INTÉGRATION DE DONNÉES *OMICS* EN UTILISANT UNE DISTANCE BASÉE SUR LE VOISINAGE DES GÈNES

Philippe BORDRON, Damien EVEILLARD and Irena RUSU

ComBi, LINA - UMR 6241, Université de Nantes

Le 08/09/2010

UNIVERSITÉ DE NANTES

lina LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

cnrs

MOTIVATIONS

Une fonction métabolique est provoquée par des gènes proches sur le génome.

- Galperin et Koonin (2000)
- Rison et al. (2002).
- Simeonidis et al. (2003)
- Kovacs et al. (2009)
- ...

MOTIVATIONS

Une f Métabolique *ue est provoquée par des gènes*
es sur le génome.

- Galperin et Koonin (2000)
- Rison et al. (2002).
- Simeonidis et al. (2003)
- Kovacs et al. (2009)
- ...

MOTIVATIONS

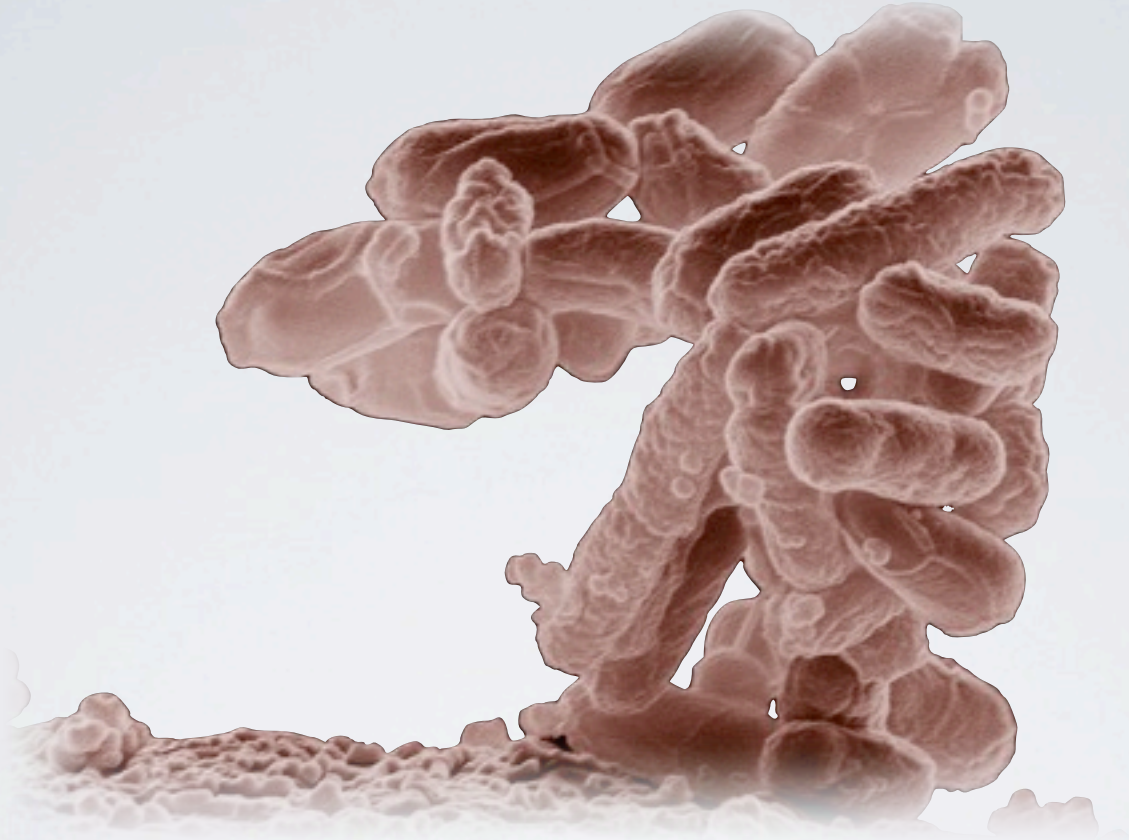
Une f Métabolique *ue est pr* Génomique *ènes*
es sur le g

- Galperin et Koonin (2000)
- Rison et al. (2002).
- Simeonidis et al. (2003)
- Kovacs et al. (2009)
- ...

Une p *gènes*

Approche intégrative

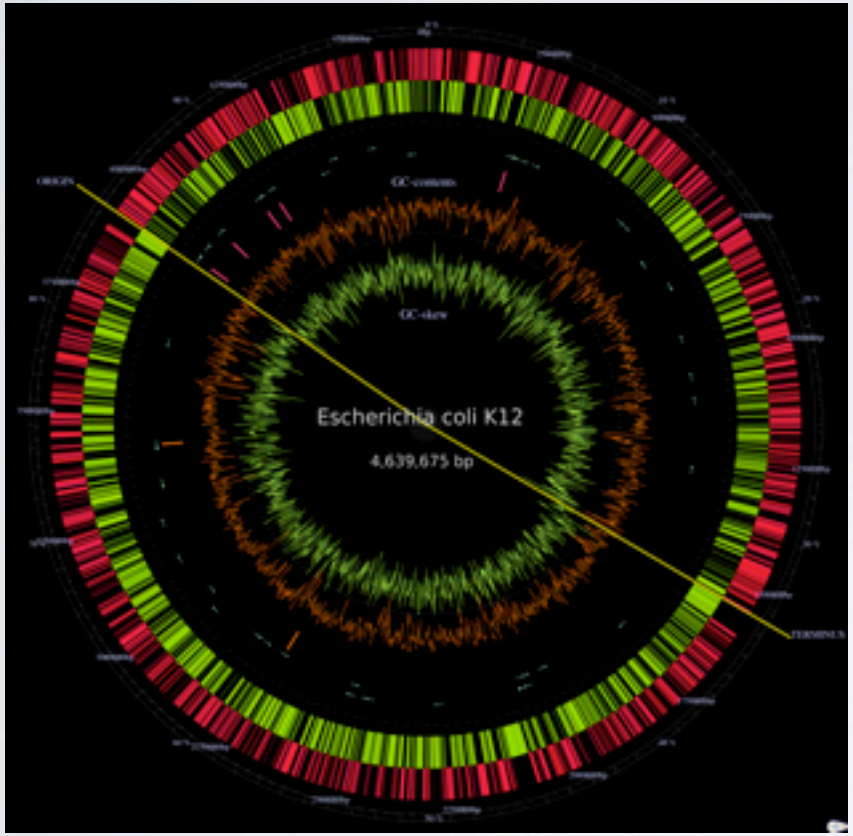
- Galper
- Rison et
- Simeonic
- Kovacs et al. (2009)
- ...



Low-temperature electron micrograph of a cluster of E. coli bacteria, magnified 10,000 times. Each individual bacterium is oblong shaped (ARS Image Gallery Image Number K11077-1).

