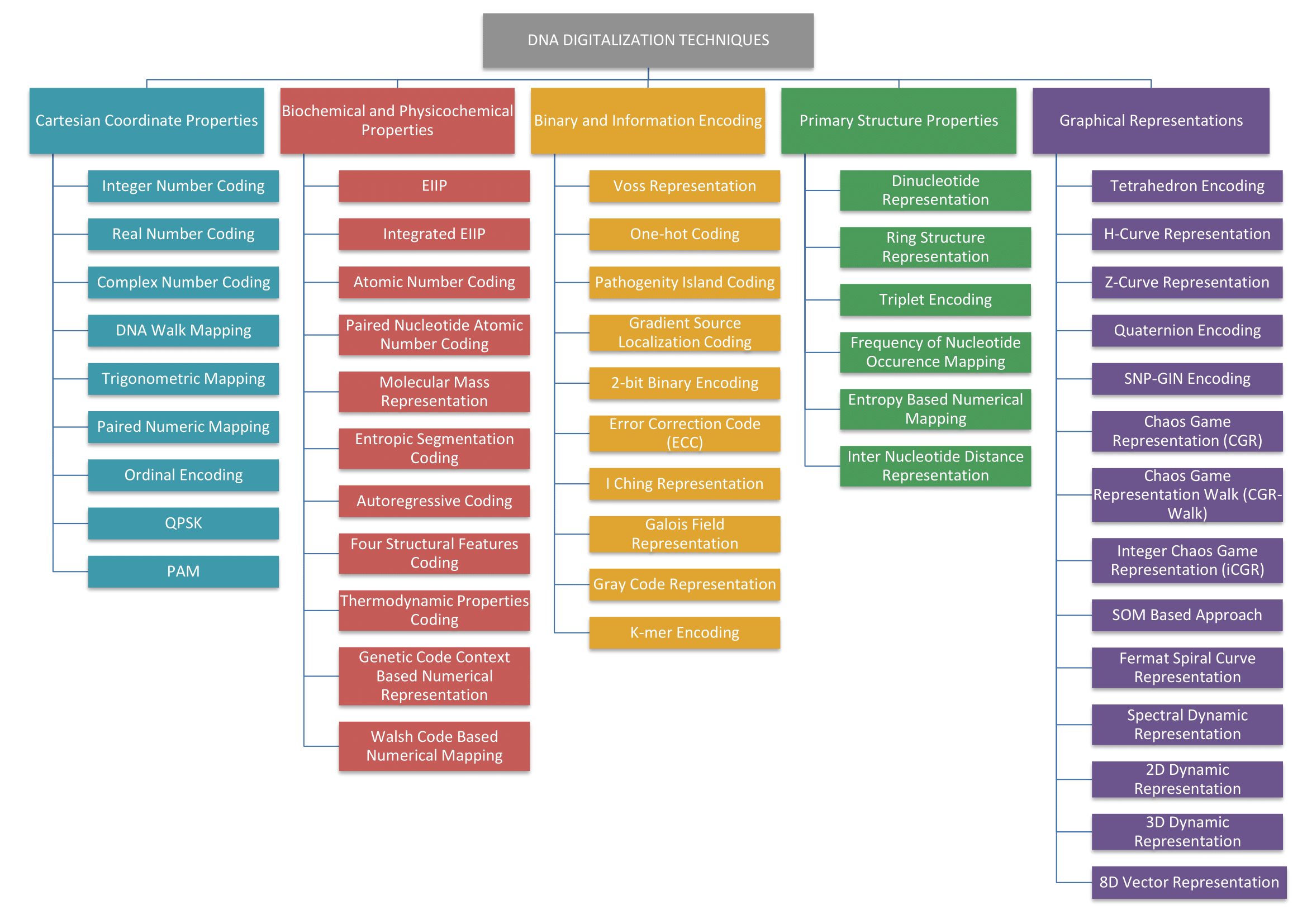
# DNA Sayısal Haritalama Tekniklerinin İncelenmesi ve Sayısal Temsilleri

Bu bölümde DNA dizilerinin sayısallaştırılması için geliştirilen kodlama teknikleri beş ana başlık altında kapsamlı bir şekilde incelenmektedir. Literatürde kodlama teknikleri sayısal haritalama teknikleri, sayısal yöntemler ve kodlama şemaları gibi farklı isimlerle de anılmaktadır. Ancak, tüm isimlendirmeler aynı anlama gelmektedir. Son 10 yılda geliştirilen 50 sayısal haritalama tekniği, genel özelliklerine göre beş gruba ayrılmaktadır. Bu gruplar kartezyen koordinat kodlama teknikleri, biyokimyasal ve fizikokimyasal kodlama teknikleri, ikili ve bilgi kodlama, birincil yapı kodlama teknikleri ve grafik olarak temsil edilen kodlama teknikleridir. Bu beş grubun her birinin sonunda o grupta incelenen sayısal kodlama tekniklerinin toplu sayısal işaret grafiği ve tabloları bulunmaktadır. Kodlama teknikleri kullanılarak sayısallaştırılmış DNA sinyallerinin grafikleri, NCBI veri tabanından NR\_131216.1 referans numarası ile DNA fasta formatındaki veri setine uygulanan her kodlama tekniği ile sayısallaştırılmış DNA dizisinin temsillerini içermektedir. Bazı haritalama tekniklerinin grafiksel değer aralıkları farklı olduğundan, bu tekniklerin sayısal gösterimi ayrı şekillerde verilmiştir. İncelenen grupların sonundaki genel tablolarda, o gruptaki sayısallaştırma tekniği, kodlama şeması ve bu kodlama tekniğinin örnek bir DNA dizisine uygulanan sayısal versiyonu hakkında kısa bir açıklama bulunmaktadır. Şekil 4.1, tüm DNA haritalama tekniklerinin hiyerarşik şemasını göstermektedir.



**Şekil 4. 1.** Tüm DNA haritalama tekniklerinin hiyerarşik şeması

## Kartezyen-Koordinat Özellikli Teknikler

DNA sayısal haritalama tekniklerinin ilk grubu, kartezyen koordinat özellikleri (KKÖ) (CCP-Cartesian Coordinate Properties) sayısallaştırma teknikleridir. Bu grup içerisinde Tam Sayı Kodlama, Gerçek Sayı Kodlama, Karmaşık Sayı Kodlama, DNA-Yürüyüş Kodlama, Trigonometrik Kodlama, Eşleştirilmiş Sayısal Kodlama, Sıralı Kodlama, QPSK (Dörtlü Faz Kaydırmalı Anahtarlama) Kodlama ve PAM (Darbe Genlik Modülasyonu) Kodlama teknikleri olmak üzere dokuz tane kodlama şeması incelenmektedir.

### Tam Sayı Kodlama Tekniği

Tam sayı kodlama tekniği, DNA nükleotidlerinin tam sayılarla kodlanması ile gerçekleştirilen bir boyutlu, 0, 1, 2, 3 gibi keyfi tam sayı değerlerin kullanıldığı bir haritalama tekniğidir. DNA diziliminde eğer pürin *(A, G)>*pirimidin *(C, T)* ise 4 baza T=0, C=1, A=2, G=3 değerleri verilir. *T>A* ve *G>C* ise bazlara A=0, C=1, T=2, G=3 değerleri verilir [20]. Bu tekniğin avantajı, tek boyutlu sayısal dizi elde edildiğinden hesaplama karmaşıklığı azdır. Dezavantajı ise keyfi atama ile yapılan kodlamalar gerçek DNA sinyalleri sağlamamaktadır [11].

### Gerçek Sayı Kodlama Tekniği

Reel sayı kodlamasında DNA dizisindeki bazlar A=-1.5, T= 1.5, C= 0.5, G=-0.5 reel sayıları ile kodlanır [20]. Bu tekniğin avantajı sinyal değerlerinin ortalaması sıfır ve sapmalarının simetrik olmasından dolayı sinir ağı uygulamalarında, veri eğitimi ve özellik öğreniminde fayda sağlamasıdır [11]. Ayrıca daha az gürültülü çıkış elde edilmesini sağlamaktadır. Ancak DNA sinyalini keyfi reel sayılarla temsil ettiğinden DNA yapısını çok iyi yansıtamaması bu tekniğin bir dezavantajıdır.

### Karmaşık Sayı Kodlama Tekniği

Karmaşık sayı kodlama, A-T ve C-G çiftlerinin tamamlayıcı yapısına göre karmaşık sayılar ile gerçekleştirilen iki boyutlu sayısal haritalama tekniğidir. A= 1+j, C= -1+j, G= -1-j, T= 1-j değerleri ile kodlama yapılır. Nükleotidler iki boyutlu kartezyen koordinat düzleminde farklı köşelere yerleştirilerek A, T, G, C için farklı kodlama değerleri verilir. Bu tekniğin avantajı, üç boyutlu sayısallaştırmayı iki boyuta indirgemesidir. Düzlemler değiştirilerek iki farklı sayısallaştırma temsili oluşturulabilir. Birincisinde C-G ve A-T nükleotid çiftleri matematiksek olarak karmaşık eşlenik değerlere sahip, pürin (A, G) ve pirimidin (C, T) eşit karmaşık değerlere ve zıt işaretli gerçek sayı değerlerine sahiptir (A= 1+j, C= -1-j, G= -1+j, T= 1-j). İkincisinde DNA’nın iki tamamlayıcı ipliği eşit karmaşık değerlere ve zıt işaretlere sahiptir (A= -1+j, C= -1-j, G= 1+j, T= 1-j). Böylece cebirsel toplamları sıfır olur ve tamamlayıcı yapıdadır, bu da sinyal işlemede avantaj sağlar [20]. Kartezyen koordinat düzleminin 45 derece döndürülmesi ile oluşturulan kodlama şeması (A= -1, C= -j, G= j, T= 1) değerlerini alır [11]. Karmaşık sayı kodlama, dört ikili Voss kodlama şemasının hesaplama yükünü %75 oranında azaltan tek işaret dizisidir [14].

### DNA-Yürüyüş Kodlama Tekniği

DNA Yürüyüş kodlama, bir pirimidin (C-T) için yukarı (+1) ve pürin (A-T) için aşağı (-1) değerleri ile DNA dizisi boyunca grafiksel olarak bir yolu gösteren iki boyutlu haritalama tekniğidir. Bir DNA dizisinde nükleotid bileşimi, baz çifti desenleri evrim değişikliklerini görselleştirmek için kullanılabilir. Bu tekniğin avantajı, DNA dizilerinin genel istatistiksel özelliklerini gösterebilmektedir. Bu nedenle birçok yöntemin temelini oluşturur [11]. Denklem 3.1, bu tekniğin sayısal temsilinin formülünü göstermektedir [16].

= (4. 1)

DNA yürüyüşünde her nükleotid matematiksel olarak temsil edildikten sonra DNA dizisi, kümülatif toplam alınarak haritalanır [17]. Ekzon bölgelerinde daha belirgin tepe noktaları elde edilmesini sağlayabilir ancak bu tekniğin dezavantajı 1000 bazdan uzun DNA dizileri için kullanışlı olmamasıdır.

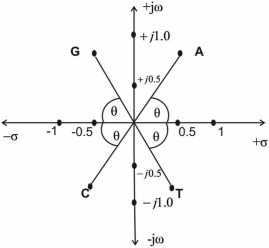
### Trigonometrik Kodlama Tekniği

Trigonometrik kodlama, DNA nükleotidlerine trigonometrik değerlerin atanması ile gerçekleştirilir. Denklem 3.3’te trigonometrik kodlamanın sayısal temsilinin formülü gösterilmektedir [27].

A = C =

G = T = (4. 2)

Bu tekniğin avantajı ekzon konumlarının tanımlanmasında kullanılan Gauss-Newton ayarlı uyarlanabilir Kaiser penceresine çok uygun olacak şekilde tasarlanmış olmasıdır ve uzun DNA dizilimleri için iyi performans göstermektedir [8]. Ancak bu tekniğin dezavantajı ise spektral analiz için uygulandığında bazı iyileştirmelere ihtiyaç duymasıdır [27]. Trigonometrik haritalamanın grafiksel gösterimi Şekil 4.2’deki gibidir.



**Şekil 4.2.** Trigonometrik kodlamanın grafiksel gösterimi [27]

### Eşleştirilmiş Sayısal Kodlama Tekniği

Eşleştirilmiş sayısal kodlama tekniğinde, DNA iki ipliğinde bulunan A-T ve G-C baz çiftleri +1 ve  değerleri ile kodlanır. Bu tekniğin avantajı DNA’nın karmaşıklığını azaltmasıdır. Nükleotid çiftlerini bu teknikle sayısallaştırmak için 7 kural bulunmaktadır. Bu kurallar içinde çalışmalarda en çok kullanılanı purin pirimidin kuralıdır. Pürin-pirimidin kuralı (RY kuralı): Eğer ni nükleotid dizisi pürin (A-G) ise ui = 1, ni pirimidin (C-T) ise ui = -1 dir. [20]. Bu sayısallaştırma tekniği en çok gen ve ekzon tahmininde kullanılır. Bazların ekzon ve intronlarda bulunma sıklığından faydalanmaktadır. İntronlarda A ve T bazları, ekzonlarda G ve C bazları daha fazla bulunur. Böylece eşleştirilerek verilen değerler ile bu ayrımdan faydalanılabilir [20].

### Sıralı Kodlama Tekniği

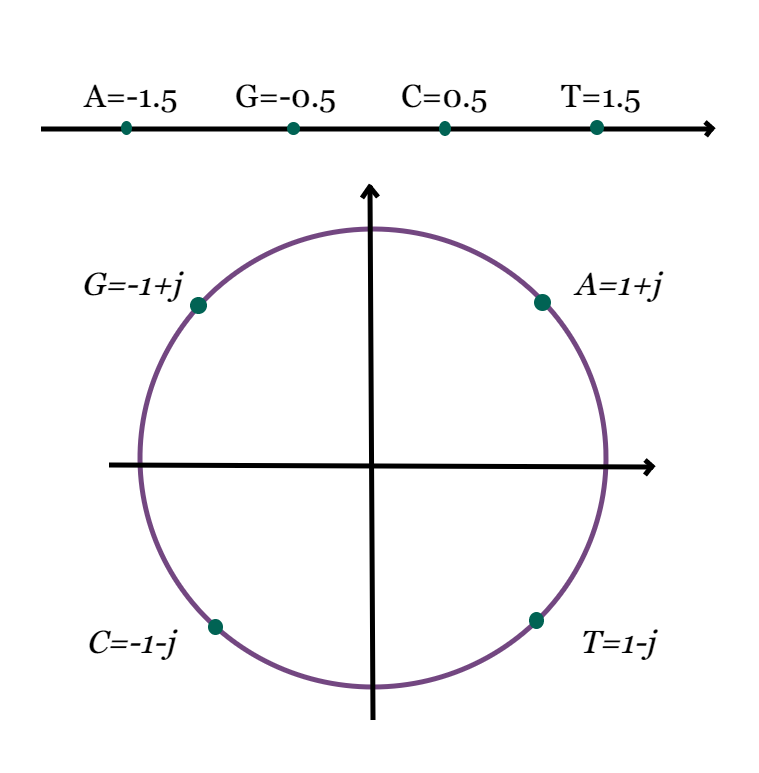
Sıralı kodlama tekniği, A= 0.25, C= 0.50, G= 0.75, T= 1.00 değerleri nükleotidlere verilerek gerçekleştirilir. Bilinmeyen nükleotid için 0.00 değeri verilir. Bu tekniğin avantajı sinir ağı kullanılan sistemlerde giriş dizisinin boyutunun azaltılmasını sağlamasıdır. Bu tekniğin dezavantajı ise uzun eğitim süresine neden olur ve daha çok kısa diziler için uygulanabilmesidir [28].

