

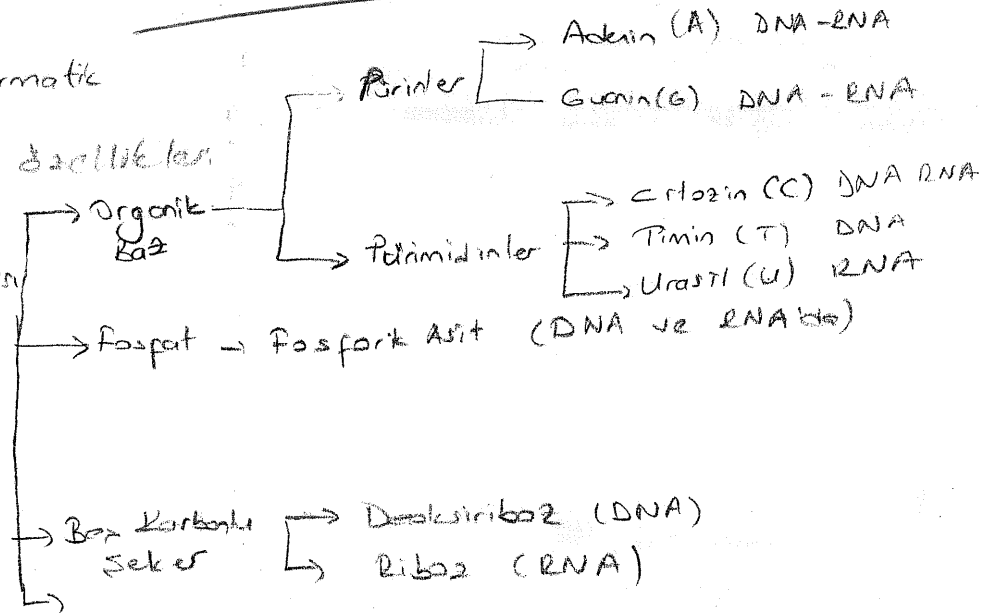
BIYOBİFORMATİKBiyoinformatik :

Biyolojik veritabanları + Informatik

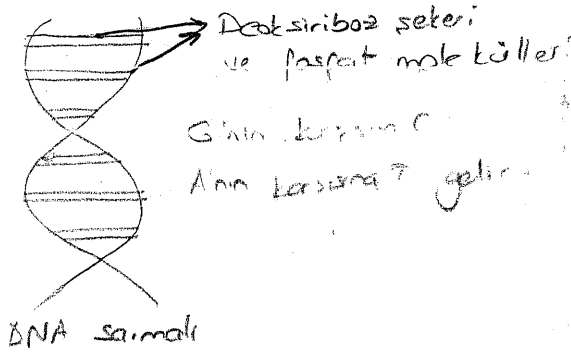
Diğer veritabanlarından farklı özellikler var.

Nükleik asitlerin Genel Yapısı

Nükleik Asitler → Nükleotid (DNA ve RNA)



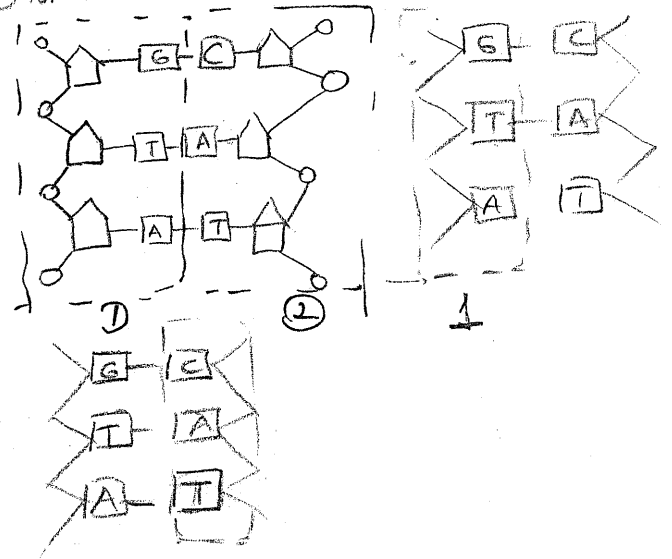
RNA'lar bir zincirden, DNA iki zincirden oluşur.

