BURSA TEKNİK ÜNİVERSİTESİ
MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ

BİM0101 – Hesaplamalı Biyolojiye Giriş
Bitirme Sınavı

Ad&Soyad	**:** CEVAP ANAHTARI
Öğrenci Numarası	**:**

Akademik yıl : **2021-2022**

Dönem : **Bahar**

Tarih : **30 Haziran 2022 – 09:00**

Sınav süresi : **100 dakika**

Öğr. görevlisi : **Dr. Öğr. Üyesi Ergün GÜMÜŞ**

Soru	1	2	3	4	5	Toplam
Puan	15	25	20	20	20	100
Not						

KURALLAR

- Sınava başlamadan önce Ad&Soyad ve Öğrenci numarası alanlarını doldurunuz.
- Sınav öncesinde ve süresince sınav gözetmenlerinin tüm uyarılarına uymanız gerekmektedir.
- Sınav öncesinde cep telefonlarınızı KAPATINIZ!
- Soruları yanıtlamak için sadece sınav kâğıdınızı kullanmanız gerekmektedir. Yanıtlarınız açık ve okunaklı olmalıdır.
- Sınav boyunca masanızın üzerinde bulunabilecek malzemeler sadece sınav kâğıdınız, kalem ve silgidir.
- Sınav süresince herhangi bir nedenle birbirinizle konuşmak ve malzeme (silgi, kalem, kâğıt vb.) alışverişi yasaktır.
- Bu kuralların herhangi birine uymamak kopya çekmeye yönelik bir hareket olarak değerlendirilir ve ilgili makamlara bildirilir.

1) [15p] Aşağıdaki kavramları tanımlayınız.
a) Edit/Levenshtein mesafesi b) Minkowski mesafesi c) Transkriptom (Transcriptome)

a) Bir sekansı başka bir sekansa dönüştürmek için gerekli olan substitution ve indel sayısı.

b) Kitabımızın 271. sayfasında bahsettiğimiz metrik. x_i ve x_j, p boyutlu veri uzayındaki iki noktadır. a = 1 ise Manhattan, a = 2 ise Öklid mesafesi.

$$d_{ij} = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|^a \right]^{1/a}$$

c) Belirli bir fizyolojik durum altında bulunan bir hücrede gözlemlenen RNA türlerinin kümesine denir.

2) [25p] A, B, C, D, E isimli beş adet canlı türü ve her canlıyı ifade etmek için de X_1, X_2, X_3, X_4, X_5 şeklinde beş adet karakter olduğunu, canlı-karakter matrisinin de aşağıdaki gibi olduğunu kabul edelim.

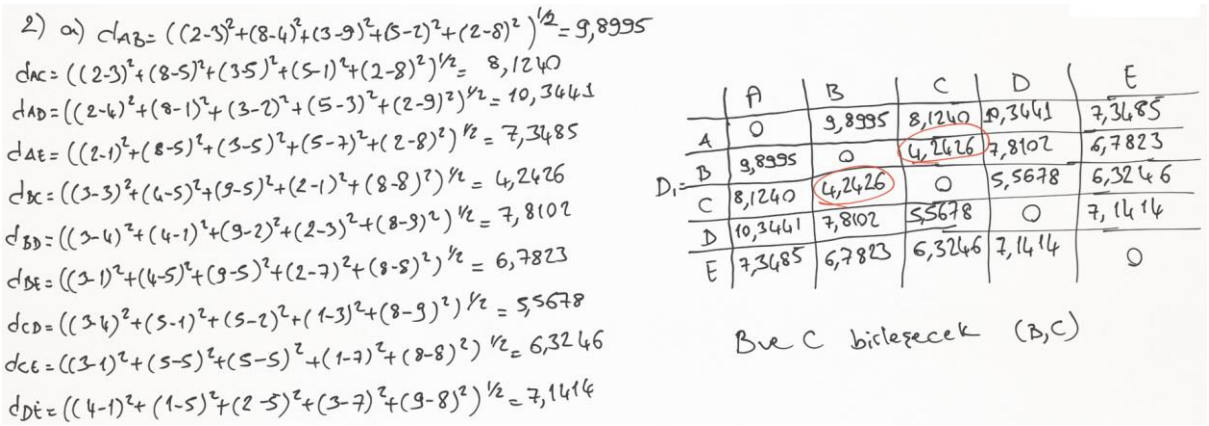
	X_1	**X_2**	**X_3**	**X_4**	**X_5**
A	2	8	3	5	2
B	3	4	9	2	8
C	3	5	5	1	8
D	4	1	2	3	9
E	1	5	5	7	8

