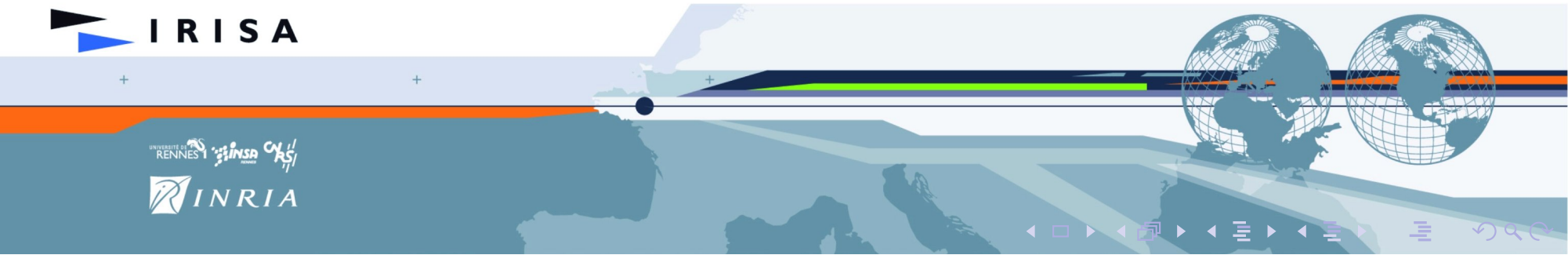
Interaction graph, modules and large scale networks

A. Siegel, P. Veber, C. Guziolowski, M. Le Borgne

CNRS - IRISA

24 juin 2007



Modules ?

A module exists by the question it is associated to

A module is a set of molecules and interactions that satisfy a specific property

Biological criteria

- Biological modules : nodes belongs to the same biological process
- Spacial modules : elements share the same cellular localization
- Time-scale modules : nodes have the same time-scale dynamics

Dynamical criteria

- Time-scale modules : nodes have the same time-scale dynamics
- In-out systems : Their exists a monotonicity relation in the variable behavior (monotonic systems)
- Circuit analysis : Loops that have an influence on the bistability or homeostasie of the system
- Path analysis : Decomposition into minimal paths (Flux Balance Analysis)

Several methods from *dynamical systems* that can be applied to networks with a few hundreds nodes

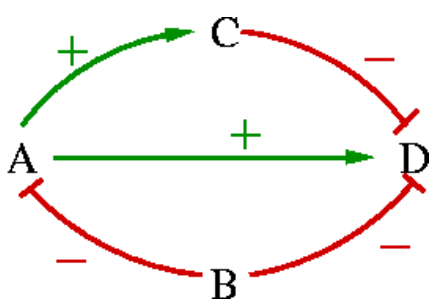
Static analysis and modules

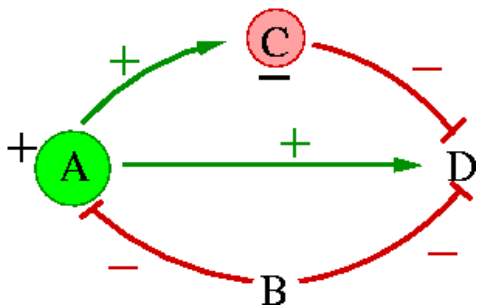
Main object of the talk

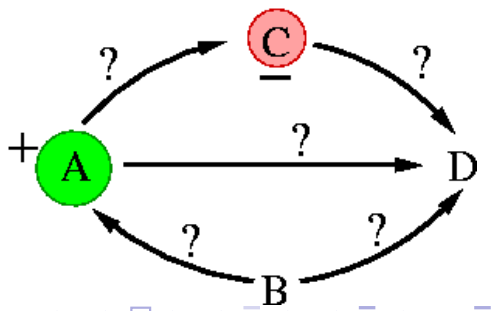
An analysis of shift equilibria allows to identify meaninful modules in large scale networks

Used data and models

- Interaction graph : $A \longrightarrow B$ if A induces a change in the production of B.
  - Signed... : signed interactions are obtained by the reading of the litterature
  - Or unsigned interactions : Chip-Chip experimentations or inference process provide unsigned interactions
- Qualitative variation datasets between two stationnary states : DNA Chip, experimental stress or mutant







Questions...

