

École thématique CNRS

Modélisation Formelle de Réseaux de Régulation Biologique

Ile de Porquerolles du 4 au 9 juin 2023

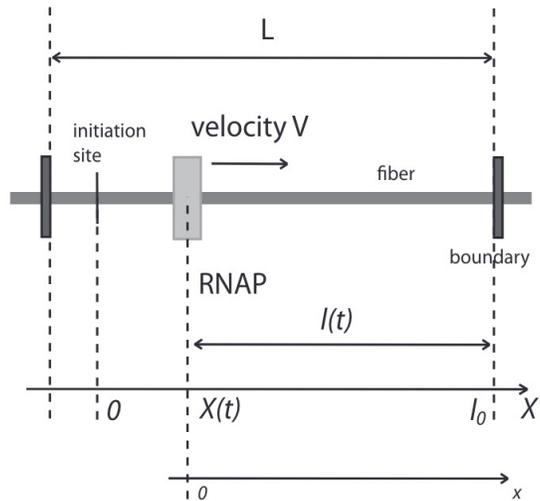
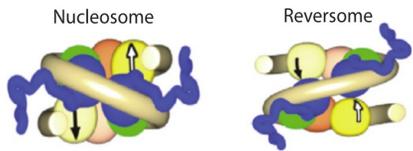
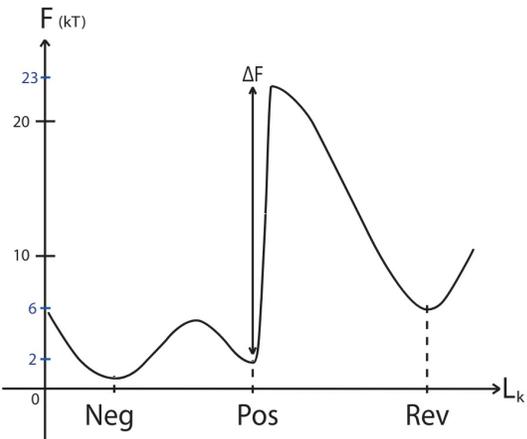
Christophe Bécavin, PhD
Maître de conférences

UNIVERSITÉ
CÔTE D'AZUR 

Machine learning pour la reconstruction des
réseaux biologiques à l'aide de données omiques



Développement de pipeline bioinformatique pour l'analyse de données multi-omiques



$$\frac{\partial \Theta}{\partial t} = D \frac{\partial^2 \Theta}{\partial x^2},$$

Becavin et al., BioPhysical Journal 2010

Highly dynamic genome viewer

Easily build your multi "omics" website

RNASeq, Arrays, Proteomics, ChipSeq, etc...

HeatMap viewer with common table analysis tools

Network visualization tools

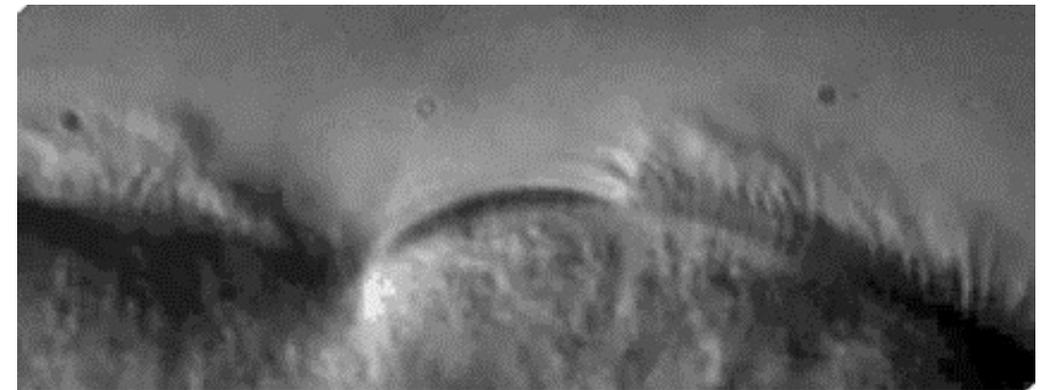
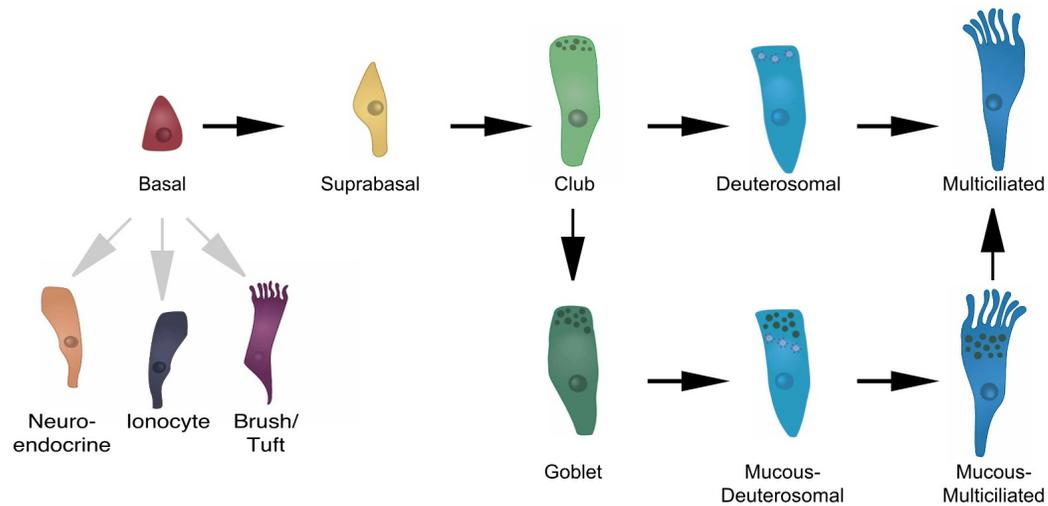
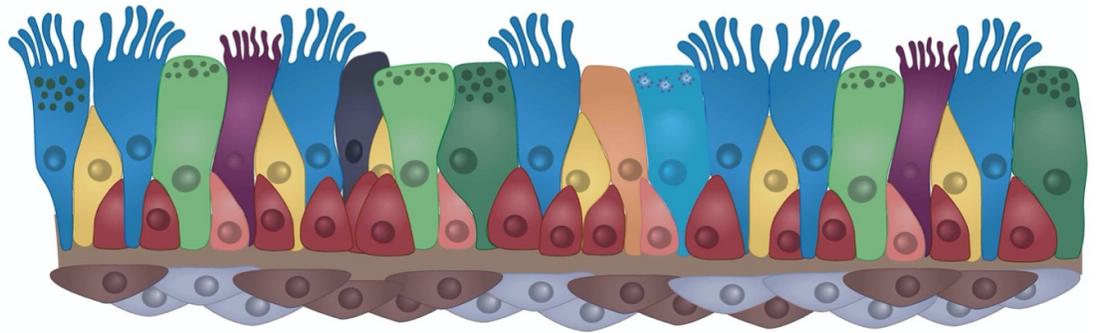
Single sourcing – Same code
Eclipse RCP -> Eclipse RAP

Web Application Desktop Application

Easy import of genomes and annotations

Bacnet: Une plateforme d'intégration de données omiques (Danès et al. Bioinformatics 2021)

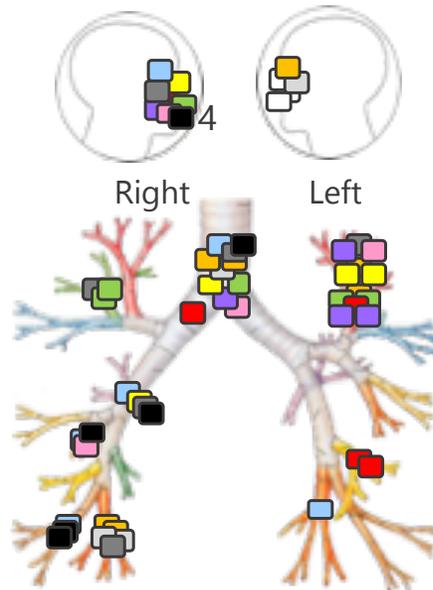
The human upper airway epithelium



Human Lung Cell Atlas

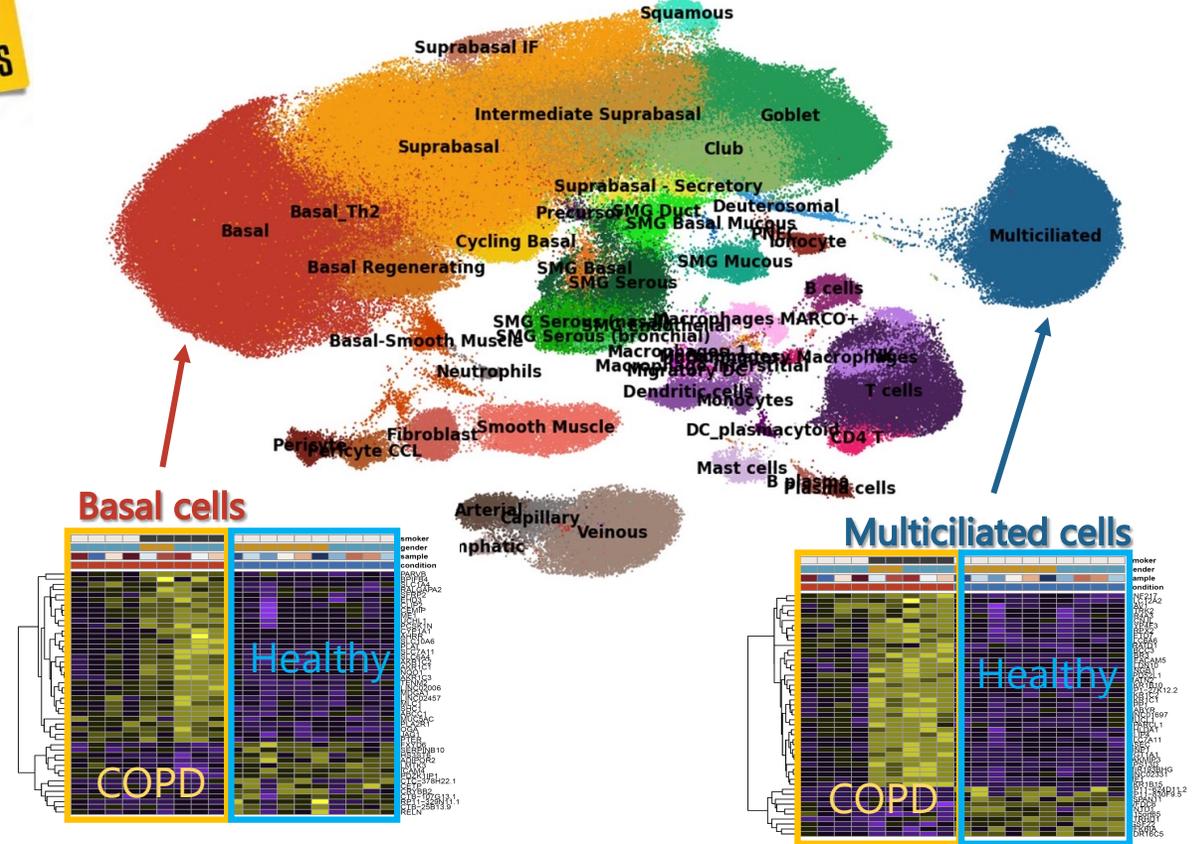


HUMAN
CELL
ATLAS



(CHU of Nice)

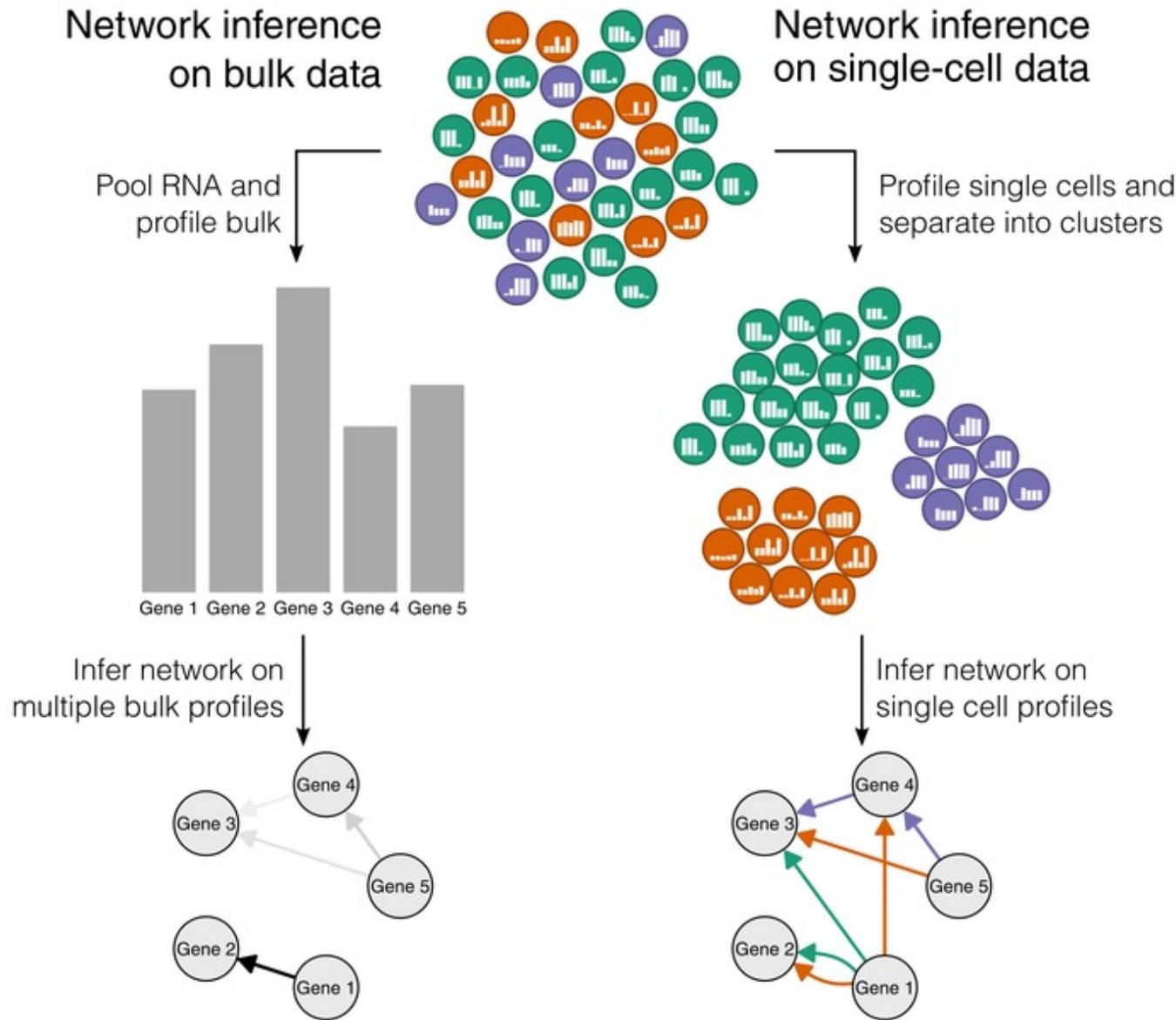
WORK IN
PROGRESS



- Deprez et al, AJRCCM 2020
- Wungnak et al., Nature Medecine 2020
- Muus et al., Cell 2020
- Sikkema, Nature Medecine (In Press)

Reconstruction d'atlas cellulaire des voies respiratoires

Reconstruction de réseaux de régulations à l'aide de données omiques



Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

- Reconstruction de réseaux biologiques à l'aide de données omiques
- Reconstruire par analyse de co-expression (TP)
- Etat de l'art des réseaux biologiques les plus étendus
- Apport du deep-learning à la biologie des systèmes



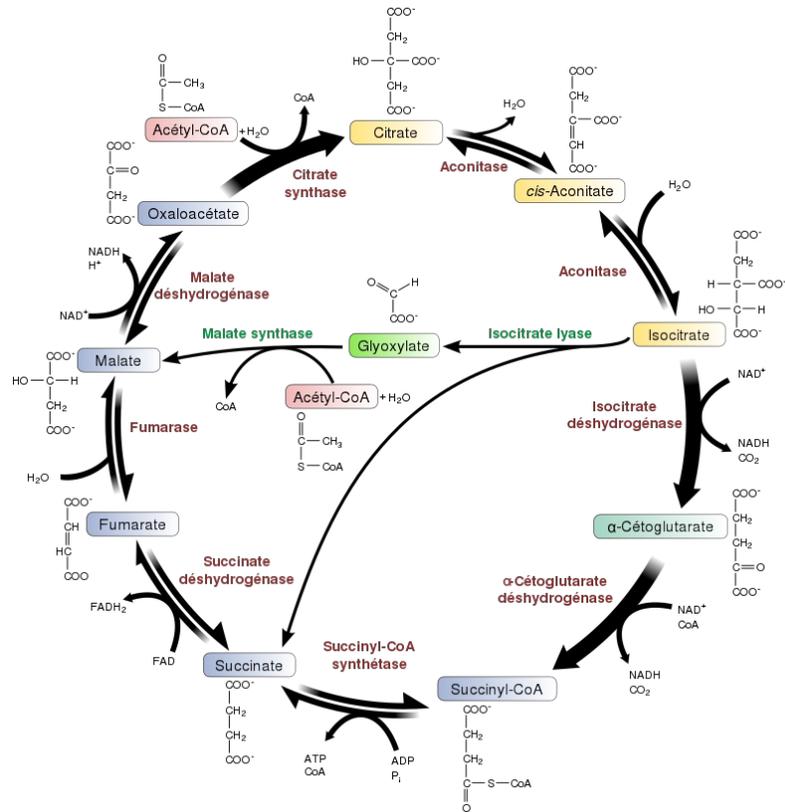
Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

- **Reconstruction de réseaux biologiques à l'aide de données omiques**
- Reconstruire par analyse de co-expression (TP)
- Etat de l'art des réseaux biologiques les plus étendus
- Apport du deep-learning à la biologie des systèmes



Les prémices de la biologie des systèmes

Réseau métabolique



Albert Szent-Györgyi-1930
Hans Adolf Krebs - 1937

Réseau génétique

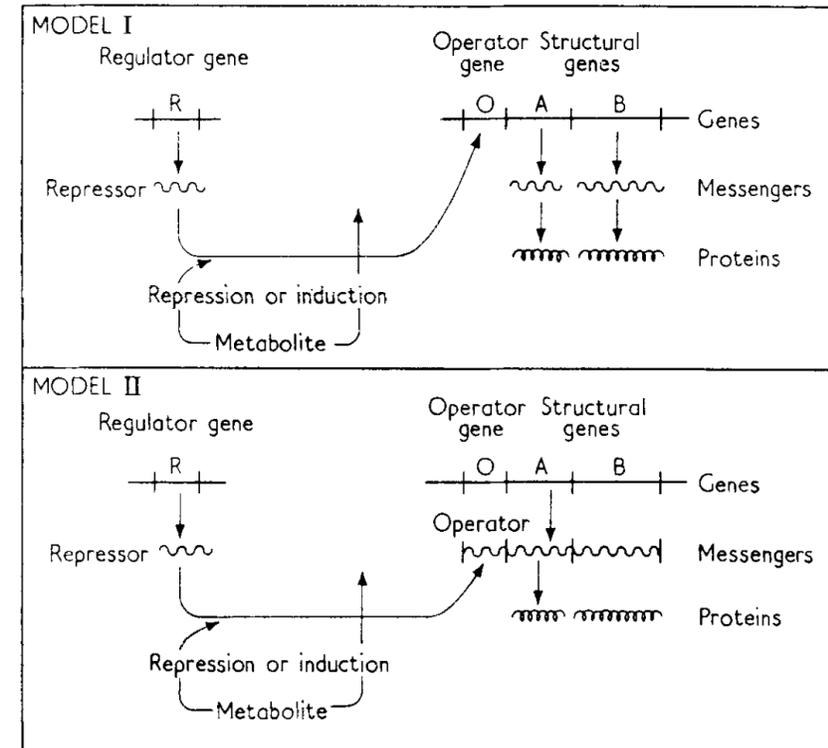


FIG. 6. Models of the regulation of protein synthesis.

Genetic Regulatory Mechanisms in the Synthesis of Proteins
F. Jacob, J. Monod, J. Mol. Biol. 1961

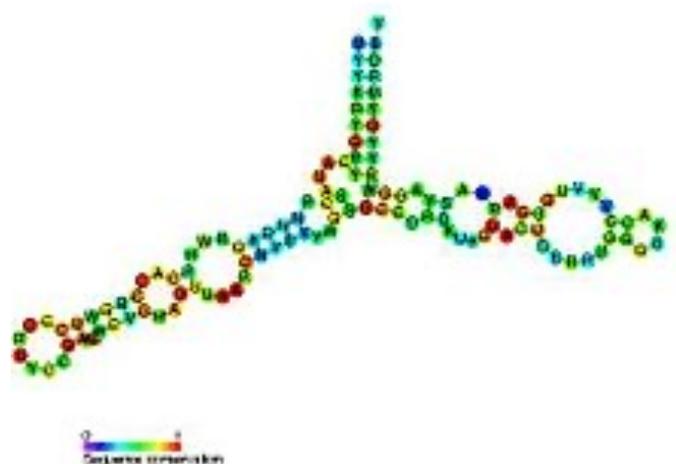
Décoder l'information biologique

Protéine

A chain	B chain	
Gly	Phe	1
Ile	Val	
Val	Asn	
Glu	Gln	
Gln	His	5
Cys	Leu	
Cys	Cys	
Ala	Gly	
Ser	Ser	
Val	His	10
Cys	Leu	
Ser	Val	
Leu	Glu	
Tyr	Ala	
Gln	Leu	15
Leu	Tyr	
Glu	Leu	
Asn	Val	
Tyr	Cys	
Cys	Gly	20
Asn	Glu	
	Arg	
	Gly	
	Phe	
	Phe	25
	Tyr	
	Thr	
	Pro	
	Lys	
	Ala	30

Insuline de bovin
1955

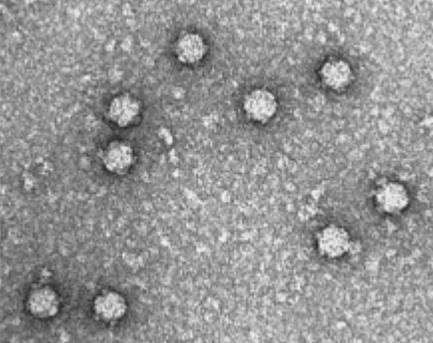
ARN



5S rRNA
120 nucleotides
1965

Frederick Sanger (1918 – 2013)

ADN



E. Virus ΦX174
5386 nucléotides
1977

“Sanger sequencing method”
Mitochondrie 16569 bp
Bacteriophage λ 48502 bp
1977

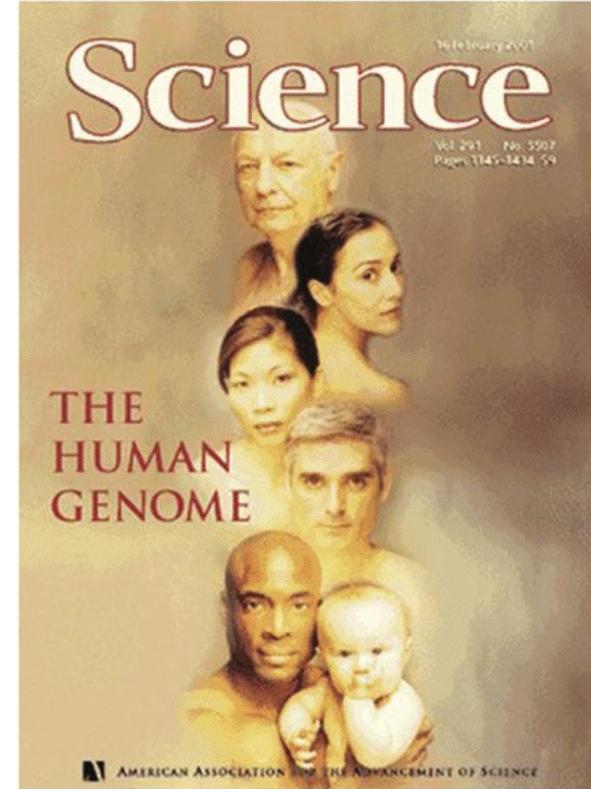
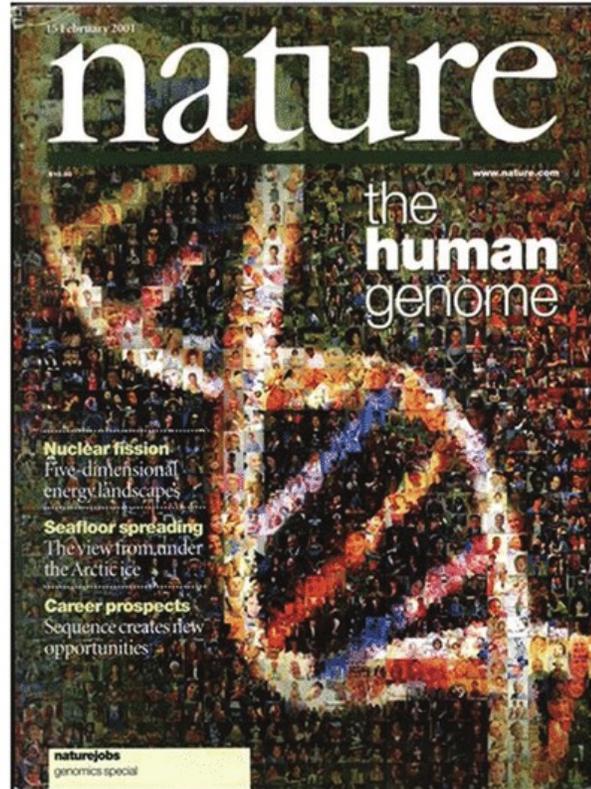
Le séquençage du génome humain

Human Genome Project

Lancé en 1985
par le gouvernement
américain

1989-2001

5 milliards de dollars
20 Instituts
6 pays



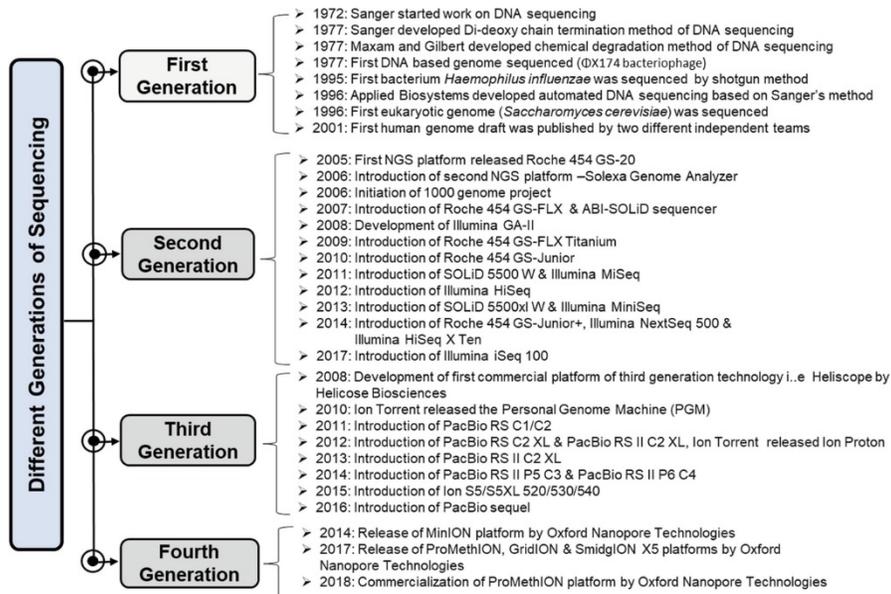
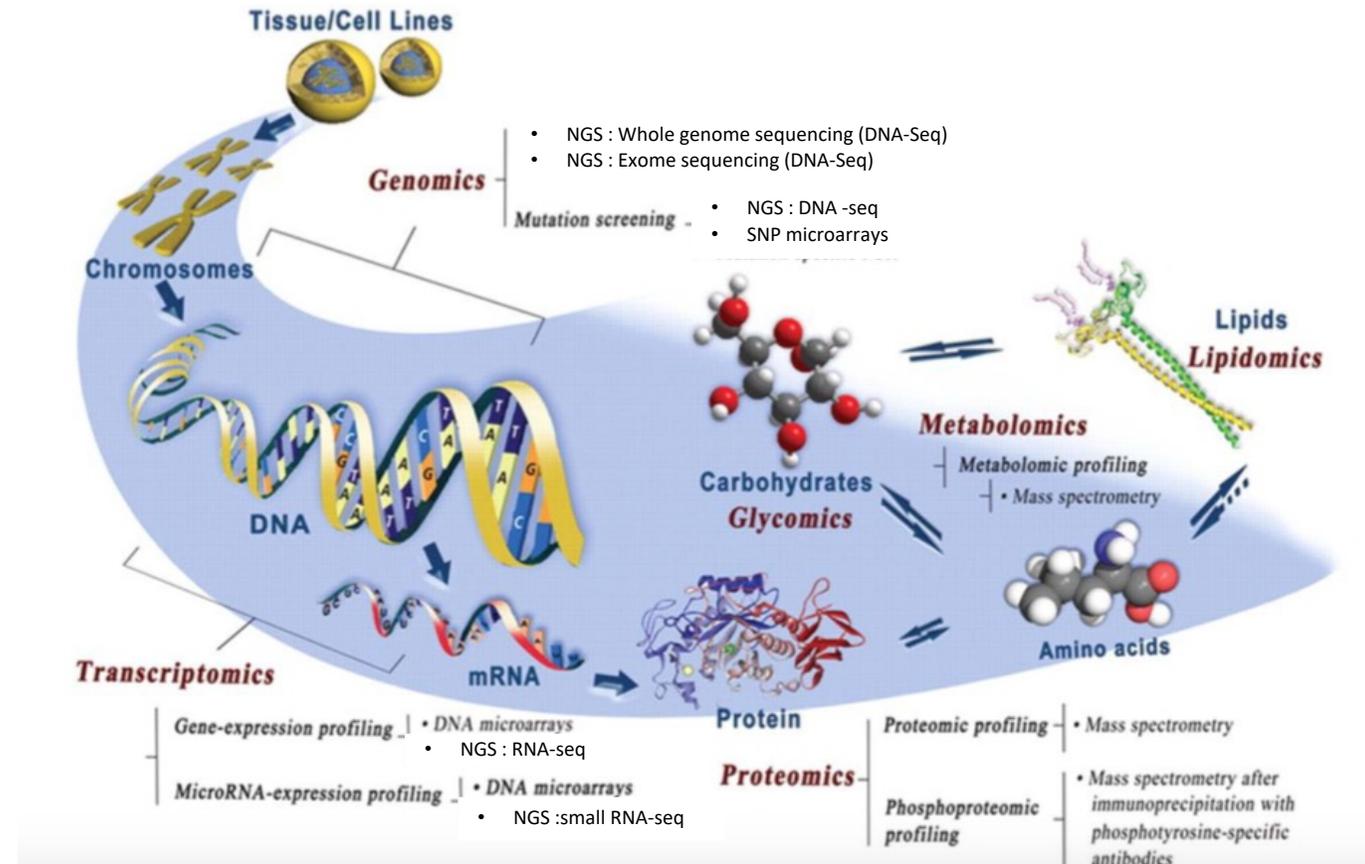
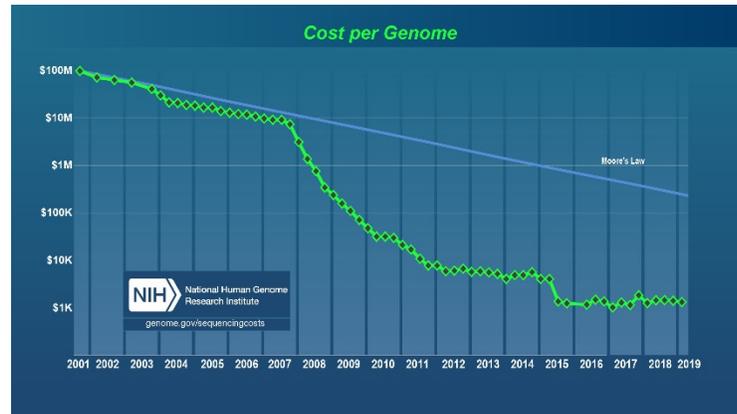
Celera Genomics
(J. Craig Venter)

1998-2001

300 millions dollars
(Shotgun sequencing)

Février 2001

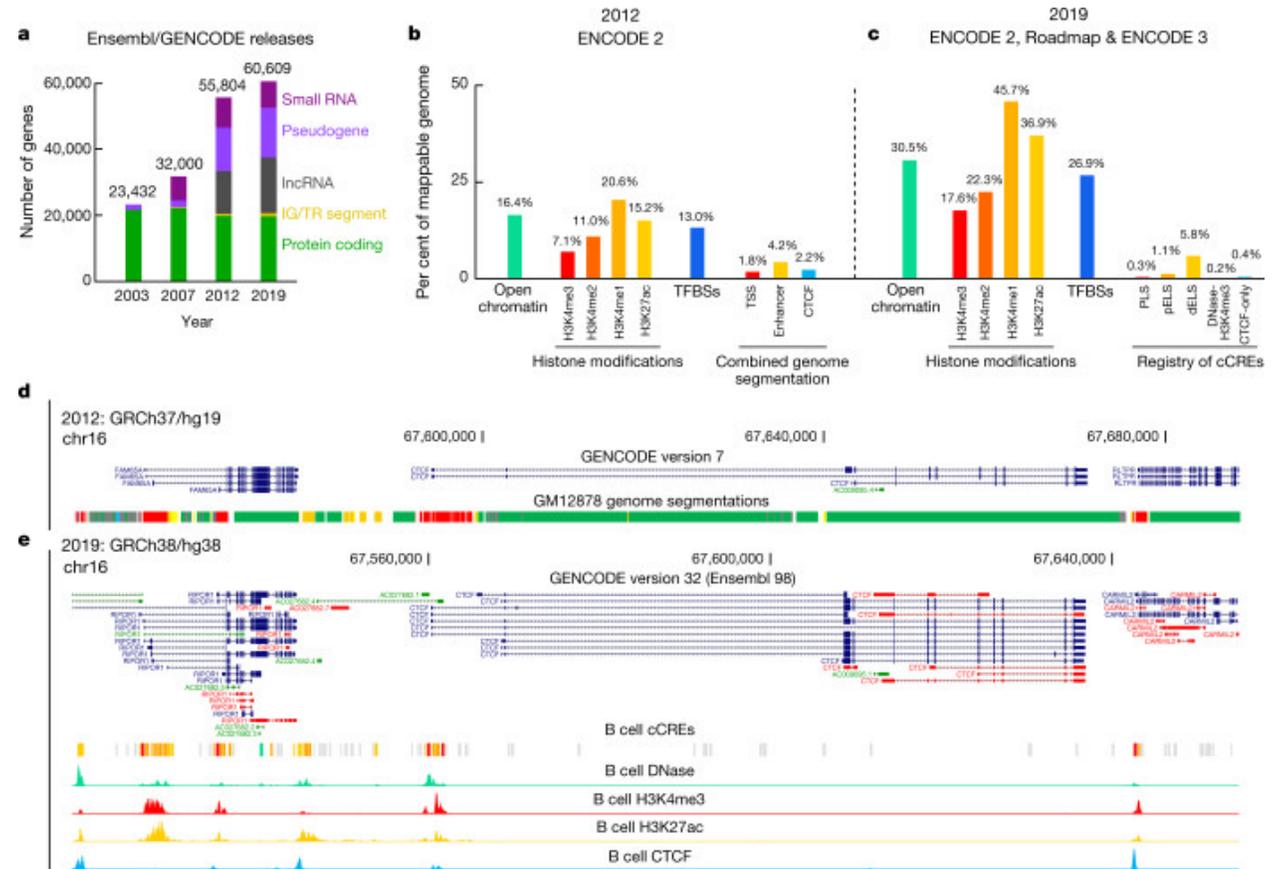
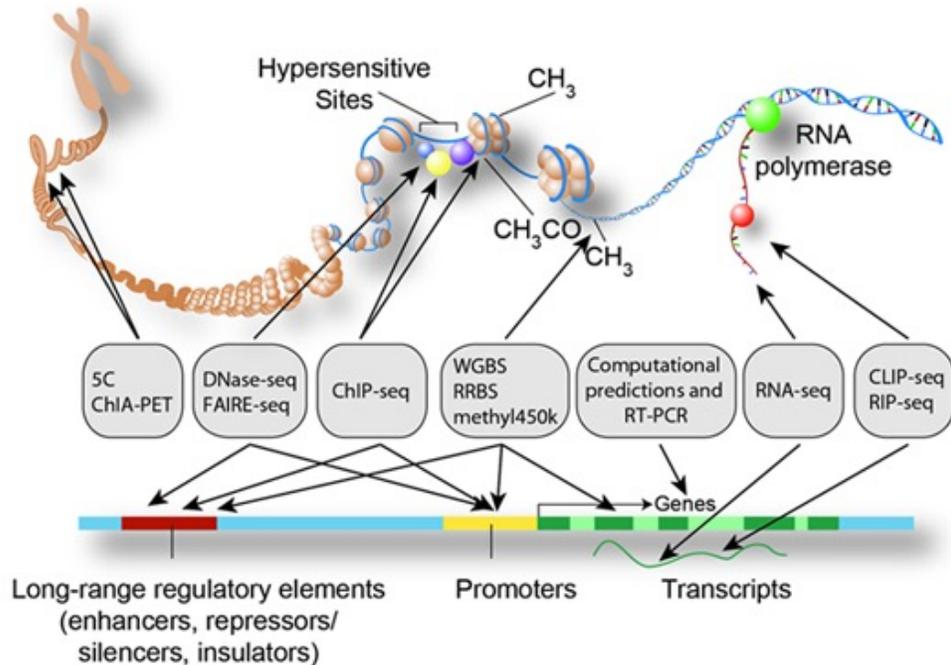
La révolution omique



Wu R.Q., J. dent. Research, 2010

Projet ENCODE

Encyclopedia of DNA Element 2003 - Present



Réseau de régulation transcriptionnel chez *Escherichia coli* (en 1998)

Table 1. Information contained in RegulonDB release 1.0

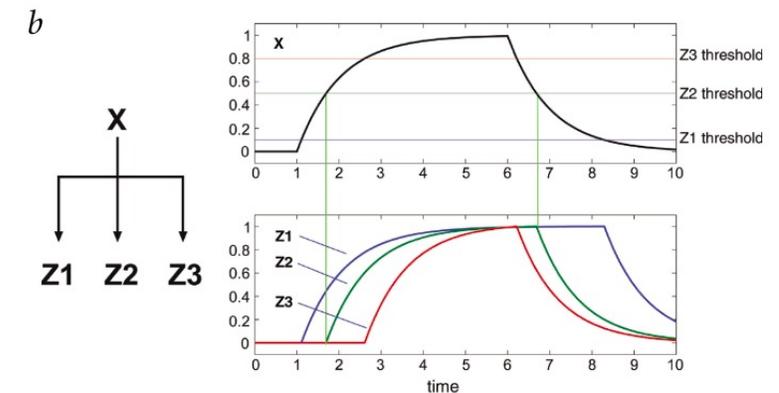
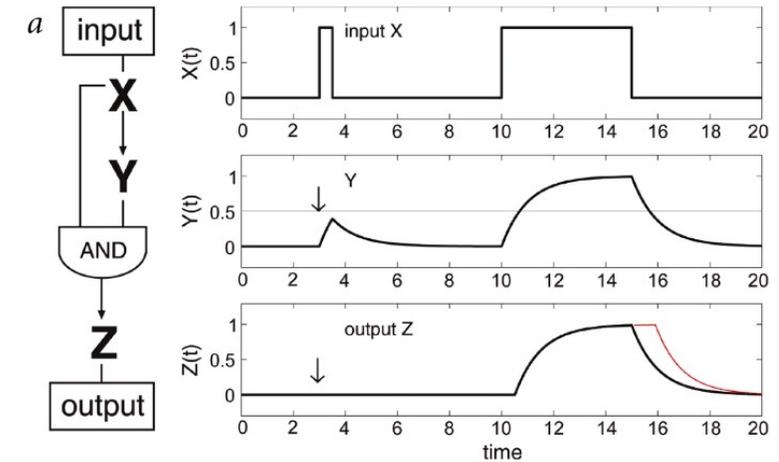
RegulonDB: a database on transcriptional regulation in *Escherichia coli*.

A. M. Huerta et al., N.A.R. 1998

Object	Number
Regulons	99
Regulatory interactions	533
Polypeptides	192
Protein complexes	99
Genes	542
Operons	292
Promoters	300
Ext_DB_References ^a	2050
Authors	298
Signals	35

4 types de source pour les interactions :

- (i) mutational experiments proving the protein–DNA binding site interaction
- (ii) specific binding of the purified protein
- (iii) evidence of binding with non-purified protein
- (iv) simple sequence similarity with other sites for a regulatory protein.



Network motifs in the transcriptional regulation network of *Escherichia coli*,
Shen-Orr et al., Nature 2002

Le réseau d'interaction protéine-protéine chez la levure (en 2000)



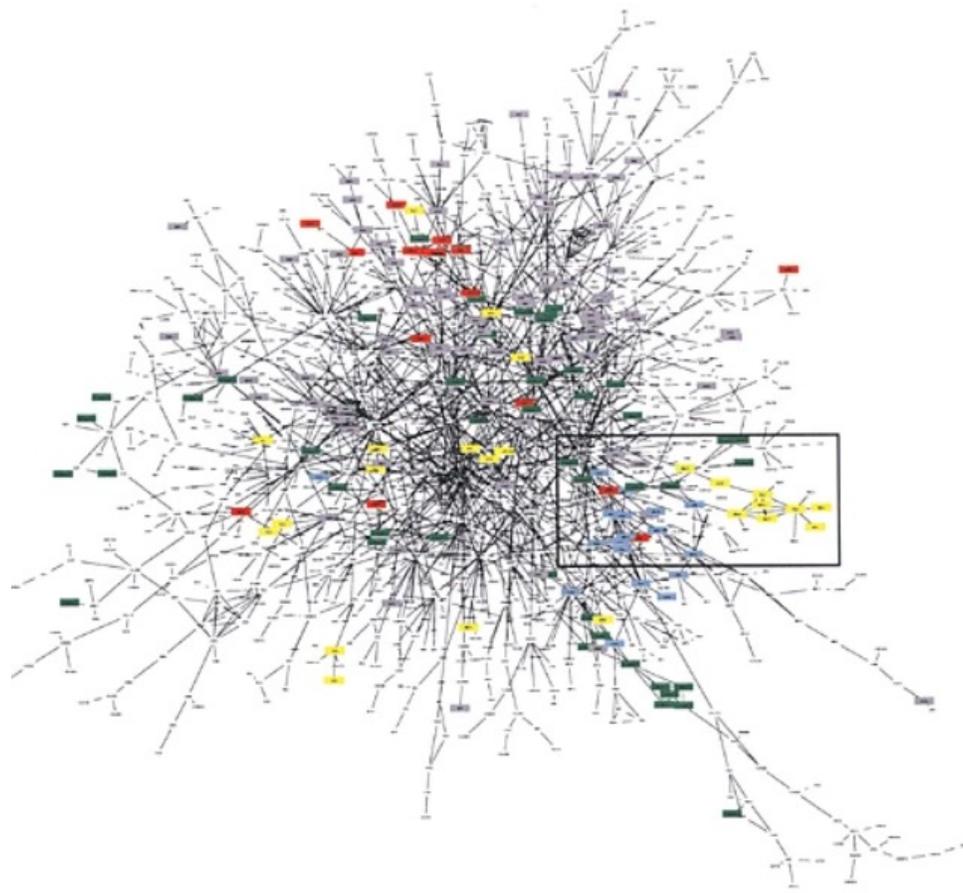
Réseau incluant 1548 protéines (sur 6000 prédites)

Les 2538 interactions du réseau ont été reconstruites en combinant :

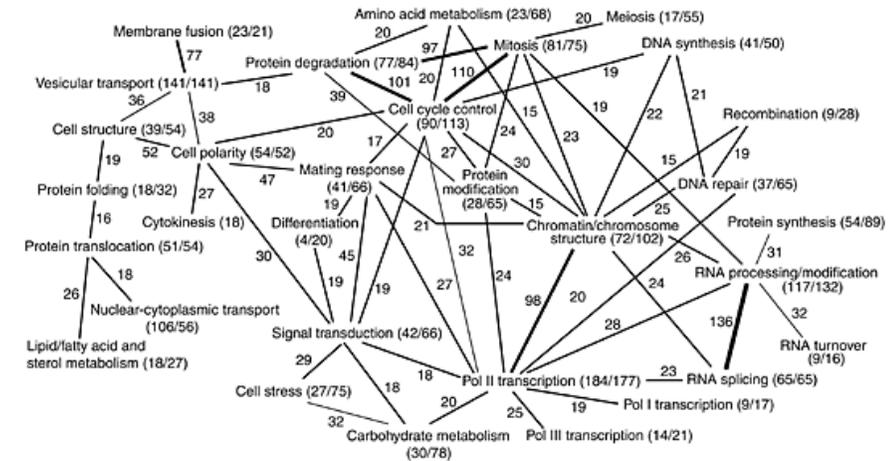
- Interaction mesurée expérimentalement (Gel2D, Spectrométrie de masse)
- Yeast two hybrid

A network of protein-protein interactions in yeast
Schwikowski, Nature Biotechnology, 2000

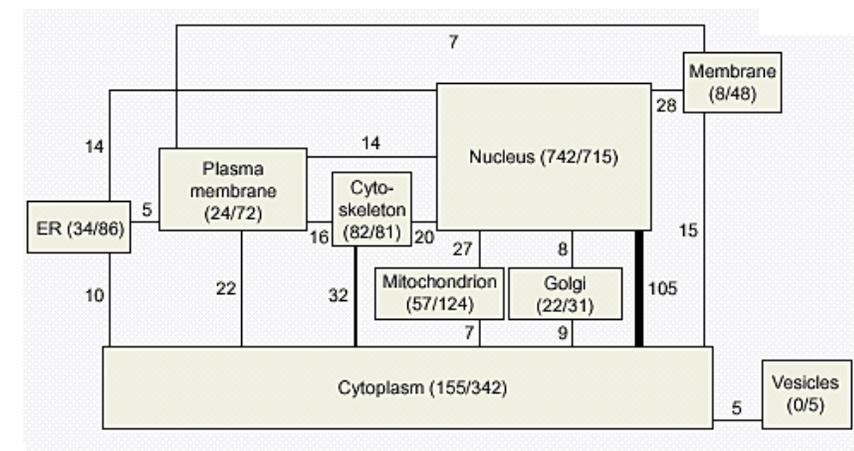
Le réseau d'interaction protéine-protéine chez la levure (en 2000)



A network of protein-protein interactions in yeast
Schwikowski, Nature Biotechnology, 2000

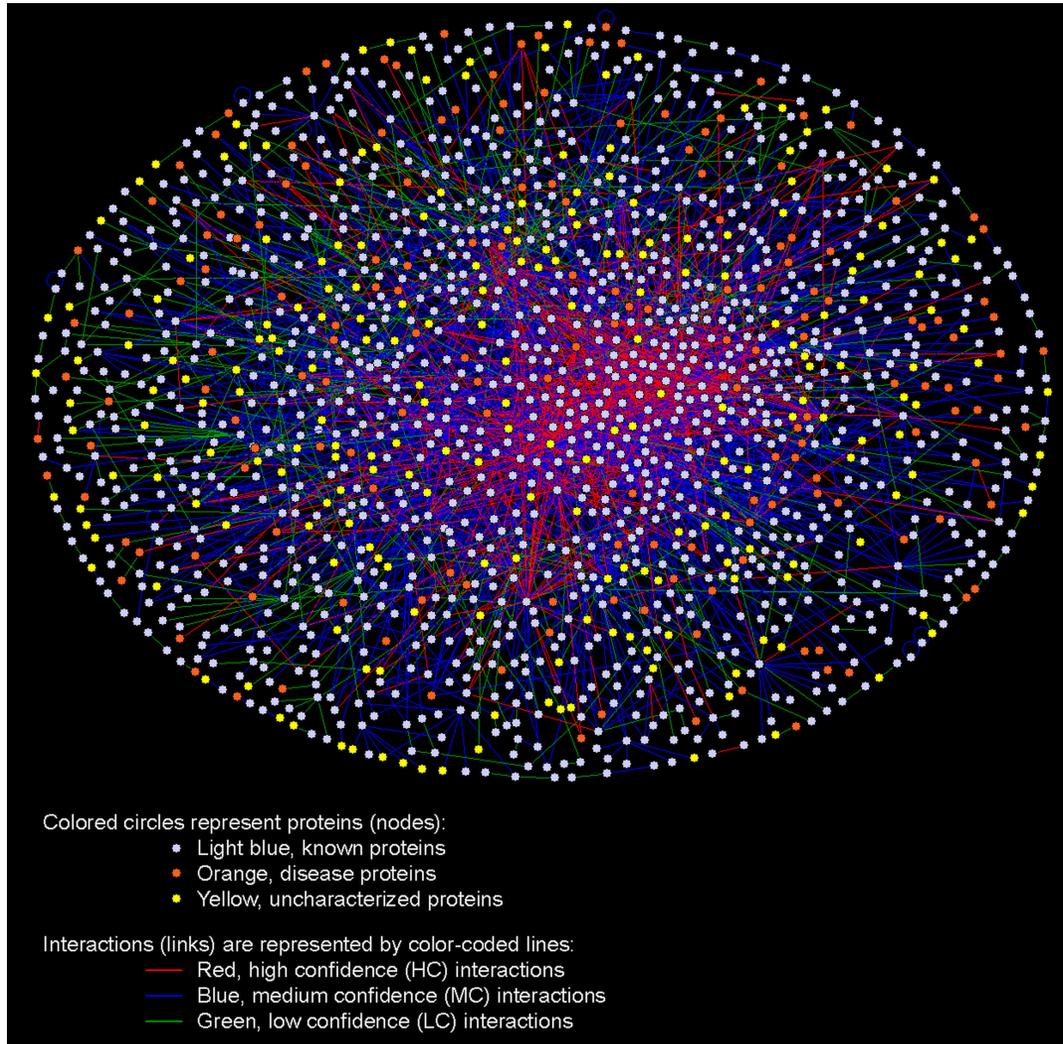


Classification par groupe fonctionnel



Classification par compartiment cellulaire

Un réseau « simple » d'interaction protéine-protéine chez l'humain



1705 protéines (sur 43000 prédites)

3186 interactions

25 millions de yeast two-hybrid

+ Interaction évaluées à l'aide de
Gene Ontology

A human protein-protein interaction network:
A resource for annotating the proteome
Stelzl et al., Cell, 2005

Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

- **Reconstruction de réseaux biologiques à l'aide de données omiques**
- Reconstruire par analyse de co-expression (TP)
- Etat de l'art des réseaux biologiques les plus étendus
- Apport du deep-learning à la biologie des systèmes



Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

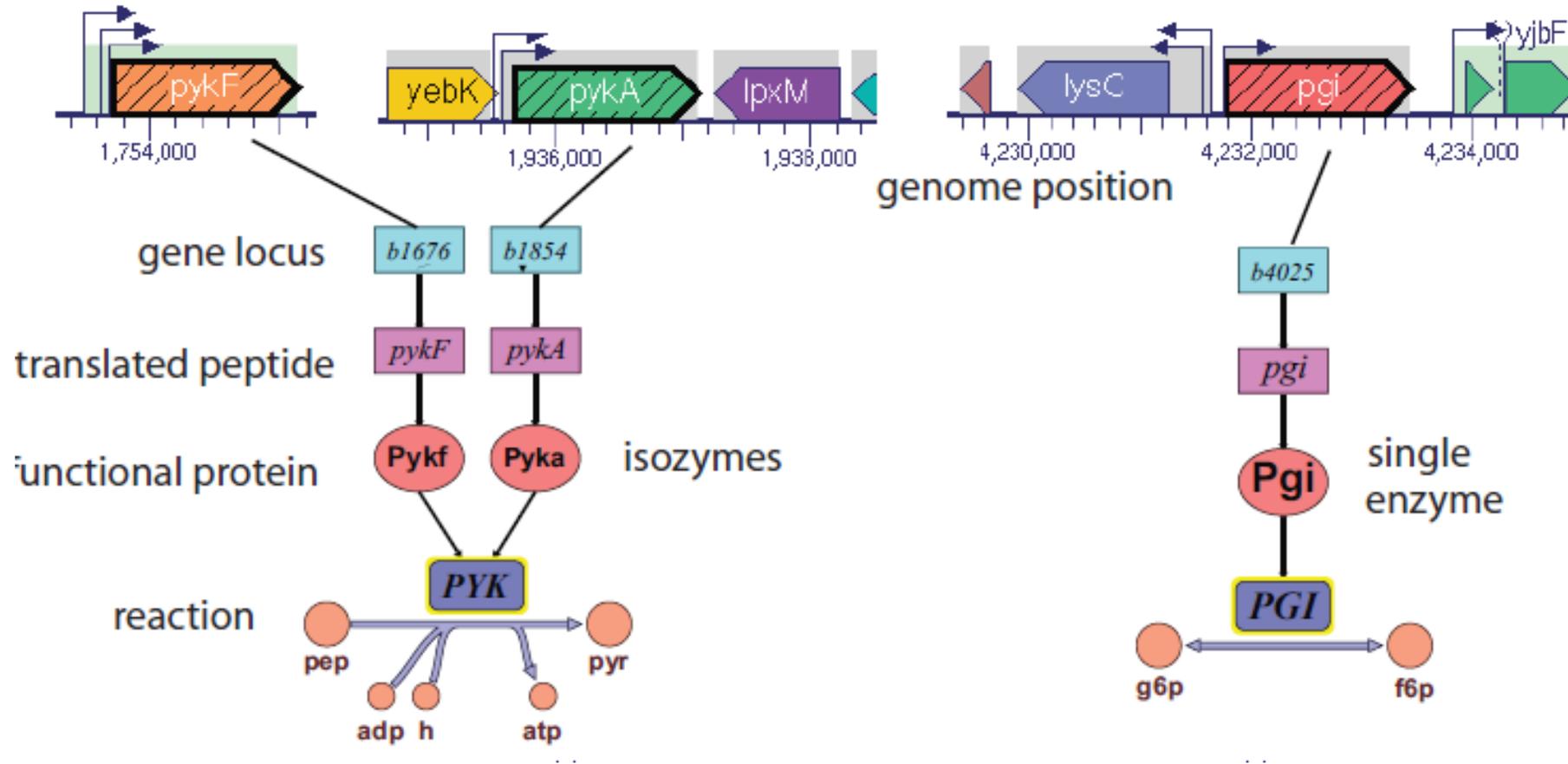
- **Reconstruction de réseaux biologiques à l'aide de données omiques**

1. Quels sont les éléments à reconstruire ?
2. Quel type de réseaux peut-on reconstruire ?
3. Pour quel type de modélisation ?
4. Les différents types de reconstruction
5. Reconstruction directe
6. Les données omiques
7. Reconstruction indirecte
8. Les limites de la reconstruction de réseaux biologiques



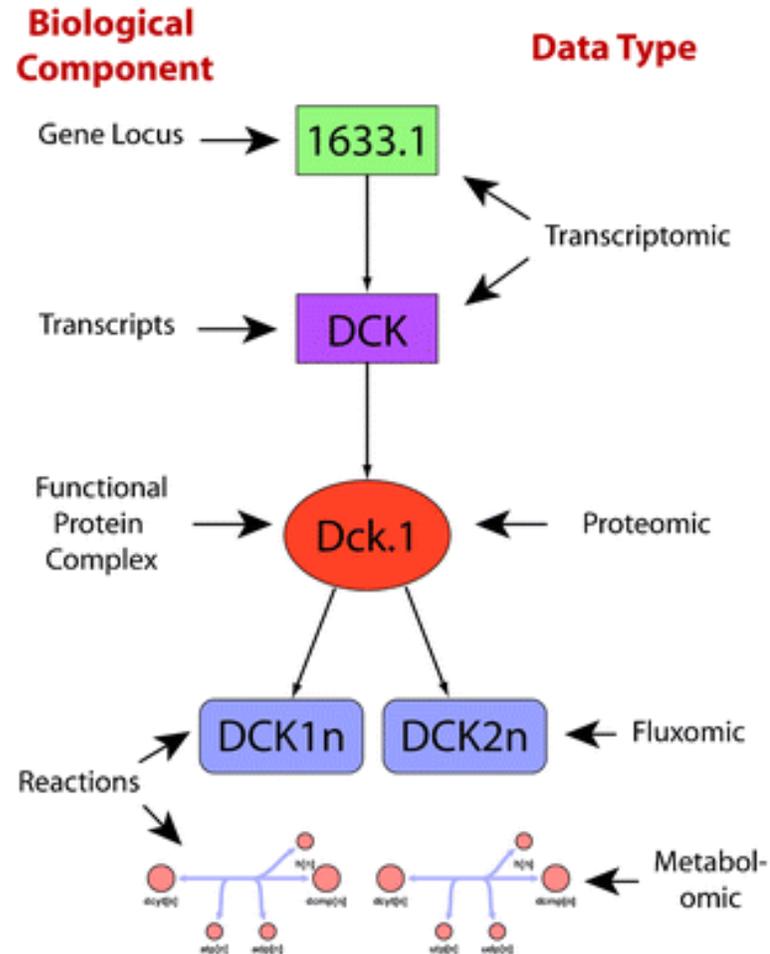
Quels sont les éléments à reconstruire ?

