Genetics and Molecular Biology, 31, 4, 974-981 (2008)

www.sbg.org.br

Research Article

Ant-Based Phylogenetic Reconstruction (ABPR): A new distance algorithm for phylogenetic estimation based on ant colony optimization

Karla Vittori[1], Alexandre C.B. Delbem[1] and Sérgio L. Pereira[2]

[1]*Departamento de Ciência da Computação, Universidade de São Paulo, São Carlos, SP, Brazil.*
[2]*Department of Natural History, Royal Ontario Museum, Toronto, ON, Canada.*

Abstract

We propose a new distance algorithm for phylogenetic estimation based on Ant Colony Optimization (ACO), named Ant-Based Phylogenetic Reconstruction (ABPR). ABPR joins two taxa iteratively based on evolutionary distance among sequences, while also accounting for the quality of the phylogenetic tree built according to the total length of the tree. Similar to optimization algorithms for phylogenetic estimation, the algorithm allows exploration of a larger set of nearly optimal solutions. We applied the algorithm to four empirical data sets of mitochondrial DNA ranging from 12 to 186 sequences, and from 898 to 16,608 base pairs, and covering taxonomic levels from populations to orders. We show that ABPR performs better than the commonly used Neighbor-Joining algorithm, except when sequences are too closely related (*e.g.*, population-level sequences). The phylogenetic relationships recovered at and above species level by ABPR agree with conventional views. However, like other algorithms of phylogenetic estimation, the proposed algorithm failed to recover expected relationships when distances are too similar or when rates of evolution are very variable, leading to the problem of long-branch attraction. ABPR, as well as other ACO-based algorithms, is emerging as a fast and accurate alternative method of phylogenetic estimation for large data sets.

Key words: phylogenetic estimation, distance algorithms, ant colony optimization.

Received: March 25, 2008; Accepted: July 24, 2008.

Introduction

Phylogenetic estimation from protein or nucleic acid sequences can be performed using optimization and distance algorithms. *Optimization algorithms* using parsimony, likelihood and Bayesian (BA) inference deal with discrete characters and evaluate the fit of phylogenetic trees under a specific optimization criterion. For example, Maximum Parsimony (MP) algorithms search for the tree that minimizes the number of substitutions to explain the variation observed in the aligned sequences (Camin and Sokal, 1965; Eck and Dayhoff, 1966; Cavalli-Sforza and Edwards, 1967; Fitch, 1971). Maximum Likelihood (ML) algorithms search for the tree or set of trees (H) that maximizes the likelihood of generating the observed data (D) given a model (θ) of nucleotide substitution or amino acid replacement (Felsenstein, 1973, 1981; Felsenstein and Churchill, 1996), or formally, $P[D,\theta|H]$. BA inference, a criterion used by ML algorithms, evaluates $P[H|D,\theta]$, *i.e.*, the probability of the hypothesis being correct, given the data and model of substitution (Huelsenbeck *et al.*, 2001). Moreover, BA inference algorithms incorporate prior information regarding model parameters (such as tree topology, rates of substitution, base frequency, among other) to estimate $P[H|D,\theta]$ (Huelsenbeck *et al.*, 2001).

These optimization criteria have their own limitations. Parsimony can be severely biased if homoplasies (*i.e.*, identical character states that do not share a common ancestor) in the data set are common, leading to the "long branch attraction" effect in which long, unrelated branches tend to attract each other (Felsenstein, 1978). Likelihood and BA inference, on the other hand, deal well with homoplastic characters, however, they are computationally expensive and the final outcome can be influenced by the evolutionary model of DNA or amino acid substitution used in tree searching (*e.g.*, Posada and Buckley, 2004). Additionally, choice of priors for BA inference (*e.g.*, Alfaro and Holder, 2006) is a problem in itself; however, if no prior knowledge is available for certain parameters, the use of flat priors (*e.g.*, all trees are equally likely *a priori*) is always advised (Huelsenbeck *et al.*, 2001).

Distance algorithms, such as Neighbor-Joining (NJ) and Unweighted Pair Grouping Method with Arithmetic means (UPGMA), convert aligned sequences into evolutionary distances and search for a tree that best fits these distances. These algorithms have been largely used for phylogenetic estimation since they are computationally

Send correspondence to Karla Vittori. Avenida Orlando Gomes 1845, Piatã, 41650-010 Salvador, BA, Brazil. E-mail: karla@cimatec. fieb.org.br.

easy to implement and fast in building a tree. In contrast to optimization algorithms for phylogenetic estimation that evaluate several trees, distance algorithms only evaluate a single tree. The computational efficiency and the adequate phylogenetic trees produced by distance algorithms have kept these algorithms in vogue, particularly for problems including a large number of sequences or genomic data. If distances can be estimated exactly (Saitou and Nei, 1987) or if they harbor very small errors (Atteson, 1997), the correct phylogenetic tree (which would be obtained under a set of assumptions if all the molecular data from the set of taxa were available) can be obtained. However, the performance of distance algorithms can be affected by the model chosen for estimating distances.

