BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DÜZENLEYİCİ DNA MOTİFLERİNİN TAHMİNİ

KEREM YILDIZ

YÜKSEK LİSANS TEZİ
2009

DÜZENLEYİCİ DNA MOTİFLERİNİN TAHMİNİ

PREDICTION OF DNA REGULATORY MOTIFS

KEREM YILDIZ

Başkent Üniversitesi
Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin
BİLGİSAYAR Mühendisliği Anabilim Dalı İçin Öngördüğü
YÜKSEK LİSANS TEZİ
olarak hazırlanmıştır.

2009

Fen Bilimleri Enstitüsü Müdürlüğü'ne,

Bu çalışma, jürimiz tarafından **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan	:
	Prof. Dr. Ziya AKTAŞ
Üye (Danışman)	:
	Y. Doç. Dr. Mustafa SERT
Üye	:
	Y. Doç. Dr. Mustafa DOĞAN

ONAY

Bu tez 03/02/2009 tarihinde, yukarıdaki jüri üyeleri tarafından kabul edilmiştir.

..../02/2009

Prof.Dr. Emin AKATA

FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRÜ

ÖZ

DÜZENLEYİCİ DNA MOTİFLERİNİN TAHMİNİ

Kerem YILDIZ

Başkent Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Gen ifadelerini düzenleyen mekanizmaların anlaşılması, moleküler biyolojideki önemli araştırma konularından birisidir. Bu konudaki önemli problemlerden birisi, transkripsiyon (yazım) faktörleri için Deoksiribonükleik Asit'te (DNA) bulunan bağlanma konumları gibi düzenleyici elemanları (motifleri) tanıma işlemidir. Son yıllarda bu amaç doğrultusunda birçok araç tasarlanmıştır. Önerilen bu araçlara rağmen DNA motiflerinin tahmini hala anlaşılmayan bir konu olarak kalmaya devam etmektedir. Bu çalışmada, Olasılıksal Sonek Ağacı (OSA) kullanılarak yeni bir motif tahmin yöntemi önerilmiştir. Deneysel sonuçlar başka motif bulma araçları ile karşılaştırmalı olarak değerlendirilmiştir. Elde edilen sonuçlar, önerilen yöntemin fare ve insan canlılarına ait motiflerde karşılaştırılan diğer yöntemlerden daha iyi sonuçlar verdiğini göstermiştir.

ANAHTAR SÖZCÜKLER: DNA dizilimi, motif, düzenleyici elemanlar, bağlayıcı konumlar, yazım faktörü, olasılıksal sonek ağacı

Danışman: Y.Doç.Dr. Mustafa Sert, Başkent Üniversitesi, Bilgisayar Mühendisliği Bölümü.

ABSTRACT

PREDICTION OF DNA REGULATORY MOTIFS

Kerem YILDIZ

Baskent University Institute of Science
Department of Computer Engineering

A major study in molecular biology is to understand the mechanisms that regulate the expressions of genes. An important challenge in this study is to identify regulatory elements (motifs), notably the binding sites in deocsiribonucleic acid (DNA) for transcription factors. Over the past few years, numerous tools have become available for this task. Despite the large number of these proposed tools, the prediction of DNA motifs still remains as a complex challenge. In this study, a novel motif prediction method using Probabilistic Suffix Tree (PST) is proposed. Experimental results are evaluated comparatively with other motif prediction tools. Experimental results show that, the proposed method gives a better recognition rate than the compared motif prediction tools for human and mouse genomes.

KEYWORDS: DNA sequence, motif, regulatory elements, binding sites, transcription factor, probabilistic suffix tree

Supervisor: Asst. Prof. Dr. Mustafa Sert, Baskent University, Department of Computer Engineering.

Aileme

TEŞEKKÜR

Bu çalışmanın gerçekleşmesinde katkılarından dolayı, aşağıda adı geçen kişilere içtenlikle teşekkür ederim.

Tez danışmanım Mustafa SERT hocama tez çalışmam boyunca bilgi ve deneyimini benimle paylaştığı için...

Tezime birlikte başladığım hocam Hasan OĞUL'a verdiği destek için...

Annem Nebahat YILDIZ, babam Tuncer YILDIZ ve ağabeyim Özcan YILDIZ'a tez esnasında verdikleri manevi destek için...

İş arkadaşlarım Öykü EREN ve Levent ÖZPARLAK'a tezim esnasında bana verdikleri farklı bakış açıları ve fikirleri için...

Tez jüri üyelerim Mustafa DOĞAN ve Ziya AKTAŞ hocalarıma verdikleri fikir ve önerileri için…

Başkent Üniversitesi Bilgisayar Mühendisliği Bölümündeki tüm iş arkadaşlarıma ve hocalarıma yarattıkları çalışma ortamı ve teşvikleri için...

İÇİNDEKİLER LİSTESİ

Sayfa No

ÖZ.. i
ABSTRACT....ii
İÇİNDEKİLER LİSTESİ....v
ÇİZELGELER LİSTESİ....vii
SİMGELER VE KISALTMALAR LİSTESİ....viii
1. GİRİŞ....1
1.1. Biyobilişim 1
1.2. Biyoloji Altyapısı 4
2. LİTERATÜR ARAŞTIRMASI 9
2.1. Düzenlenmiş Genlerin Başlatıcı Dizilimlerini Kullanan Yöntemler....11
2.1.1. Kelime tabanlı yöntemler....11
2.1.2. Olasılıksal yöntemler....13
2.2. Filogenetik Ayak İzine Dayalı Yöntemler....19
2.3. Düzenlenmiş Genlerin Başlatıcı Dizilimleri ve Filogenetik Ayak İzi Tabanlı Yöntemler....22
2.4. Motif Bulma Yöntemlerinde Performans Değerlendirmeleri 25
3. OLASILIKSAL SONEK AĞACI....29
3.1. OSA'nın Eğitilmesi....33
3.2. OSA Kullanarak Tahminde Bulunma....36
3.3. Zaman ve Alan Karmaşıklığı 37
4. DENEY DÜZENEĞİ VE YAPILAN ÇALIŞMALAR 38
4.1. Veri Kümesi 38
4.1.1. Veri kümesi A....38
4.1.2. Veri kümesi B....41
4.2. Uygulanan Yöntem....44
4.2.1. OSA'nın eğitilmesi....47
4.2.2. OSA kullanarak tahminde bulunma ve motifleri tespit etme....52
4.2.3. Karşılaştırma yapma 55
5. DENEYSEL SONUÇLAR....57
6. DEĞERLENDİRME VE GELECEK ÇALIŞMA PLANI 62
7. KAYNAKLAR 65
8. ÖZGEÇMİŞ 77

ŞEKİLLER LİSTESİ

Sayfa No

Şekil 1.1 – DNA Molekülü……..5

Şekil 1.2 – Yazım (Transkripsiyon)……...6

Şekil 1.3 – Yazım Faktörü Bağlanma Konumları ve Protein Sentezi……..7

Şekil 3.1 – OSA…….30

Şekil 3.2 – OSA ile SA (Sonek Ağacı)'nın Genel Yapıları…….31

Şekil 4.1 – Veri Kümesi A Biçim Örneği……40

Şekil 4.2 – Veri Kümesi B……42

Şekil 4.3 – Veri Kümesi B Biçim Örneği……42

Şekil 4.4 – DNA Motifleri Tahmin İşlemlerinin Altyapısı……..45

Şekil 4.5 – En İyi Parametrelerin Seçimi Örneği……..50

Şekil 4.6 – Tüm Alt Gruplardaki Motiflere Aynı Parametre Bulma İşleminin Uygulanması……..51

Şekil 4.7 – OSA Tahmin Sonucu Çıktısı……53

ÇİZELGELER LİSTESİ

Sayfa No

Çizelge 2.1 – Motif Tahmin Araçlarının Kronolojiksel Gösterimi……23

Çizelge 3.1 – OSA'nın Zaman ve Alan Karmaşıklığı……37

Çizelge 4.1 – Veri Kümesi A'da Bulunan Dosyaların İsimleri…….41

Çizelge 4.2 – Veri Kümesi B'de Bulunan Motif Grupları……..43

Çizelge 4.3 – Değerlendirme Yöntemi……48

Çizelge 4.4 – Parametrelerin Eşik Değerleri ve Artış-Azalış Miktarları…….49

Çizelge 5.1 – Sineklere Ait Motiflerin Tahmin Sonuçları……..59

Çizelge 5.2 – Farelere Ait Motiflerin Tahmin Sonuçları……60

Çizelge 5.3 – İnsanlara Ait Motiflerin Tahmin Sonuçları……..60

Çizelge 5.4 – Mayalara Ait Motiflerin Tahmin Sonuçları……..61

Çizelge 5.5 – Veri Kümesindeki Tüm Canlılara Ait Motiflerin Tahmin Sonuçları…….61

SİMGELER VE KISALTMALAR LİSTESİ

Kısaltma	Açıklama
A	Adenin
AlignACE	Aligns Nucleic Acid Conserved Elements
BLAST	Basic Local Alignment Search Tool
bp	base-pair
C	Cytosin
CRP	C-Reactive Protein
DNA	Deoksiribonükleik Asit
EM	Expectation Maximization
EMD	Ensemble Motif Discovery
FMGA	Finding Motifs by Genetic Algorithm
FNR	Fumarate Nitrate Reduction
G	Guanin
GEMFA	Genetic-Based EM Motif-Finding Algorithm
GST	Generalized Suffix Tree
ILP	Integer Linear Programming
MAP	Maksimum A Priori Log-Likelihood
max	Maksimum
MCMC	Markov Chain Monte Carlo
MEME	Multiple EM for Motif Elicitation
MOGAMOD	Multi-Objective Genetic Algorithm for Motif Discovery
mRNA	Motor Ribonükleik Asit
nCC	Nucleotide Level Correlation Coefficient
OSA	Olasılıksal Sonek Ağacı
PhyloCon	Phylogenetic Consensus
PPV	Positive Prediction Value
RNA	Ribonükleik Asit
SA	Sonek Ağacı
T	Timin
U	Urasil
YMF	Yeast Motif Finder

1. GİRİŞ

Genetik, biyokimya, hücre biyolojisi ve biyofizik alanlarındaki hızlı gelişme biyolojinin bir alt dalı olan moleküler biyolojinin önemini artırmıştır. Canlı organizmalarda hayati önemleri oldukça fazla olan nükleik asit, protein ve enzim yapılarının tamamen aydınlatılması bu bilim dalının ilgi alanıdır.

Nükleik asitler, bütün canlılarda ve virüslerde bulunan ve nükleotid olarak adlandırılan birimlerden oluşan polimerlerdir. En çok bilinen nükleik asitler Deoksiribonükleik Asit (DNA) ve Ribonükleik Asit (RNA)'tir. DNA, tüm organizmalar ve bazı virüslerin canlılık işlevleri ve biyolojik gelişmeleri için gerekli olan genetik talimatları taşıyan nükleik asit olarak tanımlanabilir [82].

DNA molekülleri nükleotid dizilerinin birleşmesiyle bir araya gelen genlerden oluşur. Gen, kalıtımın temel fiziksel ve işlevsel birimidir. Genlerden oluşan ifadeler, organizmalar ve bazı virüslerle ilgili birçok bilgi içermektedirler. Gen ifadelerinin anlaşılabilmesinin canlılar ve bazı virüslerle ilgili birçok bilinmeyen bilgiye ışık tutacağına inanılmaktadır [82].

Diğer yandan, gen ifadelerini düzenleyen mekanizmanın tanımlanması moleküler biyolojideki önemli problemlerden birisi haline gelmiştir. Transkripsiyon (yazım) faktörleri için DNA bağlanma konumlarında bulunan düzenleyici motiflerin anlaşılması çözülmesi istenen bu problemdeki önemli uğraşlardan birisidir.

Bu çalışmada, DNA motif tahmini için Olasılıksal Sonek Ağacı (OSA) yöntemini kullanan yeni bir yöntem önerilmiştir. Uygulanan yöntemin doğruluğunun test edilmesi için, literatürdeki on dört adet motif tanıma aracı tarafından da kullanılan veri kümeleri kullanılmış ve elde edilen sonuçlar bu yöntemlerle karşılaştırılmıştır.

1.1. Biyobilişim

Biyobilişim genel olarak biyolojik problemlerin, özellikle moleküler biyolojideki problemlerin çözümünü bilgisayar teknolojisi ve bununla ilişkili veri işleme

yöntemleri ile gerçekleştiren bilimsel disiplinin genel ismidir. Matematik, bilişim ve yaşam bilimlerini birleştirerek gen ve protein işlevlerini anlamayı hedefler. Bu bilim dalı protein ve gen dizilimleri ile ilgili bilgilerin işlenmesi için gerekli yöntemleri ve büyük miktardaki bu verileri veritabanında depolamak için gerekli modelleri araştırır. Bilişim teknikleri kullanılarak çeşitli biyoloji veri bankalarından gelen bilgi anlaşılır ve kullanılabilir bir hale getirilir. Bilgisayarların moleküler biyolojide kullanımı üç boyutlu moleküller yapıların grafik temsili, moleküler dizilimler ve üç boyutlu moleküler yapı veritabanları oluşturulmasıyla başlamıştır. Daha sonra kısa süre içerisinde bu alandaki gelişmeler hızla artmıştır. Çok yüksek miktarda veri üretilmesi, endüstri düzeyinde gen ifadesi, protein-protein ilişkisi, biyolojik olarak aktif molekül araştırmaları, bakteri, maya ,hayvan ve insan genom projeleri gibi biyolojik deneylerin doğurduğu taleple bu alana verilen önem artmıştır.

Biyobilişim araçların kullanıldığı araştırma konularından bazıları şunlardır:

- DNA dizilimleri
- Protein dizilimleri
- Protein-protein ilişkileri
- Karmaşık genetik fonksiyon ya da regülasyon faaliyetlerinin tanımlanması
- İnsan genom projesi
- Genetik faktörlerin,hastalık yatkınlığına olan etkileri
- Etkileşimli genler için bilgi ağları oluşturulması
- Heterojen biyolojik veritabanlarının entegrasyonu
- Bilgisayarlı veri analizleri
- Makromoleküler yapıların üç boyutlu dizilimleri ve üretimi
- Biyolojik bilginin paylaşımının kolaylaştırılması
- Biyolojik olayların simülasyonu
- Metabolik yol izleri ve hücre algılama modellemesi
- Protein familyalarının nasıl evrimleştiği mekanizmasının anlaşılması
- Hücre ve doku proteinlerinin haritalarının çıkarılması
- Protein yapı ve fonksiyonunun belirlenmesi
- Herhangi bir biyolojik fonksiyonu artıran veya engelleyen küçük moleküllerin
- tasarlanması

Yeni genlerin bulunması, genlerin yapı analizinden fonksiyonlarının tayini ve bir genin yapısındaki değişmenin hastalıklarla ilişkisinin araştırılmasında dizi analizleri kullanılmaktadır. Günümüzde Biyobilişim insan genomundaki genlerin dizilimlerinin ve haritalarının elde edilmesinde kullanılmakta ve yeni bilgilerin analizlerinin yapılması ile uğraşmaktadır. Yapılan bu çalışmalarla elde edilen bilgiler değişik genetik ve diğer hastalıkların daha iyi anlaşılmasına ve yeni ilaçların belirlenmesine fayda sağlayacaktır. Sonuçta Biyobilişim; ilaç tasarımı, gen terapisi, biyokimyasal işlemler gibi biyoteknoloji alanlarında uygulama bulan bir disiplin olarak kendini gösterir.

Biyolojik Veritabanları:

Araştırıcıların nükleotidlerle ilgili bilgilere ulaşabilmesi ve yeni veriler girebilmeleri için biyolojik veritabanları oluşturulmuştur. Bu veritabanlarında depolanan milyonlarca nükleotidin organizasyonu yer almaktadır. Biyobilişimde nükleotid dizi bilgilerinin organizasyonunu ve depolanmasını gerçekleştiren kuruluşlardan bazıları şunlardır:

1. GenBank (Gen Bankası- Maryland, ABD)
2. EMBL (Avrupa Moleküler Biyoloji Laboratuvarı – Hinxton, İngiltere)
3. DDBJ (DNA Japonya Veritabanı – Mishima, Japonya)
4. TRANSFAC (Yazım Faktörü Veritabanı)
5. MIPS (Münih Protein Dizilimleri Bilgi Merkezi)

Protein dizi verileri ile ilgili hizmetleri sağlayan kuruluşlar ise şunlardır:

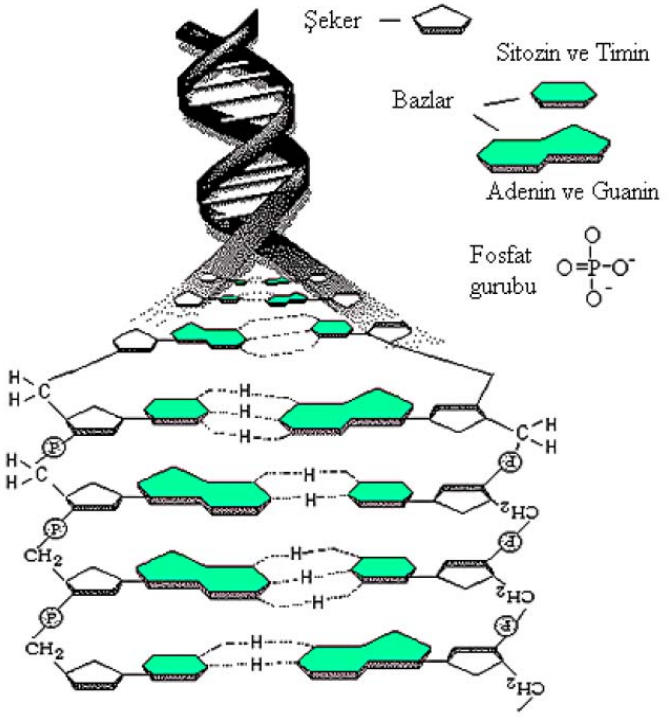
- GenBank
- EMBL
- PIR İnternational (Protein Tanımlama Kaynağı)
- Swiss-Prot.
- MIPS

MIPS veritabanı hem nükleotid hem de protein dizi verileri ile ilgili organizasyon ve depolama hizmeti sağlamaktadır.

1.2. Biyoloji Altyapısı

Nükleik asitler genetik bilginin depolanması ve ifade edilmesinden sorumlu moleküllerdir. Kimyasal olarak değerlendirildiğinde iki tür nükleik asit mevcuttur. Bunlardan biri deoksiribonükleik asit olarak nitelendirilen DNA, diğeri ise ribonükleik asit olarak nitelendirilen RNA'dır. Her iki nükleik asit de yapılarında nükleotidler bulundurmaktadırlar. Makromoleküler yapıda şeker ve fosfat birimleri fosfodiester bağı ile birbirine bağlanarak molekülün ana omurgasını oluşturur [82]. Azotlu bazlar ise iki omurgayı bir arada tutmaktadır. Nükleik asitler birden fazla yapı taşının bir araya gelmesiyle oluşmaktadır. Bu yapı taşları, şekerler, pürin ve primidin bazları, nükleozidler, nükleotidler ve polinükleotidler olarak verilir.

Ebeveynlerimizden miras aldığımız ve çocuklarımıza verdiğimiz genetik bilgiler eksik oksijenli çekirdek asidi ya da diğer bir adıyla DNA olarak verilen uzun moleküller tarafından taşınmaktadır. DNA, iki uzun tele sahiptir, bu tellerden her biri kimyasal çekirdekler, fosfat, dioksiriboz şekerler ve nükleotidlerin bir dizi şeklinde birbirine bağlanması ile oluşmaktadır [82]. DNA yapısında bulunan nükleotidler dört çeşittir, Adenin, Guanin, Sitozin, Timin sırasıyla A, G, C ve T olarak kısaltılır. Bir DNA molekülü, bir çift helis oluşturmak üzere birbirine anti paralel konumlanan iki telin birleşiminden oluşur. Bu adaptasyon katı temel eşleşme kurallarına sahiptir. Örneğin, A sadece T ile, G sadece C ile eşleşebilir. Bu nedenle, gerçekte her tel diğerinin tamamlayıcı dizisidir. DNA helisinde bir tele kalıp diğerine ise kılavuz adı verilir. DNA molekülünün örnek görünümü Şekil 1.1'de gösterilmiştir.



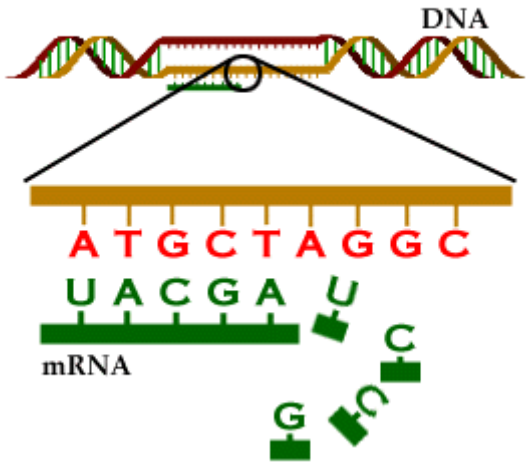
Şekil 1.1 – DNA Molekülü

Hücre içerisinde gerçekleşen birçok kimyasal reaksiyon, başlıca proteinlerin sonucudur. Herhangi bir canlı için, DNA molekülünün temel rolü, protein sentezini belirleyerek hücre içerisindeki faaliyetleri kontrol etmektir. Ancak, bir DNA doğrudan protein üretmez, bunun yerine sırayla protein sentezini kodlaması için RNA teli formunda bir kalıp üretir. Genelde, bilgi, DNA moleküllerinden proteinlere RNA molekülleri sayesinde gönderilir.

DNA molekülünden RNA molekülü sentezlenmesine yazım (transkripsiyon) denilmektedir [82]. mRNA, diğer adı ile haberci RNA, DNA'da saklı genetik bilginin, protein yapısına aktarılmasında kalıp görevi yapan aracı moleküldür. DNA molekülünden aldığı genetik şifre, sentezlenecek proteinin aminoasit sırasını tayin eder. Her mRNA molekülü, DNA üzerinde bulunan belirli bir gen dizisine karşı tamamlayıcı özelliğe sahiptir.