E. COLI : ORGANISME D'ÉTUDE

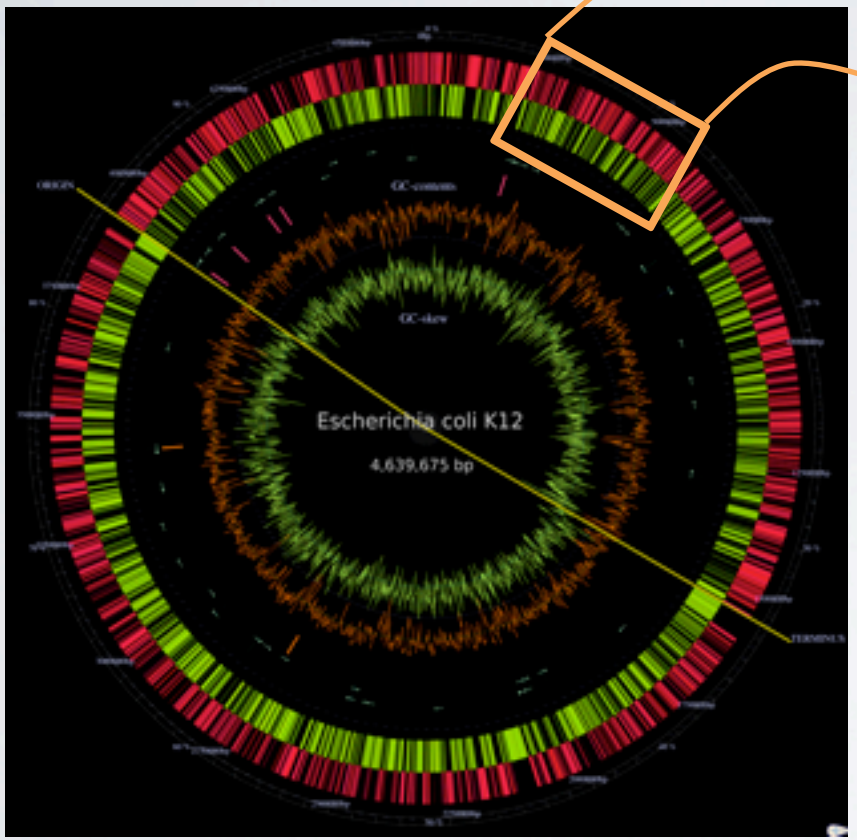
LE GÉNOME

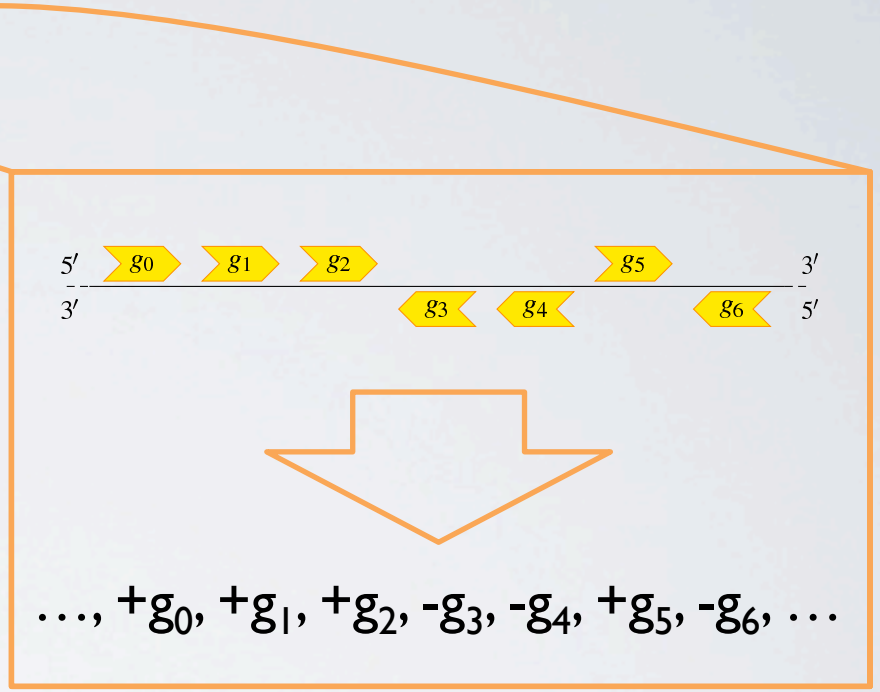


E. coli K-12 MG1655 (31/03/2008)

- GenBank id : U00096
- 4242 gènes codants.

LE GÉNOME

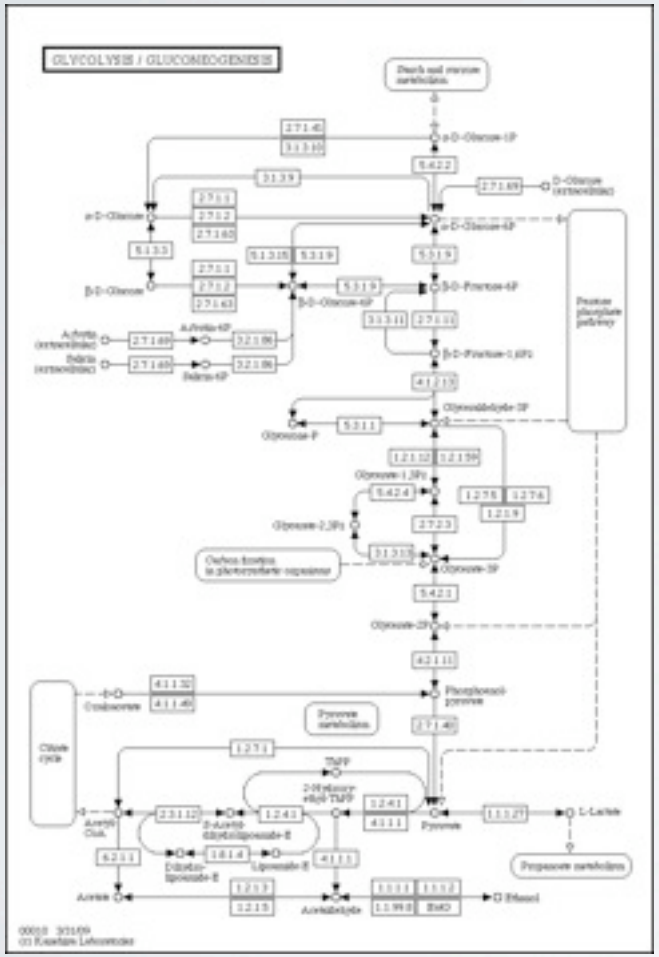




E. coli K-12 MG1655 (31/03/2008)

- GenBank id : U00096
- 4242 gènes codants.

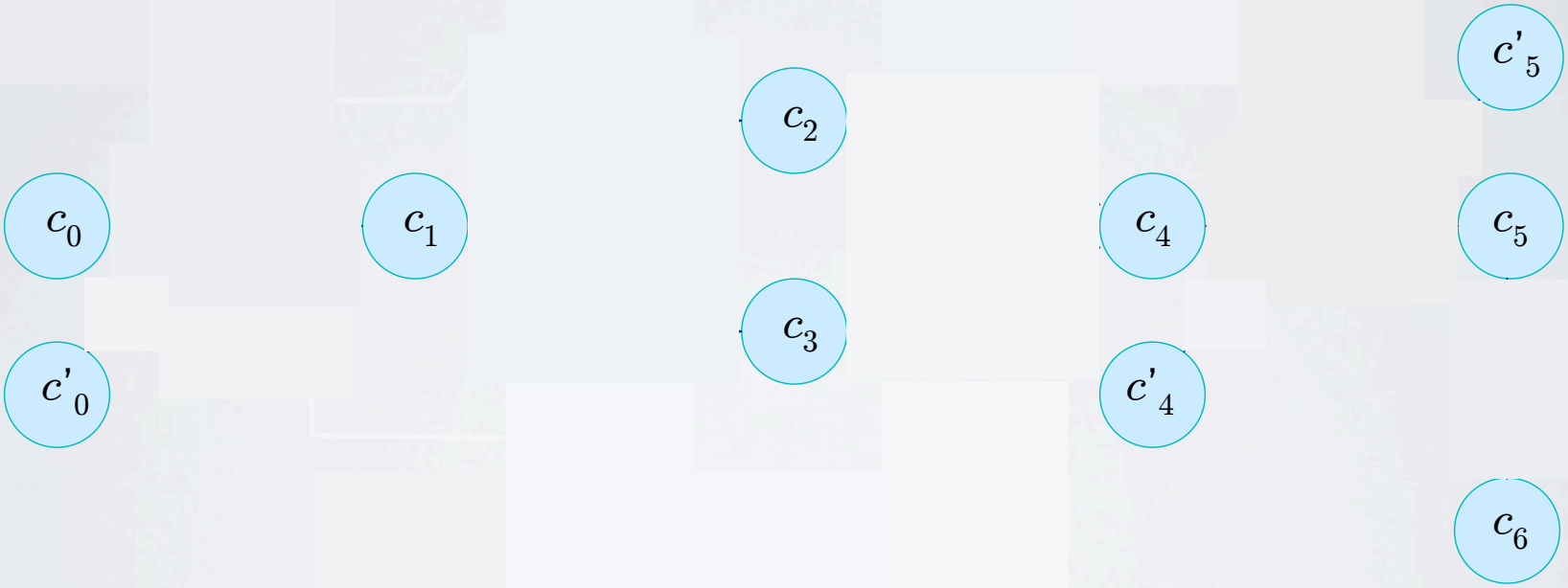
LE RÉSEAU MÉTABOLIQUE



KEGG (21/08/2008)

