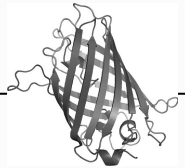
Moonlighting proteins



Emmanuel Noutahi

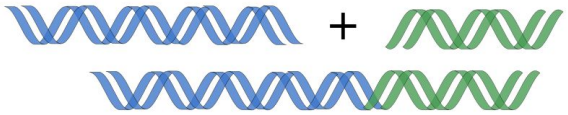
Plan de la présentation

- Résumé de la revue de synthèse
- Projet de recherche

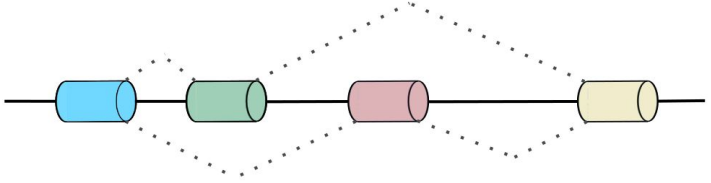
1 Revue de Synthèse

Moonlighting proteins

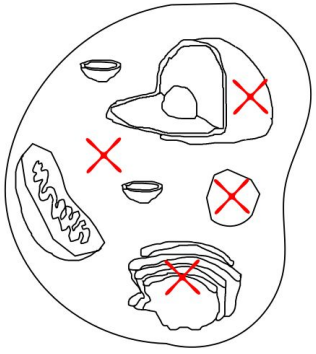
Protéines exercant plusieurs fonctions **indépendantes** à partir de la **même chaîne polypeptidique**



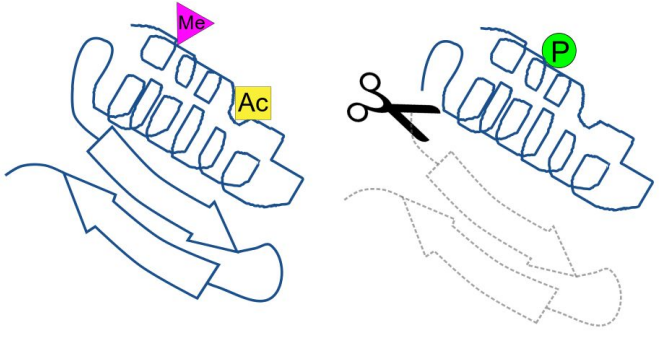
Fusion de gènes



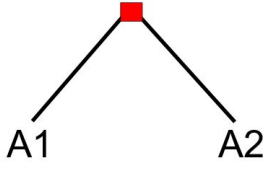
Epissage alternatif



Même fonction
multiples localisations



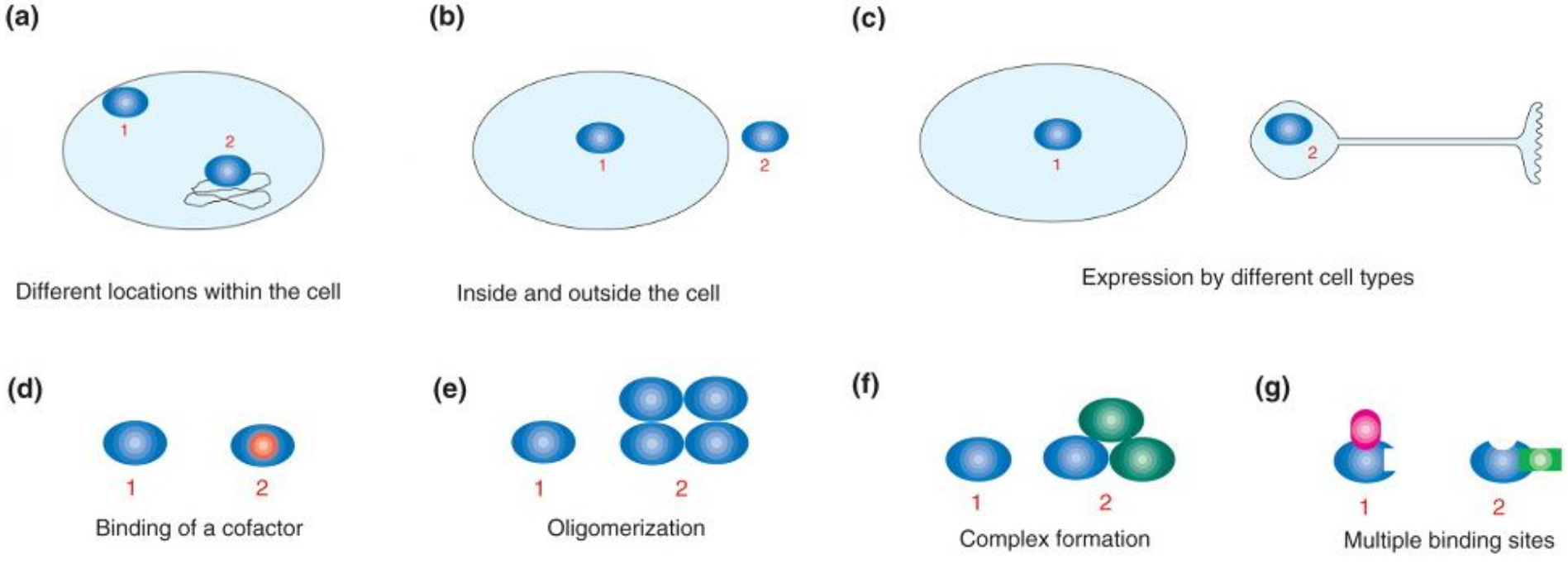
Différentes modifications post-traductionnelles



Homologues

Toute protéine multifonctionnelle n'est pas moonlighting

Moonlighting : mécanismes



Adapté de : [Jeffery, 1999, doi:10.1016/s0968-0004(98)01335-8]

