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.

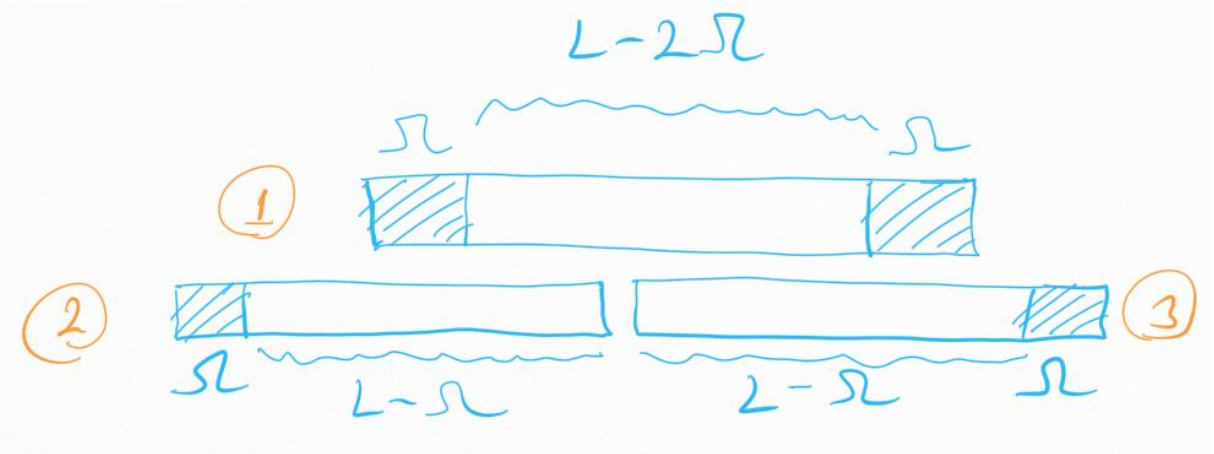
İpucu: [a, b] kapalı aralığındaki sayıların her birini $X_i = a+k$, $0 \leq k \leq b-a$ eşitliğine uygun ayrık sayılar olarak düşünebilirsiniz.

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

YANITLAR

1) a)

$$P(L_T = k) = \sum_{j \in \{A,C,G,T\}} \sum_{i \in \{A,C,G,T\}} P(L_T = k, L_{T-1} = j, L_{T-2} = i)$$

Yani, $P(L_T = k)$ marjinal olasılığını bulabilmek için matrisin sütun bazında toplanması gereklidir.

$P(L_T = A) = 0,3$ [2p]
$P(L_T = C) = 0,204$ [2p]
$P(L_T = G) = 0,249$ [2p]
$P(L_T = T) = 0,247$ [2p]

b)

$$P(L_T = k \mid L_{T-2} = i, L_{T-1} = j) = \frac{P(L_T = k, L_{T-1} = j, L_{T-2} = i)}{P(L_{T-2} = i, L_{T-1} = j)}$$

$$P(L_{T-2} = i, L_{T-1} = j) = \sum_{k \in \{A,C,G,T\}} P(L_{T-2} = i, L_{T-1} = j, L_T = k)$$

Yani, $P(L_{T-2} = i, L_{T-1} = j)$ bileşke olasılığını (ij dimerinin frekansını) bulabilmek için matriste satır bazında toplama yapmamız gerekli.

