POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

DeepTFBS: Transkripsiyon faktörü bağlanma bölgeleri tahmini için derin öğrenme yöntemleri kullanan hibrit bir model

DeepTFBS: A hybrid model using deep learning methods for transcription factor binding sites prediction

***Yazar(lar) (Author(s))**: Ayşegül HATİPOĞLU[1], Volkan ALTUNTAŞ[2]*

ORCID[1]: 0000-0003-1584-0945

ORCID[2]: 0000-0003-3144-8724

To cite to this article: Hatipoğlu A., Altuntaş V., "DeepTFBS: Transkripsiyon faktörü bağlanma bölgeleri tahmini için derin öğrenme yöntemleri kullanan hibrit bir model", ***Journal of Polytechnic***, *(*): *, (*).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Hatipoğlu A., Altuntaş V., "DeepTFBS: Transkripsiyon faktörü bağlanma bölgeleri tahmini için derin öğrenme yöntemleri kullanan hibrit bir model", ***Politeknik Dergisi***, *(*): *, (*).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1509329

DeepTFBS: Transkripsiyon Faktörü Bağlanma Bölgeleri Tahmini için Derin Öğrenme Yöntemleri Kullanan Hibrit bir Model

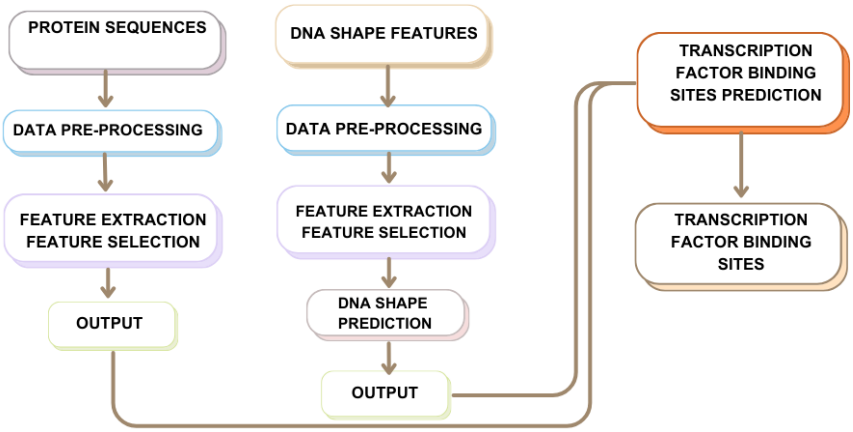
DeepTFBS: A Hybrid Model Using Deep Learning Methods for Transcription Factor Binding Sites Prediction

Önemli noktalar (Highlights)

- *Transkripsiyon Faktörlerinin DNA ile olan etkileşimi / Interaction of Transcription Factors with DNA*
- *Transkripsiyon Faktörlerinin DNA şekli ile ilişkisi / Relationship of Transcription Factors with DNA shape*
- *Derin öğrenme mimarilerin kombinasyonu / Combination of deep learning architectures*

Grafik Özet (Graphical Abstract)

Protein dizileri ve DNA şekil bilgilerden Transkripsiyon Faktörü Bağlanma Bölgelerinin Tahmini / Prediction of Transcription Factor Binding Sites from protein sequences and DNA shape information



Şekil. Uygulama mimarisi / **Figure.** Application architecture

Amaç (Aim)

Protein dizileri ve DNA şekil özelliklerinden Transkripsiyon Faktörü Bağlanma Bölgelerinin Tahmininde hibrit bir model geliştirme. / Development of a hybrid model for prediction of transcription factor binding sites from protein sequences and DNA shape features.

Tasarım ve Yöntem (Design & Methodology)

Transkripsiyon faktörü bağlanma bölgeleri tahmini için hibrit bir derin öğrenme mimarisi önerilmiştir. / A hybrid deep learning architecture is proposed for transcription factor binding site prediction.

Özgünlük (Originality)

Transkripsiyon faktörü bağlanma bölgeleri tahmini için literatürdeki başarılı yöntemlerden oluşan hibrit bir çalışma gerçekleştirilmiştir. Önemli derin öğrenme mimarileri birleştirilerek özgün bir yaklaşım sergilenmiştir. / A hybrid study of successful methods in the literature for transcription factor binding site prediction has been performed. A novel approach is demonstrated by combining important deep learning architectures.

Bulgular (Findings)

Çalışmada önerilen yöntem, literatürdeki diğer yöntemler ile kıyaslandığında yüksek başarı elde etmiştir. / The method proposed in this study has achieved high success compared to other methods in the literature.

Sonuç (Conclusion)

Transkripsiyon faktörü bağlanma bölgeleri tahmini için derin öğrenme mimarilerinin kombinasyonundan oluşan başarılı bir yöntem geliştirilmiştir. / A successful method consisting of a combination of deep learning architectures has been developed for the prediction of transcription factor binding sites.

Etik Standartların Beyanı (*Declaration of Ethical Standards*)

Bu makalenin yazar(lar)ı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler. / The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

DeepTFBS: Transkripsiyon Faktörü Bağlanma Bölgeleri Tahmini için Derin Öğrenme Yöntemleri Kullanan Hibrit bir Model

Araştırma Makalesi / Research Article

Ayşegül HATİPOĞLU[1*], Volkan ALTUNTAŞ[2]

[1]Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Bilecik Şeyh Edebali Üniversitesi, Türkiye
[2]Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, Bursa Teknik Üniversitesi, Türkiye

(Geliş/Received: 02.07.2024; Kabul/Accepted: 26.11.2024; Erken Görünüm/Early View: 28.11.2024)

ÖZ

Moleküler seviyede genetik verinin oluşum, aktarım ve düzenlenme süreçleri anlaşılması zor karmaşık kombinasyonel süreçlerden oluşmaktadır. Bu süreçlerin temelini oluşturan transkripsiyon faktörleri genetik bilginin DNA'dan RNA'ya kopyalanmasını sağlayarak hücrelerin özellik ve fonksiyonlarını belirlemede kritik rol oynar. Özellikle sinir sistemi gibi karmaşık yapıları kontrol eden transkripsiyon faktörleri, gen ifadesini düzenleyerek hastalık, sağlık gibi durumların belirlenmesinde hayati rol oynarlar. Proteinlerin DNA üzerinde bağlandıkları bölgeler, gen ifadelerinin kritik noktalarını belirler ve hücrelerin çeşitli koşullara uyum sağlamasına katkıda bulunur. Genetik hastalıkların teşhis edilmesi ve tedavi edilmesi süreçleri için önemli bir adım olan transkripsiyon faktörü bağlanma bölgelerinin tahmini amacıyla literatürde çeşitli yöntemler geliştirilmiştir. DNA'nın dizi ve şekil özelliklerinin beraber kullanımıyla başarılı sonuçlar elde edilen çeşitli çalışmalar geliştirilmiştir. Bu çalışmada DNA dizileri ve şekillerine dayalı olarak transkripsiyon faktörü etkileşimlerini belirlemek için farklı derin öğrenme teknolojileri birleştirilerek hibrit bir yöntem önerilmiştir. Çalışmada 165 doğrulanmış CHIP-Seq veri kümesi kullanılmıştır.

Anahtar Kelimeler: Derin öğrenme, transkripsiyon faktörü, transkripsiyon faktörü bağlanma bölgeleri.

DeepTFBS: A Hybrid Model Using Deep Learning Methods for Transcription Factor Binding Sites Prediction

ABSTRACT

The formation, transmission and regulation of genetic data at the molecular level are complex combinatorial processes that are difficult to understand. Transcription factors, which form the basis of these processes, play a critical role in determining the properties and functions of cells by copying genetic information from DNA to RNA. Transcription factors, which control complex structures such as the nervous system, play a vital role in determining conditions such as disease and health by regulating gene expression. The binding sites of proteins on DNA determine the critical points of gene expression and contribute to the adaptation of cells to various conditions. Various methods have been developed in the literature for the prediction of transcription factor binding sites, which is an important step for the diagnosis and treatment of genetic diseases. Several studies have been developed with successful results obtained by using DNA sequence and shape features together. In this study, a hybrid method is proposed by combining different deep learning technologies to identify transcription factor interactions based on DNA sequences and shapes. 165 validated CHIP-Seq datasets were used in the study.

Keywords: Deep learning, transcription factor, transcription factor binding sites prediction.

