Microcanonical optimization algorithm for the Euclidean Steiner problem in $\mathbb{R}^n$ with application to phylogenetic inference

Flávio Montenegro,[1] José R. A. Torreão,[2] and Nelson Maculan[1]
[1]*COPPE–Programa de Engenharia de Sistemas e Computação, Universidade Federal do Rio de Janeiro,
21941-972 Rio de Janeiro RJ, Brazil*
[2]*Programa de Pós-Graduação em Informática Aplicada, Pontifícia Universidade Católica do Paraná, 80215-901 Curitiba PR, Brazil*

(Received 28 May 2003; published 13 November 2003)

The Euclidean Steiner tree problem in $\mathbb{R}^n$ (ESTP) is that of finding the shortest interconnecting network spanning p given nodes in the Euclidean $\mathbb{R}^n$, with the possible use of extra nodes. Combinatorial explosion precludes the use of exact methods for large high-dimensional ESTP instances, but very few heuristic approaches have so far been proposed for them. Here we introduce a microcanonical optimization algorithm that works over a topology-describing data structure associated to the ESTP solutions, and which is proven able to find close-to-minimum Steiner trees in reasonable computational time, even for configurations of up to $p=50$ points in $n=50$ dimensions. Moreover, its performance is shown to increase with n, which makes it especially suited for high-dimensional clustering problems such as those of phylogenetic inference, an instance of which is considered here.

DOI: 10.1103/PhysRevE.68.056702 PACS number(s): 02.60.Pn, 02.70.−c, 05.20.−y, 89.75.Hc

I. INTRODUCTION

The Euclidean Steiner tree problem in $\mathbb{R}^n$ (ESTP) can be defined thus: given p points in $\mathbb{R}^n$, with Euclidean metric, find a minimum tree which spans them, using or not using extra points, called *Steiner points*. This is a problem with a long history in the annals of mathematics, details of which can be found in Ref. [1]. It is also a very hard computational problem, its decision version having been proven NP-complete [2].

The ESTP solution trees in $\mathbb{R}^2$ and $\mathbb{R}^3$ find several applications in network design [1,3], and approaches to protein folding have also been based on them [3,4]. In higher dimensions, the ESTP is associated with general clustering problems, including those of phylogenetic inference, such as deriving evolutionary trees: the *method of minimum evolution* [5] formulates the latter as a problem of finding minimum-length Steiner trees [6,7].

An exact enumerative scheme for solving the ESTP was proposed by Smith [8], while Maculan *et al.* [9] formulated the problem as a nonconvex mixed-integer program, introducing a Lagrangian dual, which also leads to an exact branch-and-bound solution. A number of heuristic approaches in $\mathbb{R}^2$ have also appeared (see Ref. [10] for a survey), while heuristics for $n\geqslant 3$ are rarer, but can be found in Refs. [6,3,11,12].

Here, we introduce a heuristic approach for the ESTP in $n\geqslant 3$, using a metaheuristic called the *microcanonical optimization* (MO) *algorithm* [13,14], which explores the parallel between statistical-physics systems and high-dimensional optimization problems, along similar lines as in the pioneering work by Kirkpatrick *et al.* [15]. Our heuristic performs a local search over the space of the topology-describing vectors which result from Smith's enumerative scheme [8], and has consistently yielded good solutions, for instances of up to 50 given points in 50 dimensions. Since the performance of our algorithm is found to increase with n, we believe that it might provide a suitable tool for high-dimensional clustering problems, including those of phylogenetic inference, an instance of which is treated here.

II. ESTP—BASIC CONCEPTS

The solutions to the ESTP, called *Steiner minimal trees*, present the following properties [16]. (1) Given p points $x^i \in \mathbb{R}^n$, $i=1,2,\ldots,p$, the maximum number of Steiner points is $p-2$. (2) A Steiner point has degree (valence) equal to 3. (3) The edges emanating from a Steiner point lie in a plane, and have mutual angles of 120°.

