

# Intégration de données hétérogènes Approches

Master 2

Bioinformatique et Biologie des Systèmes

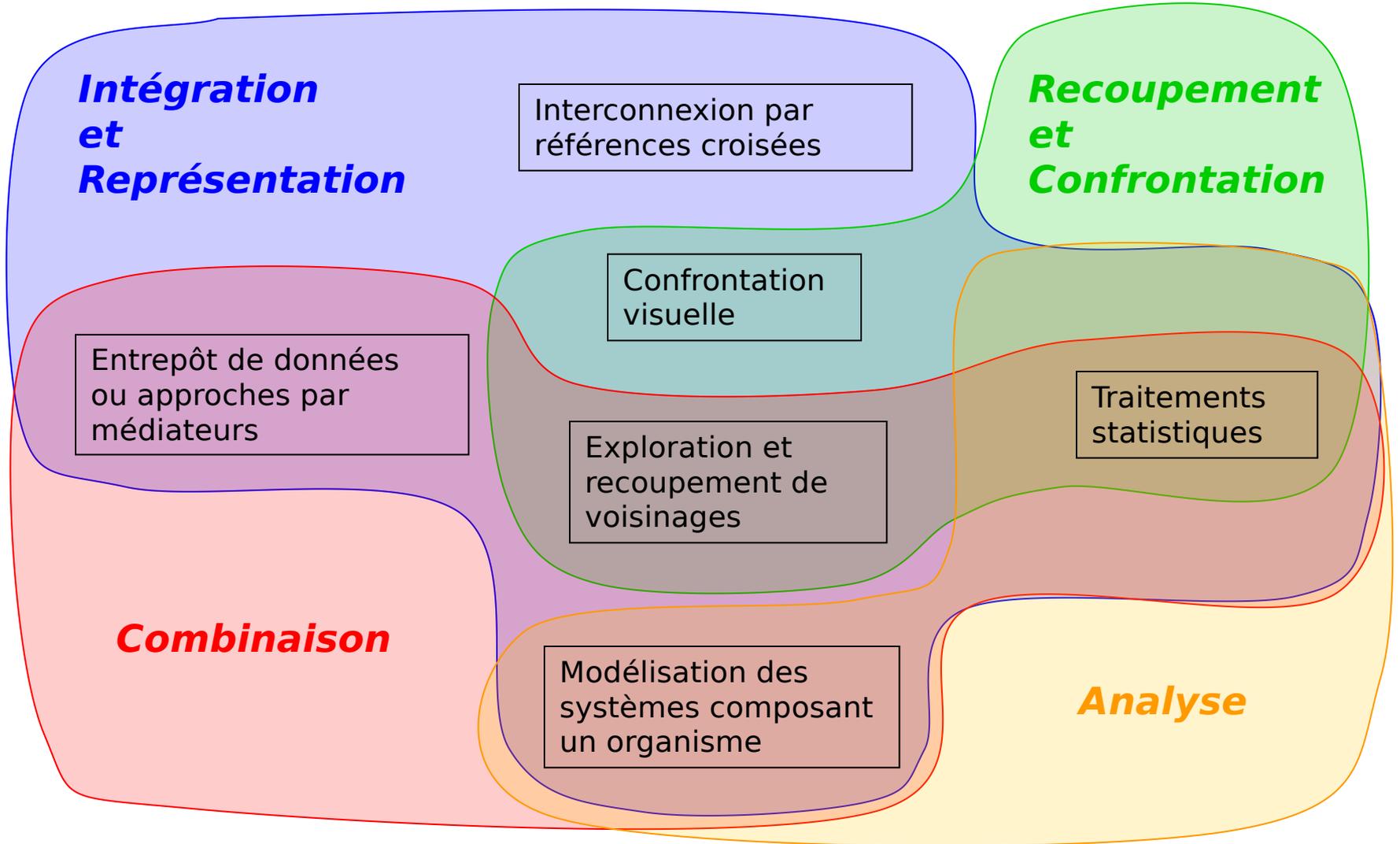
- Pourquoi ?
- Qu'est-ce que l'intégration ?
  - Interconnexion
  - Fusion
  - Médiation
  - Modélisation
  - Confrontation
  - Recoupement



