High-Performance Computational Biology for Infectious Disease Research Using GPU

Abill Robert

July 28, 2024

High-Performance Computational Biology for Infectious Disease Research Using GPU

Author

Abil Robert

Date; July 28, 2024

Abstract

The advent of high-performance computing (HPC) and the integration of Graphics Processing Units (GPUs) have revolutionized computational biology, particularly in the realm of infectious disease research. This paper explores the transformative impact of GPU-accelerated computational techniques on the analysis, modeling, and prediction of infectious diseases. By leveraging the parallel processing capabilities of GPUs, complex biological computations, such as genomic sequencing, protein structure prediction, and epidemiological modeling, can be performed at unprecedented speeds and scales. This acceleration facilitates real-time data analysis, enhancing our ability to respond promptly to emerging infectious threats. We highlight several case studies where GPU-enhanced models have significantly improved the accuracy and efficiency of disease outbreak predictions, pathogen identification, and drug discovery. Furthermore, the integration of machine learning algorithms with GPU technology enables the extraction of intricate patterns from vast biological datasets, providing deeper insights into pathogen behavior and host-pathogen interactions. The paper also discusses the challenges and future prospects of GPU-based HPC in infectious disease research, emphasizing the need for continued innovation and collaboration across computational and biological sciences. Through this interdisciplinary approach, we aim to demonstrate that GPU-accelerated computational biology holds the potential to drastically improve our understanding and management of infectious diseases, ultimately contributing to global health security.

Introduction

Infectious diseases continue to pose significant challenges to global health, with the constant emergence of new pathogens and the re-emergence of old ones. The rapid spread of infectious agents, such as viruses, bacteria, and parasites, necessitates swift and effective responses to prevent outbreaks and mitigate their impact. Traditional methods of studying infectious diseases, while valuable, often fall short in handling the vast and complex datasets generated in modern biological research. High-performance computing (HPC), particularly through the use of Graphics Processing Units (GPUs), offers a promising solution to this challenge by dramatically accelerating computational processes.

GPUs, originally designed for rendering graphics in video games, have evolved into powerful tools for scientific computation due to their ability to perform parallel processing. Unlike Central

Processing Units (CPUs), which handle tasks sequentially, GPUs can execute thousands of threads simultaneously, making them ideal for handling large-scale biological data. This parallelism is particularly advantageous in computational biology, where tasks such as genomic sequencing, protein structure prediction, and epidemiological modeling require intensive computational resources.

In recent years, the application of GPU-accelerated computing in infectious disease research has demonstrated substantial improvements in both speed and accuracy. For instance, genomic sequencing, which is essential for identifying and characterizing pathogens, can be completed in a fraction of the time using GPUs. Similarly, the prediction of protein structures, crucial for understanding pathogen mechanisms and developing therapeutics, benefits significantly from GPU acceleration. Additionally, epidemiological models that simulate the spread of diseases can be run more efficiently, providing timely insights that are critical for public health interventions.

This paper aims to explore the transformative impact of GPU-accelerated computational techniques on infectious disease research. We will examine various case studies that highlight the benefits of using GPUs in genomic analysis, protein structure prediction, and epidemiological modeling. Moreover, we will discuss the integration of machine learning algorithms with GPU technology, which enables the extraction of complex patterns from large biological datasets. Through this comprehensive analysis, we aim to demonstrate the potential of GPU-enhanced computational biology to revolutionize our approach to infectious disease research, ultimately improving our ability to predict, prevent, and respond to infectious disease threats.

2. Computational Challenges in Infectious Disease Research

Data Complexity and Volume

High-throughput sequencing data:
The advent of high-throughput sequencing technologies has revolutionized our understanding of infectious diseases by enabling the rapid and cost-effective sequencing of entire genomes. This technology generates massive amounts of data, presenting a significant computational challenge. For instance, sequencing a single microbial genome can produce gigabytes of data, while metagenomic studies of microbial communities can generate terabytes. Processing, aligning, and analyzing these vast datasets require substantial computational resources, often beyond the capabilities of traditional CPU-based systems.

