ANKARA ÜNİVERSİTESİ
BİYOTEKNOLOJİ ENSTİTÜSÜ

BİYOİNFORMATİK
YÜKSEK LİSANS TEZİ

DE NOVO TÜM GENOM SEKANS VERİSİ ÜZERİNDE GİZLİ MARKOV MODELİ TABANLI GEN BULMA YÖNTEMLERİNİN UYGULANMASI

Zeynep ÖZKESERLİ

Danışman Öğretim Üyesi
Prof. Dr. H. Gökhan İLK

MAYIS
2014

ETİK BEYAN

Bu tez çalışmasının; akademik kural ve etik ilkelere bağlı kalınarak hazırlandığını, çalışmada yararlanılan ve bu çalışma ürünü olmayan bütün bilgiler için kaynak yayınlara atıfta bulunulmuş olduğunu beyan ederim.

Zeynep ÖZKESERLİ

İmza

ÖZET

Yüksek Lisans Tezi

de novo Tüm Genom Sekans Verisi Üzerinde Gizli Markov Modeli Tabanlı Gen Bulma Yöntemlerinin Uygulanması

Zeynep ÖZKESERLİ

Ankara Üniversitesi Biyoteknoloji Enstitüsü

Danışman: Prof. Dr. H. Gökhan İLK

Nükleotid dizileme teknolojilerindeki gelişime paralel olarak artan genom projesi sayısı, yeni nesil dizileme teknolojilerinin geliştirilmesiyle daha da ivmelenmiştir. Gelişmiş laboratuvar teknikleri sayesinde elde edilen nükleotid dizisi verisinin yine ıslak laboratuvar çalışmaları ile anlamlandırılması çok maliyetlidir. Bu öngörüyle, seksenli yılların başından beri nükleotid dizileri üzerinde hesaplamsal yöntemler ile gen bulmaya yönelik çalışmalar gerçekleştirilmektedir. Bu yöntemlerden gizli Markov modeli tabanlı olanlar, teknik özellikleri ile ön plana çıkmaktadır. Tez çalışması kapsamında, Ankara Üniversitesi Biyoteknoloji Enstitüsü'nde BOREN projesi kapsamında yeni nesil sekanslama yöntemi ile gerçekleştirilen *Bacillus boroniphilus* genom projesinde elde edilen genom verisi üzerinde, çeşitli gen bulma algoritmaları ve anotasyon akış hatları kullanılarak yapılan gen bulma işlemine ilişkin sonuçlar değerlendirilmiştir. Genom projelerinin dinamik yapısı ve karmaşıklık düzeyi göz önünde bulundurulduğunda, birden fazla yöntemin birbiriyle uyumlu olarak kullanılması gerektiği, ancak gizli Markov modeli tabanlı çekirdek algoritma ve bu algoritmayı kullanan anotasyon akış hattının anotasyonlarının daha güvenilir olacağı gözlenmiştir.

(2014, 85 sayfa)

Anahtar Kelimeler: Gen bulma, *de novo* dizileme, genom anotasyonu, GeneMark, PGAAP, RAST, Glimmer

ABSTRACT

MSc. Thesis

Application of Hidden Markov Model Based Gene Finding Methods on *de novo* Whole Genome Sequence Data

Zeynep ÖZKESERLİ

Ankara University Biotechnology Institute

Supervisor: Prof. H. Gökhan İLK

Advances in nucleotide sequencing technology has led to faster sequencing of the genomes. Thanks to this advent, even small sized centers can now perform genome projects. Besides the data generation step, one of the key steps in genome sequencing projects is the genome annotation step, which is consisted of "gene finding" and adding attributes to the sequence parts which are most probably coding regions. The need for computational gene finding algorithms is known since 80's, and many gene finding algorithms are developed. In this thesis study, two annotation pipelines (PGAAP, RAST) and their core algorithms which are based on Markov models (GeneMarkS, Glimmer) are used to perform gene finding on *de novo* genomic sequence of *Bacillus boroniphilus,* and the results are compared. Because of the dynamic structure and the complexity of genome projects, it can be said that using various methods can be beneficial for obtaining different kinds of information at the post annotation step. It can be said that the method which uses hidden Markov model (GeneMarkS) and the pipeline uses this core algorithm (PGAAP) provided more reliable results for *B. boroniphilus* genome.

(2014, 85 pages)

Keywords: Gene finding, *de novo* genome sequencing, genome annotation, GeneMark, PGAAP, RAST, Glimmer

TEŞEKKÜR

2005 yılında İstatistik Bölümü'ne "Bir gün, nasıl olacağı hakkında bir fikrim olmasa da, moleküler biyoloji üzerinde çalışacağım" hayali ile başladığımda, bir genom projesinin ilk veri analizinin bana emanet edilebileceği aklımın ucundan geçmezdi. Bu onuru yaşamamda en büyük teşekkürü sanırım elimden tutan ilk büyüğüm olan değerli hocam Prof. Dr. Fikri Öztürk'e (Ankara Üni. Fen Fak. İstatistik Böl.) borçluyum. Seneler önce kendisine sorduğum "Hocam, bu genom hangi bazdan başlayarak proteinleri kodluyor, yani bir üçlü var ama, hangi üçlü, genomun başı neresi, sonu neresi, nereden biliyoruz?" sorusunun yanıtını bir tez çalışması kapsamında öğreneceğimi hiç düşünmemiştim.

Diğer büyük teşekkürü de beni henüz İstatistik Bölümü'nde öğrenciyken laboratuvarına kabul eden, Türkiye'nin ilk genom projelerinden birinde çalışmama olanak sağlayan ve değerli ekibiyle çalışma fırsatı sunan değerli hocam Prof. Dr. Hilal Özdağ'a (Ankara Üni. Biyoteknoloji Ens.) ve değerli danışman hocam Prof. Dr. H. Gökhan İlk'e (Ankara Üni. Elektrik Elektronik Müh. Böl.) borçluyum. Eminim kendileri de bu kadar uzun süre o laboratuvardan çıkmayacağımı tahmin etmemişti. Bu uzun fakat "eğitici" laboratuvar sürecinde yanımda olan, her birinden moleküler biyoloji, hayat ve *projeler* konusunda çok şey öğrendiğim değerli dostlarım Dr. Dilay Çığlıdağ Düngül, Dr. Nevin Belder, Dr. Nilgün Tekin, Dr. Günseli Çubukçuoğlu Deniz, Doç. Dr. Serkan Durdu, Dr. Aynur Karadağ, Uzm. Hülya Sümer Çelebi, Uzm. Semih Dalkılıç, Uzm. Seda Taşır Yılmaz, Uzm. Özge Cumaoğulları, Uzm. Melike Özbilgin Öztürk ve Uzm. Devrim Aydın'a yürekten teşekkür ederim. Makale sunumları, proje değerlendirmeleri ve tez süreci zorlukları ile mücadele konusunda sizden öğrendiklerime paha biçilemez.

Apayrı bir teşekkür bölümü de, bu tek hücrelinin dev genomu ile mücadelede bizzat yanımda olan ve haftalar süren zorlu genom dizileme deneyinin çok değerli sonuçlarını benimle paylaşan, bir büyük kızkardeş kadar çok sevdiğim değerli Hocam Yrd. Doç. Dr. Yeşim Doğan için. Birleştirme heyecanı ve anotasyon incelemelerinde yanımda olduğu, heyecan ve zorluk anlarını deneyimi ve güven verici varlığıyla benimle paylaştığı ve tez yazımında destek ve eleştirileri ile her daim yanımda olduğu için sonsuz teşekkürler.

Genom üzerinde farklı gen bulma araçlarının uygulanması ile elde edilen sonuçların karşılaştırılmasında tüm teknik bilgisi ve desteğiyle yanımda olan Yrd. Doç. Dr. E. Doruk Engin'e (Ankara Üni. Biyoteknoloji Ens.) tez çalışmasındaki ihtiyaca yönelik olarak geliştirdiği *Overlap* yazılımı ve sonuçların değerlendirilmesindeki desteği için çok teşekkür ederim. *Overlap* sonuçlarının özetlenmesindeki yardımları için yazılımcı dostum Serkan Eren'e ayrıca teşekkürler.

Bambaşka bir teşekkür bölümü de "tezi ne zaman bitiriyorsun", "biz yaşlanmadan biter mi" sorularıyla ve "bir an önce bitir" stratejisiyle sürekli yanımda olan, beni destekleyen değerli ailem ve değerli dostlarım için. Maddi, manevi destekleri için onlara nasıl teşekkür edeceğimi bilemiyorum. Eğitim hayatımı sürdürebilmem için hiçbir desteklerini esirgemeyen annem Nurhayat Özkeserli, babam İsmail Özkeserli ve kızkardeşim Uzm. Pınar Özkeserli'ye, gece – gündüz yanımda olan ve hiçbir zaman, hiçbir şartta vazgeçmeme izin vermeyen Uzm. Tuba Demirtaş ve Zeynep Şenkesen'e, benimle evini paylaşan sevgili anneannem Fatma Kanık ve teyzem Uzm. Gülten Kanık'a sonsuz teşekkürler. Son dönemde kötü günleri iyi günlere dönüştürme çabamda en büyük destekçim olan değerli dostum Ece Tathan'a da ayrıca teşekkürler.

Son teşekkürüm de bu tezi gelecekte okuyacaklarını ve faydalanacaklarını umduğum arkadaşlarıma. Sorularınız olduğu takdirde bana yazarsanız çok sevineceğimi belirtmeliyim.

İÇİNDEKİLER

1. GİRİŞ 1
2. KURAMSAL TEMELLER 4
2.1. Genom Projeleri 4
2.2. Dizileme 6
2.2.1. Shotgun Dizileme ve Bu Yöntemle Elde Edilen Verinin Yapısı 10
2.2.2. Paired-end Dizileme 12
2.3. Gen Bulma 16
2.3.1. Protein Kodlayan Bölgeler ve Moleküler Biyolojinin Santral Dogması 17
2.3.2. Hesaplamsal Gen Bulucuların Geliştirilmesi 18
2.4. Genom Anotasyonu 19
2.4.1. Prokaryotik Genomların Özellikleri ve Prokaryotik Genomlarda Gen Bulma 20
2.5. Gizli Markov Modelleri ve Prokaryotlarda Gen Bulma 28
2.5.1. Gizli Markov Modelleri 29
2.5.2. GeneMark Algoritma Serisi 36
2.5.3. GLIMMER 38
2.5.4. Anotasyon Akış Hatları 39
3. GEREKÇE ve AMAÇ 41
4. MATERYAL VE YÖNTEM 43
4.1. Materyal - Genom Dizileme Verisi 43
4.2. Yöntem 47
4.2.1. Gen Koordinatlarının Elde Edilmesinde Kullanılan Yöntemler 49
4.2.2. Gen Koordinatlarının Karşılaştırılmasında Kullanılan Yöntemler 53
4.2.3. Gerçekleştirilen Karşılaştırmalar 54
4.2.4. Gen Bulma Sonuçlarının ve Gerçekleştirilen Karşılaştırmaların Değerlendirilmesinde Kullanılan Yöntemler 55
5. ARAŞTIRMA BULGULARI 57
5.1. Gen Bulma İşlemi ile Elde Edilen Sonuçlar 60
5.2. Yöntemler Arası Karşılaştırma Bulguları 62
5.2.1. İkili Karşılaştırmalar 64
5.2.2. Genom Üzerindeki Kodlamayan RNA Genleri 67
6. TARTIŞMA ve SONUÇ 70
KAYNAKLAR 77

ŞEKİLLER DİZİNİ

Şekil 2.2 Shotgun dizileme sürecinin şematik gösterimi 8
Şekil 2.3 Yıllara göre dizileme işlem hacmi (2)(7) 9
Şekil 2.5 Paired-end dizileme 13
Şekil 2.9 Shotgun ve paired-end okumaların birleştirilmesi 16
Şekil 2.10 Moleküler biyolojinin santral dogması 17
Şekil 2.11 RAST ile anotasyonu gerçekleştirilmiş bir nükleotid dizisi 20
Şekil 2.12 Gen bulma işlemi için geliştirilen örüntü tanıma algoritmalarında kullanılan ölçüler 26
Şekil 2.13 Gen bulma işlemi için geliştirilen örüntü tanıma algoritmalarında kullanılan ölçülerin birbirleriyle ilişkileri 27
Şekil 2.16 DNA dizisinde kullanılabilecek örnek birinci derece homojen Markov modeli 33
Şekil 2.17 5' ayıklama bölgesi tanıma problemi 34
Şekil 2.18 GeneMarkS algoritmasının çalışma prensibi 38
Şekil 2.19 PGAAP ve RAST anotasyon akış hatları ile gen bulma süreci 40
Şekil 4.1 Dizileme işleminin en genel hatlarıyla şematik gösterimi 43
Şekil 4.4 Anotasyon akışhatları ve algoritmaların süreçteki işlevleri 50
Şekil 4.5 Süperkontig01'de bulunmuş bir gen için örnek PGAAP çıktısı 51
Şekil 4.6 Süperkontig01'de bulunmuş bir gen için örnek RAST çıktısı 51
Şekil 5.2 Yöntemlere göre gen uzunluğu dağılımlarını gösteren histogramlar 1. PGAAP, 2. RAST, 3. GMSC+GMHMM, 4. Glimmer 62
Şekil 5.3 Yöntemlerin bulduğu gen sayısına göre ikili karşılaştırma sonuçları 65
Şekil 5.4 SK01, SK02 ve SK12 üzerinde gerçekleştirilen tüm karşılaştırmalarda ortak ve farklı bulunan genler 66
Şekil 5.5 Başlangıç kodonu aynı olan, bitiş kodonu aynı olan ya da birebir aynı olan gen sayısına göre ikili karşılaştırma sonuçlar 67

ÇİZELGELER DİZİNİ

Çizelge 2.1 Organizmalara göre genom projeleri 6
Çizelge 2.2 Bazı gen kestirim yazılımları 23
Çizelge 4.1 Dizileme ile elde edilen veri 44
Çizelge 4.2 Dizileme sonucu elde edilen kontiglere ilişkin istatistikler 45
Çizelge 4.3 Süperkontig uzunluk ve kontig içerikleri 46
Çizelge 5.1 Her bir süperkontig için dizi özellikleri ve bulunan gen sayıları 60
Çizelge 5.2 Süperkontiglere göre yöntemlerin buldukları gen sayılarına göre sıralanması 61
Çizelge 5.3 Yöntemlere göre bulunan genlere ilişkin betimsel istatistikler 61
Çizelge 5.4 PGAAP ve RAST ile bulunan rRNA'lar 69
Çizelge 5.5 PGAAP, RAST ve tRNAScan-SE ile bulunan tRNA'lar 69

1. GİRİŞ

Bir canlının tüm genom dizisi, o canlının DNA'sını meydana getiren tüm nükleotidlerin dizisidir. DNA dizileme işlemi ile canlının genomunda yer alan nükleotidlerin sırasının belirlenmesiyle molekül üzerinde yer alan birincil bilgiyi taşıyan "nükleotid dizisi" elde edilmektedir. Nükleotid dizisinin eldesiyle, genom üzerindeki kodlayan bölgeler, kodlamayan bölgeler, düzenleyici bölgeler gibi yaşamla ilgili tüm genomik birimlerin, yaklaşık olarak, "kaynak koduna" erişilmektedir. İşlem bu yönüyle yaşam bilimleri alanında gerçekleştirilen araştırmaların önemli bir parçası olma özelliğini taşımaktadır (15). Eğer dizileme işlemi tüm genom üzerinde yapılacak olursa, elde edilen veri araştırmacıya canlının biyolojik özelliklerinin genel bir tablosunu sunmaktadır. Ancak bu tabloya ulaşılabilmesi için elde edilen nükleotid dizisinin anote edilmesi gerekmektedir. Bu da, genel anlamda, genom üzerindeki fonksiyonel bölgelerin konumlarının belirlenmesi ve ardından bu birimlerin fonksiyonlarına göre adlandırması işleminin gerçekleştirilmesi anlamına gelmektedir.

Günümüzde özellikle biyolojik özellikleriyle dikkat çeken canlıların genomların dizilenmesi (patojenler, ekstremofiller vb.), dünyanın her yerinde pek çok araştırma birimince sürdürülmektedir. Gelişen teknoloji ile özellikle prokaryotik genomların kısa sürede ve düşük maliyetlerle dizilenmesi mümkün olmaktadır. Bunun gibi organizmanın tüm genom dizisini elde etmeye ve elde edilen genom üzerindeki genomik birimlerin konumlarının belirlenmesine yönelik çok aşamalı araştırmalara genom projesi adı verilmektedir.

Genom projeleri genel anlamda *de novo* (baştan) ve yeniden dizileme projeleri olarak sınıflandırılabilir. Bu iki proje türünü birbirinden ayıran en önemli özellik, elde edilen verinin niteliğine bağlı olarak birleştirme ve gen bulma işlemlerinde izlenecek yöntemlerin farklılığıdır.

De novo tüm genom dizileme, daha önce genom dizisi belirlenmemiş bir türün genomunun dizilenmesidir. Yeniden tüm genom dizileme ise, daha önce genom dizisi belirlenmiş bir türe ait başka bir canlının genomunun dizilenmesidir. Örneğin insan genomu daha önce dizilenmiş olduğundan, bundan böyle gerçekleştirilecek her insan genomu dizileme işlemi bir yeniden dizilemedir. Benzer şekilde daha önce genom dizisi belirlenmiş bir bakterinin yeni bir suşunun dizilenmesi de bir yeniden dizilemedir. Fakat, bu tez çalışmasına konu olan *Bacillus boroniphilus* gibi, yeni bir türün tüm genom dizisi, bir dizi *de novo* genom dizileme işleminin gerçekleştirilmesiyle elde edilmektedir.

Genom projelerinde dizileme işleminin gerçekleştirilmesi kadar önemli olan diğer aşama, elde edilen genom dizisi üzerindeki genomik bileşenlerin konumlarının kestirilmesi ve bu bölgelerin işlevlerinin belirlenmesidir. Dizi verisinin biyolojik bilgiye dönüştürüldüğü bu aşamanın gerçekleştirilmesinde pek çok yöntem kullanılmaktadır. Bu yöntemler genel olarak genom içerik istatistikleri ve homoloji bilgisine dayanmaktadır. Deneysel verilerle elde edilmiş olan kodlayan ve kodlamayan bölgelere ait içerik istatistikleri aracılığıyla geliştirilen modelleri temel alan algoritmalar kullanılarak, elde edilen genom dizisi üzerindeki genlerin koordinatları elde edilebilmektedir. Bu işlemin ardından kodlayıcı olabileceği belirlenen bu bölgeler homoloji bilgisi temel alınarak işlevlerini ifade eden belirteçler ile isimlendirilmektedir.

Bu tez çalışması kapsamında, Ankara Üniversitesi Biyoteknoloji Enstitüsü Genombilim Birimi ve Muğla Üniversitesi Moleküler Biyoloji Bölümü işbirliği ve BOREN desteği ile gerçekleştirilen *Bacillus boroniphilus* Genom Projesi kapsamında elde edilen *de novo* tüm genom verisi kullanılmıştır. Elde edilen taslak genom üzerinde gizli Markov modeli tabanlı gen bulma algoritmaları kullanılarak gen koordinatlarının kestirimi yapılmış ve bu işlemden elde edilen sonuçlar değerlendirilmiştir.

Tez içeriğinde kuramsal temeller bölümünde kısaca genom projelerinden bahsedilmiş, ardından genom projelerinin ana basamakları olan dizileme ve prokaryotik genomların özellikleri ve bu genomlarda gen bulma çerçevesinde genom anotasyonu konularına değinilmiş, genom anotasyonunda kullanılan algoritmalar genel hatlarıyla ele alınmış ve bu algoritmaların kullanıldığı daha karmaşık yöntemler olan anotasyon akış hatlarının işleyişi anlatılmıştır.

Materyal bölümünde *B. boroniphilus* genomunun dizilenmesiyle elde edilen taslak genomun özellikleri verilmiş, yöntem bölmünde ise gen bulmada kullanılan (ve daha sonra sonuçları karşılaştırılan) yöntemlerin hangi parametrelerle çalıştırıldığı belirtilmiş, ardından yöntemlerin hangilerinin ne amaçla karşılaştırıldığı anlatılmış, karşılaştırmaların neye göre yapıldığı ve nasıl değerlendirildiği gösterilmiştir.

Araştırma bulgularında, öncelikle gen bulma işlemi ile elde edilen sonuçlara değinilmiş, ardından gerçekleştirilen yöntemler arası karşılaştırmaların sonuçları verilmiştir.

2. KURAMSAL TEMELLER

2.1. Genom Projeleri

Canlılığın anlaşılmasında hücrenin yönetim merkezini oluşturan organizmanın genomunun deşifre edilmesi gerekmektedir. Bu gereklilik kalıtım biyomolekülü olan DNA'nın çift sarmallı yapısının çözüldüğü 1950'li yıllardan itibaren ortaya konmuştur. Genom projelerine giden süreçte, canlıların genomlarına küçük parçaların (örneğin genlerin) dizilenmesi başarılmış, ardından dizileme işlem hacminin büyümesiyle daha kısa sürede daha çok nükleotid dizisi elde edilebilmeye başlanmıştır. Ancak bir organizmanın genomunun deşifre edilmesi işlemi yalnızca genomu oluşturan nükleotid diziliminin ortaya çıkarılmasıyla tamamlanmamaktadır. Dizileme işleminin yanısıra elde edilen dizinin biyolojik bağlamda anlamlandırılması da gerekmektedir (33).

Bir nükleotid dizisinin biyolojik bağlamda anlamlandırılabilmesinde ilk aşama, üzerindeki kodlayan bölgelerin konumlarının kestirilmesidir. Bu işlem gerçekleştirildikten sonra, bu konumlardaki nükleotid parçaları, kodladıkları proteinin işlevini ifade eden belirteçlerle isimlendirilmektedir. Yapılan isimlendirme, veri tabanlarında yer alan, kürasyonu manüel gerçekleştirilerek doğrulanmış dizi verisi ile elde edilen protein koleksiyonlarına dayanmaktadır (33).

Organizmanın genomunda yer alan tüm nükleotidlerin dizisini elde etmeye, üzerindeki fonksiyonel bölgelerin konumlarının kestirilmesi ve bu bölgelerin isimlendirilmesine yönelik olarak gerçekleştirilen çok aşamalı araştırmalara genom projesi adı verilmektedir. Bugün itibariyle Amerika Department of Energy Joint Genome Institute GOLD (Genomes Online Database) 2014 Nisan kayıtlarına göre dünya üzerinde toplam 43583 genom projesi bulunmaktadır. Bunlardan ondokuz bine yakını tamamlanmamış, yirmiüçbinden fazlası tamamlanmıştır. Tamamlanmış genom projelerinden onaltı bine yakını kalıcı taslaktır, üç bin civarında ise bitirilmiş proje bulunmaktadır (5). (genomesonline.org)

Tüm genom dizileme dolayısıyla genom projeleri devri, görece daha küçük boyutlu ve yapısal olarak daha az karmaşık olmaları nedeniyle, öncelikle mikrobiyal genom dizileme işlemleriyle başlamıştır. Asıl hedefin insan genomunun dizilenmesi olduğu bu süreçte, mikrobiyal genom dizilemeyi ökaryotların dizilenmesi izlemiş böylece *Saccharomyces cerevesiae* (maya), *Caenorhabditis elegans* (solucan), *Drosophila melanogaster* (sirke sineği), *Arabidopsis thaliana* (hardal tohumu) gibi ökaryotik genomlar dizilenmiştir. Tüm bu denemelerin neticesinde elde edilen deneyimlerden yararlanılarak insan genomunun dizilenmesi gerçekleştirilmiştir .

İnsan genom projesi biyolojinin ilk geniş ölçekli projesidir. Özellikle sağlık alanında sağlayacağı gelişmeler doğrultusunda çok önemli bulgular vaadeden proje, dünya üzerindeki pek çok ülkeden merkezin ve araştırmacının katılmasıyla on yılı aşkın sürede tamamlanmıştır. İnsan genomunun kompleks yapısı ve projeden beklentiler göz önünde bulundurulduğunda, sonuçların en kısa zamanda duyurulmaya başlaması ve sonuçların mükemmeliyeti özellikleri ön plana çıkmaktadır. Sonuçların daha kısa sürede elde edilmesini sağlayacak shotgun dizileme yöntemi bu dönemde Craig Venter ve ekibi tarafından İnsan Genom Projesi'ne uyarlanmış ve genom projesi eş zamanlı olarak ikinci bir strateji ile de sürdürülmüştür. Bu ikinci strateji, tüm itiraz ve endişelere rağmen dizileme işleminin sonuçlarının daha kısa sürede ve daha düşük maliyetle elde edilebilmesini sağlamıştır.

DNA'nın çift zincirli yapısını ortaya koyan iki araştırmacıdan biri olan James Watson, insan genomunun olabilecek en kısa zamanda ortaya konmamasının öncelikli olarak "ahlak dışı" olduğunu söylemiştir. İnsan Genom Projesi'nin gerçekleştirilmesi sırasında elde edilen bilgi ve deneyimler sayesinde bir dizi yeni teknoloji geliştirilmiş, pek çok organizmanın fiziksel, genetik ve transkript haritaları elde edilmiş, biyoetik çalışmaları yoğunlaşmış ve insan genomunun herkes tarafından ulaşılabilecek dizisi veritabanlarında yerini almıştır (11). Günümüzde bu deneyimler doğrultusunda geliştirilen yeni dizileme teknolojileri ve anotasyon teknikleri ile genom projeleri büyük bir hızla ve görece düşük maliyetlerle devam etmektedir. Çizelge 2.1'de organizmalara göre genom projeleri isatistikleri verilmiştir. http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html

Çizelge 2.1 Organizmalara göre genom projeleri

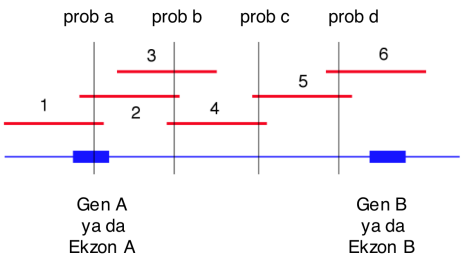
Organizma	Tamamlanmış	Taslak birleştirme	Devam eden	Toplam
Prokaryotlar	1117	966	595	2678
Archaea	100	5	48	153
Bacteria	1017	961	547	2525
Ökaryotlar	36	319	294	649
Hayvanlar	6	137	106	249
Memeliler	3	41	25	69
Kuşlar		3	13	16
Balıklar		16	16	32
Böcekler	2	38	17	57
Yassı Solucanlar		3	3	6
Roundworms	1	16	11	28
Amfibiler		1		1
Sürüngenler		2		2
Diğer Hayvanlar		20	24	44
Bitkiler	5	33	80	118
Kara Bitkileri	3	29	73	105
Yeşil Algler	2	4	6	12
Mantarlar	17	107	59	183
Asklı Mantarlar	13	83	38	134
Basidiomisetler	2	16	11	29
Diğer Mantarlar	2	8	10	20
Protistler	8	39	46	93
Sporlular	3	11	16	30
Kinetoplastlar	4	3	2	9
Diğer Protistler	1	24	28	53
Toplam	1153	1285	889	3327

2.2. Dizileme

DNA dizileme, kalıtım materyeli olan DNA'yı meydana getiren nükleotidlerin dizisinin belirlenmesidir. Araştırmacıların araştırma amaç ya da bulguları doğrultusunda genomun yalnızca araştırmaya konu olan bölgeleri dizilenebilir. Örneğin popülasyonda gözlenen "normal" yapısına göre hasta bireyde bir genin nükleotid düzeyinde farklılık gösterip göstermediğinin araştırılmasında, genom üzerinde yalnızca o genin yer aldığı bölgenin dizilenmesi gibi.

