Available online at www.sciencedirect.com
SCIENCE DIRECT

Theory in Biosciences 123 (2005) 277–299

Theory in Biosciences
www.elsevier.de/thbio

MATH/CHEM/COMP2004

Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information

Daniel Merkle*, Martin Middendorf

Department of Computer Science, University of Leipzig, Augustusplatz 10-11, D-04109 Leipzig, Germany

Received 10 January 2005; accepted 24 January 2005

Abstract

In this paper, we present a method and a corresponding tool called Tarzan for cophylogeny analysis of phylogenetic trees where the nodes are labelled with divergence timing information. The tool can be used for example to infer the common history of hosts and their parasites, of insect–plant relations or symbiotic relationships. Our method does the reconciliation analysis using an event-based concept where each event is assigned a cost and cost minimal solutions are sought. The events that are used by Tarzan are cospeciations, sortings, duplications, and (host) switches. Different from existing tools Tarzan can handle more complex timing information of the phylogenetic trees for the analysis. This is important because several recent studies of cophylogenetic relationships have shown that timing information can be very important for the correct interpretation of results from cophylogenetic analysis. We present two examples (one host–parasite system and one insect–plant system) that show how divergence timing information can be integrated into reconciliation analysis and how this influences the results.

Keywords: Reconciliation analysis; Phylogeny; Cospeciation; Hosts; Parasites

*Corresponding author. Tel.: +49 341 9732275; fax: +49 341 9732329.
E-mail addresses: merkle@informatik.uni-leipzig.de (D. Merkle), middendorf@informatik.uni-leipzig.de (M. Middendorf).

1431-7613/$ - see front matter © 2005 Elsevier GmbH. All rights reserved.
doi:10.1016/j.thbio.2005.01.003

Introduction

Coevolutionary systems like hosts and their parasites or plants and insects that feed or breed on these plants are commonly used model systems for evolutionary studies. One central aspect in the study of these systems is the evolutionary history of the relations between the two involved groups of species. Often it is possible to determine the phylogeny of each group of species (e.g., the parasite species and the host species) using genetic or morphological data. The current relations between the species from both groups (e.g., which parasite lives on which host species) are often known from field studies or laboratory experiments. An interesting problem is then to reconstruct the common history from the known phylogenies and the current relationships (for an overview see, e.g., Page, 2002). One approach to solve this problem is to use an evolutionary model that describes the set of possible events that can happen during the coevolution and to assign costs to every event. The problem is then to find a minimum-cost history. Such methods have been called event-based methods in Ronquist (2002). The following four different types of events are typically considered for studying different cophylogenetic systems, e.g., host–parasite systems: cospeciation events, duplication events, sorting events and switching events. Cospeciation events refer to simultaneous host and parasite speciation, duplication events are independent parasite speciations, sorting events correspond to lineage sorting, and switches correspond to host shifts. There exist two main tools that are commonly used and provide event-based methods for the analysis of coevolving species associations: TreeMap and TreeFitter.

TreeMap has been developed by Charleston and Page (2002). The tool takes into account all the four above-mentioned coevolutionary events. It uses a data structure that is called Jungle (see Charleston, 1998). A Jungle is a graph-based structure that is used to store possible pairs of associations of hosts and their parasites. A reconstruction of a common cophylogenetic history corresponds to a subgraph in the Jungle. The newest version 2.0.2β of TreeMap takes, unlike its predecessors, timing information on the host tree or the associate tree into account. The timing model is based on ranking information of the nodes and is computed from an ultrametric on the trees. Unfortunately, not much details are given how the ranking information is used exactly. An advantage of TreeMap is that it supports some statistical analysis on the frequency of evolutionary events. A disadvantage of TreeMap is that it is very slow and uses much memory so that scenarios of moderate size with more than 3 switches might not be feasible with usual computational resources (e.g., Sorenson et al., 2004).

TreeFitter (Ronquist, 2001) is a program that can handle arbitrary cost assignments where duplication, sorting, and switch events have zero or positive costs associated with them. Cospeciation events can be associated with either positive or negative cost (or zero cost). Unfortunately, TreeFitter has no graphical user interface that makes the use of the program not very comfortable. An advantage of TreeFitter is that it can resolve associations of parasites with multiple hosts.

A problem with the existing tools for event-based cophylogeny analysis is that they base the computation of the cost minimal reconstructions of the coevolutionary history nearly only on the number of different types of events but can include additional information only in a very restricted sense. But for coevolutionary studies additional information is often important. In particular, divergence timing information is often relevant to obtain realistic scenarios. There exists several cophylogenetic studies (e.g., Jeong et al., 1999; Percy et al., 2004; Sorenson et al., 2004) where divergence timing information shows that the minimal cost reconstructions that are obtained with the existing tools are not possible.

In this paper, we present a new tool called Tarzan for event-based cophylogeny analysis. In contrast to the existing tools Tarzan can handle divergence timing information that includes inexact information in the form of time zone intervals. Tarzan also uses a more general model of evolution than the only other tool TreeMap that considers timing information. Hence, Tarzan can find reconstructions for many relevant cases when TreeMap does not find a possible reconstruction. For cost minimal reconstructions only associations between nodes of two phylogenetic trees are allowed that are not impossible according to the timing information. Since exact time measures are often not available we assume that divergence timing information is given in the form of time zones and nodes can be assigned to time zones. Thus, every node in the given phylogenetic trees can have an associated time zone. In this case, associations between two nodes of the trees are valid only when they are in the same zone. Associations between a node of one tree and an edge in the other tree are only valid when the time zone of the first node lies between the time zones of the two nodes that are connected by the edge.

In applications it might be difficult to decide about the exact time zone in which a divergence event that corresponds to a node has happened. Therefore, Tarzan allows to assign intervals of time zones to the nodes in one of the trees, e.g., the parasite tree. But the nodes in the other tree, e.g., the host tree, have to be assigned to a single time zone. The reason for this is that the reconstruction problem becomes much more complex when nodes in both trees are assigned to time zone intervals.

We apply Tarzan to analyse two cophylogenetic systems that have been studied in the literature and show how the concept of time zones can be used. One system is the cophylogeny of African brood parasite finches (*Vidua* sp.) and their finch hosts (family Estrildae) and the information on this system is taken from Sorenson et al. (2004). Information on the other system stems from Percy (2001) which consider phytophagous Psyllids (Aritaininae, Hemiptera) and their Legume hosts (Genisteae, Leguminosa). The results for both systems show that the cophylogenetic analysis done in both papers can be extended using Tarzan since the divergence timing information that is available in the papers can then be included in the reconciliation analysis. This leads to more realistic reconstructions than can be obtained with the existing tools.

This paper is organized as follows. In section Coevolutionary model and finding reconstructions we describe our model of coevolution, how divergence timing information is included and how cheapest reconstructions are computed in Tarzan. The tool Tarzan is described in Section Tarzan. The application of Tarzan for the analysis of the two example coevolutionary systems is presented in Section Applications. Conclusions are given in the last section.

Coevolutionary model and finding reconstructions

In this section, we describe the method how a cheapest reconstruction of the common phylogeny of two phylogenetic trees (i.e., rooted binary trees) can be computed when information about divergence times is given. As an example we use a host tree and a parasite tree. Before the method used by Tarzan to find cheapest reconstructions is described we discuss coevolutionary events, reconstructions, and the concepts for representing divergence timing information. It is assumed in this section that two phylogenetic trees H and P are given. H and P will be called host tree, respectively, parasite tree in the following. Further, a mapping ϕ of the leaves of P to nodes of H is given which describes the relationship of the extent species in P to species in H. Usually, ϕ is a mapping into the leaves in H, but in some cases it can be convenient to consider more general mappings.

Coevolutionary events

Event-based methods for the reconstruction of the coevolutionary history of two phylogenies H and P are based on a set of coevolutionary events. A cost measure is used for each type of events and a possible common history of H and P is evaluated using its cost, i.e., the sum of the cost of all its events. A typical question is then to find a cheapest common history of H and P that satisfies ϕ. The most often studied events are cospeciation, duplication, sorting, and switch (see Fig. 1 and also, e.g., Page, 2002).

