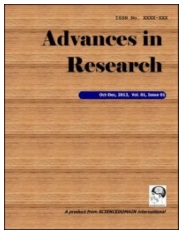

Advances in Research
2(3): 145-152, 2014, Article no. AIR.2014.003

SCIENCEDOMAIN *international*
www.sciencedomain.org

Rooted *Phylogenetic* Networks for Exploratory Data Analysis

David A. Morrison[1*]

[1]*Department of Biomedical Sciences and Veterinary Public Health, Swedish University of Agricultural Sciences, 75189 Uppsala, Sweden.*

Author's contribution

The author DAM conceived the idea, and wrote, read and approved the manuscript.

Method Article

Received 18th December 2013
Accepted 6th February 2014
Published 22nd February 2014

ABSTRACT

Rooted evolutionary networks have previously been used solely for representing explicit hypotheses of evolutionary history (i.e. a phylogeny). However, they also have potential for exploratory data analysis. An example is presented here that highlights this use. This involves a study of possible transmission histories of a nematode parasite among cattle farms in Sweden. The network adds considerably to the practical information that can be gleaned from a study of the transmission of the pathogen, thus emphasizing the practicality of this use of *phylogenetic* networks for exploratory analysis.

Keywords: Exploratory data analysis; phylogenetic networks; evolutionary networks.

1. INTRODUCTION

*Phylogenetic*s is the study of the genealogical history of the ancestry and descent of organisms. Reconstructing the evolutionary history of a group of contemporary organisms, called a phylogeny, is based on a study of their genotypic and phenotypic attributes. Each evolutionary event creates novel attributes, and it is the pattern of shared attributes among the organisms that provides evidence of those events.

**Corresponding author: E-mail: David.Morrison@slu.se;*

The phylogeny is represented as a connected line graph. It will be a tree-like graph if the genealogical history has been dominated by so-called vertical evolutionary processes, involving the passage of hereditary information directly from parent to offspring. Otherwise, it will be a network, with reticulations representing so-called horizontal processes, such as recombination, hybridization, introgression, horizontal gene transfer and genome fusion, all of which transfer genetic information in more complex ways than by simple inheritance [1]. *Phylogenetic* networks are explained in more detail by Morrison [2].

There are two types of *phylogenetic* network [1,2]: (i) rooted evolutionary networks, in which the internal nodes represent inferred ancestors of the leaf nodes (which represent contemporary organisms), and the directed edges represent historical pathways of transfer of genetic information between ancestors and their descendants; and (ii) unrooted affinity networks, in which the internal nodes do not represent ancestors, and the undirected edges represent similarity relationships among the leaf nodes.

The latter are becoming quite common in the literature. They are used principally as a form of exploratory data analysis (EDA) [3], in which the evolutionary events are studied without imposing any pre-conceived notions about the structure of the phylogeny [4-7]. EDA traditionally involves both graphical displays of the data and numerical summaries of the data [8]. The objective is not just to summarize the patterns in the data but also to visualize them in a way that is meaningful in a *phylogenetic* context. These ways should highlight features of the data that are relevant to the question at hand, as well as identifying any potential problems that should be investigated further.

Rooted evolutionary networks are usually intended to explicitly represent an hypothesis of the genealogical history [9]. However, it is possible that they might also be useful for exploratory data analysis, in addition to the unrooted networks. For example, Alroy [10] constructed a hybridization network in order to remove spurious edges from a *phylogenetic* tree, which were due to phenotypic convergence. He noted (p. 163) that “the use of reticulations clarifies the phylogeny by factoring out apparent convergence, even though there is no reason to think that actual hybridization or introgression has occurred”.

The essential difference between the rooted and unrooted networks is that the former have nodes that represent events, and the directed edges represent the temporal order in which those hypothesized events occurred. The original intention for developing the rooted networks was that the “events” would involve horizontal evolutionary processes such as hybridization and recombination. However, this is not actually a requirement, and the events could be any one of many phenomena. For example, they could be other types of phenotypic change (such as convergence, as used by Alroy [10]), or they could be geographical processes, such as dispersal.

