# Différentes sémantiques pour les règles : une définition aux propriétés intéressantes

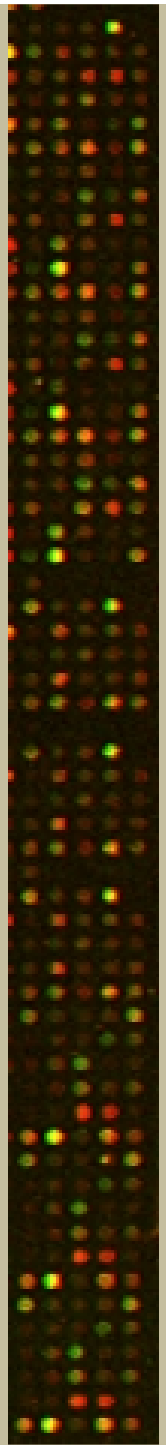
Marie Agier

LIMOS - Université Clermont-F[d] II

Société DIAGNOGENE

Centre de Lutte contre le Cancer Région Auvergne

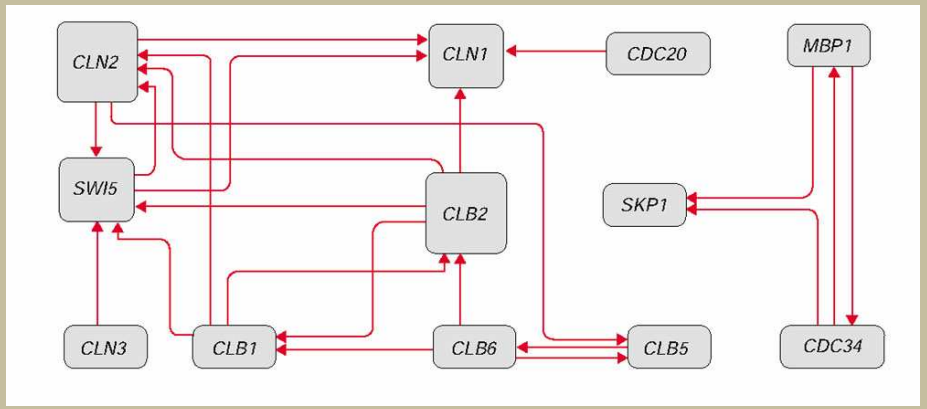
Groupe 3.4 GDR I3 – Journée Fouille de Données – 20/05/05



# Motivations

- ***Données d'expression de gènes***
  - Peu de lignes
  - Beaucoup de colonnes
  - Données bruitées
- ***Différents objectifs***
  - Recherche de gènes différentiellement exprimés entre plusieurs types de cellules
  - Construction de modèles de prédiction de tumeurs
  - Découverte de réseaux de gènes
  - …

# Réseaux de gènes

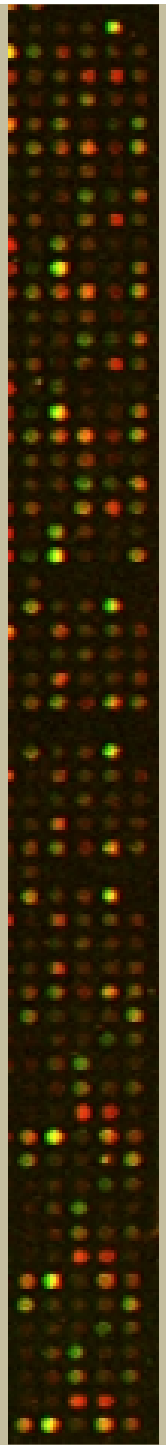



- Trouver les ***relations*** entre les gènes i.e. comment le niveau d'expression de chaque gène affecte le niveau d'expression des autres gènes

- Quel ***sens*** donner aux relations entre les gènes ?