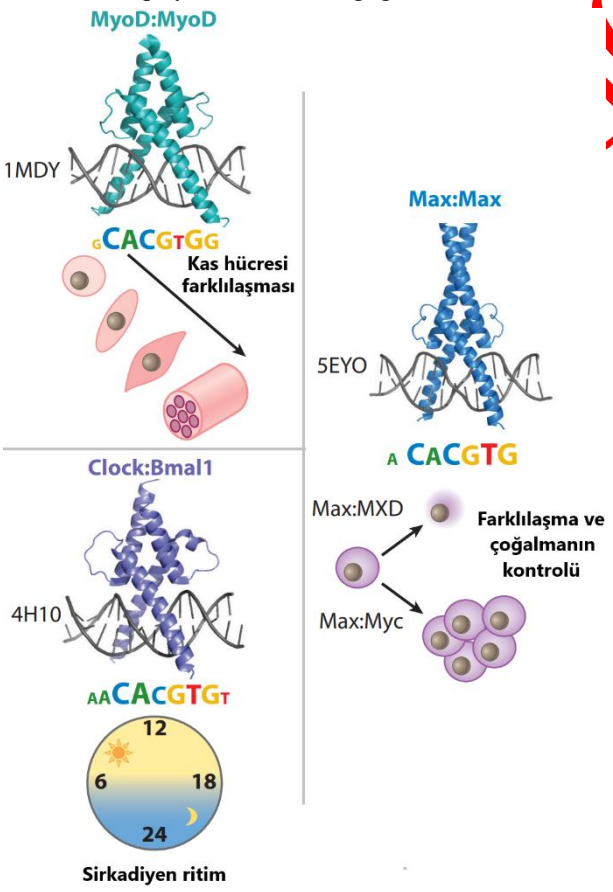
1. GİRİŞ (INTRODUCTION)

Gen, genetik özelliklerin kuşaklar boyunca aktarılmasını sağlayan ve DNA (Deoksiribonükleik asit) molekülü üzerinde bulunan yapılardır. Temelde canlıların biyolojik işlevlerini ve fiziksel özelliklerini belirlemekte görevli olan genler protein üretiminde aktif rol almaktadır. Genler üzerinde meydana gelen mutasyonlar canlılarda çeşitli genetik hastalıklara neden olmaktadır. Transkripsiyon faktörü (Transcription factor, TF) olarak adlandırılan gen ifadelerini düzenleyen proteinler genomik verileri yorumlayarak DNA dizisinin kodunun çözülmesi için gereken ilk adımları gerçekleştirir. Transkripsiyon faktörleri ve bu faktörlerin bağlanma bölgelerindeki mutasyonlar canlılarda görülen birçok genetik hastalığın temelini oluşturmaktadır. Örneğin insan, fare, tavuk, deniz kestanesi gibi metazoan dünyasındaki türlerde bulunan Ets ailesi proteinleri transkripsiyon faktörleri ile iş birliği içerisinde çalışmaktadır. Ets proteinlerindeki düzensizlikler, çeşitli kanser türlerinin oluşumuna neden olmaktadır. Lösemi, Ewing tümörü ve meme kanseri gibi ölümcül sonuçlar doğurabilecek hastalıklar, Ets protein ailesine ait proteinlerin anormal regülasyonu ile ilişkilendirilmektedir [45].

*Sorumlu Yazar (Corresponding Author)
e-posta : aysegul.hatipoglu@bilecik.edu.tr

Genlerin çalışma mekanizmalarının anlaşılması transkripsiyonel süreçlerin anlamlandırılması için önemlidir. Transkripsiyon faktörlerinin gen düzenlenmesi üzerinde biyolojik olarak önemi oldukça açıktır. Transkripsiyon faktörleri, DNA dizilerini tanıyarak bağlanan ve kromatinlerin yapılarını değiştirerek transkripsiyonu düzenleyen proteinlerdir. Gen üzerinde bulunan ve promotör olarak adlandırılan bölgeye bağlanarak genin ifade seviyesini düzenlerler. Hücrelerin tipine, gelişim aşamasına, hastalık durumuna göre transkripsiyonun başlatılması ve düzenlenmesinden sorumludurlar. Transkripsiyon faktörü bağlanma bölgeleri ise transkripsiyon faktörleri ile DNA üzerindeki bağlanma alanları arasında oluşan birleşme pozisyonudur. Dolayısıyla bu bağlanma noktalarının başarılı bir şekilde tahmini genetik hastalıkların teşhis ve tedavisi için önemli bir adımdır. Son araştırmalar transkripsiyon faktörü bağlanma bölgelerinin belirlenmesinde önemli gelişmeler olmasına rağmen transkripsiyon faktörü bağlanmalarının çok daha karmaşık olduğunu ve başlangıçta bilinenden daha fazla düzenleyici ve yapısal değişiklik barındırdığını göstermektedir [7, 8, 9, 10].

Aynı yapısal ailelerden gelen transkripsiyon faktörleri çoğunlukla benzer DNA dizilerine bağlanma eğilimi gösterirler ancak her birinin kendine özgü işlevleri vardır. Şekil 1'de 3 faklı temel bHLH (basic helix-loop-helix) ailesi transkripsiyon faktörü örneği gösterilmektedir.



Şekil 1. Transkripsiyon faktörü bağlanma mekanizmaları (Transcription factor binding mechanisms) [44]

Bu yapılar benzer simetrik motifleri (örneğin CACGTG motifi) tanımakta ancak birbirlerinden farklı görevleri yerine getirmektedirler. Şekil 1'de ki yapılara bakıldığında MyoD proteini kas hücrelerini belirlerken Clock:Bmal1 sirkadiyen ritmini (uyku uyanıklık döngüsü, vücut ısısının ayarlanması, hormon salınımı vb. süreçler) düzenlemekte, Max ise hücre çoğalması ve farklılaşmasını kontrol etmektedir. Bu yapılarda meydana gelen anomaliler yalnızca canlının yaşam kalitesini etkilemekle kalmaz aynı zamanda hayati bir rol de oynayabilir. Ets ailesinden gelen Ets1 proteini de farelerde onkojenik potansiyeli olmasına rağmen insanlarda onkogenezde rol oynadığına dair bir kanıt bulunmamaktadır. Hatta Ets1'in bir varyantının insanda kolon kanserini baskıladığı bilinmektedir [45].

Aynı yapısal ailelerden gelen, benzer DNA dizilerine bağlanma eğilimi gösteren ancak farklı canlılarda farklı işlevleri yerine getiren hatta bazı durumlarda zıt roller oynan Transkripsiyon faktörlerinin çalışma mekanizmalarını, fizyolojik rollerini anlamak, genomların spesifik özelliklerinin kodlarını çözmek ve haritalamak için önemlidir [4]. Transkripsiyonel süreçlerin anlaşılmasıyla potansiyel bağlanma mekanizmalarını ve sonraki hücresel fonksiyonların modellenmesini anlama sürecinde önemli bir adım atılmaktadır. Çeşitli makine öğrenmesi ve derin öğrenme yöntemlerinin kullanılması ile başarı oranı artırılan transkripsiyon faktörü bağlanma bölgelerinin tahmini, biyoinformatik alanındaki popüler ve zorlu çalışma konularındandır. Transkripsiyon faktörlerinin olası DNA bağlanma bölgelerini tahmin etmek, biyolojik sistemlerin karmaşık yapılar içermesi nedeniyle hesaplamalı biyolojide zor bir araştırma konusudur. Bu nedenle bahsedilen tahmin problemine hesaplama teknikleri kullanarak çözüm geliştirmek aktif bir çalışma konusudur [5].

Günümüzde yapay zekâ ve veri bilimi alanlarında yaşanan gelişmeler, bu alanlarla etkileşim halinde bulunan diğer bilim dalları üzerinde de önemli etkiler oluşturmuştur. Bu gelişmelerin etkilediği alanlar arasında tıp, genetik, biyoinformatik bilimi gibi insan sağlığı konusunda çalışmalar yapan alanlar da bulunmaktadır. Yapay zekâ ve veri bilimi alanındaki gelişmeler ve hızla artan veri miktarı, insan sağlığı için önemli olan hastalıkların teşhis ve tedavi süreçlerinde önemli yenilikler sunarken, biyoinformatik alanında ise genetik verinin anlaşılması ve biyolojik süreçlerin analizi gibi konulara katkı sağlamaktadır. Bahsedilen bu gelişmelerle beraber veri miktarının da hızla artması sonucu, geleneksel makine öğrenmesi yöntemleri yerini büyük veri konusunda oldukça başarılı çıktılar üreten derin öğrenme yöntemlerine bırakmıştır. Yüksek verimli teknolojilerin gelişimiyle genom dizileri, protein yapıları ve tıbbi görüntüleme alanındaki ilerlemeler sonucu biyoinformatik alanında birikmiş büyük miktarda biyomedikal verinin derin öğrenme uygulamaları ile birleşiminden elde edilecek çıktılar hem akademik camianın hem de endüstri camiasının dikkatini çekmektedir. Büyük miktardaki biyomedikal verinin

anlamlı bilgi haline dönüştürülmesi için, verilerin depolanmasından analiz ve yorumlanmasına kadar etkili ve verimli hesaplama araçlarına ihtiyaç duymaktadır. Derin öğrenme yöntemlerindeki karmaşık yapılar bahsedilen bu veri setlerinden özellik çıkarımı sürecini daha etkili hale getirmekte böylelikle tıbbi verilerin analizinde bu metotların tercih edilmesi giderek yaygınlaşmaktadır [1, 2, 3].

