EasyChair Preprint

№ 13758

High-Performance Computing for Protein Structure Prediction Using ML and GPUs

Abill Robert

July 2, 2024

High-Performance Computing for Protein Structure Prediction using ML and GPUs

AUTHOR

ABILL ROBERT

DATA: June 28, 2024

Abstract

The field of protein structure prediction has witnessed significant advancements due to the integration of high-performance computing (HPC) and machine learning (ML) techniques, especially with the use of Graphics Processing Units (GPUs). This paper explores the transformative impact of HPC and ML on predicting protein structures, which is crucial for understanding biological functions and developing therapeutic interventions. The introduction of GPUs has revolutionized computational biology by offering parallel processing capabilities that significantly reduce the time required for complex calculations. Machine learning models, particularly deep learning algorithms, have demonstrated unprecedented accuracy in predicting protein folding and interactions by analyzing vast datasets of known protein structures. This synergy between ML and HPC facilitates the development of predictive models that are not only faster but also more accurate and scalable. The paper discusses key methodologies, including the use of neural networks and advanced optimization techniques, and highlights successful case studies where GPU-accelerated ML models have outperformed traditional approaches. Additionally, it addresses the challenges associated with data quality, computational costs, and the integration of ML models into existing HPC frameworks. The findings underscore the potential of combining ML and GPUs in accelerating biomedical research and drug discovery, paving the way for innovations in personalized medicine and biotechnology.

Introduction

Protein structure prediction stands at the forefront of computational biology, holding the key to unlocking numerous biological mysteries and advancing drug discovery. Understanding the three-dimensional structure of proteins is essential for elucidating their functions, interactions, and roles in various biological processes. Traditionally, methods such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy have been employed to determine protein structures. However, these techniques are often time-consuming, expensive, and limited by the size and complexity of the proteins they can analyze.

The advent of high-performance computing (HPC) has dramatically transformed the landscape of protein structure prediction. HPC systems, equipped with powerful processors and large-scale parallel computing capabilities, enable the handling of massive datasets and the execution of complex algorithms at unprecedented speeds. Among the various advancements in HPC, the utilization of Graphics Processing Units (GPUs) has emerged as a game-changer. Originally designed for rendering graphics in video games, GPUs possess a parallel architecture that excels in performing multiple computations simultaneously, making them ideal for the intensive computational tasks required in protein structure prediction.

Simultaneously, machine learning (ML), particularly deep learning, has revolutionized the field by providing sophisticated models capable of learning intricate patterns from extensive datasets. ML algorithms, especially neural networks, have demonstrated remarkable proficiency in predicting protein folding and structure by leveraging vast amounts of experimental and simulated data. The integration of ML with HPC, facilitated by GPU acceleration, has resulted in significant breakthroughs, enabling the development of models that can predict protein structures with remarkable accuracy and efficiency.

This paper delves into the synergy between HPC, ML, and GPUs in the realm of protein structure prediction. It explores the advancements in computational techniques, the implementation of ML models on GPU-accelerated platforms, and the transformative impact of these technologies on biological research. By examining key methodologies, challenges, and case studies, this paper aims to highlight the potential of HPC and ML in revolutionizing our understanding of proteins and advancing the field of computational biology. Through this exploration, we seek to underscore the critical role of high-performance computing in addressing the complex and computationally demanding task of protein structure prediction, ultimately paving the way for innovations in biotechnology and personalized medicine.

II. Protein Structure Prediction

A. Overview of Protein Structure

Proteins are complex molecules that play critical roles in virtually all biological processes. The function of a protein is intricately linked to its structure, which is organized into four distinct levels: primary, secondary, tertiary, and quaternary structures.

1. **Primary Structure**: This is the linear sequence of amino acids in a protein, linked by peptide bonds. The primary structure dictates the protein's unique characteristics and serves as the foundation for higher-level structures.
2. **Secondary Structure**: The local folding of the protein chain into structures such as alpha helices and beta sheets, stabilized by hydrogen bonds. These secondary structures form the building blocks of the overall protein architecture.
3. **Tertiary Structure**: The three-dimensional arrangement of the secondary structures into a single polypeptide chain, stabilized by various interactions, including hydrogen bonds, disulfide bridges, hydrophobic interactions, and ionic bonds. The tertiary structure is crucial for the protein's biological function, as it determines the spatial arrangement of its active sites.

4. **Quaternary Structure**: The assembly of multiple polypeptide chains or subunits into a larger functional complex. The quaternary structure is essential for the function of many proteins, such as hemoglobin, which relies on the interaction between its subunits for oxygen transport.

