# Intégration de données hétérogènes
# Approches

Master 2

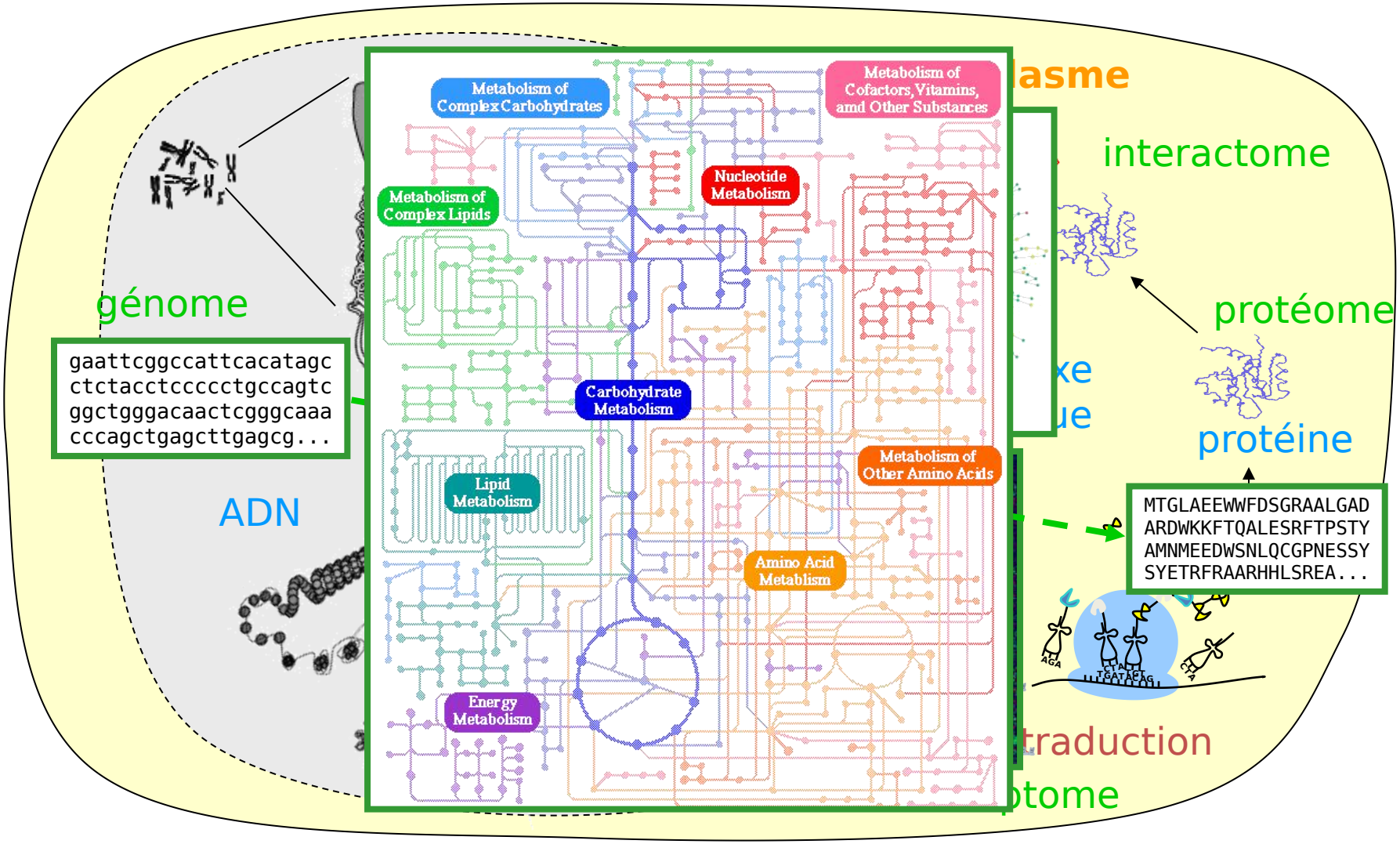
Bioinformatique et Biologie des Systèmes

# Introduction

- Pourquoi ?

