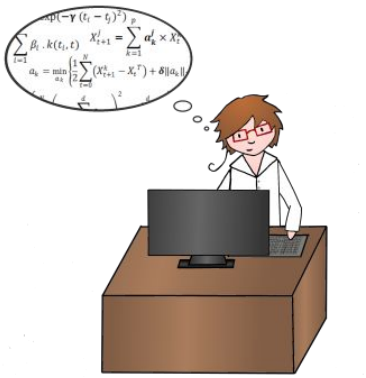
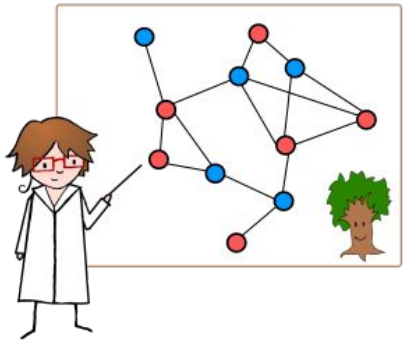
NETBIO 2017

Inférence de réseau de gènes pour comprendre la réponse du peuplier aux flexions





Lise Pomiès

INRA SCIENCE & IMPACT
UNIVERSITE Clermont Auvergne
U.M.R. PIAF

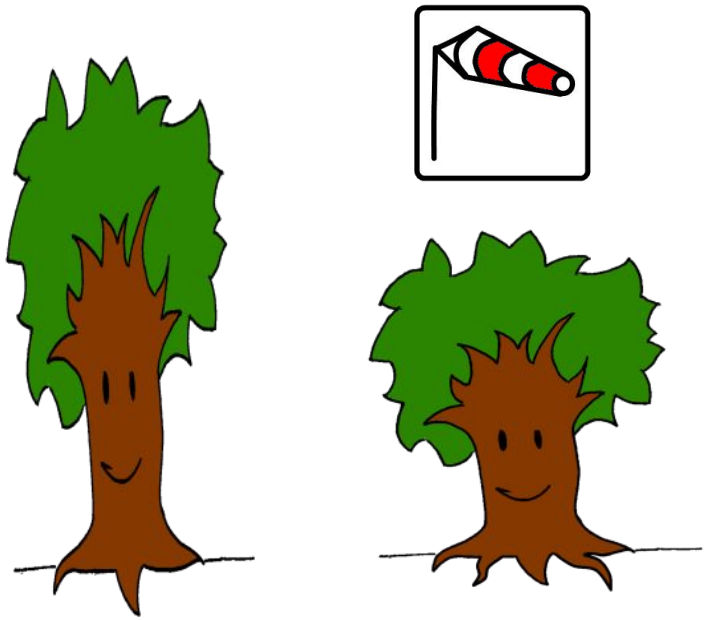
Acclimatation des arbres au vent

Changement climatique :

Nombre de de tempêtes ++

Intensité des tempêtes ++

Vents journaliers --



Modifications morphologiques

Croissance différente :

Longitudinale --

Radiale ++

Racinaire ++

Acclimatation au vent : modifications mises en places pour mieux résister aux vents futurs

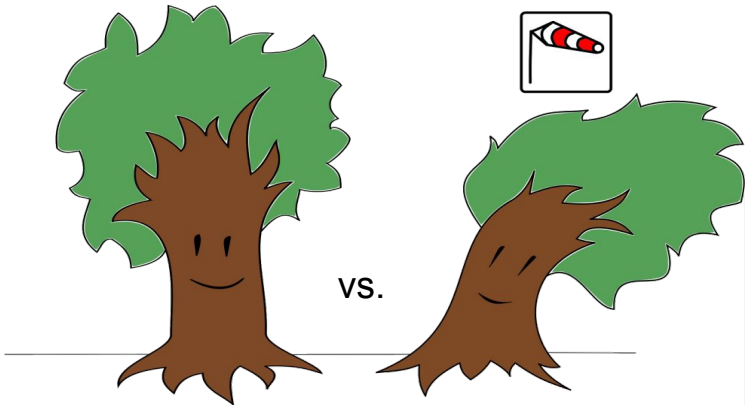
Photo : Chloé Pomiès

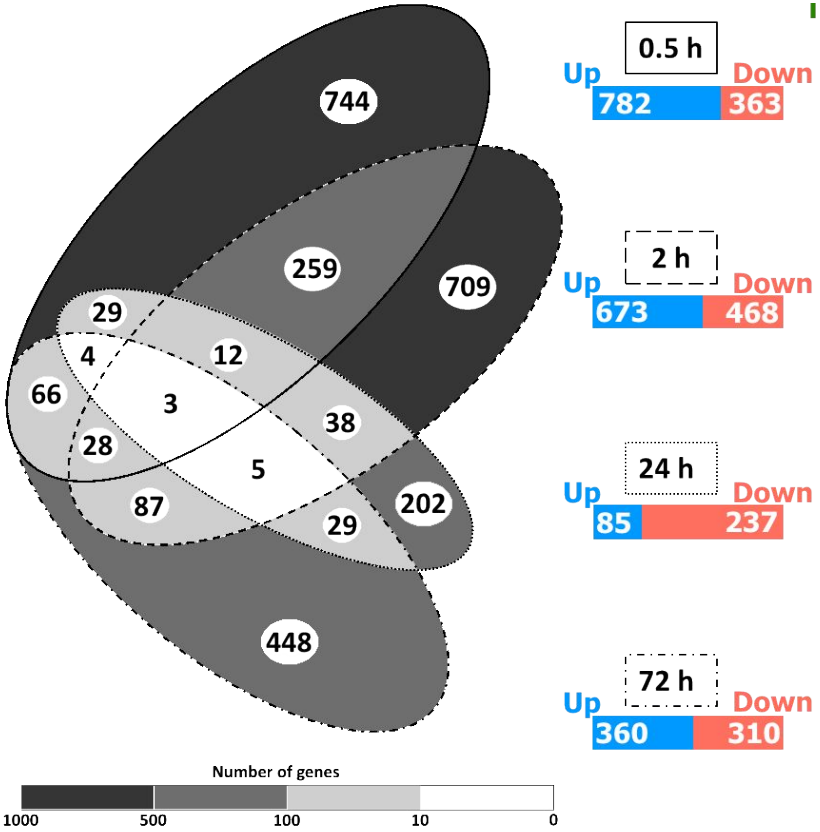
Réponse transcriptomique à une flexion transitoire

Vent → flexion tiges perçue par l'arbre

Analyse du transcriptome du peuplier après **une flexion** transitoire de la tige (10 s)

4 temps de mesures : **0.5 h | 2 h | 24 h | 72 h**





2 663 gènes différentiellement exprimés (**~6,4 % du génome**)

