We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

16

Centralities Based Analysis of Complex Networks

Giovanni Scardoni and Carlo Laudanna
Center for BioMedical Computing (CBMC), University of Verona
Italy

1. Introduction

Characterizing, describing, and extracting information from a network is by now one of the main goals of science, since the study of network currently draws the attention of several fields of research, as biology, economics, social science, computer science and so on. The main goal is to analyze networks in order to extract their emergent properties (Bhalla & Iyengar (1999)) and to understand functionality of such complex systems. Two possible analysis approaches can be applied to a complex network: the first based on the study of its topological structure, the second based on the dynamic properties of the system described by the network itself. Since "always structure affects function" (Strogatz (2001)), the topological approach wants to understand networks functionality through the analysis of their structure. For instance, the topological structure of the road network affects critical traffic jam areas, the topology of social networks affects the spread of information and disease and the topology of the power grid affects the robustness and stability of power transmission. Remarkable results have been reached in the topological analysis of networks, concerning the study and characterization of networks structure, and even if far from being complete, several key notions have been introduced. These unifying principles underly the topology of networks belonging to different fields of science. Fundamental are the notions of scale-free network (Barabasi & Albert (1999); Jeong et al. (2000)), cluster (Newman (2006)), network motifs (Milo et al. (2002); Shen-Orr et al. (2002)), small-world property (Watts & Strogatz (1998); Watts (1999); Wagner & Fell (2001)) and centralities. Particularly, centralities have been initially applied to the field of social science (Freeman (1977)) and then to biological networks (Wuchty & Stadler (2003)). Usually, works regarding biological networks rightly consider global properties of the network and when centralities are used, they are often considered from a global point of view, as for example analyzing degree or centralities distribution (Jeong et al. (2000); Wagner & Fell (2001); Wuchty & Stadler (2003); Yamada & Bork (2009); Joy et al. (2005)). A node-oriented approach have been used analyzing attack tolerance of network, where consequences of central nodes deletion are studied (Albert et al. (2000); Crucitti et al. (2004)). But also in this case the analysis have been concentrate on global properties of the network and not on the relevance of the single nodes in the network. Similarly, available software for network analysis is usually oriented to global analysis and characterization of the whole networks. To identify relevant nodes of a biological network, protocols of analysis integrating centralities analysis and lab experimental data are needed and the same for software allowing this kind of analysis. Cytoscape is an excellent visualization and analysis tool with the analysis features greatly enhanced by plug-ins. Plug-in such as NetworkAnalyzer (Assenov et al. (2008)) computes some node centralities but does not allow direct integration with experimental data. Applications such as VisANT

(Hu et al. (2005)), and Centibin (Junker et al. (2006)) calculate centralities, although they either calculate fewer centralities or are not suitable to integration with experimental data. Starting from these general considerations, the first part of this chapter concern the application of network centralities analysis to complex networks from a perspective oriented to identify relevant nodes, with a particular attention to biological networks. Necessary steps to do this are illustrated above through an example of protein-protein interaction network analysis. The aim of the first part of this chapter is to face the centralities analysis of a protein interaction network from a node oriented point of view. The same approach can then be extended to several kinds of complex networks. We want to identify nodes that are relevant for the networks for both centralities analysis and lab experiments. To do this, the following steps have been done:

- Some centralities that we consider significant have been detected. A biological meaning of these centralities have been hypothesized.
- A protocol of analysis for a protein network based on integration of centralities analysis and data from lab experiments (activation level) have been designed.
- A software (CentiScaPe) for computing centralities and integrating topological analysis results with lab experimental data set is presented.
- A human kino-phosphatome network have been extracted from a global human protein interactome data-set, including 11120 nodes and 84776 unique undirected interactions obtained from public data-bases.
- CentiScaPe have been applied to this human kino-phosphatome network and activation level (in threonine and thyrosine) of each protein obtained performeing lab experiments have been related to centrality values.
- Proteins important from both topological analysis and activation level have been easily identified: the attention of successive experiments and analysis should be focused on these proteins.

A further step have been introduced. Once we have identified relevant proteins in a network, we are interested in identifying non-obvious relation between these and other proteins in the network. In any network structure, the role of a node depends, not only on the features of the node itself, but also on the topological structure of the network and on the other nodes features. So even if centralities are node properties, they depend also on other nodes. We know that in a protein network nodes can be added or deleted because of different reasons as for example gene duplication (adding) or gene deletion or drug usage (deleting). More generally, If we delete a relevant node in a complex network, the effects of the deletion have impact not only on the single node and its neighbors, but also on other part of the network. For instance, if you are close friend of an important politician of your town, you have a central role in the social network of the town, and consequently your friends have a central role. But if this politician looses his central role, or if he is completely excluded from the political life of the town, for instance because they put him in prison (this correspond to a deletion on the social network), also you loose your central role in the network and the same for those people related to you. The idea is that the impact of an adding or deletion of a node can be measured through the variation of centrality values of the other nodes in the network. Such notion we introduced have been called “network centralities interference”. It allows to identify those nodes that are more sensitive to deletion or adding of a particular node in the network. The Interference Cytoscape plugin have been released to allow this kind of analysis (Scardoni & Laudanna (2011)).

Section 2 consists in a review of some centralities considered important with particular consideration for biological networks. For each centrality a possible biological meaning have been treated and some examples illustrate their significance. Section 3 introduce the CentiScaPe software, the Cytoscape plug-in we implemented for computing network centralities. Main feature of the software is the possibility of integrating experimental data-set with the topological analysis. In CentiScaPe, computed centralities can be easily correlated between each other or with biological parameters derived from the experiments in order to identify the most significant nodes according to both topological and biological properties. In section 4 the protocol of analysis is introduced through an example of analysis of a human kino-phosphatome network. Most relevant kinases and phosphatases according to their centralities values have been extracted from the network and their phosphorylation level in threonine and tyrosine have been obtained through a lab experiment. Centrality values and activation (phosphorylation) levels have been integrated using CentiScaPe and most relevant kinases and phosphatases according to both centrality values and activation levels have been easily identified. Section 5 introduce the Interference software to measure the changes in the topological structure of complex networks.

2. Node centralities: definition and description

In this section, some of the classical network centralities are introduced. For each centrality, we present the mathematical definition, a brief description with some examples, and a possible biological meaning in a protein network. A good and complete description of network centralities can be found in (Koschützki et al. (2005)), where also some algorithms are presented. For many centralities indices it is required that network is connected, i.e. each node is reachable from all the others. If not, some centralities can results in infinity values or some other not properly correct computation. Besides some centralities are not defined for directed graph (except of trivial situation), so we will consider here only connected undirected graph.

Preliminary definitions

Let $G = (N, E)$ an undirected graph, with $n = |N|$ vertexes. $deg(v)$, indicate the degree the vertex. $dist(v, w)$ is the shortest path between v and w. σ_{st} is the number of shortest paths between s and t and $\sigma_{st}(v)$ is the number of shortest paths between s and t passing through the vertex v. Notably:

- Vertex = nodes; edges = arches;
- The "distance" between two nodes, $dist(v, w)$ is the shortest path between the two nodes;
- All calculated scores are computed giving to "higher" values a "positive" meaning, where positive does refer to node proximity to other nodes. Thus, independently on the calculated node centrality, higher scores indicate proximity and lower scores indicate remoteness of a given node v from the other nodes in the graph.

2.1 Degree ($deg(k)$)

Is the simplest topological index, corresponding to the number of nodes adjacent to a given node v, where "adjacent" means directly connected. The nodes directly connected to a given node v are also called "first neighbors" of the given node. Thus, the degree also corresponds to the number of adjacent incident edges. In directed networks we distinguish in-degree, when

the edges target the node v, and out-degree, when the edges target the adjacent neighbors of v. Calculation of the degree allows determining the "degree distribution" $P(k)$, which gives the probability that a selected node has exactly k links. $P(k)$ is obtained counting the number of nodes $N(k)$ with $k = 1, 2, 3 \ldots$ links and dividing by the total number of nodes N. Determining the degree distribution allows distinguishing different kind of graphs. For instance, a graph with a peaked degree distribution (Gaussian distribution) indicates that the system has a characteristic degree with no highly connected nodes. This is typical of random, non-natural, networks. By contrast, a power-law degree distribution indicates the presence of few nodes having a very high degree. Nodes with high degree (highly connected) are called "hubs" and hold together several nodes with lower degree. Networks displaying a degree distribution approximating a power-law, $P(k) \approx k^{-\gamma}$, where γ is degree exponent, are called scale-free networks (Barabasi & Albert (1999)). Scale-free networks are mainly dominated by hubs and are intrinsically robust to random attacks but vulnerable to selected alterations (Albert et al. (2000); Jeong et al. (2001)). Scale-free networks are typically natural networks.

In biological terms

The degree allows an immediate evaluation of the regulatory relevance of the node. For instance, in signaling networks, proteins with very high degree are interacting with several other signaling proteins, thus suggesting a central regulatory role, that is they are likely to be regulatory "hubs". For instance, signaling proteins encoded by oncogenes, such as HRAS, SRC or TP53, are hubs. Depending on the nature of the protein, the degree could indicate a central role in amplification (kinases), diversification and turnover (small GTPases), signaling module assembly (docking proteins), gene expression (transcription factors), etc. Signaling networks have typically a scale-free architecture.