Transkripsiyonel süreçleri anlamak için gerçekleştirilen çalışmalarda gelinen son nokta henüz bu süreçleri tamamen anlamak için yeterli seviyede olmadığımızı gözler önüne seriyor. Bu sebeple Transkripsiyon Faktörü Bağlama Bölgeleri (Transcription Factor Binding Sites, TFBS) Tahmini alanında gerçekleştirilen çalışmalar hala geliştirilmeye açıktır. Ancak bu çalışmalar da karşılaşılan bir diğer problem ise biyolojik deneyler yapmanın pahalı olduğu ve uzun zaman gerektirdiği gerçeğidir. Bu gibi sebepler ve veri biliminin gelişimiyle artan biyolojik veri setlerinin varlığı, araştırmacıları gen dizilimlerini analiz etmek için derin öğrenme teknolojileri kullanmaya ve hesaplamalı modeller geliştirmeye yöneltmektedir [6].

Bu çalışmada transkripsiyon faktörü bağlama bölgelerini tahmin etmek için farklı mimarileri birleştirdiğimiz hibrit bir yöntem öneriyoruz. DeepTFBS (Deep Transcription Factor Binding Sites) adını verdiğimiz modelde farklı derin öğrenme teknolojilerini kullanarak farklı mimarilerin güçlü yanlarını bir araya getiriyoruz. Modelimiz literatürde geliştirilen diğer yöntemlerle karşılaştırıldığında başarılı çıktılar elde etmektedir.

2. LİTERATÜR TARAMA (LITERATURE REVIEW)

Makine öğrenmesi ve derin öğrenme teknolojileri son yıllarda kullanıldığı birçok alanda olduğu gibi genetik alanında da büyük başarılar elde etmiştir. Tıbbi görüntülerin ve biyomedikal verilerin sağlık sektöründeki araştırma ve uygulama potansiyeli endüstri sektörüyle beraber akademik camianın da ilgi odağı haline gelmiştir. Transkripsiyon faktörü bağlanma bölgeleri tahmini için literatürde gerçekleştirilen çalışmalar incelendiğinde çeşitli yöntemler geliştirildiği ancak makine öğrenmesi ve derin öğrenme yöntemlerinin son yıllarda sıklıkla tercih edildiği gözlemlenmiştir. Alipanahi ve arkadaşları 2015 yılında DeepBind adını verdikleri metotla derin öğrenme teknolojilerini kullanarak DNA ve RNA bağlayıcı proteinlerin dizi özelliklerini tahmin etmeye yönelik bir çalışma sunmuşlardır. Çalışma hedef bağlanma motiflerinin tahmini için esnek ve birleşik bir hesaplama yöntemi sunmaktadır [11].

Hassanzadeh ve Wang 2016 yılında gerçekleştirdikleri çalışmayla protein bağlanma bölgeleri tahmini için başarılı derin öğrenme tekniklerinden biri olan Uzun kısa süreli bellek (Long Short-Term Memory, LSTM) ve Evrişimsel sinir ağı (Convolutional Neural Network, CNN) mimarilerini kullandıkları bir model olan DeeperBind'ı önermişlerdir [12].

Quang ve Xie 2016 yılında gerçekleştirdikleri çalışma ile CNN ve çift yönlü uzun kısa süreli bellek (BiLSTM) yapısını birleştiren hibrit bir yaklaşım önermişlerdir. Önerdikleri modele DanQ adını veren araştırmacılar evrişim katmanı ile düzenleyici motifleri yakalarken, yineleme katmanı sayesinde motifler arası uzun vadeli bağımlılıkları yakalamaktadır. DanQ çeşitli ölçümlerde diğer modellere kıyasla gelişme göstermiştir [13].

Zhang ve arkadaşları 2018 yılında yaptıkları çalışmada HOCNN (High-order Convolutional Neural Network) adını verdikleri bir yöntem önermişlerdir. Önerilen çalışmada nükleotitler arasındaki yüksek dereceli bağımlılıkları ele almak için yüksek dereceli bir kodlama yöntemi ve farklı uzunluktaki motif özelliklerinin keşfi için yüksek dereceli bir evrişimsel sinir ağı mimarisi geliştirmişlerdir. Araştırmacılar çalışmalarını ChIP-seq (165 ENCODE chromatin immunoprecipitation sequencing) veri kümesi üzerinde değerlendirmişler ve elde ettikleri sonuçların en gelişmiş diğer mimarilerden daha iyi performans gösterdiğini belirtmişlerdir [14].

Trabelsi ve arkadaşları 2019 yılında DNA ve RNA dizisi bağlanma özellikleri tahmini için deepRAM adını verdikleri derin öğrenme aracını sunmuşlardır. Çalışmada derin ve karmaşık mimarilerin yeterli eğitim verileri sağlandığında avantajlı olduğunu ve CNN/RNN mimarilerinin diğer yöntemlere göre daha iyi performans gösterdiği kanıtlanmıştır [15].

Abdollahyan ve arkadaşları 2018 yılında gerçekleştirdikleri çalışmalarında TFBS tahmini amacıyla grafik tabanlı bir yaklaşım önermişlerdir. Grafikleri hizalamak ve TFBS dizilerini belirlemek için dinamik bir programlama algoritması kullanmışlarıdır [16].

Zhang ve arkadaşları 2018 yılında WSCNN (Weakly-Supervised Convolutional Neural Network) olarak adlandırdıkları bir mimariyi önermişlerdir. Çalışmalarında çok örnekli öğrenme (Multiple-Instance Learning, MIL) ile CNN yapısını bir araya getiren bir yaklaşım benimsemişlerdir. Çalışma DNA dizilerini birden fazla alt diziye böler ve CNN yardımıyla her örneği ayrı ayrı modeller. Son olarak Max, Avarage, Linear Regression ve Top-Bottom Instances olmak üzere dört füzyon yöntemiyle aynı torbadaki tüm örneklerin tahmin puanlarını birleştirir [17].

Zhang ve arkadaşları 2019 yılında yaptıkları çalışmada protein-DNA bağlanmasını modellemek için çoklu örnekli öğrenmeyi hibrit bir derin sinir ağı mimarisiyle birleştirdikleri bir yöntem tanıtmışlardır. Çalışmada DNA dizilerini dönüştürmek için K-mer kodlaması kullanmışlar ve yüksek dereceli bağımlılıkların görüntü benzeri girdilerine kodlamışlardır [18].

2019 yılında Chen ve arkadaşları yaptıkları çalışmada DeepGRN adını verdikleri derin öğrenme modelini tanıtmışlardır. Bu çalışma da evrişimli sinir ağlar (CNN), tekrarlayan sinir ağları (RNN) ve dikkat mekanizması gibi teknikleri birleştirerek hibrit bir yöntem geliştirmişlerdir. Yaptıkları testlerde sundukları modelin diğer başarılı modeller ile kıyaslandığında başarı performans ortaya koyduğu görülmektedir [19].

Zhou ve arkadaşları 2019 yılında gerçekleştirdikleri çalışmada histon modifikasyon özelliklerine CNN uygulayarak daha yüksek dereceli bağımlılıkları çıkarmak için bir yöntem önermişlerdir. Düşük ve yüksek dereceli bağımlılıkları birleştiren modeli CNN_TF olarak adlandırmışlardır. Çalışma son teknoloji birkaç yöntem ile karşılaştırıldığında CNN_TF, AUPR'de diğer metotları %5,3 oranında geride bırakmaktadır [20].

Zhang ve arkadaşları 2020 yılında gerçekleştirdikleri çalışmada DeepSite adını verdikleri bir mimari önermişlerdir. DeepSite mimarisinde DNA'daki dizi motifleri arasındaki uzun vadeli bağımlılıkları yakalamak için çift yönlü uzun kısa süreli bellek (Bi-LSTM) ve CNN yapısı beraber kullanılmaktadır. Sistem 690 Chip-seq deneyleri ile test edilmiştir ancak araştırmacılar yöntemlerinin farklı DNA dizilerindeki DNA-protein bağlanma bölgelerini bulmak için de uygulanabileceği belirtmişlerdir [21].

Aziz ve Rashid 2021 yılında gerçekleştirdikleri TFBS tahmini çalışmasında Destek vektör makineleri (Support Vector Machine, SVM) ve Kernel Lojistik Regresyon (Kernel Logistic Regression, KLR) yöntemlerini kullanmışladır [22].

Wang ve arkadaşları 2022 yılında yaptıkları çalışmada transkripsiyon faktörü bağlanma bölgeleri tahmini için Evrişimli Sinir Ağları (Convolutional Neural Networks, CNNs) ile Tekrarlayan Sinir Ağlarının (Recurrent Neural Networks, RNNs) bir çeşidi olan Uzun kısa süreli bellek (Long short-term memory, LSTM) yapısını birleştirmişledir. Araştırmacılar çalışmalarını başarılı diğer yöntemler ile kıyasladıklarında daha iyi performans elde etmişlerdir [23].

Song ve Du 2022 yılında DS-SSB adını verdikleri bir yöntem önermişlerdir. DS-SSB yöntemi Çift akışlı çoklu örnek ağını (Dual-stream multiple instance network) çoklu özelliklerle birleştirir. 690 ChIP-seq veri kümesiyle gerçekleştirilen deneyler, TFBS tahmininde başarılı sonuçlar elde edildiğini göstermiştir [24].

Cheng ve çalışma arkadaşları 2022 yılında AttBind adını verdikleri CNN, Bi-GRU ve Dikkat mekanizmasını birleştirdikleri bir öğrenme algoritması önermişlerdir. Deneysel sonuçlar modelin ENCODE DREAM Challenge'da mevcutta bulunan ilk 5 yöntemden daha başarılı olduğunu göstermiştir [25].

