



FOUILLE DE DONNÉES SUR LES RÉSEAUX (Fouille de graphes)

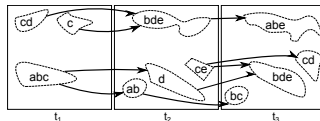
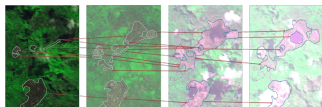
Claude Pasquier

Laboratoire I3S

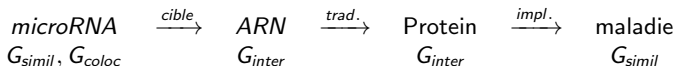
6 décembre 2018

Types de structures

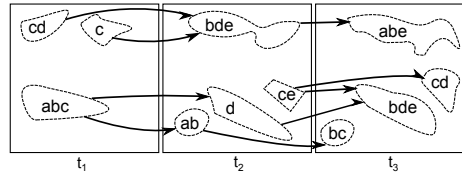
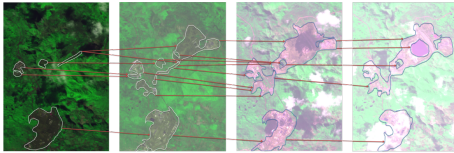
- Des éléments (nœuds) reliés par des liens
- les liens peuvent être orientés (arcs) ou non orientés (arêtes)
- les liens et les nœuds peuvent être :
 - ▶ valués
 - ▶ multi-valués
 - ▶ associés à des valeurs discrètes (attributs)
- les nœuds et liens peuvent être typés
- exemple : données spatio temporelles



- exemple : données biologiques



THÈME 1 : Recherche de motifs fréquents dans des graphes attribués

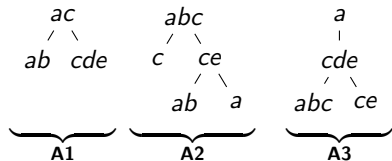


Spécificités de notre problème

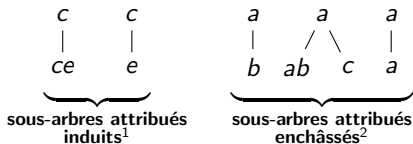
- les itemsets associés aux nœuds ne sont pas forcément identiques
- les motifs peuvent être enchâssés (implicites)

Motifs fréquents dans les arbres attribués

Base de données en entrée



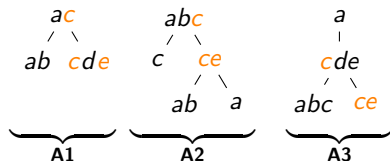
Sous-arbres attribués communs



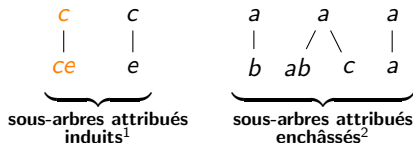
¹sous-arbre attribué isomorphe préservant la relation parent-enfant

²sous-arbre attribué isomorphe préservant la relation ancêtre-descendant

Base de données en entrée



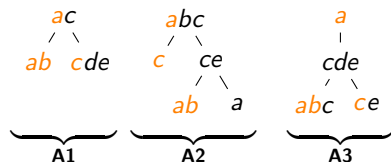
Sous-arbres attribués communs



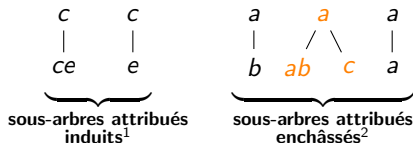
¹sous-arbre attribué isomorphe préservant la relation parent-enfant

