Nevşehir Bilim ve Teknoloji Dergisi

Derleme Makalesi(Review Article)

Makale Doi: **10.17100/nevbiltek.1005534**

Geliş Tarihi:06-10-2021 Kabul Tarihi:25-01-2022

Yeni Nesil Dizileme Verilerinin Analizinde Bulut Teknolojisi

Sema KARABUDAK [1*], Meryem Sena AKKUŞ [2]

[1]*Ankara Yıldırım Beyazıt Üniversitesi Merkez Araştırma Laboratuvarı Uygulama ve Araştırma Merkezi,Ankara*

ORCID ID: 0000-0002-3646-0442

[2]*Ankara Yıldırım Beyazıt Üniversitesi Merkez Araştırma Laboratuvarı Uygulama ve Araştırma Merkezi, Ankara*

ORCID ID: 0000-0003-2550-550X

Öz

Yeni nesil dizileme (YND) araçları, büyük miktarda veri üretme kapasitesine sahiptir ancak bu araçların hesaplama ve depolama kapasiteleri büyük ölçekli dizileme verilerinin analizi için yetersiz kalmaktadır. Bulut bilişim altyapılarını kullanmak YND verilerinin analizi, depolanması ve aktarılması ile ilgili sorunlara alternatif bir seçenek olmuştur. Bulut bilişim, kullanıcılara dizileme verilerinin analizi için gerekli hesaplama kapasitesi ve bilişim altyapılarına erişim sunmakta ve biyoinformatik altyapıları için gerekli olan ön sermaye harcamalarının çoğunu ortadan kaldırmaktadır. Bu çalışmada yeni nesil dizileme yöntemi hakkında bilgi verilmiş ve elde edilen dizileme verilerinin analizinde kullanılan yazılımlar bulut bilişim servis modellerine ve son kullanıcılara göre sınıflandırılarak özetlenmiştir.

Anahtar Kelimeler: Bulut Bilişim, Veri Analizi, Yeni Nesil Dizileme, Biyoinformatik

Cloud Computing in Next Generation Sequencing Data Analysis

Abstract

Next-generation sequencing (NGS) tools are capable of generating large amounts of data, but their computational and storage capacity is insufficient for analysis of large-scale sequencing data. Using cloud computing infrastructures has been an alternative option to problems related to analysis, storage and transfer of NGS data. Cloud computing provides users with the computing capacity and access to computing infrastructures required for analysis of sequencing data, and eliminates most of the up-front capital expenditure required for bioinformatics infrastructures. In this study, information is given about the next generation sequencing method and the software used in the analysis of the obtained sequencing data are summarized by classifying them according to cloud computing service models and end-users.

Keywords: commaCloud computing, Data Analysis, Next Generation Sequencing, Bioinformatics

[*] **Sorumlu yazar**: skarabudak@ybu.edu.tr

1. Giriş

Yeni nesil dizileme (YND) teknolojisi tüm genomun veya genomun bir kısmının nükleotid dizisini belirlemek için kullanılan yüksek verimli dizileme yöntemidir [1]. Günümüzde yaygın olarak kullanılan YND teknolojisinin kökleri 1953 yılında Watson, Crick ve Franklin'in DNA'nın yapısını keşfetmelerine kadar uzanmaktadır. 1964'te Richard Holley tRNA'nın dizilimini gerçekleştirerek nükleik asitleri dizilemeye yönelik ilk girişimde bulunmuştur. Walter Fiers 1972'de, 2 boyutlu fraksiyonlama yöntemini kullanarak bakteriyofaj MS2'nin kaplama proteinini ve bundan dört yıl sonra da tam genomunu dizilemiştir. 1977 yılında DNA dizilimi için Fredrick Sanger zincir sonlandırma yöntemini önerirken aynı yıl Maxam ve Gilbert, DNA'nın kimyasal modifikasyonuna dayanan dizileme yöntemini tanıtmışlardır. Applied Biosystems, 1987 yılında yarı otomatik DNA dizileme yöntemini geliştirmiş ve 1996 yılında dizileme ve fragment analizi yaklaşımları için tasarlanmış otomatik tek kapiler genetik analiz cihazı ABI 310'u tanıtmıştır. 2004 yılında 454 Life Sciences, Roche GS20 olarak adlandırılan ve pazardaki ilk YND platformu olan piro-dizileme teknolojisini pazarlamıştır. 2007'de SOLiD sistemi tanıtılmış ve 2011 yılında hidrojen iyonlarını tespitine dayalı olarak nükleotidleri tespit eden Ion Torrent teknolojisi geliştirilmiştir [2 ve 3].

