BURSA TEKNİK ÜNİVERSİTESİ
MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ

BİM0101 – Hesaplamalı Biyolojiye Giriş
Yıliçi Sınavı

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

Akademik yıl : **2021-2022**

Dönem : **Bahar**

Tarih : **27 Nisan 2022 – 12:00**

Sınav süresi : **75 dakika**

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

Soru	1	2	3	4	5	6	Toplam
Puan	25	15	20	15	10	15	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 kâğıdınızla beraber verilen kâğıtları 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.

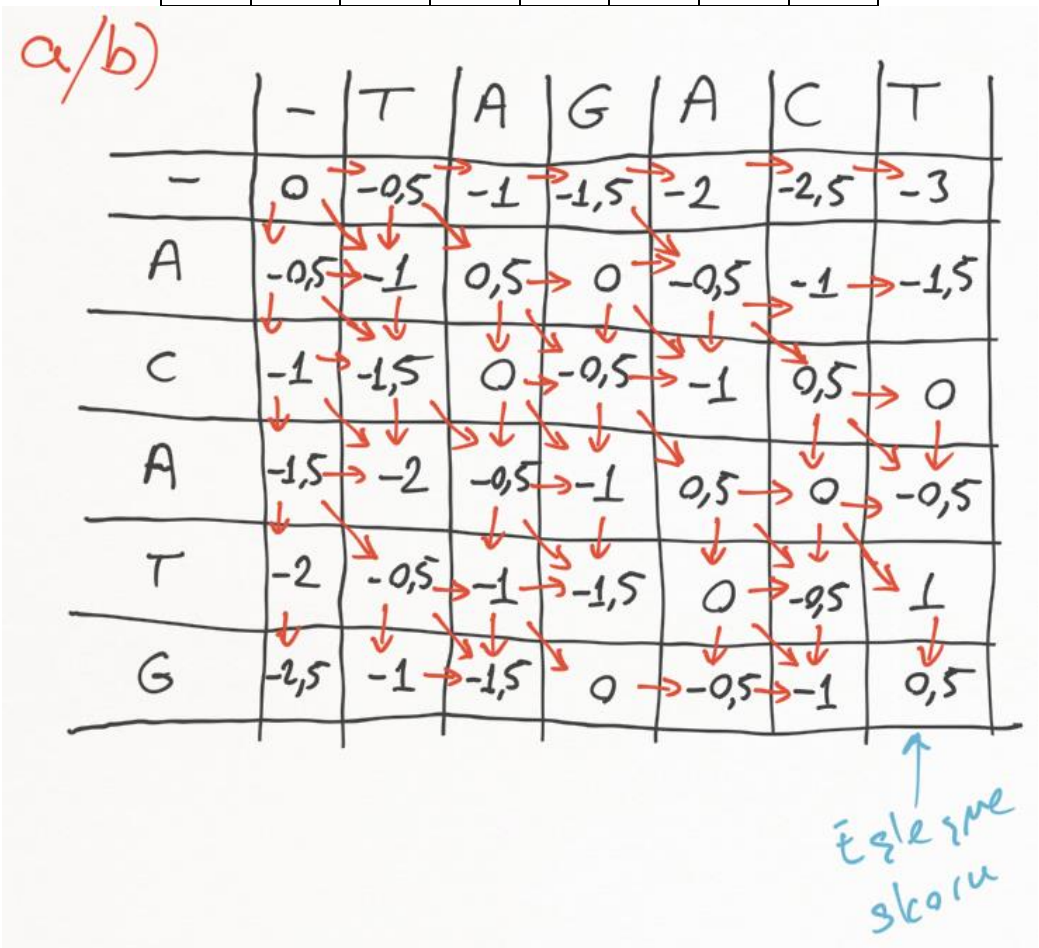
Sorular

1) [25p] TAGACT ve ACATG sekanslarını küresel hizalama (global alignment) yöntemi ile eşleştirmek istiyoruz. Buna göre aşağıdaki skor matrisini;

a) Her hücreye, o hücreye gelene kadarki skoru yazarak doldurunuz.
b) Her hücreye hangi yönden gelindiğini gösteren yön oklarını çiziniz.