# Apperçu des données disponibles

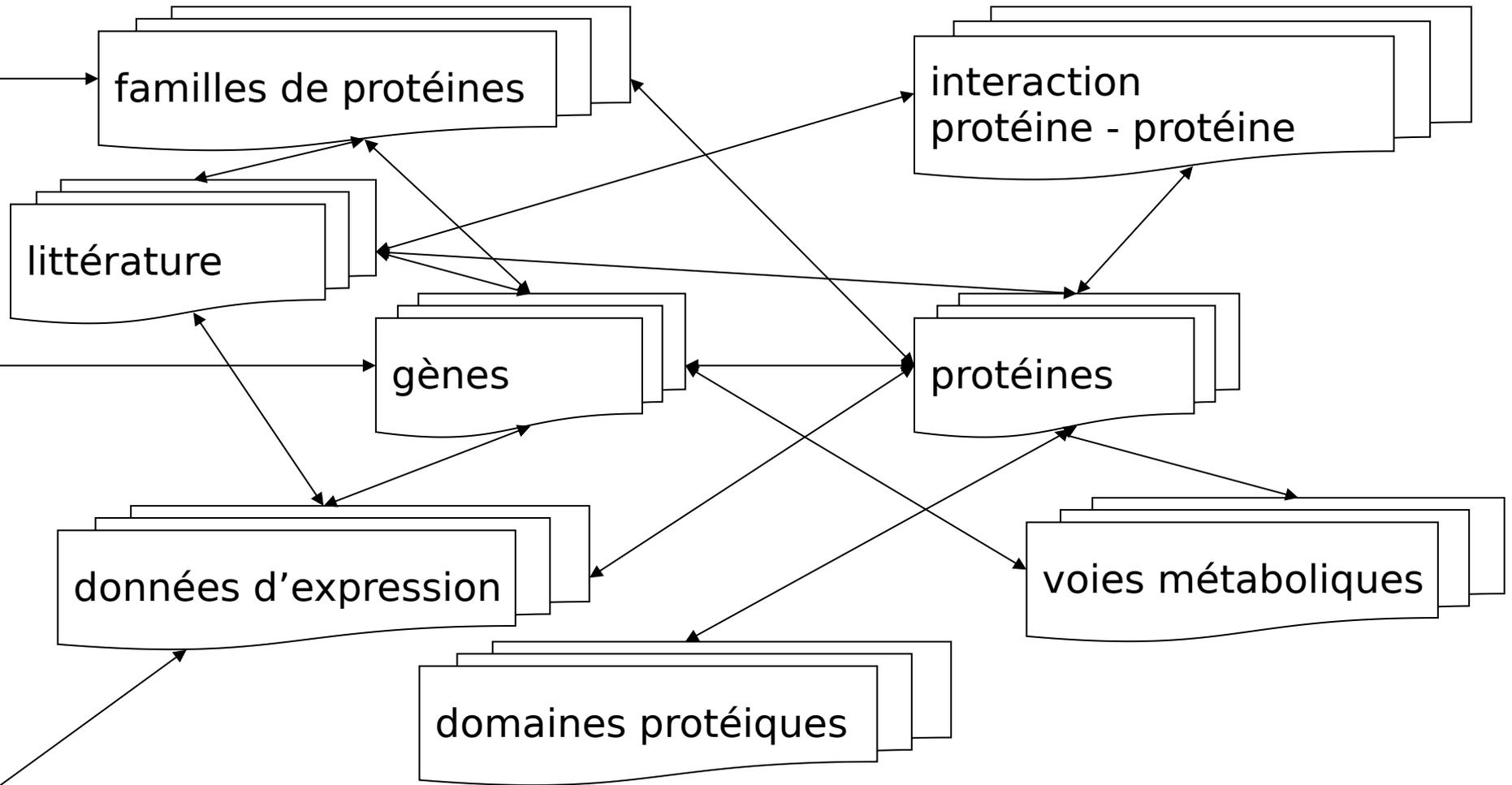
---

- En quantité
- Dispersées
  - gènes, protéines, expression, interaction, ...
  - NCBI, EBI, KEGG, SIB, ...
- Hétérogènes : type, structure et sémantique
  - mots : séquence génome, gène, protéine
  - attributs
    - nominaux : mots-clés, ontologies, vocabulaires contrôlés
    - numériques :
      - niveaux d'expression,
      - usage des codons
  - graphes : interaction protéique, réactions enzymatiques, transduction du signal, structures classificatoires
  - texte
    - vocabulaire contrôlé
    - littérature



