KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

YENİ NESİL DİZİLEME YÖNTEMİYLE ELDE EDİLEN SEKANSLARDA OKUMA HATALARININ TESPİTİ VE DÜZELTİLMESİ

YÜKSEK LİSANS TEZİ

GÖKHAN DİLEK

MART 2021
TRABZON

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

YENİ NESİL DİZİLEME YÖNTEMİYLE ELDE EDİLEN SEKANSLARDA OKUMA HATALARININ TESPİTİ VE DÜZELTİLMESİ

Gökhan DİLEK

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsünce
"BİLGİSAYAR YÜKSEK MÜHENDİSİ"
Unvanı Verilmesi İçin Kabul Edilen Tezdir.

Tezin Enstitüye Verildiği Tarih : 04 / 03 /2021
Tezin Savunma Tarihi : 19 / 03 /2021

Tez Danışmanı : Dr. Öğr. Üyesi İbrahim SAVRAN

Trabzon 2021

ÖNSÖZ

Bu tez çalışmasında DNA dizileme yapılma işlemi sırasında oluşabilecek okuma hatalarının tespit edilmesi ve bu hataların giderilip daha doğru gen dizilimini elde etmek için hata düzeltme algoritması önerilmiştir. Önerilen veri okuma ve düzeltme yönteminin, genetiksel bozukluklara neden olan DNA mutasyonlarının tespitine ve çözümüne yardımcı olabilmesi amaçlanmaktadır. Bu çalışmanın biyoinformatik alanındaki çalışmalara katkı sağlaması beklenmektedir.

Bu çalışmanın tamamlanmasında, değerli zamanını bana ayırıp bilgisini ve tecrübesini benimle paylaşarak her konuda desteğini eksik etmeyen kıymetli danışman hocam Dr. Öğr. Üyesi İbrahim SAVRAN'a teşekkür ederim. Çalışma boyunca bilgisini paylaşan Bilgisayar Mühendisi Safa AKBULUT'a teşekkür ederim.

Gökhan DİLEK

Trabzon 2021

TEZ ETİK BEYANNAMESİ

Yüksek Lisans Tezi olarak sunduğum “Genetik Verilerde Okuma Hatalarını Düzeltmek için Yöntem” başlıklı bu çalışmayı baştan sona kadar danışmanım Dr. Öğr. Üyesi İbrahim SAVRAN’ın sorumluluğunda tamamladığımı, verileri kendim topladığımı, analizleri ilgili laboratuvarlarda yaptığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim. 19/03/2021

Gökhan DİLEK

İÇİNDEKİLER

Sayfa No

Yüksek Lisans

ÖZET

YENİ NESİL DİZİLEME YÖNTEMİYLE ELDE EDİLEN SEKANSLARDA OKUMA HATALARININ TESPİTİ VE DÜZELTİLMESİ

GÖKHAN DİLEK

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Dr. Öğr. Üyesi İbrahim SAVRAN
2021, 111 Sayfa, 6 Sayfa Ek

Biyoinformatik çalışmalar; genetik hastalık araştırmaları, hastalık tespiti ve tespit edilen hastalıklara çözüm üretilebilmesi amacıyla DNA dizilimleri yöntemleri üzerinde yoğunlaşmıştır. Genetik verilerin analizindeki temel problemler, bu veri dizilimlerinin boyutlarına bağlı karmaşıklıklarından meydana gelmektedir. Bu karmaşıklık, Yeni Nesil Dizileme (YND) cihazları tarafından yapılan veri okuma hatasından kaynaklanmaktadır. Büyük veri dizilimlerinin tek seferde okunması mümkün olmadığından, verilerin parçalar halinde dizileme yapılabilmektedir. YND cihazları yardımıyla büyük genetik verilerinin okunması mümkün kılınmıştır. Ayrıca YND cihazları, genetik veri dizilimlerinin okunması sırasında %1 ile %3 arasında hatalı okuma gerçekleştirmektedir. Bu çalışmada, günümüzde en yaygın kanser hastalıklarından biri olan Braf Mürin Sarcoma Viral Onkogen Homolog B1 (BRAF) gen mutasyonunun tespiti için sıkça rastlanan veri okuma hatasının düzeltilmesi üzerine bir yöntem önerilmiştir. Bu yöntemde, Ulusal Biyoteknoloji Bilgi Merkezi (NCBI) üzerinden paylaşılan sağlıklı BRAF geni kullanılmıştır. YND okumaları simüle edilerek oluşan hata oranları ile orantılı sentetik veri üretilmiştir. Hatalı gen, belirlenen derinlik boyutunda okunmuş ve fasta formatında kaydedilmiştir. Referans dizilim ile sekanslar karşılaştırılmış ve hatalar tespit edilerek düzeltme işlemleri uygulanmıştır.

Anahtar Kelimeler: Büyük Veriler, Biyoinformatik, Yeni Nesil Dizileme, DNA Dizileme(Sekanslama), K-mer yöntemi, Bloom Filter, Hash Fonksiyonları, BRAF Geni

Master Thesis

SUMMARY

DETECTION AND CORRECTION OF READ ERRORS IN SEQUENCES OBTAINED WITH THE NEW GENERATION SEQUENCE METHOD

GÖKHAN DİLEK

Karadeniz Technical University

The Graduate School of Natural and Applied Sciences
Computer Engineering Graduate Program
Supervisor: Asst. Prof. Dr. İbrahim SAVRAN
2021, 111 Pages, 6 Pages Appendix

Bioinformatics studies have focused on genetic disease research, disease detection and DNA sequencing methods in order to find solutions to detected diseases. The main problems in the analysis of genetic data arise from the complexity of these data sequences due to their size. This complexity is due to the data reading error made by Next Generation Sequencing (YND) devices. Since large data arrays cannot be read at once, data can be sequenced in chunks. With the help of YND devices, it has been made possible to read large genetic data. In addition, YND devices perform erroneous reading between 1% and 3% during the reading of genetic data sequences. In this study, a method for the detection of Braf Murin Sarcoma Viral Oncogene Homologous B1 (BRAF) gene mutation, which is one of the most common cancer diseases today, has been proposed. In this method, the healthy BRAF gene shared through the National Biotechnology Information Center (NCBI) was used. Synthetic data were produced proportional to the error rates by simulating YND readings. The faulty gene was read at the specified depth size and recorded in fasta format. The sequences were compared with the reference sequence, errors were detected and corrections were applied.

Key Words: Big Data, Bioinformatics, Next-Generation Sequencing, DNA Sequencing, K-mer method, Bloom Filter, Hash Functions, BRAF Gene

ŞEKİLLER DİZİNİ

Sayfa No

TABLOLAR DİZİNİ

Sayfa No

SEMBOLLER DİZİNİ

A	: Adenin
bp	: Baz Çifti
C	: Sitozin
DNA	: Deoksiribo Nükleik Asit
dsDNA	: Çift Zincirli DNA
E. coli	: Escherichia coli
FP	: False Positive
G	: Guanin
H-Bağı	: Hidrojen Bağı
INSDC	: Uluslararası Nükleotid Dizisi Veritabanı İşbirliği
NCBI	: National Centre for Biotecnology Information (Ulusal Biyoteknoloji Bilgi Merkezi)
RNA	: Ribo Nükleik Asit
ssDNA	: Tek Zincirli DNA
T	: Timin
tRNA	: Taşıyıcı Ribo Nükleik Asit
U	: Urasil
YND	: Yeni Nesil Dizileme

1. GENEL BİLGİLER

1.1. Giriş

Genetik ya da kalıtım bilimi, biyolojinin organizmalardaki kalıtım ve çeşitliliği inceleyen bir dalıdır. Türkçeye Almancadan geçen genetik sözcüğü 1831 yılında Yunanca genetikos("genitif") sözcüğünden türetilmiştir. Bu sözcüğün kökeni ise genesis("köken") sözcüğüne dayanmaktadır [1]. Genetik alanındaki çalışmaların temeli Gregor Mendel'in 1866 yılında bezelye bitkisi üzerinde yapmış olduğu melezleme deneyleri ile başlamıştır [2]. Mendel, 1856'da on yıl süreceği bir kalıtım türlerinin araştırması projesine başlamıştı. Araştırmalarına farelerle başlayıp daha sonrasında bitkilere ve bal arılarına geçmişti, en sonunda temel model organizma olarak bahçe bezelyelerine karar vermiştir [3]. Model organizmalar, araştırmacıların, belirli bilimsel sorunlar hakkında en rahat şekilde araştırmalar yapabileceği laboratuvarda üretilen organizmalardır. Araştırmacılar, model organizma üzerinde çalışmalar yaparak, insanlar üzerinde çalışma yapılması kısmen daha zor olan organizmalar veya biyolojik sistemler için de geçerli olan genel prensipleri saptayabilirler. Mendel, bezelyeler üzerindeki çalışmalarında, boy, tohum rengi, çiçek rengi ve tohum şekli gibi yedi farklı özellikte kalıtımları incelemiştir. Bu incelemede öncelikle uzun boy-kısa boy gibi farklı özelliğe sahip bezelye filizlerini sınıflandırmıştır. Bu filizleri, saf döl yani ebeveyne benzeyen yavru döl elde edilinceye kadar sürekli yetiştirmişti. Sonrasında bu saf dölleri çiftleştirerek özelliklerin nasıl kalıtıldığını incelemişti. Mendel, bitkilerin her kuşakta nasıl göründüklerine ek olarak da, her bir özelliği gösteren bitkileri tek tek saymıştı. Çarpıcı olarak, üzerinde çalışmış olduğu yedi özelliğin tümün de kalıtımsal olarak birbirine benzediğini fark etmişti [4]. Yapmış olduğu deneyler sonucunda da günümüzde kullanılmakta olan başlıca terimlerin de temeli atılmıştır. Mendel tarafından bezelye çalışmasında ki faktör olarak adlandırılan renk, boy ve şekil özellikleri günümüzde "gen" olarak adlandırılmaktadır [5].

Friedrich Miescher tarafından 1869 yılında hücrelerin çekirdeklerinde asidik özellikli yeni maddeler keşfetmiştir. Bu maddeler de günümüzde "nükleik asit" olarak adlandırılmaktadır. Friedrich v.a. tarafından 1871 yılında DNA'yı (‘nüklein’) şeklinde tanımlayan ilk yayınları basılmıştır. 1882 yılında Walther Flemming kromozomları

tanımlamıştır ve bu kromozomların hücre bölünmesi sırasında davranışlarını incelemiştir. 1884-1885 yılları arasında Oscar Hertwig v.a, hücrenin çekirdeğinin kalıtımın temelini içerdiğini bağımsız olarak kanıtlamıştır. Richard Altmann 1889 yılında "nüklein" i "nükleik asit" şeklinde yeniden adlandırmıştır. 1909 yılında Wilhelm Johannsen kalıtım birimlerinin tanımlanması için "gen" kelimesini kullanmıştır. 1928 yılında Frederick Griffith, yaptığı çalışmalar sonucunda "dönüşüm prensibi" olarak adlandırılan bir bakteri türünün özelliklerinin bir başka bakteriye aktarıldığını kabul etmiştir. 1929 yılında Phoebus Levene adenin (A), timin (T), guanin (G) ve sitozin (C) olmak üzere dört adet baz içeren DNA'nın yapı taşlarını tanımlamıştır. 1944 yılında, yaklaşık on yıllık özenli bir deneyden sonra Avery ve meslektaşları Maclyn McCarty ve Colin McLeod tarafından "dönüşüm ilkesi"nin deoksiribonükleik asitten (DNA) yapıldığını göstermişlerdi. Bu durum kalıtsal bilginin DNA ile birlikte taşındığını ve DNA'nın genetik materyal olarak işlev görebileceğini düşünüyorlardı. Bunun sonucunda DNA'nın bir parçası olan gen kavramı da daha çok anlam kazanmıştır. Colette v.a. 1949 yılında germ hücrelerinin çekirdeklerinin somatik hücrelerdeki DNA miktarının yarısını içerdiğini keşfetmişti. Bu durum eşey hücresi oluşumu durumunda kromozom sayısındaki azalmaya paraleldir. DNA'nın genetik materyal olması gerçeğini daha fazla kanıt sağlamaktadır. 1949-1950 yılları arasında Erwin Chargaff, DNA baz yapısının türler arasında değiştiğini tespit etmiştir, ancak bir tür içinde DNA'daki bazların her zaman sabit oranlarda bulunduğunu belirlemiştir: T'lerle aynı sayıda A ve G'lerle aynı sayıda C olarak ifade etmiştir [5].

1953 yılında James Watson ve Francis Crick, DNA'nın moleküler yapısını keşfetmiştir: C'nin her zaman G ile ve A'nın her zaman T ile eşleştiği bir çift sarmal yapısının olduğu bilgisini söylemiştir. Bu durum da nükleik asitler daha çok anlam kazanmış ve DNA dizilim sisteminin de temel yapısı belirlenmiştir. 1956 yılında Arthur Kornberg, DNA'yı kopyalayan enzim olan DNA polimerazı keşfetmişti. Francis Crick 1957 yılında, DNA'daki bilgilerin RNA yoluyla proteinlere çevrilmesini (central dogma) önermiştir. DNA'daki üçlü bazın her zaman protein içinde bir amino asit belirlediğini tahmin etmiştir [6].

1961-1966 yılları arasında Robert W. Holley ve meslektaşları 74 nükleotitten oluşan tRNA molekülünü keşfederek genetik kodu kırmayı başarmışlardı [7]. Bunun sonucunda DNA analizinin önemi artarak daha kolay analiz yapılabilmesi için DNA analiz yöntemleri keşfetmeye çalışılmıştır. Bu analiz yöntemlerinin amacında daha doğru gen dizilim genetik

kodunu bularak kalıtsal hastalık tespiti ve hastalığa bağlı olarak ilaç üretimi, kimlik tespit edilmesi gibi sorunlara çözüm üretilmesi amaçlanmaktadır.

1977 yılında Allan Maxam, Walter Gilbert ve Frederick Sanger, DNA dizileme yöntemleri geliştirmiştir [8]. Walter Gilbert ve Allan Maxam, birinci nesil DNA dizileme yöntemi olarak bilinen kimyasal kırılma yöntemi geliştirmiştir. Frederick Sanger ve meslektaşları, birinci nesil DNA dizileme yöntemi ile paralel olan daha uzun veri boyutunda DNA parçalarının dizilimini sağlayan Sanger yöntemini(zincir sonlandırma yöntemi) geliştirmiştir [9]. Sanger yöntemi, Maxam ve Gilbert yöntemine göre daha çok verimli olduğu ve daha az kimyasal madde ve radyoaktivite gereksiniminden dolayı kısa zamanda daha çok yaygınlaşmıştır [10]. Ancak Sanger yöntemi, bütün genomun veya metagenomun (mikrobiyal topluluğun “kolektif genomu”) sekanslanması(dizilemesi) gibi büyük ölçekli projeler için verimsiz ve pahalıdır [11]. Bu gibi görevler için ise daha ucuz ve daha hızlı olan yeni, büyük ölçekli sıralama teknikleri geliştirilmesi düşünülmüştür.

Yeni nesil dizileme (YND) olarak bilinen devasa paralel sekanslama teknolojisi biyolojik bilimlerde devrim yaratmıştır [12]. Ultra yüksek performansı, ölçeklenebilirliği ve daha hızlı olması ile YND, araştırmacıların çok çeşitli uygulamaları gerçekleştirmelerini ve biyolojik sistemleri daha önce hiç mümkün olmayan bir düzeyde çalışmalarını sağlamıştır.

1990 yılında, insan genetik eğitim setindeki üç milyar kimyasal ünitenin tümünü sıralamak ve tanımlamak, hastalığın genetik kökenlerini bulmak ve daha sonra tedaviler geliştirmek amacıyla "İnsan Genom Projesi" başlatılmıştır. Bu proje göre; insan genomunun yaklaşık 3.3 milyar baz çifti olduğu belirlenmiş ve proje bir "mega proje" olarak kabul edilmiştir [13]. Bunun sonucunda büyük verilerin analizinin yapılması problemine çözüm odaklı çalışmalar yapılmıştır. Çünkü geleneksel Sanger yöntemi, büyük verilerin dizileme analizi çok zaman almakla birlikte bu kadar büyük veriyi işlemekte de yetersiz kalmaktadır. İnsan genom projesinin tamamlanmasından sonra, yeni nesil dizileme (YND) teknolojileri olarak adlandırılan yüksek verimli ve hızlı dizileme platformlarının önemi çok fazla artmıştır. Böylece büyük verilerin analizi ile kanser ve diyabet gibi yaygın hastalık riskini artıran genetik varyantların tanımlanması ve tanımlanan hastalığa yönelik çözümlerin üretilmesinde yardımcı olması planlanmıştır.

1.2. Tez Çalışmasının Amacı ve Önemi

Yeni nesil dizileme(YND) yöntemler ile DNA diziliminin saklanabilmesi ve veri üzerinde elde edilen sonuçların analizi için cihazlar geliştirilmektedir. Geliştirilen bu cihazlarda DNA parçalarının analizleri çok daha kısa sürede gerçekleştirilmektedir. Fakat bu cihazlar, çok büyük veri olan DNA'yı parçalı şekilde işlem yaparak DNA dizilimi oluştururken rastgele bölümlerde nükleotidler eksik, yanlış veya fazla okumaktadır.

Bu tez çalışmasında, yeni nesil dizileme yöntemleri kullanılarak analizi yapılan verinin üzerinde oluşan okuma hatalarının giderilmesi için bir yöntem önerilmiştir. Bu yöntemin aşamaları:

I. Analizi yapılacak DNA verisinin belirlenmesi
II. DNA dizileme işlemlerinin analizi
III. Dizileme işleminin sekanslara ayrılması
IV. Benzersiz alt diziler olan k-mer yönteminin uygulanması
V. Bloom Filter yapısının boyut ve sayısının belirlenmesi
VI. Uygunluk değerine göre Hash Fonksiyonlarının uygulanması
VII. Kullanılacak verinin belirlenen boyuttaki Bloom Filter üzerine kayıt işlemin yapılması
VIII. Sekansların k-mer yöntemi ile tekrarlı okuma işlemi yapılması için derinlik belirlenmesi
IX. K-mer yöntemine göre ayrılan sekans verilerinin belirlenen derinlikte tekrarlı okunması ile birlikte Bloom Filter üzerine kayıt işleminin yapılması
X. Sekanslama işlemi sonrasında her bir sekansın filtre içerisinde sorgulamasının yapılması ve derinlik ile orantılı şekilde uygun görülen değer doğrultusunda filtre üzerinden kontrol analizinin yapılması
XI. Yapılan analiz sonucunda hata tespiti ve düzeltilmesi işleminin gerçekleştirilmesi

Bu tez kapsamında, önerilen yöntemin DNA üzerinde kalıcı hasar oluşturabilen mutasyonların tespitinde daha doğru sonuçlar üretilebilmesine yardımcı olunacağı ve bu konu üzerinde çalışma yapacaklara da yol göstereceği umulmaktadır.

1.3. Temel Kavramlar

1.3.1. Genetik ve Kalıtımın Kısa Tarihi

Genetik, saç rengi, göz rengi ve hastalık riski gibi özelliklerin ebeveynlerden çocuklarına nasıl aktarıldığı kalıtsal bir çalışmadır. Canlının genetik bilgileri "genetik kod" veya “genom” olarak adlandırılmaktadır. Genomlar deoksiribonükleik asit (DNA) adında bir kimyasaldan oluşmaktadır ve vücut içinde hemen hemen her hücrede saklanılmaktadır.

Genetiğin kökenleri evrim teorilerinin gelişimine dayanmaktadır. 1858 yılından sonra Charles Darwin ve Wallace'ın araştırma çalışmasından sonra türlerin kökeni ve tür değişkenliğinin nasıl geliştirildiği anlaşılmıştı. 1866'da “Modern Genetiğin Babası” olarak adlandırılan Gregor Johann Mendel, bezelye içindeki özelliklerin mirasına ilişkin gözlemlerini tanımladığı “Bitki Hibridizasyonunda Deneyler” adlı makalesini yayınlamıştır [14]. Kalıtımın temellerini üç temel miras kanunu ile önermiştir:

- Eşleştirilmiş kalıtsal belirleyicilerin gamet oluşumu sırasında eşit olasılıkta ayrıldığını belirten ayrışma yasası;
- Baskın ve resesif gen kavramını tanıtan egemenlik yasası;
- Tüm kalıtım faktörlerinin birbirinden bağımsız çalıştığını öne süren bağımsız çeşitlilik yasası.

Carl Correns, Hugo de Vries ve Erich von Tschermak'ın Mendel’in çeşitli türler üzerindeki gözlemlerini bağımsız şekilde yeniden keşfettiği ve doğruladığı durumu 1900 yılına kadar unutulmuştur. Hugo de Vries, farklı özellikleri “pangenes” in mirasından sorumlu birimler seçmiştir. Günümüzde "gen" olarak kısaltılmıştır [14]. 1875 yılında Eduard Strasburger, Walther Flemming ve Edouard van Beneden, daha sonra W. von Waldeyer-Hartz tarafından “kromozom” olarak adlandırılan bir hücresel maddeyi tanımlamışlardı, çünkü bu materyal bazofilik anilin boyalarını güçlü bir şekilde emdiğini göstermiştir [15]. Bu gözlem Mendel’in ayrışma yasası ve kalıtsallığın temeli olarak

kromozomlar kurarak bağımsız çeşitlilik yasası için moleküler açıklamalar sağlanmıştır [16]. Daha sonra, Thomas Morgan ve Alfred Sturtevant, genlerin kromozom üzerinde doğrusal bir düzende olduğunu belirten ve Drosophila'daki ilk doğrusal gen haritasını oluşturduklarını belirten gen teorisini önermiştir [17]. 1928 yılında Frederick Griffith, fareler üzerinde kalıtımın temelini incelemek için farklı patojenik özelliklere sahip iki farklı Streptococcus pneumoniae suşu (R ve S) kullanmıştır. Ölümcül S türü, deneğin bağışıklık sistemine karşı koruyan ve pnömoniye neden olan bir polisakarit kaplama üreterek “pürüzsüz” koloniler oluşturmuştur. Patojenik olmayan R suşu, bu kaplamayı üretme yeteneğine sahip değildir ve bunun yerine, virüssüz bakterilerin "pürüzlü" bir kolonisini oluşturmuştur. Griffith bir fareye ısı ile öldürülen bir S suşu veya canlı bir R suşu enjekte ettiğinde, fare pnömoni geliştirmemiştir, ancak bu iki suşun bir kombinasyonu enjekte ettiğinde ise fare enfekte olmuştur ve ölmüştür. Sonuç olarak, Griffith, ısı ile öldürülen S suşundan canlı R suşuna, bir polisakkarit kaplamanın üretimine ve daha sonra ölümcül pnömoninin gelişmesine yol açan bir bileşen transferi (Griffith tarafından "dönüştürme prensibi" olarak adlandırılmış olan) olduğunu ileri sürmüştür [18]. Kalıtsal materyalin bakterilerde DNA, RNA veya protein olup olmadığı net değildir. Mükemmel nitelikteki in vitro(laboratuvar ortamında hazırlanmış) enzim sindirim deneyi olarak kullanan Oswald Avery, Maclyn McCarty ve Colin MacLeod, Griffith’in bakteriyel dönüşüm prensibinin proteaz ve ribonükleaz ile sindirildiğinde patojenik şekilde olduğunu ve deoksiribonükleaz ile sindirildiğinde aktivitenin kaybolduğunu göstermiştir [19]. Bu durum üzerine bakterilerde kalıtsal materyalin DNA olduğuna kanaat getirmişlerdir. Ayrıca, 1952'de Alfred Hershey ve Martha Chase, memeli hayvanların kalın bağırsağında yaşayan bakteri türlerinden biri olarak adlandırılan Escherichia coli'yi (E. coli) enfekte eden bakteriyofaj T2 virüsü (faj) ile izotopik etiketleme deneyleri kullanarak DNA'nın bir virüs içinde kalıtsal materyal olduğunu göstermiştir [20]. Fajın sülfür açısından zengin protein ve fosfor açısından zengin DNA içerdiği bilinmektedir. Bu nedenle, protein veya DNA'nın kalıtsal materyal olup olmadığını belirlemek için faj ve E. coli'yi hem "32P" ortamında hem de "35S" ortamında ayrı olarak kültürlenmiştir. Döl fajları, ilgili izotoplarla etiketlendikten sonra, yeni E. coli ile inkübe edilmişti, burada etiketli fajlar, kalıtım materyallerini çoğaltma için E. coli'ye enjekte edilerek etiketsiz E. coli'yi enfekte edilmektedir. Daha sonra enfekte olmuş E. coli'nin dışında bırakılan eski faj katları, E. coli içindeki kalıtsal materyalden ayırt etmek için bir mutfak karıştırıcısı kullanılarak ortamdan ayrıştırılmıştır. Son olarak, yeni döl fajları E. coli'den ayrıldığında, analiz edildi ve "32P"

ortam kültüründen kaynaklanan fajın, orijinal "32P"'nin %30'undan fazlası korunurken, "35S" ortam kültüründen kaynaklanan fajın orijinal etiketlemenin %1'inden daha azını muhafaza etmektedir. Sonuç olarak, DNA'nın bakteriyofaj T2 virüsü için Avery ve ark.'nın yapmış oldukları keşifi destekleyen kalıtsal materyal olduğu sonucu çıkarılmıştır. Bu keşifler ayrıca DNA'nın evrensel kalıtsal materyal olabileceğini ileri sürülmüştür [14,19]. Devam eden çalışmalar sonrasında da karşılaştırmalı anatomiden onkolojiye kadar uzanan disiplinlere genetik kavramlar, genetik prensipler ve genetik metodolojiler aşılanmaktadır.

Yirmi birinci yüzyılın ilk yirmi yılında bu eğilim sadece genişletilmiştir. Bu nedenle, bilim adamlarının, bilimin nasıl çalıştığını, dünya hakkında ne söylediğini ve genlere çok fazla odaklanan bir toplumda yaşayan insanlar üzerinde ne gibi etkileri olduğunu anlamak için genetiğe daha çok yönelimi olmaktadır. Yirmi birinci yüzyıl biliminde genetik, gebelik öncesinde laboratuvar ortamında oluşturulan embriyolar üzerinde yapılan bir genetik tarama testi olan preimplantasyon genetik teşhisi(PGT) gibi üreme tarama teknolojilerinde, hangi türlerin tehlikede olduğunun değerlendirilmesinde, antibiyotiğe dirençli bakterileri, kalıtımsal bozuklar gibi birçok hastalığa yön göstermek için birçok araştırmalara konu olmaktadır. Bu süreci anlamak ve potansiyel olarak kontrol etmek için de bir nesilden diğerine özellik aktarımının kalıplarına ve mekanizmalarına odaklanılması gerekmektedir. Bu durum genetiğin benzersiz bir çalışma alanı olarak şekillendirilmektedir [21]. Bilim adamları genetik çalışmaları ile bireyden alınan genetik bilgiler yardımıyla birçok hastalığı teşhis etmek, tedavi etmek, önlemek ve iyileştirmek için kullanmayı amaçlanmaktadır.

1.3.2. DNA Araştırmalarının Kısa Tarihi

Bilim adamları, kalıtsal maddi kimlik arayışının yanı sıra DNA araştırmalarını ilerletmek için inanılmaz ilerleme kaydedilmiştir. 1869 yılında Friedrich Miescher, yakınında bulunan bir cerrahi kliniğinden topladığı kirli bandajlar üzerinde bulunan irin içinde beyaz kan hücrelerinden "nüklein" adını vermiş olduğu asidik materyali izole edilmiştir. O dönemde yaygın olarak kullanılan bir bileşim analizi tekniği olan yanma deneyleri kullanılarak, nükleinin protein ve diğer organik moleküllerde yaygın olarak bulunan karbon, hidrojen, azot ve oksijen elementlerinden oluştuğunu keşfedilmiştir.

Çalışma sonucuna göre, protein içinde önemli bir bileşen olmayan, ancak proteinde yaygın olan kükürt bulunmayan fosfor bulunmuştur. Miescher ayrıca, asit/baz ilavesi ile proteaz sindirimi ve çözünürlük testleri ile nükleinin bilinen protein türlerinden veya bilinen herhangi bir molekülden farklı bir madde kategorisi olduğuna çalışmalar sonucunda ikna olmuştur [14,22].

Miescher'in çalışmalarının ardından Albrecht Kossel, nükleik asidin yapı taşlarını bazlar, şekerler ve fosforik asitler olarak tanımlanmıştır. Julius Bodo Unger ile birlikte hidroliz ve yanma deneylerini kullanarak dört farklı DNA bazını keşfedilmiştir. Örnek kaynaklarına göre: guanin("guano" olarak bilinen deniz kuşlarının dışkısından) [22], adenin(öküz pankreasından keşfedilmiştir, Yunancada "aden-" önekinden gelen), timin(buzağı timusundan olduğu gibi) [23] ve sitozin (hücrelerden olduğu gibi "sito-" öneki) [24] olarak azot içeren bileşikler olarak adlandırılmıştır.

1929'da Phoebus Levene, çeşitli kimyasal deneyler tarafından belirlenen moleküler formüle dayanan bir "tetranükleotid" yapısı önerilmiştir; burada dört farklı DNA nükleotidi, bilimsel toplum tarafından yaygın olarak kabul edildiğinde sıralanmaktadır [25].

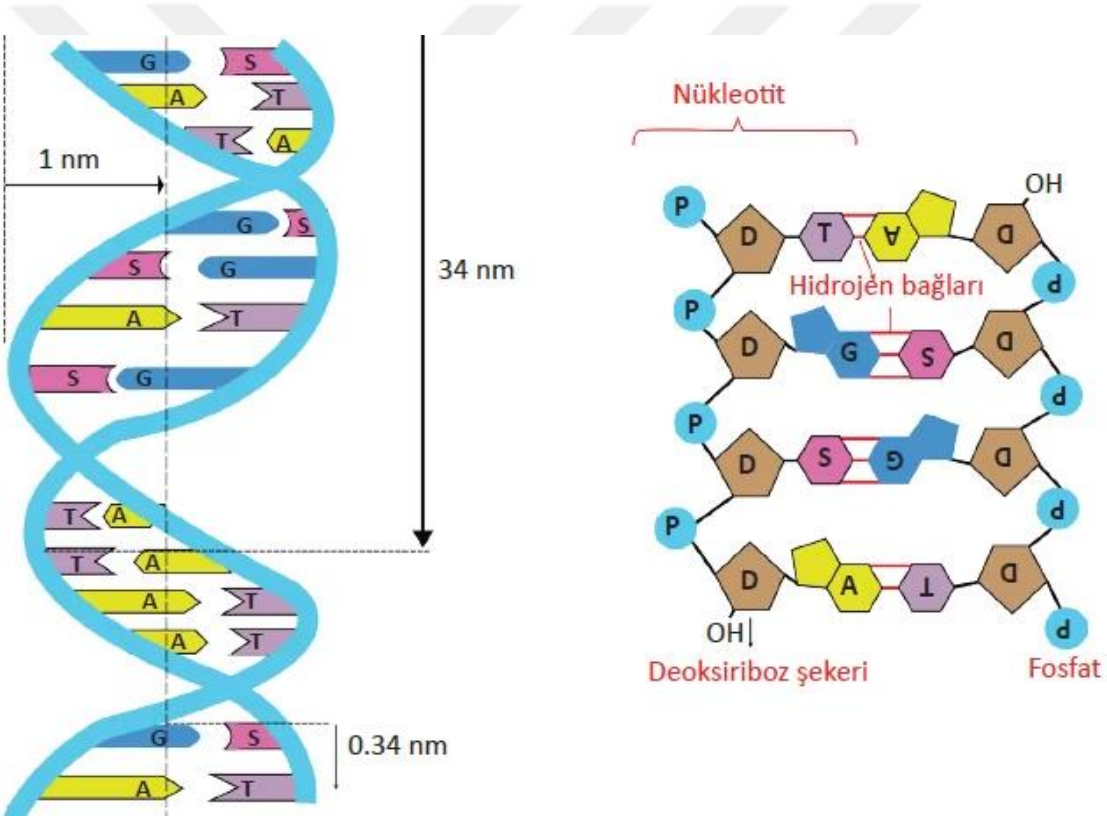
Colette Vendrely ve Roger, akıl hocası olan Andre Boivin ile birlikte DNA yapısının kalıtsal materyal olduğu önermesine daha fazla destek sağlamışlardır. Aynı canlının tüm somatik hücreleri eşit miktarda DNA bulundurduğunu ve gametogenez esnasında daha önce gözlemlenmiş olan kromozom eylemi ile uyum sağlayarak, sperm hücreleri çekirdeğinde bulunan DNA miktarının iki katını içerdiğini kanıtlamışlardır [26].

1950'de Erwin Chargaff, Levene'nin "tetranükleotid" hipotezini reddetmişti ve adenin (A) timin (T) ve sitozin (C) guanin (G) miktarlarının her zaman kağıt kromatografisi kullanılarak eşit miktarlarda meydana geldiğini ancak ATGC'nin nispi miktarı birçok türde 1:1:1:1 olmadığını göstermektedir. Bu gözlem sayesinde, doğru DNA yapısının çıkarılmasının yolunu açmaktadır. Ayrıca Chargaff DNA'nın C+G içeriği farklı türlere göre %22 ila 73 arasında değişebildiğini, fakat aynı organizmada tüm hücrelerde sabit kaldığını gözlemlemiştir [27]. 1952 yılında Rosalind Franklin tarafından elde edilmiş olan B-formu DNA'nın biçimli X-ışını kristalografi sonuçları üzerine Cambridge Üniversitesi'nden Francis Crick ve James Watson, DNA yapısının şifrelerini çözmek için önceki çalışmalardan önemli kanıtları birleştirmişlerdir [28]. Çalışma sonrası 1953'te Eagle

Pub'da ünlü DNA çift sarmal konformasyonunu önerilmiştir. Bu önerilen modelde, A ile T baz çifti, C ile G baz çifti hidrojen bağı olsa da eşit olduğunu göstermiştir, bu da Chargaff'ın eşit miktarda A: T ve C: G gözlemi için iyi bir açıklama sağlamaktadır. Watson ve Crick'in 1953'teki Nature makalelerinde doğru bir şekilde belirttiği gibi: "Yayınladığımız spesifik eşleştirmenin genetik materyal için olası bir kopyalama mekanizmasını önerdiği dikkatimizden kaçmamıştır" şeklinde açıklamışlardır [6, 28]. O yıllardan beri, birçok DNA yapısı farklı şartlar altında X-ışını kristalografisi ile çözülebilmektedir. Çift katlı DNA (dsDNA) için en yaygın yapı, nispeten düşük tuz ve yüksek nem koşulları altında oluşmakta olan sağ dönüşlü B-formu şeklinde belirtilmektedir. Yapı içerisinde bazlar, geniş bir ana oluk (tabanların kenarı) ve dar bir küçük oluk (fosfat omurgasının ve şekerlerin kenarı) ile sarmal eksene dik olacak şekilde ifade edilmektedir. Her bir sarmal dönüş için 10.5 baz çifti (bp) oluşmaktadır [29]. DNA kristalleri yüksek tuz ve dehidrasyon koşullarında büyütüldüğünde, A-formu DNA gözlenebilmektedir [30]. Bu DNA sarmal dönüş başına 11 bp ile sarmal eksene göre eğimli tabanlarla sağ el dönüşlü olarak kullanılmıştır. B-DNA ile karşılaştırıldığında, A-DNA'nın ana oluğu derin ve dar olarak belirtilmiş, küçük oluk sığ ve geniş şekilde gösterilmiştir. DNA'nın fizyolojik koşullar altında yaygın olarak oluşamadığı düşünülse de A-formu için, şeker üzerindeki 2'-hidroksil grubuna bağlı olacak şekilde çoğu çift zincirli RNA (dsRNA) konformasyonu olarak açıklanmıştır. 1979'da Alexander Rich ve meslektaşları, belirli sekans bağlamında veya belirli aşırı koşullar altında (3 M MgCl2 veya NaCl gibi veya alkol ilavesi ile) DNA'nın aşırı sığ bir ana oluk ve her turda 12 bp ile çok derin ve dar bir küçük oluk ile sola dönüşlü bir konformasyon alabilmesini açıklamışlardır. Fosfat omurgaları A- ve B-form DNA'sında bulunan düz çizgiler yerine zikzak çizgiler içerdiğinden, bu yeni DNA konformasyonuna Z-DNA adı verilmektedir [31]. Oluşum için zorlu koşullar nedeniyle, birkaç protein ailesinin bu yapıya yüksek özgüllüğü olduğu tespit edilene ve biyolojik rolüne işaret eden Z-DNA'nın biyolojik önemi şüphe edilmektedir [32].

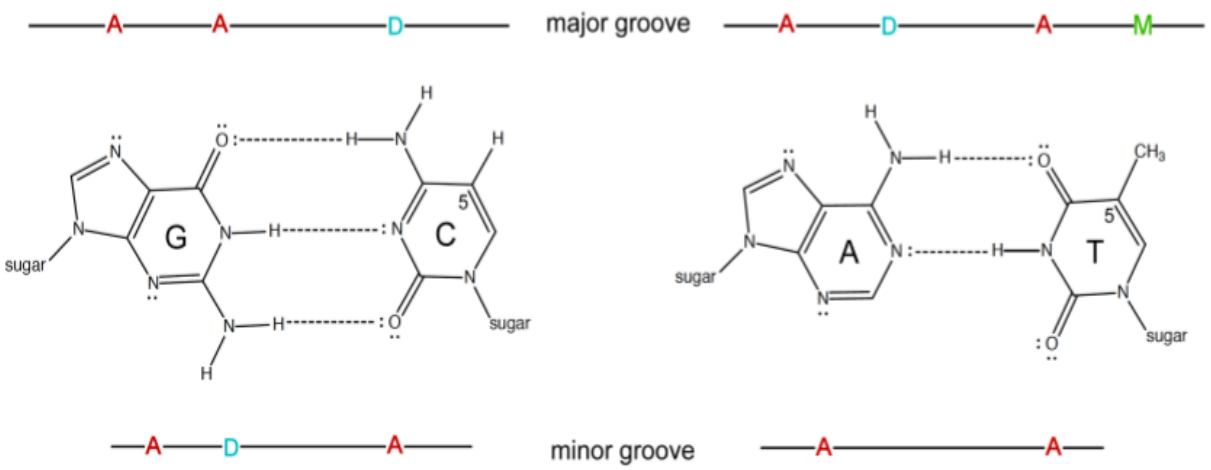
1.3.2.1. Protein Tanıma İçin DNA Yapısal Temelleri

Hücrelerdeki genomik DNA, çift sarmal yapı şeklinde bulunmaktadır ve genetik bilgilerin aktarılması ve korunması için birincil substrat görevi görmektedir. DNA içine gömülü olan bilgilerin, hücresel faaliyetleri düzenlemek için proteinler tarafından okuma işlemi yapılması gerekmektedir. Özel protein tanıma olayı, dolaylı ve doğrudan okumaların kombinasyonu şeklinde gerçekleştirilmiştir. Şekil 1.1'de DNA molekülü sarmal yapısı gösterilmektedir [33-35].



