Jamie Stevens
is a lecturer in parasitology and evolution in the Department of Biological Sciences, University of Exeter. His research interests in molecular evolutionary parasitology range from human trypanosomiasis to the evolution of myiasis in ectoparasitic blowflies.

Computational aspects of host–parasite phylogenies

Jamie Stevens

Date received (in revised form): 14th June 2004

Abstract

Computational aspects of host–parasite phylogenies form part of a set of general associations between areas and organisms, hosts and parasites, and species and genes. The problem is not new and the commonalities of exploring vicariance biogeography (organisms tracking areas) and host–parasite co-speciation (parasites tracking hosts) have been recognised for some time. Methods for comparing host–parasite phylogenies are now well established and fall within two basic categories defined in terms of the way the data are interpreted in relation to the comparison of host–parasite phylogenies, so-called *a posteriori*, eg Brooks' Parsimony Analysis (BPA), or *a priori*, eg reconciled trees and other model-based methods, as implemented in the program TreeMap; the relative merits of the two philosophies inherent in these two approaches remain hotly debated. This paper reviews the computational methods currently available to analyse host–parasite relationships.

Keywords: *co-evolution, co-speciation, host–parasite, tracking, Brook's parsimony analysis, TreeMap*

INTRODUCTION

Computational aspects of host–parasite phylogenies form part of a set of general associations between: areas and organisms, hosts and parasites, and species and genes. Within each of these associations one lineage is associated with another, and can be thought of as tracking the other over evolutionary time with a greater or lesser degree of fidelity.[1] The problem is not new and certainly the commonalities of exploring vicariance biogeography (organisms tracking areas) and host–parasite co-speciation (parasites tracking hosts) have been recognised for some time.[2,3]

The evolutionary history of any organism must be considered in relation to its environment and the selective and evolutionary forces within such an environment; in the case of a parasite, where its phylogeny is often intimately linked to that of its host(s) and where host-switching, 'arms race' interactions and resource tracking may have acted to affect parasite evolution over time, this requirement is perhaps even more poignant. Consequently, methods for comparing host–parasite phylogenies are now well established and fall within two basic categories defined in terms of the way the data are interpreted in relation to the comparison of host–parasite phylogenies: *a posteriori*, eg Brooks' parsimony analysis (BPA),[3,4] or *a priori*, eg reconciled trees and other model-based methods, as implemented in the programs COMPONENT[5,6] and latterly, TreeMap.[7] As outlined below, the relative merits of the two philosophies inherent in these two approaches remain hotly (sometimes very hotly) debated.[8–12]

In this paper the methods available to analyse host–parasite relationships will be reviewed, focusing in particular on currently available computational resources. However, given the ongoing debate concerning the philosophy underlying available methods and arguments concerning the abilities of such methods to deal with different evolutionary scenarios, it is left to the reader and would-be user to decide which method is most suitable for analysing their own data.

Jamie R. Stevens,
Department of Biological Sciences,
University of Exeter,
Prince of Wales Road,
Exeter EX4 4PS, UK

Tel: +44 (0) 1392 263775
Fax: +44 (0) 1392 263700
E-mail: j.r.stevens@ex.ac.uk

PHILOSOPHIES UNDERPINNING METHODS FOR STUDYING HOST–PARASITE PHYLOGENIES

***a posteriori* versus *a priori* approaches**

As noted, methodologies for studying host–parasite phylogenies fall into two main schools of thought. Those advocating an *a posteriori* interpretation of parasite phylogenies without imposition of a predefined model of host–parasite associations have tended to advocate the use of BPA, while those who subscribe to the idea of predefining models of parasite evolution against which to assess a range of possible evolutionary scenarios have tended to adopt *a priori*, model-based methods. The underlying philosophies of these two schools of thought will now be reviewed.

Brooks parsimony analysis

No underlying model

BPA is based on the assumption that there need not be any model-like regularities in phylogenesis and that (co)evolutionary processes are so contingent on history that no *a priori* model will be sufficient for capturing all the relevant detail.[13] Unlike model-based methods – its proponents suggest – it does not seek to maximise fit to any predetermined hypothesis, but instead offers a framework to ask such questions as: How many co-speciation patterns that do exist are due to mutual modification leading to mutual speciation, and how many are simply by-products of vicariant speciation? Significantly, BPA is designed to assess co-speciation among multiple parasite clades in the context of their hosts, but without specifying a host phylogeny *a priori*.[3] Indeed, Brooks[3] developed BPA specifically to overcome some of the obstacles pointed out by Hennig[14] concerning the non-random association between hosts and parasites, advancing the idea that if multiple parasite clades are analysed simultaneously with respect to their hosts, co-speciation patterns can be inferred from phylogenetic congruence among portions of the parasite phylogenies and host-switching can be inferred from incongruence;[3,10] indeed, it is a characteristic of BPA that incongruence implies host-switching. Subsequently, the host cladogram produced by such multi-clade analysis could be tested for congruence with a host phylogeny generated using independent (non-parasite) data.