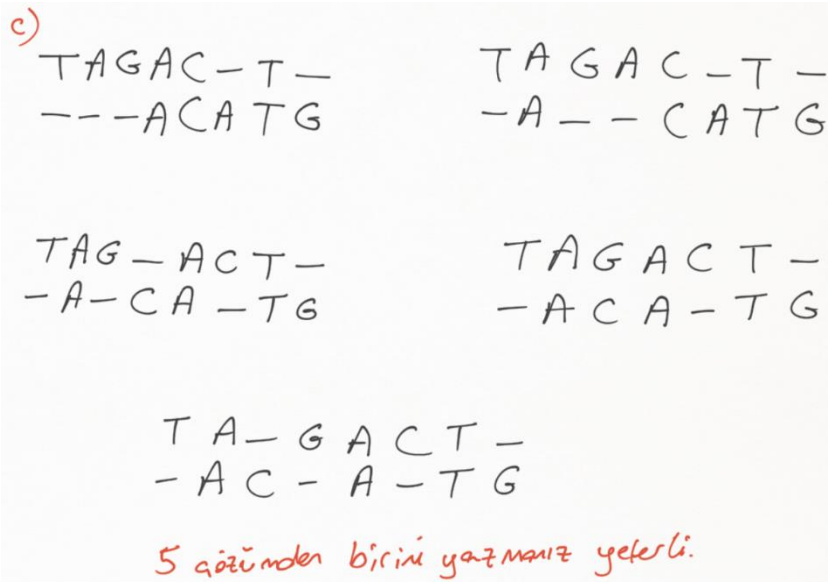
Hizalama işleminde aşağıdaki ödül/ceza puanlarını kullanınız.
Eşleşme (match): +1 Eşleşmeme (mismatch): -1 Boşluk (indel): -0,5

	-	T	A	G	A	C	T
-							
A							
C							
A							
T							
G							



Her hücre, skor ve kendisini gösteren oklarla beraber 0,5 puandır

c) Gerekli boşlukları (-) ekleyerek sekansların birbirleriyle nasıl eşleştiklerini aşağıya yazınız [4p].



2) [15p] Aşağıdaki soruları cevaplayınız.

a) n parametresi deneme sayısını, p parametresi de başarı olasılığını göstermek üzere Binom dağılımının olasılık fonksiyonunu yazınız [5p].

Bunu, kitabımızdaki Denklem 2.17'de görmüştük.

$$\mathbb{P}(N=j)=\binom{n}{j}p^j(1-p)^{n-j}, j=0,1,2,\ldots,n. \qquad (2.17)$$

b) Binom olasılık fonksiyonundan faydalanarak Poisson dağılımının olasılık fonksiyonunun nasıl türetileceğini adım adım gösteriniz [10p].

Bunu, kitabımızın 74. sayfasında görmüştük.

$$\mathbb{P}(N=j)=\frac{n!}{(n-j)!j!}p^j(1-p)^{n-j}$$

$$\mathbb{P}(N=j)=\frac{n(n-1)(n-2)\cdots(n-j+1)}{j!(1-p)^j}p^j(1-p)^n$$

n, toplam deneme sayısını göstermekte olup j ile kıyaslandığında çok büyük bir sayıdır. Bu nedenle,

$$n(n-1)(n-2)\cdots(n-j+1)\approx n^j$$

kabulü yapılabilir. Aynı zamanda p sayısının da küçük bir sayı olduğunu kabul edecek olursak

$$(1-p)^j\approx 1$$

olur. p'nin neden küçük bir sayı olacağını şu örneğe bakarak anlayabiliriz:

Örnek: n = 10^9 bp uzunluklu bir sekansta j = 30 tane AACTGA görme olasılığımız nedir?

Bu soruda, herhangi bir pozisyondaki bazların görülme olasılıklarının eşit ve 0,25 olduğunu kabul edersek p = P(AACTGA) = $0{,}25^6$ = 0,00024 çıkar.

Bu noktada birisi "madem p 0'a yakın bir sayı, neden $(1\text{-}p)^n$ ifadesine de 1 deyip geçmiyoruz?" diyebilir. Unutmayın, j sayısı $(1\text{-}p)^j$ ifadesini 0'a indiremeyecek kadar küçük bir sayı. O nedenle bu ifadeyi kabaca 1 olarak kabul ettik. Ancak, j'nin aksine n büyük bir sayı olduğu için $(1\text{-}p)^n$ ifadesinin değeri 1 kalamayabilir. 0'a yaklaşacaktır ama tam 0 olma garantisi de yoktur (Bu noktada p'nin değeri önemli).