Şekil 1.1. DNA molekülünün çift(ikili) sarmal yapısı

Doğrudan protein okuması, hidrojen bağlanması (H-bond) ve Van Der Waals'ın protein yan zincirleri arasındaki etkileşimleri ve esas olarak DNA baz çiftlerinin ana oluğu ile belirtilen yapı tamamlayıcılığı ile elde edilmektedir.



Şekil 1.2. DNA'nın hidrojen bağı verici/alıcı modeli

Şekil 1.2 DNA'nın büyük ve küçük oluklarındaki hidrojen bağı verici / alıcı modelini göstermektedir. H-bağı donörleri mavi “D”, H-bağı alıcıları kırmızı “A” ile, hidrofobik metil grubu yeşil “M” ile işaretlenmiştir.

Proteinler, B-DNA'nın ana oluğuna bir "tanıma" alfa sarmalı ekleyerek DNA ana oluğuna doğru işaret edilen DNA fonksiyonel grupları tarafından sergilenen benzersiz H-bağı verici / alıcı modelini sorgulayabilme işlemi yapabilmektedirler [36]. Şekil 1.2'de gösterildiği gibi, H-bağı verici / alıcı modeli, baz çiftlerinin kimliği ve yönlülüğü ile değişmektedir. T'nin -CH3 grubu, Van Der Waals etkileşimlerine katkıda bulunan hidrofobik bir gruptan oluşmaktadır; sitozin -C4 ve adenin -C6'nın amin gruplarının hidrojenleri H-bağı donörleri(verici) olarak gösterilmiştir(Şekil 1.2'de D); N7 ve guaninin -C6'sının karbonil oksijeni, adenin N7'si ve timinin -C4'ünün karbonil oksijeni H-bağı alıcıları olarak gösterilmiştir (Şekil 1.2'deki A). H-bağı modeli her iki yönde bulunan her baz çifti için benzersiz şekilde gösterilmektedir. Küçük oluk desenleri daha az değişkenliktedir. Sadece G:C / C:G baz çiftlerini A:T / T:A baz çiftinden H'nin bağlanan N2'sinden H-bağı donörü(verici) şeklinde ayrılmaktadır. Üst üste bindirilmiş kristal yapılar, G:C / C:G ve A:T / T:A'nın yönlülüğü, bu iki baz çifti için iki yönlü H-bağı alıcılarının neredeyse aynı konumları sebebiyle ayırt edilemediğini göstermektedir [37].

Dolaylı şekilde protein tanıma özgüllüğü, bağlayıcılık için gerekli doğru konformasyonun varsayılması kolaylığıyla elde edilebilmektedir. Bu, DNA dizisi bağlamı

tarafından temelde modüle edilen dsDNA'nın DNA esnekliğine ve yapısına (oluk genişliği, baz çifti büküm, vb.) bağlı olarak ifade edilmektedir.

1.3.3. RNA Araştırmalarının Kısa Tarihi

1868 yılında Nükleik asitler Friedrich Miescher tarafından keşfedilmiştir [5, 38]. Daha sonra çekirdeği olmayan prokaryotik hücrelerin de nükleik asitler içerdiği keşfedilmiştir. RNA'nın protein sentezindeki görevinin 1939 yılında olduğu şüphelenilmiştir [39]. Severo Ochoa, laboratuvarda RNA sentezi yapabilen bir enzim keşfinden sonra Arthur Kornberg ile paylaşmış olduğu 1959 Nobel Tıp Ödülü'nü kazanmıştır [40]. Sonraki zamanlarda, Ochoa tarafından keşfedilen bu enzimin RNA sentezinden değil RNA bozulması yapabildiği gösterilmiştir.

1965 yılında Robert W. Holley tarafından bir maya taşıyıcı RNA(tRNA)'sının 77 nükleotitinin dizilimi bulunmuştur [41]. Bundan dolayı Holley, Marshall Nirenberg ve Har Gobind Khorana ile birlikte 1968 Nobel Tıp Ödülü'nü kazanmıştır.

1967'de Carl Woese, RNA'nın katalitik olabileceğini ileri sürmüştü ve ilk yaşam formlarının (kendi kendini kopyalayan moleküller) hem genetik bilgi taşımak hem de biyokimyasal reaksiyonları katalize etme görevini RNA'nın yapabileceğini öne sürmüştür [42].

1970'lerin başlarında ters transkriptaz ve Retro virüsler keşfedilmiştir. İlk defa enzimlerin RNA'yı DNA'ya kopyalayabildiğini, genetik bilgilerin iletimi için genel rotanın tersine olduğu gösterilmiştir. Bu çalışma ile David Baltimore, Howard Temin ve Renato Dulbecco 1975 yılında Nobel Ödülü'ne layık görülmüştür. Walter Fiers ve ekibi 1976 yılında, bir RNA virüs genomunda bakteriyofaj MS2'nin ilk tam nükleotit dizilimini belirlemişlerdir [43].

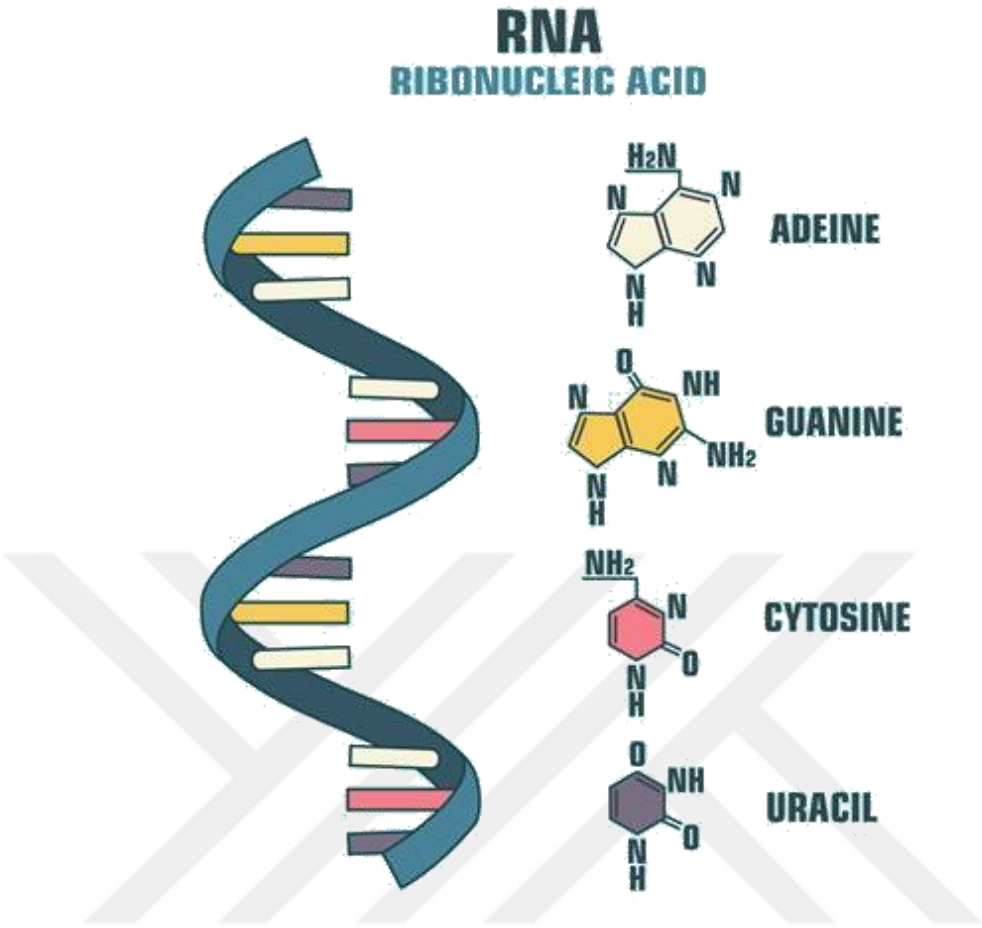
1977 yılında hem hücresel genlerde hem de memeli virüslerinde RNA birleştime ve intronlar keşfedilmiştir, bunun sonucunda Philip Sharp ve Richard Roberts'a 1993 yılında nobel ödülü verilmiştir. 1980'lerin başında katalitik RNA molekülleri(ribozimler)

keşfedilmiştir ve bunun üzerine 1989 yılında Thomas Cech ve Sidney Altman'a Nobel ödülü verilmiştir. 1990 yılında Petunya‘da, tanıtılan genlerin, şu anda RNA müdahalesinin bir sonucu olduğu bilinen, bitkinin kendisinin benzer genlerini susturduğu bulunmuştur [44, 45].

RNA, bilinen bütün yaşam formlarında gerekli olan üç ana makro-moleküllerden (protein ve DNA ile birlikte) biri olarak bilinmektedir. RNA'nın kimyasal yapısı DNA'ya çok benzemektedir, ancak iki temel farkı bulunmaktadır:

a) RNA şeker ribozu içerirken DNA farklı şeker olan deoksiribozu (bir oksijen atomu olmayan bir tür riboz) içermektedir.
b) RNA nükleobaz urasil içerirken DNA timin içermektedir (urasil ve timin benzer baz-eşleştirme özelliklerine sahiptir).

Mesajcı RNA (mRNA), hücrelerden protein sentezi bölgeleri bulunan DNA'dan ribozoma bilgi taşıyan RNA çeşididir. mRNA'nın kodlama yapısı üretilen proteindeki amino asit dizilimini belirlemektedir. Bununla birlikte, birçok RNA proteini kodlamamaktadır. Bu durum transkripsiyonel çıktının yaklaşık olarak %97'si ökaryotlarda protein olmayan kodlamalar içermektedir. Bu kodlanamayan RNA'lar (ncRNA) kendi RNA genleri tarafından kodlanılabilmektedir, fakat mRNA intronlarından da türetilebilmektedir. Kodlanamayan RNA'ların en belirgin örnekleri, her ikisi de çevirim işleminde yer alan transfer RNA (tRNA) ve ribozomal RNA (rRNA) şeklindedir. RNA işleme, gen düzenleme ve diğer görevlerde rol alan kodlayıcı olmayan RNA'lar da bulunmaktadır. Bazı RNA'lar, diğer RNA moleküllerinin bağlanması ve kesilmesi gibi kimyasal reaksiyonları ve ribozomdaki peptit bağları oluşumunu katalize edebilmektedirler; bunlar ribozimler olarak da adlandırılmaktadır [46]. Şekil 1.3’te RNA molekülünün sarmal yapısı gösterilmektedir [47].



Şekil 1.3. RNA molekülünün sarmal yapısı

1.3.4. DNA ve RNA Karşılaştırması

DNA ve RNA, bütün yaşamın temelini oluşturan genetik bilgilerin depolanması ve okunması işleminden sorumlu olan hücre biyolojisinin en önemli molekülleri olarak bilinmektedir. Her ikisi de bazlar, şekerler ve fosfatlardan oluşan doğrusal polimerlerdir, fakat ikisini ayıran bazı önemli farklılıklar bulunmaktadır. Bu ayırıcı durumlar, iki molekülün birlikte çalışabilmesini ve temel görevlerini yerine getirebilmesini sağlamaktadır. Tablo 1.1'de DNA ve RNA karşılaştırılması gösterilmektedir [48].

Tablo 1.1. DNA ve RNA karşılaştırılması

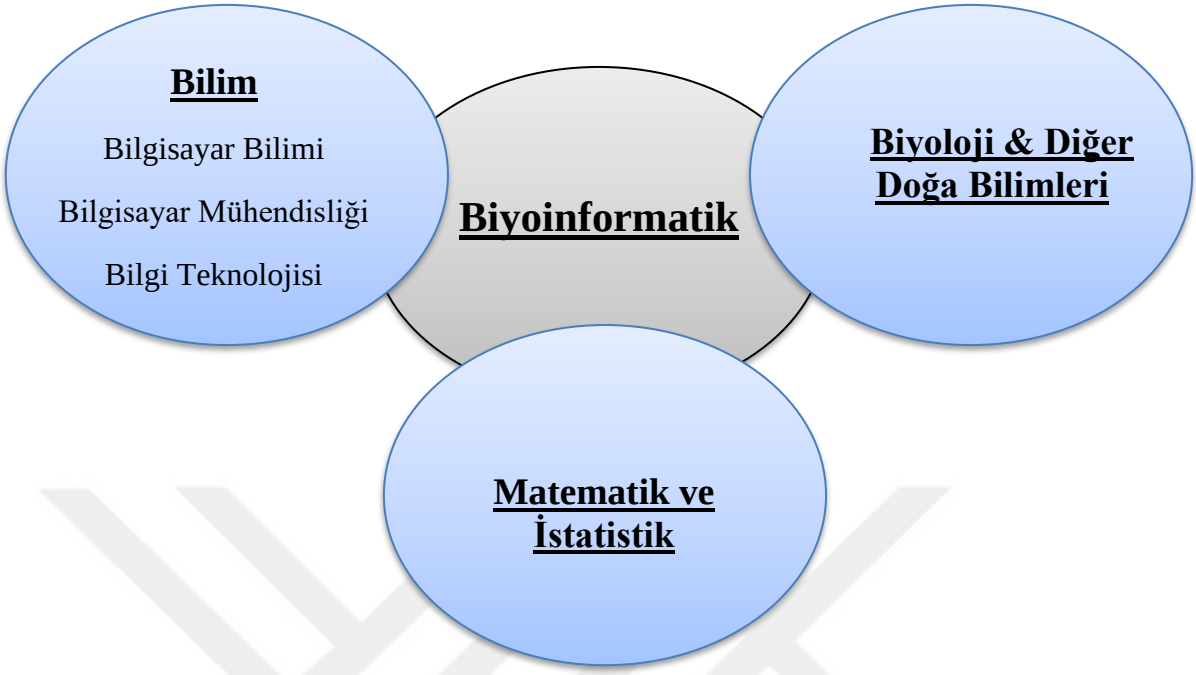
KARŞILAŞTIRMA	DNA	RNA
Fonksiyonlar	Bir organizmada bulunan bütün genetik bilgiler için bir taslaktır. Genetik bilgileri depolanır ve çoğaltılır.	DNA içinde bulunan genetik bilgileri, proteinlerin oluşması için kullanılabilir bir formata dönüştürür, daha sonra ribozomal protein bölümüne taşınılır.
Yapı	Çift sarmal olarak düzenlenmiş iki iplikten oluşmaktadır. Bu iplikler nükleotit olarak adlandırılan alt birimlerden oluşmaktadır. Her nükleotid için bir 5-karbon şeker molekülü, bir fosfat ve bir azotlu baz bulunmaktadır.	Sadece tek ipliği bulunmaktadır, fakat DNA gibi nükleotitlerden oluşmaktadır. DNA dizilerine göre RNA dizileri daha kısadır. RNA bazen ikincil bir çift sarmal yapı oluşturur, ancak sadece aralıklı olarak.
Uzunluk	RNA'ya göre çok daha uzun bir polimerdir. Örnek olarak; bir kromozom, çözüldüğünde birkaç santimetre uzunluğunda olabilecek tek, uzun bir DNA molekülü şeklindedir.	RNA moleküllerinin uzunluğu değişkenlidir, fakat uzun DNA polimerlerine göre çok daha kısadır. Büyük RNA molekülleri sadece birkaç bin baz çifti uzunluğunda olabilmektedir.
Şeker	DNA'daki şeker grubu, RNA'nın ribozundan bir tane daha az hidroksil grubu bulunan deoksiribozdur.	RNA, deoksiribozun hidroksil modifikasyonları bulunmayan riboz şeker molekülleri içermektedir.
Baz	DNA'daki bazlar; Timin("T"), Adenin("A"), Sitosin ("C") ve Guanindir("G").	RNA'daki bazlar; Adenin ("A"), Urasil ("U"), Guanin ("G") ve Sitosindir ("C").
Baz Çiftleri	Adenin ve Timin çifti (A-T) Sitosin ve Guanine çifti (C-G)	Adenin ve Urasil çifti (A-U) Sitosin ve Guanine çifti (C-G)

Tablo 1.1’in devamı

KARŞILAŞTIRMA	DNA	RNA
Konum	Çekirdekte bulunmaktadır, mitokondride az miktarda DNA da bulunmaktadır.	Nükleolusta oluşmaktadır. Oluşan RNA tipine göre sitoplazmanın özel bölgelerine doğru hareket etmektedir.
Reaktivite	Daha az oksijen bulunan hidroksil grubu içeren deoksiriboz şekeri sebebiyle DNA, genetik bilgiyi güvende tutma adına bir molekül için yararlı olan RNA'ya göre daha kararlı bir moleküle sahiptir.	Riboz şekeri içeren RNA, DNA'ya göre daha reaktiftir. Alkalin koşullarında kararlı değildir. RNA’nın daha büyük sarmal kanalları, enzimlerin saldırısına daha kolay maruz kaldığı anlamına gelmektedir.
Ultraviyole (UV) Hassasiyeti	Ultraviyole ışıklarından kaynaklanan hasarlara karşı savunmasızdır.	UV ışığından kaynaklanan hasara DNA'dan daha dirençlidir.

1.3.5. Biyoinformatik

Biyoinformatik, birçok açıdan biyolojinin diğer dallarından farklı olmayan büyüleyici ve yenilikçi bir bilim olarak nitelendirilmektedir. Ulusal Biyoteknoloji Bilgi Merkezi (NCBI), biyoinformatiği; bilgisayar bilimi, biyoloji ve bilgi teknolojisi olarak Şekil 1.4’te gösterildiği gibi tek bir disiplin sisteminde birleştiği bilim alanı olarak tanımlamıştır. Alanın nihai amacı, yeni biyolojik kavrayışların yanı sıra biyolojide birleştirici ilkelerin ayırt edilebileceği küresel bir bakış açısı oluşturmaktır. Özellikle veri setlerinin büyük ve karmaşık olması biyolojik verilerin anlaşılması için yazılım araçları ve yöntemler geliştiren bir alandır. Biyolojik verilerin analiz edilmesi ve yorumlanması için biyoloji, bilgi mühendisliği, bilgisayar bilimi, matematik ve istatistikleri birleştirmektedir [49].



Şekil 1.4. Perspektifte biyoinformatik

Biyoinformatik geliştiricileri matematikçiler, istatistikçiler, bilgisayar bilimcileri, yazılım mühendisleri arasında değişim gösterebilmektedir. Biyoinformatik veri kullanıcıları, son yirmi yılda oluşturulan geniş gen veri tabanlarını ayıklamak için geliştiriciler tarafından oluşturulan araçları kullanarak belirli problemlerin çözümlerinde yararlı bilgileri elde etmek için hem endüstriden hem de akademik çevreden bilim adamlarıdır. Biyoinformatik verileri, pratik olarak sınırsız sayıda görev için kullanılmaktadır. Örnek olarak [49, 50];

(1) Bir filogeni boyunca sekansın evriminin incelenmesi, genetik sapma yoluyla hangi bölümlerin değiştiğini (genellikle bu bölgeler "önemsiz DNA" veya vasat fonksiyonel öneme sahiptir) ve doğal seleksiyonla değiştirilen veya sürdürülen (bu bölgeler genel olarak yüksek fonksiyonel öneme sahiptir) çalışmalar,
(2) Homolog olup olmadıklarını belirlemek için yeni bir sekansın bilinen sekanslarla karşılaştırılması,
(3) Evrimin büyük ölçekli modellerini incelemek için tüm genomların karşılaştırılması,

(4) Eksprese Edilen Sekans Etiketleri (EST) sekanslarından genlerin yeniden yapılandırılması; EST, klonlanan ve dizilenen - daha sonra genel gen veri tabanlarına bırakılan eksprese edilen kısa gen parçalarıdır. Bir 'sıfır veritabanı' verildiğinde veya başlamak için hiçbir bilgi verilmediğinde, hangi dizilerin birbirine uyması gerektiğini öngörmek için örtüşme bölgeleri kullanılarak tüm tamamlayıcı DNA (cDNA) moleküllerinin yeniden yapılandırılması mümkündür. Ayrıca, farklı bir türden bilinen bir sekans göz önüne alındığında, EST'ler bitişik veya bitişik sekansın oluşturulmasına izin vermek için üst üste binen uçları ile bu şablona hizalanabilmektedir.

(5) Proteinlerin ailelere gruplanması; Genler tarafından kodlanan proteinleri süper aileler ve aileler olarak sınıflandırmak için çok fazla çalışmalar yapılmaktadır. Protein sekansları giderek artan sayıda yöntem kullanılarak karşılaştırılabilir, bunların ilişkisi ölçülebilir ve yakından ilişkili proteinlere aileler atanabilir olabilmektedir. Gruplama "Gen Bankası" 'ndaki bir sekansın diğerleriyle etkili bir karşılaştırmasını içerse de, bu durum (2)'nin bir uzantısıdır. Gen Bankasındaki dizi sayısı çok fazla olduğundan, bu uzun zaman almaktadır, bu nedenle grupları daha az sayıda diziye göre modellemek için akıllı yöntemler kullanılmaktadır. Protein Aileleri (Pfam), proteinleri ailelere gruplandırmak ve bundan elde edilen bilgileri hem kullanışlı hem de erişilebilir hale getirmek için matematik kullanımının son derece iyi bir örneğidir.

1.3.5.1 Biyoinformatik Verilerin Yapısı

Biyoinformatik alanında kullanan iki önemli büyük ölçekli faaliyet Genomik ve Proteomiktir. Genomik, genomların analiz edilmesini ifade etmektedir. Bir genom, nesilden nesile aktarılan kalıtsal materyali kodlayan DNA dizilerinin tamamı olarak düşünülebilmektedir. DNA, iki konjuge ipliğin çift sarmalı olarak bulunmaktadır ve (A, C, G, T) dört harften oluşan tek boyutlu sembolik bir dizi olarak ele alınmaktadır. Bu DNA dizileri, genomun içindeki tüm genleri (ebeveynden yavrulara geçirilen fiziksel kalıtım birimi ve fonksiyonel) ve transkriptleri (genetik bilginin kodunun çözülmesi için ilk adım olan RNA kopyalarını) içermektedir. Bundan dolayı genomik, bir organizmadaki

transkriptler ve genler de dahil olmak üzere bütün genomik varlıkların analizini ve dizilimi şeklinde ifade edilmektedir. Ayrıca Proteomik, bütün protein veya proteom setinin analizini göstermektedir [51].

Aşağıda ki örnekte Büyük Benekli Kiwi (Apteryx Haastii)'nin bir DNA parçası gösterilmiştir;

... GA TCCTGACCA TGAACCTAAGCTTCTTCGACCAATT...

Bir proteinin amino asit sekansı, [A (Adenin), G (Guanine), C (Sitosin), T (Timin)] Tablo 1.2'de gösterildiği gibi baz üçlüsü kullanılarak kodlanılmaktadır. Bir nükleotit sekansının şifresinin çözülmesi, sekansın üçlü (yani 3 nükleotit bloğuna) şekilde birleştirilmiş ve daha sonra aşağıda gösterilen tablodan bakılmasını gerektirmektedir.

Tablo 1.2. Evrensel genetik kod

	T	**C**	**G**	**A**	
T	Phe	Ser	Tyr	Cys	**T**
	Phe	Ser	Tyr	Cys	**C**
	Leu	Ser	STOP	STOP	**A**
	Leu	Ser	STOP	Trp	**G**
C	Leu	Pro	His	Arg	**T**
	Leu	Pro	His	Arg	**C**
	Leu	Pro	Gin	Arg	**A**
	Leu	Pro	Gin	Arg	**G**
A	lle	Thr	Asn	Ser	**T**
	lle	Thr	Asn	Ser	**C**
	lle	Thr	Lys	Arg	**A**
	Met	Thr	Lys	Arg	**G**
G	Val	Ala	Asp	Gly	**T**
	Val	Ala	Asp	Gly	**C**
	Val	Ala	Glu	Gly	**A**
	Val	Ala	Glu	Gly	**G**

Tekrarlamanın başlayacağı üç farklı okuma karesi vardır ve böylece iki Yön amino asitlerin altı aday dizisi Ek Tablo 1'de gösterilmiştir [49, 51].

Ek Tablo 1'de birinci yön: "SOL" 'dan elde edilen diziler;

- Birinci Dizi = (Asp, Leu, Asp, His, Glu, Pro, Lys, Leu, Leu, Arg, Pro, Ile)
- İkinci Dizi = (Ile, Leu, Thr, Met, Asn, Leu, Ser, Phe, Phe, Asp, Gln)
- Üçüncü Dizi = (Ser, STOP, Pro, STOP, Thr, STOP, Ala, Ser, Ser, Thr, Asn)

İkinci yön: "SAĞ" 'dan elde edilen diziler;

- Dördüncü Dizi = (Leu, Thr, Ser, Phe, Phe, Glu, Ser, Lys, Tyr, Gln, Ser, STOP)
- Beşinci Dizi = (STOP, Pro, Ala, Ser, Ser, Asn, Pro, Ser, Thr, Ser, Pro)
- Altıncı Dizi = (Asn, Gln, Leu, Leu, Arg, Ile, Gln, Val, Pro, Val, Leu)

Bir DNA sekansı, aynı nükleotit bileşenine sahip olanlar da bile belirli bir türü tanımlamaktadır. Bu nükleotit yapısı, en temel yapı seviyesidir. Bundan dolayı yapının daha yüksek seviyelerini de tanımlamaktadır. Bir DNA sekanslama işleminden sonraki adım, Tablo 1.3'te gösterildiği gibi içinde farklı fonksiyonel bölgeler bulmaktadır [52].

Bir protein dizisi yirmi farklı Amino Asit türünden oluşmaktadır. Amino asitler proteinlerin temel yapısal birimleridir. Bir alfa-amino asit, bir amino grubu, bir hidrojen atomu, bir karboksil grubu, ve bir karbon atomuna bağlı şekilde, R-karboksil grubuna bitişik olduğu için alfa-karbon adında ayırt edici bir R grubundan oluşmaktadır [52].

Tablo 1.3. Amino asit kodları

Amino Asit	**3-Harf Kodu**	**1-Harf Kodu**	**Amino Asit**	**3-Harf Kodu**	**1-Harf Kodu**
Alanin	Ala	A	**Lösin**	Leu	L
Arjinin	Arg	R	**Lizin**	Lys	K

Tablo 1.3'ün devamı

Amino Asit	3-Harf Kodu	1-Harf Kodu	Amino Asit	3-Harf Kodu	1-Harf Kodu
Asparajin	Asn	N	**Metiyonin**	Met	M
Aspartat	Asp	D	**Fenilalanin**	Phe	F
Sistein	Cys	C	**Prolin**	Pro	P
Glutamin	Gln	Q	**Serin**	Ser	S
Glutamat	Glu	E	**Treonin**	Thr	T
Glisin	Gly	G	**Triptofan**	Trp	W
Histidin	His	H	**Tirozin**	Tyr	Y
İzolösin	Ile	I	**Valin**	Val	V

Tipik bir veritabanı yönetim sistemi, tablolarda depolanan verileri sorgulamak ve almak için güçlü mekanizmalara sahiptir, ancak yukarıdaki örnekteki gibi (büyük benekli kivi DNA yapısı) büyük dizelerde arama yapmak için ideal değildir. Bu, biyoinformatiklerin veri işleme için kendi özel araçlarına ihtiyaç duyduğu anlamına gelmektedir.

1.3.5.2 Biyoinformatik Araştırmaları İçin Yazılım Araçları

Çok fazla mevcut veri sebebiyle, günümüzde biyomedikal verilerin bilgisayardan alınabilmesi ve analizi çok önemli hale gelmiştir. Biyoinformatik için araştırmayı kolaylaştıran yazılım araçları genellikle;

(1) Veri alma araçları,

(2) Sekans karşılaştırma araçları,
(3) Model keşif araçları ve
(4) Görselleştirme araçları şeklinde dört kategoriye ayrılabilir.

1.3.5.2.1 Veri Alma Araçları

Veri alımı için ana araç olan Entrez sistemi kullanılmaktadır. Bu sistem, NCBI tarafından geliştirilen, literatür, nükleotid ve protein dizileri, tam genomlar ve ilgili veriler dahil olmak üzere çok çeşitli veri alanlarına entegre erişim sağlayan entegre bir veri erişim sistemi olarak nitelendirilmiştir. Her Entrez Gen kaydında, belirli bir gen ve organizma için çeşitli bilgiler bulunmaktadır. Mümkün olduğunda, bilgiler dizi verileri üzerinde yapılan analizlerin sonuçlarını içermektedir. Sunulan bilgi türü ve miktarı, belirli bir gen ve organizma için neye mevcut olduğuna bağlıdır ve ayrıca şunları içerebilmektedir [53]:

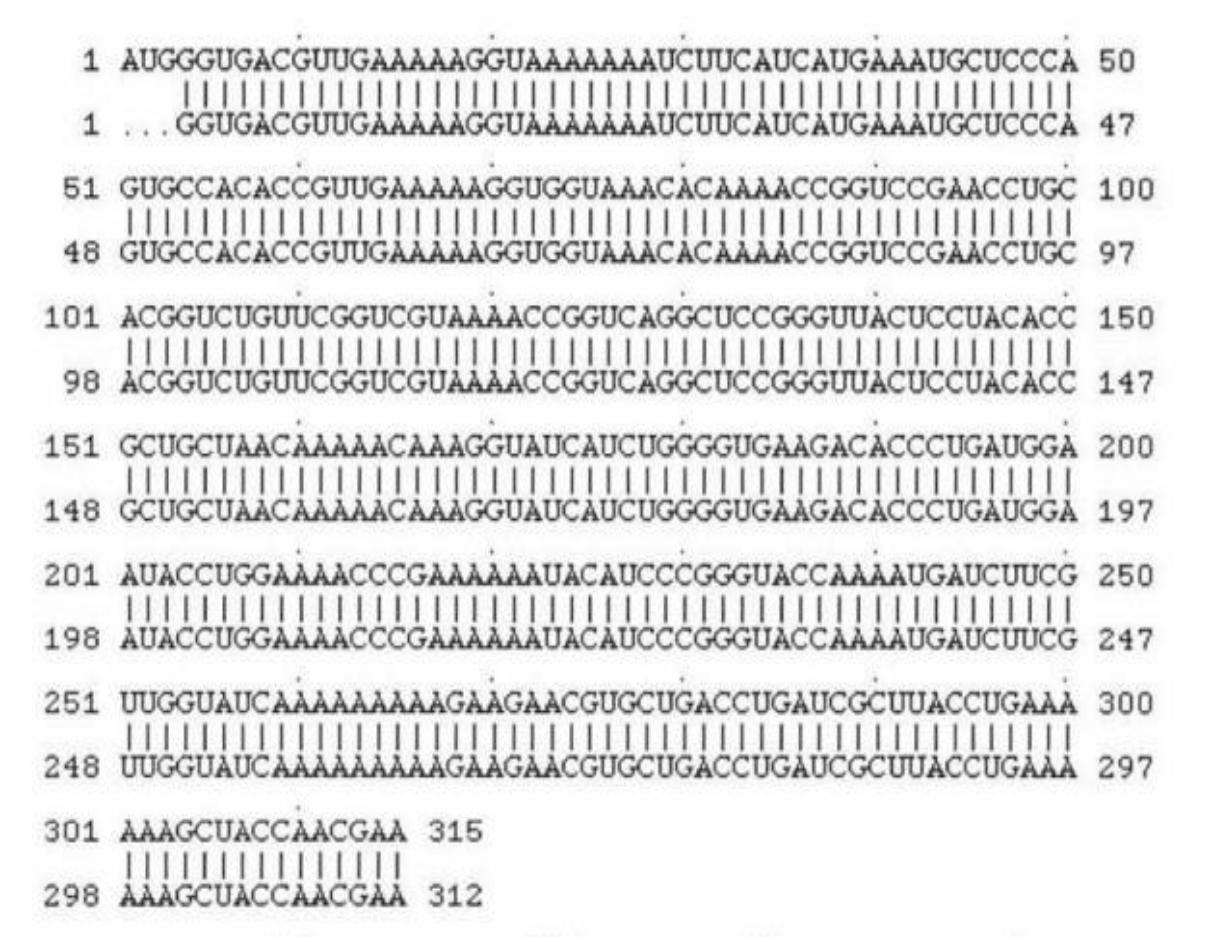
(1) Genomik bağlamın, intron / ekon yapısının ve yan genlerin grafik özeti,
(2) mRNA sekansının, kodlama dizisi(CDS) ve Tek nükleotid polimorfizmleri(SNP) gibi biyolojik özellikleri gösteren grafik görünümüne bağlantı oluşumu,
(3) Gen ontolojisi ve fenotipik bilgilere bağlantı oluşumu,
(4) Karşılık gelen protein sekansı verilerine ve korunmuş alanlara bağlantılar,
(5) Mutasyon veri tabanları gibi ilgili kaynaklara bağlantılar.

1.3.5.2.2 Sekans Karşılaştırma Araçları

Temel Yerel Hizalama Arama Aracı (BLAST) ve HIZLI Hizalama (FASTA), yaygın olarak kullanılan dizi karşılaştırma ve hizalama araçlarıdır. BLAST, gen ve protein dizilerini genel veri tabanlarındaki diğer kişilerle karşılaştırmak için kullanılır ve Konuma Özgü Yinelemeli BLAST(PSI-BLAST), Desen Vuruşlu başlatılan BLAST(PHI-BLAST) ve BLAST 2 sekansları dahil olmak üzere çeşitli türlerde kullanılmaktadır. Özel BLAST'lar ayrıca insan, sıtma, mikrobiyal ve diğer genomların yanı sıra

antikorlar(immünoglobulinler), vektör kontaminasyonu ve geçici insan konsensüs dizilimleri için de mevcut olabilmektedir. Genel özniteliklerde kullanılan FASTA formatı, hızlı bir protein veya nükleotit karşılaştırılması için kullanılabilmektedir. Program, bir yer değişme matrisi kullanarak yerel hizalamalar için optimize edilmiş aramalar gerçekleştirerek yüksek hız elde etmektedir [54].

Şekil 1.5'teki harfler, DNA'nın yapı taşları olan amino asitleri temsil etmektedir [49]. Diziler, vücudun oksijeni nasıl kullandığı ile ilgili bir gen olan Sitokrom C'yi temsil etmektedir. Üst küme dizisi şempanze amino asitleridir, alt küme dizisi de insanlara aittir. Her " | " şembolü bir amino asit eşleşmesini temsil etmektedir. Bu durumda, Şempanze ve İnsan'ın Sitokrom C genleriyle %100 eşleşmeleri bulunmaktadır.



Şekil 1.5. Şempanze ve insana özel gen karşılaştırması

1.3.5.2.3 Model Bulma Araçları

Verilerdeki modelleri veya özellikleri aramak için kullanılan araçlardır. Küme Analizi bu amaç için en yaygın araç olarak kullanılmaktadır. Belirli bir veri kümesindeki gruplamaları bulmak için kullanılmaktadır, böylece aynı gruptaki nesneler farklı gruplardaki nesneler birbirine benzeme durumu ortadan kalkmış olur. Dizi analizinde model keşif araçları da önemlidir. Spesifik alt sekansların, yapıları ve işlevsel siteleri bulmak için gelişmiş matematiksel modellemeler ve istatistiksel yöntemler kullanılmaktadır [55].

1.3.5.2.4 Görselleştirme Araçları

Bu araç, genomik verilerin etkileşimli bir grafiksel gösterimini sağlamaktadır. Expression Profiler ve GeneQuiz bunlara örnektir ve bunlar bir görselleştirme aracı içermektedir. Çoğu görselleştirme aracı internette ücretsiz olarak mevcuttur; bunlara örnek olarak TreeView, BioViews ve Protein Explorer gösterilebilmektedir. Şu anda biyoinformatikte devam eden gelişmelerden kaynaklanan daha geniş veri setlerini etkileşimli olarak analiz etmek için kullanılabilecek araçlar geliştirmek için çok fazla çalışmalar bulunmaktadır [56]. Bu konu kendi başına araştırma açısından çok zengindir.

1.3.5.3 Biyoinformatik Verilerle İlgili Zorluklar

Genomik ve proteomik alanlarında kullanılan veri tabanlarının büyüklüğü, karmaşıklığı ve sayısının yanı sıra yapı ve diğer çalışmalar, endüstrinin bu sorunları ele

alma ihtiyacına uyum sağlamada geri kalmalarına neden olmuştur. Hem bilgisayar teknolojisi hem de biyoloji, standardizasyon gibi umut verici çözümler oluşturmak için yeni ortak ebeveynlik teknikleri ve biyoinformatik olarak daha iyi iletişim kurmayı öğretme aracı olarak geliştirilmiştir. Hücre biyolojisi ve moleküler için veri yönetimi, veri modelleme, geleneksel veri üretimi ve kazanımı, veri analizi ve entegrasyonu alanlarını içermektedir. Endüstri alanında, son birkaç yıl için ana odak noktası, özellikle gen ifadesi ve DNA dizisi verileri için yüksek verimli veri analizini destekleyen yöntem ve teknolojilerin geliştirilmesi şeklinde olmuştur. Yeni teknolojik platformlar biyolojik veri üretebilmek için, büyük miktarlarda deneysel veriler için yakalama, düzenleme, yorumlama ve depolama gereksinimlerinden kaynaklanan veri yönetimi zorluklarını sunmaktadır. Platformlar, daha yüksek yoğunluklu diziler ve mikro diziler için daha iyi prob seçimi gibi teknolojik gelişmelerden faydalanan yeni sürümlerle gelişmeye devam etmektedir. Bu evrim, hem bu verilerin entegre edilmesi hem de analiz edilmesi gerektiğinde karşılaşılan, aynı platformun farklı sürümleri kullanılarak oluşturulan potansiyel olarak uyumsuz verilerin toplanması sorununu gündeme getirmektedir. Diğer zorluklar için, doğası gereği kesin olmayan teknikler ve araçlar kullanılarak oluşturulmuş verilerin nitelendirilmesini ve çeşitli, zayıf korelasyonlu depolarda bulunan verilerin entegre edilmesi olayı yüksek yer almaktadır. Deneysel veriler için veri tutarsızlığı veya belirsizliği ile baş etmek, veri yönetiminden ziyade istatistiksel yöntemlerin gerekli olduğu nitelendirilmiştir; örnek olarak, istatistiksel yöntemleri çeşitli ayrıntılı düzeyde gen ifadesi veri analizlerine uyarlama yöntemleri son yıllarda yoğun bir şekilde araştırma ve geliştirme üzerine olmuştur. Veri semantiği alanında, örneğin bir ifade tahmini değeri ve bunların ilişkileri, özellikle sürekli değişen platformlar ve gelişen biyolojik bilgi bağlamında uygun niteliklere sahip veri değerleri alanında en zor problemlerle karşılaşılmıştır. Veri üretiminden veriyi toplama ve entegrasyon etme, verinin analizine kadar bütün veri yönetimi alanlarında bunun gibi sorunlarla karşılaşılsa da, çözümlerin etki alanına özgü bilgiler ve kapsamlı veri tanımlama ve iyileştirme çalışmaları gerektirmektedir; veri yönetiminde ise bu sorunları giderebilmek için yalnızca çerçeve örneği kontrollü dizim ve ontolojileri sağlanmaktadır [57].

