Available online at www.sciencedirect.com
SCIENCE DIRECT
Molecular Phylogenetics and Evolution 39 (2006) 263–270
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Short communication

On conditioned reconstruction, gene content data, and the recovery of fusion genomes

C. Donovan Bailey [a,*], Mathew G. Fain [b], Peter Houde [a]

[a] Department of Biology, New Mexico State University, P.O. Box 30001, Las Cruces, NM 88003, USA
[b] Department of Biology, University of New Mexico, 255 Castetter Hall, Albuquerque, NM 87131, USA

Received 14 June 2005; revised 23 November 2005; accepted 28 November 2005
Available online 18 January 2006

1. Introduction

Conditioned reconstruction (CR)[1] represents a new phylogenetic method that has been presented as a means of utilizing vast amounts of gene absence/presence data to reconstruct phylogenetic relationships and to directly study the influence of genome fusion on evolution (Lake and Rivera, 2004; Rivera and Lake, 2004; Simonson et al., 2005). In the first direct application of CR, the results were stated to unambiguously support the role of an archaeal and eubacterial genome fusion event in the ancestry of the eukaryotic genome (McInerney and Wilkinson, 2005; Rivera and Lake, 2004; Simonson et al., 2005). In the present manuscript, we outline the basic components of CR, discuss the logic behind their application, and subsequently identify concerns with aspects of data interpretation and the current use of conditioning genomes and aligned networks in the study of genome fusion.

Conditioned reconstruction begins with the selection of a conditioning genome (CG) that will represent the full set of orthologous genes coded during matrix development (Lake and Rivera, 2004). Strong emphasis has been placed on the strict use of relatively small genomes as conditioning genomes (Lake and Rivera, 2004). This recommendation is based on the observation that the use of relatively large conditioning genomes can induce an artifact referred to as "big genome attraction," which can mislead phylogenetic inference. All other genomes that will be used in the analysis are scored for the absence or presence of a putative ortholog to each gene in the CG. Orthologs present in other genomes included in the analysis, but absent from the CG, are not considered and the CG is excluded from phylogenetic analysis. This approach was implemented to simplify the gene selection process and because the inclusion of the CG in analyses would not allow Absence → Absence (A → A) transformations to be defined (Lake and Rivera, 2004), which represents a problem for the estimation of novel conditioned pairwise distances favored by Lake and Rivera (2004). Thus, it was argued that the use of a conditioning genome allows for uniquely defined probabilities of all character state transformations (e.g., for two taxa A → A, A → P, P → A, P → P).

The matrix developed through this procedure (referred to here as the "conditioned matrix") is subject to multiple bootstrap resamplings and phylogenetic analysis to identify both optimal and suboptimal topologies supported by the data. While various approaches are applicable in the analysis of the conditioned matrix (e.g., parsimony and simple distance), Lake and Rivera (2004) concluded that a novel Markov-based approach used to estimate conditional probabilities of gene insertion/deletion, and that required the development of the conditioned matrix, produced results that were least susceptible to "big genome attraction" (additional discussion below). Subsequent to phylogenetic analysis an attempt is made to align all networks recovered from the analysis (see Fig. 3 in Lake and Rivera, 2004). Network alignment is the successful rotating of networks around nodes or inverting networks such that the operational terminals from two or more different networks may be partially or fully overlapped in a repeat linear pattern. Aligned networks are accepted as evidence of reticulate evolution rather than strict divergence. Repetition of this process using alternative single CGs and the continued recovery of results that are consistent with the same fusion hypothesis are accepted as robust evidence of fusion (Rivera and Lake, 2004—supplemental material).

* Corresponding author. Fax: +1 505 646 5665.
E-mail address: dbailey@nmsu.edu (C.D. Bailey).

[1] *Abbreviations used:* CR, conditioned reconstruction; CG, for conditioning genome.

1055-7903/$ - see front matter
doi:10.1016/j.ympev.2005.11.020

While not specifically noted by Lake and Rivera (2004), the logic behind aligning networks as a means of identifying genome fusion is based on the behavior of terminals traditionally referred to as "wildcards" (Nixon and Wheeler, 1992). Wildcards are individual terminals that have more than one distinct resolved position in different supported trees. Terminals of hybrid origin (e.g., Funk, 1985; McDade, 1992) and those that lack sufficient data (e.g., Nixon, 1996) are both well known to behave as wildcards in phylogenetic analysis. In the former case instability in position is induced by conflicting signal in the hybrid, just as it is with fusion genomes. The specific wildcards sought in CR are fusion genomes that combine orthologous gene content from divergent genomes and therefore resolve with both of their parental genomes and/or parental lineages in alternative topologies (see Fig. 3 in Lake and Rivera, 2004). In our view, methods that have been applied to the study of hybridization are relevant to the study of fusion because fusion is simply hybridization between highly divergent taxa.

