Intl. Trans. in Op. Res. 12 (2005) 325–338

INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

A GRASP/VND heuristic for the phylogeny problem using a new neighborhood structure

Celso C. Ribeiro and Dalessandro S. Vianna

Department of Computer Science, Catholic University of Rio de Janeiro, Rio de Janeiro, RJ 22453-900, Brazil
E-mail: celso@inf.puc-rio.br [C.C. Ribeiro]; vianna@inf.puc-rio.br [D.S. Vianna]

Received 23 May 2003; received in revised form 14 May 2004; accepted 19 October 2004

Abstract

A phylogeny is a tree that relates taxonomic units, based on their similarity over a set of characters. The phylogeny problem consists in finding a phylogeny with the minimum number of evolutionary steps. We propose a new neighborhood structure for the phylogeny problem. A greedy randomized adaptive search procedure heuristic based on this neighborhood structure and using variable neighborhood descent for local search is described. Computational results on randomly generated and benchmark instances are reported, showing that the new heuristic is quite robust and outperforms the other algorithms in the literature in terms of solution quality and time-to-target value.

Keywords: phylogeny problem; phylogenetic trees; evolutionary trees; GRASP; local search; VND; heuristics

1 The phylogeny problem

A phylogenetic tree (or a phylogeny) relates groups of species, populations of distinct species, populations of the same species, or homologous genes in populations of distinct species, indistinctly denoted by taxons (Ayala, 1995; Swofford and Olsen, 1990; Wiley et al., 1991). These relations are based on the similarity over a set of characters. Leaves represent the taxons. Interior nodes represent hypothetical ancestors.

Characters are independent attributes used to compare taxons. Each character takes values on a finite set of possible states. Each taxon is defined by its character states. Binary characters are those who have only two possible states, which represent the presence or the absence of some attribute. Instances of the phylogeny problem with binary characters are characterized by 0–1 matrices, in which each element (i, j) corresponds to the state of character j within taxon i.

Figure 1 illustrates an example extracted from Kitching et al. (1998) defined by a set of four taxons and six binary characters: (a) paired fins, (b) jaws, (c) large dermal bones, (d) fin rays, (e) lungs, and (f) rasping tongue. For each pair (i, j), an entry equal to one means that character

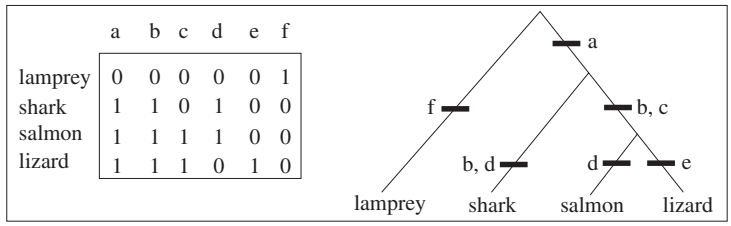

Fig. 1. Example with four taxons and six characters.

j appears in taxon *i*. We also show in this figure a possible phylogeny for these taxons, assuming that all of them have a common ancestor represented by the root of the tree, in which all characters do not appear. We also indicate the branches of this tree in which one or more characters change their states.

An evolutionary step is associated with each change of state along a branch of a phylogenetic tree. The evaluation of a phylogenetic tree can be done using different optimization criteria. Some criteria are based on stochastic models, on the comparison of distances in a metric space (which is often used in genome studies, see e.g. Araújo and Almeida, 2002; Gallut et al., 2000; Wang and Warnow, 2001), on the compatibility of the data or, most often, using the *parsimony criterion* (Swofford and Olsen, 1990). The latter states that the best (i.e., the most parsimonious) phylogeny is that explained by the minimum number of evolutionary steps (Edwards and Cavalli-Sforza, 1964; Hennig, 1966). It is often said that the parsimony criterion can be legitimated as the best one in the construction of phylogenetic trees, provided that the probability of the occurrence of evolutionary changes is small (Penny et al., 1982; Sober, 1987). The parsimony criterion is used in this work.

Figure 2 shows a phylogenetic tree for the above example. There are three hypothetical intermediary taxons (000000, 100000, 111000) used to explain the evolutionary changes represented in Fig. 1. A total of eight evolutionary steps are marked in the branches. The phylogenetic tree in Fig. 3 has more appropriate hypothetical intermediary taxons (000000, 110100, 111100) associated with the internal nodes, corresponding to a smaller parsimony value with only seven evolutionary steps.

Farris and Fitch (Farris, 1970; Fitch, 1971; Fitch and Farris, 1974) proposed polynomial algorithms running in time $O(mn)$ for computing the best parsimony value for a given phylogeny, where n is the number of operational characters and m is the number of binary characters. Given a set of taxons and a set of characters, the *phylogeny problem* studied in this work is that of finding a phylogenetic tree with the minimum number of evolutionary steps. It is NP-hard in general and in common restricted cases (Bodlaender et al., 1992; Day et al., 1986; Foulds and Graham, 1982a, 1982b).