### QPSK Kodlama Tekniği

QPSK (Quadrature Phase Shift Keying) (Dörtlü Faz Kaydırmalı Anahtarlama) kodlama, gen dizisinin tamamlayıcı özelliğine dayalı olarak bazlara A= 1+j, G= -1+j, C= -1-j, T= 1-j karmaşık değerlerin atanmasıyla gerçekleştirilmektedir. Sinyal işlemede yaygın olarak kullanılan takım yıldızı diyagramlarından biri olarak kabul edilmektedir. İki boyutlu düzlemde DNA’yı temsil eder ve genetik kodların simetrisini sağlar [11]. Bu tekniğin avantajı bir sembol haritalamasının bilgi aktarımının takip ettiği dijital iletişimde, DNA dizisinin görselleştirilmesine yardımcı olmasıdır. Ayrıca bu teknik, DNA dizilerinin analizi, DNA’nın hata düzeltme yeteneği gibi bazı yönlerini ortaya çıkarabilir [29].

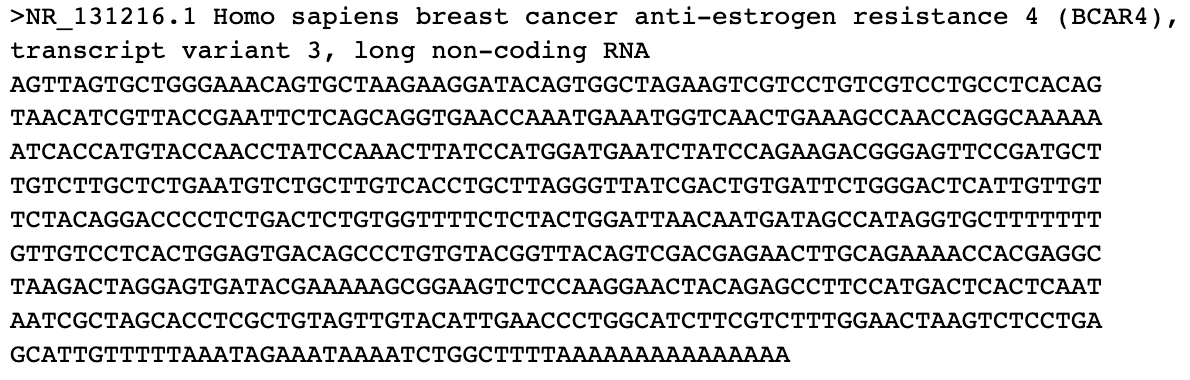
### PAM Kodlama Tekniği

PAM (Pulse Amplitude Modulation) (Darbe Genlik Modülasyonu) kodlama, nükleotidlerin gerçek sayılarla temsilini gösterir. Bir veri dizisindeki bitlerin darbe seviyelerinin kodlanmasıdır [30]. PAM’da, QPSK gibi bir takım yıldızı diyagramıdır. Bazlara A=-1.5, G=-0.5, C= 0.5, T= 0.5 değerleri verilerek kodlama yapılmaktadır. Bu tekniğin avantajı tek boyutlu düzlemde gerçek sayılarla genetik kodların simetrisini sağlamasıdır [11]. Şekil 4.3, gerçek ve karmaşık sayı gösterimleri için takımyıldız diyagramını göstermektedir.



**Şekil 4.3.** Gerçek ve karmaşık sayı gösterimleri için takımyıldızı diyagramı

Şekil 4.4, NCBI veri tabanından alınan NR\_131216.1 referans numaralı DNA diziliminin fasta formatını göstermektedir.



**Şekil 4.4.** NR\_131216.1’in fasta formatının gösterimi

Genbankalarındaki örnek DNA veri setleri, fasta formatında bulunmaktadır ve analog sinyallerdir. Bu dizilimlerin ilk 100 bazının “Kartezyen Koordinat Özellikli (KKÖ)” grubundaki kodlama teknikleriyle sayısallaştırılmış sinyal gösterimleri Şekil 4.5 ve Şekil 4.6‘da gösterilmektedir. Tablo 4.1, KKÖ grubundaki haritalama tekniklerinin tamamının kısa bir özet bilgisini vermektedir.



**Şekil 4.5.**  Kartezyen koordinat özellikli (KKÖ) grup tekniklerinin sayısal gösterimleri



**Şekil 4.6.** Kartezyen koordinat özellikli (KKÖ) grup tekniklerinin karmaşık düzlem sayısal gösterimleri

**Tablo 4. 2.** Kartezyen koordinat özellikli gruptaki tüm sayısal kodlama tekniklerinin özeti

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tekniğin Adı | Kodlama Şeması | Sayısal Temsil | Tanım | |
| Tam Sayı Kodlama | Purin>Pirimidin ise T=0, C=1, A=2, G=3  T>A ve G>C ise  A=0, C=1, T=2, G=3 | *X*=[AGCTACCGTG]  =[2, 3, 1, 0, 2, 1, 1, 3, 0, 3] | | Nükleotidler tam sayılarla temsil edilir. |
| Gerçek Sayı Kodlama | A=-1.5, T=1.5,  C=0.5, G=-0.5 | *X*=[AGCTACCGTG]  =[-1.5, -0.5, 0.5, 1.5, -1.5, 0.5, 0.5,  -0.5, 1.5, -0.5] | | Nükleotidler gerçek sayılarla temsil edilir. |
| Karmaşık Sayı Kodlama | A= -1, C= -j,  G= j, T= 1 | *X*=[AGCTACCGTG]  =[*1, –1, j, j, 1, –j, –j, –1, j, –1*] | | Nükleotidler karmaşık sayılarla temsil edilir. |
| DNA Yürüyüş Kodlama | Integer temsil;  A= -1, C= 1  G=- 1, T= 1  Complex temsil;  A=1, C= -j  G= -1, T= j | *X*=[AGCTACCGTG]  = [-1, -2, -1, 0, -1, 0, 1, 0, 1, 0] | | Nükleotidlere tam sayı veya karmaşık sayılar atanarak DNA dizisi boyunca değerleri toplanarak kodlanır |
| Trigonometrik Kodlama | A= cos(θ) + j ×sin(θ)  C= -cos(θ) - j ×sin(θ)  G= -cos(θ) + j ×sin(θ)  T= cos(θ) - j ×sin(θ) | *X*=[AGCTACCGTG]  = [0.5+0.8660i, -0.5+0.8660i,  -0.5+0.8660i,0.5+0.8660i, 0.5+0.8660i, -0.5+0.8660i,  -0.5+0.8660i, -0.5+0.8660i  0.5+0.8660i, -0.5+0.8660i] | | Nükleotidlere trigonometrik denklemler atanarak kodlanır. |
| Eşleştirilmiş Sayısal Kodlama | Purin(A&G)= 1  Pirimidin(C&T)= -1 | *X*=[AGCTACCGTG]  = [1, 1, -1, -1, 1, -1, -1, 1, -1, 1] | | Nükleotidlere yapısal özelliklerine göre değerler atanarak kodlanır. |
| Sıralı Kodlama | A= 0.25, C= 0.50  G= 0.75, T= 1.00 | *X*=[AGCTACCGTG]  = [0.25, 0.75, 0.50, 1.00, 0.25, 0.50, 0.50, 0.75, 1.00, 0.50] | | Nükleotidlere sıralı, doğrusal değerler atanır. |
| QPSK Kodlama | A= 1+j, G= -1+j  C= -1-j, T= 1-j | *X*=[AGCTACCGTG]  = [1+j, -1+j, -1-j, 1-j, 1+j, -1-j, -1-j, -1+j, 1-j, -1+j] | | DNA’nın tamamlayıcı özelliğine göre 2D QPSK takımyıldızı karmaşık sayı değerleri atanır. |
| PAM Kodlama | A= -1.5, G= -0.5  C= 0.5, T= 0.5 | *X*=[AGCTACCGTG]  = [-1.5, -0.5, 0.5, 0.5, -1.5, 0.5, 0.5, -0.5, 0.5, -0.5] | | Nükleotidler 1D gerçek sayılarla temsil edilir. |

## Biyokimyasal ve Fizikokimyasal Özellikli Teknikler

DNA sayısal haritalama tekniklerinin ikinci grubu, biyokimyasal ve fizikokimyasal özellikli (BFÖ) (BPP-Biochemical and Physicochemical Properties) sayısal tekniklerdir. Bu grup içerisinde EIIP Kodlama, Entegre EIIP Kodlama, Atom Numarası Kodlama, Eşleştirilmiş Nükleotid Atom Numarası Kodlama, Moleküler Kütle Kodlama, Entropik Bölümleme Kodlama, Otoregresif Kodlama, Dört Yapısal Özellik Kodlama, Termodinamik Kodlama, Genetik Kod Bağlam Tabanlı (GCC) Kodlama ve Walsh Koduna Dayalı Kodlama teknikleri olmak üzere on bir tane kodlama şeması incelenmektedir.

### EIIP Kodlama Tekniği

Biyomolleküler arasındaki elektromanyetik salınımdan kaynaklanan elektromanyetik rezonansa EIIP denir [8]. EIIP (Electron-Ion Interaction Pseudopotential) kodlamada dört nükleotid için C= 0.1340, T= 0.1335, A= 0.1260, G= 0.0806 değerleri verilir. EIIP değerleri ile kodlanan DNA dizisi, serbest elektronların psödopotensiyel enerjilerinin DNA dizisi boyunca dağılımını gösteren sayısal diziyi oluşturur. Bu tekniğin avantajı, sinir ağı, dalgacık dönüşümü ve genomik sinyal işleme gibi alanlarda iyi performans göstermesidir [11]. EIIP kodlama, Voss tarafından sunulan dört ikili işaret dizisi olan kodlamanın hesaplama yükünü %75 azaltmaktadır [14]. Ancak bu tekniğin dezavantajı ise bazı genomlarda protein kodlayan bölgeleri tam olarak tespit edememesidir.

### Entegre EIIP Kodlama Tekniği

Entegre EIIP kodlama tekniğinde, bazlar üçlü nükleotidler şeklinde kodonlara elektron-iyon psödopotensiyel enerji değerlerin atanması ile gerçekleşmektedir. EIIP haritalamasında biyolojik bilgiyi kodlamak için tek nükleotidin elektron potansiyeli kullanılıyordu. Entegre EIIP’de her üç nükleotidin EIIP değerleri toplanır ve kodona toplam sayısal değer atanır. Sonra kaydırılarak birer birer hareket ettirilir ve aynı sayısallaştırma her kodon için devam ettirilir [31]. Bu tekniğin avantajı EIIP tekniğinin ekzon bölgelerini belirlemedeki eksikliklerini daha iyi tamamlayabilmektedir. Dezavantajı ise bu tekniğin de protein kodlayan bölgelerde %100 bir başarı elde edememesidir. Tablo 4.2, entegre EIIP kodlaması için kodonlara verilen sayısal değerleri göstermektedir.

**Tablo 4. 3.** Entegre EIIP kodon değerleri

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| A (0.1260) | IEIIP | C (0.1340) | IEIIP | G (0.0806) | IEIIP | T (0.1335) | IEIIP |
| AAA | 0.3780 | ACA | 0.3860 | AGA | 0.3326 | ATA | 0.3885 |
| AAC | 0.3860 | ACC | 0.3940 | AGC | 0.3406 | ATC | 0.3935 |
| AAG | 0.3326 | ACG | 0.3406 | AGG | 0.2872 | ATG | 0.3401 |
| AAT | 0.3885 | ACT | 0.3935 | AGT | 0.3401 | ATT | 0.3930 |
| CAA | 0.3860 | CCA | 0.3940 | CGA | 0.3406 | CTA | 0.3935 |
| CAC | 0.3940 | CCC | 0.4020 | CGC | 0.3486 | CTC | 0.4015 |
| CAG | 0.3406 | CCG | 0.3486 | CGG | 0.2952 | CTG | 0.3481 |
| CAT | 0.3935 | CCT | 0.4015 | CGT | 0.3935 | CTT | 0.4010 |
| GAA | 0.3326 | GCA | 0.3406 | GGA | 0.2872 | GTA | 0.3401 |
| GAC | 0.3406 | GCC | 0.3486 | GCG | 0.3935 | GTC | 0.3481 |
| GAG | 0.2872 | GCG | 0.2952 | GGG | 0.2418 | GTG | 0.2947 |
| GAT | 0.3401 | GCT | 0.3481 | GGT | 0.2947 | GTT | 0.3476 |
| TAA | 0.3855 | TCA | 0.3935 | TGA | 0.3401 | TTA | 0.3930 |
| TAC | 0.3935 | TCC | 0.4015 | TGC | 03481 | TTC | 0.4010 |
| TAG | 0.3401 | TCG | 0.3481 | TGG | 0.2947 | TTG | 0.3476 |
| TAT | 0.3930 | TCT | 0.4010 | TGT | 0.3476 | TTT | 0.4005 |

Örnek DNA diziliminin (NR\_131216.1) EIIP ve Integrated EIIP teknikleriyle sayısallaştırılmış sinyal gösterimleri Şekil 4.7’ de gösterilmektedir.



**Şekil 4.7.** EIIP ve Entegre EIIP tekniklerinin sayısal gösterimleri

### Atom Numarası Kodlama Tekniği

Atom numarası kodlama tekniği, DNA dizisindeki her nükleotide atom numaralarının atanması ile gerçekleştirilir. C=58, T= 66, A= 70, G= 78 atom numaraları verilerek kodlanır. Bu tekniğin avantajı ekzon ayrımında, sinyal işleme tekniklerinde ve türlerin genomik dizileri arasındaki fraktal boyut farklılıklarını ölçmek yaygın olarak kullanılmaktadır [11]. Dezavantajı ise bazlara sayısal bir değer ataması yaptığı için DNA’nın yapısını tam olarak yansıtamamasıdır.

### Eşleştirilmiş Nükleotid Atom Numarası Kodlama Tekniği

Eşleştirilmiş atom numaralarının temsili ile kodlama tekniği, her baza atom numaralarının verilmesi ile değil pürin ve pirimidinlere atom numarası verilerek gerçekleştirilir. G-A = 62, C-T = 42 değerleri ile kodlanır. Bu tekniğin avantajı, ATCG nükleotid dalgalanmasının tespiti üzerine çalışmalarda tercih edilmesidir [14]. Ancak bu teknikte atomic number kodlama gibi DNA’ın özelliklerini tam olarak ortaya çıkaramamaktadır.

