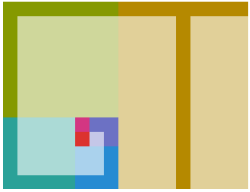

Maths – Bio

Arbres discrets pour la biologie évolutive

Marc Manceau

Séance du 19 novembre 2015

Résumé

Ce document a pour but de donner un petit aperçu des objets mathématiques utilisés pour représenter des arbres, et d'aborder de petits problèmes de combinatoire ainsi que des exemples d'applications pour la biologie. Il s'agit d'un support pour un cours informel de 2h dans le cadre du *GT Maths-Bio* de l'École normale supérieure (`http://www.gt-mathsbio.biologie.ens.fr`), un groupe de travail qui réunit des étudiants biologistes intéressés par les maths.

Il repose en majorité sur deux références dont Mike Steel est auteur : Steel [2014] en particulier, qui donne un aperçu général des questions abordées, sans tout développer de façon très formelle, et Semple and Steel [2003] qui expose beaucoup plus en profondeur et de manière plus formelle toutes les notions.

1 Introduction : l'arbre, un graphe particulier

1.1 Un graphe : des sommets et des arêtes

1.1.1 Vocabulaire de base sur les graphes

Définition 1 *Un graphe $G = (V, E)$ est la donnée d'un ensemble non-vide de sommets V et d'un ensemble d'arêtes $E \subset \{\{x, y\} : x, y \in V^2\}$.*

Par exemple, sur la figure 1A, on a $G = \big(\{\nu_1, \nu_2, \nu_3, \nu_4, \nu_5, \nu_6\}, \big\{\{\nu_1, \nu_2\}, \{\nu_2, \nu_2\}, \{\nu_3, \nu_3\}, \{\nu_3, \nu_4\}, \{\nu_4, \nu_5\}, \{\nu_4, \nu_6\}, \{\nu_5, \nu_6\}, \{\nu_5, \nu_6\}\big\}\big)$.

On introduit à présent un peu de vocabulaire afin de simplifier cet objet très général pour les applications qui nous intéressent :

- Une boucle est une arête qui joint un sommet à lui-même. Sur la figure 1A, $\{\nu_2, \nu_2\}$ est une boucle.

- Des arêtes qui rejoignent les mêmes sommets sont appelées arêtes parallèles. Sur la figure 1A, $\{\nu_5, \nu_6\}$ et $\{\nu_5, \nu_6\}$ sont des arêtes parallèles.

- Un graphe simple est un graphe sans boucles ni arêtes parallèles. Le graphe de la figure 1B est simple.

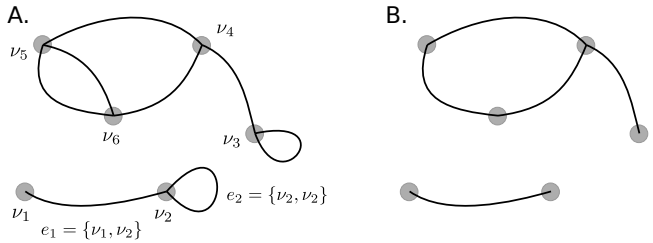

FIGURE 1 – *Deux premiers exemples de graphes. A) Graphe avec boucles et arêtes parallèles. B) Graphe simple.*

Nous aurons également besoin d'un peu de vocabulaire pour décrire les relations entre sommets de nos graphes :

- Si $e = (u, v) \in E$, on dit que u et v sont adjacents, et que e est incident à u et v.
- Si $v \in V$, on appelle degré de v, noté $d(v)$, le nombre d'arêtes de G incidentes à v.
- Un sommet de degré zéro est dit isolé.

Notons que si on somme les degrés de tous les sommets d'un graphe simple, on compte deux fois chaque arête, d'où l'égalité :

$$\sum_{v \in V} d(v) = 2\,|E|$$

1.1.2 Graphe connexe, cycles

Un *chemin* est un uplet de sommets $(\nu_1, \nu_2, \ldots, \nu_k)$ tel que $\forall i \in \{1, 2, \ldots, k-1\}$, ν_i et ν_{i+1} sont adjacents. Sur la figure 1A, $(\nu_3, \nu_4, \nu_5, \nu_6)$ est un chemin. Si, de plus, ν_1 et ν_k sont adjacents, $(\nu_1, \nu_2, \ldots, \nu_k)$ est un *cycle*. Toujours sur la figure 1A, (ν_4, ν_5, ν_6) est un cycle.

Un graphe G est *connexe* si tout couple de sommets peut être rejoint par un chemin. Le graphe de la figure 1A n'est pas connexe car, par exemple, ν_1 et ν_6 ne peuvent pas être rejoints par un chemin. En revanche, les graphes de la figure 2 sont connexes.

Lemme 1 *Soit $G = (V, E)$ un graphe connexe. Alors :*

$$|V| \leq |E| + 1$$

La démonstration se fait par récurrence sur $|V|$.
Initialisation : pour $|V| = 1$ ou $|V| = 2$, ça fonctionne.
Récurrence : supposons que l'inégalité est vraie pour tout graphe vérifiant $|V| = n$. Prenons un graphe $G = (V, E)$ tel que $|V| = n + 1$. Choisissons une arête e dans G et construisons le graphe G/e en "contractant" l'arête e, c'est-à-dire en supprimant e et en identifiant ses deux extrémités. Ce graphe est toujours connexe. Il possède n sommets et $|E|-1$ arêtes. On a donc :

$$n \leq |E| - 1 + 1$$
$$n + 1 \leq |E| + 1$$

□

1.1.3 Graphe orienté

Définition 2 *Un graphe orienté $D = (V, A)$ est la donnée d'un ensemble non-vide de sommets V et d'un ensemble d'arcs $A \subset \{(x, y) : x, y \in V^2\}$.*

Si $a = (u, v)$ est un élément de A, on dit que u est la queue et v est la tête de l'arc. L'arc a est dirigé de u vers v et on le représente avec une flèche.

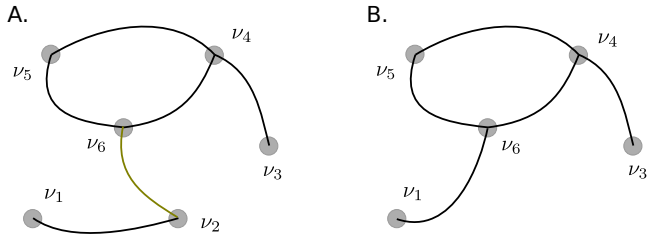

FIGURE 2 – *A) Exemple de graphe connexe. B) Graphe obtenu en contractant l'arête verte du graphe A ($\{\nu_2, \nu_6\}$), c'est-à-dire en la supprimant et en identifiant ν_2 et ν_6.*

Tout le vocabulaire présenté précedemment s'adapte aux graphes orientés en considérant que l'on ajoute l'orientation sur un graphe non-orienté pour lequel on a défini notre terminologie. Par exemple, le graphe de la figure 3A est non-connexe, et présente des boucles et arêtes parallèles. Celui de la figure 3B est simple et connexe.