Bir endüstri alanında, veri yönetimi zorlukları için çözümlerin maliyet, karmaşıklık, performans, sağlamlık ve ürüne özel gereksinimler şeklinde dikkate alınması gerekmektedir. Biyolojik veri yönetimi sorunlarına etkili çözümler tasarlamak için

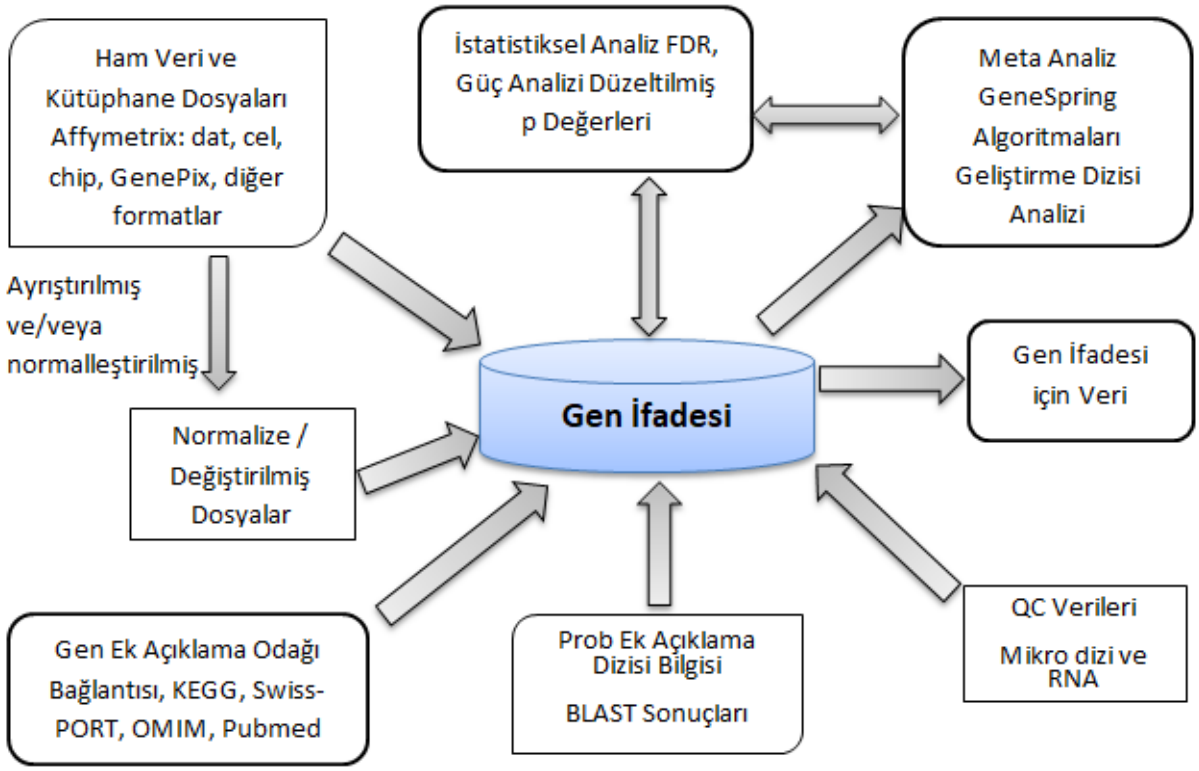
biyolojik uygulamanın, veri yönetimi alanının ve sorunların dikkate alındığı genel bağlamın ayrıntılı bir şekilde anlaşılması gerekmektedir [58].

1.3.5.3.1 Biyoinformatik Problemler

Entegre bilişimi, ilgili etki alanları arası analiz sonuçlarını ve arka plan bilgilerini düzenlemek için şemaların dinamik olarak oluşturulmasıdır. Bunu belli sorunlar üzerinde gerçekleştirmeye çalışırken zorluklar oluşmaktadır. Bazı problemler şunlardır [59]:

- Çeşitli uyumsuz kaynaklardan çok fazla entegre olmayan veri,
- Her biri özel bir tarama ve sorgulama mekanizmasına sahip standart bir adlandırma kuralı yoktur (ortak arabirim yok),
- Diğer veri kaynakları ile zayıf etkileşim.

Belirli bir veri kümesinin değeri, veri analizlerinin, en azından arayüz düzeyinde, harici olarak geliştirilmiş, yapılandırılmış bilgilerle birleştirilmesiyle büyük ölçüde artırılabilir. Klasik olarak, entegrasyonu gerçekleştirmenin birkaç yöntemi vardır [59]:



Şekil 1.6. Veri entegrasyonu

Şekil 1.6'da veri entegrasyonu özellikleri [49, 59]:

1. Bağlantılı, dizine alınmış veri sistemleri: Düz dosya veritabanlarını web [HTIP] bağlantıları ve dizinleri kullanarak bağlanması;
2. Serbest entegrasyon: Verileri ortak bir şema olmadan ancak ortak bir sorgu mekanizmasına sahip çoklu veritabanı sistemine sistematize edilmesi;
3. Görünümlerle sıkı entegrasyon: Ortak bir şema ve merkezi bir sorgu mekanizması ile bir veritabanı ilişkilendirmesi ile organize edilmesi;
4. Gerçekleştirilmiş verilerle serbest entegrasyon: Ortak bir şema olmadan bir veri ambarında düzenlenir ve tüm verileri düzenli olarak merkezi bir konuma yüklenilir.
5. Gerçekleştirilmiş verilerle sıkı entegrasyon: Ortak bir şemaya sahip bir veri ambarı oluşturulur ve tüm veriler periyodik olarak merkezi bir konuma yüklenilir.

Bu yöntemler, uygulama karmaşıklıkları, ölçeklenebilirlikleri ve sürdürülebilirlikleri bakımından önemli ölçüde farklılık göstermektedirler. Tam fiziksel ve mantıksal ayrımdan, fiziksel olarak tutarsız veri kümelerinin mantıksal birliğine ve son olarak entegre bir bütüne uzanan bir bant boyunca uzanmaktadırlar.

Uzun vadeli görünümde veri entegrasyonu, fonksiyonel ve kullanışlı biyoinformatik sistemler geliştirmek için önemli bir unsur olarak nitelendirilmektedir. Verileri ve sonuçları arşivlemek ve yayınlamak çok önemli olsa da, veri türü ile ilgili veri tabanlarının yanı sıra çok sayıda önemli veri türü de ilham vericidir. Ham veriler, her türlü nükleik asit, protein ve varsayımsal protein sekansları, bunlar arasındaki benzerlik sonuçlarını, sayısız kaynaktan fonksiyonel verileri ve çeşitli protokolleri, genetik ve fiziksel haritaları, üreme verilerini, ilişkili çevresel bilgileri ve fenotipik bilgileri içermektedir. Raporlama sisteminin doğasında bulunan karmaşıklığı azaltmak için hesaplama araçları kullanılarak, raporlama yönteminin ayrıntılarına odaklanmak yerine gerçekten ilgi çekici olan sistemin boyutsallığına ve karmaşıklığına odaklanılması sağlanmaktadır [49, 59].

Bu kapsamda biyoinformatik, sadece mevcut durumu değil, aynı zamanda gelecekteki ihtiyaç ve gelişmeleri tartışmak için ortak bir temel geliştirmek amacıyla sunulması önemli olmuştur.

1.3.5.4. Biyoinformatik Veri Bankaları

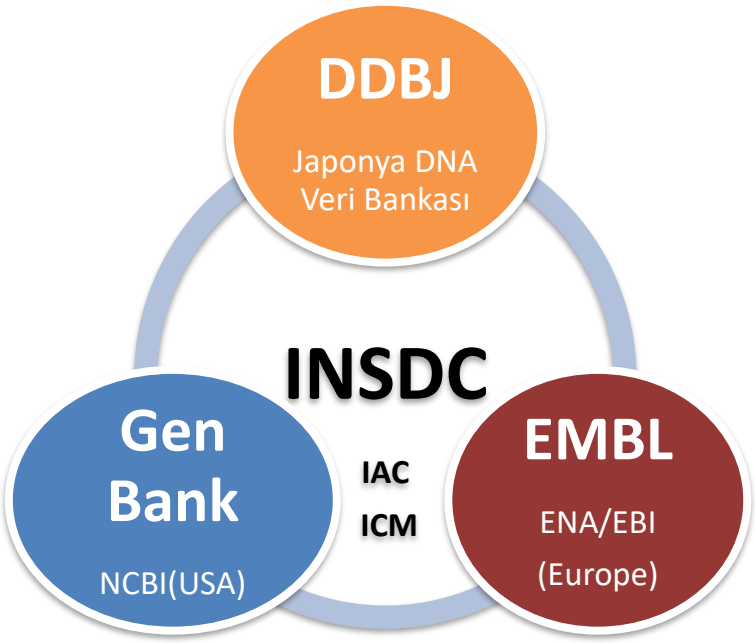
İlk biyoinformatik (biyolojik) veri tabanları, ilk protein dizilerinin piyasaya sürülmesinden birkaç yıl sonra oluşturulmuştur. Bildirilen ilk protein dizisi, 1956 yılında 51 kalıntıdan oluşan büyükbaş insülin dizisidir. Yaklaşık on yıl sonra, yetmiş yedi bazlı maya alanin tRNA'sına ait ilk nükleik asit sekansı rapor edilmişti. Dayhoff bir yıl sonra, ilk biyoinformatik veritabanını oluşturabilmek için tüm dizi verilerini toplamıştır [60].

Protein Veri Bankası, 1972'de 10 X-ışını kristalografik protein yapısının bir koleksiyonuyla izlemişti ve 1987'de SWISS PR OT protein dizisi veritabanı başlamıştı. Farklı türde ve boyutlarda çok çeşitli ıraksak veri kaynakları artık kamuya açık domain alanlarda veya ticari üçüncü taraflar için kullanımı mevcuttur. Tüm orijinal veritabanları,

veri girişlerinin giriş başına bir tane veya tek bir büyük metin dosyası olarak düz dosyalarda saklanmasıyla çok basit bir şekilde düzenlenmiştir. Başlık bilgilerinin uygun anahtar kelime aramasına olanak tanımak için daha sonra yeniden yazma dizinleri yeniden eklenmişti. Farklı veri tabanlarındaki biyoinformatik verileri, çoğunlukla kıyaslama ve kalite kontrol mekanizmalarının ilkel olması nedeniyle, genellikle gereksiz görülmüştür [54, 61].

Uluslararası Nükleotid Dizisi Veritabanı İşbirliği'nin (INSDC) bilinen üç iyi veri tabanları EMBL Bank, GenBank ve DDBJ olarak bilinmektedir. Bu üç kuruluşun veri gönderimi için ayrı siteleri bulunmaktadır ve günlük olarak veri alışverişi yapılmaktadır. Bu, stil ve ek açıklama biçimindeki bir değişikliğe rağmen, üç veri tabanındaki dizi bilgisinin herhangi bir zamanda aynı olduğunu göstermektedir. Temel sorumluluğu üç ayrı veri tabanının mevcut veri paylaşım politikasını onaylamak ve teyit etmek olan Uluslararası Danışma Komitesi'ne sahiptir [61].

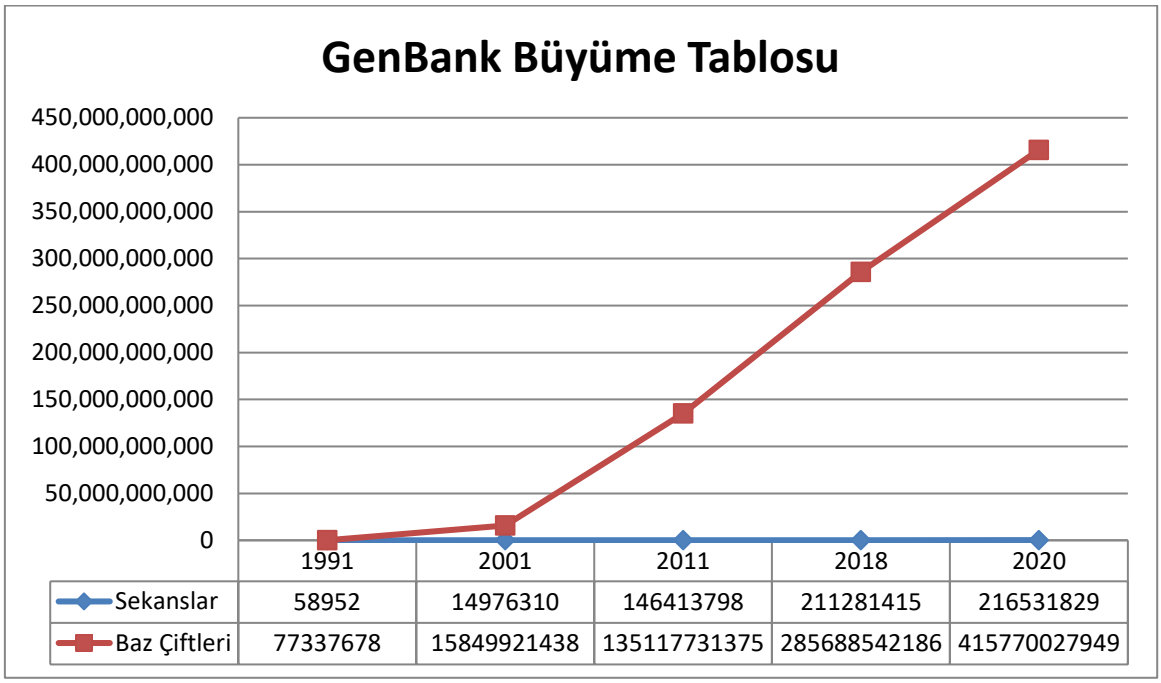
INSDC biyolojik topluluk tarafından yaygın olarak kullanılmaktadır. Şekil 1.7'deki gibi veri tabanlarına veri gönderen bireylerin uyması gereken bir takım politikalar bulunmaktadır [62].



Şekil 1.7. Uluslararası nükleotid dizisi veritabanı işbirliği(INSDC)

1.3.5.4.1. GenBank

Genetik Dizi Veri Bankası (GenBank) [63], Ulusal Sağlık Enstitüsü (NHI) genetik dizi veri tabanıdır ve halka açık tüm DNA dizilerinin açıklamalı bir koleksiyonudur. Bilinen genetik sekansların en hızlı büyüyen depolarından birisi olarak görülmektedir. GenBank, Japonya'nın DNA DataBank (DDBJ), Amerika Birleşik Devletleri Ulusal Biyoteknoloji Bilgi Merkezi(NCBI), Avrupa Moleküler Biyoloji Laboratuvarı (EMBL) ve kendisi GenBank'tan oluşan Uluslararası Nükleotid Dizisi Veritabanı İşbirliği sisteminin bir parçasıdır [64]. Her GenBank girişi, dizinin kısa bir açıklamasını, kaynağın bilimsel adını, taksonomisini, transkripsiyon ünitelerini, kodlama bölgelerini, mutasyon veya modifikasyon bölgelerini ve tekrarlar gibi biyolojik önemi bulunan diğer bölgeleri tanımlayan bir öznitelik tablosunu içermektedir. Şekil 1.8 veri bankasının yıllara göre büyümesini göstermektedir [65].



Şekil 1.8. GenBank büyümesi

1.3.5.4.2. Japonya DNA Veri Bankası (DDBJ)

1986 yılında DDBJ, Ulusal Genetik Enstitüsü'nde (NIG) DNA veri bankası faaliyetlerine hızlı ve ciddi bir şekilde başlamıştır. DDBJ, NCBI/GenBank ve EBI/EMBL ile işbirliği içinde olan uluslararası nükleotit sekans veritabanı olarak faaliyet göstermektedir [66].

DNA sekansı organizmaların evrimini diğer biyolojik materyallere göre daha doğrudan kaydedebilmektedir ve bu sebeple sadece yaşam bilimleri araştırmalarını değil, aynı zamanda insan refahı için de paha biçilemez bir sistemdir. Veri tabanları, tabiri caizse, insanların ortak bir hazinesi olarak ifade edilebilir. Bunu göz önünde bulundurarak, bu veri tabanlarına serbestçe ve her zaman online erişilebilmektedir.

1.3.5.4.3. Avrupa Moleküler Biyoloji Laboratuvarı (EMBL)

EMBL Nükleotid Sekans Veritabanı, verilerin bilimsel literatür ve patent uygulamalarından toplanması ve doğrudan araştırmacılar ve dizileme gruplarından gönderilen kapsamlı DNA ve RNA sekansı veri tabanı şeklinde tanımlanmaktadır. Veri toplama GenBank (ABD) ve Japonya DNA Veritabanı (DDBJ) işbirliği ile yapılmaktadır. Veritabanı şu anda boyutu her 18 ayda bir iki katına çıkabilmekte ve Haziran 1994 itibariyle 182.615 dizi girişi ile yaklaşık 2 milyon baz içermekteydi. EMBL-Bank olarak da bilinen EMBL Nükleotit Sekans Veritabanı, Avrupa'nın birincil nükleotit sekans kaynağını oluşturmaktadır [67].

2. DNA DİZİLEME VE DİZİLEME YÖNTEMLERİ

2.1. DNA Dizileme

DNA'nın temel yapısında, nükleotidler art arda bir iplikçik, yani bir DNA iplikçiliği oluşturmaktadır. Bir iplikçikteki nükleotidlerin dizisi, hücre fonksiyonunu kontrol eden proteinlerin üretildiği genetik kod olarak adlandırılmaktadır. Bir nükleotit zincirinin baz sekansını belirlemeye sekanslama denilmektedir [68].

DNA dizilime, DNA içinde bulunan bazların dizilimini belirleyebilmek için kullanılan bir laboratuvar yöntemi şeklinde belirtilmiştir. İnsan genomundaki yaklaşık 3 milyar baz çiftinin dizilimindeki farklılıklar, her birey için eşsiz genetik yapısını göstermektedir. Tıp alanında, DNA dizileme işlemleri, hastalıkların teşhisi ve tedavisi de dâhil birçok amaç için kullanılmaktadır. Genel olarak sekanslama (dizileme) işlemleri, sağlık uzmanlarının bir gen verisini düzenlenmesinde bir genin veya bölgenin, herhangi bir bozukluğa bağlı olarak varyantlar veya mutasyonlar şeklinde değişiklikleri içerip içermediğinin belirlenmesinde yardımcı olabilmektedir.

İlk DNA dizileme yöntemleri 1970'lerde geliştirilmiştir. En iyi bilinenleri Maxam-Gilbert kimyasal yöntemi ve Sanger enzim yöntemidir. Günümüzde, Sanger en yaygın kullanılan sekanslama yöntemidir. Kimyasal yöntem, sadece yavaşlığı ve gerekli toksik kimyasallar nedeniyle istisnai durumlarda kullanılmaktadır [69].

Moleküler biyolojinin hızlı gelişimi sayesinde, sekanslama yöntemleri son yıllarda gelişme göstermektedir. Dizileme daha ucuz, daha hızlı ve hatasızdır, bu nedenle kullanımları genişlemiştir. Sekanslama, moleküler biyolojinin hemen hemen tüm alanlarında, örneğin mikrobiyoloji ve adli tıpta kullanılmaktadır. Dizileme, özellikle genetik mühendisliğinde önemli bir araç olarak kullanılmaktadır. İnsan genomu projesi, 1990'larda başlatılan ve insan genomunun tamamının dizileme yoluyla araştırıldığı bir projedir. Proje 2003 yılında tamamlanmıştır. Projeden bu yana çalışmalar devam etmiştir ve diğer organizmaların genomları da haritalanmıştır [70].

Teknolojinin gelişimine bir örnek, insan genomunun haritalandırılmasıdır. Orijinal İnsan genom projesinde genomun haritasını çıkarmak 13 yıl ve 2.7 milyar dolar almıştır. 2008 yılında benzer çalışmalar beş ayda 1,5 milyon dolar maliyetle yapılmıştır. Aynı yıl, ABD Ulusal İnsan Genomu Araştırma Enstitüsü, insan genomu dizilişini 1000 dolardan daha az bir oranda dizilemesini sağlamayı amaçlayan bir program finanse etmeye başlamıştır [71].

Son yıllarda, genlerin işleyişini, DNA ve proteinlerin etkileşimini ve genomun hastalıklar üzerindeki etkisi daha iyi anlaşılmıştır. Yöntemlerin geliştirilmesi, yöntemlerin kullanılması için gereken süreyi kısaltmıştır ve böylece maliyet etkinliğini de artmıştır. Yöntemler birinci, ikinci ve üçüncü nesil dizileme yöntemleri olarak sınıflandırılabilir [72, 73]. Sekanslama teknikleri sürekli olarak gelişmesine rağmen, yeni yöntemler öncüllerinin yerini almamıştır, ancak her yöntem kendi uygulama alanını uyarlamıştır. Dizileme yöntem seçimi, istenen performans ve hassasiyetten ve sekanslanacak DNA dizisinin uzunluğundan etkilenmektedir [74].

2.1.1. Birinci Nesil Dizileme

1970'lerin sonlarında geliştirilen iki birinci nesil yöntem vardır: Sanger yöntemi ve Maxam ve Gilbert yöntemi [75]. Maxam ve Gilbert tarafından geliştirilen kimyasal yöntem yaklaşık 100 bazlık bir alanın yapısını belirleyebilmiştir. Sanger yönteminin kök stoku hala kullanımda ve şablon olarak tek iplikli DNA kullanmaktadır. Yöntem 800-1000 baz uzunluğunda bir alanın yapısını belirlemek için kullanılabilmektedir. Sanger yönteminin avantajı, kısa DNA ipliklerini dizilemedeki etkinliğidir. Yöntem doğru ve hızlıdır, bu nedenle analiz başına maliyet de avantajlı konumdadır. Aslında, Sanger yöntemi genellikle ilk olarak başka bir yöntemle hedeflenen bir alanı doğrulamak veya tamamlamak için kullanılmaktadır [73, 76].

2.1.2. İkinci Nesil Dizileme

21. yüzyılın başında, daha yüksek performansa sahip daha fazla sayıda DNA molekülünü paralel olarak dizileme yapabilen ikinci nesil yöntemler geliştirilmiştir. Bu yöntemler yüz milyonlarca sekansı hızlı bir şekilde belirleyebilmektedir, ancak sekanslar sadece 35 ila 600 baz şeklinde kısa uzunluktadır. Dizilerin kısalmasına ek olarak, yöntemlerin zayıflığı, Sanger yöntemine kıyasla doğruluğu bozulmaktadır [77].

İkinci nesil yöntemlerden en yaygın olarak kullanılan, on binlerce karşılıklı özdeş DNA parçasının replikasyon (kopyalama) için eklendiği döngüsel grup sıralama yöntemidir. İşlenen veriler tek bir analizde 300 gigabase'e kadar olabilmektedir, bu da verilerin işlenmesini ve birleştirilmesini zorlaştırmaktadır. Aynı anda ve hızlı bir şekilde birden fazla gen hakkında bilgi almak istendiğinde, ikinci nesil yöntemler kullanılmaktadır. Bu yöntemler, diğerlerinin yanı sıra, örneğin patojenleri tanımlamak için kullanılabilecek olan eksonların dizisini incelemek için de kullanılmaktadır [74, 77].

2.1.3. Üçüncü Nesil Dizileme

Birinci ve ikinci nesil sekanslama yöntemlerinde DNA, sekanslama için parçalara ayrılmaktadır veya dizileme sırasında sekanslanmaktadır ve sekanslar da kısa kalmaktadır. Bir DNA molekülünü tek seferde bölünmesi veya bölünme olmadan dizilemenin yeni bir yöntemi, üçüncü nesil bir yöntem olarak adlandırılabilmektedir. Ancak, üçüncü nesil yöntem terimi henüz tam olarak belirlenmemiştir. Üçüncü nesil dizileme; Illumina Tru-seq Sentetik Uzun Okuma teknolojileri, Pacific Biosciences (PacBio) Tek Moleküllü Gerçek Zamanlı (SMRT) dizilemesi ve Oxford Nanopore Technologies şeklinde DNA dizileme yöntemleridir. Klonal amplifikasyon veya tek molekül dizilimi ve uzun molekül dizilimi kullanılarak, her üç teknoloji için de 5.000bp ila 15.000bp arasında ortalama uzun okumalar üretebilmektedir ve bazı özel okumalar için 100.000bp'yi aşabilmektedir. Bunların en köklüsü, 2010 yılında ticari amaçla piyasaya sürülmüş PacBio SMRT teknolojisi olarak belirtilmiştir. Yakın zamanda piyasaya giren PacBio Sequel cihazı,

üretimi 7 kata kadar artırmayı hedeflemektedir. Okumaların %10-15 oranında ham hata oranı vardır, ancak yine de, yeterli kapsama alanı ile nükleotid başına doğruluğu %99,99'un üzerine veya daha fazla artırabilen çeşitli algoritmik teknikler geliştirilmiştir. "Kendi kendini düzeltme" yaklaşımı için yaklaşık 50 kat uzunluğunda bir okuma kapsamı gerekirken, uzun okumaların düzeltilebilmesi için ek yüksek kapsama alanı kısa okuma dizilemesinden faydalanarak hibrit hata düzeltme algoritmaları ile daha düşük kapsama alanı şeklinde etken olarak kullanılabilir. PacBio dizileme yöntemi ile ilgili temel sınırlama, çok sayıda genomu analiz etme uygulamasını sınırlayan ikinci nesil yaklaşımlara göre maliyettir. Bugüne kadar yüzlerce proje, mantar, mikrop, hayvan ve bitki türlerinin çok yüksek kaliteli genomlar oluşması yanı sıra tüm insan genomlarının çok yüksek kaliteli de novo işlemleri de dâhil PacBio dizilemesi başarıyla kullanmıştır [78].

Tablo 2.1. DNA dizileme teknolojilerinin kıyaslanması

	1.Nesil Dizileme	**2.Nesil Dizileme**	**3.Nesil Dizileme**
Temel Teknoloji	Özellikle son etiketli DNA parçalarının boyut ayrımı, Sentez Yoluyla Dizileme (SBS)	Yıkama ve Tarama, Sentez Yoluyla Dizileme (SBS)	Tek Molekül Gerçek Zamanlı Dizileme (SMRT)
Kullanılan Model DNA	DNA'nın parçalara ayrılmış kopyaları	DNA'nın parçalara ayrılmış kopyaları	DNA'nın tek zinciri
Okuma Doğruluğu	Yüksek	Yüksek	Yüksek
Geçerli Okuma Uzunluğu	800-1000 baz çifti	Kısa (Sanger metoduna göre)	1000 baz çifti ve daha fazlası
Verim	Düşük	Yüksek	Yüksek
Maliyet	Baz başına yüksek maliyet, Çalışma başına düşük maliyet	Baz başına düşük maliyet, Çalışma başına yüksek maliyet	Baz başına düşük maliyet, Çalışma başına yüksek maliyet
RNA Dizileme Yöntemi	cDNA Dizileme	cDNA Dizileme	cDNA Dizileme ya da direkt RNA molekülü

Tablo 2.1'in devamı

	1.Nesil Dizileme	2.Nesil Dizileme	3.Nesil Dizileme
Sonuç Zamanı	Saatler	Günler	Dakikalar
Örnek Hazırlama	Kısmen karmaşık, PCR Amplifikasyonu gerekli değildir	Karmaşık, PCR Amplifikasyonu gereklidir	Teknolojisine göre basit-karmaşıktır, PCR Amplifikasyonu gerekli değildir
Veri Analizi	Rutin	Karmaşık (büyük veri hacimleri ve kısa okumalar nedeniyle)	Karmaşık

DNA dizileme teknolojilerinin nesillerine göre özellik farkları Tablo 2.1'de gösterilmektedir [9, 73, 79].

2.1.4. Dizileme Yöntemlerinin Kanser Araştırmalarında Kullanılması

İkinci nesil yöntemler çeşitli mutasyonları tanımlamak için kullanılan yöntemlerdir. Kendi çalışmalarında, Leary, Kinde, Diehl, Schmidt ve Clouser ve arkadaşları (2010) rektal ve meme tümörlerinden translokasyonları aydınlatmışlardır [80]. Shah, Morin, Khattra, Prentice ve Pugh ve ark. (2009), lobüler meme kanserlerinde nokta mutasyonlarını üzerinde çalışmışlardır [81]. Sekanslama yöntemleri, kanser tedavisinin izlenmesinde de yararlı olmuştur, örneğin; kemoterapinin kötü öngörülen meme kanseri tedavisinde yararlılığı araştırılmıştır [82]. Kendi araştırmalarında, Mantere, Winqvist, Kauppila, Grip ve Jukkola-Vuorinen ve diğerleri (2016), Finlandiya popülasyonunda meme kanserinin kalıtsallığını incelerken elde edilen araştırma sonuçlarını desteklemek ve doğrulamak için hem birinci hem de ikinci nesil yöntemleri kullanmıştır [83].

SMRT teknolojisi kanser araştırmalarında birçok farklı şekilde kullanılabilmektedir. Kendi çalışmalarında Flusberg, Webster, Lee, Travers ve Olivares ve arkadaşları (2010)

kimyasal DNA metilasyonlarını saptamak için SMRT yöntemini kullanmışlardır [84]. DNA metilasyonu, bir metil grubunun, yaşam boyunca tamamen normal bir olay olan DNA'daki bir sitozin nükleotidine bağlanmasını ifade etmektedir. Bununla birlikte kanser hücrelerinde, metilasyon genlerin farklı bölgelerinde, hipo veya hipermetilasyonda bozulmuştur [85].

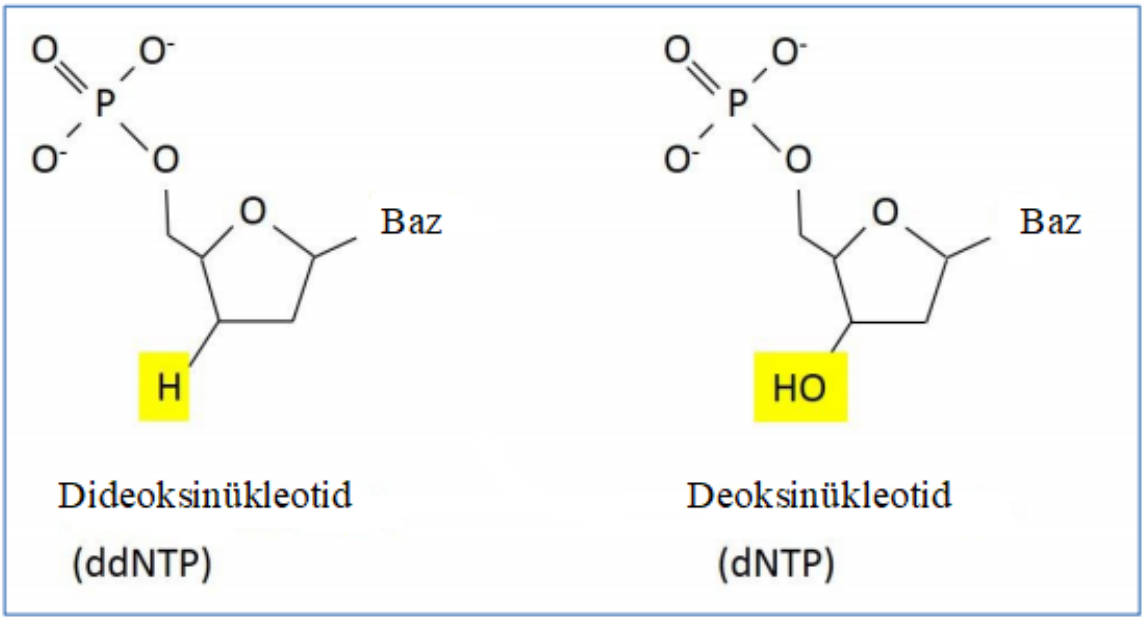
RNA dizilimi yönteminin de katkısıyla, Shuning Ding, Xiaosong Chen ve Kunwei Shen (2020) göğüs kanserinde tek hücreli RNA dizilimi yardımıyla tümör heterojenliğini anlamak ve kişiselleştirilmiş tedaviye yardımcı olmak için çalışmalar yapmışlardır. Meme kanseri araştırmaları alanında son yıllarda yapılan scRNA-seq çalışmaları, kötü prognoz ve ilaç direnci ile ilişkili olabilecek farklı popülasyonları belirlemek için farklı moleküler alt tiplere sahip meme kanseri hücre popülasyonlarını kümelendirilmektedir [86].

2.2. DNA Dizileme Yöntemleri

2.2.1. Zincir Sonlandırma Yöntemi (Sanger Yöntemi)

1977'de biyokimyacı Frederick Sanger meslektaşlarıyla birlikte geliştirdiği enzimatik DNA'yı dizilemek için bir yöntem yayınlamıştır. Yöntem, farklı uzunluklardaki DNA ipliklerinin fragmanlarını (parçalarını) rastgele oluşturmak için kullanılan polimeraz enzimi ve dideoksinükleotitlere dayanmaktadır [87]. Fragmanlar boyutlarına göre ayrılarak DNA iplikçiğindeki nükleotit sekansını bulunabilmektedir [88].

Bu sekanslama yöntemi, başka bir nükleotidin bir bağ oluşturabileceği 3' OH grubundan yoksun dideoksinükleotitlerin Şekil 2.1’de gösterildiği gibi kimyasal yapısını kullanmaktadır. Her nükleotid için bir pentoz halkası, bir baz ve bir fosfat grubu şeklinde oluşmaktadır. Deoksinükleotidlerin pentoz halkasında 3' karbonu üzerinde bir OH grubu bulunmaktadır. Dideoksinükleotitlerde, bunun yerine ve başka bir nükleotidin fosfat grubunun bağlanmasını ve bir polimer oluşumunu önleyen H bulunmaktadır [89].

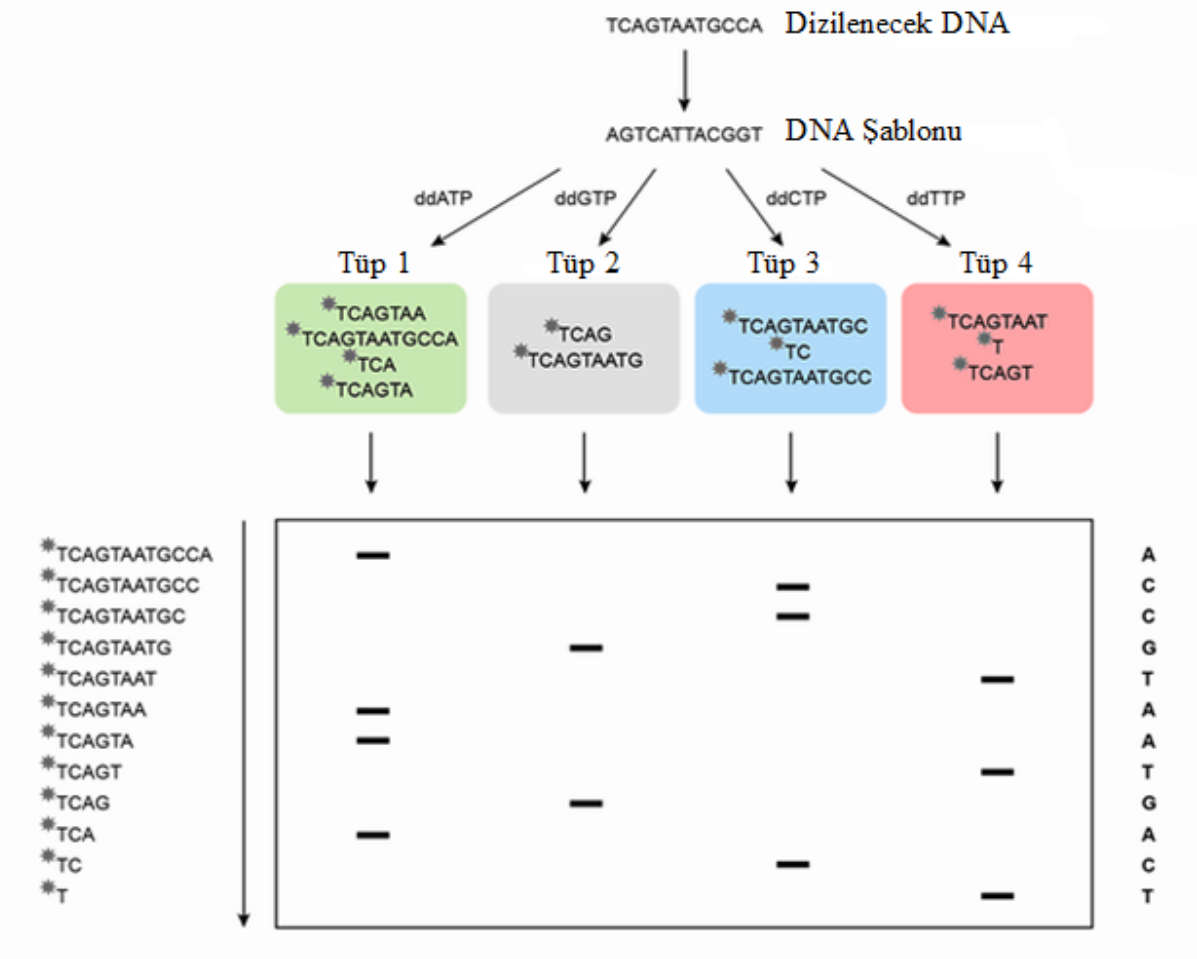


Şekil 2.1. Deoksinükleotid ve Dideoksinükleotid yapısal formülü

Normal durumda, bir DNA molekülünün birbirine bağlı iki nükleotit şeridi bulunmaktadır. Çift iplikli bir yapı oluşturmak için birbirleriyle iç içe geçmiştir. Sekanslama için, sekansı belirlenecek olan DNA tek iplikli bir forma getirilmelidir, yani iplikçikler birbirinden ayrılmış olması gerekmektedir. DNA çift sarmalın açılması ve iplikçiklerin ayrılması denatürasyon olarak adlandırılmaktadır. Bunu başarmanın birçok yolu vardır, örneğin; ısıl işlem gibi [68, 89]. Bir sonraki adımda, tek iplikçikli DNA'nın yanında yeni bir iplik oluşturulmaktadır. Denatüre (bozulan) DNA bir şablon görevi görmektedir. Reaksiyonlar dört farklı test tüpünde gerçekleşmektedir. Şablon tüp, dört farklı deoksinükleotid (dATP, dTTP, dCTP ve dDTP), bir oligonükleotid primeri ve bir polimeraz enzimi her tüpe eklenmektedir. Reaksiyon ürünlerinin daha sonraki bir aşamada görüntülenmesi için ya bir deoksinükleotid ya da primer, örneğin bir radyoaktif bileşik ile etiketlenmiş olmaktadır [68, 90].

