A Parameter-Adaptive Dynamic Programming Approach for Inferring Cophylogenies*

Daniel Merkle[1], Martin Middendorf[2] and Nicolas Wieseke[2]†

[1]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
daniel@imada.sdu.dk
[2]Parallel Computing and Complex Systems Group, Department of Computer Science, University of Leipzig, Germany
{middendorf, wieseke}@informatik.uni-leipzig.de

Abstract

This paper introduces a new algorithm and a corresponding tool called CoRe-PA, that can be used to infer the common history of coevolutionary systems, e.g., hosts and their parasites or insect-plant relations. The proposed method utilizes an event-based concept for reconciliation analyses where the possible events are cospeciations, sortings, duplications, and (host) switches. All known event-based approaches so far assign costs to each type of cophylogenetic events in order to find a cost-minimal reconstruction. CoRe-PA uses a new parameter-adaptive approach, i.e., no costs have to be assigned to the coevolutionary events in advance. This is interesting, because from a biological point of view, reasonable cost values can often be estimated only very roughly. Experimental results are presented for several cophylogenetic test systems and it is shown that CoRe-PA produces high quality results for the test systems.

Introduction

Due to the immense increase of available molecular data and the methodological improvements in computer science to handle this data, methods for analyzing the coevolution of large data sets of two groups of species become more and more sophisticated. Examples of such coevolutionary systems are hosts and their parasites, insect-plant relations, or symbiotic relationships. Different methods for reconstructing the common host parasite relations have been proposed in the literature (for an overview see, e.g., [4, 11]). One common approach is to use an evolutionary model that describes the set of possible types of events that happened during coevolution, and to assign costs for the different types of events. The problem is then to find a reconstruction of the common history with a minimal sum of event costs.

*Working Paper PACOSY 2009-01, August 2009, Submitted

†This work was supported by the German Research Foundation (DFG) through the project "Deep Metazoan Phylogeny" within SPP 1174.

Algorithms that employ this idea are called event-based methods [16]. Typically the four different types of events that are considered are cospeciation events, duplication events, sorting events and switching events. The tools that are most commonly used in biological studies and that use event-based methods for the analysis of coevolving species associations are TreeMap [3] and TreeFitter [15]. Notable is also Tarzan [9], as it can handle more complex timing information about the phylogenetic trees than th other methods. 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. Whereas these tools differ regarding several aspects, e.g., efficiency, the possibility to include timing information, or the availability of a graphical user interface, they all have in common that the event-based approach requires a cost assignment for the coevolutionary events in advance in order to compute a cost minimal reconstruction.

In this paper a new algorithm and the corresponding tool called CoRe-PA are presented. The new method is based on a dynamic programming formulation for the cophylogenetic reconstruction problem and has significant new features compared to the current state-of-the-art methods TreeFitter, TreeMap, and Tarzan (compare also the recent paper [8] where also a dynamic formulation is used). Algorithm CoRe-PA can handle associations of parasites with multiple hosts, it includes the handling of divergence timing information, unlike most other tools it can handle multifurcations in the input trees, and is very efficient also for large phylogenetic trees due to a dynamic programming formulation for the reconstruction problem. Most notably however is the parameter-adaptive reconstruction approach of CoRe-PA. Different form the other event-based methods, in CoRe-PA no costs have to be assigned to the coevolutionary events in advance. This is achieved by a careful definition of an underlying optimization criteria.

The paper is structured as follows. Basic definitions are given in Section 1. Section 2 introduces the dynamic programming approach utilized in CoRe-PA. The parameter-adaptive approach is described in Section 3. How randomized tests are performed in CoRe-PA is explained in Section 4. Results for several cophylogenetic systems are presented in Section 5. Conclusions are drawn in Section 6.

1 Basic Definitions

Let P and H be two phylogenetic trees. H and P will be called host tree, respectively, parasite tree. Let $\varphi : L(P) \times L(H)$ be a relation over the set of leaf nodes of the parasite tree and the leaf nodes of the host tree. φ is used to describe known host-parasite interactions. A toy example for a cophylogenetic system of four hosts and four parasites and their associations is given in Figure 1 (left).

