Une approche transversale pour la prédiction de relations fonctionnelles entre gènes

Julie Chabalier

E.A. “Modélisation Conceptuelle des Connaissances Biomédicales”
Université de Rennes 1

UNIVERSITE RENNES 1

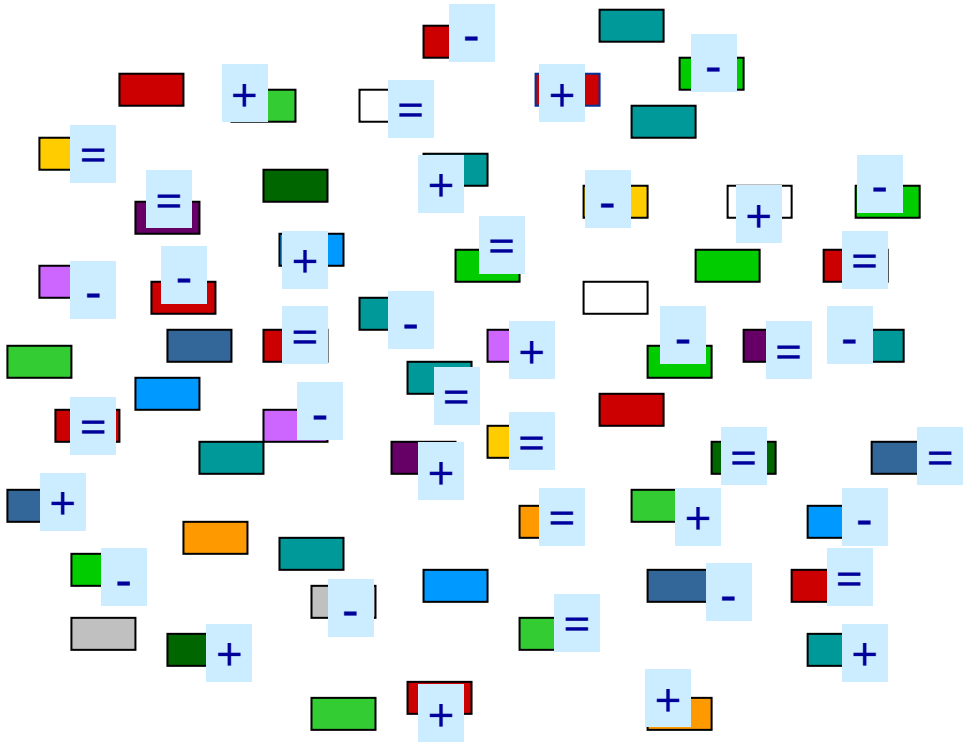
UMR 6061 - CNRS
Université Rennes 1

PLAN

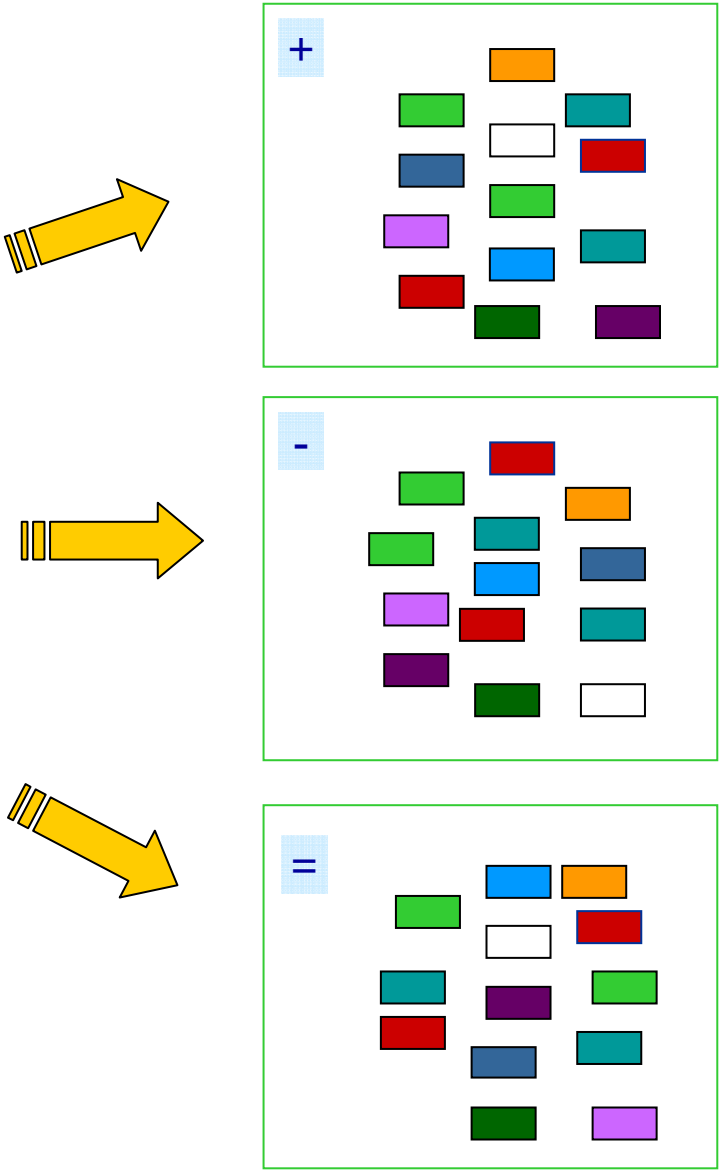
- **L'approche classique**
  - **Une approche en deux étapes**
- **L'analyse transversale**
  - **Annotation biomédicale**
  - **Regroupement des gènes**
  - **Similarité sémantique**
- **Application**
- **Conclusion**

