

582746 Modelling and Analysis in Bioinformatics

Lecture 2: Network motifs

2.11.2015

Outline

Network motifs

Statistical significance of motifs

Network motifs

- ▶ Network motifs are a way to analyze the local structure of a network:
 - ▶ What kind of local substructures (motifs, graphlets) does the network have
 - ▶ Assessing the statistical significance of these substructures

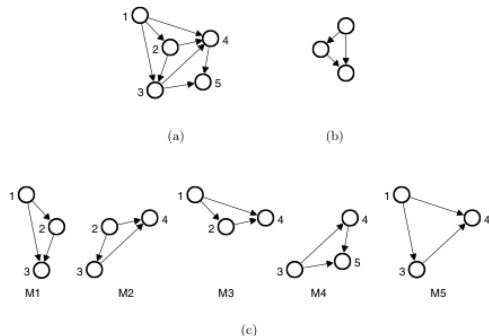


Fig. 2. An example graph (a), a pattern (b) and all different matches of the pattern (c, $M_1 - M_5$). The vertices of the graph and of the matches are numbered consecutively for identification purposes.

What is a motif?

- ▶ A *motif* is a statistically overrepresented pattern of local interactions in the network
- ▶ Overrepresentation = occurring more frequently than expected by chance
- ▶ The rationale is that overrepresentation may denote possible function
 - ▶ The motif has emerged several times
 - ▶ and it has been conserved in the evolution of the network

What is a motif?

- ▶ A *motif* is a small connected subgraph $G' = (V', E')$
- ▶ Size of motif is measured either by the number of vertices or the number of edges

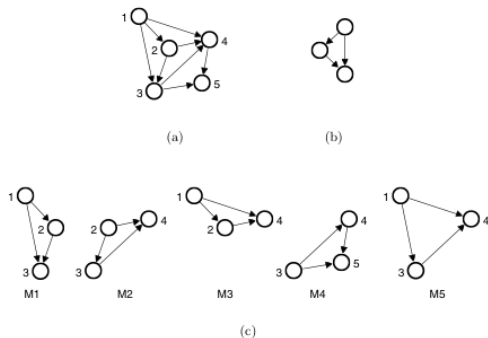


Fig. 2. An example graph (a), a pattern (b) and all different matches of the pattern (c, $M_1 - M_5$). The vertices of the graph and of the matches are numbered consecutively for identification purposes.

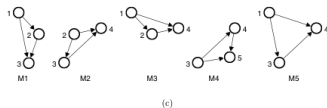
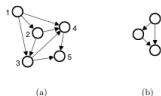
Types of motifs

- ▶ Motifs can be

- ▶ Directed or undirected
- ▶ Cyclic (loopy) or acyclic

matching the type of underlying network to be analyzed, e.g.

- ▶ Protein-protein interactions: undirected
- ▶ Gene regulatory interactions: directed, cyclic



Matching motifs

- ▶ A *match* of a motif G' in the target graph $G = (V, E)$ is a subgraph $G'' = (V'', E'')$ which is *isomorphic* to motif G'
- ▶ Two graphs G' and G'' are isomorphic if there is a bijective mapping between the edge and vertex identities
 - ▶ i.e. G' is transformed to G'' by changing the vertex and edge identities

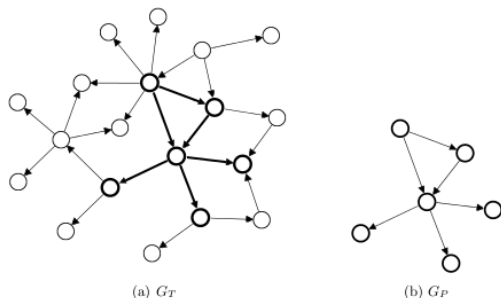
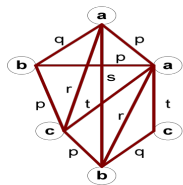


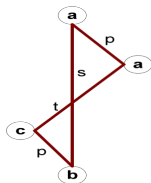
Fig. 1. (a) A graph with a randomly selected subgraph (highlighted with bold lines). This subgraph is isomorphic to the graph G_P shown in (b). The highlighted subgraph in G_T is also a match of G_P in G_T .

