**La comparaison de réseaux complexes en biologie : mission impossible ?**

De nos jours de plus en plus de systèmes complexes sont représenté à l’aide de réseaux. La science des réseaux s'applique à tous les domaines, qu'il s'agisse de réseaux sociaux, de réseaux électriques, de réseaux terroristes ou même de réseaux financiers. La biologie n’échappe pas à la mode. Chaque système biologique complexe peut être représenté par un réseau. Les exemples sont nombreux et relèvent de tous les domaines de la biologie: réseaux trophiques, réseaux neuronaux, réseaux d'interactions protéines-protéines ou de régulation géniques, etc. Cependant, quand on compare des grands réseaux, nous devons faire face à plusieurs obstacles tels que la taille du réseau. L’analyse de ces réseaux nous permettrait de mieux comprendre les mécanismes biologiques. Par exemple, on pourrait comparer le réseau épidémiologique de la transmission d'une même maladie à deux moments donnés, ou encore deux réseaux représentants deux maladies différentes au même moment. La comparaison de réseaux épidémiologiques pourrait permettre de développer un plan de contingence de la maladie.

Lors de la comparaison de réseaux il existe trois grandes tendances :

* L’étude des distributions des degrés des nœuds. On s’intéresse alors aux proportions de nœuds ayant un degré k.
* L’étude des longueurs des chemins (séquence de nœuds adjacents). On calcule souvent la moyenne du chemin plus court (nombre d’arrêtes du chemin plus court connectant deux nœuds).
* Etude des structures locales du réseau. Dans ce cas on étudie les motifs, c’est-à-dire à un sous-graphe d’un réseau plus grand, composé d’un sous ensemble de nœuds (3 ou 4). Les motifs sont la base pour construire un réseau. La plupart des études concernant les motifs, s’intéressent à la distribution des motifs dans un réseau. Ainsi, on calcule la distribution des motifs qui sont plus récurrents dans les réseaux biologiques par rapport aux réseaux aléatoires

Dès lors, une des approches permettant la comparaison de deux réseaux différents est l’étude des motifs qui les composent. Ce sont les «blocs de construction de réseaux complexes". Dans cette perspective nous nous intéressons plus spécifiquement à la comparaison des motifs de taille 4 qui servent à discriminer la topologie des réseaux. Nous proposons divers indices de similarité pouvant servir à comparer les réseaux définis sur un même ensemble d'objets.

Lors de ce projet on s’est plutôt intéressé à la comparaison de motifs composés de 4 nœuds. On travaille alors sur deux graphes ayant les mêmes nœuds, on choisit les mêmes quatre nœuds dans les deux réseaux puis on compare les motifs formés. Dans cette perspective on pourrait donc choisir plusieurs ensembles de 4 nœuds aléatoires pour pouvoir comparer la plus grande quantité de motifs présents. Le défi est comment comparer ces motifs. Il semble facile de déterminer que deux motifs ne sont pas égaux juste en regardant leur forme géométrique, mais il beaucoup plus complexe de savoir quel est leur degré de similarité. Pour cela il est nécessaire de préciser ou de créer des indices qui nous donnent un pourcentage de similarité entre les motifs.

Ainsi, notre projet consiste à créer un programme sur la plateforme R qui nous permette de sélectionner des ensembles de 4 nœuds dans deux graphes puis qui calcule les indices de similarité pour nous dire à quel point les deux motifs sont semblables ou différents. Pour cela on procède par différentes étapes. Tout d’abord, on doit définir quels sont les possibles indices qui vont nous permettre de comparer nos motifs. Une fois les indices sélectionnés, il faut passer par une étape de programmation dans laquelle on crée un programme qui nous permette de réaliser les calculs des indices pour des motifs sélectionnés aléatoirement. Une fois le programme créé il faut alors réaliser une étape de tests. Il s’agit ici d’utiliser des réseaux aléatoires, et de choisir des motifs pour calculer les indices. Avec ces résultats on pourra effectuer des tests statistiques et trier nos indices. Cela permet d’identifier les indices qui nous donnent la même information et les indices qui sont les plus discriminants pour la comparaison.