Enfin, pour un sommet $v \in V$, on parle de degré entrant (resp. sortant) pour compter le nombre d'arcs dont la tête (resp. la queue) est v.

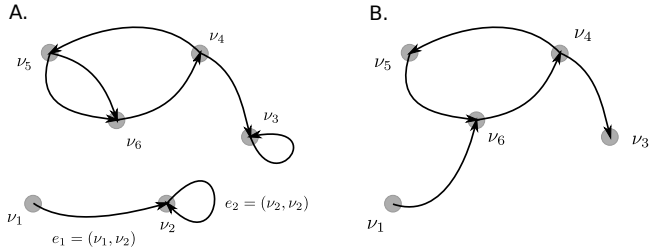

FIGURE 3 – *Exemples de graphes orientés. A) Une orientation possible du graphe de la figure 1A, B) Une orientation possible du graphe de la figure 2B.*

1.2 L'arbre, un graphe particulier

1.2.1 Caractérisation d'un arbre

On arrive à ce qu'est un arbre à ce niveau :

Définition 3 *Un arbre $T = (V, E)$ est un graphe connexe sans cycle.*

Théorème 1 *Soit $G = (V, E)$ un graphe. Les phrases suivantes sont équivalentes :*

1. *G est un arbre.*
2. *$\forall u, v \in V^2$, il existe un unique chemin dans G qui connecte u et v.*

3. G est connexe et $|V| = |E| + 1$.

Prouvons, par exemple, (1) $\Leftrightarrow$ (2) et (1) $\Leftrightarrow$ (3).

Commençons par (1) $\Leftrightarrow$ (2) : supposons que G est un graphe connexe sans cycle. Par définition de la connexité, cela signifie que $\forall u, v \in V^2$, il existe un chemin de u à v. Supposons, de plus, qu'il existe $u, v \in V^2$ tels que deux chemins différents relient u à v. Sans perte de généralité, on considère que ces deux chemins sont composés de sommets tous distincts $(u, \nu_1, \ldots, \nu_k, v)$ et $(u, \nu'_1, \ldots, \nu'_l, v)$ (s'ils ne sont pas tous distincts, on peut trouver un autre couple de sommets dans ces deux chemins qui soit relié par deux chemins constitués de noeuds distincts). Alors $(u, \nu_1, \ldots, \nu_k, v, \nu'_l, \ldots, \nu'_1, u)$ est un cycle. Donc G sans cycle $\Rightarrow$ unicité du chemin de u à v. Enfin, si G a un cycle, alors deux sommets contenus dans le cycle possèdent au moins deux chemins distincts les connectant. Donc l'unicité du chemin $\Rightarrow$ absence de cycle dans G.

Montrons par récurrence sur $|V|$ que (1) $\Rightarrow$ (3) : pour $|V| = 2$, en supposant (1), on a 2 sommets et 1 arête, donc (3) est vrai. Et en supposant (3), on a bien (1). Supposons à présent que $|V| \geq 3$, et

que (1) $\Rightarrow$ (3) pour tout graphe ayant un sommet de moins. G étant connexe sans cycle, il existe au moins un sommet ayant une seule arête incidente. Si on enlève ce sommet et son arête, on se retrouve avec un graphe (V', E') ayant un sommet de moins, vérifiant $|V'| = |E'|+1$, donc on a bien $|V| = |E|+1$.

Montrons, enfin, que (3) $\Rightarrow$ (1) : supposons (3). G est donc connexe, et n'est pas un arbre si il contient un cycle. Supposons que ce soit le cas. Alors il existe une arête contenue dans ce cycle. Supprimons-la du graphe sans aucun autre changement. Le graphe (V, E') en résultant est toujours connexe, et vérifie (par le lemme 1) : $|V| \leq |E'| + 1$. Or, $|E'| = |E| - 1$, donc on obtient $|V| \leq |E|$, ce qui entre en contradiction avec $|V| = |E| + 1$. Donc G est bien un arbre. □

1.2.2 Vocabulaire autour des arbres

En biologie évolutive, les arbres peuvent être orientés ou non, selon les applications :

Orienté on ne s'intéresse alors qu'aux arbres orientés dont toutes les arêtes sont dirigées vers l'extérieur d'un sommet particulier : la racine. On dit alors que l'arbre est *enraciné*.

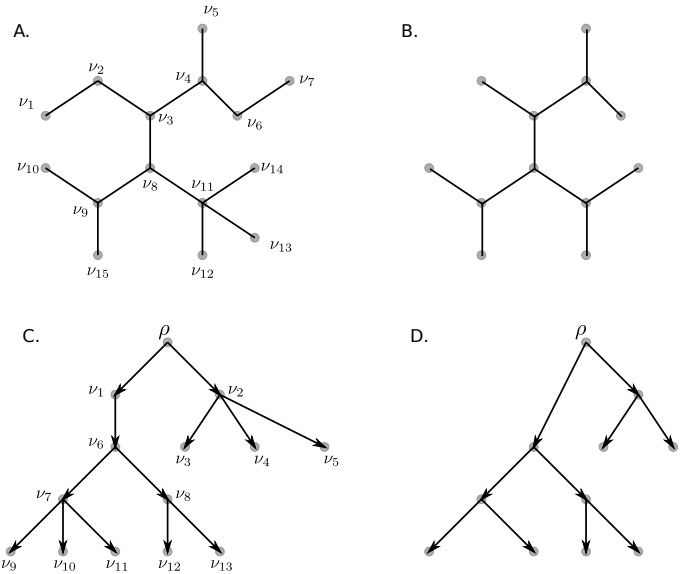

FIGURE 4 – *Vocabulaire pour les arbres. A) Arbre non-enraciné non-binaire.* $\nu_2, \nu_3, \nu_4, \nu_6, \nu_8$ *sont des exemples de sommets intérieurs,* $\nu_1, \nu_5, \nu_7, \nu_{10}, \nu_{15}$ *sont des feuilles. B) Arbre non-enraciné binaire. C) Arbre enraciné non-binaire.* ρ *est la racine,* $\nu_1, \nu_2, \nu_6, \nu_7, \nu_8$ *sont les sommets intérieurs, tandis que* $\nu_3, \nu_4, \nu_5, \nu_9$ *sont des feuilles. D) Arbre enraciné binaire.*

Non-orienté on ne distingue pas de racine, et l'arbre est alors dit *non-enraciné*.