Alternative definition: Induced subgraph

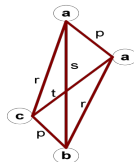
- ▶ Isomorphic induced subgraph (graphlet): a subgraph $G'' = (V'', E'')$ in $G = (V, E)$ is accepted as a match only if it contains all edges of the original graph between the nodes in V'' : mathematically we require that if $e = (n_i, n_j) \in E$ and $n_i, n_j \in V''$ then $e \in E''$
- ▶ Motivation: leaving out interactions from the motif may give false ideas of the biological function



(a) Labeled Graph



(b) Subgraph



(c) Induced Subgraph

Hardness of isomorphism problems

- ▶ The complexity of graph isomorphism is in the 'grey area' of complexity:
 - ▶ It belongs to NP class of problems (problems where solution is easy to verify once found)
 - ▶ It is not known if graph isomorphism belongs to P class of problems (problems that can be solved efficiently)
 - ▶ It is not known if graph isomorphism is NP-complete (problems that are believed to be hard to solve but easy to verify)
- ▶ Subgraph isomorphism, checking if a subgraph G'' that is isomorphic to given graph G' exists in a larger graph G , is known to be NP-complete
- ▶ No hope for really fast algorithms for finding motifs.

Motif frequency

- ▶ How many times a motif occurs in the network to be analyzed?
- ▶ Depends on
 - ▶ Definition of a match (subgraph or induced subgraph)
 - ▶ Counting schemes for matches

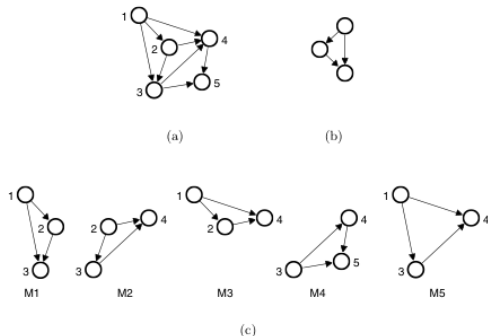


Fig. 2. An example graph (a), a pattern (b) and all different matches of the pattern (c, $M_1 - M_5$). The vertices of the graph and of the matches are numbered consecutively for identification purposes.

Counting schemes

- ▶ Possible schemes for counting matches
 - ▶ \mathcal{F}_1 : Two matches may overlap so that they share vertices and edges
 - ▶ \mathcal{F}_2 : Two matches may overlap so that they share vertices but not edges
 - ▶ \mathcal{F}_3 : Two matches may not overlap, they need to have disjoint sets of vertices

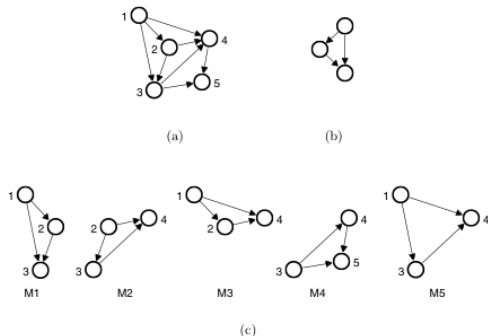


Fig. 2. An example graph (a), a pattern (b) and all different matches of the pattern (c, $M_1 - M_5$). The vertices of the graph and of the matches are numbered consecutively for identification purposes.

Counting schemes

Concept	Graph elements shared by different matches		Values for the example in Fig. 2	
	Vertices	Edges	Frequency	Selected matches
\mathcal{F}_1	yes	yes	5	$\{M_1, M_2, M_3, M_4, M_5\}$
\mathcal{F}_2	yes	no	2	$\{M_1, M_4\}$ or $\{M_3, M_4\}$
\mathcal{F}^*	no	yes	—	—
\mathcal{F}_3	no	no	1	one of $\{M_1, M_2, M_3, M_4, M_5\}$

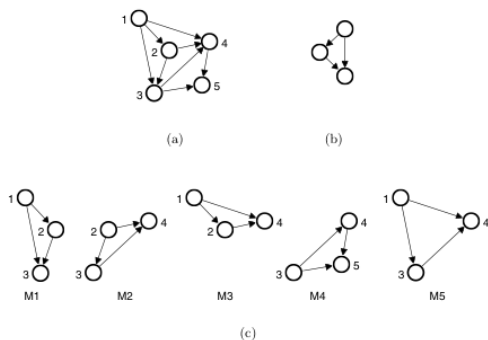


Fig. 2. An example graph (a), a pattern (b) and all different matches of the pattern (c), $M_1 - M_5$). The vertices of the graph and of the matches are numbered consecutively for identification purposes.