YND platformları tüm genom dizileme, ekzom dizileme, transkriptom dizileme (RNA-dizileme), protein-DNA etkileşimlerinin analizi (Chip-Seq) ve dizileme tabanlı DNA metilasyon analizi gibi pek çok farklı uygulama alanına sahiptir. Geliştirilen veri analizi araçlarıyla birlikte YND platformlarının artan kullanımı; biyolojik araştırmalar, klinik tanı ve tıbbi uygulamaların kapasitesini önemli ölçüde arttırmıştır. [4]. YND aracılığıyla üretilen veriler, hastalıkların biyo-belirteçlerinin doğru bir şekilde tanımlanması, kalıtsal bozuklukların tespit edilmesi, tedavilere verilen yanıtların tahmin edilmesine yardımcı olabilecek genetik faktörlerin belirlenmesi ve kanserin erken saptanması gibi önemli faydalar sağlamaktadır [5]. Hızlı ve maliyeti uygun biçimde dizilemeye olanak veren teknolojileri tanımlamada kullanılan YND, yüksek verimli dizileme olarak da bilinmektedir [6].

Son yıllarda, YND teknolojilerindeki teknik iyileştirmelere bağlı olarak dizileme maliyetlerinin önemli ölçüde düşmesi dizileme çalışmalarını kolaylaştırmış ve üretilen veri hacminde muazzam bir artışa neden olmuştur [7]. Büyük hacimli YND verilerinin işlenmesi, analizi, depolanması ve yönetilmesi için yüksek kapasiteli sabit disk ve yüksek düzeyde hesaplama gücüne ihtiyaç duyulmaktadır. YND cihazları tarafından üretilen büyük miktarda genomik verinin depolanması, aktarılması ve analizi ile ilgili problemler, iyi bilinen "büyük veri" sorununun bir örneğidir [6 ve 8]. Üretilen verinin hacmi, üretim hızı ve formatı (çeşitliliği) veriden alınacak değeri etkilemektedir. Bu nedenle, verileri güvenli bir şekilde depolayabilen ve yeterli bilgi işlem gücüne sahip altyapının seçimi, dizilemeden sonraki adımlarda kritik önem taşımaktadır. Bulut bilişim, birden fazla bilgisayardan yararlanarak hesaplama ve depolama kaynaklarını internet üzerinden sanal bir şekilde sağlamaktadır. Son yıllarda geliştirilen YND yazılım sistemlerinin çoğu, bulut tabanlı platformlar üzerinden bulut tabanlı hizmet sağlamaktadır. Bulut bilişim teknolojisi, hesaplama kaynaklarının sonsuz kullanılabilirliği sayesinde hızlı veri işlemeye olanak sağlayarak veri işleme ile ilgili sorunları çözmeyi sağlayan önemli bir teknoloji haline gelmiştir [6 ve 9]. Karşılaştırmalı genomik, genom analizi, SNP araştırmaları ve gen ekspresyonunun düzenlenmesi gibi çalışmalar YND verilerinde bulut bilişimin kullanıldığı başlıca alanlara örnek olarak verilebilir [8].

2. Bulut Bilişim Teknolojisi

Bulut bilişim kavramı, John McCarthy'nin 1965'te "hesaplamanın bir gün kamu hizmeti olacak şekilde organize edilebileceği" fikrini dile getirmesiyle ortaya çıkmıştır [4]. ABD Ulusal Standartlar ve Teknoloji Enstitüsü bulut bilişimi "minimum yönetim çabası veya hizmet sağlayıcı etkileşimi ile hızla sağlanabilen ve serbest bırakılabilen, ortak bir yapılandırılabilir bilgi işlem kaynakları havuzuna her yerde, uygun, isteğe bağlı ağ erişimi sağlayan bir model'' olarak tanımlamaktadır. Kısaca ifade etmek gerekirse, bulut bilişim, kullanıcıların bilgi işlem kaynaklarını satın almak yerine bunları kolayca kiralayabilmeleri için organize edilmiş bir modeldir [10].

Bulut bilişim modelinde, işlemciler ve sabit diskler gibi hesaplama kaynakları, bir sağlayıcıdan kiralanabilen yardımcı programlar olarak düşünülmektedir. "Bulut sağlayıcı" terimi çoğunlukla Amazon Web Services (AWS), Google Cloud Platform veya Microsoft Azure gibi ABD merkezli başlıca ticari hizmetleri tanımlamak için kullanılır. Sağlayıcılar, veri merkezlerinde düzenlenen geniş bilgisayar ve depolama havuzlarını kontrol ederler. Kullanıcılar ise bu kaynakları talep eder, kullanır ve iş tamamlandığında bu kaynakları havuza geri verirler [10].

Bulut bilişim; geniş ağ erişimi, hızlı esneklik, kaynak havuzu, isteğe bağlı self servis ve ölçülebilir hizmet olarak beş temel özellikten oluşur. Bulut bilişimde genel bulut, özel bulut, hibrit bulut ve topluluk bulutu olmak üzere dört dağıtım modeli bulunmaktadır.

Bulut bilişim modelinin temel özellikleri:

1. İsteğe bağlı self servis: Müşteri, hizmet sağlayıcıyı dahil etmeden bilgi işlem kaynaklarını tek taraflı olarak sağlayabilir.
2. Geniş Ağ Erişimi: Bulut bilgi işlem kaynaklarına bilgisayar ve akıllı telefon gibi çeşitli müşteri platformları tarafından ağ üzerinden erişilebilir.
3. Kaynak Havuzu: Birden çok kullanıcı aynı kaynaklardan geçici ve ölçeklenebilir hizmet alabilmektedir.
4. Hızlı esneklik: Sağlayıcı tarafından esnek yönetim sayesinde kullanıcı sınırsız bilgi işlem kaynaklarına erişebilir.
5. Ölçülebilir Hizmet: Kaynak kullanımı otomatik olarak izlenebilir ve kontrol edilebilir [4].

