BURSA TEKNİK ÜNİVERSİTESİ
MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ

BİM0101-Hesaplamalı Biyolojiye Giriş
Yıliçi Sınavı

Ad&Soyad	**:**
Öğrenci Numarası	**:**

Akademik yıl **: 2020-2021**

Dönem **: Bahar**

Tarih **: 18 Nisan 2021 – 17:00**

Sınav süresi **: 90 dakika + 15 dakika yükleme süresi**

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

KURALLAR

- Çözümlerinizi A4 kağıtlar üzerine kendi el yazınız ile yapınız. Sınav sonunda bu kağıtları taratınız ya da okunabilecek bir şekilde fotoğraflarını çekerek "ogrencino.rar" şeklinde isimlendirilmiş bir dosya içerisinde ekampus sitesindeki sayfamıza yükleyiniz.
- Ekampus yükleme limiti 10MB'dır. Bu nedenle fotoğrafları yüksek çözünürlükte çekmeyiniz. Aksi takdirde siteye yükleyemeyeceksiniz.
- Sınav süresi sonunda herhangi bir nedenle cevaplarını ekampus sayfamıza yükleyemeyen öğrenciler sınava girmemiş sayılacaktır. E-mail ile atılan çözümler kabul edilmeyecektir. Bu nedenle sınav sürenizin 90 dakika olduğunu kabul edip buna göre davranın.
- Öğrencilerin sınav süresince herhangi biriyle iletişime geçmesi kopya işlemi olarak değerlendirilir. Sınav kağıdında, kopya olarak değerlendirilebilecek birebir aynı doğru çözüm ya da birebir aynı hata olan öğrencilerin sınavları iptal edilerek sınav kağıtlarıyla beraber dekanlığa bildirileceklerdir.

1) (1 puan) Öğrenci numaranızı yazınız.

2) (16 puan) Aşağıdaki terimleri kısaca açıklayınız. (Her biri 2 puan)

i- Diploid canlı
ii- Otozom
iii- Alel
iv- Transkripsiyon
v- Translasyon
vi- Restriksiyon endonükleaz
vii- İntron
viii- Homolog Rekombinasyon

3) (21 puan) DNA'nın kopyalanması sürecince DNA diziliminde bazı değişiklikler meydana gelebilmektedir. Bu değişiklikleri maddeler halinde yazınız ve örnek dizilimler üzerinde anlatınız.

4) (25 puan) Aşağıdaki koddaki 1234 sayısını öğrenci numaranız ile değiştirerek https://octave-online.net/ sitesindeki en alttaki boş satıra yapıştırıp çalıştırınız. Kod 1000bp uzunluğunda rasgele bir DNA dizisi üretecek ve bu dizi içerisindeki 2-gram (dimer) ve 1-gram (baz) sayısını hesaplayarak ekrana yazacaktır. Kod MATLAB'de çalışmaz, denemeyiniz.

```
rand('seed', 1234);
bazlar = 'ACGT';
sekans = bazlar(randi(4, 1, 1000));
dimerler = ['AA';'AC';'AG';'AT';'CA';'CC';'CG';'CT';'GA';'GC';'GG';'GT';'TA';'TC';'TG';'TT'];
dimer_sayi = zeros(1,16);
for i = 1:16
  temp = strfind(sekans,dimerler(i,:));
  dimer_sayi(i) = length(temp);
  disp([dimerler(i,:) ': ' num2str(dimer_sayi(i))]);
end

fprintf('\n\n\n');

for i = 1:4
   temp = strfind(sekans, bazlar(i));
   disp([bazlar(i) ': ' num2str(length(temp))])
end
```

Ekranda görünen 2-gram (dimer) ve 1-gram (baz) sayılarını cevap kağıdınıza yazınız ve ardından aşağıdakileri yapınız:

a) 1-gram (baz) sayılarını kullanarak p(A), p(C), p(G), p(T) olasılıklarını hesaplayınız. (5p)

b) 2-gram (dimer) frekanslarını hesaplayarak kitaptaki (2.27) denklemindekine benzer bir matris formatında yazınız. (5p)

c) b şıkkında hesapladığınız frekansları kullanarak tahmini geçiş matrisini (estimated transition matrix) hesaplayıp yazınız. Bu matris her 2-gram için şartsal olasılık (mesela $P(X_t = A \mid X_{t-1} = A) = ?$ gibi) hesabıyla bulunabilir. Bu şıktan puan alabilmeniz için 16 şartsal olasılığın tek tek nasıl hesaplandığını göstermeniz gerekmektedir. (15p)

5) (15 puan) Aşağıdaki koddaki 1234 sayısını öğrenci numaranız ile değiştirerek https://octave-online.net/ sitesindeki en alttaki boş satıra yapıştırıp çalıştırınız. Kod, öğrenci numaranıza göre 10 aminoasitten oluşan bir protein dizisi ve her amino asiti kodlamak için kullanılan kodonu raporlayacaktır. Kod MATLAB'de çalışmaz, denemeyiniz.

```
rand('seed', 1234);
AA = 'ARNDBCQEZGHILKMFPSTWYV';
kodonlar={ {'GCT','GCC','GCA','GCG'};
           {'CGT','CGC','CGA','CGG','AGA','AGG'};
           {'AAT','AAC'};
           {'GAT','GAC'};
           {'AAT','AAC','GAT','GAC'};
           {'TGT','TGC'};
           {'CAA','CAG'};
           {'GAA','GAG'};
           {'CAA','CAG','GAA','GAG'};
           {'GGT','GGC','GGA','GGG'};
           {'CAT','CAC'};
           {'ATT','ATC','ATA'};
           {'CTT','CTC','CTA','CTG','TTA','TTG'};
           {'AAA','AAG'};
           {'ATG'};
           {'TTT','TTC'};
           {'CCT','CCC','CCA','CCG'};
           {'TCT','TCC','TCA','TCG','AGT','AGC'};
           {'ACT','ACC','ACA','ACG'};
           {'TGG'};
           {'TAT','TAC'};
           {'GTT','GTC','GTA','GTG'};
           };
r = randperm(length(AA));
sel = AA(r(1:10));
fprintf('Kullanacaginiz dizilim: %s\n', sel);

for i=1:10
    fprintf('%s: %s\n',sel(i),kodonlar{r(i)}{randi(length(kodonlar{r(i)}))});
end
```