Protéines MP connues

Protein	Organism	Functions
Animals		
Aconitase	*Homo sapiens*	TCA cycle enzyme
		Iron homeostasis
ATF2	***Homo sapiens***	**Transcription factor**
		DNA damage response
Crystallins*	Various	Lens structural protein
		Various enzymes
Cytochrome *c*	Various	Energy metabolism
		Apoptosis
DLD	*Homo sapiens*	Energy metabolism
		Protease
ERK2	***Homo sapiens***	**MAP kinase**
		Transcriptional repressor
ESCRT-II complex*	***Drosophila melanogaster***	**Endosomal protein sorting**
		bicoid mRNA localization
STAT3	***Mus musculus***	**Transcription factor**
		Electron transport chain

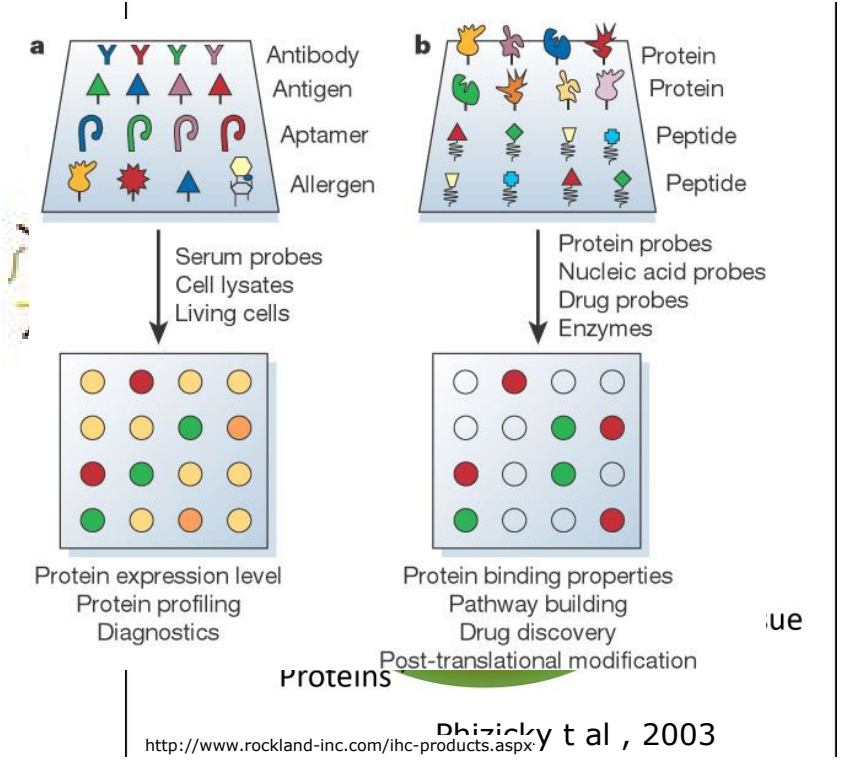
Adapté de Huberts et Van der Klei, 2010

- Impliquées dans plusieurs maladies génétiques et infectieuses
- Compliquent les associations génotype - phénotype

Identification des protéines MP

Expérimental

- MS/MS
- Immunohistochimie
- Puce à protéine
- etc



http://www.rockland-inc.com/ihc-products.aspx

Phizicky t al , 2003

In sillico identification des protéines MP (1)

Gomez et collègues

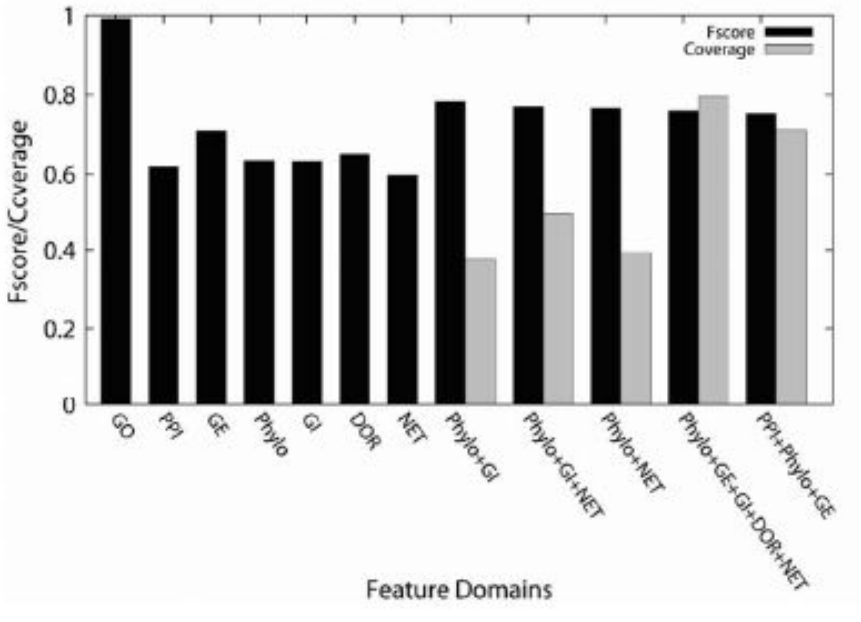
- Homologie (PSI-BLAST) **
- Conservation de motif/domaines (ProDom, PFAM) **
- Réseau d'interactions protéine-protéines (PPI)
- Structure 3D
  - Prediction de domaines fonctionnels
  - Région désordonnées
- Analyse mutation corrélation (Mistic)
- Prédiction de localisation subcellulaire

In sillico identification des protéines MP (2)

Groupe de Kihara

- Sequence (Homologie vs conservation de motif)
  - **Homologues distants ****
- Réseau d'interactions protéine-protéines
  - Interaction avec des protéines ayant des fonctions diverses
- "Profile phylogénétique"
- Region désordonnées
- Interaction génétique
- Co-expression génique

Prédiction de protéines *moonlighting* **



Khan et Kihara, 2016

Première étude de prédiction utilisant une approche d'apprentissage (Random Forest + imputation de données)

Table 1. Genome-wide prediction of moonlighting proteins

Genome	# Proteins	Cov. (%) [a)]	Known MPs Predicted [b)]	MPs (%) [c)]
yeast	6,718	69.56	22/27 (81.4%)	10.97
C.elegans	20,133	79.82	1/1 (100%)	2.73
human	20,098	67.91	33/45 (73.3%)	7.82

a) The fraction of proteins that were subject to the prediction among all the proteins in the genome; b) the number of known MPs in MoonProt predicted as MPs; c) the fraction of predicted MPs among the proteins in the genome.