Ek olarak, her tüp için az miktarda dideoksinükleotitler gerekmektedir. Dideoksinükleotitler normalde normal nükleotitlere benzer, ancak yapıları molekülün 3' ucundaki hidroksil grubunu bir hidrojen atomu ile değiştirerek modifiye edilmiştir. Şekil 2.2'deki gibi bir tüpe sadece bir tip dideoksinükleotid (ddATP, ddTTP, ddCTP veya

ddDTP) eklenmektedir ve sonunda oluşan parçalar büyüklüklerine göre dizileme işlemi yapılmıştır [89-91].



Şekil 2.2. Zincir sonlandırma yönteminde sonlanan parçaların dizileme işlemi

Tüplerde, primerler DNA şablonlarının uçlarına bağlanmaktadır. Primerler, dizilenen DNA'ya özgü olacak şekilde seçilmelidir. Bağlandıktan sonra, polimeraz enzimi, primerin bir uzantısı olarak şablonun yanında yeni bir deoksinükleotit dizisini sentezlemeye başlamaktadır. Nükleotitlerin bağlanması her zaman baz çifti kuralına göre gerçekleşmektedir: A ve T ile G ve C gibi birbirine bağlanmaktadır. İpliğin sentezi polimeraz dideoksinükleotide iplikçik ile birleşene kadar devam etmektedir. Dideoksinükleotidin bağlanması, telin uzamasını sonlandırmaktadır, çünkü ona yeni nükleotidler eklenememektedir. Polimeraz enzimi etki ettiğinde, bir deoksi veya

dideoksinükleotitin zincire bir sonraki bağlanıp bağlanmadığı ve polimerazın şablonu kopyalamak için zamana sahip olduğu durum rastgele olmaktadır. Sonuç olarak, farklı boyutlarda, “bitmiş” ve “daha az bitmiş” ürünler oluşturulmaktadır. İlk başta yeterli şablon iş parçacığı varsa, tüm olası alternatiflerin en az bir kopyasının ortaya çıkması muhtemel görülmektedir [88, 90, 91].

Ürünlerin tespiti, yani ürünlerin gözlemi, jel elektroforezi, yani bir numuneyi bir poliakrilamid jel üzerindeki bir elektrik alanında çalıştırarak gerçekleşmektedir. Her numune tüpünün içeriği jel üzerinde kendi yerinde pipetlenmektedir. DNA molekülü elektriksel olarak yüklendiğinden, numuneler elektrik akımı geçerken jelden geçmeye başlamaktadır. Daha uzun parçalar jel içinde kısa olanlardan daha yavaş hareket etmektedir, bu nedenle çalışma yeterince uzun sürdüğünde farklı boyutlardaki parçalar öne çıkmaktadır. Parçaların dizisi radyoaktif etiketler kullanılarak X-ışını filminde görüntülenebilmektedir. Parçaların dizisi radyoaktif etiketler kullanılarak X-ışını filminde görüntülenebilmektedir. Nükleotit sekansı, parçaların jel üzerine yerleştirildiği sıraya göre okunmaktadır ve fragmanın(parçanın) bulunduğu örnek tüpüne bağlı olarak, söz konusu baz belirlenmektedir [88, 90].

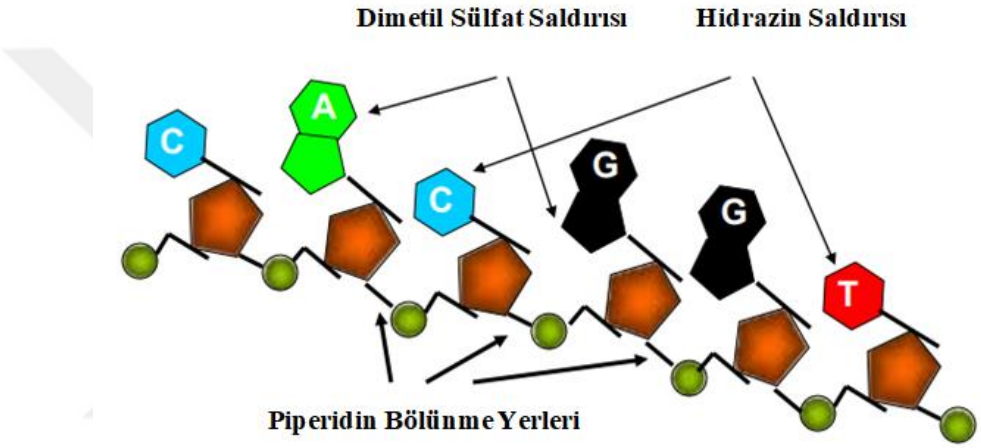
2.2.2. Maxam – Gilbert Dizileme Yöntemi

Allan Maxam ve Walter Gilbert, piperidin ve seçici olarak pürin ve pirimidinlere saldıran iki adet kimyasal bulunduran iki adımlık katalitik işlemden faydalanarak tek iplikçik DNA'yı dizileme işlemi yapabilmek için bir yöntem geliştirmiştir [76]. Bu iki adım:

(1) Pürinler dimetil sülfat ile reaksiyona girmesi ve pirimidinler, riboz şekeri ile bazın yerini alan baz arasındaki glikozit bağını kıracak şekilde hidrazin ile reaksiyona girmesidir.

(2) Piperidin daha sonra bazın yer değiştirdiği fosfodiester bağ ayrılmasını katalize etmesidir.

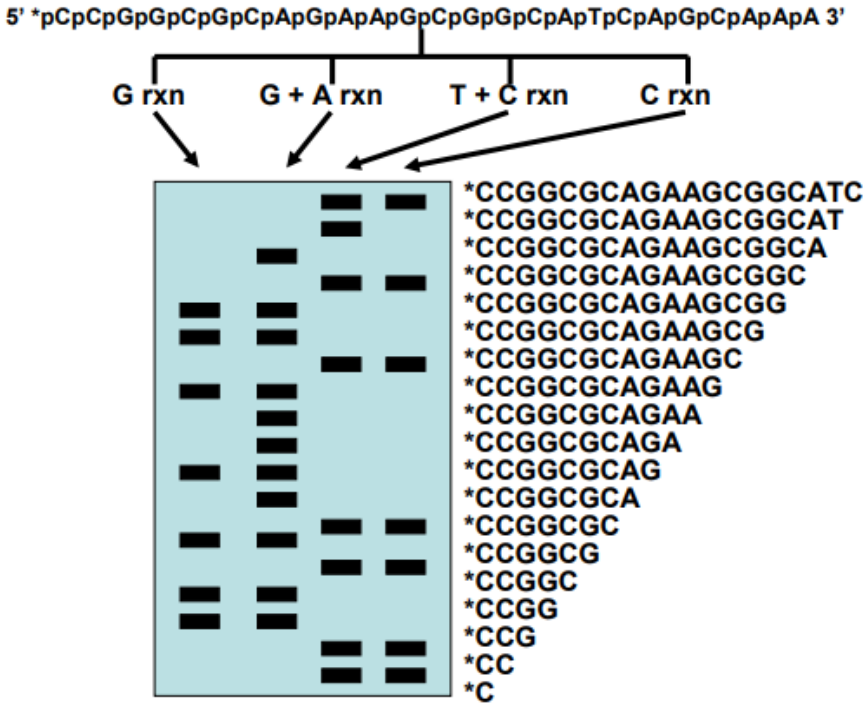
Ayrıca, dimetil sülfat ve piperidin tek başına guanin nükleotitlerini seçici olarak ayırmaktadır, ancak formik asit içindeki dimetil sülfat ve piperidin hem guanin hem de adenin nükleotitlerini ayırmaktadır. Benzer şekilde, hidrazin ve piperidin hem timini hem de sitozin nükleotitlerini ayırmaktadır, halbuki Şekil 2.3'te görüldüğü gibi 1.5M NaCl içindeki hidrazin ve piperidin sadece sitozin nükleotitlerini ayırmaktadır. Dimetil Sülfat veya hidrazin, sırasıyla pürin veya pirimidin halkalarına saldırmaktadır ve piperidin, 3' karbondaki fosfat bağını parçalamaktadır [89, 92].



Şekil 2.3. Maxam-Gilbert DNA dizileme kimyasal hedefler

Bu seçici şekildeki reaksiyonların DNA dizileme işleminde kullanımı, 5' ucunda radyoaktif etiket taşıyan tek parça için bir DNA substratı oluşturulmasını içermektedir. Bu etiketlenmiş şekilde bulunan substrat, her biri bilinen nükleotidlerle biten şeklinde bir etiketlenmiş olan bölünme ürünü popülasyonu oluşturacak dört ayrı bölünme reaksiyonuna girmektedir. Reaksiyonlar yüksek oranda poliakrilamid jellerine yüklenmektedir ve fragmanlar elektroforez ile çözülmektedir. Bu jel ışık geçirmez bir X-ışını (X-Ray) film kasetine, bir parça üzerine yerleştirilen X-ışını filmi ve kaset birkaç gün süresince bir dondurucuda saklanmaktadır. Etiketli olan parçanın jel üzerinde durduğu bölgede radyoaktif etiketli partikül çürümesi (otoradyografi) sebebiyle filmi ortaya çıkarmaktadır. Elektroforez, ister akrilamid ister agaroz matrisinde nükleik asit fragmanlarını ters uzunluk sırasına göre çözeceğinden, daha küçük fragmanlar jel matrisinde daha büyük

fragmanlardan daha hızlı çalışacağından, Şekil 2.4'te görüldüğü gibi filmdeki koyu otoradyografik bantlar aşağıdan yukarıya doğru okunduğunda 5'→ 3' DNA sekansı temsil etmektedir. Baz arama işlemi, dört kimyasal reaksiyona göre bantlama (modelinin) paterninin yorumlanmasını içermektedir. Hedef DNA radyoaktif olarak etiketlenmektedir ve daha sonra dört kimyasal bölünme reaksiyonuna ayrılmaktadır. Her reaksiyon bir poliakrilamid jel üzerine yüklenip, çalıştırılmaktadır. Son olarak, jel otoradyografiye tabi tutulup, baz sırası aşağıdan yukarıya doğru ilerlemektedir [92].



Şekil 2.4. Maxam-Gilbert manuel dizileme düzeni

Örneğin, şeritlerde sadece C'ye ve C + T reaksiyonlarına karşılık gelen bir bant C olarak adlandırılmaktadır. Bant C + T reaksiyon şeridinde mevcuttur fakat sadece C reaksiyon şeridinde olmasaydı buna T olarak adlandırılmış olurdu. Aynı şekildeki karar süreci için sadece G ve G + A reaksiyon şeritleri elde edilebilmektedir. Diziler, aynı jel üzerinde çoğaltma reaksiyonları yapılarak ve çoğaltılan diziler arasındaki otoradyografik modeller karşılaştırılması sonucu doğrulaması yapılmaktadır. Eğer radyoaktif etiketleme işlemi işe yaramışsa ve bölünme reaksiyonları da beklenildiği gibi yapılmışsa, jel istenilen

şekilde ayarlanmışsa, elektroforez çalışmıştır demektir, jel transfer işlmeminde yırtılmamış veya başka bir şekilde yok edilmemiş ve X-ışını film geliştiricisi bozulmamış ise, birkaç günde bir doğrulanmış şekilde yaklaşık 200-300 bazlık DNA dizilimi elde edilmesi beklenmektedir. Bu önemli bilgi doğrultusunda değişim, 35S veya 32P gibi oldukça fazla miktarda radyoaktif malzeme kullanılması ile oluşmuştur ve sürekli olarak büyük, kağıt ince akrilamid jellerin dökülmesi gerekmektedir ve hidrazin bir nörotoksin olmaktadır. Bunun yanısıra, engellere rağmen, DNA sekansları birçok genden ve organizmadan birikmeye başlamaktadır. İlk keşiflerden biri olan, ökaryotik gen organizasyonun prokaryotik gen organizasyonu ile aynı olduğu varsayımının çöktüğünü göstermektedir. Breathnach ve diğerleri ile Jeffries ve Flavell, tavukta ovalbumin kodlayan genin ve tavşanda β-globin kodlayan genin, kodlama bölgelerinde kodlayıcı olmayan boşluklar içerdiğini keşfettiğini açıklamışlardır [92-94]. Bu boşluklar iki gende aynı dinükleotitler tarafından kuşatılmıştır; boşlukların 5' ucunda GT ve boşlukların 3' ucunda AG olarak ifade edilmiştir. Kısa süre sonra Breathnach ve Chambon, bu GT / AG kuralının bir dizi kodlama dizisi boşluğuna uyulduğunu ve ökaryotik genlerin kodlama ve kodlamayan bölgelerini tanımlamak için genetik sözlüğe intron ve ekson terimlerinin eklendiğini bildirmiştir [95].

2.2.3. Shotgun Dizileme Yöntemi

Shotgun Dizileme yöntemi, DNA dizilerini rastgele birçok küçük parçaya bölmeyi ve daha sonra örtüşme bölgelerini arayarak diziyi yeniden birleştirmeyi içermektedir. Tüm genomlar gibi büyük DNA sekanslarını dizilemek için kullanılan bir yöntem olarak bilinmektedir. Genomun birçok kopyası önce milyonlarca küçük parçaya bölünmektedir. Her parça sıralandıktan sonra, güçlü bilgisayarlar parçaları orijinal sıralarına monte etmek için çakışan bölümler kullanmaktadır.

DNA dizilemede zincir sonlandırma yöntemi (Sanger dizileme) sadece 100-1000 baz çiftli kısa DNA sekansları için kullanılmaktadır. Bu boyut sınırından dolayı, daha uzun sekanslar ayrı gruplarda daha küçük fragmanlara ayrılmaktadır ve bu sekanslar, toplam sekansı oluşturacak şekilde birleştirilmektedir. Bu parçalanma ve dizileme işlemi için iki

ana yöntem kullanılmaktadır: Primer yürüme (kromozom yürüme) tüm iplik boyunca parça parça ilerlerken, Shotgun dizilimi rastgele parçalar kullanan daha hızlı ama daha karmaşık bir işlem şeklinde yürütülmektedir. Shotgun dizilemesi için DNA; okuma işlemi için zincir sonlandırma yönteminden faydalanılarak çok sayıda küçük parçalara rastgele bölünmektedir. Hedef DNA için çoklu okumalar(contig), bu fragmantasyon ve sekanslama için belirlenen veri uzunluğuna göre belli tur sayısı ile gerçekleştirilerek elde edilebilmektedir. Sonrasında bilgisayar programları düzenli bir diziye monte edebilmek için farklı okumaların üst üste gelen uçlarını kullanmaktadır. Shotgun dizilimi, tüm genom dizilemesini sağlamaktan sorumlu öncü teknolojilerden biri olarak çıkarılmıştır. Örneğin; Tablo 2.2'de iki tane Shotgun Dizileme okuması ele alınmıştır [96].

Tablo 2.2. Shotgun dizileme yöntemi okuma örneği

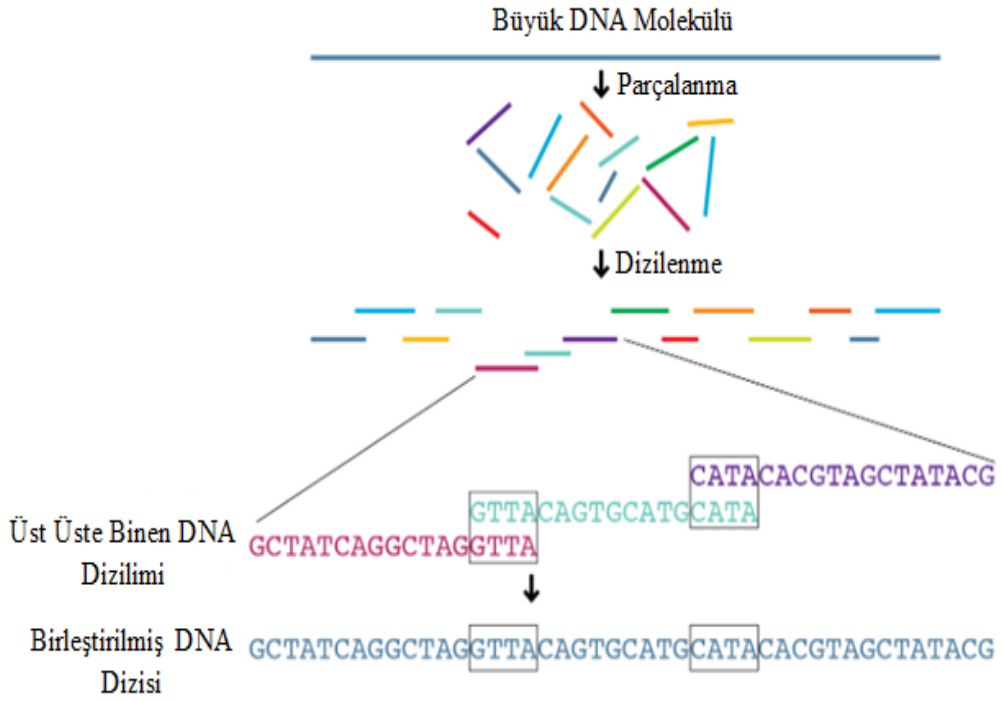
VERİ İŞLEM	**DİZİ**
Orijinal Veri	AGCATGCTGCAGTCATGCTTAGGCTA
İlk Shotgun Okuma Dizisi	AGCATGCTGCAGTCATGCT---------- ----------TAGGCTA
İkinci Shotgun Okuma Dizisi	AGCATG---------- ----------CTGCAGTCATGCTTAGGCTA
Yeniden Yapılandırma	AGCATGCTGCAGTCATGCTTAGGCTA

Bu basitleştirilmiş örnekte, okumaların hiçbiri orijinal dizinin tam uzunluğunu kapsamamaktadır, ancak dört okuma, hizalamak ve sıralamak için uçlarının üst üste binmesi kullanılarak orijinal diziye monte edilebilmektedir. Gerçekte, bu süreç belirsizlikler ve dizileme hataları ile dolu çok büyük miktarda bilgi kullanmaktadır. Karmaşık genomların birleştirilmesi ayrıca tekrar eden sekansların fazla olması nedeniyle karmaşıktır, yani sekansın tamamen farklı kısımlarından benzer kısa okumalar gelebilmektedir [96].

Bu zorlukların üstesinden gelmek ve diziyi doğru bir şekilde monte etmek için orijinal DNA'nın her segmenti için üst üste gelen birçok okuma gerekmektedir. Örneğin,

İnsan Genom Projesini tamamlamak için, insan genomunun çoğu 12X veya daha fazla kapsama alanında dizilmiştir; yani son sekanstaki her baz ortalama 12 farklı okumada bulunmaktadır. Buna rağmen, mevcut yöntemler 2004'ten itibaren (ökromatik) insan genomunun yaklaşık %1'i için güvenilir sekansı izole edememiş veya bir araya getirememiştir [97].

Büyük DNA verisinin parçalanma, dizilenme ve birleşme adımları Şekil 2.5'te gösterilmiştir. Başlangıç DNA'sı rastgele birçok küçük parçaya bölünmektedir, bir çeşit Shotgun tarzında, bu parçaların her biri daha sonra ayrı ayrı dizilenmektedir. Farklı parçalardan üretilen sonuç dizisi okumaları daha sonra, bir bilgisayar programı tarafından analiz edilmektedir ve birbiriyle örtüşen farklı okumalardan dizi uzunlukları incelenmektedir. Örtüşen bölgeler belirlendiğinde, iki sekans okumasının birbirine eklenmesine izin verecek şekilde birbirleriyle birleştirilmektedir. Bu bilgisayar işlemi defalarca tekrarlanmaktadır ve sonunda DNA'nın başlangıç parçasının tüm sekansı oluşmaktadır [97, 98].



Şekil 2.5. Shotgun dizileme yöntemi örtüşen parçalar ve birleşme işlemi

2.2.4. Polimeraz Zincir Reaksiyonu (PCR) Yöntemi

1983 yılında Polimeraz zincir reaksiyonu (PCR) yöntemi, Kary Mullis tarafından geliştirilmiştir ve 1985 yılında patentlenmiştir [99, 100]. Bu yöntem, bir nükleik asit sekansının in vitro olarak katlanarak çoğaltılabildiği bir yöntem olarak kullanılmaktadır. Genel anlamda, bir DNA şablonu belirli bir DNA fragmanının on milyarlarca kopyasını üretilebilmektedir. PCR'nin gücü, matris DNA sayısının teorik açıdan sınırlayıcı bir faktör olmadığı önermesine dayanmaktadır. Bu nedenle nükleotit sekanslarını sonsuz miktarda DNA ekstraktından çoğaltılabilmektedir. Bu işlemler doğrultusunda PCR bir saflaştırma veya klonlama tekniği olarak adlandırılabilmektedir. Bir organizmadan veya çeşitli türlerin DNA numunesinden ekstrakte edilmiş bir DNA doğrudan analizi yapılamamaktadır. Birçok nükleotid dizisi bulunmaktadır. Bundan dolayı, hem normal bir genin hemde kodlayıcı olmayan dizilerin, ilgilenilen sekansların izole edilebilmesi ve saflaştırılabilmesi gerekmektedir. Matris DNA'yı oluşturan böyle bir dizi kütleden, PCR böylece bir veya daha fazla dizi seçebilmektedir ve bunları on milyarlarca kopyalayarak çoğaltabilmektedir. Reaksiyon tamamlandığı zaman, ilgilenilen bölgede olmayan matris DNA sayısı değişmemektedir. Aksine, amplifiye edilmiş sekansların (ilgilenilen DNA) miktarı çok büyük olacaktır. PCR, bir arka plan gürültüsü üzerinden sinyalin yükseltilmesini olanak sağlamaktadır, bundan dolayı moleküler bir klonlama teknolojisidir. PCR'ın birçok uygulama alanı bulunmaktadır. Moleküler ve hücresel biyolojide çok önemli bir yöntem olarak kullanılmaktadır. Özellikle birkaç saat içinde, bir DNA parçasının standart moleküler klonlama teknikleriyle birkaç gün süren otomatik bir sistem aracılığıyla "hücresel klonlama" yapmasını mümkün kılmıştır. Öte yandan PCR, biyolojik bir sıvı içinde organizmaların spesifik bir DNA sekansının varlığını belirlenmesinde yaygın olarak kullanılmaktadır. Aynı zamanda, bir kişinin adli soruşturma bağlamında genetik kimlik tanımlanması veya gıda kalitesi testi, teşhis veya çeşit seçimi için hayvan, bitkinin veya mikrobiyal tanımlanması olsun, genetik parmak izleri yapmak için de kullanılmaktadır [99, 100]. PCR, sekanslama veya bölgeye yönelik mutagenezi gerçekleştirmek için hala gerekli bir yöntemdir.

Günümüzde moleküler biyoloji için birçok çalışmanın işleyişi PCR tekniğine dayanmaktadır. PCR tekniğini için alt akışında birkaç uygulama:

(1) Belirlenen hayvan ırklarının tüm genom dizisinin oluşturulması;
(2) Genom boyunca belli başlı dağınık polimorfizmleri ölçebilen SNP saptama yöntemleri gibi birçok teknolojinin geliştirilmesi;
(3) Büyük boyutlardaki gen transkripsiyonunu ölçebilmek için mikrodizi teknolojilerinin geliştirilmesi.

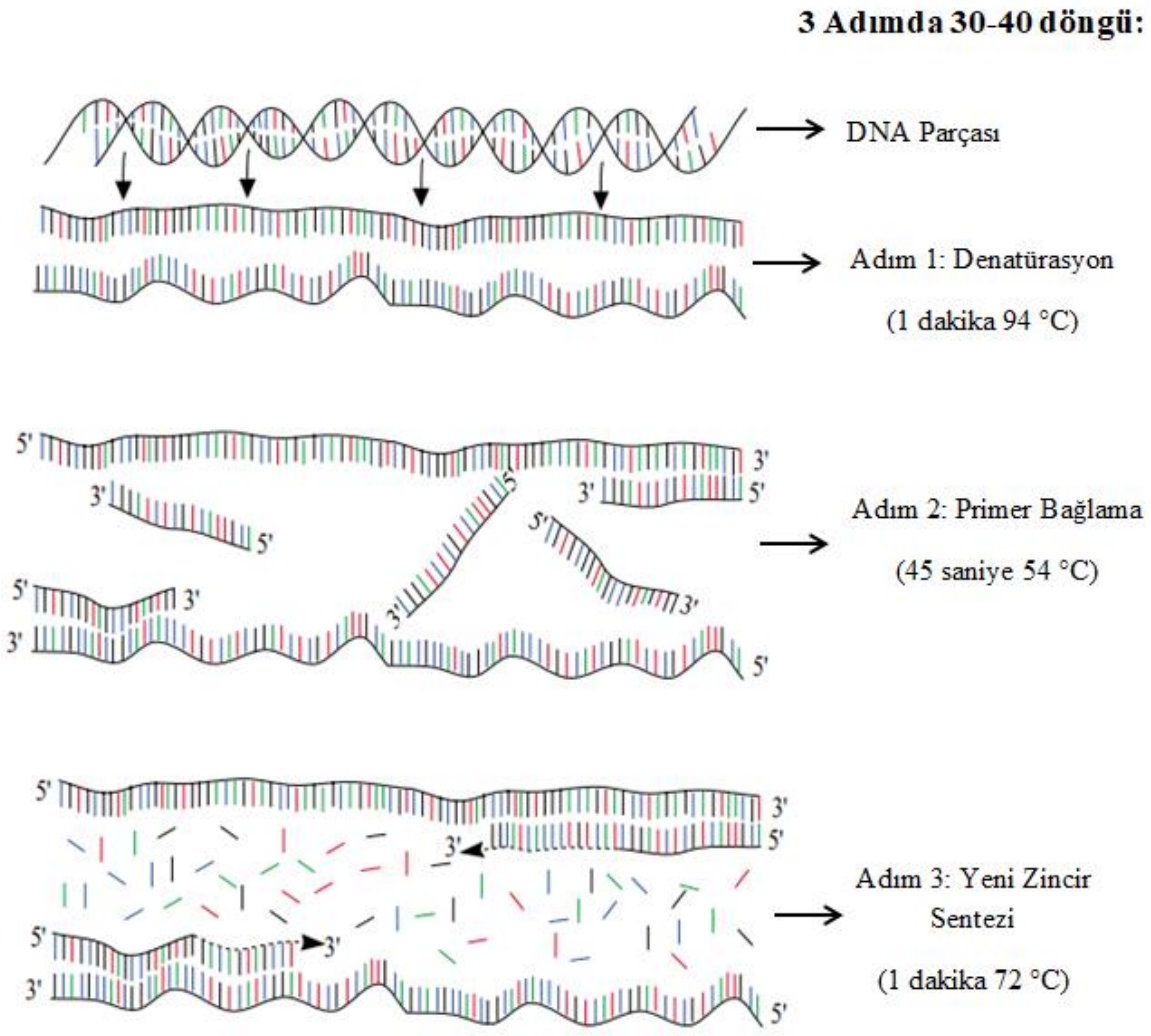
Biyolojik karmaşıklık çalışmaları, yüksek verimlilikteki moleküler teknolojileri, yüksek hız ve bilgisayar belleği, disiplinlerarası becerilerin ve veri analizine yeni yaklaşımların entegrasyonunu oluşturabilen sınır şeklinde nitelendirilmiştir [99, 100].

2.2.4.1. PCR Prensibi

PCR, in vitro replikasyon yoluyla bir DNA parçasının bir şablonundan çoklu kopyalarını elde etmeyi mümkün kılmaktadır [100]. Matris DNA, bir haberci RNA ekstraktından (poli-A RNA) veya hatta mitokondriyal DNA'dan Ters transkripsiyon PCR (RT-PCR) ile elde edilen tamamlayıcı DNA'nın yanı sıra genomik DNA da olabilmektedir. Bir DNA örneğinden büyük miktarlarda spesifik bir DNA sekansı elde etmek için kullanılmaktadır. Bu amplifikasyon işlemi, çift iplikli DNA şablonunun replikasyonuna (çoğaltılmasına) dayanmaktadır. Şekil 2.6'da gösterildiği gibi 3 adımdan oluşmaktadır:

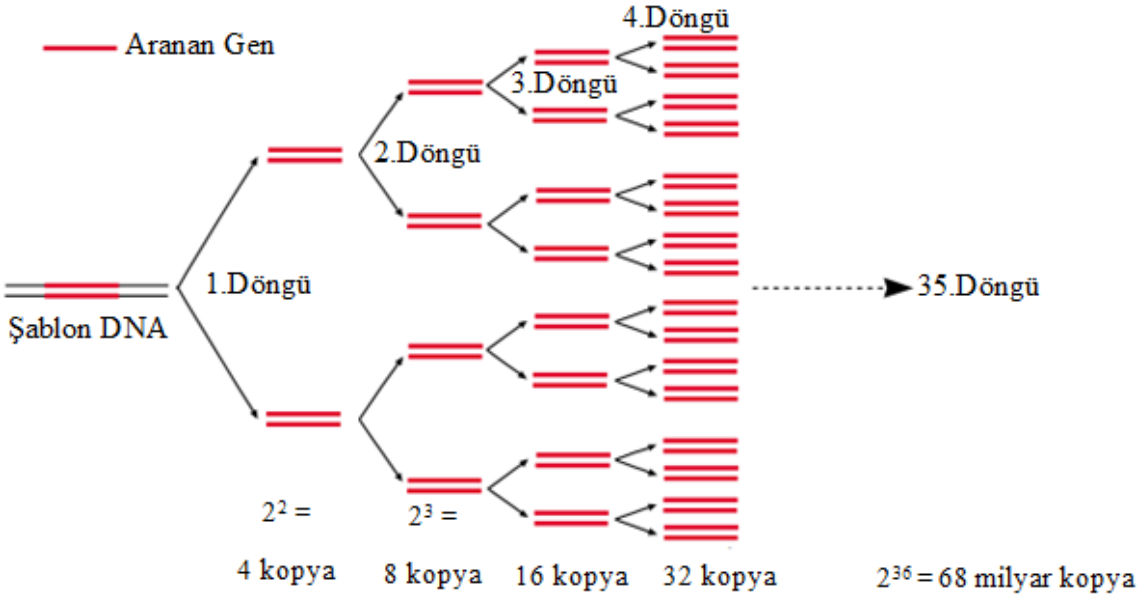
(1) Denatürasyon,
(2) Primer Bağlama,
(3) Yeni Zincir Sentezi.

Her sentez ürünleri için, gösterilen adımlar bir şablon olarak görev almaktadır, bundan dolayı Şekil 2.7'de gösterildiği üzere üstel amplifikasyon elde edilebilmektedir [101].



Şekil 2.6. PCR amplifikasyon yöntemi adımları

PCR, bir tampon çözeltisi içinde fazladan DNA ekstresi (şablon DNA), Taq polimeraz, primerler ve dört deoksiribonükleosid trifosfat (dNTP) içeren bir reaksiyon karışımında gerçekleştirilmektedir [102]. Karışım reaksiyonunu içeren tüpler, bir termal döngüleyicinin ısıtma bloğunda (örnek tüplerinin biriktiği bir muhafazaya sahip olan ve sıcaklığın Peltier etkisiyle 0'dan 100°C'ye kadar çok hızlı ve hassas bir şekilde değişebileceği cihaz) tekrarlanan sıcaklık döngülerine maruz bırakılmaktadır [101, 103]. Cihaz, sıcaklık adım döngü süresini ve art arda programlanma işlemine olanak sağlamaktadır. Her döngü için birkaç on saniyelik üç periyot gerekmektedir.



Şekil 2.7. PCR üstel amplifikasyonu

Her döngüden sonra, yeni sentezlenen DNA zincirleri bir sonraki döngüde şablon olarak kullanılabilmektedir. Şekil 2.7'de gösterildiği gibi, bu üstel reaksiyonun ana ürünü, terminleri oligonükleotid primerlerinin 5' termini ile tanımlanan ve uzunluğu primerler arasındaki mesafe ile tanımlanan bir dsDNA segmentidir [101, 102]. Başarılı şekilde ilk amplifikasyonun ürünleri, uzunlukları iki primerin birbiriyle bağlanma yeri arasındaki mesafeyi aşabilen dinamik heterojen boyutlu DNA molekülleridir. İkinci turda, bu moleküller daha sonraki amplifikasyon turlarında üstel bir şekilde birikecek ve reaksiyonun baskın ürünlerini oluşturacak olan tanımlı uzunlukta DNA zincirleri üretmektedir. Böylece, hedef dizinin son kopya sayısı olarak amplifikasyon, aşağıdaki denklemle ifade edilmektedir:

$$(2^n - 2n)x \tag{2.1}$$

Hedef dizinin son kopya sayısı olarak amplifikasyon hesaplaması Denklem 2.1'de gösterilmiştir. Denklemde kullanılmış olan parametreler:

- **n:** Döngü sayısı,
- **2n:** Birinci turdan sonra elde edilen ilk üründür ve tanımlanmamış uzunlukta ikinci turdan sonra elde edilen ikinci üründür,
- **x:** Orjinal şablon kopya sayısıdır.

Potansiyel olarak, 20 döngüden sonra, her döngü sırasında %100 verimlilik varsayımıyla 220 kat bir amplifikasyon oluşmaktadır. Bir PCR'nin etkinlik durumu, şablondan şablona ve uygulanan optimizasyon derecesine göre değişmektedir [101-103].

Denatürasyon Fazı, sıcaklığı artırılarak elde edilen iki DNA gen dizisinin ayrılmasıdır. İlk olarak denatürasyon sıcaklığı adımı 94°C'lik bir sıcaklık değerinde gerçekleştirilmektedir. Bu sıcaklık için, replikasyon aşamasında matris görevi yapan matris DNA denatüre edilmektedir ve hidrojen bağları 80°C'den yüksek sıcaklık değerinde tutulamamaktadır, ayrıca çift zincirli DNA molekülü, tek zincirli olan DNA'ya (ssDNA) denatüre edilmektedir [102, 103].

İkinci adım ise Primer Bağlama hibridizasyonu olarak isimlendirilmektedir. Primer Bağlama sıcaklığı genellikle 40 ila 70°C arasında bir sıcaklık değerinde gerçekleştirilmektedir. Sıcaklığın azaltılması hidrojen bağlarınında yeniden şekillenme sağlamaktadır ve böylece tamamlayıcı ipliklerin de melezleşmesine izin verilmiş olmaktadır. Primerler, amplifiye edilecek DNA'yı çevreleyen bölgeleri tamamlayan kısa tek zincirli diziler, uzun zincirli matris DNA'dan daha kolay hibridize edilmektedir. Hibridizasyon sıcaklığı ne kadar yüksek değerde olur ve hibridizasyon ne kadar seçici olursa, işlem o kadar spesifik olmaktadır [102, 103].

Üçüncü adım olan, primer uzaması sıcaklığı 72°C'lik sıcaklıkta gerçekleştirilmektedir. Tamamlayıcı iplikçik sentezidir. 72°C'de Taq polimerazlar, primer tek iplikli olan DNA'lara bağlanmaktadır ve reaksiyon karışımı işleminde bulunan deoksiribonükleosid trifosfatlarından yararlanılarak replikasyonu katalize etmektedir. Böylece şablon DNA'nın primerlerin alt akışındaki bölgeleri seçici bir şekilde sentezlenmektedir. Bir sonraki döngü işleminde, önceki döngüde sentezlenmiş olan

fragmanlar (parçalar) dizili matris olmaktadır ve birkaç döngü sonrasında baskın tür, primerlerin hibritleşme işlemi bölgeleri arasındaki DNA veri dizisine karşılık gelmektedir. Analiz edilebilir miktarda DNA (yaklaşık 0.1 μg) sentezlemek 20-40 döngü sürmektedir. Her döngüde teorik olarak önceki döngüde oluşan DNA verisi miktarını iki katına çıkarmaktadır. PCR reaksiyonu oldukça hızlıdır, sadece birkaç saat sürmektedir (30 döngü PCR için 2-3 saat) [102, 103].

2.2.5. Masif (Kitlesel) Paralel Dizileme Yöntemi

Yeni nesil dizileme teknolojilerinin ilki olan Kitlesel Paralel Dizilemesi (MPS), 1990'larda Lynx Therapeutics'te geliştirilmiştir [104]. Klasik Sanger sekanslaması, bir DNA zincirinin sentezi sırasında DNA sekansının, zincir sonlandırıcı modifiye nükleotitlere bağlı nükleotide spesifik floroforların dahil edilmesi yoluyla belirlendiği güvenilir bir standart yöntemdir. Üretilen etiketli parçaların elektroforetik ayrıştırma ve tespiti, numunenin DNA dizisini ortaya koymaktadır. Düşük hata oranlarına sahip olmasına rağmen, seri yapısı nedeniyle genomun büyük segmentlerini dizilemek için nispeten yavaş ve pahalı olan bir yöntemdir. Birçok hedefi (farklı belirteçler ve bireyler) dizilemesi uygun fiyatlı bir alternatif, son yirmi yılda gelişen yeni nesil dizileme(YND) yöntemidir [105].