Asked by biologists

- Consistency between knowledge and data
- Corrections of the model ?
- Prediction of new information
  - Variation for nonobserved products
  - Proposition for the signs of interaction when unknown
- Key nodes
  - For the validation of the model
  - For the understanding of behaviors
  - For the analysis of supplementary material (eQTL)

The main idea

To each question we can associate a type of module that can be computed quite efficiently

Method : Setting constraints depending on the type of avalaible data

Variables

- signs of the variation of products $\Delta X(i,\eta)$ in each considered experimentation

  (underlying hypothesis : data concern stationnary state shifts)

- signs of interactions $s(i \to k)$

  (underlying *restrictive* hypothesis : every actor has a constant action on its target)

Constraints

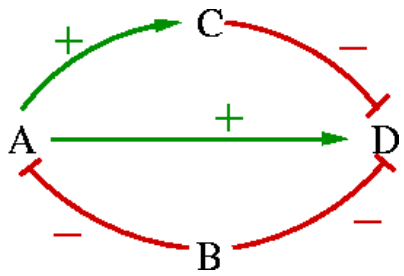
- litterature knowledge set up the signs of some interactions
- qualitative data set up the sign of some variations
- General constraint : the variation of an *internal* product is explained by the variation of one of its predecessors

$$sign(\Delta X(i,\eta)) \simeq \sum_{k \neq i, k \to i} sign(s(i \to k)) \times sign(\Delta X(k,\eta)).$$

Example 1 : interaction signs are known

$$sign(\Delta X(i,\eta)) \simeq \sum_{k\neq i, k\to i} sign(s(i\to k)) \times sign(\Delta X(k,\eta)).$$

Usual sign rules and additional rules : $+ + - = ?$ $\qquad + \not\approx -$



- The variation of C is given by the variation of A
  $sign(\Delta C) \approx sign(\Delta A)$
- the variation of A is the opposite of the variation of B
  $sign(\Delta A) \approx -sign(\Delta B)$
- the variation of D must be equal to the variation of A, $-B$ or $-C$.
  $sign(\Delta D) \approx sign(\Delta A) - sign(\Delta B) - sign(\Delta C)$

$sign(\Delta C) \approx sign(\Delta A)$
$sign(\Delta A) \approx -sign(\Delta B)$
$sign(\Delta D) \approx sign(\Delta A) - sign(\Delta B) - sign(\Delta C)$

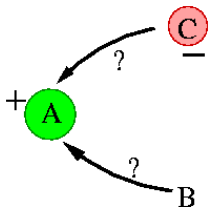
A possible solution to the system :

$$+ \approx +$$
$$+ \approx -(-)$$
$$+ \approx + - (-) - (+)$$

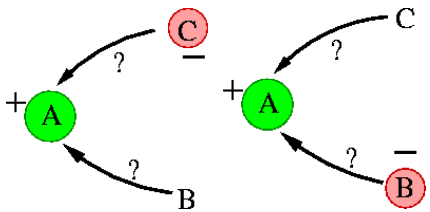
There are 4 sets of solutions (among 16 possible)

