Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in Fisica

Network Diffusion Methods for Omics Big Bio Data Analytics and Interpretation with Application to Cancer Datasets

Relatore:
Prof. Gastone Castellani

Presentata da:
Tommaso Matteuzzi

Anno Accademico 2016/2017

Contents

Abstract

A wide range of biological and empirical evidence supports the so called *"disease module"* or *"local"* hypothesis. According to this hypothesis, within the cell, functional similarity between molecular entities is associated to network proximity and entities involved in the same disease are more likely to interact with one another, in other words, the cellular components associated with a disease tend lie in the same neighbourhood of the network of interactions between proteins.

In order to understand the molecular mechanisms underlying a complex disease a central issue in current biomedical research is the identification of the interactome *modules* related to the disorder, also called *disease modules*. However, due to the incompleteness of interaction data and to the wild variability of genes associated to a disease, the identification of such modules is not straightforward. Physical methods based on network diffusion efficiently tackle this problem by simulating the behaviour of a random walker on a network.

In the first part of this thesis, we review the theory underlying diffusive process on network finding some interesting connection to clustering techniques and to other physical processes described by Laplacian dynamics.

Then, we exploit a novel diffusion technique, recently proposed in literature, for detection of sub-modules of protein-protein interactions network enriched in altered genes. Starting from a set of query nodes, such as the set of altered genes in a cancer, the aim of the method is to find a set of other genes related to the query set and forming with it a connected subnetwork. We applied such technique, for the first time, to three different real cancer datasets.

The pipeline is organized in two part. First, we select a set of source

genes and we associate to each of them an initial information which reflects its *"degree" of alteration*, *i.e.* gene expression fold change or frequency of mutation. The diffusion algorithm propagates such information within the network and allow us to define a network-based ranking of the BioPlex nodes, according to their proximity to the sources.

The enriched sub-modules identification is carried out by a network resampling procedure, based on the minimization of an objective function, starting from the network-based ranking list.

The application of this method allow us to retrieve some genes that in literature are already associated to the cancer types under examination and new genes, forming with them disease modules, that are likely to have a role in the pathologies because of their proximity to the sources. We finally assess, by a *gene set enrichment analysis*, the association of the detected disease modules to known biological pathways.

Riassunto

Nella attuale ricerca biomedica esistono numerose evidenze sperimentali a supporto della così detta *ipotesi locale*. Secondo questa ipotesi la similarità funzionale tra entità molecolari all'interno della cellula è strettamente correlata alla loro vicinanza sull'*interattoma*, il network delle interazioni tra proteine. Entità coinvolte in una stessa malattia sono concentrate in zone contigue della rete ed hanno perciò una maggiore probabilità di interagire tra loro.

Un passo fondamentale verso una comprensione sistematica dei meccanismi alla radice di una malattia complessa è costituito dalla identificazione dei *disease modules*, cioè quei sottonetwork connessi dell'interattoma con un alto numero di alterazioni correlate alla malattia. Tuttavia, l'incompletezza del network, dovuta a limiti sperimentali, e l'elevata variabilità dei geni alterati rendono la soluzione di questo problema non banale.

I metodi fisici che sfruttano le proprietà dei processi diffusivi su network, dei quali mi sono occupato in questo lavoro di tesi, sono tra quelli che, attualmente, consentono di ottenere le migliori prestazioni.

Nella prima parte del mio lavoro, ho indagato la teoria relativa alla diffusione ed ai random walk su network, trovando interessanti relazioni con le tecniche di clustering e con altri modelli fisici la cui dinamica è descritta dalla matrice laplaciana del network.

Ho poi implementato un tecnica basata sulla diffusione su rete, recentemente proposta in letteratura, applicandola a dati di espressione genica e mutazioni somatiche di tre diverse tipologie di cancro.

Il metodo è organizzato in due parti. Nella prima parte, dopo aver selezionato un sottoinsieme dei nodi dell'interattoma, associamo ad ognuno di essi un'informazione iniziale che riflette il "grado" di alterazione del gene,

per esempio la sua frequenza di mutazione o il fold change della espressione genica. L'algoritmo di diffusione propaga l'informazione iniziale nel network raggiungendo, dopo un transiente, lo stato stazionario. A questo punto, la quantità di fluido in ciascun nodo è utilizzata per costruire un ranking dei geni. Nella seconda parte, i *disease modules* sono identificati mediante una procedura di network resampling.

L'analisi condotta ci ha permesso di identificare un numero consistente di geni già noti nella letteratura relativa ai tre tipi di cancro studiati, nonché un insieme di altri geni correlati a questi, attualmente non citati in letteratura, che potrebbero essere interessanti candidati per ulteriori approfondimenti in studi futuri. Attraverso una procedura di *Gene Set Enrichment* abbiamo infine testato la correlazione dei moduli identificati con pathway biologici noti.

Chapter 1

Introduction

1.1 Diffusion and Laplacian Dynamics on Network

In physics, biology, and social sciences, many systems of interest can be modelled as network whose nodes represent the system components and links the interactions among them. Two main branches of networks theory concern, on the one hand, the characterization of their structural properties, on the other hand, the study of dynamical processes defined on them. The connection between the structure of a network and the behavior of dynamical processes, such as the diffusion of a substance, is a cross-cutting and expanding field of research. The general objectives are to characterize the dynamics of a given system by changing the network topology and, conversely, the extrapolation of certain structural characteristics of the network by means of a particular dynamical system defined on the network itself [15] [24].

In addition to having a considerable theoretical interest, this field has important practical implications, particularly in the biomedical field. In fact, the advent of high-throughput experiments has provided a huge number of large biological network databases, for example, the network of proteins in the cell (PPI network) or the complete mapping of neuronal connections in the brain, with the consequent need for increasingly effective techniques to extract useful information.

Various methods that exploit the properties of dynamical systems defined on a biological network have been introduced for the analysis and integration of omics data. In particular, several variants of systems which simulate the diffusion of a quantity within a network have been successfully proposed for the identification of disease causal genes.

The central idea is the following. An initial information, which reflects its *"degree" of alteration*, is associated to each gene. The diffusion algorithm propagates such information within the PPI network until it reaches a steady state. The stationary distribution is than used to define a similarity measure between genes.

Beyond their biological applications, the study of diffusive systems on network is interesting from a more general point of view. In fact, they are strictly related to the network laplacian matrix and to other physical systems like resistors and coupled oscillators networks. Moreover, the transient states of the diffusive dynamics turned out to have a role in network clustering.

1.2 Biological Motivation of the Study

A central issue in current biomedical research is the identification of causal genes underlying human disease.

In the last decades, various kinds of large-scale biological data have been made available by high-throughput experiments. They contain useful information for understanding how living systems work and have been proved useful for developing new methods in disease diagnosis and treatment.

These data can be classified in two main groups. On the one hand we have *biological network data* such as protein-protein interaction networks, gene regulatory networks or metabolic pathways. They represent the patterns of interaction among molecular entities in living systems and their study allows the functional characterization of such entities.

On the other hand, we have data provided by *genome wide studies* such as Gene Expression, Somatic Mutation and Methylation profiles. Such data provide large lists of altered genes related to a query disorder or a biological process.

Integration of genome-scale molecular information obtained by genome wide studies and biological networks allows to tackle several issues in current system biology.

Firstly, only a small fraction of genes among those obtained by high-throughput experiments are truly relevant to the disease or the biological process of interest. This is particularly true for tumours where altered genes vary wildly between patients. In order to get useful information from such data we need to rank genes according to prior biological knowledge.

From a more general point of view, a central goal in systems biology is the definition of the network regions associated to biological functions and disease. In this problem given a set of molecular entities of interest, for example the proteins associated to a partially known pathway, we want to find other entities related to it.

1.2.1 Cancer Altered Gene Prioritization

Cancer, as well as neurological disorders and diabetes, is a complex or multifactorial medical condition. These conditions are referred to as *multifactorial* because they are attributed to the combinatorial effects of genetic variation at a number of different genes.

In cancer research high-throughput data are being increasingly used. In order to get a better knowledge of such diseases researcher are profiling cancer at different layer of *omic* information (e.g. somatic mutation or gene expression).

These studies have revealed that the set of altered genes in cancer (sometimes referred to as *cancer genome landscape*, [37], fig.1.1) consists of a small number of genes altered in a high percentage of tumors (for example the oncogene TP53) and a large number of genes altered infrequently. Moreover, combinations of involved genes change enormously between patients.

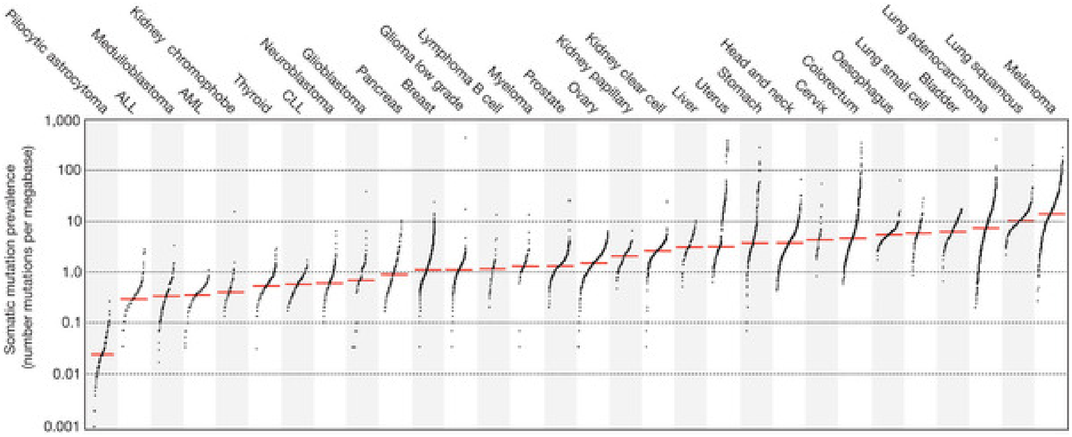


FIGURE 1.1: *Distribution of the number of mutations in cancer [1]*

This heterogeneity makes the identification of altered driver genes (genes that contain mutations or are expressed aberrantly in a fashion that confers a selective growth advantage to the tumor cells) from ones altered in a passenger fashion (alteration occurring during the growth of normal cells that have no effect on the cell growth) a challenging problem.

When a gene is altered in a high number of samples any statistic will indicate that the gene is extremely likely to be a driver gene. However, less mutated genes are more numerous. In these cases because of the high variability of the background rates of mutation among different patients, the frequency of alteration is not a good measure of gene importance, especially with the low number of samples available in actual cancer studies.

1.2.2 Complex Membership Issue and *"Hot"* Modules Identification

Within a living system, i.e cell, molecular entities rarely work alone. Many times they interact with one another to carry out biological functions. Functional similarity is associated to network proximity and entities involved in the same disease are more likely to interact with one another. For this reason, network proximity measures are valuable tools for prediction of molecular species function and disease association.

In general, the complex membership issue, can be formulated as follow. Given a set of query genes (or proteins), such as the mutated genes in a cancer, we are interested in finding a set of other genes (or proteins) highly related to these, possibly ranked according to their proximity to the query complex.

More specifically, in cancer research one is interested in finding connected network regions, usually referred to as "Hot Subnetwork" or "Disease modules" carrying the most important molecular alterations.

Cancer is a disease of pathways. In other words, there are relatively few key pathways whose perturbation transforms a normal cell into a tumor cell. Cancer can perturb each pathway employing many different combinations of gene alterations. "Hot subnetworks" are likely to be related to such key pathways.

This also explains the wild variability of alteration landscape in the same cancer and shows the link between the gene prioritization issue and the search for altered subnetworks. An altered gene is more likely to be a tumour driver if it is in the neighbourhood of other altered genes.

Both problems have been addressed integrating genome-wide molecular profiles with biological networks.

1.3 Outline of the Work

In the first part of this thesis, we review the theory underlying diffusive process on network, finding some interesting connection to clustering techniques and to other physical processes described by Laplacian dynamics. Moreover, we review the existing network diffusion algorithms for *disease modules* detection.

Then, we test a novel diffusion technique, recently proposed in literature, for detection of sub-modules of PPI network enriched in altered genes. We applied the algorithm to Somatic Mutation and Gene expression profiles of three different cancer types: Acute Myeloid Leukemia, Colon adenocarcinoma and Gastrointestinal Stromal tumour.

Finally, we discuss the results and we assess, by a *gene set enrichment analysis*, the association of the detected disease modules to known biological pathways.

Chapter 2

Network Laplacian Dynamics

In this chapter we introduce the theory underlying the network diffusion algorithms for omics data mining.

As a first step we define the *Combinatorial Laplacian* of a graph and other matrices that are strictly related to it. Then, in order to have a better insight into the their physical meaning we introduce some simple physical models, focussing on random walk and diffusion processes. Finally, we show the connections among these models and their link to a widely used clustering algorithm, based on the *Fiedler vector*.

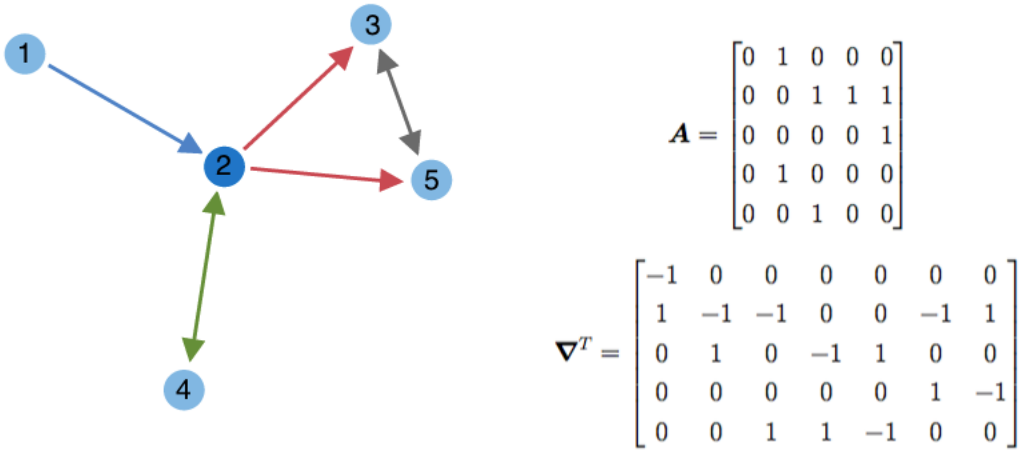
2.1 Definitions

2.1.1 Combinatorial Laplacian

Given a network $G(V, E)$, directed or undirected, with n vertices and m links there are two ways to represent it in a matrix form, the *adjacency matrix* and the *incidence matrix*. The *adjacency matrix*, $\boldsymbol{A}$, is a $n \times n$ square matrix that has elements $A_{ij} = 1$ if nodes i and j are neighbours and zero otherwise. The *incidence matrix*, $\boldsymbol{\nabla}$ is an $m \times n$ matrix in which rows represent links and columns represent nodes. The j-th element of the l-th row is $\nabla_{lj} = 1$ if l is an incoming link of node j, while $\nabla_{lj} = -1$ if l is an out-coming link of node j, fig. 2.1.

If $f(i)$ for $i \in V$ is a function defined on each node, the incidence matrix associates at a given link, $l \in L$, the difference of the values of f at the two end-points of the link:

$$\Delta_l = \sum_i \nabla_{li} f_i \tag{2.1}$$



$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\nabla^T = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & -1 & 1 \\ 0 & 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 & -1 & 0 & 0 \end{bmatrix}$$

(a) (b)

FIGURE 2.1: *Directed graph with its Adjacency* ($\boldsymbol{A}$) *and Incidence* ($\boldsymbol{\nabla}$) *Matrix*

Therefore, $\boldsymbol{\nabla}$ can be considered as a kind of *"discrete differential"* operator on the graph. It appears natural to define the analogous of the continuous laplacian operator on a graph as follow:

$$\partial \cdot \partial f \longrightarrow \boldsymbol{\nabla}^T \cdot \boldsymbol{\nabla} \vec{f} = \boldsymbol{L} \vec{f} \tag{2.2}$$

L is called *Combinatorial Laplacian* or simply *Laplacian* of the network. Note that it is symmetric and that $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$:

$$\boldsymbol{L} = \boldsymbol{\nabla}^T \cdot \boldsymbol{\nabla} = \boldsymbol{D} - \boldsymbol{A} = \begin{bmatrix} d_1 & -a_{12} & \dots \\ -a_{21} & d_2 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{2.3}$$

where $\boldsymbol{D}$ is the diagonal degree matrix. Moreover, we have:

$$(\boldsymbol{L} f)(i) = \sum_{j \sim i} [f(i) - f(j)] \tag{2.4}$$

i.e., the laplacian of a function at a node i is the sum of the differences between the function at i and the function at its neighbours [6].