If a tree (minimal or not) satisfies such conditions, we call it a *Steiner tree*, and call the graph that represents such a tree a *Steiner topology*. The total number of distinct *full Steiner topologies*—i.e., topologies with $p-2$ Steiner points—is $(2p-5)!!$, where the double exclamation mark stands for double factorial. A Steiner tree of minimum length for a given topology is called a *relatively minimal tree*, and has been proven unique in a Euclidean space of any dimension [16]. We may consider all (connected) nonfull tree topologies as full Steiner topologies where one or more Steiner points coincide with given points. Thus, it suffices to focus our attention on the relatively minimal trees for full topologies, when looking for heuristic solutions to the ESTP.

III. TOPOLOGY-DESCRIBING VECTORS

The enumerative scheme by Smith [8] is based on a one-to-one correspondence between full Steiner topologies with $p\geqslant 3$ given points, and $(p-3)$-vectors $\mathbf{a}$, whose ith entry a_i is an integer in the range $1\leqslant a_i\leqslant 2i+1$. Each topology-describing $(p-3)$-vector can be constructively obtained, starting from an initial null vector () related to a full Steiner topology for three given points connected to a single Steiner point, labeled $p+1$ (see Fig. 1).

The introduction of a fourth point in the initial topology of Fig. 1(a) is made through the Steiner point $p+2$, that must be inserted in one of the three original edges, for in-

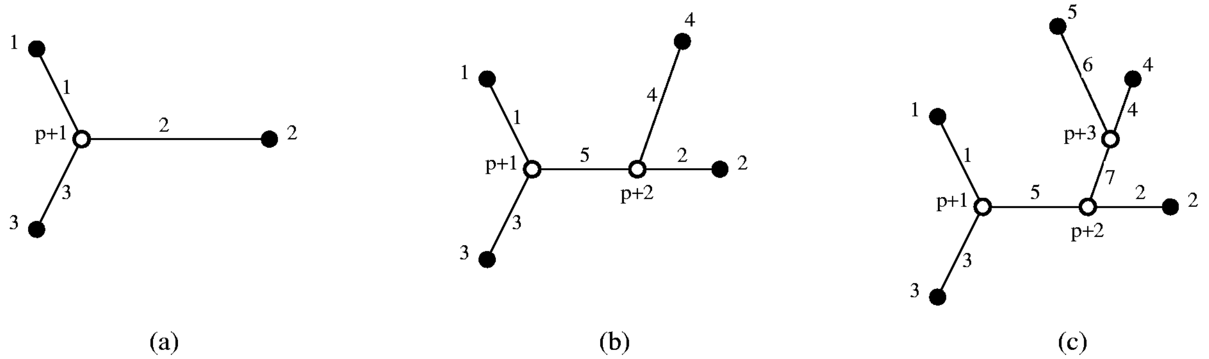

FIG. 1. The initial null vector () corresponds to the topology presented in (a); connection of given points 4, through the edge 2 in (b), and 5, through the edge 4 in (c), will give rise to new topologies, corresponding respectively to vectors (2) and (2,4).

stance, edge 2. This gives rise to new edges, numbered 4 and 5 in Fig. 1(b). Since we have chosen edge 2 to insert Steiner point $p+2$, the resulting topology will correspond to the topology-describing vector (2). Now, among the five available edges, we may choose, say, edge 4, to insert the Steiner point $p+3$ that will connect the fifth given point. The new topology [Fig. 1(c)] corresponds to the topology-describing vector (2,4), and two more edges, 6 and 7, arise. Proceeding thus, we obtain a complete topology-describing $(p-3)$-vector.

It is possible to describe all full Steiner topologies by just combining the possible $p-3$ entries of the vector $\boldsymbol{a}$. Working in the context of an exact branch-and-bound algorithm, Smith employed a backtracking technique to obtain the $\boldsymbol{a}$ vectors, whose corresponding topologies he then minimized [8]. However, the time requirement of such process is very high, already making it unfeasible for configurations with $p \approx 15$ given points. As an alternative, we propose using local-search heuristics for generating the topology vectors, keeping only the minimization step of Smith's approach. Thus, we can obtain good approximate solutions in reasonable time. To assess the quality of a solution, we use the so-called *Steiner ratio*, defined as $\rho = L_{ST}/L_{MST}$, where L_{ST} is the length of the (exact or heuristic) Steiner tree, and L_{MST} is the length of the *minimum spanning tree*, which is the shortest tree connecting all the given points without the use of extra points, which can be found in polynomial time [17]. We therefore look for solutions with low ρ values.

