, dass das Erweitern eines Alignments abgebrochen wird, sobald der erreichte Score hinreichend "schlecht" ist. Damit hat jede Position die Chance, Anfang eines lokalen Alignments zu werden. Diese Begründung gilt auch für Maßnahme 2). Änderung 3) ergibt sich automatisch, da das Ende eines lokalen Alignments beliebig in der Matrix liegen kann.

f) Die Umstellung auf den SW-Algorithmus setzt eine entsprechende Skalierung des Scoring-Systems voraus. Wie muss dieses beschaffen sein? Zeigen sie Konsequenzen ungenügender Normierung auf: Welche Ergebnisse erwarten Sie für das Alignment von Zufallsequenzen? (3 Punkte)

Der Algorithmus beruht auf der Annahme, dass der Score unter Null fallen kann. Wäre dies nicht der Fall, würde das Erweitern von Alignments nicht abgebrochen. Dies geschieht, wenn der Erwartungswert für Scores, der sich aus der Scoring-Matrix ergibt, kleiner als Null ist. Diese Bedingung lässt sich durch Addieren eines konstanten Terms zur Scoring-Matrix einstellen.

g) Geben Sie Techniken an, die eingesetzt werden, um den paarweisen Sequenzvergleich in heuristischen Verfahren zu beschleunigen. Erläutern Sie die Funktionsweise und mögliche Nachteile. (3 Punkte)

Es werden in der Regel drei Techniken angewandt, um den Vergleich zu beschleunigen: 1) *Preprocessing*, d.h. das Ablegen von Indizes auf ähnliche Teilsequenzen; basierend z. B. auf Scoring-Matrizen wie Blosum. 2) Identifizieren von Bereichen starker lokaler Sequenzähnlichkeit als Ausgangspunkt für die Scoreberechnung. 3) Abbrechen der Berechnung lokaler Scores, wenn es sich nicht mehr lohnt, weiterzumachen.

Mit diesen Techniken wird erreicht, dass nicht sämtliche Teilergebnisse berechnet werden müssen. Konsequenz aus 1) ist, dass der Datenbestand an Sequenzen indiziert werden muss, diese Tabellen müssen bei jedem Update aktualisiert werden.

Aufgabe 2 Profile (7 Punkte)

a) Definieren Sie den Begriff des Profils und begründen Sie, weshalb ein Profil eine Proteinfamilie präziser beschreibt als eine Konsensus-Sequenz. (2 Punkte)

Ein Profil ist eine Matrix, in der positionsspezifisch Häufigkeiten/Scores für das Vorkommen der Symbole in der Sequenzfamilie angegeben werden. Meist werden die spezifischen Häufigkeiten normiert, d.h. durch mittlere Symbolhäufigkeiten dividiert. Eine Konsensus-Sequenz beschreibt genau eine Sequenz, Variationen, die an den einzelnen Positionen vorkommen können, sind nicht mehr sichtbar. In einem Profil hingegen sind diese Modifikationen präzise dokumentiert. Anforderungen an die einzelnen Residuen werden genau aufgezeigt.

b) Beschreiben Sie, wie aus einem multiplen Sequenzalignment positionsspezifische log-odds Scores abgeleitet werden können. Wozu dienen hierbei Pseudocounts? (3 Punkte)

Sei $M = [1,n,\ 1,k]$ ein Multiples Sequenzalignment (MSA) bestehend aus k Sequenzen der Länge n.
Sei $\Sigma = \{ a_1 \ldots a_m \}$ das Alphabet der in M vorkommenden Symbole.
Sei $f(a_i)$ die Häufigkeit für das Vorkommen von a_i in M.
Sei $f(a_i, k)$ die Häufigkeit für das Vorkommen von a_i an Position k des MSAs.

Dann ist $S[a_i, k] = log(f(a_i, k) / f(a_i))$ der Score für das Vorkommen von a_i an Position k.

Ist die Stichprobenmenge klein, so kann es passieren, dass die ermittelten Vorkommen aufgrund der begrenzten Anzahl von Beobachtungen nicht mit den "wahren" Häufigkeiten übereinstimmen. Zur Korrektur dieses Effekts werden Pseudocounts eingeführt, d.h. die Anzahl von Beobachtungen wird korrigiert.

c) Welche Vorteile hat eine logarithmierte Angabe und wie wird diese Vorgehensweise theoretisch begründet? Tipp: Wie wird beim Alignment der Score für die Zeichenkette berechnet? (2 Punkte)

Die theoretische Begründung für die Verwendung von log-odds Scores stammt aus der Testtheorie, genauer aus dem Vergleich zweier Modelle im Neyman-Pearson Test:

$$f(a_1)\, f(a_2) \ldots f(a_n) \,/\, g(a_1)\, g(a_2) \ldots g(a_n)\ .$$

In beiden Modellen wird Unabhängigkeit der Einzelereignisse vorausgesetzt, deswegen werden im Zähler und Nenner zur Berechnung der Wahrscheinlichkeit für das Auftreten der Sequenzen $a_1 \ldots a_n$ Einzel-Häufigkeiten multipliziert. Geht man zur logarithmierten Darstellung über und betrachtet die Terme positionsweise, so kommt man zur Addition von log-odds-Scores. Ein derartiges, additives Bewertungsverfahren benötigt man z. B. bei der Berechnung von Alignments mittels dynamischer Programmierung.