Outline

Network motifs

Statistical significance of motifs

Statistical significance of motifs

- ▶ The frequency of a motif in some network does not directly tell us its importance
- ▶ Testing for statistical significance is more informative
 - ▶ How often we would expect to see this motif by chance in a similar random network
- ▶ Need to formulate a *null hypothesis* and check the probability of the motif occurring as frequently under the null hypothesis

Testing for statistical significance

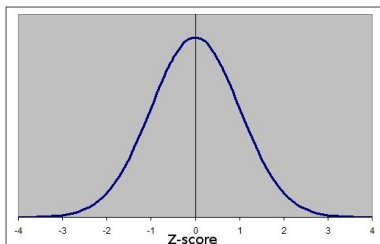
- ▶ For a null hypothesis:
- ▶ Estimate the probability distribution of the frequency of the motif in random networks
 - ▶ Analytically using a network model (e.g. ER networks)
 - ▶ By generating an ensemble of random networks
- ▶ Measure the statistical significance with Z -score or p -value

Measures of motif significance: Z-score

- ▶ Denote by $\mathcal{F}(m)$ the frequency of motif m and by $\overline{\mathcal{F}_r(m)}$ and $\sigma_r(m)$ the average and standard deviation of the motif frequency among the randomized networks.
- ▶ Z-score: "how far above the mean of the random networks"

$$Z(m) = \frac{\mathcal{F}(m) - \overline{\mathcal{F}_r(m)}}{\sigma_r(m)}$$

- ▶ Z-score above 2.0 is generally considered significant ("two standard deviations")

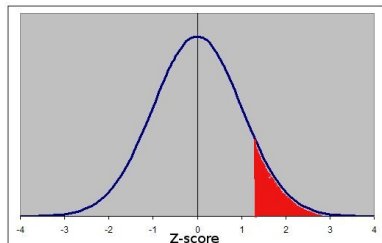


Measures of motif significance: P-value

- ▶ P-value: "how often a random network has more motif occurrences"

$$P(m) = \frac{1}{N} \sum_{r=1}^N \mathbf{1}_{\{\mathcal{F}_r(m) \geq \mathcal{F}(m)\}}$$

- ▶ $\mathbf{1}_{\{A\}}$ denotes the indicator function, $\mathcal{F}_r(m)$ denotes the motif's frequency in r 'th randomized network
- ▶ Requires a large number of randomized networks (≈ 1000) to be accurate
 - ▶ Estimating the tail of the distribution is harder than estimating its mean (as in Z-score)



Analytical approach using ER networks

- ▶ In ER networks an edge is present between two vertices with probability p
- ▶ Here we also allow self loops and also these edges are present with probability p
- ▶ The ER network should have a similar number of vertices and edges as the real network and so

$$p = \frac{E}{N^2}$$

where E is the number of edges in the real network and N is the number of vertices in the real network.

- ▶ Note that a directed network allowing self loops can have at most N^2 edges.

Probability distribution of self loops in ER networks

- ▶ The probability of having exactly k self loops is

$$P(k) = \binom{N}{k} p^k (1-p)^{N-k}$$

- ▶ The probability distribution is thus binomial with mean:

$$\langle N_{\text{self}} \rangle = Np = N \frac{E}{N^2} = \frac{E}{N}$$

- ▶ and with variance (approximation via Poisson distribution)

$$\sigma_{\text{self}} = \sqrt{\frac{E}{N}}$$



Z-score

- ▶ The *E. coli* transcriptional network has 424 vertices and 519 edges (note that this is a different version of the network than what we use in the exercises):

$$\langle N_{\text{self}} \rangle = \frac{E}{N} = \frac{519}{424} = 1.2$$

$$\sigma_{\text{self}} = \sqrt{\frac{E}{N}} = 1.1$$

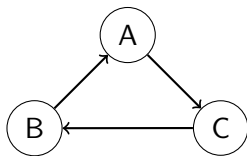
- ▶ The real network has 40 self loops:

$$Z = \frac{N_{\text{self}} - \langle N_{\text{self}} \rangle}{\sigma_{\text{self}}} = \frac{40 - 1.2}{1.1} = 32$$

