ARAŞTIRMA MAKALESİ / Research Article

Mikrobiyotada 16S rRNA ve Basit Biyoinformatik Analizler

16S rRNA In Microbiota and Basic Bioinformatics Analysis

Muhammed Kamil Turan[1], Özlem Cesur Günay[1], Seyit Ali Kayış[3], Mustafa Çörtük[3]

[1] Karabük Üniversitesi, Tıp Fakültesi, Tıbbi Biyoloji AD
[2] Karabük Üniversitesi, Tıp Fakültesi, Tıp Bilişimi AD
[3] Karabük Üniversitesi, Tıp Fakültesi, Göğüs Hastalıkları AD

Yazışma Adresi / Correspondence:
Muhammed Kamil Turan
Karabük Üniversitesi, Tıp Fakültesi, Tıbbi Biyoloji AD. Karabük, Türkiye
T: **+90 505 299 01 26** E-mail: **kamilturan@karabuk.edu.tr**

Orcid:
Orcid: https://orcid.org/0000-0002-1086-9514

Geliş Tarihi / Received : **27.01.2018** Kabul Tarihi / Accepted : **02.02.2018**

Turan MK, Günay ÖC, Kayış SA, Çörtük M.
Mikrobiyotada 16S rRNA ve Basit Biyoinformatik Analizler
J Biotechnol and Strategic Health Res. 2018;2(1):23-34.

Özet

İnsan genomunun keşfinden sonra yeni bir çağa giren biyoloji bilimi, barsak florasındaki bakterileri çok önceden bildiği halde floranın önemini henüz anlamlandırmaya başlamıştır. Barsak florasının genom yükü, insan genomundan kat ve kat fazladır. Bu flora günümüzde mikrobiyota olarak tanımlanır. Geleneksel kültür metotları ile analizinin neredeyse imkansız olduğu mikrobiyota, 16S rRNA genlerinin yüksek oranda değişken bölgelerinin dizilenmesi ve dizileme sonuçlarının biyoinformatik analizi ile rahat bir şekilde çalışılabilir hal almıştır. Ancak hala dizi boyları değişkenliği, yüksek oranda değişken bölgelerin başlangıç ve bitiş sahaları, dizileme sonucu elde edilen markerlar ile mikroorganizmaların eşleştirilmesi gibi çözüm bekleyen önemli bazı biyoinformatik sorunlar mevcuttur. Bu çalışmada, belirtilen sorunlara çözüm sunabilme amacıyla açık kaynak kodlu bir python çatısı geliştirilmiş ve bu çatının çıktılarının yorumlanması üzerinde durulmuştur. Geliştirilen yazılımı için gerekli olan 16S rRNA gen dizilerine Greengenes veritabanı kullanılarak erişim sağlanmıştır. Cins bazında tiplendirmenin oldukça zor olduğu Clostridium'lar hedeflenmiş ve geliştirilen python çatısı ile basit biyoinformatik analizler yapılmış, korunmuş diziler ortaya konarak olası primer aday bölgeleri saptanmaya çalışılmıştır. İlave olarak, araştırmacılara hedefledikleri mikroorganizmalar için genel primerler yerine çalışmaya özel primerleri geliştirebilmek için bir araç sunularak bu konuda geliştirilecek biyoinformatik araçlara öncülük etmek amaçlanmıştır. Genel olarak, bu çalışmanın mikrobiyota konusunda çalışmaya yeni başlayan araştırma gruplarına klavuzluk yapacak nitelikte olabilmesi amaçlanmıştır.

Anahtar Kelimeler: Mikrobiyota, biyoinformatik, RNA, 16S rRNA, çoklu dizi hizalama, korunmuş dizi analizi

Abstract

Although the presence of intestinal flora is previously well-known, the biology which is entering a new era after the discovery of the human genome has recently started to appreciate the importance of bacterial flora of the intestines. The genome load of the gut flora is excessive when compared to the human genome. This flora is now recognized as microbiota. Microbiota, which is almost impossible to analyse by conventional culture methods, has been studied with the sequencing of highly variable regions of 16S rRNA genes and bioinformatics analysis of resulted sequencing data. However, sequence length variability, starting and ending points of high-order variable regions, mapping of microorganisms with markers obtained from sequencing results are still problematic. In this study, an open source program coded in python framework was developed and the output of this framework was interpreted to provide a degree of understanding to such important bioinformatics problems. The 16S rRNA gene sequences required for the software developed were accessed using the Greengenes database. As an example, Clostridiums, which are difficult to type on Genus level, were targeted and simple bioinformatics analysis were performed with the developed python framework to reveal potential primer binding sites based on sequence conservation. Preserved nucleotide sequences were identified, and possible candidate regions of primers were proposed. As a result, researchers in the field are supplied with a tool to develop their own specific primers instead of utilising general primers for targeted microorganisms. Finally, it is aimed that this study will guide the research groups that start to work on microbiota.

Key Words: Microbiota, bioinformatics, RNA, 16S rRNA, multible sequence alignment, consensus sequence analysis

Gram boyama ve kültür yöntemleri enfeksiyon etkenlerinin tanısında kullanılan temel stratejilerdir. Fakat; klinik vakalarda kültür yöntemleri ile teşhis edilemeyen patojen mikroorganizmaların sayıca artmakta olması, geleneksel yöntemlerle teşhis edilseler dahi, önerilen antibiyotik tedavilerinin bazı durumlarda etkili olamaması, kommensallerin de var olduğu karma enfeksiyonların varlığı ve klinik örneklerde yalnızca kommensallerin teşhis edildiği durumlar enfeksiyonun gerçek etkenlerinin tespitinde moleküler yaklaşımların önemini ortaya çıkarmıştır [1]. Moleküler metodlardan Polimeraz Zincir Reaksiyonu (PZR-PCR) ve DNA dizileme teknolojileri, klasik kültür metodları ile saptanamayacak derecede mikroorganizma çeşitliliğine sahip olan barsak florasında, mikroorganizma çeşitliliğinin tespitine olanak verir. Ribozomal RNA (rRNA), GroEL chaperonin, RNA polymerase beta subunit (rpoB), DNA gyrase beta subunit (gyrB) vb. genler türler arasında yüksek oranda korunduğu için filogenetik analizlerde sıklıkla kullanılan stabil marker genlerdir[2]. Bu bağlamda, bakteride bulunan ve 'housekeeping' (ev işçisi) olarak adlandırılan 16s rRNA geni birçok moleküler uygulamada popüler bir hedef olarak seçilmiştir. Bu gen dizilerinin ulaşılabilir olması ve mevcut veri tabanlarında GenBank (www.ncbi.nlm.nih.gov), European Molecular Biology Laboratory (www.embl.org) vb. nükleotit dizilerinin varlığı moleküler yaklaşımlarda tercih edilme sebeplerindendir.