Here, I explore this idea further, by presenting a very different example of the use of a rooted evolutionary network as a heuristic tool, for exploring a particular empirical dataset in which the representation of time is an important feature. In evolutionary analysis, network reticulations are interpreted as representing horizontal evolutionary processes. However, in the work of Alroy [10] the network reticulations were interpreted as convergence, and in my example they are interpreted as movements of parasites among hosts.

2. EXAMPLE ANALYSIS

2.1 Background

In infection biology and epidemiology, one main interest is in the transmission of pathogens from one host to another, possibly in geographically distant locations [11]. It is usually assumed that when pathogens (*viruses*, *bacteria*, *protists*, *microfungi*, *helminths*) with the same genotype are found in different locations this represents transmission from a single source location. Conversely, a mixture of genotypes at a single location is assumed to represent multiple sources of infection at that location, possibly at different times. This type of analysis is a combination of population genetics and *phylogenetics*.

Such transmission studies can produce quite complex results, even to the extent of having different pathogen genotypes simultaneously in the same host. Data analysis is usually based on either an unrooted haplotype network or a rooted tree [12], but it can also conveniently be studied using a rooted reticulation network. Here, I am investigating the latter option.

2.2 Methods

The data analyzed here represent the mitochondrial genotypes of an endoparasite species. They are 1,544 aligned nucleotides from 72 samples of the nematode *Dictyocaulus viviparus*, which is the parasitic lungworm of domestic cattle.

The data are concatenated genes of two mitochondrial proteins (Cox3, Nad5), plus mitochondrial large-subunit rRNA and tRNA gene sequences. Höglund et al. [13] describe the data collection and processing. Many of the data patterns described below are repeated across several of these genes, but some of them are unique to a single gene. This means that multiple gene sequences are needed in order to study the epidemiology of this nematode species.

The experiment involves relationships among 64 worm samples from Sweden: 8 worms from each of 8 farms (Farms 29, 34, 36, 38, 49, 65, 68 and 76), plus 8 samples from an isolate that had been maintained for several years in the laboratory at Intervet in the Netherlands (labeled L). The latter is used as the out group to root the network. The nucleotide sequence data are taken from GenBank under the accession numbers DQ299539-DQ299826.

For the rooted network, the data were analyzed using the reticulation network method of Huson & Klöpper [14], based on splits generated by the Median network [15], using the program Splits Tree 4.11.3 [16]. Since the character data are essentially binary in this dataset (with two exceptions), this happens to produce exactly the same result as for a recombination network [14].

Briefly, a median network is a splits graph [17] that displays all of the bipartitions of the set of samples that are formed by the characters. That is, an edge is added to the network for every sample bipartition formed by every character. If two (or more) characters form the same bipartition, then the relevant edge length is increased, instead. The reticulation network is formed from this splits network by first adding a root to the unrooted graph, and then finding the minimum number of reticulation nodes needed to replace each netted region (a set of independent biconnected components) with a tangle (a set of dependent

reticulations) [18]. The objective is to find compatible reticulations that can be pooled, reducing the data incompatibilities to simpler reticulation events [19].

For the unrooted network used as a comparison, the data were analyzed using the Neighbor Net algorithm [20], based on the hamming distance matrix, also using the program Splits Tree.

2.3 Results and Discussion

A common unrooted EDA network procedure is to produce a Neighbor Net network Fig. 1. This network represents the genetic similarity of the samples from the different farms — samples that cluster together are similar genetically. The graph is rather tree-like, and forms 11 distinct clusters of the sampled genotypes, plus one collection of samples at an internal node (arrowed).

In most cases these clusters reflect the different locations at which the parasites were collected, as the eight samples from each of Farms 34, 36, 38 and 76 each form single clusters, as do the eight laboratory samples (L). However, the other farms have more complex patterns. The eight samples from Farm 29 form two separate clusters (of five and three samples), while six of the samples from Farms 49 and 68 are each clustered together. However, the two remaining samples from Farm 49 are clustered with the samples from Farm 76, and the other two samples from Farm 68 are clustered with the samples from Farm 38. Finally, the samples from Farm 65 are placed in four separate locations in the network.

