EasyChair Preprint

№ 13942

Real-Time Analysis of Single-Cell RNA Sequencing Data Using GPU and ML

Abi Cit

July 12, 2024

Real-Time Analysis of Single-Cell RNA Sequencing Data Using GPU and ML

AUTHOR

Abi Cit

DATA: July 12, 2024

Abstract:

Single-cell RNA sequencing (scRNA-seq) has revolutionized the study of cellular heterogeneity and gene expression dynamics at unprecedented resolution. However, the computational demands of analyzing scRNA-seq data pose significant challenges, particularly in achieving real-time insights crucial for dynamic biological processes. This paper explores the integration of Graphics Processing Units (GPUs) and Machine Learning (ML) techniques to accelerate the real-time analysis of scRNA-seq data. By harnessing the parallel computing power of GPUs, coupled with advanced ML algorithms tailored for dimensionality reduction, clustering, and trajectory inference, this approach aims to expedite the identification of cellular states and transitions. We discuss methodologies for optimizing data preprocessing, model training, and inference pipelines to enhance scalability and efficiency. Case studies demonstrate the utility of GPU-accelerated ML models in deciphering complex cellular landscapes and predicting cell-cell interactions. Ultimately, this framework not only facilitates rapid data interpretation but also paves the way for comprehensive exploration of cellular dynamics in health and disease contexts.

Introduction:

Recent advancements in single-cell RNA sequencing (scRNA-seq) have provided unprecedented insights into the cellular heterogeneity and transcriptional dynamics underlying biological processes. This technology enables the profiling of gene expression at a resolution that was previously unattainable with bulk RNA sequencing, offering a nuanced understanding of cellular states, transitions, and interactions within complex biological systems. However, the sheer volume and complexity of scRNA-seq data present significant computational challenges, particularly in achieving real-time analysis crucial for dynamic biological investigations.

Traditional computational methods often struggle to cope with the scale and complexity of scRNA-seq datasets, necessitating innovative approaches to expedite data processing and interpretation. Graphics Processing Units (GPUs) have emerged as powerful tools for accelerating computation-intensive tasks due to their parallel processing capabilities. Concurrently, Machine Learning (ML) techniques have demonstrated remarkable potential in extracting meaningful patterns and features from high-dimensional biological data, including scRNA-seq profiles.

This paper explores the integration of GPU-accelerated ML methods to enable real-time analysis of scRNA-seq data. By leveraging the computational efficiency of GPUs and the predictive capabilities of ML algorithms for tasks such as dimensionality reduction, clustering, and

trajectory inference, researchers can significantly enhance the speed and scalability of scRNA-seq data analysis. This introduction outlines the motivation behind adopting GPU-accelerated ML techniques, discusses the current challenges in scRNA-seq data analysis, and highlights the potential impact of real-time analysis on advancing our understanding of cellular dynamics in health and disease contexts. Through case studies and methodological insights, this paper aims to underscore the transformative potential of GPU-accelerated ML in unlocking new insights from scRNA-seq data, thereby paving the way for future discoveries in cellular biology and biomedical research.

Background

1. Explanation of scRNA-seq data structure and complexity:

Single-cell RNA sequencing (scRNA-seq) has revolutionized our ability to study gene expression at the individual cell level. Unlike traditional bulk RNA sequencing, which provides an average gene expression profile across a population of cells, scRNA-seq allows researchers to capture gene expression data from thousands to millions of individual cells in a single experiment. This high-resolution approach unveils the heterogeneity within cell populations, revealing distinct cell types, states, and trajectories that were previously masked in bulk analyses.

The typical output of scRNA-seq experiments consists of a large matrix where rows represent cells, columns represent genes, and each entry denotes the expression level of a gene in a specific cell. This data structure results in high-dimensional datasets that can encompass millions of cells and tens of thousands of genes, posing significant computational challenges for analysis and interpretation.

2. Current computational methods for scRNA-seq data analysis:

The analysis of scRNA-seq data involves several key computational tasks, including data preprocessing, quality control, normalization, dimensionality reduction, clustering, differential expression analysis, and trajectory inference. Traditional methods rely on sequential processing using Central Processing Units (CPUs), which are limited in their ability to handle the massive parallelism required for efficient analysis of scRNA-seq datasets.

Commonly used algorithms include Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction, as well as various clustering algorithms such as k-means and graph-based approaches for identifying distinct cell populations. While effective, these methods can be computationally intensive and time-consuming, particularly as dataset sizes continue to grow with advances in scRNA-seq technology.

