Evaluation of Host Parasite Reconciliation Methods using a new Approach for Cophylogeny Generation

Stephanie Keller-Schmidt*[1], Nicolas Wieseke*[2] , Konstantin Klemm[1] and Martin Middendorf[2]

[1]Bioinformatics Group, Institute of Computer Science, University Leipzig, Germany
[2]Parallel Commputing and Complex Systems Group, Institute of Computer Science, University Leipzig, Germany

Email: Stephanie Keller-Schmidt*- keller-schmidt@informatik.uni-leipzig.de; Nicolas Wieseke*- wieseke@informatik.uni-leipzig.de; Konstantin Klemm - klemm@bioinf.uni-leipzig.de; Martin Middendorf - middendorf@informatik.uni-leipzig.de;

*Corresponding author

Abstract

Background: Coevolution between species is a common phenomenon in biology: species interact across groups such that the evolution of a species from one group can be triggered by a species from another group. Most prominent examples are systems of host species and their associated parasites. Typically in this field, phylogenetic trees for both groups of species can be constructed from sequence data or/and morphological data. In addition, the host parasite interactions between the extant taxa are known empirically. The problem is then to reconcile the common history of both groups of species and to predict command line the associations between ancestral hosts and their parasites. Some algorithmic methods have been developed in recent years to solve this reconciliation problem. Only few host parasite systems, however, have been analyzed in sufficient detail to serve as benchmarks for the evaluation of the reconstruction methods.

Results: We propose to tackle the lack of benchmarks by generating meaningful test data sets with a dedicated approach for generating cophylogenies. Our method builds on biologically motivated branching models to generate cophylogenies under the assumption of the widely used coevolutionary model. It pictures coevolution as a stochastic process with cospeciation, duplication, lineage sorting and (host) switching as discrete events. The probability of an independent parasite speciation as well as the ratio between cospeciations and sortings and between duplications and host switches are user defined parameters. We evaluate choices of reasonable parameter settings under the aspect of producing realistic coevolutionary scenarios, giving rise to a large set of

test scenarios. Based on these scenarios, we provide a detailed analysis and comparison of the common reconciliation tools TreeMap 3b, Jane 2.0, and CoRe-PA with a focus on the significance of the computed reconstructions. All three tools are based on the maximum parsimony principle but using different heuristics and cost models. To the best of our knowledge, this is an initial contribution to extensively compare methods for cophylogeny reconciliation.

Conclusions: Out of the three tools applied to the test data sets, CoRe-PA yields the most precise predictions of the associations between hosts and parasites. However, it does not optimally estimate the number of cospeciation and switching events and is the computationally most expensive method. Jane 2.0, being the fastest of the three tools, is best at estimating the correct number of cospeciations. TreeMap 3b is the only tool with the option to find the optimal reconstruction for a specified cost model. In terms of accuracy of the computed reconstructions, TreeMap falls short of the other tools.

The respective application CoRe-Gen for generating randomized coevolutionary host parasite systems is freely available at http://pacosy.informatik.uni-leipzig.de/files/19/core-gen.tar.gz

---

Background

In the research field of phylogenetics, the recent advent of large genetic data sets offers increased insight into the evolutionary histories of species. Representations of such histories are phylogenies, which typically are binary trees with leaves corresponding to extant taxa and inner nodes representing ancestral species. In order to understand the driving forces of evolution leading to a high diversity of species, the reconstruction of phylogenies is inevitable. Statistical macroevolutionary growth models are used to understand the dynamical rules of evolutionary processes such as the speciation and extinction. The simplest model, generally referred to as the null hypothesis, is the Yule model [1] (also called Equal Rate Markov model, ERM model) which describes a continuous-time branching process where each speciation is equally likely [2, 3]. But the evolution of species cannot be understood as a closed system. Species are able to interact and may mutually affect their evolution. This can be described by the more complex problem of coevolution or cophylogenetics. Examples for coevolutionary systems are relationships between hosts and their associated parasites, between predators and pray, or between groups of species with symbiotic

interactions.

Here we focus on the coevolution of parasites with their hosts. Several methods have been proposed [4–9] to infer plausible coevolutionary histories from given phylogenies for the host species and the parasite species and an assignment of the extant parasite species to their host species. Assessing the accuracy of these methods requires benchmarks, preferably based on empirically confirmed data of coevolutionary histories. However, such data are scarce. The main reason is that it is very difficult to get clear evidence about the former relations between the predecessors of the extant host and parasite species. Data from simulated coevolution might be able to fill the gap. Here we take a first step in this direction and propose a method for the generation of cophylogenies. Based on sets of cophylogenies that have been generated by this method we study the accuracy of several cophylogeny reconstruction methods that have been proposed in the literature.

For evaluation purpose Doyon et al. presented a simulation approach for coevolutionary scenarios in [5]. Therefore an ultrametric tree (i.e., the host tree) was generated with a standard birth death process. Additionally the dependent tree (i.e., the parasite tree) was created by generating coevolutionary events according to a Poisson process with respect to the rates of the respective events. Unfortunately this approach requires a dating scheme of the independently generated host tree and biologically motivated estimations of the coevolutionary event rates.

To avoid timing issues and evolutionary rates a new method of generating cophylogenetic scenarios can be used. Utilizing stochastic branching models like the ERM [1] or the age model [10] our intention was to extend these models to produce evolutionary dependencies between two simultaneously generated phylogenies. Such type of dependencies have been described in the well-known coevolutionary event-model (see, e.g., [4]). The branching models are used to generate binary trees iteratively by speciating a leaf chosen with a probability distribution given by the model. This process is combined with the four types of events that are typically used to describe host parasite coevolution, namely cospeciation, duplication, host switch, and sorting. We compare the cophylogenies that have been generated by our method using different growth models with a focus on the proper choice for the parameter values of the generation model. Furthermore, generated pairs of phylogenetic trees consisting of a host tree and a parasite tree need to be compared in the context of a cophylogenetic analysis such that biologists are able to explore the relative rate of evolution with the knowledge about the coevolution of hosts and their parasites [4]. Common cophylogenetic reconstruction tools are TreeMap 3b [11], Jane 2.0 [9], and CoRe-PA [7]. These tools are evaluated with a focus on the significance of the reconstructions that they deliver for the test sets of

cophylogenies that have been generated with the different dynamical branching models.

Methods

In the following section, some basic definitions are given and it is described how the growth models for trees can be used to accommodate a coevolutionary event model in order to generate cophylogenies. Subsequently we discuss properties of the resulting cophylogenies to assess their biological plausibility.

Basic Definitions

The definitions of phylogenetic trees and different stochastic growth models for binary trees that are used in this article are given in the following section. The principle of coevolution and the considered coevolutionary event model is also explained.

Phylogenetic Trees

Phylogenetic trees describe the evolutionary history between different organisms. Here we consider phylogenetic trees as rooted binary trees where inner nodes represent ancestral species and leaf nodes represent extant species. In this study the focus is on cophylogenies which consist of two (coevolved) phylogenetic trees, a host tree T_h and a parasite tree T_p. It describes the interaction of species across groups such that the evolution of a species from one group, i.e., the parasite, developed in dependence from a species of another group, i.e., the host.

Growth Models

The simplest stochastic growth model for the generation of binary trees is the ERM model, also known as Yule model [1]. Starting from a tree with a single node, the root, the ERM model iteratively expands the tree by choosing a leaf i of the current tree and attaching two new leaf nodes j and k to it, thereby turning i into an inner node. In a tree with n leaves, each leaf i is chosen with probability $1/n$.
As a variation of the Yule model, the age model [10] uses a probability $p_i(t)$ of choosing leaf i dependent on the age of leaf i. The age $\tau_i(t) = t - t_i$ is the number of time steps passed between the current time t and the time t_i, when leaf i was generated. Specifically, the probability is chosen inversely proportional to age,

$$p_i(t) \propto \tau_i(t)^{-1} \ . \tag{1}$$

with appropriate normalization.