²sous-arbre attribué isomorphe préservant la relation ancêtre-descendant

Base de données en entrée



Sous-arbres attribués communs



¹sous-arbre attribué isomorphe préservant la relation parent-enfant

²sous-arbre attribué isomorphe préservant la relation ancêtre-descendant

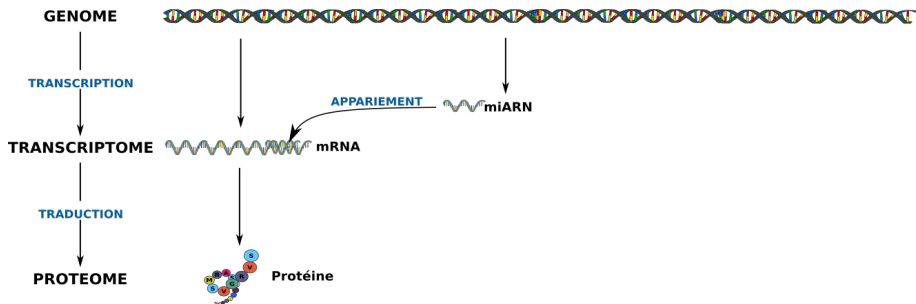
- Recherche de sous arbres attribués fréquents
 - ▶ méthode exhaustive permettant l'énumération :
 - ★ des motifs ordonnées ou non ordonnées
 - ★ des motifs induits ou enchâssés (avec distance maximale)
 - ★ des motifs concis
- Recherche de sous graphes attribués fréquents
 - ▶ prise en compte des arcs réentrants et des cycles
 - ▶ méthode d'élagage des motifs automorphes
- recherche de chemins fréquents pondérés
 - ▶ ex : $ah \xrightarrow{5} cd \xrightarrow{2} ace \xrightarrow{4} ab$
- Applications
 - ▶ données d'érosion
 - ▶ propagation de la dengue
 - ▶ pubmed 60K noeuds, 60K arcs
 - ▶ twitter 81K noeuds 1.7M arcs
 - ▶ google+ 100K noeuds 13M arcs

C Pasquier, J Sanhes, F Flouvat, N Selmaoui-Folcher, PAKDD'13

C Pasquier, F Flouvat, J Sanhes, N Selmaoui-Folcher, KAIS 2017

J Sanhes, F Flouvat, N Selmaoui-Folcher, C Pasquier, JF Boulicaut, IJCAI'13

THÈME 2 : Prédiction des associations microARN-maladies



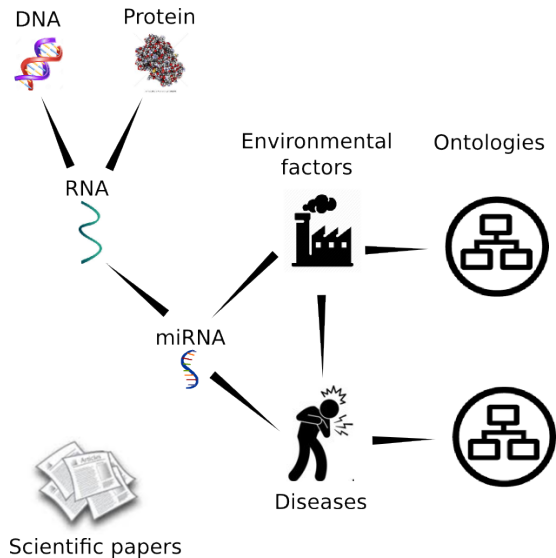
Transcriptome ensemble des ARN issus de la transcription du génome

ARN codant ARN traduits en protéine

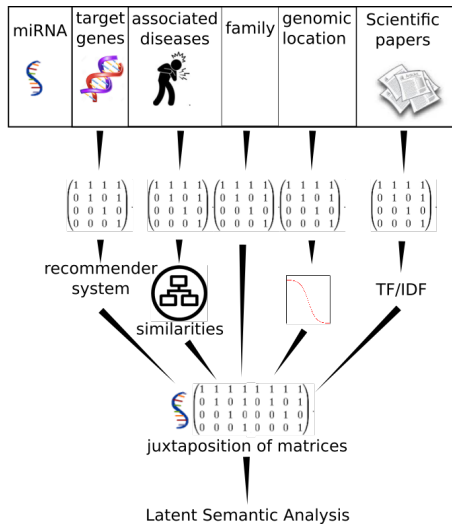
ARN non codant ARN non traduits en protéine

