BÖLÜM 2 / CHAPTER 2

GENETİK VERİ ORTAMLARININ KULLANILABİLİRLİK ANALİZİ

USABILITY ANALYSIS OF GENETIC DATA ENVIRONMENTS

Orçun TAŞAR*

*İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, Türkiye
e-posta: orcuntasar@ogr.iu.edu.tr

DOI: 10.26650/B/ET07.2020.012.02

ÖZ

Yeni nesil DNA dizileme teknolojilerinin gelişmesi ile birlikte sağlık, biyoloji ve genetik alanında yapılan çalışmalar yön değiştirmiştir. Artık bu çalışmaların büyük bir kısmını, yeni nesil DNA dizileme platformlarından elde edilen büyük veriyi işlemeyen anlamlandırmak neredeyse imkansız hale gelmiştir. Bu durum genetik verinin işlenmesi kadar onun güvenli bir şekilde saklanması ve kullanıcılara bu verinin kolay ve hızlı bir şekilde ulaştırılması ihtiyacını doğurmuştur. Bu amaçla projemizde uzun yıllardır kullanıma açık olan ve çoğunlukla DNA dizisi veri tabanı veya genom tarayıcısı olarak adlandırılırsalar da aslında bir veri tabanından daha fazlasına sahip olan iki popüler platformun kullanılabilirliğinin belirli sayıdaki katılımcının eşliğinde incelenmesi amaçlanmıştır. Elde edilen bulgu ve değerlendirmelerin ışığında genetik uzmanlarının ve genetik ile ilişkili bölüm ve anabilim dallarında lisans veya lisansüstü eğitimlerine devam eden öğrencilerin genetik verileri barındıran arayüzler ile nasıl etkileşime geçtiğini ve bu arayüzler karşısında nasıl tercihlerde bulundukları üzerine saptamalar gerçekleştirilmiştir.

Anahtar Kelimeler: Genetik, Veri, Veritabanı, İnsan Bilgisayar Etkileşimi, Kullanılabilirlik

ABSTRACT

With the evolution of new generation DNA sequencing technologies, studies in the field of health, biology and genetics have changed direction. It is now almost impossible to make sense of much of this work without processing big data from new generation DNA sequencing platforms. This situation has created the need to store it safely as well as need for the processing of genetic data and to deliver this data to the users easily and quickly. With this in mind, with the contribution of a certain number of participants, our project aims to examine the availability of two popular platforms. That platforms have been open to use for many years and are usually called DNA sequences databases or genome scanners, but actually these platforms have more than one database. In the light of the findings and evaluations obtained, determinations have been made on how geneticists and students who continue their undergraduate or graduate education in genetic related programs interact with the interfaces containing genetic data and how they choose these interfaces.

Keywords: Genetics, Data, Database, Human-Computer Interaction, Usability

GİRİŞ

Genom Verisi

1977 yılında Frederick Sanger tarafından geliştirilmiş olan Sanger dizileme metodu (Sanger, 1977) ve bu metoda dayanarak 1990 yılında başlatılan ve on üç yılın ardından 2003 yılında başarılı bir şekilde tamamlanan İnsan Genom Projesi'nin (Lander ve diğ., 2001; Venter ve diğ., 2011) ardından genetik alanında yapılan çalışmalar hız kazanmıştır. İnsan Genom Projesi çıktısı olan insan referans genomunun oluşturulmasının tetiklediği çalışmalar ile genomik bilimin devrimsel bir süreçte girmiş, son on beş yılda laboratuvarlarda ve diğer uygulama alanlarında özellikle DNA dizileme amacıyla kullanılan teknolojilerde büyük bir gelişme yaşanmıştır. İnsan Genom Projesi kapsamında iki kurum tarafından dokuz genom ile başlayan dizileme çalışmaları 2.7 milyar dolara mal olmuştur (NHGRI, 2010). Günümüzde artık kırk sekiz insan genomu tüm laboratuvar süreçleri dahil dört gün gibi bir süre içinde dizilenebilmekte olup genom başına dizileme maliyeti 2 bin dolar civarına kadar gerilemiş durumdadır. Bu maliyetin çok yakın bir zamanda 100 dolar civarına ineceği öngörülmektedir. Dizileme teknolojisinin maliyetlerinin bu derece düşmüş olması dünyanın her tarafında, birçok laboratuvar da DNA dizileme işlemlerinin kolaylıkla gerçekleştirilebilmesini sağlamaktadır. Tüm bu sebeplerden dolayı 2003 yılından itibaren üretilen biyolojik verinin boyutu üssel bir şekilde artmış vaziyettedir. Genomik veri günümüzde artık 'büyük veri' olarak sınıflandırılmaktadır. Bu durum da genomik verinin doğru ve güvenli bir şekilde saklanarak kullanıcılar tarafından kolaylıkla ulaşılmasını sağlayacak veri tabanı tasarımları ihtiyacı doğurmuştur. Projemiz de insan bilgisayar etkileşimi kapsamında kullanılabilirliğini inceleyeceğimiz iki genomik veri ortamı, insan genomu da dahil olmak üzere birçok türün genomik verisini farklı izoformlardaki nükleotit ve amino asit dizilimlerini de içerecek şekilde tutarken bu dizilimleri de yayınlanan bilimsel çalışmalar, klinik etkiler, evrimsel süreçteki korunum vb. veriler ile ilişkilendirerek saklamaktadır.

