African Journal of Biotechnology Vol. 10(60), pp. 12799-12803, 5 October, 2011
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB11.010
ISSN 1684–5315 © 2011 Academic Journals

Review

Phylogenetic networks: A tool to display character conflict and demographic history

Miro Ferreri[1#], Weijie Qu[2#] and Bo Han[1]*

[1]College of Veterinary Medicine, China Agricultural University, Beijing 100193, China.
[2]College of Animal Science and Technology, Yunnan Agricultural University, Kunming, Yunnan 650201, China.

Accepted 6 May, 2011

Evolutionary trees have the assumption that evolution and phylogeny can be represented in a strictly bifurcating manner. Firmly speaking, from one ancestral taxon, two descendant taxa emerge. Nevertheless, hybridization, recombination and horizontal gene transfer is in conflict with this straightforward concept. In such cases, evolutionary lines do not only separate from each other, but have the possibility of melting again and are called reticulations. Consequently, networks can represent evolutionary events more realistically than phylogenetic trees. Networks can display alternative topologies and co-existence of ancestors and descendants, which are otherwise not obvious when a comparison is done on several single trees or a consensus tree. Therefore, networks have the ability to visualize the conflicting information in a given data set. Moreover, the distribution, frequencies and arrangement of haplotypes in populations can reveal the phylogenetic histories of the taxa, regarding predictions from the coalescent theory. This review aims to: (1) give a brief comparison between phylogenetic trees and networks, (2) provide the overall concept of the coalescent theory, (3) clarify how phylogenetic networks can be used to display conflict data and evaluate phylogenetic histories, and (4) offer a useful starting point and guide for sequence analysis, with the aim to discover population dynamics.

Key words: Phylogenetic networks, reticulation, coalescent theory, population history, character conflict.

INTRODUCTION

During the last two decades of the genomic age, an overwhelming amount of genetic data was described and deposited on online data bases, such as GenBank. The prediction of a considerable data-intensive science, accumulated by high-throughput molecular technologies, can be described by the vision of a 'New Biology' discipline (Patterson et al., 2010). Nevertheless, increasing large amount of data generated through biological experiments is not useful when it is under-analyzed (Zahid et al., 2006).

Analyzing an increasing amount of genetic data is the challenge for one of the oldest fields of biology, called phylogenetics. Phylogenetics studies deal with the classification of taxa (species or populations) and their biodiversity with the quest for their evolutionary relation The relation between species (interspecies) can be described as hierarchical, with non-overlapping gene pools. This is caused by long time isolation, whereas intraspecies (for example, populations) relations are nonhierarchical, and are caused by sexual reproduction (Posada and Crandall, 2001) (Figure 1). The traditional way of describing and visualizing the evolutionary relation between taxa is a phylogenetic tree. A phylogenetic tree describes, in particular, the branching process when a species is divided into two separate species. This idea is a Darwinian concept with the assumption that genetic information is passed strictly and vertically from parents to offspring (Darwin, 1859). This theory leads to the supposition that there exist a unique 'tree of species', describing the evolutionary relations between them. The long-established form of visualizing such relations is a phylogenetic tree (Figure 2A). The popularity of trees has led to the design of many methods for their construction,

*Corresponding author. E-mail: hanbo@cau.edu.cn. Fax: +86-10-62737865. Tel: +86-10-62733801.