Yazım gerçekleşirken, DNA çift sarmalı açılarak, sarmallardan biri kalıp görevini üstlenir ve bu sarmala anti paralel olarak RNA sentezi gerçekleştirilir. Diğer bir

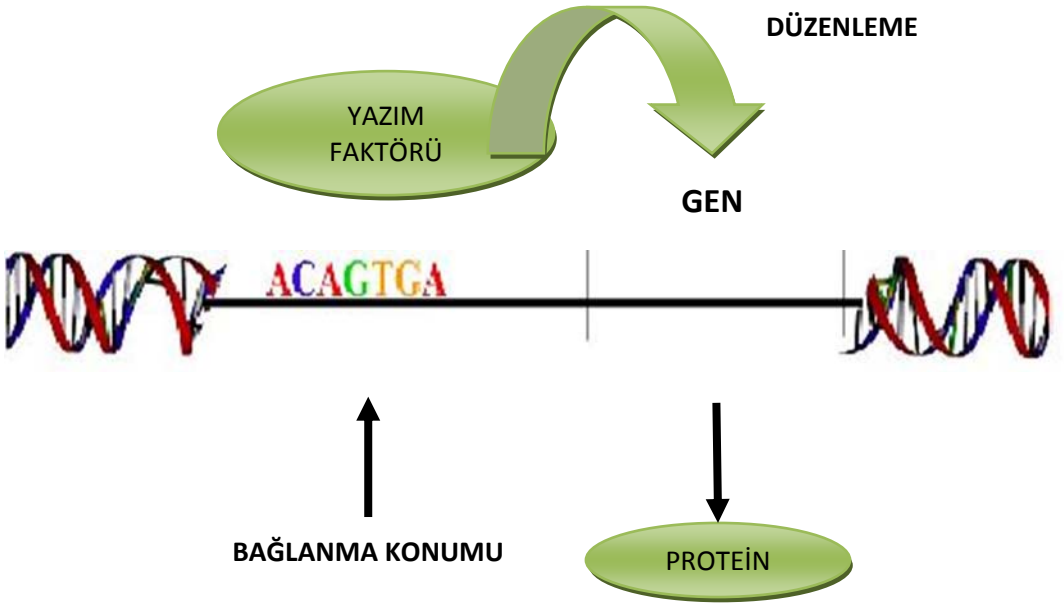
deyişle yazılılm, DNA molekülünden, RNA kalıbının üretilmesi olayıdır. Şekil 1.2'de yazım işlemi gösterilmiştir.



Şekil 1.2 – Yazım (Transkripsiyon)

Yazımdan sonra ortaya çıkan RNA, DNA çiftinin kalıp teline tamamlayıcı, kalıp olmayan diğer teline ise eş olmaktadır. DNA molekülünün kalıp olmayan teli, kod teli olarak nitelendirilir çünkü bu telin tümü daha sonra mRNA sayesinde proteinlere dönüştürülür. Ancak U ile verilen Urasil, RNA yapısında T ile yani Timin ile değiştirilir.

Gen, nükleotidlerden oluşan DNA'dan miras olarak alınmış bilginin temel birimidir ve yazım diye adlandırdığımız kopyalama işlemi için gerekli olan modelde kullanılır. Gen ifadelerindeki ana fikir, her bir genin bir protein üretmek için bilgi içeriyor olmasıdır. Gen ifadeleri çoklu protein faktörlerini DNA'da bulunan iki farklı bölge olan çoğaltıcı (enhancer) ve başlatıcı (promoter) dizilimlere bağlamakla başlar, bu protein faktörleri yazım faktörleri olarak bilinir. Yazım faktörleri, yazım mekanizmasını aktif hale getirerek ya da engelleyerek gen ifadelerini düzenlerler. Şekil 1.3'de yazılılm faktörleri bağlanma konumları ve bir genden protein üretimi işleminin düzenlenmesi gösterilmiştir.



Şekil 1.3 – Yazım Faktörü Bağlanma Konumları ve Protein Sentezi

Bir DNA motifi, yazım faktörü olarak adlandırılan düzenleyici proteinler için DNA bağlanma konumlarında olan ve biyolojiksel önem ifade eden nükleik asit örüntüleri olarak tanımlanır. Normalde örüntü oldukça kısadır. Örüntü uzunluğu 5'ten 20'ye değişen bp (base-pair) uzunluğundadır ve bu örüntünün farklı genlerin içerisinde devamlı olarak gözüktüğü kabul edilir [70]. Yani motif farklı gen dizilimleri içerisinde devamlı olarak gözüken kısa DNA parçaları olarak adlandırılır. DNA motifleri genelde proteinlerde bulunan yapısal motiflerle ilgilidir. Motifler DNA'nın her iki sarmalında da gözükebilir. Yazım faktörleri doğrudan iki sarmallı DNA'nın üzerine bağlanırlar. Dizilimlerin sıfır, bir veya çok sayıda motif kopyası olabilir. DNA motiflerinin iki özel tipi vardır:

- palindromik motifler
- spaced dyad motifler.

Palindromik bir motif, alt dizilimlerin içinde tersten okunan bir parçası yine kendisine eşit olan motiflere denir. Örnek olarak CACGTG'yi verebiliriz. Spaced dyad motifler bir boşlukçu (spacer) ile ayrılmış iki daha küçük korumalı konumlardan oluşur. Boşlukçu, motifin ortasında gözükür çünkü yazım faktörü iki küçük alt birime sahip molekül olarak bağlanır. Bu da yazım faktörünün DNA dizilimine iki ayrı temas noktası olan alt birimlerden oluştuğu anlamına gelir. Yazım

faktörünün DNA'ya bağlandığı bölümler korumalıdır fakat tipik olarak oldukça küçüktür. Uzunlukları 3-5 bp'dir. Bu iki temas noktası korumalı olmayan bir boşlukçu tarafından ayrılır. Bu boşlukçu genelde sabit uzunluktadır fakat bazen az da olsa uzunluğu değişebilir.

Bu tezin diğer bölümleri şu şekilde düzenlenmiştir: Bölüm 2'de motif tahmininde kullanılan mevcut yöntem ve araştırmalar incelenmiştir. Konu ile ilgili denenen yöntemler ve bu çalışmaların sonuçları verilmiştir. Olasılıksal Sonek Ağacı'nın yapısı, kullanım alanları ve zaman-alan karmaşıklığı bilgileri Bölüm 3'te açıklanmıştır. Bölüm 4'te düzenleyici DNA motiflerinin tahmini için önerilen yöntem ve performans ölçümleri anlatılmıştır. Önerilen yöntem ile elde edilen performans sonuçları ve bu sonuçların 14 farklı motif bulma yöntemiyle karşılaştırması Bölüm 5'te verilmiştir. Son olarak elde edilen sonuçların değerlendirmesi ve gelecek çalışma planı Bölüm 6'da verilmiştir.

2. LİTERATÜR ARAŞTIRMASI

Yazım faktörlerinin bağlandıkları bölgelerden biri olan başlatıcı bölgeden verilmiş bir DNA dizilimleri kümesinde, motif bulma problemi önceden sunulmuş motifleri tespit etme işidir. DNA motiflerini bulmak için çok fazla sayıda yöntem geliştirilmiştir. Bu yöntemlerin çoğu basit bir genomdan gelen birçok düzenli genin başlatıcı bölgesini göz önünde bulundurarak motifleri çıkarmak için tasarlanmıştır. Düzenlenmiş genler kendi düzenleyici mekanizmaları içinde bazı benzerlikleri paylaştıkça, kendi başlatıcı bölgeleri yazım faktörleri için bağlanma konumları olan bazı ortak motifleri içerebilir. Bu düzenleyici elemanları tespit etmek için düzenlenmiş genlerden oluşan bir dizi kümeyi içeren başlatıcı bölgedeki sayısal olarak önceden sunulmuş motifleri aramak mantıklı bir yaklaşımdır. Sayısal olarak sunulmuş motif, görünme sıklığı tesadüflerden çok daha fazla olan motif demektir. Sonuçta bu yöntemler başlatıcı dizilimlerin grupları içindeki önceden sunulmuş motifler için arama yapar. Ancak, bu yöntemlerin çoğu *Saccharomyces cerevisiae* olarak adlandırılan bir maya türü ve diğer düşük seviyeli organizmalarda başarılı çalışırken, yüksek seviyeli organizmalarda aynı başarıyı gösterememektedir. Çünkü bu organizmaların yapıları daha karmaşıktır. Bu zorluğun üstesinden gelmek için, son zamanlarda geliştirilmiş çapraz-tür genom kıyaslama veya filogenetik (türün evrimi ile ilgili) ayak izi tipinde veri kümesi kullanan yöntemler geliştirilmiştir [84]. Basit öncül filogenetik ayak izi fonksiyonel olmayan dizilimlerden daha düşük oranda gelişen fonksiyonel elemanlara neden olan seçici basınç olarak tanımlanır. Yani, fonksiyonel düzenleyici elemanlar veya motifler için kusursuz aday olan ortolojiksel başlatıcı bölgelerin bir kümesi içinde yer alan iyi korunmuş konum demektir. Filogenetik ayak izi tabanlı birçok motif bulma yöntemi geliştirilmiştir [6; 7; 14; 16; 17; 97]. Son zamanlarda düzenlenmiş genlerden gelen DNA dizilim verilerinin ve filogenetik ayak izini bütünleştiren yöntemler genom dizilimlerinden motif bulma işlemini önemli bir şekilde geliştirmiştir [25; 39; 57; 62; 67; 76; 77; 96]. Çalışmalar ayrıca yüksek seviyeli organizmalarda motif bulmak için önemli olan birleştirilmiş parametrelere odaklanan yöntemleri geliştirmeye başlamışlardır [30]. Stormo, DNA motif bulma için bilgisayar algoritmalarının geliştirilme ve uygulamasının tarihçesini sunmuştur [82]. Daha sonraki yıllarda DNA motif bulma yöntemlerinde kayda değer hızlı bir gelişme olmuştur ve çok sayıda DNA motif bulma yöntemi geliştirilmiş ve yayınlanmıştır.

Motif bulma yöntemleri tarafından kullanılan DNA dizilim verileri tiplerine göre üçe ayrılır:

1. Basit bir genomda bulunan düzenlenmiş genlerden gelen başlatıcı dizilimlerin kullanımı
2. Çoklu türlerden alınan basit bir genin ortolojiksel başlatıcı dizilimlerinin kullanımı (filogenetik ayak izi gibi)
3. Basit bir genomda bulunan düzenlenmiş genlerden gelen başlatıcı dizlimlerin ve buna ek olarak filogenetik ayak izinin kullanımı

Ancak ilk yapılan çalışmaların çoğu motif bulma yöntemlerini tümleşik yaklaşım olarak kullandıkları tasarımlara dayanarak iki ana grupta kategorize etmişlerdir:

1. Kelime tabanlı (katar tabanlı) yöntemler: Genelde ayrıntılı sayma işlemini yaparlar. Örnek olarak oligonükleotid frekanslarının kıyaslanması ve sayılması verilebilir.
2. Model parametrelerinin maksimum olasılık (likelihood) prensibine ya da Bayes sonuç çıkarımlarına göre değerlendirildiği olasılıksal dizilim modelleri

Kelime tabanlı sayma yöntemleri küresel optimaliteyi garanti altına alır ve kısa motifler için uygundur. Sonuçta bu yöntemler, genelde prokaryotlardaki motiflerden daha kısa olan ökaryot genomlarındaki motiflerin bulunmasında çok yararlı olurlar. Kelime tabanlı yöntemler sonek ağaçları [73] gibi en iyileştirilmiş veri yapılarının gerçekleştirildiği durumlarda çok hızlı olabilirler ve örneklerin aynı olduğu tamamıyla kısıtlandırılmış motiflerin kullanımında iyi bir seçim olurlar. Ancak, kelime tabanlı yöntemler, genelde karakteristiği kısıtlandırılmamış yani örneklerin farklı olabildiği pozisyonlara sahip yazım faktörü motifleri için sorunlu olabilirler ve sonuç sık olarak bazı kümeleme (clustering) sistemleri tarafından çeşitli işlemlerden geçirilmesi gerekir [94]. Kelime tabanlı yöntemler çok fazla gerçek olmayan motif bulmaktan dolayı da problem yaşarlar. Olasılıksal yöntem motif modeli olarak pozisyon ağırlık matrisini kullanır [11]. Pozisyon ağırlık matrisleri genelde resim yazı (pictogram) olarak görselleştirilmiştir. Resim yazının her bir pozisyonu harflerin yığıtları tarafından sunulmuştur. Harflerin boyu pozisyonun bilgi içeriği ile doğru orantılıdır [74]. Olasılıksal yöntemlerin az arama parametrelerine ihtiyaçları vardır fakat girdi verilerinin çok küçük miktarda

değişmesinden bile etkilenebilecek düzenleyici bölgelerin olasılıksal modellerine dayalıdır. Yöntemlerin çoğu yazım faktörü bağlanma konumları için gerekli olan daha uzun ve daha genel motifleri bulmak için tasarlanmış olasılıksal yöntemleri kullanırlar. Sonuçta, uzunluğu ökaryotlardaki motiflerden genelde daha uzun olan prokaryotlardaki motifleri bulmak için daha uygun yöntemlerdir. Motiflerin boyu genelde daha uzun olduğu için bu yöntemler daha iyidir. Fakat bu yöntemler, yerel arama için kullanılan Gibbs sampling (örnekleme), EM (Expectation Maximization) veya Greedy algoritmaları gibi yöntemler çıktığından beri küresel olarak en iyi sonucu bulmayı garanti etmezler.

Sonraki alt bölümlerde üç ana sınıfa ayrılmış kategorileri (DNA dizilim verilerini kullanma bakımından) temsil eden motif bulma yöntemleri incelenmiştir.

2.1. Düzenlenmiş Genlerin Başlatıcı Dizilimlerini Kullanan Yöntemler

Başlangıçta önerilen yöntemlerin çoğu sayısal olarak önceden sunulmuş motifleri tanımlayan düzenlenmiş genlerin başlatıcı dizilimlerinin bir kümesini kullanan motifleri bulmak için tasarlanmıştır. Bu motifleri bulmak için iki ana yöntem geliştirilmiştir. Sonraki alt bölümlerde bu iki ana yöntem anlatılmaktadır.

2.1.1. Kelime tabanlı yöntemler

Van Helden [91] kelime tabanlı yönteme dayalı Oligo-Analysis isimli motif bulma yöntemini geliştirmiştir. Kavramsal olarak basit olmasına rağmen, yöntem analiz edilen maya türüne (Saccharomyces cerevisiae) ait düzenleyici ailelerindeki motifleri verimli bir şekilde açığa çıkarmayı başarmıştır. Bu motifler laboratuarların deneysel analizinde daha önceden bulunmuşlardır. Buna ek olarak düzenlenmiş genlerin bulunduğu bölgelerden yeni motifler de bulunmuştur. Deneme yanılma yönteminin tersine, oligonükleotid analizi kesin ve ayrıntılıdır. Ancak, tespit etme aralığı kısa motifleri içeren görece basit örüntülerle sınırlı kalmıştır. Bu yöntemi geliştirmede kullanılan yaklaşım, düzenleyici aileleri oluşturma şeklini ve beklenen oligonükleotid frekanslarını hesaplamayı içerir. Daha sonra, van Helden [92]

spaced dyad motifleri de bulmak için bir yöntem geliştirmiştir. Çünkü spaced-dyad motiflerde kullanılan boşlukçu (spacer) başka motifler için farklı olabiliyor ve boşlukçunun uzunluğu sistematik olarak 0 ila 16 arası değişebiliyordu. Bu tip motiflerin önemi girdi verisinde bulunan korunmuş iki bölümün birleştirilmiş skorlarına ya da geri plan veri kümesindeki değerlendirilmiş dyad frekanslarına dayalı olmasıdır. Van Helden [91]'in bu yönteminin en büyük eksikliği oligonükleotidlerin hiçbir çeşitliliği olmamasıdır. Tompa [88] bu probleme DNA dizilimlerindeki kısa motifleri bulmak için tam kelime tabanlı bir yöntem öne sürerek farklı bir çözüm getirmiştir. Yöntemi ribozom bağlanma konumu problemine ayrıntılı olarak uygulanmıştır. Tompa, hem mutlak görünme sayılarını hem de geri plan dağılımını hesaba katmış, sonuçta z-skor hesabını yapmıştır. Yöntem, temeli Markov zincirine dayalı geri plan dizilimleri üzerinde verimli bir şekilde çalışmıştır.

Sinha ve Tompa [79] benzer bir yöntem kullanan YMF (Yeast Motif Finder) isimli yöntemi geliştirmişlerdir. Motif modelini mayalardaki bilinen yazım faktörü bağlanma konumları üzerinde çalışarak üretmişlerdir. Algoritma girdi olarak upstream (yazım yönünde olan) dizilimlerin bir kümesini kullanmış, motiflerdeki boşlukçusu olmayan karakterler sayılmış ve maya yukarı yön dizilimlerinin tümünden derecesi *m* olan Markov zinciri için geçiş matrisi oluşturulmuştur. Bu yöntem en büyük z-skoruna sahip motifleri başarılı bir şekilde bulmuştur. YMF maya türünün 23 adet regulonlarındaki (ortak bir düzenleyici tarafından kontrol edilen gen kümesi) düzenlenmiş aday bağlanma konumlarını tanımlamak için kullanılmıştır. 18 adet regulonda YMF'in başarılı sonuç elde ettiği fikrine varılmıştır. YMF ayrıca MIPS [60] veritabanında kayıtlı Saccharomyces cerevisiae'ye ait fonksiyonel ve mutant fenotipi kataloglarındaki gen ailelerinin motiflerini bulmak için de uygulanmış ve birçok umut vadeden yeni yazım bağlanma konumları bulunmuştur. Sinha ve Tompa [80] yapay olan mayanın başlatıcı dizilimlerini kullanarak YMF yönteminin performansını MEME (Multiple EM For Motif Elicitation) [3] ve AlignACE [72] ile karşılaştırmıştır. Karşılaştırmadan sonra YMF, maya regulonlarındaki bilinen düzenleyici elemanlarını diğer iki motif bulma aracına göre daha doğru bir şekilde tahmin etmiştir. Fakat bu çalışmadan sonra YMF'e ek olarak başka motif bulma yöntemlerinin de kullanılması gerektiği kararına varılmıştır. Çünkü yapılan gözlemde farklı motif bulma araçları başka veri kümelerinde daha iyi sonuçlar elde etmişlerdir.

Brazma [9], kelime tabanlı yönteme dayalı düzenli ifadeler tipinde örüntülerin gözükmelerini inceleyen bir motif bulma yöntemi geliştirmiştir. Yöntem, maya türüne ait genlerin yukarı yöndeki 6000 adet dizilim kümelerindeki ve mayalarda bulunan düzenlenmiş gen bölgelerindeki yukarı yöndeki örüntülerin bulunması için kullanılmıştır. En yüksek orana sahip örüntüler arasından en çok eşleşen motifler sonuç kümesi olmuştur.

Sagot [73], sonek ağacı ile dizilimlerin bir kümesinin sunumuna dayalı kelime tabanlı motif bulma yöntemini göstermiştir. Vanet [93], bakterinin bütün genomlarındaki basit motifleri aramak için sonek ağacı kullanmıştır. Marsan ve Sagot [55] bu yöntemi motiflerin kombinasyonlarını da arayabilmek için geliştirmişlerdir. Sonek ağaçları yukarı yöndeki dizilimlerin sunumunda çok sayıda muhtemel kombinasyon elde etmiştir ve gerçekleştirim hala verimliliğini korumaktadır. Motif bulma yöntemleri olan Weeder [63], MITRA (Mismatch Tree Algorithm) [19] ve GST (Generalized Suffix Tree) [61], sonek ağacı ve onun değişik biçimlerine dayanır. WINNOWER [66] ve cWINNOWER [47] motif bulma araçları, çizge kuramı yöntemine dayalı kelime tabanlı bir yöntem içermektedirler. WINNOWER yöntemi; Consensus [29], MEME [3], GibbsDNA [44] (Gibbs örnekleyicisinin DNA dizilimleri üzerinde çalışan sürümü) yöntemleri ile uzunlukları 100-1000 arası değişen motiflerde test edilerek kıyaslanmıştır ve bu üç yöntemden daha iyi sonuç üretmiştir. cWINNOWER yöntemi ise, WINNOWER yönteminin geliştirilmiş bir sürümüdür. Bu yöntem, uzunlukları 3000-12000 arası değişen motiflerde denenerek WINNOWER yöntemi ile kıyaslanmıştır ve daha iyi sonuç üretmiştir.

2.1.2. Olasılıksal yöntemler

Yazım faktörü bağlanma konumlarını matrise dayalı sunumla gerçekleştirmeyi ilk defa Hertz [28] yapmıştır. Yöntem, en yüksek bilgi içeriğine sahip konumları bulan Greedy olasılıksal dizilim modeli tabanlıdır. Bu yöntem her dizilimde bir defa sunulmuş olan ortak motifleri tanımlamak için kullanılmıştır ve yıllar geçtikçe geliştirilmiştir. En son gerçekleştirimi büyük sapma istatistiklerine göre verilmiş bilgi

içerik skorunun sayısal önemini değerlendirme işlemi yöntemini kullanarak Hertz ve Stormo [29] tarafından yapılmıştır.

Down ve Hubbard [18] NestedMICA isimli olasılıksal yönteme dayalı bir motif bulma yöntemi geliştirmiştir. Bu yöntem aynı anda birçok motifi öğrenebilen dizilim modeli tabanlıdır ve alternatif sonuç çıkarma stratejisini kullanarak tek çalıştırmada küresel olarak optimal modeli bulur. Bu yöntemin performansı yapay veri ve kas düzenleyici bölgelerin iyi karakterize edilmiş kümeleri üzerinde denenerek MEME yöntemi ile kıyaslanmıştır. Kıyaslama sonucunda NestedMICA yönteminin MEME'den daha duyarlı olduğu ve MEME'ye göre geri plan dizilimlerindeki hedef motifleri dört kat daha iyi bir şekilde açığa çıkardığı rapor edilmiştir.

Çoğu olasılıksal motif bulma yöntemleri EM, Gibbs örnekleme ve onun uzantıları gibi güçlü sayısal teknikleri kullanmaktadır.

Motif bulmak için EM yöntemi ilk olarak Lawrence ve Reilly [45] tarafından kullanılmıştır ve Hertz [28] tarafından ilk defa denenen Greedy yönteminin bir uzantısıdır. Birincil olarak protein motiflerini tespit etmek amacıyla geliştirilmiştir ancak, DNA motiflerinin tespiti için de kullanılmıştır. Konumların hiçbirinin hizalanmasına ihtiyaç yoktur ve her bir dizilimin en az bir tane ortak konum içerdiği basit model kabulüne dayanır. Konumların bulunduğu yerlerin kesin olmayışıdan dolayı, Hertz EM yöntemini geliştirmek için kayıp bilgi prensibini kullanmıştır. Bu yöntem konumların eş zamanlı tanımlanmasını ve bağlanma motiflerinin nitelendirilmesini sağlamıştır. Bailey ve Elkan [3] MEME yöntemini, EM yöntemini kullanarak hizalanmamış biyopolimer dizilimlerindeki motifleri bulmak için geliştirmişlerdir. MEME'nin amacı biyopolimer dizilim kümelerindeki yeni motifleri bulmaktır. MEME yöntemi, motif bulmak için üç yeni fikir ortaya koymuştur. Birincisi, biyopolimer dizilimlerinde gözüken alt dizilimlerin küresel olarak optimum motiflerin bulunma olasılığını artırmak için EM yönteminde başlangıç noktaları olarak kullanılmasıdır. İkincisi, her dizilim kesin bir kere gözüken silinmiş paylaşılan motifleri içermesi prensibidir. Üçüncüsü, paylaşılan motifler bulunduktan sonra onları olasılıksal olarak silen yöntemin dahil edilmesidir. Böylece farklı motifler farklı dizilimlerde gözüktüklerinde ve tek bir dizilim birçok motif içerdiğinde de birçok farklı motif dizilimleri aynı kümelerde yer alabilmişlerdir.