National Center for Biotechnology Information

National Center for Biotechnology Information (NCBI) Amerika menşeli Ulusal Sağlık Enstitüsü'ne (National Institutes of Health - NIH) bağlı ve açık erişimli veri tabanları ve biyoenformatik araçları ile tüm dünya tarafından kullanılan bir genetik veri ortamı olup 1988 yılında kurulmuştur (NIH, 2019).

Şekil 1: NCBI ana sayfası

Ensembl

Ensembl, Avrupa Biyoenformatik Enstitüsü (European Bioinformatics Institute) ile Wellcome Trust Sanger Enstitüsü'nün beraber geliştirmiş olduğu bir genom veri tabanı projesidir. Kullanıma ilk olarak 1999 yılında açılmıştır. İçinde barındırdığı genom verileri yanında büyük verilere kolaylıkla ulaşılmasını sağlayan BioMart, Perl API gibi araçları da mevcuttur (Furey, 2006).

Şekil 2: Ensembl ana sayfası

İnsan Bilgisayar Etkileşimi

İnsan Bilgisayar Etkileşimi (İBE), insan ile bilgisayarın temasa geçtiği herhangi bir arayüzü ve bu arayüzün kullanımını inceleyen multidisipliner bir alandır. İBE kapsamında psikoloji, bilgisayar bilimleri, ergonomi, sosyoloji, antropoloji veya tasarım mühendisliği gibi alanlara ait bilgi birikimi sıklıkla kullanılmaktadır. Günümüzde insanlar ile bilgisayarların hemen hemen her alanda etkileşime geçtiği düşünüldüğünde temas noktası konumunda olan her bir arayüzün kolay kullanılabilir olması oldukça önemlidir (Card ve diğ., 1980).

İBE'nin oldukça yeni bir disiplin olduğu söylenebilir. 1980'lere kadar olan süreçte bilgisayarlara erişimi olan az sayıdaki insan çoğunlukla bilgisayar uzmanı veya kod yazabilen ve bilgisayarın ardındaki teknik ayrıntıları anlayabilen kişilerdi. Kullanıcılar bilgisayarda ihtiyaç duydukları uygulamaları kendileri kodladıklarından dolayı kullanım açısından çok sıkıntı çekmemişlerdir. 1980 öncesinde bu kadar dar bir kullanıcı tabanı olan bilgisayarlar ile insanların etkileşiminin incelenmesi çok önemli bir odak noktası olmamış olmalıdır. Ancak sonrasında kişisel bilgisayarların maliyetinin düşerek yaygınlaşmaya başlaması ile birlikte uzmanlık alanı bilgisayar veya bilgisayar programcılığı olmayan insanlar da teknik ayrıntılara çok boğulmadan bu cihazları kolay bir şekilde kullanmak istediler. Her ne kadar İBE'nin yeni olarak tanımlanabilecek bir disiplin olduğunu söylesek de köklerinin çok daha eskilere dayandığı da ortadadır. Özellikle sanayi devrimi ile birlikte insan ve makineler arasındaki etkileşimin araştırılmaya başlanması İBE'nin de temellerini oluşturmaktadır (Dix ve diğ., 2004).

Kullanıcılar çeşitli arayüzler aracılığı ile bilgisayar veya bilgisayar sistemleriyle sürekli iletişim halindedir. Bu arayüz bir grafik arayüzü olabileceği gibi, klavye, mouse, hoparlör, mikrofon gibi çeşitli girdi ve çıktı aygıtları da olabilmektedir. İBE tasarımcıları bu arayüzler ile kullanıcılar arasındaki etkileşimin etkin olması için kullanıcı deneyimini ön plana çıkararak kullanılabilirlik ve erişilebilirliği dikkate alırlar. Çünkü iyi tasarlanmamış bir arayüz kullanıcıların gerçekleştirmek istedikleri işlemleri yaparken zorlanmalarına sebep olur. Bu yüzden İBE'nin amacı bilgisayar veya benzer sistemlerin kullanılabilirliğini ve erişilebilirliğini önemli ölçüde arttırmak ve geliştirmektir (Fiona ve diğ., 2006).