Kısa genomik bölgelerin dizilenmesinin yanısıra tüm genom dizisinin elde edilmesi de mümkün olmaktadır. Bu tip dizileme "tüm genom dizileme" işlemi ile gerçekleştirilmektedir. Tüm genom dizileme işlemini dizideki genomik bileşenlerin belirlenmesi amacıyla anotasyon işleminin gerçekleştirilmesi takip etmektedir. Tüm genom dizileme ve elde edilen dizinin biyolojik olarak anlamlandırılması amacıyla anotasyonunun işlemlerinin gerçekleştirildiği projelere *genom projesi* adı verilmektedir (33)

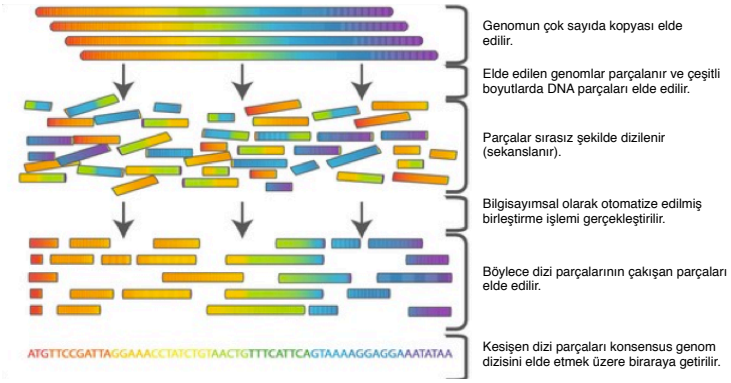
Günümüzde iki çeşit dizileme yöntemi kullanılmaktadır. Bunlardan biri *yönlenik* (directed) dizileme, diğeri ise *shotgun* dizilemedir. Yönlenik dizilemede önce nükleotid dizisi elde edilmek istenen DNA örneğinin, her bir parçasının sonu bir sonraki parçanın başıyla kesişen kopyaları elde edilmektedir (bu kopyalardan oluşan koleksiyona kütüphane denilmektedir.). Bu parçaların okunmasının (dizilenmesinin) ardından, kesişim bölgelerinden yararlanılarak genom dizisi tekrar birleştirilmektedir. Bu işlem gerçekleştirilirken genomda dizisi bilinen iki bölgenin uçlarından başlanır, her bir adımda yeni bir dizi parçası elde edilir ve bu yeni parçanın bitiş bölgesine yeni bir prob tasarlanarak yeni bir dizileme işlemiyle bir adım daha ilerlenir. Böylece adım adım hedeflenen diğer bilinen bölgeye ulaşılır. Sonuç olarak nükleotid dizisi bilinen iki bölge arasında yer alan, nükleotid dizisi bilinmeyen, tek bir kerede dizilenmesi mümkün olmayan uzun bölgenin nükleotid dizisi elde edilmiş olur (32). Şekil 2.1'de yönlenik dizileme özetlenmektedir (http://oregonstate.edu/dept/biochem/hhmi/hhmiclasses/bb451/figslett/FigBC.html).



Şekil 2.1 Yönlenik dizileme

Shotgun dizilemede ise genom fiziksel olarak rastgele parçalara bölünmekte, elde edilen rastgele parçalar çoğaltılmakta ve okunmaktadır. Ardından bu parçalar yine kesişim bilgisi yardımıyla, etkin bilgisayar algoritmaları kullanılarak genom dizisini oluşturacak şekilde bir araya getirilmektedir. Şekil 2.2'de işlem genel hatlarıyla özetlenmiştir (http://commons.wikimedia.org/wiki/File:Whole_genome_shotgun_sequencing_versus_Hierarchical_shotgun_sequencing.png). Bu işleme genom birleştirme (genome assembly) denmektedir. Genom birleştirme konusuna Bölüm 2.3'te değinilmiştir.

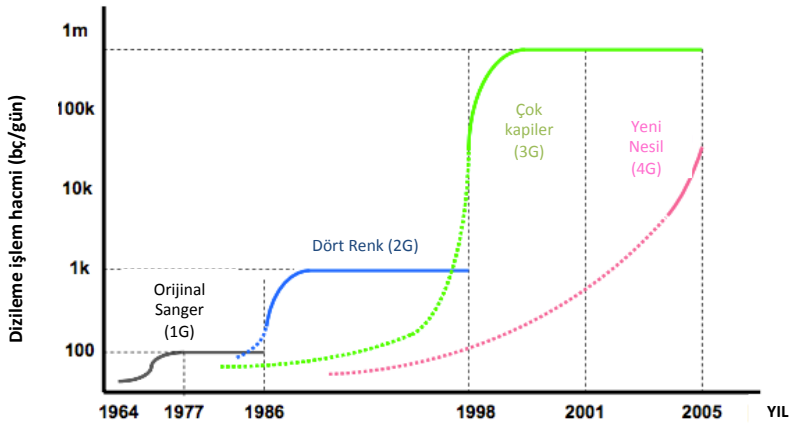
Yönlenik dizileme ve shotgun dizileme ile elde edilen genom parçalarının okunarak bütünü oluşturan daha uzun bir nükleotid dizisi haline getirilmesinde kullanılan iki farklı teknoloji bulunmaktadır. Bunlar birinci nesil ve ikinci nesil dizileme teknolojileri olarak anılmaktadır. Her iki teknoloji de günümüzde tüm genom dizileme projelerinde birbirlerini tamamlayıcı olarak kullanılmaya devam etmektedir.



Şekil 2.2 Shotgun dizileme sürecinin şematik gösterimi

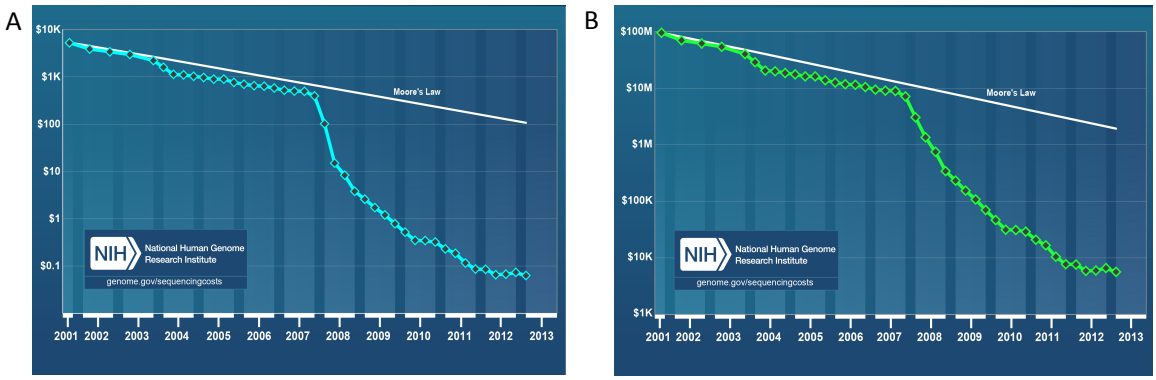
Birinci nesil dizileme teknolojilerinin tümü 1977'de Fred Sanger ve Alan R. Coulson tarafından geliştirilen seri DNA dizileme yöntemine dayanmaktadır (30) . Sanger yöntemi sağladığı teknik kolaylıklar doğrultusunda yöntemin icadını izleyen 30 yılda kullanılan tek dizileme yöntemi haline gelmiştir. Bu yöntemin icat edildiği ilk günlerdeki işlem hacmi günde ortalama 100 baz çiftiydi.

Şekil 2.3'te genom dizileme aracı endüstrisi S- Eğrileri görülmektedir. S- Eğrileri teknolojilerin yaşam döngüsünü ifade etmektedir. (http://www.improvementandinnovation.com/features/article/understanding-s-curve-innovation/) Buna göre 1964'ten 1986'ya kadar süren 1G fazında, Orijinal Sanger dizileme ile maksimum dizileme işlem hacmi günde 100bç iken, 2G fazında dört renk teknolojisiyle 12 yıl içinde bu sayı günde 1000bç'ye, 3G fazında çok kapilerli teknolojiyle işlem hacmi günde yaklaşık 1milyon baza ulaşmıştır. Yeni nesil dizileme ile artık günlük dizileme işlem hacmi milyon bazlar seviyesindedir (27).



Şekil 2.3 Yıllara göre dizileme işlem hacmi (27)

İnsan genomunun dizilenmesi sırasında projenin daha hızlı tamamlanabilmesi için ihtiyaç duyulan dizileme işlem hacminin tahmin edilenden çok yüksek olduğunun ortaya çıkmıştır. Böylece, dizileme işlem hacminin arttırılmasına yönelik olarak otomatize sistemler geliştirilmiş ve bu sistemlerin kullanıldığı fabrika benzeri dizileme merkezleri oluşturulmuştur (11). Otomatize sistemlerden ilki olan ve 1986 yılında Applied Biosystems tarafından geliştirilen ABI 370A sayesinde 70'li yılların sonunda günde ortalama 100 bç. olan dizileme işlem hacmi günde ortalama 10000 bç.'ye yükselmiştir (27). Her ne kadar bu merkezlerde birbirine rakip iki insan genom projesinin tamamlanması mümkün olduysa da, bu merkezler yine de sonraki araştırmalar için ihtiyaç duyulan dizileme işlem hacminin karşılanmasına yetmemiştir. Böylece yeni (ikinci) nesil dizileme teknolojilerinin geliştirilmesi gündeme gelmiştir. Bu yeni teknolojilerin ilk versiyonu olan 454 yeni nesil dizileme cihazı bir defada, elli adet en son teknoloji Sanger tabanlı Applied Biosystem 3730XL'nin sağlayabileceği dizileme işlem hacmini altıda bir maliyetle sağlamaktadır (31).



Şekil 2.4 A. Bir megabaz (1000 bç) uzunluğunda ham DNA dizisinin yıllara göre maliyeti B. Yıllara göre insan genomu dizileme maliyetleri

Şekil 2.4'te NIH National Human Geniome Research Institute verilerine göre bir megabaz uzunluğunda ham DNA dizisinin elde edilme maliyetleri gösterilmiştir. Grafiklerde dizileme maliyetlerinin, kabaca, bilgisayarların hesap gücünün her sene iki katına çıkacağını, böylece maliyetlerin de tersi bir trendde azalacağını Moore yasası ile gerçekleştirilen düşüş tahminlerine göre daha büyük hızda düştüğü görülmektedir http://www.genome.gov/sequencingcosts/.

2.2.1. Shotgun Dizileme ve Bu Yöntemle Elde Edilen Verinin Yapısı

Mevcut dizileme teknolojisi herhangi bir genomun bir defada kesintisiz olarak tümünün "okunabilmesine" imkan vermemektedir. Bu nedenle gerek Sanger kimyası gerek yeni nesil sekanslama yaklaşımlarında kullanılan kimyasal yaklaşımlar, genom dizilemenin ilk aşamasında genomun okunabilir daha küçük parçalara ayrılmasını gerektirmektedir. Bilinen en küçük (viral olmayan) canlı genomu 160kb (25) en büyük canlı genomu da 670Gb büyüklüğündedir (26). Sanger yöntemine dayalı teknolojilerle ~1000 bç.'ye kadar uzunlukta okumalar elde edilebilmektedir ve bu okumalar günümüz teknolojisiyle elde edilebilen en uzun okumalardır. Bu veriler göz önünde bulundurulduğunda bir canlının tüm genom dizisinin Bölüm 2.2'de bahsedilen yönlenik dizileme yöntemiyle elde edilmesinin süre ve maliyet yönünden pahalı olduğu açıkça görülebilir.

Shotgun dizileme yöntemi bu sorunun giderilmesi amacıyla İnsan Genom Projesi'nin yönlenik strateji ile yürütüldüğü dönemde gündeme getirilmiştir (11). İnsan Genom Projesine Celera Şirketinin dahil olması ile uygulanmaya başlayan bu yöntem, projenin daha kısa sürede tamamlanmasında büyük rol oynamıştır (24). Ancak bu durum değerlendirilirken Sanger dizileme teknoloji kullanılarak tamamlanan İnsan Genom Projesinde kullanılan shotgun stratejisinin 1990-1998 yılları arasında yalnızca yönlenik strateji ile elde edilmiş insan genom verisine Gen Bankası veritabanı (GenBank Database) üzerinden ulaşabildiği bilgisi unutulmamalıdır.

Yeni nesil sekanslama teknolojileri ise genom sekanslamada standart olarak shotgun dizileme ile işe başlamaktadır. Bugünkü teknolojinin işlem hacmi ve bilgisayar teknolojisinin eriştiği hesaplama gücü, artık shotgun dizileme sonucu elde edilen okumaları genomu güvenilir şekilde biraraya getirebilecek (assembly) aşamaya gelmiştir. Her ne kadar (özellikle *de novo*) genom projelerinin tamamlanması için shotgun dizileme yeterli olmasa da, bu yöntem, üzerinde gen bulma işleminin yapılabileceği kadar bilgi veren bir taslak genom dizisinin en kısa sürede oluşturulması için çok değerli bir veri sağlamaktadır.

Shotgun yöntemi, elde bir dizinin birden fazla kopyası bulunduğunda, bu dizinin rastgele parçalarının, örtüşme verisinden yararlanılarak tekrar birleştirilebileceği teorisine dayanmaktadır (34). Günümüzde, *de novo* genom projelerinde taslak genom dizisi ilk olarak shotgun dizileme ve *de novo* birleştirme algoritmaları kullanılarak elde edilmektedir. Genomda yer alan ve shotgun yöntemiyle dizilenemeyen daha karmaşık bölgelerin dizisi ise yönlenik dizileme ve Sanger yöntemiyle elde edilmekte, ardından bu iki yöntemle elde edilen diziler birleştirilerek genom dizisi tamamlanmaya çalışılmaktadır. Dizinin anotasyonu için dizileme işleminin tamamen bitirilmesi beklenmemektedir. Dizileme işleminin tamamlanması sırasında dizinin anote edilen kısımları üzerinde çalışılabilmektedir.

Bir genom projesinin bitirilmiş (finished) kabul edilebilmesi için NCBI'nın belirttiği güncel minimum standartlar şunlardır:

1- Yapısal RNA'ların (5S, 16S, 23S) her birinden, uygun uzunlukta, en az bir kopya bulunmuş olması
2- Her aminoasit için en az bir tRNA bulunmuş olması
3- Protein kodlayan gen/ genom uzunluğu oranının bire yakın olması
4- Hiçbir genin aynı zincirde ya da karşı zincirde, tamamen diğer bir genin içinde kalmıyor olması
5- Genomik bileşenlerden hiçbirinin kısmen dizilenmiş olmaması

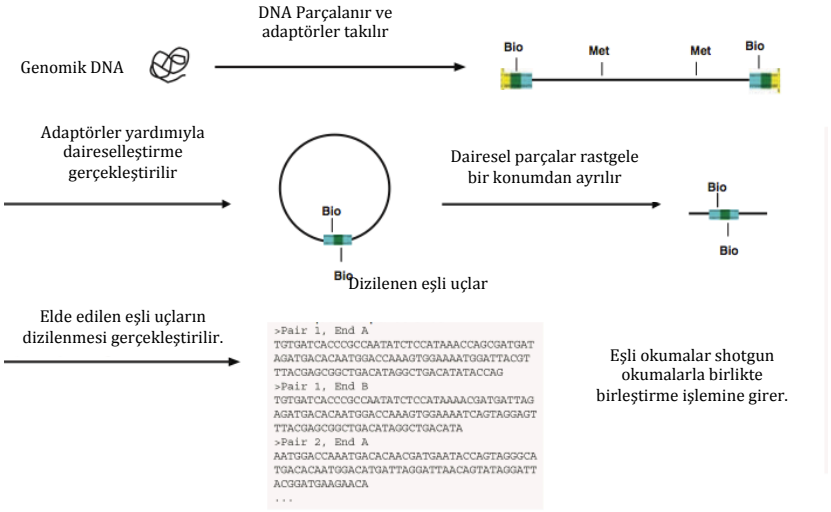
Tüm bu özelliklerin sağlanması koşuluyla bir genomun projesinin bitirilmesi yıllar sürebilmektedir. Örneğin dizilenen ilk prokaryotik organizma olan *Haemophilus influenza* genom projesi 1995 yılında başlamıştır ve ancak 2013 yılında bitirilmiştir. *H. influenza* genomu 1.8Mb büyüklüğündedir. (*B. boroniphilus* 4.6Mb.)

2.2.2. Paired-end Dizileme

Paired-end dizileme, dizileme işleminin kütüphane oluşturma aşamasında kullanılan stratejilerden biridir. Diğer bir deyişle farklı yapıda okuma verisi elde etmek amacıyla biyolojik malzemeye uygulanan çeşitli prosedürlerden biridir (17).

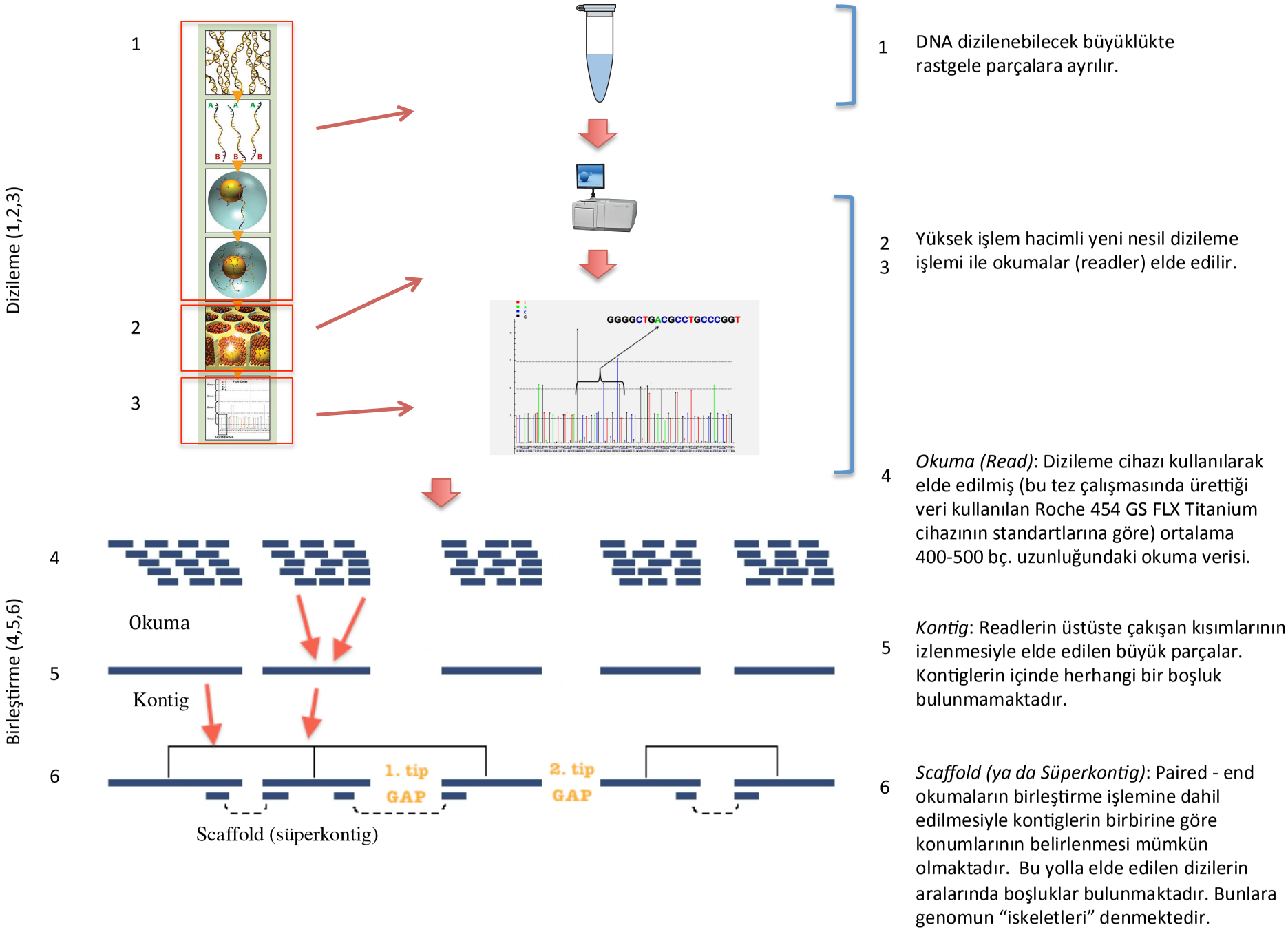
De novo dizilenen bir genomda shotgun dizilemeyle elde edilen okumalar (readler) birleştirildikten sonra, genomun ve peşpeşe gerçekleştirilen deneylerin bazı niteliklerinden dolayı nükleotd dizisinin tek parça halinde elde edilmesi mümkün olmamaktadır. Genom birbiriyle "bitişen" parçaların birleştirilmesiyle oluşan çok sayıda "kontig" halinde elde edilebilmektedir. Genom dizisini daha net ortaya koyan bir iskeletin oluşturulması için, dizileme stratejisine bağlı olarak, paired-end dizileme yöntemiyle elde edilmiş okumalardan yararlanılabilmektedir (4).

Paired-end okumalar, genomda peşpeşe gelme olasılığı düşük olan dizi parçalarının birleştirme işlemine katacağı bilgiyi kullanmak amacıyla elde edilir ve kullanılır. Bu parçalar o kadar "biricik"tirler ki, genom üzerinde iki ayrı konumda gözlendiklerinde, birleştirmede birbirlerine yakın olmaları gerektiği bilgisini verirler ve böylece kontigler üzerine haritalandıklarında, kontiglerin sıraya konarak genom iskeletini (scaffold, süperkontig) oluşturmasını sağlarlar (17).



Şekil 2.5 Paired-end dizileme

Bir genom projesinde, shotgun dizileme, (stratejiye bağlı olarak) paired-end shotgun dizileme ve Sanger dizileme yöntemlerinden elde edilen okuma verisi bir arada kullanılmaktadır. Yeterlilik, süre ve maliyet açısından dizileme işlemini tamamlayacak optimal düzenek elde edilir ve deneyler birbirini tamamlayıcı şekilde gerçekleştirilir. Yeni nesil dizileme süreci Şekil 2.6'da özetlenmiştir. 1,2 ve 3. Aşamalar ıslak laboratuvar işlemleridir. Shotgun ya da paired-end dizileme bu aşamalarda gerçekleştirilmektedir. 4,5 ve 6. basamaklar hesaplamsal yöntemlerle ilk basamakta elde edilen okumaların birleştirilmesine yönelik işlemlerdir. 6. basamakta gösterilen süperkontig içi boşlukların doldurulmasında ise Bölüm 2.2'de bahsedilen, Sanger kimyasına dayalı yönlenik dizileme işlemi gerçekleştirilmektedir.

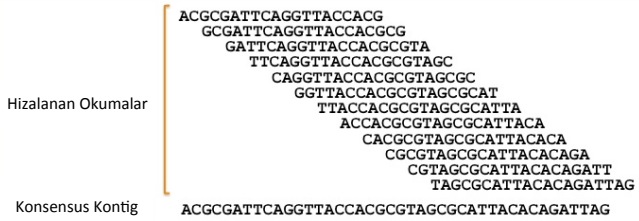


Şekil 2.6 Dizileme, birleştirme işlem akışı ve birleştirme işlemi ile ilgili terimler

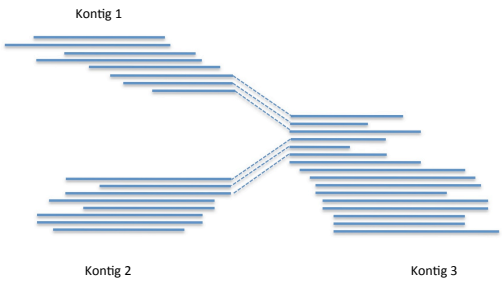
Birleştirme işlemi dizilemeyle elde edilen okumaların genom dizisini oluşturacak şekilde bir araya getirilmesini sağlayan işlemdir. Shotgun dizileme yönteminin *de novo* genom dizilemede kullanılabilmesinin en önemli nedenlerinden biri, etkin birleştirme algoritmalarının varlığıdır.

Genom birleştirme zor bir hesaplamsal problemdir. Birleştirme algoritmaları temelde her bir okumanın diğer okumalarla hizalanmasıyla elde edilen çakışan bölgeler aracılığıyla okumaların uç uca eklenmesini sağlamaktadır. Birbiriyle bitişen okumalar kontigleri oluşturmaktadır. Roche 454 GS FLX cihazından alınan okumalarla gerçekleştirilen tüm genom shotgun dizileme verisinin birleştirilmesi için kullanılan önde gelen birleştirme yazılımı GSAssembler, Newbler algoritmasını kullanmaktadır. Newbler algoritmasının çalışma prensibi teorik özelliklerine değinilmeksizin genel hatlarıyla şöyle özetlenebilir:

1- Birbiriyle örtüşme gösteren okumalar belirlenir
2- İkili olarak örtüşen okumalarla çoklu hizalamalar gerçekleştirilir

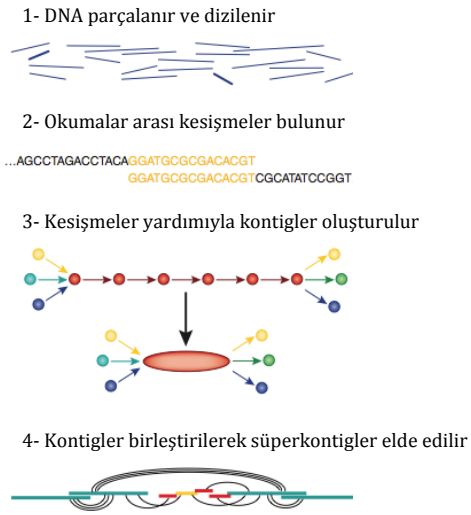


3- Çoklu hizalamalarda farklı okuma kümelerinde tutarlı farklılıkların görüldüğü yerlerden diziler parçalanır, böylece "kontig"ler elde edilmiş olur.
4- Bazı okumaların birden fazla kontigde yer alma olasılığı olduğundan kontigler arasındaki dallanma yapısı çözümlenir. Özellikle tekrar dizileri bu dallanma yapısının karmaşık olmasına neden olur.



