Bioinformatics of Pathogens

Tomáš Vinař

Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia

Abstract

Genomic sequencing has become an important tool in identification and surveillance of human pathogens. Compared to large organisms, where our goal is to obtain high-quality sequences for detailed analysis, in pathogen sequencing the emphasis is often on optimization of cost and time. Consequently, sequencing of pathogens creates interesting computational challenges and development of new methods has a potential to significantly enhance applicability of the results in epidemiology and clinical practice. In my talk, I will give two examples: plasmid identification in bacterial isolates and genomic surveillance of wastewater for SARS-CoV-2. In both cases, application of better algorithms and modeling helps to improve the quality of analysis of very noisy data.

2012 ACM Subject Classification Applied computing → Bioinformatics

Keywords and phrases sequence analysis, genome assembly, neural networks, probabilistic modeling

Digital Object Identifier 10.4230/LIPIcs.WABI.2024.1

Category Invited Talk

Funding EU Horizon 2020 project number 872539 (PANGAIA), VEGA grant agency 1/0538/22, Slovak Research and Development Agency APVV-22-0144.

Acknowledgements These results are joint work with Broňa Brejová, Cedric Chauve, Janik Sielemann, Katharina Sielemann, Mattia Sgro, Paola Bonizzoni, Askar Gafurov, Jozef Nosek, Boris Klempa, and Kristína Boršová.

1 Plasmid Identification in Bacterial Isolates

Plasmids are extra-chromosomal DNA molecules, often circular and significantly shorter than bacterial chromosomes. They typically host genes beneficial to the bacteria, such as genes assisting in resistance to antibiotics. Identification of plasmids in bacterial isolates poses an interesting challenge, as these isolates are often sequenced using short-read sequencing, and such sequencing data can only be assembled to short contigs instead of whole chromosomes.

Recently, Pu and Shamir [4] introduced an idea that classification of contigs can be improved by considering neighbouring contigs in assembly graphs. The assembly graph summarizes possible connections of contigs to longer molecules, and are typically a byproduct of the short-read assembly. Building on this idea, we have built a graph neural network framework for plasmid identification [7]. Recently, we have also explored several ways of incorporating sequence similarity to databases of known sequences to improve plasmid identification. Finally, considering multiple genome assemblies in the form of a pan-genome graph, combined with flow-based approach to analysis of these graphs [3], can help to compensate for various artifacts of a particular assembler software, and further improve plasmid identification.

2 Genomic Surveillance of SARS-CoV-2 in Wastewater

In surveillance of pathogens, such as SARS-CoV-2, wasterwater-based epidemiology has recently emerged as a cost effective alternative to sequencing individual patient samples [6, 2]. Since concentration of target molecules is typically very low, the PCR-based laboratory techniques (such as ARTIC protocol originally developed in the context of Zika virus epidemics [5]) are first used to multiply short fragments of target genomes. These fragments are then sequenced in order to estimate composition and abundance of known virus variants.

In our work [1], we have addressed this problem by designing a probabilistic mixture model that characterizes likelihood of observing particular sequencing reads for a given composition of the sample. The standard optimization techniques are then used to estimate the most likely abundances. While this method works well on many real data sets, we have observed that in some cases, our method gives incorrect predictions due to unexpected observations. These include mutations that should be present in the sequenced reads at high frequencies according to the model, yet in some cases they are completely absent. To address this challenge, we have designed a wet lab experiment in which we have been able to reproduce these observation in a controlled scenario. Based on the data obtained from the experiment, we plan to desing more complex models that will address this issue, resulting in significantly more accurate prediction on real data.

References

1 A. Gafurov, A. Balaz, F. Amman, K. Borsova, V. Cabanova, B. Klempa, A. Bergthaler, T. Vinar, and B. Brejova. VirPool: model-based estimation of SARS-CoV-2 variant proportions in wastewater samples. *BMC Bioinformatics*, 23(1):551, 2022.

2 S. E. Hrudey and B. Conant. The devil is in the details: emerging insights on the relevance of wastewater surveillance for SARS-CoV-2 to public health. *J Water Health*, 20(1):246–270, 2022.

3 A. Mane, M. Faizrahnemoon, T. Vinar, B. Brejova, and C. Chauve. PlasBin-flow: a flow-based MILP algorithm for plasmid contigs binning. *Bioinformatics*, 39(39 Suppl 1):i288–i296, 2023.

4 L. Pu and R. Shamir. 3CAC: improving the classification of phages and plasmids in metagenomic assemblies using assembly graphs. *Bioinformatics*, 38(S2):ii56–ii61, 2022.

5 J. Quick, N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, D. R. Burton, L. L. Lewis-Ximenez, J. G. de Jesus, M. Giovanetti, S. C. Hill, A. Black, T. Bedford, M. W. Carroll, M. Nunes, L. C. Alcantara Jr., E. C. Sabino, S. A. Baylis, N. R. Faria, M. Loose, J. T. Simpson, O. G. Pybus, K. G. Andersen, and N. J. Loman. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc*, 12(6):1261–1276, 2017.

6 H. R. Safford, K. Shapiro, and H. N. Bischel. Opinion: Wastewater analysis can be a powerful public health tool-if it's done sensibly. *Proc Natl Acad Sci U S A*, 119(6), 2022.

7 J. Sielemann, K. Sielemann, B. Brejova, T. Vinar, and C. Chauve. plASgraph2: using graph neural networks to detect plasmid contigs from an assembly graph. *Front Microbiol*, 14:1267695, 2023.