PROBLEM

Genetik veri çok farklı ve çok sayıdaki parametreler ile beraber tanımlanmaktadır. Boyutu yüksektir ve çok hızlı bir şekilde durmaksızın üretilmektedir. 2025 yılı itibarıyla genetik verinin çok hızlı bir şekilde büyük veri üreten diğer üç platform ve alanın (astronomi, Twitter ve Youtube) çıktılarıyla eş değer boyuta geleceği tahmin edilmektedir (Stephens ve diğ., 2015). Böyle büyük bir verinin ileriye dönük nasıl saklanacağı ve saklandıktan sonra kolay ve hızlı

bir şekilde kullanıcılara nasıl ulaştırılacağı bir problemdir. Bu problemi kullanıcı tarafında aşmanın bir yolu genetik veri tabanlarını kullanılabilirliği yüksek olacak şekilde tasarlamaktır. Bu çalışmada amaç insan bilgisayar etkileşimi kapsamında, genetik uzmanlarının kendi alanlarındaki verileri barındıran arayüzler karşısında nasıl tercihler yaptıkları ve bu arayüzleri kullanırken zorlanıp zorlanmadıklarını tespit etmektir.

YÖNTEM

Çalışmanın bu bölümünde, genetik veriyi ve bu veriler üzerinde işlem yapabilmeyi sağlayan bazı biyoenformatik araçlarını içeren iki platform incelenmiş ve kullanılabilirlik ve erişilebilirlikleri kullanıcılar ile yapılan anket ve bu ankettten çıkan sonuçlara göre üçüncü parti araçlarla sınanmıştır.

Genetik Veri Ortamı Seçimi

Çalışma kapsamında kullanılabilirliği ve erişilebilirliği incelenmiş olan platformlar genetik çalışmalar kapsamında en sık kullanılanlar arasından seçilmiştir. Dünyada bu kapsamda en sık kullanılan iki platform, NCBI ve Ensembl' dir.

Anket Çalışması

Çalışma kapsamında fakültelerin Moleküler Biyoloji ve Genetik, Genetik ve Biyomühendislik, Tıbbi Genetik, Biyoloji ve benzer bölümlerde öğrenciliğine devam eden veya bu alanlardan çıkışlı olup özel sektörde veya akademide çalışan 59 katılımcıya toplamda 11 adet soru sorulup fikirleri alınmıştır. Katılımcılara yöneltilen sorular aşağıdaki gibidir:

1. NCBI ve/veya Ensembl ortamlarından herhangi birini daha önce kullandınız mı?
 - a. Her ikisini de kullandım
 - b. Sadece NCBI kullandım
 - c. Sadece Ensembl kullandım
 - d. Hiçbirini kullanmadım
2. NCBI ve/veya Ensembl ortamlarını kullanma sıklığınızı nasıl tanımlarsınız?
 - a. Yılda birkaç kere
 - b. Yılda 3-4 kere
 - c. 1-2 ayda bir
 - d. Ayda 2-3 kere
 - e. Haftada birkaç kere
 - f. Daha sık
3. NCBI ve/veya Ensembl ortamlarında en sık gerçekleştirdiğiniz işlem/işlemler hangileridir?

4. En sık kullandığınız NCBI veri tabanları hangileridir? (Birden fazla seçenek işaretlenebilir)
 - a. Bookshelf

- b. MeSH
- c. NLM Catalog
- d. PubMed
- e. PubMed Central
- f. Gene
- g. GEO DataSets
- h. GEO Profiles
- i. HomoloGene
- j. PopSet
- k. UniGene
- l. ClinVar
- m. dbGaP
- n. dbSNP
- o. dbVar
- p. GTR
- q. MedGen
- r. OMIM
- s. Assembly
- t. BioCollections
- u. BioProject
- v. BioSample
- w. Genome
- x. Nucleotide
- y. Probe
- z. SRA
 - aa. Taxonomy
 - bb. Protein
 - cc. Diğer:

5. En sık kullandığınız NCBI araçları hangileridir? (Birden fazla seçenek işaretlenebilir)
- a. BLAST
 - b. Primer-BLAST
 - c. Variation Viewer
 - d. 1000 Genomes Browser
 - e. PubMed Clinical Queries
 - f. e-PCR
 - g. NCBI araçlarını kullanmıyorum
 - h. Diğer:
6. En sık kullandığınız Ensembl araçları hangileridir? (Birden fazla seçenek işaretlenebilir)
- a. Variant Effect Predictor (VEP)
 - b. BLAST/BLAT
 - c. File Chameleon
 - d. Assembly Converter
 - e. Data Slicer
 - f. BioMart
 - g. Ensembl araçlarını kullanmıyorum
 - h. Diğer:
7. NCBI ve/veya Ensembl ortamlarında veriye nasıl ulaşıyorsunuz?
- a. Grafik arayüz üzerinden
 - b. API, BioPython, Bioconductor vb kütüphaneleri kullanarak scriptler aracılığı ile
 - c. Her iki şekilde

