Ranking of network elements based on functional substructures

Dirk Koschützki[a*], Henning Schwöbbermeyer[a] and Falk Schreiber[a]

[a] Leibniz Institute of Plant Genetics and Crop Plant Research,
06466 Gatersleben, Germany

koschuet|schwoebb|schreibe@ipk-gatersleben.de

[*] Corresponding Author, Tel: +49 39482 5758, Fax: +49 39482 5500

Abstract

Centrality analysis has been shown to be a valuable method for the structural analysis of biological networks. It is used to identify key elements within networks and to rank network elements such that experiments can be tailored to interesting candidates. Several centrality measures have been studied, in particular for gene regulatory, metabolic and protein interaction networks. However, these centralities have been developed in other fields of science and are not adapted to biological networks. In particular, they ignore functional building blocks within biological networks and therefore do not consider specific network substructures of interest.

We incorporate functional substructures (motifs) into network centrality analysis and present a new approach to rank vertices of networks. A method for motif-based centrality analysis is presented and two extensions are discussed which broaden the idea of motif-based centrality to specific functions of particular motif

elements, and to the consideration of classes of related motifs. The presented method is applied to the gene regulatory network of *E. coli*, where it yields interesting results about key regulators.

Keywords:

- Network centrality
- Network motif
- Gene regulatory network
- Network analysis

1 Introduction

Systems biology is a major focus of current research, and the analysis of biological networks such as metabolic and gene regulatory networks is an important part of this research area (Barabási & Oltvai, 2004). Such networks consist of vertices and edges where the vertices represent, for example, genes, and edges represent functional relations between them. Of particular interest is the ranking or ordering of network elements, a concept that is known as centrality analysis (Koschützki *et al.*, 2005). Examples for the importance of this analysis are the study of lethality in protein interaction networks (Jeong *et al.*, 2001), the identification of enzymes that serve as potential drug targets (Karp *et al.*, 1999), and the examination of metabolic pathways and protein domain networks (Wuchty & Stadler, 2003).

Centrality analysis in biological networks is usually based on methods developed in other fields of science, particular social science. These methods are commonly not tailored to the biological domain. In particular, existing centrality measures consider either only the local or only the global network structure to rank elements. However, to analyse functional interactions between network elements their extended neighbourhood or specific activity patterns have to be incorporated. An example is the analysis of gene regulatory networks. Local approaches such as degree centrality are able to identify key regulators which directly regulate many other genes, but fail to find important regulators which affect many other genes mainly indirectly. Global approaches such as betweenness centrality consider the influence of every network element onto each other independent of local functional modules, and therefore, for example, tend to overemphasise intermediate regulators.

To deal with functional substructures the concept of network motifs has been introduced (Milo *et al.*, 2002). A network motif is a pattern (subnetwork) of local interconnections with particular statistical or functional properties. Several interesting motifs have been detected in complex networks of various fields (Lee *et al.*, 2002; Milo *et al.*, 2002; Shen-Orr *et al.*, 2002), and network motif analysis is particularly important for the study of biological networks such as gene regulatory or protein interaction networks (Conant & Wagner, 2003; Shen-Orr *et al.*, 2002; Wuchty *et al.*, 2003).

We present a novel approach that combines the strength of motif analysis with the concept of centralities. Based on particular functional interactions between network elements the importance of the elements with regard to

the considered interactions is calculated. This leads to, in total, three new centrality measures. The application of motif-based centrality measures to the gene regulatory network of *Escherichia coli* demonstrates their power and shows interesting insights for the study of important regulators. In particular, they facilitate a more specific analysis of their functional properties and identify certain key regulators not detected by previously applied centrality measures.

2 Motif-based Centralities

2.1 Graph, Centrality, and Motif

A directed *graph* $G = (V, E)$ consists of a *set of vertices* (or nodes) V and a *set of edges* (or arcs) $E \subseteq (V \times V)$. By $V(G)$ and $E(G)$ we denote the *vertex set* and the *edge set* of the graph G, respectively. A graph $G' = (V', E')$ is a *subgraph* of the graph $G = (V, E)$ (written as $G' \subseteq G$) if $V' \subseteq V$ and $E' \subseteq E \cap (V' \times V')$. A graph $G_1 = (V_1, E_1)$ is *isomorphic* to a graph $G_2 = (V_2, E_2)$ (written as $G_1 \simeq G_2$) if a bijective mapping $\phi\colon V_1 \mapsto V_2$ exists with $\forall v_a, v_b \in V_1\colon (v_a, v_b) \in E_1 \Leftrightarrow (\phi(v_a), \phi(v_b)) \in E_2$. Such a mapping is called an *isomorphism* and, if $G_1 = G_2$, it is called an *automorphism*.

A *centrality* is a function $c\colon V \mapsto \mathbb{R}$ that assigns every vertex a real number. A vertex v_a is said to be more important (more central) than a vertex v_b if $c(v_a) > c(v_b)$. Based on centrality values vertices can be ordered or ranked. A recent review explains different concepts for centralities and describes more than 20 measures (Koschützki *et al.*, 2005).

[Figure 1 about here.]

Small recurring subgraphs within a given graph are called motifs (Lee *et al.*, 2002; Milo *et al.*, 2002; Shen-Orr *et al.*, 2002). A *motif* M is a directed graph according to the definition of graphs above. A *match* G_M of a motif M in a target graph G is a subgraph of G ($G_M \subseteq G$) which is isomorphic[1] to the motif M ($G_M \simeq M$). See Fig. 1 for a graph, a motif and a match of the motif in the graph. The *motif match set* $\mathcal{G}_M = \{G_M \mid G_M \subseteq G \wedge G_M \simeq M\}$ of a motif M is the set of all matches of M in the graph G. It is a set of subgraphs of G and algorithms exist for its computation (Schreiber & Schwöbbermeyer, 2005; Shen-Orr *et al.*, 2002; Sporns & Kötter, 2004).

