Monday 17 December 2007

Using supernetworks to distinguish hybridization from lineage-sorting

Barbara R. Holland[1,*], Steffi Benthin[1], Peter J. Lockhart[2], Vincent Moulton[3], Katharina T. Huber[3]

[1]Allan Wilson Centre, Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand.

[2]Allan Wilson Centre, Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand.

[3] School of Computing Sciences, University of East Anglia, Norwich, UK

* Corresponding author

Email addresses: BRH b.r.holland@massey.ac.nz; SB Steffi.benthin@gmx.de; PL p.j.lockhart@massey.ac.nz; VM vincent.moulton@cmp.uea.ac.uk; KTH Katharina.Huber@cmp.uea.ac.uk

Abstract

Background: A simple and widely used approach for detecting hybridization in phylogenies is to reconstruct gene trees from independent gene loci, and to look for gene tree incongruence. However, this approach may be confounded by factors such as poor taxon-sampling and incomplete lineage-sorting.

Results: Using coalescent simulations, we investigated the potential of supernetwork methods to differentiate between gene tree incongruence arising from taxon sampling and incomplete lineage-sorting as opposed to hybridization. For few hybridization events, a large number of independent loci, and well-sampled taxa across these loci, we found that it was possible to easily distinguish incomplete lineage-sorting from hybridization using the filtered Z-closure and Q-imputation supernetwork methods. Moreover, we found that the choice of supernetwork method was less important than the choice of filtering, and that count-based filtering was the most effective filtering technique.

Conclusions: Filtered supernetworks provide a tool for detecting and identifying hybridization events in phylogenies, a tool that should become increasingly useful in light of current genome sequencing initiatives and the ease with which large numbers of independent gene loci can be determined using new generation sequencing technologies.

Background

In recent years there has been growing interest in the problem of building explicit models of reticulate evolution [1-12]. This work has to a large part been motivated by biological research highlighting the potential importance of hybridization in the origin of biotic diversity, biological invasion and rapid adaptation [13-27].

One simple and widely used approach for detecting hybridization has been to compare gene trees built from independent gene loci, and to consider gene tree incongruence as evidence for hybridization [1, 28, 29]. However, hybridization is not the only possible cause of gene tree incongruence. Other explanations include phylogenetic error [30], gene duplication and loss [31], incomplete lineage sorting [32] and lateral gene transfer [33].

In light of current genome sequencing efforts and the ease of sequencing large numbers of independent gene loci using new generation sequencing technologies, it is important to find ways to differentiate between various explanations of gene tree incongruence before attempting to reconcile gene trees. In this regard, a helpful concept might be "principal trees", the evolutionary trees upon which individual gene loci have evolved. If a phylogeny contains no hybridization or lateral gene transfer, then the expectation is for one principal or "species" tree. However, if hybridization has occurred, then there will be multiple principal trees, each one representing the evolutionary history for specific gene loci.

Here we investigate the potential of filtered Z-closure [34] and Q- imputation [35] and supernetworks to differentiate between phylogenetic signals arising from the principal trees from those caused by taxon sampling, and incomplete lineage-sorting in evolutionary histories involving hybridization. In particular, using gene trees generated under a coalescent process, we test whether these methods can be used to filter out phylogenetic signals that do not correspond to branches in principal trees, signals that have the potential of confounding efforts to reconstruct complex phylogenies.

Methods

Overview

Analogous to the supertree framework [36] our input is a set of trees on overlapping but not necessarily identical taxa. We refer to the *complete taxa set* as the union of the taxa sets of the input trees; *complete splits* are bipartitions of the complete taxa set and *trivial splits* are splits where one part consists of precisely one element. Furthermore *partial trees* and *partial splits* are trees and splits on a subset of the complete taxa set.

Our overall approach is to first generate a collection of partial trees along a hybridization network in the presence of incomplete lineage-sorting (modelled by the coalescent process), to then apply a supernetwork method to this collection of partial trees to obtain a collection of complete splits, and to then apply a filter to reduce the complexity of this collection. The reduced collection of complete splits is then compared to the splits associated with the hybridization network to determine if they have been accurately recovered. We use this approach to study two supernetwork

methods - Z-closure [34] and Q-imputation [35], and two types of filter – one counting-based and one homoplasy-based [37].

Supernetwork methods

We begin with a brief description of the supernetwork methods that will be studied. The Z-closure supernetwork approach was first described in [34] and is implemented in SplitsTree4 [38]. Our implementation differs slightly from that in SplitsTree4 in that it keeps track of multiple copies of partial splits and complete splits, as this information is required by the counting filter that we apply later. The method begins with a collection of partial trees from which a list of partial splits is obtained – each edge in each partial tree gives rise to a partial split in a well-defined way (see e.g. [8, 34, 38]). The Z-closure rule (Figure 1) is then iteratively applied to this list of partial splits by taking pairs of partial splits and either overwriting them with two splits on a more inclusive taxa set, or, if the rule does not apply, returning the same two partial splits. When the Z-closure rule can no longer be applied the method returns the list of complete splits that have been generated. The output is dependent on the order of elements in the list of partial splits, and so we repeat the procedure for 10 random orderings keeping a cumulative count of how many times each complete split appears.

The other supernetwork method we consider, Q-imputation [35], also uses partial trees as input but uses an alternative approach to generating complete splits that is based on the four-taxon subtrees (quartets) of the partial trees. For each partial tree with missing taxa - that is, taxa that are in the complete taxa set but not in the taxa set for that tree – the missing taxa are inserted in the tree. This done by processing the missing taxa in a fixed order and placing each taxon within the partial tree to

maximise the number of quartets that contain the missing taxon and are also quartets of the other partial input trees. Once all the trees have been completed the list of complete splits displayed by the completed trees is returned. (In the special case where all the trees are on identical taxa sets, the Q-imputation method reduces to the consensus network method [39]).

