Congruence entre des matrices de distance

Pierre Legendre, François-Joseph Lapointe

Département de sciences biologiques,
Université de Montréal,
Case postale 6128, succursale Centre-ville,
Montréal (Québec) Canada, H3C 3J7

RÉSUMÉ. Cet article décrit un test de congruence entre plusieurs matrices de distance (CEMD) provenant de tableaux de données destinés à être utilisés ensemble dans des analyses de données subséquentes. Ce test, qui utilise la statistique W *de Kendall, constitue une généralisation du test de Mantel au cas de plusieurs matrices. Il est appliqué ici pour la première fois pour étudier la congruence de plusieurs gènes. Le résultat permet de décider s'il convient, ou non, de les utiliser ensemble dans une même analyse phylogénétique.*

MOTS-CLÉS : analyse phylogénétique, coefficient de concordance de Kendall, combinaison de données, congruence, gènes, matrices de distance.

1 Introduction

Cet article décrit un test de congruence entre plusieurs matrices de distance provenant de tableaux de données destinés à être utilisés ensemble dans d'autres analyses de données comme des analyses de classification, des analyses phylogénétiques, ou encore des ordinations simples ou canoniques. Ce test, décrit dans [LEG 04], constitue une généralisation du test de Mantel [MAN 67] au cas de plusieurs matrices de distance. Il sera appliqué ici pour la première fois pour étudier la congruence de plusieurs gènes, afin de décider s'il convient de les utiliser ensemble pour réaliser une analyse phylogénétique.

En analyse phylogénétique, l'incongruence entre des tableaux de données portant sur les mêmes taxa peut avoir des origines diverses [WEN 98]. Même si on admet que les organismes ont une seule histoire évolutive, les données morphologiques, sérologiques et moléculaires peuvent conduire à des reconstructions phylogénétiques différentes, à cause par exemple de la convergence des caractères morphologiques. De même, les gènes nucléaires, mitochondriaux et chloroplastiques peuvent avoir des histoires évolutives différentes. Certains gènes peuvent résulter de transferts latéraux entre des branches distinctes de l'Arbre de la Vie. L'incongruence est une réalité quotidienne dans le travail des phylogénéticiens. La méthode décrite dans cet article permet de détecter la congruence et de décider s'il convient d'utiliser toute l'information dans une seule reconstruction d'arbre phylogénétique ou de réaliser des analyses séparées sur les différents jeux de données.

Il existe plusieurs méthodes statistiques permettant de tester l'incongruence des données en analyse phylogénétique (par exemple [FAR 95]). Ces méthodes se limitent cependant à la comparaison de deux jeux de données à la fois, ne s'appliquent pas aux matrices de distance et reposent sur le principe de la parcimonie. Notre approche permet de comparer plusieurs jeux de données, présentés sous la forme de distances ou non, ainsi que des matrices d'arbres reconstruits avec d'autres approche que la parcimonie.

2 Le test CEMD

L'analyse de la *congruence entre des matrices de distance* (CEMD) prend pour point de départ p matrices de distance, ou encore p tableaux de données décrivant n objets à l'aide de m_1, m_2, …, m_p variables (gènes, nucléotides ou bases en génétique, ou autres types de variables dans d'autres domaines d'application). Le test procède comme suit. Pour tester l'hypothèse nulle (H_0 : incongruence de toutes les matrices de distance) soumises au test contre l'hypothèse contraire (H_1 : au moins deux de ces matrices sont congruentes).

