15

Topological Analysis of Cellular Networks

Carlos Rodríguez-Caso and Núria Conde-Pueyo
ICREA-Complex Systems Lab, Universitat Pompeu Fabra (GRIB-PRBB).
Dr Aiguader 88, 08003,
Barcelona, Spain

1. Introduction

The description of the molecular world conforming living cells has been a long standing enterprise since Biochemistry foundation. The elucidation of biochemical pathways in the early-middle twenty century gave way to a more complete picture of genes, proteins and metabolites by the beginning of Molecular Biology. Nowadays, the ultimate deciphering of such a molecular world is now becoming to be reality by the huge biotechnological advance on high throughput analysis. Genomic, proteomic and metabolomic tools have provided a revolution of the molecular biology and biomedicine expectations in very few years, as well as, the emergence of novel disciplines such as Systems and Synthetic Biology.

One of the first conclusions of such large-scale analyses is that molecular species are networked in giant interconnected entities. The so-called *cellular networks,* -consisting of protein maps, metabolism, and gene regulatory networks- but also other systems such as food webs, internet, or social relations constitute a sort of complex networks. Contrasting with the initial thought, it was observed that their organisation strongly departs from simple random homogeneous metaphors. Interestingly, their internal organization reveals common traits that can be analysed from the perspective of modern graph theory. In this theoretical framework, a graph is a mathematical abstraction of reality that can be tackled from statistical physics and computation science perspectives.

In this chapter we will present a repertoire of methods for a standard graph analysis, particularly oriented to the study of cellular networks. These tools allow us to measure and compare different networks in order to uncover their internal organization from a statistical point of view. We will show that the network approach provides a suitable framework to explore the organisation of the biomolecular world.

2. Graph theory approach

The aim of this chapter is not to present a collection of methods but an orientation about how is the network that we are studying by a description of the most relevant descriptors of a graph. We will start with describing those descriptors to define, in a topological way, an element within a network. In second place, we will provide global descriptors to define a network.

2.1 Graph concept

A *graph* (or *network*) G is defined by a set of N vertices (or nodes) $V = \{v_1, v_2, ..., v_N\}$ and a set of L edges (or links), $E = \{e_1, e_2, ..., e_L\}$, linking the nodes. Two nodes are linked when they satisfy

Source: Data Mining in Medical and Biological Research, Book edited by: Eugenia G. Giannopoulou,
ISBN 978-953-7619-30-5, pp. 320, December 2008, I-Tech, Vienna, Austria

a given condition, such as two metabolites participating in the same reaction in a metabolic network. The graph definition does not imply that all nodes must be connected in a single component. A *connected component* in a graph is formed by a set of elements so that there is at least one path connecting any two of them. Graphs are *undirected* when the interaction between nodes is mutual and equal, as in the protein maps. On the contrary, the web is *directed* when the connection indicates that one element affect to the other but not the opposite. As we will see, this is the case of gene regulatory networks (Shen-Orr et al. 2002) and signal transduction pathways (Ma'ayan et al. 2005). Additionally, graphs can also be *weighted* when links have values according to a certain property. This is the case for gene regulatory networks, where weights indicate the strength and direction of regulatory interactions. Although graphs are usually represented as a plot of nodes and connecting edges, they can also be defined by means of the so-called *adjacency matrix*, i.e., an array A of NxN elements a_{ij}, where a_{ij}=1 if v_i links to v_j and zero otherwise. A is symmetric for undirected graphs, but not for the directed ones. For weighted nets a matrix W can be introduced, where w_{ij} indicates the strength and type of the link. The network can also be described using a list of pairs of connected nodes (edge-list), which has some computational advantages. Figure 1 summarizes the different ways of representing a graph.

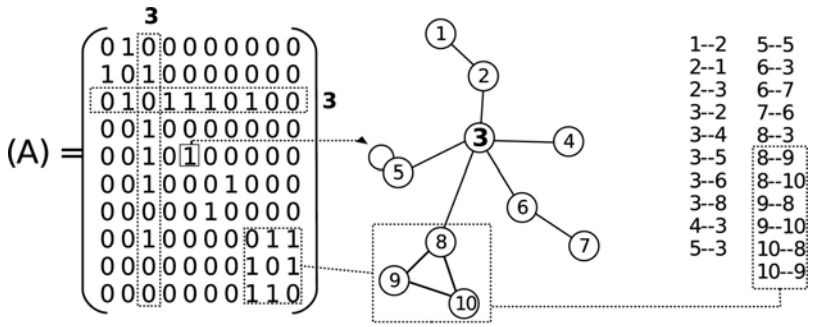

Fig. 1. Different ways of representation for a directed and unweighed graph. Left: Adjacency matrix (A). Centre: Drawn graph. Right: List of pairs (edge list). The triangle motif (in dashed box) is indicated for the three representations. The autoloop concept is represented in the vertex 5. Some examples of k, C and b values: for v_3, k_3=5, C_3=0, b_3=0.69; v_8, k_8=3, C_8=0.33, b_8=0.36; v_{10}, k_{10}=2, C_{10}=1,b_{10}=0.