Örneği bir kenara bırakıp çözüme devam edelim. $\lambda = np \rightarrow p = \lambda / n$ gibi bir kabulde bulunarak,

$$\mathbb{P}(N = j) \approx \frac{(np)^j}{j!}(1-p)^n = \frac{\lambda^j}{j!}\left(1 - \frac{\lambda}{n}\right)^n$$

ifadesi yazılabilir.

Aşağıdaki ifadenin doğruluğunu derste MATLAB üzerinden test etmiştik.

$$\lim_{n\to\infty}\left(1 - \frac{x}{n}\right)^n = \mathrm{e}^{-x}$$

Bu ifadeyi yukarıdaki denklemde yerine koyduğumuzda Poisson dağılımının olasılık fonksiyonu karşımıza çıkar.

$$\mathbb{P}(N = j) = \frac{\lambda^j}{j!}\mathrm{e}^{-\lambda}$$

3) [20p] 1000bp uzunluğundaki tek iplikçik (single strand) bir DNA dizilimi üzerinden (bu dizilime X sekansı diyelim), hesaplanan dimer (2-tuple) frekansları aşağıdaki matriste görülmektedir.

X_i \ X_{i+1}	**A**	**C**	**G**	**T**
A	0,0791	0,0711	0,0641	0,0651
C	0,0671	0,0611	0,0571	0,0721
G	0,0711	0,0631	0,0480	0,0430
T	0,0631	0,0621	0,0551	0,0581

Buna göre,
a) P(A), P(C), P(G) ve P(T) marjinal olasılıklarını hesaplayınız [4p].

Kitabımızın 53. sayfasındaki Denklem 2.26'dan da görülebileceği üzere bir grup bileşke olasılığın toplamı bize marjinal olasılığı verir. Bu matriste görülen olasılıklar da bileşke olasılıklardır.

$$\mathbb{P}(X_1 = j) = \sum_{i \in \mathcal{X}} \mathbb{P}(X_0 = i, X_1 = j)$$

Yani, X_{i+1} pozisyonunda sözgelimi C bazını görme ihtimali (P(C)), $P(X_i = A\ ,\ X_{i+1} = C) + P(X_i = C\ ,\ X_{i+1} = C) + P(X_i = G\ ,\ X_{i+1} = C) + P(X_i = T\ ,\ X_{i+1} = C) = P(AC) + P(CC) + P(GC) + P(TC)$ toplamına eşittir. Bu da, soruda verilen matrisin sütun toplamıyla elde edilebilir. Buna göre,

P(A) = 0,0791 + 0,0671 + 0,0711 + 0,0631 = 0,2804
P(C) = 0,0711 + 0,0611 + 0,0631 + 0,0621 = 0,2574
P(G) = 0,0641 + 0,0571 + 0,0480 + 0,0551 = 0,2243
P(T) = 0,0651 + 0,0721 + 0,0430 + 0,0581 = 0,2383

b) Dimer frekanslarını ve a şıkkında bulduğunuz marjinal olasılıkları kullanarak 1. seviyeden bir Markov zincirinin tahmini geçiş matrisini (estimated transition matrix) oluşturunuz [16p].

Markov zinciri, 4. sorudaki Bayes Teoremi ile doğrudan ilişkili olup tahmini geçiş matrisindeki her bir hücrenin değeri şu şekilde hesaplanır. Burada a ve b herhangi iki bazdır.

$$p_{ab} = P(X_{i+1} = b \mid X_i = a) = \frac{P(X_i = a\,, X_{i+1} = b)}{P(X_i = a)}$$