### Moleküler Kütle Kodlama Tekniği

Moleküler kütle kodlama tekniği, dört nükleotide moleküler kütlelerinin atanması ile gerçekleştirilmektedir. C= 110, G= 150, A= 134, T= 125 kütleleri atanarak DNA dizisinin haritalanması gerçekleştirilir [14], [20]. Bu tekniğin avantajı, kolay ve kullanışlı bir yöntem olması ve ek araştırmaya ihtiyaç duyulmamasıdır.

### Entropik Bölümleme Kodlama Tekniği

Entropik segmentasyon kodlama tekniği, büyük DNA segmentlerini tanımlamak için 12 harfli nükleotid oluşumunun temsil edilmesi ile gerçekleştirilir. DNA dizilerinde her kodon pozisyonundaki diferansiyel baz bileşimini 12 harfli alfabe ile tanımlar ve sınırların aranması entropik bölütleme yöntemi aracılığı ile gerçekleştirilmektedir [32,33]. 12 sembollü alfabe, kodonların içindeki nükleotid istatistiklerine dayanmaktadır. Kodonun birinci, ikinci ve üçüncü olmasına göre 12 harfli alfabe ile numaralandırılmış semboller atanır. Dizide ilerlerken okunan kodon stop kodonlarından biriyse önce hangi stop kodon olduğuna göre tablodaki sembol yeni diziye eklenir daha sonra kaçıncı baz olduğuna göre sembol atamasına devam edilir. Alfabeye göre kodlanan dizi için daha sonra entropi hesaplaması gerçekleştirilir. Bu tekniğin avantajı ekzon ve intron sınıflandırılmasında iyi performans sağlamasıdır [14].

### Otoregresif Kodlama Tekniği

Otoregresif kodlama tekniği, DNA dizilerinin yapısal özelliklerine dayanmaktadır. Kodlama yapılırken pervane büküm ve DNA eğilme sertlik değerleri kullanılmaktadır [14]. Dinükleotid pervane bükümü, derece olarak ölçülen büküm açısıdır. Pervane büküm değerinin yüksek olması o bölgede DNA sarmalının daha kuvvetli olduğunu gösterir. Pervane bükün değerinin düşük olması sarmalın nispeten daha esnek olduğu anlamına gelir. Sayısal değerler DNA dizisinin birer kaydırılarak ikili nükleotidler şeklinde okunması ile atanır ve kodlama gerçekleştirilir [34]. Bu tekniğin avantajı, bir DNA dizilimindeki kısa ekzonların 3-periyodiklik özelliklerinin bulunmasında büyük kolaylık sağlamasıdır. Tablo 4.3 pervane büküm değerlerini göstermektedir.

**Tablo 4. 4.** DNA pervane büküm değerleri (Propeller twist values)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **AA** | -18.66 | **GA** | -13.48 | **CA** | -9.45 | **TA** | -11.85 |
| **AC** | -13.10 | **GC** | -11.08 | **CC** | -8.10 | **TC** | -13.48 |
| **AG** | -14.00 | **GG** | -8.10 | **CG** | -10.03 | **TG** | -9.45 |
| **AT** | -15.01 | **GT** | -13.10 | **CT** | -14.00 | **TT** | -18.45 |

Bükülme sertliği, DNA’nın anisotropic flexibility ile dizi kolerasyonu olarak kabul edilir. Yüksek değerler DNA bölgelerinde daha katı bölgeleri gösterir, düşük değerler daha az katı bölgeleri göstermektedir [35]. Tablo 4.4 eğilme sertlik değerlerini göstermektedir.

**Tablo 4. 5.** DNA eğilme sertlik değerleri (Bending stiffness values)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| AA | 35 | GA | 60 | CA | 60 | TA | 20 |
| AC | 60 | **GC** | 85 | **CC** | 130 | **TC** | 60 |
| AG | 60 | **GG** | 130 | **CG** | 85 | **TG** | 60 |
| AT | 20 | **GT** | 60 | **CT** | 60 | **TT** | 35 |

### Dört Yapısal Özellik Kodlama Tekniği

Bu teknikte, DNA molekülünün dört fiziksel özelliği kullanılarak kodlama yapılmaktadır. DNA’nın yapısal özellik değerlerinin fiziksel modeller veya biyolojik deneyler ile elde edilmiştir [35]. DNA dizisi birer kaydırılarak dinükleotidler (DN) elde edilir ve → DNA Bükülme Sertliği (DBS) (DNA Bending Stiffness - DBS), → Dubleks Bozulma Enerjisi (DBE) (Duplex disrupt energy - DDE), → Dubleks Serbest Enerji (DSE) (Duplex free energy - DFE), → Pervane Bükümü (PB) (Propeller twist - PT) dizileri sayısal değerlere göre elde edilir. Bu tekniğin avantajı DNA’ın fiziksel özelliklerini korumasıdır. Ayrıca protein kodlama bölgelerini tanımlamak için gen tahmini uygulamasında iyi performans göstermesidir. Tablo 4.5 dört yapısal özelliğin dönüşüm tablosunu göstermektedir. Dört yapısal özellik ile kodlama tekniği ile sayısallaştırılmış sinyal gösterimi Şekil 4.8’de gösterilmektedir.

**Tablo 4. 6.** Dört yapısal özellik dönüşüm tablosu

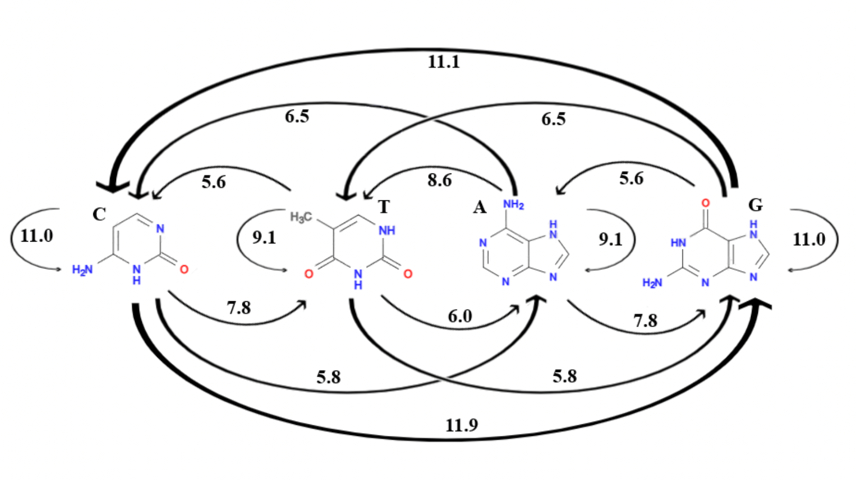
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DN | DBS | DBE | DSE | PB | DN | DBS | DBE | DSE | PB |
| AA | 35 | 19 | -12 | -1866 | **GA** | 60 | 16 | -15 | -1348 |
| AC | 60 | 13 | -15 | -1310 | **GC** | 85 | 31 | -23 | -1108 |
| AG | 60 | 16 | -15 | -1400 | **GG** | 130 | 31 | -23 | -810 |
| AT | 20 | 9 | -9 | -1501 | **GT** | 60 | 13 | -25 | -1310 |
| CA | 60 | 19 | -17 | -945 | **TA** | 20 | 15 | -9 | -1185 |
| CC | 130 | 31 | -23 | -810 | **TC** | 60 | 16 | -15 | -1348 |
| CG | 85 | 36 | -28 | -1003 | **TG** | 60 | 19 | -17 | -945 |
| CT | 60 | 16 | -15 | -1400 | **TT** | 35 | 19 | -12 | -1866 |



**Şekil 4. 8.** Dört yapısal özellik kodlamasının sayısal temsili

### Termodinamik Kodlama Tekniği

Termodinamik özelliklere göre kodlama yan yana bulunan nükleotidlerin termodinamik etkileşimlerinin entalpi değerleri ile gerçekleştirilmektedir. DNA molekülleri arasındaki termodinamik etkileşimlerin entalpi değerleri Şekil 4.9’ da gösterilmektedir [11].



**Şekil 4. 9.** Moleküller arası termodinamik etkileşimler için entalpi değerleri

Yakın nükleotid kombinasyonlarının kimyasal ve biyolojik entalpi değerlerine göre nükleotidler C, T, A, G sırasıyla yerleştirildiğinde biyokimyasal özellikleri yansıtan entalpi değerlerinin simetrik özellik göstermektedir. Bu tekniğin avantajı, moleküllerin fiziksel boyut ve ağırlıklarına göre artan sırada C, T/U, A, G olarak simetrik kodlara karşılık gelen en iyi sıralama olmasıdır[36]. Dezavantajı ise ekzon bölgelerini belirmede de tatmin edici sonuçlar vermemesidir. Tablo 4.6 dinükleotidlerin entalpi değerlerini verilmektedir.

**Tablo 4. 7.**  Dinükleotidlerin termodinamik entalpi değerleri

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| AA | 9.1 | GA | 5.6 | CA | 5.8 | TA | 6.0 |
| AC | 6.5 | **GC** | 11.1 | **CC** | 11.0 | **TC** | 5.6 |
| AG | 7.8 | **GG** | 11.0 | **CG** | 11.1 | **TG** | 5.8 |
| AT | 8.6 | **GT** | 6.5 | **CT** | 7.8 | **TT** | 9.1 |

### Genetik Kod Bağlam Tabanlı Kodlama Tekniği (GCC)

Genetik kod bağlamı ile sayısallaştırmada DNA dizisi üçlü kodonlar şeklinde okuma çerçevesi ile okunur. Okuma çerçevesi her turda bir baz kaydırılarak diziyi tekrar okur ve kodonlardan oluşan amino asitleri elde eder. Her amino asit tabloda verilen benzersiz bir karmaşık sayı ile kodlanır [18]. Tablo 4.7, 20 amino asitin GCC tabanlı sayısal değerlerini vermektedir. Örneğin bir X= AGCTACCGTGT dizisi için okuma çerçevesi ile elde edilen kodonlar aşağıdaki gibidir:

Birinci okuma çerçevesi AGC TAC CGT → [S, Y, R] amino asitleri

İkinci okuma çerçevesi GCT ACC GTG → [A, T, V] amino asitleri

Üçüncü okuma çerçevesi CTA CCG TGT → [L. P, C] amino asitleri

Amino asitlerin karmaşık sayı değerleri ile kodlanması ile elde edilen sayısal dizi:

[0.05 + 88.7i, 0.6 + 88.3i, 1.88 + 193i, 0.06 + 125.1i, 0.60 + 181.2i, 1.32 + 141.4i] şeklindedir. Şekil 4.10, NR\_131216.1 numaralı referans diziliminin GCC tabanlı kodlama tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.

**Tablo 4. 8.** Yirmi amino asitin GCC tabanlı kodlama değerleri

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Bazlar | Amino Asit | Sayısal Temsil | Bazlar | Amino Asit | Sayısal Temsil |
| GCU, GCC,  GCA, GCG | A – Alanin (Ala) | 0.61 + 88.3i | **AUG** | M –Metiyonin (Met) | 1.18 + 162.2i |
| UGU, UGC | C – Sistein (Cys) | 1.07 + 112.4i | **UAU, UAC** | Y – Tirozin (Tyr) | 1.88 + 193i |
| GAU, GAC | D – Aspartik Asit (Asp) | 0.46 + 110.8i | **UGG** | W – Triptofan (Trp) | 2.65 + 227i |
| GAA, GAG | E – Glutamik Asit (Glu) | 0.47 + 140.5i | **GUU, GUC,**  **GUA, GUG** | V – Valin (Val) | 1.32 + 141.4i |
| UUU, UUC | F – Fenilalanin (Phe) | 2.02 + 189i | **CCU, CCC,**  **CCA, CCG** | P – Prolin (Pro) | 1.95 + 122.2i |
| GGU, GGC,  GGA, GGG | G – Glisin (Gly) | 0.07 + 60i | **AAU, AAC** | N – Asparajin (Asn) | 0.06 + 125.1i |
| CAU, CAC | H – Histidin (His) | 0.61 + 152.6i | **CAA, CAG** | Q – Glutamin (Gln) | 148.7i |
| AUU, AUC,  AUA | I – İzolösin (Ile) | 2.22 + 168.5i | **CGU, CGC,**  **CGA, CGG,**  **AGA, AGG** | R – Arjinin (Arg) | 0.60 + 181.2i |
| AAA, AAG | K – Lizin (Lys) | 1.15 + 175.6i | **AGU, AGC,**  **UCU, UCC,**  **UCA, UCG** | S – Serin (Ser) | 0.05 + 88.7i |
| UUA, UUG,  CUU, CUC,  CUA, CUG | L – Lösin (Leu) | 1.53 + 168.5i | **ACU, ACC,**  **ACA, ACG** | T – Treonin (Thr) | 0.05 + 118.2i |



**Şekil 4. 10.** GCC kodlamanın sayısal temsili

### Walsh Koduna Dayalı Kodlama Tekniği

Walsh kodlama tekniğinde, k dereceli walsh kodunun temel tanımı Denklem 3.3’te verilmektedir.

(4. 3)

Burada *k = 2i, i =* {1, 2, 3, …} ve , *Wi ve W1 =* {1 veya 0}’ın tümleyenidir. i= 1 ve tohum matrisi *W1* =0 ise order-2 nin WC’ları Denklem 3.4 şeklindedir.

W2 = (4. 4)

Çekirdek matrisi *W2*, order-4 WC elde etmek için tekrarlanır. i = 2 ve k = 4 için order-4 WC kodları Denklem 3.5 şeklindedir. İkili kodların 64 farklı perfütasyonu üretilir.

(4. 5)

Denklem 3.3 ve 3.4 ‘deki gibi 4. Dereceden WC’ler belirlenir. Olası 64 kombinasyondan dört tane keyfi WC seçilir ve Denklem 3.6 ‘daki gibi nükleotidlere atanır.

(4. 6)

Walsh kodlama yöntemi WCBNE trigonometrik, karmaşık sayı, ikili, hamming mesafesi tabanlı kodlama gibi sayısallaştırma yöntemleri ile karşılaştırıldığında ekzon bölgelerini belirlemede daha yüksek performans göstermektedir [10]. Ayrıca walsh kodlar iyi kolerasyon ve otogonal özellik sağlamaktadır.

NR\_131216.1 numaralı referans numaralı dizilimlerin ilk 100 bazının “Biyokimyasal ve Fizikokimyasal Özellikli (BFÖ)” grubundaki kodlama teknikleriyle sayısallaştırılmış sinyal gösterimleri Şekil 4.11 ‘de gösterilmektedir.



**Şekil 4.11.**  Biyokimyasal ve fizikokimyasal özellikli (BFÖ) grup tekniklerinin sayısal gösterimleri

Tablo 4.8, BFÖ grubundaki haritalama tekniklerinin tamamının kısa bir özet bilgisini vermektedir.

