Review on the construction and analysis of disease networks

GIANLUCA BIASIOLO[1], RICCARDO CARANGELO[2], AND ANNA PASSARELLI[3]

[1] *University of Padova, Department of Information Engineering (DEI), M.S. Computer Engineering Student, mail: gianluca.biasiolo@studenti.unipd.it, ID: 1206433.*
[2] *University of Padova, Department of Biology, M.S. Industrial Biotechnology Student, mail: riccardo.carangelo@studenti.unipd.it, ID: 1190173.*
[3] *University of Padova, Department of Information Engineering (DEI), M.S. Computer Engineering Student, mail: anna.passarelli@studenti.unipd.it, ID: 1206153.*

Last version June 18, 2020

The target of this review is to analyze the issue of the disease networks and specifically to consider different construction approaches and understand the main features and difficulties related to the networks. In the first part we provide an introduction and a general overview on the topic, then we present some fundamental parameters used to analyze and characterize the various networks; after that we focus on the main methods used to construct a specific disease network. The reported construction methods take on account different aspects of the same problem and allow us to do different studies. Finally, we present some applications of the disease networks in recent discoveries, showing the main related problems.

Keywords: ***Disease Network, Phenotypic Disease Network, Human Disease Network, Metabolic Disease Network, micro RNA, miRNA, Cancer.***

1. INTRODUCTION

Researchers started almost fifty years ago to speculate and formulate a new theoretical model, a new perspective to look at the organization of the cell. Indeed, it changed from an unorganized set of enzymes to a new highly connected web of organelles. So, cells can be seen as integrated webs of macromolecular interactions: in other words, an "interactome" network [1].

As shown in Figure 1, the interactome network theory is applied in many different scenarios related to the cellular systems. For example, in the protein-protein interaction network each node represent a protein, while each edge stands for a physical interaction among two proteins. Also, in the virus-host network, the nodes represent viral proteins, while the edge represents the interaction among them.

In this context, a disease network is a different kind of interactome network that is useful to describe the relationship between different diseases and their associated cellular components, not only disease genes. A disease is not only due to the mutation of a single gene, but it is often the consequence of an alteration of the complex intracellular and intercellular network. The number of functional interactions between the various cellular components is very high. This means that the impact of a genetic perturbation is not only limited to the activity of the gene product that has the abnormality, but can also alter the activity of the other gene products

to which it is connected in the interactome.

Disease networks can have several biological and clinical applications. Thanks to the analysis of these networks, we may discover the molecular cause of diseases, and also the relationship between apparently unrelated phenotypes. Moreover, these studies may help to identify new disease genes and disease pathways and to define better target for drug development. They may also lead to a better disease classification and to the discovery of more accurate biomarkers, which are used to evaluate the functional status of the complex biological network.

Before explaining disease networks, it is important to note that, although great efforts have been made in recent years to increase the coverage of human interactome maps and correct some known biases, such maps are still incomplete and inaccurate. Furthermore, much research up to now has focused on intracellular interaction, ignoring intercellular interaction due to lack of systematic data.

A. Disease and network theory

Biological networks are organized according to some typical principles of network theory. Observing diseases from this point of view allows us to understand some characteristic properties of the genes that are associated with them. A known property of biological networks is the presence of a few highly connected nodes, called hubs. This suggests that the proteins represented by these hubs must have an important biological function. In effect, it turned out [2] that hub proteins tend to be encoded by essential genes, which are genes whose absence causes embryonic lethality, and that non-essential disease genes, which represent the majority of disease genes, tend to avoid hubs and locate at the functional periphery of the interactome.

Another property is the local hypothesis, according to which proteins that are involved in the same disorder show a high tendency to interact with each other [2]. The same behaviour is observed also between cellular components associated with a particular disease phenotype. This means that disease components are expected to be in the same network neighbourhood. Therefore, we say that they are located in a well-defined portion of the interactome, called disease module. Each disease module is defined on the basis of a specific disease, so each disease has its own disease module. However, cellular components (genes,proteins, metabolites or microRNAs) can be involved in many diseases and, accordingly, different disease modules can overlap.

In addition to the disease module, there are two other types of modules: the topological module and the functional module. The topological module represents a dense set of nodes that tend to interact more with each other rather than with nodes outside the module and it can be identified with clustering algorithms which group independently of the node function. Instead, the functional module is a set of nodes with similar or related function, situated in the same network neighbourhood. The function represents the role of the component in defining phenotypes. The disease module often overlaps with the other two, but it may not coincide with them. However, in many biological works there is the assumption that these three structures are equivalent, that is cellular components in a topological module have related functions whose perturbation (mutations, deletions, copy number variations or expression changes) results in a disease.

Since different disease modules can overlap, the alteration of one disease module can affect other disease modules and so diseases are not independent one another. The dependencies between different pathophenotypes and their disease modules are represented in the disease networks, which are maps whose nodes are diseases and whose edges are molecular relationships between the cellular components associated with the disease. Studying such links between diseases is very important for different aspects. First, it helps us to understand how different phenotypes, often studied by different medical branches, are connected at the molecular level. Second, it helps us to comprehend

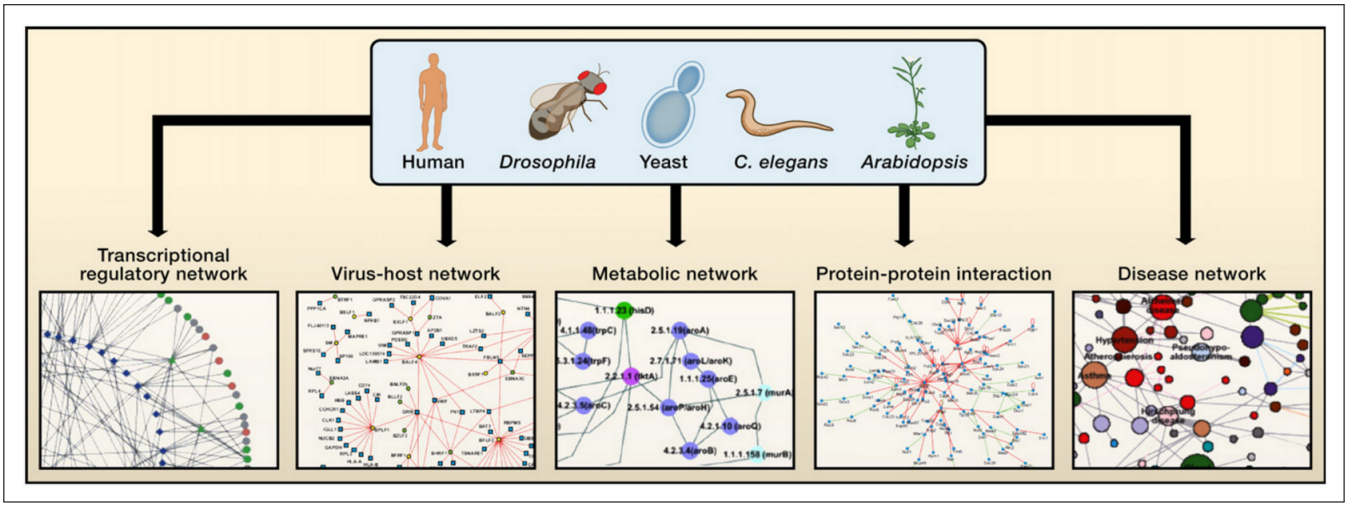

Fig. 1. The main applications of the network systems into the cellular systems horizon [1]

why some disorders often arise together. Hence, these tools may lead new approaches to disease prevention, diagnosis and treatment.

B. Biological concepts

The stunning discovery of the structural organization of the DNA molecule led by Watson and Crick in 1953 shocked and let the scientific world to get in touch with the only known molecule capable of self-replication. DNA (deoxyribonucleic acid) is a huge polymer molecule which carries a series of genetic instructions useful to carry out the entire set of functions and developmental tasks of almost every known organism.

The most important functional component of a DNA chain is the gene. The whole set of genes characterizing an organism is called genotype. A gene is a unique portion of DNA which includes a series of functional subregions aimed at initiating and managing the transcription process. The transcription process can transfer the information enclosed in the gene to an RNA equivalent, called mRNA (messenger RNA); this newly produced mRNA goes through a maturation process aimed at obtaining a mature RNA, which is then used to produce the protein by means of a specific translational machinery. In this way the genes in DNA can encode all proteins, used to perform most functions, to correctly build an organism. The overall observable features resulting from this process allow an organism to develop a specific morphology, a specific physiological and biochemical profile and a specific way to interact with the environment. The combination of all these features is called phenotype and represent the final and very complex result of the synergy between the expression genotype and the surrounding environment [3].

The RNA is not only used to carry the information needed to encode proteins, but there are RNAs with fundamental regulation roles, such as miRNAs (microRNA) and siRNAs (small interfering RNA), which can induce or inhibit the transcription of a particular gene. As it is clear from this explanation, every organism contains a complex and strictly connected system of functioning, whose failure in one or more points is responsible of several issues which can alter the phenotype and generate malfunctioning, commonly called diseases [3].

Another important concept which will be seen in this paper is the locus heterogeneity. In general, a genetic heterogeneity phenomenon arises when the same phenotype is related to different genetic events [4]. In particular, we have locus heterogeneity when different mutations taking place in distinct genomic loci cause the same phenotype. This concept can describe how the same disease phenotypic condition is given by a series of mutations in different versions of the same gene [5].

2. NETWORK ANALYSIS

A network can be a very complex structure in terms of topological profile. Therefore, studying the topology of a network is far from a simple problem. As we will see below, although there are many ways to analyze a network depending on which feature we want to examine, there are several basic parameters that can be considered essential to each kind of analysis. Indeed, using these parameters, it is possible to design many different topological studies aimed at inspecting specific properties of a given network, in order to understand the real extent of its deviation from a corresponding random network and thus to demonstrate the likelihood of a particular property. Below we define and describe several basic network properties and a series of basic parameters used to study and quantify these properties.

The main properties that we are going to explain in the following paragraphs are the modularity, centrality and assortativity [6] [7].