  $\Rightarrow$ Notion de ***règles*** entre gènes

  $\Rightarrow$ Notion de ***sémantiques*** pour les règles



# Plan

- Exemples de sémantiques pour les règles
- Cadre proposé
- Intérêts pour la découverte des réseaux de gènes
- Conclusion & Perspectives

# Contexte

- Représentation tabulaire des données

| r | $A_1$ | $A_2$ | $A_3$ | … | … | $A_n$ |
|---|---|---|---|---|---|---|
| $t_1$ | 0.0 | 1.9 | -1.9 | … | … | -0.5 |
| $t_2$ | 1.7 | 1.5 | 1.2 | … | … | 1.6 |
| $t_3$ | 1.8 | -0.7 | 1.3 | … | … | 1.7 |
| … | … | … | … | … | … | … |
| $t_m$ | -1.7 | -1.4 | 0.9 | … | … | -0.2 |

- Une ***règle*** sur U={$A_1$, $A_2$, ..., $A_n$} est une expression de la forme X $\Rightarrow$ Y avec X, Y $\subseteq$ U
- La ***sémantique*** d'une règle sur U est la *signification*, le *sens* que l'on souhaite donner à cette règle

- Remarques :
  - Notion très connue et très étudiée
  - Plusieurs sémantiques ont été proposées (règles d'association, dépendances fonctionnelles…)
  - Généralement faciles à interpréter

# Mesures d'intérêt

- Ex : confiance, erreur $g_3$, support, khi-deux, taux informationnel…
- Permettent d'ajouter des règles ***approximatives*** (presque vraies) et de se limiter aux règles ***les plus intéressantes***
- Remarques :
  - Ces indices sont très utiles
  - S'appliquent à plusieurs sémantiques
  - Renseignent sur la ***pertinence*** de la règle et non sur le ***sens*** de la règle
- Choix :

  ⇒ ***N'entrent pas dans la définition des sémantiques***

# Exemples de sémantiques pour $X \Rightarrow Y$

- **Dépendances fonctionnelles**

  $\forall t_1, t_2 \in r$, si $\forall A \in X$, $t_1[A] = t_2[A]$ alors $\forall B \in Y$, $t_1[B] = t_2[B]$

- **Règles d'association (exactes sans minsup)**

  $\forall t \in r$, si $\forall A \in X$, $t[A] = 1$ alors $\forall B \in Y$, $t[B] = 1$

- **Distance entre les valeurs**

  $\forall t_1, t_2 \in r$, si $\varepsilon_1 \leq d(t_1[X], t_2[X]) \leq \varepsilon_2$ alors $\varepsilon_1 \leq d(t_1[Y], t_2[Y]) \leq \varepsilon_2$

# Exemples de sémantiques pour $X \Rightarrow Y$

- **Sémantique $s_1$**

  $\forall t_1, t_2 \in r$, si $\forall A \in X$, $\varepsilon_1 \leq |t_1[A] - t_2[A]| \leq \varepsilon_2$ alors $\forall B \in Y$, $\varepsilon_1 \leq |t_1[B] - t_2[B]| \leq \varepsilon_2$

- **Sémantique $s_2$**

  $\forall t_i, t_{i+1} \in r$, si $\forall A \in X$, $\varepsilon_1 \leq t_{i+1}[A] - t_i[A] \leq \varepsilon_2$ alors $\forall B \in Y$, $\varepsilon_1 \leq t_{i+1}[B] - t_i[B] \leq \varepsilon_2$

- **Sémantique $s_3$**

  $\forall t \in r$, si $\forall A \in X$, $t[A] = 1$ alors $\forall B \in Y$, $t[B] = 0$

# Sémantiques pour les règles

⇒ ***De nombreuses sémantiques peuvent être définies***

⇒ ***Pour un contexte donné, plusieurs sémantiques peuvent être intéressantes***

- **Proposition**
  - Approche offrant la possibilité de choisir parmi ***plusieurs*** sémantiques
  - Possibilité de définir des ***nouvelles*** sémantiques

⇒ Définition générique d'une sémantique

⇒ Cadre général issu de l'inférence des DF

# Définition générique d'une sémantique

- Une ***sémantique s*** pour les règles peut se définir à partir :
  - de deux prédicats spécifiant une condition sur les données
  - d'une contrainte spécifiant la portée du prédicat

Exemple : $\forall t \in r$, si $\forall A \in X, t[A] = 1$ alors $\forall B \in Y, t[B] = 0$

<u>Définition 1 :</u> $r \models_s X \Rightarrow Y$ ssi $\forall\ r' \subseteq r$ vérifiant $d_c(r')$, si $Pred_1(X,r')$ est vrai alors $Pred_2(Y,r')$ est vrai