**METHODES**

Les réseaux biologiques peuvent être représentés par des graphes dans lesquels les nœuds correspondent aux composants et les arrêtes correspondent aux interactions. Il existe deux types de graphes : les graphes orientés et les graphes non orientés. Un graphe orienté permet de représenter des interactions comme celle de la régulation génique où un gène contrôle un autre, tandis que les graphes non orientés permettent de représenter des réseaux non dirigés comme par exemple le microbiome humain. Certaines mesures permettent de caractériser les nœuds et l’ensemble du réseau : le degré d’un nœud est le nombre d’arrêtes qui sont connectées au nœud.

Lors de cette étude on s’intéresse particulièrement à des graphes non orientés. Il faut ensuite déterminer le niveau d’information que les nœuds vont avoir. On travaille avec trois niveaux d’information. Le premier niveau correspond à aucune information sur nos nœuds. Le deuxième niveau correspond à une information bi-couleur, par exemple si on avait un réseau social les femmes pourraient être représentées par des nœuds roses et les hommes par des nœuds bleus. Et un dernier niveau serait les étiquettes avec les nœuds, dans ce cas on aurait les noms des personnes dans chaque nœud pour le réseau social.

Dans une première étape on s’est concentré sur la création d’indices de similarité pour comparer nos motifs. En fonction de chaque niveau d’information, on crée des indices permettant de comparer nos motifs. Il y a des indices qui peuvent être utilisés pour les trois niveaux et des indices qui sont propres à chaque niveau. Il est important de distinguer deux types d’indices en fonction de l’information qu’ils apportent. Tout d’abord, il y a des indices qui nous donnent un résultat binaire, c’est-à-dire vrai lorsque les motifs sont identiques et faux lorsqu’ils sont différents. Ensuite, on a des indices qui donnent un résultat numérique. Ainsi, pour ces derniers indices on calcule l’indice pour le motif 1 puis pour le motif 2 puis on réalise une soustraction qui reflète la différence entre les deux motifs : plus le résultat est grand plus les motifs sont différents. Un des indices à résultat binaire qu’on peut utiliser est l’isomorphisme. Il compare la forme géométrique des deux motifs, s’ils ont la même forme alors les motifs sont considérés identiques (vrai) si la forme n’est pas la même on va avoir un résultat faux indiquant que les motifs ne sont pas identiques.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nom de l'indice** | **Description** | **Indice Motif 1** | **Indice Motif 2** | **Différence entre les 2 indices des motifs 1 et 2** |
