Voisinage d'arbre évolutif appliqué au problème Maximum Parcimonie

Adrien Goëffon, Jean-Michel Richer, Jin-Kao Hao

Voisinage d'Arbre Evolutif Appliqué au Problème Maximum Parcimonie

Adrien Goëffon, Jean-Michel Richer et Jin-Kao Hao

LERIA - Université d'Angers
2 bd Lavoisier - 49 045 Angers Cedex 01
Firstname.Name@univ-angers.fr

Résumé

Le problème Maximum Parcimonie vise à reconstruire un arbre phylogénétique à partir de séquences ADN de manières à ce que le nombre de mutations génétiques survenues au cours de l'évolution soit minimal. Pour résoudre ce problème NP-complet, de nombreuses méthodes heuristiques ont été développées, pour la plupart basées sur la recherche locale. Ici, nous nous intéressons à l'influence de la relation de voisinage utilisée et introduisons le concept de voisinage évolutif. Nous montrons empiriquement que ce voisinage évolutif s'avère plus puissant et robuste que les voisinages classiques.

1 Introduction

La phylogénie peut être définie comme la reconstruction de l'évolution d'un ensemble d'espèces (ou taxons) associés à une séquence d'acides nucléiques (ADN) ou d'acides aminés (AA). Ces relations sont représentées par un arbre dit phylogénétique.

Il existe plusieurs manière de reconstruire des arbres phylogénétiques [2] : méthodes de distance, méthodes probabilistes et méthodes cladistes.

Le problème Maximum Parcimonie (MP), vise à retrouver la phylogénie qui minimise le nombre d'évènements évolutifs (score) et permet d'attribuer à chaque ancêtre hypothétique (noeud interne de l'arbre), les états possibles pris pour chaque caractère. Ce problème cladistique utilise une matrice de caractères donnée sans recourir à un modèle de l'évolution.

MP est NP-complet [4]. L'approche utilisée pour l'approximation du problème consiste à utiliser des algorithmes heuristiques dans le but de trouver le plus rapidement possible un arbre d'un score très proche de celui d'une solution optimale. Considérant les très larges espaces de recherche, il se vérifie empiriquement que des heuristiques de recherche locale stochastique sont particulièrement adaptées au problème MP, à la condition d'utiliser un voisinage approprié.

Il existe majoritairement dans la littérature trois voisinages d'arbres : NNI [8], SPR [7] et TBR [7]. Chaque recherche locale associée gagne en efficacité ou en rapidité suivant le voisinage utilisé. Notre démarche consiste à combiner les propriétés de ces voisinages intéressants afin d'obtenir une recherche locale à la fois rapide, efficace et robuste. Nous introduisons ici le concept de voisinage évolutif en tant que nouvelle approche pour la résolution du problème MP. Tous les tests effectués montrent le réel gain apporté par cette technique d'un point de vue efficacité et temps de calcul, et surtout sa capacité à converger très vite vers une solution de confiance, en évitant les pièges des optima locaux.

2 Le problème Maximum Parcimonie

Le problème MP consiste à partir d'un ensemble S de séquences, à retrouver la phylogénie optimale au sens du critère de parcimonie, c'est-à-dire un arbre dont les feuilles sont associées aux séquences de S et qui minimise le nombre de mutations.

Définition 1 *La* distance de Hamming $H(x, y)$ *entre deux séquences* $x = (x_1, x_2, \ldots, x_k)$ *et* $y = (y_1, y_2, \ldots, y_k)$ *est égale à* $|\{i : x_i \neq y_i\}|$.

Définition 2 *Le* score de parcimonie $f(T)$ *d'un arbre* $T = (V, E)$ *dont chaque noeud* v *est étiqueté par une séquence* s^v *de longueur* k *sur un alphabet* Σ *est la somme des distances de Hamming des séquences étiquetant chaque couple de noeuds séparés par une arête dans* T, *i.e.* $f(T) = \sum_{(x,y)\in E} H(x, y)$.

Etant donné un arbre T dont les feuilles sont bijectivement étiquetées par les séquences de S, Fitch a formalisé un algorithme polynomial [3] qui calcule des séquences hypothétiques (assignées aux noeuds internes de l'arbre) et le score de parcimonie de telle sorte que celui-ci soit minimal.

