KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

GENETİK DİZİLERDEKİ HATALARI DÜZELTMEK İÇİN YENİ BİR YAKLAŞIM

YÜKSEK LİSANS TEZİ

Elif ARAS

ARALIK 2018
TRABZON

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

GENETİK DİZİLERDEKİ HATALARI DÜZELTMEK İÇİN YENİ BİR YAKLAŞIM

Elif ARAS

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

Tezin Enstitüye Verildiği Tarih : 06 / 12 /2018
Tezin Savunma Tarihi : 28 / 12 /2018

Tez Danışmanı : Dr. Öğr. Üyesi İbrahim SAVRAN

Trabzon 2018

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Bilgisayar Mühendisliği Anabilim Dalında
Elif ARAS Tarafından Hazırlanan

GENETİK DİZİLERDEKİ HATALARI DÜZELTMEK İÇİN YENİ BİR YAKLAŞIM

başlıklı bu çalışma, Enstitü Yönetim Kurulunun 11 / 12 / 2018 gün ve 1782 sayılı kararıyla oluşturulan jüri tarafından yapılan sınavda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri

Başkan : Prof. Dr. Ali KARCI .....

Üye : Dr. Öğr. Üyesi İbrahim SAVRAN .....

Üye : Dr. Öğr. Üyesi Eyüp GEDİKLİ .....

Prof. Dr. Sadettin KORKMAZ
Enstitü Müdürü

ÖNSÖZ

Bu tez çalışmasında DNA dizileme işlemi sırasında oluşan hataları giderip daha doğru dizilimi elde etmek için yeni nesil cihazlarda kullanılabilecek yeni bir hata düzeltme algoritması önerilmiştir. Önerilen algoritma ile genetik bozukluklara yol açan DNA mutasyonlarının, mikrobiyal hastalıklara sebebiyet veren yeni mikroorganizma tespitinin ve biyokimyasal alanlardaki diğer sorulara çözüm bulma yolunda daha doğru sonuca ulaşılmasına yardım edilmesi amaçlanmaktadır. Bu çalışmanın biyoinformatik alanına katkı sağlaması ve yeni çalışmalara ışık tutması beklenmektedir.

Bu çalışmanın gerçekleştirilmesinde, bilgi ve tecrübesini benimle paylaşan, rehberliği ve talimatlarıyla sadece destekleyici değil aynı zamanda cesaret verici olan, çalışma boyunca beni motive eden değerli danışman hocam Dr. Öğr. Üyesi İbrahim SAVRAN'a şükranlarımı sunmayı bir borç bilirim. Çalışma boyunca benden bir an olsun sevgi, ilgi ve değerli bilgilerini esirgemeyen, karşılaştığım tüm zorlukları benimle göğüsleyen yol arkadaşım, meslektaşım, eşim Sefa ARAS'a ve hayatımın tüm evresinde benimle olan her türlü destek, sevgi ve şefkatlerini eksik etmeyen aileme sonsuz teşekkür ederim.

Elif ARAS

Trabzon 2018

TEZ ETİK BEYANNAMESİ

Yüksek Lisans Tezi olarak sunduğum “Genetik Dizilerdeki Hataları Düzeltmek için Yeni Bir Yaklaşım” başlıklı bu çalışmayı baştan sona kadar danışmanım Dr. Öğr. Üyesi İbrahim SAVRAN’ın sorumluluğunda tamamladığımı, verileri kendim topladığımı, deneyleri 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. 28/12/2018

Elif ARAS

İÇİNDEKİLER

Sayfa No

Yüksek Lisans

ÖZET

GENETİK DİZİLERDEKİ HATALARI DÜZELTMEK İÇİN YENİ BİR YAKLAŞIM

Elif ARAS

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Dr. Öğr. Üyesi İbrahim SAVRAN
2018, 80 Sayfa

Genetik, genom bilgisini öğrenip canlıların tüm yapı ve aktivitelerini incelemeyi hedefleyen bir bilimdir. Bu amaçla 20.yy başından günümüze kadar canlıların genom dizilerinin elde edilmesine çalışılmaktadır. DNA dizilimi; adli tıp, mikrobiyoloji, tıbbı tanı koyma, genetik hastalıkların tespiti ve biyokimyasal alanındaki problemlere çözüm bulmak için kullanılmaktadır. DNA dizileme çalışmalarına, geleneksel dizileme yöntemleriyle başlanmış olup, bu yöntemler düşük hatalı dizilim elde etse de uzun sürede ve yüksek maliyette kısa parçaları dizilemeye daha uygundur. Bu yöntemlerin dezavantajlarını gidermek amacıyla kısa sürede ve düşük maliyette tüm genom bilgisini elde edebilecek yeni nesil dizileme yöntemleri geliştirilmiştir. Ancak yeni nesil dizileme yöntemlerinin hata oranları geleneksel yöntemlere göre fazladır. Bu problemi çözmek amacıyla, bu çalışmada yeni nesil dizileme yöntemleriyle elde edilen genom bilgisindeki okuma hatalarını tespit edip düzeltecek yeni bir algoritma önerilmiştir. Önerilen algoritmada k-mer yaklaşımı kullanılarak okunan sekanslar gruplandırılmıştır. Aynı bölgeyi temsil eden sekanslarda çoğunluk oylaması yapılarak, hatalı nükleotidler doğru nükleotidlerle değiştirilerek sekanslar güncellenmektedir. Önerilen hata düzeltme algoritması farklı veri setleri üzerinde test edilmiş olup mevcut hata düzeltme algoritmalarına çok yakın veya daha iyi sonuçlar elde ettiği gözlemlenmiştir. Önerilen algoritmanın duyarlılık değeri [97,00-98,18] özgüllük değeri ise [99,62-99,88] aralığında bulunmaktadır. Literatürdeki mevcut algoritmalarda duyarlılık değeri [96,60-99,99], özgüllük değeri ise [48,81-100,00] aralığında değişmektedir.

Anahtar Kelimeler: DNA dizileme, Biyoinformatik, Yeni nesil dizileme, Hata düzeltme, K-mer

Master Thesis

SUMMARY

A NEW APPROACH TO CORRECT ERRORS IN THE GENETIC SEQUENCE

Elif ARAS

Karadeniz Technical University
The Graduate School of Natural and Applied Sciences
Computer Engineering Graduate Program
Supervisor: Asst. Prof. Dr. İbrahim SAVRAN
2018, 80 Pages

Genetics is a science that aims to learn the genome knowledge and study all structures and activities of living things. For this purpose, genome sequences of living beings have been studied from the beginning of the 20th century to the present. DNA sequencing; forensic medicine, microbiology, medical diagnosis, detection of genetic diseases and biochemical problems are used to find solutions. DNA sequencing studies have begun with traditional sequencing methods, and these methods are more suitable for sequencing short fragments in a long time and at high cost, even if they are misregulated. In order to eliminate the disadvantages of these methods, next generation sequencing methods have been developed which can obtain whole genome information in a short time and at low cost. However, the error rates of the next generation sequencing methods are higher than the traditional methods. In order to solve this problem, a new algorithm has been proposed to detect and correct the reading errors in genome information obtained by next generation sequencing methods in this study. In the proposed algorithm, the sequences read using the k-mer approach are grouped. In the sequences representing the same region, the majority of the sequences are performed and the sequences are updated by replacing the faulty nucleotides with the correct nucleotides. The proposed error correction algorithm has been tested on different datasets and it has been observed that it provides very close or better results to existing error correction algorithms. The sensitivity value of the proposed algorithm is [97,00-98,18] and the specificity is in the range of [99,62-99,88]. In the current algorithms in the literature, the sensitivity value is [96,60-99,99] and the specificity is in the range [48,81-100,00].

Key Words: DNA sequencing, Bioinformatics, Next generation sequencing, Error correction, K-mer

ŞEKİLLER DİZİNİ

Sayfa No

TABLOLAR DİZİNİ

Sayfa No

SEMBOLLER DİZİNİ

A	Adenin
C	Sitozin
CPU	Central Process Unit
CUDA	Compute Unified Device Architecture
DNA	Deoksiribo Nükleik Asit
FN	False Negative
FP	False Positive
G	Guanin
GPU	Graphic Process Unit
MPI	Message Passing Interface
NCBI	National Center for Biotechnology Information
RNA	Ribo Nükleik Asit
T	Timin
TN	True Negative
TP	True Positive
tRNA	Taşıyıcı Ribo Nükleik Asit
U	Urasil
YND	Yeni Nesil Dizileme

1. GENEL BİLGİLER

1.1. Giriş

Genetik alanda yapılan çalışmaların temeli 1866 yılında genetiğin babası olarak adlandırılan Gregor Mendel'in bezelye bitkisi üzerinde gerçekleştirdiği melezleme deneyleriyle başlamıştır [1]. Bu deneylerde, uzun ve kısa boylu bezelye bitkilerinin döllenmeleri sonucunda elde edilen yeni bireyler incelendiğinde atadan çocuğa aktarılan özelliklerde kalıtım tespit edilmiştir. Ayrıca gerçekleştirilen deneyler sırasında günümüzde kullanılan başlıca terimlerin temeli atılmıştır [2]. Örneğin Gregor Mendel tarafından faktör olarak adlandırılan boy, renk bilgisi günümüzde "gen" olarak adlandırılmaktadır [3]. Sonraki çalışmalar hücre bölünmesi ve yapısı üzerine yoğunlaşarak modern genetiğin temelleri atılmıştır.

1869 yılında Friedrich Miescher tarafından hücrelerin çekirdeklerinde günümüzde "nükleik asit" olarak adlandırılan asidik özellikli yeni maddeler keşfedilmiştir [2].

1944 yılında Oswald Avery yaptığı çalışmalarla kalıtsal bilginin DNA (Deoksiribo Nükleik Asit) ile birlikte taşındığı ortaya atılmıştır [4]. Böylece DNA'nın parçası olan gen terimi de anlam kazanmıştır. Watson ve Crick'in 1953 yılında DNA'nın çift sarmal yapıda olduğu bilgisini keşfedip ilan etmeleriyle birlikte nükleik asitlerin bulunması daha da anlam kazanarak DNA dizilim sisteminin temel yapısı bulunmuştur [5].

1965 yılında Robert Holley 74 nükleotidden oluşan tRNA molekülünün dizilimini bularak analiz etmiştir [6]. Bu aşamadan sonra bilim insanları, DNA analizini daha kolay yapabilmek için yeni DNA analiz yöntemleri keşfetmeye çalışmışlardır. Çünkü doğru DNA analiziyle birlikte genetik koda ulaşılarak genetik hastalık tespiti, ilaçların icadı, kimlik tespiti gibi problemlere çözüm üretilmesi amaçlanmaktadır.

1977 de Allan Maxam, Wilter Gilbert ve Frederizck Sanger tarafından DNA dizilme yöntemleri keşfedilmiştir [7, 8]. Maxam ve Gilbert birinci nesil DNA dizileme yöntemi olarak adlandırılan kimyasal kırılma yöntemini, Sanger ve arkadaşları ise birinci yönteme paralel olarak daha uzun DNA parçalarının dizilimini sağlayan zincir sonlandırma yöntemini geliştirmiştir [7]. Sanger yöntemi daha az kimyasal ve radyoaktivite gerektirdiği için daha yaygın olarak kullanılmaktadır [8]. Sanger dizileme yöntemi, yeni nesil dizileme (YND) tekniklerinin ortaya çıkmasına kadar en yaygın kullanılan yöntem olarak bilinmektedir.

YND tekniklerinde, teknolojiyle birlikte dizileme işlemi için üretilen yeni cihazlar ve geliştirilen yazılımlarla birlikte uzun DNA parçalarının depolanması ve analiz edilmesi kolaylaşmıştır. 1990 yılında uluslararası işbirliği ve birçok özel firmanın katkılarıyla insan DNA dizilimini bulmak, genetik haritayı çıkarmak için "insan genom projesi" başlatılmıştır [9]. 2001 yılında insan genom bilgisinin %96'sının elde edildiği ilan edilmiştir [9]. Buna göre; insan DNA diziliminin yaklaşık olarak üç milyar nükleotidden oluştuğu sonucuna varılmaktadır [9]. Bu bilgiler ışığında bir yandan eksik DNA bilgisi tamamlanmaya çalışılırken bir yandan da elde edilen uzun verilerin analiz edilmesi problemine çözüm arayışı incelenmektedir. Çünkü geleneksel Sanger yöntemiyle elde edilen büyük verileri dizinleyip analiz etmek çok uzun zaman almakta ve veriyi işlemekte yetersiz bulunmaktadır. Geliştirilen yeni yöntemler literatürde YND yöntemleri olarak geçmektedir [8-12].

1.2. Çalışmanın Amacı ve Önemi

Geliştirilen YND yöntemleriyle daha hızlı, güvenilir ve istikrarlı çalışmasının yanı sıra DNA diziliminin saklanması ve elde edilen sonuçların analiz edilmesi için cihazlar geliştirilmektedir. Geliştirilen cihazlarla DNA parçalarının çok daha kısa sürede dizilimleri elde edilip analiz işlemi gerçekleştirilmektedir. Bunun yanında bu cihazlarda DNA dizilimi elde edilirken rastgele yerlerde nükleotidler yanlış, eksik veya fazla okunmaktadır.

Bu çalışmada YND teknikleri kullanılan analiz cihazlarındaki hatalı okumaları gidermek için yeni bir algoritma önerilmiştir. Önerilen hata düzeltme algoritmasının işlem adımları şu şekildedir:

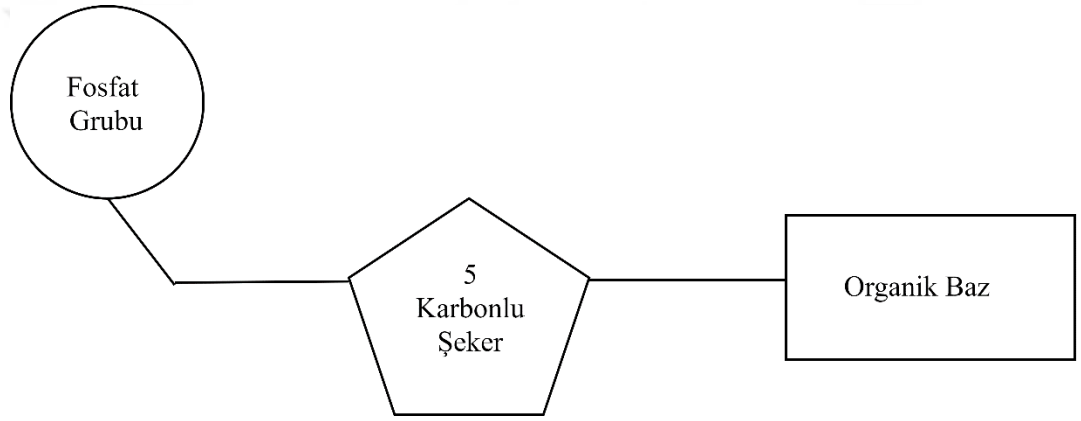
1. Analiz edilecek verinin elde edilmesi
2. En küçük benzersiz alt dizi olan k-mer'in uzunluğunun tespit edilmesi
3. DNA diziliminin sekanslara ayrılması
4. Belirlenen k-mer uzunluğunca oluşturulan alt dizinin sekaslarda aranıp, k-mer sekans grubunun oluşturulması
5. Sekans grubu üzerinde hata tespit ve düzeltme işleminin gerçekleştirilmesi

Bu çalışma kapsamında, geleneksel dizileme yöntemlerine göre hızlı çalışan fakat daha hatalı sonuçlar üreten YND cihazlarının hatalarını gidermek için yeni bir algoritma önerilmektedir.

1.3. Temel Kavramlar

1.3.1. DNA

DNA, hücrelerdeki tüm canlılık faaliyetlerinin yönetiminde önemli bir rol oynayan ve kalıtımı sağlayan moleküller topluluğu olarak tanımlanmaktadır. Ayrıca DNA'lar proteinlerin de kalıtsal bilgilerini taşıdığı için protein ve enzim sentezini yönetmektedir. Bu şekilde biyokimyasal kontrolü de sağlamaktadır. DNA nükleotid adlı moleküllerden, nükleotidler ise fosfat, şeker ve azot gruplarından oluşmaktadır (Şekil 1.1).

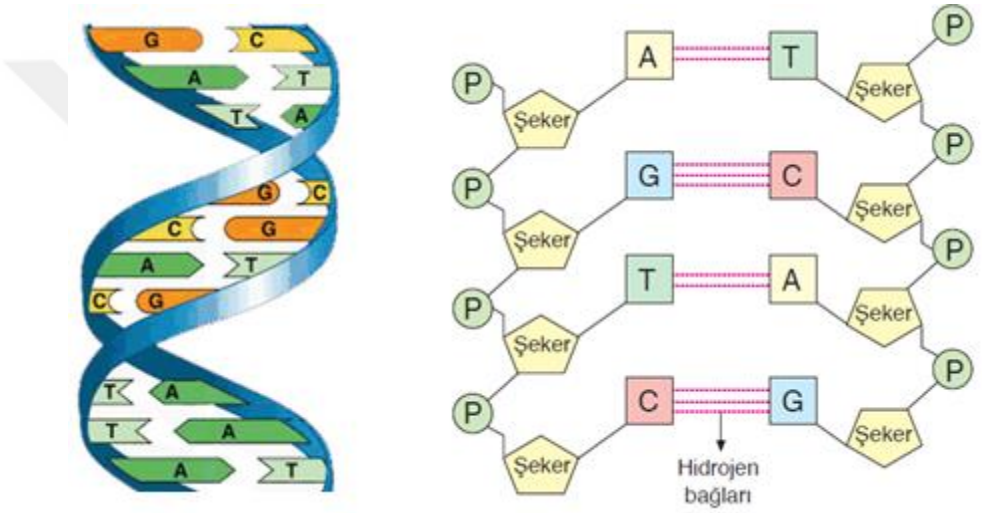


Şekil 1.1. DNA'nın moleküler yapısı

5 karbonlu şeker ile azotun glikozit bağı ile bağlanmasıyla oluşan yapıya nükleozit, ester bağıyla 5 karbonlu şekere fosfat grubunun bağlanmasıyla oluşan yapıya ise nükleotid denilmektedir. Nükleotidlerin içerisindeki Adenin (A), Timin (T), Guanin (G) ve Sitozin (C) olmak üzere 4 farklı azot bulunabilmektedir. Bu bazların dizilim şekli genetik bilgiyi belirlemektedir. Organizmanın bireysel özelliklerine ilişkin bilgileri içeren, DNA'nın anlamlı en küçük parçasına ise gen denilmektedir. Şekil 1.2'de gösterilen DNA yapısı şu şekildedir:

- DNA çift zincirin sarmal şekilde kıvrılmasıyla oluşmaktadır.
- DNA molekülleri nükleotidlerdeki bazlar arasında zayıf hidrojen bağlarının kurulmasıyla oluşmaktadır.
- Bazlar arası bağ kurulurken her zaman Adenin - Timin ve Guanin - Sitozin karşılıklı olarak eşleşmektedir.

- Adenin ve Timin arasında 2, Guanin ve Sitozin arasında 3 zayıf hidrojen bağıyla bağ kurulmaktadır.
- Bazlar arası hidrojen bağının fazla olması DNA'nın daha sağlam olmasını ifade eder. Bu nedenle DNA molekülleri içinde üçlü hidrojen bağı çok olan DNA'lar daha sağlam kabul edilmektedir.
- Moleküllerdeki toplam Adenin sayısı Timin sayısına, toplam Guanin sayısı ise Sitozin sayısına eşittir.
- Toplam pürin (A + G) sayısı pürimidin (T + S) sayısına eşittir.



Şekil 1.2. DNA molekülünün çift sarmal yapısı [13]

DNA molekülü tüm canlılarda ve bazı virüslerde bulunmaktadır. Çekirdekteki DNA molekülleri kalıtımı, organellerdeki moleküller ise bulunduğu organ hakkındaki bilgiyi saklamaktadır. DNA'nın kendini eşleyerek sahip olduğu kalıtsal bilgiyi yeni hücrelere aktarmasına DNA replikasyonu denilmektedir. DNA molekülünün sentezinde DNA polimeraz, parçalanmasında ise DNAaz enzimi etkili olmaktadır. DNA'daki toplam nükleotid sayısı (n) olduğu varsayılırsa DNA'nın nükleotidlerinin parçalanması için (n-2), tüm bileşenlerine (fosfat, şeker ve bazlar) kadar parçalanabilmesi için ise (3n-2) adet su gerekmektedir. Canlılardaki yapıyı en küçük birimden en büyüğüne kadar açıklayacak olursak sıralama; Organik baz, nükleotid, gen, DNA, kromozom, çekirdek, hücre, doku, organ, sistem, organizma şeklinde olmaktadır.

1.3.2. RNA

RNA bir polimer olup nükleotidlerden oluşmaktadır. Her nükleotid azotlu baz, fosfat ve riboz şekerinden oluşmaktadır. RNA'nın canlı vücudunda pek çok görevi bulunmaktadır. RNA (Ribo Nükleik Asit) özellikleri şu şekildedir [14]:

- DNA protein sentezinde doğrudan görev yapmak yerine ilgili bilgileri RNA'ya aktararak RNA'nın protein sentezini gerçekleştirmesini sağlamaktadır.
- RNA çekirdek, sitoplazma ve bazı organellerde bulunmaktadır.
- RNA tek nükleotid zincirinden oluştuğu için kendini eşleyip çoğalamamaktadır.
- RNA'nın DNA'dan yapısal olarak farkı; deoksiriboz şekeri yerine riboz şekeri ve Timin bazı yerine Urasil bazı taşımasıdır.

1.4. DNA Dizileme

DNA parçasındaki bazların dizilimini tespit edip hizalamaya DNA dizileme denilmektedir. DNA analizi ilk olarak, geleneksel yöntemler olarak adlandırılan birinci nesil dizileme yöntemleriyle analiz edilmiş, genetik yapı ve kontrol mekanizmaları hakkında birçok bilgi edinmemizi sağlamıştır. 2005 yılında DNA parçalarının kopyalanarak bu kopyalar üzerinde tekrarlı işlemlerle yüksek doğruluklu dizileme çalışmaları yapılmaya başlanmıştır. Bu yöntemlerde ikinci nesil DNA dizileme olarak literatürde yer almaktadır [7,15].

Birinci ve ikinci nesil dizilemelerde yürütülen çalışmalarla çok önemli sonuçlar elde edilmiş olup daha az maliyetle, daha kısa sürede ve daha doğru DNA dizilimi elde etme çalışmaları sürdürülmüş olup bu amaca uygun yeni teknikler geliştirilmiştir. Bu teknikler de üçüncü nesil DNA dizileme teknikleri başlığı altında toplanmaktadır. Bu tekniklerde ikinci nesil dizileme yöntemlerinden farklı olarak, DNA parçası katı bir yüzeye tutturularak veya bir kanal boyunca hareketi sağlanarak sentezleme aşamasında kullanılan teknolojilerle DNA dizilimi okunmaktadır. Bu işlemlerle DNA parçasının kopyaları üzerinde tekrarlı yapılarak, önceki tekniklerde karşılaşılan sorunlar giderilmiştir [7,15]. Geçmişten günümüze yaygın kullanılan DNA dizileme yöntemleri şu şekildedir:

- Zincir Sonlandırma Yöntemi
- Maxam-Gilbert Dizilemesi

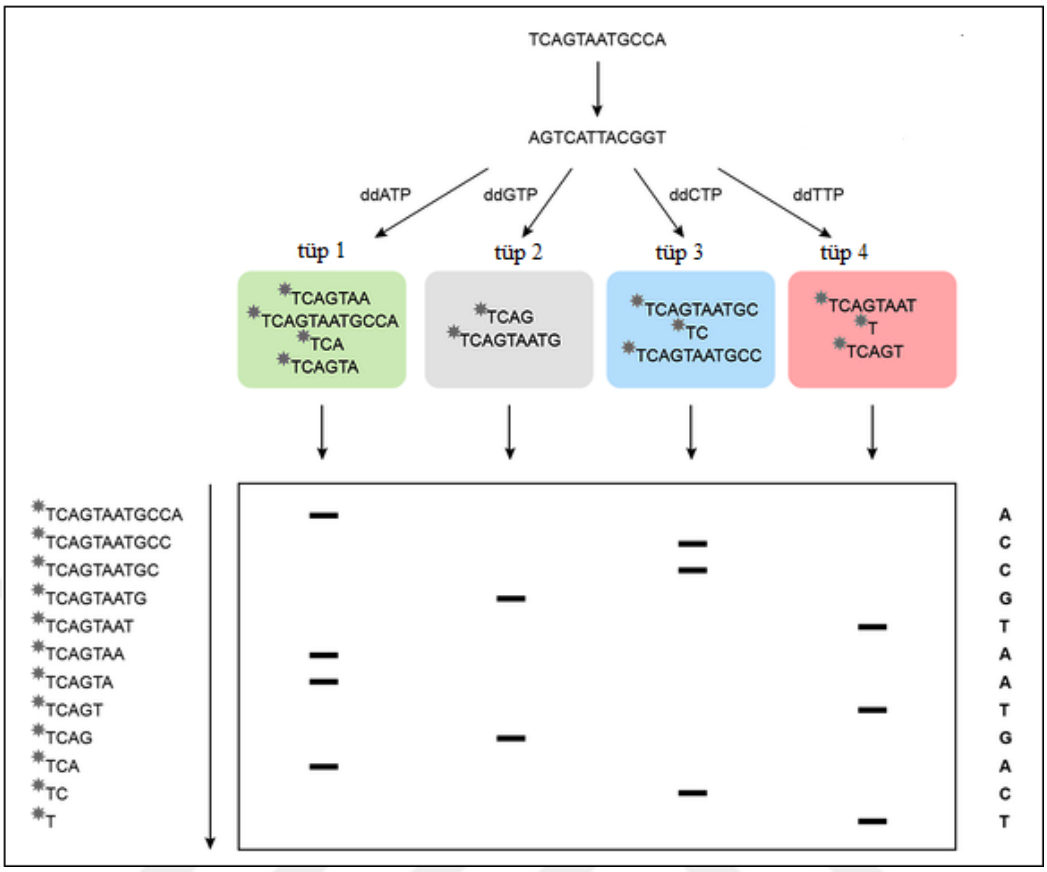
- Av Tüfeği Dizilemesi
- PCR Amplifikasyon
- Kitlesel Paralel Dizileme
- Poloni Dizilemesi
- Solexa Dizilemesi
- 454 Piro Dizilemesi
- Solid Dizilemesi
- DNA Nanotop Dizilemesi
- İyon Yarı İletken Dizilemesi

1.4.1. Zincir Sonlandırma Yöntemi

1975 yılında Frederick Sanger ve Alan Coulson tarafından geliştirilmiştir [17-19]. Sanger dizileme yöntemi uzun DNA parçalarını dizilemek için kullanılmaktadır. İlk işlemde analizi yapılamayacak kadar büyük DNA parçası daha küçük parçalara ayrılmaktadır. Sonrasında bu küçük parçalar plazmite kopyalanmaktadır. Kopyalanan plazmitler ayrı ayrı dizilenmekte ve en sonunda belirlenen analizlerle plazmitler birleştirilerek ilk aşamadaki büyük DNA parçası elde edilmektedir. Bu yöntemde gerçekleşen işlemler genel olarak aşağıda maddeler halinde ifade edilmiştir:

- DNA'nın hazırlanması
- Reaksiyonların gerçekleşmesi
- Yüksek voltajlı jel elektroforezi

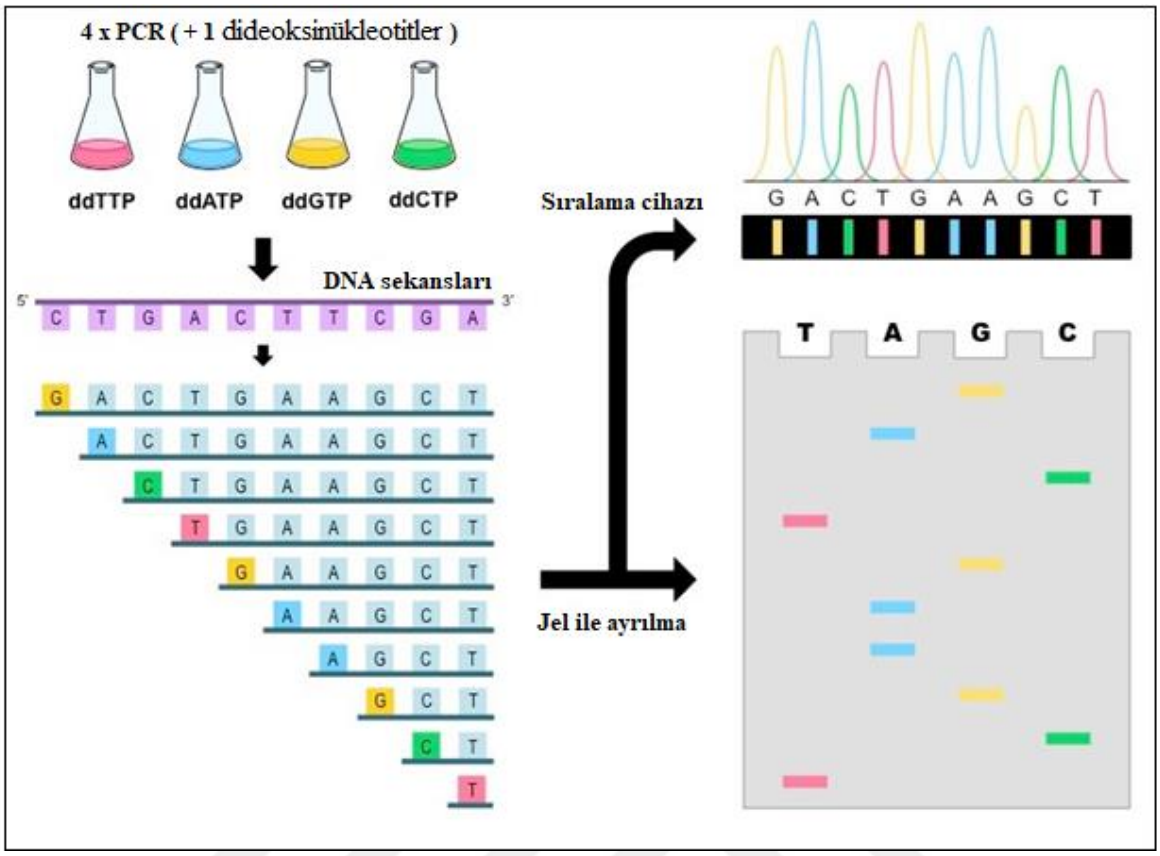
Sanger dizileme yönteminde kalıp DNA, klonlama işlemi için gerekli DNA polimeraz, tüm baz türleri için deoksinükleotitler (dNTP, dATP, dTTP, dGTP, dCTP) ve zincir büyümesini sonlandıracak olan dideoksinükleotitler (ddNTP, ddATP, ddTTP, ddGTP, ddCTP) ayrı ayrı 4 farklı reaksiyon ortamı için hazırlanmaktadır. Klonlama sırasında her nükleotid bir öncekinin 3'OH ucuna eklenerek bağ kurulmaktadır. Dideoksinükleotitlerde 3'OH grubu bulunmadığından gelen nükleotidle bağ kurulamaz ve böylece zincir sonlanmış olur. Bu klonlama işlemi tüm DNA için uygulanarak farklı uzunluklarda sonlu parçalar elde edilmektedir. Şekil 1.3'te görüldüğü gibi elde edilen sonlu parçalar da büyüklüklerine göre sıralanmaktadır [20].



Şekil 1.3. Sanger dizileme yönteminde sonlanan parçaların sıralanma işlemi [20]

Sıralanan nükleotid parçaları sonlanan nükleotid tiplerine göre dört gruba ayrılmaktadır. Örneğin; Adenin ile biten parçalar ddATP grubuna, Guanin ile biten parçalar ddGTP gruplarına ayrılmaktadır. Ayrıca her ddNTP grubu floresan işaretleyici içermektedir. Bu durum her nükleotidin farklı bir renkle gösterildiğini ifade etmektedir. Dizileme işlemi için her bir grup kendi dNTP reaksiyon karışım tüplerine eklenmektedir [21].

Şekil 1.4'de dört reaksiyon tüpünden alınan karışımlar poliakrilamid jelin üzerine dökülmekte ve oluşan sıra ile okuma işlemine başlanmaktadır. Okuma işlemi makine içindeki lazer yardımıyla floresan renklerinin yoğunluğunun algılanmasıyla gerçekleştirilmektedir [21]. Okuma işlemi sırasında eşit yoğunlukta iki renk algılanması veya hiçbir renk algılanmaması gibi durumlar oluşabilmektedir. Bu gibi durumlar belirtilen bölgede silinme, değişim ve eklenme gibi hataların veya mutasyonların oluştuğu anlamına gelmektedir. Bu hataların giderilip tam doğru dizilimlerin elde edilmesi ve mutasyonların tam doğrulukla tespit edilmesi ayrıca ilgilenilmesi gereken bir problem olarak günümüze kadar gelmiştir [22].