2.2 Node attributes

Here we summarize the measures required to describe individual nodes of a graph. They allow identifying elements by their topological properties. The *degree* -or *connectivity*- (k_i) of a node v_i is defined as the number of edges of this node. From the adjacency matrix, we easily obtain the degree of a given node as

$$k_i = \sum_{j=1}^{N} a_{ij}$$

See examples of k values in figure 1. For directed graphs, we distinguish between incoming and outgoing links. Thus, we specify the degree of a node in its *indegree*, k_i^{in}, and *outdegree*, k_i^{out}.

The *clustering coefficient* C_i is a local measure quantifying the likelihood that neighbouring nodes of v_i are connected with each other. It is calculated by dividing the number of neighbours of v_i that are actually connected among them, n, with all possible combinations excluding autoloops, i.e., $k_i(k_i\text{-}1)$. Formally, we have:

$$C_i = \frac{2n}{k_i(k_i - 1)}$$

Notice that, auto-loops, i.e., links that starts and end in the same vertex (see figure 1), are not considered in this measure. Examples of C values are illustrated in figure 1.
The *betweenness centrality* b_m for a node v_m is the fraction of *shortest pathways* Γ for each pair of nodes (v_i, v_j) also containing v_m, that is

$$b_m = \sum_{i \neq j} \frac{\Gamma(i, m, j)}{\Gamma(i, j)}$$

The ratio $\Gamma(i,m,j)/\Gamma(i,j)$ indicates how crucial v_m is relating v_i and v_j. We introduce the term *pathway* (or simply *path*) as the string of nodes relating v_i and v_j (see graph and values for b in Figure 2). This concept is similar to the metabolic pathway describing a set of coupled reactions from one metabolite to another. The shortest path connecting v_i and v_j is the one where the lowest number of nodes are involved to connect them. Such topological descriptors are useful to identify particular nodes in the network. Under this point of view, such particularities can be mapped into relevant topological properties. For instance, high k_i for a node might relate to a relevant role, since many other nodes interact with it. Alternatively, high b_i can also indicate a relevant role since it tells us that many nodes are efficiently connected through it. It is noteworthy that, b_i usually scales with degree, although this is not always true (see figure 2).

2.3 Graph attributes

For a network of size N, global measures can be defined, each one providing very different, but complementary, sources of information. The *average degree*, defined as $\langle k \rangle = 2L/N$, indicates how sparse a graph is. Real networks are sparse, i.e. $\langle k \rangle << N$. In the case of networks with auto-loops the average degree must be corrected as $\langle k \rangle = (2L\text{-}A)/N$ where A corresponds with the number of auto-loops in the network.

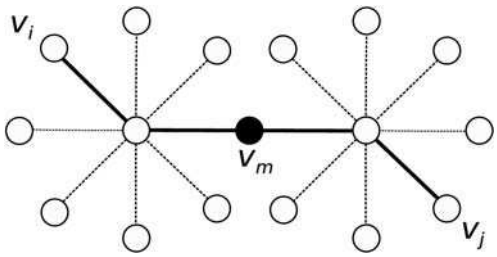

Fig. 2. Relation between degree and betweenness. The two-star graph shows a case where the two hubs support a high level of shortest pathways whereas the central node v_m shows the highest b of the graph keeping a low degree. The shortest path connecting v_i and v_j through v_m is indicated by solid lines.

The *average clustering*, $\langle C\rangle = 1/N\sum_i C_i$, provides a measure of local organization. High ⟨C⟩ indicates that neighbours of a node are likely to be linked between them. It actually gives the probability of finding triangles.

The *average path length* (*APL*) indicates the average length of the shortest pathways separating each node pair. If d_{min} is the length of the shortest path connecting nodes v_i and v_j, then *APL* is defined as:

$$APL = \frac{2}{N(N-1)}\sum_{i>j} d_{\min}(v_i, v_j)$$

Another measure is the *degree distribution* *p(k)*. It indicates the probability of a node having *k* links. Usually, because network size is restricted, the statistics are poor. It is rather difficult to get a good fitting for distribution degree from real data. A common problem in real networks is the fluctuations in the vertex abundance for very large degrees. One common solution, and in particular when we observe a power-law behaviour, is the cumulative distribution of degree frequency (Dorogovtsev & Mendes 2003), formally, $P_{cum}(k) = \sum_{k'=k}^{\infty} p(k')$.

Real networks are usually associated with the term of *scale-free.* They exhibit a degree distribution following a power-law decay, $p(k) \sim k^{-\gamma}$. Here, γ is a positive parameter that for real networks is usually in the range $2<\gamma<3$ (Albert et al. 2002). Notice that for cumulative distributions $P_{cum} \sim k^{-(\gamma-1)}$.