Le but du problème MP est de trouver un arbre dont le score de parcimonie est le plus faible parmi tous les arbres phylogénétiques possibles pour un ensemble S de séquences. MP peut alors être formulé comme un problème combinatoire de minimisation $(\mathcal{T}, f)$ tel que :

1. l'espace de recherche $\mathcal{T}$ est défini par l'ensemble des $\prod_{i=3}^{|S|}(2i-3)$ configurations possibles
2. la fonction de coût $f : \mathcal{T} \rightarrow I\!N$ qui calcule le score de parcimonie de T.

3 Recherche locale et voisinages

La méthode de descente consiste à générer une première phylogénie, puis à rechercher une phylogénie voisine (au sens d'une relation de voisinage) dont le score est inférieur, et ainsi de suite jusqu'à ce que la phylogénie courante n'ait aucun voisin dont le score soit strictement inférieur. La solution finale est alors un optimum local, qui n'est pas nécessairement un optimum global.

Cette approche de descente, qui est la méthode de recherche locale la plus simple, dépend essentiellement de la relation de voisinage à laquelle elle est associée. Même s'il existe de nombreuses techniques pour tenter d'améliorer la qualité des solutions fournies par les algorithmes de descente, ces derniers sont à la base de toutes les meilleures méthodes de résolution actuelles.

Une relation de voisinage structure l'espace de recherche sur lequel une méthode de recherche locale (par exemple descente) est appliquée. Les trois relations de voisinage d'arbres que l'on retrouve systématiquement dans la littérature sont NNI, SPR et TBR. NNI (*Nearest Neighbor Interchange*) consiste à échanger deux branches adjacentes de l'arbre, SPR (*Subtree Pruning Regrafting*) est une stratégie qui coupe une branche et la réinsère à un autre endroit de l'arbre, et TBR (*Tree-Bisection-Reconnection*) est un voisinage plus large qui casse l'arbre en deux sous-arbres qui seront reconnectés à partir d'une de leurs arêtes. On peut remarquer que $NNI \subseteq SPR \subseteq TBR$.

Une relation de voisinage réduite comme NNI possède l'avantage de favoriser la recherche à grande échelle en ne permettant que des modifications très locales sur l'arbre. Calculer la variation de coût engendrée par une transformation NNI est d'autant plus rapide que l'arbre résultant est très proche, et parcourir l'ensemble des voisins d'une configuration est également plus rapide qu'avec un voisinage plus large, car le nombre de voisins à explorer est plus petit. En revanche, une recherche locale sur un tel espace de recherche aura une faible capacité à améliorer sensiblement le coût d'une solution sur quelques pas. De plus, étant donné le faible nombre de voisins, les optima locaux seront plus fréquents sur des solutions pas nécessairement proches de l'optimum en terme de coût.

A l'opposé, une relation de voisinage plus large comme SPR ou TBR est plus coûteuse d'un point de vue calculatoire. Explorer tout le voisinage d'une configuration prend plus de temps (même si Goloboff [5] propose une méthode qui réduit la complexité du recalcul du score), et les arbres voisins peuvent subir d'importantes modifications topologiques.

4 Voisinage évolutif

Afin de combiner les propriétés intéressantes des voisinages larges et restreints, nous proposons d'effectuer une recherche locale sur un espace de recherche qui s'élargit ou se rétracte en fonction de l'avancée de la recherche, ou de la fréquence d'apparition de voisins pertinents. A titre d'exemple simple, nous définissons un schéma prédéfini de voisinage qui se rétracte que nous appliquons au problème MP.

Partir du voisinage le plus large peut s'avérer pertinent, en construisant les bases de la topologie de la future solution. Evaluer plus de voisins (avec des modifications plus sensibles) en début de recherche va permettre d'améliorer grandement le coût des solutions dès les premiers pas de la recherche locale, grâce à une recherche plus intensive. En fin de recherche, on peut imaginer n'intervenir que très localement sur la topologie de l'arbre. On peut obtenir ce schéma en réduisant petit à petit l'étendue du voisinage exploré au fil de la recherche.

Prenons deux voisinages $\mathcal{N}^1$ et $\mathcal{N}^2$ tels que $\mathcal{N}^2 \subseteq \mathcal{N}^1$, de sorte que nous puissions définir plus simplement un voisinage paramétrique $\mathcal{N}_d$ qui généralise $\mathcal{N}^1$ et $\mathcal{N}^2$. Une piste intéressante est de considérer $\mathcal{N}^1 = \mathcal{N}^{SPR}$ et $\mathcal{N}^2 = \mathcal{N}^{NNI}$. Avec SPR, on dégrafe une branche de l'arbre et on la reconnecte ailleurs, sans contrainte particulière si ce n'est d'obtenir un arbre valide et distinct. On peut voir NNI comme un SPR particulier, où une branche doit être insérée sur une arête voisine d'où elle provient dans l'arbre courant.