2. Conditioned reconstruction and causes of conflict

Regarding the discussion of potential problems with the interpretation of results from CR, we begin with the consideration of the eukaryote example presented by Rivera and Lake (2004). The eukaryotic genome has been demonstrated to include gene sets that represent both archaeal and eubacterial ancestry (e.g., Lake and Rivera, 2004; Spring, 2003). Current debate surrounding the origin of this complex assemblage of genes focuses on the competing causal mechanisms (e.g., Spring, 2003). Under the more traditional view, the eukaryotes were initially derived from an archaeal ancestor and extensive directed horizontal gene transfer from the eubacterial derived mitochondrial genome (and/or extracellular genomes) has led to the introduction of mixed gene sets in the nucleus. The alternative hypothesis suggests that a genome fusion event involving archaeal and eubacterial genomes led to the formation of the ancestral eukaryote (e.g., Lake and Rivera, 2004; Martin and Muller, 1998). Rivera and Lake (2004) state that CR can unambiguously distinguish horizontal gene transfer from genome fusion using ortholog absence/presence data.

In the absence of supported competing hypotheses, it is logical to accept genome fusion as the parsimonious explanation of the type of genome conflict observed in the eukaryotic genome. However, with the eukaryotic genome the existence of conflict is uncontroversial. Researchers are currently engaged in disentangling the influence of potential fusion from well documented widespread horizontal gene transfer as mechanisms influencing the tree of life (e.g., Creevey et al., 2004; Gogarten et al., 2002). Clearly, it is important to consider if the set of approaches that comprise CR can be used to distinguish fusion from other confounding factors such as horizontal transfer (see also, Delsuc et al., 2005). According to its authors, CR lacks sufficient information to differentiate causal mechanisms—"…genomic alignment only contains information about the probabilities of gene insertion or deletion; it cannot contain any information about the mechanism of transfer" (Lake and Rivera, 2004, p. 682). This point is difficult to reconcile with a later statement (p. 689): "In theory, conditioned reconstructions can differentiate between events in which genomes are instantaneously fused, on a genomic time scale, and those in which genes are slowly transferred from one genome to another." Conditioned reconstruction contains no information on the time scale involved in individual or whole-scale gene introductions. The method involves the bulk analysis of gene absence/presence data to identify conflict; it is unable to differentiate the cause or timing of introduced conflict. We conclude that the first assessment is correct, and that finer-scale methods accounting for the taxonomic origin of genes *and* the timing of gene introduction are necessary to distinguish the relative importance of fusion and horizontal gene transfer in the evolution of the eukaryotic genome [see Linder and Rieseberg (2004) and Creevey et al. (2004) for salient discussion on topics related to hybridization and horizontal gene transfer, respectively].

3. Gene exclusion and conditioning genomes

It has been suggested that one of the strengths of CR is its ability to incorporate an exhaustive collection of data (McInerney and Wilkinson, 2005). However, the current use of CR does not include all relevant data because each conditioned matrix only incorporates genes found in a single conditioning genome. Despite the fact that Lake and Rivera (2004) have demonstrated that CR implemented using paralinear distances slightly outperforms Jukes-Cantor and parsimony on a single empirically derived conditioned matrix that includes big genome attraction issues, no simulations have been developed to test the more general performance of CR in comparison with methods that do not require the use of a CG and the mandatory elimination of available character data. Methods that require the elimination of available data should be tested to illustrate that the loss of information is not negatively impacting the outcomes. In our view, the pattern of character selection employed through the use of a CG can lead to positively misleading results when trying to recover a fusion genome.

To illustrate this point, we have taken a known seven taxon "balanced" tree (Fig. 1A) and simulated genome evolution along its branches using Rose sequence generation software (Stoye et al., 1998). Rose can be used to simulate genome evolution by treating base pairs as orthologous genes and insertions/deletions as gains and losses of genes, respectively. The ancestral genome size was set at 2500 genes to reflect the average number of genes found in prokaryotic genomes (*sensu*, Rivera and Lake, 2004). Events were limited to single bp insertion/deletion events (indels) to simulate single gene gain/loss. To explicitly avoid branch length artifacts that can negatively influence phylogenetic