2.2 Motif-based centrality

Given a graph G, a motif M and the corresponding motif match set $\mathcal{G}_M$ a new centrality can be defined. The *motif-based centrality* $c_{mc}\colon V \mapsto \mathbb{R}$ assigns to every vertex $v \in V(G)$ the number of matches the vertex v occurs in. It is defined as $c_{mc}(v) := |\{G_M \mid G_M \in \mathcal{G}_M \wedge v \in V(G_M)\}|$. Alg. 1 shows the steps to compute this centrality. The complexity of this algorithm is mainly determined by the function computeMotifMatchSet, which has as underlying decision problem the NP-complete Subgraph Isomorphism problem (Garey & Johnson, 2003). The computation of motif-based centralities is therefore feasible for the same size of graphs and motifs that are currently investigated with existing motif analysis methods.

[1]The definition of subgraph isomorphism used here is sometimes called subgraph monomorphism.

Algorithm 1 Motif-based centrality

Input: Graph G, Motif M
Output: Centrality values $c_{mc}(v)$ for the vertices $v \in V(G)$

```
1: // Initialise result vector
2: for all v ∈ V(G) do
3:     c_mc[v] ← 0
4: // Compute motif match set with existing algorithm, see Sect. 2.1.
5: G_M ← computeMotifMatchSet(G, M)
6: // Compute motif-based centrality values
7: for all G_M ∈ G_M do
8:     for all v ∈ V(G_M) do
9:         c_mc[v] ← c_mc[v] + 1
```

[Figure 2 about here.]

As an example consider the motif *feed-forward loop* (FFL) shown in Fig. 2(b), which matches three times in the target graph shown in Fig. 2(a). Figure 2(d) shows the resulting centrality values for all vertices of this graph. Vertex v_2 is the most important vertex as it participates in all three matches of the motif.

2.3 Role-based motif-based centrality

Vertices of motifs may represent different functions. For example, in the gene regulatory network context considered in this paper, three different functions of the vertices of the FFL motif as shown in Fig. 2(c) can be identified: (1) the vertex at the top is the master regulator, this vertex regulates the other two vertices; (2) the vertex on the right side is the intermediate regulator, it is regulated by the master regulator and itself regulates together with the master regulator the vertex at the bottom; and (3) the vertex at the bottom of the drawing is regulated by both other vertices and is therefore called the regulated vertex. Such different

functions of vertices within motifs are called *roles* and three roles can be assigned to the vertices of the FFL motif (Kashtan *et al.*, 2004).

Let R be a set of roles, G be a graph, M a motif and $\mathcal{G}_M$ the corresponding motif match set. We define a function $role\colon V \times \mathcal{G}_M \mapsto R$ which assigns a role to every vertex of G under a specific match. The *role-based motif-based centrality* (short *extended motif-based centrality*) $c_{emc}(v, r)\colon (V \times R) \mapsto \mathbb{R}$ assigns to every vertex $v \in V(G)$ the number of matches the vertex v occurs in and where it has the role r. It is defined as $c_{emc}(v, r) := |\{G_M \mid G_M \in \mathcal{G}_M \wedge v \in V(G_M) \wedge role(v, G_M) = r\}|$. Considering a particular role r, the function c_{emc} is a centrality on the vertices of G.

The algorithm for the extended motif-based centrality is shown in Alg. 2. The function getRoleOfMatchingVertex returns the role of the vertex v within the motif M based on the match G_M. The result of the extended algorithm is not a single centrality vector for the vertices, as in Alg. 1, but a matrix consisting of rows and columns where the rows denote the vertices of the graph, the columns denote the roles and the entries are the centrality values. The complexity of this algorithm is in the same class as Alg. 1 as again the underlying decision problem Subgraph Isomorphism is the most complex part of the algorithm.

Figure 2(e) shows the result of the extended algorithm for the FFL motif with roles shown in Fig. 2(c) in the same graph as before, see Fig. 2(a). The column named Role A contains the number of matches for the vertices v_1 to v_5 for the role master regulator. Vertex v_2 is the most important vertex according to this role, followed by vertex v_1. A comparison with Figure 2(d) shows that, based on the FFL motif, the vertex v_1 is a more

Algorithm 2 Extended motif-based centrality

Input: Graph G, Motif M with roles R

Output: Centralities $c_{emc}(v, r)$ for the vertices $v \in V(G)$ and roles $r \in R$

```
1: // Initialise result table
2: for all v ∈ V do
3:     for all r ∈ R do
4:         c_emc[v, r] ← 0
5: // Compute motif match set with existing algorithm, see Sect. 2.1.
6: G_M ← computeMotifMatchSet(G, M)
7: // Compute extended motif-based centralities
8: for all G_M ∈ G_M do
9:     for all v ∈ V(G_M) do
10:        r ← getRoleOfMatchingVertex(v, G_M)
11:        c_emc[v, r] ← c_emc[v, r] + 1
```

important master regulator than the vertices v_3 to v_5. This is not obvious from the ranking based on the motif-based centralities without roles.

[Figure 3 about here.]

There may exist several isomorphisms between a motif and a specific match, see for example Fig. 3. As a result restrictions for the assignment of roles to vertices are necessary. If an automorphism ψ in the motif M exists with $\psi(v_a) = v_b$ for any two vertices $v_a, v_b \in V(M)$, then the roles of v_a and v_b have to be identical. In the motif shown in Fig. 3(a) there exists such an automorphism which maps the vertices B and C onto each other. The use of three different roles is therefore not allowed for this motif and identical roles have to be assigned to the two vertices at the bottom of the drawing. A definition of the allowed roles based on automorphic equivalence is given in the appendix of (Kashtan *et al.*, 2004).