5- Çoklu hizalamada kullanılan okumalarda yer alan nükleotidler için kalite ve *flow* değerleri (yeni nesil dizilemenin teknik özelliklerine bağlı bir isimlendirme. Belli bir pozisyonda hangi nükleotidden peşpeşe kaç tane okunduğunu gösteren flowgram üzerindeki her bir değere verilen isim) göz önünde bulundurularak "base calling" (yani A, T, G, C isimlerinin verilmesi) işlemi gerçekleştirilir.

Eğer deneyde paired-end okumalar da mevcutsa, birleştirme işlemine bu okumaların kontigler üzerine haritalanması basamağı da eklenmektedir. Böylece kontigler sıralı şekilde daha büyük kümeler halinde bir araya getirilmekte ve böylece taslak genom dizisini oluşturacak bir iskelet elde edilmektedir.



Şekil 2.9 Shotgun ve paired-end okumaların birleştirilmesi

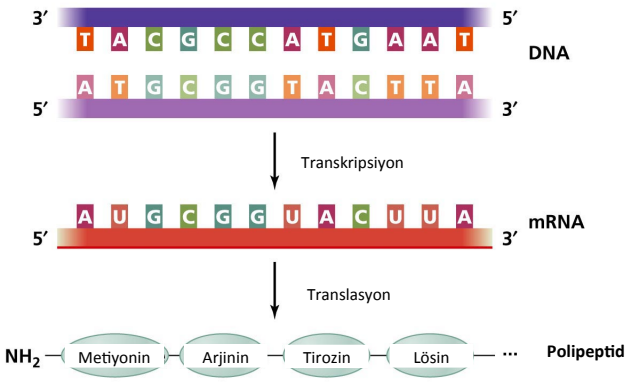
2.3. Gen Bulma

Gen bulma, bir nükleotid dizisi verildiğinde o nükleotid dizisi üzerinde yer alan kodlayıcı bölgelerin koordinatlarının bulunması işlemidir. Bir genom projesinde, nükleotid diziliminin elde edilmesinin ardından dizi üzerinde yer alan kodlayıcı bölgeler elde edilmekte ve bu bölgeler fonksiyonlarını ifade eden belirteçlerle işaretlenmektedir. Protein kodlayan bölgelerin bulunması, organizmanın yapısı hakkında genel tabloyu vereceğinden bir genom projesinin en önemli aşamalarından biridir.

2.3.1. Protein Kodlayan Bölgeler ve Moleküler Biyolojinin Santral Dogması

Bölüm 2.1'de bahsedildiği gibi, genom projelerinin dizileme işlemini takip eden ikinci önemli aşaması, elde edilen dizi üzerindeki "protein kodlayan" bölgelerin bulunması ve bu bölgelerin işlevlerini ifade eden belirteçlerle isimlendirilmesidir.

Bir organizmanın yaşamsal faaliyetleri, hücrelerde sentezlenen proteinler aracılığıyla sürdürülmektedir. DNA üzerinden RNA'lar aracılığıyla aktarılan bilgi ile sentezlenen proteinlerin primer yapısının çözümlenmesi ise ancak nükleotid dizisinin bilinmesiyle mümkün olmaktadır. DNA üzerinden RNA, RNA üzerinden de protein sentezlenmesini sağlayan bilgi akışının çözümlenmesi ilk kez 1958 yılında, DNA molekülünün yapısını Watson ile birlikte çözümleyen Crick tarafından gerçekleştirilmiştir. Bu bilgi akışı "moleküler biyolojinin santral (merkezi) dogması" olarak anılmaktadır (20).



Şekil 2.10 Moleküler biyolojinin santral dogması

Bir organizmanın genomu üzerinde işlevleri ve özellikleri birbirinden farklı pek çok bölge bulunmaktadır. Bu bölgeler genel olarak *protein kodlayan* ve *kodlamayan bölgeler* olarak gruplandırılabilir. *Gen*, bir polipeptidin ya da bir fonksiyonel RNA'nın sentezlenmesini sağlayan bir DNA dizisine verilen addır. DNA dizisi üzerinde yer alan kodlayıcı bölgeler, Şekil 2.10'da gösterildiği gibi, öncelikle mRNA'ya (mesajcı RNA) transkribe edilir. Ardından ribozomlar üzerinde tRNA'ların (taşıyıcı RNA) taşıdığı aminoasitlerin mRNA dizisindeki koda uygun şekilde proteine transle edilmesi işlemi gerçekleşir. Böylece DNA üzerinde bulunan kod, organizmanın yaşamsal faaliyetlerinde kullanılacak olan proteine dönüştürülmüş olur.

DNA dizisi üzerinde kodlayan bölgede yer alan nükleotid üçlülerine kodon adı verilmektedir. Her bir kodon, spesifik olarak tek bir aminoasite karşılık gelmektedir. mRNA aracılığıyla DNA'dan hücrenin protein üretiminden sorumlu birimi olan ribozoma ulaştırılan kodon dizisi, tRNA'ların getirdiği aminoasitlerin ribozom üzerinde bu bilgiye göre birleştirilerek polipeptidlere ve hemen ardından proteine dönüştürülmesini sağlar.

DNA üzerindeki kodlayan bölgeler genel olarak, bir açık okuma çerçevesi üzerinde bulunurlar ve organizmaya özgü belirli bir uzunluktadırlar. Bir açık okuma çerçevesi, başlangıç kodonu ile başlayan, bitiş kodonuyla biten, 3n uzunluğunda nükleotid içeren bir DNA dizisidir. Her açık okuma çerçevesi kodlayan bölge olmamakla birlikte, bir genin bir açık okuma çerçevesi üzerinde yer alması olasılığı yüksektir.

Kodlayıcı bölgelere ilişkin bu temel bilgiler ve santral dogmanın ortaya konmuş olması, hesaplamsal yöntemlerle otomatize gen bulmanın mümkün olabileceği fikrinin ortaya çıkmasına neden olmuştur. Kodlayıcı bölgelere ait daha fazla özelliğin keşfiyle etkinliği artan modeller aracılığıyla gen bulma algoritmaları her geçen gün iyileştirilmektedir.

2.3.2. Hesaplamsal Gen Bulucuların Geliştirilmesi

Önceki bölümlerde de bahsedildiği gibi, 1990 yılında İnsan Genom Projesi'nin başlamasıyla birlikte veritabanlarında bulunan insan ve model organizmalara ait genom dizisi sayısındaki artış hızlanmaya başlamıştır. 1998 yılında proje geniş ölçekli dizileme fazına girdiğinde, projenin ana hedeflerinden biri olan dizilenen insan ve model organizma genomlarındaki genlerin belirlenmesi işlemi büyük önem kazanmıştır (35).

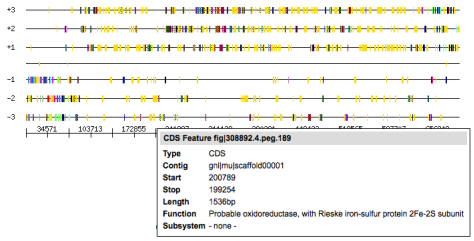
Genom anotasyonuna giden yoldaki en önemli işlem basamaklarından biri olan gen bulma işlemi hem ıslak laboratuvar yötemleri, hem de hesaplamsal (computational) yöntemler kullanılarak gerçekleştirilebilmektedir. Fakat gen bulmada kullanılan zahmetli ve tamamlanması uzun süren ıslak laboratuvar yöntemleri, elde edilen verinin hızla anlamlandırılması için gereken niteliklere sahip değildir. Islak laboratuvar çalışmaları ile gerçekleştirilen gen bulma işlemlerinde birden fazla genin homolog rekombinasyon oranlarının istatistiksel analizleri yapılarak bu genlerin kromozom üzerinde birbirine göre konumları belirlenmektedir. Pek çok ıslak laboratuvar deneyi tamamlandıktan sonra yapılan istatistiksel analizler doğrultusunda genlerin birbirlerine göre konumlarını kabaca gösteren genetik haritalar elde edilmektedir (20).

Gen bulmada kullanılan hesaplamsal yöntemler ise genomda gen kodlayan bölgelerin bir nükleotid çözünürlükte (yani kesin başlangıç ve kesin bitiş koordinatları olarak) tanımlanması hedefiyle geliştirilmektedir. Bu yöntemler günümüzde sıklıkla deneysel çalışmalardan daha iyi sonuç vermektedir (10). Metnin devamında "gen bulma" terimi hesaplamsal yöntemlerle gerçekleştirilen gen bulma işlemini ifade edecektir. Bu tez çalışmasında *de novo* genom dizilenmesi gerçekleştirilmiş prokaryotik genom verisi kullanıldığından yalnızca prokaryotik gen bulma yöntemlerinden bahsedilecektir. Gen yapısındaki farklılıklar dolayısıyla analizi de farklı olan ökaryotik genomların analizinden söz edilmeyecektir.

2.4. Genom Anotasyonu

Genom anotasyonu, genom üzerindeki kodlayıcı bölgelerin koordinatlarının belirlenmesi ve bu bölgelerin veri tabanında bulunan benzerleri aracılığıyla isimlendirilmesi sürecidir. Anotasyonu yapılmış DNA dizisi demek, üzerinde yer alan fonksiyonel bölgeler belirlenmiş ve isimlendirilmiş bir dizi demektir. Dolayısıyla genom anotasyonu kabaca iki aşamada gerçekleştirilir. Birinci aşama gen bulma, ikinci aşama ise anotasyondur.

Gen bulma genomda yer alan kodlayıcı ve düzenleyici bölgelerin kestirilmesidir. Bu öncelikle genom dizisi üzerinde yer alan işlevsel bölgelerin koordinatlarının belirlenmesi anlamına gelmektedir. Bu işlemin ardından gerçekleştirilen genom anotasyonu ise, bu işlevsel bölgelerin görevlerinin belirlenmesi ve isimlendirilmesi aşamasıdır. Bu süreç genel anlamda üç basamak olarak ele alınabilir. Bunlar nükleotid düzeyinde anotasyon, protein düzeyinde anotasyon ve biyolojik süreçler düzeyinde anotasyondur (33).



Şekil 2.11 RAST ile anotasyonu gerçekleştirilmiş bir nükleotid dizisi

Böylece canlının nükleotid dizisi, bu dizi üzerinde hangi bölgelerde hangi proteinlerin kodlandığı ve bu proteinlerin hangi biyolojik süreçlerde yol aldığı bilgileri elde edilmiş olmaktadır. Bu dizi ve bilgiler tümleşik olarak kullanılabilecek şekilde düzenlendikten sonra uluslararası veritabanlarına kaydedilerek tüm araştırıcıların kullanımına sunulmaktadır.

2.4.1. Prokaryotik Genomların Özellikleri ve Prokaryotik Genomlarda Gen Bulma

Prokaryotik genomlar küçük ve kompakt (intron taşımayan) yapıları sayesinde başlangıçta genom projeleri için iyi bir uygulama alanı olmuştur. Bu çerçevede genomu ortaya çıkarılan ilk organizma *Haemophilus influenzae* olmuştur. Prokaryotik genomların dizilerinin ortaya konmaya başlanmasıyla birlikte, bilim insanlarının elinde "kodlayan dizi" ve "kodlamayan" diziyi temsil eden veri birikmeye başlamıştır. Dolayısıyla bir süre sonra kodlayan ve kodlamayan DNA dizileri arasında nükleotid içeriğindeki değişimler yönünden istatistiksel olarak fark olup olmadığının araştırılması mümkün olmuştur. Bu bulgular, kodlayan ve kodlamayan bölgelerin otomatize gen bulmaya yönelik modellenebilirliğiyle ilgili başlangıç ipuçlarını sağlamıştır (8).

80'li yılların ortalarında Borodovsky ve ekibinin E. coli genomunun yaklaşık olarak kırkta biri uzunluğunda bir nükleotid koleksiyonu üzerinde yer alan kodlayan ve kodlamayan DNA dizileri üzerinde gerçekleştirdiği istatistiksel analizler, bu bölgeler arasında pozisyon spesifik olarak kodon gözlenme sıklıkları yönünden istatistiksel bir farklılığın bulunduğu konusunda veri sağlamış ve otomatize gen bulma sistemlerinin geliştirilebileceğini ortaya koymuştur (8).

Günümüzde, geliştirilmiş olan GeneMark ve GLIMMER gibi gen bulma algoritmaları ve bu algoritmalar üzerine geliştirilen PGAAP (Prokaryotic Genomes Automated Pipeline) ve RAST (Rapid Annotation using Subsystem Technology) gibi otomatize gen bulma ve anotasyon sistemleri sayesinde, prokaryotik genom araştırılmaları için temel oluşturan boyut ve nitelikte genom verisi hızlı, verimli ve güvenilir şekilde elde edilebilmektedir (2,3,22,29). Bu özellikte veri eldesi biyoteknoloji ve tıp alanında gerçekleştirilen araştırmalar için büyük önem taşımaktadır.

2.4.1.1. Prokaryotik Genomların Özellikleri

Canlılar prokaryot (hücre çekirdeği olmayan) ve ökaryot (hücre çekirdeği olan) olarak iki gruba ayrılır. Prokaryotik canlılara bu tez çalışmasında genomu üzerinde çalışılan *Bacillus boroniphilus*, ökaryotik canlılara da *Homo sapiens* örnek verilebilir. Bu iki grubun genom yapıları birbirinden büyük ölçüde farklıdır. En büyük farklılıklardan biri, prokaryotik genomların organizasyonun ökaryotik genomlardan daha az karmaşık olmasıdır. Bu farklılığın en temel örneklerinden biri, ökaryotların genlerinin ekzon ve intronlardan oluşması, prokaryotlarda ise bu yapının bulunmamasıdır. Prokaryotik genomların yapısının daha net anlaşılabilmesini sağlamak amacıyla aşağıda bazı özellikler özetlenmiştir:

1- En küçük ökaryotik genomlar ile en büyük prokaryotik genomlar arasında boyut bakımından örtüşmeler vardır. Fakat genel anlamda prokaryotik genomlar ökaryotik genomlardan çok daha küçüktür.
2- Bir prokaryot genomunun tamamı tek bir sirküler DNA molekülü üzerinde yer almaktadır. Başka bir deyişle prokaryotik genomlar tek bir kromozomdan oluşmaktadır.

3- Bazı prokaryotlarda bir önceki maddede bahsedilen kromozomdan bağımsız genler de bulunabilmektedir. Bu genler plazmid denilen daha küçük DNA molekülleri üzerinde yer almaktadır. Plazmidler lineer ya da sirküler olabilmektedir. Plazmidlerin üzerinde genellikle antibiyotiğe direnci ve kompleks bileşenleri kullanabilmeyi sağlayan genler yer almaktadır. Prokaryotlar plazmidlerini birbirlerine aktarabilmektedirler.
4- Prokaryotik genomlarda gen içeriği oldukça yoğundur. Genomun yaklaşık %90'ı kodlayıcı bölgelerden oluşmaktadır. Genom üzerinde aralarındaki uzaklık yalnızca bir nükleotid olan pek çok gen bulunabilmektedir.
5- Aynı biyokimyasal yolakta görevli bazı genlerin ifadesi birbirine göre düzenlenmektedir. Bu tip gen organizasyonlarına operon denmektedir.
6- Prokaryotik genler, ökaryotik emsallerinden (intronları çıkarılmış halinden dahi) çok daha kısadır.
7- Prokaryotik genlerde intron bulunmamaktadır.
8- Prokaryotik genomlarda tekrar dizileri nadiren görülmektedir. Bunun aksine ökaryotlarda uzun, çok kopyalı tekrar dizileri bulunmaktadır. Prokaryotik genomlarda insersiyon dizileri görülebilir. İnsersiyon dizileri genomda birden fazla konumda görülebilen, dönüştürülebilen kısa genomik bileşenlerdir.

Bu yapısal farklılıklar nedeniyle hesaplamsal yöntemlerle gen bulma işlemi ökaryotlara ya da prokaryotlara özel geliştirilmiş gen bulma yöntemleri kullanılarak gerçekleştirilmektedir.

2.4.1.2. Prokaryotik Genomlarda Gen Bulmada Kullanılan Hesaplamsal Yöntemler

İnsan Genom Projesi'nin başlangıcından itibaren veri tabanlarında bulunan dizi verisi hacmi büyük bir hızla artmıştır. Bölüm 2.2'de de bahsedildiği gibi dizileme işleminin gerçekleştirilmesi kadar elde edilen nükleotid dizisinin biyolojik açıdan anlamlandırılması da önemlidir.

Günümüzde, kısaca, gen bulma ve anotasyon basamaklarından oluşan bu sürecin tamamen hesaplamsal yöntemlerle gerçekleştirilebilmesini sağlayacak algoritmalar geliştirilememiştir. Ancak hem ökaryotik hem de prokaryotik genomlar için, gen ve DNA dizisi üzerinde yer alan diğer fonksiyonel bölgelerin bulunması işlemini başarıyla gerçekleştirerek araştırmacılara ıslak laboratuvar çalışmalarında yol gösteren ve bunu yüksek hassasiyet ve güvenilirlikle gerçekleştiren pek çok algoritma bulunmaktadır.

Çizelge 2.2 Bazı gen kestirim yazılımları

Yazılım	Tanım	Tür
AUGUSTUS	Ökaryotlarda gen kestirimi	Ökaryot
BGF	gizli Markov modeli ve dinamik programlama tabanlı ab initio gen kestirimi	
DIOGENES	kısa genomik diziler üzerinde kodlayıcı bölgelerin hızla bulunmasına yönelik geliştirilmiş sistem	
EUGENE	Arabidopsis thaliana'da gen kestirimi	Arabidopsis thaliana
FGENESH	Gizli Markov Modeli tabanlı gen yapısı kestirimi	Ökaryot
FRAMED	G+C zengin Prokaryotik genomlarda gen ve çerçeve kayması bulma	Prokaryot
GENIUS	Tamamlanmış genomlarda ORF'lerin protein 3D yapılarıyla bağlantılanması	
geneid	DNA üzerinde genlerin, ekzonların, bölünme bölglerinin ve diğer sinyallerin kestirilmesi	Ökaryot
GENEPARSER	DNA dizisinin intron ve ekzonlara bölümlendirilmesi	
GeneMark	Gen kestirimi yazılımları ailesi	Prokaryot
GENOMESCAN	Çeşitli organizmalarda ekzon-intron yapısının belirlenmesi	
GENSCAN	Fourier dönüşümü ile gen bulma	
GLIMMER	Mikrobiyal DNA'da gen bulma	Prokaryot
GLIMMERHMM	Ökaryotlarda gen bulma	Ökaryot
GrailEXP	DNA dizisi üzerinde ekzon, gen, promotor bölge, polyA, CpG adaları, EST benzerlikleri ve tekrar bileşenlerinin kestirimi	
MORGAN	Omurgalı DNA'sı üzerinde karar ağaçları ile gen bulma	Ökaryot
NIX	Farklı algoritmaların sonuçlarını birleştiren bir araç (GRAIL, FEX, HEXON, MZEF, GENEMARK, GENEFINDER, FGENE, BLAST, POLYAH, REPEATMASKER, TRNASCAN)	
NNPP	Sinir ağları ile promotor bölge kestirimi	
NNSPLICE	Sinir ağları ile bölünme bölgesi kestirimi	
ORF FINDER	Tüm ORF'leri bulan bir grafik analiz aracı	
Regulatory Sequence Analysis Tools	Kodlamayan diziler üzerinde düzenleyici sinyallerin belirlenmesi	
SPLICEPREDICTOR	Bayesgil istatistik modeller ile bitkide pre-mRNAlar üzerindeki bölünme bölgelerinin tanınması	Ökaryot
VEIL	Omurgalılarda gizli Markov modeli ile gen bulma	Ökaryot

Gen bulma problemi, verilen bir nükleotid dizisi üzerinde yer alan kodlayıcı bölgelerin belirlenmesi olarak tanımlanmaktadır. Prokaryot genomlarında kodlayıcı bölge başlangıç kodonuyla başlayıp bitiş kodonuyla sonlanan ve kodon kompozisyonu özellikleri yönünden kodlayıcı bölge niteliklerini taşıyan bir nükleotid dizisi olarak tanımlandığından prokaryotik nükleotid dizilerinde bu problem başlangıç ve bitiş kodonlarının konumlarının kesin olarak belirlenmesi olarak özetlenebilmektedir.

Gen bulmada kullanılan hesaplamsal yöntemler genel olarak :

1. Kodlayan bölgelerin istatistiksel özelliklerine dayanan stokastik modeller gibi yöntemler, *intrensek (ab initio) yöntemler*
2. Dizi benzerliğini araştıran yöntemler, *ekstrensek yöntemler*

olarak iki grupta toplanabilir.

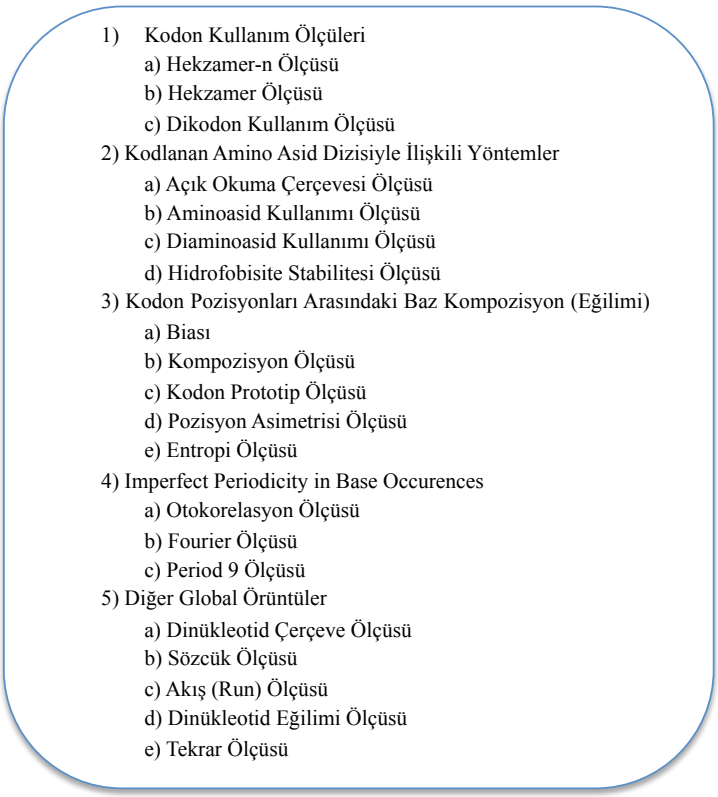
İntrensek yöntemler bir DNA dizisinin özelliklerinin üzerinde çalışılan genom nükleotid dizisinden başka DNA dizilerinden elde edilecek bilgiye başvurulmadan değerlendirilmesini sağlayan yöntemlerdir. Gen bulmada kullanılan bu özellikler kısaca ORF uzunluğu, kodon kullanımı, başlangıç kodonuna doğru uzaklıkta bir RBS dizisinin olup olmadığı, ve ifadelenen genlerin kodlamayan bölgeler ile farklılığını ifade eden daha temel (subtle) istatistik ölçüler olarak listelenebilir (12, 22).

Ekstrensek yöntemler DNA dizisinden elde edilen kodlayıcı olduğu düşünülen aminoasit dizisinin bir protein veritabanında yer alan gerçek protein dizileriyle karşılaştırılması esasına dayanmaktadır. Protein kodlayan bölgelerin evrimsel olarak korunduğu bilgisinden hareketle, fonksiyonel olup olmadığı bilinmeyen bir nükleotid dizisi, veritabanındaki kodlayan dizilerle uygun algoritmalar kullanılarak karşılaştırılmakta ve sonuçta dizinin olası görevi konusunda bilgiye ulaşılabilmektedir (9).

Ancak dizi benzerliğine dayanan yöntemlerin verimliliği doğrudan veri tabanının kapsamına ve motif arama yönteminin özelliklerine bağlı olduğundan gen bulma işlemi tamamen bu yönteme dayandırılamamaktadır. Örneğin 2002'de veri tabanlarında bulunan nükleotid dizisi koleksiyonu kullanılarak daha önce genom dizisi elde edilmemiş ve genleri deneysel yöntemlerle elde edilmemiş bir canlının genom dizisinde yer alan genlerden yalnızca %50'si homoloji taramasıyla bulunabilmekteydi (23). Her ne kadar bu oran veri tabanlarının genişlemesiyle günden güne artsa da, canlıların genomlarında daha önce bilinen hiçbir genle yüksek homoloji göstermeyen genlerinin olabileceği bilinmektedir. Bunun yanısıra iki gen homoloji gösterse bile, bu yöntemle kodlayıcı olabileceği belirlenen bir nükleotid dizisinin gen yapısı (prokaryotlar için başlama ve bitiş kodonları) tamamen tanımlanamamaktadır. Çünkü homolog proteinler kodlayıcı bölge homolojisini tüm domainlerinde göstermeyebilmektedirler. Algoritmanın nitelikleri nedeniyle bu genlerin benzerlerinin dizi hizalama ile tespit edilmesi mümkün olmamaktadır. Bu genler ancak *ab initio*, yani intrensek yöntemlerle keşfedilebilmektedir. Dolayısıyla Genom projelerinin anotasyon aşamalarında bu yöntemler birbirini tamamlayıcı olarak bir arada kullanılmaktadır (9).

İntrensek (*ab initio*) yöntemlere dayalı algoritmaların geliştirilmesine 1980'lerden itibaren başlanmıştır. Ab initio yöntemlere dayalı gen bulma algoritmaları geliştirilirken öncelikle protein kodlayan ve kodlamayan bölgelere ilişkin istatistiksel belirteçler elde edilmektedir. Bu belirteçler kodlayan ve kodlamayan bölgelerde birbirinden farklı olduğu gösterilmiş olan "kodon kullanımı" gibi genomik istatistiklerdir. Bu veriler kullanılarak, bir nükleotid dizisi üzerindeki protein kodlayan ve kodlamayan bölgelerin bulunmasına yardımcı olacak modeller oluşturulur. Ardından oluşturulan model bir örüntü tanıma algoritmasına entegre edilir (14).

