UNIVERSITÀ DEGLI STUDI DI MILANO

Dipartimento di Biotecnologie Mediche e Medicina Translazionale

Dottorato di Ricerca in Scienze Biochimiche

XXX Ciclo

APPROACHES TO THE MOLECULAR BASIS OF GENETIC DISEASES

Tutor:

Prof.ssa Laura CANTÙ

PhD candidate:

Alessio GAMBA

Matr. R11022

Anno Accademico 2016 – 2017

Index

Summary

Throughout my PhD period, I experimented different approaches to a complex problem, that is, assessing molecular basis of genetic diseases. In the present thesis I will show how both experimental and computational techniques can contribute to the study of typical biochemistry topics, providing non-conventional points of view and advanced technical and technological supports.

The first part of my PhD was focused on physico-chemical investigation, while the second part was devoted to big-data and network analysis. For this reason my thesis is divided into two chapters.

The former describes a typical bottom-up approach. I applied the physical technique of laser light scattering to study the first stages of aggregation of configurational variants of the beta-amyloid peptide, recognized to be either promoting or preventing the Alzheimer Disease, as compared with the wild type peptide. The variants correspond to genetic mutations giving rise to either familial cases of Alzheimer Disease or protection from the pathology. I carried out experiments aimed to assess similarities and differences in the kinetics of evolution of the aggregated species in very dilute solution, mimicking the physiological and pathological conditions. The outcome of this study is the discovery of a correlation between the molecular structure and the physico-chemical behavior, in this case aggregation, constituting the hallmark of the disease.

The results obtained have been published in a peer reviewed journal (Biophysical Chemistry), with the title: *"Pathogenic Aβ A2V versus protective Aβ A2T mutation: early stage aggregation and membrane interaction"* (2017).

The latter is a top-down approach. I defined the criteria and developed an algorithm for searching big databases with respect to: a) the clinical features of inherited diseases and b) the proteins that are known to be involved in

genetically determined diseases. I established searching criteria aiming at regrouping extracted data according to similarity classes in each database. Then, the developed method involves assessing the existence and degree of similarity within and between different clusters. As a result of this approach, we have discovered a correlation between similarity classes extracted from the different databases (the clinical and the biological), thus establishing or suggesting the existence of a biological basis for a genetic disease.

The obtained results have been submitted for publication in a peer reviewed journal with the title: *“The Disease Similarity Networks: Correlating the Clinical and Biological Similarity of Inherited Diseases”*, while a second manuscript is currently under preparation.

Based on different disciplines, and designed with the typical instruments and methodologies of physics and statistics, both approaches give non-conventional hints for the understanding of the molecular basis of complex genetic diseases.

Acknowledgment

I would like to express my appreciation and gratitude to all the people who helped me, in particular at the department of Medical Biotechnology and Translational Medicine (*BIOMETRA*) of University of Milan and at the department of Biochemistry and Molecular Pharmacology of the Institute of Pharmacological Research Mario Negri.

I am deeply grateful to Prof. Laura Cantù (Laboratory of Biophysics, University of Milan) for giving me the opportunity to attend the PhD program in Biochemical Sciences and for her support and patience over these years.

My gratitude goes also to all the people of the Laboratory of Biophysics, in particular Prof. Elena Del Favero, Prof. Paola Brocca and Dr. Valeria Rondelli, for their enormous help.

I am mainly indebted to Dr. Gianfranco Bazzoni (Institute of Pharmacological Research Mario Negri), who hosted me in his laboratory of Systems Biology during the last period of my thesis. I am immense grateful to him for his invaluable contribution in the work presented in this thesis and for his supervision and suggestions.

A special thanks to my family and friends, in particular Paola and Giovanni, who support me year by year, during all steps of the university path.

