In: Proceedings of 3rd. Brazilian Workshop on Bioinfomatics, Brasília (DF), 20-22/october, [CD-ROM], 2004.

Reconstruction of Phylogenetic Trees using the Ant Colony Optimization Paradigm

Mauricio Perretto, Heitor S. Lopes

Laboratório de Bioinformática / CPGEI
Centro Federal de Educação Tecnológica do Paraná - CEFET-PR
Av. 7 de setembro, 3165 – 80230-901 Curitiba (PR) – Brazil
tel. +55-41-310-4694, fax. +55-41-310-4683
http://bioinfo.cpgei.cefetpr.br
e-mails: mperretto@unicenp.br, hslopes@cpgei.cefetpr.br

Abstract. This work presents a new approach for the reconstruction of phylogenetic trees using the Ant Colony Optimization (ACO) metaheuristics. A tree is constructed using a fully-connected graph and the problem is approached similarly as the well-known Traveling Salesman Problem. The methodology is detailed as well the algorithm for constructing a phylogenetic tree using a pheromone matrix. Tests were done using two data sets: complete mitochondrial genomes from mammalians and DNA sequences of the p53 gene from several eutherians. Results show that the proposed methodology is competitive with other well-known softwares. These results are very promising and suggest more efforts for further developments.

1. Introduction

Phylogenetic trees are based on the Darwinian principle of the natural evolution of species. They aim at unveiling the evolutionary relationship among species. That is, when analyzing a set of amino acid sequences (or proteins) of different species it will be determined how these species might have been derived during their evolution. Reconstruction of phylogenetic trees is an important problem in Bioinformatics and, like others, it is still an open subject for research. This is mainly due to NP complexity of the problem [1] that leads to intractable search spaces when dealing with the phylogeny of a large number of species.

In a simple way, a phylogenetic tree can be considered a binary tree, where leaf nodes represent the species to be analyzed and inner nodes the ancestor species from which the current species have evolved. Besides, phylogenetic trees can or cannot have a root (see Fig. 1) that indicates the oldest ancestor. Usually, a rooted tree represents better the phylogenetic history of species. In the other hand, an unrooted tree represents a better correlation between species.

Considering n species, it is possible to construct NT different trees. Equation 1 shows how NT can be computed for both rooted and unrooted trees. For instance, if we would like to find the best tree using the method of maximum similarity for (only) 15 species, we should try 213,458,046,676,875 trees.

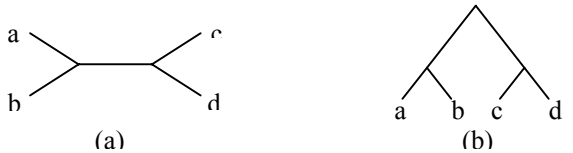

Fig. 1. Topologies of phylogenetic trees: (a) unrooted tree, (b) rooted tree.

$$NT = \begin{cases} \dfrac{(2n-3)!}{2n-2(n-2)!} & \text{for unrooted trees} \\ \dfrac{(2n-5)!}{2n-3(n-3)!} & \text{for rooted trees} \end{cases} . \quad (1)$$

Current methods for the reconstruction of phylogenetic trees can be roughly grouped in two families: feature-based methods and distance-based methods. Feature-based methods, such as maximum parsimony and maximum likelihood [2], use as input previously aligned sequences of nucleotides and they are less susceptible to errors. In the other hand, distance-based methods, such as UPGMA (Unweighted Pair Group Method using arithmetic averages) [3] and Neighbor Joining [4], use a matrix representing the distances between pairs of species and they are based in the principle of similarity.

2. Ant Colony Optimization

Social insects that live in colonies, such as ants, termites, wasps and bees develop specific tasks according to their role in the colony. One of the main tasks is the search for food. Real ants, when searching for food, can find out such resources without visual feedback (they are practically blind) and they can adapt to changes in the environment by optimizing the path between the nest and the food source. This fact is the result of stigmergy, that is, positive feedback, given by the continuous deposit of a chemical substance known as pheromone.