Olasılıksal yöntemler arasında Gibbs örnekleme yöntemi motif bulma yöntemleri için birçok kez geliştirilmiştir. Lawrence'ın [44] motif bulmak için geliştirdiği özgün Gibbs sampler'a (örnekleyici) göre yöntem DNA dizilimlerine değil protein dizilimlerine uygulanmıştır. Yöntemin özgün kabullerinden bir tanesi her dizilimdeki motifin en az bir örneği olduğudur. Yöntem bazen "konum örnekleyicisi" olarak da adlandırılmıştır. Gibbs örnekleyicisi bir MCMC (Markov Chain Monte Carlo) yöntemidir. Markov Zinciri, EM yönteminde olduğu gibi her adımdaki sonucun bir öncekine bağlı olmasından kaynaklıdır. Monte-Carlo yönteminde ise bir sonraki adımı seçme yolu belli değildir, ancak tercihen rasgele örneklemeye dayalıdır [50; 51].

Gibbs örnekleme yöntemi çeşitli dizilimlerde rasgele başlangıç pozisyonları seçilerek başlatılır. Daha sonra tahmin edici güncelleme ve devamında örnekleme adımları ile tekrarlı bir şekilde çalışır. Yani basit tekrarlayıcı prosedür temel yöntemi oluşturur. Ana fikre bakılırsa, birinci adımda ne kadar doğru bir örüntü tanımı kurulursa ikinci adımda da o örüntünün yeri o kadar doğru bir şekilde belirlenecektir.

Roth [72], Gibbs örnekleme stratejisine dayanan AlignACE (Aligns Nucleic Acid Conserved Elements) isimli motif bulma yöntemini geliştirmiştir. Bu yöntem önceden sunulmuş DNA dizilim girdi kümelerindeki motif serilerini ağılık matrisi olarak geri döndürür. Bu yöntemde bir motif, hizalanmış konum kümelerinde bilgi olarak en zengin sütunların karakteristik taban frekansı örüntüleri olarak tanımlanmıştır. Bu yöntem özgün Gibbs örnekleme yönteminden [44] farklıdır. Çünkü motif modeli değiştirilmiştir, böylece kaynak genomuna göre konumsal olmayan dizilimler için taban frekansları düzeltilmiştir. Yöntemin her bir adımında girdi dizilimlerinin iki sarmalı da aynı anda dikkate alınmıştır ve eğer konumlar farklı sarmalda iseler üst üste binen konumlara izin verilmemiştir. Tek motif arama ve tekrarlayıcı maskeleme işlemi yerine eş zamanlı birçok motif arama işlemi uygulanmıştır. Geliştirilmiş ve optimuma yakın bir örnekleme yöntemine sahiptir [72]. AlignACE örneklenen farklı motifleri değerlendirmek için MAP (maksimum a priori log-likelihood) skorlamayı kullanmıştır. MAP önceden sunulmuş motifin derecelendirilme ölçü birimi olarak da söylenebilir. Bu yöntem tarafından kullanılan

MAP skorlamanın en büyük sakıncası, yüksek olarak skorlanmış bazı motiflerin bir genomda her yerde rastlanıyor olmasıdır. Bu durum, gerçekte motif olmayan birçok nükleotid dizisinin motif olarak nitelendirilmesine yol açmaktadır. Hughes [33], AlignACE yöntemini mayalarda bulunan fonksiyonel olarak ilişkili genlerin gruplarındaki motifleri bulmak için uygulamıştır. MAP skorlamayı kullanma yerine, grup özgüllüğü (group specificity) olarak adlandırılan gelişmiş bir ölçü birimi kullanmıştır. Bu yeni ölçü birimi bütün genomların dizilimlerini hesaba katmıştır ve dikkate alınan genlerle ilişkili motifleri belirtmiştir. Bu yeni skorlama tekniği ile önceden tanımlanmış motiflere ek olarak yeni motifler de bulunabilmiştir.

Thijs [85], özgün Gibbs örnekleme yönteminin gelişmiş bir sürümünü kullanan MotifSampler isimli yeni bir motif bulma yöntemi öne sürmüştür. En büyük iki geliştirme; bir dizilimdeki motiflerin kopya sayılarını değerlendirmek için olasılıksal dağılım kullanımı ve yüksek seviyeli sıralanmış Markov zinciri geri planlı bir modelin birleştirilmesidir. Bu yöntem birçok kişi tarafından birçok veri kümesinde test edilmiştir. Bu veri kümeleri arasında G-box motifleri ve FNR (fumarate nitrate reduction) proteini tarafından düzenlenmiş bakteri genleri vardır. MotifSampler beklenen motifleri bulabilmiştir. Ayrıca yaralanmış *Arabidopsis thaliana* canlısının motifleri üzerinde de denenmiş ve bitki savunma mekanizması ile ilgili motifler bulunmuştur.

Liu [53], Gibbs örnekleme yöntemini uygulayan BioProspector isimli ve düzenlemiş genlerin başlatıcı bölgelerini kullanan bir motif bulma yöntemi geliştirmiştir. Özgün Gibbs örnekleyici yönteminden üç noktada farklıdır:

1. Kullanıcıdan ya da bir dizilim dosyasından alınmış parametreleri kapsayan 0-3 arasında değişen Markov geri plan modelini kullanır.
2. Monte Carlo yöntemi tarafından değerlendirilen motif skor dağılımı ile her bir motifin önemi belirlenir.
3. Hem spaced dyad hem de palindromik motifler için modelleme yapmaya izin verir.

Bu yöntem mayalardaki RAP1 proteininin bağlanması için gerekli motifleri, *Bacillus subtilis*'in TATA-box motifini ve *Escherichia coli*'nin bağlanma konumlarındaki CRP proteinlerini bulmada başarılı olmuştur.

Shida [75], GibbsST isimli bir motif bulma yöntemi geliştirmiştir. Gibbs örnekleme esnekliği ve çok geniş bir uygulama aralığı olduğundan en çok umut vadeden örüntü bulma yöntemlerinden biridir. Fakat yerel optima probleminde başarılı değildir [12]. Sonuçta, çözüm uzayındaki arama yöntemleri ile Gibbs örnekleme yöntemi geliştirilebilir. Örüntü bulma ve Biyobilişim alanında benzetimli tav verme (simulated annealing) yöntemi [34; 41; 43], çoğunlukla çözüm uzayında bulunan arama yöntemlerini geliştirmek için kullanılmıştır. Fakat bu yöntem kullanılarak tatmin edici sonuçlar alınamamıştır [75]. Benzetim tabanlı bu öneriler yerel optima için kullanılan Gibbs örnekleyicisinin başarısızlığını azaltmasına yardımcı olmuştur. GibbsST yönteminde benzetimli ayarlanma (simulated tempering) yöntemi kullanılarak Gibbs örnekleme yöntemi motif bulma işlemi için geliştirilmiştir. Bu yöntem yapay veri ve mayanın gerçek başlatıcı dizilimlerine uygulanmıştır ve yerel optima problemi için daha dayanıklı olduğu sonucuna varılmıştır.

Liu [49], FMGA (Finding Motifs by Genetic Algorithm) isimli genetik algoritma tabanlı bir yöntem geliştirmiştir. Bu yöntem yazım başlama konumlarının -2000 bp (base-pair) yukarı yönde (upstream) ila 1000 bp yazımın tersi yönünde (downstream) aralığında bulunan bölgelerdeki motifleri bulmak için kullanılmıştır. Genetik algoritmada mutasyon, tamamıyla korunmuş pozisyonları tersine döndürmek için pozisyon ağırlık matrislerini kullanır. Çaprazlama ise optimal çocuk örüntüleri yaratabilmek için ilk kısım ve son kısım pozisyonlarındaki genlerin yerlerinin değiştirilmesidir. Bu çaprazlama işlemi istenirse iki ortak parçalı değil daha fazla parçalara ayrılarak da yapılabilmektedir. Bu yöntem çok dengeli yerel minimumun sunumunu engellemek için pozisyon ağırlık matrisi tabanlı düzenleme kullanır. Çünkü bu çok dengeli durum yeni optimal örüntü üretmek için kullanılan işleçler için zorluk çıkarabilirler. FMGA yöntemi, Liu tarafından MEME ve Gibbs örnekleyici yöntemi ile kıyaslanmıştır. MEME'den daha kötü sonuç üretmiş ama işlem süresi daha kısa olmuştur. Gibbs örnekleyicisinden ise daha iyi sonuç üretmiş ama işlem süresi daha fazla olmuştur [44].

Liu [48], DNA ve protein dizilimlerinde motif bulabilmek için kendisini düzenleyebilen bir yapay sinir ağı geliştirmiştir. Ağ yapısı yedi katman içermektedir

ve her bir katmanda sınıflandırma işlemi yapılmaktadır. Liu tarafından yapılan benzetimler sonucunda bu yöntem MEME ve Gibbs örnekleyicisi yöntemlerinden daha iyi sonuç üretmiştir ve ayrıca uzun DNA dizilimleri için de başarılı çalışmıştır.

Kaya [38], MOGAMOD (Multi-objective genetic algorithm for motif discovery) isimli genetik algoritma tabanlı bir yöntem geliştirmiştir. Bu yöntem daha önceden sunulmuş DNA dizilimlerinin motiflerini bulmak için kullanılmıştır. Bu yöntem AlignACE [33], MEME [3] ve Weeder [63] yöntemleri ile zaman karmaşıklığı açısından kıyaslanmıştır, TRANSFAC veritabanındaki [99] insan ve maya genomlarındaki motifler üzerinde denenmiştir ve diğer üç motif bulma aracına göre zaman açısından daha hızlı performans gösterdiği Kaya tarafından belirtilmiştir. Kaya [37], bu yöntemi daha sonra yeni motif bulma işlemleri için geliştirmiştir ve aynı veritabanı verilerini kullanarak aynı motif bulma araçları ile yine zaman karmaşıklığına göre kıyaslamış ve diğer motif bulma araçlarından daha hızlı performans gösterdiğini belirtmiştir.

Bi [15], GEMFA (Genetic-based EM motif-finding algorithm) isimli genetik algoritmaya dayalı bir yöntem geliştirmiştir. Bu yöntem aynı zamanda olasılıksal yöntem çalışmalarından biri olan EM yöntemini kullanır ama, biyolojiksel dizilimlerdeki motifleri bulmak için makine öğrenme tekniklerini de yönteme ekleyerek genetik algoritma tabanlı EM motif bulma yöntemini de içeren bir yöntem içermektedir. GEMFA yöntemi; GAME [98], MEME [3] ve Bioprospector [53] isimli üç adet motif bulma aracıyla karşılaştırılmıştır. Veri kümeleri olarak üç farklı motif bağlanma konumu kullanılmıştır. Bu konumlardan ikisinde GEMFA en iyi sonucu vermiş ve diğer konumda da GAME yöntemi en iyi sonucu vermiştir. GAME yöntemi de GEMFA gibi genetik algoritma kullanmaktadır. Yani genetik algoritma kullanan yöntemler bu veri kümelerinde en iyi sonuçları vermişlerdir.

Kingsford [42], uzunluğu verilmiş olan alt dizilimleri içeren DNA motif bulma işlemi için matematiksel programlama yöntemini kullanmıştır. ILP (Integer Linear Programming), ikili alt dizilimler üzerinde mesafe metriği ayrık doğasını kullanmıştır. ILP yönteminin çözüm bulması hesaplamalı olarak zordur, dolayısıyla ILP yöntemini bulan kişiler, kısıtlamaların olduğu üssel kümeler ve verimli ayırma yöntemi ekleyerek polinom zaman çözümü ortaya koymuşlardır. Böylece ILP

yöntemi biraz kısıtlanmış ama daha verimli çalışan bir yöntem olmuştur. Bu yöntem Kingsford tarafından *E. Coli*'ye ait DNA motifleri bulunması için test edilmiş ve Gibbs örnekleyici yöntemi ile rekabet edebilecek sonuçlara ulaşmıştır.

Kaplan [35], yapı tabanlı yöntem kullanan ve DNA bağlanma verisi kullanmadan motif bulan bir yöntem geliştirmiştir. DNA dizilim verileri ve yazım faktör proteini yapı bilgisinin içeriğe özgü aminoasit nükleotid tanımlama tercihlerini çıkarabilmek için birleştirilmiştir. Bu bilgi aynı yapısal aileye ait olan yeni yazım faktörlerinden bağlanma konumları tahmin etmek için kullanılmıştır.

Birçok motif bulma yöntemi maya ve diğer organizmalarda başarılı bir şekilde çalıştıklarını ispatlamışlardır, ancak çoğu daha yüksek seviyeli organizmalarda daha düşük başarı göstermiştir [90]. Hon ve Jain [30], arama işlemini iyileştirmek için dizinleme tekniğine dayanan belirleyici bir motif bulma yöntemi geliştirmiştir. Bu yöntem insan genomuna uygulanmıştır. Hızlı arama prosedürü basit bir skorlama fonksiyonu sayesinde gerçekleştirilmiştir.

Hu [31], motif bulma yönteminin tahmin doğruluğunun artırımını geliştirmek için Ensemble yöntemini öne sürmüştür. Yöntemin birden fazla çalıştırılması sonucunda elde edilen tahminleri birleştirerek motif bulan EMD (Ensemble Motif Discovery) [32] isimli bir yeni kümeleme tabanlı Ensemble yöntemi geliştirmiştir. Hu bu yöntemi AlignACE, BioProspector, MDScan [54], MEME ve MotifSampler araçlarıyla kıyaslamışlardır. Yöntemi *E. Coli*'den türetilen veri kümeleri üzerinde test etmişlerdir. EMD yöntemi, nükleotid seviyeli tahminlerde %22,4 daha başarılı tahminde bulunmuştur. EMD yöntemi daha kısa olan girdi dizilimlerinde daha önemli başarılar elde etmiştir ancak en önemlisi daha uzun dizilimli verilerin test edildiği durumlarda en kötü ihtimalle belirli bir eşik değerinin altına düşmemiştir.

2.2. Filogenetik Ayak İzine Dayalı Yöntemler

Düzenlenmiş gen yaklaşımı üzerinde filogenetik ayak izinin en önemli avantajı düzenlenmiş genlerin tanımlanması için güvenilir bir yönteme daha az ihtiyaç duymaktadır. Filogenetik ayak izi yaklaşımının kullanımı ile sadece tek gene özgü

motifleri tanımlayabilmek mümkün olmaktadır. Fakat bu genin dikkate alınmış ortolojiksel dizilimleri arasında yeteri miktarda korunmuş olması gerekir. Çeşitli organizmalara ait genom dizilimlerinin hızlı artışıyla motif bulmak için filogenetik ayak izini kullanmak mümkün olmaya başlamıştır. Filogenetik ayak izi için kullanılan standart yöntem ortolojiksel başlatıcı dizilimler için küresel çoklu hizalama yapısı kurmak ve devamında da CLUSTAL W [86] gibi araçlar kullanarak hizalamanın içinde bulunan korunmuş bölgeleri tanımlamaktır. Ancak yapılan incelemeler sonucunda bu yöntemin filogenetik ayak izini her zaman yapamadığı sonucuna varılmıştır [7; 17; 89]. Bu sonucun en önemli nedeni, eğer türler birbiri ile yakından ilişkili ise, dizilim hizalama aşikardır fakat aydınlatıcı değildir yani fonksiyonel elemanlar etrafını sarmış olan fonksiyonel olmayan elemanlardan yeteri kadar iyi korunmuyor demektir. Diğer bir açıdan eğer türler birbirleriyle çok uzak ilişkili ise, doğru bir hizalama yapmak zor belki de imkansız olacaktır. Bu problemin üstesinden gelmek için MEME, Consensus, Gibbs örnekleyici gibi mevcut motif bulma araçları filogenetik ayak izinde kullanılmıştır. Cliften [17], AlignACE yöntemini *Saccharomyces*'in çeşitli türlerindeki kıyaslamalı DNA dizilim analizlerinde motif bulmak için kullanmış ve küresel çoklu hizalama araçlarının başarısız olduğu yerlerde bazı başarılar elde edildiğini raporlamıştır. McCue [56], Gibbs örnekleyici yöntemini protein bakterilerinin genomlarında filoenetik ayak izini kullanan motif bulma işlemi için kullanmıştır. Blanchette ve Tompa [7] tarafından bu tür motif bulma yöntemlerinin filogenetik ayak izinde problem yaratabileceği belirtilmiştir. Dolayısıyla bu tür yöntemler verilen dizilimler içinde filogenetik ilişkisi içeren işlemlerde dikkate alınmazlar çünkü bu yöntemler girdi dizilimlerini bağımsız kabul edip işlem yaparlar. Sonuçta, yakın ilişkili türlerin karışımından oluşan veri kümeleri bilinen motif seçimlerinde gereksiz yükler yaratmayacaktır. Bu yöntemler girdi dizilimlerini eşit olmayan bir şekilde tartmak için geliştirilseler bile hala soyut filogenetik ağacındaki bilgileri ele geçirememişlerdir. Bu problemin üstesinden gelmek için Blanchette ve Tompa [7] dinamik programlamaya dayalı filogenetik ayak izinden motif bulma işlemi yapan bir yöntem geliştirmişlerdir.

Berezikov [6], CONREAL isimli filogenetik ayak izi tabanlı bir motif bulma yöntemi geliştirmiştir. Bu yöntem ortolojiksel dizilimler arasında dayanak kurmak ve başlatıcı dizilim hizalamasına rehberlik etmek için gerekli motifleri kullanmıştır. CONREAL, LAGAN [10] ve AVID [8] isimli küresel hizalama programları ile

karşılaştırılmış ve bu yöntemin insan ve kemirgen gibi yakın ilişkili türler üzerinde iyi bir şekilde çalıştığı tespit edilmiştir.

Cliften [16], *Saccharomyces* deki motifleri bulabilmek için filogenetik ayak izi tabanlı bir yöntem kullanmıştır. Altı farklı *Saccharomyces* türüne ait genom dizilim üzerinde filogenetik ayak izi işlemi yaparak arama işlemi yapan CLUSTAL W isimli dizilim hizalama aracı kullanılmıştır. Bu basit hizalama tekniği ile sayısal olarak önemi olan korunmuş dizilim motifleri bulunabilmiştir.

Wang ve Stormo [97], PHYLONET isimli filogenetiksel olarak tanımlanmış korunmuş motifleri bulabilmek için birçok ilişkili genomların başlatıcı dizilimlerini analiz eden ve organizma için düzenleyici konumlar ağ yapısı tanımlayan bir yöntem geliştirmiştir. Bu yöntem her bir başlatıcı için filogenetik profil yapısı içerir ve devamında korunmuş motifleri ve onları içeren başlatıcıları tanımlamak için genomdaki bütün başlatıcıların profil uzayını verimli bir şekilde arayan BLAST (Basic local alignment search tool) tarzı yöntem kullanır. Motiflerin sayısal önemi geliştirilmiş Karlin-Altschul istatistikleri ile değerlendirilmiştir [36]. Yazarlar bu yöntemi 3524 adet mayalara ait başlatıcıların analizinde kullanmış ve 3315 adet başlatıcı ve 296 adet motif içeren iyi organize edilmiş bir düzenleyici ağ yapısı tanımlamışlardır. Bu ağ yapısı neredeyse bilinen bütün motifleri kapsar ve bilinen yazım faktörü bağlanma konumlarının %90'dan fazlasını içerir. Sonuçta bu yöntemin insan genomu gibi daha büyük genomlara uygulanabileceği kararına varmışlardır.

Carmack [14], PhyloScan isimli birçok ilişkili türden elde edilmiş ortolojiksel veri kanıtlarını bulunmuş konumlardaki kanıtlarla birleştiren bir tarama yöntemi geliştirmiştir. Ortolojiksel dizilim verileri çoklu hizalanmış, hizalanmamış veya ikisi birden olabilir. Hizalanmış verilerde, PhyloScan hizalamada payı olan türlerin filogenetik bağımlılığı için sayısal hesap yapmıştır. Hizalanmamış verilerde türler arasında filogenetik bağımsızlığın olduğu farzedilen konumlardaki kanıtlar birleştirilmiştir. Carmack bu yöntemi yedi farklı *Enterobacteriales* türlerindeki gerçek dizilim verilerinde uygulamış ve yeni yazım faktörü bağlanma konumları tanımlamıştır.

2.3. Düzenlenmiş Genlerin Başlatıcı Dizilimleri ve Filogenetik Ayak İzi Tabanlı Yöntemler

Bu iki tür yöntemi de kullanan yöntemler sonuçta tek bir olasılıksal skor üretir. Gelfand [25], düzenlenmiş genlerin başlatıcılarını ve ortolojiksel başlatıcı dizilim verilerini Archaea canlısında bulunan önceden sunulmuş motifleri bulmak için kullanmıştır. Genom dizilimlerinde ve protein benzerliği aramalarında sinyal tanımlama, tanıma profili yapılanması, aday sinyal tanımlama için Smith-Waterman algoritması kullanılmıştır. Bu çalışmada, düzenlenmiş ve ortolojiksel dizilim verileri yani iki tür veri de kullanılmıştır. Benzer şekilde iki tip veriyi McGuire [57] da kullanmıştır. Kellis [39], iki tip dizilimde bulunan karışık verilerdeki motifleri iki adımda bulan bir yöntem geliştirmiştir. İlk adımda yöntem yüksek miktarda korunmuş motifleri bulmuş ve ikinci adımda önceden sunulmuş motifleri kümelerden açığa çıkarmıştır. Prakash [67], EM yöntemli iki tip dizilimdeki karışık verileri kullanan OrthoMEME isimli bir yöntem geliştirmiştir. Bu yöntem motif boşluklarını ve motif hizalanmalarını aynı anda bulma işlemini yapmıştır. Her motifin diğer türlerde de bir kopyası olduğu farz edilmiştir. Bu yöntem iki türden ortolojiksel dizilimler üzerinde başarı göstermiştir.

PhyloCon (Phylogenetic Consensus) isimli [96], Consensus yöntemi [28] tabanlı bir yöntem Wang ve Stormo tarafından iki tür veri kümesi üzerinde işlem yapabilmesi için tasarlanmıştır. Bu yöntem ilk olarak çoklu dizilim hizalamalarında veya profillerdeki ortolojiksel dizilimlerin korunmuş bölgelerini hizalama işlemini yapmış ve devamında da sunulmuş ortolojiksel olmayan dizilimleri içeren profilleri kıyaslamıştır. Wang ve Stormo, DNA dizilim profillerini kıyaslamak için yeni bir istatistik sunmuşlar ve ortak alt profilleri aramak için Greedy yöntemini kullanmışlardır. PhyloCon, hem yapay hem de biyolojiksel verilerde başarı göstermiştir.

Sinha [77], PhyME isimli ortolojiksel dizilimlerde ve düzenlenmiş genlerde bulunan başlatıcılardaki veriler üzerinde işlem yapabilmesi için olasılıksal yönteme dayalı bir yöntem geliştirmiştir. Bu yöntemin önemli bir özelliği, ortolojiksel başlatıcılarda bulunan hem korunmuş hem de korunmamış olan bölgelerde gözüken motifleri

bulabilmiştir. İki farklı yerdeki gözükmeler içinde skorlama işlemi yapabilmiştir. Sinha bu yöntemi MEME, OrthoMEME, PhyloGibbs [76], EMnEm [62] ve GIBBS (Wadsworth Gibbs örnekleyici) [87] yöntemleriyle kıyaslamış ve PhyME yönteminin motif tespit etme başarısının diğer yöntemlerden çoğu durumda daha iyi olduğunu raporlamıştır.