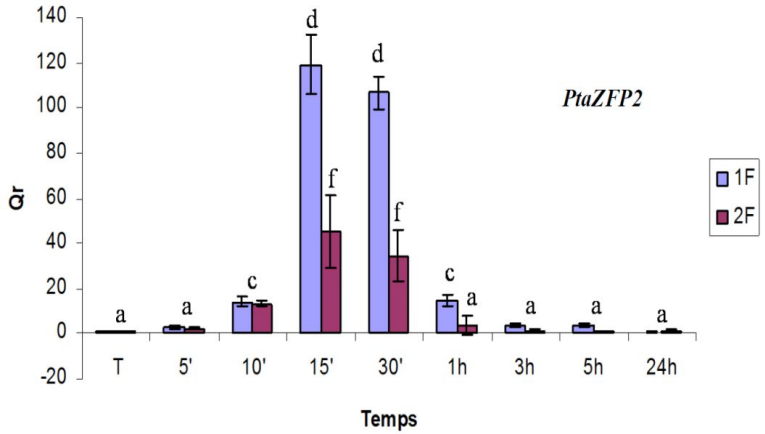
→ De très **nombreux gènes** impactés

→ Des **profils d'expression** très **différents**

Pomiès *et al.* 2017

Accommodation aux flexions répétées

Vent = plusieurs **flexions répétées**

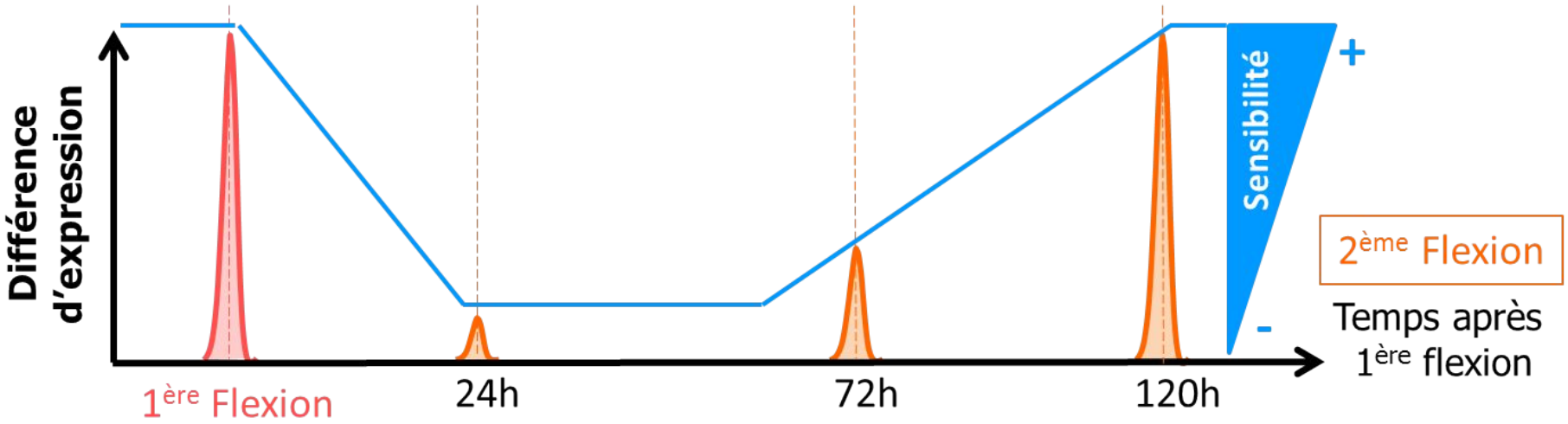


Expression induite après une 2ème flexion (24h) **plus faible** : **gène accommodé**

96 % des 1 145 DEG à **0.5h** après **1 flexion** sont **accommodés** à la 2ème flexion

Le peuplier est capable d'**adapter sa sensibilité** à la flexion :

- **perte rapide** de la sensibilité (dès 3h)
- **retour progressif** de la sensibilité (à partir de 3 jours)



Lakhal 2010, Martin *et al.* 2010

Objectif

Quel réseau de régulations transcriptionnelles régule la réponse du peuplier à la flexion et en particulier la mise en place de l'accommodation ?

~ 3 000 gènes
différentiellement exprimés

4 temps de mesures :
0.5 h | 2 h | 24 h | 72 h

→

Nombre gènes ++

Nombre mesures - -

Pas de temps irréguliers

→

≠ jeux de données idéal
pour l'inférence de réseau dynamique

Stratégie d'inférence du réseau de gènes :

1. **Réduction** du nombre de **gènes**
2. **Augmentation** du nombre de **mesures**
3. Choix d'un **modèle modulaire** adapté aux données
4. **Sélection** des meilleures **solutions**

1. Réduction du nombre de gènes

Sélection de **gènes représentants** :

Les gènes doivent représenter la **variabilité** des **profils d'expression** et des **processus biologiques** en réponse à une flexion

Comment caractériser les profils d'expression ?

Clustering des **3 000 gènes** du peuplier impliqués dans la **réponse aux flexions**

