M1 - BIM

Algorithmes sur les arbres et les graphes en bioinformatique

Réseaux biologiques et détection de motifs

Alessandra Carbone
Université Pierre et Marie Curie

Motifs, modules et réseaux hiérarchiques

•**Les motifs dans les réseaux** peuvent être définis comme des motifs d'interaction recourrant dans plusieurs parties différentes du réseau avec des fréquences plus élevées de ce que l'on trouve dans des réseaux aléatoires.

• Ces motifs sont probablement des modules fonctionnels représentant les schémas opératifs de la cellule.

• Chaque réseau réel est caractérisé par son propre ensemble de motifs distingués.

• les motifs montrent un fort degré de conservation entre espèces différentes.

• observations empiriques ont indiqué que l'agrégation de motifs amène à des larges clusters de motifs qui chevauchent et qui ne sont plus séparables.

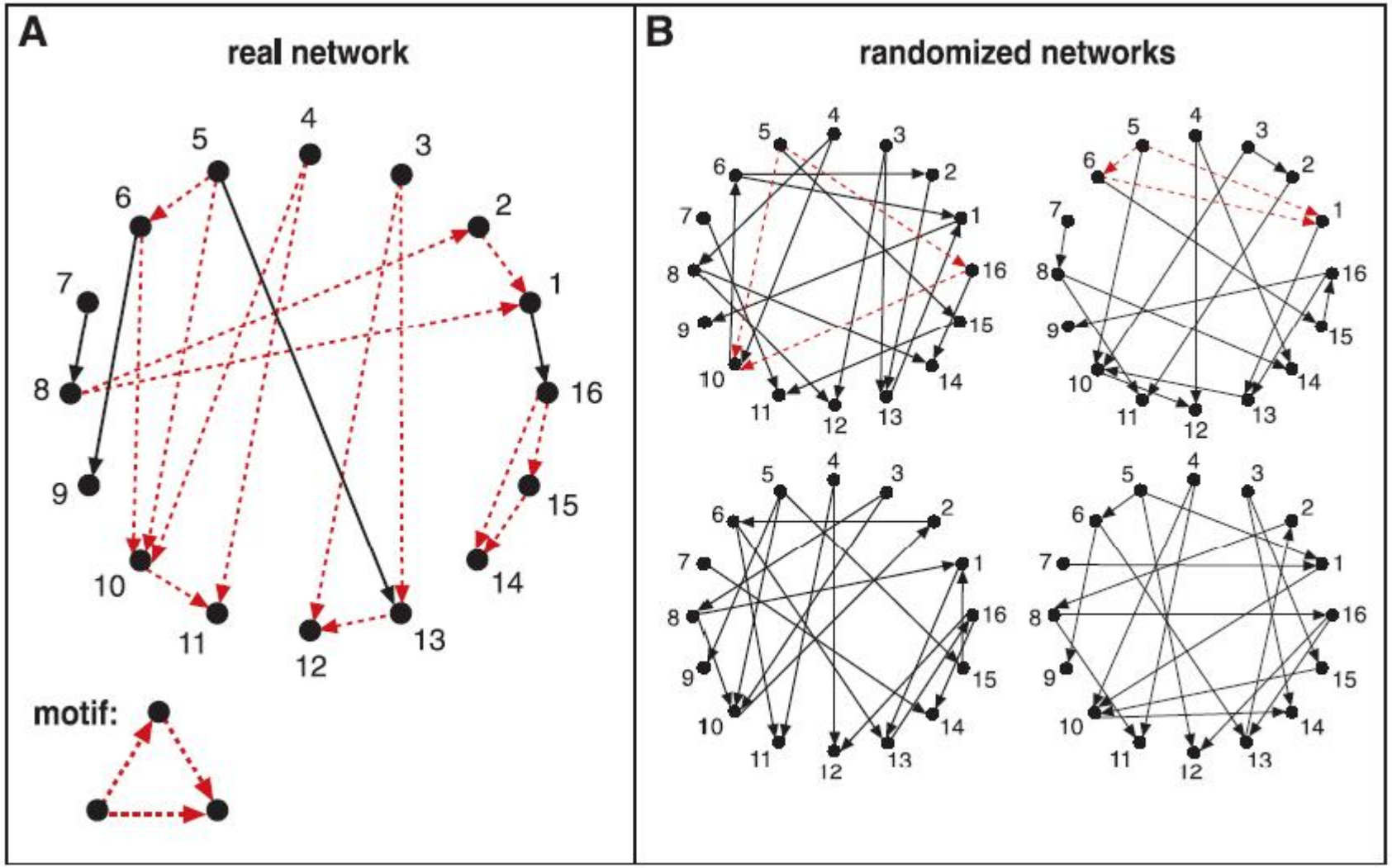
• Est-ce qu'ils existent des **modules fonctionnels plus compliqués** qui pourrons être détectés à l'aide d'algorithmes de clusterisation pour les différent modèles envisagés ?