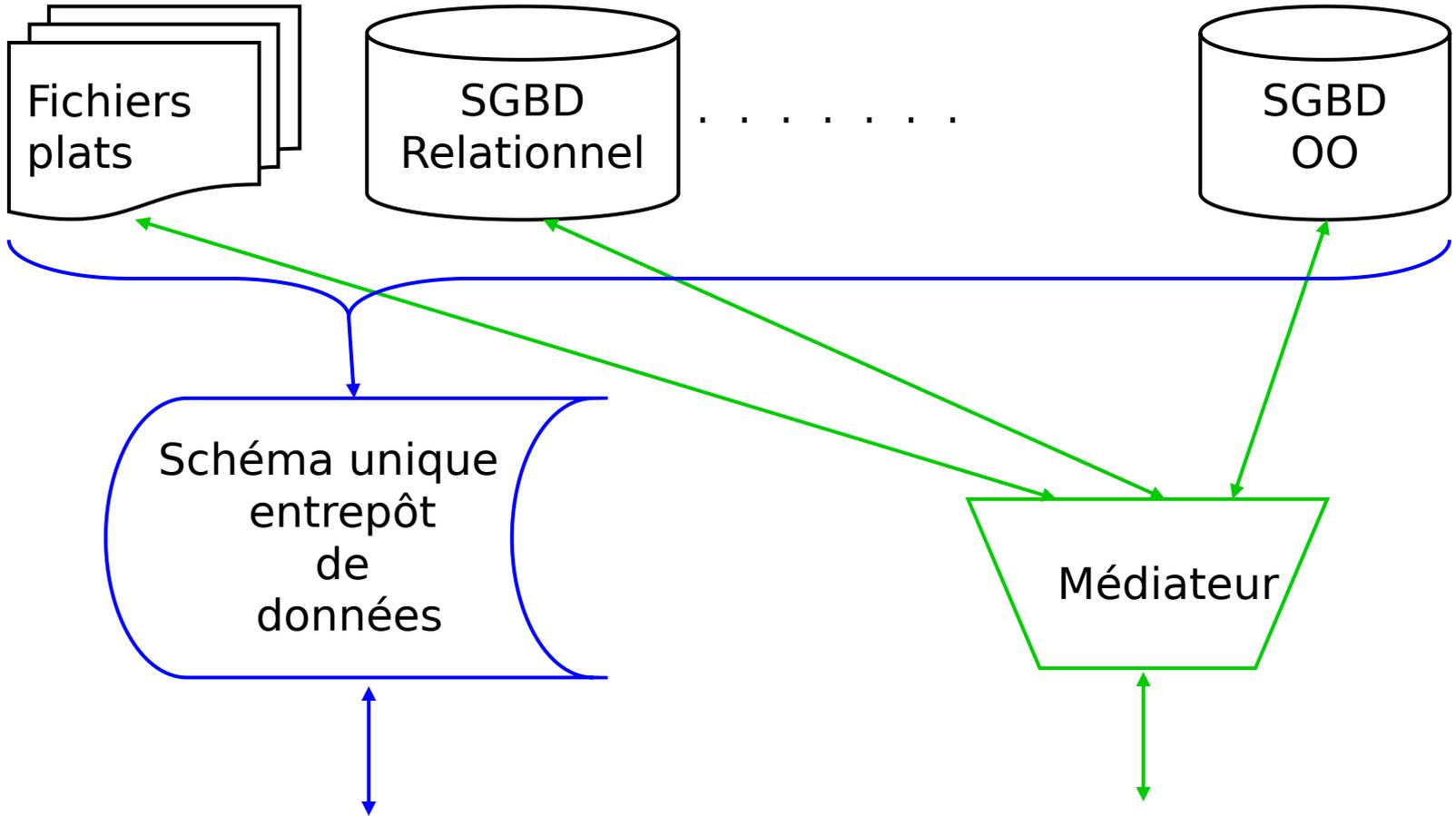
- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
  - fusion

# Intégration par interconnexion : principe



SRS [Etzold *et al.*, 1996], Entrez [Schuler *et al.*, 1996], ...

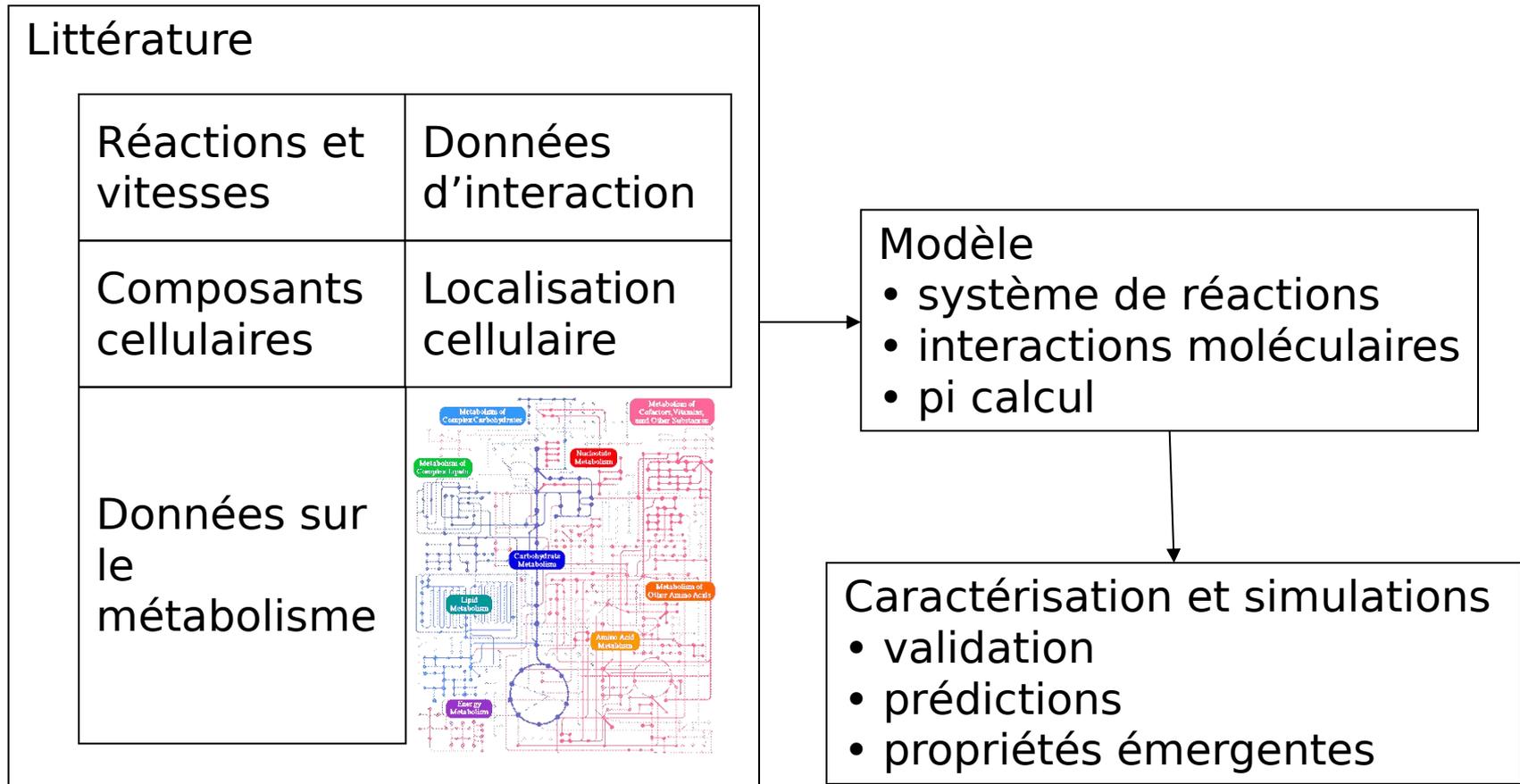
# Intégration par fusion ou par médiateurs