2.1.2 Diffusive Laplacians

A matrix closely related to the *Combinatorial Laplacian* is the *Diffusive Laplacian*, $\boldsymbol{L}'$, representing the average difference between a node i and its neighbours:

$$(\boldsymbol{L}'f)(i) = \frac{1}{d(i)} \sum_{j \sim i} [f(i) - f(j)] \tag{2.5}$$

$\boldsymbol{L}'$ is related to the adjacency matrix and to the combinatorial laplacian by the relation:

$$\boldsymbol{L}' = \boldsymbol{D}^{-1}\boldsymbol{L} = \boldsymbol{D}^{-1}(\boldsymbol{D} - \boldsymbol{A}) = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A} \tag{2.6}$$

Another related matrix is the *Symmetric Diffusive Laplacian*, $\boldsymbol{\mathcal{L}}$, [8] defined as:

$$\boldsymbol{\mathcal{L}} = \begin{cases} 1 & : i = j \\ -\frac{1}{\sqrt{d(i)d(j)}} & : i \sim j \\ 0 & : otherwise \end{cases}$$

It is related to $\boldsymbol{L}$ through the relation:

$$\boldsymbol{\mathcal{L}} = D^{-\frac{1}{2}}\boldsymbol{L}D^{-\frac{1}{2}} \tag{2.7}$$

and is similar to $\boldsymbol{L}'$:

$$\boldsymbol{\mathcal{L}} = D^{-\frac{1}{2}}\boldsymbol{L}'D^{\frac{1}{2}} \tag{2.8}$$

2.2 Random Walk and Diffusion on Network

2.2.1 Simple Random Walk (RW)

A random walk on a network is defined as follow [22] [25]. Let $G = (V, E)$ be a network where V is the set of n nodes and E is the set of m edges. We are considering unweighted edges for simplicity. Let P_0 be the probability of starting at a node v_0. If at the t-th time step we are at a node v_t, we move to a neighbouring site of v_t with probability $\frac{1}{d(v_t)}$, where $d(\cdot)$ is the degree of the node.

We denote by $P_t(i)$ the probability of being at i at time t. The matrix of transition probabilities (π^T) is given by:

$$\boldsymbol{\pi}^T_{ij} = \begin{cases} \frac{1}{d(i)} & : i \sim j \\ 0 & : otherwise \end{cases}$$

Note that $\boldsymbol{\pi}^T = \boldsymbol{D}^{-1}\boldsymbol{A} \rightarrow \boldsymbol{\pi} = \boldsymbol{A}\boldsymbol{D}^{-1}$, where $\boldsymbol{D}$ is the diagonal degree matrix and $\boldsymbol{A}$ is the adjacency matrix of the network. Moreover, $\boldsymbol{\pi}$ is column normalized. The evolution of the probability distribution for the random walker is given by:

$$\vec{P}(t + \Delta t) = \boldsymbol{\pi}\vec{P}(t) \tag{2.9}$$

Starting from the initial distribution $\vec{P}_0$, after a time $t = k\Delta t$ we have:

$$\vec{P}(t = k\Delta t) = \boldsymbol{\pi}^k\vec{P}_0 \tag{2.10}$$

The stationary distribution is given by:

$$\vec{P}_{st} = \frac{\vec{d}}{2m} \tag{2.11}$$

that is, $\vec{P}_{st}$ is proportional to the degree vector, $\vec{d}$, and does not depend on the initial distribution $\vec{P}_0$.

It is possible to define a new transition matrix as follow:

$$\boldsymbol{\pi}' = \boldsymbol{D}^{-1/2}\boldsymbol{\pi}\boldsymbol{D}^{1/2} \tag{2.12}$$

$\boldsymbol{\pi}'$ is symmetric and has the same eigenvalues and related eigenvectors of $\boldsymbol{\pi}$. If $\vec{v}$ is an eigenvector of $\vec{\pi}'$ with eigenvalue λ, then $\vec{q} = \boldsymbol{D}^{1/2}\vec{v}$ is an eigenvector of $\boldsymbol{\pi}$ with the same eigenvalue.

2.2.2 Lazy Random Walk (LRW)

It is sometimes convenient to consider a variation of random walk, in which in each step, with probability $1 - \beta$, we stay at the current vertex and only with probability β we make the usual step of random walk. This variation is called *lazy random walk*, the evolution is given by:

$$\vec{P}(t + \Delta t) = \beta\boldsymbol{\pi}\vec{P}(t) + (1 - \beta)\boldsymbol{I}\vec{P}(t) = \tilde{\boldsymbol{\pi}}\vec{P}(t) \tag{2.13}$$

with:

$$\tilde{\pi}_{ij} = \begin{cases} 1-\beta & : i = j \\ \pi_{ij}\beta & : i \neq j \end{cases}$$

It is possible to show that equations 2.9 and 2.13 have the same stationary distribution.

2.2.3 Continuous Time Random Walk (CTRW)

We want to perform the continuous time approximation of the evolution equation [25]. In the limit $\Delta t \to 0$ we expect that the probability of a walker to change state tends to zero. Starting from eq. 2.13 and taking $\beta = \beta_0 \Delta t + o(\Delta t)$, the transition probabilities become:

$$\tilde{\pi}_{ij}(\Delta t \to 0) = \begin{cases} 1-\beta_0\Delta t + o(\Delta t) & : i = j \\ \pi_{ij}\beta_0\Delta t + o(\Delta t) & : i \neq j \end{cases}$$

with the condition:

$$\sum_i \tilde{\pi}_{ij}(t) = 1 \tag{2.14}$$

The probability of being at node i at time $t + \Delta t$ can be written as:

$$P_i(t+\Delta t) = \sum_j P_j(t)\pi_{ij}\beta_0\Delta t + (1-\beta_0\Delta t)P_i(t) \tag{2.15}$$

which gives:

$$\frac{P_i(t+\Delta t) - P_i(t)}{\Delta t} = \beta_0 \sum_j P_j \pi_{ij} - P_i = \beta_0 \sum_j \left[\pi_{ij} - \delta_{ij}\right] P_j(t) \tag{2.16}$$

The *Diffusive Laplacian* is defined as:

$$\boldsymbol{L}' = (\boldsymbol{D} - \boldsymbol{A})\boldsymbol{D}^{-1} = \boldsymbol{I} - \boldsymbol{\pi} \to L'_{ij} = \delta_{ij} - \pi_{ij} \tag{2.17}$$

and introducing the vectorial notation $\vec{P}(t) = (P_i)$ for $i = 1, 2, \ldots, n$, in the limit $\Delta t \to 0$, we have:

$$\frac{d\vec{P}(t)}{dt} = -\beta_0 \boldsymbol{L}'\vec{P}(t) \tag{2.18}$$

Since the eigenvalues of $\boldsymbol{L}'$ are all positive except the null one, the stationary distribution, $\vec{P}_{st}$, is unique and attractive.

2.2.4 Open Source/Sink Model

Equation 2.18 describes a closed isolated system in which starting from any initial probability distribution the dynamics relaxes to the same stationary solution, $\vec{P}_{st}$. Introducing a *souce/sink* perturbation [4] it is possible to transform the closed system in an open one, described by the equation:

$$\frac{d\vec{\phi}(t)}{dt} + \boldsymbol{L}'\vec{\phi}(t) = \gamma_1\vec{\phi}^0 - \gamma_2\boldsymbol{I}\vec{\pi}^{out} \cdot \vec{\phi} \tag{2.19}$$

where $\vec{\phi}^0 = (\phi_1^0, \ldots, \phi_n^0)$ is the source vector, in which the ratio $\frac{\phi_i^0}{\phi_j^0}$ weights the relative importance of the sources, and $\vec{\pi}^{out} = (\pi_1^{out}, \ldots, \pi_n^{out})$ is the sink vector, with $\pi_i^{out} = 1$ if i is a sink and $\pi_i^{out} = 0$ otherwise. γ_1 and γ_2 are two constants.

The system is open in the sense that the probability conservation is no longer guaranteed. We changed the variable from $P\vec{(t)}$ to $\vec{\phi}(t)$ to highlight this fact. $\vec{\phi}(t)$ can be interpreted as an hypotetical fluid diffusing on the network with laplacian dynamics. $\vec{\phi}(t = 0)$ is the initial distribution of fluid in the nodes. Some nodes are selected to serve as sources, where the fluid is pumped in at a constant rate.

At sink nodes the system loses fluid at constant first order rate. Rewriting the previous equation as:

$$\frac{d\vec{\phi}(t)}{dt} + (\boldsymbol{L}' + \gamma_2\boldsymbol{I}\vec{\pi}^{out}) \cdot \vec{\phi}(t) - \gamma_1\vec{\phi}^0 = 0 \tag{2.20}$$

we get the stationary solution:

$$\vec{\phi}_{st} = (\boldsymbol{L}' + \gamma_2\boldsymbol{I}\vec{\pi}^{out})^{-1}\gamma_1\vec{\phi}^0 \tag{2.21}$$

Note that $\vec{\phi}_{st}$ depends on the source vector but not on the initial distribution of the fluid.

2.2.5 Random Walk with Restart (RWR)

In the random walk described above the walkers, at each time step can only move from the current node to one of its neighbors. In the *random ralk with restart* they may also choose to teleport to the start node with a given probability. If we call $(1-\alpha)$ the *restart probability* the evolution equation is:

$$\vec{P}(t+\Delta t) = \alpha\boldsymbol{\pi}\vec{P}(t) + (1-\alpha)\vec{P}_0 \tag{2.22}$$

The stationary solution of eq.2.22 is given by:

$$\vec{P}_{st} = (\boldsymbol{I} - \alpha\boldsymbol{\pi})^{-1}(1-\alpha)\vec{P}_0 \tag{2.23}$$

Note that the stationary distribution depends on the initial distribution $\vec{P}_0$. Moreover, for large networks the inversion of $(\boldsymbol{I} - \alpha\boldsymbol{\pi})$ is computationally expensive. Computing $\vec{P}_{st}$ iteratively using equation 2.22 is less expensive for large networks and converges to the solution 2.23.

Connection between RWR and the Source/Sink Model

Starting from equation 2.20 we assume a sink at each node and $\gamma_1 = \gamma_2 = \gamma$, getting:

$$\frac{d\vec{\phi}(t)}{dt} + (\boldsymbol{L}' + \gamma\boldsymbol{I})\cdot\vec{\phi}(t) - \gamma\vec{\phi}^0 = 0 \tag{2.24}$$

remembering that $\boldsymbol{L}' = \boldsymbol{I} - \boldsymbol{\pi}$ and defining $\alpha = \frac{1}{1+\gamma}$:

$$\frac{d\vec{\phi}(t)}{dt} = -(\boldsymbol{I}(1+\gamma) - \boldsymbol{\pi})\cdot\vec{\phi}(t) + \gamma\vec{\phi}^0 = -(\frac{1}{\alpha}\boldsymbol{I} - \boldsymbol{\pi})\cdot\vec{\phi}(t) + \frac{1-\alpha}{\alpha}\boldsymbol{\phi}^0 \tag{2.25}$$

Rescaling time:

$$\frac{d\vec{\phi}(t')}{dt'} = -(\boldsymbol{I} - \alpha\boldsymbol{\pi})\cdot\vec{\phi} + (1-\alpha)\vec{\phi}^0 \tag{2.26}$$

For discrete time step it is easy to see that the previous equation becomes formally equivalent to the RWR evolution equation 2.22:

$$\vec{\phi}(t+\Delta t) = \alpha\boldsymbol{\pi}\vec{\phi}(t) + (1-\alpha)\vec{\phi}^0 \tag{2.27}$$

Remark: $\vec{P}_0$ in 2.22 is the probability distribution of the walker to be at a given node at time $t = 0$ while $\vec{\phi}^0$ in 2.27 gives the incoming flow rate at each node.

In order to interpret $\vec{\phi}^0$ as a starting probability distribution it has to satisfy the normalization condition. Moreover, the total amount of fluid (probability) in the system must be conserved. This condition is satisfied by equation 2.25 due to the assumption $\gamma_1 = \gamma_2$. In fact, at the steady state the incoming flow and the outgoing flow of the system 2.20 must be equal:

$$\gamma_1 \sum_i \phi_i^0 = \gamma_2 \sum_i \phi_i^{st} \xrightarrow{\gamma_1=\gamma_2} \sum_i \phi_i^0 = \sum_i \phi_i^{st} \tag{2.28}$$

2.2.6 Diffusion or Heat Kernel

The so called *diffusion or heat kernel* can be derived starting from the continuous time random walk evolution equation. An equation equivalent to 2.18 holds for the transition matrix $\pi(t)$:

$$\frac{d\pi(t)}{dt} = -\beta_0 L \cdot \pi(t) \to \pi(t) = \pi(0)e^{-L\beta_0 t} \tag{2.29}$$

$\pi(t)$ is the propagator from $t' = 0$ to $t' = t$. If we assume $\pi(0) = I$, we get:

$$\pi(t) = e^{-\beta_0 L t} \tag{2.30}$$

$\pi(t)$ defines a global similarity metric between the nodes in the network. Note that $\beta = \beta_0 t$ is a parameter that controls the magnitude of the diffusion.

If $P(0)$ is the distribution of particles at time $t' = 0$, at time $t' = t$ we have:

$$P(t) = P(0)e^{-\beta_0 L t} \tag{2.31}$$

In a network where each node has the same connectivity degree, d, for instance an infinite regular lattice, the *Diffusive Laplacian*, $\boldsymbol{L}' = (\boldsymbol{D} - \boldsymbol{A})\boldsymbol{D}^{-1}$, is proportional to the *combinatorial Laplacian*, $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, and equation 2.30 can be rewritten as:

$$K(t) = e^{-\beta' \boldsymbol{L}} \tag{2.32}$$

where $\beta' = \beta_0 t d^{-1}$. Since the matrix $\boldsymbol{L}$ is symmetric K is also symmetric and can be interpreted a *kernel* function on the network.

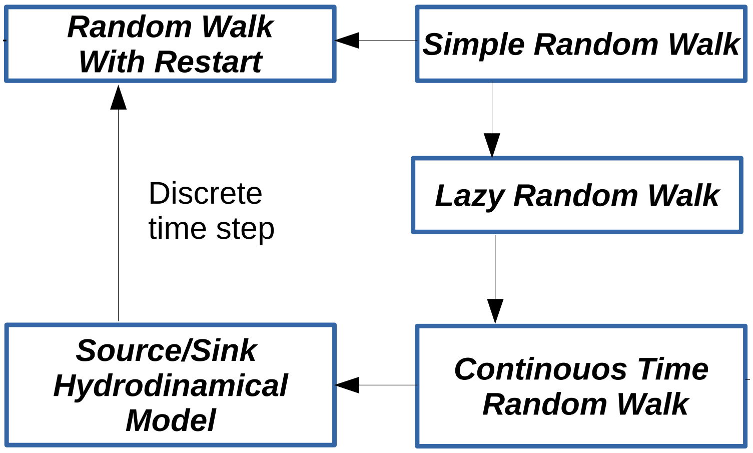


FIGURE 2.2: *Connection between various kind of random walk and Source/Sink Model.*

2.2.7 Fick's Law, Transient States and Network Clustering

Suppose we have a substance on the vertices of the network $G(V, E)$ and that it moves along each edge with a rate $r_{ij}(t) = C\,[\phi_i(t) - \phi_j(t)]$, where $\phi_i(t)$ is the amount of such substance on the vertex i at time t, therefore we have:

$$\frac{d\phi_i(t)}{dt} = -C\sum_j A_{ij}\,[\phi_i(t) - \phi_j(t)] \tag{2.33}$$

that can be rewritten as:

$$\frac{d\phi_i(t)}{dt} = -C\sum_j A_{ij}\phi_i(t) + C\sum_j A_{ij}\phi_j(t) = C\sum_j (A_{ij} - \delta_{ij}k_i)\phi_j \tag{2.34}$$

that in matrix form is:

$$\frac{d\vec{\phi}(t)}{dt} = C(\boldsymbol{A} - \boldsymbol{D})\vec{\phi}(t) = -C\boldsymbol{L}\vec{\phi}(t) \tag{2.35}$$

Note that:

$$\sum_j \frac{d\phi_j(t)}{dt} = C\sum_i \left[\sum_j (A_{ij} - \delta_{ij}k_i)\phi_j\right] = 0 \tag{2.36}$$

that is, the total amount of the substance on the network is conserved.

Moreover, equation 2.35 is equivalent to the evolution of the probability distribution of a continuous time random walk, eq. 2.18, only on *regular* networks for which $\boldsymbol{D} = q\boldsymbol{I}$, where q is the coordination number [26]. In order to derive the Fick's law as *random walk master equation* we have to take in account an *'edge centric'* random walk in which the probability for the random walker to change node is proportional to the node degree [25].

Transient States

Let us consider the eigenvalue problem for the laplacian $\boldsymbol{L}$:

$$\boldsymbol{L}\vec{v_i} = \lambda_i \vec{v_i} \tag{2.37}$$

where λ_i are the eigenvalues and v_i the related eigenvectors, it can be shown that the smallest eigenvalue is $\lambda_0 = 0$ and that $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{n-1}$. The smallest non-zero eigenvalue is called *Fiedler number* and the related eigenvector *Fiedler vector* [13].

