ANALYSIS OF METHODOLOGY IN COPHYLOGENETIC STUDIES

Arthur Yang

Biochem218 Final Paper

March 16, 2009

INTRODUCTION

Cophylogeny is "the study of the relationships between phylogenies of ecologically related groups (taxa, geographical areas, genes etc.), where one, the 'host' phylogeny is independent and the other, the 'associate' phylogeny, is hypothesized to be dependent to some degree on the host" (Charleston. "Principles."). One of the more common areas of cophylogenetic study revolves around the coevolution of two species, most of which participate in a host-parasite relationship. This relationship will be the use focus for this paper. There are several other uses of cophylogenetic studies that will be mentioned later.

One important note to make is that there is a difference when using the terms coevolution, codivergence/cospeciation, and cophylogeny. Coevolution refers to the "general process of reciprocal evolutionary change in two species or populations of interacting organisms." Codivergence or cospeciation - subsets of coevolution - refers to the "speciation of one biological entity resulting in the speciation of those entities that are associated with it." Thus, in the example of a host-parasite tandem, the host and parasite can coevolve with each other producing new traits and characteristics while interacting with each other. If those new traits are different enough to characterize a new species, then codivergence/cospeciation has occurred. Cophylogeny, then, involves the analyses and comparisons of the organisms' phylogenies to uncover patterns of any codivergence. (Charleston. "Traversing.")

Early on, studies of host-parasite relationships brought about the adoption of rules based on notable trends of those relationships. Farenholz's rule, as a primary example, claims that the "parasite phylogeny will mirror that of the host phylogeny" (Paterson). In essence, the rule narrows the coevolution possibilities to just codivergence. In studies today, the 'rule' is still frequently used as a null hypothesis measure of vertical transmission. However, it becomes immediately apparent that there are other possible biological occurrences.

Codivergence, as it turns out, is just one of the possible cophylogenetic events that can occur. The pattern of codivergence is not the absolute rule since the two species do not necessarily follow identical evolutionary paths. Four alternative scenarios add further complexity into the cophylogenetic study model.

A second scenario - other than codivergence - manifests itself as a seeming removal of the parasite species from the host species. Known as lineage sorting, this disappearance can be caused by one of two biological occurrences. One possibility is commonly known as "missing the boat," in which the host speciates but its parasite only follows one of the divergent host species, thus becoming lost on the other host species (Charleston. "Traversing."). This can happen if the host moves to a new habitat without the parasite and diverges into a new species, unlinked to the parasite. The other possibility occurs if the parasite simply goes extinct, leaving the host species without a parasitic link.

A third scenario is duplication, or intrahost speciation, and occurs when the parasite species diverges "without the stimulus of host speciation" (Paterson). This can occur if two parasite populations segregate into different niches - possibly to avoid competition - and end up speciating, though still relating to the same host species.

A four scenario involves a parasite switching completely from one host species to another. Some of the parasite may remain on the initial host species - and it frequently does - but the important distinction is that some of the parasite jumps to a new host species, independent of any host divergences.

A final scenario is merely the absence of any reaction to a host species' divergence. This "inertia" (Paterson) or "failure to diverge"(Charleston. "Traversing.") results in a multi-host parasite, a single parasite species that is able to survive on several different host species. Differing from the "missing the boat" situation, this scenario features the parasite following both the diverging host species - though still not diverging itself.

These five biological scenarios provide the basis for which host-parasite cophylogenetic studies can be done. Using these categories, a comparison of phylogenies can result in better understanding of the coevolutionary events that result in the existing taxonomical structure of species. As such, many of the cophylogenetic methods to be discussed use these categorical events to describe their results.

An important note about cophylogenetic studies is that they assume that adequate and thorough sampling of the species has been done. In addition, the construction of the taxonomies and phylogenies for each of the species is assumed to be sound. These assumptions are noteworthy because they can potentially affect analysis of cophylogeny. Whereas the scenarios described earlier are the result of natural and biological occurrences, human error in sampling and phylogenies can add other scenarios that result in incorrect or incomplete interpretations of the cophylogeny. For example, an observed extinction of a parasitic species from the host can be the legitimate result of biological evolution or a mere result of incomplete sampling. More importantly, accurate constructions of the species' phylogenies must be done to ensure

subsequently accurate analysis of cophylogeny. The smallest difference in a species' phylogeny can induce an incorrect characterization of which coevolutionary scenario has occurred.

METHODS

The many methods of cophylogenetic analysis can be broadly divided into two main categories: "event-based methods" and "global fit methods" (Desdevises). Event-based methods attempt to use the five coevolutionary scenarios to map the parasite phylogeny to the host phylogeny. This mapping can be accomplished via one of two methodologies: parsimony-based character optimization or tree reconciliation. Any incongruences in the mapping of the phylogenies can be attributed to one of the various coevolutionary scenarios mentioned earlier. Each of the two event-based methods have evolved over the last few decades and featured the detection (or lack of detection) of a variety of the coevolutionary scenarios.

