TIPE "Gene" Rank

COPPA Maxime

Année 2021-2022

1 Introduction

Etudier le niveau d'expression des gènes d'un tissu, d'un organe ou d'une cellule est très souvent utilisé en cancérologie pour étudier les causes génétiques d'un cancer.
Cependant chaque individu possède plus de 20 000 gènes dans son organisme, il convient donc de mettre en place une méthode pour mettre en avant une liste de gènes prioritaires à étudier dans un ensemble de gènes. Un telle technique permettrait alors de comprendre plus rapidement les processus biologiques de chaque organisme vivant et de rendre les études biologiques plus rapides.
Nous allons donc essayer de répondre au problème suivant :

Etant donné un ensemble S de gènes, est-il possible de mettre en avant de façon efficace, une liste de gènes prioritaires à étudier ?

Pour répondre à se problème, nous allons proposer un algorithme permettant de "noter les gènes". Pour se faire nous proposerons une représentation des gènes de l'organisme sous forme de graphe. Nous utiliserons alors cette structure pour implémenter un tel algorithme.

2 Une représentation des gènes sous forme de graphe

Pour représenter les gènes d'un organisme vivant sous forme de graphe nous allons utiliser une base de données nommée Gene Ontology GO.

Gene Ontology

Gene Ontology (GO) est une base de donnée qui structure la description des gènes et des produits géniques. GO est commune à toutes les espèces vivantes. Elle attribue des termes GO pour caractériser les gènes et leurs produits. Les gènes peuvent dès lors être annotés, caractérisés par un ou plusieurs GO term.

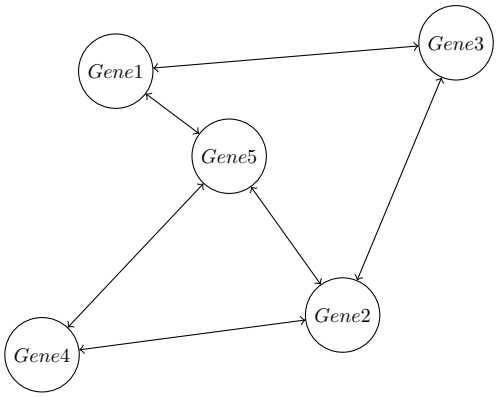
Nous allons nous servir de GO ontology pour mettre en relation plusieurs gènes. En effet, pour deux gènes données nous dirons qu'ils sont en relation si et seulement si ils possèdent un GO term en commun.
Ainsi, pour un ensemble donnée S de gèhes, on va lui associer le graphe non orienté $G = (S, E)$ où $\forall (i, j) \in S^2 \;\; (i, j) \in E \Leftrightarrow$ les gènes i et j possède un terme GO en commun.

Pour faciliter les représentations graphiques et la complexité des fonctions, nous classerons les termes GO en "high-level ontology terms" c'est à dire des termes qui regroupent un grand nombre de termes GO étant biologiquement rapprochées.

Un exemple :

Gènes	Termes GO				
	Terme 1	Terme 2	Terme 3	Terme 4	Terme 5
Gene 1	√	√			
Gene 2			√	√	√
Gene 3		√	√	√	√
Gene 4					√
Gene 5	√				√

Terme 1 = organisation des composantes cellulaires
Terme 2 = établissement de la localisation
Terme 3 = Signalement
Terme 4 = développement du système
Terme 5 = réponse à un stimulus



Remarque : Toutes les données utilisées proviennent de la base de donnée d'organisme modèle Mouse Genome Database (MGD).
Une base de donnée d'organisme modèle est une base de données biologique d'un organisme modèle (une espèce non humaine qui permet de comprendre des phénomènes biologiques ; exemple : les mouches, les souries, les serpents ...).
MGI est une base de données dont les informations sont issus de laboratoires de souris partout dans le monde.

Le problème initial revient dès lors à mesurer "l'importance" d'un noeud dans un graphe $G = (S, E)$ où S est l'ensemble des gènes et E comme défini plus haut.
Dans la suite pour faciliter les notations nous considérerons que $|S| = n \in \mathbb{N}$ et que $S = [\![1; n]\!]$

3 Mesurer l'importance des noeuds d'un graphe non pondéré : l'algorithme PageRank

3.1 Un modèle de mesure

Pour mesurer l'importance d'un noeud dans un graphe $G = (S, E)$ non pondéré nous allons utilisé le principe suivant : "plus un noeud est important, plus de noeuds importants pointent vers celui-ci." Une arête du graphe peut alors être vu comme un "vote" d'un noeud vers un autre. Plus un noeud est important plus de noeuds importants vont pointer vers celui-ci. Il faut dès lors d'attribuer à chaque noeud un "score" d'une telle popularité qui permette de mesure l'importance de ce noeud : plus le score sera élevé, plus le noeud sera jugé important.