Deoksiriboz Nükleik Asit (DNA)Human Genome Projesi

9/97'si belirlenmiş.
 9/02'sinin aynısı yapılmış.

DNA'nın Görevleri

1. Kendini eşleyerek üremeyi ve kalıtsal bilginin aktarılması
2. Protein sentezi.

DNA'nın eşlenmesi (Replikasyon = Duplikasyon)

Tüm DNA örneklerinde;

 $A + G = T + C$ 'dir.

Tüm DNA örneklerinde;

Adenin miktarı Timin miktarına

Guanin miktarı, sitosin miktarına eşittir.

Firma, satış yapan firma.

Kitap, aygıt, kâğıt, kirtasiye,

Maximum 3 kişi ; firma başına

Güvenlik, ödeme, hukuki süreç.

Proje. %70, vizyöle kadar rapor.

Ödeme, sona pas hesabı, firma banka-
cılık, bankadan bilgileri alacak. (Olursa
iyi olur.)

Stok programı, ticari program.

Bu programlardan ~~iki~~ ürün bilgilerini
getirme, iletişim bilgileri.

E-ticaret hazır paketleri

+
Stok programı ile entegre olması

LIBONÜKLEİK ASİT (RNA)

(2)

1) mRNA (mesaj RNA)

Sentezlenecek proteine ilgili bilgiyi (sifreyi) DNA'dan alır. DNA'dan

mRNA'nın sentezlenmesine transkripsiyon (yazılma) denir.

AUG ACC ACG mRNA
Kodon Kodon Kodon
Aminoasit

2) tRNA (Taşıyıcı RNA)

Tazima işi yaparlar. tRNA'nın sifreyi okuyup uygun olanları getirmesi olur.

3) rRNA (Ribozomal RNA)