| **Isomorphisme** | Comparaison de la forme géométrique des deux motifs |  |  | FAUX |
| 1. **Nombre arrêtes** | Nombre d'arrêtes dans le motif | 2 | 2 | 0 |
| 1. **Moyenne degrés** | Moyenne des degrés des nœuds | (1+1+1+1)/4 = 1 | (2+1+1+0)/4 = 1 | 0 |
| 1. **Degré minimum** | Degré minimum | 1 | 0 | 1 |
| 1. **Degré maximum** | Degré maximum | 1 | 2 | 1 |
| 1. **Différence maximum-minimum** | Différence entre le degré maximum et le degré minimum | 0 | 2 | 2 |
| 1. **Distribution degrés** | Vecteur avec un premier élément qui est le nombre de nœuds ayant un degré 0, c’est-à-dire aucun arrêt. Le deuxième élément est le nombre de nœuds ayant un degré 1 et ainsi de suite. | (0,4,0,0) | (1,2,1,0) | (1,2,1,0) 4 |
| 1. **Diamètre** | Distance maximale entre les nœuds du graphe | 1 | 2 | 1 |
| 1. **Moyenne degrés couleurs** | Vecteur avec un premier élément qui correspond à la moyenne des degrés des nœuds couleur 1 et le deuxième élément correspond à la moyenne des degrés des nœuds couleur 2 | (1 ,1) | (1.5, 0.5) | (0.5, 0.5) 1 |
| 1. **Degré minimum couleurs** | Vecteur avec un premier élément qui correspond au degré minimum des nœuds couleur 1 et le deuxième élément correspond au degré minimum des nœuds couleur 2 | (1 ,1) | (1, 0) | (0, 1) 1 |
| 1. **Degré maximum couleurs** | Vecteur avec un premier élément qui correspond au degré maximum des nœuds couleur 1 et le deuxième élément correspond au degré maximum des nœuds couleur 2 | (1 ,1) | (2, 1) | (1, 0) 1 |
| 1. **Arrêtes C1C1** | Nombre d'arrêtes entre des nœuds de couleur C1 | 1 | 1 | 0 |
| 1. **Arrêtes C2C2** | Nombre d'arrêtes entre des nœuds de couleur C2 | 1 | 0 | 1 |
| 1. **Arrêtes C1C2** | Nombre d'arrêtes entre un nœud C1 et un nœud C2 | 0 | 1 | 1 |
| 1. **Matrice d'adjacence** | Matrice dont l’élément Aij est égal à 0 si entre le nœud i et le nœud j il n’existe pas d’arrête, et Aij est égal à 1 lorsqu’il existe une arrête entre ces deux nœuds | . 1 2 3 4 1 0 0 1 0 2 0 0 0 1 3 1 0 0 0  4 0 1 0 0 | . 1 2 3 4 1 0 0 1 1 2 0 0 0 0 3 1 0 0 0  4 1 0 0 0 | . 1 2 3 4 1 0 0 0 1 2 0 0 0 1 3 0 0 0 0  4 1 1 0 0 |
| 1. **Matrice des chemins** | Matrice dont l’élément Aij est égal au nombre d'arrêtes du chemin le plus court entre le nœud i et le nœud j | . 1 2 3 4 1 0 0 1 0 2 0 0 0 1 3 1 0 0 0  4 0 1 0 0 | . 1 2 3 4 1 0 0 1 1 2 0 0 0 0 3 1 0 0 2  4 1 0 2 0 | . 1 2 3 4 1 0 0 0 1 2 0 0 0 1 3 0 0 0 2  4 1 1 2 0 |
| 1. **Distribution degrés étiquettes** | Vecteur avec un premier élément qui est le degré du nœud 1. Le deuxième élément est le degré du nœud 2 et ainsi de suite. | (1,1,1,1) | (2,0,1,1) | (1,1,0,0) 2 |