Şekil 1.4. Sanger dizileme yönteminde nükleotid sıralama işlemi [21]

1.4.2. Maxam-Gilbert Dizilemesi

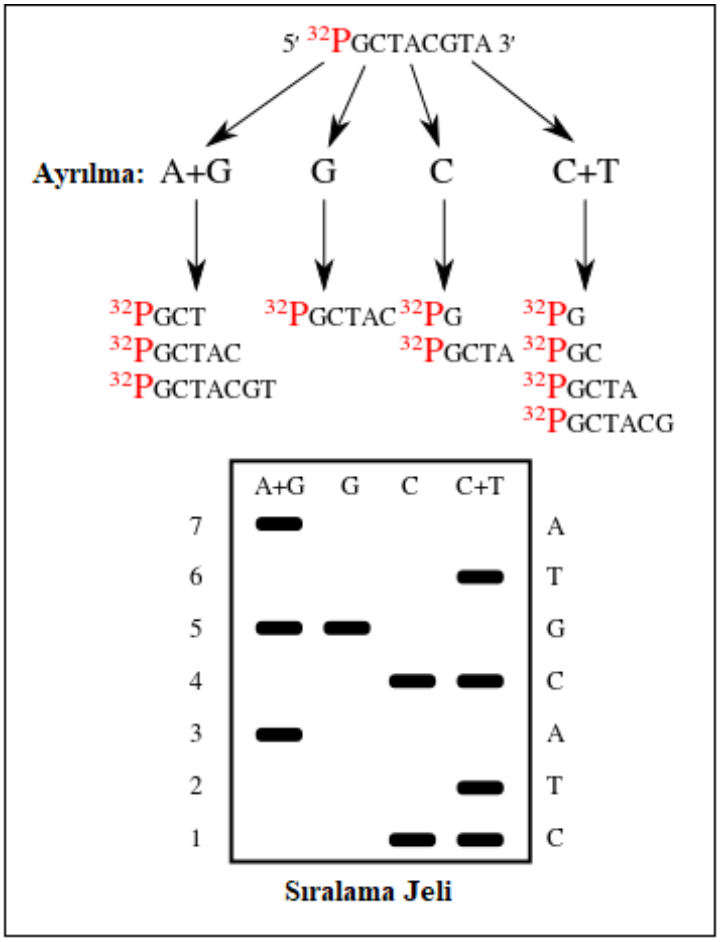
DNA parçalarının kimyasallar kullanılarak oluşturulması yöntemine Maxam-Gilbert dizilemesi denilmektedir. İlk aşamada analiz edilecek DNA parçası 5'OH ucundan radyoaktif p32 ile etiketlenerek 4 eşit parçaya ayrılmaktadır. Bu parçalar 4 ayrı tüpe konularak, zinciri farklı bazlardan ayıracak olan 4 farklı kimyasal tüpe eklenmektedir. Bu tüplerdeki kimyasal kırılma üç aşamayla gerçekleşir [23, 24]:

- Baz modifikasyonu
- Değiştirilmiş bazın şekerden ayrılması
- Şekerden DNA ipliğinin kırılması

Şekil 1.5'te görüldüğü gibi bazı bazlar karışımda birleşik durumda bulunmaktadır. Bazları birbirinden ayırmak için çeşitli kimyasal işlemler uygulanmaktadır. Hidrazin ile bazların pirimidinlerden, asit ve dimetil sülfat ile de bazların pürinlerden ayrılması sağlanmaktadır [25]. Bu işlemlerden sonra her tüpte faklı noktalardan kırılmış farklı uzunlukta DNA parçaları elde edilmektedir. Elde edilen parçalar yüksek çözünürlüklü bir

poliakrilamid jel üzerinde elektroforez uygulanarak görselleştirilmekte ve diziler oluşturulmaktadır [25].

Sanger zincir sonlandırma yöntemi, Maxam-Gilbert dizileme yöntemi ile karşılaştırıldığı zaman daha kolay yorumlanan ham veriler üretmekte ve daha yaygın kullanıldığı tespit edilmektedir [27].

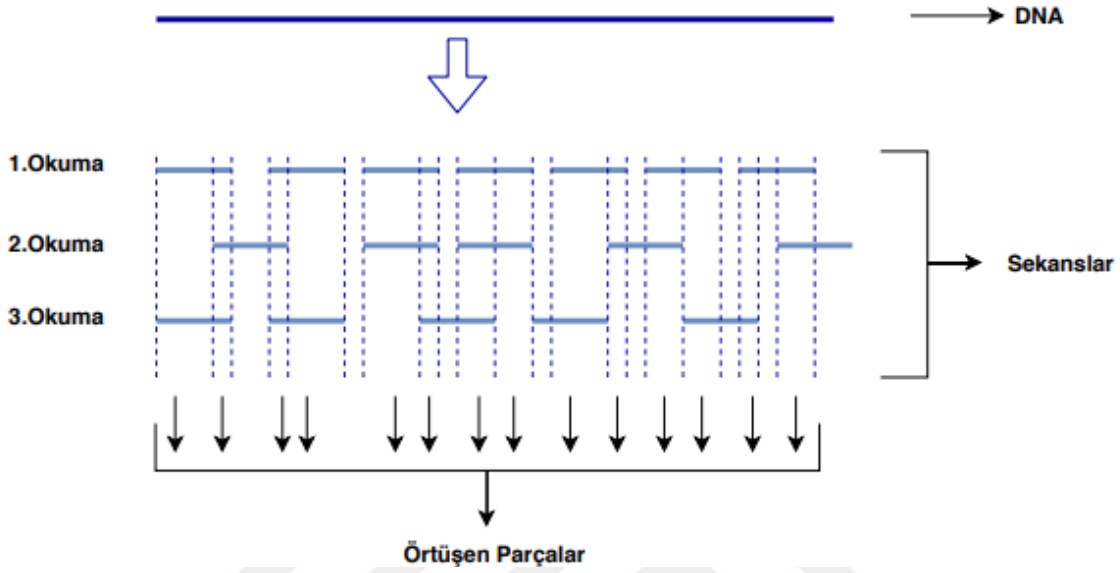


Şekil 1.5. Maxam-Gilbert dizileme yönteminde bazların ayrıştırılması [27]

1.4.3. Av Tüfeği Dizilemesi

Uzun DNA parçalarını dizilemek için kullanılan, büyük DNA parçasını rastlantısal olarak küçük parçalara-sekanslara ayırarak dizilime yapan yönteme Av Tüfeği (Shotgun) dizileme denilmektedir [28]. Bu yöntemde küçük parçalara ayırarak daha hızlı ve doğruluk oranı yüksek dizilime yapılması amaçlanmaktadır. Ayrılan küçük parçalarda üst üste gelen-örtüşen kısımlarına bakılıp birleştirme işlemi yapılarak büyük dizilim elde edilmektedir

(Şekil 1.6). Örtüşen parçaları bulmak için bir DNA parçası belirlenen sayıda okunup, örtüşen parça sayısı artırılmaktadır. Örtüşen parça sayısı ile okunan DNA parçasının doğruluk oranı doğru orantılıdır [27].



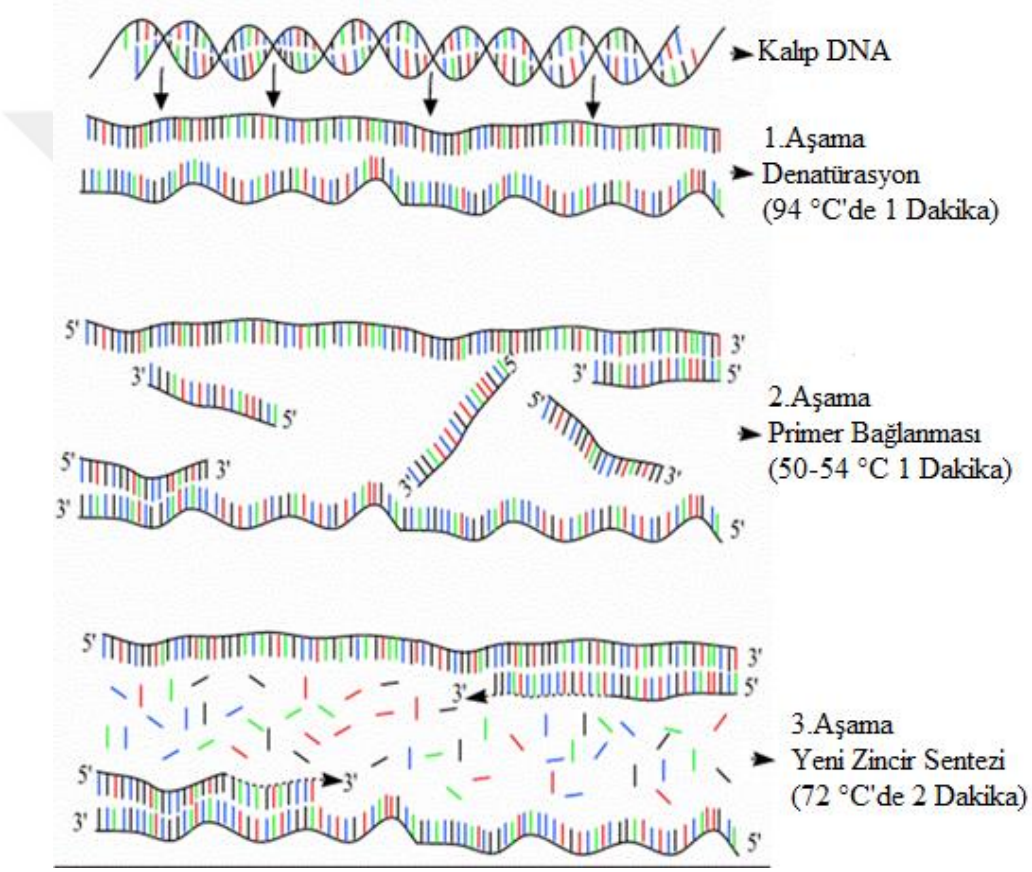
Şekil 1.6. Av Tüfeği dizileme yönteminde örtüşen parçalar [27]

Bu yöntem büyük verileri depolama ve işleme zorluğunu yanında getirmektedir. Örneğin 3 milyar uzunluklu DNA parçası 10 kez okunup her okumada 50'lik parçalara ayrılırsa 600 milyon parça elde edilmektedir. Bu parçalar arasında kesişmeyi bulmak ve analiz etmek ilkel yöntemlerle zor olmaktadır.

Av tüfeği dizileme yöntemi genellikle büyük ölçekli dizileme projelerinde tercih edilmektedir. DNA parçalarındaki örtüşen parça sayısını artırmak için tekrarlanan okuma işlemiyle sağlanan yedekleme, güvenilir sekans kalitesi sağlamaktadır [27]. Bu yöntemin uygulanması için performansı yüksek bilgisayarlar ve çeşitli programlar gerekmektedir. Av tüfeği fazından sonra kalan boşlukları doldurmak için primer yürüme yöntemi ile sıralama genellikle alt klonlar veya boşlukları dolduran PCR ürünleri üzerinde gerçekleştirilmektedir [29].

1.4.4. PCR Amplifikasyon

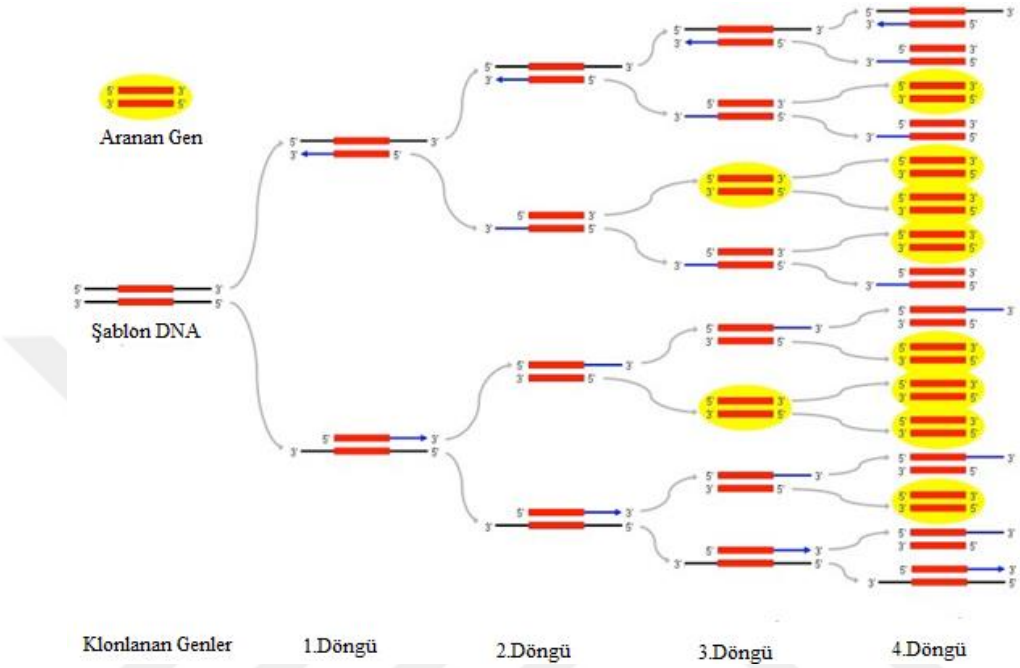
PCR Amplifikasyon yöntemi, 1983 yılında ABD'deki Cetus şirketi çalışanlarından Kary Mullis ve arkadaşları tarafında geliştirilmiştir [30]. Bu yöntem genel olarak; DNA içerisinde incelenecek alanın çeşitli enzimatik tepkimelerle çoğaltılıp, genomun tamamı bilinmese de analiz için gerekli olan DNA parçalarının yeterli miktarda elde edilmesi olarak açıklanmaktadır [31].



Şekil 1.7. PCR Amplifikasyon yöntemi aşamaları [33]

PCR amplifikasyonu, çoğaltılacak DNA parçasını kuşatan iki DNA oligonükleotid primerlerini ve DNA'nın ısı denatürasyon döngülerini, primerlerin tamamlayıcı sekanslara tavlanmasını ve tavlanmış sekansların DNA polimeraz ile çoğaltılmasını ifade etmektedir [32]. Primerler hedef dizinin karşı iplikleriyle hibritleşerek yönlendirilmektedir. Bu işlemle polimerler arası polimeraz enzimi ilerleyerek DNA sentezini gerçekleştirerek DNA parçası miktarını ikiye katlamaktadır [32]. PCR amplifikasyon yönteminin aşamaları Şekil 1.7'de

gösterilmektedir [33]. Her oluşan yeni kopya ardışık döngüde diğer enzimatik işlem için kalıp görevi görmektedir. Bu şekilde her kopya sonunda bir önceki DNA parçasının iki katı elde edilmektedir.



Şekil 1.8. PCR Amplifikasyon yöntemi döngü aşamaları [34]

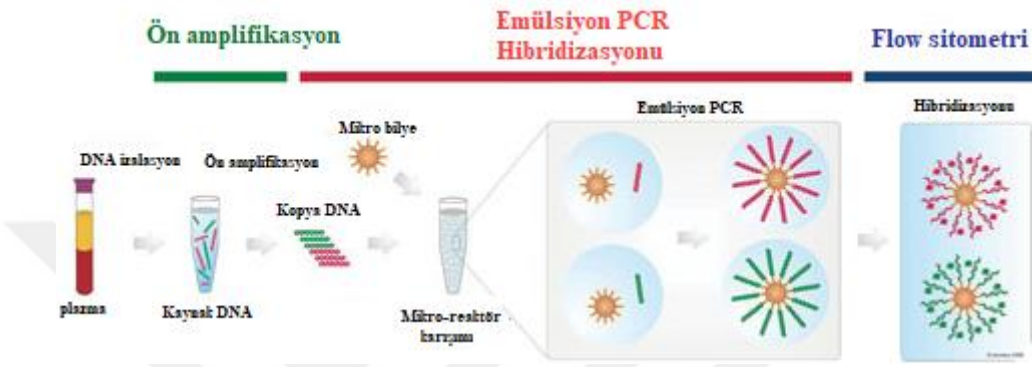
Bir DNA parçası ile başlanan PCR amplifikasyonda tüm döngüler %100 verimlikte çalışırsa 10 döngü sonunda 210 adet kopya elde edilmektedir. Bu yöntemdeki döngüler ve çoğalma miktarı Şekil 1.8'de gösterilmektedir [34].

1.4.4.1. Emülsiyon PCR

Emülsiyon PCR yönteminde emülsiyon yağı, boncuklar, PCR karışımı ve kalıp DNA'lar fragmentler oluşturmak için karıştırılmaktadır. Amplifikasyon işleminin başarılı bir şekilde sonuçlanması için oligonükleotitlerin karşılığı olan her bir fragmentin bilyelere bağlanması gerekmektedir. Bunun için de karışımın homojen olarak karıştırılması önem taşımaktadır [35].

DNA fragmentleri, mikro bilyelere bağlanarak bilyeler üzerinde çoğaltılmaktadır. Bu bilyeler ise mikro-reaktör olarak adlandırılan sıvı küre tarafından çevrilmektedir. Mikro-reaktör karışımı içerisinde kaynak DNA, primerler, boncuk ve PCR karışımı bulunmaktadır.

Mikro-reaktör içerisinde, DNA tek zincirli yapıya dönüştürülmekte ve polimeraz enzimi boncuklara bağlanan DNA fragmentlerinin kopyalarını oluşturmaktadır. Her döngü sonunda oluşan kopya fragment de bir sonraki döngü için kaynak olarak kullanılmaktadır. Bu işlem 30 ve 60 arası döngüden oluşacak şekilde devam etmektedir. İşlem sonunda da dizileme işlemi için mikro-reaktör patlatılarak uygun ortama aktarılmaktadır (Şekil 1.9).

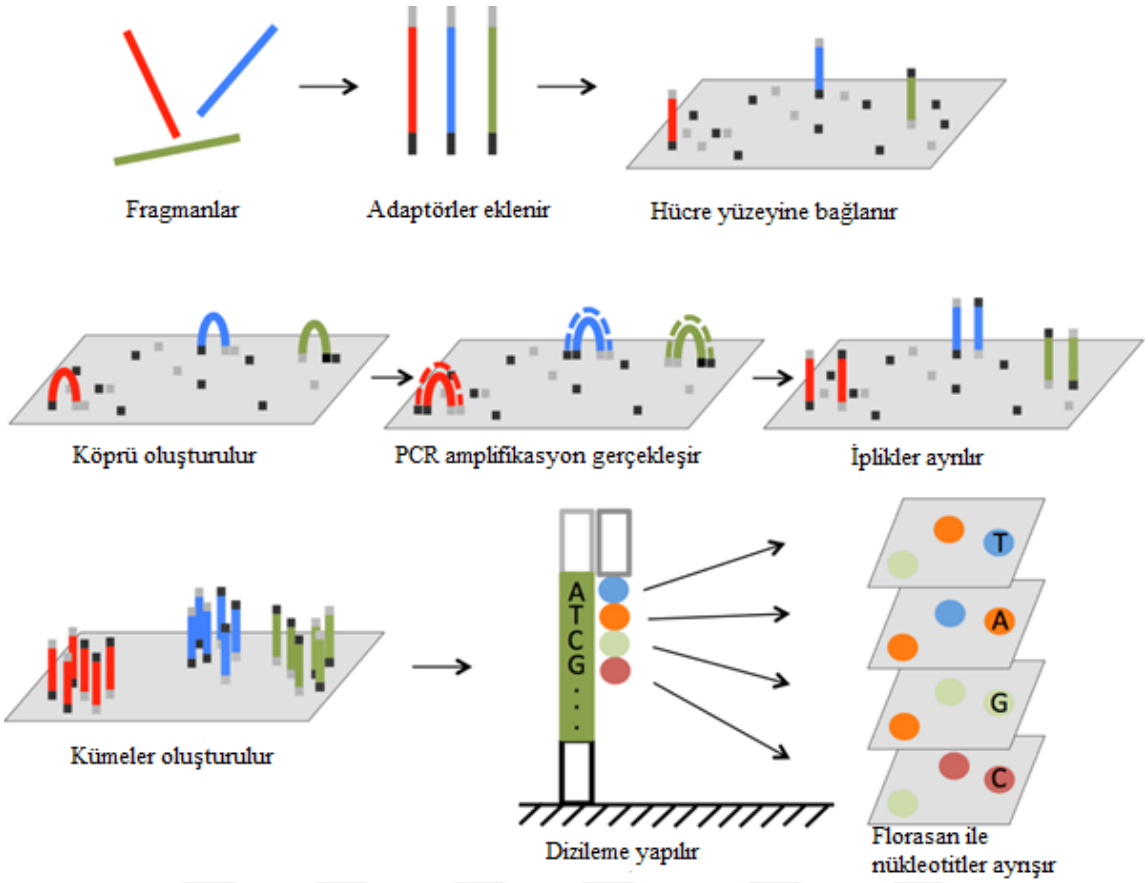


Şekil 1.9. Emülsiyon PCR yöntemi

Emülsiyon PCR yönteminde her zaman bir DNA fragmenti ve bir bilye eşleşmesi gerçekleşmeyebilmektedir. Çoğu zaman bir boncuk birden fazla fragmentle birleşmekte ve bu durumda da hatalı DNA dizilimi elde edilerek işlem sonuçlanmaktadır. Yapılan çalışmalar mikro-reaktörelerin yalnızca %15'inin başarılı eşleşme yaptığını göstermektedir. Bu nedenle, oluşan hatalı eşleşmeyi engellemek içinde ekstra işlem gerekmektedir [36].

1.4.4.2. Köprü PCR

Köprü PCR yönteminde analiz edilecek DNA parçalara ayrılmakta ve parçaların her iki ucuna primerler bağlanmaktadır. Daha sonra parçalar denatüre edilerek tek iplikçikli hale dönüştürülmektedir. Tek iplikçikli yeni DNA parçaları reaktiflere maruz kalarak bulundukları yerde rastgele olarak hücre yüzeyine bağlanmaktadır. Nükleotidler ve DNA polimeraz enzimlerin katkılarıyla da tek iplikçikli DNA parçalarının serbest uçları, primerler sayesinde hücre yüzeyine bağlanarak köprülü yapı oluşturmaktadır. Oluşan köprülü yapı içerisinde gerçekleşen amplifiye işlemiyle başlangıçtaki tek iplikçikli DNA parçasının ters dizilimine sahip DNA kopyası oluşmaktadır.



Şekil 1.10. Köprü PCR Amplifikasyonu [36]

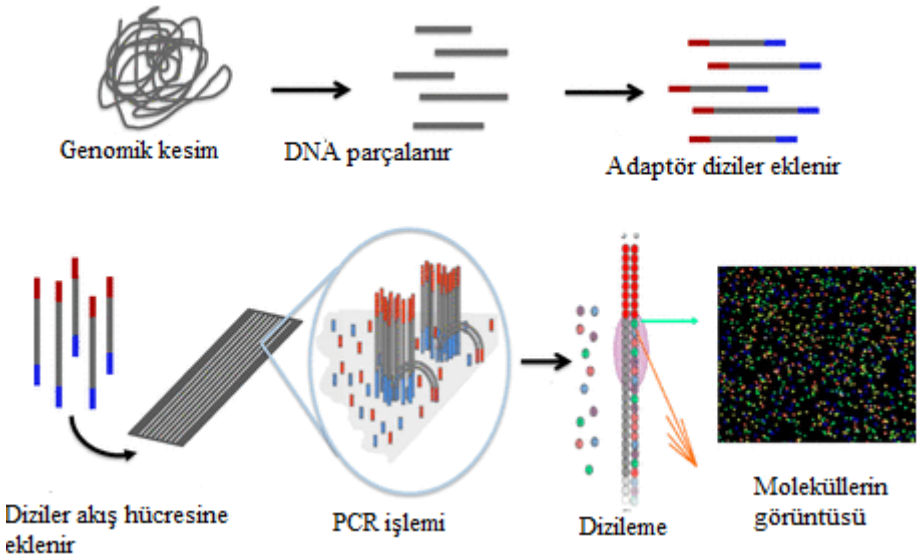
Köprü içerisinde gerçekleşen amplifikasyon işleminde, florasan etkili dNTP'ler kullanılmaktadır. Örneğin mavi florasan ışığı sönerse okunan nükleotidin Adenin, kırmızı ışık sönerse de okunan nükleotidimin Guanin olduğuna karar verilmektedir. Belirtilen ışık renkleri cihazlara göre farklılık gösterebilmektedir. Okuma işlemi tüm iplikçik için sonlana kadar dNTP gönderimi devam etmektedir [35-38]. Köprü PCR amplifikasyon işlemi adım adım Şekil 1.10'da gösterilmektedir.

1.4.5. Kitlesel Paralel Dizileme

Yeni nesil dizileme yöntemlerinden biri olan kitlesel paralel dizileme yöntemi DNA parçasının dizilimi için 1 milyon ile 43 milyar arası kısa parçalarda okuma yapmaktadır. Bu yöntemde paralel DNA dizilime tekniği kullanıldığından dolayı yüksek verimli dizilime görülmektedir. [39].

Kitlesel paralel dizileme, DNA'nın parçalanması ve bu parçalara adaptör dizilerinin eklenmesiyle başlatılmaktadır. Adaptörler, polimeraz zincir reaksiyonu kullanarak DNA'nın çoğaltılması için gerekli tüm bileşenleri içeren, yağ damlacıkları ile sarılı yaklaşık 28 mm çapında boncuklara bağlanmasını sağlamaktadır. Oluşan parçalar bu yöntemde kordon olarak adlandırılmaktadır. Yağ damlacıkları, emülsiyonu oluşturduğu için oluşan kordonlar birbirlerinden ayrı tutularak amplifikasyon karışıklığı önlenmektedir. Kopyalama işlemi sonucunda analiz edilecek DNA fragmanın yaklaşık 10 milyon kopyası oluşmaktadır [40, 41].

Sekanslama için DNA kopyası taşıyan tanecikler pikolitre reaktör kuyucuklarına yüklenmektedir. Bu yöntemde kopya DNA dizilerini tamamlayacak DNA dizisi, sentezleme yöntemleri kullanılarak oluşturulmaktadır. DNA parçası ile tamamlayıcı parça arasında oluşan bağ kimyasallar salgılamaktadır. Bu kimyasallar enzimler ve lusiferin varlığında ışık saçan reaksiyon göstermektedir. Plaka üzerinde DNA diziliminde bulunan dört nükleotid kanal boyunca akmaktadır. Daha sonra eşzamanlı olarak kanalda milyonlarca küme, bir detektör yardımıyla saçılan ışınlara bakılarak DNA dizisi okunmaktadır [40]. Okuma işleminde nükleotidlerin onarımını sağlamak için kanal boyunca tersine çevrilmektedir. Bu tüm süreç bir döngüyü oluşturmaktadır [41,42]. Döngü adımlar halinde Şekil 1.11'de gösterilmektedir [42].

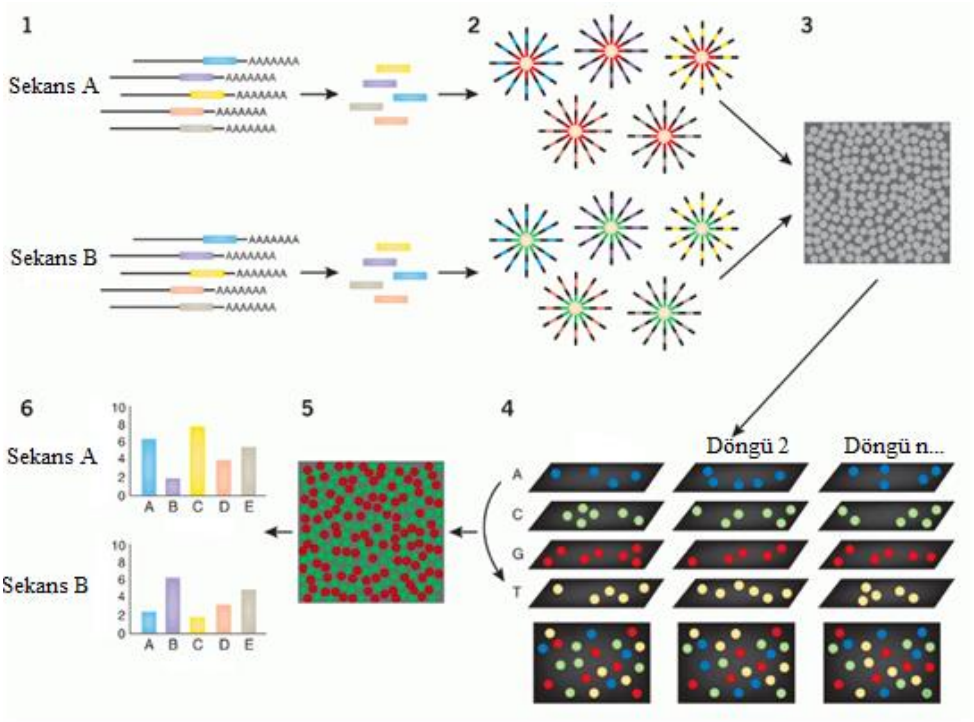


Şekil 1.11. Kitlesel paralel dizileme yönteminde döngü adımları [42]

1.4.6. Poloni Dizilemesi

Poloni dizileme yönteminde bir genomun yeniden kullanılması metodu kullanılmaktadır [43]. İlk işlem olarak özel kısıtlama kullanılarak A ve B örneklerinin transkriptlerinden izole edilmektedir. Daha sonra dizilenecek DNA küçük parçalara ayrılmaktadır. Parçalar emülsiyon PCR ile çoğaltılmaktadır. Bu genomik parçalar bir bağlayıcılar ile daire şekline getirilmektedir. Ortaya çıkan her bir polik boncuk, tek bir parçanın klonal kopyalarını içermekte ve farklı evrensel diziler tarafından kuşatılmış iki 17-18 baz çiftli genomik dizileri temsil etmektedir.

1 mm polik boncuklar üzerinde klonal şablon büyüme sağlamaktadır. Manyetik olmayan düşük yoğunluklu boncuklarına (koyu mavi) hibridizasyon, santrifüj yoluyla manyetik ePCR boncuklarının büyütülmüş fraksiyonunun (kırmızı) zenginleşmesine izin vermektedir. Polik boncuklar, otomatik dizileme için jelsiz katı cam destek üzerinde birleştirilmektedir. Her bir sıralama döngüsünde, dört renkli görüntülenme yapılarak parçaların okunmasıyla belirli bir konumdaki her büyütülmüş boncuğun içerdiği diziyi belirlemek için yüzey boyunca tarama işlemi tekrarlı olarak gerçekleştirilmektedir [44,45]. Şekil 1.12'de bu yöntemin çalışma adımları gösterilmektedir [46].



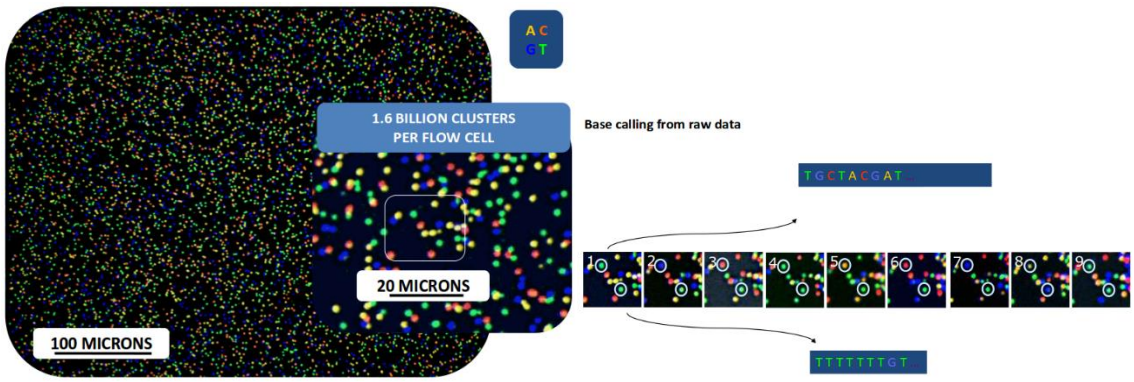
Şekil 1.12. Poloni dizileme yönteminin çalışma adımları [46]

1.4.7. Solexa Dizilemesi

Solexa dizilime zincir sonlandırma yöntemine benzemektedir. Bu yöntemde zincir sonu olarak DNA polimeraz yardımıyla eklenen terminatör nükleotidler kullanılmaktadır. Sonlandırıcı nükleotidler, floresan algılama ve etkisizleştirme sonrası polimerizasyonun devam etmesi için okuma işlemini tersine çevirebilmektedir. DNA dizisi, dört floresan etiketli tersine zincir sonlandırıcı ve DNA polimerazın karışımına eklenerek sentezleme işlemi ile elde edilmektedir [16,47].

Solexa dizilemesinde ilk olarak incelenecek olan DNA küçük parçalara ayrılmaktadır. Bu DNA parçaları katı bir yüzeye kovalent bağ ile sabitlenerek katı fazda amplifikasyon gerçekleştirilmektedir. Her bir DNA parçasına birer geri dönüşümlü terminatör nükleotid eklenmektedir. Her parça için floresan sinyal tespit edilip florofor ve geri dönüşümlü blok kaldırılmaktadır. Terminatör-enzim karışımı daha sonra bir sonraki döngüyü başlatmak için eklenerek işlem çalışmanın sonuna kadar tekrar edilmektedir [16,47].