Moses [62], EM yöntemli ve düzenlenmiş genlerdeki ve ortolojiksel dizilimlerdeki motifleri bulması için EMnEm isimli bir yöntem geliştirmiştir. Bu yöntem girdi dizilimlerinin tümünün hizalandığını kabul etmiş ancak, böyle bir kabul aralarında çok büyük evrimsel farklar olan örneğin insan ve fare gibi canlılar için uygun olmamıştır.

Siddhartran [76], PhyloGibbs isimli filogenetik ayak izi ve Bayes iskeletine entegre edilmiş Gibbs örnekleme motif bulma stratejilerini birleştiren bir yöntem geliştirmiştir. PhyloGibbs ortolojiksel dizilimlerin çoklu yerel dizilim hizalamasına uğradığı soyut koleksiyonlar üzerinde çalışmıştır. Beş farklı tür *Saccharomyces* canlısına ait yapay ve gerçek veriler üzerinde yapılan testler sonucunda MEME, Gibbs örnekleyici, PhyME ve EMnEm motif bulma araçlarından daha başarılı sonuç üretmiştir.

Çizelge 2.1'de şu ana kadar gerçekleştirilmiş motif bulma araçlarının geliştirilme zamanına göre kronolojik gösterimi verilmiştir.

Çizelge 2.1 – Motif Tahmin Araçlarının Kronolojiksel Gösterimi

Yöntemler	**Çalışma Prensipleri**	**Kaynak**
Galas	Sayım	Galas *et al*. [23]
Mengeritsky ve Smith	Sayım	Mengeritsky ve Smith [59]
Staden	Sayım	Staden [81]
EM	EM	Lawrence ve Reilly [45]
WordUP	Sayım	Pesole *et al*. [65]
Gibbs örnekleyici	Gibbs örnekleme	Lawrence *et al*. [44]
MACAW	Gibbs örnekleme	Liu [52]
MEME	EM	Bailey ve Elkan [3]
AlignACE	Gibbs örnekleme	Roth *et al*. [72]

Çizelge 2.1 Devam Ediyor

Yöntemler	Çalışma Prensipleri	Kaynak
Oligo-Analysis	Sayım	van Helden *et al*. [91]
Consensus	Ağırlık Matrisi	Hertz ve Stormo [29]
Dyad-Analysis	Sayım	van Helden *et al.* [92]
WINNOWER	Çizge	Pevzner ve Sze [66]
ANN-Spec	Gibbs örnekleme	Workman ve Stormo [100]
SMILE	Sonek Ağacı	Marsan ve Sagot [55]
Verbumculus	Sonek Ağacı	Apostolico *et al*. [2]
MobyDick	Sözlük	Bussemaker *et al.* [13]
YMF	Sayım	Sinha ve Tompa [79]
Bioprospector	Gibbs örnekleme	Liu *et al*. [53]
Co-Bind	Gibbs örnekleme	GuhaThakurta ve Stormo [26]
ITB	Sayım	Kielbasa *et al*. [40]
Weeder	Sayım	Pavesi *et al.* [63]
MotifSampler	Gibbs örnekleme	Thijs *et al*. [85]
MITRA	Sonek Ağacı/Çizge	Eskin ve Pevzner [19]
MDScan	Greedy yöntemi	Liu *et al.* [54]
Projection	Adresleme	Buhler ve Tompa [12]
Footprinter	Dinamik programlama	Blanchette ve Tompa [7]
MOPAC	Sayım	Ganesh *et al*. [24]
DMotif	Sayım	Sinha [78]
PhyloCon	Consensus	Wang ve Stormo [96]
LOGOS	EM	Xing *et al*. [101]
EC	Genetik algoritma	Fogel *et al*. [21]
GLAM	Gibbs örnekleme	Frith *et al*. [22]
Improbizer	EM	Ao *et al*. [1]
QuickScore	Consensus	Regnier ve Denise [69]
SeSiMCMC	Gibbs örnekleme	Favorov *et al*. [20]
PhyME	EM	Sinha *et al.* [77]
OrthoMEME	EM	Prakash *et al*. [67]
FMGA	Genetik algoritma	Liu *et al*. [49]
PHYLONET	Dizilim hizalama	Wang ve Stormo [97]
PhyloGibbs	Gibbs örnekleme	Siddharthan *et al*. [76]
GIMF	EM	Qi *et al*. [68]
WordSpy	Sözlük	Wang *et al*. [95]
MaMF	Sayım	Hon ve Jain [30]
EMD	Kümeleme-tabanlı Ensemble	Hu *et al*. [32]

Çizelge 2.1 Devam Ediyor

Yöntemler	Çalışma Prensipleri	Kaynak
GibbsST	Gibbs örnekleme	Shida [75]
MUSA	İkili kümeleme	Mendes *et al*. [58]
GAME	Genetik algoritma	Wei ve Jensen [98]
ALSE	EM	Leung ve Chin [46]
MotifSeeker	Üst Yaklaşım ve Kerteleme	Peng *et al.* [64]
PhyloScan	Tarama	Carmack *et al.* [14]
GST	Sonek Ağacı	Mohapatra, Mishra ve Padhy [61]
MOGAMOD	Genetik Algoritma	Kaya [38]
GEMFA	Genetik Algoritma	Bi [15]

2.4. Motif Bulma Yöntemlerinde Performans Değerlendirmeleri

Çok fazla sayıda motif bulma yöntemleri mevcuttur ve kullanıcılar motif bulma uğraşları için en iyi aracı seçerken yardıma ihtiyaçları olabilir. Fakat motif bulma araçları performans kıyaslaması üzerindeki çalışmaları yürütme işi kolay olmayan bir iştir. Tompa [90], değişik kaynaklardan gelen motif bulma araçlarının performans değerlendirmesinin zorluklarından bahsetmiştir. Araçlar değişik ve karmaşık motif modellerine göre tasarlanmış ve sonuçta bir motif bulma aracı tek başına bir tip veride daha iyi başarı gösterirken diğer tip verilerde daha kötü başarı gösterdiği sonucuna varılmıştır. Ayrıca biyolojinin hala anlaşılamamış olan bu düzenleyici mekanizmasından dolayı, motif modelleri üzerinde tahminde bulunan yöntemlerin her zaman doğru bir şekilde değerlendirildiği söylenemez.

Birçok yazar yöntemlerini diğer bazı mevcut yöntemlerle motif içeren biyolojiksel ve yapay verileri kullanarak test etmişlerdir. Pevzner ve Sze [66], kendi geliştirdikleri tümleşik yöntem yaklaşımlı SP-STAR isimli yöntemi olasılıksal yöntem kullanan GibbsDNA (Gibbs örnekleyici'nin DNA dizilimleri üzerinde çalışan sürümü), Consensus ve MEME ile karşılaştırmış ve SP-STAR'ın diğer üç yöntemden kısa motifler üzerinde daha iyi başarı gösterdiğini raporlamıştır. Sinha ve Tompa [80] YMF, MEME ve AlignACE yöntemlerinin motif bulma doğruluklarını kıyaslamıştır. Kıyaslama *S. cerevisiae* canlısına ait yapay ve gerçek düzenlenmiş

gen veri kümelerinde yapılmıştır. YMF'in diğer iki yöntemden daha iyi sonuç verdiği raporlanmıştır.

Tompa [90], on dört adet motif bulma yönteminin performanslarını değerlendirmiştir. Bu değerlendirmenin iki amacı vardır:

1. Mevcut motif bulma yöntemlerinin doğruluk başarıları konusunda yardım sunma.
2. Daha ileriki araçların değerlendirilmesine yardım etmek amacıyla veri kümeleri sunma.

Çoğu yazım faktörü ve onların hedef bağlanma konumları hakkında çok az şey bilindiği gerçeğine dayanarak, Tompa bu hesaplamalı araçların yeni düzenleyici eleman bulunması için tasarlandığı sonucuna varmıştır. Bu araçlar için kullanıcı önceden düzenlenmiş olduğuna inanılan genlerin düzenleyici bölgelerindeki kümeleri sağlamışlar ve bu araçlar sayısal olarak sunulmuş bu düzenleyici bölgelerdeki motifleri tahmin etmek için kullanılmıştır. Yazarlar tarafından değerlendirilen on dört motif bulma aracı: AlignACE, ANN-Spec [100], Consensus, GLAM [22], Improbizer [1], MEME, MEME3 (MEME'nin bir çeşidi), MITRA, MotifSampler, Oligo/Dyad-Analysis, QuickScore [69], SeSiMCMC [20], Weeder ve YMF olarak listelenebilir. Bu araçları test etmek için bağlanma konumları içeren veri kümeleri yaratılmıştır. Bilinen bağlanma konumları değerleri kullanılmadan her yazar kendi uzman olduğu aracı bu veri kümeleri üzerinde denemiştir. Bu uzmanların yaptığı tahminler daha sonra bilinen bağlanma konumları ile karşılaştırılmış ve tahminlerin doğruluğundan emin olmak için çeşitli istatistikler yapılmıştır.

Gerçek yazım faktörlerini ve onların bağlanma konumlarını seçmek için TRANSFAC veritabanı [99] kullanılmıştır. Her veri kümesi için 3 farklı tipte geri plan dizilimi kullanılmıştır:

1. Gerçek başlatıcı dizilim bağlanma konumları
2. Rassal olarak seçilen başlatıcı dizilimler
3. Derecesi 3 olan Markov zinciri ile oluşturulmuş dizilimler

Yapılan testler sonucunda programların doğruluk hesapları düşük çıkmıştır. Örneğin konum duyarlılığı en fazla 0.22 iken nCC (nucleotide level correlation coefficient) 0.20 çıkmıştır. Konum duyarlılığı tahmin edilmiş bilinen konumların parçalarını veren istatistiksel değer iken, nCC iki pozisyon kümesi (bilinen nükleotid pozisyonları ve tahmin edilen nükleotid pozisyonları) arasındaki farkı gösteren Pearson product-moment katsayısı istatistiksel değeridir. Fakat biyolojinin bu düzenleyici mekanizması hala anlaşılmazlığını korumaktadır. Bu nedenle araçların doğruluğunu test ederken mutlak bir standart eksikliği olmaktadır.

Kıyaslama deneyleri sonucunda Weeder aracı diğer araçlara göre çoğu alanda en iyi başarıyı göstermiştir. Weeder'ın üstünlüğüne karşın bazı durumlarda diğer araçlar da başarı göstermiştir. SeSiMCMC sinek veri kümesinde daha iyi başarı göstermiş, MEME3 ve YMF fare veri kümesinde daha iyi başarı göstermiştir. Yazarlar biyologlara tek bir motif bulma aracına güvenmek yerine birkaç tane motif bulma aracı kullanmaları tavsiyesinde bulunmuşlardır.

Hu [31] RegulonDB'den üretilmiş *E. coli* canlısına ait çok sayıda veri kümesi kullanan beş tane dizilim tabanlı motif bulma yöntemi performans karşılaştırması deneyi yapmıştır. Yazarlar tarafından değerlendirilen beş yöntem AlignACE, MEME, BioProspector, MDScan ve MotifSampler'dır. Yapılan testler sonucunda yöntemlerin performansı düşük çıkmıştır. 400 nükleotid uzunluklu dizilimler için %15-25 arası nükleotid seviyesinde ve %25-35 arası bağlanma konumu seviyelerinde doğruluk yüzdesi çıkmıştır. Fakat yöntemler zamanın %90'ında en az bir tane bağlanma konumu tahminini doğru bir şekilde yapmıştır. Hu [31] kıyaslama için Ensemble yönteminin en iyi sonucu verdiği kararına varmıştır. Ensemble yöntemi %52 ile popüler olan MEME'den bile daha iyi sonuç vererek en iyi performans gösteren yöntem olmuştur.

Bu tez çalışmasında kelime tabanlı sonek ağaçlarına olasılıksal bir yöntem getiren OSA yöntemi kullanılmıştır. OSA'nın yapısal olarak da sonek ağaçlarından bazı farklılıkları vardır. Bu yöntem DNA dizlimlerindeki motifleri tahmin etme işlemlerinde ilk defa kullanılmıştır. OSA yöntemi literatür çalışmasında anlatılan ilk yöntemlerin yaklaşımlarından biri olan maksimum olasılık prensibine dayanarak

motif tahmini yapmaktadır. DNA dizilim verileri olarak ilk sınıfa dahil olan yani düzenlenmiş genlerin başlatıcı dizilimlerini kullanan tipte veriler kullanılmıştır. Üçüncü bölümde OSA yöntemi anlatılmıştır. Dördüncü bölümde kullanılan veri kümelerinin özellikleri açıklanmıştır. Kullanılan veri kümeleri Tompa'nın değerlendirmiş olduğu on dört motif bulma aracının kullandığı ortak veri kümeleri ile aynıdır. Ayrıca, OSA yönteminin pratikteki kullanımı anlatılmış ve tahmin sonuçlarımızı belirlemeye yardımcı olan performans ölçüm kriterlerine yer verilmiştir. Beşinci bölümde Tompa'nın değerlendirmeye aldığı on dört motif bulma aracının elde ettiği sonuçlarla bu çalışmada önerilen yöntemin sonuçları belirli performans ölçüm kriterleri çerçevesinde kıyaslanmıştır. Elde edilen sonuçların değerlendirmesi ve gelecek çalışma planı altıncı bölümde verilmiştir.

3. OLASILIKSAL SONEK AĞACI

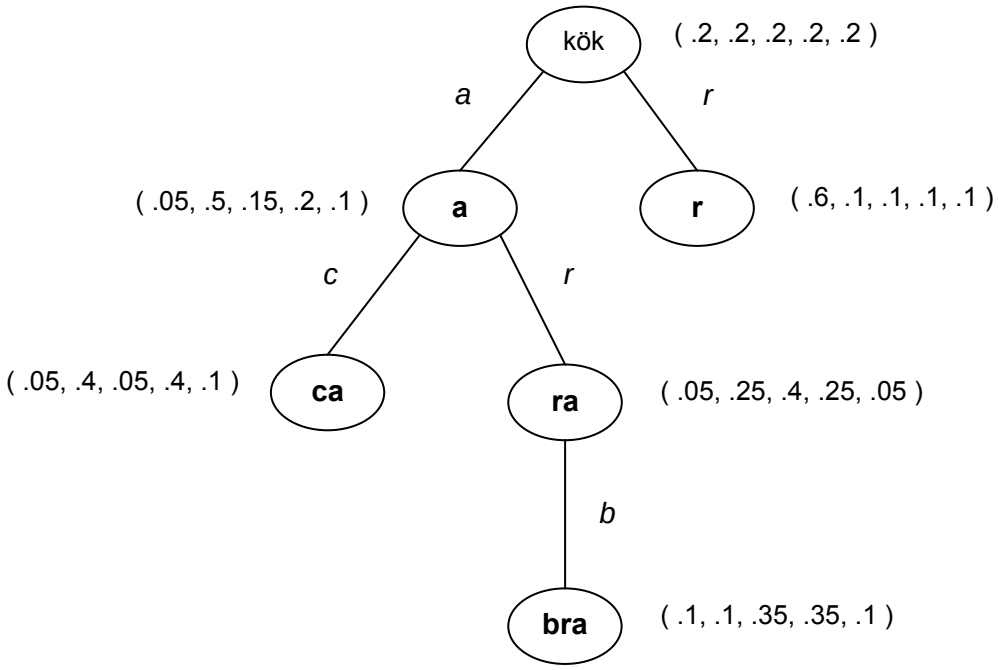
Olasılıksal sonek ağacı (OSA) ilk olarak 1996 yılında Ron tarafından öne sürülmüştür [71]. İlk çıkarılma amacı öğrenme yöntemi yaratmak olmuştur. Bu yöntem örüntü tanıma, makine öğrenme gibi alanlarda yaygın olarak kullanılmaktadır. Biyobilişim alanında da ilk olarak protein ailelerini sınıflandırmak ve hizalanmamış protein dizilimlerindeki korunmuş motifleri yani motif örüntülerini tespit amaçlı kullanılmıştır [4; 83].

Protein dizilimlerinde kullanılması için OSA değişik varyasyonlara da uğramıştır. Bu doğrultuda Bejerano ve Yona [5] biyolojik OSA fikrini öne sürmüştür. Ayrıca, ikili OSA yönteminin de başarılı sonuçlar elde ettiği rapor edilmiştir [27].

OSA, alt dizilimlerle ilişkili olasılıkları tutan ve olasılıklı model kullanan dizin yapılı sonek ağacıdır [5]. Bu yöntem çoğu biyolojiksel dizilimlerde ortak olan “kısa hafıza” ismiyle adlandırabileceğimiz bir özelliğe dayalıdır. OSA’dan önce, derecesi L (modelin hafıza uzunluğu) olan Markov zinciri ve HMM (Hidden Markov Models) yöntemleri dizilimleri modellemek için kullanılmıştır. Fakat iki yöntemin de pratik kullanımda bazı kritik kısıtlamaları vardır. Derecesi L olan Markov zinciri derece oranına göre üssel bir artış gösterir ve bu nedenle derecesi küçük olan Markov zincirleri verimli bir şekilde kullanılabilir. HMM tabanlı yöntem ise sonuçlar üzerinde öğrenme zorluğu yaşar. OSA yöntemi aynı gözleme dayalı olsa da daha büyük miktardaki kaynağı, makul miktarda hafıza kullanarak verimli bir şekilde kullanır. OSA ilk olarak 2000 yılında Benejaro ve Yona tarafından protein dizilimlerini sınıflandırmak için kullanılmıştır.

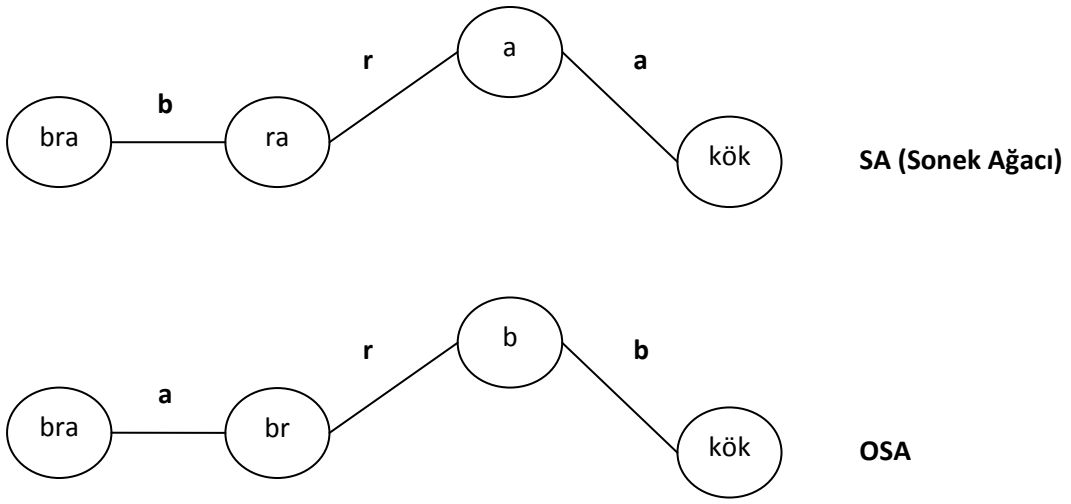
OSA uzunluğu sıfırdan farklı bir alfabe üzerinde boş olmayan bir ağaçtır ve düğüm sayıları sıfır (yapraklar için) ile alfabenin boyutu oranında değişir. Ağaçtaki her bir kenar, alfabenin bir sembolü ile adlandırılır ve hiçbir sembol ağacın dallanan kenarlarında birden fazla kez sunulamaz. Buradan her bir düğümün en fazla alfabenin boyutu kadar dallanabileceği sonucuna ulaşabiliriz. Ağacın düğümü, bu düğümden köke doğru ilerledikçe üretilen bir katar tarafından isimlendirilir. Her bir düğüm alfabe üzerinde bir olasılıksal dağılım vektörüne atanır. Olasılıksal sonek

ağacı bir sorgu katarı (yüksek olasılığa sahip parçalar) ile önemi olan örüntüleri tahmin ettiğinde, bu olasılık dağılım vektörü devreye girer.



Şekil 3.1 – OSA

Şekil 3.1'de bir OSA örneği verilmiştir. Bu örnek Σ = {a, b, c, d, r} alfabesi üzerine kurulmuştur. Kök en yukarıdaki düğümdür. Her bir düğümün yanında bulunan vektör bir sonraki sembolün olasılık dağılımıdır. Örnek olarak, alt dizilim *ra* ile ilişkili olasılık dağılımı, *a, b, c, d, r* sembolleri için sırasıyla 0.05, 0.25, 0.4, 0.25 ve 0.05'dir. Bu durumda *ra* alt diziliminden sonra en yüksek olasılıkla gözlemlenecek olan sembol 0.4 ile *c* olacaktır.



Şekil 3.2 – OSA ile SA (Sonek Ağacı)'nın Genel Yapıları

Sonek Ağacı (SA) dizilimler ile aynı yönde yapılandırılır. OSA'nın SA'dan farkı, dizilimlerin tersine göre yapılandırılmasıdır. Bu yapı en uzun sonekleri tespit etme işlemini kolaylaştırmaktadır. Bu durum şekil 3.2'de gösterilmiştir.

Şekil 3.1'deki ağaç SA olarak düşünülse, *bra* düğümünün babası *br* düğümü olurdu. Fakat olasılıksal sonek ağacına göre baba düğümü, çocuğuna göre ilk sembolü olmayan düğüm olan *ra* düğümü olur. Verilmiş bir girdi katarı için OSA'nın iskeleti (kenarlar, düğümler ve etiketler), bu katarın tersi ile ilişkili sonek ağacının alt ağacıdır.

$\sum$, ağaç için kullanılacak alfabemiz olsun. Örneğin, bu alfabe protein dizilimleri için kullanılan 20 adet aminoasit alfabesi veya DNA dizilimlerindeki 4 nükleotidi temsil eden *A, C, G, T*'den oluşabilir.

$r^1, r^2, \ldots, r^m$; $\sum$ alfabesinde bulunan *m* adet katardan oluşan örnek kümesi olsun.

i = 1 … *m* arası değişen yani toplam katar adedini gösteren değer olsun ve *i.* katarın uzunluğu l_i ile gösterilsin. Örnek verecek olursak:

$$r^i = r_1^i r_2^i \ldots r_{l_i}^i \quad , \quad r_i^j \in \sum \quad \text{olur.} \qquad (3.1)$$

İlk olarak verilmiş örnek kümedeki alt dizilim *s* nin deneysel olasılığı tanımlanmalıdır. Bu tanımlama alt dizilimin örnek kümede görülme sayısının aynı uzunluğa sahip örüntülerin toplam görünme sayılarına bölünmesi işlemi ile yapılır. Daha genel olarak ifade edecek olursak verilmiş *s* katarının uzunluğu *l* ile ifade edilirse

$$s = s_1 s_2 \dots s_l \quad \text{olur.} \tag{3.2}$$

Değişkenlerden oluşan bir küme tanımlanır. Her bir $i = 1 \dots m$ ve $j = 1 \dots l_i - (l - 1)$ için:

$$X_s^{i,j} = \begin{cases} 1 & eğer\ s_1 s_2 \dots s_l = r_j^i r_{j+1}^i \dots r_{j+(l-1)}^i\ ise \\ 0 & aksi\ takdirde \end{cases} \quad \text{olur.} \tag{3.3}$$

$X_s^{i,j}$ nin değeri 1 olur ancak ve ancak katar s, j pozisyonunda başlayan r^i nin alt dizilimidir.