Ainsi, si on pense au plus basique le premier indice qu’on pourrait imaginer est tout simplement le nombre d’arrêtes dans le motif. Cependant, on constate tout de suite que ce type d’indice ne permet pas de différencier tous les motifs. Il est alors intéressant d’avoir plusieurs indices. Une autre caractéristique des réseaux est le degré d’un nœud, on peut donc en utilisant cette mesure créé des indices. Par exemple on peut calculer la moyenne des degrés des nœuds, mais là encore cet indice ne permet pas de différencier certains motifs. On peut aussi créer un vecteur avec la distribution des degrés. Dans ce cas le premier élément du vecteur est le nombre de nœuds ayant un degré 0, c’est-à-dire aucun arrêt. Le deuxième élément est le nombre de nœuds ayant un degré 1 et ainsi de suite. On détermine la distribution des degrés des deux motifs puis on réalise une soustraction membre à membre. Cependant, même si ces indices peuvent être calculés sur des motifs des trois niveaux d’information on constate qu’ils ne sont pas toujours pertinents. D’où l’intérêt à déterminer des indices pour chaque niveau d’information. Pour le deuxième niveau d’information il faut tenir en compte les couleurs des nœuds. Dans ce cas on peut utiliser des indices qu’on a déjà déterminés en ajoutant cette information. On peut tout simplement calculer le nombre d’arrêtes qui relient deux nœuds de même couleur et le nombre d’arrêtes qui relient deux nœuds de couleur différente. On peut calculer aussi la moyenne des degrés de nœuds de chaque couleur. On crée un vecteur avec un premier élément qui correspond à la moyenne des degrés des nœuds couleur 1 et le deuxième élément correspond à la moyenne des degrés des nœuds couleur 2. Pour le troisième niveau d’information dans lequel on a les étiquettes pour chaque nœud on distingue deux indices basés sur des matrices. Le premier est ce qu’on appelle la matrice d’adjacence. Dans cette matrice chaque élément peut prendre deux valeurs soit 1 soit 0. Ainsi, l’élément Aij est égal à 0 si entre le nœud i et le nœud j il n’existe pas d’arrête, et Aij est égal à 1 lorsqu’il existe une arrête entre ces deux nœuds. Etant donné qu’on travaille avec des graphes non dirigés la matrice d’adjacence est symétrique par rapport à sa diagonale. Pour les calculs on n’utilise alors que le triangle inférieur de la matrice. Le deuxième indice suit le même principe que la matrice d’adjacence mais plutôt que regarder l’absence ou présence d’une arrête entre deux nœuds on s’intéresse au chemin le plus court entre les deux. L’élément Aij est alors le nombre d’arrêtes du chemin le plus court entre le nœud i et le nœud j.

Une fois les indices ont été déterminés, on se concentre sur la création d’un programme sur R qui nous permet de réaliser ces calculs de façon automatisé. Ainsi, on utilise le package igraph de R qui nous permet de traiter des graphes. Certains des indices qui ont été mentionnés au-dessus sont déjà accessibles sur igraph. Le programme permet alors à partir de deux graphes ayant les mêmes nœuds de choisir k ensembles de 4 nœuds aléatoires et de calculer les indices de similarité pour les paires de motifs.

Une fois le programme terminé, l’étape suivante est de calculer les indices de similarité à partir de graphes aléatoires. On réalise alors plusieurs simulations. On crée des graphes sur R à partir du modèle Erdos Renyi et Watts Strogatz. On fait varier le nombre de nœuds. Pour les graphes qui suivent le modèle Erdos Renyi on fait varier aussi la probabilité d’avoir une arrête dans un nœud : 0.2, 0.5 et 0.7. Puis à partir de ces réseaux on sélectionne aléatoirement 1000 ensemble de 4 nœuds dans deux graphes différents (mais avec les mêmes caractéristiques). Puis on calcule les indices pour les trois niveaux d’information. Les résultats sont sauvegardés dans une data frame. On réalise ainsi 19 simulations.

|  |  |  |
| --- | --- | --- |
| **Figure Numéro** | **Nombre de nœuds** | **Probabilité arrêt entre 2 nœuds** |
| 1 | 50 | 0.2 |
| 2 | 50 | 0.5 |
| 3 | 50 | 0.7 |
| 4 | 100 | 0.2 |
| 5 | 100 | 0.5 |
| 6 | 100 | 0.7 |
| 7 | 500 | 0.2 |
| 8 | 500 | 0.5 |
| 9 | 500 | 0.7 |

TABLEAU DE SIMULATIONS AVEC LE MODELE ERDOS RENYI

|  |  |  |  |
| --- | --- | --- | --- |
| **Figure Numéro** | **Nombre de nœuds** | **nei** | **p** |
| 10 | 50 | 10 | 0.8 |
| 11 | 50 | 10 | 0.3 |
| 12 | 50 | 3 | 0.3 |
| 13 | 50 | 3 | 0.8 |
| 14 | 100 | 10 | 0.8 |
| 15 | 100 | 10 | 0.3 |
| 16 | 100 | 3 | 0.3 |
| 17 | 100 | 3 | 0.8 |
| 18 | 500 | 10 | 0.8 |
| 19 | 500 | 10 | 0.3 |