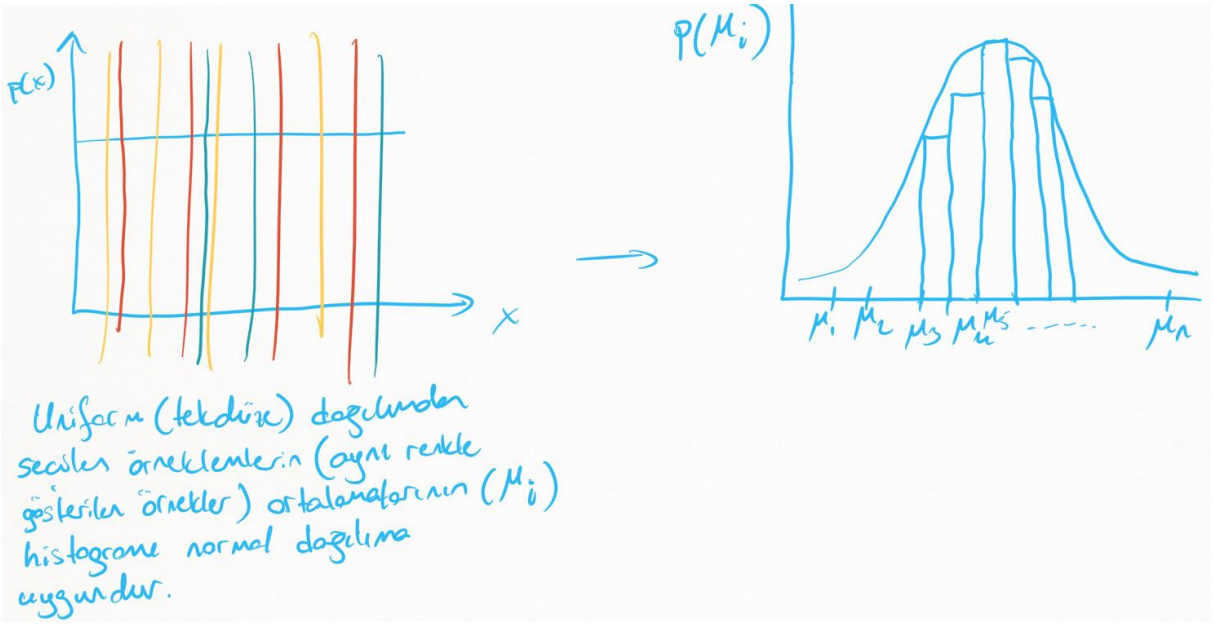
$$P(L_T = C \mid L_{T-2} = T, L_{T-1} = G) = \frac{P(L_T = C, L_{T-1} = G, L_{T-2} = T)}{P(L_{T-2} = T, L_{T-1} = G)} = \frac{0,012}{(0,018 + 0,012 + 0,015 + 0,016)} = 0,1967$$

[4p]

$$P(L_T = A \mid L_{T-2} = A, L_{T-1} = T) = \frac{P(L_T = A, L_{T-1} = T, L_{T-2} = A)}{P(L_{T-2} = A, L_{T-1} = T)} = \frac{0,023}{(0,023 + 0,015 + 0,018 + 0,018)} = 0,3108$$

[4p]

2) Merkezi limit teoremi, kaynak dağılımın türünden bağımsız olarak, kaynak dağılımdan seçilen bir dizi örneklemin ortalamalarının normal dağılıma uyduğunu gösterir. Şu şekil üzerinden (örnek şekilde uniform dağılım kullanılmıştır) gösterilebilir:



3) Soruyu çözmeden önce şu küçük örnek üzerinde çalışalım ve konuyu daha iyi anlayalım:

Örneğin P(A)=P(C)=P(G)=P(T)=0.25 olsaydı ve sadece 4bp uzunluğunda bir sekansımız olsaydı. Bu sekansta 1 tane A, 1 tane C, 1 tane G, 1 tane T olma olasılığı ne olurdu?

4bp uzunluğunda olan birbirinden farklı 4^4 = 256 adet sekans yazılabilir. Bu 256 adet sekansın sadece 4!=24 adedinde tüm bazlar birbirinden farklıdır (örneğimizde istendiği gibi). Bu durumda Örneğimizin cevabı 24/256 = 0,0937 olurdu.

Şimdi örneğimizi binom dağılımı olasılık fonksiyonu ile adım adım çözelim:

1. adım:
4 bazdan tam olarak 1 tanesinin A olma olasılığı:

P(#A=1) = $\binom{4}{1} x0{,}25^1 x0{,}75^3$=0,4219

2. adım:
A bazını eledikten sonra geriye 3 baz kaldı (C, G, T) ve bu bazların da görülme olasılıklarını güncellememiz gerekli. Üçünün de görülme olasılıkları birbirine eşit olduğuna göre → P(C)=0,33, P(G) = 0,33, P(T) = 0,33.

Kalan 3 bazdan tam olarak 1 tanesinin C olma olasılığı:

P(#C=1) = $\binom{3}{1} x0{,}33^1 x0{,}67^2$=0,4444

3. adım:
Geriye sadece 2 baz kaldı (G, T) ve bu bazların da görülme olasılıklarını güncellememiz gerekli. İkisinin de görülme olasılıkları birbirine eşit olduğuna göre -> P(G) = 0,5, P(T) = 0,5.