IV. LOCAL SEARCH OVER TOPOLOGY SPACE

In a previous work [11], we used the topology-describing vectors as the chromosomes of a genetic algorithm that yielded good results for small-sized ESTP instances ($p \approx 10$). The topology vectors are also suitable for local-search approaches, since it is easy to define neighborhood structures based on them. Here we use the following: given a topology vector $\boldsymbol{a}$, its neighbors will be the topology vectors $\boldsymbol{a}'$ that can be obtained from $\boldsymbol{a}$ by changing just one of its $p-3$ entries.

A local search may then be developed thus: Given an initial or current topology vector $\boldsymbol{s}$, related to a (relatively minimal) ESTP solution S of cost ρ, randomly choose an index i and a new value s_i, to create a neighboring vector $\boldsymbol{s}'$. Perform a minimization on $\boldsymbol{s}'$ (e.g., through Smith's minimization approach) to obtain S' and ρ'. If $\rho' < \rho$, then let $s_i \leftarrow s_i'$ (a move), $S \leftarrow S'$ and $\rho \leftarrow \rho'$, and restart the search. Else, try again with a different $\boldsymbol{s}'$. After a certain number of trials without improvement of the current solution, stop the search and output S.

Implementing this simple local search, we are able to use Smith's topology vectors with a wide variety of general purpose heuristics (metaheuristics), such as the microcanonical optimization algorithm [13,14], briefly described below.

V. MICROCANONICAL OPTIMIZATION ALGORITHM

The MO algorithm is a statistical-physics based metaheuristic that implements the simulation of a physical system evolving in equilibrium at fixed internal energy. The algorithm alternately applies two main procedures, called *initialization* and *sampling*.

The initialization performs a local search, accepting only improving solutions, either by choosing the first one that turns up, as implemented here, or by selecting the best among a given number of neighboring solutions. This phase ends when the solution gets stuck in a local minimum valley.

In the sampling phase, the MO algorithm tries to escape from the local minimum, but keeping a solution cost close to the value attained in the initialization. This is achieved through Creutz's microcanonical simulation [18], which generates samples of fixed-energy states. Creutz's technique introduces an extra degree of freedom, called the *demon*, which holds a variable (but always positive) energy load E_d that may be exchanged with the solution in such a way that the total energy $E_{total} \equiv E_s + E_d$ is kept constant, where E_s is the solution cost. Local state changes (solution moves) are attempted, and accepted whenever the demon, observing the constraint $E_d > 0$, is able to supply or to accept the ensuing energy balance $-\Delta E_s$, so as to preserve total energy. An upper bound, $E_d \leqslant E_{dmax}$, is also imposed on the demon, constraining the possible sampling solutions to evolve in a narrow energy shell. The sampling phase thus iteratively generates solutions in this shell, stopping after a preset number of iterations.

The MO algorithm then proceeds from the new current solution, alternating between initialization and sampling, until the stopping condition (see below) is achieved.

VI. IMPLEMENTATION AND RESULTS

We have developed the *C* code for a MO algorithm that implements, as its initialization phase, the local search described in Sec. IV. The sampling is likewise performed over the topology-vector space. Topology optimization is obtained through Smith's minimization procedure, as presented in Ref. [8]. Our initial solution is not arbitrary, but constructed through an adequate insertion of Steiner points in the minimum spanning tree topology, as prescribed in Ref. [19]. The main parameters of the algorithm are MaxInit, the maximum number of consecutive iterations without improvement of the initialization solution, which signals a local minimum and initialization stop; MaxSamp, the number of iterations at each sampling phase; and MaxCycles, the number of initialization and sampling cycles without improvement of the best solution so far, which signals program stop.

Following Ref. [14], we have also kept an ordered list of the moves rejected in the initialization, choosing its fifth lowest entry as both the demon's initial energy and its maximum capacity, E_{dmax}, for the subsequent sampling.