Buna göre;

a) Canlı çiftleri arasındaki mesafeyi gösteren *D1* mesafe (distance) matrisini oluşturunuz. Mesafe hesabında öklid (euclidean) uzaklığını kullanınız [10p].

b) a şıkkında oluşturduğunuz *D1* mesafe matrisini kullanarak canlı türlerini aşağıdan yukarıya yaklaşımla kümeleyiniz. Kümeleme işleminin her adımında sadece birbirine en yakın olan iki kümeyi birleştirerek bir sonraki adımın mesafe matrisini (D2, D3, …) oluşturunuz. İki kümenin yakın olup olmadıklarını belirlerken bağlama (linkage) tekniği olarak tüm bağlama (complete linkage) kullanınız. Puan alabilmeniz için kümeleme işleminin her adımını sonucuyla beraber göstermeniz gerekmektedir [10p].

c) Yaptığınız kümeleme işlemine karşılık gelen dendrogramı çiziniz [5p].



b) complete linkage : iki küme arasındaki mesafe, kümelerin birbirine en uzak olan noktaları arasındaki mesafedir.

$D_2 =$

	A	(B,C)	D	E
A	0	9,8995	10,3441	7,3485
(B,C)	9,8995	0	7,8102	6,7823
D	10,3441	7,8102	0	7,1414
E	7,3485	6,7823	7,1414	0

(B,C) ile E birleşecek → (B,C,E)

$D_3 =$

	A	(B,C,E)	D
A	0	9,8995	10,3441
(B,C,E)	9,8995	0	7,8102
D	10,3441	7,8102	0

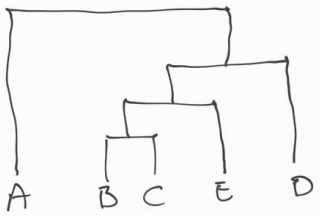
(B,C,E) ile D birleşecek → (B,C,D,E)

$D_4 =$

	A	(B,C,D,E)
A	0	10,3441
(B,C,D,E)	10,3441	0

(B,C,D,E) ile A birleşecek → (A,B,C,D,E)

c)



3) [20p] Aşağıdaki soruları cevaplayınız.

a) Standart sapma nedir? Formülünü yazıp formüldeki her ifadenin ne olduğunu açıklayınız [5p].

b) Korelasyon nedir? Pearson korelasyon katsayısını (r) hesaplamak için gereken denklemi yazıp formüldeki her ifadenin ne olduğunu açıklayınız [5p].

c) Aşağıda X ve Y isimli iki farklı gen ifade düzeyi (gene expression level) dağılımı görülmektedir. X ve Y dağılımlarının korelasyonunu (r_{XY}) hesaplayınız. Puan alabilmeniz için hesaplama adımlarını açık bir şekilde göstermeniz gerekmektedir [10p].

X = [60, 135, 176, 91, 67, 99], Y = [2422, 3077, 4824, 2810, 4349, 2241]

3) a)

$$std(X) = \sqrt{\frac{1}{N-1} \cdot \sum_{i=1}^{N} (x_i - \bar{X})^2}$$

N → dağılımın örnek sayısı; $\bar{X}$ → X dağılımının ortalaması; x_i → X dağılımının i. örneği

Standart sapma, bir X dağılımındaki noktaların dağılımın ortalamasından ortalama kaç birim saptığını gösteren bir ölçüttür.

b)

$$r_{xy} = \frac{\frac{1}{N-1} \cdot \sum_{i=1}^{N} (x_i - \bar{X}).(y_i - \bar{y})}{std(X).std(y)}$$