Bader et al. (2001) surveyed industrial applications of high-performance computing for phylogeny reconstruction. According to them, ''simple identification via phylogenetic classification of organisms has yielded more patent filings than any other applications of phylogeny in industry''. They notice that phylogenetic analysis has also been used in vaccine development. Another application of phylogenetic analysis to a practical problem is its use in studying the dynamics of microbial communities (Engelen et al., 1998). Because many microbes in such

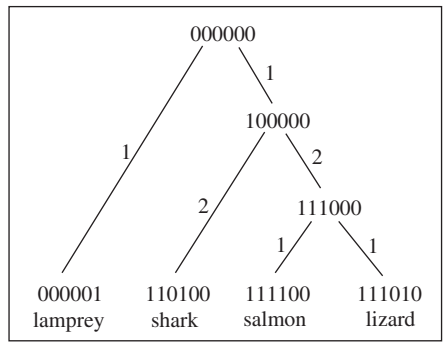

Fig. 2. Phylogenetic tree with eight evolutionary steps.

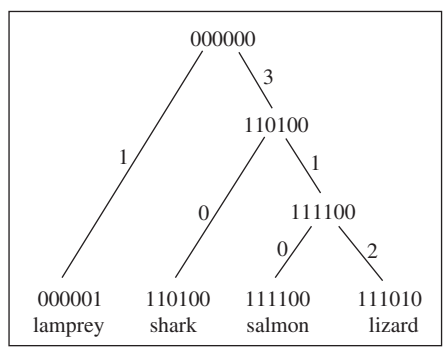

Fig. 3. A more parsimonious phylogeny with seven evolutionary steps.

population studies are novel, their gene sequences are studied phylogenetically in order to understand the composition of the community throughout the experiment. The phylogenetic distribution of biochemical pathways (Overbeek et al., 2000) is studied in the development of antibacterials and herbicides. In the pharmaceutical industry, the phylogenetic distribution of a pathway is often studied before a drug is developed in order to understand the effective range of an antimicrobial targeted at that pathway (Brown and Warren, 1998). Phylogenetic analysis is also used in the pharmaceutical industry for predicting the natural ligands for cell surface receptors which are potential drug targets (Chambers et al., 2000; Szekeres et al., 2000).

Let E be the set of all possible taxons and $W = \{x^{(1)}, \ldots, x^{(n)}\} \subseteq E$ be the set of operational taxons under analysis. A phylogenetic tree s for the operational taxons W belongs to the set S of all unrooted trees with n leaves (each leaf is in one-to-one correspondence with an operational taxon $x^{(i)} \in W$, $i = 1, \ldots, n$) and all internal nodes with degree three. Let $f : S \rightarrow \mathrm{R}$ be a function which associates each phylogeny $s \in S$ to its parsimony value. The phylogeny problem under parsimony may then be formulated as that of finding a phylogeny $s^* \in S$ such that $f(s^*) = \min_{s \in S} f(s)$.

Heuristics for the computation of phylogenetic trees are dispersed through the scientific literature (see e.g. Dress and Krüger, 1987; Luckow and Pimentel, 1985; Platnick, 1987, 1989).

Andreatta and Ribeiro (2002) reported and compared the computational results obtained by a variety of heuristics on a set of eight benchmark problems. They reported that the best-known solutions are not always found by the same heuristic. In Section 2, we propose a new neighborhood structure for the phylogeny problem and two local search procedures based on this neighborhood structure. A greedy randomized adaptive search procedure (GRASP) heuristic using a variable neighborhood descent (VND) strategy for local search is described in Section 3. Computational results obtained for randomly generated and benchmark instances are presented and discussed in detail in Section 4. In particular, we show that the GRASP heuristic using the new neighborhood structure is quite robust and outperforms the other heuristics in terms of solution quality and time-to-target value. Concluding remarks are drawn in the final section.

2 Neighborhood structure

Local search methods are based on the investigation of the solution space, by successively exploring the neighborhood of the current solution and moving to one of its neighbors.

Andreatta and Ribeiro (2002) reported and described three neighborhood relations for the phylogeny problem: *nearest neighborhood interchanges* (NNI: subtrees pending from two internal branches are swapped), *single step* (STEP: a neighbor is obtained by removing a taxon from the current solution and putting it back into another branch of the tree), and *subtree pruning and regrafting* (SPR or 1-SPR: a subtree of the current tree is disconnected and reconnected in a different position). This work is based on the SPR neighborhood.

Figure 4 illustrates an example of a move within neighborhood SPR. Neighbor solutions are obtained from the current solution as follows. First, an edge $u = (c, f)$ of the current tree is selected and eliminated (step 1, *pruning*). The subtree containing node c is called the base subtree, while that containing node f is the pending subtree. Node c is collapsed (since it has degree two in the base subtree) and an edge v of the base subtree is selected for the reconnection of the pending subtree (step 2). Edge v is eliminated and a new node h is created and joined to the two extremities of v originally in the base subtree (step 3). Finally, the pending subtree is connected to the base subtree using the newly created node h (step 4, *regrafting*). Any solution has $O(n^2)$ neighbors within this neighborhood.