A classic example of the construction of a pheromone trail in the search for a shorter path is shown in Fig. 2 and was first presented by [5]. In Fig. 2a there is a path between food and nest established by the ants. In Fig. 2b an obstacle is inserted in the path and, soon, ants spread to both sides of the obstacle, since there is no clear trail to follow (Fig. 2c). As the ants go around the obstacle and find again the previous pheromone trail, a new pheromone trail will be formed around the obstacle. This trail will be stronger in the shortest path than in the longest path, as shown in Fig. 2d.

As shown in [6], there are many differences between real ants and artificial ants, mainly: artificial ants have memory, they are completely blind and time is discrete. On the other hand, an ACS (Ant Colony System) allows the simulation the behavior of real-world ant colonies, such as: artificial ants have preference for trails with larger amounts of pheromone; shorter paths have a stronger increment in pheromone; there is an indirect communication system between ants, the pheromone trail, to find the best path.

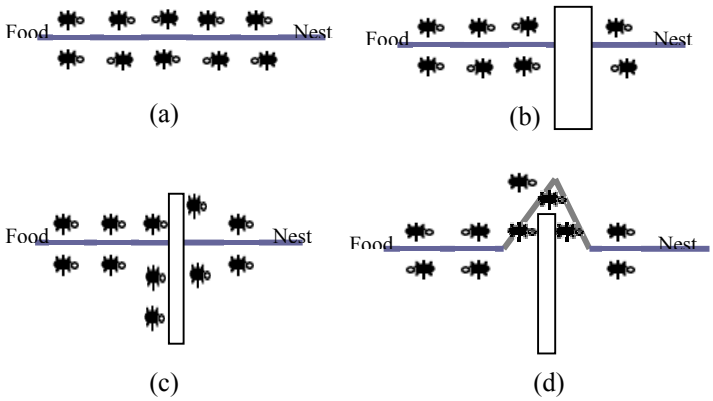

Fig. 2. a) Ants in a pheromone trail between nest and food; b) an obstacle interrupts the trail; c) ants find two paths to go around the obstacle; d) a new pheromone trail is formed through the shorter path.

3. Related Work

Korotensky and Gonnet [7] presented an alternative method named circular sum, for obtaining the sequence of branches that will give the smallest tree. This method models the problem as a circular Traveling Salesman Problem (cTSP), that is, for a complete tour, the distance from the last city and the first one is added to the tour distance. The tour corresponds to the sequence of species, and the tour distance is the smallest score for this sequence. To construct the tree, a simple idea is used: the correct tree will have the same score found by means of the cTSP. This way, a second algorithm is done, constructing trees and comparing their scores with the one found by cTSP. This search method is somewhat similar to the maximum parsimony, and, thus, requests a large computational effort for constructing a phylogenetic tree for a large number of species.

Kumnorkaew et al [8] present a new strategy for constructing trees. In the algorithm, a preprocessing step defines a number of intermediary nodes, by means of the intersection of the input species, which are the ancestor species. From this point on, input species are considered source nodes and the intermediary nodes as compulsory passing points. This strategy is similar to the well-known Steiner problem. Those authors report that equivalent trees were obtained to those constructed using the Neighbor-joining method. However, it is necessary strong preprocessing to define proper intermediary points that are underused.

4. The ACO-based Model

To define how ACO will be applied to the reconstruction of phylogenetic trees, we used a fully connected graph, constructed using the distance matrix among species.

(Fig. 3). In this graph, nodes represent the species and edges represent the evolutionary distances between species.

Species	1	2	3	4
1	0.000	0.199	0.837	0.830
2	0.199	0.000	0.893	0.886
3	0.837	0.893	0.000	0.009
4	0.830	0.886	0.009	0.000

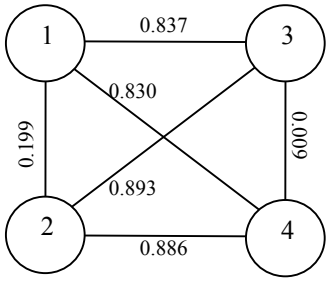

Fig. 3. Distance matrix for four species ant the corresponding graph.