In order to investigate whether there exists coevolution between hosts and their parasites, their common history is reconstructed from the phylogenies and the known current relationships. Typically, four different types of events are considered for the coevolutionary reconstruction of host-parasite systems: cospeciation events, duplication events, sorting events, and switching events. Cospeciation events refer to simultaneous speciations of host and parasite , duplication

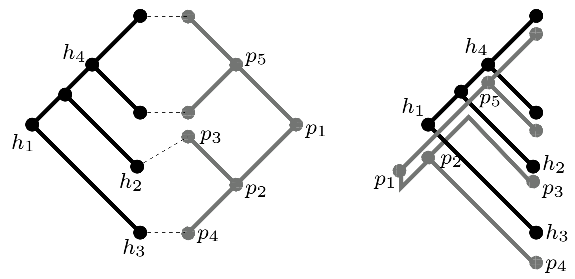

Figure 1: Left: Example for a small coevolutionary system with four host species (leaf nodes in black tree) and four parasite species (leaf nodes in grey tree); Right: example for a cophylogenetic reconstruction for the coevolutionary system; the three associations (p_3, h_2), (p_4, h_3), and (p_2, h_1) induce one cospeciation and one sorting event; the three associations (p_2, h_1), (p_5, h_4), and (p_1, h_1) induce one duplication and two sorting events; the reconstruction need two cospeciations, one duplication, and three sortings

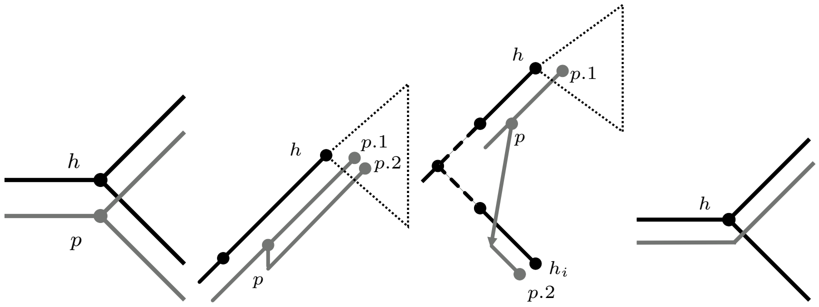

Figure 2: Coevolutionary events (depicted is only the binary case); from left to right: cospeciation (node p associated with node h); duplication (both child nodes of p are associated with a node in the subtree of H with root h); switch (only one child node of p is associated with a node in the subtree of H with root h); sorting; host tree H is depicted black, parasite tree P is depicted grey

events are independent parasite speciations, sorting events correspond to lineage sorting (i.e., a parasite species that lives on a host species remains on only one of the resulting species after a host speciation), and switch events correspond to host shifts. As has been done by other authors (e.g., [6]) we consider a switch as a speciation of the parasite where one of the resulting species switches to another host. The four event types, that are also utilized in CoRe-PA are depicted in Figure 2.

We need the following definitions. If p is a node of a tree, then $p.i$ denotes the i-th child node of p. The out-degree of node p is denoted with $\deg(p)$. An association of a parasite $p \in P$ to a host $h \in H$ is denoted as (p, h). A reconstruction R is the set of all associations of all parasites to nodes in the host tree, i.e. for each node $p \in P$ it exists an $h \in H$ such that $(p, h) \in R$. A reconstruction is valid if i) all parasite leaves are mapped to host leaves according to φ, ii) if node p is mapped to node h, then no descendant of p is associated with an ancestor of h, as this would induce an inconsistency, and iii) at least one child $p.i$ of p has to be associated with an descendant of h. We do not consider the case of a speciation of the parasite p where both child species change to hosts that are outside of the subtree with root h because such events can not be traced back (many other studies also do not allow such events, e.g., [2]).

Based on a valid reconstruction R, the events implied by the associations in R can be inferred as follows. For all non-leaf nodes $p \in P$ the association of p and of all its children $p.i, 1 \leq i \leq \deg(p)$, is considered. If for example, in the case of binary trees, the association (p, h) exists, and $p.1$ is mapped to one child of h and $p.2$ is mapped to the other child of h, then this implies either i) one cospeciation event, or ii) a duplication and two sorting events. This association triple technique has been used before in Tarzan and leads to an efficient reconstruction method (for details see [9]). A valid reconstruction for the coevolutionary system of Figure 1 (left) is depicted in Figure 1 (right). In the reconstruction the three associations (p_3, h_2), (p_4, h_3), and (p_2, h_1) induce one cospeciation and one sorting event (in general many different sets of events may be possible). The three associations (p_2, h_1), (p_5, h_4), and (p_1, h_1) induce one duplication and two sorting events. The depicted reconstruction requires two cospeciations, one duplication, and three sortings.