2.4 Role-based motif-based centrality for motif classes

Using the previously introduced concepts we can extend the method further. By assigning the same role to similar vertices of a group of similar motifs we can establish a centrality based on a class (or a group) of motifs. Consider, for example, a group of chains (see Fig. 4(a)), where all vertices at the start of such chains have a similar characteristic (no incoming edges) and all vertices at the end have another similar characteristic (no outgoing edges). For gene regulatory networks several motif classes are known. For example, the regulatory chain motif class, as in the example above, consists of a set of chains of three or more regulators in which one regulator regulates another regulator, which in turn regulates a third one and so forth (Lee *et al.*, 2002). In the motif class single input motif (SIM) a set of vertices is exclusively regulated by a single vertex (Shen-Orr *et al.*, 2002). Formally motif classes can be described by graph grammars (Engelfriet & Rozenberg, 1997). The motif generalisation described in (Kashtan *et al.*, 2004) is not able to cover the full strength of our approach, as it cannot, for example, define a group of chains as a generalisation of a motif.

The *role-based motif-based centrality for motif classes* (short: *motif-class centrality*) is computed by using Alg. 2 for the extended motif-based centrality. For each member of the motif class of interest the matrix containing centrality values for vertices and roles is computed. If the motif class consists of l different motifs, then l matrices $c^1_{emc}, \ldots, c^l_{emc}$ with centrality values are obtained. The centrality value for a vertex $v \in V(G)$ and a role $r \in R$ for a motif class is defined as the sum of the centrality values over all l matrices, $c_{mcc}[v, r] := \sum_{i=1}^{l} c^i_{emc}[v, r]$.

[Figure 4 about here.]

We demonstrate this method on the example graph shown in Fig. 4(b) and the chain motif class (see Fig. 4(a)) where we are interested in the centrality value for vertices at the top of chains (role A in Fig. 4(a)). The size of the chain motif class is given by the number of chains of different lengths in the target graph. The centrality values for the vertices of the example graph in Fig. 4(b) are given in Fig. 4(c). The vertex v_1 receives the highest centrality value, as it is the only vertex from which all other vertices are reachable.

3 Centralities for the gene regulatory network of *E. coli*

3.1 Gene regulatory network of *E. coli*, global regulators and feed-forward loop motif

In the following we analyse centralities within the gene regulatory network (GRN) of *Escherichia coli*. The network is based on the data of transcriptional regulatory interactions of genes from RegulonDB, Version 5.0 (Salgado *et al.*, 2006). Genes are represented by vertices and transcriptional regulatory interactions between genes are modelled as edges, a common approach to model GRNs. The interactions between genes represent transcriptional control of transcription factors on the transcription of regulated genes. There are a few cases where transcription factors are formed by subunits of different gene products. They are here replaced by a common identifier which corresponds to the transcription

factor, e.g. *ihfA* or *ihfB* result in *ihfAB*. The regulatory interactions of such different subunits are assigned to this new identifier, and parallel edges which occurred due to the previous operation are replaced by a single edge. The resulting network consists of 1250 vertices and 2515 edges, of which 84 edges are self-loops ($e = (v_i, v_i)$) representing autoregulation, i.e., the transcriptional control of a gene by its own gene product. It should be noted that autoregulation (self-loops) can be part of a motif.

In gene regulatory networks genes at a high level within the hierarchy of regulatory control are of particular interest due to their far reaching influence on other genes within the network. These genes are commonly called *global regulators*. Some criteria for the characterisation of global regulators have been proposed, such as the number of regulated genes, the number and type of coregulators, the number of other regulators they control, the size of their evolutionary family, and the variety of conditions where they exert their control (Martínez-Antonio & Collado-Vides, 2003).

For the motif-based centrality analysis we use the feed-forward loop (FFL) motif as described in Sect. 2.2. It has been shown that depending on the type of interactions (activating or repressing) the FFL motif acts as an accelerator or delay element in the process of gene expression and therefore has particular properties that control the expression of target genes (Mangan & Alon, 2003; Mangan *et al.*, 2003).

3.2 Motif-based centrality for the *E. coli* GRN

[Table 1 about here.]

The motif-based centrality c_{mc} based on the FFL motif is computed for the GRN of *E. coli* and the resulting centrality values are used to rank the genes, see Table 1. The genes at the top position of this ranking have important functions in the regulation of cellular processes according to *EcoCyc* (Keseler *et al.*, 2005): *crp* is the major global regulator of catabolite-sensitive operons which monitors the energy status of cells by cAMP concentration and is capable of regulating the expression of more than 200 genes, *fnr* and *narL* are transcriptional regulators for fermentation and anaerobic respiration, *arcA* is a transcriptional regulator of aerobic respiration control, and *fis* and *ihfAB* are involved in the process of DNA replication. These genes have also been characterised as global regulators (Martínez-Antonio & Collado-Vides, 2003).

The results of the motif-based centrality c_{mc} based on the FFL motif are consistent with current biological knowledge as the genes which have been characterised as global regulators are assigned to top positions in the ranking. However, by consideration of the functional roles that genes adopt within the FFL motif, an extended motif-based centrality (studied in the following section) allows a further differentiation of the genes not covered by other centrality concepts.

3.3 Extended motif-based centrality for the *E. coli* GRN

[Table 2 about here.]

The computation of the extended motif-based centrality c_{emc} based on the FFL motif with roles leads to three different rankings of the genes depending on the role under consideration, see Tables 2 (a+b). The genes

at top positions of these rankings allow an identification of important global regulators which receive a high rank for the master regulator role (role A in Table 2(b)), important local regulators which receive a high rank for the intermediate regulator role (role B in Table 2(b)) and important target genes which are controlled by at least two regulators as part of a functional feed-forward loop motif (role C in Table 2(b)).

The genes at the top positions of these rankings can be clustered into into four different groups:

1. genes that nearly exclusively adopt role A and therefore mainly act as global regulators without being controlled by many other genes (*crp*, *ihfAB*, *soxS*). The gene *crp* is the major global regulator of catabolite-sensitive operons (Busby & Ebright, 1999), *ihfAB* is involved in a wide variety of processes including DNA replication, site-specific recombination and transcription (Goosen & van de Putte, 1995), *soxS* controls cellular response to oxidative stress and is controlled by *soxR*, which in turn is regulated directly by interactions with oxidising ligands (Browning & Busby, 2004).