A cospeciation event refers to a simultaneous host and parasite speciation and can therefore be associated with one (inner) node in the host and one (inner) node in the parasite tree. A duplication event is a speciation of the parasite that is independent from a host speciation. Hence, a duplication can be associated with one (inner) node p of the parasite tree and one edge (h, h') in the host tree. For a duplication event it is assumed that both child species of p live on hosts that are within the subtree with root h'. A switch consists of a speciation of the parasite where one of the emerging parasite species then changes its host. Switches can be associated with one (inner) node p in the parasite tree and one edge (h, h') in the host tree where the speciation happens. For a switch event it is assumed that one child species of p lives on a host that is within the subtree with root h'. The other child species changes to a host that is not within this subtree. A switch consists of a speciation of the parasite where one of the emerging parasite species then changes its host. At which point of time the actual switch of the host can happen depends on the evolutionary model. In Charleston (1998), and the tool TreeMap it is assumed that the actual switch happens always at the same time as the speciation, i.e., along the same edge in the host tree. Here, we consider also the case that the actual switch happens later as explained in the next subsection. A sorting event refers to the case that a host speciation happens independent from a parasite speciation and a parasite species that was on the host before the speciation is on only one of the new emerging host species after the speciation. We do not consider the case of a speciation of the parasite p where both child species change to hosts that are outside the subtree with

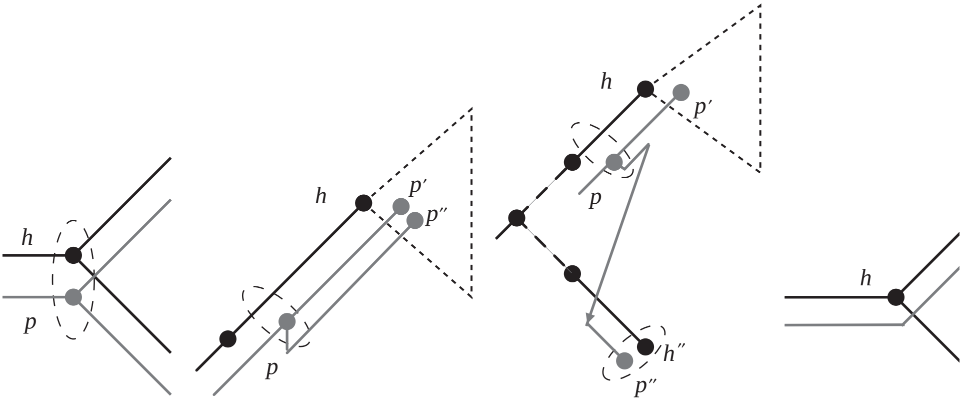


Fig. 1. Coevolutionary events (see section Coevolutionary events) and corresponding associations; from left-to-right: cospeciation $(p : h, 1)$; duplication $(p : h, 2)$ (both child nodes of p are associated with a node or edge in the subtree of H with root h); switch $(p : h, 2)$ (only one child node of p is associated with a node or edge in the subtree of H with root h); sorting; H black, P grey.

root h. The reason is that such events cannot be traced back (many other studies also do not allow such events (e.g., Charleston, 1998).

Reconstructions

In this subsection, we define what is considered to be a reconstruction of the cophylogenetic history of two phylogenetic trees. Some definitions are needed. For two phylogenetic trees H and P, a *reconstruction frame* assigns each node of P to a node or an edge of H. A reconstruction frame can be given in the form of a set of associations where each node of P is associated to a node in H and a type information of the association is given. As has been done in Charleston (1998), we call an association between a node h in H and p node in P to be of *type 1* when p is mapped onto h. It is of *type 2* when p is mapped onto the edge between h and its parent (it is assumed that the root node of H has a dummy parent). Type 2 associations are only possible for inner nodes of P. Type 1 associations are denoted $(p : h, 1)$ and type 2 associations are denoted $(p : h, 2)$.

Type 1 associations of inner nodes of P stand for cospeciation events. It is assumed for $(p : h, 1)$ that species p lives on h at the time of the speciation of h and that the speciation of h and p have happened at the same time. Type 2 associations stand for duplication events or switch events. For $(p : h, 2)$ it is assumed that the speciations of p has happened between the speciation of node h and the speciation of its parent node.

For given phylogenetic trees H and P and mapping ϕ from the leaves of P into the nodes of H a reconstruction frame is *valid* only when

(i) for each association $(p : h, 1)$ and the association $(p' : h', x)$, $x \in \{1, 2\}$ of the parent node p of p' it holds that h is a predecessor of h' and $h' \neq h$,

(ii) if in the association $(p : h, 1)$ the node p is a leaf of P then $h = \phi(p)$,
(iii) for each association $(p : h, 2)$ it holds that (a) for at least one child p' of p with association $(p' : h', x)$, $x \in \{1, 2\}$ the node h' must be a successor of h ($h' = h$ is possible) and (b) no child node of p can be associated to a proper predecessor of h.

In the following we assume that all considered reconstruction frames are valid. The evolutionary events during the coevolution of H and P that correspond to each association of an inner node of P in a reconstruction frame can easily be computed as follows. For the association $(p : h, 1)$ the corresponding event is a cospeciation. If for the association $(p : h, 2)$ the two child nodes of p are associated with successors of h then the corresponding event is a duplication. Otherwise, it is a switch.

For a reconstruction frame every pair of associations of an inner node p of P and of a child node p' of p implies a (possibly empty) set of sorting events that have happened between the corresponding events as described in the following. Let $(p : h, x)$, $x \in \{1, 2\}$ and $(p' : h', x')$, $x' \in \{1, 2\}$ be such a pair of associations. When $x = 1$ then h' is a proper successor of h and a sorting event happens at every node on the path from h to h' in H but not counting h and h' itself. Similarly, when $x = 2$ and h' is a proper successor of h then a sorting events happens at every node on the path from h to h' in H (including h) but not counting h' itself. When h' is not a successor of h then a host switch has happened. In this case let $(p'' : h'', y)$, $y \in \{1, 2\}$ be the association of the second child node p'' of p. Then h'' is a successor of h. In the evolutionary model considered in this paper the take-off of the switch can have happened on the edge between h and its parent node (as in the example of Fig. 1) or on an edge in the subtree of H with root h (as shown in the example of Fig. 2). It should be noted that our consideration of switches is an extension of the model of Charleston (1998) where it is assumed that the take-off site of a switch always happens on the edge from h to its parent node. The reason why we consider this extension is that we intend to integrate divergence timing information as explained in the next subsection. Since we consider significantly more possible (and biologicalreasonable) reconstructions we can often find time feasible reconstructions in cases where the model of Charleston (1998) says that no reconstruction exists. The landing site of the switch is assumed to be on the path between h' and the nearest common ancestor of h and h'. Then a sorting event happens at every node between h and the take-off site and between the landing site and h' (not counting h and h' itself).

We call a (valid) reconstruction frame together with an assignment of take-off and landing sites for all switches according to the model a *reconstruction*. We assume in the rest of this section that an assignment of integer costs to each of the following coevolutionary events is given: cost $\mathrm{co} \leqslant 0$ for a cospeciation, cost $\mathrm{so} \geqslant 0$ for a sorting, cost $\mathrm{du} \geqslant 0$ for a duplication, and cost $\mathrm{sw} \geqslant 0$ for a switch. The *costs* of a reconstruction is the sum of the costs of all events that correspond to inner nodes of P plus the costs of all sorting events that are implied by the reconstruction.

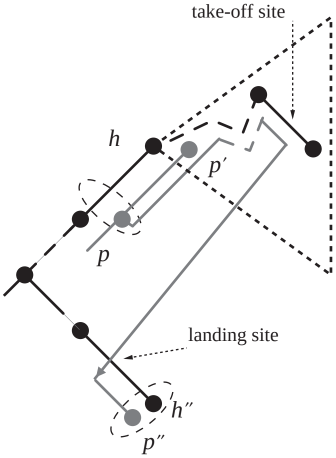


Fig. 2. Example of a switch $((p:h,2))$ where the take-off site is in the subtree of H with root h; H black, P grey.