2.2 Diameter (Δ_G)

ΔG is the maximal distance (shortest path) amongst all the distances calculated between each couple of vertexes in the graph G. The diameter indicates how much distant are the two most distant nodes. It can be a first and simple general parameter of graph "compactness", meaning with that the overall proximity between nodes. A "high" graph diameter indicates that the two nodes determining that diameter are very distant, implying little graph compactness. However, it is possible that two nodes are very distant, thus giving a high graph diameter, but several other nodes are not (see figure 1). Therefore, a graph could have high diameter and still being rather compact or have very compact regions. Thus, a high graph diameter can be misleading in term of evaluation of graph compactness. In contrast a "low" graph diameter is much more informative and reliable. Indeed, a low diameter surely indicates that all the nodes are in proximity and the graph is compact. In quantitative terms, "high" and "low" are better defined when compared to the total number of nodes in the graph. Thus, a low diameter of a very big graph (with hundreds of nodes) is much more meaningful in term of compactness than a low diameter of a small graph (with few nodes). Notably, the diameter enables to measure the development of a network in time.

In biological terms

The diameter, and thus the compactness, of a biological network, for instance a protein-signaling network, can be interpreted as the overall easiness of the proteins to communicate and/or influence their reciprocal function. It could be also a sign of functional

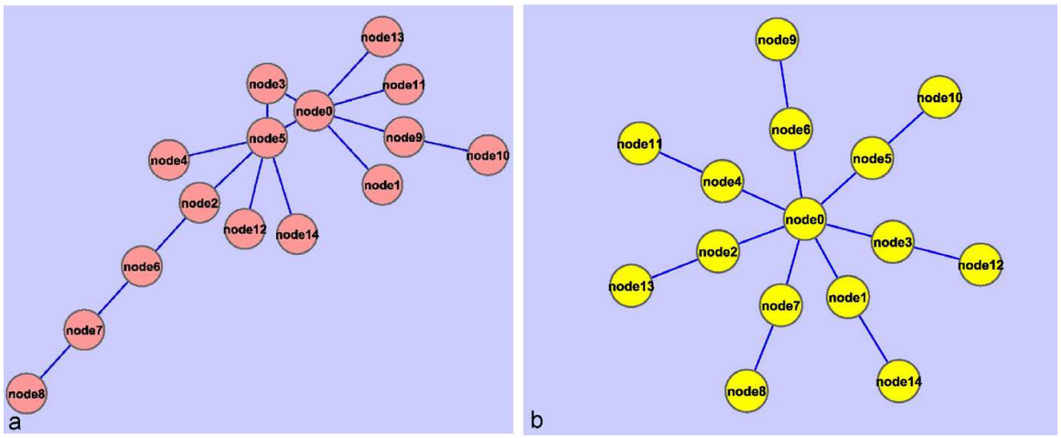


Fig. 1. a. A network where high diameter is due to a low number of nodes. b. A network with low diameter and average distance. The network is "compact".

convergence. Indeed, a big protein network with low diameter may suggest that the proteins within the network had a functional co-evolution. The diameter should be carefully weighted if the graph is not fully connected (that is, there are isolated nodes).

2.3 Average distance (AvD_G)

$$AvD_G = \frac{\sum_{i,j \in N} dist(i,j)}{n(n-1)}$$

where n is the number of nodes in G. The average distance (shortest path) of a graph G, corresponding to the sum of all shortest paths between vertex couples divided for the total number of vertex couples. Often it is not an integer. As for the diameter, it can be a simple and general parameter of graph "compactness", meaning with that the overall tendency of nodes to stay in proximity. Being an average, it can be somehow more informative than the diameter and can be also considered a general indicator of network "navigability". A "high" average distance indicates that the nodes are distant (disperse), implying little graph compactness. In contrast a "low" average distance indicates that all the nodes are in proximity and the graph is compact (figure **??**). In quantitative terms, "high" and "low" are better defined when compared to the total number of nodes in the graph. Thus, a low average distance of a very big graph (with hundreds of nodes) is more meaningful in term of compactness than a low average distance of a small graph (with few nodes).

In biological terms

The average distance of a biological network, for instance a protein-signaling network, can be interpreted as the overall easiness of the proteins to communicate and/or influence their reciprocal function. It could be also a sign of functional convergence. Indeed, a big protein network with low average distance may suggest that the proteins within the network have the tendency to generate functional complexes and/or modules (although centrality indexes should be also calculated to support that indication).

2.4 Eccentricity ($C_{ecc}(v)$)

$$C_{ecc}(v) := \frac{1}{max\{dist(v,w) : w \in N\}}$$

The eccentricity is a node centrality index. The eccentricity of a node v is calculated by computing the shortest path between the node v and all other nodes in the graph, then the "longest" shortest path is chosen (let (v,K) where K is the most distant node from v). Once this path with length $dist(v,K)$ is identified, its reciprocal is calculated $(1/dist(v,K))$. By doing that, an eccentricity with higher value assumes a positive meaning in term of node proximity. Indeed, if the eccentricity of the node v is high, this means that all other nodes are in proximity. In contrast, if the eccentricity is low, this means that there is at least one node (and all its neighbors) that is far form node v. Of course, this does not exclude that several other nodes are much closer to node v. Thus, eccentricity is a more meaningful parameter if is high. Notably, "high" and "low" values are more significant when compared to the average eccentricity of the graph G calculated by averaging the eccentricity values of all nodes in the graph.

In biological terms

The eccentricity of a node in a biological network, for instance a protein-signaling network, can be interpreted as the easiness of a protein to be functionally reached by all other proteins in the network. Thus, a protein with high eccentricity, compared to the average eccentricity of the network, will be more easily influenced by the activity of other proteins (the protein is subject to a more stringent or complex regulation) or, conversely could easily influence several other proteins. In contrast, a low eccentricity, compared to the average eccentricity of the network, could indicate a marginal functional role (although this should be also evaluated with other parameters and contextualized to the network annotations).

2.5 Closeness ($C_{clo}(v)$)

$$C_{clo}(v) := \frac{1}{\sum_{w \in N} dist(v,w)}$$

The closeness is a node centrality index. The closeness of a node v is calculated by computing the shortest path between the node v and all other nodes in the graph, and then calculating the sum. Once this value is obtained, its reciprocal is calculated, so higher values assume a positive meaning in term of node proximity. Also here, "high" and "low" values are more meaningful when compared to the average closeness of the graph G calculated by averaging the closeness values of all nodes in the graph. Notably, high values of closeness should indicate that all other nodes are in proximity to node v. In contrast, low values of closeness should indicate that all other nodes are distant from node v. However, a high closeness value can be determined by the presence of few nodes very close to node v, with other much more distant, or by the fact that all nodes are generally very close to v. Likewise, a low closeness value can be determined by the presence of few nodes very distant from node v, with other much closer, or by the fact that all nodes are generally distant from v. Thus, the closeness value should be considered as an "average tendency to node proximity or isolation", not really informative on the specific nature of the individual node couples. The closeness should be always compared to the eccentricity: a node with high eccentricity + high closeness is very

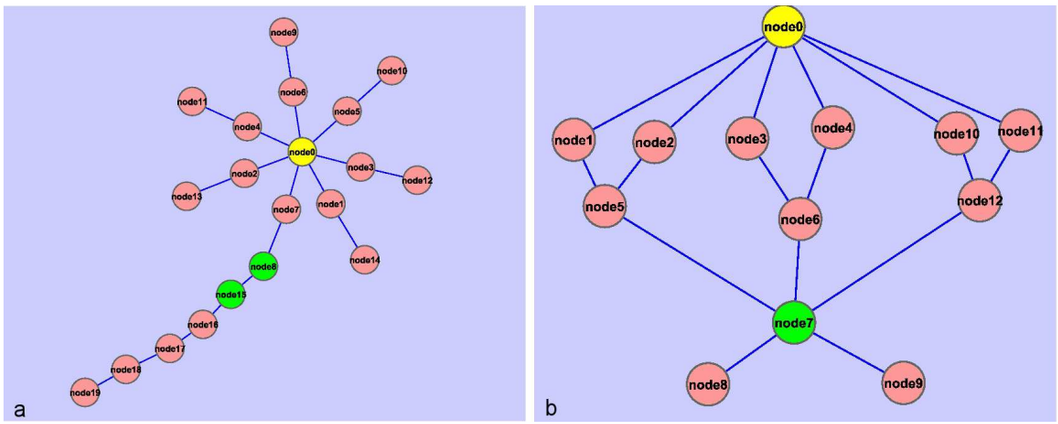


Fig. 2. a. The network shows the difference between eccentricity and closeness. The values of eccentricity are node0=0.14, node8=0.2, node15=0.2. The closeness values are node0=0.021, node8=0.017, node15=0.014. In this case node0 is closer than node8 and node15 to the most of nodes in the graph. Eccentricity value of node0 is smaller than value of node8 and node15, but this is due only to few nodes. If they are proteins this probably mean that node0 is fundamental for the most of reaction in the network, and that node8 and node15 are important only in reactions between few proteins. b. The network shows the difference between centroid and closeness. Here node0 has highest centroid value (centroid=1, closeness=0,04) and node7 has highest closeness value (centroid=-1, closeness= 0,05).

likely to be central in the graph. Figure 2 shows an example of difference between closeness and eccentricity.

In biological terms

The closeness of a node in a biological network, for instance a protein-signaling network, can be interpreted as a measure of the possibility of a protein to be functionally relevant for several other proteins, but with the possibility to be irrelevant for few other proteins. Thus, a protein with high closeness, compared to the average closeness of the network, will be easily central to the regulation of other proteins but with some proteins not influenced by its activity. Notably, in biological networks could be also of interest to analyze proteins with low closeness, compared to the average closeness of the network, as these proteins, although less relevant for that specific network, are possibly behaving as intersecting boundaries with other networks. Accordingly, a signaling network with a very high average closeness is more likely organizing functional units or modules, whereas a signaling network with very low average closeness will behave more likely as an open cluster of proteins connecting different regulatory modules.

