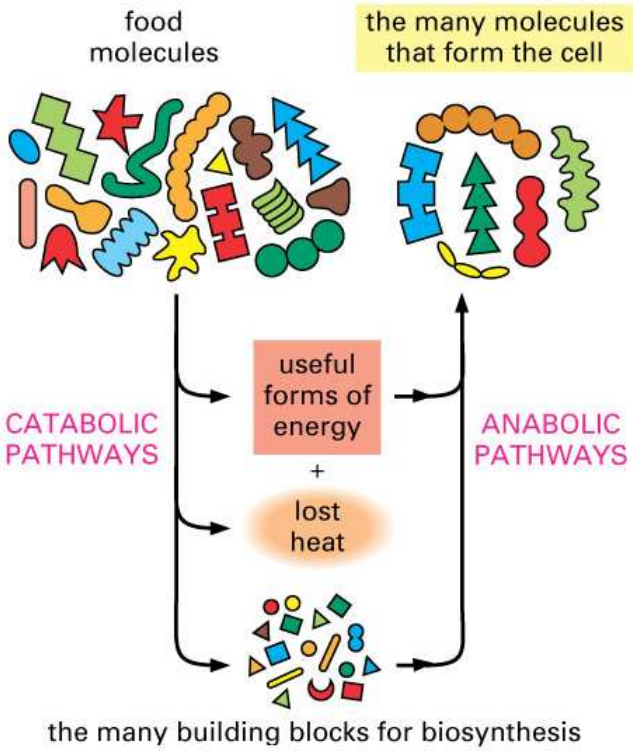
Motif Identification in Metabolic Networks

Vincent Lacroix

Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS 5558 - INRIA
Université Claude Bernard - Lyon 1

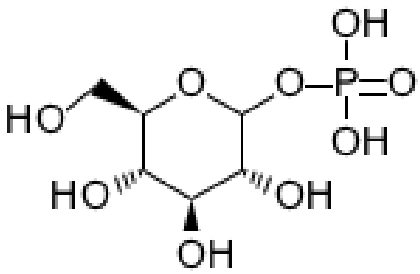
Research advisor: Marie-France Sagot

Metabolism

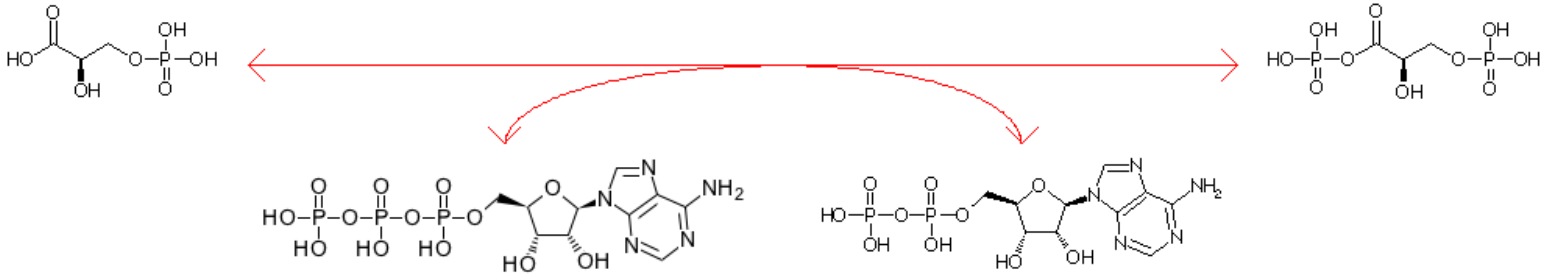

food molecules
the many molecules that form the cell
useful forms of energy
+
lost heat
CATABOLIC PATHWAYS
ANABOLIC PATHWAYS
the many building blocks for biosynthesis

Metabolites and reactions

- Metabolites (compounds)

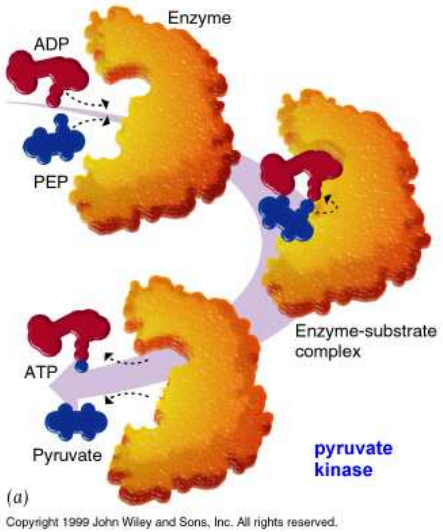


- Reactions



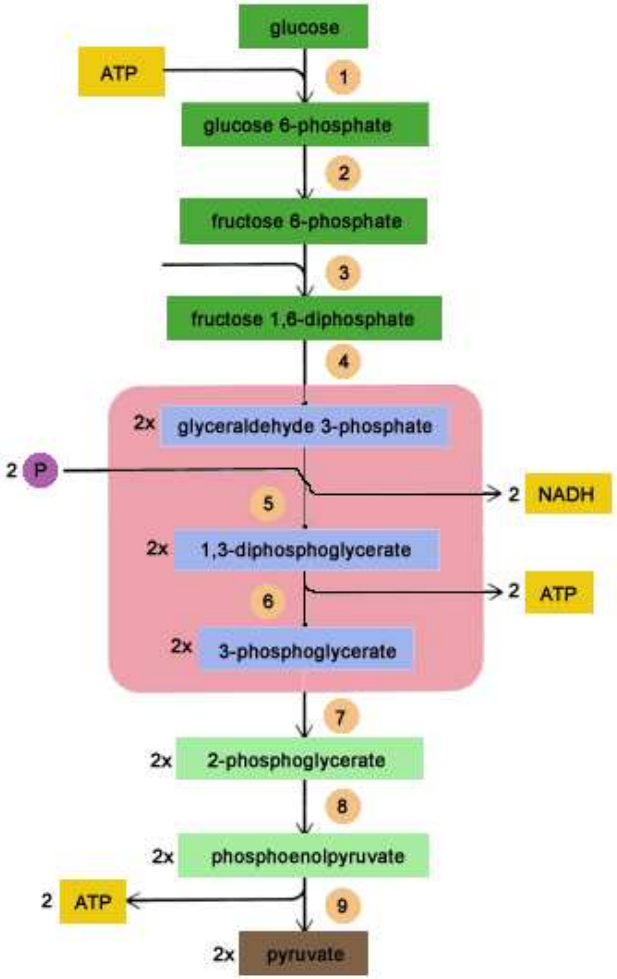
Enzymes

- Enzymes catalyse reactions

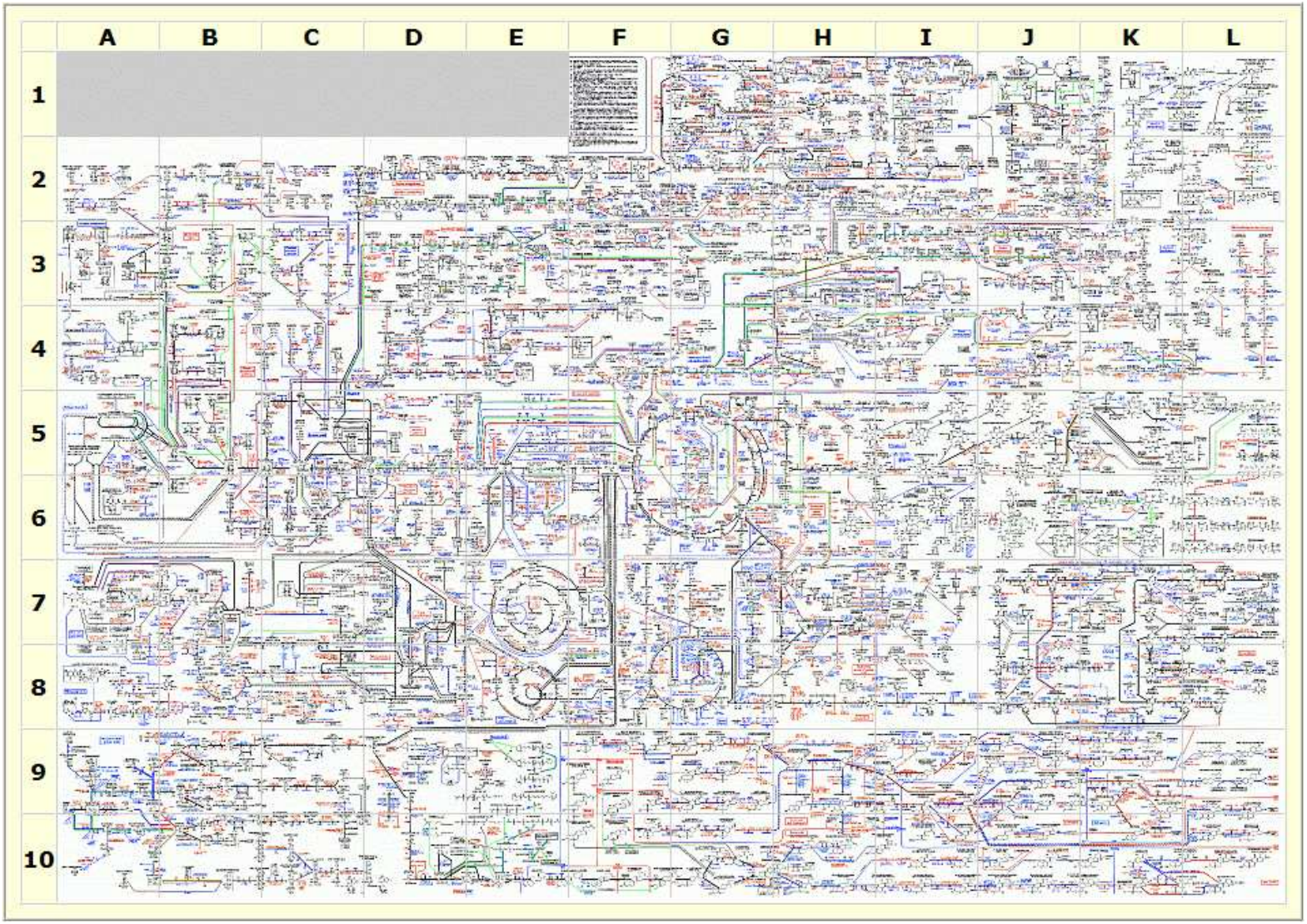


- The "EC" classification:
  Every enzyme is assigned a code with 4 numbers expressing the chemistry of the reaction it catalyses
  Ex : 1.1.2.3

