Spotlight

Cell PRESS

Networks: expanding evolutionary thinking

Eric Bapteste[1], Leo van Iersel[2], Axel Janke[3], Scot Kelchner[4], Steven Kelk[5], James O. McInerney[6], David A. Morrison[7], Luay Nakhleh[8], Mike Steel[9], Leen Stougie[2,10], and James Whitfield[11]

[1] Université Pierre et Marie Curie, Paris, France
[2] Centrum Wiskunde and Informatica, Amsterdam, The Netherlands
[3] Goethe University, Frankfurt am Main, Germany
[4] Idaho State University, Pocatello ID, USA
[5] Maastricht University, Maastricht, The Netherlands
[6] National University of Ireland, Maynooth, Ireland
[7] Sveriges Lantbruksuniversitet, Uppsala, Sweden
[8] Rice University, Houston TX, USA
[9] University of Canterbury, Christchurch, New Zealand
[10] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[11] University of Illinois, Urbana IL, USA

Networks allow the investigation of evolutionary relationships that do not fit a tree model. They are becoming a leading tool for describing the evolutionary relationships between organisms, given the comparative complexities among genomes.

Beyond trees

Ever since Darwin, a phylogenetic tree has been the principal tool for the presentation and study of evolutionary relationship among species. A familiar sight to biologists, the bifurcating tree has been used to provide evidence about the evolutionary history of individual genes as well as about the origin and diversification of many lineages of eukaryotic organisms. Community standards for the selection and assessment of phylogenetic trees are well developed and widely accepted. The tree diagram itself is ingrained in our research culture, our training, and our textbooks. It currently dominates the recognition and interpretation of patterns in genetic data.

However, many patterns in these data cannot be represented accurately by a tree. The evolution of genes in viruses and prokaryotes, of genomes in all organisms, and the inevitable noise that creeps into phylogenetic estimations, will all create patterns far more complicated than those portrayed by a simple tree diagram. Genetic restructuring and non-vertical transmission are largely overlooked by a methodological preference for phylogenetic trees and a deep-rooted expectation of tree-like evolution.

A way forward is to recognize that, mathematically, tree graphs are a subset of the broader space of general graphs (henceforth: networks). Trees are optimized, pared-down visualizations of often more complex signals. When confined to trees, we overlook additional dimensions of information in the data [1–4]. By moving beyond the exclusive use of trees, and adopting a routine application of networks to genetic data, we can expand the scope of our evolutionary thinking.

Corresponding authors: Kelchner, S. (kelchner@isu.edu); Whitfield, J. (jwhitfie@life.illinois.edu).

The future of phylogenetic networks

From 15–19 October 2012 a community of mathematicians, computer scientists, and biologists met to consider 'The Future of Phylogenetic Networks' at the Lorentz Center in Leiden, The Netherlands. The purpose of the meeting was to enhance the dialog between biologists and developers of network theory and methodology to align better the proliferation of network tools to the specific needs of evolutionary biologists. The successes and limitations of network analysis were presented and discussed, and outstanding problems in network mathematics and computer implementations were identified.

It was clear from the presentations that network methodology has advanced sufficiently to be of widespread use to biologists. Although recent textbooks on the subject [5,6] and user-friendly software [7–9] are broadening the appeal and application of network principles, not that many biologists have yet adopted network analysis. To encourage researchers to expand beyond historic tree-thinking it is important to demonstrate the advantages of modern network-thinking.

Genetic data are not always tree-like

Evolutionary networks today are most often used for population genetics, investigating hybridization in plants, or the lateral transmission of genes, especially in viruses and prokaryotes. However, the more we learn about genomes the less tree-like we find their evolutionary history to be, both in terms of the genetic components of species and occasionally of the species themselves.