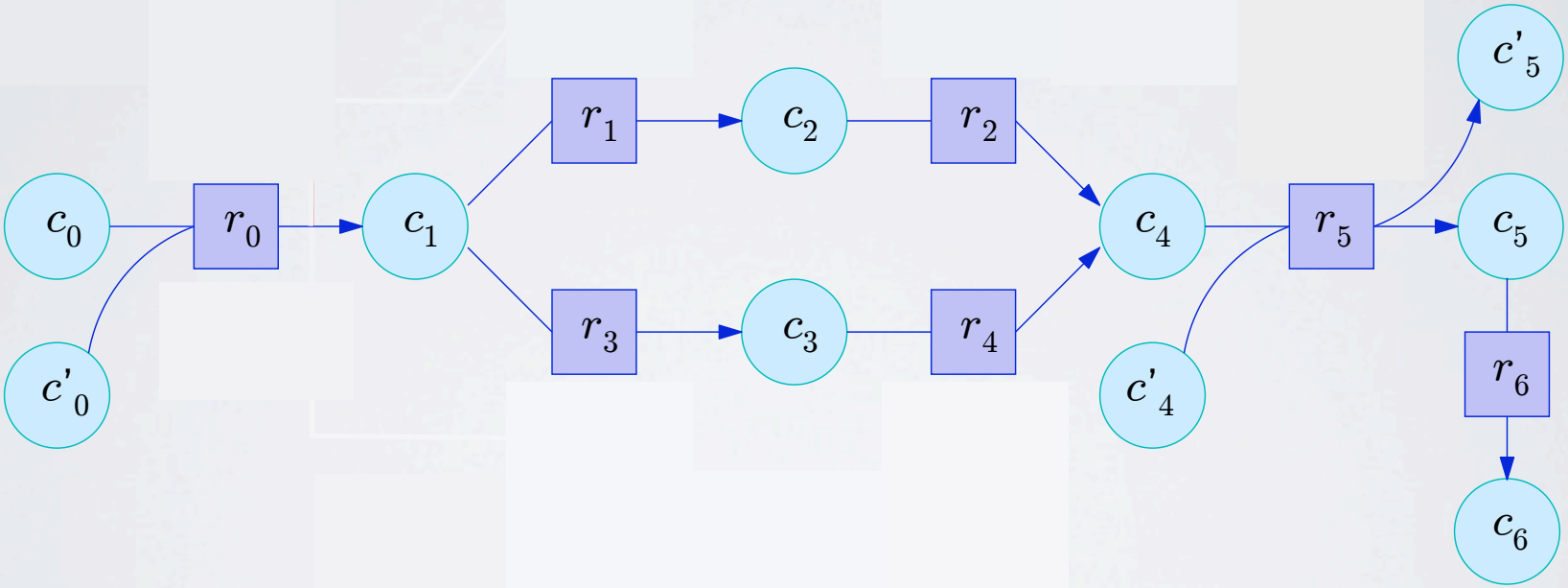
- KEGG id : eco
- 2 971 composés biochimiques
- 1131 réactions biochimiques
- 647 enzymes

INTÉGRATION : RELATION ENTRE GÉNOME ET RÉACTOME



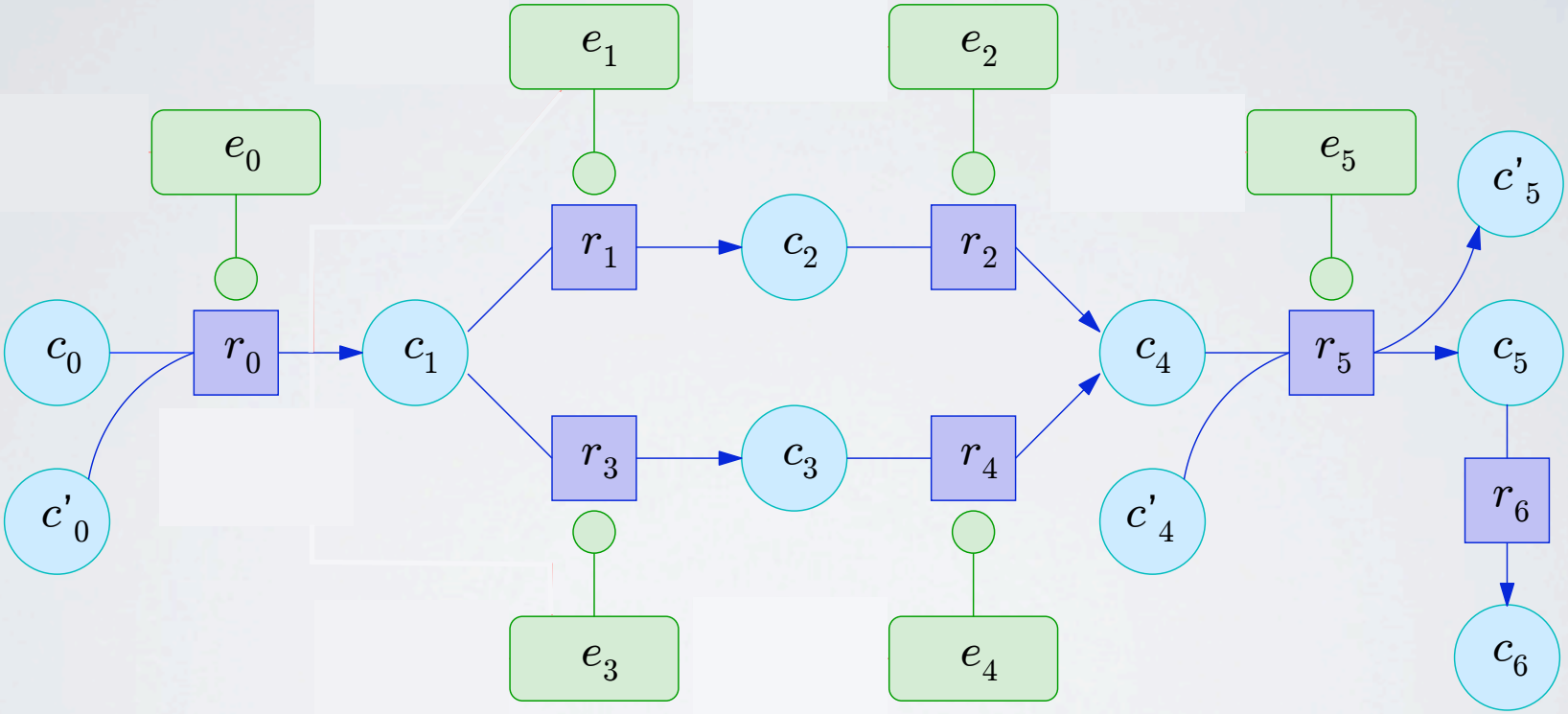
Un réseau métabolique fictif selon le formalisme SBGN

INTÉGRATION : RELATION ENTRE GÉNOME ET RÉACTOME



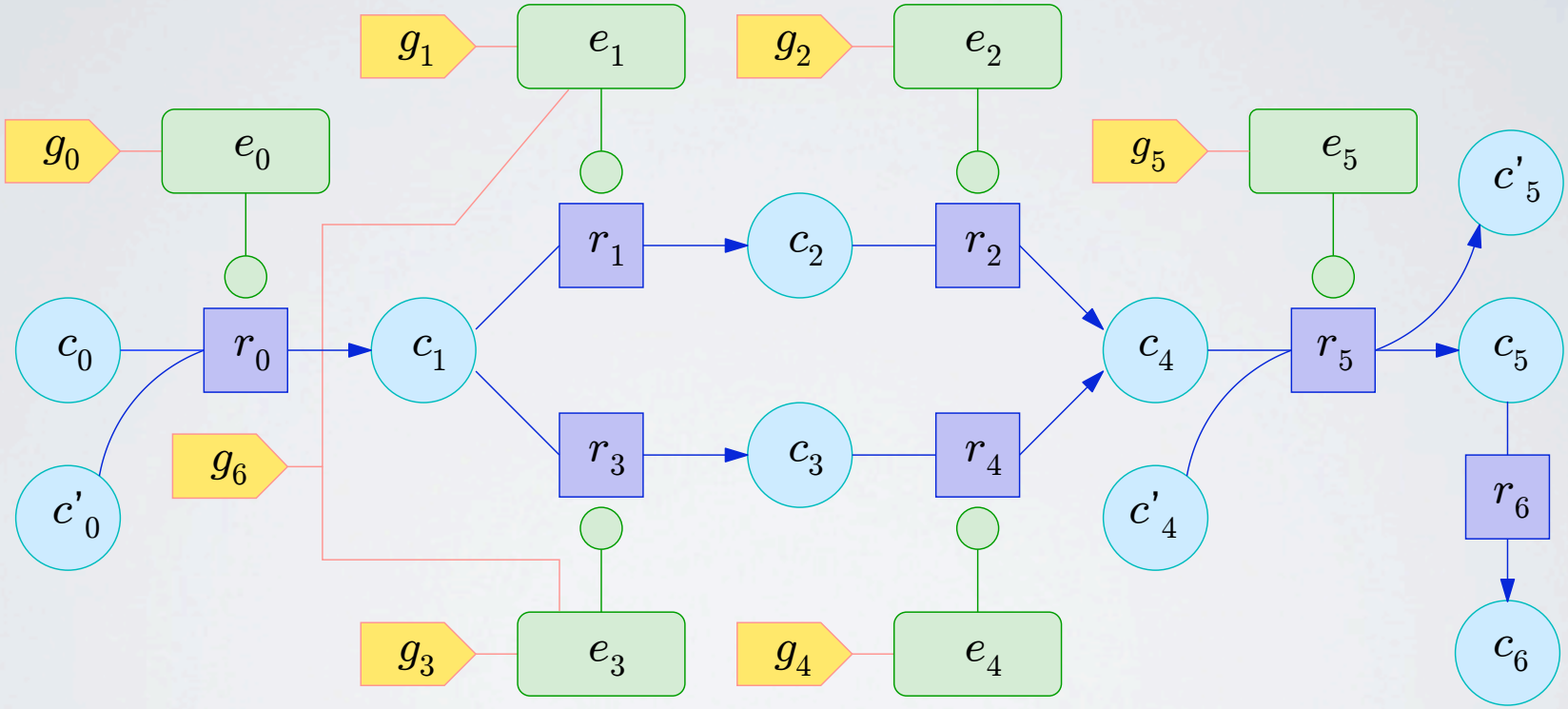
Un réseau métabolique fictif selon le formalisme SBGN