Here, we propose a new distance algorithm, named Ant-Based Phylogenetic Reconstruction (ABPR) that uses the Ant Colony Optimization (ACO) metaheuristic (Dorigo, 1992; Dorigo *et al.*, 1996, 1999; Dorigo and Stützle, 2004). The ABPR algorithm combines aspects of the NJ algorithm, parsimony and likelihood criteria to overcome the limitations cited above. Briefly, the ABPR algorithm estimates a phylogenetic tree based on pairwise distance, while also taking into account the quality of the phylogenetic tree according to the total length of multiple trees generated concomitantly. We applied the ABPR algorithm to four data sets of mitochondrial DNA varying in size and taxonomic coverage, and compared its performance with that of the NJ algorithm in terms of the total length of the produced trees, and the parsimony and likelihood scores.

Ant Colony Optimization

Most of the ideas behind ACO come from the foraging behavior observed in real ant colonies (Hölldobler and Wilson, 1990). When an ant moves through the environment and discovers a food source, it deposits a chemical substance on the ground, named trail pheromone. The pheromone leads other ants from the nest to the food source. Other ants follow the first ant's pheromone trail and reinforce it by deposition of additional pheromone. If there are several pheromone trails leading to the same food source, ants will choose probabilistically the path to follow, based on the pheromone concentrations on the existing paths. Ants that traverse the shortest path between a nest and a food source return to the nest sooner than ants that choose longer paths. Thus, after multiple traverses involving nest and food source, the shortest path will have a stronger pheromone concentration than longer paths. Consequently, ants will concentrate on this path determining cooperatively the optimal path to the food source. Ants operate under stigmergy (Grassé, 1959), a mechanism of indirect communication that coordinates the work of independent entities which have access only to local information about the environment.

In ACO, "artificial" ants cooperate in finding "optimal" solutions to relatively difficult discrete optimization problems (Dorigo, 1992; Dorigo *et al.*, 1996, 1999; Dorigo and Stützle, 2004). These problems are represented as a set of points (named states) and ants move through adjacent states. Exact definitions of state and adjacency are problem-specific. As applied to phylogenetics, a state is defined as the node visited by the ant and adjacency as the nodes connected by branches.

The original ACO approach has additional features that help to obtain satisfactory solutions to complex optimization problems. (1) *Colony of cooperating ants:* ants cooperate to find a good solution to the problem, sharing information about the environment, which is read/written in the local states as they visit them; (2) *Pheromone trail and stigmergy:* while real ants change the environment when depositing pheromone on the localities they visit, artificial ants change some numerical information about current environmental conditions, which is stored locally in the problem states they visit; (3) *Shortest path searching and local moves:* ants move through adjacent states of the environment, searching for the shortest paths joining the nest to the food source; (4) *Probabilistic transition policy:* the ants' actions are selected probabilistically based on local information about the environment and its pheromone concentration; (5) *Discrete world:* the ants' moves consist of transitions between discrete states of the environment; (6) *Internal state*: ants remember past actions; (7) *Pheromone laying:* the amount of pheromone deposited by ants is a function of the quality of the solution found, and the timing of pheromone laying is problem-dependent; and (8) *Extra search capabilities:* ants can use extra mechanisms, such as local optimization (Gambardella and Dorigo, 1997), backtracking (Di Caro and Dorigo, 1997) and look-ahead approaches (Michel and Middendorf, 1998).

The ABPR Algorithm of Phylogenetic Estimation

The algorithm we present here combines some characteristics of conventional algorithms to improve the problem of phylogenetic estimation. The ABPR algorithm evaluates multiple trees as required by parsimony and likelihood criteria. The metrics used to build a phylogenetic tree is based on the evolutionary distance between each pair of sequences and the pheromone concentration between them. As in distance algorithms under the minimum evolution criterion, the best tree is the one with the shortest summation of branch lengths. The ABPR algorithm is composed of a set of ants, moving independently and concurrently, searching for the best tree. Similarly to the NJ algorithm, ants join two species i and j in each movement through the environment. Clusters can consist of two input species (i and j), an input species (k) plus a newly merged ancestor (i - j), or two newly merged ancestors (i - j, k - l). The ABPR algorithm runs for a given number of iterations. For each iteration, a fixed number of ants is generated, and each ant builds a tree.

Selection of the species and parameter values of the ABPR algorithm

Input sequences or newly merged ancestors are considered as cities for the journey of ants. The problem is divided into (n - 2) zones, where n is the number of species. The first zone visited by the ants is named the entrance zone and the last one, the end zone. Each zone consists of (q (q - 1)/2) cities. At the entrance zone, $q = n$, and q is reduced to 1 at the end zone ($q = n, n$ - 1, , 1) (Kumnorkaew *et al.*, 2004). For example, in a data set of 20 sequences, there are 190 initial cities to start tree searching.