Par extension, nous imaginons alors un voisinage de type SPR où la distance entre l'arête supprimée et l'arête insérée soit contrainte. Si cette distance est maximale, alors il s'agit du voisinage SPR, sans contrainte. Si celle-ci elle minimale, alors nous nous retrouvons dans le cas NNI. Nous introduisons alors un paramètre d, tel que $\mathcal{N}_d^{SPR}(T)$ représente l'en-

semble des arbres obtenus par transformation SPR et dont la longueur du chemin entre l'arête dégrafée et l'arête d'insertion n'excède pas d. Réduire d durant la recherche permet de débuter avec un voisinage quadratique (SPR) et de terminer avec le voisinage NNI, linéaire par rapport au nombre de noeuds.

5 Premières expérimentations

5.1 Conditions d'expérimentation

Durant nos tests, nous avons utilisé en premier lieu 300 instances aléatoires, toutes générées suivant des combinaisons de paramètres différentes. Nous avons observé que la tendance des résultats variait très peu d'une instance à l'autre. Parmi elles et pour des raisons de lisibilité, nous avons choisi d'en extraire deux, comportant respectivement 300 séquences ADN courtes (100 caractères) et 500 séquences plus longues (1000 caractères). De plus, nous considérons l'instance *zilla* [1] majoritairement utilisée dans la littérature et réputée très difficile, composée de 500 séquences de 759 caractères.

Nous utilisons un algorithme de descente stricte sur lequel nous testons les trois voisinages : $\mathcal{N}^{SPR}$, $\mathcal{N}^{NNI}$ et $\mathcal{N}_d^{SPR}$ pour le Voisinage Evolutif tel qu'il est décrit dans la section 4 (le paramètre d est initialisé au plus petit majorant des distances entre noeuds, et est réduit de manière linéaire jusqu'à 1). Puisque $\mathcal{N}^{NNI} \subseteq \mathcal{N}_d^{SPR} \subseteq \mathcal{N}^{SPR}$, la descente classique (qui retourne obligatoirement un optimum local) utilisant $\mathcal{N}^{SPR}$ retournera majoritairement des solutions de coût meilleur ou égal. Mais déterminer un optimum local nécessite en particulier d'avoir calculé tous ses voisins. Pour mesurer efficacement l'influence du voisinage sur la qualité de la solution retournée pour un effort calculatoire équivalent (et donc des temps de calcul plus raisonnables), nous fixons un nombre maximal M d'itérations de recherche locale.

Nous avons lancé 20 fois chaque descente, à partir d'un arbre aléatoire (méthode $\mathcal{A}$ dans les tableaux de résultats), ou à partir d'un arbre construit selon une méthode de distances stochastique ($\mathcal{D}$) qui permet de débuter la recherche depuis un arbre de meilleur score (nécessaire pour l'instance difficile *zilla*).

5.2 Résultats

Les tableaux de résultats contiennent, pour chaque instance et chaque méthode, le score minimum f_b, le score moyen f_m et son écart-type σ sur l'ensemble des essais, ainsi que le *temps* d'exécution moyen en secondes (tableau 1) ou en minutes. Les tests ont été réalisés sur une station SunFire à 750 MHz avec 8 Go de Ram.

300-100	f_b	f_m	σ	*temps*
$\mathcal{A}$+$\mathcal{N}^{SPR}$	1 579	1 647,9	32,3	115
$\mathcal{A}$+$\mathcal{N}^{NNI}$	1 746	1 921,9	92,2	53
$\mathcal{A}$+$\mathcal{N}_d^{SPR}$	1 304	1 310,8	7,0	51
$\mathcal{D}$+$\mathcal{N}^{SPR}$	1 336	1 342,4	4,1	77
$\mathcal{D}$+$\mathcal{N}^{NNI}$	1 303	1 305,6	3,7	54
$\mathcal{D}$+$\mathcal{N}_d^{SPR}$	**1 302**	**1 303,4**	1,5	53

TAB. 1 – Comparaison entre les voisinages (20 exécutions limitées à 50 000 itérations)

Le tableau 1 montre que le voisinage $\mathcal{N}^{SPR}$ retourne des solutions de score très éloigné de ceux des solutions trouvées par $\mathcal{N}_d^{SPR}$. La qualité de la solution retournée avec $\mathcal{N}^{NNI}$ est fortement dépendante de l'arbre initial et peut converger très rapidement vers un optimum local, ce qui montre l'instabilité de la méthode en fonction des facteurs stochastiques. $\mathcal{N}^{NNI}$ n'est vraiment efficace (en terme de performance, robustesse et temps de calcul) que lorsque les séquences sont longues et à condition de débuter la recherche avec une *bonne* solution ($\mathcal{D}$). Notre voisinage évolutif $\mathcal{N}_d^{SPR}$ obtient de bons résultats depuis toute solution initiale, malgré le nombre réduit d'itérations comparé à la taille du problème.

