Erciyes Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 37, Sayı 2 , 2021

Erciyes University
Journal of Institue Of Science and
Technology
Volume 37, Issue 2, 2021

NanoTAX: Nanogözenek DNA Dizileme Verisi Üzerinden Hızlı Patojen Tanıma

Özkan Ufuk Nalbantoğlu*[1], Aycan Gündoğdu[2]

*[1] Erciyes Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği, KAYSERİ
[2] Erciyes Üniversitesi Tıp Fakültesi Tıbbi Mikrobiyoloji, KAYSERİ

(Alınış / Received: 23.02.2021, Kabul / Accepted: 19.05.2021, Online Yayınlanma / Published Online: 31.08.2021)

Anahtar Kelimeler
Biyoinformatik,
Hızlı patojen tanıma,
Makine öğrenme,
Yeni nesil dizileme,
Nanogözenek dizileme

Öz: Bu çalışmada gerçek zamanlı DNA dizilemesi sağlayan, düşük maliyetli ve taşınabilir bir yeni nesil dizileme teknolojisi olan nanogözenek dizileme teknolojisini kullanarak gerçek zamanlı ve düşük maliyetli patojen/etken tespit algoritmaları sunulmaktadır. Çalışma kapsamında Oxford Nanopore MinION DNA dizileyicisi ile sekanslanan bakterileri gerçek zamanlı tanıyabilecek bilgi kuramı temelli biyoinformatik teşhis algoritmaları geliştirilmiş ve performansları gerçek veri üzerinde test edilmiştir. Taksona özgü oligomer karakterizasyonunu sağlayan Bağıl Bolluk Endeksleri (ing: Relative Abundance Index-RAI) ve nükleotid dizileri içindeki korelasyonları ortaya seren Ortalama Karşılıklı Bilgi (OKB) yöntemi ile DNA karakterizasyonu yapılarak patojen tanıma algoritmaları geliştirilmiştir. Tasarlanan simülasyonlar ile ortalama teşhis koyma süreleri ve doğrulukları hakkındaki istatistikler elde edilerek bu yönde oluşturulacak sistemlerin rutin kullanım için fizibilitesi ortaya konmuştur. Önerilen OKB profili ve RAI tabanlı algoritmaların hızlı patojen tanıma konusunda yeterli doğruluk seviyesinde ve kısa sürede tanıma yapabilecek hızda olduğu ve mevcut programlarla rekabet edebilen performansta olduğu nanogözenek dizilemesi yapılan patojen paneli üzerinde gösterilmiştir. Geliştirilen yöntem kısıtlı bir patojen paneli üzerinde etkinlik göstermektedir; bu sebeple geniş çaplı kullanım için daha ileri çalışmaların yürütülmesi gerekmektedir.

Rapid Pathogen Detection Algorithms for Nanopore DNA Sequencing

Keywords
Bioinformatics,
Rapid pathogen detection,
Machine learning,
Next-Generation sequencing,
Nanopore sequencing

Abstract: In this work, real-time and cost effective pathogen detection algorithms for low-cost, real-time nanopore DNA sequencing technology are proposed. Information theory-based bioinformatics detection algorithms that can recognize the bacterial species in DNA data generated by the current nanopore sequencing technology, Oxford Nanopore MinION sequencer, were developed and tested on real-data. Novel pathogen recognition algorithms were developed based on Relative Abundance Index (RAI), which characterizes taxon-specific oligomer preferences, and also based on Average Mutual Information (AMI), which reveals the correlations within nucleotide sequences. The simulations were designed to reveal the statistics on mean detection durations and accuracies of the detection algorithms. This provided know-how on the feasibility of routine usage of the systems developed on these principles. The proposed algorithms those are based AMI and RAI based detection systems were observed to be sufficiently accurate and rapid for pathogen detection on the selected pathogen panel being competitive with state-of-the-art. While the method operates on a limited panel, further studies are needed to broaden the use cases.

*İlgili Yazar, email: nalbantoglu@erciyes.edu.tr

1. Giriş

Enfeksiyon hastalıkları, dünya ölçeğinde yetişkin ölüm sebeplerinde birinci, çocuk ölümlerinde ise ikinci sırada gelmektedir. Bir yılda dünya çapında ölen 50 milyon insanın %40'ının enfeksiyon hastalıklarından öldüğü tahmin edilmektedir [1,2]. Enfeksiyon hastalıklarında tanı ve antimikrobiyal tedavinin uygun şekilde yönlendirilebilmesi için etkenin doğrulanması ve patojenin hızlı bir şekilde identifikasyonu oldukça önemlidir. Bu amaç için kullanılan standart mikrobiyolojik yöntemler kültür, konvansiyonel identifikasyon/duyarlılık testlerini içermektedir. Her ne kadar kültür altın standart olarak tanımlansa da standart yöntemlerle sonuç alınması için en 72 saate varan sürelere ihtiyaç duyulmakta ve etken organizma her zaman identifiye edilememektedir. Özellikle sepsis gibi kritik

vakaların tanısında ya da etken mikroorganizmanın hızlı ve doğru bir şekilde tanımlanmasının kritik öneme sahip olduğu biyogüvenlik/biyoterörizm durumlarında konvensiyonel yöntemler yetersiz kalmaktadır. Kültürde hız sorununa ek olarak kan, idrar, ETA, BOS gibi vücut sıvısı örneklerinin toplanmasından başlayarak son aşamaya kadar patojen tanımlamanın doğru sonuçlandırılmasını etkileyen çok sayıda faktör bulunmaktadır. Örneğin, fastidiyöz ve yavaş üreyen patojenlerin sebep olduğu enfeksiyonlarda etken izolasyonu için kültürün duyarlılığı oldukça düşüktür. Kültürün ortaya koyduğu teknik ve diğer sınırlamaların üstesinden gelmek için, birçok ülkede 1960'lı yıllardan bu yana etken/patojen hızlı tanı yöntemleri üzerine yapılan araştırmalar devem etmektedir [3]. Bu araştırmalar başlıca, minyatürize biyokimyasal teknikler, immünolojik testler, biyosensörler ve nükleik asit temeline dayanan yöntemleri içermektedir. Fakat, hızlı sonuç vermesi için tasarlanan söz konusu yöntemlerin duyarlılıkları, hedef mikroorganizma çeşitlilikleri ve rutinde kullanımı için maliyet etkinliklerinin değerlendirilmesi önemlidir [4]. Örneğin, etken tanısında DNA temelli yöntemler henüz rutin laboratuvar kullanımına girmemiş olmakla birlikte birçok hasta grubundaki hızlı tanı ihtiyacından dolayı çok sayıda araştırmada değerlendirilmiştir. Fakat, patojen spesifik yaklaşımla -tür ya da cins düzeyinde tasarlanan- primerlerin kullanıldığı DNA temelli metotlarda hedeflenen mikroorganizma çeşitliliği oldukça sınırlıdır. Bu sebeple, gerek klinik gerekse gıda ve çevresel örneklerinde bulunan patojen mikroorganizmaların erken tespiti/ identifikasyonu için mikrobiyolojik tanıya katkı sağlamak üzere ucuz ve duyarlılığı yüksek hızlı tanı sistemlerinin geliştirilmesine ihtiyaç vardır.

Geçtiğimiz on beş yıl, yeni nesil dizileme teknolojilerinde (YNDT) gerçekleşen gelişmeler ve bunun yansıması olarak sağlık bilimleri, gıda biyoteknolojisi, endüstriyel mikrobiyoloji vb. alanlarda önemli buluşların yapıldığı bir ilerleme dönemi olmuştur. YNDT, ticari olarak kullanıma girmeye başladığı 2004 yılından bugüne değin 1 milyon nükleotidin dizileme maliyetini 1000 Amerikan dolarından 5 Amerikan senti seviyesine indirirken, tek bir seferde dizileme hacmini de Mbp (milyon nükleotid~bakteri genomu) düzeyinden Tbp (trilyon nükleotid~1000 insan genomu) düzeyine taşımıştır [5]. Günümüzde küçük ölçekli moleküler biyoloji laboratuvarlarında dahi yeni nesil dizileme ile genom düzeyinde çalışma yapılabilmekte, fenotip-genotip ilişkileri ortaya konarak hastalık teşhisi/prognozu/tedavi araştırmaları moleküler düzeyde yürütülebilmektedir.

Geçtiğimiz birkaç yılda, DNA dizileme paradigmasında devrimsel bir gelişmeye sebep olabilecek yeni bir yaklaşım bilimsel kullanıma sunulmuştur. DNA sentezinin optik (Illumina, Roche 454, PacBio teknolojileri) veya elektrokimyasal (Ion Torrent teknolojisi) tespitine dayalı teknolojiler yerine herhangi bir DNA sentezlemeye dayanmayan, nanogözenekler [6] içerisinden geçen DNA iplikçiklerinin doğrudan elektronik tespitine dayalı Oxford Nanopore MinION teknolojisi inovatif bir DNA dizileme yöntemi olarak kullanılmaya başlanmıştır. Nanogözek dizilemesi, tarihi 25 yıl öncesine dayanan bir fikir olsa da [7], DNA dizileyen ulaşılabilir bir cihaz teknolojisine dönüşmesi ancak 2014 yılında mümkün olmuştur. Herhangi bir DNA amplifikasyonu, sentezleme veya optik tarama gerektirmeyen bu teknik, ucuz ve küçük cihazlarla gerçek zamanlı, hızlı ve uzun DNA okumaları sağlayan dizilemeyi olanaklı kılmaktadır. Nanogözenek dizileme yöntemini uygulayan ilk ticari DNA dizileyicisi olarak ortaya çıkan Oxford Nanopore MinION, bu teknoloji ile kısa sürede önemli bulguların elde edilebilmesini sağlamıştır. Örneğin, nanogözenek dizileme ile elde dilen uzun okumalar bakteri tüm genomlarının büyük bir başarıyla dizilenebilmesini sağlamaktadır. *Escherichia coli* K-12 MG1655 suşunun tüm genomu 48 saatlik MinION dizilemesi sonucunda elde edilmiş ve yaklaşık %87'sinin referans genomla örtüştüğü raporlanmıştır [8]. Aynı genomun daha gelişmiş biyoinformatik teknikleriyle birleştirilmesi sonucu %99.5 referans gen benzerliği ile birleştirilmesi mümkün olmuştur [9]. Patojen türlerin antibiyotik direncinden sorumlu antibiyotik direnç geni adalarının yapı ve genom pozisyonlarının bu dizileme teknolojisi ile tespit edilebildiği *Salmonella Typhi* genomu üzerinde gösterilirken [8], *Staphylococcus aureus* ve *Mycobacterium tuberculosis* genomlarının antibiyotik direnç tahminleri sırasıyla 12 antibiyotik için ortalama % 99.4 ve % 90.55 doğrulukla yapilabildiği rapor edilmiştir. Bu örnekler dışında bir diğer başarılı gelişme, ilk ortaya çıktığı süreçte MinION verisi düşük kaliteli okumalar üretiyorken [8,10] kimyası değiştirilmiş yeni akış hücreleri sayesinde bugün neredeyse 2.nesil dizileme cihazları kadar kaliteli okumalar elde edilebilmesi olmuştur.

Bütün bu gelişmeler ışığında MinION nanogözenek dizilemesinin gerçek zamanlı tüm genom dizilemesi yapabilme yetisi, bu teknolojinin patojen tespiti konusundaki potansiyelini ortaya dökmüştür. Bu doğrultuda başlayan öncü çalışmalarda başarılı sonuçlar elde edilmiştir [8]. İngiltere'de bir hastanede başgösteren Salmonella salgınını 100 dakika süren nanogözenek dizilemesi ile tespit etmeyi başarmışlardır. 2014 Ebola salgını izolatları ile yapılan çalışmalarda bir saatten kısa sürede Ebola virusunun tespit edilebildiği, aynı teknikle chikungunya ve hepatit C viruslarının da kısa sürede MinION dizilemesiyle tespit edilebildiği görülmüştür. İnfluenza genomunun dizilendiği bir diğer çalışmada 4 saatlik bir nanogözenek dizilemesi ile tüm virus genlerinin %99 doğrulukla ortaya çıkarılabildiği ve bu sayede virus tiplemesinin yapılabildiği belirtilmiştir [11]. 2020 yılında başgösteren COVID-19 pandemisinde SARS-CoV-2 genomlarının dizilenmesinde başlıca yöntem olarak nanogözenek dizilemesi salgın boyunca öne çıkmıştır [12,13].