Vicariance biogeography

Brooks further suggested[3] that co-speciation between hosts and parasites will often be the by-product of vicariant speciation affecting host and parasite lineages simultaneously. Perhaps not surprisingly, BPA has also been also widely used in the field of biogeography and it still remains in widespread use within the discipline.

Reconciled trees, maximal co-speciation and event-cost methods

The alternative approach to BPA is to map one tree (that of the parasite) on to another tree (that of the host); a range of event–cost model-based methods designed to maximise co-speciation[15,16] can then be used to reconcile differences in the patterns between the host and parasite evolutionary trees. The philosophy of the maximum co-speciation approach is defined by Page[17] as follows:

> Co-speciation is joint cladogenesis of host and parasite. If we regard host cladogenesis as the primary cause of cladogenesis in the parasite, then the host phylogeny is the 'independent' variable and the parasite phylogeny is the 'dependent' variable. The host phylogeny explains the parasite phylogeny to the extent that speciation events in the parasite phylogeny are co-speciations. Hence a natural criterion for choosing a reconstruction is maximising the extent of co-speciation, that is, the ability of the host phylogeny to explain the parasite phylogeny.

The first methods for mapping one tree onto another explained any incongruence between the two trees by invoking the presence of unrecognised multiple

Reconciled trees

lineages in one of the trees.[15,18] Goodman *et al.*[15] developed their method to reconcile mammal phylogenies derived from protein sequence data with morphology-based trees and suggested that the incongruence could be due to some of their protein sequences being paralogous rather than orthologous, hence confounding the history of genes with the history of species. Independently, Nelson and Platnick[18] – working from a biogeographical perspective – suggested ways in which the effects of poor taxon sampling and extinction could lead to incongruence between area cladograms for different taxa. They also proposed that in the presence of two or more sympatric lineages of taxa, poor sampling could obscure the underlying area relationships in the same way that sampling paralogous genes may give a confused picture of species relationships.

Early, diverse approaches

Computer methods

Computer methods incorporating the ideas of reconciled trees and maximal co-speciation have been developed since the early 1990s, successively incorporating the ability to deal with associated issues, such as host-switching (eg COMPONENT[5,6] and TreeMap[17]). However, in part because of the optimality criteria used in some of the earlier programs (TreeMap version 1 scored each reconstruction solely by the number of co-speciation events, which will range from 1 to $n - 1$, where n is the number of parasites[9]), it was apparent that there could be many reconstructions implying the same number of co-speciation events, thus yielding multiple solutions. This then left users to trawl through large numbers of reconstructions to find the most appropriate reconstruction, using the numbers of duplications, host switches and sorting events to help choose among these reconstructions.

Cost–event-based methods

This problem has now been addressed by the development and inclusion of cost–event-based analyses[19–21] which consider and evaluate each hypothesised past association individually, to find the least costly solution (eg TreeFitter[20] and Jungles[21] in TreeMap 2.02). Indeed, the approach is exemplified by the most recent implementation of the program TreeMap 2.02[7] which attempts to explain observed relationships by producing a set of solutions that range from those that maximise co-speciation to those that include a minimum of co-speciation enforced by logical consistency – ie TreeMap 2 does not simply use co-divergence alone as an optimising criterion for evaluating solutions. Rather, solutions are based on an *a priori* model which allows the user to define event–cost assignments for a range of different potential events, eg extinctions, lineage duplications, host switching, to explain the observed pattern of host–parasite relationships (of course, placing costs on events has no direct bearing on the solutions returned, unless bounds are also placed on the total cost or on the number of specific events, such that solutions beyond these bounds are then excluded – see below for methodological details).

Gene phylogenies *v.* species phylogenies: A further complication

Early approaches to the study of host–parasite co-evolution relied on the construction and interpretation of morphology-based phylogenies, using methods influenced by the allied discipline of biogeography. However, since the classic study of Hafner and Nadler,[22] molecular data have become increasingly more widely used in the study of host–parasite co-evolution. Significantly, molecular data have provided the opportunity to test co-speciation using genetic markers evolving by, theoretically, the same processes in both parasite and host, ie sequences from homologous genes, or genes coding for interacting products, ie those involved in an 'arms race' system; however, as as been shown in lice,[23] mitochondrial DNA evolves considerably faster than vertebrate host DNA and has different substitution characteristics, which are shared with other insects. Indeed, based on what is known about molecular clocks and the

Molecular clocks

Models of DNA evolution

Phylogenetic congruence

Limitations on gene phylogenies

variability in clock speeds between different genes, even within a single species,[24–26] the use of homologous gene sequences or 'interacting' genes is essential. Thus, when comparing molecular phylogenies of a single parasite gene with a phylogeny derived from a single host gene, the potential for a lack of parity between the single gene trees and the species trees must not be overlooked. If molecular host–parasite phylogenies are not congruent, it may be due to the particular evolutionary history of one or other (or both) of the two gene phylogenies, and not necessarily to a lack of congruence in species phylogenies. Of course, lack of congruence could be due to genuine host–parasite evolutionary incongruence because of host switching or other evolutionary scenarios such as speciation independent of host, parasite extinction, non-colonisation of all host lineages or failure to speciate with host.[11]

Ultimately, since it is probably impossible to account for all factors, it should be borne in mind that single gene phylogenies have known limitations and rarely equate perfectly to overall species phylogeny, and that host–parasite systems might thus be thought of as incorporating limitations from multiple sources, ie from both the host and the parasite sides of the system under study. It remains to be explored whether this means that those host–parasite phylogenies that are seen to be congruent should be viewed with increased significance (however this may be defined), or whether the use of homologous genes in both parasite and host, which may (or may not) be under similar biological constraints in both, means an *a priori* increased likelihood of evolutionary congruence. The answer to this last particular issue may not be known until patterns of co-evolution have been evaluated for a broad range of host–parasite systems with a range of ecologies, based on a broad range of homologous and non-homologous molecular markers.