$\bar{X}$, $\bar{y}$ → dağılım ortalamaları; std(X), std(y) → dağılımların standart sapmaları (Kitapta 315. sayfa)

Korelasyon ölçüsü, değeri +1 ile -1 arasında değişen ve iki değişkendeki değerlerin artış/azalış açısından benzeyip benzemediğini gösteren bir ölçüttür.

c)

$\bar{X} = (60+135+176+91+67+99)/6 = 104,\overline{6}$

$\bar{y} = (2422+3077+4824+2810+4349+2241)/6 = 3287,\overline{16}$

$$std(x) = \left(\frac{1}{5} \cdot \left[(60-104,\overline{6})^2 + (135-104,\overline{6})^2 + (176-104,\overline{6})^2 + (91-104,\overline{6})^2 + (99-104,\overline{6})^2 \right] \right)^{1/2} = 43,912$$

$$std(y) = \left(\frac{1}{5} \cdot \left[(2422-3287,\overline{16})^2 + (3077-3287,\overline{16})^2 + (4824-3287,\overline{16})^2 + (2810-3287,\overline{16})^2 + (4349-3287,\overline{16})^2 + (2241-3287,\overline{16})^2 \right] \right)^{1/2}$$

$$= 1058,6956$$

$$r_{xy} = \frac{\frac{1}{5} \cdot \left[\begin{array}{l} (60-104,\overline{6}).(2422-3287,\overline{16}) + \\ (135-104,\overline{6}).(3077-3287,\overline{16}) + \\ (176-104,\overline{6}).(4824-3287,\overline{16}) + \\ (91-104,\overline{6}).(2810-3287,\overline{16}) + \\ (67-104,\overline{6}).(4349-3287,\overline{16}) + \\ (99-104,\overline{6}).(2241-3287,\overline{16}) \end{array} \right]}{43912 \times 1058,6956} = 0,4913$$

4) [20p] Bir canlının DNA'sının herhangi bir pozisyonundaki baz görülme olasılıkları şu şekilde olsun: P(A) = 0,23 P(C) = 0,29 P(G) = 0,25 P(T) = 0,23. Bu canlının DNA'sının herhangi bir bölgesindeki X=0 baz pozisyonunun 2 baz solu ve 2 baz sağını çevreleyen alandaki baz görülme olasılıkları ise şu şekilde verilmiş olsun:

	X=-2	**X=-1**	**X=0**	**X=+1**	**X=+2**
P(A)	0,30	0,08	0,38	0,30	0,22
P(C)	0,29	0,18	0,26	0,21	0,34
P(G)	0,17	0,40	0,03	0,26	0,08
P(T)	0,24	0,34	0,33	0,23	0,36

Buna göre;

a) X=-2 → X=+2 aralığındaki her baz pozisyonu için bağıl entropi değerlerini hesaplayınız [15p].

b) Canlının X=-2 → X=+2 aralığını ifade eden sekans logosunu oluşturunuz (Logoyu çizmeniz gerekmez, aralıktaki her baz poziyonu için o pozisyondaki baz logolarının yüksekliğini hesaplamanız yeterlidir) [5p].

4) a) $H = \sum_{a \in \{A,C,G,T\}} p_a . \log_2(p_a/q_a)$

X=-2 → $H = [0,3 . \log_2(0,3/0,23) + 0,29 . \log_2(0,29/0,29) + 0,17 . \log_2(0,17/0,25) + 0,24 . \log_2(0,24/0,23)] = 0,0351$

X=-1 → $H = [0,08 . \log_2(0,08/0,23) + 0,18 . \log_2(0,18/0,29) + 0,4 . \log_2(0,4/0,25) + 0,34 . \log_2(0,34/0,23)] = 0,2172$

X=0 → $H = [0,38 . \log_2(0,38/0,23) + 0,26 . \log_2(0,26/0,29) + 0,03 . \log_2(0,03/0,25) + 0,33 . \log_2(0,33/0,23)] = 0,3144$