INTÉGRATION : RELATION ENTRE GÉNOME ET RÉACTOME



Un réseau métabolique fictif selon le formalisme SBGN

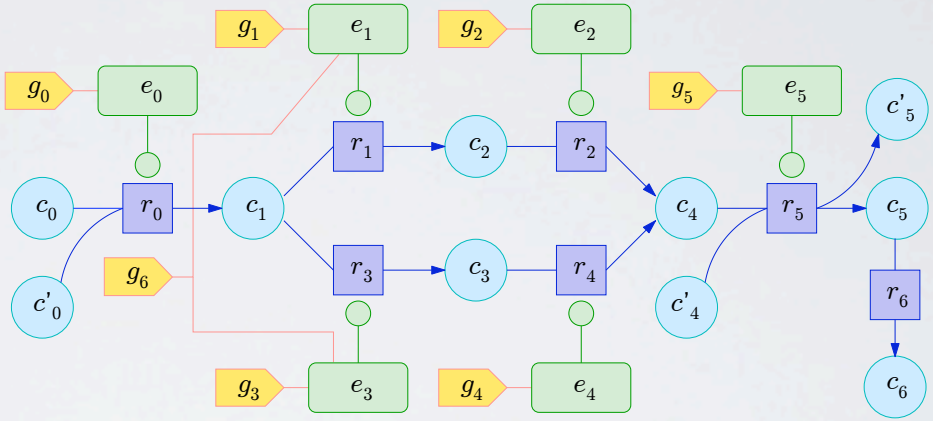
INTÉGRATION : RELATION ENTRE GÉNOME ET RÉACTOME

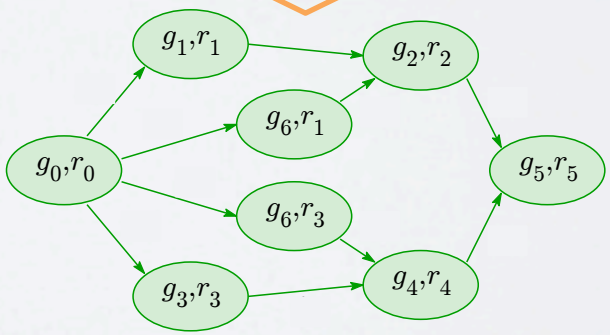


Un réseau métabolique fictif selon le formalisme SBGN

LE MODÈLE INTÉGRÉ G_{INT}

Un réseau métabolique selon le formalisme SBGN





Notre modèle intégré G_{int} résultant.

PROXIMITÉ GÉNOMIQUE

- Hypothèse:

  Une fonction métabolique est provoquée par des gènes proches sur le génome.

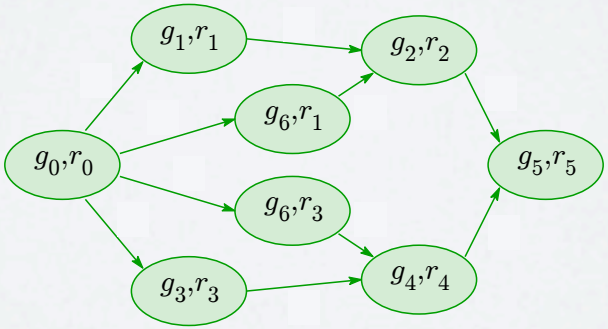
- Distance: nombre de sauts entre 2 gènes

PROXIMITÉ GÉNOMIQUE

- Hypothèse:

  Une fonction métabolique est provoquée par des gènes proches sur le génome.

- Distance: nombre de sauts entre 2 gènes

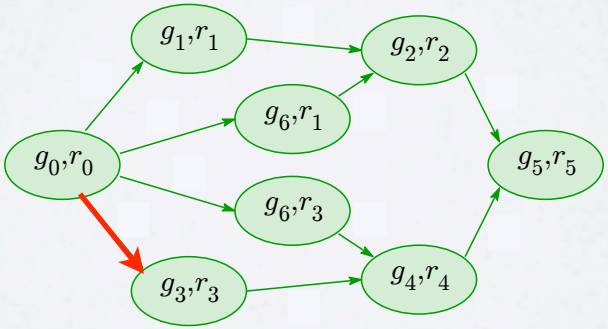


PROXIMITÉ GÉNOMIQUE

- Hypothèse:

  Une fonction métabolique est provoquée par des gènes proches sur le génome.

- Distance: nombre de sauts entre 2 gènes

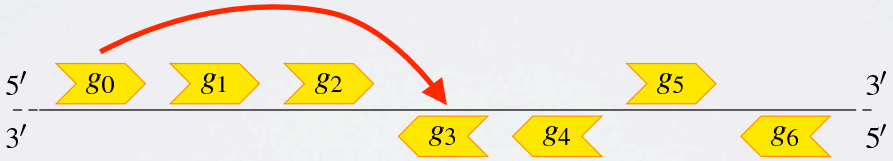


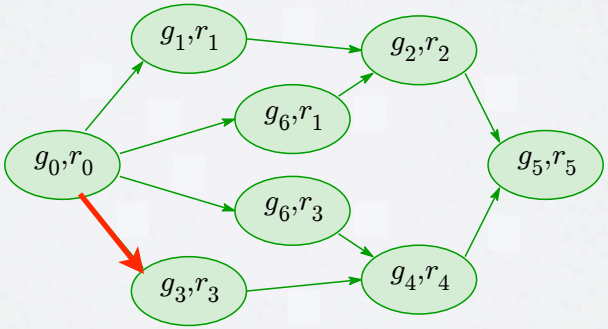
PROXIMITÉ GÉNOMIQUE

- Hypothèse:

  Une fonction métabolique est provoquée par des gènes proches sur le génome.

- Distance: nombre de sauts entre 2 gènes



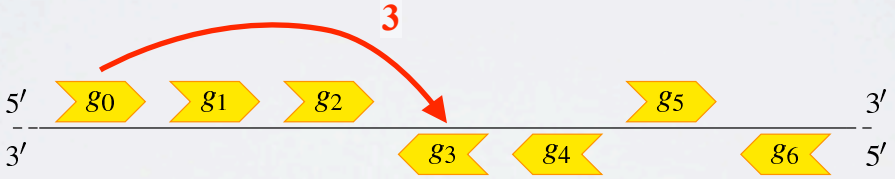


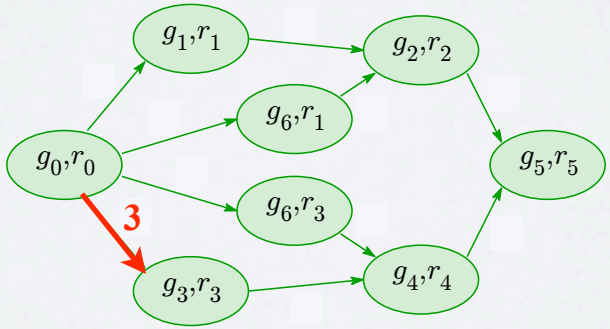
PROXIMITÉ GÉNOMIQUE

- Hypothèse:

  Une fonction métabolique est provoquée par des gènes proches sur le génome.

- Distance: nombre de sauts entre 2 gènes



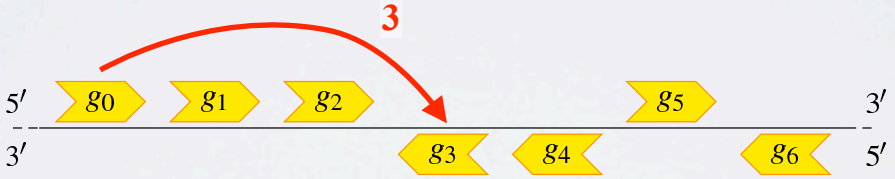


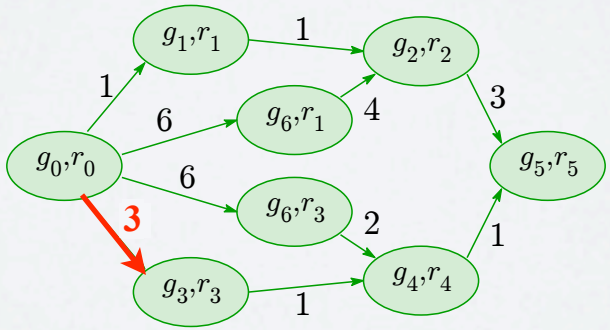
PROXIMITÉ GÉNOMIQUE

- Hypothèse:

  Une fonction métabolique est provoquée par des gènes proches sur le génome.