On va de plus considérer que plus un lien est précis plus celui-ci est important, moins un noeud possède de liens plus il aura de chance de voter pour un noeud en particulier !

Pour mesurer cette importance, nous considérerons donc un individu : Max qui se "proméne" sur le graphe. L'importance d'un noeud s peut dès lors être vu comme la probabilité qu'après un temps "infini" Max se trouve sur ce noeud. Il convient dès lors de proposer une modélisation de la position de Max à l'instant t et de l'étudier après un temps infini.

3.1.1 Le comportement de Max :

On notera dans la suite, $\mathcal{M}_n(\mathbb{R})$ l'ensemble des matrices de taille n à coefficients dans $\mathbb{R}$, pour toute matrice $M \in \mathcal{M}_n(\mathbb{R})$, $\forall (i,j) \in [\![1; n]\!]^2$ $m_{i,j}$ le coefficient de M sur la $i_{ème}$ ligne et $j_{ème}$ colonne, $\forall X \in \mathbb{R}^n$, $\forall i \in [\![1; n]\!]$ x_i ou $(X)_i$ la $i_{\text{ème}}$ composante de X et $\|.\|_1$ la norme 1 de $\mathbb{R}^n$,

Pour tout instant t, nous désignerons par un vecteur colonne $X_t \in \mathbb{R}^n$ la position de Max à l'instant t tel que $\forall i \in [\![1, n]\!]$ x_i désigne la probabilité qu'à l'instant t Max se trouve sur le sommet i. On notera $\forall t \in \mathbb{N}$, Y_t la position de Max à l'instant t.
Nous allons donc déterminer la position de Max à l'instant $t + 1$ en fonction de sa position à l'instant t.

Un premier modèle
Soit $t \in \mathbb{N}$,
En appliquant le principe que nous avons énoncé précédemment, si Max se trouve à l'instant t sur le sommet j, il va suivre de façon équiprobable un noeud émanant de ce sommet. En notant $\forall i \in S, \quad d_i$ le nombre d'arrêtes émanants du sommet i, $\forall i \in S \quad d_i = \sum 1_{(i,j)\in A}$ on peut écrire :

$$\forall (i,j) \in S^2 \quad \mathbb{P}(Y_{t+1} = i | Y_t = j) = \begin{cases} \frac{1}{d_j} \text{ si } (i,j) \in A^2 \\ 0 \text{ sinon.} \end{cases}$$

Cependant ce modèle soulève un problème : Max peut rester bloqué sur un noeud. En effet si par exemple il commence sa marche sur un noeud ne possèdant aucun autre lien, il ne pourra en bouger et va y rester.

Une amélioration de ce modèle :
Nous allons donc utiliser un modèle développé par Larry Page et Sergey Brin pour modéliser le comportement de Max. Soit $i \in S$, si Max se trouve sur le noeud i à l'instant $k \in \mathbb{N}$ alors Max a deux possibilités :

— Soit il suit une arète du graphe et se retrouve sur un sommet adjacent à i avec une probabilité α
— Soit il se téléporte et se retrouve sur un sommet quelconque du graphe avec une probabilité $1 - \alpha$

Ainsi,

$$\forall (i,j) \in S^2 \quad \mathbb{P}(Y_{t+1} = i | Y_t = j) = \begin{cases} \alpha \frac{1}{d_j} + \frac{1-\alpha}{n} \text{ si } (i,j) \in A^2 \\ \frac{1-\alpha}{n} \text{ sinon.} \end{cases}$$

Remarque :

— On remarque que le comportement de Max est indépendant du temps, on notera alors $\forall t \quad \in \mathbb{N}\mathbb{P}$("Max se trouve en i au temp 1"|"Max se trouve en j au temps t") $= \epsilon_{i,j}$
— De plus :

$$\forall j \in [\![1;n]\!] \sum_{i=1}^{n} \epsilon_{i,j} = (1-\alpha) \sum_{i=1}^{n} \frac{1}{n} + \frac{\alpha}{d_j} \sum_{(i,j)\in A} 1 = (1-\alpha) + \alpha = 1!$$

Le comportement de Max lui permet de ne pas se faire piéger par un noeud sans issue, et lui garantit d'arriver n'importe où dans le graphe.