Les sommets d'un arbre ont différents noms selon leurs relations avec les autres :

Racine quand un arbre est enraciné, c'est l'unique sommet ayant un degré entrant de 0.

Feuille tout sommet de degré 1 (ou degré entrant de 1, pour les arbres enracinés).

Sommet intérieur tout sommet de degré au moins deux.

Selon la façon dont ils se divisent, on peut qualifier un arbre de :

Chemin si tous les sommets on un degré d'au plus deux.

Arbre binaire lorsque tout sommet intérieur a un degré exactement égal à trois, sauf la racine (si l'arbre est orienté) qui a un degré égal à deux. Aussi appelés arbres ternaires, cubiques, ou trivalents.

Enfin, à titre d'exemples, on peut décrire ici trois arbres particuliers :

Arbre en peigne (ou *caterpillar* en anglais) un arbre binaire dont le graphe induit par les sommets intérieurs de l'arbre est un chemin.

Arbre en étoile un arbre ayant un seul sommet intérieur.

Arbre équilibré (*balanced* en anglais) un arbre enraciné binaire avec 2^h feuilles, chacune étant séparée de la racine par exactement h arêtes.

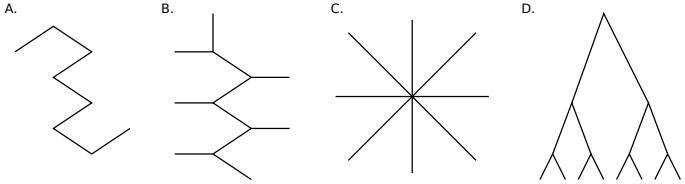

FIGURE 5 – *4 arbres particuliers. A) Chemin. B) Peigne. C) Étoile. D) Arbre équilibré.*

2 Des arbres étiquetés pour la biologie évolutive

2.1 Introduction des X-arbres

On a vu précédemment ce qu'étaient les objets que l'on appelle *arbres*. En biologie évolutive, on utilisera plutôt des arbres dont seules les feuilles sont étiquetées par le nom des espèces ou des individus. Ces étiquettes appartiennent en toute généralité à un ensemble X.

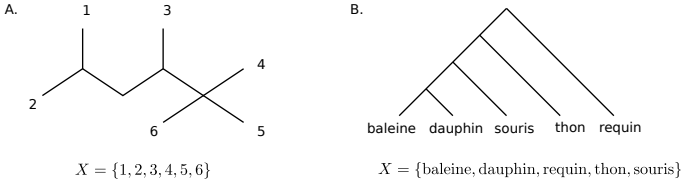

FIGURE 6 – *Deux exemples de X-arbres. A) Un X-arbre non-enraciné. B) Un X-arbre binaire enraciné.*

Définition 4 *Un X-arbre $\mathcal{T}$ est un couple (T, ϕ) où T est un arbre et $\phi : X \to V$ est une application d'étiquetage, qui est une bijection de X dans l'ensemble des feuilles de T (les sommets de degré au plus 1).*

Cette définition implique en particulier qu'on étiquette rien que les feuilles ($\phi(X) = \{\text{feuilles de } T\}$), et que chaque étiquette est attribuée à une et une seule feuille de l'arbre.

Notons que dans le livre qui sert de référence à ce petit résumé, ce que l'on nomme X-arbre ici est appelé X-arbre phylogénétique. Et leur X-arbre nécessite le fait que $\forall v \in V$ de degré au plus 2, $v \in \phi(X)$.

On parlera de X-arbre binaire, ou bien de X-arbre enraciné, non-enraciné, lorsque ces adjectifs s'appliquent à l'arbre T du couple (T, ϕ).

2.2 Nombre d'arêtes d'un X-arbre binaire

On note $\mathrm{RB}(X)$ l'ensemble des X-arbres binaires enracinés, et $\mathrm{UB}(X)$ l'ensemble des X-arbres binaires non-enracinés. Soit n le cardinal de X. On s'intéresse dans cette petite section à $r_n := |\mathrm{RB}(X)|$ et $u_n := |\mathrm{UB}(X)|$. En cherchant à dénombrer ces arbres, on va un peu apprendre à manipuler la différence entre arbres enracinés / non-enracinés.

Lemme 2 *Tout arbre binaire non-enraciné à n feuilles possède $2n-3$ arêtes. Tout arbre binaire enraciné à n feuilles possède $2n-2$ arêtes.*

La démonstration suivante ne concerne que les arbres non-enracinés. Elle peut être facilement adaptée pour les arbres enracinés. Remarquons qu'un arbre binaire non-enraciné peut être transformé en arbre binaire enraciné en coupant une arête en deux pour y mettre un sommet (la racine).

Notre arbre possède $|V| = n + i$ sommets (n feuilles et i sommets intérieurs) et $|E|$ arêtes. De plus, un graphe connexe est un arbre si et seulement si le nombre de sommets excède le nombre d'arêtes de 1 :

$$|V| = n + i = |E| + 1 \quad \Leftrightarrow \quad i = |E| + 1 - n$$

On sait aussi que la somme des degrés des sommets d'un graphe est $2\,|E|$:

$$2\,|E| = n + 3i$$

D'où l'égalité du lemme :

$$2\,|E| = n + 3\big(|E| + 1 - n\big) \quad \Leftrightarrow \quad |E| = 2n - 3$$

□

2.3 Dénombrement des X-arbres binaires, enracinés ou non

On rencontre dans la suite une notation plutôt utilisée pour faire du dénombrement, appelée *double factorielle*, qui consiste à faire suivre un entier de deux points d'exclamation. Ça ne consiste pas à prendre la factorielle de la factorielle, mais simplement à prendre "un terme sur deux" de la factorielle. Un exemple valant mieux qu'un long discours :

$$6!! := 6 \times 4 \times 2 = 48$$
$$7!! := 7 \times 5 \times 3 \times 1 = 105$$

On peut facilement se passer de cette notation, mais je l'ai laissée pour le côté "culturel" de la rencontre !

