Towards Motif Detection in Networks: Frequency Concepts and Flexible Search

Falk Schreiber   Henning Schwöbbermeyer

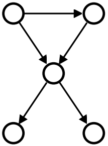
Bioinformatics Center Gatersleben-Halle,
Institute of Plant Genetics and Crop Plant Research

NETTAB Workshop, 2004

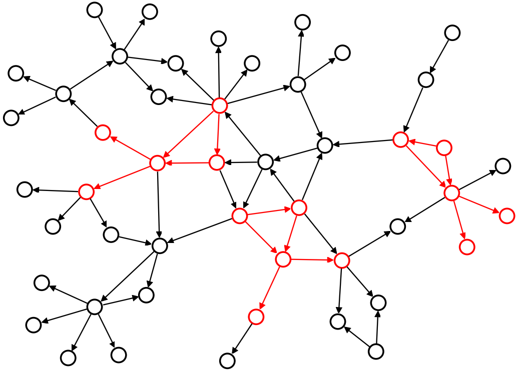
Motivation

Motif Detection in Networks:

- Searching
- Counting
- Visual Exploration



Pattern
Size: $|E_p|$



Target graph $G_t = (V_t, E_t)$ with highlighted pattern matches

Outline

Topological analysis of the network structure

- Biological processes from large and complex networks
- These networks become now available as a whole
- Topological network analysis may help to uncover important properties
  - Global network structure
  - Centrality
  - Clustering
  - Network motifs
- Interesting motifs are found in biological networks (Milo et al., 2002; Wuchty et al., 2003):
  - Gene regulatory, metabolic, protein-protein interaction, neuronal, food-webs

Topological analysis of the network structure

- Biological processes from large and complex networks
- These networks become now available as a whole
- Topological network analysis may help to uncover important properties
  - Global network structure
  - Centrality
  - Clustering
  - Network motifs
- Interesting motifs are found in biological networks (Milo et al., 2002; Wuchty et al., 2003):
  - Gene regulatory, metabolic, protein-protein interaction, neuronal, food-webs

Characterisation of network motifs

Motifs are:

- Patterns of local interconnections
- May represent the simplest building blocks of functional modules
- Varying definitions
  - Single small network (Wuchty et al., 2003)
  - Set of related networks (Shen-Orr et al., 2002)
  - Patterns with significant high frequency (Milo et al., 2002)

We define *motifs* as substructures with functional properties within a network.

For substructures where the function is currently unknown the term *pattern* is used.

Analysing networks and finding interesting patterns

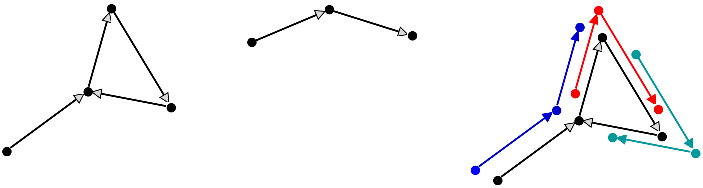
Frequent patterns in networks

- Pattern frequency: number of matches in the target graph
- Patterns with high frequency as potential candidates for functional network motifs
- Applications in other fields
  - Toxicology or carcinogenicity of chemical substances (Srinivasan et al., 1997)
  - Classification of sequential logic circuits (Milo et al., 2002)

Different concepts for frequency determination as a result of different restrictions of the reuse of graph elements

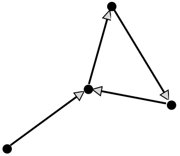
Concepts for determination of pattern frequency

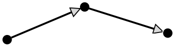
Concept	Graph element reuse		Frequency determination
	Vertices	Edges	
$\mathcal{F}_1$	yes	yes	All matches
$\mathcal{F}_2$	yes	no	Maximum independent set
-	no	yes	-
$\mathcal{F}_3$	no	no	Maximum independent set

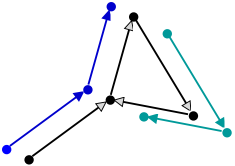


Concepts for determination of pattern frequency

Concept	Graph element reuse		Frequency determination
	Vertices	Edges	
$\mathcal{F}_1$	yes	yes	All matches
$\mathcal{F}_2$	yes	no	Maximum independent set
-	no	yes	-
$\mathcal{F}_3$	no	no	Maximum independent set

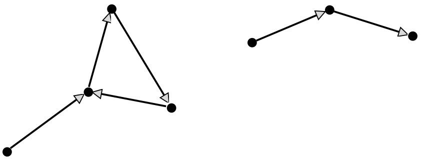






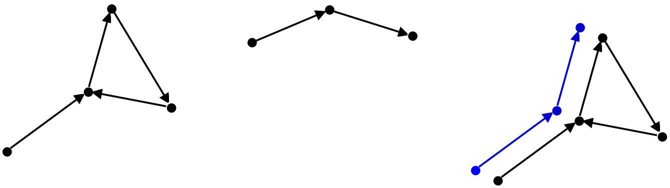
Concepts for determination of pattern frequency