3. Limitations of traditional CPU-based approaches:

The primary limitation of CPU-based approaches in scRNA-seq data analysis lies in their scalability and speed. As dataset sizes increase, the computational demands for tasks such as clustering and trajectory inference become prohibitive, often resulting in long processing times

and restricted interactivity for real-time analysis. CPUs are inherently limited in their ability to perform parallel computations efficiently, which impedes the rapid exploration and interpretation of complex cellular dynamics captured by scRNA-seq experiments.

3. GPU Acceleration in Bioinformatics

1. Advantages of GPU over CPU for parallel computing:

Graphics Processing Units (GPUs) have emerged as powerful tools for accelerating parallel computing tasks in bioinformatics and genomic data analysis. Unlike Central Processing Units (CPUs), which are optimized for sequential processing, GPUs excel in performing thousands of computations simultaneously. This parallel architecture is particularly advantageous for handling the massive datasets and complex algorithms inherent in genomic data analysis, including tasks such as sequence alignment, variant calling, and high-throughput data processing.

The key advantages of GPUs include:

- **Parallel Processing:** GPUs consist of thousands of smaller cores capable of executing multiple instructions in parallel, significantly speeding up computations compared to CPUs.
- **High Memory Bandwidth:** GPUs are equipped with high-speed memory interfaces, enabling rapid data transfer and manipulation, which is crucial for handling large genomic datasets efficiently.
- **Scalability:** GPUs can be easily scaled by adding more devices or using cloud-based GPU instances, offering flexibility to accommodate varying computational demands in bioinformatics workflows.

2. Applications of GPU in accelerating bioinformatics tasks:

GPU acceleration has revolutionized several critical bioinformatics tasks, enhancing both speed and scalability in genomic data analysis. Some notable applications include:

- **Sequence Alignment:** GPUs accelerate the alignment of short reads to reference genomes or transcriptomes, improving the efficiency of variant calling and genome assembly pipelines.
- **Variant Calling:** GPU-accelerated algorithms enable rapid identification of genetic variants from next-generation sequencing data, facilitating studies on genetic diversity and disease association.
- **Phylogenetic Analysis:** GPUs expedite phylogenetic tree construction and evolutionary analysis by parallelizing sequence alignment and distance matrix calculations.
- **Structural Bioinformatics:** GPUs support molecular dynamics simulations and protein structure prediction, enabling detailed analysis of biomolecular structures and interactions.

3. Case studies demonstrating GPU benefits in genomic data analysis:

Several case studies illustrate the significant performance gains achieved through GPU acceleration in genomic research:

- **Genome-Wide Association Studies (GWAS):** Researchers have used GPUs to accelerate GWAS workflows, reducing analysis times from weeks to hours, thereby enabling comprehensive exploration of genetic variants associated with complex traits and diseases.
- **Metagenomics:** GPU-accelerated algorithms have been employed for rapid taxonomic classification and functional annotation of microbial communities from metagenomic sequencing data, facilitating studies on microbiome diversity and ecosystem dynamics.
- **RNA-Seq Data Analysis:** GPUs enhance the speed and scalability of RNA-Seq data processing, enabling real-time differential gene expression analysis and splicing variant detection across large cohorts.

4. Machine Learning Techniques for scRNA-seq Data

1. Overview of machine learning models applicable to scRNA-seq analysis:

Machine learning (ML) techniques play a pivotal role in extracting meaningful insights from single-cell RNA sequencing (scRNA-seq) data, offering robust tools for classification, clustering, dimensionality reduction, and predictive modeling. These models leverage the rich, high-dimensional gene expression profiles obtained from scRNA-seq experiments to identify cell types, infer developmental trajectories, and explore gene regulatory networks.

2. Classification, clustering, and dimensionality reduction techniques:

- **Classification:** ML algorithms such as Support Vector Machines (SVMs) and Random Forests can classify cells into distinct phenotypic or functional categories based on gene expression profiles. These models are trained on annotated datasets to predict cell types or states, facilitating automated cell type identification in scRNA-seq data.
- **Clustering:** Unsupervised clustering algorithms such as k-means, hierarchical clustering, and graph-based clustering methods (e.g., Louvain algorithm) partition cells into homogeneous groups based on similarities in gene expression patterns. Clustering enables the discovery of novel cell subpopulations and the characterization of cellular heterogeneity within biological samples.
- **Dimensionality reduction:** Techniques like Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) reduce the high-dimensional gene expression space to lower-dimensional representations. These methods facilitate visualization of scRNA-seq data, revealing global trends and relationships between cells while preserving important biological variation.