The beta-splitting model [12] is widely used to stochastically generate binary trees that have an imbalance which is tunable by a parameter β. This kind of model, however, directly defines a probability distribution over all binary trees with a given number of leaves. In contrast to the ERM and age model, beta-splitting does not provide a dynamic rule to iteratively build up the tree. Therefore, it does not serve as the basis for defining a model of coevolution where the ordering of speciation of events is crucial.

Coevolution

The coevolution of two groups of species is studied in order to explore the combined evolutionary history. Therefore, the two phylogenetic trees T_h and T_p of both species are inferred. To this end, the observed host parasite associations in the extant species have to be known. The associations can be seen as a relation ϕ between the different leaf sets, i.e., $\phi \subset L(T_p) \times L(T_h)$. In this paper we assume that one parasite species can be associated to at most one host species. This assumption is widely used in the literature on algorithms for the analysis of coevolution. Note, however, that there are several empirical examples where this assumption does not hold. An example of an artificial coevolutionary system is given in Fig. 1 (left). A common approach for the reconstruction of coevolutionary histories establishes a mapping from the parasite tree onto the host tree. In this way, ancestral dependencies between parasites and their hosts are predicted. Coevolution is captured in terms of events. Here we employ four different types of events. The first two are host dependent events: *cospeciation* (co) and *sorting* (so) describe the reaction of a parasite if its associated host performs a speciation. The remaining two events are host independent, namely, *duplication* (du) and *host switching* (sw) where the speciation of a parasite occurs without a speciation of an associated host (see Fig. 2). In case of the cospeciation event, host and parasite speciate simultaneously. A sorting event describes the lineage sorting of a parasite across the speciation of its associated host. In this case, the parasite species remains on only one of the newly emerged host species. The duplication event describes the speciation of a parasite alone. The resulting two child species are associated to the same host as the parent species. A host switching event refers to a host shift of one of the parasite child species immediately after a speciation [13]. To each of the four event types, a cost value is assigned taking into account the likelihood of the event. Less likely events incur larger cost. Using maximum parsimony a reconstruction is sought such that the total costs of all events that occur is minimal. Depending on the chosen event costs, i.e., the cost model, different reconstructions can be optimal. A reconstruction being optimal under a certain cost model is called a Pareto optimal solution of the coevolutionary system. An example of a reconstruction for the coevolutionary system depicted in Fig. 1 (left) is given in Fig. 1 (right).

Cophylogeny Generation Model

The growth models for phylogenetic tree generation can not directly be used for the generation of cophylogenies. The reason is that it is essential that the two phylogenetic trees are generated simultaneously with respect to the intended dependencies between the corresponding groups of species. Therefore, the aim here is to adopt common growth models to meet these demands.
The starting point is a host tree and a parasite tree, both consisting of a single node. Furthermore, the parasite node is associated with the host node. Then, one node is chosen for an upcoming speciation. If the selected node is from the parasite tree, this results in a host independent coevolutionary event (i.e., a duplication or host switch). Otherwise the event is host dependent (i.e., a cospeciation or sorting). To decide which node is the next to speciate, a parameter p_{hc} is introduced, giving the probability that the node belongs to the host tree. The respective probability for selecting a parasite node is defined by

$$p_{pc} = 1 - p_{hc} \tag{2}$$

With this probability, only the type of the node (i.e., host or parasite) is chosen. The decision of which leaf in the host tree, respectively parasite tree, is taken is done according to the considered branching model. Thus it is ensured that both created trees satisfy the particular branching model. Furthermore, in each step it is clear which are the current extant species. This information is needed later for producing time consistent host switching events, as a parasite can only switch to a host which existed at the same time.
To achieve the intended dependencies between host and parasite species, additional parameters have to be considered. These parameters p_{co}, p_{so}, p_{sw}, p_{du} define the probabilities for the respective coevolutionary events cospeciation, sorting, host switch, and duplication. Thereby it holds

$$p_{co} = 1 - p_{so} \tag{3}$$

$$p_{sw} = 1 - p_{du} \tag{4}$$

It can be seen that the probability for p_{so} respectively p_{du} can be inferred from p_{co} respectively p_{sw} using Eq. 3 and 4. Therefore these parameters can be obtained from the ratio between the event frequencies of the two host dependant (respectively the two host independent) event types. Compared to the approach presented in [5] it is easier to estimate these rations than the true evolutionary rate for each of the events.
In case of a host dependent event occurring after a host node is chosen for speciation, for each associated parasite it has to be decided with probability p_{co} if the parasite speciates too, resulting in a cospeciation event. Otherwise a sorting event occurs and the parasite remains on only one of the newly emerged host

children. The respective host child is selected randomly with an equal probability. In case of a host independent event after a parasite node is chosen for speciation, at least one of the child species remains on the same host species. The other child species can switch to a randomly selected host leaf with probability p_{sw} or otherwise remains on the same host species too.

In that way the generation of both trees T_h and T_p and their respective associations is done iteratively until there exists a given total number s of extant species, i.e., $s = |L(T_h)| + |L(T_p)|$. The pseudocode describing this principle is shown in Alg. 1.

Algorithm 1: Pseudocode for the generation of a coevolutionary history

```
Input: trees T_h, T_p each with only a single node, size s, probabilities p_hc, p_co, p_du, p_sw, p_so
Output: cophylogeny composed of a parasite tree T_p associated with a host tree T_h and
        s = |L(T_h)| + |L(T_p)|
while s ≠ |L(T_h)| + |L(T_p)| do
    with uniform probability chose r ∈ [0, 1];
    if r ≤ p_hc then
        choose leave l ∈ L(T_h) w.r.t. a branching model;
        foreach parasite associated with host l do
            with uniform probability chose r ∈ [0, 1];
            if r ≤ p_co then
                do cospeciation;
            else
                do sorting;
    else
        choose leave l ∈ L(T_p) w.r.t. a branching model;
        with uniform probability chose r ∈ [0, 1];
        if r ≤ p_sw then
            do switch to a randomly selected host from L(T_h);
        else
            do duplication;
    update T_h, T_p;
```

Properties of Generated Cophylogenies

It is obvious that not all combinations of parameter values for the proposed cophylogeny generation method lead to "relevant" cophylogenies. For example choosing $p_{hc} = 0$ will result in a single host node, as no host will ever be chosen for a speciation. Thus all parasite nodes will be associated with this single host. On the other hand if the probabilities p_{hc} and p_{so} are both 1 then only host nodes are chosen and the one associated parasite does always a sorting. This results in a parasite tree with a single node associated to one of the host leaves.

To decide whether a generated host parasite system is a "relevant" data set or not properties have to be found which describe if a certain cophylogeny is similar to real biological data. The number of empirically confirmed cophylogenies does not allow a meaningful statistical analysis on that. But the host parasite systems seems to have several things in common. Studies by [13–20] have shown that the sizes of the two trees T_h and T_p differ slightly in the way that T_p is often somewhat larger. Additionally, there is usually no host taxa included which is not associated with at least one parasite. Also every host harbors approximately the same number of parasite species.

Thus, in order to evaluate the generated host parasite systems the following two characteristics are considered: The ratio between parasite tree size and host tree size and the variance of the number of associated parasites per host leaf. Generated cophylogenies with a size ratio close to 1 and variance close to 0 are considered to be more likely similar to biological cophylogenies.

Formally the ratio between the sizes of parasite and host tree (*scale*) is defined as

$$scale = \frac{|L(T_p)|}{|L(T_h)|} \tag{5}$$

The variance of the number of associated parasites (*var*) is defined as

$$var = \frac{\sum_{h_i \in L(T_h)} (x_{hi} - \mu)^2}{|L(T_h)|} \tag{6}$$

with x_{hi} being the number of parasites associated with host leaf h_i, i.e., $x_{hi} = |\{(p, h_i) \in \phi\}|$ and μ being the average number of associations per host leaf. Note, that $\mu = scale$ since we assume that each parasite leaf is associated with exactly one host leaf.