2. genes where both roles A and B are important and which selectively act as global and as local regulators (*fnr*, *arcA*, *fis*, *hns*). The gene *fnr* is the highest ranking regulator of anaerobic gene expression, as it exerts its control directly and indirectly via *arcA*. The gene *arcA* affects anaerobic gene expression and in turn regulates *fnr* (Levanon *et al.*, 2005). *fis* is a DNA-bending protein required for chromosome replication and for activation of stable RNA operons. It is involved in coupling the cellular physiology of the topology of the bacterial

chromosome (Stavans & Oppenheim, 2006). *fis* is regulated by *crp* and *ihfAB*.

3. genes that nearly exclusively adopt role B and therefore mainly act as local regulators, which are controlled by other genes (*narL*, *fur*, *hyfR*, *gadX*). For example, the gene *narL* is regulated by *fnr* and controls the expression of several genes involved in anaerobic respiration and fermentation. Here, *narL* regulates the utilisation of either nitrate or nitrite as electron acceptors, which are the preferred anaerobic electron acceptors (Darwin *et al.*, 1998). *fur* is involved in the regulation of a large number of operons that encode enzymes involved in iron transport and is regulated amongst others by *crp* and *soxS* (Keseler *et al.*, 2005).

4. genes that nearly exclusively adopt role C and which are therefore regulated genes (*marB*, *gadA*, *sodA*, see Table 2(b)). As these genes are mainly targets of regulation, they are not of interest in the analysis of regulators.

Consideration of roles of vertices allows a more detailed analysis of the functional properties of the elements within the network.

There are many cases where one pair of global and local regulators regulate a large number of different genes. For example, *crp* in role A and *fis* in role B regulate 38 different genes. However, an interesting observation is, that most genes are regulated by different pairs of global and local regulators. There are only two cases where one global regulator (*crp*) and three different local regulators (*fis*, *galS*, *flhD*) regulate the same gene (*mglA*, *mglC*).

3.4 Motif-class centrality for the *E. coli* GRN

Motifs which share similar structural properties can be combined into motif classes, see Sect. 2.4. Here we analyse regulatory chains modelled by the chain motif class (see Fig. 4(a)). Starting from a regulator at the top of the chain, the regulation is directed to its bottom via intermediate regulators. Generally, due to the hierarchical structure of gene regulatory networks regulators increase their range of control by indirectly affecting genes via intermediate regulators (Martínez-Antonio & Collado-Vides, 2003). Regulators at the top of regulatory chains are of particular interest as they start regulatory cascades.

[Table 3 about here.]

To study the regulators with the highest influence on other genes within the *E. coli* GRN the motif-class centrality for the chain motif class is computed and only the role of the vertex at the top of the chain is considered (role A in Fig. 4(a)). Comparing the ranking for the motif-class centrality based on the chain class and the motif-based centrality based on the FFL motif (see Table 3 and Table 1 respectively) shows that *crp*, *fnr*, *arcA*, *fis* and *ihfAB* are in both cases among the top six positions. The gene *narL* ranked at position 6 for the FFL-based motif-based centrality holds only position 20 for the motif-class centrality, even though it regulates a high number of genes directly it influences only a low number of genes in total by indirect regulation.

Furthermore, the composition of centrality values for individual motif chains shows some interesting characteristics. There are some genes among

the top-20 that have a very low centrality value for motif chains of size 2: *evgA* (centrality value of 4 for chains of length 2 compared to 325 for centrality c_{mcc}), *ydeO* (1 vs. 322), *soxR* (2 vs. 213), *gadW* (4 vs. 185), *cspE* (1 vs. 184) and *cspA* (2 vs. 183). Therefore, these genes have a low range of direct control. However, all these genes indirectly control a large number of other genes. For example, the genes *evgA* and *ydeO* both regulate *gadE* and are part of the same regulatory chains via *gadE*. The gene *evgA* additionally regulates three other genes which themself do not regulate any genes. These three genes (*evgA*, *ydeO*, *gadE*) build a complex regulatory circuit controlling the glutamate-dependent acid resistance system, which is remarkable in its efficiency and regulatory complexity (Ma *et al.*, 2004). Furthermore, the *evgA*-*ydeO*-*gadE* regulatory circuit also appears to have a broad impact on other genes and affects cell physiology beyond acid resistance (Ma *et al.*, 2004). Another example of genes with low direct and high indirect control are the genes *cspE* and *cspA*, which are RNA chaperones and members of the cold-shock protein family which promote cellular adaptation to low temperature (Bae *et al.*, 2000). They facilitate translation at low temperature by destabilising mRNA structures and in this way act as transcription antiterminators. Furthermore, they participate in many other aspects of gene regulation, e.g. *cspE* is involved in the regulation of the complex stress response network of the cell (Phadtare & Inouye, 2001). Also, as mentioned before, *soxR* regulates via *soxS* cellular response to oxidative stress.

These results show that the motif-class centrality with the chain class as motif family identifies genes that are important regulators within the GRN of *E. coli* which would be missed when only considering local approaches or other global approaches.

4 Discussion

Centrality analysis has been used to study different biological networks (Jeong *et al.*, 2001; Karp *et al.*, 1999; Potapov *et al.*, 2005; Wuchty & Stadler, 2003), and two centralities have already been applied for the analysis of GRNs: out-degree (c_{odeg}), that is the number of direct successors of a vertex, and shortest-path betweenness (c_{spb}), that is on how many shortest-paths the vertex of interest lies. In contrast to our study these centralities were applied to mammalian GRNs (Potapov *et al.*, 2005) which are not publicly available. A general comparison of rankings computed with these two centralities and our methods is not feasible as motif-based centralities do not give one single ranking. Instead the ranking depends on a particular motif, motif class and role. However, a comparison with specific motif-based centralities such as the motif-based centrality for the FFL motif with roles under consideration of role A (c_{emc}) and the motif-class centrality for chains under consideration of role A (c_{mcc}) is possible, see Tables 4 and 5.