**Tablo 4. 9.** Biyokimyasal ve fizikokimyasal özellikler grubundaki tüm sayısal kodlama tekniklerinin özeti

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tekniğin Adı | | Kodlama Şeması | Sayısal Temsil | Tenım |
| EIIP kodlama | C=0.1340, T=0.1335,  A=0.1260, G=0.0806 | | *X*=[AGCTACCGTG]  = [0.1260, 0.0806, 0.1340, 0.1335, 0.1260, 0.1340, 0.1340, 0.0806, 0.1335, 0.0806] | Nükleotidlere enerji değerleri atanır. |
| Entegre EIIP kodlama | EIIP codes for 64 codons | | *X*=[AGCTACCGTG]  = [0.3406, 0.3481, 0.3935, 0.3935, 0.3940, 0.3486, 0.3935, 0.2947] | DNA kodonlarına EIIP enerji değerleri atanır. |
| Atom numarası kodlama | C= 58, T= 66,  A= 70, G= 78 | | *X*=[AGCTACCGTG]  = [70, 78, 58, 66, 70, 58, 58, 78, 66, 78] | Nükleotidlere atom numaraları atanır. |
| Eşleştirilmiş atom numarası kodlama | A&G= 62, C&T= 42 | | *X*=[AGCTACCGTG]  = [62, 62, 42, 42, 62, 42, 42, 62, 42, 62] | Eşleştirilmiş nükleotidlere atom numaraları atanır. |
| Moleküler kütle kodlama | C= 110, G= 150,  A= 134, T= 125 | | *X*=[AGCTACCGTG]  = [134, 150, 110, 125, 134, 110, 110, 150, 125, 150] | Nükleotidlere moleküler kütle değerleri atanır. |
| Entropik bölümleme kodlama | 12-Symbol alphabet  A1, A2, A3, C1, C2,C3, G1, G2, G3, T1, T2, T3  Diziye göre entropi hesaplanır | | *X*=[AGTTAGTGCT]  = [A1 G2  S3 T3 S1 T1 A2 G3 T1 G2 C3] | DNA segmentleri ve stop kodonları 18 sembollü alfabe ile temsil edilir. |
| Otoregresif kodlama | Dinükleotidler için Propeller Twist ve DNA Bending Stiffness values | | *X*=[AGCTACCGTG]  Propeller Twist  = [-14.00, -11.08, -14.00, -11.85, -13.10, -8.10, -10.03, -13.10, -9.45]  Bending Stiffness  = [60, 85, 60, 20, 60, 130, 85, 60, 60] | DNA’nın yapısal özelliklerine göre pervane bükümü ve DNA bükülme sertliği değerleri ile dinükleotidler kodlanır. |
| Dört yapısal özellik kodlama | Dinükleotidler için DNA Bending Stiffness, Dublex Disrupt Energy, Dublex Free Energy, Propeller Twist values | | *X*=[AGCTACCGTG]  DNA bending stiffness  60, 60, 60, 85, 60  Dublex disrupt energy  16, 16, 13, 36, 19  Dublex free energy  -15, -15, -15, -28, -17  Propeller twist  -1400, -1400, -1310, -1003, -94 | DNA’nın dört fiziksel özelliğine göre dinükleotid; pervane büküm değeri, DNA bükülme sertliği, dubleks disrupt enerji, dubleks serbest enerji değerleri ile kodlama gerçekleştirilir. |
| Thermodinamik kodlama | TC=5.6, GA=5.6, CA=5.8, TG=5.8, TA=6.0, AC=6.5, GT=6.5, CT=7.8, AG=7.8, AT=8.6, TT=9.1, AA=9.1, CC=11.0,GG=11.0, GC=11.1, CG=11.9 | | *X*=[AGCTACCGTG]  = [7.8, 11.1, 7.8, 6.0, 6.5, 11.0, 11.9, 6.5, 5.8] | Nükleotidlerin termodinamik etkileşimlerinin entalpi değerleri atanarak kodlama gerçekleştirilir. |
| Genetik kod bağlam (GCC) tabanlı kodlama | Amino asitlere GCC temelli karmaşık sayı temsillerinin atanması (Tablo 4.7) | | *X*=[AGCTACCGTG]  Birinci çerçeve AGC TAC CGT  İkinci çerçeve GCT ACC GTG  = [0.05 + 88.7i, 0.6 + 88.3i, 1.88 + 193i, 0.06 + 125.1i, 0.60 + 181.2i, 1.32 + 141.4i] | DNA dizisi üçlü kodonlar şeklinde okuma çerçevesi ile okunur. Aminoasitlere karmaşık sayı değerleri atanarak dizi sayısallaştırılır. |
| Walsh koduna dayalı kodlama | A=WA=0000  T=WT=0011  G=WG=0101  C=WC=0110 | | *X*=[AGCTACCGTG]  = [00000101 011000110000  011001100101 00110101] | Nükleotidlere dördüncü dereceden walsh kodları atanır. |

## İkili ve Bilgi Kodlama Teknikleri

DNA sayısal haritalama tekniklerinin üçüncü grubu, ikili ve bilgi kodlama (İBK) (BIE – Binary and Information Encoding) teknikleridir. Bu grup içerisinde Voss Kodlama, Tek-Sıcak Kodlama, Patojenite Adaları Kodlama, Gradyan Kaynak Yerelleştirme Kodlama, 2-bit İkili Kodlama, Hata Düzeltma Kodu Kodlama (ECC), I Ching Kodlama, Galois Alanı Kodlama, Gray Kod Kodlama ve K-mer Kodlama olmak üzere on tane kodlama şeması incelenmektedir.

### Voss Kodlama Tekniği

Voss temsilinde genom dizilerini temsil etmek için dört ayrı ikili dizi oluşturulur. Her baz için oluşturulan dizilerde ilgili bazın olduğu dizi konumuna 1, diğer konumlara 0 değeri atanır. Bu tekniğin avantajı Fourier tekniğinde protein kodlayan bölgelerdeki 3-periyodiklik gibi kısa menzilli kolerasyonların tespitinde oldukça yüksek performans göstermesidir [11]. Kullanımı basit ve kolay olmasına rağmen, bu tekniğin dezavantajı dört dizi elde edildiği için hesaplama karmaşıklığının fazla olmasıdır. Şekil 4.12, örnek diziliminVoss kodlama tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 12.** Voss kodlama tekniğinin sayısal gösterimi

### Tek-Sıcak Kodlama Tekniği

Tek-Sıcak (One-hot) kodlama dört bit ikili kodlama olarak da adlandırılmaktadır. A, T, C, G nükleotidlerine dört bitlik ikili tabanda A=1000, T= 0100, C= 0010, G= 0001 değerlerin atanması ile gerçekleştirilmektedir. Bu tekniğin avantajı, genomlardaki belirli bölgelerin tespitinde, örneğin fare genomundaki 4mC konumlarının sınıflandırılması, kodlama bölgelerinin tespiti, sayısallaştırmaya ihtiyaç duyulan çeşitli alanlarda çok kolay bir kullanım sağlamasıdır [37].

### Patojenite Adaları Kodlama Tekniği

Pathogenicity Island kodlama tekniğinde, C ve G bazlarına 1 değeri atanır, A ve T bazlarına ise 0 değeri atanarak kodlama gerçekleştirilmektedir. Bu tekniğin avantajı genom dizilerinde G+C modelleri önemli ölçüde saptar [14]. Ekzon bölgelerinde kodlama bölgelerinin tespitinde iyi sonuçlar vermektedir [38].

### Gradyan Kaynak Yerelleştirme Kodlama Tekniği

Gradyan source localization kodlama tekniği, gösterge dizisi tasarlarken biyolojik olarak esinlenilen gradyan kaynağı lokalizasyonu göz önünde bulundurularak oluşturulmuştur. DNA dizisindeki nükleotidlere A=0, C= 1, G= 3, T= 2 değerleri verilmektedir [14].

### 2-bit İkili Kodlama Tekniği

Bu kodlama tekniğinde, dizisinin nükleotidlerine A= 00, T= 01, G= 10, C= 11 gibi iki bitlik ikili değerler atanarak kodlama gerçekleştirilir. Bu tekniğin avantajı kullanımı çok kolay olmasıdır [18].

### Hata Düzeltme Kodu Kodlama Tekniği

Hata düzeltme kodu (ECC- Error correction code) kodlama tekniği, birçok farklı şekli geliştirilmiştir. Pürin-pirimidin, güçlü-zayıf H bağı, amino-keto gruplarına göre kodonlara x ve y koordinat değerleri atanarak kodlama gerçekleştirilir. Bu tekniğin avantajı, genom kodlamasının doğasını, genom fazlalığını ve gen mutasyonlarının tespitini iyi yansıtmaktadır[11], [39].

### I Ching Kodlama Tekniği

I Ching kodlama tekniği, simetri ve periyodiklik modellerine dayanan bir genetik koddur. Nükleik asitlerin üç biyokimyasal özelliği olan H-bağları, pürin-pirimidin halkaları ve keto-enol/amino-imino tautomerine dayalı olarak düzenlenmiştir. Yirmi amino asit 43 = 64 kodon içeren I Ching tabloları ile doğrudan eşlenebilir. Bazı kodonlar aynı amino asitleri temsil eder [11]. Bazlar, 3 farklı eksende değerler almaktadır [40].

1 eksen x: 2-H bağı = 0 = A, T; 3-H bağı = 1 = C, G; y ekseni: Purin =0=A, G; Pirimidin=1=T, C

2 eksen x: Keto/Amino (G = T = 0, A = C = 1); y ekseni: Pur/ Pyr (A = G = 0, C = T = 1)

3 eksen x: H-bağları (A = T = 0, C = G = 1); y ekseni: Keto/Amino (G = T = 0, A = C = 1)

Şekil 4.13, örnek diziliminIChing kodlama tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 13.** Ching kodlama tekniğinin sayısal gösterimi

### Galois Alanı Kodlama Tekniği

Galois Field tabanlı kodlama tekniği, DNA’nın nüklotidlerinden oluşan dizinin doğrusal bir dizi olarak sayısallaştırılması gerçekleştirmektedir. Dizi (n, k) otogonal kod sözcüklerine dönüştürülür. Galois alanı ile bazlar dört tane sonlu GF alanına etiketlenmiş olur. A, T, G, C bazlarının bu teknik ile polinom temsillerine karşılık gelen sayısal atamaları Denklem 3.7 ‘de gösterilmektedir[41]. Bu tekniğin avantajı DNA dizilimlerinin kolay bir şekilde sayısallaştırmasıdır.

0 = 0 ⇔ 0 ⇔ A

x0 = 1 ⇔ 1 ⇔ C (4. 7)

x1 = x ⇔ 2 ⇔ T

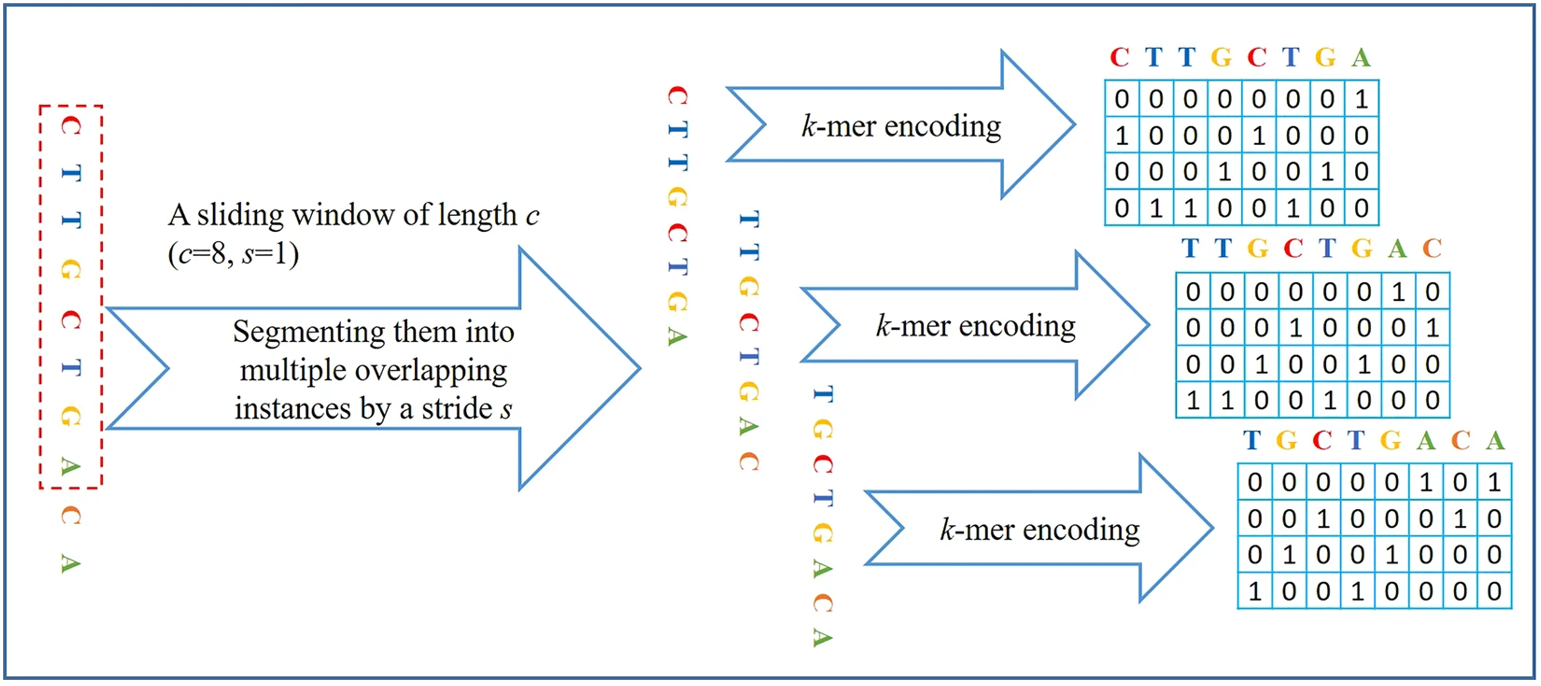
x2 = x + 1 ⇔ 3 ⇔ G

### Gray Kod Kodlama Tekniği

Bu kodlama tekniğinde, tam sayı değerleri her baza atanarak ikili sembollerle ifade edilmektedir. Değerler atanırken bazı kurallara dikkat edilmesi gerekmektedir. Bu kurallar kodların aynı olmaması veya bir sette tekrar etmemesi, herhangi bitişik iki kod bir bit ile ayrılmamasıdır. Yani Hamming mesafesi 1 olmalıdır. Bu nedenle bu kodlara tek mesafeli kodlar denir[41]. Bu tekniğin avantajı protein kodlayan bölgelerin tespitinde tahmin doğruluğunu arttırmasıdır.

### K-mer Kodlama Tekniği

K-mer kodlama tekniği, bir diziyi k bazlı çoklu alt dizilere bölmek için kullanılmaktadır. Tüm DNA dizisi bir cümle kabul edilirse, k-mer segmentleri cümleyi oluşturan kelimelerdir. 1 adım boyutu ile, *l* tabanlı bir dizi (*l* – k + 1) k-mer’e ayrılır. Örneğin, AGCCT dizisi üç 3-mer’e ayrılır: (AGC, GCC, CCT) [37], [42]. Şekil 4.14, k-mer kodlama tekniğinin çalışma mantığını göstermektedir.



**Şekil 4. 14.** K-mer kodlamanın temsili [42]

Şekil 4.15, NR\_131216.1 referans numaralı dizilimlerin ilk 100 bazının “İkili ve bilgi kodlama” grubundaki kodlama teknikleriyle sayısallaştırılmış sinyal gösterimlerini vermektedir.