Prokaryotik ribozomlar, 30S ve 50S' lik altbirimlere sahip olup, 16S rRNA geni 30S' lik ribozom alt biriminde yer alır[1]. Bu genler, mRNA'dan protein translasyonunda görev aldığı için organizmalar arasında son derece korunmuştur. Filogenetik analiz için rRNA dizilerinin ilk kullanımında canlılar Eubacteria, Archaebacteria, Eukaryotes şeklinde üç domain'e ayrılmıştır[3]. Ayrıca ribozom küçük alt birimine spesifik rRNA primerleri kullanılarak kısmi veya tam rRNA dizileri oluşturulmuş ve türler arasındaki akrabalık ilişkileri matriks metodu ile çalışılmıştır[4]. Oluşturulan bu ilk filogenetik ağaçlarda Escherichia coli (eubacterium), Halobacterium volcanii (archeabacterium), Dictyostelium discoideum (eukaryote), Saccharomyces cerevisiae (eukaryote), Zea mays (eukaryote), Mus mus (eukaryote), Xenopus laevis (eukaryote) türlerinin akrabalık ilişkileri ortaya konmaya çalışılmıştır [4]. 16S rRNA geni, oldukça kısa (~1500 bç) nükleotid uzunluğuna sahip olup, bakteri türleri arasında korunmuş olan bölgelere ilave olarak 9 adet yüksek oranda değişkenlik gösteren (V1-V9) bölgelerini de içerir [5]. 16S rRNA genlerinin yüksek oranda değişkenlik gösteren bölgeleri farklı derecelerde dizileme çeşitliliği gösterirken sadece bir bölgenin analizi bütün bakterileri ayırt edebilmek için yeterli değildir. Genellikle V2, V3 ve V6 bölgeleri cins düzeyinde ayrım için yeterli olmakla birlikte, bazı durumlarda tür düzeyinde ayrımda da kullanılmaktadır [6]. Bu sahalar yüksek oranda değişkenlik gösteren bölgeler oldukları için biyolojik çeşitliliğin sınıflandırılmasında ciddi bir potansiyeli barındırmaktadır. Diğer bölgeler yüksek derecede dizi korunumundan dolayı pek tercih edilmezler. Jeodezik uzaklık verilerine göre farklı alt-bölgeler arasındaki yakınlık ilişkisi incelendiğinde, bakteri filogenetik analizi için V4-V6 bölgeleri tüm 16S rRNA gen dizisini en iyi temsil ettiğinden en güvenilir bölgeler olurken; V2 ve V8'in en az güvenilir bölgeler oldukları bildirilmiştir [5]. Yüksek oranda değişkenlik gösteren bölgelere ilave olarak, 16S rRNA genlerinin korunmuş bölgeleride içermesi bu bölgelere de bağlanan üniversal primerlerin yazılmasına olanak sağlamaktadır. Bu özellik 16S rRNA genlerini mikrobiyota çalışmalarında oldukça popüler hale getirmiştir. Bu nedenle veritabanlarında kayıtlı 16S rRNA gen dizileri sürekli artmakta; bu da bakteri teşhisinde ve sınıflandırılmasında gen dizilerine ulaşılabilirliği artırmaktadır. Yukarıda bahsi geçen yüksek derecede değişkenlik gösteren 'V' bölgeleri için başlangıç ve bitiş pozsiyonları aşağıda Tablo-1 'de gösterilmiştir [6].

Değiken bölge numarası	Başlangıç ve bitiş numaraları
V1 bölgesi	66-99
V2 bölgesi	137-142
V3 bölgesi	433-497
V4 bölgesi	576-872
V5 bölgesi	822-879
V6 bölgesi	986-1043
V7 bölgesi	1117-1173
V8 Bölgesi	1243-1294
V9 Bölgesi	1435-1465

Tablo 1: Yüksek oranda değişkenlik gösteren bölgeler ve bu bölgelerin gen içindeki başlangıç ve bitiş pozsiyonları

Dizileme sonuçlarından elde edilen veriler ile mikroorganizmaları eşleştirmek yüksek derecede değişkenlik gösteren bölgelerin her mikroorganizmada aynı olmaması, buna bağlı olarak korunmuş sahalarında pozisyonlarının değişmesi gibi nedenlerle başlı başına bir problemdir. Bu makalede mikrobiyota çalışmaları sonucunda elde edilen 16S rRNA gen dizilerinin:

1. Basit biyoinformatik analizleri,
2. Korunmuş ve yüksek oranda değişkenlik gösteren bölgelerin nükleik asit kompozisyonlarının tespiti,
3. Tür ve cins bazında çalışmalara has özel primerler geliştirmede aday bölgeleri hesaplamak için bir çatı geliştirilmiş ve bu çatının sonuçları yorumlanmaya çalışılmıştır.

Biyoinformatik, biyolojik veri ve mikrobiyota

Yeni gelişen bir bilim dalı olarak biyoinformatik multidisipliner yapısı ile dikkat çekmekte ve ilgi odağı haline gelmektedir. Bunun altında yatan temel nedenlerden bir tanesi son yıllarda biyolojik bilginin hemen her bilim dalından beslenerek artmasıdır. Biyolojik bilgi uzayı insan genom projesinin başlaması ile gündeme gelmiş ve projenin tamamlanması ile de önemli olduğunu kanıtlamıştır. Biyolojik bilginin son yirmi yıldaki yolculuğuna bakacak olursak bilgi uzayının önemi daha anlaşılır bir hal alır. Basit birkaç soru bu önemi pekiştirecektir. Aradığım bilgi parçası bu devasa uzayın neresindedir? Ona nasıl ulaşabiliriz? Nasıl güncel bir şekilde tutabiliriz? Bu soruların en önemlisi ise ham verileri bilgiye nasıl çevirebilir ve bu sentez bilgiyi başka proje ya da sorular için nasıl ham veri halinde sunabiliriz? İşte bu sorulara yanıt veren bir bilim dalı olarak biyoinformatik, biyolojik bilimlerin önündeki hesaplama problemlerine yanıt ararken biyolojik uzayı konu alır. Laboratuvarda üretilen verilerden bilgi elde etmek amacı ile gerekli olacak algoritma ya da teknolojileri geliştirir. Biyolojik veri uzayı günümüzde o kadar büyümüştür ki artık yukarıdaki soruların cevabını verecek tek bir disiplinden bahsedilemez hal almıştır. Belki önümüzdeki yüzyılın en büyük projesi biyolojik veri uzayı için yapılacak bir genom çalışması olacaktır.