Figure 5 gives the pseudo-code of procedure LS1 applied in the search for an improving neighbor of the current solution s within neighborhood SPR. The loop in lines 1–13 searches through this neighborhood by investigating the elimination of each branch of the current solution. Each branch u is temporarily eliminated from s in line 2, creating subtrees s_1 (base) and s_2 (pending). The partial costs of subtrees s_1 and s_2 are computed in line 3 in time $O(mn)$. The reconnection of the pending subtree using each edge of the base subtree is investigated in lines 4–8. Variable f' initialized in line 4 stores the minimum incremental cost over all neighbors of s obtained by the elimination of edge u. The loop in lines 5–8 searches through all edges v of the base subtree. The incremental cost $\bar{f}$ of reconnecting the pending subtree s_2 to the base subtree s_1 using edge v is computed in line 6 in time $O(m)$. If the incremental cost $\bar{f}$ improves the currently best incremental cost, then the latter is updated in line 7 and the best edge to be used for reconnection is stored in *vbest*. The best neighbor s' of the current solution s is built in line 9 and its cost $f(s')$ is computed in line 10. If $f(s')$ is smaller than the cost of the current solution s, then s'

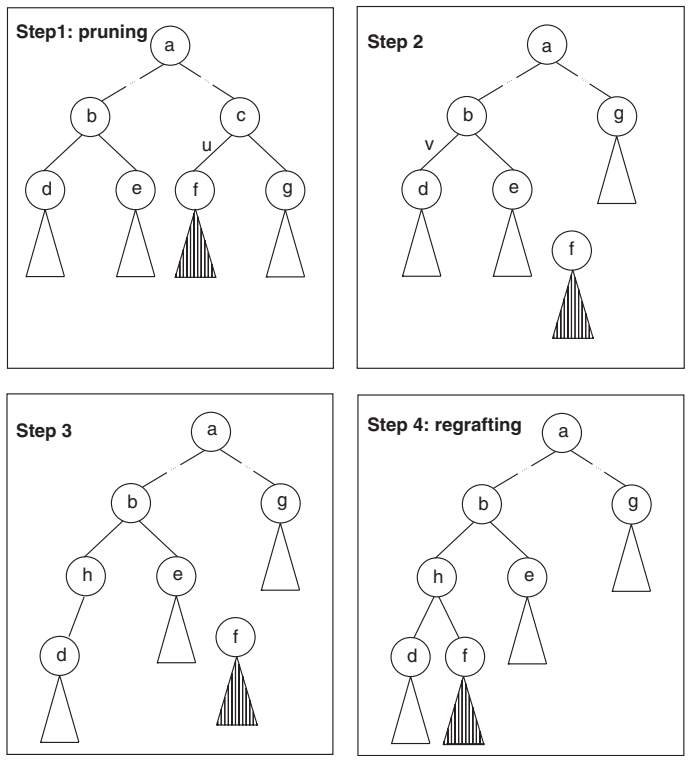

Fig. 4. Illustration of a move within neighborhood SPR.

```
procedure LS1(s)
1.    for each branch u of the current solution s do
2.        Tentatively remove u from s obtaining subtrees s1 and s2;
3.        Compute the partial costs f(s1) and f(s2) of subtrees s1 and s2;
4.        f' ← ∞;
5.        for each branch v of subtree s1 do
6.            Compute the incremental cost f̄ of reconnecting s2 to s1 using v;
7.            if f̄ < f' then f' ← f̄; vbest ← v;
8.        endfor;
9.        Obtain s' by reconnecting s2 to s1 using edge vbest;
10.       Compute the cost f(s') = f(s1) + f(s2) + f' of the new solution s';
11.       if f(s') < f(s) then return s';
12.       Restore solution s;
13.   endfor;
14.   return s;
end LS1.
```

Fig. 5. Pseudo-code of local search procedure LS1 using neighborhood SPR.

is returned in line 11. Otherwise, edge u is reinserted and the current solution s is restored in line 12. If no improving neighbor is found, the current solution itself is returned in line 14 at the end of the neighborhood search. Thus, the investigation of each neighborhood SPR can be implemented in time $O(mn^2)$, which is one order of magnitude faster than the implementation originally proposed in Andreatta and Ribeiro (2002).

The *Multiple Subtree Pruning and Regrafting* (ℓ-SPR) neighborhood is defined by the composition of ℓ successive SPR moves applied to the current solution. The case $\ell = 1$ corresponds to neighborhood SPR described above. Figure 6 gives the pseudo-code of procedure LS2 used in the search for an improving neighbor of the current solution s within neighborhood 2-SPR, in which each neighbor solution is obtained through a move involving two steps. Lines 1–11 correspond to the first step, in which intermediary solutions within neighborhood SPR are investigated. Lines 12–22 correspond to the second step, in which one additional SPR move is applied to the intermediary solution. Similarly to LS1, the investigation of each neighborhood 2-SPR can be implemented in time $O(mn^3)$.

```
procedure LS2(s)
1.   for each branch u of the current solution s do
2.        Tentatively remove u from s obtaining subtrees s1 and s2;
3.        Compute the partial costs f(s1) and f(s2) of subtrees s1 and s2;
4.        f' ← ∞;
5.        for each branch v of subtree s1 do
6.             Compute the cost f̄ of reconnecting s2 to s1 using v;
7.             if f̄ < f' then f' ← f̄; vbest ← v;
8.        endfor;
9.        Obtain s' by reconnecting s2 to s1 using edge vbest;
10.       Compute the cost f(s') = f(s1) + f(s2) + f' of the new solution s';
11.       if f(s') < f(s) then return s';
12.       for each branch u' of the intermediary solution s' do
13.            Tentatively remove u' from s' obtaining subtrees s'1 and s'2;
14.            Compute the partial costs f(s'1) and f(s'2) of subtrees s'1 and s'2;
15.            f' ← ∞;
16.            for each branch v' of subtree s'1 do
17.                 Compute the cost f̄ of reconnecting s'2 to s'1 using v';
18.                 if f̄ < f' then f' ← f̄; vbest ← v';
19.            endfor;
20.            Obtain s'' by reconnecting s'2 to s'1 using edge vbest;
21.            Compute the cost f(s'') = f(s'1) + f(s'2) + f' of the new solution s'';
22.            if f(s'') < f(s) then return s'';
23.            Restore solution s';
24.       endfor;
25.       Restore solution s;
26.  endfor;
27.  return s;
end LS2.
```

Fig. 6. Pseudo-code of local search procedure LS2 using neighborhood 2-SPR.