Integr8 [Kersey *et al.*, 2005], BioMart [Kasprzyk *et al.*, 2004],  
WInGS [Abergel *et al.*, 2004], BioKleisli [Davidson *et al.*, 1997], ...

- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
  - fusion

# Intégration par la modélisation : vers la cellule virtuelle



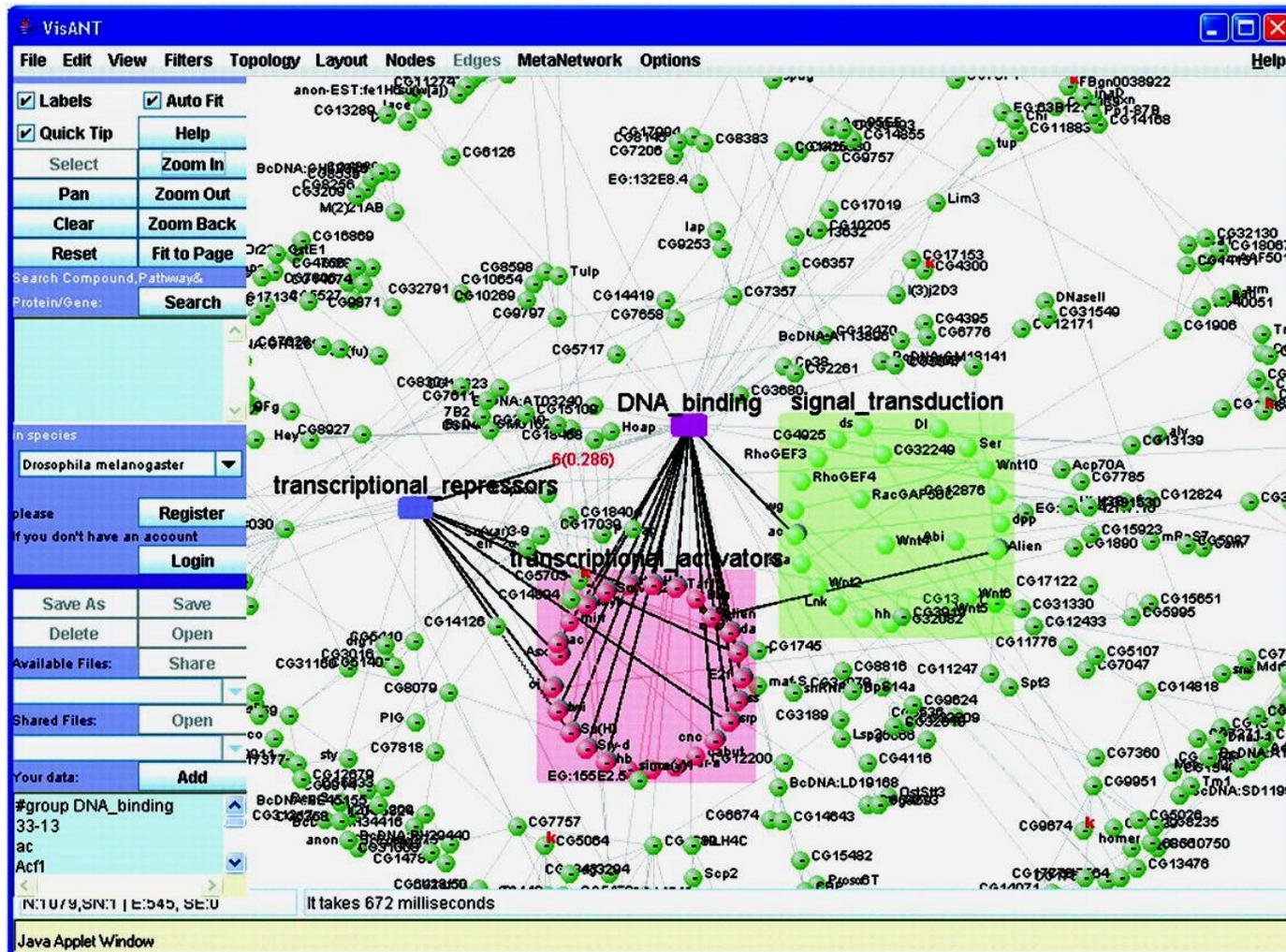
Virtual Cell [Loew et Schaff, 2001], E-CELL [Tomita *et al.*, 1999],  
 Cellerator [Shapiro *et al.*, 2003],  
 MetExplore [Cottret *et al.*, 2010], ...

- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
  - fusion



- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
  - fusion

# Confrontation visuelle



Visant [Hu *et al.*, 2005]

- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
  - fusion

**argB neighbours**

Swiss Prot Classification **Codons**

Bibliography pI Save / Print

Neighbor genes

- argA
- ybjD
- ybbB
- yeiE

**argH neighbours**

Swiss Prot Classification Codons Pathway

**Bibliography**

Select a level below to display neighbor genes

**Escherichia coli and Salmonella typhimurium cellular and molecular biology.**  
F. Neidhardt, R. Curtiss III, J. Ingraham, E. Lin, K. Brooks Low, B. Magasanik, W. Reznikoff, M. Riley, M. Schaechter and H. Umberg

**Variations on a Theme by Escherichia**

- [Genome Structure](#)
- Biosynthesis of Arginine and Polyamines**
  - [Arginine Biosynthetic Enzymes](#)
    - N-Acetylglutamokinase
    - N-Acetylglutamylphosphate Reductase
    - Argininosuccinase
  - [Arginine Reulon](#)

Neighbor genes

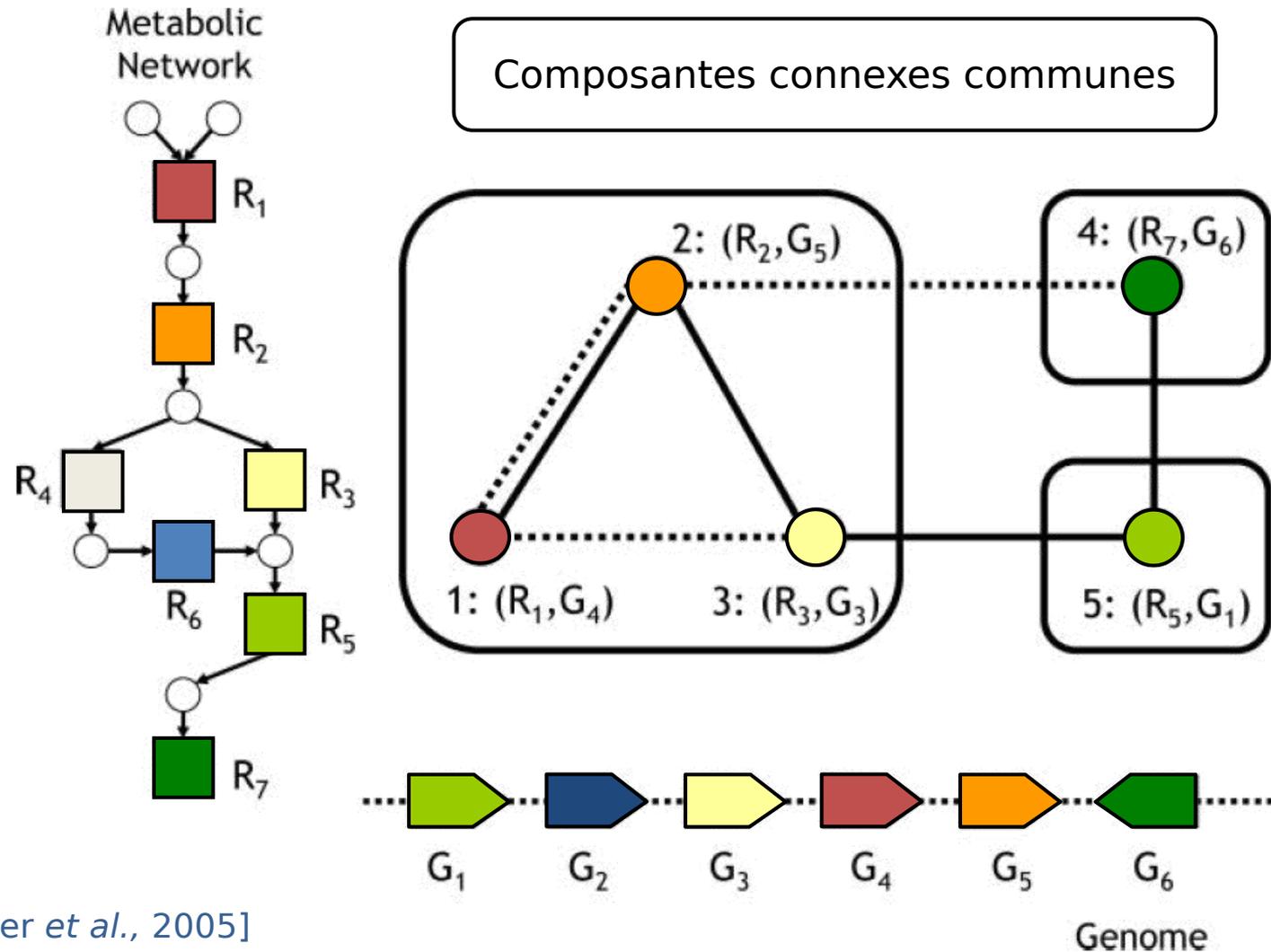
Gene Neighborhood

Codons Usage

Delete

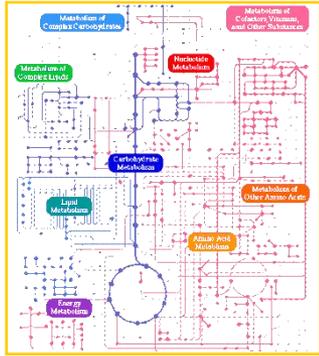
Java Applet Window

# Recoupement de voisinages : approche basée sur les graphes

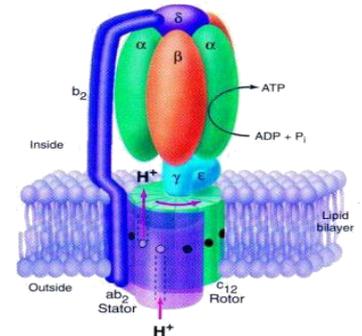


[Boyer *et al.*, 2005]