8. Basit ve hızlı bir işlemi her iki ortamda da gerçekleştirebilme imkanı var:
 - a. Tercihim NCBI olur
 - b. Tercihim Ensembl olur
 - c. Bu işlemi her iki ortamda da gerçekleştiririm
 - d. Diğer:
9. NCBI veri tabanlarında aradığım veriyi kolaylıkla bulabiliyorum:
 - a. Kesinlikle katılıyorum
 - b. Katılıyorum
 - c. Kararsızım
 - d. Katılmıyorum
 - e. Kesinlikle katılmıyorum
10. Ensembl veri tabanında aradığım veriyi kolaylıkla bulabiliyorum:
 - a. Kesinlikle katılıyorum
 - b. Katılıyorum
 - c. Kararsızım
 - d. Katılmıyorum
 - e. Kesinlikle Katılmıyorum
11. Güncel mesleki durumunuz nedir? (Birden fazla seçenek işaretlenebilir)
 - a. Öğrenci - Lisans
 - b. Öğrenci - Yüksek lisans
 - c. MSc.
 - d. PhD.
 - e. Özel sektör çalışanı
 - f. Akademi çalışanı

Kullanılabilirlik Analizleri

Proje kapsamında kullanılabilirlik performansı CogTool ile ölçülmüştür. CogTool genel amaçlı bir grafik arayüz prototipleme aracıdır (CogTool, 2014). Bir arayüz karşısında deneyimli bir insanın performansını otomatik olarak değerlendiren bir algoritmaya sahiptir. CogTool üzerinde kullanılabilirliğin ölçülebildiği yeni tasarımlar yapılabilirken aynı zamanda var olan tasarımlar da değerlendirilebilmektedir. Anket sonuçlarına göre katılımcılar tarafından sık gerçekleştirilen işlemleri temsil edecek üç işlem belirlenmiş ve bunlar hem NCBI hem de Ensembl platformları üzerinde CogTool adındaki üçüncü parti bir yazılım aracılığı ile test edilmiştir.

Şekil 3: Cogtool ile NCBI üzerinde belirli bir görevin gerçekleştirilmesi

BULGULAR

Bu bölümde Yöntem bölümünde belirtilen değerlendirme kriterleri göz önünde bulundurularak yapılan analizler ve sonucu elde edilen bulgulara yer verilmektedir.

Anket Değerlendirmeleri

Veri Tabanlarının Kullanım Durumları. Ankete katılan katılımcıların hepsi daha önce ilgili ortamların en az birini kullanmıştır. Katılımcıların %74.5'i her iki ortamı da kullandığını belirtirken, %22'si daha önce sadece NCBI ortamını kullanmıştır. Katılımcılar arasında daha önce sadece Ensembl ortamını kullananların oranı oldukça düşüktür: %1.69. Bu durum, NCBI ortamının Ensembl ortamına göre daha sık ziyaret edildiğini belirten ALEXA verileri ile de uyumluluk göstermektedir. İlgili grafik Şekil 4'te verilmiştir.

Şekil 4: Katılımcıların daha önce ilgili ortamları kullanıp kullanmama durumu

Ankete katılan katılımcıların %30'u bu ortamları haftada birkaç kereden daha sık, %28'i ise haftada birkaç kere olacak şekilde kullanmaktadır. Genetik alanında çalışma yapan insanların yarısından fazlasının oldukça sık bir şekilde bu ortamları kullandığı görülmektedir. İlgili veriler Şekil 5'te verilmiştir.

Şekil 5: Katılımcıların ilgili ortamları kullanma sıklığı

Veri Tabanlarının Kullanım Amaçları. Ankette katılımcılara sorulan üçüncü soru çalışmanın en önemli verilerinden birisini oluşturmaktadır. Bu ortamlarda yapılabilecek

işlemler yüzlerce ve çok çeşitli olabileceğinden dolayı cevap alanı serbest metin şeklinde bırakılmıştır. Bu soruya 59 katılımcıdan cevap olarak birçok farklı işlem çeşidi gelmiştir. Bu işlemler veriler elde edildikten sonra kategorilendirilmiştir. Sonuçlara göre bu her iki ortamda en çok yapılan işlem literatür taraması, FASTA formatında gen veya protein dizilerine ulaşma, anotasyon ve primer tasarımıdır. Cevapların kategorik olarak dağılım oranları Tablo 1’de verilmiştir.