Global fit methods, on the other hand, do not focus so much on the non-codivergence scenarios, but instead try to "assess the global congruence between host and parasite trees" and "identify individual host-parasite associations contributing to the cophylogenetic structure". (Desdevises) The methods evaluate the level of congruence between the two phylogenies to say something about the prevalence of codivergence: higher congruence implies higher codivergence.

Each of the computational resources available for each cophylogenetic method will be discussed while weighing the pros and cons of each.

Event-Based Methods

Character Optimization

Brooks Parsimony Analysis (BPA)

BPA is one of the earlier methods of cophylogeny that came about when Brooks took Hennig's parsimony-based phylogeny reconstruction methods and used them in host-parasite cophylogeny analyses. Unlike the other event-based methods based on tree reconciliation, BPA is not model-based and thus has no assumption about model-like regularities in phylogenesis. Further, it does not try to maximize fit into a predetermined hypothesis.

BPA takes the parasite phylogeny to be analyzed and converts it into a set of additive binary characters, which are then mapped onto the host tree via parsimony. The binary code defines each taxon by all the tree nodes representing its identity. Using those binary codes, a host phylogeny is constructed representing the theoretical phylogenetic relationships between the host taxa. That host phylogeny can then be compared to independently determined evidence - for example independently identified host relationships or geological and biogeographical evidence. In order to do the comparisons, a measure such as the consistency index is needed to determine homoplasy. Any characters that do not fit the host tree - known as homoplasious characters - can be interpreted as either host switching or duplication coevolutionary events.

Tree Reconciliation

In general, tree reconciliation methods function by mapping the parasite phylogeny to the host phylogeny. Any resulting incongruences in the mapping are then reconciled by attribution to coevolutionary events.

Component

Component was the first of the tree reconciliation methods to be developed. Started by Page, Component works by implementing several tree comparison methods, including computing consensus trees, calculating the similarity between pairs of trees, and mapping one tree onto

another. Measures such as the partition metric and quartet measures allow quantification of congruence between the host and parasite phylogenies.

The most useful part of Component and tree reconciliation methods in general is their tree-mapping abilities. The ability to compute a tree that reconciles incongruences between host and parasite trees is particularly useful. One disadvantage of Component is that it does not account for host switching as a coevolutionary event. The majority of incongruences are reconciled as duplications or lineage sorts.

TreeFitter

One of the first improvements of the tree reconciliation methods came about because of multiple solutions to the problem. Reconciling the host-parasite phylogenetic mappings often produced multiple possible reconstructions of their relationship. Having all these solutions made it difficult to determine which was representative of the actual relationship. The inclusion of cost-event-based analyses in the tree reconciliation method addressed this problem.

TreeFitter was one of the first programs to incorporate event-cost analyses into its parsimony-based tree fitting. It can take arbitrary cost assignments so that duplication, sorting, and host-switching events have a zero or positive cost association. Codivergence events can have positive, negative, or zero cost. This event-cost assignment system allows a tree reconciliation method to first return results more likely to be representative of actual phylogenetic relationships.

TreeMap

The next major development in tree reconciliation analysis was the development of TreeMap by Page. TreeMap is a direct descendent of the also Page-produced Component. The major update

in the program is the incorporation of the host-switching event as a possible explanation for incongruent host-parasite phylogenies.

The most current incarnation of TreeMap, TreeMap 2.02 adds even more features to the program. An important addition was the use of the Jungle algorithm to account for holes in the previous programs' analyses. Previously, for example, programs ignored host-switches that happened to be followed by sorting events. The new algorithm accounted for this and other previous lapses.

Another important feature of the current TreeMap is that the user can set bounds for returning feasible reconstructions. For example, being able to specify the maximum number of host switching events will eliminate some of the more unlikely biological reconstruction scenarios. The different coevolutionary event scenarios can also be weighted based on likelihood of occurrence. The result is a list of reconstructions that - in addition to the event-cost analysis - is even more representative of feasible host-parasite phylogenetic relationships.

Global-Fit Methods

As mentioned before, global-fit methods do not propose evolutionary scenarios like event-based methods do. Instead, they use statistical methods to merely assess the level of congruence between host and parasite phylogenies and identify specific associations that contribute to cophylogeny. An important aspect that global-fit methods add is that they take into account the possibility of error or in adequacy in the original phylogeny reconstructions. Using a partition homogeneity test, the probability that incorrect phylogenies are involved is calculated. In contrast, the assumption in event-based methods is that the phylogenic trees to be analyzed are complete and sound.