Ainsi, on peut écrire que $\forall t \in \mathbb{N}$:

$$X_{t+1} = MX_t$$

où $M \in \mathcal{M}_n(\mathbb{R})$ est la matrice définie par :

$$\forall (i,j) \in [\![1;n]\!]^2, m_{i,j} = \epsilon_{i,j} = \begin{cases} \alpha \frac{1}{d_j} + \frac{1-\alpha}{n} \text{ si } (i,j) \in A^2 \\ \frac{1-\alpha}{n} \text{ sinon.} \end{cases}$$

La position de Max est donc donnée $\forall t \in \mathbb{N}$ par $X_t = M^t X_0$. Etudier le comportement asymptotique de Max est donc lié aux propriétés de la matrice M, et plus particuliérement à ses éléments propres. Nous allons dès lors étudier quelques propriétés de la matrice M et le comportement asymptotique de la suite $(M^t X_0)_{t\in\mathbb{N}}$

3.1.2 Une première interprétation du paramètre α

Un surfeur aléatoire qui, à chaque étape, suit au hasard un des liens de la page sur laquelle il se trouve avec une probabilité α ou bien se téléporte avec probabilité $1 - \alpha$ suit en moyenne $\frac{\alpha}{1-\alpha}$ liens entre deux téléportations.

Preuve : On note X = " le nombre de fois que Max suit un lien entre deux téléportations successives". On a alors

$$\forall k \in \mathbb{N}, \mathbb{P}(X = k) = \alpha^k (1-\alpha)$$

X suit donc une loi géométrique de raison le paramètre α. Son espérance, qui correspond au nombre des liens qu'il suit depuis le noeuds où il se trouve avant de se téléporter, vaut donc $\frac{\alpha}{1-\alpha}$

4 Etude des valeurs propres de la matrice M

4.1 Quelques définitions

Matrices Positives

Soit $A = (a_{i,j}) \in \mathcal{M}_n(\mathbb{R})$ une matrice : - on note $A \geqslant 0$ si $\forall (i,j) \in [|1,n|]^2, \quad a_{i,j} \geqslant 0$
- on note $A > 0$ si $\forall (i,j) \in [|1,n|]^2, \quad a_{i,j} > 0$
Soit $B \in \mathcal{M}_n(\mathbb{R})$:
On note $A \geqslant B$ (resp. $A > B$) lorsque $A - B \geqslant 0$ (resp. $A - B > 0$)

Soit $X = (x_i) \in \mathbb{R}^n$:
- on note $X \geqslant 0$ si $\forall i \in [|1,n|], \quad x_i \geqslant 0$
- on note $X > 0$ si $\forall i \in [|1,n|], \quad x_i > 0$

Propriété : Soit $(A,X) \in \mathcal{M}_n(\mathbb{R}) \times \mathbb{R}^n$ tel que $A > 0$ et $X > 0$ alors $AX > 0$
Remarque : la preuve est évidente par un calcul trivial.

Matrices Stochastiques :

Soit $A = (a_{i,j \in [\![1;n]\!]^2}) \in \mathcal{M}_n(\mathbb{R})$ une matrice carré d'ordre n.
On dit que A est une matrice stochastique si et seulement si, $A \geqslant 0$ et $\forall i \in [\![1;n]\!] \sum_j a_{i,j} = 1$

On dit que A est une matrice stochastique en colonnes si et seulement si, ${}^tA$ est stochastique. Ou plus concrétement si $\forall j \in [\![1;n]\!] \sum_i a_{i,j} = 1$

Soit $X = (x_{i \in [|1,n|]}) \in \mathbb{R}^n$ un vecteur colonne.
On dit que X est un vecteur colonne stochastique si et seulement si, $X \geqslant 0$ et $\sum_j x_j = 1$

Propriété 1 :
Soit $A \in \mathcal{M}_n(\mathbb{R})$
Si A est stochastique, alors $1 \in Sp(A)$ et toutes les valeurs propres λ vérifient $|\lambda| \leqslant 1$
De même si A est stochastique en colonnes.

Preuve : Soit $A \in \mathcal{M}_n(\mathbb{R})$, une matrice stochastique.
On note $X_0 \in \mathbb{R}^n$ tel que $\forall i \in [\![1;n]\!], \quad x_i = 1$
De façon évidente il vient que $AX_0 = X_0$. Donc 1 est une valeur propre de A.

Soit $\lambda \in Sp_{\mathbb{C}}(A)$ alors $\exists X \in \mathbb{R}^n$ 0 tel que $AX = \lambda X$. On prend le $i \in [\![1,n]\!]$ tel que $|x_i| = \max(|x_j|_{j \in [\![1,n]\!]}$. On a alors :

$$|\lambda x_i| = |\lambda||x_i| = |\sum_{j=1}^{n} a_{i,j}x_j| \leqslant \sum_{j=1}^{n} |a_{i,j}x_j| \leqslant |x_i| \sum_{j=1}^{n} |a_{i,j}| \leqslant |x_i| \text{ d'où } |\lambda| \leqslant 1$$

Car A est stochastique et X est différent de 0 donc $|x_i| > 0$. Ainsi toutes les valeurs propres λ de A vérifient $\lambda \leqslant 1$

Les valeurs propres de A et ${}^tA$ étant les mêmes, en effet $\forall B \in \mathcal{M}_n(\mathbb{R}), \quad \det(B) = \det({}^tB)$, on en déduit le résultat pour les matrices stochastiques en colonnes.

