RESUME DE L'HABILITATION À DIRIGER DES RECHERCHES

délivrée par L'UNIVERSITÉ PARIS XI à Guillaume LECOINTRE, Maître de Conférences, Laboratoire d'Ichtyologie générale et appliquée et Service de Systématique moléculaire du Muséum, Muséum National d'Histoire Naturelle, 43 rue Cuvier, 75231 PARIS cedex 05

« LES RAPPORTS ENTRE L'HOMOPLASIE ET L'INCONGRUENCE DES CARACTERES EN RECONSTRUCTION PHYLOGENETIQUE »

Habilitation soutenue le 11 Décembre 1998 devant la commission d'examen composée de Hervé LE GUYADER, Professeur à PARIS XI, Pierre DELEPORTE, CR1 CNRS, Université de Rennes I, Michel MILINKOVITCH, Assoc. Prof. Université libre de Bruxelles, Armand de RICQLÈS, Professeur au Collège de France, André ADOUTTE, Professeur à PARIS XI, Jean DEUTSCH, Professeur à PARIS VI, Simon TILLIER, Professeur au Muséum National d'Histoire Naturelle

La nécessité d'évaluer la non congruence des caractères issus de corps de données indépendants apparaît comme l'une des issues du débat qui anime le milieu phylogénéticien depuis 10 ans, entre partisans de la " congruence taxonomique " et partisans de la " congruence des caractères " (" total evidence ", Carnap, 1950 ; Kluge, 1989). Actuellement, les partisans du « total evidence » semblent s'aligner sur l'approche du " prior agreement " ou encore " conditional combination " qui consiste à tester la non congruence des caractères à l'aide de tests appropriés avant de combiner les jeux de données dans une seule et même analyse phylogénétique. En effet, en raison d'un processus spécifique à l'une des partitions, il est possible que son histoire suive un " chemin " différent que celui des taxons qui la portent. C'est par exemple ce qui se passe lorsque des transferts horizontaux se produisent par recombinaison chez les bactéries ou par hybridation introgressive chez les Cyprinidae. Dans ces cas-là, la phylogénie obtenue sur la base d'un premier gène exempt de transferts peut légitimement ne pas être celle obtenue sur la base d'un second gène dont l'histoire n'est pas le traceur fidèle de celle des espèces qui le portent. Il en résulte une non congruence des caractères statistiquement significative entre les deux jeux de données. Si l'on ne veut perdre aucune des deux histoires, il ne faut pas combiner les jeux de données. Il est donc préférable de vérifier l'absence d'incongruence de ce type avant de procéder à la combinaison des données en présence.

Le travail de cette habilitation a consisté à examiner l'impact de l'homoplasie sur la mesure de la congruence des caractères, à utiliser cette mesure pour tester la non congruence entre partitions ayant subi des contraintes sélectives différentes, et enfin à définir conditions et méthode de combinaison à employer dans le cas de jeux de données ayant subi des transferts horizontaux. Enfin, les rapports entre mesure de la non congruence, mesures de l'homoplasie et pondération ont été explorés.

L'impact de l'homoplasie sur la mesure de la non congruence des caractères, telle que la produit le test ILD de Farris et al. (1995), a été évalué par simulation. Cet impact n'est pas intelligible lorsque l'homoplasie est perçue à travers ses mesures globales que sont le C.I. et le R.I., mais il le devient lorsque l'homoplasie est vue à travers ses causes, c'est-à-dire l'hétérogénéité des taux d'évolution entre sites (Yang, 1996). Une forte hétérogénéité des taux est le facteur qui va générer une non congruence significative des caractères, même artéfactuelle, c'est-à-dire même si les arbres " vrais " ayant servi à simuler les deux jeux de données sont les mêmes (on pourrait appeler cet artéfact un « faux négatif »). Par ailleurs, Sullivan (1996) souligne que si les hétérogénéités des deux jeux de données sont très différentes, on a toutes les chances d'avoir une non congruence des caractères significative. Cependant, des différentiels d'hétérogénéité élevés vont aussi provoquer l'occurrence de " faux positifs ", c'est-à-dire l'impossibilité de rejeter l'hypothèse nulle de congruence alors que les histoires sous-jacentes sont différentes ; ce paradoxe méritait d'être signalé.

Qui dit hétérogénéités différentes dit pressions sélectives sous-jacentes très différentes. Le pari de Bull et al. (1993) postule que des partitions peuvent apparaître non congruentes entre elles parce qu'elles subissent des pressions sélectives très différentes. Un second travail a consisté à tester sur des séquences mitochondriales de Crotalinae la non congruence de caractères subissant des pressions sélectives diverses, par exemple entre les premières, secondes et troisièmes positions de codon d'un même gène. Il ressort que la prédiction de Bull et al. (1993) est confirmée. La non congruence entre positions du codon d'un même gène peut être plus forte que la non congruence entre positions du codon analogues entre gènes de fonctions différentes.

