GPU MrBayes V3.1: MrBayes on Graphics Processing Units for Protein Sequence Data

Shuai Pang,[1,2] Rebecca J. Stones,[1,2] Ming-Ming Ren,*[1,2] Xiao-Guang Liu,[1,2] Gang Wang,[1,2] Hong-ju Xia,[1,2] Hao-Yang Wu,[3] Yang Liu,[3] and Qiang Xie[3]

[1]College of Computer and Control Engineering, Nankai University, Tianjin, China

[2]College of Software, Nankai University, Tianjin, China

[3]College of Life Science, Nankai University, Tianjin, China

***Corresponding author:** E-mail: renmingming@nbjl.nankai.edu.cn.

Associate editor: Sergei Kosakovsky Pond

Abstract

We present a modified GPU (graphics processing unit) version of MrBayes, called ta(MC)^3 (GPU MrBayes V3.1), for Bayesian phylogenetic inference on protein data sets. Our main contributions are 1) utilizing 64-bit variables, thereby enabling ta(MC)^3 to process larger data sets than MrBayes; and 2) to use Kahan summation to improve accuracy, convergence rates, and consequently runtime. Versus the current fastest software, we achieve a speedup of up to around 2.5 (and up to around 90 vs. serial MrBayes), and more on multi-GPU hardware. GPU MrBayes V3.1 is available from http://sourceforge.net/projects/mrbayes-gpu/.

Key words: **MrBayes, GPU, protein, phylogenetics.**

Brief Communication

The Bayesian method for phylogenetic inference, while computationally intensive, is considered one of the most reliable methods. MrBayes (Huelsenbeck and Ronquist 2001) is a popular program for performing Bayesian phylogenetic inference through Metropolis Coupled Markov Chain Monte Carlo (MC)^3 sampling. It can process DNA, RNA, and protein sequence data; in this work, we focus on protein sequence data.

Over a range of computational domains, researchers are realizing the potential of modern graphics cards, or graphics processing units (GPUs). Here, we present a modified version of MrBayes, which we call ta(MC)^3 (GPU MrBayes V3.1), designed to utilize NVIDIA brand GPUs, and even multi-GPU hardware. This software will be able to run using modern NVIDIA GPU cards with Compute Capability of 2.0 and above. Although this article focuses on protein data sets, these modifications also accelerate inference for DNA and codon data sets (see the supplementary material, Supplementary Material online, for runtime comparisons).

We will compare ta(MC)^3 against: 1) MrBayes (version 3.2.1), which is equipped with the BEAGLE library (Ronquist et al. 2012), which supports GPU accelerated computing; and 2) another modified version of MrBayes, a(MC)^3 (Bao et al. 2013) (see also Zhou et al. 2011), which is a CPU–GPU cooperative program, and ta(MC)^3's predecessor.

Our major contributions are as follows:

- To improve computational efficiency when analyzing large protein data sets, we present an efficient task mapping strategy which makes better use of GPU cores and GPU memory and reduces redundant operations.
- We introduce the option of using 64-bit variables. This enables ta(MC)^3 to process data sets that cannot otherwise be processed (we include two such data sets in our experiments below). Although, since version 3.2, MrBayes has supported some 64-bit variables, it is still incapable of analyzing larger data sets. Specifically, it uses 32-bit memory addresses, and a single such variable can only be used for 4 GB memory, so if the data set requires more than 4 GB memory per GPU to be analyzed, a(MC)^3 will not be able to perform. Modern GPUs, however, have more memory than this.
- We implement Kahan summation to reduce the accumulation of round-off errors. This results in chains converging faster compared with ta(MC)^3 without Kahan summation (some experimental results are provided in the supplementary material, Supplementary Material online).

The technical improvements, for example, task management and GPU memory access procedures, improve the program's runtime a lot; the details of these improvements will be published elsewhere.

Experimental runtimes are listed in table 1. We perform these experiments on protein data sets from a range of animals studied in phylogenetics research. The last two data sets are currently unpublished data sets used by the seventh and eighth authors for their research in of the phylogeny of insects. The platform used has the following specifications: CentOS 6.2; 1 × Intel Xeon E5645 (6 cores; 2.4 GHz); 6 × 4 GB DDR3 1333 RAM; 8 × NVIDIA GeForce GTX Titan.