Exemple : $d_c(r') = [r' = \{t\}$ avec $t \in r]$

$Pred_1(X, \{t\}) = [\forall A \in X, t[A] = 1]$

$Pred_2(X, \{t\}) = [\forall A \in X, t[A] = 0]$

# Sémantique bien-formée

- Cadre général issu de l'inférence des DF, basé sur le système d'axiomes d'Armstrong
- Intérêts pratiques :
  - Possibilité de ***raisonner*** syntaxiquement sur les règles sans accéder aux données
  - Possibilité de générer des ***couvertures*** des règles
  - Utilisation des nombreuses ***propriétés algorithmiques*** existantes pour la découverte des règles

Definition : Une sémantique *s* est ***bien-formée*** si le système d'axiomes d'Armstrong est juste et complet pour *s*

# Système d'axiomes d'Armstrong

- Système d'axiomes d'Armstrong pour un ensemble de règles F défini sur U :
  - (réflexivité) si $X \subseteq Y \subseteq U$ alors $F \vdash Y \Rightarrow X$
  - (augmentation) si $F \vdash X \Rightarrow Y$ et $W \subseteq U$, alors $F \vdash XW \Rightarrow YW$
  - (transitivité) si $F \vdash X \Rightarrow Y$ et $F \vdash Y \Rightarrow Z$ alors $F \vdash X \Rightarrow Z$

- Système d'axiomes ***juste*** et ***complet*** :
  - Justesse : les axiomes ne génèrent pas de règles incorrectes
  - Complétude : ils génèrent bien toutes les règles correctes

# Question

- Toutes les sémantiques définies avec la déf 1 sont-elles bien-formées ?

  Exemple : $\forall t \in r$, si $\forall A \in X$, $t[A] = 1$ alors $\forall B \in Y$, $t[B] = 0$

  $\Rightarrow$ Axiome de réflexivité non juste

- Comment éviter de prouver la justesse et la complétude du système d'Armstrong pour chaque sémantique ?

# Restrictions syntaxiques

Définition 2 : $r \models_s X \Rightarrow Y$ ssi $\forall\ r' \subseteq r$ vérifiant $d_c(r')$, si $\forall A \in X$, Pred(A,r') est vrai alors $\forall B \in Y$, Pred(B,r') est vrai

- $Pred_1$ et $Pred_2$ sont équivalents
- *Pred* spécifie une condition sur un attribut $A \in U$ et non plus sur un ensemble d'attributs $X \subseteq U$

Résultat : Soit *s* une sémantique entrant dans le cadre de la déf 1. La sémantique *s* est bien-formée ***si et seulement si*** la sémantique *s* entre dans le cadre de la déf 2

Exemple : $d_c(r') = [r' = \{t\}$ avec $t \in r]$

$Pred_1(A, \{t\}) = [t[A] = 1]$ $\Rightarrow$ Sémantique pas bien-formée

$Pred_2(A, \{t\}) = [t[A] = 0]$

# Exemples de sémantiques

- **Dépendances fonctionnelles**

  $d_c(r') = [r' = \{t_1,t_2\}$ avec $t_1,t_2 \in r]$ et $Pred(A, \{t_1,t_2\}) = [t_1[A] = t_2[A]]$

  $\Rightarrow$ Sémantique bien-formée

- **Règles d'association (exactes et sans minsup)**

  $d_c(r') = [r' = \{t\}$ avec $t \in r]$ et $Pred(A, \{t\}) = [t[A] = 1]$

  $\Rightarrow$ Sémantique bien-formée

# Exemples de sémantiques

- **Distance entre les valeurs**

  $d_c(r') = [r' = \{t_1,t_2\}$ avec $t_1,t_2 \in r]$ et $Pred(X, \{t_1,t_2\}) = [\varepsilon_1 \leq d(t_1[X], t_2[X]) \leq \varepsilon_2]$

  $\Rightarrow$ Sémantique pas bien-formée

- **Sémantique $s_1$**

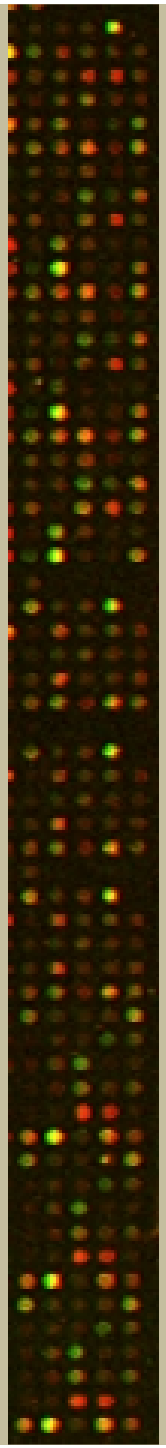
  $d_c(r') = [r' = \{t_1,t_2\}$ avec $t_1,t_2 \in r]$ et $Pred(A, \{t_1,t_2\}) = [\varepsilon_1 \leq |t_1[A] - t_2[A]| \leq \varepsilon_2]$

