Combinatorial completion for the reconstruction of metabolic networks, and application to the brown alga model *Ectocarpus siliculosus*

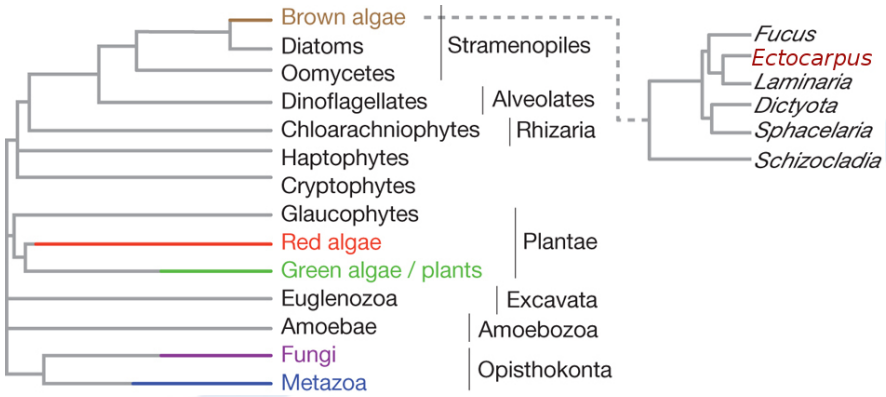
Sylvain Prigent

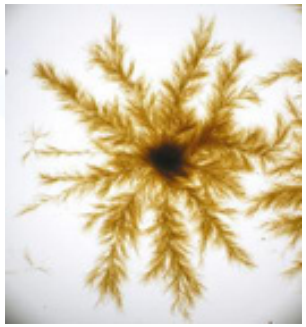
Dr Anne Siegel, IRISA
Dr Thierry Tonon, SBR

November 14th, 2014

Outlines

Ectocarpus siliculosus





- **Stramenopiles**:
  - Diverged from opisthokonta and plantae more than **1 billion years** ago
- **Secondary endosymbiosis**
  - Capture red alga $\Rightarrow$ plastids
- Evolved many **unusual characteristics**
  - **Adaptation** to intertidal zone
  - **Acclimation** to abiotic stresses

A complex evolutionary history

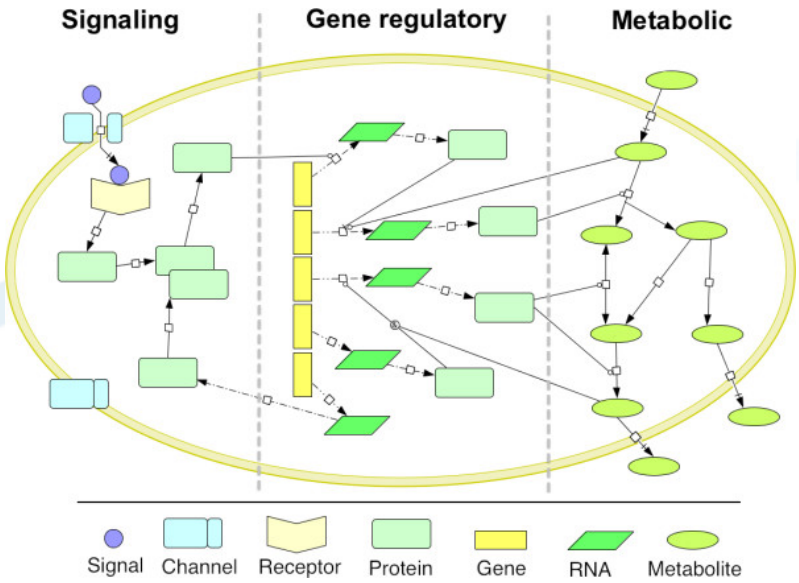
Cock et al., 2009

Ectocarpus siliculosus, available data

- An **annotated genome** (Cock *et al.*, 2010);
- **Transcriptomic** data (Dittami *et al.*, 2009);
- **Metabolite** profiling (Gravot *et al.* 2010, Dittami *et al.*, 2011);
- Knowledge on its **adaptation and acclimation** capacities to environmental changes

Can genomic data explain metabolite profiling, adaptation and acclimation capacities?

Systems biology

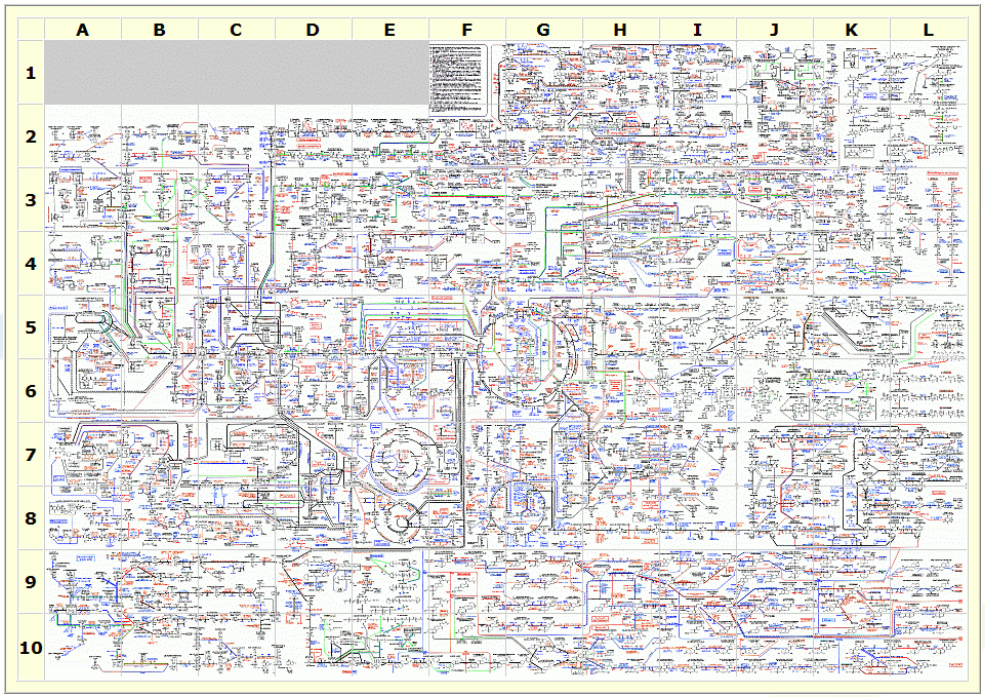


"*To understand complex biological systems requires the* **integration** *of experimental and computational research — in other words a* **systems biology** *approach.*" Kitano, 2002

Metabolic networks: relevant biological scale to study **functionality** and **adaptation**

Machado *et al.*, 2011

Metabolic networks



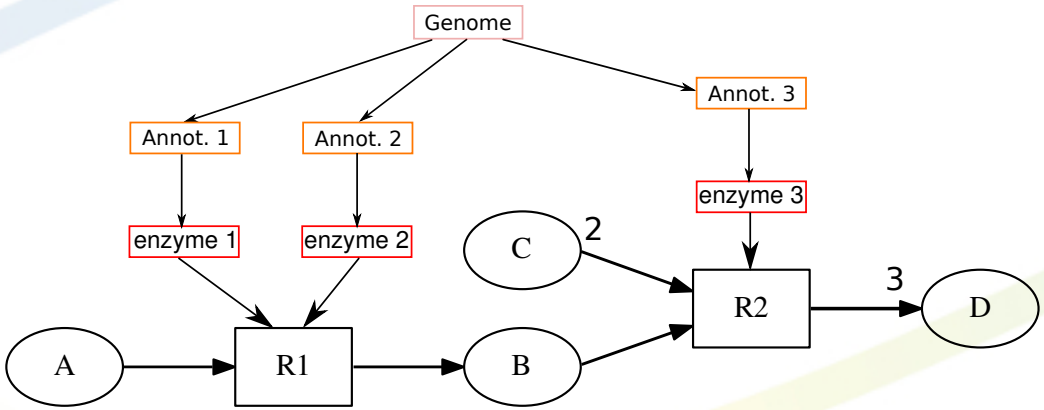
Metabolic network: complete set of metabolic reactions that determine the physiological and biochemical properties of a cell.

Large scale models of metabolic pathways

source: expasy

Metabolic networks

- Reactions:
  - R1: 1 A $\rightarrow$ 1 B
  - R2: B + 2 C $\rightarrow$ 3 D
- Network representation:



- Stoichiometric matrix:

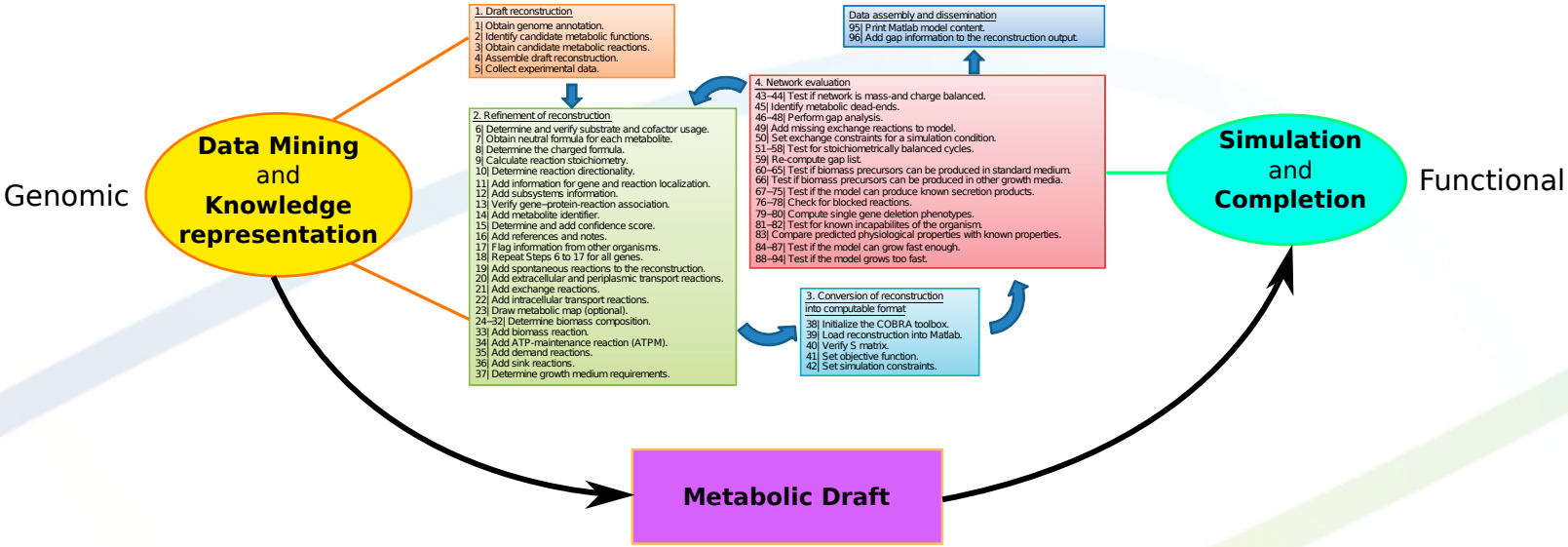
	R1	R2
A	-1	0
B	1	-1
C	0	-2
D	0	3

Studying metabolic networks using Mixed Integer Linear Programming

- Flux Balance Analysis
  - To predict **unique distribution of internal fluxes**
  - To hypothesize maximization of biomass: maximize $Z = c^T v$
- Flux Variability Analysis
  - To predict **range of fluxes related to biomass**
  - To maximize and minimize v
  - To identify 3 classes of reactions: **obligatory, blocked and alternatives**

