G. Hunault

Angers, février 2003

Regroupements phylogénétiques

NP-complets

1. Phylogénie et arbres phylogénétiques

La phylogénie est l'étude du développement dans le temps des organismes. Des termes relativement synomymes sont "histoire évolutive des organismes" et de façon plus spécifique, "généalogie des gènes dans un même organisme dérivés d'un même gène ancestral". Un phylum (pluriel phyla) est un groupe d'organismes partageant des mêmes caractères liés à un ancêtre commun, comme par exemple le phylum des arthropodes qui regroupe *tous ceux qui possèdent des pattes articulées.*

Un taxon, du latin taxum (pluriel taxa) est un ensemble d'organismes possédant des traits communs regroupés sous une même description à l'intérieur d'un même phylum ; c'est aussi le nom que l'on donne à un groupe d'organismes identifiés comme semblables. Un taxon peut ainsi être un genre, une espèce ou plus souvent un couple/genre espèce comme le taxon *Canis lupus* qui désigne l'ensemble des loups. La taxonomie est la science qui consiste à étudier les taxons, qu'il s'agisse d'en créer de nouveaux, de valider, confirmer ou infirmer d'anciens taxons. On emploie souvent le terme d'OTU (*operational taxonomic unit*) comme synonyme de "taxon de base" ou "taxon non décomposable".

Un taxonomiste évolutionniste décrit les regroupements d'organismes ou de taxons à l'aide de représentations graphiques nommées arbres phylogénétiques. Ces "arbres" correspondent à qu'on désigne en informatique par des structures arborescentes enracinées ou non, additives ou non, valuées ou non, selon qu'on s'intéresse à la topologie de l'arbre ou à la métrique. La reconstruction d'arbres phylogénétiques consiste, à partir de caractéristiques bio-physiologiques, biochimiques ou de séquences génétiques (ADN, protéines) à essayer de construire un arbre dont les séquences sont les feuilles et qui représente au mieux "l'évolution" des séquences, les noeuds ou regroupements induits par l'arbre (nommés aussi "classes") définissant alors de nouveaux taxa interprétés comme des ancêtres communs.

On sépare les méthodes de reconstruction de ces arbres à partir des données en deux catégories : les méthodes directes ou *discrètes*, lentes, qui utilisent les données et les méthode indirectes ou *comparatives*, rapides, qui travaillent sur les distances entre les données. Pour chacune de ces catégories, le problème sous-jacent est NP-complet.

On notera qu'un arbre phylogénétique construit sur des séquences génétiques fournit un guide (*guide tree*) pour le choix des séquences dans un alignement multiple progressif de ces séquences. Il existe aussi des applications linguistiques lorsque les données sont des textes, des mots...

2. Techniques directes

Pour ce genre de méthodes, les données sont des variables (ou "traits" ou "caractères") statistiques qualitatives c'est à dire ne pouvant prendre qu'un nombre fini (discret) de valeurs. La structure de données de départ est alors une "matrice des états des caractères" ou CSM ("*character state matrix*") dans laquelle les lignes correspondent aux taxa et les colonnes aux caractères, comme il est d'usage en statistiques. Lors d'un changement (*"transition"*) d'état, on parle de *convergence* s'il y a changement (parallèle) d'un état vers deux (ou plus) autres états, *réversion* ("*reversal*) s'il y a retour à l'état du parent dit "état ancestral".

Un arbre phylogénétique dont les arcs sont étiquetés par les transitions d'état est dit parfait s'il n'a aucun renversement ni aucune convergence. Le problème de trouver un arbre phylogénétique parfait associé à une matrice CSM est NP-complet lorsqu'on a plus de deux caractères avec plus de deux états possibles par caractère.

Si une méthode (quelconque) de reconstruction fournit un arbre qui n'est pas parfait, on peut essayer de le rendre parfait. Une première technique consiste à minimiser les réversions ou les convergences : c'est le critère dit de parsimonie. On utilise principalement deux algorithmes à savoir dollo qui minimise les renversements avec interdiction de convergence et camin-sokal qui minimise les convergences avec interdiction de renversement.

Une deuxième technique consiste à esssayer d'exclure un nombre minimal de caractères pour obtenir un arbre parfait : c'est le critère dit de compatibilité. Ces trois problèmes de minimisations sont NP-difficiles.