If we write $\phi_i(t)$ and $\phi_i(0)$ as linear combinations of the eigenvectors of $\boldsymbol{L}$ we have:

$$\vec{\phi}(t) = \sum_i a_i(t)\vec{v_i} \tag{2.38}$$

$$\vec{\phi}(0) = \sum_i a_i(0)\vec{v_i} \tag{2.39}$$

From equation 2.35 and 2.38 we have:

$$\frac{da_i(t)}{dt} = -C\lambda_i a_i(t) \tag{2.40}$$

which has the solution:

$$a_i(t) = e^{-C\lambda_i t} a_i(0) \tag{2.41}$$

Equation 2.38 can be rewritten as:

$$\boldsymbol{\phi}(t) = a_0(0)e^{-C\lambda_0 t}\vec{v_0} + a_1(0)e^{-C\lambda_1 t}\vec{v_1} + \ldots \tag{2.42}$$

If the graph is connected then $\lambda_i > 0$ for all $i > 0$ and the stationary solution of equation 2.35 for $t \to \infty$ is unique and is proportional to the eigenvector with null eigenvalue, $\vec{v_0} = (1, 1, \ldots, 1)^T$.

The components of the solution along the eigenvectors with positive eigenvalues are *transient states* since they tend to zero as t approaches infinity. The decay time constant of the component along the eigenvector $\boldsymbol{v}_i$ is proportional to the inverse of λ_i.

Relation to Spectral Network Clustering

Most networks display community structure, that is, their vertices are organised into groups, called clusters or modules. Detecting clusters is of great importance in network theory since it allows for a classification of vertices, according to their structural position in the modules.

A common clustering technique is the *Spectral Method* [26]. Given a network, we are interested in finding a bipartition such that the number of links between the identified subnetworks is minimum, in other words, we look for a bipartition which minimizes the cut size, R:

$$R = \sum_{i,j} \frac{1}{2} A_{ij} \tag{2.43}$$

where i and j refer to nodes in different subnetworks. It can be shown that such bipartition is given by the sign of the components of the *Fiedler vector,i.e.*, each node is assigned to a different subnetwork according to the positive or negative sign of the corresponding component. Moreover, the norm of the second eigenvalue λ_1 is a measure of the network connectivity, usually called *algebraic connectivity*, the bigger λ_1 the more connected the graph.

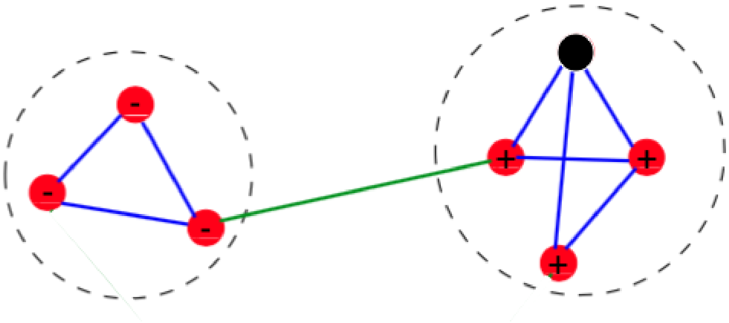


FIGURE 2.3: *Network communities. A minimum cut bipartition divides the network in two part cutting the minimum number of links. If a substance diffuses on the network starting from the black node, according to Fick's law, after a time of the order of $1/\lambda_2$, the amount of fluid will be higher in the nodes which belong to the same component of the starting node.*

Let us translate this result in the language of Fick's law. From equation 2.42 we can see that the component of the solution of Fick's law along the

Fiedler vector is the more lasting transient state. Thus, minimum cut bipartition can be interpreted as follows. Starting with the substance in one node, i, and let it to diffuse within the network, after some time, the nodes in the two minimum cut subnetworks will have, respectively, an higher and a lower content of substance with respect to the stationary distribution 2.3.

Starting from this observation, the spectral clustering method could be generalized considering the successive eigenvectors and using time as a parameter which regulates the size and the number of clusters.

2.3 Other Laplacian Physical Systems on Network

2.3.1 Network of Coupled Oscillators

It is possible to give a physical interpretation of the combinatorial laplacian considering a network in which nodes are rigid spheres and links are springs [14] [6].

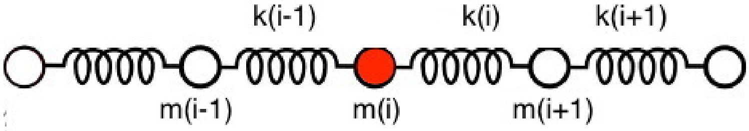


FIGURE 2.4: *1-d armonic oscillators chain*

For sake of simplicity, let us consider the 1-dimensional chain of coupled oscillators, fig. 2.4, in which the sphere i has mass m_i and the spring i has elastic constant k_i, the equation of motion for the i-th sphere is given by:

$$m_i \ddot{u}_i = F(u_i) \tag{2.44}$$

where:

$$F(u_i) = F_i - F_{i+1} = -k(u_i - u_{i-1}) - k(u_i - u_{i+1}) = -k \sum_j L_{ij} u_j \tag{2.45}$$

which in matrix form is:

$$\ddot{\vec{u}} = -\frac{k}{m} \boldsymbol{L} \vec{u} \tag{2.46}$$

The equation of motion for a vibrating string can be derived as the continuous limit for $N \to \infty$ and $\Delta x \to 0$, where N is the number of sphere and

Δx is the distance between two sphere:

$$\ddot{u} = v^2 \frac{\partial^2 u}{\partial x^2} \tag{2.47}$$

As we can see the laplacian matrix is the discrete analogous of the second derivatives.

2.3.2 Electric networks

We conclude the chapter showing the relation between resistor network and random walks.

The Dirichlet problem on Network

The continuous Laplace's equation is given by:

$$\Delta u = 0 \tag{2.48}$$

where $\Delta = \nabla^2$ is the continuous laplacian operator. A function u that satisfies Laplace's equation is called *harmonic*. Such functions have a *mean value property*: the value $u(x, y)$ equals the average of the values over any circle centred at $u(x, y)$. The converse is also true: a function satisfying the mean value property is harmonic [12].

If we take the mean value property from the continuous case as the definition of an harmonic function in the discrete case, it turns out that a function, $u(i)$, defined on the nodes of a graph is harmonic, if for every nodes i holds:

$$u(i) = \frac{1}{d(i)} \sum_{j \sim i} u(j) \tag{2.49}$$

that can be rewritten as:

$$Lu = 0 \tag{2.50}$$

where L is the combinatorial laplacian, or as:

$$L'u = 0 \tag{2.51}$$

where $L' = I_d - D^{-1}A$ is the diffusive laplacian. A function satisfying the previous equation at every nodes must be constant. A non-constant function on a finite connected graph cannot be harmonic at every node. This fact also holds in the continuous case.

An interesting problem for application consists in finding the solutions to Laplace equation satisfying a boundary condition. In general, given a partial differential equation (PDE) and a domain, D, with boundary, ∂D, the Dirichlet problem is the problem of finding a function satisfying the PDE inside the region and taking prescribed values on the boundary.

Starting from a network $G(V, L)$ we may consider the induced subgraph $G(V')$ on a subset $V' \subset V$, considering $\partial V = V/V'$ as the boundary of $G(V')$ on which the constraint $u(i) = h(i)$ is enforced.

The discrete Dirichlet problem for the Laplace equation on a graph, G, can be defined as the problem of finding a harmonic function, u, on $G(V')$, such that $u = h$ on ∂V.

It is possible to reinterpret this problem and find a solution in terms of random walk on a network with *adsorbing states*, i.e. states that once entered cannot be left.

Let us assume that $\boldsymbol{P}$ is the transition matrix for a random walk an the network G, that the nodes in ∂V are absorbing states and $n_a = \|\partial V\|$ is the number of such states. Reordering the states so that the adsorbing states come first, $\boldsymbol{P}$ can be rewritten in the form:

$$\boldsymbol{P} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \boldsymbol{R} & \boldsymbol{\pi}^T \end{bmatrix} \tag{2.52}$$

where $\mathbf{1}$ in the n_a-by-n_a identity matrix, $\boldsymbol{\pi^T}$ is the transition probability matrix (row-normalized) for a simple random walk restricted to the graph $G(V')$ and $\boldsymbol{R}$ is the matrix of probabilities of going to adsorbing states from all the non-adsorbing states.

If $u(i)$ is a function defined on the nodes of G that satisfies:

$$u(i) = \sum_j P_{ij} u(j) \tag{2.53}$$

on the nodes of $G(V')$, then u has the mean value property and thus is harmonic. This means that the right eigenfunction of eigenvalue 1 of the

Markov chain with transition matrix $\boldsymbol{P}$:

$$\boldsymbol{P}u = u \tag{2.54}$$

is the solution for the discrete Dirichlet problem. If we write $u = [\boldsymbol{u_B}, \boldsymbol{u_D}]^T$, where $\boldsymbol{u_B}$ and $\boldsymbol{u_D}$, are, respectively, the solution for boundary and non-boundary nodes, we have:

$$\begin{bmatrix} \boldsymbol{u_B} \\ \boldsymbol{u_D} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \boldsymbol{R} & \boldsymbol{\pi}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{u_B} \\ \boldsymbol{u_D} \end{bmatrix}$$

and the solutions $\boldsymbol{u_D}$ for non-boundary nodes is given by:

$$\boldsymbol{u_D} = (\mathbf{1} - \boldsymbol{\pi}^T)^{-1} \boldsymbol{R} \boldsymbol{u_B} \tag{2.55}$$

The element B_{ij} of the matrix $B = (\mathbf{1} - \boldsymbol{\pi}^T)^{-1} \boldsymbol{R}$ gives the probability that a chain starting at a non-boundary point i will end up in an absorbing state j.

Resistor Network and Random Walk

Let us consider a general resistor network [12] and assign to each link a resistance, R_{ij}. The conductance is thus given by $C_{ij} = 1/R_{ij}$. Let us define a random walk on such network with transition probabilities:

$$P_{ij} = \frac{C_{ij}}{\sum_j C_{ij}} \tag{2.56}$$

Since the network is *eulerian*, i.e. the in-degree of each vertex is equal to its out-degree, the stationary distribution, $\boldsymbol{\phi} = \boldsymbol{\phi} P$, is unique and its elements are given by:

$$\phi_i = \frac{C_i}{\sum_j C_j} \tag{2.57}$$

Since $C_{ij} = C_{ji}$, at the stationary solution holds the *detailed balance condition*:

$$\phi_i P_{ij} = \phi_j P_{ji} \tag{2.58}$$

If we put a potential difference of 1 volt between two points, A and B, of the network such that $v_A = 1$ and $v_B = 0$, it is possible to give an interpretation of the electric potential at each node in terms of probabilistic quantities.

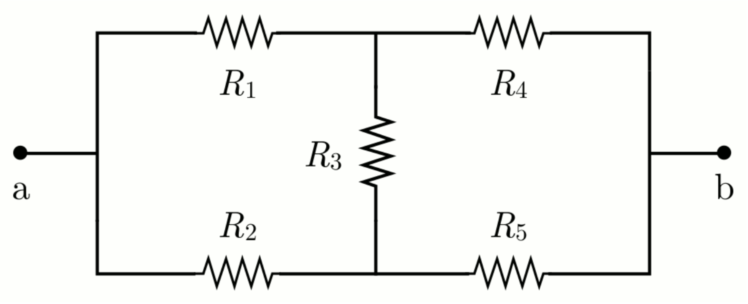


FIGURE 2.5: *Resistor Network*

The current, J_{ij}, through R_{ij}, is related to the potential difference between the end points of the resistor by Ohm's law:

$$J_{ij} = (v_i - v_j)C_{ij} \tag{2.59}$$

The Kirchoff's current Law requires that:

$$\sum_j J_{ij} = 0 \longrightarrow v_i \sum_j C_{ij} = \sum_j C_{ij} v_j \tag{2.60}$$

that can be rewritten as:

$$v_i = \sum_j \frac{C_{ij}}{C_i} v_j = \sum_j P_{ij} v_j \tag{2.61}$$

that is equivalent to say that $v(i)$ is harmonic at all points with the exception of $i = A, B$, i.e. it is the solution of the Dirichlet problem with boundary conditions $v_A = 1$ and $v_B = 0$.

As a consequence of the random walk interpretation of the Dirichlet problem the voltage at a point i in the network can be seen as the probability that a walker starting from i will and up in A before reaching B.

Currents can also be interpreted in a probabilistic fashion. Let us suppose that the electric particles enter the network at point A and leave the network at point B. A walker starts at A and make a random walk through the nodes until he reaches B. Let w_i be the probability of finding the walker at node i.

For non-adsorbing nodes, we have:

$$w_i = \sum_j P_{ji} w_j \tag{2.62}$$

From $R_{ij} = R_{ji}$ it follows that $C_i P_{ij} = C_j P_{ji}$ and thus:

$$w_i = \sum_j P_{ij} \frac{C_i}{C_j} w_j \to \frac{w_i}{C_i} = \sum_j \frac{P_{ij}}{C_j} w_j \tag{2.63}$$

this means that $\frac{w_i}{C_i}$ is harmonic and thus is the voltage at node i when we put $v_a = \frac{w_a}{C_a}$ and $v_B = 0$. The currents J_{ij} are given by:

$$J_{ij} = \frac{(v_i - v_j)}{R_{ij}} = w_i P_{ij} - w_j P_{ji} \tag{2.64}$$

The net number of times a walker passes a link can be interpreted as the current flowing between the end nodes of the link.

Chapter 3

Network Diffusion Based Methods for Omics Data Mining

Network diffusion methods exploits the global structure of a network by simulating the behaviour of a random walker. They have been introduced in graph theory to rank nodes based on their relative importance (*centrality measure*) [27].

Since genome-scale biological networks have become available many techniques have been proposed to mine these networks for functional assignment of molecular entities, disease and functional pathway discovery and community detection.

As we said above, functional similarity between molecular entities in a living system is related to network proximity. For this reason, the definition of *network proximity measures* plays a central role in biological network mining [28] [3].

The *shortest-path distance* (*SP*) is the simplest and most commonly used of such measures between two nodes on a graph. However, it has several drawbacks. Firstly, it is extremely sensitive to the insertion or deletion of single links. In biological applications, the connectivity information is derived from empirical data, and as such is subject to noise. For example, it has been hypothesized that PPI networks obtained by high throughput experiments may contain a significant number of false positive and negative interactions [11].

Moreover, nodes connected to each other via multiple paths are more likely to be functionally related than nodes that are connected via a single path [28] and the *SP* distance does not consider the global structure of the network, *i.e.* multiple paths between pairs of nodes.

For this reasons a more robust proximity measure, which involve averaging over many paths, is more suitable [32].

Random walks and diffusion on network are a good starting point for the definition of similarity measures that overcome this problems and several approaches based on them have been successfully proposed in many applications, ranging from gene prioritization to pathway detection.

In the first section of this chapter we review the most important methods reported in literature and their applications focusing ourselves on their physical meaning and connections. Then, we introduce the algorithm used in the next chapter for cancer datasets analysis.

3.1 Diffusion Methods and Biological Network

The most famous diffusion-based measure on network is *PageRank* [27] which was introduced for the ranking of web pages. The PageRank vector is the stationary distribution of a RWR (equation 2.22) in which the initial probability distribution,$\boldsymbol{P}_0$, is uniform. In other words, it allows random walkers to *"teleport"* to nodes different from their neighbours with a given constant probability.

In this section we give a detailed review of network-diffusion based methods for omics data mining. Such methods diffuse the initial information about molecular entities on a given biological network (mainly the PPI network) allowing the definition of *global similarity measures*

It is possible to arrange the methods described in literature in three main groups based on the diffusion process involved. The first group comprises techniques based on discrete time random walk with restart (RWR) and its variants such as the so called Network Propagation Algorithm (NPA) [18], [7], [16], [36] . Methods based on continuous time laplacian dynamics are in the second group [29],[35], [21] while the third group comprises *Heat Kernel Methods* [18]. As we have seen in the previous chapter the main goals of these methods are gene prioritization and identification of sub-modules, related to a given disease or a biological function, of biological interaction networks. In what follows we give a review of the main methods of the three groups and their applications.

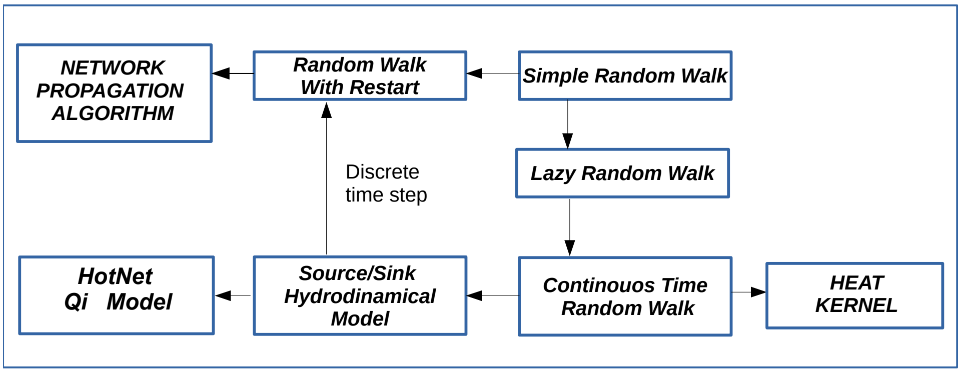


FIGURE 3.1: *Connection between random walk models on networks and algorithms for the analysis of biological network.*

RWR Techniques

This techniques are the direct generalization of the *PageRank* centrality. They have been introduced in the biological network context for protein structure prediction [38], pathway discovery and prediction of complex membership.

