Reasoning over large-scale biological systems with heterogeneous and incomplete data

Anne SIEGEL, CNRS, Rennes

Dyliss team (Inria, Univ Rennes, CNRS)

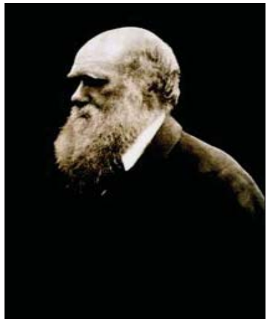
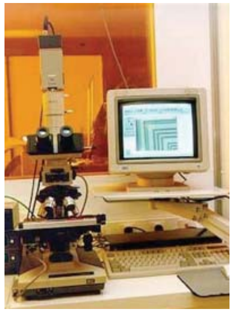
Institut de Recherche en Informatique et Systèmes Aléatoires

Short presentation

- **Reasearch field**
  - Discrete dynamical systems & fractals
  - Systems biology
  - Knowledge representation
- **IRISA & INRIA Rennes**
  - 800 members, >40 teams
  - Univ Rennes, CNRS, Inria, etc…
- **Bioinformatics@Rennes**
  - GenOuest: plateform, ressource center
  - Genscale : NGS data analysis
  - **Dyliss: Integration of heterogeneous data**

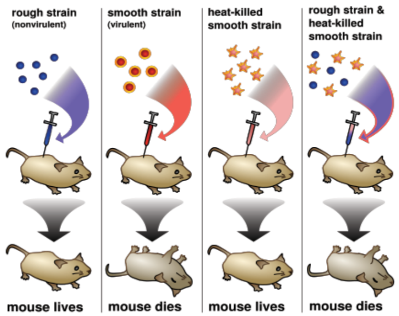
LIFE SCIENCE DATA

From life science… to data science

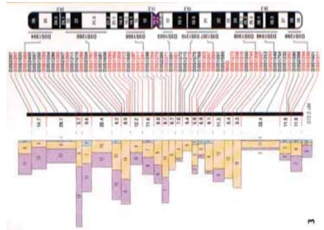
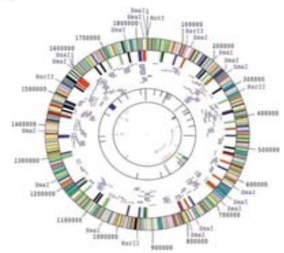

Naturalist approach

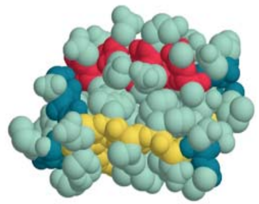
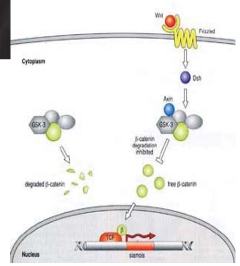
- Observing and deducing

Experimental approach

- Perturbating and observing

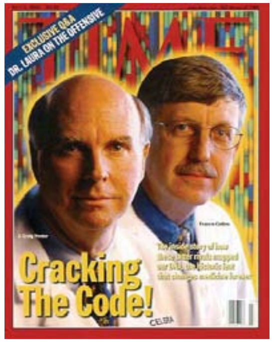



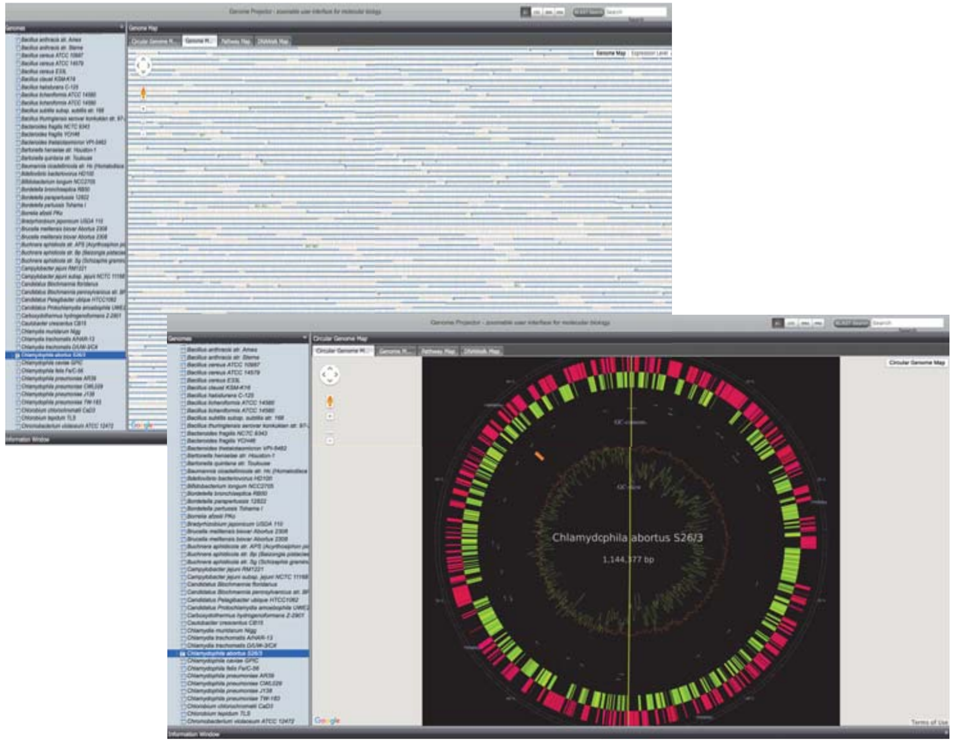
Modern biology

- Measuring at lower scales

→ **Data science !**

Biomolecular data: genomes





Genome sequencing

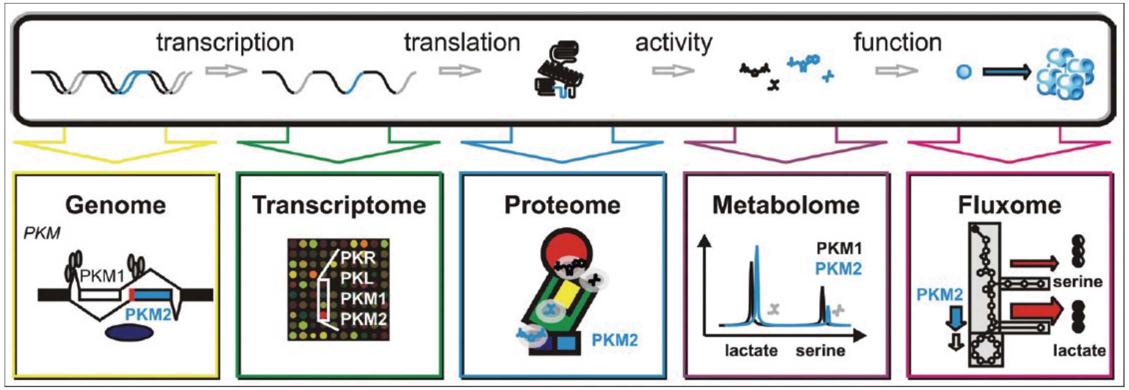
- Very smart computational issues
- Bioinformatics

Thousands of publicly available genomes

- Exploration, mapping and analysis

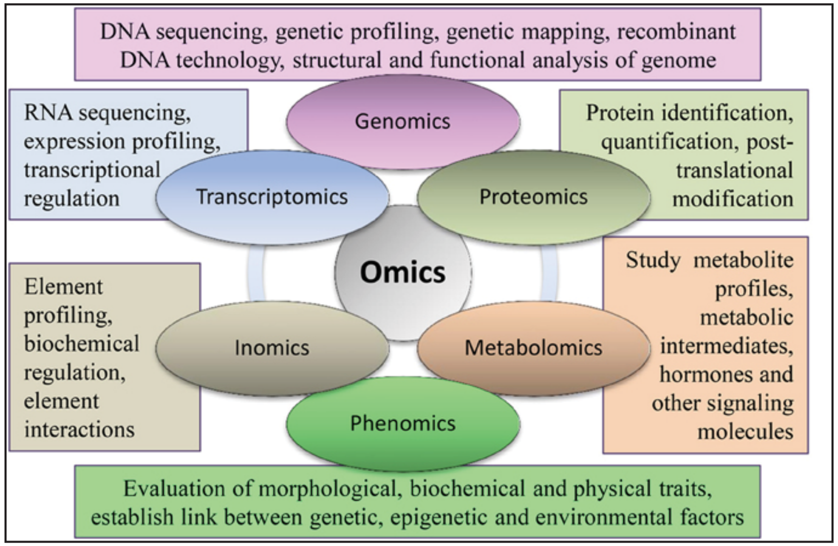
http://www.g-language.org/g3/

What do we do with genomic data ?



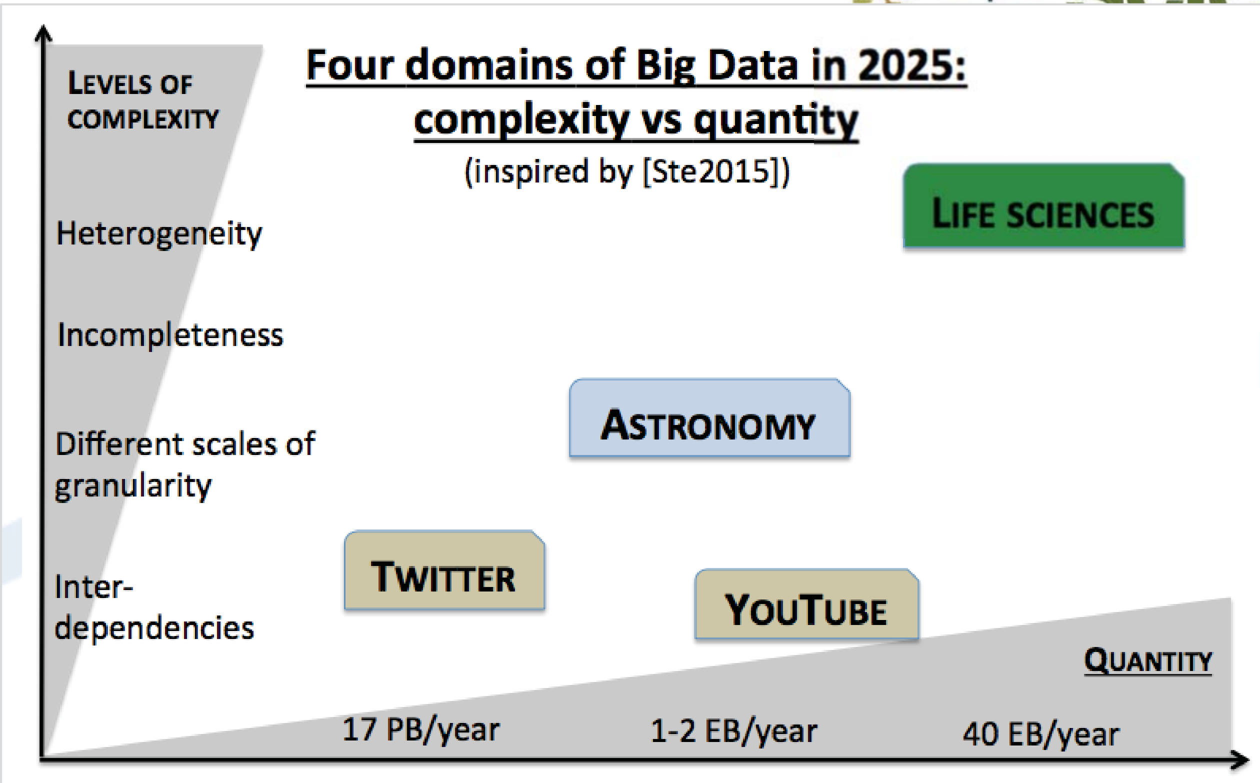
Assign a function to each DNA fragment

Develop new technologies to validate/refine the assigned functions



Data deluge !

Life science data nightmare



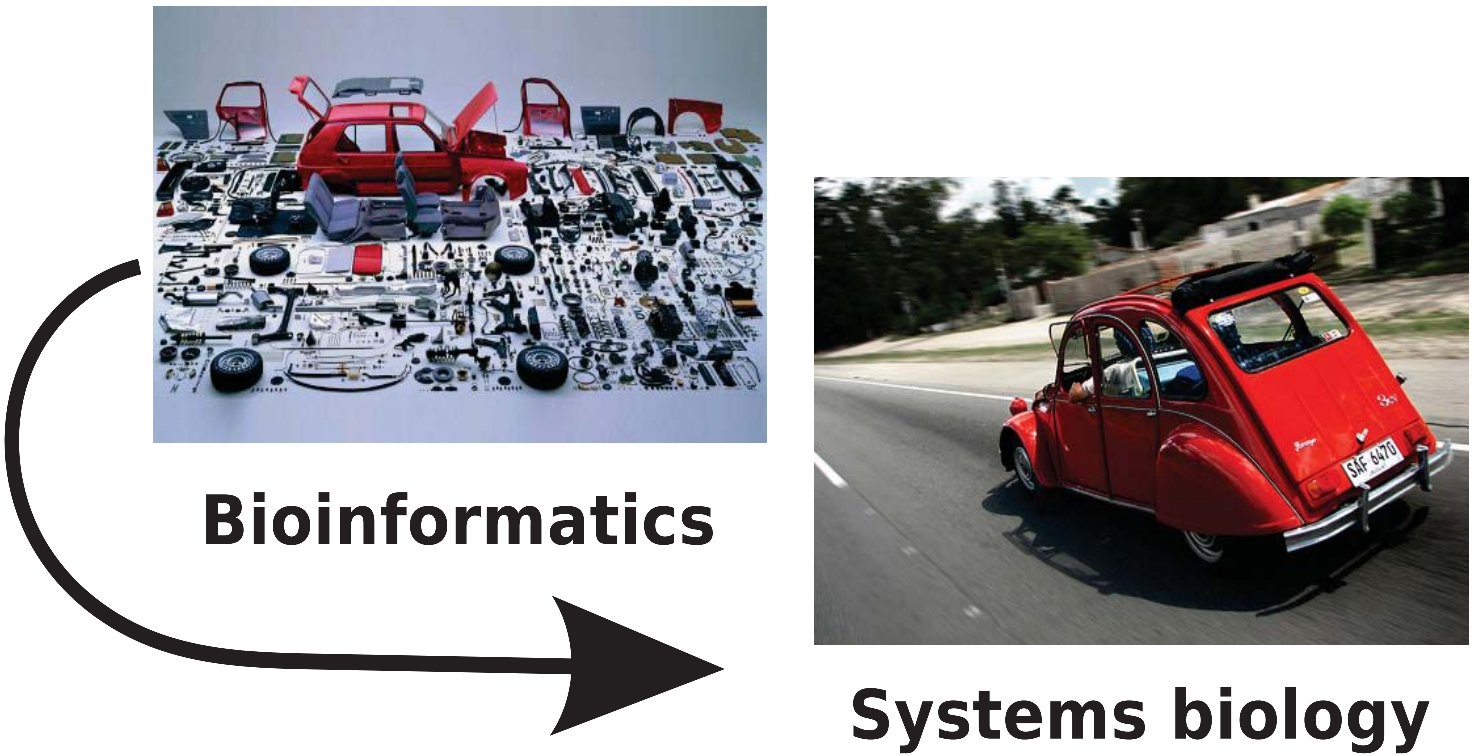
Data characteristics

- Large-scale
- Incomplete
- Inter-dependent
- Heterogeneous / multi-scale

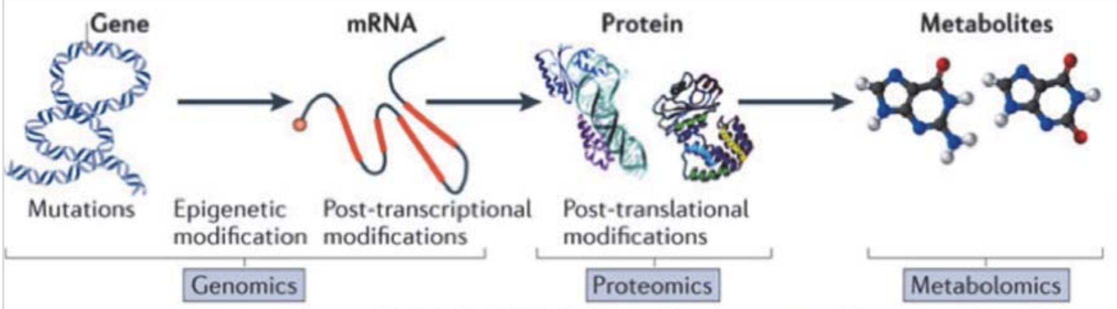
How to integrate them?

IRISA

SYSTEMS BIOLOGY



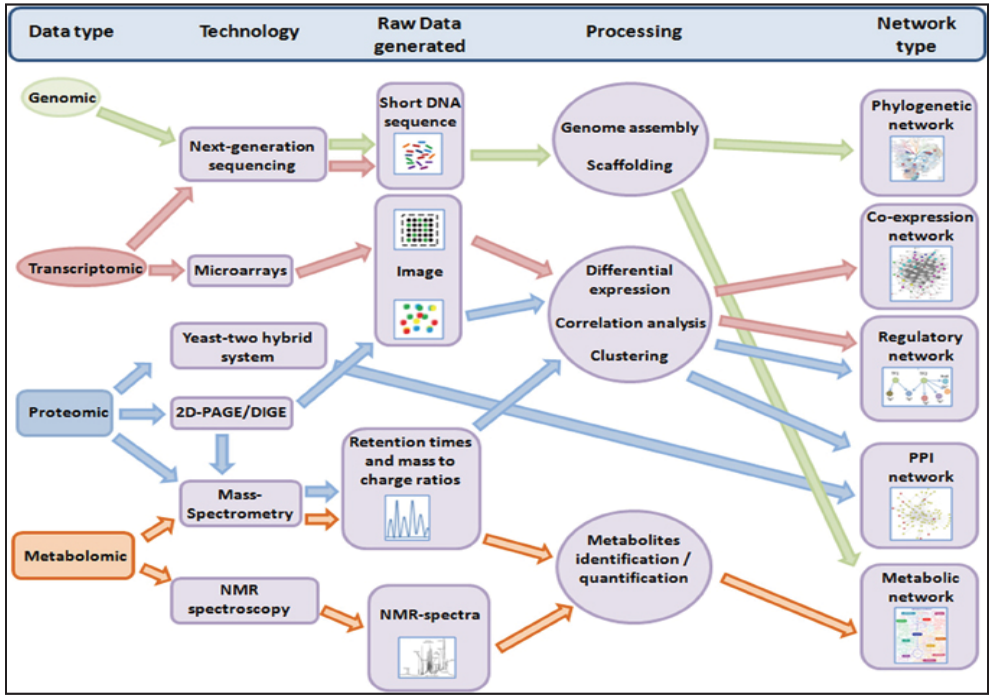
Setting all together



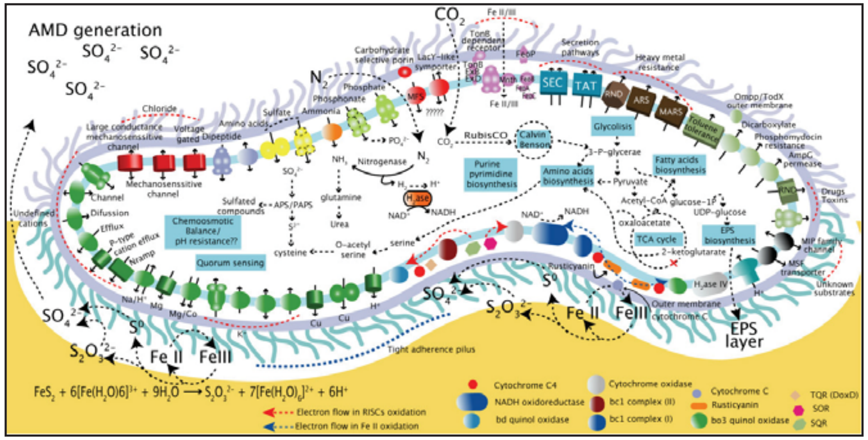
Patti et al. (2012). Metabolomics: the apogee of the omics trilogy. Nature

