Improved Solutions for the Balanced Minimum Evolution Problem

Roberto Aringhieri* Daniele Catanzaro*

*Department of Computer Science, University of Turin
Corso Svizzera 185, I-10149 Torino, Italy

*GOM, Département d'Informatique, Université Libre de Bruxelles (U.L.B.)
Boulevard du Triomphe, CP 210/01, B-1050, Brussels, Belgium

Abstract

The minimum evolution criterion is one of the most important criteria of molecular phylogenetic estimation. It states that the phylogeny of a given set of molecular data (taxa) is the one whose sum of edge weights is minimal. Finding the phylogeny that satisfies the minimum evolution criterion involves solving an optimization problem, called the Minimum Evolution Problem (MEP), whose versions are generally $\mathcal{NP}$-Hard. The most recent version of MEP is the Balanced Minimum Evolution problem (BME), which is based on Pauplin's edge weight estimation model [16]. In [1] we presented an algorithm for solving BME based on non-isomorphic enumeration of the phylogenies relative to a given set of taxa. The computational analysis reported in [2] shown the quality of the solution provided using a short amount of computing time. Furthermore, it highlighted that a Local Search can improve the solution quality and the need of increasing the dimension of the instance solved by a a parallel approach. The focus of this paper is therefore to report about the extension of our algorithm in order to include a Local Search method and the results obtained by a parallel version of the whole algorithm.

Keywords: *network design, computational biology, phylogenetics, balanced minimum evolution, local search, parallel computation*

Introduction

Molecular phylogenetics studies the hierarchical evolutionary relationships among organisms (also called *taxa*) by means of molecular data (e.g., DNA or protein sequences). These relationships are typically described by means of weighted trees, or phylogenies, whose leaves represent taxa, internal vertices the intermediate ancestors, edges the evolutionary relationships between pairs of organisms, and edge weights a measure of the dissimilarity between pairs of taxa [10].

Molecular phylogenetics provides several criteria for selecting one phylogeny from among plausible alternatives [5]. Usually, such criteria can be expressed in terms of objective functions, and the phylogenies that optimize them are referred to as optimal [5]. Each criterion adopts a set of assumptions whose ability to describe the real evolutionary process determines the gap between the real and the *true phylogeny*, i.e., the phylogeny that one would obtain under the same set of assumptions if all the molecular data from the (set of) taxa were available [10]. If the optimal phylogeny approaches the true phylogeny as the amount of molecular data analyzed increases, then the corresponding criterion is said to be statistically consistent [11].

One of the most important criteria is the minimum evolution criterion [13, 17, 18]. It states that, given a set Γ of n taxa and the corresponding $n \times n$ symmetric matrix $\mathbf{D} = \{d_{ij}\}$ of evolutionary distances [7], the optimal phylogeny for Γ is the one whose sum of edge weights, estimated from the corresponding evolutionary distances, is minimal [13]. Finding the phylogeny that satisfies the minimum evolution criterion involves the solution of the Minimum Evolution Problem (MEP), whose versions are generally $\mathcal{NP}$-Hard (see [5]). MEP can be stated, in its most general form, as follows.

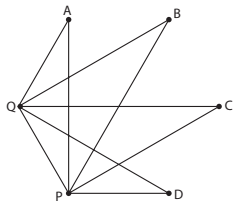

Figure 1: An example of a phylogenetic graph.

Consider a connected, unweighted, undirected graph $G = (V, \mathcal{E})$, hereafter called *phylogenetic graph* (see Figure 1), where $V = V_e \cup V_i$ is the set of vertices. V_e is the set of n *leaves* representing the n taxa in Γ, and V_i the set of $(n-2)$ *internal vertices* representing the common ancestors. By analogy, $\mathcal{E} = \mathcal{E}_e \cup \mathcal{E}_i$ is the set of $\frac{3}{2}(n-1)(n-2)$ edges, where $\mathcal{E}_e$ is the set of *external edges*, i.e., the set of edges with one extreme being a leaf, and $\mathcal{E}_i$ is the set of *internal edges*, i.e., the set of edges with both extremes being internal vertices. Then a *phylogeny* of the set Γ is any spanning tree T of G such that each internal vertex has degree three, and each leaf has degree one.