Can et.al [7], given a set of query protein on the PPI network, are interested in finding a set of other proteins ranked according to the probability of being related to the main complex. To achieve the task they use a random walk with restart defined as follow:

$$\boldsymbol{P}_{t+1} = \alpha\pi\boldsymbol{P}_t + (1-\alpha)\boldsymbol{P}_0 \tag{3.1}$$

where $\boldsymbol{P}_0$ is a vector with $P_0^i = 0$ for all its entries except for a given number, s, of query nodes where it takes the value $1/s$. As we have seen above, such a random walk converges to the stationary distribution P_{st} of eq. 2.23. Such distribution gives, for each node in the network, a global proximity measure to the query complex.

The same method can be used for the prioritization of candidate disease genes. For many disease it is possible to identify the chromosomal region in which unknown disease genes are located, but the regions could contain up to hundreds of candidate genes and it is often difficult to identify the correct disease gene by inspection of the list of genes within the interval. Kohler el al.[18] group diseases into families. For a given disease, starting from a set of known genes in its family, they use as a similarity measure the stationary distribution of equation 3.1. The initial vector $\boldsymbol{P}_0$ is constructed such that equal probabilities are assigned to the nodes representing known disease genes, with the sum of the probabilities equal to 1.

This method focus on prioritizing independent genes; however, in many cases, mutations at different loci could lead to the same disease. This genetic heterogeneity may reflect an underlying molecular mechanism in which the disease-causing genes form some kind of a functional module. Using a similar method, Vananu et al. [36] suggest a way to reveal the protein modules that are affected in a given disease. They use as starting distribution a prior knowledge function, $Y(v)$, which reflects the probability of gene v to be related to the query disease and propagate the information on the network. They choose to normalize the transition matrix, π, by the degrees of its end-points, since the latter relate to the probability of observing an edge between the same end-points in a random network with the same node degrees.

Several other application have been proposed such as: identification of biomarkers [30],[33] the study of virus-host molecular interactions and stratification of tumour mutations [16], identification of differentially enriched modules [5].

Continuous Time Source/Sink Laplacian Dynamics

An alternative global similarity measure between pairs of nodes can be defined using a continuous diffusion process, strictly related to RWR, on the interaction network [29].

Such a model considers the diffusion of a fluid on a network and is described by the equation:

$$\frac{d\boldsymbol{\phi}(t)}{dt} = -L\boldsymbol{\phi}(t) - \gamma\boldsymbol{\pi}^{out} \cdot \boldsymbol{\phi}(t) + \boldsymbol{\phi}^{0}\theta(t) \tag{3.2}$$

where θ is the step-function. Initially, all nodes contain no fluid. At $t = 0$ fluid is pumped into the source or query nodes at a constant rate and diffuses through the graph. Each node loses fluid at a constant first-order rate. After some time, the system reaches a stable state in which the amount of fluid in each node is constant. The better connected a node is to the query nodes, the more fluid it will contain at equilibrium. The final distribution of fluid in the network (ϕ_{st}, equation 2.21) gives a global similarity measure.

Hotnet [35] and *Hotnet2* [21] use this measure for cancer pathway discovery. *Hotnet* starting from somatic mutation profiles (i.e. cancer mutated genes as fluid sources) diffuses the information on the protein-protein interaction network and uses the similarity between nodes for the identification of "hot modules", i.e. connected subnetworks enriched in mutatated genes. A major drawback of *Hotnet* is due to the presence of high mutatated genes such as $TP53$. Such genes propagate their initial information to their neighboring nodes resulting in subnetworks containing many genes of poor biological interest. To overcome this drawback *Hotnet2* introduce a modified version of the diffusion model in which the transition matrix is normalized by the degree of the nodes.

Diffusion Kernel Methods

Given a set of objects $(x_1, x_2, \ldots,) \in X$ a *kernel* is a symmetic function that, given two objects x_i and x_j, returns a real number characterizing their similarity:

$$k : X \times X \mapsto \mathbb{R} \tag{3.3}$$

with:

$$k(x_i, x_j) = k(x_j, x_i)$$

If X is the set of nodes of a graph, the kernel $k(x_i, x_j)$ defines a global similarity measure between them [31]. The adjacency matrix of a network gives information only on local similarity. It expresses whether two entities are neighbors or not. A natural way of constructing a kernel from network local information is suggested by physical process of diffusion [19]. As we have seen, the transition matrix of a CTRW evolves in time according to equation 2.30. If we take $L = H = D - A$ we have:

$$K(t) = e^{-\beta H} \tag{3.4}$$

$K(t)$ is symmetric and defines a kernel on the graph called *Heat* or *Diffusion Kernel*. While in RWR and Laplacian Dynamics methods the similarity measure is given by the stationary distribution of the diffusion process, here the magnitude of the diffusion is controlled by the parameter β. Heat Kernel methods and its variants have been used for gene prioritization [18], biomarker signature discovery [9] [10] .

3.2 Diffusion Gene Prioritization and Detection of Disease Altered Subnetworks

In this section we introduce the method utilized in the next chapter for prioritization of altered cancer genes and detection of significantly altered subnetwork of PPI network. The method has been developed by Bersanelli et al. [5] for identification of differentially enriched modules, i.e. given two classes of sample the method aims to find those connected subnetworks that are altered in a different manner in the two classes. We use it in a slightly different fashion in that our data belong to a single class.

The interactions are modelled by a network $G(V, E)$. The edges, E, represent protein-protein interactions, the vertices, V, represent genes associated to individual proteins. A is the adjacency matrix of the network.

We associate to each gene initial information, described by the vector x^0. It is possible to consider different types of initial quantities, for example gene expression fold change or frequency of mutation in a set of samples. The pipeline is organized in two main steps. For the first step, the initial information is propagated in the network G according to a random walk with restart:

$$\boldsymbol{x}_{t+1} = \alpha\pi'\boldsymbol{x}_t + (1-\alpha)\boldsymbol{x}^{\mathbf{0}} \tag{3.5}$$

where π' is the symmetric transition matrix defined by $\pi' = D^{-1/2}\pi D^{1/2} = D^{-1/2}AD^{-1/2}$, in which each edge weight has been normalized by the product of the degrees of its end points. The parameter α, as we noted above, controls the relative importance of initial information ($\boldsymbol{x}_{\mathbf{0}}$) with respect to network information. According to the connection between equation 3.5 and the source/sink model we refer to the non-zero elements of $\boldsymbol{x}^{\mathbf{0}}$ as *"source"* nodes. After some time steps, the process converges to a stationary distribution, $\boldsymbol{x}^*$, that define a global similarity metric on the interaction network i.e. the quantities $\boldsymbol{x}_i^*$ rate each node according to its proximity to the source nodes.

In order to highlight the "local hypothesis", i.e. the increased probability of interaction of two genes involved in the same disease, Bersanelli et al. introduce the network smoothed index (NSI) for the gene i as follow:

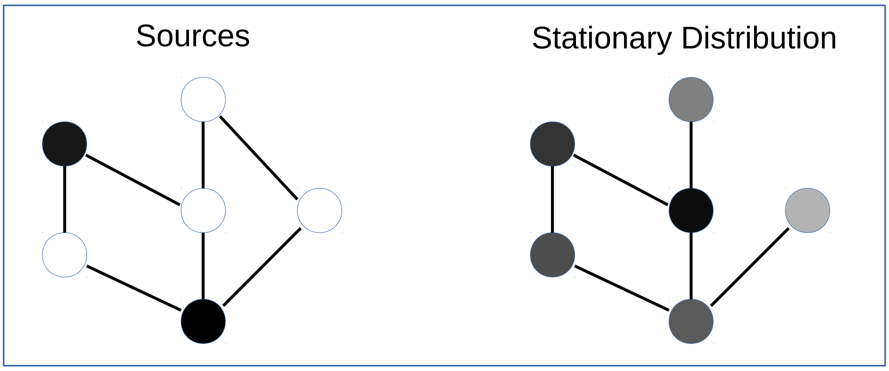


FIGURE 3.2: *Network Propagation Algorithm. Each node receives a quantity of fluid in relation to ist proximity do the sources and the topology of the network.*

$$S_i(\boldsymbol{x}^{\mathbf{0}}) = \frac{x_i^*}{x_i^0 + \epsilon} \tag{3.6}$$

Note that S_i is proportional to the gain of information of the node i. The parameter ϵ weighs the relative importance of $\boldsymbol{x}_{\mathbf{0}}$ with respect to $\boldsymbol{x}^*$. For $\epsilon \gg 1$ only $\boldsymbol{x}^*$ matters. A reasonable setting can be found in order to prioritize both sources and nodes in proximity to them. The setting depends on the amount of biological signal and has to be tuned depending on the problem under study.

As we have seen in section 3.1 a drawback of diffusion methods is the presence of hubs, i.e. nodes with high connectivity that assume high S because of the topology of the network. In order to overcome this problem two methods have been implemented. If node i is a hub, it has a high value of S_i independently of the distribution of the sources. Permuting the sources and propagating the information, we get a new value for the NSI, S_i^p. It is possible to define an empirical p-value, p_i, as the fraction of permutations with S_i^p equal or greater than the real S_i.

$$p_i = \frac{\mid (j = 1, \ldots, k \mid S_{ij}^p \geq S_i) \mid}{k} \tag{3.7}$$

where k is the number of permutations, S_{ij}^p is the permutated NSI for the node i and the j-th permutation.

At this point we can mitigate the effect of topology weighing the real S_i with the logarithm of p_i:

$$Sp_i(\boldsymbol{x^0}) = -log_{10}(p_i)S_i(\boldsymbol{x^0}) \tag{3.8}$$

Since p_i lies in $[0,1]$, $-log(p_i)$ lies in $[0,\infty]$.

The network smoothed index prioritizes PPI network nodes. The second step of the pipeline aims to find those regions in the network with the highest content of information. Starting from the list of genes, ranked according to Sp or ΔS, we define a function Ω of the top n entities of such list:

$$\Omega(n) = Sp^T(n) \cdot A_n \cdot Sp(n) \tag{3.9}$$

where the matrix A_n is the adjacency matrix of the subnetwork generated by the top n entities of the ranking. $\Omega(n)$ is the sum of the products between the NSI of connected nodes. A node not connected to any one of the other $n-1$ entities does not contribute to the sum. It the top ranked genes are related to one or more biological pathways we expect a high number of links among them. In what follows we describe a method to quantify the significance of the pattern of connection among the top n entities. Given a permutation, σ of the labels of the nodes (or equivalently, a random resampling of the existing connections which conserve the same degree distribution), we define the function:

$$\Omega_\sigma(n) = Sp^T(n) \cdot A_n^\sigma \cdot Sp(n) \tag{3.10}$$

where A_n^σ is the adjacency matrix after the permutation of the labels.

Taking k permutations, it is possible to define a *Network Resampling* p-value as the fraction of time the objective function of a permutation, $\Omega_{\sigma_j}(n)$, is greater or equal to the real network objective function, $\Omega(n)$:

$$p_{NR}(n) = \frac{\mid (j = 1, \ldots, k \mid \Omega_{\sigma_j}(n) \geq \Omega(n) \mid +1}{k+1} \tag{3.11}$$

As long as $p_{NR}(n)$ is of the order of 1 it means that a resampling of the connections among the first n top genes of the list do not change too much the strength of the subnetwork. Increasing n, it is reasonable to expect a sensible decrease of $p_{NR}(n)$ when top-scoring genes that are not connected to the previous $n-1$ ones enter the top of the list.

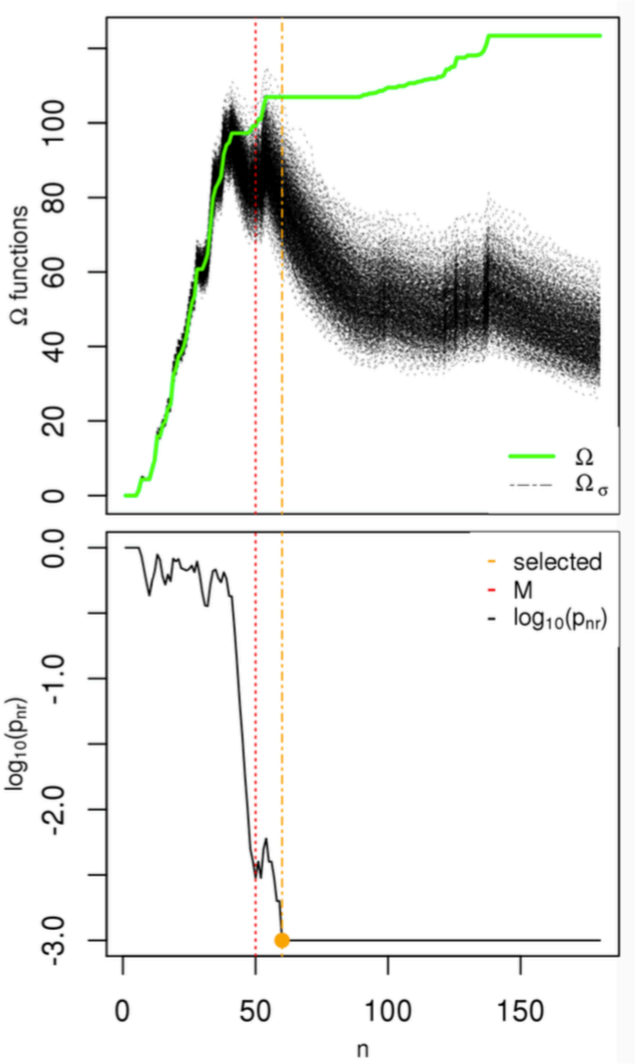


FIGURE 3.3: *Resampling procedure carried out with* 1000 *permutations [4].*

Chapter 4

Application to Cancer Datasets

In this chapter we describe the application of the network propagation technique introduced above to three different cancer datasets. At the beginning, we summarize some useful biological concepts in order to clarify the aims of the method. Then, we discuss the pre-processing of the data and the model parameters setting. Finally, we discuss the results testing the identified submodules with *Gene Pathway Enrichment Analysis.*

4.1 Biological Concepts

4.1.1 Molecular Interactions Networks

As we said above, in the cell, proteins are organized in complex structures to perform biological functions. For this reason, the mapping of interactions between molecular entities in the cell is a major step in order to understand how such structures work.

Depending on the the molecular entities involved and the cell process under examination it is possible to build several types of such mappings. Some examples are *signaling networks*, *metabolic networks*, *Gene regulatory networks* and *PPI networks*.

The diffusion methods described above can, in principle, be applied to all of these networks. In our applications we used protein-protein interaction network because it was the most suitable for our purposes. In what follow we review the main features of PPI networks.

Protein-Protein Interaction Networks (PPI Network) model individual proteins as vertices, and their interaction relationships as undirected edges. Sometimes, nodes represent, rather than the proteins themselves, the related genes, *i.e.* coding genes are considered to have a one-to-one relationship with proteins.

In general what is meant by *protein-protein interaction* is a *physical contact* between two proteins occurring in the cell. The contact must be specific, that is, with *molecular docking*, and not accidental. Moreover, excluded from the interaction network are those interactions that a protein undergoes in generic functions such as its production, folding and degradation [11].

The *interactome* is the complete map of protein-protein interactions in a cell. In humans' cell there are almost $20,000$ coding genes, *i.e.* the nodes of the interactome, and the estimated number of interaction between proteins, i.e. links, is $\sim 650,000$. Actual PPI networks has between $6,000$ to $13,000$ nodes and $25,000$ to $150,000$ links.

From a graph theory point of view PPI networks are undirected graph, i.e. their adjacency matrix is symmetric, and are connected, which means that there are no independent components.

The degree distribution of a network, $P(k)$, gives the probability that a selected node has k links. PPI-networks, as the most part of biological networks, are *scale-free*, i.e. their degree distribution approximates a *power law:*

$$P(k) \sim k^{-\gamma} \tag{4.1}$$

where k is the node degree. The name scale-free is due to the scale invariance of the distribution. Scaling the argument, k, causes a proportional scaling of the distribution P:

$$P(ck) \sim ck^{-\gamma} = c^{-\gamma}k^{-\gamma} = c^{-\gamma}P(k) \tag{4.2}$$

Due to this fact there is not a typical node in the network that can be used to characterize the rest of the network. This is in strong contrast to random networks, for which $P(k)$ is a Poissonian distribution, and the degree of all nodes is in the vicinity of the average degree, which could be considered typical.

In networks with a power-law degree distribution there are a few nodes, called hubs, with a large number of links while, most of the nodes are poorly connected. The value of the exponent, γ, determines several properties of the network. The more important the role of the hubs is in the network, the smaller is the value of γ [2].