A. Modularity

Before introducing the concept of module, we need to explain another module-related fundamental feature: the cluster. A cluster is a subgraph which presents a high level of connections and often constitutes a functional biological module. The functional modularity is the capability of the network to be segmented into dense regions that are sparsely interconnected; this segmentation represents a sort of logical divisions inside the network. Indeed, a module is a distinct subgraph component which contains elements that are functional related to each others and form together a functional distinct area. A very important example of modularity is given by all the single metabolic biochemical networks in the biochemical pathways map [8]. In this map we can see how, for example, the glycolysis forms a distinct functional module containing all the needed components to complete the glucose catabolic process [7]. One of the most important parameters used to study a network is the clustering coefficient, which is used to quantify how much each node is grouped in a network. Thus, this parameter is strictly connected with the concept of clustering and it is largely used to create models for clusterization and modularity analysis. It is possible to define a clustering coefficient for every node in a network (local clustering coefficient). For a undirected graph $G = (V, E)$ (where V is the set containing all the nodes, while E is the set containing all the edges), it is calculated as follows:

$$C_i = \frac{2|e_{jk} : v_j, v_k \in N_i, e_{jk} \in E|}{k_i(k_i - 1)} \tag{1}$$

where e_{jk} is the edge which connects two nodes j and k that are neighbors of i (i.e. that are directly connected with i); N_i is set containing all the neighbors of i (i.e. the set containing all the nodes directly connected with i) and k_i is the cardinality of N_i. This measure represents thus a quantitative level of connectivity for i, taking account of its neighborhood. For a directed graph the formula is similar:

$$C_i = \frac{|e_{jk} : v_j, v_k \in N_i, e_{jk} \in E|}{k_i(k_i - 1)} \tag{2}$$

What we have just seen are the local clustering coefficients, measures that interest the single nodes. By computing the arithmetic mean of all the clustering coefficients of all the nodes in the network, it is possible to get a global clustering coefficient:

$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n} C_i \tag{3}$$

This formula represents the basic approach to a study that is interested in finding modules and studying the properties of modularity within the network [7] [9] [21].

B. Centrality

The centrality is a characteristic that regards the single node and quantifies the position of the node in relation to the center and the periphery of the network.

The simplest measure used to study the centrality is the degree k, which indicates the number of nodes connected with an interested node.

A high degree means a node nearer to the center compared to the nodes with a lower degree.

Another relevant parameter is the degree distribution $P(\deg v = k)$, which represents how many nodes have a specific degree and can be calculated for every degree in the network. The degree distribution can be used to plot a curve that describes the network topology and gives a profile of the connectivity of the network. The distribution curve obtained from this kind of study allows us to recognize the possible presence of specific topological models in the graph. Every model has its own specific formula. Here we present, as an example, two formulas to calculate the degree distribution in two different network models:

For the Erdős – Rényi model random networks (which follows a binomial distribution) [25]:

$$P(\deg v = k) = \binom{n-1}{n} p^k (1-p)^{n-1-k} \quad (4)$$

For a scale-free model network which follows a power law distribution [21]:

$$P(\deg v = k) = k^{-\gamma} \ : \ 2 < \gamma < 3 \quad (5)$$

The others two fundamental parameters used to quantify the centrality derive from path studies, which allow to characterize the paths in the graph. A path is a series of consecutive edges that connect two nodes which are in the same connected component. It's obvious that there is more than one path in a graph. By assigning a specific numeric value to every edge, it is possible to find the shortest path as the minimum of the set of all the paths between two given nodes; this is possible thanks to a series of specific algorithms. Given this measure it is possible to calculate two important parameters [21] [22]:

Mean shortest path length:

$$\ell = \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} d(i,j) \quad (6)$$

Longest shortest path (diameter):

$$\max\{d(i,j) \mid i,j \in V, i \neq j\} \quad (7)$$

where $d(i,j)$ is the shortest path between the two nodes and n is the total number of the nodes in the network, that is $|V|$.

The last one measure is the closeness c_i, that measures how close a node i is to all the other nodes inside the network: it is computed by dividing to 1 the sum of all the distances from i to each node j of the network, calculated using shortest path measures [6] [9] [10]:

$$c_i = \frac{1}{\sum_{j=1}^{n} d(i,j)} \quad (8)$$

The longer is the distance between i and all the other nodes, the lower will be the value of the closeness.

C. Assortativity

The last important parameters are given by the hub analysis, that is the study of the nodes which present the highest number of connections in the network. This kind of analysis starts from finding the exact number of hubs, by setting a certain threshold and making an adequate analysis of the distribution degree. The next step is to understand in which extent a new node in the graph tends to connect to an existent hub, a characteristic named preferential addiction, explained by the following formula:

$$p_i = \frac{k_i}{\sum_{j=1}^{n} k_j} \quad (9)$$

where k_i and k_j are the degrees of nodes i and j.

Another property is the assortativity, which measures whether two hubs tend to connect each other preferentially. It is globally calculated using the classic Pearson correlation coefficient r:

$$r = \frac{\sum_{\substack{i,j=1 \\ i \neq j}}^{n} ij(e_{ij} - q_i q_j)}{\sigma_q^2} \quad (10)$$

where q_i and q_j are respectively the remaining degree of node i and j, while e_{ij} is the probability distribution of the remaining degrees of

both the nodes i and j. The remaining degree, while walking through a path, is the number of edges leaving the node other than the one we came from [23]. This quantity is expressed by the following formula:

$$q_j = \frac{(j+1)p_{j+1}}{\sum_{\substack{i=1 \\ i\geq 1}}^{n} i p_i} \tag{11}$$

where p_i is the probability that a randomly chosen node has the degree i. In general, the assortativity measure is described as the preference of the nodes inside the network to link to other nodes or clusters which are similar in some way[7] [23] [24].

3. CONSTRUCTION METHODS

The ways a network can be created combine methods from both molecular biology and reverse engineering. The large amount of data came from high-throughput technologies, collected through molecular biology methods, are used to generate a network using reverse engineering algorithms [11].

To build networks we start from data: they are available in public online databases, and contain multiple gene-disease associations, microRNAs and metabolic information, or hospitalized patient records.

In the following subsections we provide several construction methods, which aim to get a network model from different genetic data. These different ways to obtain a disease network can be seen as diverse points of view of the same problem, that is to discover the common genetic origin of different diseases and the relationships between them.

A. Shared gene method

Goh et al. [12] used the gene-disease associations that are stored in the OMIM (Online Mendelian Inheritance in Man) database to build a disease network. Although the disease–gene associations are still incomplete, OMIM is currently the most complete repository of all known disease genes and their related disorders. The 2,929 entries of the database were reduced to 1,284 distinct disorders by joining disease subtypes through their given disorder names. This merging procedure was automatically performed by running a string-match script. Later each entry was verified manually and each disease was provided with a specific and unique disease ID.

The 1,284 distinct diseases were afterwards classified into 20 primary disorder classes, based on the physiological system which is affected by the specific disorder. If a disorder affects more than one system, the class to which it was assigned was the most affected one. Moreover, if it is not clear which is the most affected class, the specific disease was assigned to a "Multiple" class. There were also 31 disorders with not enough information for a clear assignment and therefore they were assigned to an "Unclassified" class. Thus, the total number of class went up to 22.

Based on these data, the researchers built a bipartite graph, that is a graph where nodes are divided into two disjoint sets and edges only exist between nodes of different sets. One set of nodes of the constructed graph represents all known genetic diseases, while the other set represents all known disease genes. Two instances belonging to different partitions are linked if it is known that mutations in the considered gene are implicated in the specific disease.

Then, starting from this graph, researchers built other two different graph projections, as shown in Figure 2: the HDN (Human Disease Network), which relates the human diseases, and the DGN (Disease-Gene Network), which relates the genes. DGN and HDN offer two complementary views of the same problem.

The HDN, also known as Complex Disease Network (CDN) [7], is a highly interconnected undirected graph, where each node represents a specific disease. The dimension of the selected node depends on the number of genes that lead to that specific disease, while the color of the node represents the disease class to which the disorder belongs. Each edge connects two different nodes if, and only if, the two diseases represented by the two nodes share at least one

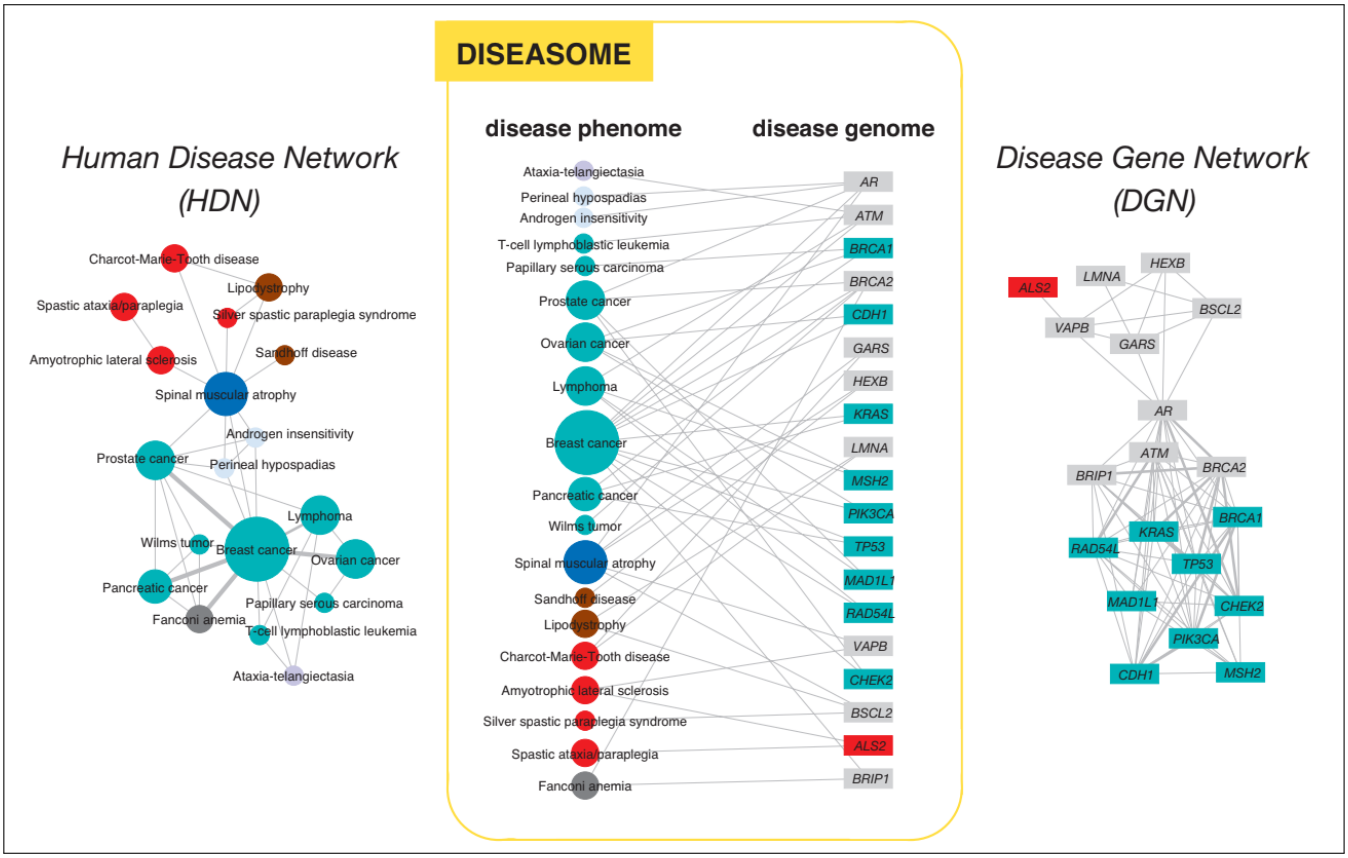

Fig. 2. Diseasome bipartite graph and relative projections [12].

common gene whose dysfunction causes both of them. The thickness of the edge is proportional to the number of shared genes between the two diseases. For example, breast cancer and prostate cancer have in common three genes and therefore the link between them has weight three.