The ant chooses two species to be joined based on the evolutionary distance and the pheromone concentration between them, regarding all pairwise comparisons. The number of candidate species joined in each iteration is calculated as in Foulds and Graham (1982) based on the Steiner's problem (Kumnorkaew *et al.* 2004). The probability $P_m(i,j)$ of the ant m to join the species i and j is calculated as:

$$P_m(i,j) = \frac{[\tau(i,j)]^{\alpha}[1/d(i,j)]^{\beta}}{\sum_u \{[\tau(u,j)]^{\alpha}[1/d(u,j)]^{\beta}\}}, \quad (1)$$

where $\tau(i,j)$ is the pheromone trail between i and j; $d(i,j)$ is the evolutionary distance between i and j; α and β are constants with limits ($1 < \alpha, \beta < 10$), and u represents other species of the environment. The constants α and β in Eq. (1) represent the weight of the pheromone concentration and the evolutionary distance, respectively, on the choice made by ant m. We experimentally tested the best value for α and β, which should provide the shortest tree, fixing one parameter while changing the other. We observed that this is achieved when $\alpha = \beta = 2.0$, (Figure 1).

Calculation of branch lengths

After species i and j are joined, the ABPR algorithm creates a new node A, the common ancestor of i and j. The lengths of the branches A_i and A_j, *i.e.*, the branch length from each species to the ancestral node, are calculated as in the NJ algorithm. This calculation allows a direct comparison between the estimated tree length obtained by the ABPR algorithm and the widely used NJ algorithm. Other metrics could also be implemented if desired (Perretto and Lopes, 2005). When using the equations applied by the NJ algorithm to calculate branch lengths, the ABPR algorithm recovers some negative branch lengths, which is biologically not plausible. This happens because the equations used by the NJ algorithm are developed to deal with deterministic choices; while the ABPR algorithm uses a probabilistic model. Likewise, the NJ algorithm can also produce negative branch lengths, which are usually set to zero, and the difference is transferred to the length of the adjacent branch. This procedure does not alter the phylogenetic tree (Kuhner and Felsenstein, 1994).

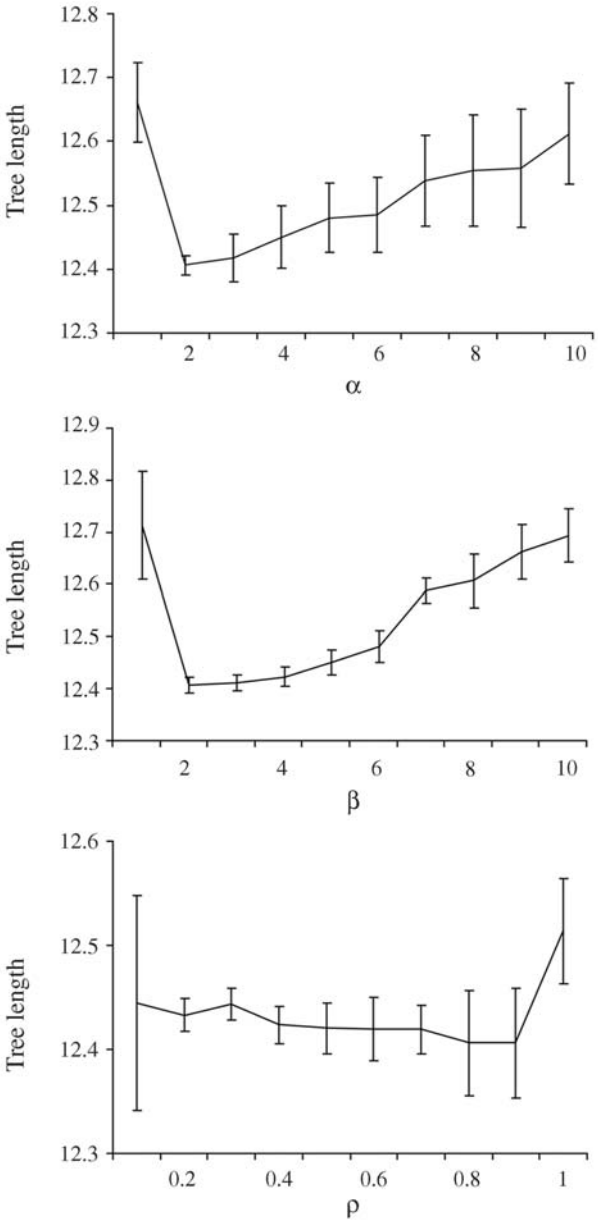

Figure 1 - Tree lengths and standard deviations produced by the ABPR algorithm for $\varphi = 0.005$, and varying α (top graphic), β (middle) and ρ (bottom).

Allowing the ABPR algorithm to recover negative branch lengths can improve the searching of the space parameter by ants (Kumnorkaew *et al.*, 2004). However, we impose that positive branch lengths only would be estimated in the ABPR algorithm. Hence, the lengths of branches A_i and A_j, represented by C_{A_i} and C_{A_j}, respectively, are calculated as:

$$C_{A_i} = d_{ij} * \frac{R_i}{R_i + R_j} \quad (2a)$$

$$C_{A_j} = d_{ij} * \frac{R_j}{R_i + R_j} \quad (2b)$$

where d_{ij} is the evolutionary distance between species *i* and *j* and R_i is the sum of the evolutionary distances between *i* and the other species; $R_i = \sum d_{ik}$ and $0 < k < n$. Similar to the condition found in the NJ algorithm, if $R_i > R_j$, the length C_{A_i} will be higher than C_{A_j}.