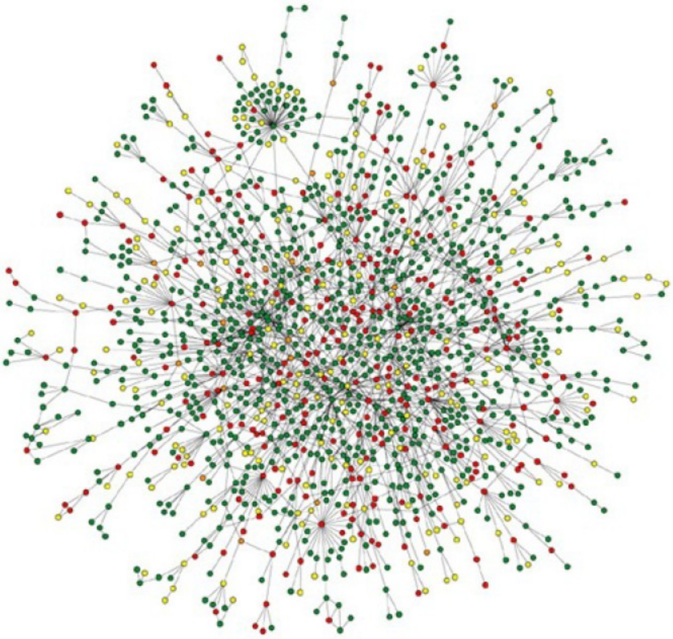


FIGURE 4.1: *Protein–protein interactions in yeast. Note the presence of a few highly connected nodes, hubs, that hold the network together, and many poorly connected nodes. [17]*

PPI networks have several drawbacks. Firstly, they represent protein interaction as static links, while physical contacts between proteins in general are not permanent and varying upon time. Secondly, they do not take into account the spatial distribution of the proteins in the cell.

Moreover, from an experimental point of view, the main drawback is the presence of an high number of both false positives and negatives [11]. As we mentioned above, due to this fact, the measure of proximity between two nodes must be chosen carefully in order to reduce the experimental noise.

4.1.2 Cancer Somatic Mutation and Gene Expression Profiles

Tumor Somatic Mutations Profile

Cancer origin and evolution are strictly related to Darwinian evolution occurring in the origins of species. Tumors can be considered to be the outcome of a process of evolution occurring among cell populations of human tissues. The two constituent processes of evolution are: the continuous acquisition of

heritable genetic variation in individual cells by *random mutation* and **natural selection** acting on the phenotypic diversity within the environments provided by the tissues of an organism [34].

Cells that have acquired deleterious mutations are removed by natural selection while, cells altered in a fashion that increases their capability to proliferate more effectively than their neighbors,*i.e* have a selective advantage, become more numerous.

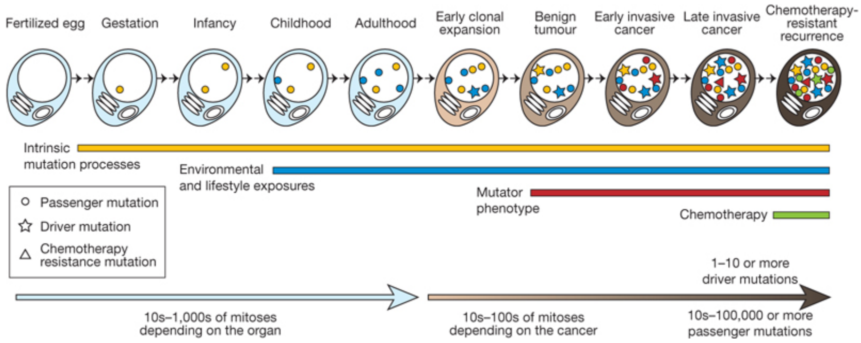

Fertilized egg
Gestation
Infancy
Childhood
Adulthood
Early clonal expansion
Benign tumour
Early invasive cancer
Late invasive cancer
Chemotherapy-resistant recurrence
Intrinsic mutation processes
Environmental and lifestyle exposures
Mutator phenotype
Chemotherapy
○ Passenger mutation
☆ Driver mutation
△ Chemotherapy resistance mutation
1–10 or more driver mutations
10s–1,000s of mitoses depending on the organ
10s–100s of mitoses depending on the cancer
10s–100,000 or more passenger mutations

FIGURE 4.2: *The lineage of mitotic cell divisions from the fertilized egg to a single cell within a cancer showing the timing of the somatic mutations acquired by the cancer cell and the processes that contribute to them. [34]*

Within an adult human cell genome there are thousands of mutation, figure 4.2. These comprise many genetic alterations, such as single-nucleotide substitutions, insertions, deletions, rearrangements, copy-number alterations, epigenetic alterations and so on. The overwhelming majority of these alterations are believed to have phenotypes with limited abnormal growth potential and are invisible or manifest themselves as benign growths. These alterations are called *passenger alterations* [23].

Only on a few occasions a cell acquires a set of sufficiently advantageous mutations that allows it to proliferate autonomously and invade tissues. These are the alterations that cause the cancer and are called driver alterations or simply tumor drivers, figure 4.3.

Because driver events are critical to cancer progression, their discovery is the primary goal of genome-wide cancer sequencing. In these studies, the genome of a sample of cancerous cells is compared to the genome of a sample of non-cancerous cells, sometimes called *germ-line cells*, of the same patient. The mutation between the two samples are called *Tumor Somatic Mutation*.

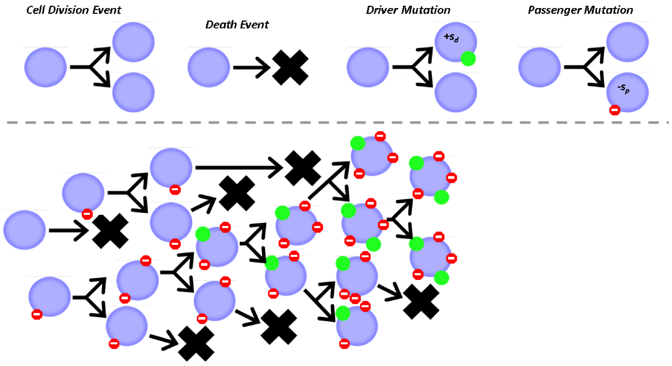


FIGURE 4.3: *Dynamics of cancer progression. Cancer progression is driven by the accumulation of a small number of genetic alterations. These few driver alterations reside in a cancer genome beside thousands of passengers mutations. [23]*

Sequencing of various cancer tissues at a genome level have found that individual cancers contain tens of thousands of somatic alterations, fig. 1.1. There are few genes altered frequently and a larger amount of genes altered much less frequently [37] [34]. Moreover, genes involved in a disease change wildly between samples. Most of this somatic mutation are passenger while a few of them are tumour drivers. As we have seen, a major challenge in system biology is the identification of tumour drivers from passengers.

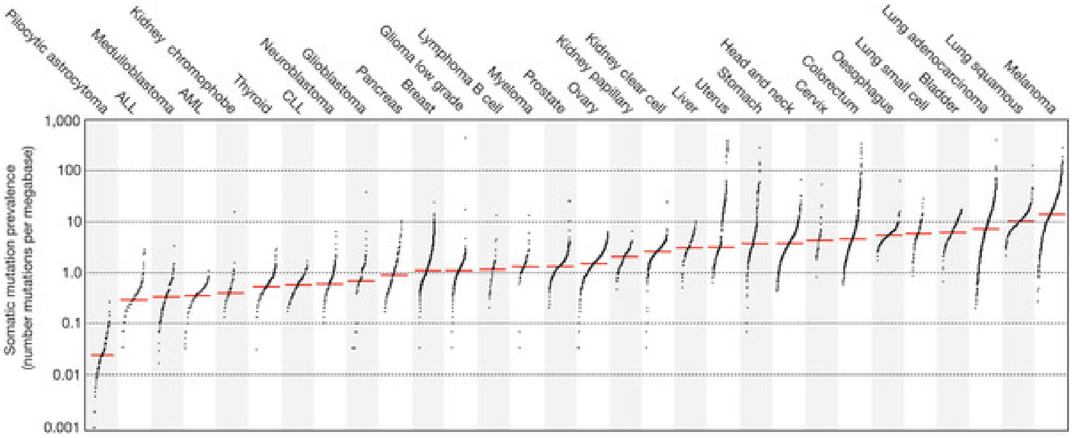


FIGURE 4.4: *Distribution of the number of mutations in various kinds of cancer [1]*

Tumor Gene Expression Profiles

Gene expression (GE) in the process by which the information contained in a gene is used to synthesize a gene product. While all cells in an organism contain the entire genome, only a small fraction of genes is expressed at a given time. The level of expression varies upon time and depends on the cell type. Since the thousands of genes expressed in a particular cell determine what that cell can do GE is the most fundamental level at which the genotype gives rise to the phenotype and hence plays a central role in cancer development.

Gene expression profiling is the measurement of the expression of thousands of genes at once. In cancer, the comparison between gene expression profiles of cancerous and germ-line cells allows us to identify those genes whose expression is altered, that is, those genes the are likely to give a selective growth advantage to tumor cells.

4.2 Materials and Methods

4.2.1 Cancer Datasets

We have applied the pipeline described in section 3.2 to the identification of molecular interaction subnetworks enriched in altered genes to three types of cancer: Colon adenocarcinoma (COAD), Gastrointestinal Stromal tumor (GIST) and Acute Myeloid Leukaemia (AML). For the last two types of tumors we analysed whole exom somatic mutation and gene expression profiles while for COAD tumour only SM profiles were available. The Colon Adenocarcinoma dataset was downloaded from The Cancer Genome Atlas website *http:// tcga-data.nci.nih.gov/docs/publications/tcga* while GIST and Acute Myeloid Leukaemia datasets have been made available from the Department of Experimental, Diagnostic and Specialty Medicine (DIMES) of the University of Bologna.

The three datasets report different kind of information. COAD and GIST datasets, for both SM and GE profiles, consist of a *sample-per-gene matrix*, *i.e.* each column of the matrix contains alteration data of a set of genes of a sample.

Somatic Mutations for each patient are represented as a profile of binary $(1, 0)$ states on genes, in which a $'1'$ indicates a gene for which mutation has occurred in the tumour genome relative to the germ-line genome in the person with the tumor.

Gene Expression is represented as a profile of real-valued states on genes, in which the real numbers indicate the CPM (Count per Million) or TPM (Transcripts per Million) for each gene.

For what concerns leukaemia datasets, they provide only the mean quantities over a set of samples. Somatic mutations are represented as a vector, $\boldsymbol{f}_{SM}$, associating to each tested gene a frequency of mutation over a set of 78 samples. In a similar fashion, gene expression profile provide the mean fold-change for each gene and an associated adjusted p-value which gives the significance of the measure.

4.2.2 Network Propagation Source Vectors

In order to perform network propagation we built, for each dataset, a "source vector" of length equal to the number of nodes in the interaction network. Such vector assigns to each gene (node) a value proportional to the relative importance of the gene. In the language of the source/sink model the vector components are the inflow rates.

Since the biological aims were slightly different for the three datasets we have pre-processed each of them in a different manner.

As regards COAD and Leukaemia somatic mutation data we were interested in finding those sub-modules of PPI network enriched in genes altered with respect to the germ-line. For this purpose we used as sources the vector of the frequencies of mutation, $\boldsymbol{f}_{SM}$. In the COAD case, starting from the *sample-per-gene matrix*, we have calculated the relative frequency of gene mutation within the set of samples summing up the columns and dividing by the number of samples.

In leukaemia gene expression data, in order to give a greater initial score to genes with a lower p-value, *i.e.* with a higher significance, we built the source vector, $\boldsymbol{g}_{GE}$, according to the following statistics:

$$g_i = -\left|FC_i\right| \log(P_i) \tag{4.3}$$

where g_i is the final score of gene i, FC_i is the fold-change and P_i i the p-value. Such definition ensures that genes with a high fold change in expression and a low p-value score very highly and vice versa. We built two different source vectors for up-regulated ($\boldsymbol{g}_{GE}^{Up}$) and down-regulated ($\boldsymbol{g}_{GE}^{down}$)

genes since the propagation method does not allows us to distinguish between them.

GIST dataset contained information related to four different cancer molecular subtypes. Table 4.1 summarizes the dataset features.

Molecular Subtypes	GE samples	SM samples
KIT	7	9
PDGFRA	5	6
Quadruplo -	11	12
SDH	2	2

TABLE 4.1: *GIST samples per molecular subtype*

We were interested in finding those modules differentially enriched for each couple of subtypes. The algorithm allows us to tackle this task in two different manners: calculating a single-class statistics or propagating the information for each subtypes and calculating the difference between the propagated quantities.

In order to have comparable results for the three dataset we chose to built a differential single-class statistics as follows. As regards GE data, we calculated the fold-change FC for each couple of tumor subtypes:

$$FC_{ij} = log_{10}\left[\frac{GE(i)}{GE(j)}\right] \tag{4.4}$$

where i, j are the tumor subtypes, GE is the gene expression level expressed in TPM.

For what concerns SM, we built a source vector for each couple of subtypes assigning a value 1 to those genes for which a mutation has occurred in a subtype relative to the other and zero to the others.

4.2.3 Network Datasets

We considered interaction data provided by *BioPlex* dataset of Protein-Protein interactions.

The BioPlex network is a map of human protein interactions identified using high-throughput affinity-purification mass spectrometry. The version we used in our analysis was downloaded from *http://bioplex.hms.harvard.edu* on 10/02/2017 and includes 10961 proteins and 56553 interactions among them.

Bioplex network has a giant component comprising 10,901 nodes and 56,519 interaction. The remaining 60 proteins are separated into 27 smaller components with two to four nodes. Far what concerns our applications we will refer only to the giant component.

The giant component of PPI network has a significantly smaller number of nodes and edges in comparison to a random network. The degree distribution follows a power law behaviour typical of scale-free networks, with an exponent of 1.947 [20].

4.2.4 Network Propagation and Parameters Setting

We applied the network propagation algorithm, equation 3.5, iteratively until convergence, i.e. until is met the condition:

$$|x_{t+1} - x_t| < t_r \tag{4.5}$$

where t_r is the convergence threshold. We set it to 10^6 in line with previous studies [5] [35].

The parameter α, as we have seen in the previous chapter, controls the magnitude of the diffusion that is, it controls how much information tends to be spread through the network versus how much is retained in the source vertices. The role of network topology is emphasized for $\alpha < 0.5$.

The optimal value of α is network dependent. Previous studies showed that α in the range $[0.5, 0.7]$ determined consistent results. We decided to set it to 0.7 in order to maximize the importance of the topology. Moreover, $\alpha = 0.7$ is a a good trade-off between computational cost, which increases as α approaches 1, and the magnitude of the diffusion.

The parameter ϵ in equation 3.6, weights the relative importance of final state of a node, $\boldsymbol{x}^*$, with respect to its initial state $\boldsymbol{x_0}$. Small values emphasize the local hypothesis, while large values emphasize only the stationary distribution state.

Bersanelli at Al. [5] analysed the performance of the network smoothing index on simulated biological dataset. They find that when the biological signal is particularly high, i.e. modules are composed of highly altered genes, the best performances are obtained for high values of ϵ, while when the modules are composed of a mixture of genes with strong and low variation, the NSI perform better for smaller values of ϵ.

Since we did not know in advance how the modules of our dataset were composed we chose ϵ in order to prioritize both sources and entities in network proximity to them. In figures 4.5 and 4.6 we report the scatter plots of the values of the initial quantities (X_0) versus S_p calculated for different values of the parameter ϵ, in leukemia dataset (The scatter plots for the GIST dataset are reported in Appendix)

We chose ϵ such that the gene i with the highest network-free score (f^i_{SM}) has approximately the same S_p of the gene, j, that has an initial network-free score, $f^j_{SM} = 0$, and assumes the highest S_p after the propagation. As shown in fig. 4.7 the dimension of the sub modules do not depend on ϵ.

