



GraMoFoNe: a Cytoscape plugin for querying motifs without topology in Protein-Protein Interactions networks

Guillaume Blin Florian Sikora Stéphane Vialette

Université Paris-Est, LIGM - UMR CNRS 8049 - France
`{gblin,sikora,vialette}@univ-mlv.fr`

BICoB March 2010

Outline

Introduction

Graph Motif : Querying motifs without topology

GraMoFoNe : a Cytoscape plugin for Graph Motif

GraMoFoNe on real data

Outline

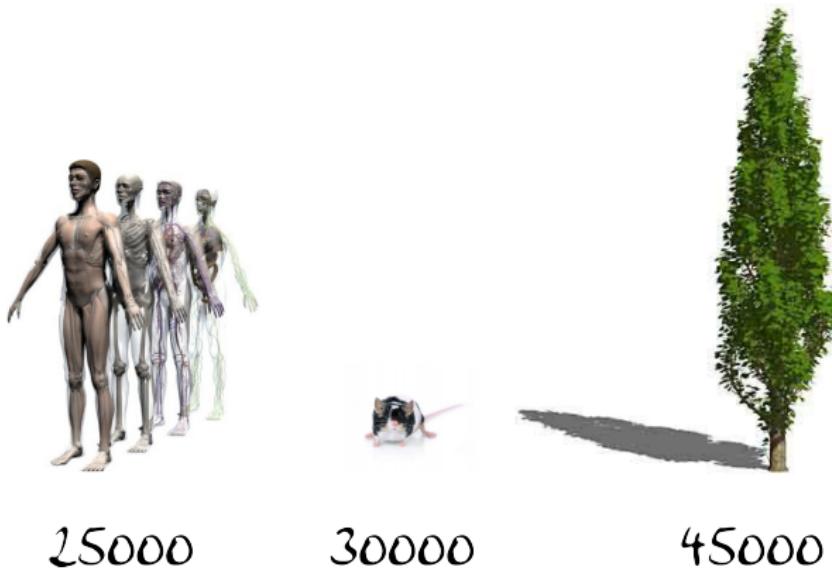
Introduction

Graph Motif : Querying motifs without topology

GraMoFoNe : a Cytoscape plugin for Graph Motif

GraMoFoNe on real data

Motivations



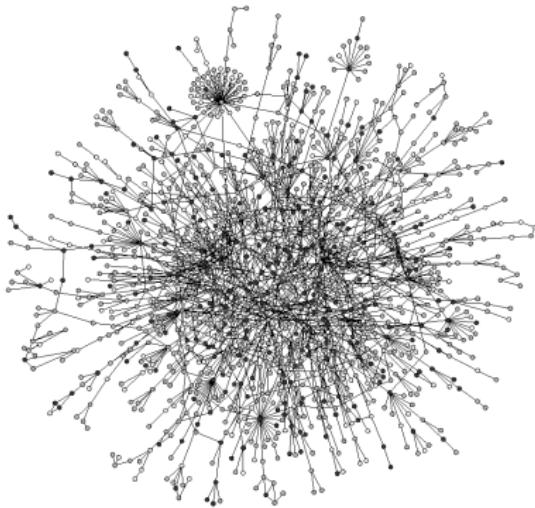
- ▶ Human complexity \Leftrightarrow # of genes ?
- ▶ Human complexity \Leftrightarrow proteins ?

Proteins..

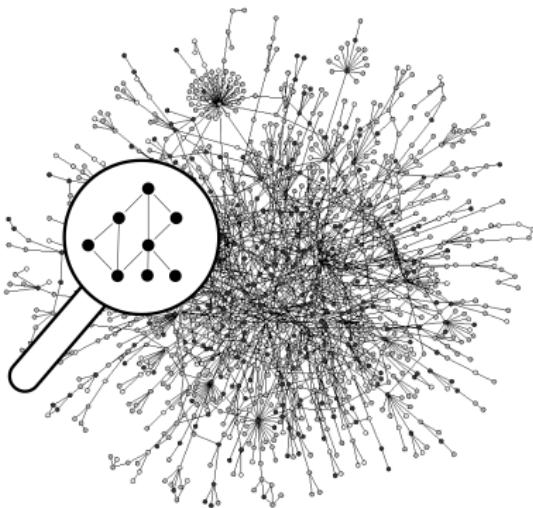
- ▶ New interest on proteins...
- ▶ ... and on their physical interactions: Protein-Protein Interactions (PPI)
- ▶ Biologically obtained... with noise !

Proteins network

- ▶ Proteins can interact with other proteins



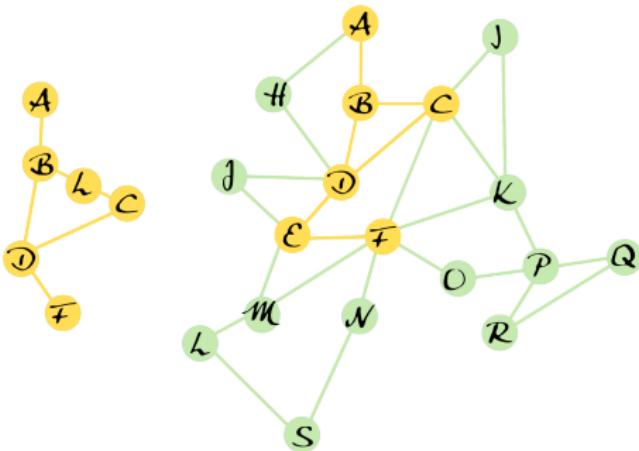
Proteins network



- ▶ Use a (weighted) graph representation
 - ▶ Proteins are nodes
 - ▶ Interactions are edges
 - ▶ Edges can be weighted by interaction probability

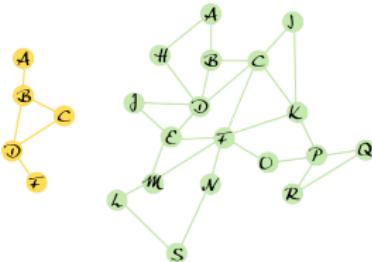
Searching patterns

- ▶ Searching patterns (set of proteins) in a PPI Network can be important to deduce information
- ▶ A protein is said to be **homologous** to another protein according to a BLAST sequence analysis



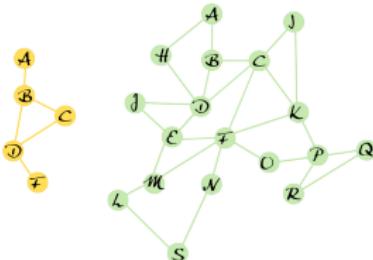
Search patterns

- ▶ Large part of the litterature deals with motif provided with a topology
 - ▶ A path
 - ▶ A tree
 - ▶ A graph under conditions
 - ▶ ...



Search patterns

- ▶ Large part of the litterature deals with motif provided with a topology
 - ▶ A path
 - ▶ A tree
 - ▶ A graph under conditions
 - ▶ ...
- ▶ New point of view : a functionnal one
- ▶ No topology given for the motif



Outline

Introduction

Graph Motif : Querying motifs without topology

GraMoFoNe : a Cytoscape plugin for Graph Motif

GraMoFoNe on real data

GRAPH MOTIF

- ▶ Fact : biological data are noisy
 - ▶ Missing informations, false negatives. About 50% [GAVIN ET AL. 2002]
 - ▶ Erroneous informations, false positives. About 65% [REGULY ET AL. 2006]
- ▶ Topology of the motif can be unknown *a priori*

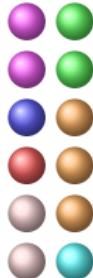
GRAPH MOTIF [LACROIX ET AL. 2006]

- ▶ Each network node is colored by its “function” → network is a vertex-colored graph
- ▶ Motif is a (multi) set of colors
- ▶ Does the motif appears as a subgraph of the network ?

