Recherche dans les arbres généalogiques

Sujet proposé par Didier Rémy

Didier.Remy@inria.fr

1 Préambule

Le projet consiste à créer et explorer un arbre généalogique. Outre des recherches simples d'informations liées aux individus et l'énumération de leurs ascendants et de leurs descendants, on s'intéressera aussi à la recherche des liens de parenté entre deux personnes. Le coeur du projet est l'implémentation d'un algorithme de calcul efficace du degré de parenté entre deux individus.

2 Détail du sujet

L'informatique est ici un avantage important par rapport au support papier, car une fois l'arbre rentré en machine il sera possible d'effectuer rapidement de multiples opérations : lire des informations liées à un individu, énumérer les enfants ou les parents d'une personne ou rechercher les parents communs à plusieurs personnes. Voire effectuer des analyses statistiques sur une base de donnée (comme par exemple, une étude sur des animaux de race).

Le but du sujet est de permettre la saisie d'un arbre généalogique à partir d'un fichier, sa modification interactive, sa sauvegarde dans un fichier, et la consultation de la base. On pourra aussi éventuellement fournir des analyses statistiques sur la base. On pourra, par exemple, s'inspirer du logiciel GeneWeb.

Le logiciel devra être robuste, *i.e.* traiter correctement les cas pathologiques et accepter des bases de donnée de taille raisonable (au minimum plusieurs milliers d'entrées). Comme base de donnée, on pourra utiliser de petits exemples concrets (tel que votre propre arbre généalogique, celle d'une famille célèbre, des rois de France, un exemple puisé dans des œuvres littéraires, voire dans la mythologie grecque...) mais aussi considérer quelques cas extrême, sans oublier de traiter également une base de plus grande taille (récupérée électroniquement ou générée par programme).

Un aspect important du projet est l'implémentation du calcul efficace des liens de parenté et plus particulièrement du degré de parenté entre deux individus.

Saisie des données La saisie devra se faire à partir d'un fichier texte ASCII, composé d'une suite d'entrées dans un format à définir. On devra vérifier que

les données fournies sont cohérentes, en particulier l'absence de cycle.

Les données devront pouvoir être sauvegardées au minimum dans le même format. (On pourra si nécessaire utiliser un second format binaire permettant une relecture rapide.)

Consultation de l'arbre généalogique La question la plus simple est la lecture des informations attachées à une personne. On écrira également une procédure qui affiche les descendants $\downarrow(x)$ et les ascendants $\uparrow(x)$ d'une personne x, éventuellement jusqu'à une profondeur donnée en paramètre. Pendant cette énumération, il sera important de ne pas répéter plusieurs fois les descendants (ou ascendants) d'une même personne par plusieurs chemins.

Recherche des liens de parenté Un chemin de longeur k entre deux individus est une suite $x_0 ... x_k$ tel que x_{k+1} est un enfant de x_k. Un chemin de longueur non nulle est dit *sous-chemin propre*. Un lien de parenté entre deux personnes x et y est un ancêtre commun a et deux chemins $a..x$ et $a..y$ qui ne comporte pas de sous-chemin propre commun (les chemins peuvent donc simplement se croiser) ; nous le noterons $x..\hat{a}..y$. La longueur d'un lien de parenté est la somme des longueurs des chemins $a..x$ et $a..y$. La hauteur d'un lien de parenté est le maximum des longueurs des chemins $a..x$ et $a..y$. Le calcul des liens de parenté de hauteur minimale ou de longeur minimale est une information pertinente que l'on peut réaliser de façon efficace. Un lien de parenté est *minimal* si les chemins ax et ay ne se croisent pas (auquel cas, le point de croisement est un chemin de parenté plus court). L'énumération de tous les liens de parenté minimaux entre deux personnes est une façon de calculer leur degré de parenté. Cette opération peut malgré tout être coûteuse, et au besoin on limitera la hauteur des liens de parenté recherchés en fonction de la hauteur minimale de tous leurs liens de parenté.