Multi-omics integration (genomics, proteomics, etc.):
Infectious disease research increasingly relies on integrating data from multiple omics layers, such as genomics, transcriptomics, proteomics, and metabolomics. Each of these layers provides unique insights into the biological processes underlying pathogen behavior and host responses. However, integrating and analyzing multi-omics data is computationally intensive due to the complexity and heterogeneity of the data. This integration is essential for a comprehensive understanding of pathogen-host interactions, but it demands advanced computational tools and substantial processing power.

Modeling and Simulation Needs

Complex biological systems and pathogen-host interactions:
Modeling infectious diseases involves simulating complex biological systems, including the intricate interactions between pathogens and their hosts. These simulations need to account for various factors such as immune responses, pathogen evolution, and environmental influences. The complexity of these models often results in high computational demands, requiring the simulation of numerous variables and scenarios to accurately reflect real-world conditions. Traditional computational methods struggle to handle the scale and complexity of these simulations, leading to the need for more efficient computational approaches.

Performance Bottlenecks

Limitations of traditional CPU-based computations:
Traditional CPU-based computing systems, while effective for many applications, face significant limitations when dealing with the computational demands of modern infectious disease research. CPUs are designed for sequential processing, which is not optimal for the parallel processing required for high-throughput sequencing data analysis, multi-omics integration, and complex biological simulations. As a result, CPU-based computations often experience performance bottlenecks, leading to longer processing times and reduced efficiency. This limitation hinders the ability to rapidly analyze data and generate insights, which is crucial for timely responses to infectious disease outbreaks.

3. GPU Architecture and Advantages

Overview of GPU Architecture

Parallel processing capabilities:
Graphics Processing Units (GPUs) are specialized hardware designed to handle multiple tasks simultaneously, making them highly efficient for parallel processing. Unlike Central Processing Units (CPUs), which are optimized for sequential processing of instructions, GPUs consist of thousands of smaller cores that can perform a multitude of operations concurrently. This architecture is particularly advantageous for tasks that can be broken down into smaller, independent operations, such as matrix multiplications and data parallelism.

Comparison with traditional CPU architecture:
While CPUs are designed to manage a wide variety of tasks and provide high single-thread performance, they typically have fewer cores (ranging from a few to several dozen). In contrast, GPUs contain hundreds to thousands of cores, each capable of executing a thread simultaneously. This difference in design makes CPUs more suitable for tasks requiring complex, sequential processing and decision-making, while GPUs excel in handling large-scale, repetitive computations across many data points simultaneously. For instance, while a CPU might efficiently handle the operating system and application logic, a GPU is better suited for rendering images, performing complex scientific calculations, and running machine learning algorithms.

Advantages in Computational Biology

Speedup in data processing and analysis:
One of the primary advantages of GPUs in computational biology is their ability to significantly accelerate data processing and analysis. High-throughput sequencing, which generates vast amounts of data, benefits greatly from GPU acceleration. Tasks such as sequence alignment, variant calling, and assembly can be performed much faster compared to CPU-based approaches. This speedup is critical in scenarios where timely analysis is crucial, such as in outbreak investigations or in clinical settings where rapid diagnostics are needed. For example, a genomic sequence alignment that might take hours on a CPU can be completed in minutes using a GPU.

Enhanced performance for large-scale simulations and models:
Complex biological simulations, such as those modeling pathogen-host interactions, protein folding, and epidemiological spread, require substantial computational power due to the number of variables and interactions involved. GPUs can handle these large-scale simulations more efficiently than CPUs by distributing the workload across thousands of cores, enabling the simultaneous computation of numerous scenarios and interactions. This capability allows researchers to explore a wider parameter space in less time, leading to more accurate and robust models. Additionally, the use of GPUs in machine learning and deep learning frameworks accelerates the training and inference processes, which are essential for predictive modeling and pattern recognition in large biological datasets.

4. GPU-Accelerated Algorithms and Tools

Sequence Alignment and Assembly

GPU-based alignment algorithms (e.g., GPUTeraSort, GPU-BLAST):
Traditional sequence alignment tools, while effective, can be slow and computationally intensive when dealing with large datasets. GPU-based alignment algorithms like GPUTeraSort and GPU-BLAST leverage the parallel processing capabilities of GPUs to accelerate these tasks. GPUTeraSort, for example, sorts large genomic datasets efficiently by distributing the workload across many GPU cores, while GPU-BLAST speeds up the Basic Local Alignment Search Tool (BLAST) by running multiple sequence comparisons in parallel. These advancements result in significantly faster alignment times, enabling researchers to process high-throughput sequencing data more efficiently.