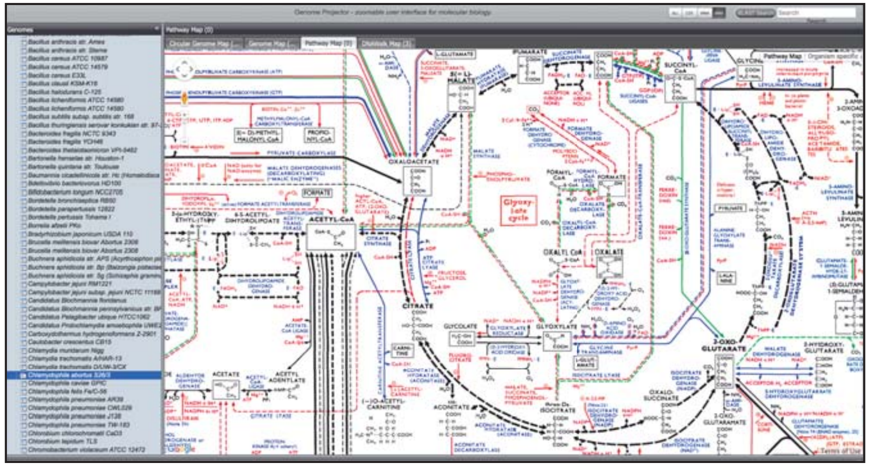
Gene function = regulation of a intra-cellular transformation procedure

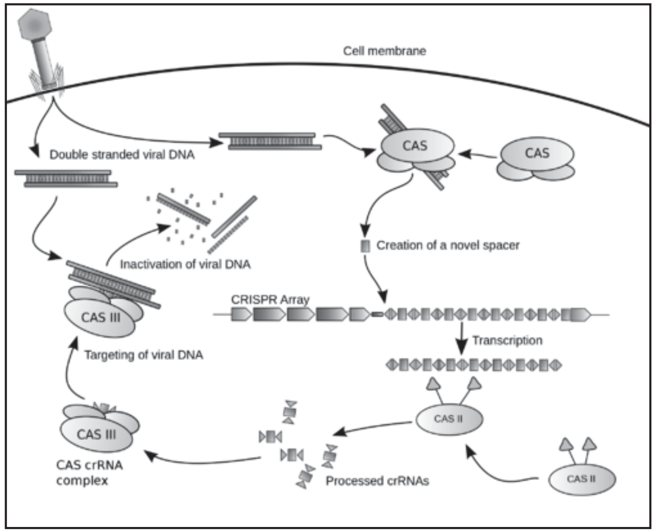
- Biological interactions !
- Graphs / networks

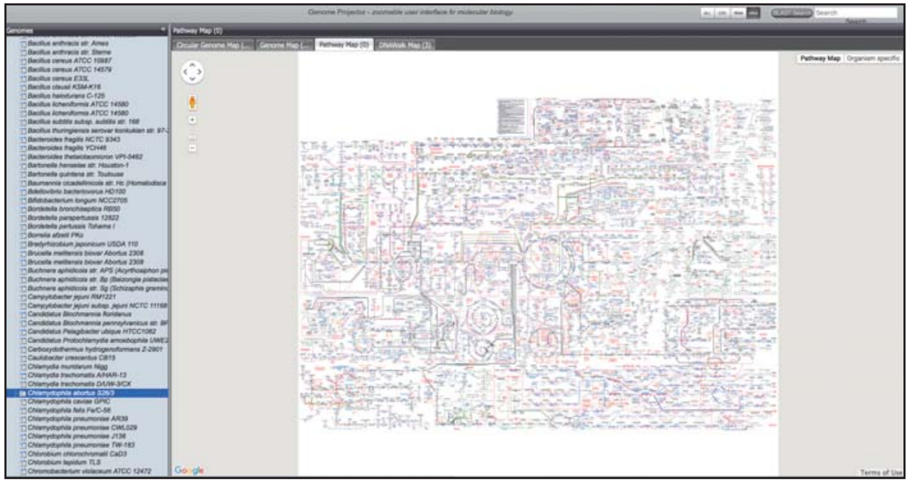


What we get…







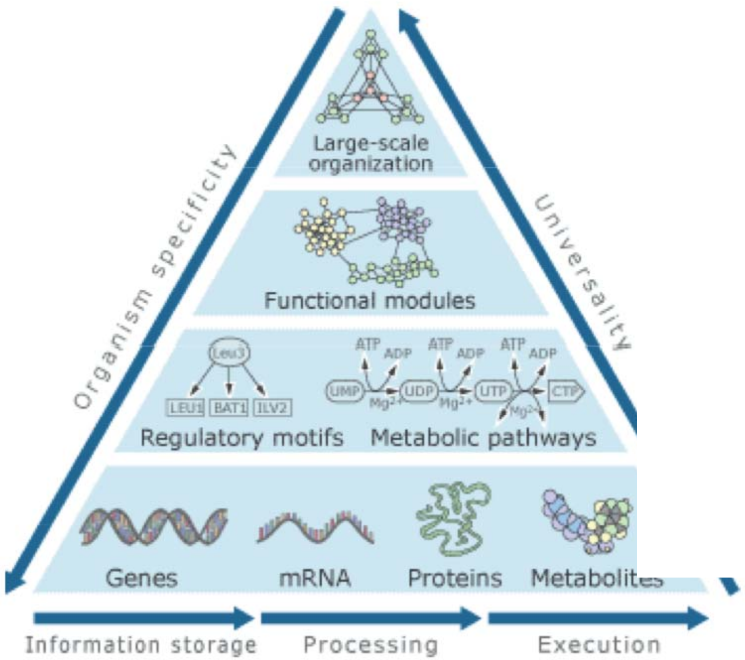


Large-scale graph description of interactions between compounds

Systems biology

Statement : **biology is a complex system**

- *« Requires to examine the structure and dynamics of a cellular function rather than the characteristics of isolated parts of a cell »* *(Kitano, 2002)*

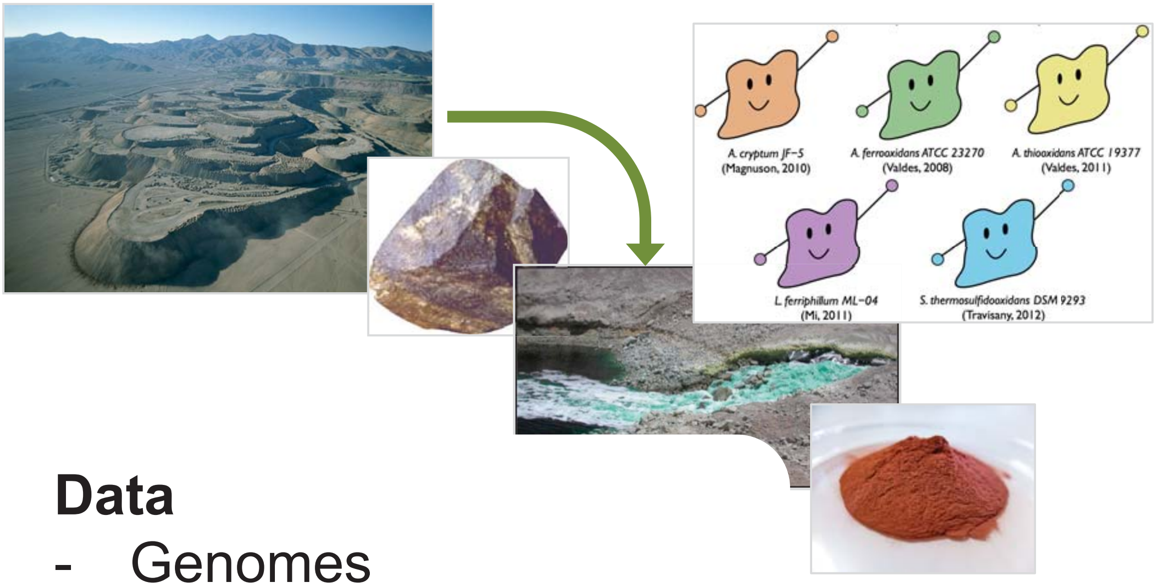


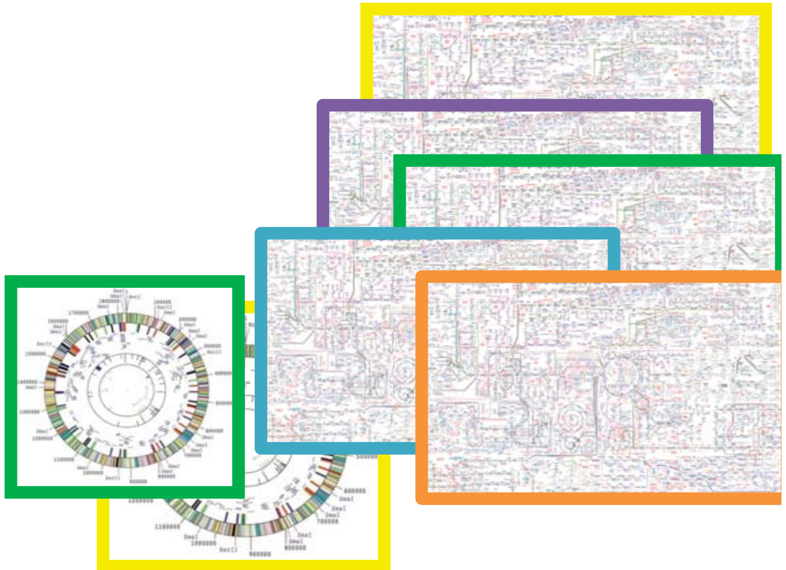
Systems biology: **Interpreting multi-layer data and graphs**

- Produce predictive statements that can be experimentally validated

Case-study: extremophile mining consortium

Role of an ***empirical taylor-made consortium*** *of bacteria in copper extraction from ore ?*





Data

- Genomes
- Expression data
- Metabolic compounds

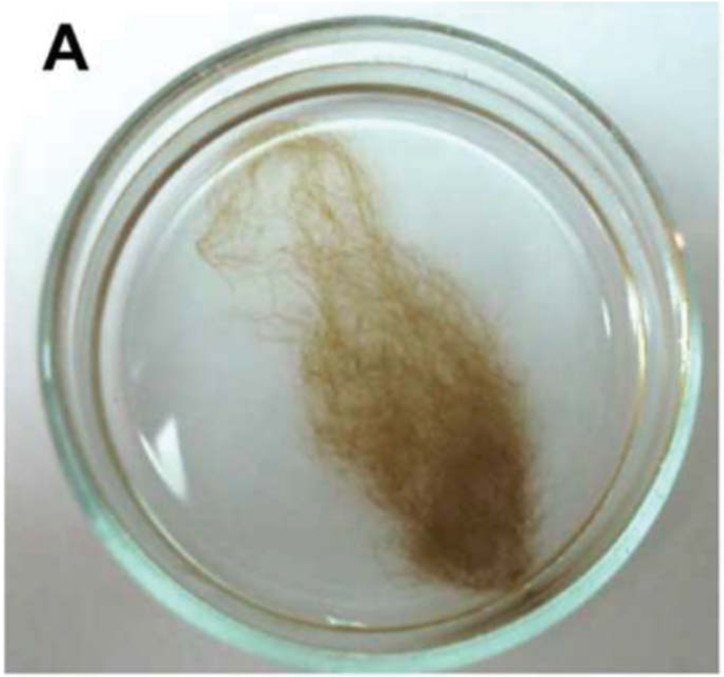
Turn data into

- genomics maps
- interaction maps

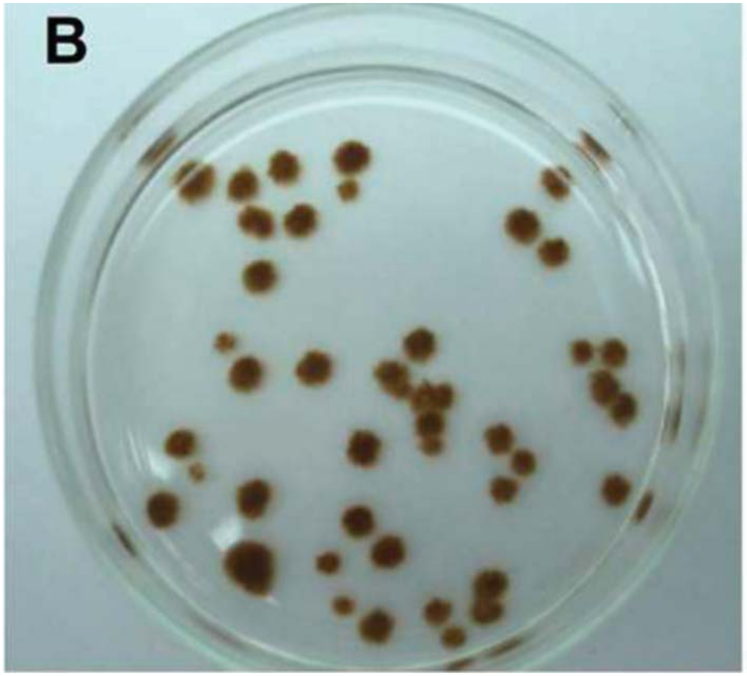
Understand the contribution of each bacteria to the complete system ?

➢ **integrative and systems biology**

A second case-study : algal metabolism



E. siliculosus

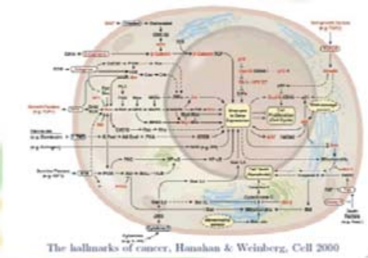


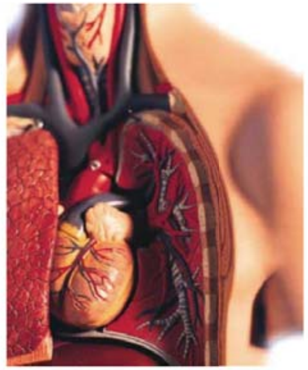
In axenic condition….

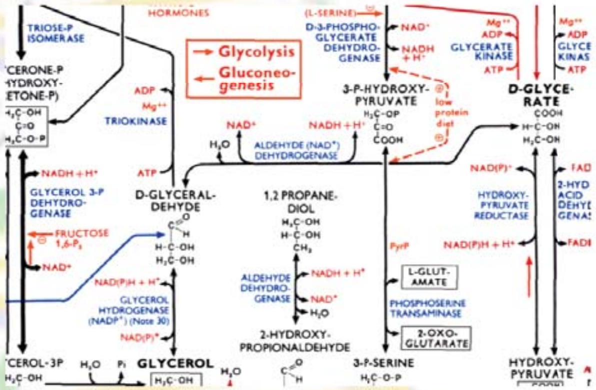
Ectocarpus
[Dittami2014, Tapia2016]

What is the role of environmental bacteria ?

Complex systems are everywhere







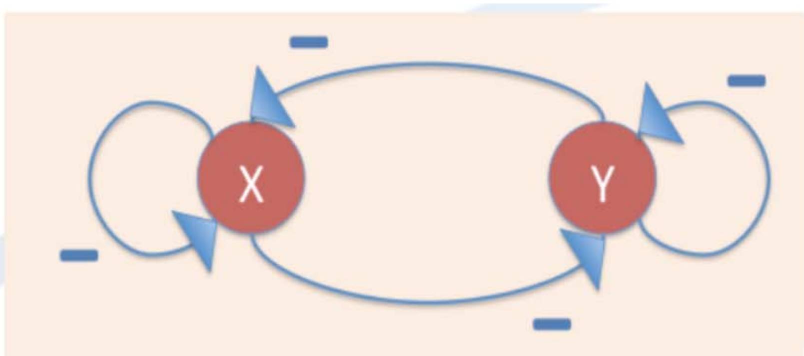
Are molecular/cellular different than others ?

Dynamical systems

Historical motivation

Model the evolution of the set of components in a system according to time.

$$
\begin{array}{lccccc}
F: & \mathbb{T} & \times & \mathbb{S} & \rightarrow & \mathbb{S} \\
 & (t & , & \mathbf{z}) & \mapsto & F(t, \mathbf{z}) \\
 & (\text{time} & , & \text{state}) & & \text{new state at time } t
\end{array}
$$



$$\frac{dX}{dt} = \frac{k}{K + Y^n} - aX$$

$$\frac{dY}{dt} = \frac{l}{L + X^n} - bY$$

Parameterized numerical system

$$f(X) \leftarrow 1 - Y$$

$$f(Y) \leftarrow 1 - X$$

Boolean model with asynchronous update scheme

Identification of a dynamical system

Find the **best function F** which parcimounously explains and describes the observed responses of a system.

Model identification since the 18th century

What has always allowed a model identification

- **A priori knowledge about the laws governing the system**
  - Predetermined shape for the function F
- **Limited number of components**
  - Reduction of the search space
- **Wide panel of sensors and perturbations**
  - Discriminate parameters

$$F : \begin{array}{ccccccc} & \mathbb{T} & \times & \mathbb{S} & \to & \mathbb{S} \\ & (t & , & \mathbf{z}) & \mapsto & F(t,\mathbf{z}) \\ & (\text{time} & , & \text{state}) & & \text{new state at time } t \end{array}$$

Where is the complexity ?

- The search space grows exponentially with the number of measured compounds

→ **The more compounds we measure, the less identifiable a system is.**

Differences between application domains

Physical sciences

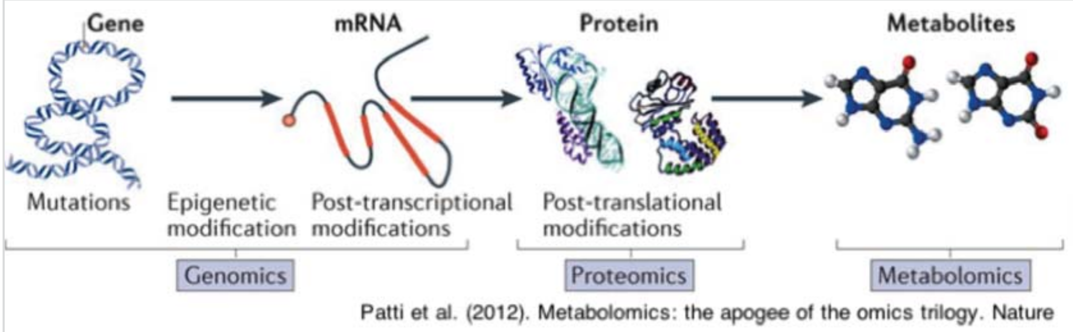
- **Knowledge.**
  Fundamental laws of physics.
- **Sensors.**
  Numerous.
- **Perturbations.**
  Various protocoles in controled frameworks.
- **System description.**
  Independent components

Biological sciences

- **Knowledge.**
  Empirical laws
- **Sensors.**
  Low quality although numerous.
- **Perturbations.**
  Quite few according to sensors
- **System description.**
  Hidden dependencies

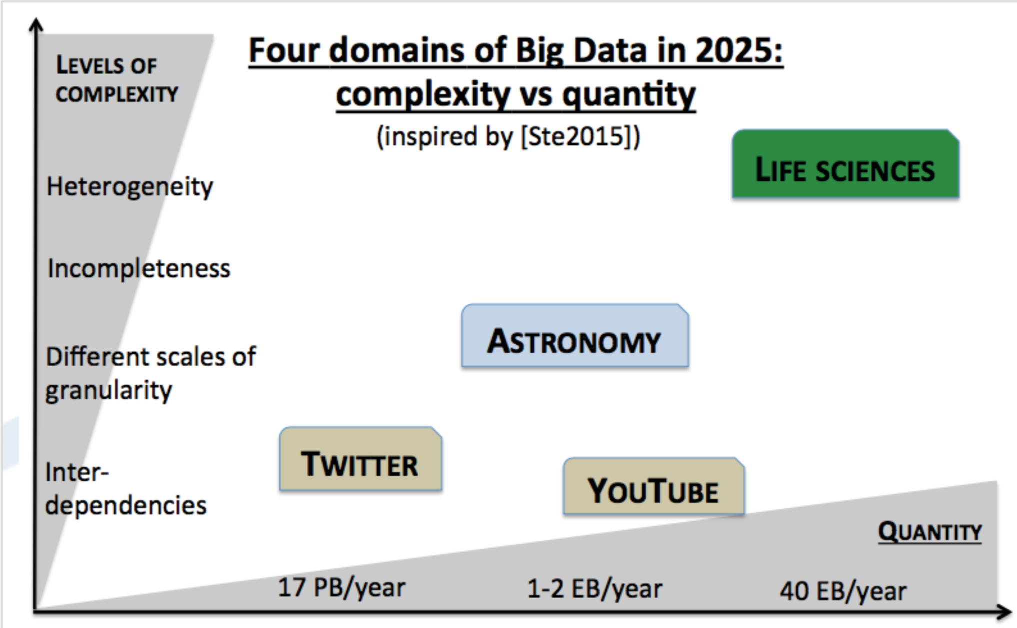
Today's molecular/cellular biological systems

Omics data.
- Large-scale
- Noisy
- Heterogeneous.