A	B	C	D
+	−	+	+
+	−	+	−
−	+	−	+
−	+	−	−

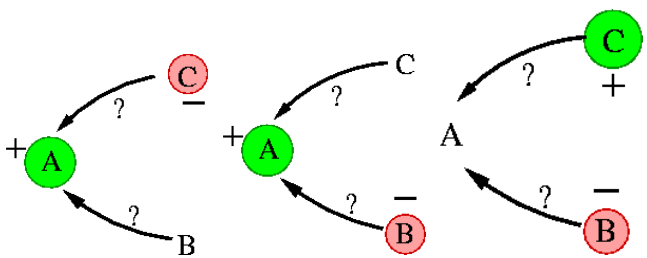
Example 2 : interaction signs are not known



$$sign(\Delta A) \simeq sign(C \to A)\, sign(\Delta C)$$
$$sign(\Delta A) = +$$
$$sign(\Delta C) = -$$



$$sign(\Delta A^{(1)}) \simeq sign(C \to A)\, sign(\Delta C^{(1)})$$
$$sign(\Delta A^{(2)}) \simeq sign(B \to A)\, sign(\Delta B^{(2)})$$
$$sign(\Delta A^{(1)}) = +$$
$$sign(\Delta C^{(1)}) = -$$
$$sign(\Delta A^{(2)}) = +$$
$$sign(\Delta B^{(2)}) = -$$



$$sign(\Delta A^{(1)}) \simeq sign(C \to A)\, sign(\Delta C^{(1)})$$
$$sign(\Delta A^{(2)}) \simeq sign(B \to A)\, sign(\Delta B^{(2)})$$
$$sign(\Delta A^{(3)}) \simeq sign(C \to A)\, sign(\Delta C^{(3)})$$
$$sign(\Delta A^{(3)}) \simeq sign(B \to A)\, sign(\Delta B^{(3)})$$
$$sign(\Delta A^{(1)}) = +$$
$$sign(\Delta C^{(1)}) = -$$
$$sign(\Delta A^{(2)}) = +$$
$$sign(\Delta B^{(2)}) = -$$
$$sign(\Delta B^{(3)}) = +$$
$$sign(\Delta C^{(3)}) = -$$

Studying constraints

Biological questions raise technical duties on systems

- Solving systems
- Eliminating variables
- Reducing systems
- Isolating subsystems

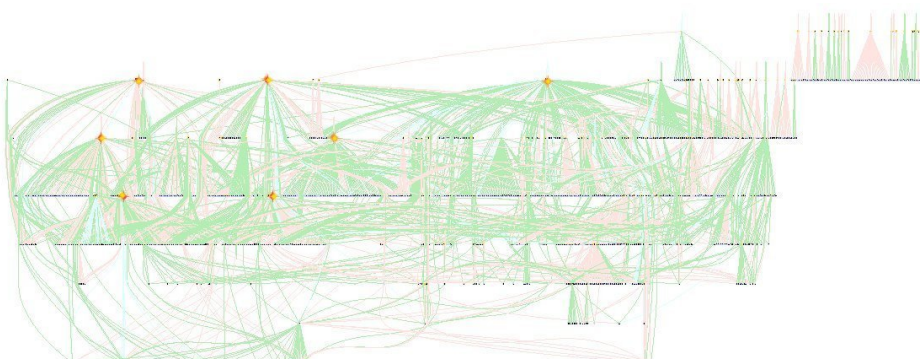
Two mains tools to realize these tasks

- Enumeration of solutions by Decision Diagrams (Pyquali)
  - Compact representation of the solutions in $\{+, -\}$
  - Elimination of variables
  - Efficient for systems of at most 400 variables.
- Solver for constraints expressed in Answer Set Programming (Clasp)
  - Provides one solution for a given set of constraints.
  - Very efficient with thousands of variables.

Question 1 : consistency

- Biological question Are the different pieces of information coherent with each other ?
- Computer scientist question Do the system of constraint admit at least a solution ?
- Solution Write an ASP program and check for the existence of a solution
- Alternative solution Check wether the system of equations has a solution with Decision Diagrams

Example : the network of transcriptional interactions for E. Coli given by Regulon DB is not internally coherent.