#These authors contributed equally to this work and share the first authorship

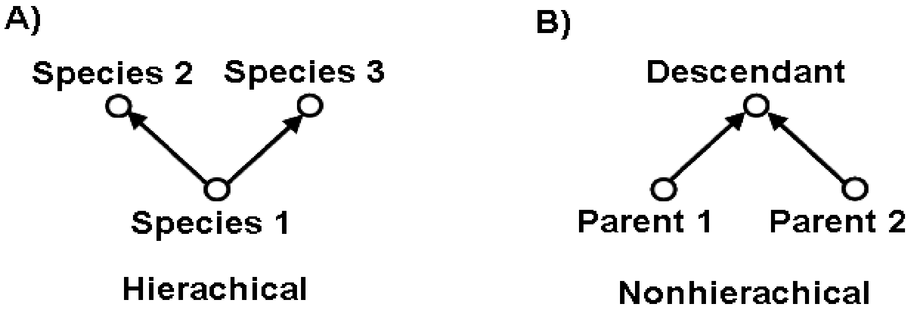

Figure 1. A) Evolutionary relationships between different species are hierarchical, as from one ancestral taxa, two descendant taxa emerge. B) In contrast, relations under the species level (intraspecies) are non-hierarchical, arising by sexual reproduction of individuals.

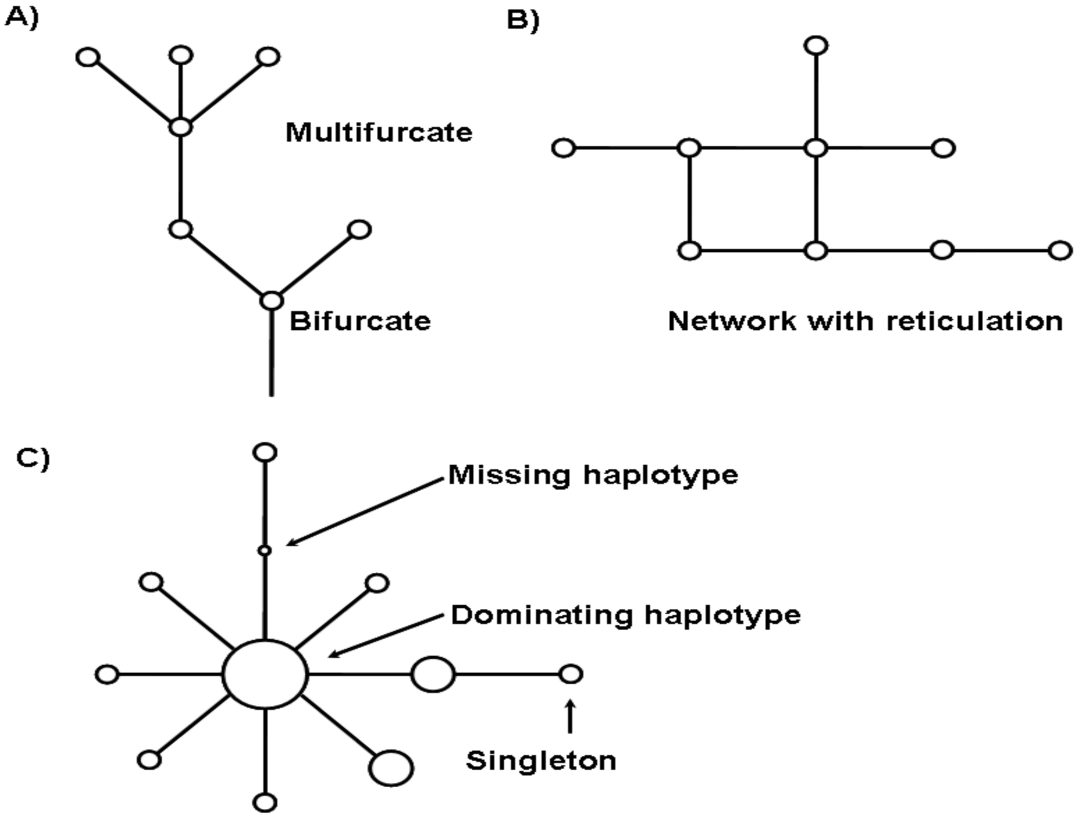

Figure 2. A) The traditional way of describing evolutionary relations between taxa is a furcating tree. B) Phlogenetics networks can display relations enhanced by a phylogenetic tree. Different parts of an information source (for example, gene or alignment part) can squabble for a different topology visualized by a loop or a cycle. C) Star-like structure of a minimum spanning network shows a numerical dominating central haplotype surrounded by several haplotypes (high haplotypes diversity), which show little differences (low nucleotide diversity). This pattern suggests that most haplotypes originate recently and is indicative of a population expansion during the recent history of the taxa, as well as, the event of the initial effect that can be hypothesized.