1. Si on désire étudier des tableaux de données brutes, on calcule, pour chaque tableau, une matrice de distance appropriée aux données qu'il contient. La mesure de distance peut varier d'un tableau à l'autre.
2. On déplie la portion triangulaire (supérieure ou inférieure) de chaque matrice de distance en un vecteur. On écrit ces vecteurs dans les lignes successives d'une matrice de travail qui comporte p lignes et $n(n–1)/2$ colonnes. Dans le cas des matrices de distance asymétriques, on écrit les portions triangulaires supérieure et inférieure de la matrice de distance dans la matrice de travail.
3. On transforme les distances en rangs, ligne par ligne.
4. On calcule le coefficient de concordance W de Kendall entre les lignes de la matrice de travail. On transforme W en une statistique χ^2 de Friedman qui fournit la valeur de référence (χ^2_{ref}) du test statistique.
5. La statistique est testée par permutations. Pour ce faire, on permute chaque matrice de distance comme dans le test de Mantel [MAN 67, LEG 00], et ce indépendamment d'une matrice à l'autre. On permute les objets de la matrice au hasard et on récrit les distances dans la matrice de travail dans l'ordre correspondant aux objets permutés. On calcule la statistique χ^2 pour les données permutées.
6. On répète l'étape 5 un grand nombre de fois (par ex. 999 fois). On assemble toutes les valeurs χ^2 ainsi obtenues, y compris la valeur χ^2_{ref}, dans une distribution, et on détermine combien de ces valeurs sont supérieures ou égales à la valeur χ^2_{ref}. La probabilité unilatérale des données sous l'hypothèse nulle du test est obtenue en divisant cette valeur par le nombre de permutations plus une (par ex. 1000).

Lorsque le test global est significatif, cela indique qu'il y a au moins deux matrices qui sont congruentes. On peut réaliser des tests *a posteriori* de la contribution de chaque matrice à la statistique χ^2 en permutant une seule matrice à la fois. Une matrice qui n'est pas congruente avec d'autres n'aura, une fois permutée, que peu d'effet sur la statistique globale. L'hypothèse nulle de ce test est H_0 : incongruence de cette matrice de distance par rapport à toutes les autres. La méthode donne un poids égal à toutes les matrices de distance dans la procédure de test. Une version pondérée du test permet de donner des poids inégaux aux matrices, par exemple en fonction du nombre de gènes ou de bases que contient chaque tableau de données [LEG 04]. Le test de Mantel entre les matrices de distance transformées en rangs fournit de l'information complémentaire, cette fois au niveau de chaque paire de tableaux de données.

3 Exemple

Mark Springer a obtenu 8 matrices de distances patristiques représentant des arbres phylogénétiques obtenus par analyse phylogénétique de gènes séquencés sur des animaux représentant 11 ordres de mammifères [SPR 99]. Les gènes sont les suivants : (1) 12S, (2) valine, (3) 16S, (4) vwf, (5) irbp, (6) a2ab, (7) cytb et (8) aquaporine. Les gènes 1, 2, 3 et 7 font partie de l'ADN mitochondrial alors que les gènes 4, 5, 6 et 8 proviennent de l'ADN nucléaire. Nous allons calculer des statistiques de Mantel et réaliser des analyses de congruence entre ces matrices de distance afin de déterminer si elles pourraient être utilisés ensemble dans une méthode de consensus pour reconstruire la phylogénie des ordres de mammifères.

Une première analyse CEMD a permis de rejeter l'hypothèse d'indépendance des 8 matrices de distance ($P = 0.0001$ après 9999 permutations des distances à l'intérieur de chaque matrice). Les tests *a posteriori* rejettent l'hypothèse d'incongruence des matrices individuelles, à l'exception des matrices 2 et 7 ($P = 0.0994$ après correction de Holm pour 8 tests simultanés). Les corrélations de Mantel sont faibles entre ces deux matrices et les matrices 4, 5, 6 et 8. Cela indique que le groupe de 8 matrices n'est pas homogène.

