Introduction à la Phylogénie moléculaire

M. Mariadassou & A. Bar-Hen

Université Paris Descartes, Paris 5, MAP5
45 rue des Saints Pères, 75270 Paris cedex 06

Résumé Cet article est une introduction au domaine de la phylogénie moléculaire et en particulier à la robustesse des arbres phylogénétiques. Nous commençons par une brève présentation historique du domaine avant de passer en revue les méthodes de reconstruction les plus populaires. Nous nous intéressons tout particulièrement à la méthode du maximum de vraisemblance. Cette méthode nécessite de construire un modèle probabiliste d'évolution des macromolécules biologiques mais fournit en contrepartie un cadre statistique propice à quantifier la variabilité de l'arbre estimé. Nous présentons tout d'abord les modèles d'évolution couramment utilisé, puis le calcul de la vraisemblance avant de montrer que la nature discrète de l'arbre rend caducs les outils traditionnels d'étude de la variabilité.

1 Le contexte de la phylogénie moléculaire

1.1 Origines du domaine

Les travaux précurseurs de Charles Darwin [4], sur lesquels a été bâtie la biologie évolutive moderne, ont radicalement changé notre compréhension de l'évolution. Darwin introduit dans son livre *De l'Origine des Espèces* la théorie de l'évolution, selon laquelle les espèces évoluent au fil des générations grâce au processus de sélection naturelle et que la diversité du vivant est obtenue grâce à l'accumulation graduelle de différences dans les sous-populations d'une espèce.

L'évolution peut être considérée comme un processus de branchement dans lequel des sous-populations d'une espèce se transforment par accumulation de différences avant de se détacher de leur espèce-mère pour former une nouvelle espèce ou s'éteindre. L'image d'arbre évolutif illustre bien le concept d'évolution et la formation de nouvelles espèces à partir d'espèces déjà existantes. Les liens de parenté qui unissent un groupe d'espèces sont communément représentés sous la forme d'arbres évolutifs, appelés "arbres phylogénétiques" ou encore "phylogénies".

Toutes les méthodes de reconstruction d'arbres phylogénétiques sont basées sur la même idée intuitive : étant donné que l'évolution intervient par accumulation de différences, deux espèces qui ont divergé récemment sont plus "semblables" que deux espèces dont la divergence est plus ancienne. La similitude entre espèces était mesurée par des critères de types morphologiques (à l'instar de la forme des os, du nombre de pattes ou du nombre de dents) jusque dans les années 50. La découverte de la structure de l'ADN par Watson et Crick en 1953 [15] et surtout les capacités de séquençage et d'analyse des molécules macrobiologiques qui ont rapidement suivis ont considérablement changé la donne en remplaçant avantageusement l'objet d'étude. Au lieu d'établir des liens de parenté à partir de critères morphologiques, pour certains fortement soumis à l'appréciation de l'expérimentateur et dont le nombre est généralement faible, les phylogénéticiens peuvent désormais s'appuyer sur des données moléculaires : des séquences génétiques (d'ADN) ou protéiques

(de protéines). Cette révolution présente trois avantages majeurs. Tout d'abord, l'évolution agit beaucoup plus finement au niveau moléculaire qu'au niveau des caractères morphologiques : certaines mutations de la séquence d'ADN sont invisibles au niveau morphologique. Ensuite, les séquences moléculaires sont moins soumises à la subjectivité de l'expérimentateur que les critères morphologiques. Enfin, les séquences moléculaires fournissent des jeux de données bien plus important que les critères morphologiques : au lieu de comparer les espèces sur quelques dizaines de critères morphologiques, on les compare sur des séquences longues de plusieurs milliers de paires de bases, voire de plusieurs millions pour les espèces dont l'intégralité du génome est connue.

Reconstruire l'histoire évolutive des espèces constitue évidemment un but en soi pour les biologistes évolutifs. Le symbole le plus emblématique en est le projet "Arbre de la Vie" (Tree of Life Project, `www.tolweb.com`), qui cherche à reconstruire l'arbre phylogénétique de toutes les espèces vivantes. Mais les arbres phylogénétiques ont aussi un intérêt majeur dans d'autres domaines de la biologie. Ils sont par exemple inestimables en génomique comparative, où ils permettent par exemple de prédire la fonction d'un gène inconnu à partir de la fonction d'un gène similaire dans des espèces proches [7, 8] ou encore de prédire si deux protéines interagissent à partir de leurs arbres phylogénétiques respectifs [13]. Mais le domaine d'application de la phylogénie ne se réduit pas à la biologie moléculaire : les phylogénies apparaissent aussi naturellement en biologie de la conservation quand, en particulier dans les études de mesure de la biodiversité [2].

1.2 Méthodes de reconstruction d'arbres phylogénétiques

Toutes les applications décrites dans la section 1.1 s'appuient sur des arbres phylogénétiques bien reconstruits. Mais reconstruire de tels arbres est une tache ardue : il s'agit de reconstituer le chemin parcouru par l'évolution à partir des empreintes qu'elle laisse sur les génomes, en sachant que ces empreintes peuvent être ténues et s'atténuent de toutes façons au fil du temps. Les systématiciens n'en reconstruisent pas moins des arbres évolutifs depuis Darwin, avec une précision étonnante.

Il existe essentiellement 5 grandes familles de méthodes pour reconstruire une phylogénie : les méthodes de parcimonie [5], les méthodes de moindres-carrés [3], les méthodes de maximum de vraisemblance [6, 9], les méthodes de distance [10] et les méthodes bayésiennes [12, 11, 14]. La contribution majeure des travaux de Cavalli-Sforza et Edwards, tous deux disciples de Fisher, est sans doute l'identification précoce de la reconstruction d'arbres phylogénétiques comme un problème d'inférence statistique.