- Large scale network with hierarchical structure (87% of genes are regulated by 13%)
- 160 doubled signed interactions
- 1100 constaints, 1258 variables

Number of nodes	1258
Number of interactions	2526
Nodes without successor	1101
Nodes with more than 80 successors	7
protein complex	4

Underlying modules

Core of a system

The core of a biological system described by its interaction graph is the smallest subgraph such that the full system of constraints admits a solution iff the constraints generated by the subgraph admit a solution

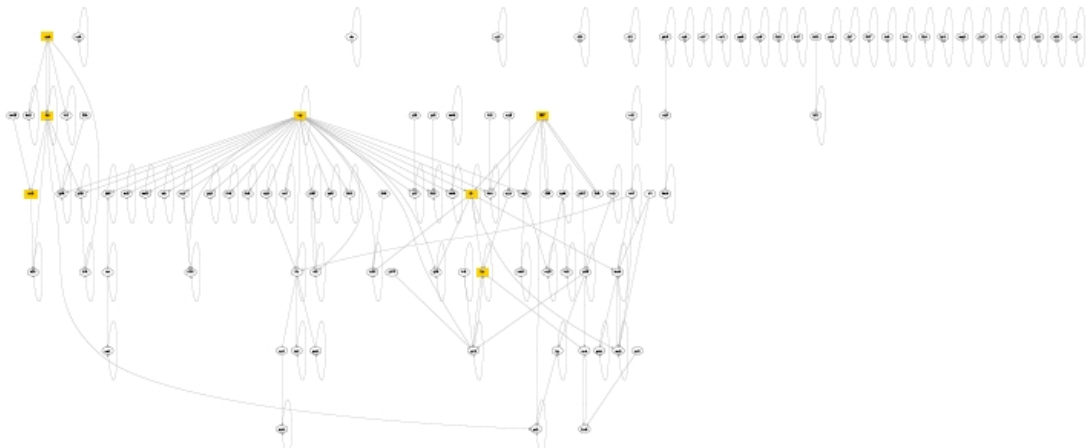
- Computation of an approximation of the graph Keep cycles of the interaction graph and their predecessors
- Used by Kauffman and Peterson to study S. Cerevisiae network.
- Concretely Recursively remove edges and nodes that do not constraint the system
- Interest The search for solutions by decision diagrams becomes possible

Morality : The core of a system contains its dynamics. The rest is static.

Underlying modules

Examples

- E. Coli network reduces from 1258 nodes to 105 nodes and 183 interactions. The central connected component contains only 28 nodes and 57 edges.



- E. Coli network and 43 stationnary phase experimental data reduces from 1258 to 148 nodes and 388 interactions
- S. Cervisiae assuming that the signs of interactions are known. Reduces from 2419 nodes and 4344 interactions to 31 nodes and 52 interactions
- S. Cervisiae with no interaction sign No reduction is possible

Question 2 : Correcting a system

- Biological question When I have contradicting data and knowledge, what should I change ?
- Origin of errors
  - Errors in experimental data or knowledge
  - Missing interaction between nodes
  - Non-constant signed action between an actor and its target
  - (Missing node)
- Computer scientist question What is the minimal set of equations that raise inconsistency ?

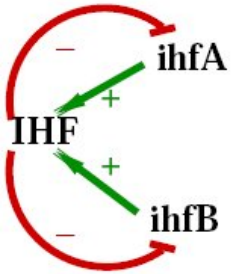
Underlying modules

An inconsistency module is a minimal subset of equations such that the remaining equations are consistent.

Strategy for computation

- Decision diagrams Recursively remove systems of size 1, 2, 3.. ; that are internally inconsistent in order to obtain a consistency system.
- ASP Look for a minimal set of corrections to the inconsistent module.

