IMPROVING PHYLOGENY RECONSTRUCTION FOLLOWING THE REQUIREMENT OF TOTAL EVIDENCE BY EXCLUDING DATA PRIOR TO SIMULTANEOUS ANALYSIS

Guillaume Lecointre and Pierre Deleporte
In Prep. (Communication at the XVIIIth meeting of the Willi Hennig Society, Goettingen, Sept. 1999)

Reducing the application of the « requirement of total evidence » (sensu Carnap, 1950) to a particular mode of phylogenetic analysis (the « combined analysis » of Bull et al., 1993, i.e. the « simultaneous analysis » of Nixon and Carpenter, 1996) introduces a confusion between the general principle and one of its possible procedures (e.g. in Kluge, 1989). After all, a consensus tree takes into account all available evidence, some way. To refer to procedures, the terms « simultaneous analysis » and « separate analyses » (Nixon and Carpenter, 1996) are preferred.

According to Carnap (1950), a part of the a priori available evidence can nevertheless be removed if one can justify that the removed data are inductively irrelevant with respect to the hypothesis at hand. Following this logic, we argue that the best processing of data for phylogenetic investigation is in three steps. First, separate analyses of the data (without consensus trees, Barrett et al., 1991), in order to detect tree reconstruction artifacts. Second, testing significance of character incongruence, for example using the ILD test (Farris et al., 1995). Third, performing simultaneous analysis (if allowed) in which, if necessary, some data are replaced by question marks. For instance, molecular sequences with an aberrant rate of change provoking a long-branch attraction artifact detected in separate analyses are replaced by question marks in the combined matrix. Another example: if the null hypothesis of character congruence is rejected by the application of the ILD test, iterative removal of taxa followed by new ILD tests (as in Lecointre et al., 1998) allow to identify the sequences responsible for significant incongruence. These sequences are also replaced by question marks in the combined matrix because possibly having experienced a process of discord sensu Maddison (1997). Such a strategy is aimed to optimally integrate all the evidence through simultaneous analysis, but only relevant evidence. Irrelevant data are 1. those neutral with respect to the hypothesis at hand (neutral ones, e.g. Mark Siddall doesn't like MacIntoshs), 2. those obscuring the answer (misleading ones, e.g. horizontal transfers). This point will be discussed. Problems linked to possible arbitrariness in the division of the data sets remains and will be discussed.Examples will be shown.

AMÉLIORER LA RECONSTRUCTION PHYLOGÉNÉTIQUE SUIVANT L'EXIGENCE DE « TOTAL EVIDENCE » EN EXCLUANT DES DONNÉES AVANT L'ANALYSE SIMULTANNÉE

Guillaume LECOINTRE et Pierre DELEPORTE
En préparation (communication donnée au dixhuitième congrès de la Willi Hennig Society, Goettingen, Sept. 1999)

Réduire le « requirement of total evidence » (au sens de Carnap, 1950) à un mode particulier d'analyse phylogénétique (l'analyse combinée de Bull et al., 1993, ou « analyse simultanée » de Nixon et Carpenter, 1996) introduit une confusion entre le principe général et l'une de ses procédures possibles (comme dans Kluge, 1989). Après tout, un arbre de consensus strict peut prendre en compte toute "l'évidence" disponible, d'une certaine manière. Pour se référer à des procédures, nous préférons les termes de « analyse simultanée » et « analyses séparées » de Nixon et Carpenter (1996). Selon Carnap (1950), une partie des données peut être écartée si l'on peut justifier pourquoi elles sont inductivement hors de propos (« inductively irrelevant ») au regard de l'hypothèse sur laquelle on doit statuer. En suivant cette logique, nous pensons que le meilleur traitement des données pour la reconstruction phylogénétique comporte trois étapes. Premièrement, procéder à des analyses séparées des données (sans arbre-consensus, Barrett et al., 1991), afin de détecter des artefacts de reconstruction. Deuxièmement, tester l'incongruence des caractères entre différents jeux de données, en utilisant par exemple le test ILD (Farris et al., 1995). Troisièmement, procéder à l'analyse simultanée (s'il est raisonnable de la faire) dans laquelle, si nécessaire, certaines données sont remplacées par des points d'interrogation. Par exemple, pour des données moléculaires, les portions de séquences aux taux d'évolution aberrants provoquant des artefacts d'attraction de branches longues détectés en analyses séparées sont remplacées par des points d'interrogation dans le jeu de données combinées. Autre exemple, si l'hypothèse nulle de congruence est rejetée à la suite du test ILD, des retraits itératifs de taxons suivis de tests ILD (comme dans Lecointre et al., 1998) permettront d'identifier celui des taxons responsable de l'incongruence : ainsi, ces séquences sont également remplacées par des points d'interrogation dans le jeu de données combinées, car elles ont probablement subi un « processus de discorde » au sens de Maddison (1997). Cette stratégie vise finalement à intégrer de manière optimale toute l'information disponible par l'analyse simultanée, mais seulement l'information raisonnablement pertinente au regard de la question posée. Les données non pertinentes sont 1. celles qui sont neutres vis-à-vis de cette question (par exemple, Pascal Tassy aime Tintin et Milou), et 2. celles qui obscurcissent la réponse à cette question (par exemple, les transferts horizontaux). Ce point sera discuté. Il demeure un problème de possible arbitraire dans la division des jeux de données (impliquant une « pondération implicite non contrôlée », ou l'association données informatives et trompeuses dans un même lot) qui sera également discuté. Des exemples seront fournis.