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

# Biologie, mesures, données et connaissances




**Cellule eucaryote**

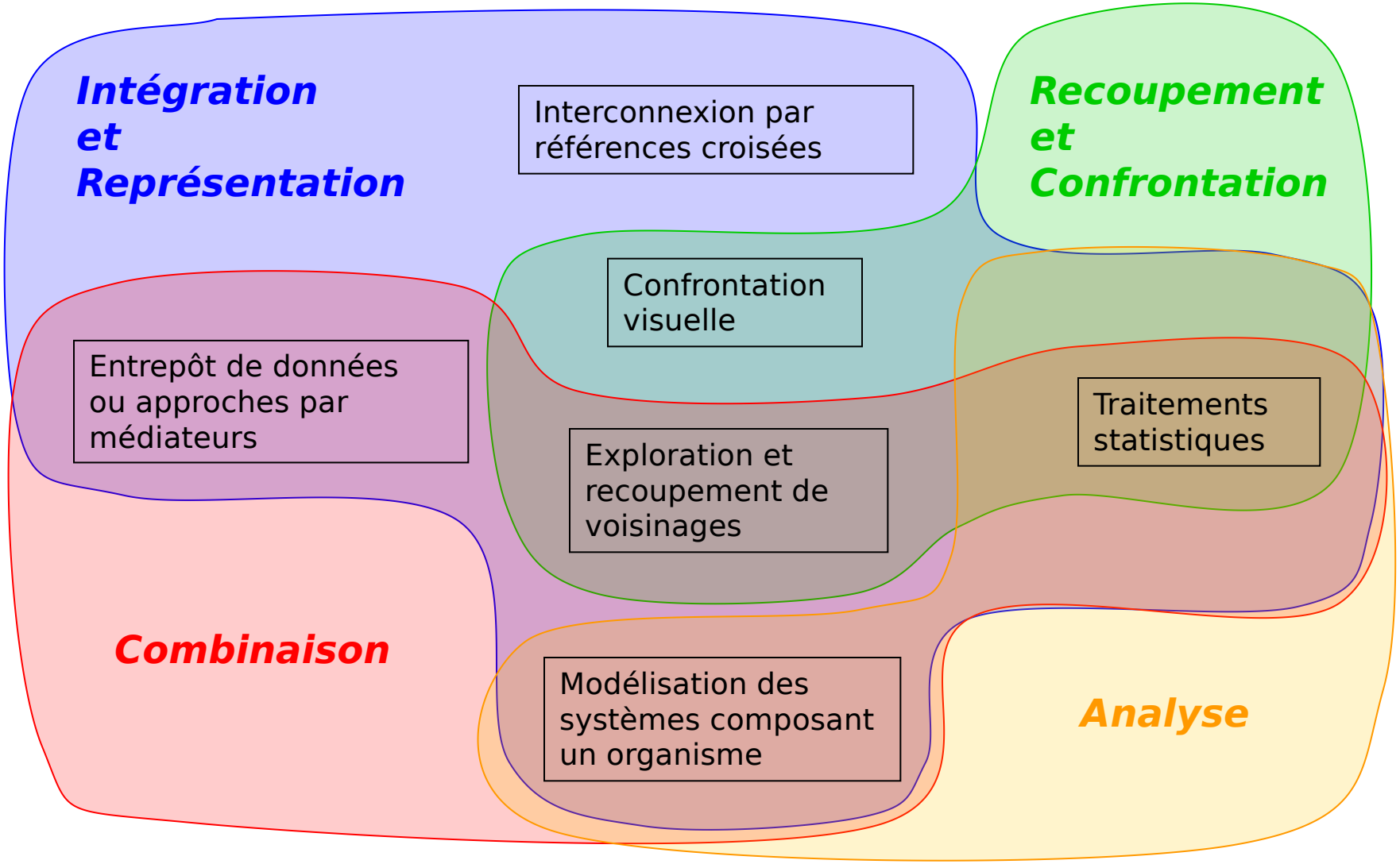
# Apperçu des données disponibles

- 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

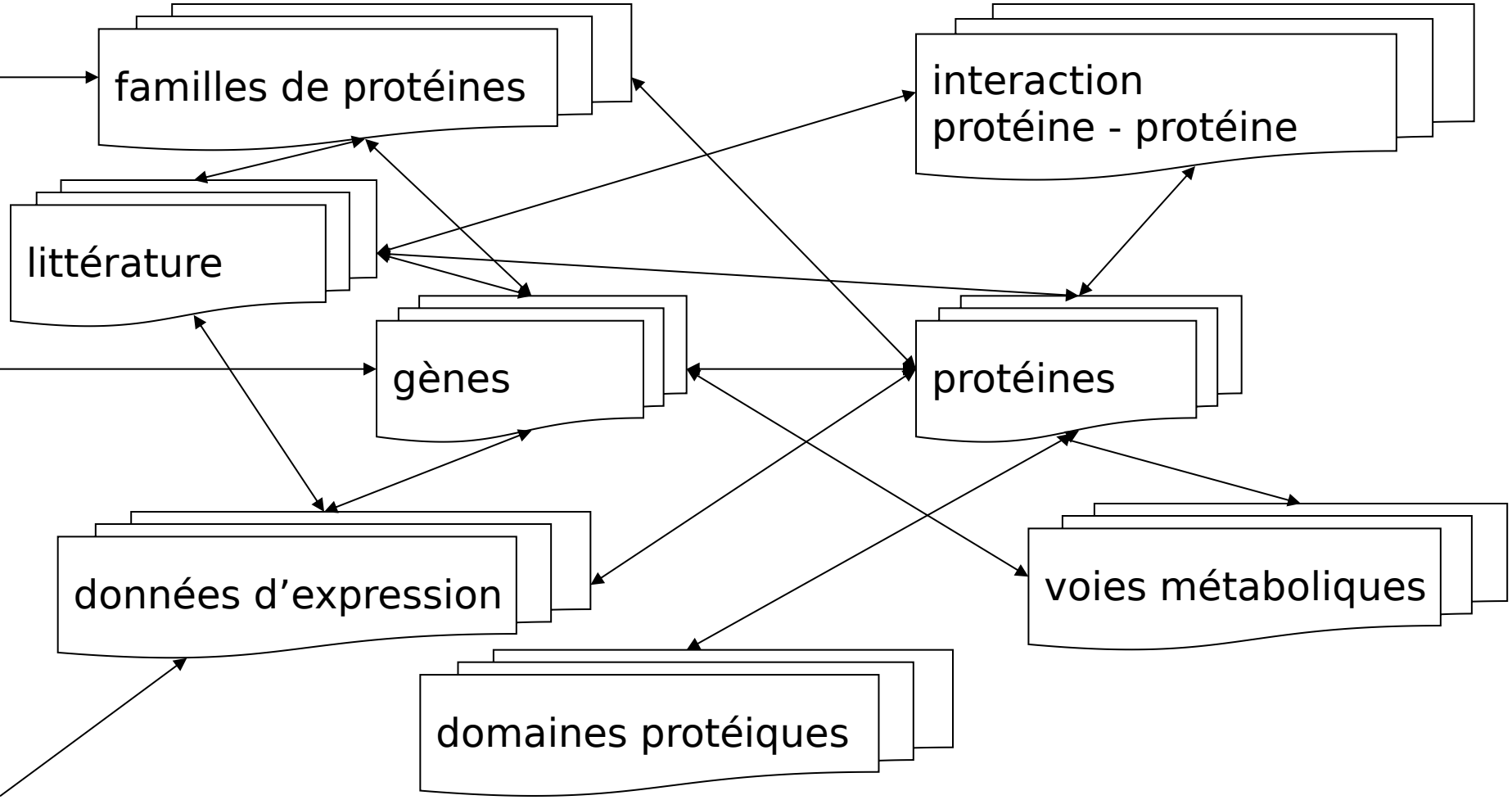
# Vue générale



# Intégration : vers une vue unifiée

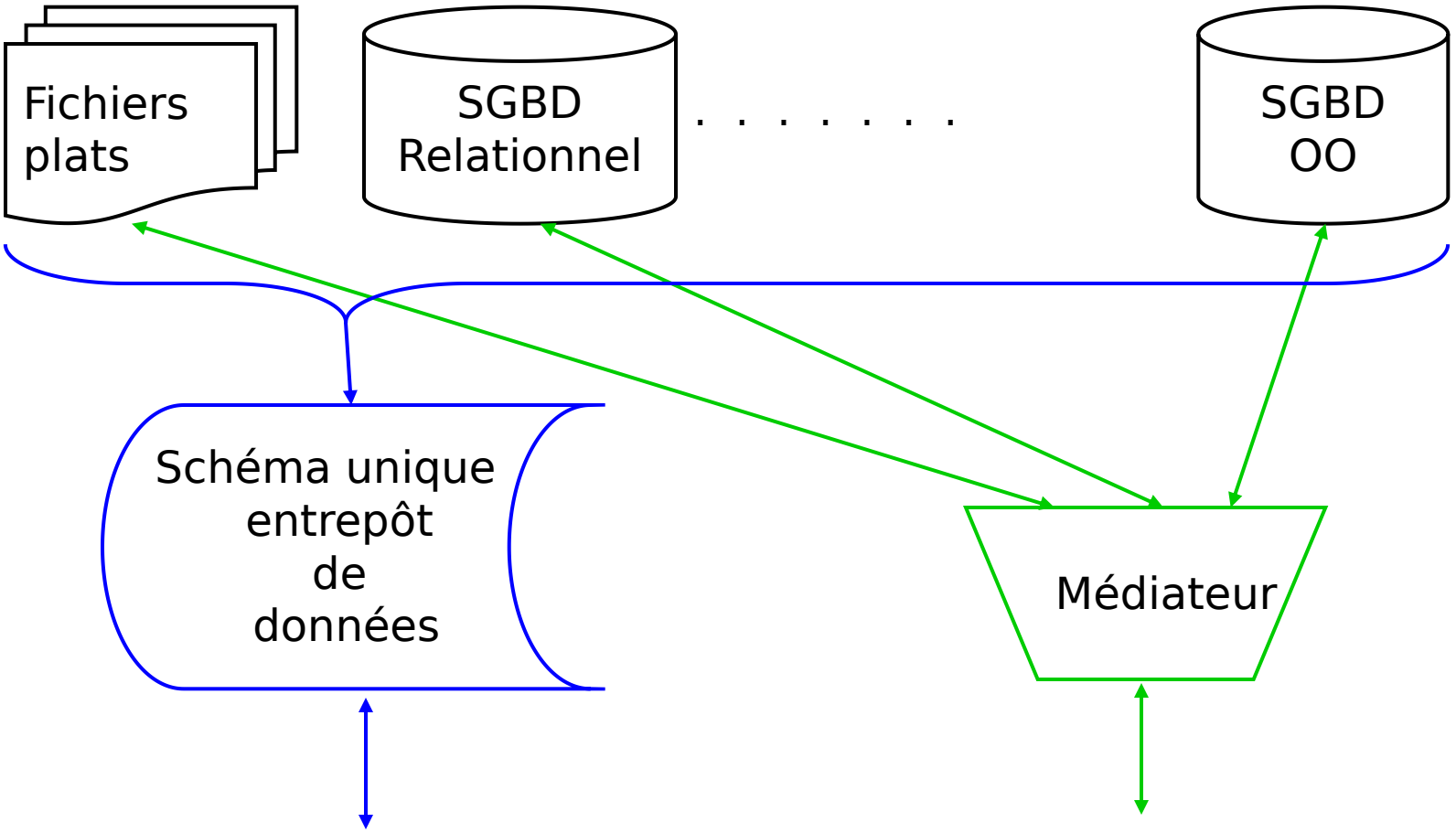
- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
 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

# Intégration par interconnexion : principe




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

# Intégration par fusion ou par médiateurs



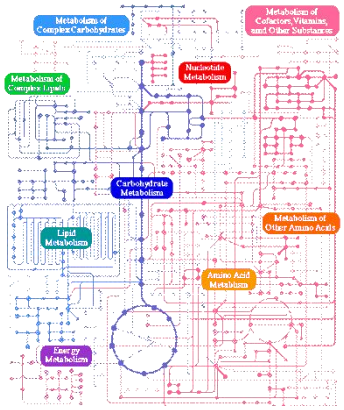