Scale-free (SF) graphs have a *p(k)* with a maximum at *k=1* (thus most elements have a single link) and rapidly decay at higher *k* values. Nevertheless, the tail of the distribution is very long and thus nodes with a very high degree are possible. Before the discovering of such an evidence, it was thought that real networks might follow a Gaussian distribution where the average degree represents a central position of a well confined distribution. The mathematical models describing this behaviour correspond with the Erdös-Renyi (ER) graph. ER graphs predict that very high *k* is exceedingly rare and unlikely to be observed at all. SF distributions have no humps and have extremely large standard deviations, which means that no confidence can be placed in a prediction of the number of links of any node sampled at random (Albert et al. 2002). Typically, real networks exhibit a mixed distribution, that is, a power-law with a sharp exponential cut-off determined by k_c in the expression $p(k) \sim k^{-\gamma} e^{-k/k_c}$ indicating that arbitrarily high degrees are not allowed (Amaral et al. 2000).

The *clustering distribution C(k)* represents C_i against *k*. ER and pure scale-free webs do not exhibit any dependency between C_i and *k*. By contrast, in so-called *hierarchical networks*, it has been associated with a decay of *C(k)* with inverse of the degree ($C \sim k^{-1}$) (Barabasi et al. 2004). This type of network exhibits modularity (nodes are preferentially linked inside clusters or modules). A *module* can be defined as a set of nodes in a connected component which tend to be more connected among them than with the rest of the network.

The *assortative mixing* (*r*) is a measure of the correlation among degrees in a graph, giving information about the likelihood to find linked nodes of a certain degree. This measure compares the correlation among degrees in the studied network (noted as G_R) with its *uncorrelated* counterpart. The expression for *r* can be obtained in (Newman 2002). Here we will only present an intuitive understanding of *assortativenness* concept. The value of *r*

ranges between -1 and 1. Here *r=0* indicates no correlation among degrees, as it occurs for example in ER graphs. Otherwise, most complex networks have been found to be *disassortative*, i.e., *r<0*, where higher degree nodes tend to be connected with lower degree ones rather than nodes with the same *k* (see Figure 3A). These networks display hubs that are not directly connected among them. It has been suggested that this situation confers network robustness (Maslov et al. 2002). When *r>0*, nodes with the same degree tend to be linked among them (see figure 3B) and the graph is called *assortative*.

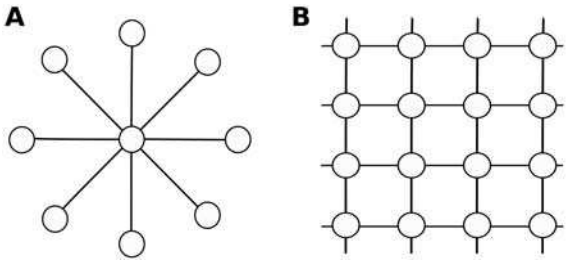

Fig. 3. Illustration of *assortativenness*. Panel A, a star graph, showing a correlation among highly connected nodes with poorly connected ones (*r<0*). Panel B, a lattice where all nodes have *k=4*. It is the extreme case where nodes with the same degree tend to be linked among them (*r>0*).

Small world pattern is a qualitatively property that exhibit most real networks. A small world criterion compares the clustering coefficient and *APL* of a real network with the respective ER model with the same average degree and size. ER graph constitutes a null model for comparison with real data. It captures the properties of a network derived from a purely random process of connection. The probability *P*, defines the likelihood that two vertices are linked among them. For ER graphs, $\langle k_{ER} \rangle = PN$, where N is the size of the network and $\langle C_{ER} \rangle = k/N$. *APL* follows the expression $APL_{ER} = \log N / \log \langle k \rangle$. When a graph G_R fulfils the conditions $APL_R \cong APL_{ER}$ but $\langle C_R \rangle << \langle C_{ER} \rangle$ then it is said that G_R exhibits a *small world* (SW) *pattern*. These networks keep their local order (high *C*) but also allow a very efficient communication (low *APL*) (Watts & Strogatz 1998).

2.4 Graph analysis and visualization software

For general purposes, the most popular visualization software is Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/), which is free for Windows operating systems. Pajek provides a graphic interface and a set of algorithms for graph analysis. Graphviz package (http://www.graphviz.org/) is another generic free package but it only provides visualization tools. Interestingly, a number of command line tools for complex network analysis for Linux/Unix platform can be found at http://www.lsi.upc.edu /~pfernandez/ software-networks.html.

Within the biological context, many databases offer a graph visualization of their content, for example KEGG database, or Transfac (http://www.biobase.de/) and Ingenuity (http://www.ingenuity.com) commercial databases.

The most interesting software for cellular network visualization and analysis is Cytoscape (http://www.cytoscape.org/). This software is supported by an open community where computer scientists can develop plugins for specific purposes: visualization methods, algorithms and the integration of the information from biological databases.