L’approche classique: 2 étapes

Etape 1

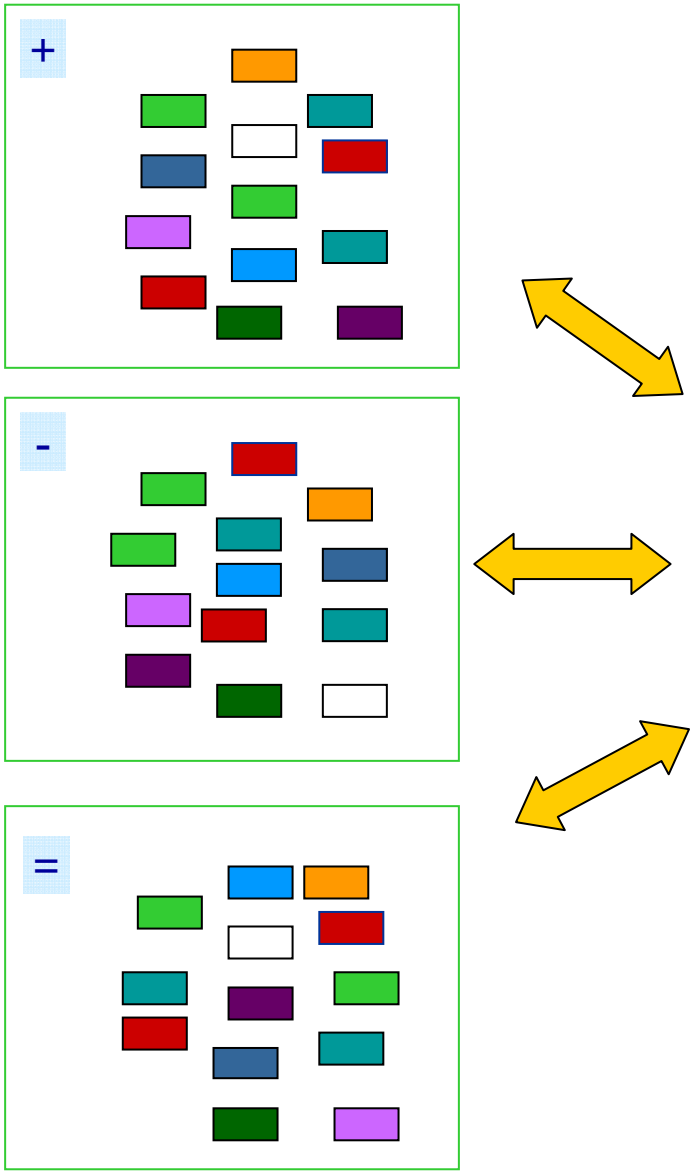


Groupes d’expression



L'approche classique: 2 étapes

Etape 2



Annotation fonctionnelle

Gene Ontology

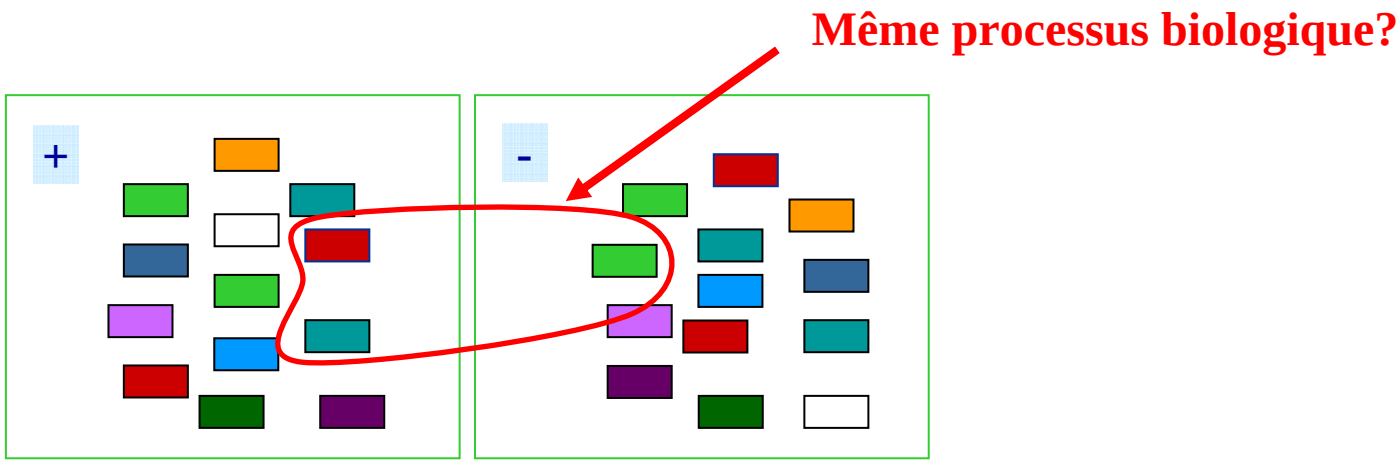
Interpretation biologique?

L'approche classique: 2 étapes

➢ **Interpretation biologique**

Regroupement des gènes par profil d'expression peut limiter l'espace d'interprétation des données

Mise en évidence de relations fonctionnelles difficile



L'analyse transversale

- **Combinaison de connaissances de nature différente**

Regroupement parallèle des gènes en fonction des connaissances :

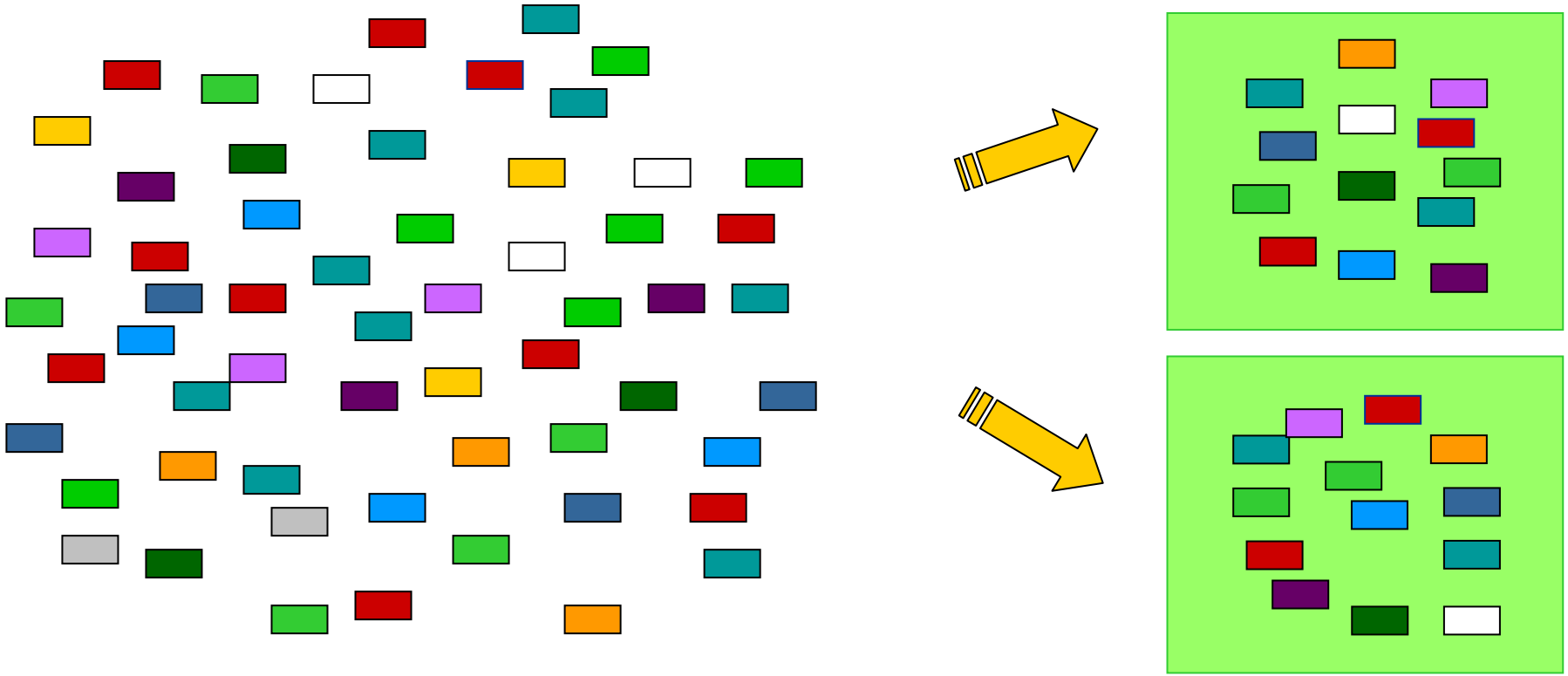
- **Expérimentales :** données d'expression
- **Biologiques :** Gene Ontology (GO)
- **Médicales :** Unified Medical Language System (UMLS)

