Bioinformation by Biomedical Informatics Publishing Group *open access*
www.bioinformation.net **Software**

PBEAM: A parallel implementation of BEAM for genome-wide inference of epistatic interactions

Tao Peng, Pufeng Du, Yanda Li*

MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China; Yanda Li - E-mail: daulyd@tsinghua.edu.cn; *Corresponding author

received February 20, 2009; accepted March 28, 2009; published April 21, 2009

Abstract:
The software tool PBEAM provides a parallel implementation of the BEAM, which is the first algorithm for large scale epistatic interaction mapping, including genome-wide studies with hundreds of thousands of markers. BEAM describes markers and their interactions with a Bayesian partitioning model and computes the posterior probability of each marker sets via Markov Chain Monte Carlo (MCMC). PBEAM takes the advantage of simulating multiple Markov chains simultaneously. This design can efficiently reduce ~n-fold execution time in the circumstance of n CPUs. The implementation of PBEAM is based on MPI libraries.

Availability: PBEAM is available for download at http://bioinfo.au.tsinghua.edu.cn/pbeam/.

Keywords: genome wide association study, epistatic mapping, Bayesian methods, parallel computing

Background:
Many genes that are responsible for Mendelian diseases have been successfully mapped during recent decades. In recent years, the number of polymorphisms considered in genetic association study has increased dramatically as the high-throughput genotyping technologies develop rapidly. This increase offers a unique chance to identify genes for complex traits through an unbiased search at a genome-wide level **[1]**. Recent studies have revealed more than 50 novel susceptible loci for many complex diseases **[2],** including obesity, age-related macular degeneration and heart disease. However, genetic polymorphisms underlying most common disease are epistatic: these variations interact with each other and interact with environment factors in a complex way. Detecting and characterizing these interactions is crucial for our understanding of these common and complex diseases. Several approaches to detect epitasis have been developed (e.g. MDR **[3],** MPVA **[4]**, BGTA **[5]**). These approaches all showed promise on small data sets, but they are quickly overwhelmed by the genome-scale data. A Monte Carlo simulation based approach, called BEAM (Bayesian Epitasis Association Mapping), which was recently introduced by Zhang and Liu (2007), can infer possible epistatic interactions among a large number of polymorphisms. In this paper, we present PBEAM (Parallel BEAM), which is an improved and parallelized version of BEAM, to facilitate the genome-wide epistasis mapping.

Methodology:
Genome-wide scale epistatic interaction inference:
BEAM [6] proposed a Bayesian partition model to describe the disease-associated markers and their interactions. For a population-based association study, all the markers across the whole genome are partitioned into three groups: group 0 contains markers unrelated to the disease, group 1 contains markers contributing independently to the disease risk and group 2 contains markers that jointly affecting the disease risk. Thus, the association study problem is transformed into inferring the composition of disease sets, that is, which markers belongs to group 1 and group 2. Given the genotype in the case and control populations, the likelihood of a possible partition I (the assignment of each marker to group 0, 1 and 2) can be theoretically calculated. With the likelihood function, the posterior probability that each marker set is associated with the disease can be estimated with Markov Chain Monte Carlo (MCMC) simulation.

The goal of the MCMC procedure is to draw the partition I from the distribution conditioned on observed case and control data. Partition I was initialized according to a simple theoretical prior. Metropolis-Hastings (MH) algorithm was then used to update the partition I. The authors used two types of update proposals: (1) randomly switch a marker's group membership (e.g. from 0 to 1 or from 1 to 2); (2) randomly exchange two markers between different groups. As a standard step in MH algorithm, the proposed update is accepted according to the likelihood ratio. BEAM algorithm also takes some annealing-like techniques to improve the sampling performance. The many samples of I drawn by MCMC can finally serve as Bayesian-style estimation or the potential 'hits' (the most probable partitions). The likelihood-based simulation avoids the computationally infeasible enumeration of all the possible combinations of makers, but is still computationally intensive. The execution time of drawing millions of samples from a MCMC chain, which is necessary for a stable estimation, has hampered some of the further studies, like permutation analysis.