[Table 4 about here.]

[Table 5 about here.]

Table 4 shows the pairwise correlation coefficients for the four centralities introduced above. The high correlation coefficient between c_{mcc} and c_{odeg} can be easily explained: 1101 out of 1250 vertices have an out-degree of zero (88.08%). Therefore, to all these vertices both centralities assign a value of 0. To obtain a sound interpretation for the group of regulatory genes we consider only the correlation of the vertices with non-zero

out-degree in Table 4(b). This proves quite different rankings between c_{odeg} and c_{mcc}.

Table 5 shows the top ranked genes for the considered centralities. In the following we concentrate on c_{mcc} as it gives especially interesting results. For the *E. coli* GRN the rankings given by c_{odeg} and c_{mcc} are very similar for the top 5 positions. However, c_{odeg} only identifies key regulators which directly regulate a large number of genes whereas c_{mcc} is able to identify important players for indirect regulation as well. The rankings given by c_{spb} and c_{mcc} are very different. c_{spb} finds some of the key regulators which are highly ranked by both c_{odeg} and c_{mcc} (*hns*, *fur*, *gadE*, *fis*, *fnr*, *narL*, *arcA*). It also identifies some genes important for indirect regulation such as *gadX*, *soxS* and *cspA*, which obtain only a low ranking based on c_{odeg}, but a high ranking based on c_{mcc}. However, the gene *crp*, which is commonly regarded as the most important global regulator and which was ranked at the top position for c_{odeg} and c_{mcc}, is not under the top 20 positions for c_{spb}. For *ihfAB* which has been characterised as a global regulator in previous studies (see Section 3.2) the same holds true: it receives position 2 and 3 for c_{mcc} and c_{odeg}, respectively, but it is not under the top 20 ranked by c_{spb}. These results are not surprising as c_{spb} assigns high centrality values to vertices that participate in many shortest-path communications. Since the gene regulatory network of *E. coli* has a hierarchical structure with global regulators on top (Martínez-Antonio & Collado-Vides, 2003), these regulators do not necessarily participate in shortest-path communications, but instead start the regulation and therefore are not captured by c_{spb}.

5 Conclusions

We presented a novel approach to rank vertices of networks based on network motifs, and discussed three particular methods. The first method (motif-based centrality) ranks vertices according to the number of motif matches such that the match contains the vertex of interest. The other two methods (extended motif-based centrality and motif-class centrality) are based on this method. The former additionally considers roles, and allows a more detailed analysis of the network of interest based on functions assigned to the vertices of the motif. The latter uses a whole group of similar motifs and therefore takes related functional network substructures into consideration. In contrast to existing centrality measures which consider either the local or the global network structure, the approach presented here deals with structural information *between* local and global information.

These methods can be applied to all kinds of networks by choosing appropriate motifs or motif classes. Here we applied them to the gene regulatory network of *E. coli*, where they yield interesting results in studying different functions of genes (by using the FFL motif) and in identifying key regulators which directly or indirectly regulate many genes (by using the chain motif-class).

In conclusion, motif-based centrality is more effective in identifying important elements in biological networks (such as key regulators in GRNs) than previously used centrality measures. By using appropriate motifs or motif classes it can be tailored to specific analysis tasks.

Acknowledgement

We would like to thank the reviewers for their helpful comments.

This work was supported by the German Ministry of Education and Research (BMBF) under grant 0312706A.

References

Bae, W., Xia, B., Inouye, M. & Severinov, K. (2000). *Escherichia coli* CspA-family RNA chaperones are transcription antiterminators. *Proceedings of the National Academy of Sciences,* **97** (14), 7784–7789.

Barabási, A.-L. & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics,* **5** (2), 101–113.

Browning, D. F. & Busby, S. J. (2004). The regulation of bacterial transcription initiation. *Nature Reviews Microbiology,* **2** (1), 57–65.

Busby, S. & Ebright, R. H. (1999). Transcription activation by catabolite activator protein (CAP). *Journal of Molecular Biology,* **293** (2), 199–213.

Conant, G. C. & Wagner, A. (2003). Convergent evolution of gene circuits. *Nature Genetics,* **34** (3), 264–266.

Darwin, A. J., Ziegelhoffer, E. C., Kiley, P. J. & Stewart, V. (1998). Fnr, NarP, and NarL regulation of Escherichia coli K-12 napF (periplasmic nitrate reductase) operon transcription in vitro. *Journal of Bacteriology,* **180** (16), 4192–4198.

Engelfriet, J. & Rozenberg, G. (1997). Node replacement graph grammers. In *Handbook of Graph Grammars and Computing by Graph Transformation*, (Rozenberg, G., ed.), vol. 1, chapter 1, pp. 1–94. World Scientific.

Garey, M. R. & Johnson, D. S. (2003). *Computers and Intractability.* W. H. Freeman and Company, New York.

Goosen, N. & van de Putte, P. (1995). The regulation of transcription initiation by integration host factor. *Molecular Microbiology,* **16** (1), 1–7.

Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature,* **411**, 41–42.

Karp, P. D., Krummenacker, M., Paley, S. & Wagg, J. (1999). Integrated pathway-genome databases and their role in drug discovery. *Trends in Biotechnology,* **17** (7), 275–281.

Kashtan, N., Itzkovitz, S., Milo, R. & Alon, U. (2004). Topological generalizations of network motifs. *Physical Review E,* **70** (3), 031909.

Keseler, I. M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I. T., Peralta-Gil, M. & Karp, P. D. (2005). EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Research,* **33**, D334–337.

Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D. & Zlotowski, O. (2005). *Network Analysis: Methodological Foundations* vol. 3418 of *LNCS Tutorial*, chapter Centrality Indices, pp. 16–61. Berlin: Springer.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K. & Young, R. A. (2002). Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. *Science*, **298** (5594), 799–804.