Highly dependent on **stoichiometry**, **structure** and **cofactors equilibrium** of the network

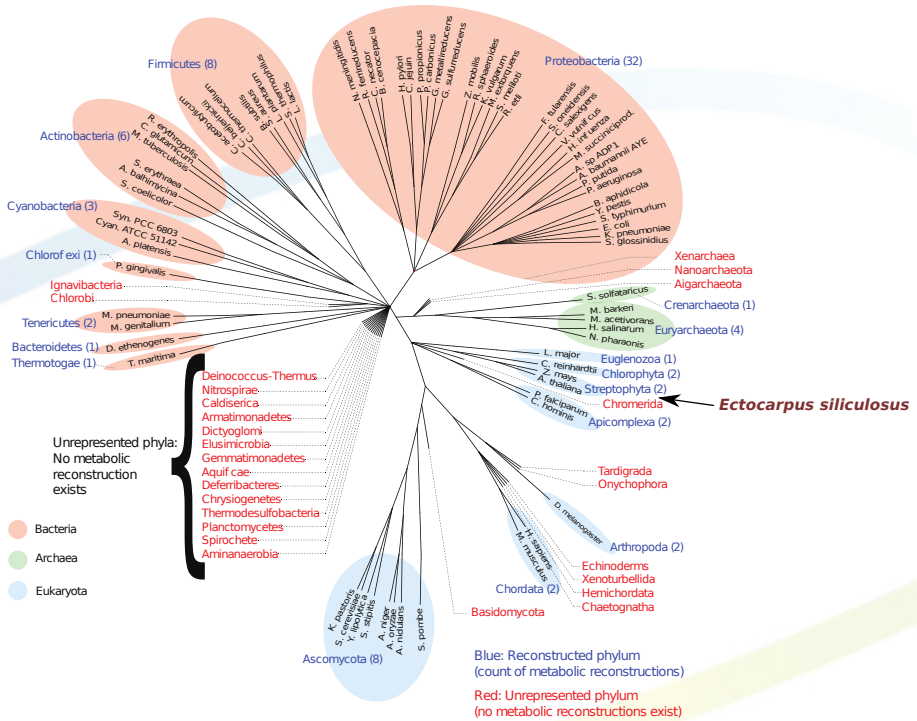
Metabolic networks reconstruction



Reconstructing metabolic networks: a task highly dependent on **data sources**

Thiele & Palson, 2010

Previous metabolic networks reconstruction



Automatic workflows exist for bacteria

- Microscope, the SEED, Pathway tools
- Rely on genome structure, genetic perturbations

Monk *et al.*, 2014

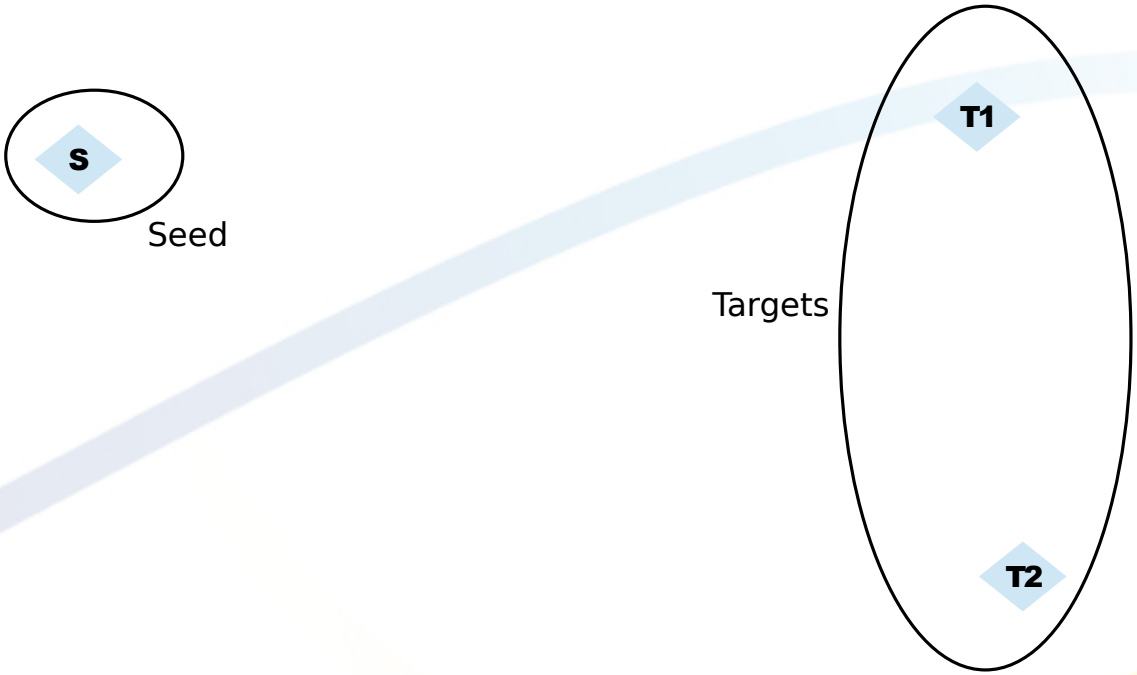
Metabolic network reconstructions for algae

Species	**Annotation format**	**Draft reconstruction**	**Functional refinement**
Chlamydomonas reinhardtii (1)	KEGG	KEGG extraction	Manual
Chlamydomonas reinhardtii (2)	Pre-existing network	Manual	No information
Ostreococcus (3)	KEGG	KEGG extraction	Automatic
Phaeodactylum tricornutum (4)	KEGG	KEGG extraction	Manual
Ectocarpus siliculosus	html pages	?	?

Data mining and **automatic refinement** are needed to reconstruct the metabolic network of *Ectocarpus siliculosus*

(1) Dal'Molin *et al.*, 2011 (2) Chang *et al.*, 2011 (3) Krumholz *et al.*, 2012 (4) Fabris *et al.*, 2012

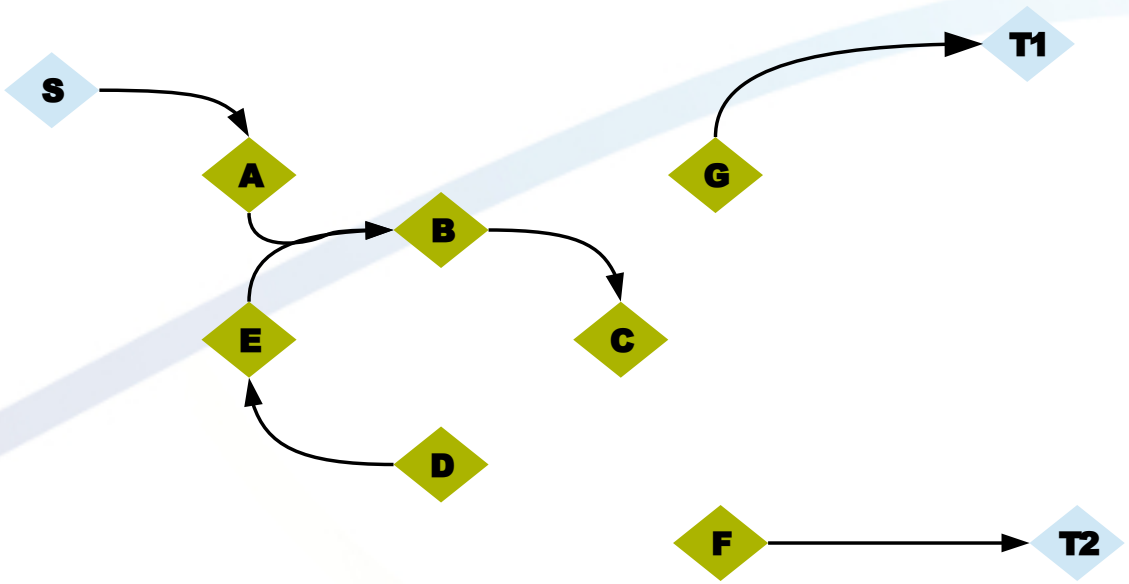
An overview of the completion problem



Problem: **minimizing** the number of added reactions to **produce** the targets from the seeds