Aufgabe 3 Genetische Algorithmen und Neuronale Netze (6 Punkte)

a) In beiden Verfahren besteht die Gefahr, in lokalen Optima des Lösungsraumes gefangen zu bleiben. Beschreiben Sie Konzepte, die verfolgt werden, um dieser Gefahr zu entgehen: Welche Vorgehensweise bietet sich bei Neuronalen Netzen an, wie versuchen Genetische Algorithmen mit dem Problem fertig zu werden? (2 Punkte)

Bei Neuronalen Netzen werden häufig mehrere Netze parallel betrieben, die mit unterschiedlichen Parametersätzen trainiert wurden. Ergeben sich übereinstimmende Lösungen, so ist mit hoher Wahrscheinlichkeit das globale Optimum gefunden.

Bei Genetischen Algorithmen wird bei der Zusammenstellung der nächsten Generation stets eine Zufallskomponente berücksichtigt, die sicherstellen soll, dass der komplette Suchraum abgetastet wird.

b) Was ist orthogonale Kodierung? Weshalb wird sie eingesetzt? (2 Punkte)

Sei k die Größe des Alphabets Σ, sei A ein n-tupel über Σ. Sei a_i das i-te Symbol aus Σ und sei a_i wie folgt durch ein k-tupel dargestellt: $OK(a_i) = ([\ 0 \mid 1\ ]^k \mid$ die i-te Stelle sei 1, sonst trete nur die 0 auf).

Dann kann A wie folgt dargestellt werden: $A = OK(1) \ldots OK(n)$.

Orthogonale Kodierung wird z.B. bei der Kodierung der Eingabe von Neuronalen Netzen oder in Genetischen Algorithmen eingesetzt, um die lineare Separabilität zu erleichtern, bzw. um die Gewichtung der Allele möglichst neutral zu halten.

c) Wo spielt lineare Separabilität eine Rolle? Wie beeinflusst sie die Architektur von Neuronalen Netzwerken? (2 Punkte)

In vielen Klassifikationsverfahren wird versucht, die Objekte so in einen hochdimensionalen Raum abzubilden, dass es möglich wird, Klassen durch Hyperebenen zu trennen. In NNs muss der letzten Schicht das Problem so präsentiert werden, dass es linear separabel ist, sonst ist das Problem mit NNs nicht lösbar. Durch die Einführung einer hinreichenden Anzahl von Schichten wird in NNs eine geeignete Umkodierung ermöglicht. Meist reichen in bioinformatischen Anwendungen drei Schichten.

Aufgabe 4 Taxonomie (5 Punkte)

a) Ein wichtiges Konzept taxonomischer Bewertung sind additive Matrizen. Beschreiben Sie kurz den einfachsten Algorithmus, mit dem überprüft werden kann, ob eine Matrix additiv ist. (2 Punkte)

Für additive Matrizen muss für jede Auswahl von 4 Einträgen i, j, k, l eine Permutation existieren, für die gilt:

$$d(i, k)+ d(j, l) \leq d(i, j)+d(k, l) = d(i,l) + d(k,j)$$

Dies kann durch systematisches Überprüfen aller Fälle nachvollzogen werden.

