Prédiction des interactions protéines-protéines

"Predicting overlapping protein complexes from weighted protein interaction graphs by gradually expanding dense neighborhoods"

Dimitrakopoulos, C. *et al., Artificial Intelligence in Medicine,* 71 (2016) 62–69

Cours INF7841

Présenté au professeur Éric Beaudry

Par Valérie Hay

Le 09 Novembre 2017

Agenda

- Le problème
- Les outils disponibles
- GENA
- Exemple appliqué
- Conclusion

Les interactions: difficile à étudier

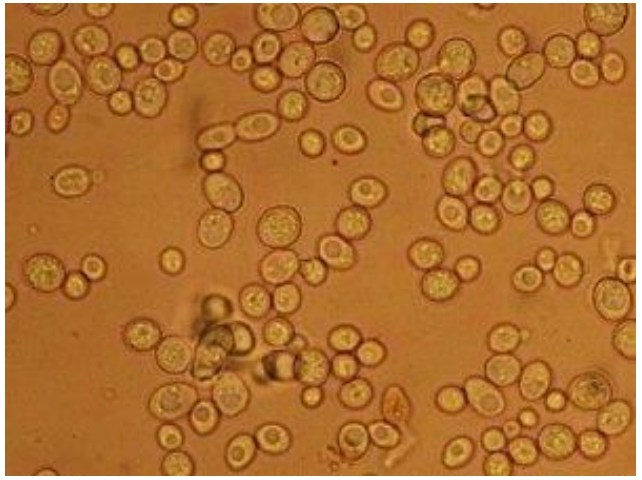
- Interactions en biologie:
  - Prédateur – proie: renard et la perdrix
  - Symbiose: légumes et le rhizobium
  - Cellule – cellule: les organes
  - Pathogènes (bactérie/virus) – hôte : *prédiction pour vaccination annuel de la grippe*
  - **Protéine – protéine :** comment le Parkinson se développe

Le défi: étude des interactions P-P

- Stabilité des complexes souvent momentanée
- Accessibilité des interactions protéiques
- Études de type “*snapshot*”
- Expérience laborieuses et coûteuses
- Expertise: personnel + équipement

Le défit des données actuelles

- Incomplètes, fragmentées
- Beaucoup de faux positifs et faux négatifs



http://www.daviddarling.info/encyclopedia/Y/yeast.html

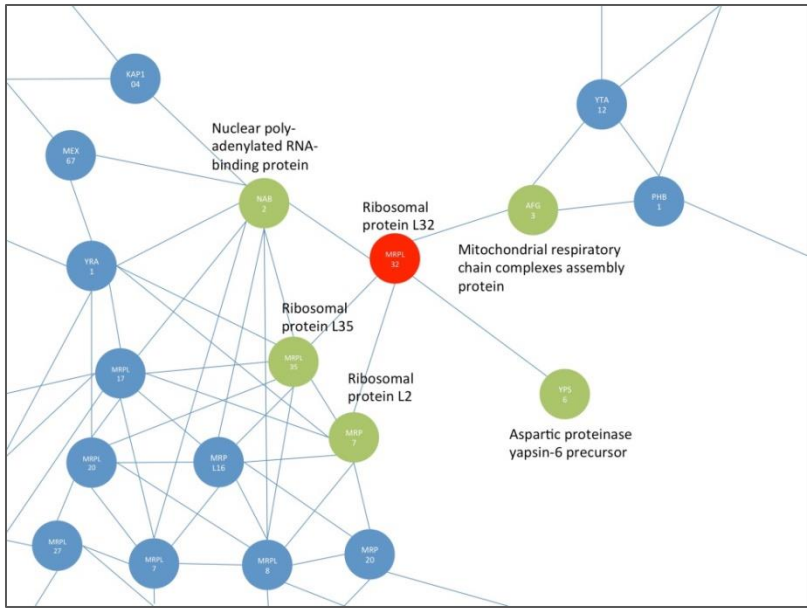
Organismes modèles:

- Homme
- La levure *Saccharomyces cerevisiae* (modèle et technique établis: double hybride)

Visualisation des interactions P-P

Données entrées ⇒ supposent interactions basées sur:

- Localisation spatiale
- Propriété physico-chimique
- Gènes co-exprimés



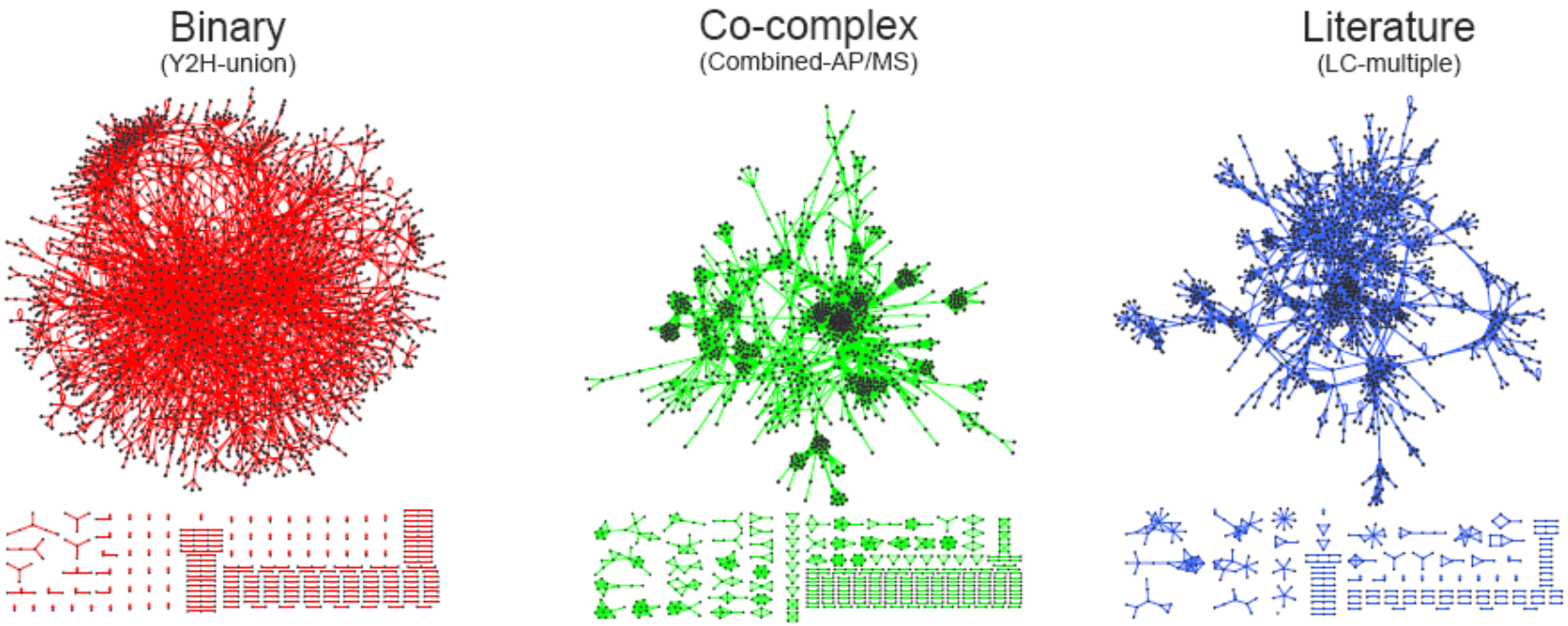
But: trouver les relations = réseaux

Ressembles aux interactions de:

Réseaux sociaux, système planétaire, route, acheteurs

https://phys.org/news/2013-03-insights-evolution-protein-networks.html

Interactome de la levure



Source: http://interactome.dfci.harvard.edu/S_cerevisiae/

Spécificité des interactions P-P

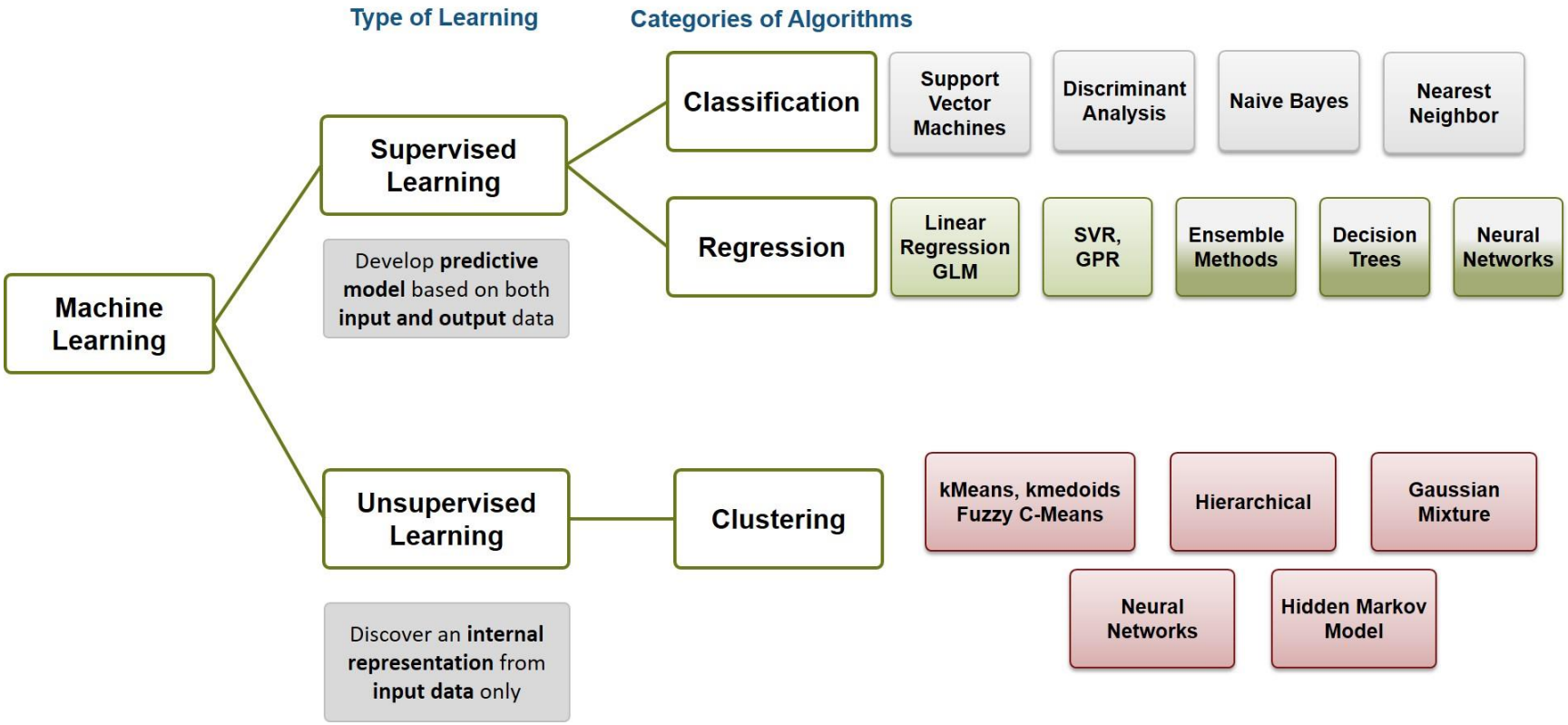
- 1 nœud = 1 protéine
- 1 vertex = 1 interaction non-directionelle
- Les nœuds (protéines) interagissent avec un certain degré de ‘force’ ou une probabilité
- Chaque nœud peut avoir plusieurs vertex
- Chaque groupe peut interagir avec d’autres groupes dans le réseaux

Article: nouvelle méthode

1. Restricted neighborhood search (RNSC)
2. Markov clustering (MCL)
3. ClusterONE
4. **GENA: Gradually Expanding Neighborhoods with Adjustment**

Première étape de ces algorithmes sont du regroupement (*clustering)* non-supervisé.