Distances and pheromone updating

When two species are merged into one, the distances from others species to the newly merged cluster change. Similar to the NJ algorithm, the ABPR algorithm estimates the new distance as:

$$d_{Ak} = \frac{d_{ik} + d_{jk} - d_{ij}}{2}, \quad (3)$$

where d_{Ak} is the distance between the new node *A* and species *k*.

If a negative branch length is produced for an ant *m* using Eq. (3), the ant is eliminated from the environment and its solution is not considered in the final decision to choose the shortest phylogenetic tree. Only about 0.5% of the generated ants had negative branch length, showing that Eq. (3) is adequate for our purpose.

After calculating the new distance between each pair of species, ant *m* selects two new species and/or clusters to be merged following Eq. (1). These steps are repeated until all ants have traversed all the (*n* - 2) zones of the environment, and each ant has built one phylogenetic tree. So, the algorithm evaluates all trees generated and stores the value of shortest total branch length in L_{best}. Then, the pheromone concentration over the pairs of species selected by ant *m* is updated as:

$$\tau(i,j) = \tau(i,j) + \Delta(i,j) \quad (4)$$

and

$$\Delta(i,j) = \varphi \frac{L_m}{L_{best}}, \quad (5)$$

where $\Delta(i,j)$ is the pheromone increment between species *i* and *j*; φ is an empirical proportionality constant ($0 < \varphi < 1$), and L_m is the total length of the tree built by ant *m*. A small value for φ avoids ants to be trapped in local minimum due to high values of pheromone concentration over some pairs of species (Figure 1). The pheromone evaporation rate ρ is calculated as:

$$\tau(i,j) = (1-\rho) * \tau(i,j), \quad (6)$$

where $0 < \rho < 1$ is the pheromone decay rate. The pheromone evaporation reduces stagnation of ants over a solution that can be a local optimum. Solutions with negative branch lengths are also considered in updating the pheromone concentration, since they are important for algorithm convergence. Parameter ρ is tested for values between 0 and 1 (Figure 1). The results show that a high rate of evaporation (0.9 or larger) largely reduced the pheromone concentration over the pairs of species, generating loss of information acquired by the ants. On the other hand, a low rate of evaporation (0.1 or smaller) does not reduce significantly the pheromone trail and does not guarantee the exploration of a wider range of solutions by the ants. When the algorithm reaches the maximum number of iteration sets, the shortest unrooted phylogenetic tree found, including its branch labels and branch lengths, is presented.

Empirical Data Sets

We inferred the phylogenetic relationships using the ABPR algorithm for: (a) a short fragment of 898 bp from the mitochondrial genome of 12 species of primates (Hayasaka *et al.*, 1988); (b) 1757 bp for 19 almost-complete mitochondrial genomes of birds (Pereira and Baker, 2006); (c) 18,748 bp for 20 complete mitochondrial genomes of mammals (Cao *et al.*, 1998; Perretto and Lopes, 2005); and (d) 16,608 bp from 186 human mitochondrial genomes (Ingman and Gyllensten, 2006). The phylogenetic relationships among the sequences included in these data sets are relatively well known, which facilitates the evaluation of the APBR algorithm in recovering "known" phylogenetic trees. Additionally, these data sets cover a taxonomic range from population to ordinal levels, allowing us to evaluate the utility of the ABPR algorithm at various taxonomic levels and different degrees of sequence divergence.

We retrieved sequences from the GenBank, in FASTA format, and aligned them in ClustalX 1.83 (Thompson *et al.*, 1997). Alignments were visually inspected for misaligned positions. We chose the best evolutionary model of DNA substitution for each data set in PAUP 4.0 (Swofford, 2001) and Modeltest 3.7 (Posada and Crandall, 1998). For all datasets, the best fitting model is GTR+I+G (general time-reversible model), assuming a proportion of invariable sites (I), and gamma distributed rate-variation across sites (G). The parameters of the model were used to estimate evolutionary distances.

For each data set, we run the ABPR algorithm 30 times, each time starting with a random seed number to avoid being locally trapped in the parameter space. We also set 150 ants for 100 iterations per run. Number of runs, ants and iterations were obtained empirically. We did not find shorter trees by using a greater number of iterations or ants, in agreement with results from a previous report applying ACO under the minimum evolution principle for phylogenetic reconstruction (Catanzaro *et al.*, 2007). In all runs of the ABPR algorithm, the parameters were set to $\alpha = 2.0$, $\beta = 2.0$, $\varphi = 0.005$ and $\rho = 0.8$, as determined previously (Figure 1).

Performance Comparison Between ABPR and the NJ Algorithms

Total tree length: the ABPR algorithm produced the shortest tree for all data sets at or above the species level (Table 1). For the 186-sequence data set representing human populations, the NJ algorithm produced the shortest trees, and some branches had negative length. This happens because distances (or branch lengths) are imposed to be perfectly additive, *i.e.*, they reflect the exact amount of changes observed in nucleotide sequences (Felsenstein, 1988; Swofford *et al.*, 1996). Although additive distances have peculiar mathematical properties, they can be biologically unrealistic because branch lengths are usually unknown parameters and need to be estimated from the data (Swofford *et al.*, 1996). Genetic distances represent the number of substitutions (differences) between nucleotide sequences, hence they cannot be negative. Further inspection of the distance matrix generated for each data set revealed that the distances for the 186-sequence data set are very similar to each other. Similar sequences may prevent the ABPR algorithm to converge properly. Moreover, the ABPR algorithm eliminates biological inconsistencies, by imposing branch lengths to be positive following Eq. (2). Therefore, this procedure can generate a tree longer than that generated by the NJ algorithm.