Computational tests were carried out for point configurations in several dimensions. Our results were compared to those obtained through Smith's exact algorithm [8] (for problems with $p\leqslant 11$), and through the *Soap Film* heuristic [12]. The latter is an extension, for dimensions $n\geqslant 3$, of a very fast algorithm, developed in the plane, which relates the ESTP to the dynamical evolution of a fluid film under surface tension forces [19]. All the results refer to implementations in a Sun Ultra 1 workstation. CPU time is reported, in order to allow speed comparison with the exact and Soap Film approaches, for which other metrics, such as number of generated cost functions, would not be meaningful.

In dimension $n=3$ (Table I), we considered four sets of 1000 different configurations of $p=8$ to $p=11$ points, randomly distributed in a unit cube. The MO solutions proved consistently superior to those of the Soap Film heuristic: their mean ρ values followed closer to those yielded by Smith's exact procedure, and the optima have actually been reached a significant number of times, compared to none, by the Soap Film heuristic. On the other hand, the relative sluggishness of our approach may be deemed reasonable, when compared to the explosive time requirement of the exact branch-and-bound algorithm.

In higher dimensions (Table II), we ran a series of tests with a fixed number of given points ($p=10$, randomly distributed in hypercubes) for several values of *n*. Table II highlights a feature of the ESTP—the progressive reduction of the mean ρ, as dimension increases, which has been observed in a series of previous experiments [12], and which seems in accordance with some conjectures about the general lower bounds of the Steiner ratio [20]. Both our exact and heuristic results display such behavior, with the MO's showing a more rapid decrease than the Soap Film's. Moreover, while, for the latter, the mean ρ remains roughly 0.5% above the exact value, irrespective of dimension, for MO algorithm it progresses from a difference of 0.18% in dimension $n=3$, to just around 0.035% above the exact ratio, in $n=10$. That is to say, the MO solutions actually improve with growing dimension.

TABLE I. Mean ρ value $(\bar{\rho})$, standard deviation (σ), mean CPU time in seconds (τ), and number of optimal solutions (Hits) found over a set of 1000 random three-dimensional distributions of *p* given points, where *p* varies from 8 to 11. MO parameter settings: MaxInit = 50, MaxSamp = 25 and MaxCycles = 5.

		p			
Algorithm		8	9	10	11
	$\bar{\rho}$	0.946761	0.946397	0.946758	0.946831
Exact	σ	0.019507	0.017754	0.017531	0.015840
	τ	1.4	7.8	43	240
	$\bar{\rho}$	0.950232	0.950600	0.951185	0.951379
Soap	σ	0.020169	0.018406	0.018003	0.016693
Film	τ	0.059	0.067	0.073	0.083
	Hits	0	0	0	0
	$\bar{\rho}$	0.947507	0.947494	0.948521	0.948724
MO	σ	0.019637	0.017884	0.017828	0.016434
	τ	22	28	34	37
	Hits	908	845	773	725

A clearer picture of this performance can be gleaned from Fig. 2, which refers to experiments with several instances of $p=50$ points, randomly distributed in hypercubes of dimensions 10, 20, 30, 40, and 50. Average time demands ranged from 25 min, in dimension $n=10$, to 90 min, in dimension $n=50$. For comparison, we also display the Soap Film results over the same problem set.

Once again, the general trend of reduction of the mean ρ values with growing *n* is evident, but more dramatic for the MO algorithm. And the difference, relative to the Soap Film

TABLE II. Results over ten-point instances in several dimensions. The number of different instances considered in each dimension appears in the last row. Notation and parameter settings as in Table I.

		n			
Algorithm		4	5	9	10
	$\bar{\rho}$	0.927784	0.911631	0.865916	0.860919
Exact	σ	0.018336	0.018687	0.020622	0.018949
	τ	120	250	1200	1500
	$\bar{\rho}$	0.932482	0.916838	0.870540	0.865272
Soap	σ	0.018823	0.019386	0.020738	0.019068
Film	τ	0.087	0.098	0.14	0.16
	Hits	42	143	17	18
	$\bar{\rho}$	0.929333	0.912812	0.866361	0.861223
MO	σ	0.018756	0.018930	0.020871	0.018936
	τ	39	41	54	62
	Hits	757	761	82	86
Instances		1000	1000	100	100