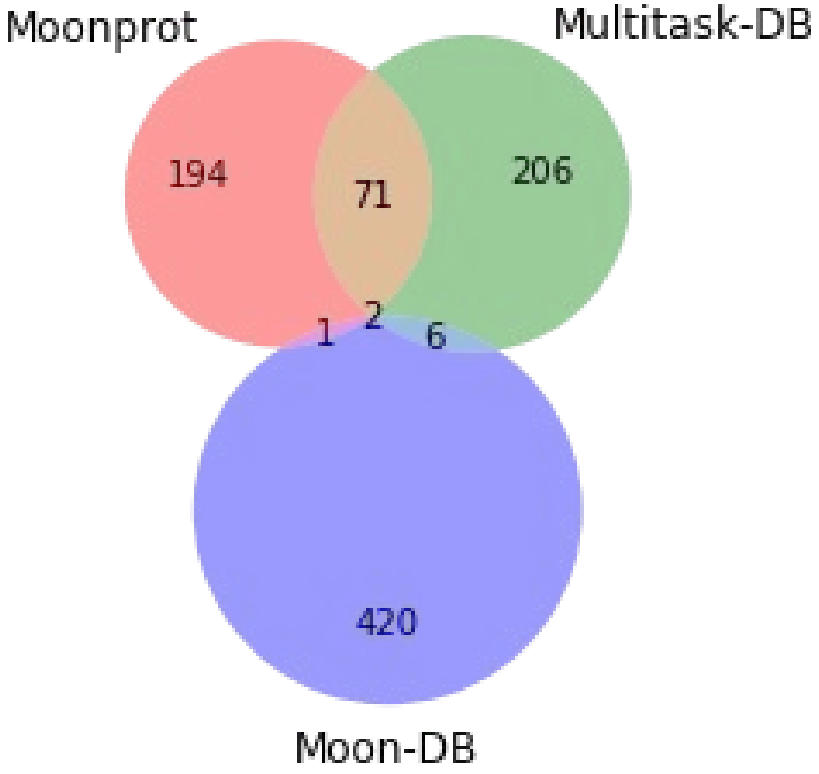
Les protéines MP prédites sont essentiellement des enzymes

1

Projet de recherche

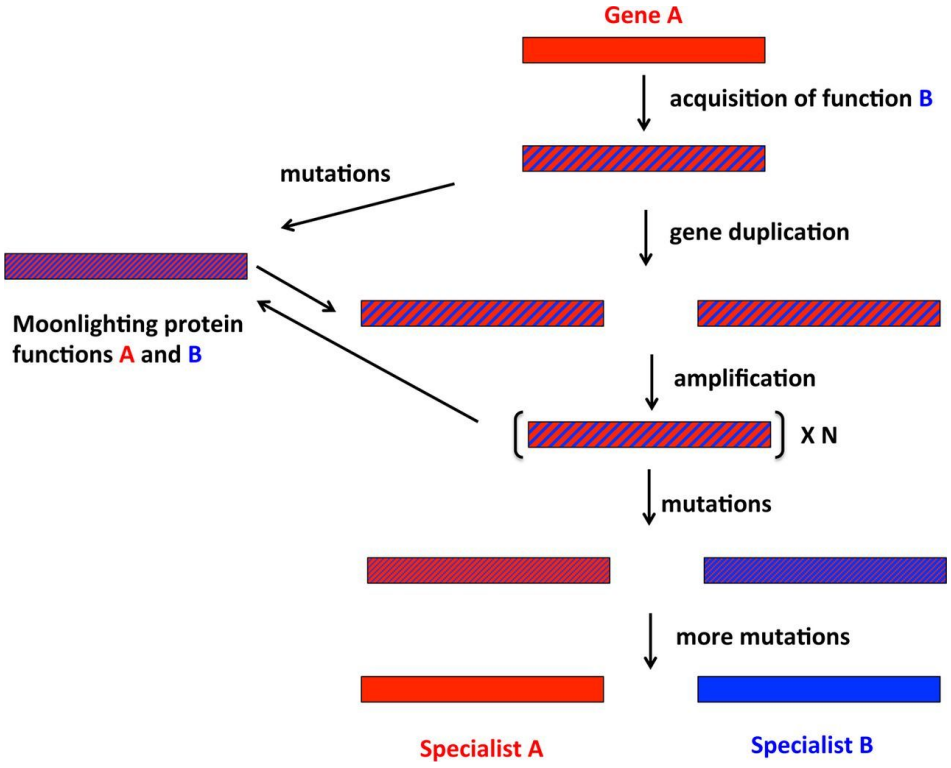
Cadre bioinformatique pour la prédiction de protéines «moonlighting » à partir de données protéomiques et phylogénomiques.

Discordance entre bases de données



Absence totale de - tentative de - validation des données

Informations phylogénomiques sous-exploitées



Copley, 2014

Partage des fonctions des protéines moonlighting entre paralogues

argininosuccinate lyase | cristalline

- MP chez canard et autriche
- Fonction partagée entre deux gènes chez le poulet

Objectifs et approches expérimentales

1. Évaluation de la qualité des bases de données de protéines MPs et validation des données afin d'établir une liste de protéines MPs fiables.