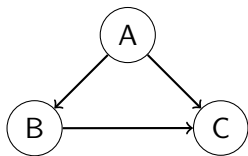
- ▶ Z-score is very high and thus the high number of self loops in the *E. coli* transcriptional network is statistically significant

Subgraphs in ER networks

- ▶ Consider a pattern graph G with n vertices and g edges
- ▶ How often would such a pattern occur in ER networks?
- ▶ We will use counting scheme \mathcal{F}_1 (vertices and edges can overlap)



Motif 1



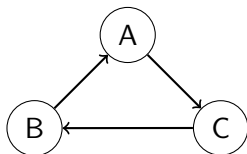
Motif 2

Subgraphs in ER networks

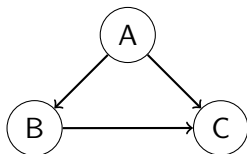
- ▶ To generate an instance of this pattern in a random graph, we need to choose n vertices and place the g edges in appropriate places:

$$\begin{aligned}\langle N_G \rangle &= a^{-1} \cdot N \cdot (N-1) \cdot \dots \cdot (N-n+1) \cdot p^g \\ &\approx a^{-1} N^n p^g\end{aligned}$$

where a is the number of permutations of vertex labels of G that give the same graph.



Motif 1
 $a = 3$



Motif 2
 $a = 1$

Subgraphs in ER networks

- ▶ The mean connectivity of a network is

$$\lambda = \frac{E}{N}$$

- ▶ and then we get

$$\begin{aligned}\langle N_G \rangle &\approx a^{-1} N^n p^g \\ &= a^{-1} N^n \left(\frac{E}{N^2} \right)^g \\ &= a^{-1} \lambda^g N^{n-g}\end{aligned}$$

- ▶ If we assume that the mean connectivity is constant regardless of the size of the network, then the number of subgraphs scales as

$$\langle N_G \rangle \sim N^{n-g}$$

Subgraphs in ER networks

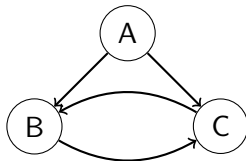
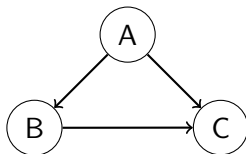
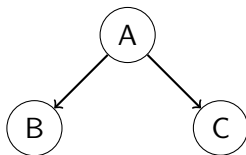
- ▶ V-shaped subgraphs (3 nodes, 2 edges) thus scale linearly with the size of the network:

$$\langle N_{V\text{-shaped}} \rangle \sim N$$

- ▶ Number of triangle shaped subgraphs (3 nodes, 3 edges) stays constant:

$$\langle N_{\text{triangle}} \rangle \sim N^0$$

- ▶ Subgraphs with 3 nodes and more than 3 edges become rarer when the network gets larger



A subgraph in ER networks and *E. coli* transcriptional regulation network

- ▶ The *E. coli* transcriptional regulation network has 424 vertices and 519 edges:

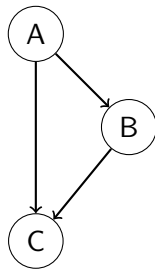
$$\langle N_G \rangle \approx a^{-1} \lambda^g N^{n-g} = \left(\frac{519}{424} \right)^3 424^{3-3} = 1.7$$

- ▶ The distribution of the motif in ER networks can be approximated by a Poisson distribution and thus the standard deviation is

$$\sigma_G \approx \sqrt{\langle N_G \rangle} = 1.3$$

- ▶ The *E. coli* transcriptional regulation network features 42 instances of the motif and so we get:

$$Z = \frac{42 - 1.7}{1.3} = 31$$



Null hypothesis from random networks

- ▶ In traditional hypothesis testing, one typically analytically formulates a probability distribution for the values of the random variable of interest (here frequency of a motif)
- ▶ In network analysis, analytically determining a suitable probability distribution may be difficult
- ▶ Instead, randomization tests are being used: a large set of random networks of appropriate structure are generated and the average frequency of the motif together with its variance is recorded.
- ▶ Computationally demanding process if the networks are large