A phylogeny T of a set of taxa Γ is said *labeled* if there exists a bijection function from the set of leaves of T to the set of taxa Γ, and *unlabeled* otherwise. Figure 2 shows an example of a labeled and an unlabeled phylogeny relative to four taxa.

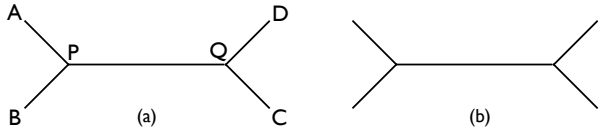

Figure 2: An example of a labeled (a) and an unlabeled (b) phylogeny relative to four taxa.

Given two unlabeled phylogenies T_1 and T_2 relative to the set of taxa Γ, we say that T_1 and T_2 are isomorphic if there exists a bijection g from the set of vertices of T_1 to the set of vertices of T_2 such that a pair of vertices u and w are adjacent in T_1 if and only if $g(u)$ and $g(w)$ are adjacent in T_2.

Denote $\mathcal{T}$ as the set of all the possible $(2n-5)!!$ phylogenies of Γ (where $n!!$ is the double factorial of n) [10]. Then, MEP consists in solving the following optimization problem [5]:

Problem 1. *The Minimum Evolution Problem (MEP)*

$$\begin{aligned} &\min_{(T,\mathbf{w})} \quad L(T,\mathbf{w}) \\ &s.t. \quad f(\mathbf{D},T,\mathbf{w}) = 0 \\ &T \in \mathcal{T}, \mathbf{w} \in \Re_{0+}^{(2n-3)} \end{aligned}$$

where $f(\mathbf{D},T,\mathbf{w})$ is a function correlating the distance matrix $\mathbf{D}$ with the phylogeny T and $\mathbf{w}$ is the $(2n-3)$-vector of edge weights associated to T, and $L(T,\mathbf{w})$ is the *length* of T, i.e., the sum of the associated edge weights. Defining the function $f(\mathbf{D},T,\mathbf{w})$ means specifying an edge weight estimation model, i.e., a model to compute edge weights provided the knowledge of $\mathbf{D}$ and T [5]. Thus, a version of MEP is completely characterized by specifying the functions $L(T,\mathbf{w})$ and $f(\mathbf{D},T,\mathbf{w})$.

Several versions of MEP are known in literature [5]: the most recent one was proposed by Pauplin [16] and it is known as the Balanced Minimum Evolution (BME) problem [8, 9]. Pauplin's edge weight estimation model makes the expression of the length of a phylogeny independent by its edge weights, and leads to the following optimization problem [5]:

Problem 2. *The Balanced Minimum Evolution Problem (BME)*

$$\min_{T \in \mathcal{T}} \quad L(T, \mathbf{w}) = L(T) = \sum_{i=1}^{n} \sum_{j=1}^{n} 2^{-\tau_{ij}} d_{ij} \tag{1}$$

where τ_{ij} are the edges belonging to the path between taxa i and j in T. In other words, BME consists in finding a spanning tree T in the phylogenetic graph G such that: (i) each leaf is a taxon, (ii) each internal vertex has degree tree, and (iii) the length function (1) is minimal. At present, deciding the complexity of BME is an open problem.

In [1, 2] we presented an algorithm for solving BME based on non-isomorphic enumeration of the phylogenies relative to a given set of taxa. The computational analysis reported in [2] shown the quality of the solution provided using a short amount of computing time. Furthermore, it highlighted that a Local Search can improve the solution quality and the need of increasing the dimension of the instance solved by a parallel approach. The focus of this paper is therefore to report about the improvement of our algorithm obtained including a Local Search method and developing a parallel version of the whole algorithm.