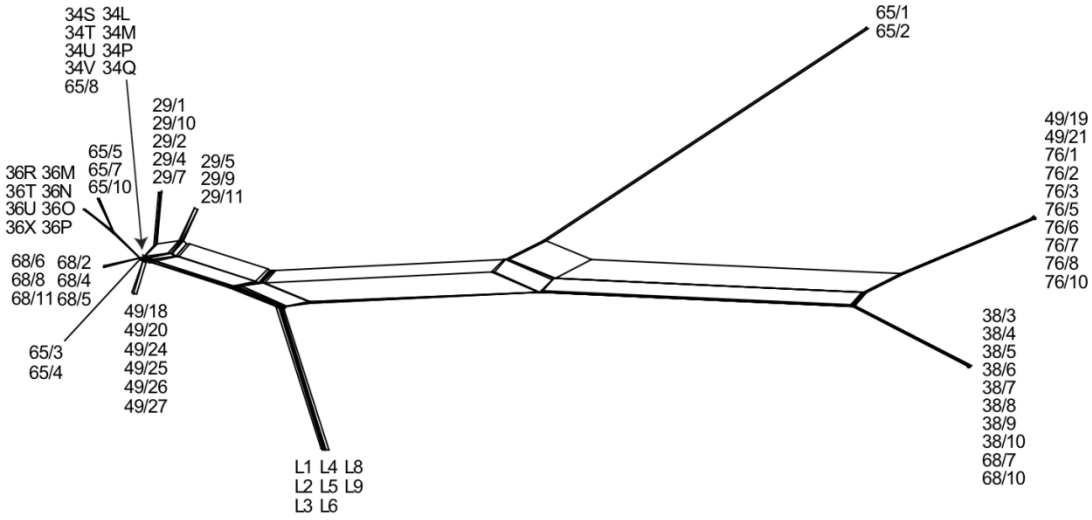

Fig. 1. Neighbor Net graph of the unrooted similarity relationships among the samples

Relationships among 72 samples of the nematode Dictyocaulus viviparus, from eight farms and one laboratory sample, based on nucleotide sequences for four genes. The first two digits of each label represent the farm from which the sample was taken.

The genetic similarity of the samples from different farms might reflect transmission of the pathogens from one farm to another, but the network is unclear about the farm relationships. In particular, without a direction to the network edges there is no indication of the hypothesized transmission history. For this purpose, a rooted network is required.

In the rooted network of the same data Fig. 2, there is a clear indication of hypothesized history. Most of the samples from within each farm seem to be closely related in a simple divergent fashion through time, as would also be conveniently displayed by a standard tree-based *phylogenetic* analysis. There are apparently two major clades of genotypes, with 6-7 subclades. (A clade is a group with a single evolutionary origin.) We can conclude from the tree-like relationships that four farms show evidence of only a single source of infection (i.e. Farms 34, 36, 38 and 76 each have only a single genotype), while two farms appear to have at least two genotypes and thus probably two sources of infection (Farms 49 and 68).