Bulut bilişim teknolojisinin dağıtım modelleri:

1. Özel Bulut: Tek bir kullanıcı veya birden fazla istemciye sahip bir kuruluş tarafından çalıştırılan bulut türüdür [11]. Diğer bulut türlerine göre performans, güvenilirlik ve güvenlik bakımından en yüksek düzeyde kontrol sunan bulut türüdür[12].
2. Genel Bulut: Bulut kaynakları üçüncü şahıslar tarafından sağlanmakta ve web tarayıcıları aracılığıyla kaynaklara erişilebilmektedir. Amazon Elastic Compute Cloud, Google Apps Engine ve Microsoft Azure bu tip bulutlara örnek olarak verilebilir [11]. Veri, ağ ve güvenlik ayarları üzerinde denetimden yoksun olmaları genel bulutların eksik yanlarıdır [12].
3. Topluluk Bulutu: Üçüncü kişiler tarafından yönetilebilen bir bulutu çeşitli organizasyonların ortaklaşa paylaştığı bulut bilişim türüdür [11]. Bu tür bulut modelinde, altyapı ve hesaplama kaynakları, tek bir kuruluş yerine ortak gizlilik, güvenlik ve yasal düzenlemelere sahip iki veya daha fazla kuruluşa özeldir [13].
4. Hibrit Bulut: Genel, özel veya topluluk bulutlarından oluşmuştur ve veri güvenliği sağlamak için daha güvenli bir seçenektir. Salesforce hibrit buluta örnek olarak verilebilir [11].

3. Yeni Nesil Dizilemede Kullanılan Bulut Tabanlı Uygulamaların Sınıflandırılması

Son yıllarda ortaya çıkan ABI SOLiD, Illumina, Ion Torrent gibi YND platformları, yüzlerce gigabayt boyutunda veri kümeleri oluşturmaktadır. Bu cihazlar dizileme sırasında veri toplamak için yeterli hesaplama ve depolama kapasitesine sahiptir. Ancak bu cihazların hesaplama gücü, genom montajı ve okumaların referans genoma hizalanması gibi dizileme sonrası veri analizi basamaklarını gerçekleştirmek için yetersizdir. Veri miktarının sürekli artması yüksek maliyetli bilişim altyapılarının kurulmasını gerektirmekte ve veri işlemeyi zorlaştırmaktadır [13] . Dizileme teknolojilerinin yaygın kullanımı ile birlikte güvenli şekilde veri depolayabilen, ulaşılabilirliği ve kullanımı kolay, büyük ölçekli verilerde hızlı analiz imkanı sunabilen, veri paylaşımına izin veren, veri görselleştirme olanağı sunan bulut

sistemlerine ihtiyaç her geçen gün artmaktadır [14]. Bulut bilişim sağlayıcılarının maliyetleri sürekli olarak düşürmesi biyoinformatik topluluğunda YND analizleri için bulut bilişimin kullanımına yönelik ilgiyi arttırmıştır [15].

Bulut bilişim, yazılım servisi (SaaS), platform servisi (PaaS), altyapı servisi (IaaS) ve veri servisi (DaaS) olmak üzere dört farklı servis modelinde hizmet vermektedir.

Bulut sağlayıcı, kendi bulutunda çalışan uygulamalar sunduğunda Yazılım Hizmeti (Software as a Service (SaaS)) terimi kullanılır [4]. SaaS modelinde, yazılım bakımları ve güncellemeleri kolaylaşmıştır. Ağ altyapısı ve bileşenleri gibi unsurları yönetemeyen kullanıcılar sadece bulut yöneticisinin izin verdiği uygulamaları kullanabilirler. Saas servis modelinde çalışan bir web portalı olan SNP2Structure, yabani ve mutant tip proteinleri karşılaştırarak sessiz tekli nükleotid değişimlerinin (msSNP) protein yapısını nasıl değiştirdiğini görselleştirmekte ve analizleri kolaylaştırmaktadır[16].

İkinci servis modeli olan Platform Hizmeti (Platform as a Service (PaaS)) modelinde müşteriye, bulut sağlayıcı tarafından desteklenen geliştirme araçlarını kullanarak uygulamalar oluşturma yetkisi verilmektedir. PaaS, hızlı uygulama geliştirme ve iyi ölçeklenebilirlik özelliklerine sahip olup büyük biyolojik verilerin analizi için özel uygulamalar geliştirmeye olanak sunmaktadır. PaaS modeli, programlama dili ortamları, web sunucuları ve veri tabanlarını kapsamaktadır [8]. Dizileme verilerinin analizinde de bu servis modelinde çalışan yazılımlar tercih edilmektedir. Örneğin, CLUSTOM-CLOUD yazılımı, bulut ortamında 16S rRNA dizi verilerini kümelemek için bellek içi veri sistemi tabanlı bir yazılım olup, Paas servis modelinde çalışmaktadır[17].

