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 literature deals with motif provided with a topology
  - A path
  - A tree
  - A graph under conditions
  - ...

New point of view: a functional one

No topology given for the motif
Search patterns

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

- New point of view: a functional 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
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*
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**

$G = (V, E)$

$M$
GRAPH MOTIF – An example

\[ G = (V, E) \]

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

$M$

$G$

$x_1$ $x_2$ $x_3$

$x_4$ $x_5$ $x_6$
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

$G : M$

- 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": $\sum_{v \in V} x_v = occ_M(c)$:
  - $x_1 + x_2 + x_6 = 2$ (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 $v$

$G: M$

$V| constraints ”one label by node in the solution” :
  - For each $v$, $x_v \Rightarrow (\sum_{i=1}^{M}| Label_v i = 1)$
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 $i$, $\sum_{v \in V} Label_v i = 1$
GraMoFoNe – Variables

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

$$\begin{align*}
&M: \quad \text{Graph} \\
&G: \quad \text{Graph}\end{align*}$$

- $|V|$ 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

\[ G : \]

\[ M \]

\[ x_1 \]
\[ x_2 \]
\[ x_3 \]
\[ x_4 \]
\[ x_5 \]
\[ x_6 \]

\[ \text{Label}_1^1 \]
\[ \text{Label}_1^2 \]
\[ \text{Label}_1^3 \]

\[ \text{Label}_2^1 \]
\[ \text{Label}_2^2 \]
\[ \text{Label}_2^3 \]

\[ \text{Label}_3^1 \]
\[ \text{Label}_3^2 \]
\[ \text{Label}_3^3 \]

\[ \text{Label}_4^1 \]
\[ \text{Label}_4^2 \]
\[ \text{Label}_4^3 \]

\[ \text{Label}_5^1 \]
\[ \text{Label}_5^2 \]
\[ \text{Label}_5^3 \]

\[ \text{Label}_6^1 \]
\[ \text{Label}_6^2 \]
\[ \text{Label}_6^3 \]
GraMoFoNe – A solution

\[ G : \]

\[ \begin{align*}
  x_1 & \quad \rightarrow \quad x_2 \\
  x_2 & \quad \rightarrow \quad x_3 \\
  x_3 & \quad \rightarrow \quad x_4 \\
  x_4 & \quad \rightarrow \quad x_5 \\
  x_5 & \quad \rightarrow \quad x_6 \\
  x_6 & \quad \rightarrow \quad x_1 \\
\end{align*} \]

\[ \begin{align*}
  Label_1 & \quad 1 \\
  Label_2 & \quad 2 \\
  Label_3 & \quad 3 \\
  Label_4 & \quad 4 \\
  Label_5 & \quad 5 \\
  Label_6 & \quad 6 \\
\end{align*} \]

\[ \text{“size” : } x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 3 \]

\[ \text{“coloration”:} \]

\[ \text{Label}_1 \quad 1 \quad \text{(white)} \]
\[ \text{Label}_2 \quad 2 \quad \text{(white)} \]
\[ \text{Label}_3 \quad 3 \quad \text{(white)} \]
\[ \text{Label}_4 \quad 4 \quad \text{(blue)} \]
\[ \text{Label}_5 \quad 5 \quad \text{(blue)} \]
\[ \text{Label}_6 \quad 6 \quad \text{(blue)} \]

\[ x_1 + x_2 + x_6 = 2 \]
\[ x_3 + x_5 = 1 \]
\[ x_4 = 0 \]
GraMoFoNe – A solution

▶ “one label by node” : \( \forall v, x_v \Rightarrow (\sum_{i=1}^{\lvert M \rvert} Label_v i = 1) \)
▶ “one node by label” : \( \sum_{v \in V} Label_v i = 1 \)
▶ “neighbor with a greater label” : 
\[ Label_v i \Rightarrow (\sum_{u \in N(v)} \sum_{j>i} Label_u j \geq 1) \]
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

![Image of GraMoFoNe GUI interface]

- **Control Panel**
  - Select network file(s) using the 'Choose Fasta File as Pattern' option.
  - Configure options such as 'Blast parameters' and 'Options' for the selected network file.

- **Import Network Files**
  - Select files from the directory 'networks'.
  - Choose files like 'SC_Torque.sif'.

