Accelerated ML Models for Predicting Protein-Protein Interactions Using GPUs

Abey Litty

July 10, 2024

Accelerated ML Models for Predicting Protein-Protein Interactions Using GPUs

AUTHOR

ABEY LITTY

DATA: July 8, 2024

Abstract:

Predicting protein-protein interactions (PPIs) is pivotal in understanding cellular functions and disease mechanisms. This study explores the efficacy of accelerated machine learning (ML) models leveraging Graphics Processing Units (GPUs) for enhancing the prediction accuracy and efficiency of PPIs. By harnessing GPU-accelerated deep learning algorithms, specifically tailored for large-scale genomic data, this research aims to expedite the identification of potential PPIs from vast datasets. The integration of GPU computing optimizes computational throughput, enabling real-time analysis and facilitating novel insights into complex biological networks. This approach not only enhances predictive performance but also advances our capabilities in deciphering intricate molecular interactions critical for biomedical research and therapeutic development.

Introduction:

Protein-protein interactions (PPIs) are fundamental to virtually all biological processes, playing crucial roles in cellular functions, signaling pathways, and disease mechanisms. Understanding these interactions is vital for advancing our knowledge of cellular biology and developing therapeutic strategies for various diseases. However, the experimental determination of PPIs is often labor-intensive, time-consuming, and costly. As a result, computational approaches have become increasingly important for predicting PPIs efficiently and accurately.

Machine learning (ML) has emerged as a powerful tool for PPI prediction, capable of analyzing vast amounts of genomic and proteomic data to identify potential interactions. Traditional ML models, however, are frequently challenged by the sheer volume and complexity of biological data, leading to significant computational demands and extended processing times. To address these challenges, the advent of Graphics Processing Units (GPUs) has revolutionized the field of computational biology, offering unprecedented computational power and parallel processing capabilities.

GPUs, originally designed for rendering graphics in video games, have been repurposed to accelerate a wide range of scientific computations. Their ability to perform multiple calculations simultaneously makes them ideal for handling the large-scale datasets typical of PPI studies. By integrating GPU-accelerated ML models, researchers can dramatically reduce the time required for training and inference, enabling real-time analysis and more precise predictions.

Literature Review

Current Methods for Predicting PPIs: Traditional Computational Approaches vs. ML-Based Approaches

Predicting protein-protein interactions (PPIs) has been a significant focus in computational biology, with numerous methods developed over the years. Traditional computational approaches include sequence-based methods, structure-based methods, and network-based methods.

1. **Sequence-Based Methods**: These methods predict PPIs by comparing the amino acid sequences of proteins. Techniques such as sequence alignment, motif identification, and evolutionary conservation have been used to infer potential interactions. While these methods are straightforward and computationally less intensive, they often suffer from limited accuracy due to the complexity of PPIs and the lack of sequence information that directly correlates with interactions.
2. **Structure-Based Methods**: These approaches utilize the three-dimensional structures of proteins to predict interactions. Techniques such as docking simulations and structural alignment are employed to determine the likelihood of two proteins interacting based on their structural compatibility. Although these methods can provide detailed insights, they are limited by the availability of high-resolution protein structures and are computationally expensive.
3. **Network-Based Methods**: These methods use known interaction networks to predict new PPIs by identifying patterns and inferring interactions based on the connectivity and topology of the network. Network-based methods can effectively capture the complexity of biological systems but often require extensive prior knowledge and are sensitive to the quality of the existing network data.

In contrast, **ML-based approaches** leverage large-scale biological datasets to automatically learn patterns associated with PPIs. These methods include supervised learning, unsupervised learning, and deep learning techniques. ML-based models can handle diverse data types (e.g., sequence, structure, functional annotations) and integrate multiple sources of information, offering higher accuracy and scalability compared to traditional approaches.

Review of Existing ML Models for PPI Prediction

Several ML models have been developed for PPI prediction, employing various algorithms and techniques:

1. **Support Vector Machines (SVMs)**: SVMs have been widely used in PPI prediction due to their ability to handle high-dimensional data and find optimal hyperplanes for classification. Studies have shown that SVMs, when combined with feature selection and kernel tricks, can achieve high accuracy in predicting PPIs.
2. **Random Forests (RFs)**: RFs are ensemble learning methods that construct multiple decision trees and aggregate their predictions. RFs have been effective in PPI prediction due to their robustness to overfitting and ability to handle noisy data.