Le travail a ensuite consisté à rechercher les conditions de combinaison pour des jeux de données susceptibles d'avoir subi des transferts horizontaux générant des non congruences « légitimes » de caractères entre jeux de données. Une méthode simple a été proposée (Lecointre et al., 1998), elle a permis d'établir une première phylogénie fiable à l'intérieur de l'espèce *Escherichia coli*, c'est-à-dire exempte de transferts horizontaux. Il faut préciser que *E. coli* n'est pas une espèce complètement clonale et qu'il était jusqu'à présent très difficile d'interpréter les phylogénies moléculaires des souches de référence en raison de la recombinaison bactérienne. La méthode proposée a également permis d'établir une phylogénie des Cyprinidae où des hybridations intergénériques sont possibles. L'inférence par parcimonie des transferts horizontaux subis par les gènes *mut* chez *E. coli* ont permis d'apporter une confirmation phylogénétique du modèle « d'évolution par bouffées » des souches d' *Escherichia coli.*

Il ressort en conclusion générale qu'il serait bon de ne pas suivre une stratégie de " total evidence " sans prendre ses précautions. Comme l'ont suggéré Larson (1994), de Queiroz et al. (1995) et Sullivan (1996), il est souhaitable de procéder à la fois aux analyses séparées et à l'analyse combinée, avec si possible un test ILD (ou autre test évaluant la non congruence *des caractères*) entre les deux. Cette stratégie permet d'évaluer la non congruence des caractères et de tirer profit de l'analyse de leur

congruence taxonomique. En effet, on montre à l'aide de données réelles que la non congruence taxonomique, évaluée par la simple observation des arbres tirés de chaque gène, révèle dans les meilleurs cas l'attraction de branches longues que le test ILD ne saurait révéler. Elle permet de mettre en oeuvre *un nouveau type de " combinaison conditionnelle ",* celle qui ôterait de la future combinaison (1) ceux des taxons qui provoquent une non congruence significative et (2) ceux des gènes pour certains taxons qui montrent une inégalité de taux d'évolution (une " branche longue "). On peut donc garder le taxon pour ses gènes aux taux similaires à ceux des autres taxons et remplacer le(s) gène(s) rapide(s) pour ce taxon par des points d'interrogation. Un retour aux textes originaux de Carnap (1950) atteste qu'une telle démarche n'est pas contraire au « requirement for total evidence » de cet auteur, tant que l'on justifie pourquoi une partie des données est « inductively irrelevant ». Signalons à cet égard que Kluge (1989) s'était trompé en présentant le « requirement of total evidence » de Carnap (1950) comme une procédure (en fait « l'analyse simultannée » de Nixon et Carpenter, 1996) au lieu de la présenter comme un principe ; et en comprenant « evidence » chez Carnap (1950) comme « data » alors qu'une relecture détaillée de Carnap indique qu'il fallait comprendre « knowledge ». Sous cet angle, Carnap ne préconisait pas, bien entendu, de mettre toutes les données dans le même sac, mais demandait à ce que tous les indices, toutes les connaissances soient prises en compte pour déterminer celles des données qui sont pertinences au regard de la question posée, et celles qui ne le sont pas. L'exclusion des données non-pertinentes (par exemple gènes importés par transfert horizontal) est donc requise par le « requirement for total evidence ». Après 10 ans de faux débats sur le « total evidence », cette méprise méritait d'être signalée.

La plupart des pondérations que l'on met en oeuvre en systématique moléculaire visent à limiter l'impact de l'homoplasie, soit en affectant un poids de zéro à celles des positions du gène qui sont saturées en mutations, soit en affectant un poids moindre à une position dotée d'un faible C.I. ou R.I. ; ceux-ci étant mesurés tous types de substitutions confondus. Comme le confirment conjointement les travaux de cette HDR et ceux d'Allard et Carpenter (1996), pondérer dans le but d'ôter de l'homoplasie n'améliore ni la congruence entre les partitions ni la résolution de l'arbre final ; car mesures d'homoplasie et mesures de non congruence sont découplées :

1. Découplage Saturation/R.I. : si un jeu de données (ou les sites d'une position du codon) est saturé en mutations, sont C.I. et son R.I. seront faibles. Mais à l'inverse, des C.I. et R.I. faibles peuvent être obtenus sur des données non saturées.

2. Découplage R.I./non congruence : Le R.I. mesure l'homoplasie *globalement* tandis que le test ILD réagit à la première non congruence locale mais significative. Par conséquent, la valeur critique du test ILD « alfa de Farris ou P-value de Swofford) et celle du R.I. ou du C.I. de chacun des deux jeux de données (ou le différentiel en R.I. des deux jeux) ne sont pas corrélées. Ceci a été montré en simulation comme sur les données réelles.