Levanon, S. S., San, K.-Y. & Bennett, G. N. (2005). Effect of oxygen on the Escherichia coli ArcA and FNR regulation systems and metabolic responses. *Biotechnology and Bioengineering*, **89** (5), 556–564.

Ma, Z., Masuda, N. & Foster, J. W. (2004). Characterization of EvgAS-YdeO-GadE Branched Regulatory Circuit Governing Glutamate-Dependent Acid Resistance in *Escherichia coli*. *Journal of Bacteriology*, **186** (21), 7378–7389.

Mangan, S. & Alon, U. (2003). Structure and Function of the Feed-Forward Loop Network Motif. *Proceedings of the National Academy of Sciences*, **100** (21), 11980–11985.

Mangan, S., Zaslaver, A. & Alon, U. (2003). The Coherent Feedforward Loop Serves as a Sign-sensitive Delay Element in Transcription Networks. *Journal of Molecular Biology*, **334** (2), 197–204.

Martínez-Antonio, A. & Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology*, **6** (5), 482–489.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks. *Science*, **298** (5594), 824–827.

Phadtare, S. & Inouye, M. (2001). Role of CspC and CspE in Regulation of Expression of RpoS and UspA, the Stress Response Proteins in *Escherichia coli*. *Journal of Bacteriology*, **183** (4), 1205–1214.

Potapov, A. P., Voss, N., Sasse, N. & Wingender, E. (2005). Topology of Mammalian Transcription Networks. *Genome Informatics*, **16** (2), 270–278.

Salgado, H., Gama-Castro, S., Peralta-Gil, M., Díaz-Peredo, E., Sánchez-Solano, F., Santos-Zavaleta, A., Martínez-Flores, I., Jiménez-Jacinto, V., Bonavides-Martínez, C., Segura-Salazar, J., Martínez-Antonio, A. & Collado-Vides, J. (2006). RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Research*, **34** (Suppl. 1), D394–397.

Schreiber, F. & Schwöbbermeyer, H. (2005). Frequency Concepts and Pattern Detection for the Analysis of Motifs in Networks. *Transactions on Computational Systems Biology*, **3** (LNBI 3737), 89–104.

Shen-Orr, S., Milo, R., Mangan, S. & Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, **31** (1), 64–68.

Sporns, O. & Kötter, R. (2004). Motifs in brain networks. *PLoS Biology*, **2** (11), e369.

Stavans, J. & Oppenheim, A. (2006). DNA-protein interactions and bacterial chromosome architecture. *Physical Biology*, **3** (4), R1–R10.

Wuchty, S., Oltvai, Z. N. & Barabási, A.-L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, **35** (2), 176–179.

Wuchty, S. & Stadler, P. F. (2003). Centers of complex networks. *Journal of Theoretical Biology*, **223**, 45–53.

List of Figures

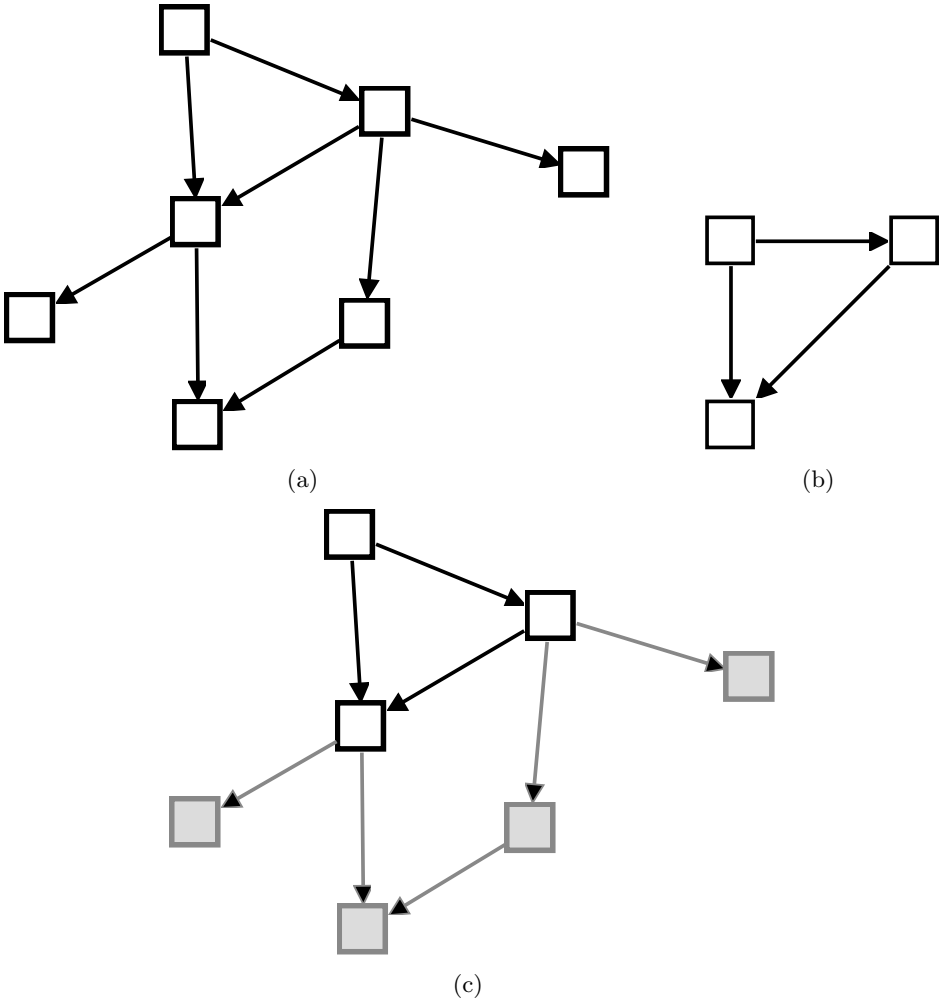
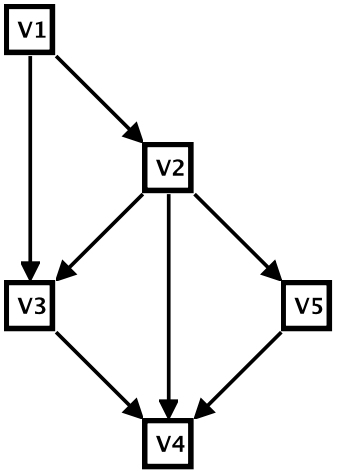
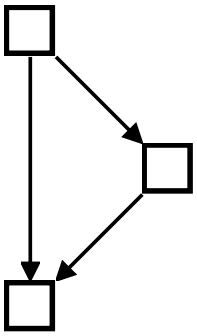