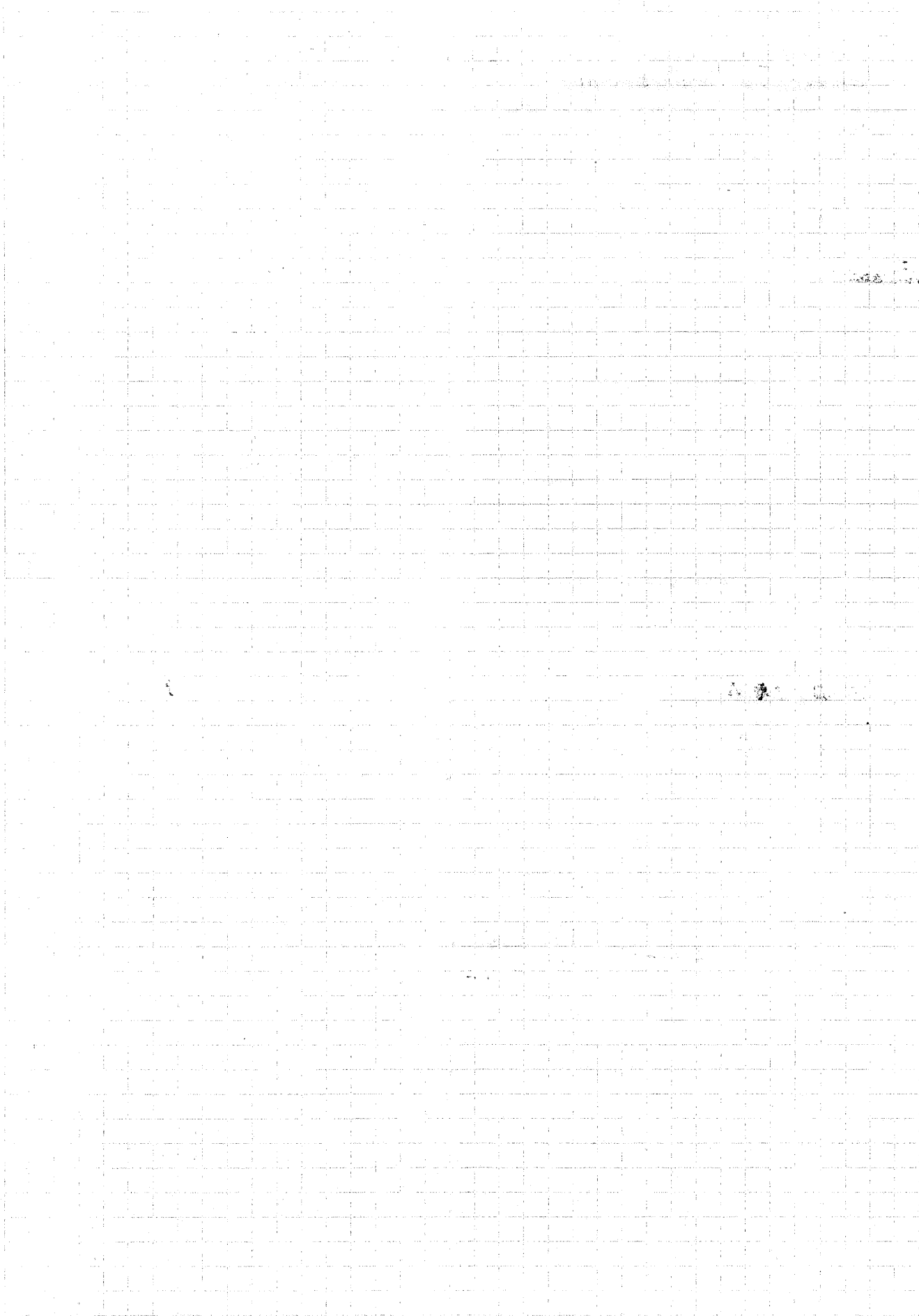
GENETİK SİFRE VE PROTEİN SENTEZİ

DNA ve RNA'da dörder harf bulunmakta ve kural olarak her kelime 3 harften oluşmaktadır. O halde DNA ve RNA dörder harf üzer üzer kombinasyon yapacak olursa $4^3 = 64$ farklı kodon oluşur.

İKİNCİ BAĞ. SİRA			
3 S R R A A S I A S I G	U C A G	U C A G	4
U	UUU UCU	UUC UCC	4
C	UUA UCA	UUG UCG	4
A			
S			
I			
A			
S			
I			
G			

$$4 \times 4 \times 4 = 64$$

Sifrelerin aminoasitler



(3)

Biyoinformatik :

Global Hizalama Kuralları :

Örn
S: ATTATCT
T: TTTCTA

Benzersizlik skor = +2

Benzersizlik skor = 0

Bosluk cezası = -1

T →

S ↓	0	-	T	T	T	C	T	A
-	0	-1	-2	-3	-4	-5	-6	
A	-1	0	-1	-2	-3	-4	-3	
T	-2	-1	0	-1	-2	-3	-2	
T	-3							
A	-4							
T	-5							
C	-6							
T	-7							7

A T T A T C T
x T T x T C T A

|| okunur
ayrılmaz.

- İlk hücre her zaman sıfır olur.
- (i-1, j-1) + Benzersizlik ya da Benzersizlik
- (j-1, j), gap (bosluk cezası)
- (i, j-1), gap (bosluk cezası)
- En köşeden başlayıp sıfıra doğru gitmeye çalışırız. Sıfırda olur, yakınında olur.
- Yatay veya dikey olması durumunda olun yöründe okun yerine boşluk okunur. Cezasız durumda iki kofte yazılır.

$$(6 \times 2) + (-1 \times 3) = 10 - 3 = 7$$

Lokal Hizalama :

Örn ACCTAAGG GGCTCAATCA

Burada globalden farklı olarak ekli sayılar yerine 0 yazılır.

		G	G	C	T	C	A	A	T	C	A
A	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	2	0	2	0	1	1	2	0
C	0	0									
T	0										
A	0										
A	0										
G	0										
G	0										

Benzersizlik = +2

Benzersizlik = -1

Bosluk = -2

Cezasız oklar önemlidir.

Yüksek skora sahip hücreden başlanır, skor pozitif kalana kadar sola/yukarıya hareket edilir.