Understanding these structural levels is vital for deciphering protein function and interactions. The shape and conformation of a protein determine how it interacts with other molecules, influencing processes such as enzyme catalysis, signal transduction, and cellular communication. Accurate prediction and characterization of protein structures are, therefore, fundamental to advancing biomedical research and developing therapeutic interventions.

B. Traditional Methods for Protein Structure Prediction

1. **Experimental Techniques**
    - **X-ray Crystallography**: This technique involves crystallizing a protein and then diffracting X-rays through the crystal. The resulting diffraction pattern is used to determine the electron density, which is then translated into the protein's three-dimensional structure. X-ray crystallography has been instrumental in solving numerous protein structures but is limited by the requirement for high-quality crystals and is often time-consuming.
    - **Nuclear Magnetic Resonance (NMR) Spectroscopy**: NMR spectroscopy exploits the magnetic properties of atomic nuclei to determine the structure of proteins in solution. It provides detailed information about the protein's atomic environment and dynamic properties. However, NMR is generally limited to smaller proteins and complexes due to its sensitivity and resolution constraints.
    - **Cryo-Electron Microscopy (cryo-EM)**: Cryo-EM involves flash-freezing protein samples and imaging them with an electron microscope. This technique has advanced significantly in recent years, allowing the visualization of large protein complexes and membrane proteins at near-atomic resolution without the need for crystallization. Cryo-EM is particularly useful for studying dynamic and flexible proteins.
2. **Computational Approaches**
    - **Homology Modeling**: Also known as comparative modeling, this approach predicts a protein’s structure based on its similarity to known structures of related proteins. By aligning the sequence of the target protein with a template structure, homology modeling builds a model that approximates the target protein’s structure. This method is effective when there are high-sequence similarities but becomes less reliable with low-sequence homology.
    - **Molecular Dynamics (MD) Simulations**: MD simulations use physical principles to model the motions of atoms in a protein over time. By solving Newton’s equations of motion, MD simulations provide insights into the dynamic behavior and conformational changes of proteins. Although computationally intensive, advancements in algorithms and computing power have made MD simulations more accessible and accurate for studying protein structures and interactions.

III. Machine Learning Approaches

A. Supervised Learning

Supervised learning in protein structure prediction involves training models on labeled datasets consisting of protein sequences paired with their known structures. This process includes several key steps:

1. **Dataset Preparation**:
    - **Protein Sequences and Known Structures**: Large databases such as the Protein Data Bank (PDB) provide extensive collections of protein sequences and their experimentally determined structures. These datasets serve as the foundation for training supervised learning models.
2. **Feature Extraction**:
    - **Amino Acid Properties**: Features derived from individual amino acids, such as hydrophobicity, charge, and size, are crucial for understanding the protein's folding patterns.
    - **Sequence Alignment**: Aligning sequences with known structures helps identify conserved regions and structural motifs, providing insights into the likely structure of the target protein.
3. **Algorithms**:
    - **Deep Learning**: Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have shown great promise in capturing complex patterns in protein sequences and predicting their structures. These models leverage hierarchical feature extraction to learn representations from raw sequence data.
    - **Neural Networks**: Various architectures, including long short-term memory (LSTM) networks and transformers, have been applied to protein structure prediction, benefiting from their ability to handle sequential data and capture long-range dependencies.
    - **Support Vector Machines (SVMs)**: SVMs are used for classification and regression tasks in protein structure prediction, such as predicting secondary structure elements or binding sites based on extracted features.

B. Unsupervised Learning

Unsupervised learning techniques are employed to discover patterns and relationships in protein structures without the need for labeled datasets.

1. **Clustering Methods**:
    - **Structural Motifs Discovery**: Clustering algorithms like k-means and hierarchical clustering are used to identify common structural motifs and domains across large sets of protein structures. These motifs can provide insights into protein function and evolutionary relationships.
2. **Autoencoders**:
    - **Dimensionality Reduction and Pattern Recognition**: Autoencoders are neural networks designed to learn efficient representations of data by compressing the

input into a lower-dimensional space and then reconstructing it. In protein structure prediction, autoencoders can reduce the complexity of structural data, facilitating pattern recognition and feature extraction.

C. Reinforcement Learning

Reinforcement learning (RL) is an emerging approach in protein structure prediction, where models learn to make a sequence of decisions to achieve a goal, such as correctly folding a protein.