Çeşitli platformlar farklı avantajlı özellikler sunmaktadır ve bu teknolojilere birlikte yeni nesil dizileme veya daha doğru bir şekilde kitlesel paralel sıralama, MPS (ve bazen de ikinci nesil sıralama olarak) olarak adlandırılmaktadır. Başka bir terim olan “üçüncü nesil dizileme”, Oxford Nanopore Teknolojisi ve Pacific Biosciences dizileme yöntemi gibi tek DNA molekülleri üzerinde uzun okuma dizilemesini kullanan bu yöntemlerin alt kümesi için de kullanılmaktadır. Temel teknolojiler değişmekle birlikte, bunların hepsi tek bir deney sırasında birden fazla örnekten büyük miktarlarda dizi verisi (milyonlarca okuma aralığında) üretmek için kitlesel (büyük ölçüde) paralel yaklaşımlar kullanılmaktadır [106]. Bu sağlam kapasite ve yüksek verim, dizilenen nükleotid başına maliyeti düşürmektedir: 2017'nin başlarında, insan boyutlu bir genomun dizileme maliyeti 1000 USD'ye yaklaşmıştır [107]. Bu teknolojiler, DNA parçasının dizilemesi için cihaz başına yaklaşık 1

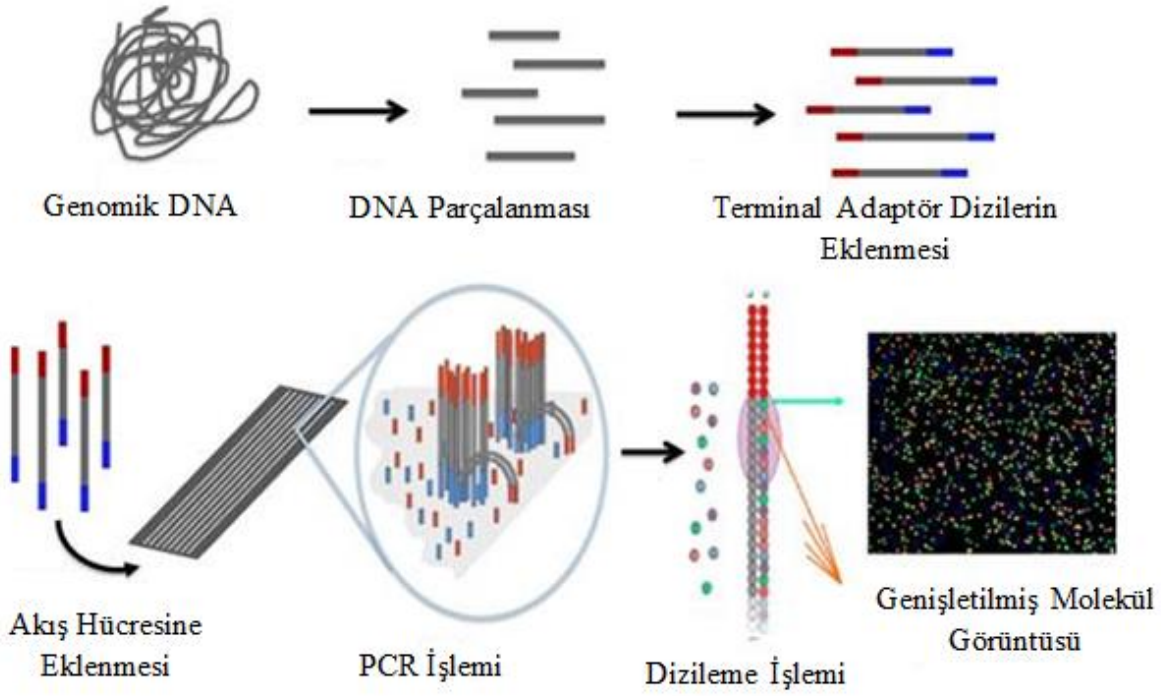
milyon - 43 milyar arasında kısa okumayı (her biri 50-400 baz) dizilemek için minyatürleştirilmiş ve paralelleştirilmiş platformlar kullanmaktadır.

Kitlesel paralel dizileme, bir çalışma sırasında eşzamanlı olarak çok sayıda çoklu okuma üreterek yüksek bir hassasiyet elde etmektedir, bu da Sanger dizilişine daha fazla ve daha güvenilir alternatifler sağlamak için bu platformların sürekli olarak geliştirildiği nispeten yüksek dizileme hatası oranlarını azaltan yüksek bir kapsama alanı sağlamaktadır.

2.2.5.1. Kitlesel Paralel Dizileme Süreci

MPS analizinde, sekanslamadan önce numune hazırlama işlemine "kütüphane hazırlama" denilmektedir. Kısa-okuma yaklaşımları için, bu, ekleme DNA'sının istenen fragman uzunluğuna ilk olarak uyarlanmasını gerektirmektedir; bu, enzimatik fragmantasyon, fiziksel kesme veya amplikon üretmek için PCR kullanılarak elde edilebilmektedir; uzun-okuma yaklaşımları için, DNA doğal parçalanmamış halde bırakılabilmektedir. Üretilen DNA fragmanları, sekanslama başlangıcı için spesifik sekanslar ve endeksler (barkod sekansları) içeren bir deneyde birkaç numunenin çoğaltılmasına olanak sağlayan terminal adaptörleri eklenerek hazırlanmaktadır [106, 108].

Sentezleme yoluyla sıralama (SBS) yöntemleri, nükleotitlerin büyüyen bir iplikçik içine dâhil edilmesinde bir sinyali algılamaktadır. Platforma bağlı olarak, bu farklı floresan sinyallerinin (Illumina® platformları) görsel olarak algılanması veya birleşme olayları sırasında meydana gelen yarı orantılı elektrokimyasal sinyallerin (İyon Torrent platformları) ölçülmesi ile olabilmektedir. Her iki yaklaşımda da, nükleotitlerin sekansları, her biri binlerce tekli molekülün klonal kopyasından oluşan milyonlarca kümeye paralel olarak kaydedilebilmektedir [108]. MPS işlem adımları Şekil 2.8'de gösterilmektedir [109].



Şekil 2.8. Kitlesel paralel dizileme yöntemi adımları

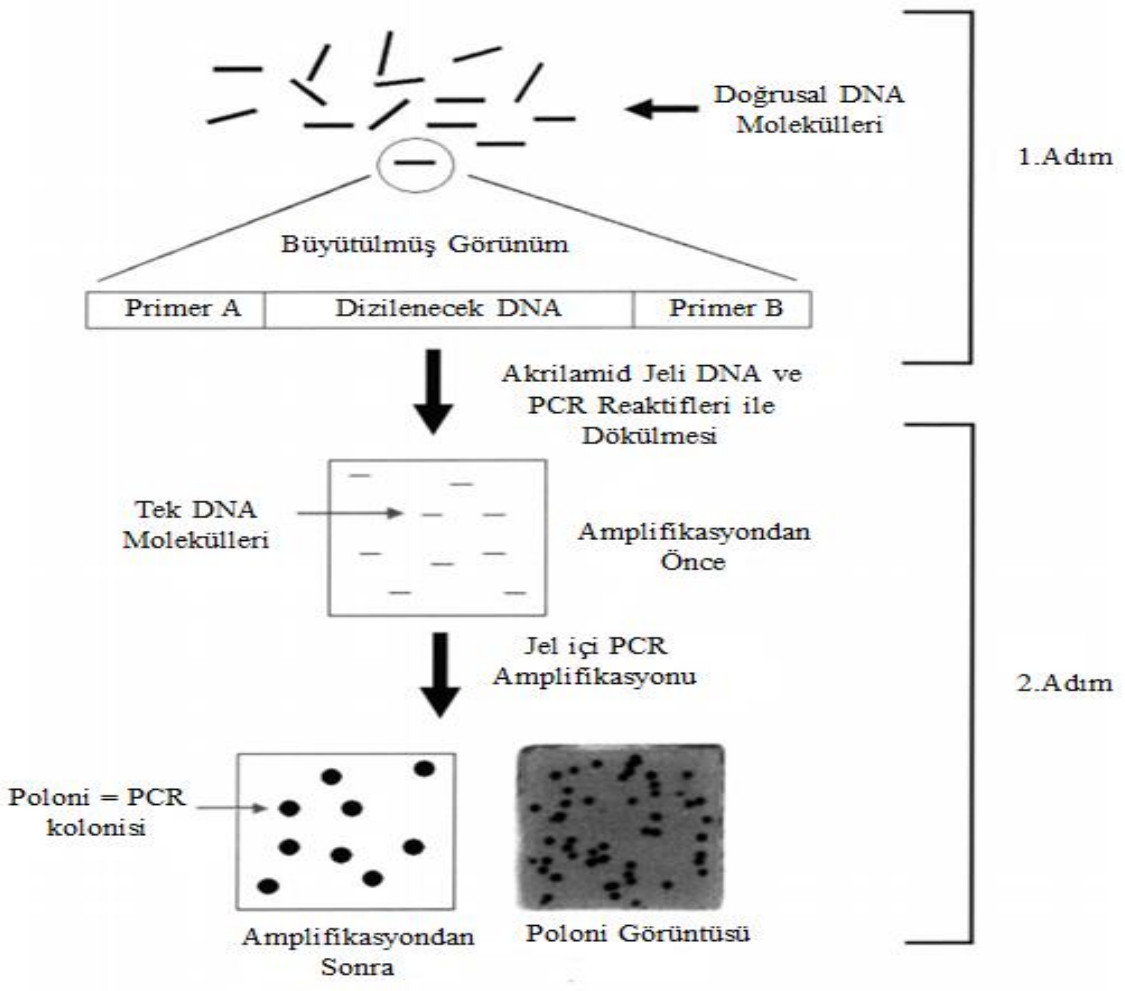
2.2.6. Poloni Dizileme Yöntemi

Poloni sekanslama yöntemi, milyonlarca hareketsizleştirilmiş DNA sekansını maliyeti düşük şekilde paralel olarak okumak için kullanılabilen, fakat oldukça hassas bir multipleks sekanslama tekniği olarak kullanılmaktadır. Bu yöntem ilk olarak Harvard Tıp Okulu'nda bulunan Dr.George Kilisesi grubu tarafından geliştirilmiştir. Poloni sekanslama yöntemi, DNA'nın bir cam mikroskop lamına tutturulmuş şekilde ince poliakrilamid film üzerinde klonlandığı, sekanslandığı ve amplifiye edildiği yeni bir dizileme teknolojisidir. Beş milyon bireysel reaksiyonu için tek bir işlem süresince paralel bir şekilde gerçekleştirileceği düşünülmektedir, böylece sekanslama hızlı ve daha ucuz gerçekleştirilebilmektedir. Poloni sekanslama genellikle, DNA şablonunun her molekülünün 135 bp uzunluğunda olduğu ve ortak dizilerle ayrılmış ve çevrili iki 17-18 bp eşleştirilmiş genomik etiketin olduğu çift uçlu etiketler kütüphanesinde

gerçekleştirilmektedir. Bu tekniğin geçerli okuma uzunluğu amplikon başına 26 baz ve etiket başına 13 baz olup, her etikette 4-5 baz boşluğu bırakmaktadır [110].

İnsan genomunun tüm dizisi 2000 yılına kadar belirlenebileceği düşünülmüştür; bununla birlikte, bu başarı DNA dizileme yöntemlerine olan taleplerde bir azalmaya sebep olmayacaktır. Aslında, DNA sekanslama yöntemlerinin kullanımı, özellikle sekans verisi için maliyeti büyük ölçüde azaltılabilmesi durumunda, önümüzdeki yıllarda artması beklenilebilmektedir.

Poloni sekanslama, bir poliakrilamid jele kovalent olarak bağlanan çok sayıda tekli DNA molekülünün amplifiye edilmesinden ve daha sonra bu amplifiye DNA'nın, tek bir deoksinükleotit ile seri genişletme yoluyla doğrudan poliakrilamide dizilmesinden oluşmaktadır. Bu işlemler Şekil 2.9’da gösterildiği gibi belirli adım aşamalarından oluşmaktadır [111].



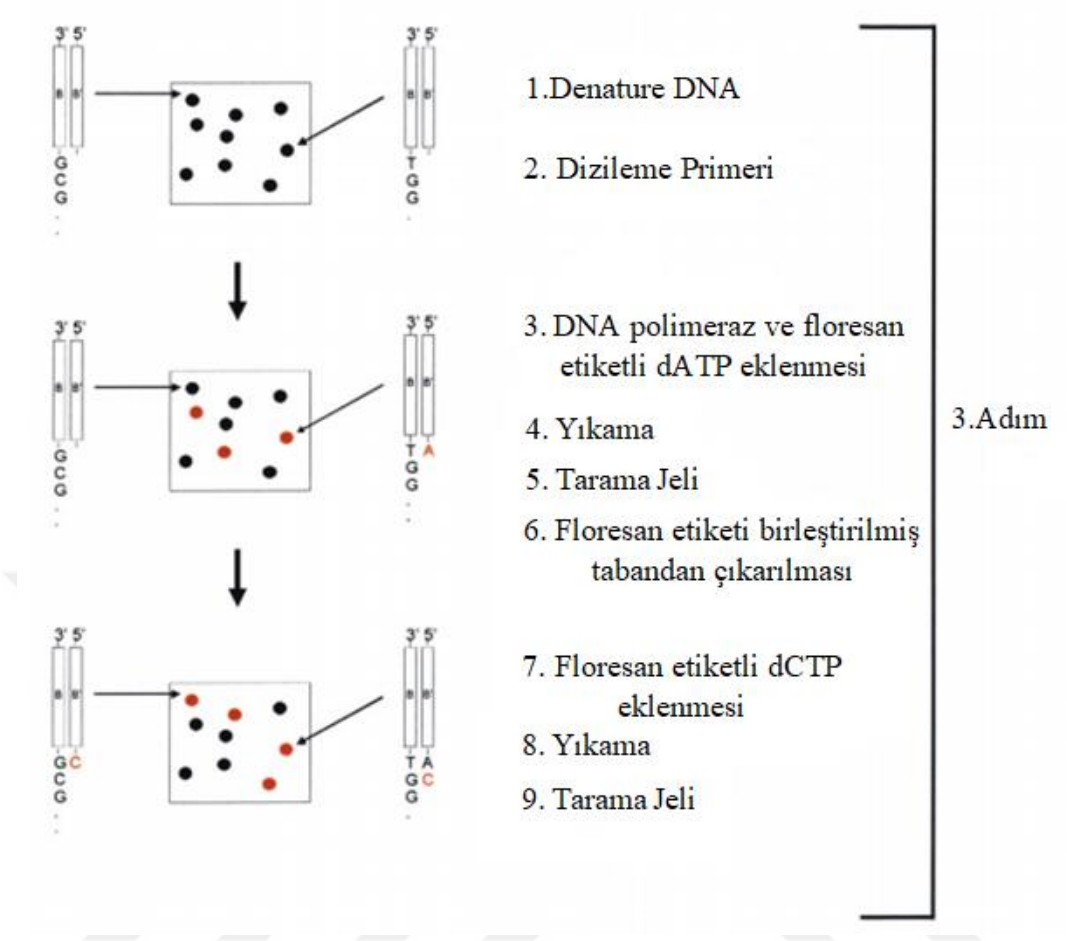
Şekil 2.9. Poloni yöntemi amplifikasyonu

Primer bölgelerini bulunduran doğrusal DNA moleküllerinin kütüphanesi, PCR ile amplifiye edilmesi için bir poliakrilamid jel içine bırakılmaktadır. Tek bir şablon molekülü ile bir polimeraz kolonisi veya polonisi oluşmaktadır. Şekil 2.9'daki adımlar [111, 112]:

1.Adım: Primer bölgeleri ile doğrusal DNA moleküllerinin kütüphanesi oluşturulmasıdır. Kütüphanedeki her bir molekül, iki sabit bölgeyle çevrili değişken bir alan içermektedir. Moleküller bu değişken bölge içinde farklı diziler içermektedir. Sabit bölgeler PCR ile amplifikasyonu sağlamak için primer bağlanma yerleri içermektedir. İlk olarak bu tür kütüphane SELEX deneyleri için oluşturulmuştur.

2.Adım: Akrilamid jel içinde polimeraz polonilerinin (koloniler) çoğaltılmasıdır. Bir cam mikroskop lamının üzerine ince bir poliakrilamid jel dökülmektedir ve polimerize olması beklenmektedir. Bu jel karışımına oligonükleotid primerleri, DNA polimeraz, nükleotid trifosfatlar ve yukarıda tarif edilen doğrusal DNA kütüphanesinin çok seyreltik miktarları (100 ila 5 milyon molekül) içermektedir. DNA, mikroskop lamlarına özel tasarlanmış termal döngüleyicilerden yararlanılarak PCR gerçekleştirilerek büyütülmesi beklenmektedir. Poliakrilamid matrisi reaksiyon aşamasında lineer DNA molekül difüzyonunu geciktirmektedir, bu sayede amplifikasyon ürünleri ilgili şablonun yakınında lokalize şekilde kalmaktadır. Reaksiyonun sonunda, her şablon bir polimeraz 'poloniye' veya koloniye yol açmaktadır. Bir işlemde 5 milyon kadar poloni yükseltilebilmektedir. Primerlerden birinin 5' ucuna bir akridit modifikasyonu dahil edilmektedir, böylece amplifiye DNA, poliakrilamid matrisine kovalent olarak bağlanmaktadır ve daha fazla enzimatik manipülasyonun gerçekleştirilmesine olanak sağlamaktadır.

3.Adım: Şekil 2.10’da gösterildiği gibi polonilerin ardışık, floresan tek bazlı uzantılarla dizilenmesi işlemidir. Öncelikle, hareketsiz hale getirilmiş DNA denatüre edilmektedir, bir iplik yıkanılıp bir evrensel primer şablonuna hibridize edilmektedir. Daha sonra jel üzerine DNA polimeraz ve tek, floresan etiketlenmiş bir nükleotid eklenmektedir. Reaksiyon işlemi birkaç dakika sürmektedir ve sonrasında birleştirilmemiş nükleotit yıkanmaktadır. Daha sonra bir tarama flüoresan mikroskobu yardımıyla jel taranmaktadır. Eğer bir poloni ilave edilen tabanı dâhil edilmişse, tavlanmış primerden hemen 3' şablon tabanının kimliğini ortaya çıkarmaktadır. Floresan daha sonra florofor ve nükleotit arasındaki bağlayıcıyı kimyasal olarak klivaj edilerek ve floroforu yıkanılarak uzaklaştırılmaktadır. Daha sonra döngü işlemine, floresanla etiketlenmiş olan farklı bir baz ilave edilmesiyle, birleştirilmemiş olan nükleotidi yıkayıp jeli tarayarak tekrarlanmaktadır. Bu şekilde, jel üzerindeki her poloni dizi verisi paralel bir şekilde belirlenebilmektedir.



Şekil 2.10. Floresan yerinde sekanslama

Poloniler denatüre edilmektedir ve bir sekanslama primeri tav verilmektedir. Poloniler, tek bir floresan nükleotidin seri bir şekilde ilave edilmeleri ile sekanslanmaktadır [111, 112].

2.2.7. Solexa Dizileme Yöntemi

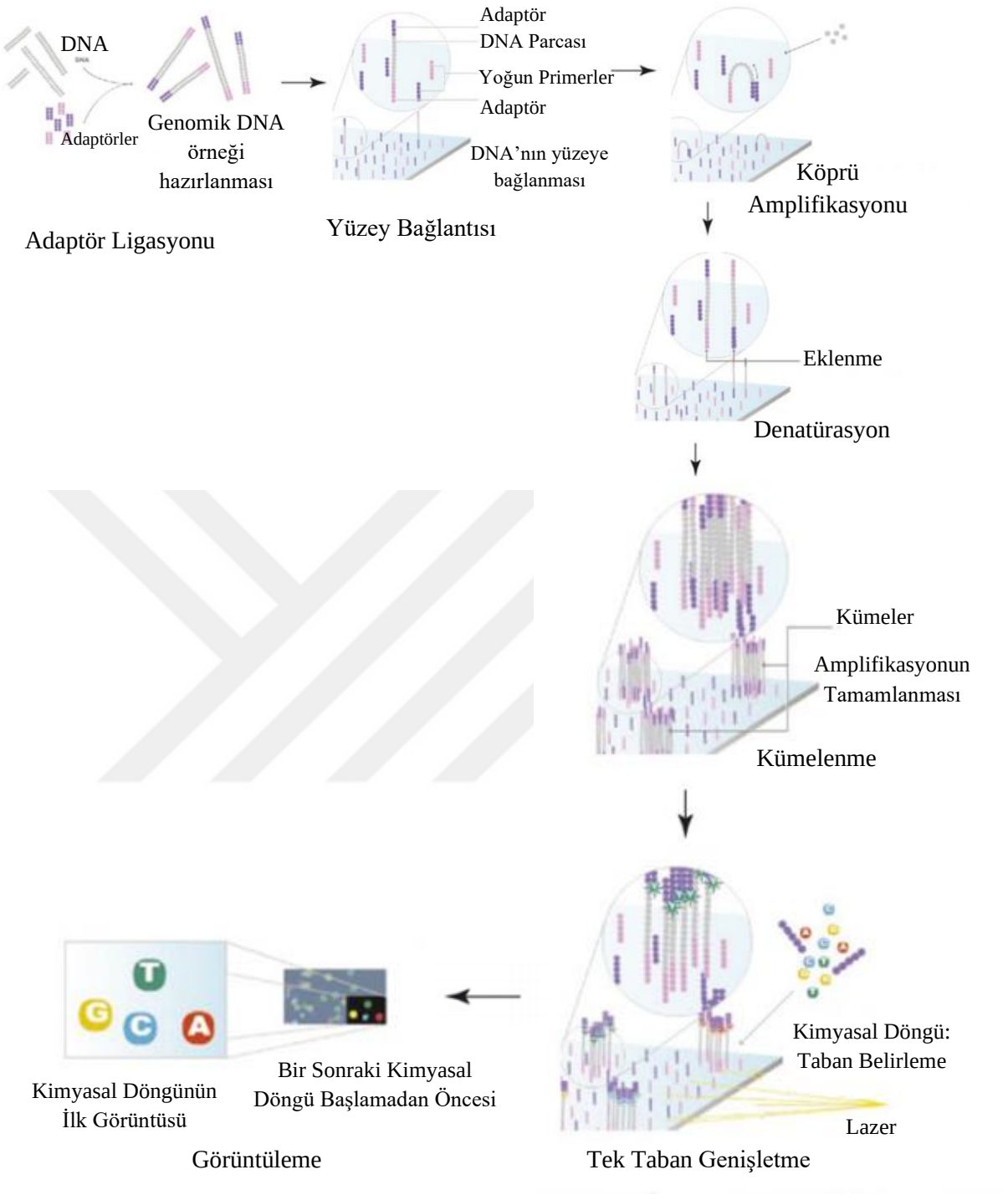
1990'ların ortalarında, Cambridge Profesörleri Shankar Balasubramanian ve David Klenerman, doğanın bir polimeraz sentez DNA'sı adı verilen makinelerinden birini izlemişlerdir. 1997'deki bir dizi tartışma, nükleotitleri katı bir yüzeye adım adım ekledikçe

tek molekülleri görselleştirerek DNA polimerazın hareketini izlemeyi çevreleyen fikirler ortaya çıkarmıştır. Daha sonra sentez teknolojisi ile sekanslama (SBS), sonunda yeni bir DNA sekanslama yaklaşımının temeli haline gelmiştir. 1997 yılında, çift Abingworth (Cambridge girişim kapitalistleri) ile DNA'nın kod çözme hızını ve maliyetini 100.000 kat arttırabileceğini söyleyerek bir araya gelmişlerdir [113].

Balasubramanian ve Klenerman, teknolojilerini daha da geliştirmek için girişim sermayesi şirketi Abingworth Management'ın finansa etmesi ile birlikte 1998 yılında Solexa'yı kurmuşlardır. Solexa büyümeye devam etmiştir ve sonunda Illumina tarafından 2007 yılında 600 milyon £ karşılığında satın alınmıştır. Günümüzde Solexa yeni nesil sekanslama teknolojisi, tüm Illumina sekanslama ürünlerinin merkezinde yer almaktadır. Solexa dizileme çok büyük ölçekli projeleri mümkün kılmıştır ve Solexa-Illumina dizileyecileri 100.000 Genom Projesi, Uluslararası Kanser Genom Projesi ve GenomeAsia 100K gibi önemli projelerin temelini oluşturan teknoloji olmuştur [114].

Solexa sekanslama, sentez (SBS) yöntemi ile sekanslama esas alınarak geliştirilmiştir. Son on yılda otomatik Sanger dizileme yönteminin yerini alan yeni nesil, yüksek verimli sekanslama teknolojilerinden biri olmuştur. Bu yöntemin çalışması DNA sentezinin gerçek zamanlı tespitine dayanmaktadır. Sekanslanacak DNA, uçlara tutturulmuş primerler ile birkaç yüz nükleotitin fragmanlarına (parçalarına) ayrılmaktadır. Parçalar, primer bağlama yerleri ile dolu bir plakaya tutturulmaktadır. Amplifikasyon aşamasında, parçanın yanında yeni bir iplik sentezlenmektedir. Elde edilen çift sarmal yay üzerinde bükülmektedir, böylece ikinci tel yakındaki bir primer bağlantı yerine yapışmaktadır. Çift iplik açılarak sonuç yeni kopyalamaya hazır iki tek zincirli zincir oluşmaktadır. Amplifikasyon fazı ilerledikçe, farklı parçalardan oluşan gruplar oluşmaktadır. Sekanslama adımında, hepsi farklı floresan boyalarla boyanmış, bazlı şablonların yanında yeni bir iplik sentezlenmektedir. Baz takıldığında, cihaz tarafından açıklanan bir floresan sinyali üretilmektedir. Farklı renklerin sinyallerine dayanarak, şablonun dizilenmiş hali çıkarılabilmektedir [115].

Şekil 2.11’de Solexa dizileme yönteminin aşamaları gösterilmiştir [116]. Adaptör Ligasyon ve benzer fragmantasyon adımlarından başlanarak, köprü amplifikasyonu için akış hücresine kütüphane eklenmektedir. Küme fragmanları denatüre edilmektedir, bir sekanslama primeri ile tav yapılmaktadır ve 30 bloke etiketli nükleotit kullanılarak sentez yoluyla sekanslamaya tabi tutulmaktadır.



Şekil 2.11. Solexa sekanslama yöntemi adımları

Tüm genom dizilemesinin yanı sıra, bu sistem çoklu örneklem çalışmaları için verilere ulaşma süresini ve maliyetini önemli ölçüde azaltmaktadır. Kütüphane hazırlığı

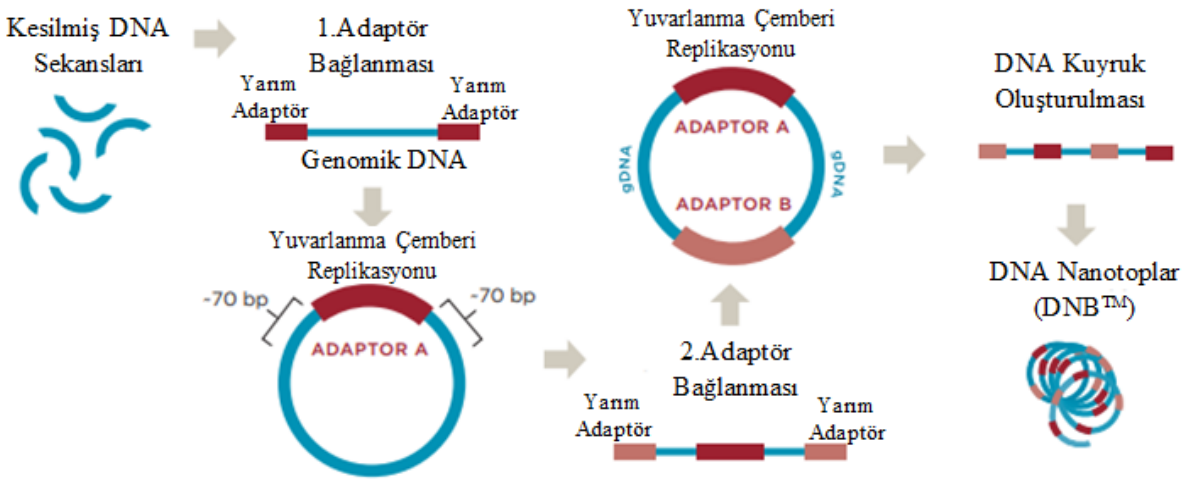
çok zahmetsizdir çünkü sistem klonal saflaştırma gerektirmemektedir. Genom büyüklüğünden bağımsız olarak tüm işlemler saatler içinde tamamlanabilmektedir [116].

Bu teknolojide en önemli aşama veri analizidir, çünkü elde edilen büyük verileri anlamak için gelişmiş bilgisayar sistemleri ve biyoinformatik araçlar gerekmektedir [117].

2.2.8. Nanotop DNA Dizileme Yöntemi

Nanotop Dizileme yöntemi, bir organizmanın tüm genomik sekansını belirlemek için kullanılan yüksek verimli bir sekanslama teknolojisi olarak kullanılmaktadır. Bu yöntemde, küçük genomik DNA fragmanlarını DNA nanotoplarına yükseltilebilmesi için yuvarlanma çemberi replikasyon işlemi kullanılmaktadır. Floresan nükleotitler, tamamlayıcı DNA'ya bağlanmaktadır. Daha sonra DNA şablonunda bilinen sekanslara bağlı olan çapa sekanslarına polimerize edilmektedir. Baz sırası, bağlı nükleotitlerin floresanı yoluyla belirlenmektedir [118]. Bu sekanslama yöntemi için, diğer yeni nesil sekanslama platformlarına göre daha düşük reaktif maliyeti için çok sayıda DNA nanotoplarının dizilenmesi sağlanmaktadır. Bununla birlikte, bu yöntemin bir sınırlaması, okumalarını bir referans genom ile eşleştirmek için zorluklar yaratan sadece kısa DNA dizileri üretmesi olarak nitelendirilmektedir [119].

İstenen DNA'nın izolasyonu ve 400-500 baz çiftine (bp) sahip küçük parçalara dönüştürülmektedir. Adaptör sekansları DNA sekansına bağlanması sağlanmaktadır ve sekansı dairesel parçalara dönüştürmektedir. Dairesel parçalar, her bir parçanın tek sarmallı olan birçok kopyasına sebep olan yuvarlanma çemberi replikasyonuyla kopyalanmaktadır. DNA kopyaları uzun bir iplikçikte başından kuyruğa birleştirilmektedir ve Şekil 2.12'de gösterildiği gibi bir DNA nanotopuna sıkıştırılmaktadır. Nanotoplar daha sonra bir sekanslama hücresine adsorbe edilmektedir. Sorgulanan her bir pozisyondaki floresan rengi, yüksek çözünürlüklü bir kameradan kaydedilmektedir. Biyoinformatik, bir baz çağrısı yapmak ve floresan verilerini analiz etmek için, ayrıyeten 50bp, 100bp veya 150bp tek veya çift uçlu okumaların nicelleştirilmesi veya haritalanması için kullanılabilmektedir [108, 120].



Şekil 2.12. Nanotop DNA dizileme yöntemi

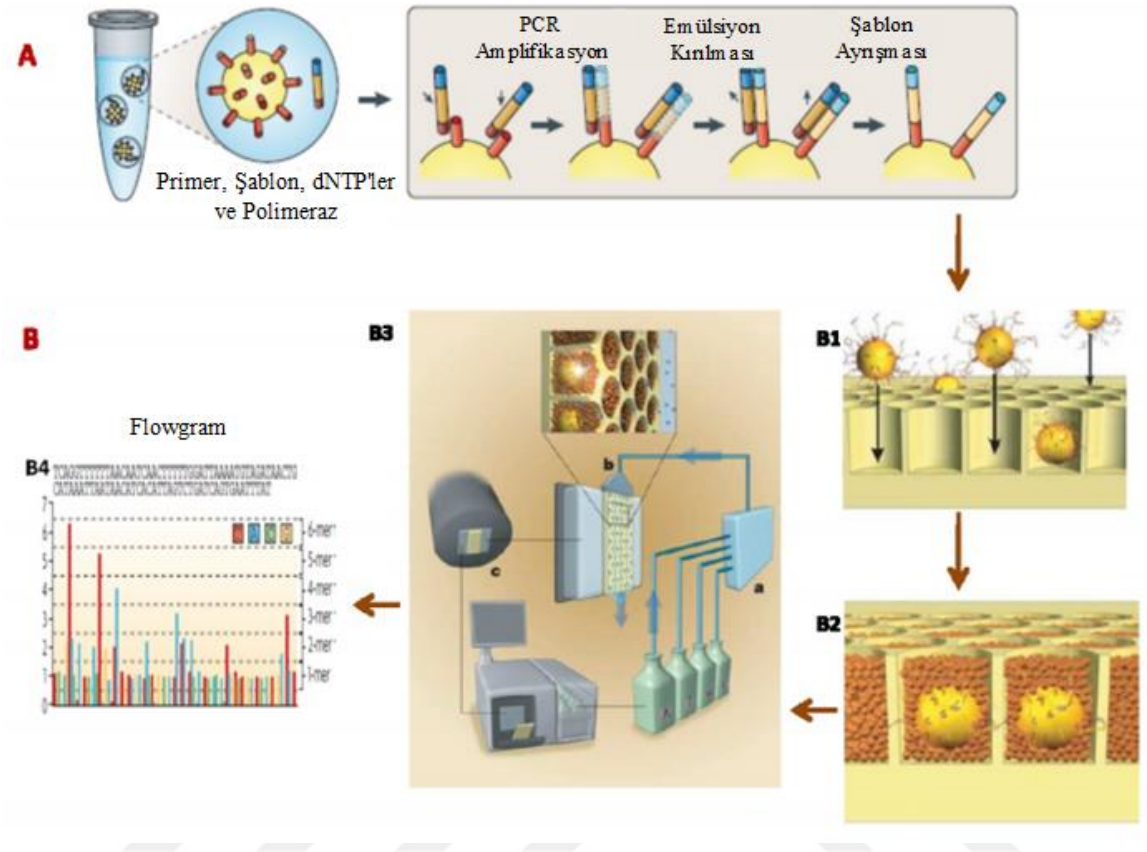
2.2.9. 454 Pirodizileme Yöntemi

454 pirodizilim teknolojisi, 454 Sciences tarafından YND teknolojileri arasından geliştirilen Roche'un bir parçası haline gelmiş ilk teknoloji olmaktadır. Bu yöntemi kullanan ilk makaleden sonraki üç yıl içinde, tüm genom dizileme, metagenomik, transkriptom profili oluşturma, nadir varyant analizi için ultra derin sekanslama ve antik DNA çalışmasını içeren yaklaşık 250 hakemli makale yayınlanmıştır [121]. Günümüzde mikrobiyal çeşitlilik ve metagenomik ile ilgili çalışmalarda 454 Genom Sequencer (GS) FLX sistemini kullanmakta olan yaklaşık 1000'den fazla yayın bulunmaktadır [122].

454 dizileme prensibi, inorganik pirofosfatın (PPi) salınımını orantılı bir şekilde elektroforetik olmayan görünür ışığa dönüştürülmesi ile ölçülebilen, biyolüminesans bir yöntem olan bir sentezleme işlemi yapan Pyrosequencing'e dayanmaktadır. Bu işleme birçok enzimatik reaksiyon işlemi ile ulaşılmaktadır. Sekanslanacak DNA fragmanı, DNA polimeraz, ATP sülfürilaz, lusiferaz ve apiraz oluşan dört enzim karışımı, eşsiz bir deoksiribonükleosid trifosfat (dNTP) bozan enzim ile inkübe (kuluçka) edilmektedir. Bir dNTP'nin uzanan DNA fragmanına bağlanması üzerine, sülfürilaz ile adenosin trifosfata

(ATP) dönüştürülen PPi salınmaktadır. Bir lusiferaz kataliz reaksiyonu ATP'yi, şarj bağlı cihaz (CCD) kamera gibi ışığa duyarlı olan bir cihaz aracılığıyla ölçülen ışığa dönüştürmektedir [123].

Yeni nesil dizileme teknolojileri dizileme için iki aşamalı bir yöntem içermektedir. İlk adım olarak, Sanger sekanslama yöntemi ile ihtiyaç bulunan bakteriyel klon kütüphanelerinin zaman alıcı üretimini yapabilen şablonların paralel toplu klonal amplifikasyonu şeklinde bilinmektedir. 454 sekanslama ile klonal amplifikasyon, yağ emülsiyon bazlı bir PCR reaksiyon karışımı şeklinde belirtilmiştir (emPCR). EmPCR, kesilmiş DNA parçaları kütüphanesinde, 5' ve 3' uçlarına bağlanabilen adaptör sekansları ile ya da amplifikasyon kütüphanesi esnasında amplifikasyon sekansı ilave edilmiş amplikon kütüphanelerinde oluşturulmasıdır. EmPCR ve primer tasarımı için 16srDNA amplikon kütüphanesinin oluşturulması (4.2.2.1) 'de tarif edilmektedir [124]. Adaptör sekansları, DNA parçalarının sekanslama ve emPCR için kullanılan taneciklere bağlanmasına olanak sağlayan 454 oligonükleotit sekansları şeklinde nitelendirilmiştir. Sınırlayıcı koşullar, her bir tanecik için tek bir DNA molekülünün bağlanmasını kolaylaştırması için yapılmaktadır. Her bir tanecik, bir yağ emülsiyonu içindeki özel bir sulu PCR reaksiyon karışımı damlacıklarında bölümlere ayrılmaktadır. Bu, tek tek moleküllerin kontaminasyon veya rakip sekanslar olmadan amplifiye edilmesine izin vermektedir ve her bir tanecik üzerinde milyonlarca klonal olarak amplifiye edilmiş sekans şablonları üretmektedir. İkinci adımda, 454 dizileme makinesi tarafından ayrı olan taneciklerin paralel dizilemesi gerçekleştirilmektedir. İşlem dört bölümden oluşmaktadır: akışkan bir düzenek, iyi içeren fiber optik slayt ile birlikte bir akış hücresi, bir CCD kamera tabanlı görüntüleme düzeneği ve bir bilgisayar şeklinde belirtilmiştir. Yaklaşık 1.6 milyon kuyu alanı bulunduran fiber optik slayt, taneciklerle yüklenmektedir ve içinde bulunan sekanslama reaktiflerinin bir akış hücresi alanına monte edilmektedir. Akışkan alt sistem, sekanslama reaksiyonu sırasında kuyu boyunca sekanslama tamponunu ve 4 nükleotidi sırayla sabit bir sırada (GS FLX titanyum serisi-200 devir) iletmektedir. Nükleotitlerin akışı sırasında, her bir oyuktaki şablon şeridine tamamlayıcı nükleotitler, sentezlenen yeni DNA iplikçiklerine dahil edilmektedir ve kamera, kemilüminesans reaksiyonu tarafından üretilen ışığı yakalamaktadır. Bu bilgiler ışığında bilgisayarda DNA baz moleküllerine çevrilmektedir. Şekil 2.13'te 454 pirodizileme yönteminin tüm akış şeması gösterilmiştir [122, 125].



