European Journal of Operational Research 244 (2015) 753–765

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Discrete Optimization

A branch-price-and-cut algorithm for the minimum evolution problem

Daniele Catanzaro[a,b], Roberto Aringhieri[c], Marco Di Summa[d,*], Raffaele Pesenti[e]

[a] *Louvain School of Management, Université Catholique de Louvain, Chausse de Binche 151, bte M1.01.01, 7000 Mons, Belgium*
[b] *Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain, Voie du Roman Pays 34, L1.03.01, B-1348 Louvain-la-Neuve, Belgium*
[c] *Dipartimento di Informatica, Università di Torino, Corso Svizzera 135, I-10149 Torino, Italy*
[d] *Dipartimento di Matematica, Università degli Studi di Padova, Via Trieste 63, I-35121 Padova, Italy*
[e] *Department of Management, Università Ca' Foscari, San Giobbe, Cannaregio 837, I-30121 Venezia, Italy*

ARTICLE INFO

Article history:
Received 13 October 2013
Accepted 10 February 2015
Available online 16 February 2015

Keywords:
Combinatorial inequalities
Branch-price-and-cut
Symmetry breaking
Tree isomorphism
Computational biology

ABSTRACT

We investigate the *Minimum Evolution Problem* (MEP), an $\mathcal{NP}$-hard network design problem arising from computational biology. The MEP consists in finding a weighted unrooted binary tree having *n* leaves, minimal length, and such that the sum of the edge weights belonging to the unique path between each pair of leaves is greater than or equal to a prescribed value. We study the polyhedral combinatorics of the MEP and investigate its relationships with the Balanced Minimum Evolution Problem. We develop an exact solution approach for the MEP based on a nontrivial combination of a parallel branch-price-and-cut scheme and a non-isomorphic enumeration of all possible solutions to the problem. Computational experiments show that the new solution approach outperforms the best mixed integer linear programming formulation for the MEP currently described in the literature. Our results give a perspective on the combinatorics of the MEP and suggest new directions for the development of future exact solution approaches that may turn out useful in practical applications. We also show that the MEP is statistically consistent.

1. Introduction

Molecular phylogenetics studies the hierarchical evolutionary relationships among species, or *taxa*, by means of molecular data such as DNA, RNA, amino acids, or codon sequences. These relationships are usually described through a weighted tree, called a *phylogeny* (see Fig. 1), whose leaves represent the observed taxa, internal vertices represent the intermediate ancestors, edges represent the estimated evolutionary relationships, and edge weights represent evolutionary processes (e.g., evolutionary rates, expected number of mutations, and so on) between pairs of taxa (Felsenstein, 2004). As usual in most of the literature, in this paper the leaves of a phylogeny are labeled in order to identify the given taxa, whereas the internal vertices are unlabeled.

Phylogenies provide fundamental information in the analysis of many fine-scale genetic data. For this reason, the use of molecular phylogenetics has become more and more frequent (and sometimes indispensable) in a multitude of research fields such as systematics, medical research, drug discovery, epidemiology, and population dynamics (Pachter & Sturmfels, 2007). For example, the use of molecular phylogenetics was of considerable assistance in predicting evolution of human influenza A (Bush, Bender, Subbarao, Cox, & Fitch, 1999), understanding the relationships between the virulence and the genetic evolution of HIV (Ou et al., 1992; Ross & Rodrigo, 2002), identifying emerging viruses such as SARS (Marra et al., 2003), recreating and investigating ancestral proteins (Chang & Donoghue, 2000), designing neuropeptides causing smooth muscle contraction (Bader, Moret, & Vawter, 2001), and relating geographic patterns to macroevolutionary processes (Harvey, Brown, Smith, & Nee, 1996). More recently, phylogenies have been used to study the evolutionary relationships of the genetic factors involved in common human diseases (Misra, Blelloch, Ravi, & Schwartz, 2011; Pennington, Smith, Shackney, & Schwartz, 2006; Sridhar et al., 2007; Sridhar, Lam, Blelloch, Ravi, & Schwartz, 2008). Similarly, phylogenies have been also employed to reconstruct a plausible progression of carcinomas over time (Riester, Attolini, Downey, Singer, & Michor, 2010; Subramanian, Shackney, & Schwartz, 2013) by using, in particular, single-cell sampled data from affected individuals (Catanzaro, Schackney, & Schwartz, 2013; Chowdhury et al., 2013). In this context, phylogenies allowed the classification of tumor cells in subfamilies characterized by specific evolutionary traits. This classification might enable a better understanding of cellular atypia over time and, on a long run, suggest new therapeutical approaches for tumor pathologies. A recent survey concerning the practical uses of molecular phylogenetics can be found in Beerenwinkel, Schwartz, Gerstung, and Markowetz (2015).