Ayrıca, mRNA ve kodlayıcı olmayan RNA analizi için gen ifade düzeylerini incelemeyi amaçlayan Bio-VLAB-MMIA-NGS yazılımı, hem PaaS hem de SaaS servis modellerinde çalışmaktadır[18].

Altyapı Hizmeti (Infrastructure as a Service (IaaS)) modelinde müşterilere hizmet olarak bilgi işlem altyapısı (IaaS), bilgisayarlar, disk alanı ve ağ bant genişliği gibi düşük seviyeli yetenekler sunulmaktadır [10]. Bu modelde müşteri, sadece izin verilen işletim sistemi ve uygulamalar üzerinde tam kontrole sahiptir [8]. Amazon EC2 başlıca ticari IaaS bulut sağlayıcılarından biridir. Metagenomik çalışmaların yazılım araçları genellikle IaaS servis modelindeki bulutları kullanmakta olup bu alanda kullanılan BLAST (Basic Local Alignment Search Tool) gibi pek çok uygulama Amazon EC2'de test edilmiştir. Cloud Virtual Resource (CloVR)-16S, CloVR-Metagenomics, CloVRMicrobe ve CloVR-Search Iaas servis modelinde çalışan uygulamalara örnek olarak verilebilir. Cloud Virtual Resource (CloVR)-16S, Sanger ve 454 dizi verilerinin 16S rRNA tabanlı mikrobiyal topluluk kompozisyon analizini sunarken [19] CloVR-Metagenomics, metagenomik tüm genom shotgun dizileme verilerinin taksonomik ve işlevsel kompozisyon analizine olanak vermektedir[20]. CloVR Microbe, Sanger, 454 veya Illumina dizi verileri için IGS Annotation Engine'i kullanan bir bakteriyel tek genom dizisi oluşturma programı iken [21] CloVR-Search, Sanger, 454 veya Illumina dizi verilerinin büyük ölçekli bir BLAST arayıcısı olarak çalışmaktadır [22]. Metagenomik çalışmalara ek olarak, hastalıklar ve genler arasındaki bağlantıda önem taşıyan Tekli nükleotid değişimlerinin (SNP'ler) araştırılmasında kullanılan Fasta, BLAT, MUMmer ve GATK gibi biyoinformatik araçları da IaaS bulutlarında konuşlandırılmıştır. Bakteriyel genlerin tanımlanması yoluyla bulaşıcı hastalıkların teşhisini kolaylaştıran ERGATIS de Iaas bulut modelinde çalışan uygulamalara örnek olarak verilebilir[23].

Son servis modeli olan veri hizmeti (Data as a Service (DaaS)) modeline göre verilere ağ üzerinden isteğe bağlı olarak erişilebilmekte ve verilerin dağıtılması sağlanabilmektedir. Kullanıcının herhangi bir yerden veri depolamasına ve verilere erişebilmesine olanak sağlayarak veri erişim sınırlamalarının üstesinden gelebilen DaaS modeli biyoinformatik çalışmalarda çok önemli bir servis modelidir. Amazon Web Service (AWS), kullanıcıların verilere erişimini sağlayan bulut tabanlı bir uygulamadır. AWS, DaaS modeline örnek olup Ensembl ve GenBank gibi büyük biyolojik veri bankaları da dahil olmak üzere pek çok kaynaktan herkese açık veri kümeleri içermektedir[24].

Son kullanıcılara göre değerlendirildiğinde, YND verilerinin bulut bilişimde analiz açık kaynak kodlu araçlar, ticari sistemler ve özelleştirilmiş sistemler olmak üzere üç seçenek bulunmaktadır. Ticari sistemlere örnek olan DNAnexus ve Seven Bridges gibi araçlar veri analizi için doğrudan kullanılabilirler. İkinci türde, ticari veya açık biyoinformatik platformları, kullanıcıların hesaplama ihtiyaçlarını karşılamak için daha da özelleştirilmiştir. Üçüncü tür olan açık kaynaklı araçlar herhangi bir özelleştirilmiş veri analizi için buluta yerleştirilebilir[8]. Şu anda bulut bilişimi destekleyen birçok işlem hattı ve iş akışı vardır. Tablo 1'de Bulut ortamında YND veri analizi için kullanılan araçlar bulut bilişim teknolojisi servis modellerine ve son kullanıcılara göre sınıflandırılmış ve kullanım amaçları özetlenmiştir.