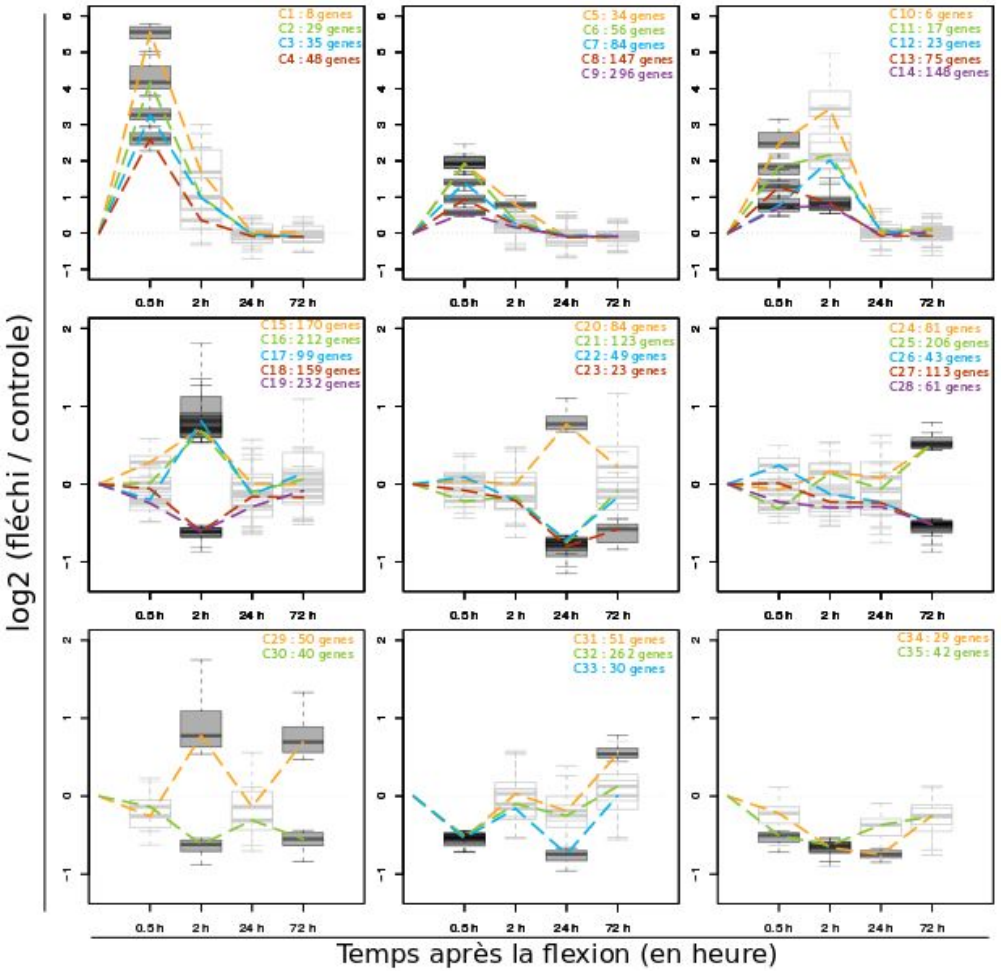
Méthode : *k-means*

Distance : *euclidienne*

- Pente entre 2 mesures
- Indice de significativité
- Différence 1F vs. 2F

→ gènes regroupés en

40 clusters d'expression



1. Réduction du nombre de gènes

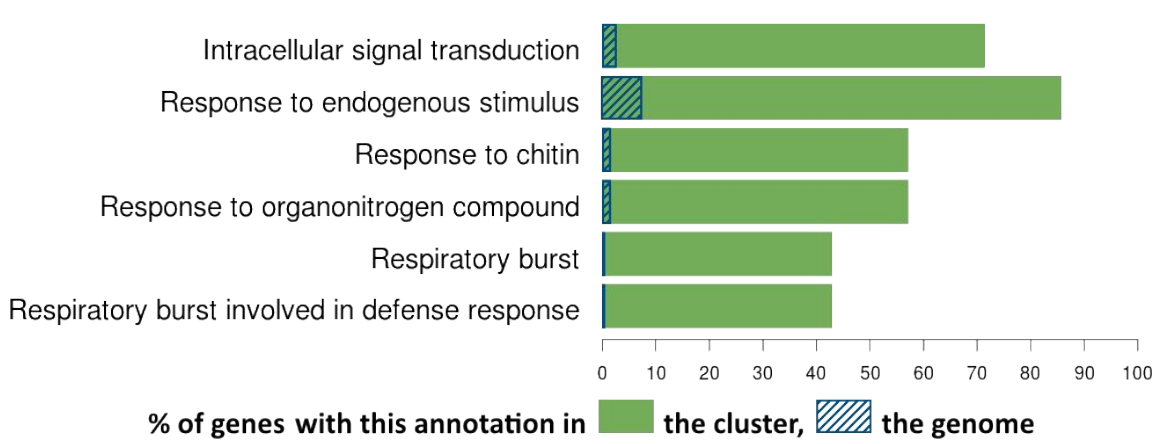
Sélection de **gènes représentants** :

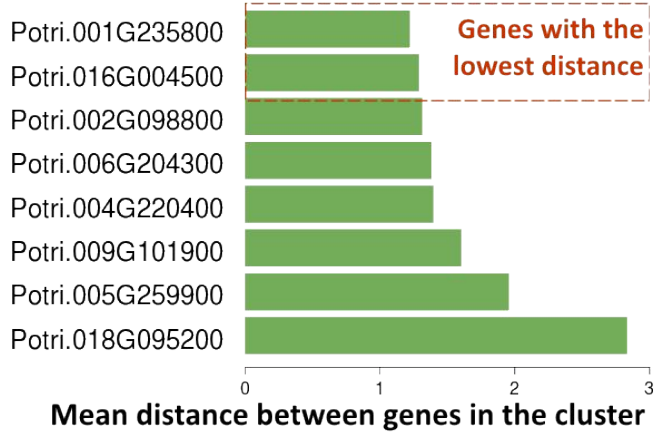
Les gènes doivent représenter la **variabilité** des **profils d'expression** et des **processus biologiques** en réponse à une flexion

Comment caractériser les processus biologiques ?