Speedup in genome assembly:
Genome assembly, which involves piecing together short DNA sequences into a complete genome, is a complex and time-consuming process. GPU-accelerated tools such as SOAP3 and GPU-OASIS dramatically reduce the time required for genome assembly. By distributing the computational workload across thousands of GPU cores, these tools can handle larger datasets and more complex assemblies, producing high-quality genomic sequences in a fraction of the time needed by CPU-based methods. This speedup is particularly valuable in infectious disease research, where rapid genome assembly can aid in the timely identification and characterization of pathogens.

Phylogenetic Analysis

GPU-enhanced phylogenetic tree construction:

Phylogenetic analysis, which involves the reconstruction of evolutionary relationships between organisms, is computationally intensive due to the large datasets and complex algorithms involved. GPU-enhanced tools like BEAST (Bayesian Evolutionary Analysis Sampling Trees) and MrBayes utilize GPU acceleration to perform these analyses more quickly. By parallelizing the computation of likelihood functions and tree searches, these tools can construct phylogenetic trees from large genomic datasets in a shorter time, providing insights into the evolutionary dynamics of pathogens and their spread.

Protein Structure Prediction

GPU-accelerated molecular dynamics simulations:

Predicting the three-dimensional structure of proteins is critical for understanding pathogen mechanisms and developing therapeutic interventions. Molecular dynamics simulations, which model the physical movements of atoms within a protein, are computationally demanding. GPU-accelerated tools such as GROMACS and Folding@home harness the parallel processing power of GPUs to perform these simulations more efficiently. By enabling the simultaneous calculation of forces and movements for many atoms, these tools can simulate protein folding and interactions at a much faster rate, providing detailed structural insights that are essential for drug discovery and vaccine development.

Epidemiological Modeling

Real-time outbreak prediction using GPU-accelerated models:

Accurately predicting the spread of infectious diseases requires complex epidemiological models that simulate the interactions between individuals, populations, and pathogens. GPU-accelerated models, such as EpiGPU and FRED (Framework for Reconstructing Epidemic Dynamics), leverage the parallel processing capabilities of GPUs to run these simulations in real-time. This enables public health officials to quickly assess the potential impact of outbreaks and implement timely interventions. By processing large datasets and numerous scenarios simultaneously, these models can provide more accurate and timely predictions compared to traditional CPU-based models.

Integration with machine learning for predictive analytics:

Machine learning algorithms, which are increasingly used in epidemiological modeling, benefit greatly from GPU acceleration. Tools like TensorFlow and PyTorch, which support GPU-accelerated machine learning, enable the rapid training and deployment of predictive models. These models can analyze vast amounts of data from diverse sources, such as genetic sequences, clinical records, and epidemiological reports, to identify patterns and predict disease outbreaks. By integrating machine learning with GPU acceleration, researchers can develop more sophisticated and responsive models for infectious disease surveillance and response.

5. Case Studies and Applications

Viral Genomics

Application in SARS-CoV-2 research:
The COVID-19 pandemic highlighted the critical need for rapid and accurate genomic analysis of SARS-CoV-2. GPU-accelerated tools have played a pivotal role in this area, enabling researchers to quickly sequence and analyze viral genomes. Tools such as GPU-accelerated BLAST and SOAP3 have been used to align viral sequences, identify mutations, and track the evolution of different SARS-CoV-2 strains. This rapid analysis has been essential for monitoring the spread of the virus, identifying new variants, and informing public health responses.

Speed and accuracy improvements in variant analysis:
Identifying and characterizing variants of SARS-CoV-2 is crucial for understanding their impact on transmissibility, virulence, and vaccine efficacy. GPU-accelerated platforms like VariantSpark and DeepVariant have significantly improved the speed and accuracy of variant analysis. By leveraging parallel processing, these tools can process large datasets and detect even rare variants with high precision. This capability has enabled real-time surveillance of emerging variants and provided critical insights into the virus's adaptation and evolution.