Ces méthodes sont lentes parce qu'elles passent en revue toutes les topologies possibles pour trouver l'arbre optimal. Dans le cas non enraciné (u_n) comme dans le cas enraciné (r_n), le nombre d'arbres croit très vite puisque

$$u_{n+1} = r_n = \prod_{k=3}^{k=n}(2k-3) = \frac{(2n-3)!}{2^{n-3}(n-2)!} > O(n!)$$

Une autre méthode utilisée est celle du maximum de vraisemblance où on évalue, en termes de probabilités numériques, l'ordre des branchements et la longueur des branches d'un modèle évolutif donné à l'aide de chaines de *Markov*.

3. Techniques indirectes

Ces techniques s'apparentent aux méthodes traditionnelles de classification non supervisée (*clustering*) en Analyse des données, de type CAH (classification hiérarchique ascendante) pour lesquelles la structure de données de départ est une matrice de distances (ou DM "distance matrix). Le principe pour ces techniques hiérarchiques ascendantes est le même : à partir des éléments à regrouper, considérés initialement comme tous isolés

- on choisit deux éléments qu'on considére comme un nouvel élément,
- on recalcule la distance entre les anciens éléments et ce nouvel élément,
- on supprime les deux anciens éléments choisis,
- on recommence jusqu'à avoir regroupé tout le monde.

Deux points sont importants pour ces techniques, hormis le choix de la distance initiale :

- le critère de sélection ou "indice d'agrégation" pour retenir deux éléments,
- la formule de recalcul des distances au nouvel élément.

Le critère de sélection est en général soit un test simple

- choix de la valeur minimale :
  . méthodes de liaison (*linkage*),
  . méthodes en *PG* (*pair group*)...
- choix de la valeur maximale...

soit un calcul plus technique

- choix de la divergence minimale par rapport aux autres groupes : *NJ*,
- critère d'inertie "à la *Ward*" : CAH/ACP, CAH/AFC ...

On peut choisir de recalculer les distances au nouvel élément de nombreuses façons. Les six principales formules de calcul sont les suivantes

- calcul simple, correspondant soit à la distance minimale, soit à la distance maximale nommées respectivement *single linkage* et *complete linkage*,

- calcul par moyenne, pondérée ou non, nommées respectivement *WPGMA*, *NJ* et *UPGMA*,

- calcul par utilisation pondérée ou non des des "centroides" de classes, nommées respectivement *WPGMC* et *UPGMC*.

Si on note $D_{x,y}$ la distance de x à y, on peut résumer toutes ces formules de recalcul de la distance de l'ancien groupe k au nouveau groupe composé de i et de j par l'équation

$$D_{ij,k} = a_i D_{i,k} + b_j D_{j,k} + c D_{i,j} + d|D_{i,k} - D_{j,k}|$$

Voici par exemple les valeurs de ces coefficients pour quelques méthodes

méthode	a_i	b_j	c	d	*interprétation*
single	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	minimum
complete	$\frac{1}{2}$	$\frac{1}{2}$	0	$+\frac{1}{2}$	maximum
upgma	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0	moyenne pondérée
wpgma	$\frac{1}{2}$	$\frac{1}{2}$	0	0	moyenne stricte
nj	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	0	moyennne réduite

Les deux principales "méthodes" considérées comme valides à cause des hypothèses biologiques sous-jacentes pour la reconstruction des arbres phylogénétiques sont

- UPGMA *unweighted pair group method with arithmetic mean*
- NJ *neighbor-joining*

Suivant le choix de la méthode on obtient :

- soit un arbre additif non enraciné (méthode NJ),
- soit un arbre ultramétrique enraciné (méthode UPGMA).

La définition de la racine de l'arbre qui correspond à l'ancêtre commun ???? Si le but est de produire un arbre phylogénétique enraciné, on utilise diverses techniques supplémentaires comme le *mid-point rooting* qui consiste à place la racine à mi-chemin du plus grand chemin dans l'arborescence ...