Enrichissement en **GO** des clusters de gènes

Exemple : Cluster n°1 : 8 gènes sur–exprimés 0,5h après flexion





6 annotations sur-représentées
PtaZFP2 annoté avec ces 6 annotations

PtaZFP2 faible distance moyenne avec les autres gènes

***PtaZFP2* représentant du cluster n°1**

1-3 représentants sélectionnés pour chaque cluster d'expression

→ **47 gènes représentants**

2. Augmentation du nombre de mesures

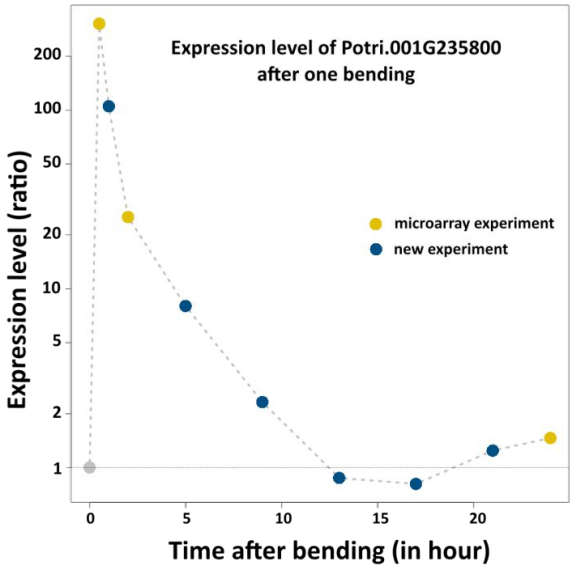
zone d'ajout de points

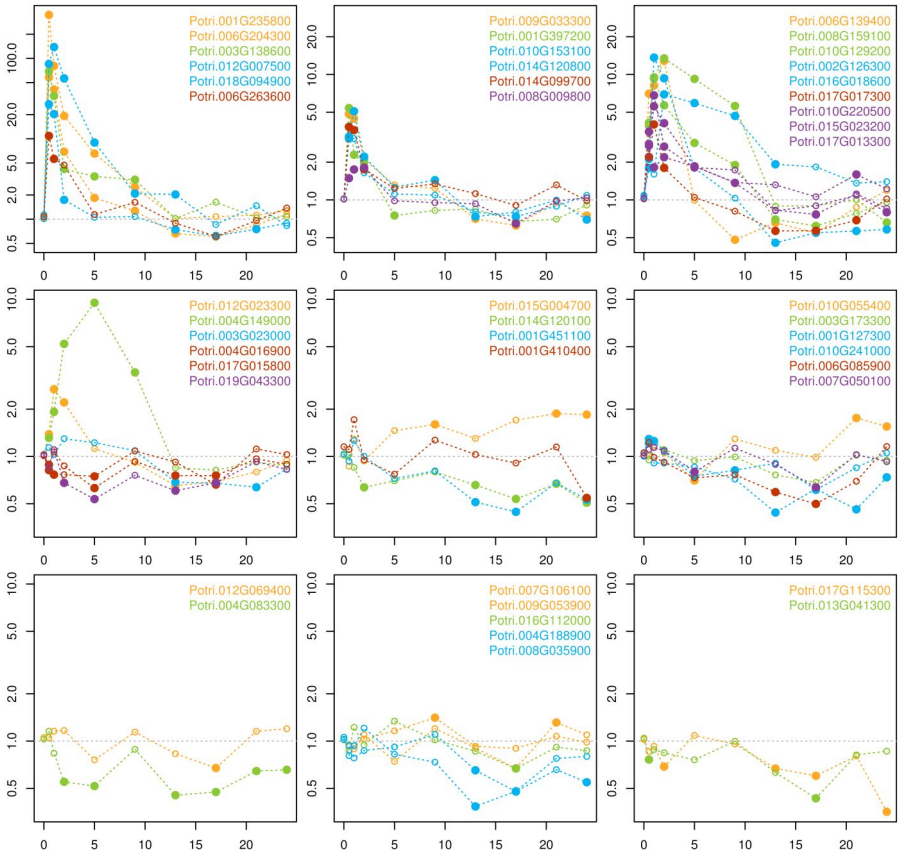
Mise en place de l'accommodation

Accommodation active

Perte progressive de l'accommodation

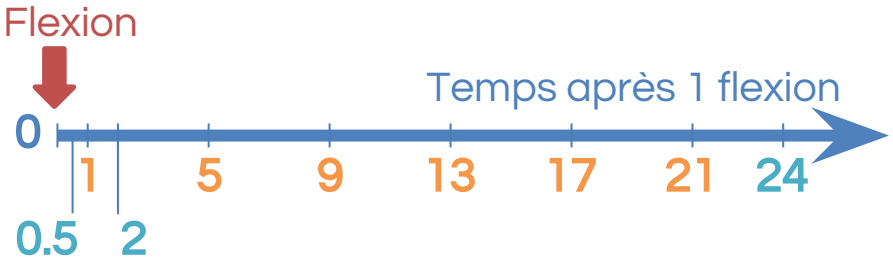
Exemple : Cluster n°1 - *PtaZFP2*





→ **10 mesures** de l'expression par gènes représentants

2. Augmentation artificielle du nombre de mesures



Échantillons récoltés pour :

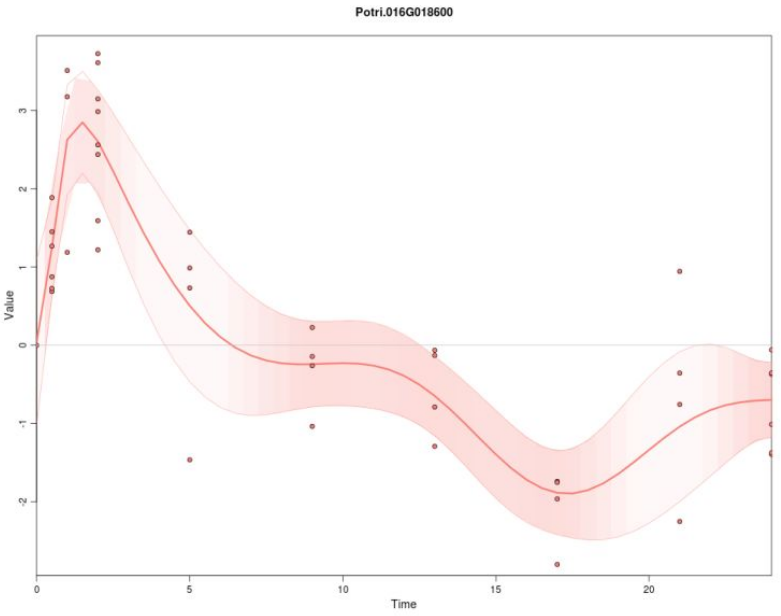
L'approche transcriptomique

L'approche qPCR

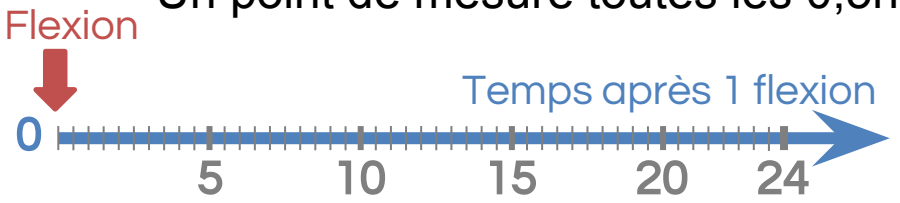
→ 10 points de mesures et seulement 7 régulièrement espacés

Augmentation artificielle du nombre de mesures en lissant les données

Méthode : **Non Stationary Gaussian Process** (package R : nsgp - Heinonen et al. 2015)



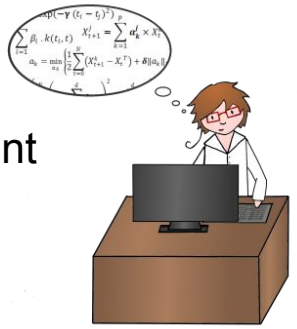
Un point de mesure toutes les 0,5h dans les 24h suivant la flexion



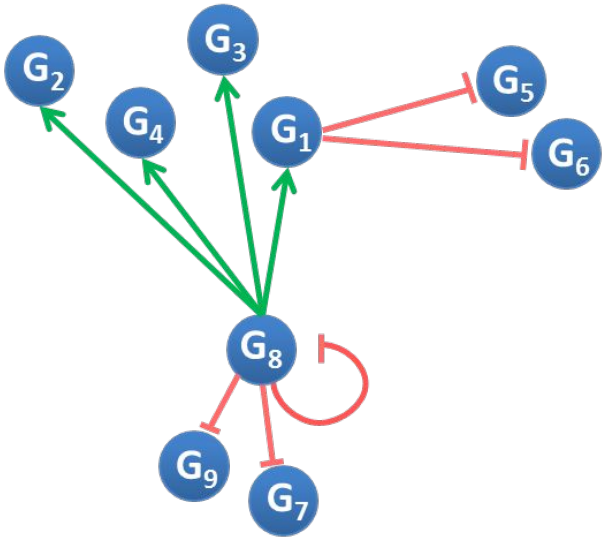
→ **49** points de **mesure** régulièrement espacés

3. Modèle modulaire : MMSIC

- Chaque gène appartient à un **module de co-régulation**
- Les gènes d'un **même module** partagent les **mêmes régulateurs**
- L'**expression** d'un **gène** à un **temps donné** dépend du **module** auquel appartient ce gène et de l'**expression** des **gènes régulateurs** de ce module