Katar kümesi { r^i } de bulunan katar *s* nin bulunma sayısı 3.4 eşitliği ile ifade edilir:

$$X_s = \sum_{i,j} X_s^{i,j} \tag{3.4}$$

{ r^i } kümesindeki $|s| = l$ uzunluğundaki alt dizilimlerin toplam sayısı 3.5 eşitliği ile gösterilir:

$$N_{|s|} = \sum_{i\ öyle\ ki\ l_i \geq 1} (l_i - (l - 1)) \tag{3.5}$$

Sonuçta gözlemlenen *s* katarının deneysel olasılığı 3.6 eşitliği ile hesaplanır:

$$\tilde{P}(s) = \frac{X_s}{N_{|s|}} \tag{3.6}$$

Kesin deneysel olasılık, *s* nin örnek kümede muhtemel gözükme sayısına bağlıdır. Genelde, katar *s* nin en fazla muhtemel gözükme sayısının hesabı daha karmaşıktır. Yaptığımız tanım alt dizilimlerin üstü üste gelme durumlarını dikkate almadığından diğer alt dizilimlerden bağımsız değildir. Ayrıca eşitlik 3.7'nin sağlandığı koşulda, uzunluğu *l* olan bütün katarlar için olasılık dağılımı verilmiştir.

$$\sum_{s \in \sum^l} \tilde{P}(s) = 1 \tag{3.7}$$

Verilen alt dizilimden hemen sonra gelen bir sembolün koşullu deneysel gözlemlenme olasılığının tanımını yapabiliriz. Bu olasılık, verilen alt dizilimden sonra sembolün toplam gözükme sayısının, herhangi bir sembolden sonra aynı sembolün toplam gözükme sayısına bölümüyle hesaplanır. X_{s*}, { r^i } katar kümesindeki *s* katarının sonekli olmayan gözükme sayısı olsun. Bu durumu eşitlik 3.8 ile ifade edebiliriz:

$$X_{s*} = \sum_{\sigma' \in \sum} X_{s\sigma'} \tag{3.8}$$

s katarının σ sembolünden sonra koşullu deneysel gözlemlenme olasılığı ise eşitlik 3.9 ile ifade edilir:

$$\tilde{P}(\sigma \mid s) = \frac{X_{s\sigma}}{X_{s*}} \tag{3.9}$$

Son olarak *suf(s)* = $s_2 s_3 \ldots s_l$ ve $s^R = s_l \ldots s_2 s_1$ tanımlanır.

3.1. OSA'nın Eğitilmesi

İlk olarak, OSA'nın hafıza uzunluğunu *L* ile adlandırarak tanımlamaya başlanır. *L,* ağaçtaki muhtemel en uzun katardır. Uzunluğu 1-*L* arası değişen arama uzayındaki bütün alt dizilimlerdeki semboller, deneysel olasılık değerleri $P_{\min}$ diye adlandırdığımız bir eşik değerine veya hafıza uzunluğu maksimum *L* uzunluk

sınırına ulaşılana kadar incelenir. $P_{\min}$, L hafıza uzunluk kapasitesi içinde arama uzayının üssel olarak çok fazla artmasını engeller.

Aramanın başında elimizde sadece OSA'nın içermiş olduğu kök düğümü bulunur. Her bir incelemeye karar verdiğimiz alt dizilimler için bir alt dizilimden sonra gelen sembolün deneysel gözlemlenme olasılığı önemsiz değildir ve ayrıca alt dizilimin en sol sembolü silindiği zaman oluşan yeni alt dizilimden sonra aynı sembolün deneysel gözlemlenme olasılığı da farklı olmalıdır. Yani bir sembolün tek başına görünme olasılığı ile bir alt dizilimden sonra görünme olasılığı farklıdır. Ayrıca bir sembolün iki farklı alt dizilimden sonra görünme olasılığı da farklıdır. Bu iki durum göz önünde bulundurulduğu zaman alt dizilim ve onu oluşturan bütün gerekli düğümler OSA'ya eklenmiş olur.

Bu iki adımın yani incelenecek olan tüm düğümlerin tanımlanması ve her birinin üstünde teker teker gezme durumunun yapılma sebebi OSA'yı budama amaçlıdır. Eğer OSA üzerindeki bir yaprağın tahmin fonksiyonu kendi baba düğümündeki tahmin fonksiyonu ile aynı veya çok yakın ise o düğüm gereksiz olarak kabul edilir. Ancak, bu durum o baba düğümü çocuklarının ileriki önemli örüntü aramaları için incelenmeyeceği anlamına gelmez. Sonuçta, OSA'nın birbirine bağlı iç içe düğümlerinde benzerlikler olabilir.

En son olarak düğüm tahmin fonksiyonları uygun koşullu deneysel olasılık değeri kullanılarak OSA iskeletine eklenir ve devamında bu olasılıklar standart bir teknikle düzleştirilir. Sonuçta verilen bir alt dizilimden sonra herhangi bir sembolün gözükme durumu imkansız olmayacaktır.

Artık ağacı kurma prosedürü uygulanmaya başlanabilir. Prosedürde kullanılan harici parametreler şunlardır:

1. L, hafıza uzunluğu
2. $P_{\min}$, katarların gözükmeleri için gerekli minimum olasılık
3. r, aday yaprak düğümünün tahmin değerinin kendisi ile doğrudan bağlantılı baba düğümüne ait tahmin değerinden farkı
4. γ_{min}, düzleştirme faktörü

5. α, bir sembolün koşullu gözükmesi için önemli olan eşik değerini tanımlayan düzleştirme olasılığı ile birlikte kullanılan parametre

Ağacı belirtmek için $\bar{T}$ sembolünü kullanalım. Kontrol etmemiz gereken katar kümelerini belirtmek için $\bar{S}$ sembolünü kullanalım ve *s* düğümü ile ilişkili olasılık dağılımını belirtmek için de $\bar{\gamma}_s$ sembollerini kullanalım. Bu durumda OSA eğitim algoritması aşağıdaki gibidir:

OSA'yı eğitme algoritması ($P_{\min}$, α, γ_{min}, *r*, *L*)

1. Başlatma: $\bar{T}$, kök düğümünü içersin ve $\bar{S} \leftarrow \{\sigma \mid \sigma \in \sum ve\ \tilde{P}(\sigma) \leq P_{min}\}$ olsun
2. OSA iskeletinin kurulumu: $\bar{S} \neq \emptyset$ iken $s \in \bar{S}$ i seç ve aşağıdaki işlemleri uygula

    a. s sembolünü $\bar{S}$ kümesinden sil
    b. Eğer $\sigma \in \sum$ olan bir sembol için

$$\tilde{P}(\sigma \mid s) \geq (1+\alpha)\gamma_{min} \tag{3.10}$$

    sağlanıyorsa eşitlik 3.11'yi gerçekleştir.

$$\frac{\tilde{P}(\sigma \mid s)}{\tilde{P}(\sigma \mid suf(s))} \begin{cases} \geq r \\ veya \\ \leq 1/r \end{cases} \tag{3.11}$$

    Sonra $\bar{T}$ ağacına *s* ile ilişkili olan ve onu oluşturan düğümlerin en derinine bu *s*'nin soneği olan düğümü ekle.
    c. Eğer $|s| < L$ ise $\{\sigma' s \mid \sigma' \in \sum ve\ \tilde{P}(\sigma' s) \geq P_{min}\}$ olan katarları $\bar{S}$ kümesine ekle

3. Olasılık dağılımını düzleştirme: $\bar{T}$ ağacındaki her bir *s* yi belirten düğümler için 3.12 eşitliğini uygula:

$$\bar{\gamma}_s(\sigma) \equiv (1 - |\sum|\gamma_{min})\tilde{P}(\sigma \mid s) + \gamma_{min} \quad (3.12)$$

Öğrenme algoritmasının son adımı (3. adım) düzleştirme işlemini yapar. Bu işlem hiçbir sembolün sıfır olasılık değeri ile tahmin edilemeyeceğini garanti altına almış olur. 3.12 eşitliğinde γ_{min} değeri DNA nükleotid alfabesi için 0.25 ve üzeri, aminoasit alfabesi için de 0.05 ve üzeri değerler olamaz. Aksi durumda sonuç sıfır veya negatif değer olur.

Şekil 3.1'e geri dönecek olursak, modelin öğrendiği veri kümesinden birkaç örnekleyici gözlemlerde bulunabiliriz:

- Eğitim kümesinden anlaşıldığı üzere *a* sembolünden sonra *b* sembolü bayağı miktarda gözlemlenmiştir. $\bar{\gamma}_a(b) = 0.5$ değeri bizi böyle düşünmemize itmiştir. Ancak başka bir durum da söz konusu olabilir. Eğer *a* sembolünden önce *r* sembolü sürekli gözlemlendiyse o zaman *a* sembolünden sonra *c* sembolünün gözükme olasılığı en fazla olacaktır. Çünkü $\bar{\gamma}_{ra}(c) = 0.4$ ile en yüksektir.
- Farz edelim ki $\gamma_{min} = 0.05$ olsun ve *ca* düğümünü incelediğimizi düşünelim. Bu durumda bu düğümün olasılık vektörü sadece 3 farklı sembol olan *b, d, r* ile ifade edilecek ve olasılık miktarları sırasıyla 7 : 7 : 1 değerleriyle orantılı olacaktır. Çünkü diğer iki sembolün değeri alt limit olan 0.05 de kalmıştır.

3.2. OSA Kullanarak Tahminde Bulunma

Verilmiş bir *s* katarının tahmini OSA üzerinden harf ve harf incelenerek yapılır. Her bir harfin olasılığı *s* katarında bulunan bu harften önce gelmiş harflerden oluşan en uzun sonek bulunarak yapılır. Fakat bu harften hemen önce sonlanmış en uzun sonek aranır.

Örneğin şekil 3.1'deki ağacı kullanarak *s* = *abracadabra* katarını tahmin edelim.

$$P^T(abracadabra)$$
$$= P^T(a)P^T(b \mid \underline{a})P^T(r \mid ab)P^T(a \mid ab\underline{r})P^T(c \mid a\underline{bra})P^T(a \mid abrac) \ldots\ldots$$

$$\ldots\ldots\ P^T(a \mid abracadabr\underline{)}$$

$$= \bar{\gamma}_{root}(a)\bar{\gamma}_a(b)\bar{\gamma}_{root}(r)\bar{\gamma}_r(a)\bar{\gamma}_{bra}(c)\bar{\gamma}_{root}(a)\ldots\ldots\ldots\ \bar{\gamma}_r(a)$$

$$= \quad 0.2 \quad 0.5 \quad 0.2 \quad 0.6 \quad 0.35 \quad 0.2 \quad 0.6$$

$$= 4.032 * 10^{-6}$$

Altı çizilmiş alt dizilimler ağaçta bulunabilmiş en uzun soneklerdir. Eğer altı çizili hiçbir karakter yoksa en uzun sonek boş bir katardır. Her bir harfin olasılığı ona tekabül eden düğümle ilişkili tahmin fonksiyonu ile birlikte verilmiştir ($\bar{\gamma}_{root}()$, $\bar{\gamma}_a()$, $\bar{\gamma}_{bra}()$ gibi).

3.3. Zaman ve Alan Karmaşıklığı

Eğitim kümesinin toplam uzunluğunu *n*, OSA'nın derinliğini *L* ve dizilimlerin uzunluğunu da *m* kabul edelim. Bu durumda OSA algoritmasının zaman ve alan karmaşıklığı çizelge 3.1'deki gibi olur:

Çizelge 3.1 – OSA'nın Zaman ve Alan Karmaşıklığı

	Zaman Karmaşıklığı	**Alan Karmaşıklığı**
Öğrenme safhası	$O(Ln^2)$	$O(Ln)$
Tahmin safhası	$O(Lm)$	$O(Ln^2)$
Toplam	$O(L(n^2+m))$	$O(Ln(n+1)) \equiv O(Ln^2)$

4. DENEY DÜZENEĞİ VE YAPILAN ÇALIŞMALAR

4.1. Veri Kümesi

Bu çalışmada iki tip veri kümesi kullanılmıştır, bu veri kümeleri 4 farklı nükleotidden oluşabilen (A, C, G, T) DNA bilgilerinden oluşmaktadır. Birincisi DNA dizilimlerinden oluşan ve OSA'nın tahmin işlemi için kullanılan veri kümesidir. İkincisi ise ilk veri kümesinde bulunan yazım faktörü için bağlanma konumlarını tutan yani motiflerin bulunduğu yerleri tutan veri kümesidir. Amaç ilk veri kümesini kullanarak ikinci veri kümesindeki motifleri tahmin etmektir.

Bu veri kümeleri http://bio.cs.washington.edu/assessment/ web adresinden alınmıştır. Kullanılan veri kümeleri özgün olarak aslında TRANSFAC veritabanından [99] alınarak bu web sitesine konmuştur. Bu web sitesinde asıl amaç olarak literatür araştırmasının anlatıldığı bölümde değinilen Tompa'nın yazmış olduğu on dört adet motif bulma aracı üzerine yapılan değerlendirme bilgileri ve test işleminden elde edilen sonuçlar da bulunmaktadır. Bu motif bulma araçları bizim kullanacağımız veri kümelerinin aynısını kullanmışlardır. Böylece bizim önermiş olduğumuz OSA'yı bu araçlarla karşılaştırabilme olanağı bulmuş olduk.

Birinci ve asıl veri kümemiz olan DNA dizilimlerinden oluşan kümeyi Veri Kümesi A olarak adlandıralım ve motiflerin yerini tutan kümeyi de Veri Kümesi B olarak adlandıralım. Önceki çalışmalarda bahsedildiği gibi motif bulma araçlarının kullandığı veri kümeleri üç ayrı sınıfta toplanabiliyordu. Birincisi düzenlenmiş genlerdeki başlatıcı dizilimler, ikincisi filogenetik ayak izi diye adlandırılan ortolojiksel dizilimler ve üçüncüsü bu iki tip dizilimin de bulunduğu dizilimlerdi. Bu çalışmada kullanılan veri kümeleri ilk sınıfa dahildir.

4.1.1. Veri kümesi A

Bu veri kümesi sınıfı OSA'nın tahmini için kullanılmıştır. Bu veri kümesinde üç tip geri plan dizilimi bulunmaktadır:

1. Gerçek başlatıcı dizilimler
2. Rassal olarak seçilmiş başlatıcı dizilimler
3. Derecesi 3 olan Markov zinciri ile oluşturulmuş dizilimler

Bu 3 tip geri plan dizilimi 4 farklı canlı türüne ait DNA dizilimleri bulundurmaktadır:

1. *Drosophila melanogaster* denilen bir sinek türüne ait DNA dizilimleri
2. Fareye ait DNA dizilimleri
3. İnsana ait DNA dizilimleri
4. *Saccharomyces cerevisiae* denilen bir maya türüne ait DNA dizilimleri

Sonuçta, 6 tanesi sinek, 12 tanesi fare, 26 tanesi insan ve 8 tanesi maya olmak üzere toplam 52 adet veri kümesi kullanılmıştır.

Bu veri kümesi FASTA olarak adlandırılan bir biçimde saklanmaktadır ve şekil 4.1'de FASTA biçiminde bir veri kümesi örneği gösterilmektedir. Şekilden görüldüğü üzere biçim oldukça basittir. *n* adet dizilimden oluşan bir veri kümesinde her bir dizilim >*seq*_ ile başlar ve _ dan sonraki kısım sıralı bir şekilde 0 dan başlayarak *n-1* e kadar devam eder. Bu bilgi, ilgili veri kümesinde bulunan DNA diziliminin kaçıncı sıradaki dizilim olduğunu belirtmektedir. Dizilim sayısının belirtiminden sonra dizilimin kendisi bulunmaktadır. Daha sonra diğer dizilimler için aynı durum devam etmektedir.

```
>seq_0
ACACAGAATAAACTGAATAAACTTTTTGTCACCCTCTGCTGCCTGAGTGTGAGTTCCTGTGG
CCCCTCCCAGTGTGACTCACACAAAAATGAGAACTAGGGTAGCCACACATCCCTGAAGCCTA
TAAAGAGGTTTTCCCAGCAGATTTCTCTCGAGCTGGATACTTTGTTGTGTTTTAGAATTTTA
GCAGCAACCTTTGCTACCACTTGATAGATGTAGTACACACATCCTGTCTTACCTTTAGCTGT
GACAATGTCTCCAAATCTAGTTTTCTCGGGGAACAAAAAACGCCCTCTCGAGTGCACACCAC
TGACCGGAATGAAAGCGGACACGATTCAAGTACAGAAAGTGAGAACCCCGTTCCTAAAACGG
GATGTTGTTGTGTAATCACACTCATGACACCATTTCTGGAGCTTCAATTTTCCCAGCTTATG
CAGATGAGCAGCGTGTTCTACTCAGAGGCAGGCGGCTCCATGCAGATGAAGGGGCGTGGCCT
GAAT
>seq_1
CGCCAAGTGGCCTGCAATTACATTGCCAGCTAACACTACACTCCCATGTTACCGTCCATAAA
TTTAACTTCTGAGAATTTACAACTACTCGGAATTGCAGTCTTCTCCAAATCTAACTTCAGCT
AGTTTGGCATTGGGGTGGCCTCCATTATAAGATCAGGCAGTCAGTCACCCACTACTCTTGTA
AACTGAATCATTCGTTCTCTCTTGCTCCCTCCTCGCTAAGCCGCTTTAAATAGGATCTCAGG
ATCTCCATCGGGGTGAAAAAAAAAATCCGGGAGAAAGCACACCCATAAAGACCCCAGGAAACA
AGAAAGTCGAAGAAAGTAGGCTGATGGGGTTGGGGGTGGGGTGGGGGCAAGAGCCGGAACAT
TTTGCGCAAGACATTTCCCAAGTCCTCCGCAGATTGTGCCACAGCTTGGATAAGCTCCGCGG
CAGCCCGAGCCGCTCTAGCCCCTCATCCCGCCACTCCGAGAGCCTGGCGCGCCGCGGGGTCC
TCCA
>seq_2
CCCCGGCCGTCGCCCGCGCCCGCCCCCTCCCTGCAGCCCGCCCCCCTGGGGCCGGGTGCGCG
GCGGAGAACTGCGGTTTGCGCGCACCGGGAGCGGCAGCAGAAGTTCGAAAATCGCCGAGGGG
GGAGCCCGCGCCGCAGCTTCCTGCCCCCGGCCCAGCCCTCTGGCCCCGGCCGCCTGCAACCC
TACTTCCTCCCCGAGCCTCGGTCGCCGCCCGCTCCCTCGGACGGGGCCGCCGCGATGGGACG
CCGCGCCCCGGCCCCTGCACGCCGCTGAGCCGAGAGCCACCTAGCCAAGCCTCGCCACACAG
CGCTGCCTGATGTAATCAGGCTCCCGGAGCCTAGTCCGCGACCCCCAGGGAAAACTGGATCT
CCGAGGCTGGAGGCGCCTGGCCGGCTGGGTGGGGACCACCATGGGCAACGCGGCTGGCAGCG
CCGAACAGCCCGCGGGCCCCACCGCGTCGCCCCCGAAGCAGCCAGCCGTCCCCAAGCAACCA
ATGC
```

Şekil 4.1 – Veri Kümesi A Biçim Örneği

Her bir veri kümesinin ismi DNA dizilimlerini bulundurduğu canlıya ait bir kısaltma kelimesiyle başlar, daha sonra kaçıncı veri kümesi olduğu bilgisi ile devam eder ve son olarak hangi geri plan dizilimine ait olduğu bilgisi ile sonlanır.

Kullanılan veri kümelerinin isimleri çizelge 4.1'de verilmiştir.

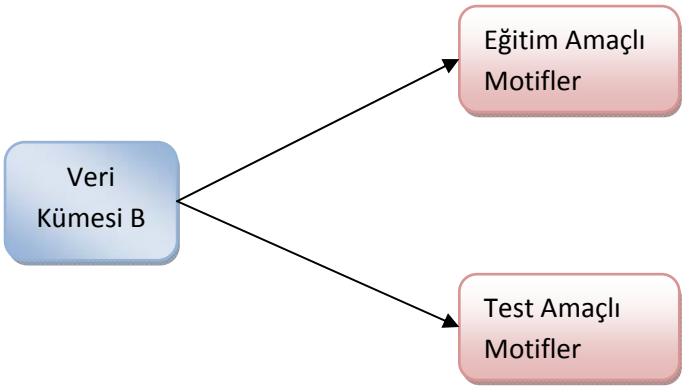
Çizelge 4.1 – Veri Kümesi A'da Bulunan Dosyaların İsimleri

Sinek	Fare	İnsan	Maya
dm01g, dm02r, dm03m, dm04g, dm05g, dm06r	mus01r, mus02r, mus03g, mus04m, mus05r, mus06g, mus07g, mus08m, mus09r, mus10g, mus11m, mus12m	hm01g, hm02r, hm03r, hm04m, hm05r, hm06g, hm07m, hm08m, hm09g, hm10m, hm11g, hm12r, hm13r, hm14r, hm15r, hm16g, hm17g, hm18m, hm19g, hm20r, hm21g, hm22m, hm23r, hm24m, hm25g, hm26m	yst01g, yst02g, yst03m, yst04r, yst05r, yst06g, yst08r, yst09g

Çizelge 4.1'de dosya isminin sonundaki bilgi olan *r* (real) gerçek başlatıcı dizilim, *g* (generic) rassal olarak seçilmiş başlatıcı dizilim ve *m* (markov) markov zinciri ile oluşturulmuş dizilim olduğu anlamına gelmektedir.

4.1.2. Veri kümesi B

Yazım faktörleri için bağlanma konumlarını tutan kümedir. Motif bilgilerini içermektedir. Bu veri kümesi şekil 4.2'de gösterildiği gibi ikiye ayrılmıştır.



Şekil 4.2 – Veri Kümesi B

Eğitim amaçlı olan motifler OSA'nın eğitimi için kullanılmıştır. Test amaçlı motifler de OSA'nın, Veri Kümesi A'dan tahmin ettiği dizilimlerdeki yüksek olasılık dağılımına sahip değerleri karşılaştırmak için kullanılmıştır.

Veri Kümesi B'nin biçim örneği şekil 4.3'de gösterilmiştir.

```
>data set
dm01
>instances
0,-869,GCCGCTGCTGCTGCATCCGTCGACGTCG,28
0,-853,CCGTCGACGTCGAC,14
0,-10,GCAGCGCTGCCGTCGCCGGCTGAGCAGC,28
2,-1243,GTCGCAGTCGCTGCC,15
2,-1217,CGCTGTTGCGGCCGACGCTGACGCA,25
2,-1133,CGCTGCCACCGCTG,14
3,-1156,CAGCGGCTGCGGA,13
>data set
dm02
>instances
0,-1566,GTTAATCCG,9
0,-1545,GAGATTATT,9
0,-1394,TATAATCGC,9
0,-1301,GGGATTAGC,9
0,-1183,GAAGGGATTAGGG,13
>data set
dm03
>instances
```

Şekil 4.3 – Veri Kümesi B Biçim Örneği

Şekil 4.3'den görüldüğü gibi her bir veri kümesinin içerdiği motif bilgileri *>dataset* kelimesi ile başlamaktadır, daha sonra hangi canlı türü ve kaçıncı veri kümesi

olduğu bilgisi verilmektedir. Devamında >*instances* kelimesi gelmekte ve bu bize motif bilgilerine geçildiğini belirtmektedir. Her bir motife ait 4 bilgi bulunmaktadır:

1. Kaçıncı dizilimde olduğu bilgisi
2. Dizilimdeki pozisyon (dizilimin sonundan başına doğru) bilgisi
3. Motifin kendisi
4. Motifin uzunluk bilgisi

Bu veri kümesindeki motif gruplarının isimleri çizelge 4.2'de verilmiştir.

