ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA

Scuola di Scienze
Corso di Laurea Magistrale in Fisica

Multiplex network analysis with application to biological high-throughput data

Relatore:
Prof. Daniel Remondini

Correlatore:
Dott. Giulia Menichetti

Presentata da:
Alice Zandegiacomo Cella

Sessione III
Anno Accademico 2014/2015

Sommario

L'analisi dei network a multiplex riscontra un sempre maggiore interesse da parte della comunità scientifica. Questi network complessi sono caratterizzati da una sovrapposizione di network semplici, chiamati layer, e vengono comunemente definiti come network di network. La struttura a multiplex permette l'analisi di dati reali provenienti dai più disparati ambiti e l'identificazione di strutture complesse non altrimenti individuabili. In questa tesi vengono studiate alcune caratteristiche dei network a multiplex; in particolare l'analisi verte sulla quantificazione delle differenze fra i layer del multiplex. Le dissimilarità sono valutate sia osservando le connessioni di singoli nodi in layer diversi, sia stimando le diverse partizioni dei layer. Sono quindi introdotte alcune importanti misure per la caratterizzazione dei multiplex, che vengono poi usate per la costruzione di metodi di *community detection* . La quantificazione delle differenze tra le partizioni di due layer viene stimata utilizzando una misura di mutua informazione. Viene inoltre approfondito l'uso del test dell'ipergeometrica per la determinazione di nodi sovra-rappresentati in un layer, mostrando l'efficacia del test in funzione della similarità dei layer. Questi metodi per la caratterizzazione delle proprietà dei network a multiplex vengono applicati a dati biologici reali. I dati utilizzati sono stati raccolti dallo studio DILGOM con l'obiettivo di determinare le implicazioni genetiche, trascrittomiche e metaboliche dell'obesità e della sindrome metabolica. Questi dati sono utilizzati dal progetto Mimomics per la determinazione di relazioni fra diverse omiche. Nella tesi sono analizzati i dati metabolici utilizzando un approccio a multiplex network per verificare la presenza di differenze fra le relazioni di composti sanguigni di persone obese e normopeso. La caratterizzazione delle differenze viene effettuata usando i metodi analitici proposti. I risultati dimostrano l'efficacia del metodo proposto, rilevando alcuni composti aventi comportamenti differenti fra il layer legato a persone normopeso e quello rappresentante persone obese.

La tesi è organizzata nel seguente modo:

- nel capitolo 1 viene introdotta l'analisi dei network; la descrizione si concentra sulle caratteristiche dei network a multiplex e delle sottostrutture (comunità) dei network.
- Nel capitolo 2 si descrive la metodologia di estrazione dei dati analizzati e gli aspetti biologici riguardanti i dataset, effettuando una breve

analisi dei dati per metterne in luce alcune caratteristiche.

- Nel capitolo 3 viene illustrato nel dettaglio il metodo implementato per l'analisi dei network a multiplex. Sono poi introdotti alcuni metodi statistici utilizzati nell'analisi dei dati reali.
- Nel capitolo 4 sono riportati i risultati ottenuti applicando il metodo proposto al set di dati metabolici reali, dimostrando sia l'efficacia del metodo, sia l'importanza di un approccio a multiplex network.

Abstract

Multiplex network analysis arouses great interest among scientific community. These complex networks are characterized by the overlapping of simple networks (or layers) and are normally defined as *networks of networks* . Multiplex structure allows the analysis of real data from different fields and the identification of complex structures, which cannot be located otherwise. The present thesis deals with some features of multiplex networks, especially the quantification of differences among multiplex layers. The differences are evaluated by observing the connections of single nodes in different layers and estimating the layer partitions. Therefore, some important measures for multiplex characterization are introduced and they are used to create community detection methods. The quantification of the differences between the partitions of two layers is estimated using a mutual information measure. Moreover, the use of hypergeometric test for the determination of over-represented nodes in a layer is deeply analysed, showing the test efficiency with regard to the layer similarity. These methods for the characterization of multiplex network properties are applied to real biological data. Data have been collected by DILGOM study in order to determine the genetic, transcriptomic and metabolic implications of obesity and metabolic syndrome. These data are used by the *Mimomics* project to determine the relations among different *omics*. The present thesis analyses the metabolic data using a multiplex network approach to verify the presence of differences among the relations of blood mixtures that belong to overweight and normal-weight people. The characterization of differences is carried out using the analytical methods described above. The results show the efficacy of the suggested method, pointing out that some mixtures have different behaviours between the layer connected to normal-weight people and the one connected to overweight people.

The structure of the thesis is described below:

- chapter 1 introduces the network analysis, in particular the description focuses on multiplex network features and on network substructures (communities).
- Chapter 2 discusses the method for the extraction of the analyzed data. Also the biological aspects regarding datasets are described, carrying out a brief data analysis in order to highlight some of their features.

- Chapter 3 includes a detailed description of the method used for the multiplex network analysis. In addition, some of the statistical methods used in the real data analysis are introduced.
- Chapter 4 describes the results that have been achieved applying the suggested method to sets of metabolic real data; showing the efficiency of the method and the importance of a multiplex network approach as well.

Contents

Introduction

Graph theory is a young mathematical field, which is attracting the attention of an increasingly wide section of the scientific community. The interest on this new analytical method is due to the fact that it permits to study complex systems, namely systems composed by a large number of interacting elements. Networks have two basic components: nodes and edges; the former are the objects under analysis, the latter are their interactions. According to this, networks can always be built if we have a dataset with interacting elements; this is why network analysis is applied to different kinds of data, such as informatics, economic, biological, genetic, social data. All these different fields show similar features when they are studied using network theory; this is the point that makes graph theory so interesting to the scientific community.

In this thesis, we will present a method for the multiplex networks analysis. It will be applied to biological data, therefore it can be included in the new field called *bioinformatics*, which applies mathematical and physical methods to genomic, metabolomic, transcriptomic, proteomic, biochemistry data. Bioinformatics is giving many interesting results, which could not be carried out only by using the classical biological methods. Indeed, complex systems can be fully characterised only using a mathematical approach.

In particular, we will analyse multi-omic data using multiplex network structures, which belong to the complex network theory. Multiplex structure allows the analysis of real data from different fields and the identification of complex structures, which cannot be located otherwise. The present thesis deals with some features of multiplex networks, especially the quantification of differences among multiplex layers. The differences are evaluated by observing the connections of single nodes in different layers and estimating the layer partitions. We will apply the developed multiplex approach on a biological dataset.

The dataset is provided by DILGOM study (*the Dietary, Lifestyle and Genetic determinants of Obesity and Metabolic syndrome*) and it is composed by metabonomic, transcriptomic, and genomic information of a Finnish cohort. Indeed, our research is linked to the investigation of metabolic peculiarities of obese individuals. The aim of our research is to evaluate if there are factors associated with *Body Mass Index* (BMI) and to the *waist-hip ratio* (who) in different omics by using network analysis. To do that, we build a multiplex network with two layers, where one layer is linked to obese individuals and the other one refers to normal weight people. Each layer displays the correlations of blood serum compounds extracted by the DILGOM study. Differences between these two *metabolic* layers are therefore linked to the BMI. In order to investigate these differences we take into consideration both differences between single nodes and differences between clusters of nodes.

In the first chapter of this thesis we will introduce network theory, and we will examine some graph measures and community detection methods which we will utilize. In the second chapter we will report a brief explanation of the method which the DILGOM study utilised for the extractions of blood serum compounds. Moreover we will explain the biological aspects linked to this research, describing the characteristics of the two datasets we will utilize. In the third chapter we will show the method which we implemented for the analysis of multiplex networks. This method is completely independent to the data to be analysed, since it is a purely physical, mathematical and statistical approach. This makes it possible to extend this procedure to the most disparate fields. We will apply the implemented method to the biological dataset, which is described in chapter 2. The results are displayed in the fourth chapter and in that chapter, a brief explanation of the results is also included.

Chapter 1

Network theory

Conventional network theory is a multidisciplinary field of study which was developed during the second half of the XX century. Initially it had connected the mathematical graph theory to social science, but in short time it was embraced by many other field to investigate complex systems. Nowadays it has been applied in the most varied disciplines including physics, computer science, electrical engineering, biology, economics and climatology and it is re-labelled as *complex networks theory.*

In this chapter we will present a brief explanation of the structures studied by network theory. These structures are discrete sets of related elements which form complex systems. We will illustrate their commonly adopted representation called *adjacency matrix* and some algebraic properties related to this algebraic description. Thereafter some useful measures will be presented, these measures are utilized for the description of the network topology and the nodes features. The first section will inspect some measures related to the whole network topology, as *distance* and *diameter*; in the second section some measures which examine properties of single nodes, that are *node degree*, *strength* and *inverse participation ratio* will be listed; and the third section will concern the underling structures of graphs, called *communities* or *clusters*. We will examine in depth some commonly adopted methods for community detection, which have been applied to the DILGOM dataset.

In the second part of this chapter the *multilayer* and *multiplex* network approaches will be introduced. These are recent extensions of graph theory which permit to investigate more complicate frameworks than the classical networks analysis. In the end of the chapter a brief description of the most studied biological networks will be proposed and we will dwell on metabolic networks.

1.1 Networks

Networks belong to the field of graph theory, which is characterised by structures with a high level of complexity, the topology of which is not intuitive. Therefore network theory is the study of complex interacting systems which can be represented as graphs equipped with some *extra structures*, which allow to define direction of the interactions or the label of graph elements. The complexity of the framework can be estimated observing diverse characteristics of network elements that are nodes and edges.

Edges represent the interactions between nodes which are the constituent elements of graphs. Therefore, a graph ($G(V, L)$) can be defined as a non-empty finite set (V) of nodes tied by a set (L) of links. Every link, or edge, connects a pair of vertices. The *relation* between the two connected nodes can be bidirectional or unidirectional: in the former case the graph is called *undirected graph*, which means that edges are symmetrical $L(x, y) = L(y, x)$; in the latter case there is an edges orientation ($L(x, y) \neq L(y, x)$) and links are called *directed*.

Undirected graphs have symmetrical links, they describe relations that do not have a preferential direction such as communication networks, chemical bonds and so on. Digraphs are graphs with directed edges, they describe hierarchical relations; these kind of graphs are used for the description of transportations, infectious diseases, citations, transcriptional networks and so on.

Links can whether or not be weighted; the weight associated to the edge usually indicates the strength of the correlation, but it can also represents a length or a cost.

Networks are classified and characterised looking at nodes and links distributions. There are measures which analyse the whole graph properties, as diameter, clustering coefficient, eigenvalues, spectral properties. Other measures concern the properties of a single node of the graph, some of them are centrality, node degree, strength and inverse participation ratio. There are also methods which analyse the presence of 'sub-structures' of highly correlated nodes inside the network, which are clustering or communities detection methods.
Graphs are usually represented using square matrices, called *Adjacency matrices* (A).

Adjacency and Laplacian matrices

The adjacency matrix (A) is a square matrix $n \times n$, where n is the number of network nodes. Each row and each column of A represents the interactions between a specific node and all the others. In an undirected unweighted graph, adjacency matrix is symmetrical and

$$A_{i,j} = \begin{cases} 1 & \text{if there is a link between node i and node j} \\ 0 & \text{otherwise} \end{cases}$$

In weighted networks, $A_{i,j} = w_{i,j}$, if there is a link between node i and node j, where $w_{i,j}$ represents the link weight. Usually self-loops are not considered, which means that the diagonal of the adjacency matrix is null ($A_{i,i} = 0$).

Another used matrix is the *Degree matrix (D)*, which is a diagonal matrix defined as:

$$D_{i,j} = \begin{cases} deg(v_i) & \text{if i=j} \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

where $deg(v_i)$ is the number of edges which terminate at the node i.

The *Laplacian matrix (L)* is a squared matrix $n \times n$ defined as:

$$L = D - A$$

where D is the degree matrix and A is the adjacency matrix of the graph. So we can deduce that:

$$L_{i,j} = \begin{cases} deg(v_i) & \text{if i=j} \\ -1 & \text{if } i \neq j \text{ and there is a link between node i and node j} \\ 0 & \text{otherwise} \end{cases} \tag{1.2}$$

L matrix has several important proprieties that help to understand the graph structure. We can note that if the graph is undirected, L is symmetrical and positive- semidefinite, that is $\lambda_i \geq 0$ for all i. Moreover, since the sum of the elements in each row of L is equal to 0, zero is an eigenvalue of L with corresponding eigenvector $e_n^t = (1, 1, ..., 1)$.

A first important information of the graph structure given by adjacency and Laplacian matrices is the number of connected components. The number of times 0 appears as an eigenvalue in the Laplacian is the number of connected components of the graph. We speak of *connected graph* if it is not divided into two or more non communicant parts, that is if L has only one eigenvalue equal to zero. This can also be shown by adjacency matrix: an undirected graph is connected if there is no any permutation (PAP^{-1}) that forms a block matrix. In a connected graph, any node can reach any other node.

1.1.1 Distance and diameter

We define *walk* a sequence of nodes and links which connects two nodes; if all the links of a walk are different then it is called *trail* and if also all nodes are different we speak of *path*.

Paths are used to calculate the *node distance*, also called *geodesic distance*. This quantity is defined as the number of edges along the shortest path which connect the two considered nodes. If the graph is weighted, the distance becomes:

$$d_{ij} = \sum_{i \to j} \frac{1}{w_{kl}}\,, \tag{1.3}$$

where w_{kl} is the weight of a link belonging to the shortest path.

We can see from equation 1.3 that if two nodes belong to disjoint components of the graph, the distance between them is infinite.

The greatest geodesic distance between a node v and any other node is called *eccentricity ($\epsilon(v)$)*.

An useful quantity to evaluate the *efficiency of information* of the network is the *average path length* (l); it is defined as the average number of steps along the shortest path for all possible pairs of network nodes.

$$l = \frac{1}{n \cdot (n-1)} \cdot \sum_{i \neq j} d_{ij} \tag{1.4}$$

where, as before, n is the number of nodes in G.

An other important quantity is the graph *diameter (d)*; it is defined as the maximum eccentricity of any node in the graph:

$$d = \max_{v \in V} \epsilon(v)\,. \tag{1.5}$$

The diameter is representative of the linear size of a network.

1.1.2 Node degree, strength and inverse participation ratio

Node degree, strength and inverse participation ratio are three graph measures which analyse the properties of a single node of the graph. Therefore, they stress the differences between nodes belonging to the same network.

Node degree (k_i) specifies the number of edges that node i has with other nodes. If the network is directed there are two different degrees, the in-degree, which is the number of incoming edges, and the out-degree, which is the number of outgoing edges. Considering an undirected network the node degree is:

$$k_i = \sum_{i \neq j} \Theta(a_{ij})\,, \tag{1.6}$$

where $\Theta(x) = 1$ if $x > 0$, otherwise $\Theta(x) = 0$.

An useful information about the network can thus be obtained from the distribution of node degrees. The degree distribution ($P(k)$) of a network is defined as the fraction of nodes in the network with degree k: $P(k) = n_k/n$. Real world networks usually have nodes with very different degree: most nodes have a relatively small degree and only few nodes have many connections. These large-degree nodes are called *hubs*.

Since the node degree does not consider the weight of links, another measure can be useful to characterize nodes of weighted networks: this measure is called *strength*. The strength (s_i) is defined as the sum of the weights of the links of node i

$$s_i = \sum_{i \neq j} a_{ij} \, . \tag{1.7}$$

It can be seen that strength is equal to the node degree for unweighted networks but, for weighted ones, it loses the information relating to the number of ties, so it is not a substitute of the node degree.

The *inverse participation ratio* Y_i is defined as the sum of the squared ratio of every edge weight (a_{ij}) of node i and the strength of the node i

$$Y_i = \sum_{i \neq j} \left(\frac{a_{ij}}{s_i} \right)^2 \, . \tag{1.8}$$

Therefore, the inverse participation ratio indicates the homogeneity of the link weights relative to a node i. As can be seen, $1 < 1/Y_i \leq k_i$; the lower limit is verified when there is a high uneven weight distribution, while the upper limit $1/Y_i = k_i$ is verified when all the edges weights of node i are equal.
The graph measures mentioned above allow to perform a classification of graphs based on their topology. The main classes of graphs are [1]:

- **Random networks**: random networks are graphs in which properties such as the number of graph vertices, graph edges, and connections between them are determined in a random way. In particular, there are two main random graph models : the Erdős and Rényi definition of random graph fixes the total number of links L: given a graph with N labelled nodes, these are connected with L randomly placed links $G(N, L)$. On the other hand Gilbert defines random graph starting from the probability p that two nodes are connected; therefore, each pair of N labeled nodes is connected with probability p, $G(N, p)$.
  Random graphs are commonly used as *null models*, that is as a term of comparison between the graph under study and a graph with some of its structural features (as number of nodes and edges) , but which is built with random models.

- **Scale free networks** are connected graphs with a degree distribution that follows a power law: $P(k) \sim k^{-\gamma}$, where usually $2 \leq \gamma \leq 3$. They can be built using the *preferential attachment technique*, that is progressively adding nodes to an existing network and introducing links to the existing nodes. In this way the probability of be linked to a given node i is proportional to the number of existing links k_i that node has: $P(linking\,to\,node\,i) \sim (k_i)/(\sum_j k_j)$.

- **Small world networks** are connected networks where the mean geodesic distance between nodes increases sufficiently slowly as a function of the number of nodes in the network. Usually the growth follows a logarithmic function: $L \propto \log N$. Nodes have usually few neighbours, but they can reach any other node with a small number or steps.

1.2 Clusters or community structures

Until now, we have illustrated some general characteristics and measures of graphs which help to classify nodes and to understand the whole graph structure. Now we will introduce another point of view used to understand the graph framework, which inspects the underling structure of graphs. Many kinds of networks are characterised by the presence of groups of nodes highly connected to each others. These groups are called *community structures* or *clusters*, nodes belonging to these groups have many connections inside the group and sparser connections between them. The process of identifying this structure in terms of grouping graph nodes is called *graph clustering* or *community detection*[2][3].

The vast majority of real networks exhibits community structures, which are considered as fairly independent compartments of a graph. These groups of nodes are expected to behave in a similar way, that is to share common properties or to carry out similar functions. Communities can thus be found both using local or global criteria. In the former case they are expected to be detected inspecting them as separated entities and, in the latter, as a part of the whole graph. Independently of the adopted detection method, graph clustering is extensively used to analyse graphs of real complex systems; examples are social networks, collaboration networks, computer science, and protein-protein interaction networks (PPI). The latter are intensively investigated by biologists, since grouping proteins which deeply interact with each others can highlight their 'collective' functions. Furthermore, community structures of PPI show differences between healthy and sick people; in particular metastatic cells have proteins which interact very frequently with each other.

Given the importance of uncover these substructures, many methods are used to investigate community structures and to evaluate their goodness. We present a summary of the most used community detection methods. The aim of community detection methods is to group nodes so that nodes which belong to the same cluster $\mathcal{C}$ have many edges within each other and few edges which connect them to nodes of other clusters. Therefore, each cluster should be connected, which means that there should be at least one path internal to the cluster connecting each pair of vertices within a cluster. A cluster can be analysed as a subgraph $\mathcal{C}(S, E_c)$ of the graph $G(V; E)$, where S is the set of vertices which belong to the cluster ($S \subseteq V$) and $E_c \subseteq E$ is the subset of edges inside the cluster. Edges starting from the nodes of a cluster are divided in *internal* and *external* edges. The *internal degree* K_v^{int} of a node $v \in \mathcal{C}$ is, from eq 1.6, the number of edges connecting v to other vertices of $\mathcal{C}$, on the other hand the *external degree* (k_v^{ext}) is the number of edges connecting v to vertices of the rest of the graph. A good cluster should have many internal edges and few external edges.

A measure which can help to evaluate the goodness of clusters is the *density.* The density of a graph $G(V; E)$ is the ratio between the number of edges of the graph and the maximum number of edges which the network could have. Considering an unweighted graph with n nodes and m edges:

$$\delta(G) = \frac{m}{\binom{n}{2}} \, . \tag{1.9}$$

The density of the subgraph composed by the nodes of the cluster (S), is called *local density.* Since a cluster has both internal and external edges, we need to define two different local densities, the *internal density* (δ_{int}) and the external density (δ_{ext})

$$\delta_{int}(\mathcal{C}_i) = \frac{|\{\{v, u\}|v, \, u \in \mathcal{C}_i\}|}{n_{c_i}(n_{c_i} - 1)/2} \, , \tag{1.10}$$

$$\delta_{ext}(\mathcal{C}_i) = \frac{|\{\{v, u\}|v \in, \mathcal{C}_i, \, u \in \mathcal{C}_j i \neq j\}|}{n_{c_i}(n - n_{c_i})} \, , \tag{1.11}$$

$$\delta_{int}(G|\mathcal{C}_1, ..., \mathcal{C}_k) = \frac{1}{k} \sum_{i=1}^{k} \delta_{int}(\mathcal{C}_i) \, , \tag{1.12}$$

where u v are nodes, $\{v, u\} = 1$ if there is a link between v and u and 0 otherwise, and $n_{c_i} = \mathcal{C}_i$ is the nodes set of the i-th cluster. If the *average internal density* (eq 1.12) is significantly higher than the external and total densities (eq 1.9) the cluster is good.

Communities detection algorithms seek to maximize the difference between internal and external densities over all clusters of the partition.

1.2.1 Community detection methods

The goal of community detection methods [4][2] is to discover subgroups of network elements which are strongly bonded together. Graph clustering normally assumes that the network of interest divides naturally into subgroups, therefore communities are network's own characteristics. According to that, methods should be able to detect communities with different size and in variable number, for this reason clustering methods are often unsupervised. Moreover, community detection methods have to consider that no good division of the network could exist.

As already said, clustering methods can be classified in local and global methods, in the former case only the subgraph of interest and its closer neighbours are inspected. In the latter case every subgraph is considered to be indispensable for the correct functioning of the whole graph. In both cases, measures used to grouping similar objects can be divided in measures based on nodes similarity and measure based on the optimisation of some quality functions.

Similarity measures are at the basis of traditional clustering methods, like hierarchical, partitional and spectral clustering; these methods place a node in the cluster whose nodes are most similar to it. Similarity is computed considering some reference properties, one of the most popular is the *distance*. Since distance does not inspect whether nodes are connected by an edge or not, a distance function cannot be in general defined for a graph which is not complete, but for weighted complete graphs as correlation matrices, it is often possible to define a distance function.

The distance between a pair of vertices can be calculated using the *Minkowski metric*

$$d^E_{XY} = \left(\sum_{k=1}^{n} (x_k - y_k)^g \right)^{1/g} \tag{1.13}$$

where $X = (x_1, ...x_n)$, $Y = (y_1, ...y_n)$ are the two nodes. The commonly used *Euclidean distance* between two objects is achieved when $g = 2$. When the graph cannot be embedded in space, a kind of distance deduced from the adjacency matrix A is often used:

$$d_{i,j} = \sqrt{\sum_{k \neq i,j} (A_{ik} - A_{jk})^2} \tag{1.14}$$

Several methods use distance optimization for clustering in order to maximize inter-cluster distance and minimize intra-cluster distance. When Euclidean distance is chosen, the inter-cluster distances are usually represented by distance of clusters centroids, which are the average of the cluster's elements. With non-euclidean metrics, methods as the shortest distance, weighted or unweighted average distance are used.

Quality measures are usually applied to assess the goodness of a graph partition. Since there are several quality functions, there is not a 'absolute best' partition, but it depends on the metric and the quality function adopted. In this thesis we will introduce two well-know quality measures which were used, at first, to evaluate the goodness of a partition.

One of the first quality function adopted is *modularity*, which was introduced by Newman and Girvan [5]. The idea behind modularity is that a good graph partition occurs when the concentration of edges within community structures compared with a random distribution of links between all nodes is significant. Therefore, the calculus of modularity requires a *null model*; different choice of how to build the null model can lead to different modularity values. For whatever null model, the modularity is defined as:

$$Q_M = \frac{1}{2m}\sum_{ij}(A_{ij} - P_{ij})\delta(C_i, C_j) \tag{1.15}$$

where A is the adjacency matrix, m the total number of edges of the graph, and P_{ij} represents the expected number of edges between vertices i and j in the null model. The function $\delta(C_i, C_j)$ is set to one if $(C_i = C_j)$, that is if nodes i and j are in the same community, zero otherwise.
Usually, the null model is built in order to maintain the degree distribution of the original graph. Therefore, a node i with degree k_i is linked to an other node j with degree k_j with probability $p_i p_j$, where $p_i = k_i/2m$ and $p_j = k_j/2m$. Considering n_c custers, the equation 1.15 can be rewritten as:

$$Q_M = \sum_{c=1}^{n_c}\left[\frac{l_c}{m} - \left(\frac{d_c}{2m}\right)^2\right] \tag{1.16}$$

where l_c is the total number of edges between nodes of the same cluster $\mathcal{C}$ and $d_c = \sum_{i \in C} k_i$ is the sum of the degrees of nodes of $\mathcal{C}$.
Equation 1.16 shows that modularity can take on values in the interval [-1, 1] and, if $\mathcal{C}$ is the whole graph, then Q become zero. Positive values of molularity indicate that the tested subgraphs represent modules, and the higher the modularity, the better defined are the modules. On the other hand, if Q assumes large negative values, it could mean that the intra-cluster edges are fewer than the inter-clusters ones, therefore cluster density is less than the whole graph density.

Another useful quality function used to find community structures is the *stability*[6], [7]. This quality measure merges the idea behind modularity (to evaluate the goodness of the partition), with an inner resolution parameter represented by the Markov time (section 3.2.4). We choose to utilize this quality function in order to evaluate different community partitions, which are obtained in different Markov time. This method will be described in chapter 3, section 3.2.4.

Originally, modularity and stability were used as evaluation criterion, the former to assess the quality of clustering methods and the latter for hierarchical partition methods. Both have rapidly become an optimisation method utilized by many clustering algorithms.

Modularity optimization methods aim to maximise modularity, since high values of modularity are assumed to indicate good partitions. Many algorithms which use modularity as optimisation method were designed, we will describe the first and famous one: the greedy Newman method [5]. This agglomerative hierarchical clustering method starts from $n_c = n$ clusters, each containing a single disjoint node (n is the number of graph nodes). Iteratively one edge of the graph is added to the nodes set, in order to maximize the modularity of the new partition (or minimize its decrease) with respect of the previous configuration. The modularity of partitions explored during this procedure is always calculated from the full topology of the graph. At each step the number of partitions can decrease or not vary, since intra-cluster edges do not merge groups and thus the modularity stays the same. The Newman greedy algorithm ends when all edges of the graph are added, consequently the number of partitions found during the procedure is n. The best partition is the one with the higher modularity.
Many other algorithms are developed in order to reduce the computing time, the most famous are: Clauset, Wakita and Tsurumi and Louvain algorithms.

1.3 Multilayer networks

Multilayer approach belongs to a recent extension of graph theory, called *complex networks theory*, which permits to investigate more complicate frameworks than the classical networks analysis. In fact, it was shown [8] that many real networks can not be exhaustively explain with a *classical* network approach, but need more complex structures. Real networks display complex topological features, which are different from those of random or regular graphs; they present community and/or hierarchical structures, high clustering coefficient and so on. These networks are classified as scale-free networks and small-world networks, which are mentioned above.

Recently, according to the increase of both available data and dataset magnitude and to the developing of a "new" complex approach, graph theory is expanded to 'networks of networks', called also *multilayer networks* . The idea behind multilayer networks is that, building multiple levels of networks, it is possible to explore various types of edges and nodes that are linked together. A multilayer network is therefore characterized by nodes, edges and layers.

In the most general case, different aspects to connect nodes are considered; for every aspect, there can be a set of layers, which are made up of *elementary layers*. Links in different set of layers specify different kinds of relationships between nodes.

An explanatory example is reported in the multilayer review published by Kivela et al [9]. They considered the Zachary Karate Club Club (ZKCC) network as a multilayer network, where nodes represent the current members of the ZKCC. These are scientists who use a particular network, the Zachary Karate Club network, as example in their conferences. In figure 1.1 nodes are labelled with the initials of the four members. Links represent interactions between the ZKCC members, in particular two aspects are analysed: the first one is the type of relationship between the scientists (talked to each other, went to a talk by the other) and the second one represents a conference in which the ZKCC trophy was awarded (and thereby passed from one recipient to the next). In this example there are thus 4 nodes (V), two aspects, and 6 elementary layers. Edges can be intra-layer or inter-layers, the latter are coupling edges because nodes are adjacent only to themselves in different layers but, in more general cases, they can connect different nodes in different layers.

A multilayer network is thus fully described by a quadruplet of components;

$$M = (V_M, E_M, V, \mathbf{L}) \tag{1.17}$$

where $\mathbf{L} = \{L_a\}_{a=1}^{d}$ is the sequence of set of elementary layers, which are L_a for each aspect a. A set of all the combinations of elementary layers can be built using Cartesian product $L_1 \times ... \times L_d$. V are the nodes of the multilayer network, a set of all combinations is given by $V = V \times L1 \times ... \times L_d$. Since the number of nodes can vary between layers, the variable V_M indicates only those nodes which are present in each layer, it is thus a subset of V : $V_M \subseteq V \times L1 \times ... \times L_d$. E_M is the edge set of the multilayer network, which specifies the starting and the ending layers of each edge in addition to the starting and the ending nodes ($E_M \subseteq V_M \times V_M$).

Going back to the previous example, $V = \{MAP, MB, YYA, AC\}$ is the set of nodes, $\mathbf{L}$= $\{$*type of relationship between the scientists, conference in which the trophy was awarded*$\}$ are the two considered aspects which produce 6 different elementary layers: {(X,A), (X,B), (Y,A), (Y,B), (Z,A), (Z,B)}. V_M indicates what nodes are present in each layer, that is: {(MAP,X,A)(MB,X,A) (YYA,X,A), (MAP,Y,A), (MB,Y,A), (YYA,Y,A), ...} . E_M is the edge set ($V_M \times V_M$), edges can be intra-layer or inter-layers. Since in this example inter-layer edges are coupling edges, therefore their corresponding matrix is diagonal.

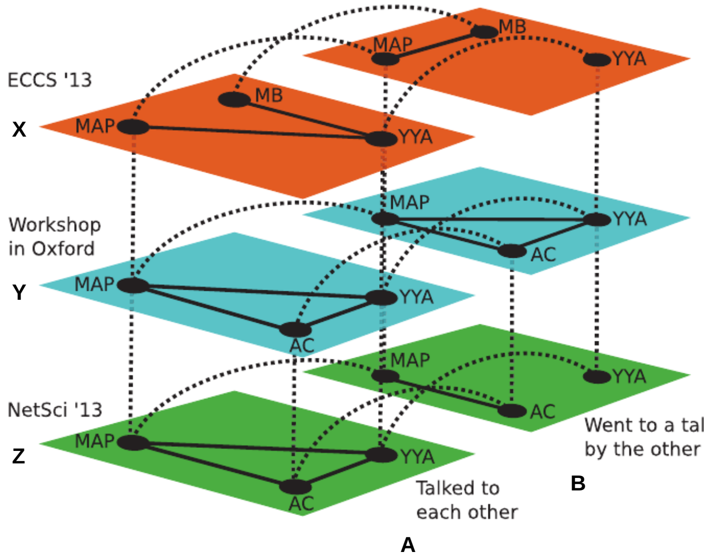


Figure 1.1: Visualization of the Zachary Karate Club Club (ZKCC) network as a multilayer network. [9]

For each multilayer network there is an underlying graph which is identified by:

$$G_M = (V_M, E_M) \tag{1.18}$$

where V_M is the subset of nodes present in layers and E_M is the edge set. The latter can be parted into intra-layer edges (E_A) and inter-layer edges E_C.

In this thesis we focus on a specific type of multilayer network called *multiplex network*, for this reason the theory underneath multiplex network will be illustrate more in depth than the theory of multilayer networks.

1.3.1 Multiplex networks

Multiplex networks [10], [11] [9] are a particular case of multilayer networks. As the latter, multiplex are composed by layers but, unlike multilayer networks, in multiplex networks the number of nodes in every layer is the same. In these networks, edges in different layers can symbolize different types of relations. The structure of a multiplex can be examined both looking at the connections between distinct nodes and at the connections between copies of the same node in different layers, trying to combine intralayer and interlayer relationships. Moreover, the layers of multiplex networks can be

study with different approaches: inspecting measures related to nodes, as node degree and strength, looking at layers sub-structure, as graph partitioning and community overlap, or comparing characteristics of the whole layers. In general, a weighted multiplex is composed of M weighted interdependent networks G_α, with $\alpha = 1, ..., M$ and each one is made up of n nodes. The number and weight of links can vary from network to network. Every network G_α represents a layer of the multiplex, which is therefore described by M adjacency matrices $n \times n$.
The elements a_{ij}^α of the adjacency matrix A^α are defined as:

$$a_{ij}^\alpha \begin{cases} > 0 & \text{if there is link of weight } a_{ij}^\alpha \text{ between node i and node j in layer } \alpha \\ = 0 & \text{ otherwise} \end{cases} \tag{1.19}$$

The adjacency matrix representative of connections between two layers is diagonal. In fact, in multiplex networks a given node i can have different edges in different layers, but it is linked only with itself in different layers.
A multiplex network can be examined evaluating properties of each layers but also reducing its dimensionality. In the latter case data from different layers are aggregated to construct a *monoplex* network, which can be analysed as a elementary graph.
Maintaining the layers structure, there are some useful graph measures which can be extended to multiplex networks. Among those regarding nodes, the *node degree*, the *strength* and the *inverse participation ratio* are very used. These measures are extended to multilayer in order to show dissimilarity between nodes in different layers, From equations 1.6, 1.7 and 1.8, they become:

$$k_i^\alpha = \sum_{i \neq j} \Theta(a_{ij}^\alpha) \tag{1.20}$$

$$s_i^\alpha = \sum_{i \neq j} a_{ij}^\alpha \tag{1.21}$$

$$Y_i^\alpha = \sum_{i \neq j} \left(\frac{a_{ij}^\alpha}{s_i^\alpha} \right)^2 \tag{1.22}$$

where, as before, $\Theta(x) = 1$ if $x > 0$, otherwise $\Theta(x) = 0$.
Comparing values of these measures for every layer (α), it is possible to evaluate the presence of significant difference between them. In order to verify the importance of a single node in all layers also the total node degree (K_i) can be useful $K_i = \sum_\alpha k_i$.