Filtering methods

We apply two different kinds of filter to the lists of complete splits obtained from the two supernetwork methods, a homoplasy-based filter and a counting-based filter. The homoplasy filter [37] requires two user-defined parameters x and y. The level of homoplasy for each complete split and partial tree is determined by recoding the split as a binary character, reducing it to the same taxa set as the partial tree, and evaluating the number of character state changes required to explain the character on the partial tree (i.e. the parsimony score). Splits that have a parsimony score greater than a given number x in more than a given number y of the partial trees are filtered out. The counting filter has one user-defined parameter n, it keeps the n splits that appear most frequently in the list of complete splits (ties are broken arbitrarily). Note that for Q-imputation this is equivalent to the filter described in [35] for some choice of threshold.

Simulations

The starting point for each simulation is a hybridization network such as the one shown in Figure 2a. Formally such networks are rooted, leaf-labelled, directed-acyclic-graphs in which the nodes are of one of four types: nodes with in-degree 2 and out-degree 1 correspond to hybridizations; nodes with in-degree 1 and out-degree 2

correspond to speciation events; nodes with in-degree 1 and out-degree 0 correspond to the extant species; and one special node of in-degree 0 and out-degree 2 is the root. Such a network can be thought of as a collection of rooted principal trees, where the trees result from picking one parent at each hybridization node and suppressing all unnecessary edges and resulting degree 2 nodes (Figure 2b). This leads to a natural definition of the collection of splits associated with a hybridization network as being the union of the splits associated with each of the principal trees of the network (Figure 2c). We will refer to such splits as the true splits of the hybridization network. The purpose of the simulations is to assess if filtered supernetworks can identify the splits present in the principal trees of the hybridization network. To be successful these splits need to be distinguishable from those arising from incomplete lineage-sorting under the coalescent process.

The main flow of our simulation is shown in Figure 3. Given a hybridization network, a collection of trees was created by sampling with replacement from the collection of principal trees (the same tree may appear multiple times). We used the software package COAL [32] to simulate trees according to the coalescent process given a principal tree with branch lengths specified in coalescent units (the number of generations divided by population size). We employed two different branch length settings, either all branch lengths had coalescent units of 1 or all branch lengths had coalescent units of 0.5. We also simulated a situation where there were no lineage-sorting effects. Each tree was then pruned of m taxa at random with the restriction that each taxon in the network must appear in at least one partial tree. These collections of partial trees were then used as input to each of the supernetwork methods.

We also applied the counting filter (CF) to both Z-closure and Q-imputation. For each hybridization network we selected n splits, where n was fixed to be the number of unique non-trivial splits associated with the principal trees of the network (Table 1).

Results and Discussion

Results were generated for each of the hybridization networks given in Table 1, but for brevity, in Figures 4-7 and Table 2 we only show results for hybridization network 7 (the network shown in Figure 2a). Results for the other hybridization networks follow the same general trends and are given in the other hybridization networks follow the same general trends and are given in the supplementary data.

Filtering

Figure 4 shows the change in the average number of split false positives and false negatives with respect to the number of gene trees. The results are averaged over 100 repetitions and the 12 combinations of number of missing taxa m and coalescent branch lengths b.

As can be seen in Figure 4, the (HF1) filter is far too stringent in combination with either Z-closure or Q-imputation; it gives almost no false positives but false negatives increase with increasing g towards the maximum value of 8. The (HF4) filter is not stringent enough in combination with either Z-closure or Q-imputation; it gives almost no false negatives but false positives increase with increasing g. Moreover, (HF3) is ineffective in combination with Z-closure as the number of false positives increases with increasing number of partial trees, in combination with Q-imputation

the average number of false positives stays close to 2 for all values of g. (HF2) is the most effective of the homoplasy filters, as for both Z-closure and Q-imputation both types of errors either decrease or stay reasonably constant with increasing g, a property that we would expect any filtered supernetwork method to satisfy. The counting filter also displays this property for both Z-closure and Q-imputation; both false positives and false negatives decrease with increasing number of input trees.

Figure 5 is similar to Figure 4 except that rather than averaging over all values of b and m we focus on the difficult case with the highest number of missing taxa and the most incongruence generated by incomplete lineage-sorting (m=3, b=0.5). While all the filtered supernetwork methods shown in Figure 5 control false negatives, false positives increase with increasing g for both Z-closure and Q-imputation using (HF3).

For the rest of this section, we restrict our attention to the best homoplasy filter (HF2) and the counting filter (CF).

Missing taxa

Figure 6 shows the trends in the number of false positives and false negatives as the number of missing taxa, m, increases from 0 to 3. Results are averaged over the 12 possible settings for number of partial trees g and coalescent branch lengths b. The (HF2) and (CF) filters exhibit very different behaviour. For (HF2) as m increases the number of false positives increases, in particular going from 2 to 3 missing taxa produces a dramatic increase. Conversely the number of false negatives decreases, this is presumably due to the fact that as the number of missing taxa gets large more splits meet the requirement of the filter. This effect was not observed for (CF) where

the total number of splits is capped; here both false positives and negatives increase with growing m.

Incomplete lineage-sorting

Recall that the parameter b effects the probability that the trees generated by COAL [32] will match the principal tree sampled from the hybridization network, $b=\infty$ corresponds to trees that match exactly. Figure 7 shows the trends in the number of false positives and false negatives for different values of b. Results are averaged over the 16 possible settings for the number of partial trees g and the number of missing taxa m. As expected, both methods and filters perform better when b is large.

Overall performance

Table 2 shows the number of false positives for hybridization network 7 (Figure 2a), which has 2 hybridization events and 8 true splits, for m=0, 1, 2 or 3, b=0.5, 1 or ∞, and g=5, 10, 15 or 20 for both Z-closure and Q-imputation with (CF). If two out of three of the conditions (g, m, or b) are favourable (i.e. many input trees, few missing taxa, and the probability that the input trees are congruent with the principal trees is high) then both methods work well. However if two or three of the conditions are unfavourable then both methods start to break down.