like the well-established neighbor joining or minimum evolution method (Saitou et al., 1987; Rzhetsky et al., 1992). More complex substitution models were developed (Kimura, 1980; Tamura, 1992) and Bayesian mathematics was integrated into the phylogenetic data analysis (Ronquist and Huelsenbeck, 2003).

The methods became more complex and sophisticated. It is about one decade ago, in which it was pointed out that the history of life cannot be properly represented as a tree (Doolittle, 1999). It has been recognized that trees oversimplify the study's view of evolution in some cases, since they can not model events such as hybrid speciation (Linder and Rieseberg, 2004), horizontal gene transfer (Bergthorsson et al., 2003; Nakamura et al., 2004) and recombination (Posada et al., 2002), which are referred to as reticulation events. These events demonstrate that a genome is not an island, which can be presented on an isolated branch of a phylogenetic tree. Moreover, reticulations break up the genomic history into different pieces, each of which has a strictly treelike pattern of descent (Maddison, 1997). This suggests that genes can have their own unique phylogenetic history. Seeing the different parts of a sequence or alignment, squabbled for a different arrangement of a tree, an obtained tree is always just a compromise that has the potential to misinterpret the phylogenetic information as great. This can be seen, as most presented trees in publications are consensus trees, and are compromised. This compromise is commonly supported due to estimation of the nodes probability by the nonparametric bootstrap procedure (Felsenstein, 1985) or Bayesian inference (Ronquist and Huelsenbeck, 2003). Alternative approaches can be performed by the Bremer support (Bremer, 1988) or the decay index (Donoghue et al., 1992). Nevertheless, an optimal supported consensus tree, by sophisticated support algorithm, is just one tree. Behind this tree, stands a forest of trees.

At this instant, the question of how the forest can be visualized or how information of different alignment parts can be simultaneously presented arises. Phylogenetic networks are able to harvest this forest and visualize it in one figure. Reticulations, indicated by loops or cycles, visualize different phylogenetic information sources (for example, different genes) by a network (Figure 2B). Therefore, it is uncertainly possible to see the phylogenetic information. The main difference of the network approaches is that they mostly work with a partition of the sequences, that is, a split, rather than the whole data set, and the network can be built by combining the splits one after the other (Bandelt and Dress, 1992). They are, definitely, a generalization of the phylogenetic trees that allow for the representation of reticulation events. This can be made clear by a reminder that a phylogenetic network, which shows no reticulations, is traditionally bifurcated and can be presented as a tree. Elsewhere, the occurrences of reticulate loops or cycles indicate a conflict in a split of the sequence. Identifying and visualizing these conflicts, as well as integrating them, is the advantage of the phylogenetic network approach.

Despite the identification of conflicts, networks are useful in the phylogeographic and phylogenetic studies, which aim to reconstruct historical events. Inferences of a possible number of networks can display ancestor and descendant haplotypes and the assumption that a mutation single-point exists in time and space. The pattern of haplotypes frequency, distribution and arrangement can reveal the phylogenetic historical events described by the coalescent theory.

This review aims to: (1) give a brief comparison between the phylogenetic trees and networks, (2) provide a theoretical background of the coalescent theory, (3) clarify how phylogenetic networks can display conflict data and evaluate phylogenetic histories, and (4) provide a useful starting point and guide for sequence analysis, with the aim to discover population dynamics.