A quantitative estimate of the overlap between layers can be obtained using *multilinks.* For each pairs of nodes i and j there is a multilink $\vec{m} = (m_1, m_2, ..., m_M)$ where

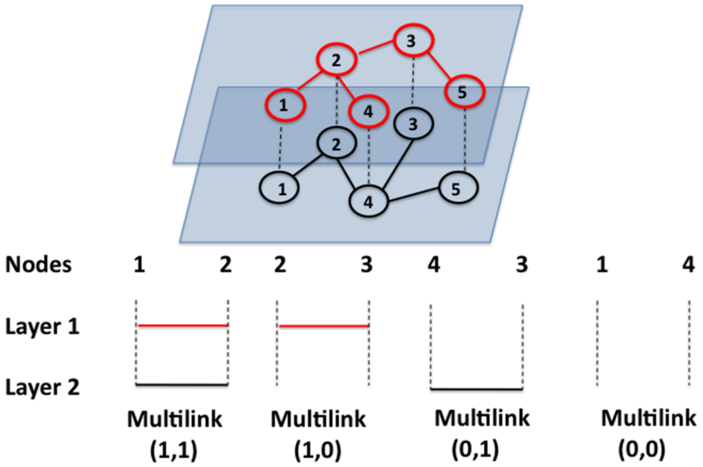


Figure 1.2: Multiplex network with M=2 layers and N=5 nodes. Connection of each pair of nodes i j can be represented by a multilink $\vec{m}$ [10].

$m_\alpha = 1$ if there is the link between i and j in the α layer and $m_\alpha = 0$ if it is not. Figure 1.2 illustrates some possible multilinks for a multiplex network with 2 layers. Using the multilink formalism the adjacency matrix of a multiplex network can be rewrite as

$$A_{ij}^{\vec{m}} = \prod_{\alpha=1}^{M} \left[\Theta(a_{ij}^{\alpha}) m_\alpha + (1 - \Theta(a_{ij}^{\alpha}) m_\alpha)(1 - m_\alpha) \right]. \tag{1.23}$$

Communities detection analysis can be extent to multiplex networks; there are more than one approach to detect clusters in different layers, as there are many clustering methods to define highly connected nodes. One way to seek multiplex communities is to start by separately detecting communities in each intra-layer network. A *multiplex community* can be defined as a set of intra-layer communities which have at least a given number of shared nodes, usually called 'support values' of the community. An other kind of clustering method is based on the *monoplex* network model. In this case community detection involves as for elementary networks, comparing nodes density of the multiplex with nodes density of a null model. As reported by Kievela et al [9], Barigozzi et al. compared intra-layer communities with communities that they found in an aggregated version of their multiplex network (monoplex network), and they observed considerable variation in the intra-layer communities across category layers. It thus seems that much of the information about multiplex communities can be lost by aggregating a multiplex network. This result underlines the importance of a multilayer approach in order to preserve as much information as possible.

1.4 Biological Networks

The application of physics to biological data is growing, according to the increase of the computing power and the admirable obtained results. In particular, we focus on the application of graph theory on biological data [12], [13], [14].
Biological networks allow a description of the complexity of biological systems using the basic components of a network: nodes and edges. Nodes represent different elements according to the biological components would be analysed, moreover various networks can be built using the same set of nodes, since many kind of relationships can be explored. The most studied biological networks are: gene regulatory networks, gene co-expression networks, protein-protein interaction networks, metabolic networks and neural networks.

Gene regulatory networks analyse the interaction between DNA, RNA and proteins; they describe gene expression as a function of regulatory inputs specified by interactions between proteins and DNA. Proteins are synthesised using informations contained in the mRNA molecules, which are transcribed from DNA. Some RNA sequences and regulatory proteins recognize specific DNA sequences and activate or repress the expression of thousands of genes. Proteins, genes, or enzymatic substrates are usually the nodes of these networks, while edges often represent direct molecular interactions, regulatory interactions or the sharing of functional properties. This kind of networks try to deduce the complex structure and dynamic behaviour of genes and proteins, often modules composed of genes with strong expression associations are found. Gene regulatory networks are usually directed, since activation or repression are one-directional relations. Moreover their topology abides by a hierarchical scale free network, there are thus few genes with a high node degree and many nodes with a lower mode degree.

Gene co-expression networks are weighted undirected graphs which evaluate the associations between genes exploiting their DNA transcripts (RNA). Therefore nodes act as genes and links as their co-expression relationships. The adjacency matrix of that kind of networks is usually a correlation matrix based on data acquired by microarray technique; a high value of adjacency matrix entries (a_{ij}) means high co-expression between the two genes (i and j). In this kind of biological networks, a clustering analysis is often performed. The aim is to associate modules to a clinical trait of interest and to study their variations in illness patients.

Protein-protein interaction networks (PPI) analyse the complex protein interactions inside cells. Hundreds of thousands of interactions are collected in biological databases which are used for the construction PPI networks. In order to display all these interactions, some networks were manually built, these databases often allow the identification of functional protein modules.

Metabolic networks examine the chemical reactions, the metabolic pathways and the regulatory interactions of metabolism. Nodes usually represent chemical compounds whereas links are the biochemical reactions which convert a compound into another. In the following section metabolic networks will be discuss, while an extensive description of the field of study of metabolic reaction, the *metabolomic*, will be presented in the next chapter.

This brief characterization of the most inspect biological networks can help to understand the usefulness of graph analysis also for medical aims. In fact, in the last decades, a great interest is turned toward network medicine [15]. This branch of biological networks studies human diseases using a more holistic approach then the scientific reductionist, which relies on single molecules or single genes to provide comprehensive description of complex diseases. Network medicine can be split in some relevant areas: the *interactome* , the *diseasome* and the *epidemics network*.

The interactome is the whole set of molecular interactions in a cell, therefore it merges informations on protein–protein interactions, co-complex memberships, regulatory interactions and metabolic network maps. These networks have permitted to discover *disease modules*, which are groups of cell components linked to disease phenotypes. More specifically, most phenotypes reflect the interplay of multiple molecular components that interact within each others, mutations inside these modules induce to disease phenotypes. Usually disease genes produce disease proteins, which tend to interact within each other forming connected subgraphs called disease modules. The research of functional disease modules has a wide biomedical application; many modules have been discover exploring diseases through their associated phenotypes [16].

Human disease networks (diseasome) are graphs linking different diseases. An edge which connects two vertices indicates shared genes among diseases. Other types of disease networks used communal metabolic pathway or communal phenotypes as links between diseases. Epidemic disease networks connect biological issues to social ones: this kind of network explores the diffusion of contagious diseases caused by biological pathogens.

All these types of networks interact and they form over-structures with a high complexity level. A notable example of these interactions was described by Barabasi [17], in respect of obesity. It is known that obesity has a genetic component, in particular it is related to the allele for the FTO gene (Fat mass and obesity-associated protein), which causes an increased risk of obesity by 30%. The risk changes into 67% if both alleles have the FTO gene variant for obesity. This example shows a high correlation between genotype and phenotype; genes are also tied to transcription factors, RNA,

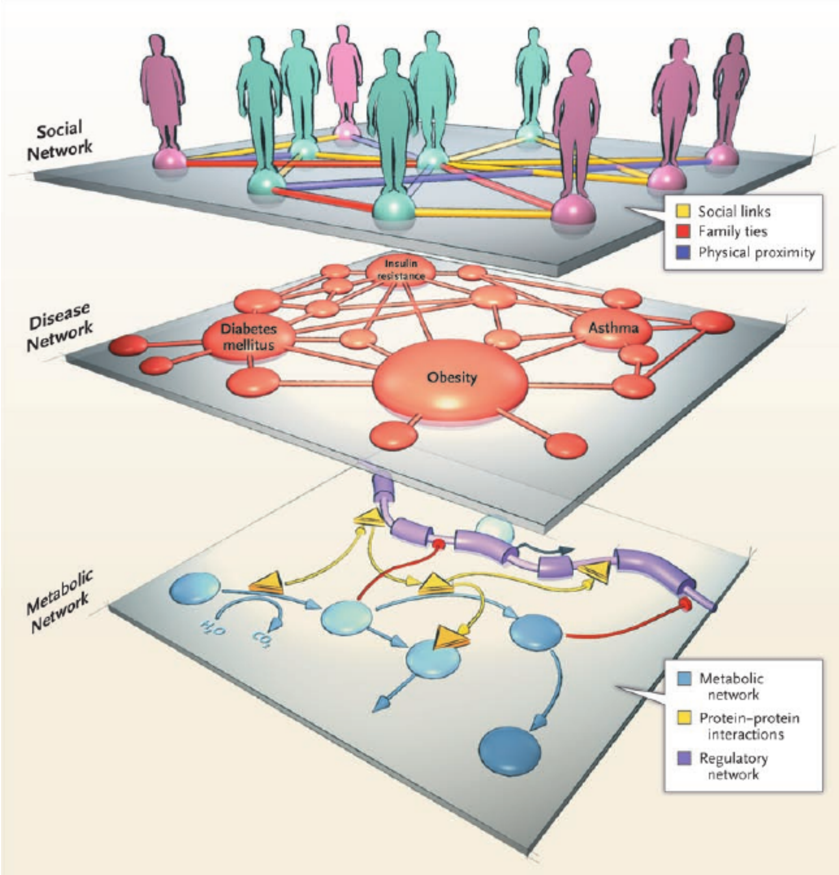


Figure 1.3: The complexity of the analysis of human diseases. Many diseases are associated with the breakdown of functional modules, which lied in different related sub-networks, as genetic, regulatory, metabolic, protein–protein and social networks [17].

enzymes and metabolites. It was also shown that friends, spouses and family members have an increase risk of obesity during a given period, if one of them became obese in that time interval. That risk is 171% for fiends and 40% for siblings. Obesity is then clustered into communities in a sort of epidemic disease network. Many genes linked to obesity are also related to other diseases. The current diseasome reveals that obesity shares genes with diabete , asthma, insuline resistence, lipodystrophy, glioblastoma and so on. This analysis of obesity shows the complexity of biological systems, where social networks are tied to genome, metabolic and disease networks. In figure 1.3 is graphically represented the overlapping of such networks.
This example permits to grasp the importance of a multilayer approach. In fact it helps to tie different kind of networks and to have a more general point of view then that offered by simple graphs.

1.4.1 Metabolic networks

Metabolic networks are a mathematical representation of the metabolism of an organism, that is the totality of chemical interactions that generates essential components such as amino acids, sugars and lipids, and the energy required to synthesize them and to use them in creating proteins and cellular structures. Therefore, the actual graph representation is a network where nodes represent metabolites and where edges are the chemical reactions which transform them.
A series of chemical reactions occurring within a cell is called *metabolic pathway*; these pathways had been manually drawn and their collection was grouped in molecular interaction and reaction networks; one of the largest pathway database is the *KEGG (Kyoto Encyclopedia of Genes and Genomes) PATHWAY Database*[18]. These metabolic pathway databases are used to build various metabolic networks where links can symbolize different kind of relations between metabolites.
One important study about topological properties and characteristics of biological networks was accomplished by Jeong et al.[19] in 2000. They analysed metabolic networks of 43 organisms representing all three domains of life, where a metabolite (node) is directly links to another metabolite if there is a chemical reaction which transforms the former in the latter. Their results show that all metabolic networks are described by scale-free networks: the probability that a node has k neighbours ($P(k)$) follows a power-law, $P(k) \approx k^{-\gamma}$. These networks have an extremely heterogeneous topology in which few nodes are highly connected (*hubs*) and most of them are less connected. Moreover, metabolic networks seem to be small world networks, since even if many nodes have few edges they can be connected with all others nodes with short paths. All the 43 networks exhibited module structures responsible for distinct metabolic functions, and the hubs were the same for every analysed specie. They also found that metabolic networks are robust against random errors, that is: removing random nodes the average distance between the remaining nodes were not affected. On the other hand, a removal of the hubs caused a fast increment of the diameter. An other particularity found by Jeong et al. is that the diameter (eq 1.5) of the metabolic network is the same for all 43 organisms, instead of increasing according to the number of nodes with the typical logarithmic trend of scale free networks. This indicates that more complex organisms with more enzymes and substrates have an increased connectivity. This permits to maintain a relatively constant metabolic network diameter.

As said in the previous section, metabolic networks can be used to build disease networks. A great example of this method is shown by Lee et al. [14] who studied *the implication of human metabolic network topology for disease comorbidity*, which

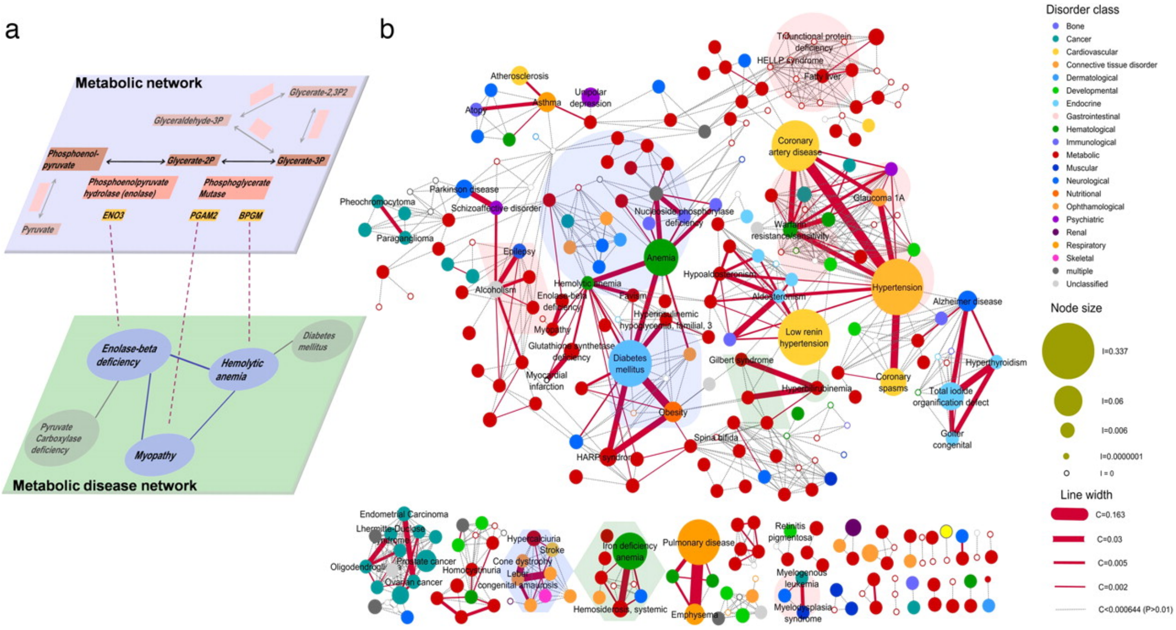


Figure 1.4: Example of the method used to build the disease network a) and disease network b) obtained by Lee et al. [14].

means that diverse disease phenotypes are coexpressed. They started looking that mutations that cause a metabolic enzyme to be nonexpressed, inactive, or functionally compromised can be associated to metabolic diseases, which are usually classified using some disease phenotypes. Moreover, since metabolites are grouped in functional modules such as carbohydrate, amino acid, fatty acid metabolism which are connected together, a dysfunction of one of them can have a cascade effect on the others. They built a disease network starting from the metabolic network. Two disease were linked if mutated enzymes associated with them catalyze adjacent metabolic reactions, as shown in figure 1.4.a). The figure 1.4.b) displays the obtained disease network. Implementing comorbidity analyses they found that connected diseases shown higher comorbidity than those that have no metabolic link between them. Moreover the strongest predictors of comorbidity are the metabolic links and not shared genes. This fact underlines the importance of the study of metabolic data, since they seem to be highly related to the phenotypes and diseases.

In this chapter we have proposed a brief introduction of the network theory, mentioning some of the most useful measures and describing some important graph features. In particular, we have illustrated the main characteristics of the complex network theory, in order to lay the foundation for the analysis we will do.

Chapter 2

Metabolomics

Metabolomics is one of the emerging disciplines of *omics* research as well as lipidomics and proteomics. The suffix *omics* is commonly used in bioinformatics for classifing a specific field of research. Therefore, the term metabolomic concerns the study of all metabolites being in the organism.

Since our study belongs to metabolomics, in this chapter we will focus on the biological and health motivations which led to the collection of the DILGOM dataset, which we will analyse with a physical and statistical approach in the following chapters. In the first section we will summarise the characteristics of *system biology*, focusing on the metabolomics branch. Consequently the reasons and the aims of DILGOM study will be illustrated; moreover, we sum up the results obtained in previous studies showing the usefulness of the network approach. In section 2.2 the main methods to extract serum compounds will be presented, in particular we will give a brief explanation of ^1HNMR, chromatography and mass spectrometry methods.

Later (section 2.3) we will show a fleeting overview of the incidence of obesity and the cardiovascular diseases correlated to that. In section 2.4 some correlations between the compounds extracted by the DILGOM study and cardiovascular diseases will be listed. These compounds are grouped using a classification based on their biological characteristics, therefore we will give a brief explanation of the their metabolic roles.

In the last section we will describe the DILGOM dataset, focusing on the phenotypes and metabolic dataset, since we will apply our multiplex analysis method on these data.

2.1 Metabolomics

Metabolomic, as lipomic and proteomic, is recent inter-disciplinary field of study. Respectively, these *omics* concern the study of all metabolites, lipids and proteins being in the organism. The *omics* have been a rapid development according to the increase of available data and computing power. The first metabolomics web database METLIN [20] was developed in 2005; its data regard metabolic profile, which is the measure of the compounds which are present in human liquids and tissue extracts. An other fundamental project for the progression of metabolomic studies is the *Human Metabolome Project* [21], which has been freely available online since 2007. The Human Metabolome Database contains detailed information about 41,993 small molecule metabolites found in the human body. These are detected and quantified using the two main techniques adopted for human compounds analysis: the mass spectrometry and the ^1H NMR. These two powerful methods will be describe later in this chapter. Other analytical technologies have been employed to extract, detect, quantify, and identify metabolites in different organisms, tissues, or fluids in order to increase informations about these small molecules.

The importance of a systematic study of the whole metabolic profile descends from the fundamental functions that these compounds have in the body, which include fuel, structure, signaling, stimulatory and inhibitory effects on enzymes, catalytic activity, defense, and interactions with other organisms. Metabolites can thus be seen as the functional readout of cellular state.

Even though the aim of the metabolomics is to understand all the chemical processes which involve metabolites, it is also central for "inter-omics" analysis, since metabolites are often connected to phenotypes, diseases, toxicity and eating habits. Metabolomic can thus be seen as a bridge between phenotype and genotype. This new branch is called functional genomics and it is applied on model organisms. The aim is to predict the function of unknown genes by comparison with the metabolic perturbations caused by deletion/insertion of known genes. An even more complex study is the nutrigenomic, which tries to establish links between genomic, transcriptomics, proteomics, metabolomics and human nutrition.

All these omics are analysed by the *sistems biology*, which is the computational and mathematical modelling of complex biological systems [22]. The ultimate goal of systems biology is to integrate these different omics using a holistic approach. This can leads to discover emergent properties and connections which are not identifiable using reductionistic methods. The idea behind this holistic approach derives from the observation of living systems: these are dynamic and complex systems and their nature is not deducible from the properties of their singular parts. Since there are hundreds of thousands of elements which participate at the human body (or cellular) functions, graph approach seems to be the most practical investigation method. Each component can be represented as a node and its interaction as link. Therefore biological networks are usually constituted by thousand of nodes connected by thousand of edges.

Metabolomic, unlike genomic and proteomic which have been developed for the last forty years, has been investigated only for the last few years. Since it is a very recent field of study, there is not a well-established background, but many articles have been published recently, highlighting the interest of the scientific community.

In the current chapter, we will focus on the biological and methodological aspects related to metabolomics, describing the method used to extract serum particles concentration and the health motivations which induced these data collections. In particular, we will examine in depth the dataset utilized in our study and the goal of that project, detailing the biological functions of the quantified metabolites and their correlation with cardiovascular disease.

2.1.1 The DILGOM study, biological aspects

The purpose of the DIetary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study is to realize a comprehensive characterization of the biological architecture of human diseases and traits[] [23]. More specifically, using a network approach the ambition is to understand how nutrition, diet, lifestyle, psychosocial factors, environment and genetics of a population are linked to obesity and to the metabolic syndrome.

As already discussed in the previous chapter, many diseases present complex subnetworks structures, which are usually characterized by the breakdown of functional modules. In particular, many studies have shown that obesity is influenced by genetics, metabolic and social factors.

For this reason the DILGOM study collected informations regarding phenotypes, metabolomics, transcriptomics and genetics of 518 Finns. The integration of these 'omics' may lead to the discover of new module structures linked to the metabolic syndorme and to assess, in a more accurately way, the risk for various vascular outcomes. The DILGOM study consists of unrelated individuals, 240 males and 278 females, aged 25–74 years, sampled from the capital region of Finland.

Using the DILGOM database, network analyses have already recognised a set of highly correlated genes, the lipid–leukocyte (LL) module, as having a prominent role in over 80 serum metabolites [24]. Using a co-expression network, Inouye et all [?] identified 11 genes with strong expression associations with HDL (High Density Cholesterol) ($P = 5.62 \times 10^{-27}$), APOB (apoliprotein B)($P = 3.06 \times 10^{-26}$), and TG levels (total triglycerides)($P = 2.44 \times 10^{-29}$). They supposed that the module composed by these 11 most strongly associated genes for APOB, HDL, and TG could be a candidate module for metabolic dysfunction, inflammation, and atherosclerosis; moreover also adiposity could be linked to the LL module. Indeed, network analyses have shown a dependence of gene co-expression leukocytes from serum metabolite concentrations, which are conditioned by environmental factors. This beginning result underlines the importance of a systematic molecular and environmental analysis.

In this thesis we analyse the phenotypes and the metabolomic dataset using a network approach. The aim is to verify if blood serum particles of obese people show a different behaviour than those of normal-weight people. The presence of compounds or modules with different behaviours in the two groups could be related to some diseases recurring in obese people.

2.2 Methods of data extraction

In the past decades, more than one method for metabolic profiling of biological specimens was developed. In fact it was proved that many diagnosis can be done from blood concentrations. Two of the most widely used analytical platforms for metabolic profiling are nuclear magnetic resonance (NMR) spectroscopy, and chromatography joined to mass spectrometry (MS).

2.2.1 ^1HNMR

NMR spectroscopy has become the most adopted technique for determining the structure and the concentration of organic compounds. It takes advantage of the intrinsic quantum property of spin (I) which all nucleons have. A spinning charge generates a magnetic field which has a magnetic momentum $\vec{\mu}$:

$$\vec{\mu} = \gamma_n \hbar \vec{I} \tag{2.1}$$

μ is proportional to the spin (I), to the Plank constant h and to the gyromagnetic ratio γ_n. The latter is the ratio of the magnetic dipole momentum and the angular momentum of a specific nuclear element n, therefore it assumes different values relative to the chemical element. The nucleons spin can be $\pm 1/2$, the overall spin of a nucleus is the sum of the nucleons spin whereof nucleus is made, and therefore it can be integral, fractional, but also zero. In the latter case, the nucleus magnetic momentum is also equal to 0 (eq 2.1) and thus nuclei of that species can not be detected using NMI.

Without an external magnetic field, the momenta of a given chemical element, such as hydrogen nuclei ${}^1H$ ($I = \pm 1/2$), are in random directions. When they are placed in a external magnetic field (B_0) they align their momentum either with or against B_0, respectively if they have $I = +1/2$ or $I = -1/2$. These two nucleus states have a little difference in energy: the state with $I = +1/2$ is aligned with B_0 and it is placed to a lower energy level in respect to the state with $I = -1/2$. Therefore, the application of an external magnetic field causes a little difference in energy between the two spin states, which is directly related to the field strength.

$$\triangle E = \frac{\mu B_o}{I} \tag{2.2}$$

Considering a set of ${}^1H$ nuclei in a magnetic field B_O, they will split in two states on the basis of the Boltzmann distribution: the lower energy state (corresponding to $I = +1/2$) will be more populated in comparison to the higher energy level. Irradiation of the sample with radio frequency (rf) energy corresponding exactly to the spin state separation ($\triangle E$) causes an excitation of those nuclei in the $+1/2$ state to the higher $-1/2$ spin state. There is thus a net transition from the lower level to the higher one which produces a net energy absorption. This absorption is quantifiable and produces the NMR signal. The resonance frequency corresponds to the *Larmor precession frequency* $\nu_L = |\gamma/2\pi|B_0$, which indicates the frequency of the precession motion of the magnetic moments $\vec{\mu}$.

For NMR purposes, the utilized frequency ranges from 20 to 900 MHz, depending on the magnetic field strength and the specific nucleus being studied.

A powerful application of NMR is the Proton NMR Spectroscopy [25]. There are more than one procedure to obtain the NMR spectrum; we will describe the simplest one which is the continuous wave (CW) method. In figure 2.1 the principal aspects of a NMR spectrometer are schematically illustrated. At first, the sample tube is oriented between the poles of a powerful magnet, and is spun to average any magnetic field variations, as well as tube imperfections. Radio frequency radiation with fixed frequency (ν_{rf}) is broadcast into the sample from the rf transmitter. A receiver coil surrounds the sample tube and the emission of absorbed rf energy is monitored by dedicated electronic devices and a computer. The magnetic field B_0 is slowly intensified, until the resonance condition is reached ($\triangle E = E_{rf} = \mu_{rf} h$) and a signal is thus detected by the radio frequency receiver. The magnetic field intensity continues to improve until the final value B_f, and the radio frequency signal is no more revealed. An equally effective technique is to vary the frequency of the rf radiation holding the

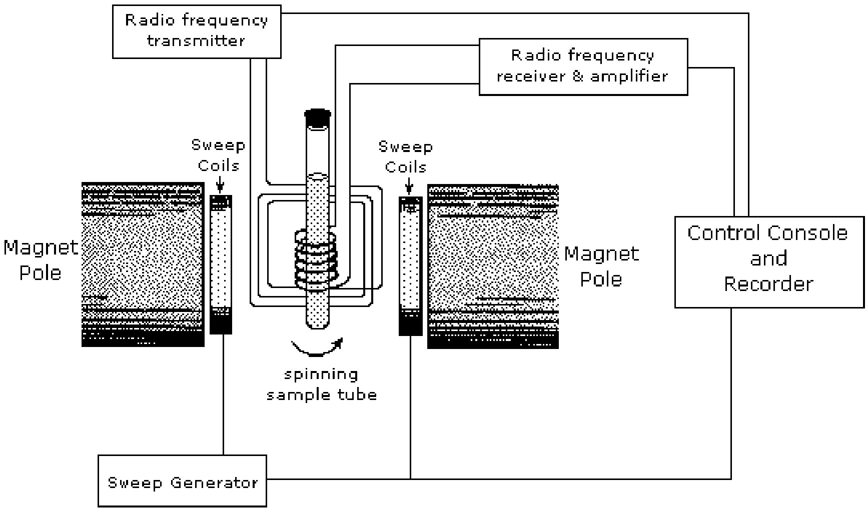

Radio frequency transmitter
Radio frequency receiver & amplifier
Sweep Coils
Sweep Coils
Magnet Pole
Magnet Pole
Control Console and Recorder
spinning sample tube
Sweep Generator

Figure 2.1: Schematic representation of a ^1H NMR spectrometer.

external field constant.

Using the ^1H NMR it is possible to detect signals from different compounds which have at least one ^1H. A classical ^1H NMR spectrum is illustrated in figure 2.4 a). The discrimination between different compounds which have all the same absorbing nucleus (^1H) is possible thank to a characteristic called *chemical shift*. Effectively, the Larmor precession frequency of a given nucleus depends not only on the spin and the gyromagnetic ratio of the element, but also on its chemical environment. This causes local microscopic magnetic fields which add themselves up to the external magnetic field B_0. The local magnetic field shields the nucleus from B_o by a factor σ called *shield factor*. The actual external magnetic field perceived by the nucleus is $B_o^{eff} = B_o(1-\sigma)$, thus the Larmor frequency is $\omega_L = \gamma_n B_o^{eff}$. The change of the external field intensity perceived by ^1H in different compounds is very small if compared with the actual external field (about 0.0042%). The chemical shift in respect to a reference compound is commonly measured in parts per million (*ppm*).

The ^1H NMR spectrum allows to reveal signal of different components but it also gives the relative ratio of the number of H for each peak. Effectively, the number of hydrogen atoms is directly proportional to the area of the peak, that is the intensity of the NMR signal.

For these reasons ^1H NMR Spectrometry has grown significantly in metabolomics in the past two decades, since it gives quantitative measurement of multicomponent in a complex mixture. Many methods are been developed in order to improve the accuracy and the precision of NMR, and now the quantitative inaccuracy of qNMR

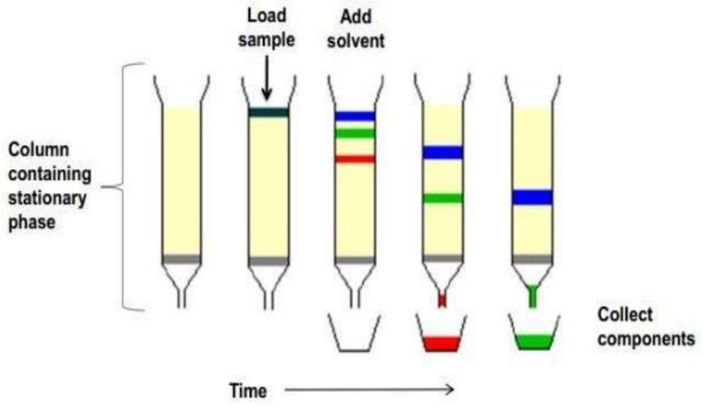


Figure 2.2: Schematic representation of the column chromatography.

has been reported to be less than 2.0% [26].

2.2.2 Chromatography and mass spectrometry

Chromatography technique is used to separate compounds of a mixture, it is often associated to mass spectroscopy in order to characterize quantitatively and qualitatively these substances. There is more than one kind of chromatography, but they are all based on the same principle of differential partitioning of components between mobile and stationary phase [27].

Considering the column chromatography, the stationary phase or adsorbent is a solid, usually silica gel, whether the mobile phase is a solvent. The stationary phase (solid adsorbent) is placed in a vertical column and the mixture to be analysed is placed inside the top of the column. The mobile liquid phase is added to the top of the column and, opening the tap at the bottom of the column, it flows down by either gravity or external pressure. Adding repeatedly fresh solvent to the top of the column, different components present in the sample start to form separate strips. This is caused by the different polarisation of the components, which tie themselves to the absorbent with different bond intensity. Since different components in the mixture have different interactions with the stationary and mobile phases, they will be carried along with the mobile phase to varying degrees, consequently separation of the strips will improve during time. The individual components, or elutants, are collected as the solvent drips from the bottom of the column. In figure 2.2 it is schematically illustrated the chromatography technique, where different colours represent different compounds of the analysed mixture sample.

The high improvement of the column chromatography is the *High Performance*

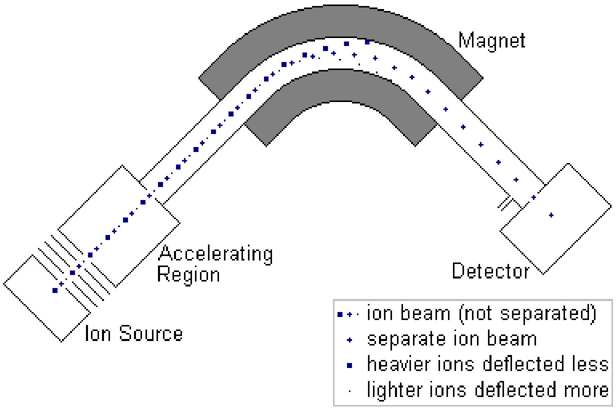


Figure 2.3: Schematic representation of mass spectrometer.

Liquid Chromatography (HPLC), a liquid chromatography which permits to separate, to identify, to purifier and to quantify each component of a mixture using an automatic process. Instead of solvent being allowed to drip through the column under gravity, it is forced through under high pressure (up to 400 atmospheres). The typical column measures are 2.1-4.6 mm diameter and 30-250 mm length; high pressure and small column dimensions allow a remarkable saving of time (an analysis usually required 10-30 minutes). Also sensitivity and resolution are extremely increased thanks to the advances in instrumentation and column technology. The separated components are automatically collected and identified using photodiode or spectrophotometer, but the most common parameter for compound identification is the retention time, that is the time of eluition. Using a chromatograph it is possible to perform chromatographic separation. This equipment produces *chromatograms*, which are graphs showing the quantity of a substance leaving a chromatography column as a function of time. The peak area is measured in order to know the amount of the compound of interest. A real chromatogram of plasma lipids is illustrated in figure 2.4.b.

Some instruments merge chromatography, which divides compounds of the studied mixture, to mass spectrometry, which quantifies the amount of compounds.

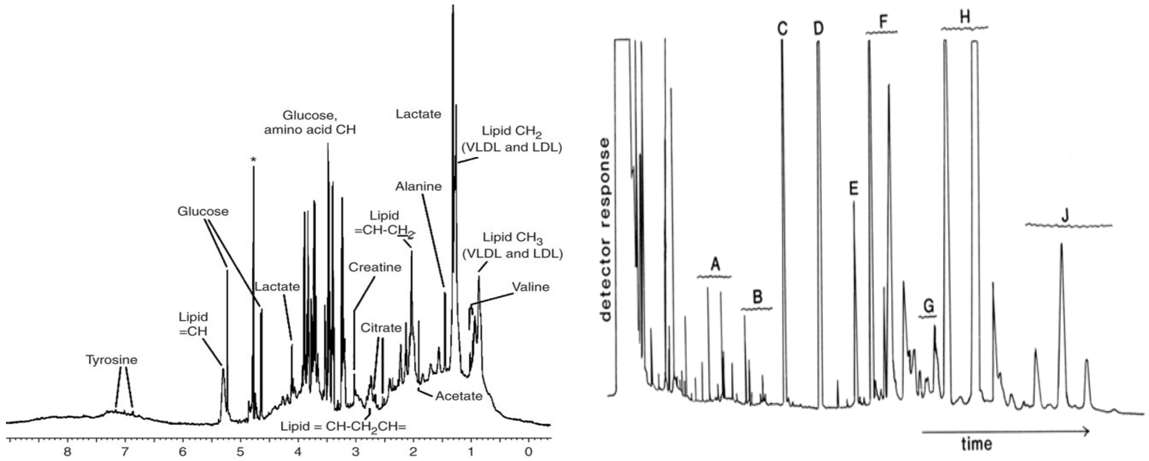
Mass spectrometry is an analytical tool which is utilised for the measurement of the mass-to-charge ratio of charged particles. This technique allows also to characterize individual molecules, such as to determine their masses, their elemental composition and their chemical structure. All mass spectrometers can be divided in five basic parts which perform the essential functions: a high vacuum system, a ion source which ionizes the sample, a mass analyser which sorts and separates the ions according to their mass and charge and a detector which implements measurements. In figure 2.3 the process conduct by a mass spectrometer is schematically illustrated. The samples are loaded into the mass spectrometer and then they are vaporised and ionised by the electrons issuing from a heated filament of the ion source. The whole analysis is conducted in vacuum since ions are very reactive and short-lived particles. The cations formed by the electron bombardment are pushed away by a charged repeller electrode whereas anions are attracted to it. The positive ions are then accelerated by electrodes and collimated in a beam. A static magnetic field (B) perpendicular to the ion beam is applied, ions of mass m and charge z moving in vacuo with a velocity v follow a circular path with radius r

$$r = \frac{mv}{Bz} \tag{2.3}$$

Ions with the same charge and momentum follow the same path. Therefore ions of different masses can be focused on the detector by varying the strength of the magnetic field (from equation 2.3). The detector is fixed at the end of the curved tube. Mass spectrometers utilize software which analyse the ion detector data. They produce graphs where the detected ions are organised by their individual mass-to-charge ratio and their relative abundance. Spectra are then compared with databases to predict the identity of the molecules.

