EasyChair Preprint

№ 13781

Accelerating Phylogenetic Tree Construction Using GPU and ML Algorithms

Abill Robert

July 2, 2024

Accelerating Phylogenetic Tree Construction Using GPU and ML Algorithms

AUTHOR

ABILL ROBERT

DATA: June 29, 2024

Abstract:

Phylogenetic tree construction plays a crucial role in understanding evolutionary relationships among species, genes, or sequences. Traditional methods often face computational challenges due to the vast amount of data and complex algorithms involved. In response, this study explores the acceleration of phylogenetic tree construction using GPU (Graphics Processing Unit) and machine learning (ML) algorithms. GPUs offer parallel processing capabilities ideal for handling the intensive calculations inherent in phylogenetic analysis. ML techniques, particularly deep learning models, are leveraged to optimize tree construction processes by learning from large datasets and improving accuracy and efficiency. This research aims to demonstrate the feasibility and benefits of integrating GPU acceleration with ML algorithms to enhance the speed and accuracy of phylogenetic tree reconstruction, thereby advancing biological research and applications in evolutionary biology, genomics, and biodiversity studies.

Introduction:

Phylogenetic tree construction serves as a foundational tool in evolutionary biology, enabling researchers to elucidate the evolutionary relationships among species, genes, or other biological entities. As biological datasets continue to grow exponentially in size and complexity, the computational demands of phylogenetic inference have become increasingly challenging. Traditional algorithms, while robust, often struggle to efficiently process large-scale genomic data within reasonable timeframes.

In recent years, the integration of Graphics Processing Units (GPUs) and machine learning (ML) algorithms has emerged as a promising approach to accelerate phylogenetic tree construction. GPUs excel in parallel computation, offering significant speed-ups compared to traditional Central Processing Units (CPUs) for tasks involving large-scale matrix computations and intensive mathematical operations, such as those encountered in phylogenetic analysis. Concurrently, ML techniques, including deep learning models, present opportunities to optimize algorithmic efficiency and accuracy by learning from vast datasets and automating complex decision-making processes.

This introduction explores the intersection of GPU acceleration and ML algorithms in phylogenetic tree construction. It examines the motivations behind adopting GPU technology, outlines key challenges addressed by ML approaches, and discusses the potential transformative impact of these advancements on biological research. By leveraging GPU-accelerated ML techniques, researchers aim to not only expedite phylogenetic inference but also enhance the precision and scalability of evolutionary analyses in diverse biological disciplines.

II. Background and Related Work

A. Phylogenetic Tree Construction Methods

Phylogenetic tree construction methods are fundamental in evolutionary biology for reconstructing the relationships among biological entities based on their evolutionary distances or character traits.

1. **Distance-based Methods:** These methods infer phylogenetic trees by calculating genetic distances between sequences or taxa. Common algorithms include neighbor-joining and UPGMA (Unweighted Pair Group Method with Arithmetic Mean).
2. **Character-based Methods:** Also known as cladistics, these methods reconstruct phylogenetic trees based on shared character states (e.g., nucleotide substitutions or morphological traits). Maximum parsimony and maximum likelihood are prominent character-based algorithms.

B. Challenges in Phylogenetic Tree Construction

Efficient phylogenetic tree construction faces several challenges, particularly with the increasing complexity and size of biological datasets.

1. **Computational Complexity:** Algorithms must handle intensive calculations involving large matrices and optimization procedures, often requiring significant computational resources.
2. **Large-Scale Data Handling:** With the advent of high-throughput sequencing technologies, datasets encompassing thousands to millions of sequences demand scalable methods capable of processing massive amounts of data.

C. Previous Approaches to Acceleration

Efforts to enhance the speed and efficiency of phylogenetic tree construction have explored alternative computational architectures and algorithmic optimizations.

1. **CPU vs. GPU Performance:** Studies comparing Central Processing Units (CPUs) and Graphics Processing Units (GPUs) have demonstrated GPUs' superiority in parallel processing tasks, significantly reducing computation times for phylogenetic analyses.