Figures 8a and 8b show the average number of false positives and false negatives respectively (averaged over m, g, and b) versus the number of true splits for hybridization networks 1-9. Hybridization network 10 is not shown in the figures, as it is an outlier with 24 true splits, but results for this network follow the same trends as the other hybridization networks. For (CF) both types of errors increase slowly

with increasing number of true splits. For (HF2) false positives appear fairly constant, but false negatives increase linearly with a slope close to one with increasing number of true splits.

Conclusions

We have evaluated the potential of Z-closure and Q-imputation filtered supernetworks to identify splits belonging to the sets of principal trees associated with hybridization networks. We have found that this approach can recover these splits when hybridization number is low. However, our results imply that (1) if the trees have many missing taxa then many trees are required; (2) if the trees are frequently incongruent with the principal trees of the hybridization network due to incomplete lineage-sorting then a large number of near complete trees is required; (3) and if there are few trees available they need to be both near complete and highly congruent with the principal trees.

Despite these limitations, with the potential now of obtaining large numbers of splits from independent gene loci using new generation sequencing technologies, our findings may nevertheless be applicable for tree like phylogenies where some degree of hybridization is inferred [40]. In such cases, filtered supernetworks can be used to identify the true splits of the underlying hybridization network. Once these are obtained, the method of Huson et al 2005 [5] can be used to convert the split system into a hybridization scenario.

One of our most interesting findings is that the choice of whether to use Z-closure or Q-imputation seems to have much less impact on accuracy with regards to recovering the splits in the underlying hybridization network than the choice of filter. The counting filter has the desirable property that as the amount of data increases (more genes or more complete gene trees) the rate of both false positives and false negatives goes down. Several settings were tried for the homoplasy-based filter and these tended to either suffer from increasing false positives or increasing false negatives as the number of gene trees increased.

In cases where there are many hybridization events, especially between individuals that are not closely related, there will be many principal trees and corresponding splits (as in hybridization network 10). Many of these splits will occur at low frequencies making them hard to distinguish from phylogenetic error. This means that phylogenetic inference will be limited, as gene tree incongruence will be extensive. In such cases, rather than attempt to reconstruct a hybridization network, it may be more appropriate to formulate objective tests to better understand the complexity of the data and the extent to which hybridization contributes to this complexity. Joly, McLenachan and Lockhart (submitted manuscript) have recently proposed such a test.

An unexplored idea worthy of study is the investigation of model based, rather than combinatorial, methods of filtering. One approach might be to consider posterior probability distributions on species trees [41]. It will be interesting to investigate whether such posterior distributions can also be analysed for evidence of distinct principal trees in cases where evolutionary relationships are complex.

Authors' contributions

BRH developed and applied the simulation scheme, implemented the modified Z-closure method, homoplasy filter and counting filter, and contributed to writing the ms, especially the methods and results section. SB conducted initial simulations with Z-closure. PL ensured biological relevance, and contributed to writing the ms, especially the introduction and conclusions. VM ensured mathematical correctness and developed the overall concept. KH ensured mathematical correctness and developed the overall concept, contributed to writing the ms, especially the methods and results section.

Acknowledgements

All authors thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK for hosting them in the context of their Phylogenetics Programme where part of the work presented in this paper was carried out. BRH and PJL thanks Marsden Fund MAU0509 and MAU0510, M thanks the Engineering and Physical Sciences Research Council (EPSRC), grant EP/D068800/1.

References

1. Linder CR, Rieseberg L: **Reconstructing patterns of reticulate evolution in plants.** *Amer J Botany* 2004, **91:**1700-1708.

2. Gusfield D, Eddhu S, Langley CH: **Optimal, efficient reconstruction of phylogenetic networks with constrained recombination.** *J Bioinf and Comp Biol 2004*, **2:**173-213.

3. Nakhleh L, Warnow T, Linder CR, St. John K: **Reconstructing reticulate evolution in species—theory and practice.** *J Comp Biol* 2005, **12:**796-811.

4. Jansson J, Sung W: **Inferring a level-1 phylogenetic network from a dense set of rooted triplets**, *Springer Lecture Notes in Computer Science* 2004, **3106:** 462-471.

5. Huson DH, Kloepper T, Lockhart PJ, Steel M: 2005. **Reconstruction of reticulate networks from gene trees.** In *RECOMB 2005.* LNBI 3500. Edited by Miyano S. Springer-Verlag Heidelberg. 2005:233-249.

6. Song YS, Hein J: **Constructing minimal ancestral recombination graphs.** *J Comp Biol* 2005, **12:**147-162.

7. Baroni M, Semple C, Steel M. **Hybrids in real time.** *Syst Biol* 2006, **55:**46–56.

8. Semple C, Steel M: *Phylogenetics.* Oxford: Oxford University Press; 2003.

9. Huber KT, Oxelman B, Lott M, Moulton V: **Reconstructing the evolutionary history of polyploids from multi-labelled trees.** *Mol Biol Evol* 2006, **23:**1784-1791.

10. Willson S: **Unique determination of some homoplasies at hybridization events.** *Bull Math Biol* 2007, **69:**1709-1725.

11. Cardona G, Rosselló F, Valiente G: **Tripartitions do not always discriminate Phylogenetic Networks.** *Math Biosci*, in press.

12. van Iersel L, Keijsper J, Kelk S, Stougie L: **Constructing level-2 phylogenetic networks from triplets**. 2007, arXiv:0707.2890v1 [q-bio.PE]

13. Linnaeus C: *Species plantarum, exhibentes plantas rite cognitas, ad genera relatas, cum differentiis specificis, nominibus trivialibus, synonymis selectis, locis natalibus, secundum systema sexuale digestas.* Holmiae: Impensis Laurentii Salvii [L. Salvius, Stockholm]; 1753.

14. Anderson E: **Hybridization of the habitat.** *Evolution* 1948, **2:**1-9.

15. Shaw KL, Stebbins GL: **The role of hybridization in evolution.** *Proc Am Philos Soc* 1959, **103**:231-251.

16. Ehrendorfer FR: **Differentiation-hybridization cycles and polyploidy**. In *Achillea. Cold Spring Harbor Symposium on Quantitative Biology* 1958, **24:**141-152.