Li ve arkadaşları 2022 yılında TFBS tahmini için Yoğun ağ (Densely network) kullanan ilk çalışma olan ve DCNN-SH olarak adlandırdıkları çalışmayla yeni bir sinir ağı mimarisi önermişlerdir. Model mevcut yöntemlere kıyasla daha küçük evrişimsel çekirdekler kullanarak çok ölçekli özellikleri dikkate alınmasını sağlamakta ve diğer birçok yönteme göre daha başarılı performans gösterdiği belirtilmektedir [26].

Cao ve arkadaşları 2022 yılında DeepARC adını verdikleri dikkat tabanlı hibrit bir sistem önermişlerdir. DeepARC, DNA2Vec ve OneHot kodlamasını birleştirerek her iki yöntemin güçlü yanlarından faydalanmıştır. Dikkat tabanlı CNN-BiLSTM ağı mimarisini kullanan DeepARC'ın mevcut son teknoloji yöntemlerden daha başarılı performans ortaya koyduğu belirtilmektedir [27].

Yu ve arkadaşları 2023 yılında gerçekleştirdikleri çalışmada DSAC olarak adlandırdıkları model ile Öz Dikkat ve Konvolüsyonu birleştirerek sadece dizi özelliklerine dayalı TFBS tahmini işlemi gerçekleştirmişlerdir [28].

2023 yılında yaptıkları çalışmada Tariq ve Amin dizi özelliklerinin çıkarılması ve TFBS'lerin tanımlanması için k-mer kodlaması ile birleştirdikleri Evrişimli Sinir Ağı (Convolutional Neural Network, CNN) modelini önermişlerdir [5].

Ding ve arkadaşları 2023 sundukları çalışmada CNN, Bi-LSTM, Transformer gibi teknolojileri birleştirdikleri DeepSTF modelini önermişlerdir. Bu hibrit model DNA'nın dizi ve şekil özelliklerini bir arada kullanarak transkripsiyon faktörü bağlanma bölgeleri tahmini için yüksek performans elde etmiştir [55].

2024 yılında Wang ve arkadaşları tarafından geliştirilen BERT-TFBS modeli transfer öğrenme yöntemini kullanarak sadece DNA dizilerine dayalı TFBS tahmini gerçekleştirmektedir. BERT-TFBS modeli DNA dizilerindeki uzun vadeli bağımlılıkları yakalamak amacıyla önceden eğitilmiş DNABERT-2 modülünü kullanırken özellik çıkarımı için CNN ve CBAM (Convolutional block attention module) yapısını kullanmaktadır [47].

Son yıllarda yapılan referans çalışmalar, derin öğrenme mimarilerinin biyoinformatik alanındaki problemlerin çözümünde sıklıkla kullanıldığını gösteriyor. Bu alandaki önemli zorluklardan birisi olan Transkripsiyon Faktörü Bağlanma Bölgeleri tahmini probleminde de DNA dizilerinin doğasının anlaşılması ve bağlanma bölgelerinin tahmini için derin öğrenme yöntemlerinin kullanımının popüler bir araştırma alanıdır. Ancak şu ana kadar geliştirilen yöntemlerin henüz yeterli düzeyde olmadığı ve gelecekte yapılacak çalışmaların umut vadedici olduğu görülmektedir.

3. MATERYAL VE METOD (MATERIAL and METHOD)

3.1. Veri Seti (Data Set)

Genomik alanında gerçekleştirilen öğrenme görevleri genellikle on binlerce ve üzerinde eğitim örneğine sahiptir. Bu kadar büyük veriler ile eğitim gerçekleştirilmesi aşırı uyum sağlamasının önüne geçer. Bu gibi büyük boyutlu eğitim örnekleri genellikle Encyclopedia of DNA Elements (ENCODE) projesi tarafından üretilenler gibi yüksek verimli verilerden çekilir [43]. Bu çalışmada da bu alanda popüler veri kümelerinden biri olan 690 ChIP-seq veri kümesinden elde edilen DNA dizisi verileri kullanılmıştır. Kullanılan 165 ChIP-seq veri kümesi ENCODE projesi kapsamında gerçekleştirilen 165 farklı kromatin immünopresipitasyon dizileme (Chromatin Immunoprecipitation Sequencing, ChIP-seq) deneyini ifade eder [29]. 165 veri seti barındıran ChIP-seq, farklı

hücre dizilerinden gelen 29 transkripsiyon faktörüne ait bağlanma bölgesi verileri içermektedir. Veri setinin İnsan lösemi hücreleri, insan akciğer adenokarsinom hücreleri, insan embriyonik kök hücreleri, memeli epitel hücreleri ve daha birçok organizmadan oluşması, hücre farklılaşması, kanser biyolojisi, bağışıklık sistemi tepkileri ve hastalık patogenezi gibi geniş bir biyolojik süreci incelemeye ve anlamaya olanak tanır. Bu çeşitlilik, yalnızca belirli bir hücre veya hastalık türüne odaklanmak yerine, farklı dokular ve kanser türlerinde genetik ve epigenetik düzenleyici mekanizmaların karşılaştırmalı olarak analiz edilmesini mümkün kılar.

Transkripsiyon faktörlerinin bağlanma bölgeleri ile ilgili DNA dizi özellikleri ve DNA şekil özellikleri birlikte bu çalışmanın veri setini oluşturmuştur. Bölüm 3.1.1 ve bölüm 3.1.2'de bu özellikler ve kullanımları ile alakalı detaylı bilgi verilmiştir.

3.1.1. DNA sekansları (DNA sequences)

DNA dizileme ilk olarak 1977'de tanıtılmıştır. Dizileme teknolojisinde geliştirilen teknikler sayesinde birçok deney daha kolay hale gelmiş ve hesaplamalı biyolojiye önemli katkılar sağlamıştır. Bir DNA dizisi A (Adenine), G (Guanine), C (Cytosine), T (Thymine) olmak üzere 4 farklı nükleotitin farklı kombinasyonlarının bir araya gelmesi ile oluşan dizilerdir. Bu dizimler canlılara ait biyolojik özelliklerin belirlenmesini sağlar ve benzersizdir [30, 31]. DNA dizilerinin bu özgün yapıları transkripsiyon faktörü bağlanma bölgeleri tahmini işleminde de önem arz etmektedir. DNA'daki nükleotit yapılarının doğru yöntemlerle incelenerek, analiz edilerek işlev özelliklerinin anlaşılması bu TFBS tahmin işleminde kritik öneme sahiptir. Bu nedenle TFBS tahmin çalışmalarında DNA sekanslarının detaylı analiz edilmesi gerekmektedir. Literatürdeki çalışmalar incelendiğinde sadece DNA dizileri kullanılarak gerçekleştirilen tahmin çalışmaları olduğu gibi farklı DNA özellikleri ile birlikte kullanıldığı çalışmalara da rastlanmaktadır.

3.1.2. DNA şekli (DNA shape)

DNA şekli DNA moleküllerinin 3 boyutlu uzaydaki yapısı olarak bilinmektedir. Sadece DNA dizileri kullanılarak gerçekleştirilen transkripsiyon faktörü bağlanma bölgeleri tahmini çalışmalarının yanı sıra motif keşfi için DNA şekil özellikleri kullanan çalışmalara da rastlanmaktadır. Yapılan birçok araştırma transkripsiyon faktörleri ve bağlanma dizileri arasındaki etkileşimin yüksek oranda korunduğunu göstermektedir. Çalışmalar bu korunmanın DNA molekülünün üç boyutlu yapısıyla bağlantılı olduğunu göstermektedir. DNA şekli transkripsiyon faktörleri ile DNA-bağlayıcı proteinlerin bağlanma tercihlerini belirlemede önemlidir. Bu nedenle bu alanda yapılan çalışmalarda TFBS tahmin doğruluğunu artırmak için DNA dizisi ve şekli entegre edilmektedir [32, 22].

DNA'nın birçok yapısal özelliği bulunmakta ve bu özellikler DNA şekline ait bazı verilerle ilişkilendirilmektedir. Dolayısıyla DNA şekil özellikleri bizlere DNA yapısı ve gerçekleştirdiği fonksiyonlar hakkında bilgiler verir. Bu şekil özelliklerinin doğru şekilde analiziyle transkripsiyon faktörlerinin bağlanma mekanizmaları hakkında detaylı bilgiler edinilebilir. Literatür incelemeleri transkripsiyon faktörü bağlanma bölgelerinin tahmininde DNA şeklinin belirlenmesi amacıyla beş temel DNA şekil özelliğinin yaygın olarak kullanıldığını ortaya koymaktadır. Aşağıda detaylandırılan bu beş DNA şekil özelliği, bazı çalışmalarda dörtlü kombinasyonlar [48, 49, 50, 51, 52, 53] halinde, bazı çalışmalarda ise beş özellik [54, 55, 56] bir arada kullanılarak tahmin sürecine dahil edilmiştir.