Chapter one

1.1 Abstract

The present study is focused on the aggregation of Amyloid beta (Aβ), a peptide that plays an important role in the onset and progression of Alzheimer disease (AD). This pathology is an increasing form of dementia, affecting today large part of the elderly population. The Aβ deposition in the brain starts with small aggregates and proceeds with formation of extracellular plaques and intracellular neurofibrillary tangles, leading to neuronal death. In our hypothesis, early aggregates of Aβ are crucial for understanding the development of AD. Among the different isoforms of Aβ peptides that are cleaved from the Amyloid Precursor Protein (APP), we used in our analysis the Aβ 1-42, the most toxic and aggregation-prone. The four amino acid residues at the N-terminal of Aβ contribute to the formation of a stable β-sheet secondary structure. For this reason, we take into consideration two mutations found in patients that affect the second position of Aβ peptides. These mutations lead to a single amino acid substitution of Alanine to Valine (A2V) and Alanine to Threonine (A2T). The carriers of A2T mutation surprisingly do not develop AD, whereas A2V mutation causes an early-onset and severe AD in homozygous, but not in heterozygous, showing (in this last situation) a protective role.

The process of aggregation for Aβ was accurately measured using static and dynamic laser scattering techniques. This approach allows to monitor the aggregation from the first steps of process even for small molecules and even working at very low concentrations (25 μM). In order to start the experiment with a monomeric sample, we used Aβ synthesized with depsipeptides method. Moreover, the temperature was maintained constantly low, slowing

down the process. Within all the duration of measurements (more than 100 h) we calculated the diameter of particles using also the “Non Negative Least Squares” (NNLS) algorithm to group aggregates in different populations by their size.

As a result, it is immediately possible to observe that the Aβ wild type (WT) sample shows an aggregation poorly detectable compared with the two mutations, Aβ A2V and Aβ A2T.

The A2V sample surely aggregates better than the other two forms, reflecting the phenotype observed in patients, who develop a severe dementia. As shown by measurements, WT and A2V present a very different aggregation kinetics and characteristic diameters. Thus, we can hypothesize that the protective role, observed in heterozygous A2V patients, is probably due to a destructive interaction between WT and A2V peptides during their aggregation.

People with A2T mutation are healthy but it is not due to the inability of Aβ to aggregate, as measured by our technique, and for this reason the causes of A2T phenotype need to be better investigated, also with other methods.

All these finds are in accordance with results obtained with parallel studies on cell toxicity, enzyme activity and membrane interaction.

1.2 Introduction

1.2.1 Alzheimer disease overview

Alzheimer disease (AD) is an increasing form of dementia, affecting today 38 million people worldwide. Among different types of dementias, this pathology result to be the most common, especially in elderly people. All the symptoms that characterize the Alzheimer disease are related to a progressive and irreversible deterioration of the higher cortical functions including mood, behavior, and memory. The cause is the degeneration of neurons, a condition depicted in the **figure 1** (as also in **figure 5**), which clearly illustrates the neuronal loss in a brain with advanced AD, compared with normal anatomy. Within ten years from the onset of the disease, a cognitive disability occurs, such as to determine the complete loss of self-sufficiency in patients. The incidence of AD is a growing function of age as well as its prevalence (which doubles every 5 years, starting from 1% around the age of 60), making this severe pathology, not only a health problem but also an economic and social plague.

Even though some molecular evidences have been emerged, the etiology of AD at present remains controversial. Among the proposed hypothesis, one of the most widely recognized regards the amyloid cascade (Hardy et al., 2002). Following this hypothesis, the fundamental anomaly of AD is the progressive deposition of Amyloid beta (Aβ) peptide, a short protein derived from the cleavage of Amyloid Precursor Protein (APP), which in turn is a transmembrane protein with a not completely understand role but with the properties of a membrane receptor. The portion of the protein corresponding to Aβ turns out to be between the extracellular portion and the transmembrane domain. The processing of Amyloid Precursor Protein starts with a splitting in the extra membrane domain followed by a further intra membrane cut by specific enzymes, and is schematically illustrated in **figure 2**.