→ **Prediction de gènes fonctionnellement reliés**

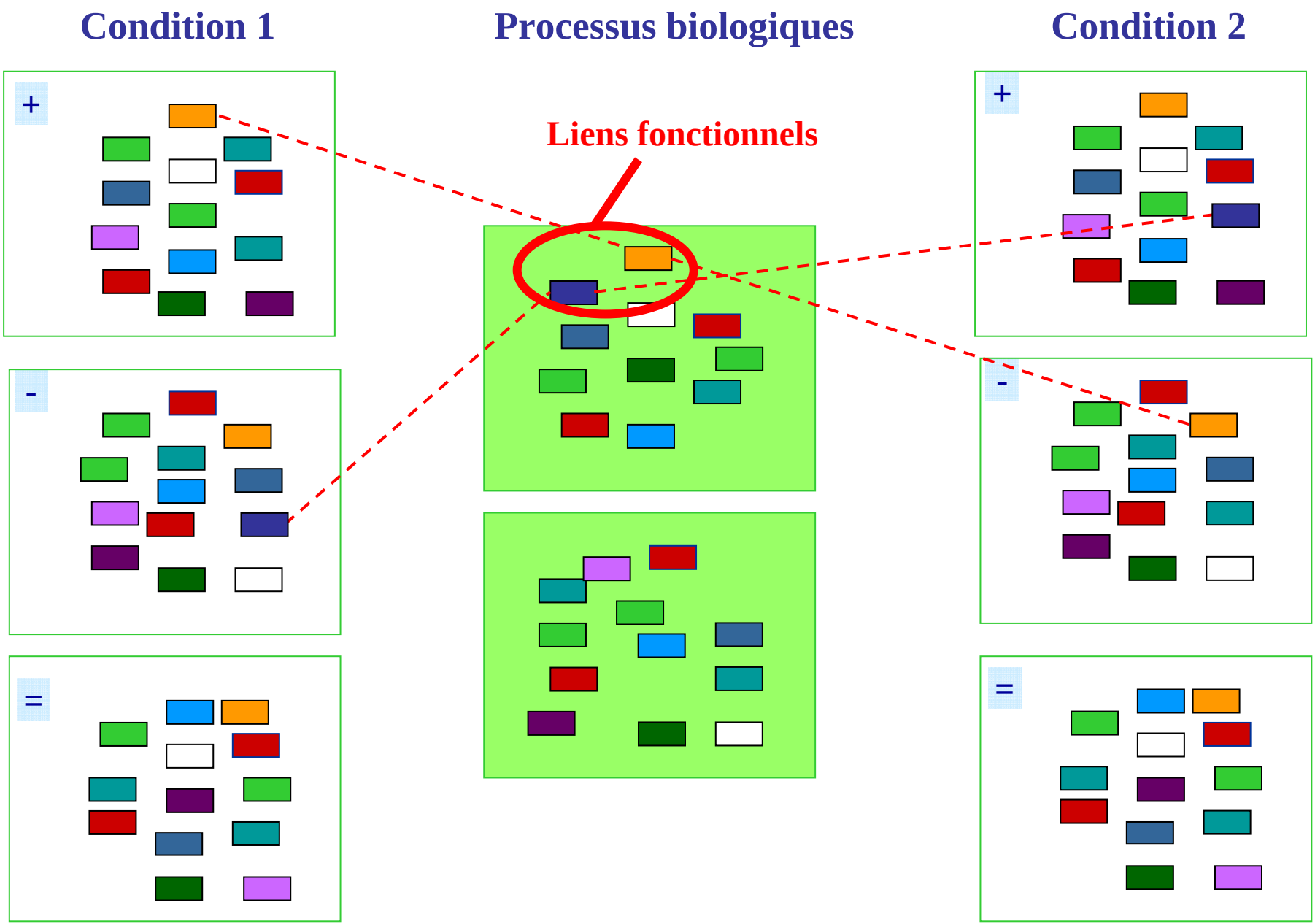
L’analyse transversale

Groupes fonctionnels

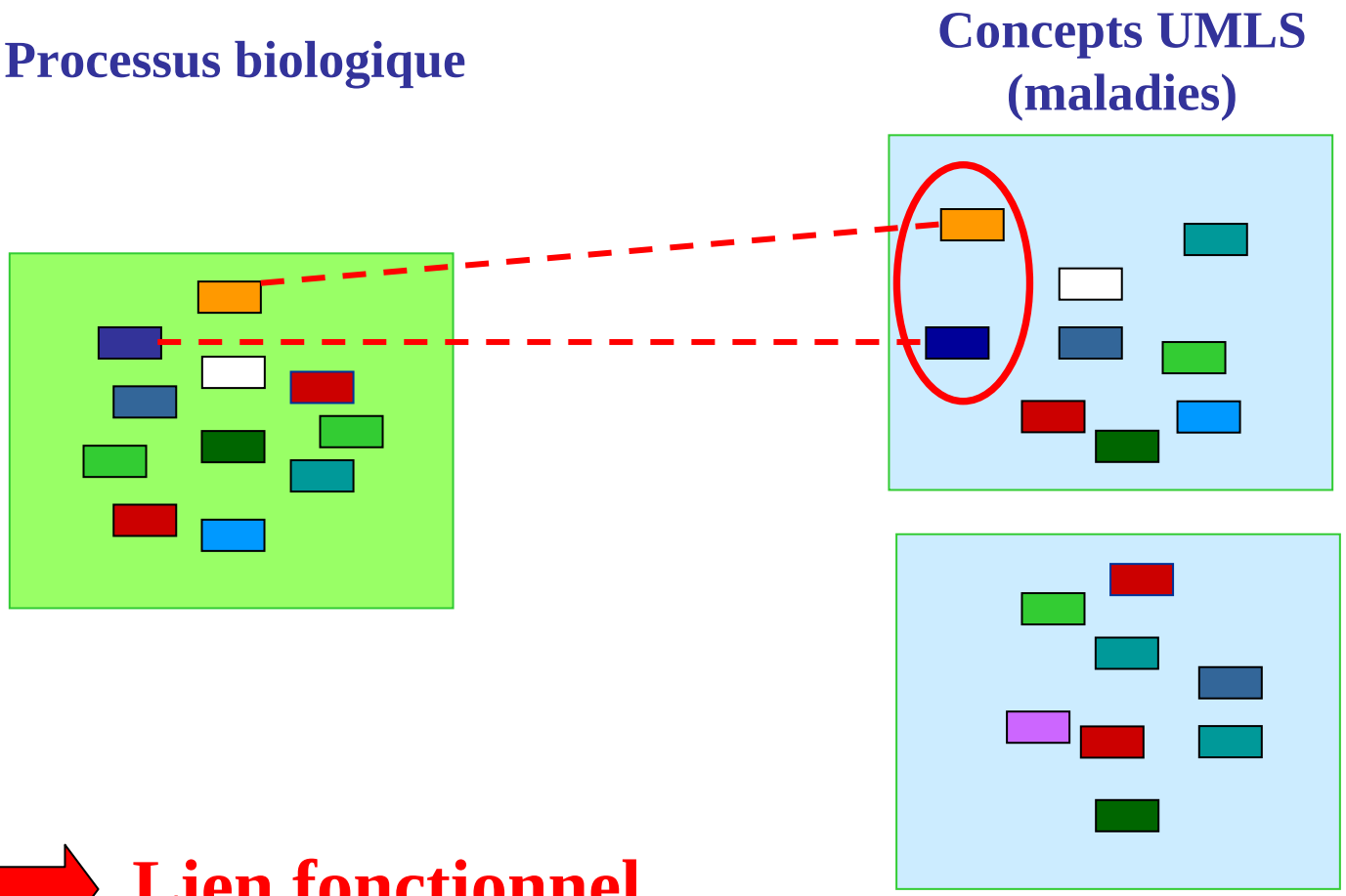
Processus biologiques GO



L’analyse transversale



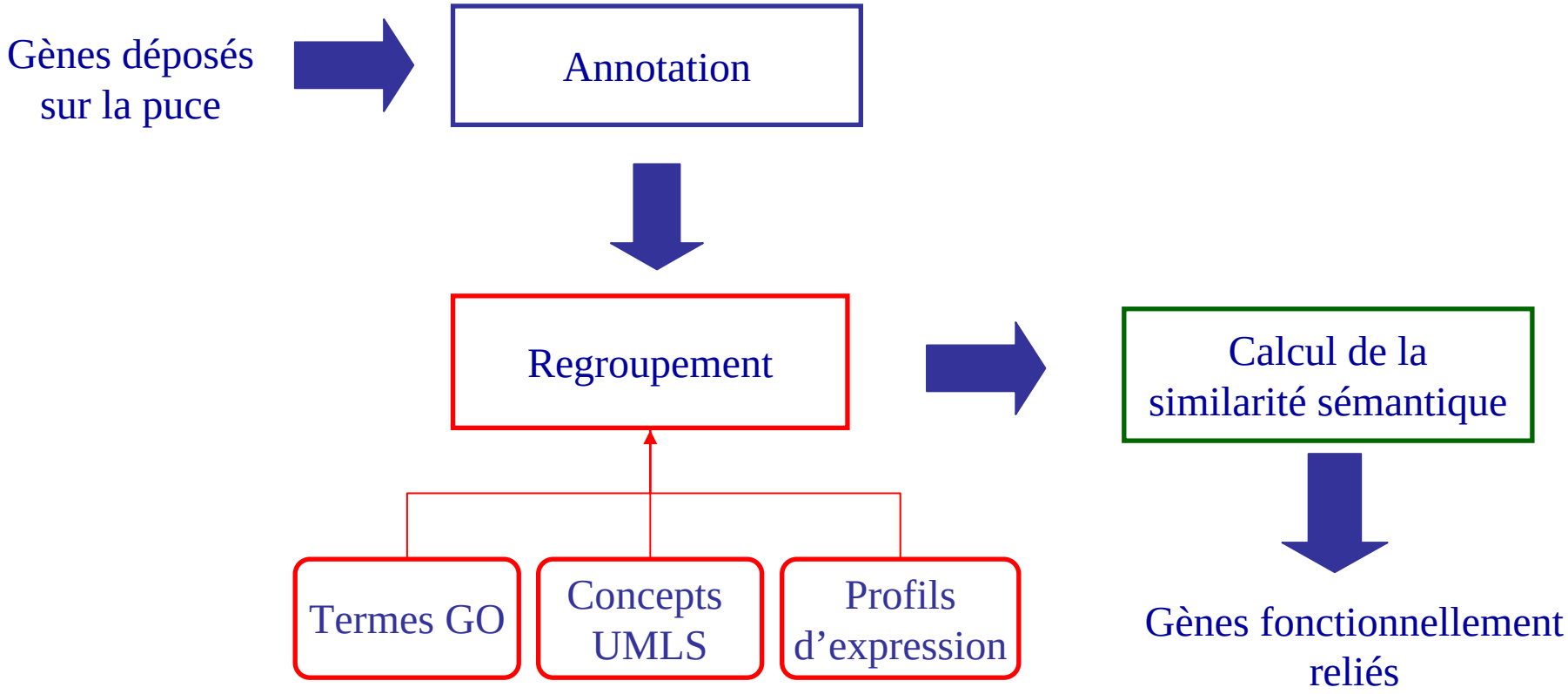
L’analyse transversale



Lien fonctionnel

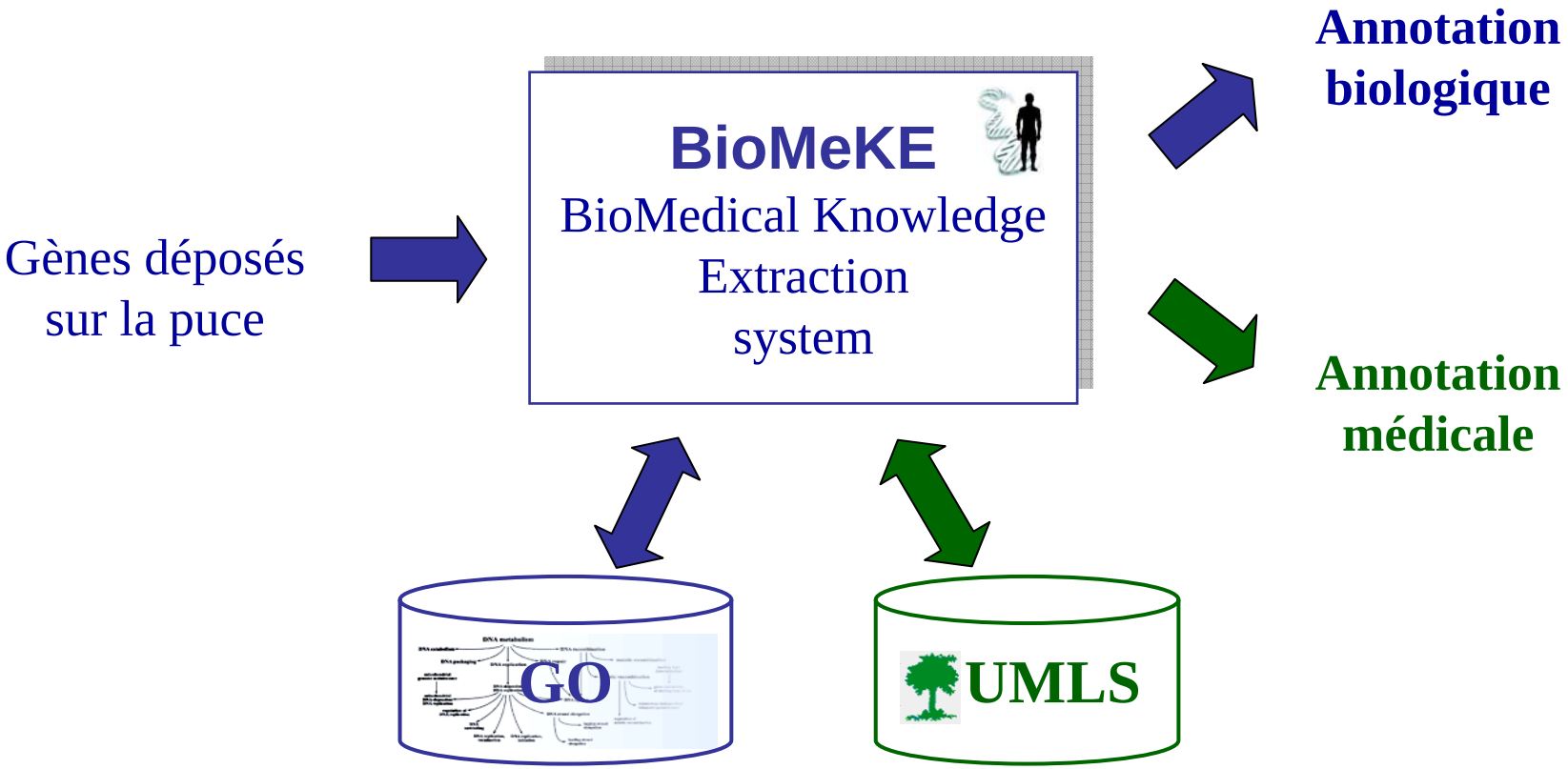
L'analyse transversale

- Stratégie



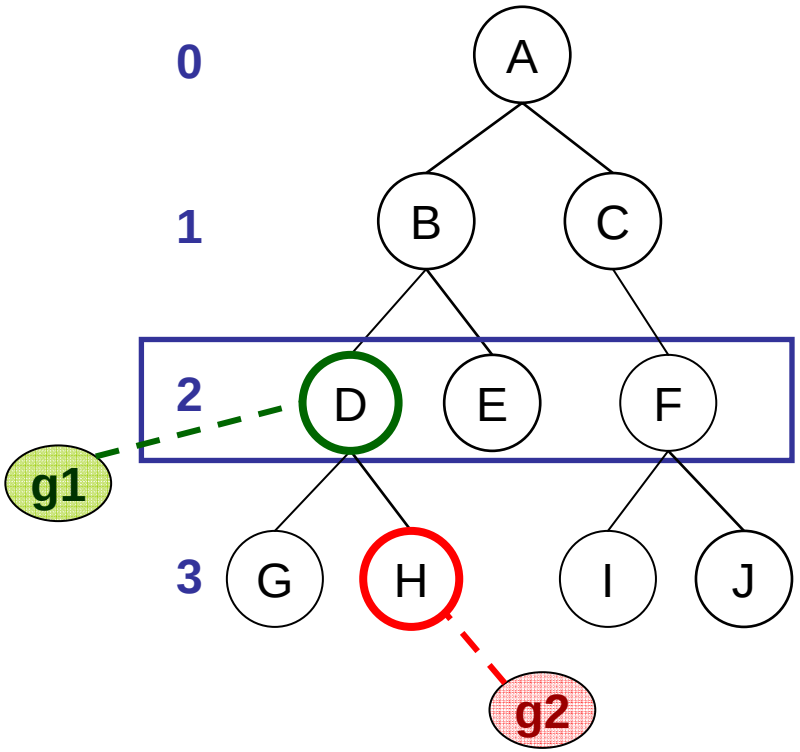
Annotation

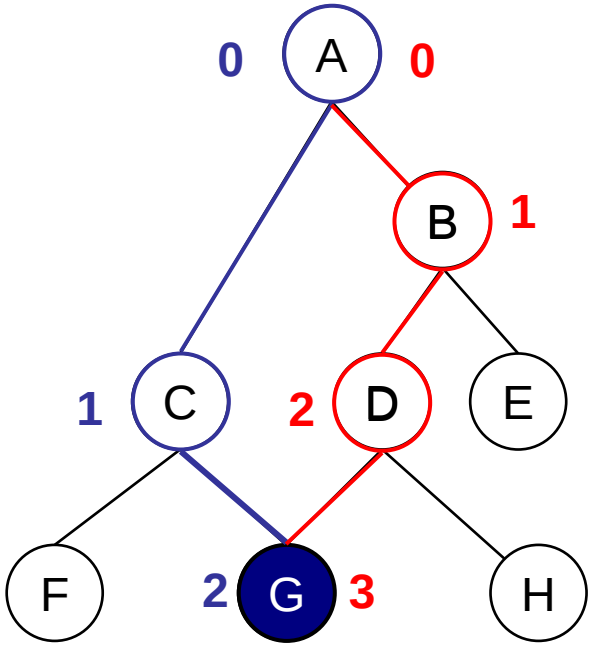
➢ Annotation biomedicale



Regroupement biologique

➢ **Annotation biologique (GO): annotations communes**



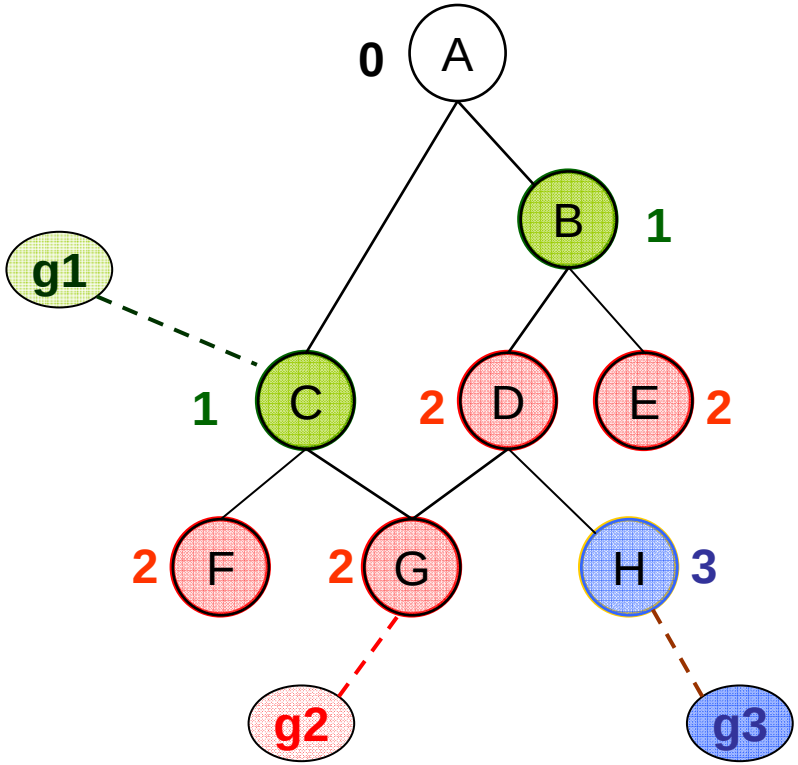