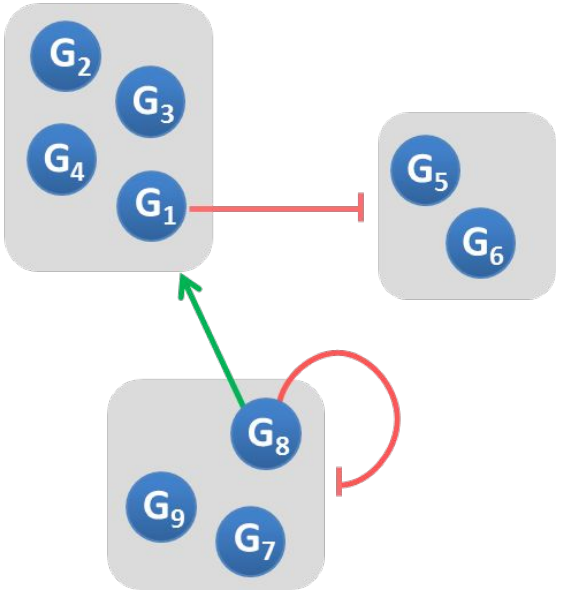


Linéaire simple



Arcs possibles : 9 x 9

MMSIC (3 modules)



Arcs possibles : 3 x 9
Modules d'appartenance : 3 x 9

Moins de paramètres à calculer avec MMSIC (nécessite moins de données)

Inconvénient : déterminer un **nombre de modules**

3. Modèle modulaire : MMSIC (F. d'Alché-Buc, J. Bedo)

Mixture of linear autoregressive **M**odels with hilbert **S**chmidt **I**ndependent **C**omponent

$$p(x_t^j|\mathbf{x}_{t-1},\theta) = \sum_{k=1}^{K} \pi_k^j \mathcal{N}(x_t^j|f_k(P_k\mathbf{x}_{t-1}),\sigma_k^2)$$

fonction linéaire (ou non)

$$f_k(\mathbf{x}) = \alpha_k^T \mathbf{x}$$

x_t^j expression du gène j au temps t

$\mathbf{x}_{t-1}$ expression de tous les gènes au temps précédent

π_k^j probabilité j ∈ k

P_k programme de régulation de k (matrice diagonale)

α_k poids des régulateurs sur l'expression des gènes de k

Méthode de **maximum de vraisemblance** pour trouver $\{P_k, \alpha_k, \sigma_k, \pi_k\}$

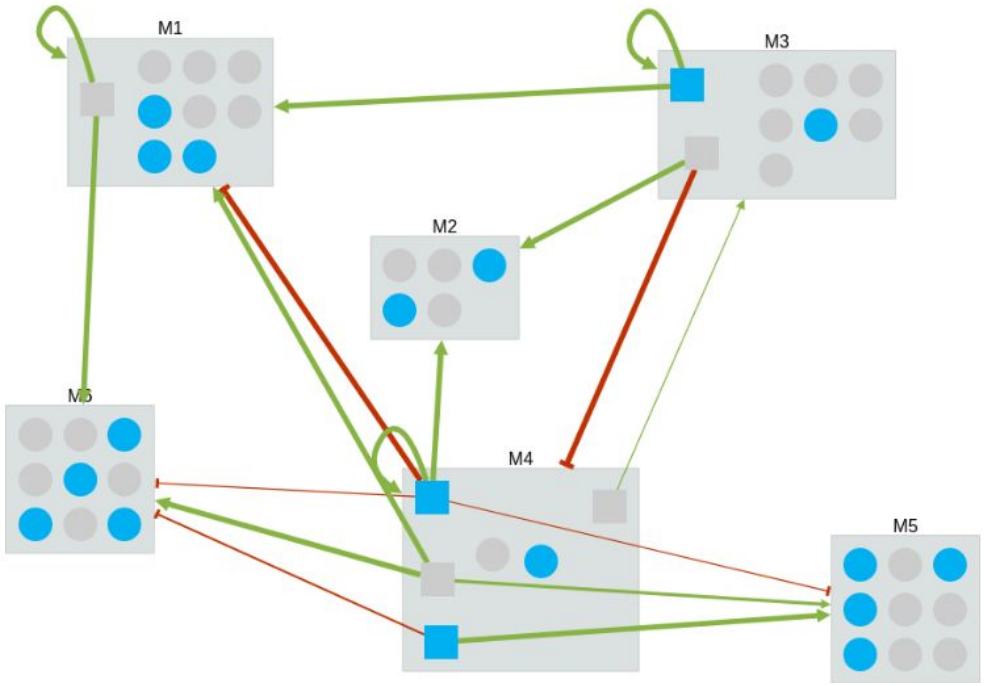
+ Terme de **pénalité HSIC** afin de favoriser la **diversité des programmes de régulation** des modules : $\sum_{k\neq\ell} \mathrm{hsic}(X, P_k, P_\ell)$

mesure l'indépendance entre les projections obtenues à partir de 2 programmes de régulations différents

3. Modèle modulaire : MMSIC

Exécution de MMSIC sur 47 gènes x 49 mesures
Modèle exécuté 1000 fois pour un nombre de modules allant de 1 à 10

→ 10 000 solutions produites



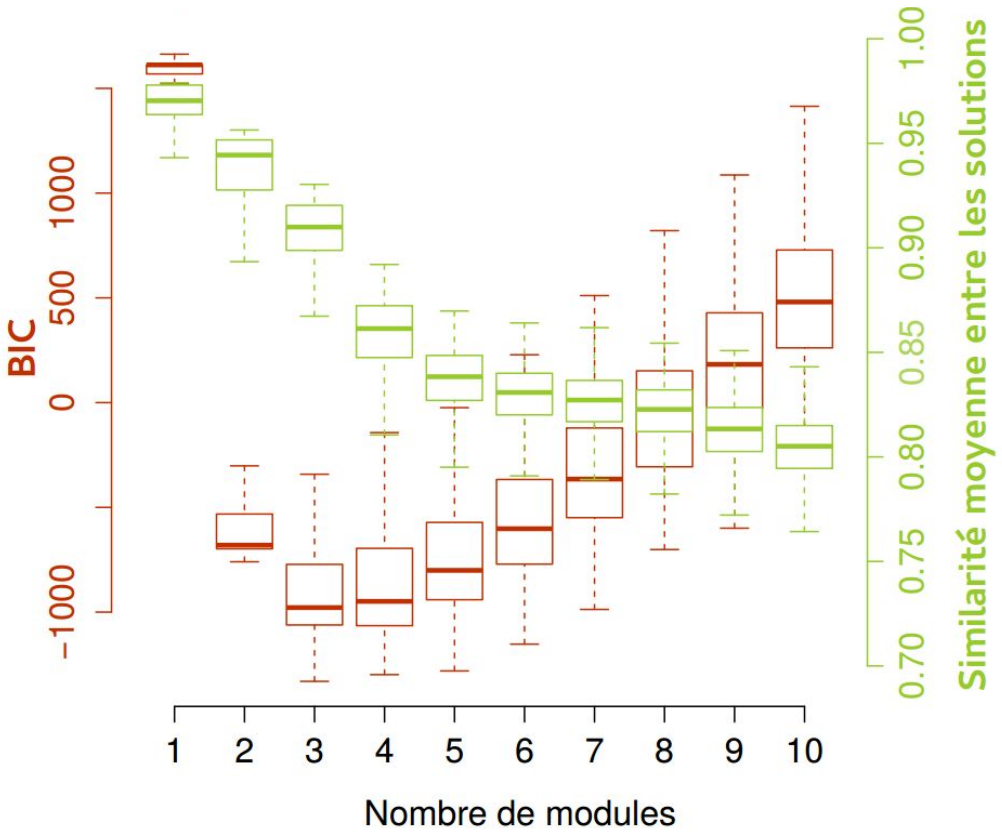
Gènes
Régulateur
Non-régulateur
FT

Relations
Induction
Répression

→ *Comment choisir les meilleures solutions ?*

Choix du nombre de modules

BIC et similarité entre les solutions du modèle MMSIC pour différents nombres de modules