To compare cophylogenies of different sizes *scale* and *var* are normalized to range from -1 to 1, respectively 0 to 1. For this purpose cut off values of 1/10 and 10 were defined for *scale* such that a value of *scale* that is 10 or larger is rated 1 and a value of *scale* which is 1/10 or below is rated -1. Furthermore, a *scale* value of 1, i.e., equal size host and parasite trees, should result in a normalized value of 0. The formal definition is given in Equation (7).

$$scale^* = \begin{cases} 1 & \text{if } scale > 10 \\ \frac{scale-1}{9} & \text{if } scale \geq 1 \wedge scale \leq 10 \\ \frac{-\frac{1}{scale}+1}{9} & \text{if } scale \geq \frac{1}{10} \wedge scale < 1 \\ -1 & \text{otherwise} \end{cases} \tag{7}$$

Accordingly a threshold of 10 is defined for *var* such that a variance of 10 or above results in a normalized value of 1. Equation (8) describes this normalization.

$$var^* = \begin{cases} 1 & \text{if } var > 10 \\ \frac{var}{10} & \text{otherwise} \end{cases} \tag{8}$$

A threshold of 10, respectively 1/10, was chosen, as this is the maximal, respectively minimal, value when considering cophylogeny systems of size 10, which are the smallest systems being analyzed in this study. Both normalizations result in a value of 0 in the best case, i.e., equal sized host and parasite trees, respectively equally distributed number of parasite associations. Conversely values of ± 1, respectively $+1$ indicates that a host parasite system is likely to be unrealistic.

To combine both measures $scale^*$ and var^* they are multiplicatively linked to obtain a *quality* value which is used as a measure of how likely a cophylogeny can be considered to be realistic. Formally, Equation (9) is defined by

$$quality = (1 - |scale^*|) * (1 - var^*) \tag{9}$$

Results and Discussion

In the following the space of parameter values for the cophylogeny generation method is analyzed in order to identify "good" sets of parameter values that lead to realistic cophylogenies. Then, an evaluation data set of cophylogenies is generated. This data set is used to evaluate the cophylogeny reconstructions that are delivered by the reconciliation tools TreeMap 3b, Jane 2.0, and CoRe-PA. The result of this evaluation is given at the end of this section.

Parameters Values

For the parameter evaluation, the modified ERM and the age model were used with the generation method to generate 100 cophylogenies for each combination of parameter values $s = \{10, 15, \ldots, 50\}$, $p_{hc} = \{0.0, 0.05, \ldots 0.95, 1.0\}$, $p_{co} = \{0.0, 0.05, \ldots 0.95, 1.0\}$, and $p_{sw} = \{0.0, 0.05, \ldots 0.3\}$. Only values up to 0.3 have been considered for p_{sw} because in typical biological host parasite systems it is much more likely for a parasite to remain on an associated host than to switch to another host. Moreover, a very high switching probability would mean that there is only a very loose relation between hosts and their parasites. Such systems are not so interesting to be analyzed with reconciliation tools. The cophylogenies generated with the different sets of parameter values have been evaluated with respect to $scale^*$, var^*, and *quality*. To analyze in more detail the influence of the system size cophylogenies have also been generated for size $s = 100$.

The *quality* of cophylogenies that have been generated with different combinations of parameters values p_{hc} and p_{co} and for different values of s are shown in Fig. 3. Recall that a set of cophylogenies may be considered to be more realistic if i) both trees are of similar size ($scale^* \approx 0$), and ii) every host is

associated with approximately the same number of parasites ($var^* \approx 0$). Cophylogenies where T_p equals T_h and each of the parasites is associated with the corresponding host can be generated using parameter values $p_{co} = 1$ and $p_{hc} = 1$. In this case no host independent events occur and there is always a cospeciation of the parasite whenever a host speciates. These perfect scenarios belong to the upper right corner of each of the *quality* plots given in Fig. 3.

It comes with no surprise that independently of all other parameter values a value for p_{hc} of at least 0.4 is needed to obtain equal size trees. Otherwise, there will be too few host speciations resulting in very small host trees. By increasing the probability of cospeciations p_{co} the parasite tree becomes larger. Hence p_{hc} must be increased simultaneously in order to obtain the same results. Surprisingly there is nearly no influence of the switching probability p_{sw} and the system size s on the ratio of both tree sizes. On the other hand, the variance of the associations varies strongly depending on the system size. In general it holds that the larger the system size s is, the higher the host choosing probability p_{hc} has to be in order to obtain a small variance. Additionally, if a higher probability of cospeciations p_{co} is chosen then smaller values of p_{hc} are possible for producing quite reasonable variances. If a higher switching probability p_{sw} is used, p_{hc} can be decreased further while retaining a small variance.

Figure 3 shows that the range of "good" parameter values strongly depends on the system size. With an increasing system size, the range of parameters leading to realistic cophylogenies shrinks to the upper right quarter of the plot. Thus for systems with 50 or more leaves, the probability p_{co} should be at least 0.4. Choosing smaller values for p_{co} is not recommended, when considering highly dependent host parasite systems. Additionally, p_{hc} should be greater than 0.7. Otherwise, the variance becomes large. Surprisingly, there is only a small influence of the switching probability p_{sw} such that the ranges of "good" values for p_{hc} and p_{co} can be larger. This means that any of the considered p_{sw} values can be chosen to produce realistic cophylogenies.

Evaluation of Reconciliation Tools

Reconciliation Methods Overview

In this section, the three software tools TreeMap 3b [21], Jane 2.0 [22] and CoRe-PA [23] are compared in terms of accuracy when reconstructing coevolutionary histories for the generated test data. TreeMap is probably the most common tool for computing reconciliations of host parasite systems. It is now in its 3rd major release and was rewritten completely in Java. Jane 2.0 and CoRe-PA are quite novel tools which offer several additional features. For instance, Jane 2.0 includes an advanced reconstruction viewer where

the user can browse easily through all possible reconstructions. CoRe-PA provides a graphical user interface for designing host parasite systems and is able to deal with non-binary species trees.
Although all three methods are based on the same coevolutionary event model, they differ in how the costs for each of the events are counted. This is due to the fact that in one approach the costs are counted per event while in the other they are counted per emerged sibling in the parasite tree. An overview on the different cost methods is given in Tab. 1.
All three approaches are based on the maximum parsimony principle. Given a certain cost vector (i.e., a cost value for each type of event) the tools search for the reconstruction which results in the minimum total cost. For that reason, the resulting reconstructions depend highly on the used cost model. Jane 2.0 uses costs $c = 0$, $d = 1$, $s = 2$ and $w = 1$ by default. But as in all three applications the cost model can also be user specified. TreeMap 3b and CoRe-PA offer more sophisticated methods to solve this issue. Since version 3b (build 1234), TreeMap uses a heuristic to find several reconstructions that are potentially optimal under a certain cost model. This set of so called Pareto optimal solutions may be huge and the reconstructions differ very much. So it is hard to decide for one of the reconstructions being the most likely. CoRe-PA also tries to find all these Pareto optimal solutions by applying multiple cost models. In addition every reconstruction is then rated by a value which indicates how good a reconstruction fitted to the appropriate cost model.
The cophylogeny reconstruction problem is NP-hard [24]. Therefore, all tools use heuristics for the optimization. Only TreeMap gives the opportunity to search for an exact solution. Depending on the size of the host parasite system the computation can be time and space intense so that only small instances can be solved in reasonable time. By default TreeMap 3b uses a heuristic, too. While CoRe-PA always finds an optimal solution, the reconstruction may be chronologically invalid, involving sets of inconsistent host switches. TreeMap 3b and Jane 2.0 always produce consistent though not necessarily optimal solutions.

Test Data Generation

To evaluate the reconciliation methods 1000 test data sets per branching model are computed. The sizes of the generated cophylogenies and the other parameter values are chosen randomly with a distribution proportional to the *quality* gathered from the parameter space evaluation discussed in the previous section. In this way it is ensured that each combination of parameters can be selected, but it is more likely that parameters are chosen that will result in cophylogenies that are similar to cophylogenies that occur in biological systems.