➢ **Un niveau fixe doit être assigné à chaque terme**

Regroupement biologique

➢ **Annotation biologicale (GO): gènes regroupés en fonction du niveau de la hiérarchie**

Niveau d'un terme = chemin minimal jusqu'au terme racine



Niveau 0

A: (g1,g2,g3)

Niveau 1

B: (g2,g3)

C: (g1,g2)

Niveau 2

D: (g2,g3)

G: (g2)

Niveau 3

H: (g3)

Comparaison de niveau

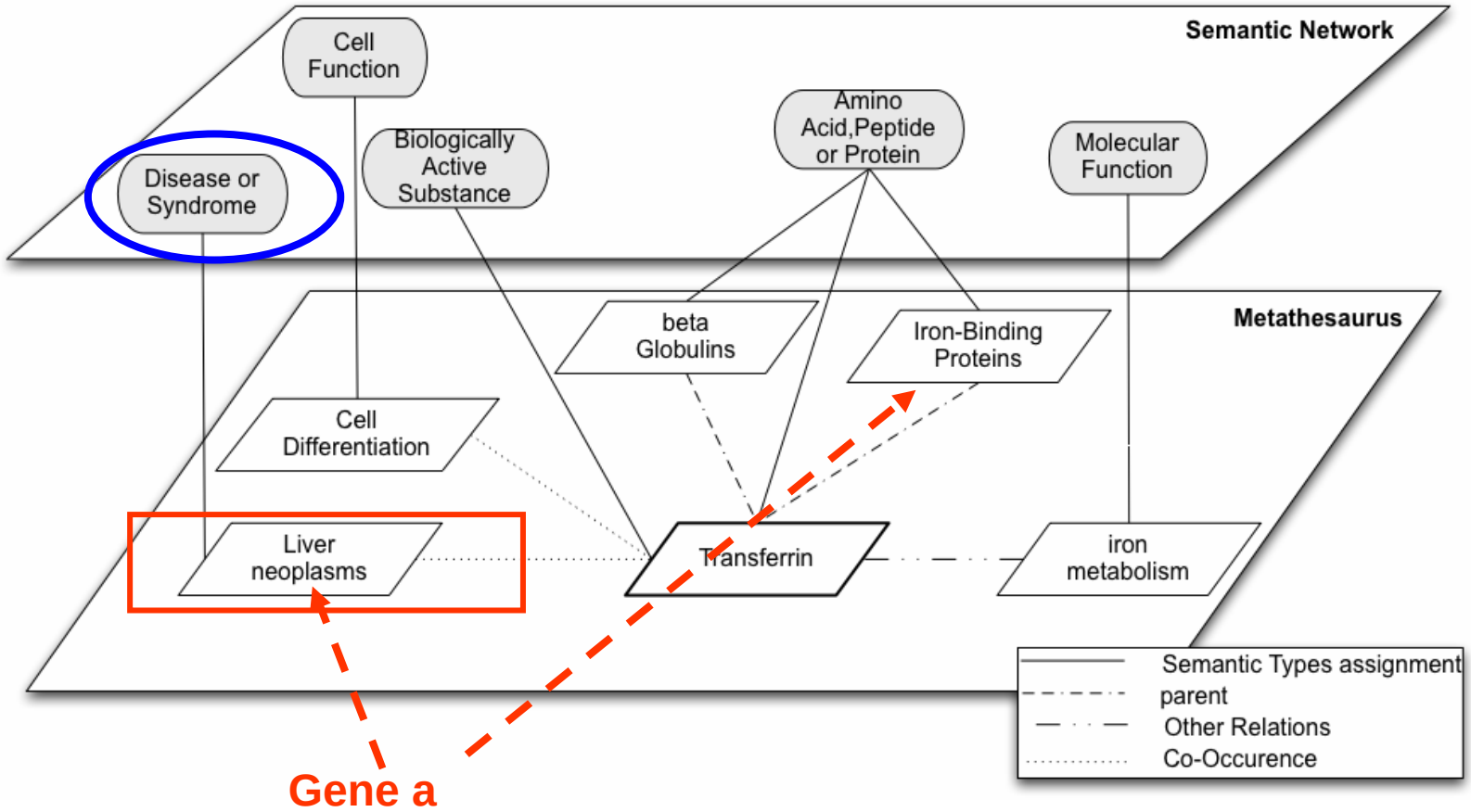
Regroupement médical

➢ Annotation médicale : UMLS

- National Library of Medicine
- Intègre plus de 100 vocabulaires
- Incluant :
  - Metathesaurus :
    - 1 million de concepts
    - 4.5 millions de termes (un concept = un groupe de termes)
  - Réseau sémantique :