Parsimony and likelihood scores

We also compared the trees produced by the ABPR and the NJ algorithms according to their parsimony (number of steps) and likelihood (logarithmic scale) scores to evaluate if their differences are statistically significant using the Kishino-Hasegawa (Kishino and Hasegawa, 1989) test as implemented in PAUP. The parsimony and likelihood scores for all 30 trees generated by the ABPR algorithm and the tree produced by the NJ algorithm are calculated using PAUP 4.0 using the commands *pscores* and *lscores*, respectively. The smaller the value of this score, the more parsimonious/likely the tree is. Figure 2 summarizes the results. The population data set including 186 human sequences are statistically more parsimonious and more likely than any of the 30 produced by the ABPR algorithm. For the primate data set, the NJ algorithm is statistically worse than 22 of the obtained by the ABPR algorithm. For birds and mammals, the differences between some of the trees produced by the ABPR and NJ algorithms cannot be considered statistically significant. Yet, for the mammal data set, the tree obtained by the NJ algorithm and one of the trees produced by the ABPR algorithm had similar parsimony scores. For the primate, bird and mammal data sets, the likelihood score for the NJ tree is larger than the best scores estimated for some of the trees produced by the ABPR algorithm, however, the differences are not statistically significant according to the Kishino-Hasegawa test.

Table 1 - Total length of the trees produced by the ABPR and NJ algorithms.

Data set	NJ	ABPR				
		Mean	SD	Median	Min	Max
12 primates	3.287	3.215	0.005	3.212	3.212	3.233
19 birds	2.949	2.960	0.008	2.959	2.940	2.974
20 mammals	4.564	4.580	0.007	4.579	4.559	4.593
186 humans	-0.128	0.136	0.0004	0.135	0.135	0.136

Min and Max = the minimum and the maximum tree length found among 30 trees produced by the ABPR algorithm.

Phylogenetic Estimation

The shortest unrooted trees inferred by the ABPR algorithm are shown in Figure 3. In general, the phylogenetic relationships recovered by the ABPR algorithm are similar to those relationships inferred based on optimization criteria such as parsimony, likelihood and BA inference. However, as expected for a distance algorithm that converts

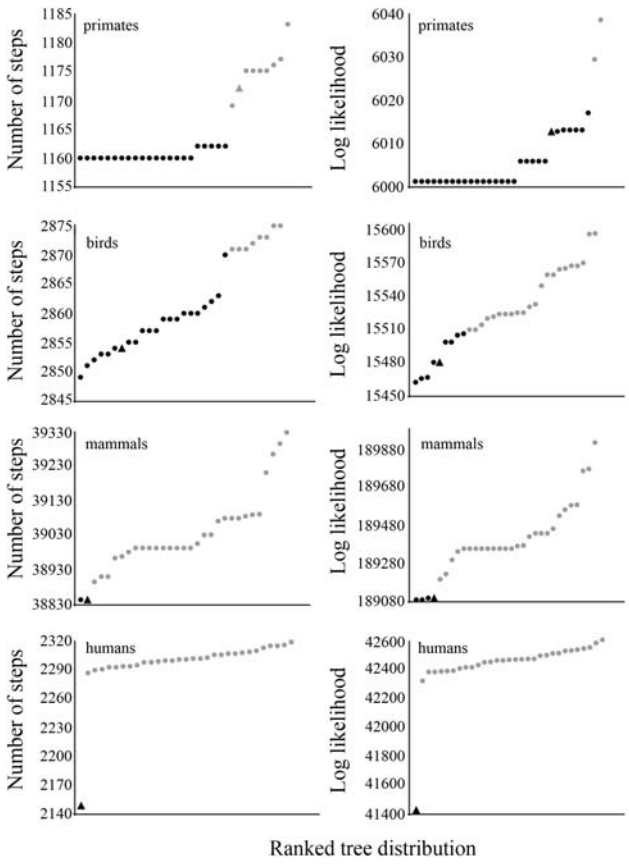

Figure 2 - Comparison of parsimony (leftmost column) and likelihood (rightmost column) scores. Trees are ranked by increasing order of scores. Filled circles are scores for each one of the 30 trees obtained by the ABPR algorithm. Triangle represents score for the tree produced by the NJ algorithm. Trees with score differences that are statistically worse according to the Kishino-Hasegawa test are colored in grey and non-significant differences in black.

discrete character data to genetic distances, some unusual relationships emerge (Nodes marked A-E in Figure 3). Optimization algorithms also fail to recover these same relationships with high bootstrap or posterior probabilities. For the primate data set, both the ABPR and NJ algorithms recover the same topology as in the original study (Hayasaka *et al.*, 1988), which used only the NJ algorithm. The sister relationship between lemurs and tarsiers (Node A in Figure 3) is still debated. Analysis of nuclear and mitochondrial sequence data by distance and optimization algorithms have also recovered lemurs and tarsiers as sister groups (*e.g.*, Eizirik *et al.*, 2001), however, analysis of rare genomic events such as insertion of Short Interspersed Elements (SINEs) suggest a closer relationship between tarsiers and other primates, in exclusion to lemurs (*e.g.*, Schmitz *et al.*, 2005).