A problem with switches in a reconstruction is that they induce a timing relation between the take-off site and the landing site. A consequence is that the occurrence of several switches in a valid reconstruction can lead to timing relations that are not possible (compare Charleston, 1998). Observe that for a cheapest (but not necessarily feasible) reconstruction it is necessary to place the landing site of a switch as late as possible to minimize the number of sorting events that are induced by the switch. When the timing relations that are implied by such switches make the reconstruction infeasible a possibility is to change the reconstruction by "moving back some landing sites" (Note that the corresponding reconstruction frame is not changed). This means for a switch from a node p to its child node p' that the landing site can be placed nearer to the nearest common ancestor of p and p'. A pair of switches in a reconstruction which induces timing relations that are not possible but where this incompatibility can be solved by moving back their landing sites is called *weakly incompatible* (Charleston, 1998). Note, that in the model of Charleston (1998) the take-off site of a switch cannot be changed and that in our model it is possible to move forward in time the take-off site of a switch. It is an interesting problem that was posed in Charleston (1998) to find for a given reconstruction frame where the cheapest reconstruction has weakly incompatible switches the cheapest reconstruction that is feasible. When only move back operations of landing sites are allowed we call this problem the Moving Back Landing Sites Problem. Here we state the following result:

Theorem. *The Moving Back Landing Sites Problem is NP-complete.*

It should be noted that the proof is not difficult but technical. Therefore, we give only a sketch. The proof can be done by a reduction from the Feedback Arcs Set Problem (FASP) on directed graphs (see Garey and Johnson, 1979). This problem is to find for a given directed graph a smallest set of arcs that contains at least one arc from each directed circle. The basic idea is to define for each circle in the given graph of the FASP a circle of switches so that each switch corresponds to one arc and so that the switches imply infeasible timing relations. The circles can be defined such that the move back of the landing site of any switch in such a circle by one node (i.e., introducing one additional sorting event) will destroy the infeasibility implied by this circle. Then for each set of arcs that solves the FASP the set of corresponding switches solves the Moving Back Landing Sites Problem when each switch in the set is moved back by one node and vice versa.

Divergence timing information

In this subsection, it is described how divergence timing information is integrated into our model. We assume that timing information about a divergence event is given in form of a time zone in which this event has happened. No general assumptions are made about how long the span of a time zone is. It can be small when the timing information is exact or it can be larger when only vague timing information exists. We assume that the time axis is partitioned into time zones (possibly with different time spans). Each time zone is denoted by an integer and these integers are chosen monotonically increasing from older time zones to younger ones. It should be noted that similar models of time are used for the construction of supertrees (see, e.g., Bininda-Edmonds, 2004; Semple and Steel, 2003).

Timing information about divergence events can be included into a phylogenetic tree by labelling each node by the time zone when the corresponding divergence has happened. Such a node labelling is *feasible* only when the label of each node is at least as large as the label of its parent node. We call such a feasible node labelling a *time zone labelling*. Since it might be difficult to decide for given time zones to which time zone a node belongs, we also consider the case that every node can be labelled by a pair of integers $[s,t]$, $s \leqslant t$ that denote a time zone interval. This means that the corresponding divergence event has happened in any of the time zones $s,\dots,t$. Such a labelling is *feasible* only when for each node with label $[s,t]$ and label $[s',t']$ of its parent node $s \geqslant s'$ and $t \geqslant t'$ holds. We call such a feasible labelling a *time interval labelling*. For a node p let $l(p)$ be its label.

In this paper, we consider the case that for the host tree H a time zone labelling is given and for the parasite tree P a time interval labelling is given. Then we call a reconstruction frame *time-valid* when for every association $(p:h,1)$ with $l(p)=[t_1,t_2]$ and $l(h)=s$ it holds that $t_1 \leqslant s \leqslant t_2$. A reconstruction is called *time-valid* when the underlying reconstruction frame is *time-valid* and when for each switch with take-off site on edge (h_1,h_1') and landing site on edge (h_2,h_2') the time zone intervals $[l(h_1),l(h_1')]$ and $[l(h_2),l(h_2')]$ have a nonempty intersection.

Computing cheapest reconstructions

The method how cheapest reconstructions of the common phylogeny of a host and parasite tree, where nodes are labelled with divergence timing information, can be computed is described in this subsection. Similar as done in TreeMap the computation of a cheapest reconstruction is based on a data structure that contains relations of associations between nodes of the parasite tree and the host tree. In TreeMap this data structure is called Jungle (see Charleston, 1998). A Jungle is a directed graph where the nodes correspond to possible associations of nodes in the hosts tree with nodes or edges in the parasite tree. The edges of the Jungle correspond to pairs of associations that can possibly be found in the same reconstruction. For example, it is required that for the two parasite nodes in such a pair of associations one is the parent of the other. Each edge is associated with the costs of the corresponding coevolutionary events. A reconstruction of the common phylogenetic history corresponds to a subgraph of the Jungle.

Instead of using pairs of associations the tool Tarzan works with a candidate set of triples of associations that can possibly be included in the same reconstruction. In every triple in the candidate set always one of the involved nodes from the parasite tree is the parent of the other two involved nodes. Each triple is also assigned the cost of its associated coevolutionary events, i.e. for a triple $((p,h,x),(p',h',x'),(p'',h'',x''))$, $x,x',x'' \in \{1,2\}$ where p is the parent of p' and p'' these are the costs for the event that corresponds to the association (p,h,x) (cospeciation, duplication, or switch) and the sorting events that occur between h and h' as well as sortings between h and h''. Only association triples that are possible in a valid reconstruction frame are included in the candidate set. When divergence timing information is given, only association triples with feasible associations are included in the candidate set. It should be noted that we do not use divergence timing information only to remove unfeasible triples from the candidate set that is computed for the case when no timing information is given. Instead, we consider also association triples that are not considered when no divergence timing information is given. We discuss the two interesting cases that have to be considered in the following. To this end consider an association triple $((p,h,x),(p',h',x'),(p'',h'',x''))$ with $x,x',x'' \in \{1,2\}$ where p is the parent of p' and p''.

(i) Switch: When a switch event happened at (p,h,x) (say the host is changed between p and p') and no divergence timing information is given then the take-off site of the switch is always assumed to be on the edge between h and its parent and the landing site is on the edge between h' and its parent. Hence, the minimal number of sorting events is assumed. When timing information is given then the minimal number of sorting events for a switch is determined differently as described in the following. All edges between h'' and the nearest common ancestor between h' and h'' are considered for the take-off site. For the landing site the edge between h and its parent node (as in the example of Fig. 1) or an edge in the subtree of H with root h (as shown in the example of Fig. 2) are considered to be possible. For the determination of the cost of the association triple a combination of edges for the take-off and landing site is chosen that is either possible according to the divergence timing information (i.e. both time zone intervals that correspond to the edges have a nonempty intersection) and that implies the smallest number of sortings. An example is given in Fig. 3.

(ii) Cospeciation and duplication: When no divergence timing information is given it can be assumed for a cheapest reconstruction that a cospeciation or duplication always happened so that h is the nearest common ancestor of h' and h'' (cmp.,

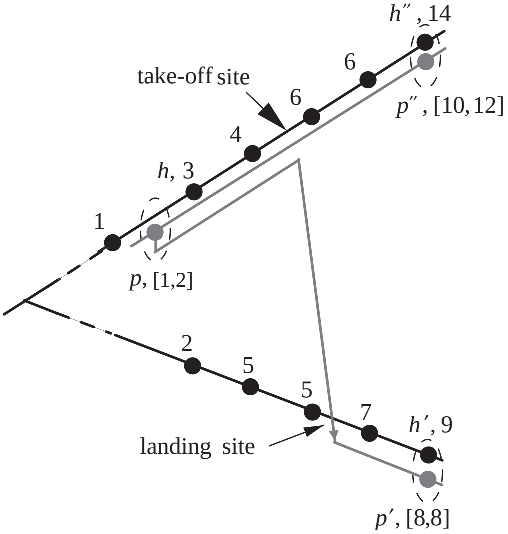


Fig. 3. Example of host switch when using divergence timing information: to compute the costs for the triple of associations $((p : h, 2)(p' : h', 1)(p'' : h'', 1))$ the take-off site and landing site are chosen so that both are in the same time zone — 5 or 6 and the smallest number of sorting events — 3 is implied; H black with time zone labelling, P grey with time interval labelling.

Charleston, 1998). With divergence timing information we consider all nodes h on the path from the nearest common ancestor of h' and h'' to the root of H as possible. In order to minimize the implied sorting events h is chosen so that it is the deepest node on this path for which the time interval $l(p)$ has a nonempty intersection with the time interval $[l(h), l(h^*)]$ where h^* is the parent of h. Examples are given in Fig. 4. Note that when the chosen h is not the nearest common ancestor between h' and h'', the event that corresponds to the association triple is always a duplication.