The obtained HDN surprisingly shows many links between both individual diseases and disorder classes. In particular, 867 of the 1284 diseases are linked to one or more other diseases. This highlights that many pathologies have a common genetic origin. The graph is dominated by a single giant connected component consisting of 516 nodes, which is opposed to a series of peripheral disjoint connected components of smaller size and isolated from each other.

Two parameters were used in the quantitative analysis of the HDN:

- The number of genes associated with a disease, that is widely distributed and indicates that most diseases are related to a few genes, whereas a small amount of diseases is related to dozens of genes;
- The degree distribution, which indicates that most diseases are connected to a few other diseases, while a few disorders such as colon cancer and breast cancer have a high degree ($k = 50$ and $k = 30$ respectively) and therefore form hubs.

From this analysis it's clear that the resulting HDN has a significant topological organization, which naturally reflects the major disorder classes. However, the disease classes have peculiar and often different topologies. Some classes form a single cluster consisting of numerous nodes closely connected to each other, while other classes form many clusters isolated from the giant component, which have few nodes. It was possible to quantify this difference by computing two measures: the locus heterogeneity of each disease class and the fraction of diseases that are connected to each other in the HDN.

Two representative examples of this topological characterization are the cancer cluster and the metabolic diseases. The cancer cluster is extensive in the HDN and has both great dimensions and degrees, resulting well represented in the giant connected component and poorly represented in the isolated components; the subgraph of this cluster is one of the most connected disorder classes and has a high locus heterogeneity. On the contrary, the clusters of the metabolic diseases are reduced in the HDN and have both small size and degree, resulting underrepresented in the giant connected component and well represented in the isolated components; the subgraph of metabolic disorders is one of the least connected and has low locus heterogeneity.

The second graph projection that was built by the researchers is the DGN graph, also known as Complex Gene Network (CGN) [7], where each node represents a different gene, that is cause of at least one disorder. The dimension of the node is proportional to the number of the associated diseases, while the color represents the class of the caused diseases. If a gene is implicated in more than one disorder class its color is grey. Each edge connects two nodes if, and only if, the two genes represented by the nodes cause at least a common disease. The thickness of the edge is proportional to the number of common diseases associated with the two genes.

The DGN contains 1777 nodes, and therefore disease genes, and many links between them. In particular, 1377 nodes show at least one connection with other nodes. Also in this case, observing the graph, it is possible to note a single giant connected component, which contains 903 genes and is opposed to a series of disjoint peripheral components of smaller size. In the DGN, the number of genes involved in multiple disorders decreases rapidly with the increase in the number of associated diseases. Therefore, there are few genes involved in many diseases, while many genes are involved in a small number of diseases.

It is also important to note that the edges of the DGN graph can be used as a measure of the phenotypic correlation between two genes. The phenotypic relatedness data can be used in studies of protein–protein interactions, transcription factor-promoter interactions or metabolic reactions.

To demonstrate how the topologies of the HDN and the DGN are significant and different from those of relative random networks, the researchers made further analysis, randomly shuffling the associations between disease and genes but maintaining the same number of edges per each node in the bipartite graph. The results show that the giant component of 104 randomized disease networks contains on average 643±16 nodes, significantly more than the 516 nodes of the giant component of the HDN. The same observation holds also for the giant component of the DGN. Indeed, the average size of the giant component of the randomized gene networks is 1087±20, unlike 903 of the actual DNG. This quantitatively significant deviation from randomness supports the hypothesis regarding the presence of a significant functional clustering in the disease network. Indeed, in current networks a disorder (or a gene) is more likely related to a disorder (or a gene) if both belong to the same disease class.

Next, researchers investigated whether disease-associated genes identify distinct functional modules. They started from the following general hypothesis: if a disease is caused by mutations in multiple genes, then the proteins encoded by them are likely to be implicated in the same functional module. Although this hypothesis had been demonstrated for some pathologies, it remained unclear to what extent most diseases can be related to distinct functional modules in the cellular network. To study this general hypothesis, secondary hypotheses were considered and verified:

1. If disease-associated genes encode proteins that interact within the same functional module, then the proteins within these disease module should more likely interact with each other than with proteins external to the module;

2. If the HDN shows modular organization,

then a group of genes associated with the same disease should share common cellular and functional characteristics, as already noted in Gene Ontology (GO);

3. Genes encoding proteins that interact within common functional modules should tend to be expressed in the same tissue;

4. Disease genes participating in the same functional module should show high correlation of their expression profile.

Researchers proved the first hypothesis by overlapping the DGN on a network of physical protein–protein interactions and discovered that 290 interactions overlap between the two graphs, that is a 10-fold increase compared to the expected number of interactions of the random control.

The second hypothesis was demonstrated by calculating the GO homogeneity of each disorder separately for each branch of GO. The Gene Ontology Resource is a bioinformatic project which aims to make a unified resource for genes and for their relative products. The database is divided into three branches (terms):

1. cellular component;

2. molecular function;

3. biological process.

Each component analyzes the properties of the genes and of their products from a different point of view [13]. GO homogeneity is the maximum fraction of genes causing the same disease which share at least one of the three GO terms. The analysis of the GO homogeneity resulted in significantly high values.

Instead, the third hypothesis was measured by calculating the tissue-homogeneity coefficient (i.e. the maximum ratio between the number of genes that are expressed in a specific tissue and the total number of genes associated with the disorder) for each disease and the result was that 68% of disorders presents almost perfect tissue-homogeneity, compared with the random expectation of 51%.

Finally, to verify the validity of the latest hypothesis, researchers computed the distribution of Pearson correlation coefficients (PCCs) for the co-expression profiles of pairs of genes implicated in the same disease and found that it is shifted toward higher values compared with that of a random control. Furthermore, also the average PCC over all pairs of genes associated with a given disorder shows a relevant shift from the random reference.

From these observations, it is possible to deduce that the genes involved in the same disease:

- Have a higher tendency to generate products that interact with each other through protein-protein interactions;

- Have a greater tendency to be expressed together in specific tissues;

- Have greater tendency to show high levels of co-expression;

- Show synchronized expression, and therefore results in a simultaneous expression;

- Share terms of the GO project.

All these findings support the general hypothesis of a correlation between diseases and functional modules. This can be useful to identify which genes are involved in the same functional module.

Then, Goh et al. found that most disease-related genes (78%), which are nonessential, show no tendency to encode hub proteins and that the expression pattern of non-essential disease genes is dissociated from the overall expression pattern of all the other genes in the cell. On the contrary, essential genes show a tendency to encode hubs and to have highly synchronized expression with the rest of the cell. Finally, it was observed that nonessential disease genes have a tendency to be expressed in a few tissues, unlike essential genes which are more likely to be expressed in many tissues. In conclusion, unexpectedly most nonessential disease genes are placed in the periphery of the cellular network and do not have a central functionality.

This discovery can be explained using evolutionary knowledge. Indeed, mutations in topologically central and widely expressed genes are likely to result in situations of severe developmental impairment, leading the individual to death in the prenatal or early stages of postnatal life and introducing a high probability that the mutation is deleted from the population. For this reason, due to the natural selection, disease-related mutations in the functionally and topologically peripheral regions of the cell give a higher chance of survival for an individual, since they are functionally tolerable. However, genes related to diseases whose mutations are somatic should not be subject to the natural selection just mentioned. Furthermore, the somatic mutations that lead to serious pathological phenotypes should influence the functional center of the network. To demonstrate this hypothesis, the researchers studied the properties of somatic cancer genes, confirming that these genes: are more likely to encode hubs; show higher co-expression with the rest of the genes in the cell; are more represented among housekeeping genes, which are expressed in all tissues. This observed functional and topological centrality supports the current understanding of the key role of cancer genes in cellular growth and development.

At the end, Goh et al. expanded their study including also genes that satisfy the less stringent criterion that the phenotype has not been mapped to a specific locus, in order to verify the robustness of their work. This expansion increased the number of disease genes from 1,777 to 2,765, but also introduced noise in the data, because the association between many of the added genes and diseases is less stringent. However, no findings were affected by this extension.

B. Shared metabolic pathway method

Lee et al. [15] built a bipartite human disease association graph, where the nodes represent different diseases and two nodes are linked with an edge if, and only if, the associated metabolic reactions are "adjacent" (they share a substrate). Then the study focuses on the connected disease pairs that show a flux-rate, and further analyses are carried on.

Gene-based approach, in order to study the relationships between disease phenotypes and the underlying cellular dysfunctions, is the main framework used to tackle this important biology challenge. The impact of variations inside the genes are usually not limited to a single phenotypic phenomenon; indeed, they affect the functional activity of many other cells that, otherwise, do not show any irregularity. Usually, a metabolic disease causes the mutation of a gene which, following this mutation, is no longer able to code for a protein that can correctly fulfill its specific task, for example absorbing (or producing) a certain substrate. Therefore, the disease implies that the substrate cannot be absorbed and, upon reaching a certain threshold, becomes dangerous for the organism (conversely, an enzyme may not be able to produce an essential element). Most of the times the effects of these enzyme defects have a waterfall effect on all the subsequent cellular reactions and mechanisms: for this reason, sets of consecutive reactions that are correlated are analyzed, because the failure on an enzyme can affect the overall "flux-rate".

The data came from both the Kyoto Encyclopedia of Genes and Genomes (KEGG) Ligand database, and the BiGG database, which contain the list of all the human metabolic reactions and the underlying catalyzing enzymes. Also, the Online Mendelian Inheritance in Man (OMIM), that contains the disorder-gene association list, is used to associate enzymes and disorders (737 metabolic reactions over 1493 in KEGG, and 1116 over 3742 in Bigg, are associated at least with one disease); also, 337 disorders over 1437 OMIM diseases have at least one association with a metabolic in KEGG, and 378 in BiGG.

The main idea underlying the Metabolism-Based Human Disease Network (MDN) is that if two reactions share the same substrate, on which the action of an enzyme is expressed, the fluxes of these reactions may be affected: indeed, the impaired activity of some metabolic enzymes is related to specific diseases. So, the hypothesis is that if two diseases can result from defects

affecting the reactions coupled by an enzyme, their cause can be related.

The Metabolic Disease Network in Figure 3.a is built from the data, where each edge represents the metabolic reaction and the linked nodes represents the enzymes and catalysts needed for that specific reaction. Starting from this point, each edge is labeled with the gene that encodes the enzyme that allows the reaction to be executed, and each gene is linked to the disease that its mutation can cause. The complete MDN network in Figure 3.b contains a giant disease cluster called "giant component", and is composed of 308 disease, which represent the nodes, and 878 metabolic links, which stands for the edges. Two nodes are linked if, and only if, the two diseases are "adjacent" (i.e. they share a common enzyme) and so, probably, their fluxes are coupled. Moreover, the color of the nodes indicates the disease class that reflects the distinct organism metabolic functionality, while the node size is proportional to its prevalence inside the dataset. From the network statistics point of view, the average node degree is 5 (each disease is directly linked to other 5) and the degree distribution (the probability distribution of the degrees among the entire network) is much different than a random network: some nodes act like a hub, linking more than 20 diseases, while the majority of the nodes have only few connections.