Parallelization of the Bayesian algorithm:
To reduce the execution time of BEAM, we developed a parallel version, called PBEAM. The design of PBEAM takes the advantage that samples can be drew from several Markov chains simulated simultaneously on different computers. We analyzed the source code of BEAM and adapted the algorithm framework and data structure of BEAM to the distributed memory parallelism. In the PBEAM, the jobs of Monte Carlo simulations are

distributed to different processes using MPI, which can be executed on various heterogeneous platforms.

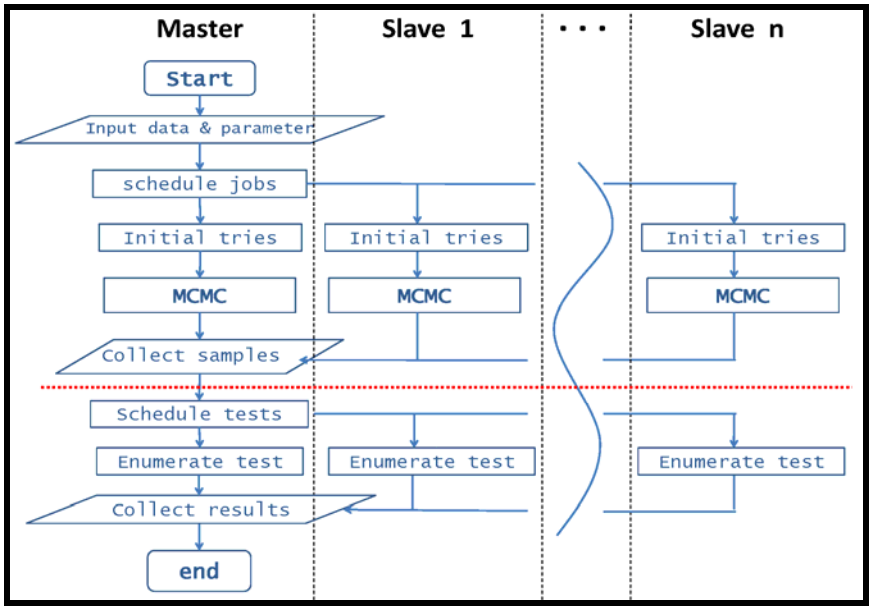

Figure 1: Parallelization scheme of PBEAM.