Concept	Graph element reuse		Frequency determination
	Vertices	Edges	
$\mathcal{F}_1$	yes	yes	All matches
$\mathcal{F}_2$	yes	no	Maximum independent set
-	no	yes	-
$\mathcal{F}_3$	no	no	Maximum independent set



Concepts for determination of pattern frequency

Concept	Graph element reuse		Frequency determination
	Vertices	Edges	
$\mathcal{F}_1$	yes	yes	All matches
$\mathcal{F}_2$	yes	no	Maximum independent set
-	no	yes	-
$\mathcal{F}_3$	no	no	Maximum independent set



Properties of the frequency concepts

$\mathcal{F}_1$

- Does not exclude matches
- Shows the full potential of the pattern

$\mathcal{F}_2$

- Matches does not share relation of elements
- Shows the maximum number of instances of a particular pattern which can be "active" at the same time

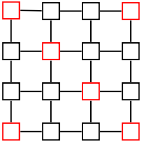
$\mathcal{F}_3$

- Matches can be seen as non-overlapping clusters
- Allows specific analysis and navigation methods
  - Folding and unfolding of clusters
  - Pattern preserving layout of the matches

Maximum independent set (MIS)

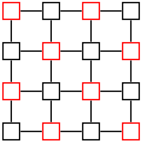
Independent set:

- Set of vertices V' (a subset of V) such that for every two vertices in V', there is no edge connecting the two.



Maximum independent set:

- Largest independent-set in a graph G
- Calculation is NP-complete



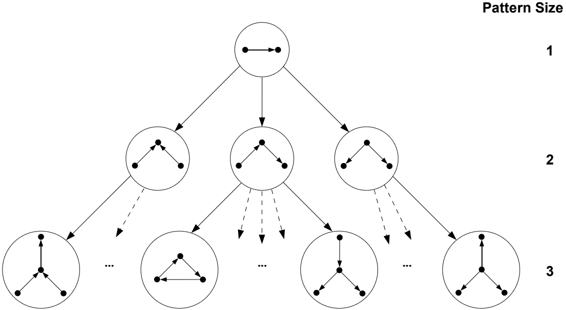
A fast heuristic is used for approximation of the MIS.

Frequent pattern finding algorithm

- Basic algorithm presented by *Kuramochi and Karypis* (2004) for search of pattern of given size with maximum frequency in undirected graphs.
- Several extensions
  - Application of different frequency concepts
  - Search for patterns in directed graphs
  - Full control over the search, e.g. define frequency threshold
  - Parallel implementation of the search algorithm
- Outline of the involved tasks in frequent pattern finding:
  - Generation of the different patterns of target size
  - Isomorphism testing of generated patterns
  - Determination of pattern matches
  - Calculation of the frequency of the patterns

Traversal of the pattern space

- Each pattern is assigned to parent pattern of size $n-1$
- Generation of the patterns supported by the target graph
- Depth first traversal of the pattern tree
  - Allows pruning of infrequent branches for $\mathcal{F}_2$ and $\mathcal{F}_3$



Data : Graph $G = (V, E)$, target size t, frequency concept $\mathcal{F}$
Result: Pattern with maximum frequency

Start of search with the pattern of size 1
$\longrightarrow$ Each edge of target graph used to create a match

while *Patterns for extension are left* **do**

Extend next pattern by addition of one edge
$\longrightarrow$ Combine matches with each incident edge to new pattern
If ancestor is generating parent $\rightarrow$ keep pattern

Compute frequency for each new pattern
If frequency is above threshold $\rightarrow$ keep pattern

Log patterns of target size as result, adjust threshold

end

Problems and complexity

Problem

- Many different patterns

Rev. edges	-	+	-	+
Self loops	-	-	+	+
Pattern size				
2	3	4	5	6
3	10	12	18	21
4	39	53	76	97
5	169	237	361	478
6	876	1306	1978	2762
7	4834	7537	11658	17002
8	29316	47913	74494	113528
9	189054	322253	505277	801966

Solution

- Only consider patterns supported by target graph
- Pruning of pattern tree

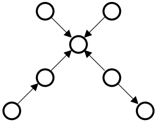
Problems and complexity (continued)

Problem	Solution
Many different matches for one pattern • $O(\lvert E_t \rvert^{\lvert E_p \rvert})$	Only extension of matches of parent pattern
Maximum independent set	Use of heuristic
Graph isomorphism	Canonical labelling
In worst case computational expensive	In practise for moderate sized networks applicable