A wide variety of evolutionary processes lead to mosaic patterns of relationships among taxa: sex in eukaryotes, recombination in its variety of forms, gene conversion between paralogs, intron retrohoming, allopolyploidization, partial non-orthologous replacement, the selection of new genetic assemblages leading to modular entities as in operon formation, the emergence of new families of transposons, independent lineage-sorting among alleles, and unequal rates of character loss between lineages, among others (Table 1). Reticulate patterns can also

Table 1. The pay-offs of network-based studies[a]

	Phylogenetic Tree	Phylogenetic network	Similarity network [1,14]
Data display	*A priori* highly constrained: acyclic connected graph	*A priori* constrained: acyclic or cyclic connected graph	*A priori* less constrained: acyclic or cyclic, connected or disconnected graph
Evolutionary scope	Conserved families of homologs (e.g., of aligned sequences)	Conserved families of homologs (e.g., of aligned sequences)	Conserved and/or expanded families of homologs (e.g., of aligned sequences and their distant homologs), and composite families (e.g., component and composite sequences)
Focus	1 process (vertical descent) or averaging of *n* processes	≥1 process (vertical descent and introgressive descent)	≥1 process (vertical descent and introgressive descent)
Objects of study	Groups of non-mosaically related entities, sharing a last common ancestor (e.g., clades)	Groups of non-mosaically related entities, and/or of mosaically related entities (e.g., clades and hybrids)	Groups of non-mosaically related entities, mosaically related entities, and/or of mosaically unrelated entities (e.g., clades, hybrids, and coalitions)

[a]The use of networks enriches data display, allowing the elaboration and testing of a greater number of evolutionary hypotheses. It also enhances the scope of evolutionary analyses because distant homologies, additional objects of studies, and multiple processes can be represented and compared in a network framework.

emerge from improper data processing and analysis, such as model misspecification, data management error, and poor alignment of sequences.

Although many single-gene datasets might produce a tree unaffected by these processes, it is less likely that multiple genes in a combined dataset would do so. In the context of the special problems presented by phylogenomic data, members of the Leiden meeting discussed a recent call from *Nature* for greater accuracy in analyzing and interpreting genomic data [10]. Tree-based genomic analysis is proving to be an accuracy challenge for the evolutionary biology community, and although genome-scale data carry the promise of fascinating insights into tree-like processes, non-treelike processes are commonly observed. Network analysis is a readily available and ideal tool that reduces the danger of misinterpreting such data.

Tackling error with networks

There are long-standing controversies regarding the evolutionary history of many taxonomic groups, and it has been expected by the community that genome-scale data will end these debates. However, to date none of the controversies has been adequately resolved as an unambiguous tree-like genealogical history using genome data. This is because quantity of data has never been a satisfactory substitute for quality of analysis. Many of the underlying data patterns are not tree-like at all, and the use of a tree model for interpretation will oversimplify a complex reticulate evolutionary process.

A pertinent example is the 2003 genomic dataset [11] from yeast (*Saccharomyces*) which has proved problematic for tree thinking. It involves a large amount of heterogeneity among the 106 individual gene trees, which leads to unreliability in the estimate of the species tree. Many tree-based approaches to resolving the evolutionary analysis have been tried, but with little success: the resulting trees are sensitive to data-coding methods and the model of sequence evolution used, and there seem to be no identifiable parameters to predict systematically the phylogenetic signal within and among genes. In this case a species tree becomes only a mathematical average estimate of evolutionary history, and even if it is supported it suppresses conflicting phylogenetic signals. Network thinking better accommodates the multiple evolutionary processes involved in these genetically mosaic entities. Importantly, network analysis has provided the insight that genome hybridization is a much more likely explanation for the differences between gene trees in the *Saccharomyces* dataset [12].