Her DNA parçasına eklenen floresanlarla birlikte akıştaki DNA parçaları Şekil 1.13'de gösterilmektedir [48].



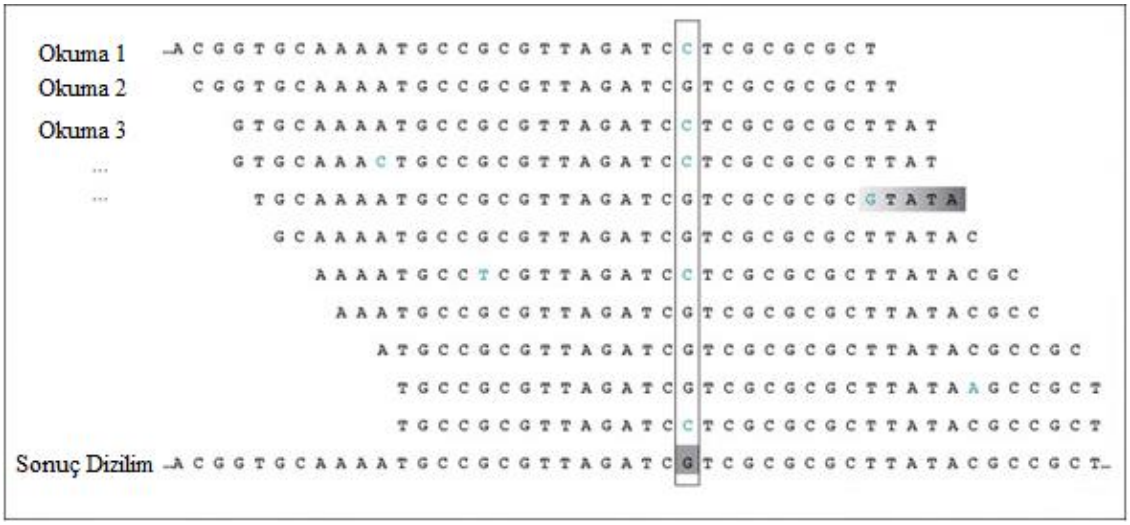
Şekil 1.13. Solexa dizileme yönteminde DNA parçaları [48]

Dört nükleotid için de reaksiyon mevcut olduğu için kavrama doğruluğu artmaktadır. Bu sistemler ayrıca bakteriyel klonlama aşamalarından kaçınarak potansiyel olarak bakteriler içinde çoğaltılamayan DNA'dan sekans verileri üretebilmektedir [16, 47].

Solexa yazılımı alanında uzman araştırmacıların işbirliği içinde çalışması ile geliştirilmiştir. Yazılım geliştirilirken analiz aşamasının kolay ve en doğru şekilde gerçekleşmesi için, eksiksiz ve az kullanıcı müdahalesi ile gerçekleşen veri toplanması,

işlenmesi ve analiz edilmesi modüllerini kapsamaktadır. Dizileme işlemi tamamlandıktan sonra veri dosyaları, baz çağrılarının üretilmesi, taban tabana güven puanlarının oylanması ve referans veri tabanına yeniden hizalanması için görüntü analizi temelli çalışan bir bilgisayara aktarılmaktadır. Bu bilgisayarda operatör verileri, özet istatistikleri ve hata analizleri görüntülenebileceği gibi ayrıntılı çalışma ve karşılaştırmalar yapılabilmektedir [47].

Şekil 1.14'de her bir okuma işleminde elde edilen dizilemeler karşılaştırılmakta ve aynı indislerdeki farklılıklar incelenerek oluşan hatalar giderilmeye çalışılmaktadır [49].



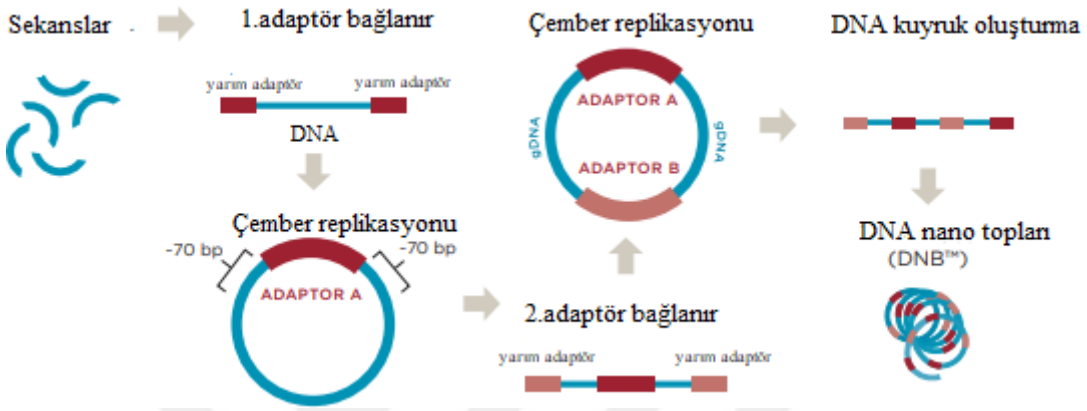
Şekil 1.14. Solexa dizileme yönteminde karşılaştırma işlemi [49]

1.4.8. DNA Nanotop Dizilemesi

DNA Nanotop dizileme yöntemi bir organizmanın tüm genom diziliminin elde edilmesi için kullanılan yüksek verimliliğe sahip teknoloji olarak tanımlanmaktadır. Bu yöntemde dizilenecek DNA, 100-350 baz çiftinden oluşacak şekilde fiziksel veya enzimatik reaksiyonlarla birlikte parçalara ayrılmaktadır. Parçaların her birinin uçlarına sıralaması bilinen adaptör DNA sekansları eklenmekte ve eklenen sekanslar daire oluşturacak şekilde uçlarından birleştirilmektedir. Bu işlemler tüm parçalar için gerçekleştirilerek dairesel DNA şablonları oluşturulmaktadır [50, 51].

Başlangıçta küçük DNA dizilimine sahip olan iplik, eklenen adaptör dizilimlerle uzun DNA dizilimli ipliğe yükseltilmektedir. Yeni sentezlenen iplik, yuvarlama çember replikasyonu kopyalama işlemine tabi tutularak çoğaltılmaktadır. Birbirinin kopyası olan

iplikler, aralarındaki uzaklık yaklaşık olarak 300 nanometre olacak şekilde birleştirilerek DNA topu haline getirilmektedir. Her iplik grubu kendi DNA topunu oluşturur ve oluşan toplarda birbirinden ayrılarak karışıklık önlenmektedir [50 - 53]. Şekil 1.15'de DNA Nanotop dizilemesi gösterilmektedir [54].

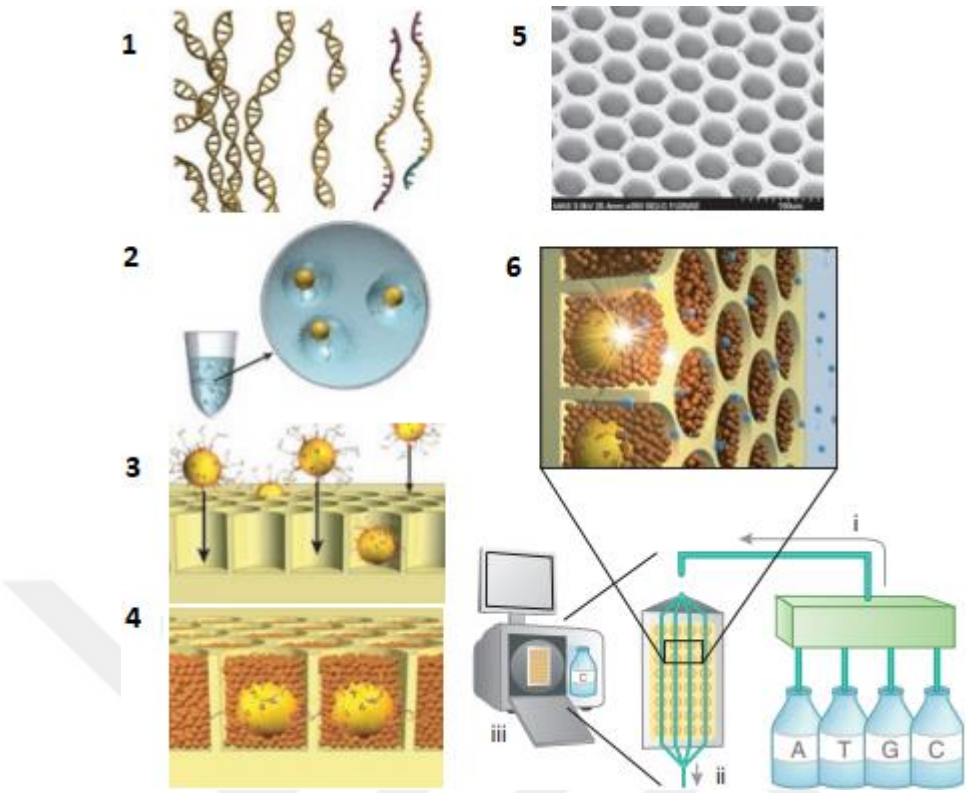


Şekil 1.15 DNA Nanotop dizileme yöntemi [54]

DNA dizilimi elde etmek için akış hücresine DNA nanotopları bağlanmaktadır. Adaptör dizide bulunan floresan tutulan lazer yardımıyla ışığın dalga boyunu uyarmaktadır. Nanotoplardan gelen floresan ışını yüksek çözünürlüklü bir kamera ile görüntülenerek analiz edilmektedir. Bu şekilde her bir nantoptan gelen renge göre dizileme işlemi gerçekleştirilmektedir. [50, 52, 53, 55].

1.4.9. 454 Prodizilemesi

DNA dizileme teknolojisinde geleneksel olarak kullanılan yöntem, DNA'nın parçalara ayrılarak tamamlayıcı bir iplikçiğin sentezi ile floresan boya veya radyoaktif izotop ile işaretlenmiş dideoksi-nükleotidleri içeren karışımla reaksiyonuna dayanmaktadır [56]. Sanger teknolojisinden günümüze kadar DNA diziliminde maliyeti azaltmak ve okuma doğruluğunu artırmak için çalışmalar sürmektedir. Bu hedef doğrultusunda yeni teknoloji olarak "prokidrasyon" prensibine dayalı teknik geliştirilmiştir. Prodizileme yöntemi DNA sentezi sırasında salınan profosfatın saptanmasını temel alarak çalışmaktadır [57].

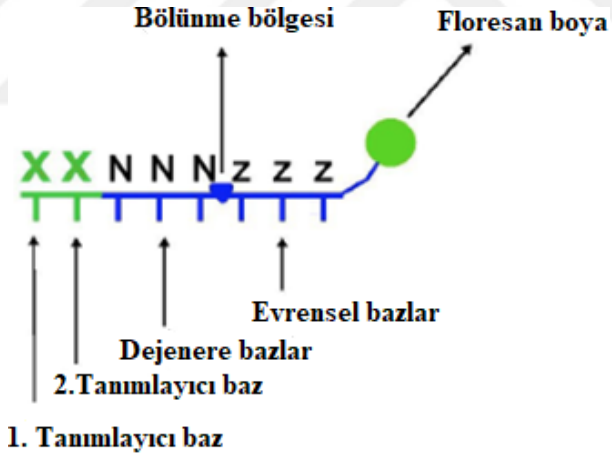


Şekil 1.16. 454 Prodizileme yöntemi [59]

454 prodizileme yönteminde incelenecek DNA parçaları, yüz ve katları olacak şekilde boyutlara ayrılarak tamamlayıcı DNA dizilimlerine eklenmektedir. Oluşan şablonda, DNA parçalarının her birine mikro boncuklar eklenmektedir. DNA parçaları doğrudan bu mikro boncuklara bağlanmakta ve bağlanma esnasında çift sarmal yapı kırılarak DNA dizisi tek iplikçili halini almaktadır. Mikro boncuklar, şablon olarak kullanılan DNA parçalarının PCR ile amplifikasyonu için emülsiyon damlacıklarına bağlanarak boncuklar üzerindeki DNA parçalarının kopyalanmasına neden olmaktadır. Oluşan mikro parçalar dört nükleotidin sırayla ekleneceği kuyucuklardan oluşan fiber optik slayda eklenmektedir. Bu ortamda mikro parçalara eklenen nükleotid çeşidine göre ışık yayımı gerçekleşmektedir. Yayılan ışığın sinyal yoğunluğu, tek adımda eklenen nükleotid sayısını temsil etmektedir [58,59]. Örneğin işlem yapılan dizide art arda 4 adet Timin bulunuyorsa yayılan ışın 1 Timin bazı içeren parçanın 4 katı olmaktadır. Tüm dizilim için ışık yoğunluklarına bakılarak DNA dizilimi elde edilmektedir. 454 prodizileme teknolojinin adımlar halinde anlatımı Şekil 1.16'da gösterilmektedir [59].

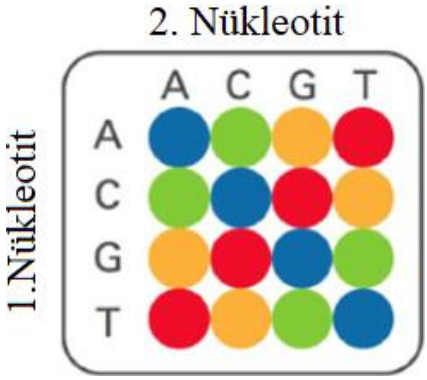
1.4.10. Solid Dizileme

Solid dizileme yöntemi literatürde yapışma ile dizileme olarak da geçmektedir. Bu teknikte kütüphane hazırlama ve PCR ile amplifikasyon aşamaları 454 prodizileme ile aynı şekilde gerçekleşmektedir. Analiz edilecek DNA parçası 25-50 bazı çifti arası nükleotid dizilimine sahip parçalara-sekanslara ayrılmaktadır. Bu parçalar ilk olarak tamamlayıcı DNA iplikleriyle daha sonra da üzerindeki emülsiyon PCR ile amplifikasyon sağlanacak manyetik boncuklarla birleştirilmektedir. Boncuklar sıralama işlemi için akışkan ortama işlenmiş cam slaydın üzerine tutturulmaktadır. Bu aşamadan sonra 454 prodizileme tekniğiyle farklılık başlamaktadır. 454 prodizileme teknolojisinde her okumada 1 baz okunup diziye eklenirken Solid dizilemede iki baz çifti okunup diziye eklenmektedir. PCR amplifikasyonu işlemi sayesinde manyetik boncuklar üzerinde şablon DNA parçasının çok sayıda kopyası bulunmaktadır. Okuma işlemi tüm kopyalar üzerinde gerçekleşmektedir [60, 61].



Şekil 1.17. Solid dizileme yönteminde algılayıcı tasarımı [60]

Bir çevrimde okunan iki baz çifti gerçek dizi ile uyuşmadığı sürece yapışma ve ışıma gerçekleşmemekte ve böylece doğruluk oranı %99,9 olan DNA dizilim elde edilmektedir [60, 62]. Solid dizilemede okunacak olan baz çiftleri için algılayıcı tasarlanmıştır. Bu algılayıcılar tanımlayıcı baz çiftleri, dejenere bazlar, evrensel bazlar ve flöresan molekülünden oluşmaktadır [61 - 63]. Tasarlanan algılayıcı Şekil 1.17'de gösterilmektedir [60].

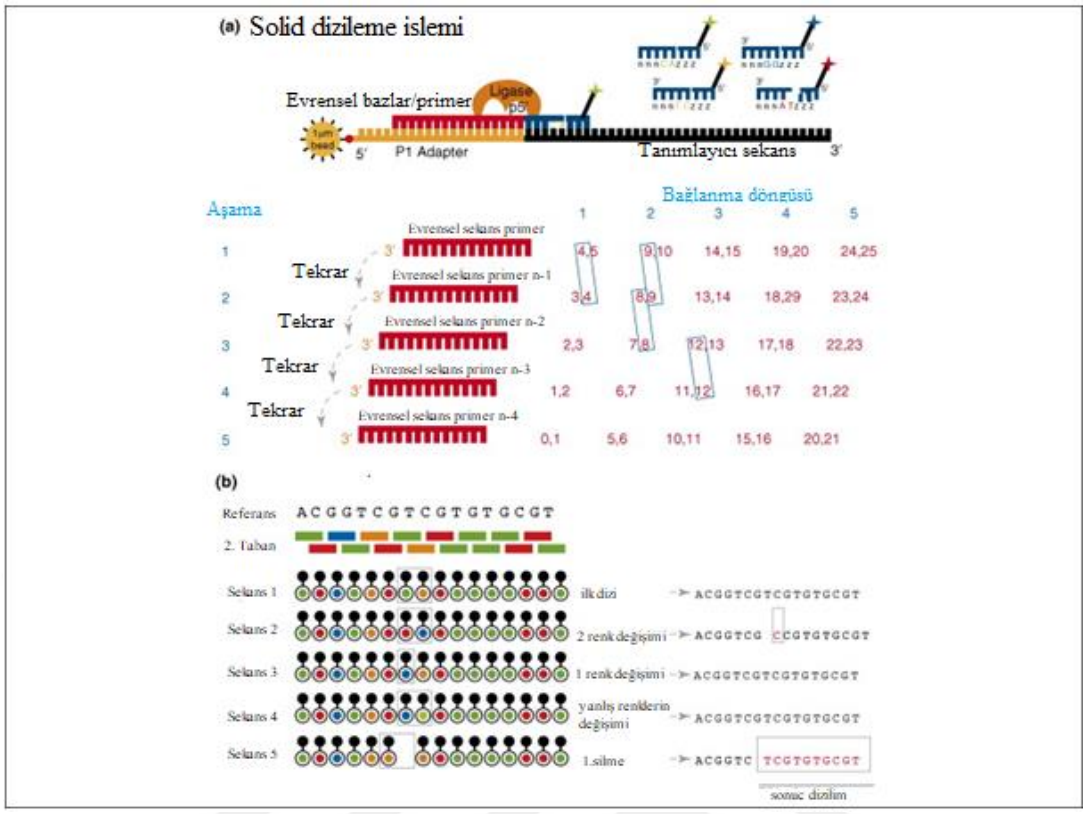


Şekil 1.18. Baz çiftlerini temsil ettiği ışın renkleri [60]

Floresandan yayılan ışın 4 nükleotidi temsil edecek şekilde 4 farklı renk ile gösterilmektedir. Işın renklerinin hangi baz çiftlerini temsil ettiği Şekil 1.18'de gösterilmektedir.

Dizilenecek DNA parçasının bulunduğu kuyucukta ligaz enzimi, algılayıcılar ve hedef dizideki adaptör diziye bağlanacak primer tabaka bulunmaktadır. İlk olarak primer adaptöre bağlanır ve bu primere ligaz ile dizideki uygun baz çifitine sahip bir algılayıcı dizi bağlanmaktadır. Daha sonraki adımda floresan molekül ayrılarak ışıma yapar ve bu ışıma kamera tarafından tespit edilmektedir [60 - 62]. Solid dizileme yöntemi çalışma aşamaları Şekil 1.19'da gösterilmektedir [61].

Floresan molekülünün yaydığı ışın belirtilen indekste hangi nükleotidin olduğunu kesin olarak göstermese de olasılıkları 4'e indirmektedir. Bu yüzden evrensel bazlar algılayıcı tasarımdan çıkarılmakta ve yeni algılayıcı dizilime eklenerek ışıma işlemi tekrarlanmaktadır. Bu işlemler dizinin sonuna kadar tekrarlı olarak devam etmektedir. Bu aşamadan itibaren aralarında 3'er bazlık boşluk bulunan dizilim için veri elde edilmektedir. Aralardaki boşluğu kapatmak için ise sentezlenen dizi tamamen atılmaktadır. Yeni primer öncekinden bir baz geride olmak üzere yeniden diziye bağlanmaktadır. Oluşan yeni dizilimde aynı işlemler yapılarak süreç devam etmektedir. Tüm bu işlemlerin tekrarıyla birlikte her tekrarla tanımlayıcı baz çiftleri de bir baz geriden bağlanarak algılayıcıların bağlanma sırasını değiştirmektedir. Bu şekilde ışıma renkleri değişmektedir. Her tekrarda primer bazlar geri kaydırılarak dizilim tamamlanmaya başlanarak beş çevrimde dizideki tüm boşluklar kapanmaktadır [60 - 63].

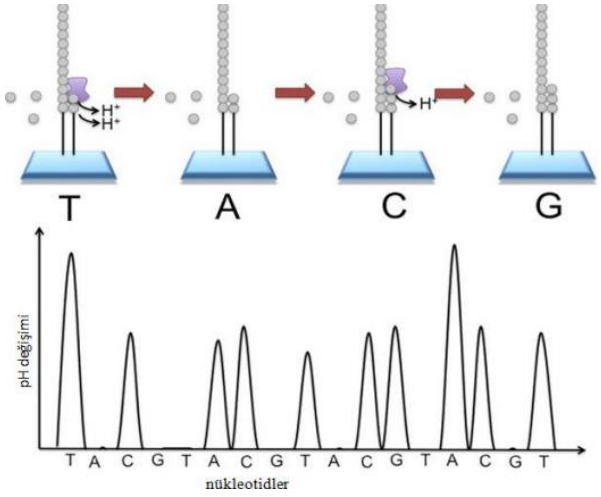


Şekil 1.19. Solid dizileme yöntemi çalışma aşamaları [61]

1.4.11. İyon Yarı İletken Dizileme

İyon yarı iletken dizileme yöntemi diğer yöntemlerin aksine boya etiketli nükleotidlerin tespiti üzerine değil, iyon tespitine dayalı olarak geliştirilmiştir. Yöntemin genel çalışma prensibi; transistörlerle geliştirilmiş sensör cihazı yardımıyla DNA zincir polimerazasyonu sırasında açığa çıkan hidrojen iyonun belirlenmesidir [64].

Dizileme işleminde ilk olarak dizilenecek DNA ipliğinin bulunduğu fragmentin çoğaltılıp DNA havuzu hazırlanmaktadır. Bu adımda dizilenecek olan fragmentin nükleotid eklenmesiyle elde edilecek sinyallerin daha doğru algılanması için emülsiyon PCR ile fragmentlerin çoğaltılıp havuz oluşturulmaktadır. Diğer aşamada ise polimeraz enzimi ile oligonükleotitler framentlere eklenmektedir. Sentez işlemi sonunda hidrojen iyonu ortama salınarak ortamın pH değeri değiştirilmektedir. Kullanılan iyon sensörü sentez işlemini tespit etmektedir [65].



Şekil 1.20. İyon yarı iletken dizileme yöntemi nükleotid eklenmesi [66]

Sentezleme işlemi sırasında havuzda birçok kopyası bulunan dizilenecek kalıp DNA ipliği, sırasıyla dört nükleotid DNA polimeraz enzimiyle reaksiyona girerek birleşmekte ve bu işlem döngüsel olarak devam ettirilmektedir (Şekil 1.20)[66]. Kalıp DNA ipliğinde eşleşen dNTP'ler bulununca sentezleme işlemi gerçekleşerek ortama H+ iyonu salınmakta ve bu salınımla ortamın pH değeri değişmektedir. Yarı iletken sensörler ile bu değişimi algılanarak sentezleme işlemi tespit edilmektedir. Bir sonraki döngüye geçilmeden önce ortamda bulunan eşleşmemiş dNTP'ler atılarak süreç devam etmektedir [64, 65, 67].

Geliştirilen iyon yarı iletken dizileme yönteminin hızlı, hazırlık aşaması basit ve maliyeti düşük olduğu vurgulanmaktadır. Yöntem eş zamanlı kullanılarak 100-200 nükleotid uzunluklu DNA parçasının yaklaşık 1 saatte okuduğu öne sürülmektedir. Sanger dizileme yöntemine göre ise daha kısa dizilimleri okuyabilmektedir [65].

1.5. DNA Dizileme Yöntemlerine Genel Bakış

1975'de geliştirilen Sanger dizileme yöntemi zamanla iyileştirilip ve makinelerle otomatikleştirilse de yüksek maliyetli olması ve elde edilen bilginin azlığı nedeniyle günümüzde kullanılan yeni nesil dizileme tekniklerinin geliştirilmesine ihitiyaç duyulmuştur. Yeni nesil dizileme yöntemleri geleneksel yöntemlere göre DNA başına 40.000 kat daha az maliyetle çalışarak 600.000 kat daha fazla bilgi edinmektedir [69].

Sanger dizileme yönteminde 500-700 nükleotid uzunluklu DNA parçaları üzerinde analiz çalışması yapılabilirken yeni nesil cihazlarda 80-200 nükleotid uzunluklu DNA

parçaları üzerinde işlem yapılabilmektedir. Bu parçalar işlem sonucunda birleştirilerek istenilen DNA dizilimi elde edilmektedir. Daha küçük dizilimler üzerinde işlem yapmak kolay ve etkin bir yöntem olsa da bu kısa parçaların birleşim aşamasında hatalar oluşmaktadır. Bu hatalar yeni nesil dizileme yöntemlerinin dezavantajı olarak gösterilmektedir [70]. Bu aşamadan sonra geliştirilen yeni yöntemler daha az maliyetli ve daha hızlı sürede sonuca ulaşmaya çalışsa da hata oranı düşük dizilimler elde eden yöntemlerin geliştirilmesi önem kazanmıştır.

Tablo 1.1'de yaygın kullanılan yeni nesil cihazların toplam analiz edebileceği DNA uzunluğu, oluşturulan sekans uzunluğu, uygulanan teknik ve çalışması sırasında oluşabilecek hata türleri gösterilmektedir [69].

Tablo 1.1. Yeni nesil dizileme cihazlarının çalışma şartları [69]

Cihaz	**Okuma Uzunluğu**	**Toplam DNA Uzunluğu**	**Teknik**	**Hata Türü**
Illumina	36, 50, 100	105-600 Gb	Geri dönüşümlü terminatör	Değişim
Applied Biosystems	35, 60, 75	7-9 Gb	Ligasyon ile sıralama	Değişim
Helicos BioSciences	25, 55	21-31 Gb	Tek molekül dizilemesi	Ekleme - Silme
454 Life Sciences	700	700 Mb	Sentez ile dizileme	Ekleme - Silme
IonTorrent	200	1 Gb	İyon yarıiletken dizileme	Ekleme - Silme

Yeni nesil dizileme cihazlarında nükleotid değişimi, silinmesi ve eklenmesi gibi hatalarla karşılaşılmaktadır. Bu hataların giderilmesi ile dizileme ve genomik çalışmaların performansı artırılmaktadır [69]. Bu nedenle DNA dizilime cihazlarında oluşan hataları gidermek için programlar geliştirilmektedir. Hata düzeltme işleminde RACER [71], Coral [72], HiTEC [73], SHREC [74,75] ve DecGPU [76] algoritmaları yaygın olarak kullanılmaktadır.

1.6. Hata Düzeltme Algoritmaları

Hata düzeltme algoritmaları yeni nesil DNA dizileme cihazları için kritik bir görev üstlenmektedir. Çünkü yüksek doğruluk oranlı sıralı okumalar daha yüksek kaliteli sonuçlara erişilmeyi sağlamaktadır. Hata düzeltme; bir dizileme platformu tarafından oluşturulan hatalı dizilimi tekrarlı okumalardan yararlanarak, her bir bazın doğru veya yanlış okunduğuna karar vermekte ve gerekli düzeltilmeleri yaparak doğru dizilimin geri döndürülmesini amaçlamaktadır. Bu bölümde DNA dizilimlerindeki oluşan hataların düzeltilmesi için geliştirilen algoritmalar anlatılarak karşılaştırılması yapılmaktadır.

1.6.1. Racer Algoritması

Okumalarda hataların hızlı ve doğru düzeltilmesi olarak verilen Racer k-mer sayma algoritmaları sınıfında kabul edilmektedir. Algoritmanın çalışma adımları şu şekildedir:

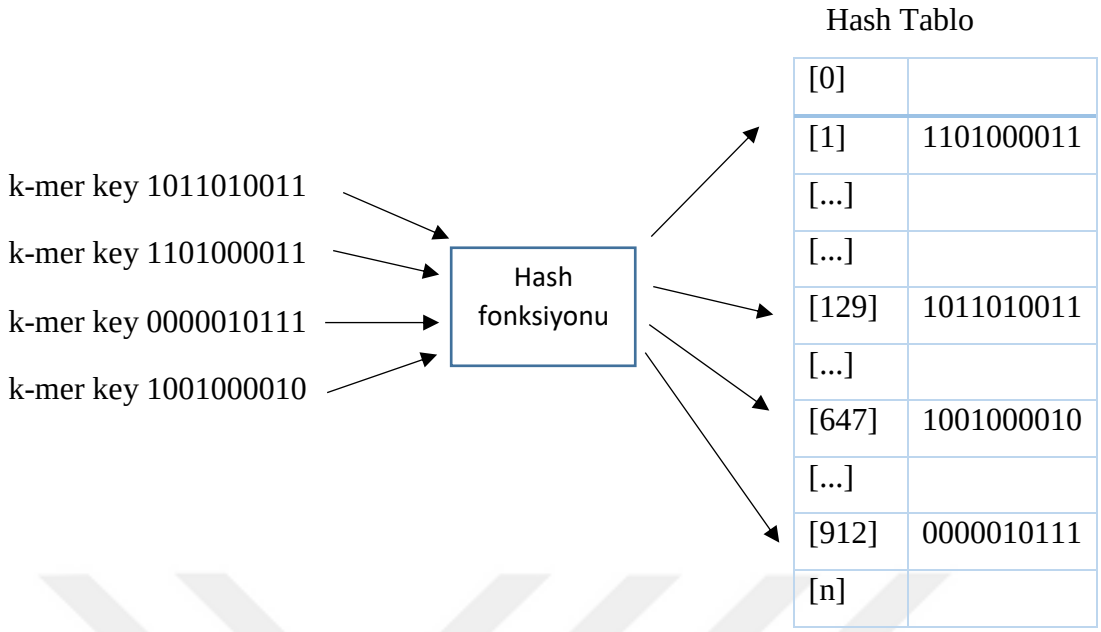
1. DNA dizilimleri 2 bitlik yazılır
2. Eşik ve k-mer uzunluğu belirlenir
3. Hash tablo oluşturulur
4. K-mer'lerin ve incelenecek nükleotidlerin saklanması
5. Hataların tespiti ve düzeltilmesi

Racer algoritmasında nükleotidler 2 bit, A (00), C (01), G (10) ve T (11) olacak şekilde temsil edilmektedir. Çünkü bitlerle işlem yapmak karakterle işlem yapmaktan daha hızlı olduğu ifade edilmektedir. Bu algoritmada okunamayan nükleotidlere rastgele değerler atanmaktadır. TAGCATTG DNA diziliminin ikili gösterimi Tablo 1.2'de gösterilmektedir.

Tablo 1.2. Racer algoritmasına göre DNA diziliminin ikili dönüşümü

T		**A**		**G**		**C**		**A**		**T**		**T**		**G**	
1	1	0	0	1	0	0	1	0	0	1	1	1	1	1	0

Algoritmada belirlenen k-mer uzunluğundaki DNA parçaları her biri 64 bitlik tam sayılarla ifade edilerek depolamada ve erişimde kolaylık sağladığı için hash tablosunda tutulmaktadır.



Şekil 1.21. Racer algoritmasında k-mer'lerin hash tablosunda saklanması [69]

Racer algoritmasında hataları belirlemek için bir eşik değeri kullanılmaktadır. Belirlenen k-mer, sekanslar içerisinde aranmakta ve bulunan parçalar saklanmaktadır. Saklama işleminde k-mer'den önceki ve sonraki gelen nükleotid kullanılmaktadır. Belirlenen k-mer için işlem bittiğinde, k-mer'den önce ve sonra gelen nükleotidlerin sayısı belirlenmektedir. Grup içerisinde ve belirlenen eşik değerinden yüksek olan nükleotidler doğru olarak kabul edilmektedir. Grup içerisinde olmayan ve belirlenen eşik değerinden küçük nükleotidler, doğru nükleotidler ile yer değiştirilmektedir [71, 77, 78]. Bu şekilde bir sonraki k-mer'e geçmeden önce tablo ve dizilim güncellenmektedir.