Reconstruction de réseau métabolique



Quels sont les éléments à reconstruire ?

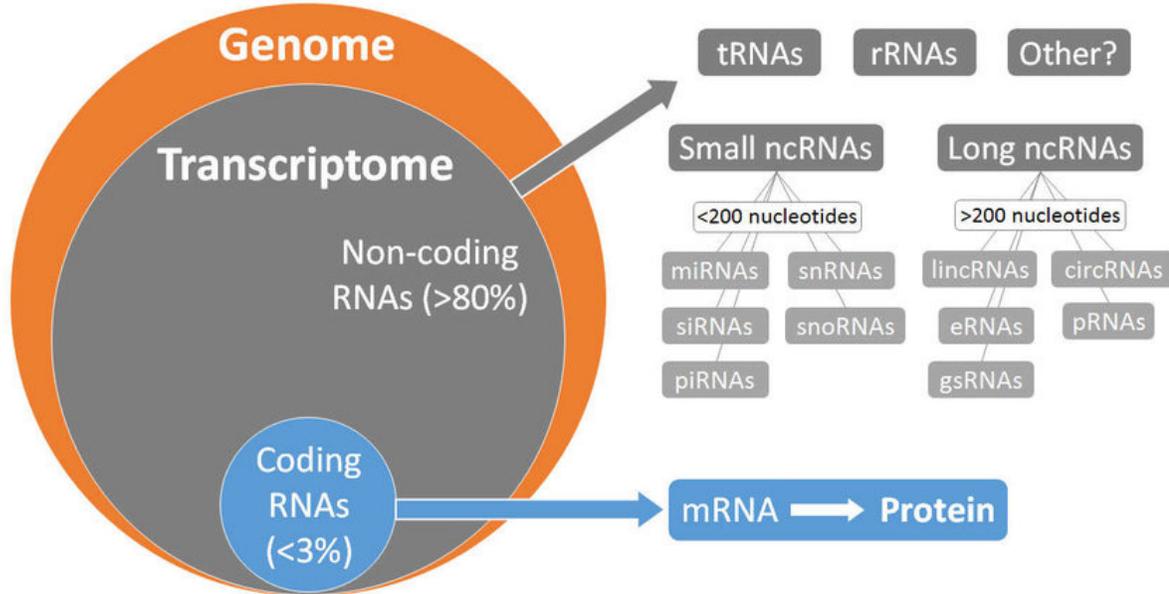
Les éléments fondamentaux du réseau



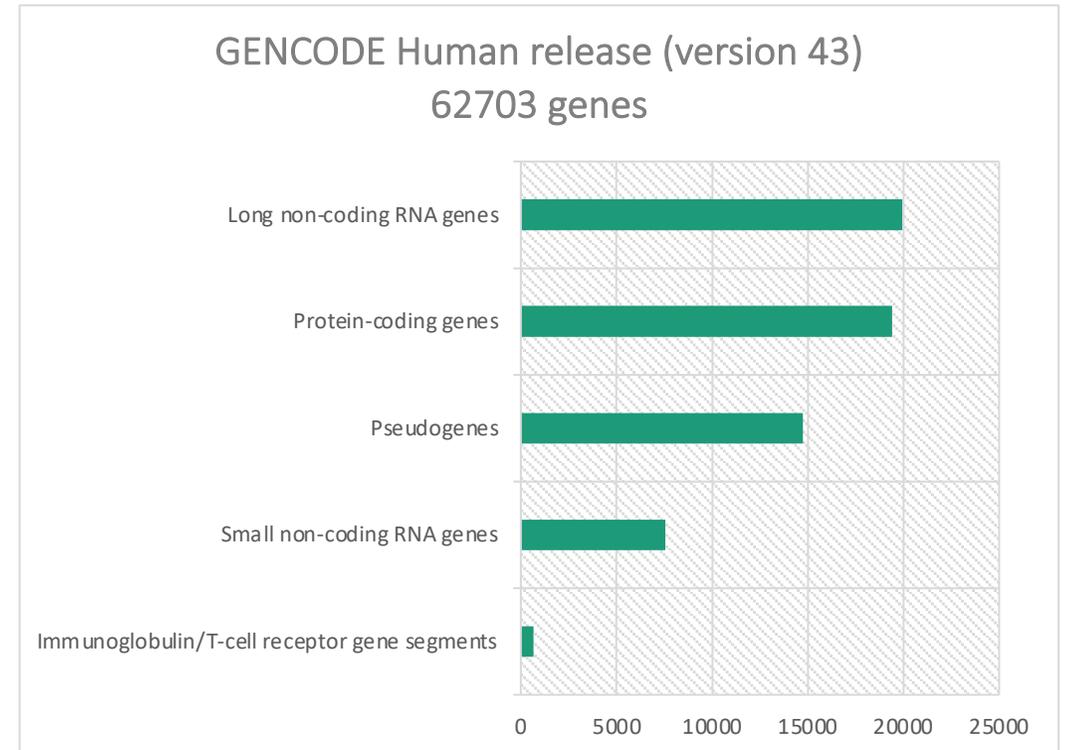
- Les gènes
- Les transcrits
- Les protéines
- Les complexes de protéines
- Les métabolites
- Les réactions enzymatiques

Quels sont les éléments à reconstruire ?

La notion de gène



Pedrosa et al., The CardioRNA COST Action 2019



GENCODE 21, Frankish et al., N.A.R. 2021

Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

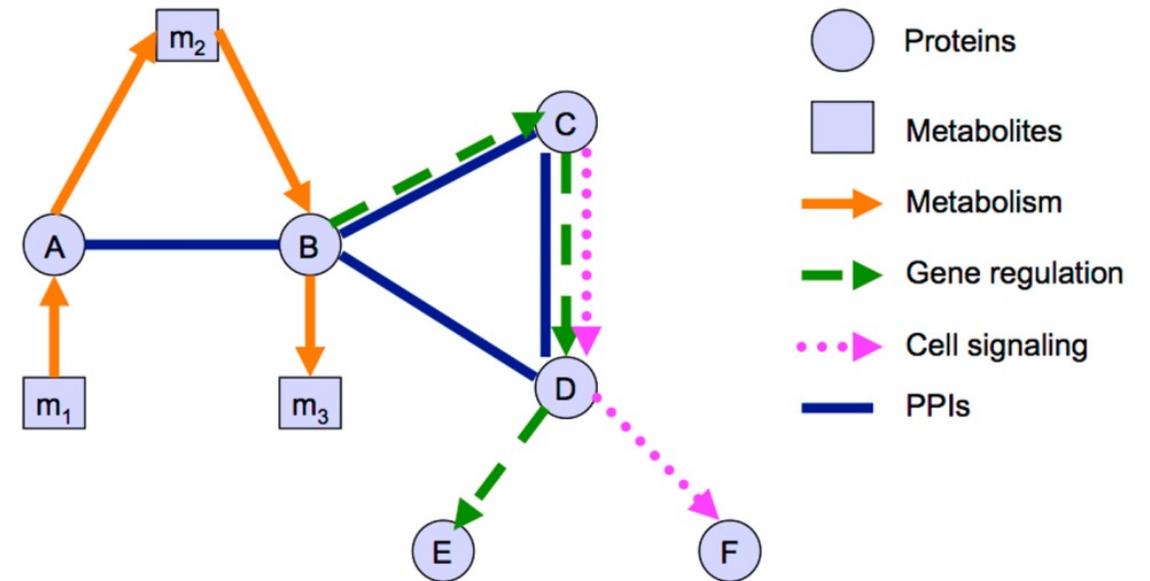
- **Reconstruction de réseaux biologiques à l'aide de données omiques**

1. Quels sont les éléments à reconstruire ?
2. **Quel type de réseaux peut-on reconstruire ?**
3. Pour quel type de modélisation ?
4. Les différents types de reconstruction
5. Reconstruction directe
6. Les données omiques
7. Reconstruction indirecte
8. Les limites de la reconstruction de réseaux biologiques

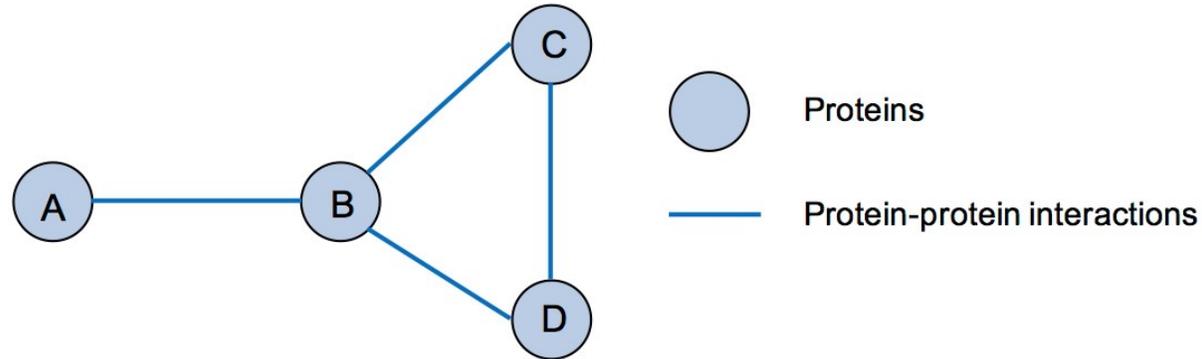


Quels types de réseaux peut-on reconstruire ?

- Réseau d'interaction protéine-protéine (PPI)
- Réseau métabolique
- Réseau de régulation génétique
- Réseau de signalisation cellulaire

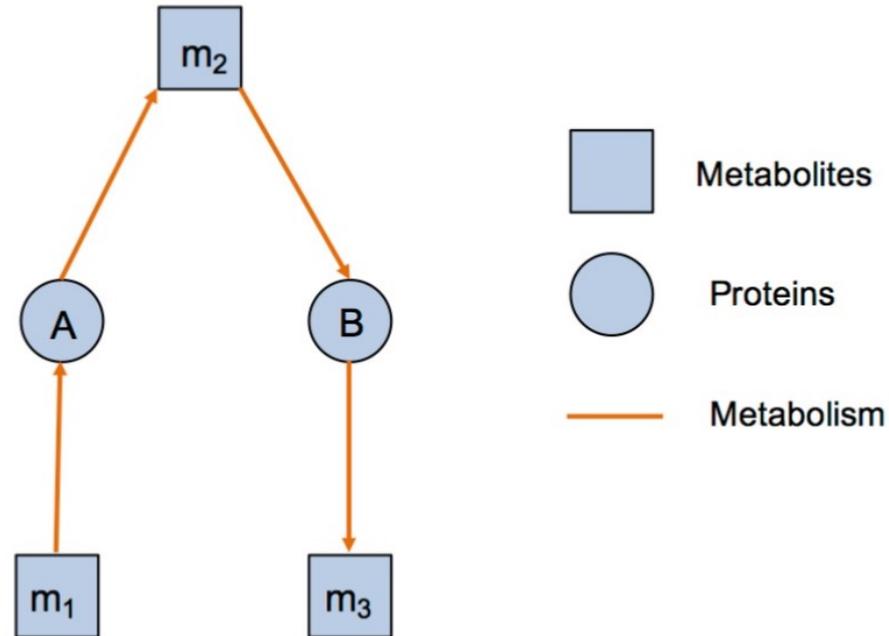


Réseau d'interaction protéine-protéine



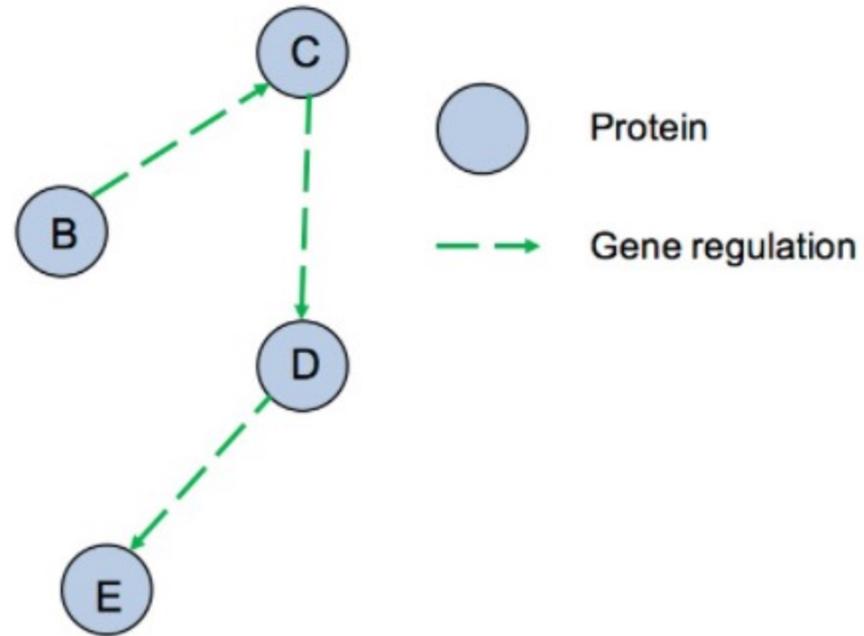
- Chaque nœud est une protéine
- Chaque arête représente un lien d'interaction physique entre protéines
- Le graphe est non orienté
(Mais peut le devenir si on considère les sites de docking)

Réseau métabolique



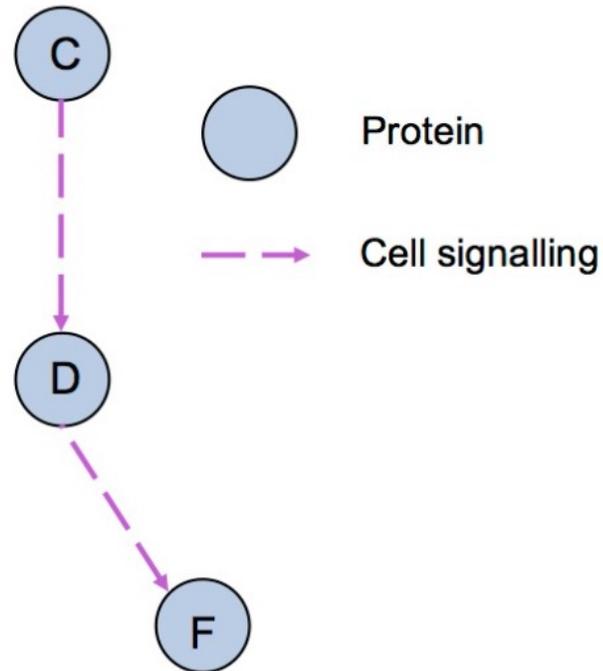
- Les nœuds sont les métabolites et les enzymes
- Les arêtes représentent les réactions métaboliques et leurs flux
- Le graphe est orienté

Réseau de régulation génétique



- Les nœuds sont les gènes et les facteurs de transcription
- Les arêtes représentent les régulations transcriptionnelles
- Le graphe est orienté

Réseau de signalisation cellulaire



- Les nœuds sont les protéines, mais aussi les gènes et métabolites
- Les arêtes orientées représentent les voies de signalisation cellulaire
- Les autres réseaux peuvent être vu comme des sous-graphes de celui-ci

Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

- **Reconstruction de réseaux biologiques à l'aide de données omiques**

1. Quels sont les éléments à reconstruire ?
2. Quel type de réseaux peut-on reconstruire ?
3. **Pour quel type de modélisation ?**
4. Les différents types de reconstruction
5. Reconstruction directe
6. Les données omiques
 1. Reconstruction indirecte
 2. Les limites de la reconstruction de réseaux biologiques

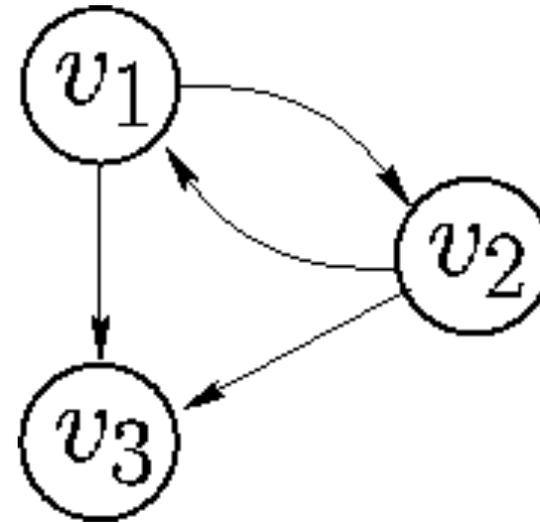


Pour quel type de modélisation va-t-on reconstruire le réseau ?

- Modélisation de réseau booléen
- Modélisation par règles
- Modélisation par contraintes
- Modélisation par équations différentielles

Modélisation de réseau booléen

- Un graphe orienté non pondéré
- Les interactions sont binaires :
Présentes – absentes
- Parcours en temps discret du réseau



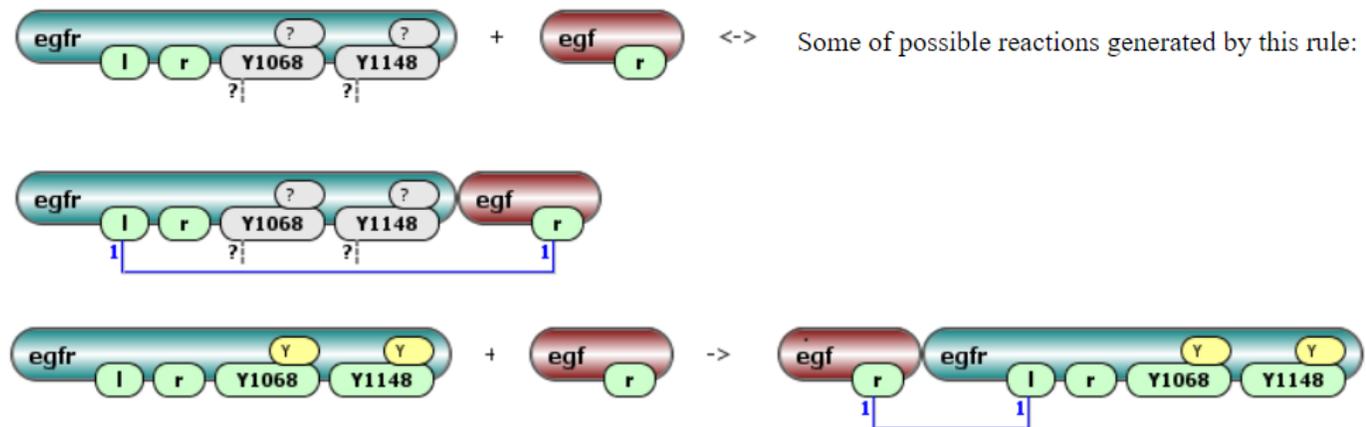
v_2	v_1
0	1
1	0

v_1	v_2
0	0
1	1

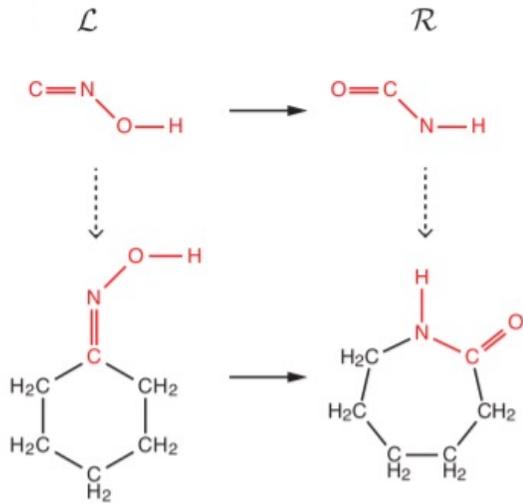
v_1	v_2	v_3
0	0	0
0	1	1
1	0	0
1	1	0

Modélisation par règle

- Chaque nœud a plusieurs états accessibles
- Les interactions sont des règles logiques entre les états des nœuds

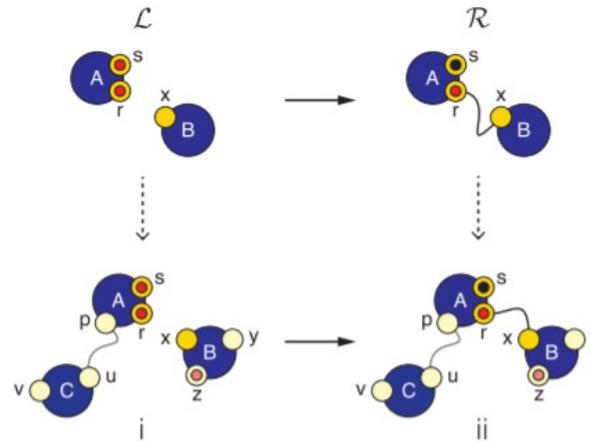


Modélisation par règle

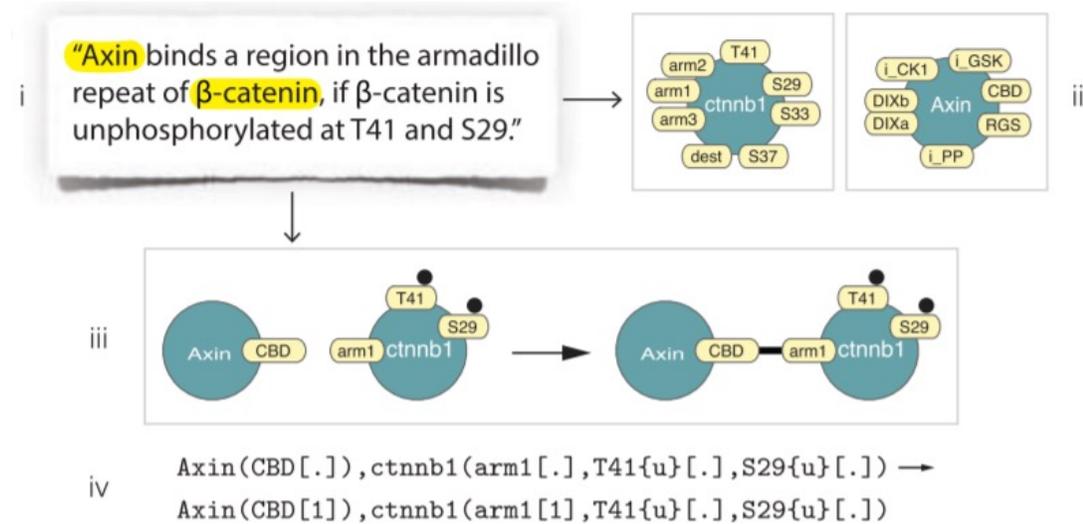


Equation chimique

Définition d'une règle



Boutillier et al., Bioinformatics 2018



Ecriture en langage kappa

On définit une probabilité pour chaque règle

On modélise en calculant « l'ensemble des possibles » pour une initialisation donnée
Cad: L'ensemble des états possibles ayant une probabilité d'accessibilité donnée

Modélisation sous contrainte

Il est trop couteux de modéliser l'évolution de tous les constituants du système



On impose une contrainte sur le système. Celle-ci va réduire la quantité d'état accessibles



La modélisation deviendra alors possible en un temps raisonnable

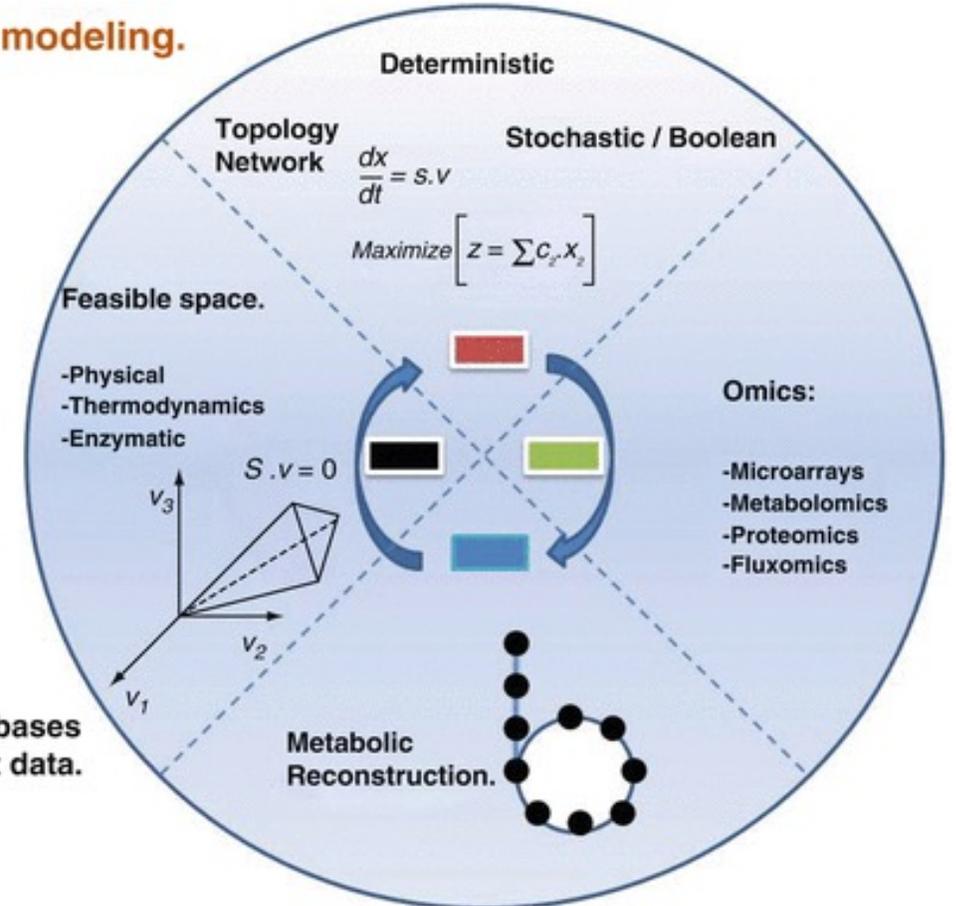
Constraint-based modeling.

 Mathematical representation

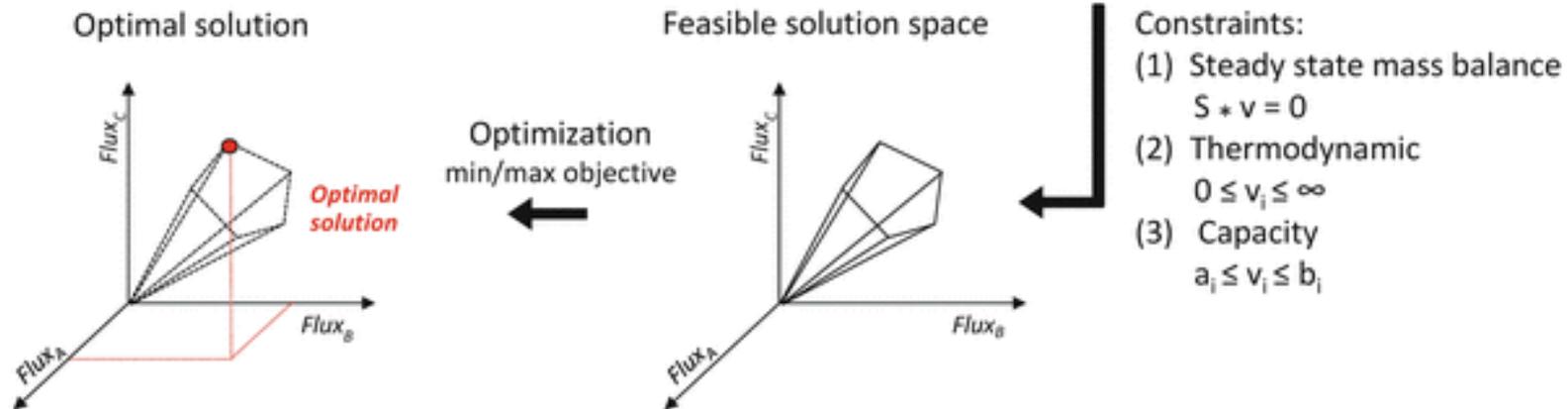
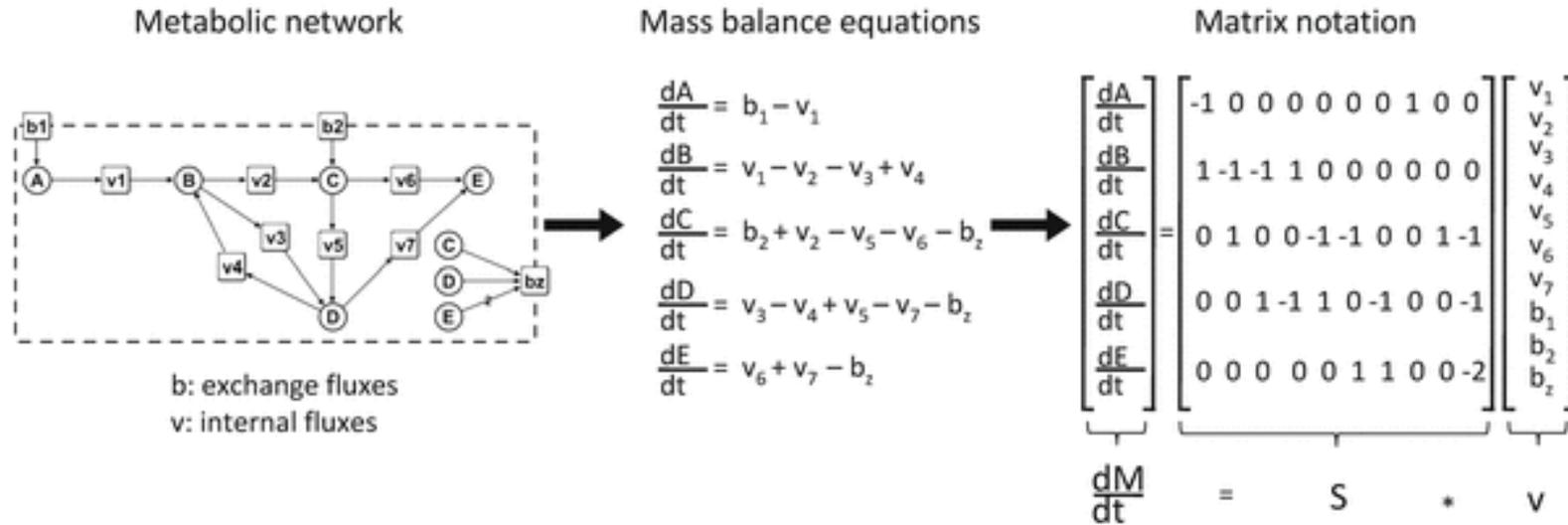
 In silico Modeling.

 Experimental assessment.

 Integration databases high-throughput data.



Modélisation sous contrainte : Flux balance analysis



Modélisation par contrainte = la contrainte nous permet de réduire le champ des possibles

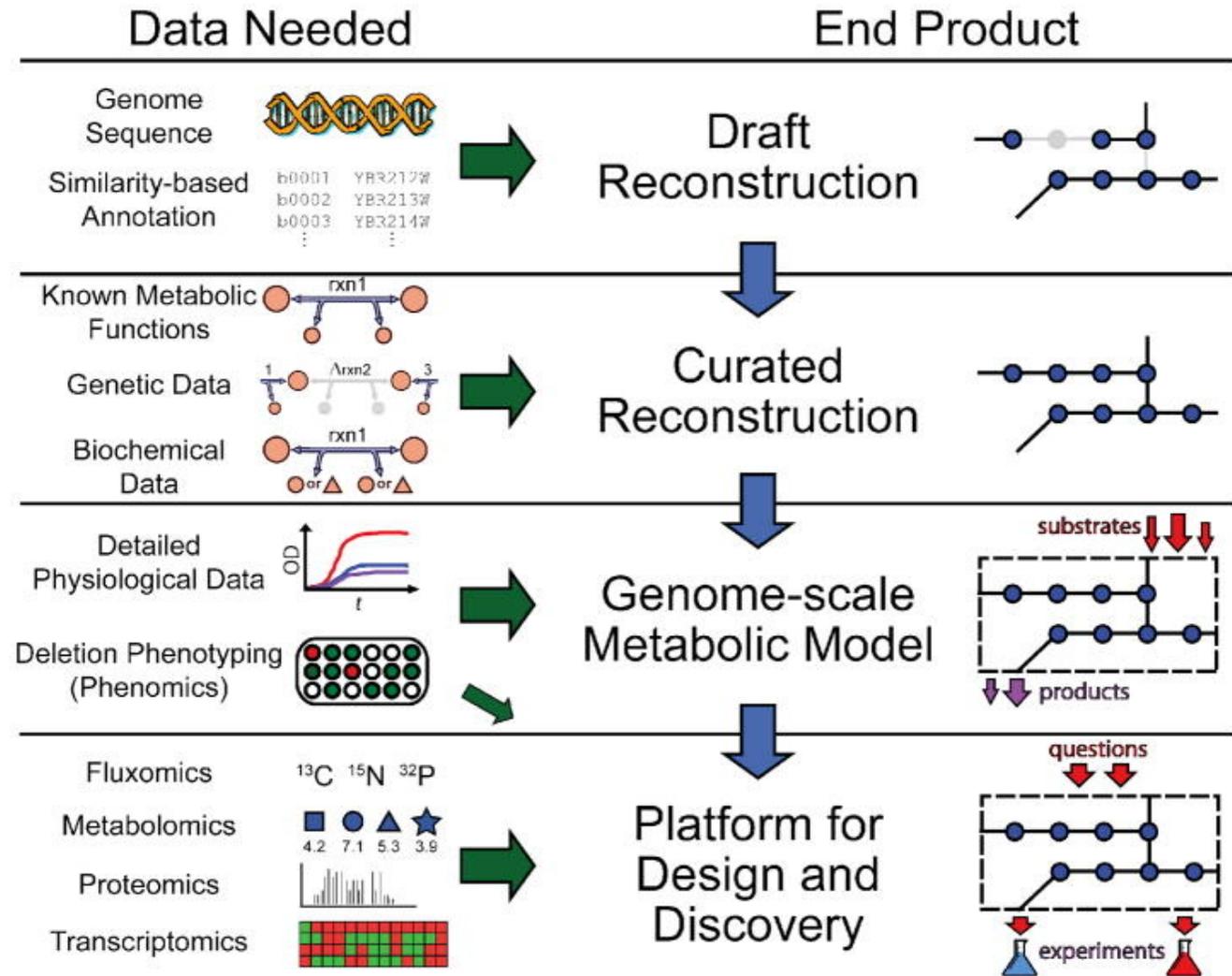
Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

- **Reconstruction de réseaux biologiques à l'aide de données omiques**

1. Quels sont les éléments à reconstruire ?
2. Quel type de réseaux peut-on reconstruire ?
3. Pour quel type de modélisation ?
4. **Les différents types de reconstruction**
5. Reconstruction directe
6. Les données omiques
7. Reconstruction indirecte
8. Les limites de la reconstruction de réseaux biologiques

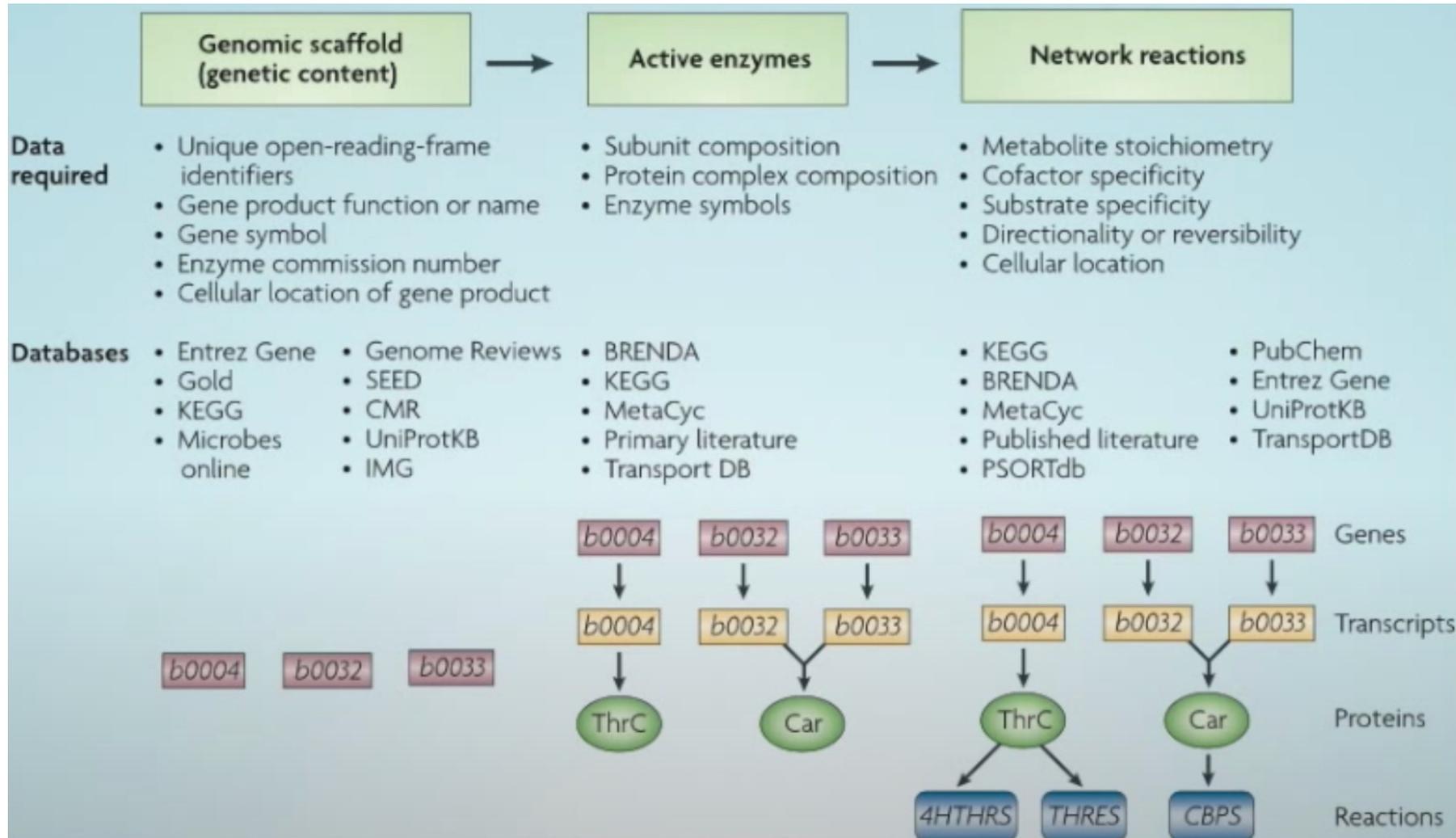


Reconstruction du réseau métabolique d'un organisme



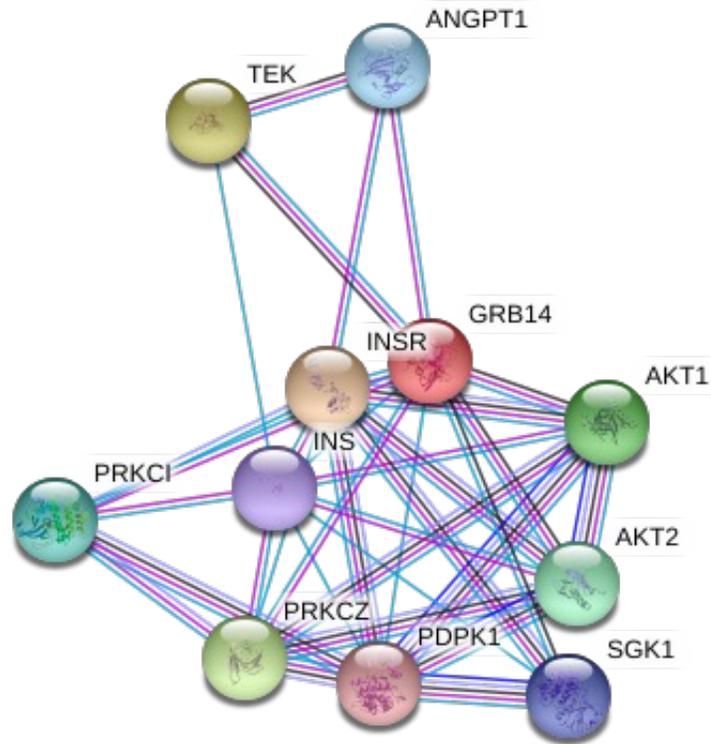
Reconstruction of biochemical Networks in Microbial Organisms
 Palsson et al., *Nature Review Microbiology*, 2009

Reconstruction du réseau métabolique d'un organisme



Reconstruction of biochemical Networks in Microbial Organisms
 Palsson et al., *Nature Review Microbiology*, 2009

Les différents types de reconstruction



Known Interactions

- from curated databases*
- experimentally determined*

Predicted Interactions

- gene neighborhood*
- gene fusions*
- gene co-occurrence*

Others

- textmining*
- co-expression*
- protein homology*

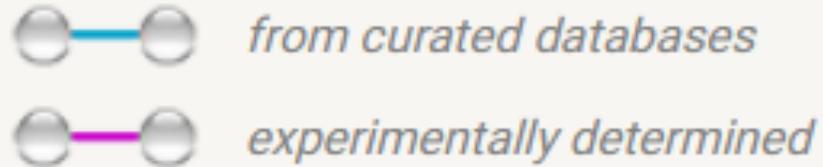
Réseau d'interaction protéine-protéine
de l'insuline prédite par string-db.org

Les différentes types de reconstruction

L'exemple d'interaction protéines-protéines

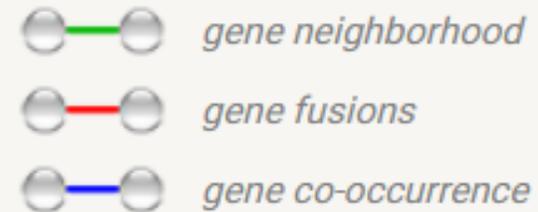
Reconstruction directe

Known Interactions

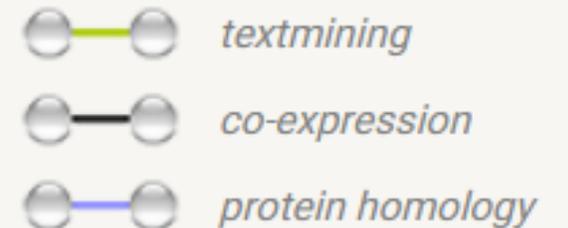


Reconstruction indirecte

Predicted Interactions



Others



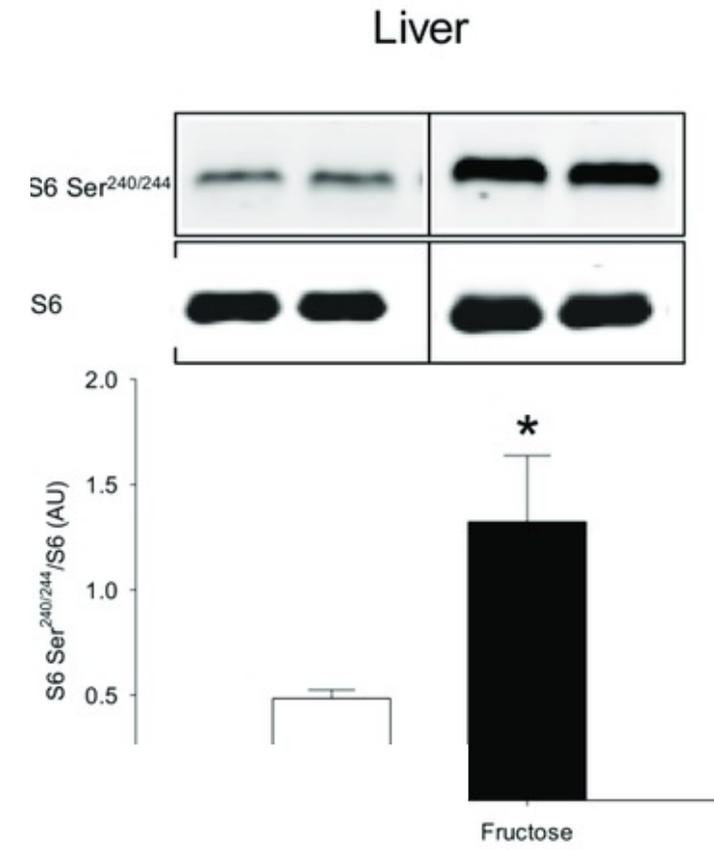
Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

- **Reconstruction de réseaux biologiques à l'aide de données omiques**

1. Quels sont les éléments à reconstruire ?
2. Quel type de réseaux peut-on reconstruire ?
3. Pour quel type de modélisation ?
4. Les différents types de reconstruction
5. **Reconstruction directe**
6. Les données omiques
7. Reconstruction indirecte
8. Les limites de la reconstruction de réseaux biologiques



Reconstruction directe des réseaux : Les Blots



Reconstruction directe des réseaux métaboliques

Top-level EC numbers^[5]

Class	Reaction catalyzed	Typical reaction	Enzyme example(s) with trivial name
EC 1 <i>Oxidoreductases</i>	To catalyze oxidation/reduction reactions; transfer of H and O atoms or electrons from one substance to another	$AH + B \rightarrow A + BH$ (reduced) $A + O \rightarrow AO$ (oxidized)	Dehydrogenase, oxidase
EC 2 <i>Transferases</i>	Transfer of a functional group from one substance to another. The group may be methyl-, acyl-, amino- or phosphate group	$AB + C \rightarrow A + BC$	Transaminase, kinase
EC 3 <i>Hydrolases</i>	Formation of two products from a substrate by hydrolysis	$AB + H_2O \rightarrow AOH + BH$	Lipase, amylase, peptidase, phosphatase
EC 4 <i>Lyases</i>	Non-hydrolytic addition or removal of groups from substrates. C-C, C-N, C-O or C-S bonds may be cleaved	$RCOCOOH \rightarrow RCOH + CO_2$ or $[X-A+B-Y] \rightarrow [A=B + X-Y]$	Decarboxylase
EC 5 <i>Isomerases</i>	Intramolecule rearrangement, i.e. isomerization changes within a single molecule	$ABC \rightarrow BCA$	Isomerase, mutase
EC 6 <i>Ligases</i>	Join together two molecules by synthesis of new C-O, C-S, C-N or C-C bonds with simultaneous breakdown of ATP	$X + Y + ATP \rightarrow XY + ADP + P_i$	Synthetase

Les 6 classes d'enzymes

Enzyme Commission number for enzymes

https://en.wikipedia.org/wiki/List_of_enzymes

Reconstruction directe des réseaux métaboliques

Les 6 classes d'enzymes

- EC 1 Oxidoreductases
- EC 2 **Transferases**
- EC 3 Hydrolases
- EC 4 Lyases
- EC 5 Isomerases
- EC 6 Ligases

ENZYME entry: EC 2.7.1.137
Phosphatidylinositol 3-kinase.
1-phosphatidylinositol 3-kinase.
PI3-kinase.

- EC 2.1 - Transferring one-carbon groups
- EC 2.2 - Transferring aldehyde or ketonic groups
- EC 2.3 - Acyltransferases
- EC 2.4 - Glycosyltransferases
- EC 2.5 - Transferring alkyl or aryl groups, other than methyl groups
- EC 2.6 - Transferring nitrogenous groups
- EC 2.7 - **Transferring phosphorus-containing groups**
- EC 2.8 - Transferring sulfur-containing groups
- EC 2.9 - Transferring selenium-containing groups
- EC 2.10 - Transferring molybdenum- or tungsten-containing groups

- .../...
- 2.7.1.127 - inositol-trisphosphate 3-kinase
 - 2.7.1.130 - tetraacyldisaccharide 4'-kinase
 - 2.7.1.134 - inositol-tetrakisphosphate 1-kinase
 - 2.7.1.136 - macrolide 2'-kinase
 - 2.7.1.137 - **phosphatidylinositol 3-kinase**
 - 2.7.1.138 - ceramide kinase
 - 2.7.1.140 - inositol-tetrakisphosphate 5-kinase
 - 2.7.1.142 - glycerol-3-phosphate—glucose phosphotransferase
 - 2.7.1.143 - diphosphate-purine nucleoside kinase
 - 2.7.1.144 - tagatose-6-phosphate kinase
 - 2.7.1.145 - deoxynucleoside kinase
 - 2.7.1.146 - ADP-specific phosphofructokinase
 - 2.7.1.147 - ADP-specific glucokinase
- .../...