17. Rattenbury JA: **Cyclic hybridization as a survival mechanism in the New Zealand forest flora.** *Evolution* 1962, **16:**348-363.

18. Levin DA: **Hybridization and evolution – a discussion.** *Taxon* 1970, **19:**167-171.

19. Grant V: *Plant speciation, 2nd edition.* New York: Columbia University Press; 1981.

20. Arnold ML: *Natural hybidization and evolution.* New York: Oxford University Press; 1997.

21. Barrier M, Baldwin BG, Robichaux RH, Purugganan MD: **Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of the Hawaiian Silversword alliance (Asteraceae) inferred from floral homeotic gene duplications.** *Mol Biol Evol* 1999, **16:**1105-1113.

22. Ellstrand NC, Schierenbeck KA: **Hybridization as a stimulus for the evolution of invasivness in plants** *Proc Natl Acad Sci USA* 2000, **97:**7043-7050.

23. Barton NH: **The role of hybridization in evolution**. *Molecular Ecology* 2001, **10:**551-568.

24. Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, Durphy JL, Schwarzbach AE, Donovan LA, Lexer C: **Major ecological transitions in annual sunflowers facilitated by hybridization.** *Science* 2003, **301**:1211-1216.

25. Seehausen O: **Hybridization and adaptive radiation.** *Trends Ecol Evol* 2004, **19:**198–207.

3. Nakhleh L, Warnow T, Linder CR, St. John K: **Reconstructing reticulate evolution in species—theory and practice.** *J Comp Biol* 2005, **12:**796-811.

4. Jansson J, Sung W: **Inferring a level-1 phylogenetic network from a dense set of rooted triplets**, *Springer Lecture Notes in Computer Science* 2004, **3106:** 462-471.

5. Huson DH, Kloepper T, Lockhart PJ, Steel M: 2005. **Reconstruction of reticulate networks from gene trees.** In *RECOMB 2005.* LNBI 3500. Edited by Miyano S. Springer-Verlag Heidelberg. 2005:233-249.

6. Song YS, Hein J: **Constructing minimal ancestral recombination graphs.** *J Comp Biol* 2005, **12:**147-162.

7. Baroni M, Semple C, Steel M. **Hybrids in real time.** *Syst Biol* 2006, **55:**46–56.

8. Semple C, Steel M: *Phylogenetics.* Oxford: Oxford University Press; 2003.

9. Huber KT, Oxelman B, Lott M, Moulton V: **Reconstructing the evolutionary history of polyploids from multi-labelled trees.** *Mol Biol Evol* 2006, **23:**1784-1791.

10. Willson S: **Unique determination of some homoplasies at hybridization events.** *Bull Math Biol* 2007, **69:**1709-1725.

11. Cardona G, Rosselló F, Valiente G: **Tripartitions do not always discriminate Phylogenetic Networks.** *Math Biosci*, in press.

12. van Iersel L, Keijsper J, Kelk S, Stougie L: **Constructing level-2 phylogenetic networks from triplets**. 2007, arXiv:0707.2890v1 [q-bio.PE]

13. Linnaeus C: *Species plantarum, exhibentes plantas rite cognitas, ad genera relatas, cum differentiis specificis, nominibus trivialibus, synonymis selectis, locis natalibus, secundum systema sexuale digestas.* Holmiae: Impensis Laurentii Salvii [L. Salvius, Stockholm]; 1753.

14. Anderson E: **Hybridization of the habitat.** *Evolution* 1948, **2:**1-9.

15. Shaw KL, Stebbins GL: **The role of hybridization in evolution.** *Proc Am Philos Soc* 1959, **103:**231-251.

16. Ehrendorfer FR: **Differentiation-hybridization cycles and polyploidy**. In *Achillea. Cold Spring Harbor Symposium on Quantitative Biology* 1958, **24:**141-152.

17. Rattenbury JA: **Cyclic hybridization as a survival mechanism in the New Zealand forest flora.** *Evolution* 1962, **16:**348-363.

18. Levin DA: **Hybridization and evolution – a discussion.** *Taxon* 1970, **19:**167-171.

19. Grant V: *Plant speciation, 2nd edition*. New York: Columbia University Press; 1981.

20. Arnold ML: *Natural hybidization and evolution*. New York: Oxford University Press; 1997.

21. Barrier M, Baldwin BG, Robichaux RH, Purugganan MD: **Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of the Hawaiian Silversword alliance (Asteraceae) inferred from floral homeotic gene duplications.** *Mol Biol Evol* 1999, **16:**1105-1113.

22. Ellstrand NC, Schierenbeck KA: **Hybridization as a stimulus for the evolution of invasivness in plants** *Proc Natl Acad Sci USA* 2000, **97:**7043-7050.

23. Barton NH: **The role of hybridization in evolution**. *Molecular Ecology* 2001, **10:**551-568.

24. Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, Durphy JL, Schwarzbach AE, Donovan LA, Lexer C: **Major ecological transitions in annual sunflowers facilitated by hybridization.** *Science* 2003, **301:**1211-1216.

25. Seehausen O: **Hybridization and adaptive radiation.** *Trends Ecol Evol* 2004, **19:**198–207.

26. Wissemann V: **Plant evolution by means of hybridization.** *Systematics and Biodiversity* 2007, 5:243–253.

27. Mallet J: **Hybrid speciation.** *Nature* 2007, **446:**279-283.

28. Palmer JD, Shields CR, Cohen DB, Orton TJ: **Chloroplast DNA evolution and the origin of amphidiploid Brassica species.** *Theor Appl Genet* 1983, **65:**181-189.

29. Sang T, Zhong Y: **Testing hybridization hypotheses based on incongruent gene trees.** *Syst Biol* 2000, **49**:422-434.

30. Cusimano N, Zhang L-B, Renner SS: **Biology revaluation of the cox1 group I intron in Araceae and Angiosperms suggests a history dominated by loss rather than horizontal transfer.** *Mol Biol Evol*, in press.

31. Hallet M, Lagergren J, Tofigh A: **Simultaneous identification of duplications and lateral transfers.** In *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology* San Diego; ACM. 2004:347-356.