Moreover, as noted, while early approaches to the study of host–parasite co-evolution relied on the interpretation of morphology-based trees, molecular data have now become the main form of data on which new phylogenic reconstructions are based. And, although it is difficult to formulate models that accurately reflect the cost of the loss or gain of morphological characters, we do now have robust, accurate models of DNA sequence evolution that can usefully be incorporated into phylogenetic analyses. There can be no doubt that the ability to formulate, test and re-formulate such models offers a major new tool to the hypothesis-driven approach. The fact that BPA is not a model-based method means that it is an excellent discovery-based method, useful for uncovering patterns in nature that can be used to evaluate different models;[13] whether in the long run this offers enough scope for its development within the field, remains to be seen.

AVAILABLE COMPUTATIONAL RESOURCES

The development of methodologies for exploring relationships between host–parasite phylogenies, has proceeded since the late 1970s first by *a posteriori*, non-model-based interpretations of parasite phylogenies (BPA), through reconciled tree methods (COMPONENT) successively incorporating the ability to deal with host-switching (TreeMap) to cost–event-based search systems which consider and evaluate each hypothesised past association individually, to find the least-costly solution (Jungles, in TreeMap 2.02). See Dowling *et al.*[10] for chronological details of conceptual and methodological developments in comparative studies of host–parasite associations. A summary of available computational resources is provided in Table 1.

Brooks' parsimony analysis

BPA converts the associate (parasite tree) into a set of additive binary characters and then maps them onto the host tree by parsimony. Specifically, half + 1 of the

Table 1: Available computational resources

Brooks' parsimony analysis (BPA) BPA is based on the assumption that there need not be any model-like regularities in phylogenesis[3,4] and, unlike model-based methods, does not seek to maximise fit to a predetermined hypothesis. BPA converts the parasite tree into a set of additive binary characters and then maps them onto the host tree by parsimony. A host phylogeny is now constructed based on the binary codes representing the phylogenetic relationships of the parasite taxa. The resulting cladogram can then be reviewed in the light of independent evidence, eg independently constructed host relationships. A suitable measure to identify homoplasy within the data can then be calculated, eg the consistency index (CI), and used to identify problematic relationships. BPA can be implemented using standard parsimony programs, eg PHYLIP[27] or PAUP,[28] for matrix analysis. A new algorithm, PACT (Phylogenetic Analysis for Comparing trees), which will replace existing BPA-based methods, is undergoing development and is scheduled for general release in 2005 (D. R. Brooks, personal communication).
Statistical tests of congruence/incongruence Tests to perform assessments of congruence/incongruence between tree topologies may be used to explore relationships between parasite and host phylogenies. These include the Kishino–Hasegawa (K-H) test[29] and the incongruence-length difference (ILD) test;[30] both are implemented in the current version of PAUP.[29] Statistical tests developed specifically for assessing congruence between parasite and host phylogenies are also available. Huelsenbeck *et al.*[31] proposed two tests to examine the null hypothesis that host and parasite trees are identical, based on phylogenetic estimates obtained using either maximum likelihood or maximum posterior probability (i.e. Bayesian inference). The ML approach uses a likelihood ratio test to examine the null hypothesis, H_0, that the host and parasite trees are identical, H_1 being that the host and parasite trees are not identical. Bayesian inference is used to calculate the posterior probabilities of host and parasite phylogenies allowing, in turn, calculation of the probability that the host and parasite trees are identical.[32] The probability of each individual phylogeny is calculated by Bayesian inference using an appropriate program, eg MrBayes.[33] ParaFit, developed by Legendre *et al.*,[34] is a matrix permutation test of co-speciation, which aims to test the significance of a global hypothesis of coevolution between parasites and hosts. Test statistics are functions of the host and parasite phylogenetic trees and of the set of host–parasite association links. ParaFit is available from http://www.fas.umontreal.ca/biol/legendre/.
Reconciled tree methods COMPONENT implements a variety of tree comparison methods. Comparisons can be made between two individual trees, a set of trees, or two sets of trees in different input files. The tree-mapping ability allows the computation of a tree reconciling incongruent parasite and host trees,[5,6] and is of particular value for the exploration of host–parasite relationships. COMPONENT does not allow host switching, but relies on the calculation of either duplications and losses, or 'items of error'[35] to reconcile incongruent trees. COMPONENT has now been largely superseded by TreeMap 2.02; however, it retains a number of useful features and runs in Microsoft Windows, so it will be described here. COMPONENT uses the NEXUS[36] format and is compatible with PAUP[28] and MacClade,[37] and tree files produced by these programs (together with those produced with PHYLIP[27]) can be read directly into it. See the COMPONENT 2.0 manual[6] for full instructions; the program is available at http://taxonomy.zoology.gla.ac.uk/rod/cpw.html. TreeMap 2.02 – see **'Event–cost methods'** below.
Event–cost methods TreeFitter is a simple program for parsimony-based event-cost tree fitting; it is available for both Macintosh and Windows. It can handle arbitrary cost assignments, such that duplication events, sorting events and switches all have zero or a positive cost associated with them. Treefitter uses the NEXUS format,[36] is compatible with PAUP and MacClade and is available at www.ebc.uu.se/systzoo/research/treefitter/treefitter.html. TreeMap is another program using the reconciled tree approach and is a 'direct descendant' of COMPONENT, the key difference being the ability to incorporate host-switching as an explanation of the observed pattern of host–parasite associations. Its current implementation is TreeMap version 2.02, a program that also provides an option to implement the Jungles[21] event–cost method to find all potentially optimal solutions to explain observed patterns of host–parasite association. In TreeMap 2, the Jungles algorithm is used to search for all feasible reconstructions within bounds set by the user. The user can specify the maximum number of host switches that any reconstruction can have; the program then filters the solutions to remove any that are definitely non-optimal for the given set of costs. Jungles uses four parameters to calculate the overall cost of each hypothesised past association individually; these are: co-speciation; duplication; lineage sorting; and host switching, and the user is prompted to enter values for the 'cost' of each. TreeMap 2.02 is available at http://evolve.zoo.ox.ac.uk/software/TreeMap/main.html.