The algorithm that is used in Tarzan to compute the candidate set of association triples is described in the following. In addition, the algorithm computes for each node in the parasite tree a list of its associations that are included in a triple in the candidate set. The algorithm starts with an empty candidate set and for each leaf node of p with a list of associations that contains the association of this leaf of the parasite tree with the corresponding node in the host tree as given by the mapping ϕ.

Then iteratively Tarzan builds up the lists and the candidate set so that triples of associations for a node p and its child nodes p' and p'' are included after all triples where p' and p'' with their respective child nodes have been included and the lists of associations of p' and p'' in the corresponding triples have been computed. Then for every pair of associations $(p' : h', x)$, $x \in \{1, 2\}$ and $(p'' : h'', y)$, $y \in \{1, 2\}$ from this list, the associations of p that can form a triple with the pair are computed and the candidate set and the list are updated accordingly. It has been described above that not many associations of p have to be considered. Also it has been described how the

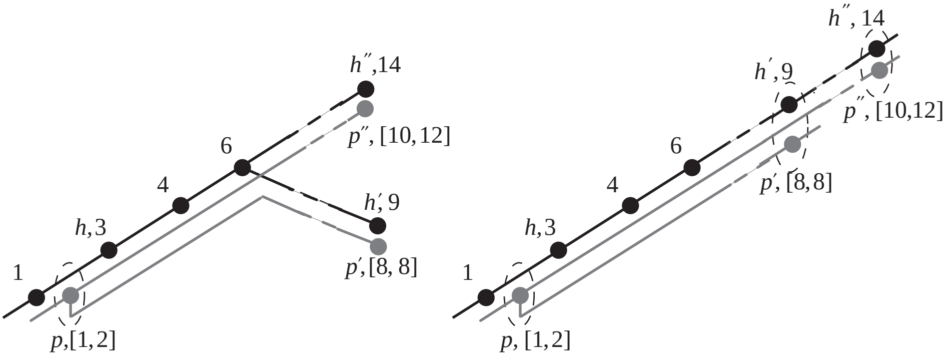


Fig. 4. Examples of duplications when using divergence timing information: for given associations $(p' : h', 1)$, $(p'' : h'', 1)$ the association of node p with a node h is chosen so that h is the deepest time feasible node on the path from the nearest common ancestor of h' and h'' to the root of H; H black, P grey.

costs for the coevolutionary events for each triple can be computed. Observe that a proper time labelling can be used to guaranty that only specific associations between the root of P and nodes or edges in H are possible, e.g., so that the root of P can only be mapped on the edge between the root of H and its (dummy) parent.

When the candidate set has been computed by Tarzan and a cost measure for the coevolutionary events has been defined, the cheapest reconstructions can be computed. The algorithm for this starts with all possible associations for the root of P. Let p be the root and p' and p'' its child nodes. Then for each association of p all triples in the candidate set with associations of p' and p'' are considered. For each such triple recursively the cheapest reconstructions for the subtrees are computed. It should be noted that hashmaps are used to store for every association of a node p the costs for the cheapest reconstruction of the corresponding subtree of p in P assuming p is mapped as in this association.

In order to show how fast Tarzan computes cheapest reconstructions we have created pairs of random phylogenetic trees of different sizes. A random tree of size n was recursively created by randomly starting at the root and assigning a random number of nodes to the right and left subtree so that a binary tree emerges. The mapping ϕ for two random phylogenetic trees H and P was determined by randomly assigning the leaves of P to leaves of H. Table 1 shows the size of the candidate set and the running time of Tarzan for creating the candidate set and computing cheapest reconstructions. Note, that the time for computing the candidate set includes the time for its visualization — which is approximately $\frac{2}{3}$ of this time for the large random trees.

Tarzan

The tool Tarzan is written in Java and consists of approximately 7500 lines of source code. The name is inspired by the fact that the tool searches relatively fast

Table 1. Number of triples in the candidate set and runtimes of Tarzan in seconds for random trees with n leaves

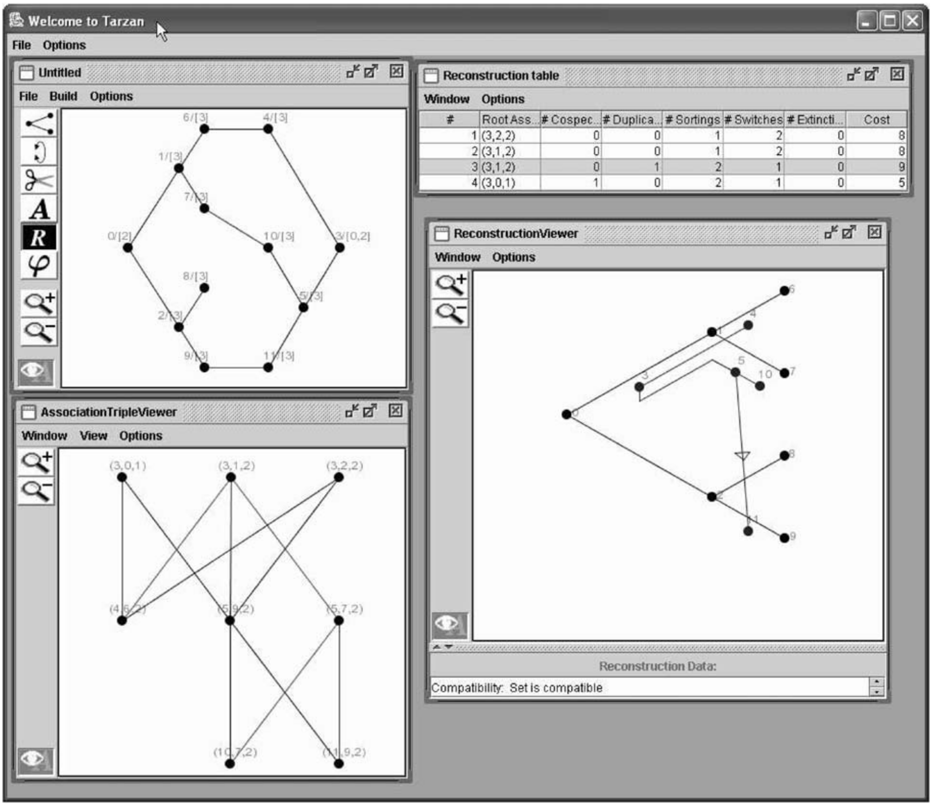
n	# Triples	Candidate set (s)	Minimum cost reconstructions (s)
25	6080	<1.0	<1.0
50	31 656	3.7	1.2
75	88 632	8.9	3.2
100	172 187	13.8	8.0
125	213 675	21.1	15.2
150	440 781	29.9	27.9
175	523 262	58.5	58.2
200	702 699	84.2	83.1

# triples: number of association triples in the candidate set; candidate set: time used by Tarzan for the construction and visualization of the candidate set; minimum cost reconstructions: time used by Tarzan for computing minimum cost reconstructions (at most 100) (when the candidate set already exists); runs where done on a PC with Intel 2.8 GHz CPU, main memory consumption in all tests runs was less than 140 MB.

through the set of possible reconstructions which were considered in Charleston (1998) as a set of trees forming a jungle. Tarzan has a graphical user interface that consists of four main windows. A screen shot of Tarzan is shown in Fig. 5. The phylogenetic trees can be defined and edited interactively in the tree editor window. The nodes of the trees can be labelled, e.g., with corresponding species names. The tree editor allows to easily include divergence times by defining a time zone labelling for one tree and a time interval labelling for the other tree. The mapping function ϕ that defines the current relations between the leaves of one tree and nodes of the other tree can simply be defined by drawing lines between the related nodes. Alternatively, the trees, their names, the divergence time information, and the mapping function can also be defined by modifying a corresponding text file.

When the phylogenetic trees and the mapping function have been defined, the candidate data structure containing the association triples can be calculated. It is presented in a corresponding association triple viewer. When the event costs have been set, all reconstructions or only the cheapest reconstructions can be calculated. The number of different types of events and the resulting costs are then listed in a reconstruction table window. By selecting a line in the reconstruction table window the corresponding reconstruction is depicted in a reconstruction viewer window. Moreover, in the association triple viewer all associations triples used for the reconstruction are marked.