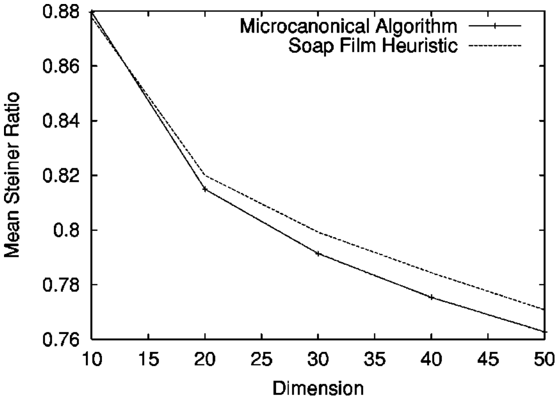

FIG. 2. Mean Steiner ratios evaluated over a set of ten 50-point instances, in dimensions 10, 20, 30, 40, and 50. The polygonal lines connect the results yielded by the MO algorithm (solid line) and by the Soap Film heuristic (dashed line). MO parameters: MaxInit = 100, MaxSamp = 50, and MaxCycles = 5.

heuristic, may not be slight: if we assume that the latter keeps on yielding good solutions—with Steiner ratios around 1% of the exact value—as n increases, then the MO results, at dimensions 30 and over, must indeed be very close to the optima, since their Steiner ratios are themselves about 1% lower than those of the Soap Film heuristic.

VII. APPLICATION TO PHYLOGENETIC INFERENCE

Other configurations and parameter settings have also been tested which corroborated the above general picture, leading us to consider the MO algorithm as a suitable approach for multidimensional clustering problems based on the Euclidean metric, such as those of phylogenetic inference.

As a real world example of this application, we took the classification problem in Ref. [21] (also treated in Ref. [7]), which deals with a set of 19 species of colonial marine invertebrates, called *gorgonians*, to be differentiated based on a set of 28 attributes. The attributes refer to the presence (1) or absence (0), in each of the species, of certain chemical compounds, the input data thus being organized as a 19×28 binary matrix.

In Ref. [21], the evolutionary tree depicted in Fig. 3 was generated from such data through a method due to Farris *et al.* [22], and proved to be in accordance with previous classifications of the gorgonians. Its internal nodes represent hypothetical ancestors in the evolutionary paths of the gorgonian species shown as leaves. The tree was rooted by explicitly introducing, as a common ancestor to all, the 28-dimensional zero vector.

In our approach to the same problem, the goal is to obtain, according to the *method of minimum evolution* [5], a minimal Euclidean Steiner tree connecting the 19 given points in a 28-dimensional space. The putative ancestors will then appear as the resulting Steiner points. With the MO algorithm, it took us 415 seconds to find a Steiner tree of ratio $\rho = 0.852\,818$.

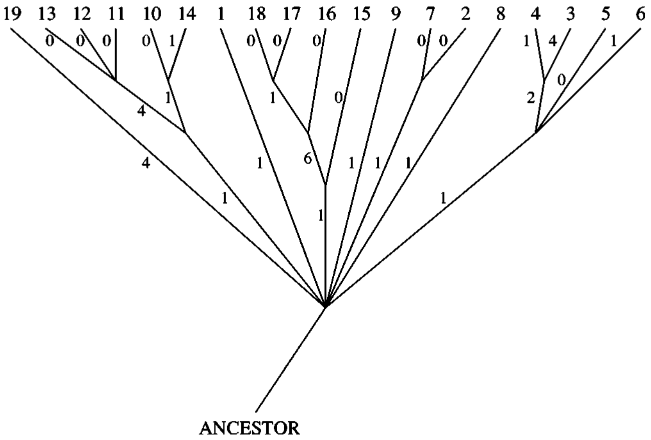

FIG. 3. Evolutionary tree for a set of 19 gorgonian species, as generated in Ref. [21]. Branch lengths represent the number of attributes not shared by the nodes at each branch end. The MO approach to the ESTP version of this problem yields exactly the same solution (Parameter settings: MaxInit = 90, MaxSamp = 20, and MaxCycles = 5).

In order to compare it to that of Ref. [21], we converted the Euclidean Steiner points to 0-1 coordinates, by rounding off to 1 the coordinate values larger than or equal to 0.5, and to 0 those smaller than 0.5. This yielded *exactly* the same tree as in Fig. 3, with the common ancestor also arising from the output data, as one of the Steiner points.