Due to space limitation we will discuss divergence timing information and incompatible reconstruction only briefly in this article and refer to [2] and [9]. Considering again an association (p, h), where one child $p.i$ is mapped to a node h', and h' is not a descendant of h, then this implies (at least) a host switch event. 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 which are not possible. CoRe-PA includes more sophisticated methods for detecting and solving these so-called incompatible (in contrast to compatible) reconstructions than, for example, Tarzan. However we will focus on the parameter-adaptive reconstruction approach in this article. Furthermore, we point out that CoRe-PA includes the same handling of divergence timing information as Tarzan, i.e., nodes can be labeled with divergence timing information and an association (p, h) is only allowed, if the timing information of p and h do not disallow this association.

2 Dynamic Programming Approach

In the following a DP formulation for the reconstruction problem is given, which is a key component of CoRe-PA. We briefly discuss how the usage of divergence timing information is included, and explain details of runtime optimization techniques that are used. We omit a detailed discussion of how multifurcations and multiple-host parasites are handled (multifurcations, e.g., can either be resolved by iterating over all possible binary subtrees or by introducing artificial new cophylogenetic events, e.g., a cospeciation of multiple host and parasite species).

2.1 Initial DP Formulation

The basic idea of the dynamic programming approach is to traverse the parasite tree P in a bottom-up manner. The cheapest cost $C_{p,h}$ for a node p of P, that is mapped on a node h of H, is stored in the dynamic programming table. If p is a leaf node, then the mapping for p is defined by the relation φ and induces no costs as no coevolutionary event occurs. In the recursive step of the dynamic programming we map all children $p.1, \ldots, p.\deg(p)$ of p to nodes in H. The mapping of the nodes $p.i$ to nodes $h_i \in H$ induces i) the recursively computed cost $C_{p.i,h_i}$ for each of the association, plus ii) the cost from the cheapest set of events due to p being associated with h, and the nodes $p.1, \ldots, p.\deg(p)$ being associated with the corresponding h_i. Note that there may exist several possibilities for this set of events to explain the given associations, and the cost-wise cheapest of those is taken. These costs are denoted by $\min(E(h, h_1, \ldots, h_{\deg(p)}))$. Let us consider again the binary example where h_1 and h_2 are children of h (i.e. $h_1 = h.1$ and $h_2 = h.2$, or $h_1 = h.2$ and $h_2 = h.1$). In this example $\min(E(h, h_1, h_2))$ refers either to the costs for one cospeciation event or to the costs for one duplication and two sorting events. The dynamic programming formulation is as follows:

$$C_{p,h}=\begin{cases} 0 & \text{if } p\in L(P),\ (p,h)\in\varphi \\ \infty & \text{if } p\in L(P),\ (p,h)\notin\varphi \\ \min\limits_{\substack{(h_1,\ldots,h_{\deg(p)}) \\ \in H^{\deg(p)}}} \left(\left(\sum\limits_{i=1}^{\deg(p)} (C_{p.i,h_i}) \right) + \min(E(h,h_1,\ldots,h_{\deg(p)})) \right) & \text{otherwise} \end{cases} \tag{1}$$

2.2 Inclusion of Divergence Timing Information

Similar to the approach in [9], algorithm CoRe-PA allows to assign intervals of time zones to the nodes in one of the trees, e.g., the parasite tree. 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 [9]. For each possible association (p, h) we define a value $Z_{p,h}$. The value of $Z_{p,h}$ is 0 if the association is valid with respect to the timing information, and it is ∞ otherwise. For the revised DP formulation we add the value $Z_{p,h}$ in the recursion step of Eqn. 1.

