**"GENETİK KOD ÜZERİNDE MOTİF ANALİZİ VE KULLANIM ALANLARI "**

Rumeysa DİKMEN

Kütahya Sağlık Bilimleri Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği, Kütahya, Türkiye

[rumeysa.dikmen@ogr.ksbu.edu.tr](mailto:rumeysa.dikmen@ogr.ksbu.edu.tr)

+905514533113

# ÖZ

Bu çalışma, *Python* programlama dili ile *Biopython* kütüphanesi kullanılarak gerçekleştirilen bir DNA sekansı analizi üzerine odaklanmaktadır. Araştırmanın temel amacı, belirli bir DNA sekansını k-merlerine göre ayırmak ve elde edilen k-merler üzerinden motif analizi gerçekleştirmektir. Bu süreç, biyoinformatik araştırmalarına dayanmaktadır.

Metodoloji olarak, *Biopython* kütüphanesi içinde bulunan *Seq* ve diğer ilgili modüller kullanılarak DNA sekansı manipülasyonları gerçekleştirilmiştir.

**Keywords:** Biyoinformatik, DNA, Python, Biopython, DNA Pattern, K-mer, N-gram

# Giriş

Genetik bilimlerdeki ilerlemelerle birlikte biyolojik veri analizi önemli bir alan haline gelmiştir. Bu alandaki araştırmalar, genetik materyali anlamak, analiz etmek ve yorumlamak amacıyla çeşitli bilişim araçlarını kullanmayı içermektedir. Bu çalışma, bir Python programlama dili ve Biopython kütüphanesi ile gerçekleştirilen bir projeyi ele almaktadır. Temel amaç, belirli bir DNA sekansını k-merlere göre bölmek ve motif analizi yapmaktır. Bu adım, genetik bilgiyi anlamak ve gelecekteki analizlere temel oluşturmak açısından önemlidir.

Bu projede, biyoinformatik ve makine öğrenimi disiplinleri arasındaki ortak noktalara da değinilmesine karşın makine öğrenimi yöntemleri doğrudan kullanılmamıştır. Bunun yerine, genetik veri analizi için geleneksel biyoinformatik araçları ve yöntemleri tercih edilmiştir. Bu yaklaşım, biyoinformatik ve makine öğrenimi arasındaki potansiyel ilişkilerin anlaşılmasına bir adım daha yaklaşmak adına bir temel oluşturmaktadır.

Çalışmanın devamında, kullanılan biyoinformatik yöntemleri ve elde edilen sonuçların daha detaylı bir şekilde ele alınacak, ayrıca genetik veri analizi ile makine öğrenimi arasında kurulan bağlantılara odaklanılması hedeflenmektedir. Bu bağlamda, makine öğrenimi yöntemlerinin potansiyel uygulama alanları ve bu alanlarda karşılaşılan zorluklar incelenecek; biyoinformatik ve genetik veri analizi ile nasıl entegre edilebileceği konusunda bir görüş sunulacaktır.

# Materyal ve Metotlar

1. Veri Toplama

Bu çalışma için kullanılan DNA sekansı/sekansları, [*Kaggle*](https://www.kaggle.com/) adlı veri bilimi platformundan temin edilmiştir. Veri setinin içeriği, DNA sekanslarının özellikleri ve toplam veri seti büyüklüğü bu kaynaktan elde edilmiştir[[1]](#footnote-1).

1. Programlama Dili ve Kütüphane

Çalışmanın temel programlama dili [*Python*](https://www.python.org/) olup, genetik verilerin analizi için [*Biopython*](https://biopython.org/) kütüphanesi kullanılmıştır. Python, genetik veri analizi için yaygın olarak tercih edilen dillerden biri olup, Biopython kütüphanesi de genetik bilgi işleme işlemlerini kolaylaştırmak için geliştirilmiştir. Bu çalışmada bu kütüphanenin seçilme nedeni, sayesinde çevrimiçi veri tabanlarına ulaşıma kolaylık sağlaması, DNA ve RNA sekanslarını tersine çevirme ve bu sekansları saklayan FASTA vb. gibi büyük boyutlu dosyaları kolaylıkla okuması dahil birçok konuda kullanıcının işini kolaylaştırmasıdır[[2]](#footnote-2).