Çizelge 4.2 – Veri Kümesi B'de Bulunan Motif Grupları

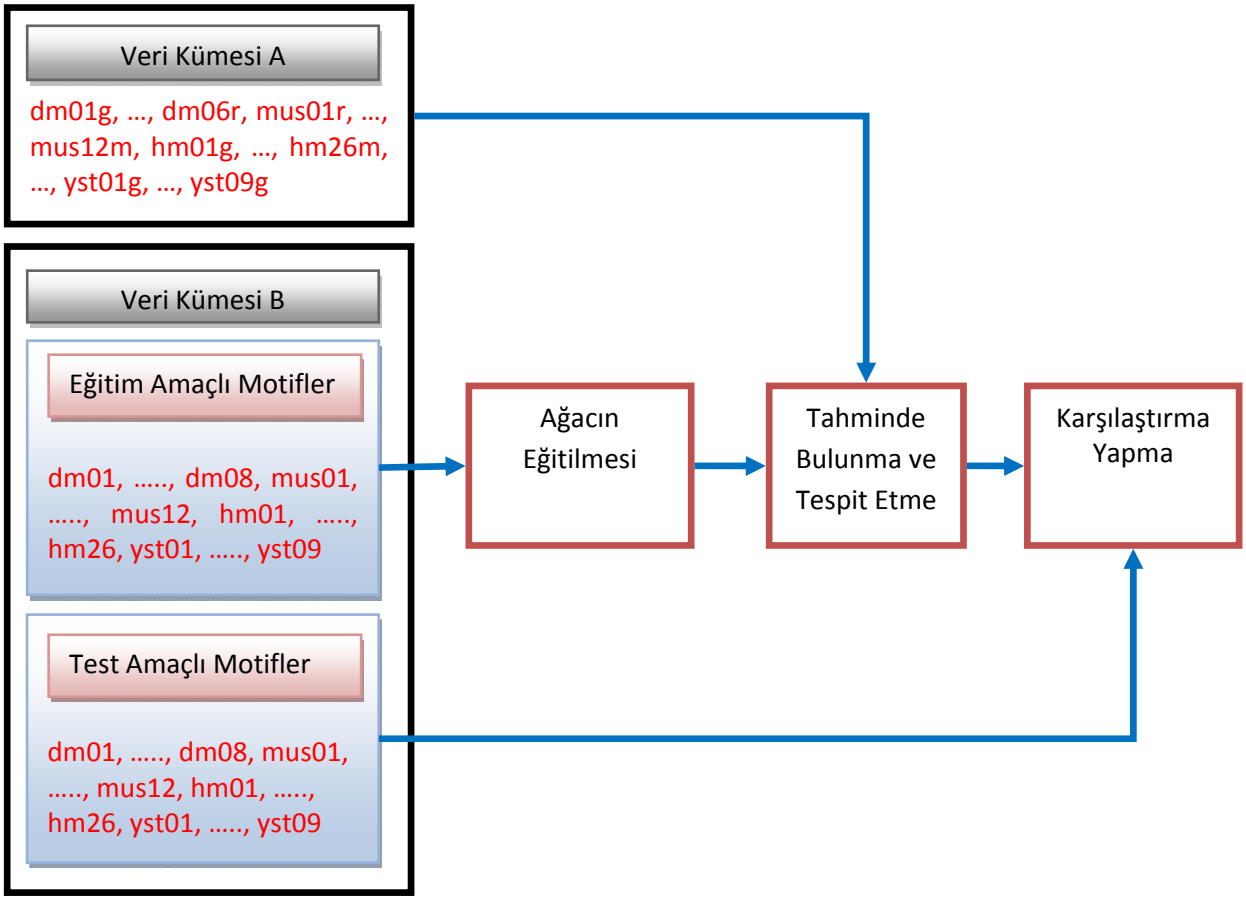
Sinek	Fare	İnsan	Maya
dm01, dm02, dm03, dm04, dm05, dm06	mus01, mus02, mus03, mus04, mus05, mus06, mus07, mus08, mus09, mus10, mus11, mus12	hm01, hm02, hm03, hm04, hm05, hm06, hm07, hm08, hm09, hm10, hm11, hm12, hm13, hm14, hm15, hm16, hm17, hm18, hm19, hm20, hm21, hm22, hm23, hm24, hm25, hm26	yst01, yst02, yst03, yst04, yst05, yst06, yst08, yst09

Bu veri kümesi 52 adet motif grubu ve toplam 517 adet motif içermektedir. Yani bu küme 4 farklı canlı türüne ait yazım faktörleri bağlanma konumlarını tutmaktadır. OSA herbir grup için eğitilerek tahminler elde edilmiştir. Veri kümesinin kullanım biçimi şu şekildedir: Sineklere ait 23 tane motif eğitim, 28 tane motif test için kullanılmıştır. Farelere ait 44 tane motif eğitim, 49 tane motif test için kullanılmıştır. İnsanlara ait 141 tane motif eğitim, 157 tane motif test için kullanılmıştır. Mayalara ait 35 tane motif eğitim, 40 tane motif test için kullanılmıştır. Toplamda 243 tane motif eğitim, 274 tane motif test için kullanılmıştır.

Herhangi bir motif grubu içindeki toplam motif sayısı çift sayı ise eğitim ve test için kullanılacak motif sayıları eşit olarak ikiye bölünmüştür. Fakat toplam motif sayısı tek sayı ise test edilecek motif sayısı eğitim için kullanılacak motif sayısından bir fazla olacak şekilde ikiye bölünmüştür. Böylece test amaçlı kullanılacak motif sayısı daha fazla olmuştur.

4.2. Uygulanan Yöntem

Linux işletim sistemi ortamında C programlama dili kullanılarak OSA veri yapısı oluşturulmuş ve bu veri yapısı kullanılarak Veri Kümesi B sınıfındaki eğitim amaçlı ayrılmış motifler ile bu ağaç yapısı eğitilmiştir. Sonra OSA, Veri Kümesi A sınıfındaki DNA dizilimlerini kullanarak Veri Kümesi B sınıfındaki test amaçlı ayrılmış motifleri tahmin etmeye çalışmıştır. Son olarak tahmin sonrasında tespit edilmiş motifler test amaçlı ayrılmış motifler ile karşılaştırılmıştır. Şekil 4.4'de DNA motiflerinin tahmini için yapılacak işlemler adım adım gösterilmiştir. Ana program akışı Sözde Kod 1'de verilmiştir.



Şekil 4.4 – DNA Motifleri Tahmin İşlemlerinin Altyapısı

Ağacı kullanarak motif tahmini maksimum olasılık prensibine dayanmaktadır. Ağacın eğitimi ve tahmini 4 canlı grubu (*dm*, *mus*, *hm* ve *yst*) içinden her seferinde sadece bir motif grubu için (*dm01*, *hm13*, *yst02* gibi) yapılmıştır. Daha sonra yeni bir motif grubu için işleme başlanmadan önce ağaç temizlenmiştir. Bu işlem Veri Kümesi B'de bulunan motifler bitene kadar devam etmiştir. Son olarak tahmin sonucunda tespit edilen motifler, test amaçlı ayrılmış motiflerin gerçek konumları ile karşılaştırılmış ve başarı oranı hesaplanmıştır (Sözde Kod 2).

Sözde Kod 1: Ana program

```
1: s ← tüm motifler, len_s ← s'nin uzunluğu
2: while i ≤ len_s do
3:     m ← motif grubu
4:     len_m ← m'nin uzunluğu, i ← i + len_m
5:     len_eğitim ←  eğitim amaçlı motif sayısı
6:     for i ← 1, len_eğitim + 1 do
7:         harici parametrelerin değerlerini ata
8:         harici parametrelerin artış-azalış değerlerini ata
9:         for j ← 1, 5 do
10:            f ← girdi dosyasını hazırla
11:            T ← ağacı_eğitme(p_min, alpha, gamma_min, L_max, p_ratio)
12:            tahmin_yapma(T, f)
13:            for k ← 5, 75 do
14:                karşılaştırma_yapma(k)
15:            end for
16:            hesaplanan nCC skor değerine göre parametrelerin değerlerini güncelle
17:        end for
18:    end for
19:end while
```

OSA veri yapısının eğitilmesi ve tahmin için kullanılan algoritmalar 3. Bölümde detaylı olarak incelenmiştir.

Sözde Kod 2: Eşleştirme (karşılaştırma) algoritması

```
1: test ← test edilecek motifler, tahmin ← tahmin edilmiş motifler
2: len_test ← test edilecek motiflerin uzunluğu
3: for i ← 1, len_test do
4:     nTP'ye test edilen ile tahmin edilmiş motifin ortak nükleotit pozisyon sayılarını ekle
5:     nFN'e geriye kalan tahmin edilmiş motifin nükleotit pozisyon sayısını ekle
6:     nFP'ye geriye kalan test edilen motifin nükleotit pozisyon sayısını ekle
7:     nTN'e dizilimdeki diğer nükleotit pozisyon sayılarını ekle
8:     if(tahmin(i), test(i) tarafından en az 0.25 oranında üstüste çakışıyorsa)
9:         sTP ← sTP + 1
```

```
10:     else
11:             sFN ← sFN + 1
12:     end if
13:     if(test(i), tahmin(i) tarafından en az 0.25 oranında üsütüste çakışmıyorsa)
14:             sFP ← sFP + 1
15:     end if
16:end for
```

4.2.1. OSA'nın eğitilmesi

OSA'nın motif tahmini yapabilmesi için ilk olarak benzer motiflerle eğitilmesi gerekir. OSA, Veri Kümesi B isminde sınıflandırdığımız ve içinde eğitim amaçlı olarak ayırdığımız motiflerle eğitilmiştir. Daha önceden de belirtildiği gibi ağacın her eğitimi her bir motif grubu için baştan yapılmıştır.

Doğru parametrelerin ve motif uzunluğunun seçilmesi

OSA'nın genel tanımını yaptığımız bölümde ağacın eğitilebilmesi için beş tane harici parametrenin belirlenmesi gerektiği belirtilmişti. Bu nedenle ağacı en doğru şekilde eğitebilmek için en doğru parametre değerlerinin seçilmesi gerekmektedir.

Bunlar daha önceden de belirtildiği gibi, $P_{\min}$, α, γ_{min}, r ve L parametreleridir.

En iyi parametreyi bulabilmek için eğitim amaçlı kullanılan motifler arasında kısa tahminlerin yapılması gerekmektedir. Elimizde bir motif grubu içinde *n* adet eğitim amaçlı ayrılmış motif olsun. Kısa tahmin için OSA *n-1* tane motif ile eğitilir geriye kalan 1 tane eğitim amaçlı motif bu ağaç tarafından tahmin edilir. Ancak ilgili veri kümesinin bütün dizilimlerine bakılmaz, çünkü tahmin edilecek motifin kaçıncı dizilimde olduğu bellidir. Bu nedenle sadece ilgili dizilim içinde tahmin yapılır ve bu işleme "kısa tahmin" adı verilir. Parametrenin nasıl sonuç verdiği 4.1 eşitliğindeki *nCC* (nucleotide level correlation coefficient) ile hesaplanmıştır.

nCC, bilinen motife ait nükleotid pozisyonları ile tahmin edilen motife ait nükleotid pozisyonları arasındaki farkı gösterir ve -1 ile +1 arasında değerler alır. -1 e doğru gidildikçe tahmin kötüleşir, +1 e doğru gidildikçe tahmin iyileşir.

Eşitlikte yer alan değişkenlerin değerlendirme yöntemindeki kullanımları çizelge 4.3'de verilmiştir.

1. nTP hem gerçek konumda hem de tahmin eden sistemin konumunda bulunan nükleotid pozisyonları,
2. nFN gerçek konumda bulunan ama tahmin eden sistemin konumunda bulunmayan nükleotid pozisyonları,
3. nFP gerçek konumda bulunmayan ama tahmin eden sistemin konumunda bulunan nükleotid pozisyonları,
4. nTN ne gerçek konumda ne de tahmin eden sistemin konumunda bulunan nükleotid pozisyonları.

Çizelge 4.3 – Değerlendirme Yöntemi

Sistem / Gerçek	**Motif (+)**	**Motif Olmayan (-)**
Motif (+)	Doğru-Pozitif	Yanlış-Negatif
Motif Olmayan (-)	Yanlış-Pozitif	Doğru-Negatif

$$nCC = \frac{nTP.nTN - nFN.nFP}{\sqrt{(nTP+nFN)(nTN+nFP)(nTP+nFP)(nTN+nFN)}} \qquad (4.1)$$

Parametrelerden bir tanesi olan *L* değeri tam sayı olmak zorundadır, çünkü ağacın hafıza uzunluğunu tutmaktadır. Bununla birlikte diğer 4 parametre ondalıklı değer olabilmektedir.

Muhtemel olasılıkları düşünecek olursak, sonsuz tane değişik parametre kombinasyonu olabilir. Bu nedenle her bir parametre için alt ve üst eşik değerleri belirlenmelidir. Fakat her ne kadar alt ve üst eşik değeri verilse bile ondalık değerler için yine sonsuz sayıda kombinasyon olabilir. Bu nedenle her bir parametre için artış veya azalış miktarı da belirlenmelidir. Çizelge 4.4'de kullanılan eşik değerleri ile artış-azalış miktarları gösterilmiştir.

Çizelge 4.4 – Parametrelerin Eşik Değerleri ve Artış-Azalış Miktarları

Parametreler	Alt Eşik Değeri	Üst Eşik Değeri	Artış-Azalış Miktarı
P_{min}	0.0001	0.1	0.0001
α	0.0	5.0	0.1
γ_{min}	0.001	0.2	0.001
r	0.1	4.0	0.1
L	5	25	1

Eşik değerleri ve artış-azalış miktarları da belirlendikten sonra en uygun değerler tek seferde 4 parametre sabit tutularak sadece 1 parametrenin değeri değiştirilerek yapılmalıdır. Her bir parametre kombinasyonu için motif uzunluğu 5-75 arası değerler ile denenmiştir.

Amaç en iyi parametreyi bulmak olduğu için ağacın her bir parametre değişikliği için temizlenip aynı motifler ile yeniden eğitilmesi ve yine aynı motifi bu yeni parametreyle tahmin etmesi gerekmektedir.

Fakat alt ve üst eşik ve artış veya azalış değeri belirlememize rağmen bu durum da bizim için verimli olmayacaktır. Çünkü her bir parametre için eşik değerleri arasındaki tüm artış veya azalış miktarı kadar olan her değere bakmak zaman açısından verimli olmayacaktır. Bu nedenle ikili arama yöntemine (binary search method) benzeyen bir yöntem kullanılmıştır. Şekil 4.5'de verilen bu yöntemin ikili arama yönteminden çok küçük bir farkı vardır.

0.0001	0.0	0.001	5	0.1
0.0001	0.0	0.001	5	5.0
0.0001	0.0	0.001	5	2.6
		.		
		.		
0.0001	0.0	0.001	5	1.1
0.0001	0.0	0.001	25	1.1
0.0001	0.0	0.001	15	1.1
		.		
		.		
0.0001	0.0	0.001	12	1.1
0.0001	0.0	0.2	12	1.1
0.0001	0.0	0.1	12	1.1
		.		
		.		
0.0001	0.0	0.003	12	1.1
0.0001	5.0	0.003	12	1.1
0.0001	2.5	0.003	12	1.1
		.		
		.		
0.0001	0.1	0.003	12	1.1
0.1	0.1	0.003	12	1.1
0.05	0.1	0.003	12	1.1
		.		
		.		
0.0002	**0.1**	**0.003**	**12**	**1.1**

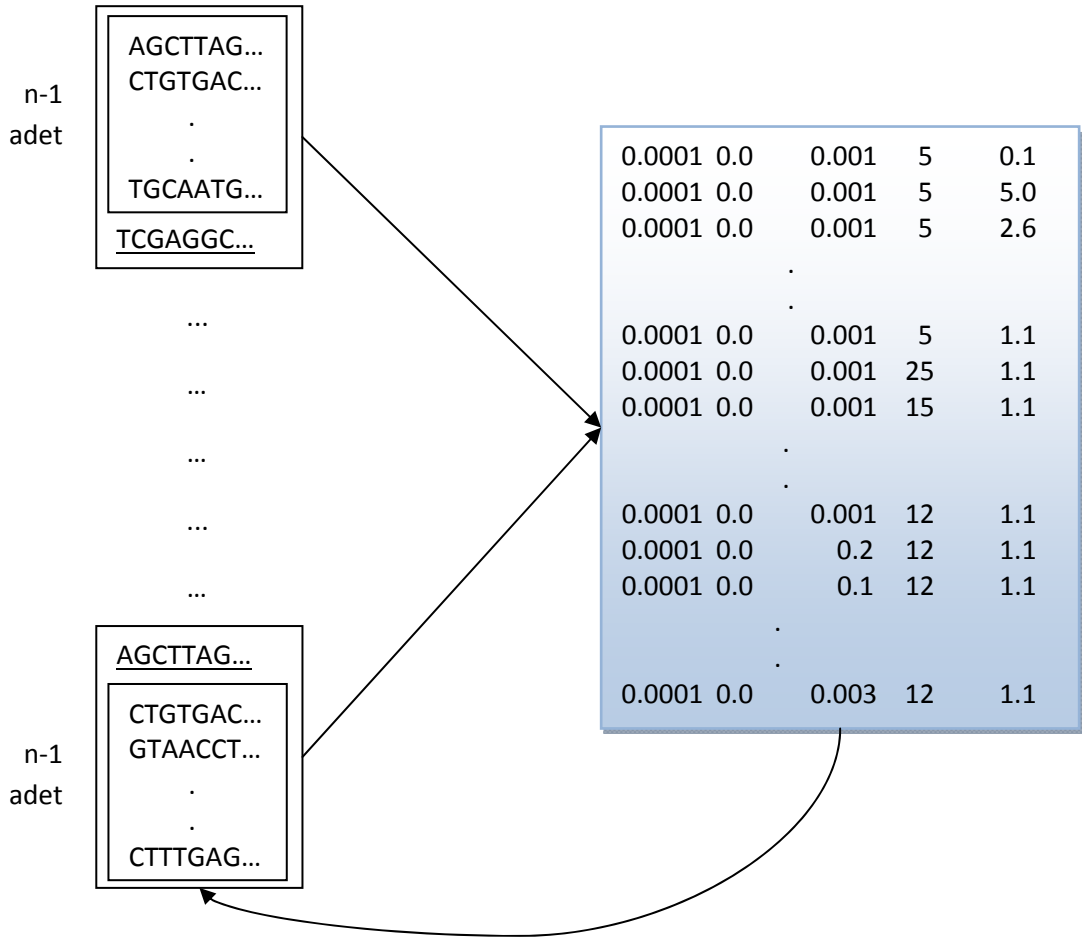
Şekil 4.5 – En İyi Parametrelerin Seçimi Örneği

Şekil 4.5'den de farkedileceği üzere ilk olarak alt ve üst eşik değerleri daha sonra bu eşik değerlerinin ortalamaları hesaba katılmaktadır. Daha sonra herbir parametreden elde edilen *nCC* skor değerine bakılarak hangi aralığa bakılacağına karar verilmiş ve alt-üst eşik değerleri bir sonraki parametre kombinasyonu için güncellenmiştir.

Bir motif grubu içinde eğitilmek için ayrılmış *n* adet motifler arasından *n-1* adet motif kullanılarak her farklı parametre için ağaç yeniden eğitilmiş ve aynı motif kısa tahmin yöntemi ile tahmin edilmiştir. Bu şekilde en yüksek *nCC* skoru veren beş parametre ve motif uzunluğu bulunmuştur. Bilindiği gibi motif tahmininde motifin uzunluğu da önemlidir. Parametre ve motif uzunluğu belirleme işlemini sadece bir motifin tahminine göre belirlemek doğru olmayacaktır. Bu nedenle ağaç, bu işlemi diğer muhtemel *n-1* adet kombinasyon için de tekrarlamalıdır. Bilindiği gibi *n* adet

eleman içeren bir kümede *n* tane farklı *n-1* adet eleman içeren alt kümeler ve *n* tane farklı 1 adet motif olabilir (Eşitlik 4.2).

$$\binom{n}{n-1} = n,\ \binom{n}{1} = n \tag{4.2}$$



Şekil 4.6 – Tüm Alt Gruplardaki Motiflere Aynı Parametre Bulma İşleminin Uygulanması

Şekil 4.6'da görüldüğü gibi ikili arama yöntemine benzeyen yöntem her bir *n* adet alt motif grubu için uygulanmıştır.

Her *n* adet durumdan elde edilmiş maksimum *nCC* değerlerini veren *n* farklı beş parametre ve motif uzunluğu bu *nCC* skorlarının yoğunluklarına göre 4.3 eşitliğindeki işleme tabi tutulmuştur. X değeri burada beş parametrenin her birini ve motif uzunluğunu temsil etmektedir. Yani bu işlemin her bir parametre ve motif uzunluğu için uygulanması gerekmektedir.

$$X = \frac{X_1(\max nCC_1+1)+X_2(\max nCC_2+1)+\cdots+X_n(maxnCC_n+1)}{n+\max nCC_1+\max nCC_2+\cdots+\max nCC_n} \quad (4.3)$$

Doğru parametreleri ve motif uzunluğunu bulduktan sonra ağaç temizlenmiş ve motif grubundaki *n* adet eğitim amaçlı motiflerin tümü ağacın eğitilmesi için kullanılmıştır. Ağacın motif tahmini için eğitiminde bu sefer doğru olarak seçildiği kabul edilen beş parametre kullanılacaktır.

4.2.2. OSA kullanarak tahminde bulunma ve motifleri tespit etme

Ağacın doğru parametrelerle eğitiminden sonra ilgili motif grubundaki test amaçlı ayrılmış motiflerin bulunduğu konumlara ve dizilimlere bakılmaksızın tahmin ve tespit işlemine geçilebilir. Test işleminin tahmin edilecek motiflere ait hiç bir bilgiye sahip olmadan yapılması gerekmektedir. Tahmin işlemi Veri Kümesi A olarak sınıflandırılan ilgili veri kümesindeki DNA dizilimleri üzerinde yapılmıştır.

OSA, tahmin işleminin sonuçlarını veri kümesindeki her bir DNA dizilimi nükleotidleri için teker teker yapacaktır. Her bir nükleotid için kullandığı sonek uzunluğunu ve olasılık dağılımını bir dosyaya alt alta yazacaktır.