Kalan 2 bazdan tam olarak 1 tanesinin G olma olasılığı:

P(#G=1) = $\binom{2}{1} x0{,}5^1 x0{,}5^1$=0,5

4. adım:
Geriye kalan son bazın T olma olasılığı 1'e eşittir. Hesaplamaya gerek yoktur.

Şimdi bulduğumuz olasılıkları çarpalım:
P(#A=1) x P(#C=1) x P(#G=1) x P(#T=1) = 0,4219 x 0,4444 x 0,5 x 1 = 0,0937

Bu değer örneğimizin başlangıcında bulduğumuz olasılıkla aynı. Buradan anlamamız gereken şey binom olasılık formülünü kullanırken her seferinde bir senaryoyu (bir bazı) elememiz ve kalan senaryolar üzerinden, yeniden hesaplanan olasılıklar ile aynı adımları devam ettirmemiz gerektiğidir.

Benzerini bizim sorumuza uygulayacak olursak:

1. adım:

P(#A=23) = $\binom{100}{23} x0{,}35^{23} x0{,}65^{77}$= 0,0032

2. adım: Kalan elemanlar olan C, G ve T olasılıklarını güncelleyelim.

P(C)= 0,2 / (0,2+0,25+0,2) = 0,3077
P(G)= 0,25 / (0,2+0,25+0,2) = 0,3846
P(T)= 0,2 / (0,2+0,25+0,2)= 0,3077

P(#C=32) = $\binom{77}{32} x0{,}3077^{32} x0{,}6923^{45}$= 0,0125

3. adım: Kalan elemanlar olan G ve T olasılıklarını güncelleyelim.

P(G)= 0,3846 / (0,3846+0,3077) = 0,5555
P(T)= 0,3077 / (0,3846+0,3077)= 0,4445

P(#G=30) = $\binom{45}{30} x0{,}5555^{30} x0{,}4445^{15}$= 0,0395

4. adım:

P(T)=1

P(#T=15) = $\binom{15}{15} x1^{15} x0^{0}$= 1

P(#A=23)xP(#C=32)xP(#G=30)xP(#T=15) = 0,0032x0,0125x0,0395x1= $1{,}58x10^{-6}$

Sorudan tam puan alabilmek için adımlardaki denklemleri doğru kurmak yeterlidir. Sayısal sonuç şart değildir [15p]

4)

a) $E(X) = \sum_{X_i \in [a,b]} X_i \cdot p(X_i)$

$= \sum_{k=0}^{b-a} (a+k) \cdot \frac{1}{(b-a+1)} = \sum_{k=0}^{b-a} \frac{a}{(b-a+1)} + \frac{1}{(b-a+1)} \cdot \sum_{k=0}^{b-a} k$

$= (b-a+1) \cdot \frac{a}{(b-a+1)} + \frac{1}{(b-a+1)} \cdot \frac{(b-a) \cdot (b-a+1)}{2}$

$= \frac{2a+b-a}{2} = \frac{a+b}{2}$ [5p]

b) $Var(X) = \frac{1}{N} \cdot \sum_{X_i \in X} (X_i - \overline{X})^2$, $\overline{X} = E(X) = \frac{a+b}{2}$, $N = b-a+1$

$Var(X) = \frac{1}{(b-a+1)} \cdot \sum_{k=0}^{b-a} \left(a+k-\frac{(a+b)}{2}\right)^2 = \frac{1}{(b-a+1)} \cdot \sum_{k=0}^{b-a} \left(\frac{a+2k-b}{2}\right)^2 =$

$= \frac{1}{4 \cdot (b-a+1)} \cdot \sum_{k=0}^{b-a} \left(\underbrace{2k}_{P} + \underbrace{(a-b)}_{R}\right)^2 = \frac{1}{4 \cdot (b-a+1)} \cdot \sum_{k=0}^{b-a} (P^2 + R^2 + 2PR) =$

$= \frac{1}{4 \cdot (b-a+1)} \cdot \left[\underbrace{\sum_{k=0}^{b-a} 4k^2}_{D1} + \underbrace{\sum_{k=0}^{b-a} a^2 - 2ab + b^2}_{D2} + \underbrace{\sum_{k=0}^{b-a} 4(a-b) \cdot k}_{D3} \right]$

(D1) $= \frac{1}{4 \cdot (b-a+1)} \cdot 4 \cdot \sum_{k=0}^{b-a} k^2 = \frac{1}{(b-a+1)} \cdot \frac{(b-a) \cdot (b-a+1) \cdot (2b-2a+1)}{6}$