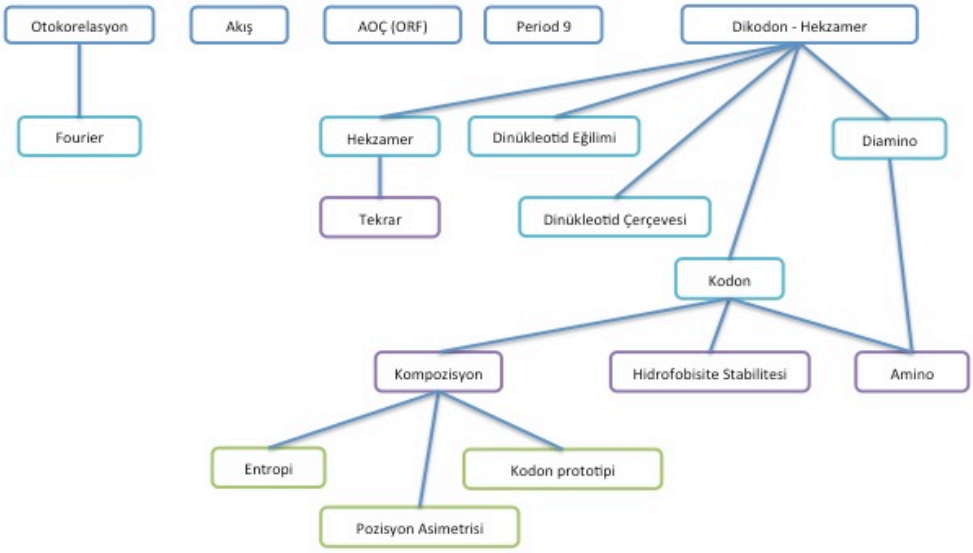
Başka bir veriye gereksinim duymaksızın yalnızca genom dizisinin protein kodlayan bölgelerinin özelliklerini ifade eden istatistiksel ölçülere dayanan hesaplamsal gen bulma yöntemlerinin geliştirilmesinin gündeme gelişinden sonra, kodlayıcı bölgeler hakkında ipucu verebilecek pek çok ölçü geliştirilmiştir (14). Bunlar Şekilde listelenmiştir:



Şekil 2.12 Gen bulma işlemi için geliştirilen örüntü tanıma algoritmalarında kullanılan ölçüler

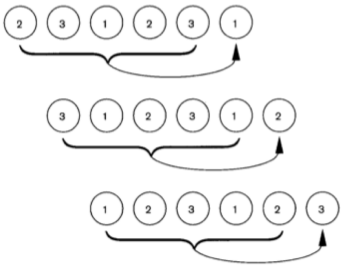
Bu ölçülerin herbiri farklı gen bulma algoritmalarında tek başına ya da kombinasyonlar halinde kullanılmıştır. Bu ölçülerin sayısı yirmileri bulduğunda Fickett ve ekibi ölçülerin dizi üzerinde kodlayıcılıkla ilgili bilgiyi özetleme gücü hakkında kapsamlı bir araştırma gerçekleştirmişlerdir. Bu araştırma sayesinde gelecekte geliştirilecek gen bulma algoritmalarında öncelikli olarak hangi ölçülerin denenmesi ve kullanılması gerektiği konusuna açıklık kazandırılmıştır.

Ölçülerin aynı gen bulma prosedüründe çalıştırılıp sonuçlarının kapsamlı olarak karşılaştırıldığı bu araştırmada yirmi kadar ölçüden otokorelasyon ve Fourier ölçüleri gibi, bazılarının birbirine çok benzer olduğu bir kısmının da bir diğerinden elde edilebilir ya da bir diğerinin özelleşmiş hali olduğu belirtilmiştir (Bkz. Şekil 2.11). Yapılan detaylı karşılaştırmalar sonucu bu ölçülerden en etkinlerinin Hekzamer-n, run, Fourier ve ORF (Açık Okuma Çerçevesi) ölçüleri olduğu gösterilmiştir (14).



Şekil 2.13 Gen bulma işlemi için geliştirilen örüntü tanıma algoritmalarında kullanılan ölçülerin birbirleriyle ilişkileri

Bu ölçüler arasında Hekzamer-n ölçüsü daha sonra geliştirilen algoritmalarda sıklıkla tercih edilen bir ölçü olarak öne çıkmıştır. Hekzamer-n ölçüsü, n = 0,1,2 için, bir test kodonunda başlangıç bazından itibaren n kadar ötelenmiş - kaydırılmış her hekzamerin sayılmasıyla elde edilmektedir.



Şekil 2.14 Hekzamer-n ölçüsünün şematik gösterimi

Gen bulma probleminde kullanılan intrensek yöntemler genellikle yukarıda bahsi geçen ölçülerin bir örüntü bulma algoritmasına entegre edilmesiyle geliştirilmektedir. Örüntü bulma algoritmaları, bir örüntünün otomatik olarak bulunmasını sağlayan algoritmalardır. Böylelikle DNA dizisi üzerindeki kodlayıcı bölgeler, kullanılan ölçü tarafından istatistiksel modele sağlanan olasılık değerleri sayesinde otomatik olarak belirlenebilmektedir .

Kodlayıcı bölgelerin otomatik olarak belirlenmesindeki hassasiyetin arttırılması amacıyla araştırmacılar pek çok istatistiksel model geliştirmişlerdir. Bunlar arasında en öne çıkan algoritmalar Markov modeli tabanlı algoritmalardır. Bu algoritmaların öncüsü *Escherichia Coli* (*E.Coli*) genomu üzerindeki genlerin bulunması için geliştirilen ECOPARSE'tır (18). Ardından EasyGene, GLIMMER, GeneMark ve GeneHacker Plus algoritmaları yine benzer teorik altyapı kullanılarak geliştirilmiştir. Bu tip algoritmalar hem prokaryotik hem de ökaryotik genomlarda gen bulma işleminde standart olarak kullanılmaya başlanmıştır.

2.5. Gizli Markov Modelleri ve Prokaryotlarda Gen Bulma

Bölüm 2.4'te bahsedildiği gibi dizileme teknolojisinin gelişimine paralel olarak DNA dizi verisinin birikim hızının artmasıyla birlikte elde edilen koleksiyon üzerinde DNA dizisinin kodlayan ve kodlamayan bölgelerinin istatistiksel özelliklerinin incelenmesi mümkün olmuştur. Nükleotid dizilerine ilişkin istatistiksel modellerin yapı ve parametrelerinin tanımlanmasında büyük öneme sahip olan istatistiksel örüntü arama algoritmalarının geliştirilmesi ise 80'li yıllardan itibaren gerçekleştirilmeye başlanmıştır.

Bir nükleotid dizisinin protein kodlayan bir bölge olma potansiyelinin otomatize gen bulma işleminde kullanılacak en önemli belirteçlerden biri olduğu ilk olarak 1982 yılında Staden ve McLachian tarafından ortaya konmuştur. Onların bulgularını destekleyen ve geliştiren çalışmalar 1984 yılında Staden ve 1992 yılında Fickett ve Tung tarafından gerçekleştirilmiştir. 1994 yılında Krogh ve ark. çalışmasıyla ilk gizli Markov modeli tabanlı gen bulucu olan ECOPARSE algoritması geliştirilmiştir. Bu öncü gelişmenin ardından gen bulmada Markov modeli yaklaşımını kullanan GeneMark algoritmasının GeneMark.hmm versiyonu geliştirilmiş, ECOPARSE'a göre performansı daha yüksek olan bu algoritma ilerleyen yıllarda diğer algoritmaların önüne geçmiştir (22).

Borodovsky ve ekibi, ilk otomatik dizileme sisteme ABI 370A'nın kullanılmaya başladığı yıl olan 1986 yılında *E. coli* genomunun kodlayan (79.900 bç) ve kodlamayan (42.600 bç) dizilerinden oluşan 135.000 bç'lik bir koleksiyonu üzerinde kodlayan ve kodlamayan DNA dizilerinin istatistiksel özelliklerini araştırmışlardır. Bu çalışmanın sonucunda, *E. coli* genomu üzerinde, kodlayan bölgeler ile kodlamayan bölgeler arasında peşpeşe gelen nükleotidlerin gösterdiği ilişkinin birbirinden farklı olduğu gösterilmiştir. Bu ve benzeri bulgular gen bulucuların geliştirilmesinde uygun istatistiksel modellerin seçiminde yol gösterici olmuştur.

Bu çalışmayla genomik DNA dizisinin düzgün olmayan (nonuniform) Markov model ile modellenmesinin uygun olduğu ortaya konmuştur. Süreç içinde bu algoritmanın yapı ve parametrelerinin değiştirildiği pek çok farklı versiyonu otomatize gen bulma problemine çözüm olarak önerilmiştir. Bu algoritmalar GeneMark, GeneMark.hmm, Heuristic Models ve GeneMarkS'tir. Tüm bu algoritmalar "GeneMark Ailesi" olarak anılmaktadır. GeneMarkS, süreç içinde, otomatize gen bulma probleminin her bir basamağına özel çözümler sunan GeneMark algoritmalarının bir arada kullanıldığı bir yöntemdir. Algoritma yapısının daha iyi anlaşılabilmesini sağlamak üzere sonraki bölümde gizli Markov modelleri konusuna değinilmiştir.

2.5.1. Gizli Markov Modelleri

Gen bulma probleminin bir örüntü bulma problemi olarak ele alınması fikri 80'li yıllarda ortaya çıkmıştır. Örüntü bulma temel olarak, bir örüntünün otomatik olarak tanınması, bulunması, gruplanması ya da sınıflandırılmasını sağlayacak bir algoritmanın geliştirilmesidir. Örüntü bulma yöntemleri biyoloji, psikoloji, tıp, pazarlama gibi pek çok disiplinde, veri madenciliği, konuşma tanıma, parmak izi tanıma, yüz tanıma, el yazısı tanıma gibi işlemlerde sıklıkla kullanılmaktadır.

Örüntü tanıma algoritmaları problemin yapısına göre şekillendirilmektedir. Başlıca örüntü tanıma yaklaşımları şablonla eşleme (template matching), istatistiksel yaklaşım, sözdizimsel (syntactic) veya yapısal (structural) eşleme, sinir ağları olarak listelenebilir. Örüntü tanıma algoritmalarından başlıcaları istatistiksel yaklaşıma dayanmaktadır. İstatistiksel öğrenmeye dayalı örüntü bulma algoritmaları genel olarak verinin istatistiksel modellenmesine dayanmaktadır. Heuristic (sezgisel, deneysel) yaklaşımın tersine, istatistiksel model aracılığıyla, olasılık ve karar teorileri uygulanarak bir örüntü tanıma algoritması geliştirilmektedir. Örüntü tanıma ve sınıflandırma sisteminin iki basamaktan oluşan bir çalışma prensibi bulunmaktadır:

1. Öğrenme: Girdideki örüntüleri ifade edecek uygun özniteliklerin bulunması
2. Sınıflandırma, test etme: Girdide bulunan örüntünün öğrenme aşamasında eğitilen sınıflandırıcı tarafından ölçülen özniteliklere dayanarak örüntü sınıflarından birine atanması.

Örüntü tanımada, genel anlamda, bir ölçümün olasılıksal kaynağının belirlenmesi işlemi gerçekleştirilmektedir. Bir "örüntü", *d* öznitelikten oluşan bir küme ile ifade edilmekte ve *d*- boyutlu bir öznitelik vektörü olarak gözlenmektedir. Örüntüleın sınıflandırılmasıyla örüntü sınıfları elde edilmektedir. Örüntü sınıfları arasındaki geçişler, istatistiksel karar kuramında kullanılan konseptler kullanılarak belirlenmektedir. Bu işlemin gerçekleştirilmesinde ihtiyaç duyulan tek istatistiksel model, bir ölçü verildiğinde o ölçünün belirtilen sınıfa ait olması koşullu olasılığıdır. Bu koşullu model her sınıftaki ölçümlerin modellenmesiyle gizli Markov modelleri ya da Mixture modeller gibi bir tümleşik modelden elde edilebilir, ya da lojistik regresyon, genel lineer sınıflandırıcılar ve en yakın komşu (nearest- neighbour) gibi doğrudan öğrenilebilir (16).

Biyolojik dizi analizinde öne çıkan uygulamalardan biri her bir rezidünün ilgili belirteçle "etiketlenmesi"dir. Gen bulma problemi de basitçe bu şekilde ifade edilebilir: Bir DNA dizisi verildiğinde bu dizi üzerinde yer alan kodlayıcı bölgelerin başlangıç ve bitiş koordinatlarının belirlenmesi, yani kodlayan bölgelere "kodlayan", kodlamayan bölgelere de "kodlamayan" etiketlerinin verilmesi. Bunun yanısıra bir çoklu dizi hizalama problemi de bir etiketleme problemi olarak ele alınabilir. Burada da sorgulanan dizinin hedef veritabanı dizisindeki homolog bölgelerle ilişkilendirilmesi sözkonusudur.

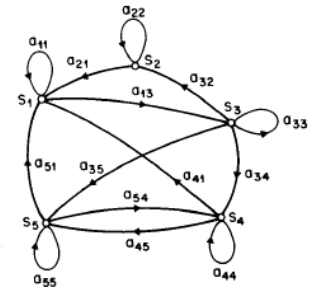
Bunlar gibi çeşitli problemler için *ad hoc* (soruna yönelik) programlar geliştirilebilir. Ancak bu tip problemlerin çözümüne yönelik ortak bazı sorunlar sözkonusudur. Bu sorunlardan biri, etiketleme probleminin çözümünde heterojen bilgi kaynaklarının birleştirilmek istenmesidir. Buna örnek olarak gen bulma probleminde içeriği tek bir skorlama sisteminde kombine olarak kullanılmak istenen bilgi kaynakları şöyle listelenebilir: ayıklama bölgesi (splice site), konsensus, kodon eğilimi, ekzon/intron uzunluğu tercihi ve ORF analizleri. Tüm bu verinin tek bir modelde birarada kullanılması çözüm getirilmesi gereken sorunlardan biridir.

İkinci bir sorun olarak sonuçların olasılıksal olarak değerlendirilmesinden bahsedilebilir. Elde edilen olası etiketlemelerin skorlanması ve bu skorlardan en iyisinin bulunması, geliştirilen bu skorlamanın ne anlama geldiğinin ifade edilmesi, bulunan "en iyi" skorun gerçekten en iyisi olup olmadığının anlaşılması da olasılıksal değerlendirmenin zorlukları olarak listelenebilir.

Diğer bir sorun da genişletilebilirliktir. Gen bulma probleminin çözümüne yönelik olarak geliştirilen gen bulucu "mükemmel" hale getirildiğinde aynı zamanda translasyon başlama konsensusu, alternatif ayıklama (splicing) ve poliadenilasyon sinyali gibi bileşenlerin de modellenmiş olması umulur. Ancak bu kadar çok beklenti gen bulucunun performasını yükseltmektense düşürmektedir.

Bahsedilen bu yöntemlerden dolayı gizli Markov modelleri, lineer dizi etiketleme problemlerinin olasılıksal modellerini oluşturmada kullanılan temel yaklaşımlardır. Özellikleri gereği gen bulma işleminde yukarıda bahsi geçen üç problemin giderilmesinde etkili olabilmektedir. Dolayısıyla gizli Markov modelleri gen bulma, profil arama, çoklu dizi hizalama, düzenleyici bölge tanımlama gibi işlemlerde kullanılan programların temelinde yer almaktadır (13).

Genel olarak bir Markov modeli, Markov özelliğini varsayan bir stokastik modeldir. Bir stokastik süreçte eğer gelecekteki durumların (yapıların) gözlenme koşullu olasılık dağılımları dizide gözlenen önceki durumlara (yapılara) değil, yalnızca mevcut duruma (yapıya) bağlı ise, o süreç Markov özelliğini gösteriyordur. Özetle Markov özelliğini gösteren bir süreçte bir "sembol"ün gözlenme olasılığı yalnızca bir önceki sembole bağlıdır. Şekil 2.15'te beş durumlu bir Markov modeli ve seçilmiş durum geçişleri gösterilmektedir (28).



Şekil 2.15 Beş durumlu (S1,... , S5) bir Markov modeli ve seçilmiş durum geçişlerinin (a11, ..., a55) şematik gösterimi 77

$S_1, S_2, \dots, S_N$ süreçte gözlenen durumlar (yapılar) olsun. $t = 1,2, \dots$ zaman değişkeni, q_t t anında gözlenen state, a_{ij} geçiş olasılıkları olmak üzere, birinci dereceden bir Markov modeli,

$$P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, \dots)$$
$$P(q_t = S_j \mid q_{t-1} = S_i)$$

Formül 2.1 Birinci dereceden Markov modeli

olarak ifade edilmektedir. Geçiş olasılıkları da,

$$a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i)$$

$$1 \leq i, j \leq N$$

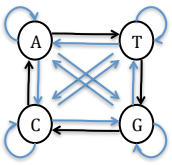
$$a_{ij} \geq 0$$

$$\sum_{j=1}^{N} a_{ij} = 1$$

Formül 2.2 Markov modelinde geçiş olasılıkları

özelliklerini sağlamaktadır.

DNA dizi verisi için benzer bir model (birinci derece homojen Markov modeli) durumlar A, T, G, C olmak üzere Şekil 2.16'daki gibi gösterilebilir.



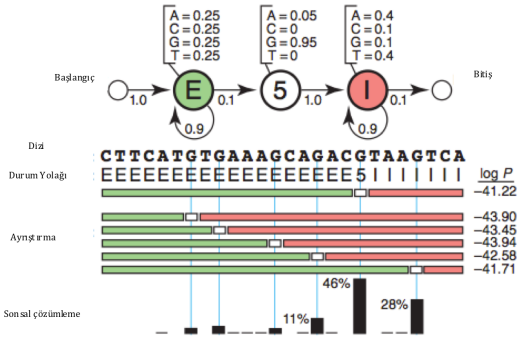
Şekil 2.16 DNA dizisinde kullanılabilecek örnek birinci derece homojen Markov modeli

Bir Markov sürecinin derecesi bir sonraki durumun (yapının) gözlenme olasılığının önceki kaç basamağa bağlı olduğu ile ifade edilir. *n.* dereceden bir Markov süreci aşağıdaki gibi ifade edilmektedir:

$$P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, \ldots, q_{1=} S_s) = P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, \ldots, q_{t-n=} S_p)$$

Formül 2.3 n. dereceden Markov süreci

Basit bir örnek olarak aşağıdaki 5' ayıklama bölgesi (splice site) tanıma problemini ele alalım. (Ayıklama bölgesi kabaca bir gen üzerinde ekzon ve intronun birbirinden ayrıldığı bölge olarak tanımlanabilir.) Elimizde bir ekzonla başlayan, 5' ayıklama bölgesi olan ve bir intronla biten bir DNA dizisi olduğunu varsayalım. Problem ekzondan introna geçişin nerede olduğunu yani 5' ayıklama bölgesinin konumunu belirlemek olsun (13).



Şekil 2.17 5' ayıklama bölgesi tanıma problemi

Daha önce de gösterildiği gibi, ekzonların, intronların ve ayıklama bölgelerinin istatistiksel özellikleri farklıdır. Bu istatistiksel farklılıklardan basit olanları örneklendirelim. Diyelim ki ekzonların baz kompozisyonu uniform (yaklaşık herbir baz %25), intronlar A/T oranı yönünden zengin (%40 A/T, %10 G/C), ve 5’ ayıklama bölgesi konsensus nükleotidi neredeyse her zaman G’dir (diyelim ki %95 G, %5 A) (13).

Bu bilgilerden yola çıkılarak bir gizli Markov model oluşturulması mümkündür. Bu gizli Markov modeli her bir nükleotide atanacak etiketler için 3 durumdan (yapıdan) oluşsun: E (ekzon), 5 (5’SS), ve I (intron). Her bir durumun (yapının) yani ekzonların, intronların ve 5’ayıklama bölgesindeki G’nin baz kompozisyonunu özetleyen kendine özgü nükleotid dağılım olasılıkları vardır (Bkz. Şekil 2.17). Bunun yanısıra herbir durumun geçiş olasılıkları (oklarla gösterilmiş) vardır. Bu durumlar arası geçişlerin hangi olasılıklarla gerçekleşeceğini göstermektedir. Geçiş olasılıkları durumların gerçekleşmesini beklediğimiz lineer sırayı tarif eder (13).

Modelde “gizli” olanın ne olduğunun daha iyi ifade edebilmesi için bir gizli Markov modelinin bir dizi “ürettiğini” hayal edelim. Bir duruma (yapıya) gelindiğinde, o durumda (yapıda) yer alabilecek nükleotidlerin olasılık dağılımından bir sonraki nükleotidin A, T, G, C’den hangisi olabileceği seçilir. Ardından geçiş olasılıkları dağılımından bir sonraki basamakta hangi duruma geçileceği seçilir. Böylece model iki bilgi dizisi açığa çıkarır. Bunlardan biri durumdan duruma (yapıdan yapıya) geçerken beliren durum yolağı (etiketlerden oluşan-Şekil 2.17) diğeri de gözlenen DNA dizisidir ve diğer bir deyişle durum yolağından olasılıksal olarak elde edilen nükleotidlerdir (13).

Durum yolağı (state path) bir Markov zinciridir (Şekil 2.17). Yani bir sonraki adımda hangi duruma (yapıya) gidileceği yalnızca içinde bulunulan yapıya bağlıdır. Elimizde yalnızca gözlenen dizi olduğundan altta yatan bu dizi gizlidir. Elde edilmeye çalışılan nükleotid etiketleri bunlardır.

Gen bulma probleminde verilen DNA dizisi üzerinden etiketleri içeren gizli durum yolağı elde edilmek istenir. Aynı diziyi oluşturacak pek çok durum yolağının bulunması olasıdır. Burada problem bu olası yolaklar arasından en yüksek olasılıkla doğru etiketleri gösteren yolağın bulunmasıdır. Bu problemde genelde herbirinin skorunun hesaplanamayacağı kadar çok sayıda olası yapı dizisi bulunmaktadır. Çok sayıda olası durum yolağı arasından en iyisinin bulunması işleminin gerçekleştirilmesinde Viterbi algoritması kullanılmaktadır (28).

Viterbi algoritmasının bulduğu sonuçları doğrulamak üzere ileriye (forward) ve geriye (backward) doğru algoritmaları kullanılarak posteriyör/geriye dönük şifre çözümü (posterior decoding) işlemi gerçekleştirilir. Örnekte bu işlemle elde edilen sonuçlar altta bar grafiği ile gösterilmiştir. (Bkz. Şekil 2.17)

2.5.2. GeneMark Algoritma Serisi

GeneMark Algoritma Serisi, hesaplamsal yöntemlerle gerçekleştirilen gen bulma işleminde karşılaşılan etkinlik ve hassasiyet gibi sorunlara çözüm üretmek üzere birbirini temel alarak geliştirilmiş olan bir dizi gen bulma algoritmasıdır. GeneMark serisi içinde markov zincir modellerini kullanan GeneMark (8), GeneMark'ın gizli Markov modellerine uyarlanmasıyla elde edilen GeneMark.hmm (22), bu algoritmalarda kullanılan öğrenme kümesi, eğitici küme (training set) ile ilgili sorunu gidermek üzere geliştirilen Heuristic Models (6) ve GeneMarkS (7) algoritmaları yer almaktadır. GeneMarkS, tüm GeneMark serisi bileşenlerinin gen bulma işleminin uygun noktalarında en uygun şekilde kullanılmasını sağlayan bir algoritmadır.

1993 yılında Borodovski *ve ark.* tarafından geliştirilen GeneMark, genom projelerinde kullanılabilecek kadar etkin ve hassas olduğu kabul edilen ilk gen bulma algoritmasıdır. GeneMark tüm genom dizilenmesi gerçekleştirilen ilk bakteri olan *Haemophilus influenzae* ve tüm genom dizisi elde edilen ilk arkea olan *Methanococcus janaschii* genom projelerinde kullanılmıştır.

GeneMark algoritması nükleotid dizisi üzerindeki protein kodlayan bölgeler için türe özel homojen olmayan 3 periyodik 5. mertebeden Markov zincir modellerini, kodlamayan bölgeler için ise homojen Markov zincir modellerini kullanmaktadır. Yani, okuma çerçevesini de göz önünde bulunduran GeneMark, bir bazın kodlayan ya da kodlamayan dizide görülme olasılığını hesaplarken o bazdan önce gelen 5 bazı kullanmaktadır.

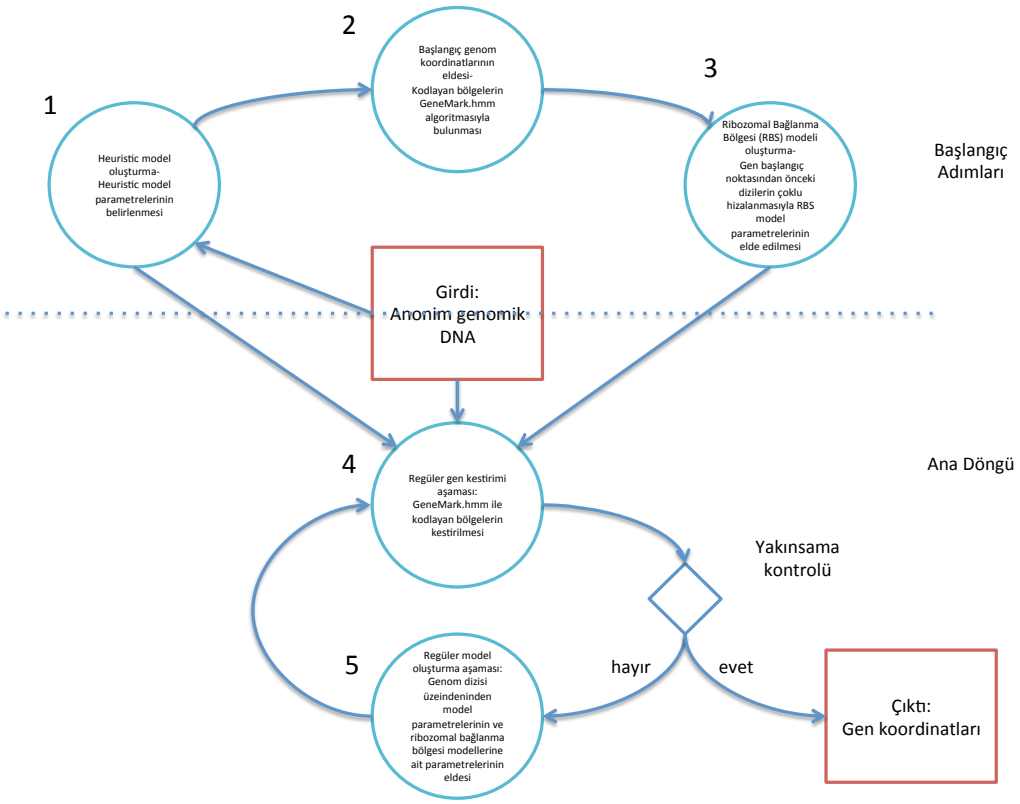
GeneMark'ın otomatize gen bulma işlemine getirdiği bir yenilik de, gen bulma işlemini gerçekleştirirken her iki DNA iplikçiği üzerindeki dizileri eş zamanlı olarak değerlendirmesidir. Bunu yaparken, kodlayan dizinin komplementeri olan "gölge dizi" üzerinde de homojen olmayan Markov modeli çalıştırmaktadır. Ardından elde edilen kodlayıcı bölge olma olasılığı olan dizilere birer skor vermektedir. Böylece daha önce geliştirilen yöntemlerin her iki iplikçik üzerinde ayrı ayrı gerçekleştirdiği gen bulma işlemi sonucu elde edilen hatalı sinyaller giderilmiş olmaktadır (8).