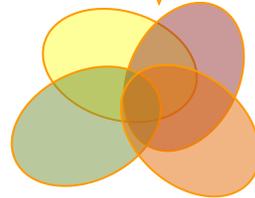
# Recouplement de voisinages : approche ensembliste



voies métaboliques



complexes protéiques



ensembles de gènes

Nucleic Acids Research Advance Access published May 28, 2008

Nucleic Acids Research Advance Access published May 28, 2008

Nucleic Acids Research Advance Access published May 28, 2008  
Nucleic Acids Research, 2008, 1-8  
doi:10.1093/nar/gkn235

**ENDEAVOUR update: a web resource for gene prioritization in multiple species**

Léon-Charles Tranchevent<sup>1</sup>, Roland Barriot<sup>1</sup>, Shi Yu<sup>1</sup>, Steven Van Vooren<sup>1</sup>, Peter Van Loo<sup>1,2,3</sup>, Bert Coessens<sup>1</sup>, Bart De Moor<sup>1</sup>, Stein Aerts<sup>3,4</sup> and Yves Moreau<sup>1,\*</sup>

<sup>1</sup>Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, <sup>2</sup>Human Genome Laboratory, Department of Molecular and Developmental Genetics, VIB Leuven, <sup>3</sup>Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine and <sup>4</sup>Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, Leuven (Belgium)

Received February 7, 2008; Revised April 30, 2008; Accepted May 7, 2008

**ABSTRACT**  
Endeavour (<http://www.esat.kuleuven.be/endeavour> web; this web site is free and open to all users and there is no login requirement) is a web resource for the prioritization of candidate genes. Using a training set of genes known to be involved in a biological process of interest, our approach consists of (i) inferring several models based on various genomic data sources, (ii) applying each model to the candidate genes to rank those candidates against the profile of the known genes and (iii) merging the several rankings into a global ranking of the candidate genes. In the present

**BACKGROUND**  
With the recent improvements in high-throughput technologies, many organisms have seen their genomes sequenced and, more importantly, annotated. This process leads to the generation of a large amount of genomic data and the creation and maintenance of corresponding databases. However, covering genomic data into biological knowledge to identify genes involved in a particular process or disease remains a major challenge. Nevertheless, there is much evidence to suggest that functionally related genes often cause similar phenotypes (1-3). To identify which genes are responsible for which phenotype, association studies and linkage analyses are often used, resulting in large lists of candidate genes. In