$= \frac{2b^2 - 2ab + b - 2ab + 2a^2 - a}{6} = \frac{2b^2 + 2a^2 - 4ab + b - a}{6}$

$$D2 = \frac{1}{4.(b-a+1)} \cdot \sum_{k=a}^{b} a^2-2ab+b^2 = \frac{1}{4.(b-a+1)} \cdot (b-a+1).(a^2-2ab+b^2)$$

$$= \frac{a^2-2ab+b^2}{4}$$

$$D3 = \frac{1}{4.(b-a+1)} \cdot 4.(a-b). \sum_{k=a}^{b} k = \frac{1}{(b-a+1)} \cdot (a-b). \frac{(b-a).(b-a+1)}{2}$$

$$= \frac{ab-a^2-b^2+ab}{2} = \frac{-a^2-b^2+2ab}{2}$$

$$Var(X) = D1 + D2 + D3 =$$

$$= \frac{2b^2+2a^2-4ab+b-a}{6} + \frac{a^2-2ab+b^2}{4} + \frac{-a^2-b^2+2ab}{2} =$$

(4) (3) (6)

$$= \frac{4b^2+4a^2-8ab+2b-2a+3a^2-6ab+3b^2-6a^2-6b^2+12ab}{12} =$$

$$= \frac{b^2+a^2-2ab+2b-2a}{12} = \frac{(b-a)^2+2(b-a)}{12} \cong \frac{(b-a)^2}{12}$$ [15]

5)

a) $P(L_{T-2}=j, L_{T-1}=i) = \sum_{k \in \{A,C,G,T\}} P(L_{T-2}=j, L_{T-1}=i, L_T=k)$

$P(AA) = 0,027+0,019+0,021+0,022 = 0,089$
$P(AC) = 0,020+0,013+0,015+0,014 = 0,062$
$P(AG) = 0,022+0,014+0,018+0,020 = 0,074$
$P(AT) = 0,023+0,015+0,018+0,018 = 0,074$

} iid modele göre frekanslar

1000 bp uzunluğundaki sekansta 999 dimer vardır.

$\#AA = 999 \times 0,089 = 89$ [2p]
$\#AC = 999 \times 0,062 = 62$ [2p]
$\#AG = 999 \times 0,074 = 74$ [2p]
$\#AT = 999 \times 0,074 = 74$ [2p]

b) Herhangi bir baz pozisyonu için bazların görülme olasılıklarının eşit olduğu varsayımına göre, beklentimiz, her dimerin görülme olasılığının 1/16 olması yönündedir.

#AA = #AC = #AG = #AT = 999/16 = 63 [4p]

c)

c) $\chi^2 = \frac{(O-E)^2}{E}$

$\chi^2_{AA} = (89-63)^2/63 = 10,7301$

$\chi^2_{AC} = (62-63)^2/63 = 0,0158$

$\chi^2_{AG} = \chi^2_{AT} = (74-63)^2/63 = 1,9206$

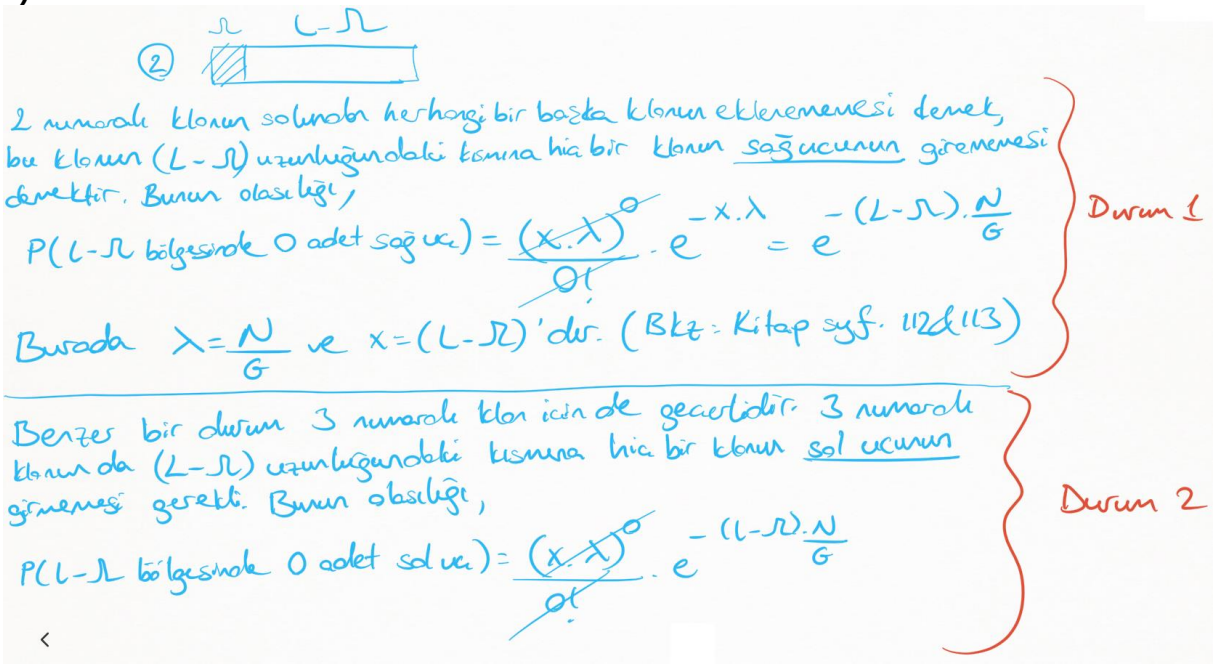
$C_{AA} = 1 + 2\times0,25 - 3\times(0,25)^2 = 1,3125$