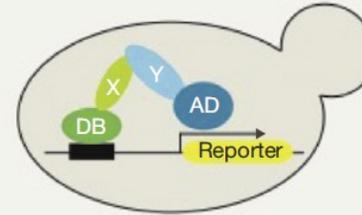
<https://www.ebi.ac.uk/enzymeportal/>

Reconstruction directe des interactions protéines-protéines

(a)

Binary mapping

Yeast two-hybrid (Y2H)

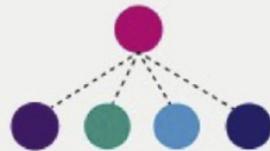


Interaction 2 à 2

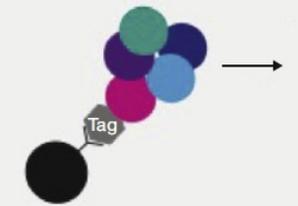
(b)

Co-complex mapping

Affinity purification followed by mass spectrometry (AP-MS)



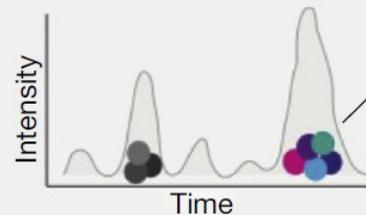
Identification by mass spectrometry



Co-fractionation followed by mass spectrometry



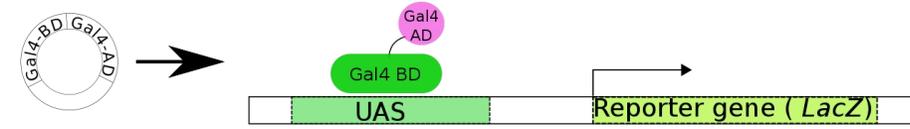
Identification by mass spectrometry



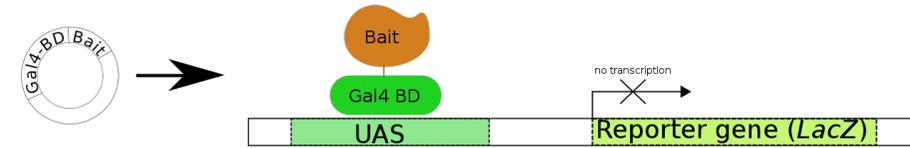
Complexe protéique

● Protein — Direct physical interaction Protein association

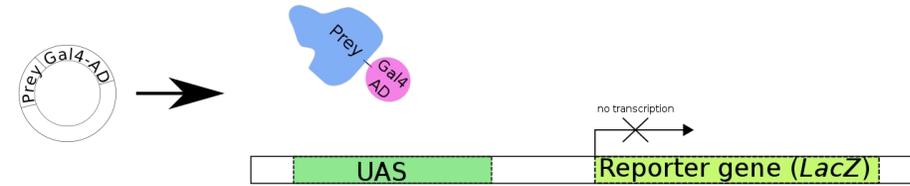
Reconstruction par mesure directe : Yeast-Two-Hybrid



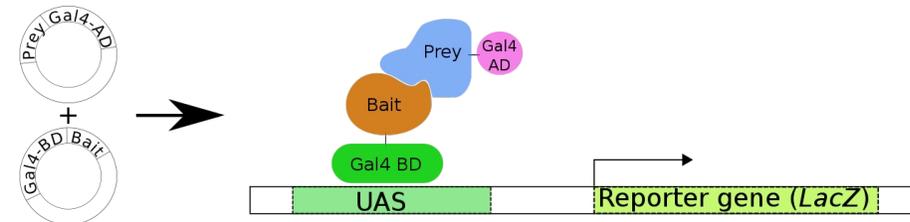
A. Regular transcription of the reporter gene



B. One fusion protein only (Gal4-BD + Bait) - no transcription

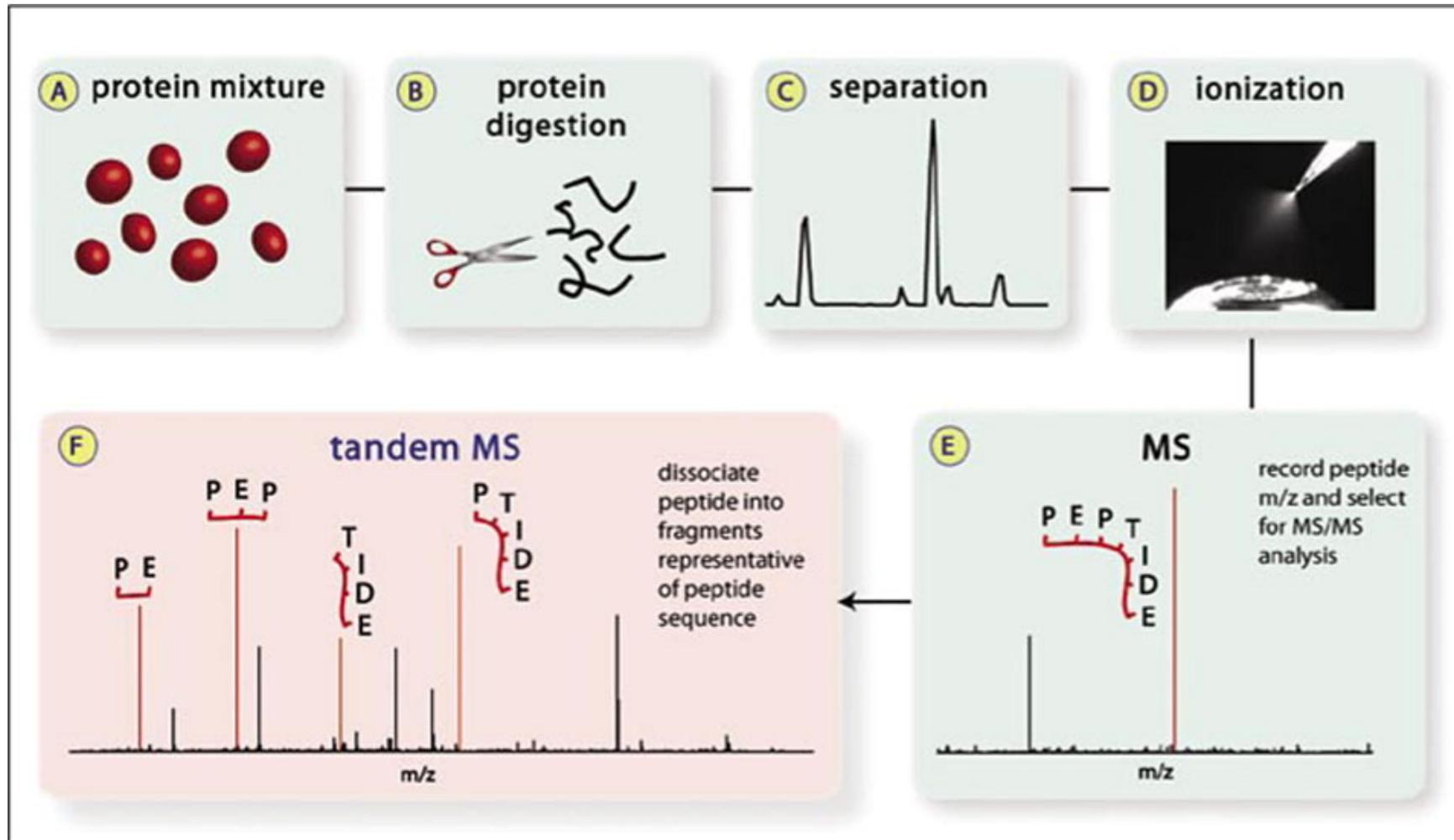


C. One fusion protein only (Gal4-AD + Prey) - no transcription



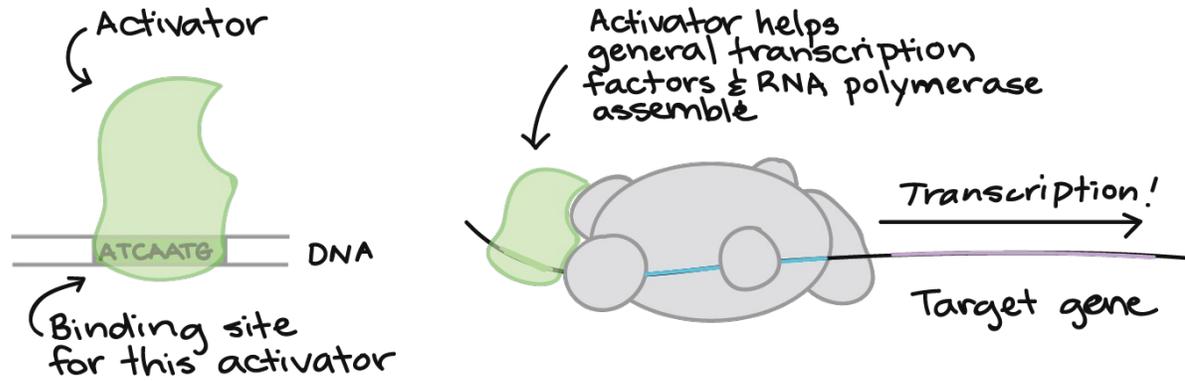
D. Two fusion proteins with interacting Bait and Prey

Reconstruction par mesure directe : Spectrométrie de masse

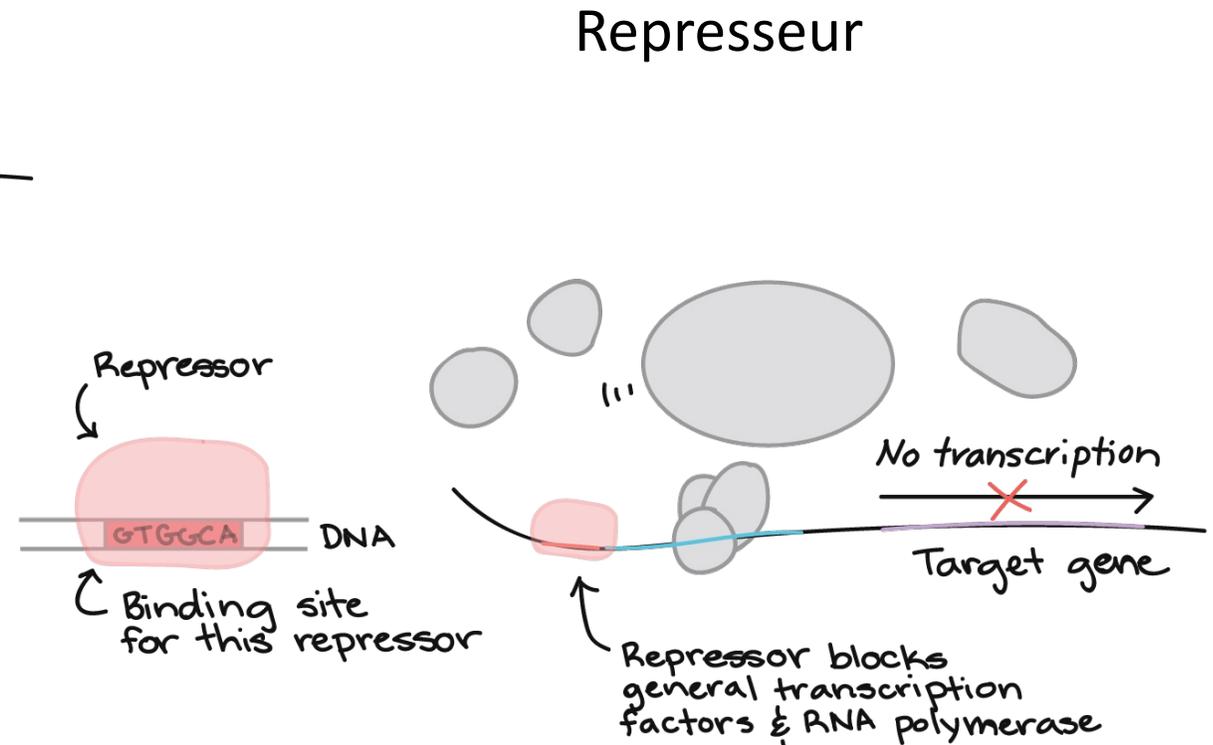


Reconstruction directe des réseaux génétiques

Détection des facteurs de transcription



Activateur



Reconstruction directe des réseaux génétiques

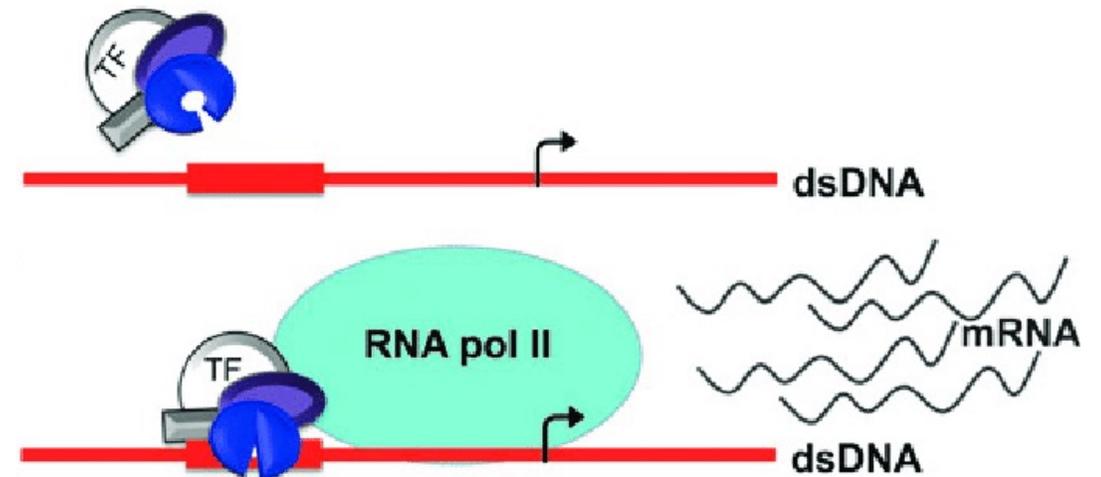
Détection des facteurs de transcription

Présence de motif



Domaine de liaison à l'ADN

Transcription Factor	Motif	Motif Fold Enrichment
<i>Neurod</i> family	gCCATCTG _{cc}	2.39
<i>Lhx / Lmx</i> family	TIAATTAA	2.42
<i>Nfi</i> family dimer	TGGCA _{ca} GCCAA	4.14
<i>Rfx</i> family dimer	GT _T CCATGG _{ac}	3.33
Novel <i>Hox</i> dimer	TAAT _{ta} TAAT _{ta}	2.32
Novel <i>Nfi</i> dimer	TGCCA _{ca} TGCCA	2.06



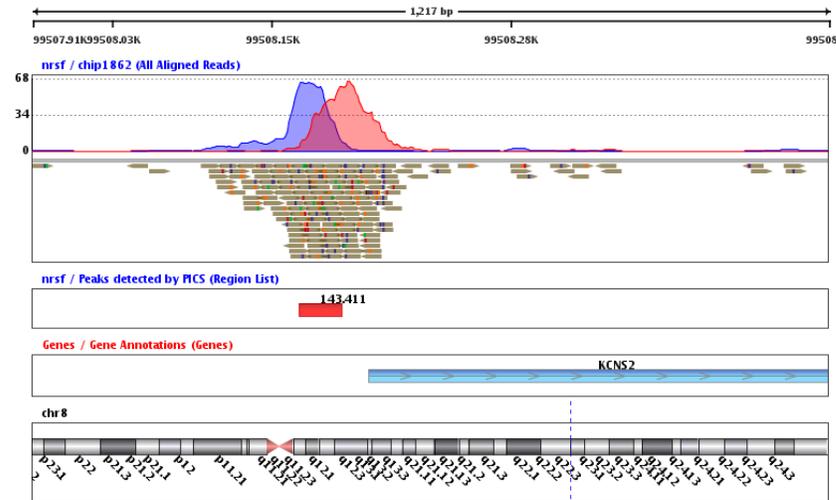
Reconstruction directe des réseaux génétiques

Détection des facteurs de transcription

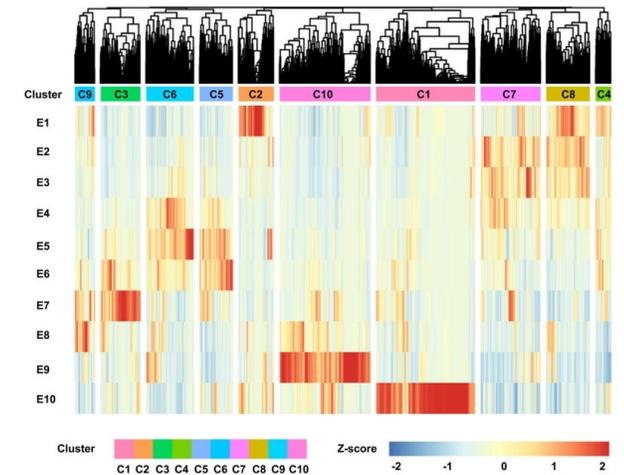
Recherche de motif

Transcription Factor	Motif	Motif Fold Enrichment
<i>Neurod</i> family		2.39
<i>Lhx / Lmx</i> family		2.42
<i>Nfi</i> family dimer		4.14
<i>Rfx</i> family dimer		3.33
Novel <i>Hox</i> dimer		2.32
Novel <i>Nfi</i> dimer		2.06

Méthode CHIPSeq



Validation des TF



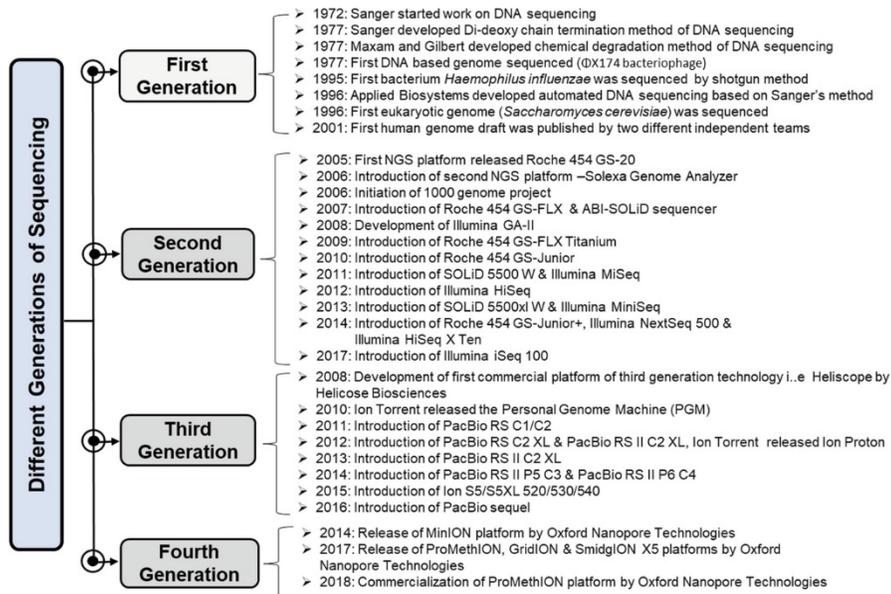
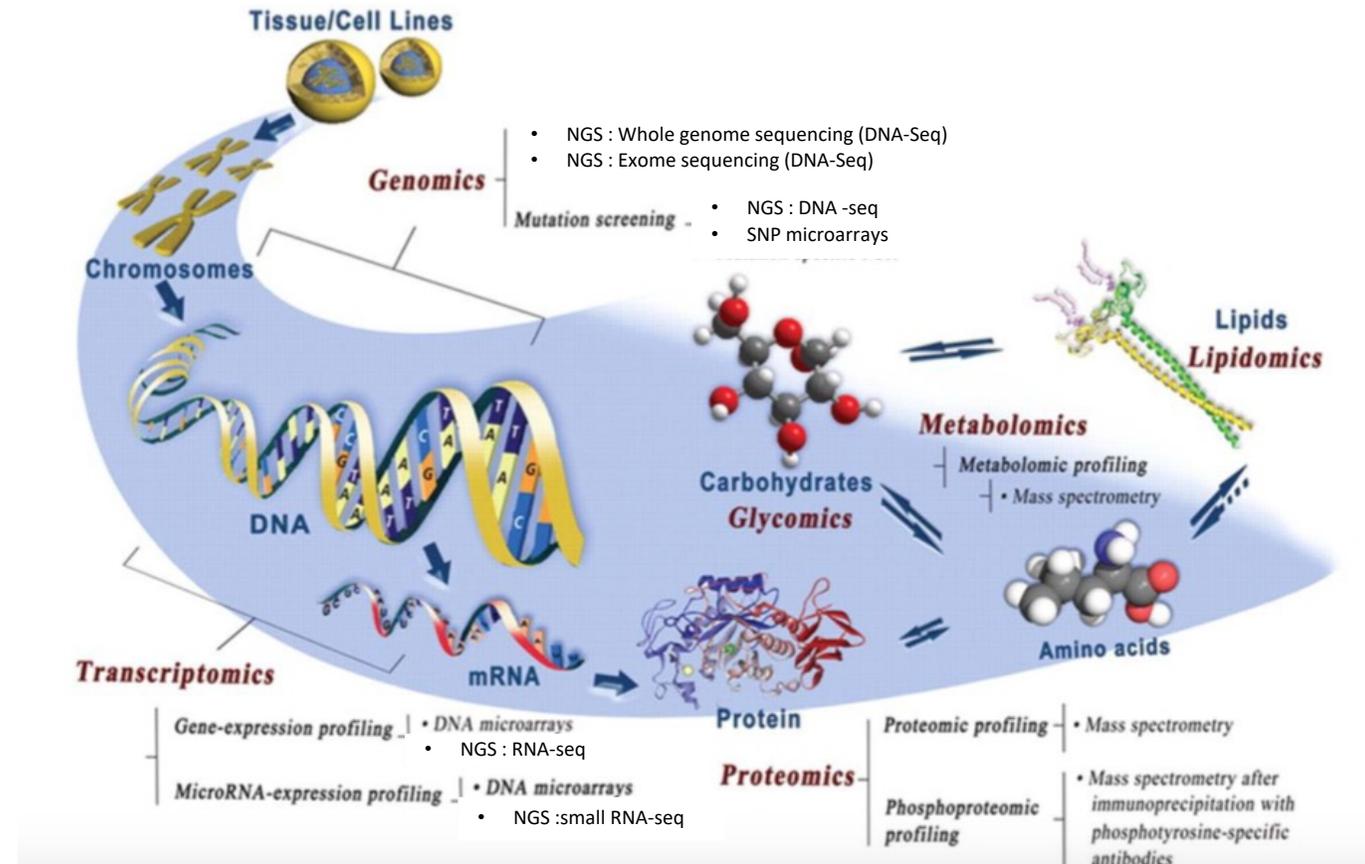
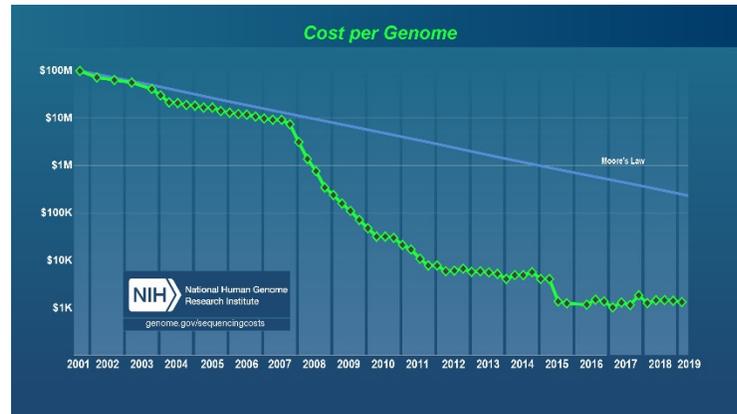
Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

- **Reconstruction de réseaux biologiques à l'aide de données omiques**

1. Quels sont les éléments à reconstruire ?
2. Quel type de réseaux peut-on reconstruire ?
3. Pour quel type de modélisation ?
4. Les différents types de reconstruction
5. Reconstruction directe
6. **Les données omiques**
7. Reconstruction indirecte
8. Les limites de la reconstruction de réseaux biologiques



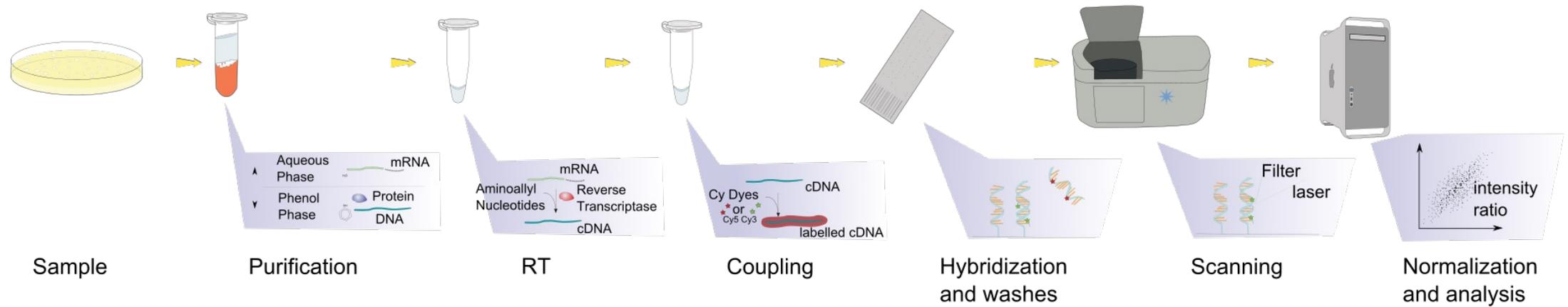
La révolution omique



Wu R.Q., J. dent. Research, 2010

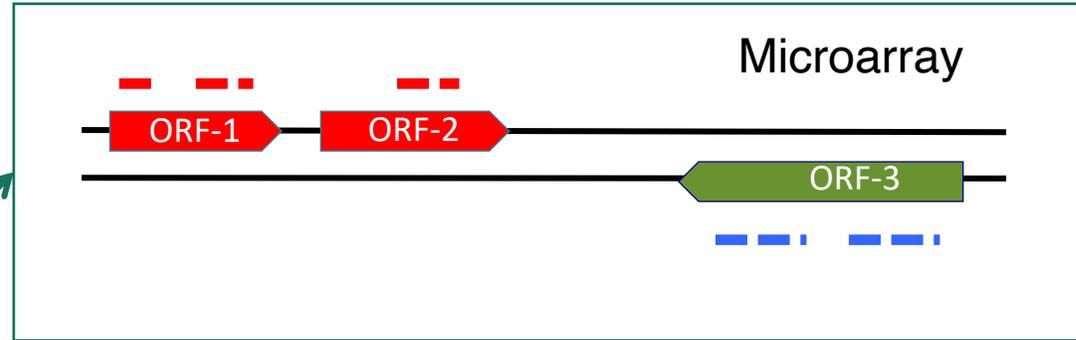
Evolution de la transcriptomique

Les prémices avec les puces ADN

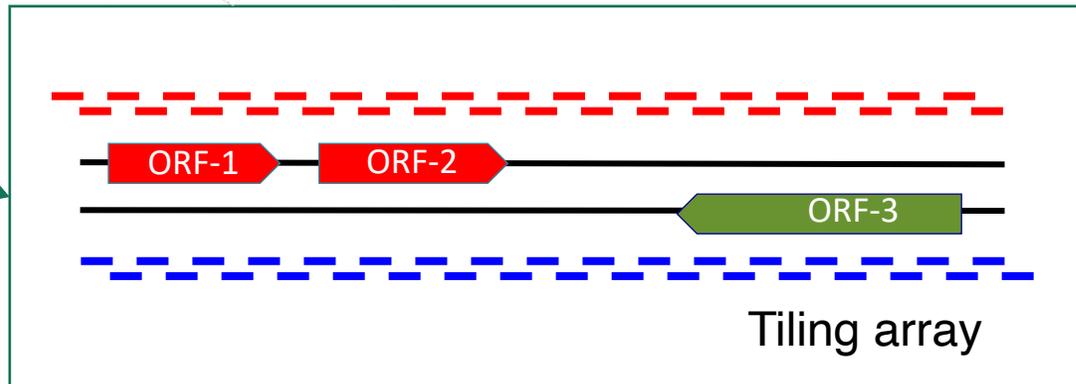


- Expression des ARNs
- Comparaison de génomes
- SNP
- ChIP

Les puces ADN



Extraction mRNA



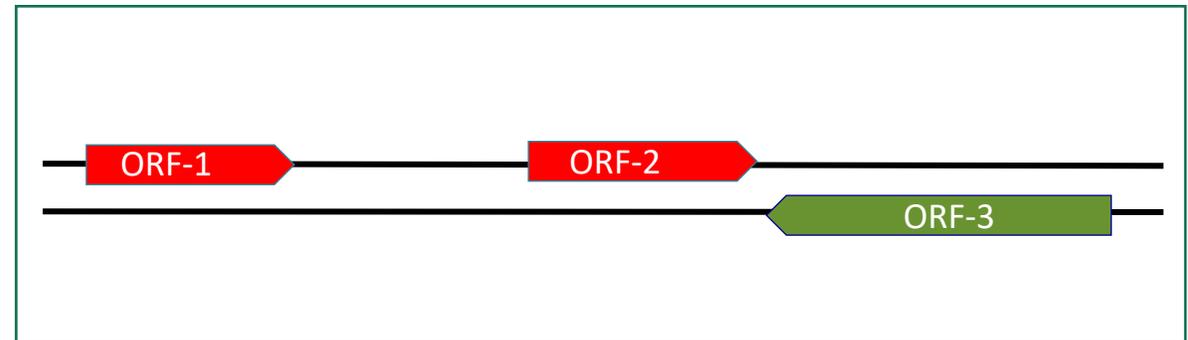
Séquençage haut débit - short-read



mRNA extraction



Reads mapping



For all you seq...

RNA

RNA Transcription

RNA-Protein Interactions

RNA Modification

RNA Low-Level Detection

RNA Structure

Sequencing

Reporting by Platforms

illumina

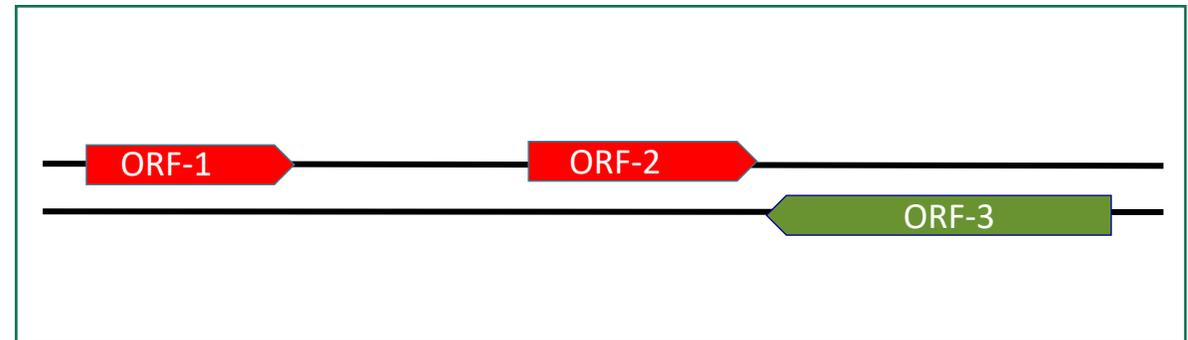
Séquençage haut débit - long-read



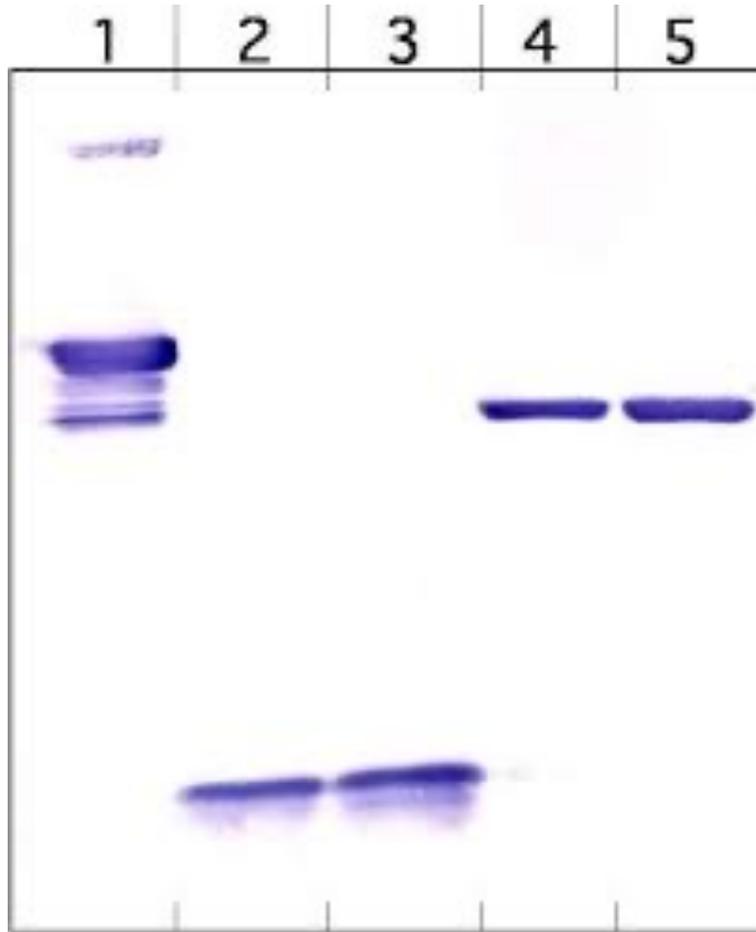
mRNA extraction



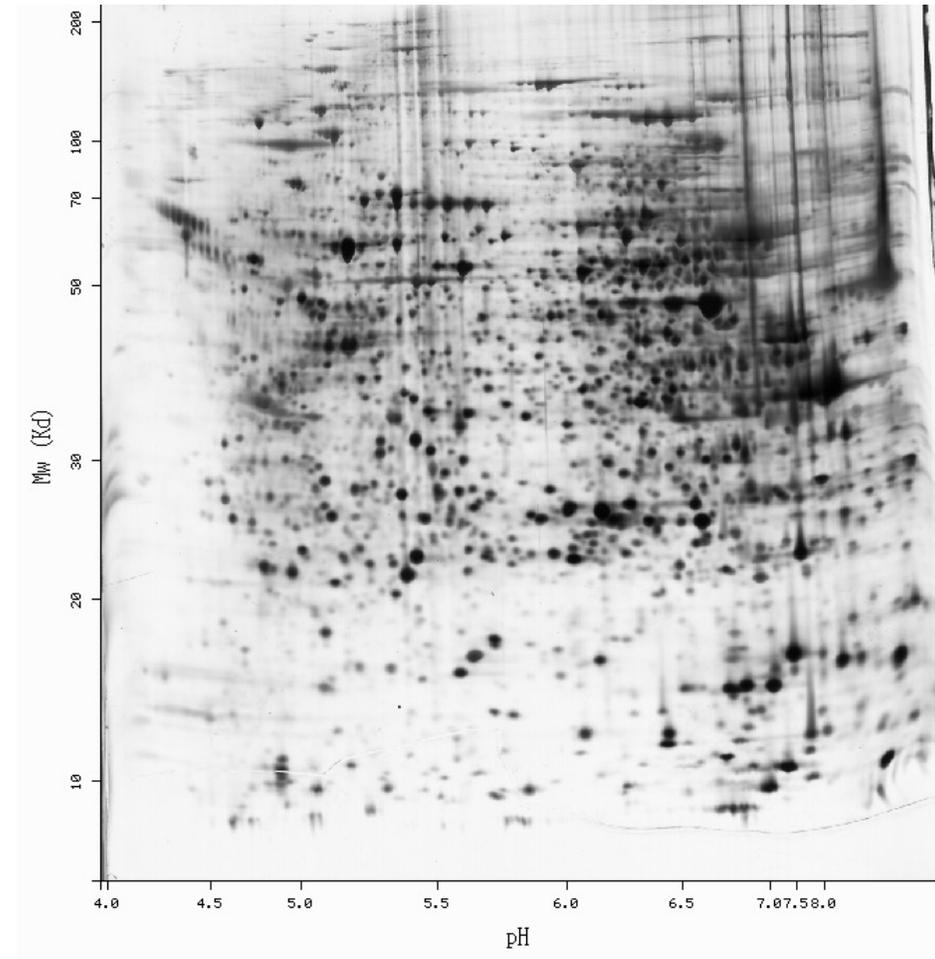
Reads mapping



Evolution de la protéomique

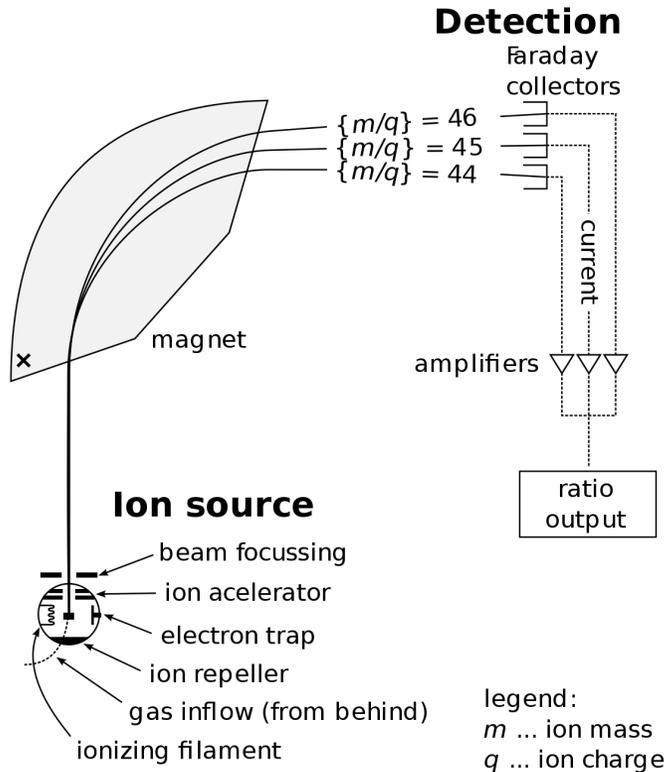


Western blot : Mesure de protéine unique



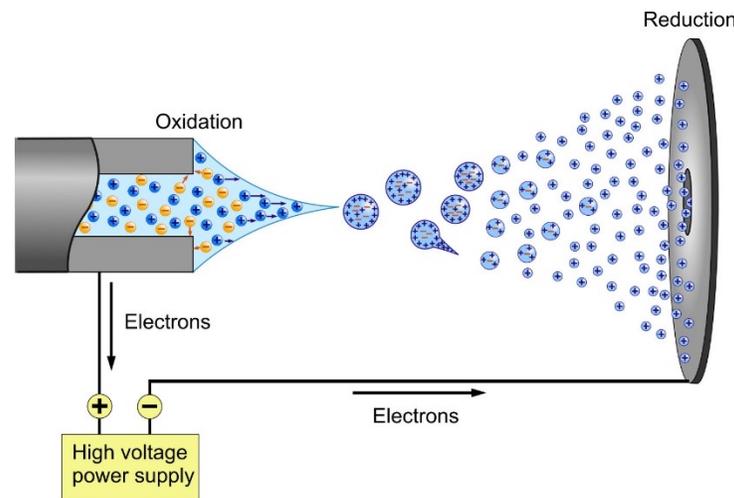
Gel 2D : Mesure de plusieurs protéines

Evolution de la protéomique

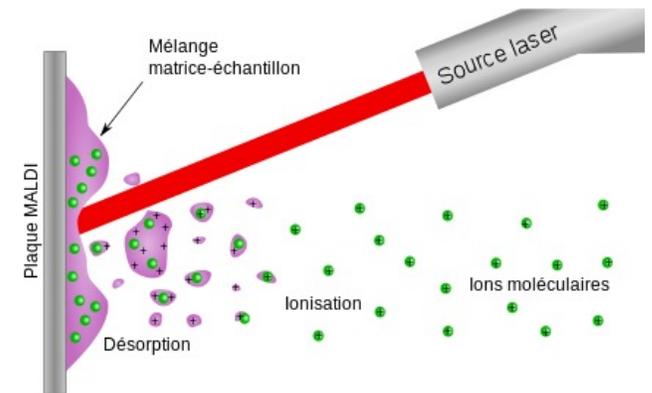


Spectrométrie de masse
1888 - 1918

Pour pouvoir passer les protéines dans un spectromètre
il faut les ioniser



Ionisation des protéines par spray
ESI - 1968



Ionisation des protéines par laser
MALDI - 1985

Evolution de la protéomique

Les protéines ionisés sont injectés dans le spectromètre de masse.

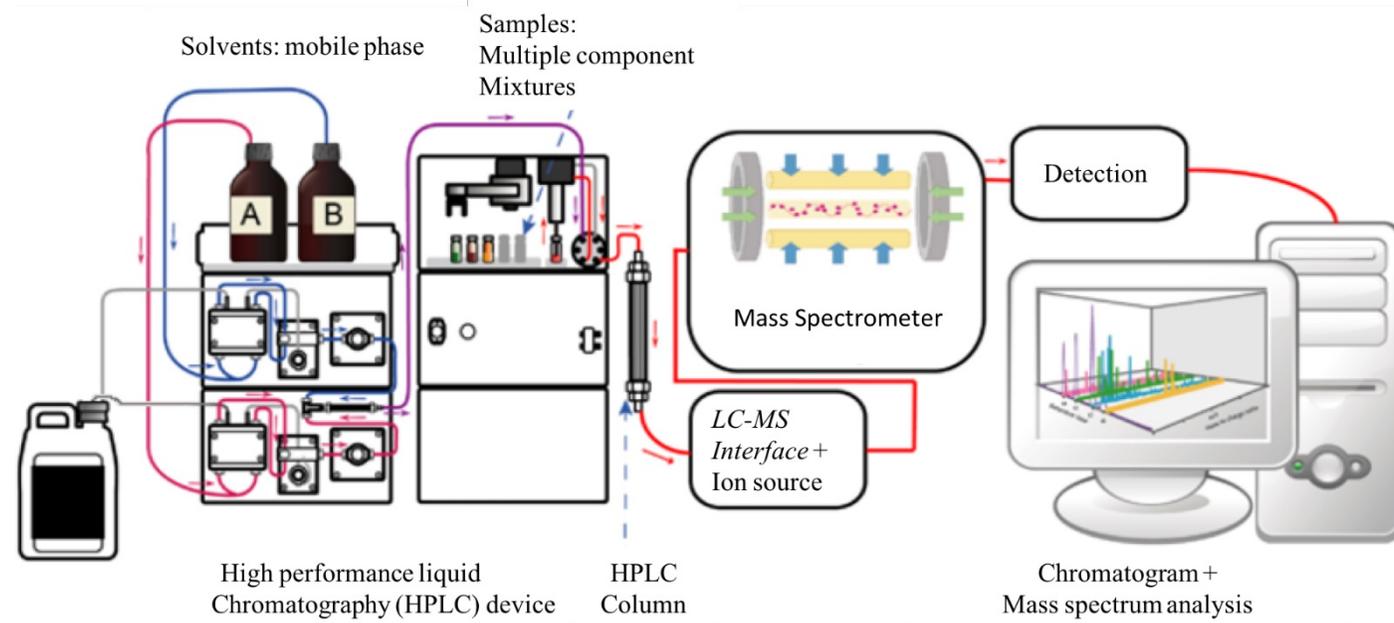
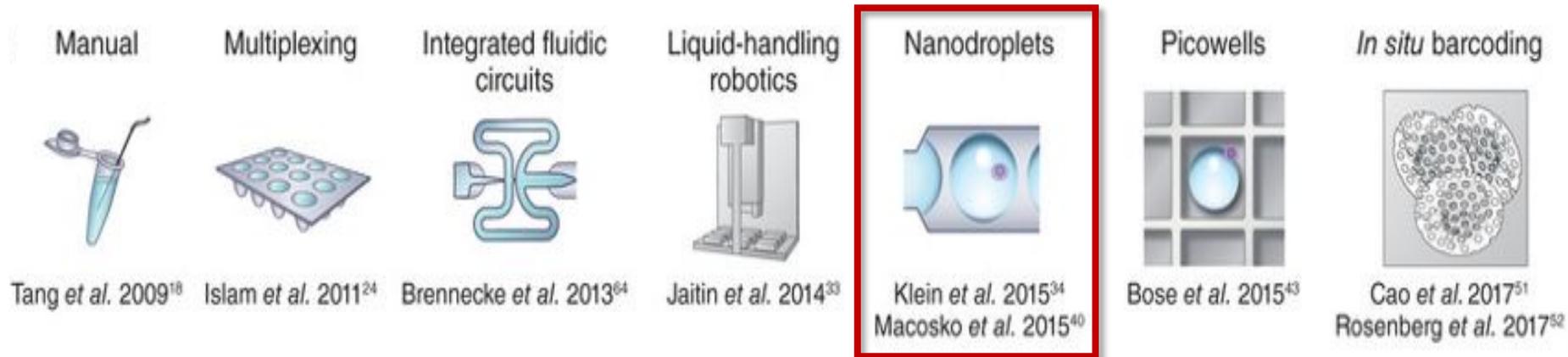
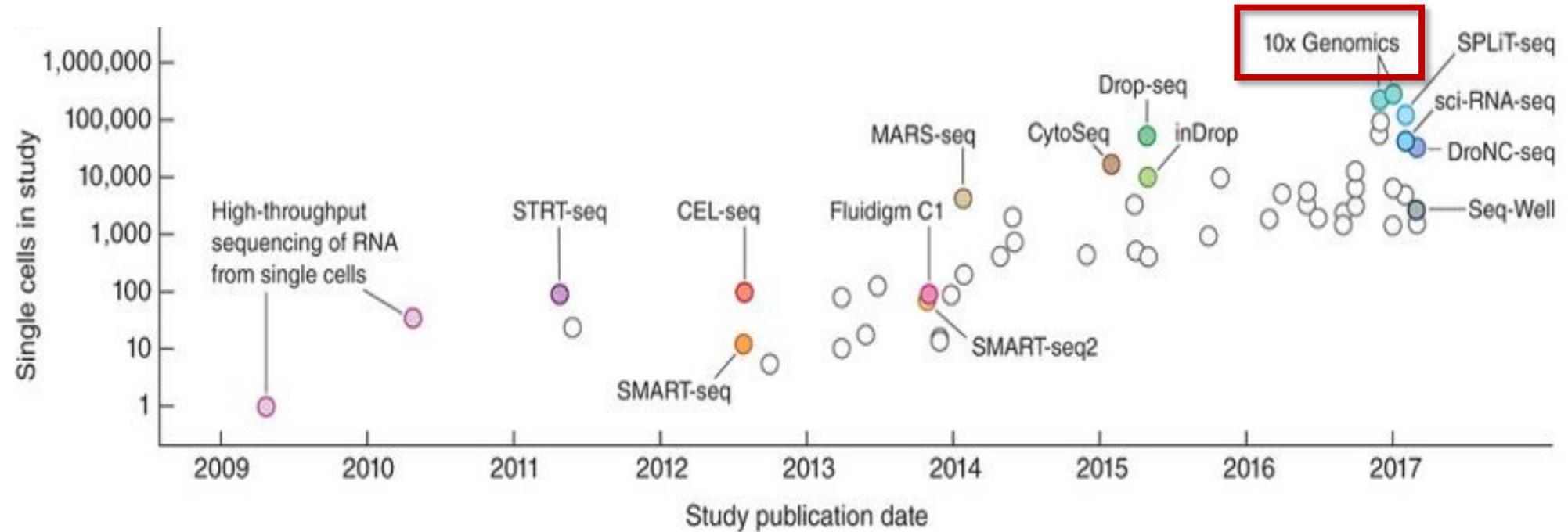


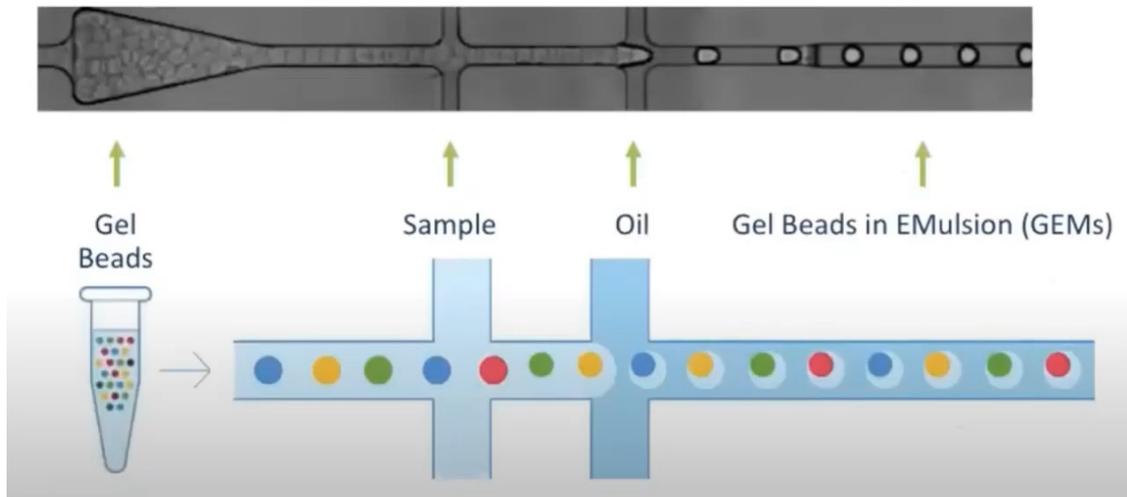
Schéma du LiquidChromatography/MassSpectrometry LC/MS - 2009

On rajoute un fractionnement des peptides pour avoir une meilleure précision
LC/MS/MS