Tablo 1: İlgili Ortamlarda En Sık Gerçekleştirilen İşlemlerin Tüm İşlemlere Oranları

İşlem	Oran (%)
Literatür taraması	33.3
FASTA formatında gen veya protein dizilerine ulaşma	21.1
Varyant etkisi inceleme, anotasyon	14.4
Primer tasarımı	12.2
Dizi hizalama, haritalama	10
Basit sorgular ile veri çekimi	6.6
Diğer	3.3

Çalışmada dördüncü sorusunda katılımcılara NCBI platformunda bulunan veri tabanlarından en sık hangilerini kullandıkları sorulmuştur. Sonuçlar üçüncü sorudan elde edilen bulgular ile doğru orantılıdır. Katılımcıların %86.5’i literatür taraması yapmak amacıyla PubMed veri tabanını kullanmaktadır. FASTA formatında genlerin dizilimlerinin elde edilebileceği Gene veri tabanı ise katılımcıların yaklaşık %68’i tarafından kullanılmaktadır. Varyantların klinik etkilerinin incelenebileceği ClinVar ve dbSNP veri tabanları ise katılımcıların yaklaşık üçte biri tarafından kullanılmaktadır. Katılımcıların en sık kullandığı NCBI veri tabanları grafiksel olarak Şekil 6’da verilmiştir.

Şekil 6: Katılımcıların NCBI ortamında en sık kullandıkları yedi veri tabanı

Çalışmanın beşinci sorusunda katılımcılara en sık kullandıkları NCBI araçlarının hangileri olduğu sorulmuştur. Elde edilen bulgular Tablo 1’de verilen NCBI’da en sık gerçekleştirilen işlemlerle uyushmaktadır. Katılımcıların %70’i BLAST aracını kullanarak dizi hizalama işlemlerini gerçekleştirirken, %47’si ise Primer-BLAST aracını kullanarak primer tasarlamaktır. Katılımcıların en sık kullandıkları NCBI araçları Şekil 7’de verilmiştir.

Şekil 7: Katılımcıların NCBI ortamında en sık kullandıkları biyoenformatik araçlar

Katılımcılara aynı zamanda Ensembl ortamında kullandıkları araçlar da sorulmuştur. Bu soruya gelen cevaplara göre Katılımcıların %46'sı BLAST/BLAT adlı aracı kullanarak dizi hizalama işlemlerini gerçekleştirmekteyken, %25'i de Variant Effect Predictor (VEP) adlı araç ile klinik anotasyon yapmaktadır. Katılımcıların %32'si ise herhangi bir Ensembl aracını kullanmadığını belirtmiştir. Katılımcıların en sık kullandıkları Ensembl araçları Şekil 8'de verilmiştir.

Şekil 8: Katılımcıların Ensembl ortamında en sık kullandıkları biyoenformatik araçlar

Veri Tabanlarının Tercih Edilme Durumları. Elde edilen bulgular neticesinde NCBI ve Ensembl'da birbirlerinin eşlenikleri olarak gözüken, bir DNA motifinin başka bir diziye lokal olarak hizalanmasını sağlayan BLAST ve BLAST/BLAT araçlarının kullanımında, bu araçların kullanan katılımcıların %96'sı NCBI ortamında bulunan BLAST adlı aracı kullanırken, %57'si ise Ensembl ortamında bulunan BLAST/BLAT adlı biyoenformatik aracı kullanmaktadır. Katılımcıların %53'ü ise her iki aracı da kullanmaktadır. Sadece BLAST kullananlar katılımcıların %43'ünü oluştururken, sadece BLAST/BLAT kullananların oranı %4 civarındadır. Aynı zamanda Ensembl araçlarını hiç kullanmayan katılımcıların oranı %32 iken, NCBI araçlarından herhangi birini kullanmayan katılımcıların oranı ise %12 olarak tespit edilmiştir. İlgili veriler Şekil 9 ve 10'da verilmiştir.

Şekil 9: Katılımcıların NCBI ve Ensembl ortamlarında dizi hizalama araçlarını kullanım oranları

Şekil 10: Katılımcıların NCBI ve Ensembl ortamlarında herhangi bir araç kullanmama oranları

Katılımcılara sekizinci soruda her iki ortamda da gerçekleştirilebilecek bir işlem için hangi ortamı tercih edecekleri soruldu. Elde edilen bulgulara göre katılımcıların yaklaşık %46'sı bu işlemi NCBI ortamında gerçekleştirmeyi tercih ederken, %15'i ise Ensembl ortamını tercih etmektedir. %36'lık bir kesim ise bu işlemi her iki ortamda da gerçekleştirmektedir. İlgili veriler Şekil 11'de verilmiştir.