32. Degnan JH, Salter LA: **Gene tree distributions under the coalescent process.** *Evolution* 2005, **59:**24-37.

33. Doolittle WF, Bapteste E: Pattern pluralism and the Tree of Life hypothesis. *Proc Natl Acad Sci USA* 2007, 104:2043-2049.

34. Huson DH, Dezulian T, Kloepper TK, Steel MA: **Phylogenetic supernetworks from partial trees**. *IEEE Trans. Comput. Biol. Bioinform.* 2004, **1:**151–158.

35. Holland BR, Conner G, Huber K, Moulton V: **Imputing supertrees and supernetworks from quartets.** *Syst Biol* 2007, **56:**57-67.

36. Bininda-Emonds ORP: **The evolution of supertrees**. *Trends Ecol Evol* 2004, **19:**315-22.

Figure legends

Fig 1

An example of two applications of the Z-rule which underpins the Z-closure supernetwork method, where two partial splits displayed in the input trees (A) are extended to full splits as shown in (B) and (C). The bold lines that form the 'Z' shape indicate that the intersection of the taxon sets is non-empty, eg in (B) $\{C,D\} \cap \{D,M\} = \{D\}$, $\{D,M\} \cap \{M,P,T\} = \{M\}$, $\{M,P,T\} \cap \{O,P,T\} = \{P,T\}$, but $\{C,D\} \cap \{O,P,T\} = \emptyset$ so the Z-rule can be applied.

Fig 2

(A) A hybridization network (number 7 from Table 1) with two hybridization nodes. (B) The principal trees of the hybridization network - these are found by choosing a single parent at each hybridization node and then suppressing the resulting internal nodes of degree 2. (C) The splits associated with the hybridization network are those displayed by the principal trees in (B). (D) A split network displaying the splits in (C).

Fig 3.

Flowchart indicating the steps used in the simulation study.

Fig 4.

False positives (A) and false negatives (B) with increasing numbers of input trees for Z-closure (ZC) and Q-imputation (Q) keeping splits with no homoplasy on any tree (HF1), keeping splits with no homoplasy on 75% or more of the trees (HF2), keeping splits with no homoplasy on 50% or more of the trees (HF3), keeping splits with a

homoplasy score of 1 or less on all of the trees (HF4), or keeping the 8 highest weight splits (CF) for hybridization network 7. Values are averages over the 12 combinations of b and m.

Fig 5.

False positives (A) and false negatives (B) with increasing number of input trees for the highest setting of missing taxa (m=3) and the smallest setting for coalescent branch lengths (b=0.5). Abbreviations are as descibed in Fig 4.

Fig 6

False positives (A) and false negatives (B) as the number of missing taxa increases from 0 to 3. Results are averaged over the 12 possible settings for g and b. Abbreviations are as descibed in Fig 4.

Fig 7.

False positives (A) and false negatives (B) for the two different branch length settings using in the coalescent simulation (b=0.5 and b=1), and for the control without incomplete lineage-sorting (b=∞). Results are averaged over the 16 possible settings for g and m. Abbreviations are as descibed in Fig 4.

Fig 8.

False positives (A) and false negatives (B) averaged over 48 combinations of m, g, and b versus the number of true splits for hybridization networks 1-9.

Tables

Table 1. Hybridization networks used in simulations. The column H gives the number of hybridization events, and the column S gives the number of unique non-trivial splits contained in the principal trees.

ID	H	S	Principle trees (given in Newick format)
1	0	5	(((a,b),(c,d)),((e,f),(g,h)));
2	1	8	(((a,b),(c,d)),((e,f),(g,h))); (((a,b),c),(((e,d),f),(g,h)));
3	1	7	(((a,b),(c,d)),((e,f),(g,h))); (((a,(b,d)),c),((e,f),(g,h)));
4	0	5	(((((((a,b),c),d),e),f),g),h);
5	1	7	(((((((a,b),c),d),e),f),g),h); ((((((a,c),(b,d)),e),f),g),h);
6	1	10	(((((((a,b),c),d),e),f),g),h); ((((((a,c),d),e),f),(g,b)),h);
7	2	8	(((a,b),(c,d)),((e,f),(g,h))); (((a,b),(c,d)),(((e,f),g),h)); ((a,((b,c),d)),((e,f),(g,h))); ((a,((b,c),d)),(((e,f),g),h));
8	2	9	(((a,b),(c,d)),((e,f),(g,h))); (((a,b),c),(((d,e),f),(g,h))); ((((a,b),(g,h)),(c,d)),(e,f)); ((((a,b),(g,h)),c),((d,e),f));
9	3	9	((((a,b),(c,d)),(e,f)),(g,h)); ((((a,b),(c,d)),e),(f,(g,h))); (((a,b),(c,d)),((e,f),(g,h))); (((a,b),(c,d)),(e,(f,(g,h)))); ((((a,b),c),(d,(e,f))),(g,h)); ((((a,b),c),(d,e)),(f,(g,h))); (((a,b),c),((d,(e,f)),(g,h))); (((a,b),c),((d,e),(f,(g,h))));
10	3	24	((((b,e),(a,c)),((d,f),g)),h); (((b,(a,c)),((e,g),(d,f))),h); (((a,(b,c)),(((c,d),f),g)),h); (((a,b),(((c,d),f),(e,g))),h); (((a,c),((((b,e),d),f),g)),h); (((a,c),(((b,d),f),(e,g))),h); ((a,((((b,e),(c,d)),f),g)),h); ((a,(((b,(c,d)),f),(e,g))),h);

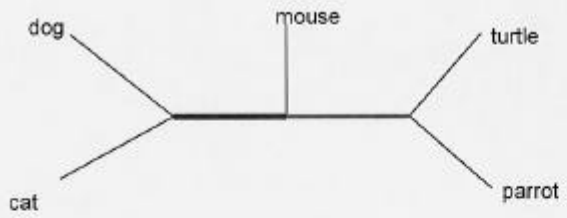
Table 2

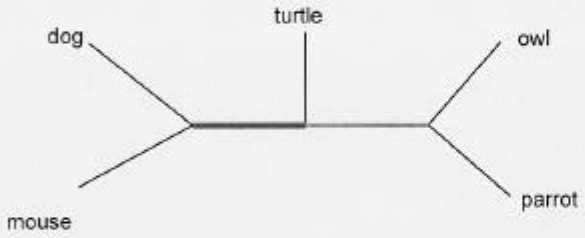
False positives for hybridization network 7 using the counting filter to select the 8 highest weight splits.