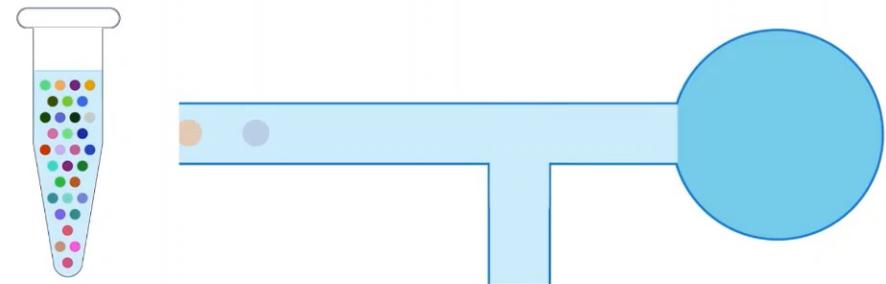
Evolution du single-cell



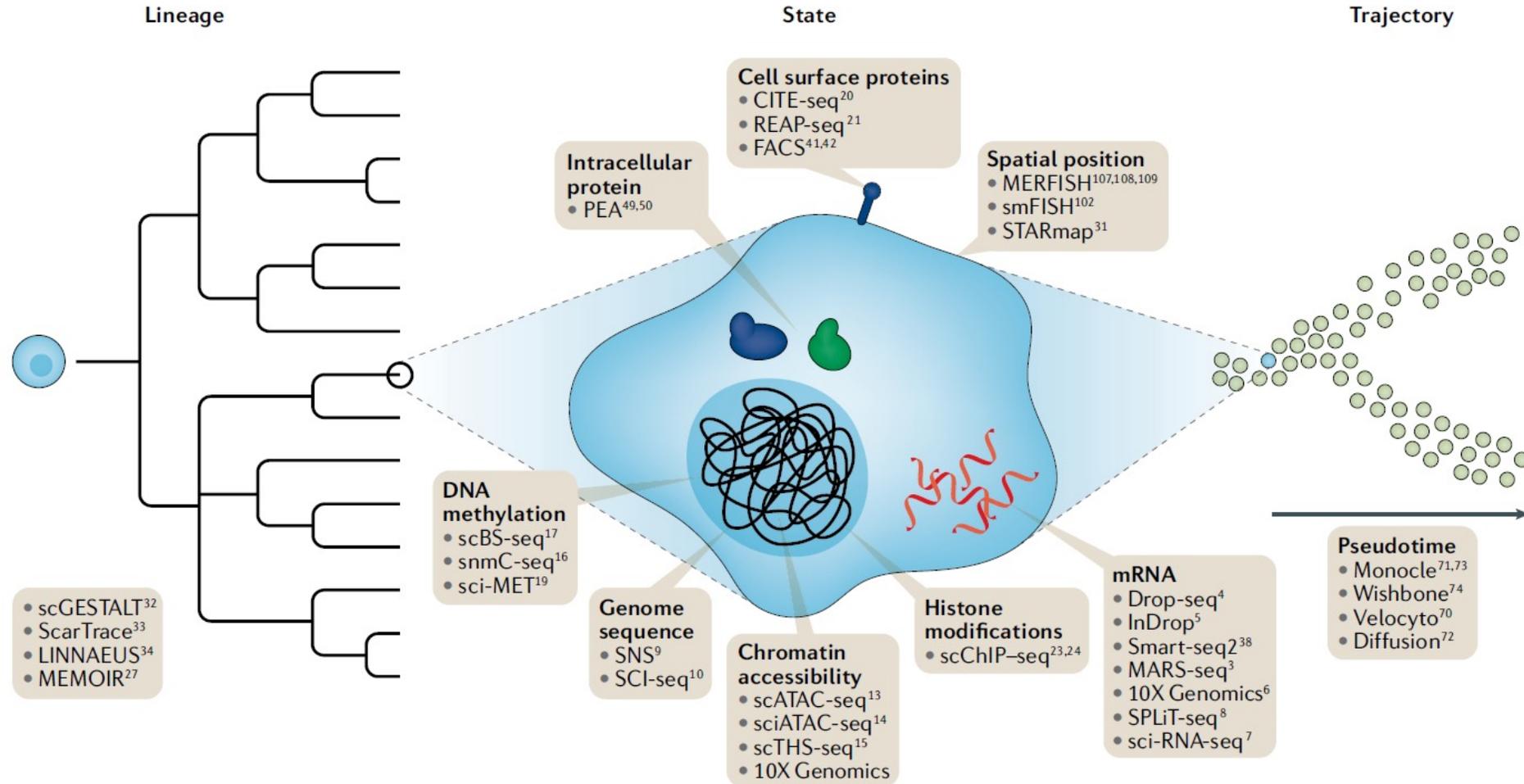
Séquençage à cellule unique



10x Next GEM Technology for Single Cell Partitioning



Une variété de techniques single-cell



Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

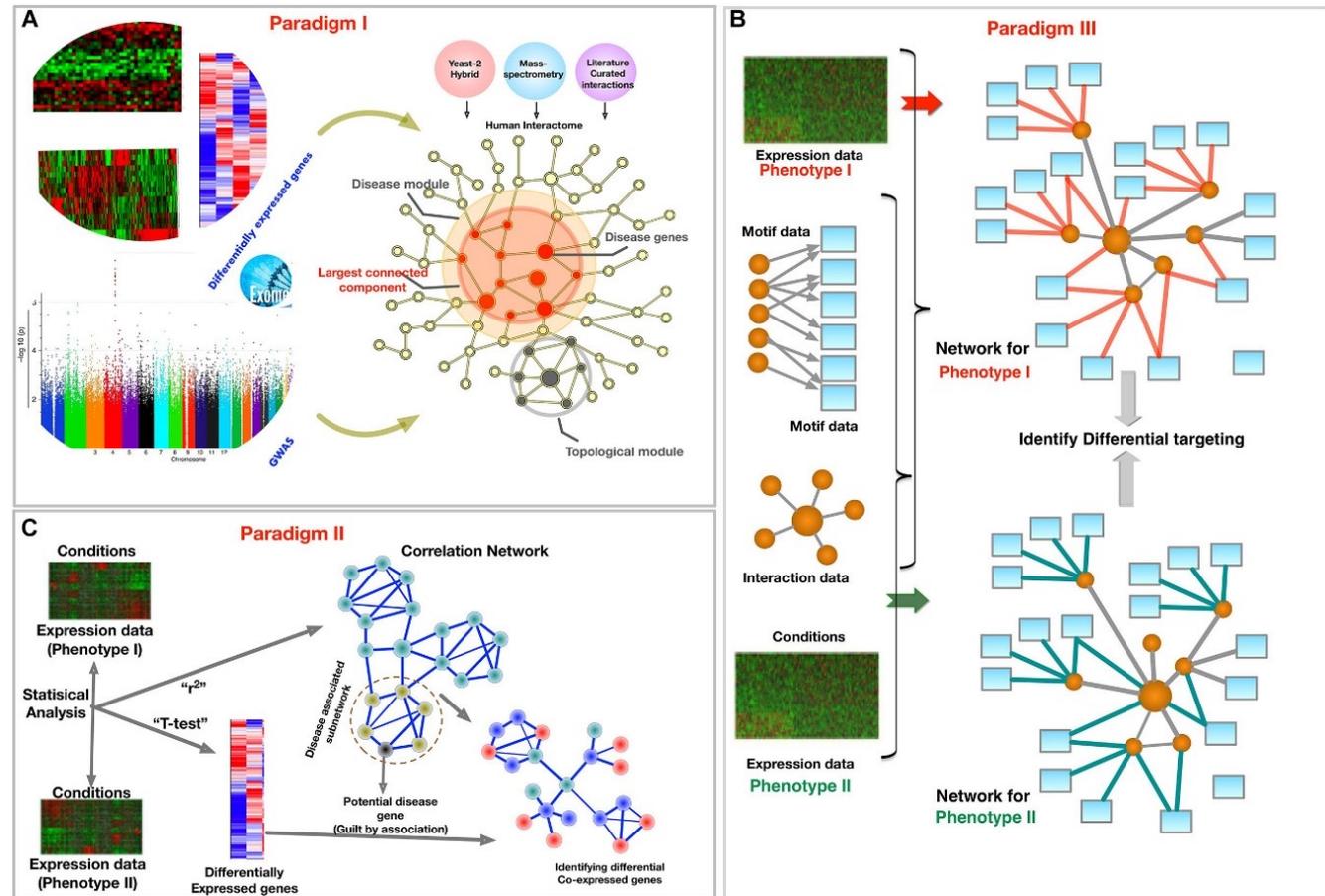
- **Reconstruction de réseaux biologiques à l'aide de données omiques**

1. Quels sont les éléments à reconstruire ?
2. Quel type de réseaux peut-on reconstruire ?
3. Pour quel type de modélisation ?
4. Les différents types de reconstruction
5. Reconstruction directe
6. Les données omiques
7. **Reconstruction indirecte**
8. Les limites de la reconstruction de réseaux biologiques

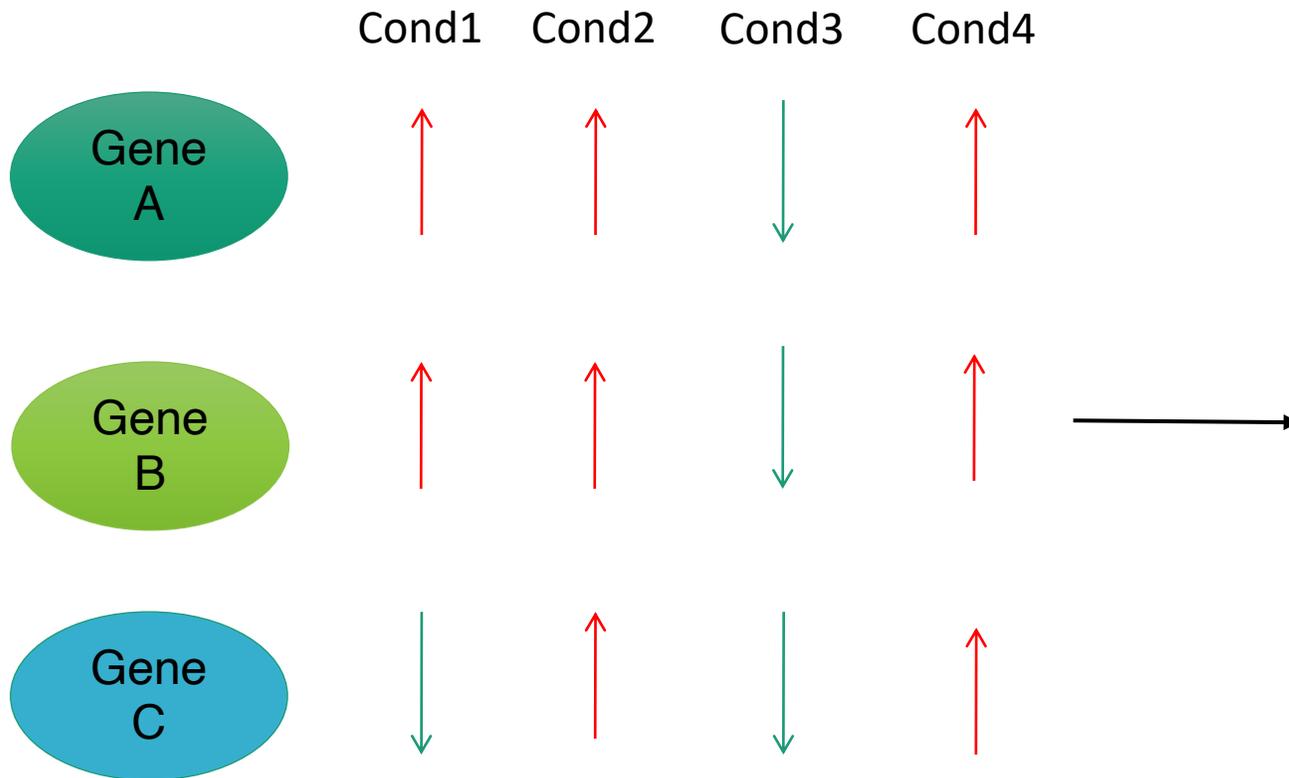


Reconstruction indirecte des réseaux biologiques

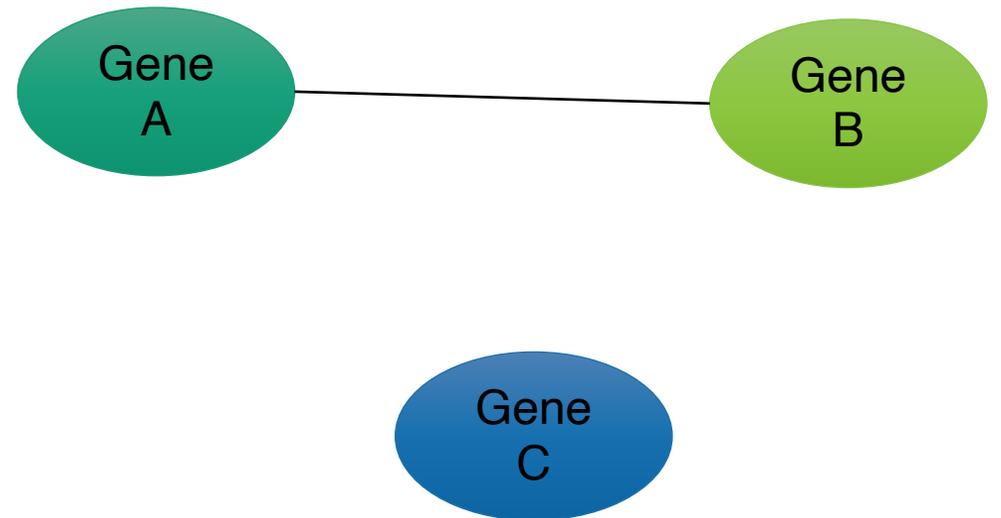
- Corrélation de données multi-omiques
- Réseau de co-expression
- Corrélation de différents réseaux biologiques



Réseau de co-expression



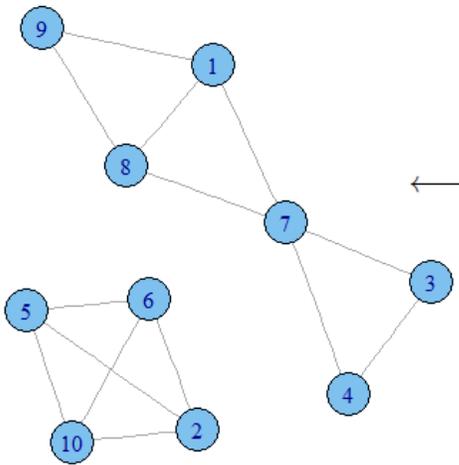
Gene Regulatory Network (GRN)



« Guilt by association »

Réseau de co-expression

	$ r(G_i, G_j) $ Pearson correlation																																																																																																																																																																						
<table border="1" style="border-collapse: collapse; text-align: center;"> <thead> <tr> <th></th> <th>S_1</th> <th>S_2</th> <th>S_3</th> </tr> </thead> <tbody> <tr><td>G_1</td><td>43.26</td><td>40.89</td><td>5.05</td></tr> <tr><td>G_2</td><td>166.6</td><td>41.87</td><td>136.65</td></tr> <tr><td>G_3</td><td>12.53</td><td>39.55</td><td>42.09</td></tr> <tr><td>G_4</td><td>28.77</td><td>191.92</td><td>236.56</td></tr> <tr><td>G_5</td><td>114.7</td><td>79.7</td><td>99.76</td></tr> <tr><td>G_6</td><td>119.1</td><td>80.57</td><td>114.59</td></tr> <tr><td>G_7</td><td>118.9</td><td>156.69</td><td>186.95</td></tr> <tr><td>G_8</td><td>3.76</td><td>2.48</td><td>136.78</td></tr> <tr><td>G_9</td><td>32.73</td><td>11.99</td><td>118.8</td></tr> <tr><td>G_{10}</td><td>17.46</td><td>56.11</td><td>21.41</td></tr> </tbody> </table> <p style="text-align: center;">Gene expression values</p>		S_1	S_2	S_3	G_1	43.26	40.89	5.05	G_2	166.6	41.87	136.65	G_3	12.53	39.55	42.09	G_4	28.77	191.92	236.56	G_5	114.7	79.7	99.76	G_6	119.1	80.57	114.59	G_7	118.9	156.69	186.95	G_8	3.76	2.48	136.78	G_9	32.73	11.99	118.8	G_{10}	17.46	56.11	21.41	\rightarrow	<table border="1" style="border-collapse: collapse; text-align: center;"> <thead> <tr> <th></th> <th>G_1</th> <th>G_2</th> <th>G_3</th> <th>G_4</th> <th>G_5</th> <th>G_6</th> <th>G_7</th> <th>G_8</th> <th>G_9</th> <th>G_{10}</th> </tr> </thead> <tbody> <tr><td>G_1</td><td>1.00</td><td>0.23</td><td>0.61</td><td>0.71</td><td>0.03</td><td>0.35</td><td>0.86</td><td>1.00</td><td>0.97</td><td>0.37</td></tr> <tr><td>G_2</td><td>0.23</td><td>1.00</td><td>0.63</td><td>0.52</td><td>0.98</td><td>0.99</td><td>0.29</td><td>0.30</td><td>0.46</td><td>0.99</td></tr> <tr><td>G_3</td><td>0.61</td><td>0.63</td><td>1.00</td><td>0.99</td><td>0.77</td><td>0.53</td><td>0.93</td><td>0.56</td><td>0.41</td><td>0.51</td></tr> <tr><td>G_4</td><td>0.71</td><td>0.52</td><td>0.99</td><td>1.00</td><td>0.69</td><td>0.41</td><td>0.97</td><td>0.66</td><td>0.52</td><td>0.40</td></tr> <tr><td>G_5</td><td>0.03</td><td>0.98</td><td>0.77</td><td>0.69</td><td>1.00</td><td>0.95</td><td>0.48</td><td>0.09</td><td>0.27</td><td>0.94</td></tr> <tr><td>G_6</td><td>0.35</td><td>0.99</td><td>0.53</td><td>0.41</td><td>0.95</td><td>1.00</td><td>0.17</td><td>0.41</td><td>0.57</td><td>1.00</td></tr> <tr><td>G_7</td><td>0.86</td><td>0.29</td><td>0.93</td><td>0.97</td><td>0.48</td><td>0.17</td><td>1.00</td><td>0.83</td><td>0.72</td><td>0.16</td></tr> <tr><td>G_8</td><td>1.00</td><td>0.30</td><td>0.56</td><td>0.66</td><td>0.09</td><td>0.41</td><td>0.83</td><td>1.00</td><td>0.98</td><td>0.42</td></tr> <tr><td>G_9</td><td>0.97</td><td>0.46</td><td>0.41</td><td>0.52</td><td>0.27</td><td>0.57</td><td>0.72</td><td>0.98</td><td>1.00</td><td>0.58</td></tr> <tr><td>G_{10}</td><td>0.37</td><td>0.99</td><td>0.51</td><td>0.40</td><td>0.94</td><td>1.00</td><td>0.16</td><td>0.42</td><td>0.58</td><td>1.00</td></tr> </tbody> </table> <p style="text-align: center;">Similarity (Co-expression) score</p>		G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}	G_1	1.00	0.23	0.61	0.71	0.03	0.35	0.86	1.00	0.97	0.37	G_2	0.23	1.00	0.63	0.52	0.98	0.99	0.29	0.30	0.46	0.99	G_3	0.61	0.63	1.00	0.99	0.77	0.53	0.93	0.56	0.41	0.51	G_4	0.71	0.52	0.99	1.00	0.69	0.41	0.97	0.66	0.52	0.40	G_5	0.03	0.98	0.77	0.69	1.00	0.95	0.48	0.09	0.27	0.94	G_6	0.35	0.99	0.53	0.41	0.95	1.00	0.17	0.41	0.57	1.00	G_7	0.86	0.29	0.93	0.97	0.48	0.17	1.00	0.83	0.72	0.16	G_8	1.00	0.30	0.56	0.66	0.09	0.41	0.83	1.00	0.98	0.42	G_9	0.97	0.46	0.41	0.52	0.27	0.57	0.72	0.98	1.00	0.58	G_{10}	0.37	0.99	0.51	0.40	0.94	1.00	0.16	0.42	0.58	1.00
	S_1	S_2	S_3																																																																																																																																																																				
G_1	43.26	40.89	5.05																																																																																																																																																																				
G_2	166.6	41.87	136.65																																																																																																																																																																				
G_3	12.53	39.55	42.09																																																																																																																																																																				
G_4	28.77	191.92	236.56																																																																																																																																																																				
G_5	114.7	79.7	99.76																																																																																																																																																																				
G_6	119.1	80.57	114.59																																																																																																																																																																				
G_7	118.9	156.69	186.95																																																																																																																																																																				
G_8	3.76	2.48	136.78																																																																																																																																																																				
G_9	32.73	11.99	118.8																																																																																																																																																																				
G_{10}	17.46	56.11	21.41																																																																																																																																																																				
	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}																																																																																																																																																													
G_1	1.00	0.23	0.61	0.71	0.03	0.35	0.86	1.00	0.97	0.37																																																																																																																																																													
G_2	0.23	1.00	0.63	0.52	0.98	0.99	0.29	0.30	0.46	0.99																																																																																																																																																													
G_3	0.61	0.63	1.00	0.99	0.77	0.53	0.93	0.56	0.41	0.51																																																																																																																																																													
G_4	0.71	0.52	0.99	1.00	0.69	0.41	0.97	0.66	0.52	0.40																																																																																																																																																													
G_5	0.03	0.98	0.77	0.69	1.00	0.95	0.48	0.09	0.27	0.94																																																																																																																																																													
G_6	0.35	0.99	0.53	0.41	0.95	1.00	0.17	0.41	0.57	1.00																																																																																																																																																													
G_7	0.86	0.29	0.93	0.97	0.48	0.17	1.00	0.83	0.72	0.16																																																																																																																																																													
G_8	1.00	0.30	0.56	0.66	0.09	0.41	0.83	1.00	0.98	0.42																																																																																																																																																													
G_9	0.97	0.46	0.41	0.52	0.27	0.57	0.72	0.98	1.00	0.58																																																																																																																																																													
G_{10}	0.37	0.99	0.51	0.40	0.94	1.00	0.16	0.42	0.58	1.00																																																																																																																																																													

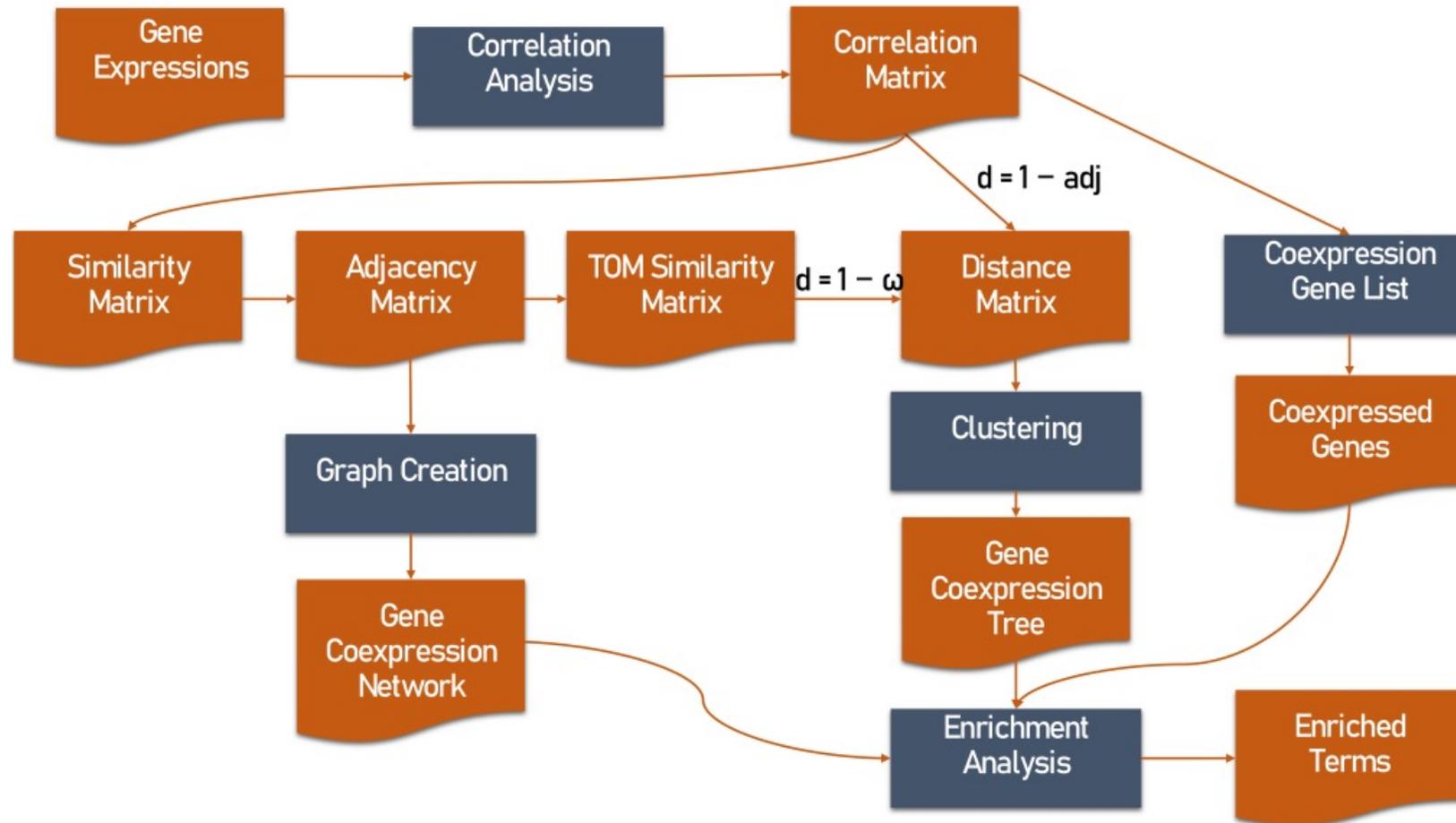


	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	0	0	0	0	0	0	1	1	1	0
G_2	0	0	0	0	1	1	0	0	0	1
G_3	0	0	0	1	0	0	1	0	0	0
G_4	0	0	1	0	0	0	1	0	0	0
G_5	0	1	0	0	0	1	0	0	0	1
G_6	0	1	0	0	1	0	0	0	0	1
G_7	1	0	1	1	0	0	0	1	0	0
G_8	1	0	0	0	0	0	1	0	1	0
G_9	1	0	0	0	0	0	0	1	0	0
G_{10}	0	1	0	0	1	1	0	0	0	0

$|r(G_i, G_j)| \geq 0.8$
 Significance threshold

Network adjacency matrix

Reconstruction de réseau de co-expression



4 classes de calculs de distances pour la reconstruction de réseau de co-expression

- Les méthodes probabilistes (ex: Methodes bayésiennes)
- Les méthodes basées sur la corrélation (ex: Pearson correlation)
- Les méthodes sur des calculs de distance statistiques (Gaussian graphical model)
- Les méthodes basées sur les calculs de distance en théorie de l'information (ex: information mutuelle)

Corrélation de Pearson

Pour deux vecteurs d'expression génique X et Y

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad (\text{Eq.2})$$

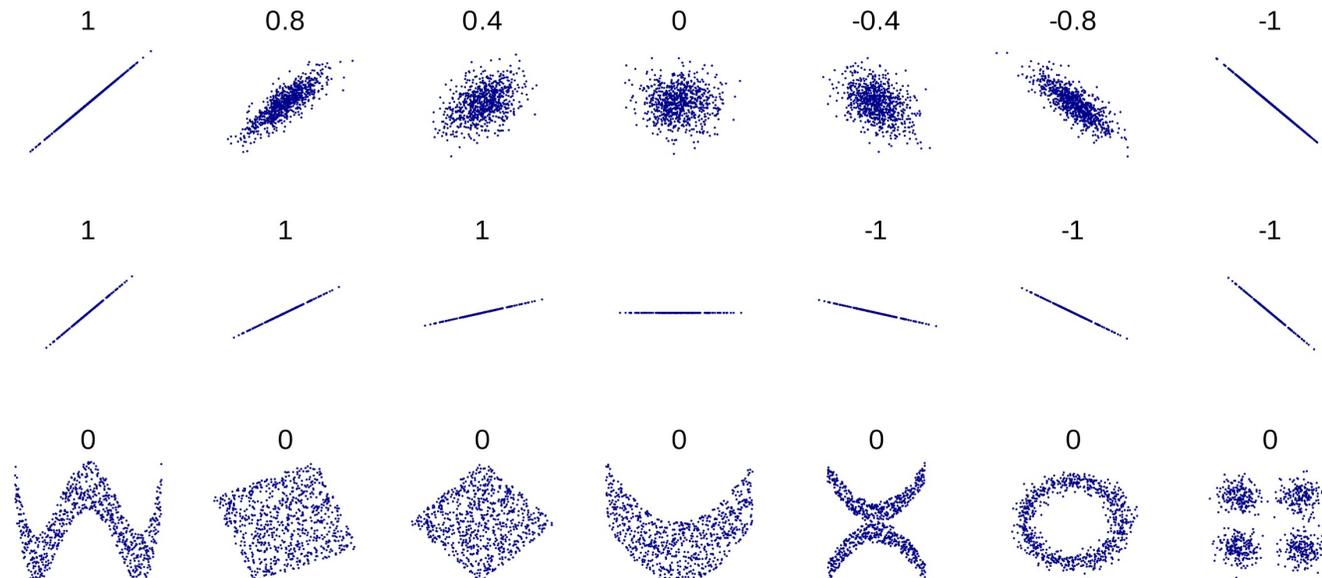
σ_X l'écart-type de X

σ_Y l'écart-type de Y

μ_X la moyenne de X

μ_Y la moyenne de Y

E l'espérance

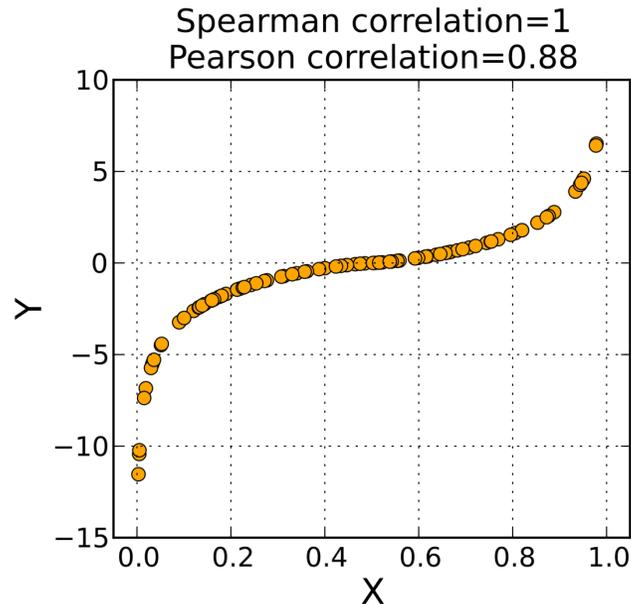


Autres métriques pour la co-expression

Corrélation de Spearman

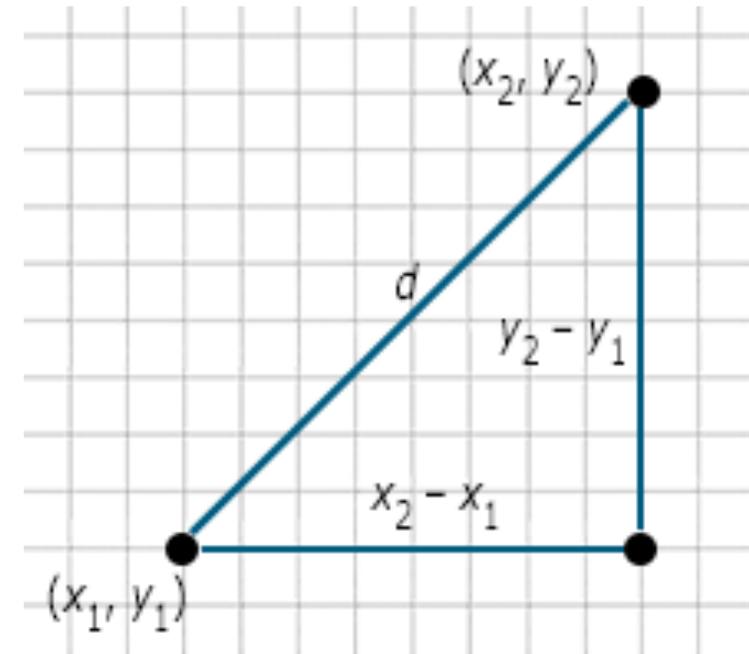
$$r_s = \frac{\text{cov}(\text{rg}_X, \text{rg}_Y)}{\sigma_{\text{rg}_X} \sigma_{\text{rg}_Y}}$$

Corrélation de rang



Distance euclidienne

$$d(x^i, x^j)^2 = \sum_k (x_k^i - x_k^j)^2$$



Information mutuelle

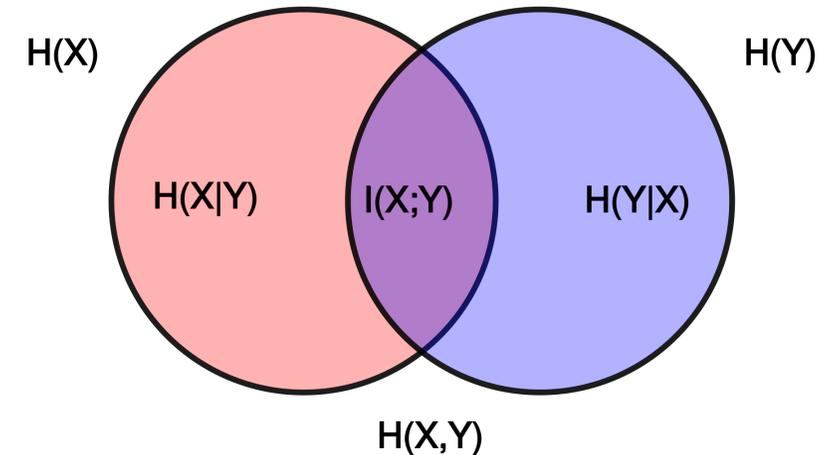
Mesure la dépendance statistique de deux variables aléatoires

$$I(X; Y) = \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x) P(y)},$$

Différents algorithmes existent pour estimer cette valeur I

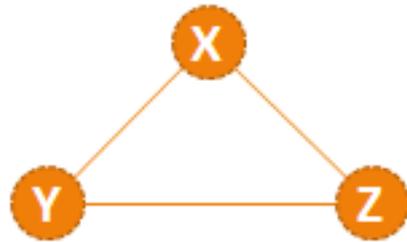
Théorie de l'information (Shannon et Weaver 1948)

$$H(X) = - \sum_i^n p_i \log_2(p_i)$$

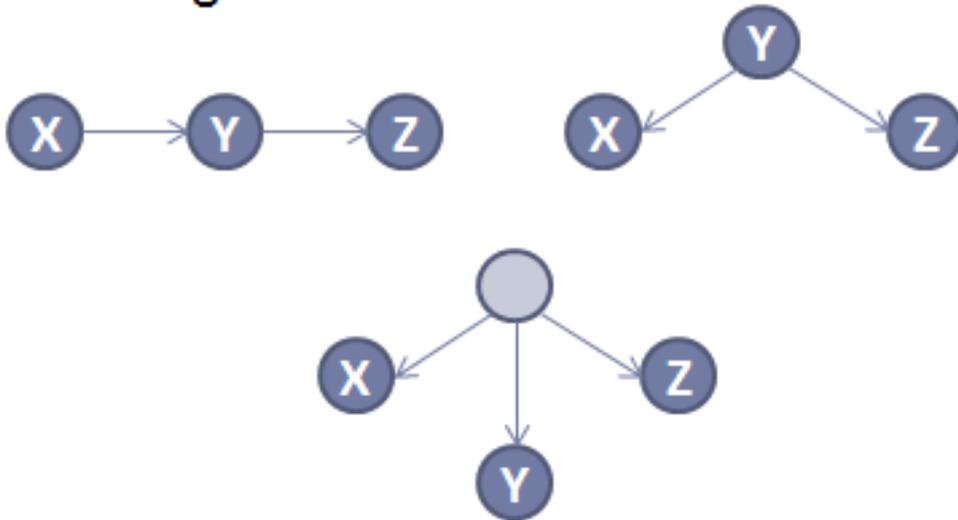


Corrélation = régulation ?

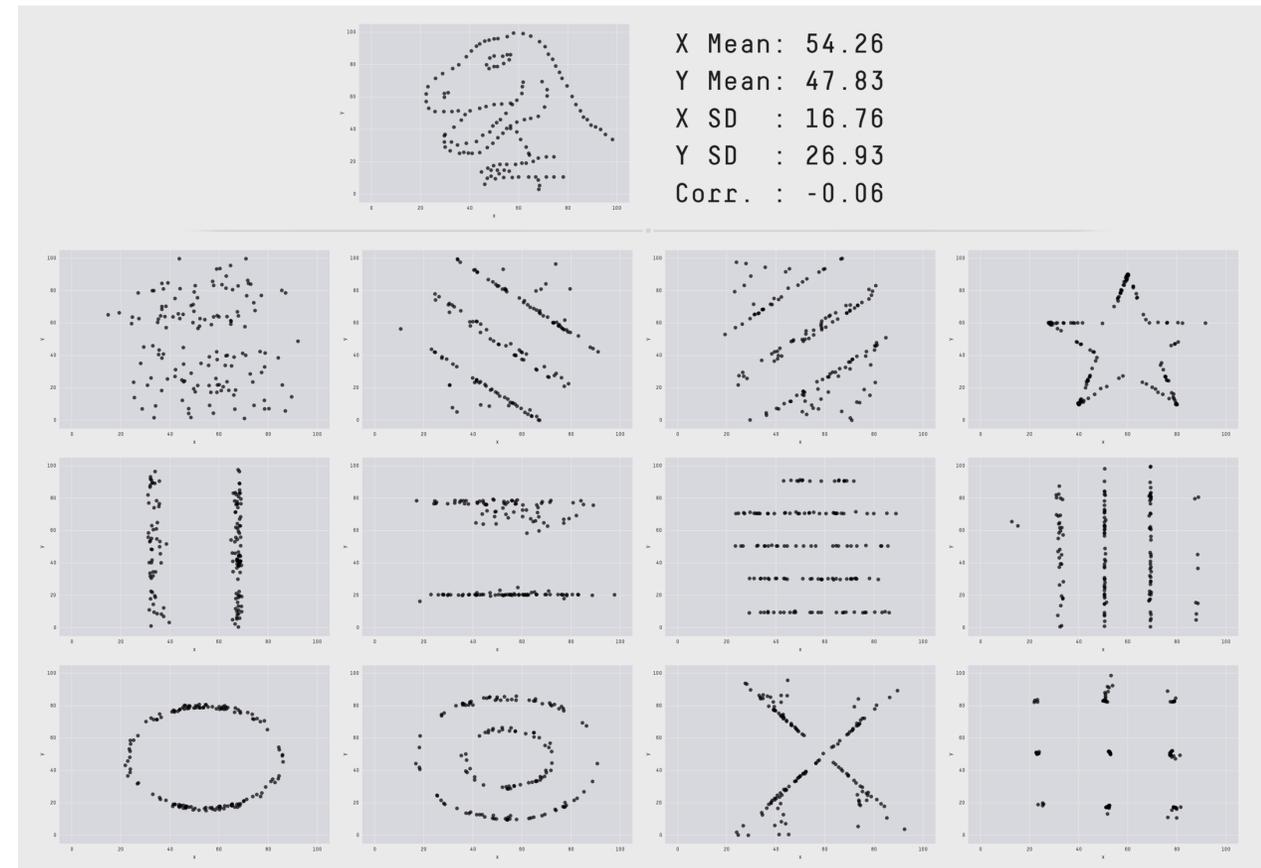
Gene Co-expression



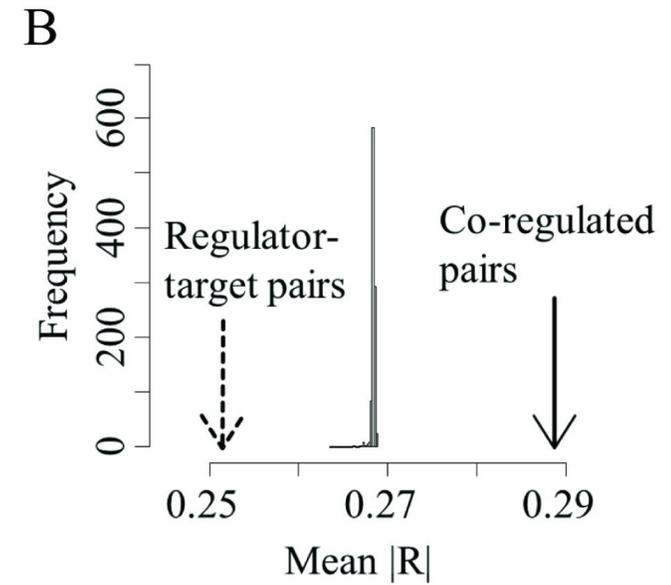
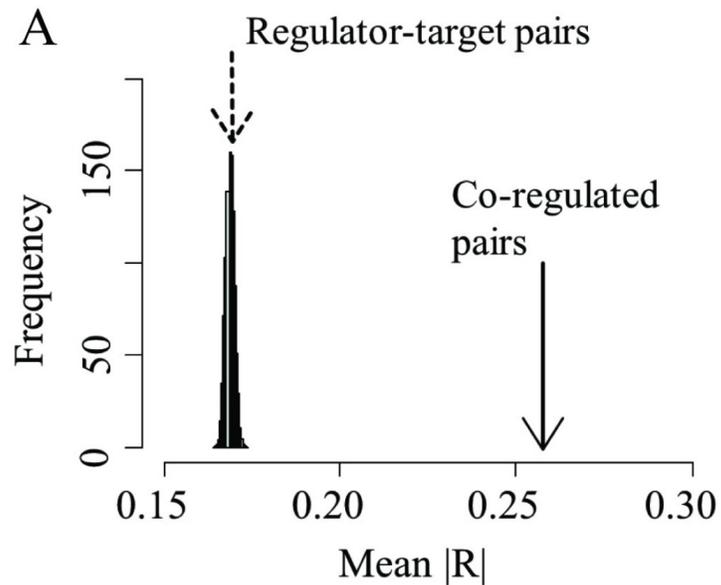
Gene Regulation



- Pas d'orientation dans les réseaux de co-expression
- Corrélation n'est pas causalité !



Est-ce une
bonne
approche ?



RESEARCH ARTICLE

Emergence of co-expression in gene regulatory networks

Wencheng Yin¹, Luis Mendoza^{2*}, Jimena Monzon-Sandoval³, Araxi O. Urrutia^{4,5},
Humberto Gutierrez^{6*}

Plos ONE 2021

Reconstruction à l'aide d'information génétique

- Gene co-occurrence

- Gene fusion

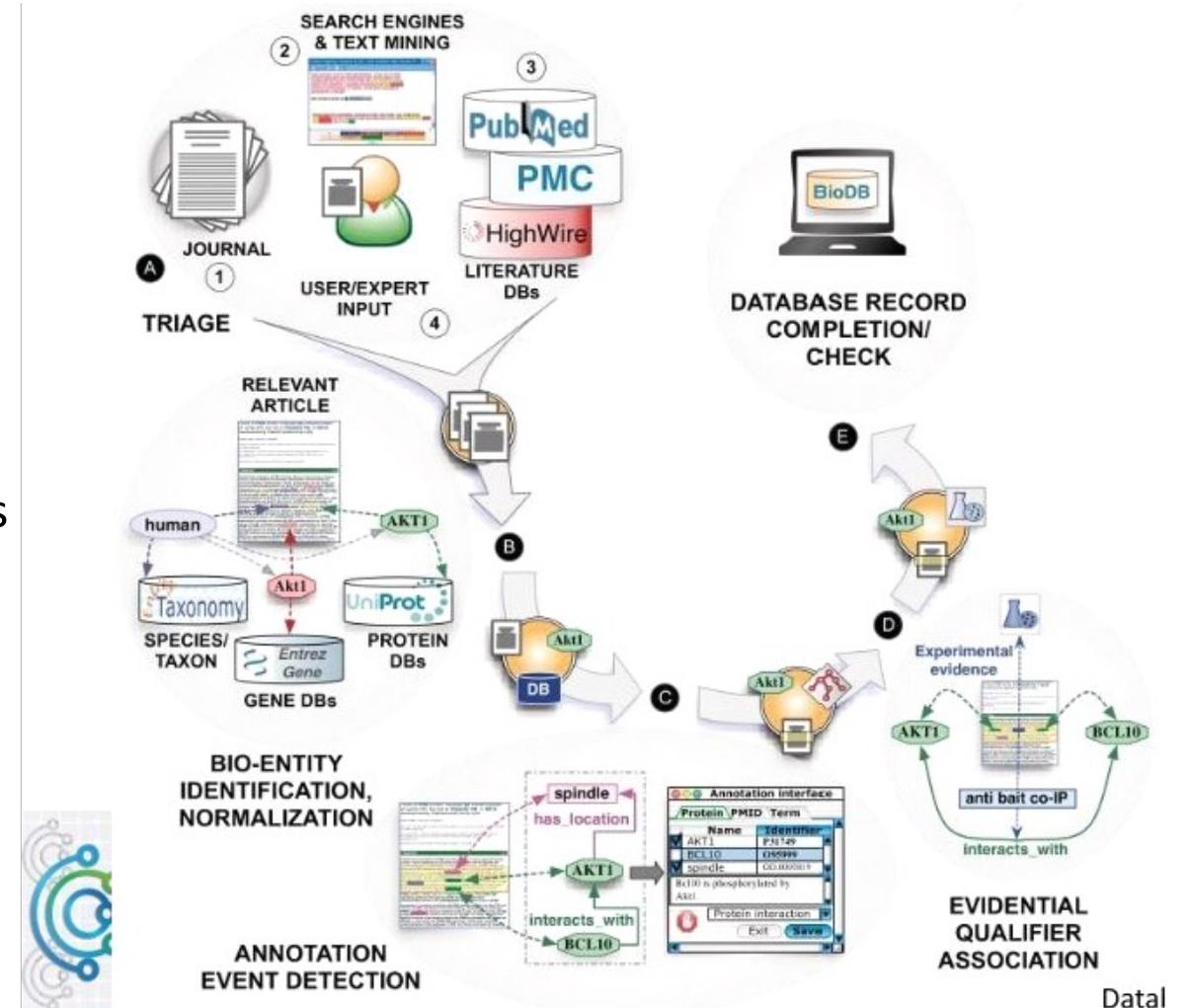
- Protein Homology

- Synthetic Genetic Array

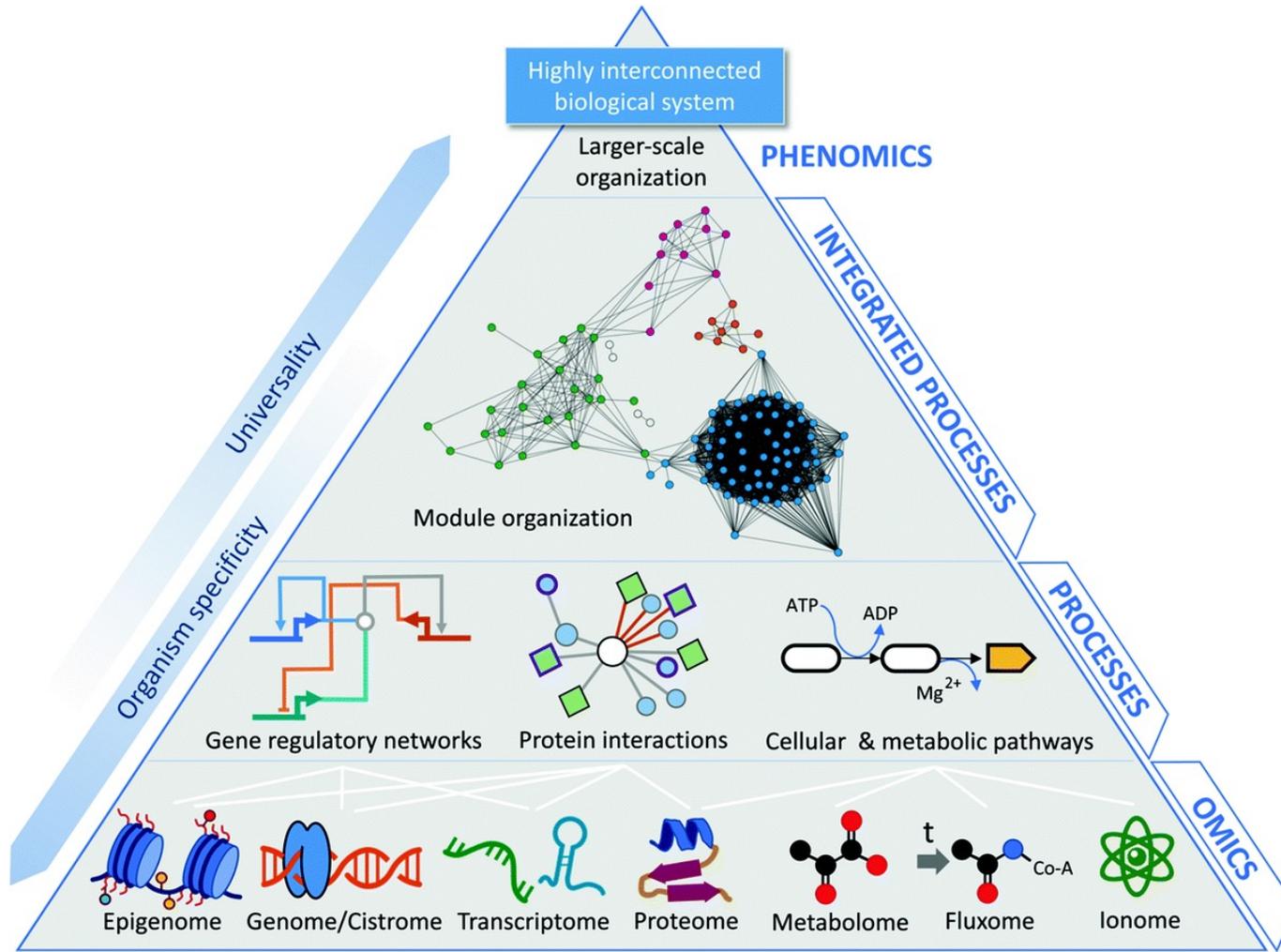
TheCellMap.org: A Web-Accessible Database for Visualizing and Mining the Global Yeast Genetic Interaction Network. Usaj et al., G3 (Bethesda) 2017

Reconstruction indirecte par Text-Mining

- Extraire les mots clés dans les publications et bases de données
- Nettoyer la liste de mots clés = Molécules, gènes, protéines
- Annoter ces éléments avec informations disponibles dans bases de données
- Chercher les co-occurrence dans les publications
- Mettre à jour les bases de données avec ces nouvelles informations



Reconstruction des réseaux biologiques



Il faut multiplier les types de reconstruction et combiner les différents réseaux obtenus

Il n'y a finalement qu'UN réseau global à reconstruire

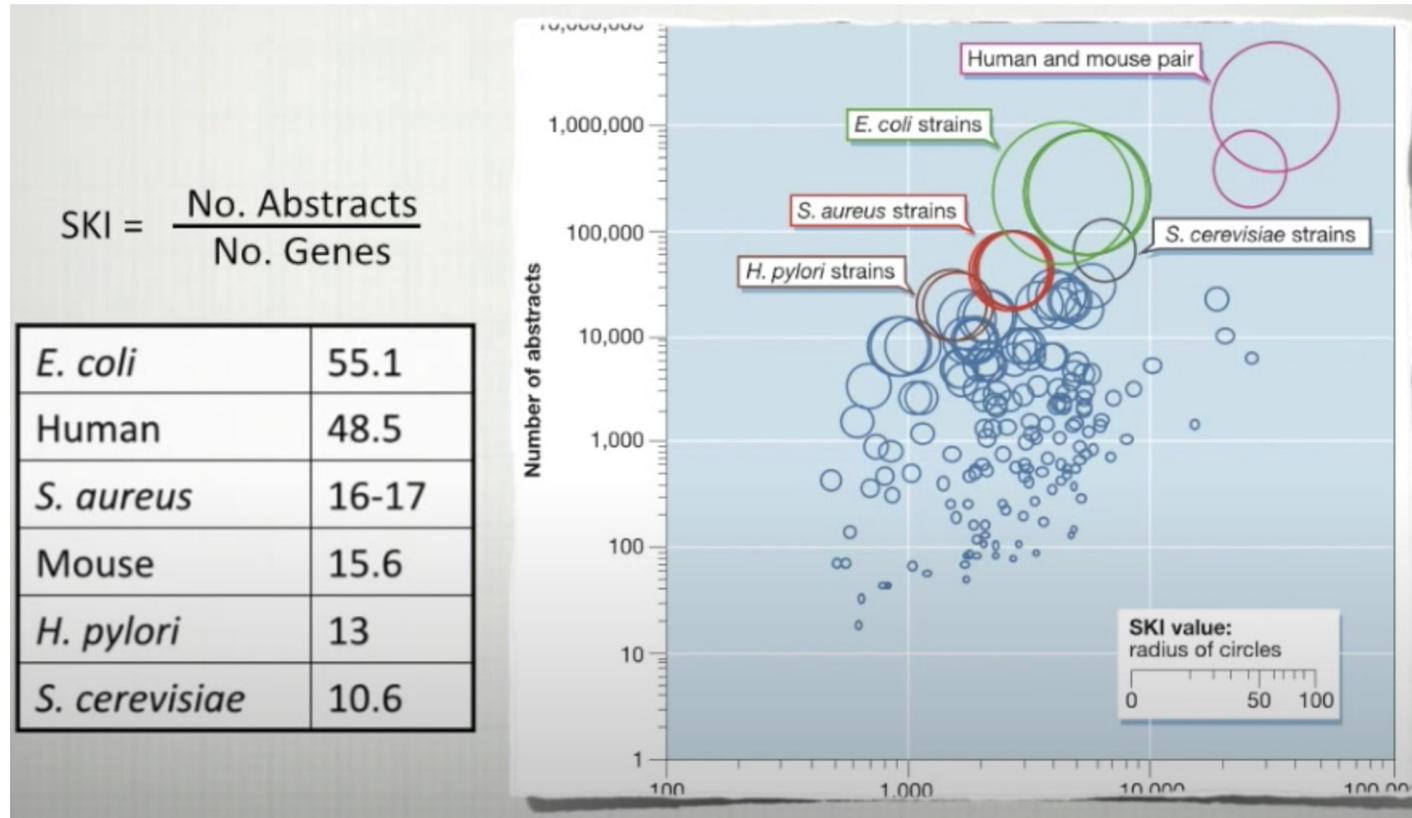
Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

- **Reconstruction de réseaux biologiques à l'aide de données omiques**

1. Quels sont les éléments à reconstruire ?
2. Quel type de réseaux peut-on reconstruire ?
3. Pour quel type de modélisation ?
4. Les différents types de reconstruction
5. Reconstruction directe
6. Reconstruction indirecte
7. **Les limites de la reconstruction de réseaux biologiques**



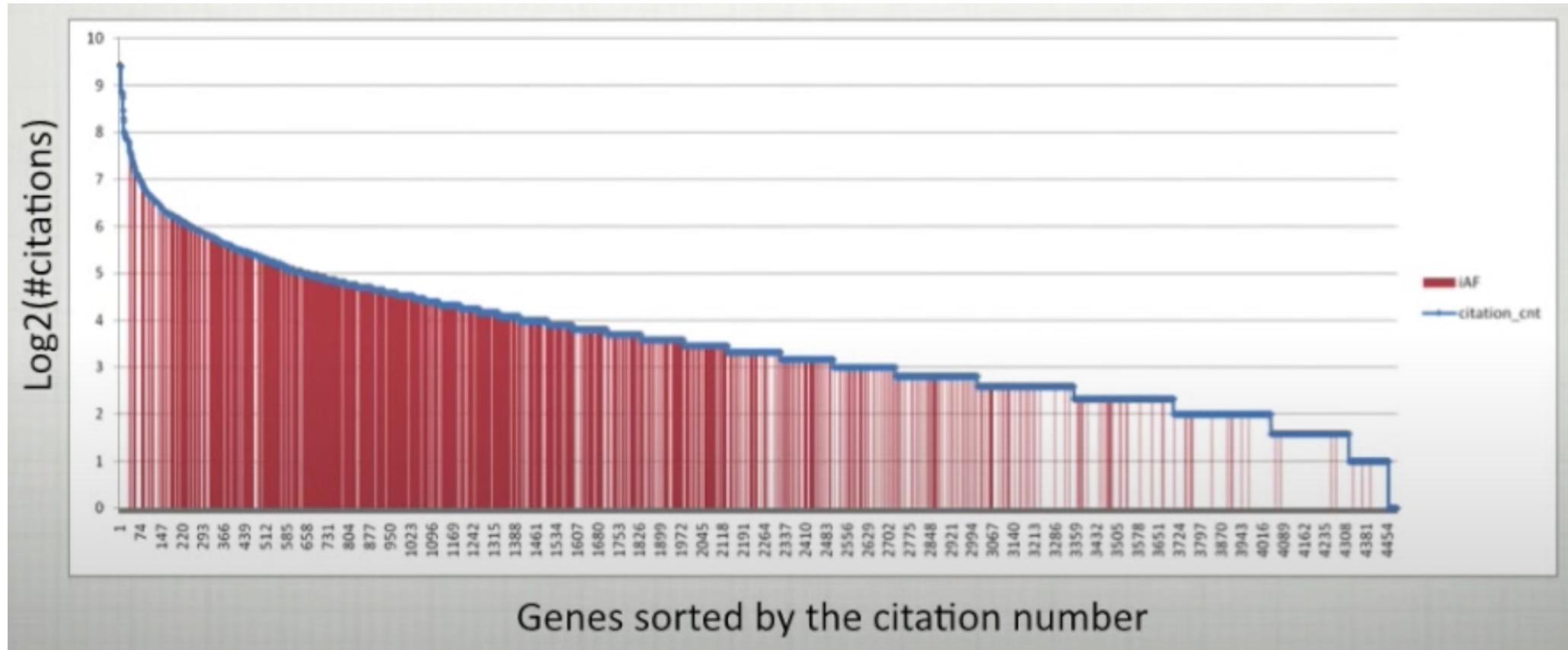
Les limites de la reconstruction de réseaux biologiques