Consanguinité et parenté Le génôme d'un individu est constitué d'un grand nombre de gènes, qui si l'on ignore les mutations, se reproduisent de façon identiques. Les gènes peuvent donc servir à mesurer l'identité d'une personne. Les gènes sont disposés à des emplacements précis, appelés locus. Chaque individu possède pour chaque locus deux gènes transmis l'un par sa mère, l'autre par son père, et transmet à ses enfants une copie de l'un des deux gènes. La *consanguinité* d'un individu x est la probabilité $cg(x)$ de trouver à un locus donné deux gènes identiques. Le *degré de parenté* de deux individus x et y est la probabilité $pr(x, y)$ de trouver à un même locus deux gènes identiques. Un calcul de probabilité montre que si

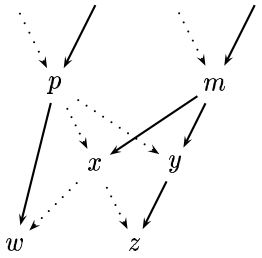
- la consanguinité $cg(x)$ est égale à la parenté $pr(p_x, m_x)$ des parents p_x et m_x de x ;
- si $x..\hat{a}..y$ est un lien de parenté entre x et y, alors il contribue à la parenté de x et de y d'un facteur $1/2^{n+1}$ où n est la longueur du lien de parenté $x..\hat{a}..y$; la parenté de x et de y est la somme des contributions de tous leurs liens de parenté.

En fait, on peut combiner les deux formules ci-dessus et se restreindre dans le calcul de la parenté à des liens de parenté minimaux :

- si $x..\hat{a}..y$ est un lien de parenté minimal entre x et y, alors il contribue à la parenté de x et de y d'un facteur $1/2^{n+1}(1 + cg(a))$ où n est la longueur du lien de parenté $x..\hat{a}..y$; la parenté de x et de y est la somme des contributions de tous leurs liens de parenté minimaux.

L'avantage de cette dernière formule est de réduire considérable le nombre de chemin à considérer. À un lien de parenté minimal $x..\hat{a}..y$ correspond n liens de parenté se coupant en a où n est le nombre de liens de parentés issus des parents de a. Par récurrence, le nombre total de liens de parentés entre deux individus peut être l'exponentiel par le nombre de générations du nombre de liens de parentés minimaux entre ces deux individus.

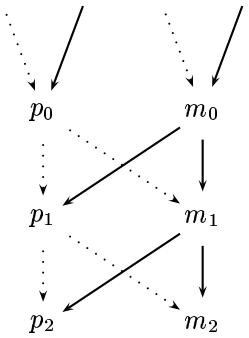
Voici une illustration du calcul du degré de parenté sur un exemple très incestueux...



$$\begin{aligned} pr(p,m) &= 0 \\ pr(x,y) &= 1/4 \\ pr(p,x) &= 1/4 \\ pr(w,z) &= 5/16 \\ pr(p,w) &= 3/8 \end{aligned}$$

Considérons, par exemple, le calcul de $pr(w,z)$. Les liens de parenté minimaux sont $w\hat{p}xz$, $w\hat{p}yz$, $w\hat{x}z$, $wx\hat{p}yz$ et $wx\hat{m}yz$. Les sommets de ces liens de parenté x, p et m ont une consanguinité de 0, car leurs parents n'ont pas de lien de parenté. Ainsi $pr(w,z)$ est égal à la somme $1/8 + 1/16 + 1/16 + 1/32 + 1/32$, soit $5/16$. Pour en savoir plus, on pourra lire par exemple les chapitres 1 et 6 de [Jac70].

Un exemple pathologique intéressant (ci-contre) est celui de frère p_n et sœur m_n dont les deux parents sont frère p_{n-1} et sœur m_{n-1}, *etc.* remontant ainsi jusqu'à n-générations à deux parents p_0 et m_0 d'origines inconnues.



Notons pr_n la parenté $pr(p_n, m_n)$ et cg_n la consanguinité de p_n qui par symétie est aussi égale à celle de m_n. Notons également $\mathcal{Q}_n$ l'ensemble de tous les chemins de parenté entre p_n et m_n, et $|q|$ la longueur du chemin q.