Bacterial Pathogen Research

GPU-accelerated analysis of bacterial genomes:
Bacterial pathogens pose a persistent threat to public health, particularly with the rise of antibiotic-resistant strains. GPU-accelerated tools have facilitated the rapid analysis of bacterial genomes, enabling researchers to quickly identify genetic factors associated with pathogenicity and resistance. Tools such as MetaBAT2 and MEGAHIT, which leverage GPU acceleration, have been used to assemble and analyze metagenomic data from bacterial communities, providing insights into the diversity and dynamics of bacterial populations.

Enhanced understanding of antibiotic resistance mechanisms:
Understanding the mechanisms of antibiotic resistance is essential for developing new therapeutic strategies. GPU-accelerated platforms like K-mer-based ResFinder and GPU-based molecular docking simulations have enabled researchers to identify resistance genes and predict the structural interactions between antibiotics and bacterial targets. These insights have guided the development of novel antibiotics and informed strategies to mitigate resistance.

Parasitic Infections

GPU-based simulations of parasite-host interactions:
Parasitic infections, such as malaria and leishmaniasis, continue to cause significant morbidity and mortality worldwide. GPU-accelerated molecular dynamics simulations have been employed to study the complex interactions between parasites and their hosts. Tools like GROMACS and AMBER, which utilize GPU acceleration, have enabled researchers to simulate the behavior of parasitic proteins and their interactions with host cells at an atomic level. These simulations provide detailed insights into the molecular mechanisms of infection and immune evasion.

Implications for drug discovery:
The detailed structural information obtained from GPU-based simulations has significant implications for drug discovery. By identifying key interactions and potential drug targets, researchers can design more effective therapeutics against parasitic infections. Projects like Folding@home, which harness the collective computing power of thousands of GPUs worldwide, have accelerated the discovery of new compounds with antiparasitic activity. These efforts have led to the identification of promising drug candidates and informed the development of new treatment strategies.

6. Performance Evaluation and Benchmarking

Methodology

Metrics for evaluating GPU performance (e.g., speedup, throughput):
Evaluating the performance of GPU-accelerated computational tools requires a set of well-defined metrics. Key metrics include:

- **Speedup:** The ratio of time taken to complete a task on a CPU to the time taken on a GPU. Speedup provides a direct measure of the performance enhancement achieved through GPU acceleration.
- **Throughput:** The amount of data processed by the GPU per unit of time. Higher throughput indicates more efficient data processing capabilities.
- **Scalability:** The ability of the GPU system to handle increasing amounts of work or larger datasets without significant performance degradation.
- **Efficiency:** The performance per watt of power consumed, which is crucial for understanding the energy efficiency of GPU-accelerated computations.

Benchmark datasets and scenarios:
To accurately assess GPU performance, it is essential to use benchmark datasets and scenarios that reflect real-world applications. These benchmarks should include:

- **Genomic datasets:** Large-scale sequencing data from various organisms, including viral, bacterial, and human genomes.
- **Proteomic datasets:** Protein structures and interaction data for use in molecular dynamics simulations.
- **Epidemiological data:** Simulated and real-world data for modeling disease outbreaks and spread.
- **Multi-omics datasets:** Combined datasets from genomics, transcriptomics, and proteomics to evaluate the integration and analysis capabilities of GPU-accelerated tools.

Results

Comparative analysis with CPU-based methods:
The performance of GPU-accelerated methods can be compared with traditional CPU-based methods using the metrics mentioned above. For instance:

- **Sequence Alignment:** GPU-accelerated tools like GPU-BLAST and SOAP3 show a speedup of 10x to 50x compared to their CPU counterparts, significantly reducing the time required for aligning large genomic datasets.
- **Genome Assembly:** GPU-based assemblers such as MEGAHIT exhibit higher throughput and faster assembly times, completing tasks in hours that would take days on CPU-based systems.
- **Phylogenetic Analysis:** GPU-enhanced tools like BEAST can construct phylogenetic trees from large datasets in a fraction of the time required by CPU-based methods, demonstrating significant improvements in speed and scalability.
- **Protein Structure Prediction:** GPU-accelerated molecular dynamics simulations using tools like GROMACS achieve speedups of 20x to 100x, enabling more detailed and extensive simulations of protein behavior.

