Trees within trees: Phylogeny and historical associations

Roderic D. M. Page and Michael A. Charleston

Roderic D. M. Page is at the Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow, UK G12 8QQ (r.page@bio.gla.ac.uk); Michael A. Charleston is at the Department of Zoology, University of Oxford, Oxford, UK OX1 3PS (michael.charleston@zoology.oxford.ac.uk)

Keywords:reconciled trees, phylogeny, biogeography, coevolution, gene trees, molecular evolution, horizontal transfer

The association between two or more lineages over evolutionary time is a recurrent theme spanning several different fields within biology, from molecular evolution to coevolution and biogeography. In each `historical association', 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. Examples include genes tracking organisms, parasites tracking hosts, and organisms tracking geological and geographical changes. Parallels among these problems raise the tantalizing prospect that each is a special case of a more general problem, and that a single analytical tool can be applied to all three kinds of association.

Evolutionary associations among genes, organisms and areas have traditionally been studied by biologists from different disciplines, with little interaction between them. Consequently, recognition of the fundamental similarity of the problem faced by molecular systematists, parasitologists and biogeographers has been slow in coming[1-3]. This is particularly true of the parallels between the relationship between gene and organismal phylogeny, and the macroevolutionary associations studied by parasitologists and biogeographers. The analogy between vicariance

biogeography (organisms tracking areas) and host-parasite cospeciation (parasites tracking hosts) has been recognised for some time[4]; for a parasite the host can be thought of as an `area', hence host speciation is equivalent to a vicariance event (Fig. 1). The suggestion that these macroevolutionary patterns are analogous to the relationship between gene and species trees is a more recent development[1,3].

Kinds of historical associations

Historical associations can be divided into three basic categories (Table 1): genes and organisms, organisms and organisms, and organisms and areas. At the molecular level, each gene has a phylogenetic history that is intimately connected with, but not necessarily identical to, the history of the organisms in which the gene resides[5,6]. Processes such as gene duplication, lineage sorting and horizontal transfer can produce complex gene trees that differ from organismal trees[3,7,8]. Associations between organisms, such as between hosts and their parasites[9] (including viruses[10]), endosymbionts and their hosts[11], and insects and plants[12,13], may have a long evolutionary

history, which is reflected in similarities between their evolutionary trees[14]. At a larger scale still, organisms may track geological history such that sequences of geological events (e.g. continental break-up) are directly reflected in the phylogenies of those organisms[15].

In each association, one entity (the `associate') tracks the other (the `host') with a degree of fidelity that depends on the relative frequency of four categories of events: codivergence, duplication, horizontal transfer, and sorting (Box 1). Joint cladogenesis of host and associate is codivergence. If the associates undergo cladogenesis independently and both descendants remain associated with the host then we have a duplication of associate lineages. Cladogenesis accompanied by one descendant colonising a new host is horizontal transfer. `Sorting event' is a generic term for the apparent absence of an associate from a host.

The analogies among the categories of events for the different kinds of association (Table 1) need not imply close analogy among the processes; rather, the analogy is among the patterns these processes produce. For example, although the processes of gene duplication and allele divergence are different, the resulting pattern is the same – more than one gene lineage in the same organismal lineage.

Reconstructing the history of an association

Despite the relative lack of interaction among these different disciplines, strikingly similar concepts have arisen independently from them. Parasitologists[16,17] recognized the problem of multiple parasite lineages decades before Fitch's[8] analogous distinction between paralogous and orthologous genes[18]. Molecular systematists[19] and cladistic biogeographers[20] independently developed similar methods for interpreting the history of gene trees and biogeographic patterns, respectively.

One implication of the parallels among the different kinds of association is that they can be studied using the same methods. Reconciled trees (Box 2) originated in molecular systematics[19] but have been applied to both host-parasite coevolution[21] and biogeography[22]. In addition to visualizing the relationship between host and associate, reconciled trees provide a quantitative measure of the extent to which the host and its associate's phylogenetic histories have diverged. Hence, reconciled trees have been employed in two different ways: to document the history of an associate where both the host and associate relationships are presumed to be known, and to infer host relationships based on the associate phylogeny. The inference of species

trees from gene trees is the paradigm instance of the latter, but there is a long history of parasitologists attempting to infer host phylogeny from parasite phylogeny[23], and cladistic biogeographers aim to infer geological history from organismal phylogeny[15].