1. **Application in Folding Simulations and Conformational Sampling**:
    - RL algorithms simulate the process of protein folding by iteratively sampling conformations and optimizing the folding pathway. These simulations aim to discover the most stable and accurate protein structures.
2. **Reward Mechanisms**:
    - **Structural Accuracy and Stability**: In RL, reward functions are designed to incentivize accurate and stable protein conformations. Rewards can be based on criteria such as the similarity of predicted structures to known native structures, energy minimization, and the satisfaction of physicochemical constraints.

IV. High-Performance Computing with GPUs

A. GPU Architecture and Capabilities

1. **Comparison with Traditional CPUs**:
    - **Central Processing Units (CPUs)** are designed for general-purpose computing with a focus on sequential task execution. They have a few powerful cores optimized for single-thread performance, making them suitable for a wide range of applications but less efficient for tasks requiring massive parallelism.
    - **Graphics Processing Units (GPUs)**, on the other hand, are specialized for parallel processing. They contain thousands of smaller, less powerful cores designed to handle multiple tasks simultaneously. This architecture makes GPUs exceptionally well-suited for the large-scale parallel computations needed in tasks such as protein structure prediction.
2. **Parallel Processing Power and Memory Bandwidth**:
    - **Parallel Processing Power**: GPUs excel in executing many operations in parallel, significantly speeding up computations that can be broken down into smaller, independent tasks. This parallelism is ideal for training machine learning models and performing large-scale simulations.
    - **Memory Bandwidth**: GPUs offer higher memory bandwidth compared to CPUs, enabling faster data transfer between the memory and processing units. This is critical for handling the large datasets involved in protein structure prediction and other HPC applications.

B. Software Frameworks and Tools

1. **CUDA (Compute Unified Device Architecture)**:
    - Developed by NVIDIA, CUDA is a parallel computing platform and programming model that allows developers to utilize NVIDIA GPUs for general-purpose processing. It provides a comprehensive suite of tools and libraries to develop GPU-accelerated applications.
2. **TensorFlow**:
    - TensorFlow, an open-source machine learning framework developed by Google, supports GPU acceleration through CUDA. It is widely used for building and deploying machine learning models, including those for protein structure prediction.
3. **PyTorch**:
    - PyTorch, an open-source machine learning library developed by Facebook's AI Research lab, also supports GPU acceleration. It is known for its flexibility and ease of use, making it a popular choice for research and development in deep learning.
4. **Specialized Libraries**:
    - **cuDNN (CUDA Deep Neural Network library)**: A GPU-accelerated library for deep neural networks, cuDNN provides highly optimized implementations of standard routines such as forward and backward convolution, pooling, normalization, and activation layers.
    - **NCCL (NVIDIA Collective Communication Library)**: NCCL is designed for multi-GPU and multi-node communication. It provides fast and efficient primitives for collective communication, crucial for training large models on multiple GPUs.

C. GPU-Accelerated Algorithms

1. **Implementation of ML Models on GPUs**:
    - Implementing machine learning models on GPUs involves adapting algorithms to leverage parallel processing. Techniques such as data parallelism, where data is distributed across multiple GPU cores, and model parallelism, where different parts of the model are processed in parallel, are commonly used.
2. **Performance Optimization Techniques**:
    - **Parallelization**: Breaking down computations into parallel tasks that can be executed simultaneously on multiple GPU cores. This involves optimizing the distribution of tasks and managing dependencies to maximize concurrency.
    - **Memory Management**: Efficient memory usage is crucial for maximizing GPU performance. Techniques include minimizing memory transfers between the CPU and GPU, optimizing memory access patterns to reduce latency, and using shared memory to speed up data retrieval.

V. Case Studies and Applications

A. AlphaFold and AlphaFold 2

1. **Overview of Google's DeepMind Achievements**:
    - **AlphaFold**: Developed by DeepMind, AlphaFold made a significant breakthrough in protein structure prediction by utilizing deep learning techniques to predict 3D protein structures from amino acid sequences with unprecedented accuracy. AlphaFold's approach integrates evolutionary information from multiple sequence alignments with advanced neural network architectures.
    - **AlphaFold 2**: The second iteration, AlphaFold 2, further enhanced the accuracy and reliability of predictions. It introduced innovations such as the attention-based neural network architecture and the use of a novel end-to-end differentiable approach. AlphaFold 2 achieved remarkable success at the CASP14 (Critical Assessment of protein Structure Prediction) competition, significantly outperforming other methods.
2. **Impact on Protein Structure Prediction Accuracy and Reliability**:
    - AlphaFold and AlphaFold 2 have set new benchmarks in the field, demonstrating that deep learning can predict protein structures with near-experimental accuracy. This has accelerated research in structural biology, enabling scientists to obtain high-confidence models for proteins that were previously intractable. The accuracy and reliability of AlphaFold 2's predictions have opened new avenues for understanding protein functions, interactions, and mechanisms, profoundly impacting drug discovery and biomedical research.