Regroupement (*clustering*)

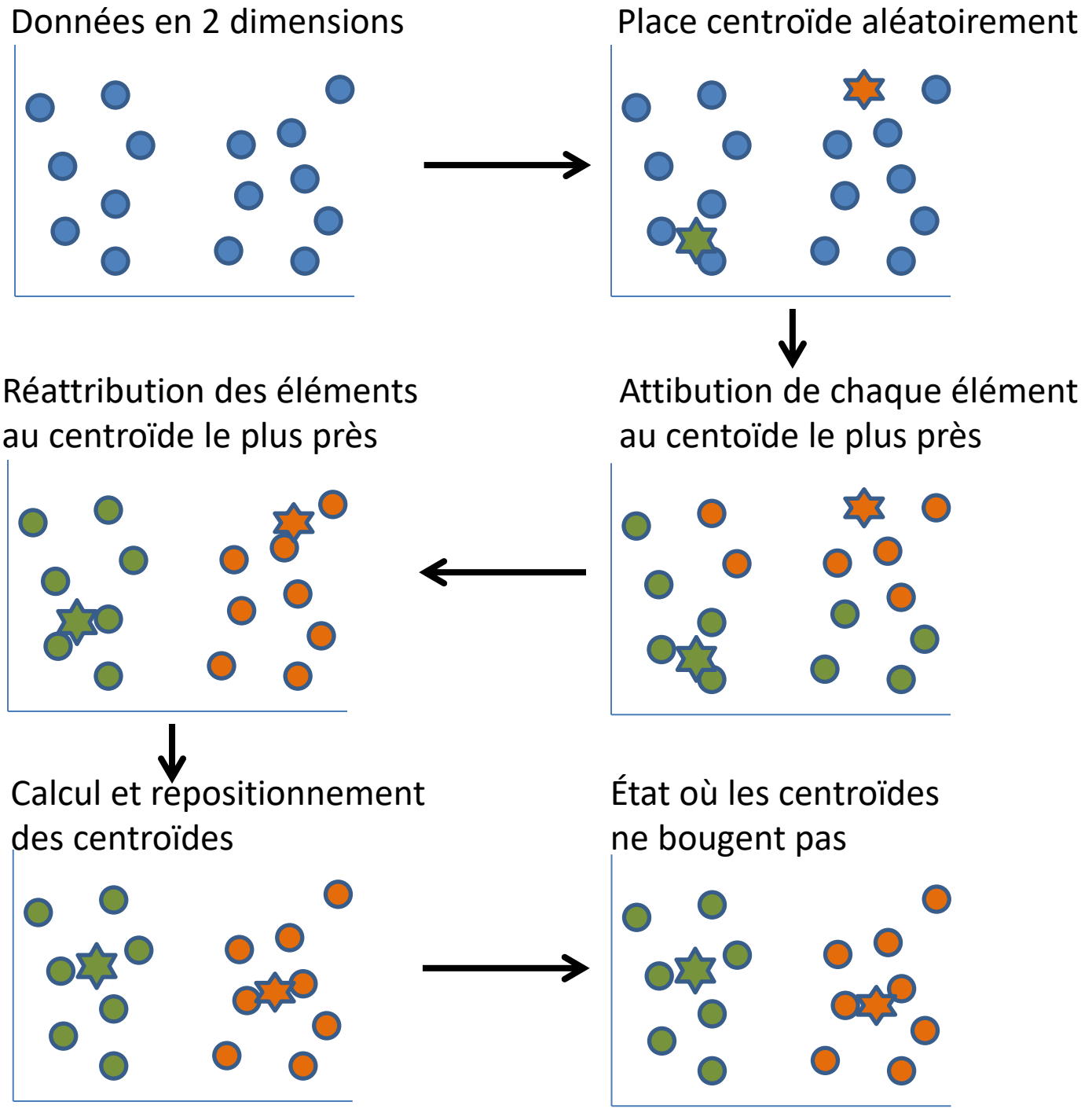


http://www.embedded-computing.com/embedded-computing-design/analytics-driven-embedded-systems-part-2-developing-analytics-and-prescriptive-controls