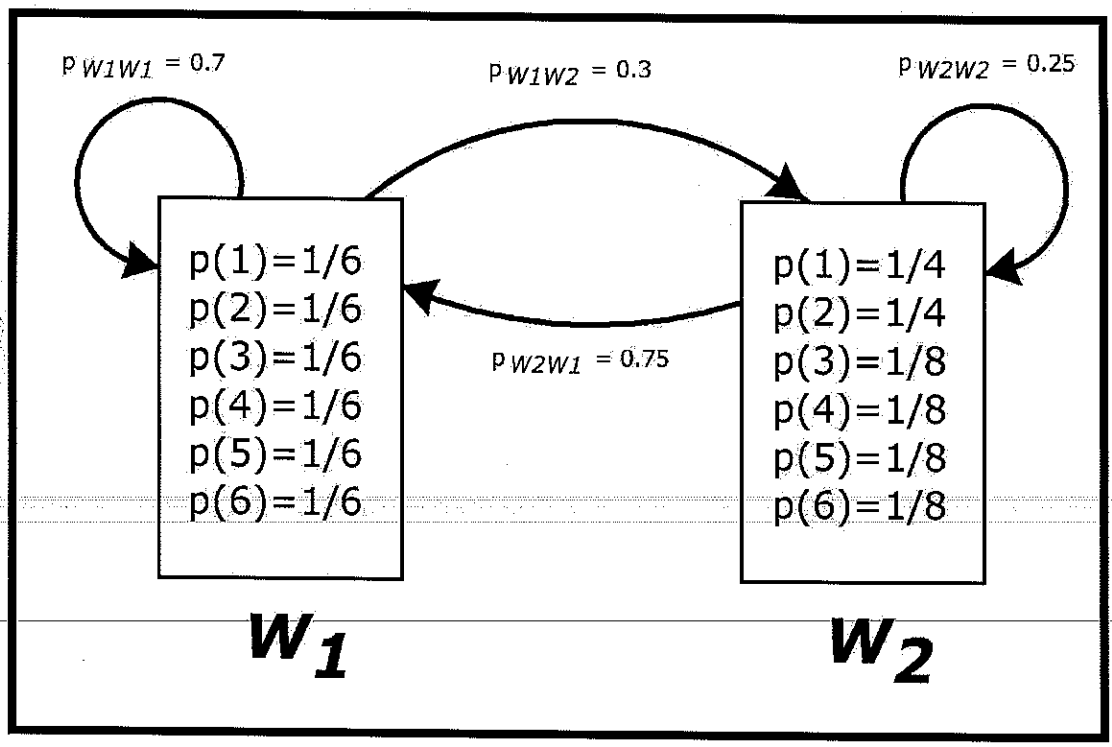
b) Beschreiben Sie kurz (max. zwei Sätze) die Idee, die dem Neighbour-Joining-Algorithmus zugrunde liegt. (3 Punkte)

Beim Neighbour-Joining Algorithmus werden jeweils diejenigen zwei Objekte (Gruppen) zu einem weiteren Objekt zusammengefasst, die a) möglichst isoliert und b) zueinander am nächsten liegen. Objekte werden solange verknüpft, bis nur noch ein einziges vorliegt.

Aufgabe 5 Hidden-Markov-Modelle (12 Punkte)

In einem unehrlichen Casino werden zwei Würfel W_1 und W_2 im Wechsel verwendet. Beim Würfel W_1 treten die Augenzahlen jeweils mit gleicher Wahrscheinlichkeit auf, beim Würfel W_2 kommen die 1 und die 2 mit jeweils $p = ¼$, vor, die restlichen Wahrscheinlichkeiten sind gleichverteilt. Wird Würfel W_1 verwendet, so wird im folgenden Wurf mit $p = 0.3$ W_2 eingesetzt. Auf W_2 folgt mit $p = 0.25$ wiederum W_2.

a) Zeichnen Sie ein Zustandsdiagramm, geben Sie sämtliche Wahrscheinlichkeiten an und beschreiben Sie das System. (5 Punkte)



Das System hat zwei Zustände W_1 und W_2, die Symbole "1" bis "6" werden mit den angegebenen Wahrscheinlichkeiten emittiert. Zwischen den beiden Zuständen wird mit den eingetragenen Wahrscheinlichkeiten $p_{w_i w_j}$ gewechselt.

b) Geben Sie die Gesamtwahrscheinlichkeit für folgende Kette von Zuständen / Beobachtungen in Form einer Folge von Multiplikatoren an: (2 Punkte)

Augenzahl	2	2	6	6	1
Würfel	W_1	W_2	W_2	W_1	W_1

p = 1/6 x 0.3 x 1/4 x 0.25 x 1/8 x 0.75 x 1/6 x 0.7 x 1/6

c) Was ist bei dieser Anwendung "hidden" ? (2 Punkte)

In dieser Anwendung ist nicht bekannt, welcher der beiden Würfel verwendet wird. Der Beobachter sieht nur die Folge der Augenzahlen.

d) Ein wichtiges Konzept bei HMMs ist der Viterbi-Pfad. Geben Sie an, was die zu den Viterbi-Variablen gehörenden Werte jeweils ausdrücken. Wie werden sie berechnet? Zu welcher Klasse von Verfahren rechnet man den Algorithmus? (3 Punkte)

Die Viterbivariable gibt an, mit welcher Wahrscheinlichkeit der betrachtete Zustand auf dem wahrscheinlichsten Pfad erreicht wird, wobei die Folge der betrachteten Emissionen auftritt. Der Algorithmus benutzt die Technik der dynamischen Programmierung ähnlich wie Alignmentverfahren.

Ende