B. RosettaFold

1. **Integration of Deep Learning and Molecular Modeling**:
    - RosettaFold, developed by the Baker Lab at the University of Washington, combines deep learning with the established Rosetta molecular modeling suite. RosettaFold utilizes a multi-scale approach, integrating information from both sequence and structure prediction models to refine protein structures iteratively.
    - The system leverages deep learning models to predict inter-residue distances and orientations, which are then used by the Rosetta software to generate and optimize 3D protein models. This integration enhances the accuracy of predictions, particularly for de novo protein design where no homologous structures are available.
2. **Contributions to De Novo Protein Design**:
    - RosettaFold has significantly contributed to the field of de novo protein design by enabling the creation of novel proteins with specified functions. The ability to predict and design new protein structures with high precision allows researchers to engineer proteins for various applications, including therapeutics, industrial enzymes, and synthetic biology. RosettaFold's impact extends to advancing our understanding of protein folding principles and guiding the rational design of proteins with desired properties.

C. Other Notable Models and Frameworks

1. **Open-Source Projects and Collaborative Research Initiatives**:
    - **ESMFold**: Developed by Meta AI, ESMFold uses language models for predicting protein structures. It leverages the Evolutionary Scale Modeling (ESM) approach to interpret protein sequences and predict their 3D conformations, contributing to the growing pool of tools available for protein structure prediction.
    - **OpenFold**: A community-driven effort to create an open-source implementation of AlphaFold, OpenFold aims to provide accessible and reproducible protein structure prediction models. This initiative supports collaborative research and development, fostering innovation and knowledge sharing in the field.
    - **ColabFold**: Leveraging the computational resources of Google Colab, ColabFold allows researchers to run AlphaFold predictions easily. This platform democratizes access to advanced protein structure prediction tools, enabling scientists worldwide to utilize cutting-edge models without requiring extensive computational infrastructure.

VI. Challenges and Future Directions

A. Computational Challenges

1. **Scalability Issues and Resource Requirements**:
    - **Scalability Issues**: As protein structure prediction models grow more complex, scaling these models to handle larger datasets and more intricate proteins presents significant challenges. Efficient parallelization and load balancing are crucial to leverage the full potential of high-performance computing systems.
    - **Resource Requirements**: High-performance computing (HPC) resources, including GPUs and large-scale clusters, are essential for training and deploying deep learning models for protein structure prediction. The demand for computational power can be prohibitive, particularly for smaller research institutions and organizations.
2. **Managing Large Datasets and Complex Models**:
    - **Large Datasets**: Handling vast amounts of protein sequence and structure data requires robust data management strategies. Efficient storage, retrieval, and preprocessing of data are critical to maintaining the performance and accuracy of machine learning models.
    - **Complex Models**: Deep learning models used in protein structure prediction are often highly complex, involving millions of parameters. Training these models requires substantial computational resources and sophisticated optimization techniques to ensure convergence and avoid overfitting.

B. Accuracy and Validation

1. **Ensuring Predictive Accuracy and Structural Fidelity**:
    - **Predictive Accuracy**: Achieving high predictive accuracy remains a central challenge. Continuous improvements in model architectures, training

methodologies, and integration of additional biological data are necessary to enhance the precision of protein structure predictions.
    - **Structural Fidelity**: Ensuring that predicted structures accurately reflect real-world conformations is vital. This involves not only predicting the correct fold but also capturing subtle details such as side-chain positions and dynamic conformational states.
2. **Experimental Validation and Benchmarking Against Known Structures**:
    - **Experimental Validation**: Validating computational predictions with experimental data is crucial for confirming their accuracy. Techniques like X-ray crystallography, NMR spectroscopy, and cryo-EM provide the gold standard for structural validation, though they are time-consuming and resource-intensive.
    - **Benchmarking**: Consistent benchmarking against known protein structures is essential to evaluate and compare the performance of different prediction models. Initiatives like CASP (Critical Assessment of protein Structure Prediction) provide a platform for systematic assessment and improvement of predictive methodologies.