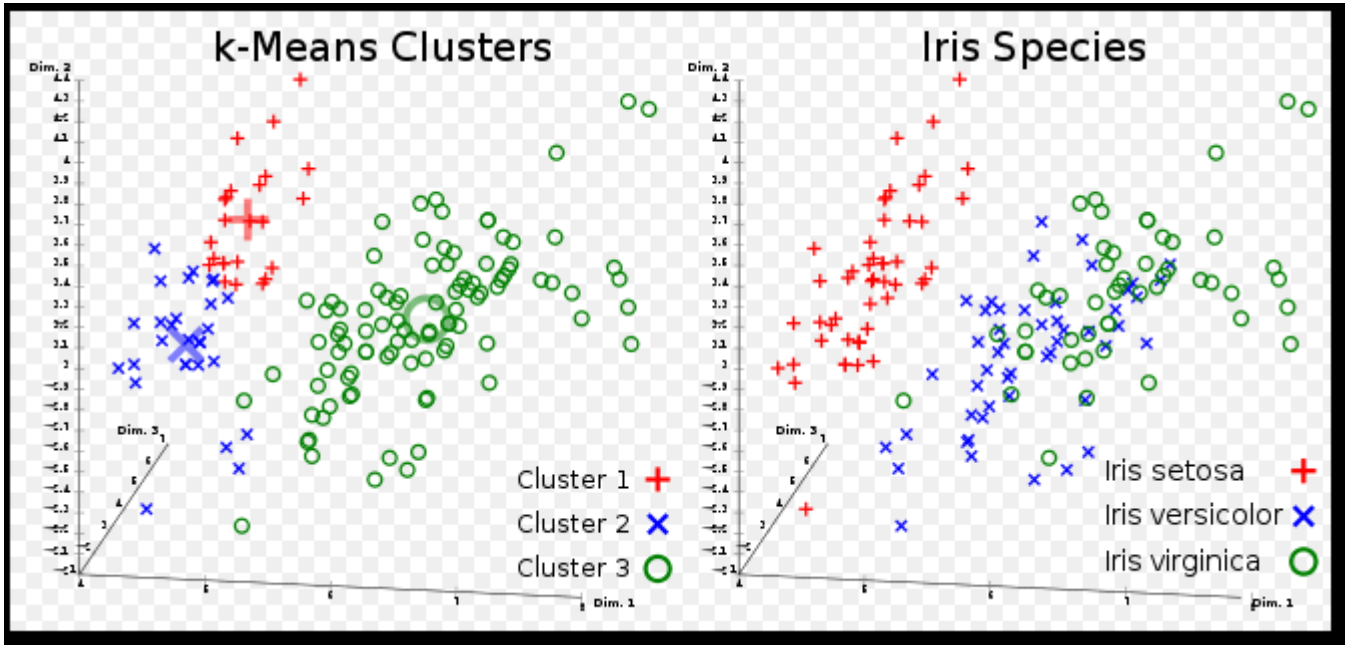
Regroupement (*clustering)*

- **But:** est de regrouper des objets qui ont des caractéristiques similaires
- Plusieurs algorithmes ont été développés:
  - Non-supervisé:
    - *K-mean*: modèle centroïde, connaissance apriori des données
    - Apriori pour apprentissage par règle d'association: détermination de relations dans de grand jeux de données (analyse de marchés)
  - Semi-supervisé/Supervisé:
    - Fourni un jeux de données d'entrainement identifié (exemple des Iris)
    - Fourni les inconnus qu'on demande de classifier
    - Réseaux de neurones

k-moyen (*k-means*)



Efficacité de K-moyenne



Source: https://en.wikipedia.org/wiki/K-means_clustering

RNSC: restricted neighborhood search

Utilise graph non-dirigé et sans poids et le but est de minimiser la fonction de coût

1. Distribution aléatoire des éléments dans des groupes (*clusters*)
2. Déplace les nœuds de manière aléatoire
   - On peut détruire les groupes et redistribuer les éléments
   - On peut mémoriser les mouvements afin d'évaluer nombre de changement
3. Fonction d'évaluation: basé sur un calcul d'amélioration

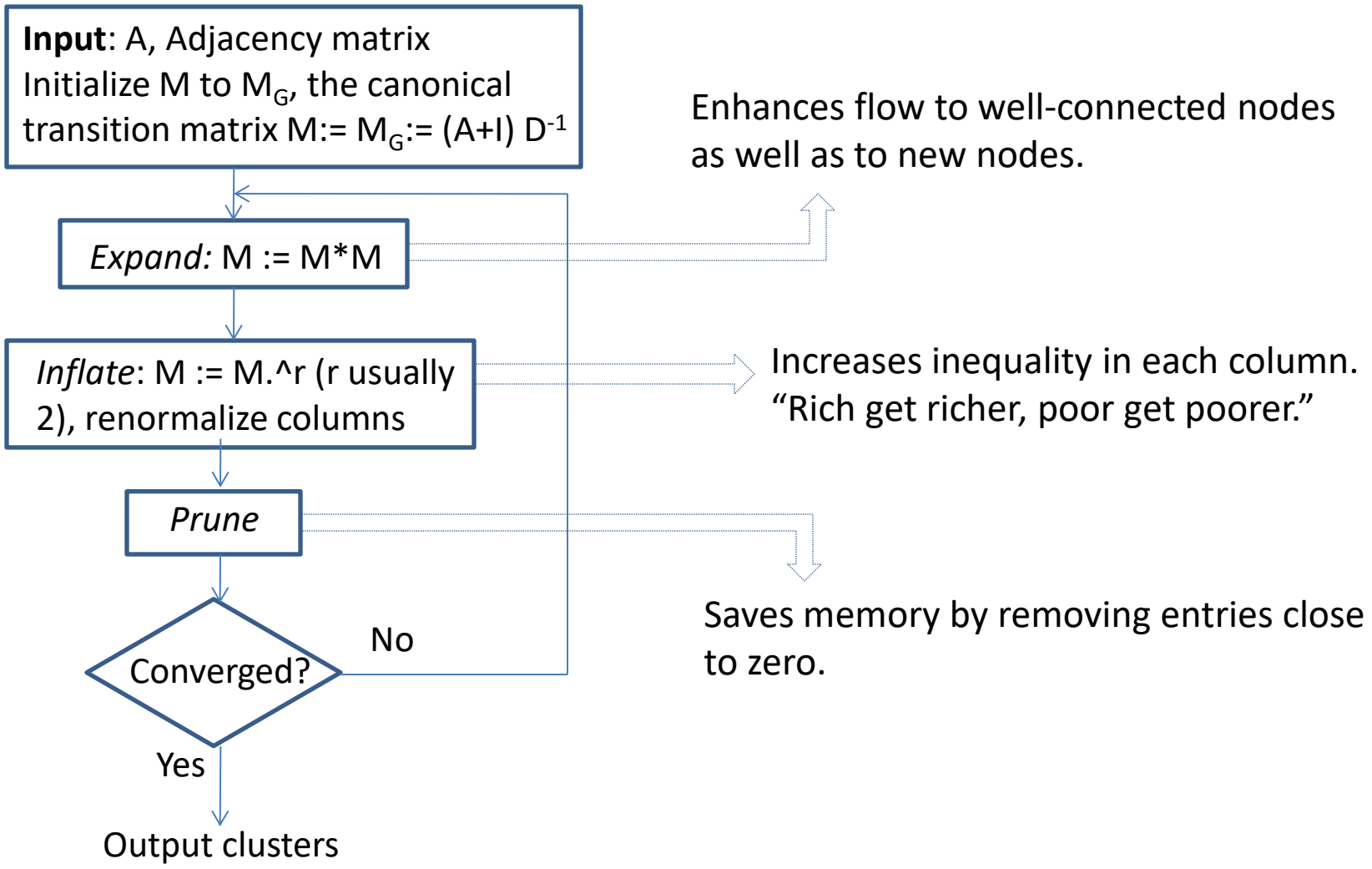
Version sur le web: http://floresta.eead.csic.es/rsat/rnsc.php