$C_{AC} = C_{AG} = C_{AT} = 1 - 3\times(0,25)^2 = 0,8125$

$\frac{\chi^2_{AA}}{C_{AA}} = \frac{10,7301}{1,3125} = 8,1753$ [3p], $\frac{\chi^2_{AC}}{C_{AC}} = \frac{0,0158}{0,8125} = 0,0194$ [3p], $\frac{\chi^2_{AG}}{C_{AG}} = \frac{\chi^2_{AT}}{C_{AT}} = \frac{1,9206}{0,8125} = 2,3638$ [6p]

	Gözlem sayısı	Beklenti	X^2/c
AA	89	63	8,1753
AC	62	63	0,0194
AG	74	63	2,3638
AT	74	63	2,3638

6)



2 numaralı kromozomun solunda herhangi bir başka kromozom eklenmemesi demek, bu kromozom $(L-\ell)$ uzunluğundaki kısmına hiç bir kromozom sağ ucunun girmemesi demektir. Bunun olasılığı,

$P(L-\ell$ bölgesinde 0 adet sağ uç$) = \frac{(x.\lambda)^0}{0!} . e^{-x.\lambda} = e^{-(L-\ell).\frac{N}{G}}$ Durum 1

Burada $\lambda = \frac{N}{G}$ ve $x = (L-\ell)$'dir. (Bkz: Kitap syf. 112&113)

Benzer bir durum 3 numaralı kromozom için de geçerlidir. 3 numaralı kromozomda $(L-\ell)$ uzunluğundaki kısmına hiç bir kromozom sol ucunun girmemesi gerekli. Bunun olasılığı,

$P(L-\ell$ bölgesinde 0 adet sol uç$) = \frac{(x.\lambda)^0}{0!} . e^{-(L-\ell).\frac{N}{G}}$ Durum 2

L uzunluğundaki kablo $L-2\Omega$ uzunluğundaki parçasında tam olarak 2 uç görülmeliktir.

$P(L-2\Omega \text{ bölgesinde 2 uç}) = \frac{(x.\lambda)^2}{2!} \cdot e^{-x\lambda} = \frac{1}{2} \cdot \left((L-2\Omega).\frac{N}{G}\right)^2 \cdot e^{-(L-2\Omega).\frac{N}{G}}$ } Durum 3

3 durumun da aynı anda gerçekleşmesi gerekli.

$P(\text{üçlük}) = P(\text{Durum1}) \times P(\text{Durum2}) \times P(\text{Durum3})$

$= \frac{1}{2} \cdot \left((L-2\Omega).\frac{N}{G}\right)^2 \cdot e^{-\frac{N}{G}.(3L-4\Omega)}$ [9p]

i) $P(\text{üçlük}) = \frac{1}{2} \cdot \left((1000-2\times100).\frac{100000}{30000000}\right)^2 \cdot e^{-\frac{100000}{30000000}.(3\times1000-4\times100)}$

$= 6,1238 \times 10^{-4}$ [3p]

ii) $100000 \times 6,1238 \times 10^{-4} \cong 61$ adet üçlük [3p]