Integr8 [Kersey *et al.*, 2005], BioMart [Kasprzyk *et al.*, 2004],
WInGS [Abergel *et al.*, 2004], BioKleisli [Davidson *et al.*, 1997], ...

# Intégration : vers une vue unifiée

- 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

Littérature

| | |
|---|---|
| Réactions et vitesses | Données d'interaction |
| Composants cellulaires | Localisation cellulaire |
| Données sur le métabolisme |  |



Modèle
- système de réactions
- interactions moléculaires
- pi calcul

Caractérisation et simulations
- validation
- prédictions
- propriétés émergentes

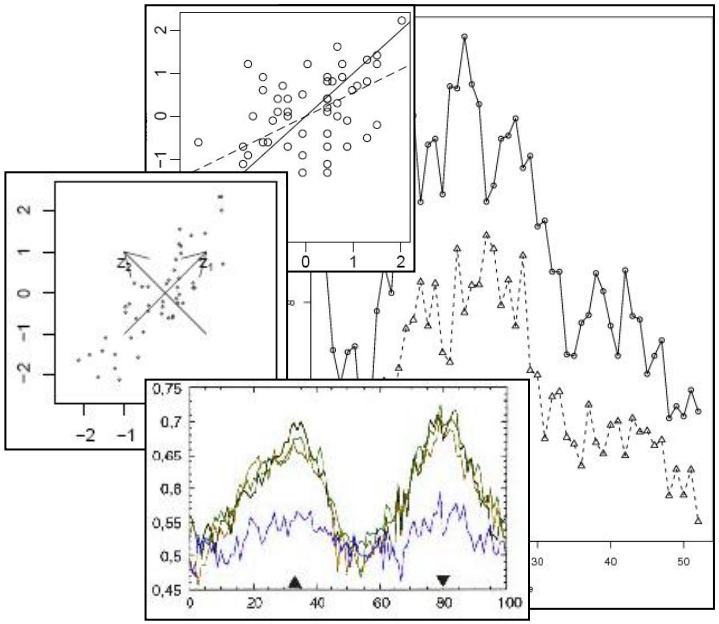
Virtual Cell [Loew et Schaff, 2001], E-CELL [Tomita *et al.*, 1999],
Cellerator [Shapiro *et al.*, 2003],
MetExplore [Cottret *et al.*, 2010], ...