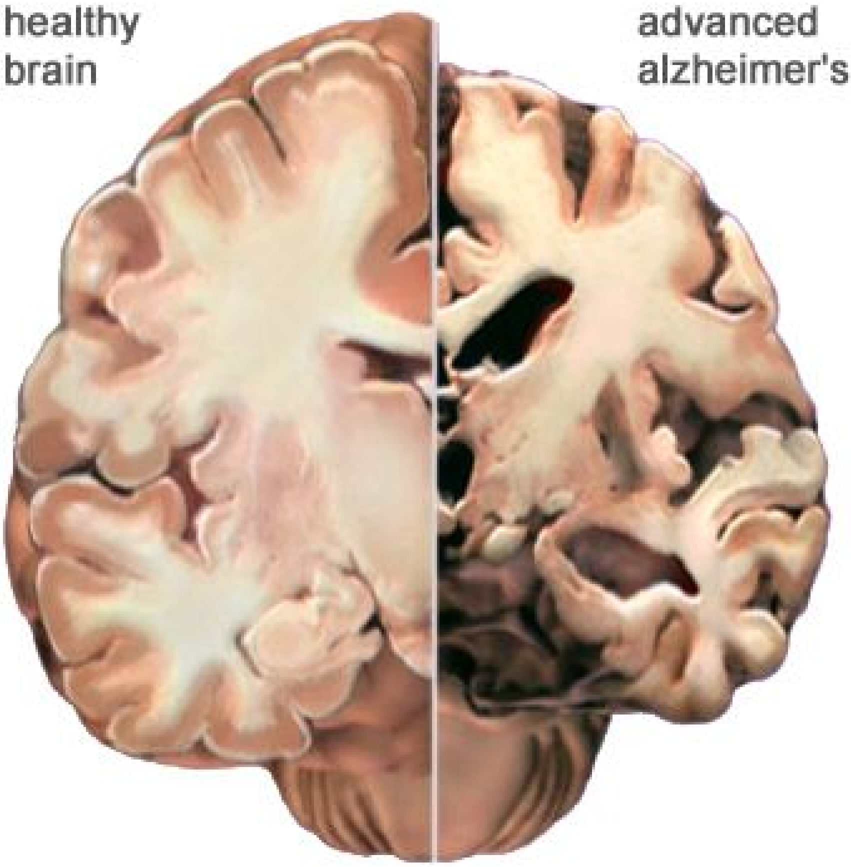


Figure 1: picture of a normal brain (on the left) and a brain in condition of advanced Alzheimer disease (on the right) where a dramatic reduction of neuronal mass is observable.