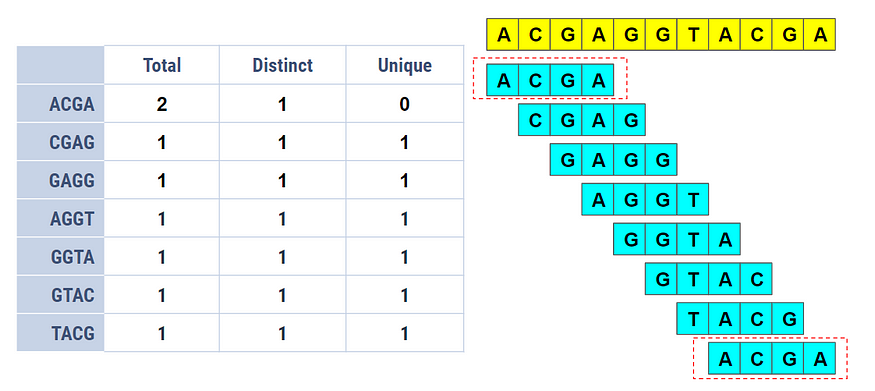
1. Analiz Yöntemleri

DNA sekansı, motif analizi için bir tür N-gram analizi olan, ancak özellikle genetik dizilerin analizi için tasarlanan K-mer Encoding algoritması[[3]](#footnote-3) aracılığıyla Biopython kütüphanesinin sunduğu imkanların da kullanılmasıyla analiz edilmiştir.

K-mer Encoding algoritması, motif analizi için kullanıldığında farklı k-merlerin frekanslarını belirleyerek bir histogram çıkarılabilir. Daha sonra elde edilen bu veriler, makine öğrenimi ve model tasarlamada bir yapı taşı olarak kullanılabilir.

*Denklem 1. K-mer Encoding algoritması formülü[[4]](#footnote-4)*

*N* = DNA uzunluğu (baz sayısı) ve *k* = mer uzunluğu olmak üzere, bir DNA sekansındaki k-mer sayısı *Denklem 1.* ‘de gösterilen formülle bulunabilir. K-mer Encoding algoritmasının işleyiş şekli ise aşağıdaki şekilde gösterildiği gibidir.

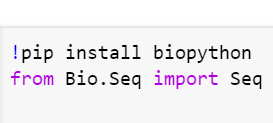


Şekil . K-mer Encoding algoritmasının işleyiş görseli[[5]](#footnote-5)

*Şekil 1*.’de görüldüğü gibi k boyutunda bir pencere oluşturulup, bu pencerenin her defasında birer karakter kaydırılarak soldan sağa verilen sekansın okunmasıyla K-mer Encoding algoritması gerçekleştirilmiş olur.[[6]](#footnote-6) Bu yöntem, genetik verilerin etkili bir şekilde işlenmesini ve çeşitli analizlerin gerçekleştirilmesini sağlamaktadır.

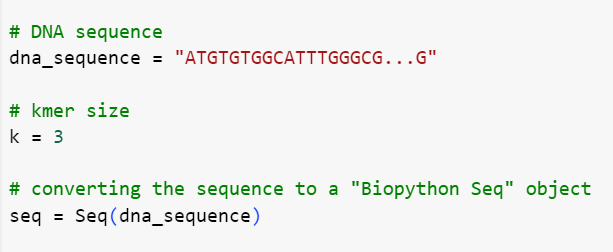
1. Araştırma ve Kod Geliştirme

Çalışma yürütülürken internet üzerinde yapılan araştırmalar sonucu edinilen bilgiler doğrultusunda DNA sekanslarını parçalarına bölmek için Python programlama dilinde kod geliştirilmiştir. Kodun geliştirilme süreci, araştırmalardan elde edilen yönergeleri takip ederek ve Biopython kütüphanesinden sağlanan Seq sınıfı kullanarak gerçekleştirilmiştir.



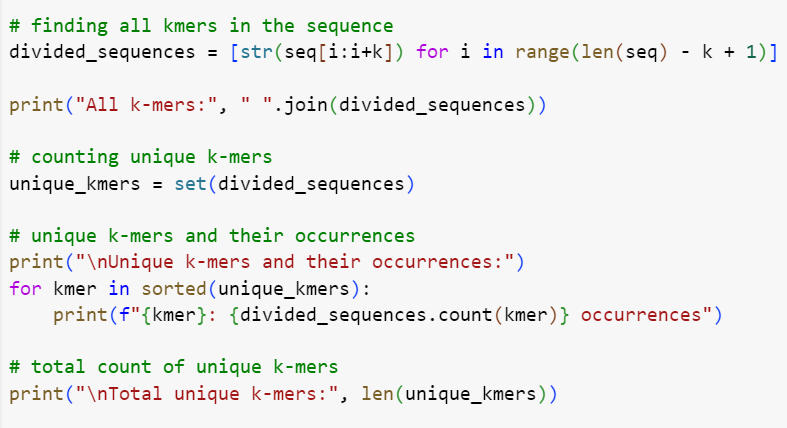
Şekil . Biopython kütüphanesinin indirilmesi ve Seq sınıfının içeriye aktarılması

Seq sınıfı, Biopython kütüphanesinin genetik çalışmalarda sağladığı faydalardan biri olup, ilgili kütüphanenin DNA veya RNA fark etmeksizin sekansları temsil etmek için kullanılan bir yapı sağlar[[7]](#footnote-7).