3 A GRASP/VND heuristic

GRASP (Feo and Resende, 1995; Resende and Ribeiro, 2003) is a multi-start metaheuristic, in which each iteration consists of two phases: construction and local search. The construction phase builds a feasible solution using a greedy randomized algorithm, whose neighborhood is investigated until a local minimum is found during the local search phase. The best overall solution is kept as the result.

3.1 Greedy randomized construction

A phylogeny $s \in S$ relating the taxons in W can be built in $n = |W|$ iterations, as outlined in the pseudo-code in Fig. 7. In the k-th iteration, a partial phylogeny $s^{(k)}$ (defined on a subset $U \subset W$ of operational taxons) is modified by the introduction of a new taxon $t \in W \setminus U$. The algorithm stops when $U = W$. Variants of this algorithm differ by the criteria they use to select a new taxon to be inserted and by the way in which $s^{(k)}$ is modified to obtain $s^{(k+1)}$. The set of modifications applied to $s^{(k)}$ to reach $s^{(k+1)}$ is called an increment. The increase in the cost function $f(.)$ because of the increment leading from $s^{(k)}$ to $s^{(k+1)}$ can be computed in time $O(mk)$, where m is the number of binary characters and k the number of taxons in the current partial phylogeny.

Increments are defined by the insertion of a new taxon into a branch of the current partial solution, as illustrated in Fig. 8. In this case, there are three possible alternatives for the insertion of taxon D into a partial solution formed by three taxons A, B, and C.

Andreatta and Ribeiro (2002) conducted a detailed evaluation study of construction algorithms for the phylogeny problem, in which they tested and compared several variants of the basic construction algorithm described in Fig. 7. We used algorithm Gstep_wR (greedy step with randomness) in the construction phase of our GRASP heuristic, since in general it found the best solutions (although at the cost of computation times one order of magnitude higher than the other algorithms). A pair taxon-branch is randomly selected from among all those with cost 10% higher than the most parsimonious increment value. Since there are still $n-(k-1)$ unselected taxons in iteration k and $2k-5$ possible branches for each insertion, the overall complexity of each construction using algorithm Gstep_wR is $O(mn^4)$.

```
procedure BuildPhylogeny(s);
1.   s ← ∅; U ← ∅; k ← 1;
2.   while k ≤ n do
3.        Select a taxon t ∈ W\U;
4.        U ← U ∪ {t};
5.        Modify the partial phylogeny s by inserting taxon t;
6.        k ← k + 1;
7.   endwhile;
8.   return s;
end BuildPhylogeny.
```

Fig. 7. Basic construction algorithm.

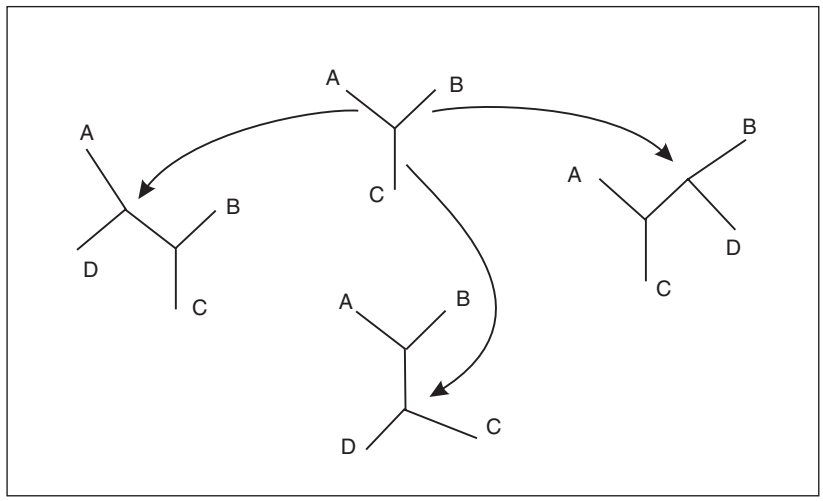

Fig. 8. Alternatives for the insertion of a new taxon into a phylogeny with three taxons.

3.2 Local search using VND

Let ℓ-SPR, $\ell = 1,\ldots,\ell_{\max}$, be a neighborhood structure for the phylogeny problem. We use a local search procedure based on variable neighborhoods with $\ell_{\max} = 2$, which essentially is a variant of the VND strategy proposed by Hansen and Mladenović (Hansen and Mladenović, 1999, 2002, 2003; Mladenović and Hansen, 1997). Successful applications of VND within GRASP are reported e.g. in Festa et al., 2002; Martins et al., 2000; Ribeiro and Souza, 2002; Ribeiro et al., 2002.