Toutes les méthodes évoquées ci-dessus peuvent être décomposées en trois parties :

1. Un *critère d'optimalité*, qui mesure l'adéquation des données à un arbre phylogénétique donné (par exemple : la parcimonie, la vraisemblance, les sommes de carrés, etc) ;
2. Une *stratégie de recherche* pour identifier l'arbre optimal (par exemple : recherche exhaustive, descente de gradient, etc)
3. Des hypothèses sur le *mécanisme d'évolution* des données.

Il n'existe pas de méthode supérieure à toutes les autres, chacune à ses forces et ses faiblesses et le débat sur les mérites comparés de deux méthodes n'est pas clos. Pour certains groupes d'espèces, le choix de la méthode importe peu : toutes les méthodes

reconstruisent le même arbre phylogénétique. Il s'agit évidemment du cas optimiste, rarement rencontré en pratique. La méthode du maximum de vraisemblance est nettement plus lente que ses concurrentes mais fournit un cadre de travail naturel tant pour tester des hypothèses que pour quantifier la variabilité de l'arbre estimé.

1.3 Validation de l'arbre

Comme dans la majorité des procédures d'inférence statistique, l'arbre estimé dépend des données : la même procédure d'estimation appliquée à différents jeux de données va donner différents arbres. Il est essentiel de quantifier cette variabilité, en prouvant par exemple que l'arbre estimé n'est pas très différent du vrai arbre. La façon standard de faire en est de prouver un théorème limite sur l'arbre estimé, généralement un théorème de normalité asymptotique sur la distance entre l'arbre estimé et le vrai arbre.

Mais un arbre phylogénétique est un paramètre inhabituel : il est composé d'une topologie (une forme d'arbre) discrète et de longueurs de branches continues, qui dépendent de la topologie de l'arbre. L'espaces des arbres phylogénétiques a de plus une structure complexe [1] qui rend inopérants les outils de convergence utilisés pour établir des théorèmes limites.

Faute de théorèmes limite, la variabilité de l'estimateur est généralement quantifiée à l'aide de technique de rééchantillonnages, telles que le bootstrap ou le jackknife qui miment des échantillons indépendants.

Enfin, il est nécessaire de valider la robustesse de l'arbre estimé. Les erreurs d'alignement et de séquençage peuvent en effet engendrer de petites modifications du jeu de données. Quelle est l'influence de ces petites modifications sur l'arbre estimé ? Si leur influence est faible, l'arbre estimé est robuste aux erreurs de séquençage et d'alignement : il est légitime de s'en servir dans des analyses ultérieurs. Dans le cas contraire, l'arbre est peu robuste : les analyses basées sur cette arbre sont peu fiables. Là encore, les méthodes de bootstrap et de jackknife permette de quantifier la robustesse de l'arbre. Dans le cas d'arbres non robustes, il est intéressant d'identifier les données erronées pour les corriger ou les supprimer du jeu de données. Les erreurs de séquençage et d'alignements ont tendance à créer des données exceptionnelles, très différentes du reste du jeu de données et inattendu. Le cadre du maximum de vraisemblance permet non seulement de quantifier la variabilité de l'arbre estimé mais aussi le caractère exceptionnel ou non d'une donnée. Il est donc particulièrement propice à la détection de données erronées.

Références

[1] Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Math.*, 27 :733–767, 2001.

[2] Magnus Bordewich, Allen G Rodrigo, and Charles Semple. Selecting taxa to save or sequence : desirable criteria and a greedy solution. *Syst Biol*, 57(6) :825–834, Dec 2008.

[3] L. L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetics analysis : Models and estimation procedures. *American Journal of man Geneics*, 19 :233–257, 1967.

[4] Charles Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (1st ed.), London : John Murray,.* John Murray, London, 1 edition, 1859.

[5] A. W. F. Edwards and L. L. Cavalli-Sforza. The reconstruction of evolution. *Annals of Human Genetics*, 27 :105–106, 1963.

[6] A. W. F. Edwards and L. L. Cavalli-Sforza. *Phenetic and Phylogenetic Classification*, chapter Reconstruction of evolutionary trees, pages 67–76. Systematics Association Publ. No. 6, London, 1964.

[7] J. A. Eisen. Phylogenomics : improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*, 8(3) :163–167, Mar 1998.

[8] Jonathan A Eisen and Martin Wu. Phylogenetic analysis and gene functional predictions : phylogenomics in action. *Theor Popul Biol*, 61(4) :481–487, Jun 2002.

[9] J. Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology*, 22 :240–249, 1973.

[10] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155(760) :279–284, Jan 1967.

[11] S. Li. *Phylogenetic tree construction using Markov chain Monte Carlo.* PhD thesis, Ohio State University, Ohio, 1996.

[12] B. Mau. *Bayesian phylogenetic inference via Markov chain Monte Carlo.* PhD thesis, University of Wisconsin, 1996.

[13] F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 14(9) :609–614, Sep 2001.

[14] B. Rannala and Z. Yang. Probability distribution of molecular evolutionary trees : a new method of phylogenetic inference. *J Mol Evol*, 43(3) :304–311, Sep 1996.

[15] James D. Watson and Francis Crick. Molecular structure of nucleic acids : A structure for deoxyribose nucleic acid. *Nature*, 171 :737–738, 1953.