# Intégration : vers une vue unifiée

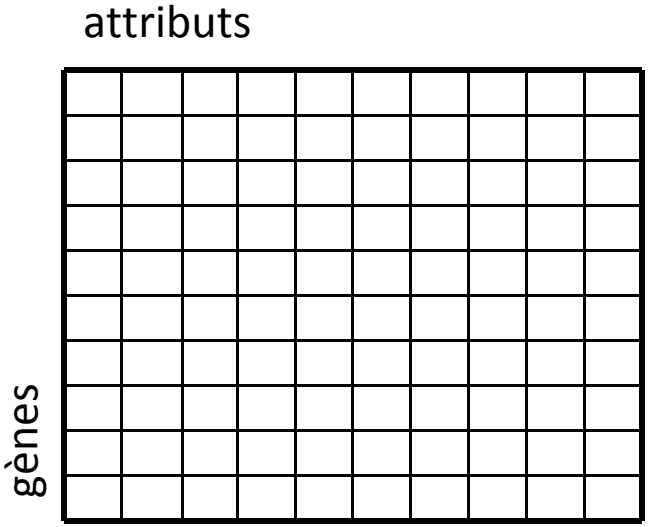
- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
 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

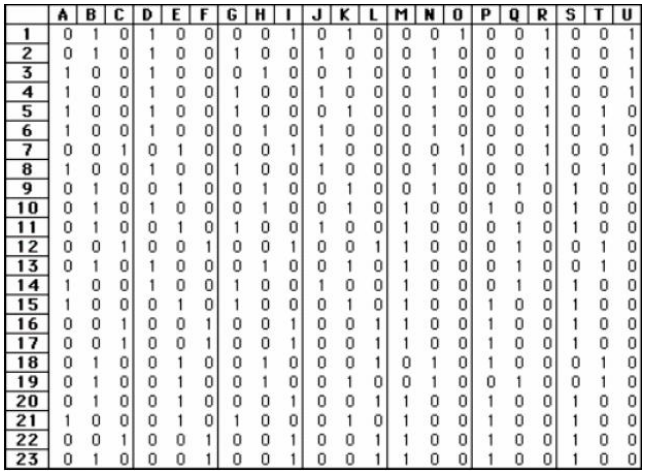
# Statistiques

- Analyse de la variance
- Analyse en composantes principales
- Analyse factorielle des correspondances
- Analyse des correspondances multiples
- …



integrOmics devenu
MixOmics [Lê Cao *et al.*, 2009]