- Distance: nombre de sauts entre 2 gènes





LE MODÈLE INTÉGRÉ DE E. COLI

Statistiques topologiques

- 2343 sommets
- 13288 arcs

Statistiques biologiques

- 1049 réactions nettoyées
- 779 gènes
- 558 enzymes

LE MODÈLE INTÉGRÉ DE E. COLI

Statistiques topologiques

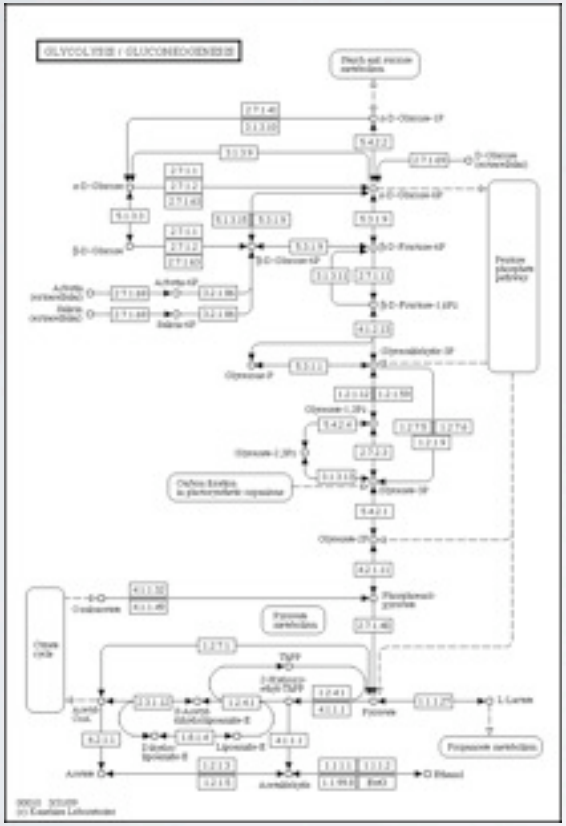
- 2343 somme
- 13288 a

Statistiques bio

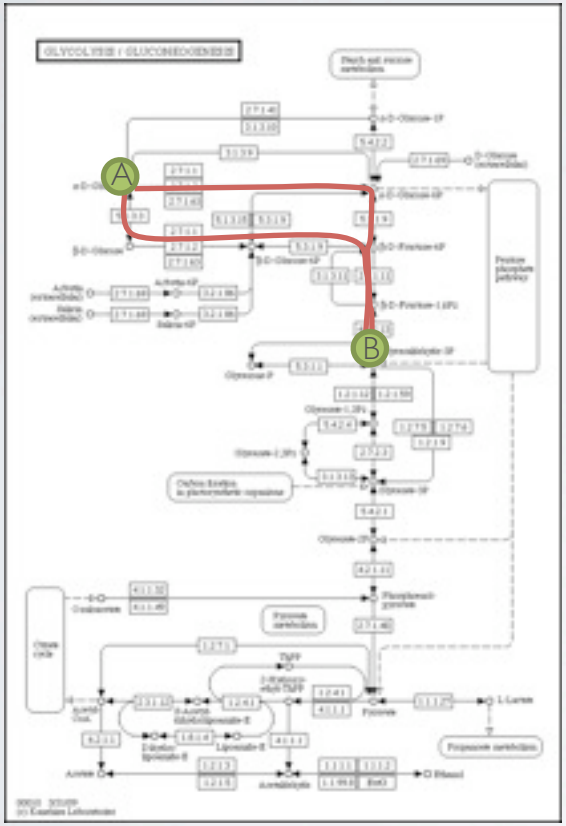
- 1049 ré
- 779 gènes
- 558 enzymes

Comment récupère-t-on de l'information biologique?

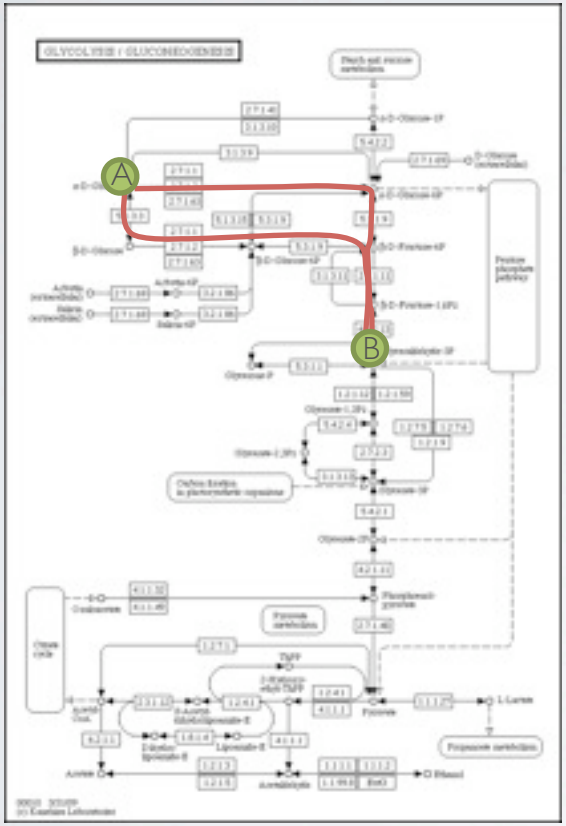
INTEGRATED PATHWAYS (IPS)

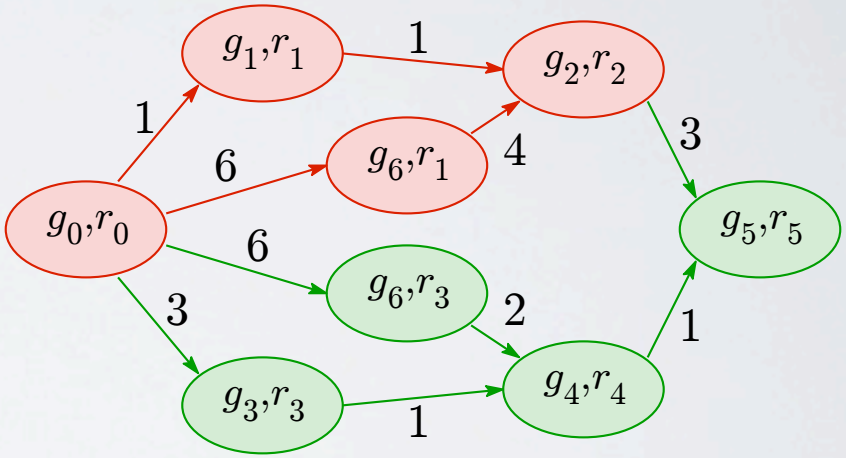


INTEGRATED PATHWAYS (IPS)