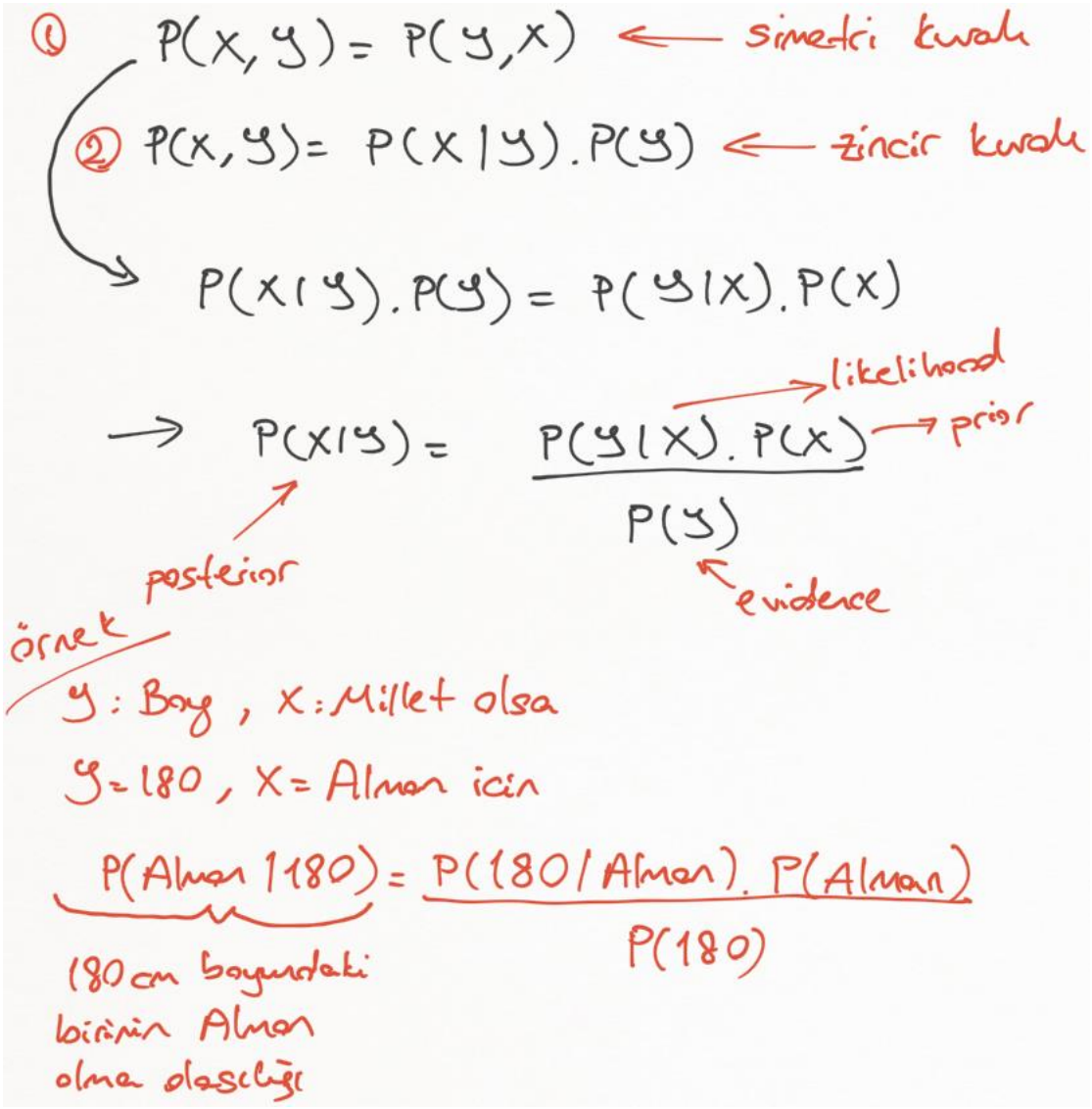
Bu hesaba dair bir örnek kitabımızın 54. sayfasında verilmiştir.

Bu durumda olası tüm a-b çiftleri için ETM'nin hücreleri şu tablodaki gibi hesaplanabilir:

a	b	İşlem	Sonuç
A	A	P(AA)/P(A)	0,0791/0,2804=0,2821
A	C	P(AC)/P(A)	0,0711/0,2804=0,2536
A	G	P(AG)/P(A)	0,0641/0,2804=0,2286
A	T	P(AT)/P(A)	0,0651/0,2804=0,2322
C	A	P(CA)/P(C)	0,0671/0,2574=0,2607
C	C	P(CC)/P(C)	0,0611/0,2574=0,2374
C	G	P(CG)/P(C)	0,0571/0,2574=0,2218
C	T	P(CT)/P(C)	0,0721/0,2574=0,2801
G	A	P(GA)/P(G)	0,0711/0,2243=0,3170
G	C	P(GC)/P(G)	0,0631/0,2243=0,2813
G	G	P(GG)/P(G)	0,0480/0,2243=0,2140
G	T	P(GT)/P(G)	0,0430/0,2243=0,1917
T	A	P(TA)/P(T)	0,0631/0,2383=0,2648
T	C	P(TC)/P(T)	0,0621/0,2383=0,2606
T	G	P(TG)/P(T)	0,0551/0,2383=0,2312
T	T	P(TT)/P(T)	0,0581/0,2383=0,2438

X_i \ X_{i+1}	**A**	**C**	**G**	**T**
A	0,2821	0,2536	0,2286	0,2322
C	0,2607	0,2374	0,2218	0,2801
G	0,3170	0,2813	0,2140	0,1917
T	0,2648	0,2606	0,2312	0,2438

4) [15p] Olasılık teorisindeki simetri ve zincir özelliklerini kullanarak Bayes Teoreminin nasıl türetildiğini adım adım gösteriniz. Prior (önsel olasılık), posterior (sonsal olasılık), likelihood (olabilirlik) ve evidence (delil) kavramlarını denklem üzerinde gösteriniz.



5) [10p] 6 yüzü olan hilesiz bir zarın,
a) Beklentisinin (expectation) nasıl hesaplandığını gösteriniz [5p].

$$E_{zar} = \sum_{i=1}^{6} P(i) \times i = \frac{1}{6}\times 1 + \frac{1}{6}\times 2 + \frac{1}{6}\times 3 + \frac{1}{6}\times 4 + \frac{1}{6}\times 5 + \frac{1}{6}\times 6 = 3{,}5$$

b) 10 atış sonucunda sırasıyla 3, 6, 6, 1, 4, 5, 2, 1, 5, 2 yüzlerinin geldiğini kabul edelim. Bu atışların varyansını hesaplayınız [5p].

$$X_{ort} = \frac{3+6+6+1+4+5+2+1+5+2}{10} = 3{,}5$$

$$Var_{10ATIŞ} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - X_{ort})^2$$
$$= \frac{1}{9} \times \{(3-3{,}5)^2 + (6-3{,}5)^2 + (6-3{,}5)^2 + (1-3{,}5)^2 + (4-3{,}5)^2 + (5-3{,}5)^2 + (2-3{,}5)^2 + (1-3{,}5)^2 + (5-3{,}5)^2 + (2-3{,}5)^2\} = 3{,}8333$$

veya

$$Var_{10ATIŞ} = \frac{1}{N}\sum_{i=1}^{N}(X_i - X_{ort})^2$$
$$= \frac{1}{10} \times \{(3-3{,}5)^2 + (6-3{,}5)^2 + (6-3{,}5)^2 + (1-3{,}5)^2 + (4-3{,}5)^2 + (5-3{,}5)^2 + (2-3{,}5)^2 + (1-3{,}5)^2 + (5-3{,}5)^2 + (2-3{,}5)^2\} = 3{,}45$$

6) [15p] Aşağıdaki kavramları kısaca açıklayınız (her biri 3 puan).

Sözlüğümüzdeki tanımlarıyla beraber aşağıda verilmiştir.

-Transkripsiyon: DNA'deki gen bölgelerinden mRNA (Messenger RNA) üretim işlemi.

-Gonozom: Eşey (cinsiyet) kromozomlarına verilen addır. X ve Y şeklinde iki türü vardır. Erkekler için XY, Kadınlar için XX çifti söz konusudur.

-Filogenetik ağaç: Canlı türlerinin genetik akrabalıklarını gösteren bir şema.

-Haplotip: Alellerin beraber oluşturduğu bir tiptir. Örneğin; yeşil gözlü, kısa boylu ve diyabet hastası bir bireyin çocuklarında ve torunlarında da bu özelliklerin beraber gözlemlendiğini düşünecek olursak bu durumda göz rengi, boy ve diyabetle ilgili özellikleri belirleyen genlerin alt soya beraber transfer edildiği ve bu gen alellerinin bir arada bir haplotip oluşturduğu söylenebilir. Haplotipler, sadece birlikte hareket eden genlerle değil, birlikte transfer edilen varyasyonlarla da (SNP'ler) oluşabilir.

-Histon: Kromozomun iskeletidir. Kromozom içerisindeki DNA, histona tutunmuş vaziyetteki "oktomer" adlı 8'li protein yapısına sarılıdır.