3. Découplage Saturation/ non congruence : Vidal et Lecointre (1998) montrent que deux types de substitutions également saturées (TS3 de cytb et TS3 de ND4) se comportent très différemment en termes de congruence vis à vis de chaque autre partition contre lesquelles elles sont testées.

Aucune des trois mesures -R.I., saturation, non congruence- n'est superposable aux deux autres, et à peine déductible des deux autres. *Il faudra donc choisir : pondérer en fonction de l'une de ces deux mesures de l'homoplasie ou en fonction de la non congruence ?* L'efficacité des méthodes de pondération visant à diminuer l'impact de l'homoplasie est faible ou nulle, en termes de gain de résolution de l'arbre le plus parcimonieux ou en termes de gain de robustesse, et les raisons en sont rappelées. Le présent travail, accompagné des conclusions de Philippe et al. (1996), tendrait à laisser de côté toute pondération visant à diminuer l'homoplasie ou la saturation au profit d'une pondération visant à améliorer la congruence entre les partitions de caractères. Ceci peut être fait soit en pondérant comme expliqué dans Hassanin et al. (1998) ; soit en affectant un poids de zéro aux partitions que le test ILD aura montré comme non congruentes avec au moins l'une des autres, soit en ôtant seulement celui ou ceux des taxons responsables de cette non congruence. Cette " philosophie " de la pondération est applicable à l'intérieur d'un même jeu de données comme entre jeux de données afin de les combiner. Dans ce dernier cas, coupler l'analyse de la congruence des caractères à celle de la congruence taxonomique permet d'ôter non seulement les taxons/gènes non congruents mais aussi ceux qui sont sujets à des artefacts de reconstruction (notamment s'affranchir des attractions de branches longues révélées par la congruence taxonomique). Dans un tel schéma, chaque retrait de données se trouve alors justifié. La complexité des processus biologiques nous enseigne qu'un " total evidence aveugle " est très risqué : la combinaison des données doit être réalisée après une série de vérifications. Oui au principe du " total evidence ", à condition que son application brutale ne nous prive pas du principal objectif de tout systématicien : obtenir une hiérarchie du vivant qui, tout en restant ouverte aux tests, puisse être suffisamment dépourvue d'artefacts de toutes sortes pour qu'à terme puisse émerger un consensus entre chercheurs.

Guillaume LECOINTRE

Allard, M.W. and Carpenter, J.M. 1996. On weighting and congruence. *Cladistics* 12: 183-198.

Bull, J.J., Huelsenbeck, J.P., Cunningham C.W., Swofford, D.L., and Waddell P.J. 1993. Partitioning and combining data in phylogenetic analysis. *Syst. Biol.* 42 (3): 384-397.

Carnap, R. 1950. Logical foundations of probability. Univ. of Chicago Press, Chicago.

de Queiroz, A., Donoghue, M.J. and Kim, J. 1995. Separate versus combined analysis of phylogenetic evidence. *Annu. Rev. Ecol. Syst.* 26: 657-681.

Farris, J.S., Källersjö, M., Kluge, and Bult C. 1995. Testing significance of incongruence. *Cladistics* 10: 315-319.

Hassanin, A., Lecointre, G. and Tillier, S. 1998. The " Evolutionary Signal " of homoplasy in Protein coding gene sequences and its consequences for a priori weighting in phylogeny. *Comptes Rendus de l'Académie des Sciences* 321: 611-620.

Kluge, A.G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Syst. Zool*. 38(1): 7-25.

Larson, A. 1994. The comparison of morphological and molecular data in phylogenetic systematics. *In* Schierwater B., Streit, B., Wagner G.P., and Desalle, R. Molecular Ecology and Evolution. Approaches and Applications. Birkhauser Verlag. Basel. Switzerland. pp. 371-390.

Lecointre, G., Rachdi, L., Darlu, P. and Denamur, E. 1998. *Escherichia coli* molecular phylogeny using the Incongruence Length Difference test. *Molecular Biology and Evolution.* 15 (12) : 1685-1695.

Philippe, H., Lecointre, G., Lê, H.L.V. and Le Guyader, H. 1996. A critical study of homoplasy in molecular data using morphologically based cladograms and its consequences for character weighting. *Molecular Biology and Evolution* 13 (9) : 1174-1186.

Sullivan, J. 1996. Combining data with different distributions of among-site rate variation. *Syst. Biol*. 45(3): 375-380.

Vidal, N. et Lecointre, G. 1998. Weighting and congruence : a case study based on three mitochondrial genes in pitvipers. *Molecular Phylogenetics and Evolution* 9 (3): 366-374.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol* (11):105-109.