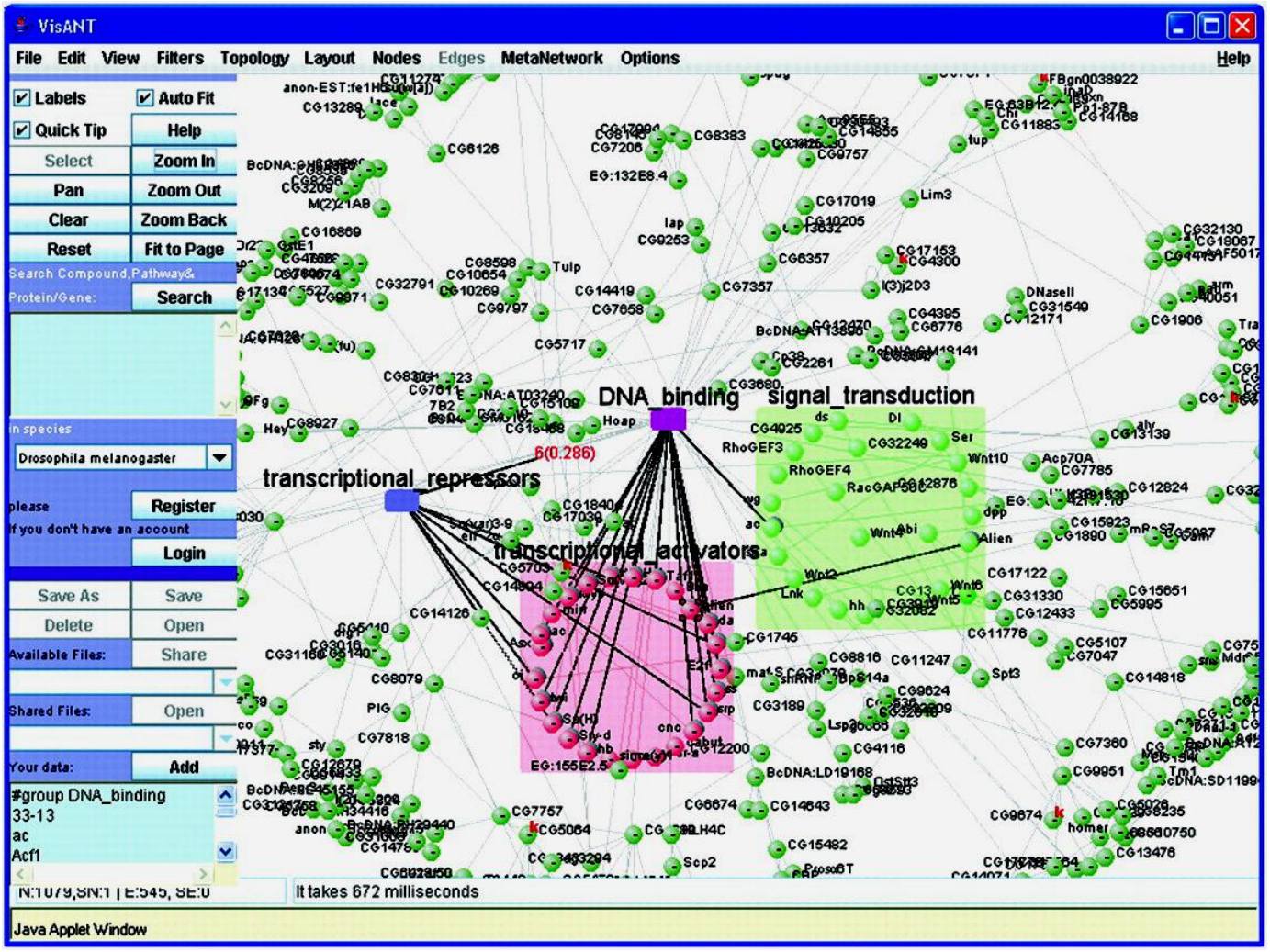



# Intégration : vers une vue unifiée

- 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





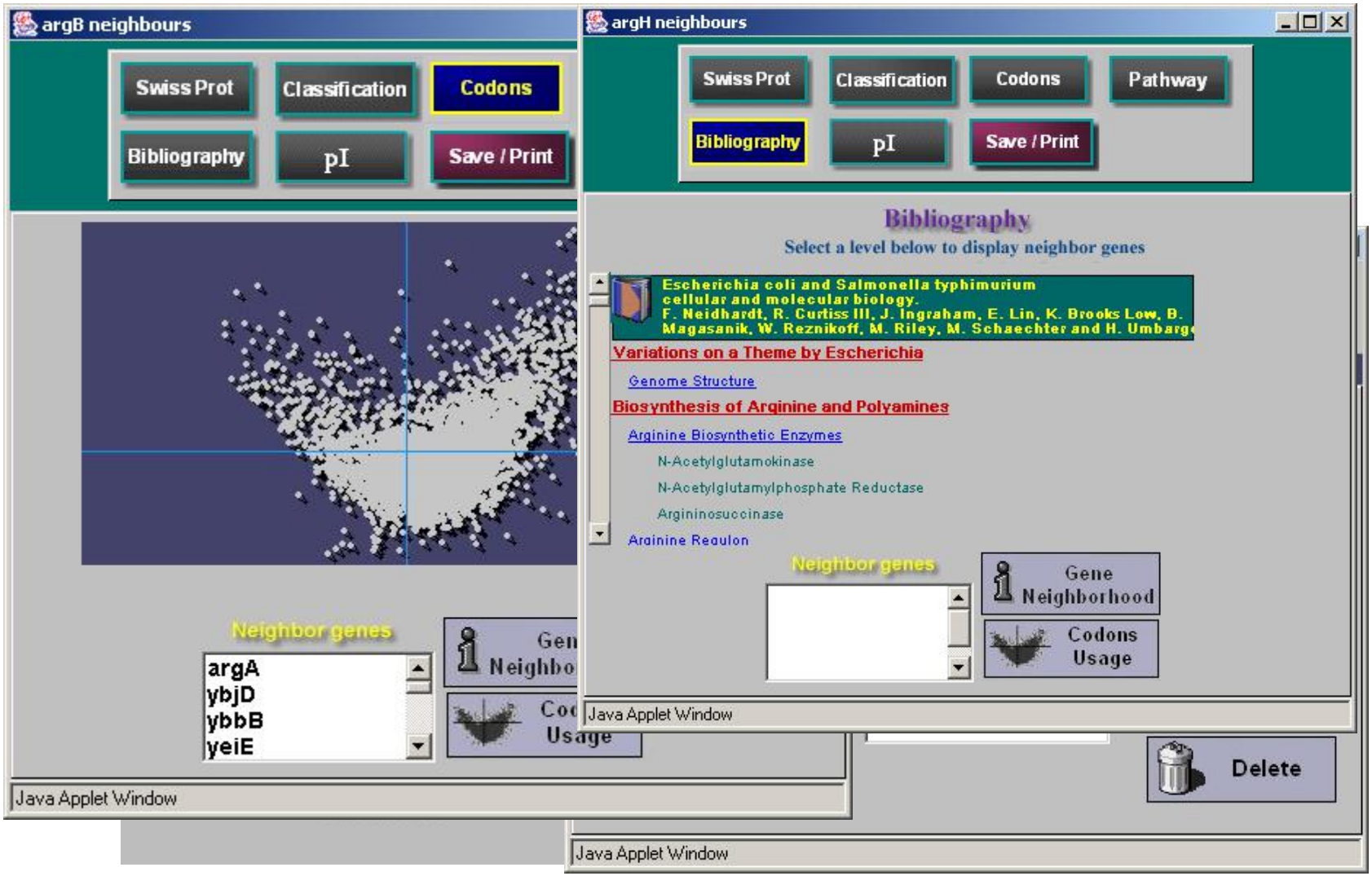

Visant [Hu *et al.*, 2005]