  $\Rightarrow$ Sémantique bien-formée

- **Sémantique $s_2$**

  $d_c(r') = [r' = \{t_i,t_{i+1}\}$ avec $t_i,t_{i+1} \in r]$ et $Pred(A, \{t_i,t_{i+1}\}) = [\varepsilon_1 \leq t_{i+1}[A] - t_i[A] \leq \varepsilon_2]$

  $\Rightarrow$ Sémantique bien-formée



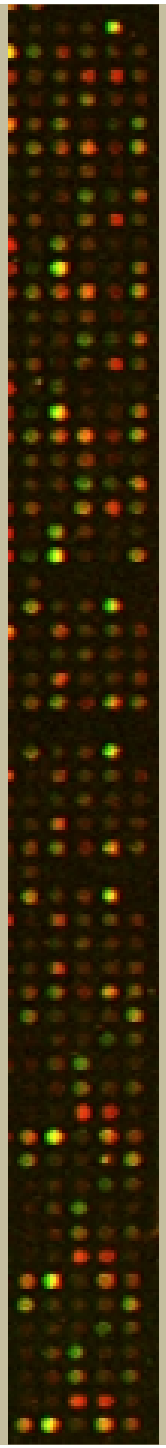
# Application aux réseaux de gènes

⇒ ***Proposer aux biologistes plusieurs sémantiques pour leurs réseaux***

⇒ ***Leur donner la possibilité de définir des nouvelles sémantiques en fonction de leurs objectifs***

- Remarques :
  - Le nombre de règles (i.e. la taille des réseaux) ne doit pas être grand
  - Importante phase de validation par les experts

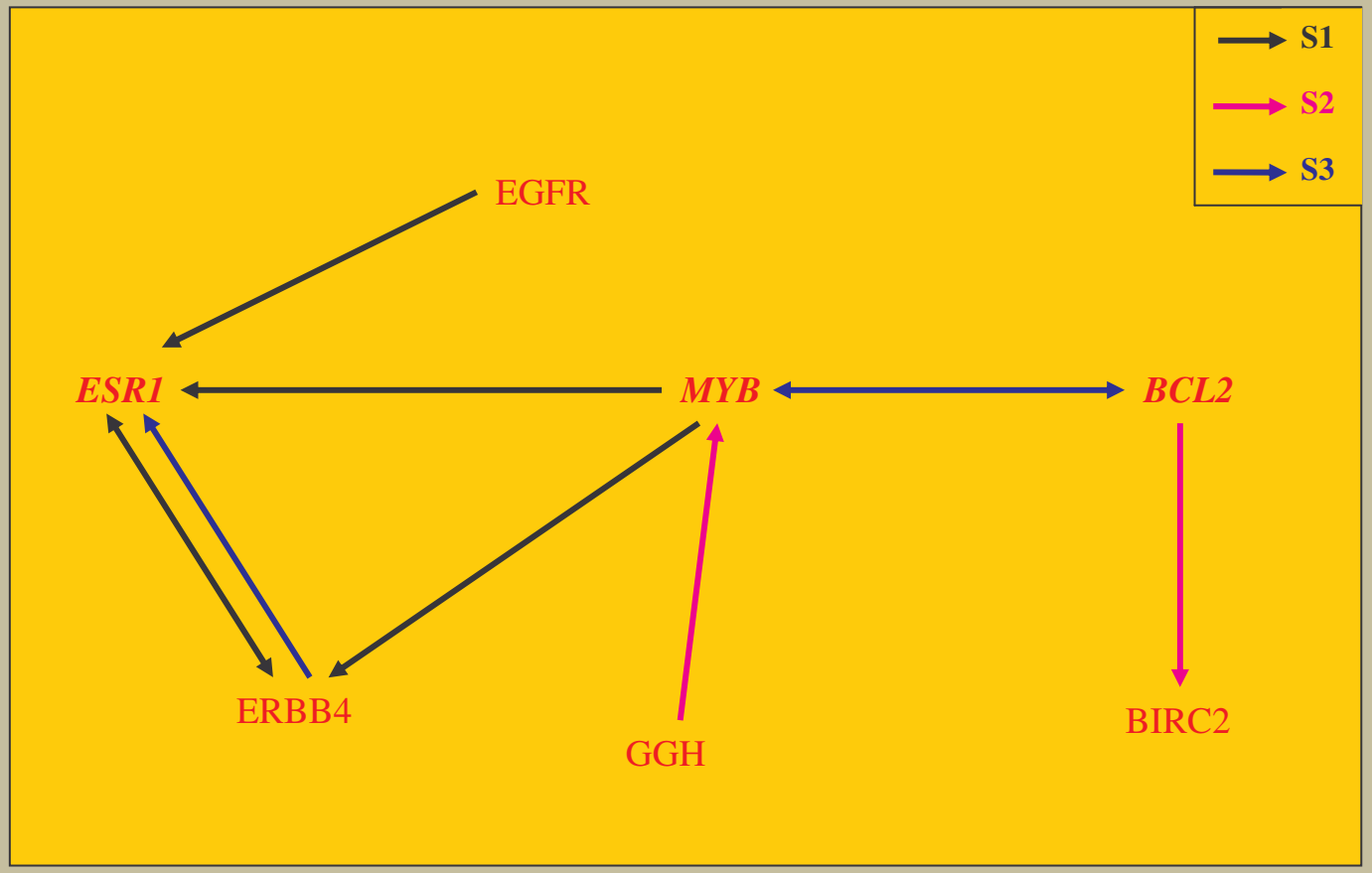
  ⇒ Demander aux experts les ***gènes centraux*** des réseaux