Study of biological compounds are often realised using the mass spectrometry associated with the chromatography, the latter is typically performed for the separation of the mixture to be analysed. In the last years sophisticated methods for the ionization of macromolecule have been developed. Moreover, mass accuracy and resolution have been improved since biological molecules are complex. Current methods allow to quantified compounds both in relative and absolute terms. Technology behind mass spectrometer is constantly growing to accommodate large-scale, high-throughput experiments [28].

DILGOM study [23] used both NMR and chromatography techniques to acquire the serum compounds concentrations of $\sim$ 140 metabolic measures. Since the multi-metabolic nature of serum, ^1H NMR spectra can contain signal overlap. The metabolite content and concentrations can be extracted by appropriate experimental settings and advanced computational techniques. The work frequency of the ^1H NMR spectrometer was set at 500.36 MHz and the temperature of the samples was approximately 0.01^oC. High-performance liquid chromatography were used to calibrate the 14 subclasses of bad cholesterol, which are commonly classified using the average particle diameter (table 2.5). Concentrations of such subclasses were also realized by ^1H NMR, which permits to quantify some lipoprotenin characteristics, such as total triglycerides free and esterified cholesterol and total cholesterol. The spectrum of human serum lipids LIPO were used to arise different lipid molecules in various lipoprotein particles. Some of them are also visible in the LMWM spectrum, which contains mainly glucose reso-



a) b)

Figure 2.4: ${}^1H$ NMR signals of low molecular weight metabolites a) as well as of larger molecules such as lipoproteins. Real chromatogram b) of plasma lipids, Abbreviations: A, TMS ester derivatives of free acids; B, TMS ethers of monoacylglycerols; C, TMS ether of cholesterol; D, tridecanoin; E, TMS ether of 16:0-sphingosine ; F, TMS ethers of diacylglycerolsand ceramides; G, more TMS ethers of ceramides; H, cholesterol esters; J, triacylglycerols (from Nature protocols).

nance.

2.3 Biological and health aspects

The goal of the DILGOM study is to understand how nutrition, diet, lifestyle, psychosocial factors, environment and genetics of a population are linked to obesity and to the metabolic syndrome. In this section the implications of such syndrome will be described, in order to underline the health importance of that kind of studies.
Over-weight and obesity are defined as abnormal or excessive fat accumulation that presents a risk to health. More specifically, overweight refers to an excess amount of body weight that may come from muscles, bones, fat, and water, when obesity refers only to an excess amount of body fat. Both obesity and overweight result from an energy imbalance and can be caused by several factors, as genetic aspects, eating habits, geography, emotions, dysfunctions and so on [29].

Many studies show that overweight and obese people have a major risk of chronic diseases, including hypertension, diabetes and musculoskeletal disorders. Obesity enhances the risk of certain forms of cancer and, above all, cardiovascular diseases (mainly heart disease and stroke) which are the leading cause of death. Every year, an estimated

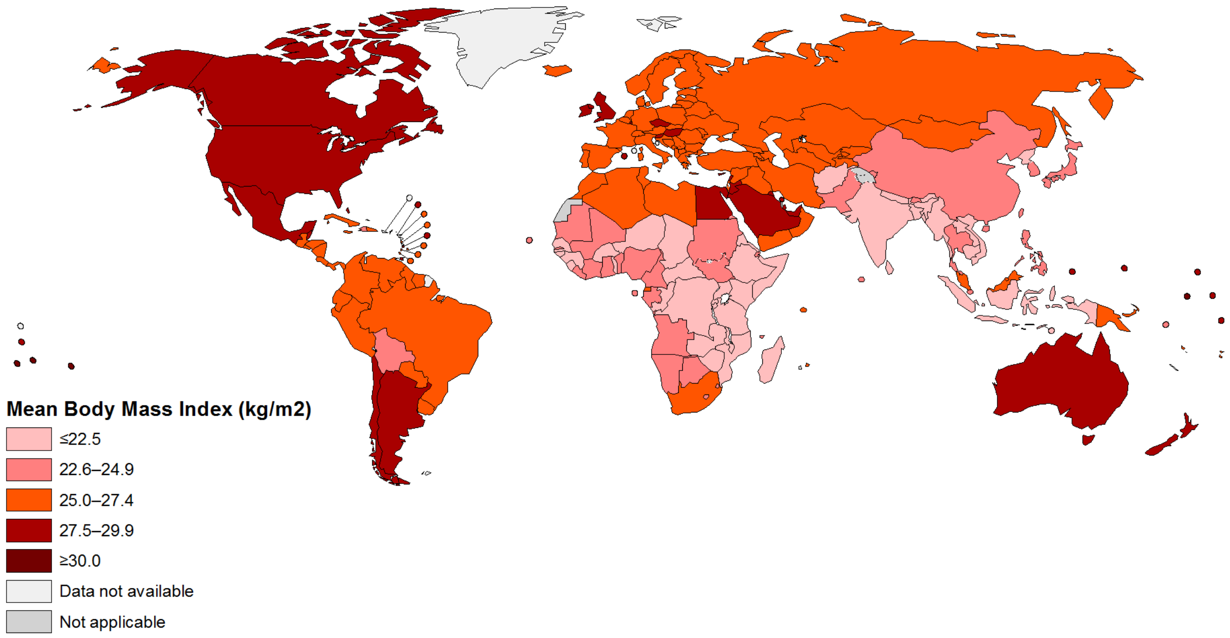


Figure 2.5: Planisphere where colours are associated to the mean BMI of each country (Figure of WHO [30])

.

17 million people globally die of cardiovascular diseases (CVD) and, approximately, the 50% of deaths in the European Region are caused by cardiovascular diseases. Mortality rate increases with increasing degrees of overweight, which is in perpetual raise: in 2014, more than 1.9 billion adults, 18 years and older, were overweight; among these over 600 million were obese. In figure 2.5 is displayed a planisphere where colours are associated to the mean body mass index in each country (image of WHO). On this basis one can understand the scientific interest about these risk factors.

There are some phenotypic measures used to classify people as obese. The most adopted are the body mass index (BMI) the waist-hip ratio (WHR) and the waist circumference. BMI is calculated as a person's weight (in kilograms) divided by the square of his height (in metres) and WHR is the ratio between waist and hip circumference. For these parameters the World Health Organisation (WHO) has established thresholds in order to classify individuals as normal weight, overweight and obese [31]. The threshold values for the three mentioned parameters are reported in table 2.1. Figure 2.7 shows the histogram of the BMI of the individuals analysed in the DILGOM study.

	gender	BMI [kg/m^2]	WHR	waist circumference [cm]
obese	F	>29	>0.85	>88
	M	>30	>0.95	>102
normal weight	F	<24	-	< 80
	M	<25	-	< 94

Table 2.1: Threshold values established by WHO to classify normal weight, overweight and obese individuals.

2.3.1 Risk factors and Cardiovascular Diseases

The presence of significant differences between the behaviour of blood serum particles in obese and in normal weight people. If they are present, it would help to diagnose risk of cardiovascular diseases. Even though overweight and obesity are the most relevant risk factors of cardiovascular diseases, other factors are important too: high blood cholesterol and triglyceride levels, high blood pressure, diabetes and prediabetes, smoking, lack of physical activity, unhealthy diet, stress, age, gender and family history [32]. All these risk factors are related together; for example, obesity typically raises pressure and cholesterol levels and lowers HDL levels. It predisposes to type 2 diabetes and it is correlated with unhealthy diet and physical inactivity. The presence of these risk factors enhances the probability of contracting Cardiovascular diseases (CVDs) [33].

CVDs are disorders related to the cardiovascular system, that is heart and blood vessels. In particular, they are categorized into: coronary heart disease (heart attacks), cerebrovascular disease (strokes), peripheral arterial disease, rheumatic heart disease, congenital heart disease, deep vein thrombosis and pulmonary embolism.

Cardiovascular diseases are globally the first cause of death; it is estimated that 31% of all global deaths is due to CVDs in 2012, as reported by the *world Health Organisation.*

2.4 Correlations between diseases and serum substances

Many biological and medical studies have found correlations between serum values and diseases.

The DILGOM study analysed 134 substances of human serum, which are divided into lipoproteins, metabolites and amino acids, fatty acids and others derived measures. We report the connections found between the blood elements analysed in the DILGOM study and the diseases for which over-weight and obesity are risk factors.

2.4.1 Lipoproteins

Lipoproteins are particles comprising several thousands molecules. They have a single layer of phospholipid molecules and apoliproteins on their outside, surrounding a central core of triglycerides and cholesterol. Lipoproteins are synthesized in the small intestine and in the liver; their role is to transport substances contained in their core to all body tissues.

They are divided into VLDL (Very low density cholesterol), IDL (intermediate-density lipoprotein), LDL (Low Density Lipoprotein) and HDL (High Density cholesterol).

VLDL is a lipoprotein made by the liver and it is composed of 55-65% triglycerides, of 10-15% cholesterol, of 15-20% phospholipid, and of 5-10% protein. Its density is between 0,95 and 1,006 g/ml and its diameter is around 30-80 nm. Once in the capillars, VLDL interacts with HDL and cedes triglycerides and phospholipid to muscles and takes cholesteryl esters. Then its density changes, the percentage of triglycerides becomes 50% and it becomes IDL.

IDL's density varies between 0,95 and 1,063 g/ml, and its diameter is around 25-35 nm. The IDL particle can be removed from the blood by the liver or converted to LDL; usually a half is removed. During the conversion to LDL, much of the remaining triglycerides are removed and LDL triglycerides are around 20%.

LDL particles have a density between 1,006 and 1,063 g/ml and a diameter around 18-25 nm. They transport mainly cholesterol and cholesteryl esters to pheripherical tissues, but if LDL's particles concentration exceeds the amount of cholesterol required by cells, then accumulate themselves within the walls of arteries.

HDL (High Density Lipoprotein) particles have a higher density, which ranges from 1,063 to 1,210 g/ml and they have a diameter around 8-11 nm. They are synthesized in blood by plasma enzymes from free cholesterol and by some apoliproteins produced by liver.The main function of HDL is to incorporate cholesterol from tissues and others lipoproteins and to transport them to the liver.

Correlation between Lipoproteins and CVD

Many studies [34] have found a high correlation between plasma lipoproteins and cardiovascular disease, specifically coronary artery diseases (CAD). A study carried out in [35] on 182 patients found a significant correlation between coronary atherosclerosis diseases, and serum lipid and lipoprotein concentrations. In particular, they found that CAI correlated significantly and positively with total, VLDL, IDL and LDL cholesterol ($p<0.01$, $p < 0.05$, $p < 0.01$ and $p < 0.01$, respectively) and negatively with HDL cholesterol ($p < 0.01$).

In general, high values of VLDL, IDL and LDL lipoproteins are positively correlated with CVDs, while HDL cholesterol correlates negatively with CVDs. The most used parameter is LDL blood concentration, since persistent high blood levels of LDL produce plaques, that is a thick, hard deposit that can clog arteries and make them less flexible. This condition is called arteriosclerosis. The most relevant consequences of arteriosclerosis are: heart attack, stroke and peripheral artery diseases.

2.4.2 Fatty Acids

Fatty acids are lipids implied in many biological processes. They are used for energy production by citric acid cycle, and are also converted into triglycerides, phospholipids, hormones and others essential molecules. They are classified into unsaturated and saturated fatty acids, according to the presence or the absence of carbon–carbon double bonds. The former help to lower the levels of total cholesterol and LDL cholesterol in the blood, whereas the latter have a high positive correlation with blood LDL cholesterol. There are also two essential fatty acids: alpha-linolenic acid (an omega-3 fatty acid) and linoleic acid (an omega-6 fatty acid). These two fatty acids are not synthesized by human metabolism, therefore they must be introduced by means of the diet. Free fatty acids come from the breakdown of a triglyceride within adipose tissue. They can be used as an immediate source of energy by many organs and can be converted by the liver into ketone bodies (see Acetoacetate 2.4.3).

Fatty acids are contained in many types of food, in particular unsaturated fats are present in avocados, nuts, and vegetable oils, (especially olive oil); animals fats usually contain both saturated and unsaturated fats.

Disease correlations

The most important health authorities, such as the World Health Organization, the American Dietetic Association, the British Dietetic Association, the American Heart Association, the World Heart Federation and the Food and Agriculture Organization (FAO), highlight the role of diet, claiming that 'adequate amounts of dietary fat are essential for health'. Moreover it is universally acknowledged that a diet rich in saturated fatty acids is directly related to coronary heart disease, due to increasing LDL cholesterol caused by fatty acid [36]. A study carried out in 2008 [37] found a correlation between blood levels of free fatty acids and obesity. In particular it is proved that obesity is closely associated with insulin resistance. Free fatty acids (FFA) cause both insulin resistance and inflammation in the major insulin target tissues and thus are an important link between obesity, insulin resistance, inflammation and the development

of type 2 diabetes, hypertension and other diseases.
On the other hand, a large body of scientific research suggests that higher dietary omega-3 fatty acid intakes are associated with reductions in cardiovascular disease risk, especially coronary hearth diseases [38].

2.4.3 Metabolites and amino acids

Metabolites are the product of enzyme-catalyzed reactions that occur naturally within cells. To be classified as a metabolite, a compound must have some characteristics as: a finite half life, in order not to be accumulated in cells, and a useful biological function [39]. Thus, metabolites are compounds that intervene between the start and the end of a pathway, where a *pathway* indicates a series of chemical reactions that occur within living cells. In figure 2.6 some important reactions which involves metabolites are shown, in particular, the image exemplifies the most important connections between the metabolites take into account in the DILGOM study.
Amino acids are organic compounds which are used both as building blocks of proteins and as intermediates in metabolism. All proteins are built using 20 amino acids, which are classified as *essential* if they must be supplied by means of food and *non essential* if the body can produce them. The 10 amino acids the body can build are: alanine, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, glycine, proline, serine and tyrosine; while the essential amino acids are arginine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine.

Metabolites and amino acids extracted from the blood serum by DILGOM study are described in the following points; for each compound the principal biological functions and its correlation with cardiovascular disease are outlined.

- **Alanine (ALA)** Alanine is a non essential amino acid and a hydrophobic molecule. It is produced within the body from the conversion of the carbohydrate pyruvate or the breakdown of DNA. Alanine is highly concentrated in muscles and is one of the most important amino acids released by muscles, functioning as a major energy source. It is also an important participant as well as regulator in glucose metabolism and has an important role in lymphocyte reproduction and immunity. It behaves as an inhibitory neurotransmitter in the brain. Alanine is highly concentrated in meat products and other high-protein types of food like wheat germ and cottage cheese.
  It has been shown [40] that alanine levels are related with blood sugar levels in both diabetes and hypoglycemia, and alanine reduces both severe hypoglycemia and ketosis of diabetes.

- **Albumin (ALB)** Albumin is a protein made by the liver and it is the most prevalent protein of blood plasma. One of albumin's principal functions is to support the oncotic pressure, which aids in keeping blood within the circulation. Albumine is also an important circulating antioxidant and it has enzymatic properties. It serves as carrier for molecules of low water solubility isolating their hydrophobic nature, including lipid-soluble hormones, bile salts, free fatty acids (apoprotein), calcium, ions (transferrin), and some drugs.
  Kuller et al [41] found a highly significant inverse relation between serum albumin level and risk of coronary heart disease. Lower albumin levels may be a marker of persistent injuries to arteries and progression of atherosclerosis and thrombosis.

- **Acetate (ACE)** Acetic acid is one of the simplest carboxylic acids and it is produced by certain bacteria. The acetyl group, derived from acetic acid, is fundamental to the biochemistry of virtually all forms of life. Acetic acid perform an important role in the metabolism of carbohydrates and fats.

- **Acetoacetate (ACACE)** Acetoacetate is a weak organic acid, it is produced in the human liver when there is an excessive fatty acid breakdown. This explains its presence under a prolonged physical exertion or during starvation. Acetoacetate is excreted either in urine or through respiration. It provides acetoacetyl-CoA and acetyl-CoA for synthesis of cholesterol, fatty acids, and complex lipids.
  The correlation between acetoacetate and diabetes mellitus type 2 was demonstrated several times [42], Ketone bodies, as acetoacetate and acetone, are released into the blood from the liver when hepatic lipid metabolism has changed to a state of increased ketogenesis. A relative or absolute insulin deficiency is present in all cases.

- **Citrate (CIT)** Citrate is a weak acid that can be produced by human cells or introduced with diet. It takes part in citric acid cycle, a series of chemical reactions that generate energy from carbohydrates, fats and proteins. Citric acid is found in citrus fruits, most concentrated in lemons and limes.
  The evaluation of plasma citric acid is scarcely used in the diagnosis of human diseases.

- **Creatinine (CREA)** Creatinine is a breakdown product of creatine phosphate in muscles, kidneys, liver and pancreas. The loss of a water molecule from creatine results in the formation of creatinine, which is transferred to the kidneys by blood plasma, whereupon it is eliminated from the body. Creatinine is usually produced at a fairly constant rate by the body.

Serum creatinine test is the most commonly used indicator of renal function. A rise in blood creatinine levels is thus related to renal failure, but it was also demonstrated [43] that serum creatinine value, obtained in normotensive, nonobese, normoglycemic survivors of a myocardial infarction without preexistent renal disease or heart failure, provides independent prognostic information regarding subsequent overall and atherosclerotic coronary heart disease mortality.

- **Glucose (GLC)** Glucose is a monosaccharide and it is the primary source of energy for living organisms. In animals glucose arises from the breakdown of glycogen; it is synthesized in the liver and the kidneys from non-carbohydrate intermediates, such as pyruvate and glycerol. Glucose can be broken down and converted into lipids or used for synthesize other important molecules such as vitamin C. It also supplies almost all the energy for the brain, so its availability influences psychological processes. Glucose is found in fruits and other parts of plants in its free state.
  It is obviously correlated with diabetes, a metabolic disorder caused by lack of insulin, that doesn't allow to use glucose as an energy resource. It was also found [44] that coronary-heart-disease mortality was approximately doubled for subjects with impaired glucose tolerance (IGT), defined as a blood-sugar above the 95th centile ($\geq$ 96 mg/dl).

- **Glutamine (GLN)** Glutamine is a non essential amminoacid. It is synthesized by the enzyme glutamine synthetase from glutamate and ammonia and the most relevant glutamine-producing tissue is the muscle mass. Glutamine is fundamental for protein synthesis; it is also a regulator for acid-base balance in the kidneys, and a energy source.
  Glutamine is found in food high in proteins, such as fish, red meat, beans, and dairy products. There is a significant body of evidence that links glutamine-enriched diets with intestinal effects; aiding maintenance of gut barrier function, intestinal cell proliferation and differentiation [45].

- **phenylalanine (PHE)** Phenylalanine is an essential amino acid and the precursor for the amino acid tyrosine. It is highly concentrated in the human brain and plasma.
  Phenylalanine is highly concentrated in high protein food, such as meat, cottage cheese and wheat germ. A new dietary source of phenylalanine is found in artificial sweeteners containing aspartame.
  Low phenylalanine diets have been prescribed for certain cancers with mixed re-

sults. Some tumors use more phenylalanine (particularly melatonin-producing tumors called melanoma). One strategy is to exclude this amino acid from the diet. Wannemacher et al [46] found a correlation between serum phenylalanine-tyrosine and myocardial infarction.

- **Isoleucine (ILE), Leucine (LEU), Valine (VAL)** Valine, isoleucine and leucine are Branched chain amino acids (BCAA), which are essential amino acids. These three amino acids are critical to human life and are particularly involved in stress, energy and muscle metabolism. Valine is used in carbohydrates metabolism, leucine in fats metabolism and isoleucine in both. BCAA, particularly leucine, stimulate protein synthesis and reduce protein breakdown. Furthermore, leucine can be an important source of calories. Leucine also stimulates insulin release, which in turn stimulates protein synthesis and inhibits protein breakdown.
  These amino acids exhibit different deficiency symptoms. Valine deficiency is marked by neurological defects in the brain, while isoleucine deficiency is marked by muscle tremors. Norrelund et al [47] found a significant correlation (P-value < 0.01) between the BCAA blood concentration comparing patients with chronic heart failure and healthy people.

- **Lactate (LAC)** Lactate is constantly produced in muscles from pyruvate, as a product obtained from glucose breakdown.
  Lange at al [48] found that plasma lactate concentrations exceeded the reference range in all the cases of mesenteric ischaemia ($n = 20$) and general bacterial peritonitis ($n = 15$). They concluded that a raised plasma lactate concentration is always a sign of an acute life-threatening condition, and usually indicates the need for an emergency operation. As a marker of mesenteric ischaemia its sensitivity was 100% and its specificity 42%, they suggest that a raised serum lactate concentration would be the best marker of mesenteric ischaemia to date.

- **Pyruvate (PYR)** Pyruvic acid is an intermediate compound in the metabolism of carbohydrates, proteins, and fats. Pyruvic acid can be made from glucose through glycolysis, converted back to carbohydrates (such as glucose) via gluconeogenesis, or to fatty acids through acetyl-CoA. It can also be used to construct the amino acid alanine and be converted into ethanol. Pyruvic acid supplies energy to living cells through the citric acid cycle when oxygen is present (aerobic respiration), and alternatively ferments to produce lactate when oxygen is lacking (fermentation).

- **Tyrosine (TYR)** Tyrosine is a non-essential amino acid, which can be synthesized within the body from phenylalanine. It readily passes the blood-brain barrier and, once in the brain, it is a precursor for the neurotransmitters dopamine, norepinephrine and adrenalin. Tyrosine is also the precursor for hormones, thyroid and the major human pigment, melanin. It is not found in large concentrations throughout the body, probably because it is rapidly metabolized.
  It is found in many high-protein food products such as chicken, turkey, fish, milk, yogurt, cheese, seeds.
  Some adults develop elevated tyrosine levels in their blood. This indicates a need for more vitamin C. More tyrosine is needed under stress, and tyrosine supplements can cure biochemical depression. Norrelund et al [47] found a significant increase of tyrosine blood concentration (P-value < 0.01) between patients with chronic heart failure and healthy people. Wannemacher et al [?] found a relation between phenylalanine-tyrosine ratio and myocardial infarction.

- **UREA** Urea is a highly soluble organic compound formed in the liver by the urea cycle. It is the principal end product of protein catabolism; it has no physiological function and constitutes about one half of the total urinary solids. It is dissolved in blood (in humans in a concentration of 2.5 - 7.5 mmol/liter) and excreted by the kidney in the urine.

- **Histidine (HIS)** Histidine is an essential amino acid, its has anti-oxidant, anti-inflammatory and anti-secretory properties, it is also the precursor of neurotransmitter histamine. Histidine increases histamine in the blood and probably in the brain.
  Elevated blood histidine is accompanied by a wide range of symptoms, from mental and physical retardation to poor intellectual functioning, emotional instability, tremor, ataxia and psychosis.
  Serum histidine levels are lower and are negatively associated with inflammation and oxidative stress in obese women. Histidine supplementation has been shown to reduce insulin resistance, reduce BMI and fat mass and suppress inflammation and oxidative stress in obese women with metabolic syndrome.

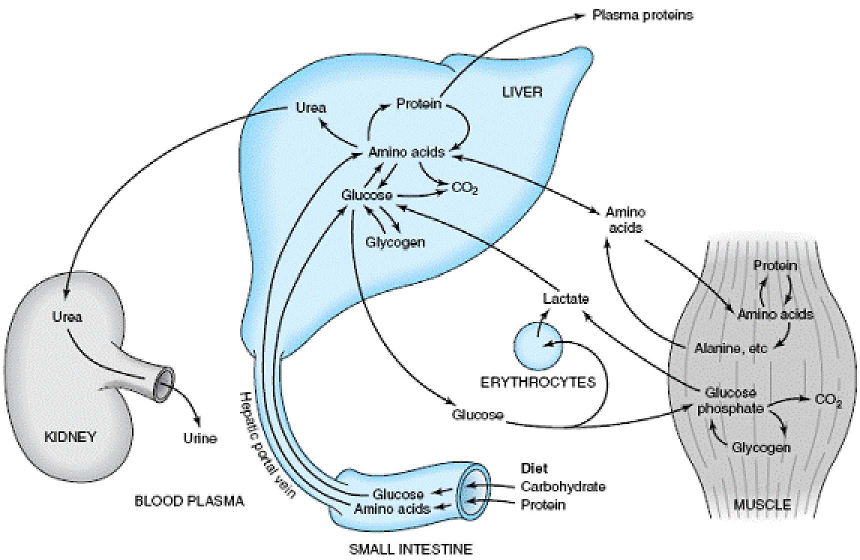


Figure 2.6: Transport and fate of major carbohydrate and amino acid substrates and metabolites.

2.5 DILGOM dataset

The Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study collected sample concerning metabolomics, transcriptomics and genomics of a finish cohort in order to investigate possible relations between lipids, immune cells in circulation and cardiovascular diseases.

Study participants were aged 25–74 years and were drawn from the Helsinki/Vantaa area of southern Finland. In order to avoid aberrant values of blood serum compounds, participants were asked to fast overnight for a period of at least 10 h before giving a blood sample. The extraction of serum from blood samples was performed by centrifugation, this operation allow to remove clotting factors. DNA and RNA were then extracted, identified and quantised using microarray technique. Here we will describe the two utilized dataset, which regard phenotypes and serum compounds.

2.5.1 Phenotypes dataset

The word phenotype indicates observable physical or biochemical characteristics of an organism, as determined by both genetic and environmental influences. Phenotypes are therefore directly linked to genetic, but also to nutrition, diet, lifestyle and psychosocial factors. For this reason they play a fundamental role in the DILGOM study which, starting from the typical obesity phenotypes, aims to identify connections between some 'omics', that are metabolomics, transcriptomics and genomics. The phenotypes listed by the DILGOM study are thus related to obesity, as the BMI, WHR (waist-hip ratio) and waist circumference. These measures are commonly used to characterized a person as normal-weight, over-weight or obese, as discussed in section 2.3. The phenotypes dataset includes also some general informations of the Finnish cohort, such as: age, gender, systolic and diastolic pressure, total cholesterol and HDL cholesterol. These features are relevant for the evaluation of the cardiovascular diseases risk. Specifically, total cholesterol, age and systolic pressure are directly related to the increase of cardiovascular disease risk, in contrast, high HDL levels lower that risk.

In the DILGOM phenotypes dataset there are also some variables which permit to identify people who already present cardiovascular or, more generally, health problems. Effectively, some individuals of the Finnish cohort are affected by insulin resistance or diabetes. These people have two to four times more likely to develop cardiovascular disease than people without diabetes, as reported by the world health federation. Looking at the diseasome in figure 1.3, it can be seen that diabetes shares disease genes with obesity; in fact, both these diseases are risk factors for cardiovascular diseases. Cardiovascular diseases are, indeed, the leading cause of mortality for people with diabetes. In the DILGOM cohort there are also people with high values of fasting glucose, which is commonly considered a pre-diabetic state, and which is associated with insulin resistance and increased risk of cardiovascular pathology. Other people are under cholesterol treatment, their risk to develop CVDs is thus lowered by medicines, which modify their LDL cholesterol level. Some individuals of the Finnish cohort take blood pressure medications, which modify systolic and diastolic pressure values.

The DILGOM study suggests to select a subset of individuals that haven't aberrant observations. Specifically, as reported in the file *DILGOMdate. Update for Case Study*, they identifyed some individuals with 'aberrant' variables values in the phenotypes dataset; that are diabetics, individuals under cholesterol treatment and individuals with fasting glucose anomalies. In table 2.2 we report the variable nouns and the used thresholds. Individuals with variables values higher then the corresponding thresholds are removed from the list, in order to harmonize the use of the DILGOM datasets. The DILGOM study states as discretionary the correction for blood pressure medication, this medication does not condition the values of blood components but only the corresponding phenotypes of systolic and diastolic pressure.

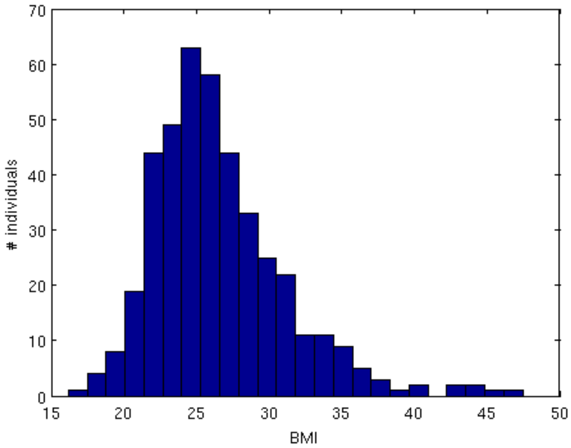


Figure 2.7: Hisogram of the BMI of the individuals analysed in the DILGOM study.

variable noun	meaning	threshold
FR07_38	diagnosed diabetes	FR07_38 >1
SL_GLUK_0H	fasting glucose	SL_GLUK_0H>10
K34	cholesterol medication	K34=1

Table 2.2: Variables of the phenotypes dataset which indicates some diseases.

Phenotypes correlation

In order to realize an initial evaluation of the phenotypes dataset, we selected a subset of 11 features. Some of them are notoriously correlated to health as BMI, WHR, waist circumference, systolic and diastolic pressure, total and HDL cholesterol and fasting glucose; others help to classified individuals, as age and gender, and others show if people use blood pressure medications.

We calculated the correlation between 10 variables for female and male individuals using the Kendall's τ. As expected, some features have high correlations, as WHR, BMI and waist circumference and systolic and diastolic pressure, less relevant correlations are those linked to total cholesterol. HDL cholesterol displays only inverse correlations, in agreement with its definition of "good cholesterol", the positive relation with total cholesterol is due to the fact that total cholesterol is the sum of bad (LDL) and good (HDL) cholesterol. Also age seems to be related to WHR, but with a higher value for male individuals. The correlation values are reported in tables 2.3 and 2.4 for female and male individuals respectively. This subset of phenotypes was used to classify obese and normal-weight individuals, in order to perform a multilayer analysis with two

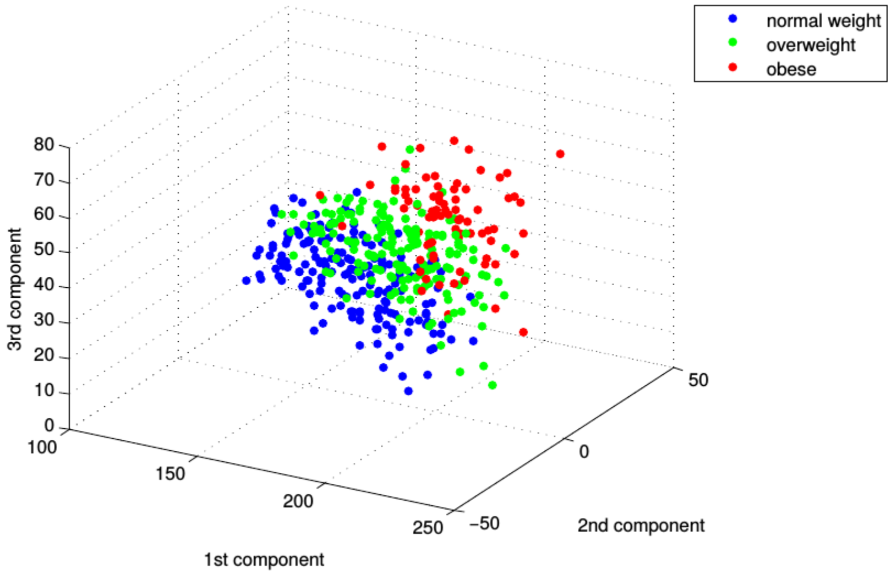


Figure 2.8: Scatter plot where every point represents an individual, coordinates are obtained using PCA on the 10 features selected. BMI is used for the classification, as reported in table 2.1.

layers, one representing normal-weight people and one for the obese ones. In figure 2.8 is displayed the correlation between the selected phenotypes subset and the BMI: individuals are represented as points, coordinates are obtained using the Principal Component Analysis (PCA) on the phenotypes subset except BMI, used for classify individuals in normal-weight (blue), over-weight (green) and obese (red). It can be qualitatively seen that the subset permits a quite good classification of individuals. The methods tested to perform a good classification of individuals will be analysed in the chapter 4 and in the appendix A.

2.5.2 Blood serum particles dataset

The DILGOM study aims to characterize the serum metabolites relationships linked to human diseases. The ultimate objective is to understand the potentially causative and reactive relationships between serum metabolites and geonomic and transciptomic networks. To this end, the DILGOM study extracts the serum metabonomes of 518 individuals from a population-based cohort using proton nuclear magnetic resonance spectroscopy (NMR). These serum compounds are divided in: lipoproteins, serum lipid extracts, amino-acids and metabolites. The full list of these compounds is reported in appendix B.

				Female				
	tot chol	HDL	fasting gluc	syst press	diastPress	BMI	waist circ	whr
age	0.19	-0.01	0.18	0.33	0.19	0.23	0.24	0.17
tot chol		0.21	0.07	0.08	0.12	0.13	0.13	0.05
HDL			-0.04	-0.06	-0.04	-0.18	-0.20	-0.21
fasting gluc				0.18	0.14	0.28	0.28	0.21
sys pres					0.47	0.22	0.22	0.19
dias press						0.22	0.21	0.22
BMI							0.75	0.38
waist circ								0.56

Table 2.3: Kendall correlation of some selected phenotypes used for classification. Upper triangular matrix refers to female correlations.

				Male				
	tot chol	HDL	fasting gluc	syst press	diastPress	BMI	waist circ	whr
age	0.05	-0.02	0.18	0.20	0.07	0.14	0.29	0.35
tot chol		0.15	0.02	0.08	0.15	0.10	0.11	0.09
HDL			-0.12	0.00	-0.05	-0.20	-0.20	-0.17
fasting gluc				0.09	0.12	0.22	0.28	0.27
sys pres					0.36	0.15	0.20	0.19
dias press						0.16	0.18	0.16
BMI							0.68	0.43
waist circ								0.66

Table 2.4: Kendall correlation of some selected phenotypes used for classification. Upper triangular matrix refers to male correlations.