An overview of the completion problem

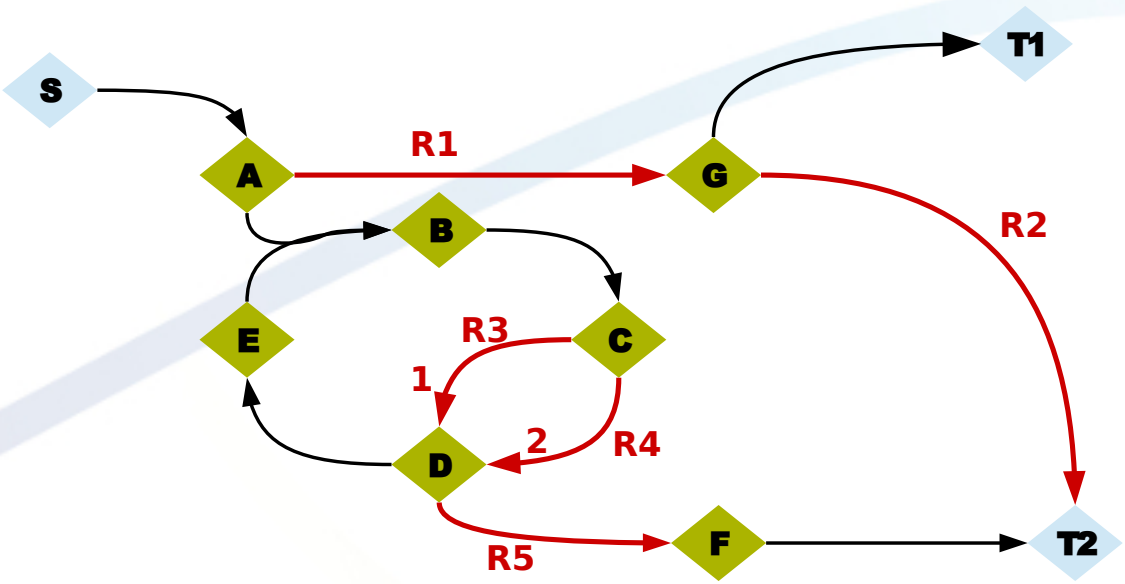
$\longrightarrow$ Draft $\longrightarrow$ Putative



Problem: **minimizing** the number of added reactions to **produce** the targets from the seeds

An overview of the completion problem

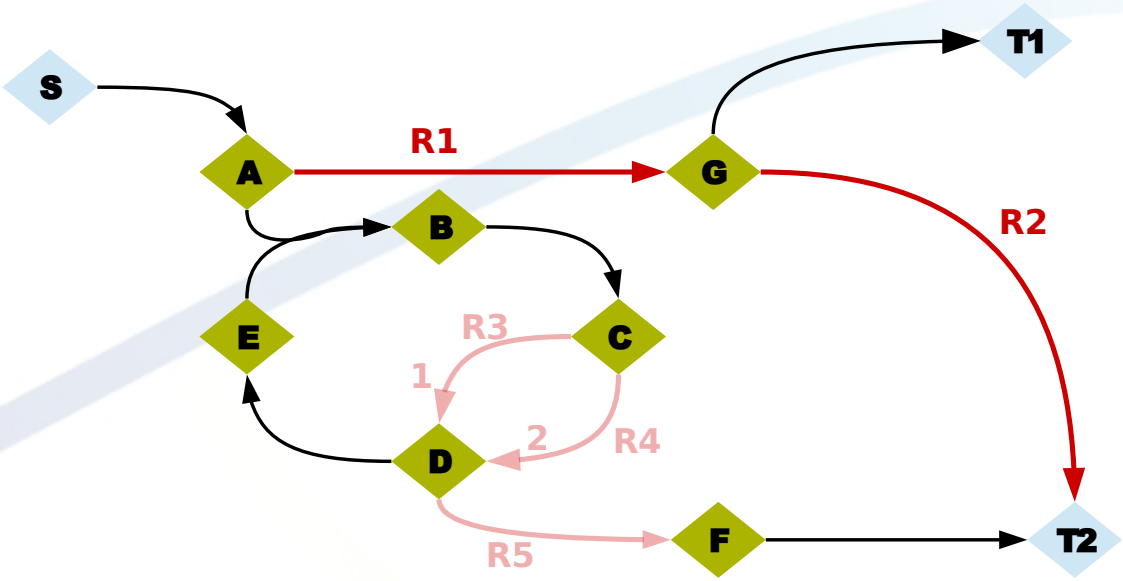
$\longrightarrow$ Draft $\longrightarrow$ Putative



Problem: **minimizing** the number of added reactions to **produce** the targets from the seeds

An overview of the completion problem

$\longrightarrow$ Draft $\longrightarrow$ Putative

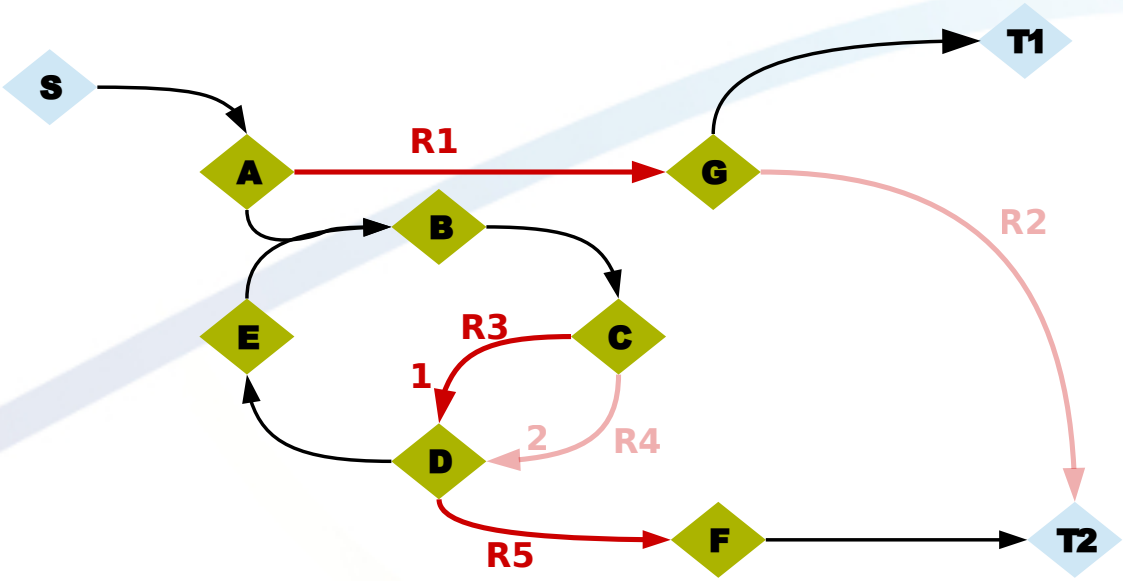


Completion	Minimality	Functional?
R1 R2	Cardinal	Yes

Problem: **minimizing** the number of added reactions to **produce** the targets from the seeds

An overview of the completion problem

$\longrightarrow$ Draft $\longrightarrow$ Putative

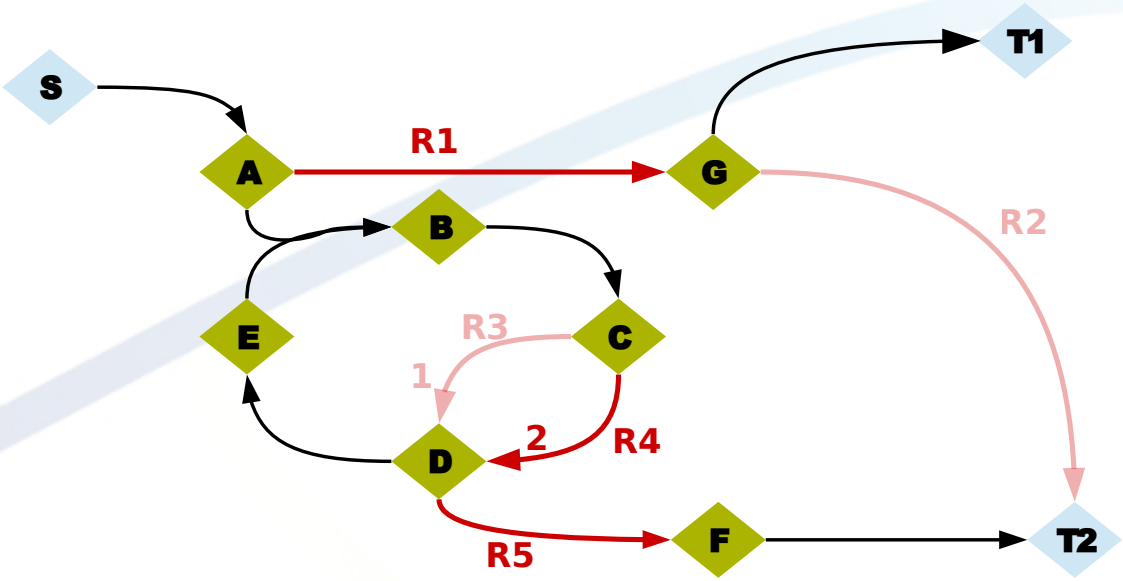


Completion	Minimality	Functional?
R1 R2	Cardinal	Yes
R1 R3 R5	Subset	No

Problem: **minimizing** the number of added reactions to **produce** the targets from the seeds

An overview of the completion problem

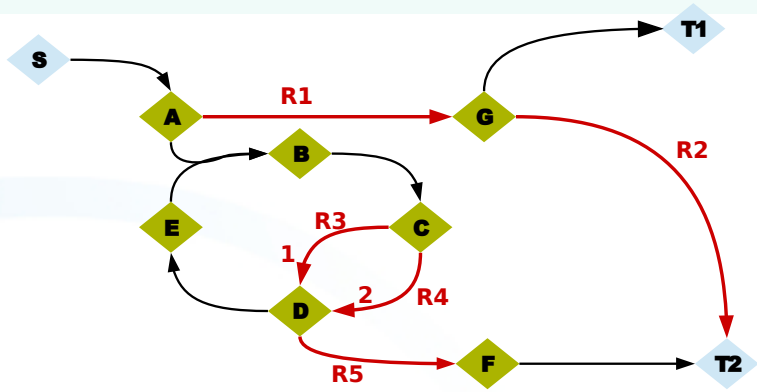
$\longrightarrow$ Draft $\longrightarrow$ Putative



Completion	Minimality	Functional?
R1 R2	Cardinal	Yes
R1 R3 R5	Subset	No
R1 R4 R5	Subset	Yes

Problem: **minimizing** the number of added reactions to **produce** the targets from the seeds

Description of the problem

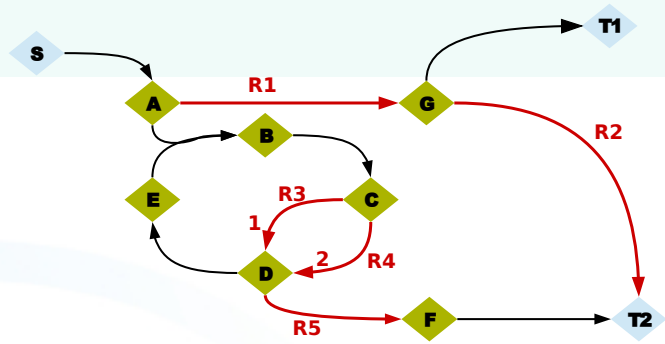


- Search space:
  - A *metabolic draft*: Directed bipartite graph R_{draft};
  - A *database of reactions*: R_{db};
  - A group of *metabolic seeds*: $M_{seed} \subset M$;
  - A group of *metabolic targets*: $M_{target} \subset M$;
  - The research space: $R = R_{draft} \cup R_{db}$
- Completion:
  - A group of reactions $R_{completion} \subseteq R_{db} \setminus R_{draft}$ such that:
    - M_{target} is reachable from M_{seed} in the network $((R_{draft} \cup R_{completion}) \cup (M_{draft} \cup M_{completion}), E_{draft} \cup E_{completion})$

Problem: find a minimal completion

Highly dependent on reachability

Metabolic network gap-filling



Name	Producibility	Minimality criteria	Completeness
Optstrain (1) & SMILEY (2)	FBA	Cardinal	Unique solution
GapFill (3)	FBA	Cardinal	Unique solution
Christian *et al.* (4)	Topologic	Subsets	Sampling
Network-expansion (5)	Topologic	Cardinal	Exhaustive

Are topologic studies precise enough to perform gap-filling?

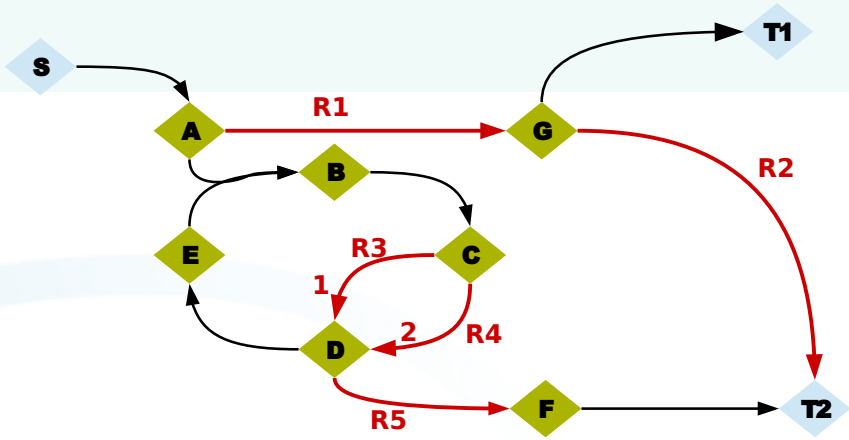
(1) Pharkya *et al.*, 2004 (2) Reed *et al.*, 2006 (3) Satish Kumar *et al.*, 2007 (4) Christian *et al.*, 2009 (5) Schaub and Thiele, 2009

Conclusion

- How can we perform accurate and exhaustive **gap-filling** that scales to targeted applications?
- Which kind of metabolic network **reconstruction pipeline** can we propose for non-classical species?
- Which **biological knowledge** do we gain by reconstructing the metabolic network of *Ectocarpus siliculosus*?

Outlines

Combinatorial problem



- A **topological study** of the network and the database
  - Implement producibility criteria proposed in (1)
  - Solve combinatorial problem

Reachability:

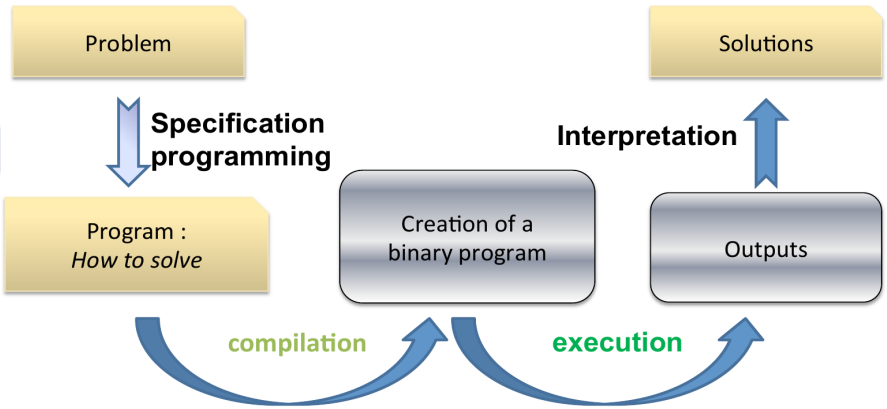
- A metabolite is producible iff:
  - It is a seed
  - It is a product of a reaction
    - If all reactants of this reaction are producible

Problem: find a minimal completion with respect to reachability

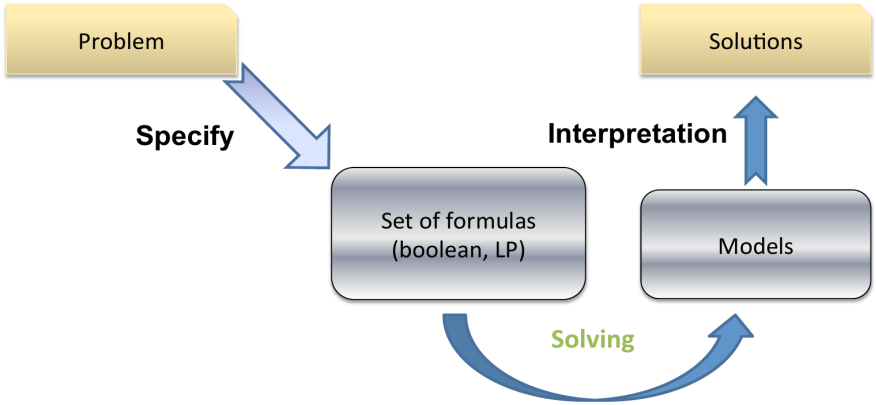
(1) Ebenhöh *et al.*, 2004

How to solve combinatorial problems ?

Dedicated Algorithm

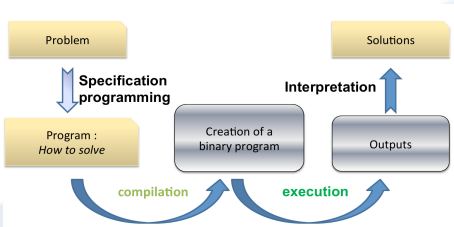


Use constraints solvers

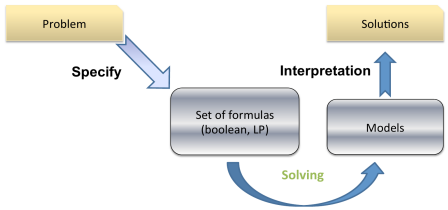


How to solve combinatorial problems ?

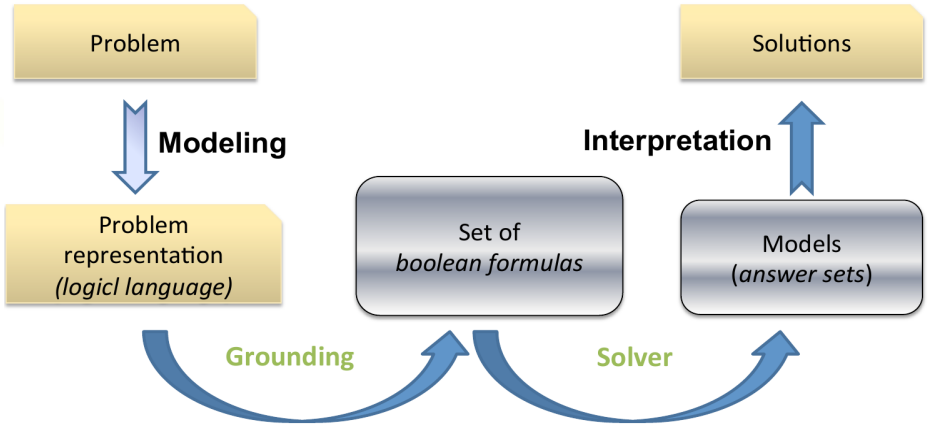
Dedicated Algorithm



Use constraints solvers



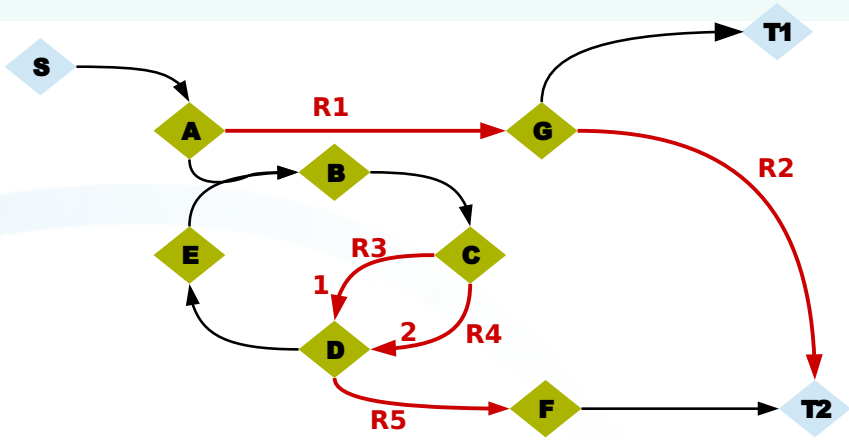
Answer Set Programming



Answer Set Programming in a nutshell

- Declarative programming
- High-level modeling language (ASP $\simeq$ Prolog expressivity)
  - The order of rules has no impact
  - **No infinite loops** in the resolution
- High performance solving capabilities (ASP $\simeq$ SAT, ILP)
  - SAT & deductive databases technics for ASP
  - Optimisation with **different heuristics**
- Different reasoning modes
  - Enumeration
  - Intersection
  - Union

Reachability in *Network-expansion*



- Topological study of the network

Reachability:

- A metabolite is producible iff:
  - It's a seed
  - It's a product of a reaction
    - If all reactants of this reaction are producible
- Computing scope of seeds in ASP:

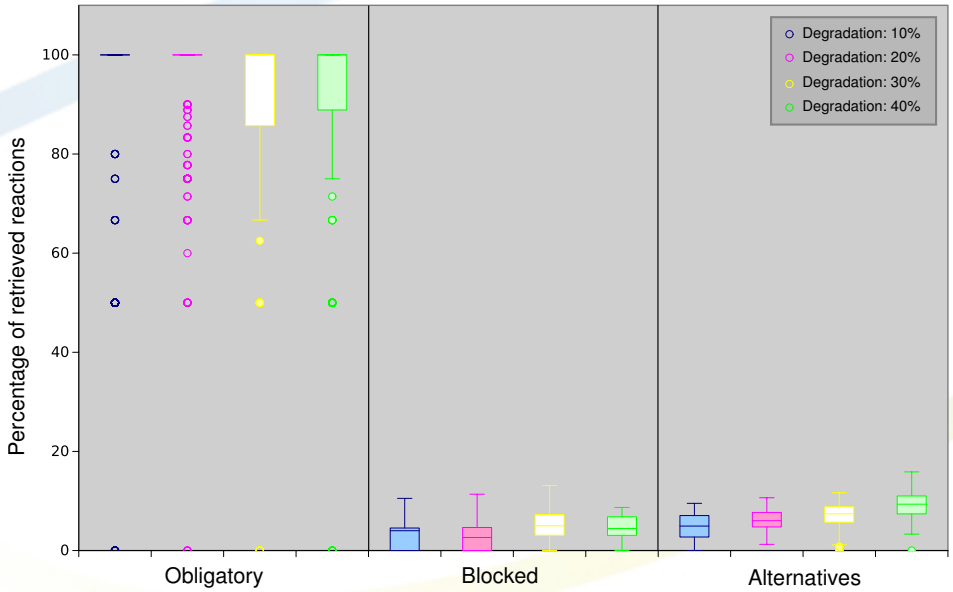
```
scope(M) :- seed(M).

scope(M) :- product(M,R), reaction(R,N), scope(M2) : reactant(M2,R).
```

Thiele & Schaub, 2009

Topologic completion VS stoichiometric studies?

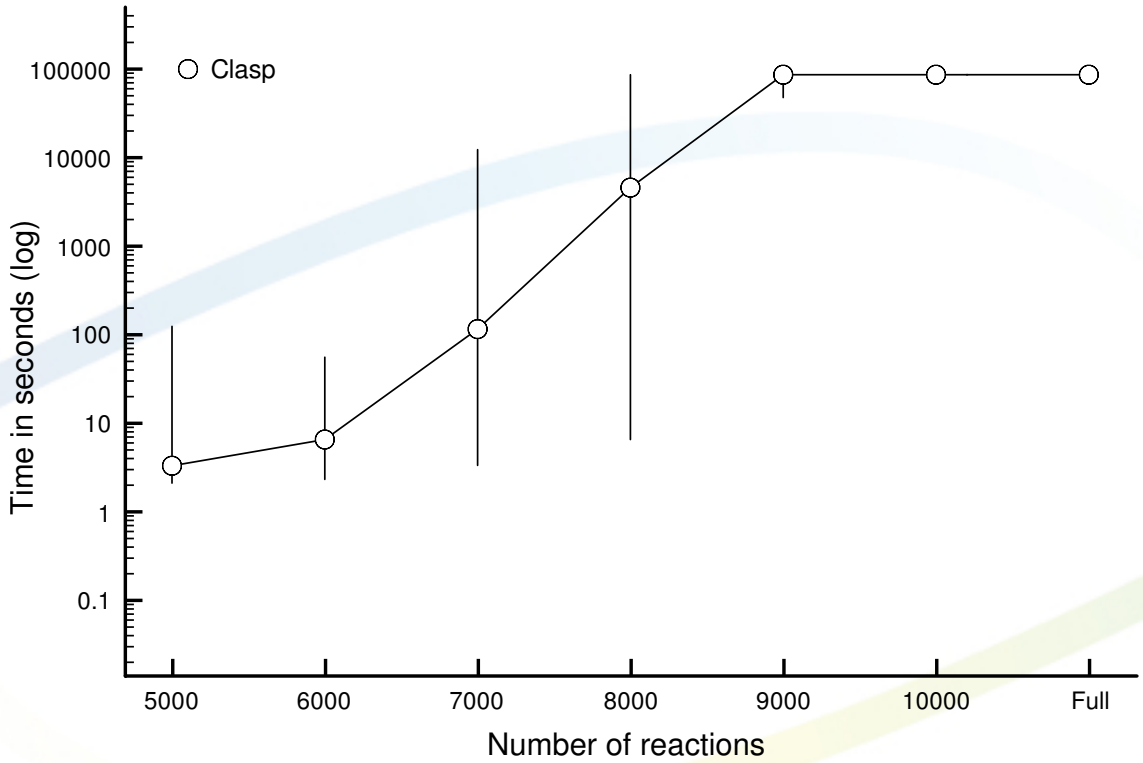
- Benchmark on Palsson's *E. coli* network
  - FVA: identification of **obligatory, blocked and alternative** reactions
  - Degradation of the network $\rightarrow$ 3.600 replicates



- Most of obligatory reactions are identified
- Blocked reactions are missed

Topological criteria are precise enough to recover functionality

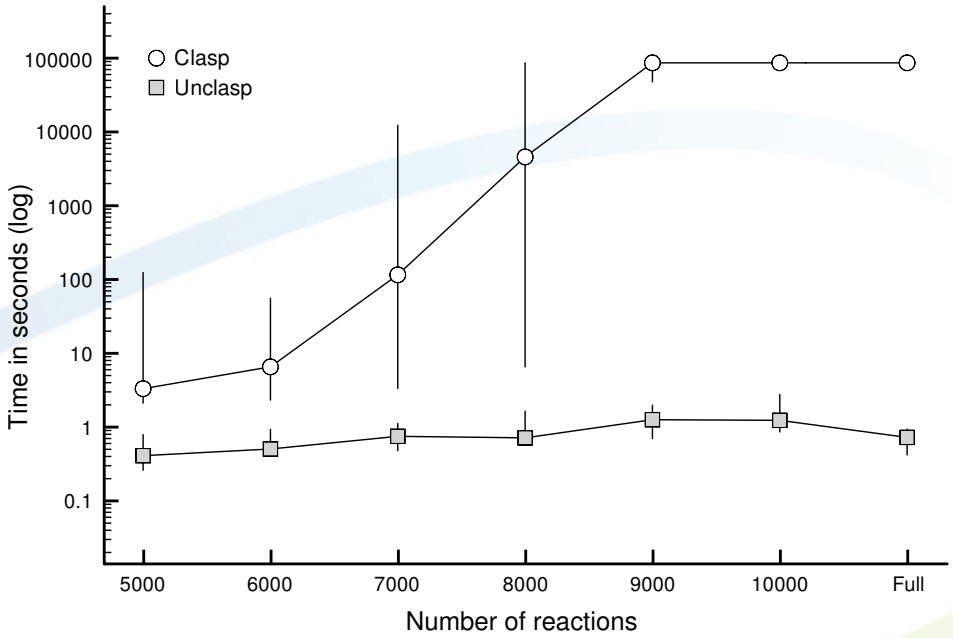
Limitations of *Network-expansion*



- **Do not scale** for large metabolic reactions databases
- Reversible reactions are splitting into two reactions

Improvements are mandatory

Changing solver (LPMNR 2013)



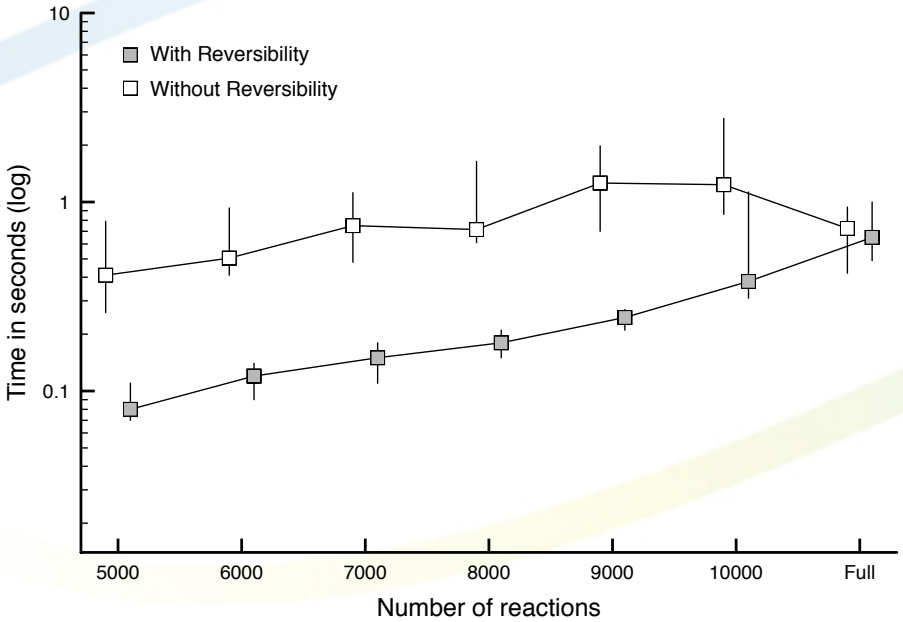
- Solution size is small ($\sim$ 10-100) with respect to size of the search space ($\sim$ 10.000)
- Use of a new ASP solver: constraints relaxations

Using **unsatisfiable cores** enables finding optima in **linear time**

Reversibility (LPMNR 2013)

New representation of reversibility in the encoding

- Fit with biological reality
- Smaller solution space



Improving biological relevance

Conclusion

- Topological criteria are efficient to do the completion
- Computation time improved by changing the solver
- Biological relevance improved by changing encoding of reversibility
- Collet *et al.*, LPNMR, 2013

$\Rightarrow$ **Meneco**

- Packaged into a python package
- Available online
  - http://mobyle.genouest.org/
  - http://bioasp.github.io/meneco/

Outlines

Building a metabolic draft

Functional annotation

- Genome annotations are not standardized
  - May loose information

Orthology research from cousin species

- Gene sequences have derived
  - Orthology search may fail

Pathway Tools

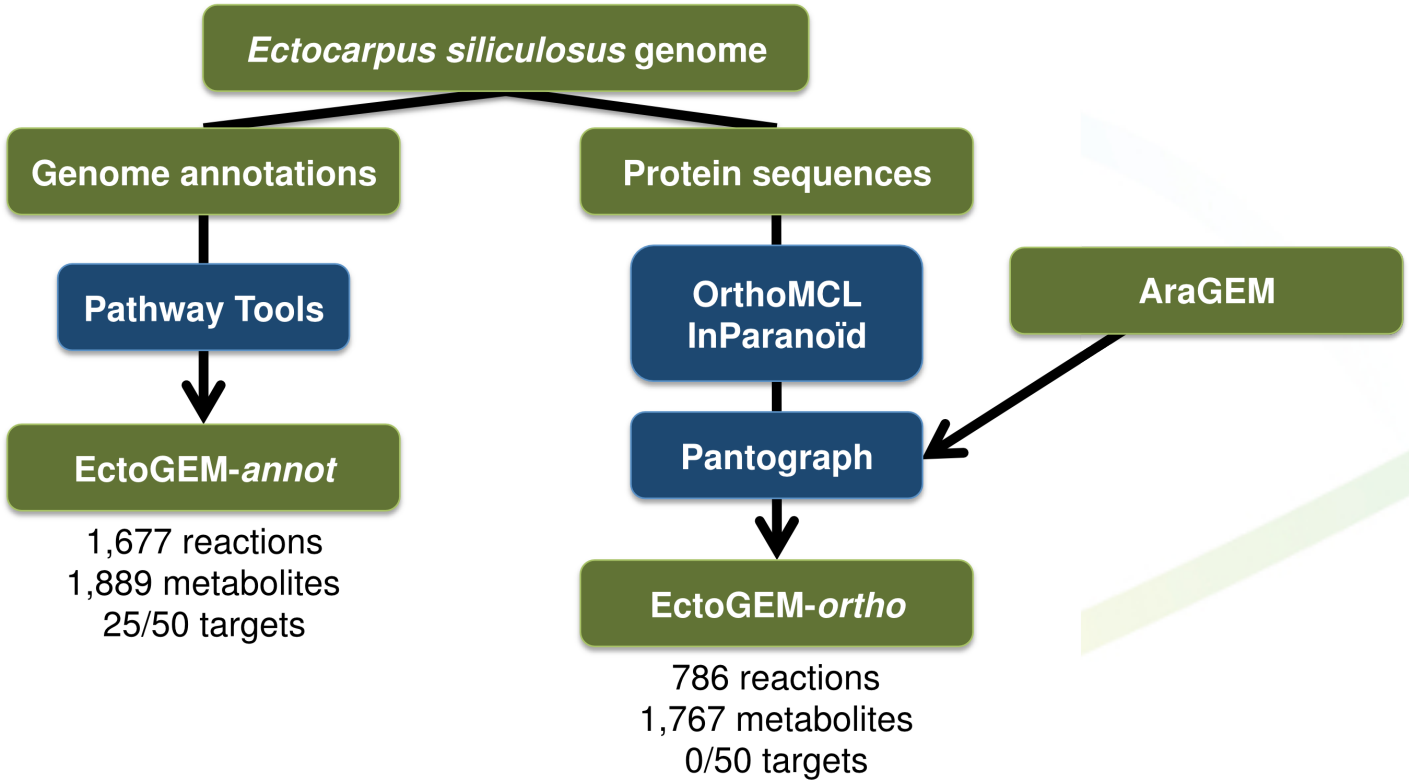
OrthoMCL DB
Ortholog Groups of Protein Sequences

InParanoid7
Eukaryotic Ortholog Groups

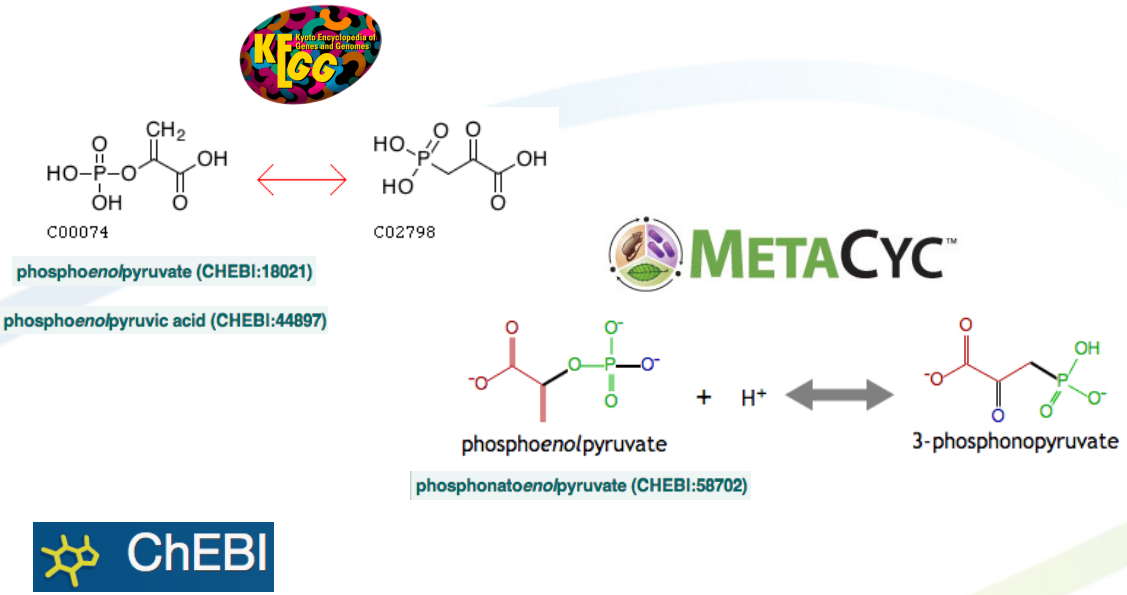
PANTOGRAPH

Combining annotations and orthology information to improve draft reconstruction

Building a metabolic draft for *Ectocarpus siliculosus*



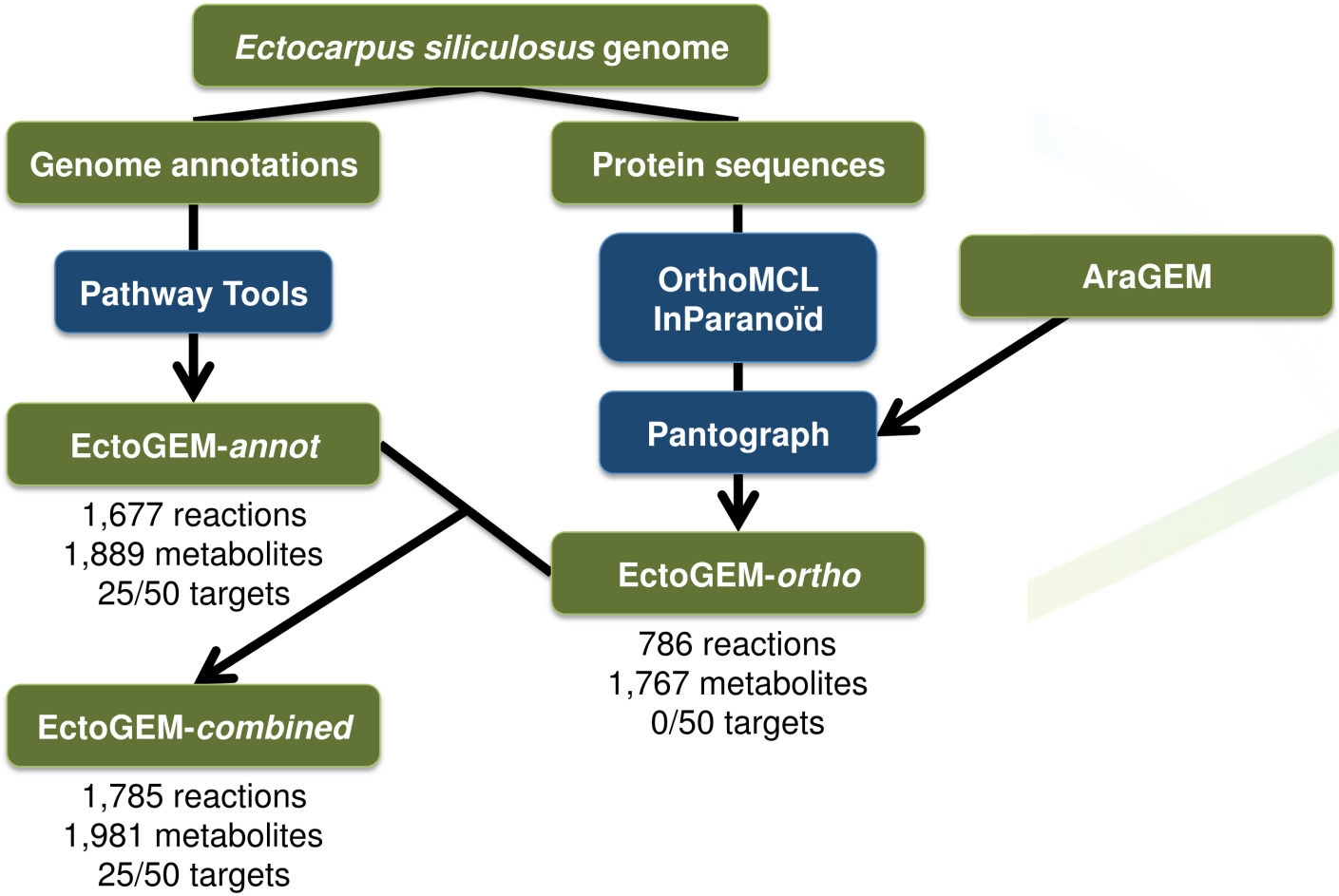
Merging two metabolic drafts



- If both draft are not based on the same database
  - Unification of identifiers needed
  - Cross-references
  - Same reactants & products ⇒ same reaction

⇒ **MeMap/MeMerge**

Merging metabolic drafts for *Ectocarpus siliculosus*



Meneco

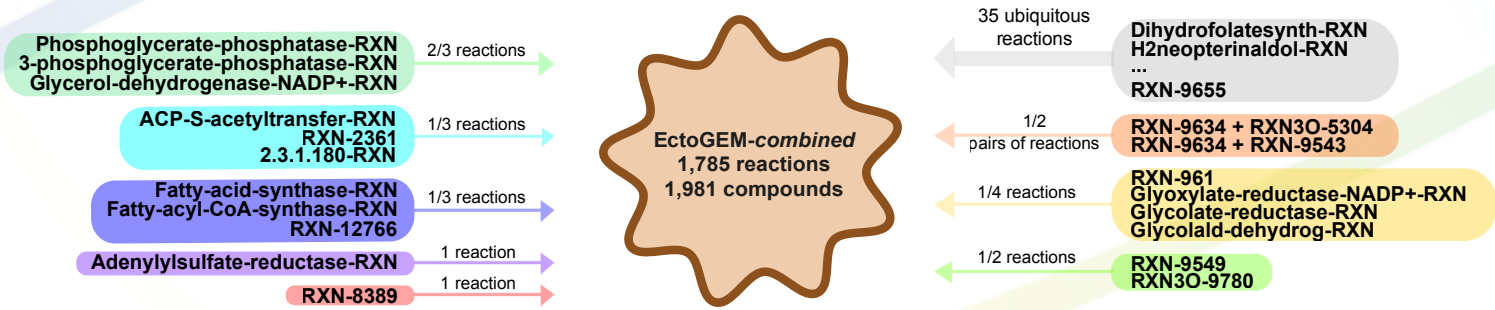
- 25 of 50 target metabolites not producible
- Completion using MetaCyc database and Meneco
  - $\sim$ 1 hour for the union

- **Minimal number** of reactions to add in the network: 44
- 4.320 **different sets** of 44 reactions can fill the network
- **Union** of these sets: 60 reactions

Completion is highly combinatorial

Semantic analysis of the 4.320 solutions

- 35 reactions are ubiquitous
- Some reactions are mutually exclusive
  - Never present together in the same completion
  - Should be biologically equivalent



- Before: 60 reactions, 4.320 completions
- After: 56 reactions, 432 completions

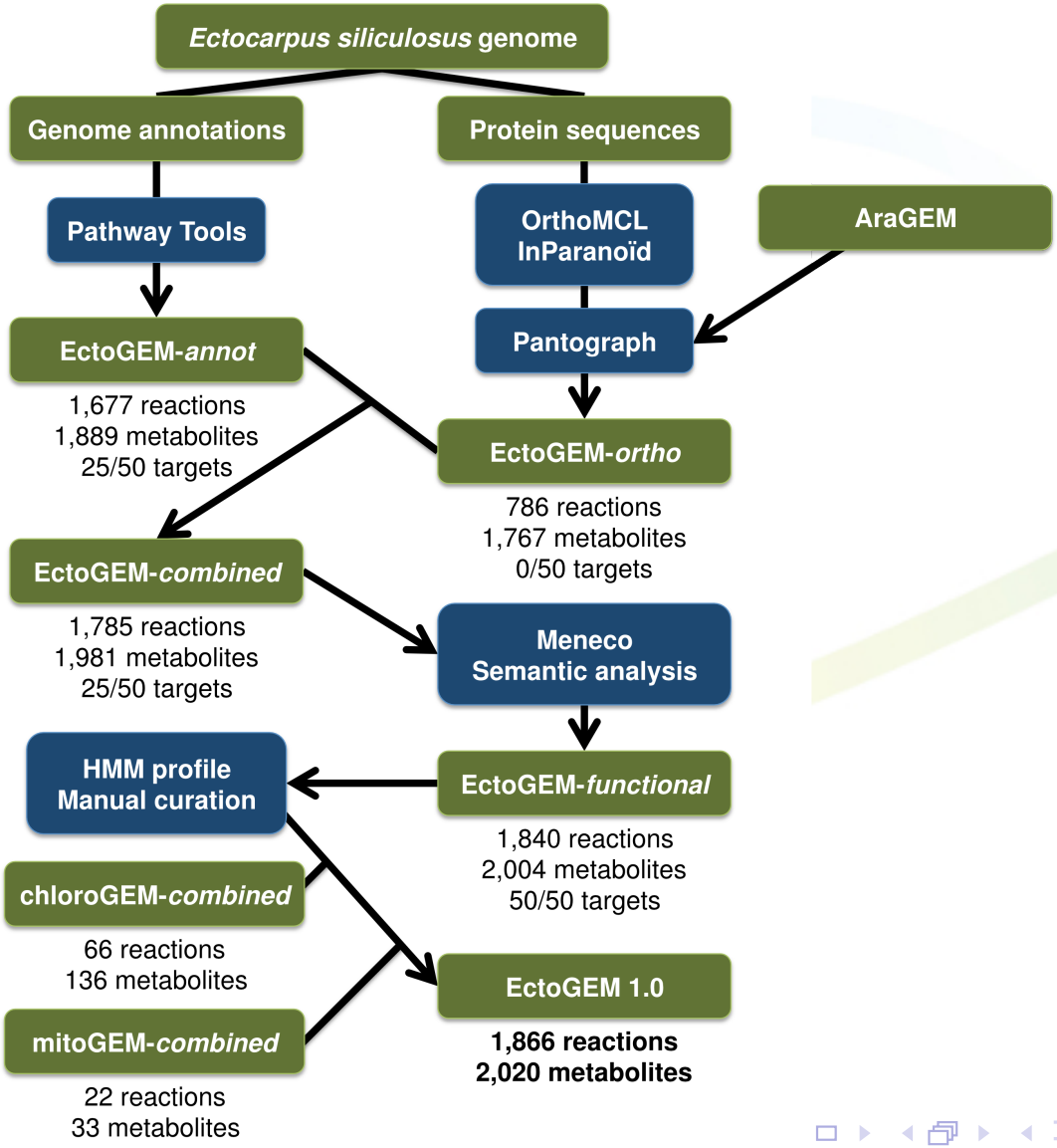
Semantic analysis reduced combinatorial of the completion

Looking for enzymes in the genome

- Proposed completions should have a biological relevance
- For each reaction:
  - Construct an Hidden Markov Model based on existing sequences
  - Search for this model in the genome
- If match found:
  - Gene previously not or badly annotated
  - Helping manual curation

Focus on particular enzymes provides new insights into the reannotation of the genome

EctoGEM 1.0


Ectocarpus siliculosus genome
Genome annotations
Protein sequences
Pathway Tools
OrthoMCL
InParanoïd
AraGEM
EctoGEM-*annot*
Pantograph
1,677 reactions
1,889 metabolites
25/50 targets
EctoGEM-*ortho*
786 reactions
1,767 metabolites
0/50 targets
EctoGEM-*combined*
1,785 reactions
1,981 metabolites
25/50 targets
Meneco
Semantic analysis
HMM profile
Manual curation
EctoGEM-*functional*
1,840 reactions
2,004 metabolites
50/50 targets
chloroGEM-*combined*
66 reactions
136 metabolites
EctoGEM 1.0
mitoGEM-*combined*
1,866 reactions
2,020 metabolites
22 reactions
33 metabolites

Conclusion

- Data mining and knowledge representation
  - Combining data sources
- Automatic combinatorial completion
  - Many solutions but not so much reactions
  - Scaling
- Towards an automatic workflow
- Helping manual curation

Pre-treatement and post-treatment of data are **mandatory**

Outline

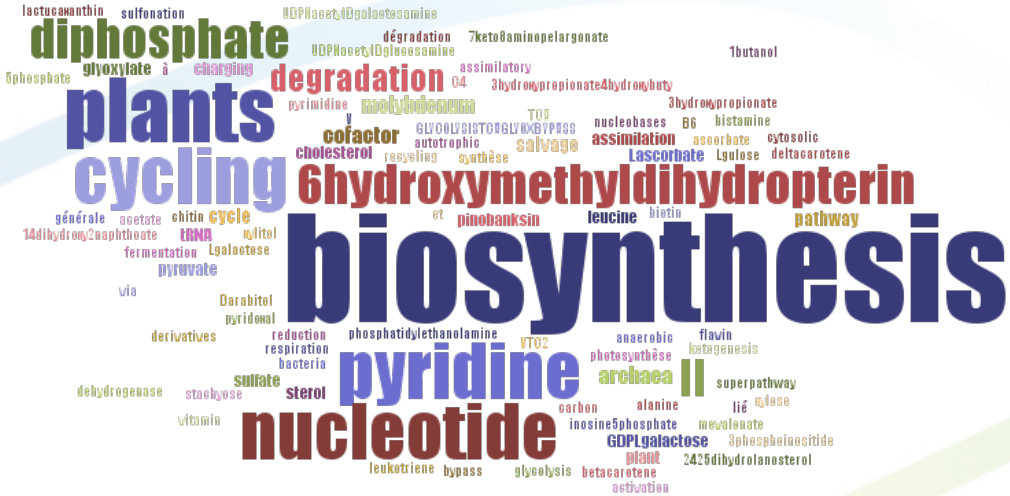
Functionality of the obtained network

- Development of a specific biomass function
  - Bibliographic study
  - 30 metabolites
- Flux Balance Analysis study

Network functionally valid

Topologic completion was sufficient to have a functional network

Reannotations



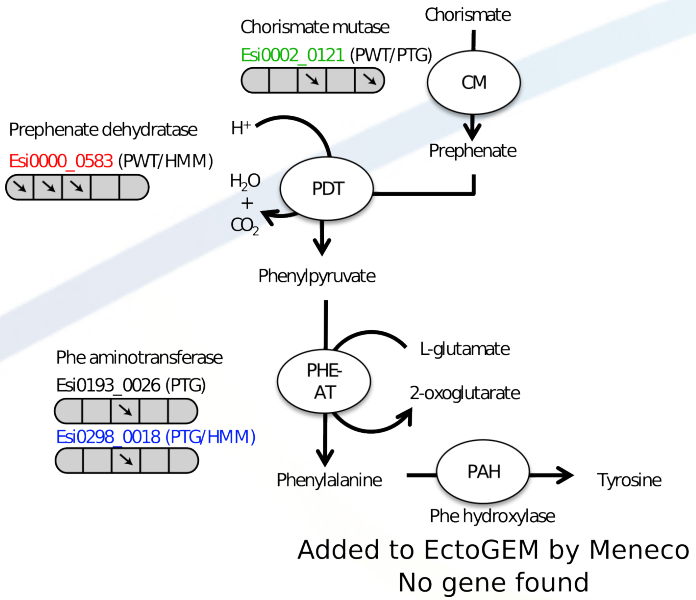
Words proportion of pathways in which genes are involved

56 genes reannotated

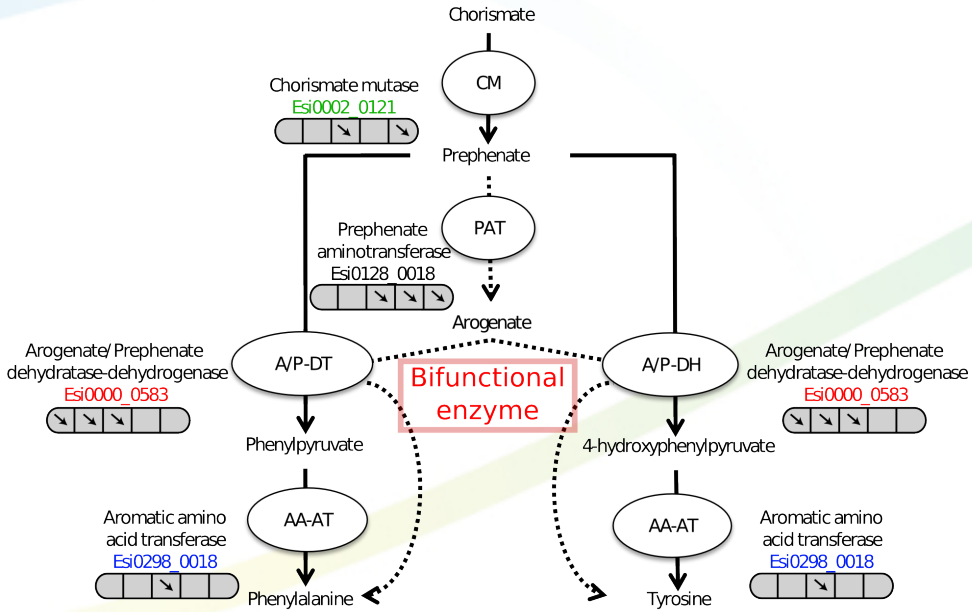
Reannotation of biosynthesis pathways

Aromatic amino acid biosynthesis

Proposed after automatic workflow

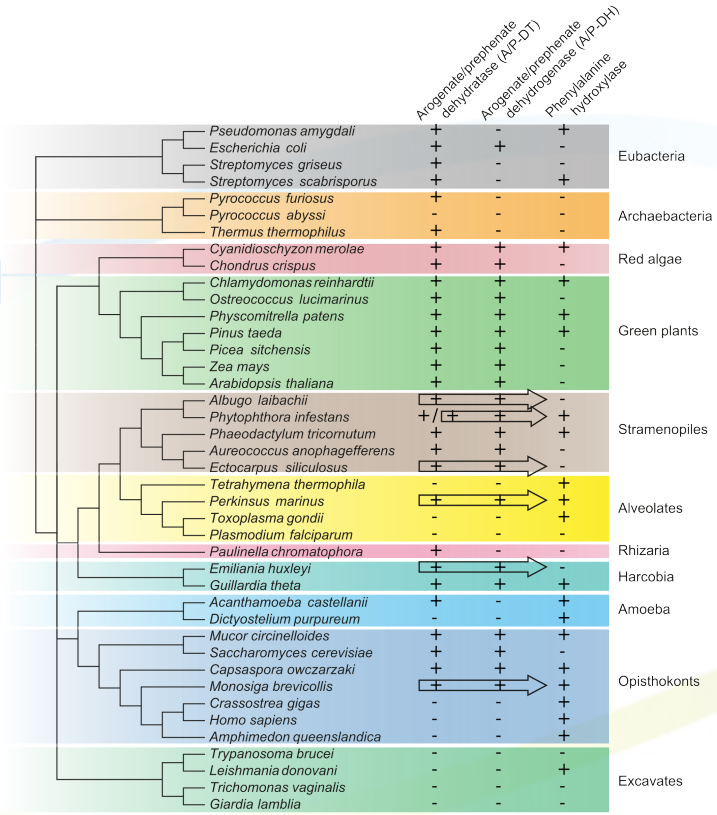


Proposed after manual curation



Reconstruction of metabolic network pinpoints a different pathway when compared to other stramenopiles

Aromatic amino-acid biosynthesis



Arrows: bifunctional enzymes

New insights into the evolution of aromatic amino acids synthesis

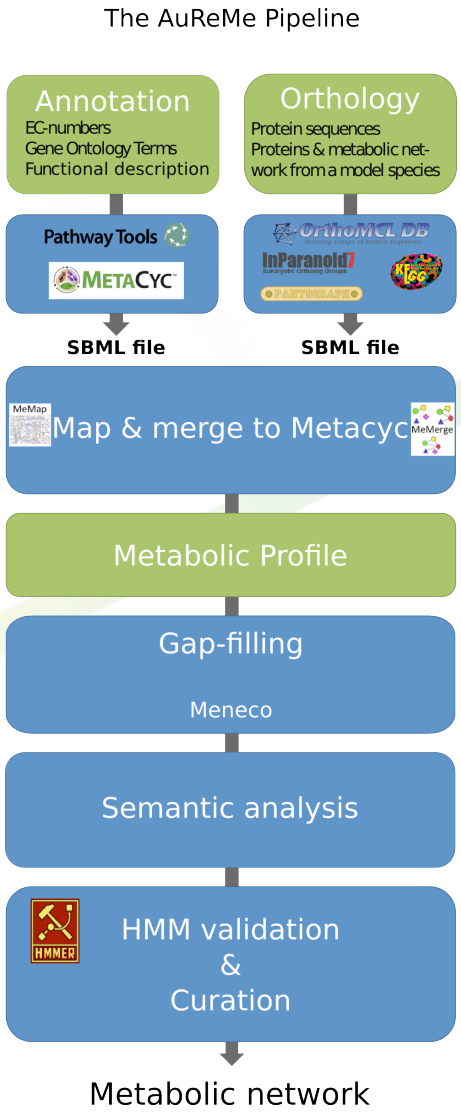
Conclusion

- Reconstruction process provides new insights into the physiology of organisms
- Reconstruction of *Ectocarpus siliculosus* metabolic network enables a better understanding of:
  - Metabolism of *Ectocarpus siliculosus*
  - Evolution of aromatic amino acid biosynthesis

Outline

Conclusion

- Topologic completion
  - Sufficient to obtain a functional network
- Semi-automatic pipeline to reconstruct metabolic networks
- New insights into the evolution of *Ectocarpus siliculosus*
- Reannotation of the genome



Perspectives in bioinformatics

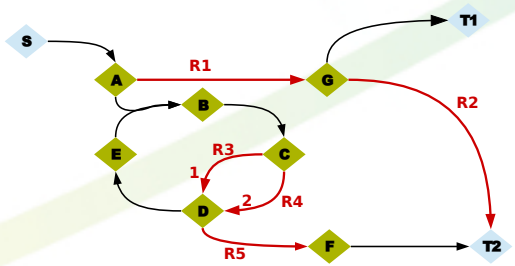
Improving the network

- New metabolite profiling
  - Better completion
- New RNA-seq data
  - How to include them in the pipeline?
- *Ectocarpus siliculosus* associated with bacteria
  - Study the association between metabolic networks of different origins
  - Holobiont metabolic network

Perspectives in computer science

Continue improvements on *Meneco*

- Studying subset minimality
  - New incremental solving in ASP
- Deepest study of effect of cycles
- New semantics of productibility
  - Preliminary results: totally different completions

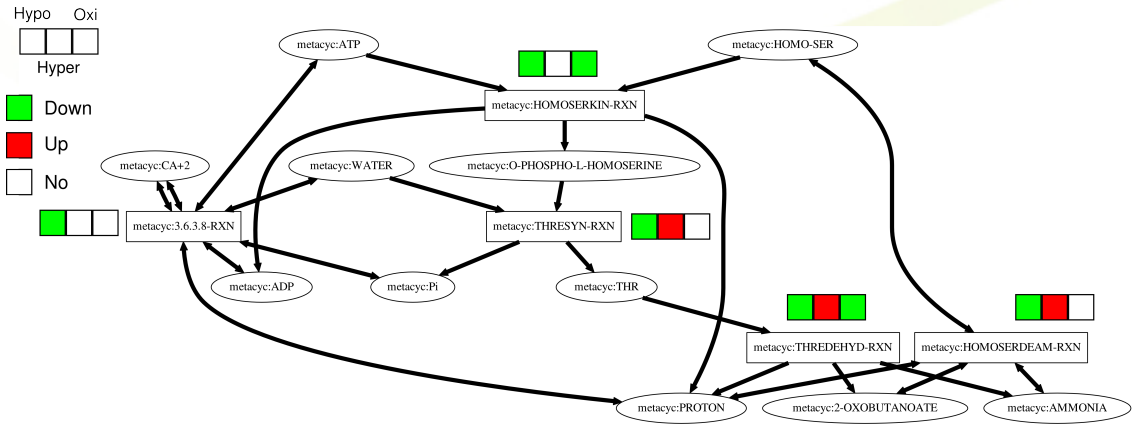


Perspectives in biology

Study adaptation and acclimation

- Freshwater species of *Ectocarpus* is sequenced
  - Reconstruct its metabolic network
  - Comparative metabolic network analysis

- Use transcriptomic data
  - Decomposition of the network
  - Found up- or down-regulated parts of this network in response to stress



Thanks for your attention