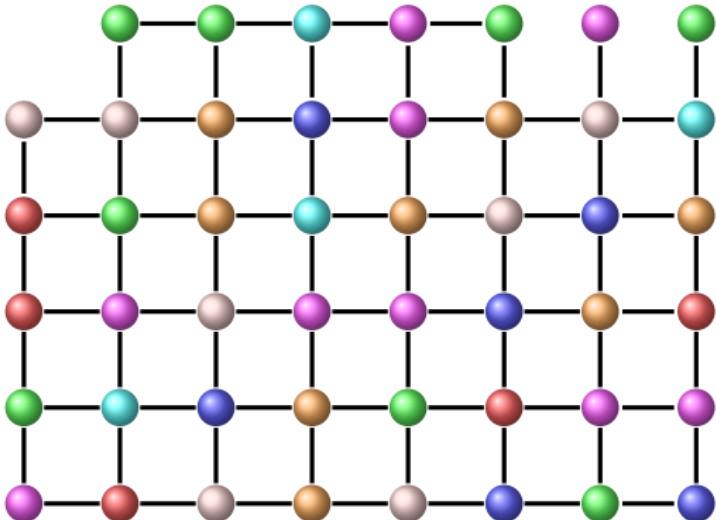
GRAPH MOTIF [LACROIX ET AL. 2006]

- ▶ Each network node is colored by its “function” → network is a vertex-colored graph
- ▶ Motif is a (multi) set of colors
- ▶ Does the motif appears as a subgraph of the network ?
- ▶ Applied to different biological networks [LACROIX ET AL. 2006]
 - ▶ In the case of PPI networks, each motif protein gets a color
 - ▶ A network node has the color c if it is homologous to the protein colored by c in the motif
- ▶ Can be used for other networks (social networks,...)
[BETZLER ET AL. 2008]

GRAPH MOTIF – An example

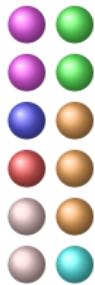


M

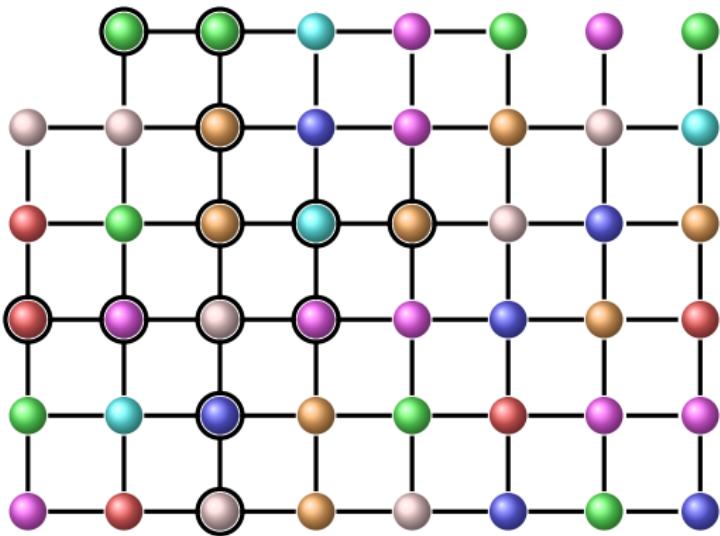


$G = (V, E)$

GRAPH MOTIF – An example

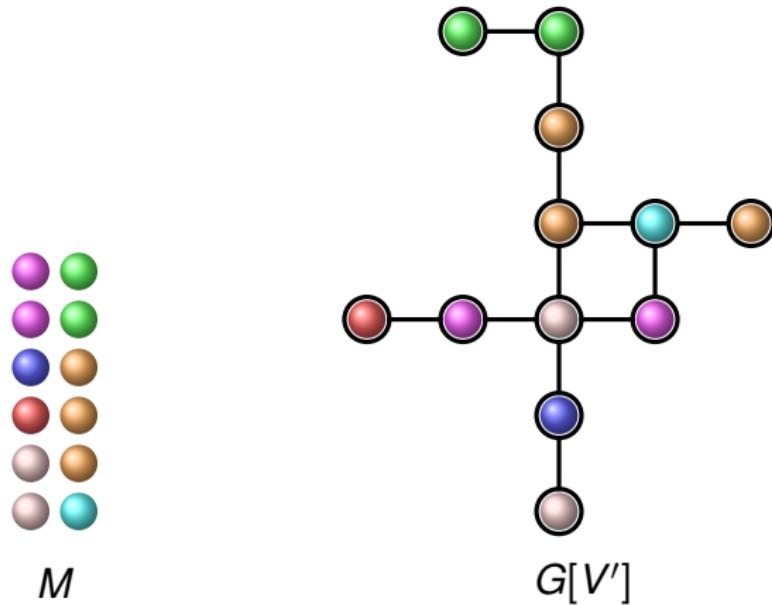


M



$G = (V, E)$
a possible $V' \subseteq V$

GRAPH MOTIF – An example



Softwares for GRAPH MOTIF

- ▶ A lot of theoretical results exists... (NP-Complete even for strong restrictions, FPT, W[1]-hard,...)
- ▶ Only two softwares
- ▶ Torque [BRUCKNER ET AL. 2009] : web service (only) considering colorful motifs
- ▶ GraMoFoNe : a Cytoscape plugin

Outline

Introduction

Graph Motif : Querying motifs without topology

GraMoFoNe : a Cytoscape plugin for Graph Motif

GraMoFoNe on real data

GraMoFoNe – Cytoscape (2002)

- ▶ A free platform, open-source, in Java for
 - ▶ importation / exportation from/to a lot of format, DB,...
 - ▶ visualizing, analyzing interaction networks
 - ▶ integrating annotations, gene expression profiles and other state data to the networks
- ▶ Widely used ("hundreds" of articles cited cytoscape for analyzing...)
- ▶ Up to date...
- ▶ Plugins !



GraMoFoNe – PB

- ▶ We modelize GRAPH MOTIF with Linear Pseudo-Boolean programming
- ▶ *i.e.* linear programming with boolean variables

GraMoFoNe – PB

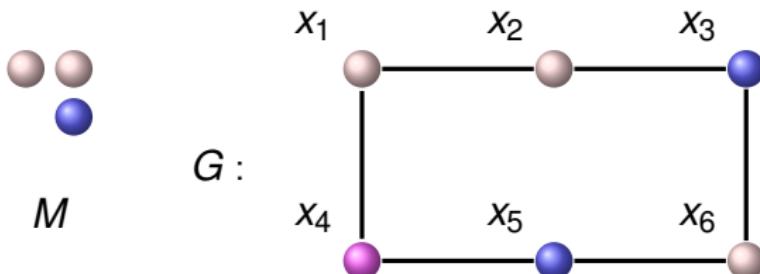
- ▶ Find a variable assignment which satisfies constraints and maximizes an objective function
- ▶ A simple example :
 - ▶ **Variables** : $x_i \in \{0, 1\}, \forall i = 1, 2, 3$
 - ▶ **Objective** : $\max x_1 + 2x_2 - x_3$
 - ▶ **Constraints** :
 1. $2x_1 - 2x_2 + 3x_3 \geq 2$
 2. $x_1 + x_2 + x_3 = 1$
- ▶ Solution : $x_1 = 1, x_2 = 0, x_3 = 0$

GraMoFoNe – PB

- ▶ We use 23 constraints and 9 domains of variables
- ▶ Recall : we look for an occurrence of a motif M in a graph G
- ▶ To respect :
 1. Solution size
 2. Solution coloration equals to the motif
 3. Connectedness of the solution (hardest part)