    - Concepts sont catégorisés par 135 types sémantiques
    - Types sémantiques sont aggrégés en 15 groupes sémantiques

Regroupement médical

➢ **Annotation médicale (UMLS) : filtrage sur le groupe sémantique « Disorders »**



Regroupement médical

➢ **Groupes de gènes suivant leur annotation: recherche des gènes fonctionnellement liés**

Processus 1
Gene a
Gene b
Gene c
…

Maladie 1
Gene a
Gene b
Gene e
…

Gene a – Gene b

➡ **Calcul de la similarité sémantique entre les gènes**

Similarité sémantique

- **Recherche de relation fonctionnelle entre gènes**

Comparaison de profils d'annotation par la méthode des

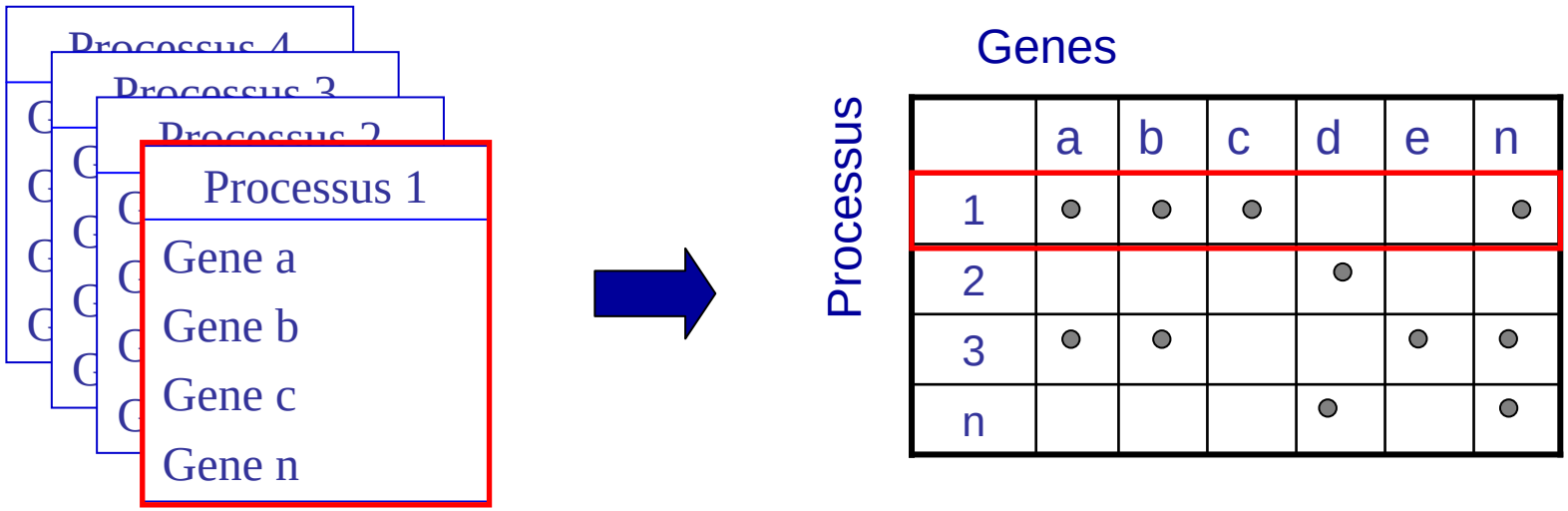
« modèles d'espace vectoriel »