Şekil 11: Katılımcıların aynı işlem için tercih ettikleri ortamlar

Aynı zamanda katılımcılara NCBI ve Ensembl veri tabanlarında aradıkları veriyi kolay bulup bulamadıkları da sorulmuştur. Bulgulara göre katılımcıların %73'ü NCBI veri tabanlarında aradıkları veriyi kolaylıkla bulabildiklerini belirtmiş olup bu oran Ensembl kullanan katılımcılarda %49 civarında kalmıştır. Burada dikkat çekmek istediğimiz bir nokta da verilerin bu veri tabanlarında kolaylıkla bulabildiği görüşlerinin küçük kısmının “Kesinlikle Katılıyorum” cevabından gelmekte olduğudur (NCBI - %16.95, Ensembl - %13.56). İlgili veriler Şekil 12 ve 13'te verilmiştir.

Şekil 12: NCBI veri tabanlarında aranılan verilerin kolay veya zor bulunabiliyor olması

Şekil 13: Ensembl veri tabanlarında aranılan verilerin kolay veya zor bulunabiliyor olması

Kullanılabilirlik Analizleri

Çalışma kapsamında kullanılabilirlik analizleri, bilişsel deneyimi ölçen CogTool aracı yardımıyla gerçekleştirilmiştir. CogTool üzerinde gerçekleştirilen görevler, anket sonuçlarına göre kullanıcıların en sık gerçekleştirdiği işlemlere paralel olarak tasarlanmıştır. CogTool sonuçları Tablo 2’de verilmiştir.

Tablo 2: CogTool ile İlgili Veri Ortamlarında Gerçekleştirilmiş Üç İşlem

İşlem	NCBI (sn)	Ensembl (sn)
Ana sayfa üzerinden belirli bir genin belirli bir transkriptine ait kodlayan dizisine ulaşılması	16.2	17.4
Ana sayfa üzerinden DNA hizalama araçlarına giderek işlemi gerçekleştirmek (DNA dizisinin önceden hafızaya kopyalandığı kabul edilmiştir.)	14.7 + (7)	15.08 + (49)
Ana sayfa üzerinden bir varyantın klinik etkilerinin gözlenebileceği sayfaya ulaşılması	18.8	27.2

SONUÇ VE TARTIŞMA

Genetik veriyi barındıran veri tabanlarının günümüzde kullanım oranları, artan ve birikmekte olan genetik verinin miktarıyla birlikte artmaktadır. Bu da genetik veri tabanlarının veriyi güvenli bir şekilde saklamalarının yanında araştırmacılara kolay ve hızlı ulaşılabilir bir şekilde sunmalarının gerekliliğini doğurmuştur.

Bu çalışma kapsamında genetik ve biyoloji alanında eğitim almış veya almakta olan 59 katılımcıya 11 soru sorulmuş ve genetik alanında çalışmalar yapan insanların dünya üzerinde

en çok kullanılan global iki genetik veri ortamı ile etkileşime geçerken yaptıkları işlemleri ve tercihleri irdelenmiştir. Yapılan bu anketin sonuçlarından yola çıkarak bu veri ortamlarının benzer işlemler için kullanılabilirlikleri CogTool aracı ile de analiz edilmiştir.

Elde edilen bulgulara göre 59 katılımcının hepsi bahsi geçen platformlar ile farklı sıklıklarla da olsa etkileşim halindedir. %59'luk bir kesim haftada birkaç kere veya daha fazla olacak şekilde bu ortamları kullanmaktadırlar. Bu durum NCBI ve Ensembl ortamlarının genetik alanında yapılan çalışmaların çok büyük bir kısmının ayrılmaz bir parçası olduğunu göstermektedir. Katılımcıların %75'i daha önce her iki ortamı da kullandığını belirtirken, %22'si ise daha önce sadece NCBI ortamını kullandığını belirtmiştir. Daha önce sadece NCBI ortamını kullanmış katılımcıların %55'i lisans veya yüksek lisans öğrencisidir. Henüz akademik çalışma yapmaya ciddi ölçüde başlamamış katılımcıların biyoenformatik üzerine aldıkları ilk derslerin NCBI ortamı üzerinden anlatılmış olması ve raporları için gerekli literatür taramasını NCBI üzerinden yapıyor olmaları bu durumun sebeplerinden birisi olmalıdır.