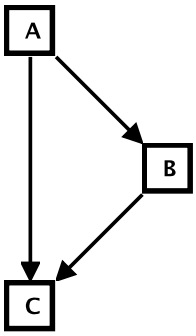
Figure 1: (a) a graph G, (b) a motif M and (c) a match G_M of M in G. The motif (b) occurs once in the graph (a). This occurrence is called a *match* and vertices and edges not participating in the match are marked in grey (c). Basically, motif-based centralities count matches of motifs in graphs.



(a)



(b)



(c)

Vertex	Centrality c_{mc}
v_1	1
v_2	3
v_3	2
v_4	2
v_5	1

(d)

	Centrality c_{emc}		
Vertex	Role A	Role B	Role C
v_1	1	0	0
v_2	2	1	0
v_3	0	1	1
v_4	0	0	2
v_5	0	1	0

(e)

Figure 2: (a) a target graph, (b) the feed-forward loop (FFL) motif, (c) the feed-forward loop motif with three different roles A, B and C. The tables (d) and (e) show the result of the centrality computation for the graph in (a): (d) the motif-based centrality given by the FFL motif without roles, and (e) the extended motif-based centrality given by the FFL motif with roles.

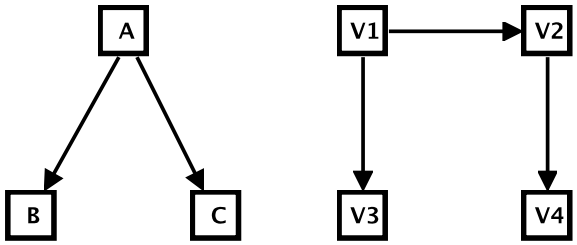
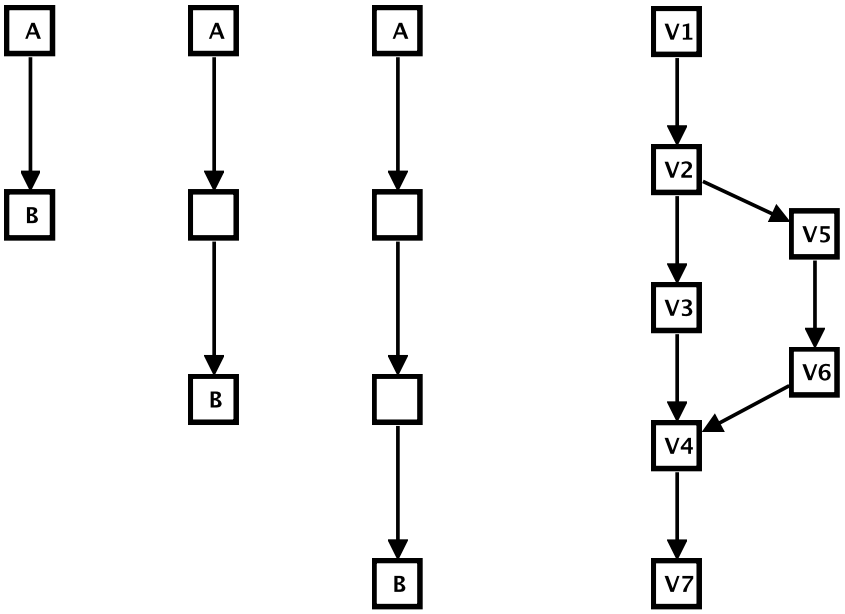


Figure 3: (a) a motif and (b) a target graph. The use of two different roles (B and C) for the vertices at the bottom of the motif is not allowed. In this example two matches between the motif in Fig. 3(a) and the subgraph of the target graph in Fig. 3(b) given by the vertices v_1, v_2 and v_3 exist. In the first match $\phi_1(A) = v_1$, $\phi_1(B) = v_2$ and $\phi_1(C) = v_3$, and in the second match $\phi_2(A) = v_1$, $\phi_2(B) = v_3$ and $\phi_2(C) = v_2$ holds. Obviously, it is not clear which role should be assigned to the vertices v_2 and v_3 as B and C are both candidates. Therefore restrictions for the assignment of roles to vertices are necessary.



(a) (b)

		Centrality Length of chains				
Vertex	c_{mcc}	2	3	4	5	6
v_1	8	1	2	2	2	1
v_2	7	2	2	2	1	0
v_3	2	1	1	0	0	0
v_4	1	1	0	0	0	0
v_5	3	1	1	1	0	0
v_6	2	1	1	0	0	0
v_7	0	0	0	0	0	0

(c)

Figure 4: (a) a sketch of the chain motif-class with roles A and B, (b) an example graph, and (c) the centrality values for the motif-class centrality for the motif-class chains. The length of the chain is defined as the number of vertices in the chain. The c_{mcc} values are the centrality values for role A for the chain motif-class, that is all different chains are considered and c_{mcc} is the sum of the centrality values of the different chains.