Example : E. Coli (step 1)

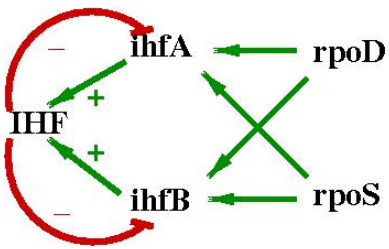


$$IHF \approx ihfA + ihfB \quad (1)$$
$$ihfA \approx -IHF \quad (2)$$
$$ihfB \approx -IHF \quad (3)$$

Automatic finding of inconsistent system (no solution)

ihfA	ihfB	IHF	Conflict
+	+	+	(2), (3)
+	+	−	(1)
+	−	+	(1)
+	−	−	(1)
−	+	+	(1)
−	+	−	(1)
−	−	+	(1)
−	−	−	(2), (3)

Manual curated answer : Adding new interactions (sigma factors)



$$IHF \approx ihfA + ihfB$$
$$ihfA \approx -IHF + rpoD + rpoS$$
$$ihfB \approx -IHF + rpoD + rpoS$$

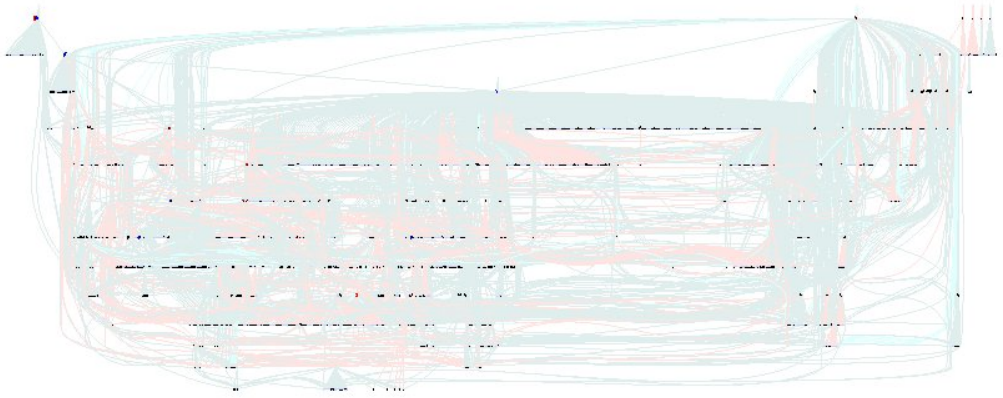
Protein	Gene	Function
σ^{70}	rpoD	Transcribes most genes in growing cells
σ^{38}	rpoS	The starvation/stationary phase sigma-factor
σ^{28}	rpoF	The flagellar sigma-factor
σ^{32}	rpoH	The heat shock sigma-factor
σ^{24}	rpoE	The extracytoplasmic stress sigma-factor
σ^{54}	rpoN	The nitrogen-limitation sigma-factor
σ^{19}	fecI	The ferric citrate sigma-factor

Consistent system (18 solutions among 32)

rpoD	rpoS	ihfA	ihfB	IHF
+	+	+	+	+
+	+	+	−	+
+	+	−	+	+
−	−	−	−	−
−	−	−	+	−
−	−	+	−	−
+/−	−/+	+	+	+
+/−	−/+	+	−	+
+/−	−/+	+	−	−
+/−	−/+	−	+	+
+/−	−/+	−	+	−
+/−	−/+	−	−	−

Example : E. Coli (step 2)

New (consistent) model and data on exponential phase



Number of nodes	1529
Number of interactions	3883
Nodes without successor	1365
Nodes with more than 80 successors	10
sigma-factors	6
protein complex	4

gene	effect	gene	effect	gene	effect	gene	effect	gene	effect
acnA	+	csiE	+	gadC	+	osmB	+	recF	+
acrA	+	cspD	+	hmp	+	osmE	+	rob	+
adhE	+	dnaN	+	hns	+	osmY	+	sdaA	−
appB	+	dppA	+	hyaA	+	otsA	+	sohB	−
appC	+	fic	+	ihfA	−	otsB	+	treA	+
appY	+	gabP	+	ihfB	−	polA	+	yeiL	+
blc	+	gadA	+	lrp	+	proP	+	yfiD	+
bolA	+	gadB	+	mpl	+	proX	+	yihI	−

Model and data are inconsistent !