For each model we differentiated between the complete cophylogenies as they were generated and a pruned version. In this pruned cophylogenies host nodes are removed which have no assigned parasites. This was done due to the fact that most biological studies also disregarded hosts without associated parasites. So one might ask if this lack of information would have a measurable impact on the reconstructions.
This results in four test set-ups, one for each combination of ERM or age model with complete or pruned cophylogenies. But not each of the 1000 generated cophylogenies per set-up could be considered for reconstruction. Due to the wide range of possible parameter values combinations 7% to 24% of the datasets were cophylogenies with one of the trees having less than 3 nodes. These trivial instances were not included in the analysis. Very few cophylogenies could not be processed with TreeMap 3b resulting in an "out of memory" error. These cophylogenies were also removed. Altogether between 771 to 920 cophylogenies were used per test set-up.

Reconstruction Evaluation

The reconstructions computed with TreeMap 3b were done with the default heuristic trying to find different Pareto optimal solutions. Although in some rare cases more than 500 different reconstructions were found for a single data set, the dominant number of computations (around 60%) produces only three or less distinct reconstructions. The command line version of Jane 2.0 was used with its default cost model, producing exactly one reconstruction per data set. CoRe-PA was configured to evaluate 2500 different cost models and the best rated reconstructions were considered for the analysis. In most cases (more than 90%) CoRe-PA produced a single reconstruction. In the other cases up to seven different solutions were found, all having the same event distribution.
To measure the accuracy of each tool, the amount of correctly predicted host parasite associations were analyzed with respect to the generated cophylogenies. If more than one solution was found by one of the tools (TreeMap 3b or CoRe-PA), the average amount of correct hits was taken. As TreeMap 3b tries to find different Pareto optimal reconstructions the solution with the highest, respectively lowest, number of hits were analyzed additionally. But it should be noted that for a determination the best (or worst) of the solution knowledge about the exact history is necessary (which will not be available for the application to biological data). For normalization purposes, the fraction of the exact hits compared to the total number of associations - including false positives and false negatives - was used.
Figure 4 depicts the strip chart of the sorted fraction values with one dot for each data set and method. Only the results of the complete age model data set are shown. For the results of the pruned cophylogenies

and the ERM model we refer to the supplement, as these are quite similar.

CoRe-PA turns out to be the most precise method in this analysis. Depending on the used branching model it produces significantly more exact hits than Jane 2.0. It comes with no surprise that the average fraction of hits computed from the multiple solutions of TreeMap 3b is much lower. By considering multiple Pareto optimal solutions there are many reconstructions which differ very much from the corresponding generated cophyogeny. This obviously lowers the average fraction of hits. On the other hand one would assume that by considering only the most similar of these reconstructions the fraction of hits would be much better, especially because the solutions of Jane 2.0 and CoRe-PA are Pareto optimal too. This leads to the assumption that the heuristic used within TreeMap 3b misses a significant amount of Pareto optimal solutions, not reaching the results of Jane 2.0 and CoRe-PA.

Additionally the reconstructed events were analyzed. This was done by computing the difference between the number of reconstructed and generated events. The difference was normalized by division with the parasite tree size. It turns out that each method has its advantages and disadvantages. Using the default cost model Jane 2.0 results in a good estimation for the number of cospeciation events. But it underestimates the number of sortings and duplications and slightly overestimates the number of host switches. Both methods, TreeMap 3b and CoRe-PA, overestimate the number of cospeciations. Whereas TreeMap 3b overestimates the number of sortings and underestimates the number of duplications CoRe-PA is quite exact in predicting the total number of both types of events. On the other hand, CoRe-PA seems to produce to few host switches whereas TreeMap 3b tends to produce slightly too many of them. Fig. 5 shows boxplots of the deviations of the number of events for each type of event and application gathered from the complete set of cophylogenies of the age model data set.

By comparing the runtime of the three tools TreeMap 3b and Jane 2.0 perform quite similar on the test data with the complete phylogenetic trees, but Jane 2.0 is significantly faster on reconstructing the pruned test data set. On average TreeMap 3b needs around 3 to 15 times longer, but this was due to the fact that there were several instances where TreeMap 3b had exceptional long runtimes. CoRe-PA was around 40 to 100 times slower compared to Jane 2.0. But it should be noted that Jane 2.0 considers only a single cost model whereas CoRe-PA analyzes 2500 different cost models per computation. Fig. 6 shows boxplots of the runtimes for each application required for the reconstructions of the complete cophylogenies with the age model data set.

It is interesting that the different branching models seem to have only a small impact on the accuracy of the reconstructions. However, when considering pruned cophylogenies the deviation between the

reconstructed number and the original number of the host dependent events (cospeciations and sortings) becomes larger. This does not hold for duplications or host switches. Hence, for coevolutionary studies it might be useful to enrich the host parasite systems with data from host species without associated parasites to obtain more precise reconstructions.

Conclusions

In this work, a method for generating cophylogenies that describe the common evolution of two groups of species were presented. In particular, the case of cophylogenies that can describe the coevolution of hosts and their parasites have been considered. Existing branching models for creating phylogenetic trees have been combined with a coevolutionary event model considering cospeciation, duplication, lineage sorting, and host switching events. The influence of different parameters (e.g., the probabilities for different types of coevolutionary events) on the characteristics of the generated cophylogenies have been analyzed. It was shown which parameter values are relevant for generating cophylogenies that have similar properties to cophylogenies found in biological systems. Based on this analysis, different sets of cophylogenies have been generated, that can be used as test data for reconciliation tools. These data sets have been used to make the first systematic study to evaluate the common reconciliation tools TreeMap 3b, Jane 2.0, and CoRe-PA on test data.

The evaluation has shown that on the generated data sets CoRe-PA is the most precise of the three tools in predicting the correct host parasite associations. But CoRe-PA is not best in estimating the correct number of cospeciation and switching events. Furthermore, CoRe-PA is the computational most intense method. Jane 2.0 is best to estimate the correct number of cospeciations and is the fastest of the three tools. A disadvantage is that it always relies on a single user specified cost model. TreeMap 3b is the only tool which can be configured so that it always finds the optimal reconstruction for a specified cost model. Using the implemented heuristic it is much faster, but the accuracy of the computed reconstructions is not as good as that of the other tools. Additionally TreeMap 3b sometimes computes several hundred solutions making it hard to decide for the best without any further evaluation.

Authors contributions

All authors made substantial intellectual contributions to the published study. MM and KK initiated this study. SKS and NW conceived the study, developed and implemented the methods and wrote the draft of this paper. All authors improved the draft version and approved the final manuscript.

Acknowledgements

This work was supported by the German Research Foundation (DFG) through the project "Deep Metazoan Phylogeny" within SPP 1174 and by Volkswagen Stiftung through the initiative Complex Networks as a Phenomenon across Disciplines.