Genetik bilgi uzayı araştırmalar sonucunda elde edilmiş verilerin veri tabanları şeklinde organize edilip araştırmacıların kullanımına sunulması neticesinde oluşmuş ve büyümüştür. Bu uzay 1993 yılında 24 veri tabanı ile sınırlı iken 2000 yılında veri tabanı büyüklüğü 1230 olmuştur. 2000 yılından günümüze kadar ise biyolojik veri uzayı katlanarak büyümesine devam etmiştir[7,8]. En sık gezilen

ait dizi sayısı ise milyarlarla ifade edilmektedir [9]. Veri uzayının büyüme ivmesi çok yüksektir. Günümüz hesaplama gücü ile elde edilen ham verilerin tamamının bilgi haline dönüştürülmesi neredeyse imkansız bir hal almıştır. Bu nedenle her biten proje pek çok yeni projenin başlamasına vesile olmaktadır. Son zamanlarda oldukça popüler hal alan mikrobiyota çalışmalarıda veri uzayında yerini almıştır. Bu konuda Silva ve Greengenes veritabanları öne çıkmaktadır [10]. Silva veritabanına https://www.arb-silva.de adresinden, Greengenes veritabanına ise http://greengenes.lbl.gov/Download/ adresinden ulaşılabilir.

Materyal

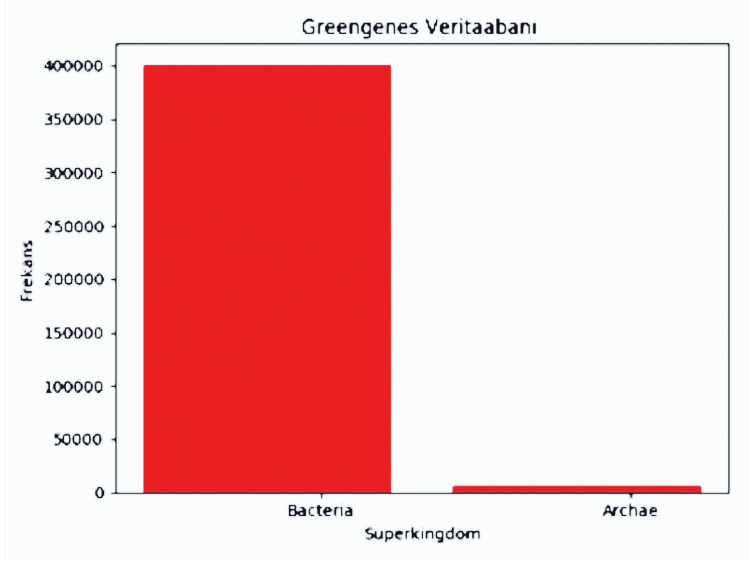
Analiz için Greengenes veri tabanından yararlanılmıştır. Greengenes veritabanı programatik erişime izin vermesi, sorgu sonuçlarını basit metin dosyaları şeklinde kullanıcısına sunması, istendiği durumlarda diğer veri değişim formatlarını desteklemesi gibi öne çıkan bazı özellikleri nedeni ile seçilmiştir. Çatı geliştirilme ortamı Python sürüm 3.5.2 64 bit olarak belirlenmiştir. Python tercih edilme nedeni söz diziminin rahat bir kullanıma sahip olması, hızlı şekilde prototip yazılımların üretilebilmesi, bilimsel içerik desteği ile özgür yazılım kategorisinde bulunması sayılabilir. Greengenes veritabanından elde edilen çıkartım bilgilerinin yerel veritabanında tutulması ve gerektiğinde sorgulanması için SqliteDatabase3 sürüm 0.2.0 kullanılmıştır. SqliteDatabase3 modülü, ikli temel özelliği nedeni ile seçilmiştir. Bunlardan ilki kaydedilen verileri yerel disk üzerinde tek bir dosya içinde tutması, ikincisi ise verilerin bir yerden bir başka yere çok rahat olarak göç ettirilebilmesidir.

Metod

16S rRNA gen dizileri elde edildikten sonra bu diziler üzerinde superstring çalışması yapılabilmesi için olabilecek en uygun biçimde hizalanmaları gerekmektedir. Bu hizalama seçilen ya da ilgi duyulan bir kısım organizma üzerinde olabileceği gibi tüm organizmalar üzerinden de yapılabilir. Bu noktada araştırmacıları kısıtlayan durum çoklu dizi hizalama işlemeninin uzun zaman alması ve çok hızlı bilgi işlem kaynağı gerektirmesidir. Bu çalışmada klasik global dizi hizalama algoritması olan Needleman-Wunch algoritması yerine genetik algoritma kullanan çoklu dizi hizalama metodu kullanılmıştır [11]. Bu sayede en iyilenmiş bir dizi hizalama sonucu yerine en iyilenmiş birden çok hizalama sonucu elde edilmiştir [12,13]. Çoklu dizi hizalama sonuçları üzerinde superstring çalışması yapabilmek için Weblogo3 kullanılmıştır. Weblogo3, superstring dizgelerinin görselleştirilmesi için sıklıkla kullanılan bir sunum tarzıdır (https://github.com/WebLogo/weblogo; 2017).