• C'est possible que les fonctions biologiques soient réalisées de façon modulaire et que « forte clusterisation => « forte modularité »

Qu’est-ce que un “motif dans un reseau” ?

Les motifs dans un réseau sont définis comme des motifs d’interconnection qui sont présent dans plusieurs parties différentes d’un réseau à des fréquences plus grandes que dans les réseaux aléatoires.

Schéma de détection des motifs sur les réseaux



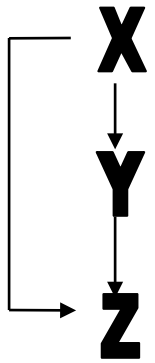
Identification de motifs

- La recherche exhaustive de sous-graphes est combinatoirement non faisable
- Alternative 1 – identification d'un groupe de noeuds fortement connectés et corrélation de cette valeur au rôle fonctionnel potentiel.
- Alternative 2 – regarder autour d'un noeud ayant un degré bas.
- Alternative 3 – méthodes de clustering – homogénéité et séparation

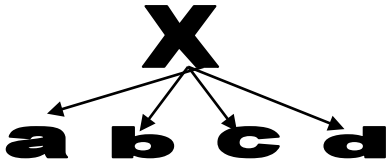
- Problèmes avec les méthodes de clustering
  - Différentes méthodes prédisent les frontières entre modules qui ne sont pas clairement séparés
  - En changeant un paramètre interne dans la méthode on peut se retrouver avec des modules plus larges ou plus petits.

Exemples de motifs

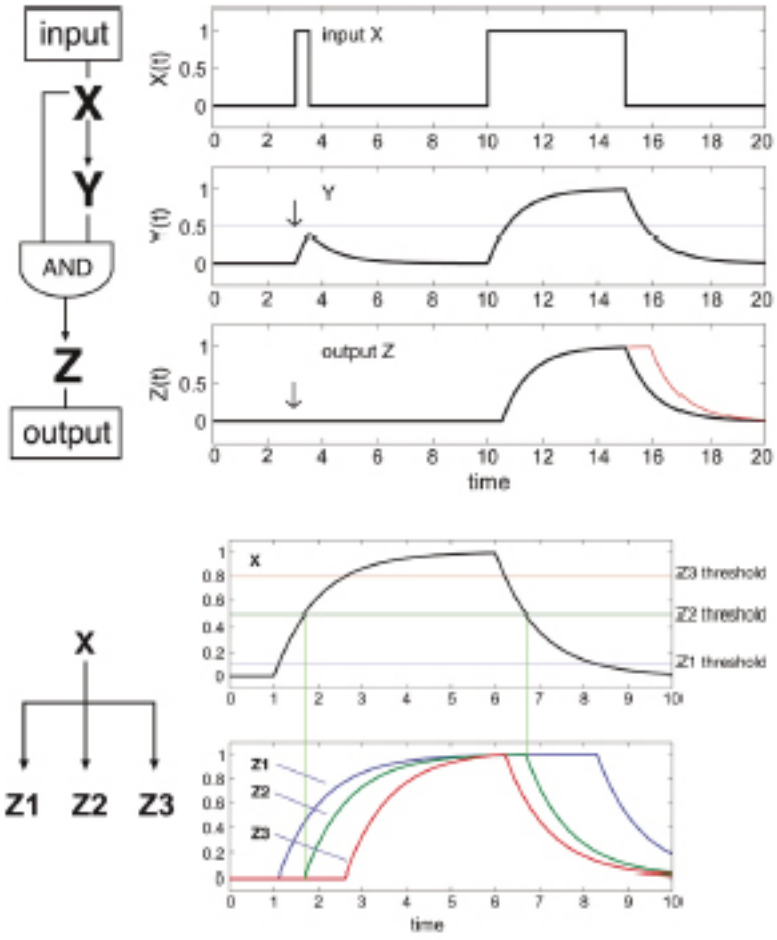
- FeedForward Loop
  - Un facteur agit sur un autre et les deux ensembles sur un troisième.
  - Le plus souvent se comporte comme un **et** logique.
  - 40 occurrences dans le réseau de régulation de *E.coli*, au lieu de 4, 4 ± 3 dans un graphe aléatoire.
  - Il se trouve très représenté dans les réseaux de neurones et il semble être utilisé pour neutraliser le “bruit biologique”.