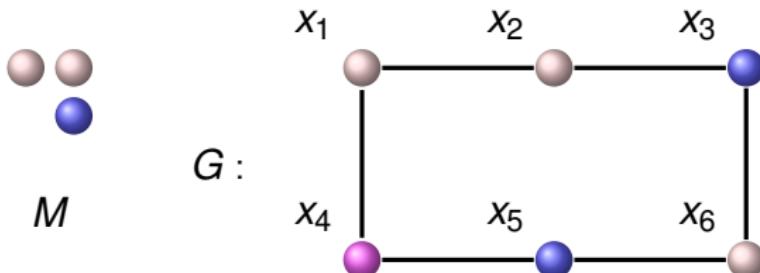
GraMoFoNe – Variables

- ▶ A variable x for each node



GraMoFoNe – Variables

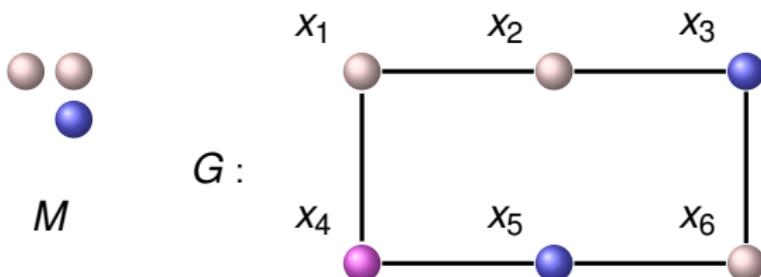
- ▶ A variable x for each node



- ▶ Constraint "solution size" : $\sum_{v \in V} x_v = |M| :$
 - ▶ $x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 3$

GraMoFoNe – Variables

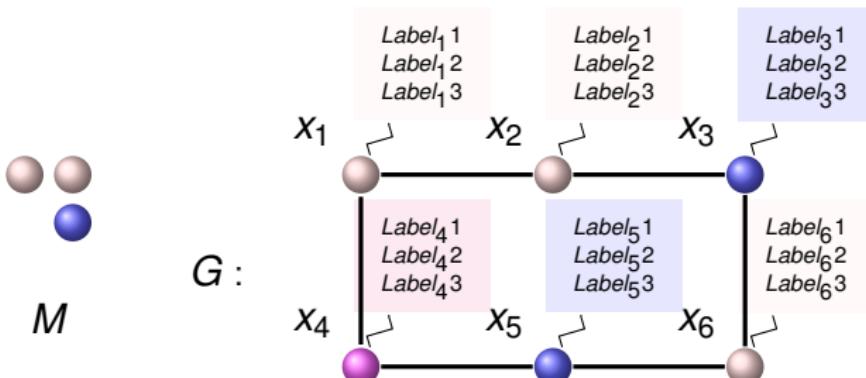
- ▶ A variable x for each node



- ▶ Constraint "solution size" : $\sum_{v \in V} x_v = |M| :$
 - ▶ $x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 3$
- ▶ Constraints "coloration" $\sum_{c \in col(v)} x_v = occ_M(c) :$
 - ▶ $x_1 + x_2 + x_6 = 2$ (white)
 - ▶ $x_3 + x_5 = 1$ (blue)
 - ▶ $x_4 = 0$ (pink)

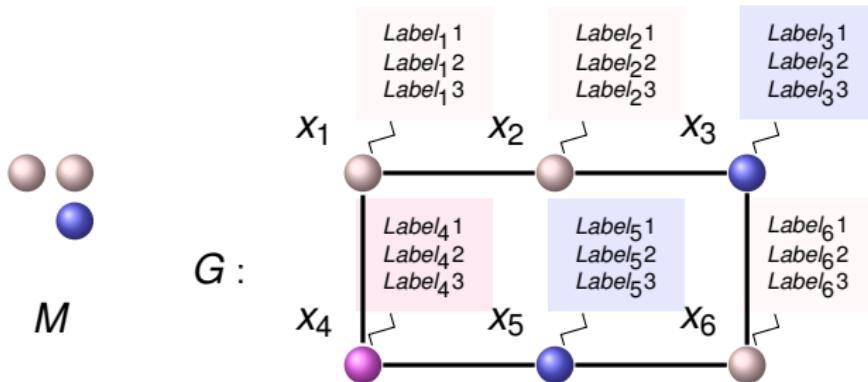
GraMoFoNe – Variables

- Connected solution : $|M|$ variables $Label_v$ for each node v



GraMoFoNe – Variables

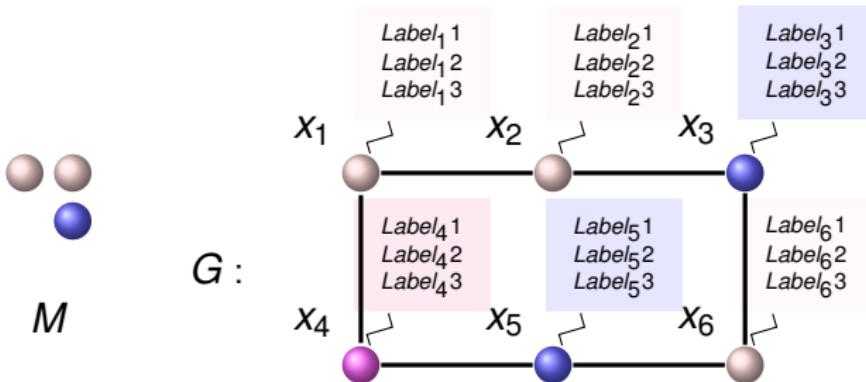
- Connected solution : $|M|$ variables $Label_v$ for each node v



- $|V|$ constraints "one label by node in the solution" :
 - For each v , $x_v \Rightarrow (\sum_{i=1}^{|M|} Label_{vi} = 1)$

GraMoFoNe – Variables

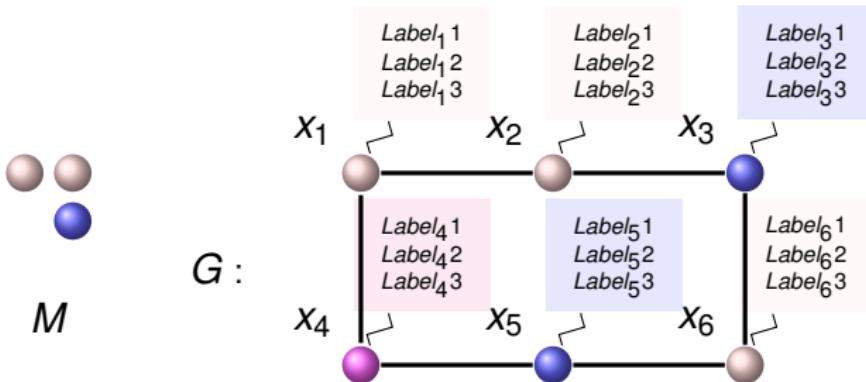
- Connected solution : $|M|$ variables $Label_v$ for each node v



- $|V|$ constraints "one label by node in the solution" :
 - For each v , $x_v \Rightarrow (\sum_{i=1}^{|M|} Label_{v,i} = 1)$
- $|M|$ constraints "a node for a given label"
 - For a given label i , $\sum_{v \in V} Label_{v,i} = 1$