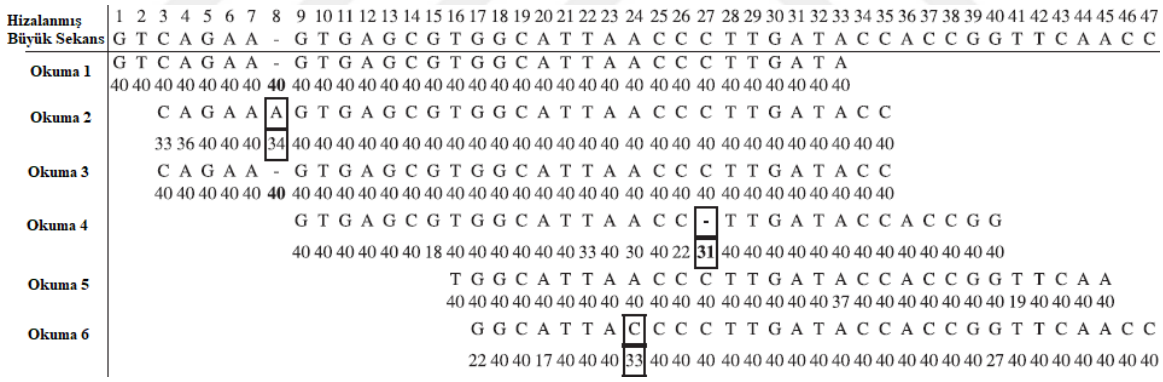
Analiz edilecek DNA büyüklüğü, k-mer'lerin sayısı ve eşik değeri başlangıçta belirlenmektedir. Bu değişkenlerin kombinasyonları deneysel sonuçlarla elde edilmiş olup bu sonuçlara uygun olarak belirlenmektedir [71]. Racer algoritmasının her k-mer işleminin sonunda dizilimi düzelterek ilerlemesi, bir sonraki işleme düzeltilmiş bir dizilimle başlaması avantaj olarak görülmektedir. Ayrıca farklı uzunluklu DNA dizilerini düzeltebildiği gibi seri ve paralel çalışabilmektedir [69].

1.6.2. Coral Algoritması

Coral, dizilimlerdeki hataları düzeltmek için kısa okumaların çoklu hizalamalarına dayalı yaklaşım kullanmaktadır. Algoritma çeşitli dizileme cihazlarının oluşturduğu farklı

hatalar için hata bulma modelini düzenleyebilmektedir. Bu sayede nükleotid değişimi, eklenmesi ve silinmesi olmak üzere oluşabilecek tüm hataları düzeltebilir niteliği taşımaktadır. Temel mantık ortak k-mer içeren sekansların tek tek hizalanıp nükleotidlerin puanlanması ve bu puanlar ile hataya karar verip doğru dizilim elde edilmesine dayanmaktadır [72, 78].

Belirlenen k-mer içeren sekanslar k-mer'in ilk bazının bulunduğu indis bilgisi ile beraber tutulmaktadır. Algoritmada ortak k-mer içeren parçalar o k-mer'in mahallesi olarak adlandırılmaktadır. Sonraki aşamada aynı mahalledeki sekanslar tek tek Needleman and Wunsch hizalama algoritması [79] ile hizalanmaktadır. Hizalanan sekanslar, çakışan k-mer alanları ve k-mer başlangıç indisler yardımıyla uzun sekans elde edilmektedir. Okumada eklenme hatası varsa o konumda uzun sekans parçasına sütun eklenmektedir. Eklenen sütuna hangi nükleotidin geleceğine, aynı sütunda yer alan nükleotidlerin var oluş sayısına bakılarak karar verilerek zaten hizalanmış olan uzun sekans parçası güncellenmektedir [72]. Bu yöntem okumalar benzer olduğunda iyi çalışırken, okuma benzerlikleri az olduğunda hizalama işlemi başarısı düşebilmektedir [72].



Şekil 1.22. Coral algoritmasında sekansların hizalanması [72]

Algoritmanın modeli oluşturulurken, uygun k-mer'ler, nükleotid eşleşme ödülü, nükleotidlerin uyumsuzluk cezası ve boşluk cezası için uygun değerler belirlenmektedir. Uygun k-mer'ler belirlenirken çoklu hizalamalarda hata düzeltmenin doğru olması için k-mer mahallindeki sekansların olabildiğince yakın olması gerekmektedir. Belirlenen k-mer'in genomda sadece bir kez tekrarlanması için k-mer boyunun seçimi önem kazanmaktadır. Şekil 1.22'de aynı k-mer mahallindeki sekanslar belirlenen eşleşme ödülü, uyumsuzluk ve boşluk cezası parametrelerine göre Needleman and Wunsch hizalama algoritması [79] ile

hizalanmakta ve bu algoritmadan dönen skorlar ile gösterilmektedir. Oluşan skorlara göre de hatalar tespit edilmektedir.

1.6.3. HiTec Algoritması

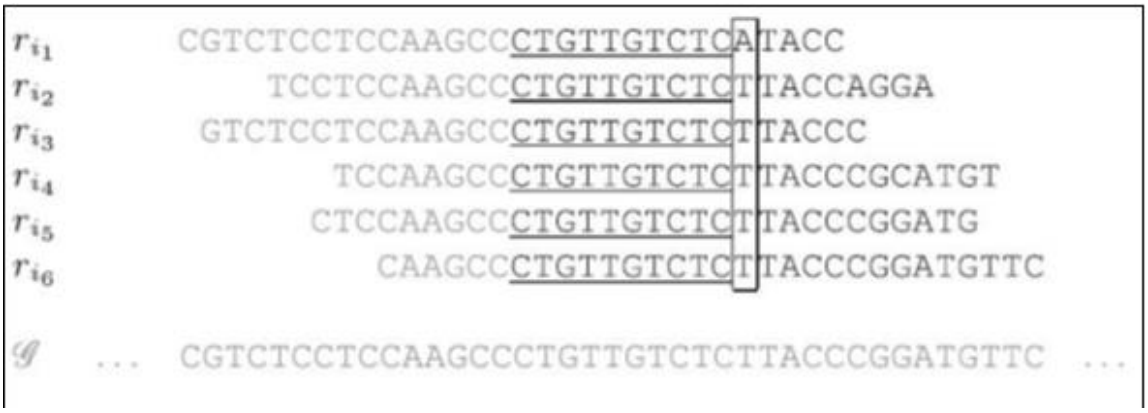
HiTec, yüksek verimli hata düzeltme algoritması, okuma dizileri ve ters tamamlayıcıları kullanarak oluşturulan son ek dizilerini kullanarak incelenen konumda en çok görülen nükleotide göre hataların giderilmesini sağlamaktadır. Algoritma istatiksel analiz kullanarak parametrelerini otomatik olarak belirlemekte ve her okuyuşta okuduğu sekanstaki hataları düzelterek ilerlemektedir [73, 78].

Ters tamamlama işleminde DNA dizilimindeki nükleotidlerin bağlandığı karşı nükleotidlerle yer değiştirme işlemi yapılarak son adımda tüm dizilim tersten yazılmaktadır. Örneğin s=CAT dizilimde A→T, C→G, T→A, G→C olacak şekilde yer değiştirme yapılmaktadır. Son olarak oluşan GTA dizisi tersten yazılarak s'nin ters tamamlayıcı dizisi elde edilmektedir [73].

HiTec, L uzunluklu genom için her okuma bir olacak şekilde p hataya sahip r_1, r_2, r_3, r_4, r_5, … r_n nükleotid dizileri oluşturmaktadır. A, T, G, C dışında oluşan bir bazı temsil etmeyen harfleri diziden çıkarmaktadır Sekanslama cihazları okuyamadığı nükleotidlerin yerine N yazmaktadır. Okumalar ve ters tamamlayıcılar $R=r_1\$r_1\$r_2\$r_2\$...r_n\$r_n\$$ şeklinde saklanmaktadır.

Düzeltme algoritması her bir r diziliminin j pozisyonundan başlayarak k pozisyonunda hata bulundurduğunu ve bu pozisyondan önceki w pozisyonlarının doğru olduğunu varsaymaktadır. Dizilimlerde tespit edilen hatanın diğer benzer dizilerde doğru olarak yazıldığı kabul edilmektedir. Benzer dizilerde ilgili indislerinde bulunan baz çeşidi, bulunma sayıları ve destek değerleri tespit edilerek indiste bulunması gereken nükleotide karar verilmektedir [73].

Şekil 1.23 incelendiğinde; kutu içerisindeki indisler karşılaştırıldığında r_1 okumasında A, diğer okumalarda T nükleotidi olduğu görülmektedir. Yani belirlenen indiste destek değerleri A=1 ve T=5 olarak bulunduğu tespit edilerek T nükleotidin doğru olduğuna karar verilmektedir.



Şekil 1.23. HiTec algoritması sekans karşılaştırma işlemi [60]

HiTec algoritmasının son ek dizi yapısı incelendiğinde aynı nükleotidi içeren tüm okumaların ardışık olduğu ve buna bağlı olarak da destek değerlerinin hesaplanmasının kolay olduğu sonucuna varılmaktadır [73].

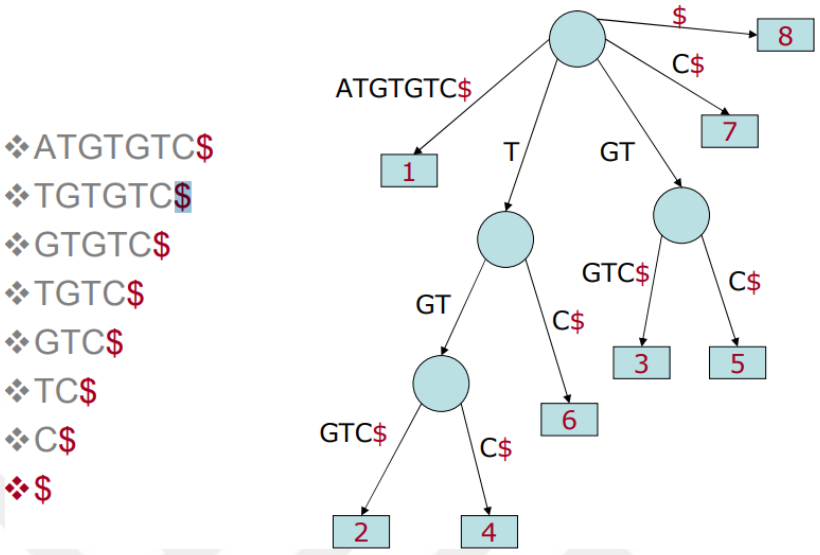
Bu algoritmada aynı hatanın birçok yerde bulunacağından dizilimlerdeki tekrarlar problem olarak kabul edilmektedir. Tekrarlar sebebi ile hatalı nükleotidin de destek değeri fazla çıkacağından doğru nükleotid ayrımını yapmak zorlaşmaktadır. Bu problemi çözmeye yönelik bir eşik değer belirlenerek nükleotidin doğru olduğuna eşik değerin aşılmasına göre karar verilmektedir. İstatiksel analiz ile eşik değeri hesaplanmaktadır. Eşik değer belirleyici bir unsur olana kadar deneysel sonuçlarla değiştirilerek elde edilmektedir [73]. Diğer bir problem ise dizilimde ardışık hataların bulunmamasından kaynaklanmaktadır. Bu hatayı gidermek için ise algoritmanın hata tespit edemediği okumaları saklayarak bu okumalarda w alanlarını küçülterek hata bulunma olasılığını artırmaktadır [73].

Bu alanda kullanılan en yaygın algoritmalardan biri olan Shrec ile karşılaştırıldığında daha az zamanda daha az yer kullanarak daha verimli sonuçlar elde ettiği gözlemlenmektedir. Dezavantaj olarak da okumalardaki belirsiz nükleotidleri düzeltememesi, aynı uzunluktaki veri setlerini düzeltmesi ve paralel modda çalışmaması gösterilmektedir [69].

1.6.4. Shrec Algoritması

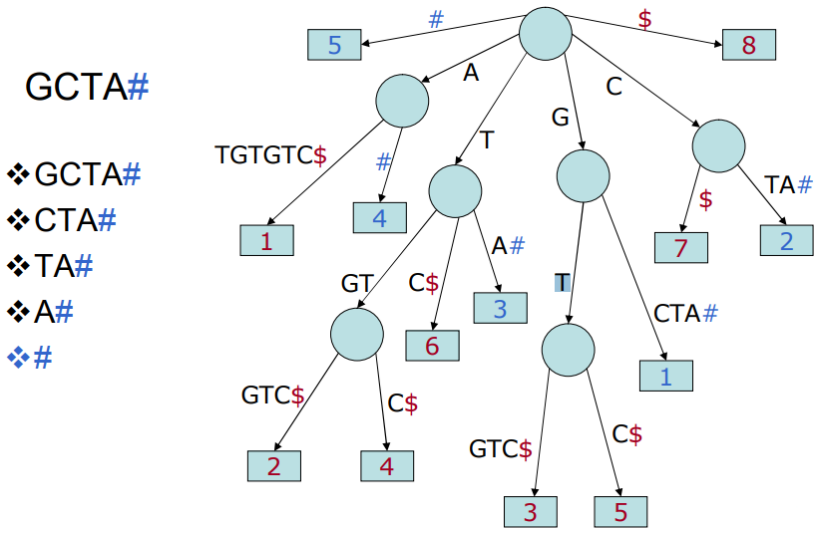
Shrec, kısa okuma hatası düzeltme algoritması, yeni nesil dizileme cihazı tarafından üretilen tüm okumalar için genelleştirilmiş bir son ek ağacı oluşturularak ve ağacın içindeki

düğümleri analiz ederek dizilerdeki nükleotid değişim hatalarını düzeltmeyi hedeflemektedir [74, 75].



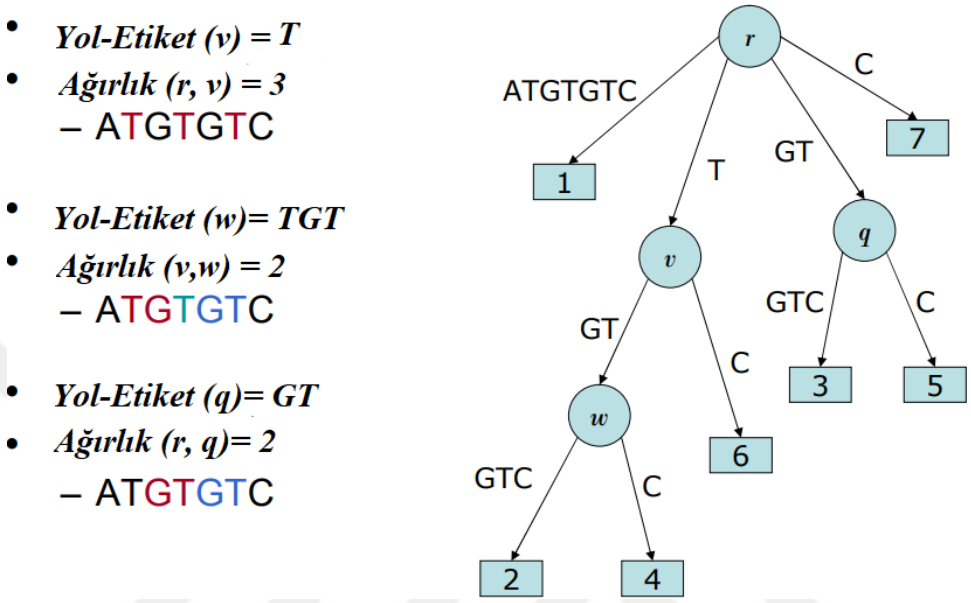
Şekil 1.24. Shrec algoritmasında son eklerin ağaç yapısında gösterimi [75]

Son ek ağaçları, analiz edilecek DNA parçasının tüm dizilerinin tüm eklerini içermektedir. Eklerin son ek olup olmadığı ayrımı benzersiz bir karakterle ($, #) mantıksal olarak sonlandırılarak belirlenmektedir. Kökten yaprağa giden bir yoldaki kenar etiketlerinin birleşiminin benzersiz olması son ek anlamına gelmektedir. ATGTGTC diziliminin son ek ağacı Şekil 1.24'de gösterilmektedir.



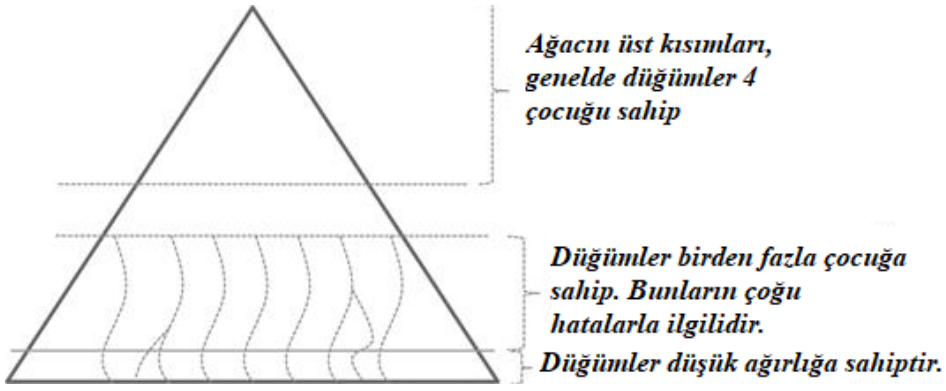
Şekil 1.25. Son ek ağacına yeni bir DNA parçasının sok ekinin eklenmesi [75]

Analiz edilecek DNA dizilimi birden fazla parçadan oluşarak her parçadaki son ekler ağaca eklenmektedir. Şekil 1.24'de gösterilen örneğe yeni bir DNA parçasının son ekleri eklendiğinde ağacın görünümü Şekil 1.25'de gösterilmektedir.



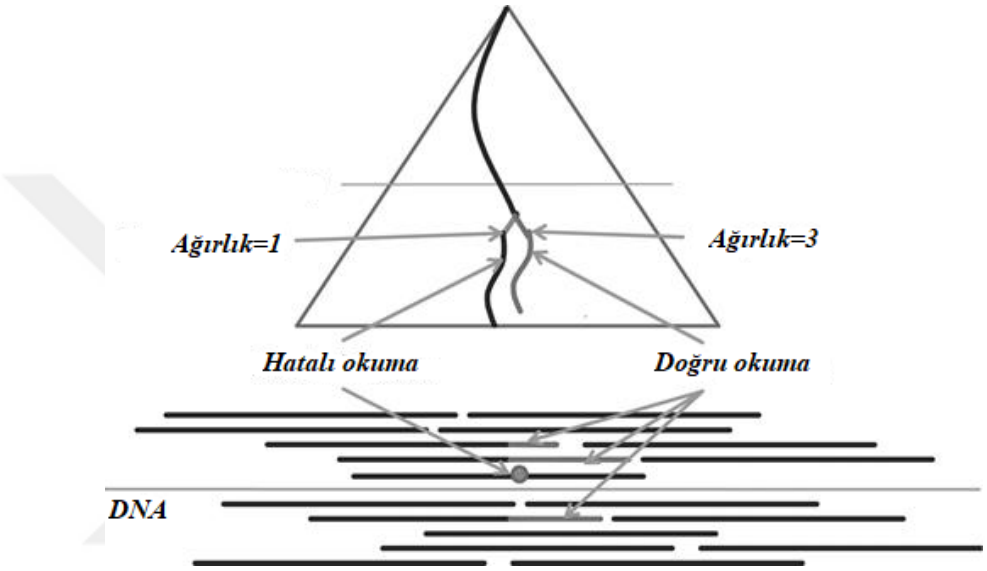
Şekil 1.26. Shrec algoritmasında ağaç düğümlerinin ağırlıklarının hesaplanması [75]

Şekil 1.26'da Shrec algoritmasının ağacın düğümlerinin ağırlıklarının tespit edilmesi gösterilmektedir. Her düğümün ağırlığı ilişki kenarının ağırlığına bağlı olarak değişmektedir. Düğümler arasındaki yol (v, w) w ile ilişki olarak gösterilmektedir. Ağırlık, dizi yolu etiketinin (w) okuma kümesinde bir alt son ek dizisinin görüldüğü sıklığı temsil etmektedir.



Şekil 1.27. Shrec algoritmasında son ek ağacında hataların tespiti [74]

Hata köke ne kadar yakın olursa tespiti o kadar zor olmaktadır. Çünkü köke yakın kısımlarda hemen hemen her düğümün dört çocuğu bulunmaktadır. Daha çok çocuğa sahip olan düğümlerin oranı, yüksek seviyelerde daha fazla olmaktadır. Buradaki düğümler doğru ve hatalı okumalar arasında güvenilir ayrım yapılmasına olanak sağlamaktadır. Dallanmanın en uç kısımlarına yaklaştıkça ise ağırlık değerleri çok küçüleceğinden hata tespitine karar vermek zorlaşmaktadır (Şekil 1.27) [74, 75].



Şekil 1.28. Shrec algoritmasında dallanmaların yorumlanması [74]

Ağacın düğümlerindeki dallamaların hatalı veya doğru olduğunu tespit etmek için düğümlerin ağırlıkları hesaplanmaktadır. Ağırlık hesaplanırken dizilerin sekanslarda kaç kez tekrarlandığı sayılarak ağırlığı fazla olan doğru kabul edilmektedir (Şekil 1.28) [74, 75].

Hataların tespit edilmesiyle hatalı son ek ağaçtan çıkartılmaktadır. Düzeltilmiş tüm okuma ekleri ağaca tekrar yerleştirilmektedir. Böylece bir sonraki hatanın tespiti kolaylaşmaktadır. Bu yöntem sürekli güncelleme sağladığından fazla zaman alsa da daha doğru çalıştığı gözlemlenmektedir [74].

1.6.5. DecGPU Algoritması

DecGPU algoritması, CPU ve GPU'nun paralel programlama modellerini birlikte kullanarak, yüksek verimli küçük okuma parçaları (DNA sekansı) için ilk hata düzeltme

algoritması özelliği taşımaktadır [76]. CPU tabanlı sürüm için MPI ve OpenMP, GPU tabanlı sürüm için CUDA ve MPI paralel programlama modellerinden yararlanmaktadır. GPU ve CPU'nun beraber kullanılmasıyla hesaplamaların performansı en üst düzeye çıkartılmaktadır [76]. DecGPU algoritması 4 aşamadan oluşmaktadır:

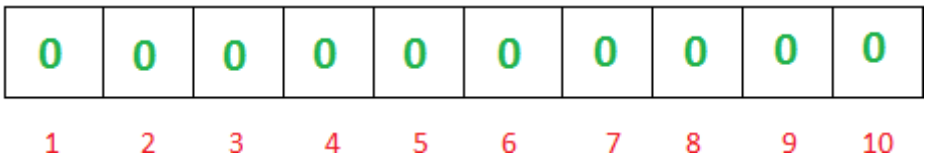
1. Dağıtık k-mer spektrumunun oluşturulması
2. Hatasız okumaların filtrelenmesi
3. Hatalı okumaların düzeltilmesi
4. Hatalı okumaların çıkartılması

Birden fazla hatayı gidermek için 2. ve 4. aşamalar arasında tekrarlı çalışmalar gerçekleştirmektedir.

1.6.5.1. Dağıtık K-mer Spektrumların Oluşturulması

DecGPU, Brujin graflarını kullanarak her bir düğümün tüm olası k-mer'lerini temsil edecek şekilde bloom filtresi kullanarak k-mer spektrumlarını dağıtmaktadır. Spektrum hizalama yaklaşımı, k-mer setlerini kullanarak bir genomdaki hataları tespit etmekte ve düzeltmektedir. Bu işlem için ilk adım olarak sekanslar üzerinde tüm olası k-mer'lerin seti belirlenmektedir [76].

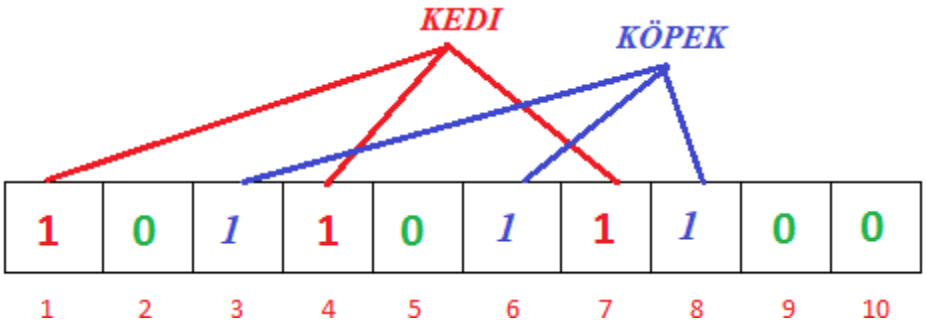
Bloom filtresi, bir elemanın bir küme içerisinde yer alıp almadığının olasılığını göstererek bellek ve süreyi verimli kullanan veri yapısı olarak tanımlanmaktadır. Aranan değer için "kümede" veya "kesinlikle kümede değil" sonucu dönmektedir. Başlangıçta tüm değerleri 0 olan bit vektörü veya bit listesi ile gösterilmektedir (Şekil 1.29)[80 - 82].



Şekil 1.29. DecGPU algoritmasında bloom filtresi başlangıç değerleri

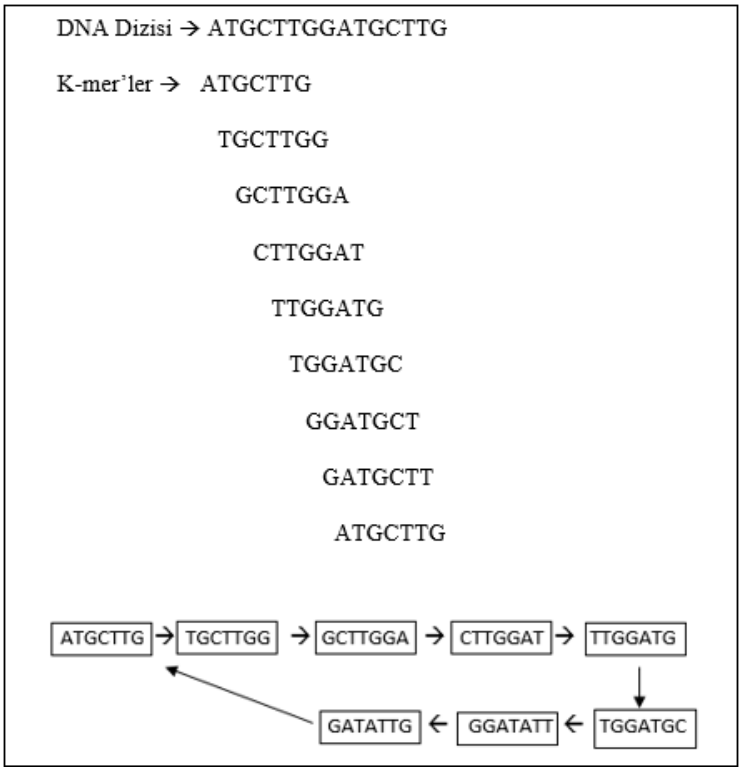
Yeni bir öğe eklemek için belirtilen konumlardaki değerler "1" olarak değiştirilmektedir. Aranan değerin küme içerisinde olup olmadığına karar vermek içinse indisteki karşılığına bakılmaktadır. "0" ise kesinlikle kümede yok, "1" ise kümede olabilir

sonucuna varılmaktadır. Şekil 1.29'daki listeye kedi ve köpek girdilerinin eklenmesi Şekil 1.30'da gösterilmektedir.



Şekil 1.30. DecGPU algoritmasında bloom filtresinin güncellenmesi

"Kedi" arandığı zaman listede 1., 4., ve 7. indislere bakılmaktadır. Bu indislerdeki değer "1" olduğu için "belki de 'kedi' zaten listeye eklenmiş" sonucuna ulaşılmaktadır. Fakat kedinin 1., 4. ve 7. indislerde değil de 1., 7. ve 9. indislerde çıktısını verseydi 9. indisteki değer "0" olduğundan kedinin daha önce listede işaretlenmediğine ve listede "kesinlikle bulunmadığı" sonucuna varılmış olurdu.



Şekil 1.31. K-mer'lerin Brujin graflarında gösterilmesi

Brujin grafları, k boyutlu bir A dizisinde tüm olası n uzunluklu alt dizileri içeren bir çevrim dizisi olarak tanımlanmaktadır. Bu veri yapısı geniş ölçekli problem için kullanılmanın yanı sıra gen dizisi elde etme ve kısa DNA parçalarını işlemek için de kullanılmaktadır. Şekil 1.31'de k=15 boyutlu DNA dizisinin n=7 boyutlu k-mer'leri tespit edilip Brujin grafında gösterilmektedir [83].

Tablo 1.3. DecGPU algoritmasında kullanılan hash tablosu örneği

K-mer	Sayaç
ATGCATT	2
TGCATTA	1
GCATTAA	5
CATTAAG	1

DecGPU algoritmasında her sekansın tüm k-mer'leri üzerinde arama işlemi yapılacağı göz önünde tutularak bloom filter mantığı ile hash tablosu kullanılmaktadır. Her bir k-mer {A, C, G, T} tabanlarını {0, 1, 2, 3, 4, ...} sayısal değerlerine eşleştirilerek hash tablosunda tutulmaktadır. Her oluşturulan k-mer, tabloda aranmakta ve varsa kendisine karşılık gelen değer artırılmaktadır. Kullanılan hash tablosuna örnek Tablo 1.3'de gösterilmektedir [76].

1.6.5.2. Hatasız Okumaların Filtrelenmesi

Birinci aşamada tüm k-mer'ler için hash tablosu sayaç bilgisi ile doldurulmaktadır. K-mer'lerin sayaçları kullanıcı tarafından belirlenen bir eşik değerinden az yani zayıf değilse bu k-mer sağlam olarak nitelendirilmektedir. Oluşturulan tüm k-mer'ler için sağlamlık kontrolü yapılmaktadır. Sağlam olarak belirlenen k-mer'ler hatasız kabul edilmektedir [76].

1.6.5.3. Hatalı Okumaların Düzeltilmesi

Bu aşamada sağlamlık kontrolü ile zayıf olduğu gözlemlenen k-mer'leri sağlam olanlara dönüştürülmektedir. Bu amaç doğrultusunda zayıf k-mer'lere oylama algoritması

uygulanmaktadır. Oylama algoritması, k-mer'in her bir pozisyonundaki tüm olası bazlarını değiştirerek sonuçtaki k-merin sağlamlıklarını kontrol ederek doğru tabanı bulmaya çalışmaktadır [76].

Algoritmanın temel çalışma prensibi, hatalı okunmuş sekansın soldan sağa tüm k-mer'lerin varlığının taranmasına dayanmaktadır. K-mer'lerin hash tablosunda sayaç değerleri kontrol edilerek doğru yada yanlış olduğuna karar verilmektedir. Taranan k-mer hatalı ise oylama matrisi ve dağıtık k-mer Brujin grafı kullanılarak düzeltilmektedir. Düzeltme işleminden sonra sekans ve k-mer'in hash tablosundaki sayaç bilgisi güncellenerek sağlam olarak belirlenmektedir [76].

1.6.5.4. Hatalı Okumaların Çıkartılması

Tespit edilen hatalı okumalar üzerinde geçekleştirilen hata düzeltme aşamasından sonra geride kalan sabit-okunmamış parçalarda kırma işlemi uygulanmaktadır. Bu aşamada kullanıcı tarafından tahmin edilen en uzun alt dizilim bulunmaya çalışılmaktadır. Dizilim bulununca sabit parçalardan kesilerek çıkartılmaktadır [76].

1.7. Algoritmaların Değerlendirilmesi

Algoritmaların değerlendirilmesinde duyarlılık ve özgüllük test yöntemlerinden faydalanılmaktadır. Bu yöntemlerle, geliştirilen algoritmanın okumaları hatasız veya hatalı olarak algılama ve hatalı okumaları düzeltme becerisi test edilmektedir. Duyarlılık testi, gerçek pozitif oranı olarak da adlandırmaktadır. Bu test hastalığı taşıyanların hangi oranla doğru olduğunu tespit etmektedir. Özgüllük testi ise gerçek negatif oranı olarak adlandırılmakta ve hasta olmayan bir hastayı doğru tespit etmektedir [84].

Açıklanan hata düzeltme algoritmaları ve bu çalışmada önerilen algoritmalar duyarlılık ve özgüllük test sonuçları ile değerlendirilmektedir. Hata düzeltme algoritmalarında duyarlılık, hatalı nükleotidlerin hangi oranla doğru tespit edildiğini, özgüllük doğru nükleotidlerin hangi oranla doğru tespit edildiğini test etmektedir [65].

$$Duyarlılık = \frac{TP}{TP + FN} \tag{1.1}$$

$$\text{Özgüllük} = \frac{TN}{TN + FP} \tag{1.2}$$

Duyarlılık hesaplaması Denklem 1.1'de, özgüllük hesaplaması Denklem 1.2'de gösterilmektedir. Denklemlerde kullanılan parametreler şunlardır:

- TP: Hatalı kabul edilen hatalı nükleotidlerin sayısı
- FP: Doğru olduğu halde hatalı kabul edilen nükleotidlerin sayısı
- FN: Hatalı olduğu halde doğru kabul edilen nükleotidlerin sayısı
- TN: Doğru tespit edilen doğru nükleotidlerin sayısı

2. YAPILAN ÇALIŞMALAR

2.1. Yöntem

Bu çalışmada yeni nesil dizileme cihazlarının DNA dizilerini tespit ederken oluşturduğu hataları gidermek amacıyla yeni bir algoritma önerilmiştir. Kullanıcıya kolaylık sağlamak için arayüz oluşturulmuştur. Oluşturulan arayüz ile analiz edilecek DNA parçası, hata oranı ve okuma sayısı dinamik olarak kullanıcıdan alınarak algoritma analiz sürecine başlanmaktadır. Önerilen algoritma ile yüksek doğruluk oranına sahip DNA parçası elde edilmesi amaçlanmış olup detaylar bu bölümde açıklanmaktadır.