2. Caractérisation des protéines MPs fiables à partir de données protéomique et phylogénomique.

3. Prédiction de nouvelles protéines MPs au sein des génomes d'organismes modèles pour lesquels les informations protéomiques et génomiques sont souvent disponibles.

Objectifs et approches expérimentales

1. Évaluation de la qualité des bases de données de protéines MPs et validation des données afin d'établir une liste de protéines MPs fiables.

- **Text-mining** : identifier les évidences biochimiques permettant de valider les protéines recensées dans les bases de données
  - **Vérification manuelle en cas de doute**
  - **Protéines non-issues de fusion de gènes**
- **Utilisation de *knockout database* pour confirmer les résultats**

Objectifs et approches expérimentales

2. Caractérisation des protéines MPs fiables à partir de données protéomique et phylogénomique.

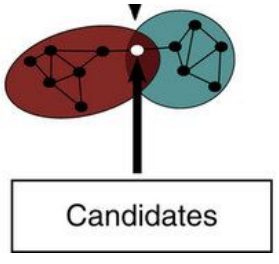
- Utilisation et extension des données ***omiques*** précédemment mentionnées
  - Interaction protéine-protéine, interaction génétique, co-expression, synténie, réseau métabolique, localisation cellulaire (base de données+ text mining), structure 3D (disponibilité de surface exposée au solvant, region fonctionnelle)
- Reconstruction de l'histoire évolutive des gènes
- Diversité fonctionnelle des protéines homologues (modélisée par un graphe d'interaction)

Objectifs et approches expérimentales

3. Prédiction de nouvelles protéines MPs au sein des génomes d'organismes modèles pour lesquels les informations protéomiques et génomiques sont souvent disponibles.

- Utilisation + amélioration de l'approche de (Khan et Kihara, 2016)
  - Score de dissimilarité de **Chapple et al**
  - **'Centralité d'inter-cluster-médiarité'**

$$c = \sum_{s \in C_1, t \in C_2, t!=s!=v} \frac{\delta_{st}(v)}{\delta_{st}}$$



- Sélection des "**prédicteurs importants**"
- Prédiction des protéines MP chez les organismes modèles

Conclusion

- Les protéines **moonlighting** représentent des atouts potentiels en médecine et en biologie (meilleure compréhension des mécanismes évolutives des fonctions de protéines)
- Abondance de ces protéines difficile à estimer
- Nouvelles approches pour leur identification à large échelle
- Caractérisation avec des données phylogénomiques pourrait aider à mieux comprendre leur mécanisme d'évolution

Merci!

Questions ?

1 Diapos supplémentaires

Méthodologie (Khan et Kihara, 2016)

Construction du dataset

Source : humain, la souris, la levure et *E. coli*

Protéines Moonlighting

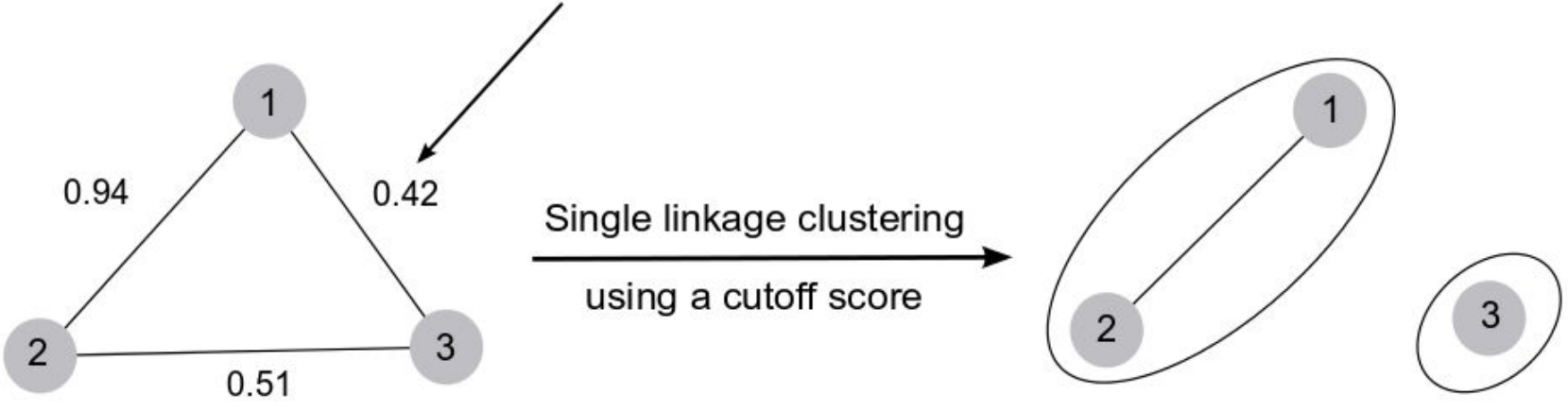
- 268 protéines provenant de la base de données MoonProt
- Moonprot est manuellement organisée

Protéines non-Moonlighting

- 162 protéines
- Au moins 8 annotations GO
- Annotations similaires pour BP et MF.