In the beginning, ants start in a randomly selected node. Then, they travel across the structured graph and at each node a transition function (Equation 2) determines its direction. This equation represents the probability that the *k*-th ant, being at node *i*, goes to node *j* in its next step.

$$P_k(i,j) = \frac{[\tau(i,j)]^{\alpha} . [1/d(i,j)]^{-\beta}}{\sum_{u \in J_i^k} \left([\tau(i,u)]^{\alpha} . [1/d(i,j)]^{-\beta} \right)} \quad . \tag{2}$$

where: $P_k(i,j)$ is the probability of transition between node *i* and *j*; τ is the pheromone trail between two nodes; $d(i,j)$ is the evolutionary distance between nodes *i* and *j*; J_i^k is the set of nodes connected to node *i* and already visited bye the *k*-th ant; α and β are arbitrary constants.

Equation 2 is composed by two terms: the first is based on the evolutionary distance between species *i* and *j*, and the second is based on the accumulated experience - the pheromone trail. This trail is represented as matrix (like that for the distance between species), whose values are dynamically changed by the algorithm, and determined according to the paths chosen by ants. Therefore, $\tau(i,j)$ represents the attractiveness of node *j*, while the ant is at node *i*. Therefore, the objective of a given ant is to find a path in the graph that maximizes the transition probabilities, thus obtaining a sequence of species that produces the smallest evolutionary distance.

Differently of a traditional ACO, where moves are done between nodes, in this work we create an intermediary node between the two previously selected ones. This node will represent the ancestor species of the other two, and it will not be in the list of nodes (species) to be set in the tree. Using such intermediary node, distances to the remaining nodes (species) are recomputed by means of Equation 3.

$$d_{nu}(i,j) = \begin{cases} d(i,u) + [d(i,u) - d(j,u)].\eta, & \text{if} \quad d(j,u) \geq d(i,u) \\ d(j,u) + [d(j,u) - d(i,u)].\eta, & \text{if} \quad d(i,u) > d(j,u) \end{cases} . \tag{3}$$

where *u* is a node that does not belongs to the set of nodes connected to node *i* and already visited by the *k*-th ant (J_i^k); $d_{nu}(i,j)$ is the distance between the new node *n* and node *u,* based on the previous distances between *(i,u)* and *(u,j)*; $d(i,u)$ is the distance

between nodes i and u; η is a scale constant that defines the distance between the new node n and its descendents i and j.

The previous procedure is repeated until all nodes belong to the list of already visited nodes, and then, a path is constructed. The score of this path is given by the sum of the transition probabilities of the adjacent nodes of the path.

Paths constructed by the ants are used for updating the pheromone trail. The increment of the pheromone trail is done in all nodes belonging to at least one path, created in an execution cycle. This key point avoids fast convergence to a local maximum. The pheromone trail matrix is updated according to equation 4:

$$\tau(i, j) = \rho \cdot \tau(i, j) + (1 - \rho) \cdot \Delta\tau(i, j) \ . \tag{4}$$

where ρ is the rate of evaporation of the pheromone, which reduces the persistence of the environment to the ants. In this work, the rate of increment of pheromone, $\Delta\tau(i,j)$, was modified to allow an increment proportional to all the obtained paths, given by the division of the current path and the best path, as shown in Equation 5:

$$\Delta\tau(i,j) = \begin{cases} \sum_{t=0}^{k} S_{c(t)} \cdot (S_{best})^{-1} & \text{, if } (i,j) \in c(t) \ . \\ 0 & \text{, otherwise} \end{cases} \tag{5}$$

where k is the number of ants, $c(t)$ is the path constructed by an ant up to time t; $S_{c(t)}$ is the score of path $c(t)$ and S_{best} is the score of the best path found up to now.

Using this procedure, ants travel through the graph and, at the end of a predefined number of cycles, it is possible to reconstruct the tree using the best path found.