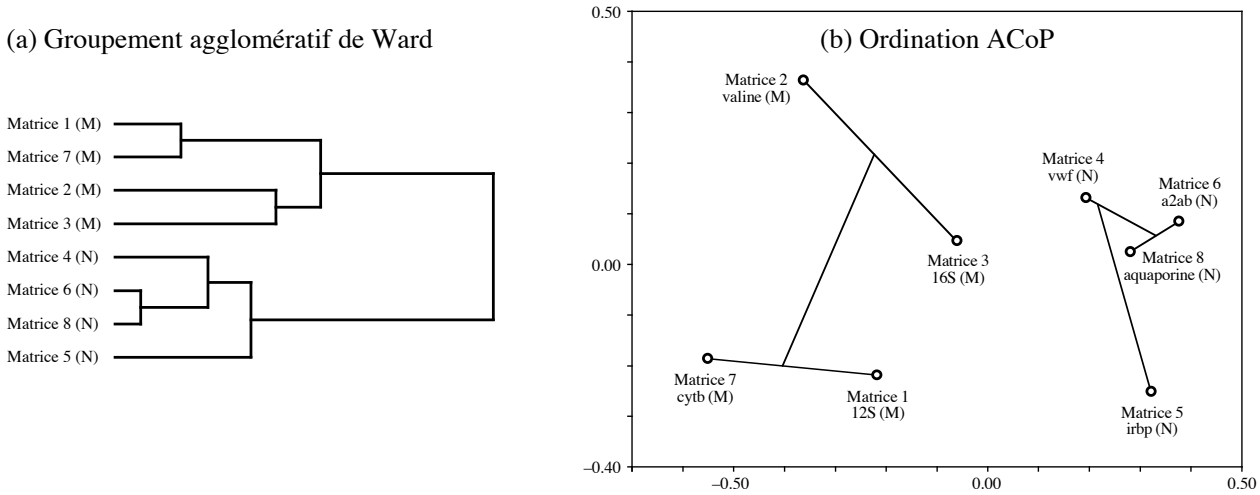

Figure 1. Groupement et ordination des matrices de distance représentant les 8 arbres phylogénétiques, basés sur les corrélations de Spearman entre matrices de distance. (a) Groupement agglomératif de Ward. (b) Ordination par analyse en coordonnées principales (ACoP). La topologie du groupement est dessinée sur l'ordination, à l'exclusion de la fusion finale. N : gène nucléaire. M : gène mitochondrial.

Un groupement agglomératif de Ward a permis d'identifier deux groupes (Fig. 1a) : les matrices 1, 2, 3 et 7 d'une part qui représentent les arbres dérivés des gènes mitochondriaux et les matrices 4, 5, 6 et 8 d'autre part qui représentent les arbres obtenus des gènes nucléaires. Puisque les statistiques de Mantel sont toutes positives, elles se comportent comme des similarités. Après transformation des similarités en distances, une ordination par analyse en coordonnées principales du tableau des statistiques de Mantel montre clairement la séparation des matrices en deux groupes (Fig. 1b).

Tableau 1. Résultats des tests CEMD globaux ainsi que des tests *a posteriori* portant sur les matrices de distance individuelles. P = probabilité (après 9999 permutations) sans correction, P_H = après correction de Holm pour 4 tests simultanés.

(a) Matrices correspondant aux gènes 1, 2, 3 et 7	**(b) Matrices correspondant aux gènes 4, 5, 6 et 8**
Test CEMD global. H_0 : les 4 mat. sont incongruentes	Test CEMD global
Statistique W de Kendall = 0.64755	Statistique W de Kendall = 0.78021
χ^2 de Friedman = 139.87013, P=0.0001 (rejet H_0)	χ^2 de Friedman = 168.52597, P=0.0001 (rejet H_0)
Tests *a posteriori*. H_0 : incongruence de cette matrice	Tests *a posteriori*
Matrice 1 : P = 0.0003, P_H = 0.0012 (rejet de H_0)	Matrice 4 : P = 0.0001, P_H = 0.0004 (rejet de H_0)
Matrice 2 : P = 0.0385, P_H = 0.0385 (rejet de H_0)	Matrice 5 : P = 0.0001, P_H = 0.0004 (rejet de H_0)
Matrice 3 : P = 0.0161, P_H = 0.0322 (rejet de H_0)	Matrice 6 : P = 0.0001, P_H = 0.0004 (rejet de H_0)
Matrice 7 : P = 0.0084, P_H = 0.0252 (rejet de H_0)	Matrice 8 : P = 0.0001, P_H = 0.0004 (rejet de H_0)
Statistiques de Mantel calculées sur les rangs	Statistiques de Mantel calculées sur les rangs
Matrice 1 1.0000 0.3976 0.6193 0.7793	Matrice 4 1.0000 0.6303 0.7670 0.7061
Matrice 2 0.3976 1.0000 0.5002 0.4672	Matrice 5 0.6303 1.0000 0.6763 0.5636
Matrice 3 0.6193 0.5002 1.0000 0.4167	Matrice 6 0.7670 0.6763 1.0000 0.8984
Matrice 7 0.7793 0.4672 0.4167 1.0000	Matrice 8 0.7061 0.5636 0.8984 1.0000