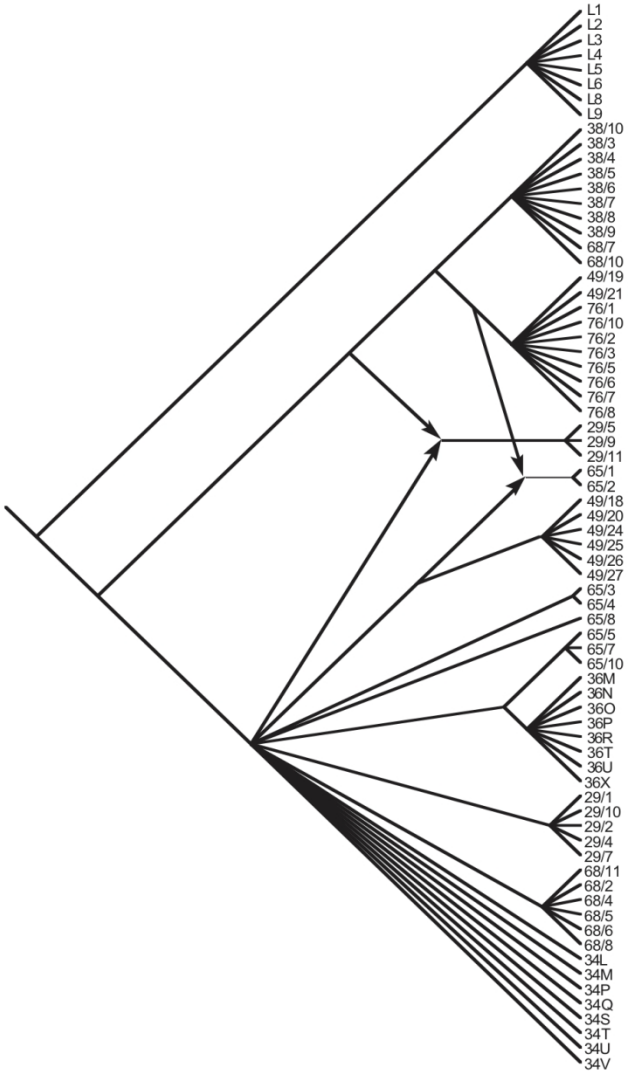

Fig. 2. Rooted evolutionary network for the samples

See Fig. 1 for an explanation. All of the edges without arrows have an implicit direction away from the root, which is at the left.

However, the other two farms show more complex patterns than this, which would not be revealed by a simple tree-based analysis. These two farms have groups of samples that descend from reticulation nodes (indicated by the arrows in Fig. 2, thus suggesting the pooling of two distinct sources of genetic material. Note that there is no suggestion that these reticulations represent either recombination or hybridization, given that the data are from mitochondrial genes (which are usually considered to be inherited as a single unit). This analysis is best treated as exploratory (EDA), highlighting within-farm genotypic complexity that warrants further biological investigation, rather then providing an explicit hypothesis of evolutionary history.

Farm 29 is shown as having one unique genotype (5 individuals) plus another genotype (3 individuals) that has elements possibly related to both of the major clades of genotypes. Perhaps these latter 3 individuals represent an earlier infection, given their apparent association with the basal branches of the two clades.

Farm 65 appears to be even more noteworthy. There are 3 individuals that are apparently related to those on Farm 36, plus 3 individuals of somewhat uncertain relationship. Then there are 2 individuals with elements possibly related to the genotypes on Farms 76 and 49. This is clearly a very interesting farm, from the point of view of lungworm infection and transmission, with at least three possible infection sources. This is important information that needs to be taken into account for possible strategies of managing the lungworm infections on the farms.

3. GENERAL DISCUSSION

I have used this empirical example to illustrate the use of a rooted evolutionary network in epidemiology. The outcome of the analysis is a set of hypotheses about the historical spread of the nematodes among the Swedish farms. It has not been intended as an explicit phylogeny, but as an exploratory data analysis. It generates hypotheses that deserve further investigation, biologically, in any study of the transmission of the pathogen.

The use of *phylogenetic* methods in the study of infection biology and epidemiology is becoming more widespread [21]. This has particularly been so in virology, to study the origin and spread of epidemics associated with new viral strains [22]. These methods have focused on the use of rooted *phylogenetic* trees as a summary of evolutionary information. However, if the evolutionary history is more complex, then a network is more appropriate, as shown by the example presented here.

The use of evolutionary networks for EDA is not restricted to the field of epidemiology, as this is simply an opportune example. It is possible to produce a *phylogenetic* analysis of any group of objects that vary in their intrinsic characteristics, and where those characteristics can be inferred to vary through time. If one wishes to generate hypotheses based on those temporal patterns, then use of an evolutionary network will be a viable method of data analysis.

I have used one particular algorithm to produce the evolutionary network in my example, although there are potentially many available. At the moment, evolutionary networks are an active area of algorithm development [9], and the one used here was chosen for convenience, as it is included in a software package. Unfortunately, there are currently few available programs. As noted by Huson et al. [23]: “there are many promising directions to follow and rudimentary software implementations, [but] there is no tool currently available

that biologists could easily and routinely use on real data." This will change in the near future.

4. CONCLUSION

Rooted evolutionary networks have heretofore been used almost solely for representing explicit hypotheses of evolutionary history (i.e. a phylogeny), in which the reticulations represent horizontal genotypic events. The use of a rooted network analysis for exploratory data analysis seems to have rarely been considered. Nevertheless, the example shown here highlights the potential use of these networks for EDA, as well, in which the reticulations represent other historical processes, such as phenotypic changes or geographical movements. In the example presented, the network added considerably to the practical information that could be gleaned regarding the transmission of the pathogen. Other uses can easily be imagined.