- Recherche d'information: similarité entre documents, similarité requête-document.
- Recherche de relations entre les concepts de Gene Ontology: liens entre les hiérarchies (processus biologique – fonction cellulaire)
- Recherche de gènes fonctionnellement liés en fonction de leur annotation

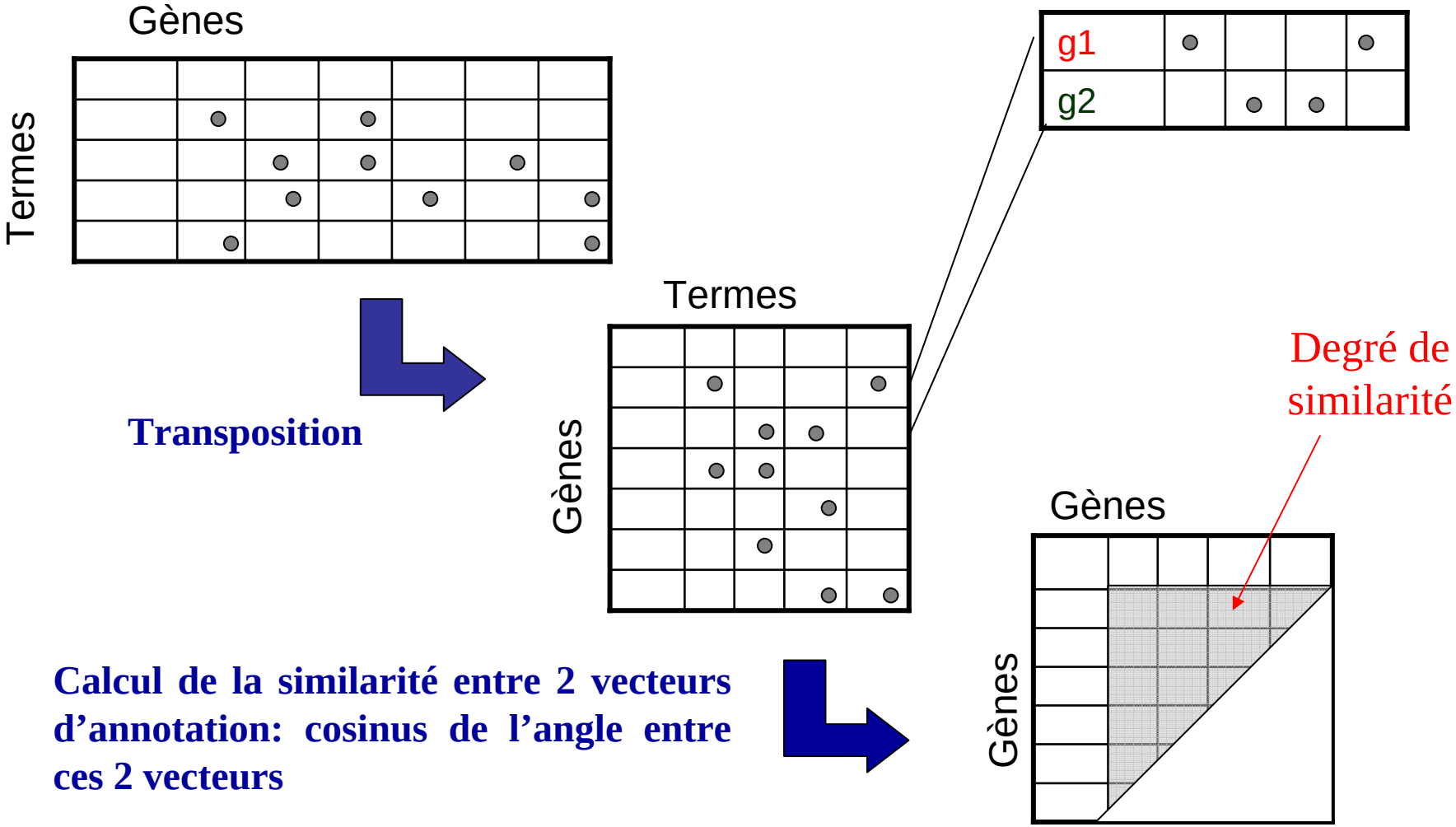
Similarité sémantique

➢ Méthodologie

Construction d'une matrice de termes/gènes à un niveau spécifique de la hiérarchie de Gene Ontology à partir des groupes de gènes



Similarité sémantique



Calcul de la similarité entre 2 vecteurs d'annotation: cosinus de l'angle entre ces 2 vecteurs

Application

- 186 gènes impliqués dans la différentiation entérocytaire

(Bedrine-ferrand *et al.*, 2005)

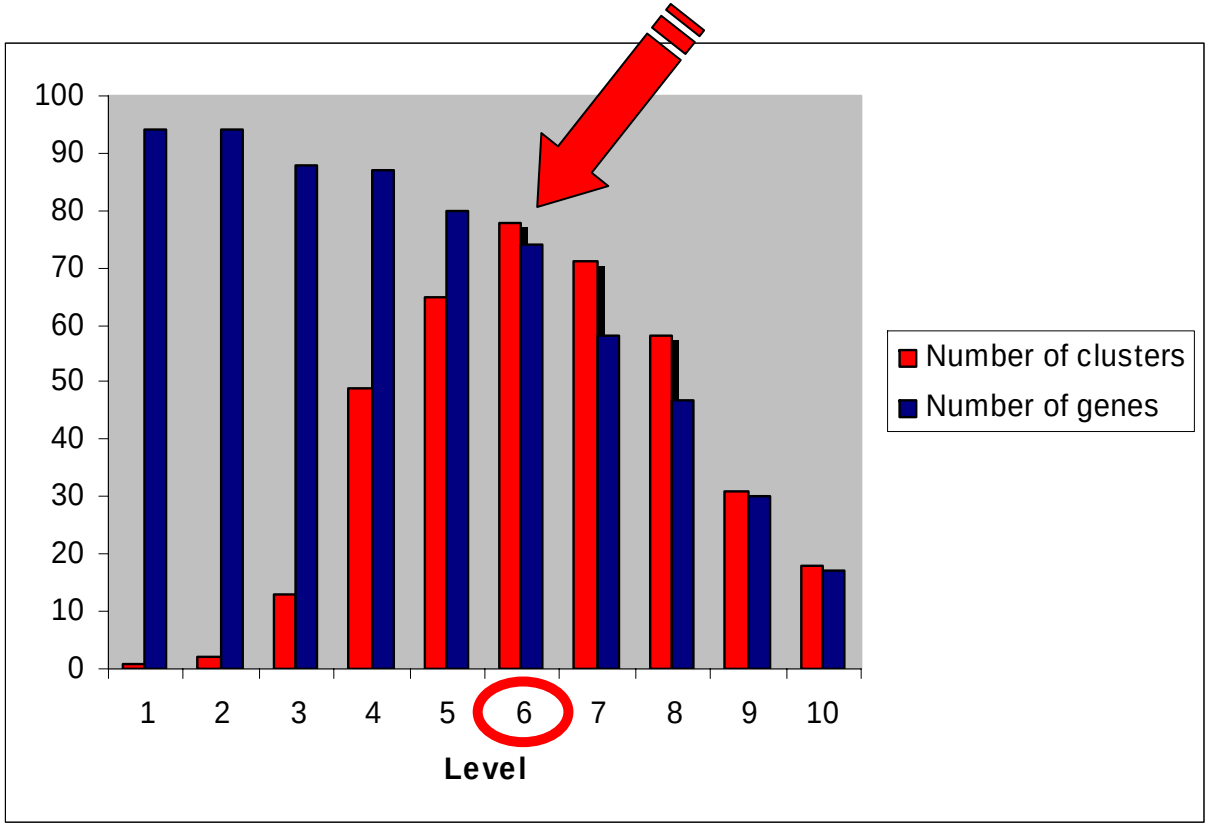
Recherche des gènes impliqués dans des processus biologiques communs