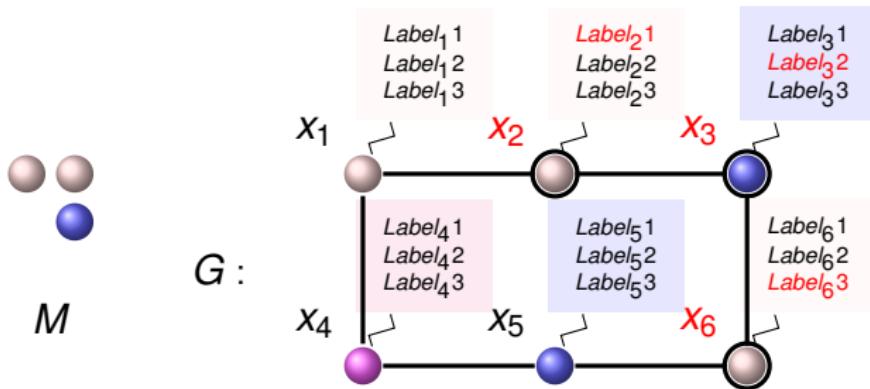
GraMoFoNe – Variables

- Connected solution : $|M|$ variables $Label_v$ for each node v

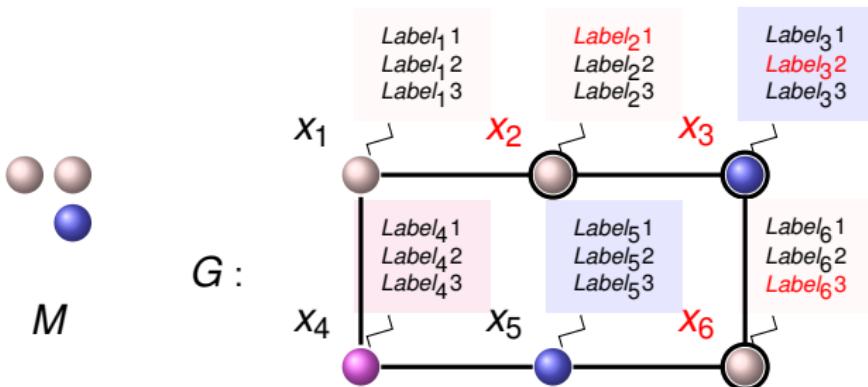


- $|V|$ constraints "one label by node in the solution" :
 - For each v , $x_v \Rightarrow (\sum_{i=1}^{|M|} Label_{v,i} = 1)$
- $|M|$ constraints "a node for a given label"
 - For a given label i , $\sum_{v \in V} Label_{v,i} = 1$
- $|V|.|M|$ constraints "one node with a label has a neighbor with a greater label" (except the last one)
 - $Label_{v,i} \Rightarrow (\sum_{u \in N(v)} \sum_{j > i} Label_{u,j} \geq 1)$

GraMoFoNe – A solution

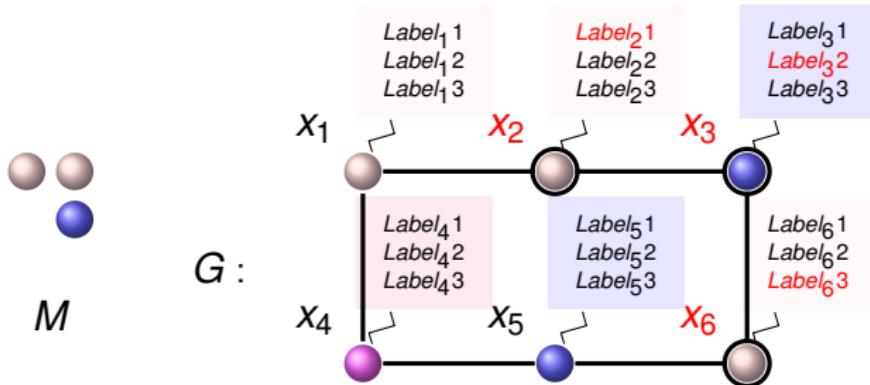


GraMoFoNe – A solution



- ▶ “size” : $x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 3$
- ▶ “coloration”:
 - ▶ $x_1 + x_2 + x_6 = 2$ (white)
 - ▶ $x_3 + x_5 = 1$ (blue)
 - ▶ $x_4 = 0$ (pink)

GraMoFoNe – A solution



- ▶ “one label by node” : $\forall v, x_v \Rightarrow (\sum_{i=1}^{|M|} Label_{vi} = 1)$
- ▶ “one node by label” : $\sum_{v \in V} Label_{vi} = 1$
- ▶ “neighbor with a greater label” :
 $Label_{vi} \Rightarrow (\sum_{u \in N(v)} \sum_{j > i} Label_{uj} \geq 1)$

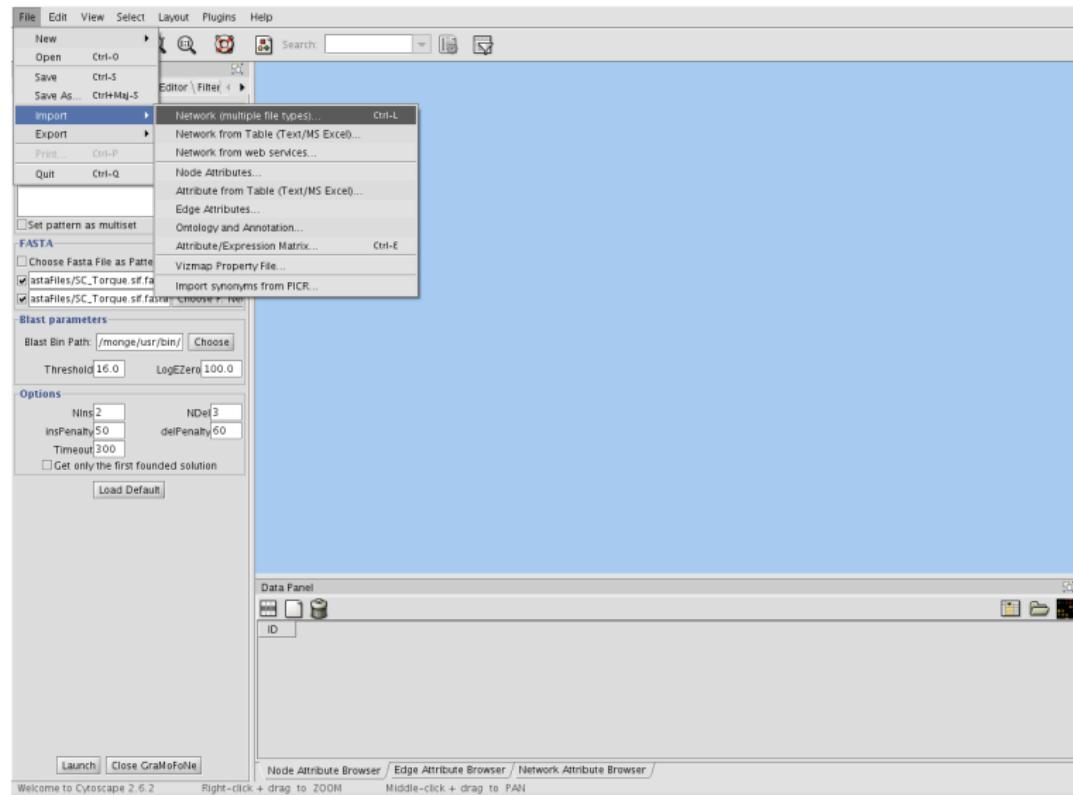
GraMoFoNe – Extensions

- ▶ With a solver over these variables and constraints
 - ▶ “Classic” GRAPH MOTIF (set or multiset motifs)
- ▶ With Pseudo-Boolean Programming, we get a set of possible solutions