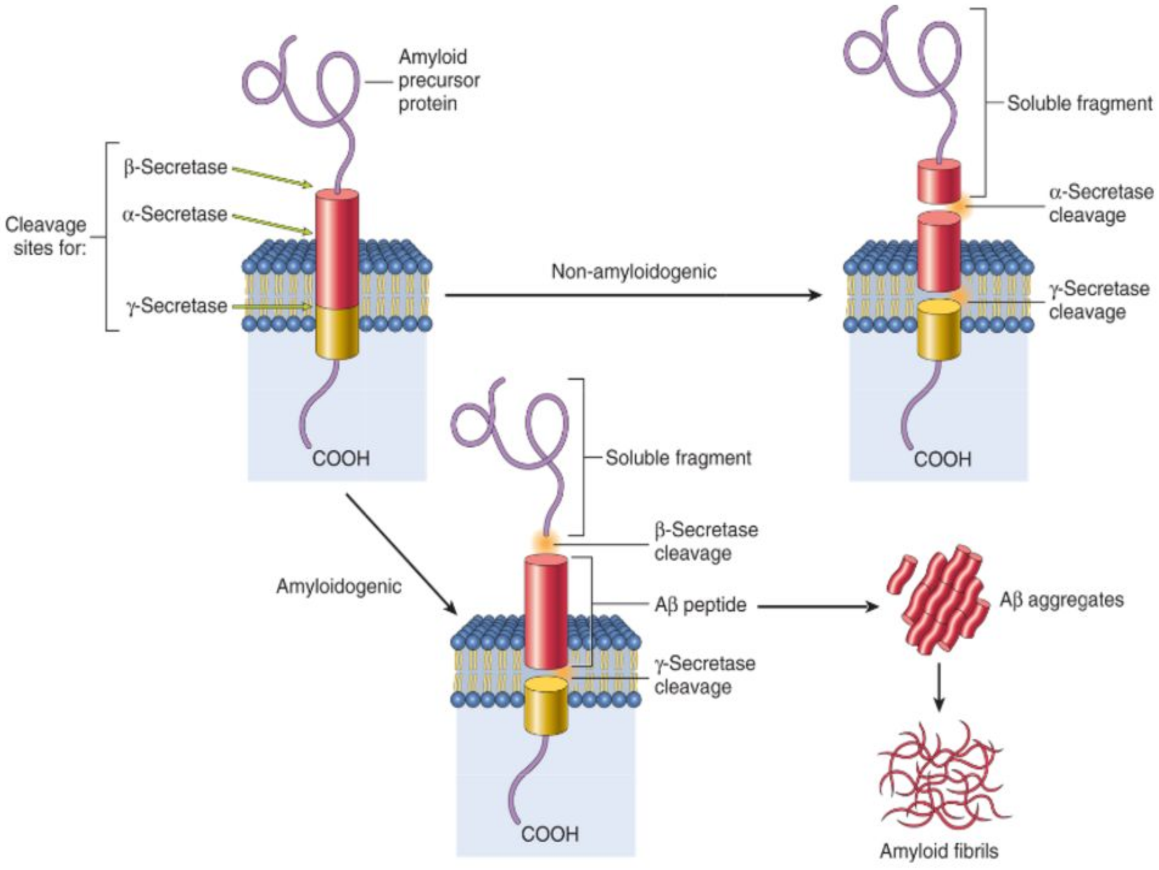


Figure 2: a schematic picture of different processing paths (amyloidogenic or non amyloidogenic) of APP.

Depending on the type of extracellular proteolysis it may be either amyloidogenic or not: if the cleavage occurs by the enzyme alfa-secretase, at the level of the external cellular surface, followed by a second proteolytic event, by gamma-secretase, the generated peptides are soluble and not amilodogenic. If APP is processed by beta-secretase at the level of the N-terminal region and then submitted to the second proteolytic event, by gamma-secretase, the generated peptides are called Amyloid beta (Aβ) peptide. The exact location of this cut may vary by giving origin to several variants of different length of Aβ peptide. The principal isoforms are two: Aβ 1-40 (40 amino acid of length) that is the most abundant and Aβ 1-42, that is recognized to be the most toxic. The **figure 3** illustrates the protein Aβ derived from the APP with its complete amino acid sequence.

The Aβ peptide, once generated, is very prone to aggregating. This process is characterized by the formation of oligomers and then protofibrils and fibrils. There are many evidences that even small aggregates can be directly neurotoxic. In particular both small and large aggregates tend to stimulate an inflammatory response, the consequences of which include alterations of the phosphorylation of the tau protein, inducing the generation of neuro-fibrillar clusters.

This cascade appears to offer a number of points at which to interfere with the sequence of pathology; however, to date all strategies have faced serious hurdles in clinical application, where the tested drugs have been ineffective or caused severe adverse events. Thus, there is still an urgent need for safe and effective molecules for AD treatment. The amyloid cascade hypothesis is depicted in **figure 4.**

