

Parallel numerical methods for inferring Phylogenies

Felipe Fernandes Albrecht IME/EB - Rio de Janeiro, RJ - Brazil, felipe.albrecht@gmail.com

Nelson Borges IME/EB - Rio de Janeiro, RJ - Brazil, nborges@de9.ime.eb.br

Abstract

Recent years have shown a sensible increase in the search for exact and complete phylogenetic trees with employment of numerical techniques. But the huge amount of computing load needed to reach such trees with a good accuracy and a high level of taxons leads to considering parallel processing as the suitable way to tackle this problem, [1,2]. In this work we deal with phylogenetic methods with distance matrices for a large amount of taxons. These methods were first introduced by Cavalli-Sforza and Edwards[3] and by Fitch and Margoliash[4]. Other methods have stepped on them, as was the case of those ones developed by Wolf Y.I. *et al.*[5] and Clarke G.D. *et al.*[6]. Observe that the method due to Grishin N.V. *et al.*[7], as well as that one by Cavalli-Sforza[3], both require the solution of linear systems of algebraic equations with dimensions that grow linearly with respect to the number of taxons.

We have used parallel computing techniques so as to reach shorter processing time for the solution of the systems associated to the length of the branches of the phylogenetic trees, the distance matrices method being employed.

In Albrecht *et al.*[10] an algorithm was proposed: phylogenetic trees are built and distributed through a group of processes, where at each iteration the best trees are identified, then kept. This algorithm, which is based in Felsenstein[8]. In each process the trees construction needs synchronization only to discard those trees with the worst results. These algorithms compute the length of the branches through a process called pruning. The analysis of the implementations carried over by [10] and [8] with the tool GNUProf in [11] has lead to the conclusion that the pruning process for the branches was the main time processing spending for these software. Another difficulty this process leads to is that to have it in parallel, a huge cost is asked by the synchronizing task. Our aim is thus to use different numerical methods for the solution of the linear systems required by the least-squares methods to get the branches length, as described in [12].

The matrix associated to the least-squares method, being symmetric and positive-definite, allows the use of the conjugated gradient method (CGM). We know that in the worst case convergence is guaranteed in n steps, being n the linear system order. Since each iteration for the CGM requires $n^2 + O(n)$ multiplications and divisions, we deduce that in this case we need $n^3 + O(n^2)$ operations, which is the same amount needed for a complete inversion of the matrix with the LL^t decomposition.

In order to improve convergence for CGM, we get hold of the matrix profile. It turns out that the matrices that appear are all block-matrices, being all blocks also sym-

metric and positive-definite. We can then use the Preconditioned Conjugate Gradient method with block Jacobi as a preconditioner – this amounts to incomplete block decomposition. It is known that Jacobi method has a rather slow convergence but, when employed as a preconditioner, overall convergence is speeded up. Besides, parallelization of block-Jacobi preconditioned conjugate gradient method (BPCGM) may be obtained very efficiently.

We first used the sequential versions of block-Jacobi method (JBM), conjugate gradient (CGM) and preconditioned conjugate gradient with block-Jacobi as preconditioner (BPCGM). It turns out that convergence is much better for the latter, although its speed-up with respect to the second one is rather small. It seems encouraging to use a parallel version of BPCGM. Nevertheless we must take in account its rather high data transmission rate as well as the additional cost to invert the block matrices B_k required for the algorithm start-up. Anyway, we can claim that as long as inversion of the matrix A is untractable, the parallel version of BPCGM is available for the rescue.

References

- [1] Keane T.M. *et al.* DPRml: distributed phylogeny reconstruction by maximum likelihood. *J. Bioinformatics*, 21, No. 7, pp 969-974, 2005.
- [2] Stamatakis A. *Distributed and parallel algorithms and systems for inferring of huge phylogenetic trees based on the maximum likelihood method.*, Ph.D. thesis, Technischen Universitt Mnchen, 132 p., 2004.
- [3] Cavalli-Sforza, L.L. and Edwards, A.W. Phylogenetic analysis: models and estimation procedures. *Am. J. Human Genetics*, 19(3), p. 233-257, 1967.
- [4] Fitch, W.M. and Margoliash, E. Construction of phylogenetic trees. *Science*, Science, 760(157), p.279, 1967.
- [5] Wolf Y.I. *et al.* Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.*, 1, p. 8, 2001.
- [6] Clarke G.D. *et al.* Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLAST scores. *J. Bacteriol.*, 184(8), p. 2072-2080, 2002.
- [7] Grishin N.V. *et al.* From complete genomes to measures of substitution rate variability within and between proteins. *Genome Res.*, 10, p. 991-1000, 2000.
- [8] Felsenstein, Joseph. An alternating least squares approach to inferring phylogenies from pairwise distances. *Systematic Biology*, 46(1), p. 101-111, mar. 1997.
- [9] David W. Mount. *Bioinformatics: sequence and genome analysis Cold Spring Harbor Laboratory Press*, 2004. 692 p.
- [10] Albrecht, F.F.; Hubner, J.F. and Davila, A. A Distributed Algorithm for Phylogenetics Inference. *BSB2007 Poster Proceedings*, 2007.
- [12] Felsenstein, Joseph. *Inferring phylogenies. Massachusetts: Sinauer Associates*, 2003. 580 p.