Hypothèse: gènes appartenant à un même groupe de processus devraient appartenir à un même groupe médical

- Annotations BioMeKE
- Regroupement des gènes par processus biologiques et par concepts UMLS
- Comparaison des niveaux d'annotation GO – application du filtre UMLS

Application

➢ Comparaison des niveaux d'annotation GO



le niveau 6 contient un maximum de termes annotant les gènes

La similarité sémantique est calculée pour le niveau 6

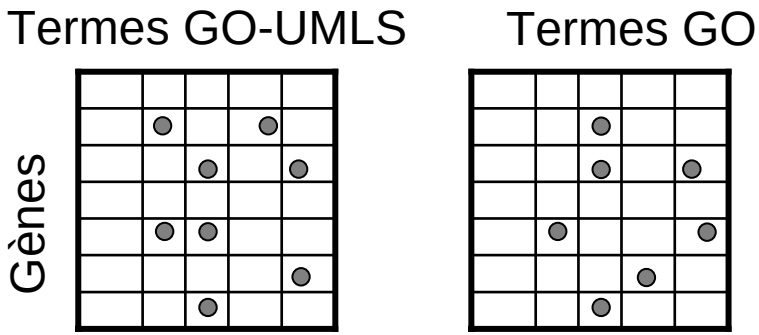
Application

➢Filtrage UMLS

13 gènes sont associés avec 1568 concepts

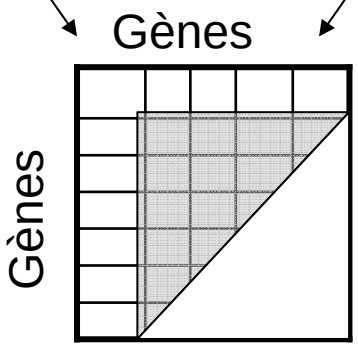
Incorporation des concepts UMLS dans la matrice augmente la disparité entre gènes annotés et non annotés

2 matrices sont construites



Vecteurs GO-UMLS

Vecteurs GO



Similarité sémantique en fonction du type d'annotation

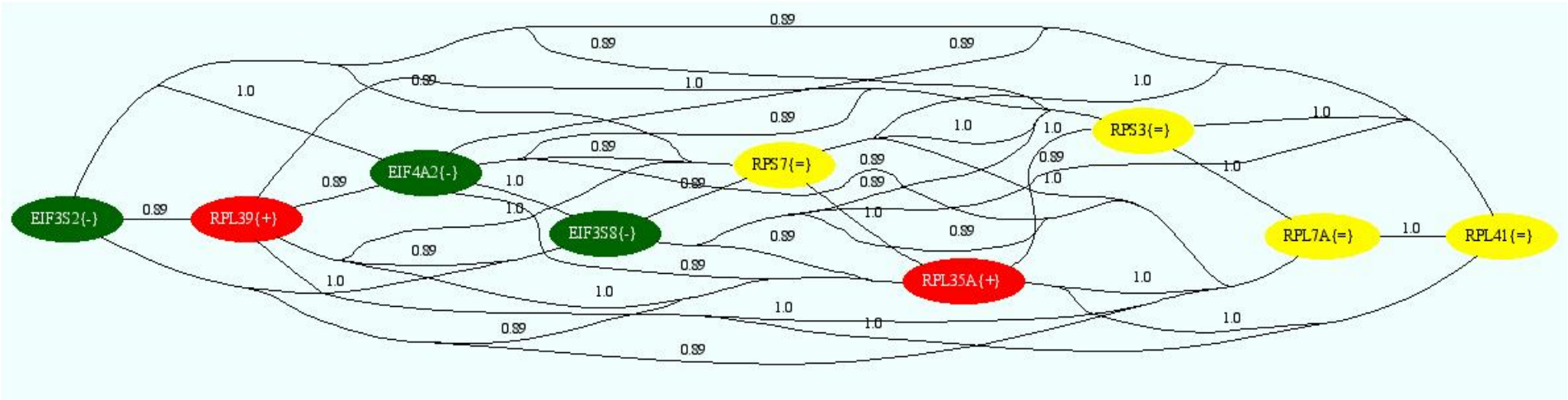
Résultats

- **3 réseaux fonctionnels avec un seuil de 0.6**

Biosynthèse de protéines

Gènes codant des facteur d'initiation à la traduction

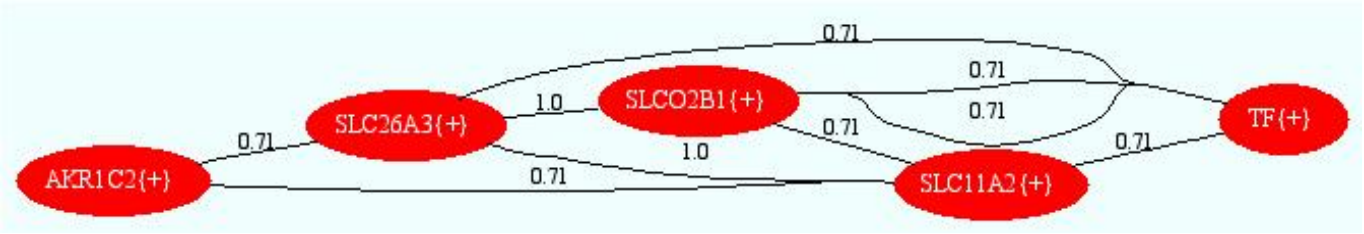
Gènes codant pour des protéines ribosomales



Résultats

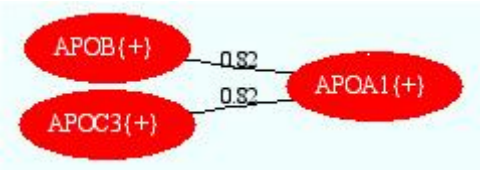
Transport d'ions

Groupe de gènes surexprimés



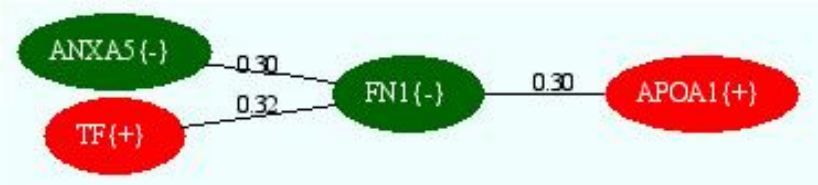
Gènes codant pour des apolipoprotéines

Groupe de gènes surexprimés



Résultats

➢ **Un réseau supplémentaire basés sur les concepts UMLS avec un seuil plus bas**



Ces gènes sont impliqués dans plusieurs maladies mais ne sont pas connus pour être impliqués dans des processus biologiques communs.

Conclusion

- **L'approche transversale :**
  - Une nouvelle approche pour interpréter les données d'expression
  - Basée sur le regroupement des gènes en fonction des termes GO (processus biologiques), UMLS (Maladies), et des profils d'expression (résultats expérimentaux)
  - La similarité sémantique, calcultée à partir des groupes de gènes, permet de construire plusieurs réseaux fonctionnels de gènes.
- **Application sur des gènes impliqués dans la différenciation entérocytaire :**
  - 4 réseaux de gènes différentiellement exprimés
  - Tous ces réseaux sont pertinents (impliqués dans la surcharge en fer)

Perspectives

➢ Prendre en compte la hiérarchie à un même niveau de GO

• Attribuer un poids spécifique aux termes parents dans un même vecteur d'annotation.