2.6 Radiality ($C_{rad}(v)$)

$$C_{rad}(v) := \frac{\sum_{w \in N}(\Delta_G + 1 - dist(v,w))}{n-1}$$

The radiality is a node centrality index. The radiality of a node v is calculated by computing the shortest path between the node v and all other nodes in the graph. The value of each

path is then subtracted by the value of the diameter $+1$ ($\Delta G + 1$) and the resulting values are summated. Finally, the obtained value is divided for the number of nodes -1 ($n - 1$). Basically, as the diameter is the maximal possible distance between nodes, subtracting systematically from the diameter the shortest paths between the node v and its neighbors will give high values if the paths are short and low values if the paths are long. Overall, if the radiality is high this means that, with respect to the diameter, the node is generally closer to the other nodes, whereas, if the radiality is low, this means that the node is peripheral. Also here, "high" and "low" values are more meaningful when compared to the average radiality of the graph G calculated by averaging the radiality values of all nodes in the graph. As for the closeness, the radiality value should be considered as an "average tendency to node proximity or isolation", not definitively informative on the centrality of the individual node. The radiality should be always compared to the closeness and to the eccentricity: a node with high eccentricity + high closeness+ high radiality is a consistent indication of a high central position in the graph.

In biological terms

The radiality of a node in a biological network, for instance a protein-signaling network, can be interpreted as the measure of the possibility of a protein to be functionally relevant for several other proteins, but with the possibility to be irrelevant for few other proteins. Thus, a protein with high radiality, compared to the average radiality of the network, will be easily central to the regulation of other proteins but with some proteins not influenced by its activity. Notably, in biological networks could be also of interest to analyze proteins with low radiality, compared to the average radiality of the network, as these proteins, although less relevant for that specific network, are possibly behaving as intersecting boundaries with other networks. Accordingly, a signaling network with a very high average radiality is more likely organizing functional units or modules, whereas a signaling network with very low average radiality will behave more likely as an open cluster of proteins connecting different regulatory modules. All these interpretations should be accompanied to the contemporary evaluation of eccentricity and closeness.

2.7 Centroid value ($C_{cen}(v)$)

$$C_{cen}(v) := min\{f(v,w) : w \in N\{v\}\}$$

Where $f(v,w) := \gamma_v(w) - \gamma_w(v)$, and $\gamma_v(w)$ is the number of vertex closer to v than to w. The centroid value is the most complex node centrality index. It is computed by focusing the calculus on couples of nodes (v,w) and systematically counting the nodes that are closer (in term of shortest path) to v or to w. The calculus proceeds by comparing the node distance from other nodes with the distance of all other nodes from the others, such that a high centroid value indicates that a node v is much closer to other nodes. Thus, the centroid value provides a centrality index always weighted with the values of all other nodes in the graph. Indeed, the node with the highest centroid value is also the node with the highest number of neighbors (not only first) if compared with all other nodes. In other terms, a node v with the highest centroid value is the node with the highest number of neighbors separated by the shortest path to v. The centroid value suggests that a specific node has a central position within a graph region characterized by a high density of interacting nodes. Also here, "high" and "low" values are more meaningful when compared to the average centrality value of the graph G calculated by averaging the centrality values of all nodes in the graph.

In biological terms

The centroid value of a node in a biological network, for instance a protein-signaling network, can be interpreted as the "probability" of a protein to be functionally capable of organizing discrete protein clusters or modules. Thus, a protein with high centroid value, compared to the average centroid value of the network, will be possibly involved in coordinating the activity of other highly connected proteins, altogether devoted to the regulation of a specific cell activity (for instance, cell adhesion, gene expression, proliferation etc.). Accordingly, a signaling network with a very high average centroid value is more likely organizing functional units or modules, whereas a signaling network with very low average centroid value will behave more likely as an open cluster of proteins connecting different regulatory modules. It can be useful to compare the centroid value to algorithms detecting dense regions in a graph, indicating protein clusters, such as, for instance, MCODE (Bader & Hogue (2003)).

2.8 Stress ($C_{str}(v)$)

$$C_{str}(v) := \sum_{s \neq v \in N} \sum_{t \neq v \in N} \sigma_{st}(v)$$

The stress is a node centrality index. Stress is calculated by measuring the number of shortest paths passing through a node. To calculate the "stress" of a node v, all shortest paths in a graph G are calculated and then the number of shortest paths passing through v is counted. A "stressed" node is a node traversed by a high number of shortest paths. Notably and importantly, a high stress values does not automatically implies that the node v is critical to maintain the connection between nodes whose paths are passing through it. Indeed, it is possible that two nodes are connected by means of other shortest paths not passing through the node v. Also here, "high" and "low" values are more meaningful when compared to the average stress value of the graph G calculated by averaging the stress values of all nodes in the graph.

In biological terms

The stress of a node in a biological network, for instance a protein-signaling network, can indicate the relevance of a protein as functionally capable of holding together communicating nodes. The higher the value the higher the relevance of the protein in connecting regulatory molecules. Due to the nature of this centrality, it is possible that the stress simply indicates a molecule heavily involved in cellular processes but not relevant to maintain the communication between other proteins.

2.9 S.-P. Betweenness ($C_{spb}(v)$)

$$C_{spb}(v) := \sum_{s \neq v \in N} \sum_{t \neq v \in N} \delta_{st}(v)$$

where

$$\delta_{st}(v) := \frac{\sigma_{st}(v)}{\sigma_{st}}$$

The S.-P. Betweenness is a node centrality index. It is similar to the stress but provides a

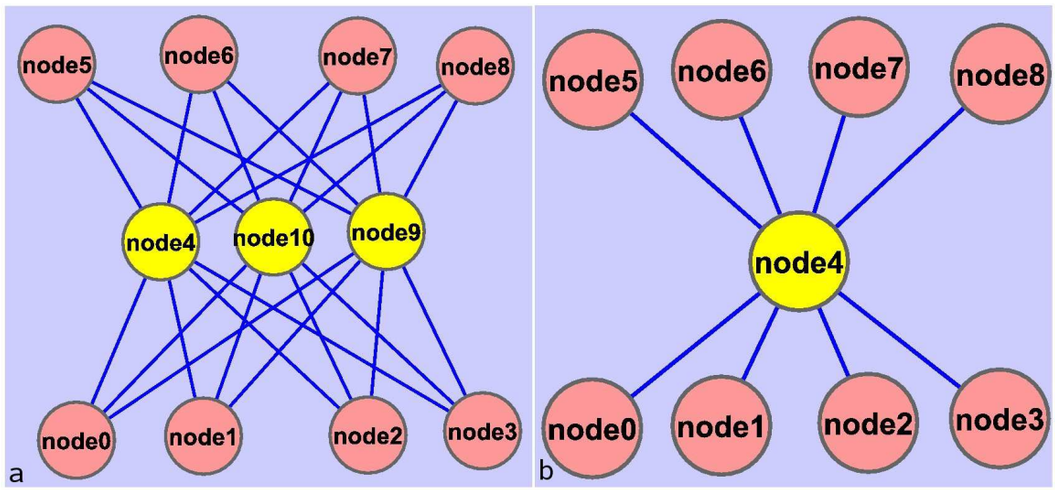


Fig. 3. Betweenness vs Stress. In fig. a node4, node10, and node9 present high value of stress (= 56), and the same value of betweenness (=18.67). In fig.b, node4 presents the same value of stress of fig.a and higher value of betweenness(=56). This is because the number of shortest paths passing through node4 is the same in the two network. But in the second network node4 is the only node connecting the two parts of the network. In this sense betweenness is more precise than stress giving also information on how the node is fundamental in the network. If we remove node4 in fig.a, the connection between the node in the network don't change so much. If we remove node 4 from fig.b the network is completely disconnected.

more elaborated and informative centrality index. The betweenness of a node n is calculated considering couples of nodes $(v1, v2)$ and counting the number of shortest paths linking $v1$ and $v2$ and passing through a node n. Then, the value is related to the total number of shortest paths linking $v1$ and $v2$. Thus, a node can be traversed by only one path linking $v1$ and $v2$, but if this path is the only connecting $v1$ and $v2$ the node n will score a higher betweenness value (in the stress computation would have had a low score). Thus, a high S.-P. Betweenness score means that the node, for certain paths, is crucial to maintain node connections. Notably, to know the number of paths for which the node is critical it is necessary to look at the stress. Thus, stress and S.-P. Betweenness can be used to gain complementary information. Further information could be gained by referring the S.-P. Betweenness to node couples, thus "quantifying" the importance of a node for two connected nodes. Also here, "high" and "low" values are more meaningful when compared to the average S.-P. Betweenness value of the graph G calculated by averaging the S.-P. Betweenness values of all nodes in the graph.

In biological terms

The S.-P. Betweenness of a node in a biological network, for instance a protein-signaling network, can indicate the relevance of a protein as functionally capable of holding together communicating proteins. The higher the value the higher the relevance of the protein as organizing regulatory molecule. The S.-P. Betweenness of a protein effectively indicates the capability of a protein to bring in communication distant proteins. In signaling modules, proteins with high S.-P. Betweenness are likely crucial to maintain functionally and coherence of signaling mechanisms.