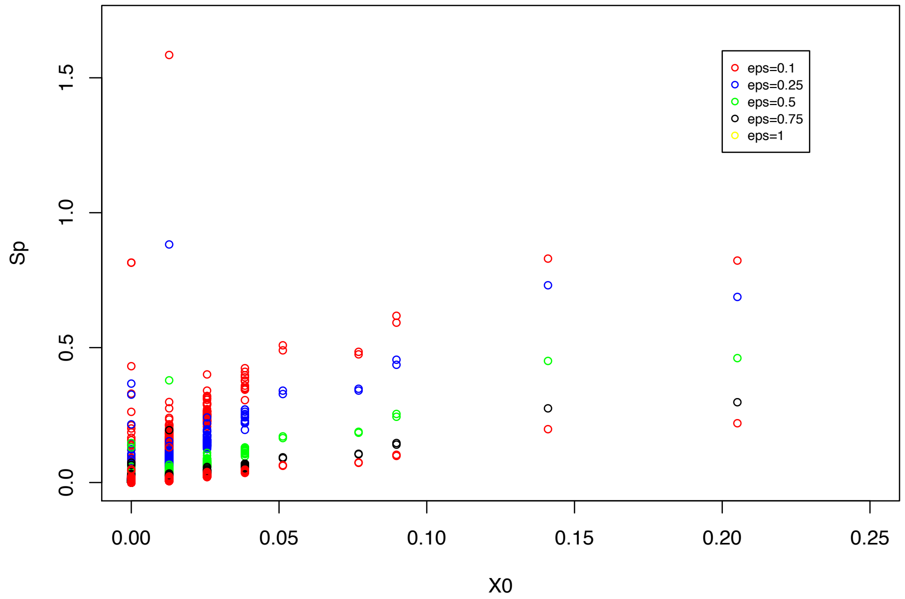
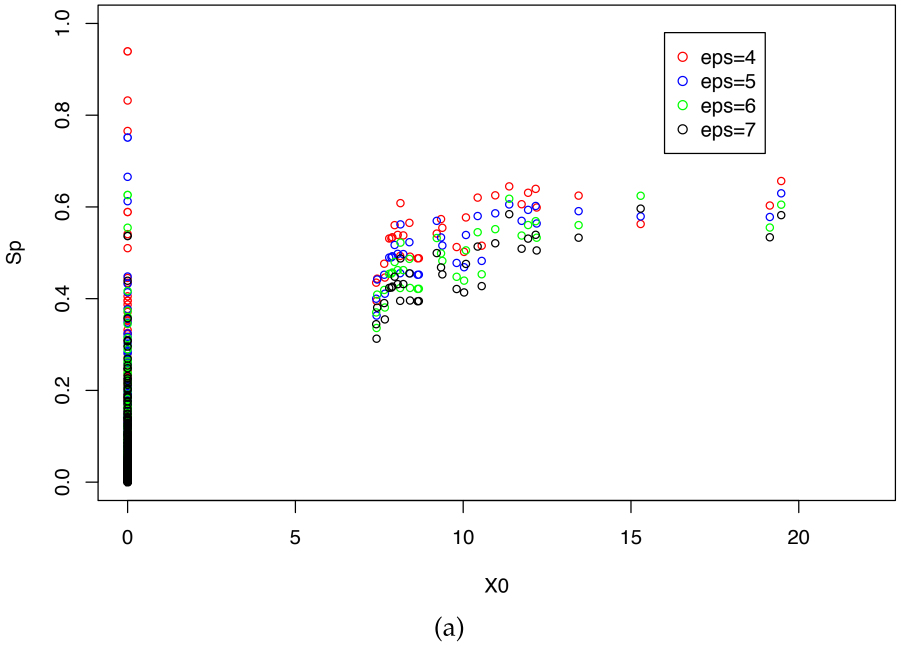
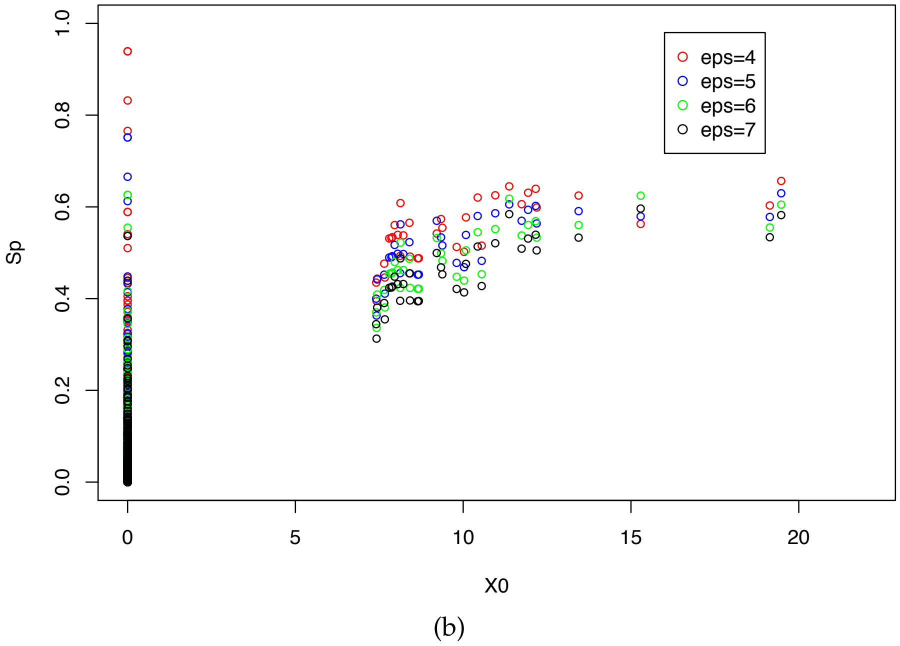


FIGURE 4.5: *Scatter plot with $\boldsymbol{f}_{SM} = X0$ values versus S_p calculated on Leukemia Somatic Mutation data for different values of the parameter ϵ.*

At this point, we have to cut the list of genes ranked byS_p, in order to identify the most significant submodules. We performed the network resampling procedure described at the end of the previous chapter with a number of permutation $k_{NR} = 500$ and $k_{NR} = 1000$. The graphs in fig.4.8 show the network resampling p-value (p_{NR}) calculated for each rank (n). Vertical lines



(a)



(b)

FIGURE 4.6: *Scatter plots with* $\boldsymbol{g} = X0$ *values versus* S_p *calculated on leukemia GE data for different values of the parameter* ϵ*. Up-regulated genes, fig.*(a)*, and Down-regulated genes fig.*(b)*, were analysed separately.*

indicate the top ranking genes selected to be part of the corresponding gene modules.

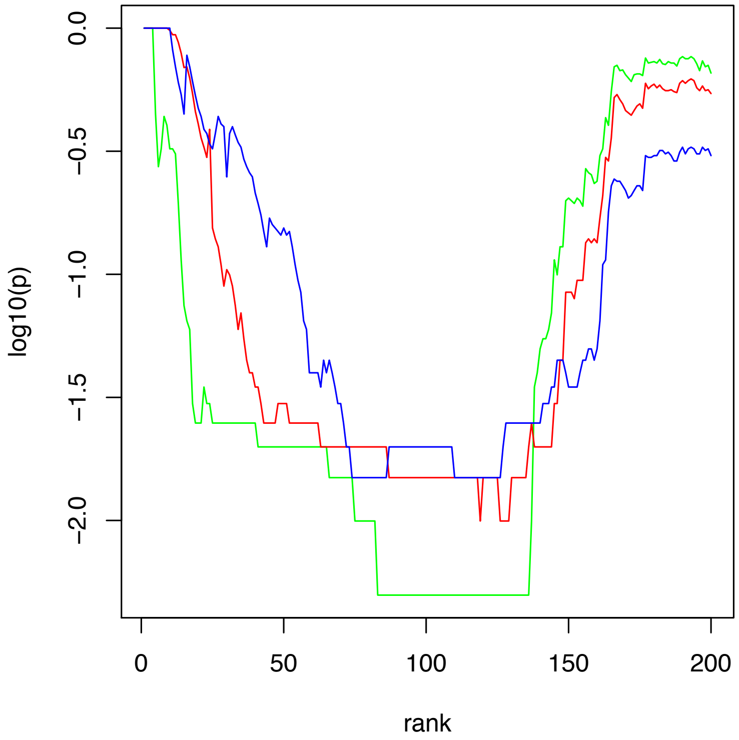


FIGURE 4.7: *Network resampling p-value (p_{NR}) versus the number n of top list entities for $k_{NR} = 500$, varying ϵ. As we can see the dimension of the submudules depends weakly on the value of ϵ*

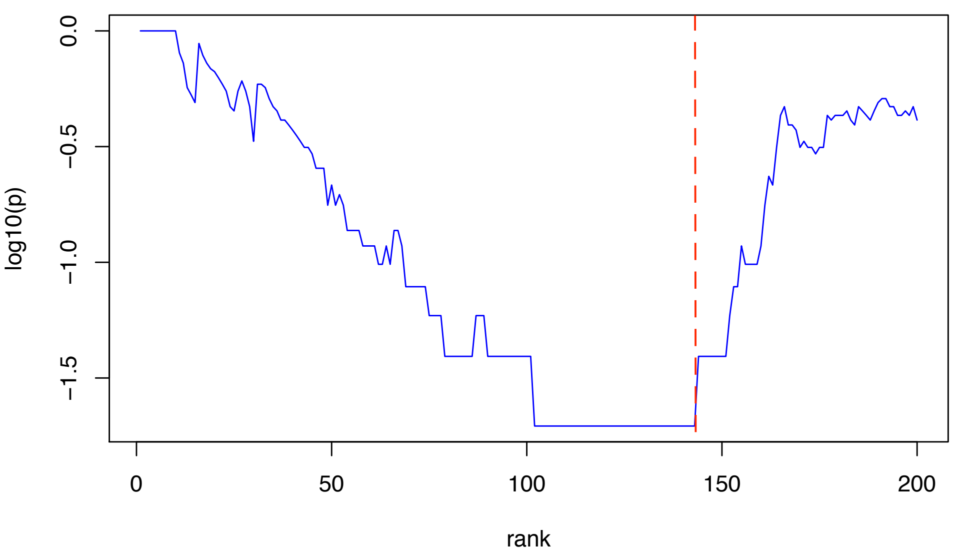
4.3 Results

In this section we discuss in detail the results obtained applying the propagation algorithm to Acute Myeloid Leukemia dataset. We summarize the properties of the detected disease modules and we assess their association to leukemia by data mining of literature and by gene set enrichment analysis.

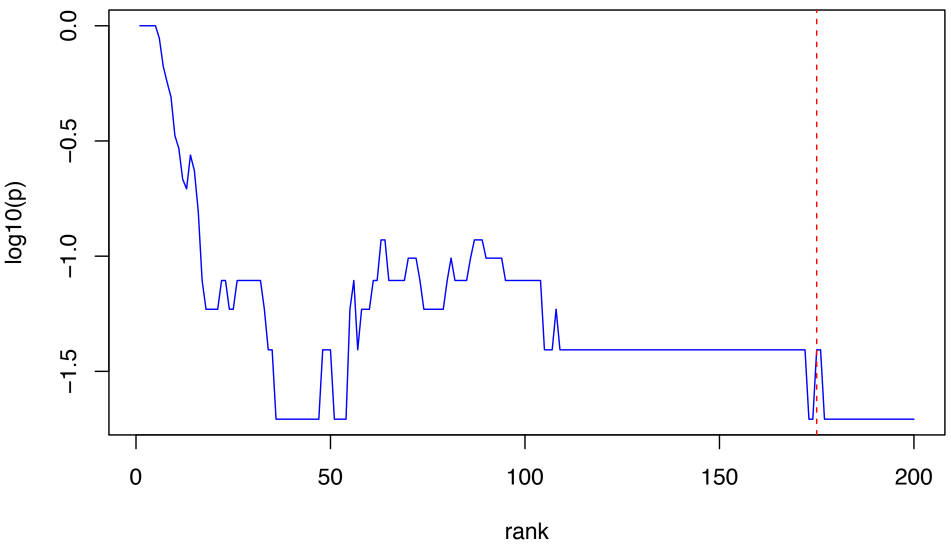
4.3.1 Acute Myeloid Leukemia Results

We applied the network propagation starting from three different source vectors: $\boldsymbol{f}_{SM}$, $\boldsymbol{g}_{GE}^{down}$ and $\boldsymbol{g}_{GE}^{up}$. As we said above, varying the parameter ϵ we tuned the overlap between network-free and network-based gene rankings.

In figures 4.11 and 4.11 are reported the disease modules identified in SM and GE datasets carrying out the network resampling procedure to the list of genes ranked in decreasing order with a number of permutation $k_{NR} = 500$.



(a)



(b)

FIGURE 4.8: *Network resampling p-value* (p_{NR}) *versus the number* n *of top list entities for Down-regulated (a) and up-regulated (b) genes in leukemia. For Down-regulated genes we chose* $\epsilon = 5$ *while for up-regulated* $\epsilon = 0.8$. $k_{NR} = 500$. *The dashed vertical lines indicate the selected number of nodes.*

As expected genes ranked according to S_p form bigger and more connected modules than genes ranked according to network-free quantities, see figure A.4.

In the three dataset we identified a total of 59 disease modules that have

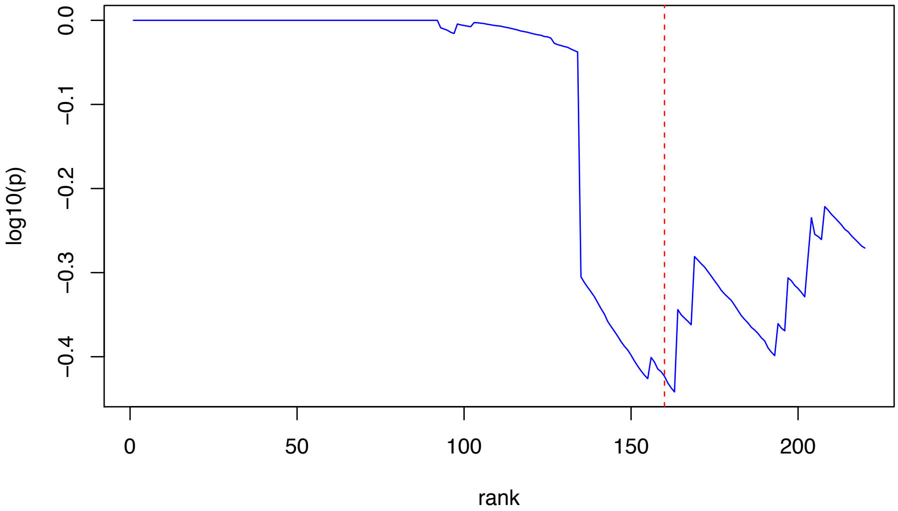


FIGURE 4.9: *Network resampling p-value* (p_{NR}) *versus the number* n *of top list entities in leukemia SM.* $k_{NR} = 500$. *We chose* $\epsilon = 0.25$. *The dashed vertical line indicates the selected number of nodes.*

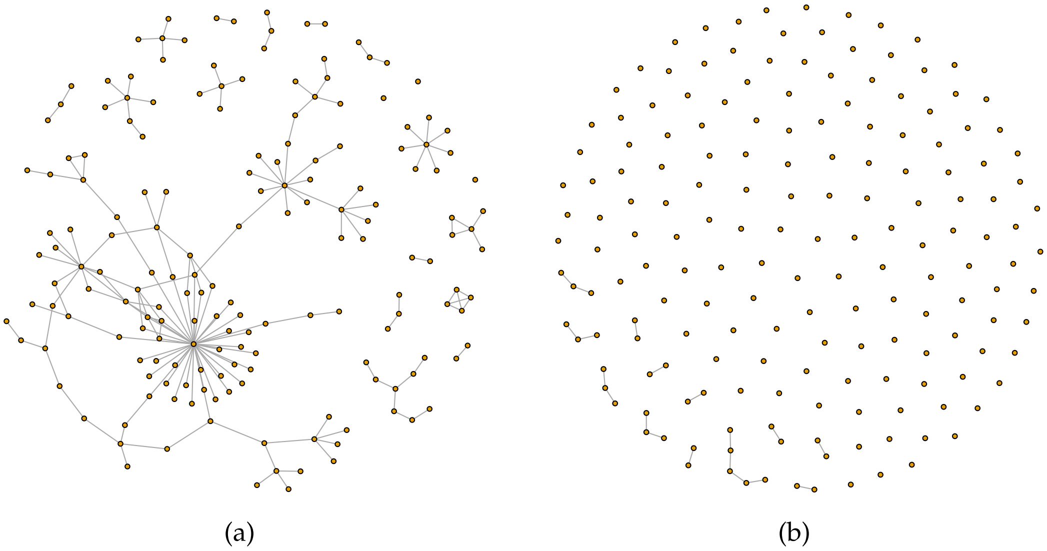


FIGURE 4.10: *GE up-regulated genes data. Comparison between the first 175 genes ranked by* S_p*, fig. (a), and by the network-free statistics* $\boldsymbol{g}_{GE}^{up}$*. As aspected, top ranking genes ordered by the network based quantity are more connected and form bigger networks than genes ordered by the network-free quantity*

size between 2 and 112 nodes. Only 16 modules were composed of more than 4 nodes. In table 4.2 we report the number of modules per size interval in GE *Up/Down regulated* and SM data. For each dataset we find one *main disease module* with more than 10 connected nodes (112 for *GE Up-regulated*,

AML	*GE Up*	*GE Down*	*SM*
number of modules	16	24	19
$2 \leq n \leq 3$	8	20	14
$4 \leq n \leq 5$	4	2	0
$5 \leq n \leq 10$	3	1	4
$n > 10$	1	1	1

TABLE 4.2: *Overview of the module size in the SM, GE up-regulated and GE down-regulated dataset, n is the number of nodes in the module.*

11 for *GE down-regulated* and 15 for *SM*). A total of 219 different genes occur in the 16 modules with more than 4 nodes. Only a few genes are shared among identified modules from different datasets.

Both SM and GE modules contain an high number of genes that are highly cited in cancer and leukemia literature, some of which were specifically prioritised using network-based quantities, table 4.3. For example, FBXW7 and RORC are two genes related to leukemia that had a very low initial score and are high ranked due to their proximity to sources.

Moreover, a number of new genes, that is, genes that are not related to leukemia, have been identified. This genes are likely to have a role in the pathologies because of their proximity to the sources and could be interesting candidates for further studies.

In order to identify the cellular pathway regulated by identified genes we carried out set enrichment analysis. The results are summarized in the next section.

4.3.2 Pathway Enrichment Analysis

In order to check our results pathway analysis was carried out using the *Gene Set Enrichment Analysis* approach on the identified submodules.