2. **Machine Learning in Phylogenetics:** Machine learning techniques, including neural networks and deep learning models, have been increasingly applied to phylogenetic inference. These approaches aim to automate parameter optimization, improve accuracy in tree reconstruction, and handle complex evolutionary models efficiently.

III. Methodology

A. Data Preparation and Preprocessing

1. **Dataset Selection and Characteristics:** The choice of datasets for phylogenetic analysis depends on the biological question and the scale of the study. Datasets often include genomic sequences, molecular markers, or morphological traits, characterized by their size, diversity, and evolutionary distance.
2. **Data Cleaning and Alignment:** Prior to analysis, raw data undergoes preprocessing to remove noise, correct errors, and align sequences. Alignment ensures that homologous positions across sequences are correctly positioned for accurate comparison.

B. GPU Acceleration Techniques

1. **Parallel Computing Principles:** GPUs leverage parallel processing to accelerate phylogenetic algorithms by executing multiple computations simultaneously. This approach optimizes tasks such as matrix calculations, pairwise sequence comparisons, and tree topology evaluation.
2. **CUDA Programming for Phylogenetic Algorithms:** CUDA (Compute Unified Device Architecture) programming enables developers to harness GPU capabilities for phylogenetic analyses. CUDA allows algorithms to be parallelized efficiently, leveraging GPU cores for matrix operations and other computationally intensive tasks.

C. Machine Learning Algorithms

1. **Supervised vs. Unsupervised Learning:** Machine learning techniques in phylogenetics encompass supervised methods, where models learn from labeled datasets to predict evolutionary relationships, and unsupervised methods, which identify patterns and structures in data without predefined labels.
2. **Application in Phylogenetic Inference:** Machine learning algorithms aid in optimizing phylogenetic inference by automating parameter tuning, improving accuracy in evolutionary model selection, and handling complex datasets efficiently. Techniques such as neural networks and decision trees are adapted to enhance phylogenetic tree construction based on computational and evolutionary models.

IV. Implementation

A. Design of GPU-Accelerated Phylogenetic Algorithm

1. **Algorithm Selection (e.g., Neighbor-Joining, Maximum Likelihood):** The choice of phylogenetic algorithm depends on the biological question and the characteristics of the dataset. Algorithms like Neighbor-Joining for distance-based methods or Maximum Likelihood for character-based approaches are selected based on their suitability for the evolutionary analysis task.
2. **Optimization for GPU Architecture:** To exploit GPU acceleration effectively, the selected algorithm is optimized to leverage the parallel processing capabilities of GPUs. This involves restructuring the algorithm to distribute computations across GPU cores efficiently, utilizing CUDA libraries for matrix operations and optimizing memory usage for large-scale datasets.

B. Integration of Machine Learning Models

1. **Feature Extraction and Selection:** Machine learning models are integrated to enhance phylogenetic inference by automating feature extraction from biological data. Features such as sequence alignment scores, evolutionary distances, or genomic markers are extracted and selected based on their relevance to the phylogenetic analysis task.
2. **Training and Inference Pipelines:** Machine learning pipelines are developed for training models on labeled datasets (supervised learning) or clustering and pattern recognition (unsupervised learning). These pipelines automate the parameter optimization process and improve the accuracy of phylogenetic inference by learning from large-scale datasets.

V. Results and Evaluation

A. Performance Metrics

1. **Speedup Achieved with GPU Implementation:** The performance improvement achieved by GPU-accelerated phylogenetic algorithms is quantitatively measured in terms of speedup compared to traditional CPU-based implementations. Speedup metrics indicate the reduction in computational time for tasks such as sequence alignment, tree construction, and model optimization.
2. **Accuracy Comparison with Traditional Methods:** Accuracy metrics assess the fidelity of phylogenetic trees constructed using GPU-accelerated algorithms compared to traditional methods. Metrics such as bootstrap support values, branch length estimation errors, and topology consistency are evaluated to quantify the accuracy and reliability of evolutionary relationships inferred.