Markov clustering algorithme (MCL)

- Algorithme rapide et évolutif (*scalable*) de regroupement non-supervisé
- Accepte des ‘poids’ sur les vertex
- Initialisation: matrice de proximité
- 2 étapes:
  - Dilation (*Expansion*) = élever à la puissance ‘*n*’
  - Gonflement (*Inflation*) = renormalisation de la matrice

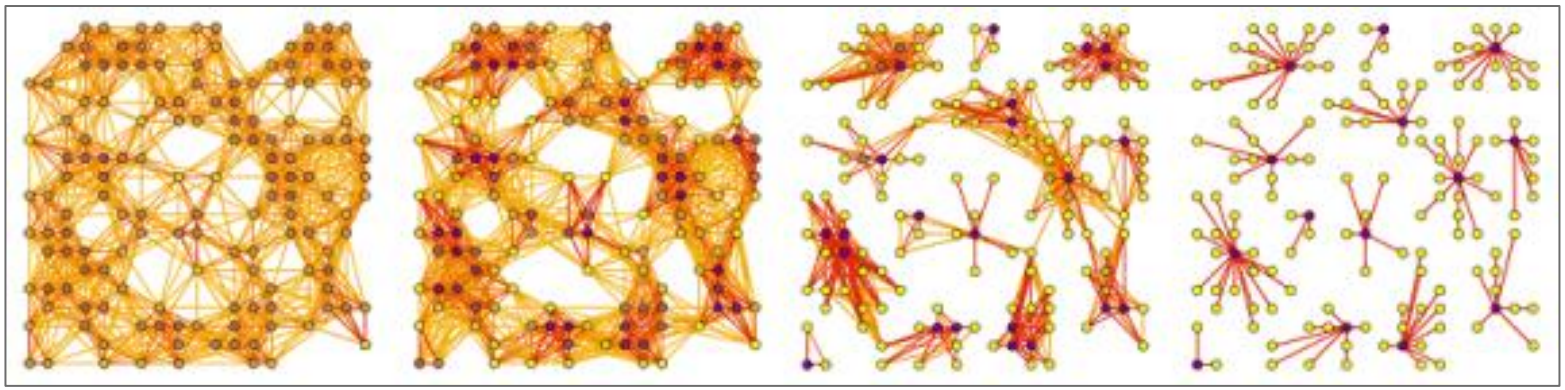
Groupes s’enrichissent

The MCL algorithm



https://www.cs.ucsb.edu/~xyan/classes/CS595D-2009winter/MCL_Presentation2.pd

Markov clustering algorithme (MCL)



Animation: https://micans.org/mcl/ani/mcl-animation.html

https://micans.org/mcl/

ClusterONE

Fonction de '**cohesiveness**' = ƒ(V):

- ω^{in} = poids total des vertexes du groupe et
- ω^{bound} = poids total des vertexes connectant le group au réseau et
- $\rho|V|$ est la pénalité modélisant les incertitudes du modèle

$$f(V) = \frac{w^{in}(V)}{w^{in}(V) + w^{bound}(V) + p \mid V|}$$

Article: http://www.nature.com.proxy.bibliotheques.uqam.ca:2048/nmeth/journal/v9/n5/full/nmeth.1938.html

ClusterONE

3 Étapes:

- Ensemencement avec des nœuds ayant le plus de connections. Groupe s'élargi avec fonction de 'cohesiveness'. Prochaine itération = prends nœuds solitaires et recommence
- Évalue le chevauchement des groupes
- Élimine les groupes les moins denses

Notez: algorithme disponible pour utilisation dans Cytoscape (http://cytoscape.org/)

Article: http://www.nature.com.proxy.bibliotheques.uqam.ca:2048/nmeth/journal/v9/n5/full/nmeth.1938.html

Algorithmes dérivés

- MCODE (Molecular Complex Detection):
  - Trouves les complexes protéiques
  - Évalue inter-connectivité des regroupements
  - Disponible dans Cytoscape
- CMC (clustering-based on maximal cliques):
  - Poids du vertex = fiabilité de l'interaction
  - But: renforcer les interactions reconnus fiables
  - Amélioration: diminue le bruit de fond des interactions