- Single-Input Module
  - Un seul facteur agit seul sur plusieurs gènes et sur lui-même.
  - Il se trouve très représenté dans les réseaux de régulation des gènes : 68 occurrences dans le réseau de régulation de *E.coli*, au lieu de 28 ± 7 dans un graphe aléatoire.

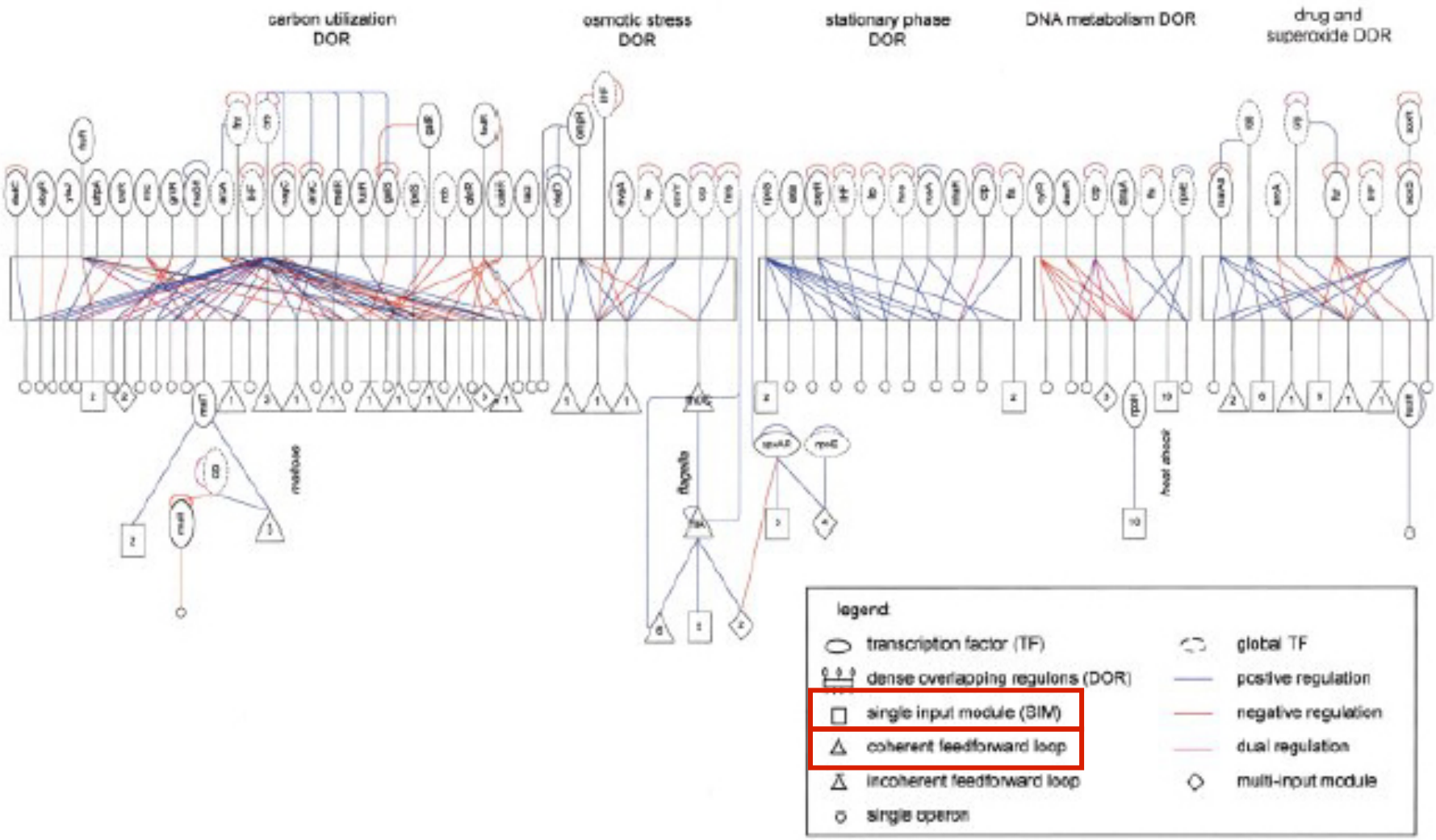