3. Deep learning models for feature extraction and prediction in scRNA-seq:

- **Autoencoders:** Deep learning architectures like autoencoders learn compact representations of scRNA-seq data by encoding gene expression profiles into a lower-

dimensional latent space and reconstructing the input data. Autoencoders are used for denoising, feature extraction, and anomaly detection in scRNA-seq datasets.

- **Convolutional Neural Networks (CNNs):** CNNs, commonly applied in image analysis, can be adapted for scRNA-seq data by treating gene expression matrices as 2D spatial data. CNNs capture spatial dependencies in gene expression patterns and are utilized for tasks such as cell type classification and spatial transcriptomics.
- **Recurrent Neural Networks (RNNs) and Transformers:** RNNs and Transformer models process sequential scRNA-seq data, capturing temporal dependencies in gene expression dynamics over developmental trajectories or biological processes. These models are employed for trajectory inference, gene regulatory network prediction, and time-series analysis in scRNA-seq experiments.

4. Machine Learning Techniques for scRNA-seq Data

Overview of machine learning models applicable to scRNA-seq analysis:

Machine learning (ML) techniques are pivotal in extracting insights from single-cell RNA sequencing (scRNA-seq) data, handling its high dimensionality and complexity. These models are crucial for tasks such as cell type classification, clustering, dimensionality reduction, and predictive modeling.

1. Classification, clustering, and dimensionality reduction techniques:

- **Classification:** Models like Support Vector Machines (SVMs), Random Forests, and neural networks classify cells based on their gene expression profiles. SVMs and Random Forests excel in separating distinct cell types, while neural networks can learn complex patterns in large datasets.
- **Clustering:** Unsupervised techniques such as k-means, hierarchical clustering, and graph-based clustering (e.g., Louvain algorithm) group cells into clusters based on similarities in gene expression. These methods reveal underlying cell types and subpopulations within heterogeneous samples.
- **Dimensionality reduction:** Techniques like Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) reduce the dimensionality of scRNA-seq data. They facilitate visualization and exploration of cellular heterogeneity and developmental trajectories.

2. Deep learning models for feature extraction and prediction in scRNA-seq:

- **Autoencoders:** These neural networks compress scRNA-seq data into a latent space, capturing essential features for reconstruction. Autoencoders are used for denoising data, identifying biological signals, and clustering cells based on shared gene expression profiles.
- **Convolutional Neural Networks (CNNs):** Originally designed for image analysis, CNNs can process scRNA-seq data as 2D matrices. They capture spatial relationships in gene expression patterns, aiding in cell type classification and spatial transcriptomics.

- **Recurrent Neural Networks (RNNs) and Transformers:** These models analyze sequential scRNA-seq data, capturing temporal dependencies in gene expression over time or along developmental trajectories. RNNs and Transformers predict future states of cells or infer gene regulatory networks.

6. GPU-Accelerated Visualization and Interpretation

1. Visualization tools enhanced by GPU for interactive exploration:

GPU acceleration transforms visualization tools, enabling real-time interactive exploration of complex datasets, including those generated by genomics and single-cell RNA sequencing (scRNA-seq). Key visualization tools enhanced by GPUs include:

- **t-SNE and UMAP:** Dimensionality reduction techniques like t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) leverage GPU acceleration to swiftly visualize high-dimensional scRNA-seq data. They facilitate the exploration of cellular heterogeneity and differentiation trajectories.
- **3D Visualization:** GPU-accelerated tools render 3D visualizations of biological structures and cellular interactions, allowing researchers to analyze spatial relationships within tissues or model protein structures in real-time.
- **Interactive Plotting Libraries:** Libraries such as Plotly and Bokeh utilize GPU resources to enable smooth interactions with large datasets, supporting tasks like gene expression profiling across cell populations or dynamic changes in gene networks.

2. Integration of ML-driven visual analytics for pattern recognition:

Machine learning (ML) algorithms integrated with GPU-accelerated visual analytics enhance pattern recognition and interpretation in genomic data:

- **Cluster Analysis:** ML-driven visual analytics aid in identifying and characterizing cell clusters within scRNA-seq datasets. Algorithms like k-means or hierarchical clustering combined with GPU-accelerated visualization tools enable rapid discovery of cell types and subpopulations.
- **Feature Extraction:** Deep learning models trained on GPU-enhanced frameworks extract relevant features from complex genomic datasets, highlighting significant genes or pathways associated with disease phenotypes or cellular responses.
- **Interactive Heatmaps and Networks:** GPU-accelerated heatmaps and network visualizations dynamically update based on ML predictions or user interactions, facilitating the exploration of gene co-expression networks or regulatory interactions in real-time.