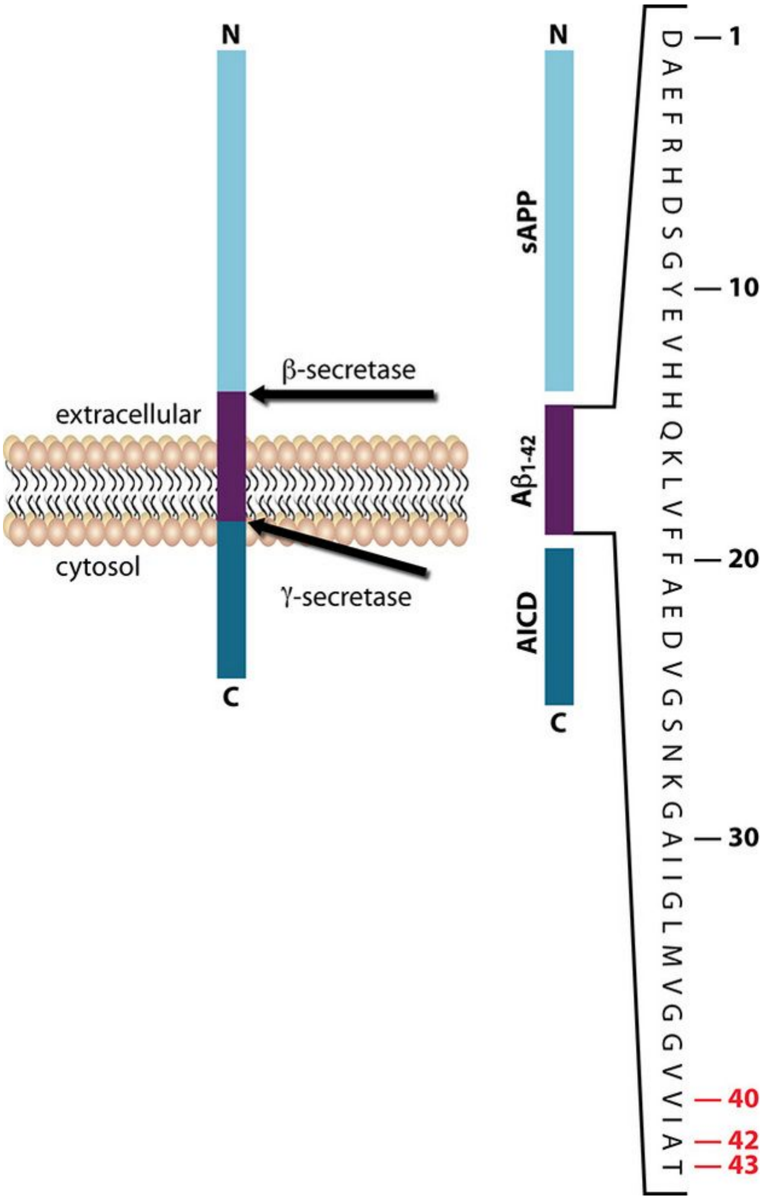

N
extracellular
β-secretase
cytosol
γ-secretase
C
N
sAPP
Aβ1-42
AICD
C
D A E F R H D S G Y E V H H Q K L V F F A E D V G S N K G A I I G L M V G G V V I A T
1
10
20
30
40
42
43

Figure 3: the sequence of Amyloid beta (Aβ). Derived from APP, the Aβ peptides can result with different length, depending by the variable position of the gamma-secretase cleavage.

AICD: APP intracellular domain; sAPP: soluble APP.