Kodun ürettiği protein dizisine, diziyi oluşturan kodonlara ve aşağıda verilen kodon kullanım oranlarına göre CAI (Codon Adaptation Index) değerini hesaplayınız. Sorudan puan alabilmeniz için tüm denklemi açık bir şekilde yazmanız gerekmektedir.

Amino asit	Kodlamada kullanılabilecek kodonlar	Kodon frekansları (Kodon sırasına göre)
A	GCT ; GCC ; GCA ; GCG	0.2924 ; 0.1541 ; 0.4897 ; 0.0637
R	CGT ; CGC ; CGA ; CGG ; AGA ; AGG	0.0650 ; 0.2230 ; 0.2182 ; 0.2425 ; 0.1560 ; 0.0954
N	AAT ; AAC	0.5452 ; 0.4548
D	GAT ; GAC	0.3502 ; 0.6498
B	AAT ; AAC ; GAT ; GAC	0.1314 ; 0.4292 ; 0.2406 ; 0.1988
C	TGT ; TGC	0.5027 ; 0.4973
Q	CAA ; CAG	0.5803 ; 0.4197
E	GAA ; GAG	0.6696 ; 0.3304
Z	CAA ; CAG ; GAA ; GAG	0.0541 ; 0.2914 ; 0.3253 ; 0.3292
G	GGT ; GGC ; GGA ; GGG	0.2448 ; 0.2771 ; 0.1681 ; 0.3100
H	CAT ; CAC	0.6761 ; 0.3239
I	ATT ; ATC ; ATA	0.3191 ; 0.2101 ; 0.4708
L	CTT ; CTC ; CTA ; CTG ; TTA ; TTG	0.1380 ; 0.2230 ; 0.1776 ; 0.0966 ; 0.3637 ; 0.0012
K	AAA ; AAG	0.8431 ; 0.1569
M	ATG	1
F	TTT ; TTC	0.7826 ; 0.2174
P	CCT ; CCC ; CCA ; CCG	0.3359 ; 0.0256 ; 0.3383 ; 0.3001
S	TCT ; TCC ; TCA ; TCG ; AGT ; AGC	0.0149 ; 0.2459 ; 0.2194 ; 0.1755 ; 0.1440 ; 0.2002
T	ACT ; ACC ; ACA ; ACG	0.0473 ; 0.3793 ; 0.4368 ; 0.1366
W	TGG	1
Y	TAT ; TAC	0.1857 ; 0.8143
V	GTT ; GTC ; GTA ; GTG	0.1950 ; 0.1094 ; 0.4393 ; 0.2563

6) (15 puan) Aşağıdaki koddaki 1234 sayısını öğrenci numaranız ile değiştirerek https://octave-online.net/ sitesindeki en alttaki boş satıra yapıştırıp çalıştırınız. Kod, öğrenci numaranıza göre biri 5 baz, diğeri 4 baz uzunluğunda iki DNA sekansı üretip ekrana yazacaktır. Kod MATLAB'de çalışmaz, denemeyiniz.

```
rand('seed', 1234);
bazlar = 'ACGT';
sekans1 = bazlar(randi(4,1,5));
sekans2 = bazlar(randi(4,1,4));
fprintf('Sekans1: %s\n', sekans1);
fprintf('Sekans2: %s\n', sekans2);
```

Buna göre Sekans1 ve Sekans2'yi küresel hizalama (global alignment) tekniğini kullanarak hizalayınız. Bunun için bir hizalama matrisi oluşturup tüm hücrelerini sayılarla doldurunuz. İstisnasız her hücrenin üzerine, o hücreye hangi hücreden geldiğinizi gösteren yön oklarını ekleyiniz. Hizalamada aşağıdaki ödül/ceza değerlerini kullanınız:

Eşleşme (Match): +2
Eşleşmeme (Mismatch): -1
Indel (gap): -0.5
Indel (gap) uzatma: -0.5

Bu sorunun çözümünde kağıdınıza çizmeniz/yazmanız istenenler:

a) Hizalama matrisi (Kitaptaki Computational Example 6.2'dekine benzer şekilde),

b) Hizalama sonucu,

c) Hizalama skoru.

7) (7 puan) Bir DNA diziliminde herhangi bir baz pozisyonunda C bazını görme ihtimali p(C) = 0.41'dir. Buna göre 30bp uzunluğundaki bir DNA diziliminde tam olarak 17 tane C bazı görme ihtimali nedir? Hesaplama adımlarınızı gösteriniz.

Not: https://octave-online.net/ sitesi çalışmazsa aşağıdaki alternatif siteleri de deneyebilirsiniz:

https://www.jdoodle.com/execute-octave-matlab-online/
https://rextester.com/l/octave_online_compiler
https://ideone.com/ (Alt kısımda dili Java'dan Octave'a çevirmeniz gerekir.)