Signification dynamique dans le réseau



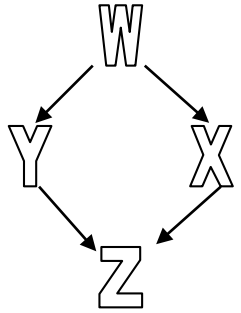
- Le gène ne s'exprime que si le premier facteur est present assez longtemps.
- Le differentes proteines se produisent presque au même moment

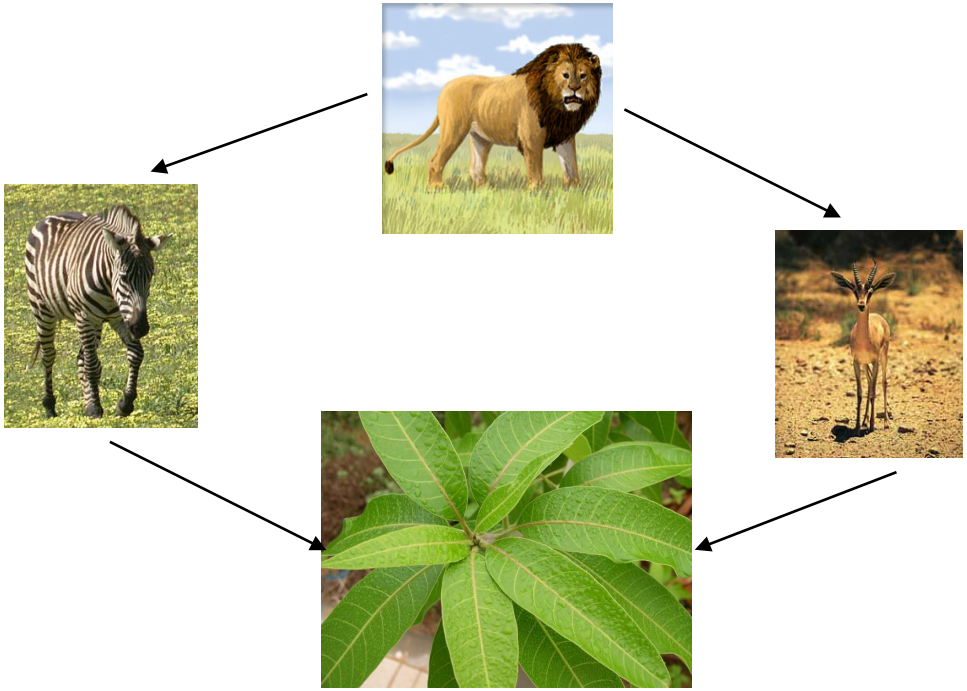
Exemple: réseau de régulation de *Ecoli*