Tüm bu özelliklerine rağmen, Markov modeli tabanlı çalışan GeneMark, özellikle başlangıç ve bitiş kodonlarının hassas kestiriminde başarılı olamamaktadır. Başlangıç ve bitiş bölgelerinin doğru kestirilmesi, otomatize gen bulma probleminde önemli bir ölçüttür. Bu problemin giderilmesinde kodlayıcı bölgelerin sınırlarını "durum"lar arası geçişler olarak modelleyen gizli Markov modelleri kullanılmıştır. Böylece GeneMark modellerinin gizli Markov modeli çerçevesine oturtulmasıyla daha etkin bir algoritma olan GeneMark.hmm algoritması geliştirilmiştir (22).

GeneMark.hmm algoritmasının çekirdeği Viterbi algoritmasıdır. Bunun yanısıra gen başlangıçlarının daha hassas olarak belirlenmesini sağlamak amacıyla algoritmada ribozom bağlanma bölgesi modeli de kullanılmıştır. Bu yeni yaklaşım sayesinde gen bulmada GeneMark'tan daha hassas bir araç elde edilmiştir.

GeneMark'ın ve ardından GeneMark.hmm'nin geliştirilmesinin ardından, modellerin oluşturulmasında kullanılan "özellikleri bilinen dizi" kısıtlamasının giderilmesini sağlamak üzere, GeneMark'ta ve GeneMark.hmm'de kullanılacak model parametrelerinin dizi üzerinden öğrenilmesine olanak sağlayan "Gen Bulma İçin Model Oluşturmada Heuristic Modeller" yaklaşımı geliştirilmiştir (6).

Bu gelişme özellikle *de novo* genom projelerinin anotasyon aşamaları için büyük önem taşımaktadır. Çünkü elde edilen genom eğer daha önce genom dizilenmesi hiç yapılmamış yeni bir türe ait ise, diğer genomlarla benzerliği diğer genomlardan elde edilen bilgi ile yeni genom üzerindeki genleri hassasiyetle konumlandırmaya yarar bir modelin oluşturulmasına elverişli olmamaktadır. Dolayısıyla algoritmanın öğrenim kümesi olarak üzerinde gen aradığı genomun kendi parçalarını kullanması, yöntemin dış veriye bağımlılığını ortadan kaldırmada önemli bir rol oynamaktadır. Tez çalışması kapsamında GeneMarkS'in Heuristic Models ile kombine edildiği ve gen bulmada GeneMark.hmm'nin çalıştırıldığı bir sistem kullanılmıştır.



Şekil 2.18 GeneMarkS algoritmasının çalışma prensibi

2.5.3. GLIMMER

GLIMMER, Bölüm 2.5.3'te bahsedilen GeneMark algoritmasında kullanılan beşinci mertebeden Markov zincir modeline alternatif olarak İnterpole Markov Modeli'ni kullanan, benzer bir gen bulma algoritmasıdır (29). GeneMark'ın yaklaşık olarak tüm özelliklerini taşıyan algoritma, GeneMark'ın sabit olarak kullandığı beşinci mertebeden Markov zincir modeli yerine, sekizinci mertebeye kadar tüm olasılıkların hesaplandığı ve ağırlıklandırılarak bir arada değerlendirildiği bir sistem kullanmaktadır.

GLIMMER'ın GeneMark'tan farklı olan diğer önemli bileşeni de öğrenme algoritmasıdır. GLIMMER dizi üzerinde yer alan ORF'leri analiz etmekte, daha sonra da kodlayıcı olma potansiyeli yüksek olan belirli bir uzunluğun üzerindeki ORF'ler ile model parametrelerini elde etmektedir. Bu sistem GeneMark.hmm ve GeneMarkS ile kullanılan Heuristic Models yaklaşımından çok daha temel bir yaklaşımdır.

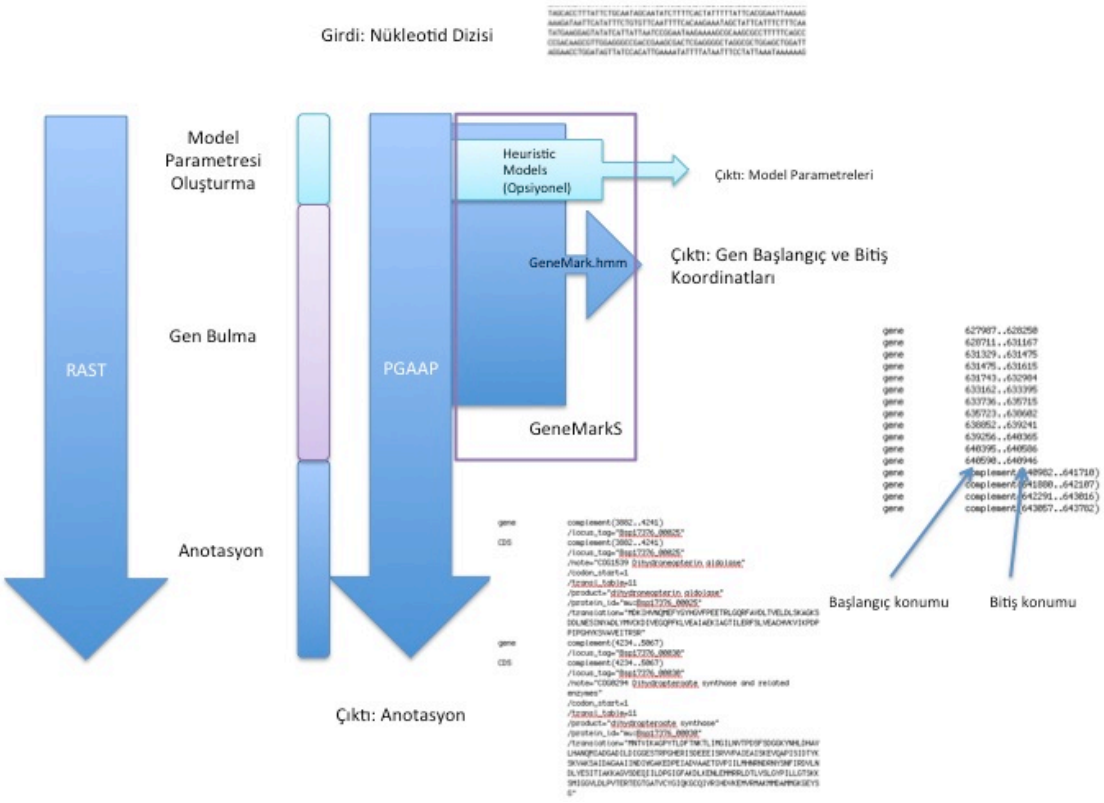
Glimmer algoritmasının performansı GeneMark'tan daha iyi, ancak GeneMarkS'ten daha düşüktür. (7, 29) Bu algoritma pek çok genom projesinde kullanılmıştır. Bunun yanısıra RAST anotasyon akış hattının çekirdek algoritması olarak da kullanılmaktadır.

2.5.4. Anotasyon Akış Hatları

Gen bulma işleminin ardından gerçekleştirilen anotasyon işlemi, konumları belirlenen fonksiyonel bölgelerin biyolojik bağlamda anlamlandırılmasını sağlamaktadır. Bu işlem fonksiyonel dizi parçalarının veri tabanlarında yer alan anotasyon bilgisi içeren dizilerle karşılaştırılarak isimlendirilmesi sürecidir.

Genom üzerinde yer alan fonksiyonel bölgelerin keşfinde hesaplamsal yöntemlerin ve homoloji aramasının bir arada birbirini destekleyici şekilde kullanılmasının daha verimli olduğunun belirlenmesiyle, bu işlemlerin bir kendine özgü bir mantık çerçevesinde bir arada gerçekleştirilmesini sağlayan akış hatları tasarlanmıştır.

Prokaryotlarda anotasyon işleminin gerçekleştirilmesini sağlayan pek çok akışhattı geliştirilmiştir. Bunlardan en önemlileri olarak NCBI (National Center for Bioinformatics Information) tarafından geliştirilmiş olan ve altyapısında GeneMark serisini kullanan PGAAP (Prokaryotic Genomes Automated Annotation Pipeline) ile NMPDR (National Microbial Pathogen Data Resource) tarafından geliştirilmiş olan RAST (Rapid Annotation using Subsystem Technology) sıralanabilir . Şekil 2.19'da PGAAP ve RAST anotasyon akış hatları ile gen bulma sürecinin algoritmik bileşenleri, girdi ve çıktıları gösterilmektedir.



Şekil 2.19 PGAAP ve RAST anotasyon akış hatları ile gen bulma süreci

Bir gen bulucu yalnızca verilen DNA dizisi üzerindeki kodlayan bölgelerin koordinatlarını bulmaktadır. Bir akış hattı için gen bulma işlemi anotasyon sürecinin temel bir parçasıdır. Anotasyon sürecinde kodlayıcı bölgelerin koordinatları elde edildikten sonra, bulunan nükleotid dizisinden aminoasit dizisi elde edilmekte ve bu diziler protein veritabanları ile karşılaştırılmaktadır. Böylece yeni dizilenmiş olan DNA dizisi üzerinde yer alan olası kodlayıcı bölgelere, homoloji bilgisinden faydalanılarak olası görevler atanmış olur. Bu da verinin biyolojik anlamlandırılmasında en önemli başlangıç basamaklarından biridir. Bahsi geçen otomatize isimlendirilmelerin, yani anotasyonun doğrulanması hem ıslak laboratuvar çalışmaları ile hem de veri tabanlarındaki küratörlerce gerçekleştirilmektedir. Bunun yanısıra anotasyon akış hatları devamlı olarak geliştirilmekte ve veri tabanlarında yer alan DNA dizileri periyodik olarak geliştirilmiş algoritmalarla tekrar tekrar anotasyon işlemine tabi tutulmaktadır.

3. GEREKÇE ve AMAÇ

Teknolojik gelişmeler doğrultusunda günümüzde *de novo* genom dizileme işlemi, dolayısıyla da *de novo* genom projeleri, Biyoteknoloji Enstitüsü Merkez Laboratuvarı gibi orta-küçük işlem hacimli laboratuvarlarda dahi gerçekleştirilebilir hale gelmiştir. Elde edilen genom dizisi verisinin biyolojik bağlamda anlamlandırılması işleminde hesaplamsal yöntemlerden yararlanılması, hızlı ve ucuz elde edilen dizi verisinin değerlendirilmesindeki en önemli basamaklardan biridir. Bahsi geçen değerlendirme işlemi, genom üzerindeki işlevsel bölgelerin koordinatlarının belirlenmesi ve bu koordinatlarda yer alan dizi parçalarına biyolojik görevlerin atfedilmesi basamaklarından oluşmaktadır. Ancak kullanılan hesaplamsal yöntemler çok basamaklı yazılımsal ve istatistiksel yöntemlere dayalıdır. Bu nedenle genom verisini elde eden araştırmacıların, özellikle *de novo* veri üzerindeki hakimiyetlerini arttırmak amacıyla en azından gen bulma algoritmalarının nasıl çalıştığını bilmesi önem taşımaktadır.

Tez çalışması kapsamında, Bacillus boroniphilus genomunun *de novo* dizilenmesi ile elde edilen dizisi üzerinde çeşitli gen bulma yöntemleri uygulanmış ve bu yöntemlerin uygulanmasıyla elde edilen fonksiyonel bölge koordinatları karşılaştırılarak elde edilen sonuçların karşılaştırılması amaçlanmıştır. Böylelikle:

- BTEML'de *de novo* dizilenmesi gerçekleştirilmiş *Bacillus boroniphilus* genomu üzerinde gen bulma ve anotasyon işlemleri gerçekleştirilmiştir. Bu bilgi sayesinde *Bacillus boroniphilus*'a özgü özelliklerin kaynağı olabilecek genomik özelliklerin araştırılabilmesini sağlayacak ön bilginin elde edilmesi,
- Gen bulma işleminde bir algoritmanın tek başına kullanılması ile bir akışhattı aracılığıyla anotasyon verisinin gen bulma işlemine dahil edilmesinin sonuçlar üzerindeki etkisi gözlenmesi,
- *Bacillus boroniphilus* genomu üzerinde gizli Markov modeli tabanlı bir yöntemle gerçekleştirilen gen bulma işlemi ile anotasyona dayalı gen bulma işlemi ile elde edilen sonuçlar arasındaki farklılıkların ortaya çıkarılmasıyla, hem sonuçların tutarlılığının incelenmesi hem de sonuçların kapsamının genişletilmesi,

- Gizli Markov modeli tabanlı gen bulma algoritmalarının çalışma yönteminin incelenmesi ve algoritmayla ilgili temel bileşenlerde gerçekleştirilen seçimlerin gen bulma sürecini nasıl etkilediğinin gözlenmesi amaçlanmıştır.

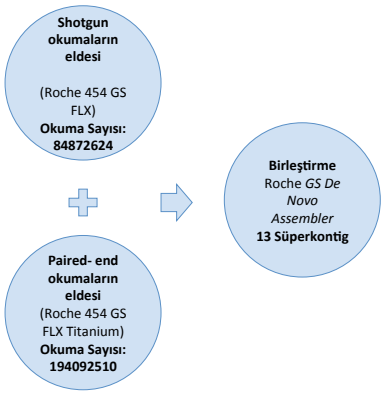
Bu amaç ve gerekçeler doğrultusunda sonraki bölümlerde detaylarından bahsedilen uygulamalar gerçekleştirilmiş, bulgular sunulmuş ve tartışılmıştır.

4. MATERYAL VE YÖNTEM

4.1. Materyal - Genom Dizileme Verisi

Tez çalışmasında *Bacillus boroniphilus* genom projesi kapsamında Roche 454 GS FLX Titanium cihazı ile elde edilen okumaların GS De Novo Assembler yazılımı kullanılarak birleştirilmesi ile elde edilen nükleotid dizisi kullanılmıştır. *Bacillus boroniphilus* genom projesi bir *de novo* genom dizileme projesidir. Bu bakteri diğer canlıların yaşam faaliyetlerini sürdüremediği yüksek bor konsantrasyonlarında varlığını sürdürebilen, bunun yanında metabolik aktiviteleri için ortamda bor varlığına ihtiyaç duyan bir bakteri türüdür (1). Bu tip bakteriler ekstremofil bakteriler olarak anılmaktadır. Ekstremofil bakteriler olağanüstü doğa koşullarında (kimyasalların yüksek konsantrasyonları, aşırı sıcak, aşırı soğuk, yüksek basınç vb.) varlıklarını sürdürebilen canlılar olduğundan, bu canlıların genom dizilerinden elde edilen bilgiler biyoteknolojik ürün geliştirilmesinde (ilaç, endüstriyel kimyasallar, tanısal ürünler vb.) sıklıkla kullanılmaktadır.

Genomun *de novo* dizilenmesinde birleştirme işleminin veriminin arttırılmasına yönelik olarak iki farklı yöntem kullanılmış ve böylece iki tip okuma elde edilmiştir. Bu yöntemler shotgun dizileme ve paired-end dizilemedir. Bölüm 2.2.1 ve Bölüm 2.2.2'de detaylarıyla bahsedilen bu yöntemlerle elde edilen okumalar tez kapsamında sırasıyla "shotgun okumaları" ve "paired-end okumaları" olarak isimlendirilmiştir. Tez çalışmasında kullanılan verinin eldesinde izlenilen yöntem anahatlarıyla Şekil 2.6'da özetlenmiştir. Bu yöntem son derece karmaşık ıslak laboratuvar protokollerine dayalı olduğundan tezde işlemin detaylarına değinilmemiştir.



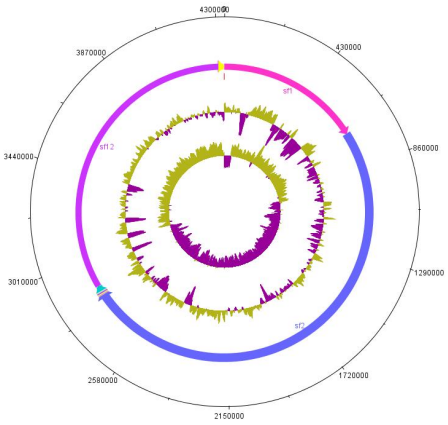
Şekil 4.1 Dizileme işleminin en genel hatlarıyla şematik gösterimi

Bacillus boroniphilus genom projesi kapsamında yürütülen shotgun ve paired-end dizileme işlemleri sonucunda toplam 345022 okuma ve 177891982 baz elde edilmiştir. Projenin birleştirme işleminde kullanılan okumalardan 346022'i shotgun, 502393'ü paired-end dizilemeyle elde edilmiştir. Shotgun okumaları toplam 847652235, paired-end okumaları ise toplam 93126747 baz içermektedir.

Çizelge 4.1 Dizileme ile elde edilen veri

Toplam Okuma Sayısı	848415
Shotgun Okumalarının Sayısı	346022
Paired-end Okumalarının Sayısı	502393
Toplam Baz Sayısı (bç)	177891982
Shotgun Dizileme İle Elde Edilen Baz Sayısı (bç)	847652235
Paired-end Dizileme İle Elde Edilen Baz Sayısı (bç)	93126747

Shotgun ve paired-end dizileme yöntemleriyle elde edilen okumalar GS De Novo Assembler yazılımı kullanılarak birleştirilmiştir. Bu yazılım Roche 454 GS FLX cihazından elde edilen *de novo* dizileme okumalarının "taslak" genom dizisini oluşturmak üzere etkin şekilde birleştirilmesi işlemine özel olarak geliştirilmiş, bu verinin birleştirme işlemini en yüksek verimle gerçekleştiren yazılımdır. Birleştirme işlemi sonucu 254 kontig ve 13 süperkontig elde edilmiştir. Şekil 4.3'te *B. Boroniphilus* genomunu oluşturan süperkontigler genom atlası formunda görülmektedir. Dıştaki çizgiler süperkontiglerin uzunluklarını, içteki daireler sırasıyla GC içerik ve biasını göstermektedir.



Şekil 4.2 *B. boroniphilus* genomunu oluşturan süperkontigleri gösteren genom atlası

Çizelge 4.3'te dizileme sonucu elde edilen süperkontiglere ilişkin istatistikler verilmiştir. Buna göre birleştirme işlemi sonucu 254 kontig, bu verinin paired-end verisiyle birleştirilmesiyle 13 süperkontig elde edilmiştir. Tahmini genom büyüklüğü 5.4Mb'dir. Genomun %GC içeriği %42 olarak hesaplanmıştır.

Çizelge 4.2 Dizileme sonucu elde edilen kontiglere ilişkin istatistikler

Elde Edilen Kontig Sayısı	254
Elde Edilen Süperkontig Sayısı	13
Tahmini Genom Büyüklüğü	5.4 MB
%GC İçeriği	%42

Birleştirme işlemi sonucunda başka kontiglerle birleştirilememiş olan kontigler de tek başlarına birer süperkontig olarak değerlendirilmektedir. Süperkontig içinde yer alan kontiglerin sayısı, süperkontig içinde yer alan boşlukların sayısı hakkında bilgi vermektedir. Çizelge 4.4'te elde edilen süperkontiglerden her birinin boyutu ve süperkontiglerin içerdiği kontig sayısı görülmektedir.

Çizelge 4.3 Süperkontig uzunluk ve kontig içerikleri

Süperkontig no.	Uzunluk (bç)	İçerdiği Kontig Sayısı
Süperkontig01	772038	36
Süperkontig02	2300059	70
Süperkontig03	2910	1
Süperkontig04	2916	1
Süperkontig05	2274	1
Süperkontig06	3056	1
Süperkontig07	2360	1
Süperkontig08	2768	1
Süperkontig09	2483	1
Süperkontig10	2641	1
Süperkontig11	2058	1
Süperkontig12	1552928	62
Süperkontig13	2425	1
Toplam	4650916	178

Şekil 4.3'teki genom atlasında da görüldüğü gibi, tahmini boyutu 5.4 MB olan genomun dizilenmiş olan kısmı 4650916 bç.'dir. Genomun dizilenen en büyük parçaları birinci, ikinci ve onikinci süperkontiglerdir. En büyük süperkontig olan ikinci süperkontig 1552928 bç., ikinci büyük süperkontig olan onikinci süperkontig 1552928 bç. ve üçüncü büyük süperkontig olan birincisüperkontig 772038 bç. uzunluğundadır. Gizli Markov modeli tabanlı gen bulma algoritmaları *Bacillus boroniphilus* genomunun yukarıda belirtilen yöntemlerle elde edilen ve yukarıda belirtilen özelliklere sahip nükleotid dizisi üzerinde uygulanmıştır.

4.2. Yöntem

de novo dizilenme işlemi gerçekleştirilmiş *Bacillus boroniphilus* genomu üzerinde Markov modeli tabanlı gen bulma algoritmalarının uygulanmasında Şekil 4.3'te gösterilen aşağıda listelenen işlem basamakları izlenmiştir:

I. Genom dizisi üzerinde gen koordinatlarının bulunması:

GeneMarkS Combined + GeneMark.hmm, GLIMMER algoritmaları ve *PGAAP, RAST* analiz akış hatları ile her bir büyük süperkontig üzerinde gen bulma işlemi gerçekleştirilmiş, sonuç olarak nükleotid dizisi üzerinde yer alan olası genlere ait başlangıç ve bitiş koordinatları elde edilmiştir.

II. Gen koordinatlarının karşılaştırılması:

I'de bahsedilen yöntemlerin üç büyük süperkontig üzerinde çalıştırılmasıyla elde edilen gen başlangıç ve bitiş koordinatları *"PGAAP vs. RAST", "PGAAP vs. (GeneMarkS Combined + GeneMark.hmm)", "RAST vs. GLIMMER", "(GeneMarkS Combined + GeneMark.hmm) vs. GLIMMER"* kombinasyonlarıyla karşılaştırıldıktan sonra elde edilen karşılaştırma sonuçları değerlendirilmiştir.

Gen koordinatlarının elde edilmesinde kullanılan yöntemlerden Bölüm 4.2.1'de, sonuçların karşılaştırılmasında kullanılan yöntemlerden Bölüm 4.2.2'de, gerçekleştirilen karşılaştırmalardan Bölüm 4.2.3'te bahsedilmektedir.

4.2. Yöntem

de novo dizilenme işlemi gerçekleştirilmiş *Bacillus boroniphilus* genomu üzerinde Markov modeli tabanlı gen bulma algoritmalarının uygulanmasında Şekil 4.3'te gösterilen aşağıda listelenen işlem basamakları izlenmiştir:

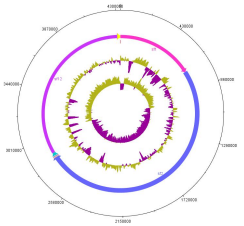
I. Genom dizisi üzerinde gen koordinatlarının bulunması:

GeneMarkS Combined + GeneMark.hmm, GLIMMER algoritmaları ve *PGAAP, RAST* analiz akış hatları ile her bir büyük süperkontig üzerinde gen bulma işlemi gerçekleştirilmiş, sonuç olarak nükleotid dizisi üzerinde yer alan olası genlere ait başlangıç ve bitiş koordinatları elde edilmiştir.

II. Gen koordinatlarının karşılaştırılması:

I'de bahsedilen yöntemlerin üç büyük süperkontig üzerinde çalıştırılmasıyla elde edilen gen başlangıç ve bitiş koordinatları *"PGAAP vs. RAST", "PGAAP vs. (GeneMarkS Combined + GeneMark.hmm)", "RAST vs. GLIMMER", "(GeneMarkS Combined + GeneMark.hmm) vs. GLIMMER"* kombinasyonlarıyla karşılaştırıldıktan sonra elde edilen karşılaştırma sonuçları değerlendirilmiştir.

Gen koordinatlarının elde edilmesinde kullanılan yöntemlerden Bölüm 4.2.1'de, sonuçların karşılaştırılmasında kullanılan yöntemlerden Bölüm 4.2.2'de, gerçekleştirilen karşılaştırmalardan Bölüm 4.2.3'te bahsedilmektedir.



Genom Dizisinin Elde Edilmesi

```
   1 TTCATTTTGT TGCTCTGTCA GAAGCAACTC TTATATATTA TCAGGTTTAA AACCCGATGT
  61 CAATAAGTTT TTTGTTTTCT TTGTTGCTAA CTAAGTAGCT CGCAGCAACA GATATAAATA
 121 TAGCACCTTT ATTCTGCAAT AGCAATATCT TTTCACTATT TTTATTCACG GAATTAAAAG
 181 AAAGATAATT CATATTTCTG TGTTCAATTT TCACAAGAAA TAGCTATTCA TTTCTTTCAA
 241 TATGAAGGAG TATATCATTA TTAATCCGGA ATAAGAAAAG CGCAAGCGCC TTTTTCAGCC
 301 CCGACAAGCG TTGGAGGGCC GACCGAAGCG ACTCGAGGGG CTAGGCGCTG GAGCTGGATT
 361 AGGAACCTGG ATAGTTATCC ACATTGAAAA TATTTTATAA TTTCCTATTA AATAAAAAAG
 421 GCAAAACGGC CTGTGCCGTT TTGCCTTTAC TCTATCTATT AACGGTGACG CATCAATGGG
 481 AATAATAGTA CGTCCCTGAT GGATGGTGAA TTCGTCAGCA GCATTACCAA TCGATCGATA
 541 CCGATTCCGA GTCCGCCAGT TGGAGGCATT CCATATTCCA GTGCTTCGAT GAAATCGTCG
 601 TCCATTTCGT GCGCTTCGTC ATTTCCTTGT TCTTTTTCTT TTAGCTGAGC TTCAAAACGC
 661 TCTCTTTGGT CGATCGGATC ATTAAGCTCA GTAAATGCGT TCGCATGCTC CCTTGCTACG
 721 ATGAACAATT CGAAACGGTC TGTAAAGCGT TGGTCTTCGT CATTCTTTTT CGCAAGAGGA
```

Gen Koordinatlarının Elde Edilmesi

```
gene            627987..628250
gene            628711..631167
gene            631329..631475
gene            631475..631615
gene            631743..632984
gene            633162..633395
gene            633736..635715
gene            635723..638602
gene            638852..639241
gene            639256..640365
gene            640395..640586
gene            640590..640946
gene            complement(640982..641710)
gene            complement(641880..642107)
gene            complement(642291..643016)
gene            complement(643057..643782)
```

Başlangıç konumu Bitiş konumu

Gen Koordinatları

	Model Oluşturma	Gen Bulma	Anotasyon
1	(GeneMarkS)	PGAAP	BLAST / Clusterprot*, Bactprot
7	(GLIMMER)	RAST	Mapping to “subsystems”**
3	GeneMarkS Combined	GeneMark.hmm	-
4	GLIMMER	GLIMMER	-

Gen Koordinatları

Başlangıç konumu Bitiş konumu

```
Gene            4270968..4271600
                /locus_tag="Bsp17376_22705"
CDS             4271776..4272528
                /locus_tag="Bsp17376_22710"
                /note="COG1349 Transcriptional regulators of
                sugarmetabolism"
                /codon_start=1
                /transl_table=11
                /product="HTH domain, deoR family
                transcriptionalregulator"
                /note="*protein_id: mu:Bsp17376_22710"
                /translation="MLTPEHHQLILQLIKEKPIVKIQDLVDLTEASESTIRRDLTILE
                EGKFLKRVHGGAARLMGKLQEPSMDEKSTKFLQEKKQIAQYAASIVEEGDSIYLDAGS
                TVFEIINSLQVKDIVVVTNGLMHLPQLLERNIETYVIGGYAKPKTNAIIGRGALASLE
                QYRFDKCFLGVNGIHPHSGYTTPDQEEAMIKQKAITLSRESFVLADDSKFSEITFAKI
                ADLHEATIITNAIDEDHKGQYTSKTSIKVVTT"
```

Gen Koordinatlarının Karşılaştırılması

1	PGAAP		vs.	RAST		
2	PGAAP		vs.	GeneMarkS Combined	+	GeneMark.HMM
3	RAST		vs.	GLIMMER	+	GLIMMER
4	GeneMarkS Combined	+ GeneMark.HMM	vs.	GLIMMER	+	GLIMMER

Model Oluşturma + Gen Bulma

Model Oluşturma

Gen Bulma

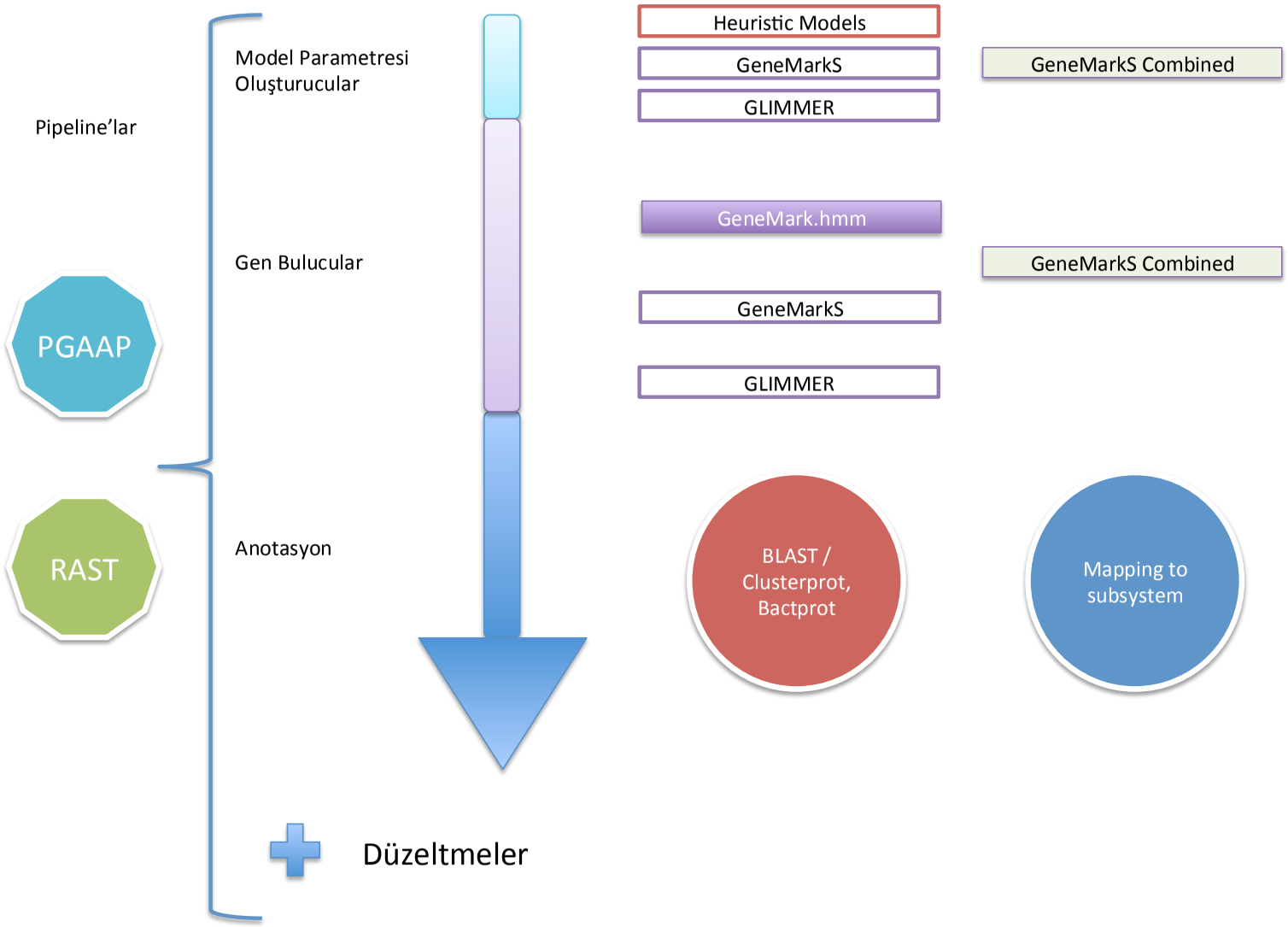
4.2.1. Gen Koordinatlarının Elde Edilmesinde Kullanılan Yöntemler

Tez çalışmasında Bölüm 4'te bahsedildiği ve Şekil 4.3'te de gösterildiği gibi gen koordinatlarının elde edilmesinde aşağıda listelenen yöntemler kullanılmıştır:

1. PGAAP
2. RAST
3. GeneMarkS Combined + GeneMark.hmm (GMSC + GMHMM)
4. GLIMMER

Markov modeli tabanlı gen bulma algoritmaları ile gen bulma işlemi, Bölüm 2.5'te detaylarıyla anlatıldığı gibi, "model parametrelerinin elde edilmesi" ve "gen bulma (gen koordinatlarını elde etme) işleminin gerçekleştirilmesi" olarak iki aşamadan oluşmaktadır. GeneMark.hmm gen bulma işlemini gerçekleştirmek için dışarıdan verilen model parametrelerine ihtiyaç duymaktadır (22). Bu nedenle model parametrelerinin eldesi aşamasında, ilk model parametrelerinin heuristic olarak dizinin kendisi üzerinden elde edilmesini sağlayan ve ardınan iteratif olarak bu modeli en iyileyen GeneMarkS Combined algoritması ile birlikte kullanılmıştır. Benzer şekilde GLIMMER da model parametresi eldesini, dizi üzerindeki uzun ORF'lerin analizi ile gerçekleştirmekte, ardından gen bulma işlemini yapmaktadır (7, 29).

GeneMarkS Combined algoritması, GeneMarkS algoritmasının Heuristic modeller algoritması ile birlikte kullanılan versiyonunu ifade etmektedir (6, 7). Heuristic modeller algoritması Bölüm 2.6.2'de de bahsedildiği üzere daha önce "benzeri" dizilenmemiş türlere ait nükleotid dizileri üzerinde gen bulma işlemi gerçekleştirilirken kullanılabilecek model parametrelerinin üretilmesini sağlamak üzere geliştirilmiş bir yöntemdir. Bu nedenle GeneMark.hmm ile birlikte kullanılmaktadır. *de novo* bir dizi için GeneMark serisinde Heuristic Models tarafından üretilmiş bir modelle başlanmaksızın etkin bir gen bulma işlemi yapmak mümkün değildir. PGAAP (Prokaryotic Genomes Automated Annotation Pipeline) çekirdek algoritma olarak GeneMarkS'i kullanmaktadır. RAST gen bulmada interpolated Markov modellerine dayanan GLIMMER algoritmasını kullanmaktadır (2, 3).



Şekil 4.4 Anotasyon akışhatları ve algoritmaların süreçteki işlevleri

4.2.1.1. PGAAP

Bir genom anotasyonu akış hattı olan PGAAP, gen bulmada çekirdek algoritma olarak GeneMark serisindan algoritmalar kullanmaktadır. Model parametrelerinin eldesinde GeneMarkS'i, gen bulmada ise GeneMark.hmm'i kullanan PGAAP, dizi girdisi alıp GenBank formatında anotasyonu tamamlanmış dizi çıktısı vermektedir. PGAAP ile *B. Boroniphilus* üzerindgen bulma işlemi, GeneMarkS Combined ile GeneMark.hmm (2.6m) kullanılarak gerçekleştirilmektedir (2, 7).

```
LOCUS       scaffold00001         691422 bp    DNA     linear   BCT 02-JUN-2011
FEATURES             Location/Qualifiers
     source          1..691422
                     /organism="Bacillus sp. 17376"
                     /mol_type="genomic DNA"
                     /strain="17376"
                     /db_xref="taxon:977905"
     gene            complement(459..1943)
                     /gene="lysS"
                     /locus_tag="Bsp17376_00005"
     CDS             complement(459..1943)
                     /gene="lysS"
                     /locus_tag="Bsp17376_00005"
                     /EC_number="6.1.1.6"
                     /note="COG1190 Lysyl-tRNA synthetase (class II)"
                     /codon_start=1
                     /transl_table=11
                     /product="lysyl-tRNA synthetase"
                     /protein_id="mu:Bsp17376_00005"
                     /translation="MSHEELNDQLRVRRDKMSSLREKGMDPFGKRFERTHLTEELIGE
                     YGELEKEEIEAKNVSVKIAGRIMTKRGKGKAGFAHIQDLAGQIQIYIRQDAVGEEQYE
                     VFDSADLGDIIGIEGTLFKTKVGELSIKAQDFVFLTKALRPLPEKFHGLKDVEQRYRQ
                     RYLDLITSNDSKTTFINRSRIIQSMRRYLDGQGYLEVETPLMHSIAGGASARPFITHH
                     NALDMQLYMRIAIELHLKRLIVGGLEKVYEIGRVFRNEGVSTRHNPEFTMLELYEAYA
                     DWRDIMSLTENMVAYIAQDVLGATTIQYGEYEIDLKPEWKRVHMVDAIKEYTGVDFWP
                     QMSTEEARALAKEHGVEITEHMQYGHIVNEFFEQKVEEHLIQPTFIYGHPVDISPLAK
                     KNDEDQRFTDRFELFIVAREHANAFTELNDPIDQRERFEAQLKEKEQGNDEAHEMDDD
                     FIEALEYGMPPTGGLGIGIDRLVMLLTNSPSIRDVLLFPLMRHR"
```

Şekil 4.5 Süperkontig01'de bulunmuş bir gen için örnek PGAAP çıktısı

4.2.1.2. RAST

Bir genom anotasyonu akış hattı olan RAST, gen bulmada çekirdek algoritma olarak Markov modeli tabanlı bir algoritma olan GLIMMER'ı kullanmaktadır. RAST çıktı olarak anotasyonu tamamlanmış GenBank formatında dizi çıktısı vermektedir (3, 29).

```
LOCUS       gnl|mu|scaffold00001       691422 bp    DNA     linear   UNK
DEFINITION  Contig gnl|mu|scaffold00001 from Bacillus boroniphilus 17376
ACCESSION   unknown
FEATURES             Location/Qualifiers
     source          1..691422
                     /mol_type="genomic DNA"
                     /db_xref="taxon: 308892"
                     /genome_md5=""
                     /project="tortoiseegg_308892"
                     /genome_id="308892.4"
                     /organism="Bacillus boroniphilus 17376"
     CDS             complement(459..1943)
                     /db_xref="GO:0004824"
                     /translation="MSHEELNDQLRVRRDKMSSLREKGMDPFGKRFERTHLTEELIGE
                     YGELEKEEIEAKNVSVKIAGRIMTKRGKGKAGFAHIQDLAGQIQIYIRQDAVGEEQYE
                     VFDSADLGDIIGIEGTLFKTKVGELSIKAQDFVFLTKALRPLPEKFHGLKDVEQRYRQ
                     RYLDLITSNDSKTTFINRSRIIQSMRRYLDGQGYLEVETPLMHSIAGGASARPFITHH
                     NALDMQLYMRIAIELHLKRLIVGGLEKVYEIGRVFRNEGVSTRHNPEFTMLELYEAYA
                     DWRDIMSLTENMVAYIAQDVLGATTIQYGEYEIDLKPEWKRVHMVDAIKEYTGVDFWP
                     QMSTEEARALAKEHGVEITEHMQYGHIVNEFFEQKVEEHLIQPTFIYGHPVDISPLAK
                     KNDEDQRFTDRFELFIVAREHANAFTELNDPIDQRERFEAQLKEKEQGNDEAHEMDDD
                     FIEALEYGMPPTGGLGIGIDRLVMLLTNSPSIRDVLLFPLMRHR"
                     /product="Lysyl-tRNA synthetase (class II) (EC 6.1.1.6)"
                     /EC_number="6.1.1.6"
```

Şekil 4.6 Süperkontig01'de bulunmuş bir gen için örnek RAST çıktısı

4.2.1.3. GeneMarkS Combined ve GeneMark.hmm

GeneMarkS Combined algoritması, GeneMarkS'in iterasyonun ilk adımında model parametrelerini Heuristic modeller algoritması ile elde edilmiş bir modeli kullandığı bir algoritmayı ifade etmektedir. Heuristic modeller algoritması kullanıcıya daha önce çeşitli GC içerikleri ve çeşitli genetik kod niteliklerine göre elde edilmiş bir model parametresi koleksiyonu sunmaktadır. Kullanıcı bu koleksiyondan kendi genomunun özelliklerine uygun olan modeli seçerek GeneMarkS ile birlikte kullanmaktadır (6, 7).

Öncelikle GeneMarkS (4.6m) ve Heuristic modeller kullanılarak en uzun süperkontig olan SK12 üzerinde model elde edilmiş, ardından GeneMark.hmm (2.6p) ile gen bulma işlemi tüm süperkotiglerde gerçekleştirilmiştir.

```
LOCUS       gnl|mu|scaffold00      691422 bp
FEATURES             Location/Qualifiers
     gene            complement(459..1943)
     gene            complement(2153..3160)
     gene            complement(3197..3400)
     gene            complement(3352..3879)
     gene            complement(3882..4241)
     gene            complement(4234..5067)
     gene            complement(5064..5939)
     gene            complement(5939..6529)
     gene            complement(6526..7923)
     gene            complement(8041..8229)
     gene            complement(8258..9184)
     gene            complement(9442..10326)
     gene            complement(10348..11226)
     gene            complement(11280..12107)
     gene            complement(12340..14310)
     gene            complement(14446..14988)
     gene            complement(15013..16395)
```

Şekil 4.7 Süperkontig01'de bulunmuş bir gen için örnek GMSC+ GMHMM çıktısı

4.2.1.4. GLIMMER

GLIMMER algoritması ile öncelikle en uzun süperkontig olan SK12 üzerinde "öğrenme" işlemi gerçekleştirilmiştir. Dizi *de novo* bir dizi olduğundan, gen bulma işleminin daha etkin olabilmesi amacıyla öğrenme işleminin dizinin kendi üzerinde gerçekleştirilmesi gerekmektedir. Bu amaca yönelik olarak GLIMMER algoritmasının stratejisi izlenerek öncelikle dizi üzerinde uzun ORF'ler bulumuş (tüm ORF'ler içinden yeterli uzunlukta ve örtüşmeyen ORF'ler seçilmiş), ardından ICM (Interpolated Context Model) elde edilmiştir. Sonraki basamakta elde edilen model kullanılarak tüm süperkontigler üzerinde gen bulma işlemi gerçekleştirilmiştir.

```
LOCUS       scaffold00001              0 bp    DNA     linear
ACCESSION   UqazfaOA
VERSION     UqazfaOA.1
KEYWORDS    .
FEATURES             Location/Qualifiers
     gene            complement(47..169)
     gene            complement(459..1943)
     gene            complement(2153..3160)
     gene            complement(3197..3400)
     gene            complement(3352..3885)
     gene            complement(3882..4217)
     gene            complement(4234..5103)
     gene            complement(5064..5939)
     gene            complement(5939..6529)
     gene            complement(6526..7932)
     gene            complement(8258..9193)
     gene            complement(9442..10326)
     gene            complement(10348..11226)
     gene            complement(11280..12107)
     gene            complement(12340..14247)
     gene            complement(14446..14988)
```

Şekil 4.8 Süperkontig01'de bulunmuş bir gen için örnek Glimmer çıktısı

4.2.2. Gen Koordinatlarının Karşılaştırılmasında Kullanılan Yöntemler

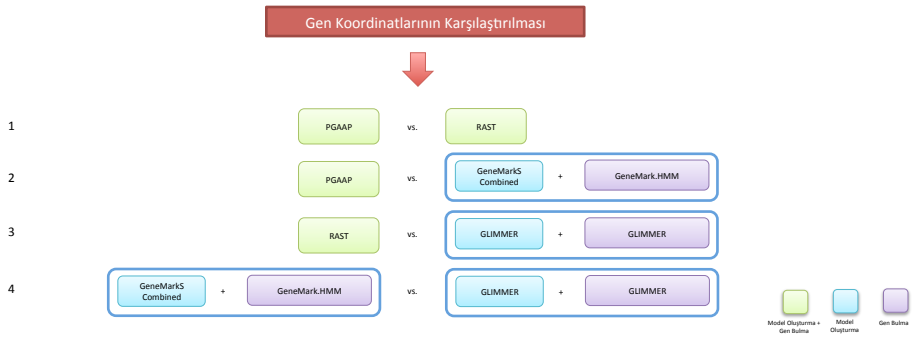
Dizi üzerinde gen bulma algoritmaları çalıştırıldığında çıktı olarak bulunan genlerin konum verisi elde edilmektedir. Konum verisi her bir süperkontig üzerinde bulunan genlerin başlangıç ve bitiş koordinatlarıdır. Konumsal karşılaştırma yöntemi iki yöntemin aynı bölge üzerinde bulduğu genlerin konumlarını karşılaştırmayı hedeflemektedir. Bunun için aynı genom dizisi üzerinde iki farklı yöntem kullanılarak elde edilmiş gen koordinatlarını kullanmaktadır.

Yöntemde karşılaştırma sonuçlarının eldesi için iki diziden birinin sonuçları referans alınarak diğeri yöntem tarafından aynı bölgede bulunan sonuçların başlangıç ve bitiş koordinatlarına ait pozisyonel değerlendirme yapılmaktadır. Bu yöntem kullanularak her bir karşılaştırma üç büyük süperkontig üzerinde ayrı ayrı gerçekleştirilmiştir.

4.2.3. Gerçekleştirilen Karşılaştırmalar

Çalışmada kullanılan PGAAP analiz akış hattı GeneMarkS combined ile birlikte GeneMark.hmm algoritmasını, RAST ise GLIMMER algoritmasını kullanmaktadır. GeneMark ve GLIMMER sistemleri, Markov modeli tabanlı gen bulma algoritmalarıdır. Yapılan karşılaştırmalar listelenen amaçlara yönelik olarak gerçekleştirilmiştir:

1- PGAAP ve RAST akış hatlarından elde edilen gen koordinatları arasındaki farkların incelenmesi,
2- PGAAP'ın çekirdek algoritması olan GeneMark ile elde ettiği gen koordinatlarının farklılıklarının incelenmesi,
3- RAST'ın da çekirdek algoritması olan GLIMMER ile elde ettiği gen koordinatlarının farklılıklarının incelenmesi,
4- GeneMark ve GLIMMER algoritmalarının gen bulma sonuçları arasındaki farklılıkların incelenmesi.



Şekil 4.9 Bulunan gen koordinatlarının karşılaştırılmasında kullanılan yöntemler

Algoritmalar teorik olarak benzer bir temele dayanıyor olmasına rağmen algoritmaların nitelikleri gen bulma işleminin sonuçlarını etkilemektedir. Bunun yanısıra, anotasyon akış hatları, çekirdek algoritma tarafından elde edilen gen koordinatları üzerinde, özellikle homoloji verisine dayanarak, çok daha kapsamlı düzeltmeler yapmaktadır. Bu düzeltmeler, gen bulucu tarafından "kodlayan bölge" olarak işaretlenmiş bir bölgeyi kodlayıcı olmayan olarak işaretlemek, ya da başlangıç - bitiş kodonunun koordinatını değiştirmek, peşpeşe bulunan iki ayrı genin aslında tek kodlayıcı bölge olduğunu bulmak gibi düzeltmelerdir. Bu değişimin gözlemlenmesi amacıyla algoritmalardan elde edilen sonuçlar karşılaştırılmıştır.

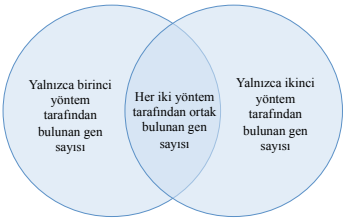
4.2.4. Gen Bulma Sonuçlarının ve Gerçekleştirilen Karşılaştırmaların Değerlendirilmesinde Kullanılan Yöntemler

Onüç süperkontig üzerinde Bölüm 4.2.1'de bahsedilen yedi yöntem ile gerçekleştirilen gen bulma işlemi sonucu elde edilen gen koordinatları öncelikle sayı ve uzunluk yönünden değerlendirilmiştir. Bu değerlendirmelerin ardından Bölüm 4.2.2'de bahsedilen ve Şekil 4.9'da gösterilen yöntemlerle gerçekleştirilen karşılaştırma sonuçları incelenmiştir. Sayı ve uzunluk yönünden gerçekleştirilen değerlendirmeler betimsel istatistikler olarak verilmiştir. Yapılan değerlendirmeler şunlardır:

1. Üzerinde gen bulma işlemi gerçekleştirilen süperkontiğin uzunluğu
2. Süperkontiğin %GC içeriği
3. Süperkontiğin üzerinde yer alan ORF sayısı
4. Bulunan gen sayısı
5. Maksimum gen uzunluğu
6. Minimum gen uzunluğu
7. Ortalama gen uzunluğu
8. Medyan gen uzunluğu

Gen koordinatlarının ikili karşılaştırılmasıyla aşağıdaki değerlendirmeler gerçekleştirilmiştir:

1. Birinci – ikinci yöntem tarafından bulunan genlerin sayısı
2. Yalnızca birinci – yalnızca ikinci yöntem tarafından bulunan ve ortak bulunan genlerin sayısı

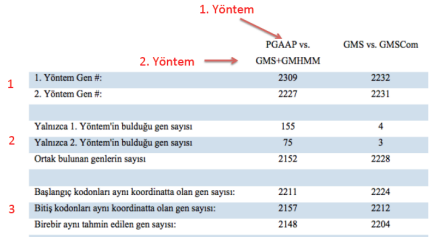


Şekil 4.10 Ortak ve özgün bulunan genler

3. Başlangıç kodonları / bitiş kodonları aynı koordinatta bulunan ve birebir aynı tahmin edilen genlerin sayısı

Şekil 4.11 Başlangıç/ bitiş koordinatları aynı olan genlerin olası pozisyonları

Bölüm 4.2.2'de bahsedilen karşılaştırma yöntemi ile Bölüm 4.2.3'te bahsedilen karşılaştırmalar gerçekleştirilmiş ve sonuçlar Şekil 4.12'deki formatta rapor edilmiştir. Bu tabloda sütunların başında yazan yöntem isimlerinden ilki "birinci yöntem", ikincisi "ikinci yöntem" olarak değerlendirilmiştir. 1 no'lu bölümde yöntemlere göre gen sayıları, 2 no'lu bölümde yöntemlerin özgün ve ortak buldukları genlerin sayıları, 3 no'lu bölümde ise aynı başlangıç kodonunu, aynı bitiş kodonunu paylaşan ya da birebir aynı bulunan genlerin sayıları verilmiştir.