Pour $n \geq 2$, l'ensemble $\mathcal{Q}_n$ se décompose en

- les chemins $(p_n\hat{p}_{n-1}m_n)$ et $(p_n\hat{m}_{n-1}m_n)$, qui ont une contribution de $2 \times 1/2^3 = 1/4$ à pr_n ;
- les chemins $(p_np_{n-1}qm_{n-1}m_n)$ et $(m_np_{n-1}qm_{n-1}m_n)$ lorsque q parcourt

$\mathcal{Q}_{n-1}$; l'ensemble de ces chemins apporte une contribution de :

$$2 \times \sum_{q \in \mathcal{Q}_{n-1}} 1/2^{|q|+2+1} = 1/2 \times \sum_{q \in \mathcal{Q}_{n-1}} 1/2^{|q|+1} = (1/2) \times pr_{n-1}.$$

– les chemins $(p_n p_{n-1} p_{n-2} q m_{n-2} p_{n-1} m_n)$, $(p_n p_{n-1} m_{n-2} q p_{n-2} p_{n-1} m_n)$, $(p_n m_{n-1} p_{n-2} q m_{n-2} m_{n-1} m_n)$, et $(p_n m_{n-1} m_{n-2} q p_{n-2} m_{n-1} m_n)$, lorsque q parcourt $\mathcal{Q}(p_{n-2}, m_{n-2})$; l'ensemblde de ces chemins apporte une contribution de

$$4 \times \sum_{q \in \mathcal{Q}_{n-2}} 1/2^{|q|+4+1} = 1/4 \times \sum_{q \in \mathcal{Q}_{n-2}} 1/2^{|q|+1} = (1/4) \times pr_{n-2}.$$

Nous avons donc la formule récurrente, $pr_n = 1/4 + 1/2 pr_{n-1} + 1/4 pr_{n-2}$ avec $pr_0 = 0$ et $pr_1 = 1/4$. Il est facile de vérifier que pr_n (ou cg_n) tend vers 1 lorsque $n \longrightarrow \infty$.

On pourra tester à la fois la correction du calcul de consanguinité et la robustesse de l'algorithme en vérifiant expérimentalement cette convergence.

3 Calcul efficace de la consanguinité

On note N est le nombre d'individus dans la base de donnée généalogique et H le nombre maximum de générations. (Notez que N peut facilement atteindre 10 000 ou 100 000, alors que 200 générations permet de remonter à Jésus-Christ (pour une généalogie humaine). On remarquera que le nombre d'arcs est au plus $2N$ donc en $O(N)$. (Un autre coefficient important est le nombre Q de générations maximun entre un individu et ses enfants qui est le rapport entre l'âge minimal et l'âge maximal de procréation ; par exemple, pour les mammifères, il est raisonable de considérer que Q est compris entre 4 et 6 selon les espèces.)

Étant donné une base généalogique où la consanguinité de chaque individu est connue, il est possible de calculer le lien de parenté entre deux individus quelconques de la base en temps $O(N)$. On peut donc, en visitant les nœuds dans un ordre chronologique, calculer la consanguinité pour toute une population en temps $O(N^2)$.

Soit x et y les individus dont on calcule le degré de parenté. Le lien de parenté entre x et y est défini à partir des liens de parenté minimaux, mais ne nécessite pas de calculer ceux-ci explicitement, car seule leur contribution importe. En particulier, lorsqu'un nœud appartient à plusieurs liens de parenté minimaux, il est possible de calculer sa contribution au degré de parenté de $pr(x, y)$ en factorisant sa contribution dans tous les chemins issus de x ou de y. Globalement, on peut ainsi ne visiter ainsi chaque nœud qu'une seule fois.

Pour cela, on parcourt en parallèle les ancêtres de x et de y dans un ordre tel que les fils sont toujours visités avant leur père. Au passage, on annote chaque nœud z avec :

- Une couleur choisie parmi quatre : la couleur initiale des nœuds non visités, la couleur des ancêtres de x, la couleur des ancêtres de y et la couleur des ancêtres communs à x et y.
- Un coefficient z_x mesurant la contribution potentielle de l'ensemble des chemins issus de x égale à la somme $\Sigma_{p\in[x,z]}1/2^{\ell_p}$ où $[x,z]$ est l'ensemble de chemins issus de x et ℓ_p est la longueur de p.
  Un coefficient z_y similaire pour les chemins issus de y.
- Un coefficient correctif z_c, mesurant la contribution des chemins comportant des sous-chemins propres.