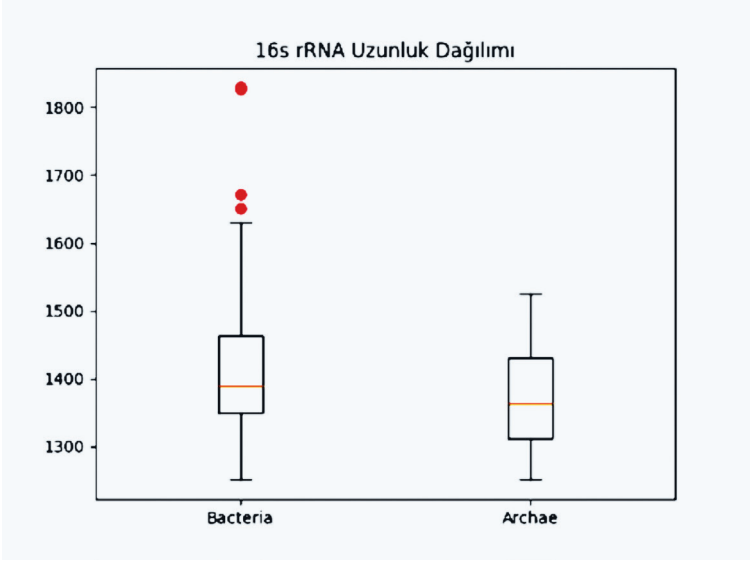
Sonuç

Greengenes veritabanı üzerinden programatik sorgu (superkingdom, phylum, class, order, family ve genus seviyesinde) yapılarak sonuç dosyaları ilişkisel yerel veritabanı içine kaydedilmiştir. İncelenen veritabanı toplam iki superkingdoma bölünmüş durumdadır. İlki Archea, ikincisi ise Bacteria olmak üzere toplam 406996 giriş bulunmaktadır. Archaea superkingdomu için 6738 kayıt ve Bacteria superkingdomu için 400258 kayıt listelenmiştir (Şekil 1). Ayrıca phylum seviyesinde 80, class seviyesinde 134, order seviyesinde 214, family seviyesinde 384, genus seviyesinde 1271, species seviyesinde ise 2374 farklı kayıt bulunmaktadır.



Şekil 1: Greengenes veri tabanı superkingdom seviyesinde girişlerin dağılımı

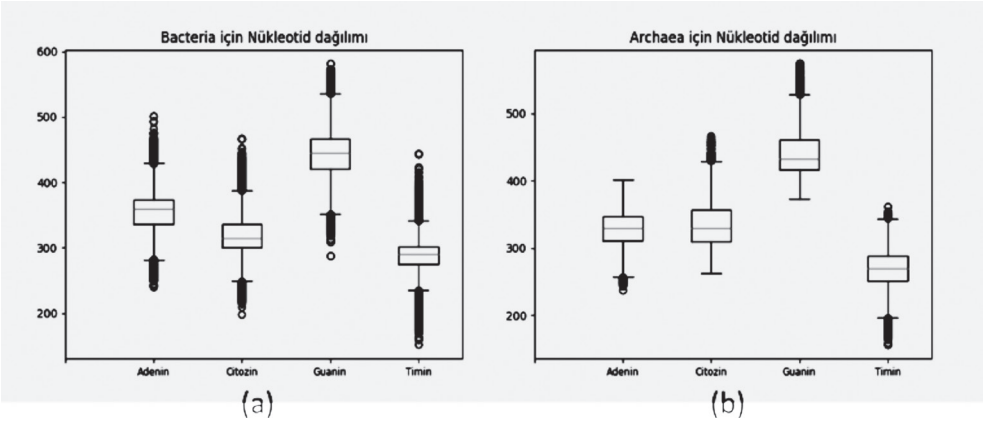
Ribozomal RNA genlerinin ortalama uzunlukları 1500 nükleotid kadardır. Fakat, biyoinformatik ve programatik açıdan bakıldığında türler arası gen uzunluklarının oldukça farklı oldukları görülmüş ve bu farklılık hesaplamalarda karşımıza ciddi bir problem olarak çıkmıştır. Aradığımız mikroorganizma grubunda, amplifiye edilmek istenen yüksek derecede değişken sahanın tam olarak hangi nükleotitde başlayıp hangi nükleotitde bittiğini bulabilmek için net bir şekilde genlerin uzunluklarına ihtiyaç vardır. Bacteria ve Archae superkingdomu için 16S rRNA genlerinin uzunluk dağılımları analiz edildiğinde uzunluklar bir birinden oldukça farklı olduğu görülmüştür (Şekil 2).



Şekil 2: 16S rRNA genlerinin uzunluk dağılım karakteristiği

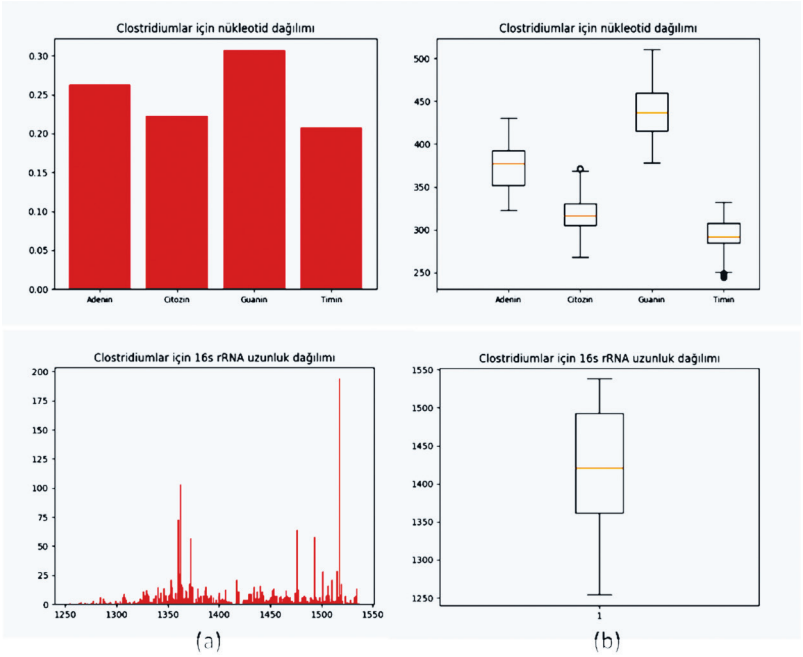
Dağılım tablosu incelendiğinde Bacteria ve Archaea superkingdomlarının 16S rRNA gen uzunluğu ortalama ve standart sapmalarının (ortalama±standart sapma) sırasıyla 1405.19±65.6 ve 1369.9±66.9 nükleotid olduğu görülmektedir (Şekil 2). Problemin doğasını daha rahat anlaya-

bilmek için Bacteria ve Archaea superkingdomlarında nükleotid dağılımları sırasıyla. Şekil 3 (a) ve (b)'de gösterilmiştir.