Recall that it can happen that switches lead to timing incompatibilities within a reconstruction. Therefore, Tarzan automatically checks every reconstruction for switch incompatibilities and tries to resolve them by pulling back the landing site of switches so that only a minimal number of sortings have to be introduced. But because the corresponding problem is NP-complete and to have a fast tool it is not



#	Root Ass...	# Cospec...	# Duplica...	# Sortings	# Switches	# Extincti...	Cost
1	(3,2,2)	0	0	1	2	0	8
2	(3,1,2)	0	0	1	2	0	8
3	(3,1,2)	0	1	2	1	0	9
4	(3,0,1)	1	0	2	1	0	5

Compatibility: Set is compatible

Fig. 5. Tarzan tool; shown are the tree editor window (top left), reconstruction table window (top right), the association triple viewer (bottom left), and the reconstruction viewer.

guaranteed that Tarzan can resolve all incompatibilities then. Incompatibilities between switches that have been resolved and the corresponding possible move back operations are listed by Tarzan. Tarzan uses the following method to detect incompatibilities. For every switch it adds two directed edges to the directed host tree (edges are directed from the parent to the child node) to include the timing constraints implied by the switch as follows. For every switch which has its take-off site on an edge (h_1, h'_1) and its landing site on an edge (h_2, h'_1) the edges (h_1, h'_2) and (h_2, h'_1) are added (see Fig. 6). Tarzan computes then whether the directed graph that is obtained by adding these edges contains a directed circle. When there exists directed circles, timing incompatibilities are found. Tarzan states for every reconstruction that is listed in the reconstruction table window whether it contains timing incompatibilities, which switches have introduced the timing incompatibilities, and wether the timing incompatibilities could have been resolved.

Some additional features of Tarzan should be mentioned. (1) Tarzan offers not only the possibility to compute any cheapest reconstruction but can also compute a

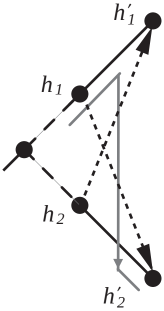


Fig. 6. Additional arcs (dotted) (h_1, h'_2) and (h_2, h'_1) that are introduced to the host tree to model the timing constraints introduced by the switch with take off site on edge (h_1, h'_1) and landing site (h_2, h'_2); *H* black, *P* grey.

cheapest reconstruction which has a minimum (or a maximum) number of cospeciations. (2) The maximal number of cheapest reconstructions that are be computed by Tarzan can be set by the user. (3) Tarzan can also list all possible reconstructions which could be interesting for cases where not too many reconstructions exist.

The Tarzan tool is available from the website of our group at http://pacosy.informatik.uni-leipzig.de/idxeng.html.

Applications

In this section, we show how the handling of phylogenetic trees with divergence timing information in Tarzan can be applied to cophylogenetic analysis. We discuss two examples of cophylogenetic systems where the data about phylogeny and divergence times are taken from the literature. One example is about host–parasite relations and the other example concerns insect–plant relations. It should be noted that all sets of cheapest reconstructions that have been computed by Tarzan and are mentioned in this section contain reconstructions without switch incompatibilities.

Host–parasite relations

The cophylogeny of African brood parasite finches (*Vidua* sp.) and their finch hosts (family Estrildae) has been studied in Sorenson et al. (2004). The parasite finches are host specialists that mimic the songs and the nestling mouth markings of their hosts. The phylogeny of the *Vidua* species was compared in Sorenson et al. (2004) with the phylogeny of the estrildid finch hosts and compared with divergence time estimates for both groups. Since hosts and parasites are sister groups they could have been combined in a single phylogenetic analysis. The divergence times were estimated from the rates of sequence evolution. The tests done in Sorenson et al.

(2004) have shown the existence of different rates of sequence evolution in the different clades. The Langley–Fitch method as implemented in the program r8s (Sanderson, 2003) was used to obtain ML estimates of the divergence times under a local clock model with different rates of sequence evolution in the clades.

An analysis of the *Vidua* and Estrildae phylogenies and their current host–parasite relations was done in Sorenson et al. (2004) with TreeMap 1.0b. The results indicate that 15 cospeciation events have occurred. It was shown that this is significantly more than expected by chance. It has been noted in the paper that care must be taken with such results that indicate a significant congruence between the host and parasite topologies. It is possible that another mechanism than cospeciation could possibly lead to similar results (cmp., Holmes, 2003). If phylogenetically related host species tend to live sympatrically, then parasites will tend to jump between closely related host species. Also, the ability to jump over species boundaries may be dependent on the phylogenetic distance between hosts, so that it is easier to jump to a closely related host species than to a more distantly related one. If in either case such events occur at sufficient frequency, then the host and parasite trees may often match, giving a false impression of cospeciation. Information about divergence times can in such cases be very useful to decide whether cospeciation events could have been the reason for the similarity between the host and parasite trees or not.

Additional analysis has been done in Sorenson et al. (2004) with TreeMap 2.0.2β (setting the costs of a cospeciation event to zero and the costs for a duplication, sorting, or switch event to one). The tree reconciliation with no host switching suggests that 8 cospeciations have occurred and tests have shown that this in not significantly higher than the expected number in a random scenario. When considering only the primary host for each parasite, 9 cospeciation events were reconstructed and it was shown that this is significantly more than in a random scenario. However, this scenario required 55 sorting events and 9 duplications and associated up to seven parasites onto the same host. It was argued in Sorenson et al. (2004) that biological reasons make duplications and multiple parasites per hosts very unlikely. Allowing an increasing number of host switches the reconstructions in TreeMap 2.0.2β had increasingly lower costs. The authors state that computational limitations prevented them from exploring reconstructions with more than 3 switches and suggest that such solutions might have even less costs. By considering information about divergence times they have shown that most cospeciation events in the reconstructions are not possible.

Since the authors of Sorenson et al. (2004) could neither explore reconstructions with more than 3 switches nor include their information about divergence times into the analysis with TreeMap 2.0.2β we have analyzed their data with Tarzan. In order to translate the timing information into labels we defined 5 time zones (0–1.5 Myr, >1.5–3.0 Myr, >3.0–4.5 Myr, >4.5–6.0 Myr, >6.0 Myr). All nodes of the estrildid host tree and the *Vidua* parasite tree as given in Fig. 4 in Sorenson et al. (2004) have been assigned labels. See Fig. 7 for the trees that were used by Tarzan. Three nodes of the parasite tree that are near the border between two time zones have been given interval labels (nodes indicated with 1, 6 (7) in Fig. 4 of Sorenson et al. (2004) have been assigned to time zone intervals 4–5 (respectively, 2–3). All nodes of the *Vidua* parasite tree are in the first four times zones. Only the primary hosts have been considered. Tarzan requires that all nodes have outdegree at most 2. Since two nodes in the parasite tree have outdegree more than 2 — this are the nodes corresponding to the sets of species {*V. raricola*, *V. maryae*, *V. larvaticola*, *V. wilsoni*} and to {*V. funerea*, *V. purpurascens*, *V. codringtoni*, *V. chalybeata.S*} — the parasite tree has been changed slightly. Each of the two nodes with outdegree 4 has been replaced by a subtree with 3 nodes of outdegree 2. Each subtree was chosen so that it implies the same relative phylogenetic distances as the corresponding hosts have in the host tree.

The cost assignments that have been used to compute cheapest reconstructions with Tarzan have cost 2 for a duplication and 1 for a sorting. The cost assignments differ in the costs for cospeciations and switches. Cost assignments Γ_1, Γ_2, and Γ_3 have cospeciation cost 0 and switch costs 2, 10, 100, respectively. Cost assignment Γ_4 has cospeciation cost -100 and switch cost 2. Note, that cost assignment Γ_1 can be considered in general as the most realistic one because it is commonly used in the literature. Cost assignments Γ_3 and Γ_4 are more unrealistic and have been used to find solutions with a minimum number of switches, respectively, maximum number of cospeciations. Note that the variables for defining a cost assignment in TreeMap (c for cospeciation, d for duplication, s for sorting, and h for switch) do not equal the cost of the corresponding event. If co, du, so, sw are the costs for cospeciation, respectively duplication, sorting, and switch in our model the following relations holds: $\mathrm{co} = 0.5 \cdot c$, $\mathrm{du} = 0.5 \cdot d$, $\mathrm{so} = s$, $\mathrm{sw} = d + h$. Thus cost assignment Γ_1 corresponds to the TreeMap variables $c = 0$, $d = 1$, $s = 1$, $h = 1$ (this cost assignment has been used in Sorenson et al. (2004)).

