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,
Markus Bauer,
Eva Lange, Erasmus Krause

17. Februar 2006

Algorithmische Bioinformatik

WS 2005/06

Klausur

Ergebnis:

	Punkte	max.
1		4
2		8
3		8
4		10
5		12
6		12
7		8
8		6
9		10
10		10
11		6
12		8
Σ		102

Name, Vorname	Matr.-Nr.

Zur Bearbeitung der Klausur stehen Ihnen 90 Minuten zur Verfügung. Die hier gestellten Aufgaben sind jedoch für eine längere Bearbeitungszeit konzipiert (102 min). 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.

1. [*4 Punkte*] Gegeben seien zwei Sequenzen der Längen 1000 und 10000. Bei einem Sequenzvergleich wird ein Alignment mit einem Score von 25 *bit* gefunden. Entscheiden Sie durch eine grobe Abschätzung, ob dieser Wert signifikant ist und begründen Sie ihre Antwort.

2. [*8 Punkte*] In der Vorlesung wurde eine Generalisierung des paarweisen *dynamic programming* Alignment—Algorithmus auf mehrere Sequenzen besprochen.

   (a) Schätzen Sie Zeit- und Platzbedarf (in O-Notation) für die Berechnung eines multiplen optimalen Stringalignments von k Strings der Länge n mit einfachem dynamischen Programmieren und WSOP-Kostenfunktion ("*weighted sum of pairs*") ab. Begründen Sie Ihre Antwort. [*3 Punkte*]

   (b) Erklären Sie genau, wie man es beim exakten MSA vermeiden kann, die gesamte *dynamic programming*-Matrix aufbauen zu müssen. [*5 Punkte*]

3. [*8 Punkte*] Der Quasar Algorithmus findet sogenannte *local approximate matches*, d.h. er findet alle gemeinsamen Substrings von Text und Pattern der Größe w, die höchstens k mismatches besitzen. Der Algorithmus basiert auf dem in der Vorlesung besprochenen q-gram Lemma.

   (a) Geben Sie das q-gram Lemma an.[*4 Punkte*]

   (b) Die Wahl von q ist wichtig in Quasar und nicht offensichtlich. Welchen Grund gibt es q möglichst klein zu wählen? Welchen Grund gibt es, q möglichst groß zu wählen?[*4 Punkte*]

4. [*10 Punkte*] Das aus der Vorlesung bekannte Motif-Suchproblem ist folgendermaßen definiert:

   Gegeben seien t Sequenzen, und ein sogenanntes (l, d) Motif der Länge l: Jede der t Sequenzen enthält einen Substring, der zum Motif nicht mehr als d Unterschiede aufweist. Der PROJECTION Algorithmus sucht nach einem solchen (l, d) Motif.

   (a) Erklären Sie, welche Bedeutung die 3 Parameter k, s und m im PROJECTION Algorithmus haben? [*3 Punkte*]

   (b) Beschreiben Sie den Algorithmus in Pseudocode. (Hilfe: Sie können eine Funktion Hash als gegeben annehmen) [*7 Punkte*]

5. [*12 Punkte*] Gegeben sei die folgende Liste der Sequenzstücklängen eines PDP (d.h. alle paarweisen Distanzen):

$$E = \{1, 3, 4, 5, 6, 7, 8, 9, 10, 15\}$$

Rechnen Sie den Skiena Algorithmus nach Eingabe von E durch und geben Sie die resultierende Menge X von Restriktionsstellen an. Geben Sie bei jedem Schritt die berechneten Mengen von Restriktionsstellen und Stücklängen an.[*10 Punkte*]

Die berechnete Lösung ist nicht eindeutig. Welche andere Lösung gibt es noch?[*2 Punkte*]

6. [*12 Punkte*]

(a) Was ist die Zielfunktion bei der Rekonstruktion phylogenetischer Bäume mittels Maximum Parsimony?[*3 Punkte*]

(b) Wozu wird Bootstrapping in der Phylogenie benutzt? [*3 Punkte*]

(c) Gegeben seien vier Objekte a, b, c und d sowie die unten abgebildete Distanzmatrix. Bildet die Matrix eine Metrik, eine additive Metrik oder eine Ultrametrik? Warum?

