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

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

Network Motifs

Concepts for Pattern Frequency

Frequent Pattern Search

Application on Biological Data

Motif Analysis and Visualisation
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

<table>
<thead>
<tr>
<th>Concept</th>
<th>Graph element reuse</th>
<th>Frequency determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{F}_1$</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>$\mathcal{F}_2$</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>-</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>$\mathcal{F}_3$</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

**Motif Detection in Networks**
### Concepts for determination of pattern frequency

<table>
<thead>
<tr>
<th>Concept</th>
<th>Graph element reuse</th>
<th>Frequency determination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Vertices</td>
<td>Edges</td>
</tr>
<tr>
<td>$\mathcal{F}_1$</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>$\mathcal{F}_2$</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>-</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>$\mathcal{F}_3$</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

![Graph Element Diagram](image-url)
## Concepts for determination of pattern frequency

<table>
<thead>
<tr>
<th>Concept</th>
<th>Graph element reuse</th>
<th>Frequency determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{F}_1$</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>$\mathcal{F}_2$</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>$\mathcal{F}_3$</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Vertices</th>
<th>Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{F}_1$</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>$\mathcal{F}_2$</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>$\mathcal{F}_3$</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

![Graph example](image.png)
### Concepts for determination of pattern frequency

<table>
<thead>
<tr>
<th>Concept</th>
<th>Graph element reuse</th>
<th>Frequency determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{F}_1$</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>$\mathcal{F}_2$</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>-</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>$\mathcal{F}_3$</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

**Graphs:**
- **$\mathcal{F}_1$:** Triangle
- **$\mathcal{F}_2$:** Path
- **$\mathcal{F}_3$:** 3-clique
Properties of the frequency concepts

\( F_1 \)
- Does not exclude matches
- Shows the full potential of the pattern

\( F_2 \)
- 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

\( F_3 \)
- 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


- 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

$\rightarrow$ 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
  $\rightarrow$ 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

<table>
<thead>
<tr>
<th>Rev. edges</th>
<th>-</th>
<th>+</th>
<th>-</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self loops</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pattern size</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>10</td>
<td>39</td>
<td>169</td>
<td>876</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>12</td>
<td>53</td>
<td>237</td>
<td>1306</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>18</td>
<td>76</td>
<td>361</td>
<td>1978</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>21</td>
<td>97</td>
<td>478</td>
<td>2762</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>21</td>
<td>97</td>
<td>478</td>
<td>2762</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>21</td>
<td>97</td>
<td>478</td>
<td>2762</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>21</td>
<td>97</td>
<td>478</td>
<td>2762</td>
</tr>
</tbody>
</table>

- Many different patterns

#### Solution

- Only consider patterns supported by target graph
- Pruning of pattern tree
### Problems and complexity (continued)

<table>
<thead>
<tr>
<th>Problem</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Many different matches for one pattern</td>
<td>Only extension of matches of parent pattern</td>
</tr>
<tr>
<td>• $O(</td>
<td>E_t</td>
</tr>
<tr>
<td>Maximum independent set</td>
<td>Use of heuristic</td>
</tr>
<tr>
<td>Graph isomorphism</td>
<td>Canonical labelling</td>
</tr>
<tr>
<td>In worst case computational expensive</td>
<td>In practise for moderate sized networks applicable</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Rank $\mathcal{F}_1$</th>
<th>Frequency $\mathcal{F}_1$</th>
<th>Frequency $\mathcal{F}_2$</th>
<th>Frequency $\mathcal{F}_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3175</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2785</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>2544</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2422</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>2106</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>2006</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>1627</td>
<td>12</td>
<td>5</td>
</tr>
</tbody>
</table>
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/
Network Motifs
Concepts for Pattern Frequency
Frequent Pattern Search
Application on Biological Data
Motif Analysis and Visualisation

Motif Detection in Networks
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