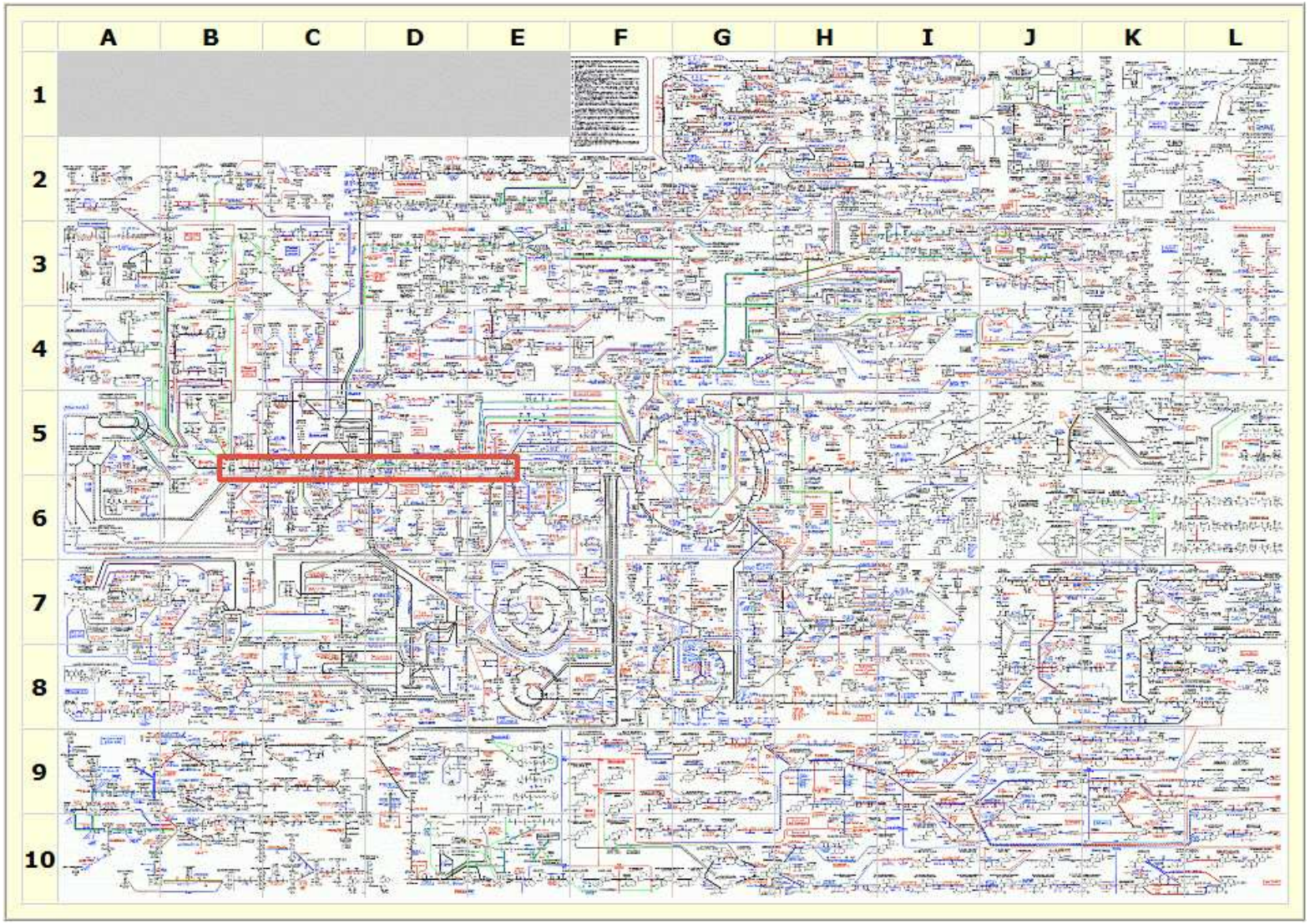
Metabolic Pathway: Glycolysis



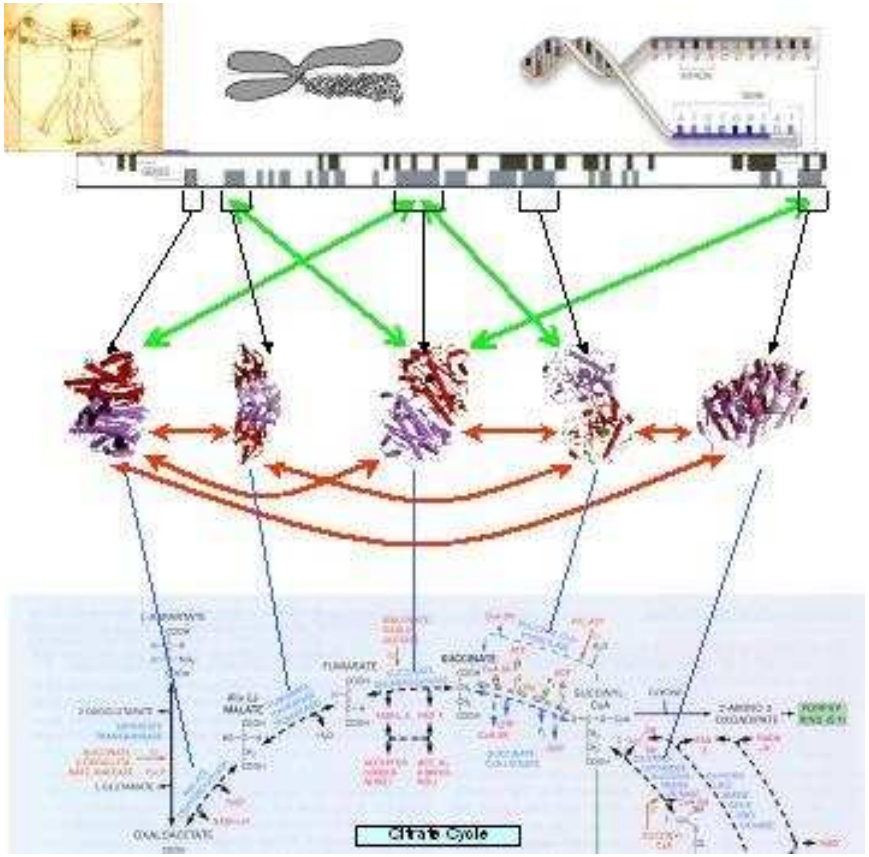
Metabolic Network



Metabolic Network



Biological networks



Gene Regulatory Network

Protein Interaction Network

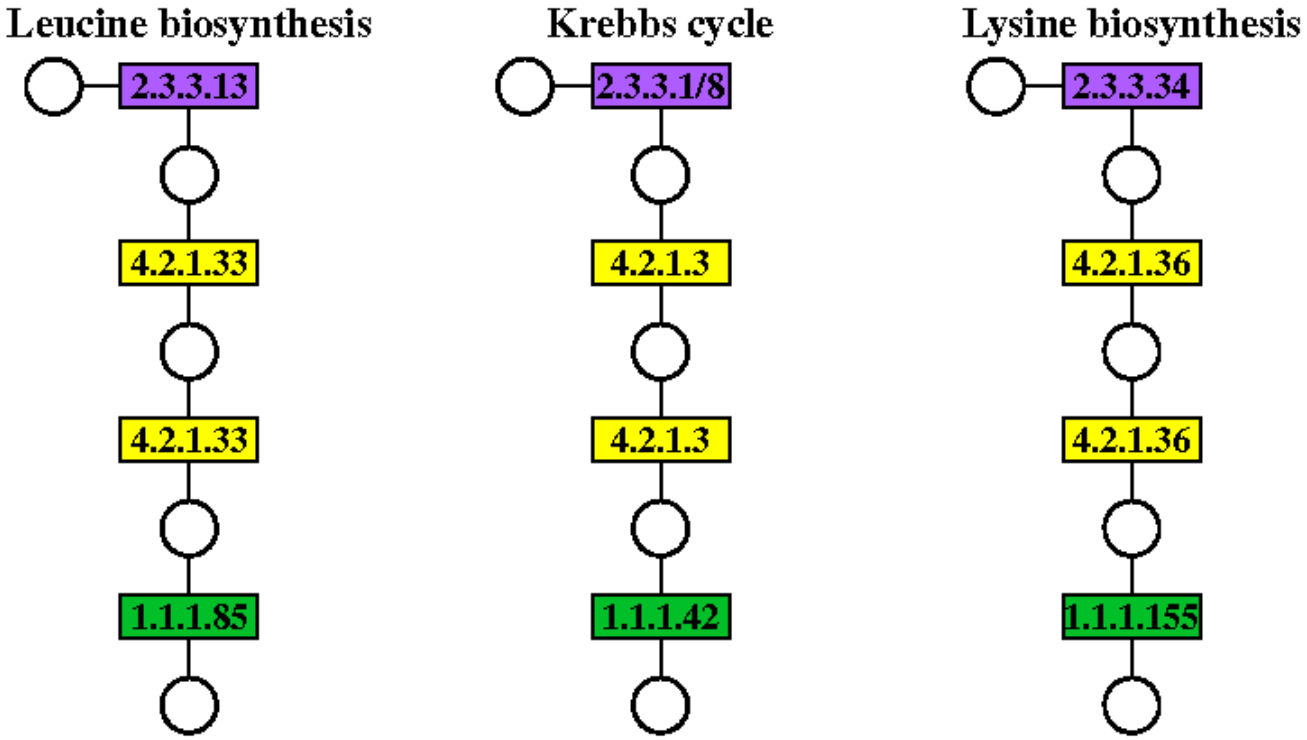
Metabolic Network

General motivation

General Motivation: understand the structure of the metabolic network, and the way it has been set up in the course of evolution

Question : can we define and identify functional and/or evolutionary units in a metabolic network ?

Repeated elements in metabolic networks



- Velasco, A.M., Leguina, J.I. and Lazcano, A. (2002) Molecular Evolution of the Lysine Biosynthetic Pathways, *J. Mol. Evol.*, **55**, 445-459.

Repeated elements in metabolic networks

Can we detect such regularities in a systematic way ?

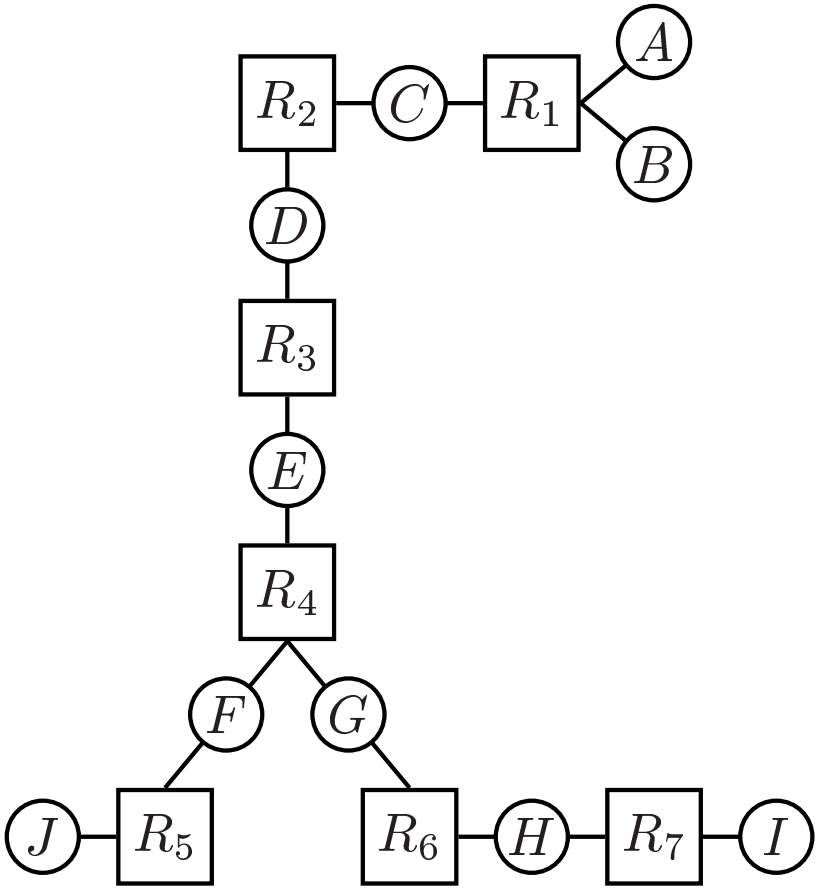
Network models

Several types of models have been proposed for metabolic networks:

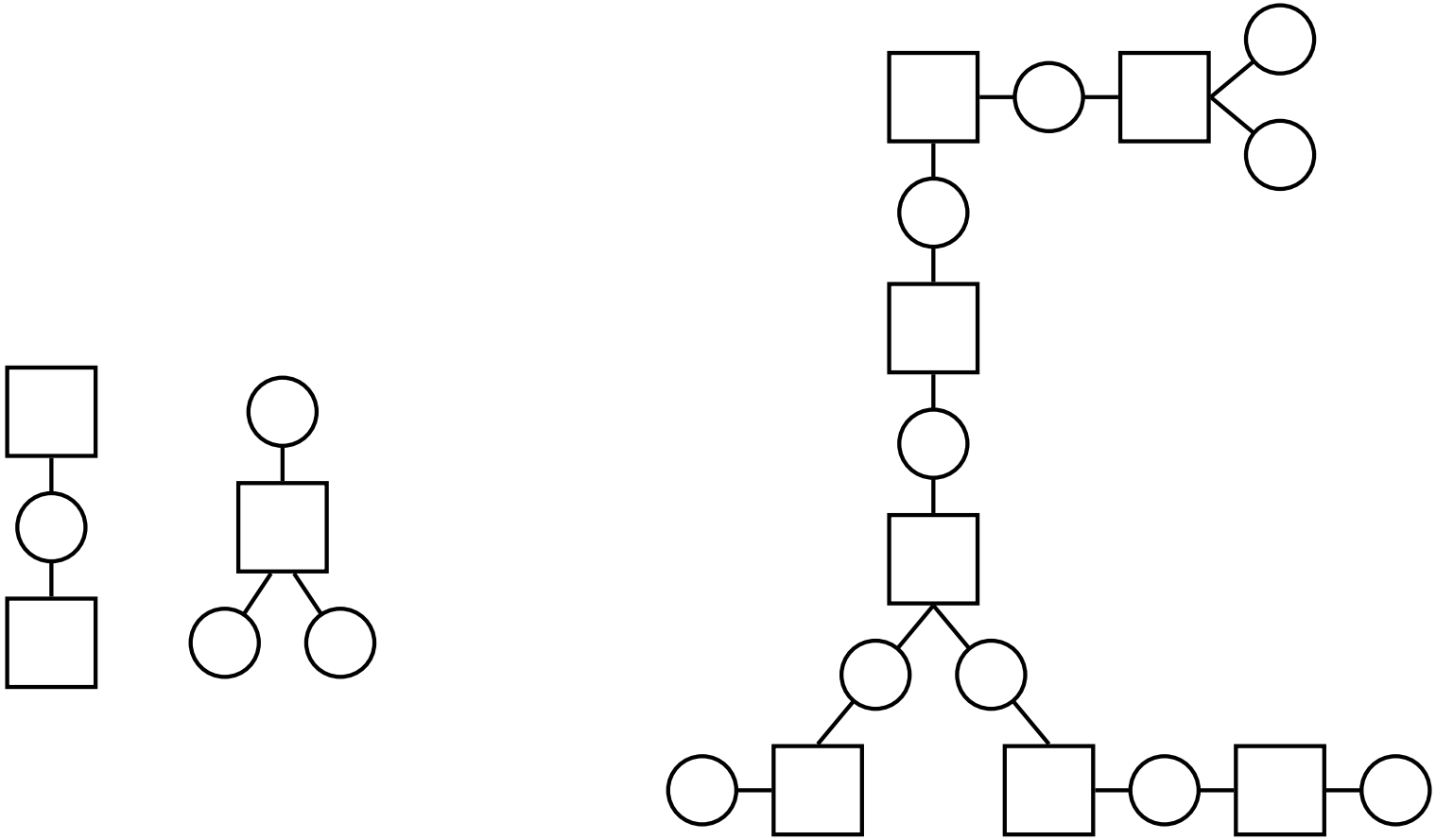
- quantitative models (differential equations)
- constraint-based models, petri-Nets, π-calculus
- qualitative models (graphs)

Graph models

$$
\begin{array}{lcl}
R_1 : A + B & \longleftrightarrow & C \\
R_2 : C & \longleftrightarrow & D \\
R_3 : D & \longleftrightarrow & E \\
R_4 : E & \longleftrightarrow & F + G \\
R_5 : F & \longleftrightarrow & J \\
R_6 : G & \longleftrightarrow & H \\
R_7 : H & \longleftrightarrow & I
\end{array}
$$

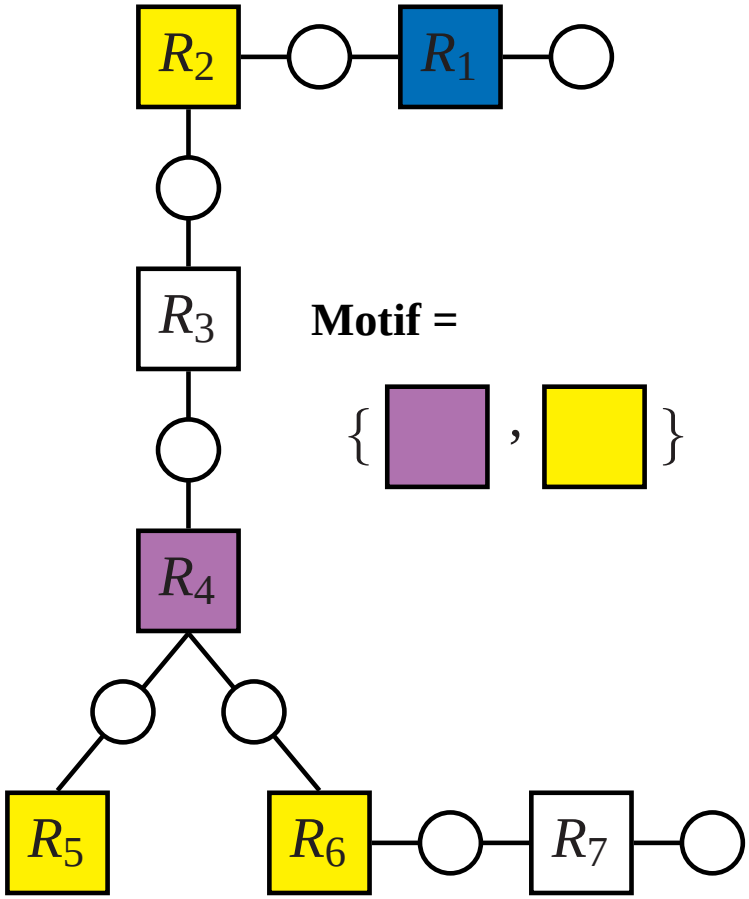


Motif models: Inadequacy of topological definition



- R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs : simple building blocks of complex networks. Science, 298(5594) :824-827, Oct 2002.