3. Cellular networks

Cellular network is the term commonly used for the current interacting molecular sets within cells (Albert, 2005; Barabasi & Oltvai, 2004). It includes mainly protein-protein interactions, metabolism, gene transcriptional regulatory networks and signal transduction pathways. All of them are different subsets of a single large-scale cellular network, since they are eventually cross-linked.

3.1 Protein-protein interaction networks

Protein-protein interaction (PPI) networks, interactomes and protein maps make reference to the collection of proteins interacting by physical contact. Proteins are the nodes and physical interactions among them are the links in the graph.

PPI networks are undirected graphs where two connected proteins are mutually affected. They exhibit a power-law decay with an exponential cut-off and small world behaviour. Interestingly, as it occurs in most cellular networks, vertices do not represent an individual but a molecular species. For this reason the appearance of auto-loops is justified since they represent the ability to make homo-multimers.

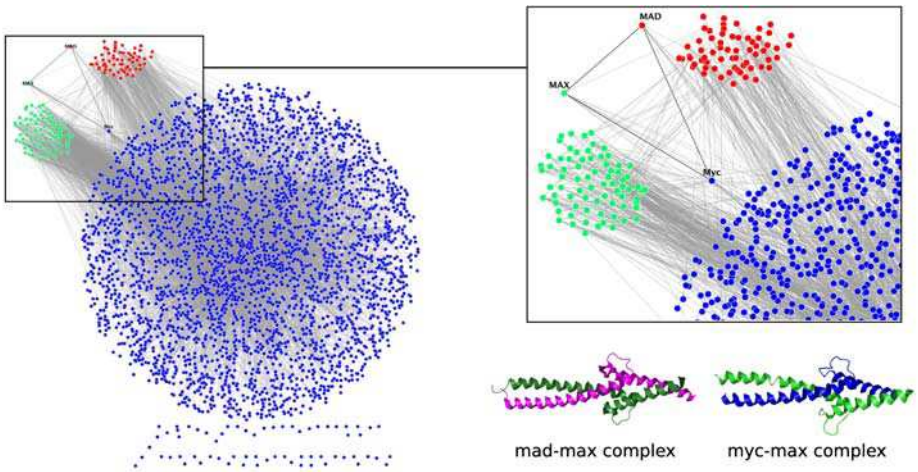

Fig. 4. Fraction of human PPI network filtered by nuclear localization criteria using Gene Ontology annotation (http://www.geneontology.org/). All proteins and interactions are expected to be in the nucleus. In red are represented those proteins marked as transcriptional co-suppressors. In green, transcriptional co-activators. As an example of interaction, the relation between mad/max and myc/max transcriptional cancer related protein complexes are depicted. Notice that databases may contain artefacts not observed in nature (e.g. myc/mad interaction). Below zoom box, protein complex representation from crystal structure. Data obtained from HPRD database. Graph generated with Cytoscape.

Measures such as clustering coefficient or average degree do not consider such a circumstance. This can be a source of errors in our analysis depending on how the measures have been implemented. To be consistent with the theory we must avoid auto-loops for such measures.

At low scale, *cliques*, i.e. full connected subgraphs within the network, constitute a way to complex protein detection (Yu, et al. 2006). The smallest clique that can be observed is the triangle, suggesting a possible hetero-trimer complex. However, the biological conclusions derived from a specific network configuration at the scale of a very few number of elements must be contrasted with different information sources. In general, this approximation must be considered as a methodology for the inference of potential biological relations among proteins to be tested experimentally. We must remain that, in spite of the analysis of different PPI networks reveals robust results in their global parameters, database information can contain artefacts. This is relevant when our aim is to focuss on functional/biological relations of a particular part of the network. After a first identification by automatic filters, a manual curation of the database for our study system is recommended (see figure 4).

To this day, large-scale studies have explored the proteome structure in viruses (McCraith et al., 2000), yeast (Uetz et al., 2000; Ito et al., 2001; Ptacek et al., 2005), the worm *Caenorhabditis elegans* (Walhout et al., 2000; Li et al., 2004), *Helicobacter pylori* (Rain et al., 2001), Drosophila melanogaster (Giot et al., 2003) and more recently in humans (Rual et al., 2005; Stelzl et al., 2005). Protein map elucidation is obtained mainly by two large-scale experimental approaches, namely, the yeast two-hybrid (Y2H) (Uetz & Hughes, 2000) and the tandem affinity purification (TAP) followed by mass spectroscopy (Gavin et al., 2002). Such information is collected in annotated databases. Different databases such as MIPS (http://mips.gsf.de/), DIP (http://dip.doe-mbi.ucla.edu/), Intact (http://www.ebi.ac.uk/intact/site/index.jsf) and in particular for humans HPRD (www.hprd.org/) are the main repositories commonly used for the acquisition of current protein maps.