2.10 Normalization and relative centralities

Once centralities have been computed, the question that arise immediately is what does it means to have a centrality of, for example, 0.4 for a node? This clearly depends on different parameters as the number of nodes in the network, the maximum value of the centrality and on the topological structure of the network. In order to compare centrality scores between the elements of a graph or between the elements of different graphs some kind of normalization of centrality values is needed. Common normalizations applicable to most centralities are to divide each value by the maximum centrality value or by the sum of all values. We will use the second defining it as the relative centralities value. So, given a centrality C, $C(G,n)$ is the value of the centrality of node n in the network G. We define the relative centrality value of node n as:

$$relC(G,n) = \frac{C(G,n)}{\sum_{i \in N} C(G,i)}$$

So a relative centrality of 0.4 means that the node has the 40% of the total centrality of the network.

2.11 Conclusions

A review of nodes centralities have been presented. The centralities introduced have been chosen for their biological relevance, and a possible biological meaning for each centrality have been hypothesized. Normalization of centralities, useful for comparison between nodes in a network and between nodes of different networks have also been considered.

3. CentiScaPe a software for network centralities

In this section we describe the CentiScaPe software (Scardoni et al. (2009)), a Cytoscape (Cline et al. (2007); Shannon et al. (2003)) plugin we implemented to calculate centralities values and integrating topological analysis of networks with lab experimental data. Main concepts of this section have been published on (Scardoni et al. (2009)).

The vast amount of available experimental data generating annotated gene or protein complex networks has increased the quest for visualization and analysis tools to understand individual node functions masked by the overall network complexity. Cytoscape is an excellent visualization and analysis tool with the analysis features greatly enhanced by plug-ins. Plug-in such as NetworkAnalyzer (Assenov et al. (2008)) computes some node centralities but does not allow direct integration with experimental data. Applications such as VisANT (Hu et al. (2005)), and Centibin (Junker et al. (2006)) calculate centralities, although they either calculate fewer centralities or are not suitable to integration with experimental data. Figure 1 shows a comparative evaluation of CentiScaPe and other applications. CentiScaPe is the only Cytoscape plug-in that computes several centralities at once. In CentiScaPe, computed centralities can be easily correlated between each other or with biological parameters derived from the experiments in order to identify the most significant nodes according to both topological and biological properties. Functional to this capability is the scatter plot by value options, which allows easy correlating node centrality values to experimental data defined by the user. Particularly this feature allows a new way to face the analysis of biological network, integrating topological analysis and lab experimental data. This new approach is described in section 4. At present version 1.2 is available and it is downloaded with a rate of about hundred downloads for month (see Cytoscape

Feature	Centiscape	Network analyzer	Visant	Centibin
Degree	Yes	Yes	Yes	Yes
Radiality	Yes	Yes	No	Yes
Closeness	Yes	Yes	No	Yes
Stress	Yes	Yes	No	Yes
Betweenness	Yes	Yes	No	Yes
Centroid value	Yes	No	No	Yes
Eccentricity	Yes	Yes	No	Yes
Scatter plot between centralities	Yes	No	No	No
Scatter plot with experimental data	Yes	No	No	No
Plot by node	Yes	No	No	No
Highlighting filter	Yes	No	No	No

Table 1. Features of CentiScaPe versus Network Analyzer, Visant, Centibin

website for download statistics). Several results using CentisCaPe have been published in Arsenio Rodriguez (2011); Sengupta et al. (2009a); Lepp et al. (2009); Sengupta et al. (2009b); Biondani et al. (2008); Sengupta et al. (2009c); Feltes et al. (2011); Schokker et al. (2011); Ladha et al. (2010); Choura & Rebaï (2010); Venkatachalam et al. (2011); Webster et al. (2011).
Availability: CentiScaPe can be downloaded via the Cytoscape web site:
http://chianti.ucsd.edu/cyto_web/plugins/index.php.
Tutorial, centrality descriptions and example data are available at:
http://www.cbmc.it/%7Escardonig/centiscape/centiscape.php

3.1 System overview

CentiScaPe computes several network centralities for undirected networks. Computed parameters are: Average Distance, Diameter, Degree, Stress, Betweenness, Radiality, Closeness, Centroid Value and Eccentricity. Plug-in help and on-line files are provided with definition, description and biological significance for each centrality (see section 2). Min, max and mean values are given for each computed centrality. Multiple networks analysis is also supported. Centrality values appear in the Cytoscape attributes browser, so they can be saved and loaded as normal Cytoscape attributes, thus allowing their visualization with the Cytoscape mapping core features. Once computation is completed, the actual analysis begins, using the graphical interface of CentiScaPe.

3.2 Algorithm and implementation

To calculate all the centralities the computation of the shortest path between each pair of nodes in the graph is needed. The algorithm for the shortest path is the well known Dijkstra algorithm (Dijkstra (1959)). There are no costs in our network edges, so in our case the algorithm keeps one as the cost of each edge. To compute Stress and Betweenness we need all the shortest paths between each pair of nodes and not only a single shortest path between each pair. To do this the Dijkstra algorithm has been adjusted as follows. Exploring the graph

when calculating the shortest path between two nodes s and t, the Dijkstra algorithm keep for each node n a predecessor node p. The predecessor node is the node that is the predecessor of n in one of the shortest paths between s and t. So in case of the Dijkstra algorithm, only one predecessor for each node is needed. To have all the shortest paths, we replace the predecessor p with a set of predecessors for each node n. The set of predecessors of the node n is the set of all the predecessors of the node n in the shortest paths set between s and t, i.e. one node is in the set of predecessors of n if it is a predecessor of n in one of the shortest paths between s and t containing n. Once the predecessors set of each node n has been computed, also the tree of all the shortest paths between s and t can be easily computed. Once we have computed all the shortest paths between each pairs of nodes of our network, the algorithm of each centralities comes directly from the formal definition of each centrality. Computational complexity for each centrality value is shown in table 2. A well done description of this and

Centrality	Computational complexity
Diameter	$O(mn + n^2)$
Average distance	$O(mn + n^2)$
Degree (deg(v))	$O(n)$
Radiality (rad(v))	$O(mn + n^2)$
Closeness (clo(v))	$O(mn + n^2)$
Stress (str(v))	$O(mn + n^2)$
Betweenness (btw(v))	$O(n^3)$
Centroid Value (cen(v))	$O(mn + n^2)$
Eccentricity (ecc(v))	$O(mn + n^2)$

Table 2. Computational complexity for each centrality value. n is the number of nodes and m is the number of edges in the network.

other centralities algorithms can be found in (Koschützki et al. (2005)). CentiScaPe is written in Java as a Cytoscape plugin, in order to exploit all the excellent features of Cytoscape and to reach the larger number of users. The Java library JFreechart (Gilbert (n.d.)) has been used for some graphic features.

3.3 Using CentiScaPe

Once CentiScaPe have been started, the main menu will appear as a panel on the left side of the Cytoscape window as shown in figure 4. The panel shows to the user the list of centralities and the user can select all the centralities or some of them. A banner and a node worked count appear during the computation to show the computation progress. The numerical results are saved as node or network attributes in the Cytoscape attributes browser, depending on the kind of parameters, so all the Cytoscape features for managing attributes are supported: after the computation the centralities are treated as normal Cytoscape attributes. The value of each centrality is saved as an attribute with name "CentiScaPe" followed by the name of the centrality. For example the eccentricity is saved in the Cytoscape attributes browser as "CentiScaPe Eccentricity". Since the Cytoscape attributes follow the alphabetical order this make it easy to find all the centralities in the attributes browser list. There are two kind of centralities: network centralities, and node centralities.

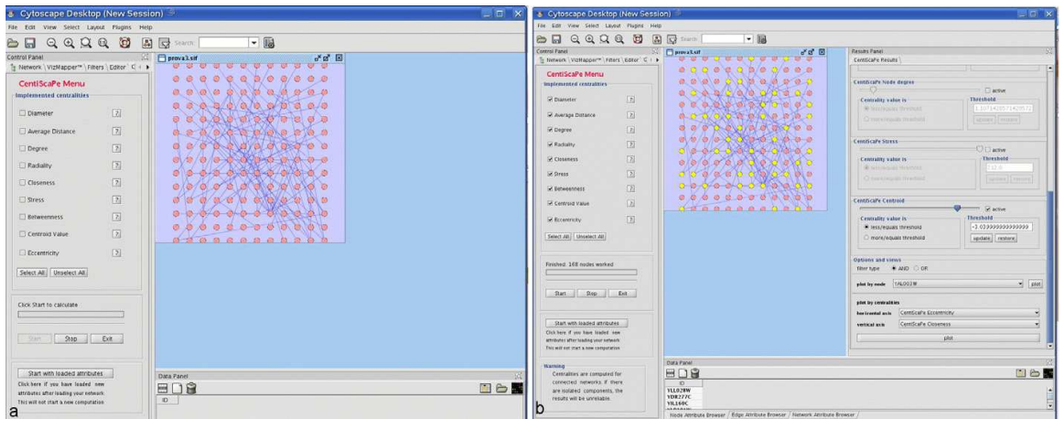


Fig. 4. a. CentiScaPe starting panel. On the left side the main menu appears, to select the centralities for computation. b. A computation results of CentiScaPe. All nodes having centrality values more/equal than the corresponding threshold (AND operator) are highlighted.

Network centralities

The network centralities concern the entire network and not the single nodes. They are the Diameter and the Average Distance. They will appear on the data panel selecting the Cytoscape network attribute browser.