**Şekil 4. 15.** İkili ve bilgi Kodlama (İBK) grubu tekniklerinin sayısal gösterimleri

Tablo 4.9, “Binary and Information Encoding” grubundaki haritalama tekniklerinin tamamının kısa bir özet bilgisini vermektedir.

**Tablo 4. 10.** İkili ve bilgi kodlama grubundaki tüm sayısal kodlama tekniklerinin özeti

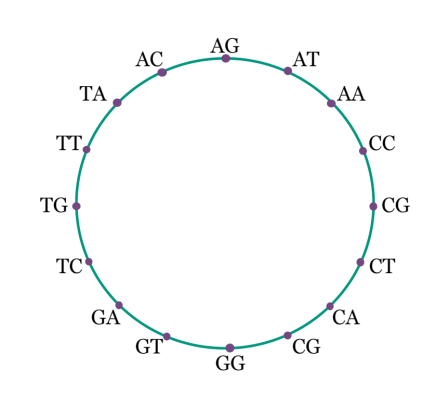
|  |  |  |  |
| --- | --- | --- | --- |
| Tekniğin Adı | Kodlama Şeması | Sayısal Temsil | Tanım |
| Voss kodlama | S=[C, G, A, T],  Cn=[1,0,0,0],  Gn=[0, 1, 0, 0],  An=[0, 0, 1, 0],  Tn=[0, 0, 0, 1] | *X*=[AGCTACCGTG]  A20 = [1000100000]  G20 = [0100000101]  C20 = [0010011000]  T20 = [0001000010] | Dört ayrı dizi oluşturularak her nükleotidin bulunup bulunmamasına göre ikili değerler atanır. |
| Tek-Sıcak kodlama | A:1,0,0,0  T:0,1,0,0  C:0,0,1,0  G:0,0,0,1 | *X*=[AGCTACCGTG]  = [1000, 0001, 0010, 0100, 1000, 0010, 0010, 0001, 0100, 0001] | Nükleotidler dört bit ikili değerler ile kodlanır. |
| Patojenite adaları kodlama | C&G= 1, A&T= 0 | *X*=[AGCTACCGTG]  = [0110011101] | Patojenite adalarının varlığı ve yokluğuna göre ikili değerler atanır. |
| Gradya kaynak yerelleştirme kodlama | A= 0, C=1, G=3, T= 2 | *X*=[AGCTACCGTG]  = [0, 3, 1, 2, 0, 1, 1, 3, 2, 3] | Gradyan kaynağı lokalizasyonuna göre nükleotidlere tam sayı değerleri atanır |
| 2-bit ikili kodlama | A= 00, G= 10,  T= 01, C= 11 | *X*=[AGCTACCGTG]  = [00, 10, 11, 01, 00, 11, 11, 10, 01, 10] | Nükleotidlere iki bit ikili değerler atanır. |
| Hata düzeltme kodu (ECC) kodlama | Purine-Prymidine, Weak-Strong H bond, Amino-Keto group kodonlarına göre x ve y koordinat değerleri | *X*=[AGCTACCGTG]  Purine-Prymidine  = [(1,1), (2,6), (3,5), (4,3), (5,6), (6,5), (7,3), (8,7)]  Weak-Strong H bond  = [(1,1), (2,2), (3,4), (4,7), (5,1), (6,3), (7,2), (8,5)]  Amino-Keto  = [(1,3), (2,7), (3,3), (4,6), (5,4), (6,5), (7,2), (8,0)] | Nükleotidlere biyokimyasal özelliklerine göre x ve y koordinat değerleri atanır |
| IChing kodlama | Amino asitlere göre 3 farklı I Ching tablosu ile binary kodlama | *X*=[AGCTACCGTG]  = [110 100 001 010 100 001 010 101] | Nükleik asitlerin üç biyokimyasal özelliklerine göre oluşturulan I Ching tabloları ile kodlama gerçekleştirilir. |
| Galois alanı kodlama | 0=0 ⬄ 0 ⬄A,  x0=1 ⬄1⬄C,  x1=x ⬄ 2 ⬄T,  x2=x+1 ⬄ 3 ⬄ G | *X*=[AGCTACCGTG]  = [0, 3, 1, 2, 0, 1, 1, 3, 2, 3] | Nükleotidlere ikinci dereceden polinom temsillerine karşılık gelen sayısal değerler atanır. |
| Gray kod kodlama | A= 00, T=01,  C=10, G= 11 | *X*=[AGCTACCGTG]  = [0010011100  0101101110] | Nükleotidlere iki bit ikili kodlar ex-or işlemi yapılarak atanır. |
| K-mer kodlama | 1-mer coding  A → [1,0,0,0]  C → [0,1,0,0]  G → [0,0,1,0]  T → [0,0,0,1] | *X*=[AGCTACCGTG]  = [1000, 0010, 0100, 0001, 1000, 0100, 0100, 0010, 0001, 0010] | DNA dizisi k-mer degmentlere ayrılarak sıfır ve birlerle kodlanır. |

## Birincil Yapı Özellikli Kodlama Teknikleri

DNA sayısal haritalama tekniklerinin dördüncü grubu, birincil yapı özellikteki (BYÖ) (PSP – Primary Structure Properties) sayısallaştırma teknikleridir. Bu grup içerisinde Dinucleotide Mesfesi Kodlama, Halka Yapısı Temsili Kodlama, Üçlü Kodlama, Nükleotid Oluşum Sıklığı Kodlama, Entropi Tabanlı Kodlama, Nükleotidler Arası Mesafe Kodlama olmak üzere altı tane kodlama şeması incelenmektedir.

### Dinükleotid Mesafesi Kodlama Tekniği

Dinükleotid gösteriminde iki komşu bazın birleşiminden oluşan dinükleotid setlerine, bir birim dairenin çevresine yerleştirilen dinükleotidlerin açı değerleri ile kodlama gerçekleştirilir. DNA’nın A, T, G, C bazlarının yan yana gelmesi ile 16 tane ikili dinükleotid seti oluşturulur: {AA, AT, AG, AC, TA, TT, TG, TC, GA, GG, GC, GG, CA, CT, CC, CT}. İkili nükleotidleri birim çemberin çevresine eşit olarak dağıtmak için “magic circle” yaklaşımı oluşturulmuştur [43], [44]. Birim çemberin çevresine eşit olarak dağıtılan dinükleotidlere bulunduğu çember konumunda karşılık gelen kutup açısı ile sayısal değeri atanır [43]. Şekil 4.16, birim çemberdeki dinükleotidlerin yerleşimini vermektedir. Tablo 4.10, birim çember etrafındaki ikili nükleotidlerin açı değerlerini göstermektedir. Şekil 4.17, örnek diziliminDinükleotid mesafesi kodlama tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir. Bu tekniğin avantajı ikili dinükleotid yapısının kullanıldığı tüm genomik alanlarda avantaj sağlar. Dezavantajı ise sayısallaştırmada DNA’nın yapısını çok iyi yansıtmamasıdır.



**Şekil 4. 16.** Birim çemberde dinükleotitlerin düzenlenmesi

**Tablo 4. 11.** Birim çemberdeki dinükleotitlerin açı değerleri

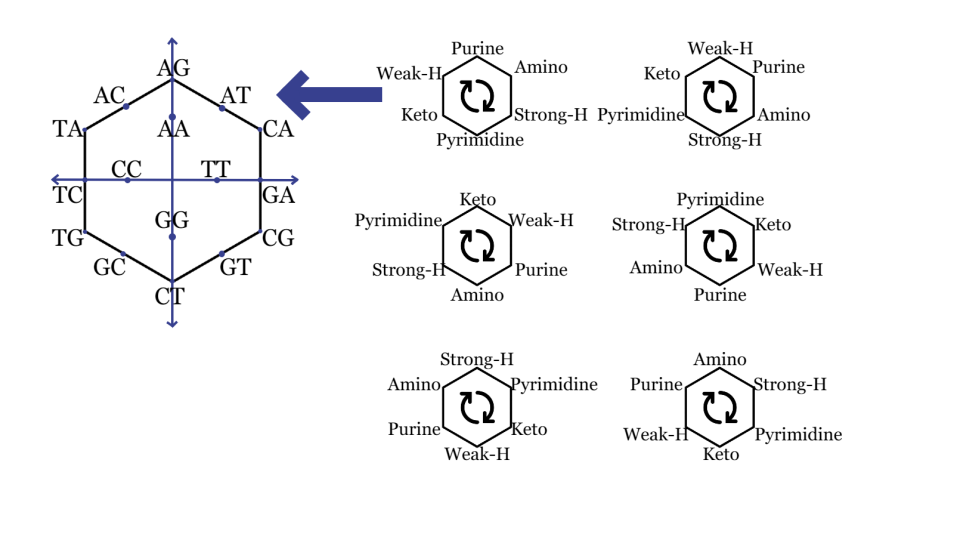
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| AA | π/4 | GA | 5π/4 | CA | 7π/4 | TA | 3π/4 |
| AC | 5π/8 | **GC** | 13π/4 | **CC** | π/8 | **TC** | 9π/8 |
| AG | π/2 | **GG** | 3π/2 | **CG** | 2π | **TG** | π |
| AT | 3π/8 | **GT** | 11π/8 | **CT** | 15π/8 | **TT** | 7π/8 |



**Şekil 4. 17.** Dinükleotid mesafesinin sayısal gösterimi

### Halka Yapısı Temsili Kodlama Tekniği

Halka yapısı temsili, dinükleotid kodlamasının farklı bir çeşididir. DNA bazlarının halka yapıları ve bunlara karşılık gelen moleküler kütleleri kullanılarak kodlama gerçekleştirilmektedir. Pürin-pirimidin, amino-keto, güçlü hidrojen bağı-zayıf hidrojen bağı sınıflarının farklı sıralarda köşelere yerleştirilmesi ile altı farklı kombinasyon elde edilmektedir. Şekil 4.18’de her çizim farklı bir kodlama sistemine karşılık gelmektedir [11]. Nükleotidin kimyasal özelliğinin ve konum bilgisinin korunması için altıgenin yüzeyine 16 dinükleotid yerleştirilir. Bu tekniğin avantajı, DNA dizilerinde benzerlik karşılaştırılmasında yüksek doğrulukta sonuçlar vermesidir. Ayrıca farklı türler arasındaki evrimsel ilişki tespiti için filogenetik analizde iyi sonuçlar vermektedir [45]. Tablo 4.11’de, Şekil 4.18’da bulunan ilk altıgendeki grupların yerleşmesine göre kodlama değerleri bulunmaktadır.



**Şekil 4. 18.** Kodlama sistemi

**Tablo 4. 12.** Birinci halka yapısı kodlama değerleri

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| AA | (0, 1) | GA | (1, 0) | CA | (1,1) | TA | (-1, 1) |
| AC | (-0.5, 1.25) | **GC** | (-0.5,-1.25) | **CC** | (-0.5, 0) | **TC** | (-1, 0) |
| AG | (0, 1.5) | **GG** | (0, -1) | **CG** | (1, -1) | **TG** | (-1, -1) |
| AT | (0.5, 1.25) | **GT** | (0.5,-1.25) | **CT** | (0,-1.5) | **TT** | (0.5, 0) |

Şekil 4.19, örnek diziliminRing Structure kodlama tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 19.** Halka Yapısı Temsil Kodlamasının sayısal gösterimi

### Üçlü Kodlama Tekniği

Üçlü kodlama, nükleotid üçlülerinin ve amino asit kodonlarının özelliklerine dayanarak ağırlıklara göre sayısallaştırma gerçekleştirmektedir. Diziler arasındaki mesafe ölçülmektedir. Birinci kodon (X1, Y1) ile ikinci kodon (X2, Y2) ile kodlanırsa aralarındaki mesafe Denklem 3.8’de gösterildiği gibi hesaplanmaktadır.

| ψ(X1) – ψ(Y1) | <|ψ(X2) – ψ(Y2) | (4. 8)

Burada *ψ* kodonların ağırlık haritalamalarıdır. Ağırlık değerleri tam sayı ve kesirli sayı kısımları olarak iki bölümden oluşmaktadır. Tam sayı kısmı amino asidi, kesirli sayı kısmı kodonu temsil eder. Örneğin Alanine amino asidi tam sayı olarak 1 ile ifade edilir. Alanine’in ilk kodonu GCT 1.1 ağırlığa sahipken ikinci kodonu GCC 1.2 ağırlığa sahiptir. Kodonlar bu şekilde ağırlık değerleri ile sayısallaştırılmaktadır [11], [46]. Bu tekniğin avantajı diziler arası benzerliğin hesaplanmasında oldukça iyi performans göstermesidir.

### Nükleotid Oluşum Sıklığı Kodlama Tekniği

Nükleotid oluşum sıklığına göre kodlama tekniği, DNA nükleotidinin ekzon ve intron gibi bölgelerde farklı frekanslara sahip olması fikrinden yararlanılmaktadır. Nükleotid oluşum frekansları istatistiksel olarak hesaplanabilir kesirli değerler atanarak kodlama gerçekleştirilmektedir. Tek nükleotidlerin sıklığına göre sayısal değerler atanabileceği gibi dinükleotidlerin veya üçlülerin frekans değerleri de hesaplanarak sayısallaştırma için kullanılabilmektedir [11], [18].  Bu tekniğin avantajı protein kodlama bölgesinin tespiti gibi spesifik alanların belirlenmesinde iyi sonuçlar vermektedir. Şekil 4.20, örnek diziliminnükleotid oluşum sıklığı kodlama tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 20.** Nükleotid Oluşum Sıklığı Kodlamasının Sayısal Gösterimi

### Entropi Tabanlı Kodlama Tekniği

Entropiye dayalı sayısal haritalama tekniği Shannon entropisinin kesirli bir türevi olan yeni bir entropi tabanlı sayısallaştırmanın nükleotid dizilimi analizinde, ekzon, intron ve aminoasitlerin dağılımının analizinde ve birçok genomik alanda kullanılmasına dayanmaktadır. Kesirli entropi denkleminde (Denklem 3.9) α ve *p*(*xi*) değerlerinde değişiklikler yapılmıştır. DNA kodon dağılımları için entropi hesaplamaları yapılması ilkesine dayanmaktadır. Her kodon için entropi değeri 64 olası kodon arasındaki dengeyi yansıtmaktadır.

(4. 9)

Burada *p*(*xi*) DNA dizisindeki her kodonun tekrarlama frekansıdır ve α için Denklem 3.10’daki formül kullanılmaktadır.

(4. 10)

α değeri kodonun tekrarlama frekansının logaritmasının 1’e bölümüdür. Bu tekniğin avantajı uygulamalarda, ekzon bölgelerinin tespitinde, hastalık yapıcı genlerin teşhisinde ve filogenetik analizde yüksek doğruluk sağlamasıdır [47].