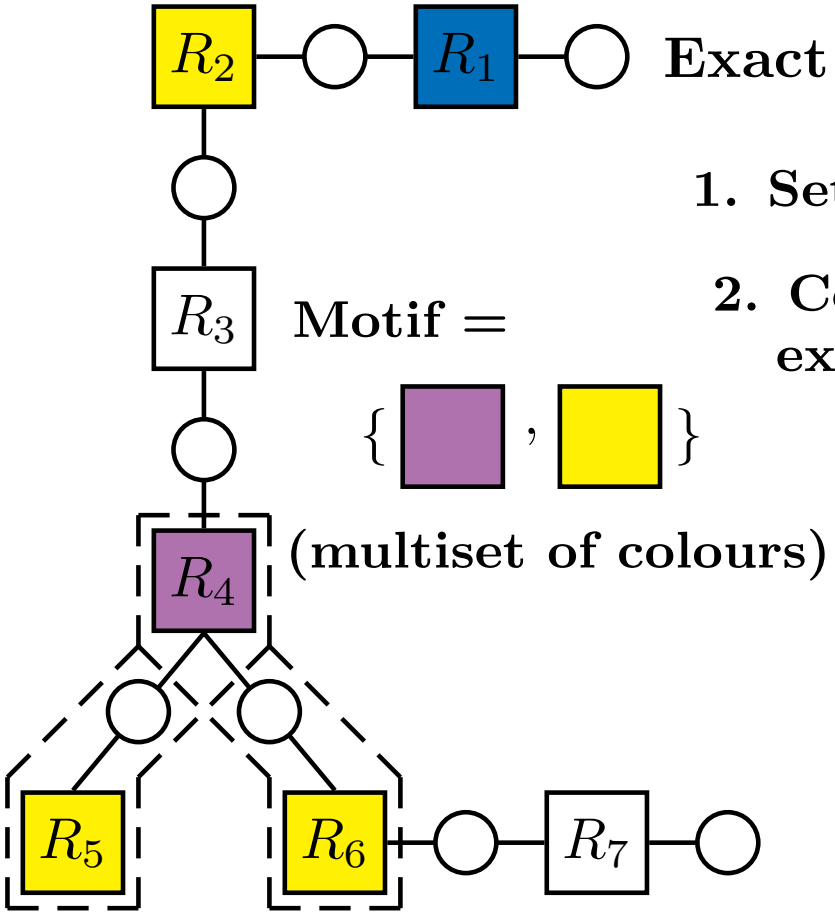
Motif models: A topology-free definition



Motif = multiset of colours

No constraint on the order nor on the topology.

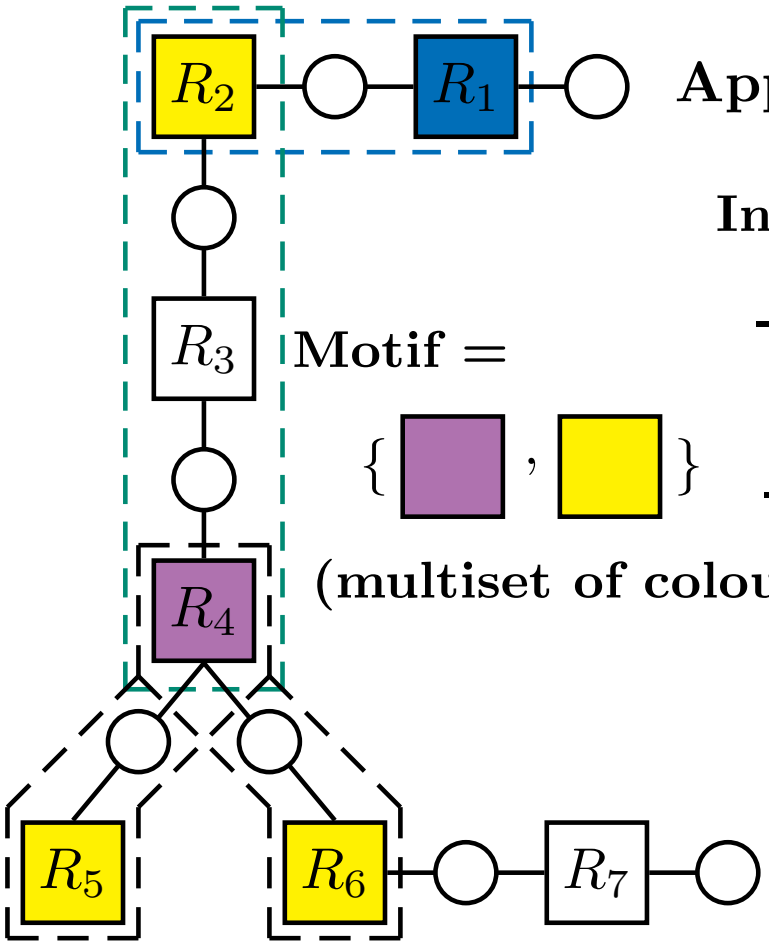
Definition of occurrence



Exact match

1. **Set of connected reactions**
2. **Colours of the reactions exactly match the motif**

Introduction of flexibility



Approximate match

Introduction of flexibility

- similarity of colours instead of identity

- possibility of insertions

Motif =

(multiset of colours)

Similarity between reactions

The "EC" classification:
Every enzyme is assigned a code with 4 numbers expressing the chemistry of the reaction it catalyses
Ex : 1.1.2.3

Similarity measure:
Two enzymes are considered similar if their codes are identical down to a given depth
Ex : 1.1.2.3 is similar to 1.1.2.1 (for threshold 3)

Search problem formulation

Search problem: given a motif and a threshold for comparison, find all occurrences of that motif in the metabolic graph

Hardness Results

INPUT : Vertex-labelled graph G and a multiset of colours M
QUESTION : Does G contain a connected subset of vertices with a bijection between its colours and M ?

TYPE OF GRAPH	PATH	TREE	GRAPH
COMPLEXITY	polynomial	NP-complete, FPT in k	NP-complete

Exact Algorithm

Graphs considered are sparse ($|V| \sim 3000$, $|E| \sim 15000$), therefore an exact algorithm can run in acceptable time.
(in practice, 8 μs for motifs of size 3 on AMD 64, 1.8 GHz, 2 Go)

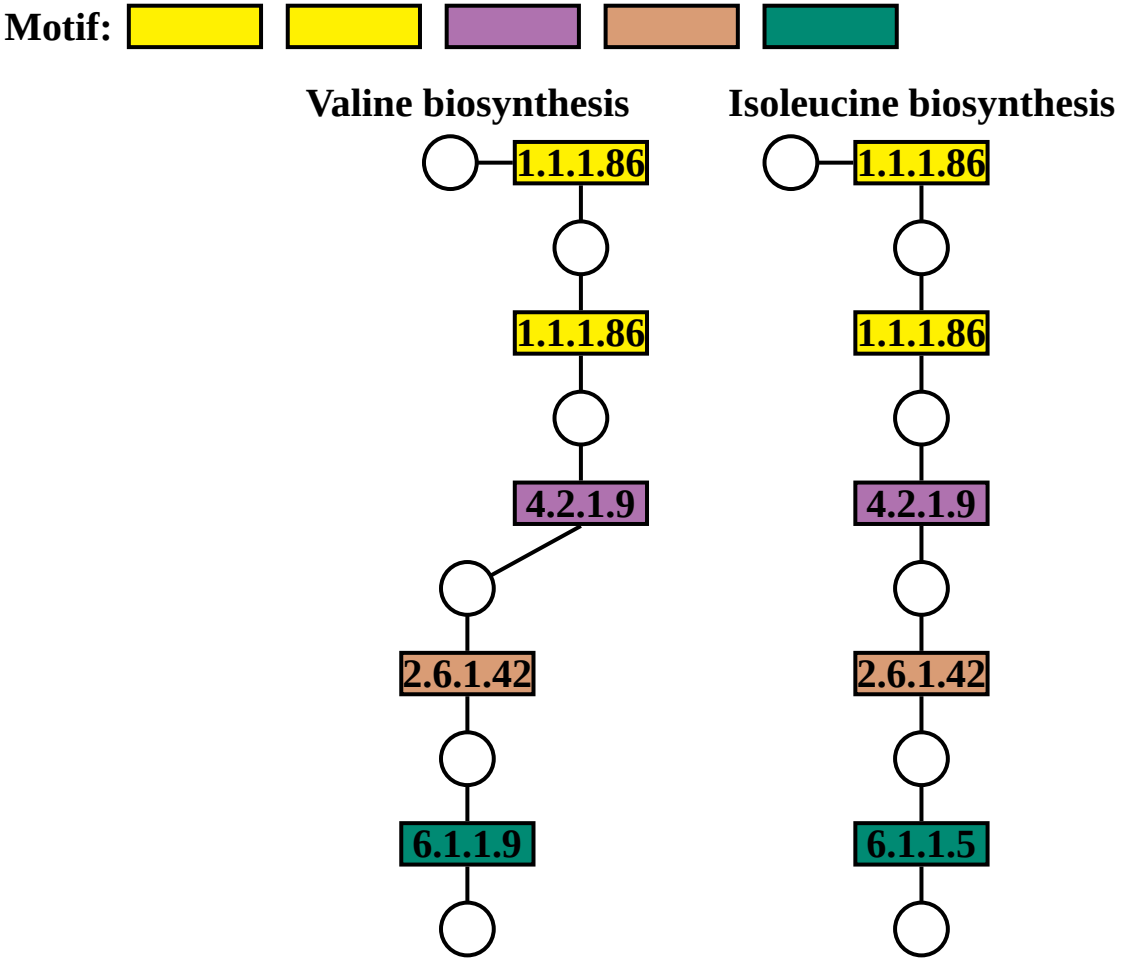
Exact Algorithm

Main ideas:

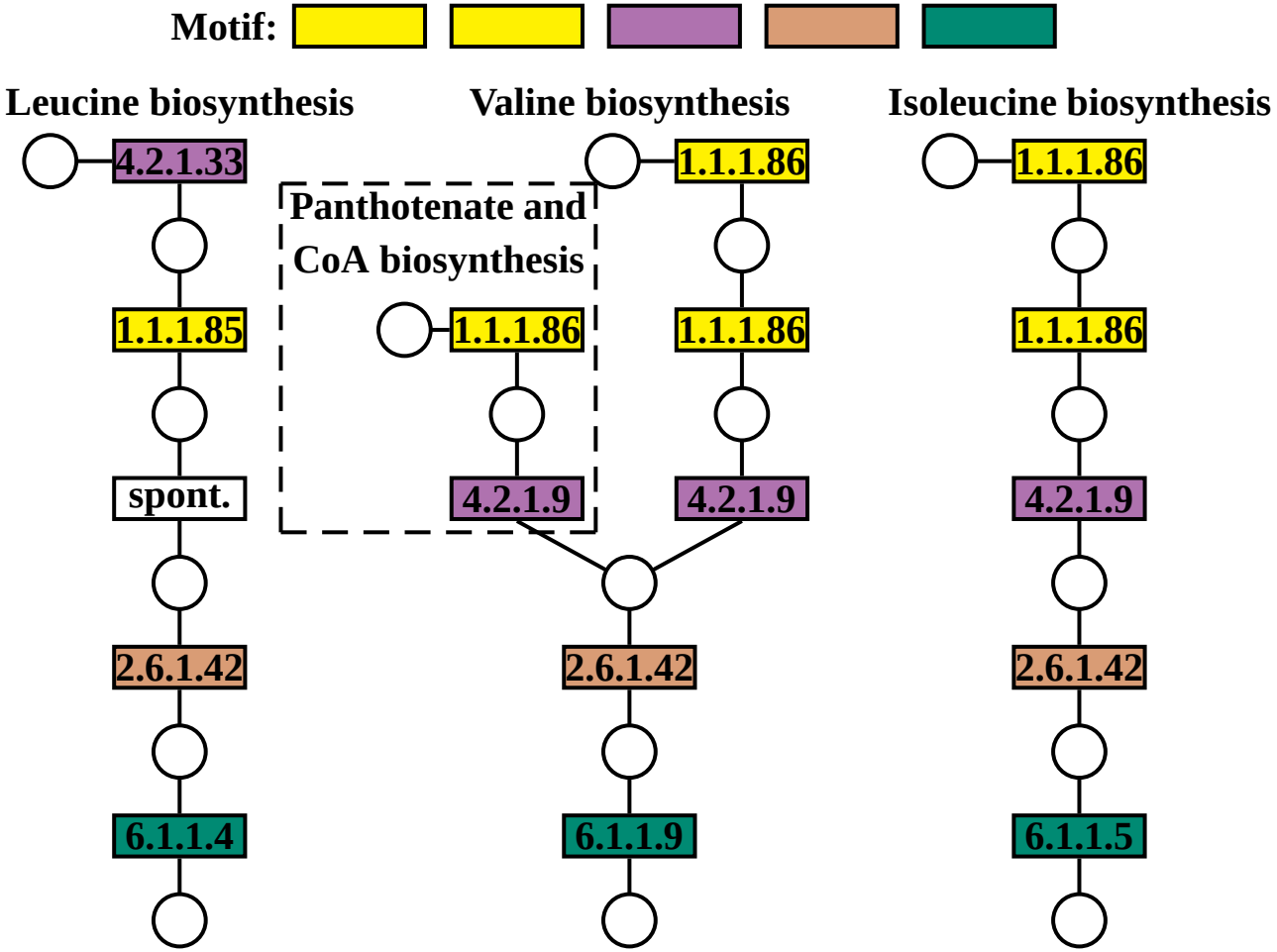
1. **Filter**: Only nodes with colours from the motif are kept.
2. **Candidate generation**:

   For each node:
   - Enumerate all sets of k connected nodes containing it (using breadth-first search (bfs) and backtrack) and test the colour condition.
   - Eliminate the node.
3. **Speed-ups**:
   - **Colour Pruning**: During candidate generation, if a set of nodes does not satisfy the colouring condition, then all sets containing this subset will not be tested.
   - **Seed Choice**: The bfs only starts from vertices with less frequent colours

Initial application to pathway evolution



Application to pathway evolution



Conclusion - so far

- **Modelling**:
  A 'coloured motif' is a multiset of colours (reaction types)
- **Algorithmics**:
  Searching for all occurrences of such motifs is NP-complete but we implemented an exact algorithm which appears to be fast in practice
- **Application**:
  Occurrences of a motif can be given a biological interpretation in some cases (evolution of metabolic pathways, alternative pathways)

Lacroix V, Fernandes CG, Sagot M-F Reaction motifs in metabolic networks. Proceedings of WABI '05, Springer-Verlag, Lecture Notes in Computer Science, 2005, vol. 3692, pp. 178-191.