Transcriptomique étude de l'abondance des ARN

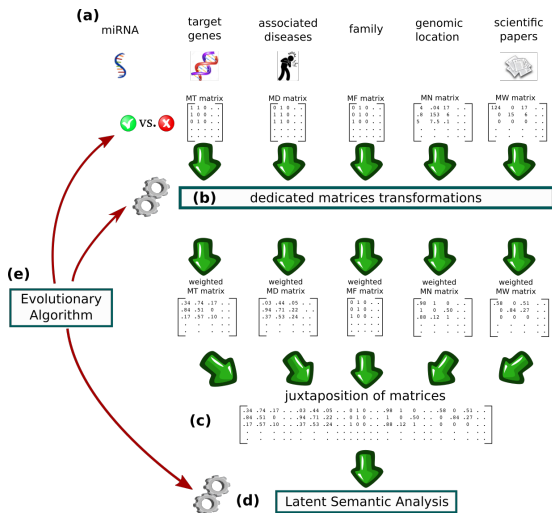
THÈME 2 : Prédiction des associations microARN-maladies



Méthode basée sur la sémantique distributionnelle (MiRAI)



Optimisation de MiRAI à l'aide d'un algorithme évolutionniste



35 paramètres binaires
 2^{35} possibilités

évaluation coûteuse de
chaque configuration

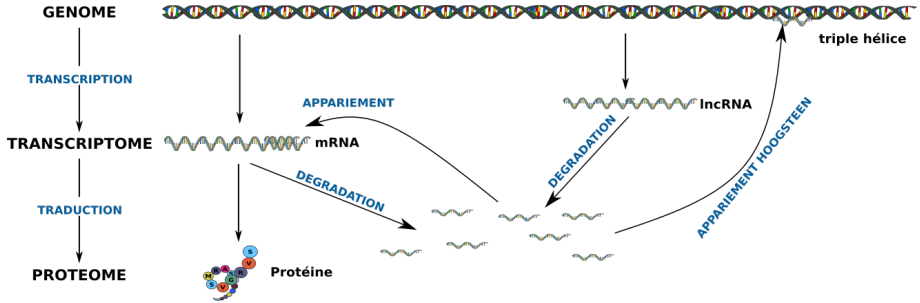
⇒ évolution différentielle
adaptée à un espace binaire