Clustering term GO

$$sim_{Rel}(c_1, c_2) = \max_{c \in S(c_1, c_2)} \left(\frac{2 \cdot \log p(c)}{\log p(c_1) + \log p(c_2)} \cdot (1 - p(c)) \right)^*$$



Données utilisées

PPI

Réseau d'interaction protéine-protéines (clustering du réseau à des seuils variés)

DOR

Consensus de prédiction de régions intrinsèquement désordonnées (nombre, taille)

Phylo

Voisinage des gènes, (clustering du réseau à des seuils variés)

GI

Interaction génétique

(clustering du réseau à des seuils variés)

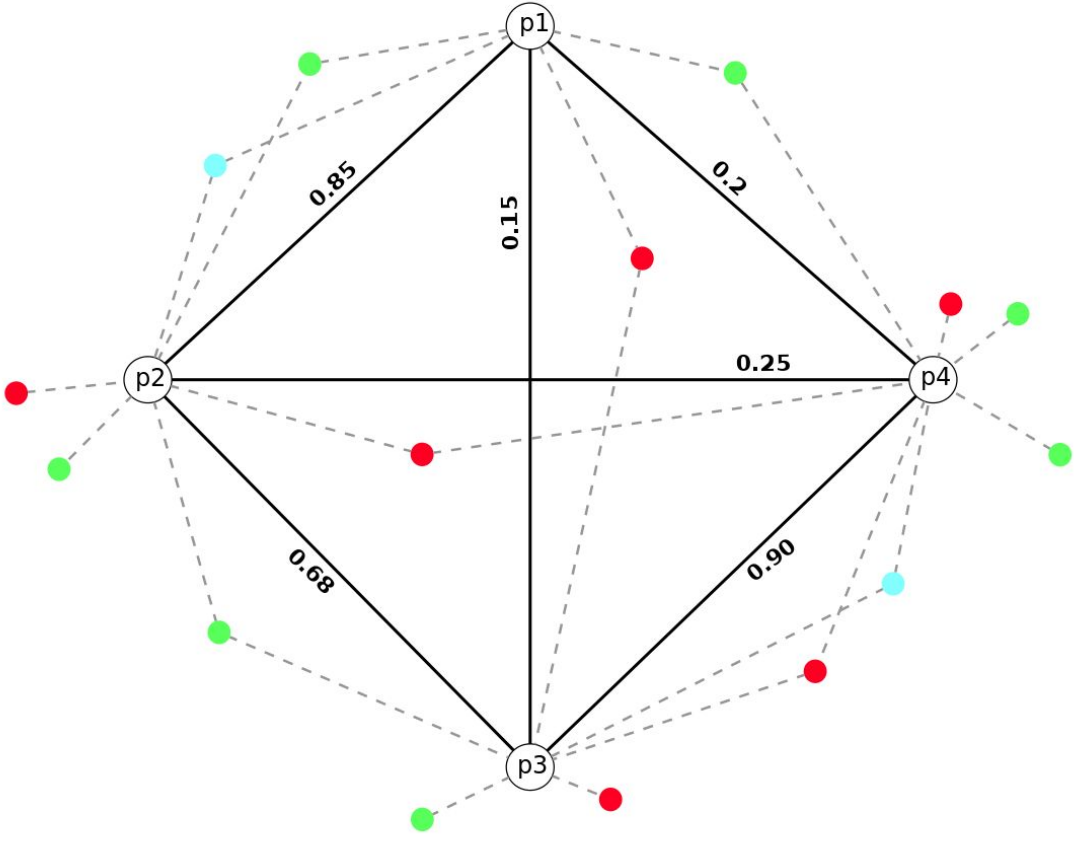
GE

Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.

NET

Propriété du réseau PPI (centralité de proximité, d' intermédiarité et de degré)

Réseau d'interaction



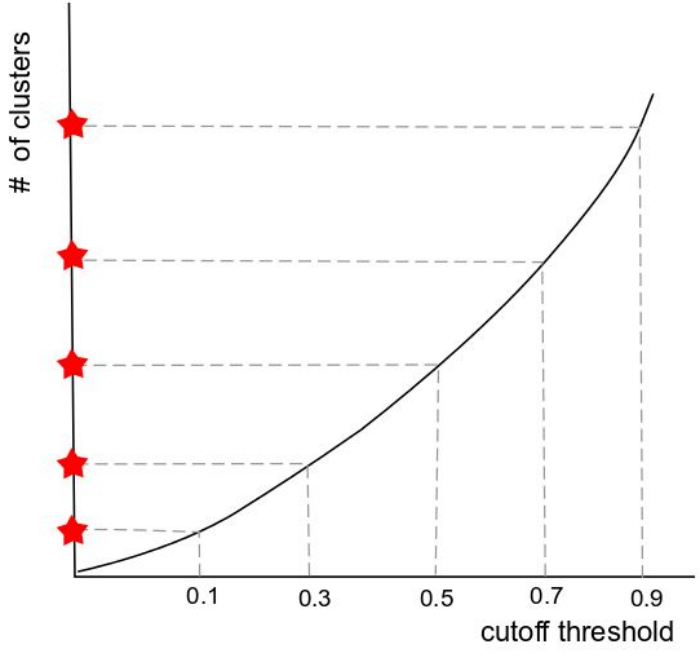
Go terms

- Biological Process (BP)
- Molecular Function (MF)
- Cellular Component (CC)

Score de similarité sémantique
(Schlicker et al., 2006)

- Réseau PPI
- Co-expression génique
- Interaction génétique
- "Profile phylogénétique"