Lipoproteins are classified according to their density: a high fat to protein ratio causes large and low dense lipoproteins. The acronym of these sub-group are: VLDL (Very Low Density Lipoprotein), IDL (intermediate-density lipoprotein), LDL (Low Density Lipoprotein) and HDL (High Density Lipoprotein). Each density-based class of lipoproteins is also divided in subclasses based on diameter differences of the particle, as reported in table 2.5. For each lipoproteins subclasses the DILGOM study extracted values relating to: total cholesterol (-C), cholesterol esters (- CE), free cholesterol (-FC), total lipids (-L), phospholipids (-PL), triglycerides (-TG), concentration (-P) and average diameter (-D) of the particles. An example of the used acronym for VLDL particles is reported in table 2.6.

Amino-acids are the building blocks of proteins in the body, but they can also

lipoprotein	sub-class acronym	name	average particles diameter (nm)
VLDL	XXL-VLDL	chylomicrons and extremely large VLDL	least 75 nm
	XL-VLDL	very large VLDL	64.0
	L-VLDL	large VLDL	53.6
	M-VLDL	medium VLDL	44.5
	S-VLDL	small VLDL	36.8
	XS-VLDL	very small VLDL	31.3
IDL			28.6
LDL	L-LDL	large LDL	25.5
	M-LDL	medium LDL	23.0
	S-LDL	small LDL	18.7
HDL	XL-HDL	very large HDL	14.3
	L-HDL	large HDL	12.1
	M-HDL	medium HDL	10.9
	S-HDL	small HDL	8.7

Table 2.5: Lipoproteins classification and respective average diameter.

be found in plasma as free amino acids, according to their role in neurotransmitter functioning, cholesterol and carbohydrate metabolism, and detoxification processes. The amino acids extracted by the DILGOM project are: alanine (ALA), glutamine (GLN), glycine (GLY), histidine (HIS), isoleucine (ILE), leucine (LEU), phenylalanine (PHE), tyrosine (TYR), valine (VAL).

Metabolites and are the intermediate products of metabolic reactions catalyzed by various enzymes that naturally occur within cells. The blood amino acids quantified by the DILGOM study are: 3-hydroxybutyrate (BOHBUT), acetate (ACE), acetoacetate (ACACE), citrate (CIT), creatinine (CREA), Lactate (LAC), pyruvate (PYR), glucose (GLC), glycerol (GLOL), glycoprotein acetyls, mainly a1-acid glycoprotein (Gp). Others serum compound listed are albumin (ALB), CH2 groups of mobile lipids (MobCH2), CH3 groups of mobile lipids (MobCH2), double bond protons of mobile lipids (MobCH), urea (UREA).

Serum lipids are lipids in the blood, either free or bound to other molecules. They are mainly fatty acids, the fatty acids values extracted by DILGOM project are: n-3 fatty acids (FAw3), n-6 fatty acids (FAw6), n-7 n-9 and saturated fatty acids (FAw79S), total fatty acids (TotFA), Monounsaturated fatty acids; 16:1, 18:1 (MUFA). Other serum lipid measures concern the serum cholesterol, as Esterified cholesterol (Est-C)

acronym	meaning	unit
L-VLDL-C	Total cholesterol in large VLDL	$(mmol/L)$
L-VLDL-FC	Free cholesterol in large VLDL	$(mmol/L)$
L-VLDL-PL	Phospholipids in large VLDL	$(mmol/L)$
L-VLDL-TG	Triglycerides in large VLDL	$(mmol/L)$
L-VLDL-CE	Cholesterol esters in large VLDL	$(mmol/L)$
L-VLDL-L	Total lipids in large VLDL	$(mmol/L)$
L-VLDL-P	Concentration of large VLDL particles	(mol/L)
VLDL-D	Mean diameter for VLDL particles	(nm)

Table 2.6: Values extracted from the VLDL particles.

and Free cholesterol (Free-C).

These listed compounds were used as nodes to build a multiplex network, where one layer represents the correlation between them in normal- weight people, and another layer shows the correlations between serum compound for obese individuals.

Chapter 3

Method

In this chapter we will present the method which were implemented for the samples classification and the graph analysis. This method requires multi-omic data, since it reveals if there are significant differences in one omic (B), based on a classification performed by another omic (A). Namely it says, if two or more omic are related to each other.

Differences inside the omic (B) are detected using complex networks analysis, which is the core of this thesis. In particular we build a multiplex network (see section 1.3.1) in order to stress the differences inside the omic B between two groups classified using the omic A. The implemented method is totally independent of the data to be analysed, since it is a purely physical, mathematical and statistical approach. This makes it possible to extend this procedure to the most disparate fields. This method can be particularly useful to reveal significant differences between two apparently similar groups, as we will see. During this chapter we will often refer to the biological case of study, which is presented in chapter 2, in order to clarify the procedure.

The implemented method can be divided into two main steps. The first step is related to the classification of two groups from the whole dataset. Looking at our case of study, this step regards the extraction of obese individuals, that have an increased CVD risk, and normal-weight individuals, and it is therefore based on the phenotype dataset (omic A), which is described in section 2.5.1). The goal of this initial processing is to obtain two sub-groups which have different characteristics in one omic (A). We opt for two different methodology of classification: one based on the thresholds suggested by the WHO (section 2.3, table 2.1) and one on *linear models* (lm) (appendix A). These approaches lead to a division into three classes, which are linked to obese, over-weight and normal-weight people. The two groups associated to obese and normal-weight people were used for the set up of the multiplex network.

The second and principal step is focused on the setting-up of the multiplex network and its analysis. The used dataset is that related to the concentrations of blood compounds (omic B), relative to each individual, which is described in section 2.5.2). We built a weighted multiplex network made up of two layers, which have the same number of nodes, but different edges between nodes. In our case, nodes of these layers represent compounds, whereas links between them are their correlations. The first layer concerns relations of compounds of the obese group and the second layer the relations between nodes of the normal-weight group. The adjacent matrices of the two layers (section 1.1) are therefore the correlation matrices of compounds of the two groups. Once set up the two graphs, we evaluated differences and similarities between them in order to highlight some diverse behaviours of compounds. This analysis evaluates both single nodes, (using the *hypergeometric test*), and communities behaviours (section 1.2).

Looking at our case of study, this method highlights compounds (nodes) which have different intra-layer connections, these compounds are interesting from a biological point of view since they could be related to some dysfunctions which cause obesity, and which increase the cardiovascular disease risk.

3.1 Classification

We use two methods for the classification of the individuals. Both methods use some features of phenotypes dataset (omic A) which are linked to CVD risk and which are described in section 2.5.1. These two approaches focus on different principles to classify people.

The first classification method is based on the BMI and whr (*waist hip ratio*) thresholds suggested by the WHO. These thresholds are listed in table 2.1) in section 2.3. We focus our analysis on this classification method, but we present also another method which uses the *linear model.*

The *linear model* (lm) classifies individuals into two groups, one is associated to obese individuals and the second one to non-obese ones. Obese and normal-weight individuals are established looking at their residuals. In our analysis, a residual of an individual, is defined as the difference between his actual BMI and the value of his BMI estimated by the lm. This method is presented in the appendix A.

3.2 Multiplex analysis

In this section we will explain the method which is utilized for the setting-up of the multiplex network and its study. As already said, this is the core of our analysis and it is completely independent from the data to be analysed. We will refer to our case of study to clarify the process.

In our metabolic study this analysis is based on the compounds dataset, which includes the concentrations of blood compounds of each participant of the DILGOM study. An accurate description of DILGOM compounds dataset is illustrated in section 2.5.2.

We built a weighted multiplex network with two layers, based on the classification performed with another omic (A), in order to evaluate if the the two groups maintain differences also in the current omic (B). In this instance we want to evaluate if the obese group and the the normal-weight group have differences also from a metabolic point of view. These layers are constituted by the same number of nodes (n), which represent the compounds extracted from the blood of each individual by the DILGOM study. An edge between two compounds specifies a positive correlations between them, and its weight is a measure of the correlation strength. The adjacency matrix (sec. 1.1) of each layer is obtained applying the CLR algorithm on the correlation matrix of compounds; this step will be described in section 3.2.2.

The assessment of the differences between the two layers is achieved looking at the dissimilarities of the single node edges and of community structure. The implemented procedure is schematically illustrated in the block-diagram in figure 3.1. A brief explanation of each step is reported in the following:

1. The starting point of our multiplex analysis consists of two 'sub-datasets' whose difference are to be evaluated. Therefore these 'sub-datasets' must have the same features to be compared. In our application the two sub-datasets contain information about 107 compounds, one is related to the concentration of these compounds in obese individuals, while the second sub-dataset is relative to the concentration of the same compounds in normal-weight individuals. The two sub-dataset are composed by a different number of individuals, but the compounds (n) are the same for the both groups.

2. for each 'sub-dataset' the correlation between features (that are compounds in our example) is computed. We choose *Kendall's* τ correlation (sec 3.2.1) since it is a non-parametric measure of relation between two variables. This step leads to a $n \times n$ symmetric matrix (τ) with values ranging from -1 to 1, where a value

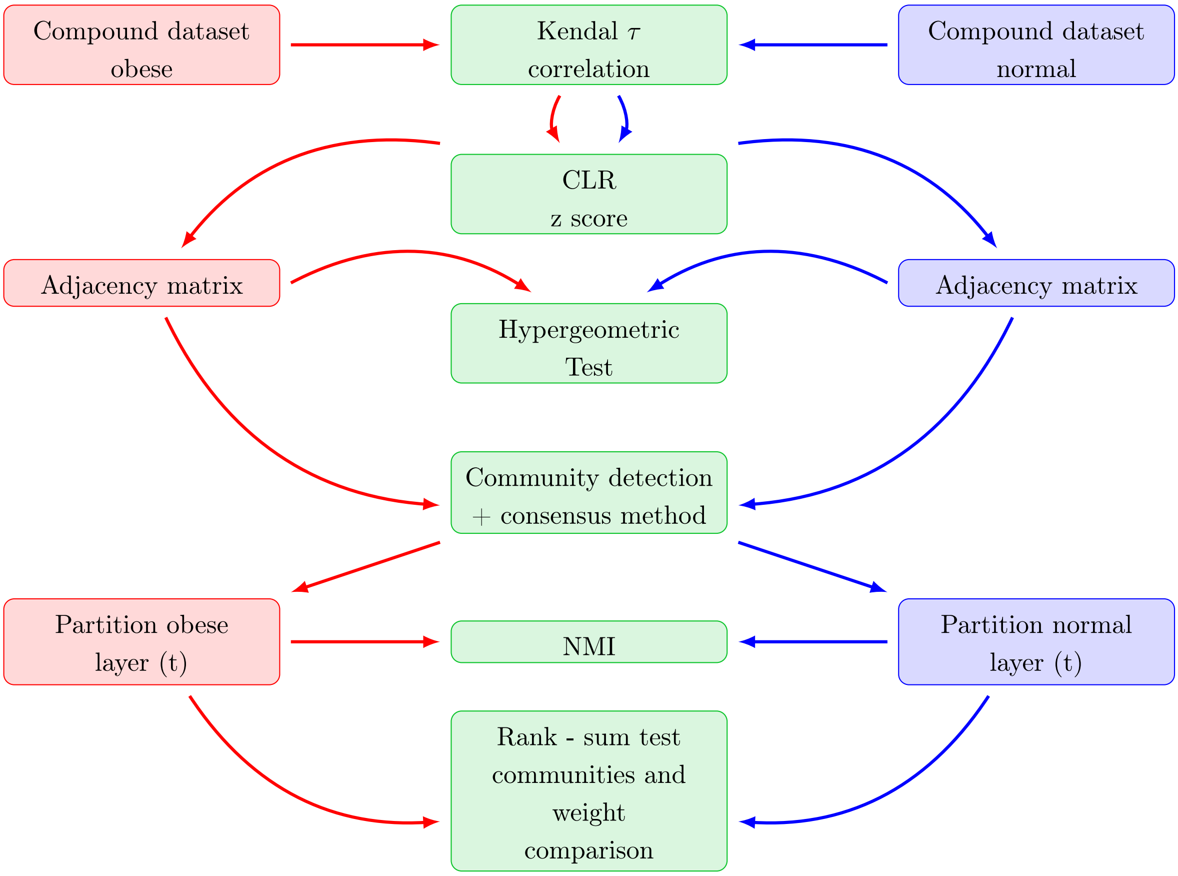

Compound dataset obese
Kendal τ correlation
Compound dataset normal
CLR z score
Adjacency matrix
Hypergeometric Test
Adjacency matrix
Community detection + consensus method
Partition obese layer (t)
NMI
Partition normal layer (t)
Rank - sum test communities and weight comparison

Figure 3.1: Scheme of the implemented method.

equal to ±1 means a perfect correlation or anticorrelation, while a correlation coefficient value towards 0, means a weak relationship. In our example, the τ_{ij} coefficient of the τ matrix indicates the Kendall's correlation between compound i and compound j, the diagonal of τ matrix is set to zero.

3. Once obtained the correlation matrix, we apply the *Context likelihood ratio* (CLR), a method used to enhance meaningful correlations. The CLR algorithm builds a z-score matrix where the contrast between the physical interactions and their indirect relationships is enhanced [49]. CLR algorithm considers only positive correlations between compounds. This method is explained in greater detail in section 3.2.2). The $n \times n$ z-score matrix resulting from the CLR is set as the adjacency matrix (section 1.1) of the considered layer, the diagonal is set to zero since self loops have no sense for this analysis.

As already said, the nodes of the two layers are the same, therefore differences between the two layers lie in their different edges weights. In other words, the same node can be linked to different nodes and with different edge weight in the two layers. Consequently, we can deduce that the feature associated to that compound has different behaviours among the two groups. In order to highlight differences in the nodes behaviour, we evaluate both single nodes and communities.

4. The evaluation of the single node connections is performed using the *hypergeometric test* (section 3.2.3). To carry out this test it is necessary to make the weighted adjacency matrices becoming topological matrices. We fix a threshold $\tilde{z}$ on the z-score values: coefficients of the adjacency matrices with z-score greater than or equal to the threshold ($z_{ij} \geq \tilde{z}$) are set to one, others ($z_{ij} < \tilde{z}$) to zero. For each node i (i.e. the $i-$th row of the $n \times n$ topological matrices), edges are divided into three groups: edges belonging only to layer 1 ($L_{10}(i)$), edges belonging only to layer 2 ($L_{01}(i)$) and edges belonging to both ($L_{11}(i)$). The hypergeometric test uses the hypergeometric distribution to assess whether the observations are statistically significant. Specifically the test assesses whether there is a significant enrichment or a depletion in the number of links of a node (eg. $k_{01}(i)$) lying only in one layer beyond what might be expected by chance, which is established looking at the total L_{01} and L_{11} links and $k_{11}(i)$ links.

5. The evaluation of the differences between the communities of nodes is performed comparing edges in the same communities in the two layers.

    - For each layer we carry out a community detection (section 1.2.1) in order to join subgroups of nodes which are strongly bonded together, using the stability as quality function. Algorithms based on stability, unlike those based on modularity, allow to evaluate the goodness of clusters through different Markov times, which are different partition scales. The evolution of the communities among time is computed separately for the two layers.

    - We apply the consensus clustering at each Markov time. This is a data analysis method used to generate stable results out of a set of partitions delivered by stochastic methods [50] (section 3.2.5). Therefore this method shall enhance the stability and the accuracy of the communities detection algorithm.

    - We evaluate the evolution of the communities of each layer with the resolution parameter represented by the discrete Markov time. As the Markov time increases, the number of communities usually decreases since a stable partition is obtained. The stability of the partition can be evaluated considering the stability value (eq 3.8) at each Markov time, but also looking at the duration of that partition along the Markov chain. These two aspects are taken into account in order to choose the best nodes partition.

- If the same communities are detected in both layers, we evaluate differences between same communities in different layers. In this case, to select the best partition we consider also the communities overlap. The overlap is computed using the *normalised mutual information* (NMI), this quantity ranges between zero and one, high NMI values means a good overlap. The complete explanation of the NMI is reported in section 3.2.6.
- Once selected the Markov time and consequently the nodes belonging to each community, we compare the weights of the links belonging to the same community, but in different layers. Considering a community made up of k links ($k < n$), we extract the corresponding $k \times k$ adjacency matrix, which represents only the connections inside that community.

  To estimate the significance of the differences between the community edges of the two layers we perform the Wilcoxon rank sum test (section 3.2.7). This is a non-parametric test which is used to verify if two statistical independent samples come from the same population. We consider also the distribution of weights in the same community in different layers. Moreover, we evaluate the robustness of these differences comparing the results obtained for the two layers with a distribution of those values which is computed extracting individuals in a random way.

This procedure allows to evaluate both single node differences and communities differences between the two layers. In the following sections we will described in great depth each step listed in this brief explanation.

3.2.1 Kendall's τ

Kendall's rank correlation [51] [52] is a non-parametric measure of dependence between two variables. It is based on the number of concordances and discordances in paired observations, therefore the two samples, X and Y, have to be of the same size n. The total number of possible pairings observations of X with Y is $n(n-1)/2$.

Two observations (x_i and y_i) and (x_j and y_j) are concordant if they are in the same order with respect to each variable. That is, if $x_i < x_j$ and $y_i < y_j$, or if $x_i > x_j$ and $y_i > y_j$. They are discordant if they are in the reverse ordering for X and Y, and they are tied if $x_i = x_j$ and/or $y_i = y_j$. Defined S as the difference between the total number of concordant(n_c) and discordant n_d pairs ($S = n_c - n_d$), Tau (τ) is related to S by:

$$\tau = \frac{S}{n(n-1)/2}$$

If there are tied (same values) observations then τ_b is used:

$$\tau_b = \frac{S}{\sqrt{\left[n(n-1)/2 - \sum_{i=1}^{t} t_i\left(t_i - 1\right)/2\right]\left[n(n-1)/2 - \sum_{i=1}^{u} u_i\left(u_i - 1\right)/2\right]}}$$

where t_i is the number of observations tied at a particular rank of X and u_i is the number tied at a given rank.

We compute the correlation between the n features for each 'sub-dataset': this lead to a $n \times n$ symmetric correlation matrix (τ), where each entry τ_{ij} indicates the Kendall's correlation between the features i and j. Going back to our biological application, we compute two correlation matrices, one for the 'sub-dataset' of obese individuals and another one for normal-weight group. These matrices represent the correlation of the n compounds in the two groups.

A network with an adjacency matrix equal to a correlation matrix would be with high probability completely connected, since the chance of having a correlation exactly equal to zero is very small. Moreover it would have both positive and negative links. It is possible to remove weak correlations fixing a threshold on the correlation values; an example is displayed in Fig 3.3.a) and 3.3.b). Figure 3.3.a) shows the kendall's τ correlation matrix of the obese layer, colours indicate the strength of the correlation between two compounds: blue represents high negative correlations whereas red means high positive ones. From this image we can qualitatively deduce that, in our biological application, negative correlations are generally weaker then positive ones, since the highest negative correlation is -0.5. The application of a threshold on the correlation matrix of figure 3.3.a) is illustrated in figure 3.3.b). Many relations are set to zero, and three main blocks are identified. The first 10 nodes are completely isolated using a threshold on the Kendall's correlation matrix. Therefore they would be not considered by our analysis if we build a network with this threshold correlation matrix as adjacency matrix. The first disadvantage of the correlation matrix is that it considers also indirect relationships: if the nodes a and b are highly related to the node c, the correlation between a and b will be high, even if they are not directly related. We fix these problems applying the CLR algorithm (section 3.2.2) to the correlation matrix.

3.2.2 Context likelihood ratio (CLR)

The *Context likelihood ratio* (CLR) is a method introduced by Gardner et al [49]. Starting from a correlation matrix C, Gardner et al used a mutual information matrix, the CLR algorithm builds a z-value matrix where the contrast between the physical interactions and their indirect relationships is enhanced. Considering an element i, which can be a gene or, in our case, a compound, the CLR algorithm compares the correlation value (c_{ij}) between that element i and another element j with their background distribution. The background distribution (C_i) of the i element is gather by all the correlation values of i, therefore it corresponds to a row or column of the correlation matrix C. In the same way it is constructed the background distribution of the j element. Most correlation values of C are caused by the random background, (e.g., due to indirect network relationships). The random background is approximate as a joint normal distribution, considering the two background distributions C_i and C_j as independent variables. A schematic representation of the CLR algorithm is symbolised in figure 3.2.
The comparison between the correlation value c_{ij} and its two marginal background distributions gives two z-score values:

$$z_i = \max\left(0, \frac{c_{ij} - \mu_i}{\sigma_i}\right) \qquad z_j = \max\left(0, \frac{c_{ij} - \mu_j}{\sigma_j}\right) \tag{3.1}$$

where μ_i and μ_j are the means of C_i and C_j which are defines as: $\mu_i = 1/N(\sum_i C_i)$ for a random variable vector made up of N scalar observations. The standard deviations σ_i and σ_j are defined as $\sigma_i = \left(1/(N-1)\sum_i |C_i - \mu_i|^2\right)^{1/2}$.
The elements of the CLR matrix are calculated as

$$f(z_i, z_j) = \sqrt{z_i^2 + z_j^2} \tag{3.2}$$

where $f(z_i, z_j)$ is the joint likelihood measure. Gardner et al applied the CLR algorithm to transcriptional profiles of E.Coli, concerned diverse set of conditions, in order to determine transcriptional regulatory interactions. They showed that, comparing several different network inference methods, CLR was the top-performing method for their case of study.

We apply the CLR method on metabolic data [53] in order to enhance the contrast between the compounds interactions and their indirect relationships. This allows to increase the significance of some quite weak correlations with respect to their background. An example is displayed in figure 3.3: figure 3.3 c) represents the z-score matrix resulting from the application of the CLR algorithm on the Kendall's correlation matrix in figure 3.3.a). We fix a threshold on the z-scores of the CLR matrix in

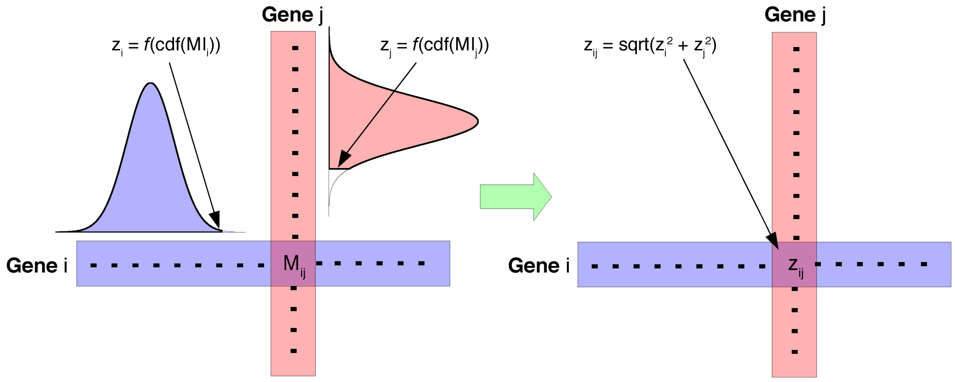


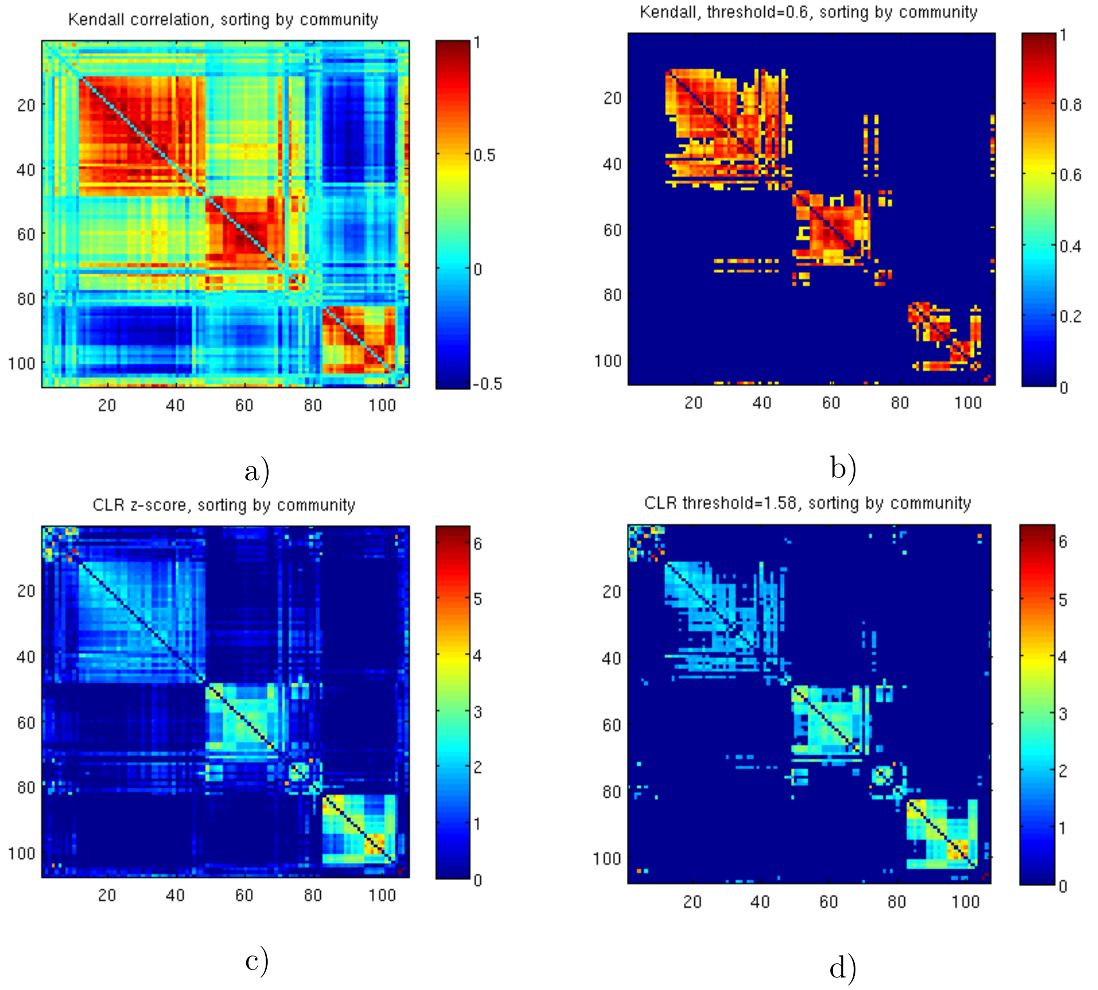
Figure 3.2: Graphic representation of the CLR algorithm. The left image symbolises the starting correlation matrix ($\tau - kendall$ in our case and MI in the article of by Gardner et al [49]). The right image symbolises the final CLR matrix.

figure 3.3.c, the resultant matrix is displayed in figure 3.3.d. This matrix shows the presence of four main blocks. Therefore there is a significant difference between the two thresholded matrices (figures 3.3,b) and 3.3.d)). While the Kendall thresholded matrix causes the disconnection of approximately 20 nodes and the presence of three blocks, the second one builds a completely connected graph with four blocks. It should be noted that the two thresholded matrices have about the same number of non-zero entries: the the Kendall thresholded matrix in figure 3.3b) has 1686 values different to zero, while the thresholded z-score matrix (fig 3.3d)) has 1680 non-zero entries.

These images permit to see that this method allows to enhance the contrast between the physical compounds interactions and their indirect relationships. However, the implementation of the CLR method causes a loss in information, since negative correlations are set to zero. This can be seen confronting images 3.3 a) and 3.3 b): blue areas in figure a) are linked to negative correlations, while in figure b) they represent null correlations. Since the CLR algorithm builds a z-score positive matrix, correlations can range from 0 to infinite.

The CLR matrix gives a completely connected network where indirect relationships are lowered. We choose to use the z-score matrices obtained with the CLR algorithm as adjacency matrices, therefore we consider only positive correlations between nodes.

Looking at our case of study, the CLR allows to evaluate all compounds extracted from the blood serum, which are usually characterised by weak correlations, as all metabolic elements. In fact, the majority of Kendall's correlations ranges between 0 and 0.4. This is one of the most problematic aspects of the metabolic studies, which require methods to enhance the most significant of these weak correlations.



a)

b)

c)

d)

Figure 3.3: Correlation matrices of: a) Kendall's correlation, b)Kendall's correlation with threshold at 0.6, c) CLR z-values matrix, d) CLR z-score matrix with a threshold at z=2.0.

3.2.3 Hypergeometric test

The hypergeopetric test is performed to evaluate the most significant differences between single nodes of the two layers. This test is based on the hypergeometric distribution.

The hypergeometric distribution [54] describes the experiment where elements are picked randomly without replacement . The initial condition is that there are N elements out of which M have a certain attribute (and $N - M$ have not). Randomly picking n elements without replacement from the whole set N, we can compute $p(m)$, which is the probability that exactly m of the selected elements (n) come from the group with the attribute (M).

The Hypergeometric distribution is given by:

$$p(m; n, N, M) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \tag{3.3}$$

where the discrete variable m has limits from $max(0, n - N + M)$ to $min(n, M)$. The parameters n ($1 \leq n \leq N$), N ($N \geq 1$) and M ($M \geq 1$) are all integers.

Hypergeometric test

The hypergeometric test [55] uses the hypergeometric distribution to assess whether the observations (n) are statistically significant; that is whether there is a significant enrichment or a depletion in the number of drown elements with a certain attribute (m) beyond what might be expected by chance.
We consider only a test for over-representation of successes in the sample; in this case the hypergeometric p-value indicates the probability of randomly drawing m or more elements with a certain attribute from the population N in n total draws. In a test for under-representation, the p-value is the probability of randomly drawing m or fewer elements. The p-value for over-representation of discrete variables is then:

$$P(at\,least\,m; N, M, n) = 1 - \sum_{i=0}^{m-1} p(i; N, M, n) \tag{3.4}$$

We set the significance level at 0.05, that is we reject the null hypothesis if the probability of obtaining a result equal to or 'more extreme' than what was actually observed is less than 5%.

We perform the hypergeopetric test to evaluate the most significant differences between nodes connections of two layers. This test evaluates significant enrichment or a depletion in the number of 'drawn elements' which are, in our case, the connections of a node.

In order to have topological matrices which permit to count the amount of links of each node, we impose a threshold on the weighted adjacency matrices. In fact, the matrices which came out from the CLR algorithm (3.2.2) are weighted. In particular, they are positive z-score matrices since negative correlations are set to zero by the algorithm. The entry $\tilde{a}_{ij}$ of a topological matrix is set to one if the z-value of the weighted adjacency matrix a_{ij} is greater than or equal to the fixed threshold value $\tilde{z}$. This process leads to two topological matrices $n \times n$, one relative to the layer 1 ($\tilde{A}^{(10)}$) and the other one to the layer 2 ($\tilde{A}^{(01)}$). Considering the *multilink* formalism explained in section 1.3.1, we classify edges of a node i in:

- edges lying only in layer 1 (i.e obese): $k_{10}(i) = \sum_{j=1}^{n}(a_{ij}^{(10)})$

- edges lying only in layer 2 (i.e normal): $k_{01}(i) = \sum_{j=1}^{n}(a_{ij}^{(01)})$
- edges lying in both layers: $k_{11}(i) = \sum_{j=1}^{n}(a_{ij}^{(11)})$

Where $a_{ij}^{(10)}$ is the ij entry of the layer-specific adjacency matrix. It is equal to one if there is an edge between i and j in layer 1 and there is not in layer 2, in other cases $a_{ij}^{(10)} = 0$. The total number of edges which occur only in the layer 1 is

$$L_{10} = \sum_{i=1}^{n} l_{10}(i) \tag{3.5}$$

In the same way is computed the total number of edges which occur only in the layer 2 (L_{01}) and the total number of shared edges L_{11}. The total number of links in the layer 1 is $L_1 = L_{10} + L_{11}$, and the total number of links in the layer 2 is $L_2 = L_{01} + L_{11}$. The total number of edges relative to the node i in the layer 1 is $k_1(i) = k_{10}(i) + k_{11}(i)$.

The hypergeometric test to evaluate the enrichment of connections of the node i in the obese layer becomes (from eq 3.4):

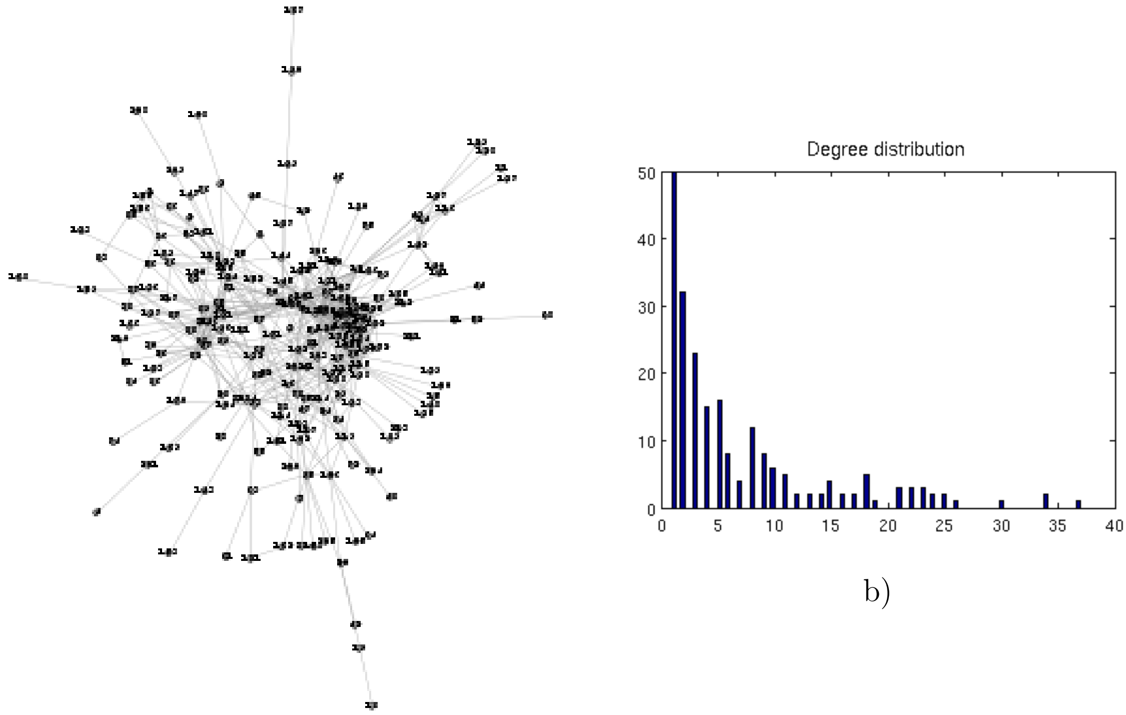
$$P(at\, least\, k_{10}(i); L_1, L_{10}, k_1(i)) = 1 - \sum_{k=0}^{k_{10}(i)-1} \frac{\binom{L_{10}}{k}\binom{L_1 - L_{10}}{k_1(i)-k}}{\binom{L_1}{k_1(i)}} \tag{3.6}$$

The null hypothesis is that there is not enrichment in the number of links of node i which occur only in layer 1. We reject the null hypothesis at the 5% significance level. Therefore this test assesses whether there is a significant enrichment in the number of links of a node ($l_{01}(i)$) lying only in one layer beyond what might be expected by chance, which is established looking at the total L_{01} and L_{11} links and $l_{11}(i)$ links. It should be noted that the same node i can be over-represented in both layers, this fact suggests that the node i has very *different behaviours* in the two layers, where *different behaviours* means different connections.