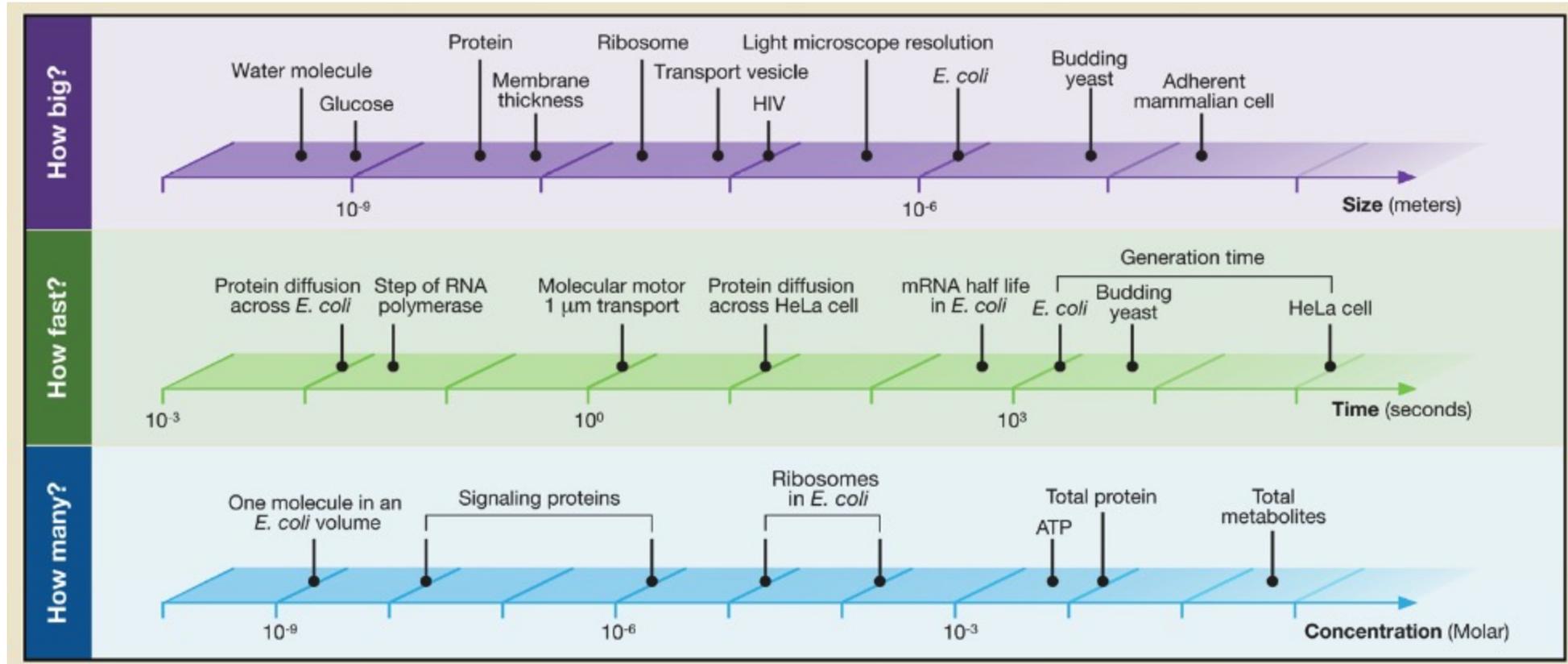
Lecture 3. Network Reconstruction: The Process
Bernhard Palsson

Biais de connaissance dans l'ensemble des gènes

Biais de connaissance dans l'ensemble des gènes



Les différentes échelles du vivant



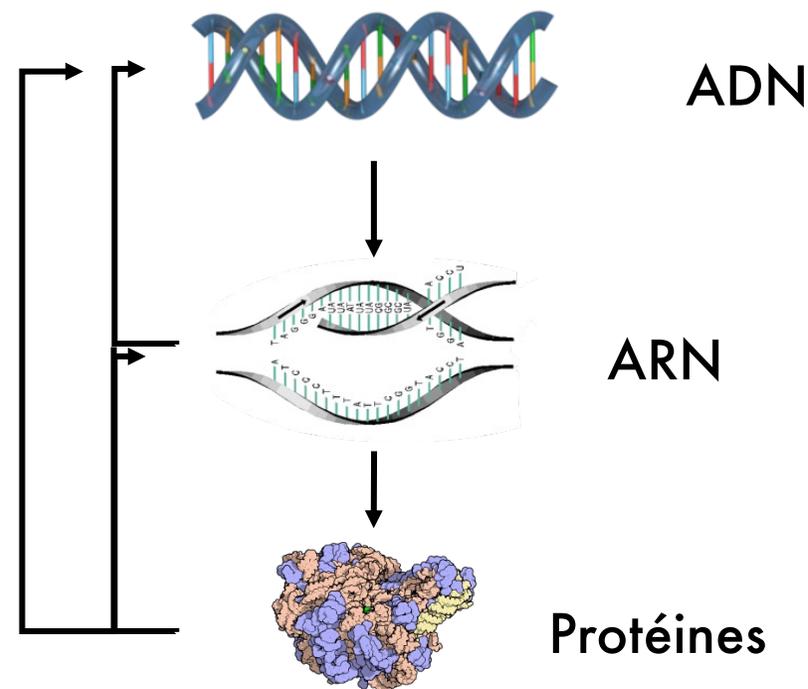
<https://bionumbers.hms.harvard.edu/>

Les différences d'échelles temporelles

Division, Replication, Transcription, Translation & Degradation Rates

at 37°C with a temperature dependence Q_{10} of $\approx 2-3$

9. Cell cycle time (exponential growth in rich media): *E. coli* $\approx 20-40$ min; yeast 70-140 min; human cell line (Hela): 15-30 hours
10. Rate of replication by DNA polymerase *E. coli* $\approx 200-1000$ bases/s; human ≈ 40 bases/s. Transcription by RNA polymerase 10-100 bases/s
11. Translation rate by ribosome 10-20 aa/s
12. Degradation rates (proliferating cells): mRNA half life $<$ cell cycle time; protein half life \geq cell cycle time



Transcriptomique vs Protéomique

Transcriptomique

- Seuil de détection bas
- Si la molécule est présente on doit la détecter
- On peut amplifier le signal facilement (PCR)

On mesure une grande majorité des ARNs présents dans l'échantillon

Protéomique

- Seuil de détection plus haut
- Certaines molécules ne vont pas être détectées
- Plus compliqué d'amplifier le signal

On mesure moins de la moitié des protéines présentes dans l'échantillon

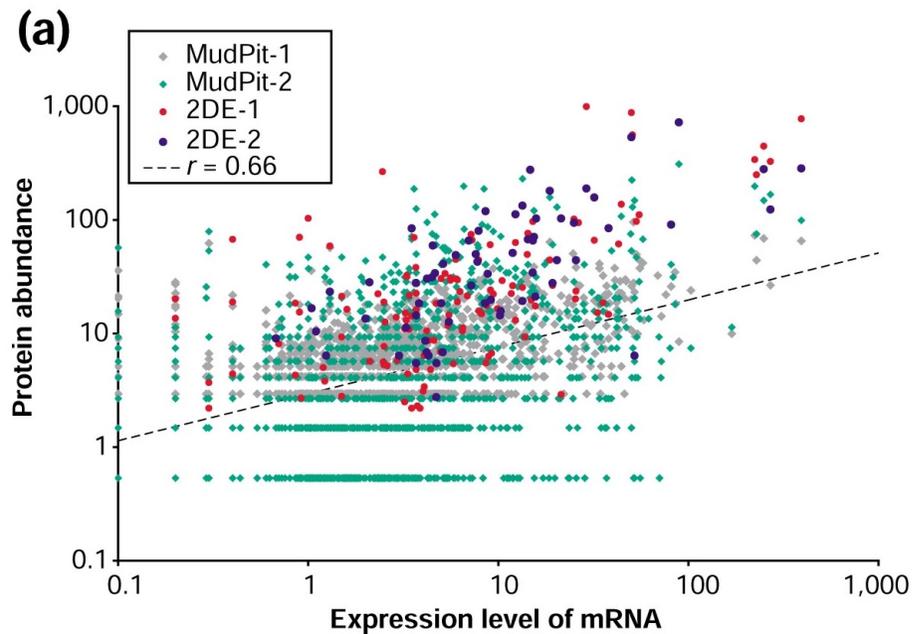
Comment corréler les échelles omiques dans ce contexte ?

Comparing protein abundance and mRNA expression levels on a genomic scale

[Dov Greenbaum](#), [Christopher Colangelo](#), [Kenneth Williams](#) ✉ & [Mark Gerstein](#) ✉

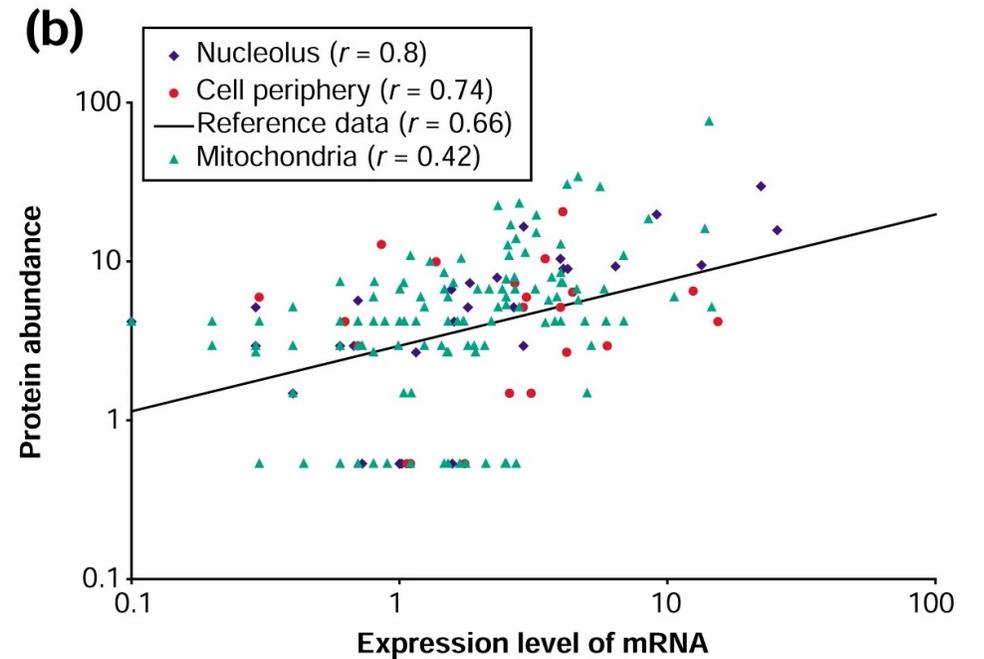
[Genome Biology](#) **4**, Article number: 117 (2003) | [Cite this article](#)

67k Accesses | **1144** Citations | **4** Altmetric | [Metrics](#)



$R = 0.66$

Il s'agit de la corrélation globale calculées sur toutes les protéines



$0.42 < R < 0.8$

Pour certains groupes GO la corrélation augmente

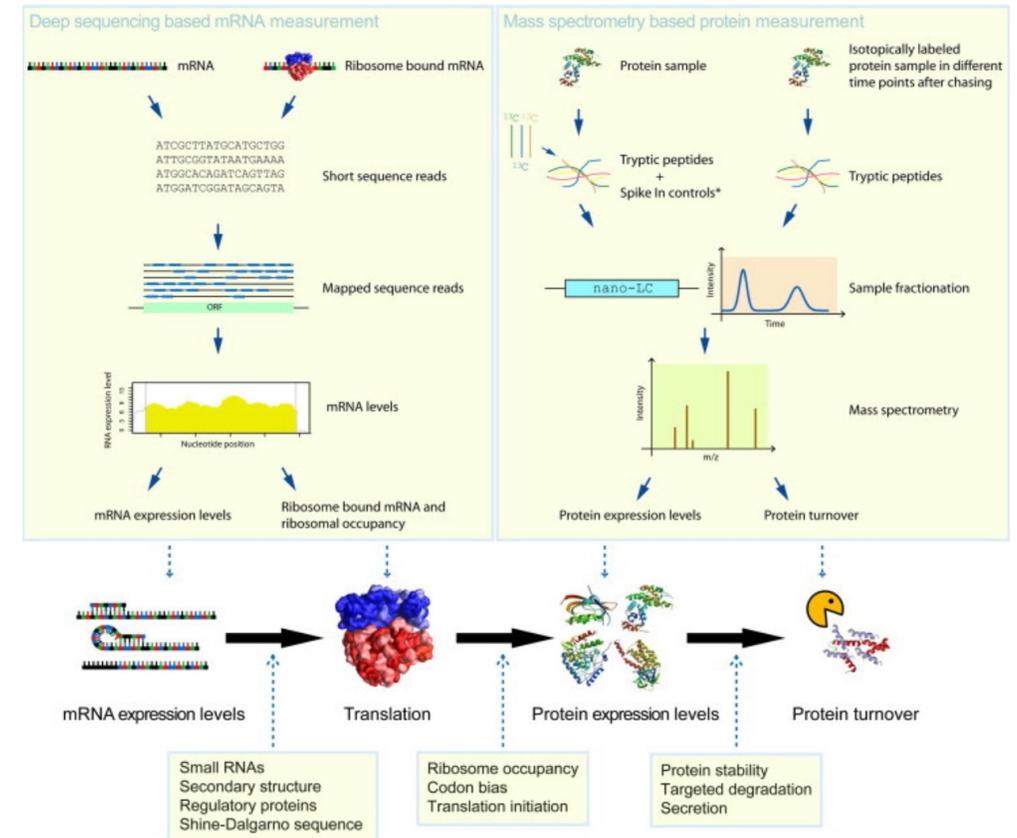
Review

Correlation of mRNA and protein in complex biological samples

Tobias Maier*, Marc Güell, Luis Serrano

Center for Genomic Regulation, Systems Biology Unit, Dr. Aiguader 88, 08003 Barcelona, Spain

Organism	r_p	r_s	Data-set size	Reference
<i>Saccharomyces cerevisiae</i>	0.36	n.d.	73	[40]
<i>Saccharomyces cerevisiae</i>	0.76	0.74	148	[39]
<i>Mus musculus</i>	0.59	n.d.	425	[46]
<i>Saccharomyces cerevisiae</i>	n.d.	0.45	678	[43]
<i>Desulfovibrio vulgaris</i>	0.50	n.d.	703	[45]
<i>Escherichia coli</i>	0.57	0.50	1103	[32]
<i>Schizosaccharomyces pombe</i>	0.58	0.61	1367	[44]
<i>Saccharomyces cerevisiae</i>	0.66	n.d.	2044	[41]
<i>Saccharomyces cerevisiae</i>	n.d.	0.57	4251	[38]



Comparative Analysis of Proteome and Transcriptome Variation in Mouse

2011

Anatole Ghazalpour^{1,*}, Brian Bennett¹, Vladislav A. Petyuk², Luz Orozco³, Raffi Hagopian¹, Imran N. Mungrue¹, Charles R. Farber⁴, Janet Sinsheimer³, Hyun M. Kang⁵, Nicholas Furlotte⁶, Christopher C. Park⁷, Ping-Zi Wen¹, Heather Brewer², Karl Weitz², David G. Camp II², Calvin Pan³, Roumyana Yordanova⁸, Isaac Neuhaus⁸, Charles Tilford⁸, Nathan Siemers⁸, Peter Gargalovic⁹, Eleazar Eskin⁶, Todd Kirchgessner⁹, Desmond J. Smith⁷, Richard D. Smith², Aldons J. Lusis^{1,3,10,11}

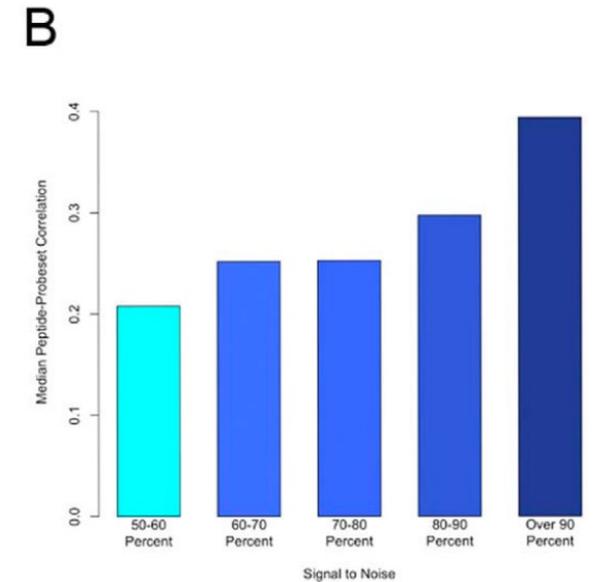
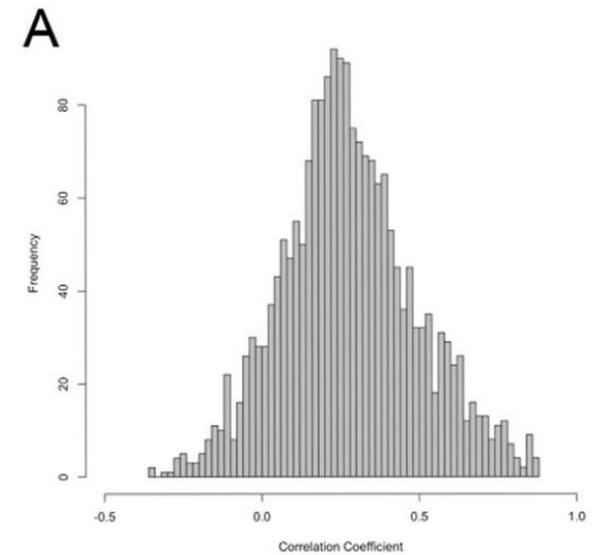
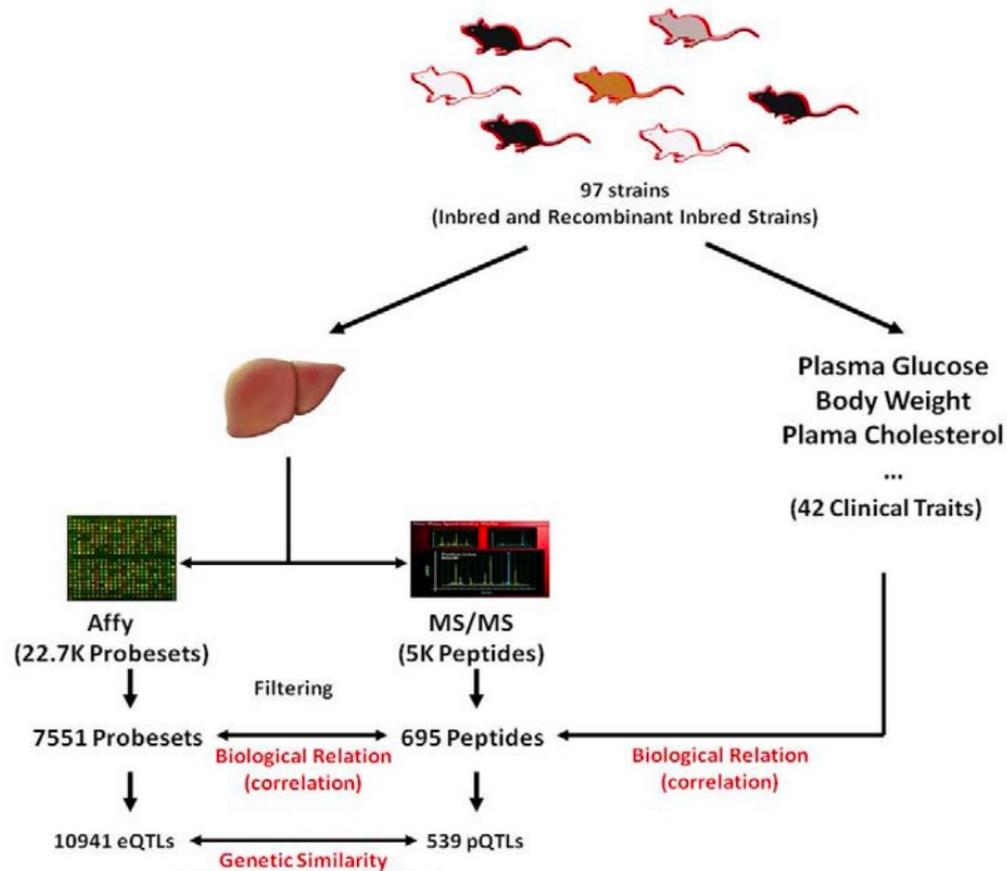
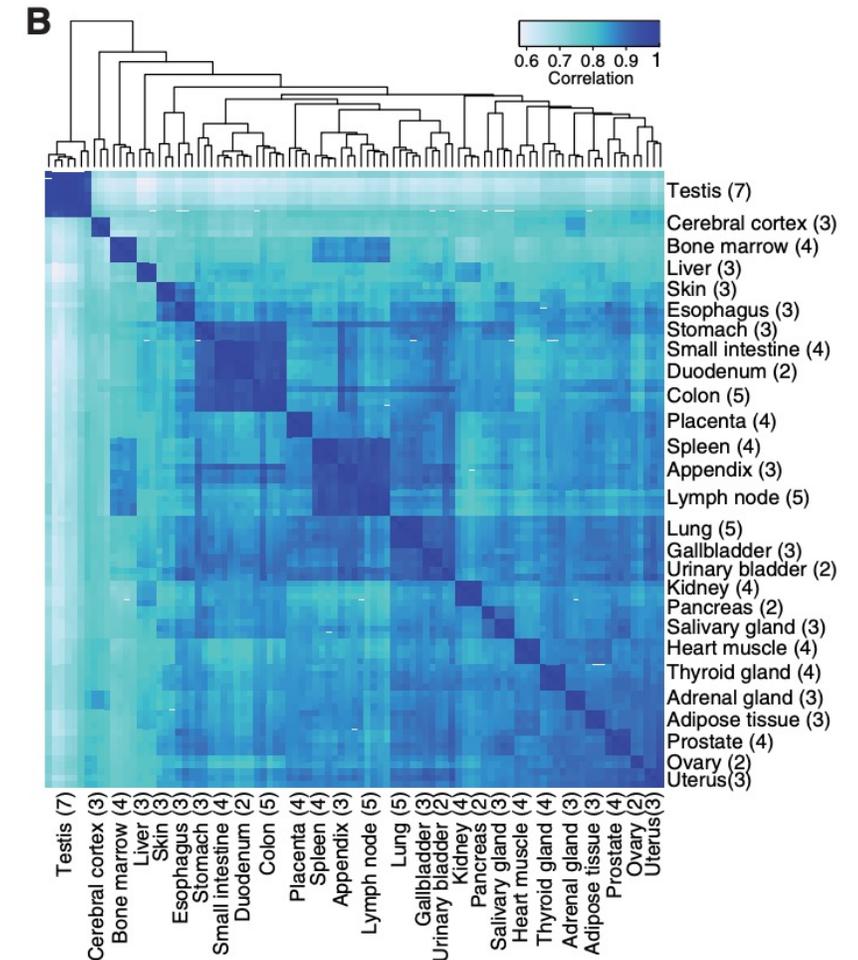


Figure 3. Relationships between protein levels and transcript levels. A) Histogram of correlation coefficients computed peptides and probesets representing the same gene. The median correlation coefficient is 0.27. B) Classification of correlations between probeset-peptides based on signal to noise ratio in the peptide data (larger signal to noise depicts less technical variation in the peptide measurement). doi:10.1371/journal.pgen.1001393.g003

Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics*

Linn Fagerberg†, Björn M. Hallström†, Per Oksvold†, Caroline Kampf§, Dijana Djureinovic§, Jacob Odeberg†, Masato Habuka†, Simin Tahmasebpoor§, Angelika Danielsson§, Karolina Edlund§, Anna Asplund§, Evelina Sjöstedt§, Emma Lundberg†, Cristina Al-Khalili Szigyarto†, Marie Skogs†, Jenny Ottosson Takanen||, Holger Berling||, Hanna Tegel||, Jan Mulder††, Peter Nilsson†, Jochen M. Schwenk†, Cecilia Lindskog§, Frida Danielsson†, Adil Mardinoglu¶, Åsa Sivertsson†, Kalle von Feilitzen||, Mattias Forsberg†, Martin Zwahlen†, IngMarie Olsson§, Sanjay Navani**, Mikael Huss†, Jens Nielsen†¶, Fredrik Ponten§, and Mathias Uhlén†¶¶¶



As expected, the RNA expression levels as measured by RNA-Seq and the protein levels detected by staining with immunohistochemistry display low correlation for many genes. This is not surprising, because immunohistochemistry based on enzymatic amplification technology is not quantitative (27) and in addition often yields off-target binding to unrelated proteins.

Méthodes d'intégration de données multi-omiques

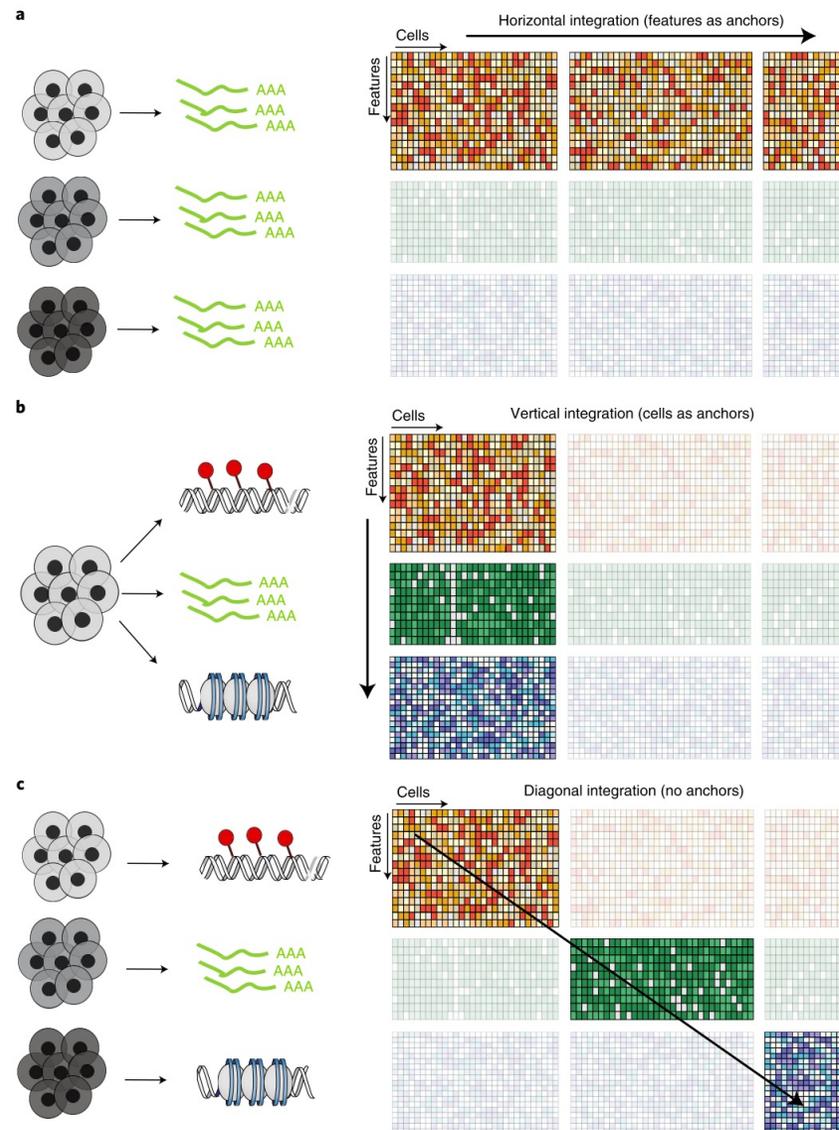
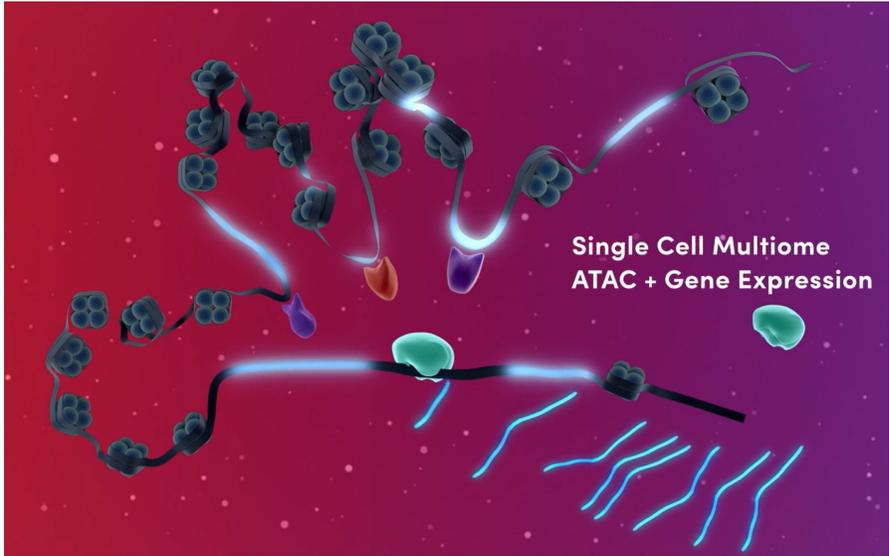


Table 1 | Overview of common data integration methods classified according to their anchor choice

Integration task	Method	Ref.
Vertical (global)	CCA	112
Vertical (global)	JIVE	70
Vertical (global)	PLS	71
Vertical (global)	MCIA	113
Vertical (global)	MOFA+	65
Vertical (global)	scAI	114
Vertical (global)	iNMF	38
Vertical (global)	Seurat v4	11
Vertical (local)	Spearman's rank correlation coefficient	50
Vertical (local)	LMM	51
Horizontal	MNN	21
Horizontal	Seurat v3	22
Horizontal	LIGER	23
Horizontal	Harmony	24
Horizontal	Scanorama	29
Horizontal	BBKNN	25
Horizontal	scVI	26
Horizontal	scmap	28
Horizontal	conos	27
Diagonal	MATCHER	77
Diagonal	MMD-MMA	78
Diagonal	SCIM	115
Diagonal	UnionCom	116
Diagonal	coupledNMF	117

Mesure simultanée sur cellule unique



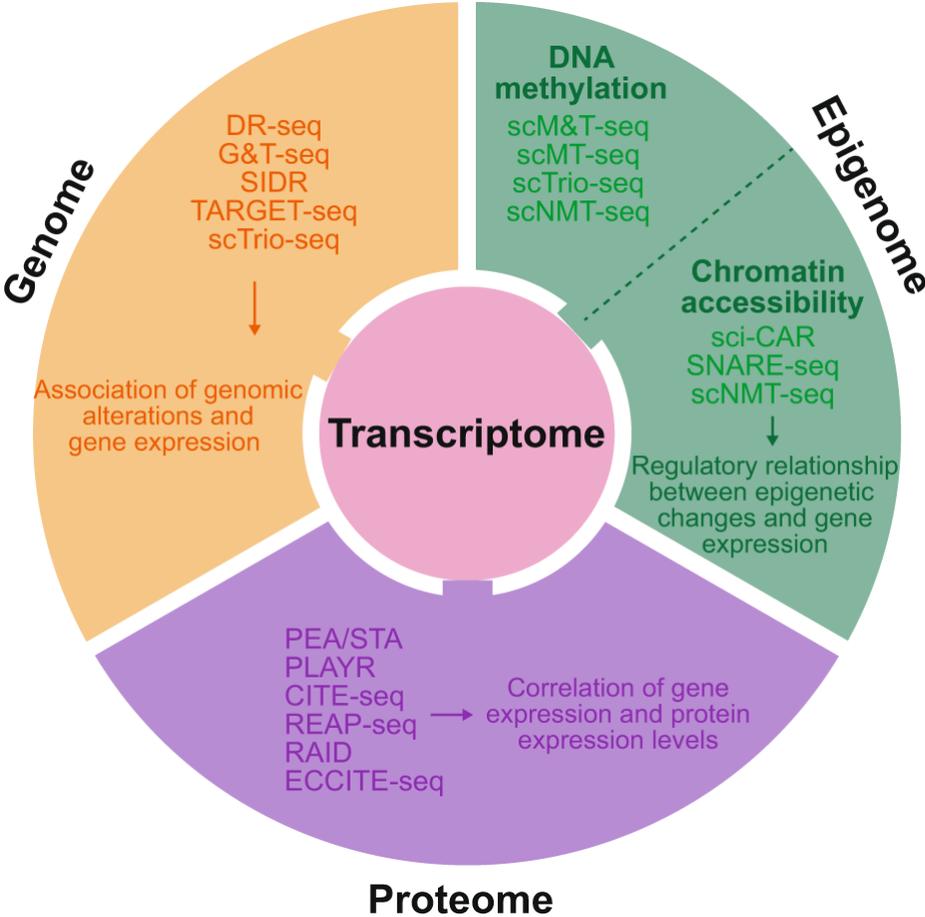
Single Cell Multiome GEX library:



Single Cell Multiome ATAC library:



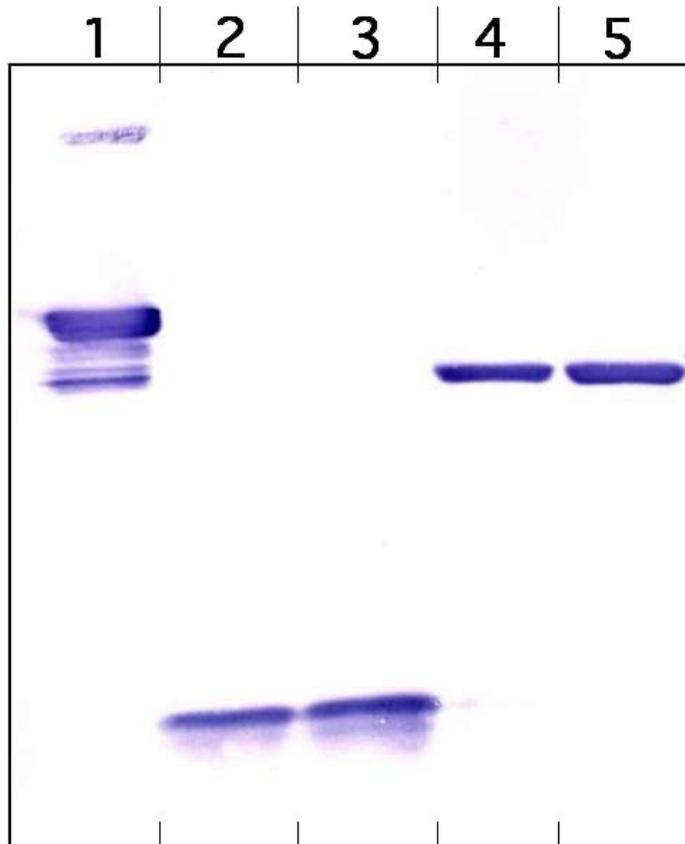
10X Multiome technology



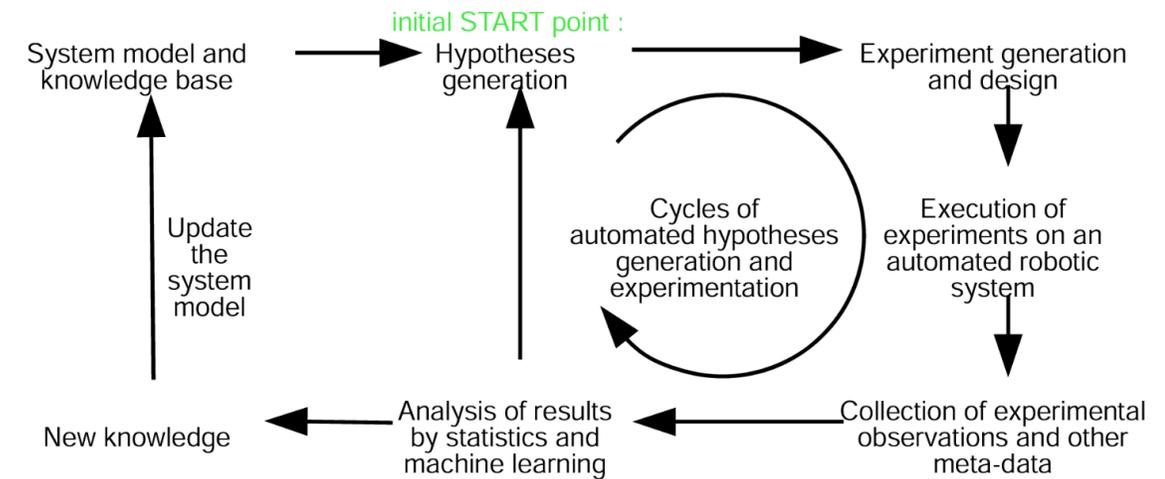
Single-cell multiomics: technologies and data analysis methods
Exp. Mol. Medecine, 2020 (52)1428-1442

Validation des réseaux biologiques

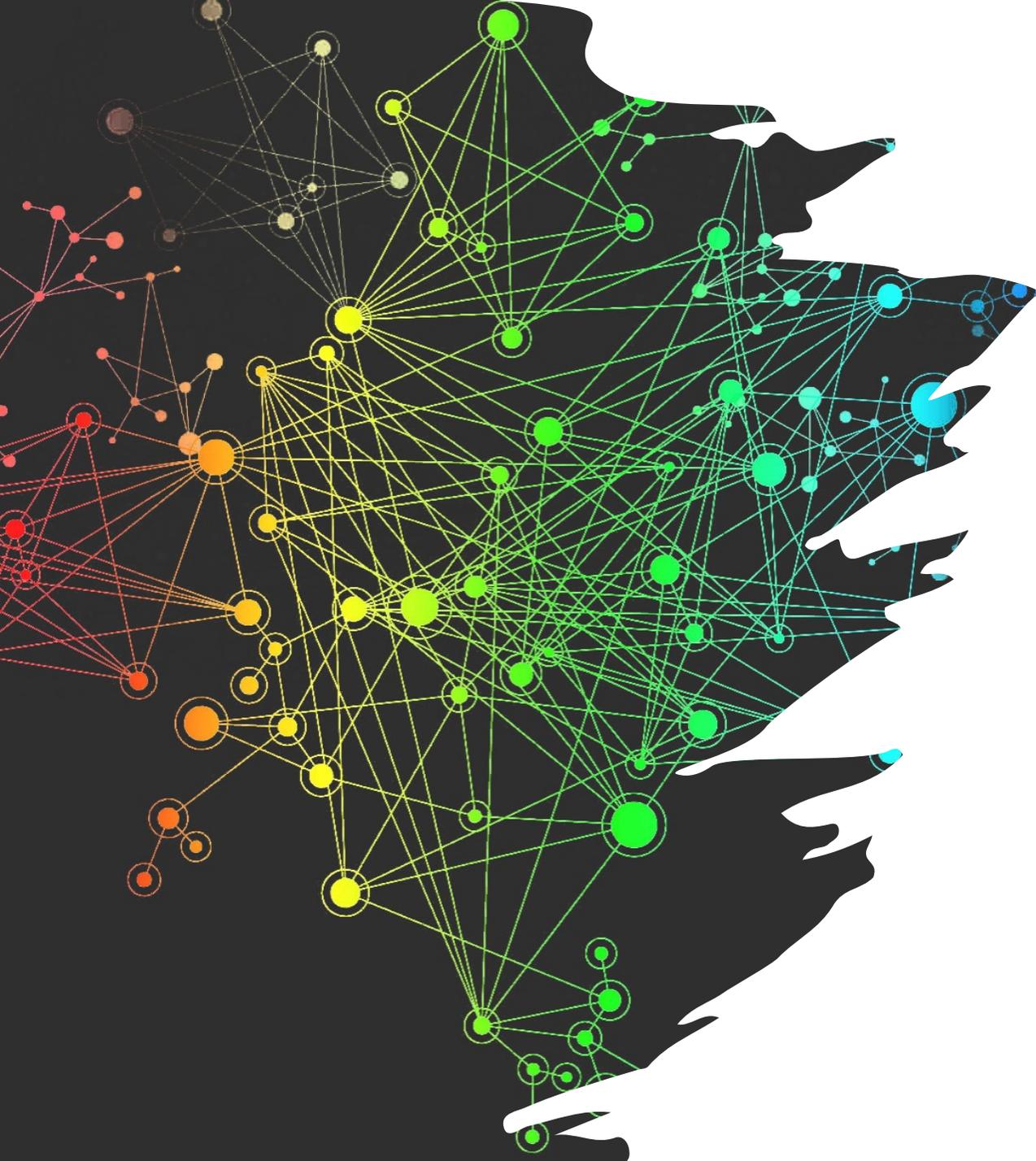
Validation expérimentale



Validation automatisée



Ross D King, Science 2009



Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

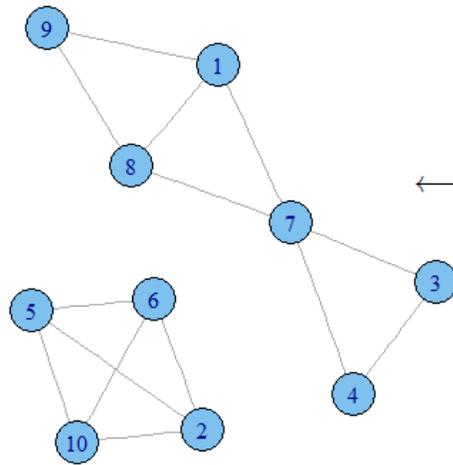
- Reconstruction de réseaux biologiques à l'aide de données omiques
- **Reconstruire par analyse de co-expression (TP)**
- Etat de l'art des réseaux biologiques les plus étendus
- Apport du deep-learning à la biologie des systèmes

Réseau de co-expression

	S_1	S_2	S_3		G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	43.26	40.89	5.05	$\xrightarrow{ r(G_i, G_j) }$ Pearson correlation	1.00	0.23	0.61	0.71	0.03	0.35	0.86	1.00	0.97	0.37
G_2	166.6	41.87	136.65		0.23	1.00	0.63	0.52	0.98	0.99	0.29	0.30	0.46	0.99
G_3	12.53	39.55	42.09		0.61	0.63	1.00	0.99	0.77	0.53	0.93	0.56	0.41	0.51
G_4	28.77	191.92	236.56		0.71	0.52	0.99	1.00	0.69	0.41	0.97	0.66	0.52	0.40
G_5	114.7	79.7	99.76		0.03	0.98	0.77	0.69	1.00	0.95	0.48	0.09	0.27	0.94
G_6	119.1	80.57	114.59		0.35	0.99	0.53	0.41	0.95	1.00	0.17	0.41	0.57	1.00
G_7	118.9	156.69	186.95		0.86	0.29	0.93	0.97	0.48	0.17	1.00	0.83	0.72	0.16
G_8	3.76	2.48	136.78		1.00	0.30	0.56	0.66	0.09	0.41	0.83	1.00	0.98	0.42
G_9	32.73	11.99	118.8		0.97	0.46	0.41	0.52	0.27	0.57	0.72	0.98	1.00	0.58
G_{10}	17.46	56.11	21.41		0.37	0.99	0.51	0.40	0.94	1.00	0.16	0.42	0.58	1.00

Gene expression values

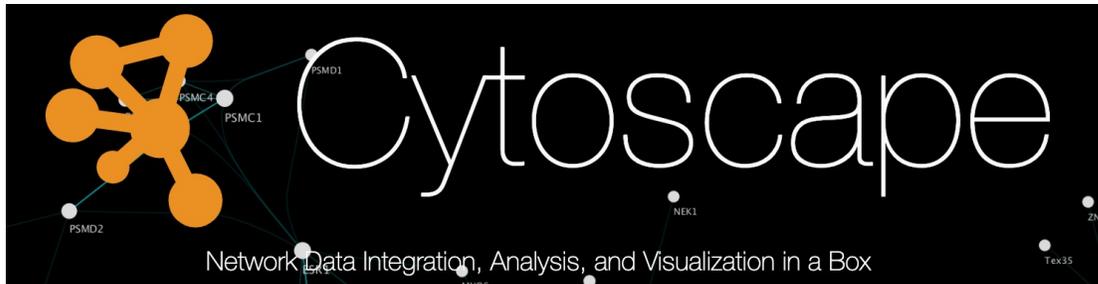
Similarity (Co-expression) score



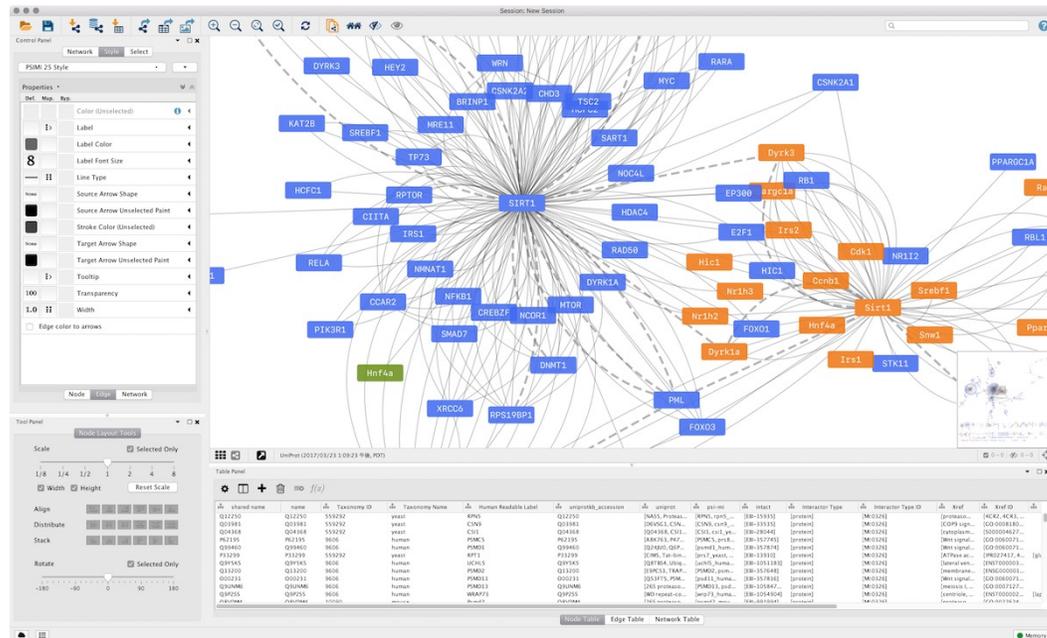
	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}	
G_1	0	0	0	0	0	0	1	1	1	0	$\xleftarrow{ r(G_i, G_j) \geq 0.8}$ Significance threshold
G_2	0	0	0	0	1	1	0	0	0	1	
G_3	0	0	0	1	0	0	1	0	0	0	
G_4	0	0	1	0	0	0	1	0	0	0	
G_5	0	1	0	0	0	1	0	0	0	1	
G_6	0	1	0	0	1	0	0	0	0	1	
G_7	1	0	1	1	0	0	0	1	0	0	
G_8	1	0	0	0	0	0	1	0	1	0	
G_9	1	0	0	0	0	0	0	1	0	0	
G_{10}	0	1	0	0	1	1	0	0	0	0	

Network adjacency matrix

TP : Co-expression dans un transcriptome bactérien



Nous avons mesuré dans différentes conditions biologiques l'expression des 2855 gènes codants et 142 ARN non-codants de la bactérie *Listeria monocytogenes* EGD-e

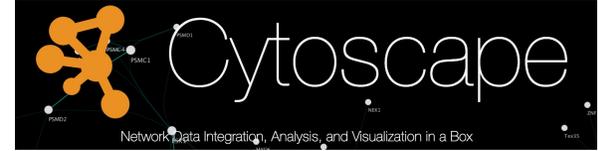


Le tableau d'expression se trouve ici : github.com/becavin-lab/BioRegul/blob/main/TP-Co-expression/Listeria-Expr-full-TP.txt

On va reconstruire un réseau de co-expression à l'aide du logiciel **Cytoscape**

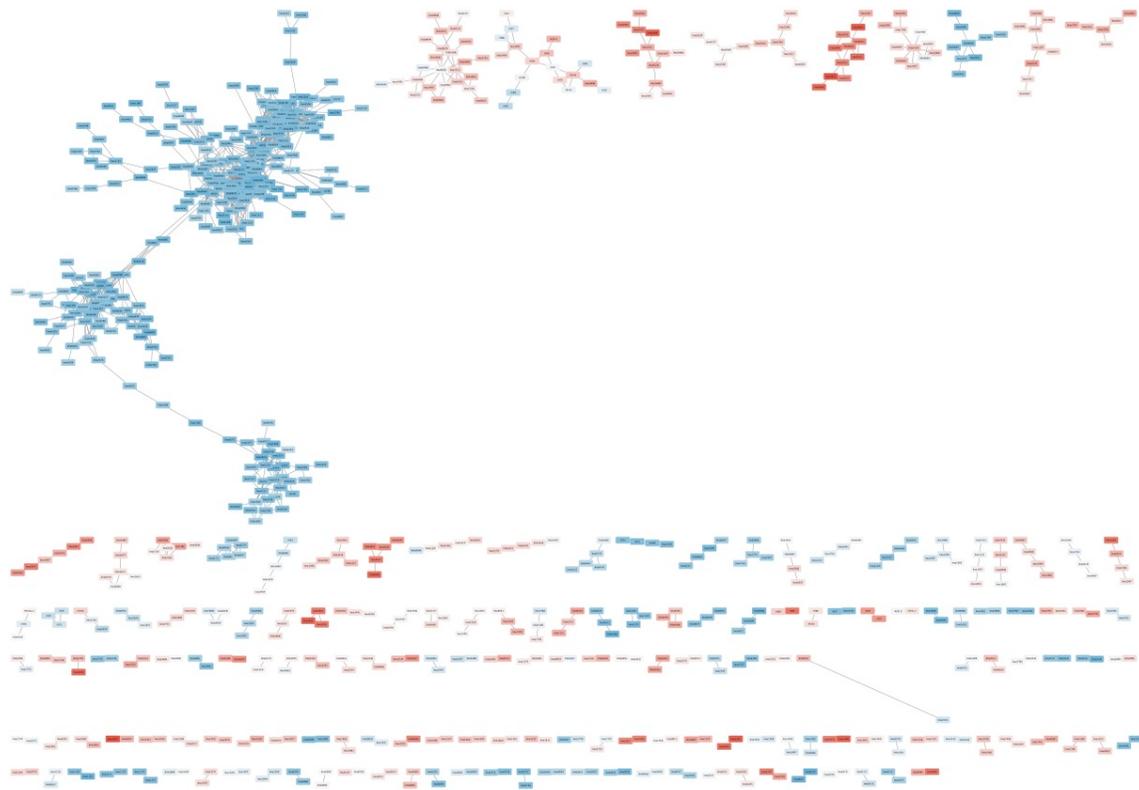
<https://cytoscape.org/download.html>

TP : Co-expression dans un transcriptome bactérien

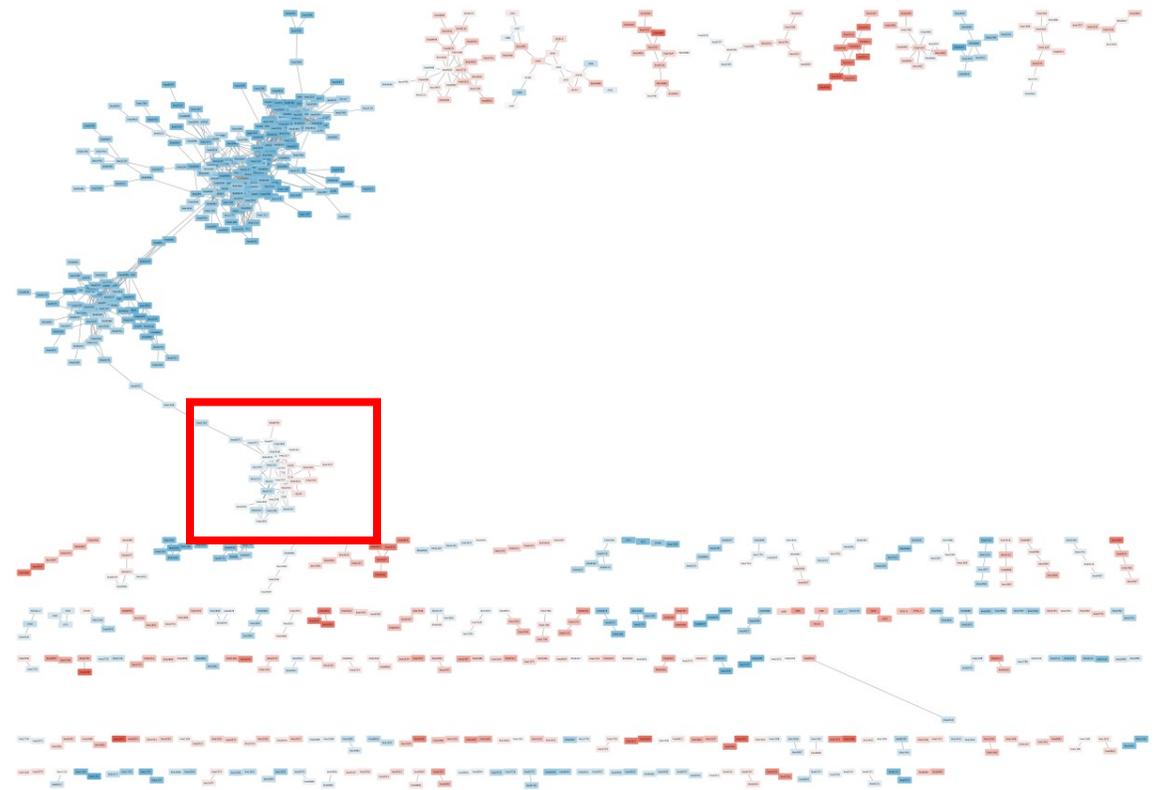


- Télécharger l'APP Cytoscape **ExpressionCorrelation**
Apps > App Store > Show App Store
- Charger le table **Listeria-Expr-full-TP.txt** dans Cytoscape
Import Table from File > To unassigned table
- Créer les réseaux de co-expression
Apps > ExpressionCorrelation > ConstructCorrelationNetworks
- Visualiser les réseaux en choisissant un **layout dépendent du parameter "strength"**
Layout > prefuse Force Directed Layout
- Tracer la distribution des degrés. Réseau sans échelles ?