İlgili ortamlarda en sık gerçekleştirilen işlemin literatür taraması olması şaşırtıcı değildir. Özellikle NCBI dahilinde bulunan PubMed veri tabanı sağlık alanında yapılan çalışmaların listelendiği en büyük veri tabanıdır. Bunun haricinde ilgili genlerin nükleotit dizilerine ulaşılması ve dolaylı olarak buna bağlı olan primer tasarımı yine katılımcıların bu ortamlar üzerinde en sık gerçekleştirdiği işlemlerdendir. Varyant etkisinin klinik açıdan incelenmesi genetik alanında yapılan çalışmalarda sıklıkla bir önceki iki işlemin ardından gerçekleştirilmektedir. DNA hizalama ise eldeki bir DNA dizisinin genom üzerinde hangi bölge ile eşleştiğinin bulunmasını sağlamakla birlikte fonksiyonel çalışmalarda önemi büyük bir metottur; çalışma kapsamındaki katılımcılar arasında da sıklıkla gerçekleştirildiği görülmüştür.

Genetik uzmanları arasında betik dilleri ile yazılmış kütüphanelerle veri çekme işlemini gerçekleştirenlerin oranı da şaşırtıcı bir şekilde azımsanmayacak kadar büyük gözükmektedir: %43. Python gibi öğrenmesi ve uygulaması diğer dillere kıyasla daha kolay olan betik dillerinin yaygınlaşmasıyla birlikte yeni nesil genetik uzmanlarının da çok büyük bir çoğunluğunun ileride programlama yoluyla daha çok biyoenformatik alanına yöneceklerini öngörmek zor değildir.

NCBI ortamında en sık kullanılan veri tabanları en sık gerçekleştirilen işlemlerle paralellik göstermektedir. PubMed ve Gene veri tabanları katılımcıların %86'si ve %68'i tarafından kullanılmaktadır. Literatür taraması ihtiyacı PubMed kullanımını zorunlu kılarken, genler içindeki farklı izoformların nükleotit dizilimlerine ulaşmak için ise Gene veri tabanının sıklıkla kullanıldığı görülmüştür. ClinVar, dbSNP, OMIM gibi veri tabanları ise varyantların klinik etkileri ile genlerin ilişkili oldukları hastalıkların barındırıldığı veri tabanları olarak

kullanım sıklıkları diğer iki veri tabanının arkasından gelmektedir. En sık kullanılan iki NCBI aracı olarak da en sık yapılan işlemlerde de kendisine yer bulmuş olan DNA hizalama ve primer tasarımı işlemlerinin gerçekleştirilmesini sağlayan araçlar olduğu görülebilir: BLAST ve Primer-BLAST. Aynı oran sık kullanılan Ensembl araçlarında da görülmektedir. DNA hizalamaya yarayan BLAST/BLAT ve varyantların klinik etkisini inceleyebilen VEP, Ensembl ortamında en sık kullanılan iki araçtır. Bu noktada iki ortam arasındaki en büyük farklardan biri kullanılmayan araçlar oranıdır. Buna göre, katılımcıların %12'si hiçbir NCBI aracını kullanmaya gerek duymazken, bu oran Ensembl ortamı için %32'lere çıkmaktadır. Bu durumun NCBI araçları ile Ensembl araçları arasındaki çeşitlilik farkından kaynaklanıyor olması muhtemeldir. Fakat her iki ortamda da DNA hizalama aracı en sık kullanılan ilk araç olarak görülmektedir. Buna rağmen katılımcıların %96'sı NCBI dahilindeki BLAST adlı aracı kullanırken, Ensembl dahilindeki BLAST/BLAT adlı araç katılımcıların %57'si tarafından kullanılmaktadır. Sadece BLAST/BLAT'ın kullanım oranı %4 civarında olduğuna göre, bu aracı kullanan çok büyük bir kesmin aslında aynı işlem için -doğrulama yapmak amacıyla- NCBI dahilindeki eşleniğini de kullanıyor olmalıdır. Katılımcıların bu işlem için daha çok NCBI tercihinde bulunmaları tutarlılık, hız ve güvenilirlik gibi farklardan kaynaklanıyor olabilir. Katılımcılara sorulmuş olan sekizinci sorunun cevap dağılımları ile birlikte bu durumun daha da netleştiği görülebilir; aynı işlem için daha çok tercih edilen ortam %46'lık oranla NCBI olarak gözükmemektedir. Bu oran Ensembl için %15'tir. %36'lık bir kesim ise ilgili işlemi her iki ortamda da gerçekleştirmeyi tercih ederek hata oranını azaltmak istemektedir.