The data set for birds includes 19 species, most of them belonging to the order Galliformes (chicken, turkeys, and allies). The unrooted tree recovers the expected relationships at the ordinal level, and for most of basal relationships within Galliformes (*e.g.* Groth and Barrowclough, 1999; Pereira and Baker, 2006; Crowe *et al.*, 2006). However, the relationships among New World quails, guineafowl and Phasianidae (pheasants, turkeys, grouse and allies; Node B in Figure 3) are not concordant among the ABPR (this study) and optimization algorithms (Pereira and Baker, 2006; Crowe *et al.*, 2006). Moreover, the ABPR algorithm is not able to recover curassows as a sister group to all other galliforms in exclusion to megapodes (Node C in Figure 3), as expected based on mitochondrial and/or nuclear sequences (Groth and Barrowclough, 1999; Pereira and Baker, 2006; Crowe *et al.*, 2006). Optimization algorithms have also had problems to solve the above mentioned relationships confidently, which are assumed to have occurred in a short period.

Among mammals, the ABPR algorithm recovers most of the expected ordinal relationships (Springer *et al.*, 2004; Murphy *et al.* 2007), with two exceptions. First, Cetacea is placed as a sister group to Primates instead of composing a monophyletic clade with Artiodactyla (Node D in Figure 3), and second, Rodentia is positioned outside placental mammals instead of being a sister group to Primates (Node E in Figure 3). These unexpected relationships have proved hard to resolve with sequence data, likely due to varying rates of evolution and short radiation time that confound phylogenetic estimation (Springer *et al.*, 2004). Rare genomic events or complete genomes may be able to solve these issues (Kriegs *et al.*, 2006; Wildman *et al.*, 2007).

The tree produced by the ABPR algorithm recovered for human sequences has too many terminals to be shown here. However, the sequences are mostly grouped by geographic groups. Unfortunately, these sequences have not been placed in a phylogenetic framework before, and hence it is difficult to evaluate the performance of the ABPR algo-

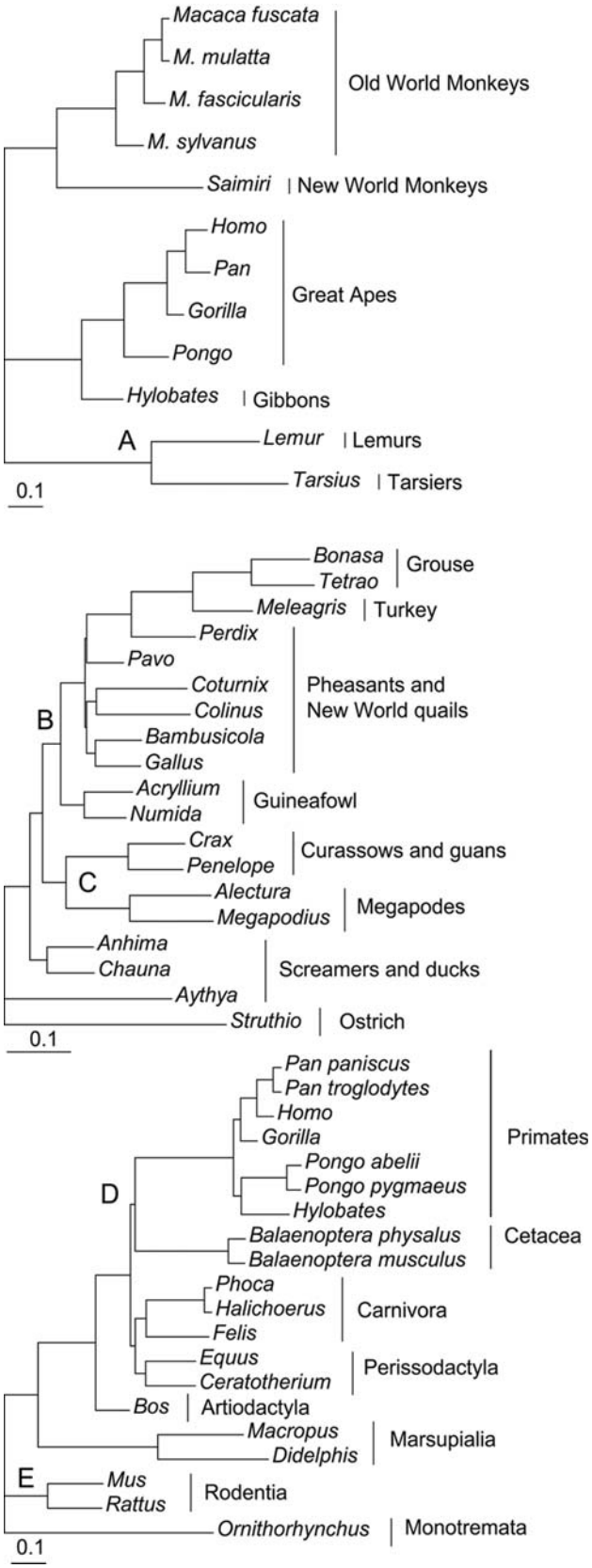

Figure 3 - Shortest unrooted phylogenetic trees recovered by the ABPR algorithm. Nodes marked A to E are incongruent with current views of phylogenetic relationships based on molecular data and are discussed in text.

rithm in recovering a "known" phylogenetic tree in this case. Additionally, as mentioned above, the ABPR algorithm may not be appropriate for building phylogenetic trees at the population level.