Dosya çıktı örneğinin bir kısmı şekil 4.7'de verilmiştir.

```
1000  >seq_0
0     0.22727273
1     0.23076923
1     0.26923077
1     0.28400000
2     0.26000000
2     0.21428571
2     0.20000000
1     0.25937500
3     0.20000000
2     0.20000000
1     0.23125000
1     0.22000000
1     0.24615385
3     0.20000000
1     0.31739130
3     0.20000000
2     0.24545455
1     0.22000000
1     0.25384615
1     0.25937500
3     0.20000000
3     0.34000000
```

Şekil 4.7 – OSA Tahmin Sonucu Çıktısı

İlk satır ilgili DNA diziliminin uzunluğunu ve kaçıncı dizilim olduğu bilgisini vermektedir. Bu satırın altından itibaren ilk pozisyondan sona doğru tüm nükleotidlerin kullandığı sonek uzunluğu ve olasılık dağılım değeri sırasıyla verilmektedir. Daha sonra eğer başka dizilimler varsa aynı biçimde alt satırdan devam etmektedir. Görüldüğü üzere ilk pozisyondaki nükleotid için sonek uzunluğu sıfırdır. Çünkü öncesinde başka bir nükleotid yoktur.

Bu olasılık dağılımları arasından motifleri bulma işlemi için ağacın eğitilmesi sonucunda elde ettiğimiz motif uzunluğu verisinden faydalanılacaktır. Motif uzunluğu *k,* dizilim üzerinde bulunan o anki pozisyon numarası da *p* kabul edilirse:

$$MAX(toplam)$$

$$= sonekuzunluğu_p olasılıkdağılımı_p + sonekuzunluğu_{p+1} olasılıkdağılımı_{p+1} + \cdots$$

$$\ldots + sonekuzunluğu_{p+k-1} olasılıkdağılımı_{p+k-1}$$ elde edilir. (4.4)

Eşitlik 4.4'de görüldüğü üzere dosya üzerindeki en yüksek toplama sahip motif uzunluğu adedinde ardışık sonek uzunluğu ve olasılık dağılımı çarpımlarının

toplamı yapılmıştır. Bulunan en yüksek değerin ilk pozisyonu *p*, tahmin edilen motifin başlangıç pozisyonu kabul edilmiştir.

Olasılık dağılımı önemli bir kriterdir, bize o anki nükleotidin önündeki soneğe de bağlı olarak ne kadar sıklıkla görüldüğünü göstermektedir. Ama sadece olasılık dağılımının hesaba katılması doğru olmayacaktır, çünkü kullanılan sonek uzunluğu da önemlidir. OSA'nın tanımı gereği, ne kadar uzun bir sonek bulunursa o kadar çok benzeyen bir örüntü yakalanmış demektir. Bu nedenle sonek uzunluğu aynı satırdaki olasılık dağılımı ile çarpılarak iki kriterin de yoğunluğu hesaba katılmış olunur.

Bir sonraki adımda, eşitlik 4.4, test edilecek motif sayısı kadar işleme sokulacaktır. İşleme sokulmadan önce bulunan motiflerin başlangıç pozisyonları hesaba katılmayacaktır. Ancak bu yeterli bir çözüm olmayacaktır. Çünkü en yüksek toplama sahip motifin bulunduğu konumun komşu pozisyonlarında bulunan nükleotidlerin olasılık dağılımı ve sonek uzunluğu çarpımlarının toplamları, bulmuş olduğumuz en yüksek toplama yakın değerlerde olacaktır. Eğer bu konumları hesaba katarsak bir sonraki tespit edilecek motifin bu konumlarda çıkma olasılığı yüksek olacak ve asıl bulmak istediğimiz motif konumlarının göz ardı edilmesine neden olacaktır. Bu nedenle tahmin edilen motifin başlangıç pozisyonuna ek olarak belirli bir uzunluğa kadar bu pozisyonun önündeki ve arkasındaki nükleotid pozisyonları da bir sonraki motif tespiti için hesaba katılmamalıdır. Böylece daha doğru bir motif tespiti yapılabilecektir. Tespit edilen motifin başlangıç pozisyonuna ek olarak 100 öncesindeki ve sonrasındaki nükleotid pozisyonları da işlem dışında bırakılmıştır. Bu değer denenmeden önce 50, 75, 150, 200 ve en doğru parametrenin seçiminden elde edilmiş olan motif uzunluğu gibi değerler miktarında nükleotid pozisyonları hesap dışı bırakılmıştır. Ancak, 50 değeri gibi küçük değerler seçilince yeni tahmin edilecek motif önceden tahmin edilen motif yakınlarındaki nükleotid pozisyonlarında tespit edilmiştir. Bu durum bildiğimiz gibi yanlış tahmin olmuştur. 200 değeri gibi büyük değerler seçilince yeni tahmin edilecek motiflerin gerçek konumları hesap dışı kalmıştır. Bu durum da bildiğimiz gibi yanlış tahmin olmuştur.

4.2.3. Karşılaştırma yapma

Veri Kümesi B'deki test amaçlı ayrılan motiflerin konumları tespit edildikten sonra, bu motiflerin gerçekte bulundukları konumlarla karşılaştırılmaları yapılmıştır. Böylece motifin bulunduğu gerçek konum ile tahmin edilen konum arasındaki farklardan faydalanılarak motif tahmin doğruluğunun oranı belirlenmiştir. Karşılaştırma için motif araçlarının performans ölçümlerinde kullanılan denklemlerden faydalanılmıştır. Bunların arasında ağacın eğitimi sırasında kullandığımız eşitlik 4.1'deki *nCC* denklemi ve *nTP, nTN, nFP, nFN* değişkenleri de bulunmaktadır. Bunlara ek olarak;

- Eşitlik 4.5 ve 4.6'deki duyarlılık (sensitivity) parametresi (*nSn* ve *sSn*): Doğru tahmin edilen motif nükleotidlerinin ve konumlarının oranını göstermektedir.
- Eşitlik 4.7 ve 4.8'deki pozitif tahmin parametresi (*nPPV* ve *sPPV*): Doğru pozitif tahmin olasılıkları.
- Eşitlik 4.9'deki belirlilik (specificity) parametresi (*nSP*): Doğru tahmin edilen motif olmayan nükleotidlerin oranını göstermektedir.
- Eşitlik 4.10'daki performans katsayısı (*nPC*) parametresi
- Eşitlik 4.11'deki ortalama konum performansı (*sASP*) parametresi hesaplanmıştır.

Denklemlerde kullanılan üç değişken daha vardır:

- *sTP*, tahmin edilen motif konumları tarafından üst üste binen gerçek motif konumlarının sayısını gösterir.
- *sFN*, tahmin edilen motif konumları tarafından üst üste binmeyen gerçek motif konumlarının sayısını gösterir.
- *sFP,* gerçek motif konumları tarafından üst üste binmeyen tahmin edilen motif konumlarının sayısını gösterir.

$$nSn = \frac{nTP}{nTP+nFN} \tag{4.5}$$

$$sSn = \frac{sTP}{sTP+sFN} \tag{4.6}$$

$$nPPV = \frac{nTP}{nTP+nFP} \quad (4.7)$$

$$sPPV = \frac{sTP}{sTP+sFP} \quad (4.8)$$

$$nSP = \frac{nTN}{nTN+nFP} \quad (4.9)$$

$$nPC = \frac{nTP}{nTP+nFN+nFP} \quad (4.10)$$

$$sASP = \frac{sSn+sPPV}{2} \quad (4.11)$$

Bu denklemler Tompa'nın değerlendirmiş olduğu on dört adet motif bulma aracının performans ölçümünde de kullanılmıştır. Böylece bu çalışmada önerilen yöntemin başarısını diğer motif bulma araçlarıyla karşılaştırabilme imkanını bulmuş olduk.

5. DENEYSEL SONUÇLAR

Yapılan çalışmada OSA yöntemi kullanılarak DNA motifleri tahmin edilmiştir. Oluşturulan veri kümesindeki eğitim amaçlı motifler kullanılarak OSA eğitilmiş ve bu eğitilen ağaç kullanılarak test amaçlı ayrılmış motifler bulundukları DNA dizilimleri üzerinden tahmin edilerek tespit edilmeye çalışılmıştır. Tahmin edilen motiflerin sonuçları, motiflerin gerçek konumları ile karşılaştırılmıştır.

Elde edilen sonuçlar on dört farklı motif bulma aracı ile kıyaslanmıştır:

1. AlignACE
2. ANN-Spec
3. Consensus
4. GLAM
5. Improbizer
6. MEME
7. MEME3
8. MITRA
9. MotifSampler
10. Oligo/Dyad Analysis
11. QuickScore
12. SeSiMCMC
13. Weeder
14. YMF

Dört farklı canlı türüne ait motifler üzerinde çalışılmıştır:

1. *Drosophila melanogaster* denilen bir sinek türüne ait DNA dizilimleri
2. Fareye ait DNA dizilimleri
3. İnsana ait DNA dizilimleri
4. *Saccharomyces cerevisiae* denilen bir maya türüne ait DNA dizilimleri

Çizelge 5.1 OSA yönteminin ve diğer on dört motif bulma araçlarının sinek türünün motiflerine uygulanmış tahmin sonuçlarını göstermektedir.

Çizelge 5.2 OSA yönteminin ve diğer on dört motif bulma araçlarının fare türünün motiflerine uygulanmış tahmin sonuçlarını göstermektedir.

Çizelge 5.3 OSA yönteminin ve diğer on dört motif bulma araçlarının insan türünün motiflerine uygulanmış tahmin sonuçlarını göstermektedir.

Çizelge 5.4 OSA yönteminin ve diğer on dört motif bulma araçlarının maya motiflerine uygulanmış tahmin sonuçlarını göstermektedir.

Çizelge 5.5 OSA yönteminin ve diğer on dört motif bulma araçlarının dört canlı türünün bütün motiflerine uygulanmış tahmin sonuçlarını göstermektedir.

Çizelge 5.3'e bakıldığında uygulamış olduğumuz yöntem olan OSA'nın insan genomunda çok iyi başarı gösterdiğini görmekteyiz. nCC değeri %4.9 motif bulma başarı oranı ile kıyaslanan 14 motif bulma araçlarının 9 tanesinden daha iyi sonuç vermiştir. sASP değeri (motif konumları bazında duyarlılık ve pozitif tahmin değeri ortalaması) insan genomunda %12.1 ile 12 motif bulma aracından daha iyi sonuç vermiştir. Buna ek olarak OSA, insan genomunda nükleotid pozisyonları bazında duyarlılık değeri (nSn) %16.1 ile diğer motif araçları arasında en iyi sonucu vermiştir. Bu oran ile literatürdeki duyarlılık başarı yüzdesini %7.1 artırmıştır.

Çizelge 5.2'e bakıldığında uygulamış olduğumuz yöntem fare genomunda da başarı elde etmiştir. %2.3 motif bulma başarı oranı ile 5 motif bulma aracından daha iyi sonuçlar elde edilmiştir. Ayrıca insan genomundaki başarısı gibi fare genomunda da nSn %11.8 ile diğer tüm motif bulma araçlarından daha iyi bir sonuç vermiştir. Bu oran ile literatürdeki duyarlılık başarı yüzdesini %1 artırmıştır.

Çizelge 5.5'e bakıldığında uygulamış olduğumuz yöntem tüm canlılara ait genomlarda başarı elde etmiştir. %2.7 motif bulma oranı ile fare genomunda olduğu gibi 5 motif bulma aracını geride bırakmıştır. Ayrıca yine fare ve insan genomunda olduğu gibi nSN %10.8 ile tüm motif bulma araçlarından daha iyi sonuçlar elde edilmiş ve literatürdeki başarı yüzdesini %2.1 ile yukarı çekmiştir.

Çizelge 5.1 ve 5.2'ye bakıldığında önermiş olduğumuz yöntem sinek ve maya genomlarında literatürün gerisinde kalmıştır. Bu durumun, kullanılan test ve eğitim verilerinin yetersizliğinden kaynaklandığı düşünülmektedir. Buna ek olarak farklı canlı türleri farklı gen dizilimlerine sahiptirler. Bu nedenle bir motif bulma aracının tüm canlı gruplarına ait motiflerde başarı gösterebilmesi söz konusu değildir. Bu konuya ayrıntılı olarak bir sonraki bölümde değinilecektir.

Performans kriterlerinden olan nPPV (nükleotid bazında pozitif tahmin değeri), sSn (motif konumu bazında duyarlılık), nSP (doğru tahmin edilen motif olmayan nükleotid pozisyonları) ve nPC parametrelerinde hedeflenen sonuçlar elde edilememiştir.

Çizelge 5.1 – Sineklere Ait Motiflerin Tahmin Sonuçları

dm (sinek)	nSn	sSn	nPPV	sPPV	nSP	nPC	sASP	nCC
SeSiMCMC	0.101	0.098	0.054	0.125	0.972	0.037	0.112	0.054
MEME	0.042	0.059	0.042	0.056	0.985	0.021	0.057	0.027
MEME3	0.037	0.059	0.026	0.045	0.978	0.016	0.052	0.013
Weeder	0.012	0.020	0.035	0.035	0.995	0.009	0.027	0.011
ANN-Spec	0.025	0.020	0.018	0.009	0.978	0.011	0.015	0.002
Improbizer	0.015	0.020	0.018	0.023	0.987	0.008	0.021	0.002
MotifSampler	0.005	0	0.008	0	0.992	0.003	0	-0.006
AlignACE	0	0	0	0	0.997	0	0	-0.006
MITRA	0	0	0	0	0.996	0	0	-0.008
GLAM	0.003	0	0.005	0	0.990	0.002	0	-0.009
Consensus	0	0	0	0	0.992	0	0	-0.011
YMF	0	0	0	0	0.988	0	0	-0.014
O/D* analysis	0	0	0	0	0.986	0	0	-0.015
QuickScore	0	0	0	0	0.984	0	0	-0.016
OSA	**0**	**0**	**0**	**0**	**0.943**	**0**	**0**	**-0.018**

O/D* = Oligo/Dyad

Çizelge 5.2 – Farelere Ait Motiflerin Tahmin Sonuçları

mus (fare)	nSn	sSn	nPPV	sPPV	nSP	nPC	sASP	nCC
YMF	0.101	0.204	0.202	0.182	0.988	0.072	0.193	0.125
MEME3	0.106	0.194	0.170	0.200	0.984	0.070	0.197	0.114
Improbizer	0.108	0.225	0.122	0.161	0.976	0.061	0.193	0.089
Weeder	0.062	0.123	0.176	0.180	0.991	0.048	0.151	0.088
MEME	0.073	0.143	0.134	0.175	0.986	0.050	0.159	0.079
Consensus	0.049	0.102	0.106	0.122	0.988	0.035	0.112	0.053
MotifSampler	0.045	0.082	0.091	0.095	0.987	0.031	0.088	0.044
O/D analsis	0.027	0.061	0.107	0.095	0.993	0.022	0.078	0.040
QuickScore	0.036	0.082	0.090	0.068	0.989	0.026	0.075	0.039
OSA	**0.118**	**0.041**	**0.016**	**0.095**	**0.942**	**0.014**	**0.068**	**0.023**
AlignACE	0.029	0.031	0.049	0.036	0.983	0.018	0.033	0.015
SeSiMCMC	0.062	0.102	0.041	0.095	0.956	0.025	0.099	0.015
ANN-Spec	0.043	0.082	0.043	0.043	0.971	0.022	0.062	0.014
GLAM	0.007	0.010	0.019	0.015	0.989	0.005	0.013	-0.007
MITRA	0.006	0.020	0.016	0.033	0.988	0.004	0.027	-0.009

Çizelge 5.3 – İnsanlara Ait Motiflerin Tahmin Sonuçları

hm (insan)	nSn	sSn	nPPV	sPPV	nSP	nPC	sASP	nCC
Weeder	0.054	0.107	0.275	0.258	0.997	0.048	0.183	0.116
O/D analysis	0.037	0.060	0.214	0.150	0.998	0.033	0.105	0.083
ANN-Spec	0.090	0.164	0.103	0.098	0.986	0.051	0.131	0.081
AlignACE	0.039	0.074	0.103	0.124	0.994	0.030	0.099	0.053
YMF	0.041	0.074	0.097	0.080	0.993	0.030	0.077	0.052
OSA	**0.161**	**0.070**	**0.021**	**0.172**	**0.973**	**0.019**	**0.121**	**0.049**
MEME	0.038	0.060	0.060	0.081	0.989	0.024	0.071	0.034
Improbizer	0.042	0.071	0.048	0.048	0.985	0.023	0.059	0.028
MEME3	0.042	0.064	0.047	0.079	0.985	0.023	0.071	0.028
MITRA	0.024	0.040	0.047	0.047	0.991	0.016	0.044	0.022
MotifSampler	0.025	0.047	0.042	0.043	0.990	0.016	0.045	0.019
GLAM	0.024	0.040	0.037	0.060	0.989	0.015	0.050	0.016
SeSiMCMC	0.046	0.067	0.028	0.063	0.971	0.018	0.065	0.014
QuickScore	0.005	0	0.010	0	0.991	0.003	0	-0.006
Consensus	0	0	NaN*	NaN	1	0	NaN	NaN

NaN* = Not a Number anlamına gelir. Sıfıra bölünme istisnasından kaynaklanır.

Çizelge 5.4 – Mayalara Ait Motiflerin Tahmin Sonuçları

yst (maya)	nSn	sSn	nPPV	sPPV	nSP	nPC	sASP	nCC
Weeder	0.293	0.520	0.534	0.550	0.995	0.233	0.535	0.386
MotifSampler	0.257	0.387	0.504	0.491	0.995	0.205	0.440	0.350
MEME	0.193	0.307	0.380	0.383	0.994	0.147	0.345	0.260
MEME3	0.210	0.320	0.300	0.304	0.990	0.141	0.312	0.238
YMF	0.144	0.280	0.330	0.339	0.994	0.112	0.309	0.208
AlignACE	0.186	0.280	0.185	0.202	0.983	0.102	0.241	0.168
O/D Analysis	0.090	0.187	0.330	0.304	0.996	0.076	0.246	0.164
ANN-Spec	0.165	0.307	0.140	0.141	0.980	0.082	0.224	0.133
Consensus	0.079	0.147	0.200	0.239	0.994	0.060	0.193	0.115
MITRA	0.111	0.160	0.153	0.154	0.987	0.069	0.157	0.115
Improbizer	0.157	0.267	0.095	0.133	0.970	0.063	0.200	0.099
SeSiMCMC	0.101	0.093	0.057	0.072	0.965	0.038	0.083	0.051
GLAM	0.071	0.147	0.059	0.058	0.976	0.033	0.102	0.043
QuickScore	0.054	0.120	0.061	0.044	0.983	0.029	0.082	0.039
OSA	**0**	**0**	**0**	**0**	**0.959**	**0**	**0**	**-0.017**

Çizelge 5.5 – Veri Kümesindeki Tüm Canlılara Ait Motiflerin Tahmin Sonuçları

Tüm Motifler	nSn	sSn	nPPV	sPPV	nSP	nPC	sASP	nCC
Weeder	0.054	0.107	0.300	0.289	0.996	0.072	0.225	0.152
YMF	0.064	0.121	0.137	0.120	0.992	0.046	0.120	0.082
MEME	0.067	0.111	0.107	0.139	0.989	0.043	0.125	0.071
O/D* Analysis	0.040	0.073	0.154	0.121	0.996	0.033	0.097	0.069
MotifSampler	0.060	0.098	0.107	0.101	0.990	0.040	0.100	0.067
AlignACE	0.055	0.088	0.112	0.122	0.991	0.038	0.105	0.066
SeSiMCMC	0.061	0.081	0.037	0.075	0.969	0.024	0.078	0.049
Consensus	0.020	0.040	0.113	0.133	0.997	0.018	0.087	0.041
MITRA	0.031	0.050	0.062	0.063	0.991	0.021	0.057	0.031
OSA	**0.108**	**0.047**	**0.014**	**0.110**	**0.965**	**0.013**	**0.079**	**0.027**
QuickScore	0.017	0.033	0.030	0.019	0.989	0.011	0.026	0.008
ANN-Spec	0.087	0.155	0.088	0.085	0.982	0.046	0.120	NaN**
MEME3	0.077	0.125	0.091	0.135	0.985	0.044	0.130	NaN
Improbizer	0.069	0.123	0.070	0.084	0.982	0.036	0.103	NaN
GLAM	0.026	0.046	0.038	0.048	0.987	0.016	0.047	NaN

6. DEĞERLENDİRME VE GELECEK ÇALIŞMA PLANI

Bu çalışmada yazım faktörleri için DNA bağlanma konumlarında bulunan motifleri tahmin eden bir yöntem geliştirilmiş ve belirlenmiş performans kriterleri çerçevesinde kıyaslanan on dört motif bulma aracından daha iyi sonuçlar elde etme hedeflenmiştir.

Uygulanan yöntem maksimum olasılık prensibine dayalı olan ve protein dizilimlerinde önceden kullanılmış ama DNA motif tahmininde ilk defa kullanılan Olasılıksal Sonek Ağacı (OSA)'dır. DNA dizilim verileri olarak düzenlenmiş genlerin başlatıcı dizilimleri tipinde veri kümesi kullanılmıştır. Bu çalışmada önerilen yöntemin sonuçları literatürdeki benzer çalışmalarla performans ve başarım açısından karşılaştırılmıştır.

Deneysel sonuçlar çerçevesinde nPPV, sSn, nSP ve nPC performans kriterlerinin bu çalışmada ele alınan canlı genomlarında önerilen yöntemle hedeflenen başarıya ulaşamadığı gözlemlenmiştir.

İnsan genomu ele alındığında, önerilen yöntemin nCC değeri diğer motif bulma araçlarının çoğuna kıyasla daha iyi bir sonuç üretmiştir. Ayrıca sASP değeri diğer motif araçlarının çoğu arasında en iyi sonuçlardan bir tanesini vermiştir. İnsan genomunda doğru bir şekilde tahmin edilen motif konumlarının sayısının fazla olmasından dolayı sASP değeri yüksek çıkmıştır. Buna ek olarak nSn performans kriteri ile, hedeflediğimiz amaçlardan bir tanesine ulaşmış olduk. Çünkü en iyi duyarlılık değerini insan genomunda OSA yöntemi vermiştir. Bu sonuca sASP değerinin de katkısı olduğu düşünülmektedir. OSA, insan genomunda diğer on dört adet motif bulma aracına göre en yüksek nSn değerini elde etmiştir. Aynı başarı veri kümesindeki bütün canlıların motiflerinde de sağlanmıştır. Ayrıca nCC değeri olarak da bütün motiflerde iyi bir başarı elde edilmiştir.

Karşılaştırılan bu on dört adet motif bulma aracı ele alındığında çizelge 5.4'te de görüldüğü üzere bu araçlar mayaya ait motiflerde diğer motiflere kıyasla daha yüksek bir başarı oranı elde etmişlerdir. Bunun sebeplerinden birisi bu motif bulma

araçlarının günümüze kadar birçok defa geliştirilmesinden kaynaklanmaktadır. Ayrıca yöntemlerin neredeyse hepsi *Saccharomyces cerevisiae* isimli maya türüne ait motiflerde sürekli olarak denenmiştir. Çünkü bu mayaların birçok türüne ait motiflerin bağlanma konumlarını bulunduran sayıca fazla DNA dizilimleri DNA veri bankalarında tutulmaktadır ve kullanıma da açıktır.