Şekil . DNA sekansının ile mer boyutunun tanımlanması ve DNA sekansının Seq objesine dönüştürülmesi

Şekil 3.’de ise DNA sekansı ve k-mer boyutunun tanımlanmasından sonra, DNA sekansı Seq objesine çevrilir ve böylece daha sonra yapılabilecek işlemler için uygun formata getirilir.



Şekil . Sekansın merlerine ayrılması ve ekrana yazdırılması

Şekil 4.’de ise çalışmanın son aşaması olan DNA sekansını merlerine ayırma ve her bir farklı merin bulunup, frekanslarının da ekrana yazdırılması için yazılan kod satırları gösterilmektedir.

ekran görüntüsü, kalıp, desen, düzen, monokrom, tek renkli içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Verilen DNA sekansının bütün k-merleri

ekran görüntüsü, metin, siyah içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . K-merlerin frekansları 1/2

ekran görüntüsü, metin, siyah içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . K-merlerin frekansı 2/2

Şekil 5., Şekil 6. ve Şekil 7.’de çalışmanın çıktıları gösterilmiş ve böylece her bir k-mer ve frekansları açık şekilde görülmektedir.

1. Değerlendirme Kriterleri

Analiz sonuçları, k-merlerin ve motiflerin belirlenmesi üzerine odaklanmıştır. Başarı kriterleri, bu genetik özelliklerin doğru bir şekilde tespit edilip edilemediği üzerinden değerlendirilmiştir.

**Bulgular**

K-mer Analizi Sonuçları

* Çalışma kapsamında gerçekleştirilen *k*=3 olmak üzere yapılan k-mer analizi, belirli bir DNA sekansını başarıyla trimerlere bölmüş ve bu trimerlerin frekanslarını belirlemiştir. Tablo 1’de, çalışma sonucu tespit edilen trimerler ve frekanslarının histogram grafiği verilmiş ve 64 trimer arasından en sık rastlananın “TTT” trimeri olduğu görülmüştür.

ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Tablo . K-merlerin frekansları

**Tartışma**

Bu çalışma, Biopython kütüphanesi kullanılarak gerçekleştirilen DNA sekansı analizi üzerine odaklanmıştır. Yapılan k-mer analizi, genetik bilgiyi anlamak ve belirli motifleri belirlemek açısından önemli bilgiler sunmuştur. Ancak, bu yöntemlerin sadece sekans tahmininde değil, aynı zamanda motif analizi algoritmalarıyla çeşitli kanser türlerinin tespiti ve aralarındaki ilişkilerin anlaşılması için de kullanılabileceği belirtilmiştir[[8]](#footnote-8).

Elde edilen k-mer analizi sonuçlarına dayanarak, belirli trimerlerin belirgin bir şekilde daha sık görüldüğü gözlemlenmiştir. Özellikle, "ATG" start kodonu gibi önemli genetik motiflerin belirlenmesi, genetik bilgi içindeki önemli bölgelerin tespit edilmesini sağlayabilir. Ancak, bu trimerlerin kanser türleri ile ilişkilendirilip ilişkilendirilemeyeceği üzerine daha fazla araştırma yapılması önemlidir.

Biopython kütüphanesinin kullanılması, genetik veri analizi için güçlü bir araç seti sunmuştur. Seq sınıfı ve ilgili modüller, DNA sekansı üzerinde çeşitli manipülasyonların kolayca gerçekleştirilmesine imkân tanımıştır. Ancak, bu yöntemlerin kanser türlerinin detaylı analizi için yeterli olup olmadığı, daha spesifik bir şekilde ele alınmalıdır.

Çalışmanın sınırlamaları göz önüne alındığında, Kaggle platformundan elde edilen veri setinin belirli bir biyolojik bağlama sahip olmaması ve sadece genetik veri analizine odaklanması, çalışmanın genel geçerliliğini sınırlayabilir. Bu nedenle, farklı kanser türlerinden elde edilen veri setleri üzerinde yapılan analizlerle çalışma sonuçlarının desteklenmesi önemlidir.

Gelecekteki çalışmalarda, makine öğrenimi yöntemlerinin genetik veri analizi üzerindeki etkilerini daha ayrıntılı bir şekilde incelemek, özellikle kanser türlerinin sınıflandırılması ve ilişkilerinin anlaşılması üzerine odaklanmak ilginç bir araştırma alanı olabilir. Ayrıca, farklı DNA sekanslarının karşılaştırılması ve evrimsel analizler, genetik çeşitlilik konusunda daha derinlemesine bir anlayış sağlayabilir[[9]](#footnote-9).