Gene Set Enrichment is a statistical test based on the *Hypergeometric distribution*. Let us recall the basic definitions. If N is the total number of genes in the network and K is the number of genes in a given pathway, that we call S, then, if we have a module of n genes, the Hypergeometric distribution gives the probability of finding k genes of our module in the pathway S by pure chance:

gene	*citations*	*citations in leukemia*
TP53	13437	1385
FBXW7	641	117
CUL9	8	1
RORC	679	16
EPHA7	175	6
MYCBP2	47	3
EPHA10	20	1
AGAP1	22	1
FAM111A	14	-
PCDHA3	3	-
HNF4A	1025	1
PCDHA12	1	-
CTNNA3	80	1
PCDHA10	1	-

TABLE 4.3: *Overview of SM main module genes literature citations. We report the total number of citations and the number of citations in leukemia literature.*

$$P(X = k) = \frac{\binom{N}{k}\binom{N-K}{n-k}}{\binom{N}{k}} \tag{4.6}$$

A statistical test based on this distribution is performed to asses the statistical significance of the overlap between genes from a known pathway and a submodule identified by our algorithm. Since multiple comparisons are done the *p-value* must be adjusted for false discovery rate.

A central role in pathway enrichment is played by the definition of the metabolic pathways. Pathway collections structure, content and functionality usually vary in different sources and Gene Set Enrichment results are strongly affected by the choice of the source database.

In order to overcome this problem we implemented a *Python* script able to perform Pathway Enrichment Analysis on the three most famous databases at the same time: Gene Ontology (GO), KEGG and REACTOME.

Our findings show that a significant number of the detected enriched modules are related to a biological process involved in cancer genesis and evolution. We focused our analysis on the 16 biggest modules. We report the outcome of the analysis or the main three modules in figures 4.13, 4.14, 4.15. In table 4.4 we report the main pathways related to the others modules.

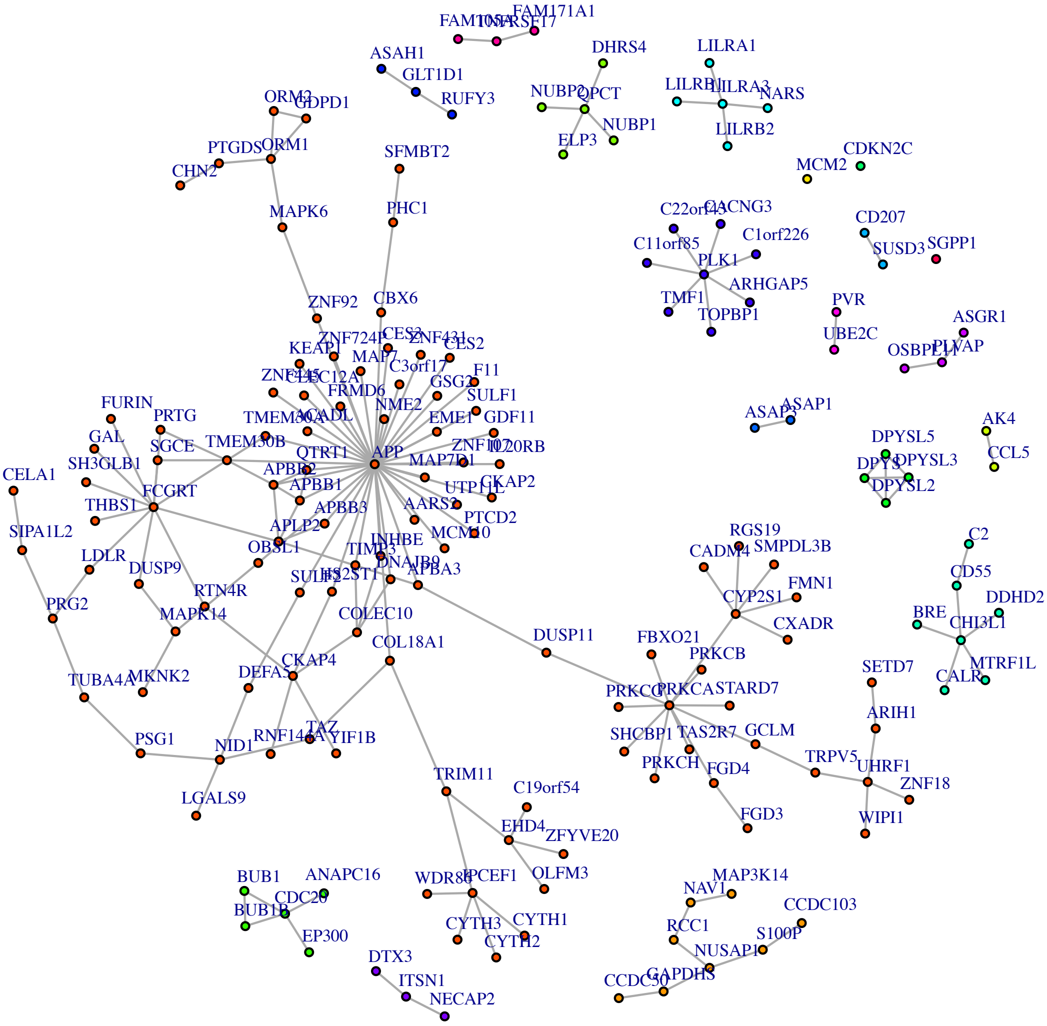


FIGURE 4.11: *Enriched Modules identified by the algorithm using as source vector the list of Up-Regulated genes in Acute Myeloid Leukemia GE. The network resampling procedure suggests to consider the first 175 genes of the S_p ranking list. As we can see there are 16 different modules (8 of them contain more than 3 nodes). The main module (red nodes) contains 112 genes. A huge number of them were specifically prioritized by the network propagation algorithm.*

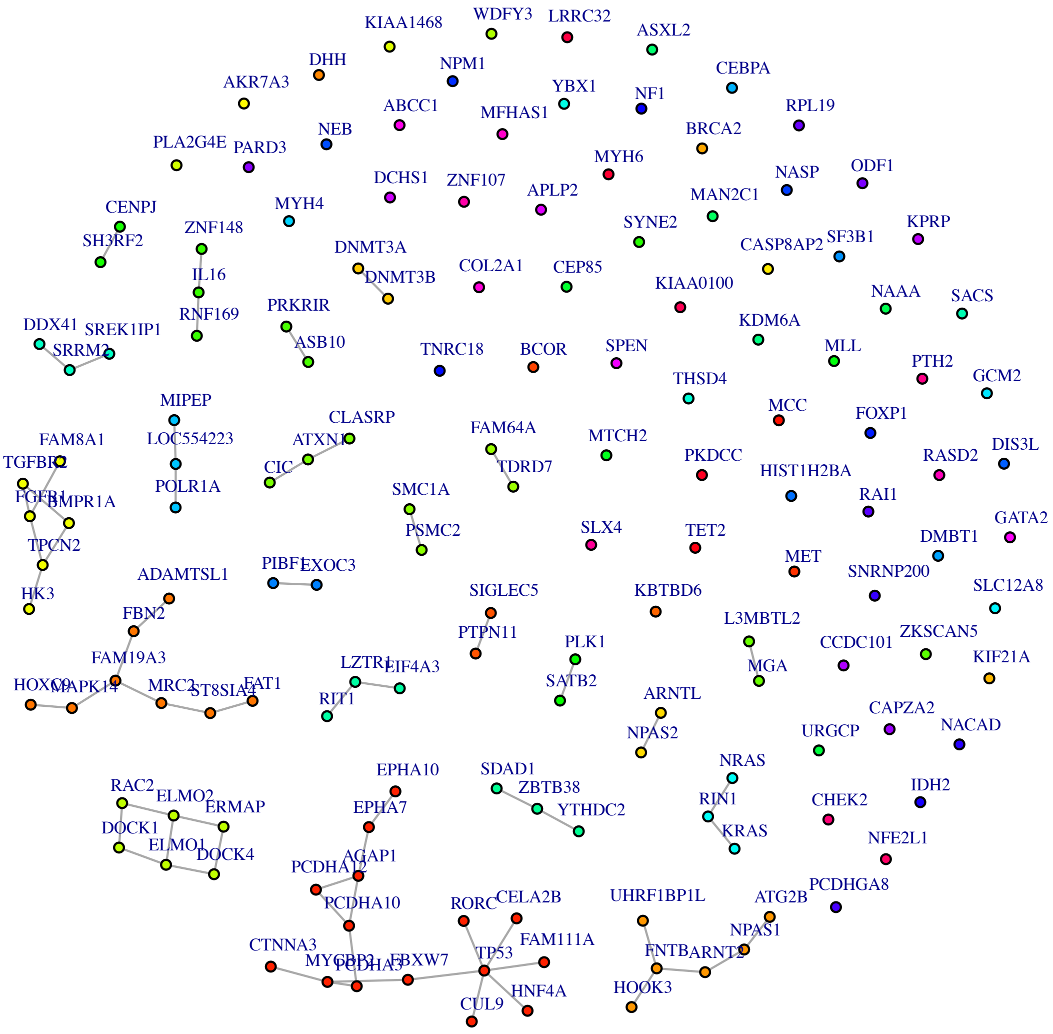


FIGURE 4.12: *Enriched Modules identified by the algorithm using as source vector the list of mutated genes in Acute Myeloid Leukemia. The network resampling procedure suggests to consider the first 150 genes of the S_p ranking list. As we can see there are 19 different modules (5 of them contain more than 3 nodes). The main module (red nodes) contains 15 genes.*

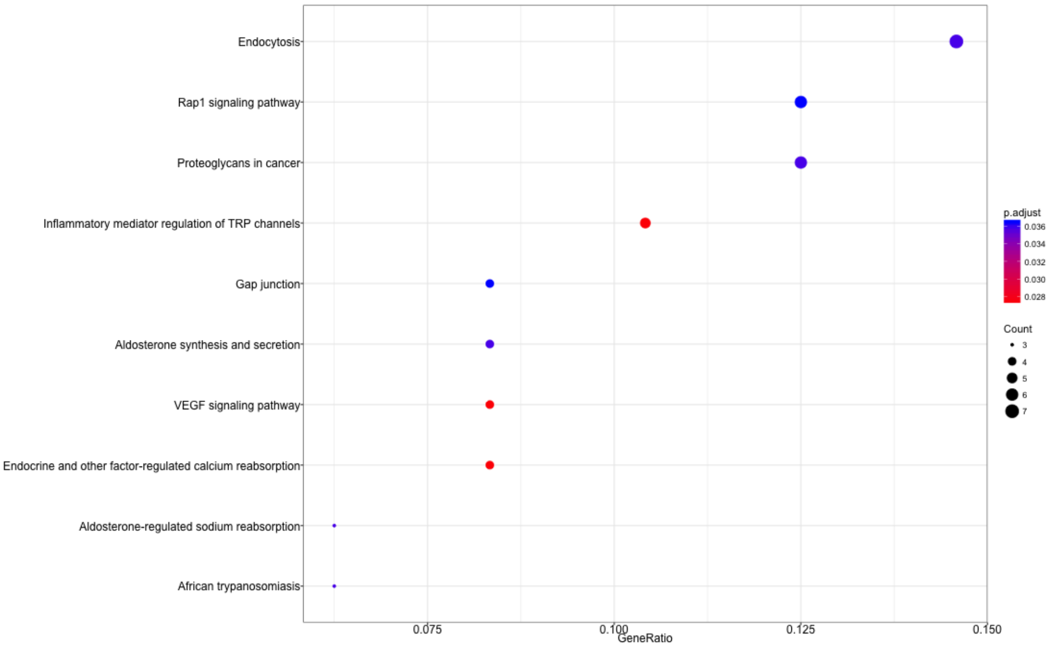


FIGURE 4.13: *Gene Set Enrichment Analysis carried out on the main enriched module in AML GE up-regulated genes dataset*

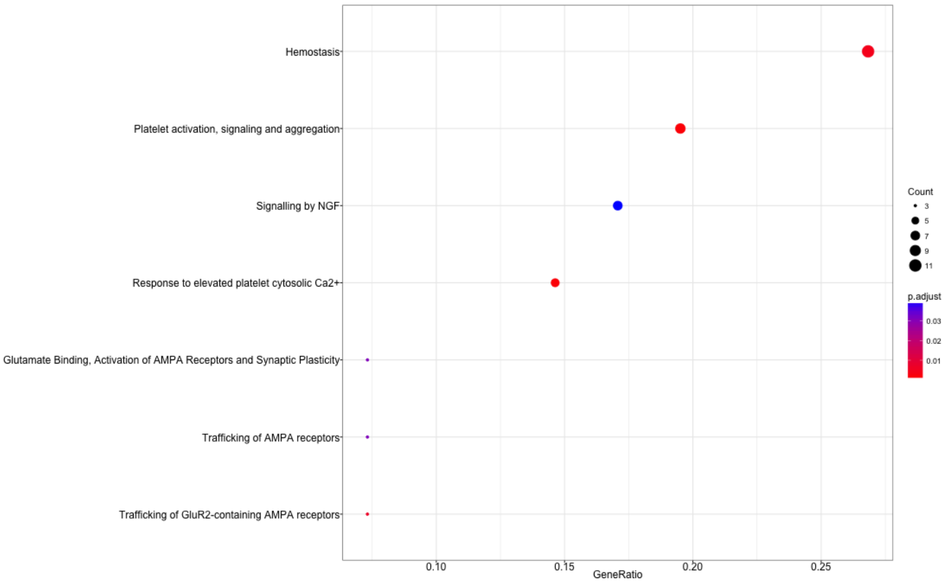


FIGURE 4.14: *Gene Set Enrichment Analysis carried out on the main enriched module in AML GE down-regulated genes dataset*

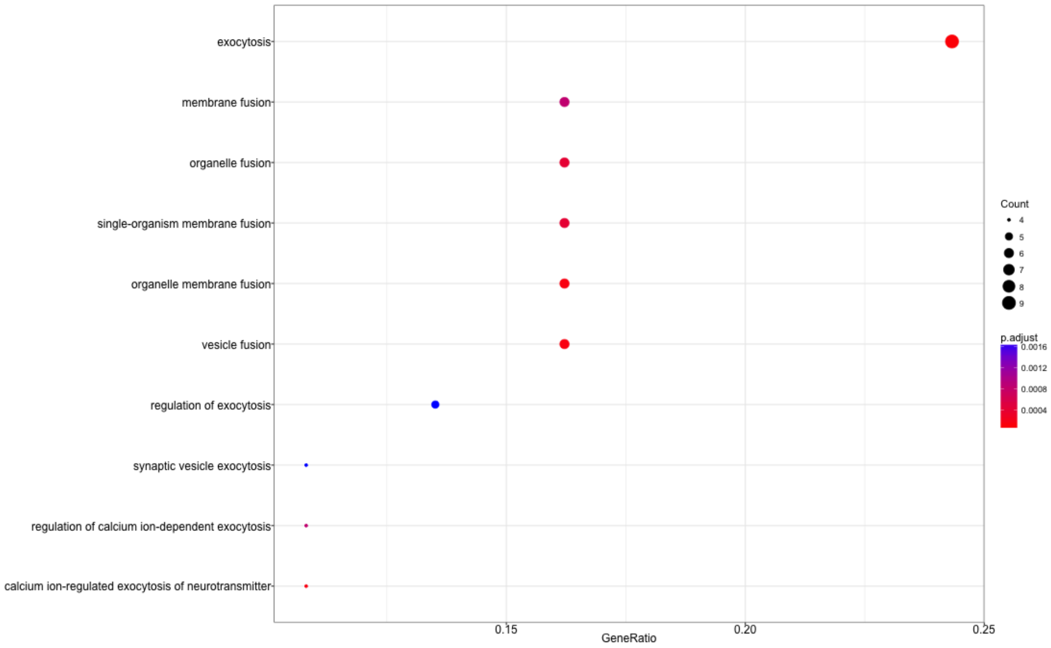


FIGURE 4.15: *Gene Set Enrichment Analysis carried out on the main enriched module in AML SM genes dataset*

AML	*GE up*	*GE down*	*SM*
main module	*Endocytosis* *RAP1 Signalling* *Proteoglycans in Cancer* *TRP channels Inflammatory Regulation*	*Exocytosis* *Membrane Fusion* *Organelle Fusion*	*Hemostasis* *Platelet Activation* -
other modules	*Microtubule Cytoskeleton Organization* *Regulation of Cell Cycle Phase Transition* *Regulation of Ubiquitin Protein ligase*	*mRNA in Cancer* *transcriptional modification in cancer* -	*Cell Cycle* *Developmental Biology* -

TABLE 4.4: *Biological pathways associated to the detected enriched modules for SM and GE AML datasets.*

Conclusions

The definition of a network region involved in a pathology is a challenging issue due to the complexity of biological networks. Several approaches have been proposed to tackle this challenge. The most effective ones exploit the behavior of a random walker, *i.e.*, the diffusion of a substance within the protein-protein interactions network.

We have reviewed the theory underlying such processes that turn out to be related to the network laplacian matrix. This relation allowed us to outline many interesting links to other physical systems defined on a network. In particular, we have found that Fick's law on network is strictly related to the behavior of a system of coupled oscillator, while its connection to random walk is not straightforward as in the continuous case. Also, we have found an interesting connection between the transient states of diffusion and spectral clustering. Further investigation of this topic can lead to a better insight into diffusion techniques especially for what concerns their application to network community detection.