Figure 9 gives the algorithmic description of procedure VND_Phylogeny which implements the VND local search starting from the solution s built in the construction phase. The initial neighborhood 1-SPR is set in line 1. The loop in lines 2–7 investigates one neighborhood at a time, until a local optimum with respect to neighborhoods 1-SPR and 2-SPR is found. The best solution s^{neighbor} within neighborhood ℓ-SPR is obtained by the application of procedure LSℓ to the current solution s in line three. In case an improving move is found, the current solution is updated in line 4 and the search resumes from the latter using neighborhood 1-SPR. Otherwise, the order of the neighborhood is increased by one in line 5, so as that the search will resume from neighborhood $(\ell+1)$-SPR. The current solution s that is now locally optimal with respect to all neighborhoods is returned in line 8.

3.3 Heuristic GRASP+VND

The pseudo-code in Fig. 10 based on the template described by Resende and Ribeiro (2003) illustrates the main blocks of a GRASP heuristic for the phylogeny problem. The algorithm takes as parameters the number MaxIterations of iterations and the value Seed used as the initial seed for the pseudo-random number generator. The loop in lines 1–5 performs MaxIterations iterations. Algorithm Gstep_wR proposed in Andreatta and Ribeiro (2002) and described in Section 3.1 is used in the construction phase in line 2 with the GRASP parameter set at $\alpha = 0.1$.

```
procedure VND_Phylogeny(s);
1.  ℓ ← 1;
2.  while ℓ ≤ 2 do
3.      s^neighbor ← LSℓ(s);
4.      if f(s^neighbor) < f(s) then s ← s^neighbor; ℓ ← 1;
5.      else ℓ ← ℓ + 1;
6.      endif;
7.  endwhile;
8.  return s;
end VND_Phylogeny.
```

Fig. 9. Variable neighborhood descent (VND) procedure for local search.

```
procedure GRASP+VND(MaxIterations, Seed)
1.  for k = 1, ..., MaxIterations do
2.      Solution ← Gstep_wR(Seed);
3.      Solution ← VND_Phylogeny(Solution);
4.      UpdateSolution(Solution,BestSolution);
5.  endfor;
6.  return BestSolution;
end GRASP+VND.
```

Fig. 10. Pseudo-code of the greedy randomized adaptive search procedure (GRASP)+variable neighborhood descent (VND) heuristic.

The VND local search strategy using $\ell_{\max} = 2$ and neighborhoods 1-SPR and 2-SPR is implemented in line 3, as described in Section 3.2. The best solution found is updated in line 4 at each iteration and returned in line 6.

4 Numerical results

All computational experiments have been performed on a 2 GHz Pentium IV processor with 512 Mbytes of RAM memory. The GRASP+VND heuristic was implemented in C using version 6.0 of the Microsoft Visual C++ compiler. We used an implementation in C of the random number generator described in Schrage (1979).

The characters of each solution are represented as integer-valued vectors, in which each position uses 32 bits. Since each character needs two bits to be represented (possible states are 0, 1, and ?, the latter standing for "undefined"), each position of this vector is able to store up to 16 characters. Thus, binary operations may handle 16 characters simultaneously. This single modification accounted for reductions of up to approximately 50% in the computation times, with respect to the original implementation using the *Searcher* framework described in Andreatta and Ribeiro (2002).

In the first part of the computational experiments, we built 20 randomly generated instances. Our generator takes as parameters the number of taxons, the number of characters, and the ratio

of indefinition, which corresponds to the fraction of undefined characters in each taxon. Instances with larger ratios of indefinition are harder. The number of taxons in these instances ranges from 45 to 75, the number of characters from 61 to 159, and the ratio of indefinition from 20% to 50%. For each instance, we first report in Table 1 its identification, the number of taxons (n), the number of characters (m), and the percentage ratio of indefinition in the characters of each taxon. For both the GRASP algorithm described in Andreatta and Ribeiro (2002) and the GRASP+VND heuristic proposed in this work (with the number of GRASP iterations fixed at MaxIterations = 50 for both algorithms), we report the computation times and the best solutions found. The new heuristic found the best solution among the two algorithms for all but only one instance (TST06). The average improvement in the solution value is approximately 1%. Moreover, the computation times observed with the new heuristic are significantly smaller for all test instances.

In the second part of the computational experiments, we used the same eight benchmark real-life test instances (Luckow and Pimentel, 1985; Platnick, 1987, 1989) already used in Andreatta and Ribeiro (2002). All benchmark instances but GOLO were obtained from the editors of the journal *Cladistics*, while the latter was provided by P.A. Goloboff. All test instances are posted at http://www.inf.puc-rio.br/ ~ celso/grupo_de_pesquisa.htm. Each instance was run ten times with different seeds. For each test problem, we first report in Table 2 its identification, the number of taxons (n), the number of characters (m), the parsimony value of the currently best-known solution, the value of the best solution found by the GRASP+VND heuristic over the ten runs,