The `BMEsolver` algorithm [1, 2]

For a fixed cyclic permutation $\hat{\pi}$ all isomorphic phylogenies share the same values $\{\tau_{ij}\}$, i.e., they identify solutions having the same length. It is worth noting that, given n taxa (say $n \geq 10$), over the 99.9% of the $(2n-5)!!$ phylogenies are isomorphic. Hence, a possible way to reduce the solution space of BME consists of generating only unlabeled non-isomorphic phylogenies and, for each of them, proceeding with finding the best possible assignment of the taxa to the leaves. The major difficulties of this approach lie on both the enumeration procedure, which is nontrivial to implement efficiently, and the solution of the assignment problem, which is a $\mathcal{NP}-$Hard problem [6].

Several authors [12, 14, 19] studied the combinatorial properties of the function (1) highlighting the relationship between BME and the TSP. Specifically, they proved that the length of the optimal solution of BME is equal to half-time the length of the shortest Hamiltonian circuit in G_Γ. In other words, the shortest Hamiltonian circuit in G_Γ identifies a way in which the taxa are ordered in the optimal phylogeny. The algorithm proposed in [1, 2] can be described as follows. Given an instance I of BME, denote T^*

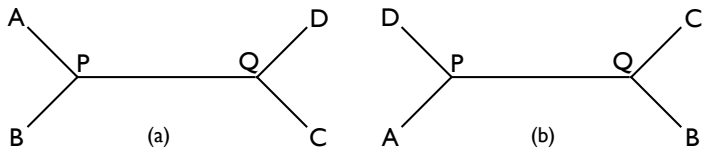

Figure 3: Example: a taxa assignment (a) and a clockwise rotation (b).

and min_{val} as the best phylogeny of I and the optimal length of T^*, respectively. Our algorithm starts solving TSP on G_Γ; the solution C so obtained identifies the taxa order in T^*. The algorithm proceeds by enumerating all the possible non-isomorphic phylogenies by using a modified version of the algorithm in [3].

For each non-isomorphic phylogeny T, the algorithm assigns to each leaf a taxa as shown in Figure 3(a); subsequently, the algorithm computes the length of the resulting phylogeny and the eventual minimum is

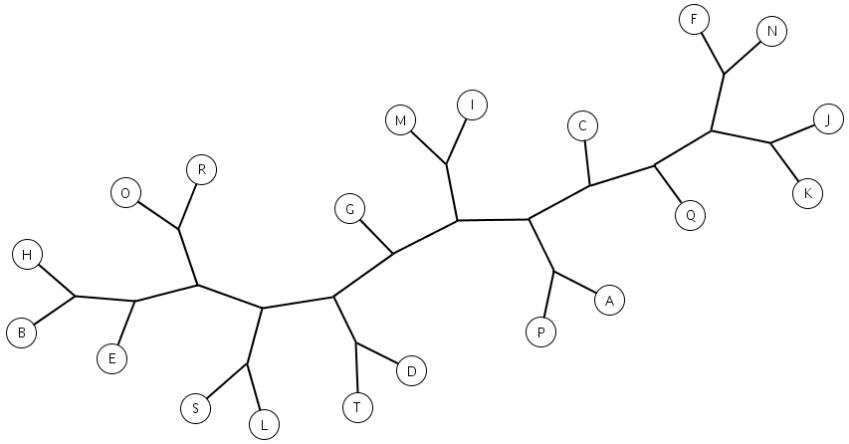

Figure 4: Best solution with 20 taxa computed by `BMEsolver`.

stored. Since the TSP solution only identifies an order of appearance of taxa in T^* but not their relative assignment to the leaves of T^*, in the subsequent $(n-1)$ steps the algorithm performs a clockwise rotation of taxa assignment to leaves of the phylogeny (see Figure 3(b)). When all the possible non-isomorphic phylogenies have been enumerated, the algorithm ends returning the best phylogeny found. A solution example is depicted in Figure 4. The pseudo-code of our algorithm is shown in Figure 5.

procedure `BMESolver`(Γ: set of taxa, **D**: distance matrix)
$C \leftarrow$SolveTSP(Γ,**D**);
T^* =NULL; $min_{val} = \infty$;
for Any possible non-isomorphic phylogeny T **do**
  for any pair of unassigned leaves i and j in T **do** Compute τ_{ij};
  $T_C \leftarrow$AssignLeaves(C,T)
  if $min_{val} \geq L(T_C)$ **then** $min_{val} = L(T_C)$; $T^* = T_C$ **end if**
  for Any clockwise shift R of C on T **do**
    $T_R \leftarrow$AssignLeaves(R,T)
    if $min_{val} \geq L(T_R)$ **then** $min_{val} = L(T_R)$; $T^* = T_R$ **end if**
  end for
end for
return T^* and min_{val}
end-procedure

Figure 5: `BMESolver` pseudo-code.