Prédicteurs du Random Forest

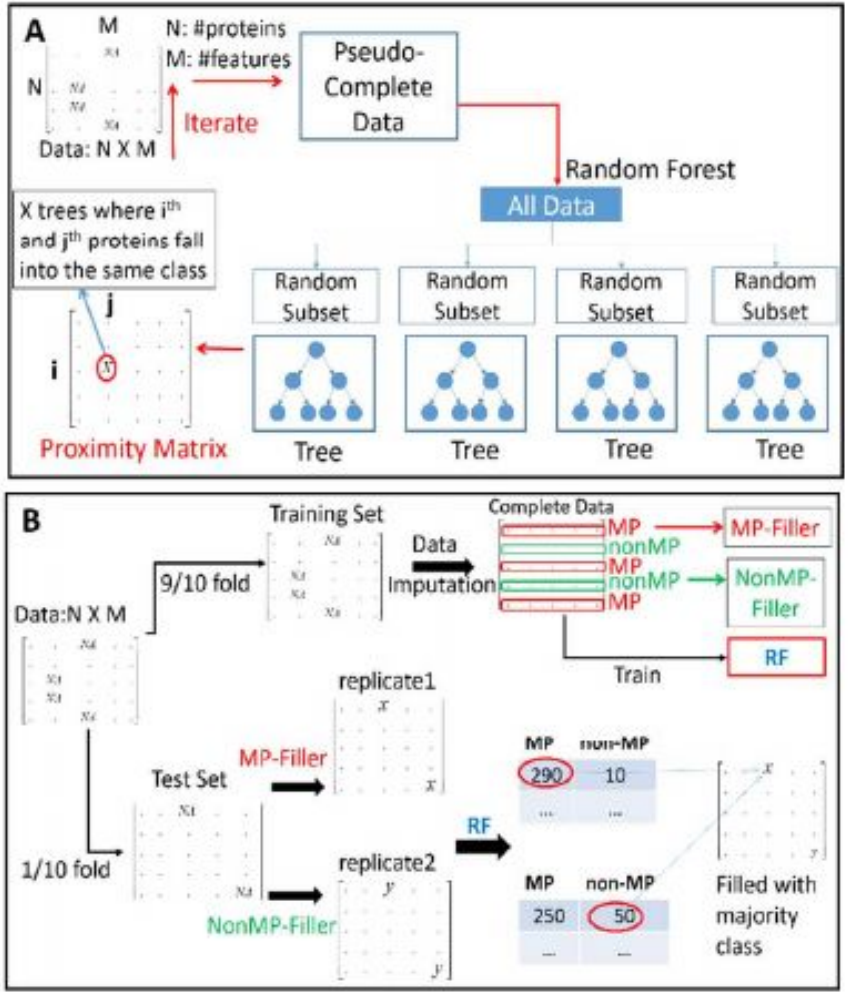


	Biological Process (BP)			Molecular function (MF)		
P1	3			4		
	2	2	3	3	3	3
P2	1			0		
	0	0	0	NA	NA	NA
	0.1	**0.5**	**0.9**	**0.1**	**0.5**	**0.9**

Random Forest + Imputation pour données manquantes

Couverture

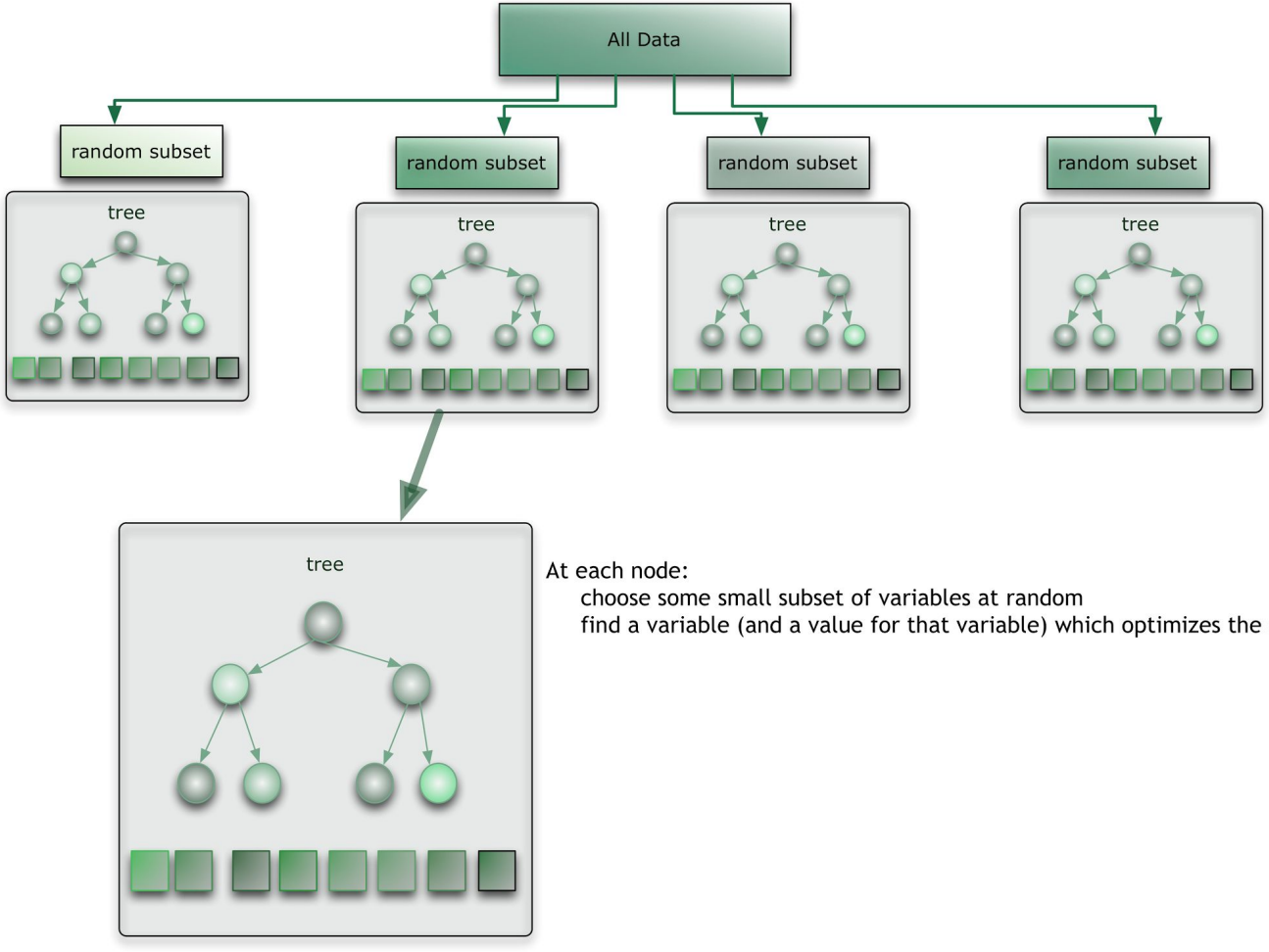
	GE			GI		
P1	2	2	3	3	3	3
P2	2	2	3	NA	NA	NA
P3	2	NA	3	NA	3	4