Multiple lineages

Duplications result in multiple lineages of associates on the same host lineages. The implications of this for inferring host phylogenies have been recognized by molecular systematists dealing with gene families[8] or multiple mitochondrial lineages[24]. If some gene lineages go extinct or are incompletely sampled, gene trees may not faithfully reflect organismal history (Fig. 2). In contrast, parasitologists and biogeographers have (with a few exceptions[17,20]) attributed discordance between host and associate trees to horizontal transfer, rather than a combination of multiple lineages and subsequent loss of associates. Host-switching and dispersal do occur, but as they might not be the sole cause of discordance their prevalence can be overestimated[18].

Applications of reconciled trees

With the increasing availability of nuclear gene sequences[25], reconciled trees may find ready application to the study of the evolution of gene diversity and the inference of organismal phylogeny from multiple, complex gene trees[26], as well as tools for database analysis[27]. Recent work [28,29] suggests that lineage sorting within a single gene may pose less of problem for phylogenetic inferrence than previously thought. However, analyses at large taxonomic scales using nuclear genes are likely to encounter problems due to gene duplication and resulting paralogy. As an example, Guigó *et al.*[30] used a variant of reconciled trees to infer eukaryote phylogeny from 53 gene trees, and discovered that only a third of the genes were perfectly consistent with the best fitting eukaryote tree. Although this study has flaws[31], it suggests that great care must be taken in using nuclear genes as phylogenetic markers – individual genes or gene families may be quite misleading about species relationships. Viewed in this light inferring organismal phylogeny from single genes becomes as fraught as inferences based on single morphological characters[3], making it essential to analyse multiple genes. In the same way, parasitologists[4] and biogeographers[32] have stressed the need

to use multiple associate trees to infer the relationships among hosts and areas, respectively.

Reconciled trees can be predictive tools. The missing lineages corresponding to sorting events (Box 2) represent associate lineages that are either extant but undiscovered, or extinct. Many sorting events in reconciled trees for genes are likely to represent undiscovered genes rather than genuine losses, given the uneven sample of sequences represented in the sequence databases. Extinct associates may also leave evidence of their previous existence. Linder and Crisp's[22] reconciliation of a phylogeny for Southern beech trees (*Nothofagus*) with a geological area cladogram required postulating the existence of a *Nothofagus* clade in areas where it is currently no longer found alive, but where fossils of that clade are known to occur.

`Jungles'

Reconciled trees have nice properties, but also some limitations, the most severe being that they do not accommodate horizontal transfer. Other methods, such as Brooks' parsimony analysis (BPA)[4], do incorporate this process, but they do not always produce biologically reasonable reconstructions[33]. Horizontal transfer poses problems that have only recently been appreciated (Box 3).

Charleston[34] has developed a solution to this problem that employs a mathematical structure called a `jungle', which contains all the possible ways in which an associate tree can be mapped into a host tree, given the four processes of codivergence, duplication, sorting, and horizontal transfer, and the extant associations known. Given `costs' for each of these processes, it is possible to find the subgraph(s) of the jungle that corresponds to the least costly (e.g., most parsimonious) reconstruction(s) of the history of the association. This also represents an improvement in the computation time required as previous methods had to rely on heuristic procedures which were not guaranteed to find optimal solutions, whereas jungles are solved using a dynamic programming approach.

Prospects

Methods for phylogenetic analysis of historical associations are still being refined, with considerable scope for future development. The analogy between the different categories of association has proved a useful heuristic tool, but detailed analogies between the processes may proved strained. More sophisticated analyses will require careful consideration of the actual processes operating in each association, especially if maximum

likelihood methods are to be developed[5,35]. Alternatively, there is a case for pushing the analogy to the limit to maximize the extent to which the apparently disparate disciplines of molecular systematics, parasitology and biogeography can employ the same analytical tools.

Acknowledgements

We thank Rob Cruickshank, Richard Griffiths, Vince Smith, Chris Simon, and two anonymous reviewers for their comments. This work was supported by NERC grant GR3/1A095 to the first author.