Lemme 3 *Les nombres de X-arbres enracinés et non-enracinés sur un ensemble X de cardinal n sont respectivement :*

$$r_n = (2n-3)!! = \frac{(2n-2)!}{2^{n-1}(n-1)!}$$
$$u_n = (2n-5)!! = \frac{(2n-4)!}{2^{n-2}(n-2)!}$$

Définissons une première application $s : \mathrm{RB}(X) \to \mathrm{UB}(X)$, qui à un X-arbre enraciné fait correspondre le X-arbre non-enraciné obtenu en supprimant

la racine. Cette application (illustrée en figure 7) est surjective : en effet, si $\mathcal{T} \in \mathrm{UB}(X)$, l'image réciproque $s^{-1}(\{\mathcal{T}\})$ est l'ensemble de tous les arbres enracinés obtenus en choisissant une arête de $\mathcal{T}$, où on positionne la racine. On a donc $|s^{-1}(\{\mathcal{T}\})| = 2n-3$ par le lemme 2. On en déduit alors la relation :

$$r_n = (2n-3)u_n$$

Définissons une seconde application $o_x : \mathrm{RB}(X\backslash\{x\}) \to \mathrm{UB}(X)$, qui, à un $X\backslash\{x\}$-arbre enraciné associe le X-arbre enraciné obtenu en attachant x à la racine par une nouvelle arête. Cette opération, illustrée en figure 7, est appelée "ajout d'outgroup x" en biologie. C'est une bijection, donc :

$$|\mathrm{UB}(X)| = |\mathrm{RB}(X\backslash\{x\})| \text{ , i.e. } u_n = r_{n-1}$$

On obtient donc en résumé :

$$r_n = (2n-3)u_n = (2n-3)r_{n-1}$$

Comme $r_2 = 1$, on obtient par récurrence $r_n = 1 \times 3 \times 5 \times \cdots \times (2n-3) = (2n-3)!!$. De plus $u_n = r_{n-1} = (2n-5)!!$. Enfin, les écritures avec les factorielles simples peuvent être checkées par le calcul. □

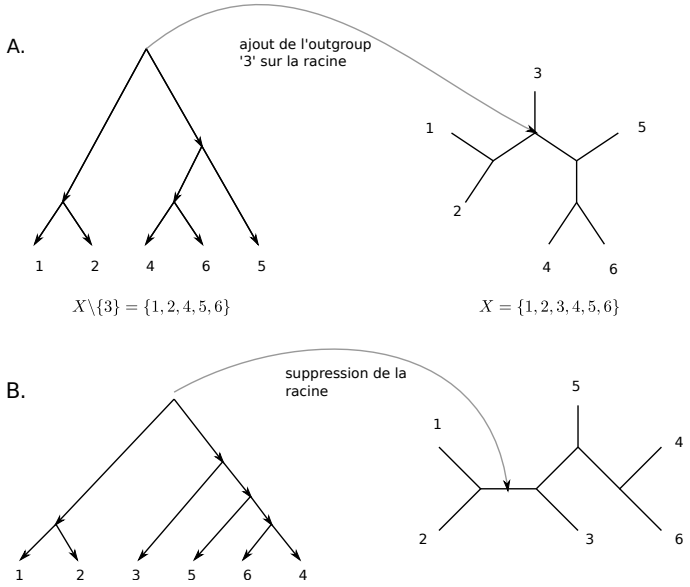

FIGURE 7 – *Relations entre X-arbres enracinés ou non. A) L'application o_3 décrite dans le texte, qui consiste à ajouter un outgroup (l'élément 3) sur l'arbre. B) L'application s décrite dans le texte, qui supprime la racine d'un X-arbre enraciné.*

Notons d'ailleurs que ces expressions peuvent être utiles pour calculer un équivalent en $+\infty$ des suites, à l'aide de la formule de Stirling. Avoir ces équivalents ne sert qu'à se rendre compte que u_n et r_n croissent très vite. Ce dont on peut se rendre compte en faisant de petites applications numériques, par exemple :

$$u_{10} = 2027025$$
$$u_{20} \approx 2.10^{20}$$

Les problèmes consistant à chercher le meilleur X-arbre satisfaisant une propriété ne peuvent donc jamais se permettre d'explorer tous les arbres possibles.

3 Deux représentations ensemblistes des X-arbres

3.1 Introduction des X-hiérarchies et X-splits

Commençons par introduire une notion ensembliste, dont on va voir qu'elle entretient un rapport très étroit avec la notion de X-arbre enraciné.

Définition 5 *Une X-hiérarchie $\mathcal{H}$ est une collection de parties de X telle que :*

1. $\forall A, B \in \mathcal{H}^2,\ A \cap B \in \{A, B, \emptyset\}$
2. $X \in \mathcal{H}$
3. $\forall x \in X,\ \{x\} \in \mathcal{H}$

Le lien entre X-hiérarchies et X-arbres enracinés est illustré en figure 8, et décrit dans les quelques lignes qui suivent.

Prenons un X-arbre enraciné, et construisons un ensemble $\mathcal{I}$ de parties de X, composé de l'ensemble des feuilles portées par chaque sommet de l'arbre. L'ensemble $\mathcal{I}$ est une hiérarchie.

De même, prenons une X-hiérarchie quelconque $\mathcal{H}$. On peut lui associer un arbre enraciné que l'on construit en faisant correspondre à chaque élément de $\mathcal{H}$ un sommet dans l'arbre, et en plaçant une arête entre deux éléments A et B de $\mathcal{H}$ si $A \subset B$ et s'il n'existe pas $C \in \mathcal{H}$, tel que $A \subset C \subset B$.

On arrive donc à un théorème précieux qui donne l'équivalence entre une hiérarchie et un arbre enraciné :

Théorème 2 *Une collection $\mathcal{H}$ de parties de X est une hiérarchie ssi $\mathcal{H}$ est l'ensemble des clusters associés à un X-arbre enraciné $\mathcal{T}$. De plus, $\mathcal{T}$ est unique à l'équivalence près.*

L'équivalence dont il est question ici est bien celle que l'on imagine intuitivement : deux X-arbres $\mathcal{T}$ et $\mathcal{T}'$ sont équivalents lorsqu'il existe une permutation des nœuds internes des arbres qui permette à $\mathcal{T}$ et $\mathcal{T}'$ d'être identiques. Cette précaution permet d'oublier les noms des nœuds internes.

Notons au passage que la plus grosse hiérarchie sur un ensemble X de cardinal n correspond à un ensemble de clusters d'un arbre binaire. Elle a donc pour taille $2n - 1$, le nombre de sommets d'un X-

arbre binaire enraciné. On notera dans le suite $\mathcal{H}(\mathcal{T})$ la hiérarchie associée aux clusters de l'arbre $\mathcal{T}$.