Propriété 2 :
Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice stochastique en colonnes et $X \in \mathbb{R}^n$ un vecteur stochastique, alors $\forall k \in \mathbb{N}, \quad A^kX$ est un vecteur stochastique.
De plus, si (A^kX) converge alors la limite est un vecteur stochastique.

Preuve :
Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice stochastique en colonnes et $X \in \mathbb{R}^n$ un vecteur stochastique. Pour démontrer le premier point nous allons montré que AX est un vecteur stochastique.
On note b_i les coordonnées de AX.

$\forall i \in [\![1; N]\!]$, $b_i \geqslant 0$ car multiplication de réels positifs.

$$\sum_{i=1}^{n} b_i = \sum_{i=1}^{n}(\sum_{j=1}^{n} a_{i,j}x_j) = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{i,j}x_j = \sum_{j=1}^{n} x_j = 1$$

En effet A stochastique en colonnes et X vecteur stochastique. D'où le résultat.

Supposons que $(A^k X)$ converge vers X_0 alors $\forall k \in \mathbb{N}$ $\|A^k X\|_1 = 1$ d'où par continuité de la norme en dimension finie on a $\|X_0\| = 1$ et comme par multiplication tous les coefficients de X_0 sont positifs on en déduit que X_0 est un vecteur stochastique. D'où le résultat.

4.2 Un théorème sur les valeurs propres de M, le théorème de Frobenius

Théorème : Soit $A \in \mathcal{M}_n(\mathbb{R})$ tel que $A > 0$ alors $\exists \lambda_0 \in Sp_{\mathbb{C}}(A)$ tel que :

1. $\forall \lambda \in Sp_{\mathbb{C}}(A)$ si $\lambda \neq \lambda_0$ alors $|\lambda| < |\lambda_0|$
2. E_{λ_0} est un espace vectoriel de dimension 1 et possède un vecteur $X \geqslant 0$ tel que $\|X\| = 1$, on appellera vecteur de Perron ce vecteur.

Une preuve :
Soit $A \in \mathcal{M}_n(\mathbb{R})$ tel que $A > 0$

Construction du λ_0
On note l'ensemble $S = \{X \in \mathbb{R}^n \text{ tel que } X \geqslant 0 \text{ et } \sum_{i=1}^{n} x_i = 1\}$.
S est borné par définiton et par caractérisation séquentielle S est fermé donc S est un fermé borné de $\mathbb{R}^n$. $\mathbb{R}^n$ étant de dimension fini, S est compact.

On définit alors $\Delta = \{\lambda \in \mathbb{R} \mid (\exists X \in S), AX \geqslant \lambda X\}$.
Δ est majorée, par exemple par le maximum de la somme des lignes de A et Δ est non vide. En effet, on peut construire un $\lambda \in \Delta$, par exemple en notant X le vecteur tel que $x_1 = 1$ et $\forall i \in [\![2, n]\!]$, $x_i = 0$, de façon immédiate $X \in S$. $AX \geqslant \lambda_1 X$ où $\lambda_1 = \min_{j \in [\![1,n]\!]} a_{1,j}$, ainsi $\lambda_1 \in \Delta$
Ainsi, par propriété de la borne supérieure Δ possède une borne supérieure λ_0

Montrons que $\lambda_0 \in \Delta$ et que λ_0 est une valeur propre de A associé à un vecteur propre $X \geqslant 0$, et X non nul.
λ_0 est la borne supérieure de Δ donc il existe $(\lambda_n) \in \Delta^{\mathbb{N}}$ tel que $\lim \lambda_n = \lambda_0$
De plus, $\forall n \in \mathbb{N}, \exists X_n \in S$ tel que $AX_n \geqslant \lambda_n X_n$

On construit ainsi une suite $(X_n) \in S^{\mathbb{N}}$. S est compact donc $\exists \phi$ strictement croissante tel que $(X_{\phi(n)})$ converge ; on note X sa limite qui est dans S car S compact.
Pour tout $n \in \mathbb{N}$ $AX_{\phi(n)} \geqslant \lambda_{\phi(n)} X_{\phi(n)}$ d'où en passant à la limite $AX \geqslant \lambda_0 X$
Ainsi, on en déduit que $\lambda_0 \in \Delta$
Il ne reste maintenant plus qu'à montrer que $AX = \lambda_0 X$

Raisonnons par l'absurde :
Si $AX \neq \lambda_0 X$, alors $\exists i \in [\![1, n]\!]$ tel que $(AX)_i > \lambda_0 X_i$, or on a comme $A > 0$, $A(AX) > \lambda_0(AX)$ d'où $\exists \epsilon > 0$ tel que $A(AX) > (\epsilon + \lambda_0)AX$ ce qui contredit la maximalité de λ_0 d'où $AX = \lambda_0 X$.