2.3 Optimization

A direct implementation of the DP formulation, as given in Eqn. 1, would not perform very well, as all possible combinations of all possible associations of nodes $p.i$ to nodes h_i would be considered in order to compute $C_{p,h}$. Therefore several improvements are included into the implementation of CoRe-PA. The most important reduces the number of combinations of associations that have to be considered significantly as described in the following. If the costs for $C_{p,h}$ are computed according to Eqn. 1, all possible mappings of each $p.i$ to all $h \in H$ are considered. Let us assume two possibilities for mappings of $p.i$, namely $p.i$ being mapped to h' and $p.i$ being mapped to h''. Let us further assume that h' and h'' are both in a subtree of H that has a child of h as a root node. As we know the values of $C_{p.i,h'}$ and $C_{p.i,h''}$ (due to the recursive approach) and as the number of sorting events induced by the pair of associations (p, h) and $(p.i, h')$ (respectively (p, h) and $(p.i, h'')$) is known, one of the associations (either $(p.i, h')$ or $(p.i, h'')$) will dominate the other (unless the costs are equal). This is true for every pair of host nodes that occur in the same subtree of H that have a child of h as root node. Therefore, only the association that induces the smallest cost in such a subtree must be considered and the number of combinations to be considered in the recursive approach is reduced significantly. This is not only true for all these subtrees, but also for the set of all other nodes that are neither h itself nor in one of the just described subtrees.

In addition to this dominance-based optimization CoRe-PA heavily utilize precomputed tables. Assume that an arbitrary parasite node p is being mapped on h and a child $p.i$ of p is being mapped on h'. A certain set of events that have to occur can be precomputed: for example, if h' is a descendant of h, the number of sorting events can be computed; in other cases host switches can be inferred beforehand. In order to perform such precomputations, it is assumed that each possible h and h' for the mapping of an arbitrary p and the child node $p.i$. Also in the case that divergence timing information is used, the best take-off and landing sites can be precomputed in the same manner. Details are omitted due to space limitations.

Let n be the maximal number of nodes in the host or in the parasite tree. It is not difficult to see, that computing a reconstruction with CoRe-PA runs in order of $O(n^3)$, if the maximal degree of the nodes in the trees is assumed to be constant.

3 Parameter-Adaptive Cophylogenetic Reconstruction

Several optimization criteria have been investigated in the literature that utilize event-based cophylogenetic reconstruction methods. Examples include the minimization of overall reconstruction costs or the maximization of the number of cospeciations. But all methods are strongly dependant on a good estimation of the cost vector, that assigns costs to the events. Often cospeciation costs are considered to be small (for example ≤ 0), and duplication and host switch costs are usually assumed to be high. However, from a biological point of view, the exact values for these costs are basically unknown. In [16] an inspiring comment is given: *"If each event is associated with a cost that is inversely related to the likelihood of the event (the more likely the event, the smaller the cost) then the most parsimonious reconstruction will also, in some sense, be the most likely explanation of the observed data."*. This comment nicely reflects the underlying idea of the parameter-adaptive approach of CoRe-PA, that will be described in the following.

Unlike other methods CoRe-PA does not require any restrictions on the cost values. However, for the parameter-adaptive approach we assume all event costs are between 0 and 1 (If the y are larger this can be achieved by muliplication with a suitable factor). Let $\overline{c} = (c_1, \ldots, c_n)$ be the cost vector for the n possible events. Based on this cost setting it is expected that the event indexed by i occurs with probability

$$p_i = \frac{1/c_i}{\sum_{j=1}^{n} 1/c_j}, \qquad (2)$$

i.e., the probability for a certain event is the normalized value of the reciprocal event cost. Based on the cost vector a cost-minimal reconstruction is inferred using the DP formulation as given in Section 2; this in turn leads to relative event frequencies r_i of the events, based on the computed reconstruction. Assume that cost vector $\overline{c}$ is used to determine a reconstruction. The obvious method to determine how good the reconstruction and the cost vector fit, is based on the sum of the differences of the probabilities p_i and the corresponding relative event frequencies r_i of the reconstruction. Formally,

$$q_{\overline{c}} = \sum_{i=1}^{n} |p_i - r_i|. \tag{3}$$

By using $q_{\overline{c}}$ as an optimization criteria, a cost vector $\overline{c}$ is sought such that $q_{\overline{c}}$ is minimized. The value $q_{\overline{c}}$ can be interpreted as a quantification of how unlikely a reconstruction is. Furthermore, if, based on some significance test, there is a strong support for coevolution, but the corresponding $q_{\overline{c}}$ is very high, then the support for the coevolutionary signal has still to be questioned.

The parameter-adaptive approach reduces the parameterized cophylogenetic reconstruction problem to a parameter-adaptive optimization problem. Of course, many sophisticated methods are known for finding a good vector $\overline{c}$, like metaheuristics [5] or utilizing the concept of a simplex (like in the Nelder-Mead downhill simplex method [10]). In order to be able to present a reasonable statistical analysis of the parameter-adaptive component of CoRe-PA and not to be biased by an underlying optimization method, we present only results that are based on randomly chosen (uniform distribution) cost vectors (although the Nelder-Mead simplex method is already included in CoRe-PA).