Şekil 2.13. 454 Pirodizileme yöntemi akış şeması

Şekil 2.13'teki 454 sekanslama akış şeması: (A), uçlara tutturulmuş adaptörlere sahip tek bir DNA molekülünün (mavi ve kırmızı uçlu sarı çubuk) paralel klonal amplifikasyonuna izin veren ayrı bir mikro-reaktör oluşturan bir yağ emülsiyonu içinde PCR reaksiyon karışımı ile birlikte mevcut olduğu emPCR'yi göstermektedir. Her tanecik, amplifikasyon işleminden sonra benzersiz olan DNA şablonunun milyonlarca kopyasını içermektedir. Daha sonra emülsiyon kırılmaktadır ve DNA denatüre edilmektedir (B) B1 Fibreoptik lam üzerine bırakılan ssDNA klonlarını taşıyan her boncuk şeklindedir. Pirofosfat dizilemesi için gerekli olan hareketsizleştirilmiş enzimleri taşıyan B2 küçük tanecikler her oyuğa bırakılmaktadır. B3, sıvı düzeneği sistemiyle (a) 454 sekanslayıcıyı, sekans slaytını (b) içeren akış hücresi odasını, görüntüleme (C) için CCD kamerayı ve bilgisayar sistemini göstermektedir. B4, bir nükleotidin ilave edilmesiyle orantılı bir şekilde üretilen ışık sinyalinden üretilmiş akış diyagramını göstermektedir [122, 125].

454 sekanslama yöntemi, mikrobiyal topluluğun hastalık ve sağlıktaki yapısını, işlevini ve sistemsel dinamiklerini anlamak için yaygın olarak kullanılmıştır. Pyrosequencing tabanlı insan mikrobiyom çalışmalarının bazı örnekleri, bağırsak mikrobiyomunun artan enerji verimi kapasitesi ile karakterizasyonu ve obezite ile ilişkisi, antibiyotik tedavisinin sağlıklı bireyler arasındaki vajinal mikrobiyal topluluktaki bağırsak mikrobiyom varyasyonu üzerindeki etkisi, cilt mikrobiyomunun bireyleri arasındaki varyasyonlar, vücut bölgelerinde, sağlık ve periodontitte bakteriyel topluluk profillerinin karşılaştırılması ve sağlık ve hastalıkta hava yolu mikrobiyomunun karakterizasyonu gibi kullanım alanları bulunmaktadır [126, 127].

2.2.10. Solid Dizileme Yöntemi

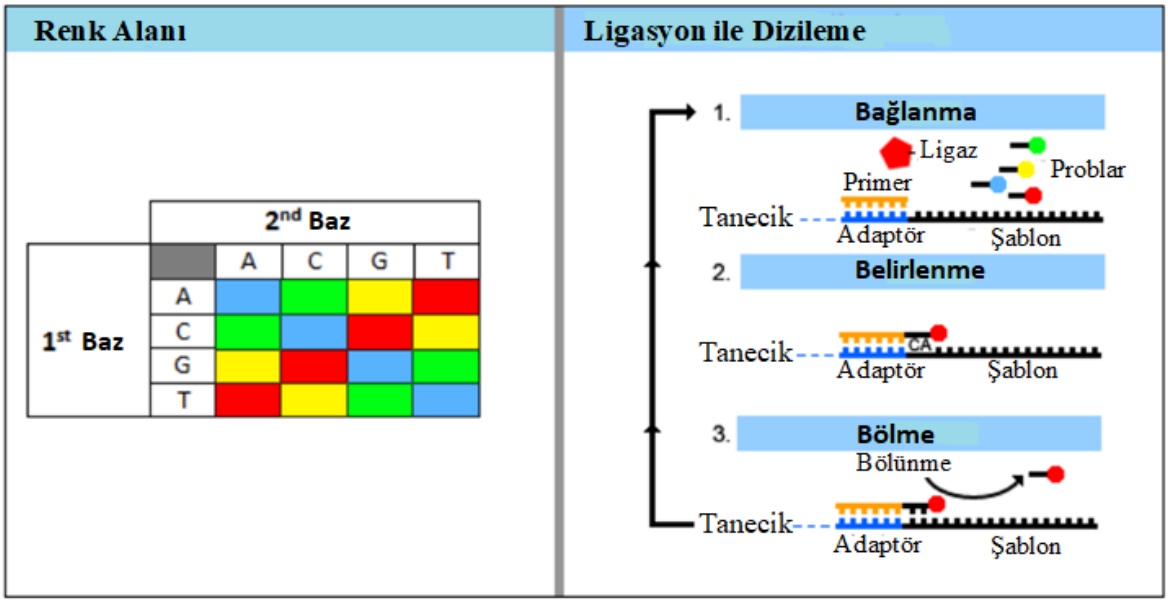
Harismendy ve arkadaşları tarafından 2009 yılında yürütülen yüksek verimli YND teknolojilerinin erken karşılaştırmaları, ABI SOLID ve Illumina sıralayıcıları gibi platformların doğruluk ve verim açısından son derece benzer olduğunu ileri sürmüştür. Bu çalışma, sekanslama platformu seçimini etkileyen ana faktörlerin zaman, maliyet ve bulunabilirlik olduğunu ortaya koymuştur. Bununla beraber, her sistemin bazı uygulamalar için küçük avantajları olduğu belirtilmiştir. ABI SOLID sistemi daha düşük kapsama değişkenliği ile düşük kapsama ortamında daha yüksek doğruluk ortaya çıkardığı gösterilmiştir. Bu durum, çok daha düşük ekspresyona sahip olan transkriptlerin, ncRNA'lar gibi nadir durum intergenik transkriptlerin karakterizasyonu ve tanımlanması için özellikle önemli bir durum şeklinde nitelendirilmiştir [128].

Applied Biosystems'in (ABI) SOLID (Oligo Ligasyonu ve Tespiti ile Sekanslama) sistemi 2007 yılında piyasaya sürülmüştür [128, 129]. Bu sistem için, floresan etiketlenmiş yarı dejenere olan oligonükleotid problarının dizili ligasyonuna dayanmakta olan benzersiz bir sekanslama yöntem metodolojisi sunmaktadır. Her prob, tek seferde iki bitişik baz pozisyon sorgulaması yapmaktadır ve dört floresan boya yardımıyla olası 16 di-baz kombinasyon kodlaması yapılmaktadır. Ligasyon aşaması sonrasında, bir floresan değeri o anki boyanın rengini kaydetmektedir. Daha sonra floresan grup, ligasyonlu oligonükleotid probu içerisinden kimyasal bölünme işlemi ile çıkarılmaktadır ve sonrasında müteakip bir

ligasyon döngüsüne izin verilmektedir. Di-baz problarının kullanımına "2 baz kodlama" denilmektedir ve SOLID sisteminin yüksek hassasiyetine katkıda bulunmaktadır. Ligasyon, algılama ve dilinim döngüsü, nihai okuma uzunluğunu belirlemektedir. Bir dizi ligasyon döngüsünü takiben, uzatma ürünü çıkarılmaktadır ve şablon, 2. tur ligasyon döngüleri için n-1 pozisyonuna tamamlayıcı bir primer ile sıfırlandırılmaktadır. Her sekans etiketi ve renk uzayında üretilen son sekans için toplam 5 döngü gerçekleştirilmektedir. Bu dizileme yöntemi, diğer sistemlere kıyasla çok sayıda avantaj sağlamaktadır [130]:

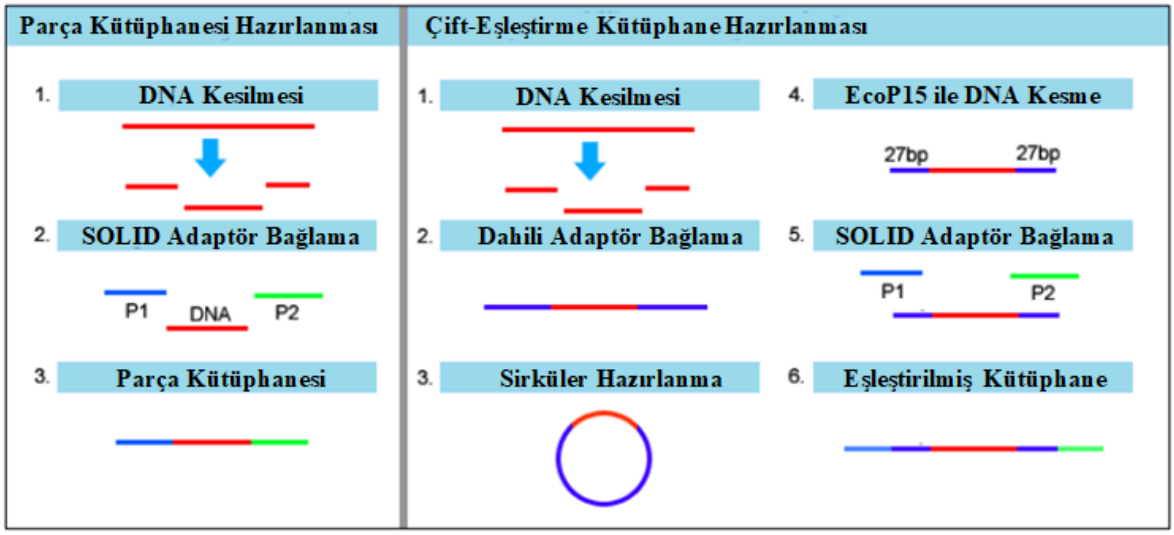
- Her ligasyonda iki bazın sorgulanması, daha fazla özgüllük sağlamaktadır.
- Her baz için iki kez sorgulama işlemi yapılarak, daha fazla güven sağlanmaktadır.
- Primer beş bağımsız uzatma turu için sıfırlanmaktadır ve sinyal-gürültü oranları iyileştirilmektedir.
- 16 olası iki baz kombinasyonunu kodlayan dört boyanın tasarımı, dahili hata kontrolünü mümkün kılmaktadır.

Sekanslama işlemi ve renk alanı formatı Şekil 2.14'te detaylandırılmıştır [130].



Şekil 2.14. Ligasyon ile renk alanı formatı ve solid sekanslama

Solid sistemi, renk alanında diziler oluşturmak için bir dizi yarı dejenere ve floresan etiketli di-baz prob kullanılmaktadır. Şekil 2.14'te di-baz sekansları renk alanına dönüştürmek için bir anahtar, sol bölmede gösterilmiştir. Solid sekanslama yöntemi Şekil 2.14'te sağ bölmede gösterilmiştir. Primerlerin kütüphane şablondaki P1 adaptör sekansına hibridizasyon işlemi ile başlamaktadır. Daha sonra dört prob, sekanslama primerine ligasyon amacıyla rekabet etmektedir. Bağlı olan prob flüoresan etiketi Solid tarafından tespit edilmektedir, daha sonra problar bölünme yoluyla çıkarılmaktadır ve işlem şablondaki bir sonraki baz için tekrarlanmaktadır. Di-baz probunun özgüllüğü, her ligasyon reaksiyonunda her 1. ve 2. bazın sorgulanmasıyla elde edilmektedir. Solid Dizilemesi için kütüphanelerin hazırlanması Şekil 2.15'te gösterilmektedir [130].



Şekil 2.15. SOLID parça kütüphanelerinin oluşturulması

Solid sekanslama sistemi iki ayrı tür kitaplık kullanabilmektedir; bunlar parçalanmış (sol bölme) ve eşleştirilmiş (sağ bölme) şeklindedir.

2.2.11. İyon Yarı İletken Dizileme Yöntemi

İyon Torrent teknolojisi 2010 yılında ticarileştirilmişti ve sekanslama makineleri 2012'de diğer birçok sekanslama teknolojisinden daha düşük bir maliyetle yaklaşık 80.000 dolara gerçekleştirilmektedir. İyon yarı iletken dizileme Kişisel Genom Makinesi (PGM), DNA dizilemesi için İyon Proton Sistemi yarı iletken yongaları ile modüler olarak yükseltilebilen bir laboratuar makinesi olarak geliştirilmiştir. Başlangıçta, sekanslama çalışmaları 100bp uzunluklarda okumalar üretmiştir. Son yongalar 400bp uzunlukta okumalar üretebilmektedir. Illumina ve Pacific Biosciences'dan farklı olarak iyon yarı iletken dizileme teknolojisi, baz arama için hidrojen iyonu algılama özelliğini kullanırken, diğer teknolojiler ışık algılama özelliğini kullanmaktadır. Çip makinesi, reaksiyon kimyasının meydana geldiği substrat ve kuyuların yanı sıra hidrojen iyonu voltaj sensörü için elektronikler içermektedir. İyon yarı iletken dizileme teknolojisi, bir tanecik üzerindeki kısa DNA parçalarını hareketsizleştirerek çalışmaktadır. Tanecik, DNA parçasının tanecik üzerinde emülsiyon PCR ile büyütüldüğü bir kuyu içine yerleştirilmektedir. Bir çip üzerinde, büyük ölçüde paralel sekanslamaya izin veren milyonlarca kuyu bulunmaktadır. DNA fragmanları polimeraz ile kopyalanmaktadır ve kopyaya her nüklelotid eklendiğinde hidrojen iyonları salınır hale gelmektedir. Kuyu, altta bir elektrik voltajı yoluyla iyon konsantrasyonunu algılayan bir sensöre sahiptir. Bir taban tanecik üzerinde yıkanırsa ve fragmanlara dahil edilmezse, hiçbir iyon açığa çıkmamaktadır ve elektrik voltajı sıfır olmaktadır [131].

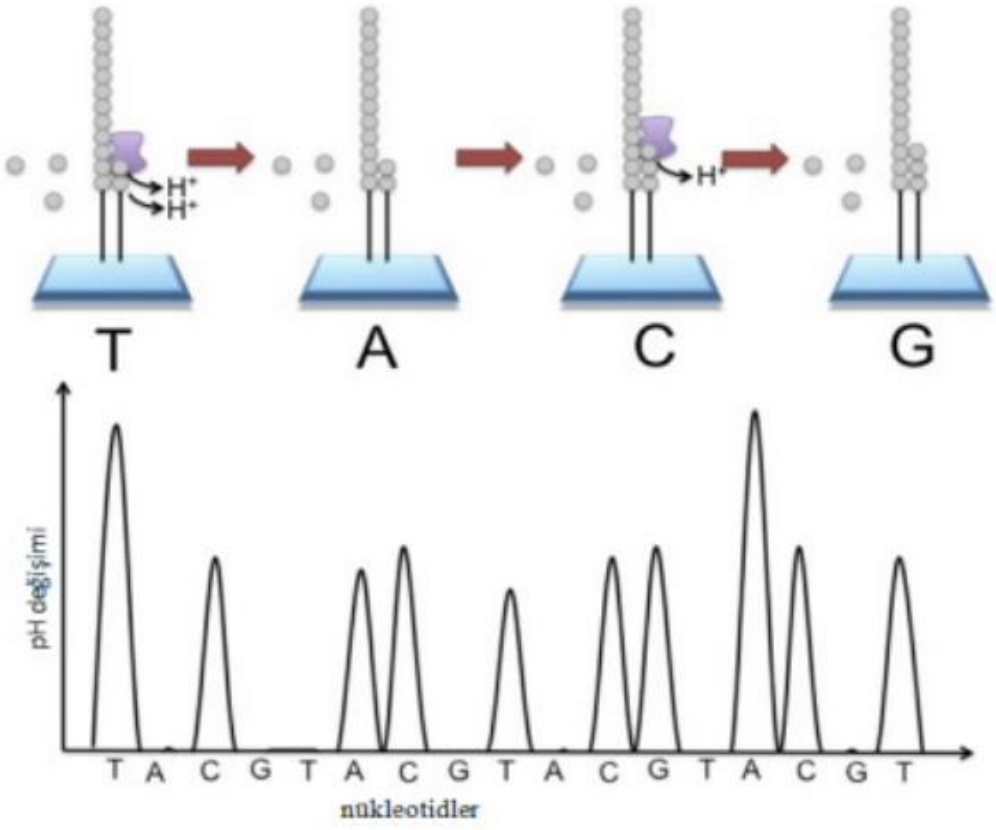
Sekanslama işlemi aşağıdaki adımlardan oluşmaktadır [132, 133]:

a) Doğal yaşamda, bir polimeraz tarafından bir nükleotid bir DNA şeridine ilave edildiğinde, bir hidrojen iyonu yan ürün şeklinde salınmaktadır.
b) İyon yarı iletken dizileme sistemi, bu biyokimyasal işlemi büyük oranda paralel bir şekilde gerçekleştirebilmek için yüksek yoğunluk değerinde bir mikro işlenmiş kuyu dizisinden faydalanmaktadır. Her kuyunun farklı DNA şablonu bulunmaktadır. Bu kuyucukların altında özel bir iyon sensörü ve iyon duyarlı bir tabaka bulunmaktadır.
c) Bir DNA şablonuna nükleotit dâhil edilirse ve sonrasında bir DNA zincirine eklenirse, bir hidrojen iyonu salınımı gözlenmektedir. Bu iyondan gelecek olan

yük, özel bir iyon sensörü yardımıyla tespit edilebilen çözeltisinin pH'ını değiştirebilmektedir. Dizileyici, kimyasal bilgiden dijital bilgiye doğrudan geçerek baz aramaktadır.

d) İyon Kişisel Genom Makinesi (PGM) dizileyeci çipin art arda bir nükleotid ile ardışık olarak taşmasına neden olmaktadır. Eğer bir sonraki çipi dolduran nükleotid bir eşleşmesi yok ise, herhangi bir baz çağrılmamaktadır ve voltaj değişikliği de kaydedilmemektedir.
e) DNA zincir yapısında iki özdeş baz var ise, voltaj iki katı artmaktadır ve çip olarak bu iki özdeş baz kaydedilmektedir.

Şekil 2.16'da İyon yarı iletken dizileme yöntemi gösterilmiştir [134].



Şekil 2.16. İyon yarı iletken dizileme yöntemi

İyon yarı iletken dizileme teknolojisinin düşük maliyetli olmasına rağmen, bazı dezavantajları nispeten düşük verim, homopolimerlerde hataya eğilimli ve yıkama hatalara neden olabilmektedir [135].

2.3. DNA Dizileme Yöntemleri Analizi ve Genel Bakış

DNA dizileme yöntemleri, son 40 yılda birçok devrimsel ilerlemeye tanık olmuştur. Sanger ve arkadaşları tarafından geliştirilmiş olan dideoxy yöntemi, genomik alanlarında gelişim açısından büyük bir yol gösterici olarak görülebilmektedir. Genel kullanımlarda hala altın standart olarak kullanılan bu yöntem, sekanslama yöntemlerindeki görünümü değiştirmiştir. Sonralarda insan genom projesinin tamamlandığının duyulması, yeni nesil dizileme teknolojileri olarak adlandırılan yüksek verimli ve hızlı dizileme platformlarının bir çağını açmıştır. İlk YND yaklaşımlarından olan pirosequencing'i, terminatör kimyasal bazlı dizileme ve ligasyon temelli sekanslama metodolojileri izlemektedir. Çok geçmeden, yeni bakış açılarına bağlı olarak üçüncü nesil dizileme sistemleri tanıtılmıştır. Üçüncü nesil dizileme yöntemlerinde, tek molekül gerçek zamanlı dizileme, tek molekül floresan sekanslama ve yarı iletken sekanslama yöntemi gibi sistemler sayesinde kullanıcılara yeni fırsatlar sunmuştur. Ayrıca, nanopore sekanslama yöntemi gibi yeni nesil teknolojileri DNA, RNA veya proteinler üzerinde daha verimli analiz yapabilme olanakları sağlamaktadır. Dördüncü nesil sekanslama yöntemleri, dokularda ve sabit hücrelerde yerinde sekanslamaya bağlı alanlarda büyük katkılar sağlanması beklenilen yeni yöntem ve çok spesifik uygulama alanı şeklinde bilinir hale gelmektedir. Mevcut sekanslama platformları sürekli olarak güncellenirken, daha ucuz, daha hızlı ve daha doğru sekanslama teknolojileri oluşturmak için araştırma ve yenilik çalışmaları hız kesmeden gerçekleştirilmektedir. Günümüzde dünyanın birçok farklı ülkelerinde nanopore dizilime sistemine dayanan birçok yeni teknoloji ve sistemler geliştirilmektedir. Nanopore sıralama alanında çalışan şirketler arasında Base4, Genia, INanoBio, Nabsys, Noblegen Biosciences, Oxford Nanopore Technologies, Quantapore ve Quantum Biosystems şeklinde listelenebilmektedir. Metodolojilerini ve tanımsal kavramlarını kısaca açıklamak için, örnek olarak Nabsys, yarı iletken tabanlı olan nanodetektörlerin etiketli bir DNA'nın

doğrudan elektriksel tespiti yapılabilmesi için kullanıldığı bir yönteme dayanırken, Genia optik tespit sistemlerine ve biyolojik membranlara bağlı olarak bir platform sunmak için hazırlanmaktadır. INanoBio sistemi teknolojisi, seçiciliği ve duyarlılığı artırabilmek adına Tamamen Boşaltılmış Üstel Olarak Birleştirilmiş (FDEC) Alan Etkili Transistör (FET) olarak nanotel sensörleri yardımıyla sinyal düşüşünü engelleyebilen bir nanopore teknolojisini bulundurmaktadır. Ayrıyeten, Quantum Biosystems firması, nanopore teknolojisini bir elektron mikroskop türü olan tünel tip elektron dedektörü ile birleştirilmesini sağlayan bir yaklaşım kullanmaktadır. Diğer taraftan, Base4, su-yağ emülsiyonu ile ayrılan ve ayrılan tek nükleotitlerin saptanması için kademeli şekilde kimyasal reaksiyonlar kullanmaktadır. Noblegen Bioscience firması, 2014 yılında optipore adında bir sistemi geliştirmeyi amaçlamıştır. Ayrıca, Quantapore nanopore tabanlı optik okuma yöntemi geliştirmektedir. Şirketlerin Ar-Ge çalışma faaliyetlerine ek olarak, devlet kurumları ve vakıfların son derece önemli hibeleri yardımıyla da yeni nesil dizileme teknolojilerinin geliştirilebilmesi adına birçok grup desteklenmektedir. Örneğin, sadece 2014 yılında NHGRI, genom sekanslama maliyetlerinde (1000 ABD dolarından az) önemli bir düşüş bekleyen ve yeni ilerlemelerin uygulanmasına izin vermek için uygulama alanlarını genişleten yeni sekanslama tekniklerinin geliştirilmesi için araştırma gruplarına yaklaşık 14,5 milyon ABD doları sağlamıştır. Bu alandaki resmi ve ticari yatırımların devam eden yıllarda artması ve böylece araştırma çalışmalarının piyasaya daha iyi platformlar getirmesini sağlaması beklenmektedir. 1000$ genom projesinin bir üyesi olan NHGRI 2004 yılından beri, bu yönde çalışması bulunan birçok grubu desteklemiştir. Bu nedenle, 2001 yılında açıklanan insan genom projesi için 3 milyar dolar olan genom dizileme maliyeti önemli ölçüde azaltılmış ve günümüzde yaklaşık 1000 dolar sınırı test edilmiştir. Öte yandan, DNA dizileme maliyetlerinin azaltılması için 2012'den beri 2007-2012 dönemine göre kıyaslanınca bir yavaşlama gözlenmektedir. Ayrıca, piyasaya sunulacak yeni teknolojilerin önümüzdeki yıllarda sürekli olarak maliyetlerin sürekli azalmasına ve veri üretim seviyelerinin artmasına neden olsa da, tüm YND teknolojileri için dizileme doğruluğunda ki tıkanıklıklar çözülmeye devam etmektedir. Buna ek olarak, verimlilikteki önemli artış, daha pratik çözümler bulmak için zaman alabilecek yeni engelleri de getireceği ön görülmektedir, fakat yeni alternatifler sunabilmek için bu konulara yönelik birçok çalışma yürütülmektedir. RNA ve DNA dizileme alanlarında yeni nesil sistemler, genomik alanındaki neredeyse her ihtiyacı karşılamak için geliştirilmektedir. Tüm genom, transkriptom, ekzom, metagenomik, metilom, ChIP, de

novo, küçük RNA, yeniden sonuçlandırma tabanlı uygulamaları yaşam bilimlerinde yaygın şekilde kullanılmaya başlanmıştır ve sonuçlar hem araştırmalarda hem de klinikte birçok kavramı değiştirmiştir. İkinci nesil dizileme yöntemleri genomik çalışmalardaki ihtiyaçların çoğunu cevaplandırmış olsa bile, üçüncü ve dördüncü nesil dizileme teknoloji platformları daha geniş çaplı uygulamalar ile doğru ve daha pratik çözümler için bir potansiyel olarak gösterilmektedir. Günümüzde DNA dizileme yöntemleri, NHGRI tarafından yaklaşık 15 yıl öncesinde belirtilmiş olan altın standartlarına neredeyse çok yakındır, fakat hala çözülmesi beklenilen önemli birçok konu bulunmaktadır. Genom Dizilemesi için 1000 $ eşik aşılmak üzeredir ve mevcut YND sistemleri ile çok daha kısa sürede (saat cinsinden) yüksek verim elde etmek daha kolay hale gelmiştir. Ayrıca, dizileme yöntemi teknolojilerinin uzun okuma probleminin tek moleküllü gerçek zamanlı sıralama yöntemleri ve nanopore gibi teknolojiler ile yakın gelecekte daha iyi ve kullanışlı bir çözüme ulaşacağı söylenebilmektedir. Doğruluk sorunu tüm yeni geliştirilen ve geliştirilmekte olan teknolojiler için en çok önem verilen konu olarak nitelendirilmektedir. Bu konuda da yüksek seviyeye atlayabilme amacıyla önemli bir değişiklik yapılması için devrim niteliğinde bir ilerleme olması düşünülmektedir. Bu doğrultuda DNA sekanslama yöntemleri, inanılmaz bir hızla gelişmektedir. Ayrıca hem bu alandaki araştırmalar hem de rutin uygulamalar için büyük yenilikler ön görülmektedir. Çok çeşitli alanların uygulama, depolama ve biyoinformatiğine yönelik yeni dizileme teknolojilerinin getirdiği önemli kolaylığın yanı sıra, platforma özgü "şu anda çözülmemiş" sorunlar için gerekli çalışmalar da yapılmaktadır. Şimdiye kadar sekanslama teknolojilerinin gelişim hızı göz önüne alındığında, yeni platformların, nesillerin ve çözümlerin, yakın gelecekte mevcut sorunların çözümü için alternatifler olarak çok daha hızlı görüneceği söylenebilmektedir [136].

2.4. DNA Dizileme Hata Düzeltme

DNA dizilemede, sekanslama hataları olarak bilinen hata yapma eğilimleri bulunmaktadır. Yaygın hatalar arasında yanlış baz birleşimi (tek bir baz değişikliğine neden olmaktadır), polimeraz kayması (eklemelere veya silmelere neden olmaktadır) ve

yanlış baz ilişkisi eşleştirmesi bulunmaktadır. Okumalardaki dağılımlar, okuma dağılımı ve dizileme hataları oranı platformlara göre değişiklik göstermektedir. Kayma gibi dizileme hataları genellikle diziye bağlıdır ve düşük karmaşıklıkta tekrarlanan dizinin hatalara neden olma olasılığı daha yüksek olarak belirlenmiştir. Örneğin, Illumina Genom Analizörü (GA) platformunda GGC sekansları ve 8bp'den uzun ters çevrilmiş tekrarlar, hataların en az %80'inini oluşturmaktadır [137]. Illumina MiSeq 2x300bp platformunda, indel (genoma bazların eklenmesi veya silinmesi) hataları yer değişme hatası oranının yüzde biri oranında meydana geliyor ve bunların %95'i okumaların ilk 10bp'sinde meydana gelmektedir. Illumina verileri, okumaların sonuna doğru artan hata oranlarını ve bir çiftin ikinci okumasında daha yüksek hata oranlarını göstermektedir [138].

Sekanslama hataları, genom derlemelerini bozan belirsizlikler, kısaltılmış bitişikler, yanlış bazlar ve yanlış birleştirilmiş bitişikler ile sonuçlanabilmektedir. Dizileme hataları ayrıca yanlış pozitif varyantlara ve hatalardan ayırt edilemeyen düşük frekanslı somatik mutasyonlara neden olabilmektedir [139].

Hata düzeltme, ilk önce Pevzner ve ark. Tarafından spektrum hizalama problemi olarak adlandırılan örtüşen okumaların fikir birliğini kullanarak okumalardaki dizileme hatalarını düzeltme işlemi olarak nitelendirilmiştir [140]. Bunu yaparak, yanlış pozitif varyant sayısı azaltılabilmektedir ve genom derlemeleri geliştirilebilmektedir. Birçok çevirici, düzeltilmiş okumalar veren kendi hata düzeltme adımlarını içermektedir. Genellikle tek temel hataları düzelten hataya odaklanılmıştır. Birkaç bağımsız hata düzeltme programı da geliştirilmiştir. Quake, tek temel yer değiştirme dizileme hatalarını düzeltmek için Jellyfish kmer sayacını kullanan mevcut bir araçtır. Quake, Illumina okumalarıyla kullanılmak üzere tasarlanmıştır. Okuma kalitesi puanlarını kullanmaktadır ve nükleotidi, nükleotid hata oranlarını öğrenmektedir [139, 141].

2.4.1. Hata Düzeltme Algoritmaları

Yeni nesil dizileme teknolojilerinde baskın olan iki tür hata bulunmaktadır; yer değiştirme hataları, tek bir taban farklı bir tabana değiştirildiğinde ve indeller, eklenen

veya silinen bir okumadaki bazlar olarak nitelendirilmiştir. Verilerdeki okuma hatalarını düzeltebilmek, YND verilerini kullanan uygulamaların performansını oldukça artırabilmektedir. Illumina platformu ağırlıklı olarak yer değiştirme hataları yapmaktadır ve hataları düzeltmek için geliştirilen baskın teknoloji olduğundan, yer değiştirme hatalarına odaklanmaktadır. Yer değiştirme hatalarını düzeltebilen en başarılı hata düzeltme programlarından bazıları: BLESS, Coral, HiTEC, RACER ve SHREC.

Yer değiştirme hatalarını düzeltebilmek için üç temel yaklaşım kullanılmaktadır; bunlar çoklu sekans hizalama, k-mer spektrumu ve k-mer sayımı yöntemleri şeklinde belirtilmiştir. Çoklu sekans hizalama yöntemi, benzer çoklu okumaları belli bir sıra doğrultusunda hizalamaya ve her bir okuma için hizalanmış olan verilerin en çok bulunan taban değerine göre diğer hizalı sekansların okumalarını düzeltmeye dayanmaktadır. Coral, okumalardaki hataları düzeltebilmek için çoklu indeks sekans hizalama yöntemini kullanmaktadır. K-mer spektrum algoritmaları, bir okumadaki k-merleri veri kümesi içinde en çok görünenlere göre diğer sekans verilerini bu düzeye çıkararak okumalardaki hataları düzeltmektedir. BLESS programı hataları düzeltmek için k-mer spektrum yaklaşımını kullanmaktadır. K-mer sayma algoritmaları ise, her bir k-mer'in bulunduğu veri kümesinde kaç defa göründüğünü saymaktadır ve bir okumadaki düşük frekanslı k-merleri veri kümesi içinde görünen değere göre düzeltmektedir. HiTEC, SHREC ve RACER, okuma hatalarını düzeltebilmek için k-mer sayma yöntemini kullanmaktadır [142].

Tüm hata düzeltme programları, hataları bulmak ve düzeltmek için bir okumada “k-mer” kullanmaktadır. Tüm okuma hatası düzeltme algoritmaları ve programları, daha doğru düzeltme işlemi yapabilmesi için belirli bir kapsama alanına ihtiyaç duymaktadır.

G: Dizilenen genomun uzunluğu,

N: Veri kümesindeki toplam baz sayısı,

Şeklinde belirtilen ifade doğrultusunda Denklem 2.2’de veri kümesinin kapsamı [142]:

$$\frac{N}{G} \tag{2.2}$$

Teorik olarak, bir veri kümesinin kapsamı c ise, genomdaki her DNA bazı, kapsama uniform olduğu varsayılarak, veri kümesinde c kez temsil edilmektedir. Gerçek veri kümeleri, uygulama üzerinde kullanılan genomun düzenli bir kapsama alanı olmamasına rağmen, bazı bölgeler diğerlerine göre daha fazla kapsanmış olması ve bazı bölgelerde hiç kapsama alanı olmaması şekilde görülebilmektedir.

Okuma doğruluğu, çıktı okumalarına uygulanabilecek tüm süreçlerde hayati bir faktör olarak bilinmektedir. Örnek olarak, YND okumalarının montajı, okuma hataları düzeltilene veya ortadan kaldırılana kadar başarılı bir şekilde gerçekleştirilememektedir. Bu nedenle, okuma hatalarını tespit etmek ve düzeltmek (veya ortadan kaldırmak), montaj işleminden önce yapılması gereken önemli bir adım olarak belirtilmiştir. Bu adım, bağımsız bir çözümle veya dolaylı olarak montaj mekanizması içinde gerçekleştirilebilmektedir. Nükleotidin frekans ve kalite değeri, hatalı olup olmadığının değerlendirilmesinde kullanılan iki ana faktör olarak nitelendirilmiştir [142].

2.4.2. Duyarlılık Ve Özgüllük

Duyarlılık ve özgüllük, kullanılan veri için bir test yönteminin kalitesini tanımlamak için kullanılan genel terim olarak nitelendirilmektedir. Teşhis açısından, bir testin hassasiyeti, belirli bir organizmanın gerçekten mevcut olduğu numuneleri doğru şekilde tanımlama yeteneğini temsil etmektedir. Tanısal özgüllük, ilgilenilen testin belirli bir organizma içermeyen numuneleri doğru şekilde tanımlama yeteneğidir. Yeni bir test veya yöntemi değerlendirirken, bir verinin gerçek doğruluk durumuna ilişkin en iyi tahmini sağlayan referans test veya altın standarda göre performans ölçümleri yapılmaktadır.

Bu doğrultuda kullanılan veri setleri için Denklem 2.3’te duyarlılık, Denklem 2.4’te özgüllük hesaplaması gösterilmektedir.

$$Duyarlılık = \frac{TP}{TP+FN} \quad (2.3)$$

$$Özgüllük = \frac{TN}{TN+FP} \quad (2.4)$$

Denklemlerde kullanılan parametreler şunlardır:

- **TP:** Hatalı kabul edilen hatalı veri sayısı,
- **FP:** Doğru olduğu halde hatalı kabul edilen veri sayısı,
- **FN:** Hatalı olduğu halde doğru kabul edilen veri sayısı,
- **TN:** Doğru tespit edilen doğru veri sayısı.

Duyarlılık ve özgüllük birlikte düşünülmelidir, çünkü test özelliklerindeki bir değişiklik genellikle birinde bir artış ve diğerinde azalma ile yansımaktadır ve doğru dengeyi bulmak genellikle kullanılacak verinin amaçlarına bağlıdır.

3. YAPILAN ÇALIŞMALAR

3.1. Yöntemler

Bu çalışmada YND cihazlarının DNA Dizilemesi yaparken oluşturduğu hataları gidermek amacıyla yeni bir yöntem önerilmiştir. Bu yöntem de BRAF geni eli alınmıştır. Sağlıklı bir BRAF genini kullanarak analizi yapılacak olan DNA parçaları üzerinde hata tespit ve düzeltilmesi amaçlanmaktadır.

3.1.1. Veri Seti

Kullanılan DNA dizisi 1988 yılında Bethesdadaki tarafından kurulan Ulusal Biyoteknoloji Bilgi Merkezi(NCBI) tarafından paylaşılan Homo sapiens(human) BRAF gen modelinin Tablo 3.1'de gösterildiği gibi FASTA formatından elde edilmiştir.

Tablo 3.1. BRAF geni fasta formatı

>NC_000007.14:c140924929-140719331 Homo sapiens chromosome 7, GRCh38.p13 Primary Assembly
CTTCCCCCAATCCCCTCAGGCTCGGCTGCGCCCGGGGCCGCGGGCCGGACCTGAGGTGGCCCAGG CTCCGCCCGCGGCGCCGCCCGGGCCGCTCCTCCCCGCGCCCCCCGCGCCCCCCGCTCCTCCGCCTC TCCGCCTCCGCCTCCCCCAGCTCTCCGCCTCCCTTCCCCCTCCCCGCCCGACAGCGGCCGCTCGGG

İlk satırdaki ">" ifadesi ile başlanarak gen veri dizilimi tanımlayan ifadeler ile devam edilmektedir. İkinci satırında DNA dizilimi yer almaktadır.

3.1.2. K-mer Yöntemi

K-mer tabanlı yöntemler, dizileme okumalarının daha küçük sabit uzunlukta (k) dizilere bölünmesini ve bu kısa diziler üzerinde analiz yapılmasını içermektedir. K-mer'ler "k – 1" bazlarıyla çakışırsa, hatalı veri oluşumunu belirtebilmektedir. Haritalama veya De-Novo montaj yaklaşımları gibi, özellikle hesaplama verimliliğinin çok önemli olduğu bazı uygulamalar için yararlı olabilecek bir yöntemdir. Örneğin; Elimizde L=9 boyutlu bir DNA dizisi ve k=3 boyutlu k-mer leri olan bir veriyi Tablo 3.2'de inceleyelim.

Tablo 3.2. K-mer metodu örneği

DNA Dizisi (L) :	ATCGATCAC
K-mer #**1**	ATC
K-mer #**2**	TCG
K-mer #**3**	CGA
K-mer #**4**	GAT
K-mer #**5**	ATC
K-mer #**6**	TCA
K-mer #**7**	CAC

L=9 olan DNA dizi boyutlu olan verinin k=3 boyutlu k-mer leri için toplam 7 tane k-mer sekansı bulunmaktadır. Buradan yola çıkarak(3.1);

$$S = L - k + 1 \tag{3.1}$$

Denklem 3.1'de kullanılan parametreler şunlardır:

- **S:** Sekans K -mer Sayısı,
- **L:** DNA Dizi Boyutu,

- **k:** K-mer Boyutu.

Bu yöntemin avantajı; bir S sekansının, A organizmasından mı, B organizmasından mı geldiğini etmek için, S'nin A'da veya B'de daha fazla k-mer içerip içermediğini(uygunluğunu) daha kolay ve daha doğru kontrol etmektir.

3.1.3. Sekans Oluşturma

Yeni nesil dizileme cihazları (YND); az zamanda daha yüksek doğruluk oranı olan sonuçlar elde edebilmesi için okuma işleminde, DNA verisi çok büyük olduğu için tüm DNA gen veri dizilimini oluşturmak yerine DNA'yı sekanslar (parçalar) halinde oluşturmaktadır. YND cihazları simüle edilerek Tablo 3.3'e göre dinamik olarak belirlenen boyutlarda sekanslar oluşturulmaktadır.