ACKNOWLEDGEMENTS

Thanks to Leo van Iersel for prompting me to undertake this analysis.

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. Morrison DA. *Phylogenetic* networks are fundamentally different from other kinds of biological networks. In Zhang WJ, editor. Network biology: theories, methods and applications, New York: Nova Science Publishers. 2013;23-68.
2. Morrison DA. Introduction to *phylogenetic* networks. Uppsala: RJR Productions; 2011.
3. Morrison DA. *Phylogenetic* networks — a new form of multivariate data summary for data mining and exploratory data analysis. WIREs Data Mining Know Discov *(*In press*);* 2014.
4. Bandelt H-J. Exploring reticulate patterns in DNA sequence data. In: Bakker FT, Chatrou LW, Gravendeel B, Pelser PB, editors. Plant species-level systematics: new perspectives on pattern and process, Königstein; Koeltz. 2005;245-269.
5. Wägele JW, Mayer C. Visualizing differences in *phylogenetic* information content of alignments and distinction of three classes of long-branch effects. BMC Evol Biol 2007;7:147.
6. Wägele JW, Letsch H, Klussmann-Kolb A, Mayer C, Misof B, Wägele H. *Phylogenetic* support values are not necessarily informative: the case of the *Serialia* hypothesis (a mollusk phylogeny). Frontiers Zool 2009;6:12.
7. Morrison DA. Using data-display networks for exploratory data analysis in *phylogenetic* studies. Mol Biol Evol 2010;27:1044-1057.
8. Tukey JW. Exploratory data analysis. Reading (MA): Addison-Wesley; 1977.
9. Morrison DA. *Phylogenetic* networks: a review of methods to display evolutionary history. Ann Res Rev Biol (in press); 2014.
10. Alroy J. Continuous track analysis: a new *phylogenetic* and biogeographic method. Syst Biol 1995;44:152-178.
11. Nelson KE, Williams CM, editors. Infectious disease epidemiology: theory and practice, third edition. Burlington (MA): Jones & Bartlett Learning; 2013.

12. Ypma RJ, van Ballegooijen WM, Wallinga J. Relating *phylogenetic* trees to transmission trees of infectious disease outbreaks. Genetics 2013;195:1055-1062.
13. Höglund J, Morrison DA, Mattsson JG, Engström A. Population genetics of the bovine/cattle lungworm (*Dictyocaulus viviparus*) based on mtDNA and AFLP marker techniques. Parasitology 2006;133:89-99.
14. Huson DH, Klöpper TH. Beyond galled trees — decomposition and computation of galled networks. Lect Notes Bioinform 2007;4453:211-225.
15. Bandelt HJ. *Phylogenetic* networks. Verhand Naturwiss Vereins Hamburg 1994;34:51-71.
16. Huson DH, Bryant D. Application of *phylogenetic* networks in evolutionary studies. Mol Biol Evol 2006;23:254-267.
17. Dress AWM, Huson DH. Constructing splits graphs. IEEE/ACM Trans Comput Biol Bioinform 2004;1:109-115.
18. Huson DH. Split networks and reticulate networks. In: Gascuel O, Steel M, editors. Reconstructing evolution: new mathematical and computational advances, pp. 247-276. Oxford: Oxford Uni. Press; 2007.
19. McBreen K, Lockhart PJ. Reconstructing reticulate evolutionary histories of plants. Trends Plant Sci 2006;11:398-404.
20. Bryant D, Moulton V. Neighbor Net: an agglomerative method for the construction of planar *phylogenetic* networks. Lect Notes Comp Sci. 2002;2452:375-391.
21. Holmes EC, Bollyky OL, Nee S, Rambaut A, Garnett GP, Harvey PH. Using *phylogenetic* trees to reconstruct the history of infectious disease epidemics. In Harvey PH, Leigh Brown AJ, Maynard Smith J, Nee S, editors. New uses for new phylogenies, Oxford: Oxford University Press. 1996;169-186.
22. McCormack GP, Clewley JP. The application of molecular *phylogenetics* to the analysis of viral genome diversity and evolution. Rev Med Virol 2002;12:221-238.
23. Huson DH, Moulton V, Steel M. Special Section: *Phylogenetics*. IEEE/ACM Trans Comput Biol Bioinform 2009;6:4-6.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=429&id=31&aid=3785