2.1.1. Verilerin Elde Edilmesi

Çalışmada analiz edilen DNA dizisi NCBI (National Center for Biotechnology Information) olarak adlandırılan ulusal biyoinformatik bilgi merkezinden FASTA ve FASTQ formatlarında elde edilmiştir.

2.1.1.1 NCBI

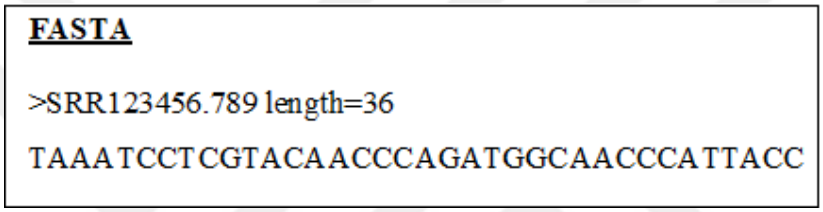
NCBI, hesaplamalı biyoloji dalında araştırmaları yürüterek, moleküler biyoloji ve tıptaki çeşitli problemlere teorik, analitik ve uygulamalı hesaplama yaklaşımlarına odaklanmaktadır. Çalışmalar, NCBI araştırmacıları, bilim insanları ve öğrencilerin yanı sıra harici araştırma toplulukları ile birlikte yürütülmektedir [85].

NCBI topluluğun uzmanlık alanlarına; DNA analizi, protein yapısı, fonksiyon analiz, kimyasal bilişim, genom analizi, hesaplama biyolojisi ve bilgi bilimde çok çeşitli konulara yoğunlaşmıştır. Bu konulara örnek olarak, veritabanı arama algoritmaları, sekans sinyali tanımlaması, evrimin matematiksel modellemesi, virolojide istatiksel yöntemler, kimyasal reaksiyon sistemlerinin dinamik davranışını, istatistiksel metin-geri kazanım algoritmaları, protein yapısını, fonksiyon tahminini, karşılaştırmalı genom, taksonomi, ağaçlar, popülasyon genetiği ve sistem biyolojisi örnek olarak verilebilir [85].

Yürütülen çalışmalar, NCBI'ın halka açık veri tabanlarınının ve yazılım uygulama araçlarının geliştirilmesine, bilimsel keşiflerle yenilikçi algoritmaların ve yeni araştırma alanlarının tespitine destek vermektedir [85].

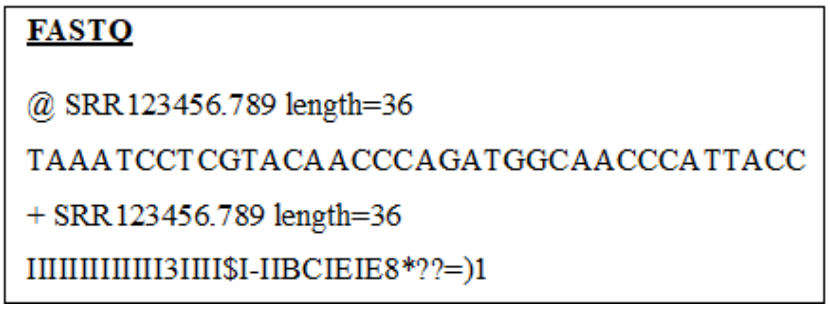
2.1.1.2. Fasta ve Fastq Veri Formatı

Fasta veri formatı iki satırdan oluşmaktadır. İlk satır '>' karakteri ile başlayarak dizilimi tanımlayan ifadelerle devam etmektedir. İkinci satırında ise DNA dizilim yer almaktadır. Fasta veri formatı Şekil 2.1'de gösterilmektedir.



Şekil 2.1. Fasta veri formatı

Fastq formatı dört satırdan oluşmakta ve Fasta formatından farklı olarak '@' karakteri ile başlamaktadır. Sonrasında DNA dizi tanıtıcısı ve isteğe bağlı açıklamayla devam etmektedir. Fastq veri formatında ikinci satırda ise DNA dizilimi yer almaktadır. Üçüncü satır '+' karakteri ile başlayarak dizi tanıtıcısı ve isteğe bağlı açıklama ile devam etmektedir. Son satırda da dizinin kalite değerleri ASCII kodları ile gösterilmektedir. Kalite kodu, her bir baz için bir karakter olacak şekilde diziyle aynı boyuta sahip harflerden oluşmaktadır. Fastq veri formatı Şekil 2.2'de gösterilmektedir.



Şekil 2.2. Fastq veri formatı

2.1.2. K-mer Uzunluğunun Belirlenmesi

L uzunluğunda bir dizinin k uzunluğundaki alt dizilerinin her biri k-mer olarak adlandırılmaktadır. Şekil 2.3'de L=15 boyutlu örnek bir DNA dizisinin k=4 bazlık k-mer'leri gösterilmektedir.

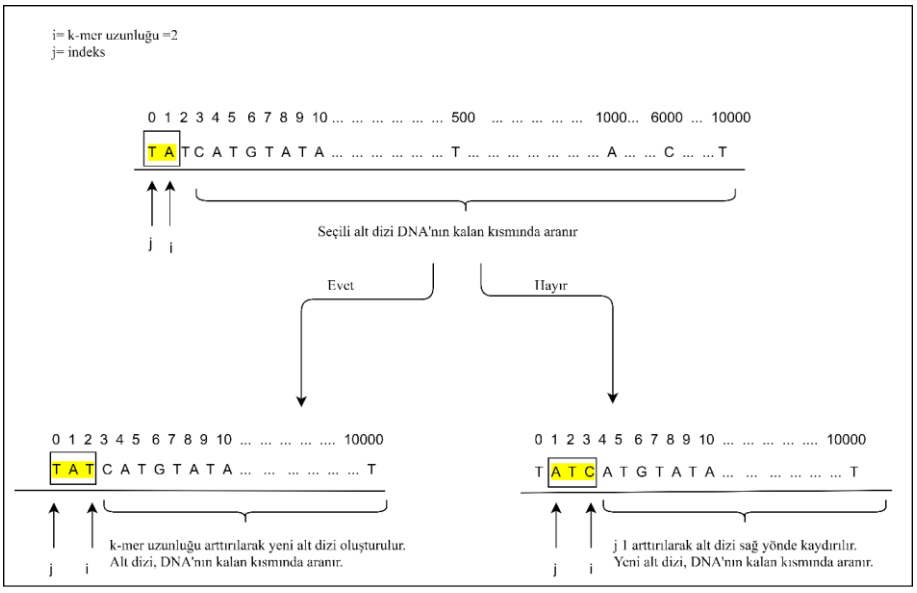
Dizi: ATTGACCATGATAGC

K-mer'ler:

ATTG	ACCA	TGAT
TTGA	CCAT	GATA
TGAC	CATG	ATAG
GACC	ATGA	TAGC

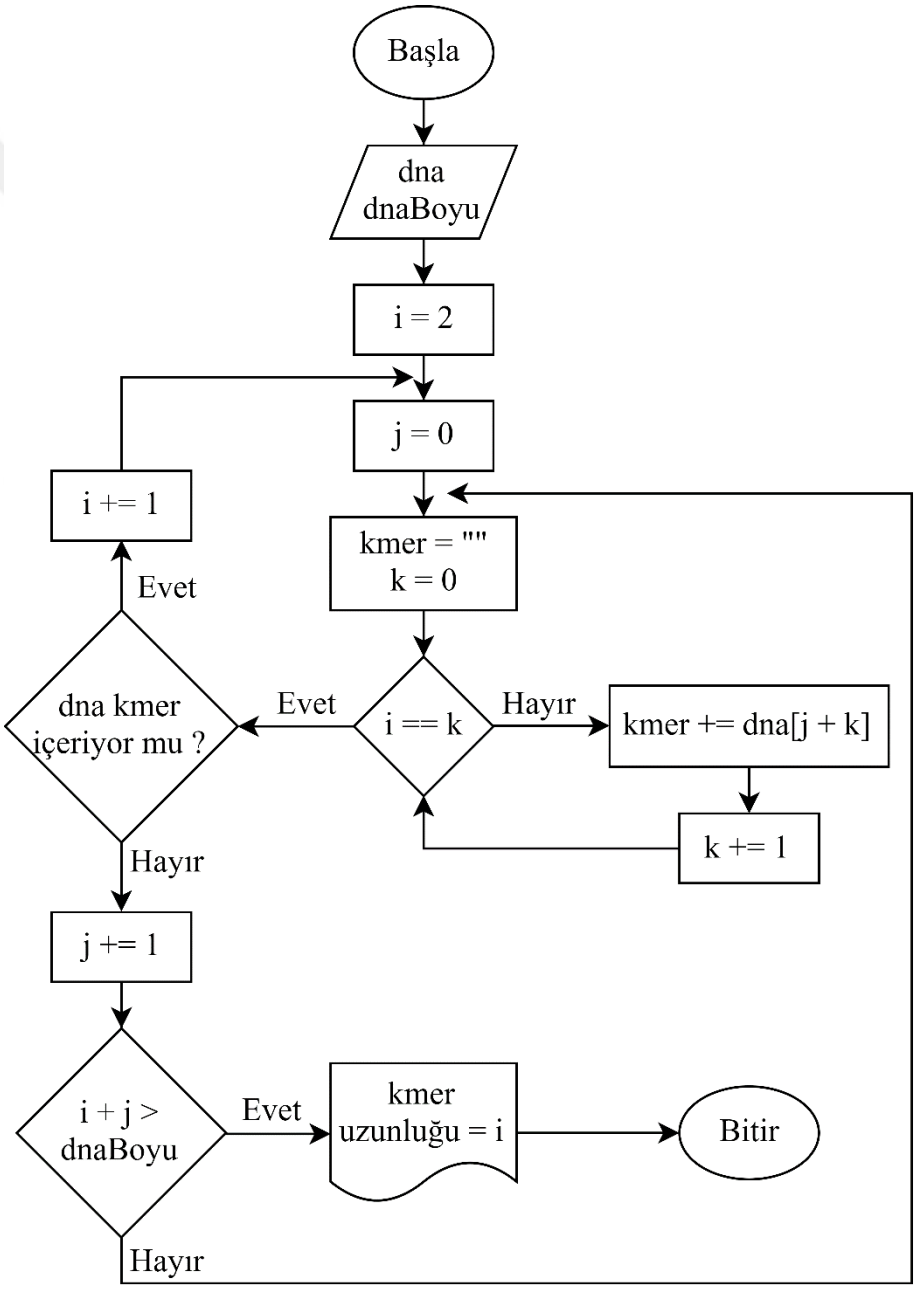
Şekil 2.3. DNA dizisinin 4 bazlık k-mer'leri

Önerilen algoritmada k-mer, her bir okuma için dizilimdeki eşsiz-tekrarsız olan en küçük alt diziyi ifade etmektedir. Bu işlem DNA'sı dizilenecek olan canlının, tekrarlı okumalar sonucunda elde edilen parçaların ortak bölgelerinin gruplandırılması için kullanılmaktadır. Gruplandırılan sekanslar kendi aralarında hata düzeltme algoritmasına tabi tutularak grubu ifade eden doğru dizilim elde edilmeye çalışılmaktadır. Şekil 2.4'de k-mer uzunluğu belirleme işlemi anlatılmaktadır.



Şekil 2.4. Geliştirilen k-mer uzunluğu bulma algoritması çalışma şekli

Örneğin k-mer uzunluğu 13 olarak bulunmuş ise sekansların 13 boyutlu alt dizileri sadece kendilerine özgü olması beklenmekte ve aynı okumadaki farklı hiç bir sekansta bulunması beklenmemektedir. Ancak analiz edilen canlı çok kez okunduğu için seçili alt dizi, ortak okunan alanlardan oluşturulan sekanslarda bulunabilmektedir. Seçili alt dizi global olarak tüm sekanslar arasında arandığında alt dizi içeren her sekans aynı konumun okunduğunu ifade ettiğinden kendi içinde gruplandırılmış olmaktadır. Geliştirilen k-mer uzunluğu bulma işlemi akış diagramı Şekil 2.5'te gösterilmektedir.



Şekil 2.5. Geliştirilen k-mer uzunluğu bulma akış diagramı

2.1.3 Sekansların Oluşturulması

Yeni nesil dizileme cihazları; az zamanda yüksek doğruluk oranlı sonuçlar elde etmek için okuma işleminde, tüm DNA dizilimini oluşturmak yerine DNA'yı parçalar-sekanslar halinde oluşturmaktadır. Önerilen hata giderme algoritmasının bu kısa veri kümeleri üzerinde daha hızlı ve doğru sonuçlar elde etmektedir.

```
sekans  1  :   GCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAAACGGTTCCG
sekans  2  :   GGCTCCGGTCGCCAATGTTGAAAGCGATGTCGGTTGTCTGGAATTGTTCC
sekans  3  :   CCAACGCTGGCATTTAAAGATTTCGGCGGTCGCTTTATGGCACAAATGCT
sekans  4  :   CCGATAAGCTGCCGTCAGAACCACGGGAAAATATCGTTTATCAGTGCTGG
sekans  5  :   AGCGTTTTTGCCAGGAACTGGGTAAGCAAATTCCAGTGGCGATGACCCTG
sekans  6  :   AAAAGAATATGCCGATCGGTTCGGGCTTAGGCTCCAGTGCCTGTTCGGTG
sekans  7  :   TCGCGGCGCTGATGGCGATGAATGAACACTGCGGCAAGCCGCTTAATGAC
sekans  8  :   CTCGTTTGCTGGCTTTGATGGGCGAGCTGGAAGGCCGTATCTCCGGCAGC
sekans  9  :   TCATTACGACAACGTGGCACCGTGTTTTCTCGGTGGTATGCAGTTGATGA
sekans  10 :   GAAGAAAACGACATCATCAGCCAGCAAGTGCCAGGGTTTGATGAGTGGCT
sekans  11 :   TGGGTGCTGGCGTATCCGGGGATTAAAGTCTCGACGGCAGAAGCCAGGGC
sekans  12 :   ATTTTACCGGCGCAGTATCGCCGCCAGGATTGCATTGCGCACGGGCGACA
sekans  13 :   TGGCAGGCTTCATTCACGCCTGCTATTCCCGTCAGCCTGAGCTTGCCGCG
  ...          ..................................................
  ...          ..................................................
  ...          ..................................................
sekans  n-2:   GCTGATGAAAGATGTTATCGCTGAACCCTACCGTGAACGGTTACTGCCAG
sekans  n-1:   CTTCCGGCAGGCGCGGCAGGCGGTCGCGGAAATCGGCGCGGTAGCGAGCG
sekans  n  :   GTATCTCCGGCTCCGGCCCGACCTTGTTCGCTCTGTGTGACAAGCCGGAA
```

Şekil 2.6. Oluşturulan DNA sekansları

Önerilen algoritmada yeni nesil dizileme cihazları simüle edilerek dinamik olarak belirlenen boyutlarda sekanslar oluşturulmaktadır. Her sekans oluşturulurken yine dinamik olarak belirlenen hata oranıyla rastgele nükleotid değişimi, silinmesi veya eklenmesi hataları yapılmaktadır. Hata oranı, gerçek cihazların çalışma prensibi göz önünde bulundurularak %1-4 arasında değişen değerlerde olmasına dikkat edilmektedir. Oluşturulan DNA sekansların görünümü Şekil 2.6'da gösterilmektedir.

Oluşturulan sekansların sayısı, coverage (okuma derinliği veya kapsama) olarak ifade edilen DNA analizi edilecek canlının okuma derinliğini ifade eden değişkene bağlı olarak hesaplanmaktadır. Oluşturulan sekansların sayısı Denklem 2.1'de gösterildiği gibi hesaplanmaktadır.

$$n = \frac{g * k}{s} \tag{2.1}$$

Denklem 2.1'de kullanılan parametreler şunlardır:

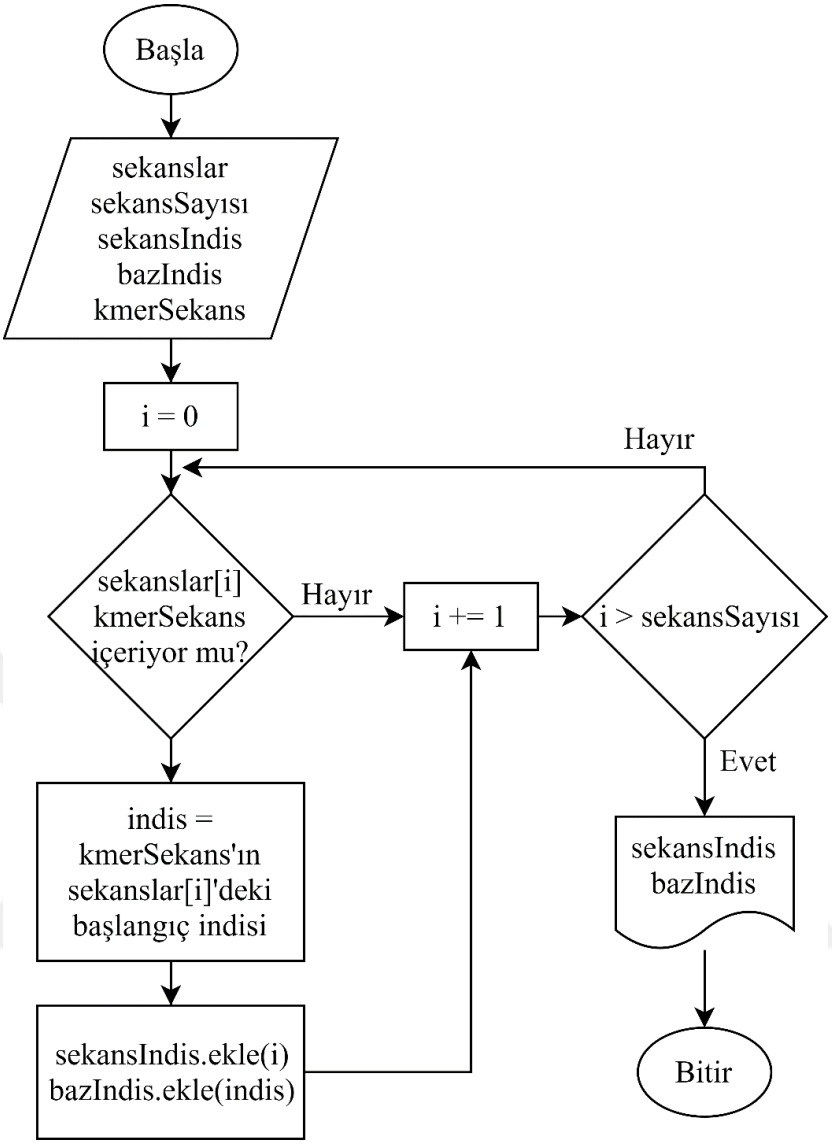
- n: Sekans sayısı
- g: Gen uzunluğu
- k: Okuma derinliği (kapsama)
- s: Sekans uzunluğu

2.1.4. K-mer Sınıflarının Oluşturulması

Hata düzeltme algoritmasına geçilmeden önceki son aşama k-mer sınıflarının oluşturulmasıdır. İlk olarak rastgele bir sekans içerisinde rastgele indiste k-mer uzunluğunda bir alt sekans belirlenmektedir. Belirlenen k-mer uzunluklu alt sekans tüm oluşturulan sekanslarda aranmaktadır. Alt sekansı içeren sekanslar DNA dizisinde aynı konumu ifade ettiğinden bulunan sekanslar bir listede toplanmaktadır. Sonrasında listede bulunan sekanslar kullanılarak hata düzeltme süreci başlatılmaktadır.

Şekil 2.7'de 13 bazdan oluşan k-mer alt sekansını içeren sekanslar k-mer'e göre hizalanmış bir şekilde gösterilmektedir. Aynı k-mer'i barından sekansların tespit edilme ve listeye eklenmesi Şekil 2.8'de gösterilen akış diagramı ile gerçekleştirilmektedir.

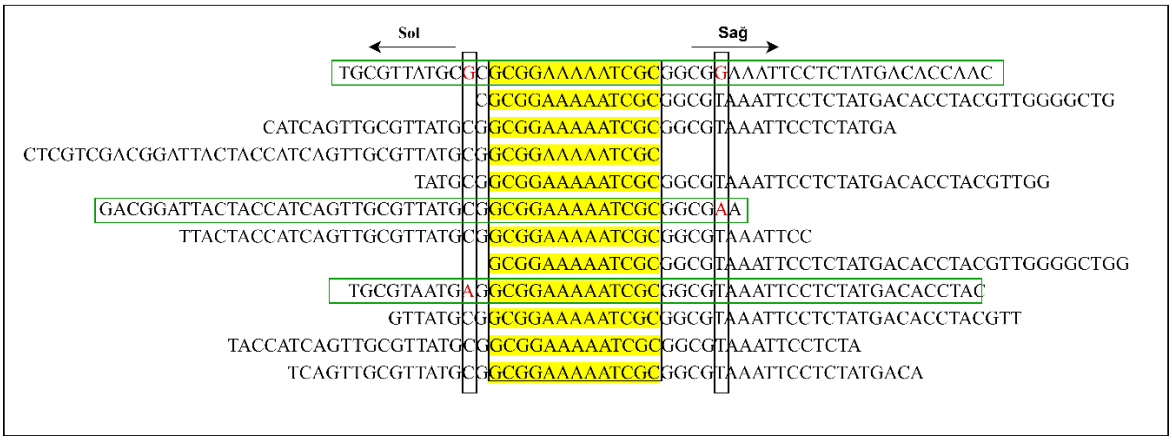
Şekil 2.7. Uzunluğu 13 olan k-mer'in sekansları sınıflandırması



Şekil 2.8. Aynı k-mer – alt sekans içeren sekansların belirleme işlemi akış diagramı

2.1.5. Hatalı Sekansın Tespiti

Aynı k-mer uzunluklu alt-sekansı içeren sekanslar k-mer konumlarına göre hizalanmaktadır. Hizalanma işleminden sonra alt-sekansın sağ ve sol yönünde hata arama süreci başlamaktadır. Bu süreçte ilk olarak kontrol sol yöne doğru başlamakta ve k-mer uzunluğu kadar devam etmektedir. Sol yönlü tarama işleminin sağ yönlü tarama işleminden farkı indisin artarak değil azalarak ilerlemesidir. Sağ yönlü hatalı sekans bulma işlemi aşağıda adımlar halinde açıklanmaktadır (Şekil 2.9):



Şekil 2.9. Hatalı sekansın tespit edilmesi

1. Sağ yöndeki ilerleme her sekansın alt-sekans (k-mer) bitişinden sonraki indisten başlamakta (*i*. indis) ve artarak devam etmektedir.
2. Sırası gelen işlem indisinde global çoğunluk oylama yapılarak o indisin olması gereken nükleotid belirlenmektedir. Yani aynı gruptaki sekansların belirlenen indiste hangi nükleotidin olması gerektiğine en fazla bulunan nükleotid sayısı ile karar verilmektedir. Örneğin 12 sekans için *i*. indiste toplam 10 adet T ve 2 adet A nükleotidi tespit edilmiş ise *i*. indiste T nükleotidi olması gerektiği sonucuna varılmaktadır.
3. *i*. indisin olması gereken nükleotidi belirlendikten sonra o nükleotidden farklı nükleotide sahip olan sekans yok ise tarama işlemi bir sonraki indise geçerek ((*i* + *1*). indis) süreç devam etmektedir (2. adım).
4. *i* indisi doğru nükleotidden farklı bir nükleotid içiren sekanslar var ise farklılığın bulunduğu ilk sekansın indisi hatalı indis, sekans ise hatalı sekans olarak belirlenmektedir. Şekil 2.9'da hatalı sekans ve indisler renkli kutucuklar içerisine alınarak gösterilmektedir.
5. Sağ yönlü tarama işlemiyle hatalı bulunan nükleotid düzeltilemez ise sağ yönlü tarama (*i* + *1*). indisten taramaya devam ederek bir sonraki hatayı bulmaya çalışmaktadır. Eğer hata düzeltilmiş ise *i* arttırılmayarak aynı indiste başka hata var mı diye kontrol edilmektedir (2. adım).
6. Bahsedilen şekilde k-mer boyu kadar indis kontrol edildikten sonra yeni bir k-mer uzunluklu alt sekans belirlenerek hata arama süreci devam ettirilmektedir (1. adım).

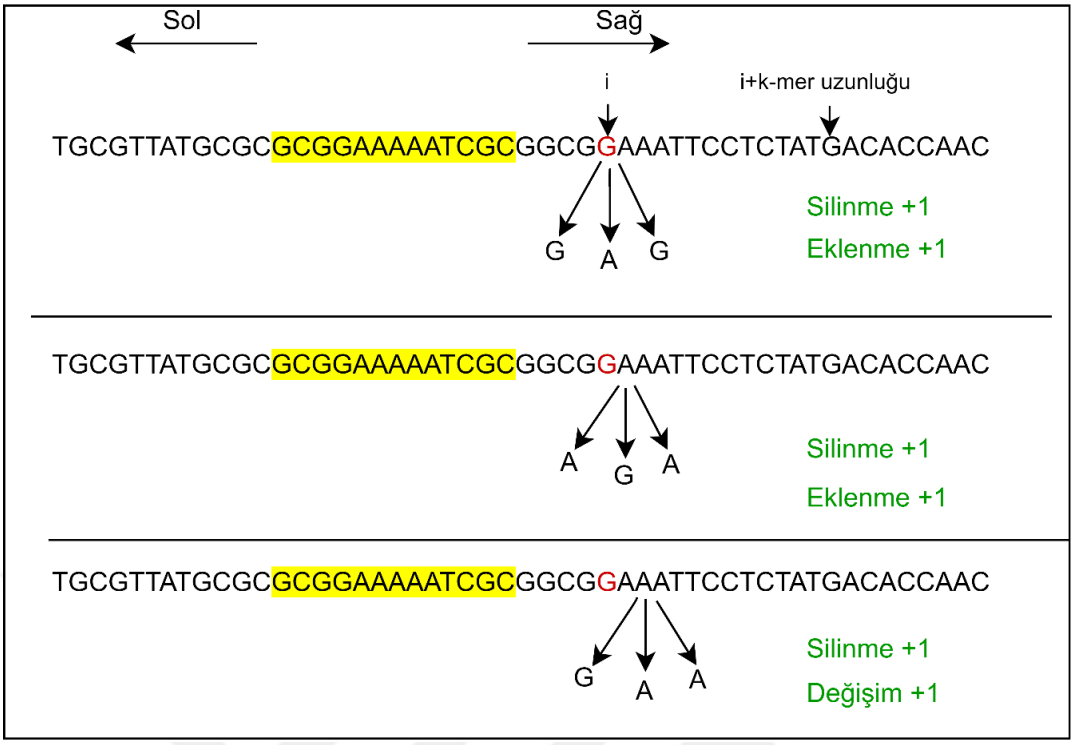
7. Sağ yönlü ilerleme yapılırken her sağ ilerleme için bir sol ilerleme yapılmaktadır. Bunun sebebi eğer aranan alt sekansta hata ver ise bunu bir sonraki alt sekans ile düzeltmektir.
8. Sağ yönlü yapılan işlemlerin hepsi benzer şekilde sol yönlü de yapılmaktadır. Sol yönlü işlemlerde *i* azaltılarak arama devam ettirilmektedir.

2.1.6. Sekanstaki Hata Tipinin Bulunması

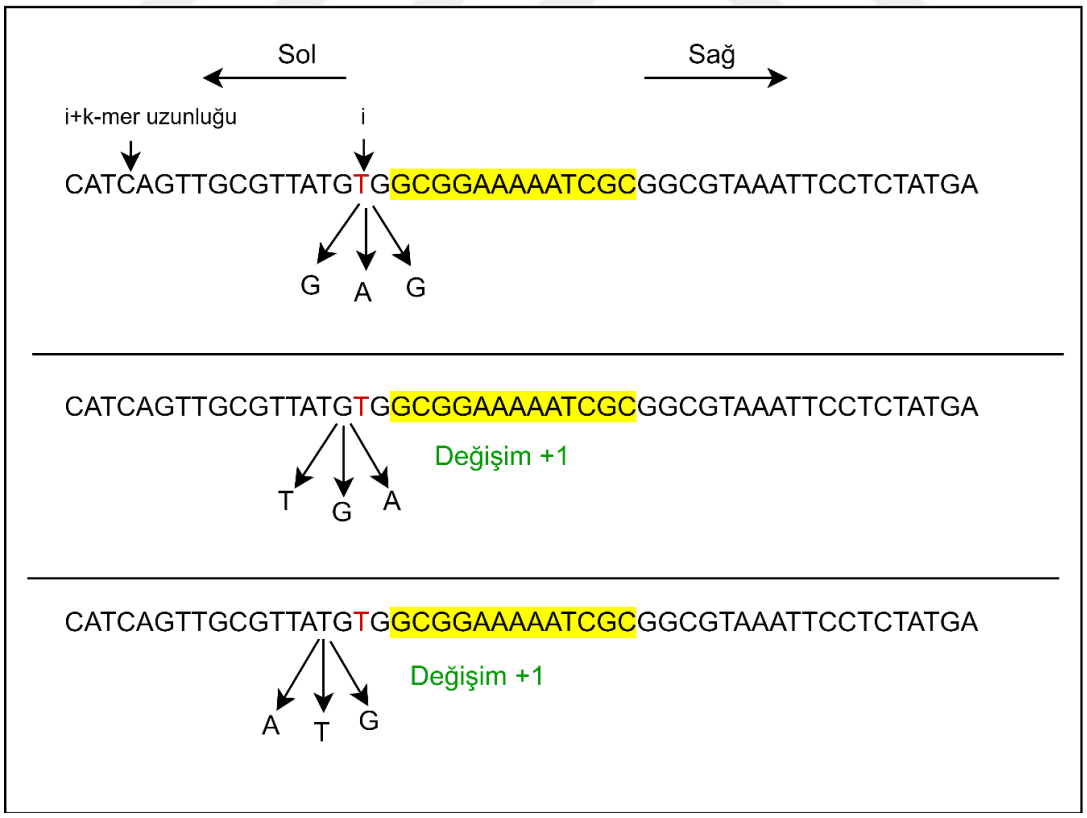
Hatalı sekans bulma işleminde olduğu gibi hata türünü bulma işlemi de sağ ve sol yönlü akışlarda küçük farklılıklar göstermektedir. İlk adım olarak gerçekleşen sol yönlü kontrol işleminde bulunan hatalı sekansın, hatalı indisi (*i*. indis) üzerinde hata tipi tespit işlemi başlamaktadır.

Hata tipinin tespiti için *i*. indisten başlayarak *i* arttırılarak karşılaştırma işlemi yapılmaktadır. Nükleotidlerin karşılaştırılması k-mer uzunluğu kadar devam etmektedir. Her bir nükleotid (*i - 1*), *i* ve (*i + 1*). indislerinin global çoğunluk oylaması sonucu doğru olarak belirlenen nükleotid ile karşılaştırılmaktadır. Karşılaştırma sonucu eşit ise ilgili hata tipi sayacı 1 artırılmaktadır. Tüm indislerdeki karşılaştırma işlemi bitince hata tiplerinin sayaçları kontrol edilerek sayacı büyük olan hata tipi gerçekleşen hata tipi olarak kabul edilmektedir. Sağ yönlü karşılaştırma; silinme, karşılıklı karşılaştırma; değişim ve sol yönlü karşılaştırma; eklenme hata tipini ifade etmektedir. Şekil 2.10'da sağ yönlü karşılaştırma ve sayaç artırma işlemi gösterilmektedir. Şekil 2.10'da tespit edilen hata tipi silinme ve silinen nükleotid A olarak belirlenmektedir.

Sol yönlü hata tespit işleminin sağ yönlüden farkı karşılaştırma işleminin *i*. indisten başlayarak *i* azaltılarak k-mer uzunluğu kadar karşılaştırılma yapılmasıdır. Sağ yönlü karşılaştırma (*i + 1*); eklenme, karşılıklı karşılaştırma (*i*); değişim ve sol yönlü karşılaştırma (*i - 1*); silinme hata tipini ifade etmektedir. Şekil 2.11'de sol yönlü karşılaştırma ve sayaç artırma işlemi gösterilmektedir. Bu şekilde tespit edilen hata tipi değişim ve değişen nükleotid A olarak belirlenmektedir.