Examining the functional relevance of the MDN, the Pearson correlation coefficient (PCC) of two diseases related with a metabolic link is higher than the coexpression between pairs without any link; moreover, the causal relationship can be extended also to the reactions whose fluxes are coupled. As we can see in Figure 4, flux coupling is divided in two types: the directional coupling between reaction i and j (i $\rightarrow$ j), where a nonzero flux for i implies a nonzero flux for j but not the reverse; this means that the reaction i is one of the reactions that makes possible the consequent j reaction. The other typology is the the full coupling (i $\leftrightarrow$ j), where a nonzero flux for i implies a nonzero flux for j and vice versa; this means that the reaction i is the only reaction that makes possible the consequent j reaction. In the BiGG database 2605 flux-coupled reactions were identified: the average co-expression (PCC) of the flux-coupled genes is 0.31, higher than the CPP for adjacent reactions 0.24 and significantly higher than the CPP for all gene pairs 0.10: this result proves the existence of functional continuity between "adjacent" and flux-coupled reactions.

Since diseases arise from the breakdown of the process chain, it's interesting to see if the metabolic activity in the network is more likely to contribute to the disease growth and comorbidity: indeed, different studies were carried on analyzing the Medicare record of 13 million elderly patients with 32 million hospital visits from 1990 to 1993. For each pair of diseases X and Y inside the KEGG and BiGG datasets, the comorbidity index C_{XY} is computed in order to understand how the same disorders occur in the patients: a positive value indicates that patients with disease X are more likely to develop also disease Y, while a negative value indicates that there is a "protecting effect from disease Y" in disease X. The average comorbidity index is calculated inside a specific population of the patients. The result of this analysis says that the metabolically connected diseases show an elevate average comorbidity index (0.0027 KEGG, 0.0023 BiGG), three times higher than the average comorbidity for all diseases (0.0009 KEGG, 0.0008BiGG). Moreover, the average comorbidity index for the fully and directional flux-coupled diseases is even higher (0.0062 KEGG, 0.0041BiGG). Also, the 17% of the disease pairs show comorbidity, while the percentage grows to 31% and 28% if the diseases are correlated with a metabolic ink or they are flux coupled.

The prevalence index I_x is calculated for each disease X and it's defined as the fraction of patients that presents the disease X: the majority of the diseases are rare cases, while a few of them affect a great number of patients. As shown in Figure 5, prevalence index and degree of the nodes in the MDN have been measured and related together, showing that the value of

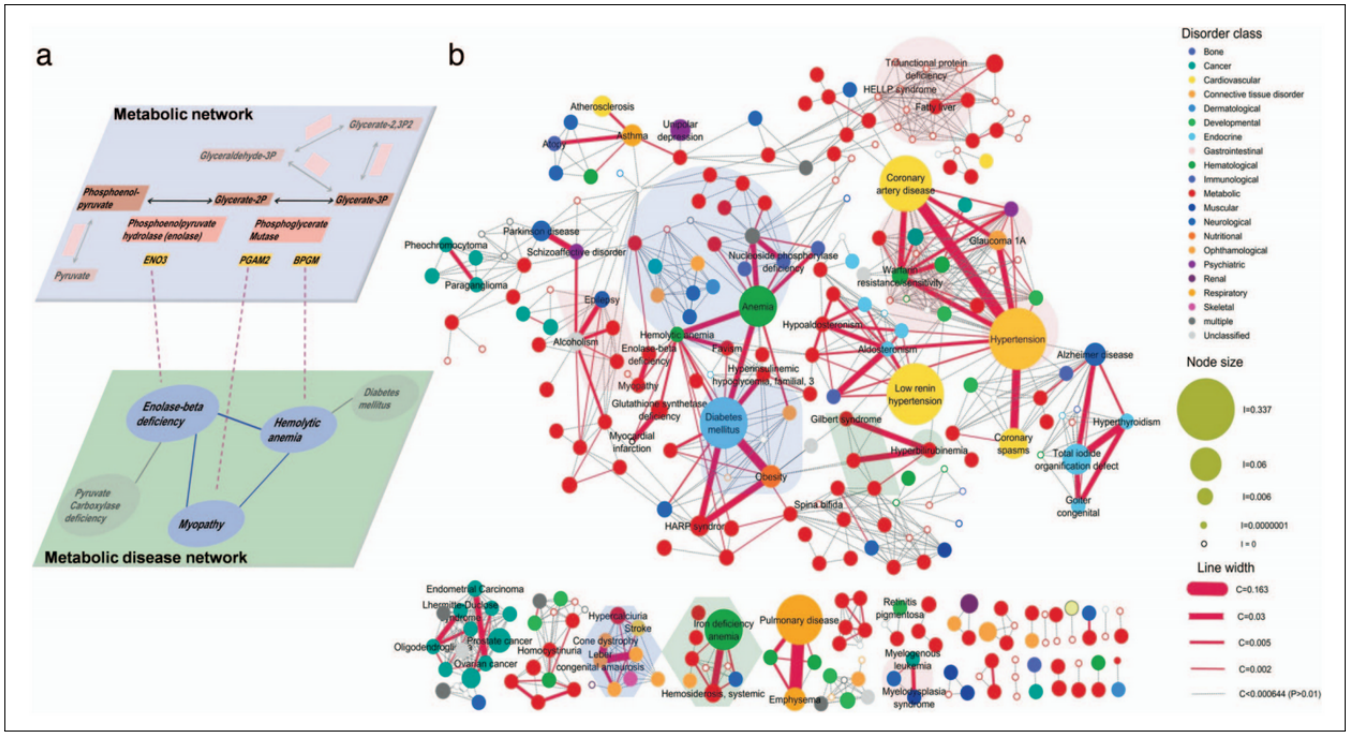

Fig. 3. Construction of a MDN [15].

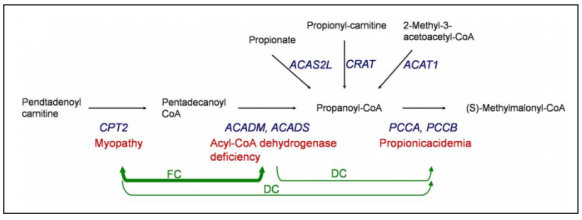

Fig. 4. Flow-coupling diagram: reactions are linked with edges, where the blue represents the genes coding for the catalyzers of the reactions, the red represents the diseases related to the mutation of the gene and in green are shown Directional and Full Coupling relations [15].

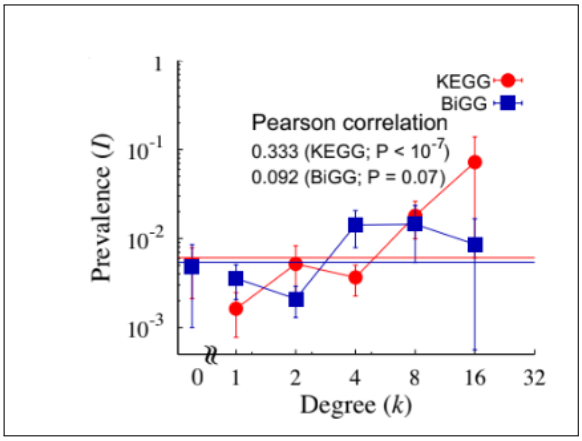

Fig. 5. Correlation between degree of the node and prevalence of the related disease [15].

prevalence increases with the degree: so, the more connected a node in the MDN, the more prevalent is the related disease inside the patient pool.

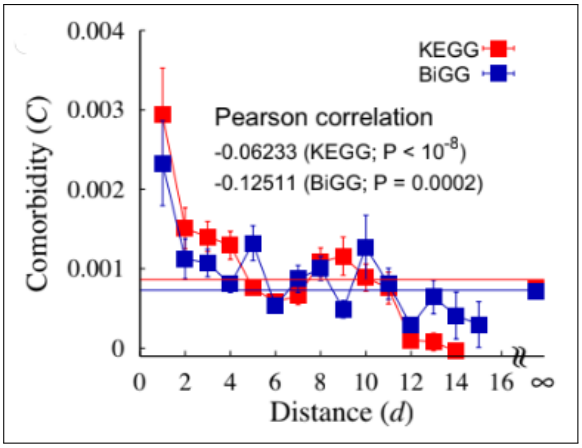

Fig. 6. Correlation between distance and comorbidity [15].

The distance of two diseases inside the MDN is calculated as the number of links of the shortest path that leads from the first disease to the second. The comorbidity effects are not limited to adjacent reactions, but they can spread along several distance inside the MDN due to a sort of transitive property: indeed, if disease X is related to disease Y, and Y to Z, we expect a comorbidity between X and Z; moreover, we can expect this waterfall effect up to three links. Indeed, as shown in Figure 6, the PCC correlation of network distance and comorbidity is -0.062 for KEGG and -0.13 for BiGG, which clearly represents that the comorbidity of two disorders decreases with the distance. Since 27% (and 12%) of reactions appear actively in all tissues, these reactions associated with the diseases may be located in the core of the network.

In Figure 7 is shown the correlation between degree and mortality, and it is easy to see that the more connection has a node, the higher mortality is associated to the related disease. Indeed, the mortality rate (the percentage of people who died inside an 8 year period after the disorder have been spotted) have been associated with each disease and, after computing the PCC correlation coefficient between mortality rate and

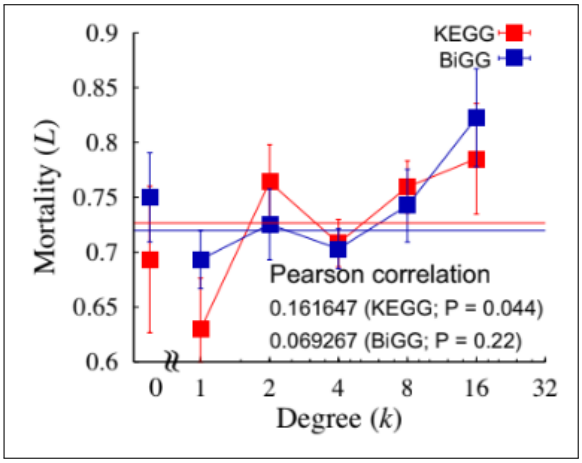

Fig. 7. Relation between degree and mortality [15].

degree of the related node, the result was a positive correlation (0.16 KEGG, 0.07 BiGG). The idea is that, if a disease is very connected, it's probable for the patient to develop another associated disease that increases the mortality of the main disease, which act like a hub.

Regarding metabolic diseases, the coupled metabolic reactions method (shared metabolites and correlated metabolic reactions) has a better predicting power than the shared gene method. After a multivariate analysis of the performance of shared genes, metabolic links and flux-coupled reactions, the results show that the comorbidity effects are better predicted by the metabolic links.