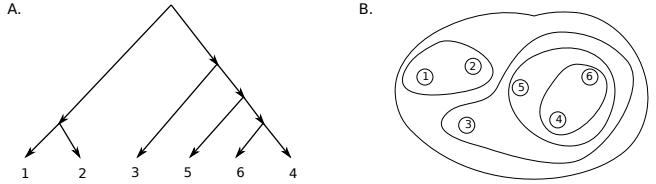

FIGURE 8 – *A) Exemple de X-arbre binaire enraciné, avec $X = \{1, 2, 3, 4, 5, 6\}$ et B) sa hiérarchie associée, où chaque cercle correspond à un ensemble qui appartient à la hiérarchie.*

On introduit à présent une seconde notion ensembliste, mieux adaptée pour travailler avec des X-arbres non-enracinés.

Définition 6 *Un X-split noté $A|B$ est une bi-partition de X en deux ensembles non-vides A et B.*

Sur tout arbre non-enraciné, si on supprime une arête e et qu'on considère les deux ensembles de feuilles séparées par cette arête, on obtient "le X-split qui correspond à e". Sur la figure 9 par exemple, le X-split correspondant à l'arête verte est $\{1, 2\}|\{3, 4, 5, 6\}$.

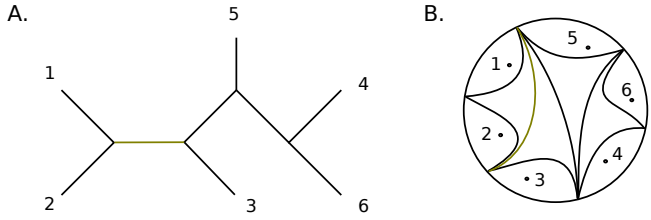

FIGURE 9 – *A) Exemple de X-arbre binaire non-enraciné, avec $X = \{1, 2, 3, 4, 5, 6\}$ et B) sa collection de X-splits induite, où l'ensemble X (cercle noir) est découpé en X-splits par des courbes noires. En vert, le X-split correspondant à l'arête verte du X-arbre.*

Notons à présent que $\forall A|A', B|B'$, deux X-splits correspondant à deux arêtes différentes sur le même X-arbre, l'une des quatre intersections "croisées" suivante est vide :

- $A \cap B$
- $A \cap B'$
- $A' \cap B$
- $A' \cap B'$

On peut s'en convaincre en regardant ce que ça signifie sur un dessin, par exemple en figure 9, ou

bien peut-être plus facilement en regardant la contraposée : si aucune des intersections est vide, on comprend vite que les deux X-splits correspondent à la même arête.

Lorsqu'une collection de X-splits satisfait cette propriété, on dit qu'elle est compatible deux à deux. Ce type de collection de X-splits joue avec les arbres non-enracinés un rôle analogue au rôle des hiérarchies avec les arbres enracinés, grâce au théorème d'équivalence qui suit :

Théorème 3 *Une collection Σ de X-splits est l'ensemble des splits d'un X-arbre non-enraciné ssi*

1. *Σ est compatible deux à deux.*
2. *Σ contient les splits triviaux : $\forall x \in X$, $\{x\}|X\backslash\{x\} \in \Sigma$.*

On admet ici ce théorème, dont la démonstration, plutôt longue, est trouvable dans Semple and Steel [2003]. Dans la suite, on appelle *collection de splits induite par $\mathcal{T}$*, et on note $\Sigma(\mathcal{T})$ l'ensemble des splits qui correspondent aux arêtes de $\mathcal{T}$.

On avait précédemment décrit une bijection de l'ensemble des $X\backslash\{x\}$-arbres enracinés dans l'ensemble des X-arbres non-enracinés, qui consistait à attacher x à la racine de l'arbre enraciné. On décrit à présent la bijection existant entre les $X\backslash\{x\}$-hiérarchies et les X-splits compatibles deux à deux.

Soit $\Sigma_{\text{tot}}(X)$ l'ensemble de tous les X-splits possibles, et $\mathcal{P}(X\backslash\{x\})$ l'ensemble des parties de X. On décrit deux fonctions :

1. $\psi_x^- : \Sigma_{\text{tot}}(X) \to \mathcal{P}(X\backslash\{x\})$, qui, à un split $A|B$ associe :

$$\psi_x^-(A|B) = \begin{array}{l} A \text{ si } x \in B \\ B \text{ si } x \in A \end{array}$$

2. $\psi_x^+ : \mathcal{P}(X\backslash\{x\}) \to \Sigma_{\text{tot}}(X)$ qui, à un sous-ensemble A de $X\backslash\{x\}$ associe :

$$\psi_x^+(A) = A|(X\backslash A)$$

On remarque que, $\forall A|B \in \Sigma_{\text{tot}}(X)$, $\psi_x^+(\psi_x^-(A|B)) = A|B$, et, $\forall A \in \mathcal{P}(X\backslash\{x\})$, $\psi_x^-(\psi_x^+(A)) = A$. Munis de cette bijection, et de nos deux théorèmes d'équivalence entre d'une part les X-arbres non-enracinés et les X-splits compatibles deux à deux,

et d'autre part entre les $X\backslash\{x\}$-hiérarchies et les $X\backslash\{x\}$-arbres enracinés, on peut démontrer le théorème suivant :

Théorème 4 *Soit x un élément de X.*

1. *Un ensemble Σ de X-splits est compatible deux à deux ssi :*

$$\{\psi_x^-(A|B) \,:\, A|B \in \Sigma\} \text{ est une } X\backslash\{x\} - \text{hiérarchie}$$

2. *Un ensemble $\mathcal{H}$ de parties de $X\backslash\{x\}$ est une hiérarchie ssi :*

$$\{\psi_x^+(A) \,:\, A \in \mathcal{H}\} \text{ est un ensemble de } X-\text{splits compatibles 2 à 2}$$

3.2 Comparer deux X-arbres

L'utilisation de représentations ensemblistes nous permet de comparer à présent facilement deux X-arbres $\mathcal{T}$ et $\mathcal{T}'$ donnés.

Définition 7 *On appelle distance de Robinson-Foulds entre deux X-arbres non-enracinés $\mathcal{T}$ et $\mathcal{T}'$, et l'on note $d(\mathcal{T}, \mathcal{T}')$, le nombre de X-splits présents dans l'un ou l'autre des deux arbres mais pas les deux.*

$$d(\mathcal{T}, \mathcal{T}') \;=\; |\Sigma(\mathcal{T}) \triangle \Sigma(\mathcal{T}')|$$

où, pour deux ensembles A et B quelconques, $A \triangle B = (A \cup B) \backslash (A \cap B)$.