Şekil 3: a) Bacteria ve b) Archaea superkingdomlarında nükleotid dağılımları

Geliştirilen çatı ile analize konu olarak Clostridium'lar seçilmiştir. Clostridium'lar için toplamda 47 farklı tür listelenmiştir. Clostridium'ların universal primerler ile tür ve cins düzeyinde tiplendirildiklerinde düşük başarı elde edildiğinden bu çalışmada hedef organizma olarak seçilmiştir[14]. Diğer mikroorganizmalara oranla daha az tiplendirme hassasiyeti olan bu tür için veri tabanındaki kayıt sayısı ise 1886 olarak listelenmiştir. 1886 kayıtın tümü hedef alınarak yapılan analizde uzunluk dağılımı Şekil 4'te gösterilmiştir. Şekil 5'te ise belirtilen tür için nükleotid dağılımı verilmiştir. Genel olarak bakıldığında superkingdom seviyesindeki sonuçlar ile uyumlu olduğu dikkati çekmiştir. Bu tespit, 16S rRNA genlerini mikrobiyota içindeki biyolojik çeşitliliği modelleme de iyi bir belirteç olarak kullanılmasına ciddi bir delil teşkil etmektedir. Benzer sonuçlar diğer türlerde de elde edilmiş fakat burada ayrıca listelenmemişlerdir.

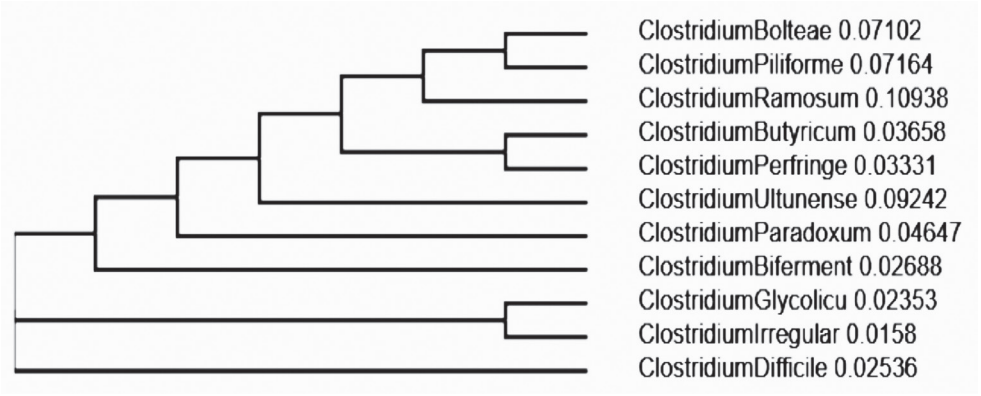


Şekil 4: a) Clostridium'lar için 16S rRNA gen uzunluklarının çubuklu diagram grafiği,
b) Clostridium'lar için 16S rRNA gen uzunlukları boxplot grafiği

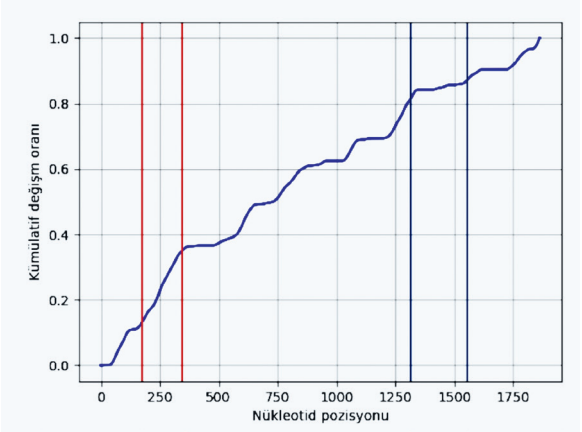
Şekil 5: Clostridium'lar için nükleotid dağılımları a) Clostridium'lar için nükleotid dağılımı (toplam nükleotid sayısına oran olarak verilmiştir), b) Clostridium'lar için nükleotid frekanslarının boxplot grafiği

Geliştirilen çatı ile mikrobiyota çalışmalarında genus seviyesinde tiplendirilmesi oldukça zor olan Clostridia türlerinden rastsal olarak seçilmiş 10 adet genus çoklu dizi hizalama metodu ile hizalanmış ve neigborhood joining metodu ile filogetik ağacı oluşturulmuştur (Şekil 6).

Şekil 6: Rastsal olarak seçilmiş Clostridia türleri için oluşturan filogenetik ağaç



Sonrasında hizalanmış veri üzerinde entropi çalışması yapılarak yazılabilecek en düzgün superstring oluşturulmuş ve ilgili Şekil Ek 1'de sunulmuştur. Superstring incelendiğinde bazı alanların oldukça yüksek entropi değerine sahip olup, oldukça korunmuş diziler olduğu tespit edilmiştir. Bazı sahalarda ise entropi değerleri oldukça düşük olarak görülmüş ve bu sahalar yüksek derecede farklılık gösteren sahalar olarak düşünülmüştür. Primerler yazılırken bu sahalara has primerlerin geliştirilmesi mikrobiyotada biyoçeşitlilik analizleri için oldukça önemlidir. Geliştirilen çatı ile yazılan superstring, araştırmacılara kendi çalıştıkları mikroorganizma gruplarına has özel primerler dizileri tasarlamaları konusunda kolaylıklar sağlıyacaktır. Literatürde tarif edilen yüksek derecede değişken bölgeler superstring üzerinden okunarak dizi haline getirilmiş ve bu diziler değişim oranları ile birlikte Tablo 2 'de gösterilmiştir. Superstring içindeki yüksek derecede değişken ve korunmuş bölgelerin tespiti için superstring 25 nükleotitlik pencereler şeklinde alt dizilere ayrılmış, bu alt diziler üzerinde n nükleotitlerinin (A,T,G ve C nükleotitlerinden herhangi biri) yüzdelik karşılıkları hesaplanmış (değişim oranı) ve grafik Şekil 7 'de sunuşmuştur.



Şekil 7: Kümülatif değişim oranı grafiği. Görselleştirmek için her 25 nükleotitlik pencere boyutundaki değişim oranlarının ardışık toplamlar şeklinde çizilmiştir. Kırmızı dikey çizgiler arası yüksek derecede değişkenlik gösteren (yüksek eğimli), mavi dikey çizgiler arası ise yüksek derecede korunmuş (plato) bölgelerdir.