Correction algorithm. There was a mistake on data provided by RegulonDB

Good variations : $ihfA = +$ and $ihfB = +$ (confirmed by the litterature)

Example : S. Cervisiae

Several unsigned networks for S. Cervisiae

- Core of S. Cervisiae [31 nodes, 52 edges]
- Unsigned interactions between transcription factors (from Chip-Chip analyses or promoteur inference) [70/83 nodes, 96/131 edges]
- Full interaction network given by Chip-Chip analyses (Lee et al, 2002) [2419 nodes 4344 edges]

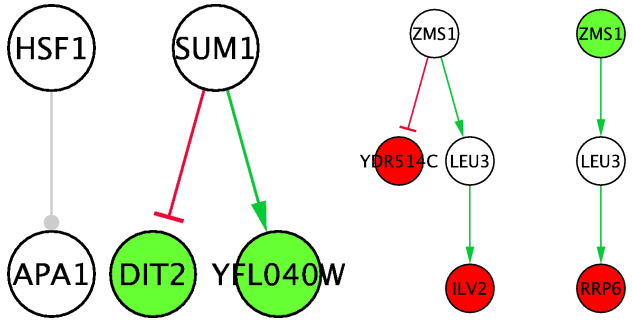
Several datasets

- 15 quite complete stress experimental datasets (YDB)
- About 300 mutant experimentations (Hugues et al, 2000)

All unsigned networks are inconsistent with the datasets

Example : S. Cervisiae (correction)

We identify inconsistent subsets for each network



Interaction network	Nodes	Edges	Number Exp.	Input/Output obs. simult.	MBM Int. TypeI	MBM Int. TypeII,III,IV
(A) Core of Lee network	31	52	15	46	3 (5.7%)	0
(B) Extended Lee network	70	96	15	70	7 (7.2%)	0
(C) Inferred network	83	131	14	91	4 (3%)	0
(D) Global network	2419	4344	14	2270	281 (6.5%)	463 (11%)

Obtaining the largest block ? To be done

Question 3 : Predictions of a system

- Biological question What do the knowledge and data predict on nonobserved signed and/or products ?
- Computer scientist question What are the variables whose sign is the same in all solutions ?

Associated module : hard component

The hard component of a system of constraints is the set of variables that are affected with the same sign in all the solutions to the constraints.

- Decision Diagram Explicitly study the tree of solutions (limited size of nodes)
- ASP For each variable, check whether the systems $\mathcal{S}$ and $(X = +)$ and $\mathcal{S}$ and $(X = -)$ have a solution (30 seconds for each node).

Example : E. Coli and 40 stationnary phase data

Allows to infer 401 new variations (that is, 26 % of the network)

Large scale network with large core

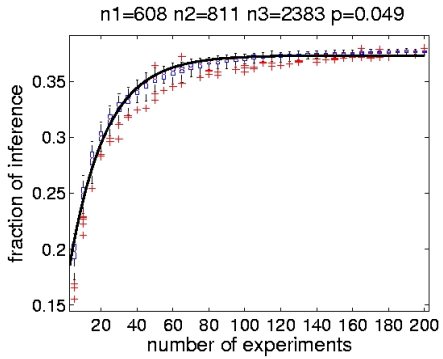
- Decision Diagrams cannot be used and constraint solvers are too long.
- The good strategy : decompose into submodules
  - Partition the set of equations into subsets of equations that share the minimum variables

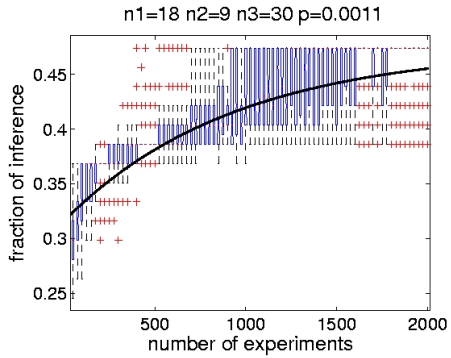
    Obtained by ASP computing
  - For each set of variables, use decision diagrams to eliminate variables outside the considered set
  - Solve the remaining constraints.