C T C A A
C T - A A

Göklü Dizi Hesaplama :

Clustal W :

1. AAAC
2. AGC
3. ACC
4. GAC

1 ve 2
AAAC
- AGC

$$\text{Benzerlik} = \frac{\text{Faleşeler}}{\text{Dizinin Uzunluğu}} = \frac{2}{4} = 0,5$$

$1 - 0,5 = 0,5 \rightarrow$ Aralarındaki uzaklık

1 ve 3
AAAC
- ACC

$$B = \frac{2}{4} = 0,5$$

$$U = 1 - 0,5 = 0,5$$

1 ve 4
AAAC
- GAC

$$B = \frac{2}{4} = 0,5$$

$$U = 1 - 0,5 = 0,5$$

2 ve 3
AGC
- ACC

$$B = \frac{2}{3} =$$

$$U = 1 - \frac{2}{3} = \frac{1}{3} = 0,33$$

2 ve 4
AGC
- GAC

$$B = \frac{1}{3}$$

$$U = 1 - \frac{1}{3} = 0,67$$

3 ve 4
ACC
- GAC

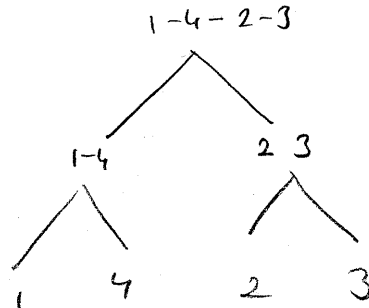
$$B = \frac{1}{3}$$

$$U = 1 - \frac{1}{3} = 0,67$$

Uzaklık az ise benzerlik daha fazla.

	1	2	3	4
1	-	0,5	0,5	0,5
2		-	0,33	0,67
3			-	0,67
4				-

	1	2 ve 3	4
1	-	0,5	0,5
2 ve 3		-	0,67
4			-



GENETİK ALGORİTUALAR 1975 J. Holland

Sonraki nesil derde güçlüler ayakta kalır, diğerleri ölür.

1989 J. Goldberg

Matematiksel modeli ^{bilinmeyen} olgular için uygun. Doğada elin'adek elde edilmiş algoritma.

$$f(x,y) = x^2 - 5xy + \frac{y^2}{4} \quad x > 0 \quad y > 0 \quad x+y \leq 15$$

$\max(f(x,y))$

116 arama başlangıç popülasyonu oluşturulur.

Bir problemi çözmek için gerekli değişkenlerin bulunduğu yapı. Birer

	y Geni	y (Gen)	
1	1100	0010	45
2	0111	0101	84
3			23
4			74
5			1
6			1
7			1
8			1
9			1
10			1
11			1
12			1
13			1
14			1
15			1
16			1
17			1
18			1
19			1
20			1
21			1
22			1
23			1
24			1
25			1
26			1
27			1
28			1
29			1
30			1
31			1
32			1
33			1
34			1
35			1
36			1
37			1
38			1
39			1
40			1
41			1
42			1
43			1
44			1
45			1
46			1
47			1
48			1
49			1
50			1
51			1
52			1
53			1
54			1
55			1
56			1
57			1
58			1
59			1
60			1
61			1
62			1
63			1
64			1
65			1
66			1
67			1
68			1
69			1
70			1
71			1
72			1
73			1
74			1
75			1
76			1
77			1
78			1
79			1
80			1
81			1
82			1
83			1
84			1
85			1
86			1
87			1
88			1
89			1
90			1
91			1
92			1
93			1
94			1
95			1
96			1
97			1
98			1
99			1
100			1

Başlangıç değerleri rasgele olarak oluşturulur

Kodlama

- İkili Kodlama
- Gerçek sayı ile Kodlama
- Gray Kodlama

$$2^4 - 1 = 15$$

$$0 \leq x \leq 14$$

$$0 \leq y \leq 14$$

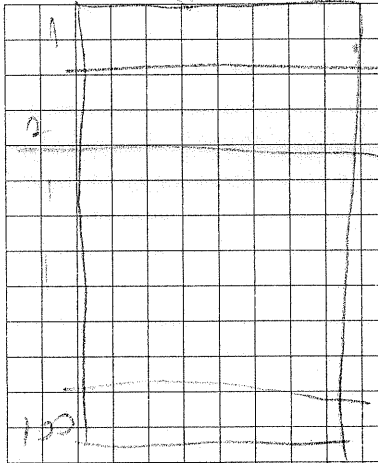
2 önemli arama var birer kodlama diğerinde uygunluk değeri (fitness)

Uygunluk Değeri (Fitness)

İnterjyonda değer yerine koyulup elde edilen değerlerden 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100. Burada daha yüksek değer bulamadığında 84'ü veren x ve y yatar-
lar denilebilir. iyiler kalır, kötüler (kötüçükler) kalır.