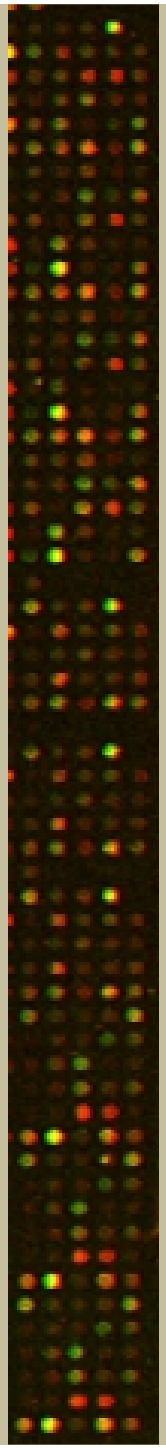


# Découverte des réseaux

- **Choix des gènes centraux des réseaux (pas plus de 2 ou 3)**
  - Les experts choisissent les gènes qui apparaîtront en partie gauche ou en partie droite des règles
- **Choix des sémantiques**
  - Les experts choisissent le sens des relations qu'ils veulent étudier dans les réseaux
- **Construction des réseaux**
  - "Soit *r* une relation et $s_1,\ldots,s_n$ des sémantiques bien-formées, découvrir les règles satisfaites dans *r* avec les sémantiques $s_1,\ldots,s_n$"
  - Plusieurs réseaux sont proposés, chacun avec une sémantique particulière
  - Un réseau global peut être proposé

# Exemple de réseau global

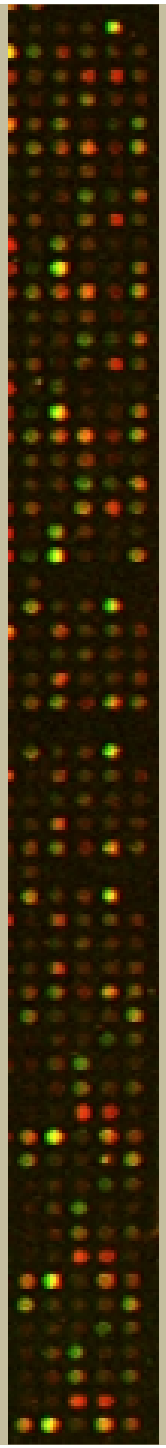




# Conclusion

- Approche offrant plusieurs sémantiques pour les règles
- Équivalence entre des restrictions syntaxiques sur la définition d'une sémantique et le fait qu'elle soit bien-formée
- Cadre identique pour toutes les sémantiques bien-formées
- Application à la découverte de réseaux de gènes

- Perspectives :
  - Finaliser l'outil de génération des réseaux
  - Validation des réseaux avec les experts
  - S'intéresser aux sémantiques qui ne sont pas bien-formées



# Merci de votre attention