

Dritte Klausur zur Übung Algorithmische Bioinformatik

Lösung

Freie Universität Berlin, WS 2004/05

Utz J. Pape · Ben Rich · Dr. Stefan Röpcke · Prof. Dr. Martin Vingron

Name:

Matrikelnummer:

1. Gegeben ist die Sequenz 5'-ACGCGTGC-3'. Ein partieller Enzymverdau von S mit einem Restriktionsenzym, dass CG in der Mitte schneidet wird durchgeführt. Welche Teilsequenzen erwarten Sie? [2 Punkte]

AC, GC, GTGC, ACGC, GCGTGC, ACGCGTGC

Schreiben Sie die Multimenge auf. [1 Punkt]

{0, 2, 2, 4, 4, 6, 8}

2. Warum genügt es beim Skiena-Algorithmus in jedem Schritt das längste Teilstück an den Rand zu platzieren (Beweisidee) [1 Punkt]?

Würde man das Stück in die Mitte platzieren, so würde das bedeuten, dass es noch ein längeres Stück geben müsste (nämlich von der rechten/linken Seite bis zum rechten/linken Ende der Platzierung).

3. Gegeben $X = \{0, 5, 11, 13, 20\}$, $E = \{1, 1, 7, 8, 12\}$, was ist das Resultat des Funktionsaufrufs `placemax(X, E)` [2 Punkte]?

$y_{max} = 12$, $\delta(12, X) = \{12, 7, 1, 1, 8\}$ ist gleich E und damit sind wir fertig.

4. STS Content Mapping

In der Tabelle 1 sehen Sie, welche Probes (Spalten) zu welchen Klonen (Zeilen) bei einem STS Content Mapping hybridisiert. Da wir bei unserem Experiment leider keine Fehlerfreiheit annehmen können, müssen wir versuchen, die minimale Anzahl von consecutive ones zu finden.

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

Tabelle 1: Hybridisierungen

- (a) Reduzieren Sie das Problem auf das TSP Problem und malen Sie den entsprechenden Graphen auf. [2 Punkte]

Graph: siehe Abbildung 1. Distanzen: $SA = 3$, $SB = 2$, $SC = 2$, $SD = 2$, $AB = 3$, $AC = 3$, $AD = 1$, $BC = 2$, $BD = 2$, $CD = 4$.

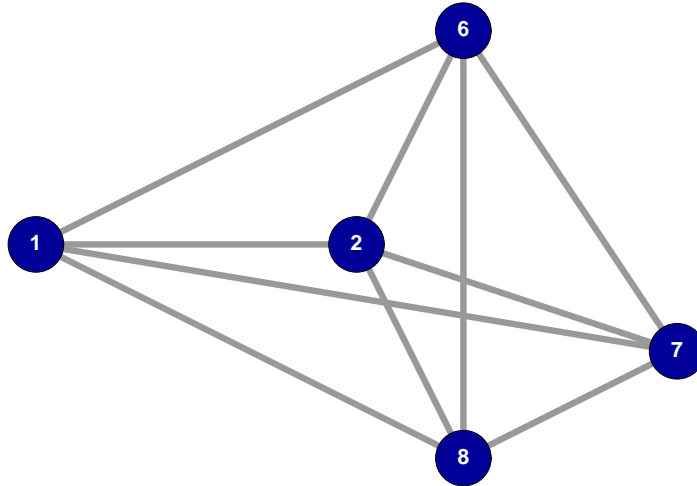


Abbildung 1: Graphen für TSP, wobei Knotenlabel 1=S, 2=A, 6=B, 7=C, 8=D.

- (b) Finden Sie einen minimalen Pfad in dem Graphen. [1 Punkt]
 Z.B. SDACB (und der Pfad im Graphen muss wieder zurück zu S gehen!)
- (c) Geben Sie das Layout der ursprünglichen Sequenz an. [1 Punkt]
 siehe Abbildung 2

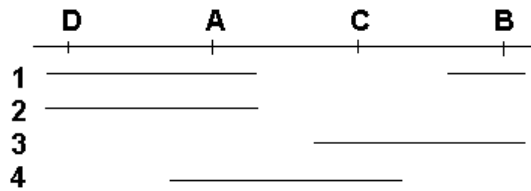


Abbildung 2: Layout

- (d) Welche Fehler können in der Hybridisierungstabelle enthalten sein? [1 Punkt]
- false positives
 - false negatives
 - chimeras

Beschreiben Sie einen davon genauer. [1 Punkt]

Z.B. false positive: Es hat keine Hybridisierung stattgefunden, aber trotzdem gibt es ein Signal.