**Sonuç ve Öneriler**

Bu çalışma, Biopython kütüphanesi kullanılarak gerçekleştirilen DNA sekansı analizi ile genetik bilimi anlamak ve ileriye dönük araştırmalar için temel bir adım oluşturur. Ancak, daha fazla araştırma ve analiz yapılması, özellikle kanser türlerinin detaylı incelenmesi, bu alandaki çalışmaların kapsamını genişletebilir.

Öneriler arasında, çalışmanın kapsamının genişletilmesi adına farklı kanser türlerinden elde edilen veri setleri ile analizlerin genişletilmesi, makine öğrenimi yöntemlerinin kullanımının detaylı bir şekilde incelenmesi ve Biopython kütüphanesi dışında diğer genetik veri analizi araçlarının da değerlendirilmesi yer alabilir.

Ayrıca, bu çalışmanın temel aldığı Kaggle platformundan elde edilen veri setinin belirli bir biyolojik ilişki içermemesi, gelecekteki çalışmalarda daha özenli veri setleri kullanılmasını teşvik edebilir. Bu, elde edilen sonuçların genel geçerliliğini artırabilir ve çeşitli kanser türlerinin analizinde daha güvenilir bulgular elde edilmesine olanak sağlayabilir.

# KAYNAKÇA

(2022, 09 30). frontiersin.org: https://www.frontiersin.org/articles/10.3389/fphys.2022.952709/full adresinden alındı

Biopython. (2024, 01 05). *Biopython Org*. Biopython.org Web sitesi: https://biopython.org/DIST/docs/tutorial/Tutorial.html#sec12 adresinden alındı

Biopython. (2024, 01 12). *Handling sequences with the Seq class.* Biopython.org: https://biopython.org/wiki/Seq adresinden alındı

Brihadiswaran, G. (2020, 07 02). *Bioinformatics 1: K-mer Counting*. medium.com: https://medium.com/swlh/bioinformatics-1-k-mer-counting-8c1283a07e29 adresinden alındı

Gardner, S. N. (2009). *Feasibility of N-Gram Data-Structures for Next-Generation Pathogen Design.* Livermore, CA: Lawrence Livermore National Lab.

*Handling Sequences with the Seq Class*. (2024, 01 11). biopython.org: https://biopython.org/wiki/Seq, adresinden alındı

Kaggle. (2020). *human-dna-data*. kaggle.com: https://www.kaggle.com/datasets/neelvasani/humandnadata/data adresinden alındı

Mahes, A. (2021, 12 21). *ML Applications for Data Mining Gene Sequences*. adityamahes.medium.com: https://adityamahes.medium.com/ml-applications-for-data-mining-gene-sequences-76fa2f70c82d adresinden alındı

Sievers, A., Bosiek, K., Bisch, M., Dreessen, C., Riedel, J., Froß, P., . . . Hildenbrand, G. (2017, 04 19). *National Library of Medicine*. ncbi.nlm.nih.gov: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5406869/ adresinden alındı

1. <https://www.kaggle.com/datasets/neelvasani/humandnadata/data> , 2020 [↑](#footnote-ref-1)
2. <https://biopython.org/DIST/docs/tutorial/Tutorial.html#sec12> [↑](#footnote-ref-2)
3. Gardner, S. N. (2009). Feasibility of N-Gram Data-Structures for Next-Generation Pathogen Signature Design (No. LLNL-TR-410145). Lawrence Livermore National Lab. (LLNL), Livermore, CA (United States). [↑](#footnote-ref-3)
4. ,5 <https://medium.com/swlh/bioinformatics-1-k-mer-counting-8c1283a07e29> , 02.07.2020 [↑](#footnote-ref-4)
5. [↑](#footnote-ref-5)
6. <https://adityamahes.medium.com/ml-applications-for-data-mining-gene-sequences-76fa2f70c82d> [↑](#footnote-ref-6)
7. <https://biopython.org/wiki/Seq>, Handling Sequences with the Seq Class, 11.01.2024 [↑](#footnote-ref-7)
8. Gupta, S., Gupta, M. K., Shabaz, M., & Sharma, A. (2022). Deep learning techniques for cancer classification using microarray gene expression data. Frontiers in Physiology, 13, 952709. [↑](#footnote-ref-8)
9. Sievers, A., Bosiek, K., Bisch, M., Dreessen, C., Riedel, J., Froß, P., ... & Hildenbrand, G. (2017). K-mer content, correlation, and position analysis of genome DNA sequences for the identification of function and evolutionary features. Genes, 8(4), 122. [↑](#footnote-ref-9)