1

Diapos supplémentaires

Divers

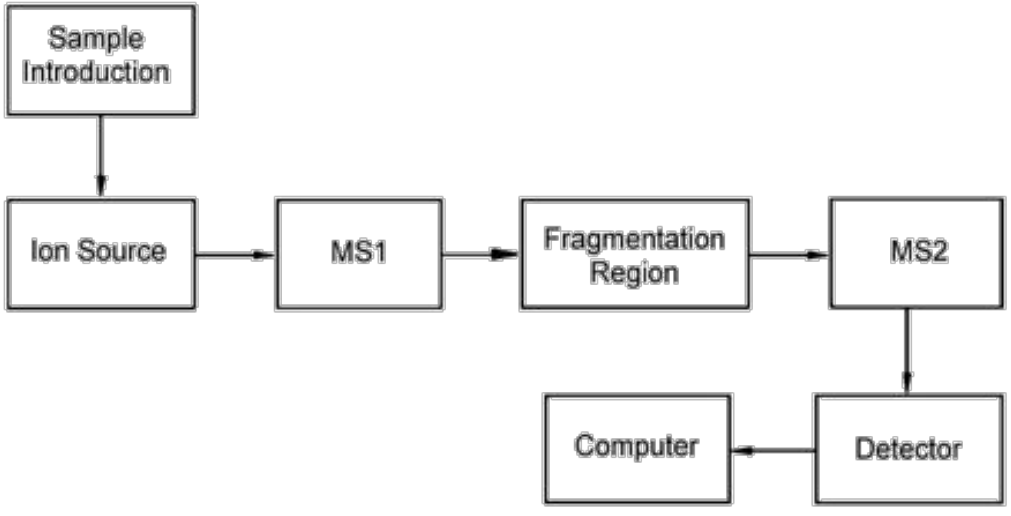


With replacement (~ ⅔)

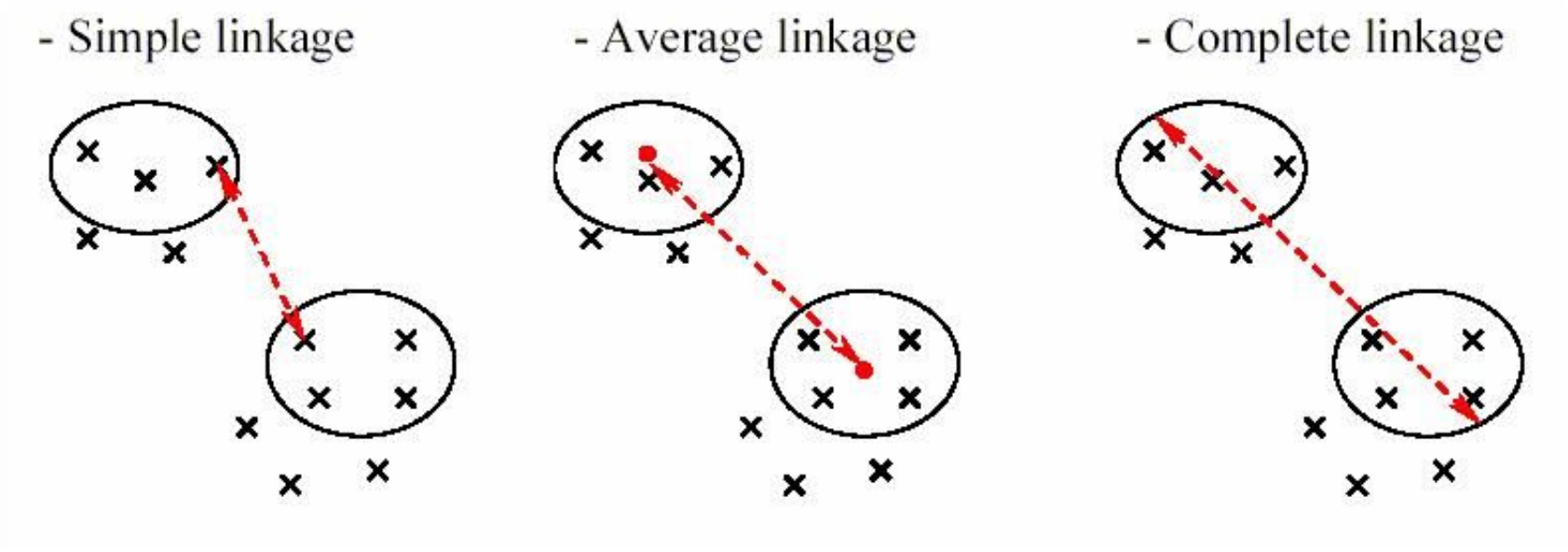
At each node:
choose some small subset of variables at random
find a variable (and a value for that variable) which optimizes the split