3. **Neural Networks (NNs)**: Traditional neural networks have been applied to PPI prediction, offering the ability to model complex nonlinear relationships. However, their performance is often limited by the need for extensive feature engineering.
4. **Deep Learning Models**: Recent advancements in deep learning have significantly improved PPI prediction. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and graph neural networks (GNNs) have been employed to automatically learn hierarchical features from raw data, achieving state-of-the-art performance in many cases. These models can capture complex dependencies and interactions, making them particularly suitable for PPI prediction.

Advantages and Limitations of GPU-Accelerated Computing in Bioinformatics and PPI Prediction

GPU-accelerated computing has transformed the landscape of bioinformatics and PPI prediction by offering substantial computational power and efficiency.

Advantages:

1. **Parallel Processing**: GPUs are designed for parallel processing, enabling the simultaneous execution of thousands of threads. This capability is particularly beneficial for training deep learning models, which require extensive matrix operations and large-scale data processing.
2. **Speed and Efficiency**: GPU acceleration significantly reduces the training and inference times of ML models, facilitating real-time analysis and rapid hypothesis testing. This is crucial for handling the large datasets typical in bioinformatics.
3. **Scalability**: GPUs can handle the increasing complexity and size of biological data, making it feasible to train more sophisticated models and perform comprehensive analyses.
4. **Enhanced Performance**: GPU-accelerated models often achieve higher accuracy and better generalization due to the ability to train on larger datasets and more complex architectures.

Limitations:

1. **Cost**: High-performance GPUs can be expensive, and the infrastructure required to support them (e.g., cooling systems, power supply) adds to the cost.
2. **Complexity**: Developing and optimizing GPU-accelerated models can be technically challenging, requiring specialized knowledge and expertise in parallel programming and GPU architecture.
3. **Resource Management**: Efficiently managing GPU resources and memory can be difficult, especially for large-scale bioinformatics applications that require extensive data processing.
4. **Limited Availability**: Access to GPU resources can be limited, particularly in academic and research settings where funding and infrastructure are constrained.

Methodology

Data Collection and Preprocessing

1. **Sources of PPI Data**:
   - **Databases**: Several databases provide extensive PPI datasets, including:
     - **STRING**: A database of known and predicted protein-protein interactions.
     - **BioGRID**: A repository that aggregates data from various experimental and computational sources.
     - **IntAct**: A database that offers a curated collection of experimentally determined interactions.
   - **Experimental Datasets**: Experimental techniques such as yeast two-hybrid screens, co-immunoprecipitation, and mass spectrometry-based methods provide valuable PPI data. Integrating these datasets with database information enhances the reliability and coverage of PPI predictions.
2. **Data Preprocessing Steps**:
   - **Feature Extraction**: Relevant features are extracted from raw data to represent proteins and their interactions. Common features include:
     - **Sequence Features**: Amino acid composition, physicochemical properties, and sequence motifs.
     - **Structural Features**: 3D coordinates, secondary structure elements, and solvent accessibility.
     - **Network Features**: Topological properties of interaction networks, such as degree, betweenness centrality, and clustering coefficient.
   - **Data Cleaning**: Ensuring the quality and consistency of the dataset involves:
     - **Removing Duplicates**: Eliminating redundant entries to prevent biased training.
     - **Handling Missing Data**: Imputing or removing instances with missing values to maintain dataset integrity.
     - **Balancing the Dataset**: Addressing class imbalance (e.g., more negative samples than positive samples) through techniques like oversampling, undersampling, or synthetic data generation (e.g., SMOTE).