Lorsqu'un nœud z est visité, il faut

- Mettre à jour la couleur et les coefficients des parents de z à partir des coeficients de z : si w est un parent de z, alors (1) $z_x/2$ est la contribution de z à w_x, similairement pour y, et $z_xz_y/4$ est la contribution de z à w_c.
- Calculer la contribution de z au degré de parenté de x et de y, égale à $(z_xz_y - z_c(1 + cg(z)))/2$.

En effet, on se convaincra facilement que z_xz_y mesure la contribution de tous les chemins $x..\hat{z}..y$, y compris ceux partageant un sous-chemin propre. Le coefficient correcteur z_c retire la contribution des chemins de la forme $x..\vec{u}\hat{z}\overleftarrow{u}y$ se terminant par un sous-chemin propre partagé. Enfin, le terme $z_ccg(z)$ retire la contribution (ultérieure) des chemins de la forme $x..\vec{u}zx'..\hat{w}..y'z\overleftarrow{u}y..z$ qui arrivent à z par un sous-chemin partagé puis se séparent en z pour se rejoindre plus tard (on note $\vec{u}$ et $\overleftarrow{u}$ deux sous-séquences non nulles de nœuds en ordre inverse.)

4 Travail minimum demandé

Structures de données et algorithmes Expliquer clairement, le format de représentation des données dans un fichier texte, puis leur représentation en mémoire. Décrire précisément les algorithmes de recherche.

Fonctionnalités minimum requises

- La lecture d'un arbre généalogique dans un fichier texte.
- La détection des incohérences et des invraisemblances lorsque les dates de naissance sont disponibles (y compris la détection des cycles).
- L'énumération des descendants et des ascendants d'une personne jusqu'à une profondeur donnée, en ne répétant pas les chemins multiples.
- La recherche des liens de parenté de longueur ou de hauteur minimale.
- L'énumération de tous les liens de parenté minimaux (éventuellement bornée par une certaine hauteur)
- Le calcul du degré de parenté de deux personnes par énumération des liens de parenté minimaux, servant de calcul de référence.
- Le calcul efficace des consanguinités de toute une généalogie.

On fournira aussi une petite interface pour saisir facilement les différentes requêtes, et imprimer les réponses de façon lisible.

Illustrations Dans le rapport, on montrera un extrait d'un arbre généalogique sur lequel on fera quelques exemples de recherche, les cas pathologiques ou extrêmes étant évidemment les plus intéressants...

5 Extensions facultatives

Le calcul de la consanguinité peut fournir des informations statistiques sur l'arbre considéré. Par exemple, on pourra chercher la consanguinité moyenne pour un individu de la famille, et prendre cette valeur moyenne plutôt que 0 comme la consanguinité des personnes dont l'un des parents est d'origine inconnue.

On pourra aussi considérer les liens de parenté par alliance. Au sens naturel, il y a alliance entre deux personnes si elles ont des enfants en commun.

On pourra éventuellement compléter l'affichage des liens de parenté par une description de la relation de parenté la plus forte. Par exemple, reconnaître les relations de parenté simples telles que x *est le beau-frère de* y ou x est le fils du neveu de y, et dans les cas difficiles *le lien de parenté est trop haut*, se contenter d'indiquer pour un lien $x..\hat{a}..y$ la hauteur minimale et la différence des hauteurs entre $a..x$ et $a..y$; par exemple, *le grand-père de* x *est cousin au 4ème degré avec* y.

La saisie ou la modification de certaines parties de l'arbre peut se faire de façon interactive pendant le déroulement du programme. Mais, cela nécessite de pouvoir sauvegarder l'arbre en mémoire dans un fichier texte qui puisse être relu plus tard afin de ne pas perdre les modifications.

Références

[Jac70] Albert Jacquard. *Structures Génétiques des Populations.* Masson & Cie, 1970.