4 Randomized Tests in CoRe-PA

In order to evaluate whether the number of different phylogenetic events of a reconstruction indicates significant coevolution, different randomization tests can be used (see, e.g., [17]). The idea of these tests is to create reconstructions for scenarios where part of the problem instance is randomly changed, e.g., the hosts and parasite associations can be changed randomly. Then the number of events in the reconstructions for the random scenarios can be compared to the reconstruction for the original host parasite scenario. Different opinions have been stated in the literature about what part of the host-parasite data should be randomized when creating random instances for a significance test. Some possibilities are to randomize the parasite tree, the host tree, both trees, or the associations between host and parasites (see [17]). It is important that the random instances are biologically plausible because otherwise the significance results that can be obtained with the tests are biologically useless. Therefore, different methods have been proposed how the random instances should be generated (see [1] for an overview).

One randomization test that is integrated in TreeMap is the most often used test in literature on host parasite coevolution (see, e.g., [12]). The test asks whether the maximum proportion of cospeciating nodes inferred is greater than the maximum proportion that can be inferred when one of the phylogenies is randomized. TreeMap allows to randomize either one tree (the host or the parasite tree) or both trees. All these possibilities have been used in the literature.

In [17] the proper use of randomization methods in order to analyze, whether the fit between hosts and parasites can be explained by coevolution, is discussed.

It was argued that for a corresponding test it is not appropriate to make random changes in the host or parasite tree. Instead it was proposed to keep the phylogenies of the hosts and the parasites as well as the number of associations. Only the associations between the hosts and parasites should be randomized. This method has been used, e.g., in [13]. For many host parasite systems it can be observed that the number of different parasite species on one host species is small. For such a system it might not be biologically meaningful if a random association between hosts and parasites is created by assigning each parasite a random host with equal probability. Therefore we propose here that random associations should be created that keep the character of the host parasite assignment in the following sense. The number of hosts that have k parasite species should be the same in the original host parasite system and the random instance for all integers k. All the discussed methods are included in CoRe-PA. In the case that random trees have to be generated, the well known β-splitting model [1] is employed. The β-splitting model includes the Markov model and the PDA model as special cases. The method for randomizing the parasite tree (resp. the host tree and both trees) is denoted by RND-parasite (resp. RND-host and RND-both); the method when associations are randomized while their character is preserved is denoted by RND-assoc.

5 Results

Six biological coevolutionary systems that have already been studied intensely in literature are used as test examples in this. Note that in coevolutionary systems multifurcations are often resolved artificially into bifurcations, although there are clear indications that the support for this based on the biological data is very weak. Furthermore, if not stated otherwise, the data sets from the literature do not contain multi-host parasites, although there is sometimes support for this in the underlying data. These restrictions are necessary in order to be able to use standard tools for cophylogenetic reconstruction; CoRe-PA would not require these restrictions. When generating random trees with the β-splitting model, we always use $\beta = -1$ as suggested in [1]. Note that all reconstructions in this section, which are suggested by CoRe-PA, are compatible.

5.1 Biological Data Sets

The test systems are gophers hosts and lice parasites (see Figures 11 and 13 in [2], denoted by S_1 in this paper), two systems of Pelecanicform bird hosts and *Pectinopygus* lice parasites (see Figure 2, 4, and 5 in [6], denoted by S_{2-ML} and S_{2-MP}), a system of hystriocognath rodents and pinworm parasites (see Figure 6.5 in [7], denoted by S_3), a system of seabirds and their chewing lice (see Figure 12.4 in [12] denoted by S_4), and a recently presented system of *Microbotyrum* funghi and their Caryophyllaceae hosts that includes multihost parasites (see Figure 4 in [14], denoted by S_5).