BPA details

characters are used to record the presence of the actual parasite taxa, ie the terminal branches within the parasite tree, while the remainder are scored such that each binary character represents an internal node of the parasite tree (Figure 1). The resulting character state matrix thus defines each taxon by a binary code that defines all the nodes leading to it in the parasite phylogeny (Table 2); each species of parasite now has a code that indicates its identity and its common ancestry, eg the code for parasite B equates to 2, 7, 9.

A host (or area) phylogeny is now constructed based on the binary codes representing the phylogenetic relationships of the parasite taxa under study. The resulting cladogram can then be reviewed in the light of independent evidence, eg independently constructed

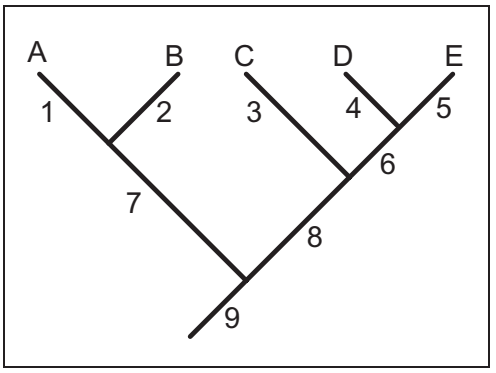

Figure 1: Phylogenetic tree for five species of parasite, with internal branches numbered for Brooks Parsimony Analysis; see also Table 2, the corresponding binary character matrix

Table 2: Binary character matrix describing the relationships between the five parasite taxa shown in Figure 1

Parasite	Binary character code
Parasite A	100000101
Parasite B	010000101
Parasite C	001000011
Parasite D	000101011
Parasite E	000011011

host relationships, geological or biogeographical evidence. A suitable measure to identify homoplasy within the data can then be calculated, eg the consistency index (CI), and used to identify problematic relationships. The problem now is to find an explanation for such homoplasy; in BPA homoplasy is explained by the process of host switching. BPA prohibits any modification of the input data and any analytical result that is logically inconsistent with any of the input data, opting for the most parsimonious result that does not violate Assumption 0,[38] which states that all information from all input parasite–host cladograms must be used without modification or deletions, and the final host cladogram must be logically consistent with all input parasite–host cladograms.

Assumption of BPA

BPA can be implemented using standard parsimony programs, eg PHYLIP2[7] or PAUP,[28] for matrix analysis. See pages 180–206 of Brooks and McLennan's recent text[29] for a detailed description of how to perform BPA.

Testing congruence/ incongruence

A new algorithm, PACT (Phylogenetic Analysis for Comparing trees), which will replace all previous BPA-based methods, including post-1990 BPA, is currently being developed and is scheduled for general release at the next international conference of the International Biogeography Society in January 2005 (D. R. Brooks, personal communication).