La borne supérieure de Δ, λ_0 est une valeur propre de A associée à un $X \in S$ et $X \geqslant 0$

Nous allons maintenant montrer que ce vecteur est unique et que $\forall \lambda \in Sp(A), |\lambda| < \lambda_0$

$$\ddots$$

Première propriété de λ_0, $\forall \lambda \in Sp(A), \lambda \neq \lambda_0$, $|\lambda| < \lambda_0$

Preuve :
Soit $\lambda \neq \lambda_0$ une valeur propre de A, $\exists Y \in \mathbb{C}^n \backslash \{0\}$ tel que $AY = \lambda Y$ or par inégalité triangulaire on a alors $AY^* \geqslant |\lambda| Y^*$ où Y^* est le vecteur tel que $\forall i \in [\![1, n]\!]$ $y_i^* = \frac{|y_i|}{\|Y\|_1}$. D'où $|\lambda| \in \Delta$ et ainsi $|\lambda| \leqslant \lambda_0$

Montrons maintenant que $|\lambda| \neq \lambda_0$
Par l'absurde, si $|\lambda| = \lambda_0$, $A > 0$ donc $\exists \delta > 0$ tel que $A_\delta = A - \delta I_n > 0$. En appliquant le même argument que précédemment on montre alors que $\lambda_0 - \delta$ est la valeur propre maximale de A_δ
Ainsi on a $\forall \lambda \in \mathrm{Sp}(A)$, $|\lambda - \delta| \leqslant \lambda_0 - \delta$
Or, on a supposé : $\lambda_0 = |\lambda| \leqslant |\lambda - \delta| + \delta \leqslant \lambda_0$
D'où $|\lambda| = |\lambda - \delta| + \delta$, ce qui n'est possible que si λ est un réel positif donc on en déduit que $\lambda = \lambda_0$ ce qui est impossible car λ avait été supposé différent de λ_0. On en déduit donc la propriété : $\lambda_0 > \lambda$

$\ddots$

Seconde propriété : Le sous-espace E_{λ_0} est un espace vectoriel de dimension 1.

Preuve :
Nous avons montré qu'il existe $X > 0$ tel que $X \in E_{\lambda_0}$.
Supposons que $\dim E_{\lambda_0} > 1$ alors $\exists Y \in E_{\lambda_0}$ tel que (X, Y) est une famille libre. Alors en posant $\mu = min\{x_i/|y_i|,\ \ y_i \neq 0\}$ on a $X - \mu Y \geqslant 0$ mais $X - \mu Y$ possède une coordonnée nulle.

Or $A > 0$ donc $A(X - \mu Y) > 0$ ie $\lambda_0(X - \mu Y) > 0$ ou encore $X - \mu Y > 0$ ce qui est en contradiction avec la définition de μ ; ainsi $\dim E_{\lambda_0} = 1$

D'où le théorème de Frobenius.

4.3 Application à notre problème

La matrice M comme défini précédemment est strictement postive car $\frac{1-\alpha}{n} > 0$; de plus M est stochastique d'où d'après la propriété 1 des matrices stochastiques $\lambda_0 = 1$, ainsi d'après le théorème de Frobenius : $\exists! R \in \mathbb{R}^n$ tel que $MR = \lambda_0 R = R$, $R \geqslant 0$ et $\|R\|_1 = 1$

Nous allons dès lors montrer que la suite $(M^t X_0)_{t \in \mathbb{N}}$ converge vers ce vecteur.

5 Convergence de la suite des $(M^t X_0)_{t \in \mathbb{N}}$

5.1 Un théorème de convergence

Enoncé du théorème : Soit $A \in \mathcal{M}_n(\mathrm{R})$, une matrice à coefficients strictement positifs et stochastiques en colonnes et R sont vecteur de Perron, alors $\forall X_0 \in \mathrm{R}^n$ vecteur stochastique, la suite $A^k X_0$ converge vers R en $\mathcal{O}(|\lambda_2|^k)$ où $\lambda_2 = \max\{|\lambda|, \lambda \in Sp(A), \lambda \neq 1\}$

Preuve :
Soit $A \in \mathcal{M}_n(\mathrm{R})$, une matrice à coefficients strictement positifs et stochastique en colonnes. On note R son vecteur de Perron.
Soit X_0 un vecteur stochastique positif.