Bugüne değin yürütülmüş çalışmalar göz önüne alındığında, patojen tanıma yöntemlerinin üretilen DNA okumalarının referans genomlar üzerine bilinen DNA hizalama programlarıyla hizalamaya dayandığı görülmektedir. Buna göre ortamdan okunan DNA dizilerinin önemli bir bölümünün, hedef bir patojenin genomuna hizalanabiliyor olması, o patojenin ortamda bulunduğuna işaret etmekte ve böylece patojen tespiti yapılabilmektedir. Uzun DNA parçalarını hizalamada kullanılan yöntemlerinden olan LAST [14] şu an nanogözenek dizilemesinde kullanılan en popüler yöntem olarak göze çarpmaktadır [8,15,16]. Bunun yanında BLAST [10,11,17] ve BWA-MEM [9,18,19] programları da hizalamada kullanıldığı rapor edilen benzer yaklaşımlardır. Bahsi geçen DNA hizalayıcıları aslen baz başına ortalama okuma hata oranı düşük olan (<< %1) Sanger ve Illumina dizileme teknolojileri için geliştirilmiş yöntemlerdir. Nanogözenek dizilemesinin mevcut teknoloji ile %15-%30 arasında değişen yüksek okuma hatası oranlarına sahip olduğu bildirilmektedir [10,20]. Hizalama algoritmalarının MinION gibi yüksek okuma hatası oranına sahip veriler üzerinde ise düşük başarıya sahip olması beklenmektedir. Nitekim rapor edilen nanogözenek dizilemesi deneylerinde üretilen veri ancak %10-%20 oranları arasında hizalanabilmiştir [20,15]. Bu oranda güncellenen yeni kimyasal prosedürlerle ufak iyileşmeler olsa da henüz ortalama olarak üretilen okumaların en fazla dörtte bire yakın bir bölümü [9] referans genomlar üzerine hizalanarak tespit için kullanılabilmektedir. Bu durum üretilen verinin büyük çoğunluğunun ortamdaki patojenlere ait önemli genom bilgileri içermesine rağmen bunların doğrudan kullanılamaması anlamına gelmektedir. Ebola, Chikungunya ve hepatit C virüslerinin tespiti için yapılan bir çalışmada MinION dizileme verisinin DNA hizalaması ile Illumina platformuna göre yaklaşık 3 kat daha az oranda hizalandığı gözlemlenmiştir [11]. Bu ise saatler olarak rapor edilen patojen tanıma süresinin üretilen verinin verimli kullanılabilmesi halinde dakikalarla ifade edilebilecek performansa erişebileceğini işaret etmektedir. Dolayısıyla patojen tanıma yöntemi olarak nanogözenek okumalarını referans genomlara hizalama yaklaşımından çok DNA kompozisyonuna dayalı örüntü tanıma ve makine öğrenimi gibi teknolojilerin kullanıldığı yeni yaklaşımlara ihtiyaç vardır. Bu tip yaklaşımlar, DNA okumalarını matematiksel modellerle karakterize ederek bilinmeyen bir DNA parçasının hangi genoma ait modelle örtüştüğünü bularak tespit yapmaya yönelik olması bakımından önemlidir. Patojen tanıma için gerekli taksonomik sınıflandırmayı hesaplamalı genom imzaları (HGİ) [21] tabanlı örüntü tanıma yöntemleriyle gerçekleştiren Phymm [22] ve RAIphy [23] gibi yaklaşımların gürültülü veriye dayanıklı olduğu gözlemlenmiştir. Bununla birlikte Ortalama Karşılıklı Bilgi (ing: Average Mutual Information-AMI) karakterizasyonun da yüksek hatalı dizilerde başarılı modelleme yapabildiği bilinmektedir [24, 25]. Bu sebeple bu yaklaşımların, MinION verisi üzerinde başarılı modelleme yapma ve dolayısıyla verimli patojen tanıma yöntemleri ortaya koyma potansiyeli bulunmaktadır.

Yürütülen çalışma ile gerçek zamanlı DNA dizilemesi sağlayan, düşük maliyetli ve taşınabilir bir yeni nesil dizileme teknolojisi olan nanogözenek dizileme teknolojisini kullanarak gerçek zamanlı ve ucuz patojen/etken tespit sistemlerinin geliştirilmesi amaçlanmıştır. Bu kapsamda, Oxford Nanopore MinION DNA dizileyicisi ile sekanslanan bakterileri gerçek zamanlı tanıyabilecek bilgi kuramı temelli biyoinformatik teşhis algoritmaları geliştirilmiş ve performansları gerçek veri üzerinde test edilmiştir.

2. Materyal ve Metot

Çalışmada izlenen metodoloji, önerilen yöntemin geliştirilmesi ve test edilmesi için gerekli (i) veri üretimi, (ii) üretilen DNA dizileme verisi üzerinden biyoinformatik yöntemler geliştirilmesi ve (iii) geliştirilen yöntemlerin tasarlanan simülasyonlarla performansının ölçülerek önerilen teknolojinin başarısının ve uygulanabilirliğinin kestirilmesi aşamalarını içermektedir.

2.1 Islak Laboratuvar Yöntemleri

Biyoinformatik tanıma algoritmalarının geliştirilmesinde gerekli eğitim verisinin üretilmesi için bu aşamada *E. coli, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Citrobacter freundii ve Enterobacter spp.* türlerine ait suşları*n* genomik DNA izolasyonu, çift iplikli DNA (dsDNA) kalite ölçümleri, dizileme kütüphanesinin hazırlanması ve MinION ile dizilemesi yapılmıştır.

2.1.1 Mikrobiyal DNA izolasyonu ve dsDNA Kalite Ölçümleri

E. coli, K. pneumoniae, A. baumannii, P. aeruginosa, C. freundii ve Enterobacter spp. suşlar uygun besi yerlerinde saf olarak kültürlenmiştir. Suşlara ait total bakteriyel DNA izolasyonu Mobio DNA İzolasyon Kiti (Mobio-PowerLyzer UltraClean Microbial DNA Isolation Kit) kullanılarak üretici firmanın prosedüründe belirtildiği şekliyle gerçekleştirilmiştir. İzolasyon sonrası elde edilen total bakteriyel DNA içerisindeki çift iplikli DNA (dsDNA) konsantrasyonu ve kalitesi Qubit dsDNA BR assay kit kullanılarak Qubit fluorometer (Life Technology) ile belirlenmiştir. Uygun kalite ve miktarda elde edilemeyen suşlar için izolasyon aşaması tekrarlanmıştır. Ölçüm sonuçlarına göre uygun kalitedeki (OD 260/280 = 1.8, OD 260/230 = 2.02.2) dsDNA örnekleri nanogözenek dizilemede kullanılmış ve saf dsDNA kullanılıncaya kadar -20 °C'de muhafaza edilmiştir.

2.1.2 Nanogözenek Tüm Genom Dizileme

Nanogözenek dizilemesi için Oxford Nanopore Technologies, MinION dizileme cihazı ve cihaza ait kontrol yazılımları kullanılmıştır. İzole edilen dsDNA, Oxford Nanopore SQK-LSK308 kiti ile firmanın önerdiği prosedüre uygun şekilde sekanslamaya hazırlanmıştır. Kit prosedürüne uygun olarak dsDNA'ları yaklaşık 8Kb'lik parçalara ayırmak için Covaris g-TUBE teknolojisi kullanılmıştır. DNA hasar ve kırıklarının onarılmasında NEBNext Ultra II End Repair/dA-Tailing, NEBNext FFPE DNA Repair Mix kitleri kulllanılmıştır. Kütüphane hazırlık işlemleri sırasında DNA saflaştırılması Dynabeads MyOne Streptavidin C1 manyetik boncuklar ile gerçekleştirilmiştir. Alternatif bir kütüphane hazırlık aşaması olarak, DNA parçalanması, hasar onarımı gibi işlemlere ihtiyaç duyulmayan kütüphane hazırlanmasına olanak sağlayan Oxford Nanopore SQK-RAD003 kiti kullanılmıştır. Her bir örnek için, MinION cihazına takılan MK-I FLO-MIN107 akış hücresi (flow cell) kullanılarak 48 saat boyunca nanogözenek sekanslaması yapılmıştır. Dizileme cihazının kontrol yazılımı olan MinKNOW programı üzerinden DNA okumaları FASTA formatında metin dosyalarına çevrilerek biyoinformatik analizler için hazır hale getirilmiştir.

2.2. Kuru Laboratuvar Süreçleri

Nanogözenek dizilemesi ile elde edilen DNA okumalarının ait olduğu mikorganizmayı veya taksonomik sınıfı hassas düzeyde tahmin ederek patojen tespiti yapan biyoinformatik yöntemler iki farklı yaklaşım üzerinden geliştirilmiştir. Bunlar Bağıl Bolluk Endeksi (ing: Relative Abundance Index-RAI) ve Ortalama Karşılıklı Bilgi (OKB) karakterizasyonlarına dayalı algoritmalardan oluşmaktadır.

2.2.1 Bağıl Bolluk Endeksi (RAI) Temelli Patojen Tanıma Algoritmaları

RAI, genomlar içerisindeki kısa oligonükleotidlerin organizmaya özgü bollukta bulundukları ve rastgele bir DNA fragmanının hangi türe (veya taksonomik birime) ait olduğunu bu karakteristiğe göre oluşturulmuş endeksler kullanılarak tespit edilebileceği gözlemine dayanan bir DNA karakterizasyon yöntemidir [23]. Bu yöntem kullanılarak geliştirilmiş olan RAIphy metagenom analiz programının mikroorganizmaların tespitinde oldukça hassas ve özgül olduğu raporlanmıştır [26,27].

RAI, bir oligonükleotidin bir genom veya taksonomik grup içindeki bağıl bolluğunu aşağıdaki şekilde hesaplar. $x_1x_2x_3...x_k$ k bazdan oluşan bir oligonükleotid ve $p(x_1x_2...x_k)$ bu oligonükleotidin genom içinde görülme olasılığı ise *i.* dereceden bağıl bolluk endeksi

$$\mathrm{rai}_i(x_1x_2...x_k)_2 \frac{p(x_1x_2...x_k)p(x_{\mathrm{k\text{-}1}}x_{\mathrm{k\text{-}2}}...x_{k-i})}{p(x_kx_{\mathrm{k\text{-}1}}...x_{k-i})p(x_1x_2....x_{k-1})} \tag{1}$$

ve toplu bağıl bolluk endeksi

$$\mathrm{rai}(x_1x_2...x_k)\sum_{\mathrm{i}=0}^{k-2}\mathrm{rai}_i(x_1x_2...x_k) \tag{2}$$

şeklinde gözlenen oligonükleotid frekanslarının farklı dereceden Markov modellerine dayanan beklenen değerlere oranları toplamı o genom (veya taksonomik ünite) içindeki her bir oligonükleotidin bağıl bolluk endeksini vermektedir. Orijini bilinmeyen bir DNA parçasının içerdiği oligonükleotidlerin bir genoma göre toplam endeks skoru, o DNA parçasının sözkonusu genoma göre RAI skorunu belirtir. Bilinmeyen fragman, RAIphy tarafından en yüksek RAI skorunu oluşturan genoma ait olarak sınıflandırılmaktadır.

RAI endeksleme algoritmasına dayanarak nanogözenek dizileri üzerinden patojen tanıması yapabilecek bir yöntem geliştirmek için öncelikle optimal modelleme yapabilen oligonükleotid uzunluklarının tespit edilmesi sağlanmıştır. Bunun için ıslak laboratuvar sürecinde elde edilen izolat genomları ve NCBI RefSeq veritabanından bugüne kadar dizilenmiş olan seçili patojen suşlarına ait genomlar indirilerek RAI veritabanlarının yapılandırılması gerçekleştirilmiştir. 7-13 arasındaki her oligonükleotid uzunluğu için farklı bir RAI veritabanı oluşturmuştur. Patojen suşların MinION dizilemesi ile üretilen DNA okumaları bu değerler için test edilerek en doğru tanımaların yapıldığı oligonükleotid uzunlukları tespit edilmiştir.

MinION dizilemesi DNA dizisinin nanogözenekler içerisinden geçerken anında görüntülenebilmesi imkanı sunmaktadır. Dizileme sırasında okunan DNA parçasının uzunluğu arttıkça çevrimiçi olarak RAI endeks skorları güncellenebilir. Gerçek profile yakınsayan (veya yanlış profillerden ıraksayan) bu türevsel skorların tanıma da kullanılması ile dizileme sırasında tespit sürelerinin elde edilmesi sağlanmıştır. Buna göre her okuma için her saniye örneklenen RAI skorlarının birinci derece farkları da sınıflandırma yöntemine entegre edilerek sınıflandırma performansına olan olumlu etkisi gözlemlenmiştir.

2.2.2 Ortalama Karşılıklı Bilgi (OKB) Temelli Patojen Tanıma Algoritmaları

OKB, bir DNA parçasındaki nükleotidlerin birbirleriyle olan korelasyon ilişkilerini ölçen bir bilgi kuramı yöntemidir. Buna göre bir DNA dizisinde birbirinden *k* nükleotid uzaklıkta bulunan iki nükleotidin (x,y) birbirleri hakkında içerdikleri bilgi

$$I_k = \sum_{x\in A}\sum_{y\in A} p_k(\mathrm{x,y})\log_2 \frac{p_k(\mathrm{x,y})}{p(x)p(y)}, \mathrm{A}=\{\mathrm{A,C,G,T}\} \quad (3)$$

şeklinde *k* nükleotid uzaktaki x ve y'nin empirik olarak kestirilmiş bileşik ve marjinal olasılık dağılımları üzerinden hesaplanabilmektedir.