Patti et al. (2012). Metabolomics: the apogee of the omics trilogy. Nature

Biological systems characteristics
- Large-scale
- Empirical laws
- Few data wrt the search space size



Biological systems observed with omics data are not uniquely identifiable

IRISA

Strategy: combine dynamical systems and constraints programming

Describe a system by a family of abstract models

- Reason over a family of models instead of selecting a single one

(Logical) knowledge representation

- Search space **description**
- Structured knowledge (link open data)

Discrete dynamical systems

- **Links** between multi-scale observations.
- **Invariants** of model families.

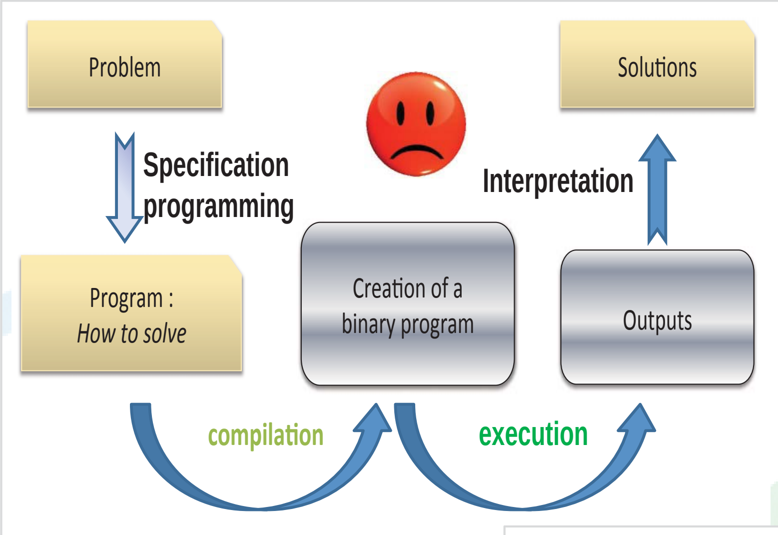
Solving optimisation problems

- **Replace laws by constraints**
- Extract robust information

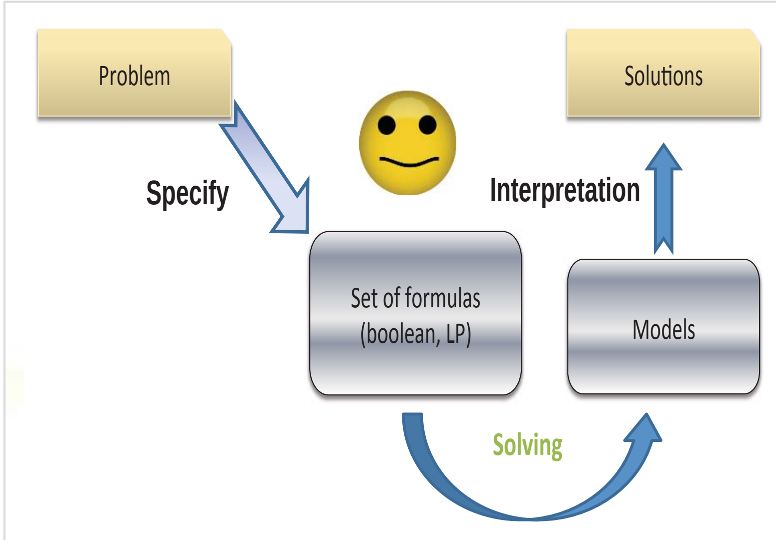
KNOWLEDGE REPRESENTATION

```
1{murderer(ms_Scarlet); murderer(colonel_Mustard)}1.
1{weapon_of_crime(revolver); weapon_of_crime(candlestick)}1.
1{place_of_crime(kitchen); place_of_crime(hall);
                              place_of_crime(dining_room)}1.

crim_record(ms_Scarlet,7). crim_record(colonel_Mustard,4).

weapon_of_crime(candlestick).
:- place_of_crime(kitchen).
place_of_crime(hall) :- murderer(colonel_Mustard), not
                                     weapon_of_crime(revolver).

sol(X,Y,Z) :- murderer(X),weapon_of_crime(Y),place_of_crime(Z).

#maximize{W , sol : sol(X,Y,Z) , crim_record(X,W) , murdered(W)}.

#show sol/3.
```

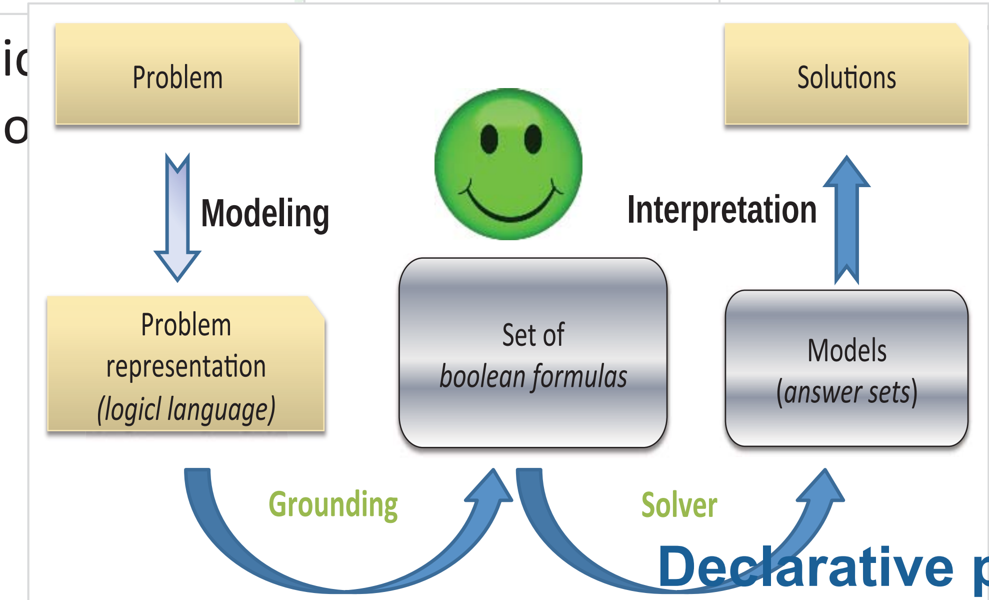
Solving combinatorial problems



Write a program whic
how the problem sho



ite (boolean, linear)
nstraints (*SAT, ILP,…)*



Declarative programing

- **Problem = axioms & rules**
- **No need of algorithm**

Answer set programming.
Describe what you want to solve

ASP logical rules : declarative programming

$$\underbrace{K \ \{ \ atom_1; \ \ldots; \ atom_n \ \} \ L}_{\text{head}} \ \underbrace{:\!-}_{\text{"smiley"}} \ \underbrace{atom_{n+1} \ ; \ldots ; \ atom_r; \ not \ atom_{r+1}; \ \ldots; \ not \ atom_s \ .}_{\text{body}}$$

If **all terms** on the **right side** are true,
then **at least K and at most L** terms are true on the **left side**.

If **nothing** on the **left side**,
then **always false.**

```
:- K{atom1, .. atomN}L.
```

If **nothing** on the **right side**,
then **always true.**

```
K{atom1, .. atomN}L.
```

Optimisation rule

```
#maximize{W,atom(X): condition(X),W}.
```

High-level model language

- Propositional logics
- Model for negation

Highly performant solving technics

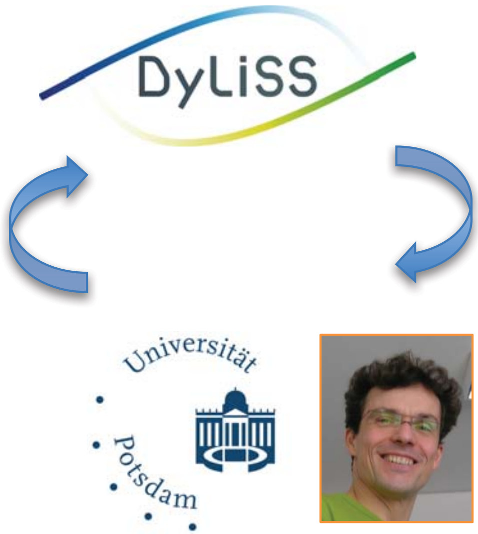
- SAT-based and deductive-DB technics
- Decidable: no infinite loop

Link with systems biology ?

Integrative and systems biology is a very relevant field to challenge ASP technologies

- Repair large-scale interaction graph with **branch and bound** solving heuristics (KR 2010)
- Scale metabolic network completion problem with **unsatisfiable core** solving strategy (LPNMR 2013)
- Design experiments with **incremental solving** (Frontiers 2015)
- Implement and benchmark **constrains propagators** (TPLP 2018)

Problem statement
& modelling



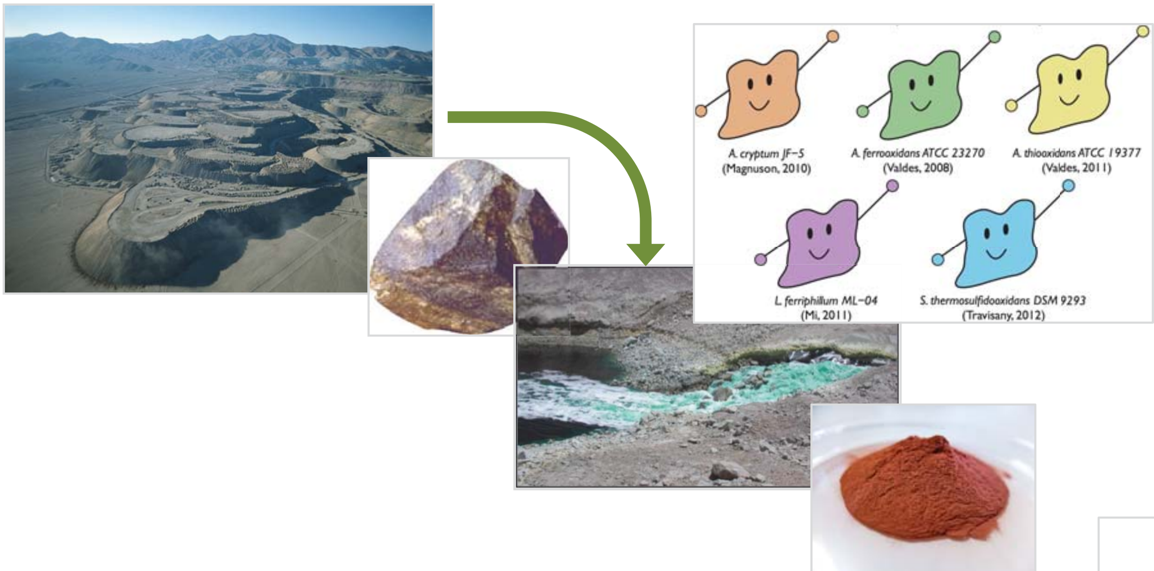
Solving heuristics
& problem reformulation

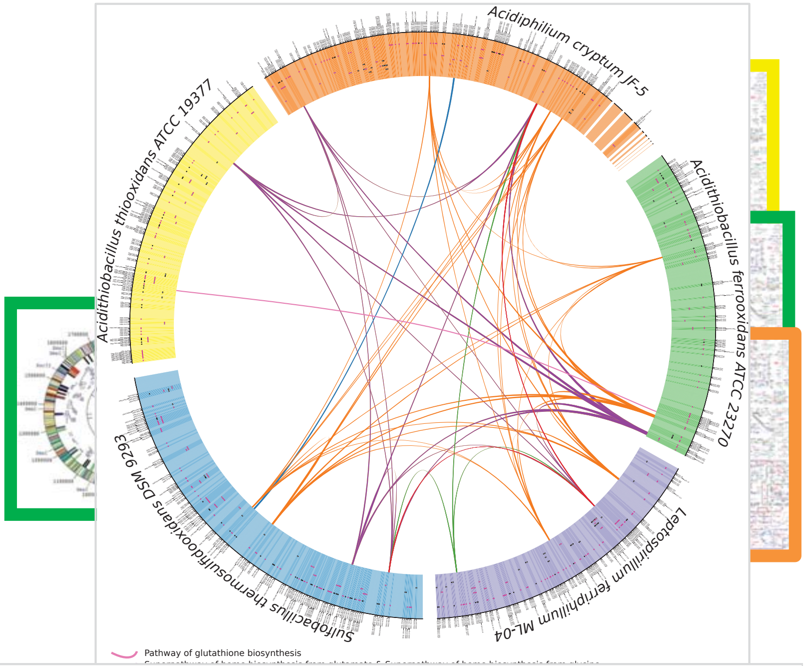
Linear constrains atoms

```
&sum{a1*x1;...;al*xl} <= k
```

Application: extremophile mining consortium

*Role of an **empirical taylor-made consortium** of bacteria in copper extraction from ore ?*





*« **NAD(H) biosynthesis** metabolic pathways of **A. Cryptum** complements metabolic functions spread between the five strains »*

ASP program

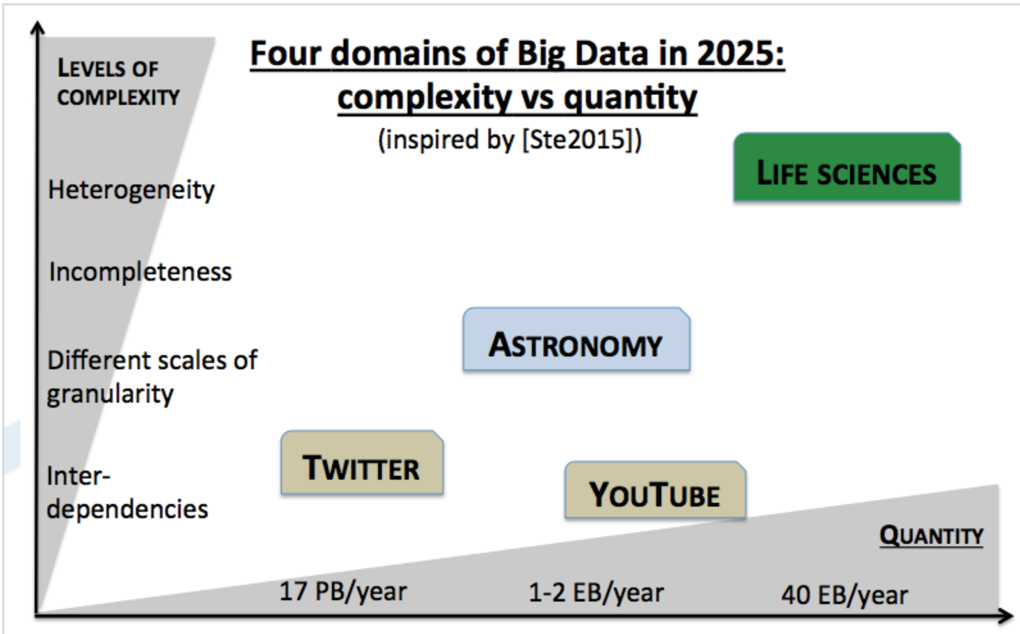
→ graph alignment / static modeling

→ chains of reactions explaining the capability of the consortium to produce the compounds (LPNRM'13, Microbiology open'15)

BACK TO DYNAMICAL SYSTEMS

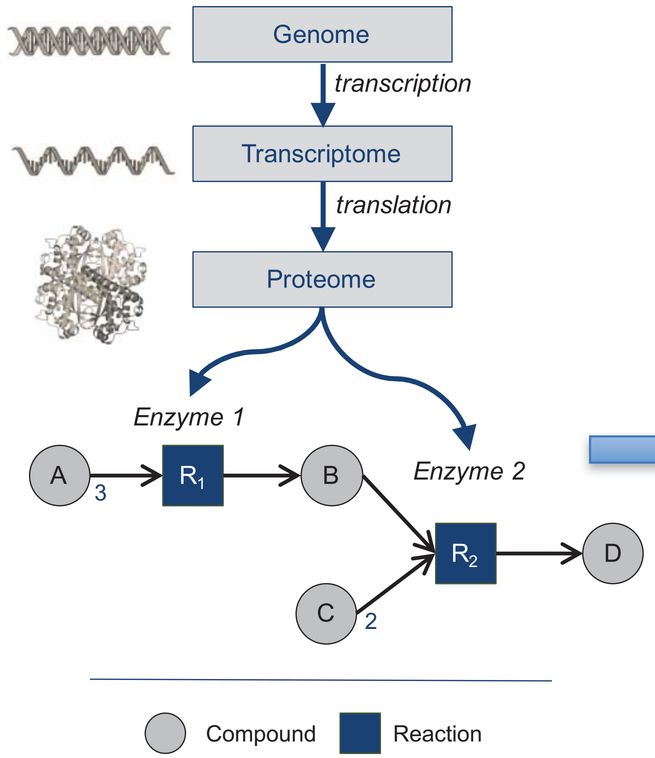
Biological systems characteristics

- Large-scale
- Empirical laws
- Few data wrt the search space size



Biological systems observed with omics data are not uniquely identifiable

Underlying tool : from genes to dynamical systems

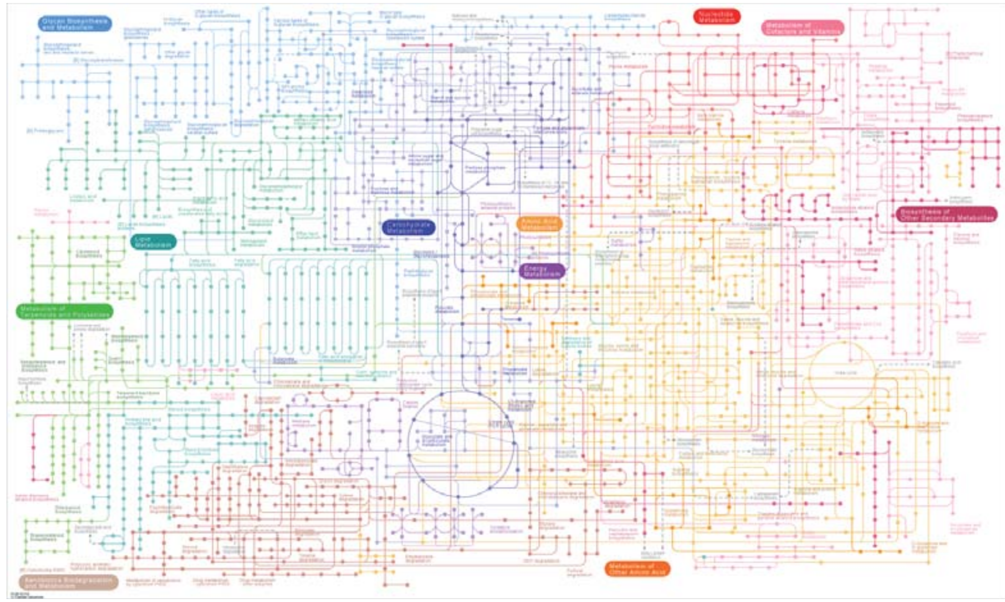


Link between genes and functions

1 genome

⇒ 1 metabolic network

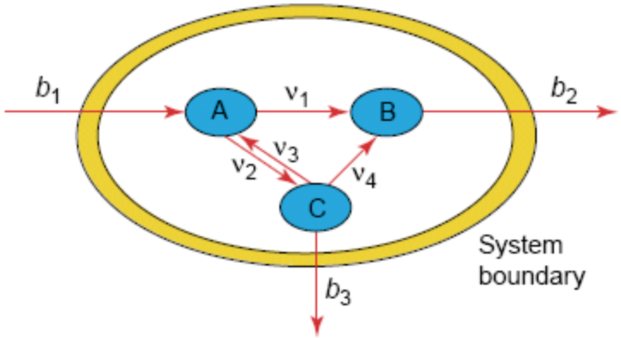
= bipartite directed graph



Large scale metabolic network

All expected metabolic capabilities of an organism

How to model fluxes ?