On appelle distance de Robinson-Foulds entre deux X-arbres enracinés $\mathcal{T}$ et $\mathcal{T}'$, et l'on note $d(\mathcal{T}, \mathcal{T}')$, le nombre de clusters présents dans l'un ou l'autre des deux arbres mais pas les deux.

$$d(\mathcal{T}, \mathcal{T}') \;=\; |\mathcal{H}(\mathcal{T}) \triangle \mathcal{H}(\mathcal{T}')|$$

On justifie aisément le fait que ce sont des distances :

Symétrie par symétrie de la différence symétrique $\triangle$, $d(\mathcal{T}, \mathcal{T}') = d(\mathcal{T}', \mathcal{T})$.

Séparation $d(\mathcal{T}, \mathcal{T}') = 0$ ssi tous les splits (resp. clusters) de $\mathcal{T}$ et $\mathcal{T}'$ sont identiques, i.e. $\mathcal{T} = \mathcal{T}'$ par les théorèmes d'équivalence précédents.

Inégalité triangulaire Montrons que $\forall \mathcal{T}, \mathcal{T}', \mathcal{T}''$, $d(\mathcal{T}, \mathcal{T}'') \leq d(\mathcal{T}, \mathcal{T}') + d(\mathcal{T}', \mathcal{T}'')$.
Ceci revient à dire que les cardinaux des différences symétriques vérifient $|A \triangle C| \leq |A \triangle B| + |B \triangle C|$. Si on en est pas persuadé, on peut le vérifier aisément en décomposant les différents ensembles en ensembles disjoints.

3.3 Construction d'un X-arbre consensus

Maintenant que l'on sait comparer deux X-arbres grâce à la distance de Robinson-Foulds, on voudrait pouvoir en combiner un certain nombre pour obtenir un arbre consensus, qui reflète ce dont on est le plus sûr dans l'histoire de X. Les hiérarchies ou les splits nous offrent une façon rapide d'exprimer ça.

Soient $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k$ k X-arbres enracinés, et $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_k$ leur hiérarchie correspondante. On construit alors le consensus majoritaire comme étant :

$$\mathcal{H}^* = \{C \in \bigcup_{i=1}^{k} \mathcal{H}_i, \text{ vérifiant } |\{j \ : \ C \in \mathcal{H}_j\}| > \frac{k}{2}\}$$

De façon analogue, si $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k$ sont k X-arbres non-enracinés, et $\Sigma_1, \Sigma_2, \ldots, \Sigma_k$ sont leurs ensembles de X-splits engendrés, on construit le consensus majoritaire comme étant :

$$\Sigma^* = \{\sigma \in \bigcup_{i=1}^{k} \Sigma_i, \text{ vérifiant } |\{j \ : \ \sigma \in \Sigma_j\}| > \frac{k}{2}\}$$

Lemme 4 *$\mathcal{H}^*$ est une hiérarchie, et Σ^* est une collection de splits compatibles deux à deux.*

On écrit la preuve dans le cas des hiérarchies, et c'est très similaire pour les collections de splits.

On peut vérifier d'abord les points 2 et 3 de la définition d'une hiérarchie. Concernant l'ensemble $X : \forall j, X \in \mathcal{H}_j$, donc $X \in \mathcal{H}^*$. Concernant les singletons, de même, $\forall x \in X, \forall j, \{x\} \in \mathcal{H}_j \Rightarrow \{x\} \in \mathcal{H}^*$.

Pour vérifier le premier point de la définition, prenons C et C' dans $\mathcal{H}^*$. L'ensemble des $\mathcal{H}_j$ qui contiennent C est de cardinal supérieur à $k/2$. De même, l'ensemble des $\mathcal{H}_j$ qui contiennent C' est de cardinal supérieur à $k/2$. Par le principe des tiroirs, il existe au moins un $\mathcal{H}_i$ tel que C et C' appartiennent à $\mathcal{H}_i$. Donc $C \cap C' \in \{C, C', \emptyset\}$. □

Notons que l'on peut définir des arbres consensus en imposant de ne prendre que les clusters ou les splits présents dans une plus grande proportion $q \geq 1/2$ de nos arbres de départ. En revanche, pour $q < 1/2$, le point utilisant le principe des tiroirs dans la démonstration ne fonctionne plus, et on ne peut pas être assuré que C et C' sont toujours "emboîtés".

Un exemple de construction de X-arbre comme consensus majoritaire de 4 arbres différents est illustré en figure 10.

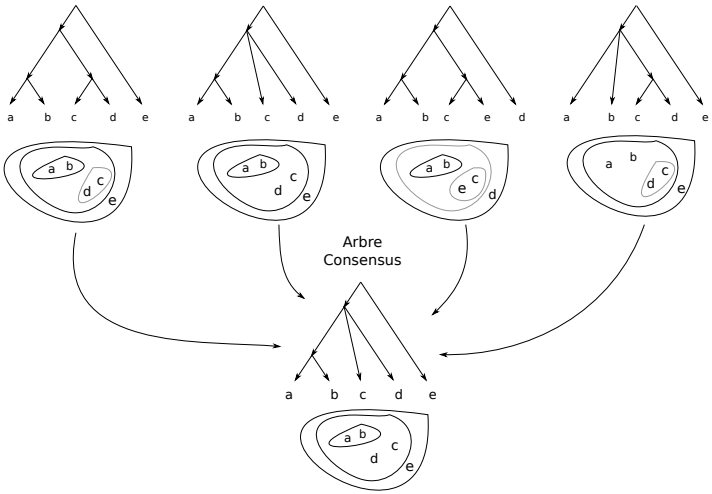

FIGURE 10 – *Création d'un arbre consensus majoritaire de 4 X-arbres enracinés (avec leur hiérarchie associée). Seuls les clusters présents chez au moins 3 arbres sont conservés pour le consensus. Notons que les singletons ne sont pas entourés dans les hiérarchie pour faciliter la lecture.*

4 Ajout de caractères sur un arbre

4.1 Introduction des caractères

Définition 8 *On appelle* caractère *à r états sur X une fonction surjective f de X dans un ensemble S de r éléments (les états du caractère).*

Pour tout $x \in X$, $f(x)$ donne l'état du caractère chez l'individu x.