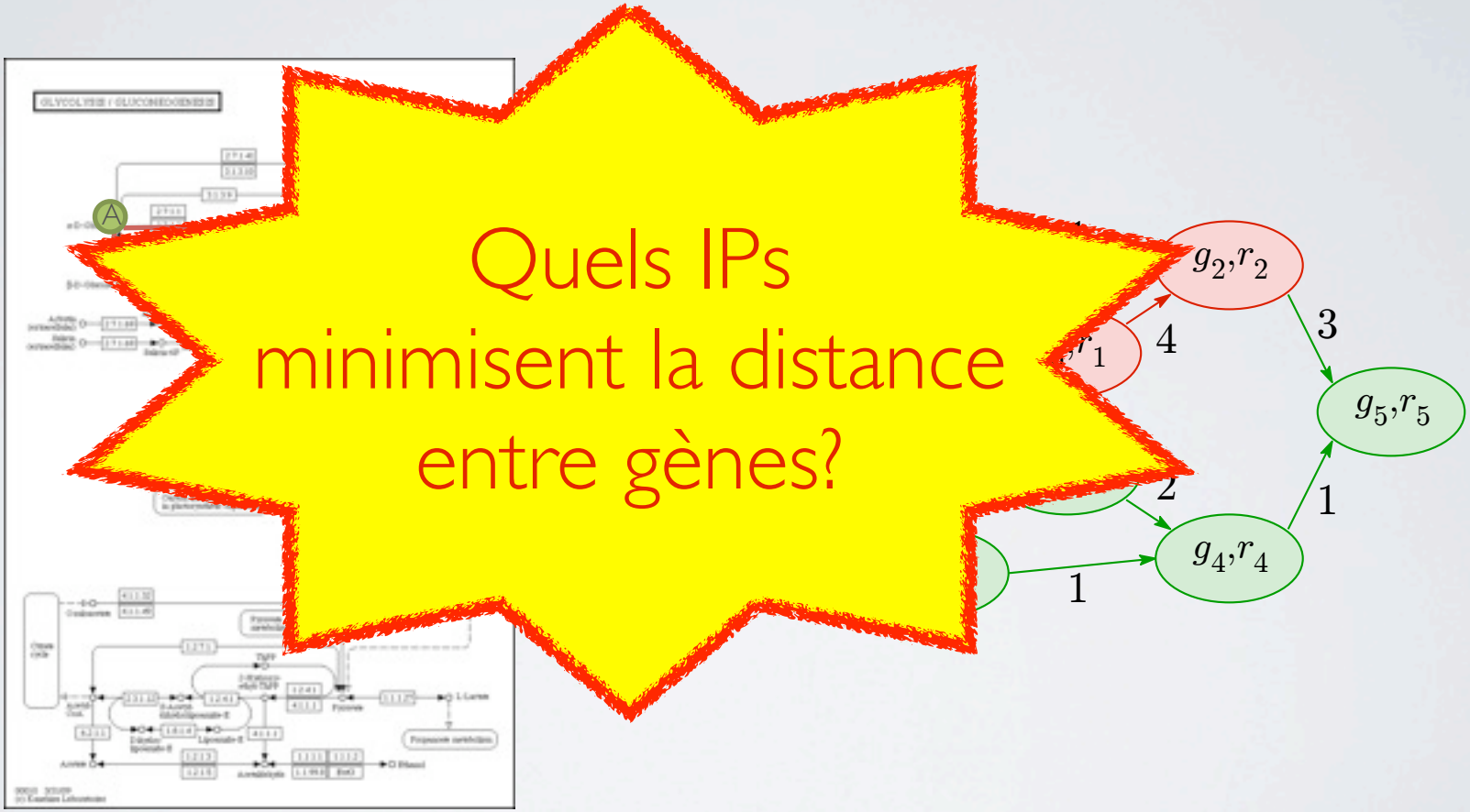


INTEGRATED PATHWAYS (IPS)





INTEGRATED PATHWAYS (IPS)



QUELLES MESURES

$$\overline{w} = \frac{\sum \text{poids des arcs des chemins}}{\#\text{réactions métaboliques utilisées}}$$

QUELLES MESURES

Mesure génomique
à minimiser

$$\overline{w} = \frac{\sum \text{poids des arcs des chemins}}{\#\text{réactions métaboliques utilisées}}$$

QUELLES MESURES

Mesure génomique
à minimiser

$$\overline{w} = \frac{\sum \text{poids des arcs des chemins}}{\#\text{réactions métaboliques utilisées}}$$

Mesure métabolique
à maximiser

QUELLES MESURES

Objectif : minimisation de $\overline{w}$

Mesure génomique
à minimiser

$$\overline{w} = \frac{\sum \text{poids des arcs des chemins}}{\#\text{réactions métaboliques utilisées}}$$

Mesure métabolique
à maximiser

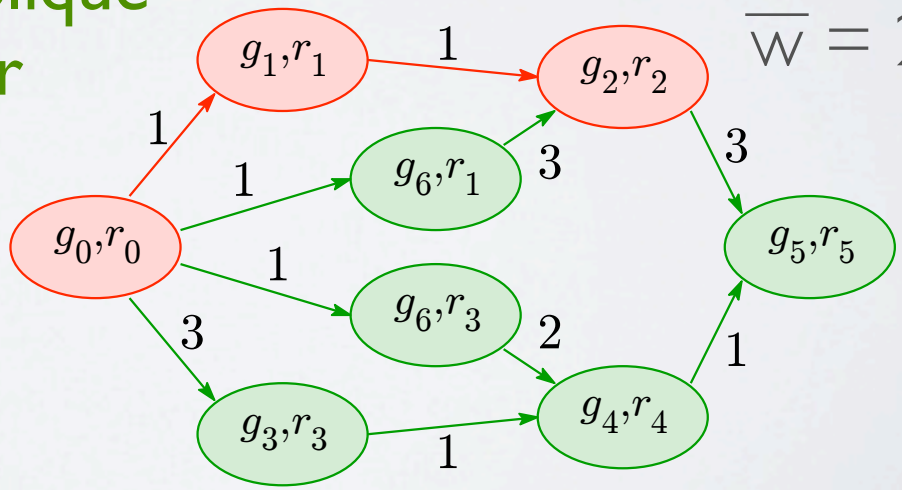
QUELLES MESURES

Objectif : minimisation de $\overline{w}$

Mesure génomique
à minimiser

$$\overline{w} = \frac{\sum \text{poids des arcs des chemins}}{\#\text{réactions métaboliques utilisées}}$$

Mesure métabolique
à maximiser



$\overline{w} = 2/3$

MINIMISATION DE $\overline{W}$

MINIMISATION DE $\overline{W}$

- Algorithme de recherche des *k* plus courts chemins sans cycle (Yen, 1970) de $\overline{w}$ minimum entre tous les couples de réactions.

MINIMISATION DE $\overline{W}$

- Algorithme de recherche des *k* plus courts chemins sans cycle (Yen, 1970) de $\overline{w}$ minimum entre tous les couples de réactions.
- 439 382 IPs obtenus

MINIMISATION DE $\overline{W}$

- Algorit… chemins sans cycle… re tous les couple…
- 439 382 IPs

Comment intéprète-t-on biologiquement ces IPs ?

ON RETROUVE DE L'INFORMATION BIOLOGIQUE

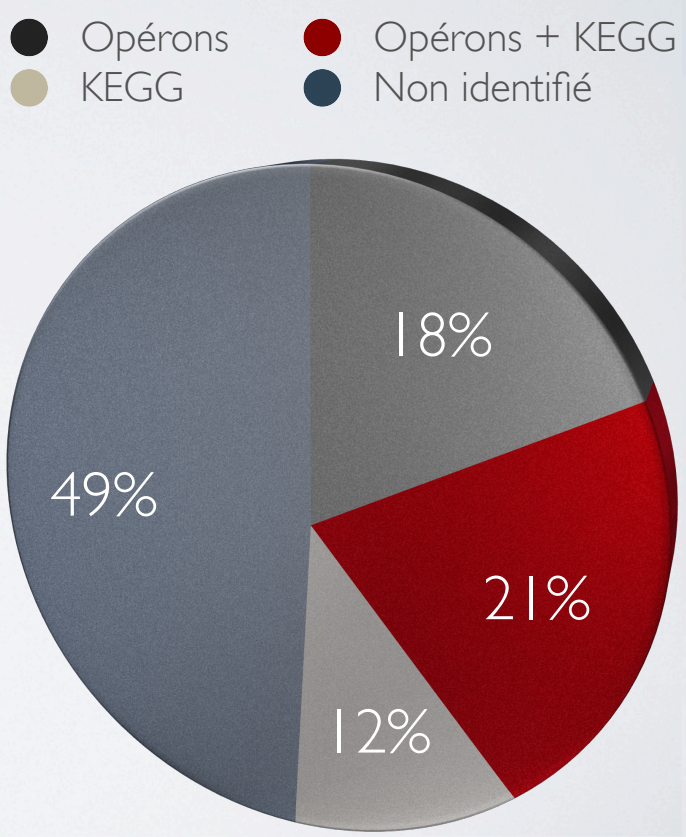
Benchmark:

- 135 opérons métaboliques de RegulonDB (30/01/2009)
- 99 modules de KEGG (21/08/2008)

ON RETROUVE DE L'INFORMATION BIOLOGIQUE

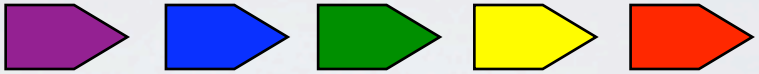
Benchmark:

- 135 opérons métaboliques de RegulonDB (30/01/2009)
- 99 modules de KEGG (21/08/2008)