⇒ modèle de substitution
pour approximer la fonction
fitness

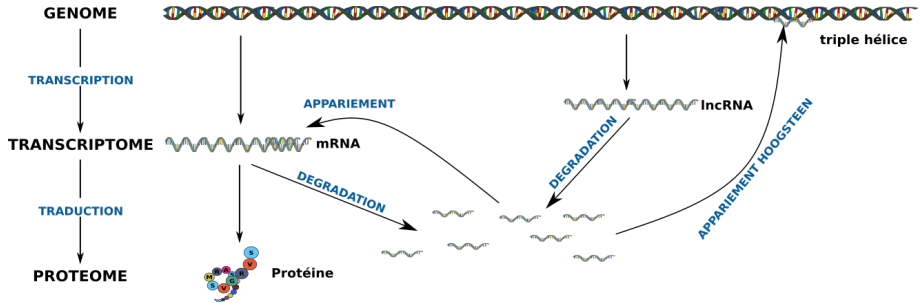
Comparaison avec d'autres méthodes

Disease name	RWR 2012	HDMP 2013	RLS 2014	MIDP 2015	MiRAI 2015	MiRAI+EA 2017
Acute myeloid leukemia	0.84	0.86	0.85	0.91	0.90	0.91
Breast neoplasms	0.79	0.80	0.83	0.84	0.86	0.86
Colorectal neoplasms	0.79	0.80	0.83	0.85	0.86	0.87
Glioblastoma	0.68	0.70	0.71	0.80	0.90	0.87
Heart failure	0.72	0.77	0.74	0.82	0.80	0.85
Hepatocellular carcinoma	0.75	0.76	0.79	0.81	0.81	0.82
Lung neoplasms	0.83	0.84	0.86	0.88	0.90	0.93
Melanoma	0.78	0.79	0.81	0.84	0.85	0.87
Ovarian neoplasms	0.88	0.88	0.91	0.92	0.87	0.91
Pancreatic neoplasms	0.87	0.90	0.89	0.95	0.93	0.92
Prostatic neoplasms	0.82	0.85	0.84	0.88	0.87	0.87
Renal cell carcinoma	0.82	0.83	0.84	0.86	0.87	0.88
Squamous cell carcinoma	0.82	0.82	0.85	0.87	0.88	0.89
Stomach neoplasms	0.78	0.79	0.80	0.82	0.82	0.85
Urinary bladder neoplasms	0.82	0.85	0.85	0.90	0.88	0.90
AVERAGE	0.799	0.816	0.827	0.862	0.867	0.880

THÈME 3 : Étude des interactions ARN $\Rightarrow$ ARN et ARN $\Rightarrow$ ADN

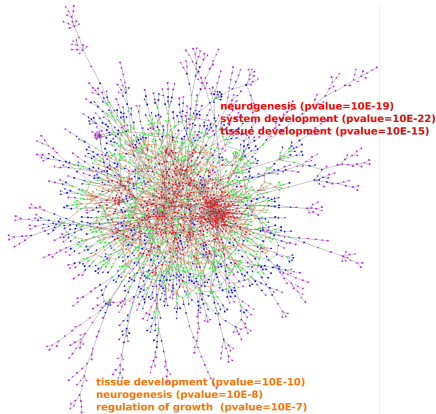


THÈME 3 : Étude des interactions ARN $\Rightarrow$ ARN et ARN $\Rightarrow$ ADN



La régulation génétique par des fragments d'ARN formant des triple hélices pourrait être d'une ampleur insoupçonnée

THÈME 3 : Étude des interactions ARN $\Rightarrow$ ARN



graphe potentiel des interactions ARN-ARN

les gènes les plus ciblés sont impliqués dans des fonctions importantes

la pertinence des annotations est corrélée avec la centralité dans le graphe

les gènes les plus exprimés ont peu ou pas d'interactions

Les résultats suggèrent l'existence d'un vaste réseau de régulation des gènes via des mécanismes d'interférence par ARN

Les résultats attendus des analyses transcriptomiques ont évolué

- Les biologistes réalisent des expérimentations plus précises
- Ils attendent des résultats beaucoup plus ciblés

Analyse transcriptomique utilisant l'optimisation multicritères

[Thèse de Leandro Correa débutée en janvier 2018]

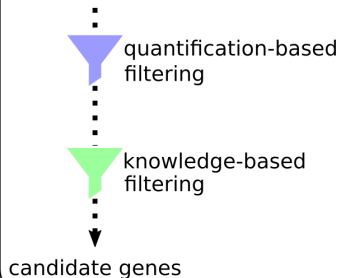
Quantification

- transcriptomique
- protéomique

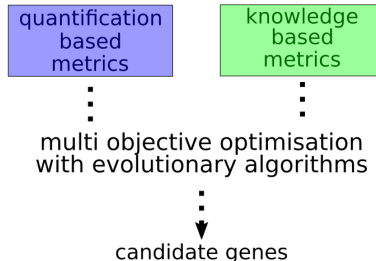
Biological knowledge

- interaction networks
- molecular pathways
- ontologies

Traditional pipeline



Our proposal



- Analyse transcriptomique basée sur l'optimisation multicritère

- ▶ collaboration avec l'iBV
- ▶ financement par la fondation ARC
- ▶ financement par l'INCa
- ▶ thèse financée par l'Université Côte d'Azur



- Prédiction des associations miARN-maladies

- ▶ collaboration avec Biomanda



- étude des interactions ARN $\Rightarrow$ ARN et ARN $\Rightarrow$ ADN

- ▶ collaboration avec l'institut Sophia Agrotech