Statistical tests of congruence/ incongruence

A number of tests to perform straightforward assessments of congruence/incongruence between tree topologies may be used to explore relationships between parasite and host phylogenies. These include the Kishino-Hasegawa (K–H) test[39] and the incongruence-length difference (ILD) test[30] which is used to assess phylogenetic homogeneity of DNA sequences and combinability of data from different sources. Both are implemented in the current version of PAUP.[28] However, statistical tests developed specifically for assessing congruence between parasite and host phylogenies are also available.[31,34]

Huelsenbeck *et al.*[31] proposed two tests to examine the null hypothesis that host and parasite trees are identical, based on phylogenetic estimates obtained using either maximum likelihood (ML) or maximum posterior probability (ie Bayesian inference). The ML approach uses a likelihood ratio test to examine the null hypothesis, H_0, that the host and parasite trees are identical, H_1 being that the host and parasite trees are not identical. Under H_0 a likelihood value is calculated under the constraint that host and parasite phylogenies are identical, but allowing the parameters of the substitution model to differ. Similarly, a likelihood value is calculated under H_1, in which the host and parasite trees are not constrained to be identical. The degree of congruence (or otherwise) between the two phylogenies is then assessed by calculating the ratio of the observed ML

values: ML H_0/ML H_1. The significance of the observed ratio is then assessed by parametric bootstrapping in which simulated data sets are generated under the assumption that the H_0 is correct. For each simulation of host and parasite data sets, new ML H_0 and ML H_1 values are generated and a new ratio calculated to build up a frequency histogram of simulated ratios with which the observed ratio can be compared and assessed against a chosen significance level, eg 95 per cent.

Bayesian methods

Bayesian inference directly calculates the probability that the host and parasite trees are identical.[32] This is achieved, by calculating the posterior probabilities of a host phylogeny – given the data – and the parasite phylogeny/phylogenies – given the data. The probability of each phylogeny is calculated by Bayesian inference using an appropriate program, eg MrBayes.[33]

Permutation tests

More recently, Legendre *et al.*[34] have developed ParaFit, a matrix permutation test of co-speciation. ParaFit[40] aims to test the significance of a global hypothesis of co-evolution between parasites and hosts, and claims to be robust in areas where other methods often encounter problems, in particular when comparing multiple host and parasite phylogenies (although individual host–parasite associations can also be tested). The test statistics employed are functions of the host and parasite phylogenetic trees and of the set of host–parasite association links.

To statistically assess the hypothesis of host–parasite co-evolution the ParaFit test combines the three types of information that are necessary to describe the situation: the phylogeny of the parasites, the phylogeny of the hosts, and the observed host–parasite associations. Each phylogeny can be described by a matrix of patristic distances among the species along each tree,[41] which in turn can be transformed into a matrix of principal coordinates. In the system described,[34] matrix B describes the parasite tree, matrix C the host tree, while matrix A represents the host–parasite associations – with the parasites in rows and the hosts in columns, a 1 is written where a parasite has been empirically found to be associated to a host, with 0 used elsewhere. If the reconstructed phylogeny for either the hosts or the parasites, or both, is uncertain (eg a phylogeny is poorly resolved or there is uncertainty among several almost equivalent trees) a matrix of phylogenetic distances computed directly from the raw data (eg morphological characters, DNA sequences) can be used instead; the distance matrix is then transformed into a rectangular matrix (B, describing the parasite tree, or C, describing the host tree) by principal coordinate analysis before being used in the ParaFit program.

As stated, the global null hypothesis, as revealed by the two phylogenetic trees and the set of host–parasite association links, is that evolution of the hosts and parasites has been independent, ie that one is random with respect to the other. ParaFit allows a statistical test of this particular global hypothesis of co-evolution and, importantly, also allows the significance of each individual host–parasite link contributing to the overall relationship to be considered and estimated. The role of particular taxa can thus be identified and earmarked for further investigation.

COMPONENT

COMPONENT[5,6] was developed by Rod Page and is primarily a program for implementing the reconciled tree approach to exploring parasite–host associations. It has now been largely superseded by TreeMap 2.02; however, it retains a number of useful features and metrics, and it runs in Microsoft Windows, so it will be described here. COMPONENT uses the NEXUS format[41] and is compatible with PAUP[28] and MacClade,[37] and tree files produced by these programs (together with those produced with PHYLIP[27]) can be read directly by COMPONENT.

COMPONENT 2.0 implements a variety of tree comparison methods, including computing consensus trees,

Reconciled trees — early approaches

calculating the similarity between pairs of trees and mapping one tree onto another, using –among others– the partition metric[42] and quartet measures.[43] These measures can then be used to quantify the similarity between trees as part of congruence studies (eg as implemented in PAUP[28]). The tree-mapping ability, which allows the computation of a tree reconciling incongruent parasite and host trees (and gene and species trees),[15,18] is perhaps of most relevance to those interested in exploring host–parasite relationships. Critically, COMPONENT does not allow any host switching, but relies on the calculation of either duplications and losses, or 'items of error'[35] to reconcile incongruent trees. Comparison of the observed pattern of duplications and losses with that expected under a model of no association between parasite and host is then made to assess host–parasite tree congruence.

Comparisons can be made between two individual trees, a set of trees, or two sets of trees in different input files. COMPONENT can also be used to generate random trees under a variety of models; these distributions can then be used as the basis for statistical tests of similarity between observed trees. COMPONENT can also generate all possible tree shapes for a specified number of taxa, a useful feature for exploring measures of shape and balance.

See the COMPONENT manual[6] for full instructions on using the program; see Slowinski[44] for an in-depth review of COMPONENT.

TreeFitter

Event–cost tree fitting

TreeFitter is a simple program for parsimony-based event–cost tree fitting; it is available for both Macintosh and Windows. It can handle arbitrary cost assignments fulfilling the requirements that duplication events, sorting events and switches all have zero or a positive cost associated with them. Co-divergence events can be associated with either positive, negative or zero cost. In TreeFitter, one type of trees are called P-trees (= parasite trees), the other type H-trees (= host trees). TreeFitter can also be used to explore relationships between gene trees and species trees, in which case the P-trees are gene trees and the H-trees are species trees.

TreeFitter has a limited number of commands but still allows a number of useful inferences to be drawn from the data sets. Treefitter uses the NEXUS format[36] and is compatible with PAUP and MacClade; a list of NEXUS format commands for use with TreeFitter are supplied in the manual (available from the TreeFitter website,[20] where a selection of pre-worked example data files can also be found). TreeFitter fits any number of P-trees to a given H-tree, and it can search for the best H-tree given a set of P-trees. It can calculate the events implied by the minimum-cost solutions. Inferences about historical constraints or the number of events of a particular type can be tested against inferences drawn from random data sets. These random data sets are drawn from the original data either by random permutation of the terminals in the P-tree, the H-tree or both. Alternatively, either the P-trees, the H-tree or both may be replaced by trees drawn at random from a tree space generated by the Markov process in which all labelled histories are equally probable. Finally, TreeFitter can examine portions of parameter space to find the combination of cost assignments giving the best chances of finding historically constrained patterns, given a set of P-trees and an H-tree. By default, TreeFitter works with the following cost assignments: co-divergence and duplication events have zero cost, sorting events have a cost of 1, and switches a cost of 2. This combination of cost assignments reportedly works well for a wide variety of problems, though potential users should consult the manual[20] for a list of situations when unexpected outcomes may result.

TreeMap

Reconciled trees — state-of-the-art

TreeMap is another program based on the reconciled tree approach developed by Rod Page and is a 'direct descendant' of COMPONENT, the key difference from COMPONENT being the ability to incorporate host-switching as an explanation of the observed pattern of host–parasite associations.

Jungles

Its current implementation is TreeMap version 2.02, a program that also provides an option to implement the Jungles[21] event–cost method to find all potentially optimal solutions to explain observed patterns of host–parasite association. The implementation of Jungles in TreeMap 2.02 avoids many of the problems associated with the use of optimality criteria in some earlier programs, where there could be many reconstructions implying the same number of co-speciation events, thus yielding multiple solutions. TreeMap 2 avoids this problem by using Jungles to search for all feasible reconstructions within bounds set by the user. Significantly, the Jungles algorithm in TreeMap 2 can explore all switches, including those that require subsequent sorting events to ensure that source and destination are contemporary (known as 'weakly incompatible switches'). Previous programs ignored these switches and only included those that obeyed the temporal rules for switches without further events ('compatible switches'). For example, the user can specify the maximum number of host switches that any reconstruction can have; the program then filters the solutions to remove any that are definitely non-optimal for the given set of costs (as noted above, placing costs on events does not cause TreeMap 2 to automatically discard solutions of low co-divergence, unless bounds are also placed on the total cost, such that solutions beyond these bounds are then excluded). To evaluate individual reconstructions the user can specify costs for each event (duplication, host switch and sorting events).

Exploring solutions

In this way the user can still explore alternative reconstructions (as in COMPONENT and TreeMap version 1), but not be swamped with many similar, but non-optimal solutions. Jungles uses four parameters to calculate the overall cost of each hypothesised past association individually; these are: co-speciation; duplication; lineage sorting; and host switching, and the user is prompted to enter values for the 'cost' of each. TreeMap 2 then estimates the significance of observed co-divergence, total cost, or the number of another event type, using randomisation tests.

Practical details

Running TreeMap

On starting TreeMap (which currently runs only in OS9 or Classic mode in OSX) a PAUP-style command log window appears. Using the standard Mac OS top menu bar one can then select to create or load an existing data file, and a number of example files are provided to begin working with. As in PAUP, the user is given the choice of loading and executing the file, or opening it in edit mode. Once executed, a series of additional windows open, in which the user can perform a range of analyses; these include viewing the tanglegram – a visual representation of the parasite tree and the host tree, and the relationships defined between parasite and host species.

Associations are defined by the user and can be edited in the 'Association' window selected from the pull-down menu; this panel contains three editable lists: hosts, parasites found on a particular (highlighted) host, parasites included in the study, but not found on the highlighted host. (Users should be aware that many menu options are related to a particular panel being active – and to a particular analysis being undertaken – and thus many are greyed-out and non-accessible at different times). With the 'Tanglegram' window active, the user is then given the option of editing the costs of evolutionary events: Duplication, Lineage loss and Host switching, relative to the cost of Co-divergence which remains fixed at 0. The user can then select from the same pull-down menu the option to 'Make Jungle. . .', using the

defined costs. All optimal solutions can then be viewed as reconciled trees (host phylogeny thick lines, parasite phylogeny thin lines) in the 'Reconstruction' window, accompanied by a separate 'Reconstruction table' window, in which the costs of all optimal solutions are listed. Thus, while a few commands still remain to be implemented, TreeMap 2.02 provides a powerful and easy-to-grasp tool for analysing host–parasite associations.

FUTURE DEVELOPMENTS

Future developments look set to focus on a maximum likelihood implementation of Jungles, such that specific questions about the likelihood of particular solutions can be asked, rather than simply focusing on maximising the optimal solution(s). Current approaches generally assume that the estimated trees are known, without error. Baysian methods, however, which allow different models to be tried and the use of statistical tests to choose among the competing models of host switching, will allow the development of programs to identify the model that best explains the observed data.[45]

Likelihood methods

Whatever, the sometimes deeply contrasting philosophies of proponents of *a posteriori* methods (ie BPA) and *a priori* methods (ie reconciled tree methods, eg TreeMap, and event–cost methods, eg Jungles) seem set to ensure that the field of host–parasite co-evolution remains much in debate for the foreseeable future.

Continuing controversy

Acknowledgments

Thanks to Dan Brooks for providing information on the as-yet unpublished PACT algorithm and program, to Rod Page for comments on various drafts of the manuscript and for insight into the workings of TreeMap 2, to two anonymous referees for constructive feedback and to Dr Lucie Evans for additional research.

References

1. Page, R. D. M. and Charleston, M. A. (1998), 'Trees within trees: Phylogeny and historical associations', *Trends Ecol. Evol.*, Vol. 13(9), pp. 356–359.
2. Rosen, D. E. (1978), 'Vicariant patterns and historical explanation in biogeography', *Syst. Zool.* Vol. 27, pp. 159–188.
3. Brooks, D. R. (1981), 'Hennig's parasitological method: A proposed solution', *Syst. Zool.*, Vol. 30, pp. 229–249.
4. Brooks, D. R. (1990), 'Parsimony analysis in historical biogeography and coevolution: Methodological and theoretical update', *Syst. Zool.*, Vol. 39(1), pp. 14–30.
5. Page, R. D. M. (1990), 'Component analysis: A valiant failure?', *Cladistics*, Vol. 6, pp. 119–136.
6. Page, R. D. M. (1993), 'COMPONENT User's manual (Version 2.0)', Natural History Museum, London (URL: http://taxonomy.zoology.gla.ac.uk/rod/cpw.html, cited 19th March 2004).
7. Charleston, M. A. and Page, R. D. M. (2002), 'TREEMAP 2.0ß: A Macintosh program for the analysis of how dependent phylogenies are related, by cophylogeny mapping' (URL: http://evolve.zoo.ox.ac.uk/software/TreeMap/main.html, cited 19th March 2004).
8. Dowling, A. P. G. (2002), 'Testing the accuracy of TreeMap and Brooks parsimony analyses of coevolutionary patterns using artificial associations', *Cladistics*, Vol. 18, pp. 416–435.
9. Page, R. D. M. and Charleston, M. A. (2002), 'Treemap versus BPA (again): A response to Dowling', *Tech. Rep. Taxonomy 02–02*, pp. 1–26 (URL: http://taxonomy.zoology.gla.ac.uk/publications/tech-reports/, cited 15th March 2004).
10. Dowling, A. P. G., van Veller, M. G. P., Hoberg, E. P. and Brooks, D. R. (2003), '*A priori* and *a posteriori* methods in comparative evolutionary studies of host–parasite associations', *Cladistics*, Vol. 19, pp. 240–253.
11. Page, R. D. M. (2003), 'Introduction', in Page, R. D. M., Ed., 'Tangled Trees: Phylogeny, Cospeciation and Coevolution', The University of Chicago Press, Chicago, IL, pp. 1–21.
12. Brooks, D. R., Dowling, A. P. G., van Veller, M. G. P. and Hoberg, E. P. (2004), 'Ending a decade of deception: A valiant failure, a not-so-valiant failure, and a success story', *Cladistics*, Vol. 20, pp. 32–46.
13. Brooks, D. R. (2003), 'The new orthogenesis', *Cladistics*, Vol. 19, pp. 443–448.
14. Hennig, W. (1966), 'Phylogenetic Systematics', University of Illinois Press, Urbana, IL.
15. Goodman, M., Czelusniak, J., Moore, G. W. *et al.* (1979), 'Fitting the gene lineage into its species lineage: A parsimony strategy illustrated by cladograms constructed from globin sequences', *Syst. Zool.*, Vol. 28, pp. 132–168.
16. Page, R. D. M. (1994), 'Maps between trees

and cladistic analysis of historical associations among genes, organisms and areas', *Syst. Biol.*, Vol. 43, pp. 58–77.

17. Page, R. D. M. (1994), 'Parallel phylogenies: Reconstructing the history of host–parasite assemblages', *Cladistics*, Vol. 10, pp. 155–173.
18. Nelson, G. and Platnick, N. I. (1981), 'Systematics and Biogeography; Cladistics and Vicariance', Columbia University Press, New York.
19. Ronquist, F. (1995), 'Reconstructing the history of host–parasite associations using generalized parsimony', *Cladistics*, Vol. 11, pp. 73–89.
20. Ronquist, F. (1999–2001) TreeFitter, Version 1.0, Department of Systematic Zoology, Uppsala University, Sweden (URL: www.ebc.uu.se/systzoo/research/treefitter/treefitter.html, cited 20th April 2004).
21. Charleston, M. A. (1998), 'Jungles: A new solution to the host/parasite phylogeny reconciliation problem', *Math. Biocsci.*, Vol. 149, pp. 191–223.
22. Hafner, M. S. and Nadler, S. A. (1988), 'Phylogenetic trees support the coevolution of parasites and their hosts', *Nature*, Vol. 332, pp. 258–259.
23. Page, R. D. M., Lee, P. L. M., Becher, S. A. *et al.* (1998), 'A different tempo of mitochondrial DNA evolution in birds and their parasitic lice', *Mol. Phylogenet. Evol.*, Vol. 9(2), pp. 276–293.
24. Stevens, J. R., Noyes, H. A., Dover, G. A. and Gibson, W. C. (1999), 'The ancient and divergent origins of the human pathogenic trypanosomes, *Trypanosoma brucei* and *T. cruzi*', *Parasitology*, Vol. 118, pp. 107–116.
25. Stevens, J. and Rambaut, A. (2001), 'Evolutionary rate differences in trypanosomes', *Infection Genet. Evol.*, Vol. 1, pp. 143–150.
26. Yoder, A. D. and Yang, Z. (2000), 'Estimation of primate speciation dates using local molecular clocks', *Mol. Biol. Evol.,* Vol. 17, pp. 1081–1090.
27. Felsenstein, J. (1993), 'PHYLIP – Phylogeny Inference Package, Version 3.5', University of Washington.
28. Swofford, D. L. (2002), 'PAUP* – Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4', Sinauer Associates, Sunderland, MA.
29. Brooks, D. R. and McLennan, D. A. (2002), 'The Nature of Diversity: An Evolutionary Voyage of Discovery', University of Chicago Press, Chicago, IL.
30. Farris, J. S., Källersjö, M., Kluge, A. G. and Bult, C. (1995), 'Testing significance of incongruence', *Cladistics*, Vol. 10, pp. 315–319.
31. Huelsenbeck, J. P., Rannala, B. and Yang, Z. (1997), 'Statistical tests of host–parasite cospeciation', *Evolution*, Vol. 51(2), pp. 410–419.
32. Huelsenbeck, J. P., Rannala, B. and Larget, B. (2000), 'A Bayesian framework for the analysis of cospeciation', *Evolution*, Vol. 54, pp. 352–364.
33. Huelsenbeck, J. P. and Ronquist, F. (2001), 'MRBAYES: Bayesian inference of phylogeny', *Bioinformatics*, Vol. 17, pp. 754–755.
34. Legendre, P., Desdevises, Y. and Bazin, E. (2002), 'A statistical test for host–parasite coevolution', *Syst. Biol.*, Vol. 51(2), pp. 217–234.
35. Robinson, D. F. and Foulds, L. R. (1981), 'Comparison of phylogenetic trees', *Math. Biocsci.*, Vol. 53, pp. 131–147.
36. Maddison, D. R., Swofford, D. L. and Maddison, W. P. (1997) 'NEXUS: An extensible file format for systematic information', *Syst. Biol.*, Vol. 46, pp. 590–621.
37. Maddison, W. P. and Maddison, D. R. (1992), 'MacClade: Analysis of Phylogeny and Character Evolution, Version 3.0', Sinauer Associates, Sunderland, MA.
38. Wiley, E. O. (1988), 'Parsimony analysis and vicariance biogeography', *Syst. Zool.*, Vol. 37, pp. 271 –290.
39. Kishino, H. and Hasegawa, M. (1989), 'Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea', *J. Mol. Evol.*, Vol. 29, pp. 170–179.
40. URL: http://www.fas.umontreal.ca/biol/legendre/
41. Lapointe, F.-J. and Legendre, P. (1992), 'A statistical framework to test the consensus among additive trees (cladograms)', *Syst. Biol.*, Vol. 41, pp. 158–171.
42. Penny, D. and Hendy, M. D. (1985), 'The use of tree comparison metrics', *Syst. Zool.*, Vol. 34, pp. 75–82.
43. Estabrook, G. F., McMorris, F. R. and Meacham, C. A. (1985), 'Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units', *Syst. Zool.*, Vol. 34, pp. 193–200.
44. Slowinski, J. B. (1993), 'Review – Component 2.0.', *Cladistics*, Vol. 9, pp. 351–353.
45. Huelsenbeck, J. P., Rannala, B. and Larget, B. (2003), 'A statistical perspective for reconstructing the history of host–parasite associations, in Page, R. D. M., Ed., 'Tangled Trees: Phylogeny, Cospeciation and Coevolution', The University of Chicago Press, Chicago, IL, pp. 93–119.

Copyright of Briefings in Bioinformatics is the property of Henry Stewart Publications and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.