* Corresponding author. Tel.: +39 049 8271 348; fax.: +39 049 8271 392.
E-mail address: disumma@math.unipd.it (M. Di Summa).

http://dx.doi.org/10.1016/j.ejor.2015.02.019
0377-2217/© 2015 Elsevier B.V. All rights reserved.

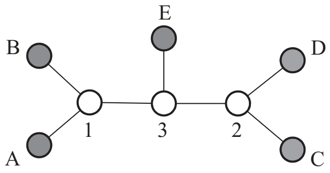

Fig. 1. A phylogeny with five taxa (A, B, C, D, E) and three internal vertices. Though the internal vertices of a phylogeny are unlabeled, here we attach numbers to them for ease of reference. Edge weights are not shown.

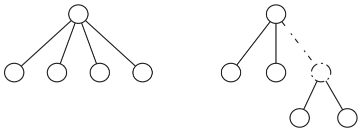

Fig. 2. The 4-ary tree (on the left) can be transformed into a phylogeny by adding a dummy vertex and a dummy edge (dashed, on the right).

The internal vertices of a phylogeny of n taxa represent speciation events occurred throughout evolution of the observed taxa and are usually constrained to have degree three. This constraint has both a biological foundation (see Felsenstein, 2004) and a mathematical motivation, as it proves helpful when formalizing the evolutionary process of taxa. In fact, it does not introduce oversimplifications, as any m-ary tree can be transformed into a phylogeny by adding "dummy" vertices and edges, e.g., see Fig. 2. Moreover, as observed in Felsenstein (2004), the degree constraint helps both in quantifying a priori the number of edges and internal vertices of a phylogeny ($2n-3$ and $n-2$, respectively), otherwise hard to determine, and in counting the overall number of possible phylogenies for a set of n taxa, (i.e., $(2n-5)!! = 1 \times 3 \times 5 \times 7 \times \cdots \times (2n-5)$). The large number of possible phylogenies for a set of n taxa entails the use of an estimation criterion to select a phylogeny from among plausible alternatives.

The literature on molecular phylogenetics proposes a number of possible criteria to estimate phylogenies from molecular data. Apart from particular cases (e.g., Bayesian Inference, Huelsenbeck, Larget, Miller, & Ronquist, 2002; Huelsenbeck, Ronquist, Nielsen, & Bollback, 2001), these criteria can be quantified and expressed in terms of objective functions, giving rise to families of optimization problems, whose general paradigm can be stated as follows:

Problem 1 (The Phylogeny Estimation Problem (PEP)). *Given a set* Γ *of n taxa, find a phylogeny* T^* *that solves the problem*

$$\begin{aligned} \textit{optimize} \quad & \Lambda(T) \\ \textit{s.t.} \quad & \Omega(\Gamma, T) \geq 0 \\ & T \in \mathcal{T}, \end{aligned}$$

where $\mathcal{T}$ *is the set of all possible phylogenies of* Γ, $\Lambda : \mathcal{T} \to \mathbb{R}$ *is a function modeling the selected criterion of phylogeny estimation, and* $\Omega : \Gamma \times \mathcal{T} \to \mathbb{R}^m$, *for some* $m \geq 1$, *is a function correlating the set* Γ *to a phylogeny T.*

A specific PEP is completely characterized by defining the functions Λ and Ω. A phylogeny T^* that optimizes Λ and satisfies the constraints described by Ω is referred to as *optimal*.

A possible version of PEP is *Minimum Evolution* (ME) (see Felsenstein, 2004, p. 161) which consists in minimizing the length of a phylogeny (i.e., the sum of its edge weights) with respect to a given measure of dissimilarity among taxa (see Felsenstein, 2004; Gascuel, 2005; Page & Holmes, 1998). In the context of ME, a phylogeny of Γ is defined to be optimal if it satisfies the following requirements:

(i) it has the shortest length, i.e., the minimum sum of edge weights;
(ii) it has nonnegative edge weights and, for each pair of distinct taxa $i, j \in \Gamma$, the sum of the weights of the edges belonging to the (unique) path from i to j in T^* is not smaller than the given measure of the dissimilarity between i and j.