Node centralities

All other centralities are node parameters and refer to the single nodes. So they will appear on the attribute browser as node attributes. Using the Node attribute browser the user can select one or more of them as normal attributes. CentiScaPe also calculates the min, max and mean value for each centrality. Since they are network parameters they appear on the Network attribute browser. As for the other attributes the user can save and load network and node parameters to/from a file. If an attribute is already loaded or calculated and the user try to recalculate it, a warning message will appear.

CentiScaPe results panel

If one or more node centralities have been selected, a result panel will appear on the right side of the Cytoscape window (figure 4b). The first step of the analysis is the Boolean logic-based result panel of CentiScaPe (figure 4b). It is possible, by using the provided sliders in the Results Panel of Cytoscape, to highlight the nodes having centralities values that are higher, minor or equal to a threshold value defined by the user. The slider threshold is initialized to the mean value of each centrality so all the nodes having a centrality value less or equal to the threshold are highlighted by default in the network view with a color depending on the selected visual mapper of Cytoscape (yellow in figure 4b). So if one centrality has been selected, all the nodes having a value less or equal the threshold for that centralities are highlighted. If more than one centralities has been selected they can be joined with an AND or an OR operator. If the AND operator is selected, the nodes for which all the values are less or equal the corresponding threshold are highlighted. If the OR operator is selected the nodes for which at least one value is less or equal the corresponding threshold are highlighted. The possibility of highlighting also the nodes that are more/equal than the

threshold is supported. So the user can select the more/equal option for some centralities, the less/equal option for others and can join them with the AND or the OR operator. If necessary, one or more centralities can be deactivated. This feature can immediately answer to questions as: "Which are the nodes having high Betweenness and Stress but low Eccentricity?" Notably, the threshold can also be modified by hand to gain in resolution. In figure **??** are highlighted all the nodes having centralities values more/equal than the corresponding threshold (AND operator). Once the nodes have been selected according to their node-specific values, the corresponding subgraph can be extracted and displayed using normal Cytoscape core features.

Graphic output

Two kind of graphical outputs are supported: plot by centrality and plot by node, both allowing analysis that are not possible with other centralities tools. The user can correlate

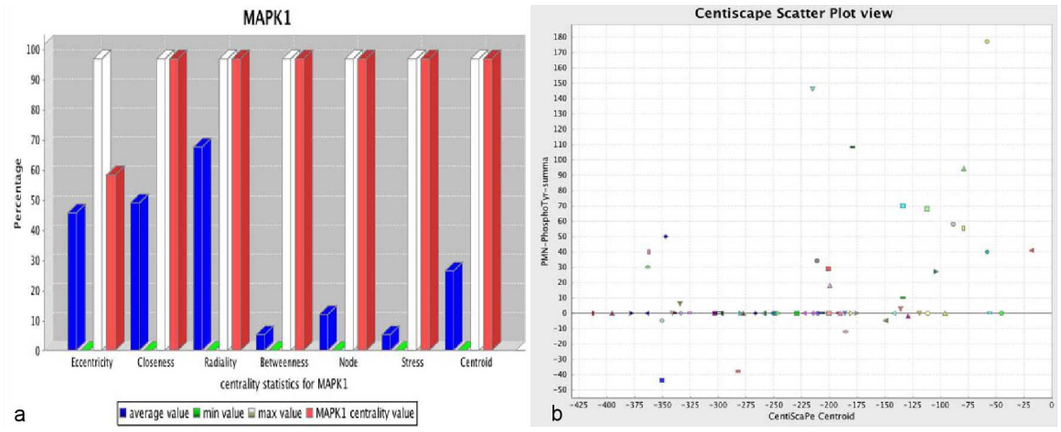


Fig. 5. a. Network analysis of human kino-phosphatome. The protein kinase MAPK1 shows high centralities values for most of the computed centralities suggesting its central role in the network structure and function. For each centrality the specific node value (red), the mean value (blue), the min value (green), and the max value (white) is shown. b. Integration of topological analysis with experimental data. Centroid values are plotted over protein phosphorylation levels in tyrosine. Relevant nodes are easily identified in the top-right quadrant.The centralities values and the node identifier appear in CentiScaPe by passing with the mouse over each geometrical shape in the plot.

centralities between them or with experimental data, such as, for example, gene expression level or protein phosphorylation level (plot by centrality), and can analyze all centralities values node by node (plot by node). Example of plot by node and plot by centrality are shown in figure 5. Graphics can be saved to a jpeg file.

Plot by centrality

The plot by centrality visualization is an easy and convenient way to discriminate nodes and/or group of nodes that are most relevant according to a combination of two selected parameters. It shows correlation between centralities and/or other quantitative node attributes, such as experimental data from genomic and/or proteomic analysis. The result of the plot by centrality option is a chart where each individual node, represented by a geometrical shape, is mapped to a Cartesian axis. In the horizontal and vertical axis, the

values of the selected attributes are reported. Most of the relevant nodes are easily identified in the top-right quadrant of the chart. Figure 5b shows a plot of centroid values over intensity of protein tyrosine phosphorylation in the human kino-phosphatome network derived from the analysis of human primary polymorphonuclear neutrophils (PMNs) stimulated with the chemoattractant IL-8 (see section 4). The proteins having high values for both parameters likely play a crucial regulatory role in the network. The user can plot in five different ways: centrality versus centrality, centrality versus experimental data, experimental data versus experimental data, a centrality versus itself and an experimental data versus itself. Notably, a specific way to use the plot function is to visualize the scatter plot of two experimental data attributes. This is an extra function of the plug-in and can be used in the same way of the centrality/centrality option and centrality/experimental attribute option. If the plot by centrality option is used selecting the same centrality (or the same experimental attribute) for both the horizontal and the vertical axis, result is an easy discrimination of nodes having low values from nodes having high values of the selected parameter (figure 6a) Thus, the main use of the "plot by centrality" feature is to identify group of nodes clustered according to combination of specific topological and/or experimental properties, in order to extract sub-networks to be further analyzed. The combination of topological properties with experimental data is useful to allow more meaningful predictions of sub-network function to be experimentally validated.

Plot by node

The plot by node option, another unique feature of CentiScaPe, shows for every single node the value of all calculated centralities represented as a bar graph. The mean, max and min values are represented with different colors. To facilitate the visualization, all the values in the graph are normalized and the real values appear when pointing the mouse over a bar. Figure 5a shows, as an example, the values for the MAPK1 calculated from the global human kino-phosphatome (see section 4).

3.4 Conclusions

CentiScaPe is a versatile and user-friendly bioinformatic tool to integrate centrality-based network analysis with experimental data. CentiScaPe is completely integrated into Cytoscape and the possibility of treating centralities as normal attributes permits to enrich the analysis with the Cytoscape core features and with other Cytoscape plug-ins. The analysis obtained with the Boolean-based result panel, the "plot by node" and the "plot by centrality" options give meaningful results not accessible to other tools and allow easy categorization of nodes in large complex networks derived from experimental data.

4. A new protocol of analysis. Centralities in the human kino-phosphatome

In this section a new protocol of analysis of protein interaction network is introduced through an example of analysis of the human kino-phosphatome (Scardoni et al. (2009)). The analysis starts with the extraction of known interaction from a protein interactome. In our case we consider kinases and phosphatases interaction i.e. those interaction regarding activation and inhibition of proteins in the network. Kinases and phosphatases are enzymes involved in the phosphorilation process: they transfer or remove phosphate groups to/from a protein regulating in this way its activity. Substantially kinases and phosphatases activate or inhibit other proteins. In a kino-phosphatome network this process generates a cascade of activations

and inhibitions of proteins corresponding to the transmissions of signals and to the control of complex processes in cells. The approach to the kino-phosphatome network is to identify the most important proteins for their centrality values and then to analyze with a lab experiment their activation level. After this, using the CentiScaPe feature of integrating topological analysis and data from lab experiments, those values are integrated and those nodes important for both centralities value and activation level are easily identified. This introduce a new way of facing the analysis of a protein interaction network based on the Strogatz assertion that in a biological network "Structure always affects function" (Strogatz (2001)). Instead of concentrating the analysis, as usual, on the global properties of the network (such degree distribution, centralities distribution, and so on) we consider in a cause-effect point of view single nodes of the network relating their centrality values (cause) with activation level (effects). Most of the contents of this section have been published on (Scardoni et al. (2009)).

4.1 Centralities analysis

The protocol used for the analysis of the human kino-phosphatome network is the following:

- The nodes of interest are extracted from the global network, resulting in a subnetwork to analyze (in our example the subnetwork of human kinases and phosphatases have been extracted from a human proteins interactome).
- The centralities values are computed with CentiScaPe. A subnetwork of proteins with all centrality values over the average is extracted.
- The lab experiment identifies which of these proteins present high phosphorilation level (in our example in tyrosine and threonine)
- Using CentiScaPe, lab experimental data and centrality values are integrated, so proteins with high level of activation and high centralities values are easily identified.
- Next experiments and analysis should be focused on these proteins.