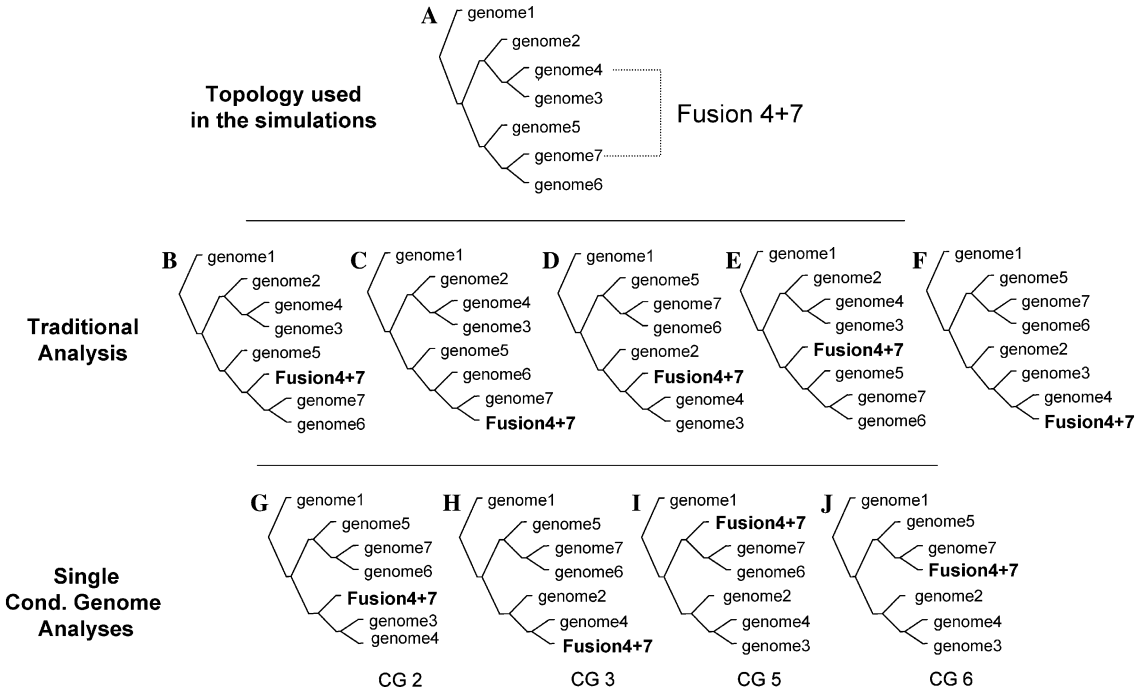

Fig. 1. Simulation of fusion genome behavior using a known divergent phylogeny and simulated gene absence/presence data. The fusion genome represents combined gene content of genomes 4&7. For the "conditioned analyses" one conditioning genome (CG) was applied to each matrix developed and these genomes were excluded phylogenetic analysis. For the traditional analysis, all characters and genomes were included. (A) The correct topology recovered for the "true tree" prior to introducing the fusion genome. (B–F) Topologies supported by the traditional approach. (G–J) Single topologies resulting from each conditioned analysis using a single genome (1G = GC2, 1H = CG3, 1I = CG5, 1J = CG6). Topologies derived from simultaneous analysis of multiple conditioning genomes are presented in supplemental materials.

inference under different methodologies (e.g., Felsenstein, 1978; Siddall, 1998), the probability of insertions and deletions was set to be equal across all branches of the tree. The sequence alignment was converted to a genome alignment by considering just two states, presence or absence of an ortholog. Parsimony-based bootstrap analyses (Felsenstein, 1985) were carried out using Winclada (Nixon, 2002, p. 481) and NONA (Goloboff, 2000). For each matrix 1000 bootstrap replicates were conducted (rep 1000, mult*10, h/10). The appropriateness of parsimony for this application is noted by "… parsimony will not be biased by unequal rate effects when it is used immediately after the fusion event. Parsimony has the particular advantage that the results may be easily interpreted" (Lake and Rivera, 2004, p. 685). The Rose output with all input parameters and the matrices developed are available online (MPE online supplemental material). For the illustrated results (below), we present every network recovered for each analysis.

Twelve matrices were developed from the simulated data. The first matrix included all genes and divergent genomes to test for the recovery of the "true tree" from the simulated gene presence/absence data prior to introducing a fusion genome. Eleven additional matrices were developed by including an extra terminal combining the genome content of genomes 4 + 7 (i.e., a fusion genome). One of these matrices included all seven divergent genomes, the fusion genome, and all characters. This is referred to as the "traditional analysis," reflecting how the data would have been assembled and analyzed prior to the advent of the conditioned approach. The remaining matrices were all conditioned with one or more conditioning genomes. These included four analyses following current use of CR (each using a single conditioning genome—2, 3, 5, and 6) and six novel analyses simultaneously including two conditioning genomes each (2&3, 5&6, 2&6, 3&6, 2&5, 3&5). While Lake and Rivera (2004) suggested that future analyses might simultaneously include multiple conditioning genomes, this approach has not been implemented in published applications of the method. The fusion genome was generated by scoring each ortholog as present if a copy occurred in either or both of the parents and absent if it is not observed in either parent. This precisely matches the specific model outlined by Lake and Rivera (2004, p. 684).