Table 2 shows the costs and the number of events in the cheapest reconstructions that have been computed with Tarzan either with or without timing information as given by the node labels.

The results show that without divergence timing information the number of switches in a cheapest reconstruction for cost assignment Γ_1 is 12–14. Thus, the conjecture from Sorenson et al. (2004) that the cheapest reconstruction might have more than 3 switches is correct. The high number of 12–14 switches in a cheapest reconstruction suggests that even without additional divergence timing information it can be conjectured that host switching plays a significant role for the Estrildae–*Vidua* host–parasite system. Without switching (cost assignment Γ_3) Tarzan computed 11 cospeciations, 9 duplications, and 69 sortings. This is only slightly different from the results in Sorenson et al. (2004) which have computed 9 cospeciations, 9 duplications, and 55 sortings when no switches were allowed. The reason for this difference is that in this paper a slightly different parasite tree has been used.

Fig. 7. Estrildae host tree (upper tree), *Vidua* parasite tree (lower tree) and mapping ϕ as used for finding reconstructions with Tarzan (after Sorenson et al., 2004); grey areas indicate the time zone labels, respectively time interval labels that have been assigned to the corresponding nodes.

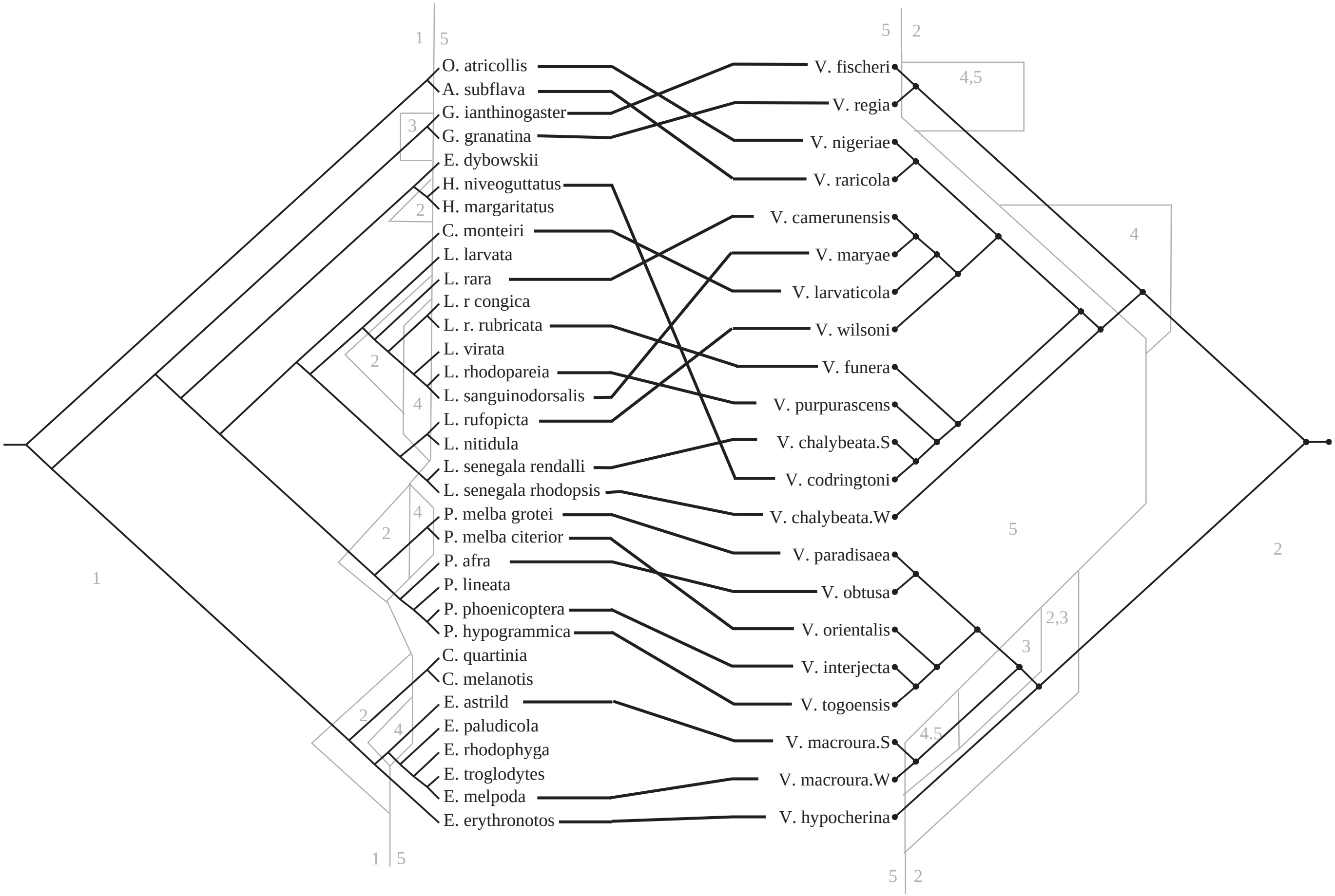

1
5
O. atricollis
A. subflava
G. ianthinogaster
G. granatina
3
E. dybowskii
H. niveoguttatus
H. margaritatus
2
C. monteiri
L. larvata
L. rara
L. r congica
L. r. rubricata
2
L. virata
L. rhodopareia
L. sanguinodorsalis
4
L. rufopicta
L. nitidula
L. senegala rendalli
L. senegala rhodopsis
4
2
P. melba grotei
P. melba citerior
P. afra
P. lineata
P. phoenicoptera
P. hypogrammica
C. quartinia
C. melanotis
2
4
E. astrild
E. paludicola
E. rhodophyga
E. troglodytes
E. melpoda
E. erythronotos
1
1
5
5
2
4,5
V. fischeri
V. regia
V. nigeriae
V. raricola
4
V. camerunensis
V. maryae
V. larvaticola
V. wilsoni
V. funera
V. purpurascens
V. chalybeata.S
V. codringtoni
V. chalybeata.W
5
2
V. paradisaea
V. obtusa
V. orientalis
2,3
3
V. interjecta
V. togoensis
4,5
V. macroura.S
V. macroura.W
V. hypocherina
5
2

Table 2. Costs and number of events for minimal cost reconstructions for host (*Vidua*) and parasite (Estrildae) phylogenies computed with Tarzan and cost assignments Γ_1–Γ_4 without timing information and with using timing information for trees as given in Fig. 7 (phylogenies and timing information from Sorenson et al., 2004)

	Without timing information					With timing information				
	Cost	#co	#du	#so	#sw	Cost	#co	#du	#so	#sw
Γ_1	29	6	0	1	14	34	4	0	2	16
	29	7	0	3	13	34	5	0	4	15
	29	8	0	5	12					
Γ_2	77	12	5	37	3	154	5	0	4	15
Γ_3	87	11	9	69	0	1504	5	0	4	15
Γ_4	−1261	13	1	25	6	−466	5	0	4	15

#co = number of cospeciations, #du = number of duplications, #so = number of sortings, #sw = number of switches.

Using the timing information the results show that the cheapest reconstructions for all cost assignments with divergence timing information have a large number of host switches. The minimal number of switches as computed with cost assignment Γ_3 is 15. The maximum number of cospeciations is only 5 compared to a maximal number of 13 when timing information is not used. Without using timing information, even with the realistic cost assignment Γ_1, the 6–8 cospeciations are more than the maximal number of 5 cospeciations obtained, when using the divergence timing information. The number of cospeciations when using timing information is only 4–5 for cost assignment Γ_1. This shows why the observation of Sorenson et al. (2004) is true, that most of the associations that have been considered by an inspection of the phylogenetic trees to correspond to possible cospeciations are not possible. It has to be noted that 3 of the 4 cospeciations in the cheapest solutions computed by Tarzan for cost assignment Γ_1 are different from the cospeciations that have been discussed in Sorenson et al. (2004). The latter cospeciations have been selected after an inspection of the topologies and then considering the timing information. Two of the 3 different cospeciations found by Tarzan include switches in their subtrees and cannot be found by a simple visual inspection of the tree topologies. This shows clearly the advantage of Tarzan to include divergence timing information into the computation of cheapest reconstructions.

