PRoMT: inferring demographic history from DNA sequences

Stephen Wooding

Department of Anthropology, University of Utah, 270 South 1400 East, Salt Lake City, UT 84112-0060, USA

Received on March 28, 1999; revised and accepted on October 4, 1999

Abstract

Summary: *I describe a parallel implementation of Rogers' mismatch algorithm, a method for making inferences about demographic history from DNA sequence data. The program is distributed on clusters of workstations, providing a substantial speedup and low execution times on large numbers of nodes.*

Availability: *Source code and documentation are available at http://mombasa.anthro.utah.edu/wooding/*

Contact: *stephen.wooding@anthro.utah.edu*

A common method for characterizing patterns of genetic variability in homologous DNA segments gathered from populations is to calculate the distribution of pairwise nucleotide differences between them—the mismatch distribution. That is, for a set of seqences, it is informative to examine the frequency distribution of pairs that differ at $i = 0, 1, 2, \ldots, k$ nucleotide positions. Simulations and mathematical models predict that mismatch distributions will be sensitive to both the timing and magnitude of population growth.

Rogers (1995) developed a theoretical framework to allow the extraction of demographic information from extant patterns of genetic diversity, as well as an accompanying computer program for analyzing data, mmci. Rogers' approach constructs a confidence interval by subjecting a sample dataset to many comparisons, each constituting a separate statistical analysis. This tool has been informative in a number of contexts (e.g. Rogers, 1995; Wooding and Ward, 1997), however its use has been limited partly due to its high computational demands. To improve the speed of execution of the mismatch test, and hence its power and flexibility, I have developed a parallel implementation of the algorithm called PRoMT (Parallel Rogers Mismatch Test).

PRoMT is a modification of Rogers' original algorithm that uses message passing to distribute the hypothesis testing process over an arbitrary number of networked workstations. The program is scalable and portable. Though designed and tested on a pair of desktop computers, it runs without modification on clusters with more complex organizations including 4- and 32-node systems. PRoMT's performance increases the rate at which Mismatch's subtests can be performed, improving the resolution and scope of the test as a whole while maintaining its availability to a broad spectrum of users.

PRoMT was adapted directly from the source code of mmci, Rogers' program for performing mismatch tests packaged with Mismatch version 4.2 (Rogers, 1999). Changes to mmci were minimized to allow future modifications to mmci to be incorporated easily into the PRoMT parallel framework. Message passing was accomplished using the MPICH 1.1 implementation of MPI (Message Passing Interface), a free, standardized API for initiating and controlling processes on several machines simultaneously (Gropp *et al.*, 1994; Message Passing Interface Forum, 1994).

The organization of Mismatch into a series of discrete statistical tests made parallelization straightforward. The task consisted mainly of orchestrating the distribution of small elements of the larger test to slave processes, and collating results at the end. Finally, PRoMT was compiled using the tools included with the MPICH distribution and the Gnu C compiler (gcc).

PRoMT's performance was measured using wall clock run times, which give an indication of the program's practical usage. Speedup, which was calculated by dividing the wall clock run time of Rogers' unmodified mmci program on one of a cluster's slave nodes by the wall clock run time of PRoMT on a varying number of slave nodes, was used to measure the performance of PRoMT relative to mmci. Benchmarking measurements were made on a 32-node cluster composed of 350 MHz Pentium II computers. The example dataset included with Mismatch (Rogers, 1999) was used in all benchmarks.

Speedup achieved with increasing cluster size was substantial. The ratio of speedup to number of nodes was approximately 0.85 (Figure 1). This pattern is consistent with PRoMT's effective load distribution and the presence of a low communications overhead. The departure of PRoMT's speedup from linearity appears to be due to

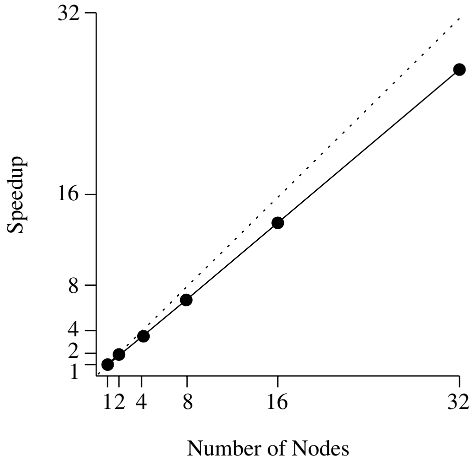

Fig. 1. Relationship between cluster size and speedup in PRoMT on a 32-node cluster. Speedup was measured by dividing the run time of Rogers' unmodified mmci program running on a single node by the run time of PRoMT running on *n* nodes.

start-up, which includes the addition of an approximately constant amount of time for each participating node.

The total execution time of PRoMT was reduced noticeably by increasing node numbers (6.5 min on 32 nodes versus 2.5 h on one node). This level of performance is indicative of the program's usefulness: with a modest outlay of network and node preparation a high level of performance can be achieved. In actual usage, the program has provided a number of benefits. In addition to increases in test speed, the program has allowed an increase in project turnaround time. Adjustments to parameters used in individual analyses, as well as comparisons between different datasets, can be performed quickly to provide an improvement in result quality as well as speed.

Acknowledgements

The author wishes to thank Alan Rogers, Henry Harpending and Brian Haymore for helpful comments. This work was supported by a NIH Genome Sciences training grant and a grant of computer time by the Center for High Performance Computing at the University of Utah.

References

Gropp,W., Lusk,E. and Skjellum,A. (1994) A high-performance, portable implementation of the MPI message passing interface standard. http://www-unix.mcs.anl.gov/mpi/mpich.

Message Passing Interface Forum (1994) MPI: A message-passing interface standard. *Int. J. Supercomp. Appl.*, **8**.

Rogers,A.R. (1995) Genetic evidence for a Pleistocene population explosion. *Evolution*, **49**, 608–615.

Rogers,A.R. (1999) *Mismatch 4.2*. Computer program available from the author. Department of Anthropology, University of Utah, Salt Lake City, Utah, USA.

Wooding,S. and Ward,R.H. (1997) Phylogeography and Pleistocene evolution in the North American black bear. *Mol. Biol. Evol.*, **14**, 1096–1105.