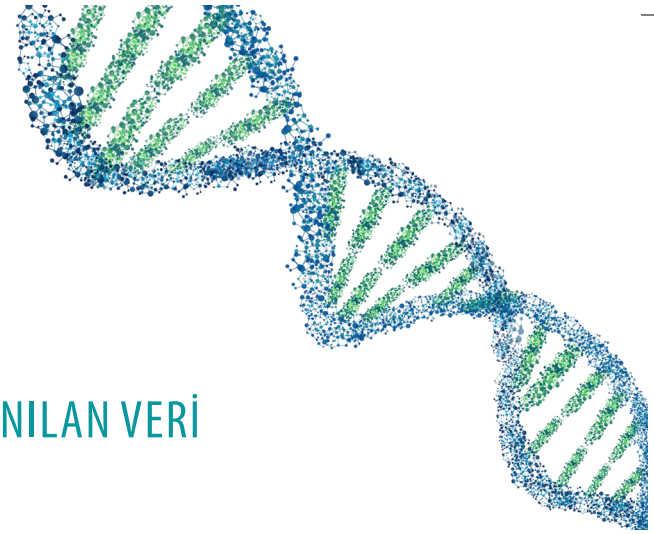

BÖLÜM 21

YENİ NESİL DİZİ ANALİZLERİNDE KULLANILAN VERİ ANALİZ SİSTEMLERİ, BİYOİNFORMATİK

Özkan Ufuk NALBANTOĞLU[1]

DNA dizileme teknolojilerinin ilk nesli olan Sanger Dizilemenin, gen seviyesinde ve düşük miktarda moleküler okuma üretmesi sebebiyle gen fragmanı hizalama, varyant analizi ve temel istatistiki analizleri, yeni nesil dizileme (YND) öncesi çağda biyoinformatik ihtiyacını karşılamaktaydı. Ancak ikinci nesil dizileme sistemlerinin mikrobiyal genomikte kullanıma girdiği 2000'li yılların ortalarından itibaren yüksek çıktılı ve *de novo* moleküler veri üretimi fırsatlar ile beraberinde çözülmesi gereken yeni hesaplamalı problemler getirdi. Bu bölümde, genel hatlarıyla en alt seviyeden en üst organizasyon seviyesine YND analizinde kullanılan biyoinformatik yaklaşımlar üzerinde durulacaktır.

DNA Okumalarından Sistem Biyolojisine

Yeni nesil dizileme sistemleri izolat veya çevresel örnek halinde mikroorganizmaların hedefli veya tüm genom/transkriptom verisinin nükleik asit fragmanları (okumalar) şeklinde elde edilebilmesini sağlamaktadır. Söz konusu ham veriler doğru şekilde işlenip analiz edildiği takdirde mikroorganizmalara ait tüm enzimatik ve yapısal proteinlerin, metabolik yolakların, sinyalleşme mekanizmalarının ve mikroorganizmaların ürettiği, yıktığı ya da dönüştürdüğü tüm küçük moleküllerin ortaya çıkarılması mümkündür. Bu amaçla yürütülecek dizi analizi, ele alınan organizmaların referans genomlarının bulunması (karşılaştırmalı genomik) ve referans olmadan (*de novo*) inceleme yapılmasına yönelik iki farklı yaklaşım ile yürütülebilmektedir.

Karşılaştırmalı Genomik

Daha önceden referans genomu elde edilmiş ve genom anotasyonu gerçekleştirilmiş organizmalara ait suşlardan elde edilen YND verisinin bu referansa kıyasla incelenmesini temel almaktadır. Referans genoma yüksek sayıdaki okumanın hizalanması ile her bir nükleotit pozisyonuna çok sayıda okunmuş baz denk gelmekte ve bunların konsensusları ile yeni dizilenen organizmanın referansla olan farklılıkları ortaya konmaktadır. Hizalama işlemi çoğunlukla yüksek performanslı kısa okuma hizalayıcıları (örneğin; Burrows-Wheeler Aligner (BWA)[1], Bowtie[2], Gem mapper[3]) ile gerçekleştirilmekte, hizalama sonrası analiz ise varyant

[1] Dr. Öğr. Üyesi Erciyes Üniversitesi Bilgisayar Mühendisliği Bölümü, nalbantoglu@erciyes.edu.tr

Plazmit tespiti: Tam genom dizilemesi için gerçekleştirilen DNA izolasyonu ile kromozomal DNA'nın yanı sıra plazmitler de izole edilmektedir. Anotasyonu yapılan genlerin kromozom üzerinde veya plazmit üzerinde taşınıyor olması ise ilişkili faktörlerin bir yatay transfer ile kolayca kaybedilebilir olup olmaması yönünden önem taşıyabilmektedir. Örneğin bir direnç geninin kromozoma entegre olması ile plazmitte taşınıyor olması mobilom aktivitesi açısından farklı değerlendirilmesi gereken durumlardır. Bu sebeple plazmitlerin tespitine yönelik spesifik analizler yürütülebilmektedir. PlasmidFinder programı[38], plazmit replikon veri tabanı ile plasmit tespiti yapabilen en yaygın ve geçerli yaklaşımlardan biridir.