		Q-imputation					Z-closure			
		m=0	m=1	m=2	m=3		m=0	m=1	m=2	m=3
b= ∞	g=5	0	0.2	0.64	2.18	g=5	0	0.23	0.56	0.66
	g=10	0	0.07	0.33	1.25	g=10	0	0.17	0.56	1.26
	g=15	0	0.08	0.25	0.86	g=15	0	0.12	0.49	0.98
	g=20	0	0.01	0.18	0.47	g=20	0	0.04	0.29	0.76
		m=0	m=1	m=2	m=3		m=0	m=1	m=2	m=3
b=1	g=5	1.62	2.05	2.29	3.38	g=5	1.61	1.97	2.25	2.25
	g=10	0.68	1.44	1.98	2.69	g=10	0.67	1.37	1.96	2.6
	g=15	0.34	0.97	1.37	2	g=15	0.33	0.84	1.43	2.1
	g=20	0.18	0.74	1.05	1.93	g=20	0.18	0.68	1.12	2.01
		m=0	m=1	m=2	m=3		m=0	m=1	m=2	m=3
b=0.5	g=5	2.85	3.24	3.74	4.56	g=5	2.79	3.04	3.71	3.78
	g=10	1.73	2.25	2.68	3.76	g=10	1.71	2.18	2.64	3.56
	g=15	1.3	1.67	2.24	3.28	g=15	1.31	1.65	2.35	3.16
	g=20	0.94	1.43	2.12	2.57	g=20	0.95	1.33	2.05	2.54

A

A





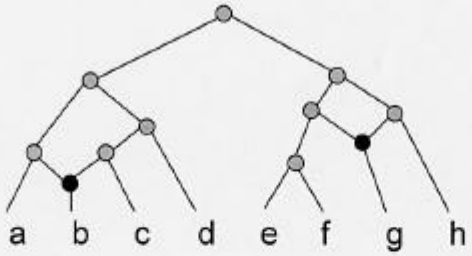
B

$$\frac{C,D}{M,P,T} \sim \frac{D,M}{O,P,T} \rightarrow \frac{C,D}{M,O,P,T} \text{ and } \frac{C,D,M}{O,P,T}$$

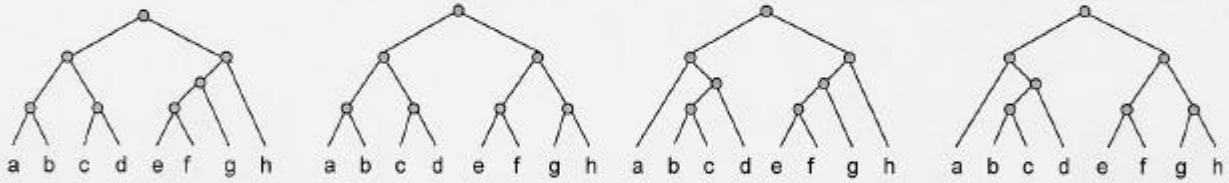
C

$$\frac{C,D,M}{P,T} \sim \frac{D,M,T}{O,P} \rightarrow \frac{C,D,M}{O,P,T} \text{ and } \frac{C,D,M,T}{O,P}$$