3.2 Gene transcriptional regulatory networks

The assembly of regulatory interactions linking transcriptions factors (TFs) to their target genes constitutes the first level of a multilayered network of gene regulation; the so called gene transcriptional regulatory networks (GTRN) (Babu et al. 2004). Genome scale approaches have provided a reliable picture of the regulatory maps for the prokaryote *Escherichia coli* (Thieffry et al. 1998; Shen-Orr et al. 2002) and the eukaryote *Saccharomyces cerevisiae* (Lee et 2002; Balaji et al. 2006). Directed graphs are the mathematical abstraction of GTRNs (Babu et al. 2004, Albert et al. 2005). The regulatory effect of a TF gene (let's say A) on a specific target gene (B) is depicted by ($A \rightarrow B$). In graph theory, A y B are vertices linked by an arrow. TFs are easily identified in the graph since they exhibit outgoing arrows. In turn, non TF genes -the target ones- only receive arrows from the TF set. The number of outgoing links of a vertex is known as *outdegree* (denoted by k^{out}) whereas the number of incoming edges corresponds with *indegree* (k^{in}). Interestingly, as a TF can be a regulatory target of other TFs, they can exhibit both incoming and outgoing arrows.

As it is occurs with PPI networks, we can find auto-loops. In this case, it means that a gene product causes a regulatory effect in its own promoter. Interestingly, the identification of network motifs in GTRN remarks the view of minimal genetic circuits as the building blocks of the networks (Shen-Orr et al. 2002; Milo et al. 2002). However, the Achilles heel of this approach is that motif analysis is restricted to previous selection criteria by the investigator which specifically must define the subgraph to be detected.

3.3 Metabolic networks

Metabolism is the best described cellular network so far. However, a global topological view of metabolism was not available until recently (Jeong et al., 2000; Ouzounis & Karp, 2000).
Metabolic pathways are composed by two types of molecular species: enzymes and metabolites. In this case, one or more than one metabolites (substrates) are transformed (in products) by enzyme mediation. The resulting graph is known as *bipartite graph*, since one type of vertices (metabolites) is always related through the other type of elements (enzymes). Therefore, no enzyme-enzyme and metabolite-metabolite interactions are found. This network definition allows defining an arrow from substrates to enzymes and from enzymes to products for irreversible reactions. However, arrow definition is not possible for reversible reactions. In spite of this graph definition is the most informative, its topological treatment results more complicated, and the graph is usually *projected* over a single type of vertex. As figure 5 shows, two types of projections can be done (Wagner & Fell, 2001). One way is considering the *substrate graph*, where each metabolite is a vertex that will be linked with those metabolites participating in the same reaction. Alternatively, a *reaction graph* is made by considering reactions as nodes and metabolites as links. This mathematical treatment has permitted to uncover the scale free (Jeong et al. 2000), small world behaviour and the hierarchical and modular organization of metabolic networks (Wagner & Fell 2001, Ravasz et al. 2002). Metabolic pathways can be found in KEGG (http://www.genome.jp/kegg/) and Reactome database (http://www.reactome.org/).

3.4 Cell signalling networks

These networks depict those processes allowing cells integrating responses to external stimuli. They are a combination of metabolic reactions and protein interactions that trigger specific changes in gene expression. Protein modifications such as phosphorylation, acetylation and ubiquitination, among others, lead to conformational changes allowing ligand-protein recognition and functional protein complexes assembling. At the present, kinases and phosphatases relations constitute the best described signalling pathways. Bibliographic sources provide the current information to reconstruct this kind of networks (Ma'ayan et al. 2005). Additionally, several databases compile this information such as the *Kinbase (*http://kinase.com/) and Reactome databases. Interestingly, this kind of networks presents a diverse type of vertices and type of connections. By this reason, its biological interpretation of topological analysis is not trivial.

3.5 Filtered networks

Network analysis can be focussed on a sub-part of the system. Figure 4 illustrates an example of this. Gene ontology annotation provides biological information about function and localization of genes. However, depending on the particular process to be considered, the heterogeneity in the quality of the gene annotation constitutes a bias. In agreement with this philosophy, several works have provided relevant biological insights about the biological meaning of the network organization (Rodríguez-Caso et al. 2005, Ravazs et al. 2002).

3.6 Feature based networks.

As we have seen, several cellular networks offer a picture that captures the biological machineries within a living cell. We observe that all of them are constructed by a well defined type of interaction. The link features that two elements are involved by physic contact (PPI and GTR networks) or transformation process (metabolism).