References

1 Page, R.D.M. (1993) **Genes, organisms, and areas: the problem of multiple lineages,** *Syst. Biol.* 42, 77-84

2 Baum, B.R. (1992) **Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees,** *Taxon* 41, 3-10

3 Doyle, J.J. (1992) **Gene trees and species trees: molecular systematics as one-character taxonomy,** *Syst. Bot.* 17, 144-63

4 Brooks, D.R. (1981) **Hennig's parasitological method: a proposed solution,** *Syst. Zool.* 30, 229-49

5 Maddison, W.P. (1997) **Gene trees in species trees,** *Syst. Biol.* 46, 523-536

6 Doyle, J.J. (1997) **Trees within trees: genes and species, molecules and morphology,** *Syst. Biol.* 46, 537-553

7 Pamilo, P. and Nei, M. (1988) **Relationships between gene trees and species trees,** *Mol. Biol. Evol.* 5, 568-83

8 Fitch, W.M. (1970) **Distinguishing homologous from analogous proteins,** *Syst. Zool.* 19, 99-113

9 Hafner, M.S. *et al*. (1994) **Disparate rates of molecular evolution in cospeciating hosts and parasites,** *Science* 265, 1087-90

10 McGeoch, D.J. *et al*. (1995) **Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses,** *J. Mol. Biol.* 247, 443-58

11 Moran, N.A., van Dohlen, C.D. and Baumann, P. (1995) **Faster evolutionary rates in endosymbiotic bacteria than in cospeciating insect hosts,** *J. Mol. Evol.* 41, 727-31

12 Farrell, B. and Mitter, C. (1990) **Phylogenies of insect/plant interactions: have *Phyllobrotica* leaf beetles (Chrysomelidae) and the Lamiales diversified in parallel?,** *Evolution* 44, 1389-403

13 Herre, E.A. *et al*. (1996) **Molecular phylogenies of figs and their pollinator wasps,** *J. Biogeog.* 23, 521-530

14 Hafner, M.S. and Nadler, S.A. (1988) **Phylogenetic trees support the coevolution of parasites and their hosts,** *Nature* 332, 258-259

15 Rosen, D.E. (1978) **Vicariant patterns and historical explanation in biogeography,** *Syst. Zool.* 27, 159-188

16 Hopkins, G.H.E. (1948) **Some factors which have modified the phylogenetic relationship between parasite and host in the Mallophaga,** *Proc. Linn. Soc., Lond.* 161, 37-39

17 Clay, T. (1949) **Some problems in the evolution of a group of ectoparasites,** *Evolution* 3, 279-99

18 Page, R.D.M., Clayton, D.H. and Paterson, A.M. (1996) **Lice and cospeciation: a response to Barker,** *Int. J. Parasitol.* 26, 213-218

19 Goodman, M. *et al*. (1979) **Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences,** *Syst. Zool.* 28, 132-168

20 Nelson, G. and Platnick, N.I. (1981) *Systematics and Biogeography: Cladistics and Vicariance*, Columbia University Press

21 Paterson, A.M. and Gray, R.D. (1997) **Host-parasite cospeciation, host switching, and missing the boat,** in *Host-Parasite Evolution: General Principles and Avian Models*, (Clayton, D. H. and Moore, J., eds.), pp. 236-250, Oxford University Press

22 Linder, H.P. and Crisp, M.D. (1995) ***Nothofagus* and Pacific biogeography,** *Cladistics* 11, 5-32

23 Hoberg, E.P., Brooks, D.R. and Seigel-Causey, D. (1997) **Host-parasite co-speciation: history, principles, and prospects,** in *Host-Parasite Evolution: General Principles and Avian Models*, (Clayton, D. H. and Moore, J., eds.), pp. 212-235, Oxford University Press

24 Avise, J.C. (1989) **Gene trees and organismal histories: a phylogenetic approach to population biology,** *Evolution* 43, 1192-1208

25 Brown, J.R. (1996) **Preparing for the flood: evolutionary biology in the age of genomics,** *Trends Ecol. Evol.* 11, 510-513

26 Page, R.D.M. and Charleston, M.A. (1997) **From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem,** *Mol. Phylog. Evol.* 7, 231-240