This protocol have been applied as follows. A global human protein interactome data-set (Global Kino-Phosphatome network), including 11120 nodes and 84776 unique undirected interactions (IDs = HGNC), was complied from public data-bases (HPRD, BIND, DIP, IntAct, MINT, others; see (Scardoni et al. (2009)) on-line file GLOBAL-HGNC.sif) between human protein kinases and phosphatases. The resulting sub-network, a kino-phosphatome network, consisted of 549 nodes and 3844 unique interactions (see (Scardoni et al. (2009)) on-line files Table S4 and Kino-Phosphatome.sif), with 406 kinases and 143 phosphatases. The kino-phosphatome network did not contain isolated nodes. We used CentiScaPe to calculate centrality parameters. A first general overview of the global topological properties of the kino-phosphatome network comes from the min, max and average values of all computed centralities along with the diameter and the average distance of the network (table 3). These data provide a general overview of the global topological properties of the kino-phosphatome network. For instance, an average degree equals to 13.5 with an average distance of 3 may suggest a highly connected network in which proteins are strongly functionally interconnected. Computation of network centralities allowed a first ranking of human kinases and phosphatases according to their central role in the network (see (Scardoni et al. (2009)) on-line files Table S6 reporting all node-by-node values of different centralities). To facilitate the identification of nodes with the highest scores we applied the "plot by centrality" feature of CentiScaPe. A first plotting degree over degree generated a linear distribution, as expected (see fig. 6). However, it is evident that the distribution is

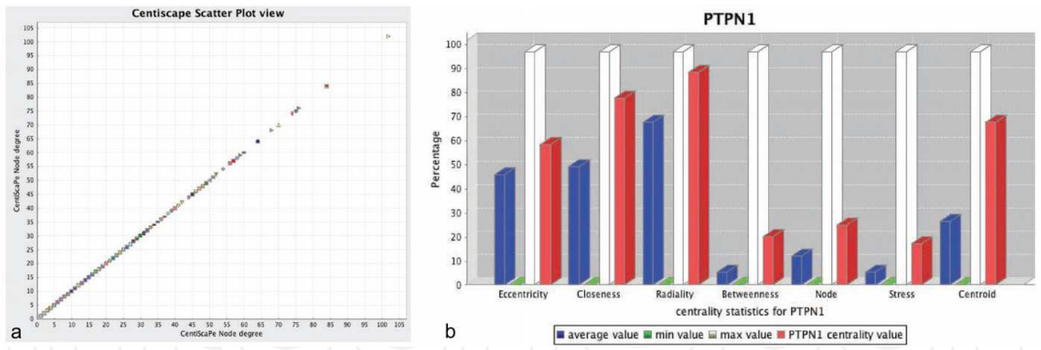


Fig. 6. a. A scatter plot degree over degree. As expected this generate a linear distribution. Notably the distribution is not uniform: many nodes display low degree and only few nodes with high degree, according to the scale-free architecture of biological network. b. The "plot by node" representation for PTPN1. The phosphatase PTPN1 presents the highest degree between all the phosphatases and a rather high score for other centralities. This suggests that PTPN1 may play a central regulatory role in the network. For each centrality the specific node value (red), the mean value (blue), the min value (green), and the max value (white) is shown.

not uniform, with the majority of nodes having a similar low degree and very few having very high degree. This is consistent with the known scale-free architecture of biological networks (Jeong et al. (2000)). The scale-free topology of the kino-phosphatome network was also confirmed with Network Analyzer (Assenov et al. (2008)). A total of 186 nodes (164 kinases and 22 phosphatases) displayed a degree over the average. The top 10 degrees (64 to 102) were all kinases, with MAPK1 showing the highest degree (102). Notably, MAPK1 displayed the highest score for most of the computed centralities (fig. 5a), suggesting its central regulatory role in the kino-phosphatome. In contrast, PTPN1 had the highest degree, 46, between all phosphatases (top 31 among all nodes) and had a rather high score also for other centralities Thus, degree analysis suggests that MAPK1 and PTPN1 are the most central kinase and phosphatase, respectively. To further support this suggestion we analyzed the centroid. Average centroid was -393. 242 nodes (206 kinases and 36 phosphatases) displayed a centroid over the average. The top 10 centroid (-79 to 8) were all kinases, with MAPK1 showing the highest centroid value (18). PTPN1 had the highest centroid value, -154, between all phosphatases (top 22 among all nodes). Thus, as for the degree, also the centroid value analysis suggests a possible scale-free distribution, with MAPK1 and PTPN1 being the most central kinase and phosphatase, respectively. This conclusion is also easily evidenced by plotting the degree over the centroid (fig. 7). Here MAPK1 appears at the top right of the plot and PTPN1 is present in the top most dispersed region of the plot, thus suggesting their higher scores. Interestingly, from the analysis is evident a non-linear distribution of nodes, with few dispersed nodes occupying the top right quadrant of the plot (i.e. high degree and high centroid): these nodes can potentially represent particularly important regulatory kinases and phosphatases. This kind of analysis can be iterated by evaluating all other centralities. To extract the most relevant nodes according to all centrality values we used CentiScaPe to select all nodes having all centrality values over the average. Upon filtering we obtained a kino-phosphatome sub-network (fig.**??**) consisting of 97 nodes (82 kinases and 15 phosphatases) and 962 interactions (see (Scardoni et al. (2009)) on-line files Table S7, and K-P sub-network.sif).

CentiScaPe Average Distance	3.0292037280789224
CentiScaPe Betweenness Max value	20159.799011925716
CentiScaPe Betweenness mean value	1112.0036429872616
CentiScaPe Betweenness min value	0.0
CentiScaPe Centroid Max value	18.0
CentiScaPe Centroid mean value	-393.07285974499086
CentiScaPe Centroid min value	-547.0
CentiScaPe Closeness Max value	8.771929824561404E-4
CentiScaPe Closeness mean value	6.175318530305184E-4
CentiScaPe Closeness min value	3.505082369435682E-4
CentiScaPe Diameter	8.0
CentiScaPe Eccentricity Max value	0.25
CentiScaPe Eccentricity mean value	0.18407494145199213
CentiScaPe Eccentricity min value	0.125
CentiScaPe Radiality Max value	6.91970802919708
CentiScaPe Radiality mean value	5.970796271921072
CentiScaPe Radiality min value	3.7937956204379564
CentiScaPe Stress Max value	210878.0
CentiScaPe Stress mean value	11537.009107468124
CentiScaPe Stress min value	0.0
CentiScaPe degree Max value	102.0
CentiScaPe degree mean value	13.5591985428051
CentiScaPe degree min value	1.0

Table 3. Global values of the kino-phosphatome network computed using CentiScaPe. The table includes min, max and mean value for each centrality and also the global parameter Diameter and Average Distance.

This sub-network possibly represents a group of highly interacting kinases and phosphatases displaying a critical role in the regulation of protein phosphorylation in human cells. Further analysis with CentiScaPe or other analysis tools, such as MCODE (Bader & Hogue (2003)) or Network Analyzer (Assenov et al. (2008)), performing a Gene Ontology database search (Ashburner et al. (2000)), or adding functional annotation data, may allow a deeper functional exploration of this sub network. The regulatory role of proteins belonging to the kino-phosphatome network may be also experimentally tested in a context-selective manner. Indeed, the centrality analysis by CentiScaPe can be even more significant by superimposing experimental data. To test this possibility, we focused the analysis on human polymorphonuclear neutrophils (PMNs).

4.2 Phosphoproteomic analysis of chemoattractant stimulated human PMNs

Human primary polymorphonuclear cells isolation

Human primary polymorphonuclear cells (PMNs) were freshly isolated form whole blood of healthy donors by ficoll gradient sedimentation. Purity of PMN preparation was evaluated by flow cytometry and estimated to about 95% of neutrophils. Isolated PMNs were kept in culture at 37^oC in standard buffer (PBS, $1mM\ CaCl_2$, $1mM\ MgCl_2$, 10% FCS, $pH7.2$) and used within 1 hour. Viability before the assays was more than 90%.

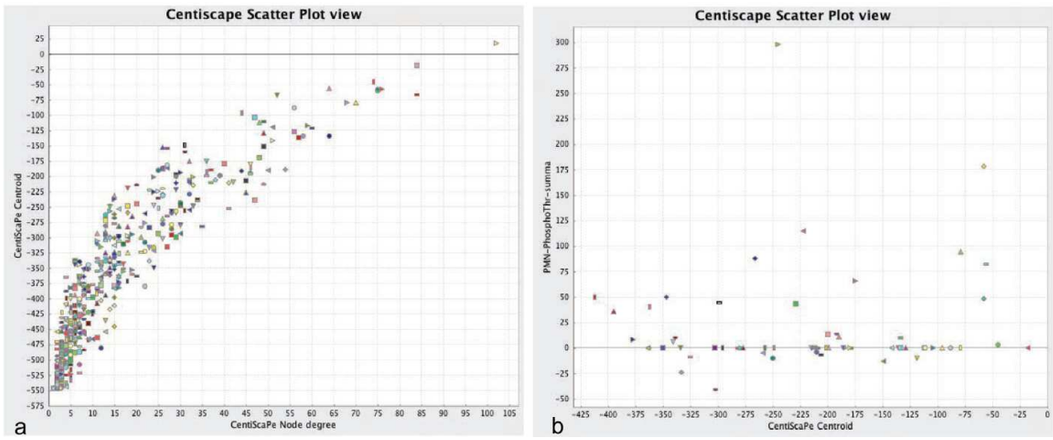


Fig. 7. a. A "plot by centralities" representation of degree over centroid. In the top right of the plot appear the nodes having high values of both degree and centroid (including MAPK1). b. Integration of topological analysis with experimental data. Centroid values are plotted over protein phosphorylation levels in threonine, experimentally determined as described in the text.