OKB temelli patojen tanıma algoritmalarının geliştirilmesi için RAI temelli algoritma geliştirme aşamasına benzer adımlar atılmıştır. Öncelikle üretilen MinION verisi üzerinden farklı profil uzunluklarında patojen tanıma performansları araştırılmıştır. OKB profillerinin tüm genom profillerine yakınsama verilerinin sınıflandırmaya olumlu etkisinin olup olmadığı yine RAI skorlarında kullanılan yöntemle araştırılmıştır. OKB profillerinin uzaklıklarının ölçümünde Öklid, L1 normu ve Pearson korelasyonu metrikleri kullanılmıştır.

2.2.2.1 OKB Profillerini Kullanan Makine Öğrenme Temelli Patojen Tanıma Yöntemleri

Farklı patojenlere ait OKB profillerinin sınıflandırılabilmesi için üretilen okumalara ait OKB profillerinden rastgele örnekleme ile altkümeler seçilmiş ve belirlenen makine öğrenme programlarında eğitilerek çapraz geçerlilik testleriyle performansları sınanmıştır. Kullanılan makine öğrenme teknikleri sırası şu şekildedir: Naif Bayes algoritması [28], Destek vektör makineleri [29], en yakın komşu algoritması [30], Karar ağaçları [31] ve Rastgele orman algoritması [32].

2.3 Hızlı Patojen Tanıma Yöntemlerinin Performansının Ölçülmesi

Bu aşamada geliştirilen patojen tanıma algoritmalarının eldeki patojenleri hangi sürede ne ölçüde tanıyabilecekleri üretilen veriler üzerinde simülasyon yapılarak test edilmiştir.

Patojen dizileme verisinin üretilmesi sırasında, DNA okumalarının okunma zamanları ve okunma sıraları MinKNOW yazılımı ile kayıt altına alınmıştır. Bir dizileme sürecin başlamasından bitişine kadar süren her saniye için o anki kümülatif veri geliştirilen yazılımla analiz edilerek her bir okumanın taksonomik tespitleri kayıt altına alınmıştır. Yapılan simülasyonlar ile profil yakınsama karakteri ve RAI skorlarının evrimi gözlenerek her bir patojen için kritik değerler belirlenmiştir. Bu değerlere erişim süreleri yine dizileme süreleri ve okuma sıraları kayıtları kullanılarak ortaya çıkarılmış, böylece yaklaşık tanıma süreleri hesaplanmıştır.

2.4 Uygulama

Geliştirilen yaklaşımla nanogözenek dizileme verisinden taksonomik analiz yapabilen yöntemler bir yazılım paketi haline getirilmiş, "nanoTAX" isimli yazılım https://github.com/nalbant/nanoTAX adresli github versiyon platformunda dağıtıma sunulmuştur.

3. Bulgular

Yürütülen altı farklı nanogözenek dizilemesi deneyi belirtilen üretici firma tarafından sağlanan MinKnow DNA dizileyici sürücü yazılımı ile gerçekleştirilmiştir. Her bir izolat için MinKnow prosedürleri içerisinde akış hücrelerini (flowcell) tam kapasite ile kullanan 48 saatlik dizileme prosedürü seçilmiş ve dizilemelerin her biri 48 saat boyunca yürütülmüştür. Deneylerde kullanılan MinION akış hücrelerinin her biri 4'erli setler halinde 512 dizileme kanalından oluşmaktadır. Verili bir dizileme anında her kanaldaki 4 nanogözeneğin o anda en sağlıklı olanı sürücü yazılım tarafından tespit edilerek kullanılmaktadır. Yürütülen 6 dizileme deneyinin tamamında başlangıç değeri olarak 300 kanaldan fazlasının sağlıklı olarak çalıştığı görülmüştür. Dizileme deneleri sonucunda deney başına yaklaşık 40.000 (ortalama 39.329 +/- 9.743) DNA okuması elde edildiği, bu okumaların ortalama olarak 10.522 baz çifti (bç) uzunluğunda olduğu görülmüştür. Okuma sayıları *E. coli* için 49.675 bç, *K. pneumoniae* için 27.975bç, *A. baumannii* için 50.062bç, *P. aeruginosa* için 41.315bç, *C. freundii* için 29.669bç ve *Enterobacter spp* için 35.476bç olarak kayıt edilmiştir.

Elde edilen okuma uzunlukları göz önüne alındığında 1000 bç'den daha uzun okumaların tüm okumaların %99,27'sini, 5000 bç'den daha uzun okumaların tüm okumaların %83,86'sını, 10.000 bç'den daha uzun okumaların tüm okumaların %49,01'ini oluşturduğu görülmektedir (Ek A). En uzun okuma ise 37.455 bç olarak kaydedilmiştir. Buna göre okumaların yaklaşık yarısının 10 Kbç'den uzun olduğu görülmektedir. Elde edilmiş olan bu DNA dizileme uzunlukları nanogözenek dizilemesi hakkında son yıllarda elde edilen raporlarla tutarlı olmakla

birlikte bir genom parçasına ait göreceli olarak uzun parçaların okunabiliyor olması biyoinformatik uygulamaların önündeki belli kısıtları ortadan kaldırılmasının önünü açabilecektir.

Çalışmada yürütülen deneylerdeki hata oranını tahmin edebilmek amacıyla dizilenen suşlar NCBI veri tabanındaki dizilenmiş genomlarla Blastn algoritması kullanılarak karşılaştırılmıştır. Tamamı hizalanabilen okumaların nükleotid varyasyonlarına bakılarak kestirilen hata oranı ortalama olarak %8,71 +/- 2.35 olarak elde edilmiştir.

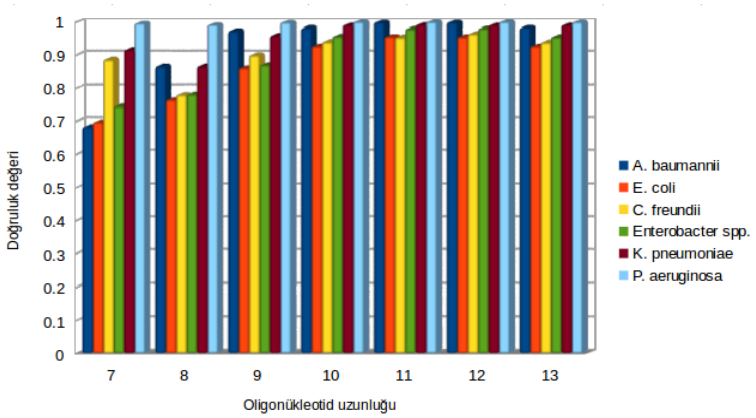
4.2 Bağıl Bolluk Endeksi (RAI) Temelli Patojen Tanıma Algoritmasının Geliştirilmesi

RAI endeksi kullanılarak bir patojen tanıma programı oluşturulmuş ve söz konusu algoritma belirlenen 6 patojen için test edilmiştir. Geliştirilen program farklı RAI uzunluğu parametreleriyle (oligonükleotid uzunluğu 7'den 13'e kadar) farklı veri tabanları oluşturmuş ve patojen tanıma deneyleri DNA okumasının ortalama doğru sınıflandırılma sayısı kriteri üzerinden test edilmiştir. Çalışma kapsamında yürütülen testlerde oligonükleotid uzunluğu 12'yi aştıktan sonra doğruluk payının doyuma ulaşıp daha da gerilemeye başladığı tespit edilmiştir (Tablo 1).

Tablo 1: Oligonükleotid uzunluğuna bağlı olarak RAI endeksi tabanlı sınıflandırma algoritmasının sınıflandırma doğruluk değerleri.

	Oligonükleotid uzunluğu						
	7	8	9	10	11	12	13
A. baumannii	0.682	0.866	0.972	0.981	0.999	0.999	0.982
E. coli	0.696	0.763	0.861	0.926	0.955	0.954	0.926
C. freundii	0.886	0.778	0.899	0.939	0.953	0.963	0.938
Enterobacter spp.	0.744	0.78	0.87	0.956	0.978	0.979	0.954
K. pneumoniae	0.916	0.866	0.958	0.99	0.991	0.99	0.99
P. aeruginosa	0.995	0.99	0.998	1	1	1	0.999

RAI endeksi 12 bazdan oluşan oligonükleotid veri tabanıyla yürütülen deneyde test edilen 6 patojenin ortalama %98 doğrulukla tespit edilebildiği gözlenmiştir. Elde edilen tüm nanogözenek okumaları üzerinde yürütülen blastn hizalaması deneylerinde ortalama okuma hata oranı %8.71 olarak bulunmuştur. Bu ise yine ortalamada 11.48 bazdan birinin hatalı olacağı anlamına gelmektedir. Bu hata oranı 11 ve 12 bazlık RAI profillerinde doğruluk artışını büyük oranda etkilemezken 13 bazlık oligonükleotid profilinden itibaren sonuçlara etki ettiği ve doğruluğu bir miktar düşürdüğü görülmektedir (Şekil 1).



Şekil 1. Oligonükleotid uzunluğuna bağlı olarak RAI endeksi tabanlı sınıflandırma algoritmasının sınıflandırma doğruluk değerleri.

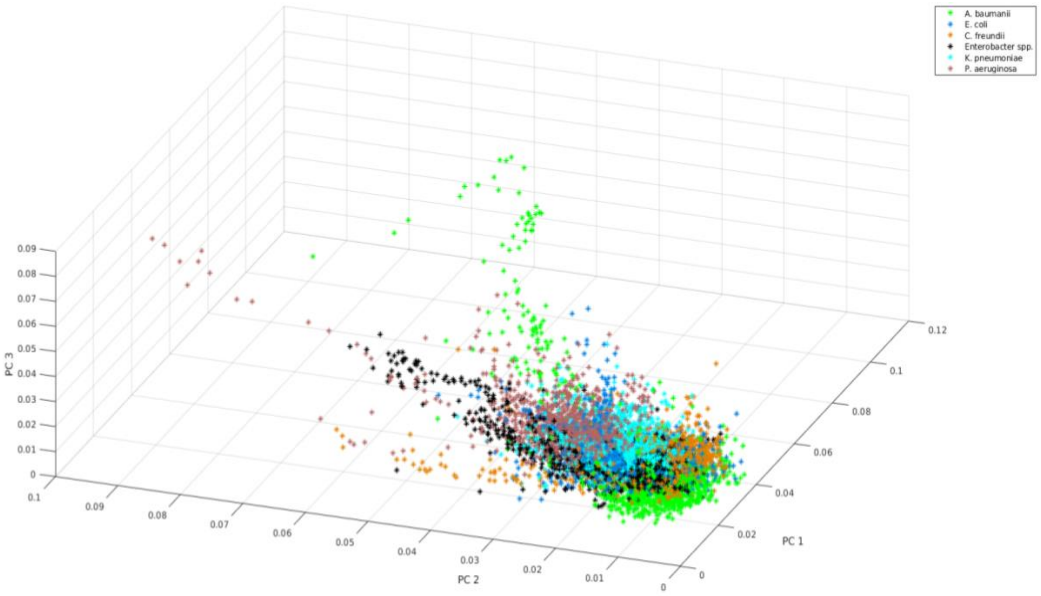
Nanogözenek DNA dizileme teknolojisi için son zamanlarda ortaya çıkarılmış olan okuma tanıma algoritmalarının en yaygın ikisi olan Minimap [33] ve WIMP (Kraken) [34] programları çalışma kapsamında üretilen 6 patojene ait deney verisi kullanılarak sınıflandırılmış ve doğruluk değerleri elde edilmiştir (Tablo 2). Elde edilen sonuçlara göre RAI endeksi tabanlı yaklaşımın WIMP programından daha başarılı Minimap'ten ise marjinal bir fark ile rekabet edebilir düzeyde daha başarılı olduğu gözlenmiştir.

Tablo 2: RAI endeksi tabanlı sınıflandırma algoritmalarının performansının WIMP ve Minimap sınıflandırma programları ile karşılaştırılması.

	A. baumannii	*E. coli*	*C. freundii*	*Enterobacter spp.*	*K. pneumoniae*	*P. aeruginosa*
RAI	0.999	0.954	0.963	0.979	0.99	1
WIMP	0.888	0.718	0.675	0.807	0.956	1
Minimap	0.944	0.897	0.911	0.9	0.923	1

3.1 OKB Temelli Patojen Tanıma Algoritmalarının Uzaklık/Benzerlik Metrikleri ile Patojen Tanımada Kullanılması

Tanıma için belirlenmiş altı patojene ait nanogözenek dizilemesi sonucunda elde edilen DNA okumaları ayrı ayrı OKB profillemesine tabii tutulmuş ve bu profillerin patojen tanıma algoritmalarındaki performansları test edilmiştir. Gözlenen DNA okumaları arası varyasyona rağmen genel OKB formunun korunuyor olması belli bir ölçekte patojen tanıma yapılabileceği görüşünü ortaya atmaktadır. Bu potansiyeli görsel olarak gözönüne sermek amacıyla elde edilen tüm OKB profilleri temel bileşenler analizi ile incelenmiş ve ilk 3 temel bileşen 3 boyutlu uzayda görüntülenmiştir. Şekil 2'de temel bileşenler analizi farklı perspektiflerden görselleştirilmektedir.



Şekil 2. Dizilenen OKB profillerinin temel bileşenler analizi ile görselleştirilmesi

İlk üç temel bileşenin görüntüsü okumaların belli bir kümeleşme gösterdiğini, ancak ayrı ayrı okumaların belli oranda iç içe geçtiklerini dolayısıyla bir patojen tanıma algoritmasında yanlış tespit sebebi olabileceklerini göstermektedir. Bununla birlikte profilleri uzayın farklı bölgelerine en çok dağılan türün A. *baumannii* olduğu ve iki ayrı öbek halinde profiller oluşturduğu görülmektedir.

Veri görselleştirmesi yolu ile elde edilen bu sonucun OKB temelli bir patojen tanıma algoritmasına nasıl yansıyacağını tespit etmek amacıyla farklı metrikler üzerinden patojen tanıma programları oluşturulmuş ve elde edilen DNA okumaları ile test edilmiştir. Sözkonusu programlar öncelikle 6 farklı patojen için ortalama OKB profillerini her bir patojenin imzası olarak kabul etmiş ve bu şekilde bir veri tabanı oluşturmuştur. Nanogözenek dizilemesi ile elde edilen bir DNA okumasının OKB profili bu veri tabanıyla karşılaştırılarak bilinmeyen bu DNA dizisinin hangi patojene ait olduğu tespit edilmeye çalışılmıştır.

İlk aşamada elde edilen ortalama genom imzaları ile her bir okuma OKB profilini Öklid vektör uzaklığı metriğiyle karşılaştırmaktadır. Tablo 3'de farklı parametreler kullanılarak Öklid metriği ile elde edilen duyarlılık ve kesinlik performansları sunulmaktadır.

Tablo 3. OKB profillerinin Öklid, L1 ve Pearson korelasyonu metrikleri kullanılarak sınıflandırılmasına ait sınıflandırma performansları.

Öklid Metriği (Duyarlılık % /Kesinlik %)						
	A. baumannii	***E. coli***	**C. *freundii***	***Enterobacter spp.***	***K. pneumoniae***	***P. aeruginosa***
1-100 bç	55,6/71,6	25,7/33,1	62,3/55,6	52,4/47,1	48/60,3	98,9/69,7
1-50 bç	57,9/69,8	23,6/31,9	63,4/54,6	50/49,6	50,9/57,4	98,6/71,7
1-10 bç	41,1/59,1	18,1/27,5	64/46,7	44,1/48,2	50,4/48,7	97,5/74,3
L1 Metriği (Duyarlılık % /Kesinlik %)						
	A. baumannii	***E. coli***	***C. freundii***	***Enterobacter spp.***	***K. pneumoniae***	***P. aeruginosa***
1-100 bç	52,2/67,5	34,3/41	63,3/55,9	50,4/46,6	46,9/57,8	85,4/62,6
1-50 bç	61,4/72,8	23/32,1	64,6/55,1	49,4/45,9	51,2/58,2	93,9/71,6
1-10 bç	44,3/60,4	17,5/26,5	64,7/47,5	43/47	54/49,8	96,8/77,4
Pearson korelasyonu (Duyarlılık % /Kesinlik %)						
	A. baumannii	***E. coli***	***C. freundii***	***Enterobacter spp.***	***K. pneumoniae***	***P. aeruginosa***
1-100 bç	31,2/89,9	7,4/40	89,6/28,2	15,1/84,4	37,1/46,5	98,1/74,5
1-50 bç	22,1/80,4	6,3/38,4	90,8/26	16/79,2	31/44	94,5/81,3
1-10 bç	4,9/45,8	0,4/5,6	90,6/23,5	21/56,9	23/31,9	76/87,5

Tablo 3'te sırasıyla 100 bç, 50 bç ve 10 bç uzaklığa kadar olan OKB profilleri test edilmiştir. Buna göre Öklid metriğine ait ölçümlerde il 100 ve 50 bç uzaklığa ait OKB profillerinin önemli ölçüde benzer sonuçlar verdiği, profilin ilk 10 elemanı göz önüne alındığında ise bu performansın bir miktar düştüğü görülmektedir. Öklid metriğine göre en başarılı patojen tanımanın P. *aeruginosa* genomu ile gerçekleştirildiği görülmektedir. *P. aeruginosa* izolatının sekanslanmasından elde edilen DNA okumalarının önemli kısmı (%98 civarı) doğru şekilde sınıflandırılmıştır, ancak bu oranın öbür türlerde oldukça azaldığı, hatta *E. coli* okumalarında *C. freundii* sınıflandırılması yapılma oranının *E. coli* oranından yüksek olduğu görülmüştür. Bunun dışında teste tabii tutulan diğer 5 patojenin de çoğunlukla doğru sınıflandırıldığı ve en yüksek tanınma oranına gözönünde bulundurarak patojen tanıma gerçekleştiren bir nanogözenek yazılımının *E. coli* dışındaki tüm patojenleri doğru tanıyabileceği gözlenmiştir.

OKB profillerinin en baskın elemanı olan profilin ilk elemanı (ardışık nükleotidlerin birbiri ile olan birlikteliğini ölçen OKB parametresi) gözardı edildiğinde elde edilecek sonuçlar incelendiğinde yani ilk 100 bç, 50 bç ve 10bç uzaklıklara kadar olan 99, 49 ve 9 OKB ölçümü profilleri test edildiğinde ortalama tanıma oranlarında bir önceki yaklaşıma oranla belirgin bir düşüş olmakla birlikte bu kez *E. coli* patojeninin çoğunlukla doğru tahmin edildiği ancak bu kez *Enterobacter spp.* okumalarının *C. freundii* ile karıştırıldığı görülmektedir. Bu trend maksimum ölçüm uzaklığı 10 bç uzaklığa indirildiğinde değişmekte ve yanlış tanımlanan patojen tekrar *E. coli* olarak ortaya çıkmaktadır (Ek A).

Aynı deneylerin L1 metriği ile tekrar edilmesi ile en başarılı tanıma yapılan patojenin P. *aeruginosa* olduğu görülmektedir. Öklid metriğinde gözlemlenen trende benzer şekilde *E. coli* okumalarında *C. freundii* sınıflandırılması yapılma oranının *E. coli* oranından yüksek olduğu görülmüştür. Bunun dışında teste tabii tutulan diğer 5 patojenin de çoğunlukla doğru sınıflandırıldığı ve en yüksek tanınma oranına gözönünde bulundurarak patojen tanıma gerçekleştiren bir nanogözenek yazılımının *E. coli* dışındaki tüm patojenleri doğru tanıyabileceği gözlenmiştir.

Ardışık baz çifti birlikteliğini ölçen ilk OKB elemanı elendiğinde ise (Ek A) 100 bç uzaklığa kadar elde edilen 99 bileşenli patojen tanıma metriğinin, ortalama tanıma başarısı düşmüş olsa da, tüm patojenleri en çok doğru patojen sınıflandırılacak şekilde doğru tanıyabildiği görülmektedir. Bu doğruluk trendi 50 bç profil uzaklığında hafifçe kaysa da 10 bç uzaklıktaki elemanların bulunduğu ölçümlerde de aynı şekilde gözlemlenmektedir. Bu yönüyle L1 metriği okumalar tek tek ele alındığında Öklid metriğinden daha başarısız olarak ortaya çıksa da, izolatlardan elde edilecek kolektif sonuçlarda tespiti daha doğru yapabileceği görülmektedir.

Son olarak her bir okumaya ait OKB profilinin veri tabanındaki patojen OKB imzalarına olan benzerlikleri Pearson korelasyon ölçümü ile belirlenmiş ve okumaların hangi patojene ait oldukları en büyük korelasyon ölçüsüne dayanılarak tespit edilmiştir. Ortalama tanıma başarılarının Pearson korelasyonu ile diğer iki metriğe göre daha düşük olduğu gözlenmektedir. Bununla birlikte diğer metriklerle bezer şekilde *P. aeruginosa* başarılı bir şekilde

tespit edilebilmekteyken yapılan altı farklı parametreli deneylerde de toplu tespitlerin bu metrik ile yanlış bir şekilde yapılacağı görülmektedir. Tüm bu sonuçlar göz önüne alındığında Öklid ve L1 metriklerinin belli ölçüde başarılı olduğu ancak Pearson korelasyonunun başarılı bir patojen tanıma yeteneğinde olmadığı sonucuna erişilmektedir.

Bir hızlı patojen tanıma senaryosunda enfeksiyon unsuru/çevresel patojen ortamında izole edilebiliyor veya önemli oranda zenginleştirilebiliyor/kültürlenebiliyor ise önerilen sistem başarılı olarak kullanılabilecektir. Ancak birçok hızlı patojen tanıma senaryosunda kültür/zenginleştirme için zaman kısıtı olduğu ve doğrudan ortamdan (örn. Hasta dokusu, serumu, çevresel örnek vb.) izole edilen DNA içeriğinin patojen tanımada kullanılabilmesi öngörülmekte ve bu yetiye sahip hızlı patojen tanıma sistemlerinin geliştirilmesi ihtiyacı duyulmaktadır. Bu durumda izole edilen DNA'nın yalnızca patojene ait olmaması, kommensal türlere ait DNA içeriğinin de dizileme sistemi ile elde edilmiş olması beklenmektedir. Böyle bir DNA karışımı /metagenom içeriği senaryosunda ise her bir okumanın önemli doğrulukla sınıflandırılabilmesi önem taşır hale gelmektedir. OKB profillerinin proje önerisinde belirtildiği ve test edildiği gibi oluşturulan veri tabanlarından belli uzaklık/benzerlik metrikleri kullanarak patojenlere sınıflandırılması dışında alternatif bir yaklaşım olarak makina öğrenme/yapay zeka teknikleri ile söz konusu DNA okumalarını daha yüksek doğrulukla tespit edebilecek zeki yazılımların üretilmesinin mümkün olup olmadığı da proje içerisinde irdelenmiştir.

3.2 OKB Profillerinin Makine Öğrenme Teknikleri ile Patojen Tanımada Kullanılması

Yürütülen 6 farklı patojene ait DNA okumalarının OKB profillerinden bir veri tabanı oluşturularak seçilen 8 farklı makine öğrenme algoritması ile bu profilleri birbirinden ayırabilecek yazılım modelleri oluşturulmuş ve ortaya çıkarılan programların performansları DNA okuması başına başarılı tespit ölçütüyle test edilmiştir.

OKB profillerinin sınıflandırması için kullanılan algoritmalar sırasıyla Naif Bayes sınıflandırıcısı (ing. Naive Bayesian), Doğrusal destek vektör makineleri (ing. Linear Support vector machines), Karesel destek vektör makineleri (ing. Polynomial -order2- Support vector machines), en yakın komşu sınıflandırıcısı (ing. 1NN algorithm), en yakın 3 komşu sınıflandırıcısı (ing. 3NN algorithm), en yakın 5 komşu sınıflandırıcısı (ing. 5NN algorithm), karar ağaçları (ing. J48 algorithm) ve rastgele orman sınıflandırıcısı (ing. Random forest classifier) olarak belirlenmiştir. Test amacıyla her bir DNA dizileme veri seti rastgele 10 parçaya ayrılarak 10 katlı çapraz geçerlilik testleri yürütülmüştür. Bu amaçla geliştirilen yazılım her model için toplam başarım sonucunu ROC eğrisi altında kalan alan ile ölçülmüştür. Yapılan her deneyden bir patojene 1000 DNA dizisi gelecek şekilde rastgele seçilen 6000 okuma ile yapılan ve bir önceki bölümde gerçekleştirilen testlerde yürütülen parametre seçimlerine paralel şekilde 100 bç uzaklığa kadar belirlenen OKB profilleri alınmış, profilin tamamı 50bç uzaklığa kadar olan kısmı, 10 bç uzaklığa kadar olan kısmı ve aynı profillerin ardışık baz çiftleri göz önünde bulundurulmadan elde edilen profilleri ayrı ayrı tüm sınıflandırma algoritmaları ile eğitilerek elde edilen programlar test edilmiştir (Tablo 4).

Tablo 4. OKB profillerinin makine öğrenme teknikleriyle sınıflandırılmasına ait ROC eğrisi altında kalan alan ölçümleri.

	Ardışık baz çiftleri ile			**Ardışık baz çiftleri olmadan**		
	1-10 bç	**1-50 bç**	**1-100 bç**	**2-10 bç**	**2-50 bç**	**2-100 bç**
Naif Bayes	0.822	0.818	0.81	0.805	0.812	0.806
Destek vektör makinesi (doğrusal)	0.865	0.911	0.918	0.857	0.908	0.914
Destek vektör makinesi (karesel)	0.854	0.918	0.927	0.838	0.913	0.923
En yakın komşu	0.871	0.922	0.929	0.858	0.921	0.926
En yakın 3 komşu	0.924	0.952	0.952	0.914	0.951	0.951
En yakın 5 komşu	0.939	0.96	0.959	0.933	0.959	0.958
Karar ağacı	0.84	0.833	0.832	0.835	0.836	0.832
Rastgele orman	0.96	0.974	0.974	0.955	0.973	0.974

Yürütülen testlere göre OKB profillerinin oluşturulduğu maksimum uzaklık arttıkça patojen tanıma performansında da belirgin bir artış trendi olmakta ve genel olarak ardışık baz çiftlerinin birliktelik bilgisinin de profil içerisinde bulundurulması daha başarılı karakterizasyona yol açmaktadır. Rastgele orman algoritması ile oluşturulan patojen tanıma programı deneyleri yapılan bakteri paneli içerisinden nanogözenek dizilemesi ile

okunmuş herhangi bir rastgele DNA parçasının ortalama %97,4 başarıyla hangi patojene ait olduğunun tespit edilebilmektedir. Bu sonuçlar ile, kültür olmaksızın ortamdan izole edilen DNA'dan enfeksiyon etkeni bir patojenin veya çevresel bir örnek içerisinde bulunan bir patojenin doğrulukla tespit edilebilmesinin mümkün olduğu, nanogözenek dizilemesi ve oluşturulan OKB profili temelli biyoinformatik yöntemin bu tespit için önemli potansiyel taşıdığı yargısına varılabilir.

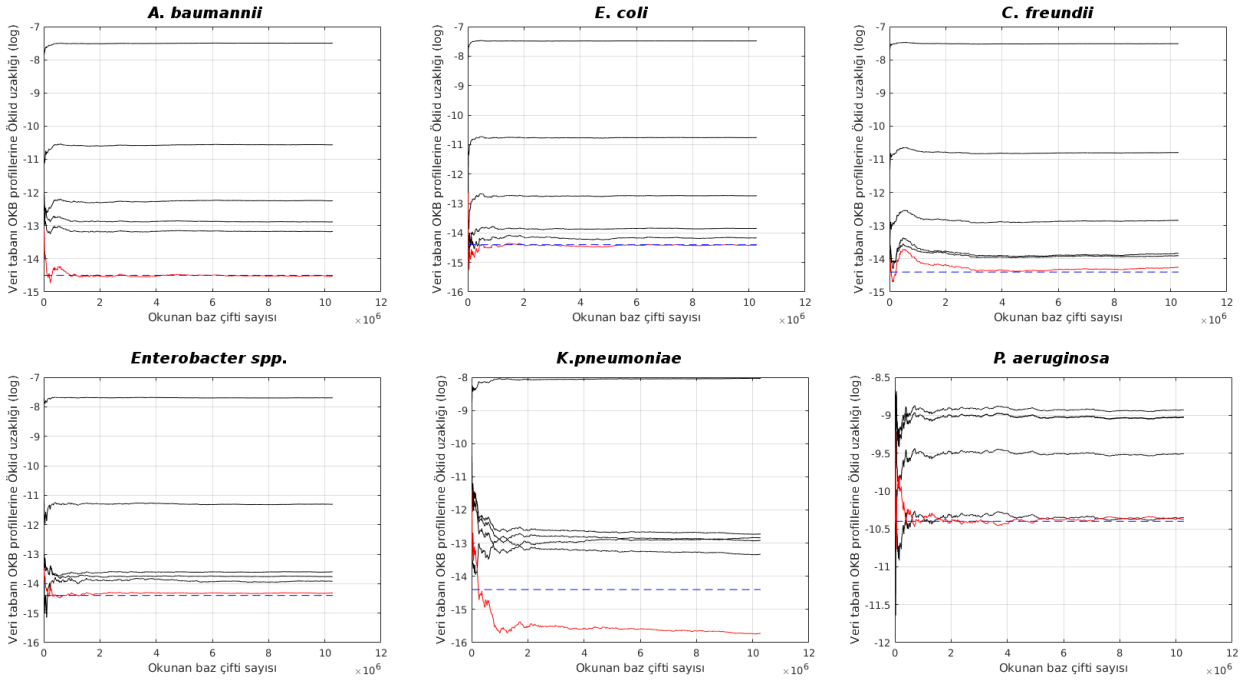
DNA içerisindeki uzak korelasyonlar ve türe özgülük

Yapılan testlerde elde edilen ilginç yan bulgulardan biri OKB profillerinde uzak nükleotid çiftlerinin birliktelik bilgisinin de patojen tanıma başarımını arttırdığı gözlemidir. Yukarıda sunular sonuçlar ele alındığında belli durumlarda genom içerisindeki 50 nükleotidden daha uzakta bulunan baz çiftlerinin arasındaki uzak korelasyonlar da tanıma başarısına katkıda bulunabilmektedir. Buradan hareketle bakteriyel genomlar içerisindeki belli uzaklıktaki nükleotid çiftlerinin arasında dahi türe özgü korelasyonlar olduğu öne sürülebilir. Bu durumu gözlemlemek amacıyla orta-uzak korelasyonları ele alan, yani ilk 20 OKB bileşenini kullanmayan, yalnızca 20 bç uzaklıktan 100bç uzaklığa kadar uzak nükleotid çiftlerinin birlikteliğini ölçen OKB profilleri kullanılarak rastgele orman algoritması ile patojen tanıma programı daha önceki prosedüre benzer şekilde tekrarlanmıştır. Elde edilen 0.958 değere sahip ROC alanı, orta-uzak nükleotid çiftlerinin benzerliğinin de türe özgü olup genom içerisinde korunduğunu işaret etmektedir.

Aynı deney uzak nükleotid çiftlerinin benzerliği için tekrarlandığında, yani birbirinden 90 nükleotidden daha uzak baz çiftleri kullanılarak tanıma deneyi yapıldığında yine 0.9 ROC alanı üzerinde bir başarımla sınıflandırma elde edildiği görülmektedir. Dolayısıyla orta-uzak nükleotid korelasyonları için elde edilen yargıya uzak nükleotidler için de varmak mümkün görünmektedir.

3.3 Hızlı Patojen Tanıma Performansı

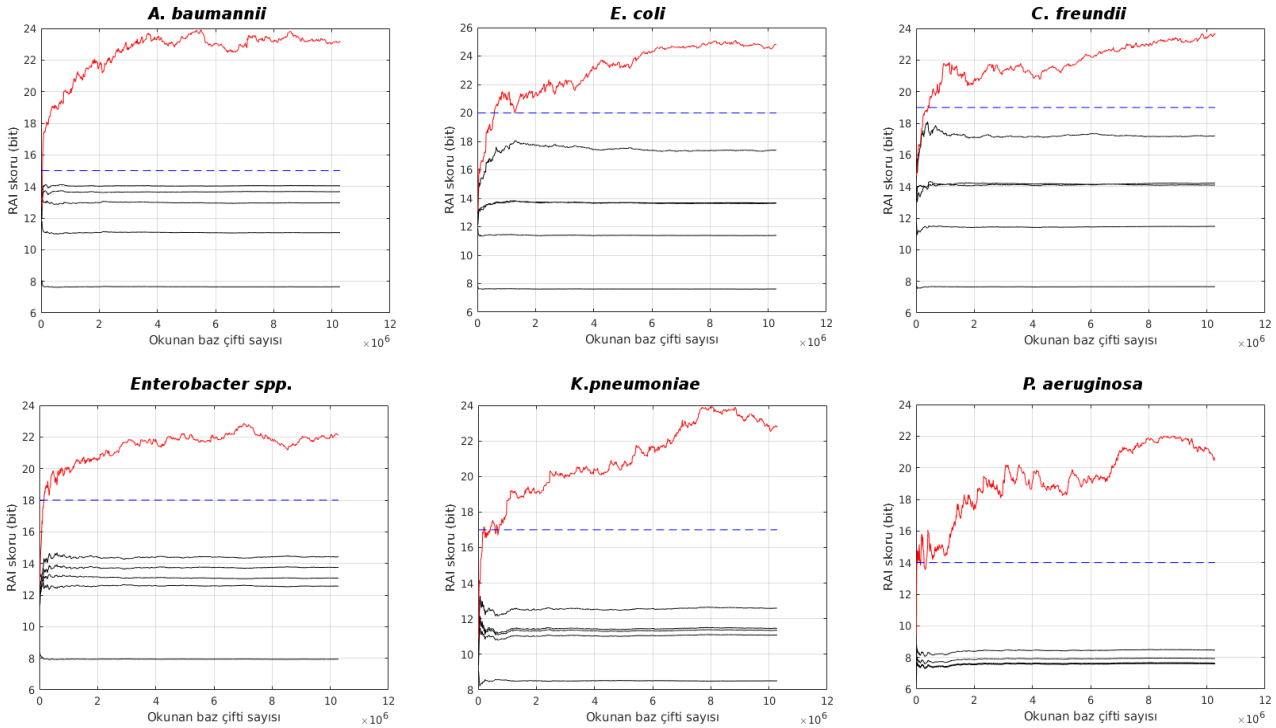
Geliştirilen OKB ve RAI temelli patojen tanıma yöntemlerinin DNA okuma tanıması performansı göz önünde bulundurularak yapılan deneylerde yeterli çoklukta dizileme verisinin bulunması halinde patojenlerin başarılı bir şekilde tespit edilebileceği yapılan deneylerle tespit edilmiştir. Proje kapsamında elde edilen patojen dizilerinin üretilme kaydı MinION cihazının sürücü yazılımı olan MinKNOW sistemi ile kayıt altına alınmış ve her saniye okunan DNA dizileri zaman imzaları (saniye:dakika:saat:gün:ay:yıl) şeklide arşivlenmiştir. Bu sayede retrospektif simülasyon mümkün olmuş ve dizilenen patojenlerin güvenli olarak tespit edilebilme süreleri elde edilmiştir. OKB ve RAI skorlarının dizileme süreci boyunca değişimine ait grafikler şekil 3-4'te görülmektedir. Buna göre dizilenen her bir patojenin eldeki veri tabanındaki patojen modellerine OKB profili uzaklıklarının (veya RAI skorlarının) değişimi toplan okunan baz çifti cinsinden verilmiş ve her patojen için yaklaşık güvenli tespit eşik değeri belirlenmiştir. Bu değer seçilirken yanlış modellerin erişemediği ve doğru modelin erişimden sonra en az 1 saat dizileme süresi boyunca güvenli bölgede kaldığı sınır değerler kriter olarak belirlenmiştir. Daha sonra doğru modelin bu kritik değere erişim süresi patojen tanıma süresi olarak belirlenmiştir.



Şekil 3. OKB profillerinin dizileme süreci boyunca veri tabanındaki ortalama profillere yakınsama grafiği. Kırmızı: hedef organizma, kesikli mavi: belirlenen kritik değer.

OKB profilleri göz önüne alındığında *A.baumannii*'nin yaklaşık 1.2 Mbç okuma sonucunda (8.64 dakika), *E. coli*'nin 2.6 Mbç okuma sonucunda (17.72 dakika), *C. freundii*'nin 4.3 Mbç okuma sonucunda (32.9 dakika), *Enterobacter spp.*'in 1.9 Mbç okuma sonucunda (13.7 dakika), K. *pneumonia*'nın 5.1 Mbç okuma sonucunda (37.4 dakika), *P. aeruginosa*'nın 270 Kbç okuma sonucunda (1.9 dakika) güvenli olarak tespit edildiği görülmüştür.

Aynı simülasyon RAI tabanlı algoritma üzerinde yürütüldüğünde ise RAI skorlarının OKB profillerine oranla daha hızlı bir şekilde yakınsadığı ve test edilen patojeni güvenli olarak tespit ettiği görülmüştür (Şekil 4).



Şekil 4. RAI skorlarının dizileme süreci boyunca veri tabanındaki ortalama profillere yakınsama grafiği. Kırmızı: hedef organizma, kesikli mavi: belirlenen kritik değer.

RAI tabanlı hızlı patojen tanıma sonuçlarına göre *A.baumannii, Enterobacter spp.* 1 dakikanın altı sürelerde tespit edilebilirken *E. coli*'nin 1.2 Mbç okuma sonucunda (8.24 dakika), *C. freundii*'nin 320 Kbç okuma sonucunda (2.3 dakika), K. *pneumonia*'nın 410 Kbç okuma sonucunda (2.85 dakika), *P. aeruginosa*'nın 240 Kbç okuma sonucunda (1.72 dakika) güvenli olarak tespit edildiği görülmüştür.

4. Tartışma ve Sonuç

Bu çalışma ile henüz yeni bir teknoloji olarak kullanılmaya başlayan ve yeni nesil DNA dizileme teknolojilerinin 3. nesili olarak kabul edilen nanogözenek dizilemesinin bir hızlı patojen tanıma aracı olarak kullanılma potansiyelini ortaya çıkarma ve bu potansiyeli işler hale getirebilecek biyoinformatik yaklaşımlarının ve algoritmaların geliştirilmesi amaçlanmıştır. Enfeksiyon etkeni veya çevresel patojen olan mikrobiyal organizmaların moleküler tekniklerle tespiti geçtiğimiz onyıllarda klinik mikrobiyolojinin ilgi odağı haline gelen konulardan biri olmuştur [8-10]. Moleküler teknikleri cazip kılan özelliklerden biri birçok uygulamada kültürlemeyi ortadan kaldırabiliyor olmasıdır. Kültür aşaması azımsanamayacak bir sıklıkta hedeflenen bakterilerin ürememe gibi sorunlar doğurduğu gibi, başarılı olduğu durumlarda dahi uzun süren kuluçka süreçleri ile hızlı patojen tespitini imkansız kılmaktadır. Günümüzde kabul gören moleküler patojen tespiti tekniklerinin önemli kısmı hedeflenmiş bölgelerin polimeraz zincir reaksiyonları ile çoğaltılması ve/veya bu bölgelerin dizilenerek karakterize edilmesine dayanmaktadır. Bu yaklaşım çeşitli hassasiyet ve özgüllük problemleri taşıdığı gibi önceden rastlanmamış suşların primer dizileri tasarlanmamış olduğundan sözkonusu patojen tespitlerini sağlayamamaktadır. Öte yandan yeni nesil dizileme teknolojileri ulaşılabilir maliyet ile tüm genom dizilemelerini olanaklı kılmakta ve diğer moleküler tekniklerinin dezavantajlarından bu yönüyle sıyrılabilmektedir [11-13]. Ancak mevcut 2. nesil DNA dizileme teknolojileri, temelinde bulunan teknik gereği dizilemenin (kimyasal reaksiyonların) tamamı bitirilmeden okunan dizileri çıktı olarak verememektedir. Bu çevrimdışı prosedür çoğunlukla 2-3 gün arası bir süreç aldığından yine mevcut 2. nesil teknolojiler ile hızlı patojen tanıma teknikleri ve prosedürleri geliştirmek teknolojinin şu anki evresinde mümkün görünmemektedir. Bununla beraber şu an yeni yaygınlaşmaya başlayan 3. nesil nanogözenek

dizilemesi ve günümüzdeki tek teknolojik uygulaması olan Oxford Nanopore sistemleri gerçek zamanlı dizilemeyi ve DNA'nın okunduğu anda işlenebilmesini olanaklı kılmaktadır [35,36]. Bu özellik sayesinde gerçek zamanlı hızlı patojen tanıma tekniklerinin geliştirilmesi bu teknoloji ile olanaklı olmuştur.

Nanogözenek dizilemesi okuma ve hata yapma karakteri açısından mevcut 2. nesil DNA dizileme teknolojilerinden önemli ölçüde ayrışmaktadır. 2. nesil teknolojilere göre daha uzun DNA parçalarından oluşan okumalar üretilebilirken hata oranı mevcut teknolojilere göre oldukça yüksek olabilmektedir [8, 10]. Dolayısıyla 2. nesil dizileme teknolojileri için geliştirilen patojen tanıma amaçlı biyoinformatik teknikler nanogözenek dizilemesinde etkisiz kalabilmektedir. Bu sebeple uzun okuma dizilerinden faydalanabilen, aynı zamanda yüksek okuma hata oranlarına dayanıklı yeni nesil tekniklere ihtiyaç duyulmaktadır.

Bu çalışma kapsamında gerçekleştirilen deneylerde tamamı hizalanabilen okumaların nükleotid varyasyonlarına göre hata oranı ortalama olarak %8,71 +/- 2.35 olarak bulunmuştur. Bu oran şu an literatürde rapor edilen hata oranına benzer olmakla beraber şu an için ikinci nesil DNA dizileme teknolojileri için tasarlanmış olan taksonomik analiz ve tanıma programlarının ve benzeri biyoinformatik yöntemlerin patojen tanıma gibi görevler için yetersiz kalabileceğine işaret etmektedir. Çalışma kapsamında geliştirilen ve test edilen biyoinformatik yöntemler hatalı okuma oranının nispeten yüksek olduğu senaryolarda da başarılı tanıma gerçekleştirebilecek dayanıklı örüntü tanıma yaklaşımları olarak tasarlanmıştır.

Yapılan çalışmada uzun okuma dizilerinin avantajı yanında yüksek okuma hata oranlarına dayanıklılık potansiyeli taşıyan iki farklı teknik olarak OKB profilleri ve RAI endeksleri tabanlı algoritmalar geliştirilip hızlı patojen tanıma deneylerinde test edilmişlerdir. Deney amacıyla klinikte yüksek sıklıkla gözlenen 6 farklı patojen izolatı MinION sisteminde dizilenerek eldeki oluşturulan iki farklı program varyasyonunda bu patojenlerin tanınabilme performansları ölçülmüştür. OKB profilleri üzerinden geçekleştirilen standart metrik ölçümleri ile toplu dizileme seviyesinde başarılı patojen tanıma sağlanırken her bir dizilenen DNA parçasının başarılı olarak sınıflandırılması gereken senaryolarda yüksek bir tespit doğruluğu gözlenememiştir. Bunun başlıca sebebinin OKB profillerindeki varyasyon olduğu ve bu varyasyonun nanogözenek dizileme teknolojisine özgü okuma hatalarının yanısıra her bir türün içsel dinamikleri ve genom dizi yapısından kaynaklandığı öngörülmüştür. Dizileme deneyleri sonucunda elde edilen ortalama OKB profilleri gözönüne alındığında oluşan profillerin farklı bit düzeylerinde ortaya çıktığı ve genom içerisindeki varyasyonların genişliği hesapta bulundurularak başarılı tanıma yapabilme potansiyeli olduğu öne sürülmüştür. Dolayısıyla, elde edilen sonuçlar itibarı ile RAI endeksi tabanlı patojen tanıma sisteminin OKB tabanlı sisteme göre belirgin ölçüde daha başarılı olduğu çıkarımına varılmaktadır.OKB tabanlı sistem tek bir patojen organizmadan oluşan izolatlar üzerinde etkili olabilirken, tanıma yeteneği kültür bağımsız metagenom örnekleri içerisindeki patojenleri tanımak için yeterli olmamaktadır. Bununla beraber RAI endeksleri tabanlı programın bir organizma karışımı içerisindeki DNA dizilerinin hangi organizmalara ait olduğunu teker teker belirleme yeteneği, dolayısıyla kültür bağımsız örnekler üzerinde kullanılma olanağı daha fazladır.

Test edilen patojen örnekleri içerisinde *P. aeruginosa* ve *A. baumannii*'nin genellikle yüksek doğrulukla tespit edilebildiği görülmektedir. Bu durum geri kalan patojenlerin aynı aileden olması (*Enterobacteriaceae*) ve filogenetik benzerliklerinin genom üzerinden oluşturulan OKB ve RAI profillerinde de benzerliğe sebep olması dolayısıyla diğer türlerin birbirlerinden daha zor ayrıştırılması ile açıklanabilir.

OKB tabanlı yaklaşım her ne kadar profil benzerliği/uzaklığı metrikleri ile kullanıldığında yüksek başarım değerlerine ulaşamasa da OKB profillerinin içsel özelliklerini verilen örneklerden çıkarımsayarak otomatik modeller oluşturan makine öğrenme teknikleri kullanıldığında OKB tabanlı patojen tanıma algoritmalarının ortaya atılabildiği gözlenmiştir. Bu duruma karşılık aslında bu profillerin patojene özgü öznitelikleri yeterince temsil ettiği, ancak tanıma algoritmalarının yeterince karmaşık olmadığı savı ortaya atılabilir. RAI endeksine dayanan algoritma herhangi bir temsil profili oluşturmadığından ve doğrudan bir endeks skoru kullandığından bu algoritma ile makine öğrenmesine tabi tutulacak bir geliştirilme sağlanamamıştır. Ancak makine öğrenme tekniklerinin patojen tanıma doğruluğu açısından getirmiş olduğu dramatik artış, bu yaklaşımlara dayanan yeni nesil hızlı patojen tanıma teknikleri açısından büyük bir potansiyel olabileceğine işaret etmektedir.

Patojen tanıma sürelerinin ölçüldüğü simülasyonlarda test edilen 6 patojen türünün de bir saatten kısa sürede, hatta çoğunlukla dizileme prosedürü başladıktan sonra dakikalar içerisinde doğrulukla tespit edilebildiği görülmüştür. Bu gözlem sonucunda nanogözenek dizilemesinin kullanılarak hayati önem taşıyan kısa sürelerde klinik suşların tanınması, çevresel patojenlerin ve salgınların tespit edilmesinde etkili ve geçerli bir teknoloji olabileceği öne sürülebilir. Bununla birlikte kısa sürede tanımanın yapılması, MinION cihazının akış hücrelerinin tekrar kullanılabilir olduğu da göz önüne alındığında örnek başına maliyeti düşük olan fizibilitesi yüksek sistemlerin oluşturulabileceğini göstermektedir.

Çalışma neticesinde elde edilen sonuçlar hızlı patojen tanıma konusunda olumlu sinyaller vermekle birlikte, bu sonuçların deneyler kapsamında incelemeye tabi tutulan 6 patojen suş için geçerli olduğunu unutmamak

gerekmektedir. Daha genel yargıların ortaya atılabilmesi için daha geniş patojen panelleri kullanılarak geniş çaplı çalışmaların yürütülmesi gerekmektedir. Geniş patojen panellerinin göz önünde bulundurulmasıyla birlikte elde edilen doğruluk sonuçlarında azalma görülmesi olasılık dahilinde bulunmakla beraber bunun ne ölçüde olacağının kestirilmesi oldukça güçtür. Yine de bu olasılık, nanogözenek dizilemesinin hızlı ve yüksek doğrulukla patojen tanıma platformu olma potansiyelini gölgelememektedir.

Çalışma kapsamında önerilen OKB profili ve RAI tabanlı algoritmaların hızlı patojen tanıma konusunda yeterli doğruluk seviyesinde ve kısa sürede tanıma yapabilecek hızda olduğu nanogözenek dizilemesi yapılan patojen paneli üzerinde gösterilmiştir. Önerilen yaklaşımların mevcut programlarla rekabet edebilen performansta olduğu yapılan testlerle görülmektedir. Gelişmekte olan nanogözenek DNA dizileme teknolojileri sahaya taşınabilen küçük portatif cihazlar olarak üretilmektedir. Dolayısıyla sahada örnek toplayıp sekanslamasının yapılması mümkün kılınmıştır. Ancak biyoinformatik analiz genellikle yüksek performanslı bilgisayarlar ve genel ağ bağlantısı gerektiren sistemlerden oluşmaktadır. Önerilen biyoinformatik algoritmalar yüksek profilli sistem gereksinimi olmayan ve çevrimiçi veri tabanları ile bağlantı kurulasını gerektirmeyen hafif sistemler olarak tasarlanmıştır. Bu sayede mobil hesaplayıcılar (örn. Tablet bilgisayar, akıllı telefon, gömülü sistemler vb.) sistemler kullanılarak sahada patojen tanıma yapılması bu projede geliştirilen ve benzer doğrultuda tasarlanan sistemler için mümkün görünmektedir.

Hızlı patojen tanımanın kısıtlı sürede, en az donanımla ve laboratuvar dışında da yapılabilmesi için dizileme ve biyoinformatik aşamaları dışında örnek hazırlama sürecinin de hızlı ve minimum kaynak gereksinimli bir hale gelmesi gerekmektedir. Mevcut koşullarda örnek hazırlama süreçleri (örn. Kültür, DNA izolasyonu, sekanslama prosedürleri) laboratuvar dışında yapılamamaktadır. Ancak geliştirilmekte olan yeni nesil sistemlerle mobil veya hızlı örnek hazırlama cihazlarının yakın bir gelecekte ulaşılabilir olacağı düşünülmektedir. Bu doğrultuda söz konusu yeni nesil teknolojilerin kısa bir süre sonra rutin analizlere de yansıması beklenmektedir.

Mevcut algoritmalar yalnızca bir patojenin taksonomik olarak kısa bir sürede tanımlanması için tasarlanmıştır. Öte yandan bu görevi üstlenen nanogözenek dizilemesinin aslında bir tüm genom dizileme tekniği olması, taksonomik tanıma ötesinde patojen karakterizasyonunun tasarlanan başka biyoinformatik süreçlerle mümkün olabileceğine işaret etmektedir. Örneğin incelenen patojenin içerdiği antibiyotik direnç genlerinin tespit edilmesi ve bunun gerçek zamanlı bir şekilde yapılması kısa sürede tedavi tercihlerinin şekillendirilmesi gereken vakalarda oldukça yararlı olabilir. Bu amaca yönelik yeni nesil biyoinformatik algoritmalarının geliştirilmesinin çalışma kapsamında ele alınan alan için önemli bir genişleme olacağı düşünülmektedir.

Teşekkür

Sunulan makale TÜBİTAK ARDEB 1002 programı kapsamında 116S083 numaralı “Nanogözenek DNA Dizilemesi ile Hızlı Patojen Tanıma Yapabilen Algoritmaların Geliştirilmesi” başlıklı projeden üretilmiştir. Desteğinden dolayı Türkiye Bilimsel ve Teknolojik Araştırma Kurumu’na teşekkür ederiz.

Kaynakça

[1] Spížek J., Novotná J., Rezanka T., Demain A. L. 2010. Do we need new antibiotics? The search for new targets and new compounds. J. Ind. Microbiol. Biotechnol., 37, 1241–1248.

[2] Fauci A. S., Touchette N. A., Folkers G. K. 2005. Emerging infectious diseases: a 10-year perspective from the National institute of allergy and infectious diseases. Emerging Infect. Dis. 11, 519–525.

[3] Maurer, J. J. 2011. Rapid detection and limitations of molecular techniques. Annual Review of Food Science and Technology, 2, 259-279.

[4] Lazcka, O., Del Campo, F. J., Munoz, F. X. 2007. Pathogen detection: A perspective of traditional methods and biosensors. Biosensors and bioelectronics, 22(7), 1205-1217.

[5] Erlich Y. 2015. A vision for ubiquitous sequencing. Genome Res., 25, 1411-1416.

[6] Stoddart D., Heron A. J., Mikhailova E. vd. 2009. Single-Nucleotide Discrimination in Immobilized DNA Oligonucleotides with a Biological Nanopore. Proceedings of the National Academy of Sciences of the United States of America, 106, 7702–7707.

[7] Kasianowicz J. J, Brandin E., Branton D., Deamer D. W. 1996. Characterization of individual polynucleotide molecules using a membrane chanel. Proc Natl Acad Sci, 93(24), 13770–13773.

[8] Quick J., Quinlan A., Loman N. 2014. A reference bacterial genome dataset generated on the MinION(TM) portable single-molecule nanopore sequencer. GigaScience, 3:22.

[9] Loman N. J., Quick J., Simpson J. T. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Methods, 12, 733–735.

[10] Mikheyev A. S., Tin M. M. 2014. A first look at the Oxford Nanopore MinION sequencer. Mol. Ecol. Resour., 14, 1097–1102.

[11] Greninger A. L., vd. 2015. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Medicine, 7 (1), 99.

[12] Bull, R. A., Adikari, T. N., Ferguson, J. M., Hammond, J. M., Stevanovski, I., Beukers, A. G., Deveson, I. W. 2020. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. Nature communications, 11(1), 1-8.

[13] Li, J., Wang, H., Mao, L., Yu, H., Yu, X., Sun, Z., Wang, X. 2020. Rapid genomic characterization of SARS-CoV-2 viruses from clinical specimens using nanopore sequencing. Scientific Reports, 10(1), 1-10.

[14] Kielbasa S. M, Wan R., Sato K., Horton P., Frith M. C. 2011. Adaptive seeds tame genomic sequence comparison. Genome Res., 21(3), 487–93.

[15] Madoui M. A., Engelen S., Cruaud C., Belser C., Bertrand L., Alberti A., vd. 2015. Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genom., 16:327.

[16] Wang J., Moore N., Deng Y., Eccles D., Hall R. 2015. MinION nanopore sequencing of an influenza genome. Front Microbiol, 6: 766.

[17] Karlsson E., Lärkeryd A., Sjödin A., vd. 2015. Scaffolding of a bacterial genome using MinION nanopore sequencing. Sci Rep., 5: 11996.

[18] Judge K., Harris S. R., Reuter S., Parkhill J., Peacock S. J. 2015. Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. J. Antimicrob. Chemother., 70, 2775–2778.

[19] Risse J., Thomson M., Blakely G., Koutsovoulos G., Blaxter M., Watson M. 2015. A single chromosome assembly of Bacteroides fragilis strain BE1 from illumina and MinION nanopore sequencing data. BioRxiv, 024323.

[20] Jain M. vd., 2015. Improved data analysis for the MinION nanopore sequencer. Nat. Methods, 12, 351–356.

[21] Nalbantoglu O. U., Sayood K. 2011. Computational Genomics Signatures. Synthesis Lectures on Biomedical Engineering, 6(2), 1-129. Morgan-Claypool Publishers, 129 sayfa.

[22] Brady S. S. vd. 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated markov models. Nat Methods, 6, 673–676.

[23] Nalbantoglu O. U., 2011. Computational Genomic Signatures and Metagenomics. Ph.D. Dissertation, University of Nebraska-Lincoln, 201 s, Lincoln, Nebraska, ABD.

[24] Otu H., Sayood K. 2003. A divide and conquer approach to sequence assembly. Bioinformatics, 19, 22–29.

[25] Bauer M., Schuster S. M., Sayood K. 2008. The average mutual information profile as a genomic signature. BMC Bioinformatics, 9, p. 48.

[26] Nalbantoglu O. U., Way S., Hinrichs S., Sayood K. 2011."RAIphy: Phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. BMC Bioinformatics, 12(41).

[27] Bazinet A. L., Cummings M. P. 2012. A comparative evaluation of sequence classification programs. BMC Bioinformatics 13: 92.

[28] Rish I. 2001. An empirical study of the naive Bayes classifier", IJCAI Workshop on Empirical Methods in AI. 4-6 Ağustos 2001, Seattle, ABD. Vol. 3, No. 22, pp. 41-46.

[29] Vapnik, V. 1995. Support-vector networks. Machine Learning. 20 (3): 273–297

[30] Altman, N. S. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 46 (3): 175–185.

[31] Utgoff, P. E., 1989. Incremental induction of decision trees. Machine learning, 4(2), 161-186.

[32] Leo B., 2001. Random Forests. Machine Learning. 45 (1): 5–32.

[33] Li, H., 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics, 32(14), 2103-2110.

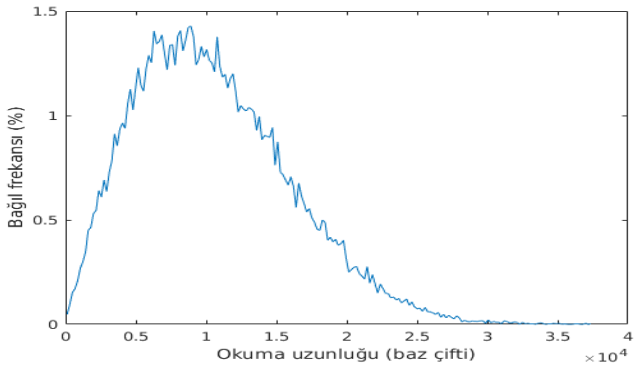
[34] Wood, D. E., Salzberg, S. L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology, 15(3), R46.

[35] Hayden E. C. 2015. Pint-sized DNA sequencer impresses first users. Nature 521, 15–16.

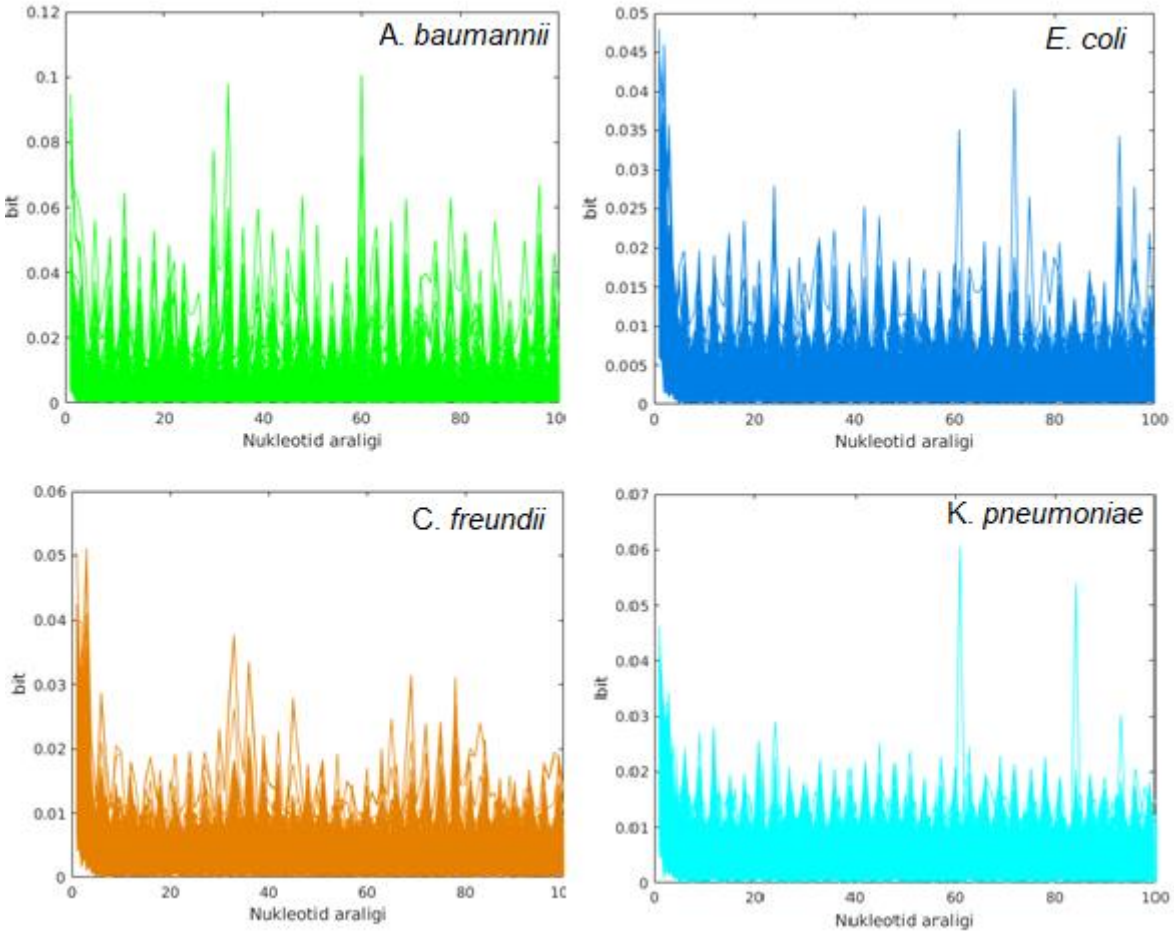
[36] Feng Y., Zhang Y., Ying C., Wang D., Du C. 2015. Nanopore-based fourth-generation DNA sequencing technology. Genomics Proteomics Bioinformatics, 13, 4–16.

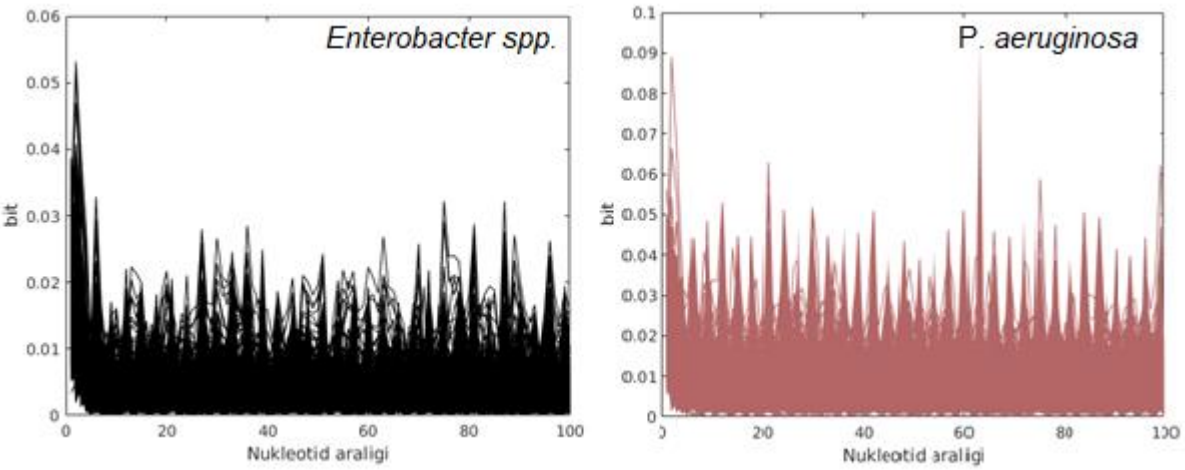
Ekler

Ek A. Yardımcı sonuçlar
OKB profillerinin sınıflandırılmasına ait karışıklık matrisleri. Nanogözenek dizilemesi sonucunda elde edilen okuma uzunluğu histogramı. Dizileme sonucu ortaya çıkan okumalara ait OKB profilleri

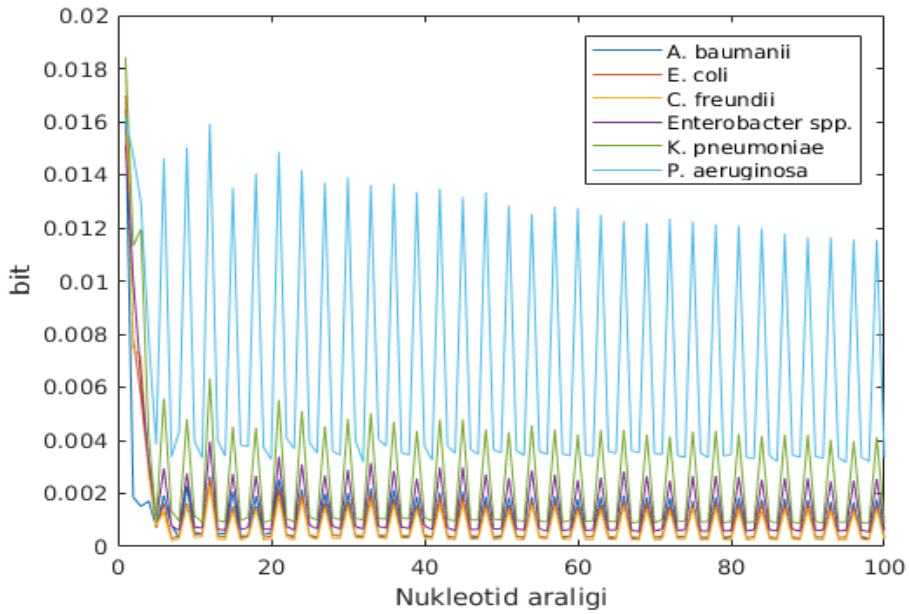


Şekil Ek 1. Nanogözenek dizilemesi sonucunda elde edilen okuma uzunluğu histogramı. Deneyler sonucunda elde edilen okuma uzunlukları histogram olarak sunulmaktadır. Buna göre okuma uzunluğu dağılımın ortalanın sağına eğimli olduğu görülmektedir.





Şekil Ek 2. Dizileme sonucu ortaya çıkan okumalara ait OKB profilleri. Tanıma için belirlenmiş altı patojene ait nanogözenek dizilemesi sonucunda elde edilen DNA okumaları ayrı ayrı OKB profillemesine tabii tutulmuş ve bu profillerin patojen tanıma algoritmalarındaki performansları test edilmiştir. Şekillerde her patojene ait elde edilen her bir DNA okuması için 1bç'den 100 bç uzaklığa kadar olan baz çiftlerinin oluşturduğu OKB profilleri görülmektedir. Profillerde OKB karakteristiği genel olarak korunsa da DNA dizileme hata oranının yüksek bir varyasyon oluşturduğu ve profilleri yerel ölçekte bozunuma uğrattığı görülmektedir.



Şekil Ek 3. Dizilenen DNA okumalaranın ortalama OKB profilleri.

Tablo Ek 1. OKB profillerinin Öklid metriği kullanılarak sınıflandırılmasına ait karışıklık matrisleri.

Öklid Metriği (1-100 bç)						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	55.6	6.6	7.9	5.1	0.6	24.2
E. coli	5	25.7	36.3	21.3	11.3	0.4
C. *freundii*	1.6	19.2	62.3	6.4	5.3	5.2
Enterobacter spp.	9.6	15.4	3.5	52.4	14.4	4.7
K. *pneumoniae*	5.4	10.7	2.1	25.4	48	8.4
P. *aeruginosa*	0.5	0	0	0.6	0	98.9
1-50 bç						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	57.9	6	9	3.5	0.3	23.3
E. coli	5.5	23.6	38.5	17.4	15	0
C. *freundii*	1.3	17.4	63.4	8	5.6	4.3
Enterobacter spp.	11.4	16.2	3.1	50	16.8	2.5
K. *pneumoniae*	5.7	10.7	2.1	21.7	50.9	8.9
P. *aeruginosa*	1.1	0	0	0.3	0	98.6
1-10 bç						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	41.1	4.1	18.5	10.8	9.1	16.4
E. coli	5.9	18.1	44.6	8.1	21.8	1.5
C. *freundii*	1.8	14.4	64	9.5	6.7	3.6
Enterobacter spp.	12.7	16.3	7	44.1	16.7	3.2
K. *pneumoniae*	5.9	12.9	2.8	18.9	50.4	9.1
P. *aeruginosa*	2.1	0	0	0.1	0.3	97.5

Tablo Ek 2. OKB profillerinin Öklid metriği kullanılarak sınıflandırılmasına ait karışıklık matrisleri (ardışık baz çiftlerinin birlikteliği olmadan).

Öklid Metriği (2-100 bç)						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	57.2	6.1	0	4.8	1.5	30.4
E. coli	5.2	40.5	26.8	9.2	12.7	5.6
C. *freundii*	2.1	19.7	55.6	6.1	10.1	6.4
Enterobacter spp.	6.4	31.3	8.6	28.6	15.4	9.7
K. *pneumoniae*	4.6	13.6	11.5	12.4	48.5	9.4
P. *aeruginosa*	22	0.8	0	2.4	0.1	74.7
2-50 bç						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	64	2.4	0	5.1	1.1	27.4
E. coli	6.6	36.4	26.8	13	12.2	5
C. *freundii*	3.6	18	54	9.2	9.8	5.4
Enterobacter spp.	7.9	29	9.4	27.4	17.5	8.8
K. *pneumoniae*	4.8	9.1	11.6	14.9	51.6	8
P. *aeruginosa*	21.4	0.4	0	2.3	0	75.9
2-10 bç						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	53.4	2.8	0.5	10.2	6.2	26.9
E. coli	10.6	16.1	27.7	18.3	24.3	3
C. *freundii*	4.1	15.3	47	11.6	17.2	4.8
Enterobacter spp.	9.8	22.3	13.3	30.3	17.1	7.2
K. *pneumoniae*	7.5	5.8	12.2	14.8	53.8	5.9
P. *aeruginosa*	17.7	0.1	0	3.3	0.8	78.1

Tablo Ek 3. OKB profillerinin L1 metriği kullanılarak sınıflandırılmasına ait karışıklık matrisleri.

L1 Metriği (1-100 bç)						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	52.2	6.7	7.9	10.3	1.7	21.2
E. coli	4.4	34.3	35.6	12.2	10.4	3.1
C. *freundii*	1.7	17.1	63.3	4.6	5.4	7.9
Enterobacter spp.	5.7	15.6	4.1	50.4	16.5	7.7
K. *pneumoniae*	4.6	10	2.3	25.1	46.9	11.1
P. *aeruginosa*	8.7	0	0	5.6	0.3	85.4
1-50 bç						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	61.4	6.6	9.1	4.2	0.5	18.2
E. coli	5.8	23	37.8	21.4	11.1	0.9
C. *freundii*	1.5	15.7	64.6	5.9	6.9	5.4
Enterobacter spp.	7.8	15.6	3.8	49.4	18.2	5.2
K. *pneumoniae*	5.1	10.8	2	23.4	51.2	7.5
P. *aeruginosa*	2.7	0	0	3.3	0.1	93.9
1-10 bç						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	44.3	4.5	15.8	10.6	8.8	16
E. coli	6.8	17.5	45.3	9.5	20	0.9
C. *freundii*	2.4	14	64.7	8.4	7.3	3.2
Enterobacter spp.	12	17.4	7.2	43	18	2.4
K. *pneumoniae*	6	12.7	3.2	18.9	54	5.7
P. *aeruginosa*	1.8	0	0	1	0.4	96.8

Tablo Ek 4. OKB profillerinin L1 metriği kullanılarak sınıflandırılmasına ait karışıklık matrisleri (ardışık baz çiftlerinin birlikteliği olmadan).

L1 Metriği (2-100 bç)						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	45.7	15	1.3	5	2	31
E. coli	3.1	39.8	27.9	9.3	13	6.9
C. *freundii*	1.9	18.5	53.9	5.6	9.4	10.7
Enterobacter spp.	4.1	25.2	12.9	30.4	14.3	13.1
K. *pneumoniae*	3.9	14.6	12.9	9.5	41.5	17.6
P. *aeruginosa*	18.2	1.5	0.2	4.2	1.4	74.5
2-50 bç						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	58.3	8.1	0.5	3	1	29.1
E. coli	5	38.2	25.5	11	14.4	5.9
C. *freundii*	2.8	18.5	51.6	8.4	10.9	7.8
Enterobacter spp.	5.4	28.5	10.9	27.4	16.5	11
K. *pneumoniae*	2.8	13	12.4	12.5	45.4	13.9
P. *aeruginosa*	19.6	0.8	0	3	0.2	76.4
2-10 bç						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	53.5	3.2	1	10.5	3.9	27.9
E. coli	10.9	16.9	28.7	16.9	22.9	3.7
C. *freundii*	5.2	11	50.8	12.6	15.4	5
Enterobacter spp.	10.6	22.8	16	27.1	15.4	8.1
K. *pneumoniae*	7	6.1	13.2	15.1	51.6	7
P. *aeruginosa*	19.5	0.6	0	3	1	75.9

Tablo Ek 5. OKB profillerinin Pearson korelasyonu kullanılarak sınıflandırılmasına ait karışıklık matrisleri.

Pearson korelasyonu (1-100 bç)						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	31.2	2.6	40.7	0.8	2.8	21.9
E. coli	0.2	7.4	79	0.5	12.8	0.1
C. *freundii*	0.1	0.6	89.6	0.2	4.9	4.6
Enterobacter spp.	2.2	5.6	53.4	15.1	21.3	2.4
K. *pneumoniae*	0.4	2.3	54.8	0.9	37.1	4.5
P. *aeruginosa*	0.6	0	0	0.4	0.9	98.1
1-50 bç						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	22.1	2.5	51.5	1.1	6.2	16.6
E. coli	0.1	6.3	82.5	0.5	10.6	0
C. *freundii*	0.3	0.7	90.8	0.8	4.6	2.8
Enterobacter spp.	2.7	5.2	59.7	16	15.6	0.8
K. *pneumoniae*	0.2	1.7	64.6	0.9	31	1.6
P. *aeruginosa*	2.1	0	0.1	0.9	2.4	94.5
1-10 bç						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	4.9	1.6	76.8	2	5.4	9.3
E. coli	0	0.4	80.6	0.8	18.2	0
C. *freundii*	0.1	0.7	90.6	4	3.4	1.2
Enterobacter spp.	0.4	2.7	65.8	21	9.9	0.2
K. *pneumoniae*	0.4	1.8	70.7	3.9	23	0.2
P. *aeruginosa*	4.9	0	1.7	5.2	12.2	76

Tablo Ek 6. OKB profillerinin Pearson korelasyonu kullanılarak sınıflandırılmasına ait karışıklık matrisleri (ardışık baz çiftlerinin birlikteliği olmadan).

Pearson korelasyonu (2-100 bç)						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	13.8	3.4	0.4	6.6	4.8	71
E. coli	1	16.4	44.6	4	24.5	9.5
C. *freundii*	0.3	1.9	69.3	1.4	17.7	9.4
Enterobacter spp.	0.5	6.8	29.1	9.8	35.7	18.1
K. *pneumoniae*	0.3	0.7	15.4	1.1	64.3	18.2
P. *aeruginosa*	0.5	0	0.2	1.2	1.8	96.3
2-50 bç						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	16	2.5	2.1	6.2	9.9	63.3
E. coli	1.4	8.4	60.5	4.5	19.3	5.9
C. *freundii*	0.3	1.7	76.7	2.4	13.1	5.8
Enterobacter spp.	0.5	4	47.2	6.2	30.8	11.3
K. *pneumoniae*	0.2	1.3	25.9	1.3	62.6	8.7
P. *aeruginosa*	0.7	0	1.2	1.3	1.7	95.1
2-10 bç						
	A. *baumannii*	***E. coli***	**C. *freundii***	***Enterobacter spp.***	**K. *pneumoniae***	**P. *aeruginosa***
A. *baumannii*	13.8	9.3	9.9	15.6	23.5	27.9
E. coli	1.5	6.4	64.7	8.3	16.3	2.8
C. *freundii*	0.1	5.4	79.1	2.5	9.6	3.3
Enterobacter spp.	0.5	6.3	64.8	8.4	15.9	4.1
K. *pneumoniae*	0.1	2.4	47.5	5.2	41.8	3
P. *aeruginosa*	1.4	2.2	1.6	4.7	11.6	78.5