27 Yuan, Y.P. *et al*. (in press) **Towards detection of orthologues in sequence databases,** *Comput. Applic. Biosci.*

28 Hoelzer, G.A. (1997) **Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees revisited,** *Evolution* 51, 622-626

29 Moore, W.S. (1997) **Mitochondrial-gene trees versus nuclear gene trees, a reply to Hoelzer,** *Evolution* 51, 627-629

30 Guigó, R., Muchnik, I. and Smith, T.F. (1996) **Reconstruction of ancient molecular phylogeny,** *Mol. Phylog. Evol.* 6, 189-213

31 Page, R.D.M. and Charleston, M.A. (1997) **Reconciled trees and incongruent gene and species trees,** in

Mathematical Hierarchies in Biology (DIMACS Series in Discrete Mathematics and Theoretical Computer Science), (Mirkin, B., McMorris, F. R., Roberts, F. S. and Rzhetsky, A., eds.), pp. 57-70, American Mathematical Society

32 Platnick, N.I. and Nelson, G. (1978) **A method of analysis for historical biogeography,** *Syst. Zool.* 27, 1-16

33 Page, R.D.M. (1994) **Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas,** *Syst. Biol.* 43, 58-77

34 Charleston, M.A. (in press) **Jungles: A new solution to the host/parasite phylogeny reconciliation problem,** *Math. Biosci.*

35 Huelsenbeck, J.P., Rannala, B. and Yang, Z. (1997) **Statistical tests of host-parasite cospeciation,** *Evolution* 51, 410-419

36 Mirkin, B., Muchnik, I. and Smith, T.F. (1995) **A biologically consistent model for comparing molecular phylogenies,** *J. Comput. Biol.* 2, 493-507

37 Delwiche, C.F. and Palmer, J.D. (1996) **Rampant horizontal transfer and duplication of rubisco genes in Eubacteria and plastids,** *Mol. Biol. Evol.* 13, 873-882

38 Page, R.D.M. (1994) **Parallel phylogenies: reconstructing the history of host-parasite assemblages,** *Cladistics* 10, 155-173

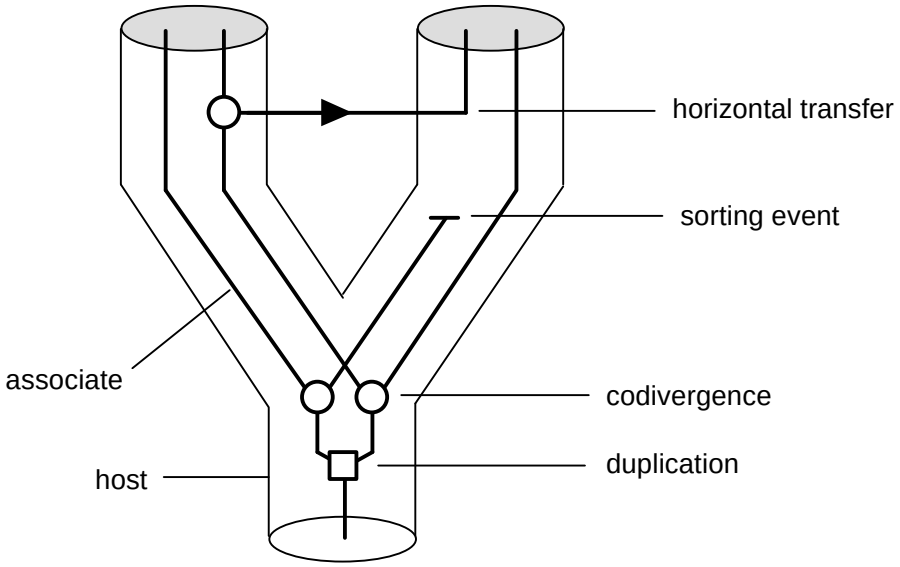
39 Ronquist, F. (1995) **Reconstructing the history of host-parasite associations using generalised parsimony,** *Cladistics* 11, 73-89

40 Takahata, N. (1989) **Gene genealogy in three related populations: consistency probability between gene and population trees,** *Genetics* 122, 957-966

Boxes

Box 1. Terminology of historical associations

Given the parallels among different kinds of historical association it is desirable to have a generic set of terms that are applicable to genes, organisms, and areas [33]. The terminology used here is illustrated below.