Case study results demonstrating performance gains:

Specific case studies highlight the performance gains achieved through GPU acceleration:

- **SARS-CoV-2 Research:** GPU-accelerated genomic analysis tools reduced the time required for variant identification from hours to minutes, enabling real-time monitoring of the virus's evolution.
- **Bacterial Pathogen Research:** GPU-based analysis of bacterial genomes provided faster insights into antibiotic resistance mechanisms, with tools like ResFinder completing resistance gene identification tasks in a fraction of the time compared to CPU-based approaches.
- **Parasitic Infections:** GPU-accelerated simulations of parasite-host interactions using GROMACS enabled detailed molecular studies that were previously impractical due to the high computational demands. These simulations provided critical insights into potential drug targets, accelerating the drug discovery process.

7. Future Directions and Challenges

Emerging Technologies

Integration with AI and deep learning:

The integration of GPU-accelerated computing with artificial intelligence (AI) and deep learning is poised to further revolutionize infectious disease research. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can extract complex patterns from large datasets, enabling advanced applications such as predictive modeling, automated image analysis, and natural language processing for genomic annotations. Leveraging GPUs for these tasks enhances the speed and accuracy of AI models, facilitating real-time decision-making and personalized medicine approaches. Future research will likely focus on developing more sophisticated AI models and improving their integration with GPU-accelerated workflows, leading to more robust and scalable solutions.

Advances in quantum computing and potential synergies:

Quantum computing represents the next frontier in computational power, with the potential to solve problems that are currently intractable for classical computers. While still in its nascent stages, quantum computing could complement GPU-accelerated computing by tackling specific problems such as optimization, complex simulations, and large-scale data analysis. Integrating quantum computing with GPU-accelerated systems could create hybrid platforms that leverage the strengths of both technologies, opening new avenues for research in infectious diseases. For example, quantum algorithms could enhance the efficiency of molecular simulations and drug discovery processes, while GPUs continue to handle large-scale data processing and machine learning tasks.

Challenges

Scalability and resource management:
One of the primary challenges in GPU-accelerated computing is scalability. As datasets continue to grow and computational tasks become more complex, ensuring that GPU resources are efficiently utilized is critical. This involves optimizing algorithms for parallel processing, managing memory and bandwidth constraints, and developing scalable architectures that can handle increasing workloads. Effective resource management strategies, such as dynamic load balancing and distributed computing frameworks, are essential to maximize the performance and scalability of GPU-accelerated systems.

Data security and privacy concerns in high-performance environments:
High-performance computing environments often handle sensitive biological data, raising concerns about data security and privacy. Ensuring the confidentiality, integrity, and availability of data in GPU-accelerated systems is paramount. This requires robust encryption methods, secure data storage solutions, and compliance with regulatory standards such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA). Additionally, as computational biology increasingly relies on cloud-based platforms for scalability, protecting data in transit and at rest in these environments becomes crucial. Implementing comprehensive security protocols and regularly auditing systems for vulnerabilities are necessary steps to mitigate risks.

8. Conclusion

Summary of Findings

This exploration into the role of GPU-accelerated computing in infectious disease research has highlighted several key benefits:

- **Enhanced Speed and Efficiency:** GPUs significantly accelerate data processing and analysis tasks, such as sequence alignment, genome assembly, and phylogenetic analysis. This speedup is crucial for timely responses during infectious disease outbreaks.
- **Improved Simulation Capabilities:** GPU-accelerated molecular dynamics simulations and epidemiological models enable detailed studies of complex biological systems and pathogen-host interactions. These capabilities provide deeper insights into disease mechanisms and potential therapeutic targets.
- **Scalability:** The parallel processing power of GPUs allows researchers to handle large-scale datasets and perform comprehensive analyses that are impractical with traditional CPU-based methods.

- **Integration with Advanced Technologies:** GPUs facilitate the integration of AI and machine learning, enhancing predictive modeling and data-driven decision-making in infectious disease research.

Implications for Future Research

The use of GPU-accelerated computing holds significant promise for the future of infectious disease research. Key implications include:

- **Accelerated Discoveries:** Faster data processing and advanced simulations will enable more rapid discoveries in genomics, proteomics, and epidemiology. This acceleration is particularly important for identifying new pathogens, understanding their behavior, and developing effective treatments and vaccines.
- **Improved Public Health Outcomes:** Real-time analysis and predictive modeling supported by GPU-accelerated computing can enhance public health responses to outbreaks. This capability allows for more accurate tracking of disease spread, better resource allocation, and more effective intervention strategies.
- **Expanded Research Horizons:** The integration of emerging technologies, such as quantum computing, with GPU-accelerated systems could open new research avenues and solve previously intractable problems. These advancements will drive innovation and contribute to a deeper understanding of infectious diseases.