Random Forest : training

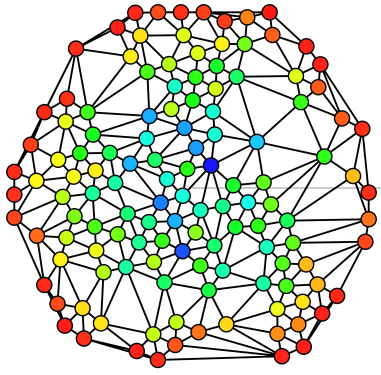
MS/MS


Sample Introduction
Ion Source
MS1
Fragmentation Region
MS2
Detector
Computer

Clustering comparison



Propriété de graphes



Centralité d'intermédiarité

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

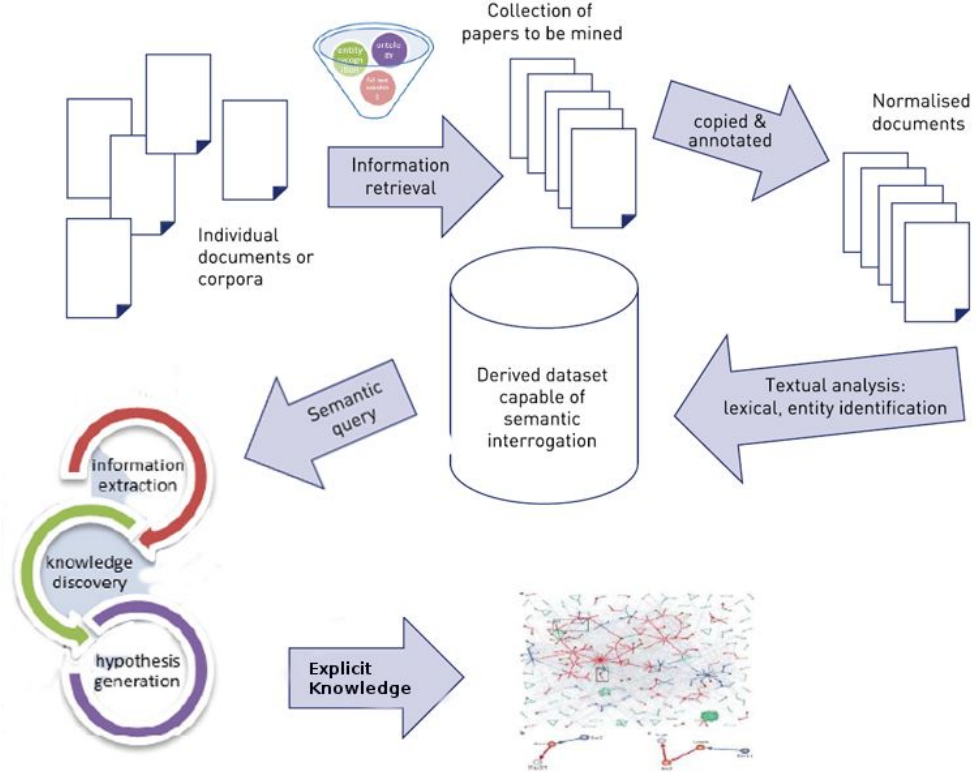
Centralité de degré

$$\mathbf{C_d}(\mathbf{v}) = \frac{\mathbf{deg}(\mathbf{v})}{\mathbf{deg}(\mathbf{v}^*)}$$

Centralité de proximité

$$C(x) = \frac{1}{\sum_y d(y,x)}.$$

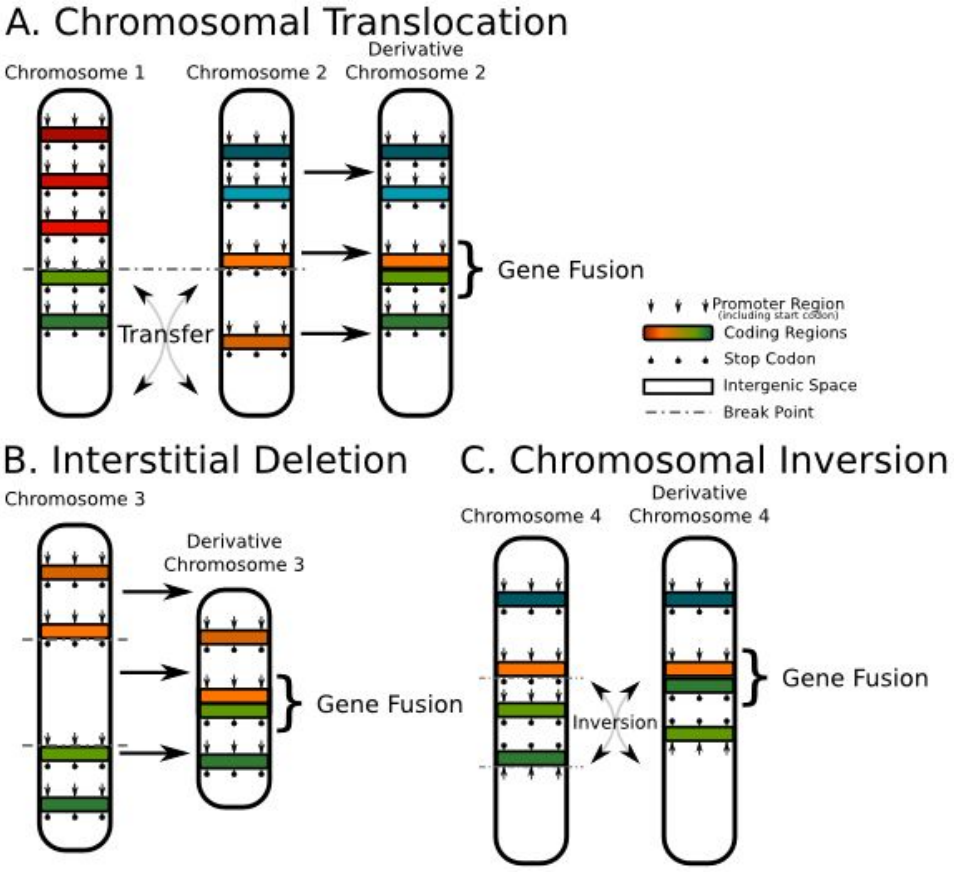
Text mining



Couramment utilisé pour l'annotation de protéines et en cancérologie

Fusion de gènes

Protéines multi-modulaires, très souvent multifonctionnelles



Gene fusion, wikipedia