VIII. CONCLUDING REMARKS

We have introduced an effective statistical-mechanics based heuristic for the high-dimensional ESTP, applying microcanonical optimization to a local search in the space of Smith's topology-describing vectors. The efficiency of our approach was illustrated with random problem sets of up to 50 points in 50 dimensions, and with a real-world example of phylogenetic tree derivation.

With our work, we believe that we have established Smith's topology vectors as a suitable data structure on which to base general optimization approaches to the ESTP, at the same time confirming the MO algorithm as an effective metaheuristic for yet one more application. Incidentally, we should emphasize that very few algorithms seem to have been developed for high-dimensional ($n > 3$) ESTP applications. We know of only one more, due to Lundy [6], which has been reported in the literature, also in the context of phylogenetic inference. Unfortunately, its results refer to two problem instances—one with $p = 20$ and $n = 19$, and another with $p = 50$ and $n = 49$—whose data were not disclosed. As a means of indirect—and certainly inconclusive—comparison, we remark that, in similar dimensions, the MO algorithm usually provides Steiner trees that are, respectively, 1.8 and 0.6 standard deviations shorter than what was reported by Lundy for his $p = 20$ and $p = 50$ solutions. We hope our work will spur the development of other heuristics that can then be compared to ours.

[1] F.K. Hwang, D.S. Richards, and P. Winter, in *Annals of Discrete Mathematics* (North-Holland, Amsterdam, 1992), Vol. 53.

[2] M.R. Garey, R.L. Graham, and D.S. Johnson, SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **32**, 835 (1977).

[3] J. MacGregor Smith, R. Weiss, and M. Patel, Network **25**, 273 (1995).

[4] J. MacGregor Smith, R. Weiss, B. Toppur, and N. Maculan, in Proceedings of the II ALIO-EURO Workshop on Practical Combinatorial Optimization, Valparaiso, Chile, 1996, Vol. 1, pp. 37–44.

[5] L.L. Cavalli-Sforza and A.W.F. Edwards, Am. J. Hum. Genet., **19**, 233 (1967).

[6] M. Lundy, Biometrika **72**, 191 (1985).

[7] F. Montenegro, P. Antonelli, N. Maculan, S. Rutz, and E. Uchoa (unpublished); F. Montenegro, S.E. Gonçalves, and N. Maculan (unpublished).

[8] W.D. Smith, Algorithmica **7**, 137 (1992).

[9] N. Maculan, P. Michelon, and A.E. Xavier, Ann. Operat. Res. **96**, 209 (2000).

[10] M. Zachariasen, Eur. J. Oper. Res. **119**, 282 (1999).

[11] F. Montenegro and N. Maculan, in *Proceedings of the X Congreso Ibero-Latinoamericano de Investigación Operativa*, edited by L. Morales (Mexico City, Mexico, 2000).

[12] F. Montenegro, N. Maculan, G. Plateau, and P. Boucher, in *Essays and Surveys in Metaheuristics*, edited by C. Ribeiro and P. Hansen, Operations Research/Computer Science Interfaces Series, Vol. 15 (Kluwer, Boston, 2001), pp. 509–524.

[13] J.R.A. Torreão and E. Roe, Phys. Lett. A **205**, 377 (1995).

[14] A. Linhares and J.R.A. Torreão, Int. J. Mod. Phys. C **9**, 133 (1998).

[15] S. Kirkpatrick, C. Gelatt, and M. Vecchi, Science **220**, 671 (1983).

[16] E.N. Gilbert and H.O. Pollak, SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **16**, 323 (1968).

[17] R.C. Prim, Bell Syst. Tech. J. **36**, 1389 (1957).

[18] M. Creutz, Phys. Rev. Lett. **50**, 1411 (1983).

[19] F. Chapeau-Blondeau, F. Janez, and J.-L. Ferrier, SIAM J. Optim. **7**, 1037 (1997).

[20] D.Z. Du and W.D. Smith, J. Comb. Theory, Ser. A **74**, 115 (1996).

[21] D.J. Gerhart, Biol. Bull. **164**, 71 (1983).

[22] J.S. Farris, A.G. Kluge, and M.J. Eckardt, Syst. Zool. **19**, 172 (1970).