Si A est diagonalisable :
On note $Sp(A) = \{\lambda_1, ..., \lambda_n\}$ les valeurs propres de A non deux à deux distinctes et telles que $|\lambda_1| = \lambda_1 = 1 > |\lambda_2| \geqslant ... \geqslant |\lambda_n|$
$\exists Y_2, ..., Y_n \in (\mathbb{R}^n)^{n-1}$ des vecteurs propres de A associés respectivement aux valeurs propres $\lambda_2, ..., \lambda_n$ et $\gamma \in \mathbb{R}$ tel que :

$$X = \gamma R + \sum_{i=2}^{n} Y_i$$

On a alors $\forall k \in \mathbb{N}$:

$$A^k X_0 = \gamma R + \sum_{i=2}^{n} \lambda_i^k Y_i.$$

Or $\forall i \in [|2, n|]\ \ |\lambda_i| < 1$ d'où $\sum_{i=2}^{n} \lambda_i^k Y_i$ converge vers 0 en $\mathcal{O}(|\lambda_2|^k)$ Ainsi :

$$\lim_{k \longrightarrow +\infty} A^k X_0 = \gamma R$$

D'après le complément sur les matrices stochastiques cette limite est un vecteur stochastique postif donc $\gamma = 1$ D'où la convergence

Sinon, dans le cas général :
On pose : $Sp_{\mathbb{C}}(A) = \{1, \lambda_2^*, \dots, \lambda_p^*\}$ où les λ_i^* sont deux à deux distinctes et $\forall i \in [\![1, p]\!]$, m_{λ_i} la multiplicité de λ_i dans χ_A.

Par propriété de cours, il existe $P \in \mathcal{GL}_n(\mathbb{C})$ telle que $A = P^{-1} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & A_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & A_p \end{pmatrix} P$

Où $\forall i \in [\![1, p]\!]$, $A_i = \lambda_i^* I_{m_{\lambda_i}} + N_i$ où $N_i \in \mathcal{M}_{m_{\lambda_i}}(\mathbb{K})$ nilpotente dont on note n_i son indice de nilpotence.

On remarque que $\forall k \in \mathbb{N}$, $M^k = P^{-1} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & A_2^k & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & A_p^k \end{pmatrix} P$

Soit $i \in [\![2, n]\!]$,

$$\forall k > n_i, A_i^k = \sum_{j=0}^{n_i} \binom{k}{j} (\lambda_i^*)^{k-j} N_i^j$$

Or,

$$\forall j \in [\![0, n_i]\!], \quad \binom{k}{j} (\lambda_i^*)^{k-j} \sim C k^j (\lambda_i^*)^k \text{ où } C = \frac{1}{j!(\lambda_i^*)^j} > 0$$

$\lambda_i^* < 1$, donc par comparaison fonction logarithmique, $\lim\limits_{k \longrightarrow \infty} k^j \lambda_i^k = 0$.
Ainsi $\forall i \in [\![2, b]\!]$, $\lim\limits_{k \longrightarrow \infty} A_i^k = 0_p$

On en déduit donc que $\lim\limits_{k \longrightarrow \infty} A^k = P^{-1} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} P$

Ainsi, la suite $(A^k X_0)_{k \in \mathbb{N}}$ converge vers le projeté de X_0 sur Vect(R).
Donc, $\exists \gamma \in \mathbb{R}$ tel que $\lim\limits_{k \longrightarrow \infty} A^k X_0 = \gamma R$, d'après le complément sur les matrics stochastiques cette limite est un vecteur stochastique strictement positif donc $\gamma = 1$.

$$\boxed{\text{Ainsi } \lim_{k \longrightarrow \infty} A^k X_0 = R}$$

5.2 Application au problème initial

La matrice M est strictement positive et stochastique en colonnes donc pour tout voyage de Max X_0, la suite $(M^t X_0)_{t \in \mathbb{N}}$ converge vers R.
Nous avons résolu le problème !

5.3 Une approche algorithmique

Il ne reste maintenant plus qu'à programmer un algorithme pour calculer $\forall t \in \mathbb{N}$, les $M^t X_0$En voici un exemple naïf.

Algorithme 1 Multiplication matricielle

Entrée: Un matrice stochastique en colonnes $A = (a_{i,j}) \in \mathcal{M}_n(\mathbb{R})$, un vecteur colonne stochastique X et un entier naturel $k \in \mathbb{N}$
Sortie: Le vecteur stochastique $V = A^k X$

```
V ⟵ X
V ⟵ vecteur nul
pour l entre 1 et k faire
    X ⟵ V
    pour i entre 1 et n faire
        pour j entre 1 et n faire
            V_i ⟵ A_{i,j} X_j + V_i
        fin pour
    fin pour
fin pour
```

La complexité de cet algorithme est en $\mathcal{O}(kn^2)$, elle est polynomiale en les paramètres d'entrée.