Tablo 3.3. DNA sekansları oluşturulması

Sekans Sayısı	Sekanslar
Sekans 1	CTTCCCCCAATCCCATCAGGCTCGGCTGC
Sekans 2	CTCCGCCCGCGGCGCCGCCGGGCCGCTC
Sekans 3	ATCAATAGGGGGCGAAACTGGGGTTGGT
Sekans 4	GTTTTACCCAGCAAATTATTTATGATTTAG
Sekans 5	TTCTGGTCTCTGGTCCTCTGTTTCCTAATG
…..	**…..**
…..	**…..**
Sekans n-1	AAAAAGATAGCTGAAATTCAGATTGAAGA
Sekans n	AAAAGGCAAAAAGAGTTCTATGTACTTGA

Kullanılan BRAF geninin boyutu:205599 olarak verilmiştir. Bu veri NCBI' da sağlıklı bir insan geni olarak belirtilmiştir. Bu veri üzerinden sekanslar oluşturulurken dinamik olarak belirlenen hata oranlarıyla rastgele nükleotid silinmesi, eklenmesi veya değişmesi şeklinde hatalar yapılmaktadır. Hata oranları, Yeni Nesil Dizileme (YND) cihazların çalışma prensibi göz önünde bulundurularak ideal değer olan %1 değer doğrultusunda oluşturulmuştur.

Oluşturulan sekansların sayısı, okuma derinliği (coverage) olarak ifade edilen DNA gen verisi analizi edilecek verinin okuma derinliğine bağlı olarak hesaplanmaktadır (3.2).

$$\mathrm{n} = \frac{g \times c}{s} \tag{3.2}$$

Denklem 3.2'de kullanılan parametreler şunlardır:

- **n:** DNA Sekans Sayısı,
- **g:** DNA Gen Boyutu,
- **c:** Okuma Derinliği,
- **s:** DNA Sekans Boyutu.

3.1.4. Sekans-Kmer Oluşumu

Sekansların k-mer sınıflarının oluşturulması; makine okuma işleminde rastgele seçilen sekans ile o sekansın belirlenen k-mer boyutunda bütün K-mer'leri oluşturulmaktadır. Belirlenen k-mer'ler o sekansın K-mer alt kümesi olmaktadır. BRAF gen verisinin tüm sekanslarını ve sekans-Kmer'leri oluşturulmaktadır.

Tablo 3.4'te gösterildiği gibi sekans boyutu=100 ve k-mer boyutları sırasıyla 16,20,24 olarak sınıflandırılmaktadır. K-mer boyutunu artırarak sekanslardan oluşan K-mer'lerin benzerlik oranlarını azaltılması planlanmaktadır.

Tablo 3.4. Sekans-Kmer boyut analizi

Sekans Boyutu:100	
K-mer Boyutu	**Sekans-Kmer Sayısı**
16	85
20	81
24	77

Denklem 3.1'e göre hesaplanan Sekans-Kmer sayısı, bir sekans için hesaplanan analiz değeridir. Bu işlem bütün BRAF gen verisinde Denklem 3.2'e göre okuma işlemi yaparken oluşan sekanslar için yapılmaktadır. Sekans oluştururken kullanılan derinlik(coverage) değeri okuma yapılan verinin daha doğru okunmasını sağlayabilmek için yapılmaktadır. Bu durum Tablo 3.5'te gösterilmektedir.

Tablo 3.5. BRAF gen sekans sayısı analizi

BRAF Gen Boyutu: 205599	
Okuma Derinliği(c)	**Toplam Sekans Sayısı(n)**
30	61679
40	82239
50	102799

Derinlik değeri arttıkça sekans sayısı da artmaktadır. Bu durum çalışma süresini bir miktar artırsa bile okuma yaparken doğru biçimde okuma olasılığını daha çok artırmaktadır. BRAF gen verisi için sekans boyutu 100 olan ve k-mer boyutu sırasıyla 16 20 24 olan değerlere göre toplam k-mer sayısı Tablo 3.6'da gösterilmektedir.

Tablo 3.6. BRAF gen toplam k-mer sayısı

Toplam Sekans Sayısı: 102799		
K-mer Boyutu	**Sekans-Kmer Sayısı**	**Toplam K-mer Sayısı**
16	85	8737915
20	81	8326719
24	77	7915523

3.1.5. Hatalı Verinin Oluşturulması

Makine okuma hatalarının analizini yapabilmek için sağlıklı BRAF geni üzerinde hatalar oluşturulmuştur. Hatalı veriler, YND cihazlarının %1 oranında ki hata değerlerine göre ayarlanmıştır ve hata türleri de literatürdeki hatalı okuma veri analizi simülasyonları sonucuna uygun olacak şekilde Tablo 3.7'deki gibi ayarlanmıştır.

Tablo 3.7. Hata ekleme oranları

Hatasız Sekans	**1 Hatalı Sekans**	**2 Hatalı Sekans**
%5	%90	%5

Hata Türü		
Ekleme Hatası	**Silme Hatası**	**YerDeğişme Hatası**
%20	%20	%60

Okuma Derinliği, sekanslar arasındaki hataların tespitinde doğruluğu artırmak için kullanılmaktadır. Derinlik arttıkça daha doğru sonuçlar elde edilmektedir. Derinlik artması verinin sekans sayısını artırmaktadır bu yüzden verinin boyutu artırmaktadır, fakat

kullanılan filtre ortamında ki sorgulamasında sekanstaki olası hatanın daha doğru şekilde tespit edilmesine olanak sağlamaktadır. Okuma derinliği analizi Tablo 3.8'de gösterilmektedir.

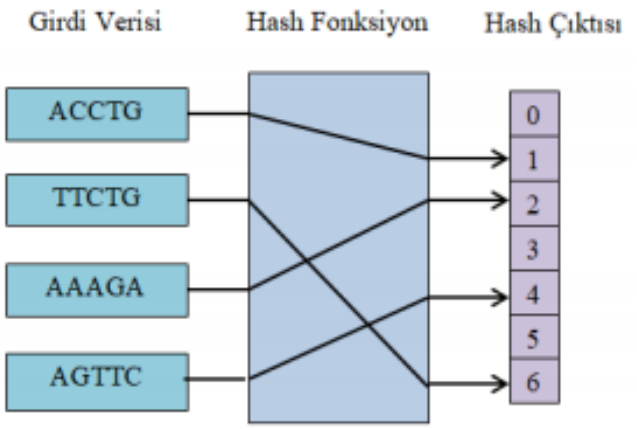
Tablo 3.8. Okuma derinliği analizi

Okuma Derinliği(c)	**Hata Değerleri (error < derinlik / 2)**	**Hata Olasılığı**
30	415	%0,46
40	166	%0,15
50	98	%0,07

Olası hatalar derinlik boyutunun yarısına kadar olan indisler baz alınarak hesaplanmıştır. Çünkü yarısından fazla olan indislerde mutasyon olma ihtimali daha yüksektir.

3.1.6. Hash Fonksiyonları

Bir hash işlevi, rastgele boyutlardaki girdileri sabit boyuttaki çıktılarla eşleyen deterministik bir işlevdir. Bu, sonsuz sayıda olası girdi olduğu, ancak yalnızca sınırlı sayıda olası çıktı olduğu anlamına gelmektedir. Şekil 3.1'de hash fonksiyonu çalışmanı prensibi gösterilmiştir.



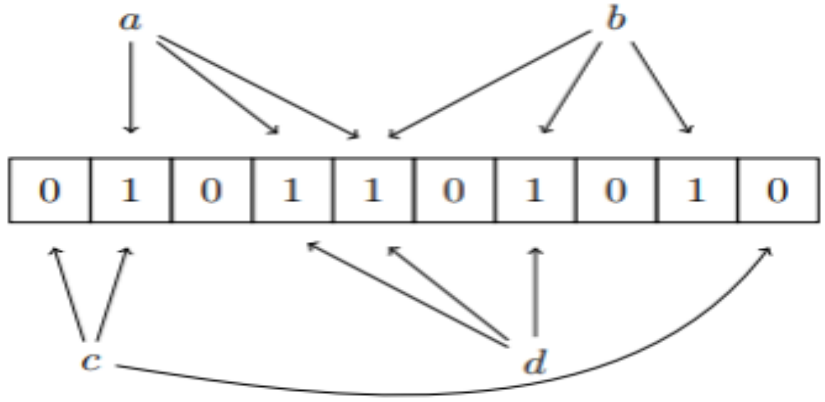
Şekil 3.1. Hash fonksiyonları

Hash fonksiyonların sayısı girdi verilerinin set edilecek indislerini belirlemek için kullanılmaktadır. Sistem için önemli olan durum kullanılacak hash fonksiyonlarının sayısının belirlenmesidir. Çünkü hash sayısının optimum değerinin bulunması girdi verilerindeki çakışmayı önlememize sağlamaktadır. Çakışma oranı ne kadar az olursa girdi verisinin kontrol edilmesi de o kadar doğru sonuç vermektedir.

3.1.7. Bloom Filter

Bloom filtresi, Burton Howard Bloom tarafından 1970 yılında tasarlanmış olup, hızlı çalışabilmesi ve bellek açısından daha verimli olabilmesi ve bir elemanın bir kümenin üyesi olup olmadığını test edebilmek amacıyla kullanılan bir olasılıksal veri yapısı olarak nitelendirilmektedir. Bloom filtresi, yanlış pozitiflerle dinamik veri kümesi için üyelik sorgularını destekleyen olasılıklı veri yapısı olarak adlandırılmaktadır [143-145]. Düşük oranda yanlış pozitiflere izin verirken, bir veri setinde birden fazla bulunan tüm K-mer'leri son derece kompakt bir şekilde tanımlamamıza olanak tanımaktadır. Bloom filtreleri bilgi işlem uygulamalarında yaygın olarak kullanılmıştır, ancak bugüne kadar biyoinformatik nadiren kullanılmıştır. Temel tanım, Şekil 3.2'de gösterilmektedir. Bloom filtresi, her pozisyonda "0" olarak başlatılan bir bit dizisi B'dir. Ayrıca, kullanılan her bir hash

fonksiyon görevi belirli k-mer'leri işlem sonucuna göre B'deki indislere eşlediği bir dizi d hash işlevi, h1, ..., hd şeklinde tanımlanmaktadır.



Şekil 3.2. Bloom Filter örneği

Şekil 3.2'de görüldüğü gibi 3 adet hash fonksiyonuna kullanılmış Bloom filtresi örneği bulunmaktadır. K-mer a ve b filtreye eklenmiştir, ancak c ve d eklenememiştir. Kullanılan hash fonksiyonların işlevi oklarla temsil edilmektedir ve devamında a ve b için hash fonksiyonuna karşılık gelen bitler filtre üzerinde 1 değeri olarak ayarlanmıştır. Bloom filtresi, tüm bitleri 1'e ayarlanmadığı için k-mer c'nin doğru şekilde eklenmediğini göstermektedir. Fakat k-mer d yanlış pozitifi göstermektedir ve eklenmemiştir, ancak filtre üzerindeki bitleri a ve b'nin eklenmesiyle 1'e ayarlandığından dolayı, Bloom filtresi sorgulama esanasından yanlış cevap verip d'nin zaten görüldüğünü bildirmektedir [12]. Bloom filtresine bir k-mer x eklemek için, B'deki tüm d'ye karşılık gelen konumları 1 olacak şekilde ayarlanmıştır; yani, i = 1,...,d için B [hi (x)] = 1 ayarlanmaktadır. Ardından, sorgulama ile bir k-mer y'nin filtreye eklenip eklenmediğini belirleyebilmek için, hash işlevi sonucunda karşılık gelen indislerin her birinin 1 olup olmadığını kontrol etmek gerekmektedir: yani, i = 1,...,d için B [hi (y)] = 1 olup olmadığı kontrol edilmesi gerekmektedir. Eğer sonuç durumu böyleyse, sorgulama sonucunda y'nin muhtemelen daha önce görüldüğü şeklinde ön görülmektedir [146]. Çalışması gereği, birden fazla bulunan her k-mer için doğru bir şekilde tanımlama yapılabilmektedir; fakat çok verimli bellek kullanımınında, y'nin daha önce görülmüş olduğu, ancak aslında görülmediği sonucuna vardığımız yanlış pozitiflerin düşük oranını kabul edilmesidir.

İlk olarak, kombinatorik bir argümanla, yanlış pozitifin p olasılığının aşağıdaki Denklem 3.3'te tahmin edildiğini gösterilmektedir:

$$p = \left(1 - e^{\frac{-kn}{m}}\right)^k \tag{3.3}$$

Denklem 3.3'te kullanılan parametreler şunlardır:

- **p:** Yanlış Pozitif(False-Positive) Olasılığı,
- **k:** Hash Fonksiyon Sayısı,
- **n:** Girdi Verisinin Boyutu,
- **m:** Bloom Filter Boyutu.

Literatürde yanlış pozitifin genel olarak optimum değeri: 0.0000001(1.0E-7) olarak kullanılmaktadır.

İkinci olarak, aşağıdaki durumlarda k'nin optimal değeri Denklem 3.4 ve Denklem 3.5'te hesaplanmaktadır:

$$k = \frac{m}{n} \ln 2 \tag{3.4}$$

Bu formülde, m / n'nin Bloom filtresindeki öğe başına bit sayısı olduğuna dikkat edilmektedir. Dolayısıyla, Denklem 3.5'te gösterilen optimal hash sayısı, eleman(b) başına bit sayısıyla doğrusal olarak artmaktadır.

$$k = b \log 2 \tag{3.5}$$

Denklem 3.3'e göre, k için optimal bir seçim olduğunu varsayarsak, Denklem 3.6 elde edilmektedir:

$$p = 2^{\frac{-m \log 2}{n}} \tag{3.6}$$

Tüm veri analizinden çıkan sonuçta Bloom Filter boyutu(m) için Denklem 3.7 oluşmaktadır;

$$m = \frac{n \log \frac{1}{p}}{(\log 2)^2} \tag{3.7}$$

Temel Bloom Filter formüllerine ek olarak bizim çalışmamızda projeye uyarlanmış Bloom Filter hesaplama formülü oluşturulmuştur. Bu Denklem 3.8 ile biz her girdi verisinin sekansları ile bu sekansların da k-mer alt sekansları üzerinden yeni Bloom Filter boyutu belirleme işlemi yapılmaktadır.

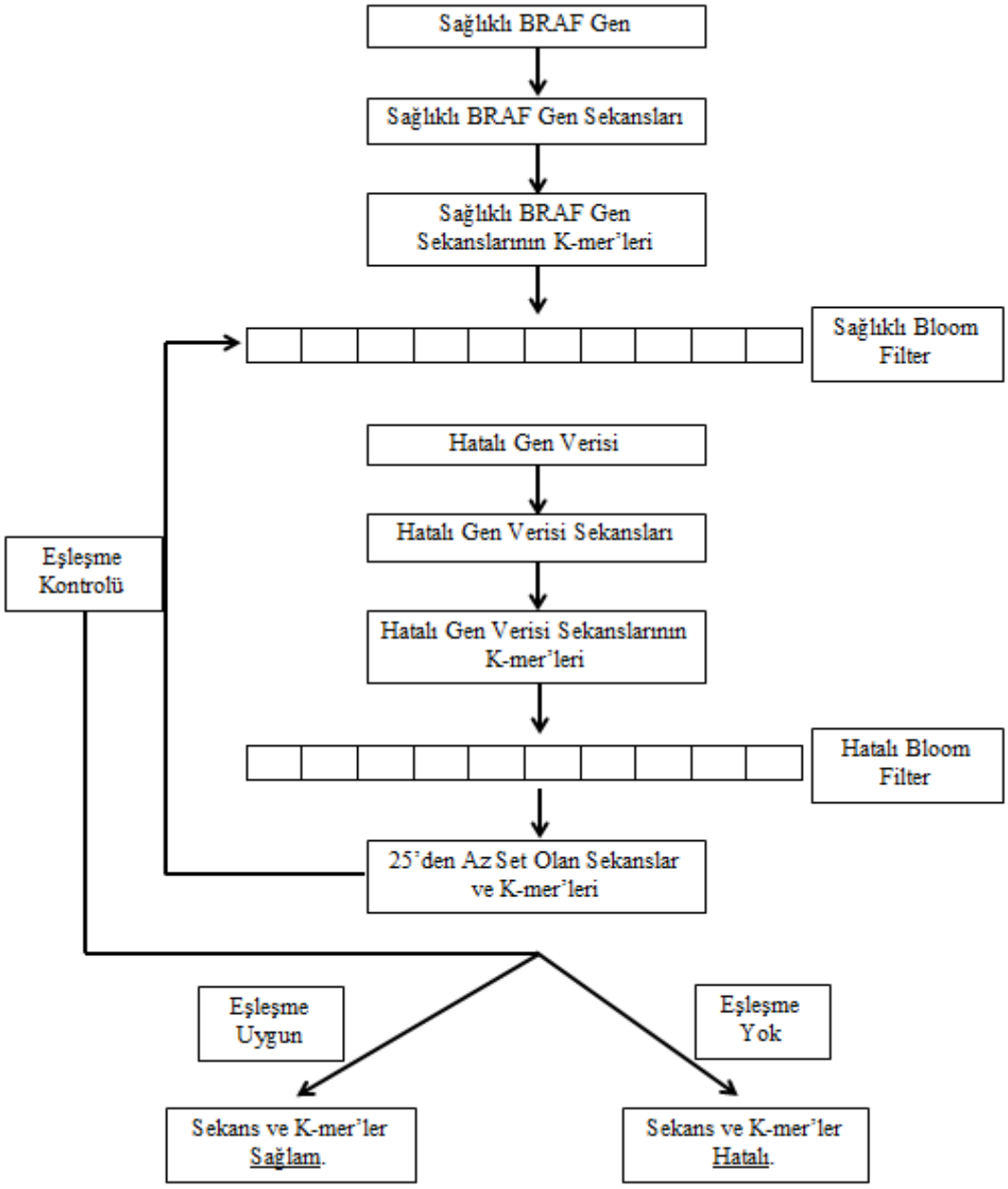
$$m = c \times n \times k \tag{3.8}$$

Denklem 3.8'de kullanılan parametreler şunlardır:

- **m:** Bloom Filter Boyutu,
- **c:** Eleman Başına Kullanılan Bit Sayısı,
- **n:** Filtrenin içereceği Beklenen Girdi Veri Sayısı,
- **k:** Hash Fonksiyon Sayısı.

3.1.8. Okuma Hatalarının Tespiti

Sağlıklı BRAF geni ile oluşturulan hatalı gen verisi arasında kıyaslama analizi yapılarak hata tespiti yapılmıştır. Bu kıyaslamayı sağlıklı gen için ve hatalı gen için oluşturulan Bloom Filter yapısıyla kontrol edilmektedir. Sağlıklı Bloom Filter içinde sağlıklı BRAF gen sekansları kaydedilmiştir. Hatalı Bloom Filter içinde hatalı gen sekansları kaydedilmiştir. Örnek olarak uyarladığımız uygun görülen derinlik değeri olan c=50'nin hatalı veri olma olasılığını c/2=25 olarak ön gördüğümüz analize göre 25'ten az set edilen sekansları sağlıklı Bloom Filter içinde sorgulama işlemi yapılarak eşleşme olup olmadığı kontrol edilmektedir. Şekil 3.3'teki gibi eşleşme olmayan sekanslarda hata olduğu anlaşılmaktadır ve tespit edilen sekans ve içindeki k-mer'ler kayıt altına alınmaktadır.



Şekil 3.3. Hatalı sekans ve k-mer tespiti

Hata tespit türleri;

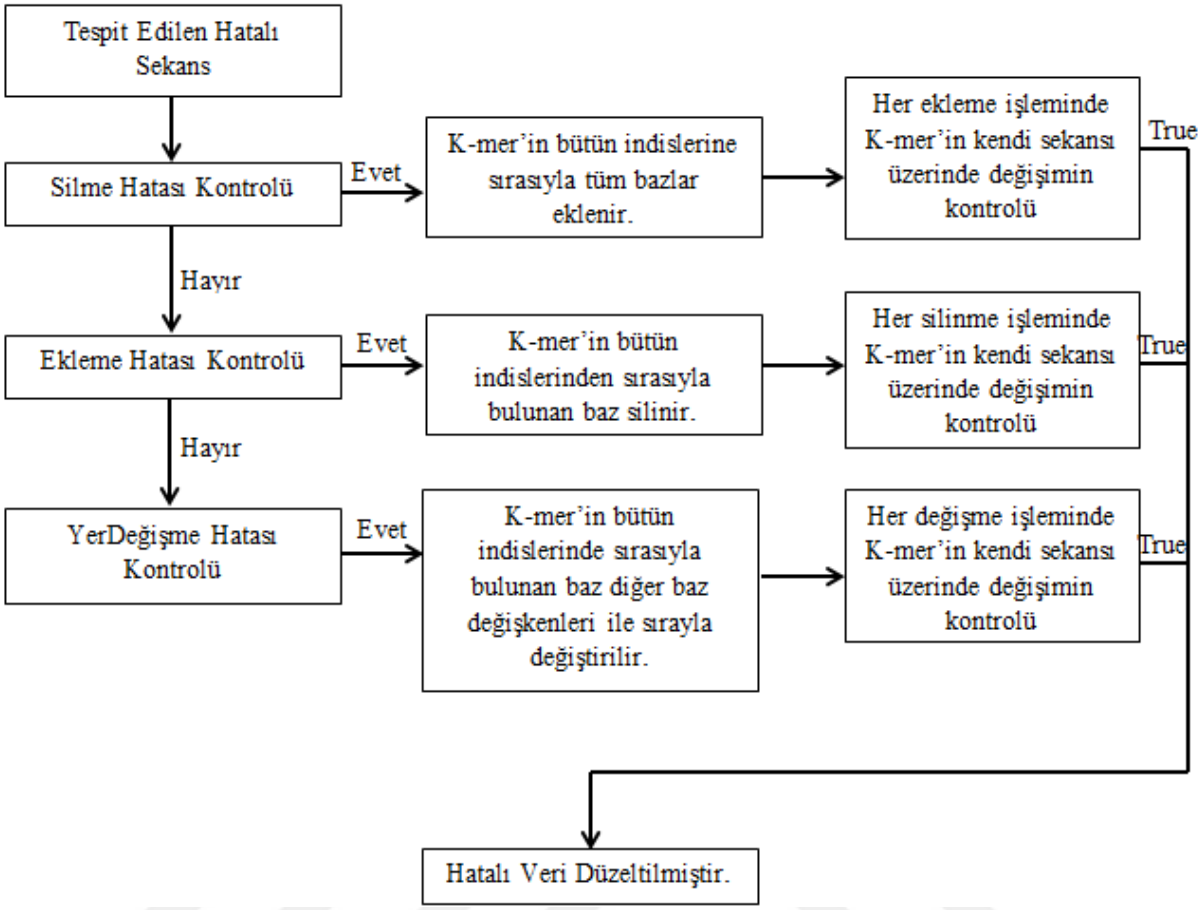
- Ekleme,
- Silme,
- Yer Değişme,

şeklinde Şekil 3.4'te örneklendirilmiştir.

Şekil 3.4. Hata türleri örneği

3.1.9. Okuma Hatalarının Düzeltilmesi

Girdi verisi üzerinde ki hatalı okumaların tespit edilme işleminden sonra belirlenen hatalı sekansın kendi içinde belirlenmiş olan k-mer boyutunda k-mer alt sekansları haline getirilmektedir. Her k-mer içindeki DNA bazları ilk indisinden başlayarak sırasıyla baz değişimi, baz eklemesi ve baz silmesi yapılarak Bloom Filter'a kaydedilen sağlıklı BRAF geni üzerinden sorgulama yapılıp kontrolü sağlanmıştır. Ayrıca k-mer üzerinde yapılan her düzeltme işlemleri ait olduğu sekans ve sağlıklı sekans ile de kontrolü yapılmaktadır. Çünkü hatalı k-mer düzeltilirken sağlıklı BRAF geni içindeki başka bir k-mer ile uyuşması olasılığı olabilmektedir. Bu yüzden her düzeltme işlemi aynı zamanda sekans olarak da kontrol edilip sekans-kmer ilişkisi olarak Bloom Filter sorgusu yapılmaktadır. Şekil 3.5'teki gibi hatalı veri kontrol sonrasında doğru veriyi bulana kadar 'false' değerini üretecektir. Yöntem sonunda 'true' değeri elde edilince hatanın hangi indiste olduğu belirlenmiş olup, düzeltilmesi yapılmaktadır.



Şekil 3.5. Hatalı verinin düzeltilmesi

4. BULGULAR

Temel Bloom Filter formülleri ile bu çalışmada kullanılan Bloom Filter formül ve özniteliklerine göre BRAF veri kümesi için ortak varyans noktası Tablo 4.1'de gösterilmiştir.

Tablo 4.1. BRAF gen veri setinde Bloom Filter boyut analizi

BRAF VERİ SETİ BOYUTU: 205599				
Hash Sayısı	**Temel Bloom Filter**		**Çalışılan Bloom Filter**	
	K-mer Boyutu	**Bloom Filter Boyutu**	**K-mer Boyutu(Ortak)**	**Bloom Filter Boyutu**
4	16	45830955	16	13157376
4	20	45830063	20	16446400
4	24	45829171	24	19735296
5	16	25302744	16	16446720
5	20	25302251	20	20558016
5	24	25301759	24	24669120
6	16	17481359	16	19736064
6	20	17481019	20	24669632
6	24	17480678	24	29602944

Tablo 4.1'e göre iki formül için olası en optimum değerler için Bloom Filter boyutunun birbirine benzerlik analizi yapılmıştır. Bu doğrultuda, Hash Sayısı=6 ve K-mer uzunluğu=16 olan değer en yakın değer olarak görülmüştür. Bu farklılıklar kullanılan veri seti için seçilen özniteliklerin yanı sıra Bloom Filter yapısının False-Positive oranının farklılığından kaynaklanmaktadır. Temel Bloom Filter formüllerine göre kullanılan veri

seti ile bu çalışmada kullanılan Bloom Filter için birbirine yakın oranlarda olması ve sekansların kullanılacak veri setine göre belirlenen k-mer boyutlu alt dizileri sadece kendilerine özgü olması gerekmekte ve aynı okumadaki farklı hiç bir sekansta bulunması beklenmemektedir. Ayrıca çakışmayı önlemek için hash fonksiyonları sayısını ve ram deki bellekteki çalışma süreleri baz alınarak uygun görülmüştür.

Bu çalışma için 2 tane Bloom Filter kullanılması ön görülmüştür. Birinci filtre NCBI veri tabanından indirdiğimiz sağlam BRAF geninin sekanslama ve hash fonksiyonu yardımıyla oluşturduğumuz yeni Denklem 3.8'e göre oluşturulmuştur. İkinci filtre ise makinenin okuma hatası yaparak okuduğu gen verisini sekanslama, derinlik boyutu ve hash fonksiyonları ile analizinden sonra yine aynı denklem ile oluşturulmuştur. Karşılaştırmalı değerler Tablo 4.2'de gösterilmektedir.

Tablo 4.2. Bloom Filter boyutu

Hash Sayısı	**K-mer Boyutu**	**Sağlam Bloom Filter Boyutu**	**Hatalı Bloom Filter Boyutu**	**Toplam Süre(sn)**
4	16	13157376	559226560	34,523
4	20	16446400	666137536	34,482
4	24	19735296	759890240	33,24
5	16	16446720	699033216	42,127
5	20	20558016	832671936	41,017
5	24	24669120	949862784	40,112
6	16	19736064	838839872	48,538
6	20	24669632	999206336	48,397
6	24	29602944	1139835328	49,508

Yapılan çalışma BRAF veri geni üzerinde farklı okuma uzunluğuna sahip, farklı derinlik değerlerinde test edilmiştir. Bu sonuçlar Ek Tablo 2'de gösterilmektedir. Ek Tablo 2'de gösterilen öznitelik değerlerine göre programın filtre boyutları ve okuma hatalarının

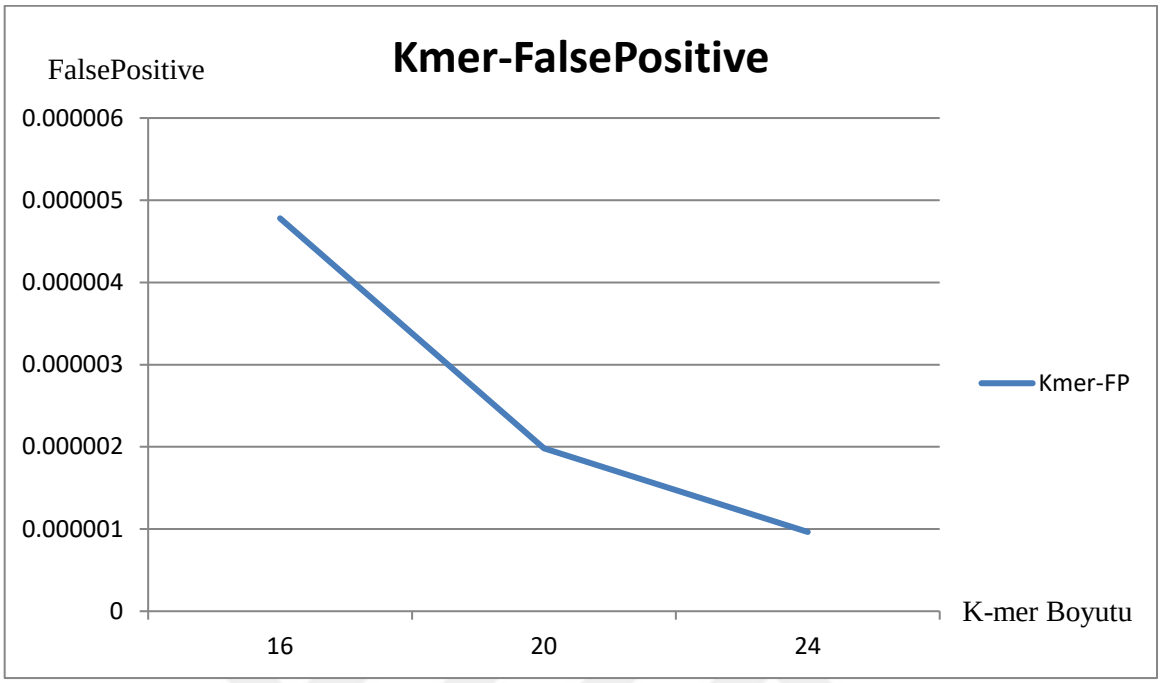
düzeltilmesi işlemi için çalışma süresi gösterilmektedir. Farklı derinliklerde, farklı okuma boyutu ve k-mer boyutuna göre hesaplanmaktadır.

Tablo 4.3. FalsePositive oranları

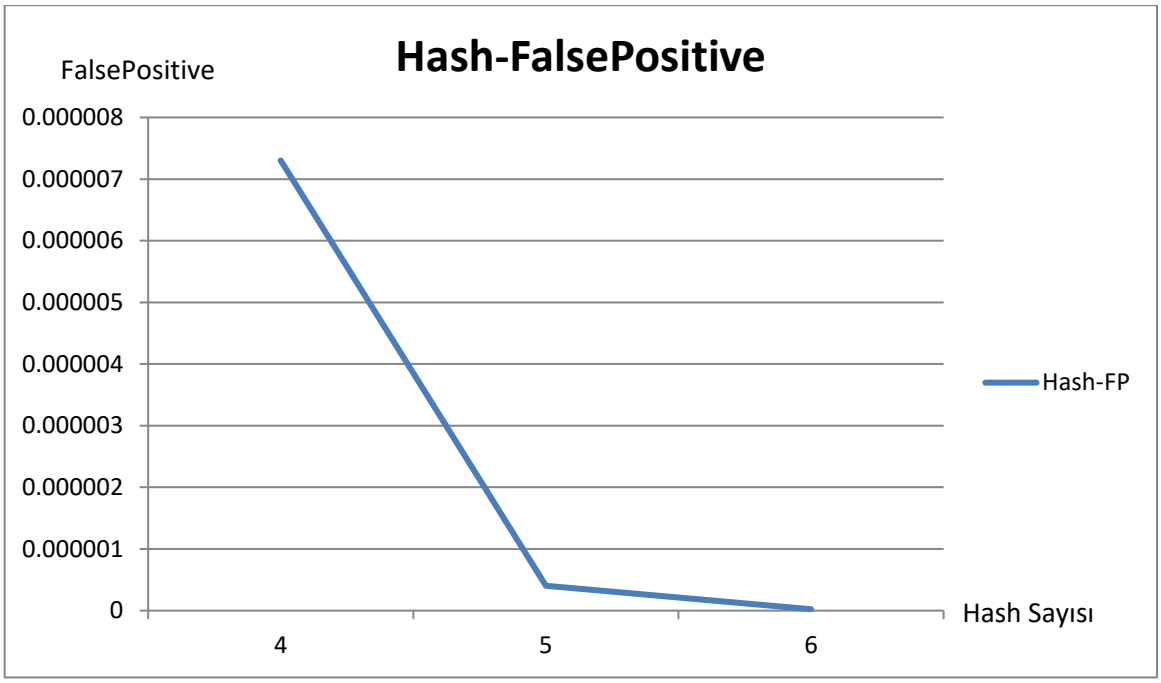
K-mer Boyutu	**Hash Fonksiyon Sayısı**	**FalsePositive Oranı**
16	4	$1{,}3475 \times 10^{-5}$
16	5	$8{,}16 \times 10^{-7}$
16	6	$4{,}9 \times 10^{-8}$
20	4	$5{,}658 \times 10^{-6}$
20	5	$2{,}76 \times 10^{-7}$
20	6	$1{,}3 \times 10^{-8}$
24	4	$2{,}774 \times 10^{-6}$
24	5	$1{,}13 \times 10^{-7}$
24	6	5×10^{-9}

Tablo 4.3'te gösterilen False-Positive oranları aynı veri seti üzerinde kullanılan k-mer boyutu ve hash fonksiyonlarının sayısına bağlı olarak programın yanılma oranını göstermektedir. K-mer boyutu ve hash fonksiyonu sayısı artması programın daha düşük hata yapma olasılığı olduğunu göstermektedir. Ancak Ek Tablo 2 verilerine göre k-mer boyutu ve hash fonksiyonun artması çalışma süresini artırmaktadır. Bu sonuçların doğrultusunda programın çalışma şartlarına göre test edilip okuma hata düzeltme becerisinin(oranın) en iyi olduğu sekans ve k-mer boyutu tespit edilmelidir.

Bloom Filter için kullanılan temel(standart) False-Positive oranı="0,0000001(1×10^{-7})" olarak kullanılmaktadır. Tablo 4.3'te gösterilen False-Positive oranları ile Bloom Filter'in genel False-Positive oranı kıyaslaması yapıldığı zaman doğruluk açısından en uygun hash fonksiyonu sayısının "6" olduğu ön görülmüştür. Çünkü hash sayısı 6 iken k-mer boyutunun da artması da normal Bloom Filter'a göre daha düşük hata yapma olasılığı göstermektedir.



Şekil 4.1. Kmer ve FalsePositive ilişkisi



Şekil 4.2. Hash sayısı ve FalsePositive ilişkisi

FalsePositive oranlarının k-mer sayısı ve hash sayısına göre ilişkisini Şekil 4.1 ve Şekil 4.2'de gösterilmiştir. K-mer ve hash sayısının artırmak FalsePositive oranını azalttığı görülmektedir. Bunun sonucunda da hata oranı azalmaktadır.

Tablo 4.4. Duyarlılık ve özgüllük sonuçları

Sekans Boyutu	**K-mer Boyutu**	**Duyarlılık**	**Özgüllük**
50	16	98,94	99,99
50	20	99,52	99,99
50	24	99,58	99,99
75	16	99,44	99,99
75	20	99,73	99,99
75	24	99,87	99,99
100	16	99,63	99,99
100	20	99,75	99,99
100	24	99,89	99,99

Tablo 4.4'te çalışmadaki belirli öznitelik değerleri için yapılan duyarlılık ve özgüllük sonuçları gösterilmektedir. Farklı sekans boyutu ve k-mer boyutuna göre duyarlılık ve özgüllük sonucu üzerindeki etkisin analizini yapılmaktadır. Sekans boyutu ve aynı doğrultuda k-mer boyutunun artması hata düzeltilme performansını artırdığı ön görülmektedir. Ancak sekans ve k-mer boyutunun artması çalışma süresini de artırmaktadır. Bu sonuçlar doğrultusunda kullanılacak gen verisi için istenilen özniteliklere göre test edilip okuma hatası düzeltme becerisinin en uygun olduğu sekans ve k-mer boyutu tespit edilmelidir.

5. TARTIŞMA

Yeni nesil dizileme cihazlarında ki okuma hatalarının düzeltilmesi için geliştirilen yöntemde %1 hata oranı baz alınarak yapılmaktadır. Kullanılan yöntemlerle bu veri seti için ön görülen öznitelik değerler analizinde hataların tespiti ve düzeltilmesi ve genel Bloom Filter False-Positive oranını değiştirerek doğruluk açısından daha elverişli olduğu gösterilmektedir. Kullanılan BRAF gen veri seti için yapılmakta olan hata tespit ve düzeltme işlemleri yapılan çalışma için süreleri gösterilmektedir. Hata tespit oranını göz önüne aldığımız zaman bu veri seti için en uygun hash sayısının "6" olduğu gösterilmiştir. Çalışma sonucuna göre seçilen öznitelik değerleri olan sekans boyutu, kmer boyutu, derinlik ve hash sayısı değerlerinin arttıkça Bloom Filter boyutunun arttığı ve bununla beraber çalışma zamanının da arttığı gösterilmektedir, fakat bu artış hata oranlarının azaldığını göstermektedir.

Tablo 5.1. Bloom Filter FalsePositive oranı analizi

Öznitelikler	**Standart Bloom Filter – FalsePositive Oranı**	**Kullanılan Yöntem Bloom Filter – FalsePositive Oranı**	**Ortalama Çalışma Zamanı(sn.)**
K16	1×10^{-7}	$4{,}9 \times 10^{-8}$	450,84
K20	1×10^{-7}	$1{,}3 \times 10^{-8}$	570,36
K24	1×10^{-7}	5×10^{-9}	684,12

Tablo 5.1'de gösterilen öznitelikler olan sekans sayısı, derinlik, kmer boyutu ve hash sayısı=6 için ortalama Bloom Filter FalsePositive oranı ile standart Bloom Filter FalsePositive oranı gösterilmiştir. Ortalama çalışma zamanı da aynı öznitelik değerleri için hesaplanmıştır. Önerilen yöntemde daha doğru sonuçlar elde edilmek istenilirse eğer, Tablo 5.1'deki veri analizine göre öznitelik değerleri; hash sayısı=6, derinlik=50, k-mer

boyutları sırasıyla 16,20,24 olacak şekilde sonuçlandırılmıştır. Tablo 4.4'de bu öznitelik değerlerine göre duyarlılık ve özgüllük değerleri gösterilmiştir. Ek Tablo 2, Tablo 4.3 ve Tablo 4.4'e göre kullanılacak gen verisi için daha doğru sonuçlar veya daha hızlı sonuçlar şeklinde analize göre değerler seçilmelidir. Elde edilen sonuçlar Intel(R) Core(TM) i7-2630QM CPU @ 2.00GHz, 8 GB RAM, AMD Mobility Radeon HD 5000 Series paylaşımlı ekran kartı olan Windows 10 Pro 64 bit işlemci sistemli notebook ile gösterilmiştir. Kullanılacak veri seti için özniteliklerin uygunluğuna göre test edilip, hata düzeltme becerisinin en uygun olduğu değerler belirlenmelidir.