Associate: a lineage that tracks another lineage, or set of historically related entities (such as geographical areas).

Codivergence: joint divergence of both host and associate. Examples include host-parasite cospeciation, and vicariance.

Duplication: independent divergence of the associate, with both descendants remaining associated with the host (e.g., gene duplication).

Horizontal transfer: transfer of an associate lineage from one host (`source') to another host (`destination') that is not itself the immediate descendant of the source host. Examples include horizontal gene transfer and host switching.

Host: the lineage or entities being tracked, such as organisms harbouring a lineage of parasites.

`Missing the boat:' If an associate is distributed over only part of the host's distribution (e.g. a patchily distributed parasite), divergence within the host lineage may yield one or more descendants that lack the associate. The associate has not gone extinct from those hosts, it simply was never there[21].

Orthologous: a pair of genes that are descendants of the same copy of a gene are orthologous.

Paralogous: a pair of genes separated by at least one gene duplication are paralogous.

Reconciled tree: the simplest embedding of an associate tree inside its host tree (Box 2).

Sorting event: the (apparent) absence of an associate in the descendants of a host that had previously had that associate. Sorting events include extinction of the

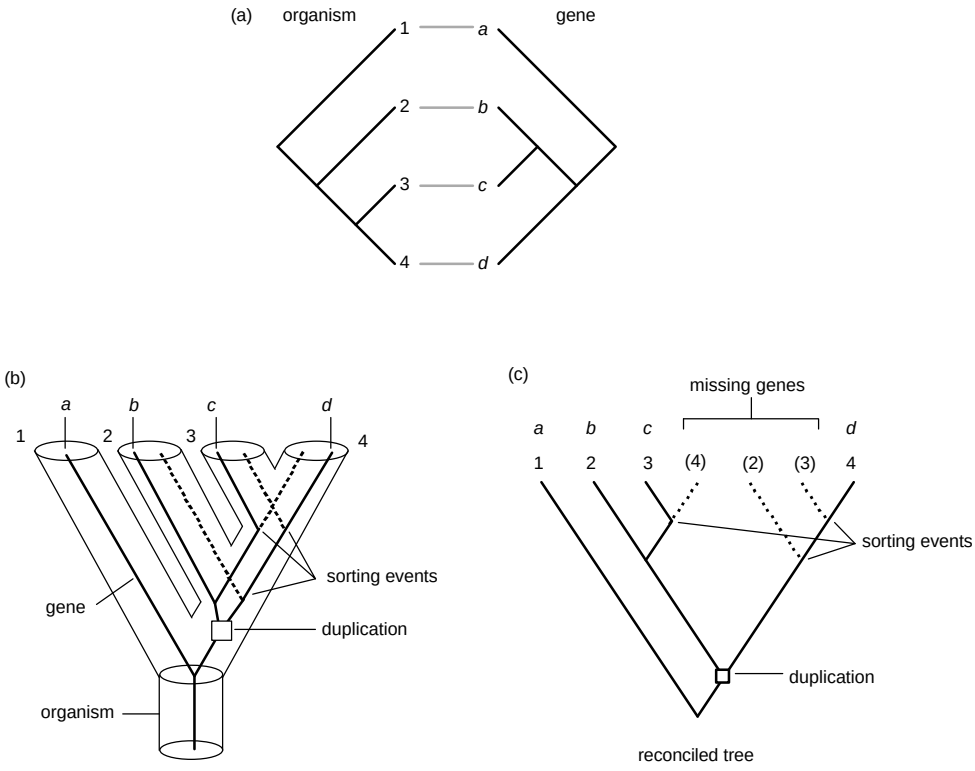
associate and `missing the boat.’ Associates can also be present but undetected, such as with rare parasites or undiscovered gene loci.

Box 2. Reconciled trees

The concept of reconciled trees dates from Goodman *et al.*'s[19] attempts to reconcile disagreements between then accepted mammalian evolutionary relationships and those obtained from haemoglobin genes. Largely neglected until recently, reconciled trees are now receiving renewed attention from biologists and mathematicians[26,27,30,31,33,36]. Suppose we have a phylogeny for four species, and four sequences sampled from those species, and that the two trees, which we believe to be correct, disagree (a).