X=1 → $H = [0,3 . \log_2(0,3/0,23) + 0,21 . \log_2(0,21/0,29) + 0,26 . \log_2(0,26/0,25) + 0,23 . \log_2(0,23/0,23)] = 0,0319$

X=2 → $H = [0,22 . \log_2(0,22/0,23) + 0,34 . \log_2(0,34/0,29) + 0,08 . \log_2(0,08/0,25) + 0,36 . \log_2(0,36/0,23)] = 0,165$

b) L yüksekliği ifade etmek üzere,

X=-2 → L(A) = 0,3 . 0,0351 = 0,0105
L(C) = 0,29 . 0,0351 = 0,0101
L(G) = 0,17 . 0,0351 = 0,0059
L(T) = 0,24 . 0,0351 = 0,0084

X=-1 → L(A) = 0,08 . 0,2172 = 0,0173
L(C) = 0,18 . 0,2172 = 0,039
L(G) = 0,4 . 0,2172 = 0,0868
L(T) = 0,34 . 0,2172 = 0,0738

X=0 → L(A) = 0,38 . 0,3144 = 0,1194
L(C) = 0,26 . 0,3144 = 0,0817
L(G) = 0,03 . 0,3144 = 0,0094
L(T) = 0,33 . 0,3144 = 0,1037

X=1 → L(A) = 0,3 . 0,0319 = 0,0095
L(C) = 0,21 . 0,0319 = 0,0066
L(G) = 0,26 . 0,0319 = 0,0082
L(T) = 0,23 . 0,0319 = 0,0073

X=2 → L(A) = 0,22 . 0,165 = 0,363
L(C) = 0,34 . 0,165 = 0,0561
L(G) = 0,08 . 0,165 = 0,0132
L(T) = 0,36 . 0,165 = 0,0594

5) [20p] G = 10^9 bp uzunluklu bir genomun sekanslandığını ve ortalama L = 2×10^3 bp uzunluklu, N = 3×10^5 adet klon (fragman) elde edildiğini kabul edelim. Buna göre aşağıdaki soruları cevaplayınız:

a) Coverage nasıl hesaplanır? Bu sekanslama işleminin sonunda coverage (c) kaç olur [5p]?

b) Genomdan rastgele seçilen bir bazın sekanslama işlemi sonunda en az 1 kez okunmuş olma ihtimali kaçtır?, hesaplayınız [5p].

c) Genomdan rastgele seçilen bir bazın sekanslama işlemi sonunda en az 1 kez okunmuş olma ihtimalinin %80 olabilmesi için en az kaç adet klona ihtiyaç vardır?, hesaplayınız [5p].

d) Klonları ucuca ekleyerek layout'lar oluşturmak istiyoruz. Bir klonun ucuyla diğer bir klonun ucunu eşleştirmek istersek bu eşleşmenin, en az klon uzunluğunun %15'i kadar olması gerektiğini varsayalım. Bu durumda hiçbir klon ile eşleşmeyen (layout oluşturmayan) kaç adet tekil (singleton) klon bulmayı bekleriz?, hesaplayınız [5p].

5) a) $c = \frac{N\,L}{G} = \frac{3.10^5 . 2.10^3}{10^9} = 0,6$

b) $f_c = 1 - (1 - L/G)^N = 1 - \left(1 - \frac{2.10^3}{10^9}\right)^{3.10^5} = 0,4511 \rightarrow \%45,11$

c) $0,8 = 1 - \underbrace{(1 - L/G)^N}_{0,2} \rightarrow (1 - L/G)^N = 0,2$

$\rightarrow \log_{(1-L/G)} (1-L/G)^N = \log_{(1-L/G)} 0,2$

$\rightarrow N = \log_{(1-L/G)} 0,2 = \frac{\log_{10} 0,2}{\log_{10}(1-L/G)} = 804718,15$

Yani en az 804719 adet klon lazım.

d) Singleton sayısı $= \underbrace{N . e^{-2(1-\theta).c}}_{\text{Kitapta 113-115 arasında}} = 3.10^5 . e^{-2.(0,85).0,6} \simeq 108178$