Table 1
Comparative results on randomly generated problems

Instance	n	m	Indefinition (%)	GRASP Time (s)	GRASP Value	GRASP+VND Time (s)	GRASP+VND Value
TST01	45	61	20	530.30	558	526.86	551
TST02	47	151	30	1560.00	1377	663.47	1364
TST03	49	111	40	1731.44	851	687.69	845
TST04	50	97	50	1614.34	605	779.39	598
TST05	52	75	20	1129.05	807	736.72	797
TST06	54	65	30	1357.53	608	960.48	609
TST07	56	143	40	3746.81	1304	1040.93	1291
TST08	57	119	50	3082.61	881	1164.59	870
TST09	59	93	20	2700.21	1167	1344.66	1152
TST10	60	71	30	2550.61	734	1454.67	733
TST11	62	63	40	2649.65	557	1520.33	553
TST12	64	147	50	6715.86	1250	2055.82	1243
TST13	65	113	20	4975.89	1545	2499.95	1532
TST14	67	99	30	5528.49	1186	2820.38	1177
TST15	69	77	40	4946.88	782	2974.00	774
TST16	70	69	50	4885.68	556	3309.26	551
TST17	71	159	20	8002.02	2481	3886.32	2468
TST18	73	117	30	8125.33	1568	3774.29	1554
TST19	74	95	40	7331.50	1042	3558.12	1036
TST20	75	79	50	6884.55	694	3884.26	682

and the average computation time in seconds taken by algorithm GRASP+VND. The number of GRASP iterations was set at MaxIterations = 500. These results show that the GRASP+VND heuristic is quite robust. It not only improved the best-known solution for three out of the eight test instances (ROPA, GOLO, SCHU), but also matched the best-known solutions for all other instances. We recall that the best-known solutions previously reported in the literature for each test problem were not found by the same algorithm.

To effectively compare the new GRASP+VND heuristic with the original GRASP algorithm in Andreatta and Ribeiro (2002), we compared the behavior of both algorithms on instances ROPA and GOLO using the methodology proposed by Aiex et al. (2002) and recently reviewed by Resende and Ribeiro (2003). One hundred independent runs for each heuristic were done for each instance. Each run was terminated when a solution of value less than or equal to a certain target value was found. The target values were set at the previously best-known solution values for each of these two instances, i.e., 326 for instance ROPA and 497 for instance GOLO. Although each of these sub-optimal values was chosen such that the slowest heuristic could terminate in a reasonable amount of computation time, the relative behavior of the two heuristics is not affected by this choice. Empirical probability distributions for the time-to-target value are plotted in Figs 11 and 12. To plot the empirical distribution for each algorithm, we follow the procedure described in Aiex et al. (2002). We associate with the i-th smallest running time t_i a probability $p_i = (i - \frac{1}{2})/100$, and plot the points $z_i = (t_i, p_i)$, for $i = 1, \ldots, 100$.

The plots in these figures show that the GRASP+VND heuristic is approximately two times of magnitude faster than the GRASP implementation in Andreatta and Ribeiro (2002), clearly showing the improvement because of the use of the VND local search strategy. The new heuristic clearly outperformed that in Andreatta and Ribeiro (2002): for a given computation time, the probability of finding a solution at least as good as the target value is much higher for the GRASP+VND heuristic.

5. Final remarks

The phylogeny problem is one of the most important problems in comparative biology. Approximate and exact (for small problems) algorithms for the computation of phylogenetic trees are dispersed through the scientific literature. We proposed in this paper a new heuristic for the phylogeny

Table 2
Results obtained by the GRASP+VND heuristic

Instance	n	m	Current	Best	Time (s)
ANGI	49	59	216	216	5099.33
GRIS	47	93	172	172	3505.43
TENU	56	179	682	682	7497.60
ETHE	58	86	372	372	10 042.73
ROPA	75	82	326	(*) 325	15 764.93
GOLO	77	97	497	(*) 496	32 836.65
SCHU	113	146	760	(*) 759	113 391.30
CARP	117	110	548	548	82 176.60

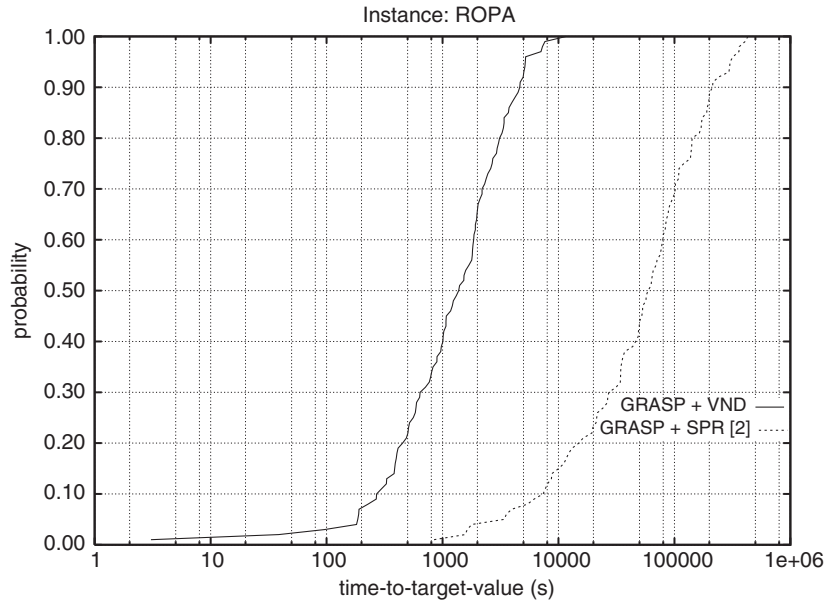

Fig. 11. Empirical probability distributions of time-to-target value on instance ROPA for the GRASP algorithm in Andreatta and Ribeiro (2002) and the GRASP+VND heuristic.