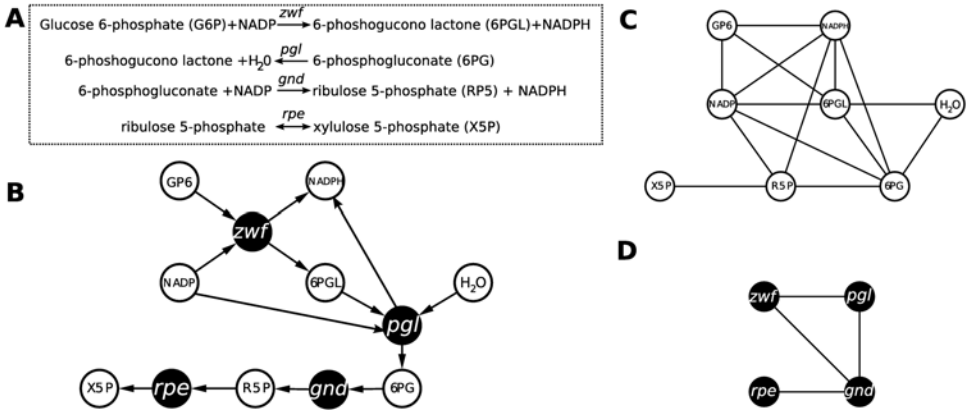

Fig. 5. Metabolic network representations (picture modified from Wagner & Fell 2001). Panel A, description of the reactions. Panel B, bipartite graph representation. Panel C, substrate projection. Panel D respective reaction projection. White vertices represent metabolites whereas black vertices represent enzymes.

Recently, network approach has been applied to define a sort of networks that captures relations in a broader sense. The purpose of these networks is not to describe truly molecular machinery but to offer a global view of some type of biological property, function or consequence. This is the case of the human disease network (Goh et al. 2007) that relates the diseases contained in OMIM database with their responsible genes. As it occurs with metabolic networks, this constitutes a bipartite graph with two kinds of entities, genes and diseases. This network, more than recovering a biological process, give us a conceptual picture of the relation between genes and diseases.

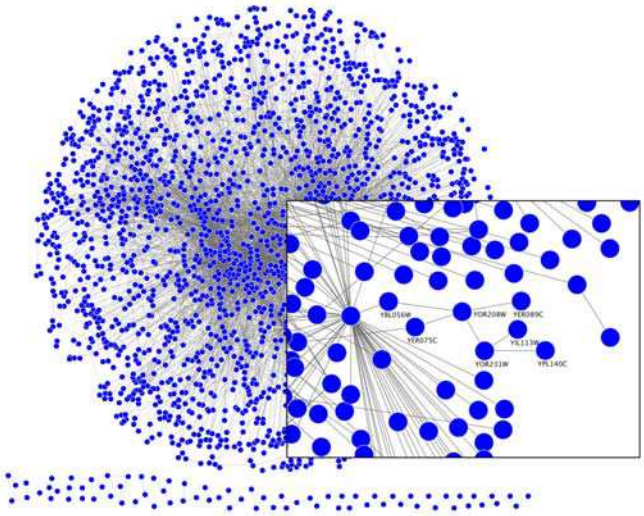

Fig. 6. Example of a feature based network. Yeast synthetic lethal network obtained from BioGRID database (http://www.thebiogrid.org/). Graph generated with Cytoscape.

In this direction, the same group goes beyond, constructing a bipartite graph composed of US Food and Drug Administration–approved drugs and proteins linked by drug–target binary associations. This drug-target protein network does not capture any biological machinery but offer a global picture of the relation between drugs and targets by the conceptualization of the problem in a graph. As another example, figure 6 illustrates the case of Synthetic Lethal network in yeast (Tong et al. 2001), that recovers the information of those pair of gene that simultaneously mutated lead to lethality but not when they are individually mutated.

4. Topological analysis of cellular networks

Since cellular networks are in constant change, here we present the *state of the art* of different cellular networks. The topological analysis is based on the previously described estimators. Table 1 summarizes the topological analysis of PPI, metabolic and gene regulatory networks. In addition, we have included the yeast synthetic lethal network as an example of featured base networks. All these networks present a single giant component and a number of very small subgraphs. The statistics are provided for the giant components. In general, these networks are sparse graphs. Remarkably, all the networks are dissasortative *(r<0)*, i.e, high degree tends to be connected with lower degree. Small world behaviour is clearly evidenced in yeast and human PPI networks. Interestingly, these two networks differ in their size but present a high similarity in their organization. The two available GTRN present a very small *APL*. The explanation if found in the biology, since only a small fraction of genes are TFs. The major fraction of vertices corresponds with terminal genes linked to one of these factors. These gene regulatory networks differ in their <C>, in other words, in their local organization.

	N	***L***	***<k>***	***<C>***	***(<C_{ER}>)***	***APL (APL$_{ER}$)***	***r***	**Source Data**
Human PPI *	9048	34876	7.71	0.16	($8.52\ 10^{-3}$)	4.26 (4.46)	-0.04	HPRD
Yeast PPI *	4842	17119	7.07	0.10	($1.46\ 10^{-2}$)	4.14 (4.34)	-0.13	DIP
E.coli GTRN **	1589	4030	5.07	0.43	($3.19\ 10^{-2}$)	2.68 (4.54)	-0.26	RegulomDB6.0
Yeast GTRN **	4441	12864	4.79	0.08	($1.08\ 10^{-2}$)	3.49 (5.36)	-0.59	Balaji *et al.* 2006
Human metabolism	2827	5988	4.23	0.00	($1.5\ 10^{-3}$)	4.55 (5.50)	-0.12	KEGG
Yeast SL *	2287	9616	8.34	0.30	($3.67\ 10^{-2}$)	3.75 (3.65)	-0.19	BioGRID

Table 1. Global descriptors for the giant component of cellular networks. Notice that, for all the cases, giant component represents almost the total number of interactions. Parenthesis shows the <C> and APL values are showed for respective ER counterparts (calculated according definition described in the text). (*) It indicates a small world pattern. (**) Notice that APL is revealed shorter than the expected in ER model. Human metabolism presents <C>=0 due to the bipartite nature of the graph. Graph descriptors were calculated by Gstats command line software available at http://www.lsi.upc.edu/~pfernandez/software-networks.html. Autoloops were eliminated for the topological analysis.