TABLEAU DES SIMULATIONS AVEC LE MODELE WATTS STROGATZ

L’étape suivante à partir de ces résultats est d’analyser nos indices avec une analyse en composantes principales. Cela nous permet de visualiser les corrélations entre nos indices. De plus ces résultats peuvent nous permettre d’observer s’il y a un groupe d’indices qui donnent la même information. On peut ainsi déterminer quels sont les indices les plus pertinents et quels sont les indices qui peuvent être regroupés. Ces analyses nous permettent d’avoir des indices ou des groupes indices qui vont être discriminants et permettre de déterminer les différences entre nos motifs.

**RESULTATS ET DISCUSSIONS**

Les analyses en composantes principales nous permettent d’avoir une représentation en deux dimensions des corrélations entre les indices calculés lors de nos simulations. Dans ce type d’analyse l’angle de séparation entre les vecteurs met en évidence la corrélation entre deux indices, plus proches sont les vecteurs plus corrélés sont les indices. A partir de nos résultats plusieurs observations peuvent être faites.

À première vue on n’observe pas un regroupement des indices en fonction du niveau d’information. On ne retrouve pas non plus un indice qui diffère visiblement de tous les autres dans toutes les conditions. Même si les indices ne sont pas tous corrélés ils ne semblent pas nous donner des informations opposées, les vecteurs ne sont pas à 180 degrés. On constate qu’il y a des figures dans lesquelles on retrouve des groupes d’indices fortement corrélés et des figures où cette corrélation semble moins évidente.

On constate tout d’abord deux indices qui sont toujours fortement corrélés dans toutes nos simulations : l’indice 1 et l’indice 2. Ces deux indices nous donnent alors toujours la même information. On pourrait alors se dire qu’il est suffisant d’utiliser un seul des indices plutôt que les deux. L’indice 1 étant plus simple à calculer il serait plus pertinent de le préférer par rapport à l’indice 2. Un troisième indice est aussi souvent corrélé avec le nombre d’arrêtes et la moyenne des degrés, il s’agit de l’indice 9. Ces trois indices correspondent à un premier groupe d’indices qui nous donnent la même information.

On remarque aussi que les indices 14 et 15 qui tiennent en compte les étiquettes des nœuds, sont aussi souvent corrélés. Cependant il existe des exceptions, dans la figure 5 on observe que les vecteurs de ces deux indices sont éloignés. Ici encore on se retrouve avec un ensemble d’indices qui nous donnent la même information. Dans ce cas, pour les indices des nœuds avec étiquettes on pourrait choisir l’indice 16 (moins corrélé avec les indices 14 et 15) et soit l’indice 15 soit l’indice 14. Etant donné que la matrice des chemins donne comme résultat certaines fois l’infini, la matrice d’adjacence est plus simple à manipuler, il serait alors plus pratique de choisir cet indice.

En outre, on remarque aussi que dans plusieurs figures les vecteurs des indices 3 et 5 sont ceux qui sont plus éloignés. C’est-à-dire que ces indices ne sont presque pas corrélés. Ils nous donnent alors des informations différentes. De plus, dans certaines figures (19, 18) l’indice 3 est le seul qui se différencie par rapport aux restes des indices qui semblent être corrélés. Cet indice nous ramène donc une information que les autres indices ne nous donnent pas.

On observe qu’il existe un groupe d’indices qui sont dans la plupart des cas corrélés les uns avec les autres ce sont les indices 1, 2, 9, 8, 11 et certaines fois le 6 et le 13. On peut ainsi déterminer un groupe d’indices qui seraient corrélés et donnent la même information. Dans ce cas il n’est pas intéressant de calculer tous les indices puis qu’ils nous donnent la même information.