DISPLAYING CHARACTER CONFLICT

Genetic relationships between individuals belonging to different species can be described hierarchically. They have non-overlapping gene pools as they are the product of reproductive isolation over long timescales, during which mutation combined with the population's divergence led to fixation of different alleles (Posada and Crandall, 2001). In contrast, the genetic relationships below the species level (populations) are different and can not be properly described by a furcating tree as they are not hierarchical. Individuals from different populations can mate, but the genes are not isolated from each other, in that the previously diverged lineages can be recombined again. The result is that different genes have their own original phylogenetic history. To go one step beyond, every part of an alignment can have its own unique phylogenetic history. If we want to work with this assumption, it would be plausible to cut the alignment in sets or splits, which share similarities rather than using the whole alignment.

This is exactly what the network approaches do, that is, they divide the haplotypes into exclusive sets or splits. Any data set can be partitioned into sets (not necessarily of equal size) of sequences or splits and a network can be built by combining them one after the other (Bandelt and Dress, 1992). When splits are incompatible, a loop is introduced in the network to indicate that there are alternative splits. The use of networks to visualize phylogeny has been realized by the Spectronet package (Huber et al., 2002), and the program Network 4.6.0.0. (Bandelt et al., 1999), as well as the java-based program SplitsTree v4.1. A simplified standard work flow to obtain a network would therefore be: (i) sequencing of a DNA fragment (for example, partial mtDNA or ptDNA), (ii) construction of an alignment by Clustal W v.2.0 (Larkin et al., 2007), (iii) transformation in a NEXUS format (Maddison et al., 1997) and transferring the alignment to the SplitsTree v4.1, Network 4.1.1.2 or Spectronet package to obtain a network based on the splits concept of the aligned sequences. Application of networks to the discovery of the population structures and their demographic history can be found in several recent publications (Cuc et al., 2011; Han et al., 2010; Huson and Bryant, 2006).

The conflicts of splits can be separately visualized by a Lento plot (Lento et al., 1995). "Lento plots" display, in a ranked order, support or conflict in a data set as a series of bars (each representing a split) extending above (support) or below (conflict) a horizontal line. The height or depth of each bar corresponds to the proportion of data patterns that either support or conflict with the split. The main advantage of the Lento plots is that they enable one to identify not only the amount of support and conflict for individual splits, but also display which individual or sequence is responsible. Finally, the identified splits can be used to draw a diagram for the median network (Bandelt et al., 1995, 2000). In case the collection of splits is compatible, the associated network is a tree, whereas incompatibilities give rise to reticulations.

ABILITY OF NETWORKS TO VISUALIZE PREDICTION FROM THE COALESCENT THEORY

The coalescent theory links the population genealogy (haplotypes diversity, frequency, etc.) with the demographic history of a taxon (Hudson, 1991; Felsenstein, 2004). Inferences of past events are possible because most mutations (or alleles) arise at a single point in time and space. The distribution of each new mutation, assuming neutrality, is influenced by dispersal patterns, population sizes and other processes. In short, if we now know how the recent genetic distribution looks like and make assumptions on how this distribution is influenced, we can open a small window in the history of a taxon. The necessary biological assumptions are quite intuitive, unlike the complicated mathematical description of the coalescent theory (Kingman, 1982). The following described some of the predictions of the coalescent theory: (a) High frequency haplotypes are most likely to be old alleles; (b) Within the network, old alleles are interior, whereas new alleles are more likely to be peripheral; (c) Haplotypes with multiple connections are most likely to be old alleles; (d) Old alleles are expected to show a broad geographical distribution because their carriers have had enough time to disperse them; (e) Haplotypes with only one connection (singletons) are likely to be connected to haplotypes from the same population, because they just came into existence and their carriers may not have had the time to disperse them.

These patterns can lead to several assumptions and can reveal the demographic histories by predictions from the described coalescent theory. Several results from the coalescent theory, related to the frequency and geographical distribution of the haplotypes, are relevant to intraspecific phylogenetics. There is a direct relationship between haplotype frequencies and the ages of the haplotypes (Watterson and Guess, 1977; Donnelly and Tavare, 1986). Therefore, high frequency haplotypes have probably been present in the population for a long time. Consequently, most of the new mutants are derived from common haplotypes, implying that rarer variants represent more recent mutations and are more likely to be related to common haplotypes than other rare variants (Excoffier and Langaney, 1989).