Lacroix V, Fernandes CG, Sagot M-F, Motif search in graphs: application to metabolic networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2006, vol. 3, pp. 360-368.

Inferring motifs

Question:
What happens if you do not know which motif to look for ?

Answer:
You can consider the inference problem: given a coloured graph, find all repeated motifs.

Inference problem formulation

Inference problem: given a metabolic graph, a number k and a threshold σ, find all repeated motifs of size k with threshold σ.

Inference algorithm

- **Algorithm**: The current implementation of the inference algorithm is merely a series of search of all possible motifs of a given size and threshold.
- **Speed-up**: If the motif $M = \{1.1, 2.3, 1.4\}$ has no occurrence then the motif $M' = \{1.1.1, 2.3.2, 1.4.2\}$ will have no occurrence either. Therefore the list of motifs to test can be pruned.

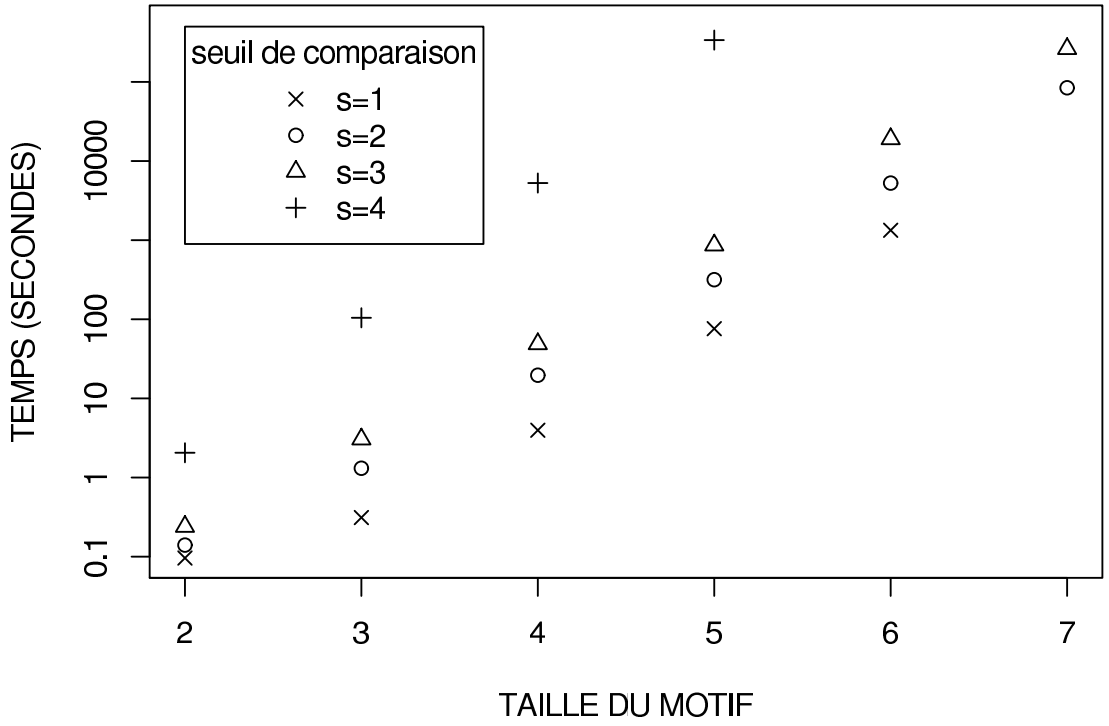
Dataset

Dataset: Small molecule metabolism of *Escherichia coli*

- **Source:** http://ecocyc.org/
- **Pre-treatment:**
  For each reaction, remove side compounds

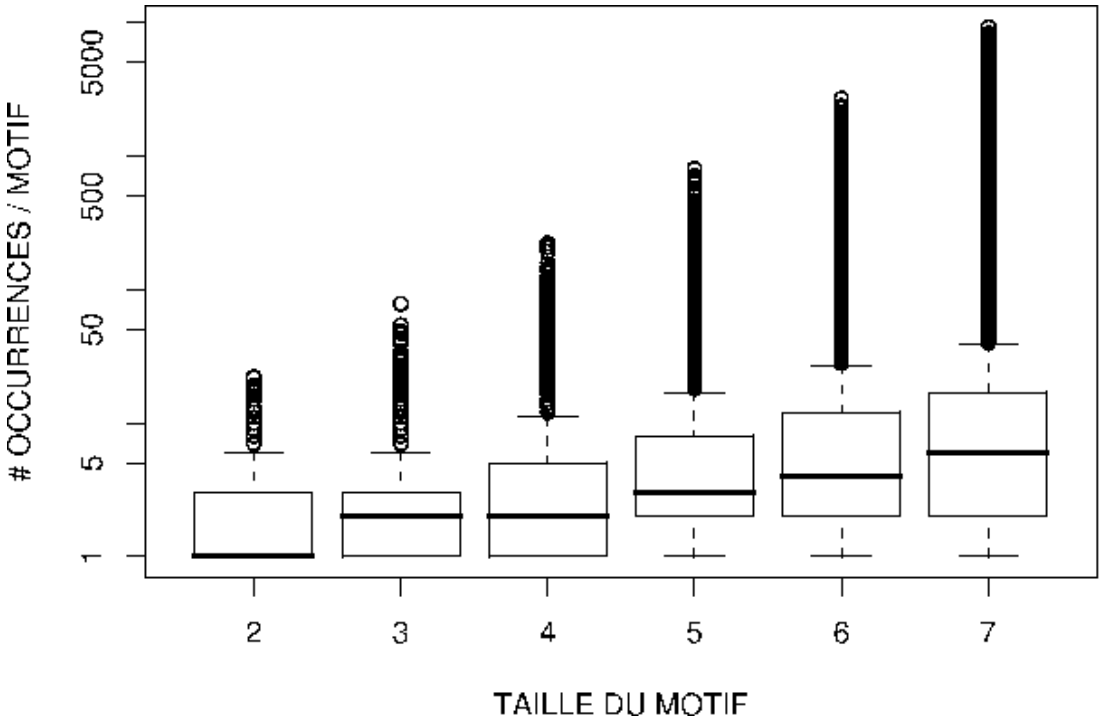
- **Characteristics**
  - number of reactions: 587
  - number of compounds: 553
  - number of EC numbers: 463
    - 428 ($\sigma = 4$), 91 ($\sigma = 3$), 40 ($\sigma = 2$), 6 ($\sigma = 1$)

Inference - time results



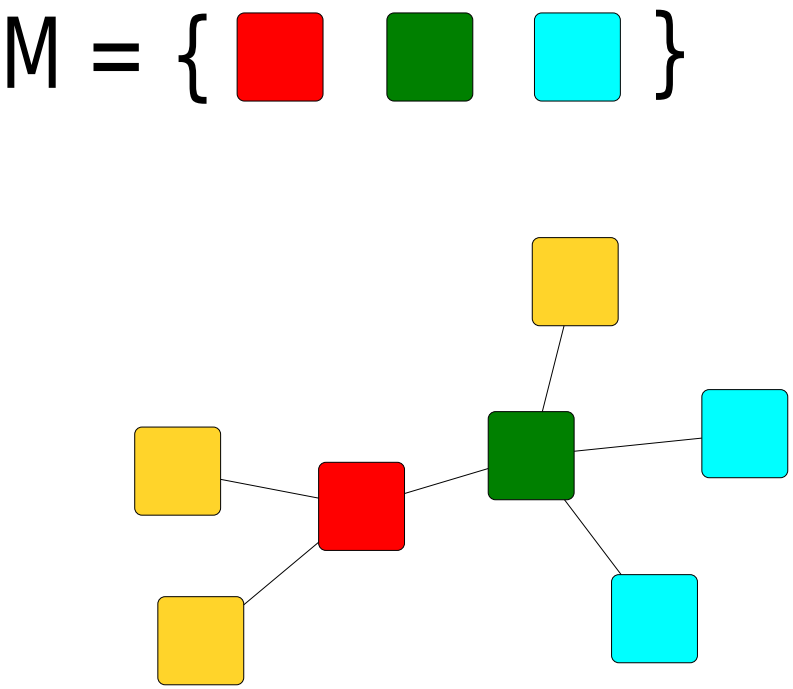
- Time grows almost exponentially with motif size
- Motifs of size 6 can be inferred in 3 hours

Inference - number of occurrences



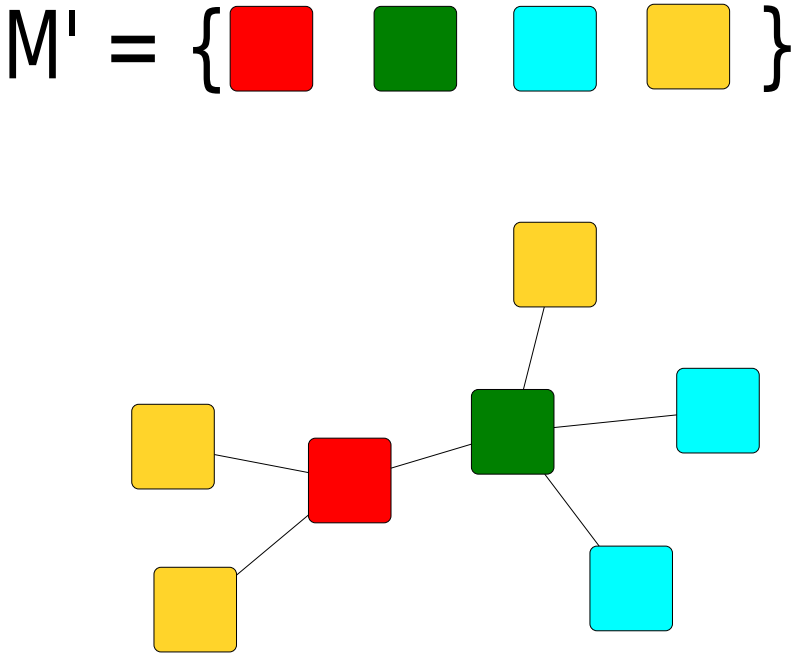
- Many motifs with few occurrences and some with a great number of occurrences
- The number of occurrences per motif tends to grow with motif size (counter-intuitive)

Larger motifs may have more occurrences



- Motif M has 2 occurrences

How can longer motifs have more occurrences ?



- Motif M' has 6 occurrences

Inference

Summary:

- Execution time is not our main limitation: motifs of size 7 can be inferred within hours
- Output size may become a problem for later interpretation: a motif of size 7 may have up to 10000 occurrences

Are all occurrences equally relevant ?

- Filter and/or group occurrences which share common features

Are all motifs equally relevant ?

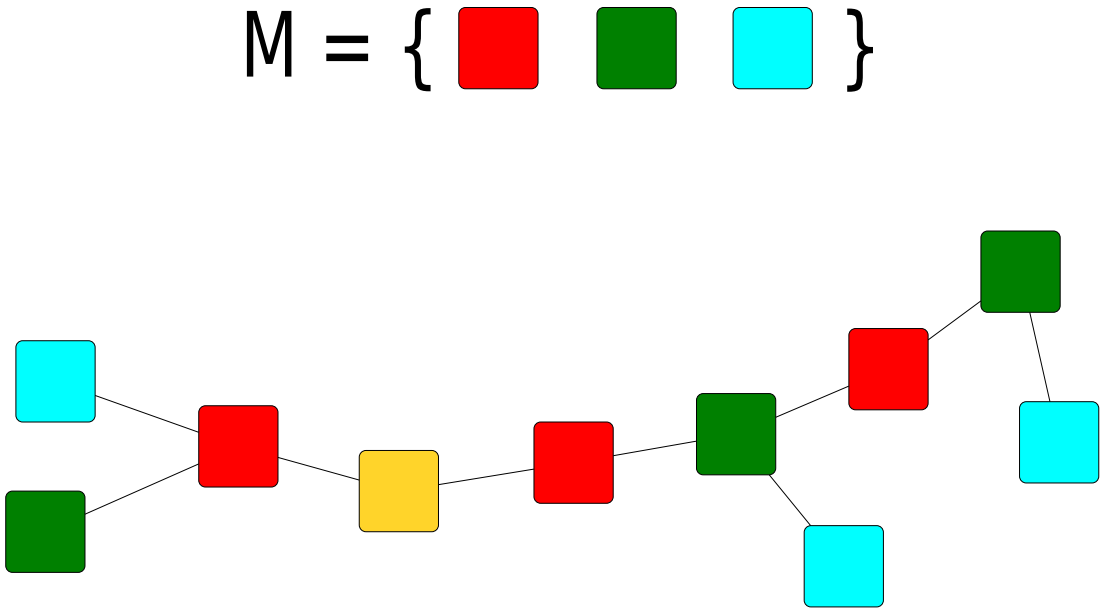
- Propose a statistical criterion for over-representation

Are all occurrences equally relevant ?