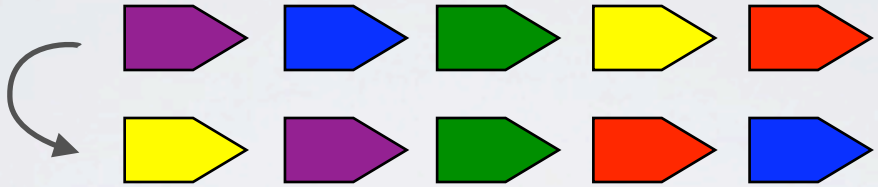
IMPACT GÉNOMIQUE

- Mélange du génome



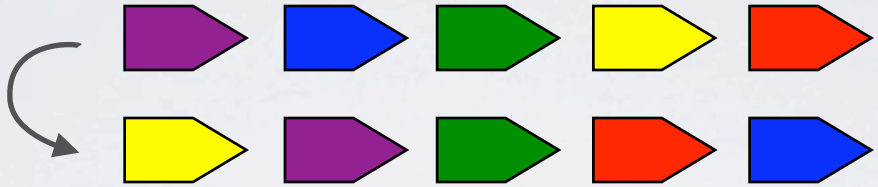
IMPACT GÉNOMIQUE

- Mélange du génome



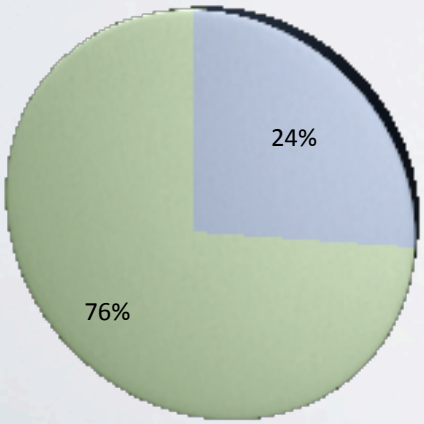
IMPACT GÉNOMIQUE

- Mélange du génome

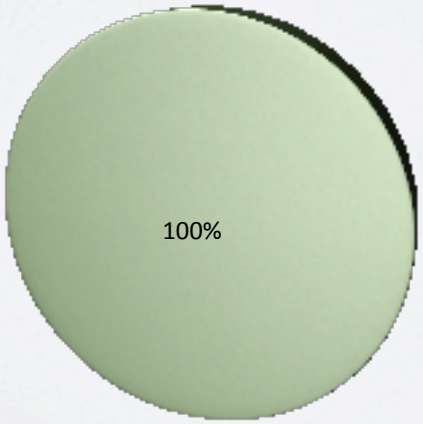


- Proportion d'opérons retrouvés

Génome non mélangé



Génome mélangé



exactemement matché (Jaccard à 100%)
non matché exactement

IMPACT MÉTABOLIQUE

- Mélange du métabolisme [Maslov et Sneppen (2002)]

  On change l'enchaînement des réactions en conservant le degré des noeuds.

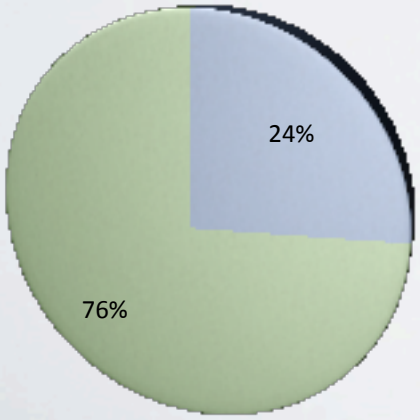
IMPACT MÉTABOLIQUE

- Mélange du métabolisme [Maslov et Sneppen (2002)]

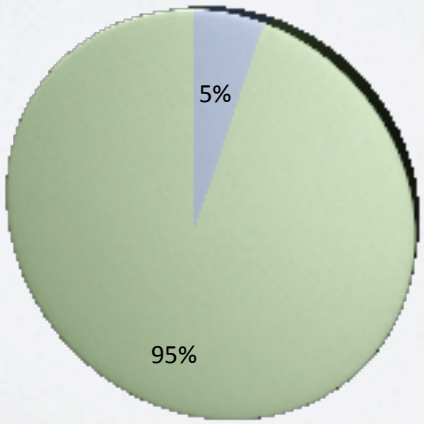
  On change l'enchaînement des réactions en conservant le degré des noeuds.

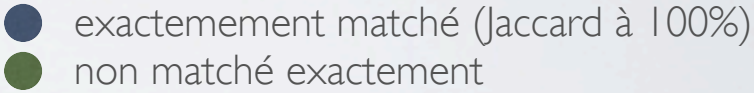
- Proportion d'opérons retrouvés

Métabolisme non mélangé



Métabolisme mélangé





$\overline{W}$ EST-IL DISCRIMINANT?

$\overline{w}$ EST-IL DISCRIMINANT?

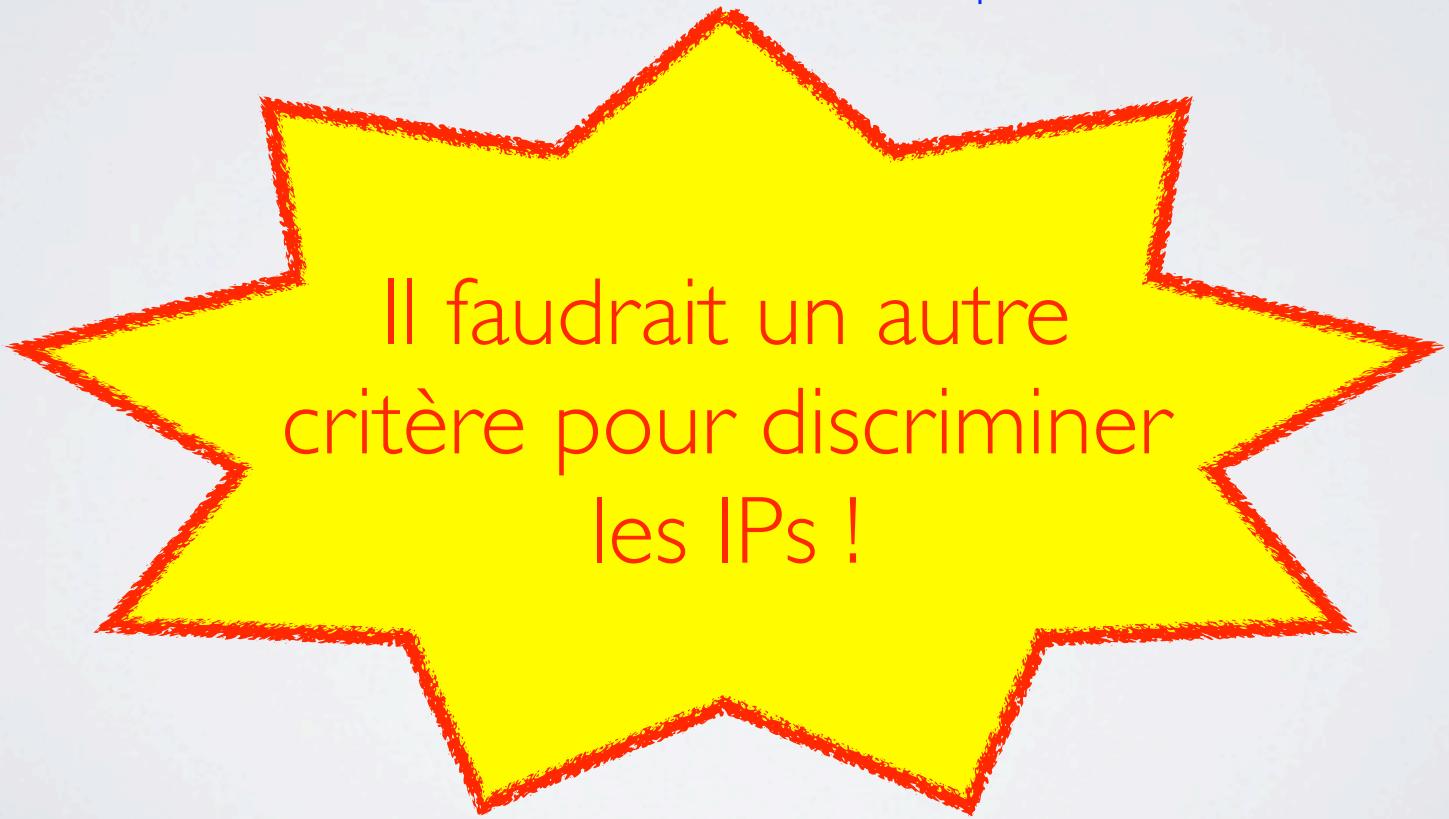
- Les opérons ont un $\overline{w}$ faible ($\overline{w} \leq 1$)

$\overline{w}$ EST-IL DISCRIMINANT?

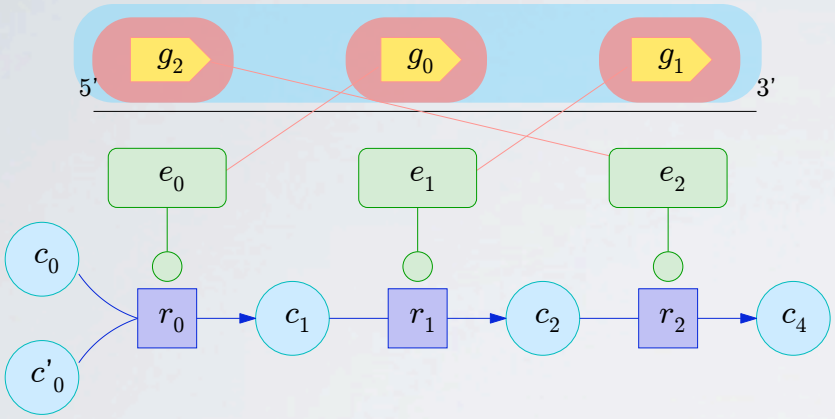
- Les opérons ont un $\overline{w}$ faible ($\overline{w} \leq 1$)
- 50% des IPs avec $\overline{w} \leq 1$ sont des opérons

$\overline{w}$ EST-IL DISCRIMINANT?