Si l'évolution des individus de X est donnée par un X-arbre $\mathcal{T} = (T, \phi)$, on voudrait assigner un état de caractère à chaque sommet de T. Il y a donc évidemment plusieurs extensions F de la fonction f à l'ensemble V des sommets de T. En biologie, on fait souvent l'hypothèse que les caractères ont évolué sans *homoplasie*, i.e. sans réversion ni évolution convergente. Formellement, on peut caractériser la possibilité d'évolution sans homoplasie de plusieurs façons, dont celle qui suit :

Définition 9 *On dit qu'un caractère sur X, f, est convexe sur un X-arbre $\mathcal{T} = (T, \phi)$ avec $T = (V, E)$ si il existe une fonction $F : V \to S$ vérifiant :*

1. $F \circ \phi = f$
2. $\forall s \in S$, *le sous-graphe induit par* $\{v \in V$ *tels que* $F(v) = s\}$ *est connexe.*

Le sous graphe induit par $\{v \in V$ t.q. $F(v) = s\}$ consiste simplement à considérer cet ensemble de sommets ainsi que les arêtes qui les lient entre eux, en oubliant tout le reste du graphe. Cette construction est illustrée à l'aide des couleurs sur les figures 11A et 11C.

On dit qu'une telle extension F, quand elle existe, représente l'histoire évolutive d'une caractère ayant évolué sans homoplasie.

Notons qu'un caractère f binaire (à deux états) induit une bipartition de X, soit un X-split. Le caractère est convexe sur un X-arbre ssi le X-split qu'il induit est un X-split du X-arbre (voir figures 11A et 11B).

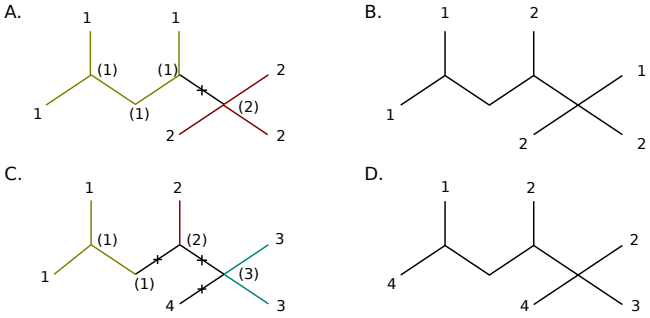

FIGURE 11 – *Quatre caractères sur un même arbre. A) Un caractère binaire convexe, et son unique extension sans homoplasie F à l'ensemble de noeuds internes, entre parenthèses. Le sous-graphe induit par $\{v \in V$ tels que $F(v) = 1\}$ est représenté en vert, tandis que le sous-graphe induit par $\{v \in V$ tels que $F(v) = 2\}$ est colorié en rouge. B) Un caractère binaire non-convexe. C) Un caractère à 4 états convexe, et une proposition d'extension sans homoplasie F à l'ensemble des noeuds, avec des couleurs différentes sur chaque sous-graphe induit par une valeur du caractère. D) Un caractère à 4 états non-convexe.*

4.2 Dénombrement de caractères convexes

À titre d'illustration, on peut se demander combien de caractères $f : X \to S$, avec $|X| = n$ feuilles et $|S| = r$ états, sont convexes sur un X-arbre binaire donné. On pourrait à priori s'attendre à ce que le résultat dépende de la forme de l'arbre, mais ce n'est en fait pas le cas.

Lemme 5 *Soit X un ensemble de cardinal n, $\mathcal{T} = (T, \phi)$ un X-arbre binaire, et S un ensemble de r éléments. Le nombre de caractères à r états $f : X \to S$ qui sont convexes sur $\mathcal{T}$ est :*

$$\chi(\mathcal{T}, r) = r! \binom{2n - r - 1}{r - 1}$$

Montrons ce résultat par récurrence sur r :

Initialisation : Pour $r = 2$, les caractères convexes sont obtenus à partir des X-splits associés à l'arbre. Et on a autant de X-splits associés à l'arbre que d'arêtes dans l'arbre : $2n - 3$. Il reste à placer les deux caractères de part et d'autre du split, ce qui donne bien le résultat recherché : $2(2n - 3)$.

Récurrence : Supposons que la propriété est vraie pour tout n, pour r inférieur ou égal à $p-1$, et montrons que c'est aussi vrai pour tout n au rang $r = p$. Sélectionnons un sommet adjacent à deux feuilles u et v. On appelle $\mathcal{T}'$ l'arbre obtenu en supprimant la feuille u, et $\mathcal{T}''$ l'arbre obtenu en supprimant u et v. Les caractères à p états sur $\mathcal{T}$ appartiennent à trois ensembles disjoints :

C_1 si u et v ont des états différents uniquement présents chez eux. On choisit ces deux états et on considère le nombre de caractères convexes à $(p-2)$ états sur le reste de l'arbre.

$$|C_1| = p(p-1)\chi(\mathcal{T}'', p-2)$$

C_2 si v a le même état que u. On considère alors le nombre de caractères convexes à p états sur $\mathcal{T}'$.

$$|C_2| = \chi(\mathcal{T}', p)$$

C_3 si uniquement l'une des deux feuilles u ou v a un état uniquement présent chez elle. On choisit s'il s'agit de u ou de v (facteur 2), puis on soustrait au nombre de caractères pour lesquels au moins une feuille a un état uniquement présent chez elle ($p\chi(\mathcal{T}', p-1)$ le nombre de

caractères pour lesquels les deux feuilles ont un état uniquement présent chez elle (cf. C_1).

$$|C_3| = 2\left(p\chi(\mathcal{T}', p-1) - p(p-1)\chi(\mathcal{T}'', p-2)\right)$$

Au final, on obtient donc :

$$\begin{aligned}
\chi(\mathcal{T},p) &= p(p-1)\chi(\mathcal{T}'',p-2)+\chi(\mathcal{T}',p)+2(p\chi(\mathcal{T}',p-1)-p(p-1)\chi(\mathcal{T}'',p-2)) \\
&= \chi(\mathcal{T}',p)+2p\chi(\mathcal{T}',p-1)-p(p-1)\chi(\mathcal{T}'',p-2) \\
&= p!\left(\binom{2(n-1)-p-1}{p-1}+2\binom{2(n-1)-(p-1)-1}{p-2}-\binom{2(n-2)-(p-1)-1}{p-3}\right) \\
&= p!\binom{2n-p-1}{p-1}
\end{aligned}$$

La dernière ligne nécessite plus de lignes de calcul, mais on s'en tire grâce à la formule de Pascal. On en conclut que la propriété est vraie pour tout n et pour tout r. □

4.3 Deux questions pour aller plus loin

4.3.1 Existe-t-il un arbre tel que des caractères donnés soient compatibles ?

Définition 10 *On dit qu'une séquence* $(f_1, f_2, \dots f_k)$ *de caractères sur* X possède une phylogénie parfaite *si et seulement si il existe un* X*-arbre sur lequel l'évolution de chaque caractère peut se faire sans homoplasie (i.e. sans réversion ni évolution convergente). Le* X*-arbre est alors appelé* phylogénie parfaite *pour ces caractères, et les caractères sont dit* compatibles.

Lorsqu'on possède uniquement une séquence de caractères sur X, il n'est pas trivial de vérifier qu'ils sont compatibles. Il existe cependant des résultats donnant des conditions nécessaires ou suffisantes pour avoir des caractères compatibles, ainsi que des algorithmes permettant de vérifier si des caractères donnés sont compatibles, en cherchant une phylogénie parfaite. Le problème général est NP-complet, mais sous certaines hypothèses simplificatrices il existe des algorithmes qui checkent ça en temps polynomial. En cas d'intérêt pour la question, se référer au livre Semple and Steel [2003].

4.3.2 Combien de caractères sont nécessaires pour définir un arbre ?

Quand une séquence de caractères possède une phylogénie parfaite, on a envie de se demander si cette phylogénie est unique. En biologie évolutive cette question est très intéressante, puisqu'elle signifie que les caractères en question permettent d'inférer avec certitude l'histoire évolutive des organismes (toujours sous l'hypothèse d'absence d'homoplasie).

On considère donc à partir de maintenant uniquement des arbres binaires. En effet, si on s'autorisait des arbres non-binaires, on pourrait au moins "résoudre" cet arbre en un arbre binaire, ce qui ferait au moins deux autres phylogénies parfaites, comme illustré en figure 12.

La question générale qu'on se pose est la suivante :

Quel est le plus petit nombre $h(n)$ tel que tout X-arbre non-enraciné binaire à n feuilles est l'unique phylogénie parfaite d'une séquence de $h(n)$ caractères ?

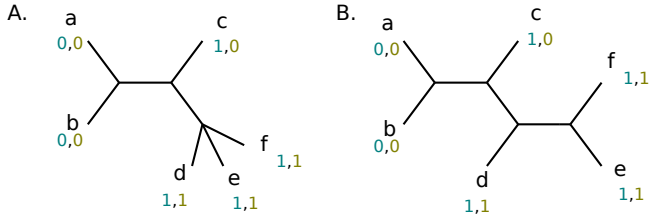

FIGURE 12 – *Histoire évolutive de deux caractères (l'un bleu, l'autre vert) à deux états. A) Une phylogénie parfaite non binaire pour cette séquence de caractères. B) Une autre phylogénie parfaite binaire associée à la même séquence de caractères. Notons que ce n'est pas l'unique phylogénie parfaite binaire.*

Évidemment, ça dépend des caractères que l'on considère. Plus les caractères ont d'états, et moins on s'attend à avoir besoin de caractères pour définir notre arbre.

Lemme 6 *Pour des caractères à deux états, on a* $h(n) = n - 3$.

En effet, la séquence de $n - 3$ caractères correspondant aux $n - 3$ splits non-triviaux de l'arbre (i.e. associés à des branches qui ne mènent pas aux feuilles) ont cet arbre comme unique phylogénie parfaite. Imaginons qu'on enlève un de ces caractères

associé à un split : la nouvelle collection de X-splits n'est plus équivalente à l'arbre de départ. Le plus petit nombre de caractères binaires permettant de définir une unique phylogénie parfaite est donc bien $n-3$. □

Ce lemme est bien illustré en figure 12 : seuls deux caractères à deux états sont alors considérés, ce qui ne permet pas de résoudre la relation existant entre d, e, et f.

Mais plus étonnamment, si on n'impose pas de limite au nombre d'états, il est possible de montrer le théorème suivant (ce que nous ne ferons pas ici !)

Théorème 5 *Pour tout X-arbre $\mathcal{T}$, il existe une séquence de 4 caractères pour laquelle $\mathcal{T}$ est l'unique phylogénie parfaite.*

Même sans aborder la démonstration, on peut toucher une vague sensation de compréhension en visualisant la construction d'une telle séquence de caractères, faite dans Steel [2014].

Prenons une orientation donnée de l'arbre, et attribuons à chaque branche qui part à gauche un attribut (l, l') et à chaque branche qui part à droite un

attribut (r, r'), avec la règle que deux arêtes avec prime ou sans prime ne peuvent se toucher entre elles. Chaque arête l est responsable d'un changement du premier caractère, chaque arête r est responsable d'un changement du second caractère, l' pour le troisième et r' pour le quatrième caractère. Chaque changement de caractère fait apparaître un nouvel état. Éventuellement, les mêmes états peuvent exister pour deux caractères.

À partir de la donnée des 4 caractères sur les feuilles, on peut remarquer qu'il est possible de reconstruire l'arbre, ainsi que l'état ancestral en tout sommet intérieur de l'arbre. Mais il n'est pas évident de montrer que notre arbre est l'unique phylogénie parfaite des 4 caractères.

Enfin, il est également possible de montrer que trois caractères évoluant sans homoplasie ne suffisent pas à définir un X-arbre. Donc 4 est bien le plus petit nombre de caractères qui permet de définir toute phylogénie.

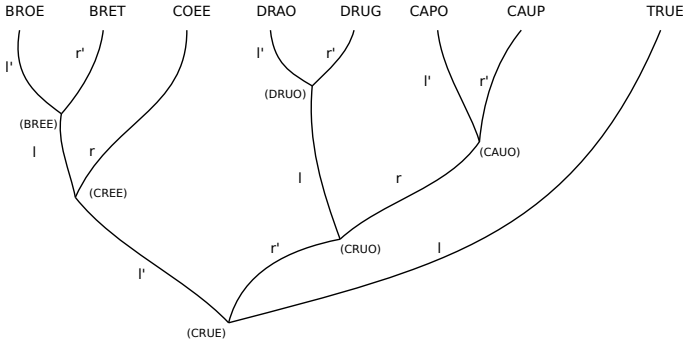

FIGURE 13 – *Quatre caractères évoluant sans homoplasie suffisent pour définir un X-arbre : exemple d'une construction de 4 caractères définissant un arbre donné, tirée de Steel [2014].*

Références

Charles Semple and Mike A Steel. *Phylogenetics*, volume 24. Oxford University Press, 2003.

Mike A Steel. Tracing evolutionary links between species. *The American Mathematical Monthly*, 121 (9) :771–792, 2014.