BIC : évalue la vraisemblance du modèle et la complexité du modèle

Similarité : évalue la ressemblance entre les différentes solutions

- similarité entre les **modules** trouvés pas 2 solutions

$$S_{mod}(\theta_1,\theta_2)=\frac{2}{p(p-1)}\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\left(1-\left|M_1(i,j)-M_2(i,j)\right|\right)$$

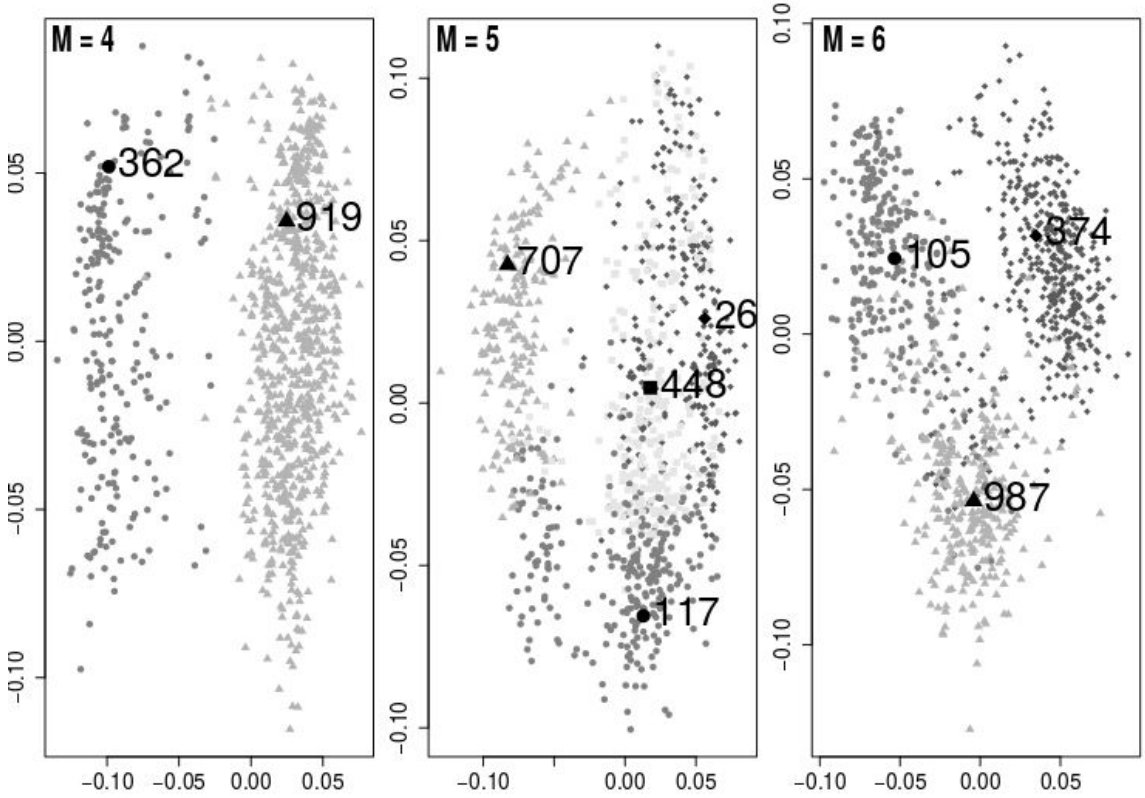
- similarité entre les **régulations** trouvées par deux solutions

$$S_{rég}(\theta_1,\theta_2)=\frac{1}{p}\sum_{i=1}^{p}h_i(\theta_1,\theta_2)$$

Sélection des solutions composées de **4, 5 ou 6 modules** (**3 000 solutions** sur les 10 000 solutions possibles)

4. Sélection des solutions

Choix des solutions



Identification par **clustering** de **groupes de solutions** proches pour chaque nombre de modules

Sélection des **solutions médianes** de ces groupes de solutions

4 modules :
M4 s362 | M4 s919

5 modules :
M5 s26 | M5 s117 | M5 s448 | M5 s707

6 modules :
M6 s105 | M6 s374 | M6 s987

Sélection de **9 solutions** parmi les 3000 solutions possibles

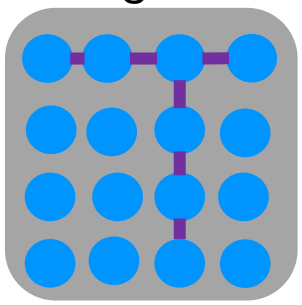
4. Discrimination des solutions par des données biologiques

1- Evaluation de la co-expression dans les modules

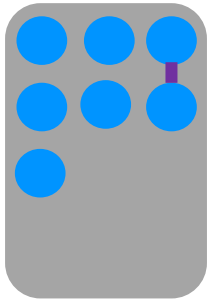
Les modules représentent des gènes co-régulés → ***existe-t-il des relations de co-expression entre ces gènes décrites dans la littérature ?***

Ex : M4 s 362

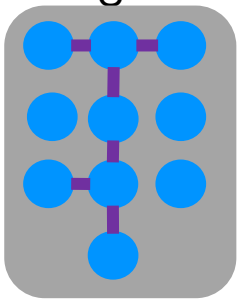
16 gènes



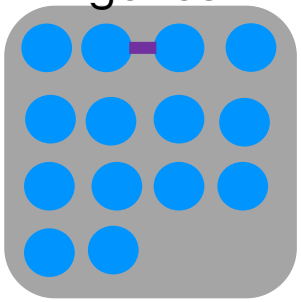
7 gènes



10 gènes



12 gènes



Co-expression retrouvée dans la littérature

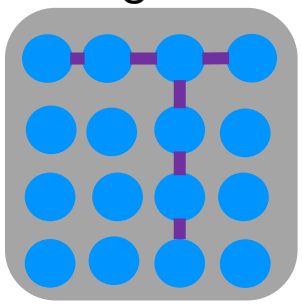
4. Discrimination des solutions par des données biologiques

2- Enrichissement en Gene Ontology

Les modules représentent des gènes co-régulés potentiellement impliqués dans un même processus biologique → ***existe-t-il un enrichissement en Gene Ontology dans les modules?***

