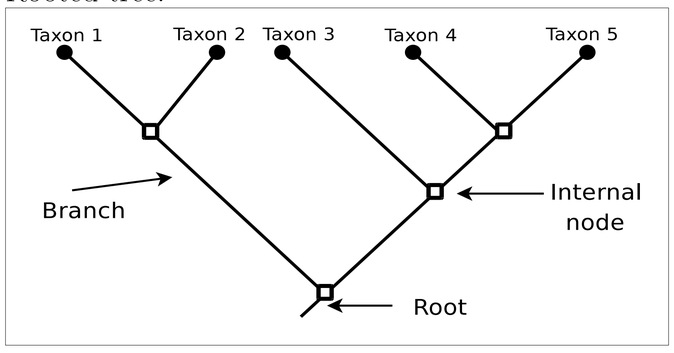
A Distributed Algorithm for Phylogenetics Inference

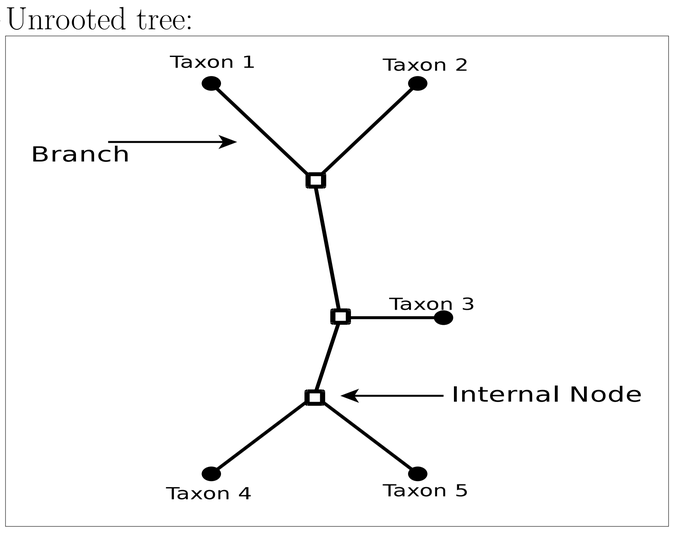
Felipe Fernandes Albrecht — Instituto Militar de Engenharia
Jomi Fred Hübner — Universidade Regional de Blumenau
Alberto M. R. Dávila — Instituto Oswaldo Cruz, FIOCRUZ

Introduction

- Molecular phylogeny is used to gain information on an organism's evolutionary relationships.
- The result of a molecular phylogenetic analysis is expressed in a so-called phylogenetic tree.
  - Rooted tree:



  - Unrooted tree:



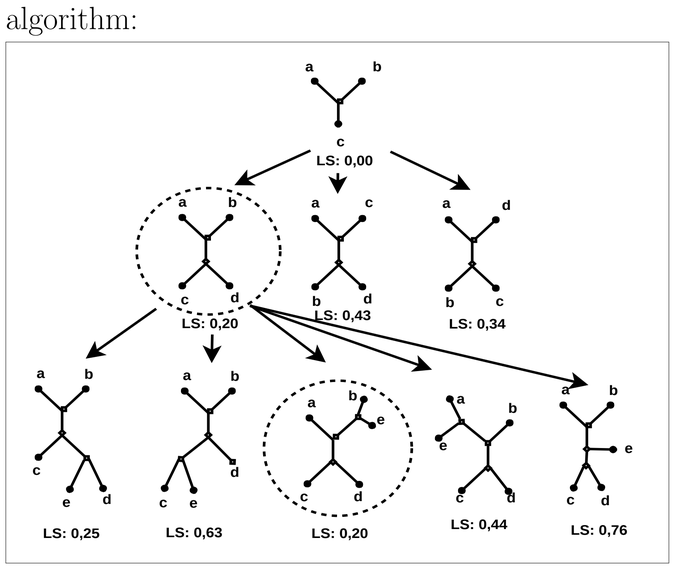
- Two main phylogenetics methods:
  - Molecular sequence:
    - maximum likelihood
    - maximum parsimony methods.
  - Distance matrix:
    - Neighbor-Joining
    - UPGMA
    - least squares methods

Motivation

- This work is based on least squares method, that is a distance matrix method where an unrooted tree is returned as result.
- This method has an objective function (equation 1) that represents the inferred tree quality and this method

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(D_{ij} - d_{ij})^2 \qquad (1)$$

- The alternating least squares algorithm proposed by Felsenstein[3] is slower than others distance matrix methods.
- This method creates trees choosing the bests, considering the least Squares, like a search algorithm:



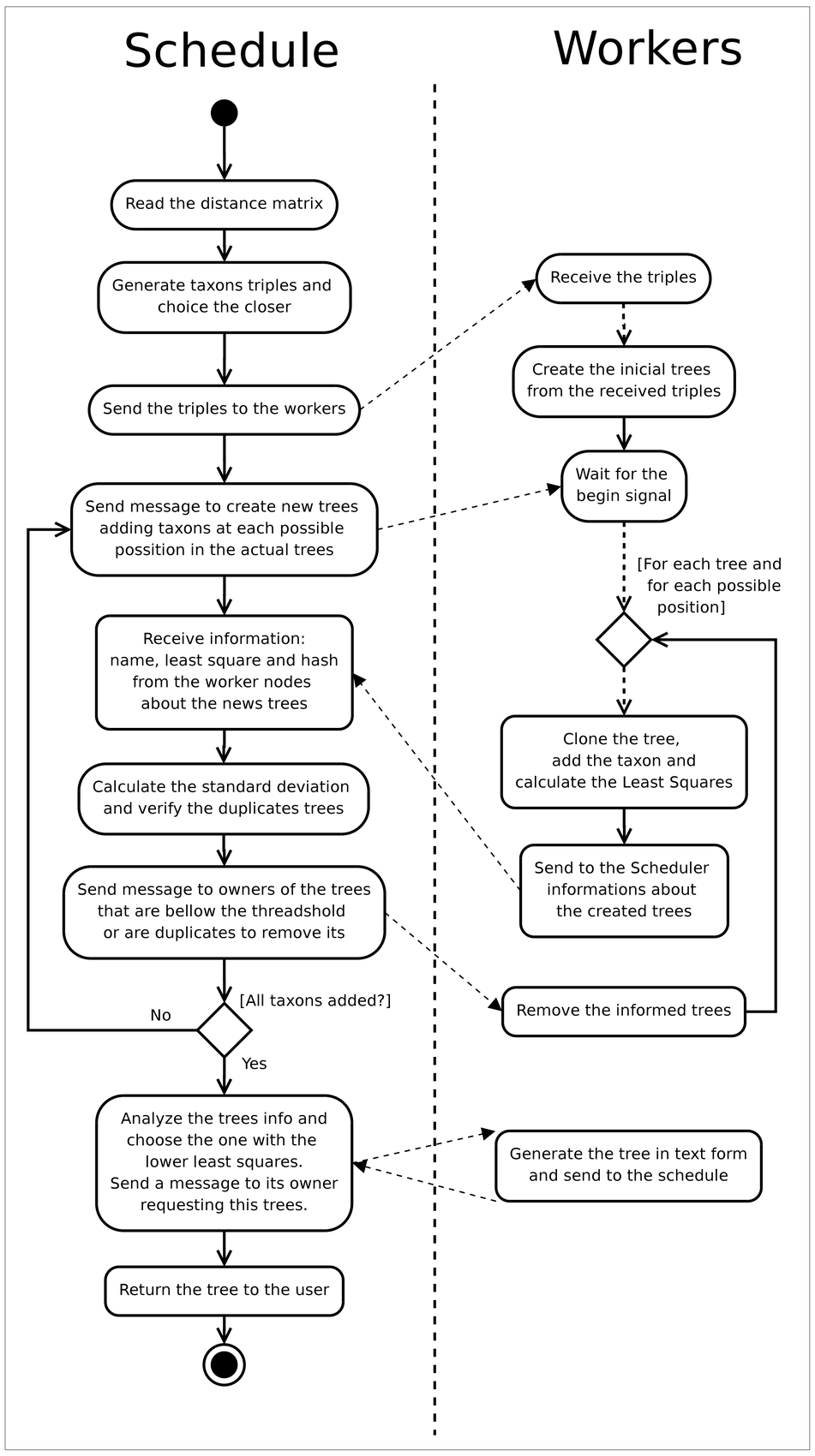
- The process: to create trees, to choose the bests, can be parallelized to improve the search time.

Objectives

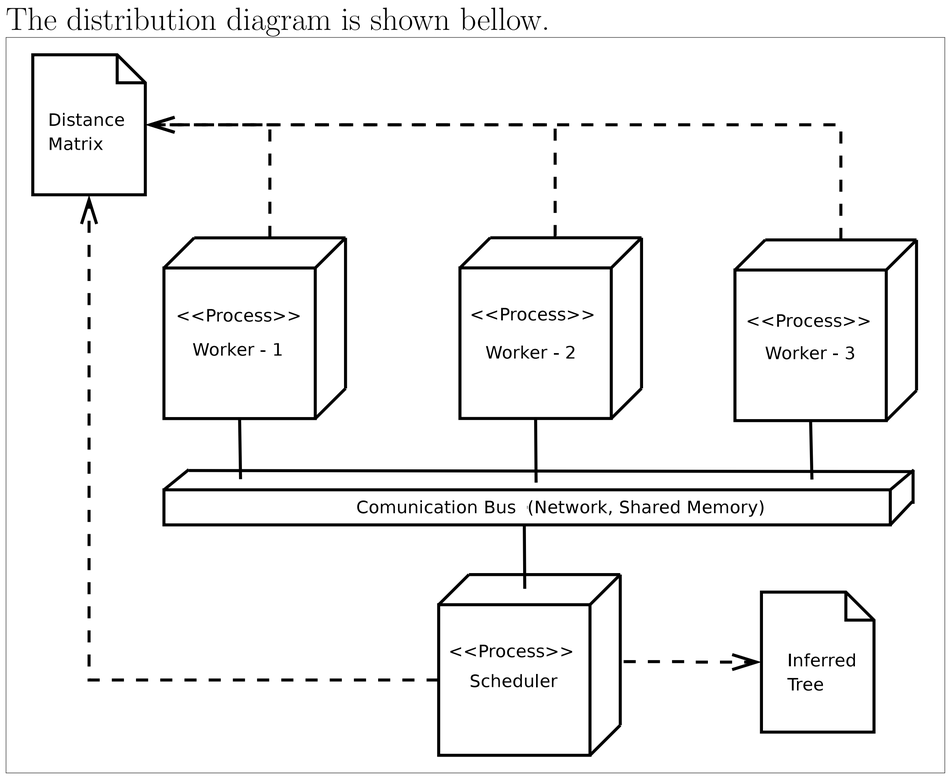
- To parallelize the algorithm proposed by Felsenstein[3]
- To give to the user options to increase or to reduce the search space.
- With more machines working, to reduce the execution total time with minor changes in the trees quality.