Conclusions and Future Research

Phylogenetic estimation is a major feature of evolutionary biology, leading to the development of multiple methods of tree reconstruction and more realistic models of DNA evolution. As computer technology advances, phylogenetic trees are built for an ever increasing number of taxa, longer sequences and even complete genomes. Because the number of possible topologies increases exponentially with the increase in number of sequences, the use of exhaustive and even heuristic algorithms to search the parameter and tree space becomes computationally intractable for optimization algorithms. Hence, distance algorithms are the strategy of choice at the moment when dealing with large data sets because a phylogenetic tree can be obtained within minutes or hours.

Here, we present a promising distance algorithm of phylogenetic estimation named Ant-Based Phylogenetic Reconstruction (ABPR). This algorithm combines properties commonly used by distance and optimization algorithms for phylogenetic estimation, allowing suboptimal solutions to be evaluated before a final phylogenetic tree is presented. As applied to the four data sets chosen for this study, which are evolving under the GTR+I+G model of substitution, it seems that the ABPR algorithm may be affected by disparate rates of evolution, rapid radiation at high-taxonomic levels, and by incomplete lineage sorting due to genetic polymorphisms at the population level. Although all these issues are known to reduce the phylogenetic signal and confound algorithms of phylogenetic estimation, the generalities of our findings need to be tested for other sequences evolving under simpler models of substitutions. Additionally, low sequence divergence is expected at the population level, and it seems to have prevented the ABPR algorithm to converge to a single answer because the distance matrix contains many cells that are too similar and, therefore, uninformative regarding the relationships among members of the same population.

On the other hand, the ABPR algorithm performed better than the NJ algorithm for phylogenetic estimation of DNA sequences at or above the species level. The phylogenetic relationships recovered for primates, birds and mammals are similar to those obtained by optimization algorithms using parsimony, likelihood and BA inference. The ABPR algorithm has the advantage over conventional distance algorithms of evaluating multiple trees during tree search. These positive results are in agreement with recent findings that ACO-based algorithms are potentially competitive with conventional distance algorithms (Perretto and Lopes, 2005; Qin *et al.*, 2006; Catanzaro *et al.*, 2007).

In the future, ABPR and other ACO-based algorithms can be improved to include, among others, measures of clade support, such as bootstrap or jackknife values, use of non-homogeneous models of DNA substitution, and concomitant tree search under multiple models of DNA substitutions. Also, networks may be a better representation of population-level relationships compared to phylogenetic trees. Thus, an ACO-based algorithm to build genetic networks should be implemented to deal with incomplete lineage sorting that affects the evolutionary relationships of populations, subspecies and incipient species.

Acknowledgments

This work was supported by CNPq. Robert Blaquière provided useful comments incorporated in the manuscript.

References

Alfaro ME and Holder MT (2006) The posterior and the prior in Bayesian phylogenetics. Annu Rev Ecol Evol Syst 37:19-42.

Atteson K (1997) The performance of the neighbor-joining method of phylogeny reconstruction. In: Mirkin B, McMorris FR, Roberts FS and Rzhetsky A (eds) Mathematical Hierarchies and Biology. DIMACS Series of Discrete Mathematics and Theoretical Computer Science, v. 37. American Mathematical Society, Rhode Island, pp 133-147.

Camin J and Sokal R (1965) A method for deducing branching sequences in phylogeny. Evolution 19:311-326.

Cao Y, Janke A, Waddell PJ, Westerman M, Takenaka O, Murata S, Okada N, Pääbo S and Hasegawa M (1998) Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J Mol Evol 47:307-322.

Catanzaro D, Pesenti R and Milinkovitch MC (2007) An ant colony optimization algorithm for phylogenetic estimation under the minimum evolution principle. BMC Evol Biol 7:228.

Cavalli-Sforza LL and Edwards AWF (1967) Phylogenetic analysis:models and estimation procedures. Evolution 21:550-570.

Crowe TM, Bowie RCK, Bloomer P, Mandiwana T, Hedderson T, Randi E, Pereira SL and Wakeling J (2006) Phylogenetics and biogeography of, and character evolution in gamebirds (Aves, Galliformes): Effects of character exclusion, partitioning and missing data. Cladistics 22:495-532.

Di Caro G and Dorigo M (1997) AntNet: A mobile agents approach to adaptive routing. Technical Report 97-12. IRIDIA, Université Libre de Bruxelles, Brussels, 27 pp.

Dorigo M (1992) Optimization, learning and natural algorithms. PhD Thesis, Politecnico di Milano, Italy.