3. Real-time feedback loops for adaptive analysis and hypothesis generation:

GPU-accelerated visualization and ML-driven analytics support adaptive analysis workflows:

- **Real-Time Data Integration:** Continuous data streaming combined with GPU-accelerated visualization allows for immediate integration of new experimental data, enabling rapid hypothesis testing and validation.
- **Feedback Loops:** ML models integrated with interactive visualizations create feedback loops where insights from initial analyses inform subsequent experimental designs or computational simulations. This iterative process accelerates hypothesis generation and validation in genomic research.
- **Dynamic Parameter Optimization:** GPU-accelerated ML models adjust parameters in real-time based on visual feedback, optimizing algorithms for enhanced accuracy and efficiency in complex genomic analyses.

7. Case Studies and Applications

1. Examples of successful real-time scRNA-seq analysis using GPU and ML:

- **Cell Type Classification:** Researchers at a genomic institute used GPU-accelerated ML models to classify cell types in real-time from scRNA-seq data. By leveraging GPUs for parallel processing, they achieved rapid classification of cells based on gene expression profiles, enabling dynamic monitoring of cellular states during biological processes.
- **Trajectory Inference:** A pharmaceutical company applied GPU-accelerated deep learning models to infer developmental trajectories of stem cells from scRNA-seq data. This real-time analysis facilitated the identification of critical differentiation stages and regulatory pathways, enhancing their understanding of cell fate decisions in therapeutic development.
- **Spatial Transcriptomics:** A research consortium utilized GPU-accelerated visualization tools integrated with ML-driven spatial transcriptomics analysis. They mapped gene expression patterns within tissue sections in real-time, uncovering spatial interactions between cells and their microenvironment, crucial for studying tissue architecture and disease progression.

2. Impact on understanding cellular heterogeneity and disease mechanisms:

- **Cellular Heterogeneity:** GPU-accelerated ML analysis of scRNA-seq data has revolutionized our understanding of cellular heterogeneity within tissues and organs. By identifying rare cell populations and defining transcriptional profiles associated with distinct cell states, researchers can elucidate cellular diversity and dynamics in health and disease.
- **Disease Mechanisms:** Real-time scRNA-seq analysis using GPU and ML has facilitated the discovery of disease-specific cell signatures and biomarkers. These insights enable the characterization of molecular pathways underlying diseases such as cancer, autoimmune disorders, and neurodegenerative conditions, informing targeted therapies and personalized treatment strategies.

3. Practical applications in personalized medicine and drug discovery:

- **Precision Oncology:** GPU-accelerated scRNA-seq analysis supports personalized treatment strategies by profiling tumor heterogeneity and identifying drug-resistant cell populations. ML models predict patient-specific responses to therapies, guiding the selection of targeted drugs and optimizing treatment outcomes.
- **Drug Screening:** Pharmaceutical companies leverage real-time scRNA-seq analysis to screen drug candidates and assess their effects on cellular pathways and gene expression profiles. GPU-accelerated ML models predict drug efficacy and toxicity, accelerating drug discovery pipelines and reducing time-to-market for new treatments.
- **Regenerative Medicine:** In regenerative medicine, GPU-accelerated scRNA-seq analysis aids in optimizing cell reprogramming protocols and assessing the quality of induced pluripotent stem cells (iPSCs). ML-driven insights into cellular differentiation pathways enhance the development of cell-based therapies for tissue repair and regeneration.

8. Challenges and Future Directions

1. Addressing scalability issues with GPU resources and cloud computing:

- **GPU Resource Management:** As scRNA-seq datasets continue to grow in size and complexity, managing GPU resources effectively becomes crucial. Optimizing parallel computing workflows and utilizing multi-GPU systems or cloud-based GPU instances can mitigate scalability challenges, ensuring rapid data processing and analysis scalability.
- **Cloud Computing:** Leveraging cloud platforms for GPU-accelerated scRNA-seq analysis offers scalability and flexibility. However, ensuring data security, minimizing latency in data transfer, and optimizing cost-efficiency are ongoing challenges. Future directions involve developing efficient data management strategies and enhancing cloud infrastructure tailored for genomic data analytics.