Listeria growth in Macrophages



BHI 2014



Macrophages 2014

Listeria growth in Macrophages

Cell



Volume 150, Issue 4, 17 August 2012, Pages 792-802

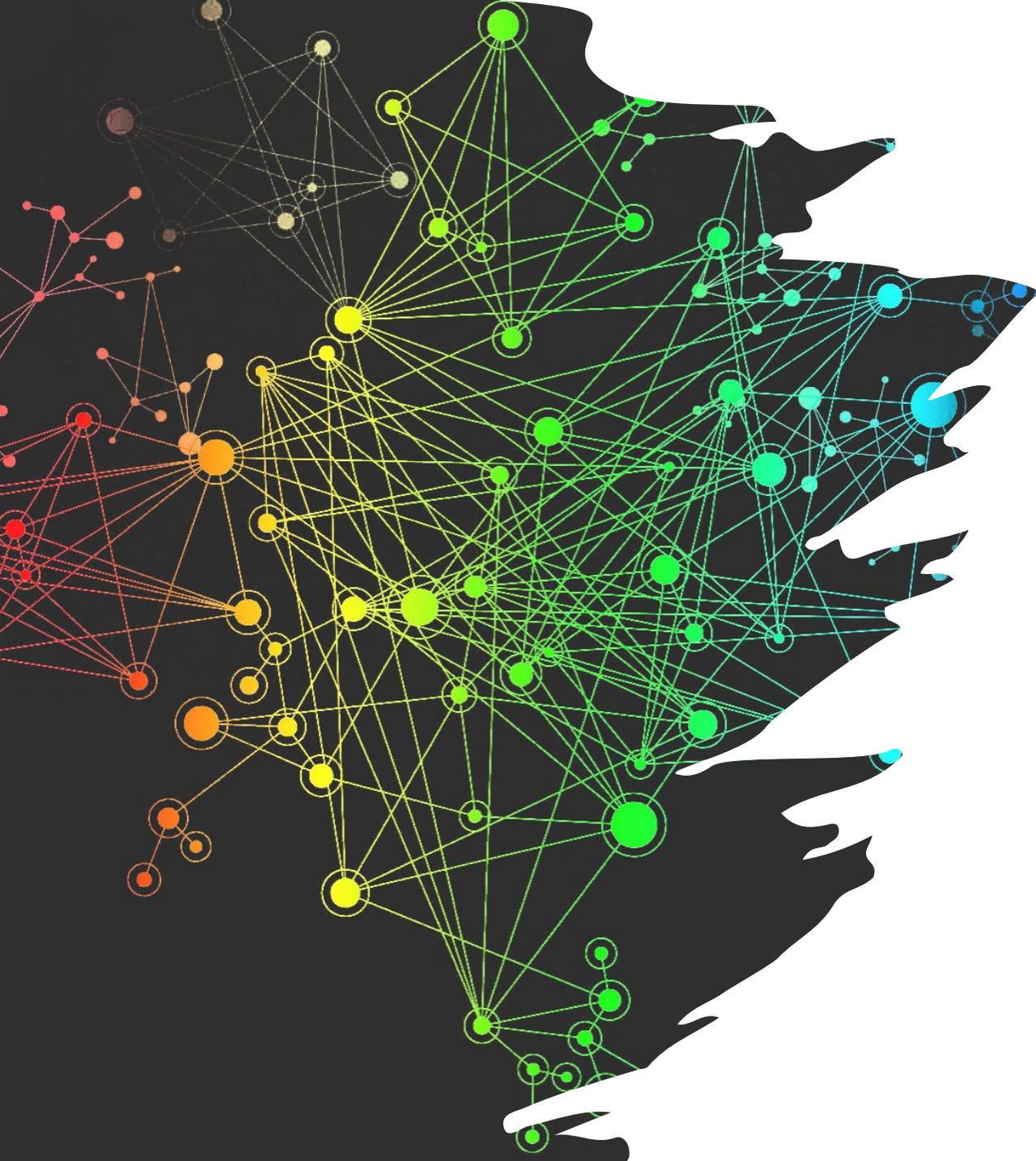
Article

Prophage Excision Activates *Listeria* Competence Genes that Promote Phagosomal Escape and Virulence

Lev Rabinovich¹, Nadejda Sigal¹, Ilya Borovok¹, Ran Nir-Paz², Anat A. Herskovits¹  



Macrophages 2014



Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

- Reconstruction de réseaux biologiques à l'aide de données omiques
- Reconstruire par analyse de co-expression (TP)
- **Etat de l'art des réseaux biologiques les plus étendus (full organism ?)**
- Apport du deep-learning à la biologie des systèmes

LacZ operon version 2023

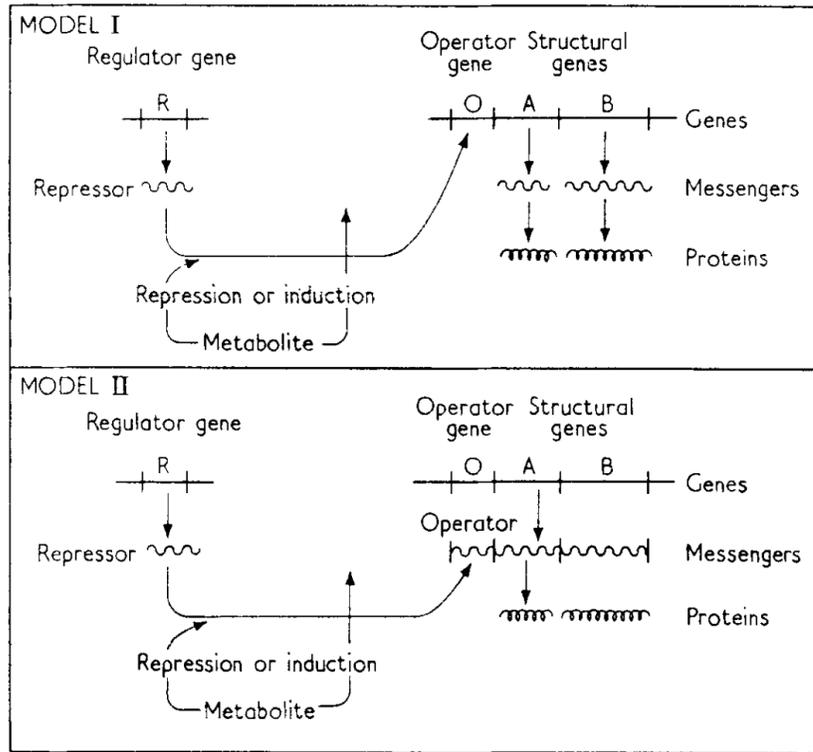
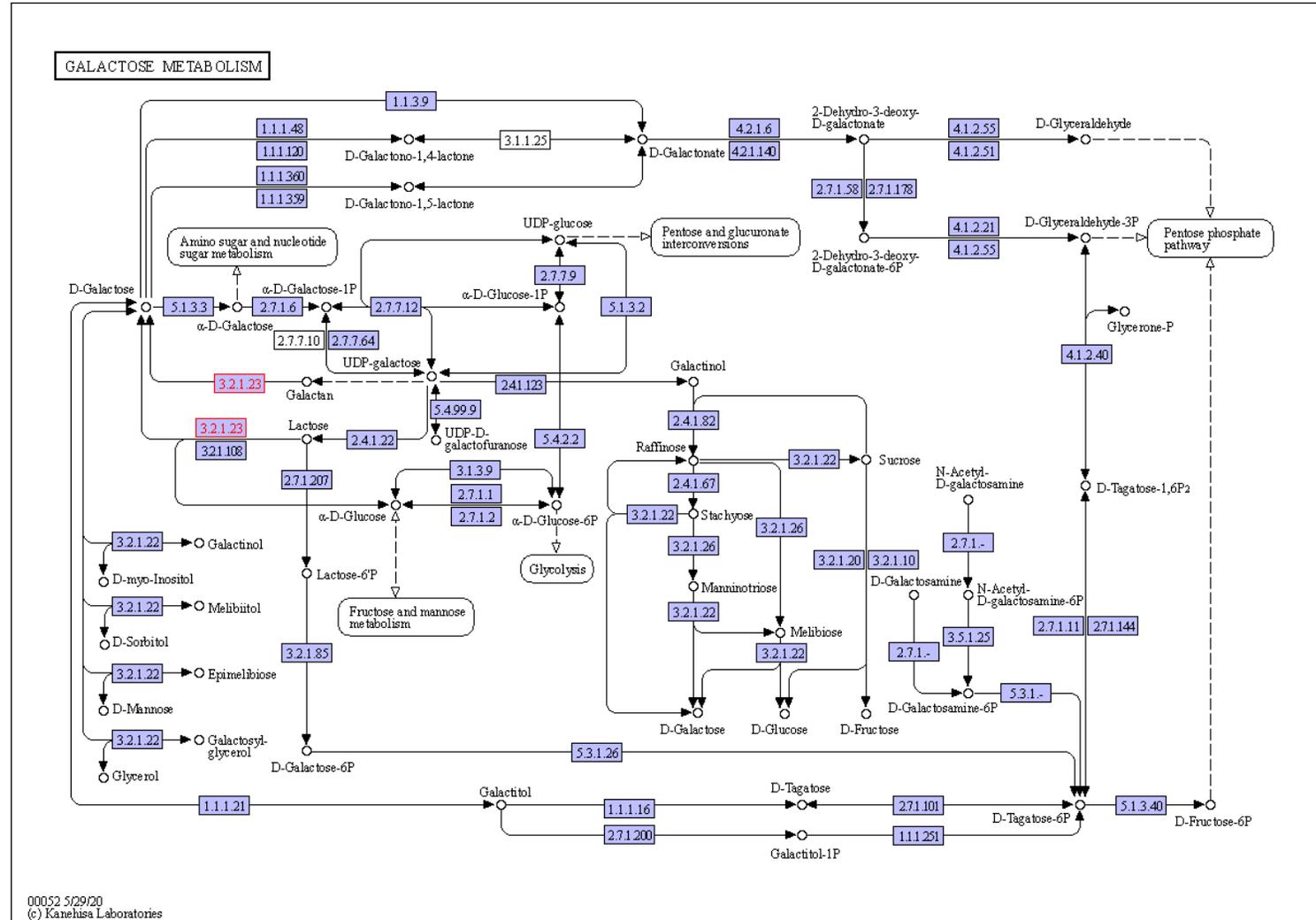


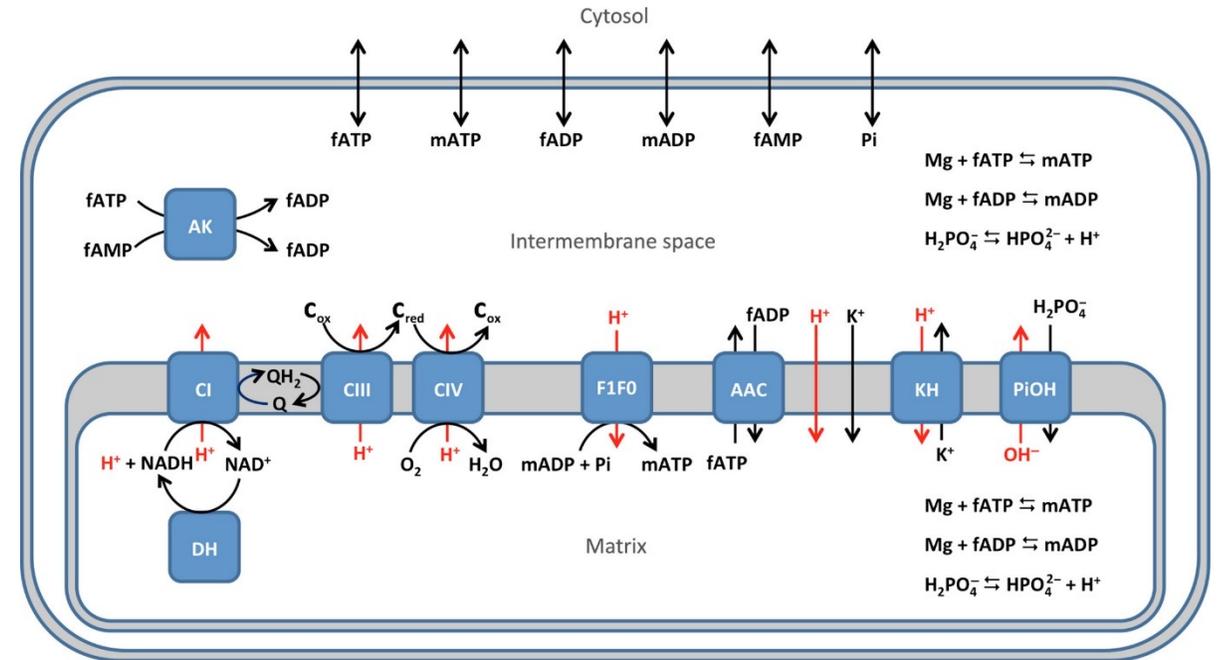
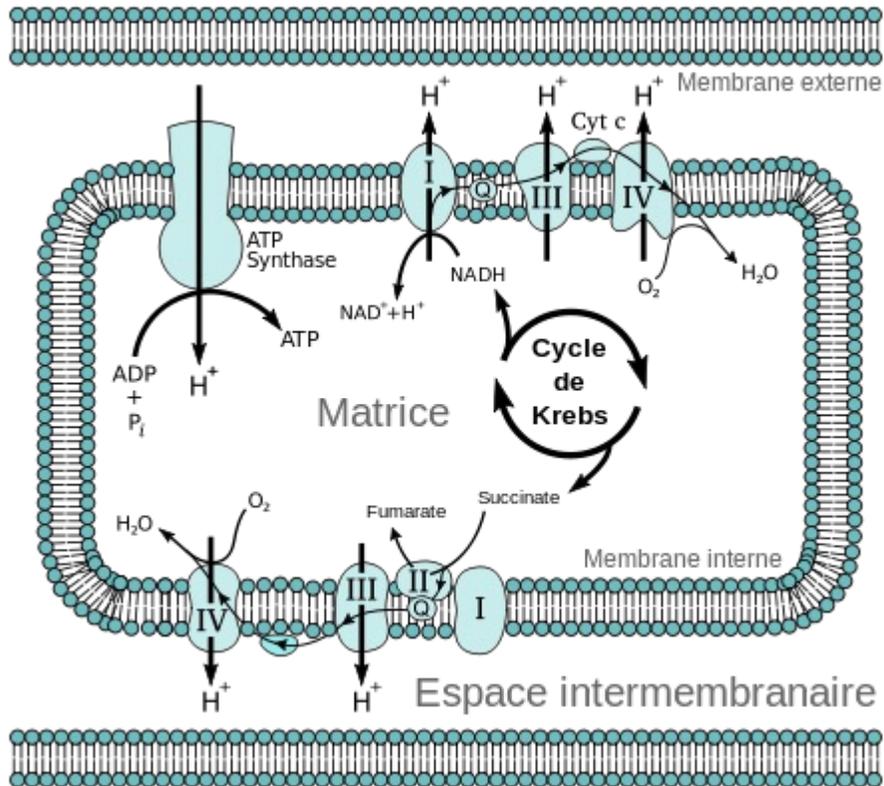
FIG. 6. Models of the regulation of protein synthesis.

Genetic Regulatory Mechanisms in the Synthesis of Proteins
 F. Jacob, J. Monod, J. Mol. Biol. 1961



https://www.genome.jp/kegg-bin/show_pathway?ko00052+K01190

Le cycle de Krebs en 2023 : Phosphorylation oxydative



Phosphorylation de l'ADP en ATP
dans les mitochondries
(wikipedia)

Heiske et al., FEBS Journal 2017

Le cycle de Krebs en 2023 : Phosphorylation oxydative

Matrix		
DH	Dehydrogenase reaction	$NAD_x + H_x \rightleftharpoons NADH_x$
MgATP _x	Mg ²⁺ binding on fATP	$fATP_x + Mg_x \rightleftharpoons mATP_x$
MgADP _x	Mg ²⁺ binding on fADP	$fADP_x + Mg_x \rightleftharpoons mADP_x$
Pdiss _x	Pi dissociation	$H_2PO_{4x} \rightleftharpoons HPO_{4x} + H_x$
Inner membrane		
C1	Complex I	$NADH_x + Q_m + 5H \rightleftharpoons NAD_x + QH_{2im} + 4H_{is}$
C3	Complex III	$QH_{2im} + 2Co_{is} + 2H_x \rightleftharpoons Q_m + 2Cred_{is} + 4H_{is}$
C4	Complex IV	$Cred_{is} + 0.25 O_2 + 2H_x \rightleftharpoons Cox_{is} + 0.5H_2O + H_{is}$
F1F0	ATP synthase	$mADP_x + P_i + n_A H_{is} + H_x \rightleftharpoons mATP_x + n_A H_x$
AAC	ADP/ATP carrier	$fATP_x + fADP_{is} \rightleftharpoons fATP_{is} + fADP_x$
PiOH	Pi/OH antiporter	$H_2PO_{4is} + OH_x \rightleftharpoons H_2PO_{4x} + OH_{is}$
KH	K ⁺ /H ⁺ antiporter	$K_{is} + H_x \rightleftharpoons K_x + H_{is}$
Hleak	H ⁺ leak	$H_{is} \rightleftharpoons H_x$
Kleak	K ⁺ leak	$K_{is} \rightleftharpoons K_x$
Intermembrane space		
AK	Adenylate kinase	$mATP_{is} + fAMP_{is} \rightleftharpoons mADP_{is} + fADP_{is}$
MgATP _{is}	Mg ²⁺ binding on fATP	$fATP_{is} + Mg_{is} \rightleftharpoons mATP_{is}$
MgADP _{is}	Mg ²⁺ binding on fADP	$fADP_{is} + Mg_{is} \rightleftharpoons mADP_{is}$
Pdiss _{is}	Pi dissociation	$H_2PO_{4is} \rightleftharpoons HPO_{4is} + H_{is}$
Outer membrane		
fATP _{om}	fATP diffusion	$fATP_e \rightleftharpoons fATP_{is}$
mATP _{om}	mATP diffusion	$mATP_e \rightleftharpoons mATP_{is}$
fADP _{om}	fADP diffusion	$fADP_e \rightleftharpoons fADP_{is}$
mADP _{om}	mADP diffusion	$mADP_e \rightleftharpoons mADP_{is}$
fAMP _{om}	fAMP diffusion	$fAMP_e \rightleftharpoons fAMP_{is}$
Pi _{om}	Pi diffusion	$P_{ie} \rightleftharpoons P_{is}$
Mg _{om} *	Mg ²⁺ diffusion	$Mg_e \rightleftharpoons Mg_{is}$
External space/cytosol		
MgATP _e *	Mg ²⁺ binding on fATP	$fATP_e + Mg_e \rightleftharpoons mATP_e$
MgADP _e *	Mg ²⁺ binding on fADP	$fADP_e + Mg_e \rightleftharpoons mADP_e$

$$\frac{d[H]_x}{dt} = x_{out} \cdot \left(\frac{+v_{DH} - 5v_{C1} - 2v_{C3} - 4v_{C4}}{+(n_A - 1)v_{F1F0} + 2v_{PiOH} + v_{leak} - v_{KH}} \right) / W_x \quad (35)$$

$$\frac{d[K]_x}{dt} = (+v_{KH} + v_K) / W_x \quad (36)$$

$$\frac{d[Mg]_x}{dt} = -v_{MgATP_x} - v_{MgADP_x} \quad (37)$$

$$\frac{d[NADH]_x}{dt} = (+v_{DH} - v_{C1}) / W_x \quad (38)$$

$$\frac{d[fATP]_x}{dt} = -v_{AAC} / W_x - v_{MgATP_x} \quad (39)$$

$$\frac{d[mATP]_x}{dt} = +v_{F1F0} / W_x + v_{MgATP_x} \quad (40)$$

$$\frac{d[fADP]_x}{dt} = +v_{AAC} / W_x - v_{MgADP_x} \quad (41)$$

$$\frac{d[mADP]_x}{dt} = -v_{F1F0} / W_x + v_{MgADP_x} \quad (42)$$

$$\frac{d[P_i]_x}{dt} = (-v_{F1F0} + v_{PiH}) / W_x \quad (43)$$

$$\frac{d[QH_2]_{im}}{dt} = (+v_{C1} - v_{C3}) / W_{im} \quad (44)$$

$$\frac{d[Cred]_{im}}{dt} = (+2v_{C3} - 2v_{C4}) / W_{is} \quad (45)$$

$$\frac{d[fATP]_{is}}{dt} = (+v_{ATP_{om}} + v_{AAC_{om}} - v_{AK_{is}}) / W_{is} - v_{MgATP_{is}} \quad (46)$$

$$\frac{d[mATP]_{is}}{dt} = +v_{mATP_{om}} / W_{is} + v_{MgATP_{is}} \quad (47)$$

$$\frac{d[fADP]_{is}}{dt} = (+v_{ADP_{om}} - v_{AAC_{om}} + 2v_{AK_{is}}) / W_{is} - v_{MgADP_{is}} \quad (48)$$

$$\frac{d[mADP]_{is}}{dt} = +v_{mADP_{om}} / W_{is} + v_{MgADP_{is}} \quad (49)$$

$$\frac{d[fAMP]_{is}}{dt} = (+v_{AMP_{om}} - v_{AK_{is}}) / W_{is} \quad (50)$$

$$\frac{d[P_i]_e}{dt} = (-v_{PiOH} + v_{Pi_{om}}) / W_e \quad (51)$$

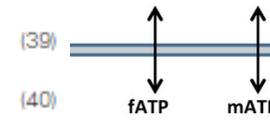
$$\frac{d\Delta\Psi}{dt} = (+4v_{C1} + 2v_{C3} + 4v_{C4} - n_A \cdot v_{F1F0} - v_{AAC} - v_{leak} - v_K) / C \quad (52)$$

$$[NAD]_x = N_{tot_x} - [NADH]_x \quad (53)$$

$$[Q]_{im} = Q_{tot_{im}} - [QH_2]_{im} \quad (54)$$

$$[Cox]_{is} = C_{tot_{is}} - [Cred]_{is} \quad (55)$$

$$[mADP]_e = \frac{1}{2} \left((K_{d_{mADP}} + ADP_{tot_e} + Mg_{tot_e}) - \sqrt{(K_{d_{mADP}} + ADP_{tot_e} + Mg_{tot_e})^2 - 4(Mg_{tot_e} \cdot ADP_{tot_e})} \right) \quad (56)$$



$$[fADP]_e = ADP_{tot_e} - [mADP]_e \quad (57)$$

$$[Mg]_e = Mg_{tot_e} - [mADP]_e \quad (58)$$

$$[Mg]_{is} = [Mg]_e \quad (59)$$

$$[P_i]_e = [P_i]_e \quad (60)$$

$$[K]_e = [K]_e \quad (61)$$

$$[H]_e = [H]_e \quad (62)$$

$$[H_2PO_4]_{is} = \frac{[H]_{is} \cdot [P_i]_{is}}{[H]_{is} + K_{d_{H_2PO_4}}}$$

$$[H_2PO_4]_x = \frac{[H]_x \cdot [P_i]_x}{[H]_x + K_{d_{H_2PO_4}}} \quad (63)$$

$$[OH]_{is} = \frac{10^{-14}}{[H]_{is}} \quad (64)$$

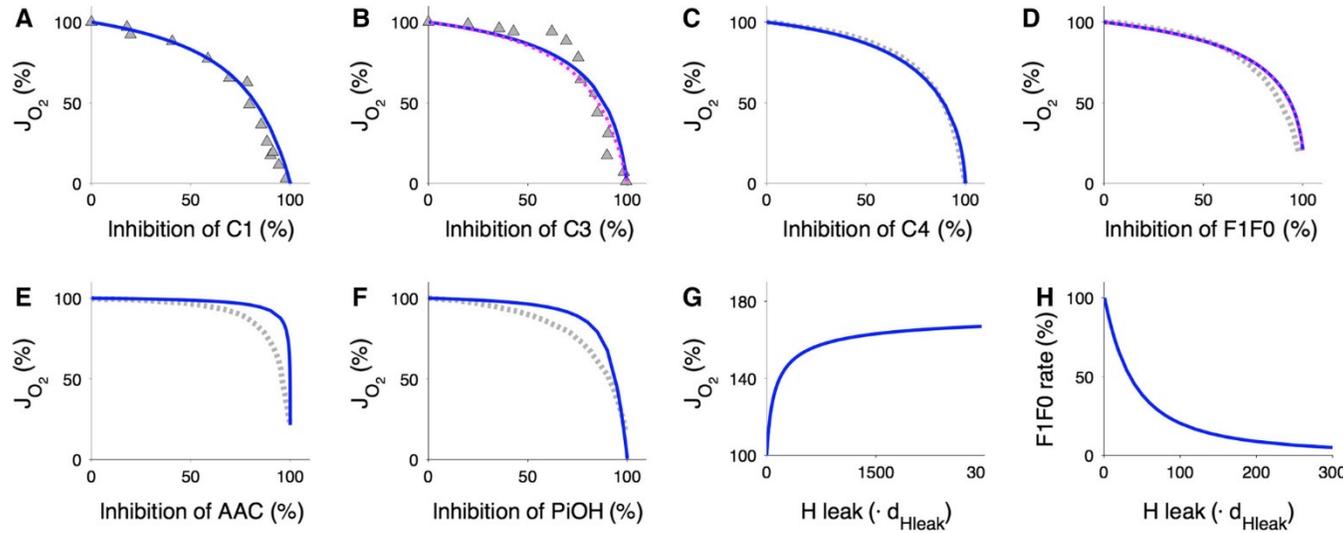
$$[OH]_x = \frac{10^{-14}}{[H]_x} \quad (65)$$

$$[OH]_x = \frac{10^{-14}}{[H]_x} \quad (66)$$

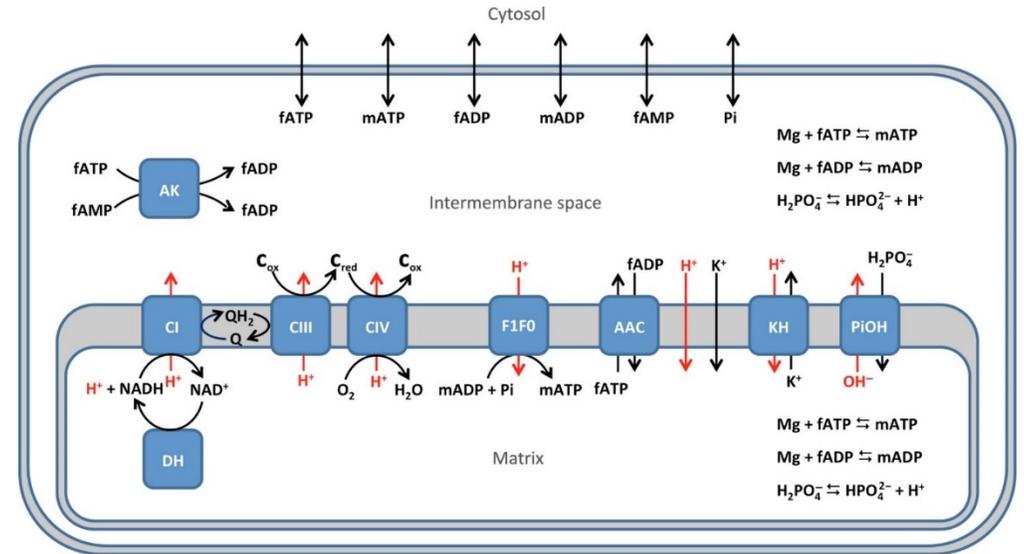
Heiske et al., FEBS Journal 2017

Le cycle de Krebs en 2023 : Phosphorylation oxydative

Optimisation non-linéaire sous contrainte



On obtient une modélisation de l'évolution de toutes les concentrations



Heiske et al., FEBS Journal 2017

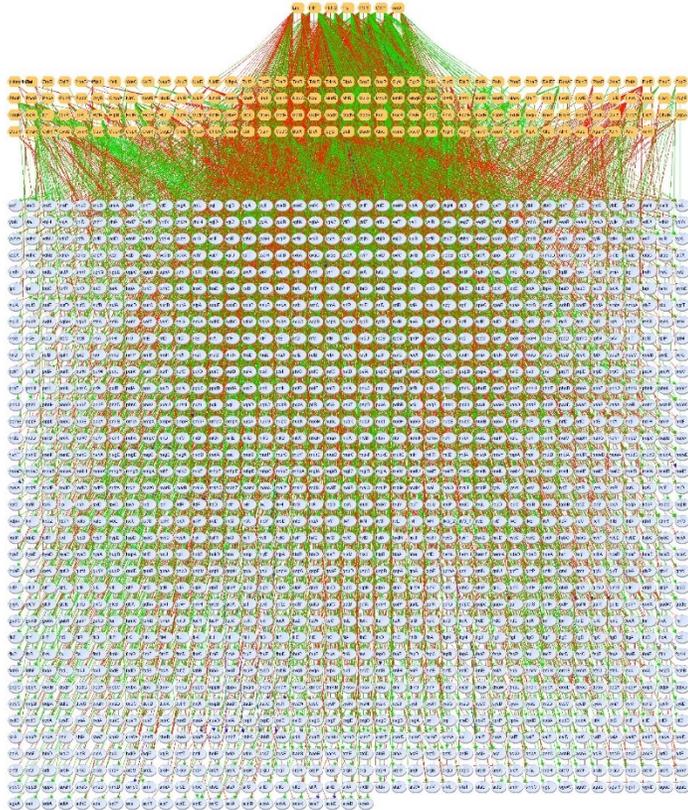
- Réglage fin de tous les paramètres du modèle
- **20 pages de description du modèle !**
- Le modèle reste « simple » ce n'est pas un organisme entier

« Base de données la plus complète sur terre »

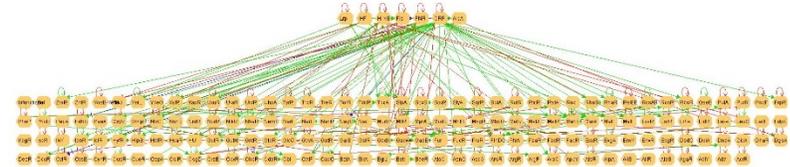
RegulonDB

Escherichia coli

Réseau TF vs gène



Réseau TF vs TF



- Réseau TF vs opéron
- Réseau Facteur sigma vs gène
- Réseau facteur sigma vs opéron

Regulon DB en 2023



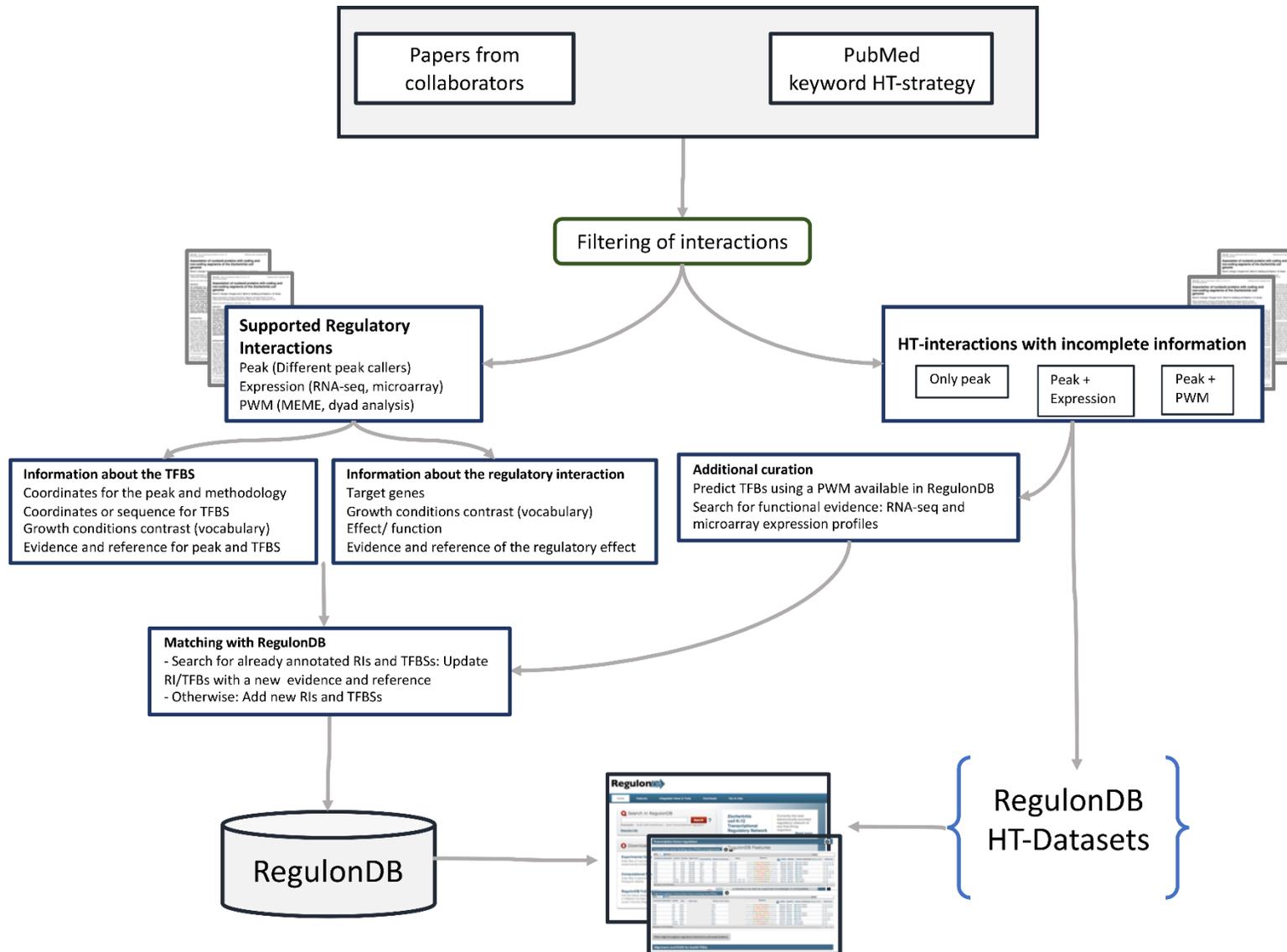
Escherichia coli K-12
Transcriptional
Regulatory Network

Currently the best electronically-encoded regulatory network of any free-living organism.

[Read more](#)

Object	Total	Weak Evidence	Strong Evidence	Confirmed Evidence	Without Evidence
Transcription Units:	3696	2768	525		403
Genes:	4736				4736
Promoter:	8795	3099	5680	16	
Operon:	2592				2592
TF binding Sites:	6958	5030	1558		370
Regulatory Interactions:	3951	3169	86		696
small RNA Interactions:	247	171	76		
Terminators:	366				366
RBSs - Shine-Dalgarno:	179				179
Transcriptional Factors:	229	96	130		3
Simple Regulons:	124				
Complex Regulons:	432				
Effectors:	138				138
Attenuators:	751				751
Riboswitches:	51				51
Synonyms:	30794				
Growth Global Conditions:	16				
Experiment Conditions:	82				
Affected Genes in Different Experimental Conditions:	316				
Gensor Unit:	53				

RegulonDB en 2023



« Base de données la plus complète sur terre »

RegulonDB

Escherichia coli

RegulonDB en 2023

Experimental datasets

Description	File
coli K-12 genome sequence used into RegulonDB	E. coli K-12 genome sequence raw format
	E. coli K-12 genebank
	E. coli K-12 genebank refseq
Sequences	Gene Sequence
	5' and 3' UTR sequence of TUs
Gene - Product	All gene products
	Gene Product Identifiers
	sRNA genes
Transcriptional Factors - Functional conformation	 Download
Regulatory Interactions	 Download
Regulatory Network Interactions	TF - gene interactions 
	TF - operon interactions 
	TF - TU interactions 
	TF - TF interactions 
	Sigma - gene interactions 
	Sigma - TU interactions 
	Alon and MA interactions 
	sRNA - gene interactions 

Promoters	All Promoters 
	Sigma 70 
	Sigma 54 
	Sigma 38 
	Sigma 32 
	Sigma 28 
	Sigma 24 
	Sigma 19 
	Unknown 
Transcription start sites experimentally determined in the laboratory of Dr. Morett	High-throughput transcription initiation mapping. Illumina directional RNA-seq experiments were total RNA received different treatments to enrich for 5' monophosphate or 5' triphosphate ends. Version 3.0. See the file description. 
	High-throughput transcription initiation mapping. See the file description. 
	5'-RACE transcription initiation mapping with specific primers. See the file description. 
Transcription Factor Weight Matrix	TF-Matrix browser
	 Download
Active and Inactive Transcription Factor Conformations	 Download
Transcription Units	 Download
Operons	 Download
Growth Conditions	 Download

RegulonDB en 2023

Computational Predictions datasets

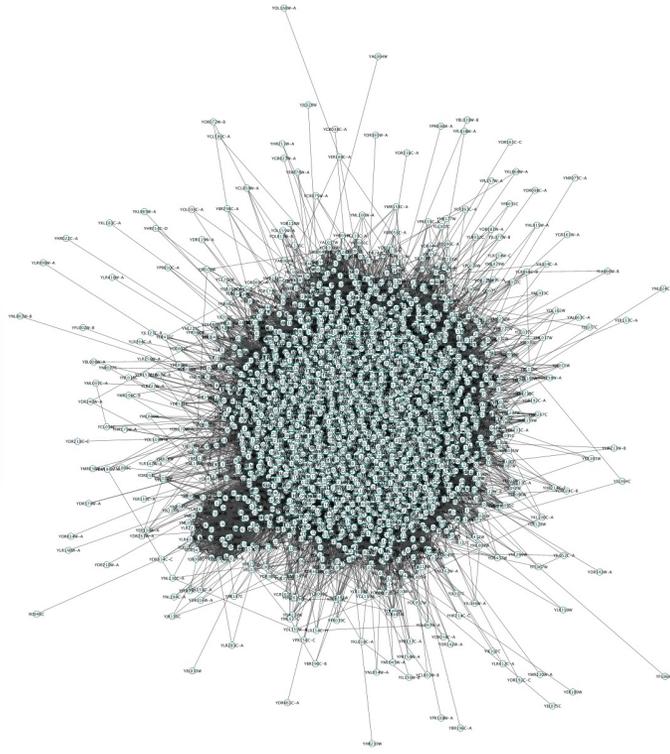
Description	Method
Promoter predictions	"We observed that real promoters occur mostly within regions with high densities of overlapping putative promoters. We evaluated several strategies to identify promoters. The best one uses an intrinsic score of the -10 and -35 hexamers that form the promoter as well as an extrinsic score that uses the distribution of promoters from the start of the gene. This high signal density is found mainly within regions upstream of genes, contrasting with coding regions and regions located between convergently transcribed genes." A.M. Huerta, J. Collado-Vides, J Mol Biol. 333:261-78 (2003).
Operon predictions	Operon prediction on (intergenic) distances Operon predictions based on (intergenic) distances and Riley's functional classification. We have previously demonstrated that genes within experimentally characterized operons of <i>Escherichia coli</i> are conserved together in other genomes more frequently than genes located at the borders of transcription units. We also show the relationship between our analyses of conservation and the inference of functional relationships from a genomic context
TF binding sites predictions	We have taken advantage of the phylogenetic proximity of <i>Escherichia coli</i> and other 16 organisms of this subdivision and the intensive search of the space sequence provided by a pattern-matching strategy. Using this approach, we complement predictions of regulatory sites made by using statistical models currently stored in Tractor_DB, and increase the number of transcriptional regulators with predicted binding sites up to 86. The original prediction approach, based on the representation of binding sites through statistical models was complemented by a new approach that uses known <i>E. coli</i> regulatory sites as the basis for a pattern matching search of regulatory sites. The use of both approaches together resulted in a more intensive exploration of the sequence space of each regulator's binding site. Computationally predicted transcription factor binding sites (TFBSs) using the evaluated weight matrix (see http://regulondb.ccg.unam.mx/menu/download/datasets/index.jsp). We scanned all upstream regions of every single gene, from +50 to -400 or from +50 to the closest upstream ORF, whatever happens first. (see the methodology)

tors Predictions	"Regulatory proteins in <i>Escherichia coli</i> with a helix-turn-helix (HTH) DNA binding motif show a position-function correlation such that repressors have this motif predominantly at the N terminus, whereas activators have the motif at the C-terminus extreme. Evidence is presented supporting a common history at the origin of this correlation. These results suggest that if shuffling of motifs occurred in Bacteria, it occurred only early in the history of these proteins, as opposed to what is observed in eukaryotic regulators." Pérez-Rueda E, Collado-Vides J. J Mol Evol. 2001 Sep;53(3):172-9.
s Prediction	For each group of orthologous proteins, the upstream regions of the first gene of each operon are taken and searched for motifs using MEME (Figure 1a). Each motif is then refined by several cycles of locating it among all upstream regions from all bacteria using MAST, and redefining a more specific motif with MEME (Figure 1b). Sequences with motifs can then be analyzed to see if they present evidence of conserved secondary structure (Figure 1c). Predicted motifs are also compared against the Rfam database to locate known structured elements and against RegulonDB to find known transcription factor binding sites. Click here to see image.
Prediction	For each predicted operon, the upstream region of the first gene is taken (Figure 1a). For every run of Us present in this region (Figure 1b), a stable structure in the adjacent region is searched for (Figure 1c). If a terminator is found, an anti-terminator is searched for, since it must be overlapping with the terminator (Figure 1d). An anti-antiterminator can be analogously located by finding a structure that overlaps with the anti-terminator (Figure 1e). For the particular case of translational attenuators, a terminator is searched for, since it overlaps with the Shine-Dalgarno site. Click here to see image.

Saccharomyces cerevisiae – YeastNet v3

[YeastNet: a network by integration of all data-type-specific networks (CC, CX, DC, GN, GT, HT, LC PG, TS)]

YeastNet v.3		
Edge list download	5808 genes	362421 links



Kim et al.,
Nucleic Acid Research 2014

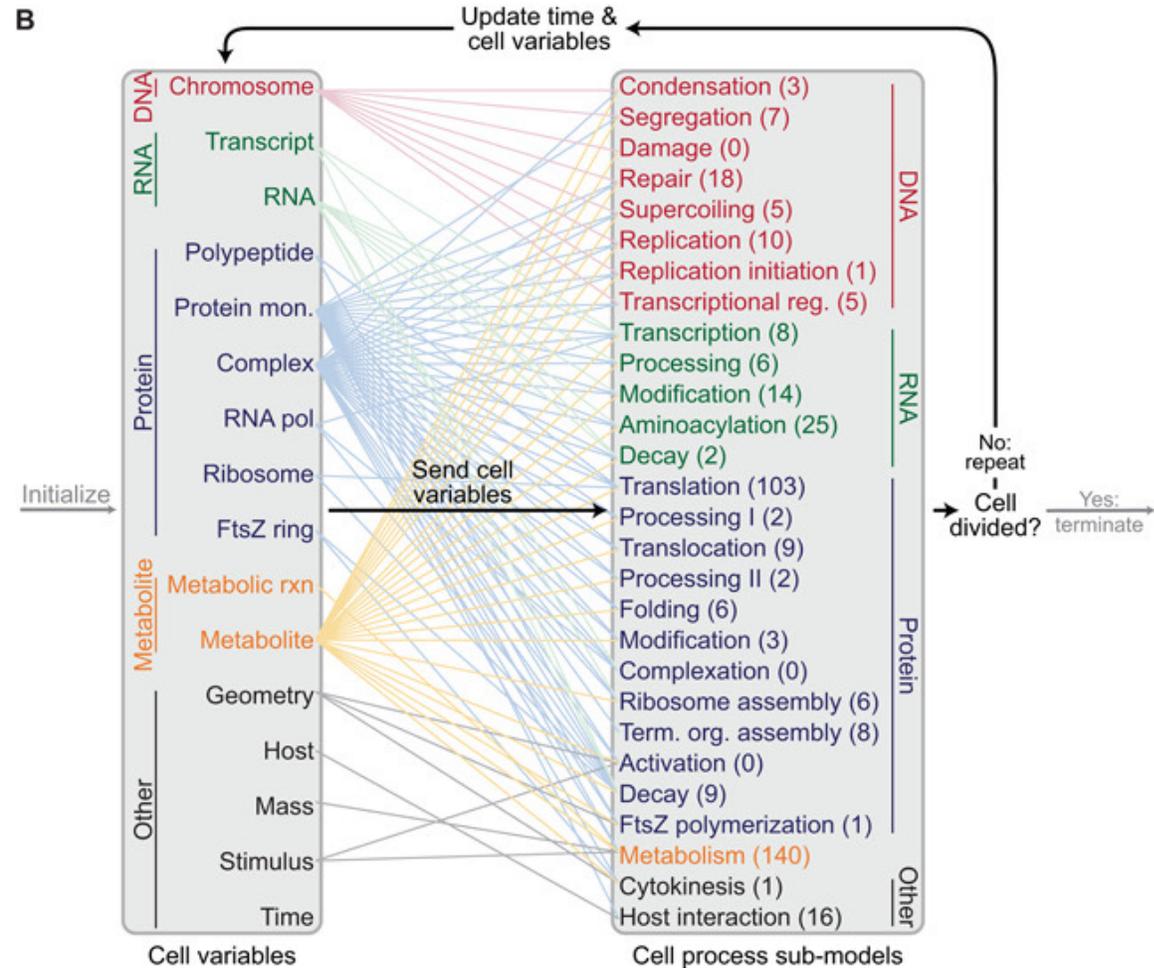
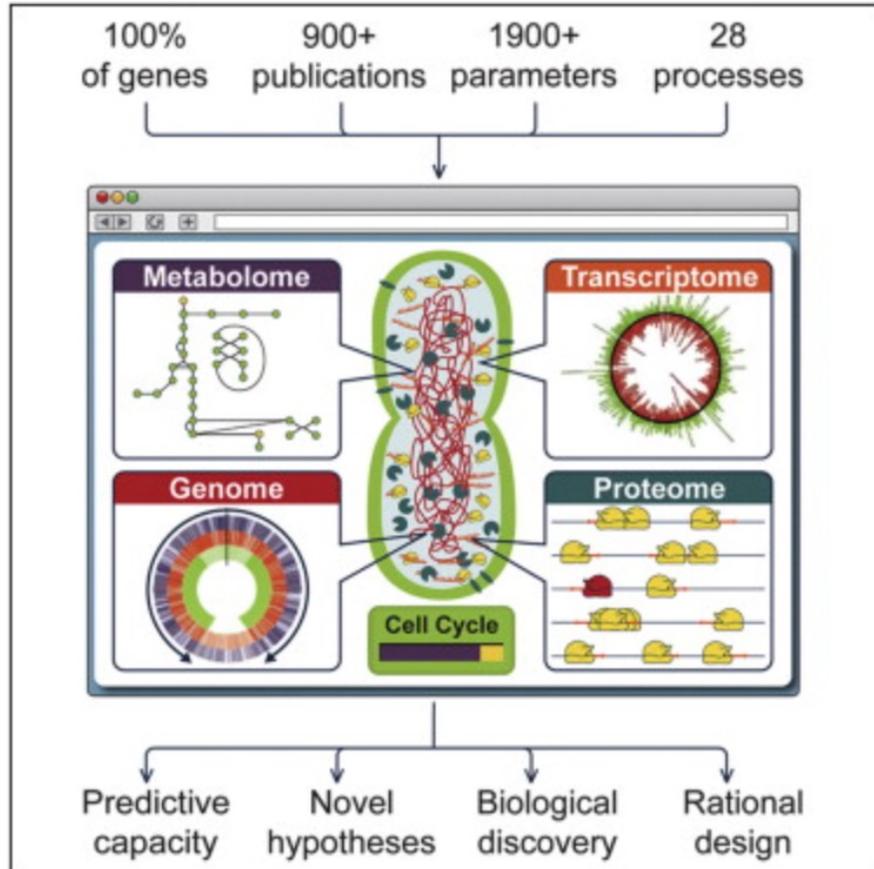
Saccharomyces cerevisiae – YeastNet v3

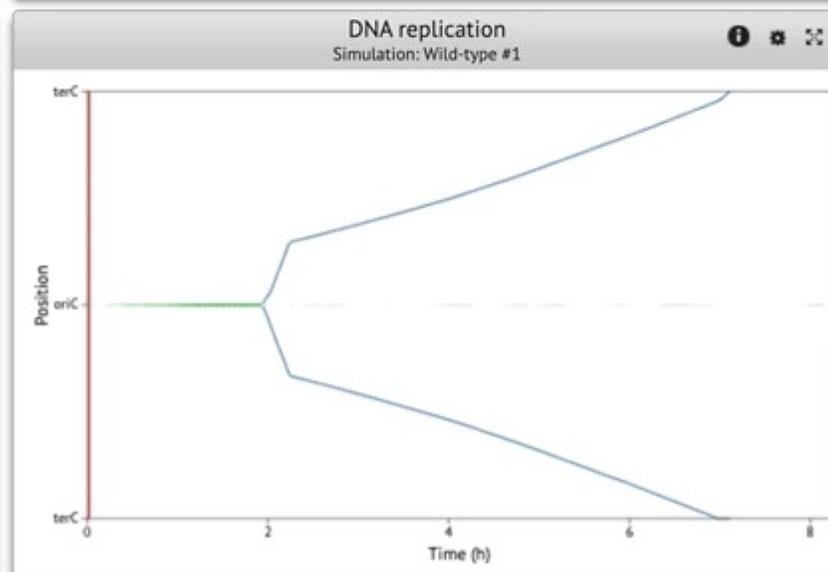
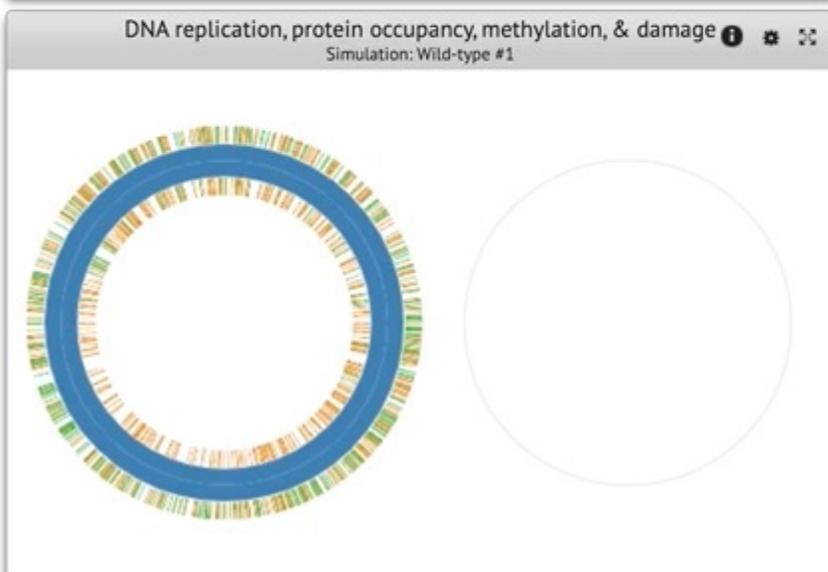
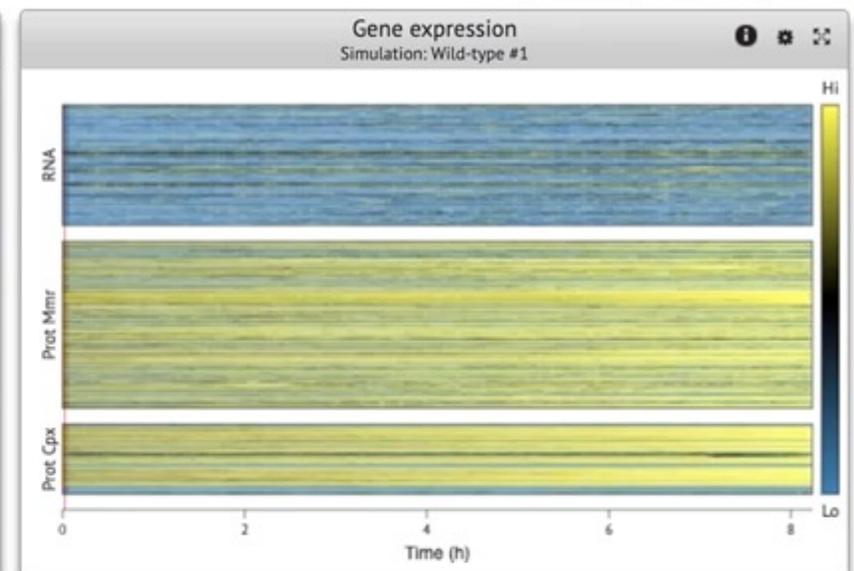
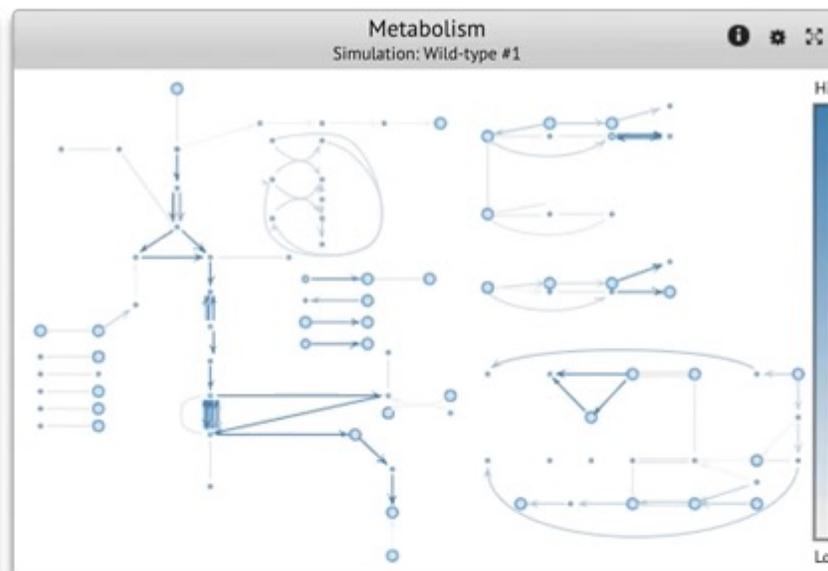
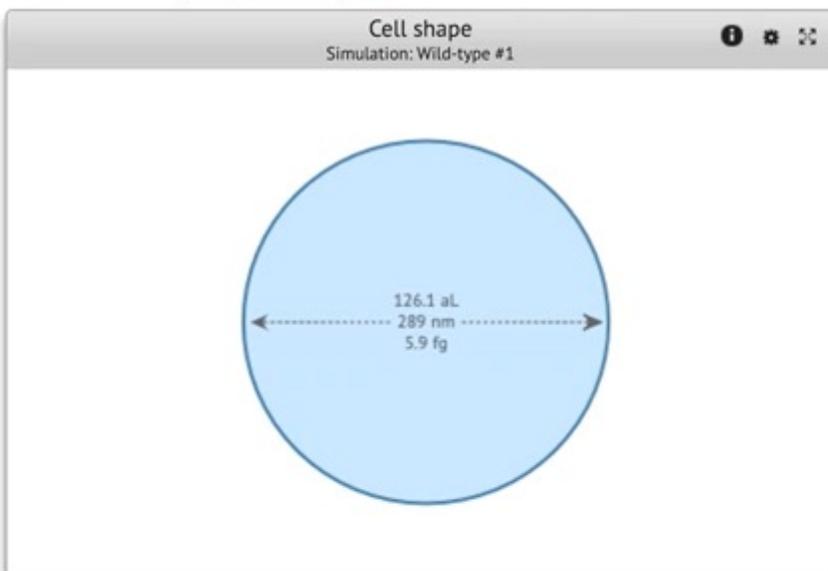
Evidence code	Data set description
CC	Inferred links by co-citation of two genes across 46,111 pubmed Medline article abstracts for yeast biology
CX	Inferred links by co-expression pattern of two genes (based on high-dimensional gene expression data)
DC	Inferred links by co-occurrence of protein domains between two coding genes
GN	Inferred links by similar genomic context of bacterial orthologs of two yeast genes

GT	Inferred links by similar profiles of genetic interaction partners
HT	Links by high-throughput protein-protein interactions
LC	Links by small/medium-scale protein-protein interactions (collected from protein-protein interaction data bases)
PG	Inferred links by similar phylogenetic profiles between two yeast genes
TS	Inferred links by 3-D protein structure of interacting orthologous proteins between two yeast proteins

28 sous-modèles de 28 processus cellulaires
 401 gènes, 722 molécules, 1,857 réactions, and 1,836 paramètres

Mycoplasma genitalium Reconstruction ?