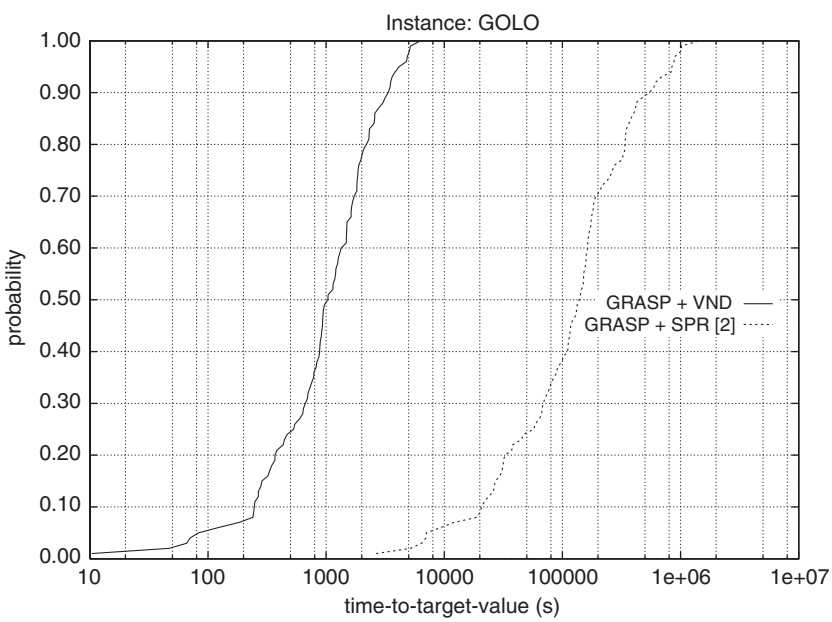

Fig. 12. Empirical probability distributions of time-to-target value on instance GOLO for the GRASP algorithm in Andreatta and Ribeiro (2002) and the GRASP+VND heuristic.

problem. This algorithm combines the GRASP metaheuristic with a VND local search strategy based on a new neighborhood structure called *multiple subtree pruning and regrafting* (ℓ-SPR).

Computational experiments on randomly generated and real-life problems are reported. The new heuristic outperformed the best algorithm in a recent survey (Andreatta and Ribeiro, 2002), finding better solutions for 19 out of 20 realistic randomly generated instances in much smaller computation times. The new heuristic is very robust. It improved the best-known solutions for three out of eight benchmark real-life instances, and matched the best results for the others.

Acknowledgments

The authors are grateful to one anonymous referee for several constructive remarks that improved the presentation of this work.

References

Aiex, R.M., Resende, M.G.C., Ribeiro, C.C., 2002. Probability distribution of solution time in GRASP: An experimental investigation. *Journal of Heuristics*, 8, 343–373.

Andreatta, A.A., Ribeiro, C.C., 2002. Heuristics for the phylogeny problem. *Journal of Heuristics*, 8, 429–447.

Araújo, G.S., Almeida, N.F., 2002. Phylogeny from whole genome comparison. *Proceedings of the 1st Brazilian Workshop on Bioinformatics,* Gramado, pp. 9–15.

Ayala, F.J., 1995. The myth of Eve: Molecular biology and human origins. *Science*, 270, 1930–1939.

Bader, D.A., Moret, B.M.E., Vawter, L., 2001. Industrial applications of high-performance computing for phylogeny reconstruction. *SPIE*, 4528, 159–168.

Bodlaender, H., Fellows, M., Warnow, T., 1992. Two strikes against perfect phylogeny. In: *Proceedings of the 19th International Colloquium on Automata, Languages and Programming*, Lecture Notes in Computer Science, 623, 273–283. Springer-Verlag, Berlin.

Brown, J., Warren, P., 1998. Antibiotic discovery: is it in the genes? *Drug Discovery Today*, 3, 564–566.

Chambers, J.J., Macdonald, L., Sarau, H., Ames, R., Freeman, K., Foley, J., Zhu, Y., McLaughlin, M., Murdock, P., McMillan, L., Trill, J., Swift, A., Aiyar, N., Taylor, P., Vawter, L., Naheed, S., Szekeres, P., Hervieu, G., Scott, C., Watson, J., Murphy, A., Duzic, E., Klein, C., Bergsma, D., Wilson, S., Livi, G., 2000. A G protein-coupled receptor for UDP-glucose. *Journal of Biological Chemisry*, 275, 10767–10771.

Day, W.H.E., Johnson, D.S., Sankoff, D., 1986. The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical Biosciences*, 81, 33–42.

Dress, A., Krüger, M., 1987. Parsimonious phylogenetic trees in metric spaces and simulated annealing. *Advances in Applied Mathematics*, 8, 8–37.

Edwards, A., Cavalli-Sforza, L., 1964. Reconstruction of evolutionary trees. *Phenetic and Phylogenetic Classification*, 6, 67–76.

Engelen, B., Meinken, K., von Wintzingerode, F., Heuer, H., Malkomes, H.-P., Backhaus, H., 1998. Monitoring impact of a pesticide treatment on bacterial soil communities by metabolic and genetic fingerprinting in addition to conventional testing procedures. *Applied Environmental Microbial*, 64, 2814–2821.