References

1. Yule GU: **A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S.** *Phil. Trans. R. Soc. B* 1925, **213**:21–87.
2. Blum MG, François O: **Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance**. *Systematic Biology* 2006, **55**:685–691.
3. Aldous D: **Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today**. *Statistical Science* 2001, **16**:23–34.
4. Charleston MA, Perkins SL: **Traversing the tangle: Algorithms and applications for cophylogenetic studies**. *Journal of Biomedical Informatics* 2006, **39**:62–71.
5. Doyon JP, Scornavacca C, Gorbunov KY, Szöllosi GJ, Ranwez V, Berry V: **An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers.** In *RECOMB-CG'10* 2010:93–108.
6. Merkle D, Middendorf M: **Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information**. *Theory in Biosciences* 2005, **123**(4):277–299.
7. Merkle D, Middendorf M, Wieseke N: **A parameter-adaptive dynamic programming approach for inferring cophylogenies**. *BMC Bioinformatics* 2010, **11**(Suppl 1):S60.
8. Ronquist F: **Three-dimensional cost-matrix optimization and maximum cospeciation**. *Cladistics* 1998, **14**(2):167–172.
9. Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R: **Jane: A new tool for the cophylogeny reconstruction problem**. *Algorithms for Molecular Biology* 2010, **5**:16.
10. Keller-Schmidt S, Tuğrul M, Eguíluz VM, Hernández-García E, Klemm K: **An age dependent branching model for macroevolution**. *arXiv:1012.3298v1* submitted.
11. Jackson AP, Charleston MA: **A cophylogenetic perspective of RNA–Virus evolution**. *Molecular Biology and Evolution* 2004, **21**:45–57.
12. Aldous D: **Probability distributions on cladograms**. In *Random Discrete Structures*. Edited by Aldous D, Pemantle R, Springer 1996:1–18.
13. Charleston MA: **Jungles: A new solution to the host/parasite phylogeny reconciliation problem**. *Mathematical Biosciences* 1998, **149**(2):191–223.
14. Hafner MS, Nadler SA: **Phylogenetic trees support the coevolution of parasites and their hosts**. *Nature* 1988, **332**:258–259.
15. Kikuchi Y, Hosokawa T, Nikoh N, Meng XY, Kamagata Y, Fukatsu T: **Host-symbiont co-speciation and reductive genome evolution in gut symbiotic bacteria of acanthosomatid stinkbugs**. *BMC Biology* 2009, **7**:2.
16. Refrégier G, Gac ML, Jabbour F, Widmer A, Shykoff JA, Yockteng R, Hood ME, Giraud T: **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* 2008, **8**:100.
17. Reed DL, Light JE, Allen JM, Kirchman JJ: **Pair of lice lost or parasites regained: The evolutionary history of anthropoid primate lice.** *BMC Biology* 2007, **7**:5–7.
18. Hughes J, Kennedy M, Johnson KP, Palma RL, Page RDM: **Multiple cophylogenetic analyses reveal frequent cospeciation between pelecaniform birds and pectinopygus lice**. *Systematic Biology* 2007, **56**(2):232–251.

19. Banks JC, Palma RL, Paterson AM: **Cophylogenetic relationships between penguins and their chewing lice.** *Journal of Evolutionary Biology* 2006, **19**:156–166.
20. Ramsden C, Holmes EC, Charleston MA: **Hantavirus evolution in relation to its rodent and insectivore hosts: No evidence for codivergence.** *Molecular Biology and Evolution* 2009, **26**:143–153.
21. Charleston MA: **TreeMap 3b** 2011, [http://sites.google.com/site/cophylogeny].
22. Libeskind-Hadas R: **Jane 2.0** 2010, [http://www.cs.hmc.edu/~hadas/jane].
23. Wieseke N, Merkle D, Middendorf M: **CoRe-PA** 2010, [http://pacosy.informatik.uni-leipzig.de/core-pa].
24. Ovadia Y, Fielder D, Conow C, Libeskind-Hadas R: **The cophylogeny reconstruction problem is NP-complete**. *Journal of Computational Biology* 2011, **18**:59–65.

Figures

Figure 1 - Example for a coevolutionary system and a corresponding reconstruction

Left: Example for a small coevolutionary system with four extant host species (leaf nodes in dark grey tree) and four extant parasite species (leaf nodes in light grey tree). Right: Example of a cophylogenetic reconstruction for the coevolutionary system. The three associations (p_5, h_5), (p_6, h_6) and (p_4, h_0) induce one cospeciation and one sorting event. The three associations (p_1, h_2), (p_4, h_0), and (p_0, h_0) induce one duplication and two sorting events. The reconstruction needs two cospeciations, one duplication, and three sortings.

Figure 2 - Coevolutionary Events

Host tree T_h (dark gray), parasite tree T_p (light gray); (a) *cospeciation*: node of T_h and T_p associated; (b) *sorting*; (c) *duplication*: both child nodes of T_p are associated with a node in the subtree of T_h; (d) *host switch*: only one child node of T_p is associated with a node in the subtree of T_h.

Figure 3 - Mean *quality* of 100 generated cophylogenies

Mean *quality* of 100 generated cophylogenies per parameter combination of p_{hc} and p_{co} for a tree sizes $s = (25, 50, 100)$ (left to right) and $p_{sw} = (0.1, 0.3)$ (top to bottom) for the age model. (See supplements for more results of further parameter combinations considering the age model as well as the ERM model.)

Figure 4 - Sorted fraction of exact predicted host parasite association

Sorted fraction of exact predicted host parasite association for each tool for the complete cophylogenies with age model data set.

Figure 5 - Normalized deviation between number of generated and reconstructed events

Normalized deviation between number of generated and reconstructed events for the complete cophylogenies with age model data set. The average deviation per event and tool is depicted in brackets in the x-axis.

Figure 6 - Runtimes of the three tools

Runtimes of the three tools for the complete cophylogenies with age model data set. The y-axis is in log scale. The average runtime per tool is depicted in brackets in the x-axis.

Tables

Table 1 - Different methods of costs for each reconciliation method

	Cospec.	Dupl.	Sorting	Switch
TreeMap 3b	$2c$	$2d$	s	$d + w$
Jane 2.0[1]	$2c$	$2d$	s	$2d + w$
CoRe-PA	c	d	s	w

Table 1: Different methods of costs assignments per event for the reconciliation tools considering specified cost values c, d, s and w

Additional Files

Additional file 1 — Data sets

The generated evaluation data sets of all test set-ups can be downloaded from

http://pacosy.informatik.uni-leipzig.de/files/19/datasets.tar.gz

Additional file 2 — Results and reconstructions

All computed reconstructions, including event distributions and runtimes of all tools as well as the raw graphics of the parameter space analysis can be downloaded from

http://pacosy.informatik.uni-leipzig.de/files/19/rawdata.tar.gz

[1]The cost method used by Jane 2.0 is the same as the one that was used in former versions of TreeMap; Jane 2.0 can also be configured to count the costs in the same way as CoRe-PA does.

Additional file 3 — Additional analysis

The analysis of the additional test set-ups for ERM/Age model and complete/pruned data sets can be downloaded from http://pacosy.informatik.uni-leipzig.de/files/19/supplement.pdf

Additional file 4 — CoRe-Gen

The command line based tool CoRe-Gen for generating test data sets can be downloaded from http://pacosy.informatik.uni-leipzig.de/files/19/core-gen.tar.gz

The tool is able to generate different formattings, .nex files which can be processed by Jane 2.0 and CoRe-PA and .tree files which is supported by TreeMap 3b and Jane 2.0 as well.