DISPLAYING DEMOGRAPHIC HISTORY

Occasionally, in the evolutionary history of a species, there are singular demographic events that can leave a lasting impression of the portioning of population genetic variation within and among populations (for example, bottlenecks, found effects, range expansion or geographical isolation, etc.). To make it simple, the his-tory of a species shapes the genetic makeup. As such, traces can be found while studying haplotype and nucleotide diversity and the haplotypes frequencies and their pattern of distribution. Most haplotypes in a population are present as copies in several individuals. If one copy mutates into a new haplotype, it is extremely unlikely that all other ancestral haplotypes are also mutated or are extincted. Strictly speaking, the ancestral haplotype can be expected to persist in the population (Posada and Crandall, 2001). The visualizing of relations between haplotypes has its limitations in a bifurcating tree; therefore, networks are the method of choice as they can present co-existence of ancestors and descendants, as well as reticulation events. In general, a network shows patterns of haplotypes distribution (Figure 2C). A network with an ancestral haplotype shows a star-like or star-burst appearance with the ancestral haplotype centered on it. Ancestral haplotypes will often give rise to multiple descendant haplotypes resulting in a multifurcating tree or star-like network. When we apply the assumptions of the coalescent theory to a network, we can predict that high frequency haplotypes have probably been present in the population for a long time. Moreover, the descendants will associate more with each other, if the haplotypes are older and are numerically dominated. Therefore, there is a direct relationship between haplotypes frequencies and the ages of haplotypes (Watterson and Guess, 1977; Donnelly and Tavare, 1986). If a star-like structure of a minimum spanning network can demonstrate that most variants of haplotypes surround the central haplotypes, then this pattern suggests that most of the haplotypes originate recently and is indicative of a population expansion during the recent history of the species (Bandelt et al., 1995). Moreover, a haplotype network can reveal not only the sampled or possible disappeared haplotypes, but also the two step- nucleotide- differences shown in Figure 2C.

CONCLUSION

The classical view of a bifurcating tree has, as a limitation, a hybrid speciation, horizontal gene transfer and recombination, which can not be demonstrated. The advantage of a network is that: (a) conflicts among different sites can be revealed and (b) persistent ancestral nodes, as well as (c) non furcations can be displayed. Phylogenetic networks have an advantage of representing data as they can incorporate predictions from the coalescent theory and can therefore reveal events of the demographic history from the taxa. Moreover, since more data are available, recent publications focus on sophisticated data analysis in spite of collecting new data. The shift of phylogenetic studies from laboratory data collection to a focus on data mining techniques and methods, used to compare and visualize the overwhelming amount of data, may predict that the post genomic age has begun.

ACKNOWLEDGEMENT

This work was supported by the Natural Scientific Funds of China through NSFC project No. 30571362 and No.30972230.

REFERENCES

Bandelt HJ, Macaulay V, Richards M (2000). Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. Mol. Phylogenet. Evol., 16(1): 8-28.

Bandelt HJ, Forster P, Röhl A (1999). Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol., 16: 37–48.

Bandelt HJ, Dress AW (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol. Phylogenet. Evol., 1(3): 242-252.

Bandelt PJ, Forster P, Sykes BC, Richards MB (1995). Mitochondrial portraits of human populations using median networks. Genetics, 141: 743-753.

Bergthorsson U, Adams KL, Thomason B, Palmer JD (2003). Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature Genetics, 424: 197-201.

Bremer K (1988). The limits of amno acid sequence data in angiosperm phylogenetic reconstruction. Evolution, 42(4): 795-803.