100 tane da en kötü 105'i atılır. 100 tane da yerine en iyiler bir tane kopulandı.

2. jenerasyon.



Genetik İlemler

1. Çaprazlama
2. Mutasyon

Çaprazlama

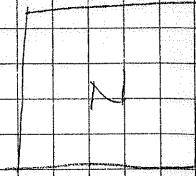
↓ rastgele seçim

0110 1000 } 0111 0111
0001 0111 } 0000 1000

Rastgele birey.

Daha güçlü oldu.

den Knapsack (Sırt Çantası)



V_1, V_2, \dots, V_k

$U_1, U_2, \dots, U_k \rightarrow$ fayda.

maximum

N_1, N_2, \dots, N_k

1 0 0 ... 1

fayda.

1 → Alındı

0 → Alınmadı

$$\sum V_i \leq N$$

$$\therefore \text{Fitness} = \max \sum U_i$$

P.C = Çaprazlama oranı = %80 → iyi bireyler

herkes %80'i

çaprazlamaya
tabi tutulur.

fitness hesaplanır.

ilk jenerasyonda daha iyi sonuçlar elde etmek büyük olasılık.

Çaprazlanmış 80

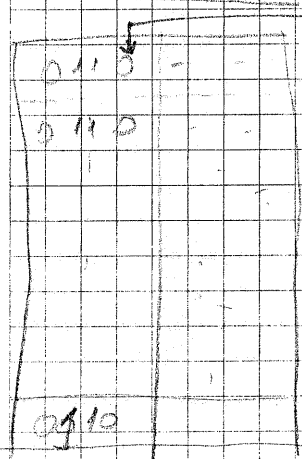
Sıralama kriteri ~~2~~ 2 farklı.

→ 100 jenerasyon ile kesin çözüm bulunur.

→ Arka arkaya iki jenerasyonda

film çekildiği varsayarak çekim sürüyor. Sonuçta en iyi bireyleri gözden alın.

2. Mutasyon



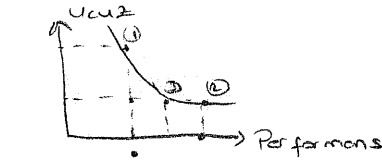
Genetik süreçte 1 olması gerek.

Rastgele bir bireyin rastgele bir bitini

değiştirip 0 → 1, 1 → 0 yapma.

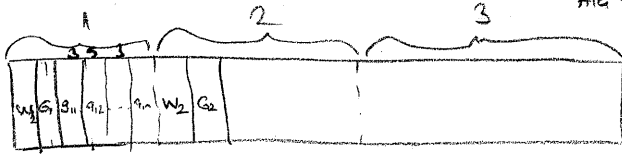
10 jenerasyonda bir olur.

Coklu dizii sıralamalarda tek çözüm sunulmaz, alternatif ^{CUMA} çözümlerde 5



Grafik Algoritma ile Çoklu Dizii Sıralama :

- S1) 10 [10, 14] } Her birinin bir ^(w) ağırlığı vardır. Başlangıçtaki en küçük değerin
 S2) 15 [5, 9] } den küçük olanlar sıralamaya alınmaz. Bazı diziler
 S3) 20 [0, 4] } zorla sıralamaya sokulmamalıdır.



ATT - - GT -
3 3 5

ATTG } ATTG
ATG } ATG

AT - - GT

q₁ değerleri : 2 2

AT - - GT

2 2 3 → AT - - G - T

q : Pozisyonu söyler. Başlukların yeri.

G : Aralığı gösterir. ^{ömek} [10, 14]
Kaç tane ek alına-
caktır. [5, 9]
[0, 4]

ATT - - } Minimum hizalama uzunluğu : 5 →

AAA - - - } Maximum hizalama uzunluğu 7 →

ATT - -
ATT T G

AAA - - -
T T T T G

En yüksek dizinin %20'si geçmiyor hizalama uzunluğu.