Çalışmamızda DNA şeklinin belirlenmesine yardımcı olan ve literatürde de transkripsiyon faktörü bağlanma bölgeleri tahmini için sıklıkla kullanımına rastlanılan 5 özellik;

1) **Sarmal Büküm (Helix twist, HelT):** DNA'nın spiral yapısını tanımlayarak baz çiftlerinin birbirine göre dönme açılarını düzenlemektedir. DNA'nın üç boyutlu yapısındaki bu değişiklikler, proteinlerin belirli DNA bölgesine bağlanmasını etkilemektedir [60].
2) **Küçük oluk genişliği (Minor grove width, MGW):** MGW transkripsiyon faktörlerinin küçük oluğa erişimiyle ilişkilendirilir ve yüksek afiniteye sahip bağlanma bölgelerinde küçük oluktaki genişleme veya daralmalar transkripsiyon faktörlerinin bağlanma tercihini etkilemektedir [59, 61].
3) **Pervane bükümü (Propeller twist, ProT):** DNA'nın bükülme ve kıvrılma şeklini değiştirmekte ve bu sayede transkripsiyon faktörlerinin spesifik bağlanma motiflerini tanımasını sağlamaktadır [60].
4) **Yuvarlanma (Rolling, Roll):** Roll açısı, baz çiftleri arasındaki eğimi tanımlamakta ve bu eğim transkripsiyon faktörlerinin bağlanma doğruluğunu etkilemektedir [60].
5) **Küçük oluk elektrostatik potansiyeli (Minor Groove electrostatic potential, EP):** DNA'nın negatif yüklü omurgası, pozitif yüklü amino asitlere sahip proteinlerle elektrostatik çekim yoluyla etkileşime girer. Bu etkileşim transkripsiyon faktörleri gibi DNA'ya bağlanan proteinlerin bağlanma bölgelerini tanımasında etkilidir [59].

DNA şekil özelliklerinin TFBS tahmini için olumlu etkisi yapılan çalışmalarla kanıtlanmıştır [57, 58]. Bu özellikler DNA sekanslarından oluşan veri setindeki verilerin, DNAShaper kullanılarak hesaplanmasından elde edilmiştir [33]. Seçilen özellikler DNA'nın üç boyutlu yapısını belirlemekte ve transkripsiyon faktörü bağlanma bölgelerini etkileyen temel yapısal ve fiziksel özellikleri kapsamaktadır. Dolayısıyla bu özellikler protein-DNA etkileşimini anlamak için kritiktir ve proteinlerin bağlanma motiflerini daha iyi tanımlamayı sağlayarak modelin tahmin performansını etkilemektedir.

Çalışmada kullandığımız DNA dizi özellikleri ve şekil özelliklerinden oluşan veri seti, modelin performansını değerlendirmek ve genel başarısını test edebilmek amacıyla eğitim veri seti ve test veri seti olarak 2 gruba ayrılmıştır. Tüm veri setinin %80'ini oluşturan eğitim verisi, modelin öğrenme sürecindeki ana veri grubudur. Eğitim verisi modelin özellikleri tanımasını ve doğru tahminler için gerekli kalıpları öğrenmesini sağlar. Tüm veri setinin %20'sini oluşturan test verisi ise modelin eğitim sırasında öğrenmediği yeni verileri içerir. Test veri seti, modelin eğitim sırasında öğrendiği bilgileri yeni verilere uygulama becerisini ölçer. Önerdiğimiz yöntem de DNA dizi özellikleri ile beraber DNA yapı özellikleri kullanılarak transkripsiyon faktörlerinin bağlanma bölgelerini tahmin işlemi için yüksek doğrulukta bir derin öğrenme mimarisi geliştirilmiştir.

3.2. Uygulama Mimarisi (Application Architecture)

DeepTFBS adını verdiğimiz mimaride, transkripsiyon faktörü bağlanma bölgelerinin tahmin edilmesi için farklı derin öğrenme mimarilerini birleştirdiğimiz hibrit bir yaklaşım benimsenmiştir. Bu yaklaşımda, DNA dizilerini alt parçalara bölmek için kullandığımız K-mer analizi, Evrişimli sinir ağı (Convolutional neural network, CNN), Geçitli yinelemeli birimler (Gated recurrent unit, GRU) ve dikkat mekanizması kullanan bir transformer mimarisi kombinasyonunu içermektedir.

3.2.1. K-mer

K-mer, biyoinformatik alanında kullanılan bir terimdir. Bir dizinin "k" uzunluğunda tüm alt parçalarını ifade etmektedir. DNA ya da RNA dizileri belirli k-mer uzunluğunda ardışık olan parçalara bölünerek biyoinformatik alanında kullanılır. Gen hataları, bireyler arasındaki genetik benzerliklerin ve farklılıkların analizi, tür çeşitliliği gibi alanlarda kullanımına rastlanılmaktadır [34, 35]. Çizelge 1'de 10 adet nükleotitten oluşan bir dizi için K-mer verileri gösterilmektedir.

Çizelge 1: GTAGAGCTGT dizisi için k-mers (k-mers for the sequence GTAGAGCTGT) [6]

1-mer	G, T, A, C
2-mer	GT, TA, AG, GA, AG, GC, CT, TG
3-mer	GTA, TAG, AGA, GAG, AGC, GCT, CTG, TGT
4-mer	GTAG, TAGA, AGAG, GAGC, AGCT, GCTG, CTGT
5-mer	GTAGA, TAGAG, AGAGC, GAGCT, AGCTG, GCTGT
6-mer	GTAGAG, TAGAGC, AGAGCT, GAGCTG, AGCTGT
7-mer	GTAGAGC, TAGAGCT, AGAGCTG, GAGCTGT
8-mer	GTAGAGCT, TAGAGCTG, AGAGCTGT
9-mer	GTAGAGCTG, TAGAGCTGT
10-mer	GTAGAGCTGT

DNA dizileri uygun alt parçalara bölündüğünde hesaplama açısından kolaylık sağlamaktadır. Ancak seçilen parça uzunlukları işlemin başarısını etkilemektedir. Bu nedenle çok uzun ya da çok kısa boyut seçmek her durumda işlevsel değildir. Problem uzayına göre uygun değerler belirlenmelidir. Farklı k değeri tercihi, dizinin ardışık motiflerle temsil edebileceği uzunluk miktarını belirler. Çalışmada DNA dizisi, belirli bir uzunlukta ve adım aralığında olan k-mer dizilerine bölünmüştür. Bu k-mer dizileri, kelime olarak ele alınıp bir kelime temsil modeli elde edilmiştir. Çalışmada farklı k değerleri için gerekli testler yapılmış ve optimal k değeri 5 olarak elde edilmiştir. k değeri ardışık motiflerin anlaşılması ve özelliklerin öğrenilmesi için seçilmiştir.

3.2.2. Geçitli yinelenen birimler (Gated recurrent unit, GRU)

Geçitli yinelen birimler zamansal veriler ve uzun, sıralı veriler için başarılı sonuçlar üreten yapılardır. 2014 yılında LSTM yapılarının karmaşıklığını azaltmak için geliştirilmişlerdir. Başarılarını mevcut girişin ve hafızalarından bulunan önceki bilgi akışının organize edilmesinden sorumlu olan geçit yapılarına borçludurlar. LSTM yapıları RNN yapılarının 3 kapı kullandığı ağlar iken GRU yapıları ise önceki durum ile ilgili işlemleri gerçekleştiren güncelleme kapısı (update gate) ve yeni giriş verileriyle önceki bilgileri birleştiren, silinmesi gereken verilere karar veren sıfırlama kapısı (reset gate) olmak üzere 2 kapıdan oluşmaktadır [36].

Bi-GRU yapıları ise iki GRU katmanından meydana gelmektedir. İleri ve geri yönlü olan bu GRU katmanları bilginin çift yönlü öğrenilmesini ve bağlamın daha iyi anlaşılmasını sağlar. LSTM, GRU gibi modeller geleneksel RNN yapılarını güçlendirirken parametre artışından dolayı hesaplama maliyetini artırırlar. GRU yapıları LSTM yapılarına göre sade ve performansı artırılmış modellerdir.

3.2.3. Evrişimli sinir ağları (Convolutional neural network, CNN)

Evrişimli sinir ağları derin öğrenme alanındaki en önemli ağ modellerinden birisidir. Bilgisayarlı görü, yüz tanıma, otonom sistemler, akıllı ilaç tedavileri gibi geniş bir çalışma alanında kullanılmaktadırlar. Barındırdıkları konvolüsyon yapılarıyla verilerden özellik çıkarabilen ileri beslemeli sinir ağlarıdır. Geleneksel yöntemlerden farklı olarak manuel özellik çıkarımına ihtiyaçları yoktur. En eski evrişimli sinir ağı modeli 1998'de önerilen leNet-5 modelidir. CNN temelli yapılar yıllar içinde geliştirilen yöntemler ile üstün başarılı çıktılar sunmuştur. Temel olarak bir CNN modeli Konvolüsyon Katmanı (Convolution Layer), Havuzlama Katmanı (Pooling Layer) ve Tam Bağlı Katmanlar (Fully Connected Layers) olmak üzere 3 ana katmandan meydana gelmektedir. Evrişim katmanları CNN mimarilerinin temel öğeleridir. Giriş verisi üzerinde filtrelerin hareketiyle özellik haritalarının oluşturulmasını sağlar. Havuzlama katmanları yardımıyla özellik haritalarının boyutları azaltılarak öğrenilecek parametre sayısı ve dolayısıyla hesaplama maliyeti düşürür. Tam bağlantılı katman ise klasik yapay sinir ağı katmanıdır. Bu katmanda bulunan her nöron önceki katmanda bulunan tüm nöronlara bağlı olduğundan bu şekilde isimlendirilmiştir. [1, 37, 38].

