FREIE UNIVERSITÄT BERLIN
Fachbereich Mathematik und Informatik
Institut für Informatik (WE 3)

FU BERLIN

Freie Universität Berlin
FB Mathematik und Informatik, Institut für Informatik, Takustr. 9, D-14195 Berlin

Prof. Dr. Knut Reinert,
Andreas Döring,
Moritz Blöcker

15. Februar 2004

Algorithmische Bioinformatik

WS 2003/04

Klausur

Name, Vorname	Matr.-Nr.

Zur Bearbeitung der Klausur stehen Ihnen 120 Minuten zur Verfügung. Die hier gestellten Aufgaben sind jedoch für eine deutlich längere Bearbeitungszeit konzipiert. Sie werden somit vermutlich nicht sämtliche Aufgaben bearbeiten können, und dies wird auch nicht von Ihnen verlangt.

Abgesehen von einem nicht-programmierbaren Taschenrechner sind keinerlei Hilfsmittel gestattet. Geben Sie auf dem Titelblatt ihren Namen und ihre Immatrikulationsnummer an. Schreiben Sie ihre Lösungen direkt auf die entsprechenden Aufgabenbögen. Sollte dort der Platz nicht ausreichen und Sie weitere Blätter benötigen, vermerken Sie dies bitte, damit wir auch den Rest ihrer Antwort finden und bei der Bewertung berücksichtigen können.

Am Ende der Klausur sind sämtliche Aufgabenblätter wieder abzugeben.

Ergebniss:

	Punkte	max.
1		8
2		4
3		6
4		6
5		9
6		9
7		8
8		6
9		8
10		6
11		6
12		6
13		8
14		6
15		6
16		3
17		3
18		5
19		4
20		3
Σ		120

1. [*8 Punkte*] Beschreiben Sie einen Algorithmus, der ein Sequenzalignment (*score* und *traceback*) mit linearen Gapkosten in linearem Platz berechnet. [15]

2. [*4 Punkte*] Gegeben seien zwei Sequenzen der Längen 1000 und 40000. Bei einem Sequenzvergleich wird ein Alignment mit einem Score von 20 *bit* gefunden. Entscheiden Sie durch eine grobe Abschätzung, ob dieser Wert signifikant ist. [5]

3. [*6 Punkte*] Blat hat eine *seed phase* und eine *extend phase*. [10]

   (a) Welche drei Strategien bietet Blat in der *seed phase* dafür an, mögliche *seeds* zu finden? [*2 Punkte*]

   (b) Geben Sie für eine dieser Strategien Formeln zur Abschätzung der Sensitivität und der Spezifizität an. [*4 Punkte*]

4. [*6 Punkte*] Schätzen Sie Zeit- und Platzbedarf (in O-Notation) für die Berechnung eines multiplen Stringalignments von k Strings der Länge n mit einfachem dynamischen Programmieren und WSOP-Kostenfunktion ("*weighted sum of pairs*") ab. Begründen Sie Ihre Antwort. [10]

5. [*9 Punkte*] [15]

   (a) Beschreiben Sie die Idee des in der Vorlesung beschriebenen Algorithmus für ein exaktes MSA ("*multiple sequence alignment*") nach dem Prinzip des *divide and conquer*. [*3 Punkte*]

   (b) Beschreiben Sie eine Methode, wie man im *divide and conquer*-Alignment effizient gute Schnittpositionen finden kann. [*3 Punkte*]

   (c) Wie kann man es beim exakten MSA vermeiden, die gesamte *dynamic programming*-Matrix aufbauen zu müssen. [*3 Punkte*]

6. [*9 Punkte*] Gegeben sei das folgende HMM mit Startzustand, Endzustand und zwei ausgebenden Zuständen 1 und 2. Die Werte in den Zuständen 1 und 2 entsprechen den Ausgabewahrscheinlichkeiten für die Zeichen A und B. Die Werte an den Kanten sind die Übergangswahrscheinlichkeiten. [15]

   Bestimmen Sie mit dem Viterbi Algorithmus einen der wahrscheinlichsten Pfade durch das HMM vom Start bis zum Endzustand, wenn bekannt ist, dass dabei genau die Zeichenkette "ABA" ausgegeben wird, und geben Sie die Wahrscheinlichkeit dafür an, dass das HMM durch diesen Pfad läuft und "ABA" ausgibt.

7. [*5 Punkte*] Gegeben die folgende Distanzmatrix. Ist dies eine additive Metrik? Ist dies eine Ultrametrik? Begründen Sie Ihre Antworten. [8]

	A	B	C	D
A	0	5	2	4
B	5	0	1	3
C	2	1	0	4
D	4	3	4	0

8. [*6 Punkte*] Gegeben sei folgende Distanzmatrix: [10]

	A	B	C	D	E
A	0	10	4	5	9
B		0	12	9	7
C			0	7	11
D				0	8
W					0

Welche Baumtopologie berechnet UPGMA für diese Distanzmatrix? Begründen Sie Ihre Antwort.