# Intégration : vers une vue unifiée

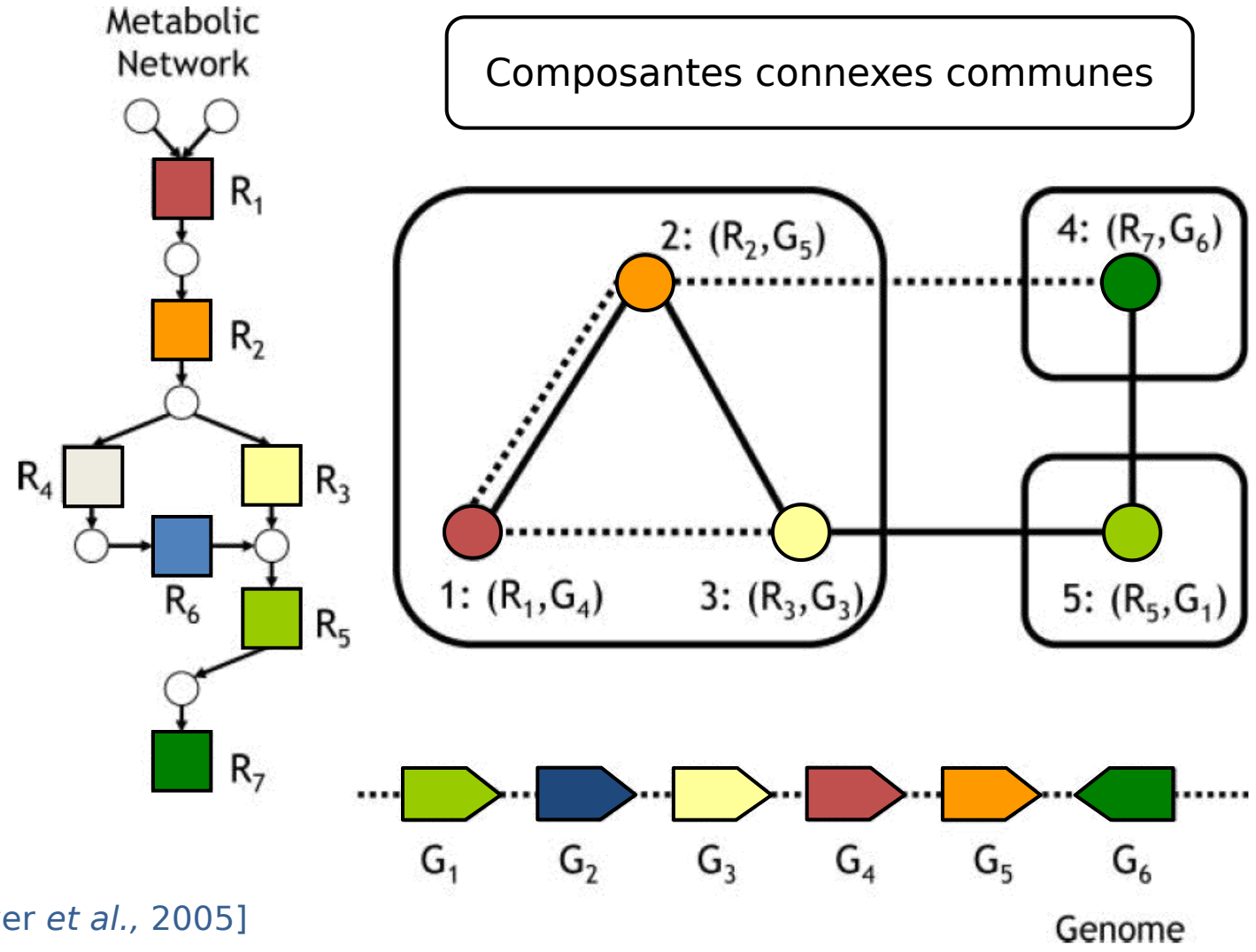
- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virutel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
 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

# Exploration visuelle de voisinages



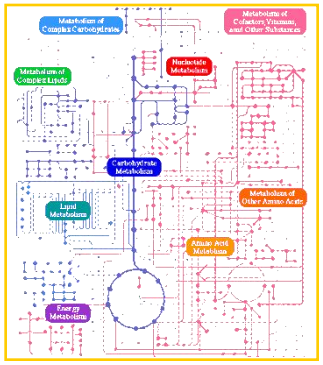


# Recoupement de voisinages : approche basée sur les graphes



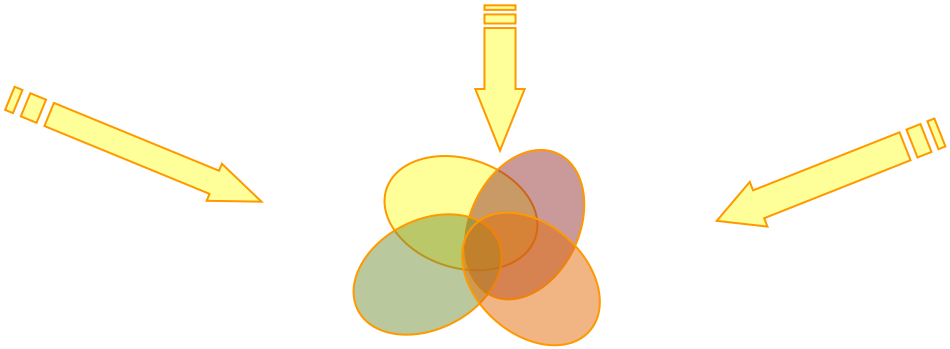

[Boyer *et al.*, 2005]

# Recoupement de voisinages : approche ensembliste

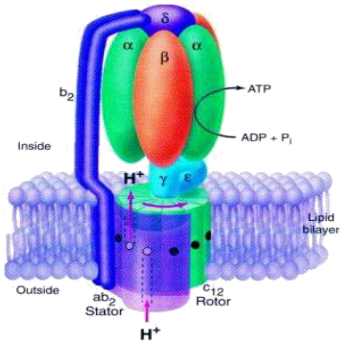


voies métaboliques

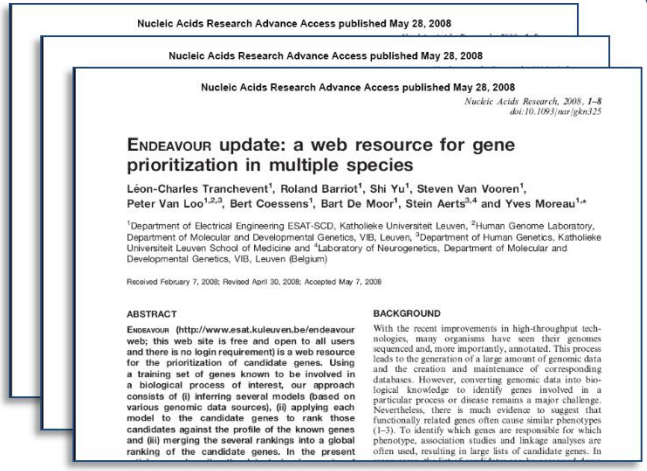
localisation chromosomique



ensembles de gènes




complexes protéiques



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/gkn325

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

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

[1]Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, [2]Human Genome Laboratory, Department of Molecular and Developmental Genetics, VIB, Leuven, [3]Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine and [4]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