C. Ethical and Practical Considerations

1. **Data Privacy and Intellectual Property Concerns**:
    - **Data Privacy**: Ensuring the privacy of sensitive biological data, especially in collaborative research and publicly accessible databases, is a critical concern. Implementing robust data protection measures and adhering to regulatory standards are necessary to safeguard confidential information.
    - **Intellectual Property**: The development and application of proprietary algorithms and models raise intellectual property issues. Balancing open scientific collaboration with the protection of proprietary technologies is a delicate but essential task.
2. **Translational Impact on Drug Discovery and Personalized Medicine**:
    - **Drug Discovery**: The predictive power of advanced models has significant implications for drug discovery. Accurate protein structure predictions can accelerate the identification of drug targets, facilitate virtual screening, and enhance the design of therapeutic molecules. However, translating computational predictions into clinically viable drugs requires rigorous validation and regulatory approval.
    - **Personalized Medicine**: Protein structure prediction can contribute to personalized medicine by enabling the design of tailor-made therapeutics based on individual genetic profiles. Integrating predictive models with clinical data to develop personalized treatment strategies poses both technical and ethical challenges.

Future Directions

Addressing these challenges will require continued advancements in computational methods, model accuracy, and ethical frameworks. Future research should focus on:

1. **Improving Computational Efficiency**: Developing more efficient algorithms and leveraging advancements in hardware, such as quantum computing, could address scalability and resource constraints.
2. **Enhancing Model Robustness**: Integrating multi-modal data, such as proteomics and genomics, with deep learning models can improve predictive accuracy and structural fidelity.
3. **Strengthening Validation Protocols**: Establishing more rigorous validation and benchmarking standards, along with closer collaboration between computational and experimental scientists, will enhance the reliability of predictions.
4. **Promoting Ethical Practices**: Developing clear guidelines for data privacy, intellectual property, and the translational application of predictive models will support the responsible advancement of the field.

VII. Conclusion

A. Summary of Key Points

The integration of machine learning (ML) and high-performance computing (HPC) has significantly advanced the field of protein structure prediction, enabling unprecedented levels of accuracy and efficiency. High-performance GPUs have played a crucial role in enhancing computational efficiency, allowing for the rapid processing of large datasets and the execution of complex models that were previously infeasible.

Key advancements include:

- **Machine Learning Approaches**: The application of supervised learning, unsupervised learning, and reinforcement learning has revolutionized protein structure prediction. Deep learning models, particularly those leveraging neural networks, have demonstrated remarkable success in predicting protein folding patterns and 3D structures.
- **High-Performance Computing with GPUs**: GPUs, with their parallel processing capabilities and high memory bandwidth, have accelerated the training and deployment of ML models. Software frameworks such as CUDA, TensorFlow, and PyTorch, along with specialized libraries like cuDNN and NCCL, have been instrumental in optimizing GPU performance for these tasks.
- **Case Studies and Applications**: Breakthroughs such as AlphaFold and RosettaFold have set new benchmarks for predictive accuracy and structural fidelity. These models have not only advanced our understanding of protein structures but also opened new avenues for drug discovery and de novo protein design.

B. Future Prospects

Looking ahead, continued advancements in ML algorithms and HPC technologies hold the promise of further breakthroughs in protein structure prediction. Future prospects include:

1. **Enhanced ML Algorithms**: Ongoing research into more sophisticated and efficient ML algorithms will continue to improve the accuracy and reliability of protein structure

predictions. Innovations in deep learning architectures and training methodologies will enable the handling of even more complex protein structures and interactions.

2. **Advances in HPC Technologies**: The development of next-generation GPUs, as well as other HPC technologies like quantum computing, will provide even greater computational power and efficiency. These advancements will support the scaling of models to tackle larger and more intricate datasets, further accelerating the pace of scientific discovery.
3. **Breakthroughs in Scientific and Medical Applications**: The ability to accurately predict and manipulate protein structures will have profound implications for various scientific and medical fields. This includes the development of novel therapeutics, personalized medicine approaches, and a deeper understanding of fundamental biological processes. The integration of predictive models with experimental and clinical data will enable more targeted and effective interventions, transforming the landscape of healthcare and biotechnology.

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).
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. Sankar S, H., Patni, A., Mulleti, S., & Seelamantula, C. S. (2020). Digitization of electrocardiogram using bilateral filtering. *bioRxiv*, 2020-05.
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. Sankar, S. H., Jayadev, K., Suraj, B., & Aparna, P. (2016, November). A comprehensive solution to road traffic accident detection and ambulance management. In *2016 International Conference on Advances in Electrical, Electronic and Systems Engineering (ICAEES)* (pp. 43-47). IEEE.
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