3.2.4. Transformer mimarisi (Transformer architecture)

2017 yılında yayınlanan çalışma ile tanıtılan transformer mimarisi, yineleme ve evrişimden soyutlanan dikkat mekanizmasına dayalı bir ağ mimarisidir. Mimari Çok Başlı Dikkat Mekanizması (Multi-Head Attention Mechanism), Kodlayıcı (Encoder) yapısı, Kod Çözücü (Decoder) yapısı, Tam Bağlantılı Katmanlardan (Fully Connected Layers) oluşmakta ve derin öğrenme modellerinde birçok zorlu görevin başarı ile üstesinden gelmektedir. Kodlayıcı giriş verilerini, embedding denen özel temsillere dönüştürür. Kod çözücü, elde edilen bu temsilleri bazı işlemlerden geçirerek yeni bir forma dönüştürür. Barındırdığı dikkat mekanizması yapısı ile uzun vadeli bağımlılıkları etkili bir şekilde ele alır ve başarılı çıktılar üretir [39]. Önerilen model eğitilirken kullanılan parametreler Çizelge 2'de gösterilmiştir.

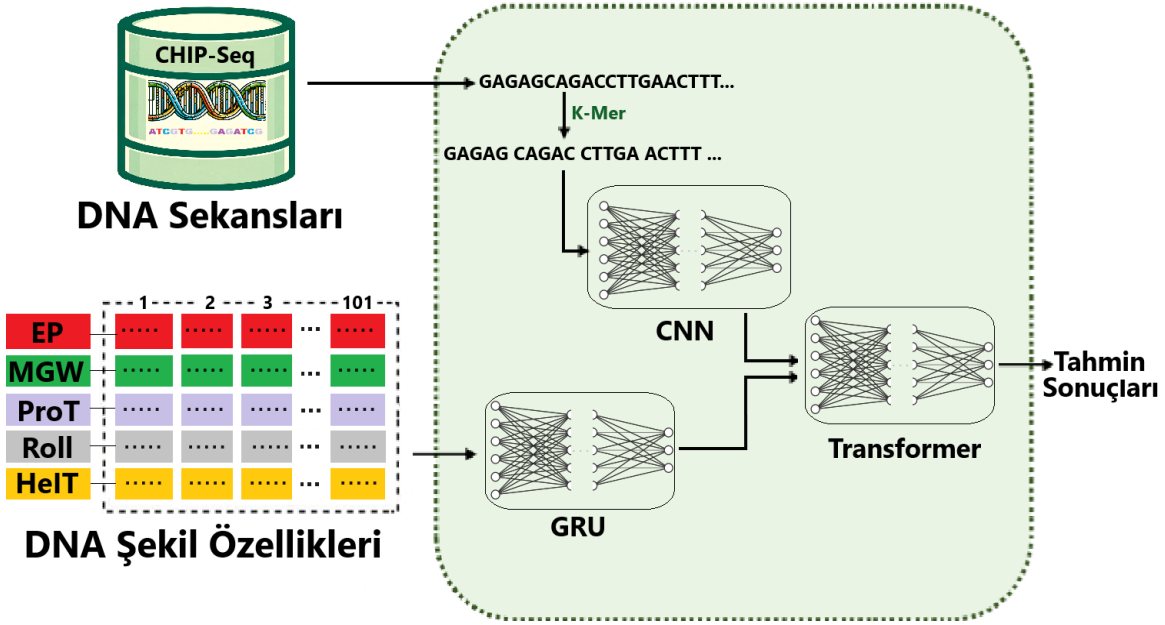
Çizelge 2: Model eğitim parametreleri (Model training parameters)

Öğrenme Oranı	Dinamik
Optimizasyon Fonksiyonu	Adam
Dropout oranı	0.2
Çekirdek Sayısı	128
Eğitim Döngüsü	50
Aktivasyon Fonksiyonu	RELU, GELU

Çalışmada dinamik öğrenme stratejileri kullanılarak modelin eğitim sürecinde hiper parametrelerin sabit tutulmadan, eğitim verisinin özelliklerine ve eğitim sürecine bağlı olarak sürekli değişmesi sağlanmıştır. Aşırı öğrenmenin önüne geçmek ve modelin genelleme kapasitesini artırmak amacıyla farklı dropout oranları test edilmiş ve en uygun oran belirlenmiştir. Eğitim döngüsü (epoch) sayısı belirlenirken erken durdurma (early stopping) yöntemi kullanılarak maksimum öğrenme sağlanmış ve en uygun noktada eğitim süreci otomatik olarak durdurulmuştur. Bu yöntemle ideal eğitim döngüsü sayısına dinamik olarak ulaşılmıştır. Optimizasyon algoritması olarak Adam optimizasyonu seçilmiş ve öğrenme oranı her parametre için otomatik olarak belirlenmiştir.

Önerdiğimiz çalışmada Evrişimli sinir ağları özellik çıkarımı için DNA sekanslarının evrişimsel özelliklerini öğrenirken, Geçitli yinelenen birimler ardışık bilgiyi modellemekten sorumludurlar. Transformer yapısı, dikkat mekanizması aracılığıyla önemli bölgelere öncelik vererek daha geniş bir bağlamın dikkate alınmasını sağlayarak Transkripsiyon faktörü bağlanma bölgelerinin tahmininde daha yüksek performans ve hassasiyet sağlamaktadır. Bu hibrit çalışma mevcut genetik mekanizmaların anlaşılmasına katkı sağlamanın yanı sıra gelecekteki çalışmalarda karmaşık genetik düzenleme mekanizmalarının anlaşılmasına ve biyolojik süreçlerin analiz edilmesine katkıda bulunabilir.

Özetle modelimiz Transkripsiyon faktörü bağlanma bölgeleri tahmini için DNA dizisinin özelliklerini çıkarırken Evrişimli sinir ağları kullanılmıştır. Bu yapı ile DNA üzerindeki motiflerin tanınması ve yapısal özelliklerinin çıkarılmasına imkân sağlanmıştır. Sonraki süreçte DNA şekil özelliklerinin kapsamlı bir şekilde çıkarılması ve TFBS tahmin doğruluğunu artırmak için Bi-GRU (Bidirectional Gated recurrent unit) ve Transformer mimarisi birleştiren bir yaklaşım geliştirilmiştir. Bu hibrit yöntemle yinelenen sinir ağlarının DNA dizileri üzerindeki güçlü öğrenme kapasiteleri ve Transformer mimarisinin uzun vadeli bağımlılıkları yakalama özelliği birleştirilerek verimlilik artırılmıştır. Şekil 2'de DeepTFBS modeline ait genel sistem mimarisi gösterilmektedir



Şekil 2. DeepTFBS sistem mimarisi (DeepTFBS system architecture)

4. DENEYSEL SONUÇLAR (EXPERIMENTAL RESULTS)

Derin öğrenme alanında geliştiren algoritma ve modellerin sonuçlarını doğru şekilde değerlendirebilmek adına herkesçe kabul gören ve mimariye uygun değerlendirme metrikleri seçimi kritik öneme sahiptir. Böylelikle benzer çalışmaların sonuçları doğru şekilde kıyaslanabilir. Çalışmamızda önerilen sistemi değerlendirmek için bu alanda sıklıkla tercih edilen ACC, ROC-AUC, PR-AUC değerlendirme metrikleri kullanılmıştır.

ACC (Accuracy): Doğru tahmin edilen değerlerin sayısının toplam tahmin edilen değerin sayısına bölünmesiyle elde edilen değerlendirme metriğine doğruluk (accuracy) değeri denir. Denklem 1 doğruluk değeri formülünü göstermektedir. DP Doğru Pozitif, DN Doğru Negatif, YP Yanlış Pozitif ve YN Yanlış Negatif anlamına gelmektedir [40].

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

ROC-AUC (Receiver Operating Characteristic, Area Under the Curve): Bir ROC eğrisi ikili sınıflandırma modelinin her sınıflandırma eşiğindeki performansını özetleyen bir grafiktir. Bu eğri, farklı sınıflandırma eşiklerinde gerçek pozitif oran (TPR) ve yanlış pozitif oran (FPR) olarak bilinen iki ölçümün grafiğini çizer. AUC ifadesini açılımı ise Eğri Altındaki Alan (Area Under the Curve) anlamına gelmektedir. Bu alan her zaman 0 ile 1 arasında bir değer olarak temsil edilir [41].