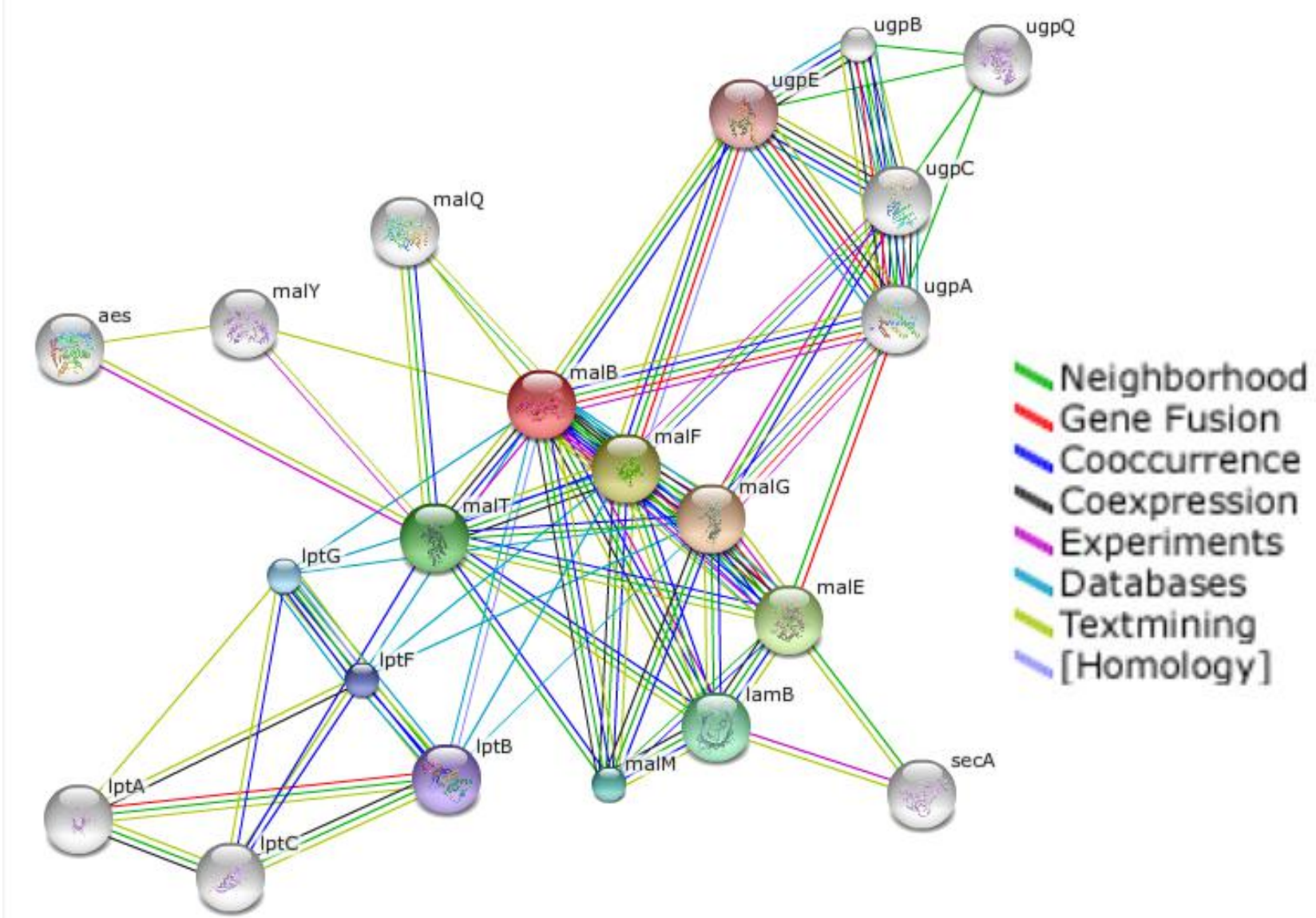
With the recent improvements in high-throughput technologies, many organisms have seen their genomes sequenced and, more importantly, annotated. This process leads to the generation of a large amount of genomic data and the creation and maintenance of corresponding databases. However, converting genomic 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