Another case is the inference of the early branching order in placental mammalian evolution, a problem that has been difficult to resolve as a bifurcating process because different genetic datasets support different trees. In particular, the question as to which one of the three placental mammalian groups, Afrotheria (e.g., elephant, manatees, hyraxes), Xenarthra (e.g., armadillos, anteaters), or Boreoplacentalia (e.g., human, mouse, dog), represents the first divergence among placental mammals has long vexed mammalian systematics. Different sets of molecular data have placed each of the three major groups as a sister group to the others. Even genome-scale analyses of more than one million amino acid sites from orthologous protein-coding genes have not rejected any of the three alternatives, despite the statistical estimate that 20 000 amino acid sites should be sufficient to resolve the question at this level of divergence given the tree structure, branch lengths, and number of substitutions. By contrast, a network analysis of retroposon insertion data provides an alternative hypothesis for the history of placental mammals: owing to incomplete lineage sorting and hybridization in the early placental mammalian divergences, the evolutionary history of placental mammals is network-like and far more intricate than a simple tree can show [13].

In both of these examples the network provides biological explanations that go beyond what can be accommodated by a simple tree model. More examples are now available in diverse taxonomic groups and they should inspire evolutionary biologists to explore networks in a much more systematic way.

Opportunities and challenges

The further improvement of networks for evolutionary biology offers many outstanding opportunities for mathematicians, statisticians, and computer scientists. Several developments were showcased at the Leiden meeting, including: (i) theoretical work addressing the extent to

which random lateral gene-transfer will either recover or obliterate signal for a central-tendency species tree; (ii) statistical methods to distinguish genuine reticulate evolution, such as hybridization, from other non-reticulate processes, such as incomplete lineage sorting; and (iii) a mathematical understanding of the number of reticulations needed to reconcile two conflicting gene trees. A network can be both a more parsimonious description of the amount of discordance between genes, and a starting point for generating hypotheses to explain that discordance. An important subject of ongoing research is to understand how far networks over-estimate the true amount of reticulate pattern in datasets.

For mathematicians, the field is ripe for advances. For evolutionary biologists, networks already provide an invaluable complement to trees that are likely to increase in robustness and importance over the next few years.

However, biologists must also keep in mind that networks are not yet free of interpretive challenges. One must knowledgeably select from the various types of network methods available to interpret properly such features as internal nodes and the meaning of taxon groupings, which differ in important ways among methods. Furthermore, community standards do not yet exist for network assessment and interpretation. As with tree methods, the responsibility remains with the researcher to understand network methodology, apply it correctly, and make valid inferences.

These challenges do not detract from the fact that networks represent an historic juncture in the development of evolutionary biology: it is a shift away from strict tree-thinking to a more expansive view of what is possible in the development of genes, genomes, and organisms through time. Something of an esoteric academic pursuit in the past, networks are now poised to become a widely used and effective tool for the analysis and interpretation of evolution.

References

1 Bapteste, E. *et al.* (2012) Evolutionary analyses of non-genealogical bonds produced by introgressive descent. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18266–18272
2 Dopazo, J. *et al.* (1993) Split decomposition: a technique to analyze viral evolution. *Proc. Nat. Acad. Sci. U.S.A.* 90, 10320–10324
3 Rivera, M.C. and Lake, J.A. (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431, 152–155
4 McBreen, K. and Lockhart, P. (2006) Reconstructing reticulate evolutionary histories of plants. *Trends Plant Sci.* 11, 398–404
5 Huson, D.H. *et al.* (2010) *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press
6 Morrison, D.A. (2011) *Introduction to Phylogenetic Network*. RJR Productions
7 Huson, D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14, 68–73
8 Than, C. *et al.* (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9, 322
9 Huson, D.H. and Scornavacca, C. (2012) Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* 61, 1061–1067
10 Editorial (2012) Error prone. *Nature* 487, 406
11 Rokas, A. *et al.* (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804
12 Yu, Y. *et al.* (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet.* 8, e1002660
13 Hallström, B.M. and Janke, A. (2010) Mammalian evolution may not be strictly bifurcating. *Mol. Biol. Evol.* 27, 2804–2816
14 Dagan, T. *et al.* (2008) Modular networks and cumulative impact of lateral gene transfer in prokaryote genome evolution. *Proc. Nat. Acad. Sci. U.S.A.* 105, 10039–10044

0168-9525/$ – see front matter
http://dx.doi.org/10.1016/j.tig.2013.05.007 Trends in Genetics xx (2013) 1–3