M2BI :
UE Génomique
UE Méthodes en bioinformatique

Support de cours

Plan du cours:

1- Introduction : la biologie des systèmes
2- Réseaux biologiques
3- Bases de données
4- Visualisation et analyse avec Cytoscape

Références:

Réseaux : Nat Rev Genet. 2004 Feb;5(2):101-113

1- Introduction

a. La biologie des systèmes

La biologie des systèmes est basée sur l'étude de la biologie de façon interdisciplinaire, qui se concentre sur *les interactions complexes des systèmes biologiques* utilisant une approche plus holistique que réductionniste (traditionnelle).

Elle est utilisée dans divers domaines tels que la recherche biomédicale.

Un des buts de la biologie des systèmes est de créer des modèles et de découvrir de nouvelles propriétés émergentes.

b. Pourquoi avons nous besoin de réseaux?

Les molécules du système vivant ne fonctionnent pas seules … mais elles interagissent en complexes (réseaux)

2- Réseaux biologiques

a. Les réseaux

Nœuds: sommets: Nodes: ce sont les objets du réseau
Liens: arrêtes: Edges: ce sont les interactions du réseau

b. Terminologie des réseaux/ Théorie des graphes

Graphe orienté (directed) (Interaction ADN-protéine, correspondance email, chaine alimentaire) vs non orienté (Interaction protéine-protéine, amis-facebook)

Liens pondérés (weighted) (Interaction protéine-protéine: score de confiance, Réseaux métaboliques : flux) ou non pondérés.

Noeuds pondérées (attribus) ou non pondérés.

Graphe dense ou épar? En mathématiques, un graphe est dense quand le nombre d'arrêtes est proche du nombre maximal d'arrêtes.
Au contraire, un graphe avec seulement quelques liens, est un graphe clairsemé. La distinction entre les graphes clairsemés et denses est plutôt vague, et dépend du contexte.

Densité d'un graphe :
- le nombre maximal de connections entre N protéines est 2N(N-1)
- on définit la densité comme:

$$d = \frac{\text{Nombre de connections}}{\text{Nombre maximale de connections}}$$

Autres notions : self-interaction (ex. homodimère), multigraphs, bipartite graph.

c. Model de réseaux

Matrix ou spoke

d. Propriétés des graphes

Degré des nœuds ou connectivité (k): Node degree or connectivity
- Moyenne des degrés (<k>)

Chemin le plus court: Shortest path
- Moyenne des chemins les plus courtes (<l>)
- Diamètre du réseau

Degré de distribution P(k): Degree distribution
- La probabilité qu'un nœud choisi au hasard a exactement k liens (= degré k)

N = total number of nodes
N_k = number of nodes with degree k
$P(k) = N_k/N$

Coefficient de clustering (C): Clustering coefficient
- Coefficient clustering (C) :

$$C_I = \frac{n_I}{\binom{k}{2}} = \frac{2n_I}{k \cdot (k-1)}$$

k: neighbors of I
n_I: edges between node I's neighbors

- Moyenne des coefficients de clustering (<C>)

C(k): moyenne des coefficients de clustering de tous les nœuds ayant k liens

e. Topologie de réseaux:

Hierarchical, scale-free, random …

f. Réseaux de protéines

Ex. avec les interactions protéine-protéine:
Dans une cellule, les protéines ne fonctionnent pas individuellement mais plutôt en module fonctionnel composé de 2 ou plusieurs protéines.
Ces modules sont mesurés expérimentalement, ce qui permet la création de scores :

- Yeast two-hybrid (Y2H)
- Mass spectrométrie (MS): Protein complex pull down

Ex. de score: interaction binaire (Y2H):

$$S(A,B)_{bin} = -\log_{10}\left((N_A + 1)(N_B + 1)\right)$$

where N_A and N_B are the numbers of non-shared interaction partners for an interaction between protein A and B, see Figure 1.

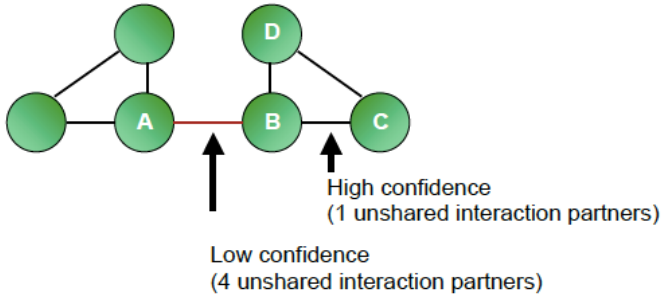
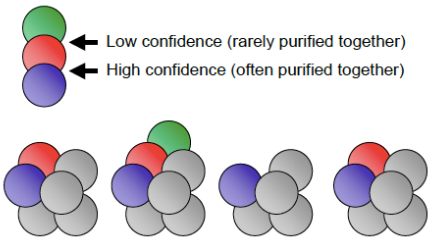


Figure 1: The reliability of a binary interaction has been found to correlate with the number of non-shared interaction partners.