Tablo 1: Bulut ortamında YND veri analizi için kullanılan araçların bulut bilişim teknolojisi servis modellerine ve son kullanıcılara göre sınıflandırılması

Servis Modeli	YND Aracı	Kullanım Amacı	Son kullanıcılara Göre Sınıflandırma	Kaynak
DaaS	Amazon Web Servisi	GenBank, Ensembl, 1000 Genomes, Model Organism Encyclopedia of DNA Elements, Unigene gibi veritabanlarında bulunan genel verilere kontrollü erişim sağlamaktadır.		
SaaS	SNP2Structure	msSNP'Lerin protein yapısına etkisinin araştırılması	Açık kaynak	[25]
	Rainbow	Bulut bilişim kullanarak büyük ölçekli tüm genom dizileme veri analizi için bir araç	Açık kaynak	[15]
	CloudBurst	MapReduce ile son derece hassas okuma haritalama	Açık kaynak	[26]
	VAT	Varyant anotasyonu	Açık kaynak	[27]
	Myrna	RNA dizileme verilerinden gen ifade farklılıklarının hesaplanması	Açık kaynak	[28]
	BlastReduce	Kısa okumaların haritalanması	Açık kaynak	[29]
	Seal	Burrows-Wheeler Aligner eşlemeleriyle tutarlı olan kısa okuma çift eşlemeleri	Açık kaynak	[30]
	CloudBrush	Yeni nesil dizileme verilerinin de novo birleştirilmesi	Açık kaynak	[31]
	Cloudgene	Genel bulutta büyük ölçekli veri işleme ve özel bulutlar üzerinde iş akışı yeniden üretilebilirliği için MapReduce tabanlı Grafik kullanıcı arabirimi çerçevesi	Açık kaynak	[32]
	Cumulus	Tek-hücre ve tek-çekirdek RNA-seq verilerinin analizi	Açık kaynak	[33]
	Peak ranger	Chip-Seq verilerinin analizi	Açık kaynak	[34]
	Atlas2	Varyantların belirlenmesi	Açık kaynak	[35]
	FX	RNA-Seq analizi aracı	Açık kaynak	[36]
	YunBe	Gen Set Analizi	Açık kaynak	[37]
	StormSeq	Okumaların haritalanması	Açık kaynak	[38]
	StormBow	RNA-Seq analizi aracı	Açık kaynak	[39]
	Jnomics	Java tabanlı dizi analizi paketi	Açık kaynak	[40]
	Seq-Map-Reduce	Dizi haritalama algoritması	Açık kaynak	[41]
	Crossbow	SNP'leri araştırılması	Açık kaynak	[42]
PaaS	Eoulsan	Yüksek Verimli Dizileme analizleri için tasarlanmış modüler ve ölçeklenebilir iş-akış motoru	Açık kaynak	[43]
	Galaxy	Biyomedikal araştırmalar için açık kaynaklı, web tabanlı platform	Açık kaynak	[44]
	GalaxyCloud	Büyük ölçekli analizlerin tekrarlanabilirliğini sağlayan genomik araştırmalar için bulut tabanlı çerçeve	Açık kaynak	[45]
	SparkSeq	RNA ve DNA dizileme verilerinin nükleotid hassasiyetiyle işlenmesi	Açık kaynak	[46]

	Cloud Biolinux	Yüksek performanslı biyoinformatik bilgi işlem için talep üzerine hızlı bir şekilde altyapı hazırlanmasına olanak tanıyan, herkese açık bir sanal makinedir.	Açık kaynak	[47]
	CloudMan	YND verilerinin, analizlerin ve analiz araçlarının paylaşılmasını sağlayan platform	Açık kaynak	[48]
	ClustomCloud	16S dizileme verilerinin analizi	Açık kaynak	[49]
	SeqPig	Hadoop'ta büyük veri dizileme kümeleri için basit ve ölçeklenebilir komut dosyası oluşturma	Açık kaynak	[50]
	DNANexus	Dizileme verileri için bulut tabanlı veri analizi ve yönetim platformu sağlar	Ticari	[51]
	BioPig	Büyük ölçekli sekans verileri için Hadoop tabanlı bir analitik araç seti	Açık kaynak	[52]
IaaS	Cloud Virtual Resource (CloVR)-16S	16S rRNA tabanlı mikrobiyal topluluk kompozisyon analizi	Açık kaynak	[19]
	CloVR-Metagenomics,	Tüm genom shotgun dizileme verilerinin taksonomik ve işlevsel kompozisyon analizi	Açık kaynak	[20]
	CloVRMicrobe	Bakteriyel tek genom dizisi oluşturma programı	Açık kaynak	[21]
	CloVR-Search	Sanger, 454 veya Illumina dizi verilerinin büyük ölçekli bir BLAST arayıcısı	Açık kaynak	[22]
	ERGATIS	Genomik verilerin analizi için işlem hatları oluşturmaya, yürütmeye ve izlemeye olanak tanıyan iş akışı yönetim sistemi	Açık kaynak	[23]
	RAP-Search2	YND verilerini kullanarak protein benzerliklerini arama aracı	Açık kaynak	[53]
	GalaxyCloudman	Amazon'un EC2 bulut altyapısında bilgi işlem kümelerini yapılandırmaya yönelik çözümler sunan bulut yönetim sistemi	Açık kaynak	[54]
	Cloud Aligner	Dizilerin haritalanması için hızlı ve tam özellikli MapReduce tabanlı bir araç.	Açık kaynak	[55]
	Genome Analysis Toolkit	YND araçları için verimli ve sağlam analiz araçlarının geliştirilmesini kolaylaştırmak üzere tasarlanmış yapılandırılmış bir programlama çerçevesidir	Açık kaynak	[56]
	MEGAN	Metagenomik ve metatranskriptomik verilerin taksonomik ve işlevsel olarak analizi	Açık kaynak	[57]
	MG-RAST	Metagenomik dizileme verilerinden Mikrobiyal Topluluk Yapısı ve İşlevinin Analizi	Açık kaynak	[58]
	DIYA	Bakteriyel genomların görselleştirilmesi için bakteriyel genom dizilerinin annotasyonu	Açık kaynak	[59]