YND Verisi Analizinin Problemleri ve Geleceği

Mevcut YND analiz paradigması, büyük ölçüde karşılaştırmalı genomik yaklaşımına dayanmaktadır. Buna göre, bir genomun içerisindeki genlerin ve yapısal elemanların tanımlanması için mevcut olarak tanımlanmış kataloglardan yakın homolojilerin tespiti yapılarak genomik işlevler tahminlenmektedir. Henüz gen ve protein uzayının tam olarak keşfedilememiş olması ve mevcut veri tabanlarında bu yönde önemli boşlukların olması, uzak homoloji tespitini verimli bir şekilde gerçekleştirebilen biyoinformatik yöntemlerin yoksunluğu sebebiyle tanımlamaların yeterince hassas olmamasına sebebiyet vermektedir. YND analizinin bu önemli probleminin aşılmasında gelecek dönemde uzak homoloji tahmini yapabilen algoritmaların geliştirilmesi önemli rol oynayabilir. Bu açıdan, son yıllarda büyük bir atılım içerisinde olan makine öğrenme ve özellikle derin öğrenme teknolojilerinin genom analizi alanında yaygınlık kazanması muhtemel bir çözüm olarak umut vadetmektedir. Bu teknolojiler yüksek hacimli veriler ve belli ölçüde standardizasyon gerektirdiğinden her geçen gün genişleyen veri setleri ve katalogların bu aşamada önemli rol oynayacağı ve dizi hizalama temelli konvansiyonel yöntemlere ek olarak matematiksel modellerin de YND analizinde önemli bir kullanım alanı bulacağı beklenmektedir.

Kaynaklar

1. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics 2009;25(14): 1754-60.
2. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology 2009;10(3):1-0.
3. Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nature methods 2012;9(12):1185-8.
4. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience 2021;10(2): giab008.
5. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 2010;1;20(9): 1297-303.
6. Zeevi D, Korem T, Godneva A, Bar N, Kurilshikov A, Lotan-Pompan M, et al. Structural variation in the gut microbiome associates with host health. Nature 2019;568(7750): 43-8.
7. Bolger A, Giorgi F. Trimmomatic: a flexible read trimming tool for illumina NGS data. Bioinformatics 2014;30(15): 2114-20.
8. Gordon A, Hannon GJ. Fastx-toolkit. FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit. 2010 Jan 20;5.
9. Brown J, Pirrung M, McCue LA. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics 2017;33(19): 3137-9.
10. Compeau PE, Pevzner PA, Tesler G. Why are de Bruijn graphs useful for genome assembly? Nature biotechnology 2011;29(11):987.
11. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology 2012;19(5):455-77.
12. Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 2012;28(11): 1420-8.
13. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 2015;31(10): 1674-6.
14. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature methods 2013;10(6): 563-9.
15. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS computational biology 2017;13(6): e1005595.
16. Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Current opinion in microbiology 2015;23: 110-20.
17. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and trans-

lation initiation site identification. BMC bioinformatics 2010;11(1):1-1.

18. Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic acids research 2005;33(suppl_2): W451-4.
19. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic acids research 1999;27(23):4636-41.
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology 1990;215(3):403-10.
21. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature methods 2015;12(1):59-60.
22. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research 2011;39(suppl_2): W29-37.
23. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 2000;28(1): 27-30.
24. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research 2019;47(D1): D309-14.
25. Bairoch A. The ENZYME database in 2000. Nucleic acids research 2000;28(1):304-5.
26. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nature genetics 2000;25(1): 25-9.
27. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005;21(18): 3674-6.
28. Huntley RP, Binns D, Dimmer E, Barrell D, O'Donovan C, Apweiler R. QuickGO: a user tutorial for the web-based Gene Ontology browser. Database 2009;2009.
29. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, et al. Pfam: The protein families database in 2021. Nucleic acids research 2021;49(D1): D412-9.
30. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic acids research 2009;37(suppl_1): D233-8.
31. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. Nucleic acids research 2016;44(14): 6614-24.
32. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic acids research 2014;42(D1): D206-14.
33. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The comprehensive antibiotic resistance database. Antimicrobial agents and chemotherapy 2013;57(7): 3348-57.
34. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome 2018;6(1): 1-5.
35. Chowdhury AS, Call DR, Broschat SL. PARGT: A software tool for predicting antimicrobial resistance in bacteria. Scientific reports 2020;10(1): 1-7.
36. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. Nucleic acids research 2005;33(suppl_1): D325-8.
37. Xie R, Li J, Wang J, Dai W, Leier A, Marquez-Lago TT, et al. DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy. Briefings in Bioinformatics 2021;22(3): bbaa125.
38. Carattoli A, Hasman H. PlasmidFinder and in silico pMLST: identification and typing of plasmid replicons in whole-genome sequencing (WGS). InHorizontal Gene Transfer 2020 (pp. 285-294). Humana, New York, NY.