4.1 Construction of the Phylogenetic Tree

The execution of the ACO algorithm, as detailed before, gives a linear sequence of species and a measure of closeness between them using the pheromone matrix. Using these elements, the phylogenetic tree can be constructed as shown by the algorithm of Fig. 4.

```
WHILE NOT (all species grouped)
    FIND i,j pair that have the largest value in the pheromone matrix
    IF (i OR j) already grouped CHANGE index by group index
    GROUP i,j pair into a new species k;
    COMPUTE the distance between current species and ancestor;
    DELETE the value of i, j pair
END
```

Fig. 4. Pseudocode for constructing the phylogenetic tree using the pheromone matrix and the sequence of species given by the ACO algorithm.

5. Computational Experiments and Results

To evaluate the methodology proposed in this work we used two data sets. The first is a set of complete mitochondrial genomes (mtDNA) from 20 species of mammalians, previously used in another studies (see, for instance, [9]). The second data set was especially constructed for this work and is based on DNA sequences of 8 species corresponding to the gene p53. The data for the last data set was found in the NCBI (*http://www.ncbi.nih.gov*).

Results of the construction of phylogenetic trees were compared with well-known PHYLIP package [10] using programs NEIGHBOR and FITCH.

The comparison of two trees is based on the analysis of their structure and the total distance between nodes (Equation 6), proposed by Kumnorkaev et al [8].

$$d_t = 110 + \sum_{i=0}^{n} \sum_{j=0}^{n} \frac{d_{obs}(i,j)^2}{d_{\exp}(i,j)} \quad . \tag{6}$$

where d_{obs} is distance obtained by the algorithm and d_{exp} is expected distance from the distance matrix, between two species, and n is the number of species. This distance measure is somewhat similar to the computation of the quadratic error.

In Fig. 5 two trees obtained with the mtDNA data set are shown. They were obtained using the proposed ACO and the neighbor-joining method, respectively. It is important to notice that, although species are similarly grouped, there are small differences in the order of groupings. This is what causes the differences in the distance between branches.

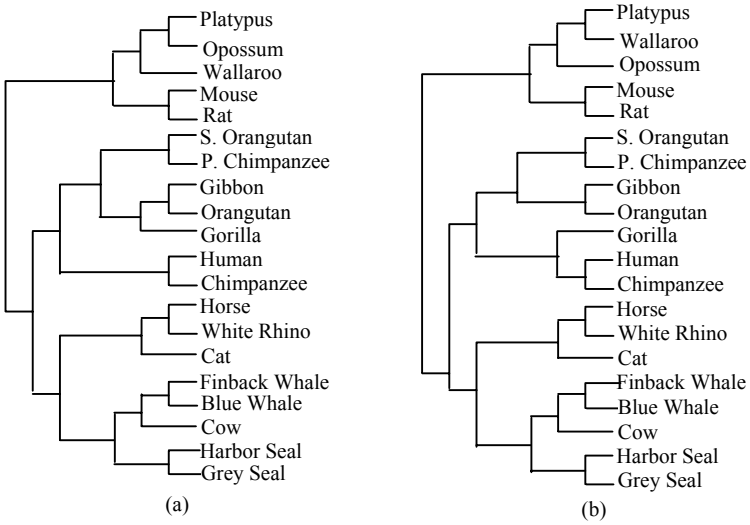

Fig. 5. Phylogenetic trees produced with: (a) the proposed ACO; (b) Neighbor-joining method.

Regarding the distance between branches, the proposed ACO obtained better values when compared with Fitch and Neighbor-joining methods, for both data sets, as shown in Table 1.

Table 1. Comparison of the total distances between branches, for both data sets

Algorithm	mtDNA	gene p53
ACO	351.56	189.98
Fitch	352.27	190.42
Neighbor-joining	354.23	190.63

5.1 Parameters Sensitivity

Several experiments were done with different parameters and, for both data sets, the best results reported here were found using parameters shown in Table 2.

Table 2. Standard parameters for the proposed ACO

Datasets	α	β	η	ρ	k	*#cycles*
mtDNA	1	2	0.3	0.6	100	60
p53	1	2	0.5	0.6	100	60

Parameter α controls the exploration of the search space, by means of weighting the importance of the pheromone trail in the decision of an ant when it arrives to a branch. The algorithm is sensible to high values of this parameter, leading to a fast convergence to a local optimum.