Article = GENA

- Améliorations
  - Une protéine peut appartenir à plusieurs groupes
  - Croissance des groupes indépendamment = chevaucher
  - Utilisation des poids sur les vertexs (probabilité d'interaction) sans modification (binarizing et élimine poids les plus faible basé sur un seuil)

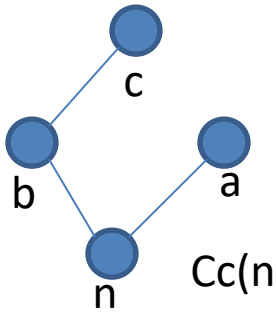
GENA

- Gradually Expanding Neighbourhoods with Adjustment: 2 grandes étapes
  1. Ensemensement avec les noeuds de manière aléatoire d'après leur coefficient de regroupement
     - Evaluation et rapprochement des noeuds avec des coefficient plus petit (valeur de: "clustering coefficient metric")
  2. Interchangement aléatoire des nœuds
     - Réévaluation du coefficient de clustering
- Équilibre quand mouvement de groupe ne peux améliorer le coefficient

Étape 1: Initiation et croissance

- Ensemensement du système basé sur plus haut coefficient de regroupement (*clustering coefficient*)

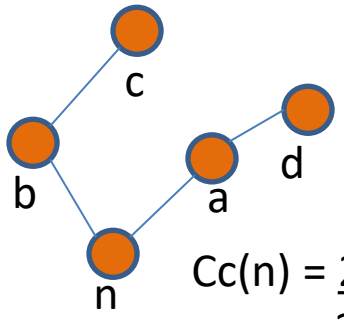
$$C_C(n) = \frac{2e_n}{k_n(k_n - 1)}$$



Cc(n) = $\frac{2*1}{2(2-1)}$ = 1

e_n = nbr de voisin connecté

κ_n = nbr de voisin du noeud



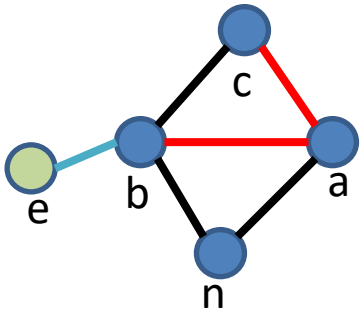
Cc(n) = $\frac{2*2}{2(2-1)}$ = 2

Dimitrakopoulos, C. *et al., Artificial Intelligence in Medicine,* 71 (2016) 62–69

Croissance des groupes
(*l-forward*, *r-backward*)

Ajout des nœuds et évaluation de la connexion (*cluster connectivity*)

$$\frac{aW_{in\_c} + (1-a)W_{out}}{|Cl \cup N_{cl}|}$$



W_{in} = vertex du groupe (=3)

W_{in_c} = complément des vertexs du groupe (donc vertex qui n'existe pas entre deux nœuds (=2))

W_{out} = vertex connectant un nœuds a un groupe différent (=1)

a = 0.6 arbitraire par les auteurs

$|Cl \cup N_{cl}|$ = Union des noeuds du groupe avec les noeuds voisins n'appartenant pas au groupe

Élimine les nœuds dont la connectivité diminue

Chaque groupe croît indépendamment

Dimitrakopoulos, C. *et al., Artificial Intelligence in Medicine,* 71 (2016) 62–69

Étape 2: Ajustement des groupes

- Espace local est exploré basé sur les nœuds initiaux
- Trois possibilités d'ajustement:
  - Nœud transféré à un groupe différent: p = 0.8
  - Nœud copié dans un autre groupe: p = 0.1
  - Nœud éliminé du groupe: p = 0.1

Étape 3: Évaluation du réseaux

Sommes de tous les groupes du réseaux

$$Eval\_cl = \sum_{i=1}^{N_{cl}} Eval(i)$$

Même fonction d'évaluation que pour les nœuds, mais on le fait pour le groupe dans le réseaux

$$Eval(i) = \frac{aW_{in\_c}(i) + (1-a)W_{out}(i)}{|Cl_i \cup N_{cl}(i)|}$$

Fin de l'ajustement = 1000000 déplacements ou que *Eval_cl* est stable depuis 1000 dernières itérations

Dimitrakopoulos, C. *et al., Artificial Intelligence in Medicine,* 71 (2016) 62–69

Évaluation du nouvel algorithme

- Compétition: RNSC, MCL, ClusterONE et GENA
- Critères d'évaluation des algorithmes:
  - Sensitivité
  - Valeur de prédictions positives (PPV)
  - Précision géométrique
- Qui à gagné???? GENA

Résultats de comparaison

Levure

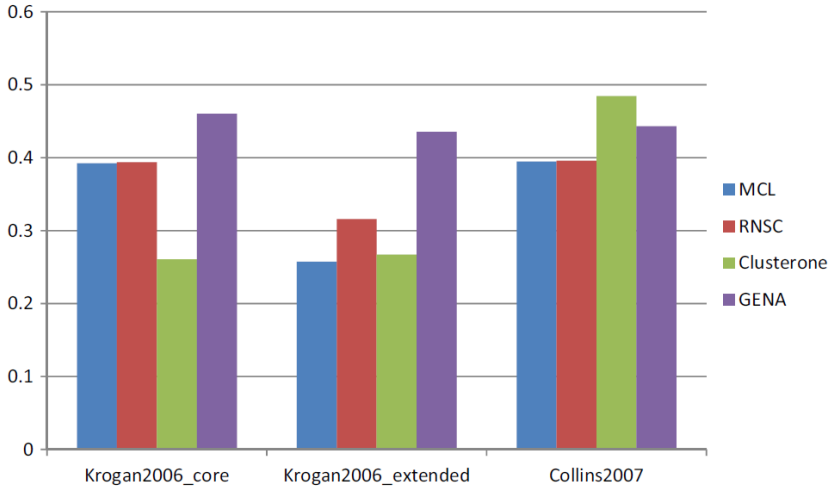


Fig. 4. Maximum matching ratio (MMR) metric for Markov clustering (MCL), restricted neighborhood search (RNSC), clustering with overlapping neighborhood expansion (ClusterONE) and gradually expanding neighborhoods with adjustment (GENA) on the three *Saccharomyces cerevisiae* input networks and the evaluation dataset of Pu.