6 Une "notation" de l'expression des gènes

Nous allons maintenant adapter l'agorithme précédent pour mesurer l'importance des gènes. Il suffit pour cela de modifier la matrice M défini précédemment et donc d'adapter la marche aléatoire de Max dans le graphe non pondéré $G = (S, E)$ au problème d'importance des gènes.

6.1 Un nouveau modèle

Dans toute la suite $G = (S, E)$ (avec $|S| = n \in \mathbb{N}$, on confondra S et $[\![1, n]\!]$) désignera le graphe défini en partie 2. Pour adapter le problème à l'expression des gènes, nous allons modifier la promenade de Max en incluant de nouveaux paramètres :

— Le "rôle" des gènes

— Les liens que possède les gènes entre eux

Ces paramètres, vont soit modifier la téléportation de Max, soit modifier la marche aléatoire. Nous allons donc proposer une façon d'inclure ces paramètres au problème.

Nous allons pour se faire nous servir du modèle défini dans la première partie pour qualifier les gènes la Gene Ontology. Pour cela nous noterons $\forall i \in S \;\; O_i$ les GO terms dont peut-être décrit i.

Le "rôle" des gènes
Plus un gène possède de GO terms plus ce gène est important puisque il participera d'avantage au fonctions moléculaires de l'individu. Le nombre de GO terms un gène va donc naturellement modifier la téléportation de Max, plus un gène i va possèdait de GO plus Max aura de chance de se retrouver téléporte en i. Ainsi on a :

$$\forall t \in \mathbb{N}, \forall i \in S \;\; \mathrm{P}_{\text{"Téléportation"}}(Y_{t+1} = i | Y_t = j) = \frac{|O_i|}{\sum_{k \in [\![1n]\!]} |O_k|} \text{ notée } EX(i)$$

Les liens entre les gènes
Plus un gène i possède des GO terms en commun avec un gène important j, plus le gène i sera important. Ainsi le nombre de GO terms en commun entre les gène i et j va modifier la marche aléatoire de Max.
Nous définissons alors ω tel que :

$$\forall (i, j) \in S^2 \;\; \omega(i, j) = \frac{|O_i \bigcap O_j|}{|O_i||O_j|}$$

Remarque, $\omega(i, j) = 0$ si i et j ne possèdent aucun lien en commun.
Dès lors,

$$\forall t \in \mathbb{N}, \;\; \forall (i, j) \in S^2 \;\; \mathrm{P}_{\text{"Non téléportation"}}(Y_{t+1} = i | Y_t = j) = \frac{\omega(j, i)}{\sum_{k \in [\![1,n]\!]} \omega(j, k)} \text{ notée } LI(j, i)$$

6.2 La notation des gènes :

Nous remarquons ainsi comme dans la partie 3 que :

$$\forall k \in \mathbb{N} \quad \forall (i,j) \in [\![1,n]\!]^2 \quad \mathbb{P}(X_{k+1} = i | X_k = j) = \alpha LI(j,i) + (1-\alpha)EX(i) \text{ notée } \epsilon_{i,j}$$

Par analogie avec la Partie 3, on définit alors la matrice $M \in \mathcal{M}_n(\mathbb{R})$ par :

$$\forall (i,j) \in [\![1;n]\!]^2, m_{i,j} = \epsilon_{i,j} = \alpha \times LI(j,i) + (1-\alpha)EX(i)$$

La matrice est strictement positive et stochastiques en colonnes. En effet,

$$\forall j \in [\![1,n]\!] \sum_{i \in [\![1,n]\!]} m_{i,j} = \alpha \sum_{i \in [\![1,n]\!]} \frac{\omega(j,i)}{\sum_{k \in [\![1,n]\!]} \omega(j,k)} + (1-\alpha) \sum_{i \in [\![1,n]\!]} \frac{|O_i|}{\sum_{k \in [\![1n]\!]} |O_k|} = \alpha + (1-\alpha) = 1$$

Ainsi on peut appliquer l'algorithme 1. à cette nouvelle matrice M pour répondre au problème.

6.3 Une simulation

Pour tester mon algorithme j'ai simulé en Caml des graphes de gènes hypothétiques à partir du modèle de "notation" de gènes définis plus haut. J'ai ensuite tester mon algorithme pour des tailles d'ensemble de gènes et des nombres d'itérations. Les algortihmes utilisés sont en annexe J'ai alors eu les résultats en moyenne suivant pour 10 essais.