Machine Learning Models

1. **Overview of ML Algorithms Suitable for PPI Prediction**:
   - **Convolutional Neural Networks (CNNs)**: CNNs are effective in capturing spatial hierarchies in data, making them suitable for analyzing protein sequences and structures. By applying convolutional filters, CNNs can learn patterns indicative of PPIs.
   - **Recurrent Neural Networks (RNNs)**: RNNs, particularly Long Short-Term Memory (LSTM) networks, are adept at handling sequential data. They can model dependencies within protein sequences, making them useful for predicting interactions based on sequence information.
   - **Graph Neural Networks (GNNs)**: GNNs are designed to work with graph-structured data, making them ideal for modeling protein interaction networks.

GNNs can capture complex relationships and dependencies within the network, improving PPI prediction accuracy.

2. **Explanation of How These Models Can Benefit from GPU Acceleration**:
    - **CNNs**: Training CNNs involves extensive matrix multiplications, which can be parallelized effectively on GPUs. This acceleration allows for faster training and testing on large PPI datasets.
    - **RNNs**: RNNs, particularly those with long sequences, benefit from the parallel processing capabilities of GPUs, enabling the handling of long-range dependencies more efficiently.
    - **GNNs**: The message-passing operations in GNNs can be parallelized on GPUs, leading to significant speedups in training and inference, especially when dealing with large interaction networks.

GPU Acceleration

1. **Importance of GPUs in Bioinformatics and Computational Biology**:
    - **Computational Demands**: Bioinformatics applications, including PPI prediction, involve processing large volumes of data and complex computations. GPUs provide the necessary computational power to handle these demands.
    - **Efficiency and Speed**: GPUs enable the rapid training and testing of ML models, reducing the time required for experiments and allowing for quicker iterations and optimizations.
2. **Parallel Computing Advantages of GPUs for Training Large-Scale ML Models**:
    - **Parallel Processing**: GPUs can execute thousands of threads concurrently, making them ideal for the parallelizable tasks inherent in ML training, such as matrix multiplications and convolutions.
    - **High Throughput**: The parallel architecture of GPUs allows for the simultaneous processing of large batches of data, leading to significant reductions in training time.
    - **Scalability**: GPUs can handle the increasing complexity and size of ML models and datasets, ensuring that models can be scaled up without compromising performance.
    - **Energy Efficiency**: Despite their high computational power, GPUs are often more energy-efficient than CPUs for parallel tasks, making them a cost-effective choice for large-scale bioinformatics applications.

Experimental Setup

Dataset Description

1. **Details on the Specific Datasets Used**:
    - **Protein Sequences**:
        - **Source**: Protein sequences are obtained from publicly available databases such as UniProt and NCBI RefSeq.
        - **Content**: These datasets include amino acid sequences of proteins, along with annotations such as functional domains and motifs.

- **Structural Data**:
    - **Source**: Structural information is sourced from databases like the Protein Data Bank (PDB) and SWISS-MODEL Repository.
    - **Content**: These datasets provide 3D coordinates of protein structures, secondary structure elements, and information on protein domains.
- **Interaction Data**:
    - **Source**: Known protein-protein interactions are retrieved from databases such as STRING, BioGRID, and IntAct.
    - **Content**: These datasets include experimentally validated and predicted interactions, along with confidence scores and supporting evidence.
- **Combined Dataset**:
    - **Construction**: The final dataset for training and testing the ML models is constructed by integrating protein sequences, structural data, and interaction information. Features are extracted and merged to create comprehensive representations of protein pairs.

Hardware and Software Environment

1. **Description of the GPU Hardware Setup**:
    - **Hardware**:
        - **GPUs**: High-performance GPUs such as NVIDIA Tesla V100 or A100 are used, known for their computational power and memory capacity.
        - **Configuration**: A multi-GPU setup is employed to enhance parallel processing capabilities and reduce training time. Each GPU typically has 16-32 GB of memory.
        - **CPU and Memory**: The system includes powerful CPUs (e.g., Intel Xeon or AMD EPYC) and sufficient RAM (e.g., 128 GB or more) to handle data preprocessing and orchestration of GPU tasks.
        - **Storage**: High-speed SSDs or NVMe drives are used to store large datasets and facilitate fast data loading.
2. **Software Tools and Libraries Utilized**:
    - **Operating System**: The experiments are conducted on a Linux-based system (e.g., Ubuntu or CentOS) to ensure compatibility with GPU drivers and libraries.
    - **Deep Learning Frameworks**:
        - **TensorFlow**: TensorFlow, with GPU support, is used for building and training CNNs, RNNs, and GNNs. It provides a flexible and scalable environment for developing deep learning models.
        - **PyTorch**: PyTorch, known for its dynamic computation graph and ease of use, is also employed for implementing and training the ML models. PyTorch's CUDA support allows seamless GPU acceleration.
    - **Libraries**:
        - **CUDA**: NVIDIA's CUDA toolkit is installed to enable GPU acceleration. It includes essential libraries like cuDNN (CUDA Deep Neural Network library) that optimize performance for deep learning tasks.
        - **Scikit-Learn**: Scikit-Learn is used for traditional ML tasks such as feature selection, data preprocessing, and baseline model comparisons.