Fig. 1A depicts the "true tree" uniquely recovered in 100% of bootstrap replicates when no complicating factors are introduced from the fusion genome. Bootstrap replicates from the traditional analysis supported five distinct topologies (Figs. 1B–F; Table 1). Comparison of these topologies to the tree lacking the fusion genome indicates that all relationships are stable except that of the fusion genome, which moves to alternative positions within the two clades. These results are consistent with instability induced by a fusion terminal (wildcard terminal) and that we would expect to see in the conditioned analyses. In contrast to the traditional analysis, the single genome conditioned analyses each recover just one topology (i.e., 100%

Table 1
Results from analysis of matrices developed from the simulation of genome evolution on a seven taxon tree

	Total genes	PI genes	Tree(s) recovered (% bootstrap support)	CI	RI
Analysis					
No fusion genome	3651	788	1A (100)	0.83	0.86
Traditional approach	3651	993	1B (0.29), 1C (81), 1D (0.29), 1E (0.09), 1F (18)	0.62	0.66
Conditioned analyses					
Cond. genome 2	2103	374	1H (100)	0.80	0.83
Cond. genome 3	2075	364	1I (100)	0.84	0.87
Cond. genome 5	2084	390	1J (100)	0.80	0.83
Cond. genome 6	2104	409	1K (100)	0.83	0.86
Cond. genomes 2&3	2727	379	*1 tree (100—consistent with 1K—result inconsistent with fusion)*	0.86	0.87
Cond. genomes 5&6	2792	415	*1 tree (100—consistent with 1I—result inconsistent with fusion)*	0.86	0.86
Cond. genomes 2&6	2935	570	*3 trees (combination consistent with fusion)*	0.70	068
Cond. genomes 3&6	2998	567	*4 trees (combination consistent with fusion)*	0.63	0.55
Cond. genomes 2&5	2838	533	*2 trees (combination consistent with fusion)*	0.85	0.76
Cond. genomes 3&5	2908	550	*3 trees (combination consistent with fusion)*	0.68	0.65

"PI" refers to parsimony potentially informative genes. Figure references are provided for Fig. 1. *Note.* The topologies resulting from novel analysis of matrices using multiple conditioning genomes (in italics) are presented in supplemental materials.

support for a single tree; Figs. 1G–J). The application of conditioning genomes 2 (Fig. 1G) and 3 (Fig. 1H) produced topologies with the fusion genome resolved in the clade with parental genome 4 while conditioning genomes 5 (Fig. 1I) and 6 (Fig. 1J) recovered topologies with the fusion genome in the clade with parental genome 7 (see also Table 1). It was only through novel simultaneous application of multiple conditioning genomes that divergent topologies were recovered through the conditioned approach (Table 1). Whenever the conditioning genomes spanned the two clades (CGs 2&5, 2&6, 3&5, 3&6) sufficient character conflict was retained and the result was consistent with the presence of a fusion genome. Note that using two CGs from the same clade (2&3 and 5&6) each produced results similar to the application of single CGs (Table 1). These results, from a relatively uncomplicated data set, do not support the supposition that multiple analyses using alternative single CGs should individually recover evidence of fusion (*sensu* Rivera and Lake, 2004).

These comparisons illustrate that the current use of a conditioning genome(s) (one genome per analysis) can reduce or eliminate the critical character conflict introduced by the addition of a fusion genome. Without this variation conditioned analyses using single CGs, or unbalanced representation of CGs, recover a single topology per CG and fails to identify the presence of a fusion genome. This point is further illustrated by the comparison of the consistency index (CI) and retention index (RI) for each result (Table 1). The CI and RI for the conditioned matrices using single CGs or multiple CGs from the same clade are essentially equal to the result derived from the matrix that did not include any fusion genome. In contrast, we observed the expected decrease in the CI and RI in the traditional analysis and the novel use of multiple CGs relative to the divergent analysis or standard conditioned analyses. It is further demonstrated by considering gene exclusion resulting from the conditioned approach. The traditional matrix included 3651 orthologous genes with 993 parsimony informative characters. Applying the use of single CGs reduced character numbers by an average of 40% (1480 orthologs excluded) with an average of just 380 parsimony informative characters (38% of the traditional matrix). Not only are genes excluded, but a high percentage of variable genes (parsimony informative) are eliminated by current application of CR.