--AT- GT- → 0 0 2 4 = G1 = 4

TTAT- GT A → 4 = G2 = 1

Amaçlar :

→ Benzerlik : Her kolonistaki max değer.

Diziler sıyfta $\frac{0,5+1+0,8+1}{4} = \text{maximize et}$

Max = 1 olur minimum 0,25

→ Başlığın Art Arada Gelme Durumu :

AAAGAAATTCA

AAAGAAATTCA

A-A-A-T-CA

TAAA-----TCA

→ bu daha değerli
Parça elde edilerek
hizalamalar yapılmalı

ilk görülen boşluk 5 ile cezalandırılır, arka arka gelirse ~~10~~
ikinci boşluklar 2 ile cezalandırılır.

a) ~~20~~ 20 b = 11

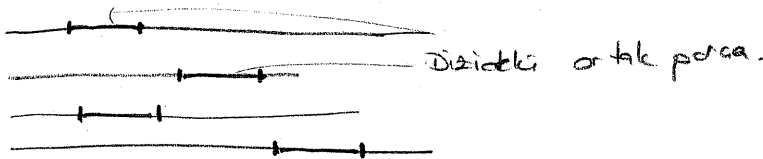
3) Destek Değeri

Olabildiğince tüm dizieler ele alınmaya çalışılır. Ancak bazı diziler
ağırlıklarından dolayı elenebilir. Ele aldığımız dizi sayısını maximize
etmeye çalışılır.

11.04.2014
CUMA

MOTIF BULMA :

Farklı uzunlukta dizilerde aynı parçaların elde edilmesine motif denir.



Bilinen motif örnekleri: TATA box (TATAAAA), BRE ((GIC)(GIC)(GIA)CGCC)

3 gen parçasının ortak alanlarını bulma motifin konusuna girer.

ACCGTA
TTGAT

→ Başlangıç olarak verip dizielerde bu parçanın varlığını arama.

Dezavantajı : motifleri ve motif uzunluğunu dışardan girmek
motif başlığını dışardan vermek
motif uzunluğunu " vermek.

Score Hesaplama :

$$PS(S_m, P_m) = \max \left\{ \sum_{i=1}^k \text{match}(S_m, P_m) / k \right\}$$

m : A/C

R : A/G

W : A/T

S : C/G

Y : C/T

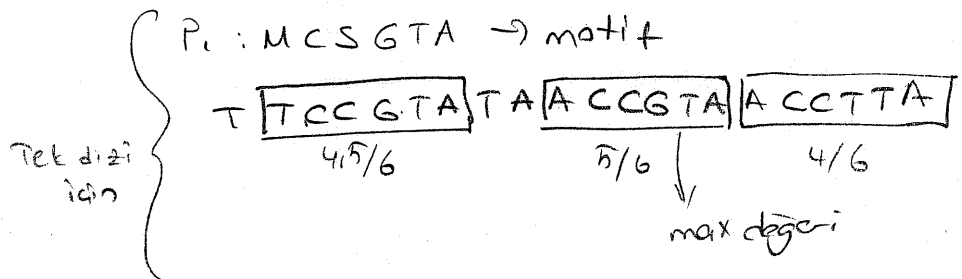
K : G/T

IUPAC Kodları
(Belirsizlik)

Tam eşleşme varsa 1

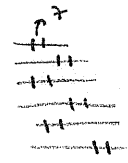
IUPAC'den biri varsa 0,5

Yoksa 0



Her dâire en uygun değeri bulup bunu maximize etmeye çalışıyoruz.

Grup 1:

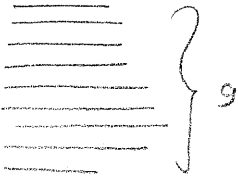


FMGA algoritmasıyla elde edilen sonuçları karşılaştırmış

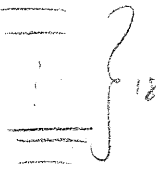
} 6

→ $6 \cdot 7 = 42$ → Toplam Fitness Skoru. %84 match demek 1 tane
match sequences: (6/6) hataya izin verme. (6/7)

Grup 2:



Grup 3:



MEME, Gibbs Sampler → diğer algoritmalar.

Alternatif bulma sebebi farklı motiflerden oluşan popülasyonlar da olabilir.