### Nükleotidler Arası Mesafe Kodlama Tekniği

Nukleotidler arası mesafe kodlama tekniği (Inter nücleotide distance coding) tekniğinde, DNA dizisindeki her bir baz, kendisinden sonra gelen aynı baz ile arasındaki taban mesafesinin değeri ile kodlanmaktadır. Örneğin dizideki ilk Adenin bazının kendisinden sonra gelecek Adenin bazı ile arasındaki mesafe değeri n ise ilk Adenin yerine n değeri verilir ve kodlama her benzer ifade için devam etmektedir. Eğer dizinin devamında benzer ifade yoksa o baz yerine kalan dizinin uzunluğu atanır [17]. Bu tekniğin avantajı peptit zincirini temsil etmek için kullanılırsa ölçülen mesafe değerleri proteinin katlanma noktaları üzerinde doğrudan etkili olmasıdır [48]. Dezavantajı ise, bu tekniğin DNA’nın bilgisini koruyan bir sayısallaştırma olmamasıdır. Şekil 4.21, örnek dizilimin, bu bölümde incelenennükleotidler arası mesafe kodlama tekniği, triplet kodlama tekniği ve entropi tabanlı kodlama tekniği ile sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 21.** Birincil yapı özellikli (BYÖ) grup tekniklerinin sayısal gösterimleri

Tablo 4.12, “Birincil Yapı Özellik” grubundaki haritalama tekniklerinin tamamının kısa bir özet bilgisini vermektedir.

**Tablo 4. 13.** Birincil yapı özellikleri grubundaki tüm sayısal kodlama tekniklerinin özeti

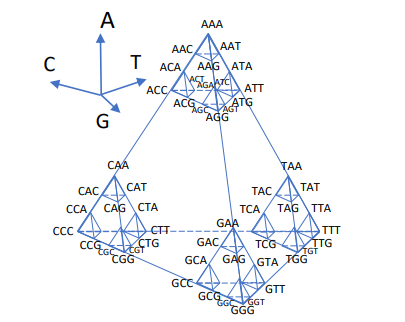
|  |  |  |  |
| --- | --- | --- | --- |
| Tekniğin Adı | Kodlama ŞEması | Sayısal Temsil | Tanım |
| Dinucleotid mesafesi kodlama | 16 dinükleotid birim çembere yerleştirilip konumlarına göre kodlanır | *X*=[AGCTACCGTG]  = [(cos(π/2), sin(π/2)),  (Cos(13π/4), sin(13π/4)),  (Cos(15π/8), sin(15π/8)),  (Cos(3π/4), sin(3π/4)),  (Cos(5π/8), sin(5π/8)),  (Cos(π/8), sin(π/8)), (cos(2π),  Sin(2π)), (cos(11π/8),  Sin(11π/8)), (cos(π), sin(π))] | Dinükleotidler bir çemberin çevresine eşit dağıtılarak bulundukları koordinat değerleri ile kodlanır. |
| Halka yapısı kodlama | AG: (0, 1.5), CT: (0, -1.5), CA:(1,1), TG: (-1, -1), CG: (1, -1), TA: (-1, 1), GA: (1, 0),GT(0.5, -1.25), GC: (-0.5, 1.25), TC: (-1, 0), AC: (-0.5, 1.25), AT: (0.5, 1.25), AA: (0, 1), TT: (0.5, 0), GG: (0,1), CC: (-0.5, 0). | *X*=[AGCTACCGTG]  = [(0, 1.5), (0,-1.5),  (-0.5, 1.25), (1, -1), (-1, -1)] | Dinükleotidler biyokimyasal özelliklerine göre ayrıldıkları altı gruba göre altıgenin köşelerine yerleştirilerek altı farklı kombinasyon ile altı kodlama şeması elde edilir ve altıgenin köşe koordinatları ile kodlanır. |
| Üçlü kodlama | 64 kodon ağırlıklarına göre kodlanır | *X*=[AGCTACCGTG]  = [15.6, 1.1, 11.3, 18.2, 16.4, 14.4, 2.1, 19.4] | Nükleotid üçlüleri ve amino asit kodonları ağırlıklarına göre sayısallaştırılır. |
| Nükleotid oluşum sıklığı kodlama | C=0.27215,T=0.2056, A=0.24300,G=0.27909 veya  CG:0.01, GC: 0.043, CC: 0.047, GT:0 .049, GG: 0.050, AC: 0.054, TC: 0.057, GA: 0.061, TA:0.067, AG: 0.070, CT: 0.071, TG: 0.074, CA: 0.074, AT: 0.081, AA: 0.097, TT: 0 .097 | *X*=[AGCTACCGTG]  = [0.070, 0.071, 0.054, 0.01, 0.074] | Nükleotidler veya dinükleotidler oluşum sıklıklarına göre frekans değerleri ile kodlanır. |
| Entropi tabanlı kodlama | 64 kodona yeni formüllere göre hesaplanan entropi değerleri atanır | *X*=[AGCTACCGTG]  = [0.7222, 0.8331, 0.9086, 0.8331, 0.8118, 0.5363, 0.6259, 0.9818, 0.9998, 0.9954] | Kodonlar Shannon entropi denkleminin değiştirilmiş ve kesirli yeni denklemi entropi değerleri hesaplanarak kodlanır. |
| Nükleotidler arası mesafe kodlama | Her baz sonraki kendisiyle aynı baz arasındaki taban mesafesi değeri ile kodlanır | *X*=[AGCTACCGTG]  = [4, 6, 3, 5, 5, 1, 3, 2, 1, 0] | DNA dizisindeki her baz kendinden sonra gelen aynı baz ile arasındaki taban mesafesi değeri ile kodlanır. |

## Grafiksel Kodlama Teknikleri

DNA sayısal haritalama tekniklerinin beşinci grubu, grafiksel olarak temsil edilen (GT) (GR – Graphical Representation) sayısallaştırma teknikleridir. Bu grup içerisinde Tetrahedron Kodlama, H-Eğrisi Kodlama, Z-Eğrisi Kodlama, Kuaternion Kodlama, SNP-GIN Kodlama, Kaos Oyunu Temsili Kodlama (CGR), Kaos Oyunu Temsili Yürüyüş (CGR-Yürüyüş), Tamsayı Kaos Oyunu Temsili Kodlama(iCGR), Kendi Kendini Düzenleyen (SOM) Haritalar Tabanlı Kodlama, Fermat Spiral Eğrisi Kodlama, Spektral Dinamik Kodlama, 2D Dinamik Kodlama, 3D Dinamik Kodlama ve 8D Vektör Kodlama olmak üzere on dört tane kodlama şeması incelenmektedir.

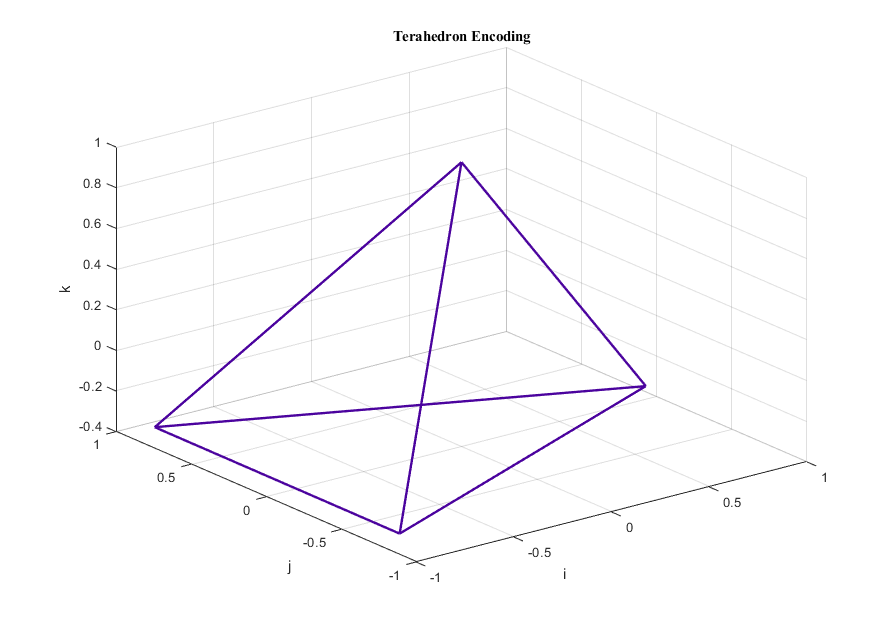
### Tetrahedron Kodlama Tekniği

Tetrahedron kodlama tekniğinde, her nükleotid düzenli bir tetrahedronun dört köşesinden birine yerleştirilmektedir. A, G, C, T bazlarının sayısal temsili Şekil 4.22’de 3D koordinat sistemi olarak temsil edilmektedir [11]. Nükleotidlerin sayısal temsilleri için A *= k, G = C = , T =*  şeklindedir. Şekil 4.22’de temsil edildiği gibi kodonlar için de tetrahedron ile sayısallaştırma gerçekleştirilebilir [11].



**Şekil 4. 22.** Tetrahedron temsili [11]

Bu tekniğin avantajı, biyomoleküler dizilerin DNA spektrumunun elde edilmesi uygulamalarında başarı elde etmektedir. Elde edilen spektrumlar bazlar için lokal frekans bilgilerini vermektedir [2]. Bu tekniğin dezavantajı ise 5000 bazdan uzun DNA dizileri için uygun olmamasıdır. Şekil 4.23, örnek dizilimintedrahedron kodlama tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 23.** Tetrahedron kodlamanın sayısal gösterimi

### H-Eğrisi Kodlama Tekniği

H eğrisi kodlama tekniği, DNA dizisindeki her nükleotidinin x, y ve z eksenlerinde bulunan i, j, k birim vektörleriyle oluşturulan fonksiyonların 3D temsilini sağlamaktadır. H eğrisi Denklem 3.11’ de tanımlanmaktadır.

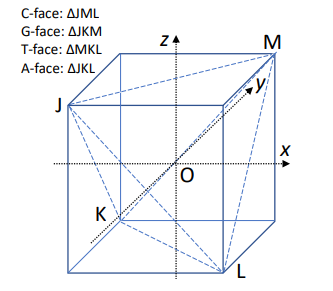
h(z) = (4. 11)

Burada *z* ∊ {C, T, A, G} ve *n* dizi uzunluğudur. Baz fonksiyonları ise sırasıyla *gw(A) = i + j – k,* *gw(T) = i – j – k,* *gw(C) = –i – j –k,* *gw(G) = –i + j – k* şeklindedir. H eğrisinin uç noktaları, dizinin nükleotid birleşimini tanımlamaktadır. Bu tekniğin avantajı, çeşitli gen yapılarını görüntülemek, karşılaştırmak ve sıralamak için basit bir yol sağlamasıdır. Dezavantajı ise, bu teknikte koordinatları hesaplamanın karmaşık olmasıdır. Bir döngü veya devre oluşma ihtimali yüksektir. Diziler benzersiz bir şekilde tanımlanmadığından bilgi kaybı gerçekleşebilir. Bu kaybın önlenmesi için kodlama fonksiyonları 2D Kartezyen koordinat düzlemine indirgenmiştir. Düzenlenen kodlama için nükleotidlere atanan fonksiyonlar sırasıyla gw(A) = , gw(T) = , gw(C) = , gw(G) = şeklindedir. Burada *w,* dizideki bir nükleotidin konumudur ve *i* ve *j*, kartezyen düzlemdeki vektörlerdir. Bir dizinin benzersizliği bu sistem ile ifade edilebilir [11], [49].

### Z-Eğrisi Kodlama Tekniği

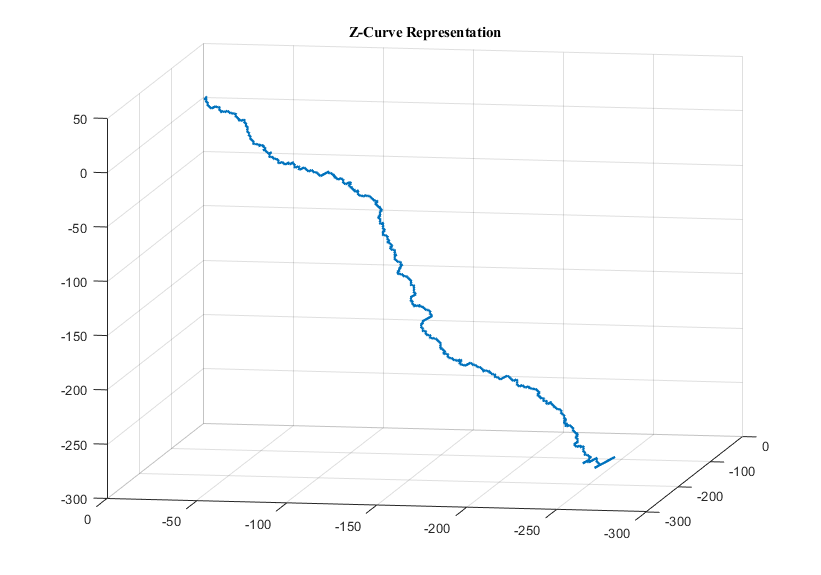
Z eğrisi gösterimi, benzersiz vektör setleri ile DNA dizilerini 3 boyutlu görselleştirmeyi ve analiz etmeyi sağlamaktadır {xn, yn, zn} bileşenleri, Z eğrisinde DNA dizisi nükleotidlerinin koordinatlarını temsil eder. An, Cn, Gn ve Tn nükleotidleri xn, yn, zn koordinat vektörlerine dönüştürülür. Şekil 4.24’de her yüzeye doğru olan vektörler nükleotidleri temsil etmektedir. Bu dört yöne hareket eden vektörlerin toplamı (n) nükleotid sayısına eşittir [11]. (xn, yn, zn) ve (An, Cn, Gn, Tn) arasındaki dönüşüm Equation 12 deki gibi temsil edilmektedir.