Explaining the behavior of fusion genomes in conditioned analyses with a single conditioning genome is straightforward. When attempting to recover the parentage of a fusion genome the critical characters are those with different character states in the two parental genomes or lineages. Using single CGs, a fusion genome will generally have inherited presence of genes from the parent that is more closely related to the CG (which has all presence characters) and absence from a more diverged parent. The orthologs that the fusion genome shares in common with the other parent and that are not present in the relative of the conditioning genome will be ignored. With dominant data, such as gene presence/absence, this inheritance pattern and sampling procedure results in the fusion genome being scored as present for most of the characters used in the conditioned matrix and therefore sharing more orthologous gene content with the parent that is more closely related to the conditioning genome. This is a direct result of excluding orthologs that are not found in the conditioning genome and it identifies that the phylogenetic position of the CG(s) is an extremely important consideration in CR. In these simulations, either the traditional analysis or two novel applications of CR must be applied to uncover the fusion genome. For the analyses that include just one CG each (*sensu* Rivera and Lake, 2004), we must combine information from the trees resulting from the different analyses (alternative CGs) to identify the fusion genome. Alternatively, one must simultaneously apply multiple conditioning genomes that represent diverse components of the tree. Neither of these applications has been previously

applied to CR and both result in reduced character sampling relative to the traditional approach.

3.1. Use of alignability

Using all genes and genomes in the traditional approach, the recovery of multiple supported trees (Figs. 1B–F) is indicative of a fusion genome; however, these trees are not alignable. The range of resolved positions for the fusion genome is consistent with discussion of hybrid genomes presented by McDade (1992). In contrast to alignability, existing approaches can be used to infer the potential presence and identity of fusion genomes within matrices. For example, Adam's consensus (Adams, 1972; Rohlf, 1982) and related approaches (Wilkinson, 1994) have clear procedures associated with their application, existing software for implementation, and these have been used to identify more extensive wildcard behavior (see, Bremer, 1990; Nixon and Carpenter, 1996) than appears to be detected using aligned networks. Moreover, an Adam's-based approach may identify a single terminal or a subset of terminals that need to be investigated rather than necessarily leaving the researcher to delete each terminal and rerun analyses to identify the culprit(s). This is illustrated by comparing strict and Adam's consensus trees derived from the topologies recovered through the traditional results (Fig. 2A). In this case the strict consensus is entirely unresolved while the Adam's is fully resolved except for the position of the fusion genome (4 + 7). The placement of the fusion genome at the base of the tree identifies it as the