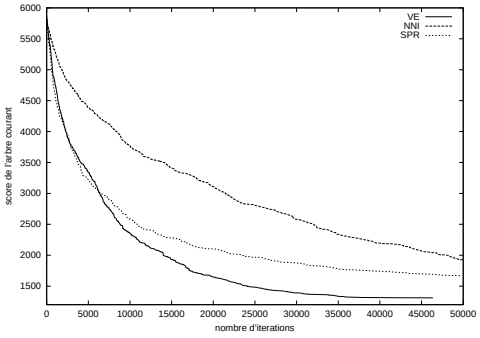

FIG. 1 – Evolution du score de la solution courante à partir d'un arbre aléatoire sur 50 000 itérations (instance *300-100*)

La figure 1 montre l'évolution du score de l'arbre courant en fonction de l'avancée de la recherche, pour l'instance 300-100 et en partant d'un arbre aléatoire (la recherche reportée est celle retournant l'arbre de score médian). La recherche locale à voisinage évolutif (VE) domine clairement SPR et NNI.

Pour une instance très large (tableau 2), il se confirme que $\mathcal{N}^{SPR}$ donne de mauvais résultats. Les scores retournés par $\mathcal{N}^{NNI}$ sont peu stables, avec des écarts-types importants. Si un allongement du temps de recherche provoque une amélioration des performances de $\mathcal{N}^{SPR}$, ce n'est rapidement plus le cas pour $\mathcal{N}^{NNI}$, qui retourne très tôt un optimum local. On remarque particulièrement l'efficacité du voisinage évolutif $\mathcal{N}_d^{SPR}$ dans le cas de larges instances. En 100 000 itérations, il parvient à retourner systématiquement,

500-1000	M	f_b	f_m	σ	$temps$
$\mathcal{N}^{SPR}$	5.10^4	36 505	37 595,6	595,1	39'
$\mathcal{N}^{NNI}$		28 478	30 072,0	1 374,8	20'
$\mathcal{N}_d^{SPR}$		**24 337**	**24 491,7**	184,7	16'
$\mathcal{N}^{SPR}$	10^5	31 615	32 029,7	319,4	75'
$\mathcal{N}^{NNI}$		**24 319**	24 460,3	188,9	29'
$\mathcal{N}_d^{SPR}$		**24 319**	**24 319**	0	33'
$\mathcal{N}^{SPR}$	$1,5.10^5$	28 961	29 490,7	407,6	116'
$\mathcal{N}^{NNI}$		**24 319**	24 460,3	188,9	29'
$\mathcal{N}_d^{SPR}$		**24 319**	**24 319**	0	44'

TAB. 2 – Comparaison entre les voisinages sur une instance large, à partir d'une solution initiale aléatoire

une solution de score égal au meilleur score trouvé pour cette instance. Sur les 20 tests, $\mathcal{N}^{NNI}$ n'a trouvé qu'une seule fois un tel score.

zilla	M	f_b	f_m	σ	$temps$
$\mathcal{N}^{SPR}$	10^5	17 089	17 116,8	24,8	145'
$\mathcal{N}^{NNI}$		16 556	16 778,9	119,2	19'
$\mathcal{N}_d^{SPR}$		**16 306**	**16 356,5**	47,8	27'
$\mathcal{N}^{SPR}$	2.10^5	16 785	16 816,8	36,6	278'
$\mathcal{N}^{NNI}$		16 556	16 778,9	119,2	19'
$\mathcal{N}_d^{SPR}$		**16 282**	**16 297,9**	9,8	57'
$\mathcal{N}^{SPR}$	3.10^5	16 590	16 645,3	57,5	395'
$\mathcal{N}^{NNI}$		16 556	16 778,9	119,2	19'
$\mathcal{N}_d^{SPR}$		**16 277**	**16 296,3**	18,0	77'

TAB. 3 – Comparaison entre les voisinages sur l'instance *zilla*, à partir d'une solution construite