= + . (4. 12)



**Şekil 4. 24.** Z eğrisinin dörtyüzlü tabanlı koordinat sistemi [11]

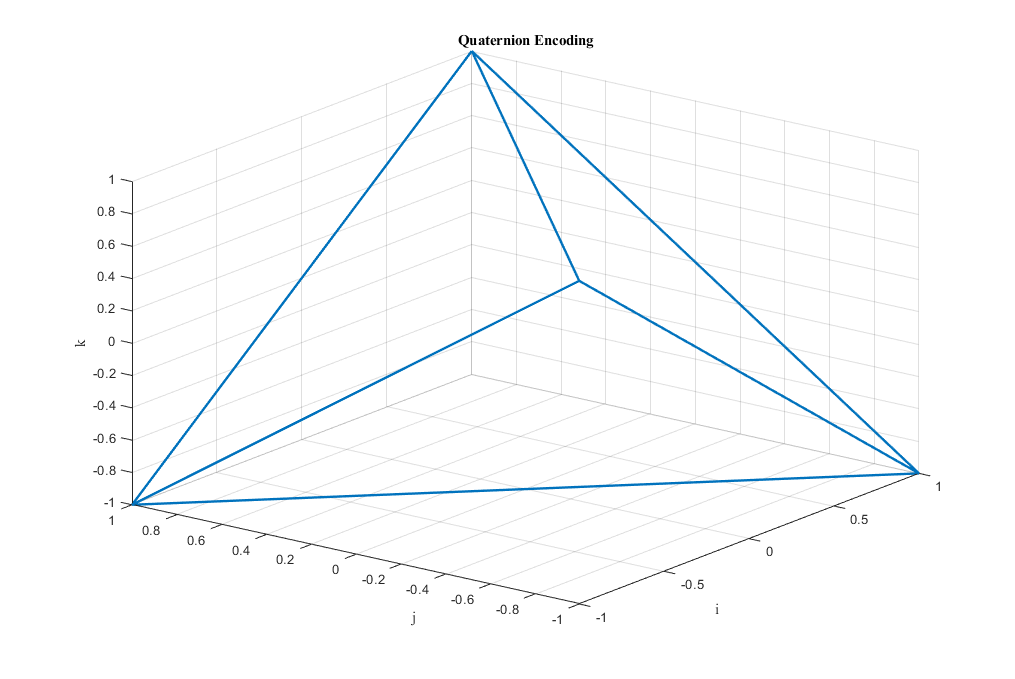
Bu tekniğin avantajı, herhangi bir dizi uzunluğuna uygulanabilmesi ve protein kodlama bölgelerinin ölçülmesi için DNA verisini eğriye dayalı bir dizi sinyale dönüştürerek yüksek doğrulukta tanımlanmasını sağlamasıdır [50]. Uzun menzilli kolerasyon gösteren intron bölgelerini belirlemede başarı sağlamaktadır [51]. Şekil 4.25, örnek diziliminZ eğrisi tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 25.** Z-Eğrisi gösteriminin sayısal gösterimi

### Kuaternion Kodlama Tekniği

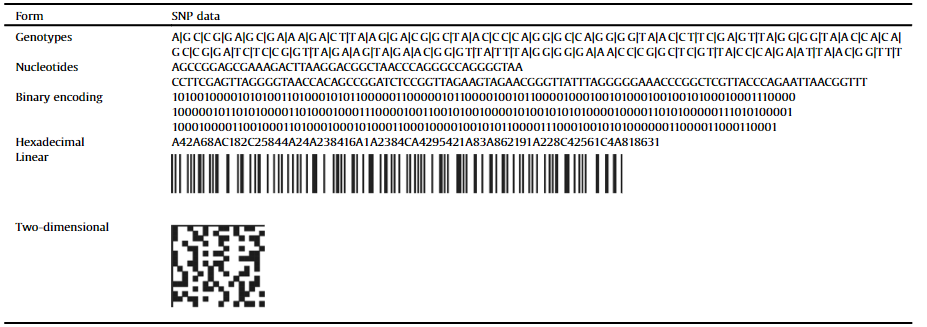
Kuaternion (Quaternion) kodlama, N boyutlu karmaşık sayı kodlamasıdır. Tetrahedron ile eşdeğerdir. *h= a0i0 + … + ajij + … +* aNi*,  N* Z+ denkleminde j = 3 olduğunda 4D kuaterniondur [11]. Protein kodlayan bölgelerin tespiti için dört tane Kuaternion çeşidi geliştirilmiştir. Bazlara sırasıyla *A = i + j + k, C = –i + j – k, G = –i –j + k, T = i – j – k* verilmektedir. Bu tekniğin avantajı, sıralı ve karmaşık sayı kodlamaya göre Kuaternion kodlama protein kodlayan bölgelerin sınıflandırılması için daha iyi performans göstermektedir [23]. Şekil 4.26, örnek diziliminKuaternion kodlama tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 26.** Kuaternion kodlamanın sayısal gösterimi

### SNP-GIN Kodlama Tekniği

SNP-GIN kodlama tekniği, bireylere kendi biyolojik özelliklerine göre benzersiz bir genetik kimlik numarası (GIN) atanması için geliştirilmiştir. SNP (Single Nücleotide Polimorphisms) tek nükleotid polimorfizmleri bireysel DNA belirteçleri olarak kullanılmaktadır. Bu tekniğin avantajı genotip frekanslarının hesaplanması kolaydır. SNP sonuçları bir bilgisayar tarafından kolayca yorumlanabilmektedir. SNP’lerin tekrarlanan elementlerle karşılaşıldığında yeniden genetik düzenlemelere karşı düşük duyarlılık göstermesi, yüksek oranda bozulmuş DNA’ya uygulanabilirliği ve hızlı genotipleme sağlaması gibi avantajları vardır. Ana avantajı, polimorfik nükleotidler üzerindeki verileri sayısallaştırma fırsatıdır. Her spesifik SNP, A, C, G, T harflerinin alfabetik sırayla düzenlendiği ACGT (SNP değeri) 4 bitlik bir kutu olarak olarak alınır. Nükleotidlerin varlığına veya yokluğuna göre 1 veya 0 değerleri atanır. Kısa çizgi (-), ikinci nükleotidin tanımlanamadığı durumu sembolize eder. Her iki eşleştirilmiş kromozomda SNP’nin bulunmaması durumunda sayısallaştırma için “0000” değeri atanır. Homozigot SNP’ler bir tane 1 ve üç tane 0 ile kodlanırken heterezigot SNP’ler iki tane 1 ve iki tane 0 ile kodlanmaktadır. Tüm SNP’lerin analiz edilmesi ile elde edilen ikili formattaki veri kolayca onaltılık formata da kodlanabilir veya grafiksel olarak temsil edilebilir ve herhangi bir kimlik belgesinde kullanılabilir. Genetik tanımlama numarası (GIN) olarak adlandırılmaktadır. Dijital ve grafik formatlarında SNP veri işleme ve GIN’lerin oluşturulması için 72 SNP’den oluşan örnek sayısallaştırma Şekil 4.17’de gösterilmektedir [52].



**Şekil 4. 27.** SNP'den oluşan örnek sayısallaştırma [52]

### Kaos Oyunu Temsili Kodlama Tekniği

Kaos oyunu kodlama (Chaos Game Representation-CGR), DNA dizisindeki her bir nükleotidi düzlemdeki ilgili konuma atayan görüntü şeklinde haritalamayı sağlayan bir tekniktir. CGR kodlama şeması, DNA dizisini birim karenin her bir köşesine nükleotidleri atayarak haritalamaktadır. Bazlara sırayla A: (0, 0), T: (1, 0), G: (1, 1) ve C: (0, 1) değerleri verilmektedir. Bu tekniğin avantajı, orijinal biyolojik dizinin ana özelliklerinin korunarak sayısal diziye dönüşüm gerçekleştirilebilmesidir ve evrimsel ilişkilerin analizinde iyi performans göstermesidir[53]. CGR kodlama şeması, DNA dizisini birim karenin her bir köşesine nükleotidleri atayarak haritalar [11]. Sonra gelen nükleotidler kendisinden önceki nükleotidin konumu ile kendisine ait tepe noktasının ortasına haritalanır. Matematiksel olarak kodlama şeması Denklem 3.13’te verilmektedir [11].

(4. 13)

Burada (gix, giy) bu nükleotide karşılık gelen tepe noktasıdır ve (Xi, Yi) ve (Xi-1, Yi-1) koordinat üzerinde sırasıyla şimdiki ve önceki çizilen noktalardır. Şekil 4.28, örnek diziliminChaos game representation tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 28.** Kaos Oyunu Temsilinin sayısal gösterimi

### Kaos Oyunu Temsili Yürüyüş Kodlama Tekniği

Chaos Game Representation Yürüyüş tekniğinde, CGR haritalama ile termodinamik özellikleri göz önünde bulunduran bir haritalama tekniğidir. Üç tane iki boyutlu CGR düzlemi oluşturulmuştur. Bu üç düzlemdeki tepe noktaları pürin-pirimidin, amino-keto, güçlü-zayıf H bağları termodinamik özelliklerine göre nükleotidlere atanmıştır. Temel kodlama yapısı CGR ile aynıdır [11]. Bu tekniğin avantajı, türlerin ekzonları arasındaki benzerlik ve farklılığın ölçülmesinde daha iyi performans göstermesidir [54], [55]. Şekil 4.29, örnek diziliminKaos oyunu temsili yürüyüş tekniğiyle pürin-pürimidin, amino-keto, güçlü-zayıf H bağları termodinamik özelliklerine göre sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 29.** Kaos Oyunu Temsil Yürüyüşünün sayısal gösterimi (RY, MK, WS)

### Tamsayı Kaos Oyunu Temsili Kodlama Tekniği

Tamsayı kaos oyunu temsili tekniği, kayıpsız bir kodlama yöntemi sağlamaktadır. ICGR ile DNA dizisi nükleotidlerin yinelenen fonksiyonu ve dizideki konumları ile temsil edilir ve kodlanıp kod tekrar geri çözülebilir. CGR kodlamasının tanımında görüldüğü gibi mevcut nükleotidin koordinatlarının önceki nükleotidin koordinatı ve kendisinin köşe koordinatı tarafından belirlendiği göz önünde bulundurulunca bu özyinelemeli ilişkiye göre bir DNA dizisinin son koordinatının tam DNA dizisinin bilgisinin içerdiği savunulmaktadır. Bazlar sırasıyla A = (1,1), T = (-1, 1), C = (-1, -1), G = (1, -1) koordinatları ile temsil edilerek bir karenin köşelerini oluşturur, bunlar CGR köşeleridir. Bu tekniğin avantajı, türler arası benzerlik tespitinde iyi performans göstermesidir. Dezavantajı ise hesaplamalardaki kayan noktalar nedeniyle uzun DNA dizileri tam olarak kurtarılamaz [56]

p i,x = p i-1,x + 2i-1 α i,x

p i,y = p i-1,y + 2i-1 α i,y

α i = S(i), α i ϵ {A, T, C, G} (4. 14)

i = 2, 3, …, n

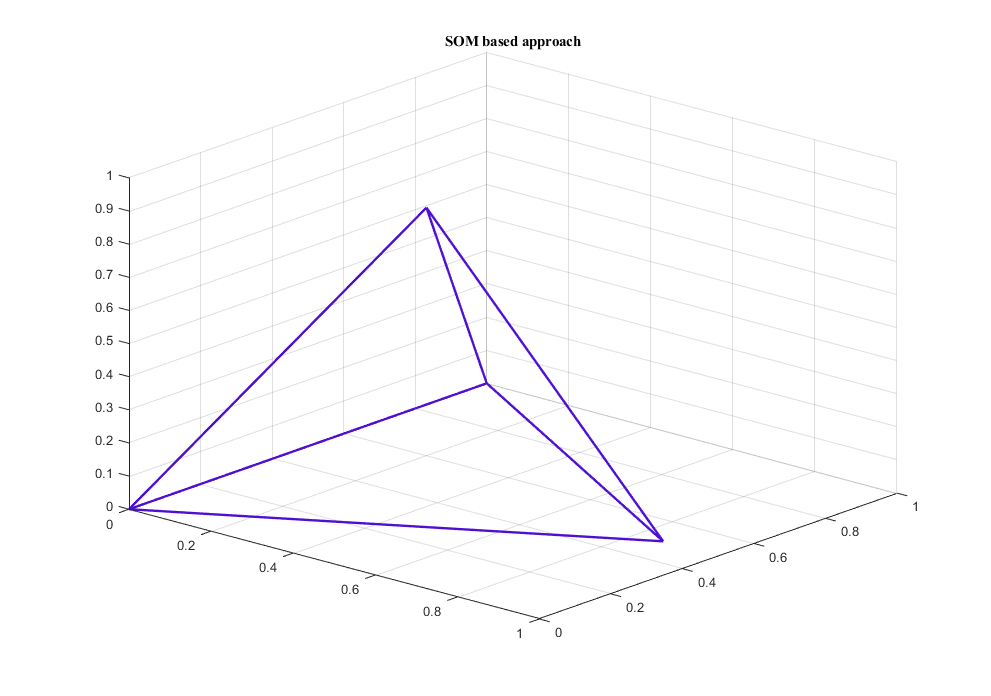
Burada i kaçıncı nükleotid olduğunu, pi mevcut nükleotidin koordinatını, pi-1 önceki nükleotidin koordinatını, αi mevcut nükleotide ait köşenoktasının değerini temsil etmektedir [56]. Şekil 4.30, örnek diziliminKaos oyun representation tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 30.** Tamsayılı Kaos Oyunu Temsilinin sayısal gösterimi

### Kendi Kendini Düzenleyen Haritalar Tabanlı Kodlama Tekniği

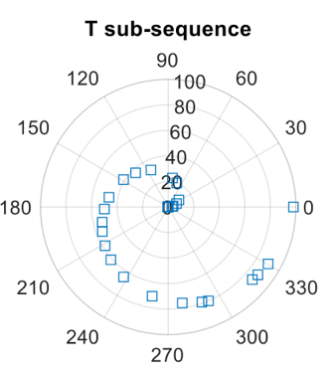
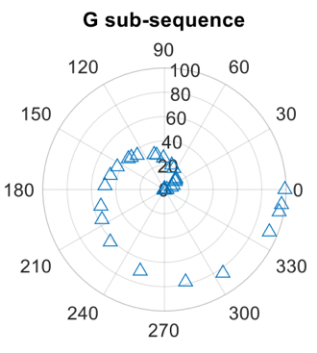
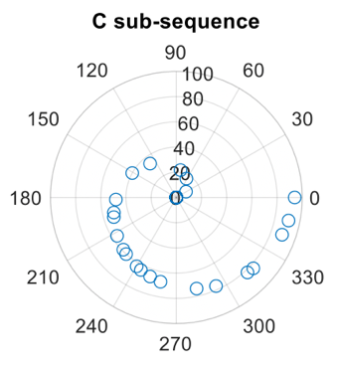
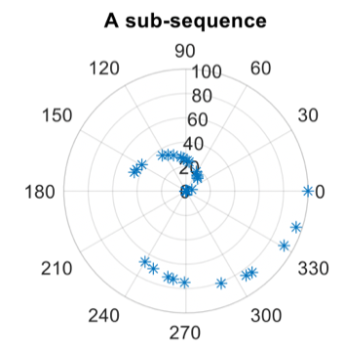
Kendi kendini düzenleyen harita (SOM – Self organizing map) tabanlı kodlamada genomik diziler sabit boyutlu ve metrik tabanlı vektörler ile kodlanmaktadır. C, T, A, G nükleotidleri düzensiz bir tetrahedronun dört köşesine eşlenir. AG ve CT köşeleri arasındaki mesafe 1, diğer çiftler arasındaki mesafe 2 olarak ayarlanır [11]. Bazlara sırayla A: (0, 0, 0), T: (0.289, 0.5, 0.816), C: (0.866, 0.5, 0), G: (0, 1, 0) değerleri verilmektedir [57]. Bu tekniğin avantajı, yapay sinir ağlarında denetimsiz öğrenmede iyi performans göstermektedir. Ayrıca SOM ile uzak türlerin filogenetik analizi çalışmalarında yeni bir sayısallaştırma sağlanmıştır [57]. Şekil 4.31, örnek diziliminSOM kodlama tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 31.** SOM tabanlı kodlamanın sayısal gösterimi