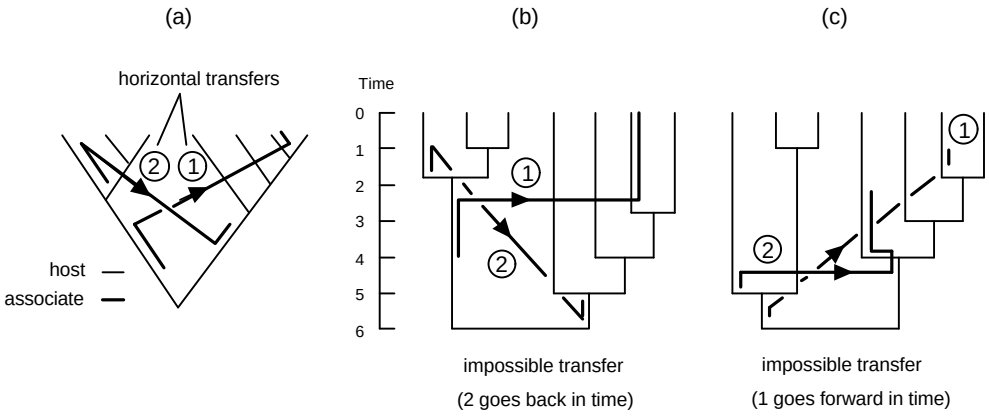
The question is, how can the trees both be true, and yet be discordant? One explanation is to embed the gene tree in the species tree (b), which requires us to postulate a number of gene duplications and subsequent losses (in this instance one duplication and three losses). This embedding can also be represented using a reconciled tree (c), which simply takes the embedded gene tree and `unfolds' it so that it lies flat on the page. The reconciled tree would depict the complete history of the gene if there had been no gene losses (i.e. the three sorting events). As a consequence of the gene duplication in the ancestor of species 2, 3 and 4, we would expect those species to each have two copies of the gene. Because they do not we must postulate three gene losses. Alternatively, the gene copies may be present but undetected. Hence the reconciled tree makes predictions

about the existence of undiscovered genes. It also suggests that genes *b* and *c* are paralogous to gene *d*, which is not apparent from the gene phylogeny alone.



Box 3. Horizontal transfer

Reconciled trees (Box 2) do not allow for horizontal transfer of associates among host lineages, and hence exclude a class of processes (horizontal gene transfer, host-switching and dispersal) which are known to occur[37]. Because postulating a horizontal transfer requires that the source and destination hosts are contemporaneous, we have to consider the relative ages of different host lineages. Failure to take this constraint into account can resulting in postulating transfers that are mutually incompatible. An example is shown in (a) where making the source and destination host lineages contemporaneous for one horizontal transfer makes the other impossible (b and c). Initial attempts[38] to incorporate horizontal transfers failed to address this problem[39], which has since been solved[34].



Tables

Table 1. Equivalent processes among different historical associations

Host/Associate	Codivergence	Duplication	Horizontal transfer	Sorting event
Organism/gene	interspecific coalescence[40]	gene duplication deep coalescence[5]	gene transfer	gene loss, lineage sorting
Host/Parasite	cospeciation	within host speciation	host-switch	parasite extinction, missing the boat
Organism/areas	vicariance	sympatry	dispersal	extinction

Figure captions

Fig. 1. Historical associations among genes, organisms and areas:(a) a gene tree embedded in a species tree; (b) a parasite cospeciating with its host; and (c) a clade of organisms diverging in concert with geological events (vicariance). In each case one entity (the `associate') may be thought of as tracking the other (the `host').

Fig. 2 The presence of multiple associate lineages on the same hosts can lead to spurious inferences about host relationships. In this example, there are two associate lineages (α and β) on the same hosts (a). If only one associate from each host is sampled [circled in (a)], then it is possible to infer incorrect host relationships (b), even though the associate relationships are correct.