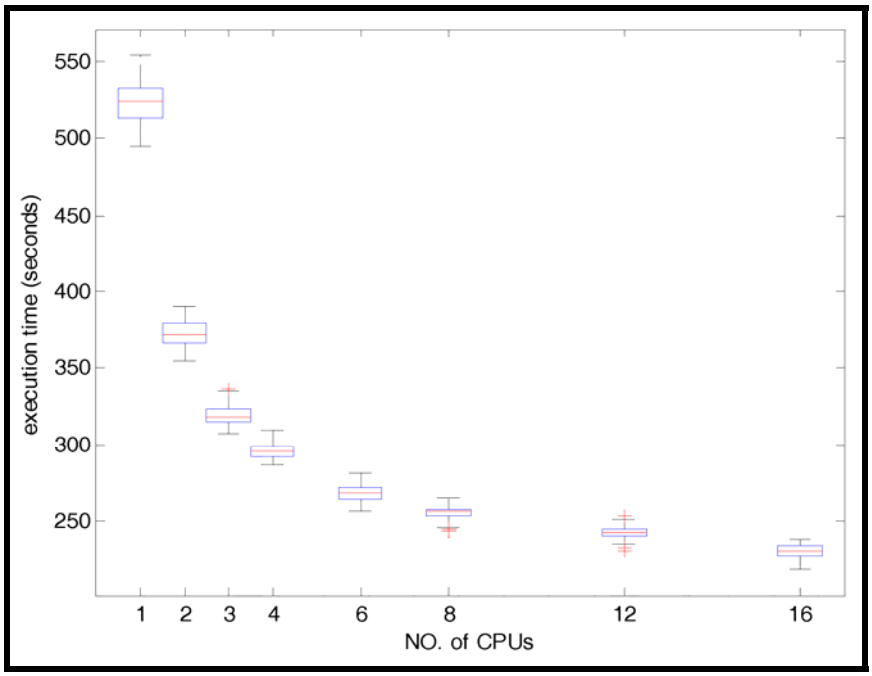

Figure 2: Average execution time for 106 iterations on dat sets of 1000 loci, 400 cases and 400 controls

We have optimized and parallelized the PBEAM program with master/slave architecture. The Master process schedules and distributes blocks of chains to the slave process, which actually perform the tasks (Fig. 1). After sampling, the master collects all the samples from slaves and adaptively schedules the calculation of B-statistics [6] on all the combination of possible disease-associated markers. Final reports are based on the results across all process.

Utility:

There is a wide spectrum of epistatic interaction models for disease-associated markers. We chose three common two-loci models (additive, multiplicative and multiplicative with threshold [7]) and simulated three data sets (1000 loci,

400 cases and 400 controls) for each model. Each data set was performed on 1 (BEAM) and multiple (PBEAM) processors of an AMD 2GHz Opteron cluster. The execution time of 10 independent repetitions for each number of CPUs are averaged for comparison. This experiment design reflects two concerns: 1) the speedup performance of PBEAM; 2) the results consistence of BEAM and PBEAM, considering result variance is intrinsic for stochastic sampling algorithms.

The parallel program PBEAM presents an almost perfect speedup. The communication cost between processes and the initial tries of each Markov chain lead to only a slight performance loss (Fig 2). The result variance of PBEAM and BEAM are in comparable range. This performance improvement is crucial for large dataset of the state-of-the-art genome wide association studies. A typical dataset[8] of these researches consists of ~4000 individuals (e.g. 2000 cases and 2000 controls) and each individual is genotyped at ~ 500K markers. For a dataset of this size, BEAM took 27.5 days to draw 10^9 samples from one Markov chain. With 16 processors, PBEAM could finish the inference procedure within 2 days.

Caveat and future development:
The detailed information about the usage of PBEAM source code and executables could be found in the readme and website. Currently, PBEAM chains treat the MCMC local mode issues independently. Future development could include determining the burn-in time and merging the samples collectively between chains.

Acknowledgements:
The authors would like to thank Wanwan Tang and Jin Gu for preparing simulation data and providing the helpful comments on PBEAM code, respectively. This work was supported by National Science Foundation of China (60775002 and 60572086).

References:
[1] J. N. Hirschhorn, M.J. Daly, Nat Rev Genet, (2005). **6**: 95-108 [PMID: 15716906]
[2] M. L. McCarthy, et al., Nat Rev Genet, (2008). **9**: 356-69 [PMID: 18398418]
[3] M. D. Ritchie, et al., Am J Hum Genet, (2001). **69**: 138-47 [PMID: 11404819]
[4] X. Sun, et al., Hum Hered, (2005). **60**: 143-9 [PMID: 16319491]
[5] T. Zheng, et al., Hum Hered, (2006). **62**: 196-212 [PMID: 17114886]
[6] Y. Zhang, J.S. Liu, Nat Genet, (2007). **39**: 1167-73 [PMID: 17721534]
[7] N. Risch, Am J Hum Genet, (1990). **46**: 222-8 [PMID: 2301392]
[8] T. W. T. C. C. Consortium, Nature, (2007). **447**: 661-78 [PMID: 17554300]

Edited by P. Kangueane
Citation: Peng et al. Bioinformation 3(8): 349-351 (2009)