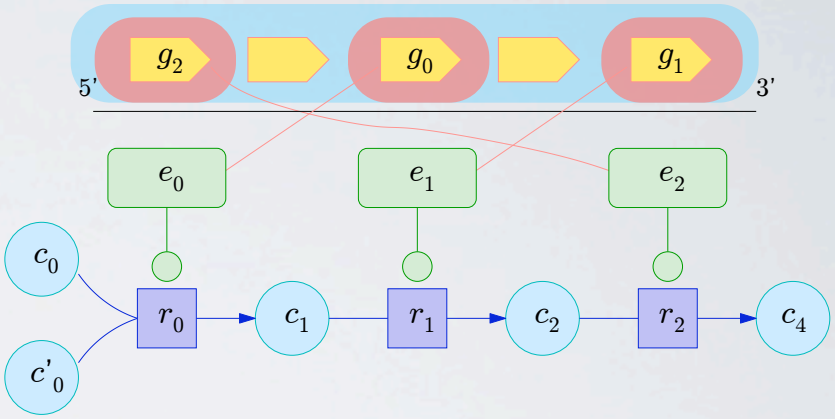
- Les opérons ont un $\overline{w}$ faible ($\overline{w} \leq 1$)
- 50% des IPs avec $\overline{w} \leq 1$ sont des opérons



DENSITÉ DES IPS

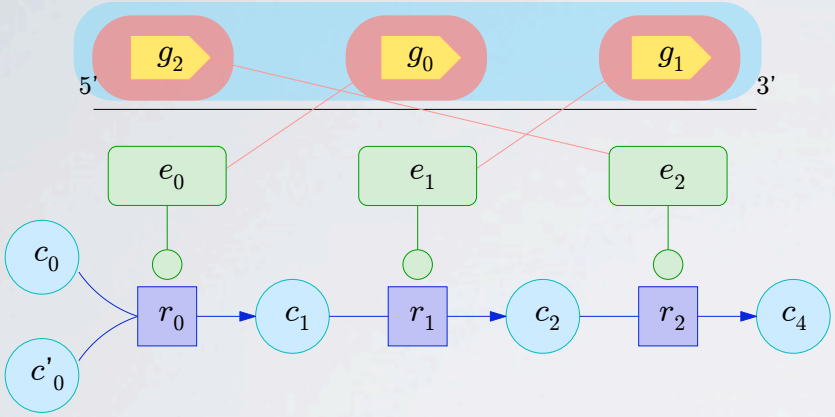


IP de densité maximale (d=1)

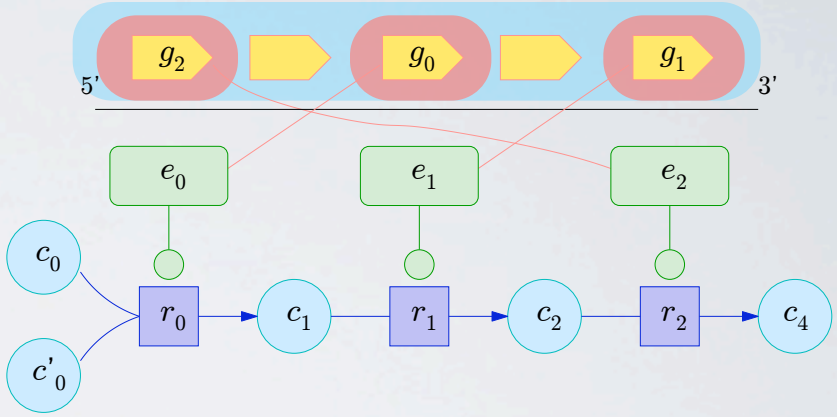


IP de densité plus faible (d=3/5)

DENSITÉ DES IPS



IP de densité maximale (d=1)

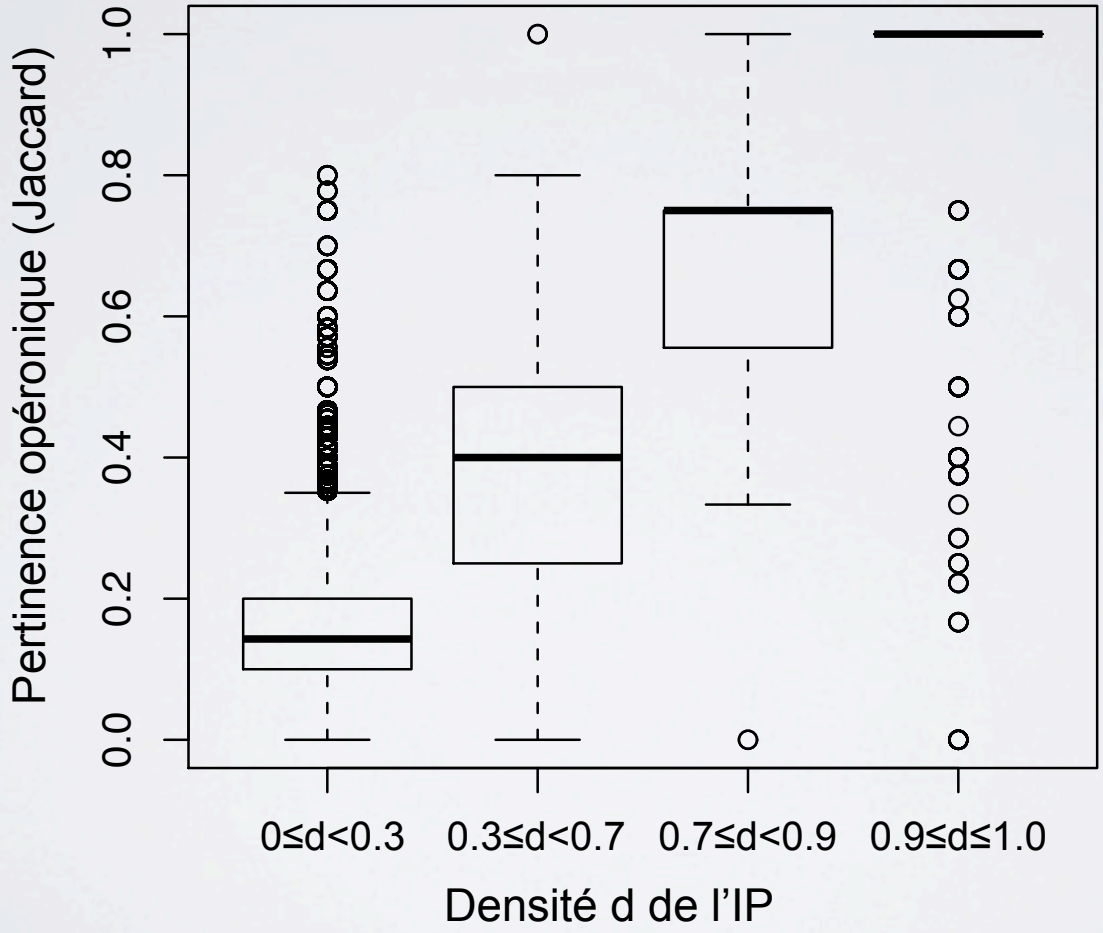


IP de densité plus faible (d=3/5)

$$\text{densité} = \frac{\#\text{gènes dans l'IP}}{\#\text{gènes dans l'intervalle correspondant}}$$

DENSITÉ DES IPS

Matching exact des IPs par classe de densité



CONCLUSION

CONCLUSION

- Approche intégrative généralisable basée sur la notion de distance.

CONCLUSION

- Approche intégrative généralisable basée sur la notion de distance.
- On retrouve de l'information opéronique de façon satisfaisante en utilisant $\overline{w}$ comme un critère de recherche et la densité pour la discriminer .

PERSPECTIVES

PERSPECTIVES

- Tester d'autres informations biologiques (KEGG).

PERSPECTIVES

- Tester d'autres informations biologiques (KEGG).
- Application de la méthode avec une distance de coexpression entre gènes.

PERSPECTIVES

- Tester d'autres informations biologiques (KEGG).
- Application de la méthode avec une distance de coexpression entre gènes.
- Approche par contraintes pour réduire le nombre d'IPs.

PERSPECTIVES

- Tester d'autres informations biologiques (KEGG).
- Application de la méthode avec une distance de coexpression entre gènes.
- Approche par contraintes pour réduire le nombre d'IPs.
- Travail avec l'INRIA de Rennes et le Pr. A. Mass (Université de Santiago au Chili) pour la prédiction microbienne de l'assimilation du cuivre.