Les test CEMD ont été répétés pour les deux groupes séparément (Tableau 1). L'hypothèse globale d'indépendance des 4 membres de chaque groupe fut rejetée, de même que celle d'indépendance des matrices individuelles lors des tests *a posteriori*. Les statistiques de Mantel sont élevées ; elles montrent le degré de corrélation de chaque paire de matrices de distance.

L'analyse CEMD a montré l'existence de deux groupes distincts de gènes parmi les 8 matrices de distance et a confirmé la congruence interne de chacun de ces groupes. Dans ces conditions, il sera préférable de réaliser une reconstruction phylogénétique pour les gènes nucléaires et une autre pour les gènes mitochondriaux. Lapointe et Cucumel [LAP 02] avaient cherché comment diviser ces mêmes arbres phylogénétiques en groupes avant des analyses de consensus. La méthode CEMD nous a permis de déterminer statistiquement quels sont les arbres dont il est intéressant de rechercher le consensus.

4 Discussion

Des simulations numériques rapportées dans [LEG 04] ont montré que le test global de même que les tests *a posteriori* ont une erreur de type I correcte et une bonne puissance. La puissance des tests augmente en fonction du nombre d'objets (n) et du nombre de matrices congruentes dans l'étude. Pour un nombre donné de matrices congruentes, la puissance diminue lorsqu'on augmente le nombre de matrices incongruentes dans l'analyse. De nouvelles simulations sont en cours pour établir la puissance de ce test pour des données phylogénétiques et la comparer à celle d'autres tests qui ont été proposés dans ce domaine [FAR 95].

Un programme d'ordinateur permettant de réaliser le test CEMD (le sigle en anglais est CADM) est disponible sur le site WWWeb <http://www.bio.umontreal.ca/legendre/>.

5 Bibliographie

[FAR 95] FARRIS J.S., KÄLLERSJÖ M., KLUGE A.G., BULT C., "Testing significance of incongruence", *Cladistics*, vol. 10, 1995, p. 315-319.

[LAP 02] LAPOINTE F.J., CUCUMEL G., "Multiple consensus trees", *in: Classification, Clustering and Data Analysis - Recent Advances and Applications*, Jajuga K., A. Sokolowski A., Bock H.-H., editors, Springer-Verlag, Berlin, 2002, p. 359-364.

[LEG 00] LEGENDRE P., "Comparison of permutation methods for the partial correlation and partial Mantel tests", *Journal of Statistical Computation and Simulation*, vol. 67, 2000, 37-73.

[LEG 04] LEGENDRE P., LAPOINTE F.-J., "Assessing congruence among distance matrices: single malt Scotch whiskies revisited", *Australian and New Zealand Journal of Statistics*, vol. 46, 2004, p. 615-629.

[MAN 67] MANTEL N., "The detection of disease clustering and a generalized regression approach", *Cancer Research*, vol. 27, 1967, p. 209-220.

[SPR 99] SPRINGER M.S., AMRINE H.M., BURK A., STANHOPE M.J., "Additional support for Afrotheria and Paenungalata, the performance of mitochondrial versus nuclear genes, and the impact of data partitions with heterogeneous base composition", *Systematic Biology*, vol. 48, 1999, p. 65-75.

[WEN 98] WENDEL J.F., DOYLE J.J., "Phylogenetic incongruence : window into genome history and molecular evolution", *in* : *Molecular systematics of plants, II. DNA sequencing*, Kluwer Academic Publishers, Boston, 1998, p. 265-296.