5.2 Parameterized Reconstruction of Random Trees

A problem with inferring cophylogenetic reconstruction based on a (standard) cost vector is that the frequencies of certain events strongly depend on the size

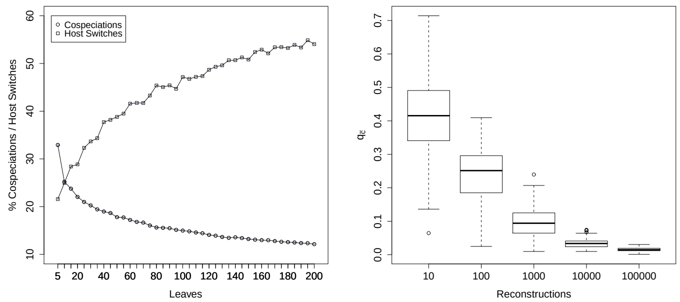

Figure 3: Left: Mean frequency of switch and cospeciation events based on random tree pairs with $5, 10, \ldots, 200$ leaf nodes; fixed costs for cospeciation, sorting, duplication, and host switches are $co = -2$, $so = 1$, $du = 2$, and $hs = 4$; Right: Convergence behavior based on $q_{\overline{c}}$ for CoRe-PA on data set S_1 when searching for the best cost vector; depicted are box plots for $q_{\overline{c}}$ for 100 independent test runs after $10, 100, 1000, 10000, 100000$ cost vectors have been choosen randomly

of the input data set. To investigate this, we created 100 random tree pairs with random associations for $5, 10, \ldots, 200$ leaf nodes (all together 4000 tree pairs). A fixed cost vector was used with cost settings for cospeciation, sorting, duplication, and host switches being $co = -2$, $so = 1$, $du = 2$, and $hs = 4$. Note that in standard cost vectors used in literature, the cospeciation event has usually higher costs whereas the switch event has usually lower costs. However, even when exaggerating these values, the frequency, for example, of host switches in reconstructions grows dramatically when using larger trees. The 40 mean values for the frequencies of the number of host switches and for the number of cospeciations, based on the 40 sets of 100 random tree pairs, are depicted in Figure 3 (left). The results clearly indicate that host switches become more and more likely when larger phylogenetic trees are used (respectively cospeciations become more and more unlikely). Reconstructions for large trees (more than ≈ 30 leaf nodes), that are based on real biological data, show very similar results: cost-minimal reconstructions tend to have many unrealistic host switches with take-off and landing sites close to the leaf nodes and the number of cospeciations becomes very small (results not given here).

5.3 Parameter-Adaptive Reconstruction

When using the parameter-adaptive approach of CoRe-PA, 100000 cost-minimal reconstructions are computed based on randomly chosen cost vectors. The reconstruction with the smallest value for $q_{\overline{c}}$ (cmp. Eqn. 3) is the reconstruction suggested by CoRe-PA. When employing randomization methods RND-{host, parasite, both, assoc}, then of course also for each randomized instance 100000 cost-minimal reconstructions are computed based on randomly chosen cost vectors, and the resulting value $q_{\overline{c}}$ refers to the best of these.

In Figure 3 (right) the convergence behavior of CoRe-PA is depicted for system S_1. Given are box plots of $q_{\overline{c}}$ based on 100 test runs that were stopped

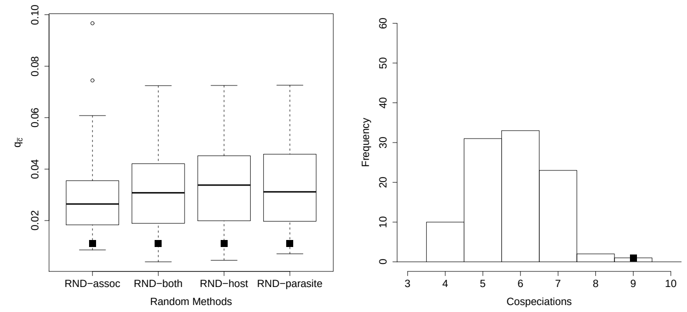

Figure 4: Left: randomization methods RND-{assoc, both, host, parasite} on system S_4; for each box plot 100 random instances were created and $q_{\overline{c}}$ was computed based on 100000 reconstructions for each instance; Right: histogram for the number of cospeciations for system S_4 when using randomization method RND-assoc; based on the original instance CoRe-PA suggested a reconstruction with 9 cospeciations; black squares indicate the outcome of CoRe-PAfor the unmodified test instance

after $10, 100, 1000, 10000$, and 100000 cost vectors have been chosen randomly in each run. The results indicate that the algorithm is in a nearly converged state after 100000 randomly chosen cost vectors were used. Of course an optimization method would improve convergence (but could bias the significance results).