Avec l'instance réelle *zilla* bien connue, trouver le score minimum calculé à ce jour (16 218) en un faible nombre d'itérations et avec un unique arbre de départ (une *réplication*) est peu probable. Pour $\mathcal{N}^{SPR}$ et $\mathcal{N}^{NNI}$, nous avons lancé 20 fois chaque méthode mais uniquement sur 3.10^5 itérations. Les valeurs pour 10^5 et 2.10^5 itérations sont données par les résultats intermédiaires des recherches locales. $\mathcal{N}^{NNI}$ retourne systématiquement un optimum local en 50 000 itérations en moyenne, c'est pour cette raison que les résultats sont identiques pour un M supérieur. Sur cette instance difficile, on remarque immédiatement la puissance du voisinage évolutif. En 200 000 ou 300 000 itérations, l'écart entre les arbres retournés et le meilleur arbre connu à ce jour varie entre $0,36\%$ et $0,69\%$, ($0,49\%$ en moyenne), tandis qu'il varie entre $2,08\%$ et $4,33\%$ ($3,31\%$ en moyenne, soit près de 7 fois plus) pour les autres voisinages connus.

La figure 2 montre le comportement des trois recherches ayant retourné le *score médian* pour les trois méthodes appliquées à l'instance *zilla* sur 300 000 itérations. On remarque clairement la supériorité du voisinage évolutif $\mathcal{N}_d^{SPR}$, noté VE sur les figures, par rapport à SPR et NNI.

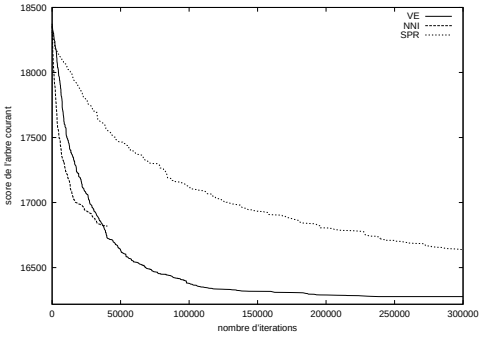

FIG. 2 – Evolution du score de la solution courante sur 3.10^5 itérations (instance *zilla*) partant d'un arbre initial construit selon une méthode de distances

6 Conclusion

Pour résoudre le problème MP, les recherches locales basées sur SPR et NNI sont parmi les méthodes les plus populaires. Bien qu'elles soient très performantes et rapides pour des petites instances (comportant moins de 100 espèces), elles apparaissent peu fiables lorsqu'on les applique à des instances plus grandes.

L'objectif ici était de comprendre l'influence du voisinage utilisé en fonction des instances et de l'avancée de la recherche locale, et de proposer une alternative qui combine les propriétés intéressantes des voisinages SPR et NNI. Nous avons alors introduit la notion de *voisinage évolutif*, et effectué une série d'expérimentations montrant un gain d'efficacité sensible par rapport à SPR et NNI, notamment sur des instances difficiles. De plus, sa robustesse entraîne un gain de temps, car elle permet de minimiser le nombre de réplications. En effet, l'arbre initial influe très peu sur la qualité des solutions retournées par le voisinage évolutif.

Remerciements : Ce travail est partiellement supporté par Ouest Genopole®.

Références

[1] M. W. Chase *et al.* Phylogenetics of seed plants : an analysis of nucleotide-sequences from the plastid gene rbcL. *Annals of the Missouri Botanical Garden*, 80 :528-580, 1993.

[2] J. Felsenstein. Inferring Phylogenies. *Sinauer*, 2004.

[3] W. Fitch. Towards defining course of evolution : minimum change for a specified tree topology. *Systematic Zoology* 20 :406-416, 1971.

[4] L. R. Foulds et R. L. Graham. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics* 3 :43-49, 1982.

[5] P. A. Goloboff. Character optimisation and calculation of tree lengths. *Cladistics* 9 : 433-436, 1993.

[6] R. R. Sokal et C. D. Michener. A statistical method for evaluating systemaic relationships *Univ. of Kansas Science Bulletin* 38 :1409-1438, 1958.

[7] D. L. Swofford et G. J. Olsen. in D.M. Hillis and C. Moritz (Ed.) Phylogeny Reconstruction. *Molecular Systematics*, chapter 11 :411-501, 1990.

[8] M. S. Waterman et T. F. Smith. On the similarity of dendograms. *Journal of Theoretical Biology* 73 :789-800, 1978.