The MDN based method shows 193 new pairs of metabolically linked diseases according to both KEGG and BiGG databases, and also show a high comorbidity; also, the analysis of this disease in pair can lead to discover the comorbidity effects and can suggest some potential "disease-modifying" factors. For example 1656 patients show both diabetes mellitus and hemolytic anemia, which is a higher result than the expected number of 1215 if they occur separately: indeed, inspecting the reaction, the researchers found that some mutated genes are also coding for enzymes that catalyze and speed up that metabolic reaction. Other situations that relate metabolic dependency and MDN to comorbidity can be found easily inside this context.

To sum up, these four main hypotheses have been confirmed

1. Connected diseases show high comorbidity than those who are not linked;
2. The more the degree of a node representing a disease, the higher is the prevalence in the population;
3. Comorbidity effects are not limited to couples of diseases, bu they can be spread along multiple edges (up to distance three);
4. For classical metabolic diseases, the best comorbidity predictor are the metabolic links and not the shared genes.

C. Shared microRNA method

Lu et al. [16] built a disease network by considering the associations between miRNAs and disorders. MicroRNAs (miRNAs) are small non-coding RNA molecules, with an average length of 22 bases, that regulate negatively the expression of genes at post-transcriptional level. It is supposed that 1–4% genes in the human genome are miRNAs and a single miRNA can regulate many other miRNAs. Several studies suggest that miRNAs have an important role in many biological processes, so their mutations, dysfunctions or dysregulations may cause different diseases.

The associations between miRNAs and diseases was retrieved from about 100 scientific papers, which allowed the creation of the Human MiRNAs-associated Disease Database (HMDD). As shown in Figure 8, starting from this data, researchers built a bipartite graph, whose nodes are divided in human miRNA-associated diseases (69 diseases) and disease-associated miRNAs (238 miRNAs). Then, based on this graph, they constructed the MiRNA-associated Disease Network (MDN) by linking diseases that share at least one common miRNA molecule.

The created MDN has a cluster structure, which reflects the classification of the disorders. Cancer diseases are grouped together, suggesting that they share common "onco-miRNAs" (ex. miR-21 is overexpressed in many cancers)

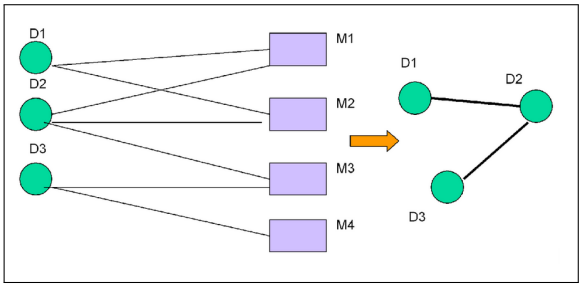

Fig. 8. From disease-microRNA associations to the MiRNA-associated Disease Network (MDN) [16], SI.

and "miRNA suppressors" (ex. miR-125a shows down-regulation in many cancers). Also cardiovascular diseases are grouped together due to miR-1 and miR-133, which are involved in almost all cardiovascular diseases. The cancer cluster and the cardiovascular cluster are well separated, except for the neointimal hyperplasia, a cardiovascular disease, which is mostly connected to the cancer cluster. Therefore, it shares more common miRNA associations with cancers than with cardiovascular diseases. The average number of connections between diseases of different groups is much lower than the average number of connections between diseases inside the same cluster, which are are highly interconnected. For example, a cancer node is linked to an average of 26 other cancer nodes and only 5 cardiovascular disease nodes.

Disease-associated miRNAs can produce various dysfunctions, so the Human MiRNAs-associated Disease Database was divided into two main groups: the "up-regulation" group, that contains the "overexpressed" and "highly expressed" miRNAs, and the "down-regulation" group, that groups the "deleted" and "low expressed" miRNAs. Researchers searched dysfunctional patterns within the same cluster and between different clusters in order to discover if miRNA-associated diseases have the same miRNA dysfunctions.

The number of the same dysfunctions (i.e. when the shared miRNA shows up-regulation or down-regulation in both diseases) and the number of different dysfunctions (i.e. when the common miRNA shows up-regulation in a

disease and down-regulation in the other) was computed intra-cluster (within cardiovascular or cancer cluster) and inter-cluster (between the two different clusters). The result is that shared miRNAs show with a high percentage the same dysfunction within the same cluster, while the percentage is lower between diseases of different clusters. For example, the percentage of same dysfunction evidences in the cancer cluster is 82% and in cardiovascular disease cluster is 77%, whereas it is only 54% between connected disorders of the cancer cluster and the cardiovascular disease cluster.

The researchers investigated also the correlation between tissue-specific miRNAs and the number of diseases associated with them. First, they collected the data of 345 miRNAs expression profiles across 40 tissues, then they computed the tissue specificity of miRNAs with the τ index (which ranges from 0 to 1 and is higher when the tissue specificity is higher) and they classified the disease-related miRNAs in different groups according to the number of associated diseases. Finally, they computed the average tissue specificity index value for each group and they found a negative correlation between it and the number of disorders in which a miRNA is implicated. This means that miRNAs associated with more diseases have a lower tissue specificity index value than miRNAs associated with few diseases. This result suggests a potential correlation between miRNA tissue specificity and disease and it may be useful to predict novel miRNA-disease associations. Indeed, if a disease occurs specifically in a given tissue, it will be very likely that the miRNAs expressed in that tissue are implicated in that disease.

Next, researchers investigated the relation between evolutionary conservation of a gene and diseases. It is expected that if a gene is evolutionary conserved, than it has more connections to other genes and its dysfunction is more likely to result in diseases. In this case, all the human miRNAs were divided into two groups: one containing miRNAs which are conserved in other species and the other containing miRNAs which are human specific. Then, the number of miRNAs associated with no diseases and with at least one disease was counted for every group. It was observed that miRNAs that are conserved in other species have a significantly higher probability to be associated with diseases. This discovery will be useful to understand the miRNAs' roles in human diseases.

The last finding reported in this work is about miRNA sets and miRNA diseases. In particular, researchers investigated whether miRNAs associated with the same disease belong to the same miRNA sets. They defined two types of sets: miRNA families that contain sets of homologous miRNAs and miRNA clusters that contain sets of neighboring miRNAs on the human genome. The results on miRNA families show that in 57% of the diseases a miRNA has at least a family member implicated in the same disease. For example, 3 of the 6 miRNAs in miR-8 family was associated with the thyroid cancer. This suggests that the miRNAs belonging to the same family might play roles in similar biological processes. Therefore, their dysfunction would lead to similar phenotype.

Moreover, Lu et al. uncovered that miRNAs in 46% of the diseases have at least one neighboring member. For example, all the 6 miRNAs belonging to the miR-17 cluster was related to hematopoietic malignancies. This indicates that neighboring miRNAs might be regulated by common regulators and work together. Therefore, their dysfunction would cause the same disease.

In conclusion, if many members of a miRNA set are implicated in a disease, there is a great probability that also the other members of the that set are associated with the same disorder. This can be used to predict novel disease-associated miRNAs.

The above analyses were performed on the miRNA-disease association data of November 2007, and they have been validated in a second study with new data of June 2008, where the main results have been confirmed.

A limitation of this study is that data regarding miRNA-disease associations are incomplete, although the recent publications. For example,

various brain diseases such as schizophrenia, Parkinson's disease, and neurodegeneration, are not linked to each other in this work, probably due to the incompleteness of the adopted dataset. However, this construction method and analysis approach will be useful for future works.

D. Common phenotype method

Another method to construct disease networks consist of linking disease pairs with significant comorbidity between them. The resulting map is called Phenotypic Disease Network (PDN). Hidalgo et al. [17] built a network involving 657 disorders from the clinical history of more than 13 million Medicare patients.

The Medicare database used in this work contains the medical claims associated with hospitalizations from 1990 to 1993 of elderly Americans aged 65 or older; thus, it contains limited information about disorders that are not common among elders from an industrialized country, such as many infectious diseases or pregnancy-related conditions. Moreover, it does not contain information about patients that were not hospitalized but were only visited in the surgery. Thus, this work must be contextualized in this specific population. Another limitation of this study is the ICD-9-CM format of the medical claims because in some cases a specific disease can be associated to more codes, whereas in other cases there are not enough codes to correctly associate all diseases.

Since each comorbidity measure has biases that over- or under-estimate the relationships between rare or prevalent diseases, researchers adopted two different measures: the relative risk (RR) and the ϕ-correlation (already described in the subsection 3.1). Both these measures have intrinsic biases that are complementary: RR overestimates relationships concerning rare diseases and underestimates those with high prevalent disorders; instead, ϕ underestimates the comorbidity between rare and common disorders, while it correctly detects the comorbidity between diseases with similar prevalence. Researchers built a PDN for each measure and analysed their respective relevance to specific disease groups.

The PDN is a graphic representation of the set of all comorbidity associations in which every node is a disease phenotype, identified by a unique ICD-9-CM code, and every edge connects two phenotypes if and only if they show significant comorbidity (RR_{ij} >20 in one network and ϕ>0.06 in the other). The color of the node identifies the ICD-9 category, while its size is proportional to the prevalence of the disease. Also the links of the PDN have different colors which represent the correlation strength.

Although the two different PDNs have many similarities, the global structure and the represented disease group reflect the specific biases of the metric used to construct the links. The network derived from the RR measure contains infrequent diseases and presents clusters that follow the ICD-9 classification (as shown in Figure 9.a); instead, the network derived from the ϕ-correlation measure contains highly frequent diseases which have many links also among distinct ICD-9 categories (Figure 9.b). Both the networks are used to study statistically significant associations at different prevalence scales and offer a complementary view of the problem. Further analyses on these PDNs are explained in the following section Disease progression.