Genetik veri tabanları ile ilgili en önemli sorulardan birisi, aranılan verinin platform üzerinde kolay bulunup bulunamıyor olduğudur. Anket sonuçlarında hem NCBI hem de Ensembl üzerinde aranılan verinin katılımcılar tarafından çoğunlukla kolay bir şekilde bulunabildiği görülmüştür. NCBI için bu oran %73 iken, Ensembl için %49'dur. Fakat dikkat edilmesi gereken bir nokta da bu veri tabanlarının konunun uzmanı veya uzman adayları olan kişiler tarafından kullanılıyor olduğudur. Özellikle Ensembl'ı göz önünde bulundurursak her iki katılımcıdan biri Ensembl üzerinde aradığı veriyi kolaylıkla bulamamaktadır. Böyle özelleşmiş bir alan için bu oran oldukça yüksektir. Ayrıca NCBI ve Ensembl'ın kolay kullanılabilir olduğuna dair gözüken %24'lük fark, daha önce tartıştığımız NCBI tercihlerinin çoğunlukta olması ile de paraleldir ve önceki bulgularımızı desteklemektedir.

CogTool aracından elde edilen veriler çoğunlukla NCBI üzerinde yapılan işlemlerin Ensembl üzerinde yapılan eşleniklerine göre küçük farklarla daha hızlı tamamlanması şeklinde oluşmuştur. Özellikle katılımcıların sıklıkla kullandıklarını belirttikleri DNA hizlama işlemi için her iki ortamda da benzer sürelerle ve hamlelerle işlem yapılmasına rağmen -CogTool

sonuçlarından bağımsız olarak- NCBI’ın aynı hizalama işlemi için yedi saniye içinde, Ensembl’in ise kırk dokuz saniye içinde yanıt verdiği görülmüştür. Bu fark hizalanması beklenen DNA dizisinin uzunluğu arttıkça daha da artmaktadır. Katılımcıların BLAST/BLAT yerine daha çok BLAST tercih etmelerinin sebeplerinden birisi bu olmalıdır. Yine de CogTool ve türevlerinin kullanılabilirlik üzerine verdiği sonuçların tartışmaya açık olabileceği de daha önce gösterilmiştir (Jorritsma ve diğ., 2015).

Bulguların çeşitlendirilmesi amacıyla ileriki aşamalarda farklı coğrafik bölgeler için kategorilendirilmiş daha fazla katılımcı ile bu çalışmanın sonuçlarının derinleştirilmesinde fayda vardır. Katılımcılara anket soruları yöneltmenin yanında onlarla birebir görüşmelerde bulunmak daha detaylı sonuçlar ortaya koymaya yardımcı olacaktır. Ayrıca fikir ve görüş almanın dışında farklı sıklıklarla ilgili veri tabanlarını kullanan kişiler ile birlikte göz izleme çalışmalarının yapılması CogTool gibi üçüncü parti yazılımlardan elde ettiğimiz verilere katkı yapacaktır. Bu çalışmanın ileriki dönemlerde detaylandırılması sadece ilgili veri tabanlarının kullanılabilirliğini ölçmek için değil, ayrıca genetik uzmanlarının biyoenformatik problemleri çözerken genel olarak nasıl bir iş akışı izlediklerini ve dolayısıyla bilgisayar karşısında uyguladıkları çalışma profilini ortaya koyması açısından önemli olacaktır.

KAYNAKLAR / REFERENCES

- Card, K., et al. (1980). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, 23(7), 396–410.
- CogTool. (2014). General purpose UI prototyping tool, version 1.2.2 specifications. Ocak 2019 tarihinde <https://github.com/cogtool/cogtool/releases> adresinden erişildi.
- Dix, A., et al. (2004). *R. Human Computer Interaction*, 3rd ed., Pearson Education, Essex, ISBN: 10: 0-13-0461091
- Fiona, N., et al. (2006). Human–Computer Interaction Research In The Management *Information Systems Discipline. International Journal of Human Computer Studies*, 64(9), 787–788.
- Furey, T. S. (2006). Comparison of human (and other) genome browsers. *Hum Genomics*, 2(4), 266–270.
- Lander, S., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921.
- National Human Genome Research Institute. The Human Genome Project completion: Frequently Asked Questions. Ocak 2019 tarihinde <https://www.genome.gov/11006943> adresinden erişildi.
- National Institutes of Health. *About NCBI*, Our Mission. Ocak 2019 tarihinde <https://www.ncbi.nlm.nih.gov/home/about/mission/> adresinden erişildi.
- Jorritsma W., et al. (2015). Predicting human performance differences on multiple interface alternatives: KLM, GOMS and CogTool are unreliable. *Procedia Manufacturing*, 3, 3725–3731.
- Sanger, F., Nicklen, S. and Coulson, A. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), 5463–5467.
- Stephens, Z., et al. (2015). Big data: astronomical or genetical. *PLoS Biol*, 13(7), e1002195
- Venter J. C., et al. (2001). The sequence of the human genome. *Science*, 291, 1304–1351.