List of Tables

Rank	Gene	Centrality c_{mc}
1	crp	254
2	fnr	203
3	arcA	111
4	fis	110
5	narL	100
6	ihfAB	61
7	hns	53
8	fur	43
9	gadX	34
10	hyfR	33
11	marA	29
12	flhD	21
13	nagC, soxS	19
14	modE, tdcA, yiaJ	18
15	gutM, ompR, srlR	17

Table 1: Top-20 out of 1250 genes of the *E. coli* GRN according to the motif-based centrality using the FFL motif with three vertices (see Figure 2(b)).

(a)

	Centrality			
	c_{mc}	c_{emc}		
Gene		A	B	C
crp	254	254	0	0
fnr	203	150	53	0
arcA	111	58	53	0
fis	110	40	70	0
narL	100	5	95	0
ihfAB	61	61	0	0
hns	53	14	39	0
fur	43	6	36	1
gadX	34	8	26	0
hyfR	33	0	33	0
marA	29	1	25	3
flhD	21	0	17	4
soxS	19	18	1	0
nagC	19	5	14	0
modE	18	18	0	0
tdcA, yiaJ	18	0	18	0
ompR	17	5	12	0
gutM, srlR	17	5	11	1

(b)

Role A	Role B	Role C
crp (254)	narL (95)	marB (8)
fnr (150)	fis (70)	gadA (6)
ihfAB (61)	arcA, fnr (53)	fumB, gadC, gadB,
arcA (58)	hns (39)	lpdA, sodA (5)
fis (40)	fur (36)	aceE, aceF, flhC and
modE, soxS (18)	hyfR (33)	27 further genes (4)
hns (14)	gadX (26)	
cpxR, fhlA, gadE (11)	marA (25)	
	tdcA, yiaJ (18)	

Table 2: (a) Top-20 out of 1250 genes (see also Table 1) centrality values for different roles computed with the extended motif-based centrality using the FFL motif see (Figure 2(c)) and (b) top ranking genes of the *E. coli* GRN shown for each role (numbers in brackets are the centrality values).

		Centrality					
		Length of chain					
Gene	c_{mcc}	2	3	4	5	6	7
crp	1592	359	525	436	212	60	0
ihfAB	667	186	215	156	82	28	0
fnr	470	206	237	27	0	0	0
arcA	470	111	215	127	17	0	0
fis	387	156	121	82	28	0	0
evgA	325	4	27	90	125	51	28
ydeO	322	1	27	90	125	51	28
gadE	321	27	90	125	51	28	0
soxR	213	2	24	92	91	4	0
soxS	211	24	92	91	4	0	0
torR	191	10	15	87	51	28	0
gadW	185	4	15	87	51	28	0
cspE	184	1	2	88	65	28	0
cspA	183	2	88	65	28	0	0
gadX	181	15	87	51	28	0	0
hns	181	88	65	28	0	0	0
oxyR	166	15	73	74	4	0	0
fur	151	73	74	4	0	0	0
modE	141	32	94	15	0	0	0
narL	109	94	15	0	0	0	0

Table 3: Top-20 out of 1250 genes of the *E.coli* GRN according to the motif-class centrality based on the chain motif class for the role A (see Fig. 4). There are no chains with a length greater than 7. The number of chains of length 2 gives the number of genes that are directly regulated, excluding autoregulation.

(a)

	c_{mcc}	c_{emc}	c_{odeg}	c_{spb}
c_{mcc}	1.00	0.18	0.98	0.49
c_{emc}	0.18	1.00	0.17	0.22
c_{odeg}	0.98	0.17	1.00	0.49
c_{spb}	0.49	0.22	0.49	1.00

(b)

	c_{mcc}	c_{emc}	c_{odeg}	c_{spb}
c_{mcc}	1.00	0.40	0.72	0.29
c_{emc}	0.40	1.00	0.47	0.36
c_{odeg}	0.72	0.47	1.00	0.32
c_{spb}	0.29	0.36	0.32	1.00

Table 4: Kendall's correlation coefficient for the centrality values of (a) all vertices and (b) vertices with out-degree greater than 0. Abbreviation: c_{mcc}: motif-class centrality for chains under consideration of role A, c_{emc}: motif-based centrality for the FFL motif with roles under consideration of role A, c_{odeg}: out-degree, c_{spb}: shortest-path betweenness.

Order	c_{mcc}	c_{emc}	c_{odeg}	c_{spb}
1	crp (1592)	crp (254)	crp (360)	hns (1039.83)
2	ihfAB (667)	fnr (203)	fnr (207)	gadX (552)
3	arcA (470)	arcA (111)	ihfAB (187)	flhD (535)
4	fnr (470)	fis (110)	fis (157)	fur (488)
5	fis (387)	narL (100)	arcA (111)	gadE (418)
6	evgA (325)	ihfAB (61)	narL (94)	fis (342.5)
7	ydeO (322)	hns (53)	hns (89)	lrp (205.5)
8	gadE (321)	fur (43)	fur (74)	rcsAB (204)
9	soxR (213)	gadX (34)	lrp (63)	soxS (165)
10	soxS (211)	hyfR (33)	glnG (44)	fnr (159)
11	torR (191)	marA (29)	narP (40)	cspA (141)
12	gadW (185)	flhD (21)	cpxR (35)	caiF (135)
13	cspE (184)	nagC (19)	phoB (35)	purR (132)
14	cspA (183)	soxS (19)	fruR (32)	narL (99)
15	gadX (181)	modE (18)	modE (32)	marA (70.33)
16	hns (181)	tdcA (18)	fhlA (29)	metJ (61)
17	oxyR (166)	yiaJ (18)	lexA (29)	malT (51)
18	fur (151)	gutM (17)	flhD (28)	arcA (50.5)
19	modE (141)	ompR (17)	gadE (28)	glnG (50)
20	narL (109)	srlR (17)	purR (28)	ompR (23)

Table 5: Top-20 genes of the *E. coli* GRN according to c_{mcc}, c_{emc}, c_{odeg} and c_{spb}. Numbers in brackets are the centrality values. Abbreviations, see Tab. 4. If genes have the same centrality value under a specific centrality they are ordered alphabetically.