Farris, J.S., 1970. Methods for computing Wagner trees. *Systematic Zoology*, 19, 83–92.

Feo, T.A., Resende, M.G.C., 1995. Greedy randomized adaptative search procedures. *Journal of Global Optimization*, 6, 109–133.

Festa, P., Pardalos, P.M., Resende, M.G.C., Ribeiro, C.C., 2002. Randomized heuristics for the max-cut problem. *Optimization Methods and Software*, 6, 1033–1058.

Fitch, W.M., 1971. Towards defining the course of evolution: minimum chances for a specific tree topology. *Systematic Zoology*, 20, 406–419.

Fitch, W.M., Farris, J.S., 1974. Evolutionary trees with minimum nucleotide replacements from amino acid sequences. *Journal of Molecular Evolution*, 3, 263–278.

Foulds, L.R., Graham, R.L., 1982a. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3, 43–49.

Foulds, L.R., Graham, R.L., 1982b. Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. *Mathematical Biosciences*, 60, 133–142.

Gallut, C., Barriel, V., Vignes-Lebbe, R., 2000. Gene order and phylogenetic information. *Computational Biology*, 1, 123–132.

Hansen, P., Mladenović, N., 1999. An introduction to variable neighbourhood search. In Voss, S., Martello, S., Osman, I.H., Roucairol, C. (eds) *Metaheuristics: Advances and Trends in Local Search Procedures for Optimization*. Kluwer, Dordrecht, pp. 433–458.

Hansen, P., Mladenović, N., 2002. Developments of variable neighbourhood search. In Ribeiro, C.C., Hansen, P. (eds) *Essays and Surveys in Metaheuristics*. Kluwer, Dordrecht, pp. 415–440.

Hansen, P., Mladenović, N., 2003. Variable neighbourhood search. In Glover, F., Kochenberger, G. (eds) *Handbook of Metaheuristics*. Kluwer, Dordrecht, pp. 145–184.

Hennig, E., 1966. *Phylogenetic Systematics*. University of Illinois Press, Urbana.

Kitching, I.J., Forey, P.L., Humphries, C.J., Williams, D.M., 1998. *Cladistics: The Theory and Practice of Parsimony Analysis*. Oxford University Press, London.

Luckow, M., Pimentel, R.A., 1985. An empirical comparison of numerical Wagner computer programs. *Cladistics*, 1, 47–66.

Martins, S.L., Pardalos, Resende, M.G.C., Ribeiro, C.C., 2000. A parallel GRASP for the Steiner tree problem in graphs using a hybrid local search strategy. *Journal of Global Optimization*, 17, 267–283.

Mladenović, N., Hansen, P., 1997. Variable neighbourhood search. *Computers and Operations Research*, 24, 1097–1100.

Overbeek, R., Larsen, N., Pusch, G., D'Souza, M., Selkov, E. Jr., Kyrpides, N., Fonstein, M., Maltsev, N., Selkov, E., 2000. WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research*, 28, 123–125.

Penny, D., Foulds, L.R., Hendy, M.D., 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature*, 247, 197–200.

Platnick, N.I., 1987. An empirical comparison of microcomputer parsimony programs. *Cladistics*, 3, 121–144.

Platnick, N.I., 1989. An empirical comparison of microcomputer parsimony programs, II. *Cladistics*, 5, 145–161.

Resende, M.G.C., Ribeiro, C.C., 2003. Greedy randomized adaptive search procedures. In Glover, F., Kochenberger, G. (eds) *Handbook of Metaheuristics*. Kluwer, Dordrecht, pp. 219–249.

Ribeiro, C.C., Souza, M.C., 2002. Variable neighborhood search for the degree-constrained minimum spanning tree problem. *Discrete Applied Mathematics*, 118, 43–54.

Ribeiro, C.C., Uchoa, E., Werneck, R.F., 2002. A hybrid GRASP with perturbations for the Steiner problem in graphs. *INFORMS Journal on Computing*, 14, 228–246.

Schrage, L., 1979. A more portable FORTRAN random number generator. *ACM Transactions on Mathematical Software*, 5, 132–138.

Sober, E., 1987. Parsimony likelihood and the principle of the common cause. *Philosophy of Science*, 54, 465–469.

Swofford, D.L., Olsen, G., 1990. Phylogeny reconstruction. In: Hillis, D.M., Moritz, C. (eds) *Molecular systematics*. Sinauer, Sunderland.

Szekeres, P., Muir, A., Spinage, L., Miller, J., Butler, S., Smith, A., Rennie, G., Murdock, P., Fitzgerald, L., Wu, H., McMillan, L., Guerrera, S., Vawter, L., Elshourbagy, N., Mooney, J., Bergsma, D., Wilson, S., Chambers, J., 2000. Neuromedin U is a potent agonist at the orphan G protein-coupled receptor FM3. *Journal of Biological Chemistry*, 275, 20247–20250.

Wang, L., Warnow, T., 2001. New polynomial time methods for whole genome phylogeny reconstruction. *DIMACS Whole Genome Comparison Workshop, Piscataway*.

Wiley, E.O., Siegel-Causey, D., Brooks, D.R., Funk, V.A., 1991. *The compleat cladist: A primer of phylogenetic procedures*, Special publication no. 19, University of Kansas, Museum of Natural History.