Insect–plant relations

The common phylogenetic history between phytophagous insects (Psylloidae) and their host plants (Leguminosae) on the Canary Islands has been studied in Percy (2001). Psyllids (Hemiptera) complete their whole life cycle on a single host plant. The phylogenies of the psyllids and legumes as well as their current relations have been analyzed with TreeMap 1.0. Geological data have been used to calibrate the phylogenies of both groups by estimating divergence times.

The results obtained with TreeMap in Percy (2001) indicate that 15 cospeciation events have occurred (considering only the primary host–parasite relations). In a later study (Percy et al., 2004) with slightly changed phylogenetic trees the authors obtained with TreeMap (when minimizing the number of switches) 16 cospeciations, 29 duplications and 220 sorting events. Moreover, it was shown that this number is significantly more than to be expected by chance. However, the information on divergence times show that the majority of the psyllids nodes are clearly younger than the legume nodes that where associated to them in the corresponding reconstruction. As concluded in Percy (2001) this indicates that the general psyllid–host pattern is not the result of cospeciations, as suggested by the reconstruction of TreeMap.

In order to investigate the psyllids and legume trees as given in Fig. 1 in Percy (2001) together with the timing information, we have defined 6 time zones (0–2 Myr, >2–4 Myr, >4– 6 Myr, >6–8 Myr, >8–10 Myr). All nodes of the legume tree and the psyllid tree have been labelled by their time zone. See Fig. 8 for the trees that were used by Tarzan. Most inner nodes of the psyllid tree have been assigned an interval of two time zones. Because some psyllids are associated with more than one host in Fig. 1 of Percy (2001) we have changed the trees used for the computations with Tarzan slightly. Psyllid species and *A.* sp *10* and *L. retamae* have been assigned to the first common ancestor of their three host species. Because legume species *T. stenSSPpauc* and *T. stenSSPmicro* as well as species *T. canariensis* and *T. osyroides* have been assigned to the same leaf node in the legume tree. Psyllid species *Ar. adenocarpi* has been assigned only to legume species *A. boudyi.*

Program Tarzan has been used to compute the cheapest reconstructions with and without using the labelling information. In all cost assignments that have been used the cost for sortings and duplications were set to 1, respectively, 2. The cost assignments differ by their cospeciation and switch costs. Cost assignments Γ_1, Γ_2, and Γ_3 have cospeciation costs -2 and switch costs 2, 10, and 100, respectively. Cost assignment Γ_4 has cospeciation costs -100 and switch costs 2. It should be noted that cost assignment Γ_1 corresponds to cost assignment $c = -1$, $d = 1$, $s = 1$, $h = 1$ of TreeMap. Note, that cost assignment Γ_1 is commonly used in the literature for cophylogenetic studies. Cost assignment Γ_3 and Γ_4 have been used to find reconstructions with a minimal number of switches, respectively maximum number of cospeciations.

Table 3 shows the costs and the number of events in the cheapest reconstructions that have been computed with Tarzan either with or without timing information as given by the node labels.

The results show that the 10 cospeciation events in the cheapest reconstructions computed with Tarzan using cost assignment (i) and using the timing information are significantly less compared to the 15 cospeciations in solutions that have been computed with TreeMap in Percy (2001). Similar to the latter case, Tarzan computed when not using the timing information between 12 and 17 cospeciations for cost assignments Γ_1–Γ_3. Even the maximal number of 10 cospeciations when using timing

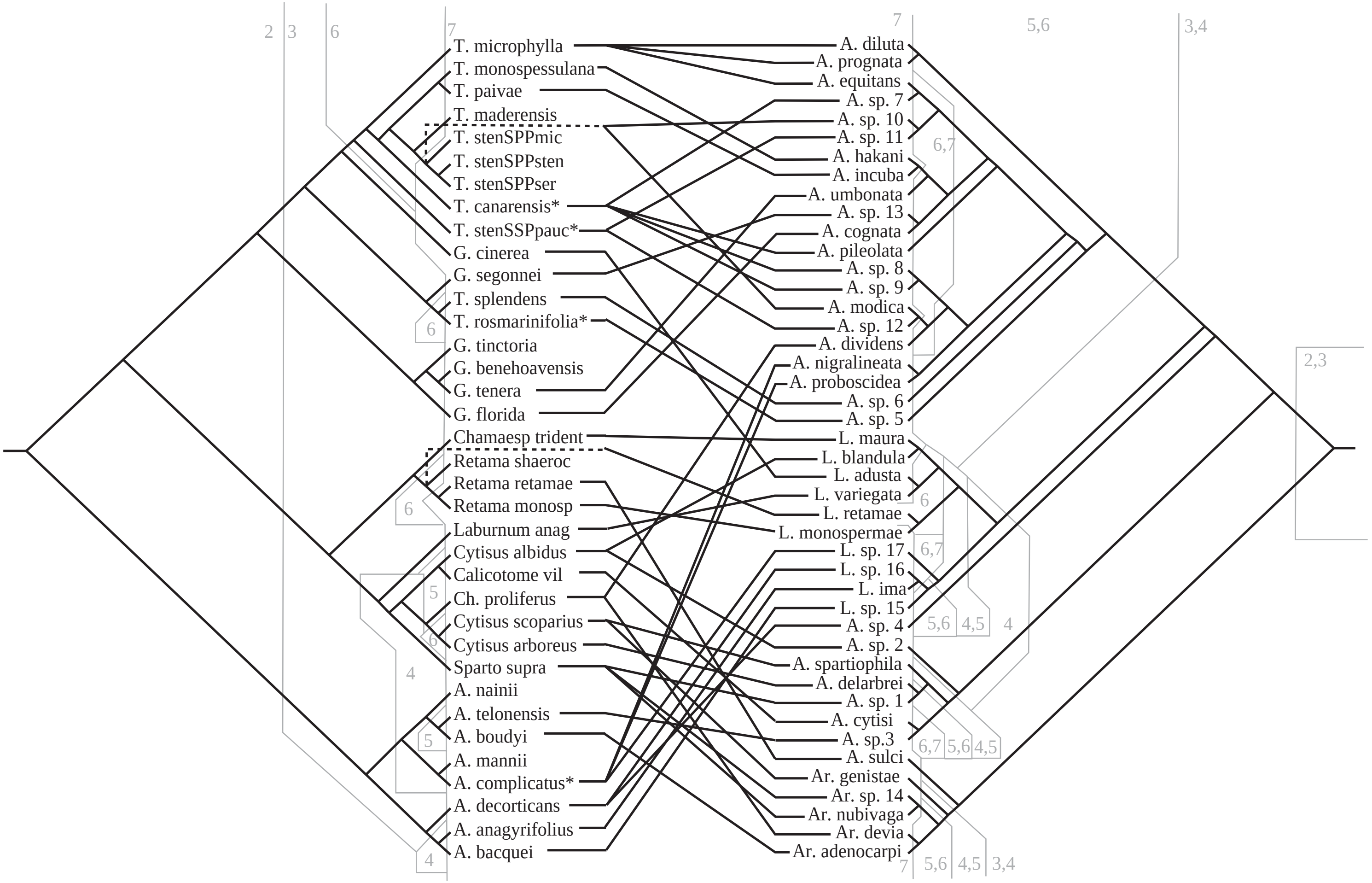

2
3
6
7
T. microphylla
T. monospessulana
T. paivae
T. maderensis
T. stenSPPmic
T. stenSPPsten
T. stenSPPser
T. canarensis*
T. stenSSPpauc*
G. cinerea
G. segonnei
T. splendens
T. rosmarinifolia*
6
G. tinctoria
G. benehoavensis
G. tenera
G. florida
Chamaesp trident
Retama shaeroc
Retama retamae
Retama monosp
6
Laburnum anag
Cytisus albidus
Calicotome vil
5
Ch. proliferus
Cytisus scoparius
6
Cytisus arboreus
Sparto supra
4
A. nainii
A. telonensis
A. boudyi
5
A. mannii
A. complicatus*
A. decorticans
A. anagyrifolius
A. bacquei
4
7
5,6
3,4
A. diluta
A. prognata
A. equitans
A. sp. 7
A. sp. 10
A. sp. 11
6,7
A. hakani
A. incuba
A. umbonata
A. sp. 13
A. cognata
A. pileolata
A. sp. 8
A. sp. 9
A. modica
A. sp. 12
A. dividens
A. nigralineata
A. proboscidea
2,3
A. sp. 6
A. sp. 5
L. maura
L. blandula
L. adusta
L. variegata
6
L. retamae
L. monospermae
L. sp. 17
6,7
L. sp. 16
L. ima
L. sp. 15
A. sp. 4
5,6
4,5
4
A. sp. 2
A. spartiophila
A. delarbrei
A. sp. 1
A. cytisi
A. sp.3
6,7
5,6
4,5
A. sulci
Ar. genistae
Ar. sp. 14
Ar. nubivaga
Ar. devia
Ar. adenocarpi
7
5,6
4,5
3,4