### Fermat Spiral Eğrisi Kodlama Tekniği

Fermat spirali eğrisi gösterimi DNA dizisinin global ve yerel konum bilgisini birleştiren grafiksel bir sayısal temsil sağlamaktadır. DNA dizisinin grafiksel temsilinde dizinin global bilgisinden tam olarak yararlanmak için A, C, G, T nükleotidlerinin karşılık geleceği konum bilgilerinin temsili için dört alt dizi oluşturulmaktadır. Alt dizilerdeki her nükleotid kümedeki bir noktaya karşılık gelmektedir. Fermat spirali eğrisine nükleotidlerin kodlanması ile DNA dizisinin grafiksel temsili gerçekleştirilir. Bu tekniğin avantajı, kutupsal koordinat sisteminde orijinal dizinin konum bilgisini tutabilen monoton olarak artan fonksiyon olmasıdır. Farklı türlerin tespitinde, ilk ekzonların karşılaştırılmasında iyi performans göstermektedir [58]. Şekil 4.32, örnek diziliminfermat spiral eğrisi kodlama tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 32.** Fermat spiral eğri kodlamasının sayısal gösterimi

### Spektral Dinamik Kodlama Tekniği

Spektral dinamik gösterim, DNA dizisindeki A, G, C, T bazlarının dağılımlarının bir dizi ayrık çizgilerle ifade edildiği grafiksel bir temsildir. Dizideki taban konumları orijinal dizinin taban konumlarına karşılık gelir ve çizgilerin uzunlukları 1’e eşittir. Bu temsilin grafiksel görüntüsü bir dizi keskin tayf çizgisinden oluşan atomik, moleküler veya yıldız tayfına benzer. Bu nedenle DNA dizisinin spektral temsili olarak adlandırılır. Böylece dinamik bir temsil elde edilmiş olur [59]. Bu tekniğin avantajı fiziğin spektroskopi ve dinamik alanlarında iyi performans göstermesidir. Şekil 4.33, örnek diziliminSpektral dinamik kodlama tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 33.** Spektral dinamik gösterimin sayısal gösterimi

### 2 Boyutlu Dinamik Kodlama Tekniği

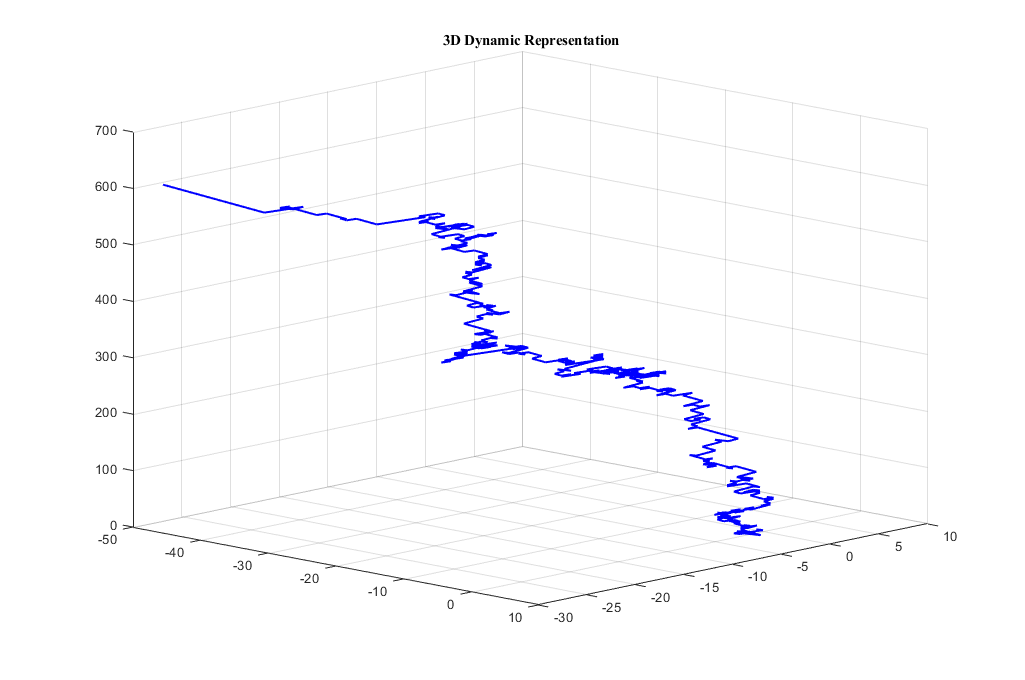
Biyolojik dizilimlerin dinamik temsili (Dynamic Representation of Biological Sequences -DRBS) yöntemlerinden biri olan 2 boyutlu dinamik temsil biyolojik dizilerin öklit uzayındaki nokta kütleleri ile temsil edilmesidir. Bu dizi DNA, RNA veya protein olabilir. DBRS yöntemlerinde her baz bir vektör ile temsil edilir. Grafik temsili kartezyen koordinat sisteminde bazları temsil eden nükleotidlerin yürütülmesi (kayma) ile gerçekleştirilir. DNA/RNA dizilerinin 2 boyutlu dinamik temsilinde bazlar sırayla A = (-1, 0), G = (1, 0), C = (0, 1), T/U = (0, -1) değerlerini alır [60]. Bu tekniğin avantajı biyolojik dizinin karakteristiğini çıkarmada etkili olmasıdır. Şekil 4.34, örnek dizilimin2 boyutlu dinamik kodlama tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 34.** 2 Boyutlu Dinamik Temsilinin sayısal gösterimi

### 3 Boyutlu Dinamik Kodlama Tekniği

DBRS yöntemlerinden bir diğeri 3 boyutlu dinamik temsildir. DBRS yöntemleri ile ilgili 2 boyutlu dinamik temsilde anlatılanlar burada da geçerlidir. 2 boyutlu temsilin dejenerasyona yol açtığı çakışmalar 3 boyutlu temsilde giderilmiştir. 3 boyutlu temsilde DNA/RNA bazları A = (-1, 0, 1), G = (1, 0, 1), C = (0, 1, 1), T/U = (0, -1, 1) şeklinde temsil edilmektedir [60], [61].  Bu tekniğin avantajı biyolojik dizinin karakteristiğini çıkarmada yine oldukça etkili olmasıdır. Şekil 4.35, örnek dizilimin3-boyutlu dinamik temsili tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 35.** 3 Boyutlu Dinamik Temsil'in sayısal gösterimi

### 8 Boyutlu Vektör Kodlama Tekniği

8 boyutlu vektör temsili, DNA dizilerinin 2 boyutlu grafik temsiline dayanarak ortalama ve varyans olasılık kavramları ile geliştirilmiştir. DNA dizisinin 8 boyutlu Öklid uzayında haritalanması sağlanır. Bu haritalama ile farklı türlerin ekzonları karşılaştırılarak benzerlik/farklılıkları incelenmiştir [62]. Bazlara sırayla (1, 0.2) A, (1, -0.2) T, (1, 0.3) C,  (1, -0.3) G vektörleri atanır. DNA dizilerinin uzayında 8D vektörü elde edilerek 8 boyutlu Öklid uzayında yeni bir harita oluşturulmuş olur. Dizilerin vektörleri belirlendikten sonra aralarındaki mesafe d=‖E1-E2‖ şeklinde hesaplanarak türler arasındaki benzerlik/farklılık ölçüsü elde edilmiş olur. Değer ne kadar küçükse türler o kadar benzerdir. Bu tekniğin avantajı türler arası benzerlik tespitinde ve protein dizilerini analiz etmede iyi performans göstermesidir [62]. Şekil 4.36, örnek dizilimin8-boyutlu vektör temsili tekniğiyle sayısallaştırılmış sinyal çizimini vermektedir.



**Şekil 4. 36.** 8 Boyutlu Vektör Temsilinin Sayısal Gösterimi

Tablo 4.13, “Grafiksel Temsil (GT)” grubundaki haritalama tekniklerinin tamamının kısa bir özet bilgisini vermektedir.

**Tablo 4. 14.** Grafik gösterim grubundaki tüm sayısal kodlama tekniklerinin özeti

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tekniğin Adı | Kodlama Şeması | Sayısal Temsil | Tanım | |
| Tetrahedron kodlama | A=  G=,  C= ,  T= | *X*=[AGCTACCGTG]  = ,  = ,  = | | Nükleotidler tetrahedronun dört köşesine yerleştirilerek köşelerin sayısal denklemleri ile kodlanır. |
| H-Eğrisi kodlama | A =  T =,  C =  G = | *X*=[AGCTACCGTG]  = [-0.3660i, 0.3660i, 1.3660i, 1.3660i, -0.3660i, 1.3660i, 1.3660i, 0.3660i, 1.3660i, 0.3660i] | | Nükleotidler x, y eksenlerindeki i, j vektörleri ile oluşturulan fonksiyonlar ile kodlanır. |
| Z-Eğrisi kodlama |  | *X*=[AGCTACCGTG]  = [-1, 0, -1, 0, -1, -2, -3, -2, -1, 0]  = [1, 0, 1, 0, 1, 2, 3, 2, 1, 0]  = [1, 0, -1, 0, 1, 0, -1, -2, -1, -2] | | Nükleotidler tetrahedronun dört yüzeyine doğru olan vektör setleri ile kodlanır. |
| Quaternion kodlama | A= i+j+k, C=-i+j-k, G=-i-j+k, T=i-j-k | *X*=[AGCTACCGTG]  = [*i+j+k, -i-j+k, -i+j-k, i-j-k, i+j+k, -i+j-k, -i+j-k, -i-j+k, i-j-k, -i-j+k*] | | Nükleotidler 4D quaternion denklemleri ile temsil edilir. |
| SNP GIN kodlama | 1000 → A|A veya A|-  0100 → C|C veya C|-  0010 → G|G veya G|-  0001 → T|T veya T|-  1100 → A|C, 1010 → A|G  1001 → A|T, 0110 → C|G  0101 → C|T, 0011 → G|T | *X*=[AGCTACCGTG]  =>Genotypes A|G C|T A|C C|G T|G  Nucleotides AGCTACCGTG  Binary Encoding 10100101110001100000  Hexadecimal A5C60 | | Nükleotidlere dört bir ikili SNP deperleri atanarak GIN’ler oluşturulur. |
| Kaos oyunu temsili kodlama (CGR) | A: (0, 0), T: (1, 0), G: (1, 1), C: (0, 1) | *X*=[AGCTACCGTG]  = [(0,0), (0,0), (0.5, 0.5), (0.25, 0.75), (0.625, 0.375), (0.3125, 0.1875), (0.1563, 0.5938), (0.0781, 0.7969), (0.5391, 0.8984), (0.7695, 0.4492)] | | DNA dizisi birim kare içerisinde koordinat değerlerine göre haritalanır. |
| Kaos oyunu temsili yürüyüş kodlama (CGR-Yürüyüş) | CGRRY: A(0, 0),T(1, 0),C(0, 1),G(1, 1)  CGRMK: A(0, 0),T(1, 0),G(0, 1),C(1, 1)  CGRWS: A(0, 0),G(1, 0),C(0, 1),T(1, 1) | *X*=[AGCTACCGTG]  Purine-Prymidine  = [[(0,0), (0,0), (0.5, 0.5), (0.25, 0.75), (0.625, 0.375), (0.3125, 0.1875), (0.1563, 0.5938), (0.0781, 0.7969), (0.5391, 0.8984), (0.7695, 0.4492)]]  Amino-Keto  = [(0,0), (0,0), (0, 0.5), (0.5, 0.75), (0.75, 0.375), (0.375, 0.1875), 0.6875, 0.5938), 0.8438, 0.7969), (0.4219, 0.8984), 0.7109, 0.4492)]  Weak-Strong  = [(0,0), (0,0), (0.5, 0), (0.25, 0.5), (0.625, 0.75), (0.3125, 0.375), (0.1563, 0.6875),(0.0781, 0.8438), (0.5391, 0.4219), (0.7695, 0.7109)] | | Kaos oyunu temsili DNA’nın termodinamik özellikleri göz önünde bulundurularak DNA-yütüyüş şeklinde gerçekleştirilir. |
| Tamsayı kaos oyunu temsili kodlama (iCGR) | A=(1,1), T=(-1,1), C=(-1,-1), G=(1,-1) | *X*=[AGCTACCGTG]  = [(1,1), (3, -1), (-1,-5), (-9,3), (7,19), (-25,-13), (-89, -77), (39,-205), (-217, 51), (295, -461)] | | Kaos oyunu temsili kayan noktalı sayılar yerine tam sayılar ile gerçekleştirilir. |
| Kensi kendini düzenleyen haritalar tabanlı kodlama (SOM) | A: (0, 0, 0), T: (0.289, 0.5, 0.816),  C: (0.866, 0.5, 0), G: (0, 1, 0) | *X*=[AGCTACCGTG]  = [(0,0,0), (0,1,0), (0.866, 0.5, 0), (0.289, 0.5, 0.816), (0,0,0), (0.866, 0.5, 0), (0.866, 0.5, 0), (0,1,0), (0.289, 0.5, 0.816), (0,1,0)] | | Nükleotidler dört köşesi ile eşlenir. AG ve CT köşeleri arasındaki mesafe değerleri ile kodlama gerçekleştirilir. |
| Fermat spiral eğrisi kodlama | A, T, C, G konumlarına göre oluşturulan dört alt dizinin fermat spiralinde gösterilmesi | *X*=[AGCTACCGTG]  (Aseq)= [1,0,0,0,5,0,0,0,0,0]  (Gseq)= [0,2,0,0,0,6,0,8,0,0]  (Cseq)= [0,0,0,0,0,0,0,0,9,0]  (Tseq)= [0,0,3,4,0,0,7,0,0,10] | | DNA’daki nükleotidlerin global ve yerel konum bilgileri fermat spirali üzerinde haritalanır. |
| Spectral dinamik kodlama | Her bazın sağılımlarının bir dizi çizgi ile temsil edilmesi | *X*=[AGCTACCGTG]  (A)= [1,0,0,0,1,0,0,0,0,0]  (G)= [0,1,0,0,0,1,0,1,0,0]  (C)= [0,0,0,0,0,0,0,0,1,0]  (T)= [0,0,1,1,0,0,1,0,0,1] | | DNA nükletidlerinin dağılımları dört ayrı ayrık çizgi grafikleri ile temsil edilir. |
| 2D dinamik kodlama | A=(-1,0), G=(1, 0), C=(0, 1), T=(0,-1) | *X*=[AGCTACCGTG]  = [(-1,0), (0,0), (0,1), (0,0), (-1,0), (-1,1), (-1,2), (0,2), (0,1) (1,1)] | | DNA dizileri 2D öklit uzayında nokta kütleleri ile temsil edilir. |
| 3D dinamik kodlama | A=(-1, 0, 1), G=(1, 0, 1),  C=(0, 1, 1), T=(0, -1, 1) | *X*=[AGCTACCGTG]  = [(-1,0,1), (0,0,2), (0,1,3), (0,0,4), (-1,0,5), (-1,1,6), (-1,2,7), (0,2,8), (0,1,9), (1,1,10)] | | DNA/RNA dizileri 3D düzlemde temsil edilir. |
| 8D vektör kodlama | A=(1, 0.2), T=(1, -0.2),  C=(1, 0.3), G=(1, -0.3)  *zi = yi / i*  K =(*mz, vz*) | *X*=[AGCTACCGTG]  = [(1, 0.2), (2, -0.2), (3, 0.2), (4,0), (5, 0.2), (6, 0.5), (7, 0.8), (8, 0.5), (9, 0.3), (10, 0)]  Slope=1.5370e-04  Variance=0.0182+0.0200i | | DNA/RNA dizilerinin zikzak grafiğinden ortalama, varyans değerleri ile 8D vektör oluşur |