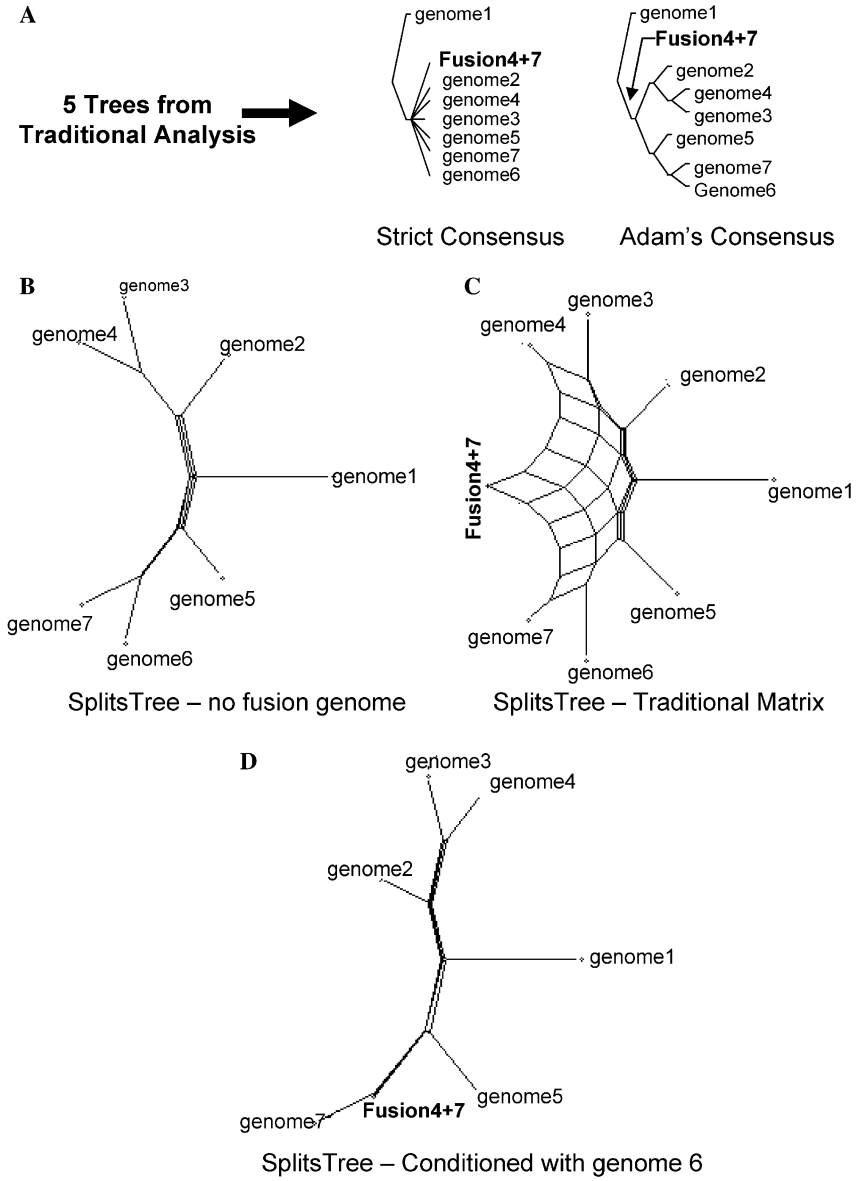

Fig. 2. Application of alternative approaches to alignability in the recovery of fusion genomes. (A) Comparison of strict and Adam's consensus trees from the set of trees resulting from "traditional analysis" (Figs. 1. (B–D) Application of SplitsTree (Dress et al., 1996) to the divergent matrix, an example conditioned matrix (using CG 3), and to the traditional matrix. Note that only the traditionally derived matrix retained sufficient character conflict to suggest the presence of a fusion genome.

wildcard/fusion genome causing the lack of resolution in the strict consensus.

In addition to combined Adam's and strict consensus approaches, previously developed software, such as SplitsTree (Dress et al., 1996), which specifically attempts to depict relationships involving reticulation, can be used to further analyze and visualize results of genome content analyses. Analysis of the traditional matrix using SplitsTree produces a network within which we can visualize the fusion genome's mixed ancestry (Fig. 2C). Furthermore, analysis of the divergent matrix (no fusion genome—Fig. 2B) and each of the single genome conditioned matrices (e.g., Fig. 2D) supported networks consistent with the results presented above (i.e., no sign of mixed ancestry).

4. Reticulation and wildcards in CR

Beyond concerns with the level of information that can be extracted from the results of CR, the current application of CGs, and the failure of alignability to uncover intragenomic conflict, there are additional concerns with the current use and understanding of CR. In this final section, we will further discuss the conditions under which fusion genomes should be expected to act as wildcards and the potential influence random resolution and homoplasy may have on results.

In searching for optimal topologies based on parsimony using morphological characters, McDade (e.g., McDade, 1990) demonstrated that even recently derived hybrids do not necessarily behave as wildcards nor do hybrids only resolve with either or both parentals. For gene absence/presence data, the failure of a fusion genome to behave as a wildcard may be attributed to biased gene loss, the fusion of genomes containing unequal numbers of genes, or both. Either of these factors should effectively reduce conflicting signal and therefore diminish the expectation that fusion genomes act as wildcards. Lake and Rivera's (2004) novel implementation of the bootstrap to identify genomic conflict has obvious advantages over strict consideration of optimal topologies in a single matrix. The resampling procedure is applied to identify potential conflict that may not be clear in the analysis of the original matrix. Nevertheless, there are important limitations to consider. The implementation of CR to study fusion assumes roughly similar genome contributions and/or a lack of subsequent bias in retained gene sets to be effective. Deviation from either of these factors is not outside the range of expectation, particularly in the case of ancient fusion. Evidence for biased gene loss is supported by recently derived hybrid eukaryotic genomes that reveal biased shifts in gene content occurring at remarkable rates (e.g., Ozkan et al., 2001; Song et al., 1995). With respect to the origin(s) of the eukaryotic genome, we see no compelling evidence to suggest that a three billion year old fusion genome, that combined sufficiently divergent genomes to be detectable today, should be assumed to have arisen from parentals of similar genome size or that their descendants would necessarily retain balanced proportions of genic complements.