Parameter β defines the relative importance of the distance between species in the transitions between nodes. In practice, we observed that it has to be higher than α. But values too high make the algorithm to converge to a tree that groups species sequentially.

The pheromone trail evaporation is controlled by the parameter ρ that is influenced by the number of ants (k) and the number of cycles. Experimentally, we observed that values higher than 0.8 do not allow the convergence to the same tree, and values lower than 0.2 make the algorithm to find trees with larger distances between branches. It is supposed that this is a consequence of the convergence to a local optimum in the beginning of the run.

Regarding the number of ants (k), we found two distinct behaviors. When k is too small (say, $k<50$) or too high (say, $k>400$) a random behavior is observed in the resulting trees for repeated runs. For intermediary, but high values of k (say, $200<k<350$) a well-defined tree can be obtained but with distances higher than those obtained by other approaches. The range for which the best trees were obtained was $90<k<120$, although we believe that this value may have some dependency with other parameters. Future work will address this issue.

The evolutionary distance between an ancestor and two descendent species is controlled by parameter η. For the p53 data set we observed that the best tree found was obtained using η=0.5, meaning that the distance between the ancestor and the two descendents is the same for both branches. For the mtDNA data set, this parameter was set to 0.3, meaning that the distance between descendents to ancestor species will be divide on 30% for the first descendent and 70% for the other.

6. Conclusions

This work presented a new method for the reconstruction of phylogenetic trees using the Ant Colony Optimization paradigm. For the particular data sets used, the preliminary results presented here shows that the proposed method obtained better results than other established algorithms. However, it cannot be claimed that this will stands for any other data set. Although it was observed that the algorithm is sensitive to parameter changes, no serious attempt was done to optimize parameters. Therefore, it is fair to expect even better performances if an optimized set of parameter can be found. Overall, results are very promising and encourage further developments.

Future work will include exhaustive tests to find a more optimized set of parameters and analysis of its behavior for different types of data sets. Also, we intend to develop an improved methodology to deal with aligned and non-aligned sequences. More experiments shall be done using different benchmarks so as to evaluate how the algorithm behaves as the size of sequence and/or the data set grow.

7. Acknowledgements

Authors would like to thanks CNPQ for the support (processes number: 350053/03-0, 475049/03-9 and 402018/03-6).

References

1. Gonnet, G.H.: New algorithms for the computation of evolutionary phylogenetic trees. In Suhai, S., Computational Methods in Genome Research. Plenum, New York (1994) 153-161
2. Felsenstein, J.: Maximum likelihood estimation of evolutionary trees from continuous characters. Am J Hum Gen 25 (1973) 471-492
3. Sokal, R.R., Michener, C.D.: A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin 28 (1958) 1409-1438
4. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4 (1987) 406-425
5. Colorni, A., Dorigo, M., Maniezzo, V.: Distributed optimization by ant colonies. Proceedings of ECAL'91 European Conf. on Artificial Life. Amsterdam, Elsevier (1991) 134-142
6. Parpinelli, R.S., Lopes, H.S., Freitas, A.A.: Data mining with an ant colony optimization algorithm. IEEE Trans Evol Comput 6 (2002) 321-332
7. Korostensky, C., Gonnet, G.H.: Using traveling salesman problem algorithms for evolutionary tree construction. Bioinformatics 16 (2000) 619-627
8. Kumnorkaew, M., Ku, K., Ruenglertpanyakul, P. Application of ant colony optimization to evolutionary tree construction. Proceedings of 15th Annual Meeting of the Thai Society for Biotechnology. Chiang Mai, Thailand (2004)
9. Cao, Y., Janke, A., Waddell, P.J., Westerman, M., Takenaka, O., Murata, S., Okada, N., Pääbo, S., Hasegawa, M. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J Mol Evol 47 (1998) 307-322
10. Fitch, W., Margoliash, E. The construction of phylogenetic trees. Science 155 (1967) 279-284