Metabolic network corresponds with the bipartite representation. This imposes a restriction on the clustering coefficient. Since two vertices of the same nature cannot be connected, $\langle C \rangle = 0$ by definition.
In spite of SL network has not a biomolecular machinery correlate, it reveals a small world pattern indicating that SL are not trivially organised. In this case, high clustering is interpreted as two synthetically lethal genes tend to make a synthetic interaction with a common third gene.
It is remarkable that some relevant properties of these examples can be explained by the consideration of the network definition. This suggests that a suitable knowledge of the study of the system besides graph theory approach provides the best system study comprehension.

5. Goals and pitfalls of network approach

Uncovering the molecular world constitutes the new frontier of biology. Large zoological and botanical expeditions at the end of nineteen century pursued the characterization of organism diversity and their relations. Nowadays, in a similar way, the molecular biologist explores the diversity inside the cell. Unfortunately, the current picture of the study system is only a sketch of the actual relations between elements and most of the biological details are still unknown. Precisely, the relations among elements are the target for graph theory approach that has been profusely applied in many real systems. During the last decade, graph view has been incorporated to a diverse number of disciplines. This approach opens the possibility of a global comprehension of the system, against the predominant reductionism of the current scientific thought. We can access to the study of very large systems even when we do not know the details. Pioneer works about scale-freeness in metabolism, proteome (see the review, Albert 2005), the diameter of the world wide web (Albert et al. 1999), well as the widely observed small world behaviour in real networks have demonstrated that the pattern of interactions encloses relevant constraints defining the internal organisation of networks.
Graph theory enables a systemic study through the statistical approximation from the collection of local interactions; nevertheless, a limitation of such a global understanding is precisely its own size. In general for any statistical approach, the larger size of our data the more reliable is the statistics. This is not an exception for the global estimators of graph theory such as degree distribution, or assortativeness. From a theoretical point of view, the graph properties derived from analytical models are established when graph size tends to infinite. Therefore, if our study system is not large enough, deviations from the theory are expected.
In any case, we must remain that the true understanding of our study system will be only successful if we exactly know what is the captured from the reality in our graph abstraction and what is not. A graph is constructed by considering some particular property that is used to link a set of elements. Both of them -elements and their relation type- must be clearly defined. Most probably, graph definition does not affect to the topological analysis but it is essential for its biological interpretation that is, in the last instance, the aim of the biologist.

6. Acknowledgements

This work has been supported by 6th EU framework SYNLET (NEST-043312), ComplexDis (NEST-043241) and NHI CA 113004 projects. We thank Dr Ricard Solé and Complex

Systems Lab members for successful comments. We thank Itziar Castanedo for her successful comments during the writing of this work.

7. References

Albert, R.; Yeong H. & Barabasi, A. L. (1999). Diameter of the world-wide web. *Nature* 401 (September 1999) 130.

Albert, R. & Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys,* 74, (January 2002) 47-97

Albert, R. (2005). Scale-free networks in cell biology. *Jcell Sci*, 118, (Pt 21) (October 2005) 4947-57.

Amaral, L.A.N.; Scala, A.; Barthél'emy, M. & Standley H.E. (2000). Classes of small-world networks. *Proc. Natl. Acad Sci*. USA, 97, 21, (September 2000) 11149-11152.

Babu, M.M.; Luscombe N.M.; Aravind, L.; Gerstein, M. & Teichmann, S.A. (2004). Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* 14, (2004 Jun) 283-291.

Balaji, S.; Babu, M.M.; Iyer, L.M.; Luscombe, N.M. & Aravind, L (2006). Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J Mol Biol,* 360, (June 2006) 213-27.

Barabasi, A.L. & Oltvai, Z.N. (2004). Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet*, 5, (February 2004) 101-13.

Dorogovtsev, S.N. & Mendes J.F.F. (2003). *Evolution of Networks. From Biological Nets to the Internet and WWW.* Oxford University Press ISBN 0-19-851590-1, Oxford UK.