Homme

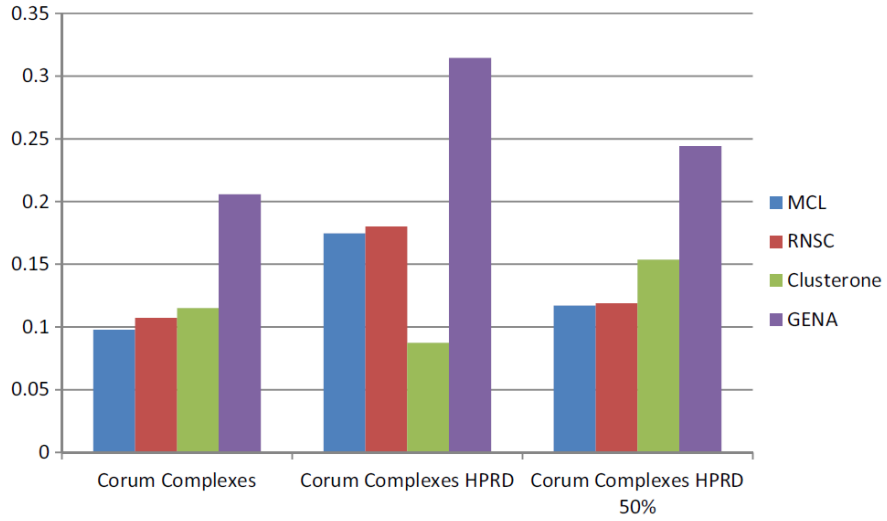


Fig. 5. Maximum matching ratio (MMR) metric for Markov clustering (MCL), restricted neighborhood search (RNSC), clustering with overlapping neighborhood expansion (ClusterONE) and gradually expanding neighborhoods with adjustment (GENA) on the three human evaluation datasets (Section 2.1) and the human protein reference database (HPRD) input network.

Majeure partie des cas, GENA est meilleure. Spéculation des auteurs: algorithme prends en compte le poids des interactions sans les altérer ('binarizing' + élimination sous un seuil critique)

Dimitrakopoulos, C. *et al., Artificial Intelligence in Medicine,* 71 (2016) 62–69

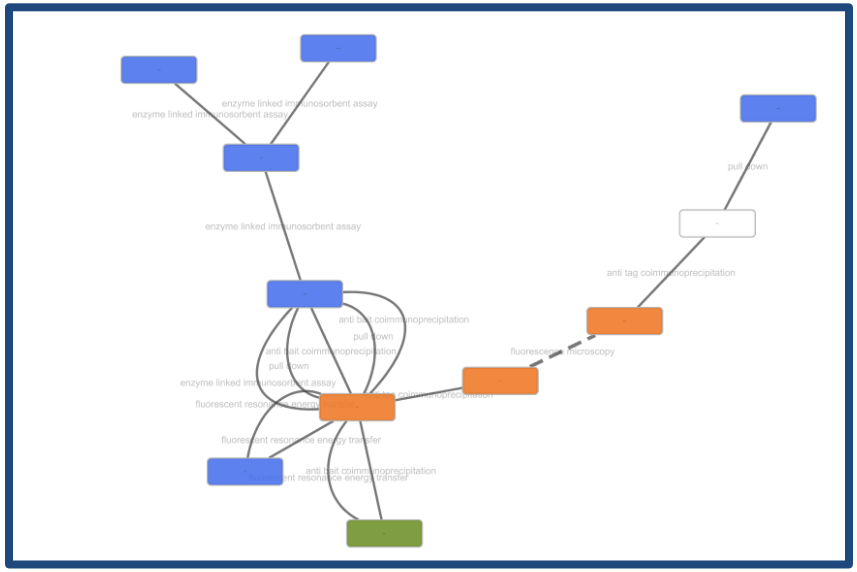
Application de GENA

Utilisé données d'expression (microarray) de 3 groupes:

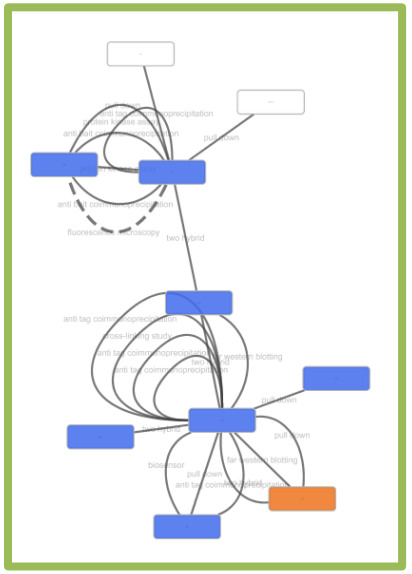
- Parkinson (PD)
- Donneur sain (HD)
- Sujet avec maladie neurodégénérative autre que Parkinson (OD)