4. Sonuç

Yeni nesil dizileme yöntemi, başta tıp olmak üzere birçok bilim dalında çeşitli amaçlarla kullanılmaktadır. Bu yöntemin önemi ve değeri her geçen gün artmaktadır. Yüksek miktardaki YND verilerinin analizi bulut bilişim sistemi kullanılarak hızlı bir şekilde yapılabilmektedir. Bu çalışmada YND veri analizinde en çok kullanılan bulut tabanlı araçlar bulut bilişim servis modeline ve son kullanıcılara göre sınıflandırılmıştır. Servis modeline göre 4 tipe ayrılan toplam 42 bulut tabanlı aracın tıbbi ve biyolojik alanlarda kullanım amaçları özetlenmiştir. İncelenen araçların çoğu açık kaynaklı olup içlerinden sadece DNANexus ticari araçlara örnek olarak verilebilmektedir. Her geçen gün düşen dizileme maliyetleri ve artan hesaplama olanakları sayesinde bulut bilişim teknolojisine olan ihtiyaç artacak ve bu çalışmada örnek verilen araçlar gibi YND verileri için geliştirilen araç sayısı da artacaktır.

5. Teşekkür ve Katkı Beyanı

S.K.: Çalışmanın dizayn edilmesi, çalışma için kaynakların taranması, makale yazımı M.S.A.: Çalışma için kaynakların taranması, makale yazımı

6. Kaynaklar

[1] Behjati S., Tarpey P. S., "What is next generation sequencing?," ***Archives of Disease in Childhood. Education and Practice Edition***, 98, 236-238, 2013.

[2] Barba, M., Czosnek, H., & Hadidi, A." Historical perspective, development and applications of next-generation sequencing in plant virology," ***Viruses***, 6, 106–136, 2014.

[3] Goodwin, S., McPherson, J. & McCombie, W., "Coming of age: ten years of next-generation sequencing Technologies," ***Nat Rev Genet***, 17, 333–351, 2016.

[4] Kwon T., Yoo W. G., W.-J. Lee W.J., Kim W., Kim D.W., "Next-generation sequencing data analysis on cloud computing," ***Genes & Genomics***, 37, 489-501, 2015.

[5] Pereira M., Malta F., Freire M., and Couto P., "Application of Next-Generation Sequencing in the Era of Precision Medicine. In Applications of RNA-Seq and Omics Strategies – From Microorganisms to Human Health", ***Intech Open***, 2017.

[6] Celesti F., Celesti A., Carnevale L., Galletta A., Campo S., Romano A., "Big data analytics in genomics: The point on Deep Learning solutions," ***22nd IEEE Symposium on Computers and Communications (ISCC), Abstract Book***, 306-309, 2017.

[7] Schmidt B. , Hildebrandt A., "Next-generation sequencing: Big data meets high performance computing," ***Drug Discovery Today***, 22, 712-717, 2017.

[8] Zhao S., Watrous K., Zhang C., and Zhang B., "Cloud Computing for Next-Generation Sequencing Data Analysis," ***InTechOpen***, 29–51, 2017.

[9] Thakur R., Bandopadhyay R., Chaudhary B., Chatterjee S., "Now and next-generation sequencing techniques: Future of sequence analysis using cloud computing," ***Frontiers in Genetics***, 3, 280-280, 2012.

[10] Langmead B. and Nellore A., "Cloud computing for genomic data analysis and collaboration," ***Nature Reviews Genetics***, 19, 208-219, 2018.

[11] Baker Q. B., Al-Rashdan W., and Jararweh Y., "Cloud-Based Tools for Next-Generation Sequencing Data Analysis," ***2018 5th International Conference on Social Networks Analysis, Management and Security (SNAMS),*** **Abstract Book** 99-105s, Valencia-Spain, 2018.

[12] Zhang Q., Cheng L., and Boutaba R., "Cloud Computing: State-of-the-art and research challenges," ***Journal of Internet Services and Applications***, 1, 7-18, 2010.

[13] Dai, L., Gao, X., Guo, Y., Xiao, J., Zhang, Z., "Bioinformatics clouds for big data manipulation," ***Biology direct,*** 7, 1-7, 2012.

[14] Goyal S., "Public vs private vs hybrid vs community - cloud computing: A critical review," ***International Journal of Computer Network and Information Security***, 6, 20-29, 2014.