4. APPLICATIONS

In this section we present and analyze some different scenarios of actual application areas of the network theory seen in the previous sections. There is a strong need for network approaches in medicine. Indeed, how we have seen in this work, a network built for specific studies is often based on novel hypothesis and thus it introduces a different approach to look into the same problem from a new point of view, depending on the characteristics of the specific assumptions. Therefore, such need for a network approach in medicine is clear when we understand the great investigation and predictive powers of the networks. Despite this, it must be considered that the vast majority of the network-based approaches is premature or newborn and needs

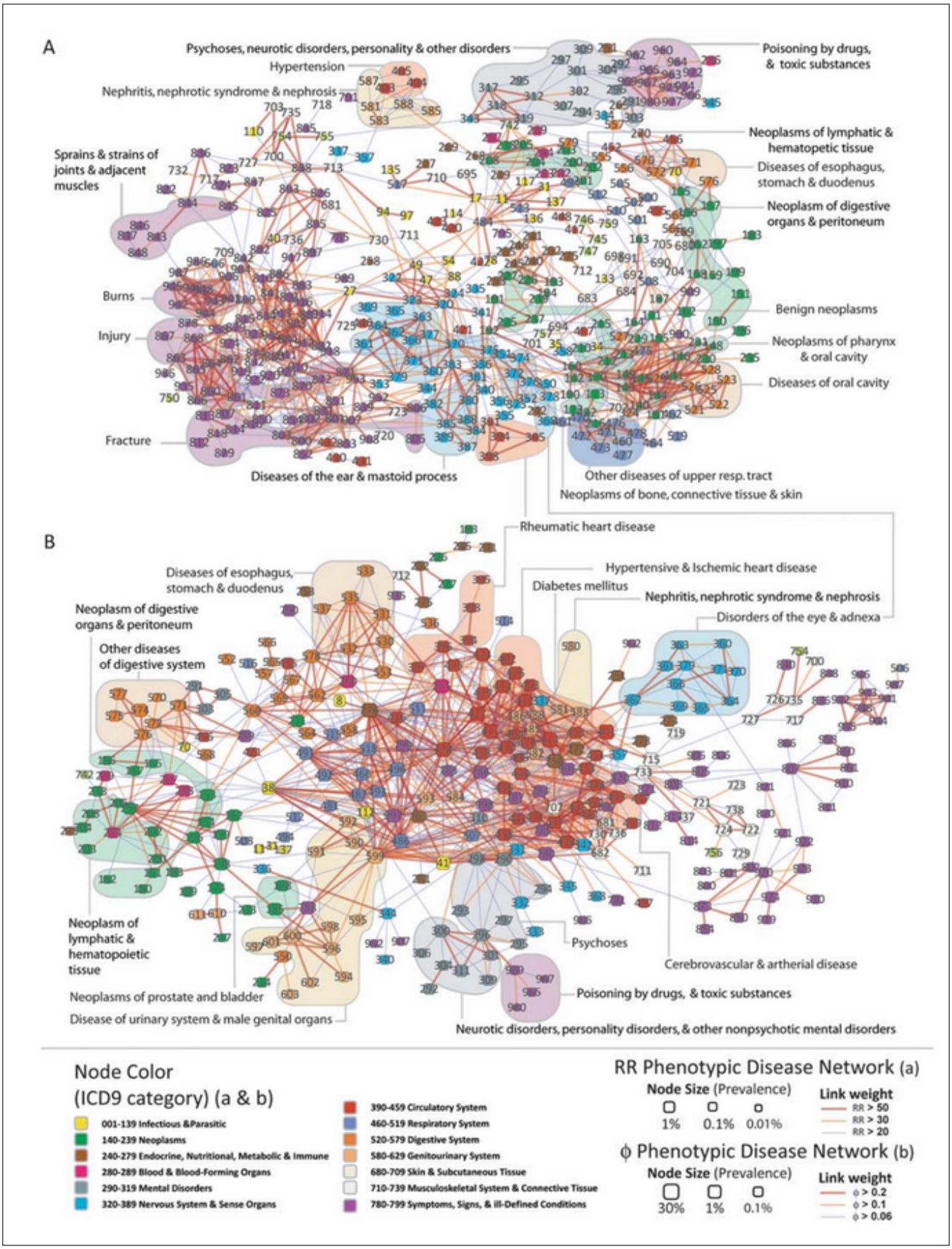

Fig. 9. Phenotypic Disease Networks (PDNs) built using *RR* (a) or ϕ-correlation measure (b) [17].

time to possibly satisfying all the expectations about its practical potential in medicine studies and for diagnostics and therapeutics. This kind of approach for now is considered mostly as a model used to make predictions based on empirical information [18]. An example of this observation is the graph constructed following the shared miRNA method [16], which, as we have shown, is still far from being a reliable tool for a serious investigation due to the lack of complete datasets.

A. Diagnosis and treatment

According to the work of Streib et al. [18] a network-based theory in medical studies may be progressively required, because of the necessity of an improvement of methods for diagnose a disease and cure it. Improvements in the network approaches could lead a better way for diagnostics and, thus, a stronger direction for treatment. For instance, a lot of diseases hide complex and unknown mechanisms, that can be successfully studied through the analysis of the underlying molecular pathways and components [18]. Simple examples of this necessity of a network-based approach in medical sciences are given by cardiovascular diseases and cancer, which have a very complex etiology and various processes. Specifically, observing the HDN, we can see that the cancer module is prominent compared to the other most connected disorders: this is due to the high degree of connections between nodes related to the common tumour factors, such as TP53 and CHEK2. It has been proved indeed how different mutations in TP53 gene (a transcription factor that regulates the cell cycle) are related with 11 types of cancer [19], demonstrating how consistent the HDN is with all the information provided by annotations. A detailed analysis of the HDN shows also that somatic cancer genes are more likely to encode hubs. Moreover, they tend to be co-expressed with the rest of the genes in the cell and they are over-represented among housekeeping genes. Also, the centrality of somatic cancer genes in the DGN is in line with the critical role played by them in the cellular development and growth [12] [19]. This important result can provide a better understanding of how a disease can affect the cell and the organism and can be useful for the creation of an efficient protocol designed for the cancer diagnosis and for the adoption of an appropriate treatment.

B. Pharmacology

As we seen in a study led by Barabási et al. [2], usually, a specific phenotype related to a disease is rarely a consequence of a single factor alteration, but rather shows a general effect caused by several complex interactions within the organism. This fact is very important to understand how a disease works at system level. In this context, the concept of disease module comes in handy; it is a subgraph containing all the cellular components implicated in a disease. This perspective can be a very useful tool because it can help the development of novel drugs, reducing the research to those molecules which induce a significant variation in the specific disease module. In addition, a network approach could provide results closer to those which we could obtain for in vivo studies. Indeed, a in vitro drug effect study analyzes the binding between a specific drug and related molecules individually, without having an overall view. For this reason any future in vivo study could lead to very different results. Instead, a in silico study could guarantee a more complete vision of the general situation in a drug effect study [2].

An interesting finding is that female breast cancer and the clusters of schizophrenia and bipolar disorder are strongly negatively correlated: this may suggest that there is a competition for the genes that regulate the production and death of the cells. Indeed, while the breast cancer presents an uncontrolled proliferation of the cells, the other two diseases are related to a genetic polymorphism that leads to an abnormal cell death. This discovery may reveal us why the “tamoxifen”, a cancer-treatment drug, is effective in treating symptoms of bipolar disorder [26].

Another considerable application is the use of

the system biology approaches to identify efficient combinations of anticancer drugs. Since anticancer therapies targeting specific molecules are not giving totally satisfying results, it is necessary to adopt a global perspective, in order to overcome all the difficulties caused by the high complexity of the interactions present in a cancer cell [2] [20].

C. Disease Comorbidity

Park et al. [14] studied the comorbidity (i.e. the coexistence of multiple diseases in a single patient) between disease pairs with shared genes in order to understand if this sharing has consequences for disease occurrence in patients. Indeed, disease-causing defects, due to cellular interaction, may affect not only the products of the mutated gene but also other molecular components and their functions, resulting in possible comorbidity effects.

In particular, the researchers analyzed the large-scale comorbidity patterns stored in the US Medicare claims database and the gene-disease associations in the HDN to understand the statistically significant comorbidity patterns at a population level. The incidence of disease i is defined as I_i and is the number of patients in the Medicare database diagnosed with the corresponding ICD-9-CM code and its sub-level codes. The ICD-9-CM is a famous hierarchical disease diagnosis code system. Instead, C_{ij} represents the number of patients who are diagnosed with both disease i and j. The comorbidity of two diseases is estimated in two ways:

- The relative risk that is $RR = C_{ij}/C^*_{ij}$, where $C^*_{ij} = I_i I_j / N$ represents the expected value of C_{ij} if the two diseases i and j were independent;
- The ϕ-correlation that is defined as $\phi = (NC_{ij} - I_i I_j)\sqrt{I_i I_j (N - I_i)(N - I_j)}$, where $N = 13039018$ is the total number of patients in the Medicare database.

Two diseases co-occur more than expected if $RR > 1$ and $\phi > 0$.

In this approach there are four main limitations. The first one concerns the mapping of the disease names from the Medicare database (ICD-9-CM format, where the first three digits of the code indicate the main disease category, while the last two show additional information) to the OMIM database (OMIM format): they are not identical, therefore a manual mapping has been drafted. This mapping is not perfect, and it can contain spurious associations or discrepancies in disease names that may result in noise in the next steps of the study. It is important to note that the noise can be also introduced in the process of assigning a specific diagnosis to a specific ICD-9-CM code in hospital. The second limitation is that only the 5% of the ICD-9-CM codes can be successfully mapped to OMIM disorders because the OMIM repository contains only diseases with validated gene–disease associations. However, this is not a great issue because the 90% of the total patients in the Medicare database was diagnosed with at least one disorder whose ICD-9-CM code has been considered in the mapping, so the study has not been strongly affected from this limitation. Then, the third limitation is that this work takes on account only the role of the cellular network on comorbidity but there are also many other factors that can contribute to comorbidity, such as the human lifestyle and behavior, the environmental situation, and the treatments of the patients. Finally, the last problem raises in the computation of the average comorbidity between all disease pairs because the value is >1, indicating that many patients develop multiple diseases, regardless of cellular network-level connections. However, this problem is solved considering the overall comorbidity as the baseline against which to compare the impact of the genetic and cellular networks.

The Figure 10 shows the procedure used to connect comorbidity and shared disease genes between the breast cancer and the bone and cartilage cancer. In the Medicare layer it is represented a table in which each entry is characterized by the patient ID and the ICD-9-CM codes of the diagnosed diseases. The blue lines

represent the incidence of each disorder: 174.1, which is a sub-level code, is correctly counted as an incidence of 174 for breast cancer. Instead, the red lines indicate the co-occurrence of the two diseases in the same patient. The OMIM layer shows the shared genes and in green the disease-gene associations. Finally, the mapping between the two different diseases-labeling schemes (ICD-9-CM and OMIM) is represented with grey lines.

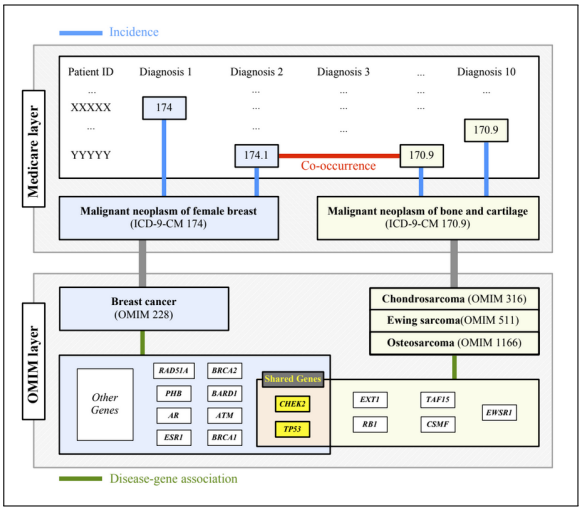

Fig. 10. Example of mapping of patients' diagnosis and co-occurrence from Medicare database to disease-gene associations in OMIM database [14].

The cellular network-level relationships between two diseases i and j are measured by the researchers with the following three cellular variables:

- n_{ij}^{g} is the number of shared genes associated with both the diseases and measures the common genetic origin;
- n_{ij}^{p} is the number of shared protein-protein interactions (PPIs) and measures the PPI network-level relationships between the two diseases;
- ρ_{ij} is the average Pearson correlation coefficient (PCC) of the co-expression between pairs of genes of the two different diseases and measures the co-expression degree of the genes implicated in the two diseases.