Différence de réseaux produit par GENA

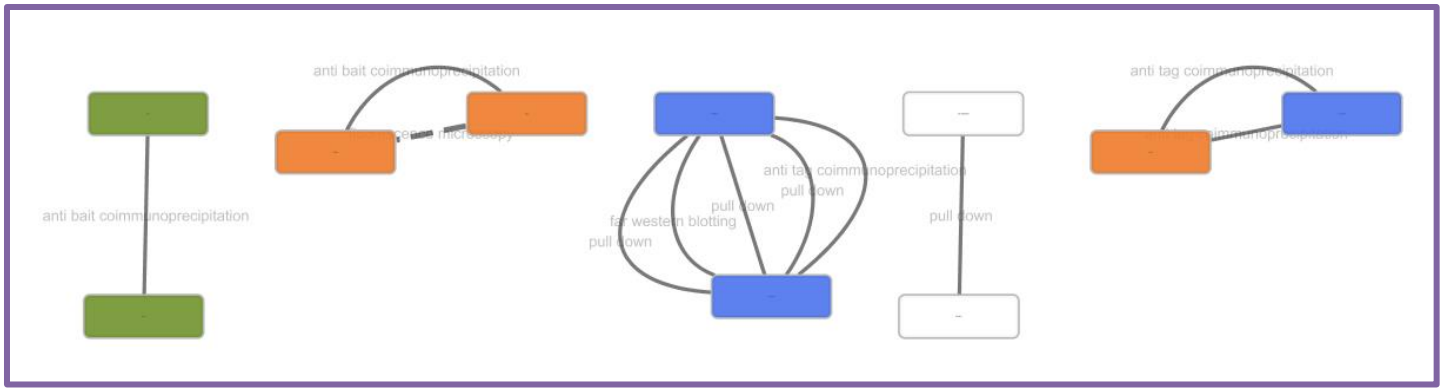
PD



HD



312 groupes existes dans HD mais pas PD ⇒ à étudier!



Conclusion: GENA ou autre?

- Une panoplie d'outils en développement
- Algorithme de classification/décision permettent
  - une extraction des données de biologie
  - un meilleur ciblage des expériences
  - une meilleur allocation des ressources
- Amélioration
  - Toujours beaucoup de faux positifs dans toutes les techniques de prédictions
  - L'ensemencement est crucial, donc peut être biaisé

Merci & Question

EXTRA

Dogme centrale de la biologie moléculaire

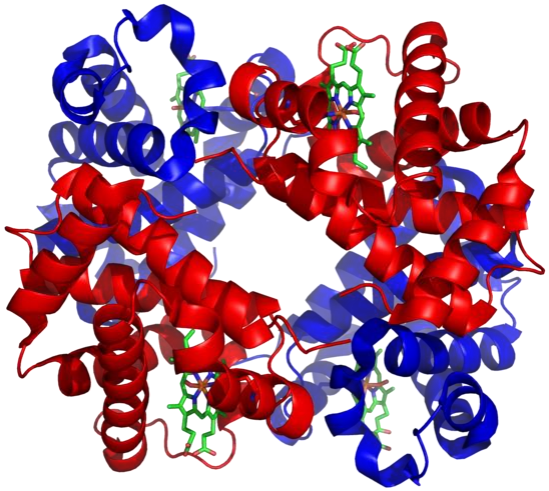
Transcription

ADN ⟶⟵ ARN ⟶ Protéines

Traduction *(Translation)*

Rétro-Transcription

Séquençage du génome humain (2003)



Structure de l'hémoglobine
https://alevelnotes.com/Protein-Structure/61

Vidéo sur le dogme (anglais):
https://www.youtube.com/watch?v=gG7uCskUOrA

Structure de l'ADN: https://www.cgtrader.com/3d-models/science/medical/dna-strand--3

Structure de l'ARN: https://biologywise.com/difference-between-dna-rna

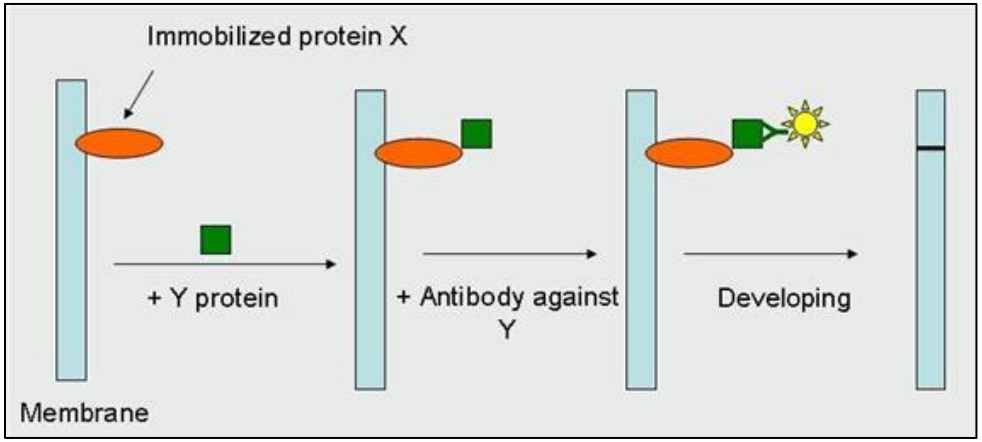
Données d’entrée

- Base de données obtenues biologiquement contenant moins de 2000 interactions chez la levure ou chez l’humain.
- Ce sont des données vérifiées par quelqu’un mais il peut y avoir des faux positifs et positifs

Étude des interactions protéines-protéines

TAP: *tandem affinity purification*

- Long
- Réactifs souvent non disponibles



MALDI-TOF & LC-MS/MS

- Coûteux
- Expertise difficile à acquérir

1) http://www.tankonyvtar.hu/en/tartalom/tamop425/0011_1A_Proteinbiotech_en_book/ch07.html