Some of the straightforward congruence tests include the Kishino-Hasegawa (K-H) test and the incongruence-length difference (ILD) test. The latter importantly assesses phylogenetic homogeneity of DNA sequences (of the host and parasites being studied) gathered from a possible variety of sources.

Maximum Likelihood

Huelsenbeck proposed two different null hypothesis tests that examine whether host and parasite phylogenies are identical. The first approach, the maximum likelihood approach, uses a likelihood ratio test, evaluating how likely they are identical versus how likely they are not. The second approach, the maximum posterior probability approach, uses Bayesian inference to directly calculate the posterior probabilities of the host and parasite phylogenies.

ParaFit

ParaFit is the latest statistical cophylogeny test, developed by Legendre. ParaFit evaluates the global hypothesis of host-parasite coevolution with a matrix permutation test of codivergence. It uses three types of information to describe the situation in matrix form: the parasite phylogeny, the host phylogeny, and a set of the observed host-parasite associations. Along with evaluating the level of congruence in the host-parasite coevolution, ParaFit allows each host-parasite association to be marked for later, more specific investigation.

CONCLUSION

Cophylogenetic methods end up affecting not only the study of host-parasite relations but many analogous relationships as well. Possible coevolution of viruses, for example, with human hosts can provide insight into medical research. Cophylogenetic methods can also be compared to studies of gene divergence across various species. Understanding of the evolution of proteomes

allows better understanding of overall organism function. Also, cophylogenetic methods are also relevant in the realm of biogeography, where better understanding of species' evolution with respect to different geographical areas can be attained.

As can be seen in the wide variety of cophylogenetic study methods, there is much debate on which method gives the best overview of a host-parasite relationship. Each methodology has its own pros and cons so they must be evaluated when choosing the best method for a specific study. Parsimony-based character optimization methods are great, simple methods for analysis though they overlook some of the coevolutionary scenarios. Event-based methods seem like the most logical approach to a problem and can take into account all scenarios. The overflow of results in a given analysis, however, makes it difficult to sift through and find the actual relationship. As a con, both of the two methods don't take into account potential phylogenic sampling and reconstruction error. Global-fit statistical methods provide a good overview of congruence in a relationship as well as taking into account phylogenic error. Their lack of linking with specific coevolutionary scenarios, however, leaves something to be desired. Overall, there is much room for improvement in all the types of methods, and if there is some way to combine the advantages of all the methods, cophylogenetic analyses of all phylogenetic relationships will greatly benefit.

REFERENCES

Banks, Jonathan C. and Adrian M. Paterson. "Multi-host parasite species in cophylogenetic studies." *International Journal of Parasitology* 35 (2005): 741-746.

Brooks, Daniel R. "Analysis of Host-Parasite Coevolution." *International Journal of Parasitology* 17.1 (1987): 291-297.

Brooks, Daniel R. "Hennig's Parasitological Method: A Proposed Solution." *Systematics Zoology* 30.3 (1981): 229-249.

Charleston, M. A. "Jungles: A new solution to the host/parasite phylogeny reconciliation problem." *Mathematical Biosciences* 149.2 (1998): 191-223.

Charleston, Michael A. "Principles of cophylogenetic maps." *Biological Evolution and Statistical Physics* 585 (2002): 122-147.

Charleston, Michael A. "Recent Results in Cophylogeny Mapping." *Advances in Parasitology* 54 (2003): 304-330.

Charleston, Michael A. "Traversing the tangle: Algorithms and applications for cophylogenetic studies." *Journal of Biomedical Informatics* 39 (2006): 62-71.

Desdevises Y. "Cophylogeny: insights from fish-parasite systems." *Parassitologia* 49 (2007): 125-128.

Legendre, Pierre, et al. "A statistical Test for Host-Parasite Coevolution." *Systematic Biology* 51.2 (2002): 217-234.

Libeskind-Hadas, Ran and Michael A. Charleston. "On the Computational Complexity of the Reticulate Cophylogeny Reconstruction Problem." *Journal of Computation Biology* 16.1 (2009): 105-117.

Page, Roderic D. M. and Michael A. Charleston. "From Gene to Organismal Phylogeny: Reconciled Trees and the Gene Tree/Species Tree Problem." *Molecular Phylogenetic Evolution* 7.2 (1997): 231-240.

Paterson, Adrian M. and Jonathan Banks. "Analytical approaches to measuring cospeciation of host and parasites: through a glass, darkly." *International Journal for Parasitology* 31 (2001): 1012-1022.

Stevens, Jamie. "Computational aspects of host-parasite phylogenies." *Briefings in Bioinformatics* 5.4 (2004): 339-349.