Going back to our metabolic application, this test shows if there is a significant enrichment in the links of a compound i, therefore if it has different relations in the obese layer and in the normal-weight one.

Toy model

In order to evaluate the significance of the hypergeometric test we performed an analysis on a toy model. We choose a toy model which respects the principal characteristics of biological networks, it is a protein-protein interaction network [56]. This graph is an unweighted undirected graph with 217 nodes and 726 links; in figure 3.4 the structure of the toy model is displayed. We built a multilayer network with two layers, in a

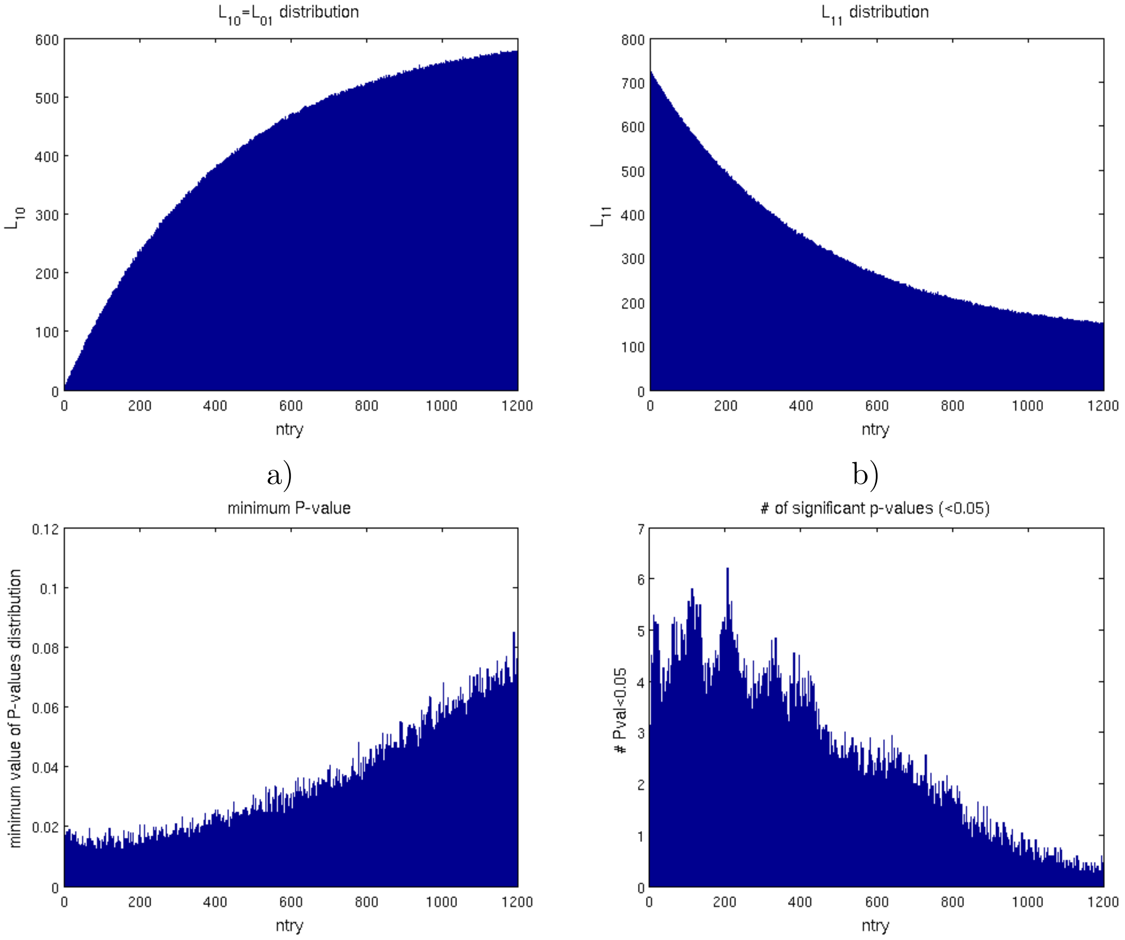


b)

a)

Figure 3.4: Graph of the toy model [56] a) and degree distribution b).

first instance both layers are identical to the null model, therefore the overlap is 100%. In order to analyse the behaviour of the hypergeometric test, we modify step by step edges of one layer, in this way there is a gradual decrease of the overlap. In particular, we decided to gradually modify edges in one layer keeping unvaried the degree of each node (equation 1.6). Therefore we change nodes connections, and we fix the number of edges of each node. With the increase of the randomization, the number of edges (ij) belonging to both layers decreases. Consequently, using the formalism introduced in the previous section, edges which occur only in one layer ($L_{10} = L_{01}$) grow. In figure 3.5.b) we report the evolution of the number of links which are shared by layers (L_{11}), whereas in figure 3.5.a) is displayed the number layer-specific links $L_{10} = L_{01}$. With the increase of randomization and, consequently, the decrease of the overlap, the number of nodes which result enriched in one layer is reduced. This fact is displayed in figure 3.5.d), where it is plotted the number of significant enriched nodes with reference to the randomization . Figure 3.5.c) displays the trend of the smallest p-value versus randomisation: it increases with the randomisation growth. In particular, there are not significant enriched nodes when the overlap is low: the minimum p-value when the overlap is low is equal to 0.07. Both the lower p-value and the number of significant



a)

b)

c)

d)

Figure 3.5: Images referred to the toy model in figure 3.4.a). We built a multilayer network with two layers, at ntry=0 (x axis) both layers are identical to the null model, therefore the overlap is 100%. We modify step by step edges of one layer, keeping the degree of each node unvaried. In this way there is a gradual decrease of the overlap. The x axis indicates an increased randomisation of edges in one layer. Figure 3.5.a) shows the evolution of $L_{10} = L_{01}$: with the increase of randomisation the overlap decrease and the number of nodes belonging only to one layer grows. Figure 3.5. b)shows the evolution of L_{11}. Figure 3.5.c) displays the trend of the smallest p-value versus randomisation, while figure 3.5.d) illustrate the number of significant enriched nodes.

enriched nodes depend on the degree of overlap.

3.2.4 Community detection: stability optimisation

In this section we will describe the community detection method adopted in our multiplex analysis. We perform a community detection separately for the two layers, in order to assess differences of nodes clustering between two networks.

As discuss in chapter 1, there are many clustering networks, which are usually based on distance measures. Distance measures are utilised by clustering methods as k-means and hierarchical clustering, which are methods that do not consider the network structure, that is the presence or the absence of an edge. Therefore we choose to use a quality measure optimisation method, in order to perform a community detection based on the structure of the network. In particular, we opt for stability as quality measure [7], since it merges the idea behind modularity (described in section 1.2.1), with an inner resolution parameter represented by the Markov time (section 3.2.4).

The stability of a graph considers the graph as a Markov chain where each node represents a state and each edge a possible state transition. As before, we consider a graph composed by n nodes and m edges, which can be weighted or not. The vector d indicates the degree of each node (eq. 1.6) and thus it is a n size vector. If the graph is weighted, d can represent the strength of each node (eq 1.7). The corresponding diagonal matrix is $D = diag(d)$. The stationary degree distribution is thus represented by $\pi = d/2m$ where, Π is the corresponding diagonal matrix $\Pi = diag(\pi)$. The transition between states, which are symbolized by graph nodes, is given by the $n \times n$ *stochastic matrix* $M = D^{-1}A$, since edges are assumed to display the possible state transitions. A community partition can be represented by a matrix H of size $n \times c$ where c is the number of communities, so that each node belong to one community. Stability is computed starting from the *auto-covariance matrix* at Markov time t, which gives the covariance of the process with itself at pairs of time points. The clustered auto-covariance matrix is defined as:

$$R_t = H^T(\Pi M^t - \pi^T \pi)H \tag{3.7}$$

Consequently, if the process is not stationary, the stability changes according to time t, so it is noted as Q_{S_t}. The stability value Q_{S_t} is given by the trace of R_t, and the global stability measure Q_S considers the minimum value of the Q_{S_t} over time, from time 0 to a given upper bound τ:

$$Q_S = \min_{0 \leq t \leq \tau} trace(R_t) \tag{3.8}$$

Therefore stability allows to evaluate the goodness of clusters toward different times, which are different partition scales. Form equations 1.16 and 3.7 it can be seen that stability at time $t = 1$ is modularity.

Stability optimization methods are inferred by modularity methods, which are described in section 1.2.1. The former, unlike modularity methods, permits to investigate the evolution of communities along the Markov time.

The stability at time t can be obtained from modularity since, for discrete time model, from eq 3.7 $Q_{S_t} = Q_M(A_t)$ with $A_t = DM^t$. Stability optimisation can thus be reduced to modularity optimisation methods with an additional resolution parameter representing by the Markov time. At each Markov time, the partition with the best stability value (Q'_S) is kept and Q_S is update as $Q_S = Q'_S$. Defining with $\triangle Q_{S_t}$ the change of stability at time t, from eq. 3.8

$$Q'_S = \min_{0 \leq t \leq \tau} (Q_{S_t} + \triangle Q_{S_t}) \tag{3.9}$$

The fastest way to approximate stability is to compute it with only one time value. As stability tends to decrease as the Markov time increases, we are seeking when the following approximation can be made:

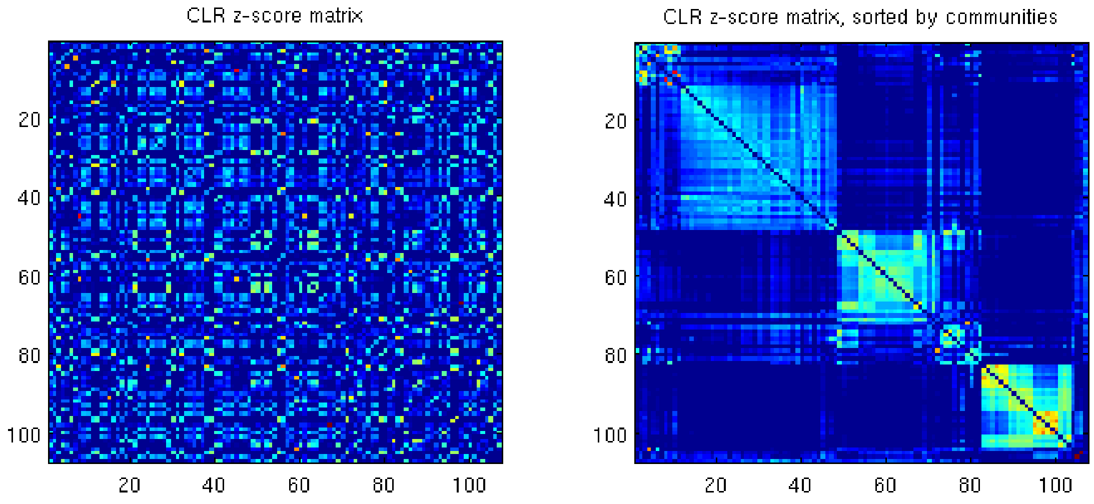
$$Q_S = \min_{0 \leq t \leq \tau} trace(R_t) \approx trace(R_\tau) \tag{3.10}$$

The computation time increases according to the time boundaries adopted, but a wide time interval permits to better evaluate the obtained clusters.
As for modularity optimisation methods, many algorithms have been implemented in order to improve the stability optimisation algorithm performance. We use the fast multi-scale detection algorithms, which are explain by Erwan Le Martelot [6]. In order to obtain the most stable results we implemented a consensus algorithm at each Markov time; this method is described in the successive section.

The goodness of a partition can be evaluated looking at the duration of that partition/ community along the Markov chain. As reported by Le Martelot [57], in the absence of knowledge on networks, the analyst can look for community structures that are consistently found on some scale intervals. These are stable partitions. Similar or identical partitions may have about the same composition of communities. For a qualitative evaluation of that parameters we can consider fig. 4.6 in chapter 4, which displays the evolution of communities with the increase of the Markov time of our case of study; while in figure 4.8 it illustrates the stability evolution. Another qualitative evaluation of the partition can be done looking at the adjacency matrix: as we described in chapter 1, the presence of groups of nodes highly connected produces a block matrix. We plot in figure 3.6 the adjacency matrix where nodes (i.e rows and columns) are sorted by the detected communities in comparison with the same adjacency matrix, where rows do not have a order criterion.

Markov chains

A Markov chain [58] is a random process used to represent sequences of states of a system. The evolution of many systems can be represented by a Markov chain: in



a) b)

Figure 3.6: Adjacency matrix with no order criterion a), same adjacency matrix with rows and column sorted by the detected communities b).

physics it is commonly adopted for the description of thermodynamic systems, since their dynamic is assumed to be time-invariant. Markov chains are widely adopted in chemistry to described chemical reactions, as those modelled by Michaelis–Menten kinetics. This statistical method has many application also in network theory, where it is used to calculate *random walks*.

Considering a set of states $S = s_1, s_2, ...s_n$, the process starts in one of these states and it moves successively from one state to another. For each pair of states s_i and s_j, there is a probability p_{ij} of going from state i to state j, where for each i, $\sum_j p_{ij} = 1$. The probabilities p_{ij} are called *transition probabilities*. These probabilities depend only upon the current state of the chain, and not upon the previous states; that is why this process is often defined as *memoryless*. Formally it becomes;

$$\begin{aligned} &P[x(t_{n+1}) = x_{n+1}|x(t_n) = x_n,\, x(t_{n-1}) = x_{n-1}, \ldots, x(t_0) = x_0] = \\ &= P[x(t_{n+1}) = x_{n+1}|x(t_n) = x_n] \end{aligned} \tag{3.11}$$

where $t_1 < t_2 < ... < t_n < t_{n+1}$.

The transition probabilities can be represented as a square matrix $\mathbf{P}$ called *transition matrix*, where p_{ij} shows the probability of being at state j at time $t+1$ if the current state is i at time t. Therefore, the ijth entry $p_{ij}^{(n)}$ of the matrix $\mathbf{P}^n$ gives the probability that the Markov chain, starting in state s_i, will be in state s_j after n steps.

3.2.5 Consensus matrix

Consensus clustering is a data analysis method introduced by Lancichinetti and Fortunato [50]. This method is utilized to generate stable results out of a set of partitions delivered by stochastic methods. We applied this method in order to enhance the stability and the accuracy of the community detection algorithm: we iterate several times the community detection at each Markov time and, at each t, we perform a consensus algorithm. In this section we briefly describe the implemented algorithm.

The consensus algorithm is applied on a set of n_p partitions, produced by a classification algorithm, starting from a network with n nodes. The consensus matrix D is a square weighted matrix $n \times n$ which entries have value between 0 and 1. The value assigned to an element D_{ij} is calculated as the ratio between the number of partitions n_s in which nodes i and j are in the same group and the number of total partitions n_p. High values of D_{ij} mean that nodes i and j appear in the same group in most partitions, on the other hand, lower weights indicate a low probability that the two nodes belong to the same group. Since the consensus matrix D is used as adjacency matrix for recalculating the n_p partitions, it is necessary to drop the low values in order to avoid that the successive consensus matrices will become to much dense. A threshold $\tilde{d}$ is then fixed, in this way random edges are not considered. If a node has all the weight edges below the threshold value, it will be disjointed from the network; in order to avoid this eventuality only the higher edge weight is maintained.

This method uses an iterative process: the consensus matrix D is used as adjacency matrix to generate other n_p partitions, then D is re-calculated from the new partitions. The process go ahead until all weights D_{ij} are 0 or 1; $D_{ij} = 1$ indicates that nodes i and j are in the same group, instead $D_{ij} = 0$ means that the two nodes belong to different groups.

As already said, we applied the consensus method to the n_p partitions generated by a community detection algorithm C . The procedure can be summarising in a sequence of steps:

1. obtain n_p partition from C using the adjacency matrix
2. compute the consensus matrix D, where $D_{ij} = n_s/n_p$
3. if $D_{ij} < \tilde{d}$ set $D_{i,j} = 0$; if a node becomes disjoint, keep the higher weight.
4. apply C on D n_p times, so to yield n_p partitions.
5. recalculate D using the new n_p partitions

6. if all entries are equal to 0 or to 1 stop, else go back to 2.

This method allows to enhance the stability and the accuracy of the community detection algorithm. In fact, there are some nodes which lie on the *boundary* of two communities, and they are assigned to different communities in different iterations of the community detection algorithm. Therefore we iterate several times the community detection algorithm at each Markov time and, at each t, we perform the consensus algorithm. The final classification during different Markov time is the most stable partition of that network.

3.2.6 Normalized mutual information (NMI)

The community detection with consensus method, described in the previous sections, is performed separately for the two layers; therefore we obtain the evolution of the communities partitions of each layer. It becomes useful to define a measure which permits to compare the sets of clusters found in different graphs, since the clustering analysis examines each single layer as a separate graph. More precisely, this measure has to evaluate how similar or different the sets are.

In multiplex networks analysis this similarity measure is often compute using the *normalized mutual information (NMI)* [59] [60].
Considering two discrete random variables $X = (x_1, ..., x_n)$ and $Y = (y_1, ..., y_n)$, which are associated to the partitions C_X and C_Y, the *mutual information (MI)* is defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right), \tag{3.12}$$

where $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively and $p(x,y)$ is the joint probability distribution function of X and Y. Mutual information can be equivalently expressed as function of the entropy. Named $H(X)$ and $H(Y)$ the marginal entropies of X and Y, $H(X|Y)$and $H(Y|X)$ the conditional entropies, and $H(X,Y)$ the joint entropy of X and Y, the equation 3.12 becomes:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3.13}$$

Consequently, the normalized mutual information $I_{norm}(X:Y)$ is defined as:

$$I_{norm}(X:Y) = \frac{H(X) + H(Y) - H(X,Y)}{(H(X) + H(Y))/2} \tag{3.14}$$

The normalisation ensures that the I_{norm} values lie in the range [0, 1]; I_{norm} equal to 0 means that the two sets are totally dissimilar, while a value $I_{norm} = 1$ indicates that they are identical.

This measure gives a quantification of the similarity of two partitions, therefore a low NMI value suggests that the two layers have different clusters. We evaluate the NMI at each Markov time of the stability optimisation method in order to highlight the time at which communities present the maximum diversity (fig 4.8).

3.2.7 Wilcoxon rank sum test

The Wilcoxon rank sum test is performed to evaluate differences of edges weights belonging to one community between two layers. More precisely, we choose the best partition of each layer; for each community of the layer 1 we evaluate if the weights distribution of edges in that community is significantly different in respect to the corresponding distribution in the layer 2. Since the partitions of the two layers are rather different, we impose the partition of the layer 1 on the layer 2 and vice versa. Here we described the main characteristics of the performed test.

The Wilcoxon rank sum test is a non-parametric test which is used to verify if two statistical independent samples of ordinal values from a continuous distribution, come from the same population.

The null hypothesis is that the two samples X and Y (obese and the normal-weight layer case of study) come from the same population, in this case their probability distributions are equal.

The Wilcoxon rank sum test ranks the combined two samples $(X+Y)$ and then it calculates the sum of the ranks for each group R_x and R_y. Indicating the two samples dimensions as n_x and n_y, the statistic U is calculated as:

$$U_x = R_x - \frac{n_x(n_x-2)}{2} \qquad U_y = R_y - \frac{n_y(n_y-2)}{2} \tag{3.15}$$

The P-value indicates what is the chance that random sampling would result in the mean ranks being as far apart as observed, if the groups are sampled from populations with identical distributions.

For small sets of observations, usually lower that 100 observations, the exact p-value is calculated by calculator, which randomly classifies n_x of the total observations in the X samples and the rest in the Y's sample . For large samples, it is usually adopted a z-statistic to compute the approximate p-value of the test. In that case, the standardized value

$$z = \frac{U - m_U}{\sigma_U} \tag{3.16}$$

where m_U and σ_U are the mean and standard deviation of U

$$m_U = \frac{n_1 n_2}{2} \qquad \sigma_U = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}} \tag{3.17}$$

The standard normal distribution gives the p-value for this z-statistic.

We use this test to evaluate if there are significant differences between edges weight which belong to the same community, but in different layers.

3.3 The software Gephi

Gephi [61] is an open-source software for networks visualization and exploration. Through an easy interactive interface it permits to manipulate graph structures and visually evaluate modification in nodes and edge arrangement. Since visualizations are useful to find features in network structure, Gephi shows in real time all manipulations and filtering. Various layout algorithms both for efficiency and quality can be used to give the shape to the graph, it is also possible to change layout settings while algorithm is running and to visualize in real-time the framework variation. Graph can be also visually analysed using the most common metrics as betweenness centrality, closeness, diameter, clustering coefficient, community detection (modularity), shortest path. In order to improve network readability, the thickness of edges is proportional to their weight ans it is also possible to show node labels.

Data are imported in two data tables, one for nodes and the other one for edges. The node table is composed by columns which indicate node IDs, the corresponding node label and other node characteristics. The edge table describes the edge characteristics: each row corresponds to one link, the first two columns contain the source and the target nodes, other columns are used to indicate if the edge is directed or not and the edge weight.

We used Gephi to visualise the network structure, in order to qualitatively evaluate our results. In the node table we specify the community label for each node, to visualize the different clusters with an assigned color. Various kind of layout algorithms were employed, in the end we opted for the *ForceAltas* algorithm [62]. This method was made to spatialize Small-World and Scale-free networks, i.e. networks of real data. The parameters utilised by ForceAtlas are: a repulsion strength, which ensures that a node rejects others, an attraction strength, which puts connected nodes near and a gravity parameter which attracts all nodes to the center to avoid dispersion of disconnected components. Since ForceAtlas permits a rigorous interpretation of the graph with the least bias possible, we choose this layout to qualitatively evaluate the goodness of partitions.

Chapter 4

Results

In this chapter we illustrate the results that we obtained using the method explained in chapter 3. We apply the complex graph theory on multi-omics biological data in order to discover complex relationships between elements of that omics. In particular our study is linked to metabolomics, since we perform a multiplex network analysis using metabolic data. All layers of the multiplex have the same nodes, which are blood serum compounds, therefore the final aim of this project is to reveal significant differences between layers, and these differences are set in the nodes connections.

4.1 Data processing

The Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study collected samples concerning metabolomics, transcriptomics and genomics of a Finnish cohort, as already said in section 2.5. They aim to investigate possible relations between lipids, immune cells in circulation and cardiovascular diseases. Therefore the DILGOM study offered four datasets: one related to genomic, one to transcriptomic (SNP), one to metabolomic and one that describes the phenotypes of the individuals. The total number of individuals analysed by the DILGOM study is 518, in particular 240 males and 278 females. Since not all the subjects have compounds, SNPs and phenotypes data, it is necessary to remove such people from the analysis.

Moreover, as written in the file *DILGOMdate. Update for Case Study* , there are some individuals with 'aberrant' variables values in the phenotypes dataset, indicating diabetics, individuals under cholesterol treatments and individuals with fast glucose anomalies. For reasons explained in section 2.5, these individuals are removed from

Case 1, not considered metabolites				
APOA1	VAL	TOTFA	PC	TGPG
APOB	ESTC	LA	SM	CH2DB
APOBAPOA1	FREEC	OTPUFA	FAW3FA	DBINFA
BOHBUT	FAW3	DHA	FAW6FA	BISDB
GLOL	FAW6	MUFA	FAW79SFA	BISFA
LEU	FAW79S	TOTPG	CH2INFA	FALEN

Table 4.1: Metabolites with Nan values. These metabolites are not considered in case 1.

the list.

In particular we try two different ways for the homogenization of the dataset:

1. intersection of SNP, compounds and phenotypes (without aberrant values) datasets. This way will permit to integrate our results, which are related to the compound dataset, with future analysis on SNP. This intersection leads to a subset of 187 individuals.

2. intersection of compounds and phenotypes datasets, without considering the SNP dataset. In this way we have a dataset as large as possible. This intersection leads to a subset of around 500 individuals.

Another reduction of the number of samples is caused by the compounds dataset. In fact, not all the subjects have all the values of compounds concentrations. Figure 4.1 displays the complete compounds dataset matrix (m individuals × n compounds), where blue points indicate Nan values. We see that Nan values are linked to 30 compounds, therefore we used two different ways to manage these Nan values: for the case 1) (dataset with 187 individuals) compounds with Nan values are deleted (i.e the columns of the matrix in figure 4.1). We choose to delete compounds in order to not further reduce the samples number, therefore the subset of case 1 is composed by 187 individuals and 107 compounds. The majority of the discarded compounds belong to the class of fatty acids, while all lipoproteins measurements are maintained. The deleted compounds are reported in table 4.1. For dataset 2, that is the dataset with approximately 500 individuals, we consider all the compounds while individuals with Nan values are discarded. In this second dataset the final subset is composed by 418 individuals and 137 compounds. We also analysed this larger dataset considering 418 individuals and the 107 compounds of dataset 1, in order to validate the results obtained for the smaller dataset.

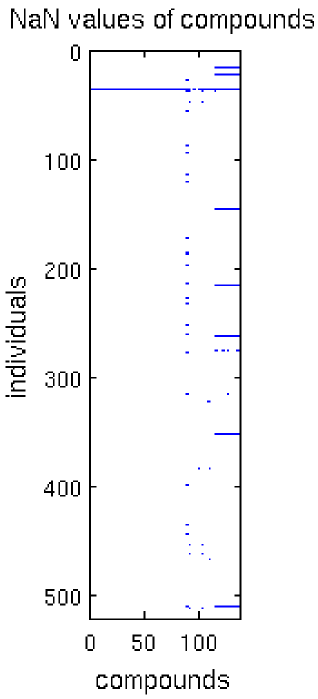


Figure 4.1: Image of the complete compound dataset *individuals* × *metabolites*, where blue points correspond to Nan values.

4.2 Classification

Classification is carried out to extract two subsets which delineate obese and normal-weight people. This classification is based on the phenotypes dataset, which is described in section 2.5.1. This classification permits to investigate possible differences between blood compounds relations for obese and normal weight people. If significant differences are found, then the metabolomics and the phenotypes are correlated.

The classification using the phenotypes dataset is thus a preliminary important step in view of performing a multiplex analysis. The final aim is to assess differences and similarities between two layers, which represent the metabolites correlations of the two defined subsets.

We tested various classification methods, in order to compare the final results of the multilayer analysis and to verify the robustness of the outcomes. Clearly, different methods can be based on slightly diverse assumptions, so they can lead to marginally different subset and results. In the end we opt for the thresholds suggested by WHO, we consider also a classification based on the glm, which is reported in appendix A.

method	dataset	class	BMI	LDL	HDL	age	systolic Pressure
WHO	1	obese	33.6 ± 3.6	5.7 ± 0.6	1.2 ± 0.3	57 ± 11	139 ± 16
		normal	22.3 ± 1.5	5.1 ± 0.9	1.7 ± 0.4	47 ± 13	125 ± 18
	2	obese	33 ± 4	5.6 ± 0.8	1.3 ± 0.3	57 ± 10	138 ± 17
		normal	22.2 ± 1.7	5.1 ± 0.9	1.6 ± 0.4	46 ± 14	126 ± 17

Table 4.2: Mean and standard deviation values of some phenotypes features, related to normal and obese groups.

method	cass	BMI	age	LDL	HDL	fast Gl	sys P.	dias P	WHO	waist circ
WHO	1	1	1	1	1	1	1	1	1	1
	2	1	1	1	1	1	1	1	1	1

Table 4.3: Results of the median test (sec. 3.2.7) between the obese and the normal-weight clusters. The null hypothesis is that the two groups belong to the same population, $H_0 = 1$ indicates that the null hypothesis is rejected. The confidence is set at $p = 0.05$.

4.2.1 WHO thresholds

We used a partition method based on the directions of the *World Heath Organization* (WHO) to classify obese and normal-weight individuals. This classification is realised looking only at the values of waist-hip ratio, BMI and waist circumference. As reported in table 2.1, WHO fixed different threshold values for female and male individuals, therefore we classify separately obese female and obese male individuals using their respective threshold values. Once the two groups of obese people (male and female) are obtained, they are grouped in one obese cluster. The same strategy is adopted for normal-weight male and female individuals. We display the main phenotypic differences between normal and obese groups in table 4.2, where the mean and standard deviation values of some phenotypes features are reported.

We also perform a median test (see section 3.2.7) for quantifying the phenotype differences between the obese and the normal-weight clusters.The null hypothesis is that the two groups belong to the same population, which means that their phenotypes are not significantly different. The confidence value is set at $p = 0.05$ and a value of $H_0 = 1$ indicates that the null hypothesis is rejected. In table 4.3 the results of the test are reported. The WHO thresholds method forms two groups that are significantly different with respect to all the phenotipic parameters. Only the gender appears to be homogeneous between obese and normal-weight groups.

4.3 Multiplex network analisys

The main purpose of this thesis is the setting-up of a multiplex network and its analysis. As enunciated in section 1.3.1, multiplex networks belong to *complex networks theory*, a recent extension of the classical graph theory which permits to investigate real networks, that are usually characterised by complex frameworks.

We build a weighted multiplex network with two layers, one for the obese group and the other one for the normal-weight group; these groups are obtained using the classification method described in section 4.2.

These layers are fully described by their $n \times n$ symmetric adjacency matrices $A^{(1)}$, $A^{(2)}$ which are constituted by the same number of nodes (n). Each node represents a blood compound extracted by the DILGOM study. Since we evaluate the correlation between compounds, each row i and each column i of A represents the interactions between a specific compound i and all the others. The differences between the two layers are shown by their coefficients $a_{ij}^{(1)}$ and $a_{ij}^{(2)}$. In order to avoid indirect relations, we apply the CLR algorithm to each correlation matrix (section 3.2.2). This method builds a z-score matrix where the contrast between the physical interactions and their indirect relationships is enhanced. The complete procedure we implemented to build the adjacency matrices with meaningful informations is described in chapter 3 where figure 3.1 shows the schematic block diagram of our method.

The comparison between the two layers is carried out looking at:

- differences of a fixed node in different layers. We implement a *hypergeometric test* (section 3.2.3) in order to highlight the enrichment of connections of a node in a specific layer.
- differences between the same community in different layers. This comparison requires a preliminary intra-layer community detection, in order to define groups of compounds which are strongly linked together. Once determined compounds clusters, we move on the parallel of the two layers.

In this section we report the results of the multiplex analysis, which ground on the study of the z-score adjacency matrix of each layer.

4.3.1 The hypergeometric test

The hypergeopetric test is performed to evaluate the most significant differences between single compounds of the normal-weight and the obese layers. More precisely, the hypergeometric test assesses if there is a significant enrichment in the number of links of a node i in a specific layer beyond what might be expected by chance.

As explained in detail in section 3.2.3, this method requires topological matrices to evaluate the enrichment; since the adjacency matrices are z-score weighted matrices, we set a threshold $\tilde{Z}$ on the adjacency matrix values. Entries of the adjacency matrices which are higher than the threshold $\tilde{Z}$ are set equal to one, while all the others are set equal to zero. Different threshold values are fixed in order to not affect results by the $\tilde{Z}$. In particular, the tested thresholds rank from $\tilde{Z} = 1.5$ to $\tilde{Z} = 3.4$ with steps of 0.1. For each threshold we obtain two topological matrices, one for the obese layer and the second one for the normal-weight layer. The hypergeometric test is performed for each layer. Referring to the formalism introduced in section 3.2.3, we evaluate whether there is a significant enrichment or a depletion in the number of links of a node ($k_{01}(i)$) lying only in one layer beyond what might be expected by chance, which is established looking at the total L_{01} and L_{11} links and $k_{11}(i)$ links. Therefore the comparison is between edges belonging only to one layer (as example the obese layer) and the shared links. Considering as example the obese layer, the hypergeometric test returns for each node (i.e compound) a value equal to one if the null hypothesis is rejected, that is if the compound is oversampled in the obese layer, and 0 otherwise. We fix the significance level at 0.05. For each layer we consider the number of times that a specific compound results to be enriched, since we perform the test using different thresholds. At the end we rank the compounds looking at their number of over-representations; the results for the different datasets are listed in table 4.4.Numbers associated to each compound indicate the number of times that every single compound has p-value ≤ 0.05. Since the threshold $\tilde{Z}$ ranks from 1.5 to 3.4, the maximum possible number is 20.

Table 4.4 shows the results obtained for the different datasets. The first two columns list the rank obtained using the dataset 1. In order to validate the results of the smaller dataset, we perform the hypergeometric test on the dataset with 418 individuals, restricted to the 107 compounds of the dataset 1. The results are listed in the third and fourth columns of table 4.4. The last two columns refer to the results obtained for the complete second dataset (418 individuals and 137 metabolites).

The results of the obese and normal-weight layers are coherent between the dataset 1 and 2. Moreover, results of the hypergeometric test show that some compounds have a different behaviour between the two layers, especially for what concerns metabolites and good cholesterol. In particular, metabolites as lactate, glucose, glutamine, histidine, phenylalanine, pyruvate, tyrosine, albumin and alanine are oversampled in the obese layer. Some of these metabolites, such as glucose and albumin, are oversampled also in the normal layer; this fact suggests that they have different links in the obese and in the normal-weight layers. Also the measures of the medium good cholesterol (MHDL-) seem to have different behaviours in the two layers, therefore they are interesting for our analysis.

It must be stressed that these results are linked to the multilayer structure that we build. In fact, a simple analysis of the blood concentrations does not reveal differences between those compounds. We analyse the differences in concentration between the obese and the normal-weight groups of individuals. Both the t-test and the rank-sum test reject the null hypothesis that the MHDL- cholesterol measures are significantly different between the two groups. Also the metabolites glutamine, histidine, albumin, creatinine, urea, acetate, acetoacetate and glycine are not significantly different in concentration between the two groups. This fact indicates that a network approach can give further or different informations with respect to an analysis based on the blood concentrations.

In figure 4.2 the evolution of the number of edges depending on the threshold is displayed. The red line refers to edges which occur only in the obese layer (L_{10}), the blue line shows links lying only in the normal-weight layer (L_{01}) and the green line indicates the number of shared links (L_{11}). Since the shared edges are many more with respect to L_{10} and L_{01}, we plot the $L_{11}/10$. As expected, L_{11}, L_{10} L_{01} decrease with the increase of the z-threshold and only the most significant edges endure. Figure 4.2.a) refers to dataset 1 (107 compounds), while Fig. 4.2.b. refers to dataset 2 (with 137 compounds). To evaluate the importance of the overlap in the hypergeometric test, we randomize the edges of the obese layer. The procedure is the same that we utilize for the toy model of section 3.2.3. We consider the dataset 1 and we utilize the topological matrices of the obese and normal-weight layers, which we obtained fixing a threshold at $\tilde{Z} = 2.1$. We chose this threshold since the two layers have approximately the same number of edges at $\tilde{Z} = 2.1$, as shown in Fig. 4.2.a). In particular, links which occur only in the obese layer (L_{10}) are 214, edges which belong only to the normal-weight layer are $L_{01} = 224$ and the number of shared links is $L_{11} = 628$. As expected, the number of significant enriched nodes ($p - val < 0.05$) decreases with the increase of the randomization; this trend is displayed in figure 4.3 c). We consider also the minimum p-value at each randomization step: the minimum p-value grows with the increase of the randomisation; this trend is shown in figure 4.3 d). In figures 4.3.a and 4.3 b). we report the evolution of L_{10} and L_{11} with respect to the increase of randomization.