A



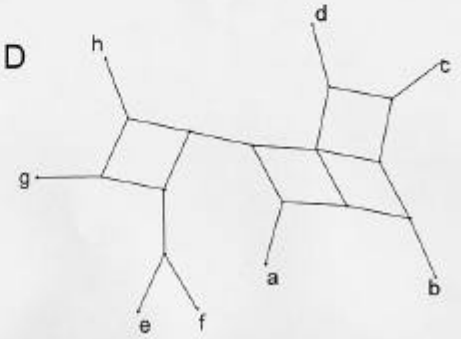
B

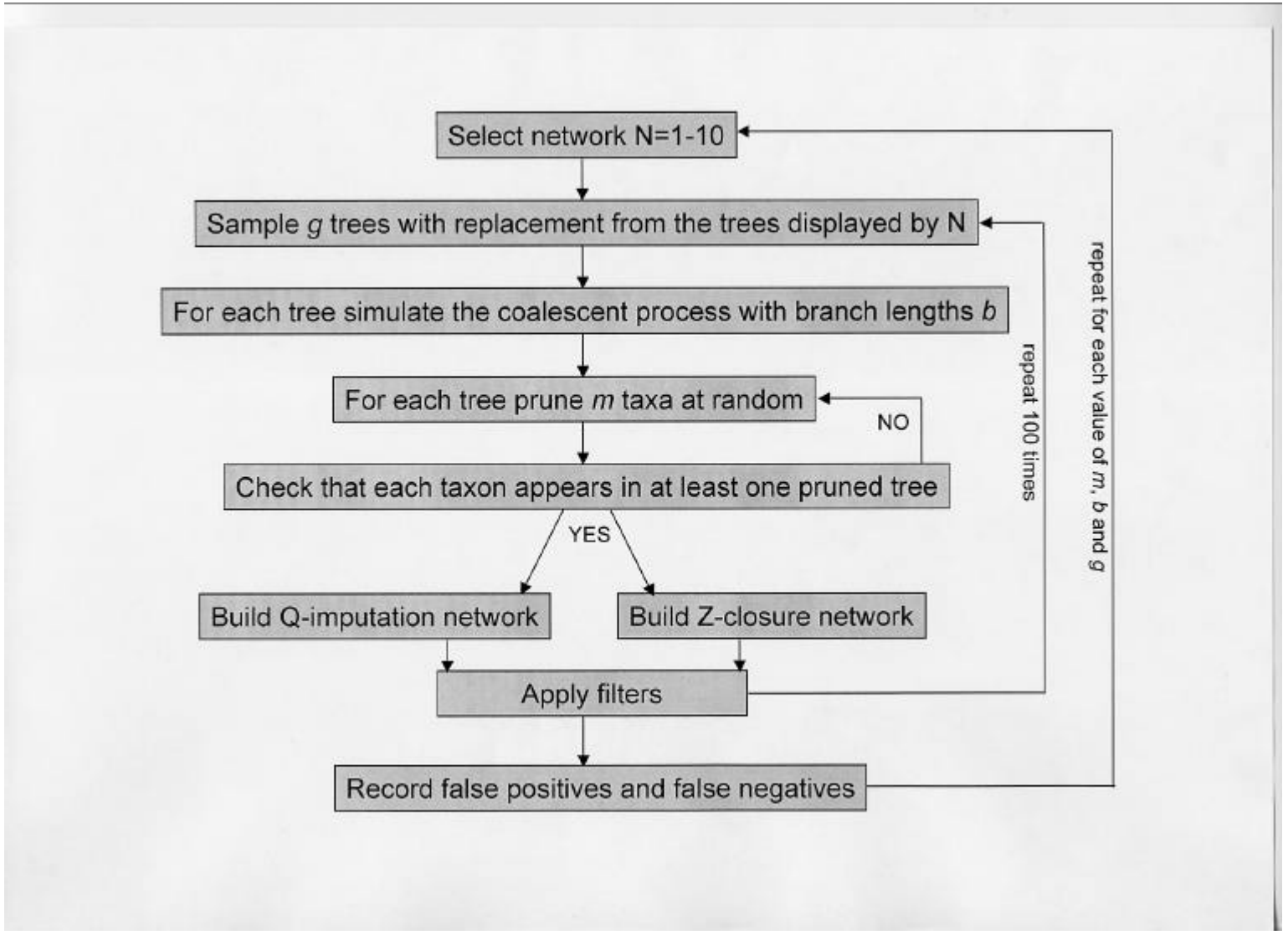


C

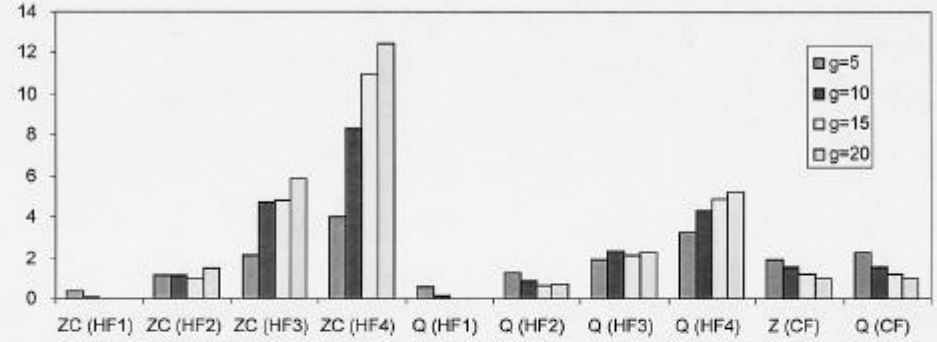
ab|cdefgh
cd|abefgh
abcd|efgh
ef|abcdgh
efg|abcdh
bc|adefgh
bcd|aefgh
gh|abcdef

D

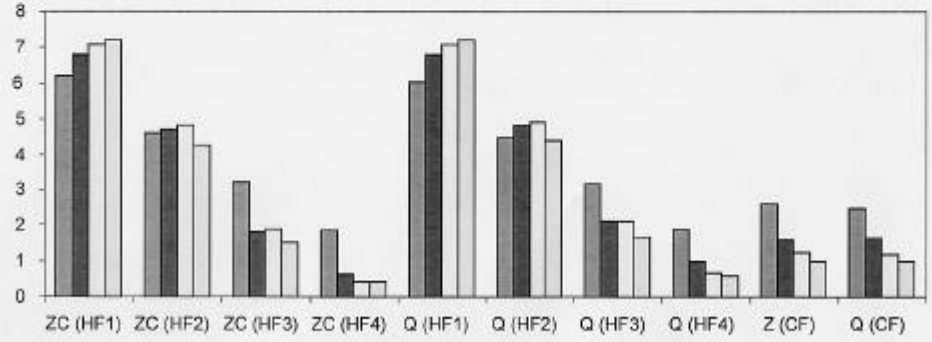



Select network N=1-10
Sample g trees with replacement from the trees displayed by N
For each tree simulate the coalescent process with branch lengths b
For each tree prune m taxa at random
NO
Check that each taxon appears in at least one pruned tree
YES
Build Q-imputation network
Build Z-closure network
Apply filters
Record false positives and false negatives
repeat 100 times
repeat for each value of m, b and g

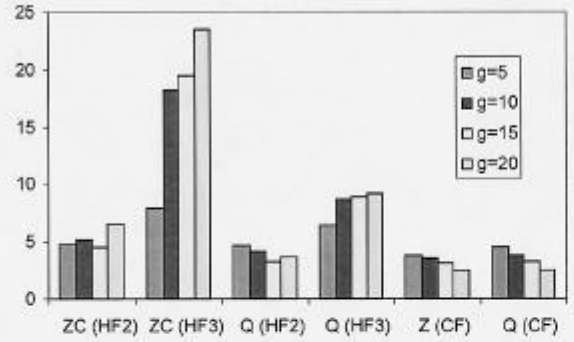
A


14
12
10
8
6
4
2
0
g=5
g=10
g=15
g=20
ZC (HF1)
ZC (HF2)
ZC (HF3)
ZC (HF4)
Q (HF1)
Q (HF2)
Q (HF3)
Q (HF4)
Z (CF)
Q (CF)