co-citation

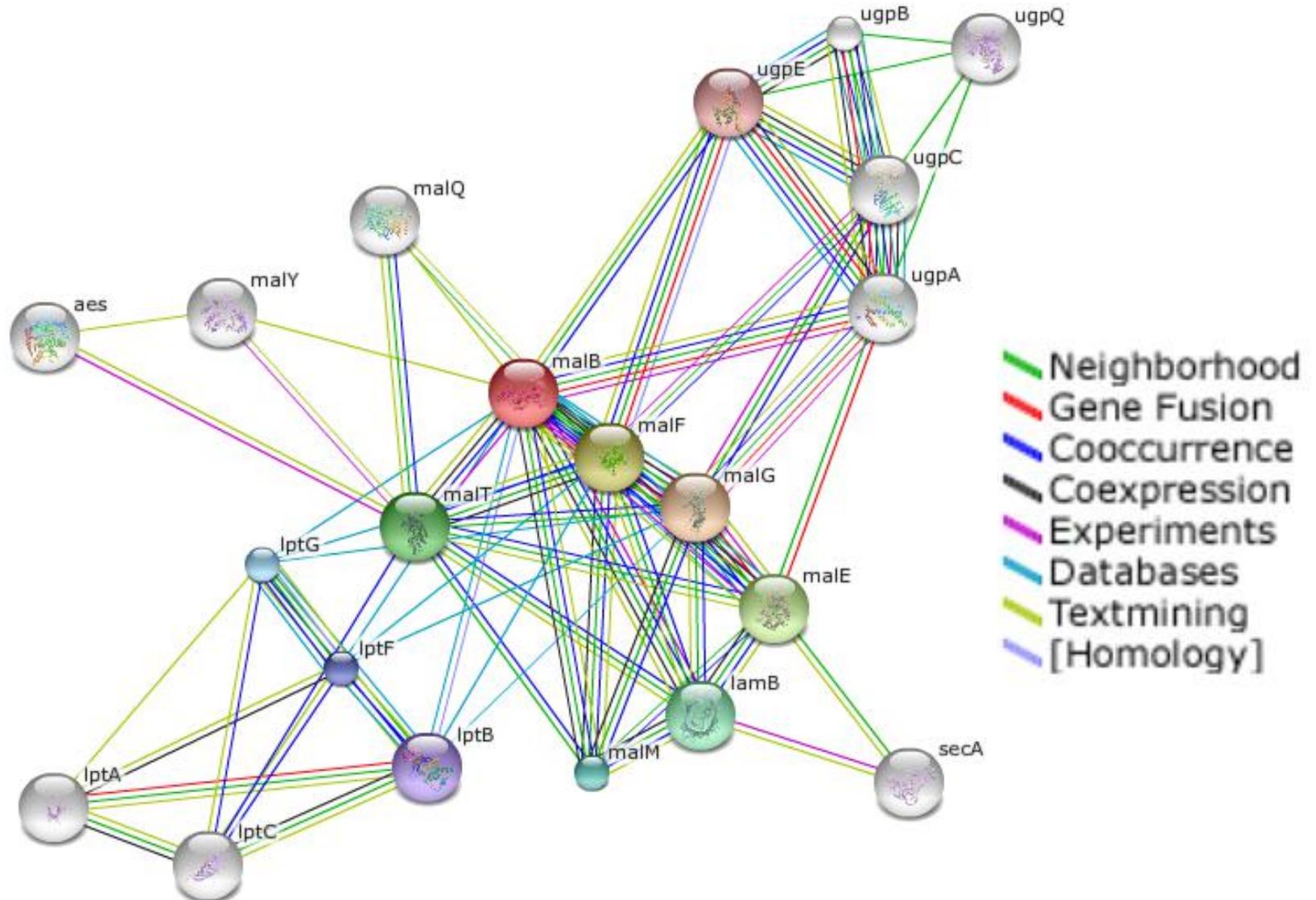


domaines protéiques

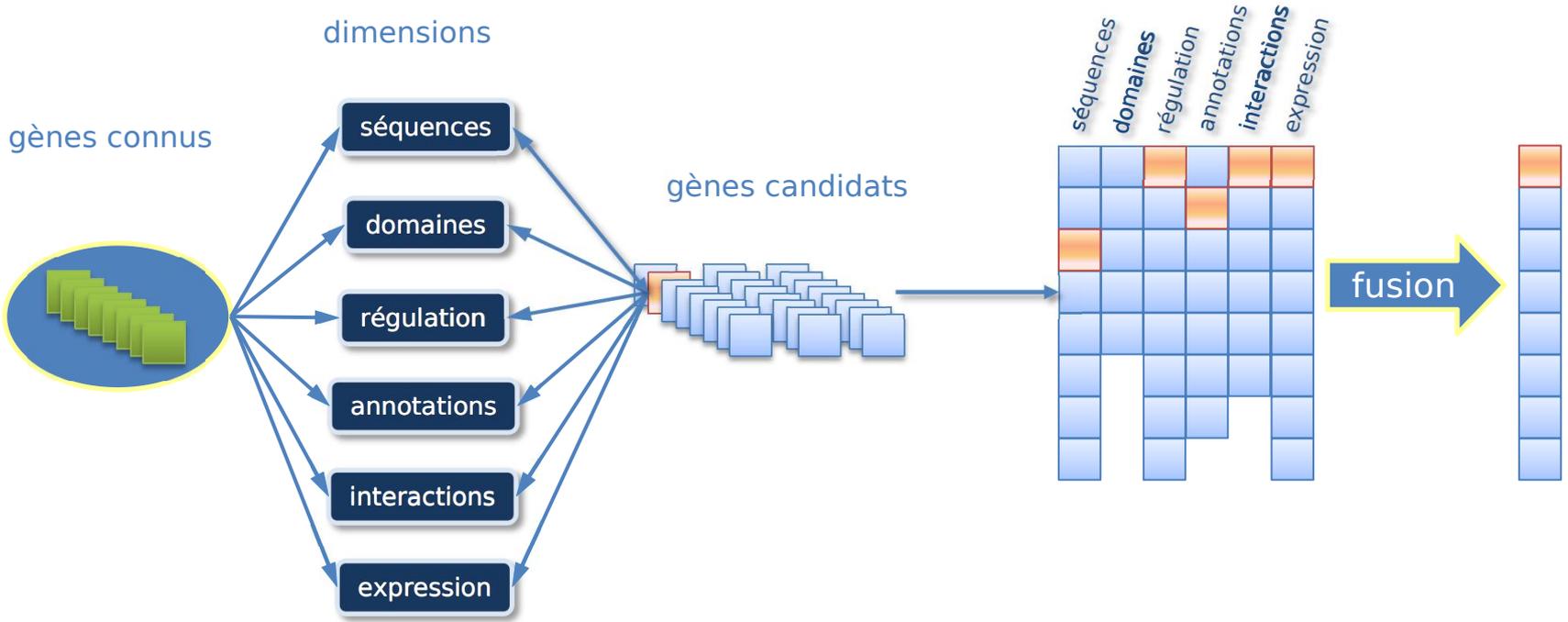


Gene Ontology

# Fusion : approche basée sur les graphes



STRINGdb [von Mering *et al.*, 2003]



- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
  - fusion

# L'hétérogénéité sous diverses ses formes

---

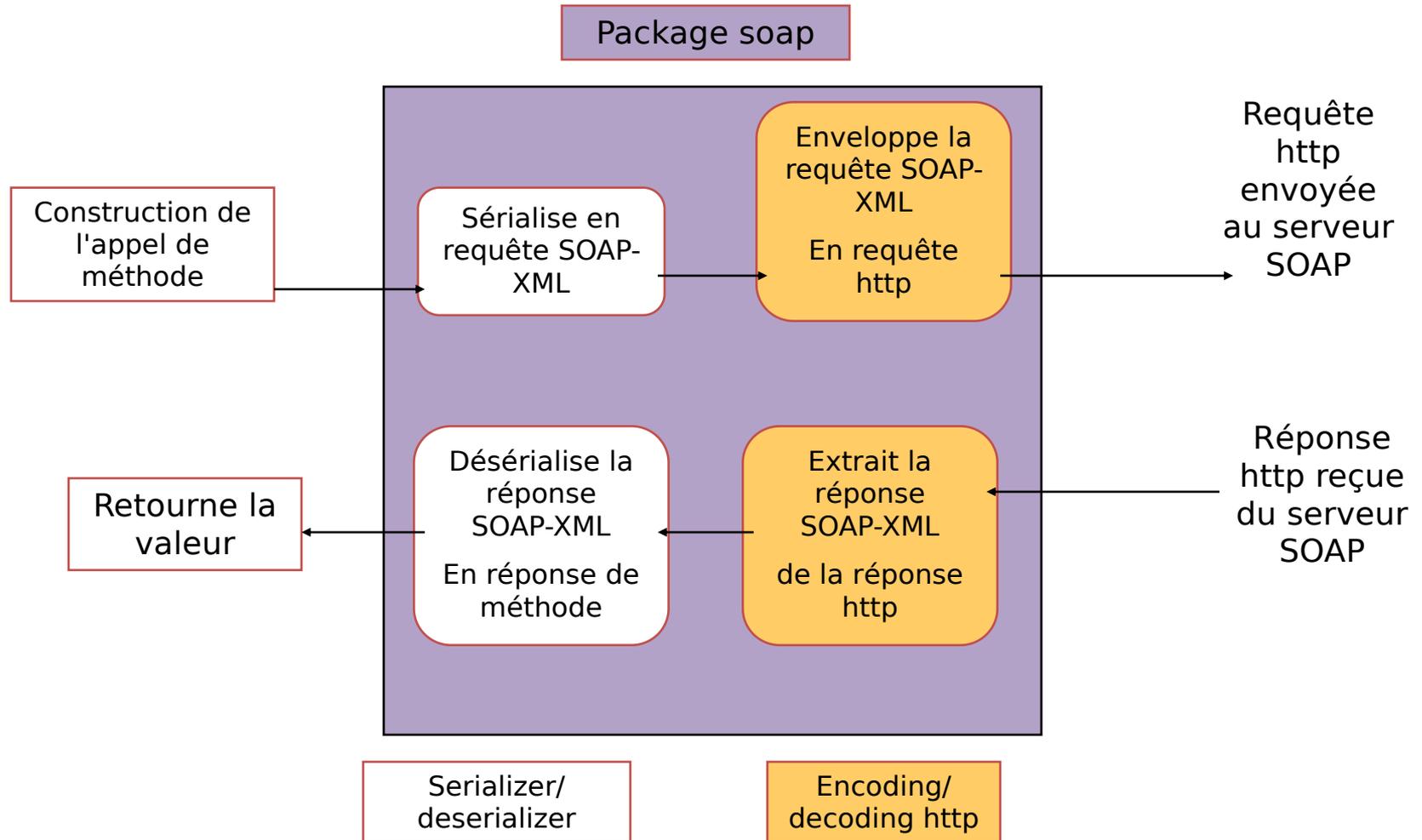
- Sémantique
  - Gene Ontology BP/KEGG pathways/BioCyc
  - structure d'un gène/peptide
- Modèle et format
  - SGBDr, SGBDoo, LDAP, fichier
    - schémas, attributs et unités
  - XML, FASTA, EMBL/GenBank, SWISSPROT
- Architecture et accès
  - SGBD, SOAP, REST, pipeline (galaxy, ergatis)

- Hétérogénéité des systèmes et des représentations
- Besoin de standards
- Plusieurs solutions :
  - Format d'échange : XML, JSON, ...
  - Protocole d'accès : services Web, SOAP, REST, ...

- 
- Principe
    - appel de procédure à distance
    - permet de faire communiquer différents systèmes (plateforme, OS, langage de programmation, ...) à travers XML
    - un fournisseur propose certaines fonctionnalités
    - description du mode d'accès à ces fonctions : WSDL (Web Service Description Language)
    - protocole de communication : SOAP (Simple Object Access Protocol)
      - couche transport (HTTP, SMTP, POP3, IMAP, ...)
      - représentation XML : encodage/décodage des données

- Côté fournisseur :
  - serveur : Web, SMTP ou autre
  - ex: script CGI (apache-perl-SOAP), ou PHP
- Côté client :
  - de nombreuses bibliothèques disponibles
    - perl, ruby, python, Java, ...

## Processus Client avec binding http



## Processus Serveur avec binding http