Then, we have exploited a novel diffusion technique, recently proposed in literature, for detection of protein-protein interactions network enriched in altered genes. Starting from a set of query nodes, such as the set of altered genes in a cancer, the aim of the method is to find a set of other genes related to the query set and forming with it a connected subnetwork.

We have applied such technique, for the first time, to three datasets containing Somatic Mutation and Gene expression profiles of different types of cancer: Acute Myeloid Leukemia (AML), Gastrointestinal Stromal Tumor (GIST) and Colon Adenocarcinoma (COAD).

We have discussed in detail the results relative to AML. Our findings show that the algorithm is able to prioritize genes that are strictly related to the pathology and provides several other genes that are likely to have a role in the pathologies, that is, genes that lie in network proximity to genes already associated to AML or in regions of the PPI network enriched in altered genes.

Moreover, the method allowed us to detect several connected components for each dataset that are likely to be associated with the pathobiological processes underlying the disease.

A deep investigation of biology underlying the identified disease modules was beyond the scope of this thesis. Nevertheless, we assessed, by gene set enrichment analysis, their relation to known biological pathways contained in three different database (Gene Ontology, KEGG and Reactome) finding correlations to a significant number of cancer related pathways.

In future work it could be interesting to investigate the structure of the single detected modules, for example by means of centrality and clustering measures, and the relations among them as parts of the interactome to asses causal relationships between alterations and pathobiological processes.

Appendix A

A.1 AML GE-Down Results

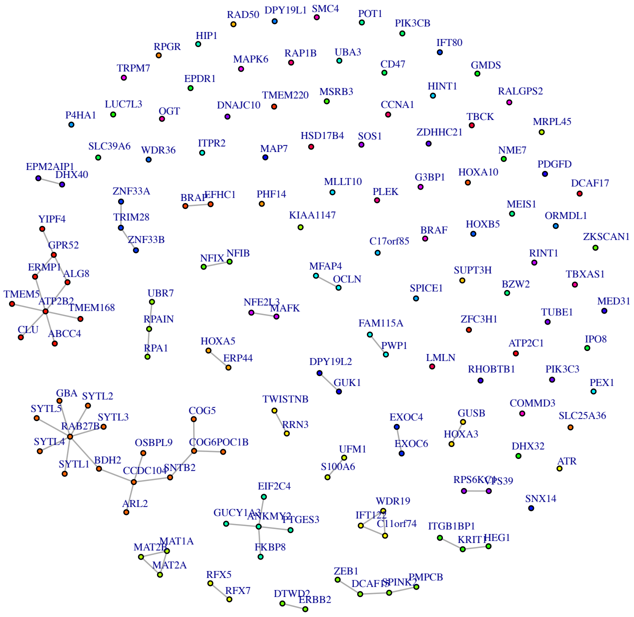


FIGURE A.1: *Enriched Modules identified by the algorithm starting from the list of down-Regulated genes in AML GE. The network resampling procedure suggests to consider the first 145 genes of the S_p ranking list. As we can see there are 24 different modules (4 of them contain more than 3 nodes). The main module (red nodes) contains 11 genes.*

A.2 GIST Results

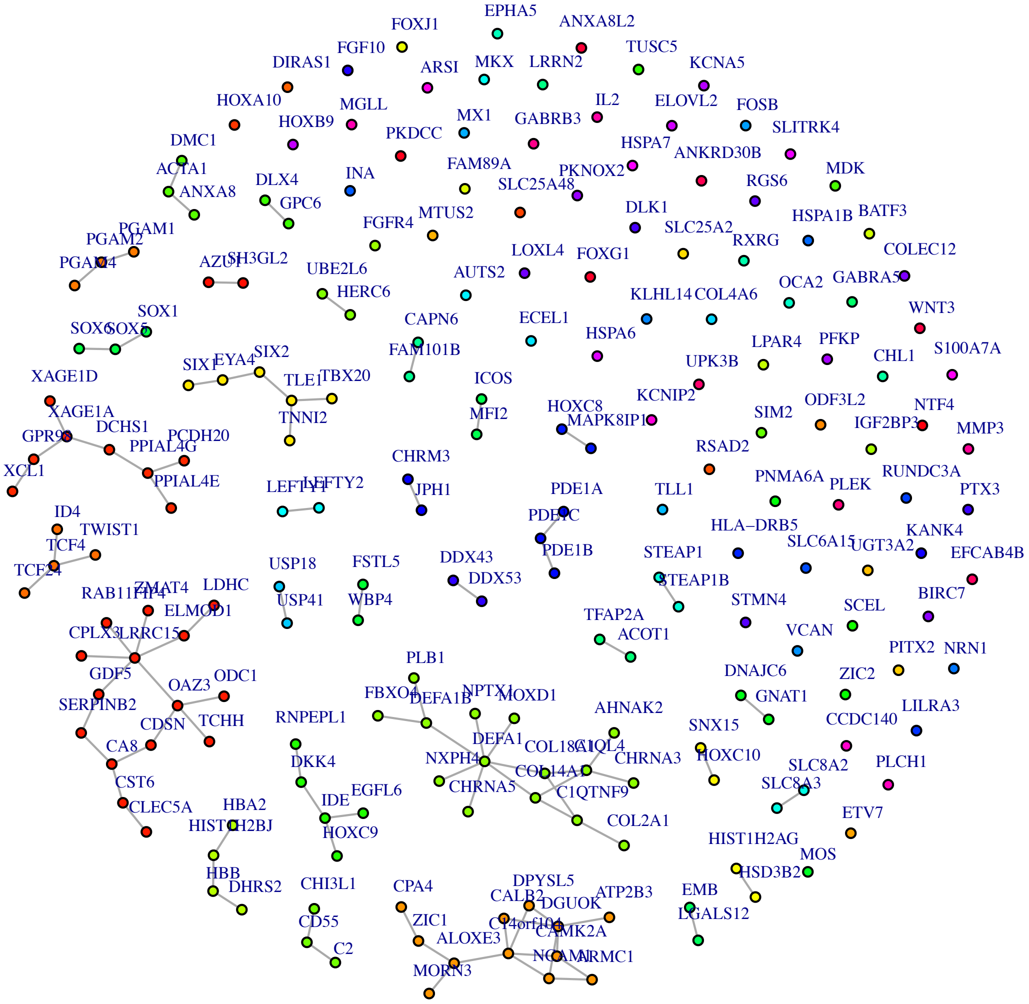


FIGURE A.2: *Enriched Modules identified by the algorithm starting from the list of up and down-Regulated genes in the comparison between KIT and Q- GIST GE . The network resampling procedure suggests to consider the first 200 genes of the S_p ranking list.*

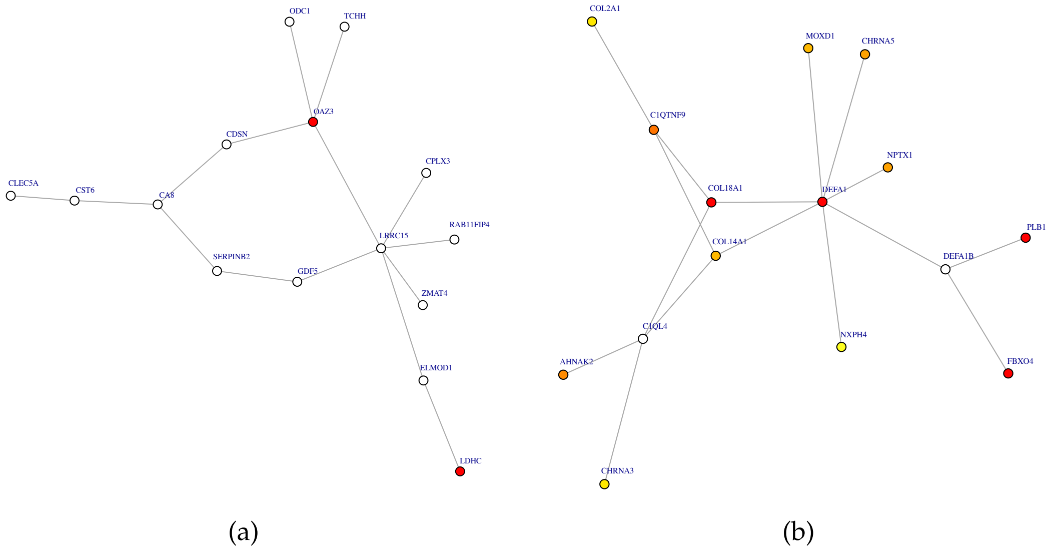


FIGURE A.3: *Main differentially enriched modules identified in GIST subtypes KIT and Q-. Red and yellow nodes are those genes specifically prioritized by the algorithm.*

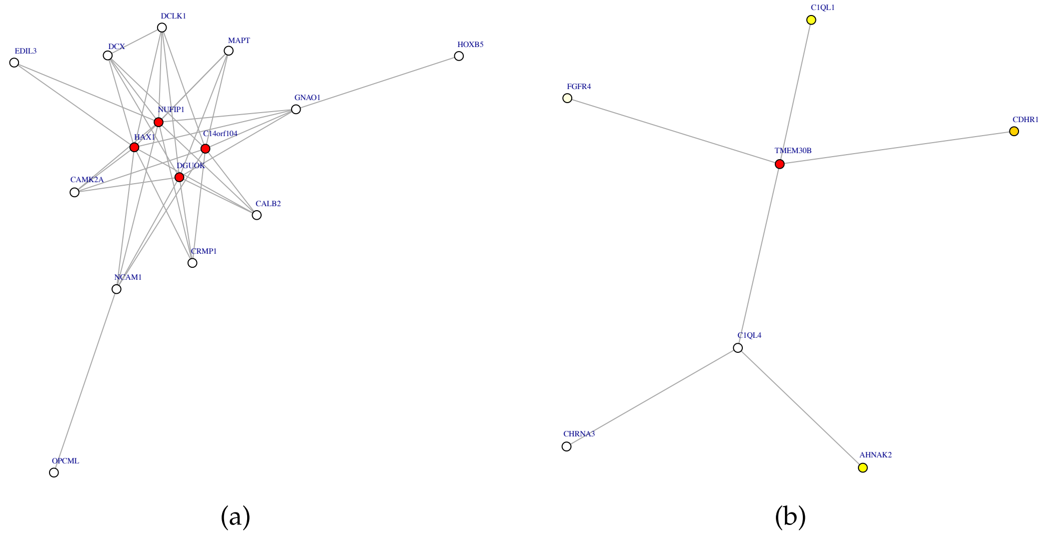


FIGURE A.4: *Main differentially enriched modules identified in GIST subtypes PDGFRA and Q-. Red and yellow nodes are those genes specifically prioritized by the algorithm.*

List of Figures

List of Abbreviations

AML	**A**cute **M**yeloid **L**eukemia
COAD	**Co**lon **Ad**enocarcinoma
GIST	**G**astro**I**ntestinal **S**romal **T**umor
SM	**S**omatic **M**utation
GE	**G**ene **E**xpression
FC	**F**old **C**hange
RW	**R**andom **W**alk
CTRW	**C**ontinuous **T**ime **R**andom **W**alk
RWR	**R**andom **W**alk **R**estart
PPI	**P**rotein **P**rotein **I**nteraction
GO	**G**ene **O**ntology
KEGG	**K**oto **E**ncyclopedia of **G**enes and **G**enomes

Bibliography

[1] L.B. Alexandrov et al. “Signatures of mutational processes in human cancer”. In: *Nature* 500 (2013), pp. 415–421.

[2] A.L. Barabasi and Z.N. Oltvai. “Network biology: understanding the cell’s functional organization”. In: *Nature Reviews Genetics* 5 (2004), pp. 100 –113.

[3] A.L. Barabasi et al. “Network medicine: a network- based approach to human disease”. In: *Nat. Rev. Genet.* 12 (2011), pp. 56–68.

[4] M. Bersanelli. “Mathematical Physics Techniques for Omics Data Integration”. In: *Doctoral Thesis* (2017).

[5] M. Bersanelli et al. “Network diffusion-based analysis of high throughput data for the detection of differentially enriched modules”. In: *Scientific Reports* 6 (2016).

[6] T. Biyikoglu, J. Leydold, and P.F. Stadler. *Laplacian Eigenvectors of Graphs*. Berlin: Springer-Verlag, 2007.

[7] T. Can, O. Camoglu, and A.K. Singh. “Analysis of protein-protein interaction networks using random walks”. In: *BIOKDD ’05: Proceedings of the 5th international workshop on Bioinformatics (New York, USA: Association for Computing Machinery)* 61-68 (2005).

[8] F. Chung. *Spectral Graph Theory*. Cambridge: CBMS Regional Conference Series in Mathematics, 1997.

[9] Y. Cun and H. Frohlich. “Netclass: an r-package for network based, integrative biomarker signature discovery”. In: *Bioinformatics* 30 (2014), pp. 1325–1326.

[10] Y. Cun and H. Frohlich. “Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics”. In: *PLoS ONE* 8(9).e73074 (2013).

[11] J. De Las Rivas and C. Fontanillo. “Protein-Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks”. In: *PLoS Computational Biology* 6(6): e1000807 (2010).

[12] P.G Doyle and Snell. J.L. “Random walks and electric networks”. In: (2006).

[13] M. Fiedler. “Algebraic connectivity of graphs”. In: *Czechoslovak Math. J.* 23 (2013), pp. 298 –305.

[14] M.E. Fisher. “On hearing the shape of a drum”. In: *Journal of Combinarics Theory* 23 (1966), pp. 105 –125.

[15] S. Fortunato and D. Hricb. “Community detection in networks: A user guide”. In: *Physics Reports* 659 (2016), pp. 1 –44.

[16] M. Hofree et al. “Network-based stratification of tumor mutations”. In: *Nature Methods* 10.11 (2013), pp. 1108–1115.

[17] H. Jeong et al. “Lethality and centrality in protein networks”. In: *Nature* 411 (2001), pp. 41 –42.

[18] S Kohler et al. “Walking the Interactome for Prioritization of Candidate Disease Genes”. In: *The American Journal of Human Genetics* 82.4 (2008), pp. 949–958.

[19] R. I. Kondor and J. Lafferty. “Diffusion kernels on graphs and other discrete input spaces”. In: *ICML* 2 (2002), pp. 315–322.

[20] Yang L. et al. “Characterization of BioPlex network by topological properties”. In: *Journal of Theoretical Biology* 409 (2016), pp. 148 –154.

[21] M.D.M Leiserson et al. “Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes”. In: *Nature Genetics* 47 (2015), pp. 106–114.

[22] L. Lovatz. “Random Walks on Graph”. In: *Combinatorics, Paul Erdos is Eighty* 2 (1993).

[23] C.D. McFarland et al. “Impact of Deleterious Passenger Mutations on Cancer Progression”. In: *PNAS* 110.8 (2013), pp. 2910 –2915.

[24] I. Mirzaev and J. Gunawardena. “Laplacian Dynamics on General Graphs”. In: *Journal of Mathematical Biology* (2013).

[25] N. Musada, M.A. Porter, and R. Lambiotte. “Random walks and diffusion on networks”. In: *arXiv:1612.03281* (2016).

[26] M. Newman. *Networks: an Intoduction*. London: Oxford University Press, 2010.

[27] L. Page et al. “The PageRank citation ranking: Bringing order to the web,” in: *Technical Report Stanford Infolab* 1999-66 (1999).

[28] J. Pandey, M. Koyutrk, and Grama A. “Functional characterization and topological modularity of molecular interaction networks”. In: *BMC Bioinformatics* 11(Suppl 1):S35 (2010).

[29] Y. Qi et al. “Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions”. In: *Genome Research* 18 (2008), pp. 1991–2004.

[30] Y. Qian et al. “Identifying disease associated genes by network propagation”. In: *BMC systems biology* 8.Suppl 1 (2014).

[31] B. Scholkopf and A.J. Smola. “Learning with Kernels: A tutoria Introduction”. In: (2001).

[32] B. Scholkopf, K. Tsuda, and Vert J.P. *Kernel Methods in Computational Biology*. Cambridge, Massachusetts: The MIT Press, 2004.

[33] M.E. Stokes et al. “The application of network label propagation to rank biomarkers in genome-wide alzheimer?s data”. In: *BMC Genomics* 15 (2014), p. 282.

[34] M. R. Stratton, P.J. Campbell, and P.A. Futreal. “The Cancer Genome”. In: *Nature* 458 (2009), pp. 719–724.

[35] F. Vandin, E. Upfal, and B.J. Raphael. “Algorithms for Detecting Significantly Mutated Pathways in Cancer”. In: *Journal of Computational Biology* 18(3) (2011), pp. 507–522.

[36] O. Vanunu et al. “Associating Genes and Protein Complexes with Disease via Network Propagation”. In: *PLoS Computation Biology* 6(1) (2010).

[37] B. Vogelstein et al. “Cancer Genome Landscapes”. In: *Science* 339.6127 (2013), pp. 1546–1558. ISSN: 0036-8075.

[38] J. Weston et al. “Protein ranking: From local to global structure in the protein similarity network”. In: *Proc. Nat. Acad. Sci.* 6569-6563.101(17) (2004).