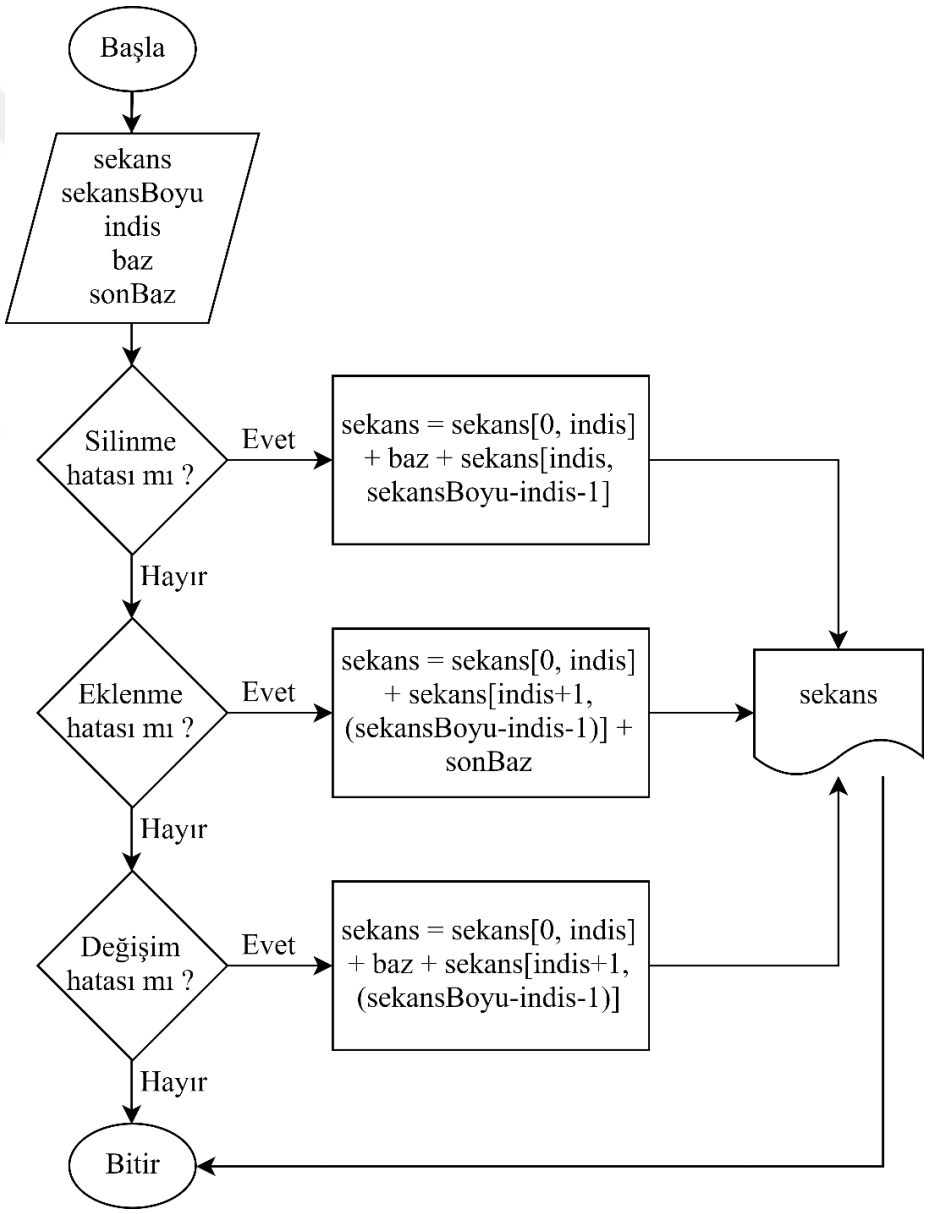
Şekil 2.10. Sağ yönünde karşılaştırma ve sayaç artırma işlemi



Şekil 2.11. Sol yönünde karşılaştırma ve sayaç artırma işlemi

2.1.7. Hatanın Düzeltilmesi

Hatalı sekans, hatalı indis (*i*. indis) ve hata türünün tespit edilmesi ile sıra hatanın düzeltilmesi işlemine gelmektedir. Düzeltme işlemi, hatalı indiste bulunması gereken nükleotidin yani doğru nükleotidin (indisin grup içinde global çoğunluk oylama ile bulunan) hata türlerine göre dizilimde düzeltilmesine dayanmaktadır. Hata türlerine göre yapılan işlemler sağ ve sol yönlü karşılaştırma için farklılık göstermektedir. Sağ yönlü karşılaştırma işleminde geliştirilen hata düzeltme işlemi akış diagramı Şekil 2.12'de gösterilmektedir.

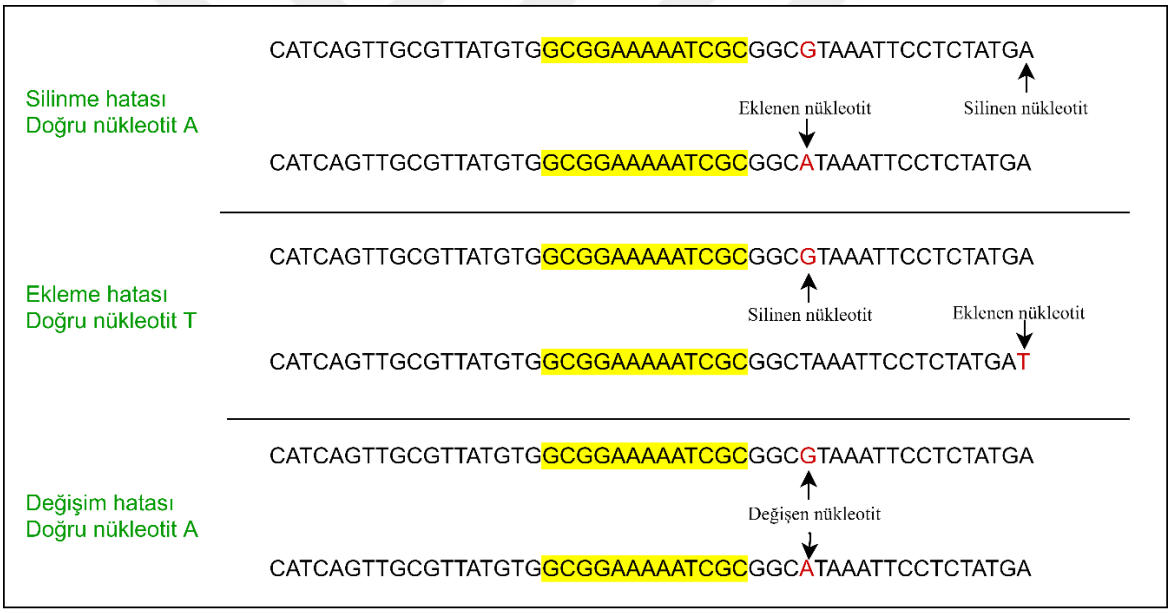


Şekil 2.12. Sağ yönlü karşılaştırma için geliştirilen hata düzeltme işlemi akış diagramı

Sağ yönlü karşılaştırma işleminde hatanın düzeltilmesi:

- Silinme hatası için doğru nükleotid, hatalı indisin bir önceki (*i - 1*) indisine eklenmektedir. Eklenme işleminde sekans uzunluğundaki aşmayı engellemek için eklenen nükleotid kadar dizilim sonundan nükleotid silinmektedir.
- Eklenme hatası için hatalı indisteki nükleotid silinmektedir. Silinme işlemi ile kısalan sekans uzunluğunu gidermek için silinen nükleotid kadar son indis için global çoğunluk oylaması yapılarak bulunan nükleotid sekans sonunda eklenmektedir.
- Değişim hatası için doğru nükleotid, hatalı indisteki nükleotid ile değiştirilerek hata giderilmektedir.

Sağ yönlü karşılaştırma için geliştirilen hata düzeltme işlemi Şekil 2.13'de gösterilmektedir.

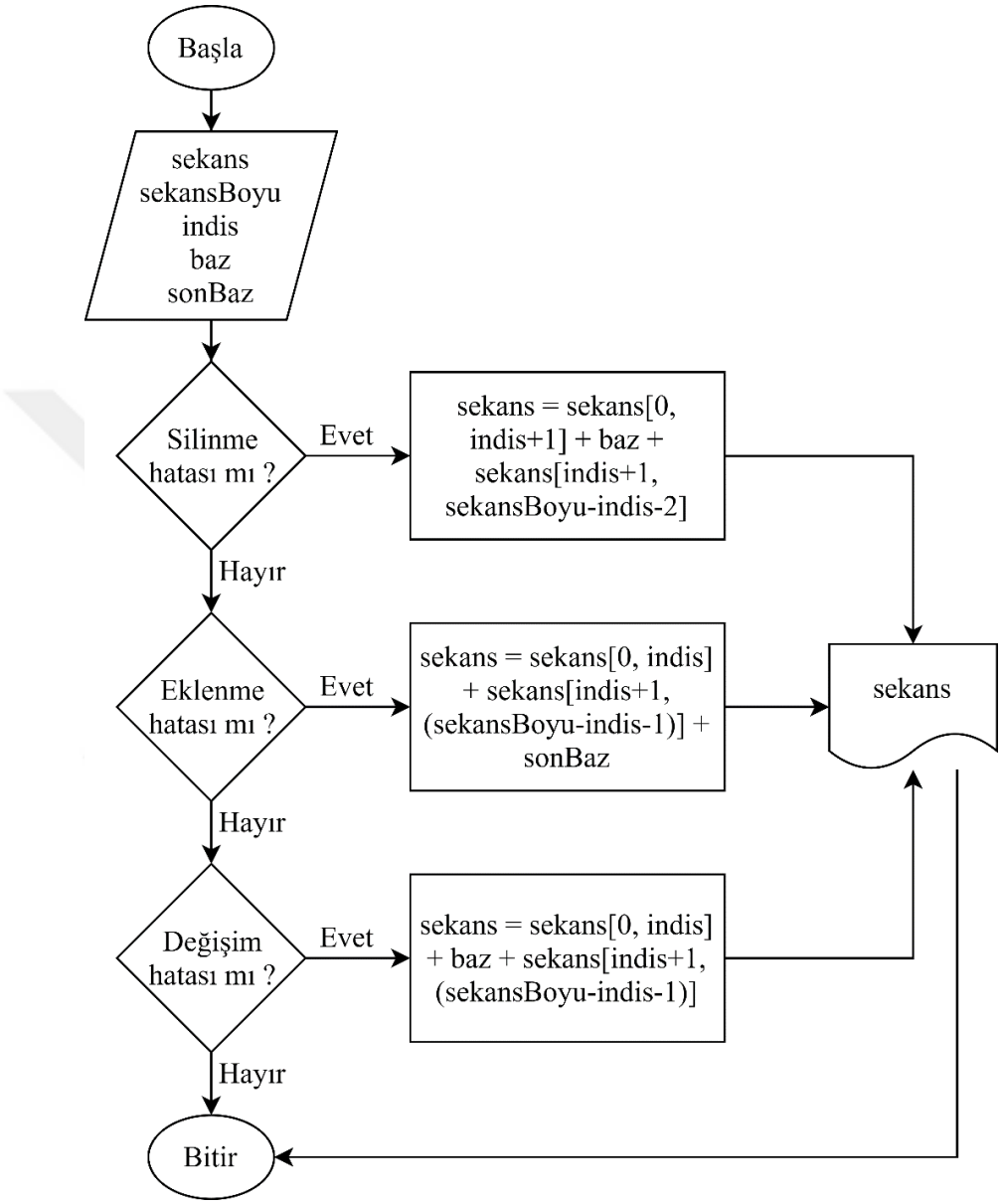


Şekil 2.13. Sağ yönlü karşılaştırma için geliştirilen hata düzeltme işlemi

Sol yönlü karşılaştırma işleminde hatanın düzeltilmesi:

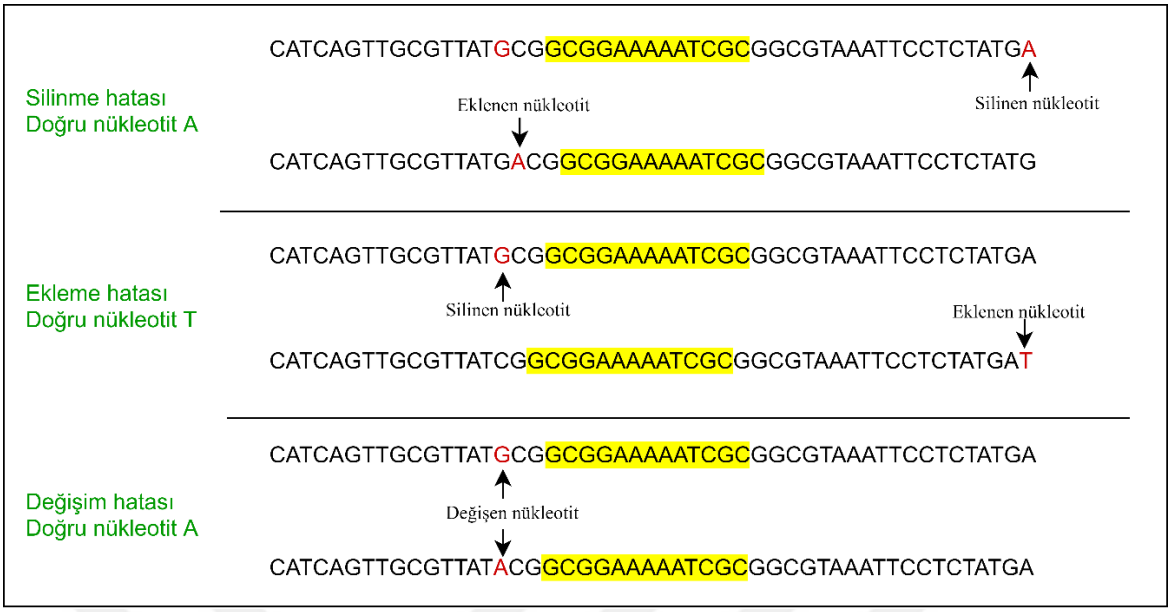
- Silinme hatası için doğru nükleotid, hatalı indisin bir sonraki (*i + 1*) indisine eklenmektedir. Eklenme işleminde sekans uzunluğundaki aşmayı engellemek için eklenen nükleotid kadar dizilim sonundan nükleotid silinmektedir.
- Eklenme hatası için hatalı indisteki nükleotid silinmektedir. Silinme işlemi ile kısalan sekans uzunluğunu gidermek için silinen nükleotid kadar son indis için global çoğunluk oylaması yapılarak bulunan nükleotid sekans sonunda eklenmektedir.

- Değişim hatası için doğru nükleotid, hatalı indisteki nükleotid ile değiştirilerek hata giderilmektedir.



Şekil 2.14. Sol yönlü karşılaştırma için geliştirilen hata düzeltme işlemi akış diagramı

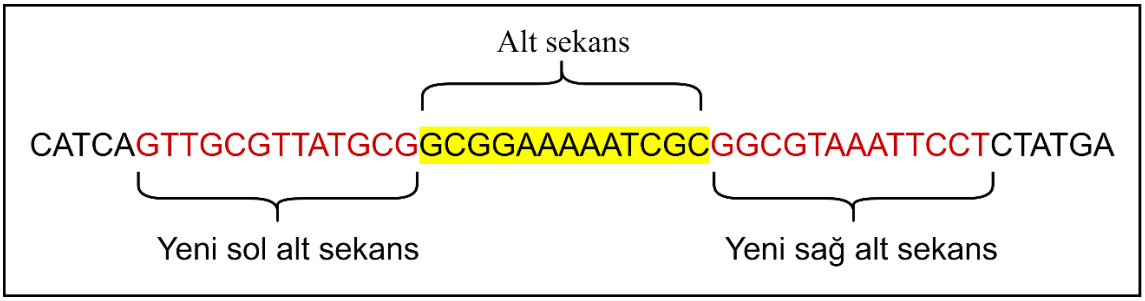
Sol yönlü karşılaştırma işleminde geliştirilen hata düzeltme işlemi akış diagramı Şekil 2.14'de gösterilmektedir. Sol yönlü karşılaştırma için geliştirilen hata düzeltme işlemi Şekil 2.15'te gösterilmektedir.



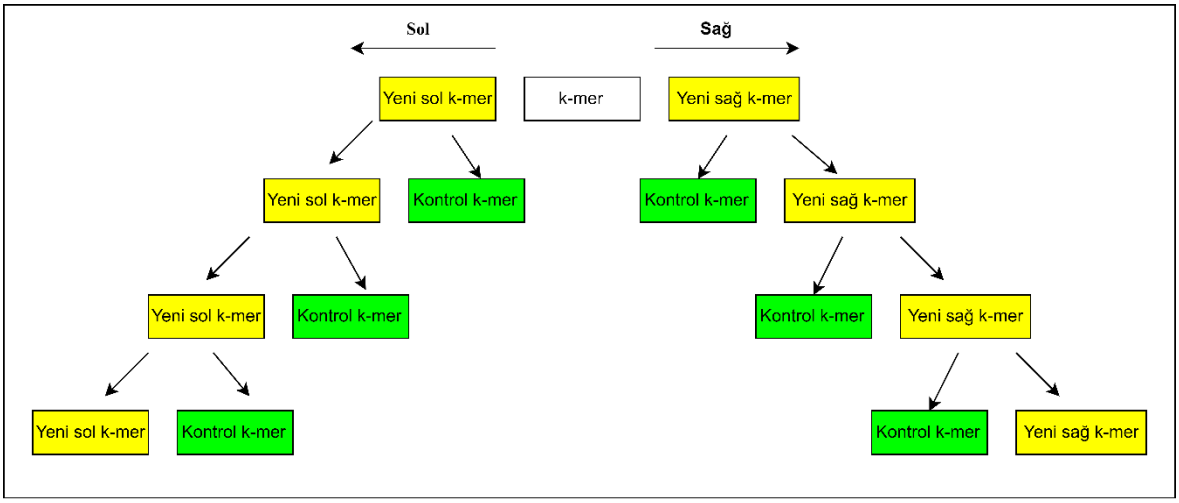
Şekil 2.15. Sol yönlü karşılaştırma için geliştirilen hata düzeltme işlemi

2.1.8. Yeni K-mer Sınıfının Belirlenmesi

Oluşturulan k-mer sınıfındaki tüm hatalı sekansların dizilimi düzeltildikten sonraki adım yeni k-mer boyu uzunluğunda alt-sekansın belirlenmesi ve yeni k-mer sınıfının oluşturulmasıdır. Geliştirilen yöntemde yeni k-mer sınıfının belirlenme işlemi Şekil 2.16'da gösterilmektedir.



Şekil 2.16. Yeni k-mer sınıfının belirlenme işlemi



Şekil 2.17. Yeni k-mer'le oluşan ağaç yapısı

Hata düzeltme işlemi bittiğinde k-mer uzunluklu alt sekansın sağ ve solundaki k-mer uzunluklu alt sekanslar yeni k-mer'ler olarak belirlenmektedir. Belirlenen k-mer'ler bütün sekanslarda aranarak sağ ve sol yönlü yeni k-mer sınıflıları oluşturulmaktadır. K-mer sınıfları oluşturulduktan sonra hatalı sekans tespit işlemi başlayarak süreç devam ettirilmektedir.

K-mer belirleme işleminde solun sol ve sağın sağ yönüne sürekli yeni k-mer'ler oluşur. Herhangi bir sekansta bir önceki k-mer dizilimi hatalı olup k-mer sekans sınıflandırılmasına dahil olamayan sekanslardaki hataları düzeltmek için sağın sağ ve solun sol yönüne kontrol k-mer'leri oluşturulmaktadır. Kontrol k-mer'leri sekanslarda aranır ve k-mer sekans sınıfı oluşturulur. K-mer'in sağ ve sol yönünde k-mer boyutu kadar alanı (Bu bölge bir önceki k-mer alanını ifade eder.) tarayıp hataları giderdiği için bir önceki k-mer'in hatalı alanı düzeltilir.

Yeni k-mer oluşturma işleminde oluşan k-mer'ler ağacın düğümlerini ifade ettiği için paralel çalışmaya uygun ağaç yapısı oluşmuş olur. Şekil 2.17'de yeni k-mer'lerin oluşturulması ile oluşan ağaç yapısı gösterilmektedir.

3. BULGULAR

Önerilen algoritma E. coli K12-MG1665 genomu referans alınarak farklı okuma uzunluğuna sahip, farklı kapsama değerlerinde ve farklı hata oranlı veri kümeleri kullanılarak test edilmiştir (Tablo 3.1). Algoritma performansı duyarlılık ve özgüllük verilerine göre değerlendirilmiştir (Tablo 3.2).

Tablo 3.1. Deneysel çalışmada kullanılan veri setleri ve özellikleri

Veri seti	**DNA Uzunluğu**	**Hata Oranı (%)**	**Sekans Uzunluğu**	**Kapsama**	**Okunan Sekans Sayısı**	**Okunan Nükleotid Sayısı**
D1	10200	1	50	20	4080	204000
D2	10200	1	50	50	10200	510000
D3	10200	1	50	75	15300	765000
D4	10200	1	75	20	2720	204000
D5	10200	1	75	50	6800	510000
D6	10200	1	75	75	10200	765000
D7	10200	1	100	20	2040	204000
D8	10200	1	100	50	5100	510000
D9	10200	1	100	75	7650	765000
D10	10200	2	50	20	4080	204000
D11	10200	2	50	50	10200	510000
D12	10200	2	50	75	15300	765000
D13	10200	2	75	20	2720	204000
D14	10200	2	75	50	6800	510000
D15	10200	2	75	75	765000	765000
D16	10200	2	100	20	2040	204000
D17	10200	2	100	50	5100	510000
D18	10200	2	100	75	7650	765000
D19	10200	3	50	20	4080	204000

Tablo 3.1'in devamı

Veri seti	DNA Uzunluğu	Hata Oranı (%)	Sekans Uzunluğu	Kapsama	Okunan Sekans Sayısı	Okunan Nükleotid Sayısı
D20	10200	3	50	50	10200	510000
D21	10200	3	50	75	15300	765000
D22	10200	3	75	20	2720	204000
D23	10200	3	75	50	6800	510000
D24	10200	3	75	75	10200	765000
D25	10200	3	100	20	2040	204000
D26	10200	3	100	50	5100	510000
D27	10200	3	100	75	7650	765000
D28	10200	4	50	20	4080	204000
D29	10200	4	50	50	10200	510000
D30	10200	4	50	75	15300	765000
D31	10200	4	75	20	2720	204000
D32	10200	4	75	50	6800	510000
D33	10200	4	75	75	10200	765000
D34	10200	4	100	20	2040	204000
D35	10200	4	100	50	5100	510000
D36	10200	4	100	75	7650	765000
D37	50000	1	50	20	20000	1000000
D38	50000	1	50	50	50000	2500000
D39	50000	1	50	75	75000	3750000
D40	50000	1	75	20	13333	1000000
D41	50000	1	75	50	33333	2500000
D42	50000	1	75	75	50000	3750000
D43	50000	1	100	20	10000	1000000
D44	50000	1	100	50	25000	2500000
D45	50000	1	100	75	37500	3750000
D46	50000	2	50	20	20000	1000000
D47	50000	2	50	50	50000	2500000

Tablo 3.1'in devamı

Veri seti	DNA Uzunluğu	Hata Oranı (%)	Sekans Uzunluğu	Kapsama	Okunan Sekans Sayısı	Okunan Nükleotid Sayısı
D48	50000	2	50	75	75000	3750000
D49	50000	2	75	20	13333	1000000
D50	50000	2	75	50	33333	2500000
D51	50000	2	75	75	50000	3750000
D52	50000	2	100	20	10000	1000000
D53	50000	2	100	50	25000	2500000
D54	50000	2	100	75	37500	3750000
D55	50000	3	50	20	20000	1000000
D56	50000	3	50	50	50000	2500000
D57	50000	3	50	75	75000	3750000
D58	50000	3	75	20	13333	1000000
D59	50000	3	75	50	33333	2500000
D60	50000	3	75	75	50000	3750000
D61	50000	3	100	20	10000	1000000
D62	50000	3	100	50	25000	2500000
D63	50000	3	100	75	37500	3750000
D64	50000	4	50	20	20000	1000000
D65	50000	4	50	50	50000	2500000
D66	50000	4	50	75	75000	3750000
D67	50000	4	75	20	13333	1000000
D68	50000	4	75	50	33333	2500000
D69	50000	4	75	75	50000	3750000
D70	50000	4	100	20	10000	1000000
D71	50000	4	100	50	25000	2500000
D72	50000	4	100	75	37500	3750000

Tablo 3.2. Duyarlılık ve özgüllük test sonuçları

Veriseti	Duyarlılık	Özgüllük	Veriseti	Duyarlılık	Özgüllük
D1	99,63	99,92	D37	97,75	99,89
D2	99,65	99,92	D38	97,77	99,92
D3	99,72	99,94	D39	99,78	99,92
D4	99,64	99,94	D40	97,77	99,90
D5	99,65	99,96	D41	97,81	99,93
D6	99,75	99,96	D42	97,88	99,97
D7	99,63	99,90	D43	97,52	99,82
D8	99, 85	99,95	D44	97,53	99,85
D9	99,70	99,98	D45	97,60	99,86
D10	98,30	99,6	D46	96,22	99,70
D11	98,75	99,88	D47	96,29	99,84
D12	99,90	99,92	D48	96,45	99,96
D13	99,00	99,96	D49	97,91	99,90
D14	99,27	99,90	D50	98,70	99,90
D15	99,41	99,95	D51	98,83	99,92
D16	98,98	99,49	D52	96,11	98,74
D17	99,00	99,57	D53	96,17	98,75
D18	99,13	99,63	D54	97,31	98,88
D19	98,22	99,78	D55	92,96	97,12
D20	98,51	99,79	D56	95,20	97,27
D21	98,80	99,82	D57	95,36	97,29
D22	98,75	98,81	D58	94,84	98,00
D23	98,77	98,83	D59	94,21	98,61
D24	98,91	98,84	D60	97,01	98,75
D25	98,12	99,01	D61	92,29	96,21
D26	98,18	99,09	D62	93,24	96,52
D27	98,24	99,15	D63	95,67	97,74
D28	94,26	96,57	D64	89,16	92,65
D29	97,13	99,14	D65	93,22	97,17
D30	98,52	99,68	D66	94,41	97,28

Tablo 3.2'nin devamı

Veriseti	Duyarlılık	Özgüllük	Veriseti	Duyarlılık	Özgüllük
D31	96,06	97,74	D67	91,60	95,00
D32	97,33	98,71	D68	97,03	98,14
D33	97,99	98,85	D69	97,24	98,08
D34	96,03	98,69	D70	93,12	93,96
D35	96,45	98,72	D71	94,11	95,32
D36	96,98	98,78	D72	95,03	96,63

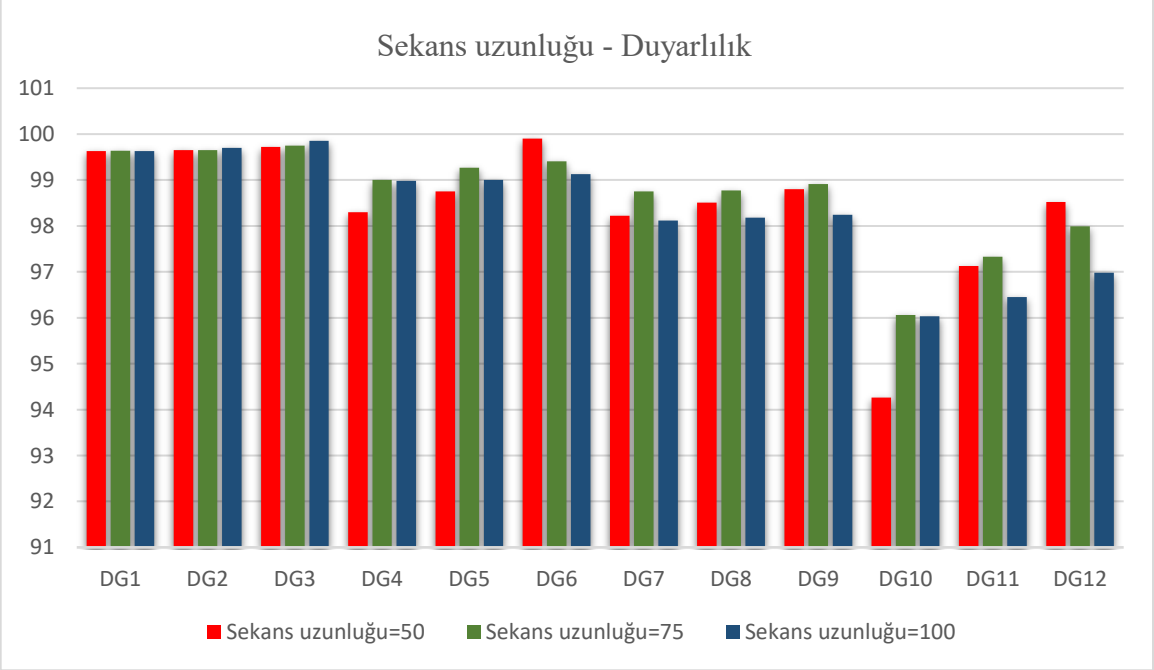
Önerilen algoritma Tablo 3.1'de özellikleri verilen veri setleri ile test edilip elde edilen dizilimlerin duyarlılık ve özgüllük değerleri hesaplanmıştır. Farklı sekans uzunluğu ile çalışıp diğer özellikleri aynı olan veri setleri gruplandırılmış ve sekans uzunluğunun duyarlılık ve özgüllük değeri üzerindeki etkisi irdelenmiştir. Tablo 3.3'te oluşturulan veri seti grupları gösterilmektedir.

Sekans uzunluğunun küçük olması çok fazla sekans arasında işlem yaptırıp işlem süreni uzatabileceği ve hata düzeltme performansını düşürebileceği ön görülmektedir. Büyük uzunluklu sekansların ise hatalı nükleotidin tespiti için fazla alt sekans-kmer üretimi gerektirdiği için işlem süresini uzatıp hatalı nükleotidin yakalanma oranını düşereceği gözlemlenmiştir. Bu sonuçlar doğrultusunda önerilen algoritma çalışma koşullarına göre test edilip hata düzeltme becerisinin en iyi olduğu sekans uzunluğu tespit edilmelidir.

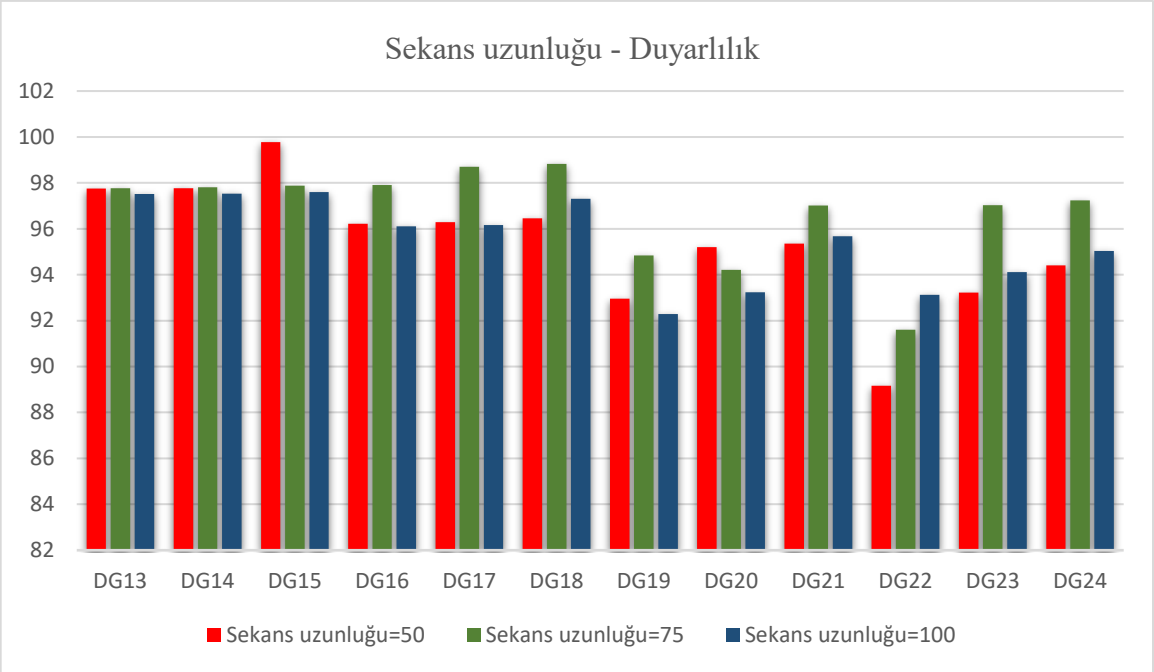
Şekil 3.1 'de 10200 uzunluklu DNA dizilimi üzerinde test işlemleri yapılmıştır. Bu test işlemleri ile belirtilen uzunluklu DNA dizilime işleminde sekans uzunluklarının duyarlılık üzerinde etkisi gözlemlenmektedir. Şekil incelendiğinde, yapılan deneylerde 50 ve 100 uzunluklu sekansların istikrarlı ve kesin yargıya varılacak sonuçları gözlemlenmese de 75 uzunluklu sekanslarla yapılan işlemlerin daha iyi sonuçlar elde ettiği gözlemlenmiştir. Şekil 3.2'de ise 50000 uzunluklu DNA dizilimi üzerinde test işlemleri yapılmıştır. Şekil incelendiğinde de yine 75 sekans uzunluklu yapılan testlerin duyarlılık değeri daha iyi olduğu gözlemlenmiştir.