Results for the four different randomization methods are given in Figure 4 (left) for system S_4. Depicted are the box plots for $q_{\overline{c}}$ (100 randomized test instances were created based on the methods RND-{host, parasite, both, assoc}). It can be seen that the method of randomization has only a small influence on the overall result of $q_{\overline{c}}$, and that $q_{\overline{c}}$ is significantly smaller for the original instance compared to randomized instances. In the rest of this result section we will only employ the method RND-assoc (all results for the other randomization methods were very similar). The frequency of the number of cospeciations that occurred in the randomized instances for S_4 (method RND-assoc) are depicted in the histogram in Figure 4 (right). This figure clearly indicates the strong support for coevolution, as no reconstruction had more cospeciations than the reconstruction suggested by CoRe-PA.

In Table 1 the overall results are given when CoRe-PA is applied to the six systems. For the solution having the smallest value for $q_{\overline{c}}$, we give the number of events, the best cost vector, and the value for $q_{\overline{c}}$. For each system 100 randomized instances were created (method RND-assoc); the column $p_{co,>}/p_{co,\geq}$ (respectively p_{qu}) denotes the probability, that a randomized instance lead to reconstructions with (an equal number or) more coevolutionary events (resp. to reconstructions with a smaller $q_{\overline{c}}$). Figure 5 (left, respectively right) depicts the box plots for the number of cospeciation (respectively for $q_{\overline{c}}$) based on the 100 randomized instances, and the value of cospeciations (resp. $q_{\overline{c}}$) for the reconstruction suggested by CoRe-PA for the unmodified test instance (indicated by the black square).

There is a strong indication for a coevolutionary history for systems S_1 and S_4 with respect to the number of cospeciations. As $q_{\overline{c}}$ is very small for these

System	event frequency	best cost vector	$q_{\overline{c}}$	$p_{co,>}$ / $p_{co,\geq}$	p_{qu}
S_1	(6, 5, 2, 1)	(0.166, 0.198, 0.512, 0.987)	0.008	0.00/0.13	0.04
S_{2-ML}	(10, 20, 5, 2)	(0.226, 0.114, 0.457, 0.989)	0.015	0.04/0.13	0.24
S_{2-MP}	(12, 18, 5, 0)	(0.007, 0.005, 0.018, 0.882)	0.036	0.00/0.00	0.78
S_3	(8, 15, 3, 1)	(0.095, 0.053, 0.268, 0.738)	0.024	0.00/0.00	0.28
S_4	(9, 11, 3, 1)	(0.040, 0.033, 0.125, 0.386)	0.011	0.01/0.03	0.05
S_5	(6, 32, 9, 4)	(0.388, 0.072, 0.252, 0.587)	0.006	0.87/0.98	0.00

Table 1: Results of CoRe-PA for coevolutionary systems $S_1, \dots, S_5$; the event order for the vectors in column 2 (absolute event frequency) and column 3 (best cost vector) is (cospeciation, sorting, duplication, host switch); $q_{\overline{c}}$ as in Eqn 3; $p_{co,>}$ (respectively $p_{co,\geq}$): probability that a reconstruction based on a randomized instance leads to more (respectively, an equal number of more) cospeciations; p_{qu}: probability that a randomized reconstruction leads to a smaller $q_{\overline{c}}$; in all test runs randomization method RND-assoc was used

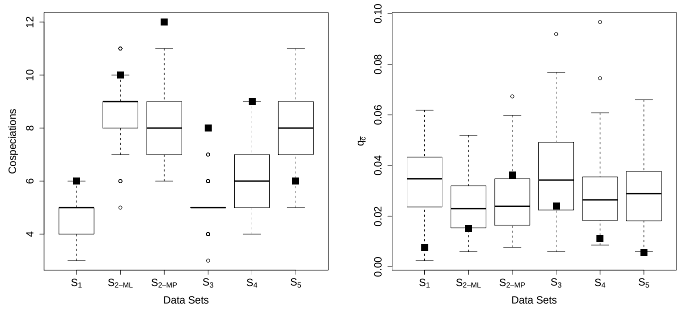

Figure 5: Box plots for the number of cospeciations (left) and $q_{\overline{c}}$ (right) based on 100 randomized test instances (method RND-assoc) for systems $S_1, \dots, S_5$; black squares indicate the corresponding value for the solution suggested by CoRe-PA

systems this outcome should be interpreted as a clear sign of coevolution. Systems S_{2-ML}, S_{2-MP}, and S_3 have also a quite strong evidence for coevolution based on $p_{co,\geq}$, but the support for this (cmp. p_{qu}) is only reasonably good for S_{2-ML} and S_3, and bad for S_{2-MP} ($p_{qu} = 0.78$). The values for system S_5 should be interpreted as a clear sign of no coevolution ($p_{co,\geq} = 0.98$) with a strong support for this result based on $p_{qu} = 0.00$. Note that the extensive studies in the literature [7, 14] for systems S_{2-ML}, S_{2-MP}, and S_5 also do not conclude that there is a clear coevolutionary signal, and the tools used showed partially contradicting results.