Modèle d'organisme entier

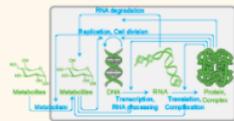


Whole-Cell Modeling
PREDICTING PHENOTYPE FROM GENOTYPE
FOR SCIENCE, MEDICINE & ENGINEERING

[Models](#) [Tools](#) [Team](#) [Learn more](#) [Get involved](#) [News](#) [About](#)

Models: Comprehensive computational models of individual cells

Archetypal bacterium



The archetypal bacterium model generator is a tool for generating WC models that represent user-specified numbers of genes, RNA, proteins, and reactions. The models generated by the model generator represent the metabolism, replication, transcription, translation, RNA and protein degradation, and cell division of a typical bacterium. The archetypal bacterium model generator is particularly useful for driving the development of WC modeling tools, as well as teaching WC modeling.

Availability: In development
Author: Karr Lab, Sinai

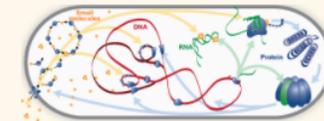
Mycoplasma genitalium



The *M. genitalium* WC model was the first model that represented each characterized gene function of an organism. The model is composed of 28 submodels of 28 cellular processes. In total, the model represents the functions of 401 genes, 722 compounds, 1,857 reactions, and 1,836 parameters. In addition to demonstrating the feasibility of WC models, the model has been used to gain new insights into cell cycle regulation, learn unknown parameters, and suggest new uses of existing antibiotics for *Mycoplasmas*.

Availability: Download
Author: Covert Lab, Stanford
More info: Docs | Source | License | Tests
Reference: Karr JR et al. *Cell* 2012.

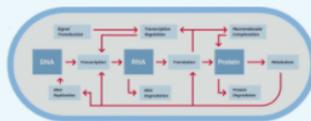
Mycoplasma pneumoniae



The *M. pneumoniae* WC model will be the most comprehensive, most systematically constructed, and most extensible WC model to date. The model will represent all of the major cellular functions of *M. pneumoniae*, including the function each characterized gene. The model will be based primarily on *M. pneumoniae* genomic data. The model will be used to drive the development of WC modeling methods, as well as to help design a reliable, energy efficient, fast-growing chassis for future bioengineering.

Availability: In development
Author: Karr Lab, Sinai

Escherichia coli



The *E. coli* WC model represents the core cellular functions of *E. coli*. The model is the most detailed and most thoroughly tested WC model to date.

Availability: In development
Author: Covert Lab, Stanford

Homo sapiens (H1-hESC)



The H1 human embryonic stem cell (hESC) model is the first step toward WC models of human cells. The model will represent the core cellular functions of all human cells including their metabolism, DNA replication, transcription, translation, protein complexation, RNA and protein degradation, and division. The model focuses on H1-hESCs because ESCs behave as individual cells, because ESC lines are karyotypically normal, because ESCs grow quickly, and because H1-hESC has been extensively characterized. In addition to demonstrating the feasibility of human WC models and driving the development of WC modeling tools, the model will be used to gain insights into how stem cells maintain pluripotency.

Availability: In development
Author: Karr Lab, Sinai

Homo sapiens cancer signaling (MCF10A)

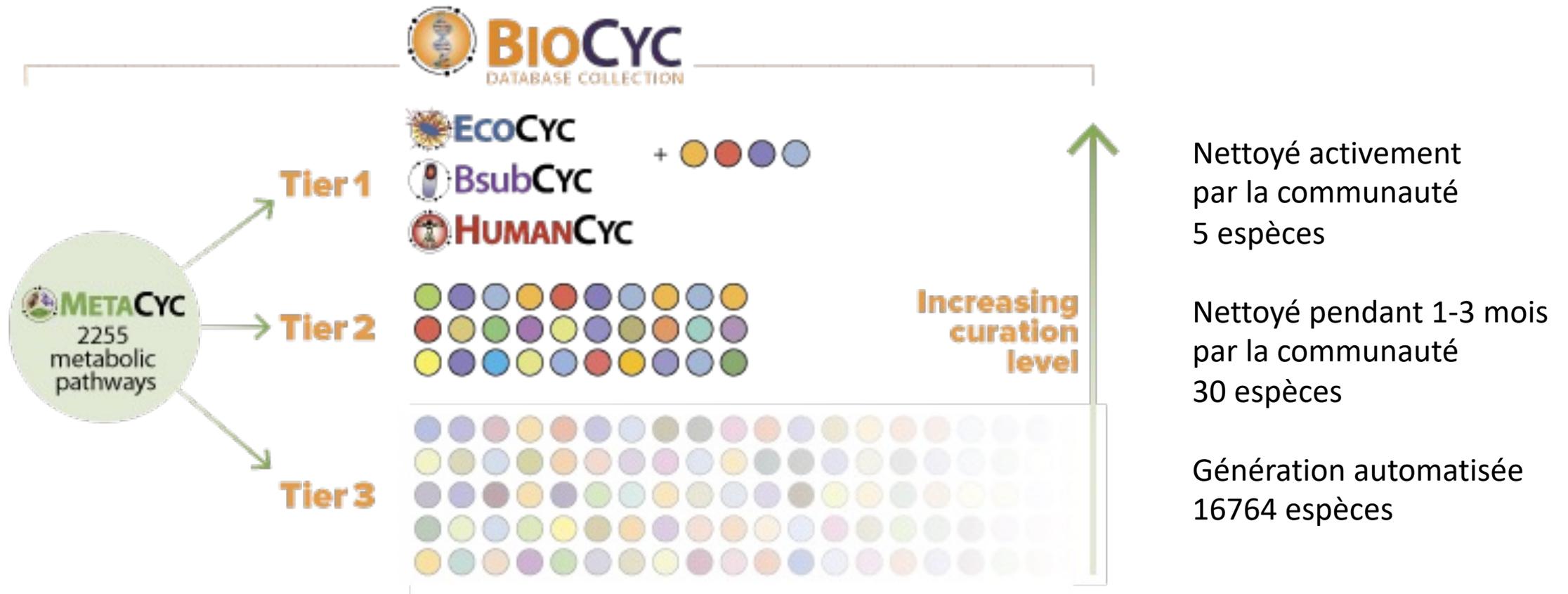


A mechanistic ordinary differential equation model describing the interactions between commonly mutated pan-cancer signaling pathways—receptor tyrosine kinases, Ras/RAF/ERK, PI3K/AKT, mTOR, cell cycle, DNA damage, and apoptosis.

Availability: Bouhaddou M et al. A mechanistic pan-cancer pathway model informed by multi-omics data interprets stochastic cell fate responses to drugs and mitogens. *PLoS Comput Biol* 26. e1005985 (2018).

Author: Birtwistle Lab, Clemson

Bases de données « nettoyées » par la communauté



Etat de l'art : BioCyc



Genes:	4,737	SmartTable	Ontology
Pathways:	365	SmartTable	Ontology
Enzymatic Reactions:	2,202		
Transport Reactions:	530		Ontology
Polypeptides:	4,466	SmartTable	Ontology
Protein Complexes:	1,166	SmartTable	Ontology
Enzymes:	1,714	SmartTable	
Transporters:	479	SmartTable	
Compounds:	2,967	SmartTable	Ontology
Transcription Units:	3,694	SmartTable	Ontology
tRNAs:	89		Ontology
Growth Media:	438		List
Transcriptional Regulation:	5,661		Ontology
Protein Features:	41,346		
Phenotype Microarray Datasets:	5		List
GO Terms:	71,124		Ontology
Gene Essentiality Datasets:	6		List



Genes:	4,541	SmartTable	Ontology
Pathways:	274	SmartTable	Ontology
Enzymatic Reactions:	1,529		
Transport Reactions:	92		Ontology
Polypeptides:	4,293	SmartTable	Ontology
Protein Complexes:	255	SmartTable	Ontology
Enzymes:	1,072	SmartTable	
Transporters:	632	SmartTable	
Compounds:	1,007	SmartTable	Ontology
Transcription Units:	1,648	SmartTable	Ontology
tRNAs:	86		Ontology
Growth Media:	1		List
Transcriptional Regulation:	864		Ontology
Protein Features:	22,636		
GO Terms:	34,078		Ontology
Gene Essentiality Datasets:	1		List



Genes:	20,997	SmartTable	Ontology
Pathways:	362	SmartTable	Ontology
Enzymatic Reactions:	2,895		
Transport Reactions:	145		Ontology
Polypeptides:	20,732	SmartTable	Ontology
Protein Complexes:	541	SmartTable	Ontology
Enzymes:	3,541	SmartTable	
Transporters:	773	SmartTable	
Compounds:	2,119	SmartTable	Ontology
tRNAs:	53		Ontology
Growth Media:	2		List
Transcriptional Regulation:	2		Ontology
Protein Features:	14		
GO Terms:	890,187		Ontology
Gene Essentiality Datasets:	1		List

Etat de l'art : MetaCyc



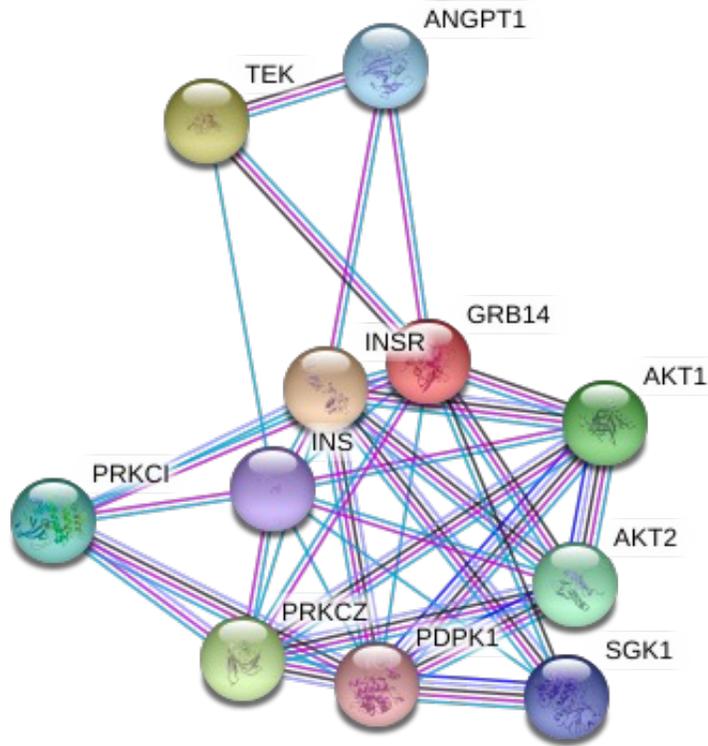
MetaCyc is a curated database of experimentally elucidated metabolic pathways from all domains of life.

MetaCyc contains pathways involved in both primary and secondary metabolism, as well as associated metabolites, reactions, enzymes, and genes. The goal of MetaCyc is to catalog the universe of metabolism by storing a representative sample of each experimentally elucidated pathway.

Genes:	14,976	SmartTable	Ontology
Pathways:	3,063	SmartTable	Ontology
Enzymatic Reactions:	18,285		
Transport Reactions:	910		Ontology
Polypeptides:	16,731	SmartTable	Ontology
Protein Complexes:	4,756	SmartTable	Ontology
Enzymes:	14,028	SmartTable	
Transporters:	638	SmartTable	
Compounds:	18,452	SmartTable	Ontology
tRNAs:	8		Ontology
Growth Media:	19		List
Protein Features:	30,982		
GO Terms:	72,124		Ontology

Automatisation de la reconstruction

Le serveur : string-db.org



Réseau d'interaction protéine-protéine de l'insuline prédite par string-db.org

Known Interactions

- from curated databases
- experimentally determined

Predicted Interactions

- gene neighborhood
- gene fusions
- gene co-occurrence

Welcome to STRING

Protein-Protein Interaction Networks
Functional Enrichment Analysis

ORGANISMS	PROTEINS	INTERACTIONS
14094	67.6 mio	>20 bln

SEARCH

Others

- textmining
- co-expression
- protein homology

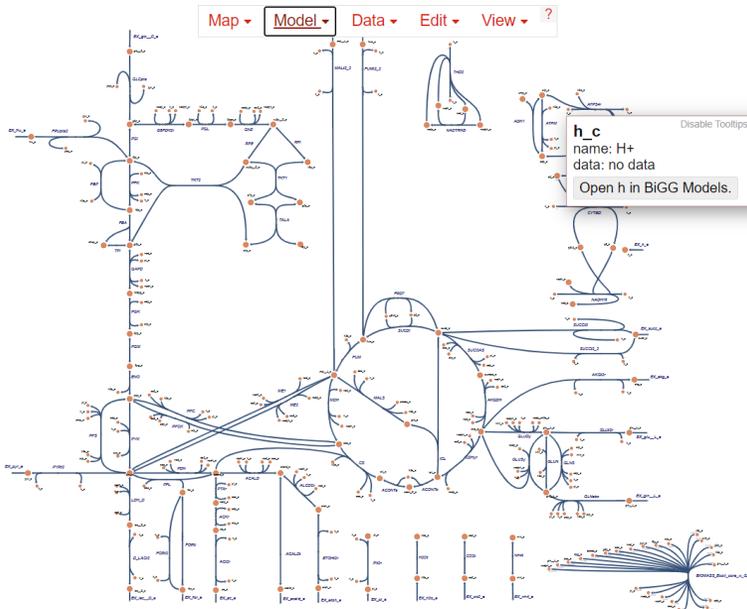
BiGG Models

Palsson et al.

Search the database by model, reaction, metabolite, or gene ?



The image shows three rectangular buttons with a blue border. The first button, labeled 'View Models', contains a diagram of a cell with labels for DNA, Ribosome, Enzyme, Membrane, Biomass precursors, and Actual biomass. The second button, labeled 'View Metabolites', shows a 3D illustration of a bacterium with various chemical structures floating around it. The third button, labeled 'View Reactions', displays a metabolic pathway diagram with enzymes like PFL, PDH, PTA, ACALD, and ALCOX.



E. Coli core



openCOBRA

The COntstraint-Based Reconstruction and Analysis Toolbox

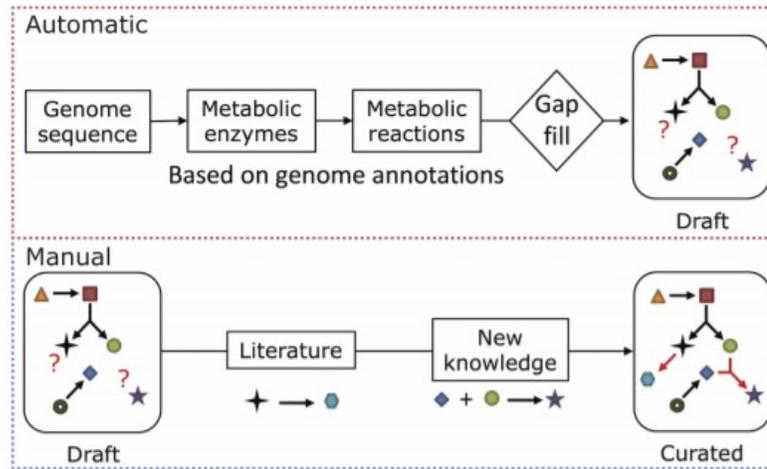
Automatisation de la reconstruction



openCOBRA

The COntstraint-Based Reconstruction and Analysis Toolbox

a) Genome-scale metabolic reconstruction



b) Flux balance analysis

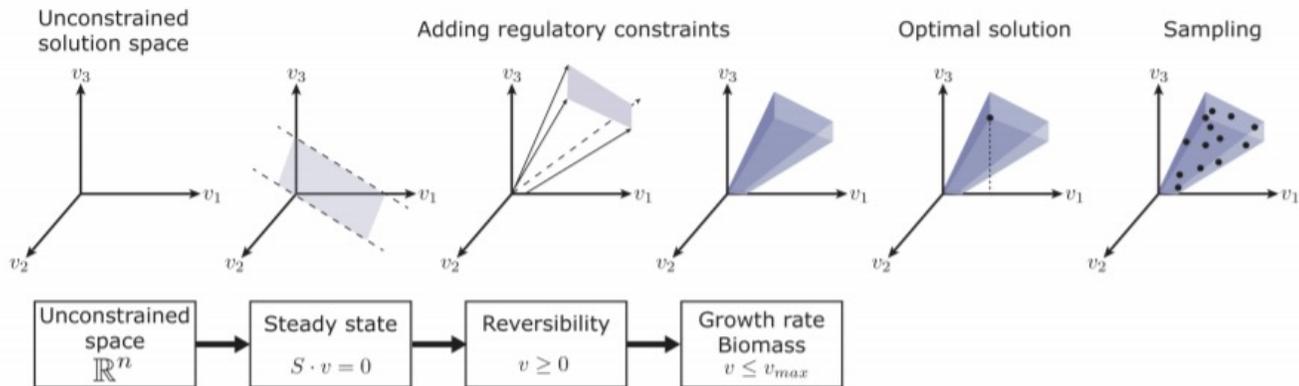
Maximize/minimize an objective function
 $Z = c_1 v_1 + c_2 v_2 + \dots + c_5 v_5$ such that:

		Reactions						
		R ₁	R ₂	R ₃	R ₄	R ₅		
Metabolites	▲	-1	0	0	0	0	⊗	$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$
	■	1	-1	0	0	0		
	✦	0	1	-1	0	0		
	●	0	1	0	0	-1		
	◆	0	0	1	0	0		
	★	0	0	0	-1	0		
		0	0	0	1	-1		
		0	0	0	0	1		

S-matrix Flux vector

and for every reaction i : $lb_i \leq v_i \leq ub_i$

c) Solution spaces

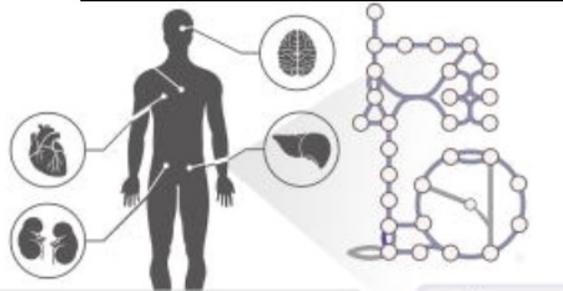


Etat de l'art : BIGG models (order by genes)

Models

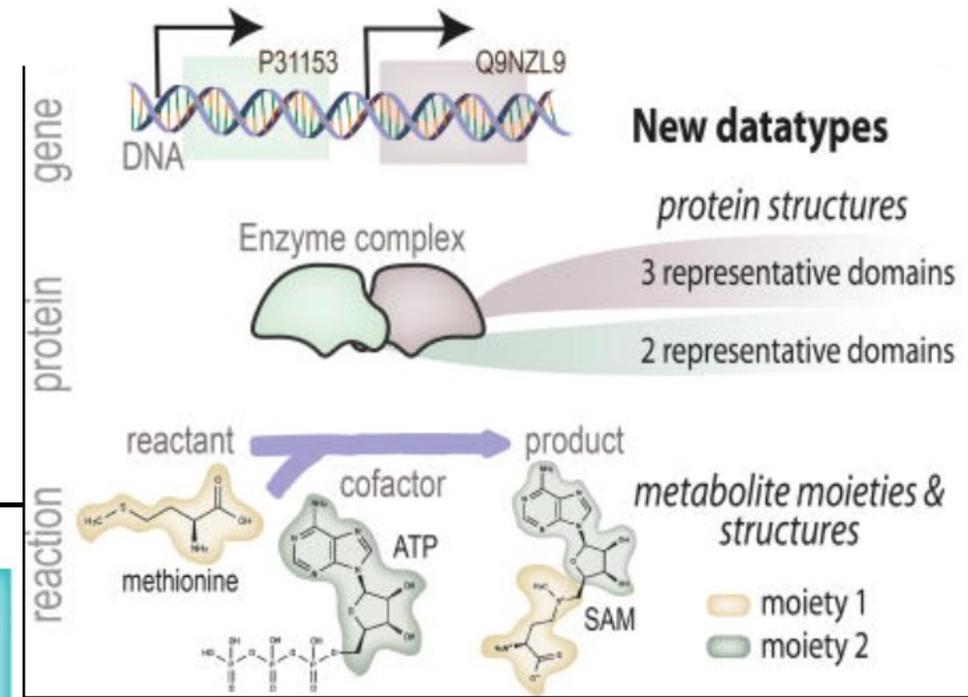
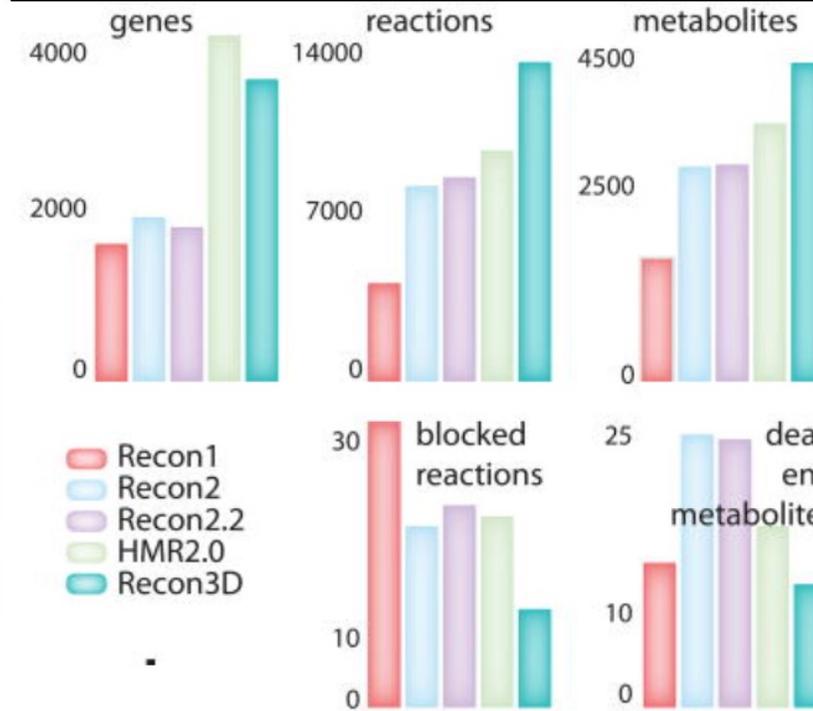
BiGG ID	Organism	Metabolites	Reactions	Genes
Recon3D	Homo sapiens	5835	10600	2248
RECON1	Homo sapiens	2766	3741	1905
iCHOv1	Cricetulus griseus	4456	6663	1766
iYS1720	Salmonella pan-reactome	2436	3357	1707
iML1515	Escherichia coli str. K-12 substr. MG1655	1877	2712	1516
iJN1463	Pseudomonas putida KT2440	2153	2927	1462
iECDH1ME8569_1439	Escherichia coli DH1	1950	2755	1439
iMM1415	Mus musculus	2775	3726	1375
iWFL_1372	Escherichia coli W	1973	2782	1372
iEC1372_W3110	Escherichia coli str. K-12 substr. W3110	1918	2758	1372

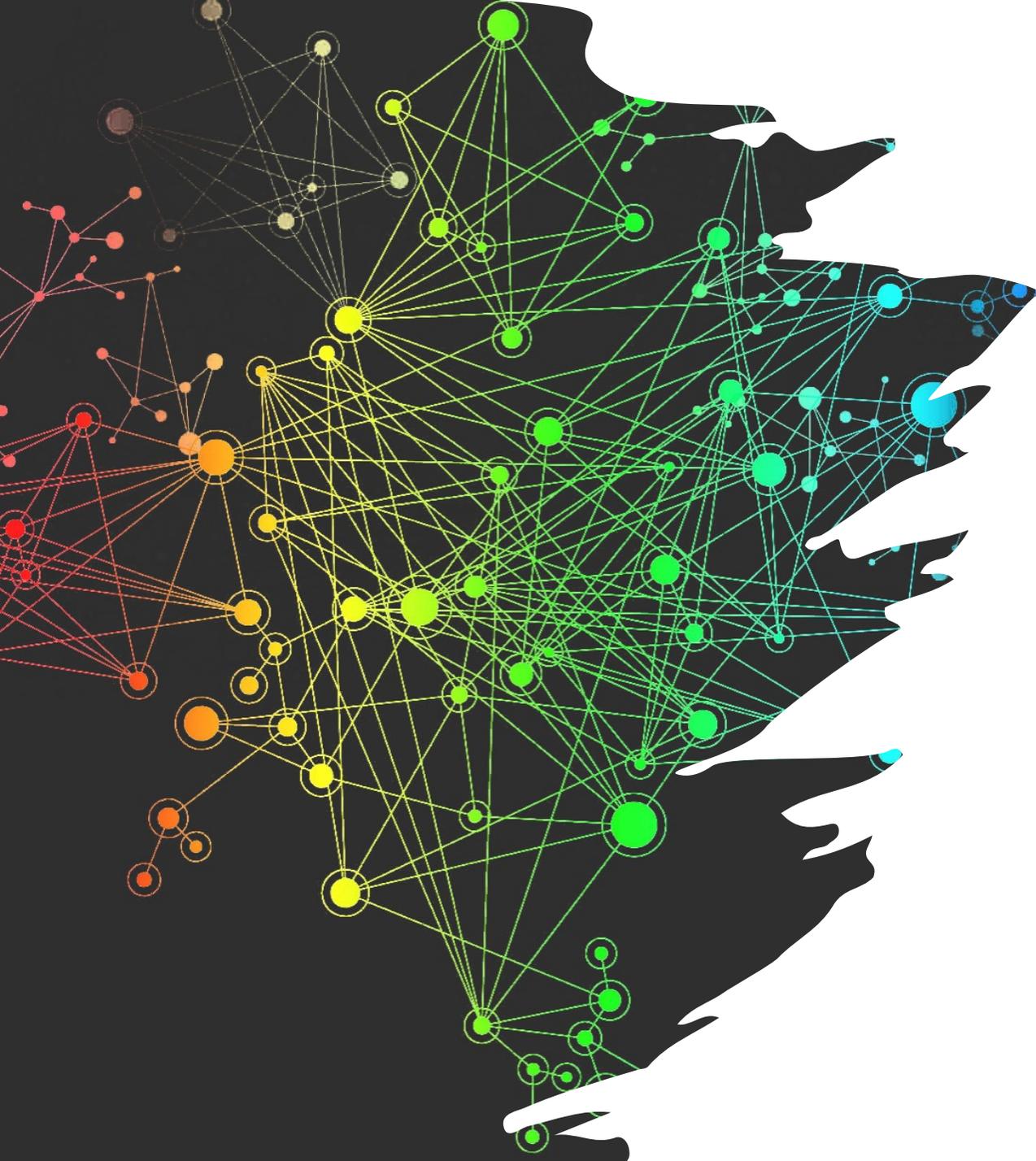
Recon3D : Human metabolism



biochemical data
genomic data
tissue-specific localization
proteomic data
metabolomic data
protein structural data
pharmacogenomic data
atom-atom mappings

Recon 3D
3288 genes
2908 domains
12890 structures
13543 reactions
4140 metabolites
3536 SNVs
8315 atom-atom mappings

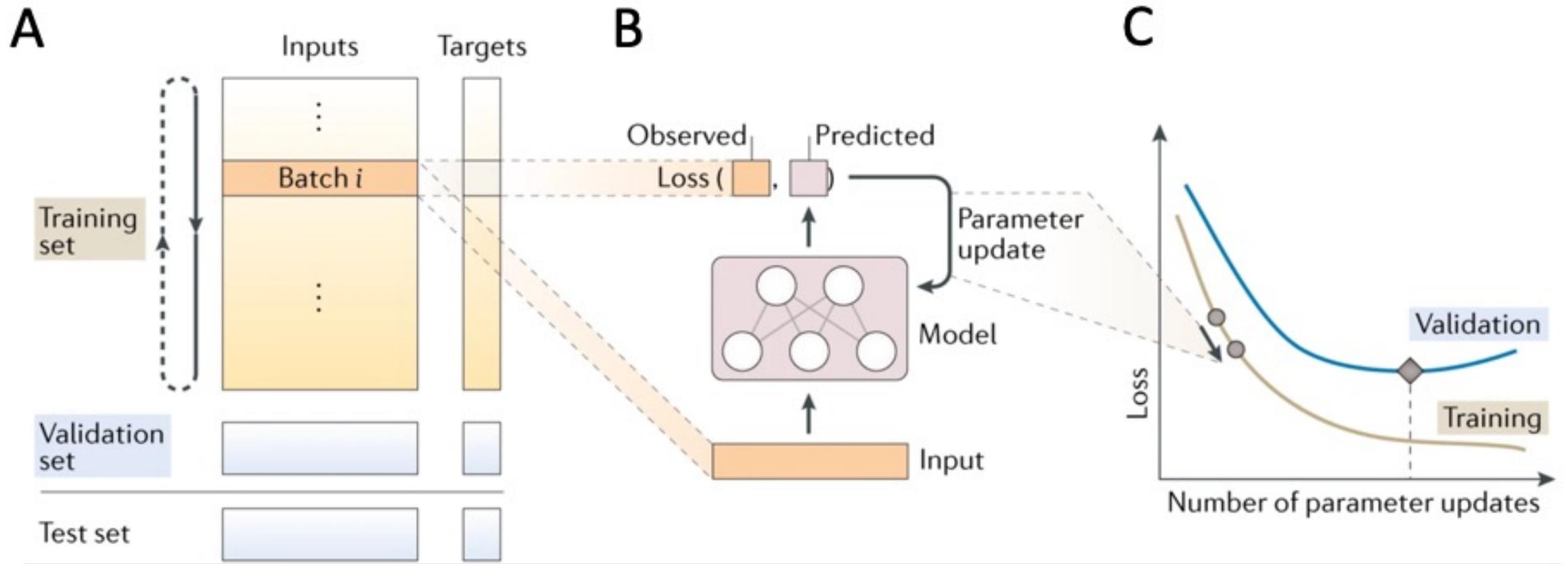




Machine learning pour la reconstruction des réseaux biologiques à l'aide de données omiques

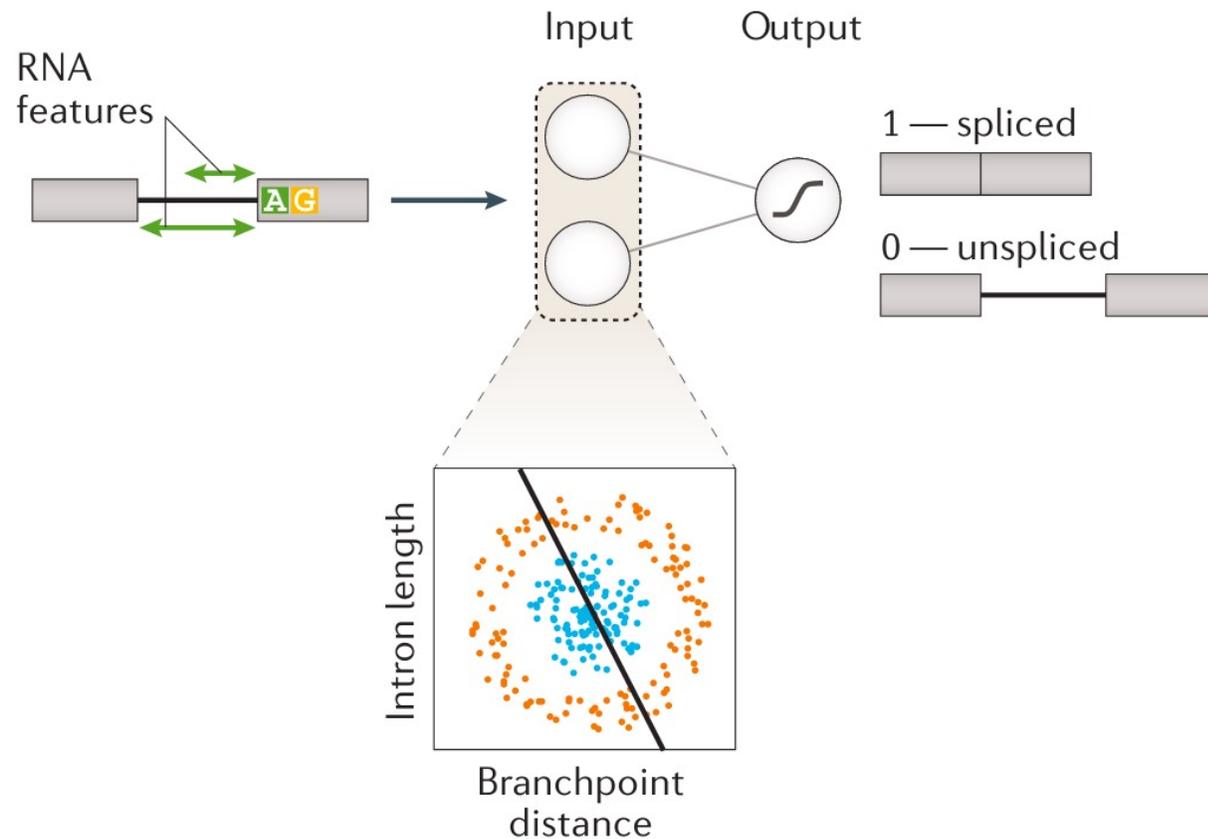
- Reconstruction de réseaux biologiques à l'aide de données omiques
- Reconstruire par analyse de co-expression (TP)
- Etat de l'art des réseaux biologiques les plus étendus
- **Apport du deep-learning à la biologie des systèmes**

Le Machine Learning : Principe

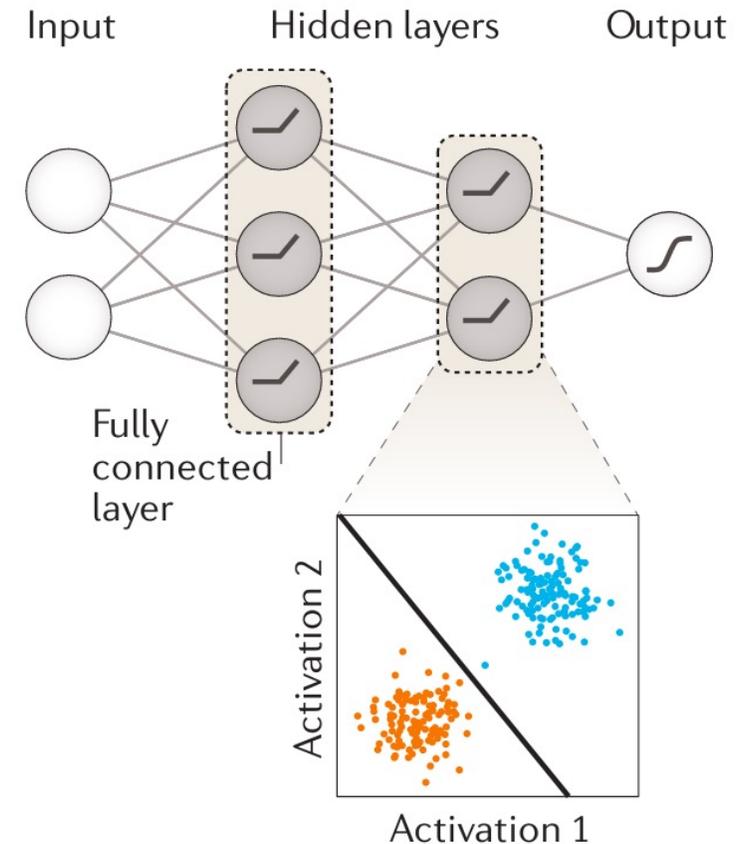


Le Deep Learning : Principe

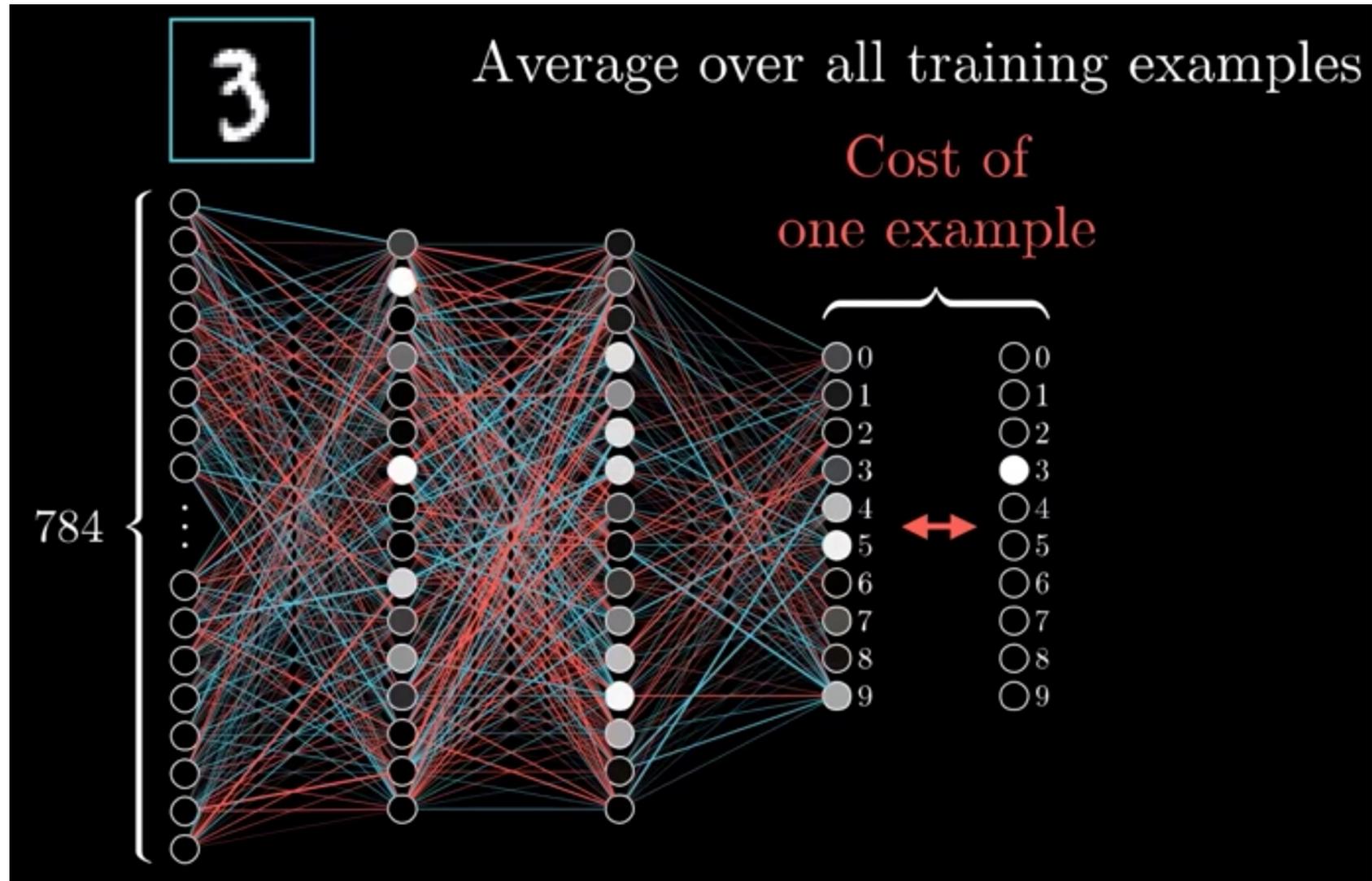
a Single-layer neural network (logistic regression)



b Multilayer neural network



Entraînement et loss fonction



Entraînement et loss fonction

Regression
Mean Square error

$$MSE = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2$$

AutoEncoder
Kullback-Liebler Divergence

$$L(y_{\text{pred}}, y_{\text{true}}) = y_{\text{true}} \cdot \log \frac{y_{\text{true}}}{y_{\text{pred}}}$$

Classification
Binary Cross-Entropy

$$CE \text{ Loss} = \frac{1}{n} \sum_{i=1}^N - (y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i))$$

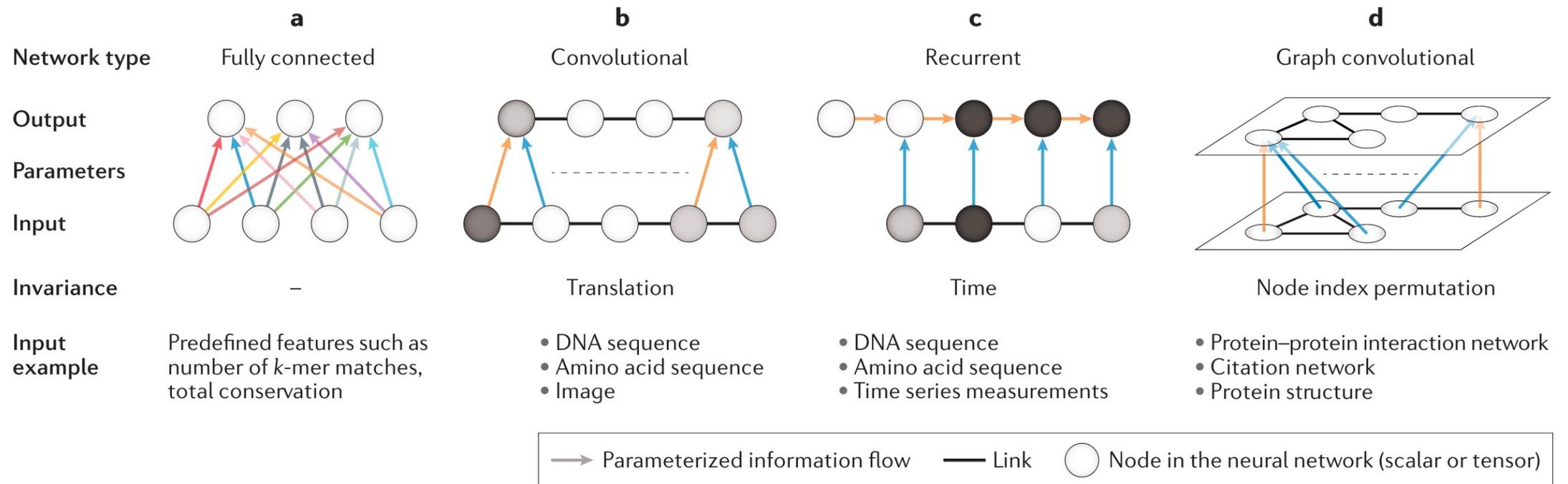
Generative Adversarial Network
Cross-Entropy

$$E_x [\log(D(x))] + E_z [\log(1 - D(G(z)))]$$

Dans cette fonction:

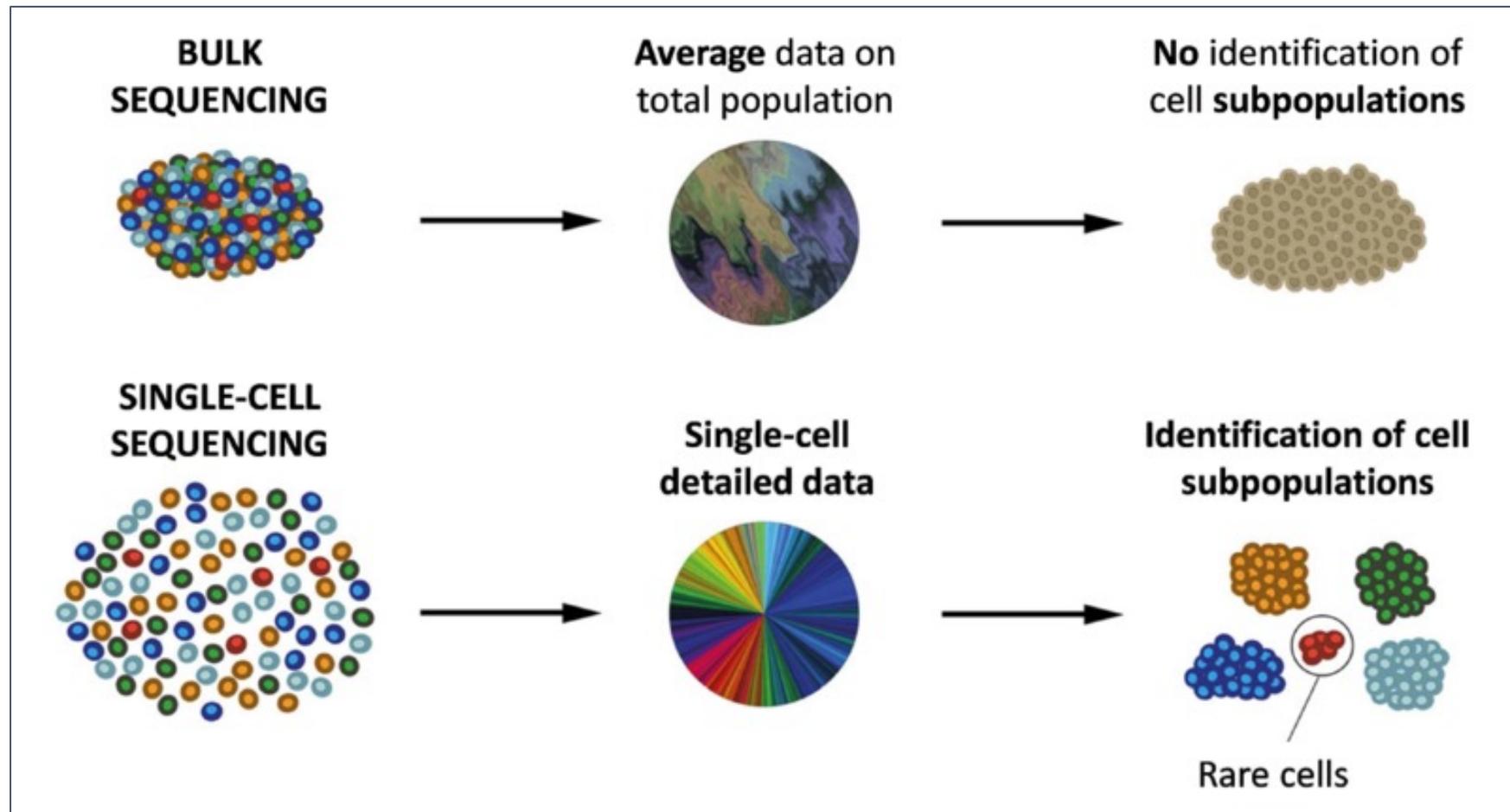
- $D(x)$ est une estimation de la probabilité que l'instance de données x réelle soit réelle.
- E_x est la valeur attendue sur toutes les instances de données réelles.
- $G(z)$ est la sortie du générateur lorsqu'il reçoit une valeur de z pour le bruit.
- $D(G(z))$ est l'estimation de la probabilité qu'une fausse instance soit réelle.
- E_z est la valeur attendue sur toutes les entrées aléatoires du générateur (en effet, il s'agit de la valeur attendue sur toutes les fausses instances générées, $G(z)$).
- La formule est dérivée de l'[entropie croisée](#) entre les distributions réelles et générées.