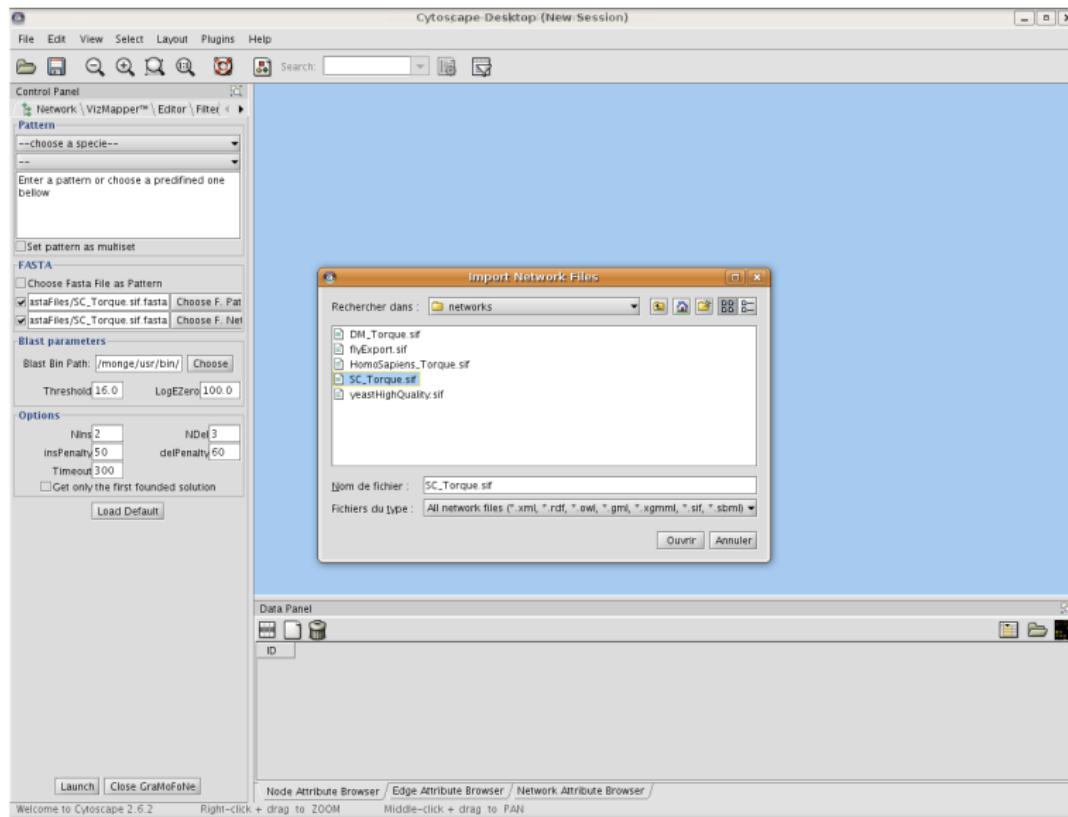
GraMoFoNe – Extensions

- ▶ With a solver over these variables and constraints
 - ▶ “Classic” GRAPH MOTIF (set or multiset motifs)
- ▶ With Pseudo-Boolean Programming, we get a set of possible solutions
- ▶ With more variables and constraints, we can manage
 - ▶ Insertions
 - ▶ Deletions
 - ▶ A set of colors associated to any graph node (a protein can have more than one function)