Randomization algorithm for Null model networks

- ▶ Typical method for null model generation is to take the original network being analyzed and make large number of randomized versions of it by modifying the network by a large number of random edit operations
- ▶ Commonly used edit operation is to rewire the network locally:
 - ▶ Take two edges (A, B) and (C, D) and replace them with edges (A, D) and (C, B)
 - ▶ Preserves degree distribution of nodes
 - ▶ If the nodes are chosen from a small neighborhood, also keeps average path length close to original

Randomization algorithm for Null model networks

Additional criteria to be preserved can be set, e.g.

- ▶ Preserve number of bidirectional edges
- ▶ Preserve number of motif of size $n - 1$ when searching for motifs of size n
- ▶ ...

As a guideline, the null model should be as close to the original as possible, but randomize the property of interest.

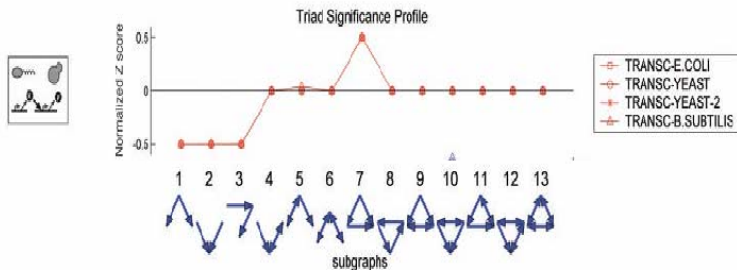
Motif significance profile

- ▶ Motif significance profile SP is a vector of normalized Z-scores for a particular set of motifs









$$SP = ((SP(m_1), \dots, SP(m_2)),$$

where $SP(m) = Z(m_i) / \sqrt{\sum_j Z(m_j)^2}$.

- ▶ Motif significance profile allows comparing different size networks in terms of the motifs they contain
- ▶ Typically, the set of motifs contains all motifs of particular size



Motifs and antimotifs in PPI networks and internet router network

Pattern	Protein interactions	Internet routers
	Not a motif $C = 0.981$	Not a motif $C = 0.977$
	Motif ($Z = 48$) $C = 0.019$	Motif ($Z = 4600$) $C = 0.023$
	Motif ($Z = 15$) $C = 0.680$	Not a motif $C = 0.931$
	Anti-motif ($Z = -19$) $C = 0.024$	Motif ($Z = 18$) $C = 0.013$
	Anti-motif ($Z = -18$) $C = 0.292$	Anti-motif ($Z = -7$) $C = 0.048$
	Not a motif $C = 0.0013$	Motif ($Z = 356$) $C = 0.004$
	Anti-motif ($Z = -4.5$) $C = 0.0019$	Motif ($Z = 137$) $C = 0.002$
	Not a motif $C = 0.0004$	Motif ($Z = ND$) $C = 0.0005$

Hardness of motif discovery

Several challenging subproblems:

- ▶ Graph isomorphism testing: required to check if two motifs are in fact the same. No polynomial time algorithm is known for this problem.
- ▶ Number of motifs: grows exponentially in the size of the motif. Especially with directed motifs grows very fast.
- ▶ Number of matches: theoretically the worst case number of potential matches is $O(|E_t|^{|E_m|})$ where E_t and E_m are the number of edges in the target and motif, respectively.
- ▶ Size of analyzed networks affects the above steps via the number of different patterns and matches that can be found.
- ▶ Calculation of statistical significance via randomization calls for generation and motif discovery from a large number of networks, multiplying the computation time of all the above points.

Study Group on Thursday

- ▶ Group 1: Students whose *first* name starts with A-L
 - ▶ F. Schreiber and H. Schwbbermeyer: Frequency concepts and pattern detection for the analysis of motifs in networks. *Trans. on Comput. Syst. Biol.* III, pp. 89–104, 2005.
 - ▶ Concentrate on section 4.
- ▶ Group 2: Students whose *first* name starts with M-Z
 - ▶ N. Kashtan, S. Itzkovitz, R. Milo and U. Alon: Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* 20(11):1746–1758, 2004.
 - ▶ Concentrate on the Methods section.