Table 3. Costs and number of events for minimal cost reconstructions for psyllid and legume phylogenies computed with Tarzan and cost assignments Γ_1–Γ_4 without timing information and with using timing information for the trees as given in Fig. 8 (phylogenies and timing information are taken from Percy (2001))

	Without timing information					With timing information				
	Cost	#co	#du	#so	#sw	Cost	#co	#du	#so	#sw
Γ_1	46	12	5	8	26	66	10	5	20	28
	46	13	5	12	25					
	46	14	5	16	24					
	46	15	5	20	23					
	46	16	5	24	22					
Γ_2	158	17	23	116	3	268	9	11	34	23
Γ_3	174	16	27	152	0	2338	9	11	34	23
Γ_4	−1805	19	5	47	19	−914	10	5	20	28

#co = number of cospeciations, #du = number of duplications, #so = number of sortings, #sw = number of switches.

information as obtained with cost assignment Γ_4 is less than the number of 12–16 cospeciations obtained with the realistic cost assignment Γ_1 when timing information is not used. The cheapest solutions for cost assignments Γ_2 and Γ_3 without using timing information have only 3 switches, respectively, no switch. For this case the number of the other events are similar to the reconstructions that have been obtained in Percy et al. (2004) for similar trees with TreeMap minimizing switches. These reconstructions have 16 cospeciations, 29 duplications, 220 sortings and no switch event. But as has been noted in Percy et al. (2004) the lack of switches would be unusual for plant–herbivore systems (see, e.g., Becerra and Venable (1999) where the cophylogeny between the beetle genus *Blepharida* and its host plant genus *Bursera* has been investigated). Moreover, the cospeciations are in nearly all cases not possible with respect to their timing. The reconstructions with Tarzan for the corresponding cost measures when using timing information have 23 switches and the cospeciations respect the timing information. This shows that the inclusion of timing information in Tarzan is important for the analysis of cophylogenetic history. Reasonable reconstructions that do not clearly underestimate the relative importance of switches and include cospeciations that are possible with respect to timing information could not have been obtained with the existing methods.

Fig. 8. Leguminosae plant tree (upper tree), psyllids insect tree (lower tree), and mapping ϕ as used for finding reconstructions with Tarzan (after Percy (2001)); grey areas indicate the time zone labels, respectively, time interval labels that have been assigned to the corresponding nodes, ⋆ indicates nodes where several species have been combined.

Conclusions

In this paper, we have presented a method for the reconstruction of the cophylogenetic history of two phylogenetic trees (e.g., a host and a parasite tree) where timing information about divergence events is considered. As a model for the timing information we used a partition of the time axis into time zones. Each node in the host tree can be labelled with a time zone. In order to allow non-exact timing information each node in the parasite tree can be labelled with an interval of time zones. Our reconstruction method is event based and considers cospeciation, duplication, sorting, and host switch events. Each evolutionary event is assigned a cost and cost minimal solutions are sought so that the timing constraints for all evolutionary events are satisfied. Out method is included in the newly presented tool called Tarzan. Tarzan can detect timing incompatibilities that might be introduced by host switches. But it is not guarantied that these incompatibilities can be resolved by the tool. We show that it is an NP-complete problem to resolve such incompatibilities with minimal additional costs. Different from an existing tool called TreeMap that uses pairs of associations between parasite nodes and nodes or edges in the host, our algorithm for computing cheapest reconstructions uses triples of associations. The use of triples allows a fast computation of cheapest reconstructions. Test results with Tarzan on random trees show that for pairs of trees with 200 nodes each the computation of a cheapest reconstruction including its visual presentation takes less than 3 min on a standard PC.

Two cophylogenetic systems (a host–parasite system and an insect–plant system) that were taken from the literature have been analysed with Tarzan. It was shown that some conjectures from the corresponding studies about the relative importance of cospeciations and switches could be backed with our results. On the other hand, Tarzan found time feasible cospeciation events that were not considered before in these studies. This shows that a cophylogenetic analysis with the current tools and a subsequent consideration of divergence timing information is not enough to fully analyse such systems.

Future work includes the integration of other types of constraints and also more complex timing constraints for the reconstructions into Tarzan. In addition, improved methods for the handling of timing incompatibilities that are caused by switches should be integrated.

Acknowledgement

We thank Roman Legat for his work on a preliminary version of Tarzan (see Legat, 2001).

References

Becerra, J.X., Venable, E., 1999. Macroevolution of insect–plant associations: the relevance of host biogeography to host affiliation. Proc. Nat. Acad. Sci. USA 96, 12626–12631.

Bininda-Edmonds, O.R.P., (Ed.), 2004. Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life. Kluwer Academic Publishers, Dordrecht.

Charleston, M.A., 1998. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. Math. Biosci. 149, 191–223.

Charleston, M.A., Page, R.D.M., 2002. TreeMap 2.0β A Macintosh program for the analysis of how dependent phylogenies are related, by cophylogeny mapping. Webpage: http://taxonomy.zoology.gla.ac.uk/ Latest version is 2.0.2β.

Garey, M.R., Johnson, D.S., 1979. Computers and Intractability—A Guide to the Theory of NP-Completeness. Freeman, New York.

Holmes, E.C., 2003. Molecular clocks and the puzzle of RNA virus origins. J. Virol. 77 (7), 3893–3897.

Jeong, S.-C., Ritchie, N.J., Myrold, D.D., 1999. Molecular phylogenies of plants and Frankia support multiple origins of actinorhizal symbioses. Mol. Phylogenet. Evol. 13, 493–503.

Legat, R., 2001. Datenstrukturen zur Analyse der Phylogenie von Parasit-Wirt-Beziehungen (data structures for the analysis of the phylogeny of parasite–host relations). Diploma Thesis, Institute of Computer Science, University of Hannover, Germany.

Page, R.D.M. (Ed.), 2002. Tangled Trees: Phylogeny, Cospeciation and Coevolution. University of Chicago Press, Chicago.

Percy, D.M., 2001. Diversification of legume-feeding psyllids (Psylloidea, Hemiptera) and their host plants (Genisteae, Leguminosae). Ph.D. Thesis, Division of Environmental and Evolutionary Biology, University of Glasgow, UK.

Percy, D.M., Page, R.D.M., Cronk, Q.C.B., 2004. Plant–insect interactions: double-dating associated insect and plant lineages reveals asynchronous radiations. Syst. Biol. 53 (1), 120–127.

Ronquist, F., 2002. Parsimony analysis of coevolving species associations. In: Page, R.D.M., (Ed.), Tangled Trees: Phylogeny, Cospeciation and Coevolution. University of Chicago Press, Chicago, pp. 22–64.

Ronquist, F., 2001. TreeFitter Version 1.0 Manual. Webpage: http://www.ebc.uu.se/systzoo/research/treefitter/treefitter.html.

Sanderson, M.J., 2003. r8s: inferring absolute rates of molecular evolution, divergence times in the absence of a molecular clock. Bioinformatics 19 (2), 301–302.

Semple, C., Steel, M., 2003. Phylogenetics. Mathematics and its Applications series, vol. 22. Oxford University Press, Oxford.

Sorenson, M.D., Balakrishnan, C.N., Payne, R.B., 2004. Clade-limited colonization in brood parasitic finches (*Vidua* spp.). Syst. Biol. 53, 140–153.