Si M est une matrice de distances, le problème qui consiste à trouver un arbre T dont la métrique d_T minimise la distance entre M est T est NP-difficile. Toutefois :

- si M est ultramétrique, la méthode UPGMA fournit la solution exacte,
- si M est additive, la méthode NJ fournit la solution exacte.

En pratique, les matrices ne sont ni forcément ultramétriques ni forcément additives et la méthode NJ est considérée comme une bonne heuristique.

4. Problèmes formels

Soit f une fonction définie sur $E \times E$. On pose $q_f(a,b,c,d) = f(a,b) + f(c,d)$. On définit les propriétés suivantes dites ≪ pour f sur E ≫ :

	f est	si et seulement si f vérife
C1	d-nulle	$f(x,x) = 0$
C2	symétrique	$f(x,y) = f(y,x)$
C3	d-triangulaire	$f(x,y) \leqslant f(x,z) + f(z,y)$
C4	séparée	$f(x,y) = 0 \Rightarrow y = x$
C5	sous-maximale	$f(x,y) \leqslant max(f(x,z), f(y,z))$
C6	buneman	$q_f(x,y,z,t) \leqslant max(\ q_f(x,z,y,t), q_f(x,t,y,z)\ )$

La propriété de *Buneman* est également nommée "condition des 4 points".

On définit par combinaison les propriétés suivantes :

nom	conditions	remarques
indice de dissimilarité	C1 C2	
pseudo-métrique (ou écart)	C1 C2 C3	
métrique (ou distance)	C1 C2 C3 C4	
ultramétrique	C1 C2 C3 C5	ce qui implique C4
métrique d'arbres	C1 C2 C3 C4 C6	ou métrique additive

La propriété d'ultramétricité est équivalente à la "condition des 3 points" :

$$f(x,z) = f(y,z) = c \Rightarrow f(x,y) \leqslant c$$

Ces définitions induisent des propriétés de même nom pour une matrice M si l'on définit f par $f(i,j) = M_{i,j}$. On peut donc parler de matrice de distances, de matrice de buneman, de matrice ultramétrique...

Soient H_i des ensembles. On dit que H_i et H_j sont hiérarchiquement compatibles *ssi* $H_i \cap H_j \in \{\ \emptyset\ , H_i\ , H_j\ \}$. La collection* des H_i a une structure de hiérarchie *ssi* Soient H_i des ensembles. On dit que les H_i ont une structure de hiérarchie *ssi* les Hi sont deux à deux hiérarchiquement compatibles. pour tous i et j on a $H_i \cap H_j \in \{\ \emptyset\ , H_i\ , H_j\ \}$.

* pour éviter de parler de l'ensemble des ensembles H_i on utilise le terme synonyme de *collection* d'ensembles.

Soit C un ensemble ordonné fini de m éléments nommés caractères. Chaque caractère c_j de C ne peut prendre qu'un nombre fini de valeurs r_j, ordonnées ou non, nommées valeurs admissibles pour c_j. r_j est nommé cardinalité de c_j. Si $r_j = 2$ le caractère est dit binaire. On appelle cardinalité (ou cardinalité maximale) de C le maximum des r_j qu'on notera pour ce qui suit r.

Une espèce s définie sur C est une suite de m valeurs telle que sa j-ième valeur, nommée état de c_j pour s, soit admissible pour c_j.

Une matrice d'états M sur C est une matrice dont chaque ligne i définit une espèce s_i sur C. La valeur en ligne i et colonne j, notée $s_{i,j}$, définit la valeur du caractère c_j pour l'espèce s_i. Sauf mention explicite contraire, on désignera par n le nombre de lignes de M, c'est à dire le nombre d'espèces.

On identifie parfois s_i et c_j à des fonctions définies par $s_i(j) = c_j(i) = s_{i,j}$.

On supposera, pour simplifier les notations et identifier les états, caractères et espèces que les valeurs prises par M sont 0, 1, $2...r_{j-1}$ ou $1,2...r_j$ quitte à recoder les valeurs de M. On convient de même que pour tout caractère binaire les valeurs admissibles sont 0 et 1 nommés respectivement valeur ancestrale et valeur dérivée. Là encore, quitte à recoder les valeurs, on considère que la valeur 0 est plus présente ou égale à la valeur 1. Une matrice d'états est dite binaire si tous ses caractères sont binaires.