[15] Zhao S., Prenger K., Smith L., Messina T., Fan H., Jaeger E., "Rainbow: a tool for large-scale whole-genome sequencing data analysis using cloud computing," ***BMC Genomics***, 14, 425-425, 2013.

[16] Wang, D., Song, L., Singh, V., Rao, S., An, L., Madhavan, S., "SNP2Structure: a public and versatile resource for mapping and three-dimensional modeling of missense SNPs on human protein structures," **Computational and structural biotechnology journal,** 13, 514-519, 2015.

[17] Oh, J., Choi, C. H., Park, M. K., Kim, B. K., Hwang, K., Lee, S. H.,Kim, K. M., "Clustom-cloud: In-memory data grid-based software for clustering 16s rrna sequence data in the cloud environment," ***PloS one,*** 11, e0151064, (2016).

[18] Chae, H., Rhee, S., Nephew, K. P., Kim, S., ”BioVLAB-MMIA-NGS: microRNA–mRNA integrated analysis using high-throughput sequencing data,” ***Bioinformatics,*** *31*, 265-267, 2015.

[19] White, J., Arze, C., Matalka, M., Team, T. C., Angiuoli, S., Fricke, W. F., “CloVR-Metagenomics: Functional and taxonomic microbial community characterization from metagenomic whole-genome shotgun (WGS) sequences–standard operating procedure,” ***Nature Precedings***, 1, 1-1, 2011.

[20] Fricke, W., White, J., Arze, Matalka,M., White,O., Angiuoli,S., ‘’CloVR-Metagenomics: Functional and taxonomic microbial community characterization from metagenomic whole-genome shotgun (WGS) sequences – standard operating procedure, version 1.0.’’ ***Nature Precedings***, 1,1-1,2011.

[21] White, O., Angiuoli, S., Fricke, W. F., Galens, K., White, J., Arze, C., Team, T. C., “CloVR-Microbe: Assembly, gene finding and functional annotation of raw sequence data from single microbial genome projects–standard operating procedure,” ***Nature Precedings***, 1, 1-1, 2011.

[22] http://clovr.org/methods/clovr-search/

[23] Orvis, J., Crabtree, J., Galens, K., Gussman, A., Inman, J. M., Lee, E., Angiuoli, S. V., “Ergatis: a web interface and scalable software system for bioinformatics workflows,” ***Bioinformatics,*** 26, 1488-1492, 2010.

[24] Dai, L., Gao, X., Guo, Y., Xiao, J., Zhang, Z., “Bioinformatics clouds for big data manipulation,” ***Biology direct,*** 7, 1-7, 2012.

[25] Wang,D., Song,L., Singh,V., Rao,S., An,L., Madhavan,S., ‘’ SNP2Structure: A Public and Versatile Resource for Mapping and Three-Dimensional Modeling of Missense SNPs on Human Protein Structures,’’ ***Computational and structural biotechnology journal***, 13, 514-519, 2015.

[26] Schatz,M., ‘’CloudBurst: highly sensitive read mapping with MapReduce’’ ***Bioinformatics***, 25,1363-1369, 2009.

[27] Habegger, L., Balasubramanian, S., Chen, DZ., Khurana,E., Sboner,A., Harmanci,A., Rozowsky,J., Clarke,D., Snyder,M., Gerstein,M., “VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment,’’ ***Bioinformatics***. 28,2267-2269,2012.

[28] Langmead B., Hansen K., Leek J., "Cloud-scale RNA-sequencing differential expression analysis with Myrna," ***Genome Biology***, 11, R83, 2010.

[29] Schatz, Michael C. "BlastReduce: high performance short read mapping with MapReduce." ***University of Maryland, http://cgis. cs. umd. edu/Grad/scholarlypapers/papers/MichaelSchatz. pdf***, 2008.

[30] Pireddu, L., Leo, S., Zanetti, G., “SEAL: a distributed short read mapping and duplicate removal tool,” ***Bioinformatics,*** 27, 2159-2160, 2011.

[31] Chang, Y. J., Chen, C. C., Ho, J. M., & Chen, C. L., “De novo assembly of high-throughput sequencing data with cloud computing and new operations on string graphs,” *In 2012 IEEE Fifth International Conference on Cloud Computing IEEE,* 155-161, 2012.

[32] Schönherr, S., Forer, L., Weißensteiner, H., Kronenberg, F., Specht, G., & Kloss-Brandstätter, A., “Cloudgene: a graphical execution platform for MapReduce programs on private and public clouds,” ***BMC bioinformatics,*** 13, 1-9, 2012.

[33] Li, Bo, Gould, J., Yang, Y., Sarkizova, S., Tabaka, M., Ashenberg, O., Regev, A. "Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq." ***Nature methods***, 17, 793-798, 2020.

[34] Nordberg H., Bhatia K., Wang K., Wang Z., "Biopig: a Hadoop-based analytic toolkit for large-scale sequence data," ***Bioinformatics***, 29, 23, 2013.

[35] Challis, D., Yu, J., Evani, U. S., Jackson, A. R., Paithankar, S., Coarfa, C., Yu, F., "An integrative variant analysis suite for whole exome next-generation sequencing data," ***BMC bioinformatics,*** 13, 1-12, 2012.