Denklem 2 ve denklem 3'te TPR ve FPR değerlerinin hesaplanması gösterilmektedir. Bu sonuçlar kullanılarak ROC eğrisi oluşturulur.

$$\text{TPR} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{3}$$

PR-AUC (Precision-Recall Area Under Curve): Türkçeye kesinlik ve duyarlılık olarak çevrilen Precision-Recall değerleri karmaşıklık matrisine dayanmaktadır. Precision-Recall eğrisinin oluşturduğu grafiğin altında kalan ise PR-AUC değerini ifade etmektedir. Modelin tüm veriler için başarısını tek bir değer ile ifade eder [42].

Denklem 4 ve denklem 5'te Precision ve Recall değerlerinin hesaplanması gösterilmektedir.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

Modelimizin literatürde bulunan diğer başarılı modeller ile karşılaştırma sonuçları Çizelge 3'te gösterilmektedir.

Çizelge 3: Deneysel sonuçlar (Experimental results)

Model	ACC	ROC-AUC	PR-AUC
DeepBind	0.785	0.853	0.858
DanQ	0.782	0.849	0.855
DLBSS	0.793	0.865	0.871
CRPTS	0.793	0.862	0.867
D-SSCA	0.793	0.867	0.871
DeepSTF	0.814	0.883	0.890
DeepTFBS	**0.794**	**0.891**	**0.895**

Çalışmada, transkripsiyon faktörü bağlanma bölgelerinin (TFBS) DNA sekanslarından tahmin edilmesi için Evrişimli sinir ağları, GRU ve Transformer yapılarını birleştiren hibrit bir yaklaşım benimsenmiştir. Çalışmada, modelimizin sınıflandırma performansını gösteren önemli metriklerden biri olan F1 skoru 0.813 olarak elde edilmiştir. Ayrıca modelimizin performansını diğer modellerle kıyasladığımızda başarılı sonuçlar elde ettiğimiz gözlemlenebilir. Özellikle doğruluk skoru 0.794, ROC-AUC değeri 0.891 ve PR-AUC değeri 0.895 olarak hesaplanmıştır. Sayısal sonuçlar modelimizin sınıflandırma görevlerinde güçlü performans sergilediğini ve veri setindeki pozitif ve negatif sınıfları ayırt etme yeteneğinin yüksek olduğunu göstermektedir. Bu karşılaştırma, modelimizin diğer modellere göre avantaj sağladığını ortaya koymaktadır.

Elde ettiğimiz sonuçlar, önerdiğimiz DeepTFBS yaklaşımının transkripsiyon faktörü bağlama bölgeleri tespiti için başarılı olduğunu göstermektedir. Gelecek çalışmalarda, modelin performansını artırmak için farklı veri setleri üzerinde kapsamlı değerlendirilmeler gerçekleştirilmesi ve genelleme yeteneğinin artırılması için daha karmaşık teknikler ile optimize edilmesi planlanmaktadır.

5. SONUÇLAR ve TARTIŞMA (RESULTS AND DISCUSSION)

Transkripsiyon faktörleri gen ekspresyonunun düzenlenmesinde önemli faktörlerden birisidir. Bu proteinler DNA üzerinde bağlanılacak bölgelerin tanınmasından ve bu bölgelere bağlanarak DNA'dan RNA'ya gen aktarımından sorumludur. Sonuç olarak hücrelerin adaptasyonu ve biyolojik süreçlerin düzgün şekilde ilerlemesini sağlarlar. Transkripsiyon faktörleri bahsedilen bu hayati görevlerinden dolayı tıp ve bilgisayar dünyasındaki birçok araştırmacının ilgisini

çeken bir konu olmuştur. Tıp biliminde deneysel çalışmalar üzerine odaklanılırken bilgisayar bilimlerinde daha çok hesaplamalı olarak çözümler geliştirilmiştir.

Yapılan araştırmalar, zaman serisi verilerinde Tekrarlayan sinir ağlarının (RNN) diğer modeller ile kıyaslandığında daha başarılı sonuçlar ürettiğini ortaya koymaktadır. Bu çalışmada da RNN türlerden birisi olan çift yönlü geçitli yinelenen birimler (GRU), K-mer analizi, evrişimli sinir ağları ve transformer yapıları birleştirilerek hibrit bir yöntem geliştirilmiştir. Deneylerimizi gerçekleştirirken bu alanda popüler veri kümelerinden olan 690 ChIP-seq veri kümesinden elde edilen DNA dizisi verileri kullanılmıştır. Deneysel sonuçlar, geliştirdiğimiz yöntemin literatür de geliştirilen diğer birkaç yöntemle karşılaştırıldığında daha iyi performansa sahip olduğunu göstermiştir.

Literatür araştırmaları ve gerçekleştirdiğimiz deneysel sonuçlar bu alanda hala geliştirmeye açık boşluklar olduğunu göstermektedir. Derin öğrenme modellerinde kısıtlı veri kümeleri ile çalışırken modelin performansı belirli aralıklar ile sınırlı kalabilmektedir. Kullanılan DNA şekil özellikleri de modelin performansının belirlenmesinde kritik rol oynamaktadır. Daha fazla DNA şekil özelliğinin modele entegre edilmesi, daha kapsamlı bir analiz yapılmasına, modelin doğruluk, genelleme yeteneği ve başarısının artırılmasına katkı sağlayabilir. Çalışmamızın ilerleyen süreçlerinde, veri setini genişleterek daha fazla sayıda şekil özelliği üzerinde ayrıntılı çalışmalar gerçekleştirerek kapsamlı bir model oluşturulması planlanmaktadır. Bu geliştirmelerle, modelin sadece DNA dizilimlerine değil, aynı zamanda DNA'nın yapısal karakteristiklerine dair daha derin bir analiz sunan, kapsamlı bir araç haline getirilmesi amaçlanmaktadır.

Bu çalışma, gelecekteki çalışmalarda daha karmaşık genetik düzenleme mekanizmalarının anlaşılmasına ve biyolojik süreçlerin derinlemesine analiz edilmesine katkıda bulunabilir.

ETİK STANDARTLARIN BEYANI (DECLARATION OF ETHICAL STANDARDS)

Bu makalenin yazarlar(ı) çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

YAZARLARIN KATKILARI (AUTHORS' CONTRIBUTIONS)

Ayşegül HATİPOĞLU: Problemin belirlenmesi ve araştırılması, makalenin yazımı, analizlerin yapılması ve sonuçların değerlendirilmesi.

Volkan ALTUNTAŞ: Problemin belirlenmesi, makalenin düzeltilmesi, yayına hazırlanması ve kontrolü.

ÇIKAR ÇATIŞMASI (CONFLICT OF INTEREST)

Bu çalışmada herhangi bir çıkar çatışması yoktur.

KAYNAKLAR (REFERENCES)

[1] Angermueller, C., Pärnamaa, T., Parts, L. and Stegle, O., "Deep learning for computational biology", ***Molecular systems biology***, 12(7), 878, (2016).

[2] Min, S., Lee, B. and Yoon, S., "Deep learning in bioinformatics", ***Briefings in bioinformatics***, 18(5), 851-869, (2017).

[3] Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., ... and Xie, Z., "Deep learning and its applications in Biomedicine", ***Genomics, proteomics and bioinformatics***, 16(1), 17-32, (2018)

[4] Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., ... and Weirauch, M. T., "The human transcription factors", ***Cell***, 172(4), 650-665, (2018).

[5] Tariq, S. and Amin, A., "Detection of DNA-Protein Binding Using Deep Learning", ***2023 IEEE International Conference on Emerging Trends in Engineering, Sciences and Technology (ICES&T)*** (pp. 1-4). IEEE, (2023).

[6] Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B., "Genome-wide mapping of in vivo protein-DNA interactions", ***Science***, 316(5830), 1497-1502, (2007).

[7] Zeng, Y., Gong, M., Lin, M., Gao, D. and Zhang, Y., "A review about transcription factor binding sites prediction based on deep learning", ***Ieee Access***, 8, 219256-219274, (2020).

[8] Clark, S. J., Argelaguet, R., Kapourani, C. A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., ... and Reik, W., "scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells**",** ***Nature communications***, 9(1), 781, (2018).

[9] Song, Y., Chi, A. Y. and Qu, J., "A graph theoretic approach for the feature extraction of transcription factor binding sites", ***Intelligent Computing Theories and Methodologies: 11th International Conference, ICIC 2015, Fuzhou, China, August 20-23, 2015, Proceedings, Part II 11*** (pp. 445-455). Springer International Publishing, (2015).

[10] Keilwagen, J. and Grau, J., "Varying levels of complexity in transcription factor binding motifs", ***Nucleic acids research***, 43(18), e119-e119, (2015).

[11] Alipanahi, B., Delong, A., Weirauch, M. T. and Frey, B. J., "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning", ***Nature biotechnology***, 33(8), 831-838, (2015).

[12] Hassanzadeh, H. R. and Wang, M. D., "DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins", ***2016 IEEE International conference on bioinformatics and biomedicine (BIBM)*** (pp. 178-183). IEEE, (2016).

[13] Li, Z., Liu, F., Yang, W., Peng, S. and Zhou, J., "A survey of convolutional neural networks: analysis, applications, and prospects", ***IEEE transactions on neural networks and learning systems***, 33(12), 6999-7019, (2021).