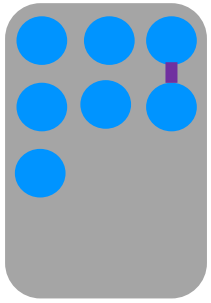
Ex : M4 s 362

16 gènes

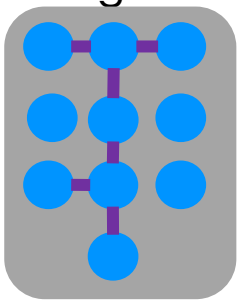


1BP
1CC

7 gènes

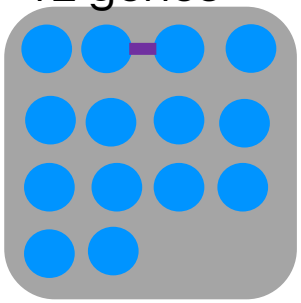


10 gènes



4 BP

12 gènes



19 BP
5 MF
5 CC

Co-expression retrouvée dans la littérature

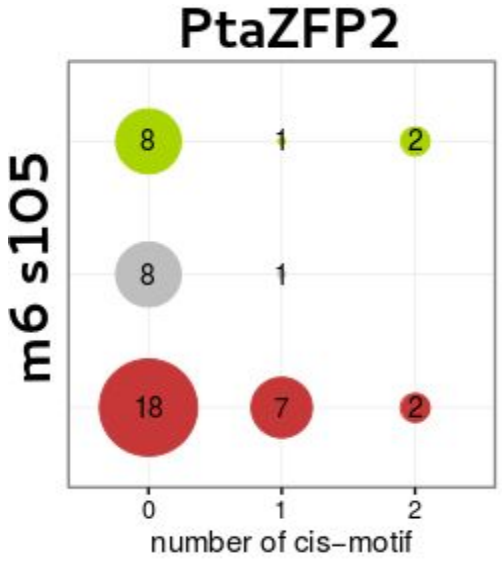
Sur-représentation GO
Biological Process
Molecular Function
Cellular Component

→ Validation de 3 (voir 4) modules pour la solution M4 s362

4. Discrimination des solutions par des données biologiques

3 - Evaluation des régulations transcriptionnelles

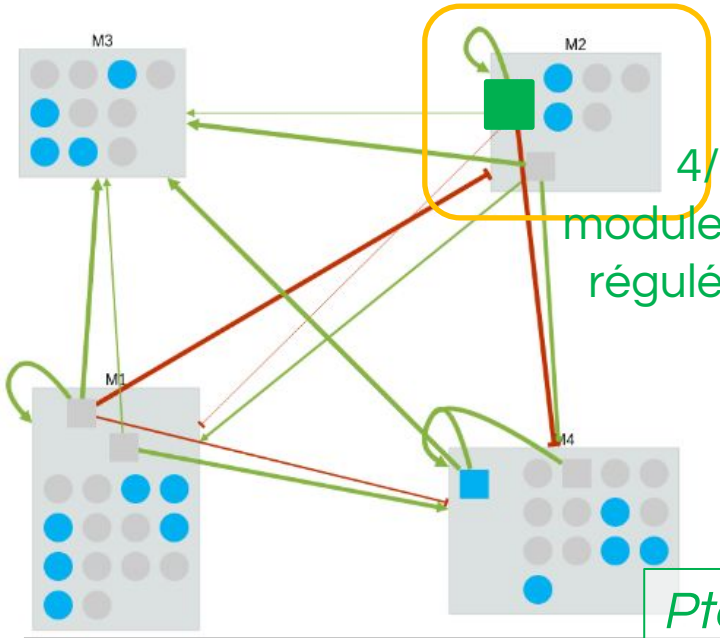
*Est-ce que les gènes régulés par des facteurs de transcription dans les solutions possèdent des **cis-motifs spécifiques de ces facteurs de transcription ?***



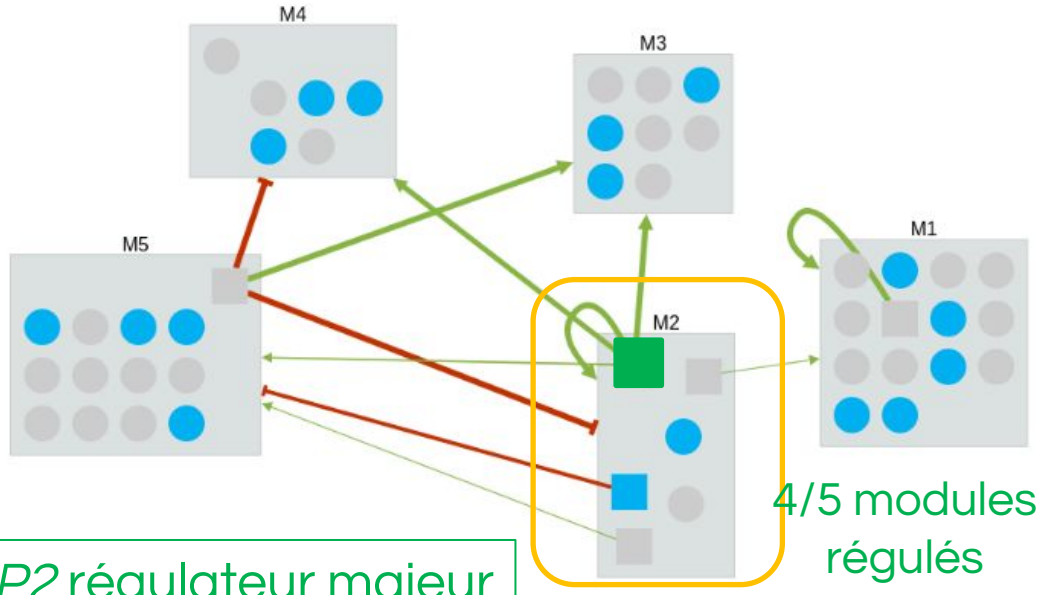
Les **gènes** possédant le **cis-motif spécifique** de *PtaZFP2* sur leur promoteur sont **majoritairement régulés** par *PtaZFP2* dans la solution m6 s105

→ A partir de ces **3 critères**, **sélection** des **solutions** dont le plus d'**éléments** (module, régulation) sont **validés**