Dorigo M and Stützle T (2004) Ant Colony Optimization. MIT Press, Harvard, 319 pp.

Dorigo M, Di Caro G and Gambardella LM (1999) Ant algorithms for discrete optimization. Artificial Life 5:97-116.

Dorigo M, Maniezzo V and Colorni A (1996) The ant system: Optimization by a colony of cooperating agents. IEEE T Syst Man Cyb B 26:29-41.

Eck RV and Dayhoff MO (1966) Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Spring, pp 161-202.

Eizirik E, Murphy WJ and O'Brien SJ (2001) Molecular dating and biogeography of the early placental mammal radiation. J Hered 92:212-219.

Felsenstein J (1973) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst Zool 22:240-249.

Felsenstein J (1978) Cases in which parsimony and compatibility methods will be positivity misleading. Syst Zool 27:401-410.

Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17:368-376.

Felsenstein J (1988) Phylogenies from molecular sequences: Inference and reliability. Annu Rev Genet 22:521-565.

Felsenstein J and Churchill GA (1996) A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol 13:93-104.

Fitch WM (1971) Toward defining the course of evolution: Minimum change for a specific tree topology Syst Zool 35:406-416.

Foulds LR and Graham RL (1982) The Steiner problem in phylogeny NP-complete. Adv Appl Math 3:43-49.

Gambardella LM and Dorigo M (1997) HAS-SOP: An hybrid ant system for the sequential ordering problem. Technical Report 11-97. IDSIA, Lugano, 22 pp.

Grassé PP (1959) La reconstruction du nid et les coordinations interindividuelles chez bellicositermes natalensis et cubitermes sp La théorie de la stigmergie:essai d'interprétation du comportement des termites constructeurs. Insectes Soc 6:41-81.

Groth JG and Barrowclough GF (1999) Basal divergences in birds and the phylogenetic utility of the nuclear RAG-1 gene. Mol Phylogenet Evol 12:115-123.

Hayasaka K, Gojobori T and Horai S (1988) Molecular phylogeny and evolution of primate mitochondrial DNA. Mol Biol Evol 5:626-644.

Hölldobler B and Wilson EO (1990) The Ants. Springer-Verlag, Berlin, 732 pp.

Huelsenbeck JP, Ronquist F, Nielsen R and Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310-2314.

Ingman M and Gyllensten U (2006) mtDB: Human mitochondrial genome database, a resource for population genetics and medical sciences. Nucleic Acids Res 34:D749-D751.

Kishino H and Hasegawa M (1989) Evolution of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoida. J Mol Evol 29:170-179.

Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J and Schmitz J (2006) Retroposed elements as archives for the evolutionary history of placental mammals. PLOS Biol 4:e91.

Kuhner MK and Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459-468.

Kumnorkaew M, Ku K and Ruenglertpanyakul P (2004) Application of ant colony optimization to evolutionary tree construction. Proceedings of 15th Annual Meeting of the Thai Society for Biotechnology Chiang Mai, Thailand, 11 pp.

Michel R and Middendorf M (1998) An island model based ant system with lookahead for the shortest supersequence problem. In: Eiben AE, Back T, Schoenauer M and Schwefel H-P (eds) Proceedings of the Fifth International Conference on Parallel Problem Solving from Nature. Springer-Verlag, Berlin, pp 692-701.

Murphy WJ, Pringle TH, Crider TA, Springer MS and Miller W (2007) Using genomic data to unravel the rooted the placental mammal phylogeny. Genome Res 17:413-421.

Pereira SL and Baker AJ (2006) A molecular timescale for galliform birds accounting for uncertainty in time estimates and heterogeneity of rates of DNA substitutions across lineages and sites. Mol Phylogenet Evol 38:499-509.

Perretto M and Lopes HS (2005) Reconstruction of phylogenetic trees using the ant colony optimization paradigm. Genet Mol Res 4:581-589.

Posada D and Buckley TR (2004) Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol 53:793-808.

Posada D and Crandall KA (1998) Modeltest: Testing the model of DNA substitution. Bioinformatics 14:817-818.

Qin L, Chen Y, Pan Y and Chen L (2006) A novel approach to phylogenetic tree construction using stochastic optimization and clustering. BMC Bioinformatics 7:S24

Saitou N and Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-426.

Schmitz J, Roos C and Zischler H (2005) Primate phylogeny: Molecular evidence from retroposons. Cytogenet Genome Res 108:26-37.

Springer MS, Stanhope MJ, Madsen O and de Jong WW (2004) Molecular consolidate the placental mammal tree. Trends Ecol Evol 19:430-438.

Swofford DL (2001) PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods) v. 4.b4. Sinauer Associates, Inc, Sunderland.

Swofford DL, Olsen GJ, Waddell PJ and Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C and Mable BK (eds) Molecular Systematics. 2nd edition. Sinauer, Sunderland, pp 407-514.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F and Higgins DG (1997) The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876-4882.

Wildman DE, Uddin M, Opazo JC, Liu G, Lefort V, Guindon S, Gascuel O, Grossman LI, Romero R and Goodman M (2007) Genomics, biogeography, and the diversification of placental mammals. Proc Natl Acad Sci USA 104:14395-14400.

Associate Editor: Louis Bernard Klaczko