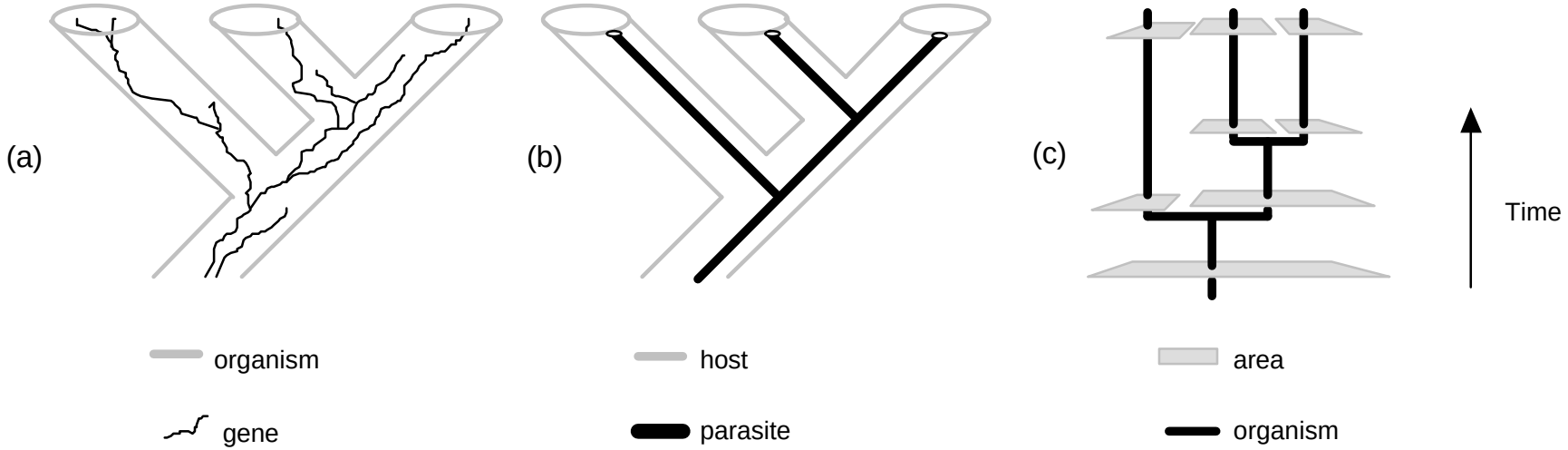


Fig. 1

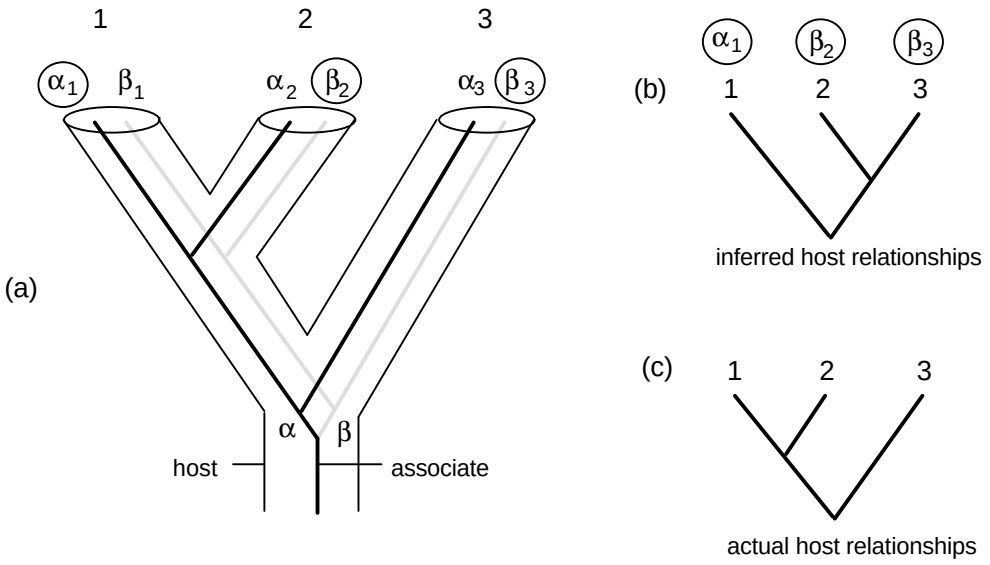

1
2
3
α1
β1
α2
β2
α3
β3
(a)
α
β
host
associate
(b)
α1
β2
β3
1
2
3
inferred host relationships
(c)
1
2
3
actual host relationships

Fig 2