6. SONUÇLAR

- Geleneksel dizileme yöntemleriyle boyutu küçük olan sekans parçalarında doğruluk oranı yüksek olan DNA dizilemeleri elde edilmektedir. Ancak bu yöntemler maliyet ve hız yönünden yeterli performansı gösterememektedir. Herhangi bir canlının analizinde kullanılacak verilerin daha kolay, hata oranı az ve hızlı elde edilebilir olması çok önemlidir. Bundan dolayı sürekli gelişen teknolojiyle birlikte biyoinformatik alanında da yeni nesil DNA dizileme cihazları üretilmeye başlanılmıştır. Bu cihazlar sayesinde geleneksel dizileme yöntemlerinden çok daha hızlı ve maliyeti düşük olan DNA dizilimleri elde edilebilmektedir.
- Yeni nesil DNA dizileme cihazlarının avantajlarının yanı sıra, geleneksel dizileme yöntemlerine göre daha fazla hatalı okuma sonuçları elde edilebilmektedir. Düşük maliyet ve yüksek hızdan vazgeçilememesi, bu probleme çözüm getirmek için hata düzletme algoritmaları geliştirilmektedir. Bu çalışmada yeni nesil dizileme yöntemiyle elde edilen sekanslarda okuma hatalarının tespiti ve düzeltilmesi için yöntem geliştirilmiştir.
- Yapılan çalışmada standart Bloom Filter metoduna ek olarak verilen formüller ile kullanılacak gen verisi ve belirlenen değerler için okuma hata oranındaki azalmalar gösterilmiştir. Kullanmış olduğumuz öznitelik değerleri olan sekanslama, derinlik okuma, hash sayısı ve k-mer boyutu gibi işlemler ile BRAF gen verisinin boyutu artırılarak ve ayrıca veri üzerinde collision yani çakışma oranını da azaltarak hataların daha kolay tespit edilebilmesi ve düzeltilebilmesi sağlanmıştır. Bloom Filter yapısı sayesinde de tespit edilen hataların düzeltme işleminde aynı sekans ve k-mer üzerindeki değişiklikleri ile kontrol edilmesi ve farklı sekanslarda da kontrolünün sağlanması ile daha doğru düzeltme işlemi yapılmaktadır. Yapılan test sonucunda da standart Bloom Filter verileri ile geliştirilmiş olan yöntemdeki Bloom Filter verisinin hata oranı analizi yapılmaktadır. Çalışma zamanı kullanılan notebook özellikleri ve kullanılan gen verisine gerekli görülen öznitelik değerlerine göre hesaplanmaktadır.
- Yapılan bu tez çalışmasında, yüksek doğruluklu gen veri dizilimleri elde edilebilir. Elde edilen gen verisinin dizilimi ile dizileme algoritmaları daha verimli bir şekilde

kullanılarak sağlıklı-hastalıklı gen tespiti analizleri daha yüksek doğrulukta oranlarla gerçekleştirilebilir.

- Önerilen veri okuma ve düzeltme yönteminin, genetiksel bozukluklara neden olan DNA mutasyonların, genetiksel hastalıkların vs. tespitine ve çözümüne yardımcı olabilmesi ve biyoinformatik alanındaki çalışmalara katkı sağlanması amaçlanmaktadır.

7. ÖNERİLER

Yapılan çalışma için gerekli öznitelik sonuçlarının analizi gözlemlenmiştir. Sonraki çalışmalarda yöntemin eksik yönleri giderilmesi planlanmaktadır. Planlanan çalışmalar:

- Çalışma zamanını daha da iyileştirmek için CPU tabanlı paralel programlamaya uyarlanan yöntem ve GPU tabanlı paralel programlamaya uyarlanan yöntem şeklinde geliştirilip kıyaslanması planlanmaktadır.
- Daha büyük gen verileri üzerinde çalışmaların yapılması ve öznitelik analizlerinin değerlendirilmesi planlanmaktadır.
- Bloom Filter formülü üzerinde iyileştirmeler yapılarak bellek kullanımının azaltılması ve hatalı okuma oranı olan FalsePositive değerinin daha da azaltılması planlanmaktadır.

Hata düzeltme algoritmaları ile oluşturulan gen verilerinin dizilim, gen veri tespiti, mutasyon tespiti ve metagenom çalışmalar gibi biyoinformatik alanlarda sorunlara çözüm üretilebilmesi ve literatüre katkı sağlaması planlanmaktadır. Bu sorunlara yeni algoritmik yöntemler ve dizileme algoritmalarının geliştirilmesi gibi çözüm odaklı çalışmalar yapılabilmesi amaçlanmaktadır.

8. KAYNAKLAR

1. https://www.etymonline.com/ Genetic. 02 Ağustos 2020.

2. Miko, I., Gregor Mendel and the principles of inheritance, Nature Education, 1,1 (2008) 134.

3. https://en.wikipedia.org/wiki/Gregor_Mendel/ Wikipedia the free encyclopedia, Gregor Mendel. 09 Eylül 2020.

4. https://tr.khanacademy.org/science/biology/classical-genetics/mendelian--genetics/a/mendel-and-his-peas/ Khan Academy, Mendel ve bezelyeleri. 10 Eylül 2020.

5. Dahm, R., Friedrich Miescher and the discovery of DNA, Developmental biology, 278,2 (2005) 274-288.

6. Watson, J.D. ve Crick, F.H., Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid, Nature, 171,4356 (1953) 737-738.

7. Nirenberg, M.W., The genetic code, Scientific American, 208,3 (1963) 80-95.

8. Sanger, F. ve Gilbert, W., Genome analysis, (1989).

9. Bolukbasi, E. ve Aras, E.S., Third generation DNA sequencing technologies, DNA, 1,3 (2015).

10. KIZMAZ, M.Z., PAYLAN, İ.C. ve ERKAN, S., DNA Dizilemenin Tarihsel Gelişimi, Gaziosmanpaşa Bilimsel Araştırma Dergisi, 6,2 (2017) 47-53.

11. https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-biotechnology/a/dna-sequencing/ Khan Academy, DNA sequencing. 24 Eylül 2020.

12. https://www.illumina.com/science/technology/next-generation-sequencing.html/ illumına, NEXT-GENERATION SEQUENCING. 22 Eylül 2020.

13. https://en.wikipedia.org/wiki/Human_Genome_Project#Human_Genome_Project/ Wikipedia, Human Genome Project. 26 Eylül 2020.

14. Li, Z., Effect of Naturally Occurring DNA Modifications on DNA Structure and Packaging, University of Cambridge, 2019.

15. Flemming, W., Zur Kenntnis der Zelle und ihrer Theilungserscheinungen, Schr Nat Wiss Ver Schlesw-Holst, 3 (1878) 23-27.

16. Gayon, J., From Mendel to epigenetics: History of genetics, Comptes rendus biologies, 339,7-8 (2016) 225-230.

17. Sturtevant, A.H., " The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association?(1913), by Alfred Henry Sturtevant.

18. Griffith, F., The significance of pneumococcal types, Epidemiology & Infection, 27,2 (1928) 113-159.

19. Avery, O.T., MacLeod, C.M. ve McCarty, M., Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III, The Journal of experimental medicine, 79,2 (1944) 137-158.

20. Hershey, A.D. ve Chase, M., Independent functions of viral protein and nucleic acid in growth of bacteriophage, Journal of general physiology, 36,1 (1952) 39-56.

21. https://plato.stanford.edu/entries/genetics/ Stanford Encyclopedia of Philosophy, Genetics. 12 Ekim 2020.

22. Unger, B., Bemerkungen zu obiger Notiz, Justus Liebigs Annalen der Chemie, 58,1 (1846) 18-20.

23. Kossel, A., Ueber eine neue Base aus dem Thierkörper, Berichte der deutschen chemischen Gesellschaft, 18,1 (1885) 79-81.

24. Kossel, A. ve Steudel, H., Weitere Untersuchungen über das Cytosin, Hoppe-Seyler s Zeitschrift für physiologische Chemie, 38,1-2 (1903) 49-59.

25. Levene, P. ve Jacobs, W., On the structure of thymus nucleic acid, Journal of Biological Chemistry, 12,3 (1912) 411-420.

26. Boivin, A., Vendrely, R. ve Vendrely, C., L'acide désoxyribonucléique du noyau cellulaire, dépositaire des caractères héréditaires; arguments d'ordre analytique, Comptes rendus hebdomadaires des seances de l'Academie des sciences, 226,13 (1948) 1061-1063.

27. Chargaff, E., Lipshitz, R. ve Green, C., Composition of the desoxypentose nucleic acids of four genera of sea-urchin, Journal of Biological Chemistry, 195,1 (1952) 155-160.

28. Franklin, R.E. ve Gosling, R.G., The structure of sodium thymonucleate fibres. I. The influence of water content, Acta Crystallographica, 6,8-9 (1953) 673-677.

29. Dickerson, R.E., The DNA helix and how it is read, Scientific American, 249,6 (1983) 94-111.

30. Fuller, W., Wilkins, M.H., Wilson, H.R. ve Hamilton, L.D., THE MOLECULAR CONFIGURATION OF DEOXYRIBONUCLEIC ACID. IV. X-RAY DIFFRACTION STUDY OF THE A FORM, Journal of molecular biology, 12 (1965) 60-76.

31. Wang, A.H.-J., Quigley, G.J., Kolpak, F.J., Crawford, J.L., Van Boom, J.H., van der Marel, G. ve Rich, A., Molecular structure of a left-handed double helical DNA fragment at atomic resolution, Nature, 282,5740 (1979) 680-686.

32. Herbert, A.G., Spitzner, J.R., Lowenhaupt, K. ve Rich, A., Z-DNA binding protein from chicken blood nuclei, Proceedings of the National Academy of Sciences, 90,8 (1993) 3339-3342.

33. Frederick, C.A., Grable, J., Melia, M., Samudzi, C., Jen-Jacobson, L., Wang, B.-C., Greene, P., Boyer, H.W. ve Rosenberg, J.M., Kinked DNA in crystalline complex with EcoRI endonuclease, Nature, 309,5966 (1984) 327-331.

34. Kool, E.T., Hydrogen bonding, base stacking, and steric effects in DNA replication, Annual review of biophysics and biomolecular structure, 30,1 (2001) 1-22.

35. https://www.bilgial.com/dna-deoksiribonukleik-asit-nedir-nasil-olusur-dnanin-gorevleri-nelerdir/ Bilgi Al, DNA (Deoksiribonükleik asit) Nedir Nasıl Oluşur? DNA'nın Görevleri Nelerdir? 15 Ekim 2020.

36. Zubay, G. ve Doty, P., The isolation and properties of deoxyribonucleoprotein particles containing single nucleic acid molecules, Journal of molecular biology, 1,1 (1959) 1-IN1.

37. Seeman, N.C., Rosenberg, J.M. ve Rich, A., Sequence-specific recognition of double helical nucleic acids by proteins, Proceedings of the National Academy of Sciences, 73,3 (1976) 804-808.

38. https://www.nature.com/scitable/definition/nucleic-acid-274/#:~:text=Nucleic%20acids%20were%20discovered%20in,novel%20type%20of%20biological%20molecule./ Scitable by nature EDUCATION, nucleic acid. 17 Ekim 2020.

39. Caspersson, T. ve Schultz, J., Ribonucleic acids in both nucleus and cytoplasm, and the function of the nucleolus, Proceedings of the National Academy of Sciences of the United States of America, 26,8 (1940) 507.

40. Ochoa, S., Künstliche Radioaktive Isotope in Physiologie Diagnostik und Therapie/Radioactive Isotopes in Physiology Diagnostics and Therapy, Enzymatic synthesis of ribonucleic acid, Springer, 960-973, 1961.

41. Holley, R.W., Apgar, J., Everett, G., Madison, J., Marquisee, M. ve Merrill, S., Penswick. JR and Zamir, Science, 147,3664 (1965) 1462-1465.

42. Siebert, S., Common sequence structure properties and stable regions in RNA secondary structures, PhD thesis, Albert-Ludwigs-University Freiburg, Institute of Computer Science, 2006.

43. Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Jou, W.M., Molemans, F., Raeymaekers, A. ve Van den Berghe, A., Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene, Nature, 260,5551 (1976) 500-507.

44. Napoli, C., Lemieux. C. and Jorgensen. R.(1990) Introduction of a chimeric chalcone sy nthase gene into petunia results in reversible co-suppression of homologous genes in trans, Plant Cell, 2,4 279-289.

45. Dafny-Yelin, M., Chung, S.-M., Frankman, E.L. ve Tzfira, T., pSAT RNA interference vectors: a modular series for multiple gene down-regulation in plants, Plant physiology, 145,4 (2007) 1272-1281.

46. https://en.wikipedia.org/wiki/RNA/ Wikipedia, RNA. 17 Ekim 2020.

47. https://www.shutterstock.com/pl/image-vector/vector-scientific-icon-spiral-rna-illustration-1145441705/ shutterstock, Vector scientific icon spiral RNA. Illustration of the structure of the RNA molecule. 18 Ekim 2020.

48. https://www.technologynetworks.com/genomics/lists/what-are-the-key-differences-between-dna-and-rna-296719/ Genomics Research from TECHNOLOGY NETWORKS, DNA vs. RNA – 5 Key Differences and Comparison. 20 Ekim 2020.

49. Mubayiwa, D., Bio-mirrors and networking security: for the partial fulfilment of Masters of Information Sciences, Information Systems major, 2006, Massey University, 2006.

50. Tomás, A.V., Michael, S.P., Dombrovskaia, L., Jorge, A.A., Felipe, C.B., Diego, C.C., Freddy, M.R., Valeria, L.R., Agulló, L. ve Macarena, C.H., Bioinformatics integration framework for metabolic pathway data-mining, International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems 2006: 917-926.

51. Chung, S.Y. ve Wong, L., Kleisli: a new tool for data integration in biology, Trends in Biotechnology, 17,9 (1999) 351-355.

52. Varfolomeev, S., Uporov, I. ve Fedorov, E., Bioinformatics and molecular modeling in chemical enzymology. Active sites of hydrolases, Biochemistry (Moscow), 67,10 (2002) 1099-1108.

53. Liu, M. ve Grigoriev, A., Fast parsers for Entrez Gene, Bioinformatics, 21,14 (2005) 3189-3190.

54. Chen, Y.-P.P., Bioinformatics technologies, Springer Science & Business Media, 2005.

55. Park, J.A., Lee, C.S. ve Park, J.C., Information visualization with text data mining for knowledge discovery tools in bioinformatics, Key Engineering Materials 2005, 277: 259-265.

56. Davenport, G., Ellis, N., Ambrose, M. ve Dicks, J., Using bioinformatics to analyse germplasm collections, Euphytica, 137,1 (2004) 39-54.

57. Richardson, D.E., Vanwye, J.D., Exum, A.M., Cowen, R.K. ve Crawford, D.L., High-throughput species identification: from DNA isolation to bioinformatics, Molecular Ecology Notes, 7,2 (2007) 199-207.

58. Thompson, J.D. ve Poch, O., Multiple sequence alignment as a workbench for molecular systems biology, Current Bioinformatics, 1,1 (2006) 95-104.

59. Achard, F., Vaysseix, G. ve Barillot, E., XML, bioinformatics and data integration, Bioinformatics, 17,2 (2001) 115-125.

60. Abbas, A.E. ve Holmes, S.P., Bioinformatics and management science: Some common tools and techniques, Operations Research, 52,2 (2004) 165-190.

61. Okubo, K., Sugawara, H., Gojobori, T. ve Tateno, Y., DDBJ in preparation for overview of research activities behind data submissions, Nucleic acids research, 34,suppl_1 (2006) D6-D9.

62. http://www.insdc.org/ International Nucleotide Sequence Database Collaboration, INSDC policies. 22 Ekim 2020.

63. https://www.ncbi.nlm.nih.gov/genbank/ National Center for Biotechnology Information, GenBank. 23 Ekim 2020.

64. Bilofsky, H.S., Burks, C., Fickett, J.W., Goad, W.B., Lewitter, F.I., Rindone, W.P., Swindell, C.D. ve Tung, C.-S., The GenBank genetic sequence databank, Nucleic acids research, 14,1 (1986) 1-4.

65. https://www.ncbi.nlm.nih.gov/genbank/statistics/ National Center for Biotechnology Information, GenBank and WGS Statistics. 25 Ekim 2020.

66. https://www.ddbj.nig.ac.jp/index-e.html/ Bioinformation and DDBJ Center, DDBJ. 26 Ekim 2020.

67. https://www.ebi.ac.uk/ena/browser/home/ European Nucleotide Archive, EMBL-EBI. 25 Ekim 2020.

68. Suominen, I. ve Ollikka, P., Yhdistelmä-DNA-tekniikan perusteet, Opetushallitus, 1997.

69. Needleman, S.B. ve Wunsch, C.D., A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of molecular biology, 48,3 (1970) 443-453.

70. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet/ National Human Genome Research Institute, DNA Sequencing Fact Sheet. 1 Kasım 2020.

71. Voelkerding, K.V., Dames, S.A. ve Durtschi, J.D., Next-generation sequencing: from basic research to diagnostics, Clinical chemistry, 55,4 (2009) 641-658.

72. McCarthy, A., Third generation DNA sequencing: pacific biosciences' single molecule real time technology, Chemistry & biology, 17,7 (2010) 675-676.

73. Schadt, E.E., Turner, S. ve Kasarskis, A., A window into third-generation sequencing, Human molecular genetics, 19,R2 (2010) R227-R240.

74. https://www.thermofisher.com/tr/en/home/life-science/sequencing/sequencing-learning-center/sequencing-basics/dna-sequencing-technologies.html/ THERMO FISHER SCIENTIFIC, DNA sequencing technologies. 2 Kasım 2020.

75. Pareek, C.S., Smoczynski, R. ve Tretyn, A., Sequencing technologies and genome sequencing, Journal of applied genetics, 52,4 (2011) 413-435.

76. Maxam, A.M. ve Gilbert, W., A new method for sequencing DNA, Proceedings of the National Academy of Sciences, 74,2 (1977) 560-564.

77. LeBlanc, V.G. ve Marra, M.A., Next-generation sequencing approaches in cancer: where have they brought us and where will they take us?, Cancers, 7,3 (2015) 1925-1958.

78. Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., McCombie, W.R. ve Schatz, M.C., Third-generation sequencing and the future of genomics, BioRxiv, (2016) 048603.

79. Schloss, J.A., How to get genomes at one ten-thousandth the cost, Nature biotechnology, 26,10 (2008) 1113-1115.

80. Leary, R.J., Kinde, I., Diehl, F., Schmidt, K., Clouser, C., Duncan, C., Antipova, A., Lee, C., McKernan, K. ve Francisco, M., Development of personalized tumor biomarkers using massively parallel sequencing, Science translational medicine, 2,20 (2010) 20ra14-20ra14.

81. Shah, S.P., Morin, R.D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R. ve Senz, J., Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution, Nature, 461,7265 (2009) 809-813.

82. Lips, E.H., Michaut, M., Hoogstraat, M., Mulder, L., Besselink, N.J., Koudijs, M.J., Cuppen, E., Voest, E.E., Bernards, R. ve Nederlof, P.M., Next generation sequencing of triple negative breast cancer to find predictors for chemotherapy response, Breast Cancer Research, 17,1 (2015) 1-10.

83. Mantere, T., Winqvist, R., Kauppila, S., Grip, M., Jukkola-Vuorinen, A., Tervasmäki, A., Rapakko, K. ve Pylkäs, K., Targeted next-generation sequencing identifies a recurrent mutation in MCPH1 associating with hereditary breast cancer susceptibility, PLoS genetics, 12,1 (2016) e1005816.

84. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J. ve Turner, S.W., Direct detection of DNA methylation during single-molecule, real-time sequencing, Nature methods, 7,6 (2010) 461.

85. Tieva, A. ve Peltomäki, P., Epigeneettiset muutokset syövässä, Duodecim, 128,1 (2012) 62-71.

86. Ding, S., Chen, X. ve Shen, K., Single-cell RNA sequencing in breast cancer: Understanding tumor heterogeneity and paving roads to individualized therapy, Cancer Communications, 40,8 (2020) 329-344.

87. Sanger, F., Nicklen, S. ve Coulson, A.R., DNA sequencing with chain-terminating inhibitors, Proceedings of the National Academy of Sciences, 74,12 (1977) 5463-5467.

88. Strachan, T. ve Read, A.P., Human molecular genetics 2/; Tom Strachan and Andrew P. Read, (1999).

89. Hudspith, K.A.Z., Application of genomic technologies for molecular diagnosis of genetic diseases, University of Oxford, 2015.

90. Russel, D. ve Sambrook, J., Molecular Cloning volume 2: A Laboratory Manual. (2001).

91. https://www.mgbiyoinformatik.com/sanger-dizileme/ MG BİYOİNFORMATİK, Sanger Dizileme. 15 Kasım 2020.

92. https://eu.idtdna.com/pages/ INTEGRATED DNA TECHNOLOGIES, DNA sequencing. 20 Kasım 2020.

93. Breathnach, R., Mandel, J.-L. ve Chambon, P., Ovalbumin gene is split in chicken DNA, Nature, 270,5635 (1977) 314-319.

94. Jeffreys, A.J. ve Flavell, R., The rabbit β-globin gene contains a large insert in the coding sequence, Cell, 12,4 (1977) 1097-1108.

95. Breathnach, R., Benoist, C., O'hare, K., Gannon, F. ve Chambon, P., Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries, Proceedings of the National Academy of Sciences, 75,10 (1978) 4853-4857.

96. Staden, R., A strategy of DNA sequencing employing computer programs, Nucleic acids research, 6,7 (1979) 2601-2610.

97. DeBoer, J.A., Fontaine, J.J., Chizinski, C.J., Pope, K.L., Sandbars, M.E., Kinzel, P., Bauer, M., Feller, M., Holmquist-Johnson, C. ve Preston, T., Natural Sciences, (2004).

98. https://www.genome.gov/genetics-glossary/Shotgun-Sequencing/ National Human Genome Research Institute, Shotgun Sequencing. 19 Kasım 2020.

99. Kadri, K., Perspectives on Polymerase Chain Reaction, Polymerase Chain Reaction (PCR): Principle and Applications, IntechOpen, 2019.

100. Mullis, K.B., The unusual origin of the polymerase chain reaction, Scientific American, 262,4 (1990) 56-65.

101. van Pelt-Verkuil, E., van Belkum, A. ve Hays, J.P., A brief comparison between in vivo DNA replication and in vitro PCR amplification, Principles and Technical Aspects of PCR Amplification, (2008) 9-15.

102. Somma, M. ve Querci, M., The Analysis of Food Samples for the Presence of Genetically Modified Organisms.

103. Haydel, S. ve Stout, V., A Kinesthetic Modeling Activity to Teach PCR Fundamentals. CourseSource. (2015).

104. Margulies M, Egholm M, Altman WE; ve diğerleri, https://tr.wikipedia.org/wiki/DNA_dizileme#Lynx_Therapeutics'in_Kitlesel_Paralel_%C4%B0mza_Dizilemesi/ Wikipedia, DNA Dizileme. 26 Kasım 2020.

105. Heather, J.M. ve Chain, B., The sequence of sequencers: The history of sequencing DNA, Genomics, 107,1 (2016) 1-8.

106. Huszar, T.I., Massively parallel sequencing of forensic markers: sequence variation and forensic application, University of Leicester, 2020.

107. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data/ National Human Genome Research Institute, DNA Sequencing Costs: Data. 27 Kasım 2020.

108. Goodwin, S., McPherson, J.D. ve McCombie, W.R., Coming of age: ten years of next-generation sequencing technologies, Nature Reviews Genetics, 17,6 (2016) 333.

109. Johnsen, J.M., Nickerson, D.A. ve Reiner, A.P., Massively parallel sequencing: the new frontier of hematologic genomics, Blood, 122,19 (2013) 3268-3275.

110. https://en.wikipedia.org/wiki/Polony_sequencing#DNA_sequencing/ Wikipedia, Polony sequencing. 21 Kasım 2020.

111. Mitra, R.D., Polony sequencing: DNA sequencing technology and a computational analysis reveals chromosomal domains of gene expression, Massachusetts Institute of Technology, 2000.

112. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. ve Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic acids research, 25,17 (1997) 3389-3402.

113. https://www.enterprise.cam.ac.uk/case-studies/solexa-second-generation-genetic-sequencing/ UNIVERSITY OF CAMBRIDGE enterprise, Solexa: second-gen genetic sequencing. 25 Kasım 2020.

114. https://www.ch.cam.ac.uk/collaboration-and-impact/solexa-sequencing/ UNIVERSITY OF CAMBRIDGE, Solexa Sequencing. 25 Kasım 2020.

115. Mardis, E.R., Next-generation DNA sequencing methods, Annu. Rev. Genomics Hum. Genet., 9 (2008) 387-402.

116. Mardis, E.R., The impact of next-generation sequencing technology on genetics, Trends in genetics, 24,3 (2008) 133-141.

117. Logares, R., Haverkamp, T.H., Kumar, S., Lanzén, A., Nederbragt, A.J., Quince, C. ve Kauserud, H., Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches, Journal of microbiological methods, 91,1 (2012) 106-113.

118. Reuter, J.A., Spacek, D.V. ve Snyder, M.P., High-throughput sequencing technologies, Molecular cell, 58,4 (2015) 586-597.

119. https://en.wikipedia.org/wiki/DNA_nanoball_sequencing/ Wikipedia, DNA nanoball sequencing. 26 Kasım 2020.

120. http://www.completegenomics.com/documents/revolocity-tech-overview.pdf/ Complete GENOMICS, Whole Genome Sequencing Technology Overview. 27 Kasım 2020.

121. Rothberg, J.M. ve Leamon, J.H., The development and impact of 454 sequencing, Nature biotechnology, 26,10 (2008) 1117-1124.

122. Haldar, K.S., Profiling of bacterial communities in chronic obstructive pulmonary disease, University of Leicester, 2015.

123. Ronaghi, M., Uhlén, M. ve Nyrén, P., A sequencing method based on real-time pyrophosphate, Science, 281,5375 (1998) 363-365.

124. Holt, R.A. ve Jones, S.J., The new paradigm of flow cell sequencing, Genome research, 18,6 (2008) 839-846.

125. Metzker, M.L., Sequencing technologies—the next generation, Nature Reviews Genetics, 11,1 (2010) 31-46.

126. Charlson, E.S., Bittinger, K., Haas, A.R., Fitzgerald, A.S., Frank, I., Yadav, A., Bushman, F.D. ve Collman, R.G., Topographical continuity of bacterial populations in the healthy human respiratory tract, American journal of respiratory and critical care medicine, 184,8 (2011) 957-963.

127. Pragman, A.A., Kim, H.B., Reilly, C.S., Wendt, C. ve Isaacson, R.E., The lung microbiome in moderate and severe chronic obstructive pulmonary disease, PloS one, 7,10 (2012) e47305.

128. Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J. ve Levy, S., Evaluation of next generation sequencing platforms for population targeted sequencing studies, Genome biology, 10,3 (2009) 1-13.

129. Pandey, V., Nutter, R.C. ve Prediger, E., Applied biosystems solid™ system: ligation-based sequencing, Next Generation Genome Sequencing: Towards Personalized Medicine, (2008) 29-42.

130. Sibthorp, C., Analysis of the Aspergillus nidulans transcriptome using high-throughput RNA sequencing, Citeseer, 2012.

131. Porter, J.S., Mapping bisulfite-treated short DNA reads, Virginia Tech, 2018.

132. Lahens, N.F., Ricciotti, E., Smirnova, O., Toorens, E., Kim, E.J., Baruzzo, G., Hayer, K.E., Ganguly, T., Schug, J. ve Grant, G.R., A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression, BMC genomics, 18,1 (2017) 1-13.

133. https://www.thermofisher.com/tr/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html/ THERMO FISHER SCIENTIFIC, Ion Torrent Next-Generation Sequencing Technology. 28 Kasım 2020.

134. Otlu, B., Klimud , https://www.klimud.org/public/kongre2013pdf/Yen%C4%B1%20jenarasyon%20sekanslama-Bar%C4%B1%C5%9F%20Otlu.pdf/ Yeni Nesil Dizileme Sistemleri 29 Kasım 2020.

135. Tork, B.A., Viral quasispecies reconstruction using next generation sequencing reads, (2013).

136. Ari, Ş. ve Arikan, M., Plant omics: Trends and applications, Next-generation sequencing: advantages, disadvantages, and future, Springer, 109-135, 2016.

137. Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A. ve Takahashi, H., Sequence-specific error profile of Illumina sequencers, Nucleic acids research, 39,13 (2011) e90-e90.

138. Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T. ve Quince, C., Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform, Nucleic acids research, 43,6 (2015) e37-e37.

139. Turner, I.H., Discovering genetic variation in populations using next generation sequencing and de novo assembly, University of Oxford, 2018.

140. Medvedev, P., Scott, E., Kakaradov, B. ve Pevzner, P., Error correction of high-throughput sequencing datasets with non-uniform coverage, Bioinformatics, 27,13 (2011) i137-i141.

141. Simpson, J.T. ve Durbin, R., Efficient de novo assembly of large genomes using compressed data structures, Genome research, 22,3 (2012) 549-556.

142. Ilie, L. ve Molnar, M., RACER: rapid and accurate correction of errors in reads, Bioinformatics, 29,19 (2013) 2490-2493.

143. Çaylak Kayaturan, G., Representing shortest paths in graphs using Bloom filters without false positives and applications to routing in computer networks, University of Essex, 2018.

144. Bloom, B.H., Space/time trade-offs in hash coding with allowable errors, Communications of the ACM, 13,7 (1970) 422-426.

145. Boldyreva, A., Threshold signatures, multisignatures and blind signatures based on the gap-Diffie-Hellman-group signature scheme, International Workshop on Public Key Cryptography 2003: 31-46.

146. Melsted, P. ve Pritchard, J.K., Efficient counting of k-mers in DNA sequences using a bloom filter, BMC bioinformatics, 12,1 (2011) 1-7.

9. EKLER

Ek Tablo 1. DNA dizisi örneği

Yön 1														
DNA Dizisi	GAT	CCT	GAC	CAT	GAA	CCT	AAG	CTT	CTT	CGA	CCA	ATT		
Amino Asit	Asp	Leu	Asp	His	Glu	Pro	Lys	Leu	Leu	Arg	Pro	lle		
Birinci Dizi = (Asp, Leu, Asp, His, Glu, Pro, Lys, Leu, Leu, Arg, Pro, Ile)														
DNA Dizisi	G	ATC	CTG	ACC	ATG	AAC	CTA	AGC	TTC	TTC	GAC	CAA	T	T
Amino Asit		Ile	Leu	Thr	Met	Asn	Leu	Ser	Phe	Phe	Asp	Gln		
İkinci Dizi = (Ile, Leu, Thr, Met, Asn, Leu, Ser, Phe, Phe, Asp, Gln)														
DNA Dizisi	G	A	TCC	TGA	CCA	TGA	ACC	TAA	GCT	TCT	TCG	ACC	AAT	T
Amino Asit			Ile	Leu	Thr	Met	Asn	Leu	Ser	Phe	Phe	Asp	Gln	
Üçüncü Dizi = (Ser, STOP, Pro, STOP, Thr, STOP, Ala, Ser, Ser, Thr, Asn)														
Tersine ve Yön 2														
DNA Dizisi	TTA	ACC	AGC	TTC	TTC	GAA	TCC	AAG	TAC	CAG	TCC	TAG		
Amino Asit	Leu	Thr	Ser	Phe	Phe	Glu	Ser	Lys	Tyr	Gln	Ser	STOP		
Dördüncü Dizi = (Leu, Thr, Ser, Phe, Phe, Glu, Ser, Lys, Tyr, Gln, Ser, STOP)														
DNA Dizisi	T	TAA	CCA	GCT	TCT	TCG	AAT	CCA	AGT	ACC	AGT	CCT	A	G
Amino Asit		STOP	Pro	Ala	Ser	Ser	Asn	Pro	Ser	Thr	Ser	Pro		
Beşinci Dizi = (STOP, Pro, Ala, Ser, Ser, Asn, Pro, Ser, Thr, Ser, Pro)														

Ek Tablo 1’in devamı

Tersine ve Yön 2														
DNA Dizisi	T	T	AAC	CAG	CTT	CTT	CGA	ATC	CAA	GTA	CCA	GTC	CTA	G
Amino Asit			Asn	Gln	Leu	Leu	Arg	Ile	Gln	Val	Pro	Val	Leu	
Altıncı Dizi = (Asn, Gln, Leu, Leu, Arg, Ile, Gln, Val, Pro, Val, Leu)														

Ek Tablo 2. BRAF veri seti ve kullanılan öznitelikler

BRAF Gen Verisi: 205599						
Sekans Boyutu	**K-mer Boyutu**	**Derinlik**	**Hash Fonksiyon Sayısı**	**Sağlam Bloom Filter Boyutu**	**Hatalı Bloom Filter Boyutu**	**Çalışma Süresi(sn.)**
50	16	30	4	13157376	276324160	189,62
50	16	30	5	16446720	345405248	307,57
50	16	30	6	19736064	414486272	327,19
50	16	40	4	13157376	368432960	330,62
50	16	40	5	16446720	460541248	441,53
50	16	40	6	19736064	552649472	455,16
50	16	50	4	13157376	460541760	384,71
50	16	50	5	16446720	575677248	532,56
50	16	50	6	19736064	690812672	544,24
50	20	30	4	16446400	305930368	236,54
50	20	30	5	20558016	382412928	367,72
50	20	30	6	24669632	458895488	381,56
50	20	40	4	16446400	407907968	353,19
50	20	40	5	20558016	509884928	514,62
50	20	40	6	24669632	611861888	558,51
50	20	50	4	16446400	509885568	453,66
50	20	50	5	20558016	637356928	691,12
50	20	50	6	24669632	764828288	696,18
50	24	30	4	19735296	319746560	286,36
50	24	30	5	24669120	399683200	438,62

Ek Tablo 2’in devamı

Sekans Boyutu	K-mer Boyutu	Derinlik	Hash Fonksiyon Sayısı	Sağlam Bloom Filter Boyutu	Hatalı Bloom Filter Boyutu	Çalışma Süresi(sn.)
50	24	30	6	29602944	479619840	464,74
50	24	40	4	19735296	426329600	387,79
50	24	40	5	24669120	532912000	642,67
50	24	40	6	29602944	639494400	667,75
50	24	50	4	19735296	532912640	435,65
50	24	50	5	24669120	666140800	677,54
50	24	50	6	29602944	799368960	707,41
75	16	30	4	13157376	315797760	138,14
75	16	30	5	16446720	394747200	215,47
75	16	30	6	19736064	473696640	231,1
75	16	40	4	13157376	421063680	286,83
75	16	40	5	16446720	526329600	308,706
75	16	40	6	19736064	631595520	325,18
75	16	50	4	13157376	526333440	262,3
75	16	50	5	16446720	657916800	397,16
75	16	50	6	19736064	789500160	428,81
75	20	30	4	16446400	368430720	182,77
75	20	30	5	20558016	460538432	304,56
75	20	30	6	24669632	552646080	316,14
75	20	40	4	16446400	491240960	266,55
75	20	40	5	20558016	614051200	417,2
75	20	40	6	24669632	736861440	432,12
75	20	50	4	16446400	614055680	346,13
75	20	50	5	20558016	767569600	542,39

Ek Tablo 2’in devamı

Sekans Boyutu	K-mer Boyutu	Derinlik	Hash Fonksiyon Sayısı	Sağlam Bloom Filter Boyutu	Hatalı Bloom Filter Boyutu	Çalışma Süresi(sn.)
75	20	50	6	24669632	921083520	548,37
75	24	30	4	19735296	410537088	226,93
75	24	30	5	24669120	513171392	363,71
75	24	30	6	29602944	615805632	396,22
75	24	40	4	19735296	547382784	339,45
75	24	40	5	24669120	684228480	533,57
75	24	40	6	29602944	821074176	534,64
75	24	50	4	19735296	684233472	420,29
75	24	50	5	24669120	855291840	668,73
75	24	50	6	29602944	1026350208	691,16
100	16	30	4	13157376	335533760	125,29
100	16	30	5	16446720	419417216	205,21
100	16	30	6	19736064	503300672	211,67
100	16	40	4	13157376	447380160	186,52
100	16	40	5	16446720	559225216	288,38
100	16	40	6	19736064	671070272	294,31
100	16	50	4	13157376	559226560	236,24
100	16	50	5	16446720	699033216	370,66
100	16	50	6	19736064	838839872	379,47
100	20	30	4	16446400	399679936	165,13
100	20	30	5	20558016	499599936	273,63
100	20	30	6	24669632	599519936	290,59
100	20	40	4	16446400	532908736	247,24
100	20	40	5	20558016	666135936	383,81

Ek Tablo 2’in devamı

Sekans Boyutu	K-mer Boyutu	Derinlik	Hash Fonksiyon Sayısı	Sağlam Bloom Filter Boyutu	Hatalı Bloom Filter Boyutu	Çalışma Süresi(sn.)
100	20	40	6	24669632	799363136	392,92
100	20	50	4	16446400	666137536	315,31
100	20	50	5	20558016	832671936	485,96
100	20	50	6	24669632	999206336	511,56
100	24	30	4	19735296	455931200	214,57
100	24	30	5	24669120	569913984	346,78
100	24	30	6	29602944	683896768	368,66
100	24	40	4	19735296	607910720	306,42
100	24	40	5	24669120	759888384	477,24
100	24	40	6	29602944	911866048	498,14
100	24	50	4	19735296	759890240	387,55
100	24	50	5	24669120	949862784	632,15
100	24	50	6	29602944	1139835328	659,49

ÖZGEÇMİŞ

İlköğretim eğitimini İsmet Paşa ilköğretim Okulu'nda, Ortaöğretim eğitimini Özel İstiklal Koleji'nde tamamlamıştır. Lise eğitimini Erzurum Anadolu Lisesi'nde tamamlamıştır. 2011 güz döneminde Atatürk Üniversitesi Bilgisayar Mühendisliği Bölümü'nde başladığı lisans eğitiminden 2015 bahar döneminde mezun olmuştur. 2017 yılının güz döneminde Karadeniz Teknik Üniversitesi Bilgisayar Mühendisliği Bölümü'nde yüksek lisans eğitimine başlamıştır. Şu an yüksek lisans öğrencisi olarak devam etmektedir.