➢ Intégrer les gènes non annotés dans les réseaux de similarités

Etude de la fonction des produits de gènes impliqués dans les réseaux

• Regrouper les gènes en fonction des **fonctions GO** (Fonctions moléculaires) et **des caractéristiques des séquences** (promoteurs, domaines)
• Intégrer des fonctions et des caractéristiques des séquences dans les vecteurs d'annotation
• Evaluer la contribution relative des différent descripteurs
• Attributer des poids en fonction des différentes contributions

EA 3888 - Faculté de Médecine
Université Rennes 1

FACULTÉ DE MÉDECINE

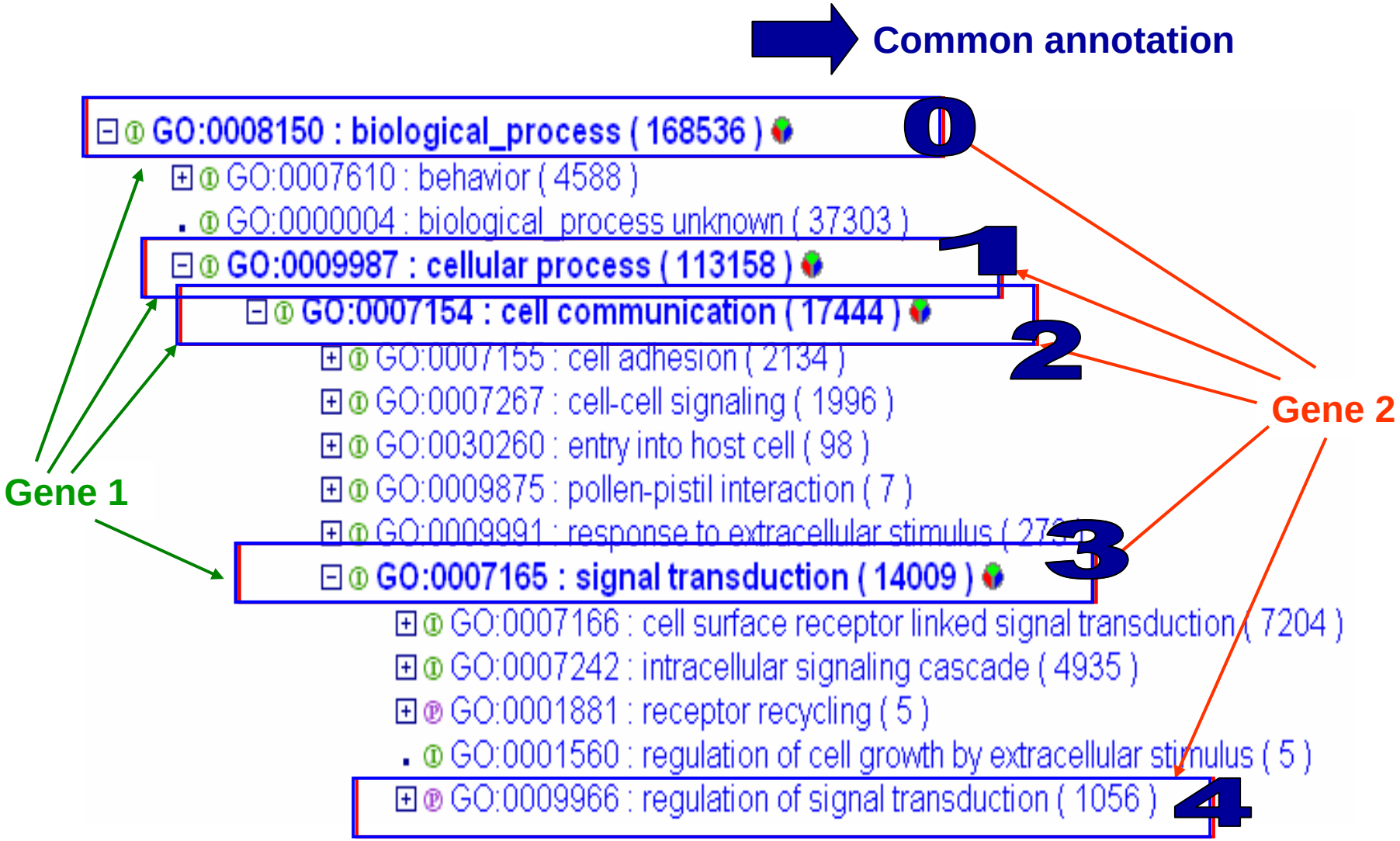
julie.chabalier@univ-rennes1.fr

UNIVERSITE RENNES 1

UMR 6061 - CNRS
Université Rennes 1

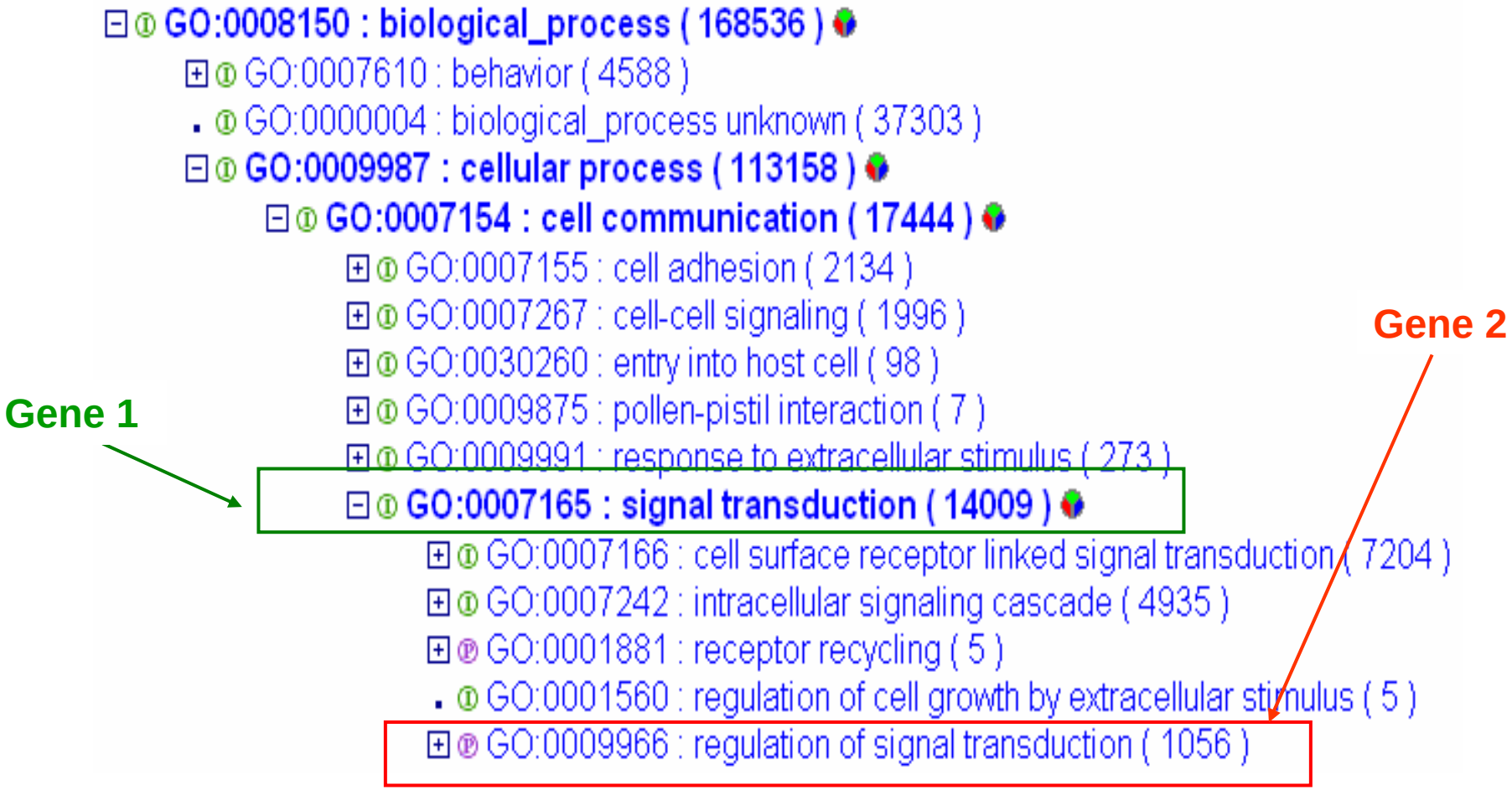
Biological clustering

- **Biological annotation (GO): ascendant retrieval**



Biological clustering

- **Biological annotation (GO): different level of hierarchies**



Related work

- **Gene clustering according to GO annotations**

  The Gene Ontology Categorizer (Joslyn *et al.*, 2005) :

  - Categorization of a set of genes using semantic similarity in GO
  - Ranking of representative terms for a gene set

- The transversal approach compares gene clusters in order to construct networks of similarity
- Specific annotations are required for each genes

- **Semantic similarity**

  (Lord *et al.*, 2003; Wang *et al.*, 2004)

  - Exploit the usage of terms in a corpus to give a measure of information content
  - Integration of this measure to compute semantic similarity

- The VSM method appears to be less precise, but complementary to these approaches
- Possibility to integrate different kinds of knowledge (not only terms) into the vector annotations (motifs, promoters)