Şekil 4.12 İkili karşılaştırma sonuçları tablosu

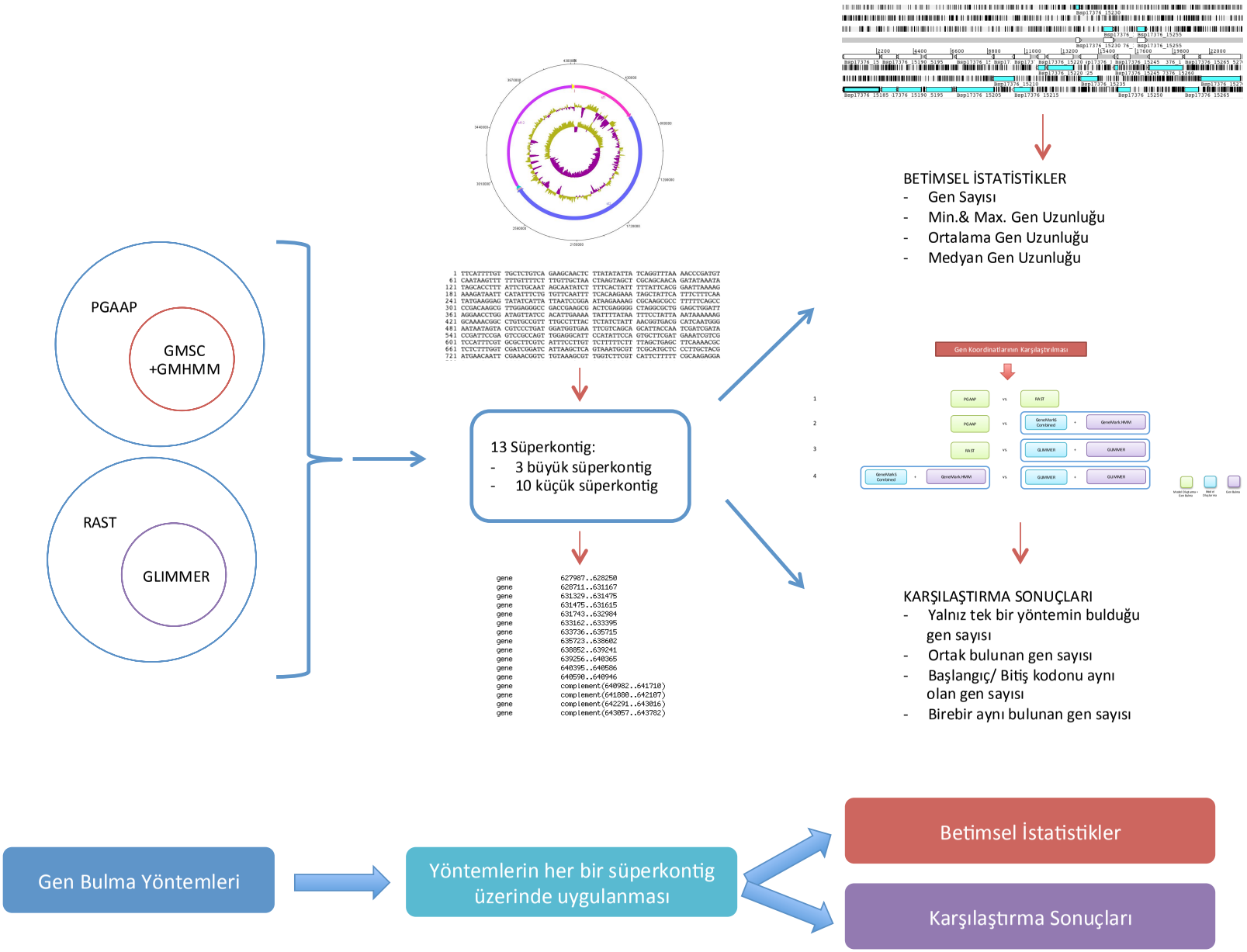
5. ARAŞTIRMA BULGULARI

Bir nükleotid dizisi verildiğinde, dizi üzerinde yer alan kodlayıcı bölgelerin koordinatlarının belirlenmesi problemi, nükleotid dizileme işlem hacminin hızla artmaya başladığı yıllardan beri önemini korumuştur. Bu problemin çözümüne yönelik olarak çeşitli gen bulma algoritmaları geliştirilmiştir. Bu algoritmalardan en etkin olanları Markov modeli tabanlı gen buluculardır (13). Bu tez kapsamında otomatize gen bulma sistemleri arasında öne çıkan iki sistem olan PGAAP ve RAST genom anotasyon akış hatları ve bu akış hatlarının çekirdek algoritmaları kullanılarak elde edilen gen koordinatları verisi değerlendirilmiştir.

Halihazırda NCBI tarafından standart otomatize prokaryotik genom anotasyon aracı olarak kullanılan PGAAP, çekirdek algoritma olarak GeneMark'tan türetilen GeneMarkS+ algoritmasını kullanırken, RAST yine Markov modeli tabanlı bir algoritma olan GLIMMER ile gen bulma işlemini gerçekleştirmektedir (2, 3). PGAAP ve RAST'ın ayrıştığı başka bir nokta ise anotasyon aşamasında kullandıkları ontoloji veri tabanlarıdır. Ancak tez kapsamında anotasyon, yani koordinatları belirlenen genlerin fonksiyonlarının saptanması işlemi ile ilgili bir karşılaştırma gerçekleştirilmemiştir. Araştırma kapsamında kullanılan tüm yöntemlerin özelliklerine ilişkin detaylara Bölüm 2.5'te değinilmiştir.

Yapılan karşılaştırmalar ile yöntemler tarafından bulunan genlerin uzunluk, başlangıç – bitiş koordinatları gibi özellikler yönünden farklılık gösterip göstermediği araştırılmıştır. Ayrıca akış hatları tarafından işlemin bir basamağında elenen, ya da başka bir nedenden listelenmeyen ancak ontolojik olarak gen olabileceğine dair kanıt bulunan bölgelerin varlığı araştırılmıştır.

Bu araştırma için ilk olarak gen bulma işlemi sonucu elde edilen gen koordinatları ile ilgili betimsel istatistikler, gen sayıları ve gen uzunluklarıyla ilgili detaylara odaklanılarak değerlendirilmiştir. Doğrudan gen koordinatlarının karşılaştırılması işlemi öncelikle otomatize akış hatları arasında gerçekleştirilmiştir. Ardından her bir akış hattından elde edilen sonuçlar, kendi çekirdek algoritmaları tarafından elde edilen sonuçlarla karşılaştırılmıştır. Böylece otomatize sistemlerce anotasyonu gerçekleştirilmemiş genlerin varlığı araştırılmıştır. Çalışma Şekil 5.1'de gösterilen işlem basamakları izlenerek gerçekleştirilmiştir.



Şekil 5.1 Tez çalışmasında izlenilen yöntem

5.1. Gen Bulma İşlemi ile Elde Edilen Sonuçlar

Gen bulma işlemi sonucunda ilk karşılaştırılan sonuç her bir yöntem tarafından bulunan gen sayıları olmuştur. Süperkontiglerin uzunlukları, %GC içerikleri, üzerlerinde bulunan ORF'ler (Açık okuma çerçeveleri) ve her bir yöntem tarafından her bir dizi üzerinde bulunan genlerin sayıları Çizelge 5.1'de verilmiştir.

Çizelge 5.1 Her bir süperkontig için dizi özellikleri ve bulunan gen sayıları

				Gen Sayıları			
	Uzunluk	%GC	ORF (min100bp)	PGAAP	RAST	GMSC+GMHMM	GLIMMER
SK01	772038	41,61	1170	678	661	673	666
SK02	2300059	41,97	3715	2309	2290	2330	2326
SK03	2910	39,73	5	4	4	4	5
SK04	2916	41,36	2	2	3	2	3
SK05	2274	48,07	4	2	1	2	2
SK06	3056	52,42	6	1	1	1	0
SK07	2360	41,4	3	2	2	2	2
SK08	2768	40,53	6	3	3	3	3
SK09	2483	40,03	4	2	2	2	2
SK10	2641	45,36	5	6	5	6	5
SK11	2058	37,03	1	1	1	2	1
SK12	1552928	42,4	2540	1589	1506	1567	1533
SK13	2425	41,32	7	4	4	5	4
Toplam	4650916		7468	4603	4482	4599	4552

Çizelge 5.1'de verilen özellikler ve gen sayıları incelenerek aşağıdaki bulgular elde edimiştir.

1. Tüm süperkontigler üzerinde toplam 7468 tane ORF, ancak en fazla 4603 gen bulunmuştur. Bu sonuçlar arasındaki fark 2865'dir.
2. Süperkontigler üzerinde bulunan gen sayıları birkaç durum haricinde her zaman bulunan ORF sayısından küçük olmuştur. Bu durum yalnızca üç küçük süperkontigde (4, 10 ve 11) değişmiştir ve diziler üzerinde yer alan ORF sayısından daha fazla gen bulunmuştur. Dördüncü süperkontigde iki adet ORF bulunmasına rağmen bazı yöntemler diziler üzerinde üç adet gen bulmuştur. Onuncu süperkontigde beş adet ORF bulunmasına rağmen PGAAP ve GMSC+GMHMM dizi üzerinde altı adet gen bulmuştur. Onbirinci süperkontigde bir adet ORF bulunmasına rağmen GeneMarkS Combined+GeneMark.hmm iki adet gen bulmuştur.

3. 6 tane ORF bulunmuş olan altıncı süperkontigde yöntemler sıfır, bir ve iki arasında değişen sayıda gen bulmuştur. Dizi üzerinde GLIMMER hiç gen bulmamıştır.
4. Tüm süperkontigler üzerinde bulunan gen sayıları değerlendirildiğinde toplamda en fazla gen PGAAP tarafından (4603) en az gen ise RAST tarafından bulunmutşur (4482). Bu iki sonuç arasındaki fark 121 gendir.
5. Büyük süperkontigler üzerinde her bir yöntem tarafından bulunan gen sayıları değerlendirildiğinde, bulunan gen sayılarının çizelgedeki gibi sıralandığı görülmüştür. Buna göre en az geni her bir büyük süperkontigde RAST, en fazla geni de PGAAP ya da çekirdek algoritması olan GMSC+GMHMM bulmuştur.

Çizelge 5.2 Süperkontiglere göre yöntemlerin buldukları gen sayılarına göre küçükten büyüğe sıralanması

	SK01	SK02	SK12
1	RAST	RAST	RAST
2	GLIMMER	PGAAP	GLIMMER
3	GMSC+GMHMM	GLIMMER	GMSC+GMHMM
4	PGAAP	GMSC+GMHMM	PGAAP

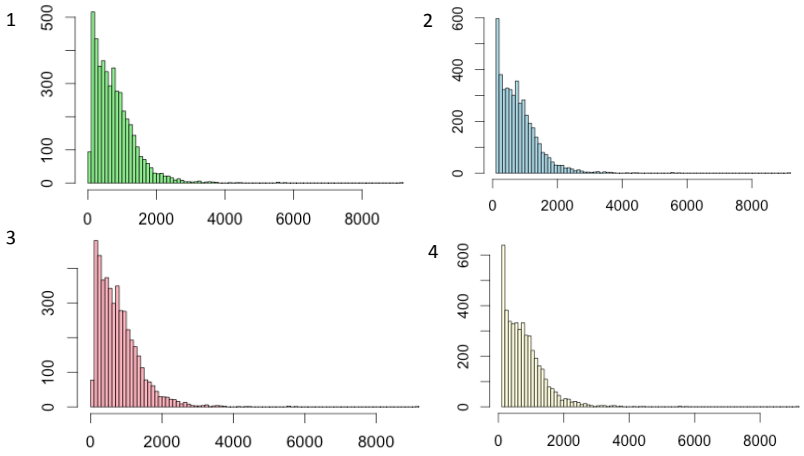
6. Küçük süperkontiglerde genellikle tüm yöntemler benzer sayıda gen bulmuştur.
7. PGAAP büyük süperkontiglerin tümünde RAST'tan daha fazla sayıda gen bulmuştur.

Yöntemler tarafından elde edilen genlerin genel olarak karşılaştırılması amacıyla gen uzunlukları elde edilmiş ve gen uzunluk dağılımları histogramları elde edilerek karşılaştırılmıştır. PGAAP, RAST, GMSC+GMHMM ve GLIMMER yöntemleri tarafından bulunan genlerin uzunluk dağılımları elde edilmiştir.

Çizelge 5.3 Yöntemlere göre bulunan genlere ilişkin betimsel istatistikler

	En Küçük	En Büyük	Ortalama	Medyan
PGAAP	68	9107	784.2	665
RAST	113	9107	801.1	695
GMSC+GMHMM	47	9107	790.6	671
GLIMMER	113	9107	790.5	677

Şekil'deki histogramlar incelendiğinde PGAAP ve GMSC+GMHMM'nin RAST ve GLIMMER'ın bulamadığı yüze yakın, boyu 100 bç'den kısa gen bulduğu görülmektedir. Anotasyon akış hatlarının kendi çekirdek algoritmaları ile gen uzunluk dağılımları benzerlik göstermektedir. Genel olarak gen uzunluk dağılımlarında çok büyük bir farklılık gözlenmemektedir.



Şekil 5.2 Yöntemlere göre gen uzunluğu dağılımlarını gösteren histogramlar 1. PGAAP, 2. RAST, 3. GMSC+GMHMM, 4. Glimmer

5.2. Yöntemler Arası Karşılaştırma Bulguları

Bölüm 4.1'de ve Şekil 4.2'de gösterildiği gibi genom dizisinin büyük bölümü elde edilen üç büyük parçada (süperkontigde) bulunmaktadır. Bunlar Süperkontig01 (SK01), Süperkontig02 (SK02) ve Süperkontig12 (SK12)'dir. Bu süperkontigler boyutları yönünden geri kalan süperkontiglerden çok farklı ve kendi aralarında benzer olan büyük süperkontigler olduğundan gen bulma sonuçlarının yöntemlere göre karşılaştırılması yalnızca bu süperkontigler için verilmiştir.

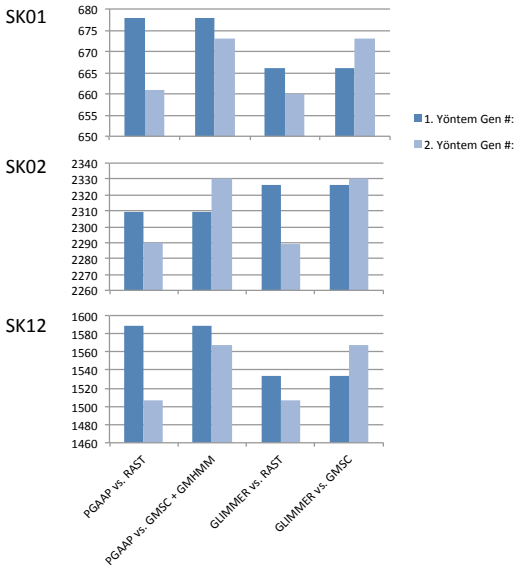
Büyük süperkontigler üzerinde gerçekleştirilen gen bulma işlemlerinin yöntemlere göre ikili karşılaştırılması sonucu, elde edilen gen koordinatlarının büyük ölçüde paralellik gösterdiği görülmüştür. Bunun yanısıra her bir yöntemin bulduğu özgün genler de bulunmaktadır. Ancak bu farklılıklar büyük bölgeleri değil görece kısa nükleotid dizilerini içerdiğinden çok önemli farklılıklara neden olmamaktadır. Aynı genin başlangıç kodonunun farklı bulunmasından kaynaklanabilecek çerçeve farklılığı problemi, tüm yöntemler boyunca birkaç kereden fazla gözlenmemiştir. Bu bölümde yapılan karşılaştırmalar ile elde edilen sonuçlardan öne çıkanlar verilmiştir.

5.2.1. İkili Karşılaştırmalar

Çalışma kapsamında listedeki ikili karşılaştırmalar yapılmıştır:

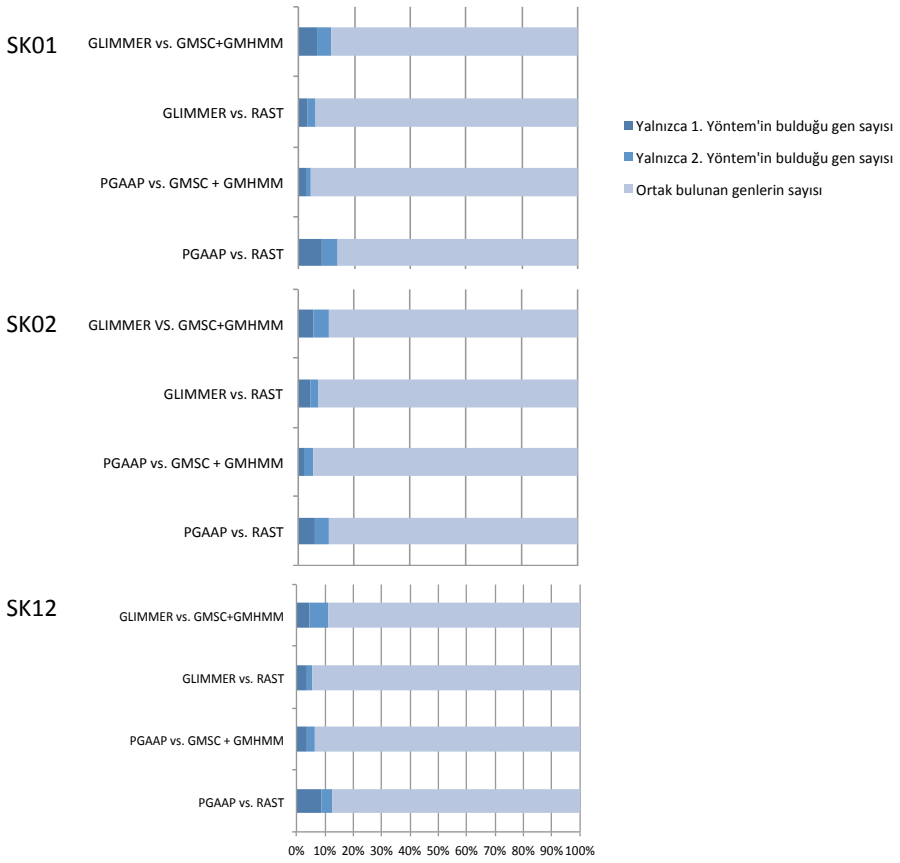
1- **PGAAP – RAST:** Anotasyon akış hatlarının bulduğu gen koordinatları arasındaki benzerlik ve farkların gözlemlenmesi
2- **PGAAP – GMSC+GMHMM:** Anotasyon akış hattı olan PGAAP ile bulunan gen koordinatlarının çekirdek algoritması olan GMSC+GMHMM tarafından bulunan gen koordinatlarıyla karşılaştırılması
3- **RAST – GLIMMER:** Anotasyon akış hattı olan RAST ile bulunan gen koordinatlarının çekirdek algoritması olan GLIMMER tarafından bul unan gen koordinatlarıyla karşılaştırılması
4- **GMSC+GMHMM – GLIMMER:** Önde gelen gen bulma algoritmalarının bulduğu gen koordinatları arasındaki benzerlik ve farkların gözlemlenmesi

İkili karşılaştırmalarda ilk olarak yöntemlerin bulduğu gen sayıları arasındaki farklılıklar incelenmiştir. Şekil ..de bu farklılıklar görülmektedir. Buna göre PGAAP SK02 haricinde hem diğer akış hattı olan RAST'tan, hem de kendi çekirdek algoritması olan GMSC+GMHMM'den daha fazla gen bulmuştur. RAST ise kendi çekirdek algoritmasından daha az sayıda gen bulmuştur. PGAAP ve RAST karşılaştırmasında tüm süperkontiglerde PGAAP'ın daha fazla gen bulduğu gibi, GMSC+GMHMM de tüm süperkontiglerde GLIMMER'dan fazla gen bulmuştur.



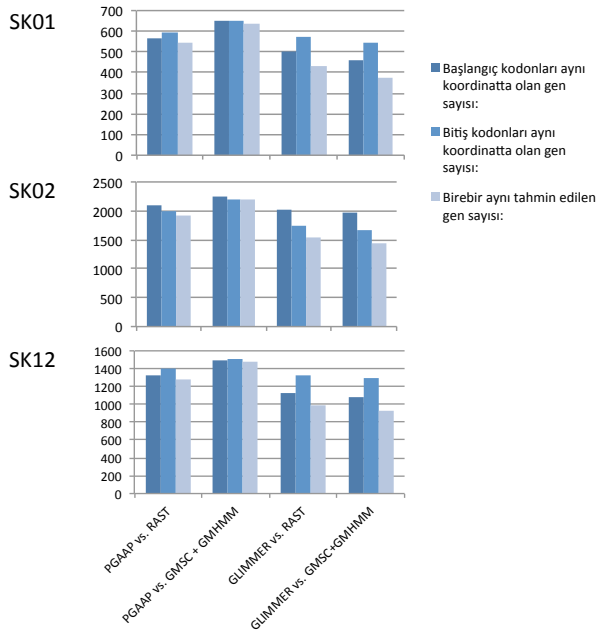
Şekil 5.3 Yöntemlerin bulduğu gen sayısına göre ikili karşılaştırma sonuçları

İkinci olarak, ikili karşılaştırması yapılan yöntemlerin özgün buldukları ve ortak buldukları genlerin sayıları incelenmiştir. Buna göre, RAST ve PGAAP'ın çekirdek algoritmalarını oluşturan GMSC+GMHMM'nin Glimmer ile yakın sonuç verdiğini, her bir algoritmanın tek başına bulduğu gen sayısının diğeri ile neredeyse denk olduğunu göstermiştir. Diğer yandan anotasyon akış hatları kendi çekirdek algoritmaları ile karşılaştırıldığında ise algoritmaların özgün buldukları gen sayısı en aza inmektedir. Bulunan özgün gen sayısında en büyük farklılık Şekil 5.2'de görüldüğü gibi iki anotasyon akış hattı birbirleriyle karşılaştırıldığında gözlemlenmektedir.



Şekil 5.4 SK01, SK02 ve SK12 üzerinde gerçekleştirilen tüm karşılaştırmalarda ortak ve farklı bulunan genler

Son olarak ikili karşılaştırmalarda yöntemler arasında aynı başlangıç kodonunu, aynı bitiş kodonunu paylaşan ve birebir aynı olan gen sayıları elde edilmiştir (Şekil 5.4). Başlangıç – bitiş ve birebir örtüşme sayısı en büyük olan yöntemler PGAAP ve çekirdek algoritması olan GMSC+GMHMM'dir. RAST ve GLIMMER arasındaki bu yönden benzerlik, GMSC +GMHMM ve GLIMMER arasındaki benzerlikten fazla değildir.

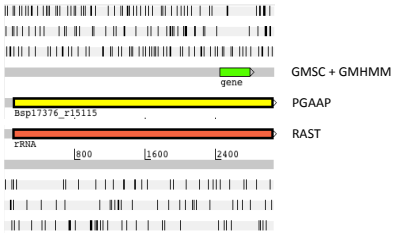


Şekil 5.5 Başlangıç kodonu aynı olan, bitiş kodonu aynı olan ya da birebir aynı olan gen sayısına göre ikili karşılaştırma sonuçlar

5.2.2. Genom Üzerindeki Kodlamayan RNA Genleri

Markov Modeli tabanlı gen bulma algoritmaları gen bulmada GC içeriği, kodlayan bölgelere ilişkin kodon kullanım istatistikleri, Ribozomal Bağlanma Bölgesi'nin varlığı, Ribozom Bağlanma Bölgesi'ne olan uzaklığın dağılım özellikleri gibi üzerinde çalışılan türe özel genomik niteliklerden yararlanmaktadır. Algoritmalar ancak türe özel "eğitildiğinde" doğru bir gen bulma işlemi gerçekleştirebileceğinden, genel olarak, öncelikle verilen dizi üzerindeki kodlayıcı bölge olma olasılığı yüksek bölgeleri belirlemekedir. Ardından bu bölgeleri geçici olarak doğru kabul ederek, kodlayıcı bölgeyi anlatan "yaklaşık" bir model oluşturmaktadır. Bu işlemi takip eden aşamalarda adım adım model iyileştirilmekte, her yeni model ile daha iyi gen koordinatları öğrenilmekte ve sonuç olarak da birbirini besleyen öğrenme ve gen bulma aşamaları sonucunda en "doğru" gen koordinatları elde edilmektedir.

Genom üzerinde protein kodlamakta görevli olan genlerin haricinde, başka fonksiyonel moleküllerin üretilmesini sağlayan genler de bulunmaktadır. Bu genleri içeren genomik bölgeler protein kodlamadıkları için "kodlamayan" RNA'lar olarak anılmaktadır. Bu bölgeler, örneğin, tRNA ve rRNA gibi proteine dönüşmeyen, ancak protein kodlanmasında aminoasitlerin taşınması ve birleştirilmesi gibi başka görevlerde yer alan RNA moleküllerini kodlamaktadır. Bu bölgeler nükleotid içeriği yönünden diğer genlerden farklıdır. Bu nedenle protein kodlayan genleri bulmak amacıyla geliştirilmiş modeller kullanılarak bu bölgelerin etkin olarak belirlenmesi mümkün olmamaktadır. Dolayısıyla yalnızca çekirdek algoritmalar ile gerçekleştirilen gen bulma işlemlerinde rRNA ve tRNA genleri yüksek olasılıkla listelenemeyecektir. Ancak anotasyon akış hatları, çekirdek gen bulma algoritmalarının dışında diğer genomik bileşenleri de isimlendirebilecek algoritmaları da kullandığından, bu genleri tespit etmiş olacaktır. tRNA'ları bulmak için tRNA yapısına özel geliştirilmiş tRNAScan-SE (21) , rRNA'ları bulmak için rRNA yapısı ile eğitilmiş, yine gizli Markov modeli tabanlı RNAmmer (19) kullanılabilir. Şekil 5.6 'da bir rRNA geninin olduğu bölgede PGAAP, RAST ve GMSC+GMHMM yöntemlerinin gerçekleştirdiği gen bulma işleminin sonuçları görülmekedir. GLIMMER bu bölgede hiç gen bulamadığından şekilde yer almamaktadır.



Şekil 5.6 Altıncı süperkontig üzerinde farklı gen bulucular tarafından bulunan sonuçlar

Çizelge 5.4'te PGAAP ve RAST tarafından bulunan rRNA genleri ve bu genlerin konumları verilmiştir. *B. Boroniphilus*'a ait 16S rRNA dizisi tez kapsamında değerlendirilmeyen küçük bir kontig üzerinde bulunmuştur. Bu nedenle Çizelge 5.4'te görülmemektedir.

Çizelge 5.4 PGAAP ve RAST ile bulunan rRNA'lar

	rRNA	Süperkontig	Başlangıç	Bitiş
PGAAP	5S	2	1468065	1467950
	5S	12	41652	41537
	5S	12	561202	561317
	23S	6	103	3041
RAST	5S	2	1468065	1467950
	5S	12	41652	41537
	5S	12	561202	561317
	23S	6	103	3041

rRNA'lar gibi, tRNA'lar da ancak kendilerine uygun algoritmalar tarafından arandığında bulunabilmektedir. tRNA'ların rRNA'lardan farkı, üç boyutlu yapılarının teorik olarak, diğer genomik bileşenlere göre, çok daha kesin hatlarla belirlenebilmiş olmasıdır. Dolayısıyla doğrudan tRNA'ya benzer diziyi bulmak için diğer genomik bileşenlerin bulunmasında kullanılan algoritmalardan daha az karmaşık bir algoritma kullanılabilmektedir. PGAAP, tRNA'ların belirlenmesinde en iyi tRNA bulma algoritması olduğu gösterilmiş tRNA Scan-SE algoritmasını kullanmaktadır. Bu bulguların doğrulanması için algoritma dizi üzerinde tekrar çalıştırılmış ve sonuçları Çizelge 5.5'te verilmiştir. PGAAP, tRNAScan-Se ve RAST tarafından, *Serin (Ser)* hariç tüm antikodonlara ait tRNA'ların bulunduğu görülmüştür.