Tablo 3.3. Önerilen algoritmanın sekans uzunluğu deneyi veri seti grupları

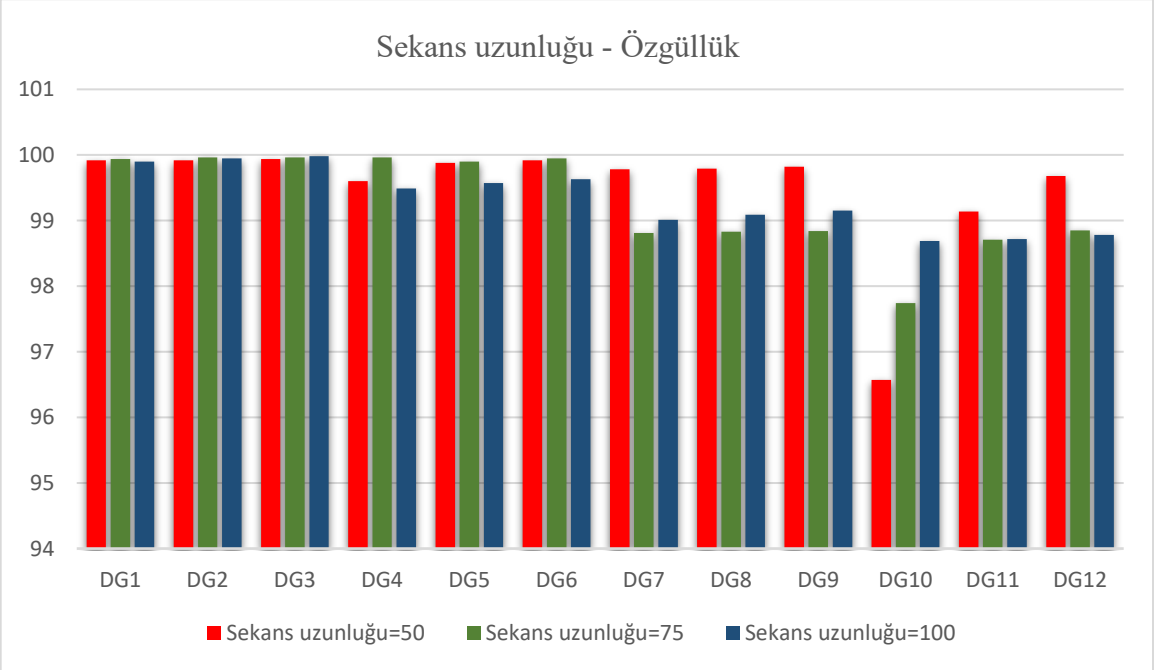
Deney grubu	Veri setleri	Deney grubu	Veri setleri
DG1	D1-D4-D7	DG13	D37-D40-D43
DG2	D2-D5-D8	DG14	D38-D41-D44
DG3	D3-D6-D9	DG15	D39-D42-D45
DG4	D10-D13-D16	DG16	D46-D49-D52
DG5	D11-D14-D17	DG17	D47-D50-D53
DG6	D12-D15-D18	DG18	D48-D51-D54
DG7	D19-D22-D25	DG19	D55-D58-D61
DG8	D20-D23-D26	DG20	D56-D59-D62
DG9	D21-D24-D27	DG21	D57-D60-D63
DG10	D28-D31-D34	DG22	D64-D67-D70
DG11	D29-D32-D35	DG23	D65-D68-D71
DG12	D30-D33-D36	DG24	D66-D69-D72



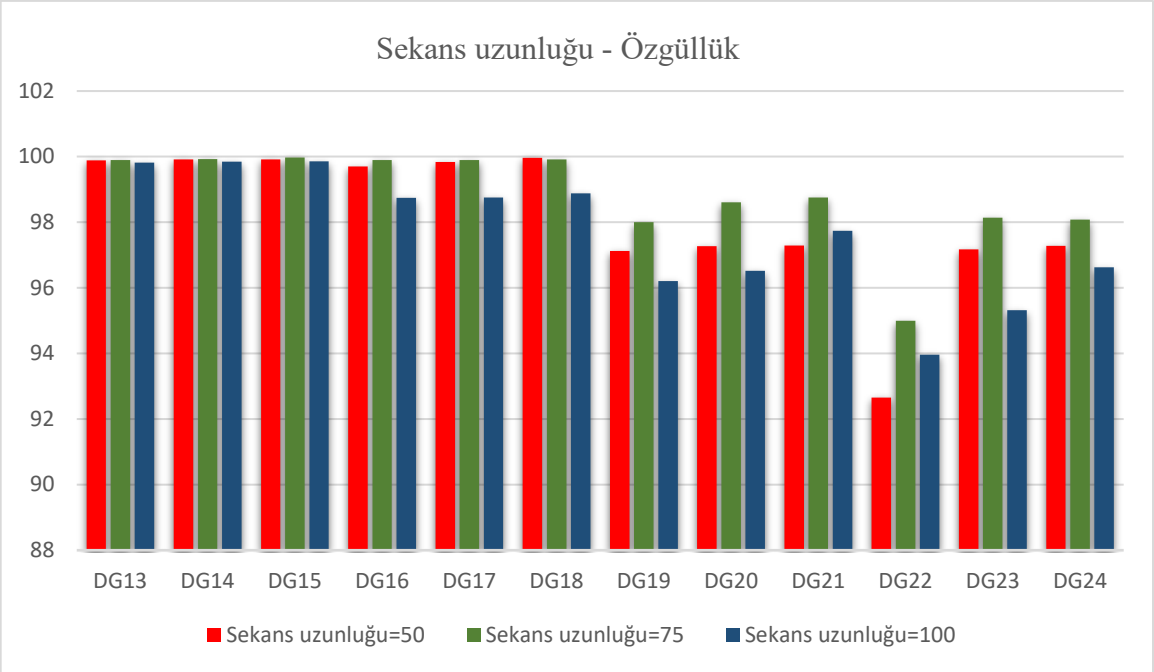
Şekil 3.1. Önerilen algoritmanın 10200 uzunluklu DNA üzerinde sekans uzunluğu – duyarlılık ilişkisi



Şekil 3.2. Önerilen algoritmanın 50000 uzunluklu DNA üzerinde sekans uzunluğu – duyarlılık ilişkisi



Şekil 3.3. Önerilen algoritmanın 10200 uzunluklu DNA üzerinde uzunluğu – özgüllük ilişkisi



Şekil 3.4. Önerilen algoritmanın 50000 uzunluklu DNA üzerinde uzunluğu – özgüllük ilişkisi

Şekil 3.3'te 10200 ve şekil 3.4'te 50000 uzunluklu DNA'lar, farklı uzunluklu sekanslarla dizilime çalışması yapılmış elde edilen dizilimin özgüllük sonuçları değerlendirilmiştir. Şekiller incelendiğinde sekans uzunluğu 75 olan deneylerin özgüllük değerleri 50 ve 100 uzunluklu sekanslarla yapılan testlerin özgüllük değerinden daha iyi sonuç vermiştir.

Şekil 3.1-3.4 incelendiğinde iki DNA uzunluğu üzerinde yapılan testlerin 75 uzunluklu sekanslarda daha iyi sonuçlar elde ettiği gözlemlenmiştir. Bu sonuçla önerilen algoritmanın 75 sekans uzunluğu ile daha istikrarlı ve iyi sonuçlar ürettiği sonucuna varılmaktadır.

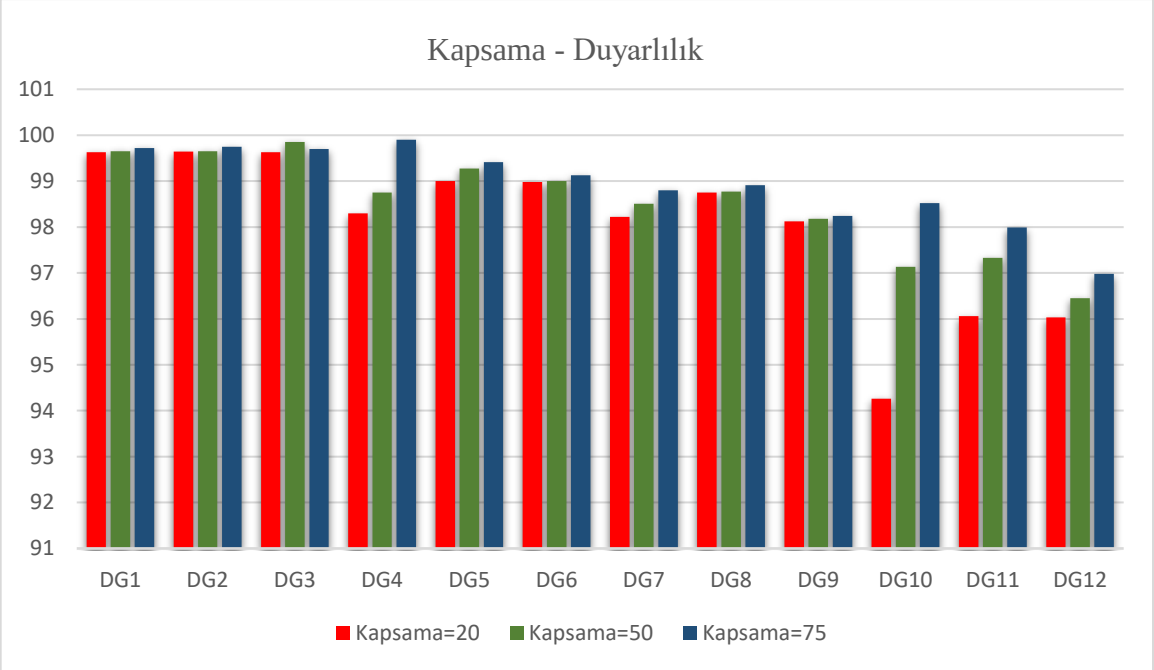
Kapsama değerinin duyarlılık ve özgüllük sonuçları üzerindeki etkisini gözlemlemek için tablo 3.2' deki deney sonuçları, farklı kapsama değeri içerip diğer özellikleri aynı olacak şekilde gruplandırılmıştır. Oluşan gruplarda tablo 3.4' te gösterilemektedir. Gruplandırılan test sonuçları şekil 3.5 – 3.8' de irdelenmiştir.

Şekil 3.5'te 10200 ve Şekil 3.6'da 50000 uzunluklu DNA dizilimli çalışmalarda kapsama değerinin artması ile elde edilen duyarlılık sonucundaki değişim gözlemlenmektedir. Şekil 3.7'de 10200 ve Şekil 3.8'de 50000 uzunluklu DNA parçalarının farklı kapsama değeri ile dizilenip elde edilen özgüllük değerleri gösterilmektedir. Şekil 3.5-3.8 incelendiğinde duyarlılık ve özgüllük değerlerinin kapsama değeriyle doğru orantılı

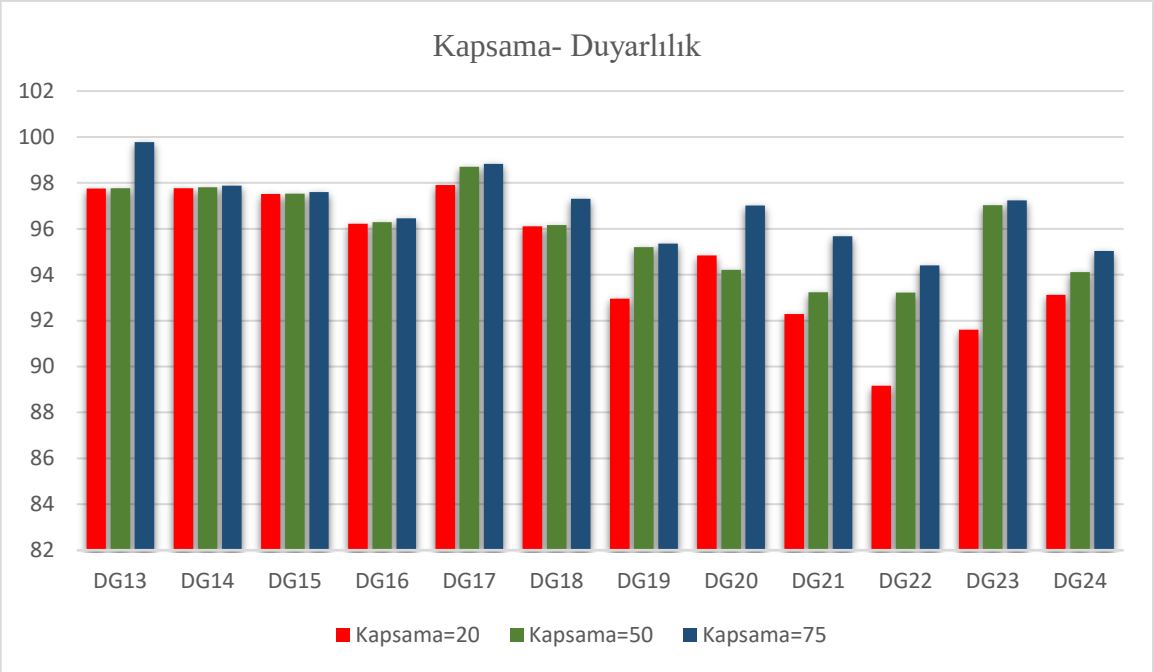
oldukları gözlemlenmektedir. Çünkü kapsama değeri fazla olan çalışmalarda sekanslar da doğru orantılı olarak fazla okunur. Bu şekilde bir bölge tekrarlı okunur. Tekrarlı okumalar ile hataların yakalanması ve düzeltilme becerisi artmaktadır.

Tablo 3.4. Önerilen algoritmanın kapsama değeri deneyi veri seti grupları

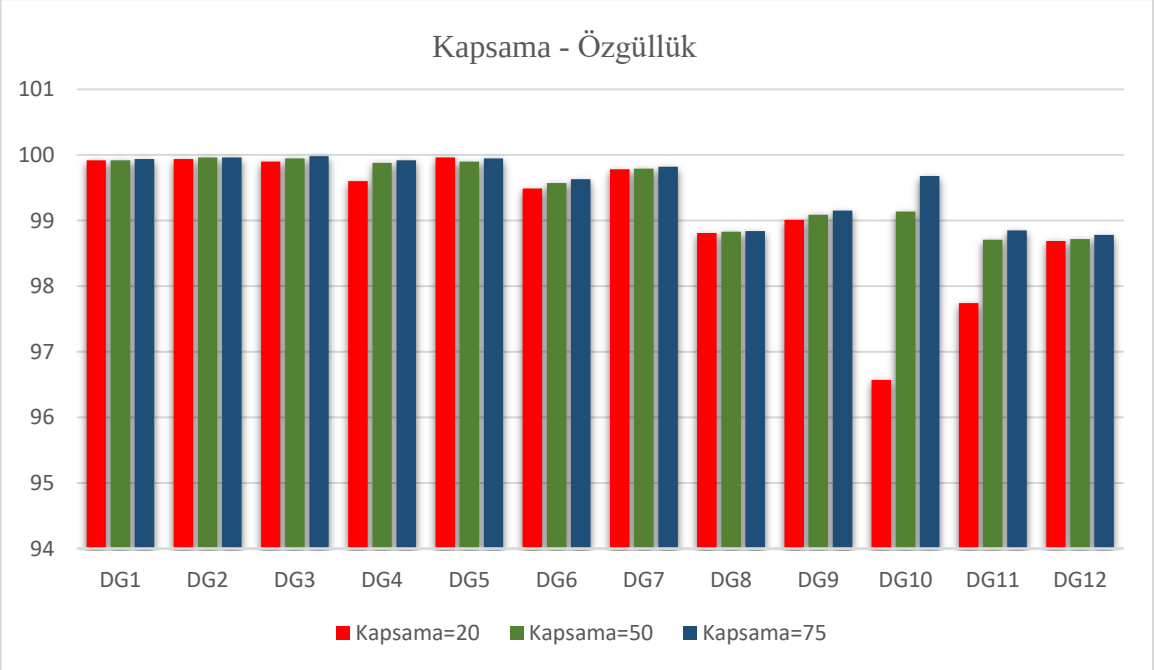
Deney grubu	**Veri setleri**	**Deney grubu**	**Veri setleri**	**Deney grubu**	**Veri setleri**
DG1	D1-D2-D3	DG9	D25-D26-D27	DG17	D49-D50-D51
DG2	D4-D5-D6	DG10	D28-D29-D30	DG18	D52-D53-D54
DG3	D7-D8-D9	DG11	D31-D32-D33	DG19	D55-D56-D57
DG4	D10-D11-D12	DG12	D34-D35-D36	DG20	D58-D59-D60
DG5	D13-D14-D15	DG13	D37-D38-D39	DG21	D61-D62-D63
DG6	D16-D17-D18	DG14	D40-D41-D42	DG22	D64-D65-D66
DG7	D19-D20-D21	DG15	D43-D42-D45	DG23	D67-D68-D69
DG8	D22-D23-D24	DG16	D46-D47-D48	DG24	D70-D71-D72



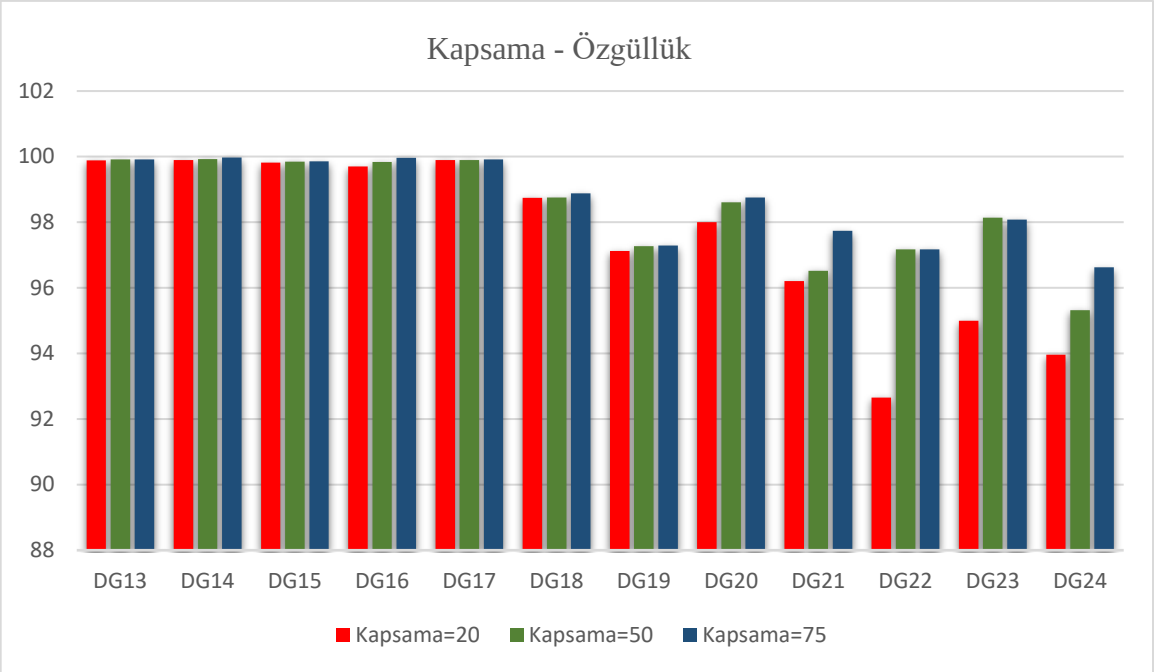
Şekil 3.5. Önerilen algoritmanın 10200 uzunluklu DNA üzerinde kapsama değeri – duyarlılık ilişkisi



Şekil 3.6. Önerilen algoritmanın 50000 uzunluklu DNA üzerinde kapsama değeri – duyarlılık ilişkisi



Şekil 3.7. Önerilen algoritmanın 10200 uzunluklu DNA üzerinde kapsama değeri – özgüllük ilişkisi



Şekil 3.8. Önerilen algoritmanın 50000 uzunluklu DNA üzerinde kapsama değeri – özgüllük ilişkisi

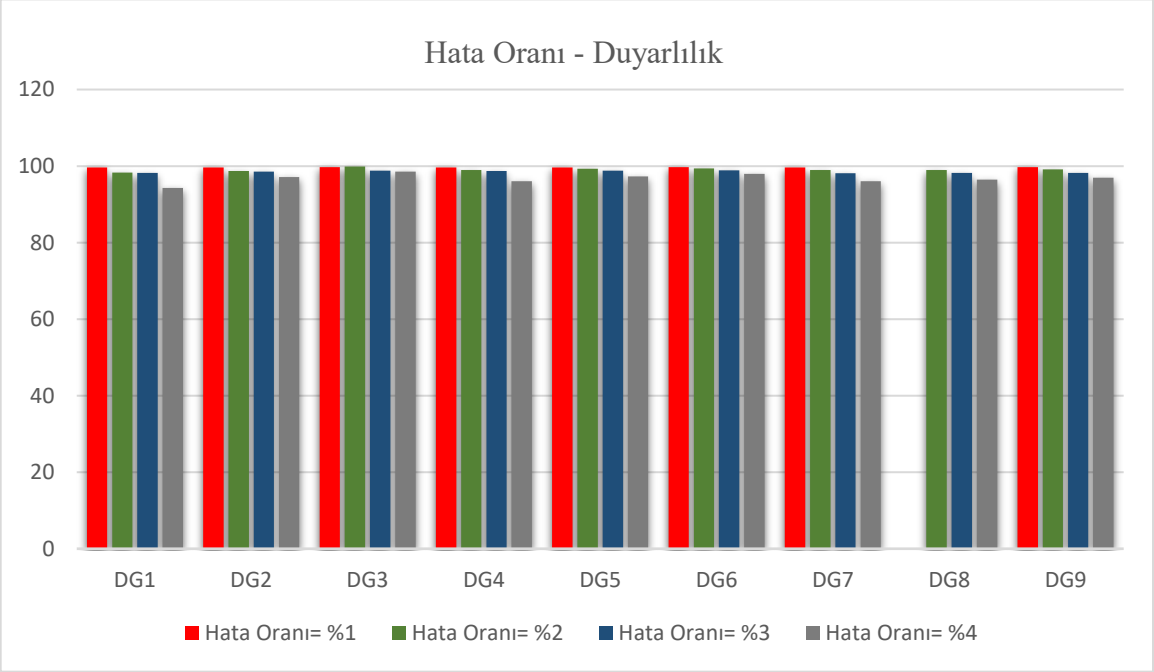
Gerçekleştirilen deney sonuçları ve sonuçlarla oluşturulan grafikler incelendiğinde önerilen algoritmanın, verilen veri setleri içinde en iyi duyarlılık ve özgüllük sonuçlarına 75 kapsama değeri ile eriştiği gözlemlenmiştir.

Hata oranının duyarlılık ve özgüllük değerleri üzerindeki etkisinin incelenmesi için Tablo 3.2'deki deney sonuçlarının hata oranları farklı diğer özellikleri aynı olacak şekilde gruplara ayrılmıştır. Oluşan veri seti grupları Tablo 3.5'te gösterilmektedir.

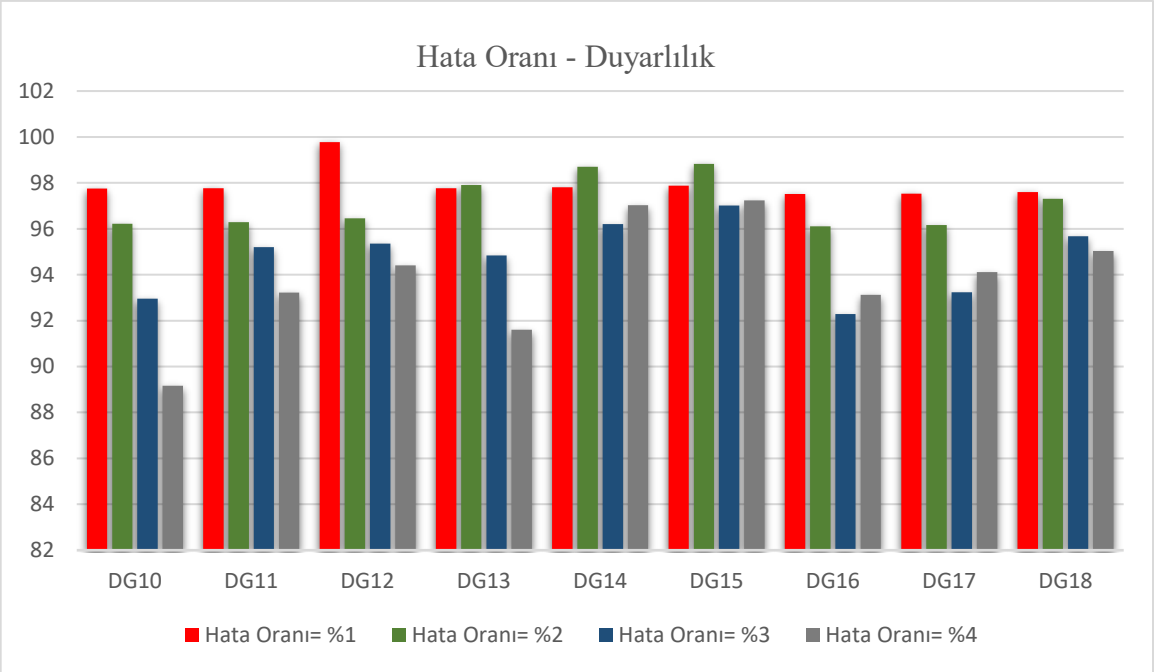
Hata oranının, duyarlılık sonuçları üzerindeki etkisi şekil 3.9 ve şekil 3.10'da, özgüllük değeri üzerinde etkisi Şekil 3.11 ve Şekil 3.12'de gözlemlenmiştir. Şekiller irdelendiğinde genel olarak hata oranı ile sonuçlar arasında doğru orantı olduğu tespit edilmiştir. Hata oranı yüksek olan yeni nesil dizileme cihazlarında sekanslar hata oranına göre belirlenen sayıda k-mer ile incelendiğinde hataların daha çok giderilmesi ön görülmektedir. Bu öneri tüm hata oranlı cihazlar için performans arttıracak olsa da lüzumsuz çalışmayı sağlayabileceği için üzerinde çalışılması gereken bir problem olarak görülmektedir. Bu probleme, yeni nesil cihazların hata oranlarının yaklaşık değeri bilindiği göz önünde bulundurularak cihazlara göre k-mer sayısı belirleme deneyleri yapılarak çözüm aranabilir.

Tablo 3.5. Önerilen algoritmanın hata oranı deneyi veri seti grupları

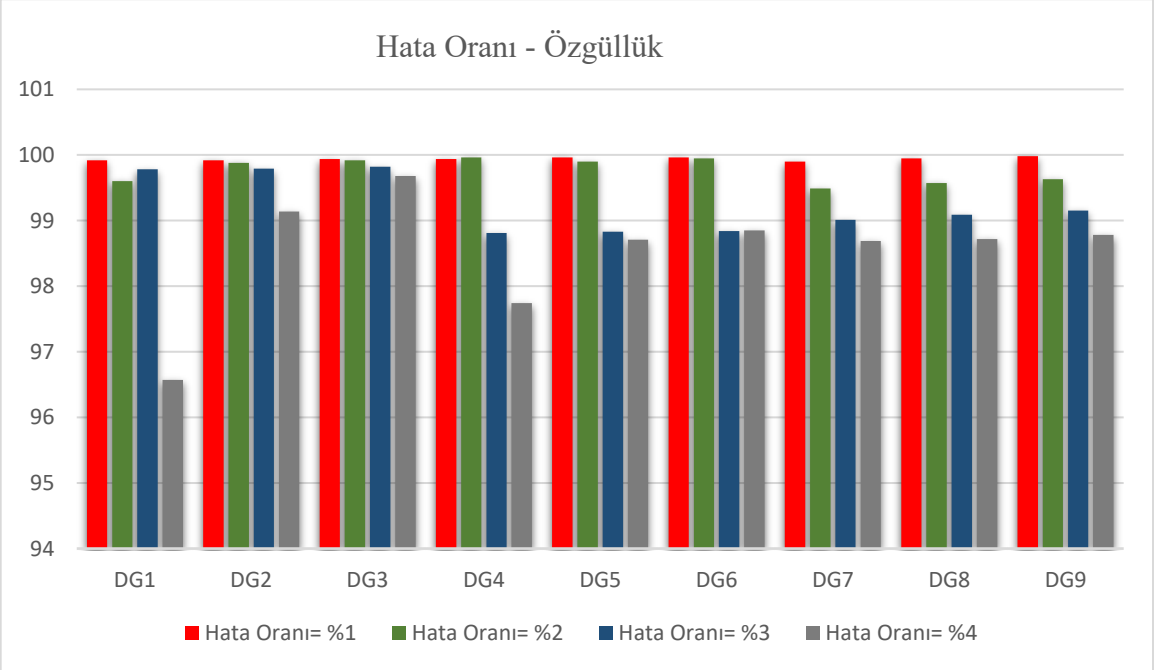
Deney grubu	Veri setleri	Deney grubu	Veri setleri
DG1	D1-D10-D19-D28	DG10	D37-D46-D55-D64
DG2	D2-D11-D20-D29	DG11	D38-D47-D56-D65
DG3	D3-D12-D21-D30	DG12	D39-D48-D57-D66
DG4	D4-D13-D22-D31	DG13	D40-D49-D58-D67
DG5	D5-D14-D23-D32	DG14	D41-D50-D59-D68
DG6	D6-D15-D24-D33	DG15	D42-D51-D60-D69
DG7	D7-D16-D25-D34	DG16	D43-D52-D61-D70
DG8	D8-D17-D26-D35	DG17	D44-D53-D62-D71
DG9	D9-D18-D27-D36	DG18	D45-D54-D63-D72



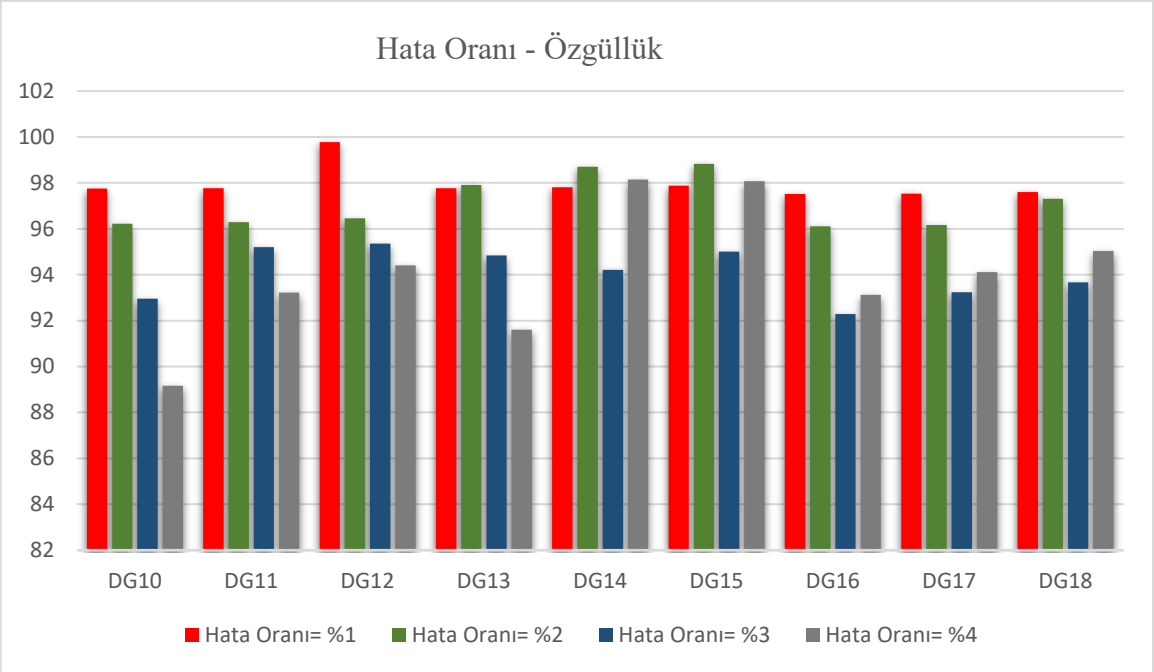
Şekil 3.9. Önerilen algoritmanın 10200 uzunluklu DNA üzerinde hata oranı – duyarlılık ilişkisi



Şekil 3.10. Önerilen algoritmanın 50000 uzunluklu DNA üzerinde hata oranı – duyarlılık ilişkisi



Şekil 3.11. Önerilen algoritmanın 10200 uzunluklu DNA üzerinde hata oranı – özgüllük ilişkisi



Şekil 3.12. Önerilen algoritmanın 50000 uzunluklu DNA üzerinde hata oranı – özgüllük ilişkisi

Hata oranının duyarlılık ve özgüllük sonuçları üzerindeki etkisinin incelendiği şekil 3.9, şekil 3.10, şekil 3.11 ve şekil 3.12 irdelendiğinde hata oranının %1 yükselmesi ile deney sonuçlarındaki düşüş aynı oranda olmadığı da gözlemlenmektedir. Özellikle önerilen algoritmanın hata oranının %3'e yükselmesiyle performansındaki ani düşüş göze çarpmaktadır. Bu düşüşü engellemek için oluşturulan k-mer'lerin sekans dizilimlerinin tamamını tarayıp taramadığı incelenebilir. Bu yöntemin yanı sıra çalışılan sekans uzunluğu değerlendirilebilir. Oluşturulan k-mer'lerin tüm sekansı dizilimini tarayacak şekilde sekans uzunluğu belirlenebilir. Bu tespitler gelecek çalışmalarda geliştirilip uygulanması, yeni nesil dizileme cihazlarındaki hataların giderilmesine büyük katkı sağlayacaktır.