9. [*8 Punkte*] Gegeben sei die folgende Liste der Sequenzstücklängen eines PDP: [15]

$$E = \{3, 10, 15, 7, 18, 8\}$$

Beschreiben Sie kurz, wie der Skiena Algorithmus nach Eingabe von E eine Menge X von Restriktionsstellen berechnet, indem Sie die bei jedem Schritt berechneten Mengen von Restiktionsstellen und Stücklängen angeben.

10. [*6 Punkte*] Gegeben sei folgende Matrix: [10]

a	b	c	d
0	1	0	1
1	0	0	1
1	0	1	1

(a) Ordnen Sie die Spalten der Matrix auf eine Weise an, dass die *consecutive ones property* erfüllt ist. [*2 Punkte*]

(b) Bei größeren Matrizen lässt sich nicht mehr auf den ersten Blick sagen, ob sich die *consecutive ones property* überhaupt durch eine Spaltenpermutation erfüllen lässt. Man möchte dann allgemeiner die Gesamtzahl der separaten Einserblöcke minimieren. Dieses Problem lässt sich als ein Travelling-Salesman-Problem formulieren. Stellen Sie die Distanzmatrix des TSPs für die oben stehende Matrix auf. [*4 Punkte*]

11. [*6 Punkte*] Geben Sie eine möglichst kurze RNA-Sequenz an, die sich zu einer Struktur falten könnte, welche mindestens eine Wölbung (*bulge*) und zwei Haarnadelschleifen (*hairpin loops*) enthält. Zeichnen Sie diese Struktur in 2 gängigen Darstellungsweisen auf. [10]

12. [*6 Punkte*] Geben Sie die *fill stage* (Initialisierung und Rekursionsformel) des Nussinov Algorithmus an. [10]

13. [*8 Punkte*] Gegeben folgendes Alignment von 4 RNAs: [15]

seq 1:	A	C	A	A	A
seq 2:	A	C	C	U	A
seq 3:	A	U	G	A	C
seq 4:	A	U	U	A	G

Berechnen Sie dazu ein RNA Sequenzlogo, d.h. geben Sie zu jeder Spalte die Höhe der einzelnen Buchstaben und die Gesamthöhe aller Buchstaben der Spalte an.

14. [*6 Punkte*] Geben Sie die Rekursionsgleichung eines Algorithmus zur Berechnung des optimalen strukturellen Alignments von zwei Sequenzen S_1 und S_2 an, die mit Sekundärstrukturen P_1 bzw. P_2 ohne *pseudo knots* annotiert sind. [10]

15. [*6 Punkte*] Gegeben sei ein Protein aus 100 Aminosäuren. Die Konformation des Proteins sei modellhaft allein durch die Φ- und Ψ-Winkel des Rückgrates beschrieben. [10]

(a) Wie viele Freiheitsgrade bestimmen die Konformation dieses Proteinmodells? [*2 Punkte*]

(b) Angenommen, für jeden der Φ- und Ψ-Winkel werden nur 10 diskrete Werte zugelassen. Wie viele Konformationen gibt es dann? [*2 Punkte*]

(c) Schätzen Sie ab, wie viele Jahre ein Cluster mit 1000 Prozessoren brauchen würde, um alle möglichen Konformationen energetisch zu bewerten, wenn jede CPU 4 $GFlops/s$ durchführen kann und pro Freiheitsgrad 200 Gleitkommaoperationen benötigt werden würde, das System aber im Schnitt nur 10% seiner theoretischen Spitzenleistung erbringt. [*2 Punkte*]

16. [*3 Punkte*] Was braucht man für die Durchführung eines Moleküldynamikexperiments? [5]

17. [*3 Punkte*] Was versteht man unter *virtual screening*? [5]

18. [*5 Punkte*] Angenommen, es werden n_1 Replikate einer Hybridisierung von wildtyp Hefe und n_2 Replikate einer Hybridisierung einer Hefemutante gemacht. Wie testet man, ob sich ein Gen signifikant verändert hat? Nennen Sie mögliche Tests und geben Sie eine Teststatistik an. [8]

19. [*4 Punkte*] Welche 4 Schritte werden in der Massenspektrometrie bei der Umwandlung von rohen MS-Daten in Stickdaten durchgeführt, und wozu dienen diese Schritte? [5]

20. [*3 Punkte*] Was sind die beiden Hauptschritte bei SCOPE? [5]