5. Wieviele RNA-Sekundärstrukturen einer Sequenz der Länge $n = 5$ gibt es? Geben Sie auch die rekursive Formel an. [2 Punkte]

$r(0) = 1$ wird definiert. Zunächst überlegt man sich die leichten Fälle: $r(1) = 1$ da es nur ungepaart sein kann. $r(2) = 2$ entweder gepaart oder ungepaart. Um und etwas rechnen zu sparen, kann man auch noch $r(3) = 4$ 'sehen', da entweder die linken oder rechten beiden gepaart sind oder ganz links mit ganz rechts oder gar nichts! Nun die Rekursionsformel (man überlege sich, dass entweder die n -te Base ungepaart ist, dann hat man den Fall $n - 1$ oder aber das n -te sich mit dem k -ten paart, daher kommt die Summe:

$$r(n) = r(n-1) + \sum_{k=1}^{n-1} r(k-1)r(n-k-1)$$

$$r(4) = r(3) + r(0)r(2) + r(1)r(1) + r(2)r(0) = 9$$

$$r(5) = r(4) + r(0)r(3) + r(1)r(2) + r(2)r(1) + r(3)r(0) = 21$$

Die Lösung lautet also 21.

6. Stellen Sie die Rekursionsformel des Nussinov-Algorithmus auf. [2 Punkte]

$$f(i, j) = \max \left\{ \begin{array}{l} f(i+1, j) \\ f(i, j-1) \\ f(i+1, j-1) + \delta(i, j) \\ \max_{i < k < j} [f(i, k) + f(k+1, j)] \end{array} \right.$$

7. Beschreiben Sie kurz aber genau, wie das Backtracking funktioniert. [2 Punkte]

Das Backtracking beginnt links unten (oder rechts oben je nach Matrix-Aufstellung) in der Matrix, man schaut, wie man zu diesem Wert gekommen ist und führt dann von diesen Feldern weiter das Backtracking durch.

8. Was ist die Laufzeitkomplexität des Nussinov-Algorithmus? [2 Punkte]

$O(n^3)$ einmal halbe Matrix durchlaufen, und bei jedem Schritt noch mal eine Zeile bzw. Spalte.

9. Gegeben ist die Sequenz *acgguc*, vervollständigen Sie die letzte Zeile in der Nussinov Score Matrix: [2 Punkte]

		i			
	0				
	1	1			
j	1	1	0		
	2	1	0	0	
	2	2	1	1	0

10. Gegeben ist die gewöhnliche Differentialgleichung $y' = \log(y^2)$. Schreiben Sie den Ansatz des Euler-Verfahrens hin. [1 Punkt]

$$\hat{y}(t + \tau) = \hat{y}(t) + \tau \log(\hat{y}^2(t))$$

11. Die Energie eines harmonischen Oszillators lässt sich als Summe der kinetischen und der potentiellen Energie des Systems beschreiben. Mit welchem Ansatz zeigt man, dass der Energieerhaltungssatz in diesem System gilt? [1 Punkt]

$$\frac{\partial E}{\partial t} = 0$$

12. Bleibt die Gesamtenergie bei der numerischen Lösung des Hamilton Systems eines harmonischen Oszillators mit dem Euler-Verfahren konstant? [1 Punkt]

Nein!

13. Welche Informationen aus der PDB-Datei nutzt man, um Sekundärstrukturen vorherzusagen? [1 Punkt]

- den Typ des Atoms
- die dreidimensionalen Koordinaten der Atome (x,y,z)

14. Welche SCOP-Klassen gibt es? [2 Punkte]

- ALL ALPHA: mehr als 80% der AA in SSE haben den SSE-Typ H.
- ALL BETA: mehr als 80% der AA in SSE haben den SSE-Typ E.
- ALPHA/BETA:
 - Das Protein gehört nicht in die ersten beiden Klassen.
 - Der grösste Beta-Sheet (2-oder mehr Strands, die über H-Bonds in Kontakt sind), besteht in der Mehrzahl aus parallelen Strands, wobei in der Sequenz der SSE zwischen zwei Strands mindestens eine Helix liegen muss.
- ALPHA+BETA:
 - Das Protein gehört nicht in die obigen drei Klassen.
 - Das Protein gehört nicht in die obigen drei Klassen.
 - Der grösste Beta-Sheet, besteht in der Mehrzahl aus antiparallelen Strands.

Beschreiben Sie eine davon genauer. [1 Punkt]

siehe oben.