How Tarzan Works - Some Information on the Recontsruction and Cost Model

Daniel Merkle and Martin Middendorf

Department of Computer Science

University of Leipzig

Augustusplatz 10-11

D-04109 Leipzig, Germany

{merkle,middendorf}@informatik.uni-leipzig.de

Fax: +49-341-9732329

Tel:+49-341-9732275

July 1, 2005

1 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.

1.1 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 Figure 1 and also, e.g., [**?**]).

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 [1] 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

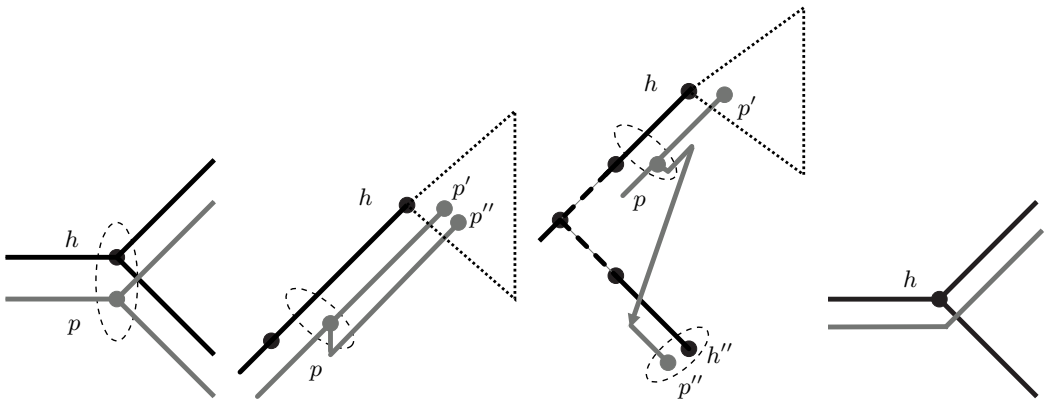

Figure 1: Coevolutionary events (see Subsection 1.1) and corresponding associations (see Subsection 1.2); from left to right: cospeciation $(p : h, 1)$; duplication $(p : h, 2)$ (both child nodes of p are associated with a node or edge in the subtree of H with root h); switch $(p : h, 2)$ (only one child node of p is associated with a node or edge in the subtree of H with root h); sorting; H black, P grey

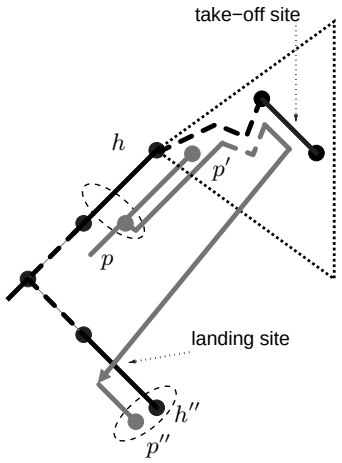

Figure 2: Example of a switch $((p : h, 2))$ where the take-off site is in the subtree of H with root h; H black, P grey

case that the actual switch happens later as explained in the next Subsection 1.2. 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 root h. The reason is that such events can not be traced back (many other studies also do not allow such events (e.g., [1]).

1.2 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 [1], 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 Figure 1) or on an edge in the subtree of H with root h (as shown in the example of Figure 2). It should be noted that our consideration of switches is an extension of the model of [1] 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 1.3. Since we consider significantly more possible (and biological reasonable) reconstructions we can often find time feasible reconstructions in cases where the model of [1] 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 $co \leq 0$ for a cospeciation, cost $so \geq 0$ for a sorting, cost $du \geq 0$ for a duplication, and cost $sw \geq 0$ 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.

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 [1]). 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 [1]. Note, that in the model of [1] the take-off site of a switch can not 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 [1] 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.

It is shown in [3] that the Moving Back Landing Sites Problem is NP-complete.

1.3 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. [**?**, **?**]).

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 $[s,t]$, $s \leq t$ that denote a time zone interval. This means that the corresponding divergence event has happened in any of the time zones $s,\ldots,t$. Such a labelling is *feasible* only when for each node with label $[s,t]$ and label $[s',t']$ of its parent node $s \geq s'$ and $t \geq t'$ 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 $t_1 \leq s \leq t_2$. 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.

1.4 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 [1, 2]). 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

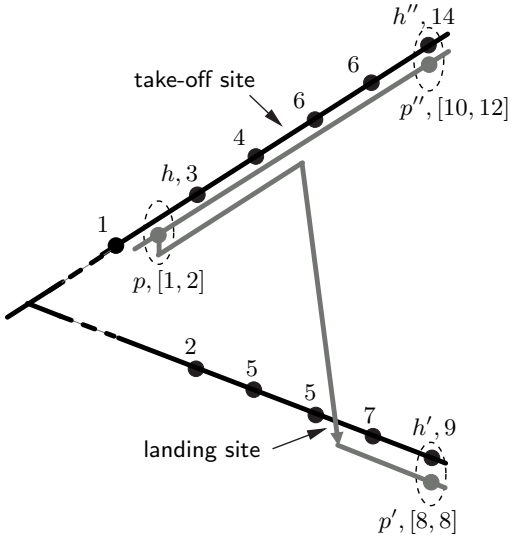

Figure 3: Example of host switch when using divergence timing information: to compute the costs for the triple of associations $((p : h, 2)(p' : h', 1)(p'' : h'', 1))$ the take-off site and landing site are chosen so that both are in the same time zone — 5 or 6 in the figure — and the smallest number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

the example of Figure 1) or an edge in the subtree of H with root h (as shown in the example of Figure 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 Figure 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. [1]). 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 Figure 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

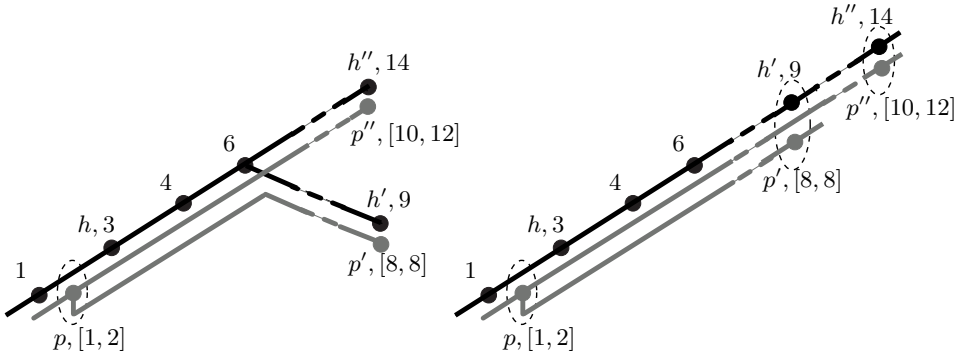

Figure 4: Examples of duplications when using divergence timing information: for given associations $(p' : h', 1)$, $(p'' : h'', 1)$ the association of node p with a node h is chosen so that h is the deepest time feasible node on the path from the nearest common ancestor of h' and h'' to the root of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

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 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.

References

[1] M.A. Charleston: Jungles: a new solution to the host/parasite phylogeny reconciliation problem. Mathematical Biosciences, 149:191–223 (1998).

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

[3] D. Merkle, M. Middendorf: Reconstruction of the Cophylogenetic History of Related Phylogenetic Trees with Divergence Timing Information. Theory of Biosciences, 4: 277-299, (2005).