Gavin, A. C.; Bosche, M & Krause, R. et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415.; (January 2002) 141-147

Giot, L.; Bader, J.S & Brouwer, C. et al. (2003). A protien interaction map of drosophila melanogaster. *Science*, 302, 5651, (December 2003) 1727-1736

Goh, K. I.; Cusick, M.E.; Valle, D.; Childs, B.; Vidal, M and Barabási, A.L. The human disease network. *Proc Natl Acad Sci USA,* 104, 21, (May 2007) 8685-8690

Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori , M. & Sakaki, Y. (2001) . A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, 98, 8, (April 2001) 4569-4574

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

Lee, T. I.; Rinaldi, N. J & Robert, F. et al. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Sciencie*, 298, 5594, (October 2002) 799-804

Li, S.; Armstrong, C.M.; Berint, N.; Ge, H.; Milstein, S.; Boxem, M.; Vidalain, P.O & et al. (2004). A map of the interactome network of the metazoan C.elegans. *Science,* 303, 5657, (January 2004) 540-543

Ma'ayan, A.; Jenkins, S.L.; Neves, S.; Hasseldine, A.; Grace, E.; Dubin-Thaler, B. & et al. (2005). Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science*, 309, 5737, (August 2005) 1078-1083

Maslov, S. & Sneppen K. (2002). Specificity and stability in topology of protein networks. *Science*, 296, 5569, (May 2002) 910-913

McCraith, S.; Holtzman, T.; Moss, B. & Fields, S. (2000). Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc. Natl. Acad. Sci. USA,* 97, 9, (April 2000) 4879-4884

Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D. & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science,* 298, 5595 (October 2002) 824-7

Newman, M. E. (2002). Assortative mixing in networks. *Phys Rev Lett*. 89, 20, (Novembre 2002)

Ouzounis, C. A. & Karp P.D. (2000). Global properties of the metabolic map of escherichia coli. *Genome Res.* 10, 4, (April 2000) 568-576

Ptacek, J.; Devegan, G.; Michaud,; G. Zhu,; H. Zhu,; X. Fasolo,; J. & et al. (2005). Global analysis of protein phosphorylation in yeast. *Nature*, 438, 7068, (December 2005) 679-84

Ravasz, E.; Somera, A.L.; Mongru, D.A.; Oltvai, Z.N. & Barabasi, A.L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297, 5586, (August 2002) 1551-1555

Rodríguez-Caso, C.; Medina, M.A. & Solé, R.V. (2005).Topology, tinkering and evolution of the human transcription factor network. *FEBSJ* 272 (December 2005) 6423-34.

Rual, J. F.; Venkatesan, K.; Hao, T.; Hirozane-Kishikawa, T.; Crico, A.; Li, N.; Berriz, G. F. & et al. (2005).Towards a proteome-scale map of the human protein-protein interaction network. *Nature,* 437, 7062, (October 2005) 1173-1178

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

Stelzl, U.; Worm, U.; Lalowski, M.; Haening, C.; Brembeck, F.H.; Goehler, H. & et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 12, 6, 957-968.

Tong, A.H.; Evangelista, M.; Parsons, A.B.; Xu, H.; Bader, G.D.; Pagé, N.; et al.(2001).Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*. 294, 5550 (2001 December) 2364-8

Thieffry, D.; Huerta, A.M.; Pérez-Rueda, E. & Collado-Vides, J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *Bioessays*, 20, 5, (1998 May) 433-40

Uetz, P. & Hughes, R. E. (2000). Systematic and large-scale two-hybrid screens. *Curr. Opin. Microbiol.* 3, 3, (June 2000) 303-308

Uetz, P.; Giot, L.; Cagney, G.; Mansfield,T.A.; Judson, R.S.; Knight, J.R.; & et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403, 6770, (February 2000) 623-627

Wagner, A. & Fell, D. A. (2001). The small world inside large metabolic networks. *Proc. Biol. Sci.* 268, 1478, (September 2001) 1803-1810

Walhout, A.J.; Boulton, S. J. & Vidal, M. (2000). Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast,* 17, 2, (June 2000) 88-94

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

Yu, H.; Paccanaro, A.; Trifonov, V. & Gerstein, M. (2006). Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 22, 7, (February 2006) 823-829

Data Mining in Medical and Biological Research

Edited by Eugenia G. Giannopoulou

ISBN 978-953-7619-30-5
Hard cover, 320 pages
Publisher InTech
Published online 01, November, 2008
Published in print edition November, 2008

This book intends to bring together the most recent advances and applications of data mining research in the promising areas of medicine and biology from around the world. It consists of seventeen chapters, twelve related to medical research and five focused on the biological domain, which describe interesting applications, motivating progress and worthwhile results. We hope that the readers will benefit from this book and consider it as an excellent way to keep pace with the vast and diverse advances of new research efforts.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Carlos Rodríguez-Caso and Núria Conde-Pueyo (2008). Topological Analysis of Cellular Networks, Data Mining in Medical and Biological Research, Eugenia G. Giannopoulou (Ed.), ISBN: 978-953-7619-30-5, InTech, Available from: http://www.intechopen.com/books/data_mining_in_medical_and_biological_research/topological_analysis_of_cellular_networks

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