Going back to the results in table (4.4), we note a clear difference as concerns the metabolites and amino acids behaviour (section 2.4.3). In particular they seem to be more oversampled in the obese layer. Therefore we represent with a network image their connections in order to qualitatively evaluate the differences. Figure 4.4 displays

WHO 187x107				WHO 418x107				WHO 418x137			
Obese		Normal		Obese		Normal		Obese		Normal	
LAC	17	CREA	1	GLC	20	SHDLL	9	PHE	20	CREA	14
GLC	15	ALB	8	PHE	18	SHDLP	7	GLC	18	SHDLL	10
TYR	14	XLHDLTG	7	XLHDLTG	11	XLHDLTG	6	GLOL	15	ACE	9
ALB	12	LDLD	7	MHDLC	11	CREA	6	UREA	15	APOB	8
PHE	12	XLHDLCE	6	ALA	11	LVLDLC	4	ALB	13	SHDLP	7
GLN	11	GLY	6	ALB	10	IDLL	4	MHDLFC	13	LVLDLC	6
IDLTG	10	UREA	6	MHDLFC	10	IDLCEFR	4	MHDLL	12	FAW6	6
PYR	10	IDLL	5	MHDLCE	10	SVLDLTG	3	LDLD	12	LA	6
XSVLDLPL	8	SHDLL	5	UREA	10	IDLP	3	MHDLC	11	XLHDLTG	5
XSVLDLL	7	XLVLDLPL	4	MHDLPL	9	SERUMC	3	MHDLPL	11	IDLCEFR	5
XSVLDLP	7	XLVLDLL	4	MHDLL	9	ACE	3	MHDLCE	10	MOBCH3	5
XLHDLTG	7	XLVLDLP	4	LDLD	9	GLC	3	MHDLP	10	GLOL	5
MHDLL	7	XSVLDLTG	4	MHDLP	8	TYR	3	ALA	10	TOTFA	5
LDLD	6	IDLFC	4	XSVLDLPL	6	ALB	2	FAW6FA	10	MVLDLFC	4
MHDLPL	5	XLHDLC	4	IDLTG	6	LVLDLFC	2	GLN	9	MVLDLP	4
MHDLCE	5	SHDLP	4	GLN	6	LVLDLCE	2	XLHDLTG	7	VLDLD	4
ALA	5	GLC	4	HIS	5	MVLDLFC	2	APOA1	7	APOBAPOA1	4
HIS	5	HIS	4	ILE	5	MVLDLPL	2	SM	7	HDL3C	4
XXLVLDLP	4	XLVLDLTG	3	XXLVLDLP	4	MVLDLTG	2	GLY	6	GLC	4
XSVLDLTG	4	IDLPL	3	XSVLDLTG	4	MVLDLL	2	FALEN	6	TYR	4
IDLP	4	IDLP	3	XSVLDLL	3	MVLDLP	2	XSVLDLPL	5	FAW79S	4
MHDLC	4	IDLCEFR	3	IDLP	3	SVLDLC	2	HDL3C	5	CH2INFA	4
MOBCH2	4	ACE	3	XSVLDLP	2	SVLDLFC	2	XXLVLDLP	4	TGPG	4
ILE	4	ACACE	3	IDLPL	2	SVLDLL	2	XSVLDLL	4	XLVLDLPL	3
XXLVLDLPL	3	CIT	3	IDLL	2	XSVLDLTG	2	LHDLFC	4	LVLDLFC	3
				LHDLPL	2	LLDLCE	2	IDLTG	4	LVLDLPL	3

Table 4.4: Rank of DILGOM compounds based on p-values of hypergeometric test. The rank is obtained considering the results of hypergeometric test for different threshold values: from z-score of 1.5 to 3.5 with steps of 0.1. We fix the significance level at 0.05. The number associated to a compound name indicates the number of times that compound has p-value ≤ 0.05. The first two columns are related to the first dataset (187 × 107), the third and the fourth columns refer to the dataset 2 with 107 compounds while the last two columns are related to the dataset 2 with 137 compounds.

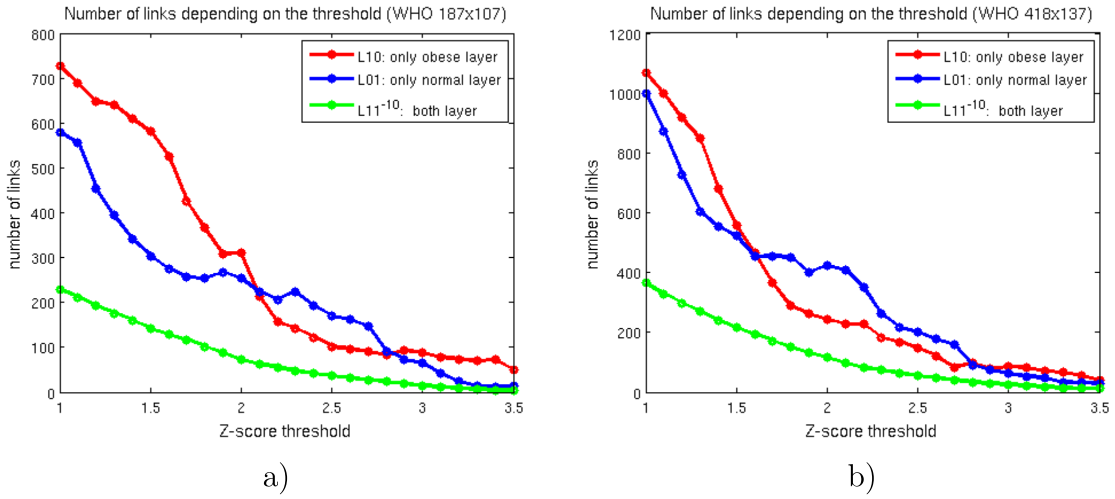
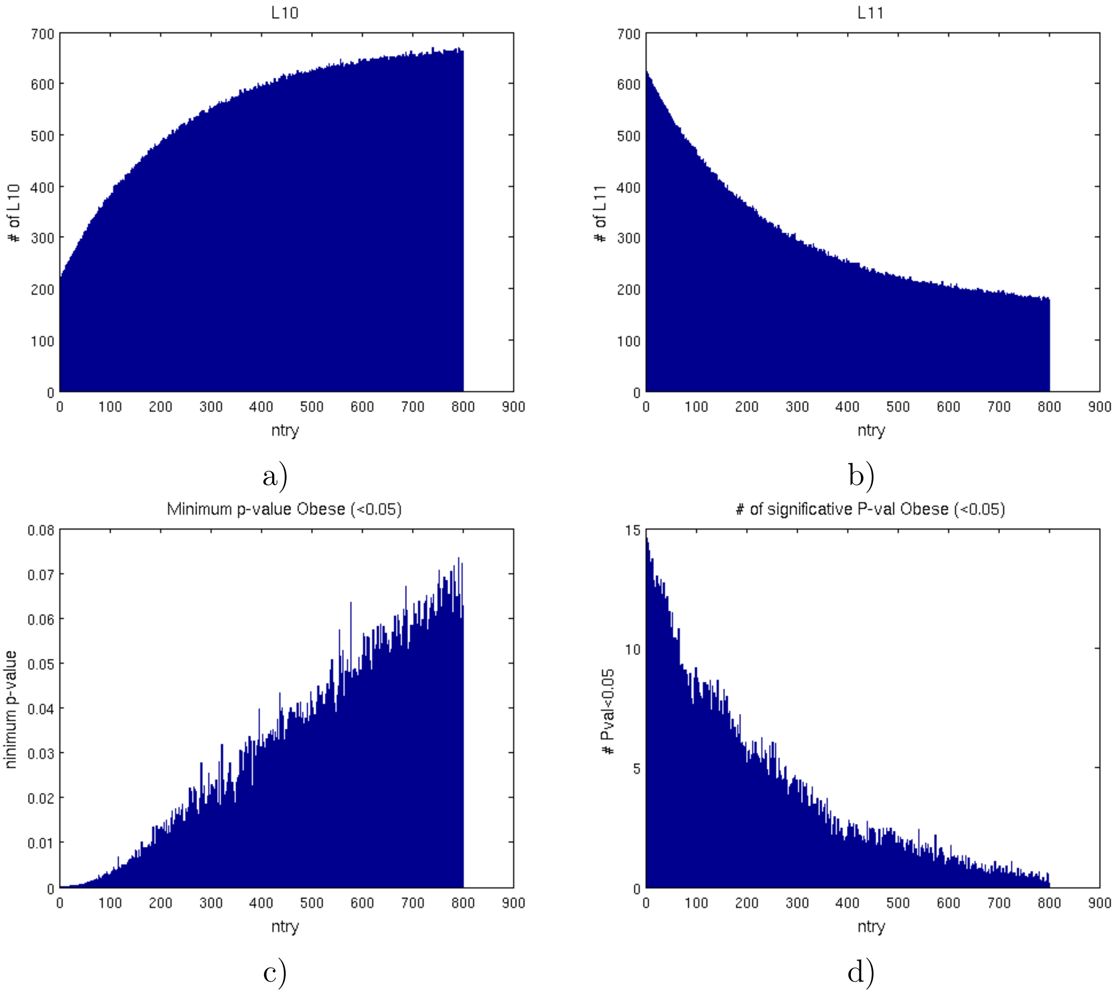


Figure 4.2: Evolution of the number of edges depending on the threshold. Red line refers to links lying only in the obese layer (L_{10}), blue line refers to links lying only in the normal-weight layer (L_{01}) and green line to the number of shared links (L_{11}). Image 1) refers to the dataset 1 (107 compounds), and image 2 to dataset 2(137 compounds); using WHO thresholds for the classification.

the graph of the overlap of metabolites and amino acids, considering the dataset 2 (418×137). Red lines refer to edges which occur only in the obese layer (L_{10}), blue lines show links lying only in the normal-weight layer (L_{01}) and green lines indicate the shared links (L_{11}). The overlap in figure 4.4a) is obtained for a threshold of $\tilde{z} = 1.6$; we choose this threshold because $L_{01}(\tilde{z} = 1.6) = L_{10}(\tilde{z} = 1.6)$, as can be seen in figure 4.2.b. Figures 4.4.c., 4.4.d. and 4.4.e display the graph in figure 4.4a) considering L_{11}, L_{10} and L_{01} separately. These images help to visualise the graph in figure 4.4a); they show that the these metabolites have more links in the obese layer than in the normal one, in particular glucose, phenylalanine, albumin, alanine and glutamine are highly connected in the obese layer. The normal-weight layer shows many relations between the amino acids: isoleucine, leucine, valine and tirosine. The urea seems to have very different behaviour between the two groups. Figure 4.4.b. shows the edges overlap for the threshold at $\tilde{z} = 2.5$. At this threshold, edges which occur only in the obese layer (L_{10}) are less then those in the normal-weight layer (L_{01}), as we deduce from figure 4.2.b. The graph 4.4.b. shows that metabolites and amino acids are stronger bounded in the obese layer than in the normal one. In particular, glucose, phenylalanine and glutamine have many links in the obese layer, but very few in the normal one. On the other hand, the amino acids: isoleucine, leucine, valine and tirosine have more links in the normal-weight layer.

This qualitative analysis suggest that the metabolites group is more linked in the



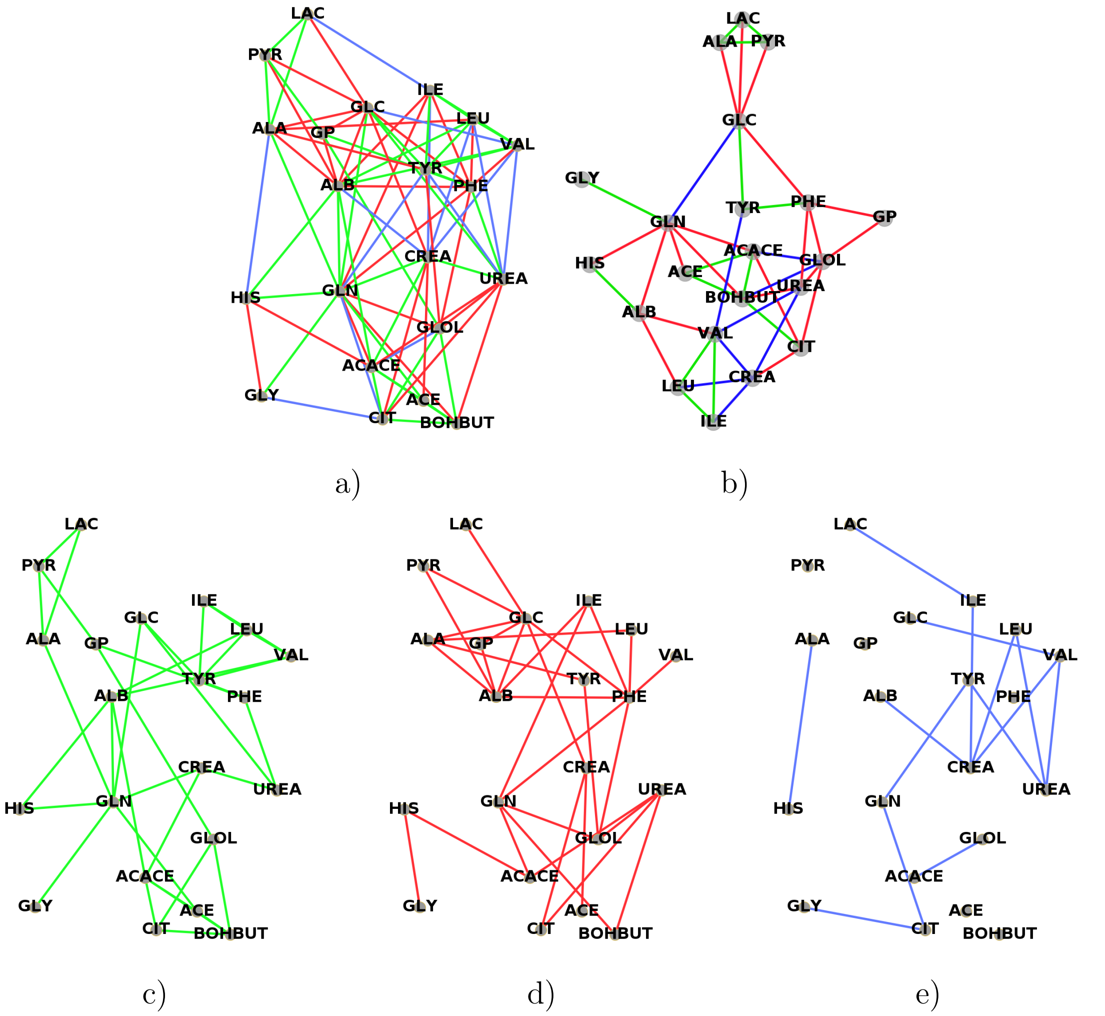
a) b) c) d)

Figure 4.3: Randomisation of links in the obese layer. The x axis indicates an increased randomisation of edges in the obese layer. Figure 4.3.a) shows the evolution of L_{10} with respect to the randomization. Figure 3.5. b) shows the evolution of L_{11}. Figure 4.3.c) display the trend of the smallest p-value versus randomisation, while figure 4.3.d) illustrates the number of significant enriched nodes.

obese group. In the next section we consider whether all compounds are linked together and in what way, that is we evaluate the communities structure of the two layers.

4.3.2 Communities analysis

We perform a community detection in order to determine subgroups of compounds which are strongly bonded together. In particular we want to evaluate how the compounds highlighted by hypergeometric test are linked. The final goal of this analysis is to highlight *structural* differences between the two layers and, precisely, between their



a)

b)

c)

d)

e)

Figure 4.4: Graphs of the overlap of metabolites and amino acids (see section2.4.3), considering the dataset 2 (418 × 137) and using classification into obese and normal layer based on the WHO thresholds method. Red lines refer to edges which occur only in the obese layer (L_{10}) (figure 4.4 d), blue lines show links lying only in the normal-weight layer (L_{01})(figure 4.4 e) and green lines indicate the shared links (L_{11})(figure 4.4 d). The overlap in figure 4.4 a) is obtained fixing the threshold at $\tilde{z}$=1.6; at this threshold $L_{01} = L_{10}$. Figure 4.4.b. shows the edges overlap for $\tilde{z}$=2.5.

communities. Therefore this study can be divided in two main steps: the first one is related to the community detection, while the second and principal step concerns the evaluation of differences between community partitions of the two layers.

Community detection

We perform a community detection separately for the two layers and we use the method described in section 3.2.4. The quality function chosen to evaluate communities is the stability (equation 3.8), since it allows to evaluate the goodness of clusters through different partition scales. In particular, the stability evaluates if the density of edges within community structures compared with a random distribution of links between all nodes is significantly different. Therefore it ranges between -1 and 1: high positive values means a good definition of the communities.

We implement the community detection algorithm which assigns each compound to a community, according to which group it belongs to. The algorithm returns different node partitions according to the given Markov time. Therefore we evaluate a *range* of partitions, from $t = 0.6$ to $t = 2.0$, with discrete steps equal to $\delta t = 0.1$. Since clustering methods can lead to slightly different partitions on the same graph, we iterate several times the community detection algorithm at each Markov time. This set of partitions are used to obtain a stable result: considering all the partitions obtained at a fixed t, we programmed a consensus algorithm (section 3.2.5) which returns the more stable partition. Therefore the evolution of the partition through Markov times that we obtain is the most stable one.

With the increase of the Markov time, the number of communities usually decreases since a stable partition is obtained. Therefore the goodness of a partition can be evaluated both looking at the stability value 3.8 at each Markov time, and at the duration of that partition during the Markov chain.

For these reasons we evaluated both the evolution of the partitions and the evolution of the stability with the increase of the Markov time. This evaluation is performed separately for each layer.

We wrote an algorithm which allows to display the evolution of partitions, this program fixes a colour for each community, the final plots are reported in figures 4.5, 4.6 and 4.7 . The first figure (4.5) displays the evolution of dataset 1 (187 individuals ×107 compounds), the second (fig 4.6) shows the communities detected for dataset 2 with 418 individuals and 107 compounds, while the latter concerns the dataset 2 with 418 individuals and 137 compounds. In these images we use same colours for communities which are composed by roughly the same compounds. In particular, compounds of the blue community are mostly VLDL measures, compounds of the cyan community are mainly IDL an LDL cholesterol measures and the green community is composed by good cholesterol (HDL) measures. Differences between the two layers are located in the orange and dark-green communities. The orange community belongs only to the obese layer and it is mainly composed by metabolites. The dark-green community lies only in the normal-weight layer and it is constituted by medium cholesterol measures (MHDL-). From a biological point of view, both the green and the dark-green communities belong to good cholesterol measures, but they differ by the diameter of the analysed good cholesterol.

Looking at the hypergeometric test results, both compounds of the orange and dark-green communities are enriched in the obese layer (table 4.4). We explain this apparent contrast studying the adjacency matrix of the normal layer. We noted that nodes belonging to the MHDL- community (dark-green) in the normal-weight layer are very highly related together, but they have few connections to other good cholesterol measures (green community). In the obese layer, compounds of the (MHDL-) community are less related together, and they form a bigger community with the other good cholesterols (green community). Therefore the (MHDL-) nodes have more links in the obese layer, even if they are less heavy. At high Markov times, the dark-green community of the normal layer is joined to the green community, forming the big community of the good cholesterol measures (green). All images illustrate a decrease in the number of communities with the increase of the Markov time, that fact is expected since t can be seen as a resolution parameter.

As mentioned earlier, an important parameter to estimate the goodness of the partition is the stability (3.8). We plot the evolution of the stability for both datasets in Fig. 4.8. In these figures the red line refers to the stability profile of the obese layer partition, while blue line indicates the stability profile of the normal-weight layer. Figure 4.8 a. refers to the dataset 1, fig. 4.8 b) is related to the dataset 2 with 107 compounds and Fig. 4.8 c) refers to the dataset 2 with 137 compounds. In table 4.5 we report the maximum stability value and the corresponding Markov time for each layer. For all the datasets, the maximum stability value is observed at small Markov time, that is when there are many communities. At these short times there is a high variability of the partitions, this fact is linked to the utilised optimization method.

Community comparison: Markov time selection

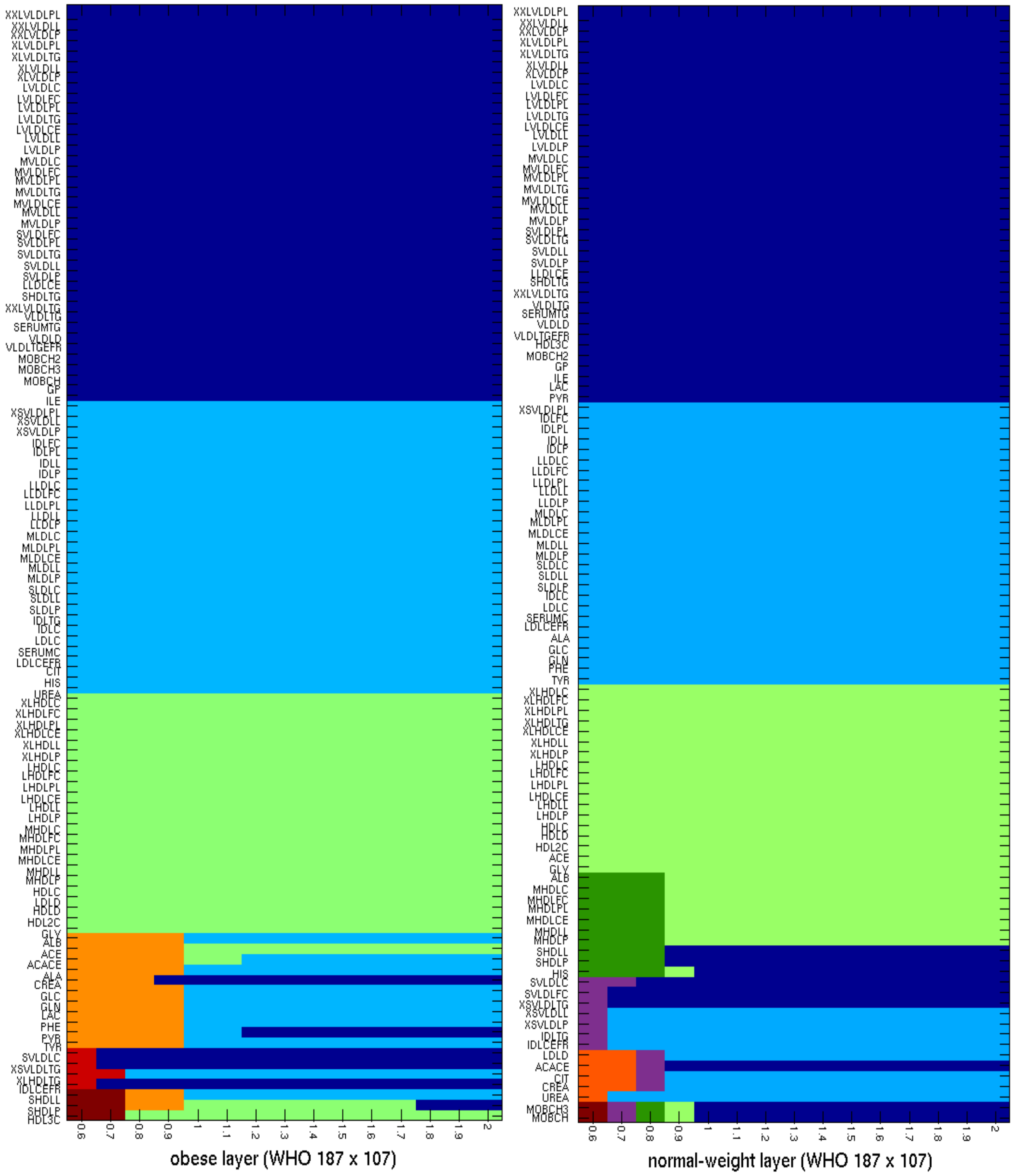
The comparison between communities of the two layers required the definition of shared communities. As can be seen from figure 4.5 both layers have the same number of communities along the Markov chain. Moreover, communities are composed by roughly the same compounds. We facilitate the visualisation of these shared communities colouring them with same colours. Also for the dataset 2 (figures 4.6 and 4.7) there is a positive overlap between communities of the two layers.

Dataset	Layer	$t(max\,stability)$	max stability
WHO 187 × 107	obese	0.6	0.52
	normal	0.7	0.49
WHO 418 × 107	obese	0.6	0.485
	normal	0.6, 0.7	0.48
WHO 418 × 137	obese	0.6, 0.7	0.45
	normal	0.7	0.45

Table 4.5: Maximum stability value and the corresponding Markov time for each layer and each dataset.

We quantitatively estimate the overlap between communities in different layers using the *normalised mutual information* (NMI), which is described in section 3.2.6. This measure of similarity ranges between 0 and 1, a value equal to 0 means that the two compared sets are totally dissimilar, while a value of 1 indicates that they are identical. The community detection method which we utilise gives different partitions according to the Markov time, therefore we compute the NMI between the partitions of the two layers at each t. The NMI value at each Markov time is illustrated by the green line in figure 4.8. For both datasets, the NMI has a depletion at small Markov times, then it increases according to t. That agrees with the informations that we can deduced from figures 4.5 and 4.6: at low times, four communities are detected, but with the increment of the Markov time one of them is lost (the orange one for the obese layer and the dark-green for the normal one). Nodes belonging to the orange and the dark-green communities are merged to the other communities, causing the increase of the partitions overlap at high Markov times. The same analysis can be done for the dataset 2 with 137 compounds (figure 4.7): at very low Markov times five communities are discover in both layers, but two of them are untied at higher times. We keep the same colours for communities which share approximately the same nodes and, as for the dataset 1, also for the dataset 2 the orange and the dark-green communities are untied. Therefore nodes of the orange and dark-green communities result to be interesting for the investigation of differences between the communities of the two layers.

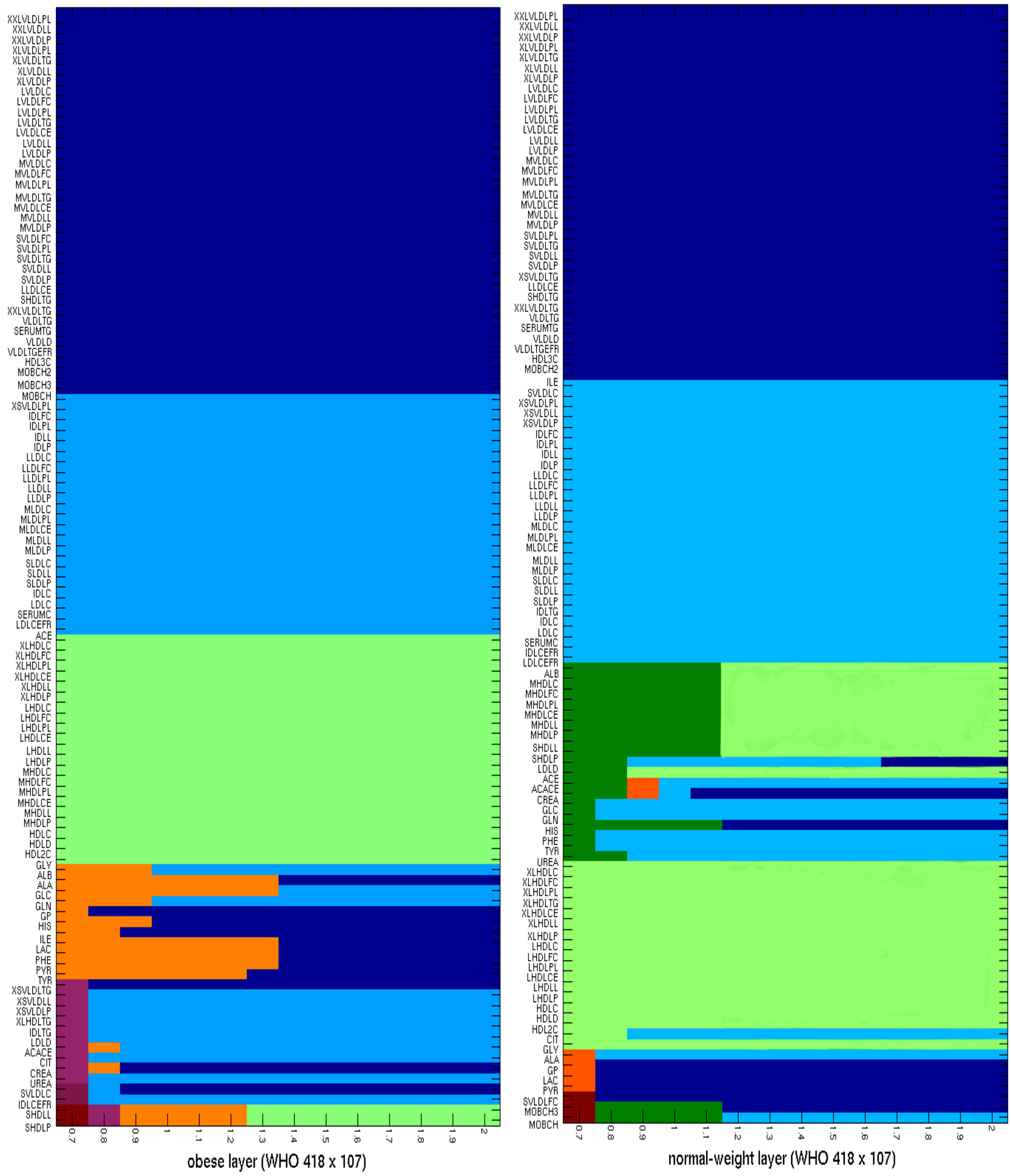
Datasets with 107 compounds The aims of the community analysis is to evaluate if there are significant differences between the partition of the obese layer and that of the normal-weight one. In the previous section we compared the partitions of the two layers, which are detected separately for the two groups. Results show that different



a) b)

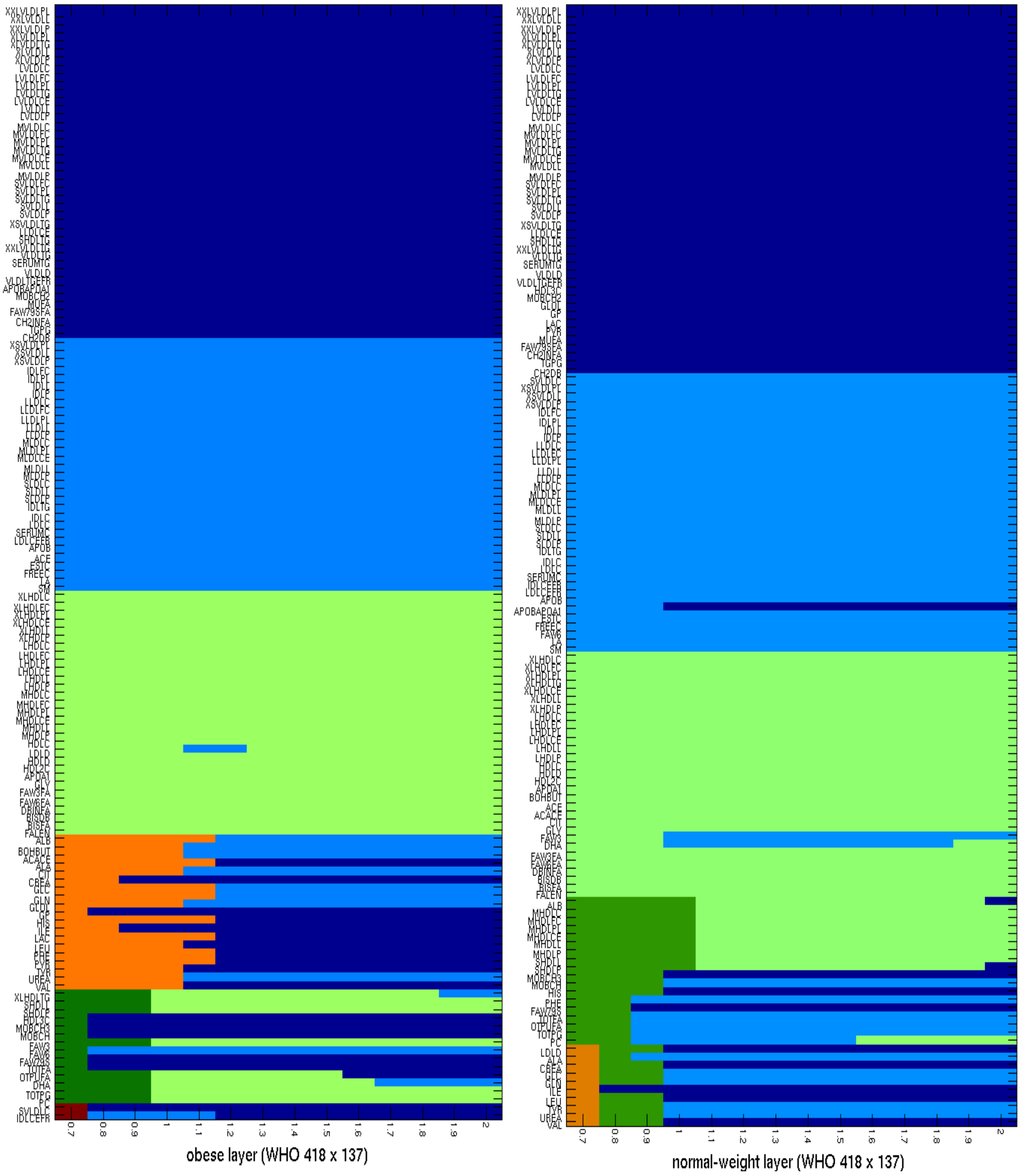
Figure 4.5: Evolution of communities with the increase of the Markov time for the dataset 1 (187 × 107), using the WHO classification. Figure a) is related to the evolution of communities in the obese layer, figure b) for the evolution in the normal-weight one.

communities are detected for the two layers. Moreover, there is an agreement with the results obtained with the dataset 1 (187 × 107) and the dataset 2 restricted to 107



a) b)

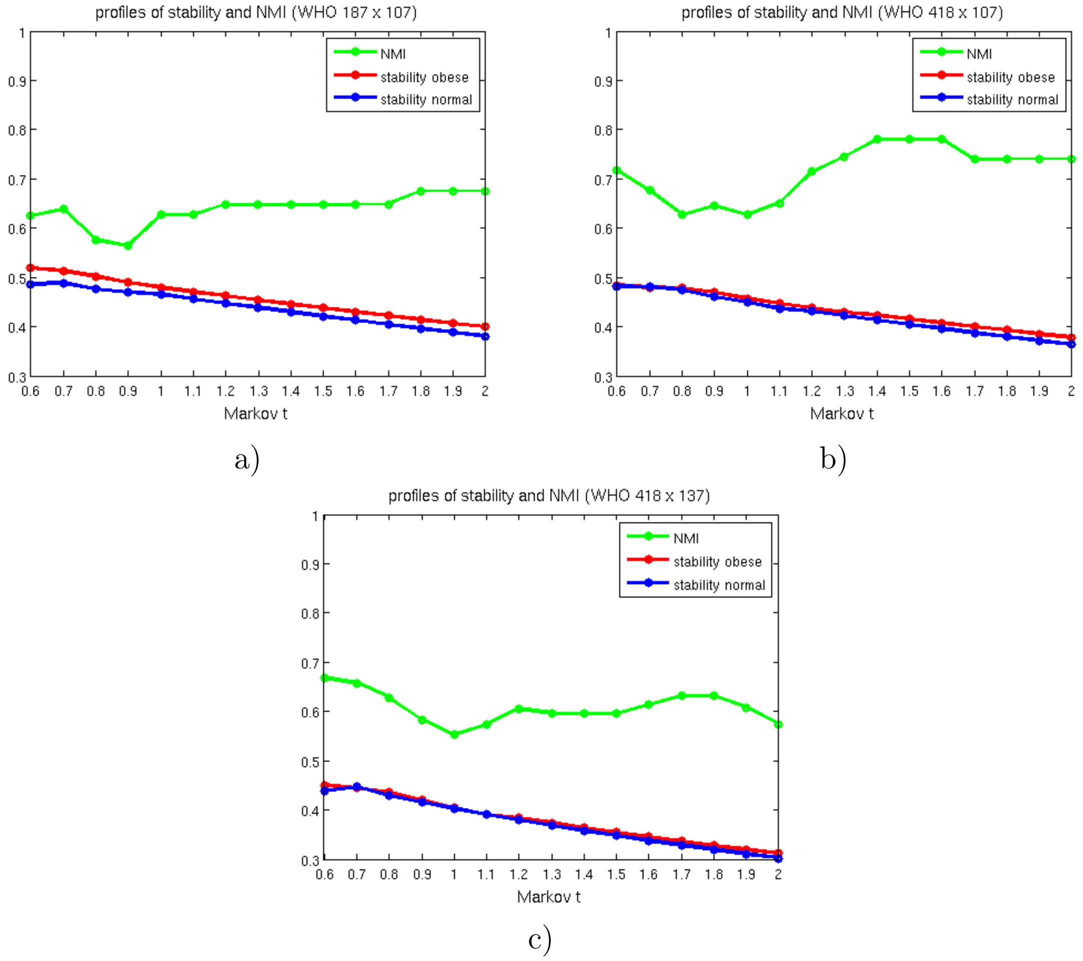
Figure 4.6: Evolution of communities with the increase of the Markov time for the dataset 2 restricted to 107 metabolites, using the WHO classification. Figure a) is related to the evolution of communities in the obese layer, figure b) for the evolution in the normal-weight one.



a) b)

Figure 4.7: Evolution of communities with the increase of the Markov time for the dataset 2 with 137 metabolites, using the WHO classification. Figure a) is related to the evolution of communities in the obese layer, figure b) for the evolution in the normal-weight one.

compounds (418 × 107). We compute the NMI between the partitions of the obese layer for the two datasets, that is we consider the obese layer of the dataset 1 and the obese



a)

b)

c)

Figure 4.8: Stability and normalised mutual information (NMI) profiles. Red line refers to the stability profile of the partition of obese layer, blue line for stability profile of the normal-weight layer. Green line shows the NMI evolution between the partitions of the two layers, along the Markov chain.