- Parallel paths

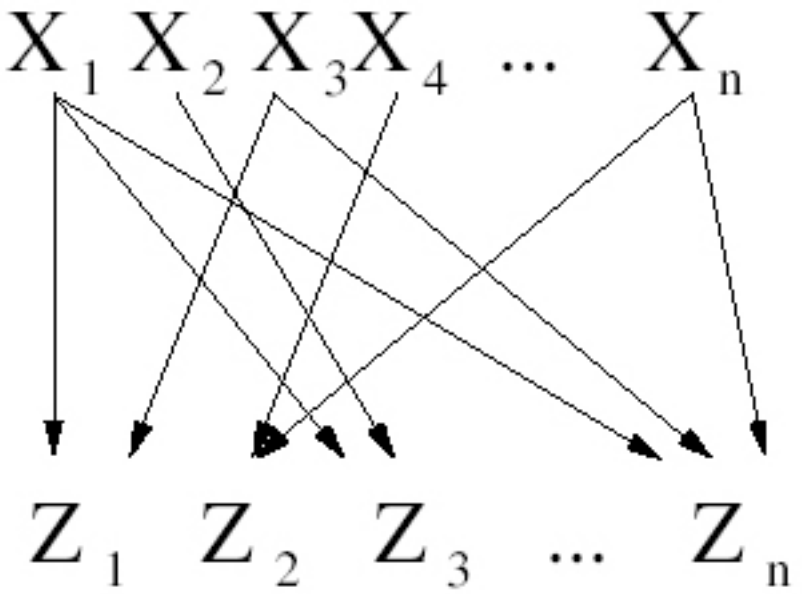
On le trouve très représenté dans les réseaux des neurones (pas trop représenté dans les réseaux de gènes)