Furthermore, no consideration of the influence of homoplasy and weak support on recovered networks in CR has been made. Poorly supported or even randomly generated networks may provide ambiguous results that are accepted as evidence of fusion. For the following discussion, we return to Rivera and Lake's (2004) eukaryotic genome example, which represents a five terminal network (e.g., the two yeast genomes are sister in all networks and therefore function as a single terminal). The five networks reported by Rivera and Lake (2004, p. 152) are incongruent if traditionally interpreted as distinct dichotomous topologies. Instead, Rivera and Lake (2004) use the fact that they can be aligned to argue that the networks support a "ring-like" pattern with 96.3% support. The individual networks aligned in their primary example received 60.5, 16.8, 10.0, 7.2, and 1.8% bootstrap support, respectively (with 3.7% excluded from consideration). However, it is critical to note that a maximum of five unique networks (from a possible 15 for five terminals) can be aligned to form a circular topology. Thus the recovery of six or more networks, as was found in the eukaryote example (the unalignable 3.7% of trees), questions support for genome fusion when applying alignability. Furthermore, any two networks generated at random for five taxa have an 87% probability of being alignable. Both of these observations raise unaddressed questions: (1) are networks supported by low bootstrap values significantly different from random noise in the data, (2) how is the summation of bootstrap values over independent topologies statistically justified as representing overall bootstrap support for a "ring," (3) how is the rejection of networks (e.g., 3.7%) that are inconsistent with alignability, justified when some networks receiving low support are accepted as evidence, and (4) are unalignable networks necessarily inconsistent with fusion (see above). From the eukaryote example (Rivera and Lake, 2004), additional concerns are noted by the fact that the majority of replicates (60.5%) support a topology in which *Bacillus* shares more orthologous gene content with archaea than with its two fellow eubacteria.

5. Conclusions

Through examples and discussion we have argued that (1) CR cannot be used to distinguish horizontal gene transfer from genome fusion when the two represent competing causal factors in the formation of mosaic genomes, (2) the current use of CR can induce bias in ortholog sampling inhibiting the discovery of fusion genomes, and (3) alignability as a criterion for identification of fusion genomes can fail relative to existing approaches. We suggest that additional investigations are warranted to better understand the performance of available methods on a variety of phylogenetic frameworks. The inclusion of additional data in existing approaches (e.g., the "traditional approach") does not violate

assumptions inherent in the methods and they appear to retain the necessary conflict to recover fusion terminals. Such analyses have been conducted for the purpose of studying divergence (e.g., Fitz-Gibbon and House, 1999; Herniou et al., 2001; Huson and Steel, 2004; Montague and Hutchison, 2000; Snel et al., 1999) but not reticulation. Alternatively, careful application of CR through simultaneous inclusion of multiple CGs representing clades that include potential progenitors to a fusion genome may be effective.

In the study of intragenomic conflict using gene absence/presence data, none of these methods should be expected to generate wildcard behavior in fusion genomes without a relative balance of parental gene components. Furthermore, the results should not necessarily be interpreted as fusion, particularly when the alternative hypothesis is horizontal gene transfer. For those who choose to use instability induced by wildcards to study intragenomic conflict from dominant data, an Adam's consensus or Splits approach can be implemented with existing software and these appear to be more effective than use of alignability (which lacks available software). While the phylogenetic pattern underlying the origin of eukaryotes may well turn out to be consistent with fusion (e.g., Campbell, 2000; Karlin et al., 1997; Rivera and Lake, 2004; Spring, 2003), we question the notion that these methods have explicitly tested the relevant questions (fusion vs. horizontal gene transfer) and the present application of CGs and alignability relative to preexisting phylogenetic tools.

Acknowledgments

Five anonymous reviewers provided productive comments on various incarnations of the manuscript. We are especially indebted to the third MPE reviewer for his/her thorough and thoughtful evaluation that countered an earlier review vehemently arguing for rejection of the ideas presented. The NMSU Department of Biology systematics discussion group, Mark Simmons, Brook Milligan, Christine Bacon, Patrick Alexander, Bernard Pfeil, Timothy Wright, and Helga Ochoterena provided helpful comments and/or discussion. Support for this research was provided by the NMSU Department of Biology.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2005.11.020.

References

Adams, E.N., 1972. Consensus techniques and the comparison of taxonomic trees. Syst. Zool. 21, 390–397.

Bremer, K., 1990. Combinable component consensus. Cladistics 6, 369–372.

Campbell, A.M., 2000. Lateral gene transfer in prokaryotes. Theor. Popul. Biol. 57, 71–77.

Creevey, C.J., Fitzpatrick, D.A., Philip, G.K., Kinsella, R.J., O'Connell, M.J., Pentony, M.M., Travers, S.A., Wilkinson, M., McInerney, J.O., 2004. Does a tree-like phylogeny exist at the tips in the prokaryotes? Proc. R. Soc. Lond. B 22, 2551–2558.

Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6, 361–375.