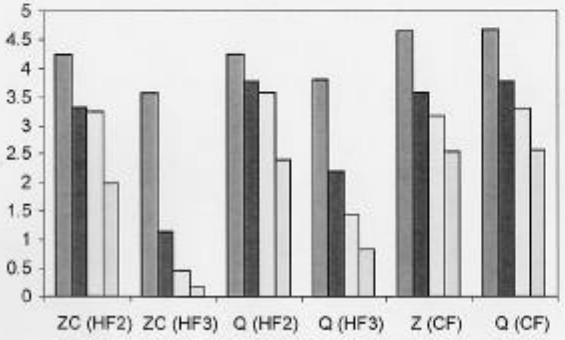
B

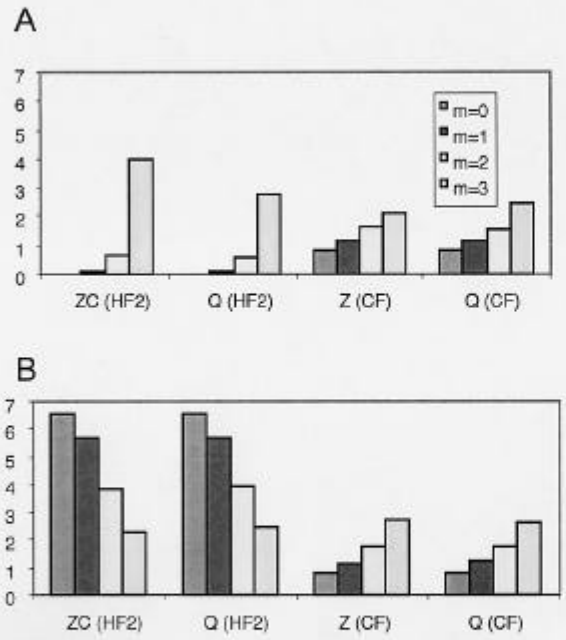

8
7
6
5
4
3
2
1
0
ZC (HF1)
ZC (HF2)
ZC (HF3)
ZC (HF4)
Q (HF1)
Q (HF2)
Q (HF3)
Q (HF4)
Z (CF)
Q (CF)

A

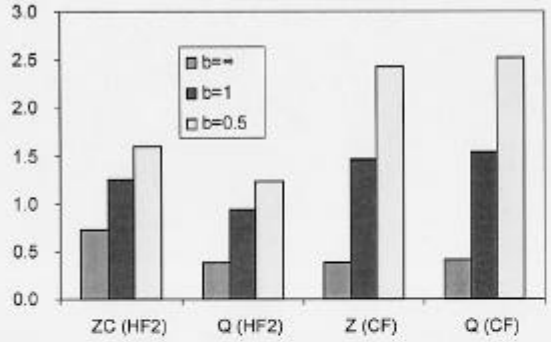

25
20
15
10
5
0
g=5
g=10
g=15
g=20
ZC (HF2)
ZC (HF3)
Q (HF2)
Q (HF3)
Z (CF)
Q (CF)

B

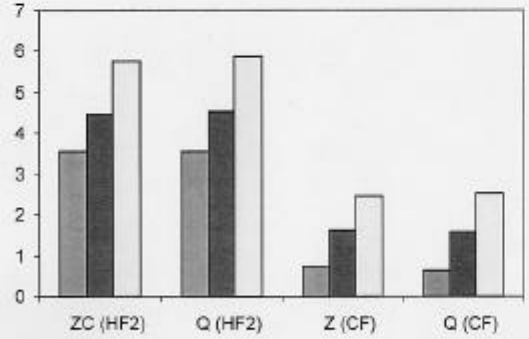

5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
ZC (HF2)
ZC (HF3)
Q (HF2)
Q (HF3)
Z (CF)
Q (CF)


A
7
6
5
4
3
2
1
0
m=0
m=1
m=2
m=3
ZC (HF2)
Q (HF2)
Z (CF)
Q (CF)
B
7
6
5
4
3
2
1
0
ZC (HF2)
Q (HF2)
Z (CF)
Q (CF)

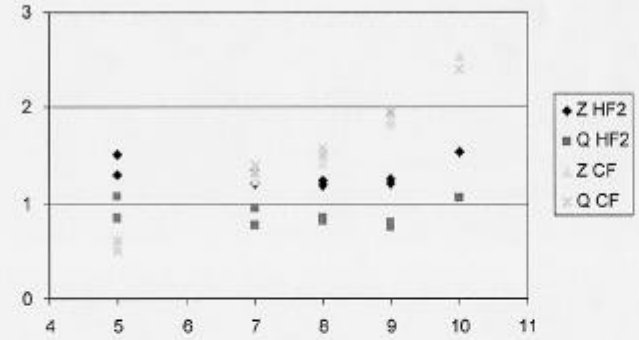
A


3.0
2.5
2.0
1.5
1.0
0.5
0.0
b=∞
b=1
b=0.5
ZC (HF2)
Q (HF2)
Z (CF)
Q (CF)

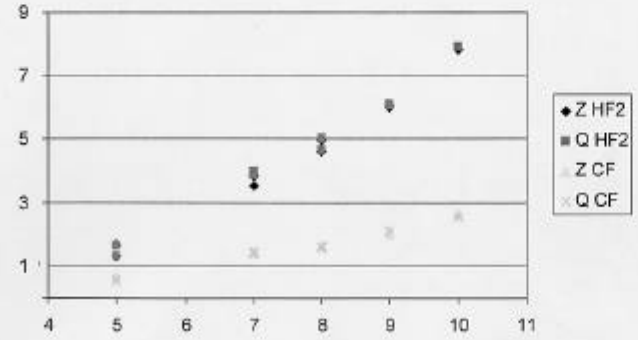
B


7
6
5
4
3
2
1
0
ZC (HF2)
Q (HF2)
Z (CF)
Q (CF)

A


3
2
1
0
4
5
6
7
8
9
10
11
Z HF2
Q HF2
Z CF
Q CF

B


9
7
5
3
1
4
5
6
7
8
9
10
11
Z HF2
Q HF2
Z CF
Q CF