We refer the interested reader to Farach, Kannan, and Warnow (1995) for a biological justification of constraint (ii). Here we just observe that if only (i) and the nonnegativity of the weights were imposed, then the problem would be trivial, as all weights would be set to zero.

A number of versions of ME have been proposed in the literature, mainly differing from one another by the way in which the edge weights are estimated. Examples include Kidd and Sgaramella-Zonta (1971), Beyer, Stein, Smith, and Ulam (1974), Rzhetsky and Nei (1992, 1993), Felsenstein (2004), Gascuel (2005). One of the earliest edge weight estimation models was proposed by Waterman, Smith, Singh, and Beyer (1977) and can be stated as follows:

Problem 2 (The Minimum Evolution Problem (MEP)). *Given a set* Γ *of n taxa and values* $d_{ij} \geq 0$ *for all pairs of taxa* $i, j \in \Gamma$ $(i \neq j)$, *find a phylogeny* $T^* \in \mathcal{T}$ *that solves the problem*

$$\min_{T \in \mathcal{T}} \mathcal{L}(T) = \sum_{e \in E(T)} w_e \tag{1}$$

$$\text{s.t.} \sum_{e \in P(i,j)} w_e \geq d_{ij} \quad \forall\, i, j \in \Gamma : i \neq j \tag{2}$$

$$w_e \geq 0 \quad \forall\, e \in E(T) \tag{3}$$

where $E(T)$ *denotes the set of edges of a phylogeny* $T \in \mathcal{T}$, w_e *is the weight of edge* $e \in E(T)$, *and* $P(i, j)$ *is the unique path in T connecting the distinct taxa* $i, j \in \Gamma$.

The generic value d_{ij} in Problem 2 is called the *observed evolutionary distance* between taxa i and j, and represents a given measure of the observed dissimilarity between i and j; the d_{ij}'s satisfy $d_{ij} = d_{ji}$ for all $i, j \in \Gamma$, $i \neq j$. These values are usually computed, e.g., by using one of the models described in Jukes and Cantor (1969), Fitch (1971), Kidd and Sgaramella-Zonta (1971), Beyer et al. (1974), Hasegawa, Kishino, and Yano (1981), Kimura (1980), Lanave, Preparata, Saccone, and Serio (1984), Rodriguez, Oliver, Marin, and Medina (1990), Waddell and Steel (1997), Galtier (2001), Huelsenbeck (2002), Lopez, Casane, and Philippe (2002), Felsenstein (2004), Catanzaro, Pesenti, and Milinkovitch (2006).

The objective function (1) models condition (i), while constraints (2)–(3) impose condition (ii). Problem 2 is a particular network design problem (see Johnson, Lenstra, & Kan, 1978; Pop, 2012) with specific degree constraints and unknown edge weights. As observed in Catanzaro (2011), all known versions of ME can be obtained from MEP by imposing further constraints on the edge weights.

The MEP can be solved in polynomial time if the observed evolutionary distances satisfy some specific properties described in Waterman et al. (1977). In contrast, if the observed evolutionary distances are generic, then the MEP is proved to be $\mathcal{NP}$-hard via a reduction from the k-coloring problem (Farach et al., 1995). Moreover, in such a case the MEP cannot be even approximated in polynomial time within ratio n^{ϵ}, for some $\epsilon > 0$, unless $\mathcal{P} = \mathcal{NP}$ (Farach et al., 1995).

The hardness of the MEP has prevented researchers from finding practical solution techniques even for small ($n \leq 10$) instances. This fact justifies the interest of operations researchers in developing approaches that can exactly solve the problem or approximate its optimal solution. In this context, Catanzaro, Labbé, Pesenti, and Salazar-Gonzáles (2009) recently presented mixed integer programming models for the MEP based on a particular encoding of phylogenies by means of edge-path incidence matrices of trees (see, e.g., Nemhauser & Wolsey, 1999). This encoding allowed the use of ad-hoc block decomposition methods capable of reducing the solution space of the problem. Unfortunately, all the proposed models proved to be unable to solve real instances of the MEP containing more than 8 taxa.