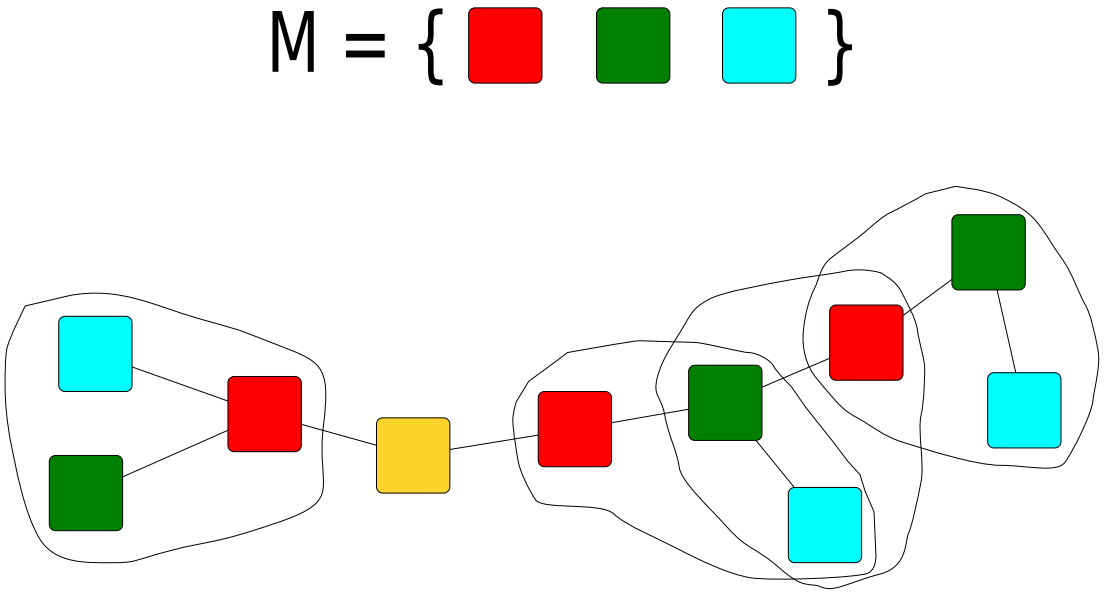
Two ways of grouping occurrences that we used:

- group occurrences which overlap (*i.e.* share a node)
- group occurrences which share the same topology

Group by overlap



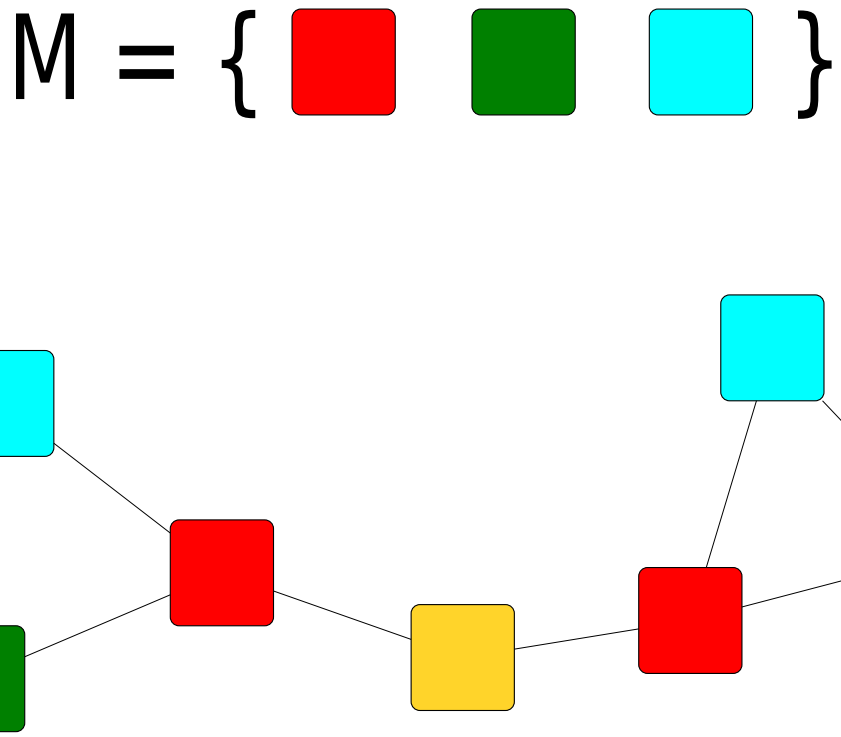
Group by overlap



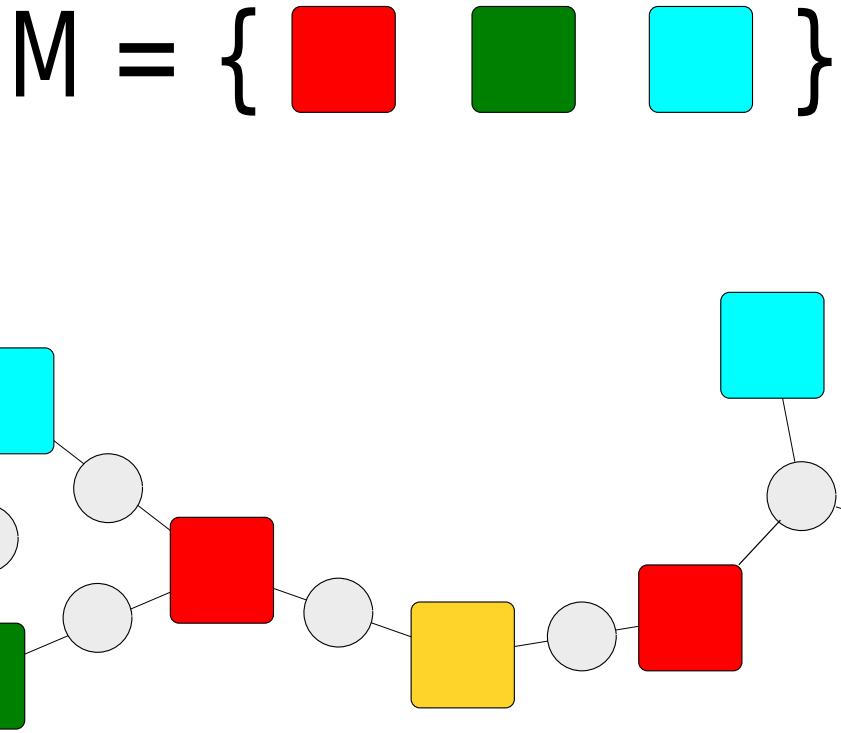
- 4 occurrences
- 2 clumps of occurrences

Overlapping occurrences may not be given the same biological interpretation as disjoint occurrences.

Group by topology



Group by topology



Need to use a bipartite graph model to discriminate more precisely

Are all motifs equally relevant ?

- Highly represented motifs are not necessarily over-represented motifs.
- An **over-represented** motif is a motif which occurs more than expected by chance.
- Need to define a null model: a random graph model

Over-represented motifs

Which random graph model should we choose ? ... an open problem

- Erdös-Rényi: all nodes are connected with the same probability p. (not realistic in biology)
- Erdös-Rényi Mixture for Graphs (ERMG): nodes belong to groups. The probability for two nodes to be connected depends on the groups.
- **Fixed topology**: the topology of the real graph is fixed but the colours are shuffled.

Daudin JJ, **Lacroix V**, Mariadassou M, Miele V, Picard F, Robin S, Sagot M-F, Uncovering structure in biological networks. RIAMS'06 , 2006.

Over-represented motifs

Once a random graph model is chosen, two approaches can be adopted:

- Exact formulae: obtain a formula for the mean and variance of the motif count in a random graph model and derive a Z-score to assess motif over-representation. (on-going work for the Erdös-Rényi model)
- **Simulations**: generate random (or randomized) graphs and count the motif in each one of them. The real count can then be compared to this obtained count distribution.

Applications

Question: what can we learn using our definition of motif ?

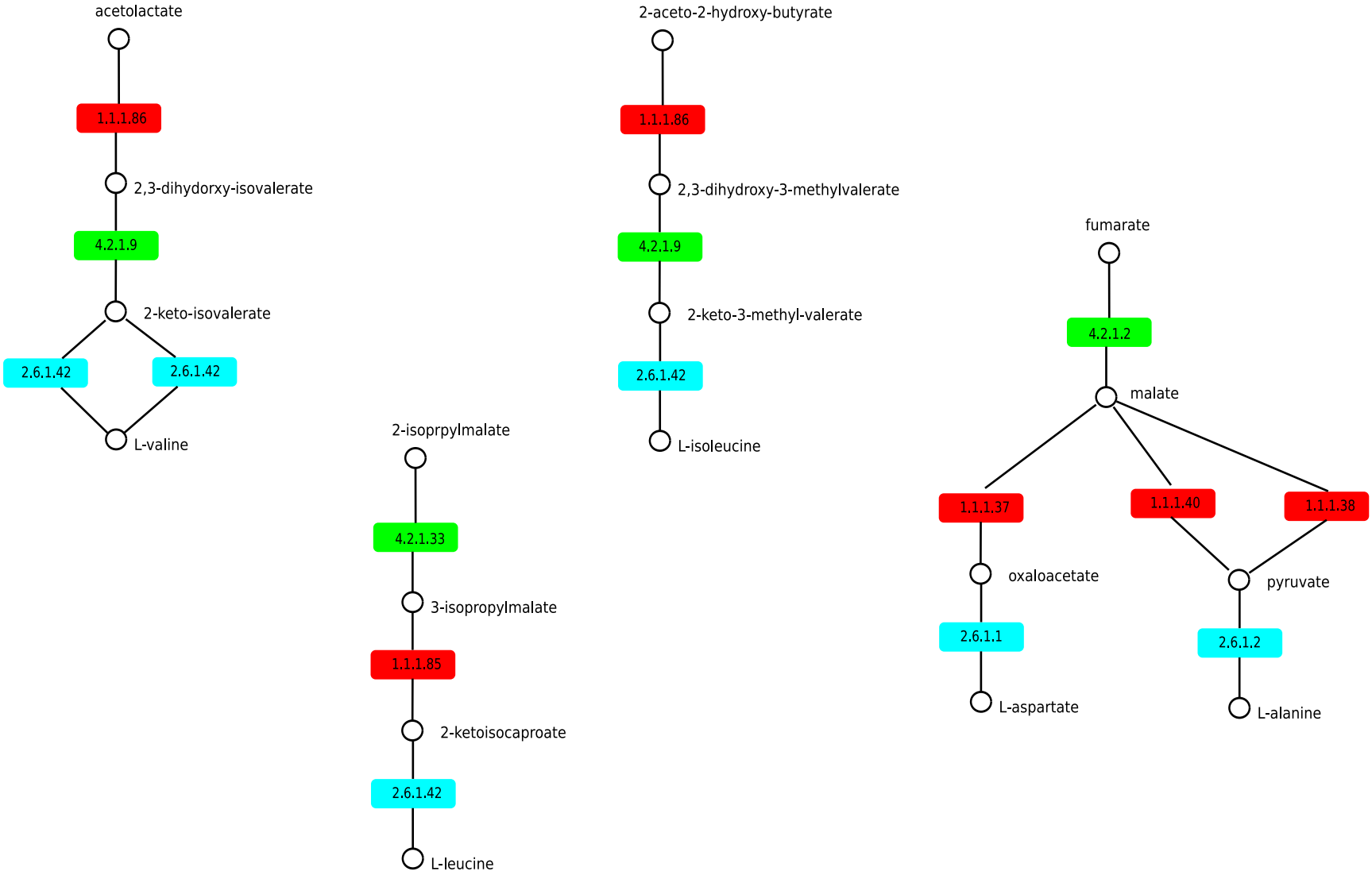
- Examine more deeply some examples
- Relate motifs to known functional structures

Examples: maximal motifs

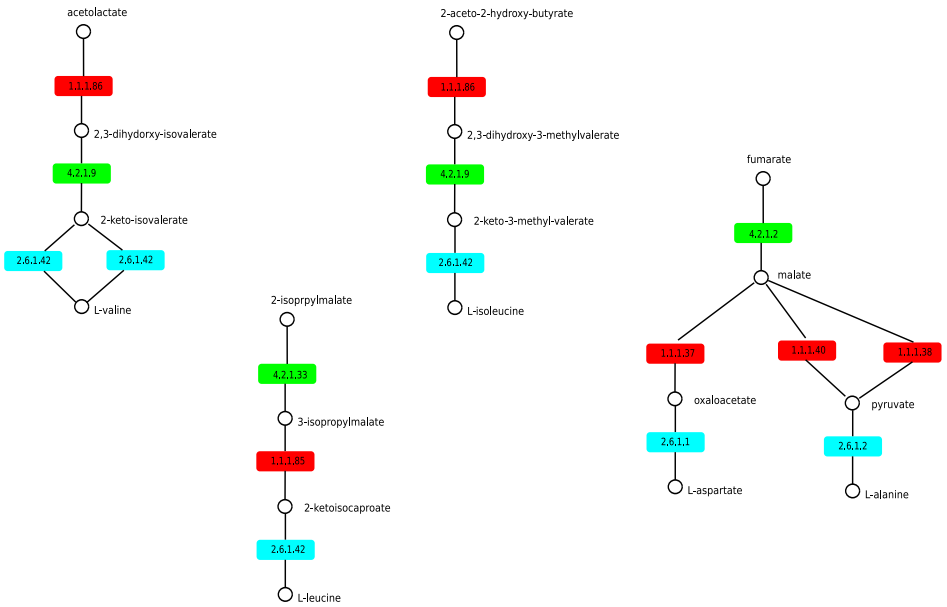
Examples have been chosen using the following rules:

1. maximum number of clumps
2. metabolic pathway of interest
3. randomly

Example for n=3, s=3

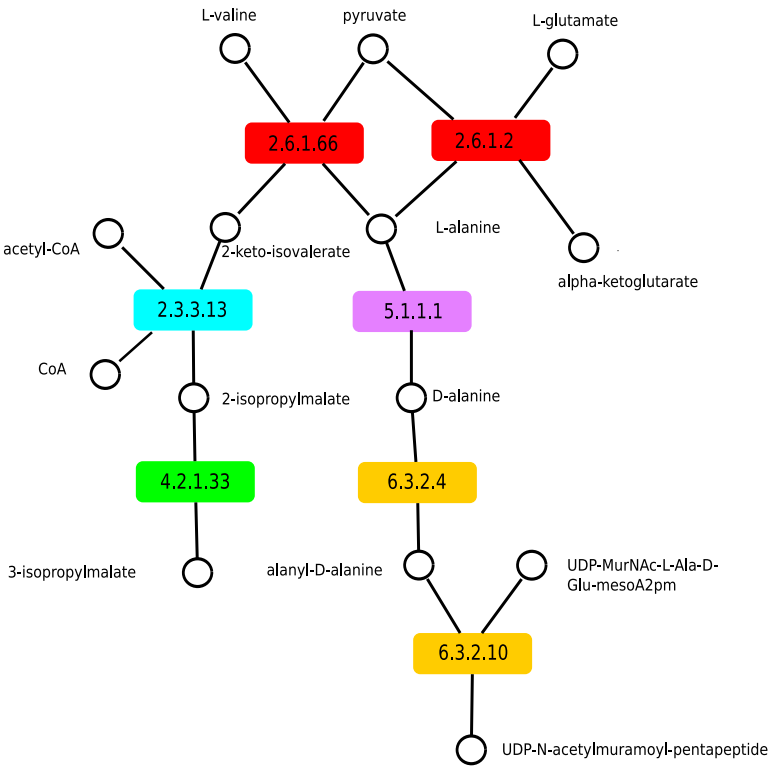


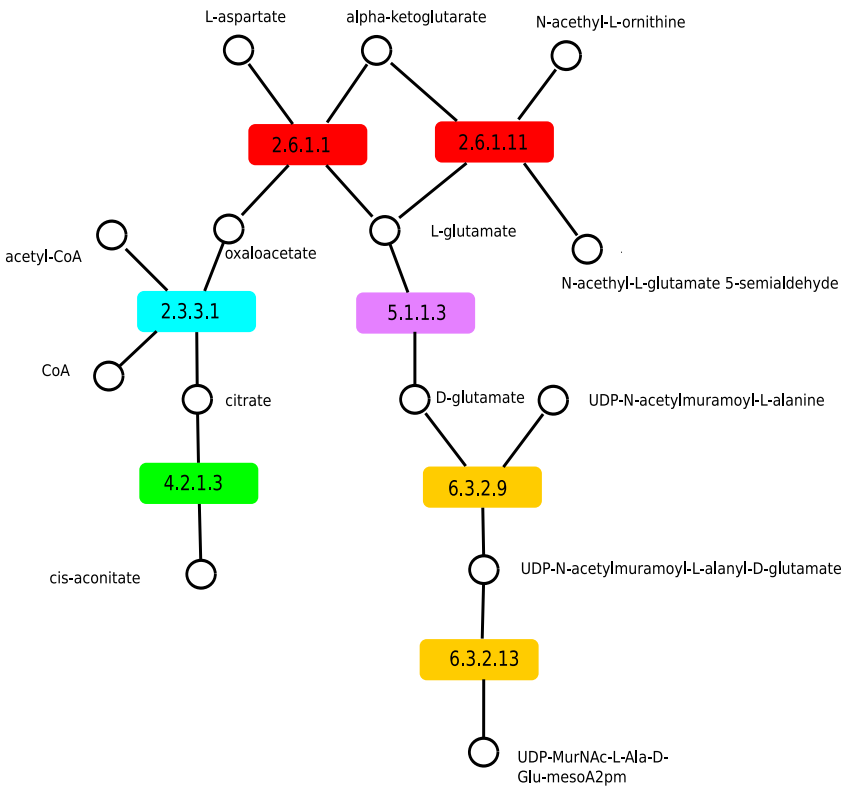
Example for n=3, s=3



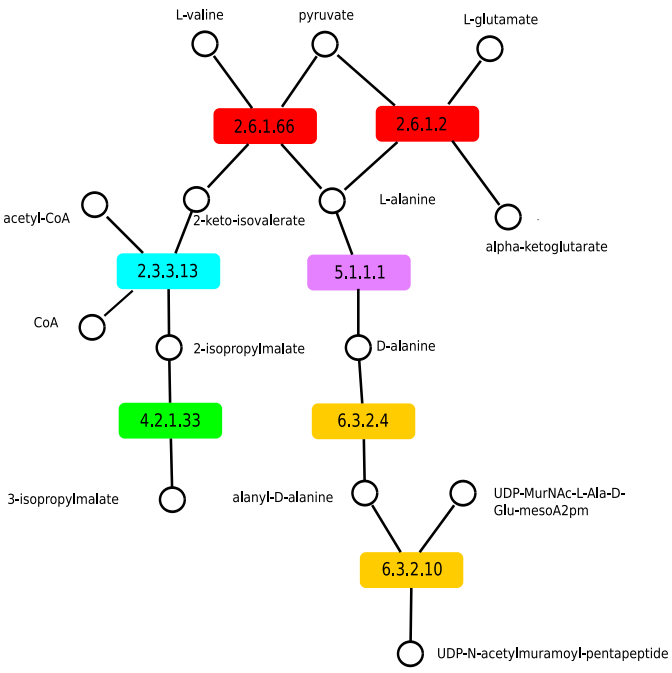
- 7 occurrences, 4 clumps
- common to 4 amino-acid biosyntheses
- the last clump is made of inter-pathway occurrences

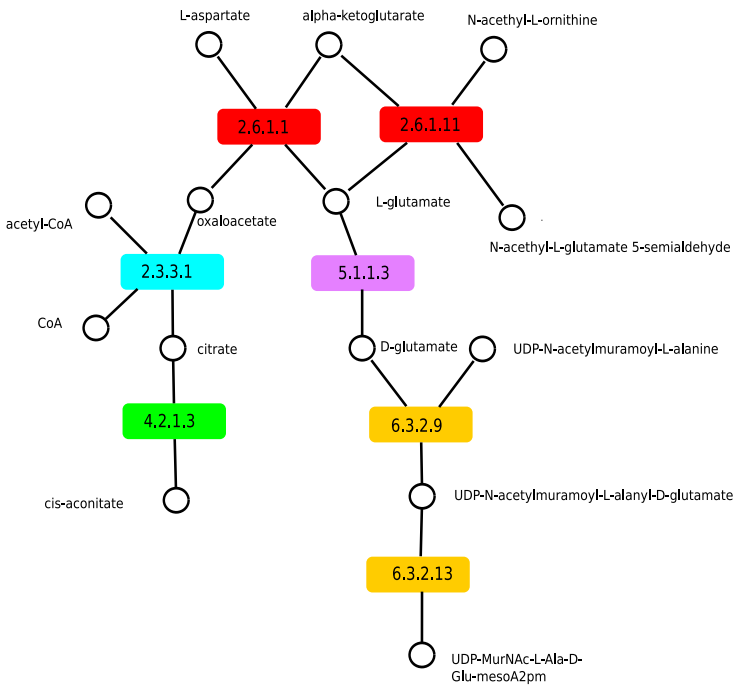
Example for n=7, s=3





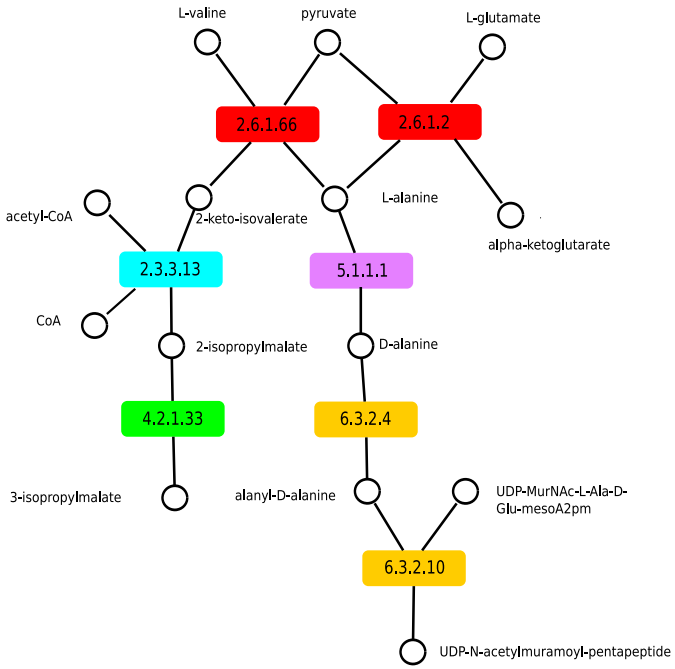
Example for n=7, s=3

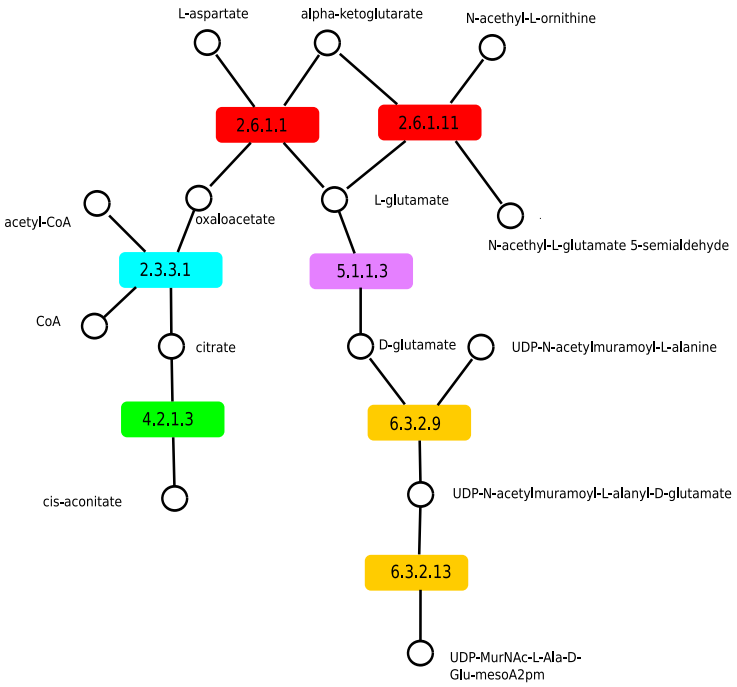




- 2 occurrences, 2 clumps
- key role of the transaminase connecting leucine biosynthesis, krebbs cycle and peptidoglycan biosynthesis

Example for n=7, s=3

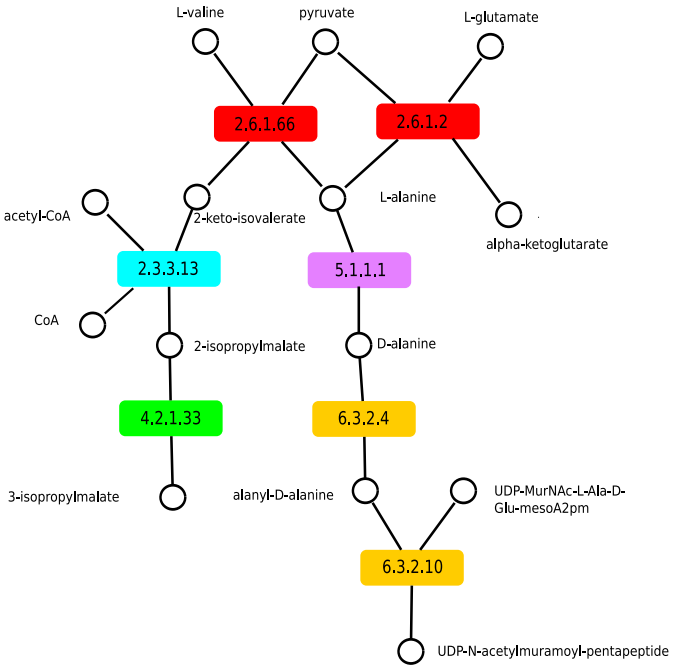


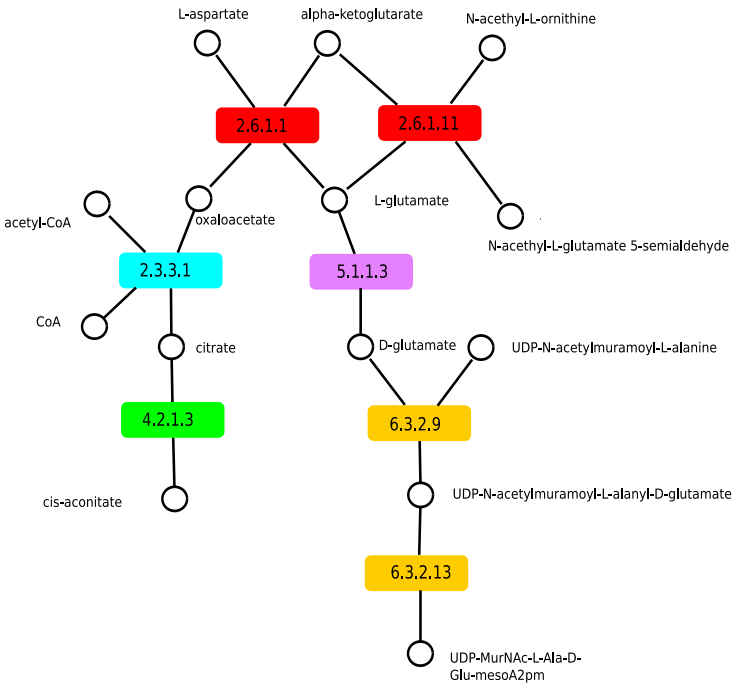


Paralogs:

- acnA (4.2.1.3) and leuC (4.2.1.33)
- leuB (1.1.1.85) and icd (1.1.1.42)
- murD (6.3.2.9), murE (6.3.2.13) and murF (6.3.2.10)

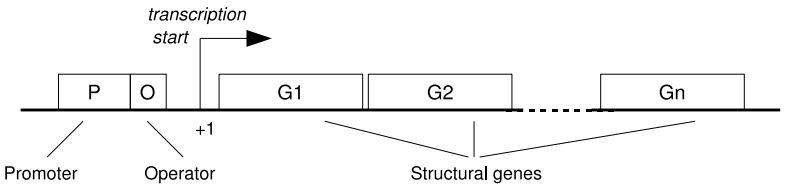
Example for n=7, s=3





Operons:

- murD, murE and murF are part of the same operon
- murF and ddlA are part of the same operon in other organisms



Relate motifs to known functional structures

Question: Are the genes involved in repeated motifs more clustered on the genome ?