- Réalisation en Caml d'un réseau de 100 gènes pour 1 000 itérations : $\boxed{\text{Temps} \approx 8.7 \text{ s}}$
- Réalisation en Caml d'un réseau de 1000 gènes pour 100 itérations : $\boxed{\text{Temps} \approx 93.6 \text{ s}}$
- Réalisation en Caml d'un réseau de 1000 gènes pour 1000 itérations : $\boxed{\text{Temps} \approx 886.3 \text{ s}}$
- Réalisation en Caml d'un réseau de 1000 gènes pour 10000 itérations : $\boxed{\text{Temps} \approx 8765.3 \text{ s}}$

Les résultats sont en accord avec la complexité de la fonction utilisée en $\mathcal{O}(kn^2)$ où k est le nombre d'itération et n le nombre de gènes étudiés.
Pour un ensemble de gènes de taille plus importante, les temps de calculs et la puissance nécessaire devenaient trop importants pour mon ordinateur. Ce problème peut être amélioré en proposant un algorithme de calcul matriciel plus efficace que notre algorithme naïf.

7 Conclusion

L'algorithme classement de gènes :

1. Nécessité de tests en laboratoires, tester l'efficacité effective de l'algorithme proposé nécessiterait des études en laboratoires, plus poussées.
2. Notre algorithme nous fournit une liste de gènes prioritaires à étudier. Ces résultats ne peuvent pas encore remplacer les mesures d'expressions des gènes actuels mais il devrait être utilisés comme complément, pour mettre en avant des gènes usuellement peu important mais noté de façon significatives !
3. Un algorithme modifiable, le modèle que j'ai utilisé : Gene Ontology, peut être modifié est remplacé par d'autres sources de données. Il s'agit alors de modifier les paramètres qui veulent être pris en compte par l'étude menée. Le principe reste néanmoins globalement le même.

Annexes

```
(* Simulation réseau de gènes *)

let simulation_genes n =
  let m = Array.make_matrix n n 0. in
  for i = 0 to n-1 do
    m.(i).(i) <-float_of_int ((Random.int 13) +
1 )
  done;
  for i = 0 to n-1 do
    for j = i+1 to n-1 do
      m.(i).(j) <- float_of_int(Random.int (min
(int_of_float m.(i).(i)) (int_of_float m.(j).
(j)) )) ;
      m.(j).(i) <- m.(i).(j)
    done;
  done;
  m
;;

let simulation_x0 n m =
  let x0 = Array.make n 0. in
  let s0 = ref 0. in
  for i = 0 to n-2 do
    x0.(i) <- (float_of_int (Random.int
(int_of_float (m -. (m *. !s0)))))/.m;
    s0 := !s0 +. x0.(i)
  done;
  x0.(n-1) <- 1. -. !s0;
  x0
;;

let reseau_genes g alpha =
  let n = Array.length g in
  let v0 = Array.make n 0. in
  let s0 = Array.make n 0. in
  let m = Array.make_matrix n n 0. in
  let w = Array.make_matrix n n 0. in
  let s = ref 0. in
  for i = 0 to n-1 do
    s := !s +. g.(i).(i) ;
  done;
  for i = 0 to n-1 do
    v0.(i) <- g.(i).(i) /. !s
  done;
  for i = 0 to n-1 do
    for j = 0 to n-1 do
      w.(i).(j) <- g.(i).(j)/.(g.(i).(i)*.g.(j).
```

```
(j))
    done;
  done;
  for i = 0 to n-1 do
    for k = 0 to n-1 do
      s0.(i) <- w.(i).(k) +. s0.(i) ;
    done;
  done;
  for i = 0 to n-1 do
    for j = 0 to n-1 do
      m.(i).(j) <- alpha*.(w.(i).(j)/.s0.(j)) +.
(1. -. alpha)*.v0.(i);
    done;
  done;
  m
;;


(* L'algorithme GeneRank *)

let est_stocha_colonne m =
  let b0 = ref true in
  let n = Array.length m.(0) in
  for j= 0 to n-1 do
    let s = ref 0. in
    for i = 0 to n-1 do
      if m.(i).(j) < 0. then b0 := false;
      s := !s +. m.(i).(j)
    done;
    if !s < 0.99 then b0 := false;
  done;
  !b0
;;

let multi_matrice m x =
  if not (est_stocha_colonne m) then
   failwith "le problème ne peut pas être
résolu";
  let n = Array.length x in
  let x0 = Array.make n 0. in
  for i = 0 to n-1 do
    for j = 0 to n-1 do
      x0.(i) <- x0.(i) +. m.(i).(j)*.x.(j)
    done;
  done;
  x0
;;
```

```
let rec gene_rank m n x0 = match n with
  | 0 -> x0
  | _ -> gene_rank m (n-1) (multi_matrice m x0)
;;


let calcul_temps n m a =
  let t1 = Sys.time () in
  let g = gene_rank (reseau_genes
(simulation_genes n) a) m (simulation_x0 n
1000.) in
  let t2 = Sys.time () in
 t2 -. t1
;;

calcul_temps 100 1000 0.85 ;;

calcul_temps 1000 10000 0.85 ;;
```