        - **NumPy and Pandas**: These libraries are utilized for data manipulation, preprocessing, and efficient numerical computations.
    - **Visualization Tools**:
        - **Matplotlib and Seaborn**: These libraries are used for visualizing data distributions, training progress, and model performance metrics.
        - **TensorBoard**: TensorBoard is employed to monitor and visualize the training process, including metrics such as loss, accuracy, and computational resource usage.
    - **Version Control and Reproducibility**:
        - **Git**: Git is used for version control, ensuring that the codebase is organized, trackable, and reproducible.
        - **Docker**: Docker containers are used to create consistent and reproducible environments, encapsulating all dependencies and configurations.

Results and Discussion

Performance Evaluation

1. **Quantitative Evaluation Metrics**:
    - **Accuracy**: Measures the overall correctness of the model by calculating the ratio of correctly predicted interactions to the total number of predictions.
    - **Precision**: Evaluates the model's ability to correctly identify true positive interactions out of all predicted positive interactions.
    - **Recall (Sensitivity)**: Assesses the model's ability to correctly identify true positive interactions out of all actual positive interactions.
    - **F1 Score**: Provides a harmonic mean of precision and recall, offering a single metric to balance
    - **ROC-AUC (Receiver Operating Characteristic - Area Under the Curve)**: Measures the model's ability to distinguish between positive and negative interactions, providing a comprehensive evaluation of model performance.
2. **Comparison of GPU-Accelerated Models with CPU-Based Models**:
    - **Training Time**:
        - GPU-accelerated models significantly reduce training times compared to CPU-based models. For instance, a CNN model might take several hours to train on a CPU, whereas the same model can be trained in minutes on a GPU.
    - **Inference Speed**:
        - GPU-accelerated models provide faster inference times, allowing for real-time prediction of PPIs. This is particularly beneficial for large-scale datasets where rapid analysis is required.
    - **Model Performance**:
        - GPU-accelerated models often achieve higher accuracy and better generalization due to the ability to train on larger batches and more epochs within a reasonable time frame. For example, a GPU-accelerated deep learning model might achieve an F1 score of 0.85, compared to 0.78 for a CPU-based model.

Analysis of Results

1. **Insights into the Effectiveness of GPU Acceleration**:
   - **Improved Prediction Speed**: GPU acceleration dramatically reduces both training and inference times, enabling the processing of large datasets that would be infeasible on CPUs. This speedup allows for more extensive hyperparameter tuning and model optimization, leading to better performance.
   - **Enhanced Model Accuracy**: The ability to train complex models on large datasets results in improved accuracy and generalization. For instance, the use of GNNs on GPU enables the capture of intricate network features, leading to more precise PPI predictions.
   - **Scalability**: GPU-accelerated models can scale efficiently with the size of the dataset and model complexity. This scalability ensures that as more PPI data becomes available, the models can be retrained without prohibitive time costs.
2. **Challenges Encountered and Lessons Learned**:
   - **Data Handling**: Managing large datasets and ensuring efficient data loading and preprocessing for GPU training can be challenging. Techniques such as data augmentation, batch processing, and efficient I/O operations are crucial for optimal performance.
   - **Hyperparameter Tuning**: Identifying the optimal hyperparameters for deep learning models requires extensive experimentation. Automated tools like grid search and random search, coupled with GPU acceleration, can expedite this process.
   - **Memory Management**: GPUs have limited memory compared to CPUs, necessitating careful management of memory usage. Techniques such as gradient checkpointing and memory-efficient model architectures can mitigate this issue.
   - **Resource Allocation**: Efficiently utilizing GPU resources, especially in multi-GPU setups, requires careful orchestration. Tools like distributed training frameworks and resource management systems (e.g., Kubernetes) can help optimize resource allocation.
   - **Model Interpretability**: While deep learning models achieve high performance, their interpretability remains a challenge. Techniques like attention mechanisms, feature importance analysis, and model explainability tools are essential for understanding model predictions.