	a	b	c	d
a	0	9	9	9
b		0	5	5
c			0	2
d				0

[*6 Punkte*]

7. [*8 Punkte*] Sei P die 1-Schritt Übergangsmatrix eines Markovprozesses und τ die Verteilung des Prozesses. Dabei sei τ sei ein Zeilenvektor und das Matrixelement P_{ij} bezeichne die Übergangswahrscheinlichkeit vom Zustand i in Zustand j. Es gelte die sogenannte *detailed balance* Gleichung

$$\tau_i P_{ij} = \tau_j P_{ji} \qquad \text{für alle} \quad i, j$$

Zeigen Sie, dass τ die stationäre Verteilung des Markovprozesses ist, also dass $\tau P = \tau$ gilt.

8. [*6 Punkte*] Geben Sie eine RNA-Sequenz an, die sich zu einer Struktur falten könnte, welche mindestens eine innere Schleife (*interior loop*) und eine Haarnadelschleife (*hairpin loop*) enthält. Zeichnen Sie diese Struktur in zwei gängigen Darstellungsweisen und kennzeichnen Sie die *interior* und *hairpin loop*.

9. [*10 Punkte*] Benutzen Sie ein vierdimensionales Feld A zur Berechnung eines optimalen Sequenz-Struktur Alignments zwischen zwei annotierten Sequenzen (S_1, P_1) and (S_2, P_2). Nehmen Sie an, dass die Annotationen Strukturen ohne Pseudoknoten sind.

   (a) Geben Sie die Initialisierung von A an. [*3 Punkte*]

   (b) Geben Sie die Rekursionsvorschrift zum Berechnen eines optimalen Sequenz-Struktur Alignments an.[*5 Punkte*]

   (c) Wie ist die Laufzeit und Platzbedarf des Algorithmus? [*2 Punkte*]

10. [*10 Punkte*] Sie haben in der Vorlesung zum Thema Kraftfeldmethoden im Bereich "Statistische Thermodynamik (Monte-Carlo Methoden)" eine Formel für die Metropolis-Akzeptanzwahrscheinlichkeit kennengelernt:

$$P_A(q \to \widetilde{q}) = \min\{1, \exp(-\beta\,(V(\widetilde{q}) - V(q)))\}.$$

In der Vorlesung wurde gezeigt, dass diese die Bedingung für ein korrektes Sampling bei "symmetrischer Vorschlagwahrscheinlichkeit", d.h. für den Fall $P_V(q \to \widetilde{q}) = P_V(\widetilde{q} \to q)$, erfüllt. Erfüllt für einen symmetrischen Vorschlagschritt auch die folgende Formel

$$P_A(q \to \widetilde{q}) = \frac{\exp(-\beta\, V(\widetilde{q}))}{\exp(-\beta\, V(\widetilde{q})) + \exp(-\beta\, V(q))},$$

die Bedingung für ein korrektes Sampling? Begründen Sie Ihre Meinung.

11. [*6 Punkte*] Bei der Beantwortung der Frage, wie real Kraftfelder sind, haben Sie drei Näherungen kennengelernt, die von einem quantenchemischen Ansatz zu den vorgestellten Kraftfeldern führen: Die Born-Oppenheimer-Näherung, die Grundzustands-Näherung und die Lokalitäts-Näherung. Beschreiben Sie jeweils mit wenigen Worten, was diese Begriffe bedeuten.

12. [*8 Punkte*] Nehmen Sie an, wir messen das "Protein" X mit der Atomkomposition CN_2. Nehmen Sie weiter an, es gibt Y viele Ionen mit Ladung 1, die sie alle in einem Massenspektrometer messen. Wieviele Anteile der Y Ionen messen Sie bei welchen Massen?

    Hilfe: Rechnen Sie mit ganzen Zahlen für Isotopenmassen, $Prob(\ ^{12}C) = 0.989$, $Prob(\ ^{13}C) = 0.011$, $Prob(\ ^{14}N) = 0.9963$, $Prob(\ ^{15}N) = 0.0037$.