Related works

Rison, S.C., Teichmann, S.A. and Thornton, J.M. (2002) Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli*. *J. Mol. Biol.* **318**, 911-932.

- There is a positive correlation between pathway distance and chromosomal distance
- This correlation is not verified for long distances
- Short distance correlation is mainly explained by operon structures

Protocol

1. We retrieved a set of known operons in *E. coli* from RegulonDB.
2. We identified the occurrences that were covered by an operon.
   - An occurrence is covered by an operon if all its reactions are covered.
   - A reaction is covered by an operon if the gene(s) coding for one of its enzymes is (are) in this operon.

Results

Parameters: size=2, threshold=3

- 249 motifs
- 1379 occurrences

Counts:

	operon +	operon -	
several clumps	77	612	689
only one clump	45	645	690
	122	1257	1379

Motifs repeated in several clumps

Question: Are occurrences of motifs repeated in several clumps more covered by operons ?

Answer: Yes. (permutation test, p=0.003)

Quantification:
63.1% of occurrences covered by operons are occurrences of repeated motifs.

Conclusion

- Known result: neighbours in the network tend to be neighbours on the genome (operon structure)
- New result: This tendency is reinforced when reactions belong to repeated motifs (several clumps)

Software

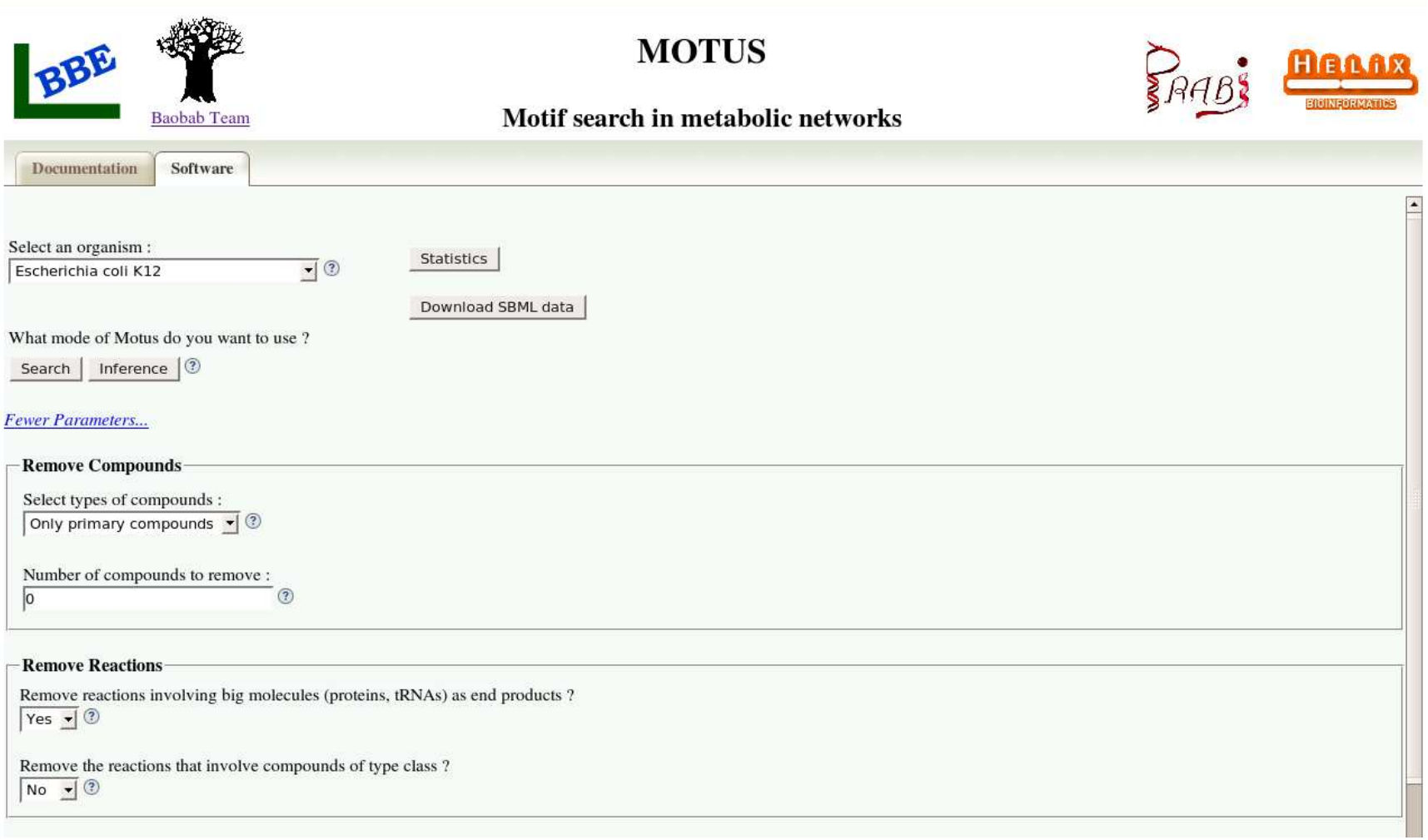
MOTUS:

http://pbil.univ-lyon1.fr/software/motus

Participants:

- Data: Ludovic Cottret (Baobab)
- Web: Odile Rogier (PRABI)
- Drawing: Fabien Jourdan (INRA toulouse)

Dataset selection



Motif search

Baobab Team

MOTUS

Motif search in metabolic networks

Documentation | Software | Search Mode | Search Results

Parameters	
Selected Organism	Escherichia coli K12
Type of Compounds	Only primary compounds
Coumpounds to remove	0
Remove reactions involving big molecules as end products ?	Yes
Remove compounds of type "class" ?	No
Number of simulations	1000
Motif	1.1.1 4.2.1 2.3.3

Results	
Occurrences Number	7
p-Value	0.038

No Occurrence	Occurrence			Pathway
	Reaction1	Reaction2	Reaction3	
1	2-ISOPROPYLMALATESYN-RXN [2.3.3.13]	3-ISOPROPYLMALISOM-RXN [4.2.1.33]	3-ISOPROPYLMALDEHYDROG-RXN [1.1.1.85]	superpathway of leucine, valine, and isoleucine biosynthesis leucine biosynthesis
2	2-ISOPROPYLMALATESYN-RXN [2.3.3.13]	DIHYDROXYISOVALDEHYDRAT-RXN [4.2.1.9]	ACETOLACTREDUCTOISOM-RXN [1.1.1.86]	superpathway of leucine, valine, and isoleucine biosynthesis
3	ACONITATEDEHYDR-RXN [4.2.1.3]	CITSYN-RXN [2.3.3.1]	MALATE-DEH-RXN [1.1.1.37]	respiration (anaerobic) superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass glyoxylate cycle superpathway of glyoxylate bypass and TCA TCA cycle

Motif inference

Baobab Team

MOTUS

Motif search in metabolic networks

Documentation | Software | Inference Mode | Inference Results

Parameters	
Selected Organism	Escherichia coli K12
Type of Compounds	Only primary compounds
Coumpounds to remove	0
Remove reactions involving big molecules as end products ?	Yes
Remove compounds of type "class" ?	No
Number of simulations	1000
Size of motif	2
Threshold	2

The inference results can be visualized by MotusViewer.

No Motif	Motif		Number of Occurrences (Occ)	p-Value (Occ)	Connected Components (CC)	p-Value (CC)	Number of Pathways per Occurrence		Number of Occurrences which are included in a single Pathway
	EC1	EC2					Mean	Variance	
1	1.2	2.2	10 More details...	< 0.0005	1	0.3975	6.1	2.88	0
2	1.2	4.1	23 More details...	< 0.0005	2	0.894	4.91	2.54	8
3	1.4	2.6	7 More details...	< 0.0005	1	0.3095	4.71	1.38	1
4	1.4	6.3	8 More details...	< 0.0005	2	0.1545	4.37	1.27	1
5	1.5	1.5	4 More details...	< 0.0005	2	0.001	1.5	0.25	2
6	1.5	2.1	9 More details...	< 0.0005	1	0.333	4.77	4.69	9

Motus Viewer

Conclusion

Motif Search

- Searching for a coloured motif in a coloured graph is NP-complete
- Metabolic networks are not so dense, which enables to run exact algorithms
- Coloured motifs may help in formulating hypotheses regarding pathway evolution

Conclusion

Motif Inference

- Time is not a limitation but the number of occurrences may become one
- Occurrences may be grouped in different ways
- Over-representation may enable to select relevant motifs
- MOTUS: available software
- Some examples have been studied in detail and provide insight
- Motifs repeated in several clumps are enriched in operons

Lacroix V, Cottret L, Rogier O, Fernandes CG, Jourdan F, Sagot M-F
MOTUS: a tool to detect coloured motifs in metabolic networks. in prep.

Perspectives

More modelling:

- Explore alternative ways of comparing reactions... towards RC numbers ?
- Number of occurrences, number of clumps ... maximum number of pairwise disjoint occurrences ?

More algorithms:

- Inference algorithm
- Largest repeated motif

More statistics:

- Assess expected motif count in available random graph models (without using simulations)
- Open problem: what is a relevant random graph model ?

Perspectives - continued

More biology:

- Explore the link between genomic position and motifs
- Explore the link between paralogy and motifs
  - Are motifs repeated in several clumps enriched in duplicated genes ?
- Relate motifs to models of pathway evolution
- Compare motifs in different organisms
- Apply the concept of coloured motif to protein interaction networks