Gerek bu çalışmada önerilen yöntem, gerekse diğer on dört motif bulma yöntemi genlerin düzenleyici mekanizmasını hala anlayamamaktadır. Çizelge 5.5'deki Weeder'ın nCC değerine bakıldığında en iyi sonucu bu yöntemin verdiği görülmektedir. Ancak literatürdeki en iyi oran olan %15.2 bile çok düşük bir değerdir. Buradan da bu düzenleyici mekanizmanın hala anlaşılamadığı sonucunu çıkarabiliriz. Diğer bir problem ise farklı canlı türlerinin motif karakteristiklerinin de farklı olmasıdır. Yani bir veri kümesinde iyi başarı elde etmiş bir yöntemin başka bir veri kümesinde de aynı başarıyı yakalayacağını söylemek çok yanlış olacaktır. Örneğin on dört motif bulma aracından biri olan Consensus insan genomunda en kötü başarıyı gösterirken, fare genomunda birçok motif bulma aracından daha iyi bir sonuç üretmiştir. Her ne kadar Weeder yöntemi genelde en iyi başarıyı gösterse bile fare genomunda YMF ve sinek genomunda SeSiMCMC yöntemi en iyi başarıyı göstermiştir.

Motif bulma araçlarını kullanmak isteyen ve hangisinin kullanılması gerektiğini bilemeyen biyologlara önereceğimiz motif bulma aracı bir tane değil birden fazla olacaktır. Çünkü bu çalışmada uyguladığımız yöntemden ve kıyaslanan diğer motif bulma araçlarından elde edilen sonuçlara dayanarak bir motif bulma aracı kullanılan veri kümesinde diğer motif bulma araçlarını tek başına geride bırakamamıştır. Farklı motif bulma araçları farklı veri kümelerinde başarı göstermiştir.

Bu çalışmada önerilen OSA yöntemi, kullanılan veri kümesindeki insan ve fare genomunda çok başarılı sonuçlar verirken, maya ve sinek genomlarında literatürün gerisinde kalmıştır. Değinilmesi gerekilen önemli bir nokta literatür çalışması başlığı altında da değinildiği üzere insanın mayadan farklı olarak (kıyaslanan on dört motif bulma aracı maya motiflerinde iyi sonuçlar elde etmiştir) yüksek seviyeli organizmaya sahip olmasıdır. Yüksek seviyeli organizmalarda

motif tahmin işlemi düşük seviyeli organizmalara göre daha zordur. Çünkü daha karmaşık bir yapıya sahiptirler. Bu çalışmada önerdiğimiz yöntem insan genomlarında başarı göstererek önemli bir işi daha başarılı bir şekilde gerçekleştirmiştir. OSA yöntemi ilk defa DNA dizilimlerine uygulanmıştır ve bu yöntemin geliştirilmesi ile daha başarılı sonuçların elde edilebileceği düşünülmektedir. Protein dizilimlerine önceden uygulanmış olan biyolojik OSA ve ikili OSA yöntemleri gibi OSA'nın geliştirilmiş halleri DNA dizilimlerinde de kullanılarak daha iyi bir başarıya ulaşacağı düşünülmektedir.

Gelecekteki çalışmalar arasında OSA'nın geliştirilmiş hallerinin kullanılması ve farklı motif gruplarındaki başarısının test edilmesi için farklı veri kümelerinin kullanılması (filogenetik ayak izi) planlanmaktadır. Ayrıca OSA yönteminin diğer on dört motif bulma aracı ile zaman ve alan karmaşıklığı açısından da kıyaslanması hedeflenmektedir.

7. KAYNAKLAR

[1] Ao W, Gaudet J, Kent WJ, Muttumu S, Mango SE: Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. Science 2004, 305:1743-1746.

[2] Apostolico A, Bock M, Lonardi S, Xu X: Efficient detection of unusual words. J Comput Biol 2000, 7:71-94.

[3] Bailey TL, Elkan C: Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning 1995, 21:51-80.

[4] Bejerano G, Yona G: Modeling protein families using probabilistic suffix trees. Annual Conference on Research in Computational Molecular Biology Proceedings of the third annual international conference on Computational molecular biology Lyon, France Pages: 15 - 24

[5] Bejerano G, Yona G: Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. Bioinformatics 2001 Jan;17(1):23-43.

[6] Berezikov E, Guryev V, Plasterk RHA, Cuppen E: CONREAL: Conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. Genome Res 2004, 14:170-178

[7] Blanchette M, Tompa M: Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res 2002, 12:739-748

[8] Bray N, Dubchak I, Pachter L: AVID: A global alignment program. Genome Res 2003, 13:97-102.

[9] Brazma A, Jonassen I, Vilo J, Ukkonen E: Predicting gene regulatory elements in silico on a genomic scale. Genome Res 1998, 8:1202-1215

[10] Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, NISC Comparative Sequencing Program, Green ED, Sidow A, Batzoglou S: LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 2003, 13:721-731.

[11] Bucher P: Weight matrix description for four eukaryotic RNA polymerase II promoter element derived from 502 unrelated promoter sequences. J Mol Biol 1990, 212:563-578

[12] Buhler J, Tompa M: Finding motifs using random projections. J Comput Biol 2002, 9:225-242.

[13] Bussemaker H, Li H, Siggia E: Regulatory element detection using a probabilistic segmentation model. Proceedings of the Eighth International Conference on Intelligent Systems on Molecular Biology, San Diego, CA 2000, 67-74.

[14] Carmack CS, McCue LA, Newberg LA, Lawrence CE: PhyloScan: identification of transcription factor binding sites using cross-species evidence. Algorithms for Molecular Biology 2007, 2:1.

[15] Chengpeng Bi: A Genetic-Based EM Motif-Finding Algorithm for Biological Sequence Analysis. This paper appears in: Computational Intelligence and Bioinformatics and Computational Biology, 2007. CIBCB '07. IEEE Symposium on Publication Date: 1-5 April 2007 On page(s): 275-282

[16] Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 2003, 301:71-76.

[17] Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH, Johnston M: Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. Genome Res 2001, 11:1175-1186

[18] Down TA, Hubbard TJP: NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. Nucleic Acids Res 2005, 33:1445-1453.

[19] Eskin E, Pevzner P: Finding composite regulatory patterns in DNA sequences. Bioinformatics 2002, 18(Suppl 1):S354-S363.

[20] Favorov AV, Gelfand MS, Gerasimova AV, Mironov AA, Makeev VJ: Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length and its validation on the ArcA binding sites. Proceedings of Fourth International Conference on Bioinformatics of Genome Regulation and Structure, Novosibirsk, Russia 2004.

[21] Fogel GB, Weekes DG, Varga G, Dow ER, Harlow HB, Onyia JE, Su C: Discovery of sequence motifs related to coexpression of genes using evolutionary computation. Nucleic Acids Res 2004, 32:3826-3835.

[22] Frith MC, Hansen U, Spouge JL, Weng Z: Finding functional sequence elements by multiple local alignment. Nucleic Acids Res 2004, 32:189-200.

[23] Galas DJ, Eggert M, Waterman MS: Rigorous pattern-recognition methods for DNA sequences: analysis of promoter sequences from Escherichia coli. J Mol Biol 1985, 186:117-128.

[24] Ganesh R, Siegele DA, Ioerger TR: MOPAC: motif finding by preprocessing and agglomerative clustering from microarrays. Proceedings of the Eighth Pacific Symposium on Biocomputing 2003, 41-52.

[25] Gelfand MS, Koonin EV, Mironov AA: Prediction of transcription regulatory sites in Archaea by a comparative genome approach. Nucleic Acids Res 2000, 28:695-705.

[26] GuhaThakurta D, Stormo GD: Identifying target sites for cooperatively binding factors. Bioinformatics 2001, 17:608-621.

[27] H.Oğul, E.Mumcuoğlu, (2006), SVM-based Detection of Distant Protein Structural Relationships Using Pairwise Probabilistic Suffix Trees Computational Biology and Chemistry 30, 292-299 (Elsevier)

[28] Hertz GZ, Hartzell GW, Stormo GD: Identification of consensus patterns in unaligned DNA sequences known to be functionally related. Comput Appl Biosci 1990, 6:81-92.

[29] Hertz GZ, Stormo GD: Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 1999, 15:563-577

[30] Hon LS, Jain AN: A deterministic motif finding algorithm with application to the human genome. Bioinformatics 2006, 22:1047-1054.

[31] Hu J, Li B, Kihara D: Limitations and potentials of current motif discovery algorithms. Nucleic Acids Res 2005, 33:4899-4913.

[32] Hu J, Yang YD, Kihara D: EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences. BMC Bioinformatics 2006, 7:342.

[33] Hughes JD, Estep PW, Tavazoie S, Church GM: Computational identification of cis-regulatory elements associated with functionally coherent groups of genes in Saccharomyces cerevisiae. J Mol Biol 2000, 296:1205-1214.

[34] Ishikawa M, Toya T, Hoshida M, Nitta K, Ogiwara A, Kanehisa M: Multiple sequence alignment by parallel simulated annealing. Comput Appl Biosci 1993, 9:267-273.

[35] Kaplan T, Friedman N, Margalit H: Ab initio prediction of transcription factor targets using structural knowledge. PLoS Comput Biol 2005, 1(1):e1.

[36] Karlin S, Altschul SF: Methods for assessing the statistical significance of sequence features by using general scoring schemes. PNAS 1990, 87:2264-2268.

[37] Kaya M: A Multi-Objective Genetic Algorithm for Discovering Non-Dominated Motifs in DNA Sequences. Hybrid Intelligent Systems, 2007. HIS 2007. 7th International Conference on Volume , Issue , 17-19 Sept. 2007 Page(s):180 – 185

[38] Kaya M: MOGAMOD: Multi-Objective Genetic Algorithm for Motif Discovery, Expert Systems with Applications, In Press, 2007.

[39] Kellis M, Patterson N, Endrizzi M, Birren B, Lander E: Sequencing and comparison of yeast species to identify genes and regulatory element. Nature 2003, 423:241-254

[40] Kielbasa SM, Korbel JO, Beule D, Schuchhardt J, Herzel H: Combining frequency and positional information to predict transcription factor binding sites. Bioinformatics 2001, 17:1019-1026

[41] Kim J, Pramanik S, Chung MJ: Multiple sequence alignment using simulated annealing. Comput Appl Biosci 1994, 10:419-426.

[42] Kingsford C, Zaslavsky E, Singh M: A compact mathematical programming formulation for DNA motif finding. Lecture Notes in Computer Science 2006, 4009:233-245.

[43] Kirkpatrick S, Gelatt CD, Vecchi MP: Optimization by simulated annealing. Science 1983, 220:671-680.

[44] Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 1993, 262:208-214

[45] Lawrence CE, Reilly AA: An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins 1990, 7:41-51

[46] Leung HCM, Chin FYL: Finding motifs from all sequences with and without binding sites. Bioinformatics 2006, 22:2217-2223.

[47] Liang S: cWINNOWER algorithm for finding fuzzy DNA motifs. IEEE Computer Society Bioinformatics Conference 2003, 260-265.

[48] Liu D, Xiong X, DasGupta B, Zhang H: Motif discoveries in unaligned molecular sequences using self-organizing neural network. IEEE Transactions on Neural Networks 2006, 17:919-928

[49] Liu FFM, Tsai JJP, Chen RM, Chen SN, Shih SH: FMGA: finding motifs by genetic algorithm. Fourth IEEE Symposium on Bioinformatics and Bioengineering 2004, 459.

[50] Liu JS: Monte Carlo Strategies in Scientific Computing. Springer Series in Statistics; 2001.

[51] Liu JS, Neuwald AF, Lawrence CE: Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. J Amer Statist Assoc 1995, 90:1156-1170.

[52] Liu JS: The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. J Amer Statist Assoc 1994, 89:958-966.

[53] Liu X, Brutlag DL, Liu JS: BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Proceedings of the Sixth Pacific Symposium on Biocomputing 2001, 127-138.

[54] Liu XS, Brutlag DL, Liu JS: An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol 2002, 20:835-839.

[55] Marsan L, Sagot M: Algorithms for extracting structured motifs using a Suffix tree with an application to promoter and regulatory site consensus identification. J Comput Biol 2000, 7:345-362

[56] McCue L, Thompson W, Carmack C, Ryan M, Liu J, Derbyshire V, Lawrence C: Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. Nucleic Acids Res 2001, 29:774-782.

[57] McGuire AM, Hughes JD, Church GM: Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. Genome Res 2000, 10:744-757.

[58] Mendes ND, Casimiro AC, Santos PM, Sa-Correira I, Oliveira AL, Freitas AT: MUSA: a parameter free algorithm for the identification of biologically significant motifs. Bioinformatics 2006, 22:2996-3002.

[59] Mengeritsky G, Smith TF: Recognition of characteristic patterns in sets of functionally equivalent DNA sequences. Comput Appl Biosci 1987, 3:223-227.

[60] Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B: MIPS: a database for genomes and protein sequences. Nucleic Acids Res 2002, 30:31-34.

[61] Mohapatra, A.; Mishra, P.M.; Padhy, S: Motif Search in DNA Sequences Using Generalized Suffix Tree. Information Technology, (ICIT 2007). 10th International Conference onVolume, Issue, 17-20 Dec. 2007 Page(s):100 – 103

[62] Moses A, Chiang D, Eisen M: Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. Proceedings of the Ninth Pacific Symposium on Biocomputing 2004, 324-335.

[63] Pavesi G, Mauri G, Pesole G: An algorithm for finding signals of unknown length in DNA sequences. Bioinformatics 2001, 17(Suppl 1):S207-S214.

[64] Peng CH, Hsu JT, Chung YS, Lin YJ, Chow WY, Hsu DF, Tang CY: Identification of degenerate motifs using position restricted selection and hybrid ranking combination. Nucleic Acids Res 2006, 34:6379-6391.

[65] Pesole G, Prunella N, Liuni S, Attimonelli M, Saccon C: WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. Nucleic Acids Res 1992, 20:2871-2875.

[66] Pevzner P, Sze S: Combinatorial approaches to finding subtle signals in DNA sequences. Proceedings of the Eighth International Conference on Intelligent Systems on Molecular Biology, San Diego, CA 2000, 269-278.

[67] Prakash A, Blanchette M, Sinha S, Tompa M: Motif discovery in heterogeneous sequence data. Proceedings of the Ninth Pacific Symposium on Biocomputing 2004, 348-359.

[68] Qi Y, Ye P, Bader JS: Genetic interaction motif finding by expectation maximization – a novel statistical model for inferring gene modules from synthetic lethality. BMC Bioinformatics 2005, 6:288.

[69] Regnier M, Denise A: Rare events and conditional events on random strings. Discrete Math Theor Comput Sci 2004, 6:191-214.

[70] Rombauts S, Dehais P, Van Montagu M, Rouze P: PlantCARE, a plant cis acting regulatory element database. Nucleic Acids Res 1999, 27:295-296.

[71] Ron D, Singer Y. and Tishby N. (1996) The power of amnesia: learning probabilistic automata with variable memoty length. Machine Learning, 25, 117-149.

[72] Roth FP, Hughes JD, Estep PW, Church GM: Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nature Biotechnology 1998, 16:939-945.

[73] Sagot M: Spelling approximate repeated or common motifs using a suffix tree. Lecture Notes in Computer Science 1998, 1380:111-127.

[74] Schneider TD, Stephens RM: Sequence logos: a new way to display consensus sequence. Nucleic Acids Res 1990, 18:6097-6100.

[75] Shida K: GibbsST: a Gibbs sampling method for motif discovery with enhanced resistance to local optima. BMC Bioinformatics 2006, 7:486.

[76] Siddharthan R, Siggia ED, van Nimwegen E: PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. PLoS Comput Biol 2005, 1:534-556.

[77] Sinha S, Blanchette M, Tompa M: PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. BMC Bioinformatics 2004, 5:170.

[78] Sinha S: Discriminative motifs. J Comput Biol 2003, 10:599-615.

[79] Sinha S, Tompa M: A statistical method for finding transcription factor binding site. Proceedings of the Eighth International Conference on Intelligent Systems on Molecular Biology, San Diego, CA 2000, 344-354.

[80] Sinha S, Tompa M: Performance comparison of algorithms for finding transcription factor binding sites. In Third IEEE Symposium on Bioinformatics and Bioengineering. IEEE Press; 2003:214-220.

[81] Staden R: Methods for discovering novel motif in nucleic acid sequences. Comput Appl Biosci 1989, 5:293-298

[82] Stormo GD: DNA binding sites: representation and discovery. Bioinformatics 2000, 16:16-23.

[83] Sun Z., Deogun J.S.: Local prediction approach for protein classification using probabilistic suffix trees. ACM International Conference Proceeding Series; Vol. 55 Proceedings of the second conference on Asia-Pacific bioinformatics - Volume 29 Dunedin, New Zealand Pages: 357 - 362

[84] Tagle D, Koop B, Goodman M, Slightom J, Hess D, Jones R: Embryonic ε and γ globin genes of a prosimian primate (Galago crassicaudatus): nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 1988, 203:439-455

[85] Thijs G, Marchal K, Moreau Y: A Gibbs sampling method to detect over-represented motifs in upstream regions of co-expressed genes. RECOMB 2001, 5:305-312.

[86] Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res 1994, 22:4673.

[87] Thompson W, Rouchka E, Lawrence C: Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Res 2003, 31:3580-3585.

[88] Tompa M: An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. Proceedings of the Seventh International Conference on Intelligent Systems on Molecular Biology 1999, 262-271.

[89] Tompa M: Identifying functional elements by comparative DNA sequence analysis. Genome Res 2001, 11:1143-1144.

[90] Tompa M, Li N, Bailey T, Church GM, De Moor B, Eskin E, Favorov A, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 2005, 23:137-144.

[91] van Helden J, Andre B, Collado-Vides J: Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 1998, 281:827-842.

[92] van Helden J, Rios AF, Collado-Vides J: Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res 2000, 28:1808-1818.

[93] Vanet A, Marsan L, Labigne A, Sagot MF: Inferring regulatory elements from a whole genome. An analysis of Helicobacter pylori σ^{80} family of promoter signals. J Mol Biol 2000, 297:335-353

[94] Vilo J, Brazma A, Jonassen I, Robinson A, Ukonnen E: Mining for putative regulatory elements in the yeast genome using gene expression data. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. AAAI Press San Diego, CA; 2000:384-394.

[95] Wang G, Yu T, Zhang W: WordSpy: identifying transcription factor binding motifs by building a dictionary and learning a grammar. Nucleic Acids Res 2005, (33 Web Server):W412-W416.

[96] Wang T, Stormo GD: Combining phylogenetic data with coregulated genes to identify regulatory motifs. Bioinformatics 2003, 19:2369-2380.

[97] Wang T, Stormo GD: Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. PNAS 2005, 102:17400-17405.

[98] Wei Z, Jensen ST: GAME: detecting cis-regulatory elements using a genetic algorithm. Bioinformatics 2006, 22:1577-1584.

[99] Wingender E, Dietze P, Karas H, Knuppel R: TRANSFAC: a Database on transcription factors and their DNA binding sites. Nucleic Acids Res 1996, 24:238-241.

[100] Workman CT, Stormo GD: ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. Proceedings of the Fifth Pacific Symposium on Biocomputing 2000, 467-478.

[101] Xing EP, Wu W, Jordan MI, Karp RM: Logos: a modular Bayesian model for de novo motif detection. J Bioinform Comput Biol 2004, 2:127-154.

8. ÖZGEÇMİŞ

Adres : Ceyhun Atuf Kansu Caddesi No:36/12
Balgat/ANKARA

Tel : *+90 312 234 10 10 - 2099*
E-posta : *kyildiz@baskent.edu.tr*

Ad ve Soyad: Kerem YILDIZ

Kişisel Bilgiler :

- 13 Mayıs 1982 / Ankara - TÜRKİYE
- Türkçe ana dil, İngilizce çok iyi
- Microsoft Office, Windows-Linux işletim sistemleri çok iyi
- Basic, Delphi, C, C++, HTML, PHP programlama dilleri çok iyi
- Lex/yacc ve flex/bison araçları çok iyi
- Java, C#, ASP, MATLAB iyi
- MySQL Server çok iyi
- MapInfo, MapBasic çok iyi
- Bekar, sağlıklı

Eğitim Durumu	:	Başkent Üniversitesi Bilgisayar Mühendisliği yüksek lisans programını **2009** yılında bitirdim	**Bağlıca Kampüsü, Ankara**
		Başkent Üniversitesi Bilgisayar Mühendisliği lisans programını **2005** yılında bitirdim	**Bağlıca Kampüsü, Ankara**
		Ömer Seyfettin Lisesi'nden **2000** yılında mezun oldum	**Balgat, Ankara**
		Özel İlkem Koleji'nden **1997** yılında mezun oldum	**Tandoğan, Ankara**
		Mustafa Kemal İlkokulundan **1993** yılında mezun oldum	**Balgat, Ankara**
Bitirme Projesi	:	Bitirme projesi çalışması kullanıcının girmiş olduğu program kodlarını yorumlama üzerinedir. Yorumlama işlemi kodun hata kontrolünü ve devamında eğer kodda hata yoksa çalıştırılması işlemlerini kapsamaktadır. Programlama kodları tipik programlama komutlarını içermektedir. (değişken tanımlama, atamalar, döngüler, if koşulları vb.) C++ Builder 6.0 ve Microsoft Access veritabanı kullanılarak bu proje geliştirilmiştir.	
İş Tecrübesi	:	Başkent Üniversitesi Bilgisayar Mühendisliği Bölümü'nde **2006** Şubat ayından bu yana araştırma	**Bağlıca Kampüsü, Ankara**

görevlisi olarak çalışmaktayım.

MapInfo Başar Ltd Sti'de **2005** yılında 2 ay süreliğine kendimi geliştirme amaçlı çalıştım. Haritada bulunan noktalar üzerinde yol çizen ve verilen mesafeye göre kamulaştırma işlemi yapan bir program geliştirdim. Bu noktalar kullanıcı tarafından belirlenmektedir. Yol çizildikten sonra, kullanıcının belirlemiş olduğu mesafeler içerisindeki alanlar kamulaştırılmıştır. Bu uygulama MapInfo ve MapBasic programları kullanılarak geliştirilmiştir. **Balgat, Ankara**

İkinci stajımı **2004** yılında 2 ay süreliğine Harbourfront Centre'da yaptım. Bu kuruluş kar amacı gütmemektedir. PHP programlama dili ve MySQL sunucusu kullanılarak olay raporlaması yapan bir web uygulaması geliştirilmiştir. Bu çalışma staj yapılan kuruluş ve mezun olduğum üniversite tarafından onaylanmıştır. **Toronto, KANADA**

İlk stajımı BOTAŞ Doğalgaz Müdürlüğü'nde **2003** yılında 2 ay süreliğine yaptım. Visual Basic 6.0 programlama dili ve Microsoft Access veritabanı kullanılarak bir misafirhane programı geliştirilmiştir. Bu çalışma staj yapılan kuruluş ve mezun olduğum üniversite tarafından onaylanmıştır. **Yapracık, Ankara**

İlgi Alanları : BiyoBilişim, Örüntü Tanıma, CBS
Bilgisayar Grafiği, İmge İşleme

Activiteler : Basketbol, masa tenisi, bilardo, koşu, bowling.

2001 yılında Başkent Üniversitesi Atatürkçü Düşünce Topluluğu ve Radyo Topluluğu üyeliği yaptım.

1997 yılında ortaokulda iken "Ah Şu Gençler" isimli tiyatroda oynadım.