Dataset	$t(min\,NMI)$	min NMI	$t(max\,NMI)$	max NMI
WHO 187 × 107	0.9	0.56	1.8; 1.9; 2.0	0.68
WHO 418 × 107	1.0	0.63	1.4; 1.5; 1.6	0.78
WHO 418 × 137	1.0	0.55	0.6	0.63

Table 4.6: Maximum and minimum NMI value and the corresponding Markov time for each layer and each dataset.

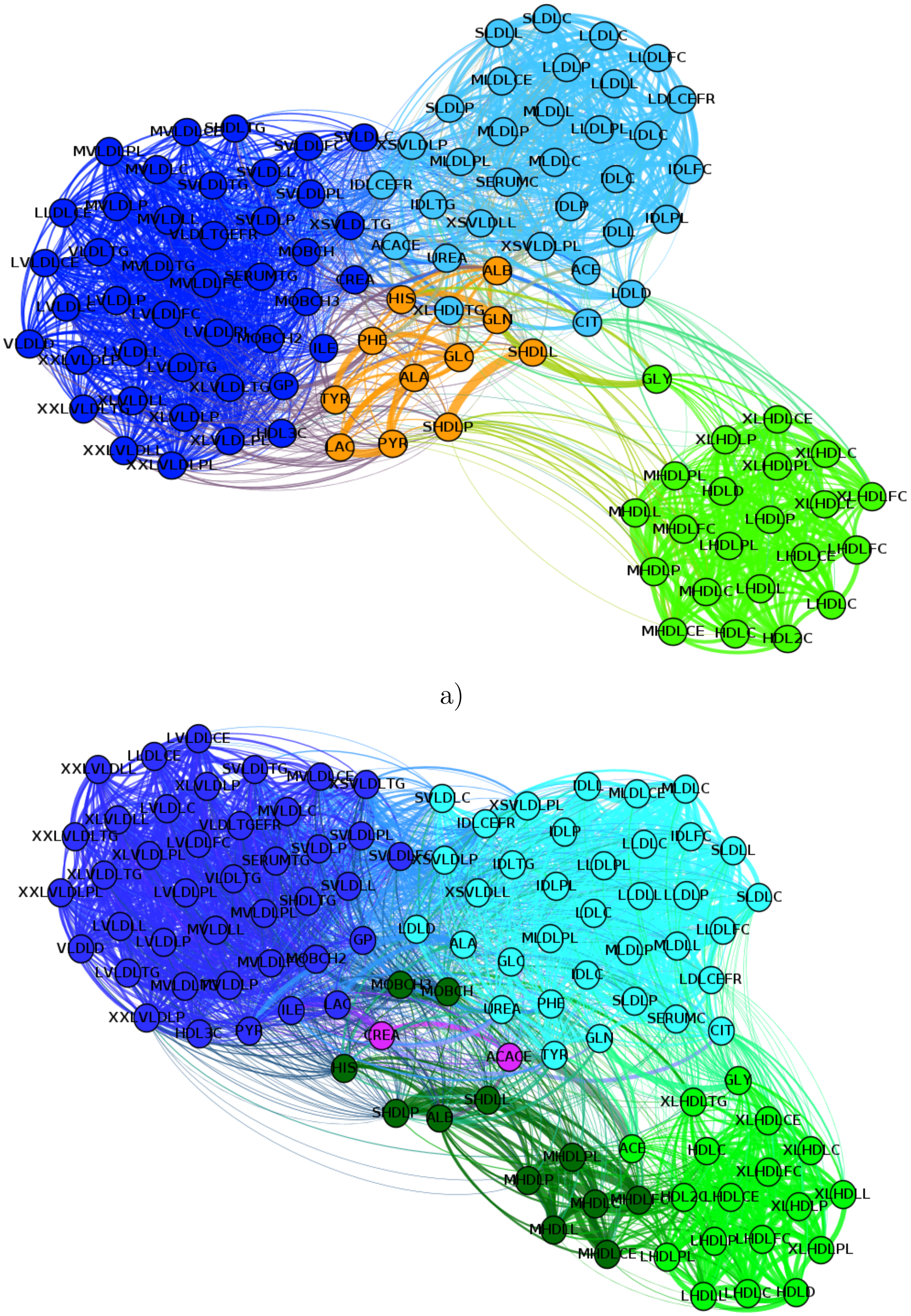
layer of the dataset 2 with 107 compounds, the same is done for the normal layer. The maximum overlap of the obese layer partitions is obtained when t=0.9 for both datasets; the mutual information between these partitions is $NMI_{obese} = 0.83$. The overlap of the normal-weight layers is greater, since the NMI results to be $NMI_{normal} = 0.93$ considering the partition at $t = 0.8$ for the dataset 1 and the partition at $t = 0.9$ for the dataset 2 .

Given the high agreement of the partition results in the two datasets, we display that with 107 metabolites and 418 individuals. We consider the partitions obtained at $t = 0.9$ for both the obese and the normal layers of the dataset 2 with 107 compounds; these are quite stable partitions, since they persist for different resolution scales. A graphic representation of the two layers is displayed in figures 4.9.a) and 4.9.b), which represent the graphs of the obese and normal-weight layer respectively. Node colours are associated to the communities detected at Markov time =0.9, which are listed in table 4.7. The layout underlines the communities found by the algorithm. In fact, nodes belonging to the same community form clusters in the images. Moreover, the orange and the dark-green communities are located between bigger communities, to which they are merged at high Markov times. We built these images using the Gephi software, which is shortly presented in section 3.3. To simplify the visualization we delete edges with weight lower than z=2.0, moreover we utilize the *Force Atlas* layout to accost nodes highly related.

In table 4.7 we report the compounds of each community at $t = 0.9$. The blue and cyan communities have a good overlap, since they differ by almost 10%. The differences are due to the compounds of the orange and dark-green communities: in the obese layer the dark-green community does not assemble and it is joined to the green community. On the other hand, the orange community is grouped only in the obese layer, while it's elements are distributed in various communities in the normal layer.

We evaluate the significance of these differences looking at the mean weight of communities. In particular we consider the blue, the cyan and the green communities equal for both layers, since their differences are caused by the elements of the orange and dark-green communities. Therefore we look at the intersection of the blue communities of layer 1 and 2, and the same for the cyan and green clusters. The final clusters which we obtain from the intersection are marked by a line in table 4.7.

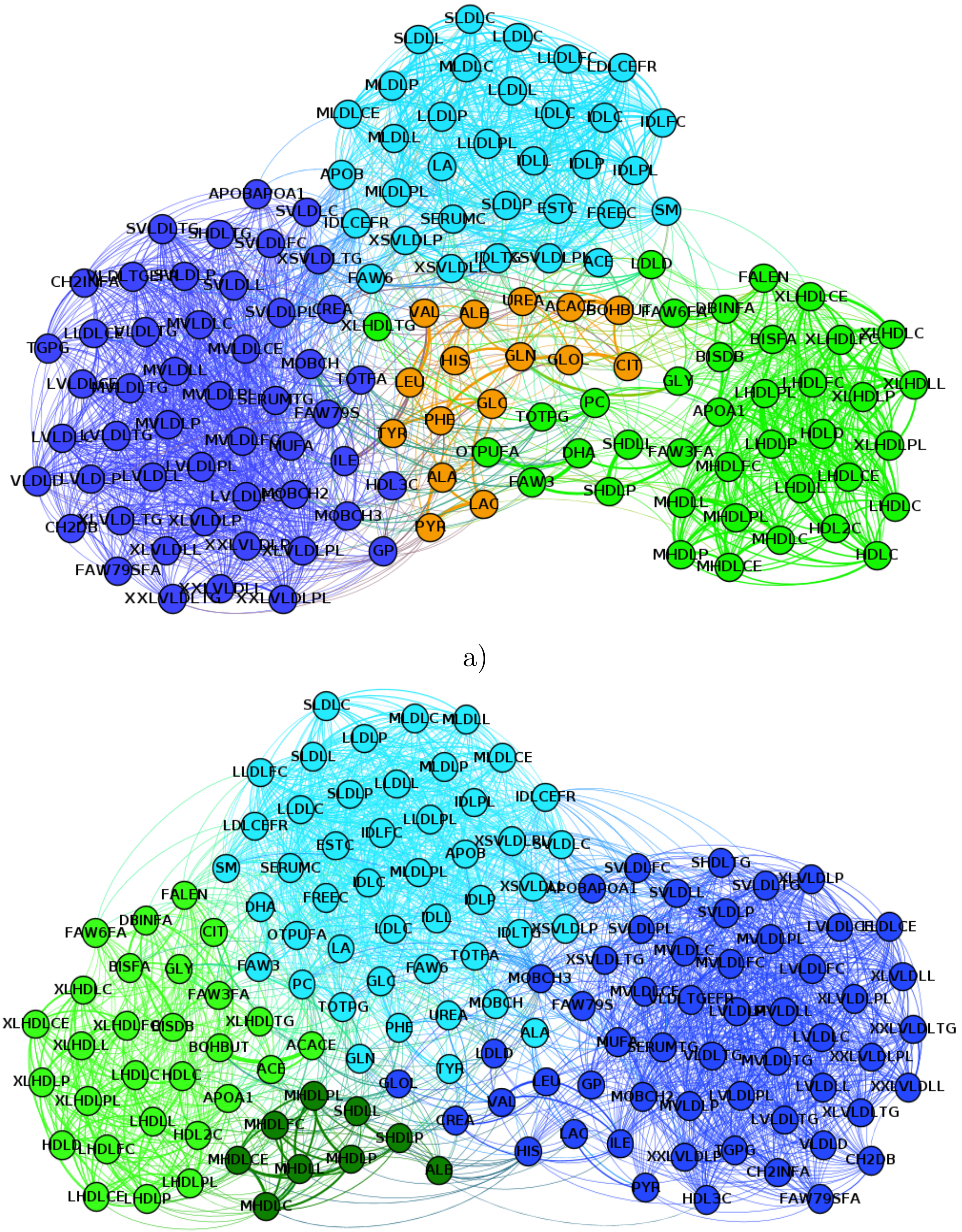
Datasets with 137 compounds The dataset 2 is composed by 418 individuals and 137 compounds. The main differences towards the dataset with 107 compounds are related to fatty acids and some metabolites nodes, which are not present in the dataset 1. Looking at the evolution of the partition (figure 4.7), these *new* fatty acids nodes are initially joined to the dark-green partition in the normal-weight layer, while in the obese layer they form a separate group. This happens at very low Markov times, then



a)

b)

Figure 4.9: Graph of the obese layer (a) and normal-weight layer (b), considering the dataset 2 with 107 compounds and using classification into obese and normal layer based on the WHO thresholds. Colours specified the communities detected at Markov time =0.9.



a)

b)

Figure 4.10: Graph of the obese layer (a) and of the normal-weight layer (b), considering the dataset 2 (418 × 137) and using classification into obese and normal layer based on the WHO method. Colours specify the communities detected at Markov time =1.0.

blue		cyan		green		orange	dark-green
obese	normal	obese	normal	obese	normal	obese	normal
XXLVLDLPL	XXLVLDLPL	XSVLDLPL	XSVLDLPL	XLHDLC	XLHDLC	ALB	ALB
XXLVLDLL	XXLVLDLL	XSVLDLL	XSVLDLL	XLHDLFC	XLHDLFC	SHDLL	MHDLC
XXLVLDLP	XXLVLDLP	XSVLDLP	XSVLDLP	XLHDLPL	XLHDLPL	SHDLP	MHDLFC
XLVLDLPL	XLVLDLPL	IDLFC	IDLFC	XLHDLCE	XLHDLCE	ALA	MHDLPL
XLVLDLTG	XLVLDLTG	IDLPL	IDLPL	XLHDLL	XLHDLL	GLC	MHDLCE
XLVLDLL	XLVLDLL	IDLL	IDLL	XLHDLP	XLHDLP	GLN	MHDLL
XLVLDLP	XLVLDLP	IDLP	IDLP	LHDLC	LHDLC	HIS	MHDLP
LVLDLC	LVLDLC	LLDLC	LLDLC	LHDLFC	LHDLFC	LAC	SHDLL
LVLDLFC	LVLDLFC	LLDLFC	LLDLFC	LHDLPL	LHDLPL	PHE	SHDLP
LVLDLPL	LVLDLPL	LLDLPL	LLDLPL	LHDLCE	LHDLCE	PYR	MOBCH3
LVLDLTG	LVLDLTG	LLDLL	LLDLL	LHDLL	LHDLL	TYR	MOBCH
LVLDLCE	LVLDLCE	LLDLP	LLDLP	LHDLP	LHDLP		HIS
LVLDLL	LVLDLL	MLDLC	MLDLC	HDLC	HDLC		
LVLDLP	LVLDLP	MLDLPL	MLDLPL	HDLD	HDLD		
MVLDLC	MVLDLC	MLDLCE	MLDLCE	HDL2C	HDL2C		
MVLDLFC	MVLDLFC	MLDLL	MLDLL	GLY	GLY		
MVLDLPL	MVLDLPL	MLDLP	MLDLP	MHDLC	ACE		
MVLDLTG	MVLDLTG	SLDLC	SLDLC	MHDLFC	XLHDLTG		
MVLDLCE	MVLDLCE	SLDLL	SLDLL	MHDLPL			
MVLDLL	MVLDLL	SLDLP	SLDLP	MHDLCE			
MVLDLP	MVLDLP	IDLTG	IDLTG	MHDLL			
SVLDLFC	SVLDLFC	IDLC	IDLC	MHDLP			
SVLDLPL	SVLDLPL	LDLC	LDLC				
SVLDLTG	SVLDLTG	SERUMC	SERUMC				
SVLDLL	SVLDLL	LDLD	LDLD				
SVLDLP	SVLDLP	IDLCEFR	IDLCEFR				
XSVLDLTG	XSVLDLTG	LDLCEFR	LDLCEFR				
LLDLCE	LLDLCE	CIT	CIT				
SHDLTG	SHDLTG	UREA	UREA				
XXLVLDLTG	XXLVLDLTG	ACACE	GLC				
VLDLTG	VLDLTG	ACE	GLN				
SERUMTG	SERUMTG	XLHDLTG	PHE				
VLDLD	VLDLD		TYR				
VLDLTGEFR	VLDLTGEFR		ALA				
HDL3C	HDL3C		SVLDLC				
MOBCH2	MOBCH2						
GP	GP		ACACE				
ILE	ILE		CREA				
SVLDLC	LAC						
MOBCH3	PYR						
MOBCH							
CREA							

Table 4.7: Communities found using the optimisation of the stability on the dataset 2 with 107 compounds. The Markov time is 0.9. Colours refer to the image 4.6.

they are merged to the blue, cyan, and green communities. therefore they seem to not affect the partition which we found for the datasets with 107 compounds. The *new* nodes are grouped in the orange community as concerns the obese layer, while they are merged to the blue and cyan communities in the normal-weight layer. In order to evaluate differences between communities of layers with 137 nodes, we chose the partitions looking at the NMI. As reported in table 4.6, the minimum overlap occurs at t=1.0 . Therefore we consider the two partitions at this Markov time. At this time, both layers have four communities and the fatty acids are grouped in the blue, cyan and green communities, and they exhibit similar behaviours. The *new* nodes are joined to the orange community in the obese layer, while the dark-green community of the

Community	median obese	mean obese	median normal	mean Normal	p-val Rank Sum	H_0
dataset 1 (187 × 107)						
Blue	1.68	1.6 ± 0.5	1.49	1.5 ± 0.5	0.000	1
Cyan	1.75	1.9 ± 0.9	2.10	1.9 ± 0.9	0.659	0
Green	2.57	2.6 ± 0.8	3.00	2.9 ± 0.5	0.000	1
Orange	2.08	2.2 ± 1.7	0.57	1.0 ± 1.0	0.003	1
Dark-green	1.37	1.9 ± 1.9	2.05	2.6 ± 1.9	0.019	1
dataset 2 (418 × 107)						
Blue	1.56	1.5 ± 0.5	1.54	1.5 ± 0.5	0.585	0
Cyan	1.79	1.8 ± 0.9	2.06	1.9 ± 0.8	0.068	0
Green	2.73	2.7 ± 0.7	3.01	2.9 ± 0.7	0.047	1
Orange	1.87	2.08 ± 1.5	0.91	1.2 ± 1.2	0.012	1
Dark-green	0.89	1.5 ± 1.7	1.54	2.4 ± 1.7	0.000	1
Green+ dark-green	2.1	1.8 ± 1.3	1.4	1.7 ± 1.5	0.33	0

Table 4.8: Median and mean edges weight values of each community and median test between layers. The null hypothesis is that edges of the two layers of a fixed community comes from the same population. The significance level is set at 0.05.

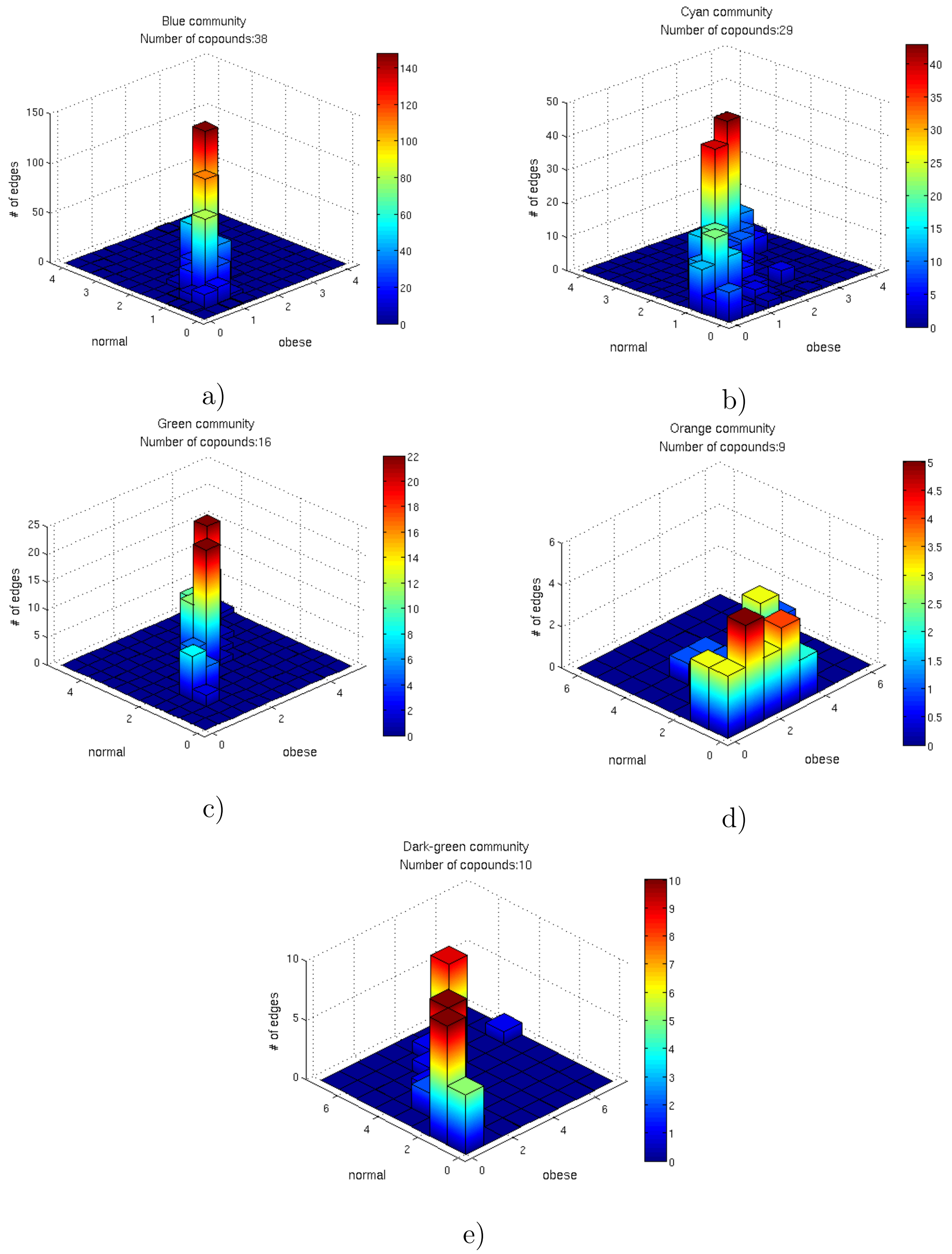
normal layer has approximately the same elements of the dataset with 107 compounds. The orange community is then composed by 16 metabolites and amino acids, which are: albumin, alanine, 3-hydroxybutyrate, acetate, acetoacetate, citrate,glucose, glutamine, glycerol, histidine, lactate, leucine, phenylalanine, pyruvate, tyrosine, urea and valine.

As for the graph with 107 nodes, we display the community structures of these layers using the software Gephi, and the *ForceAtlas* layout. The two graphs are displayed in figure 4.10.

Community comparison

This almost total overlap of the blue, cyan and green communities allows to analyse differences between the same communities in different layers. The disagreement lying in the orange and dark-green communities, since they share only 4 compounds, therefore we evaluate these clusters in a different way.

A preliminary evaluation of the differences which occur in the orange community is done looking at figure 4.4. This graph shows the connections between all the metabolites of the DILGOM dataset and it is obtained form the dataset 2 with 137 compounds. We use this graph to give a qualitative evaluation of the orange community since there are not huge differences between the metabolites connections of dataset 2 and that of



a)

b)

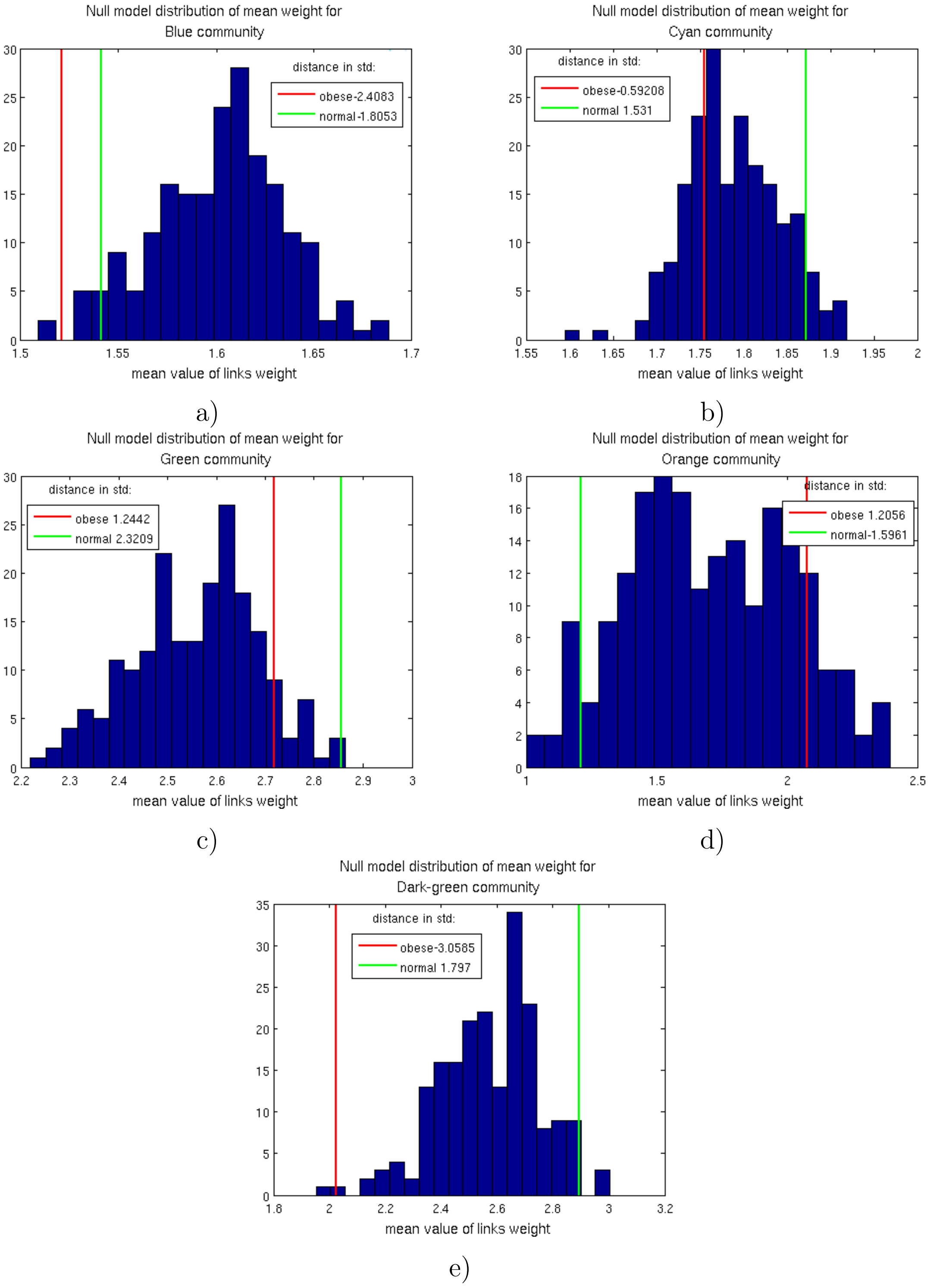
c)

d)

e)

Figure 4.11: Bivariate distribution (A^{obese}, A^{normal}) for each community. X and Y axis represent edges weights, while Z axis displays the number of edges. Bars on the diagonal indicate edges with the same weight belonging to both layers. Asymmetries reveal differences between the weight distribution of the two layers. Figure 4.11 a) shows the bivariate distribution of the blue community, b) of the cyan community, c) of the green community, d) of the orange community and e) of the dark-green one.

dataset 2 restricted to 107 compounds. In that graph red lines refer to edges which occur only in the obese layer (L_{10}), blue lines show links lying only in the normal-weight layer (L_{01}) and green lines show shared links (L_{11}). There is a clear majority of red lines, in particular if we restrict the nodes to those belonging to the orange community. In order to quantified these differences ad similarities, we perform a Wilcoxon rank sum test (section 3.2.7) between communities of the two layers. In particular we consider the intersections of the blues, cyans and greens communities, while we consider separately the orange and the dark-green communities. Since the orange community is grouped only in the obese layer, we extract the corresponding nodes in the normal layer and quantify the differences of their links weight distributions. The same is done for the dark-green community. Table 4.8 shows the results of the median test for the dataset 1 and for the dataset 2 with 107 compounds. The null hypothesis is that edges of the two layers of a fixed community come from the same population, and the significance level is set at 0.05. The mean and the median weight of edges in the orange community are significantly higher in the obese community, while the mean and the median edges weight of the dark-green community are significantly greater in the normal layer. Also the green community is significantly higher related in the normal layer. When we consider the total cluster of good cholesterol (green+ dark-green communities) there are not significant differences between the two layers. In table 4.8 we report also the mean and the standard deviation values of the edge weight distribution of each community. The median test indicates that there are significant differences between edges weights which connect nodes of the orange and dark-green communities in the two layers. Considering the orange community, we extract the adjacency matrix of *orange* nodes of the two layers (A^{ob}_{Or}, A^{nor}_{Or}) and we plot a bivariate distribution of links of the orange community. The x and the y axes of these distributions indicate the edges weight, while the z axis displays the number of edges. In particular the x axis refers to the obese layer and the y axis to the normal-weight one. The bivariate distribution considers the weight of the ij edge in the two layers: $a^{ob}_{Or}(ij)$ and $a^{nor}_{Or}(ij)$; if its weights are similar in the two layers it will be set on the diagonal of the discrete bivariate distribution. If $a^{ob}_{Or}(ij) > a^{nor}_{Or}(ij)$ it will be set in a bin of the right half of the histograms in figure 4.11; vice versa, if its weight is higher in the normal layer it will be set in the left half. Therefore asymmetries reveal differences between the weight distribution of the two layers. Figures 4.11 display the bivariate distributions of the blue, cyan, green, orange and dark-green communities. The orange distribution has a high asymmetry, in particular all bins are set in the right half of the bivariate distribution. Also in figure 4.11.e, which refers to the dark-green community, there is a high asymmetry, since all bins are in the left half of the bivariate distribution.



a)

b)

c)

d)

e)

Figure 4.12: Distribution of the mean weight of links of each community for random sample extraction. Red line indicates the mean weight of links of the obese layer, while green line of the normal one. Figure a) shows results of the blue community, b) of the cyan community, c) of the green community, d) of the orange community and e) of the dark-green one.

To evaluate the strength of the results we performed a comparison with random models. We randomly extract 85 individuals from the whole DILGOM dataset and we compute their adjacency matrix, as done for the two layers. In other words, we compute the correlation matrix using the Kendall's τ, then we compute the z-score matrix with the CLR algorithm (A_R). For each adjacency matrix we compute the mean edge weight of the five detected communities (blue, cyan, green, orange and dark-green). Considering the orange community as example, we extract from the *random* adjacency matrix A_R the nodes belonging to the orange community. Then we compute the mean edges weight of nodes of that community. This process is iterated 250 times, for each iteration we compute the mean edges weight of each community. The final distributions of the mean edges weight are displayed in figure 4.12. Figure 4.12.a shows the mean weight distribution of edges of the blue community, figure 4.12.b displays the distribution of cyan community, figure 4.12.c is related to green community and figure 4.12.d to the orange one. In each histogram we plot a red line to indicate the mean weight of links of that community for the obese layer, while the green line is set for the normal one. The distribution in figure 4.12.b seems to indicate that the mean weight of links of the cyan community is not related to the performed classification. In other words, compounds of the cyan community does not discriminate obese individuals form normal-weight ones. The blue community (figure 4.12.a) is highly related for the random samples; since we utilize the CLR algorithm, this distribution means that, for a random sampling, *blue* nodes are more significantly linked with respect to other compounds (the CLR considers the background distribution of each entry). The highest differences between the random distribution and the mean weight of the obese and normal-weight layer are linked to the green, dark green and orange communities. In particular, the mean weights of the orange community for the obese and normal layer are on opposite sides of the mean weight distribution. The mean weight of the edges of the obese layer is higher than that of the normal one, as reported in table 4.8; this result is displayed in figure 4.12.d. The opposite situation occurs for the dark green community, where the mean weight of the edges of the normal layer is significantly higher than that of the obese one.

These results provide a further confirm that there are some nodes which have different behaviours in the two layers. Specifically, these nodes are medium good cholesterol measures (MHDL-) and some blood metabolites. The former are highly related for normal individuals, while the latter form a cluster with significant mean weight only in the obese layer.

This application of multiplex analysis on real data allowed to discover differences between the two layers. These results show the efficacy of the proposed method.

CHAPTER 5

Conclusions

In this thesis we analysed multiplex network structures, which belong to the *complex network theory*. This new approach permits to investigate more complex frameworks than the classical networks analysis. Indeed, multiplex network are "networks of networks", that is multiple levels of networks. Therefore each network is a layer of a more complex structure.

We focused our analysis on the characterization of some multiplex properties. The final aim is to develop null hypothesis which can be applied to statistical analysis. We studied some methods which permit to evaluate intra-layer structures and inter-layers differences.