Although a detailed discussion of any of the reconstructions is not possible in this paper, we want to point out that for systems S_4 (respectively S_1, S_{2-ML}, and S_{2-MP}) the best reconstruction was identical (respectively very similar) to the reconstruction given in the literature, without setting any costs for the events.

6 Conclusions

We have introduced a new algorithm and a corresponding tool called CoRe-PAfor parameter-adaptive cophylogenetic analysis. Different from other event-based reconstruction methods CoRe-PA does not require any cost settings for the considered cophylogenetic events in advance, but seeks for the cheapest reconstruction in which the used costs are inversely related to the relative frequency of the corresponding event. The quality of the reconstructions obtained with CoRe-PA was analyzed experimentally on six coevolutionary systems. The results show that CoRe-PA is very useful when it is difficult or impossible to assign exact cost values to different types of coevolutionary events in advance.

References

[1] D. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. *Statistical Science*, 16(1):23–34, 2001.

[2] M. Charleston. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosiences*, 149:191–223, 1998.

[3] M. Charleston and R. Page. A macintosh program for the analysis of how dependent phylogenies are related, by cophylogeny mapping, 2002. http://www.cs.usyd.edu.au/~mcharles/software/treemap/treemap.html.

[4] M. Charleston and S. Perkins. Traversing the tangle: Algorithms and applications for cophylogenetic studies. *Journal of Biomedical Informatics*, 39:62–71, 2006.

[5] F. Glover and G. Kochenberger, editors. *Handbook of Metaheuristics*, volume 57 of *International Series in Operations Research & Management Science*. Springer, 2003.

[6] J. Hughes, M. Kennedy, K. P. Johnson, R. L. Palma, and R. D. M. Page. Multiple cophylogenetic analyses reveal frequent cospeciation between pelecaniform birds and pectinopygus lice. *Systematic Biology*, 56(2):232–251, 2007.

[7] J.-P. Hugot. New evidence for hystricognath rodent monophyly from the phylogeny of their pinworms. In R. Page, editor, *Tangled Trees: Phylogeny, Cospeciation, and Coevolution*, pages 144–173. The University of Chicago Press, 2003.

[8] R. Liebeskind-Hadas and M. A. Charleston. On the computational complexity of the reticulate cophylogeny reconstruction problem. *Journal of Computational Biology*, 16(1):105–117, 2009.

[9] D. Merkle and M. Middendorf. Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. *Theory in Biosciences*, 123:277–299, 2005.

[10] J. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[11] R. Page. *Tangled Trees: Phylogeny, Cospeciation and Coevolution*. The University of Chicago Press, 2003.

[12] A. Paterson, R. Palma, and R. Gray. Drowning on arrival, missing the boat, and x-events: How likely are sorting events? In R. Page, editor, *Tangled Trees: Phylogeny, Cospeciation, and Coevolution*, pages 287–309. The University of Chicago Press, 2003.

[13] S. Perlman, G. Spicer, D. Shoemaker, and J. Jaenike. Associations between mycophagous drosophila and their howardula nematode parasites: a worldwide phylogenetic shuffle. *Molecular Ecology*, 12(1):237–249, 2003.

[14] G. Refrégier, M. L. Gac, F. Jabbour, A. Widmer, J. A. Shykoff, R. Yockteng, M. E. Hood, and T. Giraud. Cophylogeny of the anther smut fungi and their caryophyllaceous hosts: Prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation. *BMC Evolutionary Biology*, 8(100), 2008.

[15] F. Ronquist. Treefitter 1.0, 2001. http://www.ebc.uu.se/systzoo/research /treefitter/treefitter.html.

[16] F. Ronquist. Parsimony analysis of coevolving species associations. In R. Page, editor, *Tangled Trees: Phylogeny, Cospeciation, and Coevolution*, pages 22–64. The University of Chicago Press, 2003.

[17] M. Siddall. Computer-intensive randomization in systematics. *Cladistics*, 17:35–52, 2001.