Cuc NTK, Simianer H, Groeneveld LF, Weigend S (2011). Multiple Maternal Lineages of Vietnamese Local Chickens Inferred by Mitochondrial DNA D-loop Sequences. Asian-Aust. J. Anim. Sci., 24(2): 155–161.

Darwin C (1859). On the origin of species by mean of natural selection, or the preservation of favoured races in the struggle of life. John Murray, London.

Donnelly P, Tavare S (1986). The ages of alleles and a coalescent. Adv. Appl. Prob., 18: 1-19.

Donoghue MJ, Olmstead RG, Smith JF, Palmer JD (1992). Phylogenetic relationships of dipsacales based on rbcL sequences. Ann. Missouri Bot. Gard., 79: 333- 345.

Doolittle WF (1999). Phylogenetic classification and the universal tree. Science 284(5423): 2124-2129.

Excoffier L, Langaney A (1989). Origin and differentation of human mitochondrial DNA. Am. J. Hum. Genet., 44: 73-85.

Felsenstein J (1985). Confidence limits on phylogenies: an approach using bootstrap. Evolution, 39: 783-791.

Felsenstein J (2004). Inferring Phylogenies. Sinauer Associates; Sunderland, MA.

Han L, Yub HX, Cai DW, Shi HL, Zhua H, Zhou H (2010). Mitochondrial DNA analysis provides new insights into the origin of the Chinese domestic goat. Small Ruminant Res., 90: 41–46.

Huber KT, Langton M, Penny D, Moulton V, Hendy M (2002). Spectronet: a package for computing spectra and median networks. Appl. Bioinformatics, 1(3): 159-61.

Hudson RR (1991). Gene genealogies and the coalescent process. Futuyuma, D.; Antonovics, J., editors. Oxford University Press; New York, Ny. pp. 1-44.

Huson DH, Bryant D (2006). Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol., 23(2): 254-267

Kimura M (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol., 16: 111-120.

Kingman JFC (1982). The coalescent. Stochastic Process. Appl., 13(3): 235-248.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007). Clustal W and Clustal X version 2.0. Bioinformatics, 23(21): 2947-2948.

Lento GM, Hickson RE, Chambers GK, Penny D (1995). Use of spectral analysis to test hypotheses on the origin of pinnipeds. Mol. Biol. Evol., 12(1): 28-52.

Linder CR, Rieseberg LH (2004). Reconstructing patterns of reticulate evolution in plants. Am. J. Bot., 91: 1700-1708.

Maddison DR, Swofford DL, Maddison WP (1997). NEXUS: an extensible file format for systematic information. Syst. Biol., 46 (4): 590-621.

Maddison WP (1997). Gene trees in species trees. Syst. Biol., 3(46): 523-536.

Nakamura Y, Itoh T, Matsuda H, Gojobori T (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nature Genetics, 36(7): 760-766.

Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP (2010). Names are key to the big new biology. Trends Ecol. Evol., 25(12): 686-691.

Posada D, Crandall KA (2001). Intraspecific gene genealogies: trees grafting into networks. Trends Ecol. Evol., 16(1): 37-45.

Posada D, Crandall KA, Holmes EC (2002). Recombination in evolutionary genomics. Annu. Rev. Genet., 36: 75-97.

Ronquist F, Huelsenbeck JP (2003). MrBayes 3: bayesian phylogenetic inference under mixed models. Bioinformatics, 19: 1572-1574.

Rzhetsky A, Nei M (1992). A simple method for estimating and testing minimum evolution trees. Mol. Biol. Evol., 9: 945-967.

Saitou N, Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol., 4: 406-425.

Tamura K (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. Mol. Biol. Evol., 9: 678-687.

Watterson GA, Guess HA (1977). Is the most frequent allele the oldest? Theor. Popul. Biol., 11(2): 141-160.

Zahid MAH, Ankush M, Joshi RC (2006). A pattern recognition-based approach for phylogenetic network construction with constrained recombination. Pattern Recogn., 39: 2312-2322.