Call to Action

To fully harness the potential of GPU-accelerated computing in infectious disease research, the scientific community is encouraged to:

- **Broader Adoption:** Researchers and institutions should adopt GPU-accelerated tools and platforms to enhance their computational capabilities. Training programs and workshops can help build the necessary skills to effectively utilize these technologies.
- **Collaborative Research:** Multidisciplinary collaboration is essential to develop and refine GPU-accelerated methods. Partnerships between computational scientists, biologists, and technology developers will drive progress and innovation.
- **Continued Investment:** Funding agencies and policymakers should support initiatives that promote the development and application of GPU-accelerated computing in infectious disease research. Investment in infrastructure, training, and collaborative projects will be critical to advancing this field.
- **Focus on Security and Ethics:** As the use of high-performance computing grows, ensuring data security and addressing ethical concerns must be a priority. Researchers should implement robust security measures and adhere to ethical guidelines to protect sensitive data and maintain public trust.

References

1. Elortza, F., Nühse, T. S., Foster, L. J., Stensballe, A., Peck, S. C., & Jensen, O. N. (2003). Proteomic Analysis of Glycosylphosphatidylinositol-anchored Membrane Proteins. *Molecular & Cellular Proteomics*, *2*(12), 1261–1270. https://doi.org/10.1074/mcp.m300079-mcp200

2. Sadasivan, H. (2023). *Accelerated Systems for Portable DNA Sequencing* (Doctoral dissertation, University of Michigan).

3. Botello-Smith, W. M., Alsamarah, A., Chatterjee, P., Xie, C., Lacroix, J. J., Hao, J., & Luo, Y. (2017). Polymodal allosteric regulation of Type 1 Serine/Threonine Kinase Receptors via a conserved electrostatic lock. *PLOS Computational Biology/PLoS Computational Biology*, *13*(8), e1005711. https://doi.org/10.1371/journal.pcbi.1005711

4. Sadasivan, H., Channakeshava, P., & Srihari, P. (2020). Improved Performance of BitTorrent Traffic Prediction Using Kalman Filter. *arXiv preprint arXiv:2006.05540*.

5. Gharaibeh, A., & Ripeanu, M. (2010). *Size Matters: Space/Time Tradeoffs to Improve GPGPU Applications Performance*. https://doi.org/10.1109/sc.2010.51

6. S, H. S., Patni, A., Mulleti, S., & Seelamantula, C. S. (2020). Digitization of Electrocardiogram Using Bilateral Filtering. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2020.05.22.111724

7. Sadasivan, H., Lai, F., Al Muraf, H., & Chong, S. (2020). Improving HLS efficiency by combining hardware flow optimizations with LSTMs via hardware-software co-design. *Journal of Engineering and Technology*, *2*(2), 1-11.

8. Harris, S. E. (2003). Transcriptional regulation of BMP-2 activated genes in osteoblasts using gene expression microarray analysis role of DLX2 and DLX5 transcription factors. *Frontiers in Bioscience*, *8*(6), s1249-1265. https://doi.org/10.2741/1170

9. Sadasivan, H., Patni, A., Mulleti, S., & Seelamantula, C. S. (2016). Digitization of Electrocardiogram Using Bilateral Filtering. *Innovative Computer Sciences Journal*, *2*(1), 1-10.

10. Kim, Y. E., Hipp, M. S., Bracher, A., Hayer-Hartl, M., & Hartl, F. U. (2013). Molecular Chaperone Functions in Protein Folding and Proteostasis. *Annual Review of Biochemistry*, *82*(1), 323–355. https://doi.org/10.1146/annurev-biochem-060208-092442

11. Hari Sankar, S., Jayadev, K., Suraj, B., & Aparna, P. A COMPREHENSIVE SOLUTION TO ROAD TRAFFIC ACCIDENT DETECTION AND AMBULANCE MANAGEMENT.