Bölge	Bu bölgelere ait diziler (Şekil - 8'den)	Değişim
V1	nnnnnnnAnGnnTnnnnTnAnnnnnnnnnCTTC	0.72
V2	CGGGT	0.00
V3	AGGGTGATCGGCCACATTGGAACTGAGACACGGTCCAGACTCCTACGGGAG-GCAGCAGTGGGGA	0.00
V4	GCnTCTGTCnnnCTnAAnnGGAAGAnnnnnnnnnTnnnnnnAAnnnTGnnnnnnnnnnnnACGGTAC-nnTTGnnnnnnnnAGGAGGAAnGCnCCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACG-TAnGGGGGCnAGCGTTATCCGGATTTACTGGGCGTAAAGnnGnGTGCGTAGGCGGTn-nAnnTnnTTnAAGnnTCAnnnGnnnnGTGAAAnnnnnGGCnTAnGGCTCAACCnTAnGnnnTAA-GCnnTTnnnnGAAACTGnnnGnnnnGnAGnnCTTGAGTGnCAGGAGAGGA	0.32
V5	AGCnnTTnnnnGAAACTGnnnGnnnnGnAGnnCTTGAGTGnCAGGAGAGGAnAnnGTG	0.34
V6	TGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGTACTAG	%0
V7	TCCGCCTGGGGAGTACGnTCGCAAGAnTGAAACTCAAAGGAATTGACGGGGACCCGC	0.03
V8	TCCGCCTGGGGAGTACGnTCGCAAGAnTGAAACTCAAAGGAATTGACGGGGACCCGC	0.61
V9	NnCnnTnnGnnnnACnnCnnTnCnnCnTnnnTAnnnATCnGAGnnnnTTnnn	0.06

Tablo 2: Yüksek derecede değişkenlik gösteren dizilerin elde edilen superstring üzerinden elde edilen diziler ve bu dizilerdeki değişkenlik oranları

Şekil 7 'deki grafikte eğimi yüksek sahalar yüksek derecede değişken olan alanları ve eğimi 00'ye yakın olan sahalar (plato sahaları) yüksek derecede korunmuş alanları işaret etmektedir. Grafikteki plato sahalarının toplam 9 tane olduğu ve yerleşimlerinin ise eğimin sertçe yükseldiği alanlardan önce olduğu görülmektedir. Bu grafik şekli yüksek derece değişkenlik gösteren sahalar ile korunmuş sahalar arasındaki ilişkiyi destekler tarzdadır. Plato ve yüksek bir eğimle artmakta olan sahalar değerlendirilerek yüksek oranda değişiklik gösteren sahaların başlangıç ve bitiş pozisyonları, sahaların dizileri, değişim oranı ve bölgenin önündeki 25 nükleotitlik saha yeniden hesaplanarak Tablo 3'de sunulmuştur.

No*	Başlangıç	Bitiş	Dizi	Değişim oranı	Olası primer**
I	60	127	GnnnnnnnnnnnnAnGnnTnnnnT-nAnnnnnnnnnCTTCGGnnTG-nAnnnnnnnnnTTnnnnnnnnnn	0.75	GCGTGCCTAACA-CATGCAAGTCGAG
II	174	343	nnnnnnnGGnATAAnCnAnTnnCC-nnnnnnnnnnnGAAAnnGGnAnTnGCTAATAC-nCnnnGnATAAnnnnTnnnnnnnnnTnnnnnnnnn-nAnnGnnnnnnnnnnnnnnCGCAT-GnGnnnTnnCnnTnnnAnTnnnAnnTCnnn-nAAAGnnnCTnnnGnnnGnnnCnnnnnnn	0.64	CGTGnGGTAACCT-GCCTCnTACACA
III	604	655	nnnnnTnnnnnnAAnnnTGnnnnnnnnnnn-nACGGTACnnTTGnnnnnnnnA	0.69	TCTGTCnnnCT-nAAnnGGAAGAnnn
IV	758	855	nnAnnTnnTTnAAGnnTCAnnnGnnnnGT-GAAAnnnnnGGCnTAnGGCTCAACC-nTAnGnnnTAAGCnnTTnnnnGAAACT-GnnnGnnnnGnAGnnC	0.45	GCGTAAA-GnnGnGTGCGTAG-GCGGT

V	1048	1088	nCGGnGGnGnnnnnnnnnTTACC-nnnnTCGnnnnnnnnnn	0.65	GCCGTAAACGAT-GAGTACTAGGTGT
VI	1226	1314	nAnAGnCTTGACATnCCnnCnnTnnGnnn-nACnnCnnTnCnnCnTnnnTAnnnATCn-GAGnnnnTTnnnnnnnnCnnnnnCnCnnnnn	0.60	CGAAGCAACGC-GAAGAACCTTACCT
VII	1555	1610	nnACAGAnGGGnAGCnAnnnAGnCn-nCGnTGAGnGTGGAGCnAATCCCTTAAAnA	0.28	CTACACACGTGC-TACAATGGnTGGT
VIII	1747	1800	nnAnCnnnACCCGAAGCCnnGTG-nAnCTAACnCnnnnnGnAAGnGAGnnnnCT	0.41	CCGCCCGTCACAC-CATGGGAGTTGG
IX	1855	1888	nnGGnnAnnnnnnnTGC-nnnnnnnnnnnnnnnn	0.82	GTGAAGTCG-TAACAAGGTAGC-CGTA

Tablo 3: Superstring dizinin kümülatif değişim grafiğinden yola çıkarak yeniden hesaplanmış yüksek derecede değişkenlik gösteren ve korunun sahalar ile bunların dizileri ve değişim oranları
* Bölge numaraları romen rakamları ile verilmiştir

** Olası primer sahası bölge başlangıcından hemen önceki 25 nükleotitlik kısım olup özellikle bir primer dizisi olarak hesaplanmamıştır

Bu çalışmada, yapılan biyoinformatik analizler tarafımızca geliştirilen açık kaynak kodlu Pyton çatısı kullanılarak gerçekleştirilmiştir. Kullanılan biyoinformatik prensipler ile, primer dizaynında ve 16S rRNA geninin konumlandırılmasında daha etkin çözümler getirilmiştir rRNA geninin konumunu internet kaynaklı ortalama bir değer almak yerine, biyoinformatik temelli yaklaşımlarla daha gerçekçi olarak değerlendirmek mümkün olduğu gösterilmiştir. rRNA geni kopyalanması için primer geliştirilmesi konusunda da bu değerlendirme çok önemli olacaktır. Özellikle PCR protokolünü gerçekleştirmek üzere primerler seçilirken yüksek kapsama yüzdesine sahip olması ve elde edilen PCR amplikonunun kimliklendirme için yeterli bilgiyi içermesi gereklidir. 16S rRNA geni üzerinde bütün bakteri türlerini kapsayacak özellikte üniversal diziler var olmadığından, uygun primer seçimi analiz sonuçlarını oldukça etkileyecektir.