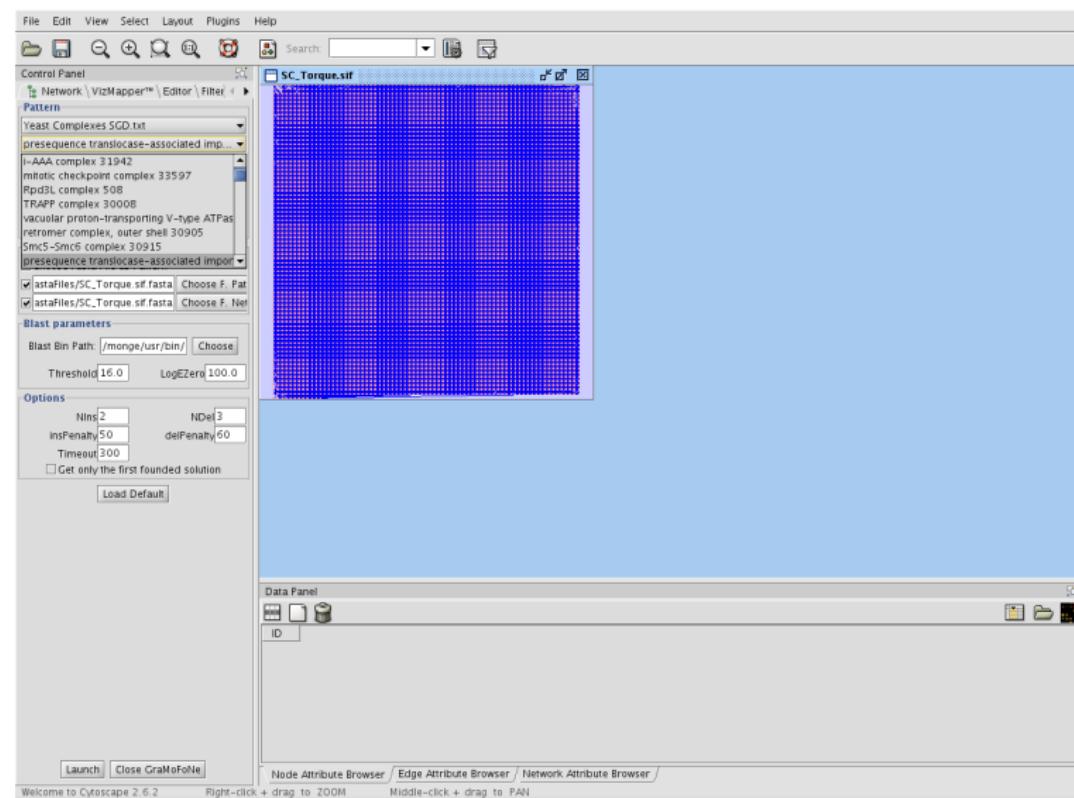
GraMoFoNe – GUI



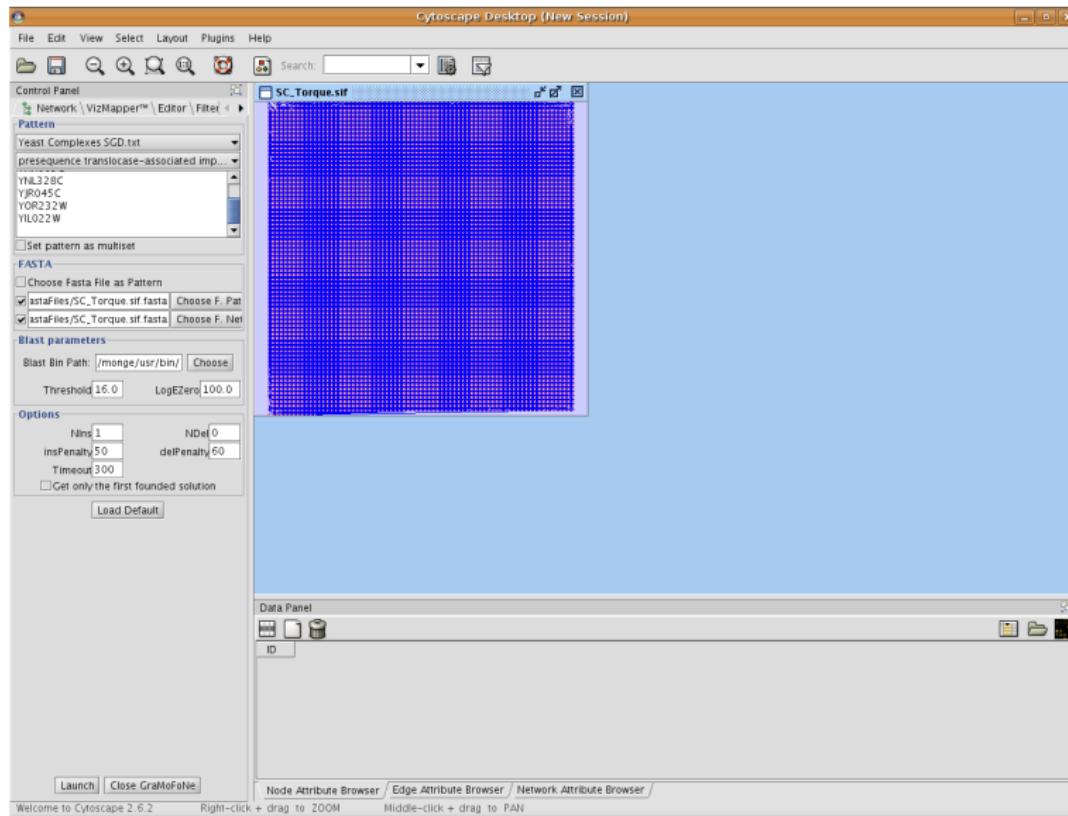
GraMoFoNe – GUI



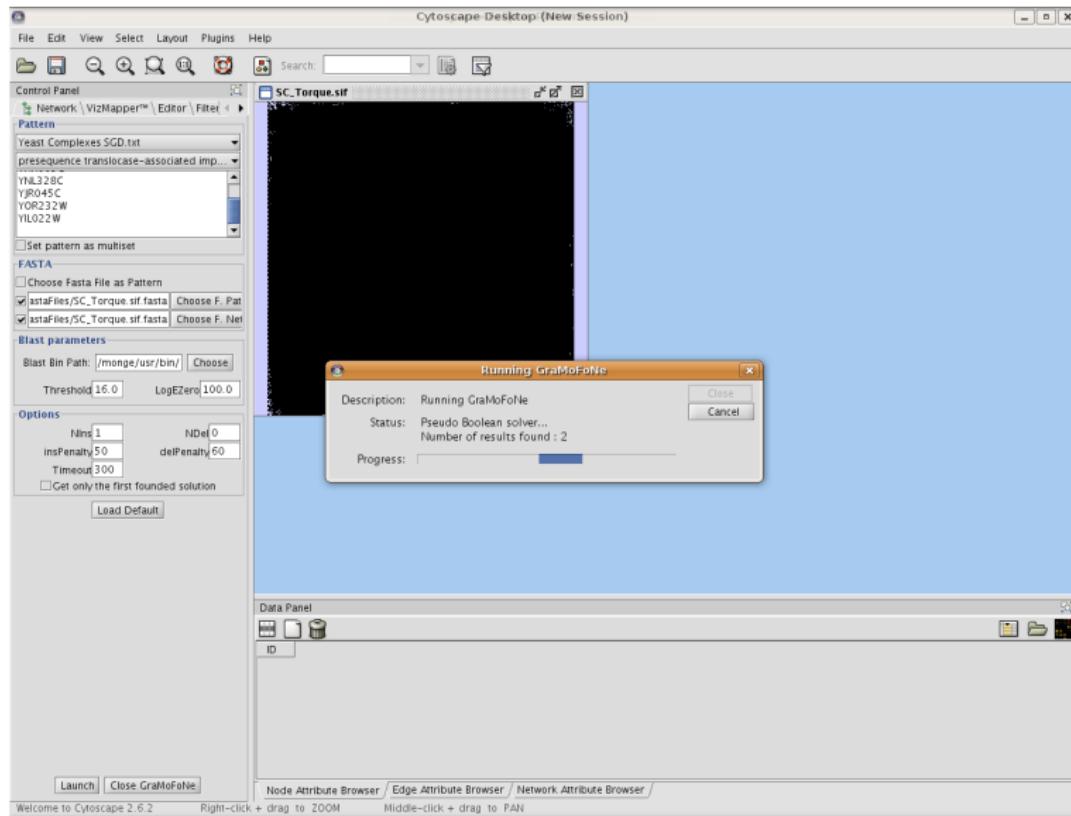
GraMoFoNe – GUI



GraMoFoNe – GUI



GraMoFoNe – GUI



GraMoFoNe – GUI

Cytoscape Desktop (New Session)

Control Panel

- Network VizMapper™ Editor Filter
- Pattern**
 - Yeast Complexes SGD.txt
 - pressequence translocase-associated imp...
 - YNL328C
 - YR045C
 - YOR232W
 - YIL022W
- Set pattern as multiset
- FASTA**
 - Choose Fasta File as Pattern
 - astaFiles/SC_Torque.sif.fasta
 - Choose F. Pat
 - astaFiles/SC_Torque.sif.fasta
 - Choose F. Net
- Blast parameters**
 - Blast Bin Path: /monge/usr/bin/
 - Choose
 - Threshold 16.0
 - LogZero 100.0
- Options**
 - Nins 1 Ndel 0
 - insPenalty 50 delPenalty 50
 - Timeout 300
 - Get only the first founded solution
-

Data Panel

Results Panel

Result	Details
Score = 15.0 Rank = 1 Nb Nodes = 8 Nb Del = 0 Nb Ins = 1 (0 C + 1 NC)	
Score = 14.0 Rank = 2 Nb Nodes = 8 Nb Del = 0 Nb Ins = 1 (0 C + 1 NC)	
Score = 13.0 Rank = 3 Nb Nodes = 8 Nb Del = 0 Nb Ins = 1 (0 C + 1 NC)	
Score = 13.0 Rank = 4 Nb Nodes = 8 Nb Del = 0 Nb Ins = 1 (0 C + 1 NC)	
Score = 13.0 Rank = 5 Nb Nodes = 8	

Node Attribute Browser / Edge Attribute Browser / Network Attribute Browser

Welcome to Cytoscape 2.6.2 Right-click + drag to ZOOM Middle-click + drag to PAN

GraMoFoNe – GUI

Cytoscape Desktop (New Session)

Control Panel

- Network \ VizMapper™ \ Editor \ Filter \
- Pattern**
 - Yeast Complexes SGD.txt
 - presequence translocase-associated imp...
 - YAL328C
 - YJR045C
 - YOR232W
 - YIL022W
- Set pattern as multiset
- FASTA**
 - Choose Fasta File as Pattern
 - astaFiles/SC_Torque.sif.fasta | Choose F. Pat
 - astaFiles/SC_Torque.sif.fasta | Choose F. Net
- Blast parameters**
 - Blast Bin Path: /monge/usr/bin/ | Choose
 - Threshold 16.0 | LogError 100.0
- Options**

Nins 1	NDel 0
insPenalty 50	delPenalty 50
Timeout 300	
<input type="checkbox"/> Get only the first founded solution	

[Load Default](#)

Data Panel

ID
YJR045C
YJL104W
YOR232W
YNR017W
YL008C
YKR065C
YIL022W
YNL328C

[Launch](#) [Close GraMoFoNe](#)

Welcome to Cytoscape 2.6.2 Right-click + drag to ZOOM Middle-click + drag to PAN

[Node Attribute Browser](#) [Edge Attribute Browser](#) [Network Attribute Browser](#)