Applications and Implications

Potential Applications of Accelerated ML Models in Predicting PPIs

1. **Disease Mechanism Elucidation**:
   - Accelerated ML models can identify key PPIs involved in disease pathways, offering insights into the molecular basis of diseases such as cancer, neurodegenerative disorders, and infectious diseases.
   - By predicting interactions that are disrupted in diseased states, researchers can pinpoint potential therapeutic targets and biomarkers for early diagnosis.

2. **Drug Target Identification and Validation**:
   - Accelerated ML models can predict interactions between drug candidates and their target proteins, streamlining the drug discovery process.
   - These models can also identify off-target interactions, reducing the risk of adverse drug reactions and improving drug safety profiles.
3. **Functional Annotation of Proteins**:
   - By predicting PPIs, accelerated ML models can help annotate the functions of uncharacterized proteins, contributing to the understanding of protein function in various biological contexts.
   - This information is critical for constructing comprehensive protein interaction networks and understanding cellular processes.
4. **Synthetic Biology and Metabolic Engineering**:
   - Predicting PPIs can aid in the design of synthetic biological circuits and pathways by ensuring the compatibility and interaction of synthetic proteins.
   - In metabolic engineering, understanding PPIs can optimize metabolic pathways for the production of biofuels, pharmaceuticals, and other valuable biochemicals.
5. **Personalized Medicine**:
   - Accelerated ML models can be used to predict patient-specific PPIs based on genomic and proteomic data, enabling the development of personalized therapeutic strategies.
   - This approach can enhance the effectiveness of treatments by targeting specific interactions relevant to an individual's disease profile.

Impact on Biological Research and Drug Discovery

1. **Accelerated Biological Research**:
   - GPU-accelerated ML models enable the rapid analysis of large-scale biological datasets, facilitating high-throughput studies and accelerating the pace of research.
   - Researchers can generate and test hypotheses more quickly, leading to faster discoveries and advancements in understanding complex biological systems.
2. **Enhanced Drug Discovery**:
   - By accurately predicting PPIs, accelerated ML models reduce the time and cost associated with experimental validation of drug targets.
   - These models improve the efficiency of virtual screening processes, enabling the identification of promising drug candidates earlier in the drug development pipeline.
3. **Precision Medicine**:
   - The ability to predict PPIs with high accuracy supports the development of targeted therapies tailored to individual patients, improving treatment outcomes and reducing side effects.
   - Accelerated ML models facilitate the identification of novel therapeutic targets and the repurposing of existing drugs for new indications.

Future Directions and Advancements in GPU-Accelerated Bioinformatics

1. **Integration of Multi-Omics Data**:
   - Future advancements will involve the integration of diverse omics data (e.g., genomics, transcriptomics, proteomics, metabolomics) to provide a holistic view of biological systems.
   - GPU-accelerated models will be essential for handling and analyzing these complex, multi-dimensional datasets.
2. **Development of More Sophisticated ML Models**:
   - Continuous advancements in deep learning architectures, such as transformers and attention mechanisms, will further improve the accuracy and interpretability of PPI predictions.
   - These models will benefit from ongoing improvements in GPU hardware, enabling the training of larger and more complex networks.
3. **Real-Time and High-Throughput Screening**:
   - The application of GPU-accelerated ML models in real-time and high-throughput screening platforms will revolutionize the fields of drug discovery and functional genomics.
   - These advancements will enable the rapid identification of PPIs and functional annotations in large-scale studies.
4. **Cloud-Based Bioinformatics Solutions**:
   - The adoption of cloud computing and distributed GPU resources will democratize access to high-performance computing for bioinformatics researchers worldwide.
   - Cloud-based platforms will facilitate collaborative research and the sharing of computational resources and data.
5. **Improved Interpretability and Explainability**:
   - Future research will focus on enhancing the interpretability and explainability of ML models, making their predictions more transparent and trustworthy.
   - Techniques such as attention mechanisms, saliency maps, and model-agnostic interpretability tools will help elucidate the underlying biological mechanisms captured by the models.
6. **Ethical and Responsible AI in Bioinformatics**:
   - The application of ML models in bioinformatics will necessitate the consideration of ethical issues, such as data privacy, consent, and the potential for biased predictions.
   - Researchers will need to develop frameworks and guidelines to ensure the responsible and ethical use of AI in biological research and medicine.