5. TARTIŞMA

Önerilen algoritma performansı, farklı hata oranlı, sekans uzunluklu, kapsama değerli ve DNA uzunluklu parçalarla oluşturulan veri setleri üzerinde gerçekleştirilen deneysel çalışmalarla değerlendirilmiştir. Veri setleri literatürde yer alan Shrec ve DecGPU algoritmalarının makalelerinin deneysel sonuçlarında verilen veri setleri içerisinden seçilmiştir. Deneylerde kullanılan veri setleri Tablo 5.1'de gösterilmektedir.

Deneyler önerilen algoritmanın Shrec ve DecGPU algoritmaları ile ayrı ayrı karşılaştırılmasıyla oluşturulmuştur. Karşılaştırma işlemi aynı veri setiyle aynı hata oranı, sekans uzunluğu ve kapsama değeri ile gerçekleştirilmiştir. Literatürde yer alan DNA dizilimlerindeki hatalı bölgeleri düzeltmeye yönelik çalışan algoritmalar ve önerilen algoritma deney sonuçları Tablo 5.2'de gösterilmektedir.

Tablo 5.1. Önerilen algoritmanın literatürdeki algoritmalarla kıyaslanmasında kullanılan veri setleri

Deney No.	**Referans Genom**	**Erişim No.**	**Uzunluk**
Deney 1, 2	Saccharomyces cerevisiae	NC_001137	576.869
Deney 3,4	Escherichia coli	NC_000913	576.869

Önerilen algoritma yüzde 2 hata oranı ile 35 ve 70 kapsama değerleri ile ayrı ayrı çalıştırılarak Shrec algoritmasıyla karşılaştırılmıştır. Deney 1 ve 2 sonuçlarında da gösterildiği gibi önerilen algoritma genel olarak Shrec algoritmasına göre daha iyi sonuçlar üretmiştir. Fakat ikinci deneyde Shrec algoritması özgüllük sonuçları daha iyi çıkmıştır.

Deney 3 ve 4'te önerilen algoritma 30 ve 75 kapsama değeri ve yüzde 2 hata oranıyla 2 teste tabi tutularak sonuçlar elde edilmiştir. Elde edilen sonuçlar DecGPU algoritmasının sonuçları ile karşılaştırılmıştır. Bu deneylerle DecGPU algoritmasının özgüllük sonuçları önerilen algoritmaya göre deney 3'de çok düşük bir oranda deney 4'te ise eşit olarak tespit edilmiştir. Ancak DecGPU algoritmasının duyarlılık değeri iki deneyde de yüksek çıkmıştır.

Tablo 5.2. Önerilen algoritmanın literatürdeki algoritmalarla kıyaslanması

Deney No.	**Okuma Derinliği**	**Hata Oranı**	**Algoritma**	**Duyarlılık**	**Özgüllük**
Deney 1	35	2	Shrec	96,60	99,40
			Önerilen Algoritma	*97,00*	*99,62*
Deney 2	70	2	Shrec	97,00	100,00
			Önerilen Algoritma	*98,18*	*99,88*
Deney 3	30	3	DecGPU	99,99	48,81
			Önerilen Algoritma	*97,40*	*99,35*
Deney 4	75	3	DecGPU	99,97	99,89
			Önerilen Algoritma	*98,86*	*99,89*

Deneylerde kullanılan algoritmaların duyarlılık değerleri birlikte değerlendirilmiştir. Gerçekleştirilen deneylerde DecGPU algoritmasının duyarlılık değerinin yani hataları tespit edebilme yeteneğinin diğer algoritmalara göre üstünlüğü ve tutarlılığı tespit edilmiştir. Önerilen algoritmanın ise Shrec algoritmasına göre üstün olduğu böylece 2. en iyi sonuçları elde ettiği sonucuna ulaşılmıştır. Ayrıca genel olarak algoritmaların yüksek okuma derinliğinde daha iyi duyarlılık değeri elde ettiği de gözlemlenmiştir.

Deney sonuçlarından elde edilen özgüllük değerlerinin karşılaştırılması da gösterilmiştir. Özgüllük değeri ile algoritmaların doğru nükleotitleri tespit etme yeteneği değerlenderilmektedir. Sonuçlara bakıldığında Deney 2'de Shrec algoritmasının özgüllük değerinin yüzde yüze ulaştığı gözlemlenmiştir. Diğer deneylerde ise önerilen algoritmanın özgünlük değerinin üstünlük gösterdiği görülmektedir.

Deneyler sonucunda duyarlılık değeri diğer algoritmalara göre üstünlük gösteren DecGPU algoritmasının aynı performansı özgüllük değerinde göstermediği de gözlenmektedir.

Tablo 5.3. Önerilen algoritmanın en iyi sonuçları üreten parametreleriyle çalıştırılıp literatürdeki algoritmalarla kıyaslanması

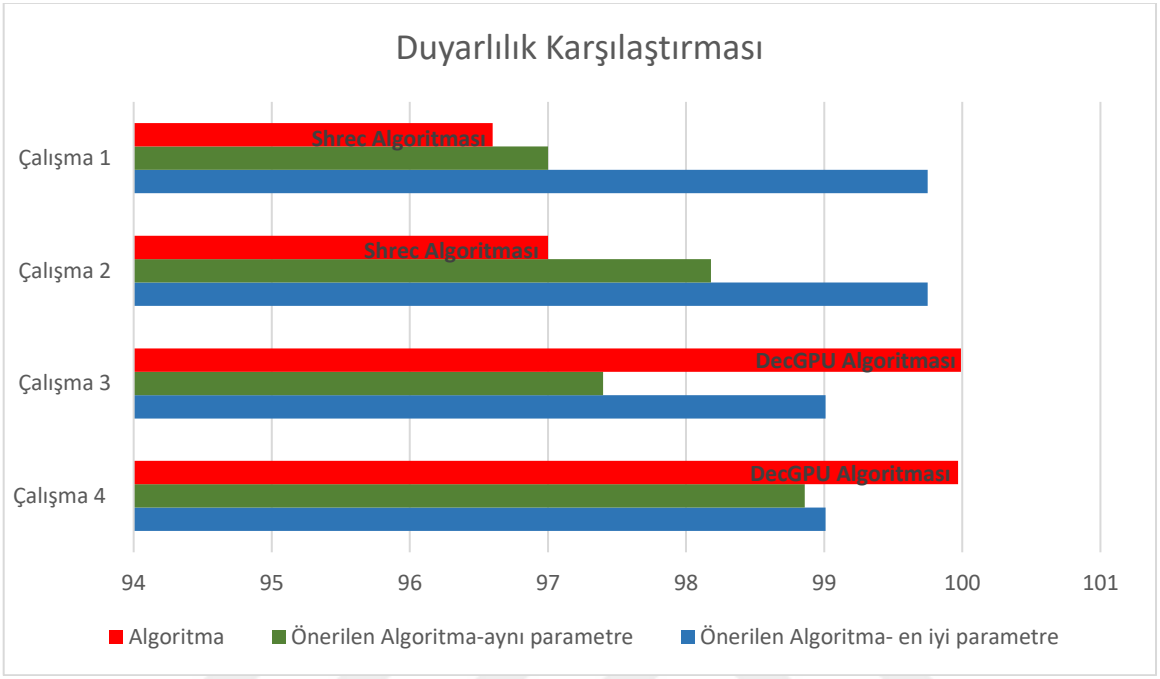
Deney No.	**Hata Okuma**	**Derinliği Oranı**	**Algoritma**	**Duyarlılık**	**Özgüllük**
Deney 1	2	35	Shrec	96,60	99,40
		75	*Önerilen Algoritma*	*99,75*	*99,98*
Deney 2	2	70	Shrec	97,00	100,00
		75	*Önerilen Algoritma*	*99,75*	*99,98*
Deney 3	3	30	DecGPU	99,99	48,81
		75	*Önerilen Algoritma*	*99,01*	*99,93*
Deney 4	3	75	DecGPU	99,97	99,89
		75	*Önerilen Algoritma*	*99,01*	*99,93*

Önerilen algoritma, bulgular kısmında yapılan deneylerle elde edilen, en iyi sonuçla çalıştığı 75 okuma derinlikli ve 75 sekans uzunluklu parametrelerle tablo 5.1'de gösterilen referans genomlar üzerinde çalıştırılmış elde edilen edilen sonuçlar aynı genomlarla kendi şartlarında çalıştırılan algoritmalarla karşılaştırılmıştır. Elde edilen sonuçlar tablo 5.3'de gösterilmektedir.

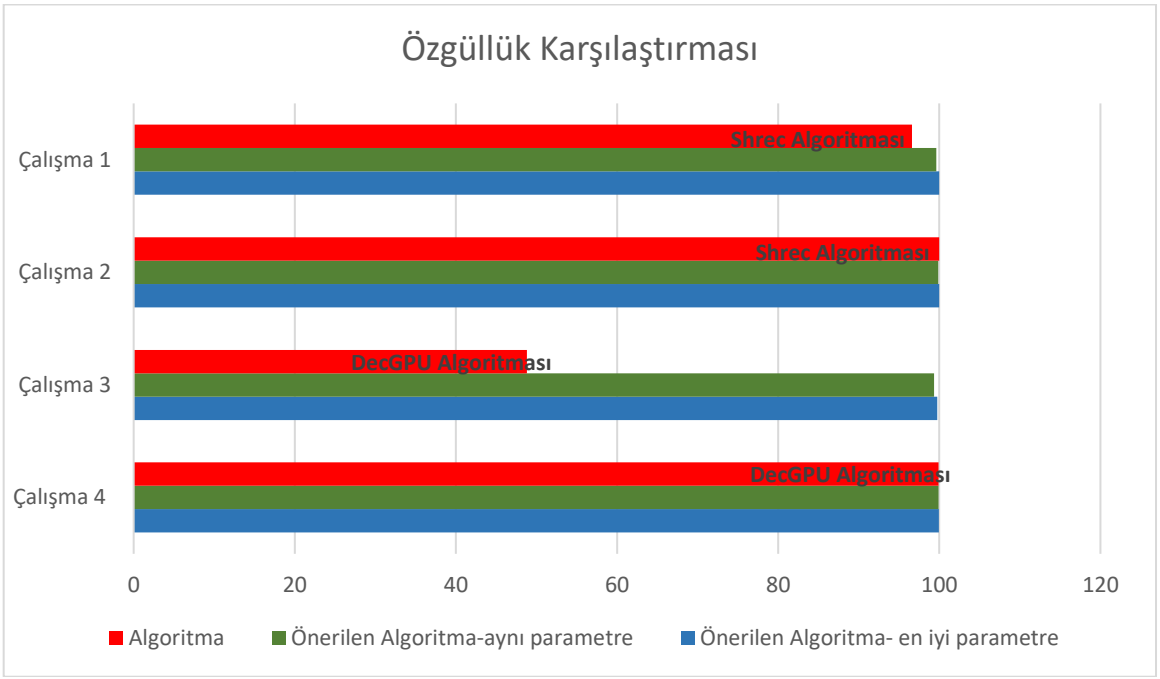
Tablo 5.2 ve tablo 5.3'deki sonuçlar algoritmalara göre gruplandırılarak şekil 5.1'de duyarlılık, şekil 5.2'de özgüllük sonuçları kıyaslanmıştır. Şekil 5.1 incelendiğinde önerilen algoritmanın her şartda shrec algoritmasından daha iyi duyarlılıkla çalıştığı gözlemlenmiştir. Özellikle en iyi parametre (kapsama ve sekans uzunluğu) ile çalıştırıldığında elde ettiği duyarlılık oranı daha da iyileşerek, shrec algoritmasına göre çok daha üstün duruma gelmiştir. Şekilde çalışma 3 ve 4 incelendiğinde ise önerilen algoritmanın en iyi parametrelerinde duyarlılık oranında yükselme olsa da decGPU algoritmasına göre tekrar geri kaldığı gözlemlenmiştir.

Şekil 5.2 incelendiğinde önerilen algoritmanın en iyi parametreleriyle çalıştırıltırıldığında, bir çalışmada shrec algoritmasından bir çalışmada da decGPU

algoritmasından daha iyi özgüllük değeri elde ederek, doğru nükleotidleri doğru tespit etme beceresinde diğer algoritmaların önüne geçtiği gözlemlenmiştir.



Şekil 5.1 Önerilen algoritmanın literatürdeki algoritmalarla duyarlılık değeri kıyaslaması



Şekil 5.2 Önerilen algoritmanın literatürdeki algoritmalarla özgüllük değeri kıyaslaması

6. SONUÇLAR

- Geleneksel dizileme yöntemleriyle küçük parçalarda doğruluk oranları yüksek DNA dizilimleri elde edilmektedir. Fakat bu yöntemler hız ve maliyet açısından iyi performans göstermemektedir. Canlılık faaliyetlerinin analizinde kullanılacak olan verilerin hızlı ve kolay elde edilebilir olması oldukça önem taşımaktadır. Bu önem doğrultusunda sürekli gelişen teknolojiye ayak uydurularak bu alanda da yeni teknolojik DNA dizileme cihazları üretilmeye başlanılmıştır. Bu cihazlarla geleneksel yöntemlerden çok daha hızlı ve düşük maliyetle DNA dizilimleri elde edilmektedir.
- Geliştirilen yeni nesil DNA dizileme cihazlarının avantajları yanında, geleneksel yöntemlerden daha fazla hatalı sonuçlar elde edilebilmektedir. Yüksek hız ve düşük maliyetten vazgeçmemek için araştırmacılar bu probleme çözüm bulabilmek için hata giderme algoritmaları geliştirmektedir. Bu çalışmada yeni nesil dizileme cihazlarının elde ettiği verideki hataları gidermek için yeni bir algoritma önerilmiştir.
- Bu çalışmada önerilen algoritma literatüre, k-mer'lerin sağ ve sol yönlerinde paralel arama tekniğini katması beklenmektedir. Ayrıca algoritma sağ ve sol yönde arama yaparken k-mer'ler ağaçtaki düğümü temsil etmektedir. Her k-mer'in bir önceki ve bir sonraki k-mer'i aranmaktadır.
- Karşılaştırma bölümünde gösterilen deney sonuçlarında ve bulgular bölümünde algoritma üzerinde yapılan testlerde görüldüğü üzere genel olarak yüksek okuma derinliğinde daha iyi duyarlılık değeri elde edildiği gözlemlenmiştir. Bu aşama da uygun okuma derinliği seçiminin önemi kanıtlanmıştır. Zaman kaybına ve gereksiz işlemlere sebebiyet vermeyecek şekilde yüksek duyarlılık elde edilebilecek okuma derinlikliğinin tespit edilmesi gerekmektedir.
- Algoritma üzerinde yapılan test sonuçlarına göre sekans uzunluğunun duyarlılık ve özgüllük testi üzerinde belirleyici olduğu ispat edilmiştir. Bu da algoritmalar uygun sekans uzunluklarının performansı maksimize edebilecek şekilde seçilmesi gerektiğini göstermektedir.
- Deney sonucunda duyarlılık - hatalı nükleotitlerin tespit edilme yeteneğinin - DecGPU algoritmasında en iyi olduğu ve önerilen algoritmanın ise onu takip ettiği

gözlemlenmiştir. Bu sonuçlar doğrultusudan önerilen algoritmanın duyarlılık performansının azımsanamayacak değerlere sahip olduğu gözlemlenmiş ve başarılı olduğu tespit edilmiştir.

- Önerilen algoritmanın başarısı özgüllük test sonuçlarında da devam etmiş sadece bir deneyde karşılaştırılan algoritmadan daha düşük değer elde etmiştir.
- Yapılan deneyler sonucunda önerilen algoritmanın tutarlı davrandığı, aynı algoritma ile farklı okuma derinliklikli karşılaştırıldığı deneylerde gözlemlenmektedir.
- Bu tez kapsamında yapılan çalışma ile yüksek doğruluklu DNA dizilimleri elde edilir. Elde edilen DNA dizilimi ile hizalama algoritmaları daha etkin bir şekilde kullanılarak yeni veya hastalıklı gen tespiti çalışmaları yüksek doğrululukla gerçekleştirilebilir.
- Çalışmadan elde edilen sonuçlar ilgili alanlarda yapılan deneylerle, DNA'da oluşan mutasyonların, genetik hastalıkların ve mikroorganizmaların tespiti, metogenom ve medikal çalışmalara da katkı sağlayacaktır.

7. ÖNERİLER

Önerilen algoritma sonuçları incelenmiş ve diğer algoritmalarla karşılaştırılarak anlamlı sonuçlar elde edildiği gözlemlenmiştir. Bundan sonraki çalışmalarda önerilen algoritmanın eksik yönleri giderilmeye devam edilecektir. Yapılması planlanan çalışmalar aşağıda maddeler halinde açıklanmaktadır.

- Hız problemine gidermek için hali hazırda paralel programlamaya uygun geliştirilen algoritma GPU üzerinde çalışabilecek şekilde geliştirilmesi planlanmaktadır.
- Daha uzun DNA dizilimleri üzerinde çalışmaların yapılması ve algortimanın bu dizilimlere uygun hale getirilmesi gerekmektedir.
- Sekanslarda düzeltilemeyen hataları gidermek ve büyük DNA dizilimini elde etmek için birleştirme işlemi yapılması gerekmektedir.

DNA dizilim hata düzeltme algoritmaları kapsamında yapılabilecek diğer çalışmalar:

- Uygun sekans uzunluğu tespiti
- Uygun kapsama değeri tespiti

Hata giderme algoritmaları ile elde edilen DNA dizilimi, gen tespiti, mutasyon tespiti ve metagenom araştırmaları gibi yeni çalışmalara olanak sağlayacağından bu alanlardaki problemlere çözüm üretilmesi literatüre büyük katkı sağlayacaktır. Bu problemlere hizalama algoritmaları geliştirilerek veya yeni bir algoritma keşfiyle çözüm aranabilir.

8. KAYNAKLAR

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

2. Dahm, R., Friedrich Miescher and the discovery of DNA, Developmental Biology, 27, 2 (2005) 274-288.

3. http://www.ilimrehberi.net/bilgi-bankas/147-m-n-harfi/937-mendelin-kaltim-calismalari.html. 1 Aralık 2018.

4. Cobb, Matthew., A Speculative History of DNA: What If Oswald Avery Had Died in 1934, PLoS Biology, 14, 12 (2016) e2001197.

5. Watson, James D. ve Francis HC Crick., Molecular structure of nucleic acids, Nature, 171, 4356 (1953) 737-738.

6. Zülal, A., İnsan Genomu, kalıtım şifresinin peşinde 136 yıl, Tübitak Yayınları (2001) 5-11.

7. Bolukbasi, E. ve Aras E. S., Third Generation DNA Sequencing Technologies, DNA, 1, 3 (2015).

8. Kızmaz, M. Z., Paylan, İ. C. ve Erkan S., DNA Dizilemenin Tarihsel Gelişimi, Gaziosmanpaşa Bilimsel Araştirma Dergisi, 6, 2 (2017) 47-53.

9. Tuğ, A., H. Hancı, ve A. Balseven, İnsan genom projesi: Umut mu, Kabus mu, Sürekli Tıp Eğitimi Dergisi, 11, 2 (2002) 56-57.

10. 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.

11. Sarıman, M., Ekmekci, S. S., Abacı, N., Çakiris, A., Paçal, F., Emrence, Z., Üstek, D. ve Öztürk Ş., Yeni Nesil Dizileme Teknolojisi ile Transkriptom Analizi, Deneysel Tıp Araştırma Enstitüsü Dergisi, 5, 10 (2015) 51-59.

12. Üstek D., Abacı N., Sırma S. ve Çakiris A., Yeni Nesil DNA Dizileme, Deneysel Tıp Araştırma Enstitüsü Dergisi, 1, 1 (2011) 11-18.

13. http://fenogretmeni.net/2015/11/15/dna-yapisi/. 1 Aralık 2018.

14. Clancy, S., Chemical structure of RNA, Nature Education, 1, 1 (2008) 223.

15. Shendure, J., ve Hanlee J., Next-generation DNA sequencing, Nature biotechnology, 26, 10 (2008) 1135.

16. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. ve Claude, T., Ten years of next-generation sequencing technology, Trends in genetics, 30, 9 (2014) 418-426.

17. Sanger, F. ve Alan, R. C., A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase, Journal of molecular biology, 94, 3 (1975) 441-448.

18. Sanger, F., Steven N. ve Alan R. C., DNA sequencing with chain-terminating inhibitors, Proceedings of the national academy of sciences, 74, 12 (1977) 5463-5467.

19. Metzker, M. L., Sequencing technologies - the next generation, Nature reviews genetics, 11, 1 (2010) 31.

20. https://www.mgbiyoinformatik.com/sanger-dizileme. 1 Aralık 2018.

21. http://www.onlinebiologynotes.com/sangers-method-gene-sequencing/. 1 Aralık 2018

22. Obenrader, S., The Sanger Method, Biology at Davidson (2003).

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

24. França, L. T. C., Emanuel C. ve Tarso B. L. K., A review of DNA sequencing techniques, Quarterly reviews of biophysics, 35, 2 (2002) 169-200.

25. Heather, J. M. ve Benjamin C., The sequence of sequencers: the history of sequencing DNA, Genomics, 107, 1 (2016) 1-8.

26. https://www.klimik.org.tr/wp-content/uploads/2017/05/Toplam-DNA-Dizisi-Saptama-Y%C3%B6ntemleri-ile-Mikobakterilerde-H%C4%B1zl%C4%B1-%C4%B0la%C3%A7-Direnci-Belirlenmesi-Tan%C4%B1l-Kocag%C3%B6z.pdf. 1 Aralık 2018.

27. Weber, J. L. ve Eugene W. M., Human whole-genome shotgun sequencing, Genome research, 7, 5 (1997) 401-409.

28. Anderson, S., Shotgun DNA sequencing using cloned DNase I-generated fragments, Nucleic acids research, 9, 13 (1981) 3015-3027.

29. Wilson, E., The sphinx in the city: Urban life, the control of disorder, and women, Univ of California Press, 1992.

30. Mullis, K. B., Erlich, H. A., Arnheim, N., Horn, G. T., Saiki, R. K. ve Scharf, S. J., Process for amplifying, detecting, and/or-cloning nucleic acid sequences, U.S. Patent No. 4,683,195, 1987.

31. http://biyokure.org/pcr-nedir-temel-basamaklari-ve-gereken-malzemeler-nelerdir/85/. 1 Aralık 2018.

32. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. ve Erlich, H. A., Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase, Science, 239, 4839 (1988) 487-491.

33. Hadidi, A. ve Candresse, T., Polymerase chain reaction, Viroids, (2003) 115-122.

34. http://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/27-dna-replication-transcri/pcr.html. 1 Aralık 2018.

35. Diehl, F. ve Smergeliene, E, BEAMing for Cancer: Detecting Tumor Mutations in Peripheral Blood Using Digital PCR, Genetic Engineering & Biotechnology News, 33, 15 (2013) 48-49.

36. Vacuik O., Analýza mitochondriální DNAz historického kosterního materiálu metodami sekvenace nové generace, Yüksek Lisans Tezi, Masarykova Univerzita, Brno, 2017.

37. 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.

38. Garrido-Cardenas, J., Garcia-Maroto, F., Alvarez-Bermejo, J. ve Manzano-Agugliaro, F., DNA sequencing sensors: an overview, Sensors, 17, 3 (2017) 588.

39. King, J. L., LaRue, B. L., Novroski, N. M., Stoljarova, M., Seo, S. B., Zeng, X., Warshauer, D. H., Davis, C. P., Parson, W. ve Sajantila, A., High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq, Forensic Science International: Genetics, 12 (2014) 128-135.

40. Rogers, Y. H. ve Venter, J. C., Genomics: massively parallel sequencing, Nature, 437, 7057 (2005) 326.

41. 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.

42. Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K. ve Wang, J., SNP detection for massively parallel whole-genome resequencing, Genome research (2009).

43. Pettersson, E., Joakim L. ve Afshin A., Generations of sequencing technologies, Genomics, 93, 2 (2009) 105-111.

44. Hall, N., Advanced sequencing technologies and their wider impact in microbiology, Journal of experimental biology, 210, 9 (2007) 1518-1525.

45. Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. ve Church, G. M., Accurate multiplex

polony sequencing of an evolved bacterial genome, Science, 309, 5741 (2005) 1728-1732.

46. https://evrimagaci.org/dna-dizileme-yontemleri-genleri-nasil-diziliyoruz-21. 1 Aralık 2018.

47. Sequencing, M. P., Protocol for Whole Genome Sequencing using Solexa Technology, BioTechniques (2006).

48. https://monsterbashseq.wordpress.com/2014/01/14/workshop-on-genomics-notes-day-2-sequencing-technologies/. 1 Aralık 2018.

49. Bentley, D. R., Whole-genome re-sequencing, Current opinion in genetics & development, 16, 6 (2006) 545-552.

50. Porreca, G. J., Genome sequencing on nanoballs, Nature biotechnology, 28, 1 (2010) 43.

51. Huang, J., Liang, X., Xuan, Y., Geng, C., Li, Y., Lu, H., Qu, S., Mei, X., Chen, H. ve Yu, T., A reference human genome dataset of the BGISEQ-500 sequencer, Gigascience, 6, 5 (2017) 1-9.

52. Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B. ve Yeung, G., Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays, Science, 327, 5961 (2010) 78-81.

53. Morey, M., Fernandez-Marmiesse, A., Castineiras, D., Fraga, J. M., Couce, M. L. ve Cocho, J. A., A glimpse into past, present, and future DNA sequencing, Molecular genetics and metabolism, 110, 1 (2013) 3-24.

54. http://www.completegenomics.com/documents/revolocity-tech-overview.pdf. 1 Aralık 2018.

55. Chrisey, L. A., Gil U. L. ve O'Ferrall, C. E., Covalent attachment of synthetic DNA to self-assembled monolayer films, Nucleic acids research, 24, 15 (1996) 3031-3039.

56. Sanger, F., Steven, N. ve Coulson, A. R., DNA sequencing with chain-terminating inhibitors, Proceedings of the national academy of sciences, 74, 12 (1977) 5463-5467.

57. Ronaghi, M., Pyrosequencing sheds light on DNA sequencing, Genome research, 11, 1 (2001) 3-11.

58. Wicker, T., Schlagenhauf, E., Graner, A., Close, T. J., Keller, B. ve Stein, N., 454 sequencing put to the test using the complex genome of barley, BMC genomics, 7, 1 (2006) 7-275.

59. Margulies M., Egholm M., Altman W. E., Attiya S., Bader J. S., Bemben L. A., Berka J., Braverman M. S., Chen Y. J., Chen Z., Dewell S. B., Du L., Fierro J. M., Gomes X. V., Godwin B. C., He W., Helgesen S., Ho C. H., Irzyk G. P., Jando S. C., Alenquer M. L. I., Jarvie T. P., Jirage K. B., Kim J. B., Knight J. R., Lanza J. R., Leamon J. H., Lefkowitz S. M., Lei M., Li J., Lohman K. L., Lu H., Makhijani V. B., McDade K. E., McKenna M. P., Myers E. W., Nickerson E., Nobile J. R., Plant R., Puc B. P., Ronan M. T., Roth G. T., Sarkis G. J., Simons J. F., Simpson J. W., Srinivasan M., Tartaro K. R., Tomasz A., Vogt K. A., Volkmer G. A., Wang S. H., Wang Y., Weiner M. P., Yu P., Begley R. F. ve Rothberg J.M., Genome sequencing in microfabricated high-density picolitre reactors, Nature, 437, 7057 (2005) 376-380.

60. Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H. ve Sidow, A., A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning, Genome research, (2008) gr-076463.

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

62. https://yasambilgibilimi.wordpress.com. 1 Aralık 2018.

63. Buermans, H. P. J. ve Den Dunnen, J. T., Next generation sequencing technology: advances and applications, Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1842, 10 (2014) 1932-1941.

64. Akyol, İ., Yıldız, M. A. ve Tutar E., Yeni Nesil Nükleotid Dizileme Metotlarının Biyokimyasal Temelleri, Kahramanmaraş Sütçü İmam Üniversitesi Doğa Bilimleri Dergisi, 20, 1 (2017) 1-15.

65. Rusk, N., Torrents of sequence, Nature Methods, 8, 1 (2010) 44.

66. http://www.ankaramikrobiyoloji.org.tr/docs/8kongre/Baris_Otlu.pdf. 1 Aralık 2018.

67. Pennisi, E., Semiconductors inspire new sequencing technologies, Science, 327, 5970 (2010) 1190-1190.

68. Purushothaman, S., Toumazou, C. ve Ou, C. P., Protons and single nucleotide polymorphism detection: A simple use for the Ion Sensitive Field Effect Transistor, Sensors and Actuators B: Chemical, 114, 2 (2006) 964-968.

69. Molnar, M. Z., Error Correction in Next Generation DNA Sequencing Data, Yüksek Lisans Tezi, University of Western Ontario, Londra, 2012.

70. Chaisson, M., Pavel P. ve Tang H., Fragment assembly with short reads, Bioinformatics, 20, 13 (2004) 2067-2074.

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

72. Salmela, L. ve Schröder, J., Correcting errors in short reads by multiple alignments, Bioinformatics, 27, 11 (2011) 1455-1461.

73. Ilie, L., Fazayeli, F. ve Ilie, S., HiTEC: accurate error correction in high-throughput sequencing data, Bioinformatics, 27, 3 (2010) 295-302.

74. Schröder, J., Schröder, H., Puglisi, S. J., Sinha, R. ve Schmidt, B., SHREC: a short-read error correction method, Bioinformatics, 25, 17 (2009) 2157-2163.

75. http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/presentations/shrec.pdf. 1 Aralık 2018.

76. Liu, Y., Schmidt, B. ve Maskell, D. L., DecGPU: distributed error correction on massively parallel graphics processing units using CUDA and MPI, BMC bioinformatics, 12, 1 (2011) 85.

77. Gomaa, S., Belal, N. A. ve El-Sonbaty, Y., H-RACER: Hybrid RACER to Correct Substitution, Insertion, and Deletion Errors, International Conference on Bioinformatics and Biomedical Engineering, Nisan 2017, Springer Cham, 62-73.

78. Saha, S. ve Rajasekaran, S., EC: an efficient error correction algorithm for short reads, BMC bioinformatics, 16, 17 (2015) S2.

79. 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.

80. https://hackernoon.com/probabilistic-data-structures-bloom-filter-5374112a7832. 1 Aralık 2018.

81. Song, H., Dharmapurikar, S., Turner, J., ve Lockwood, J., Fast hash table lookup using extended bloom filter: an aid to network processing, Acm Sigcomm Computer Communication Review, 35, 4 (2005) 181-192.

82. Geravand, S. ve Ahmadi, M., Bloom filter applications in network security: A state-of-the-art survey, Computer Networks, 57, 18 (2013) 4047-4064.

83. Kılıç, İ. Ve Karcı, A., DNA Dizilerinin De Bruijn Grafları ile İncelenmesi, Elektrik - Elektronik ve Bilgisayar Mühendisliği Sempozyumu, Aralık 2017, Eleco.

84. https://tezverianaliz.com/biyoistatistik-dershanesi/ozgulluk-ve-duyarlilik-specificity-and-sensitivity-1/. 1 Aralık 2018.

85. https://www.ncbi.nlm.nih.gov/. 1 Aralık 2018.

ÖZGEÇMİŞ

Elif ARAS, 1991 Trabzon doğumludur. İlköğretim eğitimini Özel Alparslan İlköğretim Okulu'nda ve lise eğitimini aynı okulun fen lisesinde burslu tamamlamıştır. 2009 güz döneminde Gazi Üniversitesi Bilgisayar Mühendisliği Bölümü'nde başladığı lisans eğitiminden 2014 yılı bahar döneminde mezun olmuştur. 2014 - 2016 yıllarında Mutlusoft Bilişim Teknolojileri Tic. A.Ş.'de yazılım geliştirici olarak çalışmıştır. 2016 – 2017 yıllarında Canbakkal İnşaat Turizm Madencilik Nakliye Hafriyat Ticaret Ltd.'de web geliştirici olarak çalışmıştır. 2015 yılının güz döneminde Karadeniz Teknik Üniversitesi; Bilgisayar Mühendisliği Bölümü'nde yüksek lisans eğitimine başlamıştır. 2017 yılının Eylül ayından itibaren Avrasya Üniversitesi; Bilgisayar Programcılığı Bölümü'nde öğretim görevlisi olarak görev yapmaktadır.