Ex. de score: inferred interaction (MS):



$$S(A,B)_{pul} = \log_{10}\left(\frac{(N_{A\cap B})(N_{A\cup B})}{(N_A+1)(N_B+1)}\right)$$

where:

- $N_{A\cap B}$ is the number of purifications containing both proteins, i.e. the intersection of experiment sets that find them
- $N_{A\cup B}$ is the total number of purifications that find either A or B, i.e. the union of experiments that find them
- N_A is the number of purifications containing A
- N_B is the number of purifications containing B

Note that $N_{A\cap B} \geq 1$, $N_{A\cup B} \geq 1$, $N_A \geq 1$ and $N_B \geq 1$ as we only consider protein pairs that have been detected at least once.

g. Analyse de réseaux

Détection de modules : Beaucoup de méthodes existent, certaines sont implémentées dans Cytoscape:

- MCODE (liens non pondérés, basé sur la topologie)
- ClusterONE (liens pondérés, basé sur la topologie)
- jActiveModules (nœuds pondérés)

3- Bases de données

Création d'un réseau … où trouver des données? → Bases de données

a. Base de données

C'est un ensemble de données relatives à un domaine précis, organisées par traitements informatiques, accessibles en ligne et à distance.

Il existe une multitude de bases de données, générales ou spécifiques. En voici quelques ex. vus en cours (liste non exhaustive) :

- Banques de séquences nucléiques (EMBL/ UniGene-NCBI) : Origine des données: Séquençage d'ADN et d'ARN. Les données stockées : séquences + annotations: Fragments de génomes, plusieurs gènes, séquences intergéniques, génomes complets, ARNm, ARNt, ARNr…

- Banques de séquences protéiques (UniProt/ SwissProt/TrEMBL): Origine des données: Traduction de séquences d'AND, Séquençage de protéines, Protéines dont la structure 3D est connue. Les données stockées : séquences + annotations: Protéines entières, Fragments de protéines.

- Banques de données de puce à ADN (NCBI Gene Expression Omnibus (GEO)/ ArrayExpress (à l'EBI)/ Chemical Effects in Biological Systems (CEBS) Knowledgebase) : Stockage des données d'expériences de puces à ADN.

- Banques bibliographiques (PubMed): Etat de l'art sur les connaissances actuelles

- Bases de données de chemical biologie (Stitch/ChemProt): Information entre les relations chemicals-protéines, mais aussi chemicals-diseases et pathways.

- Autres bases de données: GO (Gene Ontology), KEGG et Réactome (voies de signalisation, voies métaboliques), GeneCards (diseases) …

4- Visualisation et Analyse avec Cytosccape :

a. Cytoscape:

i. Logiciel librement accessible (open-source, java), facilement extensible (Plugins, Apps)
ii. Visualisation des réseaux (réseaux d'interactions moléculaires)
iii. Analyse des réseaux avec des profils d'expression génique et d'autres données d'état cellulaire (GO, protéomique, ...)
iv. Utilisé dans plusieurs centaines d'analyses dans la littérature récente
v. Garantie de continuité

b. Input : tout type de données:

i. Interactions physiques: Interactions protéine-protéine, Interactions protéine-ADN, Interactions métaboliques
ii. Interactions fonctionnelles : Relations de co-expression, Interactions génétiques

c. Formats supportés:

i. file: SIF (simple interaction format
ii. GML (graph markup language)
iii. XGMML (extensible graph markup and modeling language)
iv. BioPax (biological pathway data)
v. URLs, webservices (données disponibles aux téléchargements)
vi. Bases de données publiques (EMBL, NCBI …)

d. Type de cartographie:

i. Continue

Les données continuent sont représentées à l'aide d'attributs visuels continus (e.g. gene expression levels mapped to node color)
Les données continuent sont représentées à l'aide d'attributs visuels discrets (catégories) (e.g. p-value categories mapped to node shape)

ii. Discrète

Les données discrètes sont représentées à l'aide d'attributs visuels discrets (e.g. GO annotation mapped to node shape)
Les données discrètes sont représentées à l'aide d'attributs visuels continus (e.g. multiple GO terms mapped to pie coloring)

e. Analyse de réseaux dans Cytoscape:

   i. Analyses basiques : Analyze network

   ii. Analyses complexes : Plugins ou Apps : BINGO, MCODE, Netmatch, jActiveModules
   More on: http://apps.cytoscape.org/apps/all

f. Workflow:

1- Load data/network: Importer des données pour créer un réseau, ou importer un réseau
2- Load attributes: importer des informations sur les nœuds/liens
3- Visualize: modifier la visualisation du réseau (layout, shape, color, etc…)
4- Analyze: analyser votre réseau (apps)
5- Exporting output: créer des figures