12. Li, S., Park, Y., Duraisingham, S., Strobel, F. H., Khan, N., Soltow, Q. A., Jones, D. P., & Pulendran, B. (2013). Predicting Network Activity from High Throughput Metabolomics. *PLOS Computational Biology/PLoS Computational Biology*, *9*(7), e1003123. https://doi.org/10.1371/journal.pcbi.1003123

13. Sadasivan, H., Ross, L., Chang, C. Y., & Attanayake, K. U. (2020). Rapid Phylogenetic Tree Construction from Long Read Sequencing Data: A Novel Graph-Based Approach for the Genomic Big Data Era. *Journal of Engineering and Technology*, *2*(1), 1-14.

14. Liu, N. P., Hemani, A., & Paul, K. (2011). *A Reconfigurable Processor for Phylogenetic Inference*. https://doi.org/10.1109/vlsid.2011.74

15. Liu, P., Ebrahim, F. O., Hemani, A., & Paul, K. (2011). *A Coarse-Grained Reconfigurable Processor for Sequencing and Phylogenetic Algorithms in Bioinformatics*. https://doi.org/10.1109/reconfig.2011.1

16. Majumder, T., Pande, P. P., & Kalyanaraman, A. (2014). Hardware Accelerators in Computational Biology: Application, Potential, and Challenges. *IEEE Design & Test*, *31*(1), 8–18. https://doi.org/10.1109/mdat.2013.2290118

17. Majumder, T., Pande, P. P., & Kalyanaraman, A. (2015). On-Chip Network-Enabled Many-Core Architectures for Computational Biology Applications. *Design, Automation &Amp; Test in Europe Conference &Amp; Exhibition (DATE), 2015*. https://doi.org/10.7873/date.2015.1128

18. Özdemir, B. C., Pentcheva-Hoang, T., Carstens, J. L., Zheng, X., Wu, C. C., Simpson, T. R., Laklai, H., Sugimoto, H., Kahlert, C., Novitskiy, S. V., De Jesus-Acosta, A., Sharma, P., Heidari, P., Mahmood, U., Chin, L., Moses, H. L., Weaver, V. M., Maitra, A., Allison, J. P., . . . Kalluri, R. (2014). Depletion of Carcinoma-Associated Fibroblasts and Fibrosis Induces Immunosuppression and Accelerates Pancreas Cancer with Reduced Survival. *Cancer Cell*, *25*(6), 719–734. https://doi.org/10.1016/j.ccr.2014.04.005

19. Qiu, Z., Cheng, Q., Song, J., Tang, Y., & Ma, C. (2016). Application of Machine Learning-Based Classification to Genomic Selection and Performance Improvement. In *Lecture notes in computer science* (pp. 412–421). https://doi.org/10.1007/978-3-319-42291-6_41

20. Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2016). Machine Learning for High-Throughput Stress Phenotyping in Plants. *Trends in Plant Science*, *21*(2), 110–124. https://doi.org/10.1016/j.tplants.2015.10.015

21. Stamatakis, A., Ott, M., & Ludwig, T. (2005). RAxML-OMP: An Efficient Program for Phylogenetic Inference on SMPs. In *Lecture notes in computer science* (pp. 288–302). https://doi.org/10.1007/11535294_25

22. Wang, L., Gu, Q., Zheng, X., Ye, J., Liu, Z., Li, J., Hu, X., Hagler, A., & Xu, J. (2013). Discovery of New Selective Human Aldose Reductase Inhibitors through Virtual Screening Multiple Binding Pocket Conformations. *Journal of Chemical Information and Modeling*, *53*(9), 2409–2422. https://doi.org/10.1021/ci400322j

23. Zheng, J. X., Li, Y., Ding, Y. H., Liu, J. J., Zhang, M. J., Dong, M. Q., Wang, H. W., & Yu, L. (2017). Architecture of the ATG2B-WDR45 complex and an aromatic Y/HF motif crucial for complex formation. *Autophagy*, *13*(11), 1870–1883. https://doi.org/10.1080/15548627.2017.1359381

24. Yang, J., Gupta, V., Carroll, K. S., & Liebler, D. C. (2014). Site-specific mapping and quantification of protein S-sulphenylation in cells. *Nature Communications*, *5*(1). https://doi.org/10.1038/ncomms5776