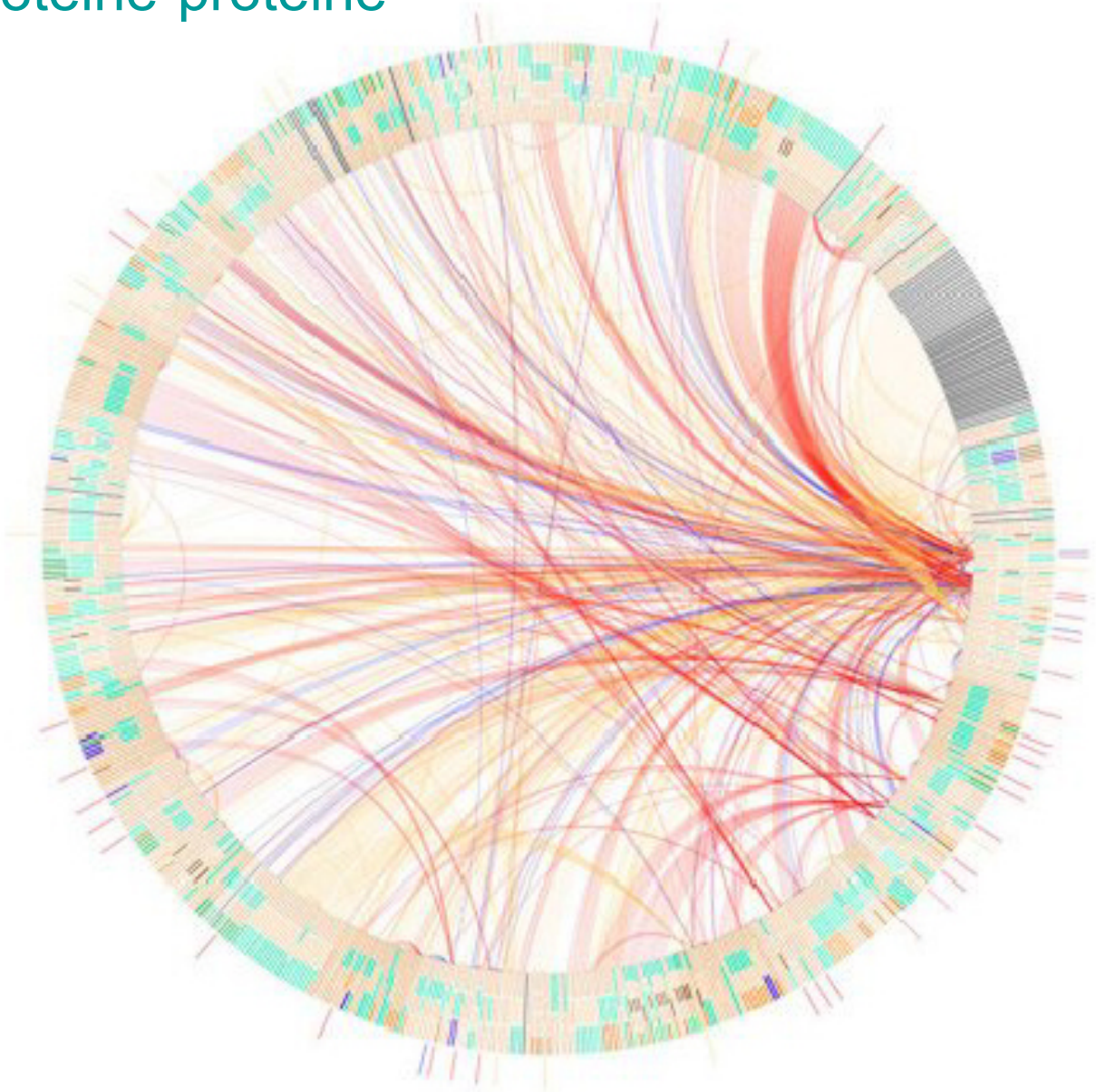
Food web

Dense overlapping regulons : ensemble de motifs



- Le graphe est decomposé en régions denses.
- Les connections entre regions denses sont peu nombreuses.
- Chaque région correspond à une fonctionalité biologique.

Réseau protéine-protéine
Ecoli



Relation fonction-structure dans les réseaux PPI

(On suppose de connaître la classification fonctionnelle des protéines)

- Classes fonctionnelles de protéines distinctes ont des propriétés du réseau différentes. Par exemple : les protéines participantes à la traduction présentent un plus fort degré de connectivité moyen que les protéines de transport, et les protéines senseurs présentent le degré de connectivite le plus bas.

- Entre tous les groupes fonctionnels, les protéines participantes à l'organisation cellulaire ont la présence la plus élevée dans les hubs et leurs suppression déconnecte le réseau PPI.

Algorithme de détection de petits motifs

Entrées : M = la matrice d'adjacence du graphe,
n = la taille des motifs à chercher.

Sortie : La liste des motifs surreprésentés dans le graphe.

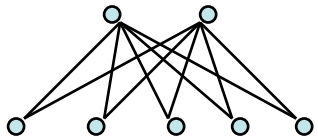
1. Parcourir l'ensemble des sous-matrices n × n de M et compter le nombre d'occurrences de chaque classe de sous-matrices.

2. Générer un grand nombre de graphes aléatoires de même taille et de même distribution de degrés que le graphe initial.

3. Pour chacun des graphes aléatoires compter les occurrences des motifs, et en déduire les espérances et écarts types.

4. Choisir les motifs qui apparaissent dans M significativement plus souvent que dans les graphes aléatoires.

5. Détecter les lignes identiques dans la matrice d'adjacence.



6. Calculer une distance entre chaque paire de facteurs et en déduire les regroupements.

Améliorations possibles de l'algorithme

Prendre un échantillon de sous-graphes au lieu de tous les calculer.

Améliorer la qualité de l'échantillonnage.

Créer des graphes aléatoires qui ont plus de caractéristiques en commun avec le graphe initial.

Construire la liste des motifs progressivement, en ne gardant à chaque étape que les motifs les plus fréquents.

Outils existants:

mfinder (2002-2005)
FANMOD (2005)
MAVisto (2005)

Graph mining problem - techniques

Frequent and closed subgraph mining methods

gSpan and CloseGraph: pattern-growth depth-first search approach

Graph indexing techniques

Frequent and discriminative subgraphs are high-quality indexing features

Similarity search in graph databases

Indexing and feature-based matching

Regarder les références à la fin de la présentation