Dress, A.W.M., Huson, D.H., Moulton, V., 1996. Analyzing and visualizing sequence and distance data using SplitsTree. Discrete Appl. Math. 71, 95–109.

Felsenstein, J., 1978. Cases in which parsimony and compatibility methods will be positively mis-leading. Syst. Zool. 27, 401–410.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Fitz-Gibbon, S.T., House, C.H., 1999. Whole genome-based phylogenetic analysis of freeliving microorganisms. Nucleic Acids Res. 27, 4218–4222.

Funk, V.A., 1985. Phylogenetic pattern and hybridization. Ann. Mo. Bot. Gard. 72, 681–715.

Gogarten, J.P., Doolittle, W.F., Lawrence, J.G., 2002. Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. 19, 2226–2238.

Goloboff, P., 2000. NONA: a tree searching program: available at http://www.cladistics.com.

Herniou, E.A., Luque, T., Chen, X., Vlak, J.M., Winstanley, D., Cory, J.S., O'Reilly, D.R., 2001. Use of whole genome sequence data to infer baculovirus phylogeny. J. Virol. 75, 8117–8126.

Huson, D.H., Steel, M., 2004. Phylogenetic trees based on gene content. Bioinformatics 20, 2044–2049.

Karlin, S., Mrazek, J., Campbell, A.M., 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. 179, 3899–3913.

Lake, J.A., Rivera, M.C., 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. Mol. Biol. Evol. 21, 681–690.

Linder, C.R., Rieseberg, L.H., 2004. Reconstructing patterns of reticulate evolution in plants. Am. J. Bot. 91, 1700–1708.

Martin, W., Muller, M., 1998. The hydrogen hypothesis for the first eukaryote. Nature 392, 37–41.

McDade, L.A., 1990. Hybrids and phylogenetic systematics I: patterns of character expression in hybrids and their implications for cladistic analysis. Evolution 44, 1685–1700.

McDade, L.A., 1992. Hybrids and phylogenetic systematics II: the impact of hybrids on cladistic analysis. Evolution 46, 1329–1346.

McInerney, J.O., Wilkinson, M., 2005. New methods ring changes for the ring of life. Trend Ecol. Evol. 20, 105–107.

Montague, M.G., Hutchison, C.A.I., 2000. Gene content phylogeny of herpesviruses. Proc. Natl. Acad. Sci. USA 97, 5334–5339.

Nixon, K.C., Wheeler, Q.D., 1992. Extinction and the origin of species. In: Wheeler, Q.D., Novacek, M. (Eds.), Extinction and phylogeny. Columbia University Press, New York, pp. 119–143.

Nixon, K.C., 1996. Paleobotany in cladistics and cladistics in paleobotany: enlightenment and uncertainty. Rev. Palaeobot. Palynol. 90, 361–373.

Nixon, K.C., Carpenter, J.M., 1996. On consensus, collapsibility, and clade concordance. Cladistics 12, 305–321.

Nixon, K.C., 2002. WinClada (Beta) version 1.00.08 Published by author. Ithaca, New York Shareware.

Ozkan, H., Levy, A.A., Feldman, M., 2001. Allopolyploidy-induced rapid genome evolution in the wheat (Aegilops—Triticum) group. Plant Cell 13, 1735–1747.

Rivera, M.C., Lake, J.A., 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431, 152–155.

Rohlf, F.J., 1982. Consensus indices for comparing classifications. Math. Biosci. 59, 131–144.

Siddall, M.E., 1998. Success of parsimony in the four-taxon case: long branch repulsion by likelihood in the Farris zone. Cladistics 14, 209–221.

Simonson, A.B., Servin, J.A., Skophammer, R.G., Herbold, G.W., Rivera, M.C., Lake, J.A., 2005. Decoding the genomic tree of life. Proc. Natl. Acad. Sci. USA 102, 6608–6613.

Snel, B., Bork, P., Huynen, M.A., 1999. Genome phylogeny based on gene content. Nat. Genet. 21, 108–110.

Song, K., Lu, P., Tang, K., Osborn, T.C., 1995. Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. Proc. Natl. Acad. Sci. USA 92, 7719–7723.

Spring, J., 2003. Major transitions in evolution by genome fusions: from prokaryotes to eukaryotes, metazoans, bilaterians and vertebrates. J. Struct. Funct. Genomics 3, 19–25.

Stoye, J., Evers, D., Meyer, F., 1998. Rose: generating sequence families. Bioinformatics 14, 157–163.

Wilkinson, M., 1994. Common cladistic information and its consensus representation: reduced adams and reduced cladistic consensus trees and profiles. Syst. Biol. 43, 343–368.