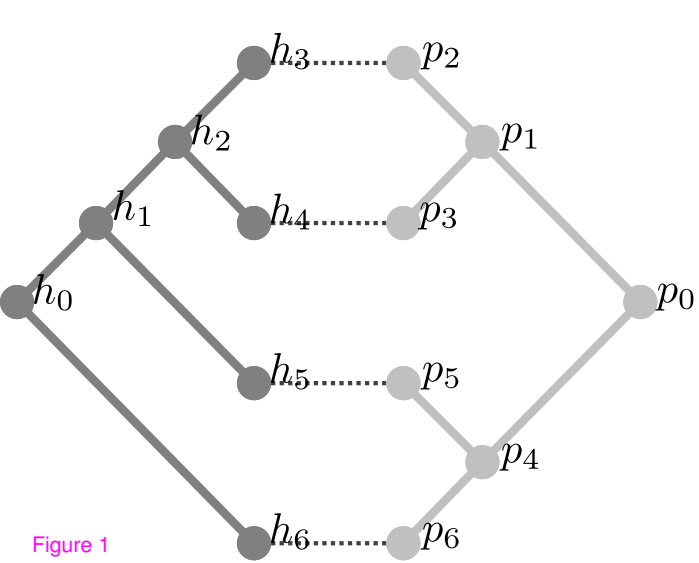
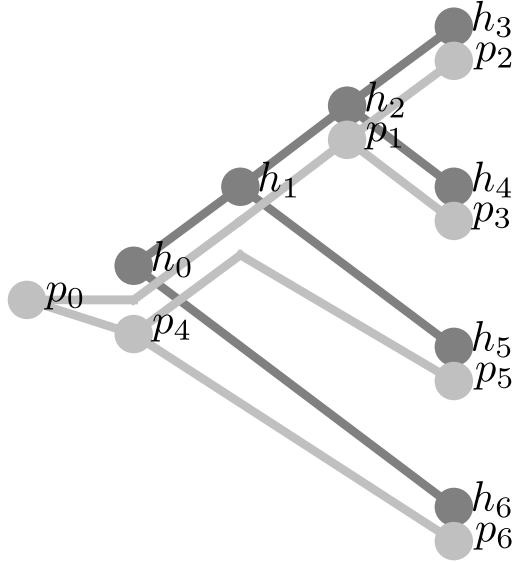
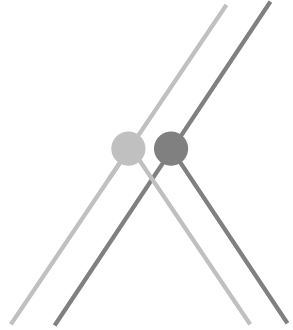
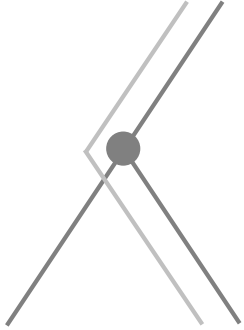

h_3
p_2
h_2
p_1
h_1
h_4
p_3
h_0
p_0
h_5
p_5
p_4
h_6
p_6

Figure 1

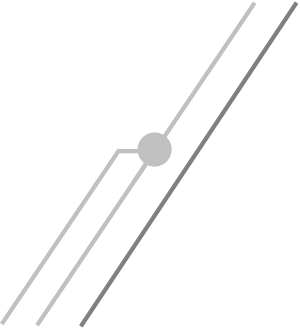

h_3
p_2
h_2
p_1
h_1
h_4
p_3
h_0
p_0
p_4
h_5
p_5
h_6
p_6



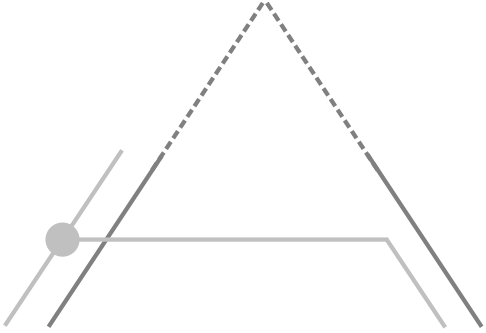
(a)



(b)



(c)



(d)

Figure 2

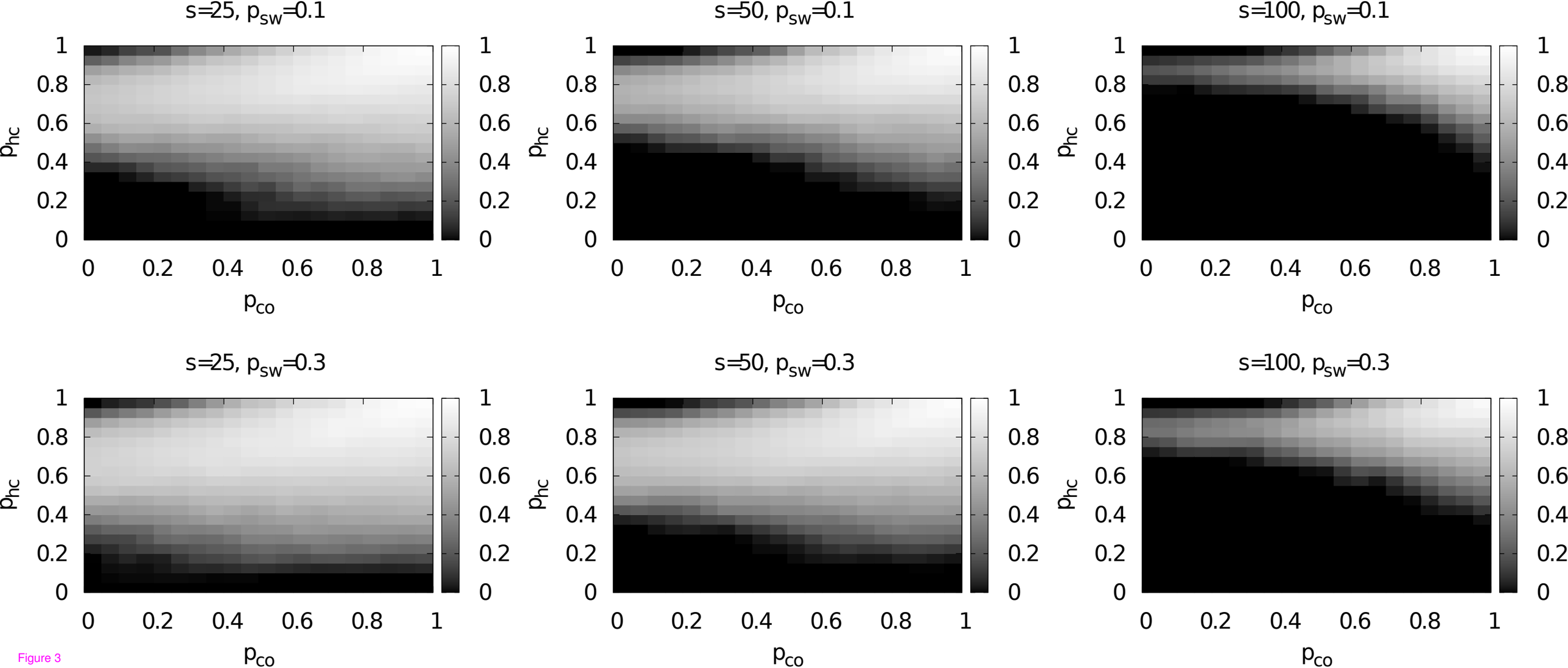

s=25, p_sw=0.1
s=50, p_sw=0.1
s=100, p_sw=0.1
s=25, p_sw=0.3
s=50, p_sw=0.3
s=100, p_sw=0.3
p_hc
p_co
0
0.2
0.4
0.6
0.8
1

Figure 3

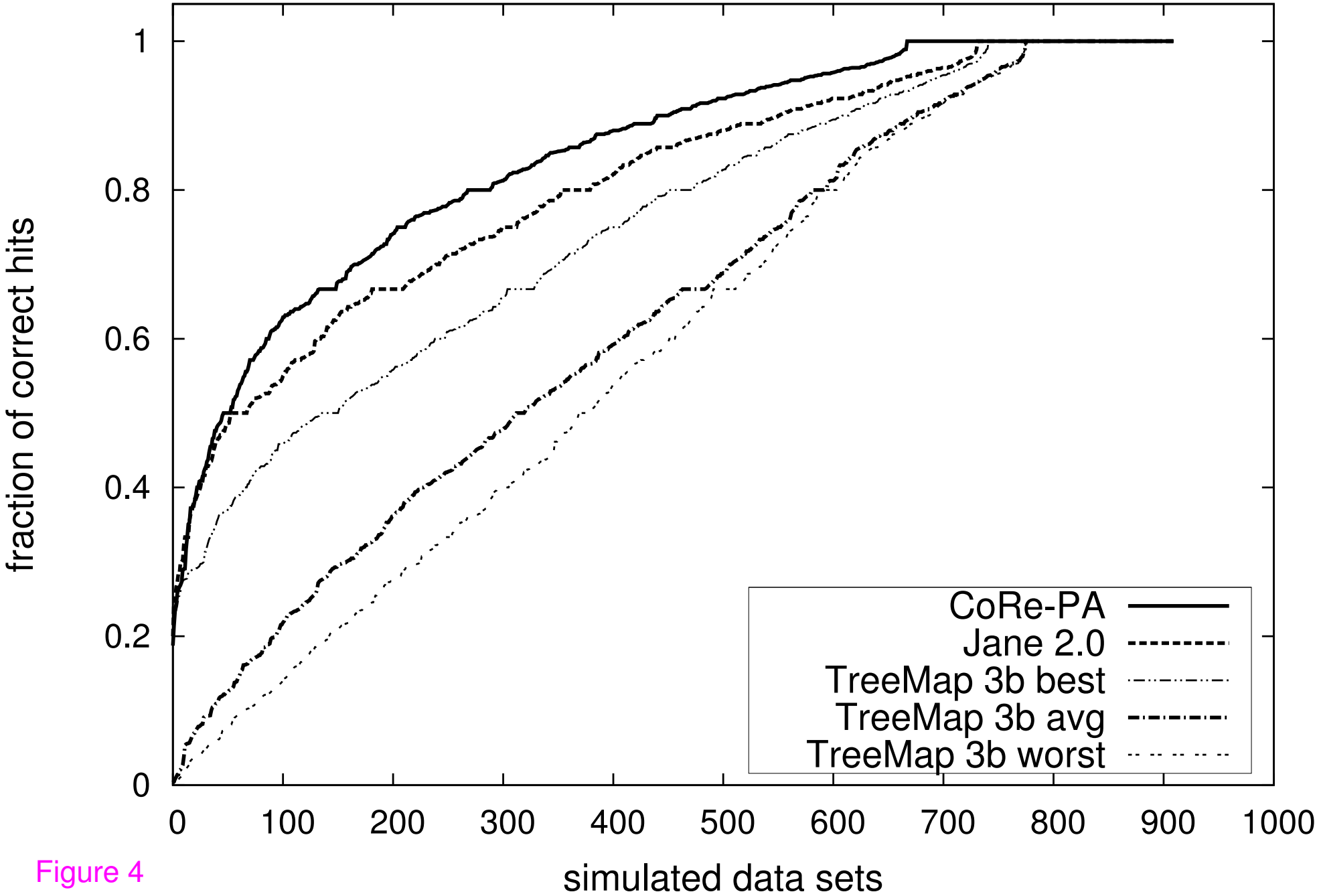
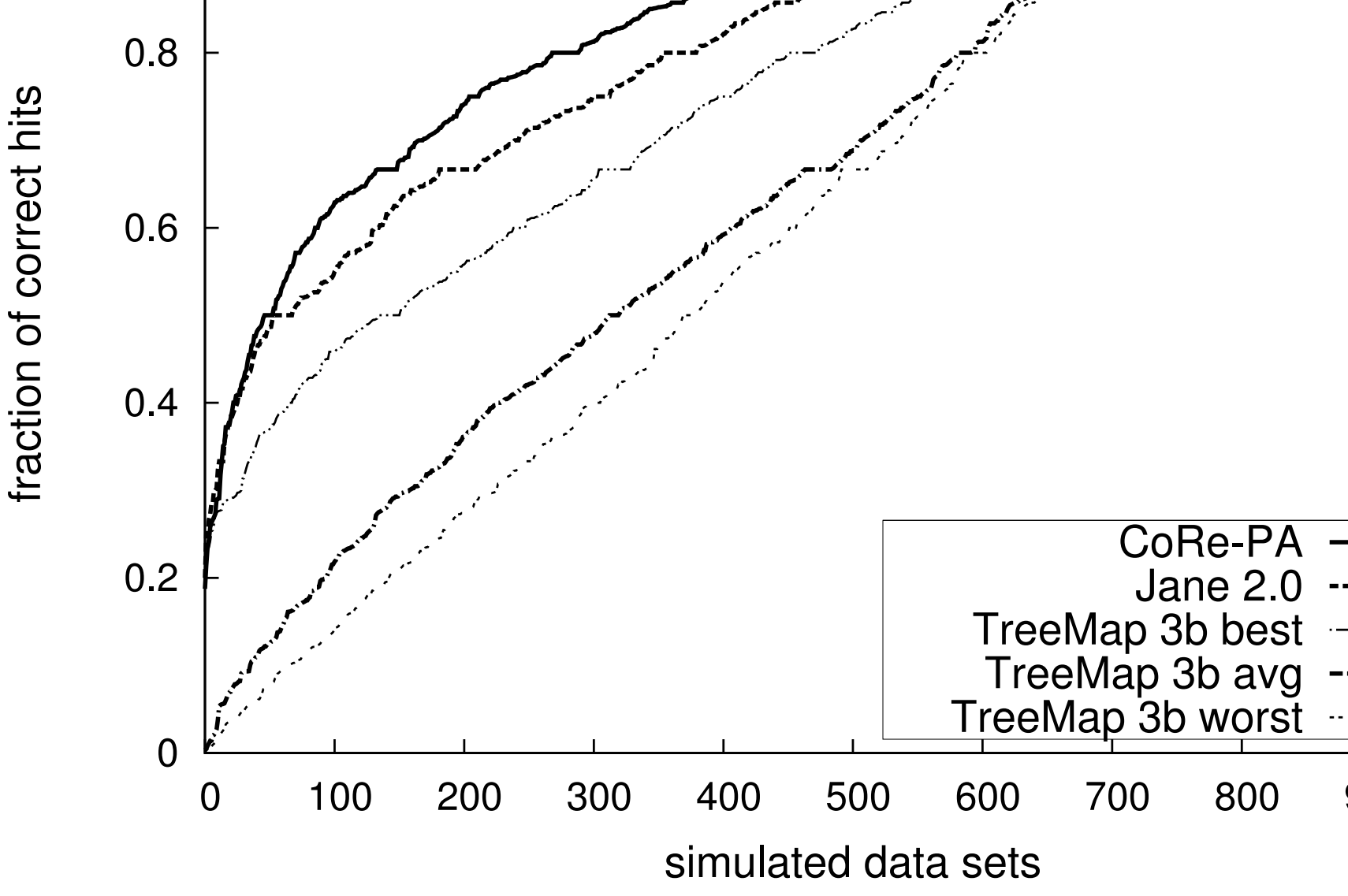

fraction of correct hits
simulated data sets
0
0.2
0.4
0.6
0.8
1
0
100
200
300
400
500
600
700
800
900
1000
CoRe-PA
Jane 2.0
TreeMap 3b best
TreeMap 3b avg
TreeMap 3b worst

Figure 4


fraction of correct hits
0
0.2
0.4
0.6
0.8
0
100
200
300
400
500
600
700
800
simulated data sets
CoRe-PA
Jane 2.0
TreeMap 3b best
TreeMap 3b avg
TreeMap 3b worst