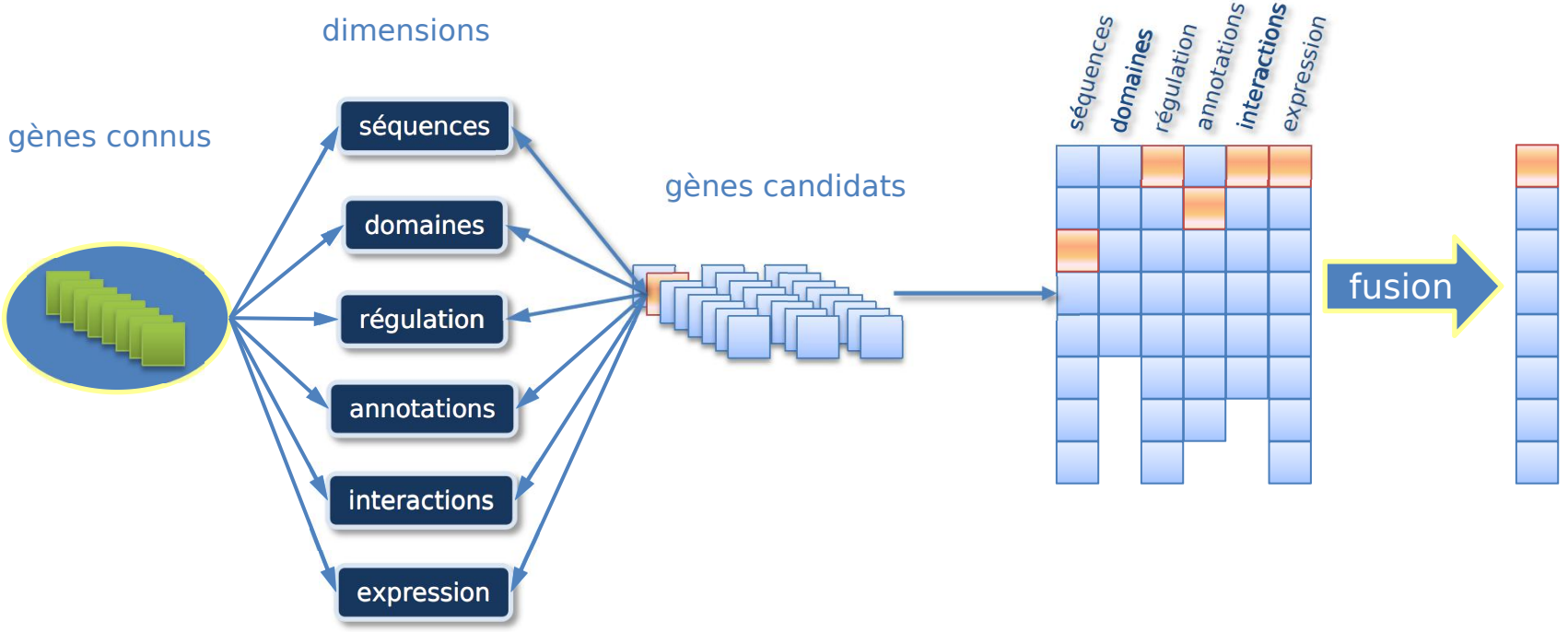

Gene Ontology

# Fusion : approche basée sur les graphes




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

# Approche apprentissage automatique / fusion : dimensions & dissimilarités. Priorisation et clustering




[Aerts *et al.*, 2006]

# Intégration : vers une vue unifiée

- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virutel : 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, attibuts et unités
  - XML, FASTA, EMBL/GenBank, SWISSPROT

- Architecture et accès
  - SGBD, SOAP, REST, pipeline (galaxy, ergatis)

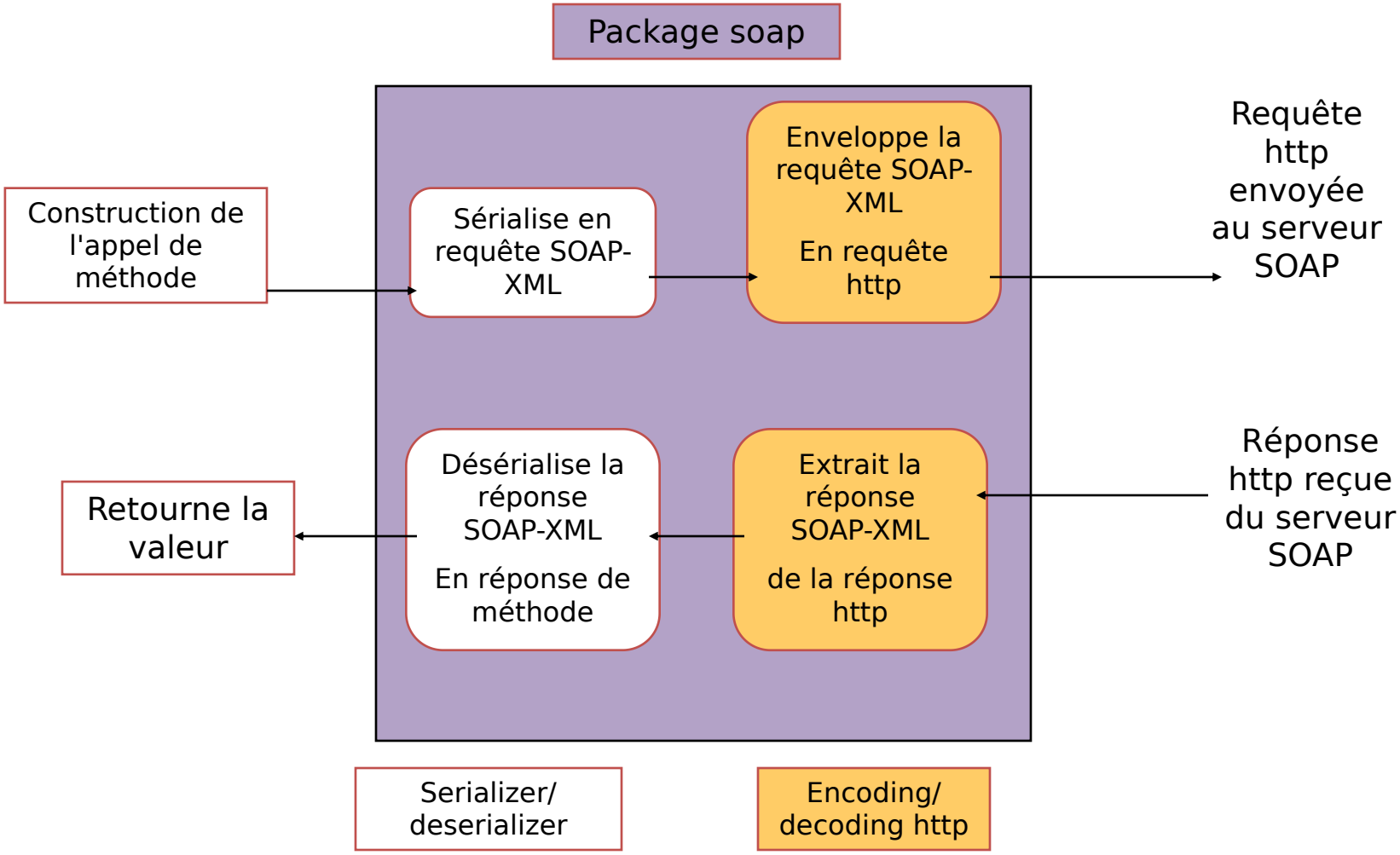
- Hétérogénéité des systèmes et des représentations
- Besoin de standards
- Plusieurs solutions :
  - Format d'échange : XML, JSON, ...
  - Protocole d'accès : services Web, SOAP, REST, ...

- 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)
  - protocol 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