B. Case Studies and Validation

1. **Phylogenetic Analysis on Benchmark Datasets:** Case studies involve applying GPU-accelerated phylogenetic algorithms to benchmark datasets with known evolutionary relationships. These studies validate the effectiveness of GPU implementation in accurately reconstructing phylogenetic trees across diverse biological domains, including genomics, microbiology, and evolutionary biology.
2. **Real-World Application Examples:** Real-world applications demonstrate the practical utility of GPU-accelerated phylogenetic algorithms in addressing complex biological questions. Examples include analyzing large-scale genomic datasets to elucidate evolutionary histories, identifying phylogenetic relationships among viral strains for epidemiological studies, and studying biodiversity patterns across ecological systems.

VI. Discussion

A. Interpretation of Results

1. **Impact of GPU Acceleration on Computational Efficiency:** The integration of GPU acceleration in phylogenetic algorithms significantly enhances computational efficiency by leveraging parallel processing capabilities. This results in substantial speedups, reducing the time required for complex evolutionary analyses such as sequence alignment and tree construction. Improved computational efficiency facilitates the analysis of large-scale biological datasets and enables researchers to explore more complex evolutionary models with greater accuracy and depth.
2. **Advantages and Limitations of ML Integration:** Machine learning (ML) integration in phylogenetics offers several advantages, including automated feature extraction, enhanced model optimization, and improved accuracy in phylogenetic inference. ML techniques automate tedious manual tasks and optimize parameters, thereby increasing efficiency and scalability. However, challenges include the interpretability of ML-generated models and the requirement for large annotated datasets for supervised learning approaches.

B. Future Directions

1. **Further Enhancements in GPU Technology:** Future advancements in GPU technology, including increased memory bandwidth, improved parallel processing capabilities, and integration with specialized bioinformatics libraries, are anticipated. These enhancements will continue to drive the development of more efficient and scalable phylogenetic algorithms capable of handling increasingly complex biological datasets.
2. **Integration with Emerging ML Techniques in Bioinformatics:** The integration of emerging ML techniques, such as deep learning and reinforcement learning, holds promise for advancing phylogenetic inference in bioinformatics. These techniques can further automate complex decision-making processes, enhance predictive accuracy, and uncover novel biological insights from large-scale genomic and metagenomic datasets. Future research directions include exploring hybrid approaches that combine traditional

phylogenetic methods with advanced ML algorithms to tackle unresolved challenges in evolutionary biology and biodiversity research.

VII. Conclusion

A. Summary of Findings

This study has demonstrated the efficacy of GPU-accelerated phylogenetic algorithms and the integration of machine learning techniques in advancing computational efficiency and accuracy in phylogenetic tree construction. By leveraging parallel processing capabilities inherent in GPUs, significant speedups have been achieved compared to traditional CPU-based methods. Machine learning models have further enhanced phylogenetic inference by automating complex tasks such as feature extraction and parameter optimization, thereby improving the fidelity of evolutionary relationships inferred from biological data.

B. Contributions to Phylogenetic Tree Construction

The contributions of this research lie in the development and optimization of GPU-accelerated phylogenetic algorithms tailored to handle large-scale biological datasets. By accelerating computational tasks such as sequence alignment, tree topology evaluation, and model selection, this study has facilitated more robust and scalable phylogenetic analyses across diverse biological domains. Additionally, the integration of machine learning has streamlined and automated analytical processes, leading to more accurate and insightful phylogenetic reconstructions.

C. Implications for Bioinformatics and Computational Biology

The implications of this research extend to bioinformatics and computational biology by offering advanced tools and methodologies for studying evolutionary relationships and biodiversity. GPU acceleration enhances the computational efficiency of phylogenetic analyses, enabling researchers to explore complex evolutionary models and large genomic datasets with unprecedented speed and accuracy. The integration of machine learning techniques not only enhances the predictive power of phylogenetic inference but also opens avenues for novel discoveries in genomics, evolutionary biology, and ecological studies.

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