[14] Zhang, Q., Zhu, L. and Huang, D. S., "High-order convolutional neural network architecture for predicting DNA-protein binding sites", ***IEEE/ACM transactions on computational biology and bioinformatics***, 16(4), 1184-1192, (2018).

[15] Trabelsi, A., Chaabane, M. and Ben-Hur, A., "Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence

binding specificities", ***Bioinformatics***, *35*(14), i269-i277, (2019).

[16] Abdollahyan, M., Elgar, G. and Smeraldi, F., "Identifying potential regulatory elements by transcription factor binding site alignment using partial order graphs"**,** ***International Journal of Foundations of Computer Science*****,** 29(08), 1345-1354, (2018).

[17] Zhang, Q., Zhu, L., Bao, W. and Huang, D. S., "Weakly-supervised convolutional neural network architecture for predicting protein-DNA binding", ***IEEE/ACM transactions on computational biology and bioinformatics***, 17(2), 679-689, (2018).

[18] Zhang, Q., Shen, Z. and Huang, D. S., "Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network", ***Scientific reports***, 9(1), 8484, (2019).

[19] Chen, C., Hou, J., Shi, X., Yang, H., Birchler, J. A. and Cheng, J., "Interpretable attention model in transcription factor binding site prediction with deep neural networks", ***bioRxiv***, 648691, (2019).

[20] Zhou, J., Lu, Q., Xu, R., Gui, L. and Wang, H., "Prediction of TF-binding site by inclusion of higher order position dependencies", ***IEEE/ACM Transactions on Computational Biology and Bioinformatics***, 17(4), 1383-1393, (2019).

[21] Zhang, Y., Qiao, S., Ji, S. and Li, Y., "DeepSite: bidirectional LSTM and CNN models for predicting DNA–protein binding", ***International Journal of Machine Learning and Cybernetics***, 11, 841-851, (2020).

[22] Aziz, F. A. and Al-Rashid, S. Z., "Prediction of DNA binding sites bound to specific transcription factors by the SVM algorithm", ***Iraqi Journal of Science***, 5024-5036, (2022).

[23] Wang, W., Jiao, X., Sun, B., Liang, S., Wang, X. and Zhou, Y., "DeepGenBind: a novel deep learning model for predicting transcription factor binding sites", ***2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*** (pp. 3629-3635). IEEE, (2022).

[24] Song, R. and Du, X., "Predicting transcription factor binding sites by dual-stream multiple instance learning network", ***2022 IEEE International Conference on Bioinformatics and Biomedicine*** (BIBM) (pp. 3391-3398). IEEE, (2022).

[25] Cheng, J., Xu, M., Liu, Y. and Huang, W., "AttBind: Prediction of Transcription Factor Binding Sites Across Cell-types Based on Attention Mechanism", ***2022 7th International Conference on Computer and Communication Systems (ICCCS)*** (pp. 135-139). IEEE, (2022).

[26] Li, B., Wang, Z., Xiong, S. and Zhang, Y., "Densely convolutional neural network for transcription factor binding sites prediction using DNA sequence and histone modification", ***2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD)*** (pp. 179-183). IEEE, (2022).

[27] Cao, L., Liu, P., Chen, J. and Deng, L., "Prediction of transcription factor binding sites using a combined deep learning approach", ***Frontiers in Oncology*****,** 12, 893520, (2022).

[28] Yu, Y., Ding, P., Gao, H., Liu, G., Zhang, F. and Yu, B., "Cooperation of local features and global representations by a dual-branch network for transcription factor binding sites prediction", ***Briefings in Bioinformatics***, 24(2), bbad036, (2023).

[29] Rhee, H. S. and Pugh, B. F., "Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution", ***Cell***, 147(6), 1408-1419, (2011).

[30] [30] Mardis, E. R., "DNA sequencing technologies: 2006–2016", ***Nature protocols*****,** 12(2), 213-218, (2017).

[31] Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A. and Waterston, R. H., "DNA sequencing at 40: past, present and future", ***Nature*****,** 550(7676), 345-353, (2017).

[32] Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H. J., ... and Mann, R. S., "Deconvolving the recognition of DNA shape from sequence", ***Cell***, 161(2), 307-318, (2015).

[33] DNAShaper, https://www.bioconductor.org/

[34] Fletez-Brant, C., Lee, D., McCallion, A. S. and Beer, M. A., "kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets", ***Nucleic acids research*****,** 41(W1), W544-W556, (2013).

[35] https://en.wikipedia.org/wiki/K-mer

[36] Dey, R. and Salem, F. M., "Gate-variants of gated recurrent unit (GRU) neural networks", ***2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*** (pp. 1597-1600). IEEE, (2017).

[37] Bakır H. and Eker S. B., "An experimental study for evaluating the performance of CNN pre-trained models in noisy environments", ***Journal of Polytechnic***, 27(1): 355-369, (2024).

[38] Gençaslan S., Utku A. and Akcayol M.A., "Derin Öğrenme Tabanlı Video Üzerinde Olay Sınıflandırma", ***Journal of Polytechnic***, 26(3): 1155-1165, (2023).

[39] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., "Attention Is All You Need", ***Advances in Neural Information Processing Systems***, (2017).

[40] https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

[41] https://builtin.com/data-science/roc-curves-auc

[42] https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/

[43] Zeng, H., Edwards, M. D., Liu, G. and Gifford, D. K., "Convolutional neural network architectures for predicting DNA–protein binding", ***Bioinformatics***, 32(12), i121-i127, (2016).

[44] Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J. and Mann, R. S., "Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes", ***Annual review of cell and developmental biology***, 35(1), 357-379, (2019).

[45] Dittmer, J. and Nordheim, A., "Ets transcription factors and human disease", ***Biochimica et Biophysica Acta (BBA)-Reviews on Cancer***, 1377(2), F1-F11, (1998).

[46] Ji, Y., Zhou, Z., Liu, H. and Davuluri, R. V., "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome", ***Bioinformatics*****,** 37(15), 2112-2120, (2021).

[47] Wang, K., Zeng, X., Zhou, J., Liu, F., Luan, X. and Wang, X., "BERT-TFBS: a novel BERT-based model for

predicting transcription factor binding sites by transfer learning**", *Briefings in Bioinformatics,*** 25(3), bbae195, (2024).

[48] Tariq, S. and Amin, A., "DeepCTF: transcription factor binding specificity prediction using DNA sequence plus shape in an attention-based deep learning model", ***Signal, Image and Video Processing,*** 1-13, (2024).

[49] Wang, Z., Gong, M., Liu, Y., Xiong, S., Wang, M., Zhou, J. and Zhang, Y., "Towards a better understanding of TF-DNA binding prediction from genomic features", ***Computers in Biology and Medicine,*** 149, 105993, (2022).

[50] Yang, J., Ma, A., Hoppe, A. D., Wang, C., Li, Y., Zhang, C., ... and Ma, Q., "Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework", ***Nucleic acids research,*** 47(15), 7809-7824, (2019).

[51] Zhang, Q., Shen, Z. and Huang, D. S., "Predicting in-vitro transcription factor binding sites using DNA sequence+ shape", ***IEEE/ACM transactions on computational biology and bioinformatics***, 18(2), 667-676, (2019).

[52] Wang, S., Zhang, Q., Shen, Z., He, Y., Chen, Z. H., Li, J., and Huang, D. S., "Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture", *Molecular* ***Therapy-Nucleic Acids***, 24, 154-163, (2021).

[53] Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S., ... and Rohs, R., "Quantitative modeling of transcription factor binding specificities using DNA shape", ***Proceedings of the National Academy of Sciences***, 112(15), 4654-4659, (2015).

[54] Wang, X., Qiao, L., Qu, P. and Yang, Q., "TBCA: Prediction of transcription factor binding sites using a deep neural network with lightweight attention mechanism", ***IEEE Journal of Biomedical and Health Informatics***, (2024).

[55] Ding, P., Wang, Y., Zhang, X., Gao, X., Liu, G. and Yu, B., "DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape", ***Briefings in bioinformatics***, 24(4), bbad231, (2023).

[56] Wei, Y., Zhang, Q. and Liu, L., "Predicting transcription factor binding sites by a multi-modal representation learning method based on cross-attention network", ***Applied Soft Computing***, 166, 112134, (2024).

[57] Mathelier, A., Xin, B., Chiu, T. P., Yang, L., Rohs, R. and Wasserman, W. W., "DNA shape features improve transcription factor binding site predictions in vivo", ***Cell systems***, *3*(3), 278-286, (2016).

[58] Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R. and Bulyk, M. L., "Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape", ***Cell reports***, 3(4), 1093-1104, (2013).

[59] Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S. and Honig, B., "The role of DNA shape in protein–DNA recognition", ***Nature,*** 461(7268), 1248-1253, (2009).

[60] Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R. and Bulyk, M. L., "Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape", ***Cell reports***, 3(4), 1093-1104, (2013).

[61] Lazarovici, A., Zhou, T., Shafer, A., Dantas Machado, A. C., Riley, T. R., Sandstrom, R., ... and Bussemaker, H. J., "Probing DNA shape and methylation state on a genomic scale with DNase I", ***Proceedings of the National Academy of Sciences***, 110(16), 6376-6381, (2013).