Frequent patterns in gene-regulatory network

- Interactions of transcription factors in yeast (Lee et al., 2002)
- 106 transcriptional regulators with 108 interactions
- Totally 1811 different patterns of size 6

	Frequency		
Rank $\mathcal{F}_1$	$\mathcal{F}_1$	$\mathcal{F}_2$	$\mathcal{F}_3$
1	3175	6	3
2	2785	10	5
3	2544	9	4
4	2422	8	4
5	2106	5	3
6	2006	8	4
18	1627	12	5

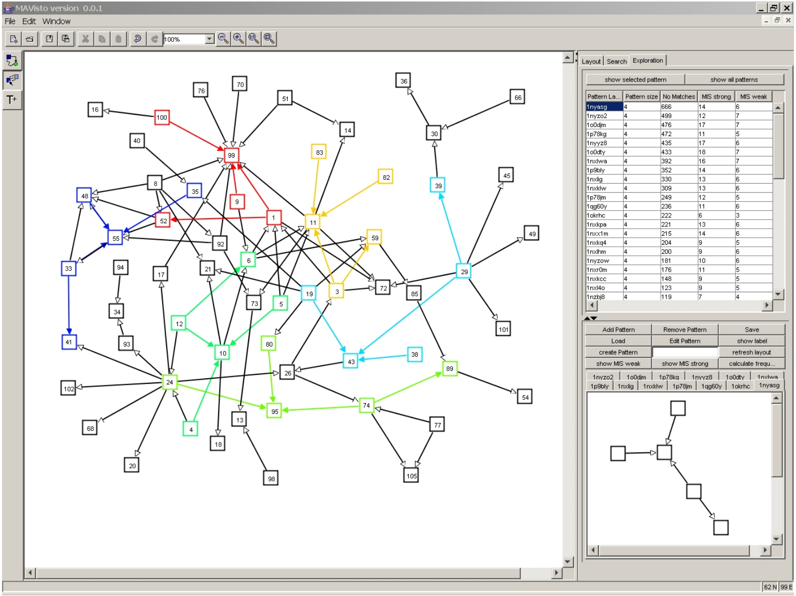


Pattern at position 1

Pattern at position 18

Mavisto - Motif Analysis and Visualisation Tool

- Implements the presented features for frequent pattern search
- Visual analysis methods for patterns in networks
- First release as Java Webstart Application in the near future
- Announcements under http://nwg.bic-gh.de/
- Based on Gravisto (Bachmaier et al., 2004)
  - Editor for graphs and a toolkit for implementing graph visualisation algorithms
  - Developed at University of Passau, Germany
  - Gravisto is licensed under the GPL
  - Available under http://www.gravisto.org/


MAVisto version 0.0.1
File Edit Window
100%
Layout Search Exploration
show selected pattern
show all patterns
Pattern La... Pattern size No Matches MIS strong MIS weak
1nyasg 4 666 14 6
1nyzo2 4 499 12 7
1o0djm 4 476 17 7
1p78kg 4 472 11 5
1nyyz8 4 435 17 6
1o0dty 4 433 18 7
1nxlwa 4 392 16 7
1p9bly 4 352 14 6
1nxlig 4 330 13 6
1nxklw 4 309 13 6
1p78jm 4 249 12 5
1qg60y 4 236 11 6
1okrhc 4 222 6 3
1nxkpa 4 221 13 6
1nxx1m 4 215 14 6
1nxkq4 4 204 9 5
1nxlhm 4 200 9 6
1nyzow 4 181 10 6
1nxr0m 4 176 11 5
1nxkcc 4 148 9 5
1nxl4o 4 123 9 5
1nzbj8 4 119 7 4
Add Pattern
Remove Pattern
Save
Load
Edit Pattern
show label
create Pattern
refresh layout
show MIS weak
show MIS strong
calculate frequ...
1nyzo2 1o0djm 1p78kg 1nyyz8 1o0dty 1nxlwa
1p9bly 1nxlig 1nxklw 1p78jm 1qg60y 1okrhc 1nyasg
62 N 99 E

Summary and outlook

Presented topics related to frequent pattern mining

- Three different frequency concepts
- Flexible algorithm to search for frequent patterns
- Mavisto - Motif Analysis and Visualisation Tool

Extensions planned for the future

- Incorporation of more information of the network elements into the search
- Improvement of visual analysis part

Acknowledgements

Members of the Network Analysis Group

- Falk Schreiber
- Dirk Koschützki
- Christian Klukas

The Gravisto Team at University Passau

Towards Motif Detection in Networks: Frequency Concepts and Flexible Search

Mavisto http://nwg.bic-gh.de

Demos on demand