$$\frac{dA}{dt} = -v_1 - v_2 + v_3 + b_1$$

$$\frac{dB}{dt} = v_1 + v_4 - b_2$$

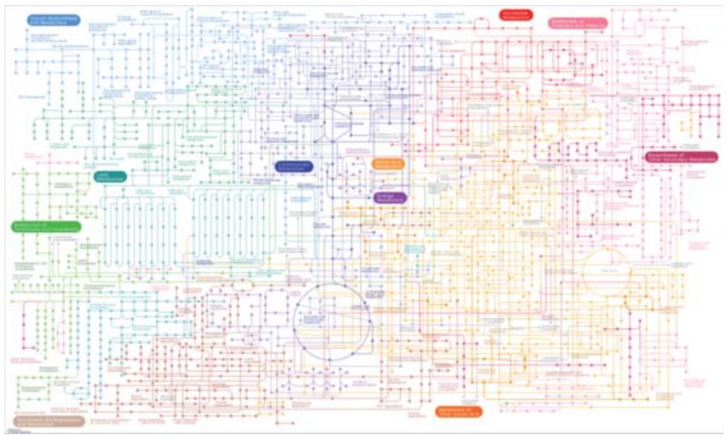
$$\frac{dC}{dt} = v_2 - v_3 - v_4 - b_3$$

$$\frac{dx}{dt} = S \cdot v(x)$$

$$v([substrat]) = Vm[Substrat]/(Km + [Substrat])$$

Back to high school chemistry

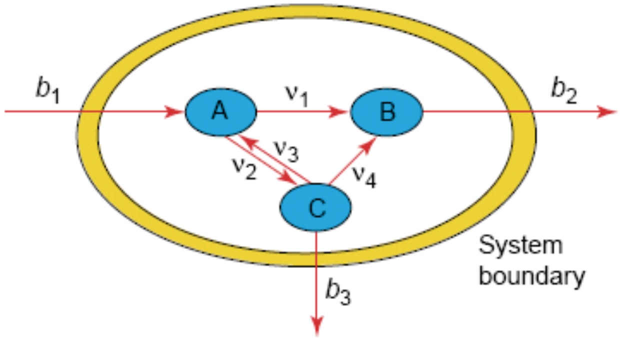
- ➢ Two parameters have to be estimated for each reaction



Intractable in practice !

- ➢ Overapproximation of the dynamics

Quasi-steady state hypothesis



$$\frac{dx}{dt} = S \cdot v(x) = 0$$

$$v([substrat]) = Vm[Substrat]/(Km + [Substrat])$$

= *constant*

Metabolic compounds do not accumulate

- Fluxes have constant values
- Fluxes are constrained by linear values
- The system optimises a global objective

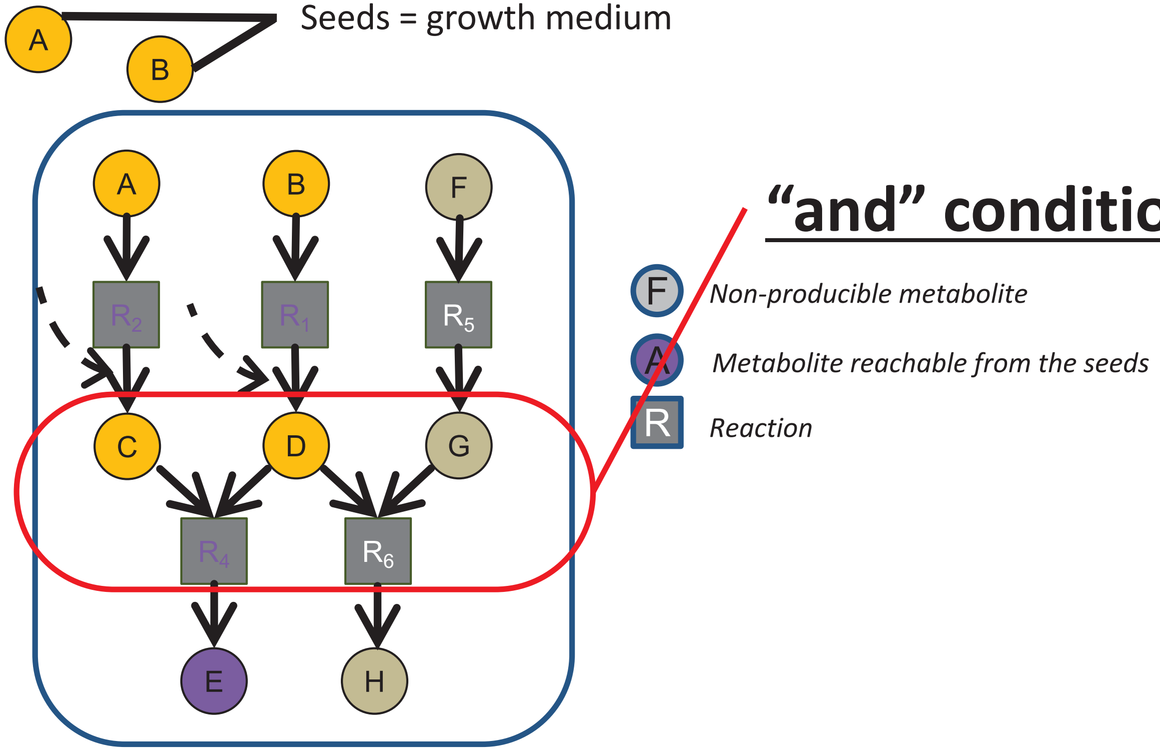
r is *active* if

$v_r > 0$ and
$s.v = 0$ and
$lb < v < ub$

Replace kinetic constants by global optimisation hypotheses

Growing phase hypothesis

Functionality: recursive graph-based semantics

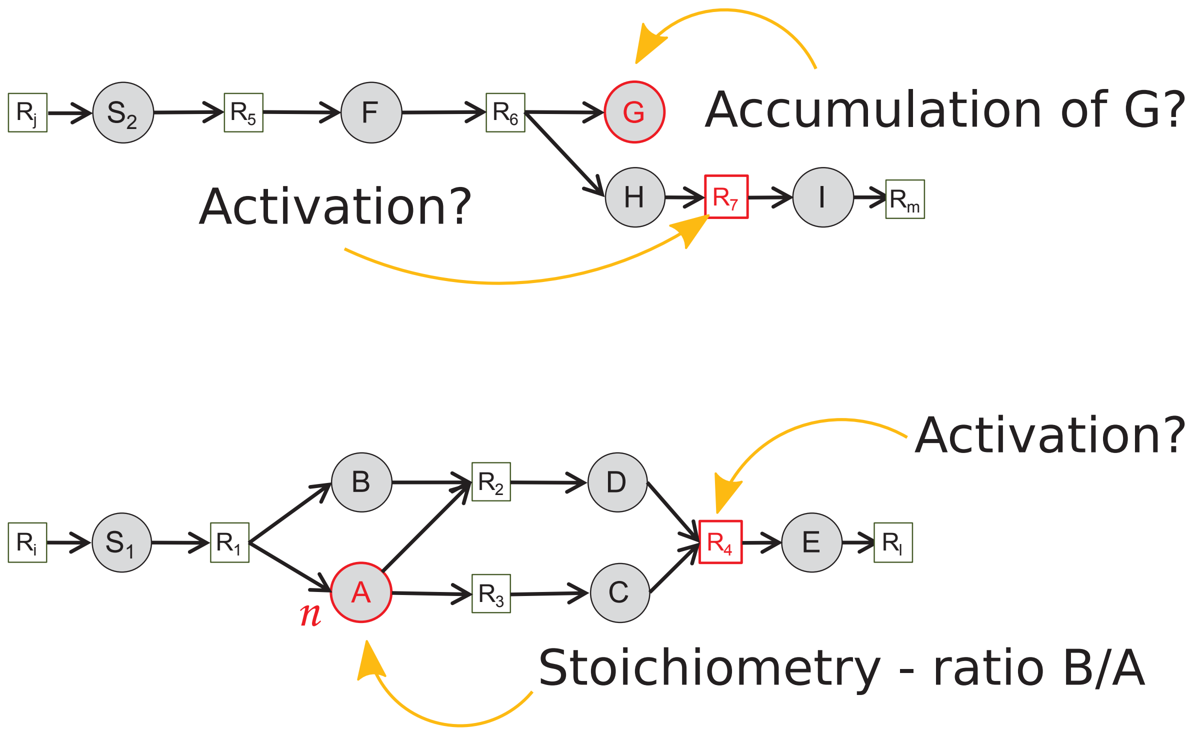


“and” condition checked recursively

```
scope(M):- seed(M).
scope(M):- product(M,R), reaction(R),  scope(M') : reactant(M',R).
```

Study paths in hypergraphs

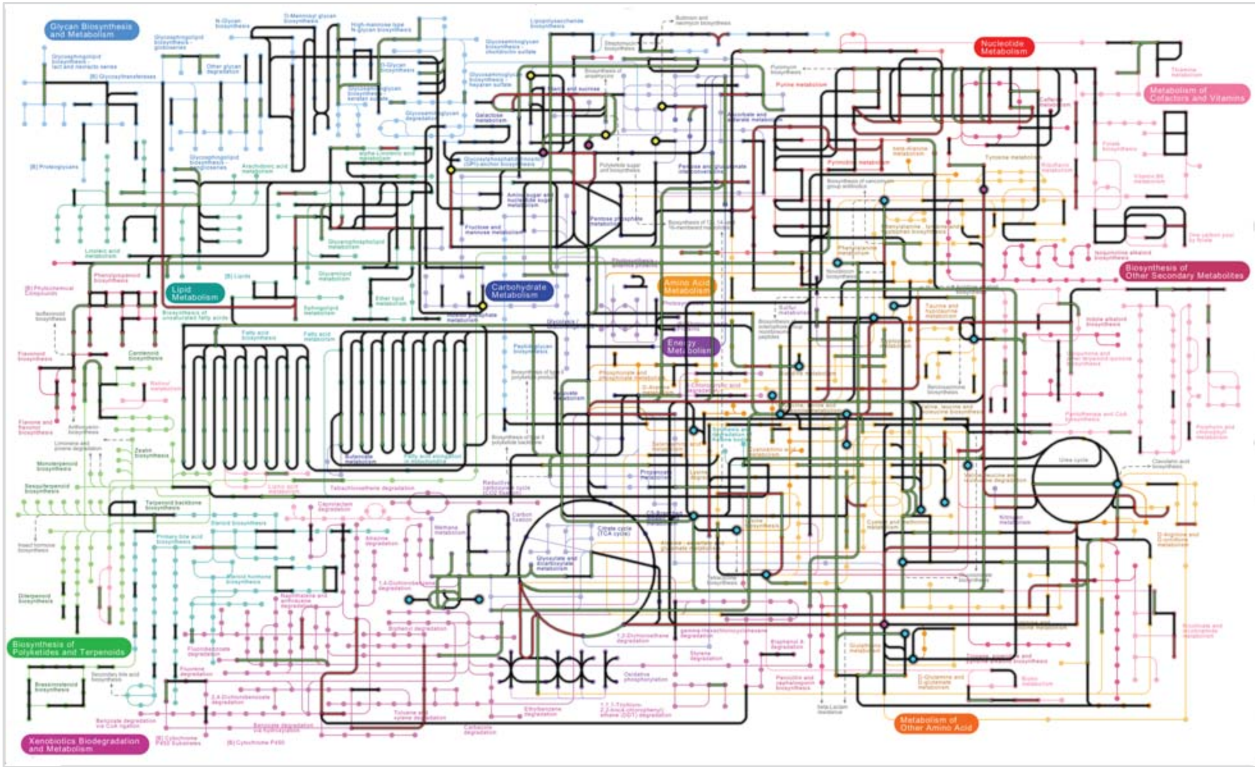
Everything is a matter of choices



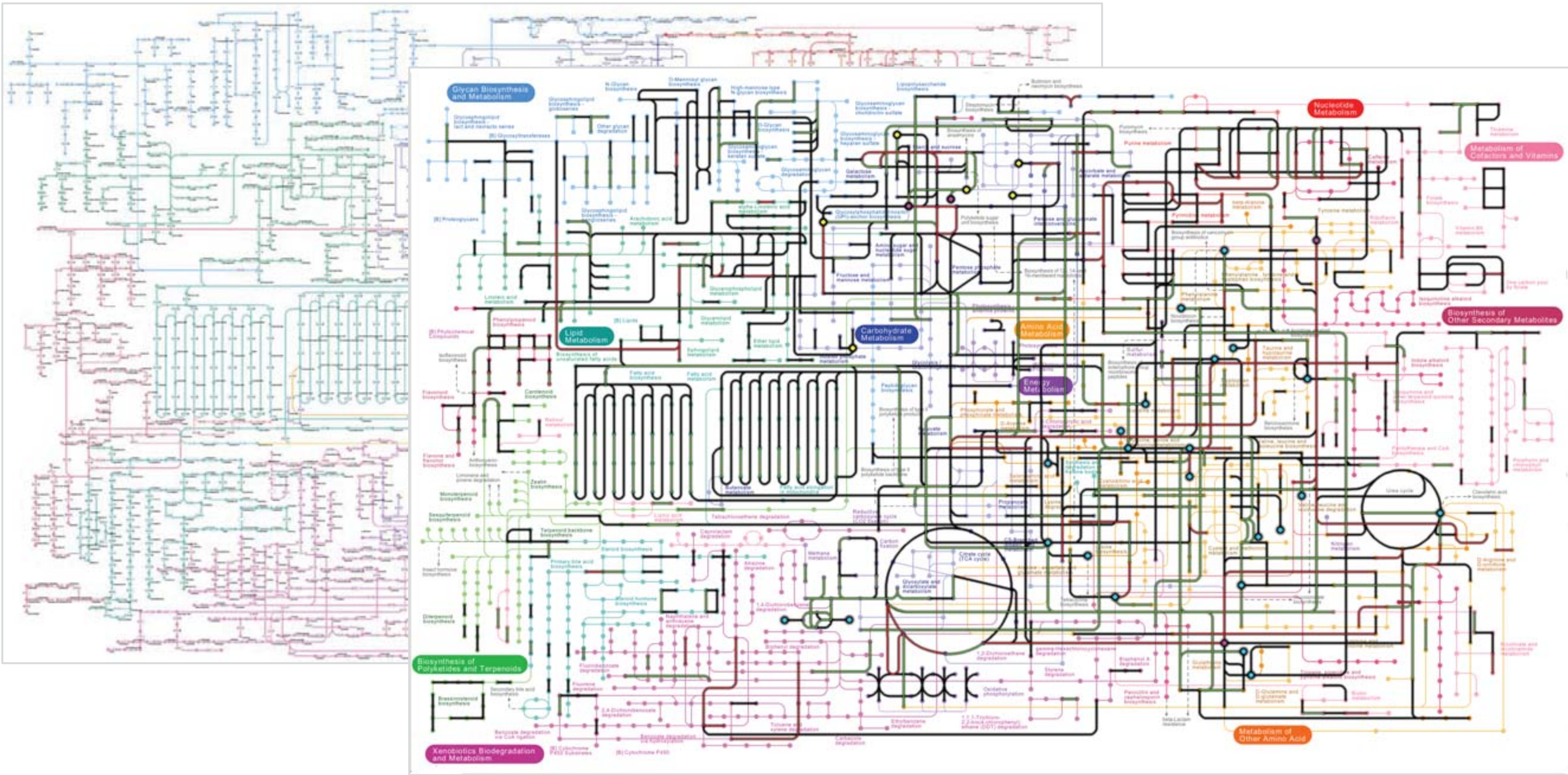
The reaction status of the reactions is different according to the approximation

- No choice but dealing with such overapproximation !
- Use the flexibility of ASP language to handle these questions

APPLICATION TO NETWORK COMPLETION



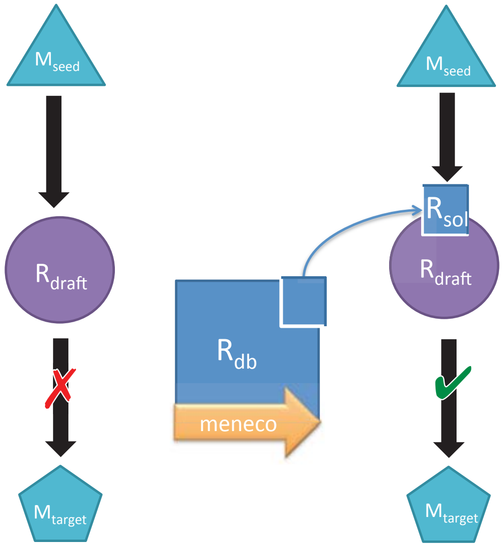
Data incompleteness



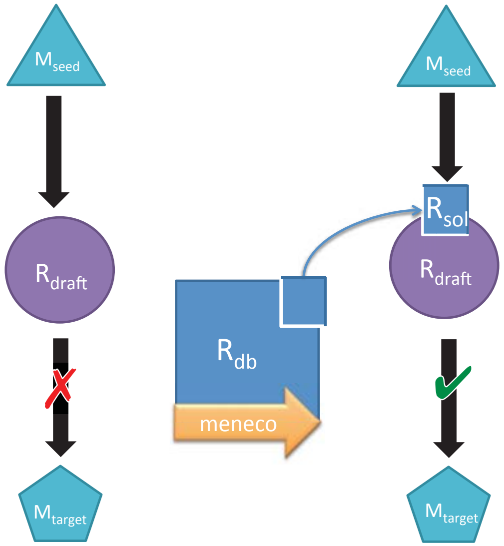
Metabolic networks built from NGS sequencing

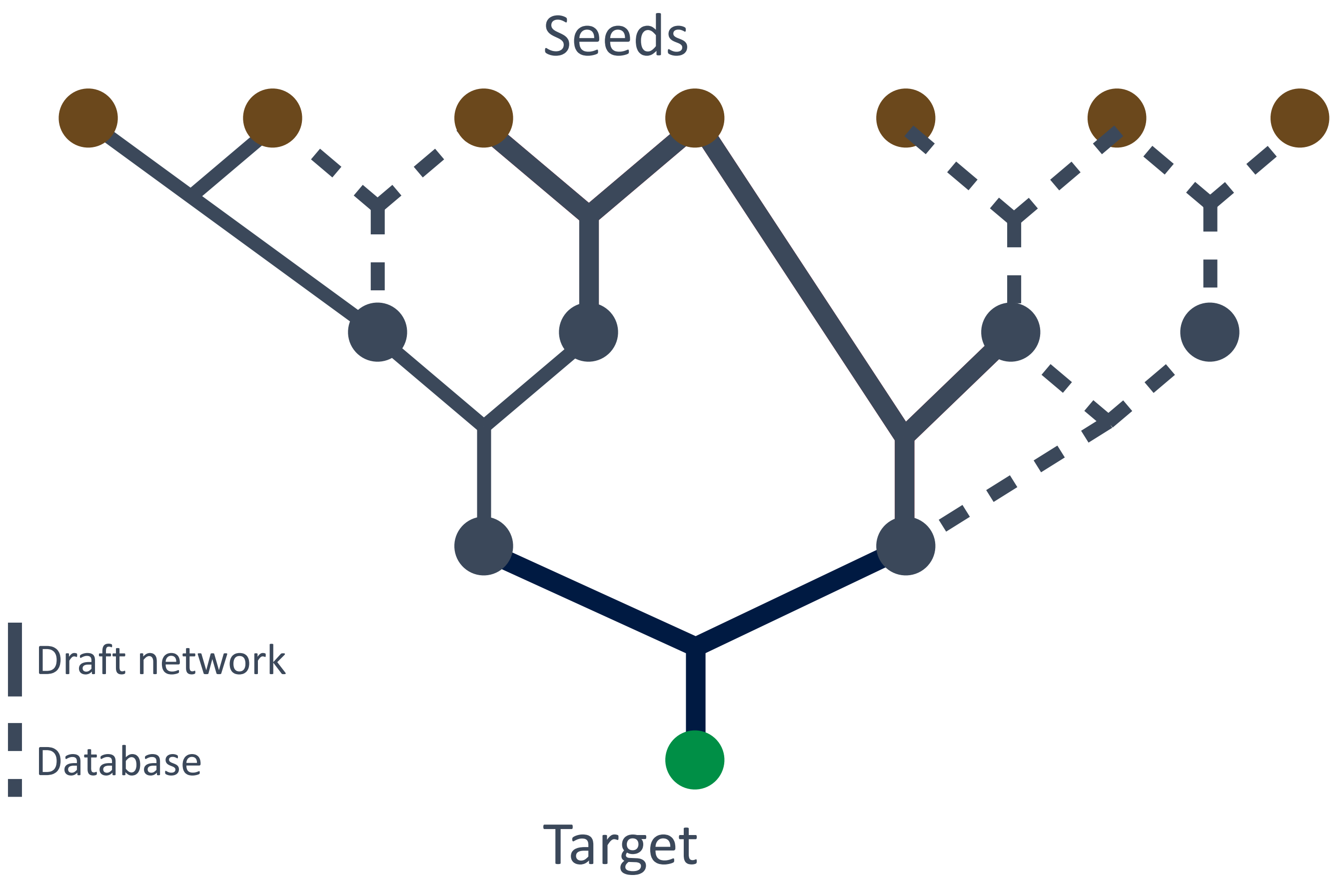
➢ no possible biomass production.

Gapfilling a metabolic network (nutshell)

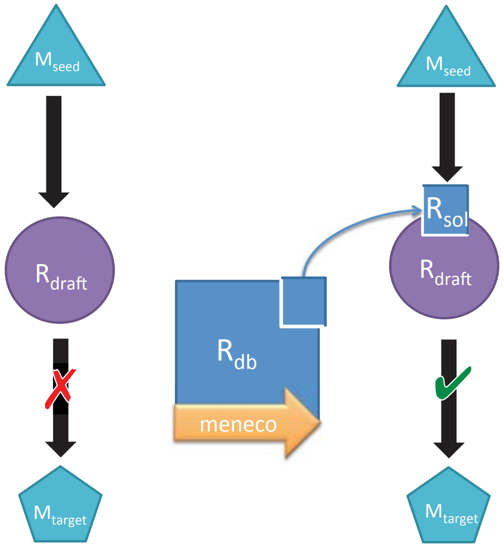


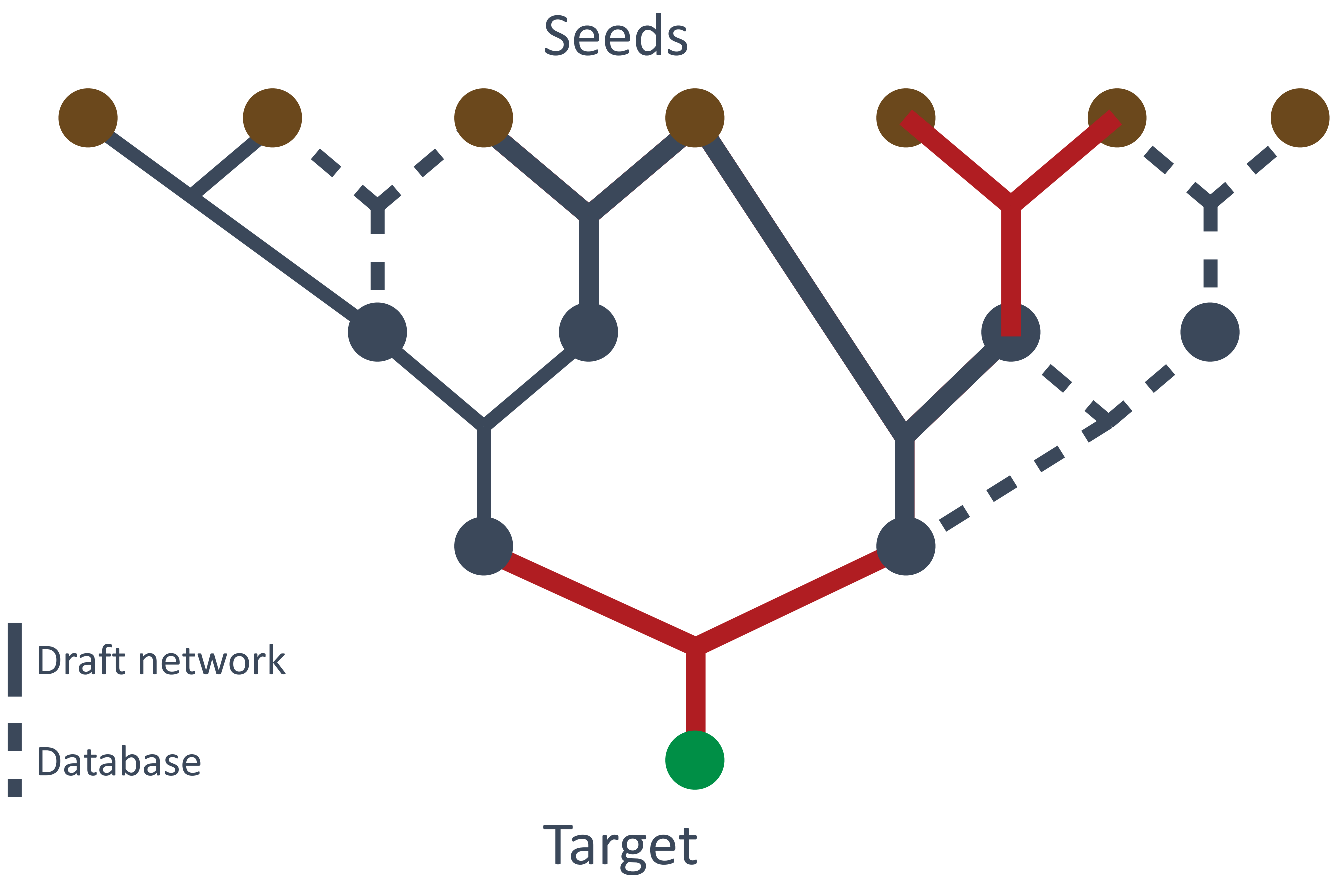
Gapfilling a metabolic network (nutshell)



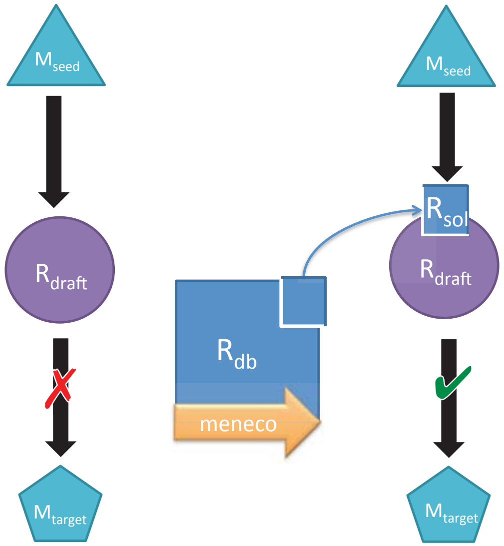


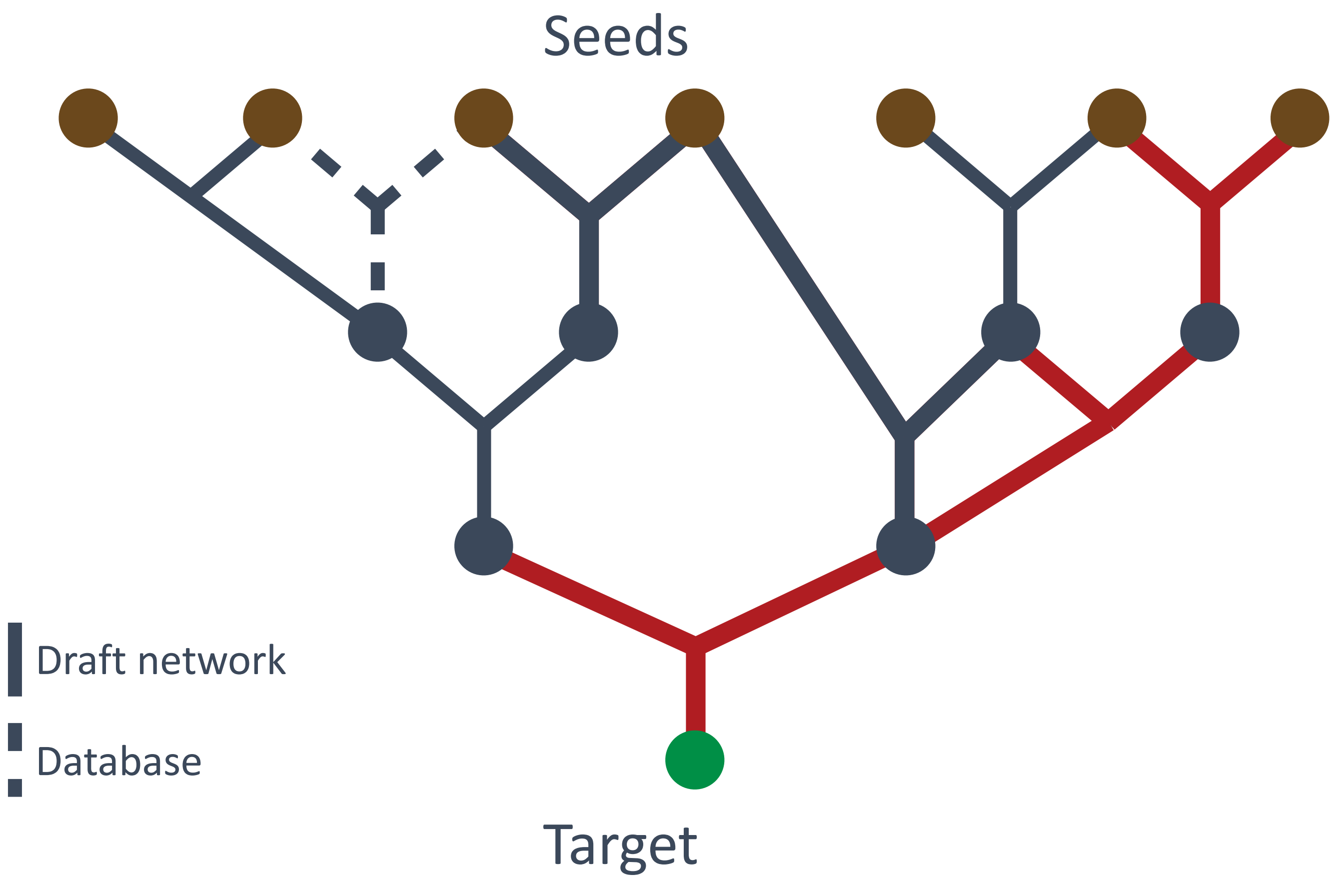
Gapfilling a metabolic network (nutshell)



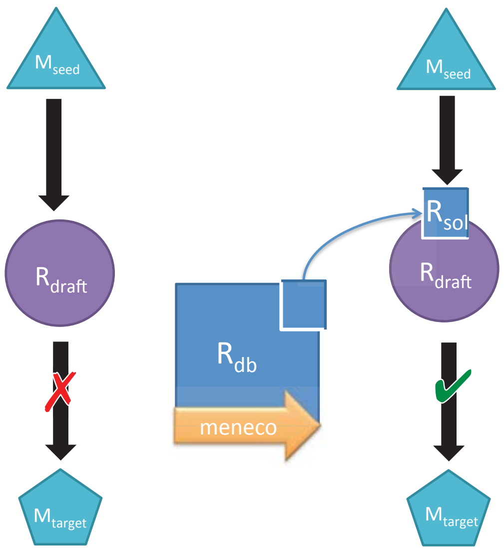


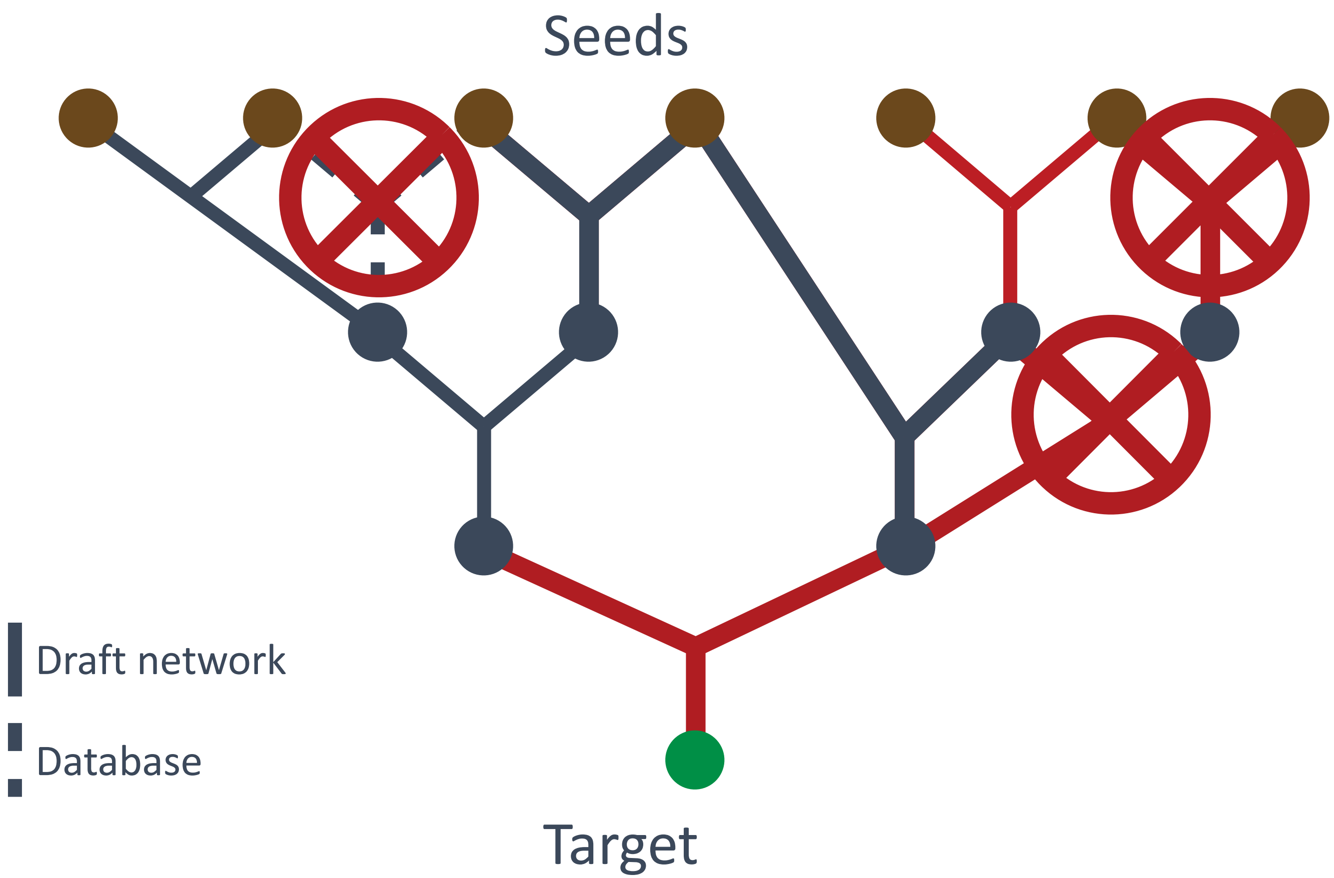
Gapfilling a metabolic network (nutshell)





Gapfilling a metabolic network (nutshell)

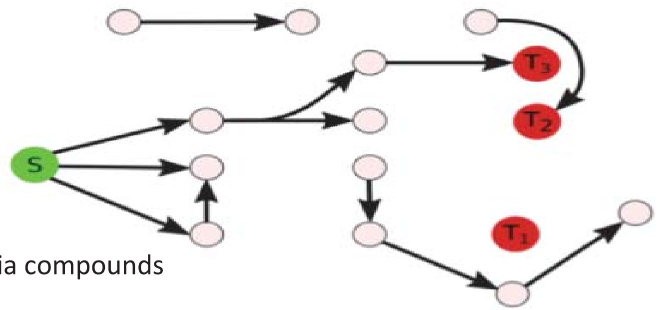




Gapfilling a metabolic network

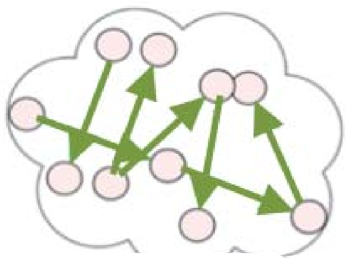
What we have

- Graph with **non-accessible target components**
- **Knowledge database** of possible edges



Media compounds

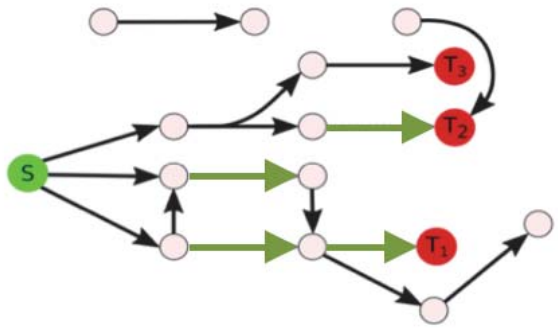
Experimentally observed compounds



Putative interactions from knowledge databases

Gap-filling problem:

- Restore target accessibility
- Minimal number of reactions



$$\mathrm{gapfilling}(S, R_T, G_1, G_{DB}) = \underset{\{R_i..R_m\} \subset G_{DB}}{\arg\min} \begin{pmatrix} size(reactants(R_T) \setminus scope(G_1 \cup \{R_i..R_m\}) \\ size\{R_i..R_m\} \end{pmatrix}$$

Meneco: ASP-based gap-filling for non-model organisms

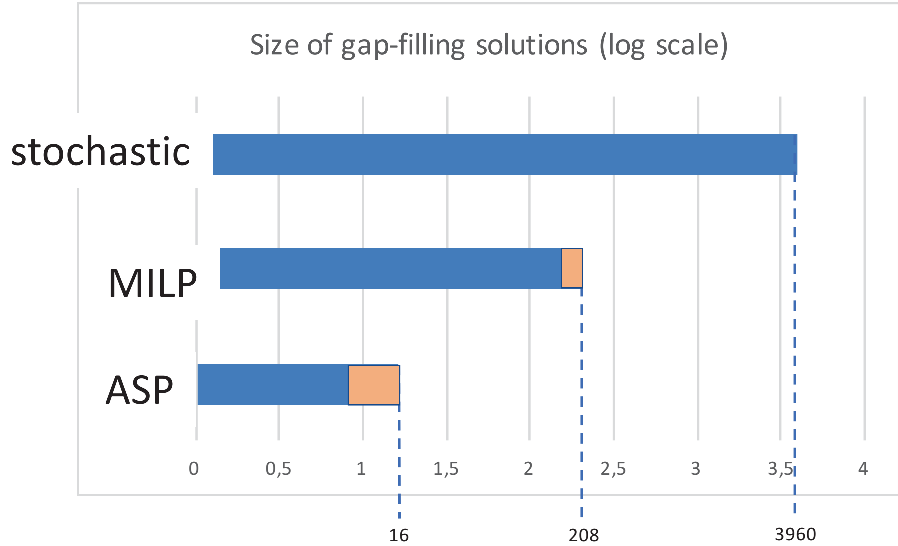
$$\mathrm{Hybgapfilling}(S, R_T, G_1, G_{DB})$$

$$\underset{\{R_i..R_m\}\subset G_{DB}}{\arg\min} \begin{pmatrix} size(reactants(R_T) \setminus scope(G_1 \cup \{R_i..R_m\}) \\ size\{R_i..R_m\} \end{pmatrix}$$

```
{reaction(r)}.
scope(M):- seed(M).
scope(M):- product(M,R), reaction(R), scope(M') : reactant(M',R).
:- target(T), not scope(T).
#minimize{ reaction(r) }.
```

16 reactions in average are sufficient to restore degraded bacterial networks (PLOS CB 2017)

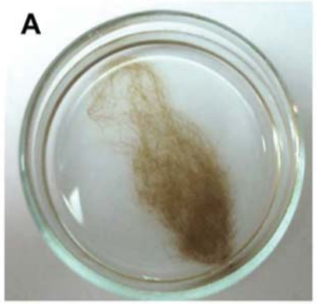
➢ MILP-based approaches required from 200 to 4000 reactions.



Benchmark of 10,800 bacterial networks

Example of application

- Genome: 1785 reactions, 1981 compounds

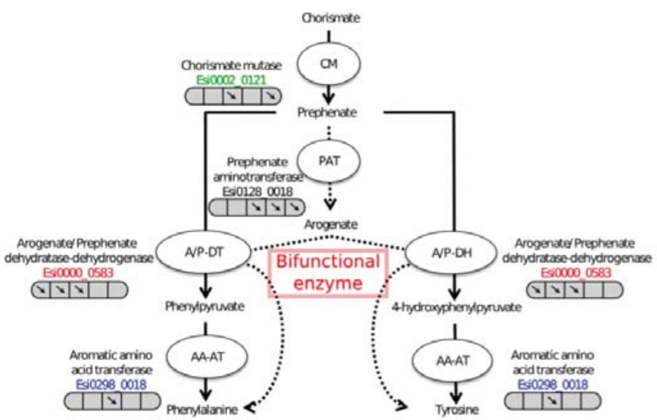


Ectocarpus siliculosus

[Tapia2016]

- 54 metabolites to produce:
  - 25 are graph-based producible
  - None is FBA-based producible.
- Gapfilling
  - MILP : 500 reactions (untractable)
  - ASP: 50 reactions added to the network
    - Sufficient for fluxes
    - Manual curation

Proposed after manual curation



New bifunctional role of a specific enzym

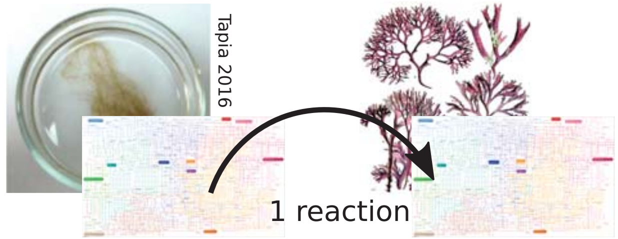
(Plant Journal 2015)

Counter-example of application

Chondrus crispus

Network analysis (G. Markov, SBR)

- 1943 reactions
- 149 reactions added by ASP
- **No way to produce biomass**

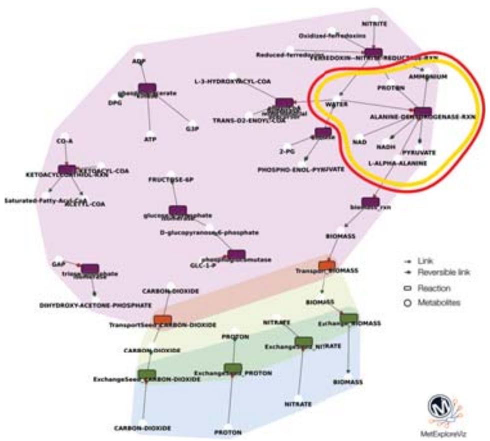


New problem to be solved

- **Hybrid problem** (TPLP 2018)
- Constraint propagator
- Reduce the database

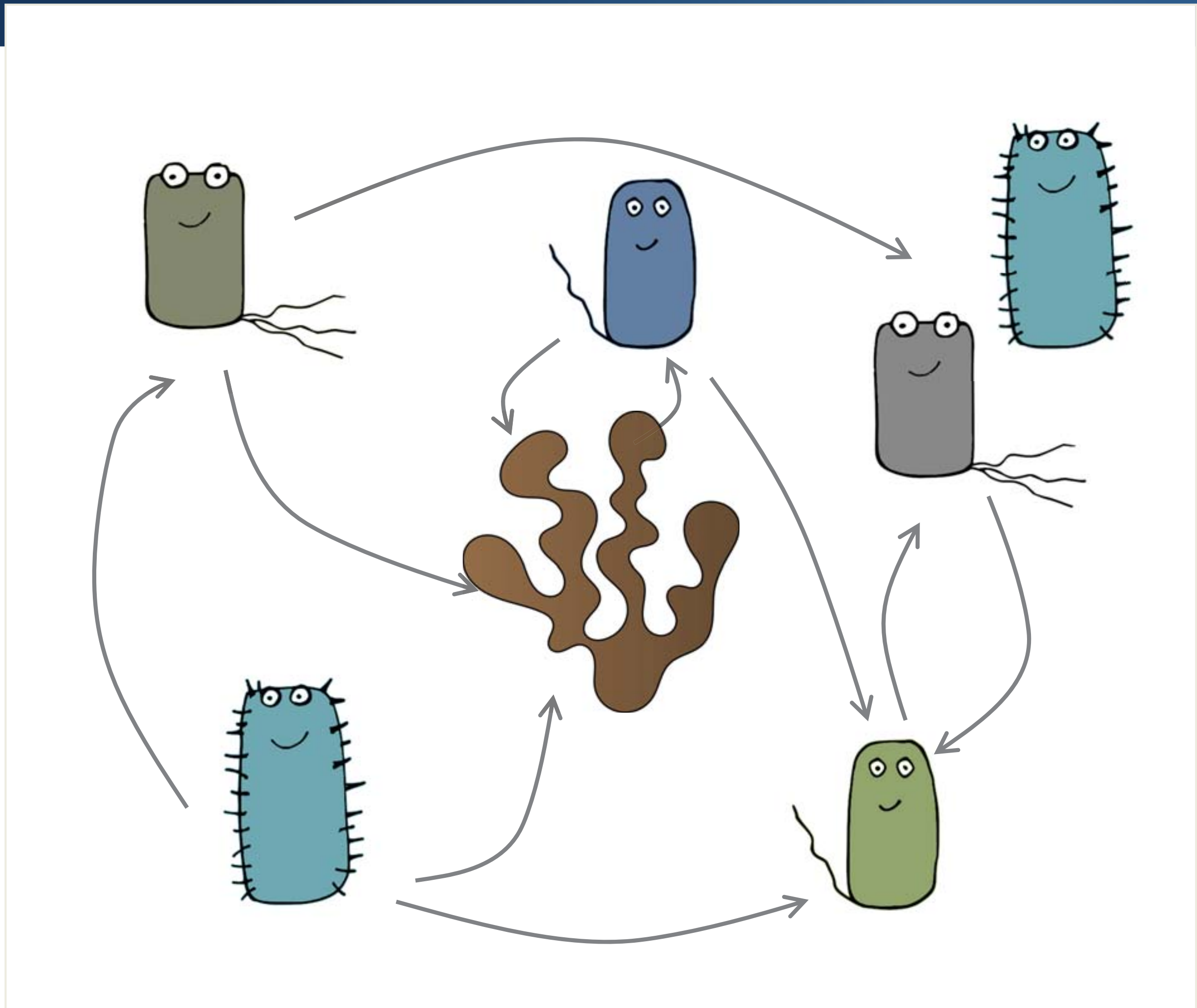
$$\mathrm{Hybgapfilling}(\mathrm{S}, \mathrm{R_T}, \mathrm{G_1}, \mathrm{G_{DB}}) = \underset{\{R_i..R_m\} \subset G_{DB}}{\arg\min} \left(\frac{size(reactants(R_T) \setminus scope(G_1 \cup \{R_i..R_m\})}{size\{R_i..R_m\}} \right)$$

$$\text{s.t} \quad s.v = 0, v_{R_T} > 0,\ lb < v < ub$$

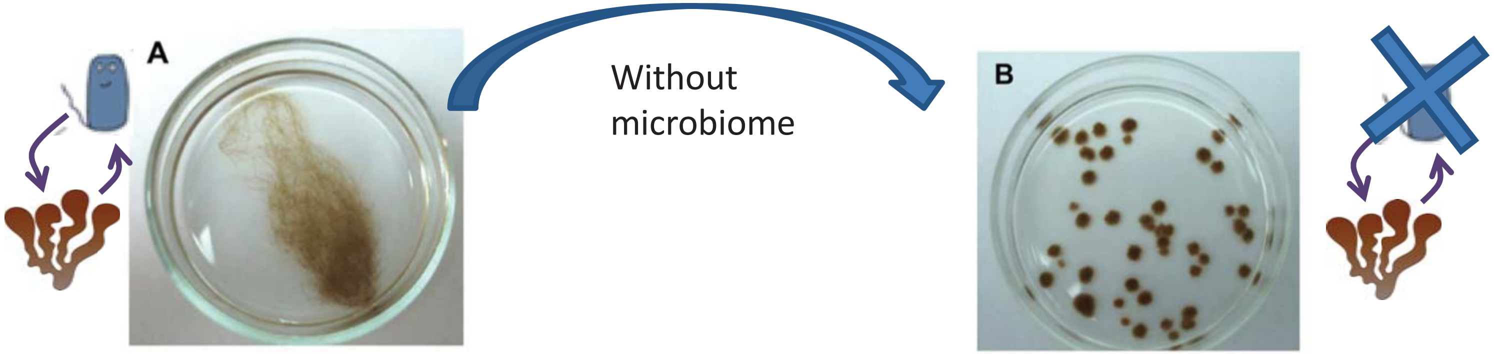


Essential reactions for alanine production in *CcrGem*

STILL MORE COMPLEXITY

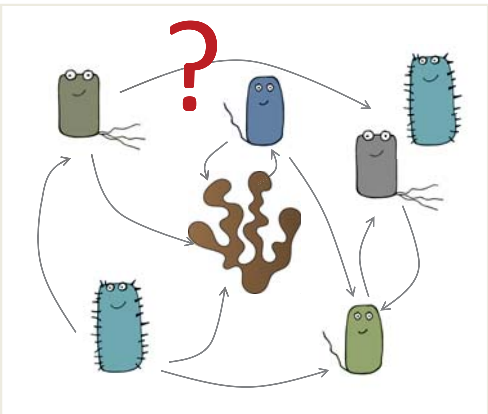


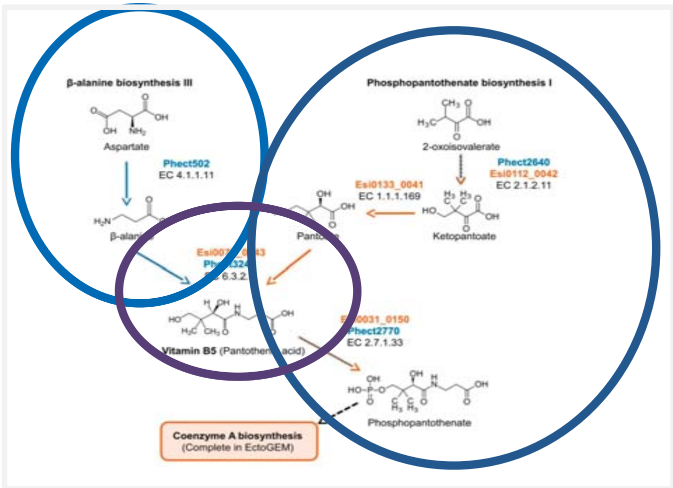
Role of environmental bacteria ?



Ectocarpus
[Dittami2014, Tapia2016, Prigent2015]

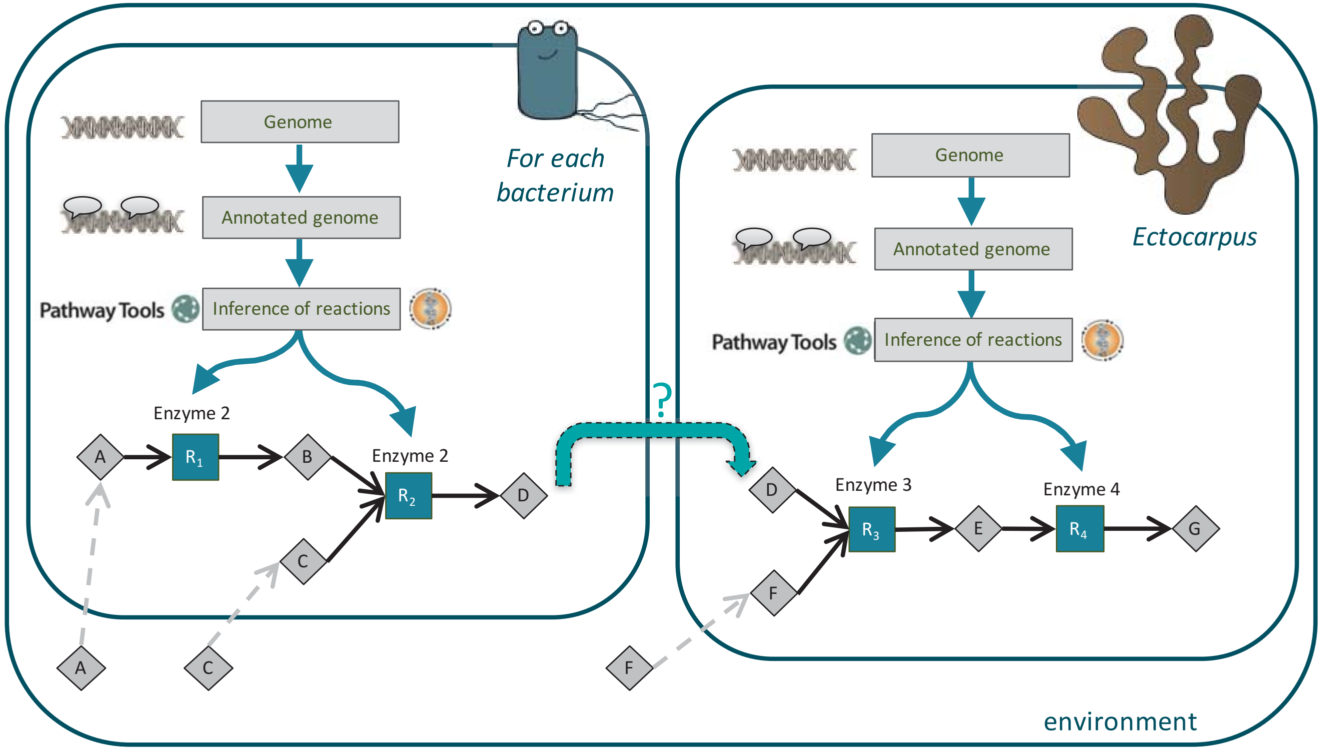
CNRS • SORBONNE UNIVERSITÉ Station Biologique de Roscoff





Metabolism may be an explanation
(PLOS CB 2017)

Systems ecology question



Can we suggest compound exchanges that could restore the production of targeted compounds ?

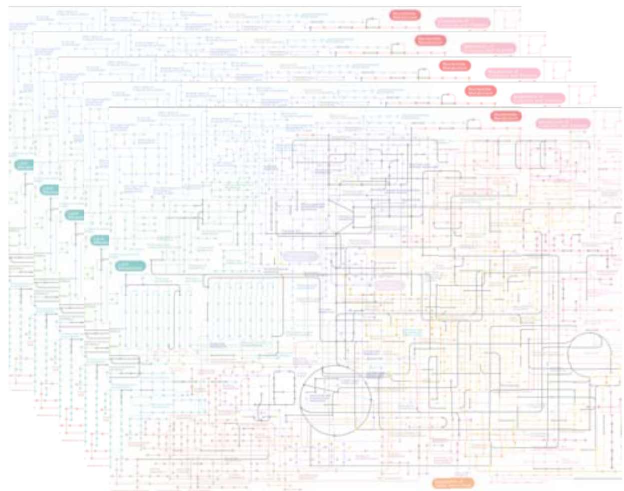
- New gap-filling problem !
- Steiner graph approach (Sagot team, 2017) or ASP implementation

Scalability…

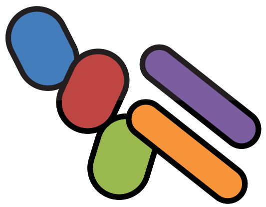
But… There are hundreds of bacteria in the environment

Marine biology

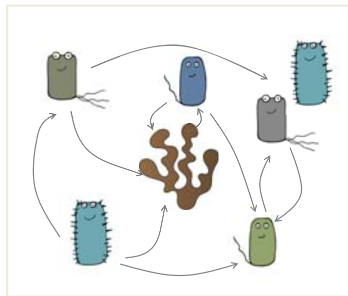
Hundreds
of bacteria



Hundreds of Genome-scale models (GSMs)



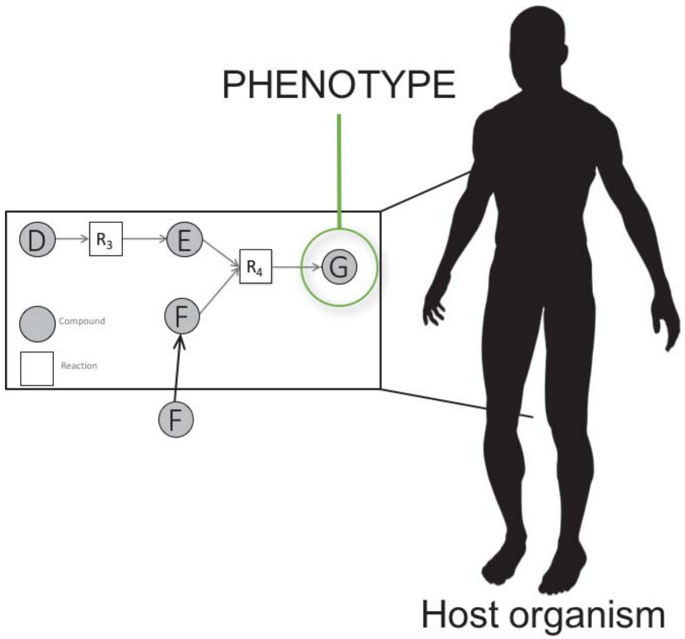
?



Happy few bacteria interact with the algae

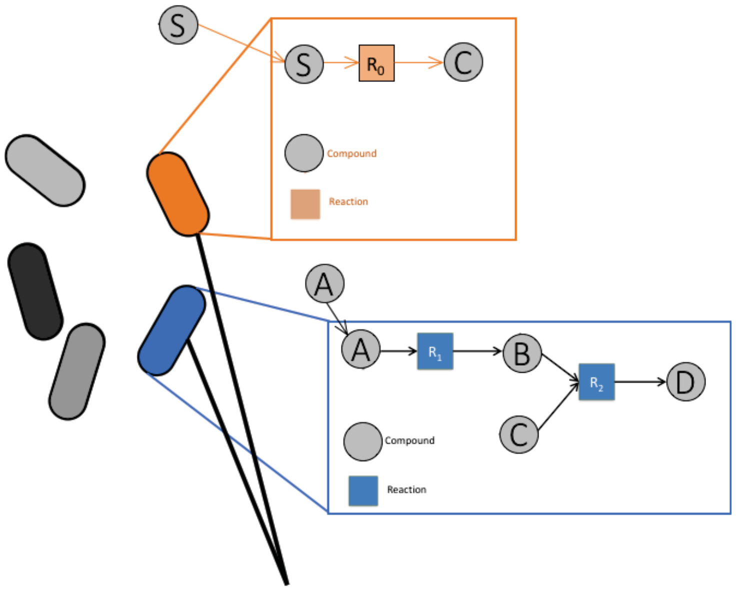
How to select communities within large microbiotas which explain the algal response to stress ?

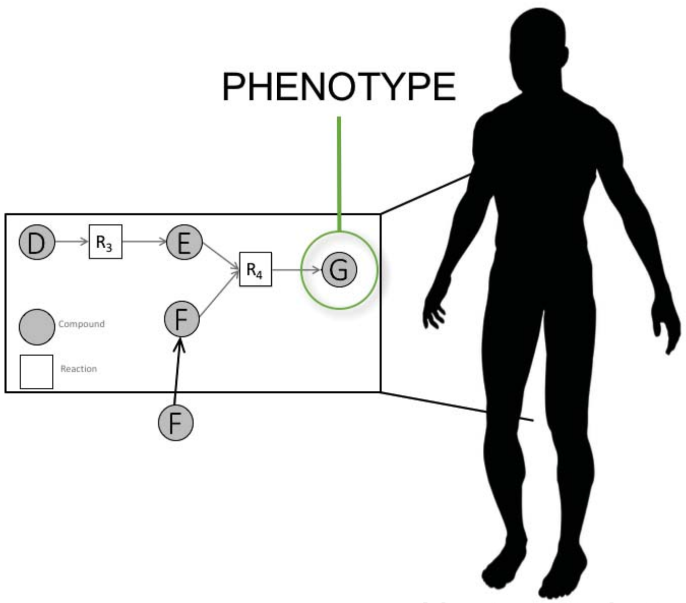
Selecting communities of interest within [large] microbiotas



The "who", "how" challenges of community selection

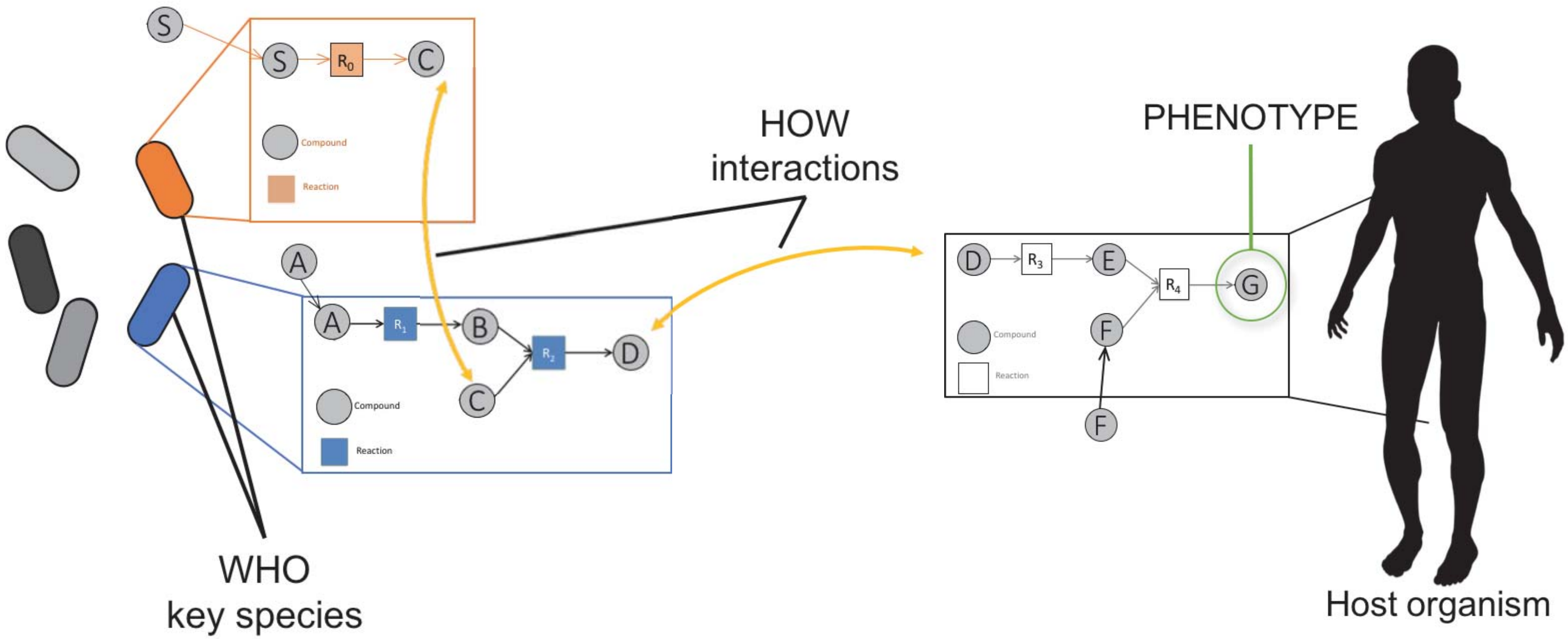
Selecting communities of interest within [large] microbiotas





The "who", "how" challenges of community selection

Selecting communities of interest within [large] microbiotas



The "who", "how" challenges of community selection

Complexity

Community selection problem

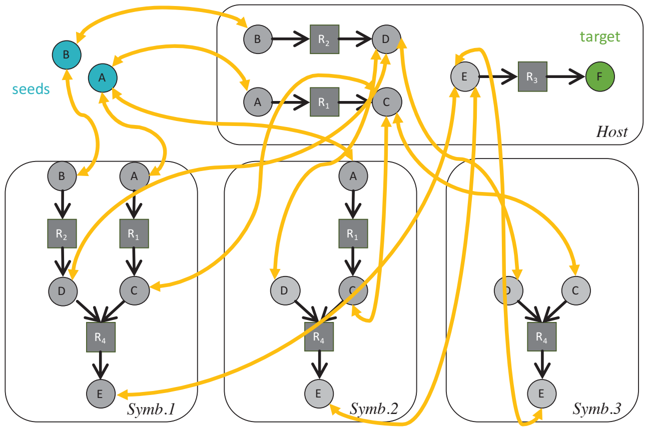
- Switch from hundreds of symbiots to 3 or 4
- Pinpoint a few number of putative cross-feedings

$$\mathrm{Comsel}(S,T,G_1..G_n) = \underset{\{exchg(G_{i_1}..G_{i_L})\}\subset\{G_1..G_n\}}{\arg\min} \begin{pmatrix} size(T \setminus MBscope(G_{i_1}..G_{i_L})) \\ size\{\varepsilon \subset exchg(G_{i_1}..G_{i_L}) \mid \\ T \cap CPscope(G_{i_1}..G_{i_L},\varepsilon,S) = \\ T \cap MBscope(G_{i_1}..G_{i_L},S)\} \end{pmatrix}$$

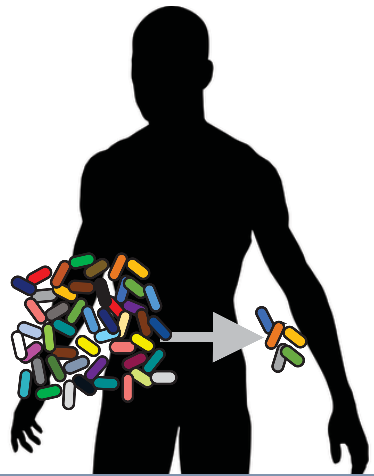
- depends on the number of hyperarcs

Size of the search space

- depends on the number of symbionts



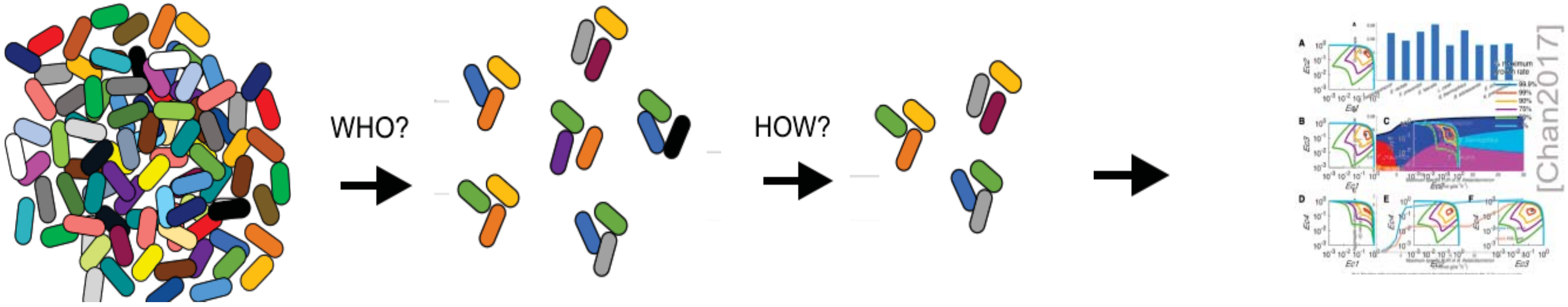
499,177 combinations of <6 exchanges



1.62.10^{81} combinations of <10 exchanges

Highly combinatorial problem

Two-step optimization procedure



Heuristics for the community selection problem

- **Who problem.**
  - Get rid of boundaries and select all minimal symbiot families

$$\mathrm{mxdbagCnity}(S,T,G_1..G_N) = \underset{\{G_{i_1}..G_{i_L}\}\subset\{G_1..G_N\}}{\arg\min} \begin{pmatrix} \mathrm{size}\left(T\setminus \mathrm{mxdbagScope}(G_{i_1}..G_{i_L},S)\right), \\ \mathrm{size}\,\{G_{i_1}..G_{i_L}\}. \end{pmatrix}$$

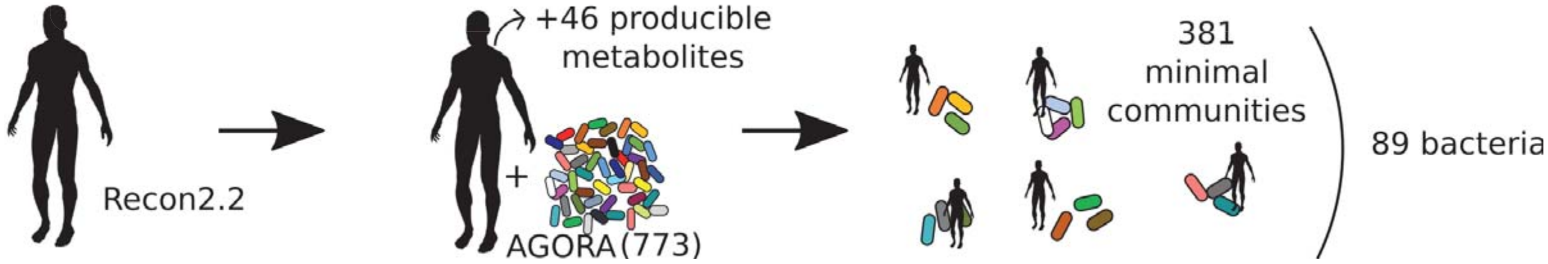
- **How problem.**
  - Sort the selected families according to the number of exchanges

$$\mathrm{cptCnity}(S,T,G_1..G_N) = \underset{\substack{\{G_{i_1}..G_{i_L}\} \\ \subset \{G_1..G_N\}}}{\arg\min} \begin{pmatrix} \mathrm{size}\left(T\setminus \mathrm{mxdbagScope}(G_{i_1}..G_{i_L},S)\right), \\ \mathrm{size}\{G_{i_1}..G_{i_L}\}, \\ \mathrm{size}\{\mathcal{E}\subset \mathrm{exchg}(G_{i_1}..G_{i_L}) \mid \\ \quad T\cap \mathrm{cptScope}(G_{i_1}..G_{i_L},\mathcal{E},S) \\ \quad = T\cap \mathrm{mxdbagScope}(G_{i_1}..G_{i_L},S)\}. \end{pmatrix}$$

- **Manual curation.**
  - Ask your favorite biologist to select the final one

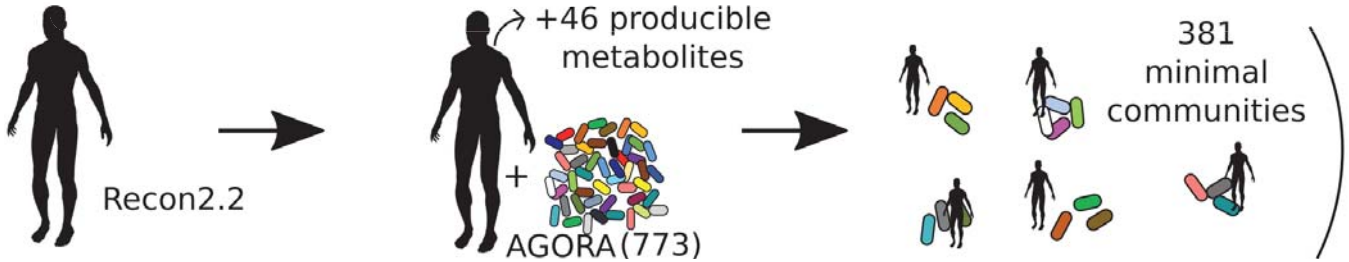
Validation/benchmarking on human microbiome project

Context of the study [Swainston et al., 2016] [Magnúsdóttir et al., 2016]

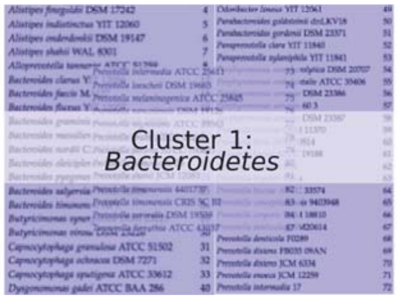


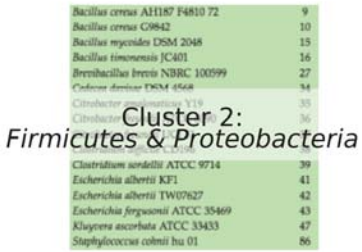
Validation/benchmarking on human microbiome project

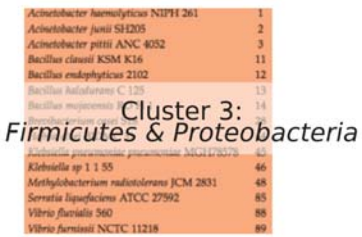
Context of the study [Swainston et al., 2016] [Magnúsdóttir et al., 2016]

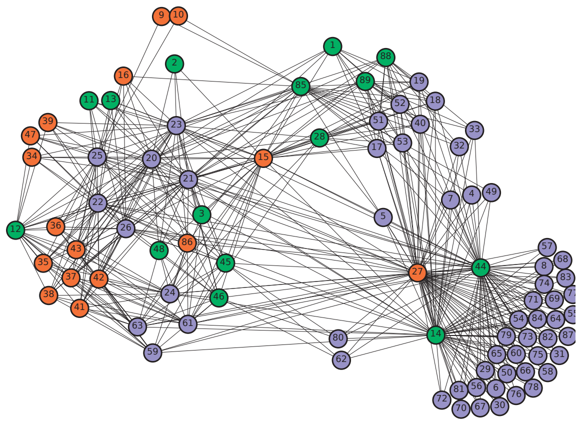


Clustering of bacteria







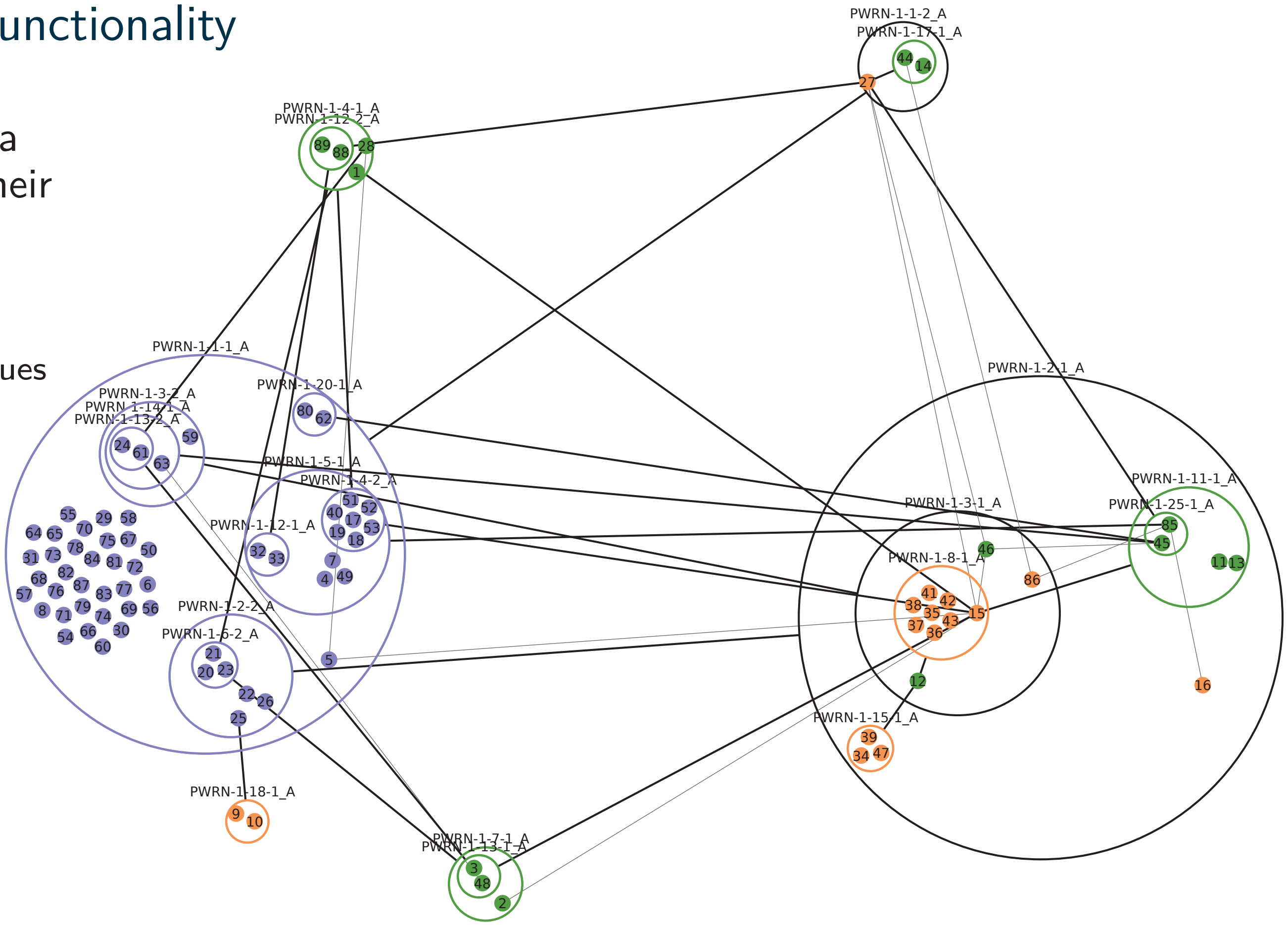


Each of the 381 communities is composed of
1 *Bacteroidetes* (/58) + 1 *Firmicute* or *Proteobacteria* (/15) + 1 *Firmicute* or *Proteobacteria* (/16)

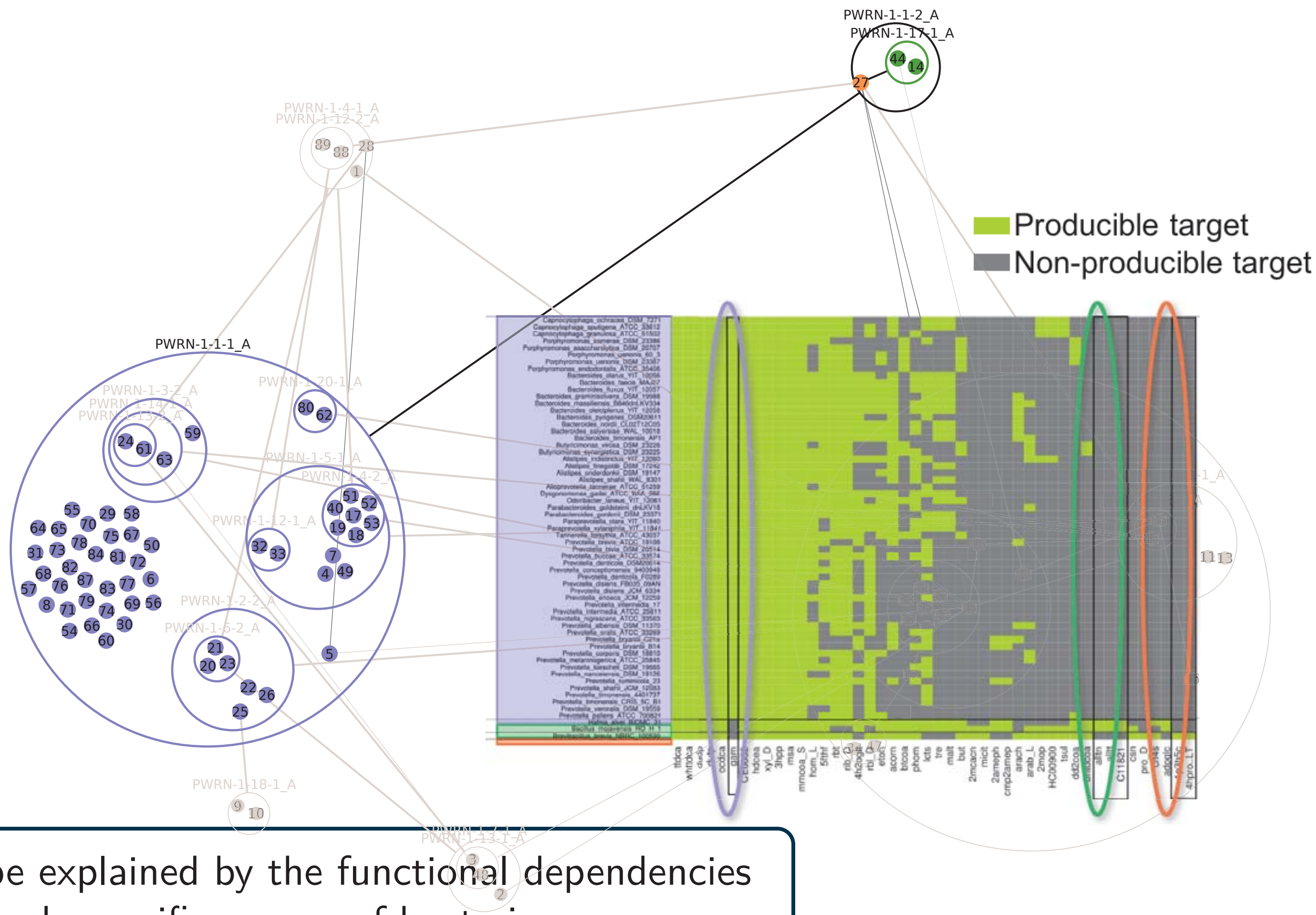
Validation/benchmarking on human microbiome project

Association of bacteria & functionality

- Groups of equivalent bacteria in clusters with respect to their associations [Bourneuf et al., 2017]
  - **Powernodes:** groups of bacteria, parts of bicliques
  - **Poweredges:** connect bicliques



Validation/benchmarking on human microbiome project



- Producibility of individual targets explains the communities → screening

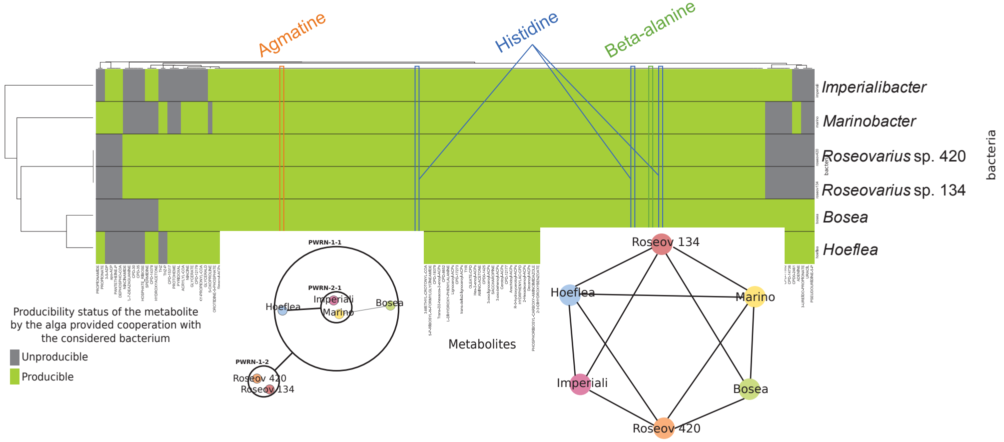
Community composition can be explained by the functional dependencies of the targets towards specific groups of bacteria

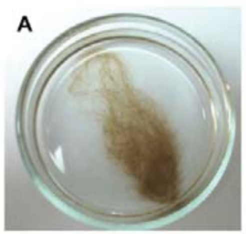
C. Frioux's Thesis. ECCB 2018

Validation/benchmarking on human microbiome project

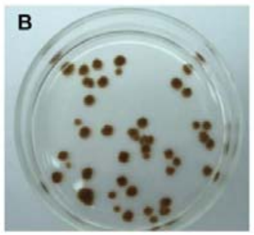
- *Ca.* P. ectocarpi not culturable
- 10 culturable bacteria → functional redundancy
- 6 equivalent communities of 3 bacteria

Joint work with Enora Fremy, Bertille Burgunter-Delamare & Simon Dittami

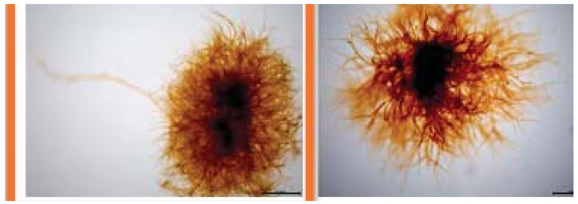




Without microbiome



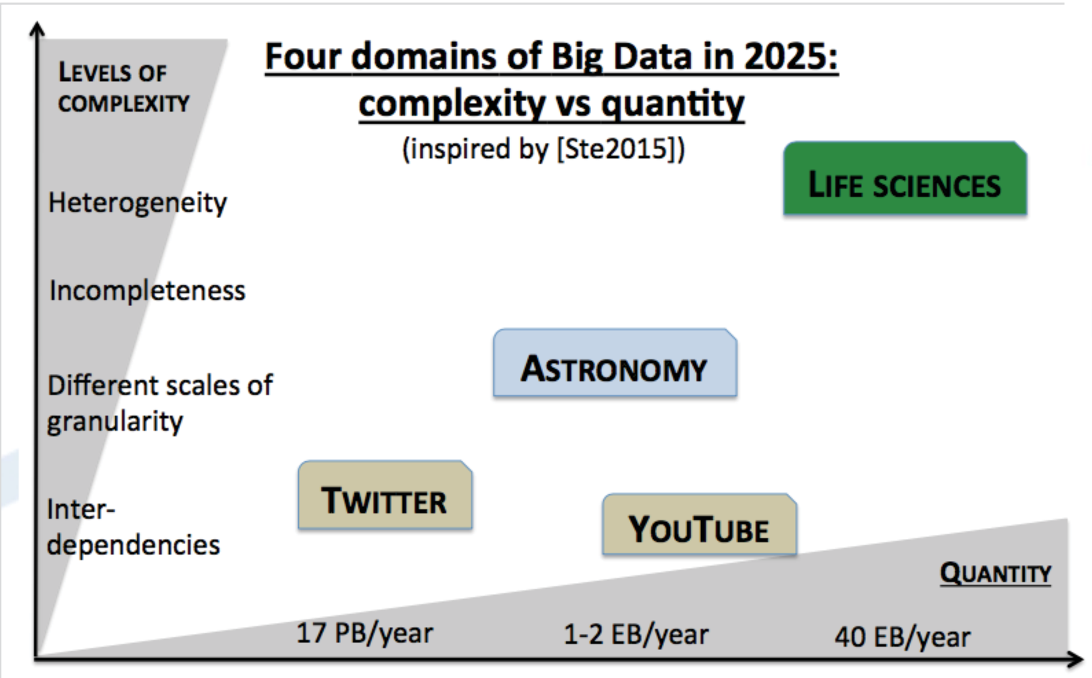
+ 3 selected bacteria among 30 cultivable bacteria



S. Dittami, Bertille Burgunter-Delamare

The algae grew again… But with strange behaviors

TOWARDS CONCLUSION



Take home messages: life science data integration ?

- **Life science data are multi-scale and heterogeneous**
  - Linked by underlying regulatory processes

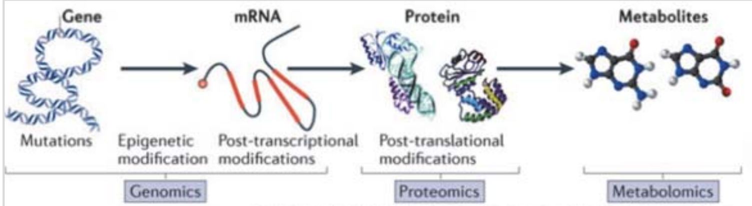
- **Systems biology ?**
  - study of complex systems which cannot be uniquely identified

- **Handling complexity for**
  - Make (dynamical) hypotheses
  - Solve optimization problems instead of identify parameters
  - Win-win collaboration with your BBF ASP-tech developers

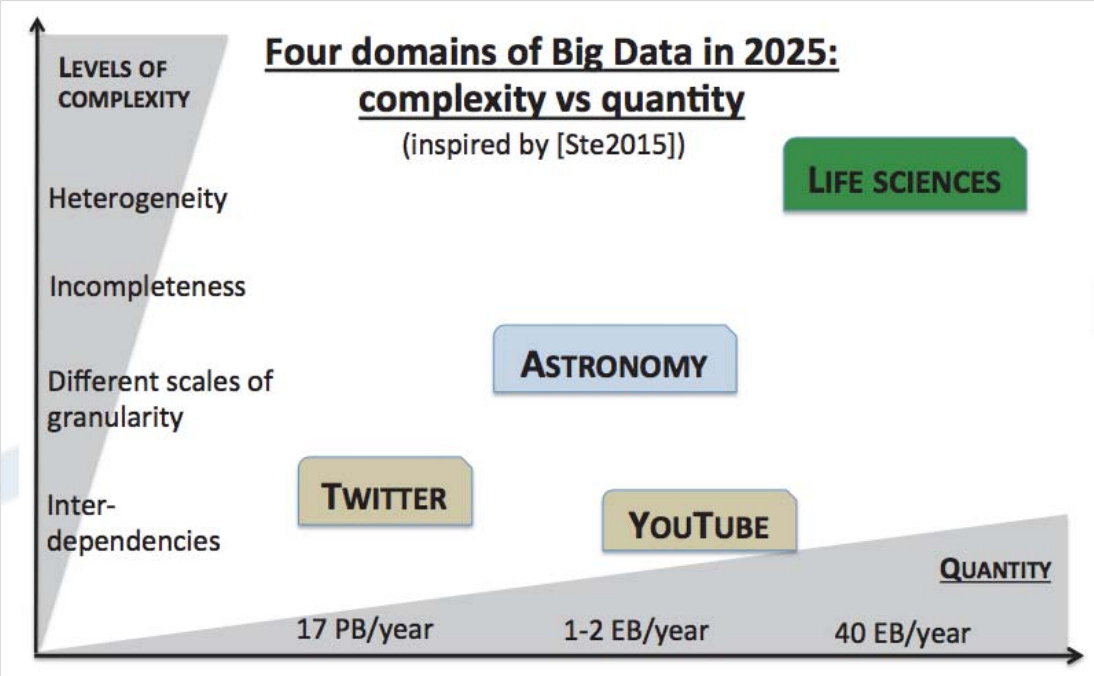
- **We will never replace biologists**

Molecular and cellular life science analysis is a user-assisted data science rather than a modeling system science

What about the future



Patti et al. (2012). Metabolomics: the apogee of the omics trilogy. Nature

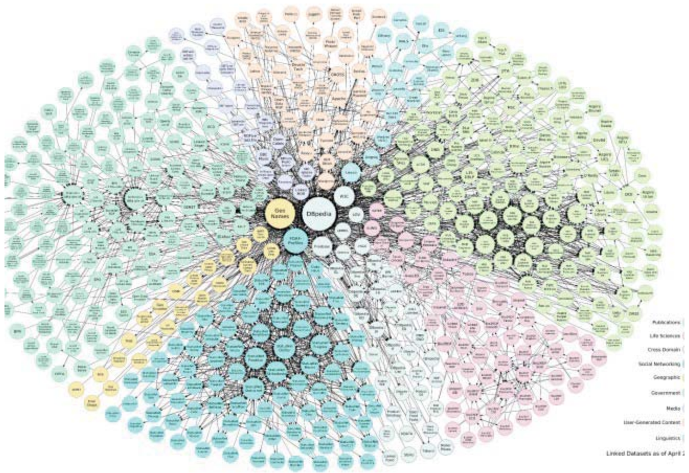


- **Size complexity**
  - Towards deep-learning ?
- **Heterogeneity complexity ?**
  - Knowledge-based methods

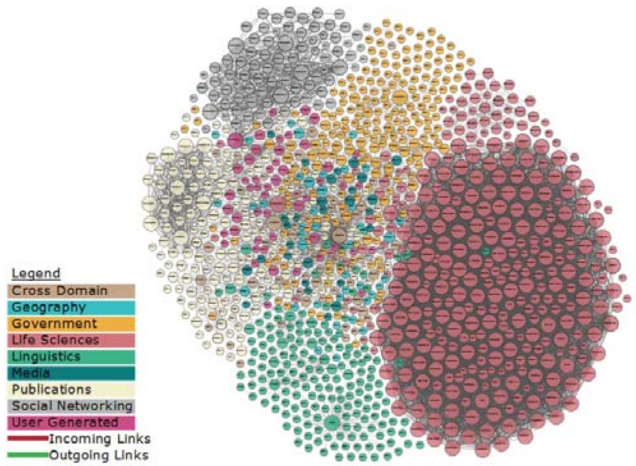
Linked open data

- **More than 1500 life science databases**
  - Gene Ontology
  - Chebi
  - KEGG
  - Swissprot…

- **Many of these DB are being linked and can be queried**
  - **Huge knowledge repositories to support reasoning**



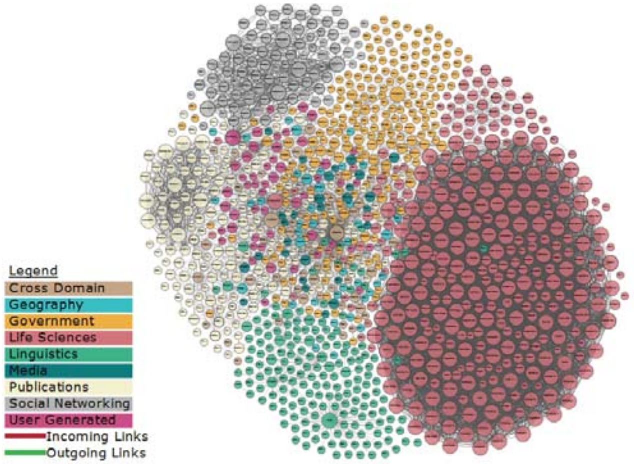
Linked Open Data initiative (2014)



Linked Open Data initiative (2017)

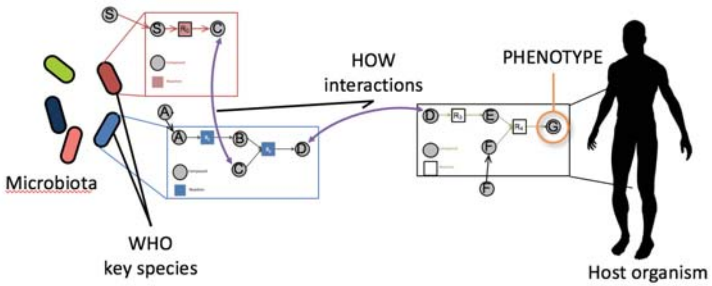
The futur of life-science data analysis ?

Machine learning : compound, function and species identification



Knowledge representation : Connect data

- Performant queries
- User-friendly interfaces



Formal approaches : explain

- Automatic reasoning
- **Assist biologists and never replace them**

Prospective

- **Our future role : facilitate and scale life science data analysis**
  - Easy exploration of search spaces
  - **Extract dynamical features as constraints** (temporal ?)
  - Use knowledge DBs

- **Always explain the results**
  - Give choices to experimentalists
  - **According to all the hypotheses that we make, biologists have to double-check our predictions.**

Acknowledgment

- **Equipe Dyliss@IRISA**

- Métabolisme@IRISA
  - C. Frioux
  - S. Prigent (INRA)
  - M. Aite
  - M. Chevallier
  - J. Got

- Algues@SBR Roscoff
  - S. Dittami
  - H. Kliijean
  - B. Burgunter
  - T. Tonon

- ASP tech@Potsdam university
  - T. Schaub's team
  - S. Thiele

- CMM@Univ Chile
  - A. Maass
  - M. P. Cortes
  - P. Bourdon

- Modélisation métabolisme@Nantes
  - D. Eveillard (LS2N)