Results Panel

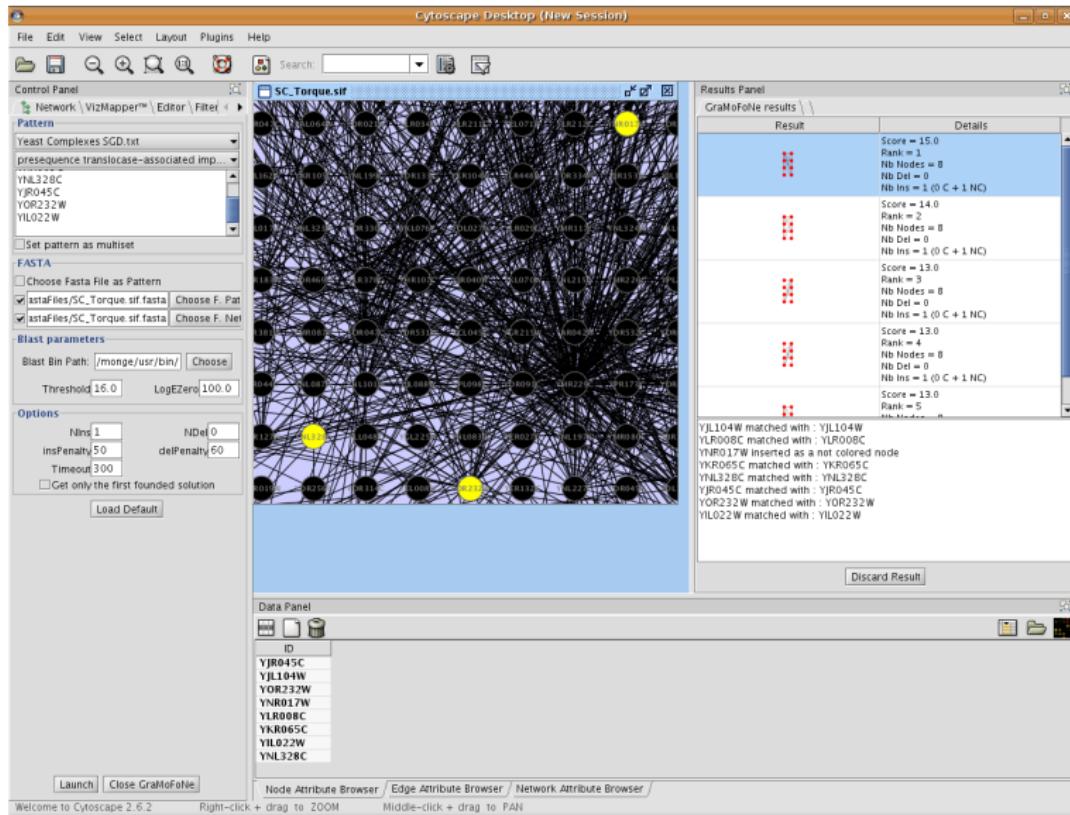
GraMoFoNe results \ \

Result	Details
Score = 15.0	Rank = 1 Nb Nodes = 8 Nb Del = 0 Nb Ins = 1 (0 C + 1 NC)
Score = 14.0	Rank = 2 Nb Nodes = 8 Nb Del = 0 Nb Ins = 1 (0 C + 1 NC)
Score = 13.0	Rank = 3 Nb Nodes = 8 Nb Del = 0 Nb Ins = 1 (0 C + 1 NC)
Score = 13.0	Rank = 4 Nb Nodes = 8 Nb Del = 0 Nb Ins = 1 (0 C + 1 NC)
Score = 13.0	Rank = 5 Nb Nodes = 8 Nb Del = 0 Nb Ins = 1 (0 C + 1 NC)

VJL104W matched with : YJL104W
 YLR008C matched with : YLR008C
 YNR017W inserted as a not colored node
 YKR065C matched with : YKR065C
 YHL232W matched with : YHL232W
 YJR045C matched with : YJR045C
 YOR232W matched with : YOR232W
 YIL022W matched with : YIL022W

[Discard Result](#)

GraMoFoNe – GUI



GraMoFoNe – GUI

File Edit View Select Layout Plugins Help

Control Panel Network VizMapper™ Editor Filter

Pattern

Yeast Complexes SGD.txt
presequence translocase-associated imp...
YNL328C
YJR045C
YOR232W
YLO22W

Set pattern as multiset

FASTA

Choose Fasta File as Pattern
 astafiles/SC_Torque.sif.fasta Choose F. Pat
 astafiles/SC_Torque.sif.fasta Choose F. Pat

Blast parameters

Blast Bin Path: /monge/usr/bin/ Choose
Threshold 16.0 LogEZero 100.0

Options

Node 1	Node 0
insPenalty 50	delPenalty 60
Timeout 300	
<input type="checkbox"/> Get only the first founded solution	

Load Default

Results Panel

GraMoFoNe results \ \

Result Details

Score = 15.0
Rank = 1
Nb Node = 8

Export in a new network
Export in a new network with neighbor

Rank = 2
Nb Nodes = 8
Nb Del = 0
Nb Ins = 1 (0 C + 1 NC)

Rank = 3
Nb Nodes = 8
Nb Del = 0
Nb Ins = 1 (0 C + 1 NC)

Rank = 4
Nb Nodes = 8
Nb Del = 0
Nb Ins = 1 (0 C + 1 NC)

Rank = 5
Nb Nodes = 8

YLR045C matched with : YLR045C
YLR065C matched with : YLR065C
YNL328C matched with : YNL328C
YJR045C matched with : YJR045C
YOR232W matched with : YOR232W
YLO22W matched with : YLO22W