To discover intra-layer structures we implemented a community detection method, which detects subgroups of network elements which are strongly bonded together. These sub-groups of highly tied nodes are commonly called *clusters* or *communities*. We perform a clustering method which utilises a quality function called *stability*. It evaluates whether the concentration of edges within clusters is significant when compared with a random distribution of links. An important aspect of this clustering method is that it permits an evaluation of the layer partition at different partition scales. The inner resolution parameter is represented by the Markov chain. Therefore, the stability of a graph considers the graph as a Markov chain where each node represents a state and each edge a possible state transition. The overlap between communities of different layers is quantified using the *normalised mutual information*. This measure quantifies how similar or different two partitions are.

The evaluation of inter-layers differences concerned also single nodes connections. Indeed, since all layers have the same nodes, differences between layers lay in their edges. Therefore we perform a hypergeopetric test to evaluate whether there is a significant enrichment or depletion in the number of layer-specific edges of a node. This method shows that only with a high overlap between edges laying in different layers some nodes are significantly enriched. When the overlap is weak, the nodes do not result to be significantly over-represented in a specific layer. Therefore this method emerges the differences between similar layer, looking at the at the layer-specific, node-specific and shared edges.

These analyses allow to characterise multiplex properties and especially intra-layer structures and intra-layer differences.

We applied this analytical method to real data. In particular, these data are collected by the DLIGOM *the Dietary, Lifestyle and Genetic determinants of Obesity and Metabolic syndrome* study and they regard metabolomic, transcriptomic, phenotypic and genomic information of a Finnish cohort. These datasets are utilised by the *Mimomics* project in order to identify some factors that might be related to obesity. The final goal of the *Mimomics* project is to reveal if there are significant differences in one or more omics, which are genomic, transcriptomic and metabolomic, looking at the differences which are observed in an another omic, that is the phenotype of obesity in this specific case. This mulit-omic approach is an innovative way which evaluates if two or more omic are related to each other.

We focus our analysis on the metabolomic, which regards the concentrations of blood serum compound of the Finnish cohort. These metabolic data are analysed using complex network analysis and, in particular, the multilayer approach. We built a multiplex network in order to stress the metabolic differences between obese and normal-weight people, which are classified on the basis of their phenotypes. This multiplex network is made up of two layers, one linked to obese individuals and the other one to normal-weight individuals. The layers are constituted by the same number of nodes (n), which represent the compounds extracted from the blood of each individual by the DILGOM study. An edge between two compounds specifies a positive correlations between them, and its weight is a measure of the correlation strength. We utilized a statistical method to avoid indirect relationships and enhance significant interactions. This method produces z-score positive matrices, that are set as adjacency matrices.

The assessment of the properties of this real multiplex network was achieved implementing the statistical analysis which we develop. Specifically, the results of the hypergeometric test show that there are some nodes which have different behaviour in the two layers. We noted a significant enrichment of the links of some metabolites and of some measures related to the medium good cholesterol (MHDL-), which seem to have different behaviours in the two layers.

These results are confirmed by the community detection. This method discovered a module which is composed by highly related metabolites, which lies only in the obese layer. These metabolites are approximately the same which result enriched in the previous analysis. This evidence suggests that those metabolites are not only highly related, but they are highly related to each other. The measures related to the medium good cholesterol gather together only in the normal-weight layer, while other communities have a perfect overlap between the two layers.

It must be stressed that these results are linked to the multilayer structure we built. In fact, a simple analysis of the blood concentrations does not bring to light differences between the concentrations of those compounds. Specifically, the concentrations of good medium cholesterol measures and some metabolites are the same for both groups of individuals.

In conclusion this multiplex analysis on real data shows the efficacy of the proposed method. Since the implemented method is totally independent of the data to analyse, it will be possible to extend this procedure to other fields of study.

.

Appendix A: Generalized linear model

We consider a second method for the classification of the individuals, in order to form the obese and the normal-weight group. This method is based on the *generalised linear model* (glm).

The Linear Regression Model, (lm) [63] is a statistic technique which studies linear, additive relationships between variables. More precisely, the goal of lm is to predict values of a scalar dependent variable y called the *criterion variable*, given n variables $\mathbf{X} = x_1, ..., x_n$ called *predictors or explanatory variables.* We consider as predictors some features of the phenotypes dataset, that are: genders, age, bad and good cholesterol (LDL and HDL), fasting glucose, systolic and diastolic pressures and the consumption of blood pressure medications. The response variable y was considered the BMI . When there are more than one predictors ($n > 1$), as in our case, the process is called *multiple linear regression.* A multiple linear regression model is:

$$y(i) = \beta_0 + \beta_1 x_1(i) + \cdots + \beta_n x_n(i) + \varepsilon_i = \beta_0 + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n, \qquad (5.1)$$

where the vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_n)$ expresses parameters called *regression coefficients* and the variable ϵ, named *noise or error variable*, represents the random error. Usually the noise terms are assumed to be uncorrelated and to have independent and identical normal distributions with mean zero and constant variance. Moreover, the responses y_i are assumed to be uncorrelated. Given these assumptions, the fitted linear function becomes:

$$\tilde{y}_i = b_o + \sum_{k_i}^{K} b_k f_k(x_1(i), x_2(i), ..., x_n(i)) \qquad i = 1, \ldots, n, \qquad (5.2)$$

Where $\tilde{y}_i$ is the estimated response and f is a scalar-valued function of the predictors. The function f might be non linear or polynomial, since the linearity is due to the coefficients β_k of equation 5.1; that is, the response variable, y, is a linear function of the coefficients β_k. The coefficients b_k are the fitted coefficients, which are computed minimising the mean square difference between the prediction vector $bf(x)$ and the true response vector y: $\tilde{y} - y$.

The difference between the observed value and the estimated value of the quantity of interest $\tilde{y} - y$ is the *residual* or *fitting deviation* of the criterion variable. The residuals, in our analysis, are the differences between the actual BMI of individuals and their estimated BMI ($\tilde{y}$) using the lm. Since they are used for the estimation of the fitted coefficients, the sum of residuals of a random sample is zero.

An extension of the lm are the *generalised linear models* (glm). These models consist of three components:

- A *random component*, which specifies the conditional distribution of the criterion variable (y_i), given the predictors $\mathbf{X} = x_1, ..., x_n$. The distribution of y_i can be Gaussian, binomial, Poisson, gamma and so on.
- A *linear predictor* which performs a linear regression, as for the lm (equation5.2).
- a smooth and invertible linearizing *link function* $g(\tilde{y})$, which transforms the expectation of the response variable $\tilde{y}$ to the linear predictor:

$$g(\tilde{y}_i) = b_o + \sum_{k_i}^{K} b_k f_k(x_1(i), x_2(i), ..., x_n(i)) \qquad i = 1, \dots, n, \tag{5.3}$$

Therefore, generalised linear models allow to perform a linear regression for criterion variables which have arbitrary distributions, moreover they introduce the link function to relate the linear model to the response variable.

We use a glm to implement a regression on our data. In particular, the *link function* is set as the logarithm and the performed regression is:

$$\begin{aligned}\log BMI =& lm(gender, age, hypertension\, medications,\, log(fasting\, glucose),\\ & systolic\, pressure,\, diastolic\, pressure,\, total\, LDL,\, total\, HDL),\end{aligned}$$

For each individual the corresponding residual is computed ($log(BMI) - log(\widetilde{BMI})$), obtaining a residual vector with length equal to the number of individuals (m).

We computed the number of normal-weight (n_{nor}) (BMI$\leq$ 25) and obese individuals (n_{ob}) (BMI $\geq$ 30), in order to classify people into a group with an increased CVD risk and a group with normal CVD risk. We sorted the m residuals and grouped individuals with the n_{ob} highest residuals in the increased CVD risk class people. In the same way we compose the normal CVD risk group, which is formed by the individuals with the n_{nor} lower residuals.

dataset	# obese	mean BMI	# normal- weight	mean BMI	total # of people in the dataset
1)	33	32.8	84	23.4	187
2)	79	32.8	172	23.5	418

Table 5.1: Number of individuals classified as obese and normal-weight using the glm method for the two datasets described in section 4.1.

Results with generalised linear model classification

As already said, we consider as predictors the following phenotypes: genders, age, total and good cholesterols (LDL+HDL and HDL), fasting glucose, systolic and diastolic pressures and the consumption of blood pressure medications. The response variable is considered the BMI and we use as *link function* the logarithm. The number of individuals labelled as obese abides by the percentage of people with BMI greater than or equal to 30. More precisely, we include in the obese group people with residuals greater than a threshold value, which is established as the percentile associated to the percentage of non-obese people. The residual of one individual is the difference between his actual BMI and the $\widetilde{BMI}$ estimated by the glm. In the same way we classify normal-weight individuals, considering normal weight people characterised by $BMI \leq 25$. We applied this classification method on both the datasets described in section 4.1:

1. For dataset 1 (the dataset composed by 187 individuals and 107 compounds) the percentage of normal-weight individuals ($BMI < 25$) is 45% and that of obese individuals ($BMI > 30$) is 17%.

2. For dataset 2 (the dataset composed of 418 individuals and 137 metabolites) the percentage of normal-weight individuals is 41% and that of obese individuals is 19%.

We report in table 5.1 the number of individuals that are grouped in the obese and normal-weight groups; for each group we associate its mean BMI.

We display the main phenotypic differences between normal and obese groups in table 5.2, where the mean and standard deviation values of some phenotypes features are reported.

We also perform a median test (see section 3.2.7) for quantifying the phenotype differences between the obese and the normal-weight clusters.The null hypothesis is that the two groups belong to the same population, which means that their phenotypes are not significantly different. The confidence value is set at $p = 0.05$ and a value of

dataset	class	BMI	LDL	HDL	age	systolic Pressure
1	obese	32.8 ± 4.5	5.5 ± 0.8	1.44 ± 0.43	54 ± 13	135 ± 16
	normal	23.4 ± 2.4	5.3 ± 0.98	1.5 ± 0.4	53 ± 12	130 ± 18
	non-ob	24.9 ± 3.1	5.30 ± 0.9	1.5 ± 0.4	53 ± 13	130 ± 18
2	obese	32.8 ± 4.9	5.3 ± 0.9	1.4 ± 0.6	52 ± 13	132 ± 15
	normal	23.5 ± 3.5	5.2 ± 0.97	1.4 ± 0.3	51 ± 13	131 ± 18
	non-ob	25.1 ± 3.1	5.7 ± 0.96	1.5 ± 0.4	51 ± 14	130 ± 18

Table 5.2: Mean and standard deviation values of some phenotypes features, related to normal and obese groups.

method	cass	BMI	age	LDL	HDL	fast Gl	sys P.	dias P	WHO	waist circ
glmBMI	1	1	0	0	0	0	0	0	1	1
	2	1	0	0	0	0	0	0	1	1

Table 5.3: Results of the median test (see sec. 3.2.7) between the obese and the normal-weight clusters, for the glm classification method. The null hypothesis is that the two groups belong to the same population, $H_0 = 1$ indicates that the null hypothesis is rejected. The confidence is set at $p = 0.05$.

$H_0 = 1$ indicates that the null hypothesis is rejected. In table 5.3 the results of the test are reported.

Both tables 4.2, 5.2 and 4.3, 5.3 show substantial differences between the cluster phenotypes which depend on the adopted classification method. In particular, the glm method forms two groups of individuals that are not significantly different with respect to some phenotypic parameters (age, good and total cholesterols, fasting glucose and systolic and diastolic pressures), but they have different BMI, whr and waist circumference.

Hypergeometric test results

We perform the hypergeopetric test to evaluate the most significant differences between single compounds of the normal-weight and the obese layers. The theoretical principles of this test are explained in section 3.2.3. The method is the same of that described in section 4.3.1 which is applied on the classification based on the WHO thresholds. As in that section, for each layer we consider the number of times that a specific compound results to be oversampled, since we perform the test for different thresholds. At the end we rank the compounds looking at their number of over-representations; the results for the different datasets and classification methods are listed in table 5.4.

glm BMI 187x107				glm BMI 418x107				glm BMI 418x137			
Obese		Normal		Obese		Normal		Obese		Normal	
ALB	20	SLDLL	13	LDLD	13	ALB	16	LDLD	14	CREA	14
LAC	20	CIT	11	ACACE	9	GLN	14	CREA	10	GLN	14
PYR	20	SLDLP	9	IDLTG	8	HIS	13	UREA	10	ALB	12
ALA	17	SLDLC	8	XLHDLTG	6	UREA	12	SM	10	UREA	11
PHE	17	XXLVLDLP	6	MHDLPL	6	CIT	9	ACACE	7	VAL	11
GLC	15	XLVLDLPL	6	MHDLP	5	ILE	8	PHE	7	ACE	10
TYR	11	XSVLDLTG	5	XSVLDLL	4	GP	5	SHDLP	6	CIT	10
XLHDLTG	10	MLDLL	5	SHDLL	4	TYR	5	IDLTG	6	FALEN	10
HIS	9	VLDLTGEFR	5	SHDLP	4	LVLDLC	4	GLC	6	GLC	7
GLY	8	ACACE	5	CIT	4	ACE	4	FAW3FA	6	GLOL	7
IDLTG	7	XLVLDLP	4	UREA	4	CREA	4	FAW6FA	6	HIS	5
ACE	6	LLDLC	4	XSVLDLPL	3	GLC	4	MOBCH	5	BISFA	5
ACACE	6	MLDLPL	4	MHDLL	3	XXLVLDLP	3	SHDLL	4	LDLD	4
UREA	6	MLDLP	4	XXLVLDLP	2	XLVLDLL	3	APOA1	4	TYR	4
MHDLPL	4	SERUMTG	4	LHDLC	2	LVLDLPL	3	LLDLPL	3	LA	4
XSVLDLPL	3	SERUMC	4	LHDLCE	2	LVLDLP	3	XLHDLTG	3	DBINFA	4
XSVLDLL	3	XLVLDLTG	3	HDLC	2	MVLDLP	3	MHDLPL	3	FREEC	3
MHDLL	3	XLVLDLL	3	ALA	2	VLDLTG	3	ESTC	3	XXLVLDLP	2
MHDLP	3	LVLDLC	3	MOBCH2	2	VLDLD	3	BISDB	3	LVLDLC	2
GLN	3	LVLDLFC	3	GLC	2	LDLD	3	LVLDLFC	2	XLHDLPL	2
XSVLDLP	2	MVLDLC	3	GLN	2	IDLCEFR	3	IDLFC	2	SHDLL	2
LDLD	2	MVLDLFC	3	GP	2	XLVLDLPL	2	LHDLC	2	APOB	2
XXLVLDLTG	1	MVLDLPL	3	ALB	1	LVLDLCE	2	HDL3C	2	GLY	2
XXLVLDLPL	0	LLDLL	3	LVLDLFC	1	MVLDLC	2	MOBCH2	2	FAW3	2
XXLVLDLL	0	LLDLP	3	LVLDLCE	1	MVLDLPL	2	GLN	2	XLVLDLPL	1

Table 5.4: Rank of DILGOM compounds based on p-values of hypergeometric test. The rank is obtained considering the results of hypergeometric test for different threshold values, in particular from z-value of 1.5 to 3.5 with steps of 0.1 using the glm classification method.

We noted that the results we obtain using the glm classification (tab 5.4) have a greater discrepancy between the dataset 1 and the dataset 2 (with 107 compounds). For this reason we choose to use only the WHO classification method for the multiplex analysis.

Appendix B: Compounds of the DILGOM study

Abbreviation	Full description	Unit
	Lipoprotein subclasses	
XXL-VLDL-PL	Phospholipids in chylomicrons and extremely large VLDL	mmol/L
XXL-VLDL-TG	Triglycerides in chylomicrons and extremely large VLDL	mmol/L
XXL-VLDL-L	Total lipids in chylomicrons and extremely large VLDL	mmol/L
XXL-VLDL-P	Concentration of chylomicrons and extremely large VLDL particles	mol/L
XL-VLDL-PL	Phospholipids in very large VLDL	mmol/L
XL-VLDL-TG	Triglycerides in very large VLDL	mmol/L
XL-VLDL-L	Total lipids in very large VLDL	mmol/L
XL-VLDL-P	Concentration of very large VLDL particles	mol/L
L-VLDL-C	Total cholesterol in large VLDL	mmol/L
L-VLDL-FC	Free cholesterol in large VLDL	mmol/L
L-VLDL-PL	Phospholipids in large VLDL	mmol/L
L-VLDL-TG	Triglycerides in large VLDL	mmol/L
L-VLDL-CE	Cholesterol esters in large VLDL	mmol/L
L-VLDL-L	Total lipids in large VLDL	mmol/L
L-VLDL-P	Concentration of large VLDL particles	mol/L
M-VLDL-C	Total cholesterol in medium VLDL	mmol/L
M-VLDL-FC	Free cholesterol in medium VLDL	mmol/L
M-VLDL-PL	Phospholipids in medium VLDL	mmol/L
M-VLDL-TG	Triglycerides in medium VLDL	mmol/L
M-VLDL-CE	Cholesterol esters in medium VLDL	mmol/L
M-VLDL-L	Total lipids in medium VLDL	mmol/L
M-VLDL-P	Concentration of medium VLDL particles	mol/L
S-VLDL-C	Total cholesterol in small VLDL	mmol/L
S-VLDL-FC	Free cholesterol in small VLDL	mmol/L
S-VLDL-PL	Phospholipids in small VLDL	mmol/L
S-VLDL-TG	Triglycerides in small VLDL	mmol/L
S-VLDL-L	Total lipids in small VLDL	mmol/L
S-VLDL-P	Concentration of small VLDL particles	mol/L

Abbreviation	Full description	Unit
XS-VLDL-PL	Phospholipids in very small VLDL	mmol/L
XS-VLDL-TG	Triglycerides in very small VLDL	mmol/L
XS-VLDL-L	Total lipids in very small VLDL	mmol/L
XS-VLDL-P	Concentration of very small VLDL particles	mol/L
IDL-C	Total cholesterol in IDL	mmol/L
IDL-FC	Free cholesterol in IDL	mmol/L
IDL-PL	Phospholipids in IDL	mmol/L
IDL-TG	Triglycerides in IDL	mmol/L
IDL-L	Total lipids in IDL	mmol/L
IDL-P	Concentration of IDL particles	mol/L
L-LDL-C	Total cholesterol in large LDL	mmol/L
L-LDL-FC	Free cholesterol in large LDL	mmol/L
L-LDL-PL	Phospholipids in large LDL	mmol/L
L-LDL-CE	Cholesterol esters in large LDL	mmol/L
L-LDL-L	Total lipids in large LDL	mmol/L
L-LDL-P	Concentration of large LDL particles	mol/L
M-LDL-C	Total cholesterol in medium LDL	mmol/L
M-LDL-PL	Phospholipids in medium LDL	mmol/L
M-LDL-CE	Cholesterol esters in medium LDL	mmol/L
M-LDL-L	Total lipids in medium LDL	mmol/L
M-LDL-P	Concentration of medium LDL particles	mol/L
S-LDL-C	Total cholesterol in small LDL	mmol/L
S-LDL-L	Total lipids in small LDL	mmol/L
S-LDL-P	Concentration of small LDL particles	mol/L
XL-HDL-C	Total cholesterol in very large HDL	mmol/L
XL-HDL-FC	Free cholesterol in very large HDL	mmol/L
XL-HDL-PL	Phospholipids in very large HDL	mmol/L
XL-HDL-TG	Triglycerides in very large HDL	mmol/L
XL-HDL-CE	Cholesterol esters in very large HDL	mmol/L
XL-HDL-L	Total lipids in very large HDL	mmol/L
XL-HDL-P	Concentration of very large HDL particles	mol/L
L-HDL-C	Total cholesterol in large HDL	mmol/L
L-HDL-FC	Free cholesterol in large HDL	mmol/L
L-HDL-PL	Phospholipids in large HDL	mmol/L
L-HDL-CE	Cholesterol esters in large HDL	mmol/L
L-HDL-L	Total lipids in large HDL	mmol/L
L-HDL-P	Concentration of large HDL particles	mol/L
M-HDL-C	Total cholesterol in medium HDL	mmol/L
M-HDL-FC	Free cholesterol in medium HDL	mmol/L
M-HDL-PL	Phospholipids in medium HDL	mmol/L
M-HDL-CE	Cholesterol esters in medium HDL	mmol/L
M-HDL-L	Total lipids in medium HDL	mmol/L
M-HDL-P	Concentration of medium HDL particles	mol/L

Abbreviation	Full description	Unit
S-HDL-TG	Triglycerides in small HDL	mmol/L
S-HDL-L	Total lipids in small HDL	mmol/L
S-HDL-P	Concentration of small HDL particles	mol/L
Total lipids		
VLDL-TG	Triglycerides in VLDL	mmol/L
LDL-C	Total cholesterol in LDL	mmol/L
HDL-C	Total cholesterol in HDL	mmol/L
Serum-TG	Serum total triglycerides	mmol/L
Serum-C	Serum total cholesterol	mmol/L
Amino acids and other metabolites		
bOHBut	3-hydroxybutyrate	mmol/L
Ace	Acetate	mmol/L
AcAce	Acetoacetate	mmol/L
Ala	Alanine	mmol/L
Alb	Albumin	mmol/L
MobCH2	CH2 groups of mobile lipids	mmol/L
MobCH3	CH3 groups of mobile lipids	mmol/L
Cit	Citrate	mmol/L
Crea	Creatinine	mmol/L
MobCH	Double bond protons of mobile lipids	mmol/L
Glc	Glucose	mmol/L
Gln	Glutamine	mmol/L
Glol	Glycerol	mmol/L
Gly	Glycine	mmol/L
Gp	Glycoprotein acetyls, mainly a1-acid glycoprotein	mmol/L
His	Histidine	mmol/L
Ile	Isoleucine	mmol/L
Lac	Lactate	mmol/L
Leu	Leucine	mmol/L
Phe	Phenylalanine	mmol/L
Pyr	Pyruvate	mmol/L
Tyr	Tyrosine	mmol/L
Urea	Urea	mmol/L
Val	Valine	mmol/L

Abbreviation	Full description	Unit
Serum lipid extracts		
Est-C	Esterified cholesterol	mmol/L
Free-C	Free cholesterol	mmol/L
FAw3	n-3 fatty acids	mmol/L
FAw6	n-6 fatty acids	mmol/L
FAw79S	n-7, n-9 and saturated fatty acids	mmol/L
TotFA	Total fatty acids	mmol/L
LA	18:2, linoleic acid	mmol/L
otPUFA	Other polyunsaturated fatty acids than 18:2	**
DHA	22:6, docosahexaenoic acid	mmol/L
MUFA	Monounsaturated fatty acids; 16:1, 18:1	mmol/L
TotPG	Total phosphoglycerides	mmol/L
PC	Phosphatidylcholine and other cholines	mmol/L
SM	Sphingomyelins	mmol/L
	Derived measures	
VLDL-D	Mean diameter for VLDL particles	nm
LDL-D	Mean diameter for LDL particles	nm
HDL-D	Mean diameter for HDL particles	nm
VLDL-TG-eFR	Triglycerides in VLDL *	mmol/L
IDL-C-eFR	Total cholesterol in IDL *	mmol/L
LDL-C-eFR	Total cholesterol in LDL *	mmol/L
HDL2-C	Total cholesterol in HDL2 *	mmol/L
HDL3-C	Total cholesterol in HDL3 *	mmol/L
ApoA1	Apolipoprotein A-I *	g/L
ApoB	Apolipoprotein B *	g/L
ApoBtoApoA1	Apolipoprotein B by apolipoprotein A-I *	
FAw3toFA	Ratio of n-3 fatty acids to total fatty acids	%
FAw6toFA	Ratio of n-6 fatty acids to total fatty acids	%
FAw79StoFA	Ratio of n-7, n-9 and saturated fatty acids to total fatty acids	%
CH2inFA	Average number of methylene groups in a fatty acid chain	
TGtoPG	Ratio of triglycerides to phosphoglycerides	
CH2toDB	Average number of methylene groups per a double bond	
DBinFA	Average number of double bonds in a fatty acid chain	
BIStoDB	Ratio of bisallylic groups to double bonds	
BIStoFA	Ratio of bisallylic groups to total fatty acids	
FALen		

Bibliography

[1] "Wolfram mathworld the web most extensive mathematics resource," http://mathworld.wolfram.com/, accessed: 2015-11-28.

[2] S. E. Schaeffer, "Survey: Graph clustering," *Comput. Sci. Rev.*, vol. 1, no. 1, pp. 27–64, Aug. 2007.

[3] S. Fortunato, "Community detection in graphs," *CoRR*, vol. abs/0906.0612, 2009.

[4] L. Rokach and O. Maimon, "Clustering methods," in *Data mining and knowledge discovery handbook.* Springer, 2005, pp. 321–352.

[5] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[6] E. Le Martelot and C. Hankin, "Multi-scale community detection using stability as optimisation criterion in a greedy algorithm." in *KDIR*, 2011, pp. 216–225.

[7] ——, "Multi-scale community detection using stability optimisation," *International Journal of Web Based Communities*, vol. 9, no. 3, pp. 323–348, 2013.

[8] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, "The structure and dynamics of multilayer networks," *Physics Reports*, vol. 544, no. 1, pp. 1–122, 2014.

[9] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of Complex Networks*, vol. 2, no. 3, pp. 203–271, 2014.

[10] G. Menichetti, D. Remondini, P. Panzarasa, R. J. Mondragón, and G. Bianconi, "Weighted multiplex networks," *PloS one*, vol. 9, no. 6, p. e97857, 2014.

[11] G. Menichetti, D. Remondini, and G. Bianconi, "Correlations between weights and overlap in ensembles of weighted multiplex networks," *Physical Review E*, vol. 90, no. 6, p. 062817, 2014.

[12] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási, "Uncovering disease-disease relationships through the incomplete interactome," *Science*, vol. 347, no. 6224, p. 1257601, 2015.

[13] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature reviews genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[14] D.-S. Lee, J. Park, K. Kay, N. Christakis, Z. Oltvai, and A.-L. Barabási, "The implications of human metabolic network topology for disease comorbidity," *Proceedings of the National Academy of Sciences*, vol. 105, no. 29, pp. 9880–9885, 2008.

[15] "The hardvard clinical and transational science center. network medicine," https://catalyst.harvard.edu/services/networkmedicine/, accessed: 2016-01-22.

[16] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases," *Scientific reports*, vol. 5, 2015.

[17] A.-L. Barabási, "Network medicine — from obesity to the "diseasome"," *New England Journal of Medicine*, vol. 357, no. 4, pp. 404–407, 2007.

[18] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[19] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[20] C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, and G. Siuzdak, "Metlin: a metabolite mass spectral database," *Therapeutic drug monitoring*, vol. 27, no. 6, pp. 747–751, 2005.

[21] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney *et al.*, "Hmdb: the human metabolome database," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D521–D526, 2007.

[22] H. Kitano, “Computational systems biology,” *Nature*, vol. 420, no. 6912, pp. 206–210, 2002.

[23] M. Inouye, J. Kettunen, P. Soininen, K. Silander, S. Ripatti, L. S. Kumpula, E. Hämäläinen, P. Jousilahti, A. J. Kangas, S. Männistö *et al.*, “Metabonomic, transcriptomic, and genomic variation of a population cohort,” *Molecular systems biology*, vol. 6, no. 1, p. 441, 2010.

[24] M. Inouye, K. Silander, E. Hamalainen, V. Salomaa, K. Harald, P. Jousilahti, S. Männistö, J. G. Eriksson, J. Saarela, S. Ripatti *et al.*, “An immune response network associated with blood lipid levels,” *PLoS Genet*, vol. 6, no. 9, p. e1001113, 2010.

[25] R. A. Wevers, U. Engelke, and A. Heerschap, “High-resolution 1h-nmr spectroscopy of blood plasma for metabolic studies.” *Clinical Chemistry*, vol. 40, no. 7, pp. 1245–1250, 1994.

[26] S. K. Bharti and R. Roy, “Quantitative 1 h nmr spectroscopy,” *TrAC Trends in Analytical Chemistry*, vol. 35, pp. 5–26, 2012.

[27] L. R. Snyder, J. J. Kirkland, and J. W. Dolan, *Introduction to modern liquid chromatography.* John Wiley & Sons, 2011.

[28] M. S. A. Sophisticated, “Mass spectrometry as an emerging tool for systems biology,” *Biotechniques*, vol. 36, no. 6, 2004.

[29] W. H. O. (WHO *et al.*, “Obesity and overweight factsheet from the who,” *World*, 2015.

[30] “Who nutrition databases,” http://www.who.int/nutrition/databases/en/, accessed: 2015-12-09.

[31] W. E. Consultation, “Waist circumference and waist-hip ratio,” 2011.

[32] R. B. D’Agostino, R. S. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro, and W. B. Kannel, “General cardiovascular risk profile for use in primary care the framingham heart study,” *Circulation*, vol. 117, no. 6, pp. 743–753, 2008.

[33] S. M. Grundy, R. Pasternak, P. Greenland, S. Smith, and V. Fuster, “Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for healthcare professionals from the american heart association and the american college of cardiology,” *Journal of the American College of Cardiology*, vol. 34, no. 4, pp. 1348–1359, 1999.

[34] R. Tatami, H. Mabuchi, K. Ueda, R. Ueda, T. Haba, T. Kametani, S. Ito, J. Koizumi, M. Ohta, S. Miyamoto *et al.*, "Intermediate-density lipoprotein and cholesterol-rich very low density lipoprotein in angiographically determined coronary artery disease." *Circulation*, vol. 64, no. 6, pp. 1174–1184, 1981.

[35] ——, "Intermediate-density lipoprotein and cholesterol-rich very low density lipoprotein in angiographically determined coronary artery disease." *Circulation*, vol. 64, no. 6, pp. 1174–1184, 1981.

[36] C. Booker and J. Mann, "The relationship between saturated and trans unsaturated fatty acids and ldl-cholesterol and coronary heart disease," *A review undertaken for Food Standards Australia New Zealand. Canberra, FSANZ*, 2005.

[37] G. Boden, "Obesity and free fatty acids," *Endocrinology and metabolism clinics of North America*, vol. 37, no. 3, pp. 635–646, 2008.

[38] H. M. Roche, "Unsaturated fatty acids," *Proceedings of the Nutrition Society*, vol. 58, no. 02, pp. 397–401, 1999.

[39] E. Harris, "Biochemical facts behind the definition and properties of metabolites," 2014.

[40] E. R. Braverman, C. C. Pfeiffer, K. Blum, and R. Smayda, *The healing nutrients within: facts, findings, and new research on amino acids.* Basic Health Publications, Inc., 2003.

[41] L. H. Kuller, J. E. Eichner, T. J. Orchard, G. A. Grandits, L. McCallum, R. P. Tracy, M. R. F. I. T. R. Group *et al.*, "The relation between serum albumin levels and risk of coronary heart disease in the multiple risk factor intervention trial," *American journal of epidemiology*, vol. 134, no. 11, pp. 1266–1277, 1991.

[42] J. R. Bales, D. P. Higham, I. Howe, J. K. Nicholson, and P. J. Sadler, "Use of high-resolution proton nuclear magnetic resonance spectroscopy for rapid multicomponent analysis of urine." *Clinical Chemistry*, vol. 30, no. 3, pp. 426–432, 1984.

[43] J. P. Matts, J. N. Karnegis, C. T. Campos, L. L. Fitch, J. W. Johnson, and H. Buchwald, "Serum creatinine as an independent predictor of coronary heart disease mortality in normotensive survivors of myocardial infarction," *Journal of family practice*, vol. 36, no. 5, pp. 497–504, 1993.

[44] J. Fuller, M. Shipley, G. Rose, R. J. Jarrett, and H. Keen, “Coronary-heart-disease risk and impaired glucose tolerance the whitehall study,” *The Lancet*, vol. 315, no. 8183, pp. 1373–1376, 1980.

[45] N. Sfetcu, *Health & Drugs: Disease, Prescription & Medication.* Nicolae Sfetcu, 2014.

[46] R. Wannemacher, A. Klainer, R. Dinterman, and W. Beisel, “The significance and mechanism of an increased serum phenylalanine-tyrosine ratio during infection.” *The American journal of clinical nutrition*, vol. 29, no. 9, pp. 997–1006, 1976.

[47] H. Nørrelund, H. Wiggers, M. Halbirk, J. Frystyk, A. Flyvbjerg, H. E. Bøtker, O. Schmitz, J. O. L. Jørgensen, J. S. Christiansen, and N. Møller, “Abnormalities of whole body protein turnover, muscle metabolism and levels of metabolic hormones in patients with chronic heart failure,” *Journal of internal medicine*, vol. 260, no. 1, pp. 11–21, 2006.

[48] H. Lange and R. Jäckel, “Usefulness of plasma lactate concentration in the diagnosis of acute abdominal disease.” *The European journal of surgery= Acta chirurgica*, vol. 160, no. 6-7, pp. 381–384, 1993.

[49] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, “Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles,” *PLoS biol*, vol. 5, no. 1, p. e8, 2007.

[50] A. Lancichinetti and S. Fortunato, “Consensus clustering in complex networks,” *Scientific reports*, vol. 2, 2012.

[51] “Kendall’s rank correlation,” http://www.statsdirect.com/help/content/nonparametric_methods/kendall_correlation.htm, accessed: 2015-12-17.

[52] “Correlation (pearson, kendall, spearman),” http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/, accessed: 2015-12-17.

[53] E. Saccenti, M. Suarez-Diez, C. Luchinat, C. Santucci, and L. Tenori, “Probabilistic networks of blood metabolites in healthy subjects as indicators of latent cardiovascular risk,” *Journal of proteome research*, vol. 14, no. 2, pp. 1101–1111, 2014.

[54] C. Walck, “Handbook on statistical distributions for experimentalists,” 2007.

[55] "Protein-protein interaction networks. structures and modules," http://lectures.molgen.mpg.de/Functional_Genomics_WS1112/PPI2.pdf, accessed: 2015-06-25.

[56] G. Menichetti, G. Bianconi, G. Castellani, E. Giampieri, and D. Remondini, "Multiscale characterization of ageing and cancer progression by a novel network entropy measure," *Molecular BioSystems*, vol. 11, no. 7, pp. 1824–1831, 2015.

[57] E. L. Martelot and C. Hankin, "Multi-scale community detection using stability optimisation within greedy algorithms," *arXiv preprint arXiv:1201.3307*, 2012.

[58] "Markov chains," https://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/Chapter11.pdf, accessed: 2015-11-28.

[59] A. F. McDaid, D. Greene, and N. Hurley, "Normalized mutual information to evaluate overlapping community finding algorithms," *arXiv preprint arXiv:1110.2515*, 2011.

[60] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, p. 033015, 2009.

[61] M. Bastian, S. Heymann, M. Jacomy *et al.*, "Gephi: an open source software for exploring and manipulating networks." *ICWSM*, vol. 8, pp. 361–362, 2009.

[62] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software," *PloS one*, vol. 9, no. 6, p. e98679, 2014.

[63] "Linear regression models," http://people.duke.edu/~rnau/regintro.htm, accessed: 2016-01-30.