2. Ethical considerations in handling large-scale genomic data in real-time:

- **Data Privacy and Security:** Real-time analysis of scRNA-seq data raises concerns about patient privacy, particularly in personalized medicine and clinical applications. Adhering to data protection regulations (e.g., GDPR, HIPAA) and implementing robust encryption methods are essential to safeguard sensitive genomic information.
- **Informed Consent and Data Sharing:** Ethical guidelines emphasize obtaining informed consent for genomic research and transparent data sharing practices. Balancing data accessibility for scientific advancement with respecting individual rights and confidentiality remains a challenge. Future directions involve developing ethical frameworks that ensure equitable access to genomic data while protecting participant privacy.

3. Emerging trends and future innovations in GPU-accelerated scRNA-seq analysis:

- **Advanced ML Algorithms:** Continued development of deep learning architectures tailored for scRNA-seq data, such as graph neural networks and reinforcement learning models, will enhance predictive accuracy and scalability in real-time analysis.

- **Integration of Multi-Omics Data:** GPU-accelerated analysis platforms will integrate scRNA-seq with other omics data (e.g., single-cell ATAC-seq, proteomics) to provide comprehensive insights into cellular function and regulatory networks.
- **Real-Time Visualization and Interpretation:** Innovations in GPU-accelerated visualization tools, augmented reality (AR), and virtual reality (VR) environments will enable immersive data exploration and collaborative analysis among researchers and clinicians.
- **AI-Driven Decision Support Systems:** AI-powered platforms integrating GPU-accelerated scRNA-seq analysis with clinical data will enable real-time diagnostic support, personalized treatment recommendations, and predictive modeling of disease progression.

9. Conclusion

In summary, the integration of Graphics Processing Units (GPUs) and Machine Learning (ML) techniques represents a transformative leap forward in advancing single-cell RNA sequencing (scRNA-seq) analysis. This synergy not only enhances the speed and scalability of data processing but also unlocks deeper insights into cellular dynamics and disease mechanisms.

Benefits of GPU and ML in advancing scRNA-seq analysis:

- **Enhanced Computational Efficiency:** GPUs enable parallel processing, significantly accelerating tasks such as dimensionality reduction, clustering, and trajectory inference in scRNA-seq data analysis.
- **Improved Accuracy and Predictive Power:** ML algorithms, optimized for GPU architectures, enhance the accuracy of cell type classification, identification of gene regulatory networks, and prediction of cellular responses to stimuli or treatments.
- **Real-Time Insights:** By facilitating real-time analysis, GPU-accelerated ML models empower researchers and clinicians to dynamically monitor cellular states, characterize cellular heterogeneity, and uncover biomarkers relevant to disease progression and treatment response.

Potential impact on accelerating biomedical research and clinical applications:

- **Biomedical Research:** GPU-accelerated scRNA-seq analysis accelerates the pace of discovery in biomedical research by elucidating complex biological processes and disease mechanisms at unprecedented resolution. This knowledge fuels advancements in personalized medicine, drug discovery, and regenerative therapies.
- **Clinical Applications:** In clinical settings, real-time genomic data analysis using GPU and ML technologies supports precision medicine initiatives. It enables rapid diagnosis, prognostication, and personalized treatment strategies tailored to individual patient profiles, improving healthcare outcomes and patient care.

Call to action for further research and development in real-time genomic data analysis:

Moving forward, continuous research and development efforts are essential to harness the full potential of GPU-accelerated scRNA-seq analysis:

- **Optimizing GPU Algorithms:** Further optimization of ML algorithms for GPU architectures and development of scalable GPU-based frameworks will enhance computational efficiency and scalability in handling large-scale genomic datasets.
- **Ethical and Regulatory Considerations:** Addressing ethical considerations surrounding data privacy, informed consent, and responsible data sharing practices is crucial to build public trust and ensure ethical use of genomic data in real-time analysis.
- **Integration of Multi-Omics Data:** Future research should focus on integrating scRNA-seq data with other omics datasets (e.g., proteomics, metabolomics) to provide comprehensive insights into biological systems and disease mechanisms.

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. Hari Sankar, S., Patni, A., Mulleti, S., & Seelamantula, C. S. DIGITIZATION OF ELECTROCARDIOGRAM USING BILATERAL FILTERING.

7. 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

8. 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

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

10. 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

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

12. 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

13. 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

14. 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

15. 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

16. 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

17. 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

18. 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

19. 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

20. 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

21. 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