Human primary polymorphonuclear cell stimulation

Human neutrophils were resuspended in standard buffer at $10^7/ml$ and stimulated under stirring at 37^oC for 1 min. with the classical chemoattractant fMLP ($100nM$). Stimulation was blocked by directly disrupting the cells for 10 min. in ice-cold lysis buffer containing: $20mM$ MOPS, $pH7.0$, $2mM$ EGTA, $5mM$ EDTA, $30mM$ sodium fluoride, $60mM$ β-glycerophosphate, $20mM$ sodium pyrophosphate, $1mM$ sodium orthovanadate, $1mM$ phenylmethylsulfonylfluoride, $3mM$ benzamidine, $5\mu M$ pepstatin A, $10\mu M$ leupeptin, 1% Triton X-100. Lysates were clarified by centrifugation at $12.000xg$ for 10 min. and kept at 80^oC until further processing.

Evaluation of protein phosphorylation

Protein phosphorylation was evaluated both qualitatively and quantitatively by using the Kinexus protein array service (Kinexus (n.d.)). Kinexus provides a complete service for high throughput proteomic and phosphoproteomic high sensitive analysis of cell lysed samples, allowing detection of more than 800 proteins, including about 200 phosphorylated proteins (about 350 phospho-sites) by means of in-house validated antibody microarrays (Kinexus (n.d.)). $100\mu l$ of frozen samples of lysed PMNs (about $1mg/ml$ protein concentration) have been sent to Kinexus for the analysis. Phosphoproteomic antibody microarray data have been delivered by email and subsequently elaborated to extract values of protein phosphorylation of control versus agonist-triggered samples. (phosphorylation data files are available on-line: see (Scardoni et al. (2009)) PMN-PhosphoSer.NA, PMN-PhosphoTyr.NA, PMN-PhosphoThr.NA).

4.3 Combining topological analysis and experimental data

Data about protein phosphorylation were used as bioinformatic probes and node attributes to extract, from the Global Kino-Phosphatome network, subnetworks of protein phosphorylation, to be analyzed with CentiScaPe Experimental data were loaded as node

attributes in Cytoscape and the computed centrality values were plotted over values of protein phosphorylation. Here, every node is represented with two coordinates consisting of a computed centrality and of experimental data regarding protein phosphorylation induced in PMNs by fMLP. In figures 5b and 7b are shown plots of centroid values over intensity of protein phosphorylation in threonine or tyrosine residues induced by fMLP triggering in human PMNs. Notably, in the plot are shown only those proteins whose phosphorylation level was experimentally determined. The two plots allow immediately evidencing that proteins phosphorylated in threonine (fig. 5b) or in tyrosine (fig. 7b) have different topological position in the network, with proteins phosphorylated in tyrosine showing a higher centrality values. This could suggest that tyrosine phosphorylation induced in PMNs by chemoattractants involves signaling proteins regulating clusters of proteins, as the centroid value may suggest. Besides, the top/left quadrant is empty in both figures 5b and 7b. So there are no nodes having low centroid value and high phosphorylation in threonine or tyrosine. This may suggest that centroid value and activation level are strictly related. Further hypotheses can be formulated by expanding the analysis to other centralities and by adding more phopshorylation data. From this type of plotting it is possible to further identify relevant nodes not only according to topological position but also to experimental outputs. Thus, groups of nodes whose regulatory relevance is suggested by centrality analysis are further characterized by the corresponding data of biological activity.

4.4 Conclusions

In this section a protocol of analysis for protein network have been proposed. The key idea is that of identify most important proteins from both topological and biological point of view. Through the example of the kino-phosphatome network, we have seen how CentiScaPe can integrate the two kinds of analysis allowing an easy characterization of most relevant proteins. The topological analysis and experimental data do confirm each other's regulatory relevance and may suggest further, more focused, experimental verifications. Combination of CentiScaPe with other bioinformatics tools may help to analyze high throughput genomic and/or proteomic experimental data and may facilitate the decision process.

5. A further step in centralities analysis: node centralities interference

As seen in the previous section, network centralities allow us to understand the role and the importance of each single node in a protein network. Next step we introduce in this section is to understand and measure changes to the topological structure of the network. The effects of mutation in the network structure, have been studied from a global point of view: nodes are removed from the network and the effects on some global parameters, as for example diameter, average distance or global efficiency are evaluated (Barabasi & Oltvai (2004); Jeong et al. (2001); Albert et al. (2000); Crucitti et al. (2004)). Our approach wants to answer to this question: "we remove or add one node in the network, how do other nodes modify their functionality because of this removal?" Since centrality indexes allow categorizing nodes in complex networks according to their topological relevance (see CentiScaPe plugin), in a node-oriented perspective, centralities are very useful topological parameters to compute in order to quantify the effect of individual node(s) alteration. We introduced the notion of interference and developed the Cytoscape plugin **Interference** (Scardoni & Laudanna (2011)) to evaluate the topological effects of single or multiple nodes removal from a network. In this perspective, interference allows virtual node knock-out experiments: it is possible to remove one or more nodes from a network and analyze the consequences on network structure,

by looking to the variations of the node centralities values. As the centrality value of a node is strictly dependent on the network structure and on the properties of other nodes in the network, the consequences of a node deletion are well captured by the variation on the centrality values of all the other nodes. The interference approach can model common situations where real nodes are removed or added from/to a physical network:

- Biological networks, where one or more nodes (genes, proteins, metabolites) are possibly removed from the network because of gene deletion, pharmacological treatment or protein degradation. Interference can be used to:
  - Simulate pharmacological treatment: one can potentially predict side effects of the drug by looking at the topological properties of nodes in a drug-treated network, meaning with that a network in which a drug-targeted node (protein) was removed. To inhibit a protein (for instance a kinases) corresponds to removing the node from the network
  - Simulate gene deletion: gene deletion implies losing encoded proteins, thus resulting in the corresponding removal of one or more nodes from a protein network
- Social and financial networks, where the structure of the network is naturally modified over time
- Power grid failures
- Traffic jam or work in progress in a road network
- Temporary closure of an airport in an airline network

5.1 How Interference plugin works

The Interference plug-in allows you identifying the area of influence of single nodes or group of nodes. Specifically, Interference:

- Compute the centralities value of the network
- Remove the node(s) of interest
- Recompute the centralities in the new network (the one where the node(s) have been removed)
- Evaluate the differences between the centralities in the two networks.

This allows identifying the differences between the two networks: the one with the node(s) of interest still present and the one where the node(s) have been removed. Some nodes will increase, whereas others will reduce, their individual centrality values. This may suggest hints on the functional, regulatory, relevance of specific node(s).

Example

Consider the two networks in figure 8 and observe the role of node1 and node5. Network B is obtained by network A removing node5. Interference notion is based on measuring the effects of such remotion: in the table 4 are reported the betweenness values of the two networks (in percentage). Consider node1: in the network A, node1 has the 27% of the total value of betweenness. In the network B (where node5 has been removed) node1 has the 64% of the total value. The betweenness interference of node5 with respect to node1 is: Betwennes of node1 in the network A - Betwennes of node1 in the network = -37% It means that the topological relevance measured by the betweenness centrality of node1 increase of the 37% of the total betweenness if we remove node5 from the network. The presence of node5 negatively interferes with the central role of node1 measured by the betweennes centrality.

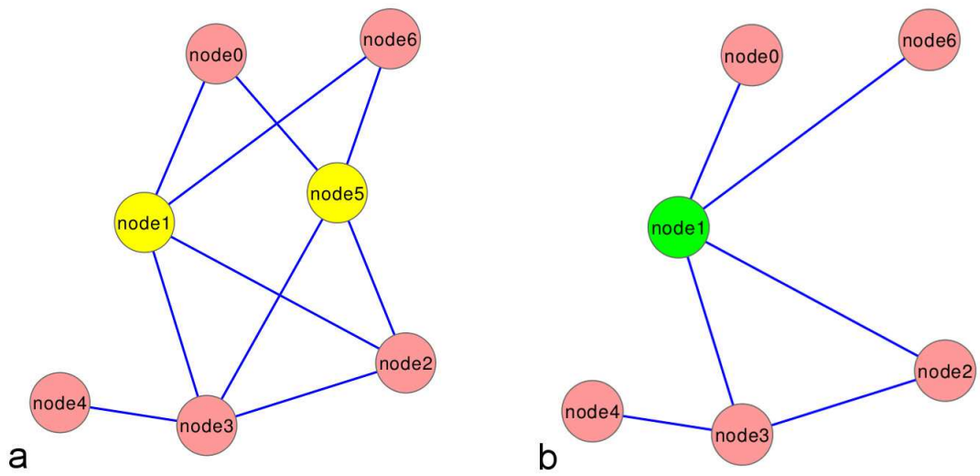


Fig. 8. Network b is obtained by network a removing node5. In this case node1 remains the only node connecting the top of the network with the bottom. Its betweenness values will increase.

Node	Network A (with node5) Betweenness (%)	Network B (no node5) Betweenness (%)	Interference value
node1	27%	64%	-37%
node4	0%	0%	0%
node0	2%	0%	2%
node2	2%	0%	2%
node6	2%	0%	2%
node3	40%	36%	4%
node5	27%		

Table 4. Betweenness and interference values expressed as percentage of the total value for the networks in figure 8. Node1 is the most sensitive to the deletion of node5.

The interference values of node5 with respect to the overall network are reported in the table. Node1 is the node mostly affected by the presence of node5 in the network. If we are considering real networks we expect that the activity of node5 strongly affects the activity of node 1.

Positive interference

If a node (A), upon removal from the network of a specific node (B) or of a group of nodes, decreases its value for a certain centrality index, its interference value is positive. This means that this node (A), topologically speaking, takes advantage (is positively influenced) by the presence in the network of the node (B) or of that group of nodes. Thus, "removal" of node (B) or of that group of nodes from the network, negatively affects the topological role of the node (A). This is called positive interference.