Discard Result

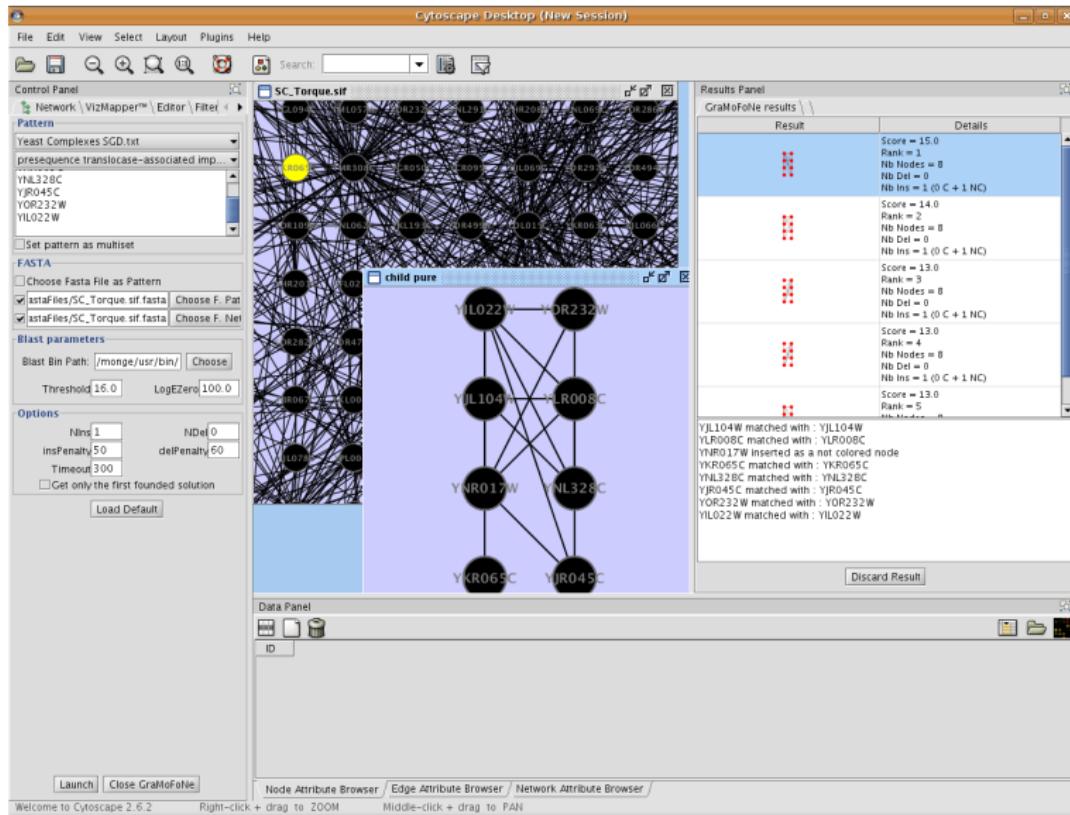
Data Panel

ID
YJR045C
YLR045W
YOR232W
YNR017W
YLR008C
YKR065C
YLO22W
YNL328C

Node Attribute Browser Edge Attribute Browser Network Attribute Browser

Welcome to Cytoscape 2.6.2 Right-click + drag to ZOOM Middle-click + drag to PAN

GraMoFoNe – GUI



Outline

Introduction

Graph Motif : Querying motifs without topology

GraMoFoNe : a Cytoscape plugin for Graph Motif

GraMoFoNe on real data

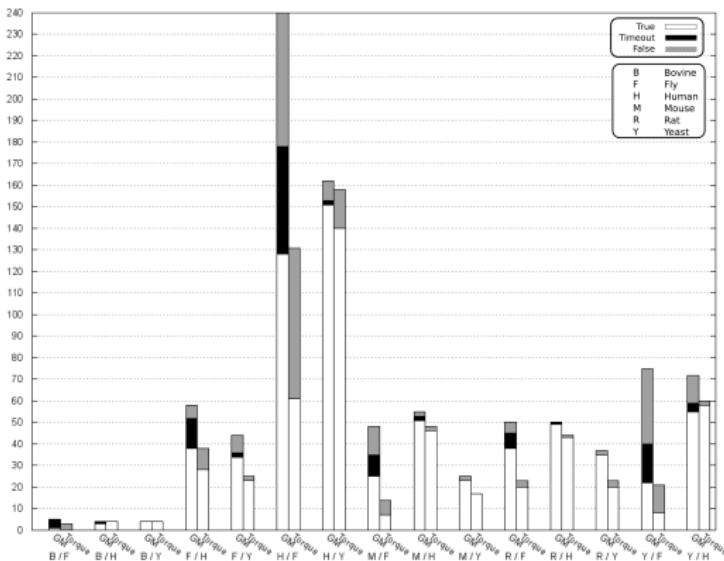
GraMoFoNe – “batch mode”

- ▶ Used for large scale tests
- ▶ Search for dozen of proteins complexes in other species networks
- ▶ Data :
 - ▶ Motifs of 6 species (yeast, fly, homo sapiens, mouse, bovine, rat)
 - ▶ Network of 3 species (yeast, fly, homo sapiens)
 - ▶ Same data as Torque (from up to date DB (SGD, AmiGo, Corum,...) and recent papers)
 - ▶ From 2 to 4 indels allowed according to the motif size

GraMoFoNe – “batch mode”

- ▶ For each motif
 - ▶ Motif found (Before the timeout)
 - ▶ Motif does not exist in the network (Before the timeout)
 - ▶ Timeout reached

GraMoFoNe – “batch mode”



- ▶ False negatives disconnect, false positives give “bad solutions”
- ▶ 5-20s (small M), 40-60s (large). Hard to predict time of PB

Conclusion

- ▶ We provide a software as a Cytoscape plugin
- ▶ Which manage the GRAPH MOTIF problem and some of its variants with Linear Pseudo Boolean Programming
- ▶ Freely available at :
 - ▶ <http://igm.univ-mlv.fr/AlgoB/gramofone/>
 - ▶ Cytoscape plugin page
- ▶ Coloration method given in terms of sequence similarity.
Other measures ?
- ▶ Other relaxations ?

Thank you !

Guillaume Blin Florian Sikora Stéphane Vialette

Université Paris-Est, LIGM - UMR CNRS 8049 - France
`{gblin,sikora,vialette}@univ-mlv.fr`

BICoB March 2010

GRAPH MOTIF [LACROIX ET AL. 2006]

- ▶ Given a (multi)-set of colors M and a vertex-colored graph $G = (V, E)$
- ▶ Find a subset $V' \subseteq V$ s.t.
 - ▶ $G[V']$ is connected
 - ▶ Colors of V' equals M (there is a bijection between the motif and solution colors)

GRAPH MOTIF – NP-Complete

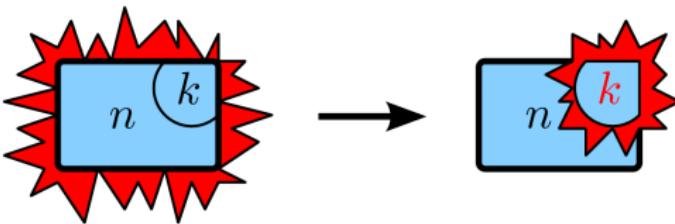
- ▶ A lot of theoretical results...
- ▶ GRAPH MOTIF is NP-Complete, even on strong conditions:
 - ▶ Network is a tree [LACROIX ET AL. 2006]
 - ▶ Network is a tree with maximum degree 3 and the motif is a colorful set [FELLOWS ET AL 2008]
 - ▶ Motif is over 2 colors and the network is a bipartite graph of maximum degree 4 [FELLOWS ET AL 2008]
- ▶ Exact solution → exponential runtime

GRAPH MOTIF – Coping with hardness

- ▶ Fact : **patterns are smaller** ($\sim 5 - 15$) than the network
(e.g. ~ 5.000 for the yeast)
- ▶ Restrict the exponential part to k instead of n :
parameterized complexity

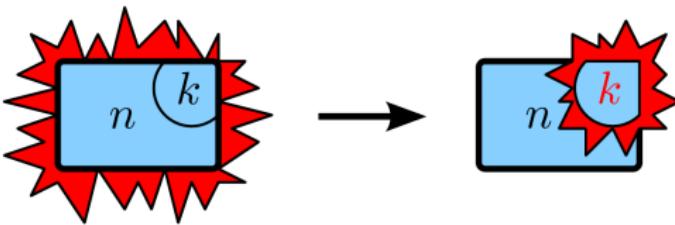
FPT Algorithms

- ▶ An FPT algorithm [DOWNY & FELLOWS 1999]:
exact algorithm **exponential** only in its **parameter k** (not in the input size n)
- ▶ $f(k) \cdot n^c$, with c a constant and f any function



FPT Algorithms

- ▶ An FPT algorithm [DOWNY & FELLOWS 1999]:
exact algorithm **exponential** only in its **parameter k** (not in the input size n)
- ▶ $f(k) \cdot n^c$, with c a constant and f any function
- ▶ Warning, $2^{2^{2^{2^{2^{2^k}}}}} \cdot n$ is FPT...



GRAPH MOTIF – Coping with hardness

- ▶ Sharp borderline between tractable and intractable instances of the problem :
 - ▶ FPT if the parameter is the size k of the motif [LACROIX ET AL. 2006]
 - ▶ $\mathcal{O}^*(2^k)$ if the motif is colorful [BRUCKNER ET AL. 2009]
 - ▶ $\mathcal{O}^*(4.32^k)$ if the motif is a multiset [BETZLER ET AL. 2008]

GRAPH MOTIF – Coping with hardness

- ▶ Sharp borderline between tractable and intractable instances of the problem :
 - ▶ FPT if the parameter is the size k of the motif [LACROIX ET AL. 2006]
 - ▶ $\mathcal{O}^*(2^k)$ if the motif is colorful [BRUCKNER ET AL. 2009]
 - ▶ $\mathcal{O}^*(4.32^k)$ if the motif is a multiset [BETZLER ET AL. 2008]
 - ▶ But: GRAPH MOTIF is W[1]-hard if the parameter is the number of colors [FELLOWS ET AL. 2008] (no FPT algorithm possible with this parameter)

GraMoFoNe – Extensions

- ▶ We have to be careful
- ▶ If there is 1 insertion and 1 deletion, the size of the solution is equals to the size of the motif
- ▶ But we can not bound the number of in/del
- ▶ We have to look for each different color...

- ▶ A node with a set of colors can not match more than one color of the motif (need a bijection)

GraMoFoNe – “batch mode”

- ▶ Preprocessing
 - 1. Proteins without homologous protein in the network are apriori “deleted”
 - 2. If a colored network node is “too far” from another colored network node, we delete it

GraMoFoNe – “batch mode”

- ▶ A motif is “feasible” if
 1. Size 4-25
 2. Less than a fixed number of motif proteins without homologous in the network
 3. A connected component with “enough” colored nodes
- ▶ For each motif
 - ▶ Before the timeout
 - ▶ Motif found
 - ▶ Motif do not exists in the network
 - ▶ Timeout reached