Parallel Algorithm and Local Search

A limitation to the computational capability of `BMESolver` is due to the increasing number of unlabeled trees: we have 11020, 565734 and 2841632 different trees having 20, 25 and 27 taxa determining a total running time of about 6, 620 and 1500 seconds, respectively. Moreover, real applications need to address instances having more than 27 taxa [4].

In order to reduce the running time needed to compute a solution for a larger number of taxa, we developed a parallel algorithm which is based on a simple communication model among $k+2$ processes depicted in Figure 6. Process 0 enumerates the unlabeled phylogenies. Then it assigns the last enumerated phylogeny to the first available `BMEsolver` process. Each `BMEsolver` (the ones having process id in $[1, \ldots, k]$) waits for a new phylogeny. When it receives a new phylogeny, it computes the best taxa assignment and the value (1). At the end of its computation, the process sends the solution computed to process $k+1$ which collects all the computed solutions in order to extract the best one.

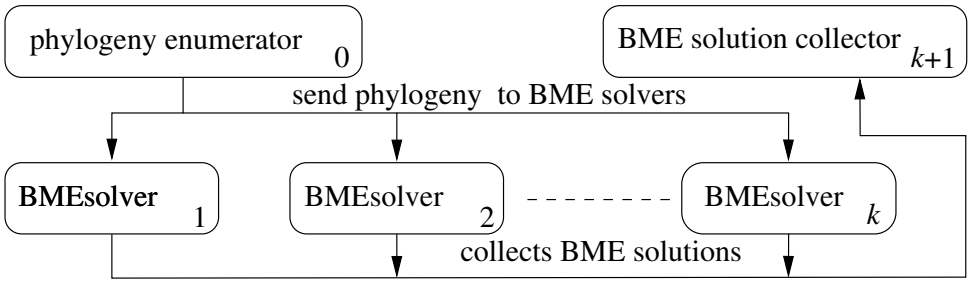

Figure 6: Description of the parallel algorithm.

We tested the sequential and parallel version of `BMEsolver` on a machine having 2 Dual Core AMD Opteron 275 2.2 GHz with 3 Gb of main memory running under Linux operating system. On the same instances reported in Table 1, preliminary computational results show an average linear speed-up: the running time decreases from 626.015 seconds required for the sequential version of `BMEsolver` to the 105.185 seconds required for the parallel version setting $k = 6$.

We now consider a 2-opt Local Search (LS) consisting in the swap of two taxa in all possible ways, at each iteration. Let T^* be the phylogeny computed by `BMESolver` giving the best BME value $L(T^*)$. Table 1 compares $L(T^*)$ with the following values: $L_1(T^*)$, obtained after 1 LS iteration starting from T^*, and $L_1(T')$, computed after 1 LS iteration applied to each phylogeny $T' \neq T^*$. Finally, Table 1 reports the number $t_{\min}$ of phylogenies T' improving T^*.

	`taxa_25`					
	$L(T^*)$	$L_1(T^*)$	gap	$L_1(T')$	gap	$t_{\min}$
01	2868.430542	2844.117065	0.85%	2814.289917	1.89%	18
02	3447.088867	3412.391602	1.01%	3341.882202	3.05%	47
03	3263.754700	3258.775391	0.15%	3254.036194	0.30%	4
04	3266.178436	3264.205292	0.06%	3256.033661	0.31%	2
05	3084.504700	3084.504700	0.00%	-	-	0
06	3181.696167	3181.696167	0.00%	3147.139465	1.09%	11
07	3601.660278	3601.660278	0.00%	-	-	0
08	2981.984192	2981.984192	0.00%	2955.996826	0.87%	2
09	2999.072113	2999.072113	0.00%	2992.837128	0.21%	1
10	3363.253296	3363.253296	0.00%	3351.924805	0.34%	3
	Avg. gaps		0.21%		1.01%	

Table 1: Instances with 25 taxa, total number of phylogenies 565734, results from [2].