The algorithm

The simplified version of the algorithm is shown bellow.



The distribution diagram is shown bellow.



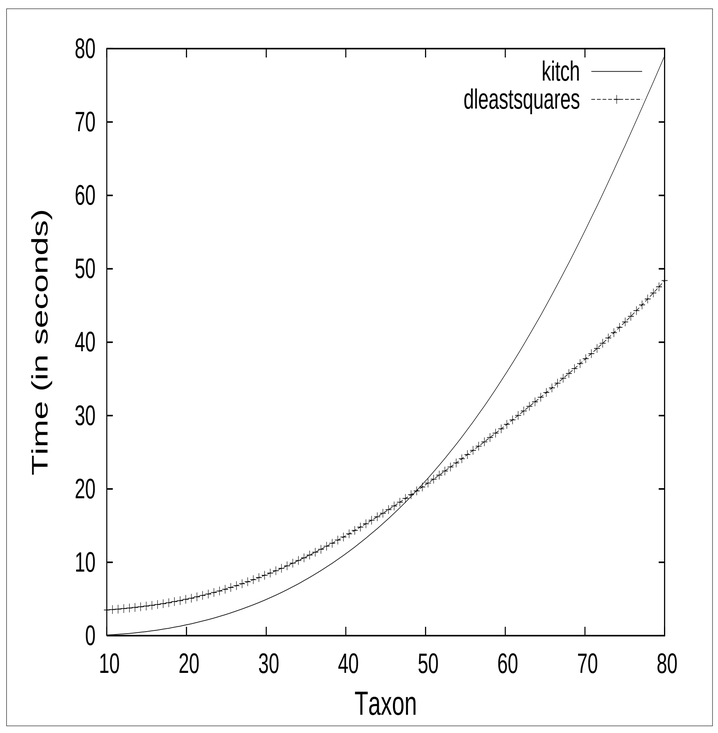
Implementation

- The algorithm is implemented in C language, with gcc[5] compiler in Linux environment.
- The MPI[6] standard with LAM[2] implementation is used for multiprocess communication.
- For multiprocessor, not multicomputer cluster, the algorithm can be implemented using Threads and shared memory.
- The resulting tree can be shown in the drawtree software, from the PHYLIP[4] package.
- The implementation software is called dleastsquares.

Results

- The dleastsquares and the kitch was executed 8 times and the matrix sizes varies from 10 taxons to 80 taxons.
- It is shown in figure below the execution time.
- Note that for a matrix with less than 50 taxons, the kitch performance is better.
- Otherwise, with matrices with more than 50 taxons, the dleastsquares performance outperforms the kitch.
- With a matrix with 80 taxons, the dleastsquares shows a time gaim of 50

The time per node count is shown bellow.



Conclusions and Related Work

- This work proposes a distributed version for least squares method.
- It is shown a time gain when the taxon number is greater or equal 80.
- The main problem is the tree inferred quality, that it is worst than sequencial implementation.
- Another distribution can be done parallelizing the branchs length, like in Albrecht and Borges [1].
- Searching the literature, it was not possible to find another distributed version for the least squares method.

References

[1] F. F. Albrecht and N. Borges. Parallel numerical methods for inferring phylogenies. In International Congress on Industrial and Applied Mathematics, Zurich, 2007.

[2] G. Burns, R. Daoud, and J. Vaigl. Lam. In Proceedings..., pages 379–386, Toronto, 1994. Supoercomputing Symposium'94, University of Toronto.

[3] J. Felsenstein. An alternating least squarers approach to inferring phylogenies from pairwise distances. Systematic Biology, 46(1):101–111, mar. 1997.

[4] J. Felsenstein. Phylip (phylogeny inference package) version 3.6, 2005.

[5] F. S. Foundation. Gcc, the gnu compiler collection, 2006.

[6] MPI. Message passing interface, 2006.