- **Data Panel**
  - Explore network attributes and data through the 'Node Attribute Browser' and 'Edge Attribute Browser'.

---

**Querying PPI Networks**
GraMoFoNe – GUI

Welcome to Cytoscape 2.6.2
Right-click + drag to Zoom
Middle-click + drag to Pan
GraMoFoNe – GUI
GraMoFoNe – GUI
GraMoFoNe – GUI

![Cytoplasm Desktop](image)

**Control Panel**
- Yeast Complexes SCO.txt
- presequence translocase-associated imp
- YNL328C
- YPD4458
- YOR232W
- YL022W
- Set pattern as multiset

**FASTA**
- Choose Fasta File as Pattern
- Choose Fasta
- Choose F. Pat
- Choose F. File

**Blast parameters**
- Blast Dir Path
- Default
- Choose

**Options**
- **Score**
- **Rank**
- **Nb Nodes**
- **Nb Del**
- **Nb Ins**
- **C+1 NC**

**Results Panel**
- GraMoFoNe results
- Score: 15.9
- Rank: 1
- Nb Nodes: 0
- Nb Del: 0
- Nb Ins: 1 (C+1 NC)

- Score: 14.9
- Rank: 2
- Nb Nodes: 8
- Nb Del: 0
- Nb Ins: 2 (C+1 NC)

- Score: 13.9
- Rank: 3
- Nb Nodes: 8
- Nb Del: 0
- Nb Ins: 1 (C+1 NC)

- Score: 13.9
- Rank: 4
- Nb Nodes: 8
- Nb Del: 0
- Nb Ins: 1 (C+1 NC)

**Data Panel**
- 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

sikora@univ-mlv.fr

Querying PPI Networks
GraMoFoNe – GUI

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

Node Attribute Browser / Edge Attribute Browser / Network Attribute Browser /

YJR045C
YLR004C matched with : YLR004C
YJR045C matched with : YOR232W

YLR104W
YLR004C matched with : YLR004C
YOR045C matched as a not colored node
YVR045C matched with : YVR045C
YLR004C matched with : YLR004C
YJR045C matched with : YJR045C
YOR232W matched with : YOR232W
YLR022W matched with : YLR022W

Score = 15.9
Rank = 1
Nb Nodes = 0
Nb Del = 0
Nb Ins = 1 · C + 1 · NC

Score = 14.9
Rank = 2
Nb Nodes = 8
Nb Del = 0
Nb Ins = 1 · C + 1 · NC

Score = 13.9
Rank = 3
Nb Nodes = 8
Nb Del = 0
Nb Ins = 1 · C + 1 · NC

Score = 13.9
Rank = 3
Nb Nodes = 8
Nb Del = 0
Nb Ins = 1 · C + 1 · NC

Score = 13.9
Rank = 3
Nb Nodes = 8
Nb Del = 0
Nb Ins = 1 · C + 1 · NC

Intro
Graph Motif
GraMoFoNe
Experiments

sikora@univ-mlv.fr

Querying PPI Networks

23 / 29
GraMoFoNe – GUI
GraMoFoNe – GUI
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)
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
Fact: patterns are smaller ($\sim 5 - 15$) than the network (e.g. $\sim 5.000$ for the yeast)

Restrict the exponential part to $k$ instead of $n$: parameterized complexity
FPT Algorithms

- An FPT algorithm [Downey & Fellows 1999]:
  - exact algorithm **exponential** only in its **parameter** $k$ (not in the input size $n$)
  - $f(k).n^c$, with $c$ a constant and $f$ any function
FPT Algorithms

- An FPT algorithm [Downey & Fellows 1999]:
  exact algorithm \textbf{exponential} only in its \textit{parameter} \( k \) (not in the input size \( n \))
- \( f(k) \cdot n^c \), with \( c \) a constant and \( f \) any function
- Warning, \( 2^{2^{2^{2^{2^k}}}} \cdot n \) 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]
    - $O^*(2^k)$ if the motif is colorful [Bruckner et al. 2009]
    - $O^*(4.32^k)$ 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]
    - $O^*(2^k)$ if the motif is colorful [Bruckner et al. 2009]
    - $O^*(4.32^k)$ 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