Les 4 architectures de références en deep learning

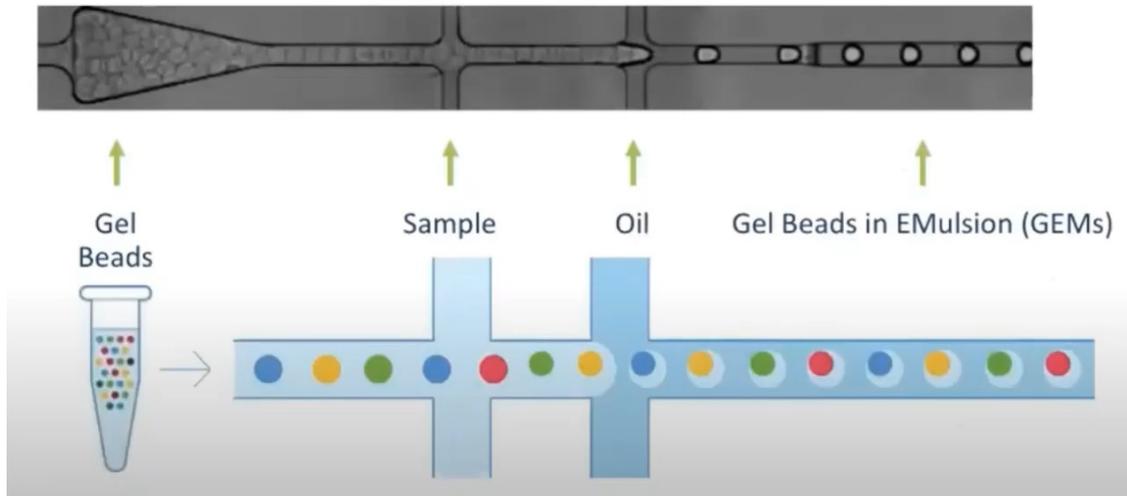


Eraslan, ... , Theis, Nature Review Genetics, 2019

Reconstruction de réseaux de régulations à l'aide du single-cell



What could possibly go wrong ? ...



- Multiple cells in droplet
- Ambient RNA in droplet
- Sampling batch
- Bad tissue dissociation
- Batch in cell bar coding
- Low sequencing quality
- Low number of reads & UMI

Méthodes d'intégration de données multi-omiques

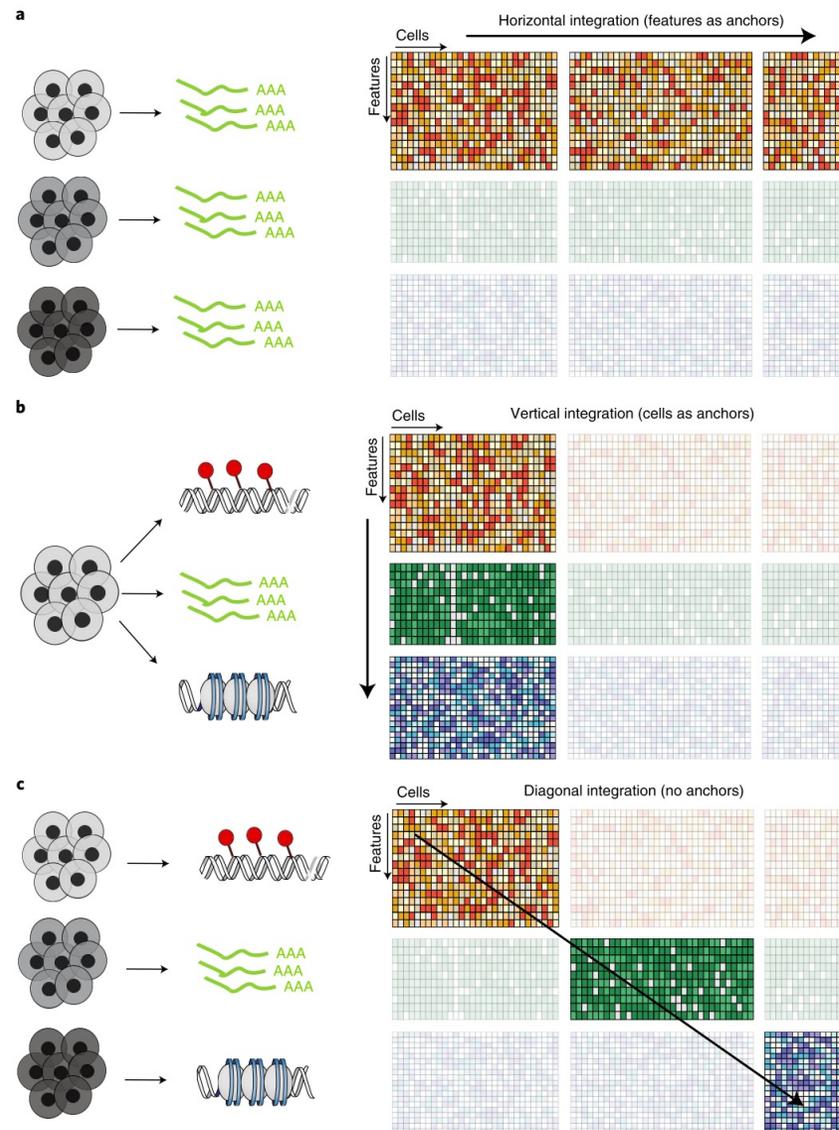
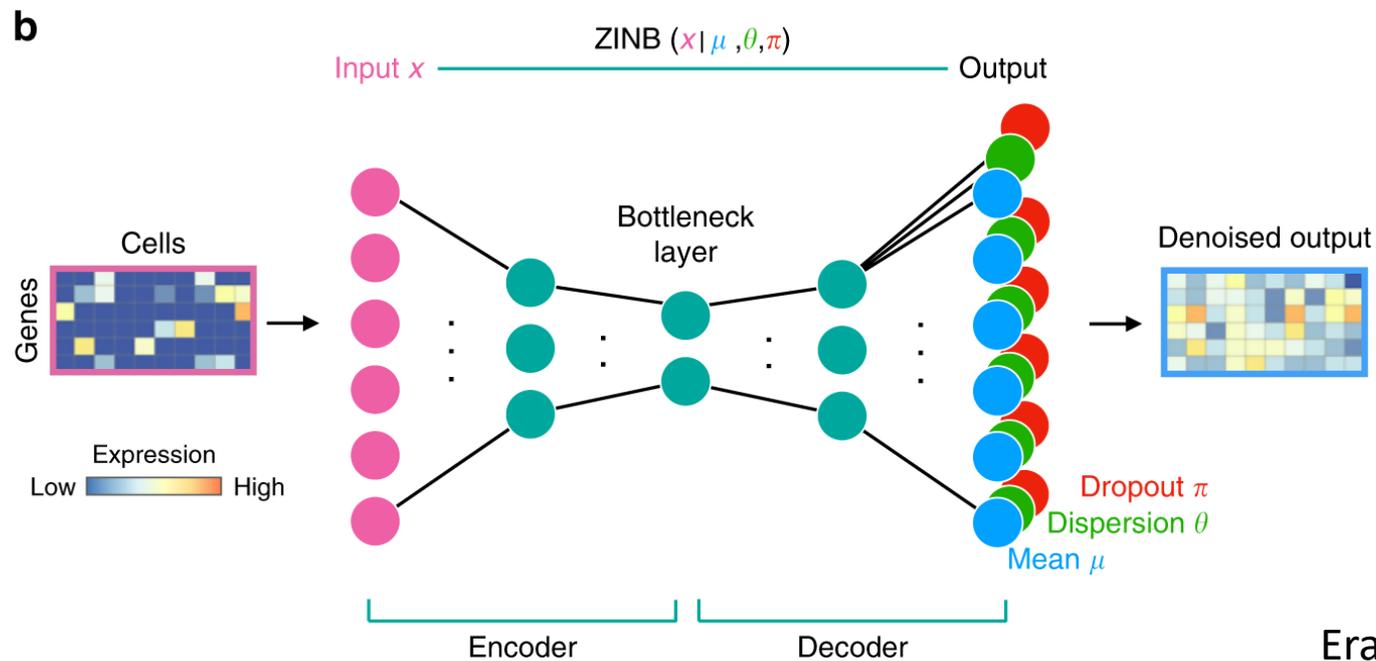
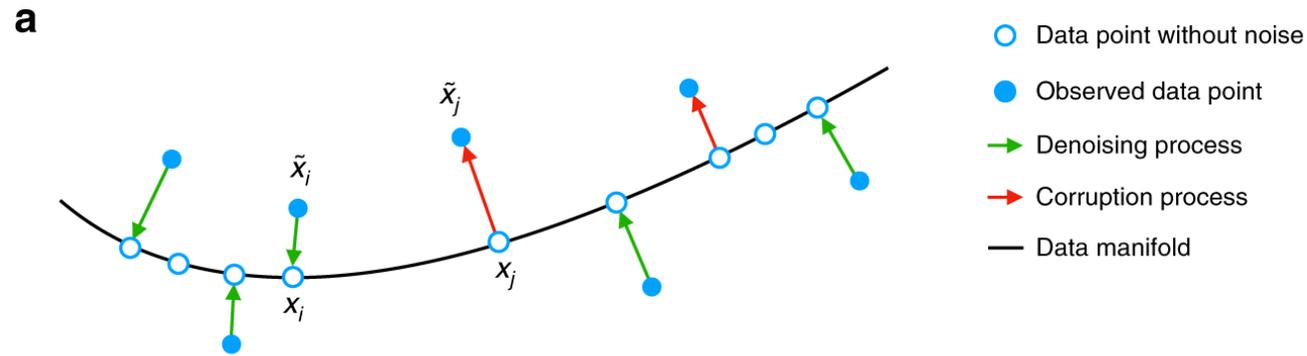


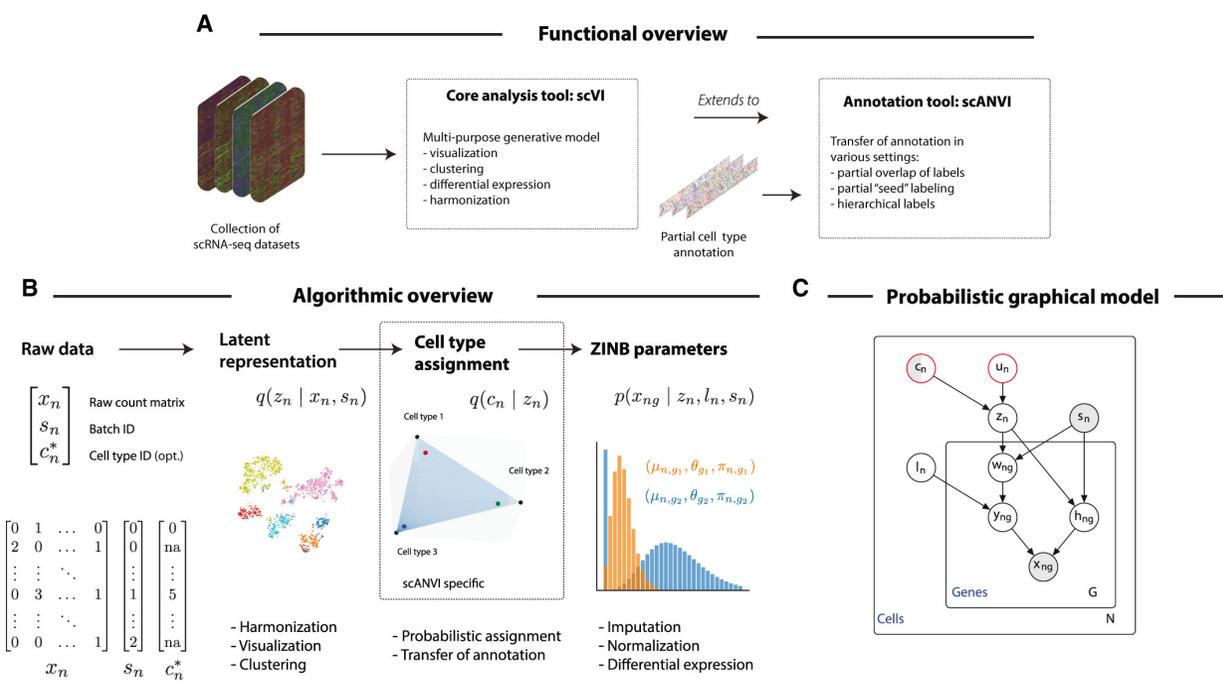
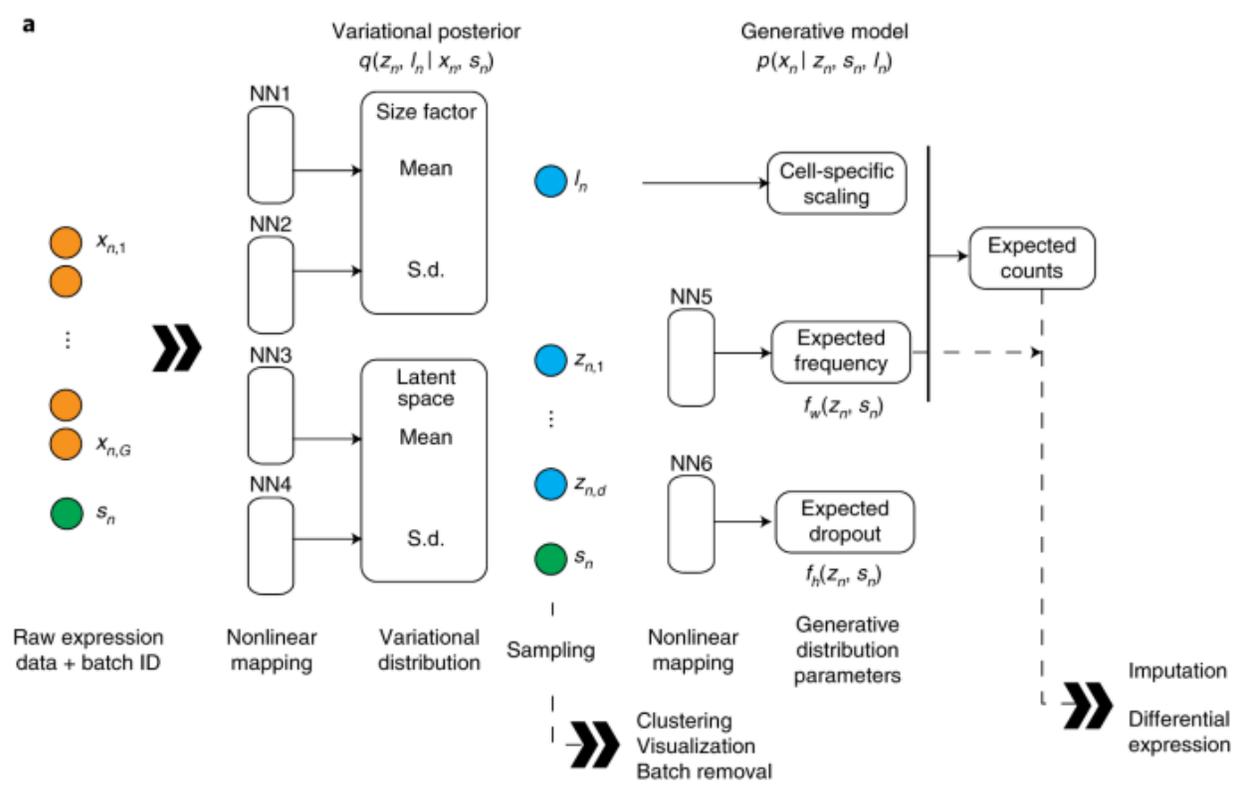
Table 1 | Overview of common data integration methods classified according to their anchor choice

Integration task	Method	Ref.
Vertical (global)	CCA	112
Vertical (global)	JIVE	70
Vertical (global)	PLS	71
Vertical (global)	MCIA	113
Vertical (global)	MOFA+	65
Vertical (global)	scAI	114
Vertical (global)	iNMF	38
Vertical (global)	Seurat v4	11
Vertical (local)	Spearman's rank correlation coefficient	50
Vertical (local)	LMM	51
Horizontal	MNN	21
Horizontal	Seurat v3	22
Horizontal	LIGER	23
Horizontal	Harmony	24
Horizontal	Scanorama	29
Horizontal	BBKNN	25
Horizontal	scVI	26
Horizontal	scmap	28
Horizontal	conos	27
Diagonal	MATCHER	77
Diagonal	MMD-MMA	78
Diagonal	SCIM	115
Diagonal	UnionCom	116
Diagonal	coupledNMF	117

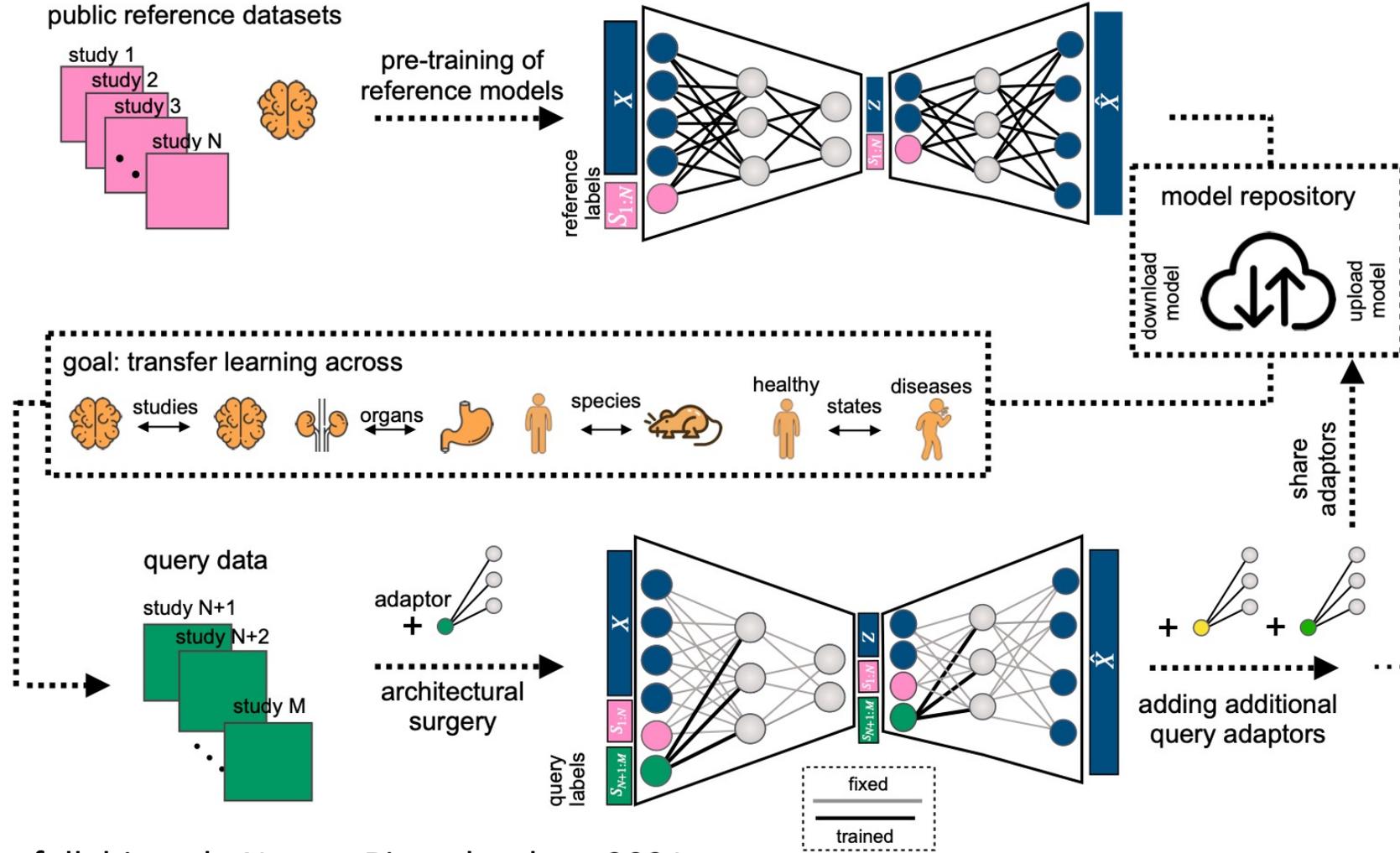
Deep count autoencoder : Denoising



scVI et scANVI : Intégration et transfert d'annotation par variational auto-encoder

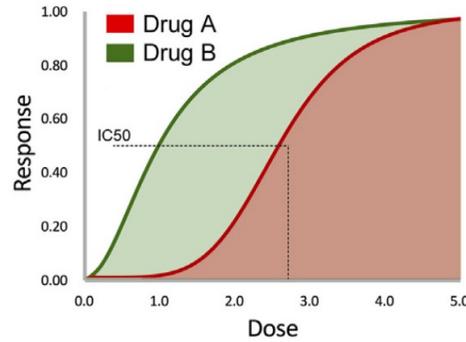
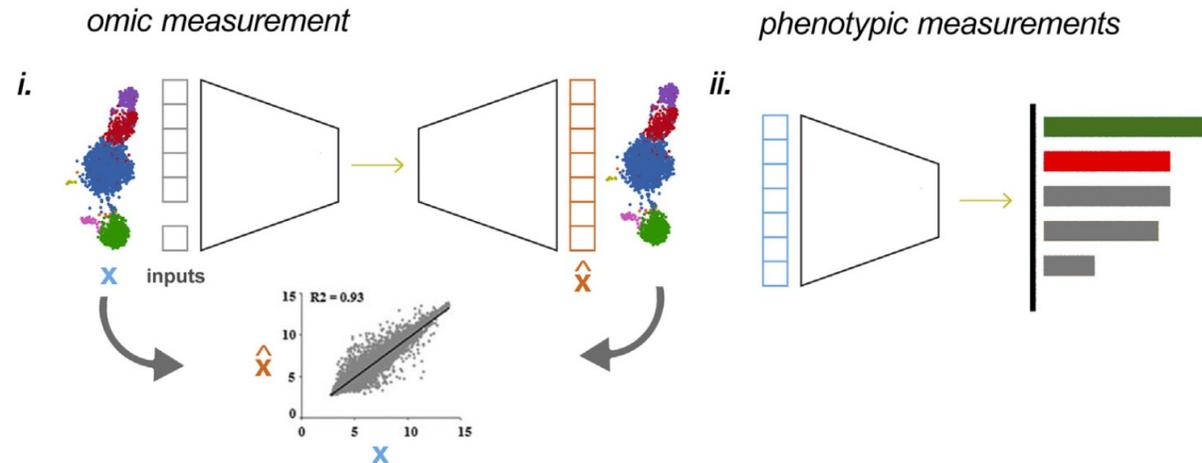


Transfer Learning : scArches

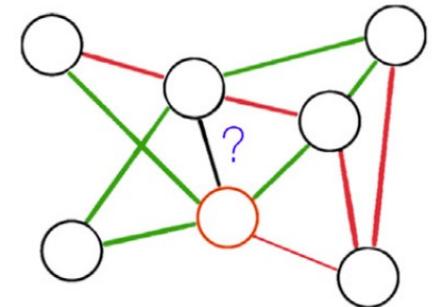
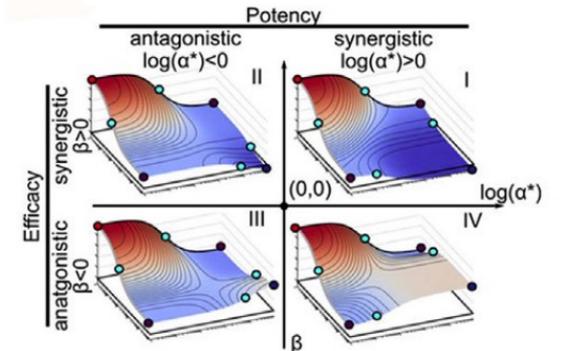
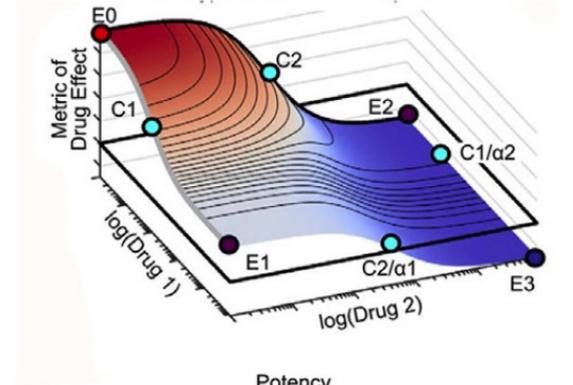


Perturbation map

A Perturbation response prediction



C Perturbation interaction prediction



Machine Learning for perturbational single-cell omics
 Yuge et al., Cell Systems, 2021

Graph neural network



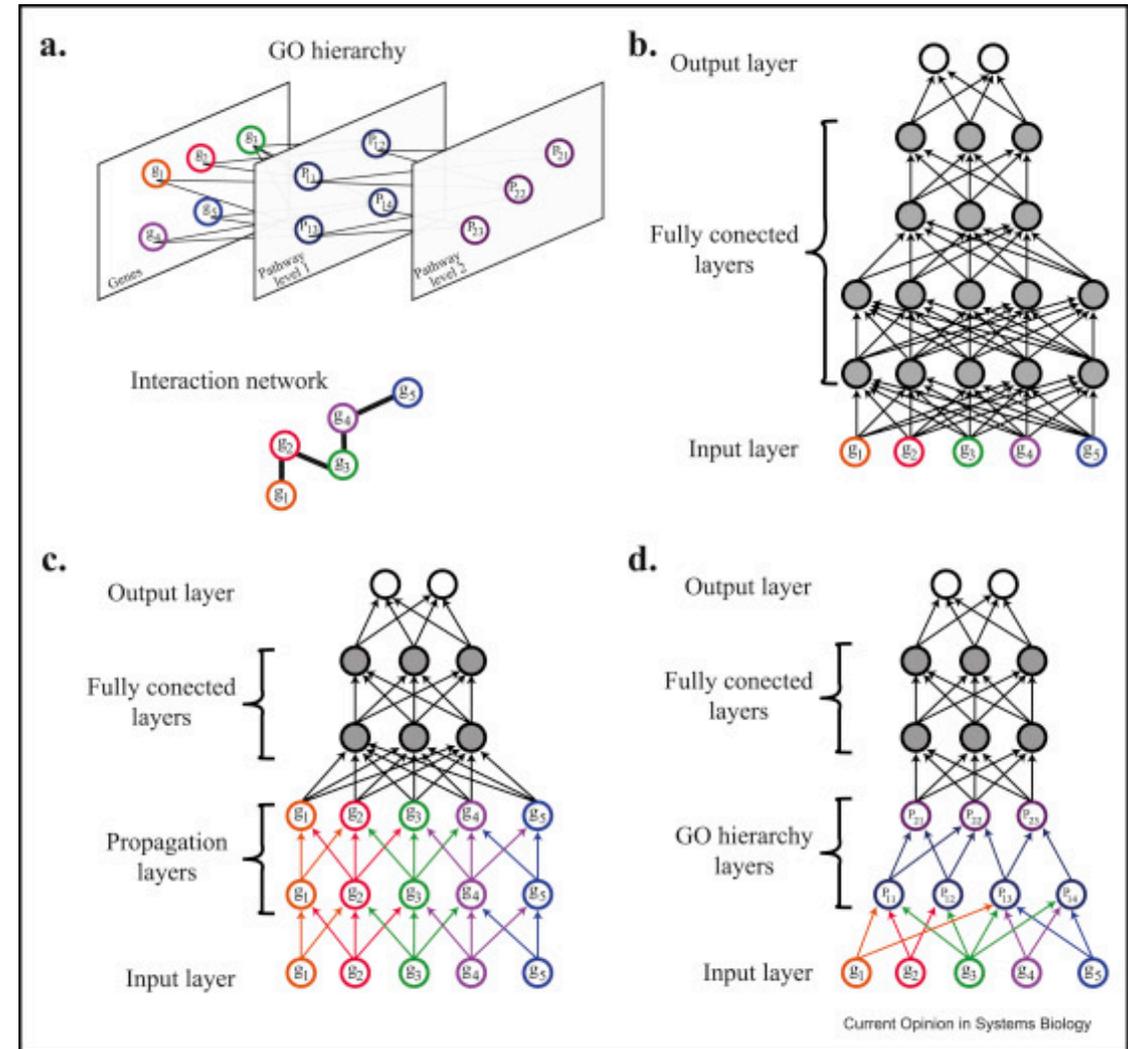
Current Opinion in Systems Biology

Volume 15, June 2019, Pages 68-73

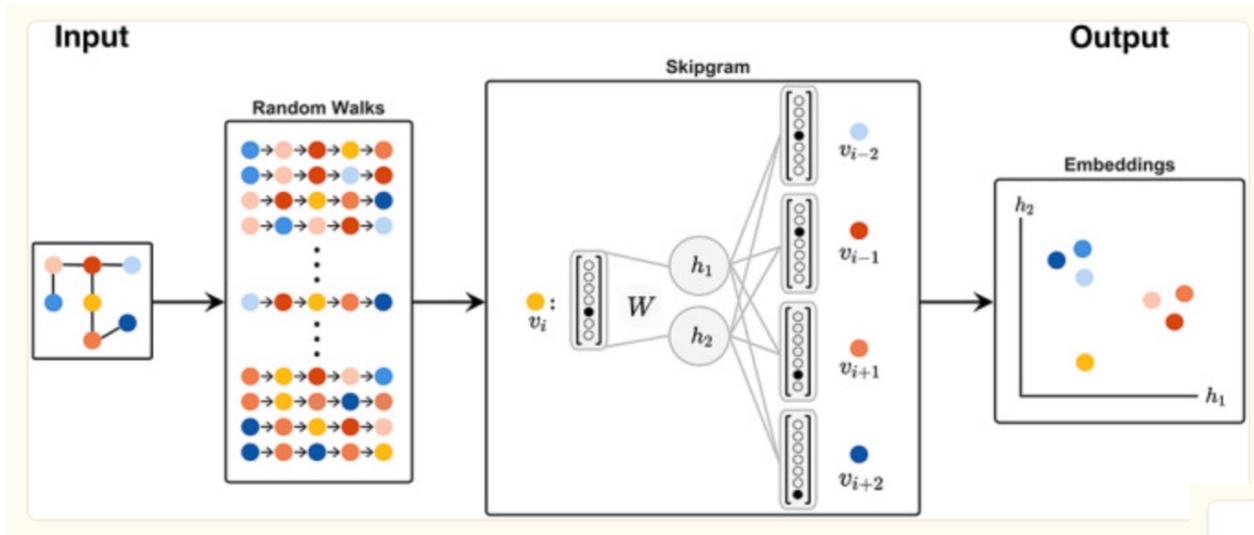


From genotype to phenotype: augmenting deep learning with networks and systems biology

Vahid H. Gazestani^{1,2,3}, Nathan E. Lewis^{1,3,4}  

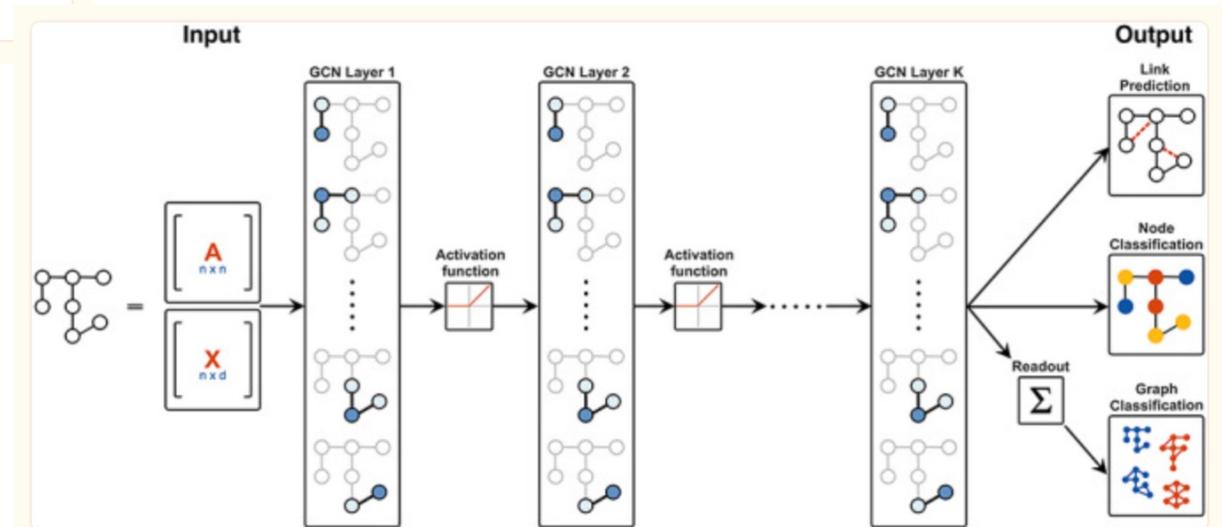


Graph Neural Network



DeepWalks

*Biological Network Analysis with Deep Learning,
Muzio et al., Briefings in Bioinformatics, 2021*



K-layer GCN

Physics informed neural networks

Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations

M. Raissi¹, P. Perdikaris² and G.E. Karniadakis¹

¹*Division of Applied Mathematics, Brown University, Providence, RI, 02912, USA*

²*Department of Mechanical Engineering and Applied Mechanics, University of Pennsylvania, Philadelphia, PA, 19104, USA*

SIAM REVIEW
Vol. 63, No. 1, pp. 208–228

© 2021 Society for Industrial and Applied Mathematics

DeepXDE: A Deep Learning Library for Solving Differential Equations*

On optimise un réseau de neurone $u(t, x)$ dont certains paramètres SONT les paramètres d'un système de PDE $f(t, x)$

$$f := u_t + \mathcal{N}[u],$$

La loss est une loss combinée

$$MSE = MSE_u + MSE_f,$$

where

$$MSE_u = \frac{1}{N_u} \sum_{i=1}^{N_u} |u(t_u^i, x_u^i) - u^i|^2,$$

and

$$MSE_f = \frac{1}{N_f} \sum_{i=1}^{N_f} |f(t_f^i, x_f^i)|^2.$$

« Encoder » le système d'équation différentiel dans la loss

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, t; \mathbf{p}),$$

$$\mathbf{x}(T_0) = \mathbf{x}_0,$$

$$\mathbf{y} = \mathbf{h}(\mathbf{x}) + \epsilon(t), \quad \epsilon(t) \sim \mathcal{N}(0, \sigma^2),$$

Systems biology informed deep learning for inferring parameters and hidden dynamics, Yazdani et al., Plos Computational Biology, 2021

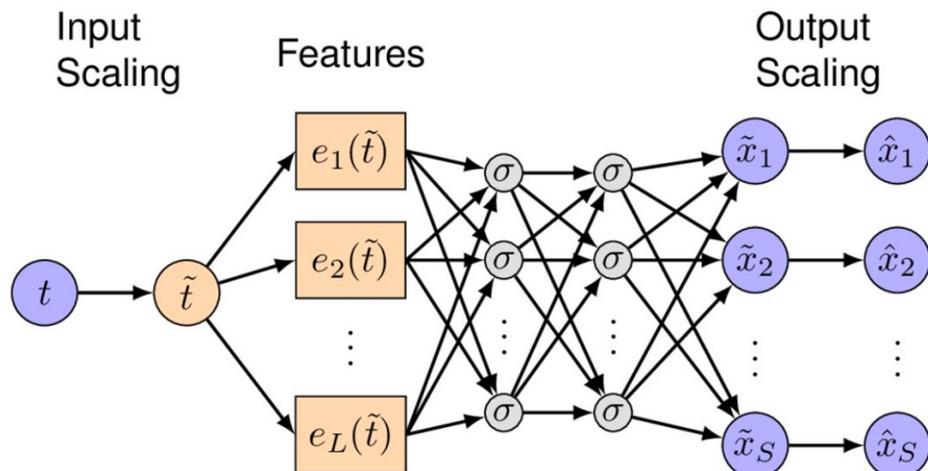
$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{p}) = \mathcal{L}^{data}(\boldsymbol{\theta}) + \mathcal{L}^{ode}(\boldsymbol{\theta}, \mathbf{p}) + \mathcal{L}^{aux}(\boldsymbol{\theta}), \quad (3)$$

where

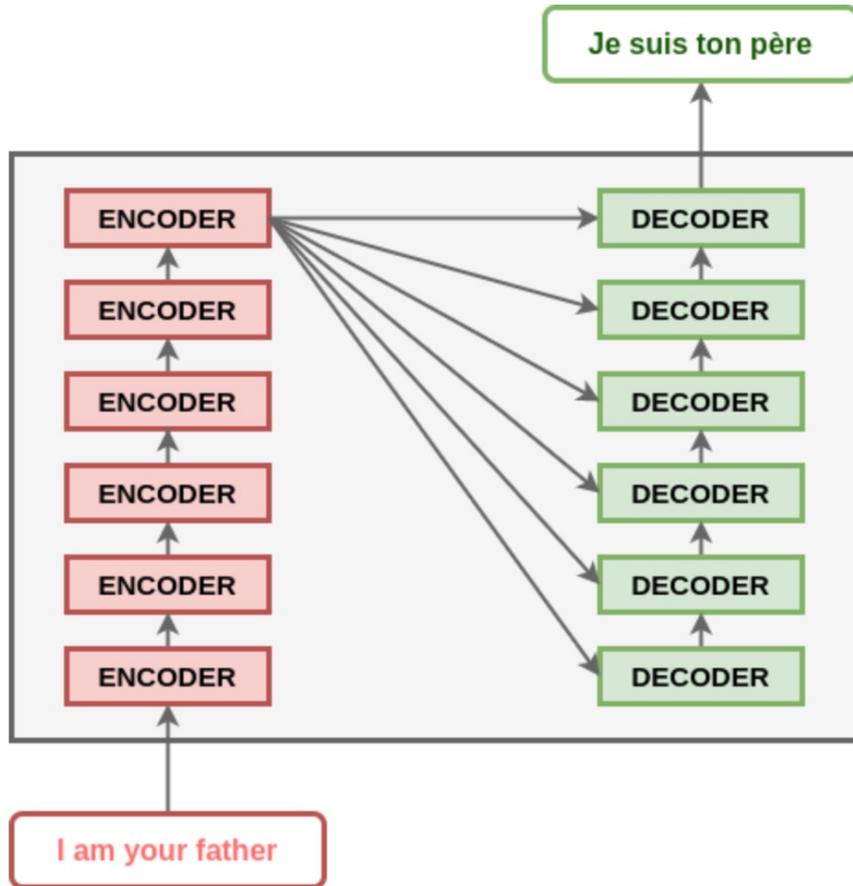
$$\mathcal{L}^{data}(\boldsymbol{\theta}) = \sum_{m=1}^M w_m^{data} \mathcal{L}_m^{data} = \sum_{m=1}^M w_m^{data} \left[\frac{1}{N^{data}} \sum_{n=1}^{N^{data}} (y_m(t_n) - \hat{x}_{s_m}(t_n; \boldsymbol{\theta}))^2 \right], \quad (4)$$

$$\mathcal{L}^{ode}(\boldsymbol{\theta}, \mathbf{p}) = \sum_{s=1}^S w_s^{ode} \mathcal{L}_s^{ode} = \sum_{s=1}^S w_s^{ode} \left[\frac{1}{N^{ode}} \sum_{n=1}^{N^{ode}} \left(\frac{d\hat{x}_s}{dt} \Big|_{\tau_n} - f_s(\hat{x}_s(\tau_n; \boldsymbol{\theta}), \tau_n; \mathbf{p}) \right)^2 \right], \quad (5)$$

$$\mathcal{L}^{aux}(\boldsymbol{\theta}) = \sum_{s=1}^S w_s^{aux} \mathcal{L}_s^{aux} = \sum_{s=1}^S w_s^{aux} \frac{(x_s(T_0) - \hat{x}_s(T_0; \boldsymbol{\theta}))^2 + (x_s(T_1) - \hat{x}_s(T_1; \boldsymbol{\theta}))^2}{2}. \quad (6)$$



Game changer : Le Transformer



Architecture du Transformer

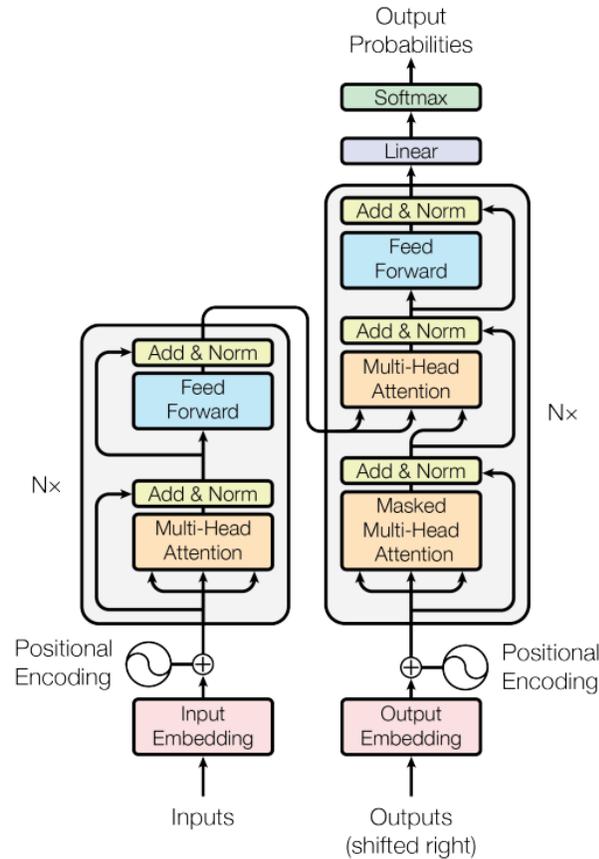
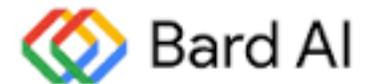


Figure 1: The Transformer - model architecture.

Attention is all you need
Vaswani et al., ArXiv, 2017

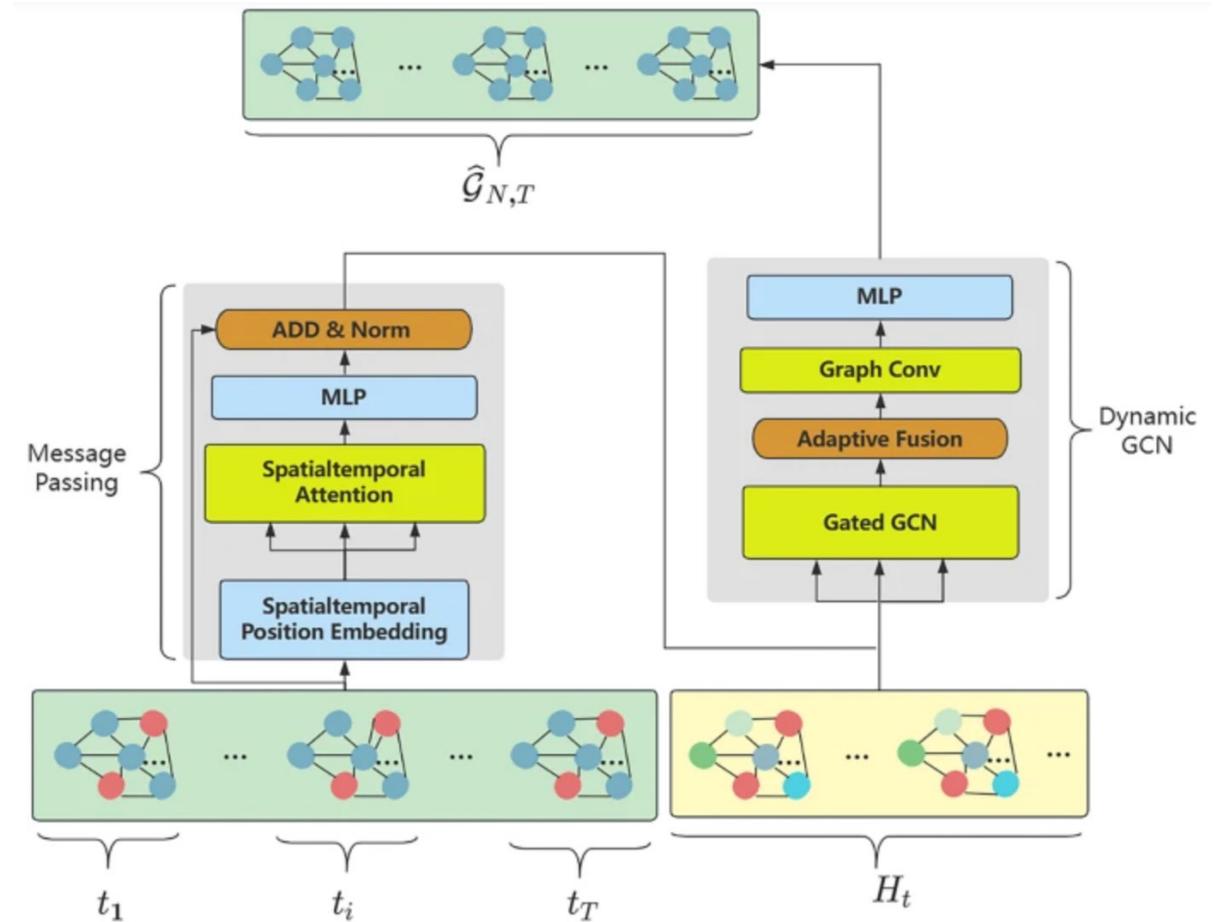
Large Language model



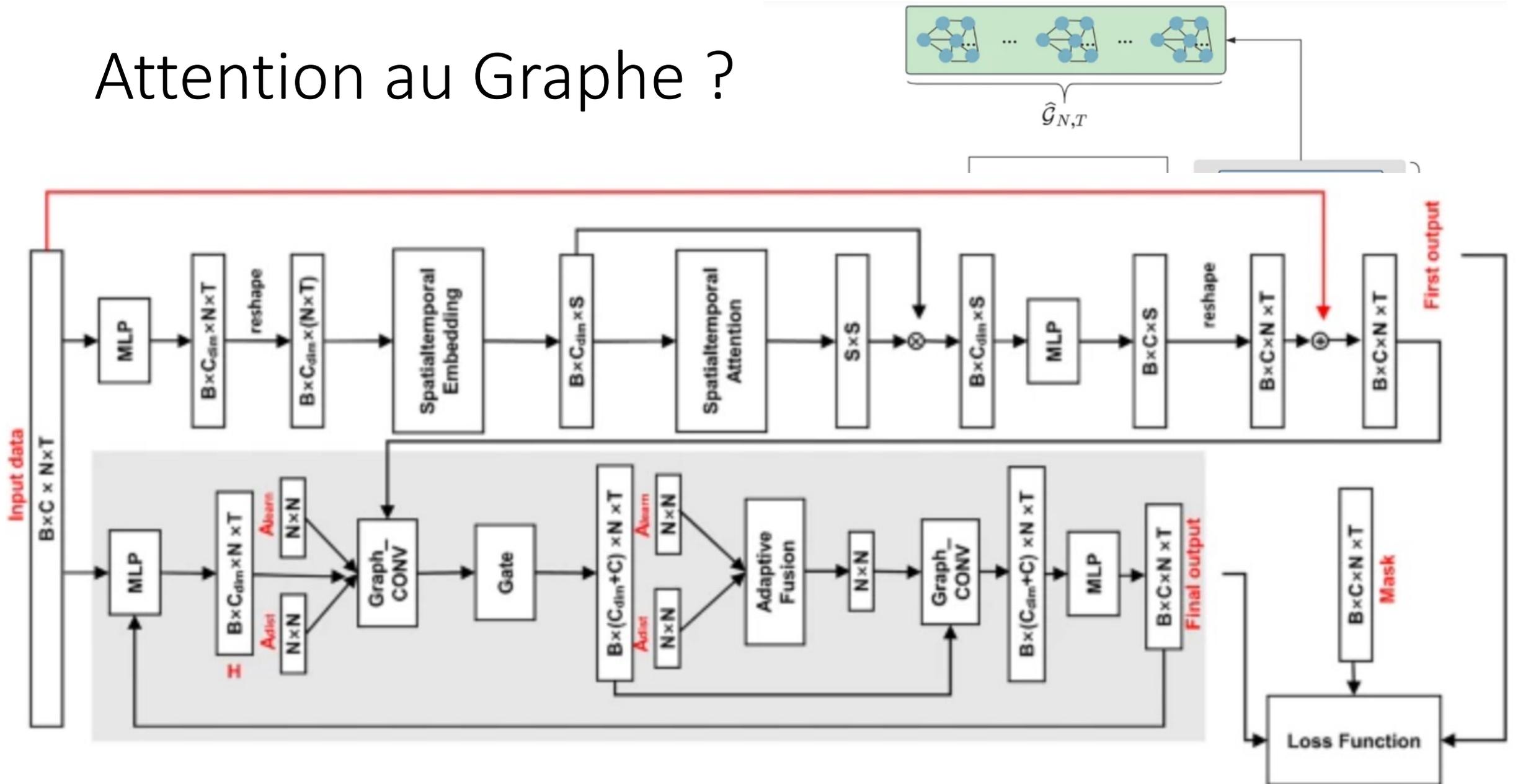
Attention au Graphe ?

Attention based message passing and dynamic graph convolution for spatiotemporal data imputation, Wang et al., Scientific Reports,

7 April 2023



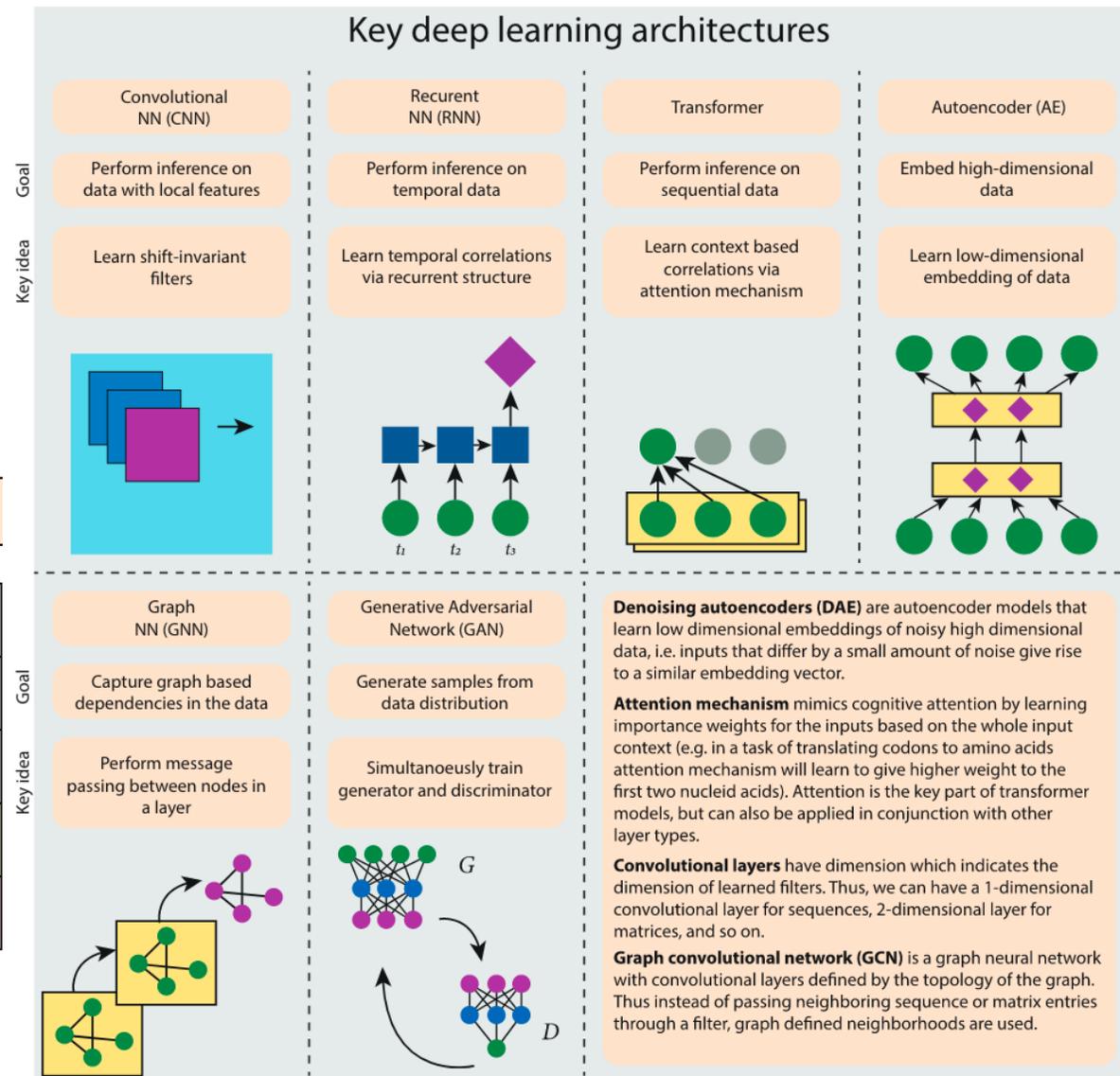
Attention au Graphe ?



Systems Biology et Deep Learning

Table 1 Impact of Deep Learning on Computational Biology.

	Protein structure prediction	Protein function prediction	Genome engineering	Systems biology and data integration	Phylogenetic inference
Paradigm shifting	✓				
Major success		✓	✓		
Moderate success				✓	
Minor success					✓



Current Progress and open challenges ifor applying deep learning across the biosciences, Sapoval et al., Nature Comm. 2022