On appelle orbite d'ordre k pour le caractère c_j l'ensemble O_j^k des espèces présentant l'état k pour le caractère c_j c'est à dire les espèces s telles que $s(c_j) = k$. Lorsque c_j est binaire, on nomme orbite [principale] de c_j l'orbite correspondant à la valeur 1 et on note abusivement O_j au lieu de O_j^1.

Deux caractères c_1 et c_2 de C sont dits compatibles par états si leurs orbites sont hiérarchiquement compatibles.

Soit s une espèce sur C. On appelle mutation (ou transition) de s tout changement d'un ou plusieurs états de s et on appelle mutation ponctuelle de s tout changement d'un seul état de s. Si s_1 et s_2 sont deux espèces sur C on appelle nombre de mutations entre s_1 et s_2 le nombre de caractères pour lesquels s_1 et s_2 sont différents.

Remarques :

Dans une matrice d'états, une espèce est identifiée à son indice de ligne, un caractère à son indice de colonne ; dans la pratique on préfére utiliser leurs noms biologiques.

Compte-tenu de ce qui précède, la définition proposée ici pour le terme espéce englobe tout ce qui peut être considéré comme objet ou organisme évolutif. Une espèce pourra donc être, selon les cas ce que les biologistes nomment un taxon, un OTU, une séquence génétique... Le vocabulaire associé à une matrice d'états pour chacun de ces cas est alors spécifique. Ainsi :

- en statistique et analyse des données classiques, les espèces sont nommées individus et les caractères variables aléatoires qualitatives discrètes,
- en mathématiques, les espèces et les caractères [ne] sont [que] des vecteurs d'entiers,
- en stemmatologie (partie de la linguistique historique), les espèces sont des manuscrits, les caractères sont des variantes et les états sont nommés variations,
- en zoologie, les espèces sont soit des "vraies" espèces biologiques au sens de la taxonomie soit des couples (genre,espèce) soit d'autres désignations respectant la nomenclature taxonomique officielle et les caractères peuvent être des traits ($features$) morphologiques comme avoir des ailes, comme le nombre de doigts,
- enfin, en biologie moléculaire, les espèces sont des séquences de nucléotides ou d'acides aminés et les caractères sont nommés sites (ou positions).

On remarquera que pour ce dernier cas une des conséquences des définitions est que pour définir une matrice d'états, les séquences doivent toutes avoir la même longueur ce qui peut obliger à un alignement multiple alors que pour définir une matrice de distances par paire, un alignement deux à deux suffit.

Les valeurs respectives de n, m, r sont très variables en fonction du domaine d'étude. Ainsi $r = 4$ pour les séquences ADN et $r = 20$, pour les séquences d'acides aminés ; lorsqu'on traite des caractères morphologiques, m dépasse rarement 50 alors que pour les séquences ADN il faut envisager $m > 1000$; certaines méthdoes de regroupements sont impraticables pour $n > 20$ alors qu'on "aimerait" pouvoir traiter des matrices d'états pour lesquels $n > 300$.

La fonction qui associe à deux espèces le nombre de mutations entre ces deux espèces est une distance S nommée classiquement distance de Hamming sur l'ensemble des espèces définies sur un même ensemble de caractères. On la note h pour ce qui suit.

Soit T un arbre binaire enraciné. On note A l'ensemble des arcs de T, S l'ensemble des sommets, F l'ensemble des feuilles, N l'ensemble des noeuds internes, c'est à dire des sommets de T qui ne sont pas des feuilles. On rappelle qu'il existe alors un chemin unique entre deux sommets. Si u et v sont deux sommets successeurs de n, on dit que n est un ancêtre [commun] de u et v. Si F contient n éléments alors on a $n-1$ noeuds internes et $2n-2$ arcs.

Soit T un arbre [non enraciné]. On dit que T est un arbre bifurcant sur n points *ssi* T possède n sommets de degré 1 (ou "feuilles"), étiquetés par $1, 2...n$ et si tous les autres sommets (ou "noeuds" [internes]) sont de degré 3. Pour un tel arbre, on a donc $2n-2$ sommets en tout et donc $2n-3$ arcs.