[36] Lu W., Jackson J., Barga R., "AzureBlast: A case study of developing science applications on the cloud," 2010. *Conference: Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, HPDC 2010,* 21-25 June 2010, 413-420, Chicago, Illinois, USA, 2010.

[37] Zhang L., Gu S., Liu Y., Wang B., Azuaje F., "Gene set analysis in the cloud," **Bioinformatics**, 28, 294-295, 2012.

[38] Karczewski, K. J., Fernald, G. H., Martin, A. R., Snyder, M., Tatonetti, N. P., Dudley, J. T.," STORMSeq: an open-source, user-friendly pipeline for processing personal genomics data in the cloud," ***PloS one***, 9, e84860, 2014.

[39] Zhao, S., Prenger, K., Smith, L., Stormbow: a cloud-based tool for reads mapping and expression quantification in large-scale RNA-Seq studies," ***ISRN Bioinformatics***, 2013, 1-8, 2013.

[40] Zhao, S., Prenger, K., Smith, L., Stormbow: a cloud-based tool for reads mapping and expression quantification in large-scale RNA-Seq studies," ***ISRN Bioinformatics***, 2013, 1-8, 2013.

[41] Li, Y., Zhong, S., "SeqMapReduce: software and web service for accelerating sequence mapping," ***Critical Assessment of Massive Data Anaysis (CAMDA)***, 2009, 1-5, 2009.

[42] Gurtowski J., Schatz M. C., Langmead B., "Genotyping in the cloud with Crossbow," ***Current Protocols in Bioinformatics***, 15, Unit15.3, 2012.

[43] Jourdren L., Bernard M., Dillies M.-A., Crom S. Le, "Eoulsan: A cloud computing-based framework facilitating high throughput sequencing analyses," ***Bioinformatics***, 28, 1542-1543, 2012.

[44] Blankenberg, D., Hillman-Jackson, J., "Analysis of next-generation sequencing data using Galaxy," *In Stem cell transcriptional networks*, Humana Press, New York, 21-43, 2014.

[45] Afgan, E., Baker, D., Coraor, N., Goto, H., Paul, I. M., Makova, K. D.,Taylor, J., "Harnessing cloud computing with Galaxy Cloud," ***Nature biotechnology***, *29*, 972-974, 2011.

[46] Wiewiórka M. S., Messina A., Pacholewska A., Maffioletti S., Gawrysiak P., Okoniewski M. J., "SparkSeq: fast, scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision," ***Bioinformatics***, 30, 2652-2653, 2014.

[47] Krampis K., Booth T., Chapman B., Tiwari B., Bicak M., "Cloud BioLinux: Pre-configured and on-demand bioinformatics computing for the genomics community," ***BMC Bioinformatics***, 13, 42, 2012.

[48] Afgan, E., Chapman, B., Taylor, J., "CloudMan as a platform for tool, data, and analysis distribution," ***BMC bioinformatics,*** *13*, 1-7, 2012.

[49] Oh, J., Choi, C. H., Park, M. K., Kim, B. K., Hwang, K., Lee, S. H., Kim, K. M., "Clustom-cloud: In-memory data grid-based software for clustering 16s rRNA sequence data in the cloud environment," ***PloS one***, 11, e0151064, 2016.

[50] Schumacher A., Pireddu L., Niemenmaa M., Kallio A., Korpelainen E., Zanetti G., "SeqPig: simple and scalable scripting for large sequencing data sets in Hadoop," ***Bioinformatics***, 30(1), 119-120, 2013.

[51] Navale V., Bourne P. E., "Cloud computing applications for biomedical science: A perspective," ***PLoS Computational Biology***, 14,1006144, 2018.

[52] Nordberg H., Bhatia K., Wang K., Wang Z., "Biopig: a Hadoop-based analytic toolkit for large-scale sequence data," ***Bioinformatics***, 29, 23, 2013.

[53] Zhao, Y., Tang, H., Ye, Y.,"RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data," ***Bioinformatics,*** 28, 125-126, 2012.

[54] Afgan, E., Baker, D., Coraor, N., Chapman, B., Nekrutenko, A., Taylor, J., "Galaxy CloudMan: delivering cloud compute clusters," ***BMC bioinformatics,*** 11, 1-6, 2010.

[55] Nguyen T., Shi W., Ruden D., "CloudAligner: A fast and full-featured MapReduce based tool for sequence mapping," ***BMC Research Notes***, 4, 171, 2011.

[56] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., DePristo, M. A., "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," ***Genome research***, 20, 1297-1303, 2010.

[57] Huson, D. H., Weber, N., "Microbial community analysis using MEGAN," ***Methods in enzymology***, 531, 465-485, 2013.

[58] Keegan, K. P., Glass, E. M., Meyer, F., "MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Microbial environmental genomics*," Humana Press, New York, 207-233, 2016.

[59] Stewart, A. C., Osborne, B., Read, T. D., "DIYA: a bacterial annotation pipeline for any genomics lab.," ***Bioinformatics,*** 25, 962-963, 2009.