Collaborations

Graph Statistics

Jean-Jacques Daudin
Stéphane Robin
Franck Picard
Sophie Schbath

Graph Drawing

Fabien Jourdan

Romain Bourqui, David Auber

Collaborations

Cristina
Gomes Fernandes

Leen Stougie

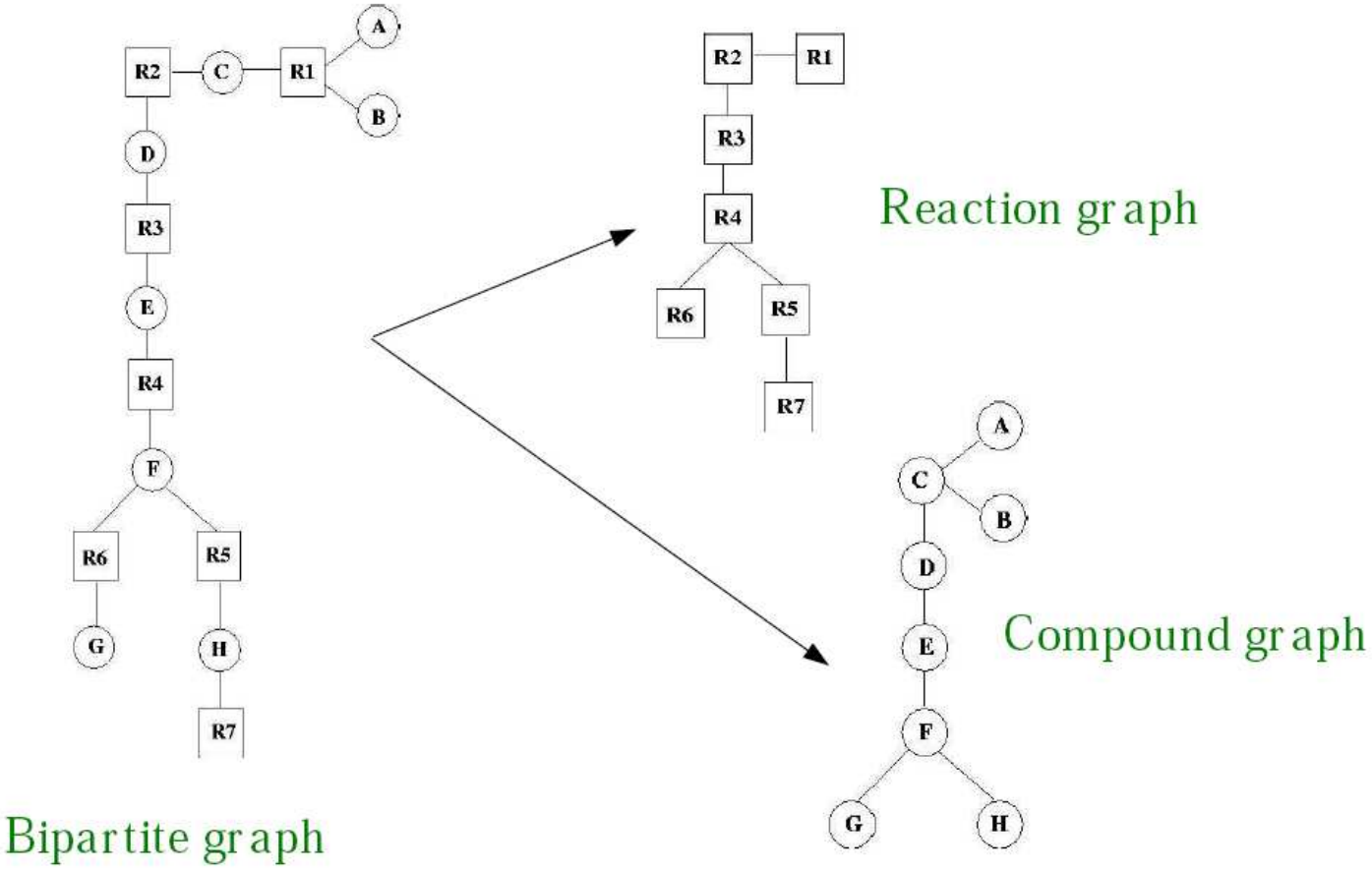
Motif Search

Elementary modes

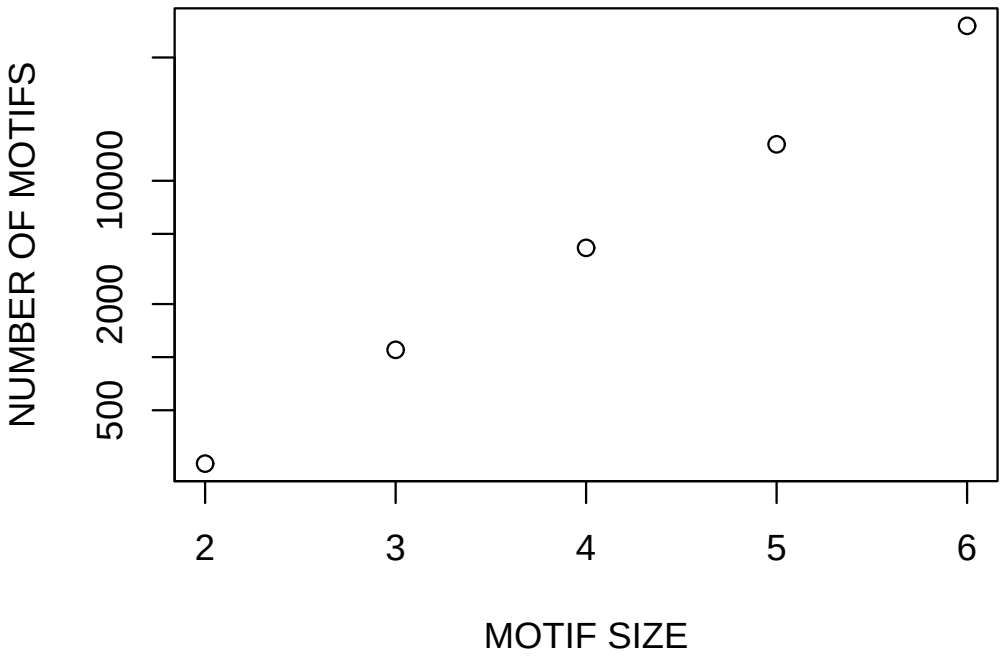
Alberto
Marchetti-
Spaccamela

Thank you !

Graph models



Inference - number of motifs

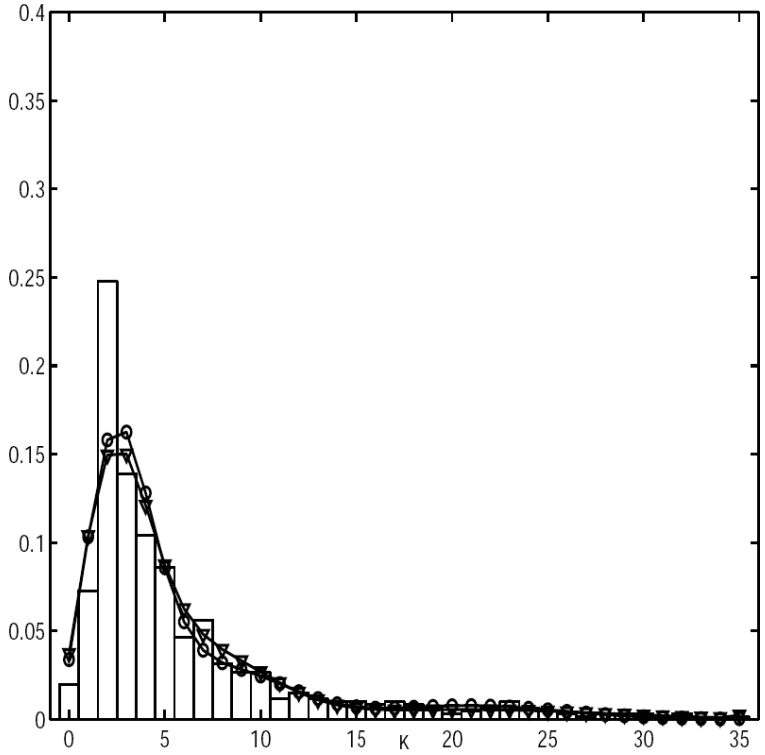


- The number of motifs grows exponentially with motif size

ERMG model

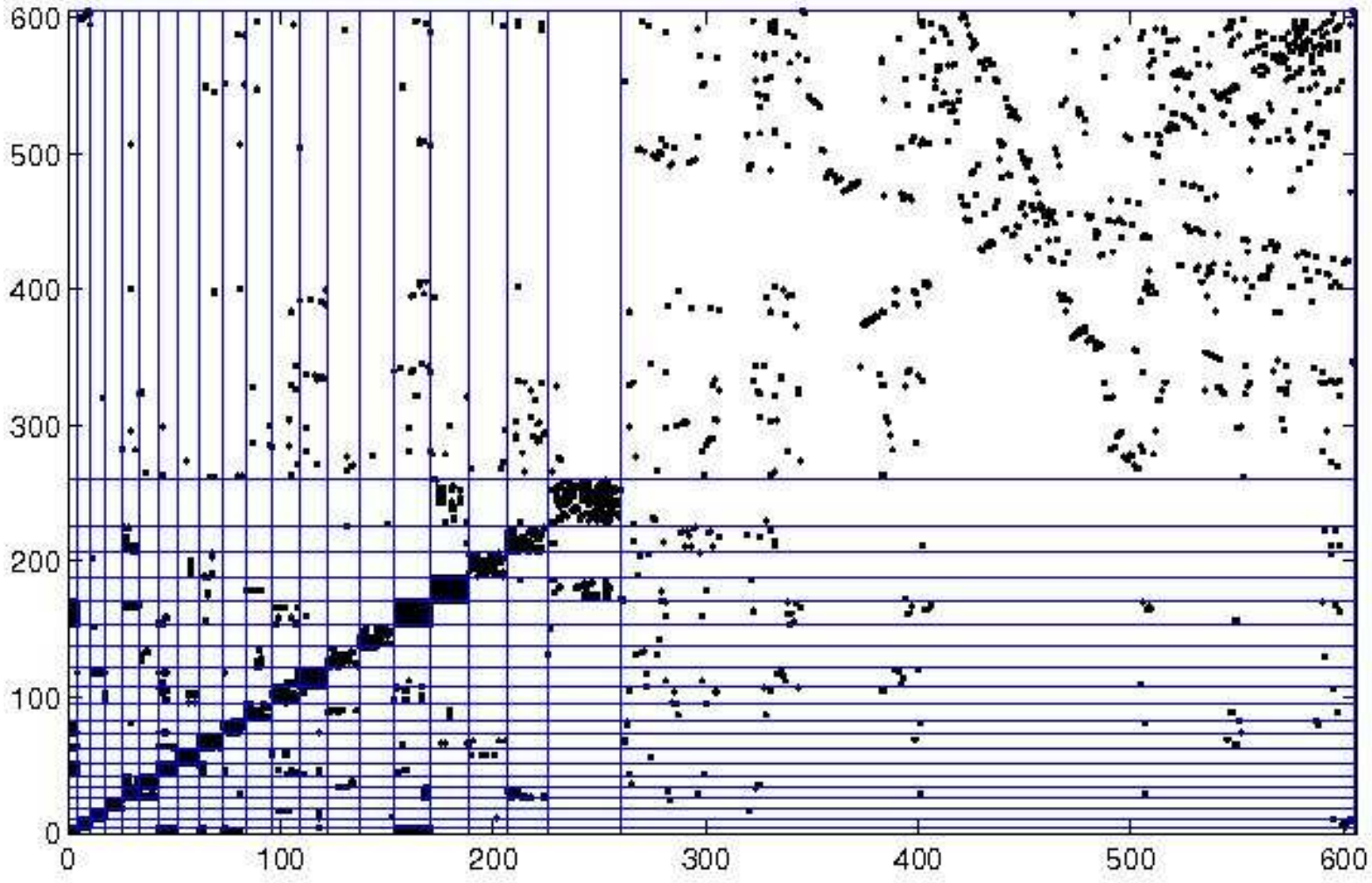
- Erdös-Rényi graph do not model well the degree distribution of real networks
  - ER model: $K_i \sim P(\lambda)$
  - observed: $K_i \sim k^{-\gamma}$
- ERMG is a generalisation of ER
- Hypothesis: there exists a hidden structure into Q classes of connectivity

ERMG model

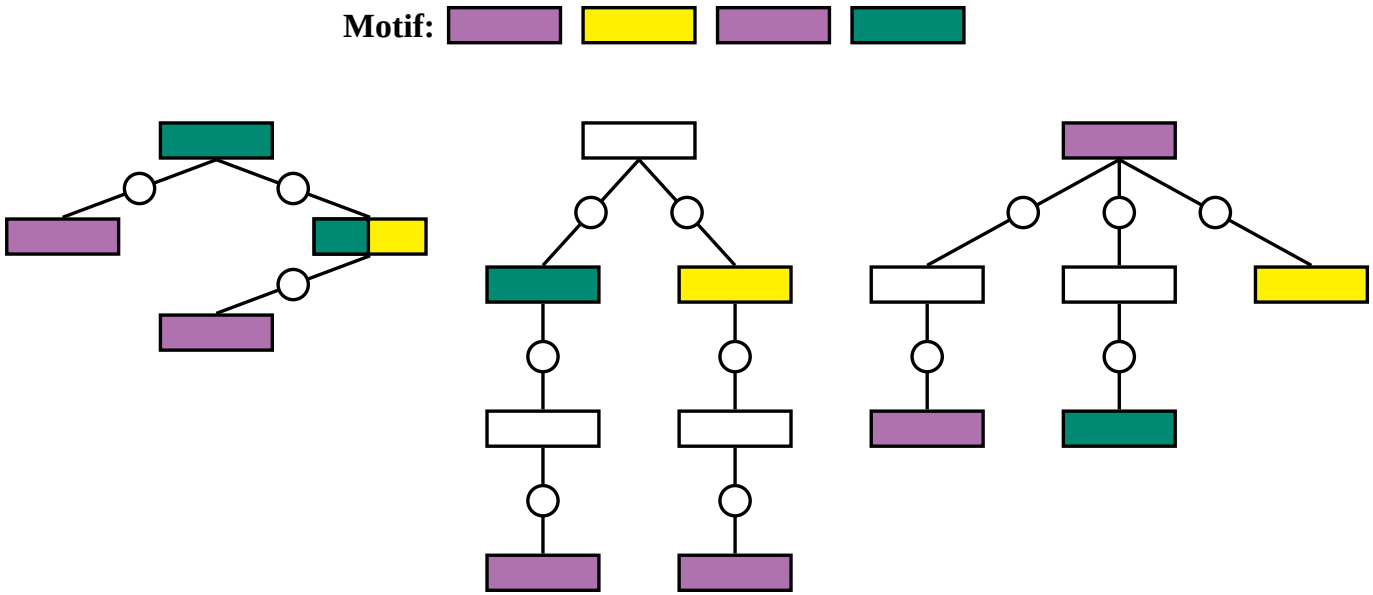


- The degree distribution is modelled correctly.

ERMG model



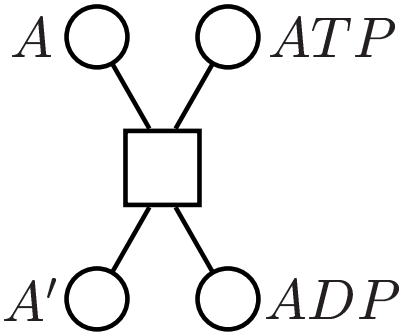
Taking gaps into account



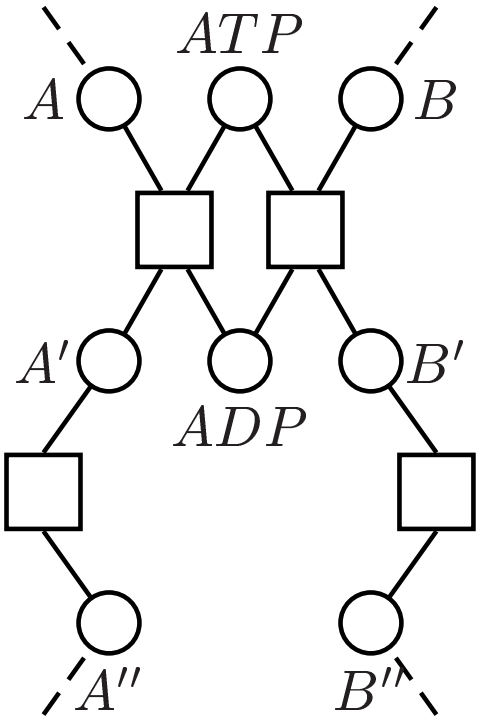
local gap = 0	local gap = 1	local gap = 1
global gap = 0	global gap = 3	global gap = 2

Managing local gap: first compute a transitive closure of the metabolic graph

Are all metabolites equivalent?



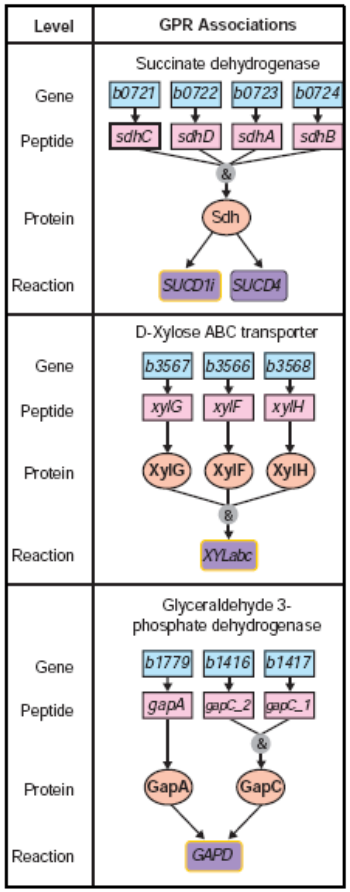
How to handle ubiquitous metabolites ?



Choice 1:
withdraw ubiquitous metabolites

Choice 2:
withdraw secondary metabolites

Gene-Protein-Reaction



- The correspondance between genes and reactions is not always 1 to 1.