Un arbre bifurcant sur n points est équivalent à un arbre binaire enraciné sur $n-1$ feuilles, la bijection s'obtenant en retirant le dernier noeud, l'arc afférent et en nommant racine le sommet restant de cet arc ou en insérant une racine sur n'importe lequel des $2n-3$ arcs.

Si $B(n)$ désigne le nombre d'arbres bifurcants et $R(n)$ le nombre d'arbres binaires enracinés alors $R(n) = (2n-3)B(n)$ et $B(n) = R(n-1)$.

Soit C un ensemble de caractères, M une matrice sur C et soit E l'ensemble des espèces. On dit que T est un arbre d'évolution pour E (sous entendu via M et C) *ssi*

- T est un arbre binaire enraciné ou un arbre bifurcant,
- toutes les espèces de E étiquettent un sommet de T,
- toutes les feuilles de T sont étiquetées par E,
- tout noeud est étiqueté par une séquence de C.

Si T est un arbre d'évolution pour E, On appelle longueur (de *Hamming*) de T la somme des distances de *Hamming* entre les sommets des arcs pour tous les arcs de T, c'est à dire le nombre total de mutations de l'arbre.

On appelle arbre de [meilleure/maximale] parsimonie pour un ensemble d'espèces E définie par une matrice d'états tout arbre d'évolution T pour E telle que sa longueur est minimale pour l'ensemble des arbres d'évolutions sur E. On appelle alors longueur de parsimonie (ou "score") de T cette longueur.

Le problème de parsimonie [large] consiste, pour une matrice d'états donnée, à trouver un arbre de meilleure parcimonie. Ce problème est NP-difficile, même pour des séquences binaires.

Le problème de parsimonie restreinte consiste, pour un arbre d'évolution donné, à trouver sa longueur de parsimonie. Ce problème peut être résolu en un temps polynomial et le problème de parcimonie large est donc polynomial en le nombre d'arbres du problème.

Théorème d'approximation : soient S une suite de séquences données et $G(S)$ le graphe complet dont les noeuds sont bijectivement étiquetés par S pour lequel le poids de l'arc (a, b) est $h(a, b)$. Si T est un arbre de poids minimum pour $G(S)$ alors la longueur de T est au plus égale au double de la longueur de l'arbre le plus parcimonieux.

On dit qu'il y a mutation convergente ou convergence au noeud n pour le caractère c *ssi* il existe deux successeurs directs de n qui sont des mutations de c vers un même état.

On dit qu'il y a mutation renverséee ou réversion *ssi* pour le sommet k au niveau du caractère c *ssi* il existe deux sommets u et v ancêtres de k tel que u est ancêtre de v avec $c(u) = c(k)$ et $c(v) \neq c(k)$; en d'autres termes : c est passé de la valeur $c(u)$ à la valeur $c(v)$ puis est revenu à la valeur $c(u)$ en k sur le chemin qui va de u à k et qui passe par v.

Soit T un arbre ; on dit que T est un arbre ppa si les arcs de T ont un poids positif ou nul.

Si T est un arbre d'évolution pour E, la longueur de parsimonie de T est définie comme la somme de toutes les distances de hamming entre feuilles de T.

Soit T un arbre ppa ; la fonction qui associe à deux feuilles i et j la somme des arcs le long du chemin (unique) qui relie i à j est une distance sur l'ensemble des feuilles, nommée métrique induite par T notéé pd_T appelée distance-chemin de l'arbre T.

Une matrice de distances M est additive s'il existe un arbre ppa T tel que la métrique induite par T sur les feuilles redonne M.

Si S est un ensemble de séquences, l'arbre de parsimonie minimale de S est un arbre T dont les feuilles sont étiquetées par S tel que la longueur de parsimonie de T est minimale.

Un arbre T est dit ultramétrique s'il est enraciné et si sa distance-chemin entre la racine et chaque feuille est constante.

5. Aspects formels du problème direct

Soit S un ensemble de n éléments que l'on nommera symboliquement espèces bien que ce puisse être des OTUS, des séquences, etc. Soit C un ensemble de m caractères prenant au plus r valeurs ordonnées ou non et soit M la matrice qui décrit l'état des caractères de C pour les n espèces : M est donc une matrice $n \times m$ tel que l'élément $m_{i,j}$ en ligne i colonne j donne l'état du caractère j pour l'espèce i. On note $PP(O, n, m, r)$ les paramètres lorsque les valeurs sont ordonnées, $PP(U, n, m, r)$ les paramètres lorsque les valeurs sont non ordonnées.