Tartışma

Genel olarak, bu çalışmada mikrobiyota konusunda çalışmaya yeni başlayan, özellikle 16S rRNA ve dizileme çalışması yapacak araştırmacılar için klavuzluk yapacak nitelikte bilgiler verilmeye çalışılmış ve açık kaynak kodlu çatı uygulaması temelleri atılmaya çalışılmıştır. Mikroorganizmaların sınıflandırılmasında 16S rRNA genlerindeki korunmuş diziler ve bu dizilere yakın lokalizasyondaki yüksek dercede değişkenlik gösteren bölgeler kullanılmaktadır (V1-V9)[5]. Bu çalışmalarda temel problem, yüksek derecede değişkenlik gösteren bölgelerin başlama ve bitiş noktalarının her mikroorganizma için birebir aynı olmamasıdır. Ayrıca kullandığımız universal primerlerin tasarladığımız çalışmalarda sonuç vermemesi diğer bir problemdir. Literatürde verilen universal primerlerin bazı türler üzerinde iyi bir sınıflandırmaya yetecek kadar tutarlı sonuç vermesi fakat bazı türlerde ise başarılı olamamasının altında yatan nedenlerden birisi bu olabilir.

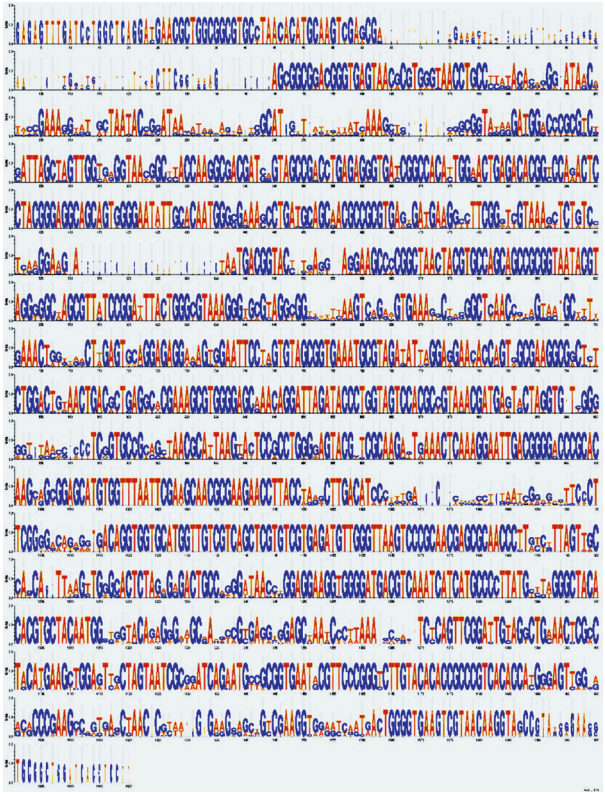
Bu çalışmada önerilen metot doğrultusunda yeniden hesaplanmış yüksek derecede değişkenlik gösteren dizilerin başlangıç ve bitiş noktalarından yararlanarak yazılacak primerler ile daha verimli sonuçlar alınabileceği gösterilmiştir. Şekil 8'de Clostridyum'lar için oluşturulmuş bir superstring dizi görülmektedir. Bu diziden yola çıkarak yüksek deredece değişkenlik gösteren bölgeler tekrar

hesaplanmış ve sonuçları Tablo 2'de sunulmuştur. V1 ve V8 bölgelerinde ki değişkenlik oranı oldukça yüksekken diğer bölgelerde bu değişkenliğin çok daha azdır. Diğer bölgelerin seçilen türde korunmuş olması iyi bir sınıflandırma yapılmasına engeldir. Bu durum seçilen tür ile ilgili olarak yapılan çalışmalarda karşılaşılan tiplendirme problemlerini açıklayabilir. Bu nedenle bahsedilen bölgelerin yeniden oluşturulması ve oluşturulan bu bölgelere göre primerlerin dizayn edilmesinin başarıyı arttıracağı gösterilmiştir.

Tablo 1 ve Tablo 2' deki bilgiler ışığında 25 nükleotitlik parçalar şeklinde superstring yeniden değerlendirildiğinde ve herhangi bir nükleotid olma olasığının 0.80'in üzerinde olduğu bölgeler ele alındağında 1855 – 1888 arasında bir bölgenin yüksek derecede değişken olduğu ve bu bölgenin hemen önündeki 25 nükeotidlik sahanın GTGAAGTCGTAACAAGGTAGCCGTA (A:8, T:5, G:8, C:4, GC%=48) şeklinde ve primer yazmak için oldukça uygun bir dizi olduğu dikkati çekmektedir. Benzer analizler 0.75, 0.70 gibi daha düşük değerler için tekrar edilerek çalışmaya özel farklı bölgeler hesaplanabilir. Bu hesaplama başarıyı arttıracaktır.

Geliştirilen çatı yazılımın, laboratuvar standartlarımıza ve çalışmayı hedeflediğimiz mikroorganizma topluluklarına göre universal primerler yerine ihtiyaçlarımız doğrultusunda geliştirilmiş primerler yazma konusundaki katkısı gösterilmiştir. Ancak, bu çalışmada 1886 Clostridiumdan rasgele seçilen sadece 11 adedi ile yapılan analiz sonuçları verilmiştir. Farklı setler ile yada farklı örnek genişliği ile çalışmanın tekrar edilmesi, farklı organizmalar ile de çalışılarak yapılacak analizlerden elde edilecek benzer sonuçlar ile çalışma kuvvetlendirilebilir. İlave olarak benzer çalışmaların artması, elde edilen teorik sonuçların laboratuvar çalışmaları ile doğrulanması ve elde edilen pratik bilgiler çerçevesinde ortaya çıkacak sonuçların sunulan yazılımın geliştirilmesine katkısı olacağı inancındayız.