Then, researchers calculated the Pearson correlation between these cellular variables (n_{ij}^{g} , n_{ij}^{p},ρ_{ij}) and comorbidities (RR and ϕ) in order to investigate whether the existence of these cellular-level connections (i.e. when these quantities are greater than zero) increases the likelihood that a patient simultaneously develops both disorders. The cellular variables were computed for all the 83924 disease pairs: 658 disease pairs are linked through shared genes and 1873 through PPIs. The Pearson correlations between relative risk, ϕ-correlation and the three genetic variables are all positive, although ng has the highest correlation values, as shown in Figure 11.

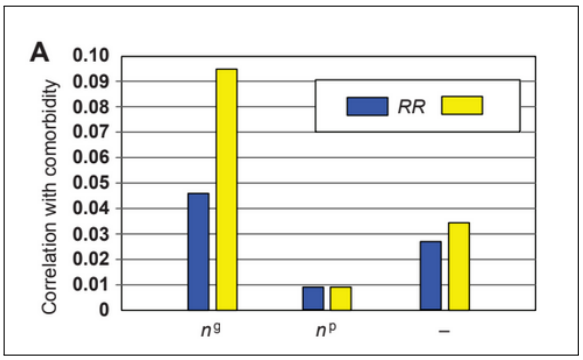

Fig. 11. Pearson correlation between cellular variables and comorbidity [14].

In order to quantify and measure the degree of comorbidity, the average values $< RR >$ and $< \phi >$ were computed over the entire set of 83924 disease pairs and were compared to the average comorbidities of the disease pairs that are linked at cellular network level. The result (see Figure 12) was that the disorders that share genes (with $n_{ij}^{g} \geq 1$) have an average comorbidity from 2 to 4 times higher than the average of the total pairs. A slightly increased comorbidity can be seen also for diseases pairs connected through PPIs (with $n_{ij}^{p} \geq 1$) and highly co-expressed genes (with $ij \geq 0.5$). Hence, the probability of developing a certain disease doubles if this disease shares one or more genes with the patient's primary disease in the HDN. However, there are some pairs of diseases that share common genes but do not exhibit a high

comorbidity: this may be due to the fact that different gene mutations present totally different phenotypic effects.

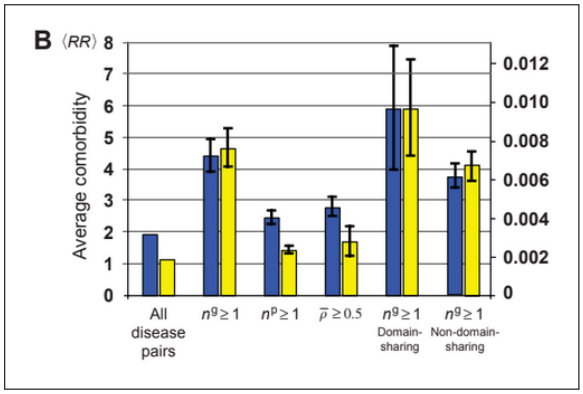

Fig. 12. Average comorbidity for disease pairs satisfying specific cellular constraints [14].

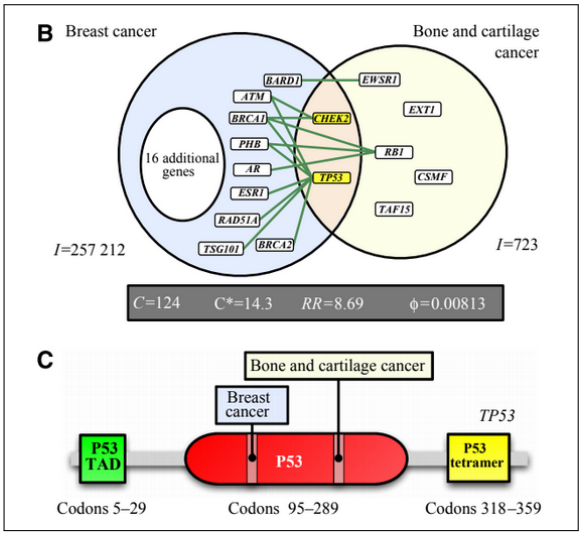

Fig. 13. Shared genes and PPI interactions between breast cancer and the bone and cartilage cancer [14].

Another finding of this work is that disease pairs which are caused by mutations in the same functional domain of the shared genes show higher comorbidity than those which are caused by mutations in different functional domains (as seen in Figure 12). To test this, researchers identified the functional domains of disease-causing mutations using the Pfam database and computed the $< RR >$ and $< \phi >$ for both disease pairs whose mutations are on the same domain of the shared gene and whose mutations are in distinct functional domains, finding more statistically significant comorbidities in the first case.

This findings suggest that new comorbidity patterns can be found combining disease information and cellular network data. In some cases the comorbidity patterns were already known, such as diabetes and obesity; in other cases the found patterns are due to the limitations previously reported, such as diabetes and hypoglycemia, since hypoglycemia is a known side effect of the treatment of diabetes. However, researchers discovered also several interesting disease pairs that are linked at cellular level and also show significant comorbidity, such as Alzheimer's disease and myocardial infarction or autonomic nervous system disorder and carpal tunnel syndrome.

In Figure 13 is shown the case of the breast cancer and the bone and cartilage cancer, which are linked at cellular level and also have a high comorbidity. They share two genes (CHEK2 and TP53) and 13 protein–protein interactions (represented by the green lines). The relative risk between the two diseases is $RR = 8.69$, indicating that the number of patients who simultaneously develop both diseases shows a seven-fold increase compared with random expectation. The Figure 13 shows also the functional domains of the TP53 protein and highlights that mutations of both disorders on the shared TP53 gene take place on the same P53 domain.

Although there may be various possible physiological explanations for some of the observed comorbidities, the method described above can be useful to formulate new, testable hypotheses about the biological basis of disease correlations. In this context it's of absolute importance the collection of large amount of health care and treatment data, that can be complemented by the well-established genomic knowledge.

D. Disease Progression

PDNs have many possible applications. In the work of Hidalgo et al. [17] they are used to study the illness progression from a dynamic network point of view: in this prospective, the PDN can be seen as a map in which patients

"jump" from one disease to another along the edges. However, it is important to observe that disease records are temporarily ordered, but their temporal progression may be the result of a limited data window. For example, if a patient is diagnosed with disease A on the first visit and with disease B on the second visit, we cannot conclude that disease A precedes B because B could have been diagnosed at any earlier time before the recorded data. Hence, researchers adopted a conservative approach.

Using a method that studies whether or not a node property extends along the edges in the network, researchers analyzed the relationship between diseases diagnosed in four subsequent visits of $N = 946580$ patients, finding that the diseases diagnosed in the first two visits are more correlated with those diagnosed in the last two visits than the results obtained with random PDNs. In particular, the inter-visit correlations are greater by a factor of 10 on average for the ϕ-PDN and of 1.5 for the RR-PDN compared with the correlations of the control case. This discovery led the authors to think that the development of patients' diseases is a spreading process over the PDN (with a more pronounced effect in the ϕ-PDN).

Although the data are not useful to say anything about the directionality of disease progression, it is possible to study how the different strength of comorbidity in patients of different genders and ethnicities can affect the dynamics of disorder progression. Researchers investigated this by calculating the odds ratio for the difference in comorbidity between two diseases (i and j) expressed in two populations (α and β), which is defined as follows:

$$OR_{ij}(\alpha,\beta) = \frac{p_{ij}(\alpha)}{1-p_{ij}(\alpha)}\frac{1-p_{ij}(\beta)}{p_{ij}(\beta)} \qquad (12)$$

where $p_{ij}(x)$ is the probability that both disorders i and j are diagnosed in a patient of population x.

Figure 14 shows a subgraph of the PDN constructed using ϕ-correlation, which contains only the nodes associated to hypertension and ischemic heart disease.

In particular, this figure highlights the differences in the strength of comorbidities between white and black males. Blue links indicate comorbidities that are strongest among black males; whereas red links indicate comorbidities that are strongest among white males. Observing the figure it is possible to note that ischemic heart disease, infarctions, hypercholesterolemia, and pulmonary complications tend to be more comorbid in white males; whereas hypertension, diabetes, and renal diseases tend to be more comorbid in black males.

This kind of study conducted on white and black male, and also on males and females, revealed that each group shows an almost distinct subgraph of more comorbid diseases, demonstrating the capability of the PDN to explore gender and ethnic differences on comorbidity.

Next, researchers explored the correlation between the lethality of a disorder and its connectivity in the PDN. To measure the lethality of a disease they calculated the percentage of patients died within the 8 years after the first recorded diagnosis; instead, the connectivity K of the disease i was computed by summing up all the comorbidity values between it and the other diseases to which it is connected:

$$K_i^{\phi} = \sum_j \phi_{ij}, K_i^{RR} = \sum_j RR_{ij} \qquad (13)$$

High values of K_i^{ϕ} and K_i^{RR} mean that disease i is strongly connected to many other disorders in the PDN.

Researchers found that there is an evident positive correlation between disease connectivity and mortality, and a negative correlation between the average connectivity of the diagnosis assigned to a patient and the number of years survived after the last diagnosis, indicating that the observed correlation between connectivity and lethality is not caused by a simple accumulation of diagnoses of sicker patients. This result means that patients diagnosed with diseases which are highly connected in the PDN tend to die sooner than those affected by less connected diseases.

