AŞAĞIDAKİ SORULARA VERDİĞİNİZ YANITLARDA ÇÖZÜMÜNÜZÜN HER ADIMINI AÇIK BİR ŞEKİLDE GÖSTERMELİ, TÜRETMENİZ GEREKEN YENİ BİR DENKLEM VARSA BU DENKLEMİ NASIL TÜRETTİĞİNİZİ AÇIKLAMALISINIZ. AKSİ TAKDİRDE ÇÖZÜMDEN PUAN ALAMAZSINIZ.

Sorular

1) [16p] Aşağıdaki tabloda bir genomik sekansın (L sekansı) içerisinde geçen trimer'lerin (3-tuple) frekansları verilmiştir. Söz konusu sekansın eş dağılım (identical distribution) özelliğini sağladığı ve herhangi bir bazın değerinin kendisinden önceki iki bazın değerlerine bağımlı olduğu (dependent) bilinmektedir.

	A	C	G	T	L_T
AA	0,027	0,019	0,021	0,022	
AC	0,020	0,013	0,015	0,014	
AG	0,022	0,014	0,018	0,020	
AT	0,023	0,015	0,018	0,018	
CA	0,018	0,013	0,015	0,016	
CC	0,012	0,009	0,016	0,011	
CG	0,015	0,010	0,013	0,013	
CT	0,015	0,010	0,012	0,012	
GA	0,022	0,015	0,018	0,018	
GC	0,015	0,010	0,012	0,012	
GG	0,019	0,013	0,014	0,014	
GT	0,019	0,012	0,016	0,015	
TA	0,022	0,015	0,019	0,019	
TC	0,015	0,011	0,012	0,013	
TG	0,018	0,012	0,015	0,016	
TT	0,018	0,013	0,015	0,014	
L_{T-2}, L_{T-1}					

T sembolü bir baz pozisyonunu göstermek üzere, tablodaki bir satırla sütunun kesişimi $P(L_{T-2} = i, L_{T-1} = j, L_T = k)$, $i,j,k \in \{A,C,G,T\}$ bileşke olasılığını vermektedir. Buna göre aşağıdaki soruları cevaplayınız:

a) [8p] $P(L_T = A) = ?$ $P(L_T = C) = ?$ $P(L_T = G) = ?$ $P(L_T = T) = ?$

b) [8p] Bayes teoremini kullanarak şu iki olasılığı hesaplayınız:
$P(L_T = C \mid L_{T-2} = T, L_{T-1} = G) = ?$ $P(L_T = A \mid L_{T-2} = A, L_{T-1} = T) = ?$

2) [10p] Merkezi limit teoremini şekil çizerek kısa/öz biçimde açıklayınız.

3) [15p] Bir X canlısının genomundaki baz dağılımının bağımsız ve eş dağılım (independent and identical distribution-iid) özelliğini sağladığını varsayalım. Bu canlının genomundaki herhangi bir baz pozisyonu için baz görülme olasılıkları P(A) = 0,35, P(C) = 0,20, P(G) = 0,25, P(T) = 0,20 şeklinde olsun. Buna göre, bu canlının genomundan rasgele seçilen 100bp uzunluklu bir sekansın içinde tam olarak 23 tane A, 32 tane C, 30 tane G ve 15 tane T bulunma olasılığını hesaplayınız (Sonucu verecek denklemi yazmanız da yeterlidir).

4) [20p] X dağılımının [a, b] sayı aralığından tekdüze (uniform) seçilen sayılardan oluştuğunu kabul edelim. Buna göre,

a) X dağılımının beklentisinin (a+b)/2 olduğunu [5p],
b) X dağılımının varyansının yaklaşık olarak $(b-a)^2/12$ olduğunu [15p] ispatlayınız.

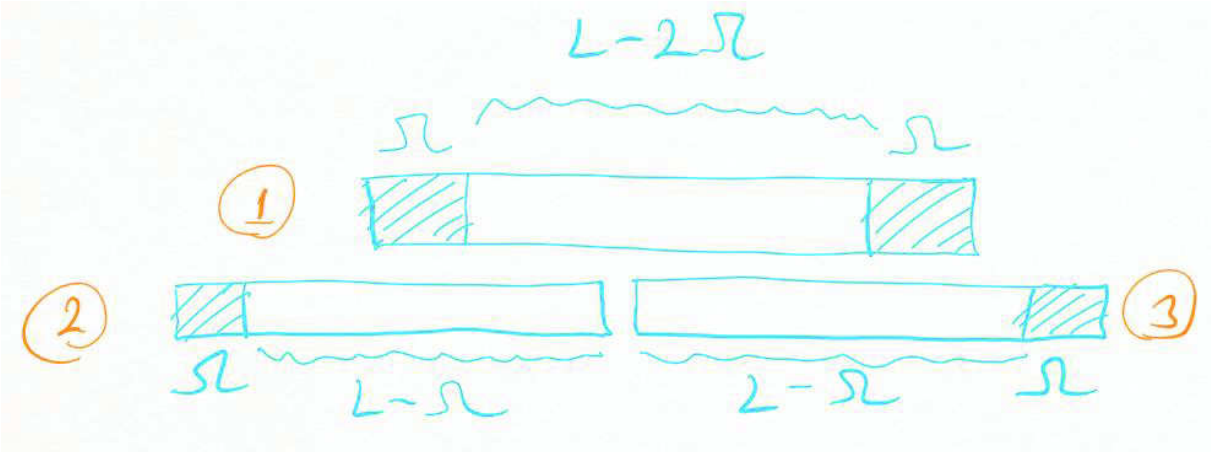
İpucu: [a, b] kapalı aralığındaki sayıların her birini $X_i = a+k$, $0 \leq k \leq b-a$ eşitliğine uygun ayrık sayılar olarak düşünebilirsiniz.

5) [24p] İlk soruda trimer frekans tablosu verilen sekansın bu sefer bağımsız ve eş dağılım (independent and identical distribution-iid) özelliğini sağladığını varsayalım. Herhangi bir baz pozisyonu için bazların görülme olasılıklarının da eşit olduğunu kabul edelim (P(A)=P(C)=P(G)=P(T)=0,25). Sekansımız 1000bp uzunluklu olsun. Buna göre aşağıdaki tabloda verilen 4 dimer için;

(i) Dimer'in frekans bilgisi üzerinden gözlem sayısını (gözlem) [8p],
(ii) Dimer'in iid modele göre gözlemlenmesi beklenen sayısı (beklenti) [4p],
(iii) Gözlem ve beklenti değerleri üzerinden X^2/c değerini [12p] hesaplayınız.

	Gözlem sayısı	**Beklenti**	**X^2/c**
AA			
AC			
AG			
AT			

6) [15p] Bu soru genom derleme (genome assembly) ile ilgilidir. Aşağıdaki şekilde üç klondan oluşan bir contig görülmektedir. 2 numaralı klonun sağ ucu ve 3 numaralı klonun sol ucu, 1 numaralı klonun L-2Ω alanına girmektedir (bu alanın daha ilerisine geçmemektedir). Aynı zamanda 2 numaralı klonun soluna ve 3 numaralı klonun da sağına başka hiçbir klon eklenememektedir. 3 klondan oluşan bu yapıya "üçlük" diyelim.



Buna göre, Poisson kestirimini kullanarak parametreleri aşağıda verilen deney ortamında,
(i) Herhangi bir contig'in bir üçlük olma olasılığını [12p],
(ii) N adet klondan kaç tane üçlük çıkmasının beklendiğini [3p] hesaplayınız.

Parametreler:
G = 30.000.000 N = 100.000 L = 1.000 Ω = 100

G: genomun uzunluğu, N: toplam klon sayısı, L: ortalama klon uzunluğu,
Ω: contig oluşması için minimum örtüşme miktarı