The reported results show that a Local Search may improve the quality of the solutions computed by `BMESolver`. The peculiar form of the objective function (1) makes its evaluation a non-easy task, however a complex data structure such as the EPT matrix [15] may turn out helpful. We also add a short-term memory tool to allow a more accurate exploration of the solution space.

Finally, we observe that the linear speed-up gained by the parallel algorithm can be also employed to allow a more deep exploration of the solution space starting from the solution computed by the original `BMESolver`. Therefore, a Local Search improvement method can be added to each one of the k `BMESolver` depicted in Figure 6.

References

[1] R. Aringhieri, C. Braghin, and D. Catanzaro. An exact approach for solving the balanced minimum evolution problem. In *Cologne-Twente Workshop on Graphs and Combinatorial Optimization*, May

2008.

[2] R. Aringhieri and D. Catanzaro. A non-isomorphic enumerative approach to solution of the balanced minimum evolution problem. Note del Polo 116, DTI - University of Milano, October 2008. Submitted to Networks.

[3] R. Aringhieri, P. Hansen, and F. Malucelli. Chemical trees enumeration algorithms. *4OR*, 1:67–83, 2003.

[4] L. Bertolotti, T.L. Goldberg, G. Amore, R. Reina, E. Grego, M. Giacobini, and S. Rosati. Heuristic approaches in bayesian evolution inferences of rna viruses: West nile virus and small ruminant lentivirus case studies. In *Facing the Challenge of Infectious Diseases – Integrating mathematical modeling, computational thinking and ICT applications*, October.

[5] D. Catanzaro. The minimum evolution problem: Overview and classification. *Networks*, In print, 2008.

[6] D. Catanzaro, M. Labbé, R. Pesenti, and J. J. Salazar. Mathematical models to reconstruct phylogenetic trees. *Networks*, In Print, 2008.

[7] D. Catanzaro, R. Pesenti, and M. Milinkovitch. A non-linear optimization procedure to estimate distances and instantaneous substitution rate matrices under the GTR model. *Bioinformatics*, 22(6):708–715, 2006.

[8] R. Desper and O. Gascuel. Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle. *Journal of Computational Biology*, 9(5):687–705, 2002.

[9] R. Desper and O. Gascuel. Theoretical foundations of the balanced minimum evolution method of phylogenetic inference and its relationship to the weighted least-squares tree fitting. *Molecular Biology and Evolution*, 21(3):587–598, 2004.

[10] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA, 2004.

[11] O. Gascuel. *Mathematics of evolution and phylogeny*. Oxford University Press, New York, 2005.

[12] G. H. Gonnet, C. Korostensky, and S. Benner. Evaluation measures of multiple sequence alignments. *Journal of Computational Biology*, 7(1-2):261–276, 2000.

[13] K. K. Kidd and L. A. Sgaramella-Zonta. Phylogenetic analysis: Concepts and methods. *American Journal of Human Genetics*, 23:235–252, 1971.

[14] C. Korostensky and G. H. Gonnet. Using traveling salesman problem algorithms for evolutionary tree construction. *Bioinformatics*, 16(7):619–627, 2000.

[15] G.L. Nemhauser and L.A. Wolsey. *Integer and combinatorial optimization*. Wiley-Interscience, New York, 1999.

[16] Y. Pauplin. Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, 51:41–47, 2000.

[17] A. Rzhetsky and M. Nei. Statistical properties of the ordinary least-squares generalized least-squares and minimum evolution methods of phylogenetic inference. *Journal of Molecular Evolution*, 35:367–375, 1992.

[18] A. Rzhetsky and M. Nei. Theoretical foundations of the minimum evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10:1073–1095, 1993.

[19] C. Semple and M. Steel. Cyclic permutations and evolutionary trees. *Advances in Applied Mathematics*, 32(4):669–680, 2004.