Çizelge 5.5 PGAAP, RAST ve tRNAScan-SE ile bulunan tRNA'lar

Kodon	PGAAP	tRNAScan-SE	Antikodonlar	RAST	Antikodonlar
Ala	2	2	GGC, TGC	2	GGC, TGC
Arg	5	5	CCG, CCT, ACG, ACG, TCT	5	CCG, CCT, ACG, ACG, TCT
Asn	3	3	GTT, GTT, GTT	3	GTT, GTT, GTT
Asp	4	4	GTC, GTC, GTC, GTC	4	GTC, GTC, GTC, GTC
Cys	2	2	GCA, GCA	2	GCA, GCA
Gln	3	3	TTG, TTG, TTG	3	TTG, TTG, TTG
Glu	4	4	TTC, TTC, TTC, TTC	4	TTC, TTC, TTC, TTC
Gly	5	5	TCC, GCC, TCC, GCC, TCC	5	TCC, GCC, TCC, GCC, TCC
His	2	2	GTG, GTG	2	GTG, GTG
Ile	1	1	GAT	1	GAT
Leu	6	6	CAA, GAG, TAG, TAG, TAA, CAG	6	CAA, GAG, TAG, TAG, TAA, CAG
Lys	2	2	TTT, TTT	2	TTT, TTT
Met	6	6	CAT, CAT, CAT, CAT, CAT, CAT	6	CAT, CAT, CAT, CAT, CAT, CAT
Phe	3	2	GAA, GAA	2	GAA, GAA
Pro	2	2	TGG, TGG	2	TGG, TGG
Ser	5	5	TGA, GGA, GCT, TGA, GCT	1	GCT
Thr	1	1	TGT	1	TGT
Trp	1	1	CCA	1	CCA
Tyr	1	1	GTA	1	GTA
Val	3	3	GAC, TAC, TAC	3	GAC, TAC, TAC

6. TARTIŞMA ve SONUÇ

Genom projeleri çağında gen bulma işlemi ıslak laboratuvar tabanlı deneysel yöntemlerle gerçekleştirilemeyecek kadar büyük işlem hacmi gerektirmektedir. Bu işlemin araştırmacıların genom üzerinde gerçekleştirecekleri çalışmaların devamını sağlayacak en önemli ara basamaklardan biri olduğu açıktır. Genom üzerinde gen bulma işlemi ve gen olma olasılığı yüksek olan bu bölgelerin fonksiyonel özelliklerle isimlendirilmesi genel olarak "genom anotasyonu" olarak anılmaktadır. Genom anotasyonunun kalitesini arttıran öncül basamak ise gen bulma işlemidir.

Günümüzde henüz genom anotasyonunu tek başına hesaplamsal yöntemlerle gerçekleştirilebilir hale getiren bir yöntem geliştirilememiştir. Ancak gen kestirim işleminin "yeterli" kalitede gerçekleştirilmesini sağlayan matematiksel ve istatistiksel teknikler mevcuttur. Bu tekniklerin gen bulma gücünün daha önce anotasyonu gerçekleştirilmiş genom dizileri üzerinden elde edilen bilgilerden yararlanılarak arttırılması ile GenBank üzerinde rutin çalıştırılan otomatize gen bulma ve anotasyon akış hatlarının geliştirilmesi mümkün hale gelmiştir. Bölüm 2.3'te bahsedildiği gibi, geliştirilen pek çok hesaplamsal gen bulma yöntemi arasından gizli Markov modeli tabanlı gen bulma algoritmaları öne çıkmıştır. Sonuç olarak bu algoritma NCBI'nın kullandığı otomatize genom anotasyonu akış hattının (PGAAP) çekirdeği olarak kullanılmaya başlanmıştır.

Hesaplamsal gen bulma yöntemleri *de novo* genom projelerinde daha büyük önem kazanmaktadır. Zira anotasyonda yararlanılabilecek homoloji verisi aynı tür/cinsten değil benzer klas/ordo/familyadan geldiği için gücü azalmaktadır. Oysaki gen bulma basamağının incelikle gerçekleştirilmesi genom verisinin daha etkin bir biçimde kullanılması için şarttır. Ancak iyi bir gen bulma işlemi sayesinde diğer canlılarda benzeri olmayan genomik bölgelerin tespiti mümkün olabilir. Benzeri olan genler içinse kesin başlangıç ve bitiş koordinatlarının belirlenmesi önemlidir.

Hesaplamsal gen bulma işlemi sayesinde araştırıcının elinde kodlayıcı olma olasılığı yüksek ancak işlevi belli olmayan (ya da homoloji verisi sayesinde işlevi yaklaşık olarak tahmin edilmiş) genomik bölgelerin başlangıç ve bitiş koordinatları olabilecek, böylece gerektiğinde araştırmacı bu veriyi kullanarak fonksiyonel deneyler tasarlayabilecektir. Yalnızca homoloji taramasıyla bu bilginin bu hassasiyette eldesi mümkün değildir.

Diğer yandan bir genomun dizilenmesi, anotasyonunun otomatize olarak gerçekleştirilmesi, o genom üzerinde gerçekleştirilecek çalışmaların tamamlanması anlamına gelmemektedir. Yeni, daha güçlü algoritmalar geliştirildikçe, genomlar üzerinden elde edilecek bilgi ve bu bilginin güvenilirliği günden güne artacaktır. Bu nedenle gen bulma algoritmaları üzerinde bilgi sahibi olmak, genom araştırmaları yapmakta olan araştırma ekipleri için çok önemlidir.

Bu tez kapsamında anotasyonu ele alınan ve farklı algoritmalarla tekrar gerçekleştirilen Bacillus boroniphilus genomunun orijinal anotasyonu NCBI'nın PGAAP'ın güncellenmesiyle geliştirdiği yeni anotasyon akış hattı PGAP ile gerçekleştirilmiş ve NCBI veri tabanında PRJNA62235 erişim numarasıyla araştırmacıların kullanımına sunulmuştur. PGAP NCBI ve GeneMark ekibinin ortaklaşa geliştirdiği ve GeneMarkS+ algoritmasına dayalı olarak çalışan yeni bir sistem kullanmaktadır.

PGAAP – RAST, GMSC – Glimmer Karşılaştırmaları

Gen bulma sonuçlarındaki benzerlik ve farklılıkların incelenmesi amacı ile yöntemler ile elde edilen sonuçlar Bölüm 5.2'de karşılaştırılmıştır. İkili karşılaştırma sonuçları incelendiğinde, GeneMarkS Combined algoritmasını kullanan PGAAP anotasyon akış hattının Glimmer algoritmasını kullananan RAST anotasyon akış hattından daha fazla sayıda gen (Şekil 5.3) ve daha fazla sayıda özgün gen bulduğu görülmüştür (Şekil 5.4). Benzer şekilde, anotasyon akış hatlarının çekirdek algoritmalarının sonuçlarının karşılaştırılmasından elde edilen bulgular da bu bulgularla paralellik göstermektedir (Şekil 5.3 ve Şekil 5.4).

Ab initio, yani nükleotid dizisi haricinde bilgi verilmeksizin aynı dizi üzerinde gen bulunmasını sağlayan algoritmaların etkinliğinde modelin öğrenme yöntemi ve modelin özellikleri ön plana çıkmaktadır. Glimmer algoritması öğrenme yöntemi olarak nükleotid dizisi üzerindeki ORF'leri kullanmaktadır. Bunu ORF'lerin kodlayan bölge olma olasılığı daha yüksek nükleotid bölgeleri olacağı varsayımıyla yapmaktadır (29). GeneMarkS ise, doğrudan bu probleme, yani üzerinde daha önce gen bulma işlemi gerçekleştirilmemiş bir nükleotid dizisi üzerinde gen bulmada kullanılacak modelin başlangıç parametrelerinin iyilştirilmesine yönelik geliştirilmiş "Heuristic Modeller" ile başlangıçta kullanılacak gen koordinatlarını bulmakta, ardından bu öncül genler üzerinde kullanılan gizli Markov modeli tabanlı algoritmayı eğitecek gen içerik istatistiklerini elde etmektedir (6). Algoritmalar, dolayısıyla da bunları kullanan anotasyon akış hatları arasındaki temel yöntem farklılıklarından biri budur. Diğer farklılık ise kullanılan Markov modelinin türüdür. GeneMarkS gen koordinatlarının belirlenmesinde beşinci mertebeden Markov modeli kullanmakta, Glimmer ise interpole Markov modelinden yararlanmaktadır.

Gen bulmanın yanısıra, kodlayan bölgelere ait başlangıç ve bitiş kodonlarının hassas olarak belirlenmesi de önem taşımaktadır. Bu amaçla yöntemler tarafından bulunan genlerin başlangıç ve bitiş kodonlarının bulunan kaç gende ortak olduğu da araştırılmıştır (Şekil 5.5). Buna göre PGAAP ve çekirdek algoritması olan GeneMarkS arasında başlangıç ve bitiş koordinatlarının benzerlik oranı yüksektir. Ancak RAST ve çekirdek algoritması Glimmer arasındaki başlangıç ve bitiş koordinatlarının benzerlik oranı, Glimmer ve GeneMarkS algoritmalarının arasındaki benzerlik oranından farklı değildir. Her iki algoritmanın da başlangıç kodonunu belirlemede RBS'ye (Ribozomal Bağlanma Bölgesi'ne) olan uzaklıkları değerlendirerek iyileştirme yaptığı göz önünde bulundurulduğunda, RAST'ın bu iyileştirmeyi daha çok homoloji arama safhasında gerçekleştirdiği, GeneMarkS'in homoloji aramasına gerçeğe daha yakın bir gen seti ile başlanmasını sağladığı söylenebilir. Tartışılan bulgulara benzer bulgular GeneMarkS'in Glimmer'a karşı performans değerlendirilmesinde de görülmektedir (7).

Anotasyonların Güncelliği

Gen bulma algoritmaları, genel anlamda, bir organizmanın genlerinin neye benzediğini öğrenerek, üzerinde hangi genomik bileşenlerin olduğu bilinmeyen bir nükleotid dizisi üzerinde, öğrendiklerine benzeyen yapıları işaretlemektedir. Başlangıç ve bitiş koordinatları belirlenen bu dizinin gerçekten kodlayan bir bölge olup olmadığını anlamak için yapılacak bir diğer hesaplamsal işlem de, bu bölge tarafından kodlanan proteinin homoloğunun veri tabanlarında var olup olmadığının araştırılmasıdır. Bu aşama tamamen veri tabanındaki güncel kayıtlara, kullanılan veri tabanına ve BLAST algoritmasında kullanılan parametrelere bağlıdır. Gen bulma algoritmasının kodlayıcı bölge olarak işaretlediği bir nükleotid dizisinin o tarihte veri tabanındaki homolog bir proteinle eşleşmeme olasılığı vardır. Bu bölgelerin gerçekten kodlayan bölge olup olmadığını gösterebilecek başka bir hesaplamsal yöntem yoktur. Bu nedenle bu tip diziler "hipotetik protein" olarak işaretlenmekte ve o genin hangi proteini kodladığı bulunana kadar bu isimle kaydedilmektedir.

NCBI'nın prokaryotik genomların anotasyonunda standart olarak kullandığı PGAAP, GeneMarkS ile bulunan genleri, tüm prokaryotik RefSeq genomlarından elde edilen global protein kümelerinden elde edilen temsilciler üzerinde BLAST ile aramaktadır. RefSeq, NCBI'nın Referans Dizi Veritabanı'dır. Bu veritabanı kapsamlı, tekrarsız (non-redundant) ve iyi anote edilmiş genomik DNA, transkript ve protein dizilerini içermektedir. Genom anotasyonu, gen kimliklendirme ve karakterizasyonu, mutasyon ve polimorfizm analizleri, ifade çalışmaları ve karşılaştırmalı analizler için stabil bir referans oluşturmaktadır. RefSeq kayıtları, halka açık dizi verisinin çok aşamalı bir validasyon sürecinden geçirilmesi ve manüel kürasyon sonucunda elde edilmektedir.

Anotasyon sürecinde, anotasyonu gerçrekleştirilen dizi RefSeq dizi koleksiyonları üzerinde BLAST'landıktan sonra elde edilen protein hizalamaları benzerlik ve simetrik örtüşme özelliklerine göre kalite değerlendirmesinden geçirilmektedir. Eğer bir dizi tutarlı olarak yeterli sayıda aynı kümeye ait olan proteinle eşleşirse, o dizi o betimleyici kümeye atanır.

RAST da yine benzer yöntem ile dizileri FIGfam protein kümeleri ile eşlemektedir. FIGfam'ler Fellowship for Interpretation of Genomes (FIG) tarafından oluşturulan protein kümeleridir. Bu yöntem NCBI'nın anotasyon sürecine alternatif olarak geliştirilmiştir.

Anotasyon akış hatları, bu homoloji aramaları sonucunda, gen bulma algoritmalarının sonuçlarında düzeltmeler de yapmaktadır. Bu nedenle gerçekleştirilen anotasyonun sürekli olarak güncellenmesi, sonuçların güvenilirliği yönünden büyük önem taşımaktadır.

Tez çalışmasında, NCBI'nın otomatize prokaryotik anotasyon aracının ilk versiyonu olan PGAAP sonuçları değerlendirilmiştir. 2013 yılında PGAAP'ın yerini, otomatize prokaryotik anotasyon aracının ikinci versiyonu olan PGAP almıştır. PGAAP ile PGAP arasında büyük farklılıklar bulunmaktadır. Sonuçlardaki en büyük farklılıklar ise, sistemin anotasyon sonuçlarında gözlenmektedir.

PGAAP'ın "Gen bulma algoritması -> Homoloji araması" yönünde ilerleyen anotasyon stratejisinin yerini PGAP'ta dizinin önce homoloji aramasından geçirildiği, ardından anotasyonların GeneMarkS+ ile daha rafine hale getirildiği bir stratejiye bırakmıştır. Yeni versiyonun anotasyon konusunda çok daha "tutucu" olduğu söylenebilir. PGAAP'ın aynı lokasyonda birebir aynı bularak farklı isimlendirdiği pek çok genin, yeni versiyonda hipotetik protein olarak işaretlendiği görülmüştür.

PGAP'ın veri kayıt prosedürü de PGAAP'tan farklıdır. PGAP artık, kullanıcının isteği doğrultusunda, NCBI'ya veri yüklenme işlemini takiben otomatik olarak çalıştırılmaktadır. Eski versiyonda ise, veri anotasyon için NCBI'ya gönderilmekte, anotasyon sonuçları geri alınmakta, ardından veri tabanına yüklenmekteydi. NCBI otomatize anotasyon işlemini periyodik olarak tüm veritabanı üzerinde gerçekleştirmektedir.

Gen Bulucular ve Kodlamayan RNA Genleri

Belirli bir genomik yapıya özel tasarlanan bir "bileşen" bulucu ile yalnızca o tipteki bölgelerin keşfi mümkün olmaktadır. Çalışmada bahsedilen gen bulucuların bulduğu bölgeler olan protein kodlayan bölgelerin dışında genom üzerinde özellikleri kodlayan bölgelerden farklı olan çok sayıda genomik bileşen bulunmaktadır. Bunlardan bazıları tRNA ve rRNA genleri gibi kodlayıcı olmayan RNA genleridir. Bu genler nükleotid içerikleri yönünden kodlayan genlerden farklıdır. Bu nedenle bu genlerin bulunmasında gen bulma algoritmalarının kullanılması uygun değildir. tRNA'ların bulunması için tRNAScan-SE ve rRNA'lar için de RNAmmer algoritmaları geliştirilmiştir. Hem PGAAP hem RAST tRNA'ların bulunmasında tRNAScan-SE algoritmasını kullanırken, rRNA'ların bulunmasında PGAAP RNAmmer algoritmasını, RAST ise *search_for_rna* adında bir araç kullanmaktadır. Her iki yöntem de hemen hemen aynı rRNA genlerini bulmuş olsa da, iki yöntemin verdiği rRNA sonuçlarında küçük bir farklılık görülmektedir (Çizelge 5.5).

Sonuç

Tüm bu bulgular değerlendirildiğinde, genom projelerinde anotasyon işleminde en etkin sonuçların elde edilebilmesi için, PGAAP ve RAST gibi en iyi gen bulma algoritmalarını etkin homoloji aramalarıyla birleştiren yöntemlerin kullanılması gerektiği görülmektedir. Bu araçlardan PGAAP gen bulma işlemini performansı daha yüksek olduğu gösterilmiş GeneMarkS ile, homoloji taramasını da RefSeq dizileri üzerinde yaptığı için daha güvenilir sonuçlar vereceği sonucuna varılmıştır. RAST ise, görece yeni bir anotasyon girişimidir, ekip tarafından proteinler üzerinde yeni bir isimlendirme çalışması sürdürülmektedir. RAST sunucusu, anotasyon sonuçlarını çok daha kullanılabilir ve araştırmacı dostu şekilde vermektedir. Bunun yanısıra, RAST'ın surduğu analiz araçlarıyla verinin karşılaştırmalı analizleri web üzerinden gerçekleştirilebilmektedir. Her ne kadar PGAAP ve RAST sonuçları birebir aynı olmasa da, sonuçlar arasındaki büyük paralellik bu iki yöntem sonuçlarının bir arada değerlendirmesinin mümkün olduğunu ve bu tip bir kullanımın araştırmacıya esneklik ve kolaylık sağlayacağını göstermektedir.

Genom dizilerinin anotasyon akış hatlarında anotasyonunda, genomun merkeze teslim edilmeden bu sistemler tarafından anotasyonu mümkün olmamaktadır. Stratejik önemi olan genom projelerinin anotasyonunun bir merkez tarafından gerçekleştirildiği durumlarda alınabilecek tek önlem genom dizi ve anotasyonunun halka açık hale geleceği tarihin belirlenmesidir. Bu noktada anotasyonun, bu sistemler kadar "mükemmel" olmasa da, anotasyonunun gerçekleştirilmesi için GeneMarkS algoritmasının çalıştırılması, ardından PGAAP'ta izlenilen stratejiye benzer otomatize bir BLAST algoritmasıyla bölgelerin isimlendirilmesi önerilebilir.

NCBI'nın PGAP'ta kullandığı GeneMarkS+ versiyonu, tek başına kullanıma sunulmamıştır (Çünkü ilk aşamada veritabanındaki hizalama sonuçlarını kullanmaktadır.). Veritabanına kaydedilen veri üzerinde yazılım güncellemeleri, yöntem değişikliği ve standart anotasyon güncellemeleri ile değişiklikler gerçekleşebilmektedir. Bu nedenle, araştırmacıların anotasyon versiyonlarını takip etmesi, versiyonlar arasındaki değişimleri izlemesi, araştırmalarının sonraki basamaklarını tasarlarken bu değişimlere dikkat etmesi gerekmektedir. Bu değişimlerin gözlenmesi NCBI GenBank revizyon kontrolü ile gerçekleşirilebilmektedir. Ancak bu araçla başka yöntemlerle gerçekleştirilmiş anotasyonlardaki farklılıklar karşılaştırılamamaktadır. Bu amaca yönelik olarak, tez çalışması kapsamında geliştirilmiş olan Overlap yazılımı kullanılabilir.

Sonuç olarak, her ne kadar genom anotasyon hizmetleri NCBI, JCVI gibi büyük laboratuvarlar tarafından sağlansa da, bu işlem ve sonuçların değerlendirilmesi konusunda teknik uzmanlık geliştirmek stratejik önem taşımaktadır. Genom araştırması yapan merkezlerin, kendi gen bulma ve anotasyon alt yapılarını kurmaları veya mevcut algoritmaların akışına hakim olmaları, özellikle *de novo* genom projelerinde anotasyonların versiyon değişikliklerinde yaşanan değişimlere özen göstermeleri, araştırmalarının devamında büyük önem taşımaktadır.

KAYNAKLAR

1. Ahmed I, Yokota A, Fujiwara T. A novel highly boron tolerant bacterium, *Bacillus boroniphilus* sp. nov., isolated from soil, that requires boron for its growth. *Extremophiles*. 2007;11:217–224. doi:10.1007/s00792-006-0027-0.

2. Angiuoli SV, Gussman A, Klimke W, et al. Toward an Online Repository of Standard Operating Procedures (SOPs) for (Meta)genomic Annotation. *OMICS: A Journal of Integrative Biology*. 2008;12(2):137–141. doi:10.1089/omi.2008.0017.

3. Aziz RK, Bartels D, Best AA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics*. 2008;9(1):75. doi:10.1186/1471-2164-9-75.

4. Baker M. De novo genome assembly: what every biologist should know. *Nat Meth*. 2012;9(4):333–337. doi:10.1038/nmeth.1935.

5. Bertsch J, United States Department of Energy, Joint Genome Institute. Genomes Online Database. Reddy T, Mallajosyula J, Thomas A, Isbandi M, eds. :43089. Available at: http://www.genomesonline.org.

6. Besemer J, Borodovsky M. Heuristic approach to deriving models for gene finding. *Nucleic Acids Res*. 1999;27(19):3911–3920.

7. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*. 2001;29(12):2607–2618.

8. Borodovsky M, McIninch J. GENMARK: parallel gene recognition for both DNA strands. *Computers & chemistry*. 1993;17(2):123–133.

9. Borodovsky M, Rudd KE, Koonin EV. Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res*. 1994;22(22):4756–4767.

10. Burge CB, Karlin S. Finding the genes in genomic DNA. *Curr Opin Struct Biol*. 1998;8(3):346–354.

11. Collins FS. The Human Genome Project: Lessons from Large-Scale Biology. *Science*. 2003;300(5617):286–290. doi:10.1126/science.1084564.

12. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*. 2006;23(6):673–679.

13. Eddy SR. What is a hidden Markov model? *Nat Biotechnol*. 2004;22(10):1315–1316.

14. Fickett JW, Tung CS. Assessment of protein coding measures. *Nucleic Acids Res*. 1992;20(24):6441–6450.

15. Gilbert W. DNA sequencing and gene structure Nobel lecture, 8 December 1980. *Biosci Rep*. 1981;1(5):353–375.

16. Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 2000;22(1):4–37. doi:10.1109/34.824819.

17. Korbel JO, Urban AE, Affourtit JP, et al. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science*. 2007;318(5849):420–426. doi:10.1126/science.1149504.

18. Krogh A, Mian IS, Haussler D. A hidden Markov model that finds genes in E.coliDNA. *Nucleic Acids Res*. 1994;22(22):4768–4778. doi:10.1093/nar/22.22.4768.

19. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*. 2007;35(9):3100–3108. doi:10.1093/nar/gkm160.

20. Lodish H, Zipursky SL. Molecular cell biology. *Biochemistry and Molecular Biology Education*. 2001;29:126–133.

21. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 1997;25(5):0955–0964.

22. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*. 1998;26(4):1107–1115.

23. Mathé C, Sagot M-F, Schiex T, Rouzé P. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*. 2002;30(19):4103–4117.

24. Metzker ML. Sequencing technologies — the next generation. *Nature Reviews Genetics*. 2009;11(1):31–46. doi:10.1038/nrg2626.

25. Nakabachi A, Yamashita A, Toh H, et al. The 160-Kilobase Genome of the Bacterial Endosymbiont Carsonella. *Science*. 2006;314(5797):267–267. doi:10.1126/science.1134196.

26. Parfrey LW, Lahr DJG, Katz LA. The Dynamic Nature of Eukaryotic Genomes. *Molecular Biology and Evolution*. 2008;25(4):787–794. doi:10.1093/molbev/msn032.

27. Penttila KMKMM. Genome Sequencer Industry Evolution. 2005:1–12.

28. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*. 1989;77(2):257–286.

29. Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*. 1998;26(2):544–548. doi:10.1093/nar/26.2.544.

30. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*. 1975;94(3):441–448.

31. Schuster SC. Next-generation sequencing transforms today's biology. *Nat Meth*.

2007;5(1):16–18. doi:10.1038/nmeth1156.

32. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008;26(10):1135–1145. doi:10.1038/nbt1486.

33. Stein L. Genome annotation: from sequence to biology. *Nature Reviews Genetics*. 2001;2(7):493–503. doi:10.1038/35080529.

34. Venter JC. GENOMICS: Shotgun Sequencing of the Human Genome. *Science*. 1998;280(5369):1540–1542. doi:10.1126/science.280.5369.1540.

35. Wang Z, Chen Y, Li Y. A brief review of computational gene prediction methods. *Genomics Proteomics Bioinformatics*. 2004;2(4):216–221.

ÖZGEÇMİŞ

Adı Soyadı: Zeynep ÖZKESERLİ
Doğum Yeri: Bursa - Türkiye
Doğum Tarihi: 25.02.1986
Medeni Hali: Bekar
Yabancı Dili: İngilizce, Almanca, İtalyanca

Eğitim Durumu
Lise: Bursa Nilüfer Milli Piyango Anadolu Lisesi – Fen Matematik Bölümü
Lisans: Ankara Üniversitesi Fen Fakültesi İstatistik Bölümü

İş Tecrübesi
Kurumu: HGM Biyoinformatik **Görevi:** Veri Analiz Sorumlusu
Yılları: 2009 - 2010

Yayınlar:
SCI'da Yer Alan Makaleler:
2014 Genome Announcements
Col, B., Ozkeserli Z., Ozdag, H., Alakoc YD.
Genome Sequence of the Boron-Tolerant and -Requiring Bacterium Bacillus boroniphilus.

Uluslararası Kongrelerde Sunulan Bildiriler:
2011 NACD-2011, Ankara, Türkiye
Günseli Çubukçuoğlu Deniz, Serkan Durdu, Yeşim Doğan Alakoç, Çağın Zaim, Zeynep Özkeserli, Hakan Gürdal, Hilal Özdağ, Ahmet Rüçhan Akar
Comparative Atrial Gene Expression Profiles of Patients with Degenerative Mitral Regurgitation: Atrial Fibrillation Versus Sinus Rhythm International Symposium on New Approaches in Cardiovascular Disorders

2012 ISCB – RGS Turkey Student Symposium
Zeynep Özkeserli, Hilal Özdağ, H. Gökhan İlk
Hidden Markov Model Based Annotation Methods and an Application on a de novo Prokaryotic Genome Sequence (Sözlü Sunum)

Ulusal Kongrelerde Sunulan Bildiriler:

2010 36. Ulusal Hematoloji Kongresi

Çetinkaya F, Eğin Y, Özkeserli Z, Doğan A, Uysal S, Deniz GD, Alakoç YD, Durdu S, Özdağ H, Akar AR, Akar N.

Antikoagülan Kullanım Endikasyonu Olan Hastalarda VKORC1 C1173T Ve G1639A Gen Polimorfizmlerinin Farmakogenetik Etkisinin Araştırılması

2011 Ulusal Biyoloji Kongresi, Aydın, Türkiye

Alakoç YD., Özkeserli Z., Özdağ H., Çöl B.

Genomic Based Approach to the Subject of Boron in Life: Draft de novo Genome Sequence of a Highly Boron Tolerant and Requiring Bacterium (Poster Sunum)