Negative interference

If a node (A), upon removal from the network of a specific node (B) or of a group of nodes, increases its value for a certain centrality index, its interference value is positive. This means that this node (A), topologically speaking, is disadvantaged (is negatively influenced) by the presence in the network of the node (B) or of that group the nodes. Thus, "removal" of node (B) or of that group of nodes from the network, positively affects the topological role of node (A). This is called negative interference.

5.2 Conclusions

As argued above, interference naturally induces cluster of proteins that are similar for their interference values due to the same node. A new clusterization algorithm can be derived if we group nodes depending on their interference value: given a node we compute its interference value and we put all the nodes having high interference in the same cluster. This interference-based modular decomposition of a network characterizes nodes for their answer to the inhibition (or adding) of a certain node in the network. If deletion of the node in a protein network is due to drug usage, the cluster of nodes having high interference value is the set of proteins where the drug has its greatest effects. In pharmacology this should permit to predict which proteins are more affected from the inhibition of another protein in the network. We can so prevent side effects of the inhibition of a node due to a drug usage.

6. References

Albert, R., Jeong, H. & Barabasi, A.-L. (2000). Error and attack tolerance of complex networks, *Nature* 406(6794): 378–382.

Arsenio Rodriguez, D. I. (2011). Characterization in silico of flavonoids biosynthesis in theobroma cacao l., *Network Biology* 1: 34–45.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium., *Nature genetics* 25(1): 25–29.

Assenov, Y., Ramirez, F., Schelhorn, S.-E., Lengauer, T. & Albrecht, M. (2008). Computing topological parameters of biological networks, *Bioinformatics* 24(2): 282–284.

Bader, G. D. & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks., *BMC Bioinformatics* 4(1).

Barabasi, A.-L. & Albert, R. (1999). Emergence of scaling in random networks, *Science* 286(5439): 509–512.

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

Bhalla, U. S. & Iyengar, R. (1999). Emergent properties of networks of biological signaling pathways, *Science* 283.

Biondani, Viollet, Foretz, Laudanna, Devin-Leclerc, Scardoni & Franceschi, D. (2008). Identification of new functional targets of *ampkα*1 in mouse red cells., *48th annual meeting of American society for cell biology*.

Choura, M. & Rebaï, A. (2010). Application of computational approaches to study signalling networks of nuclear and Tyrosine kinase receptors., *Biology direct* 5: 58+.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A. R., Vailaya, A., Wang, P.-L. L., Adler, A., Conklin, B. R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G. J., Ideker, T. & Bader, G. D. (2007). Integration of biological networks and gene expression data using cytoscape., *Nature protocols* 2(10): 2366–2382.

Crucitti, P., Latora, V., Marchiori, M. & Rapisarda, A. (2004). Error and attack tolerance of complex networks, *Physica A: Statistical Mechanics and its Applications* 340(1-3): 388 – 394. News and Expectations in Thermostatistics.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs, *Numerische Mathematik* 1: 269–271.

Feltes, B., de Faria Poloni, J. & Bonatto, D. (2011). The developmental aging and origins of health and disease hypotheses explained by different protein networks, *Biogerontology* 12: 293–308. 10.1007/s10522-011-9325-8.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness, *Sociometry* 40(1): 35–41.

Gilbert, D. (n.d.). http://www.jfree.org/jfreechart/.

Hu, Z., Mellor, J., Wu, J., Yamada, T., Holloway, D. & DeLisi, C. (2005). VisANT: data-integrating visual framework for biological networks and modules, *Nucl. Acids Res.* 33(suppl_2): W352–357.

Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. (2001). Lethality and centrality in protein networks, *Nature* 411(6833): 41–42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. (2000). The large-scale organization of metabolic networks, *Nature* 407(6804): 651–654.

Joy, M. P., Brock, A., Ingber, D. E. & Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network., *J Biomed Biotechnol* 2005(2): 96–103.

Junker, B., Koschutzki, D. & Schreiber, F. (2006). Exploration of biological network centralities with centibin, *BMC Bioinformatics* 7(1): 219+.

Kinexus (n.d.). http://www.kinexus.ca.

Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Podehl, D. T. & Zlotowski, O. (2005). Centrality indices, *in* U. Brandes & T. Erlebach (eds), *Network Analysis: Methodological Foundations*, Springer, pp. 16–61.

Ladha, J., Donakonda, S., Agrawal, S., Thota, B., Srividya, M. R., Sridevi, S., Arivazhagan, A., Thennarasu, K., Balasubramaniam, A., Chandramouli, B. A., Hegde, A. S., Kondaiah, P., Somasundaram, K., Santosh, V. & Rao, S. M. R. (2010). Glioblastoma-specific protein interaction network identifies pp1a and csk21 as connecting molecules between cell cycle-associated genes., *Cancer Res* 70(16): 6437–47.

Lepp, Z., Huang, C. & Okada, T. (2009). Finding Key Members in Compound Libraries by Analyzing Networks of Molecules Assembled by Structural Similarity, *Journal of Chemical Information and Modeling* 0(0): 091030094710018+.

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.

Newman, M. E. J. (2006). Modularity and community structure in networks, *Proceedings of the National Academy of Sciences* 103(23): 8577–8582.

Scardoni, G. & Laudanna, C. (2011). Interference: a tool for virtual experimental network topological analysis.
URL: *http://www.cbmc.it/ scardonig/interference/Interference.php*

Scardoni, G., Petterlini, M. & Laudanna, C. (2009). Analyzing biological network parameters with CentiScaPe, *Bioinformatics* 25(21): 2857–2859.

Schokker, D., de Koning, D.-J., Rebel, J. M. J. & Smits, M. A. (2011). Shift in chicken intestinal gene association networks after infection with salmonella., *Comp Biochem Physiol Part D Genomics Proteomics* .

Sengupta, U., Ukil, S., Dimitrova, N. & Agrawal, S. (2009a). Expression-based network biology identifies alteration in key regulatory pathways of type 2 diabetes and associated risk/complications., *PloS one* 4(12): e8100+.

Sengupta, U., Ukil, S., Dimitrova, N. & Agrawal, S. (2009b). Expression-based network biology identifies alteration in key regulatory pathways of type 2 diabetes and associated risk/complications., *PloS one* 4(12): e8100+.

Sengupta, U., Ukil, S., Dimitrova, N. & Agrawal, S. (2009c). Identification of altered regulatory pathways in diabetes type ii and complications through expression networks, *Genomic Signal Processing and Statistics, 2009. GENSIPS 2009. IEEE International Workshop on*, pp. 1 –4.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks., *Genome research* 13(11): 2498–2504.

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

Strogatz, S. H. (2001). Exploring complex networks, *Nature* 410(6825): 268–276.

Venkatachalam, G., Kumar, A. P., Sakharkar, K. R., Thangavel, S., Clement, M.-V. & Sakharkar, M. K. (2011). Pparγ; disease gene network and identification of therapeutic targets for prostate cancer., *J Drug Target* .
URL: *http://www.biomedsearch.com/nih/PPAR-disease-gene-network-identification/21780947.html*

Wagner, A. & Fell, D. A. (2001). The small world inside large metabolic networks., *Proceedings. Biological sciences / The Royal Society* 268(1478): 1803–1810.

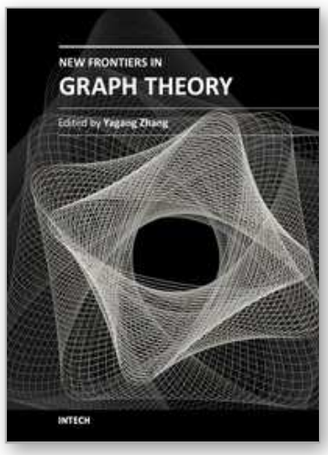
Watts, D. J. (1999). *Small worlds: the dynamics of networks between order and randomness*, Princeton University Press, Princeton, NJ, USA.

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks, *Nature* 393(6684): 440–442.

Webster, Y. W., Dow, E. R., Koehler, J., Gudivada, R. C. & Palakal, M. J. (2011). Leveraging Health Social Networking Communities in Translational Research., *Journal of biomedical informatics* .
URL: *http://dx.doi.org/10.1016/j.jbi.2011.01.010*

Wuchty, S. & Stadler, P. F. (2003). Centers of complex networks., *J Theor Biol* 223(1): 45–53.

Yamada, T. & Bork, P. (2009). Evolution of biomolecular networks: lessons from metabolic and protein interactions., *Nature reviews. Molecular cell biology* 10(11): 791–803.



New Frontiers in Graph Theory

Edited by Dr. Yagang Zhang

ISBN 978-953-51-0115-4
Hard cover, 526 pages
Publisher InTech
Published online 02, March, 2012
Published in print edition March, 2012

Nowadays, graph theory is an important analysis tool in mathematics and computer science. Because of the inherent simplicity of graph theory, it can be used to model many different physical and abstract systems such as transportation and communication networks, models for business administration, political science, and psychology and so on. The purpose of this book is not only to present the latest state and development tendencies of graph theory, but to bring the reader far enough along the way to enable him to embark on the research problems of his own. Taking into account the large amount of knowledge about graph theory and practice presented in the book, it has two major parts: theoretical researches and applications. The book is also intended for both graduate and postgraduate students in fields such as mathematics, computer science, system sciences, biology, engineering, cybernetics, and social sciences, and as a reference for software professionals and practitioners.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen