WILEY-BLACKWELL

Cladistics (2017) 1–4

Cladistics

10.1111/cla.12229

Parsimony analysis of unaligned sequence data: some clarifications

Jhon Jairo Ospina-Sarria[a,*] and Jimmy Cabra-García[a,b]

[a]Departamento de Zoologia, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP 05508-090, Brazil; [b]Departamento de Biología, Universidad del Valle, Cali, AA 25360, Colombia

Accepted 10 October 2017

Abstract

De Laet (2015) claimed that minimization of *ad hoc* hypotheses of homoplasy does not lead to a preference for trivial optimizations when analysing unaligned sequence data, as claimed by Wheeler (2012; see also Kluge and Grant, 2006). In addition, De Laet asserted that Kluge and Grant's (2006) parsimony rationale is internally inconsistent in terms of Baker's (2003) theoretical framework. We argue that De Laet used extraneous presuppositions to critique Wheeler's position and, as such, his criticism should be considered cautiously in terms of its scope. Finally, we demonstrate that considering Kluge and Grant's parsimony rationale as inconsistent rests on De Laet's misunderstanding of the ideographic character concept and the consequences of relating it to Baker's rationale.

© The Willi Hennig Society 2017.

De Laet (2015; DL henceforth) stated that it "is not correct" (p. 550) to affirm that minimization of *ad hoc* hypotheses of homoplasy always leads to a preference for trivial optimizations when analysing unaligned sequence data, as claimed by Wheeler (2012; see also Kluge and Grant, 2006). In addition, DL argued that applying Baker's (2003) philosophical framework to indels, as proposed by Kluge and Grant (2006), leads to a preference for explaining an indel of length *n* by one indel event of length *n*, rather than by *n* indel events of length one. As such, DL (p. 559) asserted that Kluge and Grant's (2006) parsimony rationale based on Baker's (2003) theoretical framework is "internally inconsistent". Considering DL's claims, where the epistemological justification of parsimony in phylogenetic inference is at stake, we assessed his affirmations in the context of Popper's (2009) characterization of scientific criticism.[1] We find that DL's position in relation to trivial alignments (i.e., alignments that are obtained by simply juxtaposing all observed sequences) does not reveal any internal contradiction within Wheeler's (2012) position and, as such, his criticism should be considered cautiously. Finally, we determine that the purported internal

*Corresponding author.
E-mail address: sarriajhon@gmail.com

© The Willi Hennig Society 2017

[1]Popper's philosophy of science has been discussed extensively within systematics (see, for example, Gattei, 2003; Rieppel, 2003a,b; Vogt, 2008, 2014; Kluge, 2009; Crother and Murray, 2015 and references therein) and diverse scientific disciplines (e.g., Hansson, 2006; Persson, 2016). We are aware of some of the alleged shortcomings of Popper's position that, according to some authors, render Popper's falsificationism fatally flawed (e.g., Vogt, 2014). Nevertheless, we are also aware that several theoretical components tightly related to Popper's philosophy of science, such as simplicity, unification and maximization of explanatory power, have been repeatedly recognized as epistemological virtues of scientific explanations (Farris, 1983; Koertge, 1992; Norton, 2000; Baker, 2003), and as such, we consider that Popper's (2009) characterization of scientific criticism, which allowed him to criticize the inductive logic and establish his demarcation criterion, is an extremely useful tool to consider in the epistemological debate.

inconsistency of Kluge and Grant's (2006) parsimony rationale rests on DL's misunderstanding of the consequences of integrating the ideographic character concept (see Grant and Kluge, 2004) with Baker's (2003) rationale.

According to Popper (2009), all scientific criticism consists of identifying contradictions. A contradiction may be a purely logical one, which can be demonstrated by highlighting internally inconsistent positions (logical method) or may be an empirical one, which should be demonstrated by contradiction with the facts—that is, with experience (empirical method; Popper, 2009). The logical and empirical methods of criticism may be called methods of immanent criticism. Popper (2009) also characterized a method of criticism called transcendent criticism, which is the confrontation of two different theses—using a contradiction between one position assumed to be true and another that is being criticized as evidence against the latter. Nevertheless, Popper (2009) warned that although the latter criticism method can sound persuasive and even be exceedingly illuminating, it will never be sufficient for a clear refutation. Having conceptualized Popper's (2009) characterization of scientific criticism, we will focus on DL's claims.

De Laet (2015) demonstrated (see also De Laet, 2005) that, under his interpretation of parsimony, trivial alignments would never be considered as optimal solutions. From this, DL concluded that asserting that minimization of *ad hoc* hypotheses of homoplasy (*sensu* Farris, 1983) always leads to a preference for trivial alignments, as claimed by Wheeler (2012; see also Kluge and Grant, 2006), is not correct. However, DL did not explicitly acknowledge that his criticism is based on an extraneous parameter (i.e., subcharacters; see the next paragraph) that is not relevant within Wheeler's (2012) explication of parsimony as minimization of total cost.

According to DL's understanding of parsimony, tree alignments that simultaneously minimize the total number of indels, substitutions and subcharacters maximize the amount of similarity that can be interpreted as homology. The number of subcharacters in a tree alignment amounts to the number of regions of the tree where a certain position is applicable and is a way to quantify the amount of compositional homology (i.e., base-to-base similarity; see also De Laet, 2005). Conversely, within Wheeler's (2012; see also Wheeler, 2001) dynamic homology framework, characters are treated as transformation events, not objects aligned according to similarity (see Kluge and Grant, 2006) and consequently subcharacters are meaningless. Wheeler's (2012) premise that minimization of *ad hoc* hypotheses of homoplasy (*sensu* Farris, 1983) always leads to a preference for trivial alignments is a logical consequence of the observation that, under the dynamic homology approach, the co-optimization of total steps and extra steps need not occur (Wheeler, 2012). That is, in traditional pre-aligned data there is a direct proportional relationship between tree length (i.e., total steps) and extra steps (i.e., homoplasy), and therefore minimizing total or extra steps makes no difference (Grant and Kluge, 2009; Wheeler, 2012). Nevertheless, when the alignment is allowed to vary this relationship is lost and minimizing steps is not the same as minimizing homoplasy, a fact recognized by Farris (2008, p. 829) as "well known". This aspect was precisely one of the motivations of Kluge and Grant (2006) to propose a novel justification of parsimony based on the anti-superfluity principle.

De Laet's (2015) criticism did not uncover any logical or empirical contradiction within the context of what is asserted by the thesis that is being criticized. Rather, DL criticized Wheeler's (2012) position by using a completely extraneous parameter, establishing a typical transcendental critique. As expected, several contradictions can emerge when a new parameter is quantified when comparing competing hypotheses, as DL underscored in his paper (see DL's figs 6, 7). Nevertheless, we consider that it is important to make clear that these differences are *a priori* expected because they are the consequence of applying two conflicting justifications of parsimony to the phylogenetic analysis of unaligned sequence data: minimizing *ad hoc* hypotheses of transformation events *sensu* Kluge and Grant (2006) versus maximizing the amount of similarity that can be interpreted as homology *sensu* DL.

De Laet's (2015) second claim is focused on the supposed internal inconsistency of Kluge and Grant's (2006) parsimony justification, as it pertains to Baker's (2003) rationale. According to DL, applying Baker's (2003) theoretical framework to indels leads to a preference for explaining an indel of length n by one indel event of length n, rather than by n indel events of length one. That said, we will now demonstrate that DL's criticism rests on a misunderstanding of the ideographic character concept[2] (ICC) originally advocated by Hennig (1966) and thoroughly discussed by Grant and Kluge (2004), and the consequences of relating the ICC to Baker's (2003) rationale.

[2]The ideographic character concept has been criticized by several authors (e.g., Assis and Brigandt, 2009). Nevertheless, these critiques do not focus its claims on the ontological status of the ideographic concept *per se* (see Göpel and Richter, 2016; for a recent endorsement of the ICC), but focus their attention on the utility of similarity as an empirical tool to recognize homology. Remarkably, Kluge and Grant (2006, p. 279) already commented on this aspect and succinctly stated: "The bottom line is that the *concept* of similarity is irrelevant to the evolutionary scientist. Similarity may be useful *operationally*, but only insofar as it facilitates the ostensive (by reference, pointing, or enumeration) or extensional (denotative) definition of character-states in the delimitation of what are hypothesized transformation series, of which the transformation event(s) are a part".

The example that Baker (2003) discussed is related to different competing explanations for the observation of a missing ½-spin following beta decay in the atomic nucleus. According to Baker (2003), the hypothesis that only postulates one neutrino with a spin of ½ (H1) maximizes the explanatory power and as such should be preferred over any other hypothesis which postulates $n > 1$ neutrinos each with a spin of $½_n$. As Baker (2003) underscored, H1 is able to explain the relevant available evidence (spin loss of ½) and even serve as the basis for a better explanation of non-observations. Moreover, Baker (2003: p. 257) clearly stated:

> Recall that the neutrino case study involves the postulation of a number of qualitatively identical individual particles which collectively explain some particular observed phenomenon. The explanation is additive in the sense that the overall phenomenon is explained by summing the individual positive contributions of each particle.

Baker (2003: pp. 257–258) also asserted that in cases where the postulated entities are not qualitatively similar (i.e., non-additive explanations), his rationale would not be applicable.

De Laet (p. 559) considered the phylogenetic explanation of different sequence lengths as a non-additive case, as he compared it with a non-additive example discussed by Baker (2003: p. 257). Hence, in DL's view, Baker's (2003) rationale is not applicable to the phylogenetic explanation of variation in DNA sequence length. Nevertheless, a key aspect misunderstood by DL is that, when the event-based ICC is considered, the additive condition as defined by Baker (2003) is satisfied.

Grant and Kluge's (2004) ICC defines characters as transformation series of which the transformation event(s) are a part. A transformation event (i.e., individual), which is a historical process that we can never directly observe, involves the modification of concrete objects which may be complex phenotypic characters or single nucleotides. Objects are causally related with events and are expressive of certain of the knowable characteristics of the event that can be exemplified in sense-experience (Woodger, 1929; Kluge and Grant, 2006). Nevertheless, only events, which are causally related to phylogeny, constitute the evidentiary entities of phylogenetic relationships; regardless of the nature of the objects that they may involve (e.g., one nucleotide, 100 nucleotides, tail morphology, head morphology), they are phenomenologically the same in the causal law of inheritance, and as such are considered identical and additive (Kluge and Grant, 2006). This rationale led Kluge and Grant (2006) to assert that by synthesizing the ICC, Baker's (2003) rationale and Farris's (1967) characterization of a phylogenetic hypothesis, **h**, it should be concluded that the explanatory power of **h** is maximized by minimizing the number of transformation events required to explain the character states of the terminal taxa as hypotheses of homology. Considering the case of the phylogenetic explanation of different sequence lengths as a nonadditive case, as DL asserted, implies a focus on objects and not on events.

Note that in the neutrino case described by Baker (2003), the relationship between object and event is necessarily one-to-one, because in the beta decay one neutrino (object) is emitted in each decay process (event); thereby the minimization of objects, and consequently causal events, imply the maximization of explanatory power. In the case of the molecular sequences, the background knowledge indicates that nucleotides (objects) are free to evolve independently of one another and indels (events) of various lengths can occur, so the relationship between event and object is not necessarily one-to-one. However, this difference does not prevent systematists from applying Baker's (2003) rationale to the phylogenetic explanation of different sequence lengths, because the event-based ICC clearly defines the evidentiary entities that need to be minimized in order to maximize the explanatory power.

Having characterized the phylogenetic explanation of differences in sequence length as an additive case, *contra* DL, in which the relationship between event and number of objects involved is unknown, it is pertinent to ask which position should be preferred *a priori* (i.e., before conducting a phylogenetic analysis): explaining an indel of length n by one indel event or by n indel events of length one? The former position, claimed by DL as a "more realistic view" (p. 559), actually goes beyond the limited background knowledge about changes of DNA sequence length throughout molecular evolution. As stated above, the background knowledge shows that single nucleotides are free to evolve independently, and that indels of different lengths can occur with different frequency (Zhang and Gerstein, 2003; Tanay and Siggia, 2008; Sung et al., 2016). Nevertheless, there is not enough empirical evidence to assume *a priori* that an indel of length n is *always* caused by one and only one indel event; in other words, that indel events *always* affect multiple nucleotides at once.

The latter position, which implies the *a priori* assumption of a one-to-one relationship between event and object (i.e., explaining an indel of length n by n indel events of length one), is able to explain all our observations using the minimum possible transformation event (i.e., an event involving one nucleotide), staying in the context of the limited background knowledge. It is important to make clear that the *ad hoc* assumption of a one-to-one relationship is not equivalent to asserting that indel events *always* affect one and only one base at a time. In fact, this assumption may allow systematists to detect *a posteriori* (i.e.,

after conducting the phylogenetic analysis) cases of non-independence in insertion or deletion events (i.e., one indel event involving several nucleotides). Nevertheless, these cases cannot be accepted as background knowledge in subsequent phylogenetic analyses, because this position entails a bias toward a subset of hypotheses (Grant and Kluge, 2005). In other words, this position does not allow the possibility of detecting cases of independence among nucleotides in insertion or deletion events. Conversely, the assumption of a one-to-one relationship, which stays in the context of the limited background knowledge and as such represents the most severe test, should always be used in additional cycles of reciprocal illumination (Kluge, 1998).

Considering parsimony as the minimization of *ad hoc* hypotheses of transformation events, or the maximization of the amount of similarity that can be interpreted as homology, is not a trivial issue and strong empirical effects can emerge when applying these contrasting epistemological justifications in phylogenetic inference. Both positions agree in that minimization of *ad hoc* hypothesis is a theoretical virtue of phylogenetic explanations, but they differ radically in their treatment of similarity. We consider that the epistemological debate concerning parsimony in phylogenetic inference will further benefit when clearly differentiating transcendental from immanent criticisms and when embracing the event-based ICC and its consequences.

Acknowledgements

We are grateful to T. Grant, W. Wheeler, A. Kluge and L. Vogt for discussion and comments on an early draft of the manuscript. We also thank two anonymous reviewers and the associate editor Mark P. Simmons for their suggestions and comments. It goes without saying, however, that we are solely responsible for all the arguments and statements in this paper. Brittany Damron improved the English text, which is gratefully acknowledged. Financial support: FAPESP: 2012/10000-5, 2014/03585-2, and 2016/25070-0 (J.J.O.S.); 2013/20262-0 and 2016/04437-2 (J.C.G.).

References

Assis, L.C.S., Brigandt, I., 2009. Homology: homeostatic property cluster kinds in systematics and evolution. Evol. Biol. 36, 248–255.

Baker, A., 2003. Quantitative parsimony and explanatory power. Br. J. Philos. Sci. 54, 245–259.

Crother, B.I., Murray, M., 2015. Testable but not falsifiable? Cladistics 31, 573–574.

De Laet, J., 2005. Parsimony and the problem of inapplicables in sequence data. In: Albert, V.A. (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, Oxford, pp. 81–116.

De Laet, J., 2015. Parsimony analysis of unaligned sequence data: maximization of homology and minimization of homoplasy, not minimization of operationally defined total cost or minimization of equally weighted transformations. Cladistics 31, 550–567.

Farris, J.S., 1967. The meaning of relationship and taxonomic procedure. Syst. Zool. 16, 44–51.

Farris, J.S., 1983. The logical basis of phylogenetic analysis. In: Platnick, N.I., Funk, V.A. (Eds.), Advances in Cladistics II. Columbia University Press, New York, NY, pp. 7–36.

Farris, J.S., 2008. Parsimony and explanatory power. Cladistics 24, 825–847.

Gattei, S., 2003. Reply to Olivier Rieppel. Cladistics 19, 172.

Göpel, T., Richter, S., 2016. The word is not enough: on morphemes, characters and ontological concepts. Cladistics 32, 682–690.

Grant, T., Kluge, A.G., 2004. Transformation series as an ideographic character concept. Cladistics 20, 23–31.

Grant, T., Kluge, A.G., 2005. Stability, sensitivity science and heurism. Cladistics 21, 597–604.

Grant, T., Kluge, A.G., 2009. Parsimony, explanatory power, and dynamic homology testing. Syst. Biodivers. 7, 357–363.

Hansson, S.O., 2006. Falsificationism falsified. Found. Sci. 11, 275–286.

Hennig, W., 1966. Phylogenetic Systematics. University Illinois Press, Urbana, IL.

Kluge, A.G., 1998. Sophisticated falsification and research cycles: consequences for differential character weighting in phylogenetic analysis. Zool. Scr. 26, 349–360.

Kluge, A.G., 2009. Explanation and falsification in phylogenetic inference: exercises in Popperian philosophy. Acta. Biotheor. 57, 171–186.

Kluge, A.G., Grant, T., 2006. From conviction to anti-superfluity: old and new justifications of parsimony in phylogenetic inference. Cladistics 22, 276–288.

Koertge, N., 1992. Explanation and its problems. Br. J. Philos. Sci. 43, 85–98.

Norton, J.D., 2000. Nature is the realisation of the simplest conceivable mathematical ideas: Einstein and the canon of mathematical simplicity. Stud. Hist. Phil. Mod. Phys. 21, 135–170.

Persson, U., 2016. Is falsification falsifiable? Found. Sci. 21, 461–475.

Popper, K.R., 2009. The Two Fundamental Problems of the Theory of Knowledge. Routledge, London, UK.

Rieppel, O., 2003a. Popper and systematics. Syst. Biol. 52, 259–271.

Rieppel, O., 2003b. Gattei on Popper and truth. Cladistics 19, 170–171.

Sung, W., Ackerman, M.S., Dillon, M.M., Platt, T.G., Fuqua, C., Cooper, V.S., Lynch, M., 2016. Evolution of the insertion–deletion mutation rate across the tree of life. G3 6, 2583–2591.

Tanay, A., Siggia, E.D., 2008. Sequence context affects the rate of short insertions and deletions in flies and primates. Genome Biol. 9, R37.

Vogt, L., 2008. The unfalsifiability of cladograms and its consequences. Cladistics 24, 62–73.

Vogt, L., 2014. Popper and phylogenetics, a misguided rendezvous. Austral. Syst. Bot. 27, 85–94.

Wheeler, W.C., 2001. Homology and the optimization of DNA sequence data. Cladistics 17, S3–S11.

Wheeler, W.C., 2012. Trivial minimization of extra-steps under dynamic homology. Cladistics 28, 188–189.

Woodger, W.C., 1929. Biological Principles: A Critical Study. Harcourt Brace, New York, NY.

Zhang, Z., Gerstein, M., 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res. 31, 5338–5348.