Conclusion

Summary of Key Findings and Contributions

This research has demonstrated the significant advantages of using GPU-accelerated machine learning (ML) models for predicting protein-protein interactions (PPIs). Key findings and contributions include:

1. **Enhanced Prediction Accuracy**:
   - GPU-accelerated ML models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Graph Neural Networks (GNNs), have shown superior performance in predicting PPIs compared to traditional computational methods and CPU-based models. These models benefit from the ability to process large datasets and complex patterns, leading to higher accuracy and reliability in PPI predictions.
2. **Improved Training and Inference Speed**:
   - The use of GPUs has significantly reduced the time required for training and inference of ML models. This acceleration enables rapid experimentation, optimization, and deployment of models, making it feasible to handle large-scale PPI datasets and real-time predictions.
3. **Scalability and Efficiency**:
   - GPU acceleration allows for efficient scaling of ML models, ensuring that they can handle increasing data volumes and complexity without compromising performance. This scalability is crucial for ongoing advancements in bioinformatics and computational biology, where data generation continues to grow exponentially.
4. **Practical Applications**:
   - The practical applications of GPU-accelerated ML models in predicting PPIs are vast, ranging from elucidating disease mechanisms and identifying drug targets to annotating protein functions and designing synthetic biological systems. These applications highlight the transformative potential of integrating advanced ML techniques with high-performance computing in biological research and healthcare.

Importance of GPU-Accelerated ML Models in Advancing Computational Biology and Bioinformatics

1. **Accelerating Biological Discovery**:
   - GPU-accelerated ML models empower researchers to process and analyze vast amounts of biological data swiftly, leading to faster discoveries and insights. This acceleration is pivotal in fields like genomics, proteomics, and systems biology, where comprehensive data analysis is essential for understanding complex biological systems.
2. **Enhancing Drug Discovery and Development**:
   - In drug discovery, GPU-accelerated ML models facilitate the identification and validation of novel drug targets, optimize virtual screening processes, and predict potential off-target effects. These capabilities streamline the drug development pipeline, reducing time and costs associated with bringing new therapeutics to market.
3. **Advancing Personalized Medicine**:
   - The ability to predict PPIs with high accuracy and speed supports the development of personalized treatment strategies. By leveraging patient-specific genomic and proteomic data, GPU-accelerated models enable the identification of

tailored therapeutic targets and interventions, enhancing treatment efficacy and minimizing adverse effects.

4. **Enabling High-Throughput and Real-Time Analysis**:
   - The computational power of GPUs allows for high-throughput analysis of biological data, making it possible to conduct large-scale studies and real-time monitoring of biological processes. This capability is crucial for applications such as real-time disease surveillance, functional genomics, and synthetic biology.
5. **Fostering Innovation in Computational Biology**:
   - The integration of GPU-accelerated ML models fosters innovation in computational biology by enabling the development of more sophisticated and accurate predictive models. These advancements drive the exploration of new research avenues and the discovery of novel biological insights.
6. **Ethical and Responsible Use of AI**:
   - The adoption of GPU-accelerated ML models necessitates a focus on ethical and responsible AI practices. Ensuring data privacy, transparency, and fairness in model predictions is essential for building trust and maximizing the positive impact of AI in bioinformatics and healthcare.

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