Solutions sélectionnées



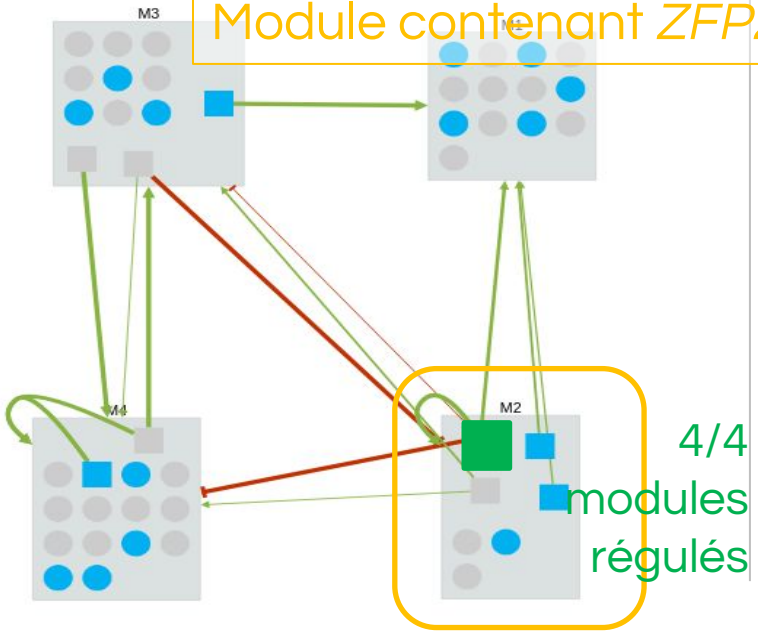
4/4 modules régulés



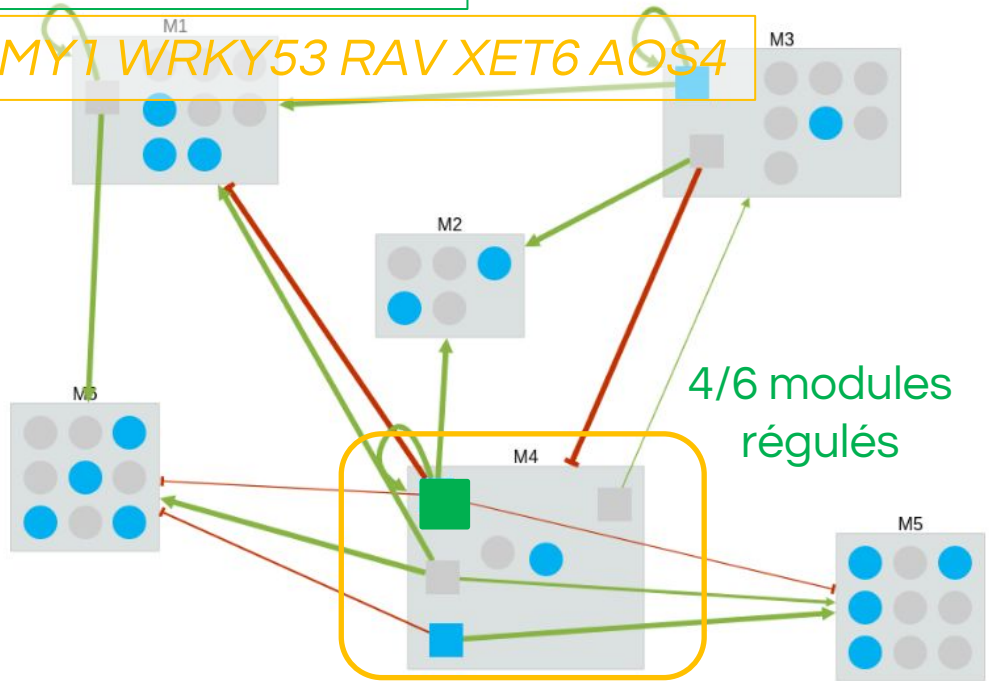
4/5 modules régulés

PtaZFP2 régulateur majeur

Module contenant *ZFP2 AMY1 WRKY53 RAV XET6 AOS4*



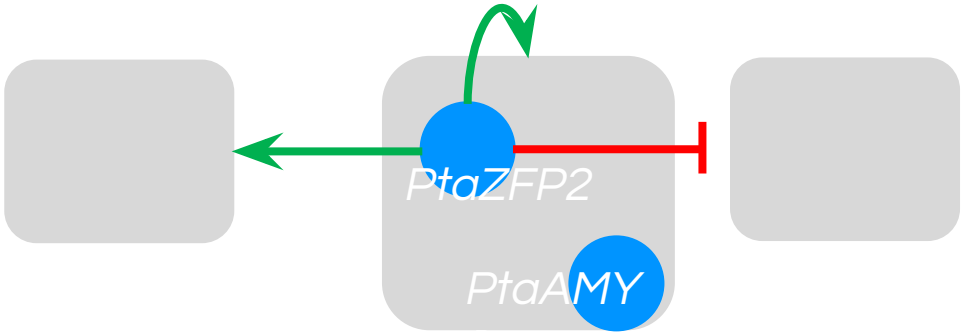
4/4 modules régulés

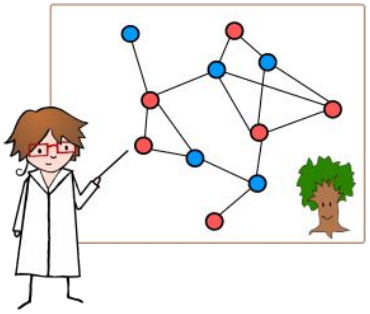


4/6 modules régulés

Interprétation biologique

Rôle central de PtaZFP2





Peuplier transgénique sur-exprimant *PtaZFP2* : PtaZFP2-OE #39

PtaAMY1 ↗ 0,5-13h après 1 flexion ↗ PtaZFP2-OE #39

→ Rôle de régulateur majeur de ***PtaZFP2***
cohérent avec les **observations biologiques**

Conclusion

Stratégie d'inférence

Nombre gènes >> Nombre de mesures

Validation des résultats

Pas de réseau existant pour comparer

Sélection de gènes	Augmentation des données	Modèle MMSIC	Sélection nombre de modules et meilleures solutions	Discrimination des solutions

- **Clustering** + **GO** pour identifier des gènes representants

Données considérées comme indépendantes (*clustering spectral*)

- **Augmentation** par lissage du **nombre de données**

Composition des **modules varie peu** avec ou sans interpolation
Réduit drastiquement le **nombre de régulateurs**

- **Modèle de mélange linéaire**

Moins coûteux en données
Réponse à la flexion : **plusieurs états stables** possibles → Modèle non linéaire

Conclusion

Stratégie d'inférence

Nombre gènes >> Nombre de mesures

Validation des résultats

Pas de réseau existant pour comparer

Sélection de gènes	Augmentation des données	Modèle MMSIC	**Sélection nombre de modules et meilleures solutions**	**Discrimination des solutions**

Utilisation de **données biologiques** disponibles dans la littérature

Modules

Relation de co-expression + Gene Ontology

Utiliser d'autres bases de données, *ex* : KEGG

Régulations

Recherche de cis-motifs

/!\ Transposition à d'autres espèces /!\

Perspectives

Validation des solutions du modèle

- Tester les **cis-motifs** sur le génome de ***P. tremula x alba***

- Mise en place des méthodes de **simple hybride** ou **double hybride** permettant de tester et valider directement les interactions prédites par les solutions du modèle (au niveau physique)

- Comparer l'expression des gènes prédite par le modèle à celle mesurée dans le cas de la transgénèse :

Transgénèse virtuelle vs. **Plante trannsgénique**

Amélioration de la stratégie d'inférence

- **Modèle non linéaire** (combiné à du *bootstrap*)

Remerciements

Membres du laboratoire PIAF

INRA SCIENCE & IMPACT

UNIVERSITÉ Clermont Auvergne

U.M.R. PIAF

Nathalie Leblanc-Fournier
Mélanie Decourteix
Bruno Moulia
Jérôme Franchel

Membres des laboratoires IBISC puis LTCI

ibisc

cnrs

LTCI
UMR 5141

TELECOM ParisTech

Florence d'Alché-Buc

Merci de votre attention

lise.pomies@inra.fr