Şekil Ek 1: Clostridium türleri üzerinde entropi çalışması yapılarak oluşturulmuş superstring



nAGAGTTTGATCnTGGCTCAGGATGAACGCTGGCGGCGTGCCTAACACATGCAAGTCGAGG
nnnnnnnnnnnnAnGnnTnnnnTnAnnnnnnnnnnCTTCGGnnTGnAnnnnnnnnnnTTnnnnnnnnnnAG
CGGCGGACGGGTGAGTAACGCGTGnGGTAACCTGCCTCnTACACAnnnnnnnnGGnATAAnCnA
nTnnCCnnnnnnnnnnnGAAAnnGGnAnTnGCTAATACnCnnnGnATAAnnnTnnnnnnnnTnnnnnnn
nnAnnGnnnnnnnnnnnnnnCGCATGnGnnnTnnCnnTnnnAnTnnnAnnTCnnnnAAAGnnnCTnnnG
nnnGnnnCnnnnnnnnGGTATGAnGATGGACCnCnnnnnnGCGTCnGnATTAGCTAGTTGGTGAGG-
TAACGGCTTACCAAGGCGACGATnCAGTAGCCGACCTGAGAGGGTGATCGGCCACATTGGA-
ACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTnnnGCACAATGGn
nGCGAAAGCnCTGATGCnnAGCAACGCCGCGTGAnGnGATGAAGGCnCTTCGGnGTCGTAA-
AGCnTCTGTCnnnCTnAAnnGGAAGAnnnnnnnnnTnnnnnnAAnnnTGnnnnnnnnnnnnACGGTACnnT
TGnnnnnnnnAGGAGGAAnGCnCCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAnGGG
GGCnAGCGTTATCCGGATTTACTGGGCGTAAAGnnGnGTGCGTAGGCGGTnnAnnTnnTTnAAGn
nTCAnnnGnnnnGTGAAAnnnnnnGGCnTAnGGCTCAACCnTAnGnnnTAAGCnnTTnnnnGAAACTG
nnnGnnnnGnAGnnCTTGAGTGnCAGGAGAGGAnAnnGTGnGAATTCCnTAGTGTAGCGGTGAA-
ATGCGTAGATATTAnGGAGGAACACCAGTGGCGAAGGCGGCnnTnTCTGGACnnTGTnAACTGAC
GCTGAGGCACGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAAC-
GATGAGTACTAGGTGTnCGGnGGnGnnnnnnnnnnTTACCnnnnTCGnnnnnnnnnnnGTGCCnnnGCA
GCnTAACGCATTAAGTACTCCGCCTGGGGAGTACGnTCGCAAGAnTGAAACTCAAAGGAATTGA
CGGGGACCCGCACAAGCAGCGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAACCTTACC-
TnAnAGnCTTGACATnCCnnCnnTnnGnnnnACnnCnnTnCnnCnTnnnTAnnnATCnGAGnnnnTTnnn
nnnnnCnnnnnCnCnnnnnTTCGGnGGACAnnGnAGnnnnnnnTnnnnGAnnCAGGTGGTGCATGGTT
GTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGTCTTTAGT
TGnCCAGCAnTTAnAGTTGGGCACTCTAGAGAGACTGCCnGnGGATAACnCnGGAGGAAGGTG-
GGGATGACGTCAAATCATCATGCCCCTTATGnCTTAGGGCTACACACGTGCTACAATGGnTGGTnn
ACAGAnGGGnAGCnAnnnAGnCnnCGnTGAGnGTGGAGCnAATCCCTTAAAnAnnnnCCAnTCTC
AGTTCGGATTGTAGGCTGAAACTCGCCTACATGAAGCTGGAGTTGCTAGTAATCGCGGATCAGA-
ATGCCGCGGTGAATACGTTCCCGGGTCTTGTACACACCGCCCGTCACACCATGGGAGTTGGnnA
nCnnnACCCGAAGCCnnGTGnAnCTAACnCnnnnnGnAAGnGAGnnnnCTnGTCGAAGGTnGnAnT-
CAnnnATGACTGGGGTGAAGTCGTAACAAGGTAGCCGTAnnGGnnAnnnnnnnTGCnnnnnnnnnnn
nnnnnn

Şekil Ek 2: Superstring dizisi, burada 'n' herhangi bir nükleotid yerine kullanılmıştır.

Journal of BSHR
2018;2(1):23-34

1. Fukuda, K., et al., Molecular Approaches to Studying Microbial Communities: Targeting the 16S Ribosomal RNA Gene. J UOEH, 2016. 38(3): p. 223-32.
2. Rajendhran, J. and P. Gunasekaran, Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res, 2011. 166(2): p. 99-110.
3. Woese, C.R., O. Kandler, and M.L. Wheelis, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A, 1990. 87(12): p. 4576-9.
4. Lane, D.J., et al., Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A, 1985. 82(20): p. 6955-9.
5. Yang, B., Y. Wang, and P.Y. Qian, Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics, 2016. 17: p. 135.
6. Chakravorty, S., et al., A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods, 2007. 69(2): p. 330-9.
7. Galperin, M.Y., X.M. Fernandez-Suarez, and D.J. Rigden, The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. Nucleic Acids Res, 2017. 45(D1): p. D1-D11.
8. Zou, D., et al., Biological databases for human research. Genomics Proteomics Bioinformatics, 2015. 13(1): p. 55-63.
9. Benson, D.A., et al., GenBank. Nucleic Acids Res, 2013. 41(Database issue): p. D36-42.
10. Quast, C., et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res, 2013. 41(Database issue): p. D590-6.
11. Xu, R. and D. Wunsch, 2nd, Survey of clustering algorithms. IEEE Trans Neural Netw, 2005. 16(3): p. 645-78.
12. Elen A., T.M.K., Genetik algoritmalar ve çoklu dizi hizalama probleminin çözümü, in XIII Tıbbi Biyoloji ve Genetik Kongresi. 2013: Kuşadası.
13. Arslan S., T.T., Karcı A., , Çoklu-dizi hizalama problemi için genetik algoritma, in ELECO-2004: Elektrik–Elektronik-Bilgisayar Mühendisliği Sempozyumu. 2004. p. 353-356.
14. Woo, P.C., et al., Clostridium bacteraemia characterised by 16S ribosomal RNA gene sequencing. J Clin Pathol, 2005. 58(3): p. 301-7.