5.1 Etude du cas binaire ordonné $PP(O, n, m, r = 2)$

Si les caractères sont binaires et ordonnés, on convient que 0 désigne la valeur ancestrale du caractère et que 1 est la valeur dérivée.

Soit T un arbre orienté dont les feuilles sont les espèces. On dit que la valeur k du caractère c_j est compatible avec l'arbre T si l'ensemble des noeuds (feuilles et noeuds intérieurs) associés à la valeur k pour le caractère c_j est un sous-arbre de T. Si toutes les valeurs de c_j sont compatibles avec T on dit que le caractère c_j est compatible avec T. Si tous les caractères de C sont compatibles avec T on dit que l'ensemble C est compatible.

Un arbre étiqueté enraciné T un arbre phylogénétique parfait pour S et C si les feuilles de T sont les éléments de S, si chaque caractère de C étiquette un arc et si cet arc correspond à la transition de ce caractère de l'état ancestral vers l'état dérivé ce qui équivaut à la condition tous les caractères de C sont compatibles avec T.

Soit O_j l'ensemble des espèces pour lesquelles C_j vaut 1 ; O_j est nommé ensemble dérivé de c_j. La matrice M de caractères binaires admet un arbre phylogénétique parfait ssi les O_j ont une structure de hiérarchie.

Il existe un algorithme de décision pour tester si une matrice M de caractères binaires admet un arbre phylogénétique parfait qui s'exécute intuitivement $\mathrm{O}(nm^2)$ mais qui peut être rendu $\mathrm{O}(nm)$: c'est celui du test de structure de hiérarchie. Lorsqu'il existe un tel arbre, l'algorithme de construction de cet arbre s'exécute en $\mathrm{O}(nm)$.

Un arbre étiqueté enraciné T est dit ultramétrique ssi

- chaque noeud interne est étiqueté par un nombre,
- chaque noeud interne a au plus deux fils,
- le long de chaque chemins de la racine à une feuile les étiquettes décroissent strictement.

Il existe un algorithme en $O(n^2)$ pour reconstruire un arbre étiqueté ultramétrique sur n feuilles.

Si une matrice M de caractères binaires admet un arbre phylogénétique parfait T alors l'étiquettage de T qui consiste à prendre pour chaque noeud interne le nombre de caractères rencontrés fait de T un arbre ultramétrique.

5.2 Etude du cas binaire non ordonné $PP(U, n, m, r = 2)$

Quitte à recoder les données, on suppose que l'état le plus fréquent est 0. Un graphe triangulé est un graphe non orienté tel que tout cycle d'au moins 4 sommets possède une corde c'est à dire un arc qui joint deux sommets non consécutifs du cycle.

A chaque collection de sous-arbres d'un arbre donné T correspond un graphe triangulé et réciproquement.

Le graphe d'intersection d'une collection C d'ensembles est le graphe G dont les sommets sont les élements de C tel qu'il y ait un arc entre deux éléments si leur intersection est non vide.

On dit qu'un graphe $G = (S, A)$ muni d'un coloriage c sur S peut être $c-$triangulé s'il existe un graphe triangulé $H = (S, B)$ tel que B contienne A et que c est un coloriage valide pour H c'est à dire que tout arc de B non présent dans A doit relier deux sommets de couleurs différentes.

Une matrice de caractères dont les caractères définissent un coloriage c admet un arbre phylogénétique parfait ssi le SIG correspondant peut être $c-$triangulé.

Une matrice de 2 caractères dont les caractères définissent un coloriage c admet un arbre phylogénétique parfait ssi le SIG correspondant est acyclique.

Il existe un algorithme de décision $\mathrm{O}(n)$ pour tester l'acyclicité et l'algorithme de construction de l'arbre phylogénétique parfait s'exécute aussi en $\mathrm{O}(n)$.

5.3 Etude du cas général $PP(n, m, r)$

6. Aspects formels du problème direct

7. Bibliographie