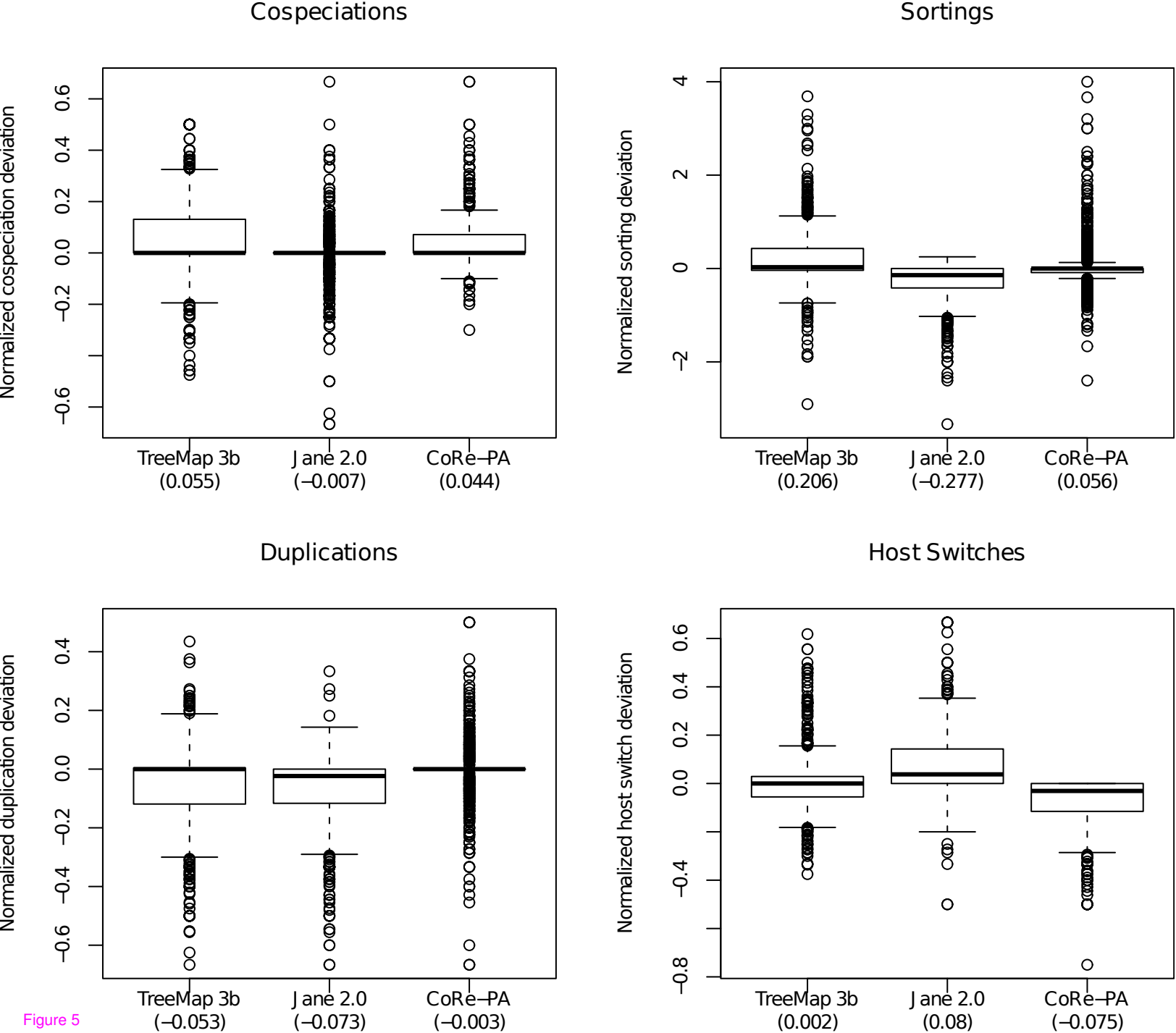

Cospeciations
Normalized cospeciation deviation
0.6
0.4
0.2
0.0
−0.2
−0.6
TreeMap 3b
(0.055)
Jane 2.0
(−0.007)
CoRe−PA
(0.044)
Sortings
Normalized sorting deviation
4
2
0
−2
TreeMap 3b
(0.206)
Jane 2.0
(−0.277)
CoRe−PA
(0.056)
Duplications
Normalized duplication deviation
0.4
0.2
0.0
−0.2
−0.4
−0.6
TreeMap 3b
(−0.053)
Jane 2.0
(−0.073)
CoRe−PA
(−0.003)
Host Switches
Normalized host switch deviation
0.6
0.4
0.2
0.0
−0.4
−0.8
TreeMap 3b
(0.002)
Jane 2.0
(0.08)
CoRe−PA
(−0.075)

Figure 5

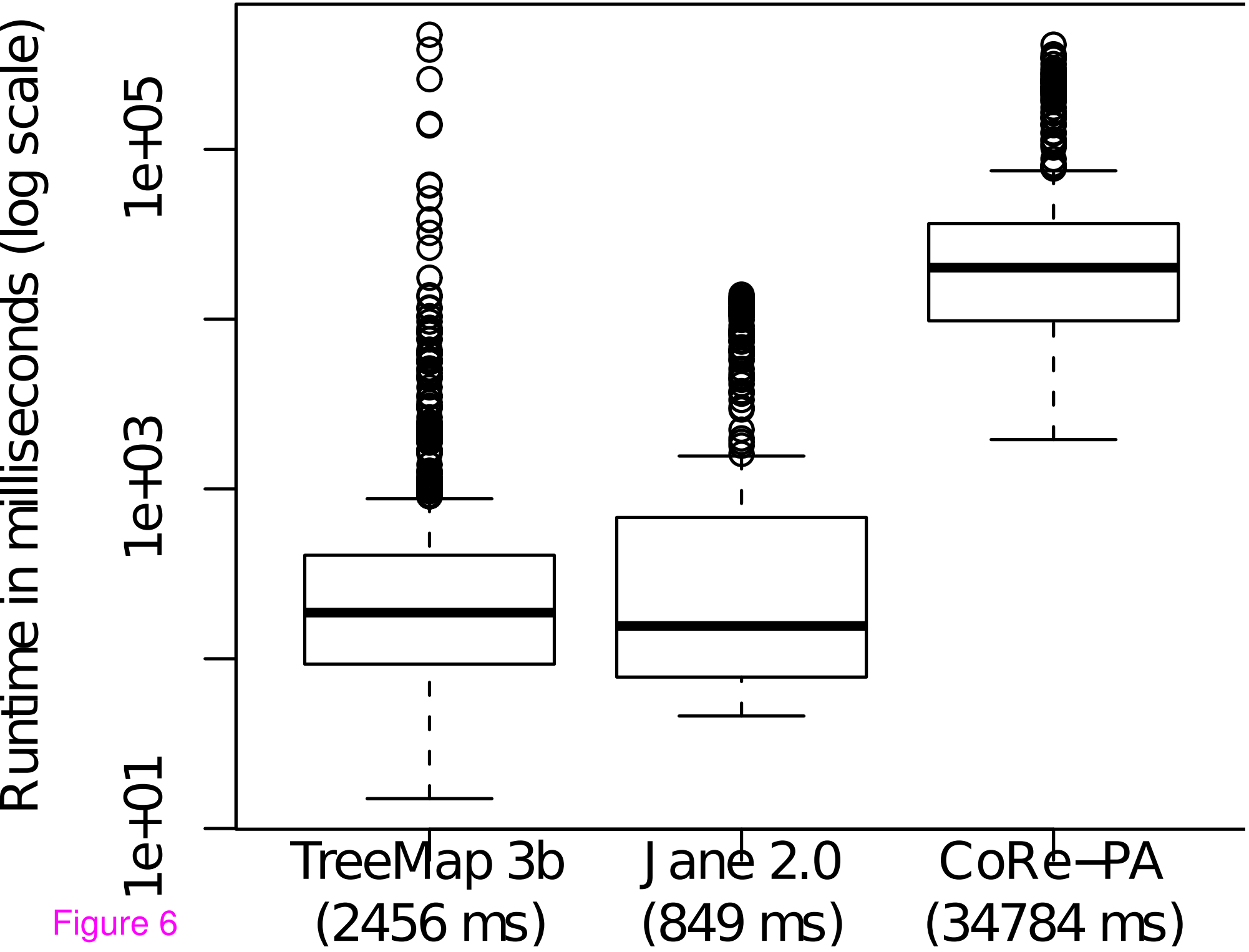

Runtime in milliseconds (log scale)
1e+01
1e+03
1e+05
TreeMap 3b
(2456 ms)
Jane 2.0
(849 ms)
CoRe–PA
(34784 ms)

Figure 6

Additional files provided with this submission:

Additional file 1: datasets.tar.gz, 2985K
http://www.almob.org/imedia/5480236357327762/supp1.gz
Additional file 2: rawdata.tar.gz, 7390K
http://www.almob.org/imedia/1777791672573277/supp2.gz
Additional file 3: supplement.pdf, 10931K
http://www.almob.org/imedia/3626712135733067/supp3.pdf
Additional file 4: core-gen.tar.gz, 164K
http://www.almob.org/imedia/2770868605733068/supp4.gz