We can see ta(MC)^3 is consistently at least twice as fast as a(MC)^3, and up to 2.5 times as fast. Further, it consistently outperforms the current version of MrBayes which utilizes the BEAGLE library (listed under MrB + BEAG). The later two data sets in table 1 are unable to be processed by MrBayes or a(MC)^3. When we analyzing data set 7 and 8 with MrBayes (version 3.1.2 and 3.2.1) and a(MC)^3, the programs will crash

Table 1. Runtimes of the Various Programs on Real-World Data Sets.

#Taxa	#Chars	#Unique site patterns	MrB (no GPU)	MrB + BEAG	a(MC)^3	ta(MC)^3	Description	TreeBASE
			Runtime (s); 100,000 generations; 1 GPU					
39	11,445	5,849	52,650	4,113	1,423	675	Lophophorata; ribosomal protein data (Helmkampf et al. 2008)	S2051
48	11,949	6,267	57,025	4,320	1,840	738	Sipuncula; ribosomal protein data (Dordel et al. 2010)	S10271
8	10,088	134	3,744	401	267	117	Insecta; transcriptome protein data (Simon et al. 2009)	S10115
32	9,377	4,782	40,216	3,212	1,183	535	Nemertea; ribosomal protein data (Struck and Fisse 2008)	S1994
85	13,087	6,987	253,953	17,663	6,194	2,919	Arthropoda (Regier et al. 2010)	
59	12,428	7,295	123,656	8,721	3,171	1,508	Myzostomida; ribosomal protein data (Bleidorn et al. 2009)	S10084
			Runtime (s); 1,000 generations; 8 GPUs					
31	360,031	124,382	—	—	—	604	Hexapoda; transcriptome protein data	
14	407,604	73,810	—	—	—	512	Hemiptera; transcriptome protein data	

for they do not use 64-bit variables to allocate memory for some larger data.

To further test the capabilities of ta(MC)^3, we run it on a GPU cluster on the Tianhe-1A supercomputer. We inspect the run-times for 2, 4, 8, and 16 GPUs, and observe a greater benefit of from multi-GPU hardware on larger data sets.

To summarize, we have presented a modified GPU-accelerated version of MrBayes for protein sequence data, which is faster than its predecessors and able to process larger data sets that its predecessors are incapable of processing. The inferred results of ta(MC)^3 are essentially the same with MrBayes version 3.1.2 and 3.2.1 (further details are included in the supplementary material, Supplementary Material online).

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was partially supported by NSF of China (grant numbers: 61373018, 11301288), Program for New Century Excellent Talents in University (grant number: NCET130301), and the Fundamental Research Funds for the Central Universities (grant number: 65141021). R.J.S. was supported by her NSF China Research Fellowship for International Young Scientists (grant number: 11450110409).

References

Bao J, Xia J, Zhou J, Liu XG, Wang G. 2013. Efficient implementation of MrBayes on multi-GPU. *Mol Biol Evol.* 30:1471–1479.

Bleidorn C, Podsiadlowski L, Zhong M, Eeckhaut I, Hartmann S, Halanych KM, Tiedemann R. 2009. On the phylogenetic position of Myzostomida: can 77 genes get it wrong? *BMC Evol Biol*. 9:150.

Dordel J, Fisse F, Purschke G, Struck TH. 2010. Phylogenetic position of Sipuncula derived from multi-gene and phylogenomic data and its implication for the evolution of segmentation. *J Zool Syst Evol Res.* 48:197–207.

Helmkampf M, Bruchhaus I, Hausdorf B. 2008. Phylogenomic analyses of lophophorates (brachiopods, phoronids and bryozoans) confirm the Lophotrochozoa concept. *Proc Biol Sci.* 275:1927–1933.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.

Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* 463:1079–1083.

Ronquist F, Teslenko M, Mark vander P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 61:539–542.

Simon S, Strauss S, Vonhaeseler A, Hadrys H. 2009. A phylogenomic approach to resolve the basal pterygote divergence. *Mol Biol Evol.* 26:2719–2730.

Struck TH, Fisse F. 2008. Phylogenetic position of Nemertea derived from phylogenomic data. *Mol Biol Evol.* 25:728–736.

Zhou J, Liu X, Stones DS, Xie Q, Wang G. 2011. MrBayes on a graphics processing unit. *Bioinformatics* 27:1255–1261.