An additional analysis regards the direction-

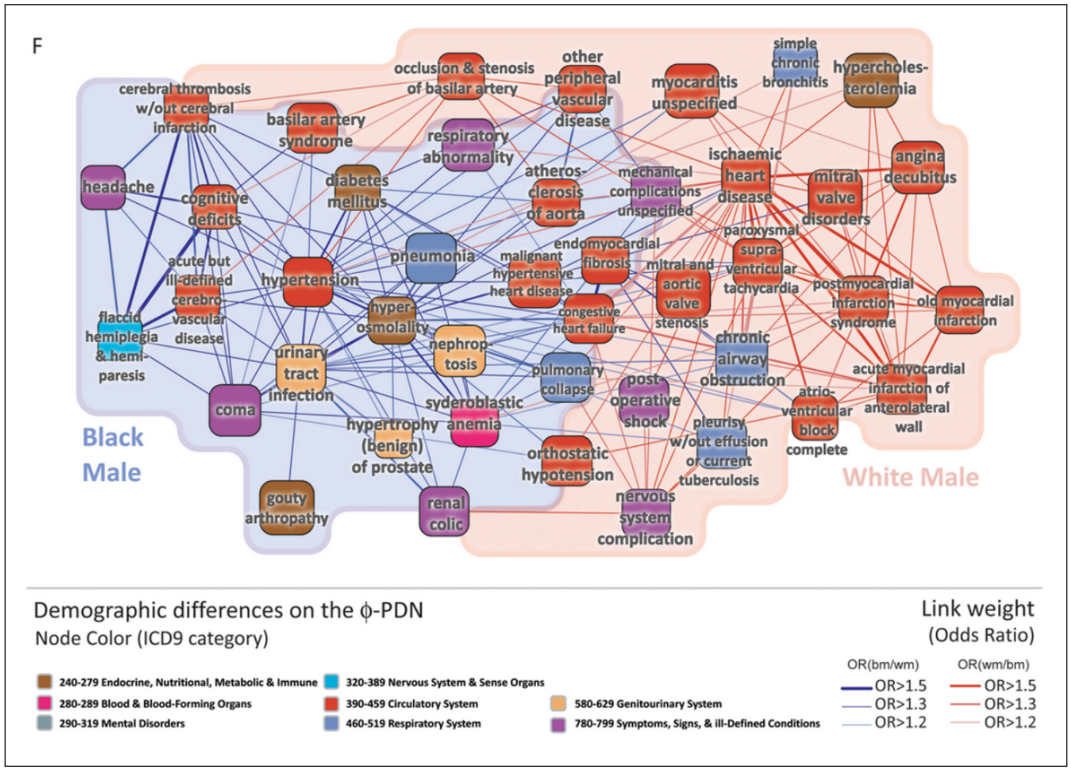

Fig. 14. Subgraph of the ϕ-PDN, which shows all diseases connected to hypertension and ischemic heart disease: the color of the link indicates if a comorbidity is stronger in black (blue) or white (red) males [17].

ality of disease progression. In order to reduce the noise due to the limited observation period of their study, researchers considered only links between diseases affecting at least the 0.2% of the patients. To define the directionality, they computed $L_{i\to j}$ by counting the number of times that disease i was diagnosed before j (the cases in which both diseases were diagnosed for the first time in the same visit are disregarded). This value was then normalized with the prevalence Pi, since a more prevalent disease has more probability to occur than the others, so it becomes $l_{i\to j} = (L_{i\to j} + 1)/P_i$. Finally, the directionality $\lambda_{i\to j}$ of the link connecting disease i to disease j was calculated as:

$$\lambda_{i\to j} = \log_{10}\left(\frac{l_{i\to j}}{l_{j\to i}}\right) \tag{14}$$

If $\lambda_{i\to j} = 1$, it means that the probability a patient is diagnosed with disease i before it is diagnosed with disease j is 10 times higher than the opposite. Whereas $\lambda_{i\to j} = 2$ means that the ratio between this probabilities is 100. The distribution of the directionality shows a peak close to $\lambda_{i\to j} = 0$, indicating that many links do not have a preferred direction.

Instead, the precedence of a disease i is the sum of the directionality of all the connections of this disease in the PDN:

$$\Lambda_i = \sum_j \lambda_{i\to j} \tag{15}$$

If $\Lambda_i > 0$ the disorder can be considered as a "source" disease, which is a starting point for other disorders, otherwise can be treated as a "sink" disease, that is an arrival point from other disorders. To obtain a precedence value which is independent of disease prevalence, we can subtract the trend from it ($\Lambda_i^* = \Lambda_i + 496.08\log_{10}(P_i) - 2446.2$). An interesting result is that this measure of precedence Λ_i^* is negatively correlated with the connectivity of a disease: this means that highly connected diseases tend to come after other diseases (thus rep-

resent advanced stages of the disease), and not before like someone may think. Furthermore, disease precedence and lethality are correlated: patients diagnosed with sink diseases tend to die faster than those diagnosed with source diseases. The conducted statistical analysis shows also that for short terms (2 years) precedence is a better predictor of lethality than connectivity, whereas for longer terms (8 years) it's the opposite.

To sum up, these following findings have been observed:

1. patients tend to develop diseases in the network neighbourhood of diseases that they have already had;
2. patients who have diseases that are highly connected in the PDN show a higher mortality than those who have diseases that are less connected;
3. patients which develop diseases that tend to be preceded by others in the PDN tend to die sooner than those diagnosed with diseases that precede others;
4. the disease progression along the links of the network is different for patients of different genders and ethnicities;
5. highly connected diseases tend to come after other diseases.

Despite the great potential of the PDNs, the lack of phenotypic data is for now a relevant problem, which limits biological progress towards understanding the origins of human disease. However, this work represents a step toward the resolution of this problem thanks to the introduction of a large and available dataset of comorbidity associations. PDN could be used to study the disease evolution of patients and represent an ideal way to visualize and represent medical health records. As we have shown, they could also help to find differences in the strength of comorbidity measured for patients of various ethnicities and gender. A possible future work could be to investigate whether differences in the comorbidity patterns expressed in different populations indicate differences in biological processes, environmental factors, or health care quality of each population.

5. DISCUSSION

The usual reductionist way of studying a complex system consists in analyzing each part of this system separately and carefully. This approach has often been chosen to analyze complex systems and has also brought positives results. An example of this approach is the study of metabolic process components, which finds all the knowable about every single element of the chain. This kind of study results from the effort of many research groups that have studied independently the proteins and the possible co-factors, the reagents, the products, and the reaction intermediates. This way, although it is very useful and permits to enrich the literature with a lot of information, suffers from some fundamental limitations. Indeed, it lacks of a global view that can provide information about the existing interactions and the relations among the components of the complex system in question. The network theory offers a new way to look at a complex system, providing an holistic approach which tends to go beyond the characteristics of the single components of a system. The study of the overall interactions among the components of a complex system can provide much more knowledge than the analysis of the same parts taken individually.

In the disease networks it's used this powerful concept to analyze with an appropriate network-based approach the human diseases all at once, rather than one at a time. This allows to have a complete view of the various associations between different diseases and the implicated genes. As we have seen before, several in-depth ways to generate a disease network were studied, with various differences between semantic relations among different diseases. In this work, we analyze four disease networks built in four different ways. In particular, we considered four different relationship profiles between two diseases, all based on the sharing of a biological element: a gene, a metabolic pathway, a molecule

of miRNA, or a specific phenotype. The various ways of constructing the disease networks provide us the tools to analyze different aspects of the same issue. Indeed, we can generate distinct disease networks that highlight different disease characteristics and problems. This consideration is essential to understand that there is not a real term of comparison among the networks: each network encloses different assumptions and consequently provides a distinct paradigm to investigate various features of the same problem. From this observation, we understand how network science can be rich in information and can be a powerful and elegant instrument to properly deal with the enormous complexity of a problem involving biological systems.

In this work we showed the outstanding predictive and analytical capability of a well-built network approach, that, unfortunately, does not yet have an adequate application background behind it and thus its potential remains mostly unexpressed. There are still not enough real cases for the researchers to recognize such potential, neither to appreciate it as a test for new approaches to medicine and science. But why there is not such a good application background? We can highlight two important issues: at first, this discipline is still premature and provide a new way of thinking about biological issues, that's why it could take time and it could need further studies to allow this sort of approach to become important enough for additional applications; secondly, a not indifferent problem concerns the incompleteness of the available data, which are usually taken from specific databases obtained through high-throughput bioinformatics methods. So, a more complete application range of the network science will include a better annotation and greater completeness of the support databases, in terms of quality and quantity.

ACKNOWLEDGEMENT

This work is presented for the final grade of the Bioalgorithms course, held by prof. Matteo Comin at the University of Padova in 2020.

REFERENCES

1. Vidal, M., Cusick, M. E., & Barabási, A. L. (2011). Interactome networks and human disease. Cell, 144(6), 986-998.
2. Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature reviews genetics, 12(1), 56-68.
3. Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., & Jackson, R. B. (2014). Campbell biology (No. s 1309). Boston, MA: Pearson.
4. McClellan, J., & King, M. C. (2010). Genetic heterogeneity in human disease. Cell, 141(2), 210-217.
5. Keith, B. P., Robertson, D. L., & Hentges, K. E. (2014). Locus heterogeneity disease genes encode proteins with high interconnectivity in the human protein interaction network. Frontiers in genetics, 5, 434.
6. Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. Nature reviews genetics, 5(2), 101-113.
7. Barrenas, F., Chavali, S., Holme, P., Mobini, R., & Benson, M. (2009). Network properties of complex human disease genes identified through genome-wide association studies. PloS one, 4(11).
8. Nelson, D. L., Cox, M. M., & Lehninger, A. L. (2008). Principles of biochemistry (p. 245). New York:: Freeman.
9. Alon, U. (2019). An introduction to systems biology: design principles of biological circuits. CRC press.
10. Freeman, L. C. (1978). Centrality in social networks conceptual clarification. Social networks, 1(3), 215-239.
11. Natale, J. L., Hofmann, D., Hernández, D. G., & Nemenman, I. (2017). Reverse-engineering biological networks from large data sets. arXiv preprint arXiv:1705.06370.
12. Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A. L. (2007). The human disease network. Proceedings of the National Academy of Sciences, 104(21), 8685-8690.

13. Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. Nucleic acids research, 32(suppl-1), D258-D261.
14. Park, J., Lee, D. S., Christakis, N. A., & Barabási, A. L. (2009). The impact of cellular networks on disease comorbidity. Molecular systems biology, 5(1).
15. Lee, D. S., Park, J., Kay, K. A., Christakis, N. A., Oltvai, Z. N., & Barabási, A. L. (2008). The implications of human metabolic network topology for disease comorbidity. Proceedings of the National Academy of Sciences, 105(29), 9880-9885.
16. Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., & Cui, Q. (2008). An analysis of human microRNA and disease associations. PloS one, 3(10).
17. Hidalgo, C. A., Blumm, N., Barabási, A. L., & Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. PLoS computational biology, 5(4).
18. Emmert-Streib, F., Tripathi, S., Simoes, R. D. M., Hawwa, A. F., & Dehmer, M. (2013). The human disease network: Opportunities for classification, diagnosis, and prediction of disorders and disease genes. Systems Biomedicine, 1(1), 20-28.
19. Vogelstein, B., Lane, D., & Levine, A. J. (2000). Surfing the p53 network. Nature, 408(6810), 307-310.
20. Azmi, A. S., Wang, Z., Philip, P. A., Mohammad, R. M., & Sarkar, F. H. (2010). Proof of concept: network and systems biology approaches aid in the discovery of potent anticancer drug combinations. Molecular cancer therapeutics, 9(12), 3137-3144.
21. Fronczak, A., Hołyst, J. A., Jedynak, M., & Sienkiewicz, J. (2002). Higher order clustering coefficients in Barabási–Albert networks. Physica A: Statistical Mechanics and its Applications, 316(1-4), 688-694.
22. Abraham, I., Fiat, A., Goldberg, A. V., & Werneck, R. F. (2010, January). Highway dimension, shortest paths, and provably efficient algorithms. In Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms (pp. 782-793). Society for Industrial and Applied Mathematics.
23. Newman, M. E. (2002). Assortative mixing in networks. Physical review letters, 89(20), 208701.
24. Newman, M. E. (2001). Clustering and preferential attachment in growing networks. Physical review E, 64(2), 025102.
25. Newman, M. E., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. Physical review E, 64(2), 026118.
26. Rzhetsky, A., Wajngurt, D., Park, N., & Zheng, T. (2007). Probing genetic overlap among complex human phenotypes. Proceedings of the National Academy of Sciences, 104(28), 11694-11699.