Unsigned E. Coli graph and predictability of signs

- Large consistent graph : 1529 nodes and 3802 edges.
- Core of the graph : 28 nodes and 57 edges.
- Random production of consistent sets of signs of variations that simulate random experimental datasets

How many signs can we predict from a given set of observations ?



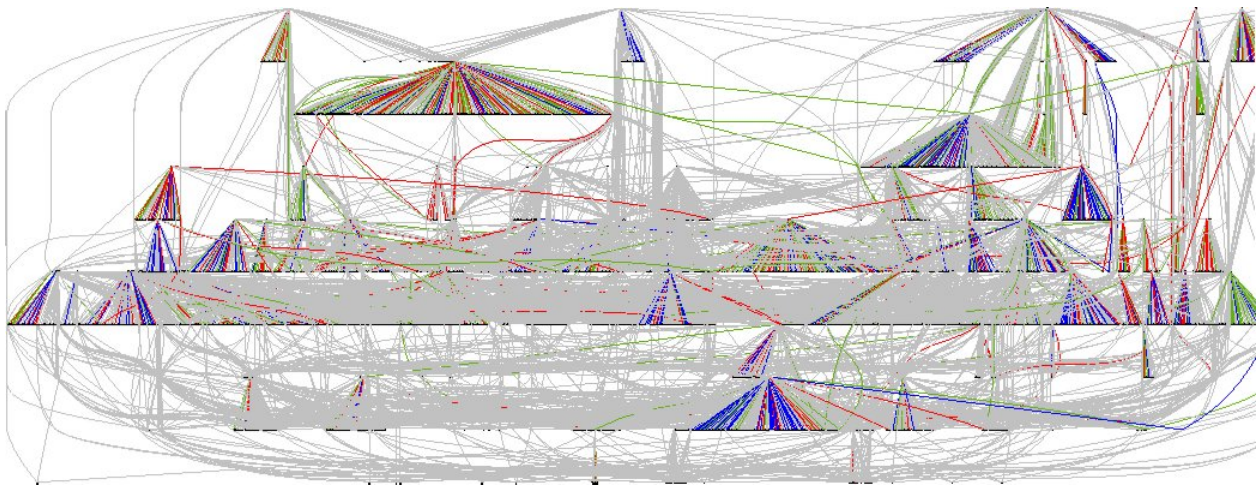


- A maximum of 40,7 % of the graph can be infered.
- In average, 30 experimental datasets are enough to infer 30% of the network.
- 600 signs can be infered from a unique suitable dataset.
- 800 signs can be infered with a probability 0.05.

- A maximum of 47,3 % of the graph can be infered.
- In average, 100 experimental datasets are enough to infer 30% of the network.
- Not all observations have equivalent impact on sign inference

Example : S. Cervisiae

- About 15% of unsigned networks are inconsistent
- About 15% of the remaining unsigned interactions can be infered from 15 datasets.



Question 4 : Key nodes ?

Validation power

- Biological question What are the most important 15 nodes to be observed to ensure that my model is good ?
- Computer scientist question What si the group of 15 nodes that belongs to the minimal number of consistent solutions ?
- Computation To be done (Decision Diagram + ASP)

Prediction power

- Biological question What are the most important 15 nodes to be observed to have the most important influence on the network ?
- Computer scientist question What is the group of 15 nodes that have the most important hard component whatever the consistent signs we consider ?
- Computation To be done (Decision Diagram + ASP)

Conclusions

- Many questions asked by biologists can be solved by using a static approach and constraints solvers
- Each question is associated with a class of modules that can often be computed
- Some of these modules are intrinsicly dynamical and other are static
- More than the size of the network, the important thing is the size of the reduced module associated to a question.