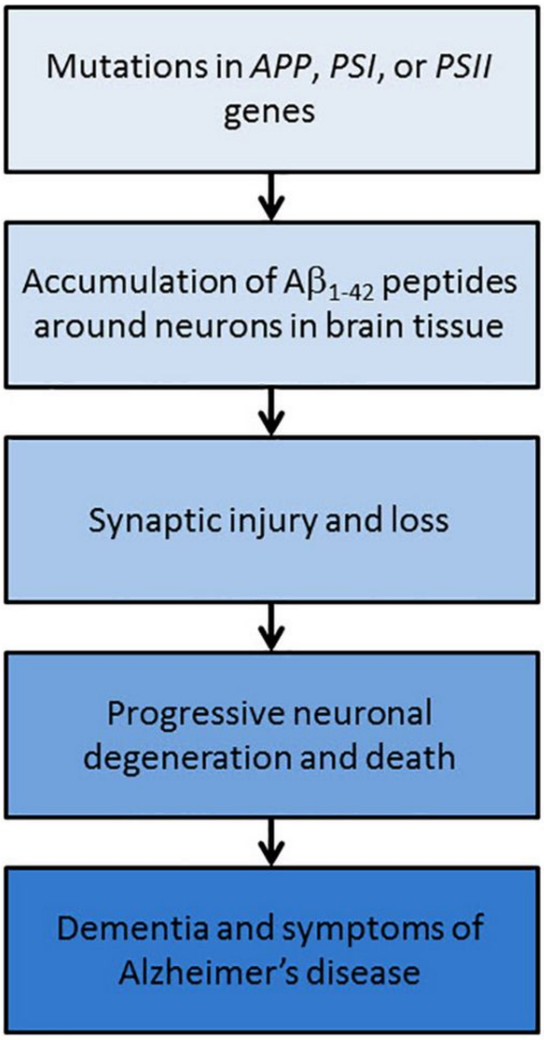

Mutations in *APP*, *PSI*, or *PSII* genes
Accumulation of Aβ1-42 peptides around neurons in brain tissue
Synaptic injury and loss
Progressive neuronal degeneration and death
Dementia and symptoms of Alzheimer's disease

Figure 4: schematic representation of progressive steps of the amyloid cascade hypothesis.

1.2.2 Familial Alzheimer disease forms

Most forms of Alzheimer disease are sporadic, that is, they manifest themselves without inheritance among generations of a family and begin after 65 years of age. In a minority of cases, however, AD manifests itself at a younger age (before 60-65 years). 60% of these early-onset forms are referred to affect different family members, in this case the disease is called familial and generally observed in two or more people belonging to the same family; 13% of them are caused by the presence of a genetic mutation present since birth.

These familial forms result to be quite rare in population and characterized by premature ages of onset and rapid evolution of the disease, representing an important source of information for genetics studies. In a high percentage of these forms, especially those characterized by an early onset and a rapid evolution, mutations have been identified on the coding genes for Presenilin 1 (PSI) and Presenilin 2 (PSII) and for the precursor of Amyloid beta, APP. The mutations so far identified are located on chromosomes 1, 14 and 21, for PSII, PSI and APP, respectively. Mutations that affect APP cause an abnormal production of amyloid beta and its pathological brain buildup in the form of “senile plaques” typical of AD patients. PSI and PSII are proteins forming the catalytic subunit of the gamma-secretase complex. All these three mutations have the same effect, leading to a preferential cleavage of the C-terminus of APP that produces Aβ 1-42, a longer form of the normal peptide (1-40). Aβ 1-42 is the neuron-toxic peptide that was originally purified in the first experiments from amyloid plaques, in 1984. These hereditary forms of Alzheimer have autosomal dominant transmission and complete penetrance (Selkoe et al., 2001).

In 2009 our research group described a new APP mutation (A673V) that causes early-onset AD when in homozygosis (Di Fede et al., 2009). The missense mutation consists of a C-to-T transition resulting in alanine-to-valine substitution at position 673 of APP, which corresponds to position 2 of Aβ peptides (Aβ A2V peptides). Notably, heterozygous individuals with A2V do not develop AD even in advanced age. Effectively, as reported in Di Fede et al. (2009), five patients with heterozygous A673V, performed well on the neuropsychological assessments, and the 88 years old aunt of the patients, in particular, showed an excellent performance on all the tests, despite she was non-educated.

Recently, another single mutation on the same alanine residue (alanine to threonine, A673T) has been reported by Jonsson (Jonsson et al., 2012). This mutation, corresponding to the position 2 of Aβ peptides (A2T) seems to protect from the growth of dementias and in particular to hinder the onset of AD. The complete mechanism of this protective effect is not yet understood, although some studies on different cell models suggest that the presence of this mutation reduces the BACE1-mediated processing of APP, then lowering the levels of Aβ production (Maloney et al., 2014).

The two punctual mutations, A2T and A2V, and the wild type of Aβ 1-42 peptide are the following:

```
WT   DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA
A2T  DTEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA
A2V  DVEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA
```

The aim of present work is to study differences in the aggregation process of Aβ peptides, in particular, among the wild type peptide and the two mutated forms, A2V and A2T. In particular we focused on the early stages of self aggregation process, by Static and Dynamic laser Light Scattering techniques when monomers evolved into oligomers and aggregated soluble forms.

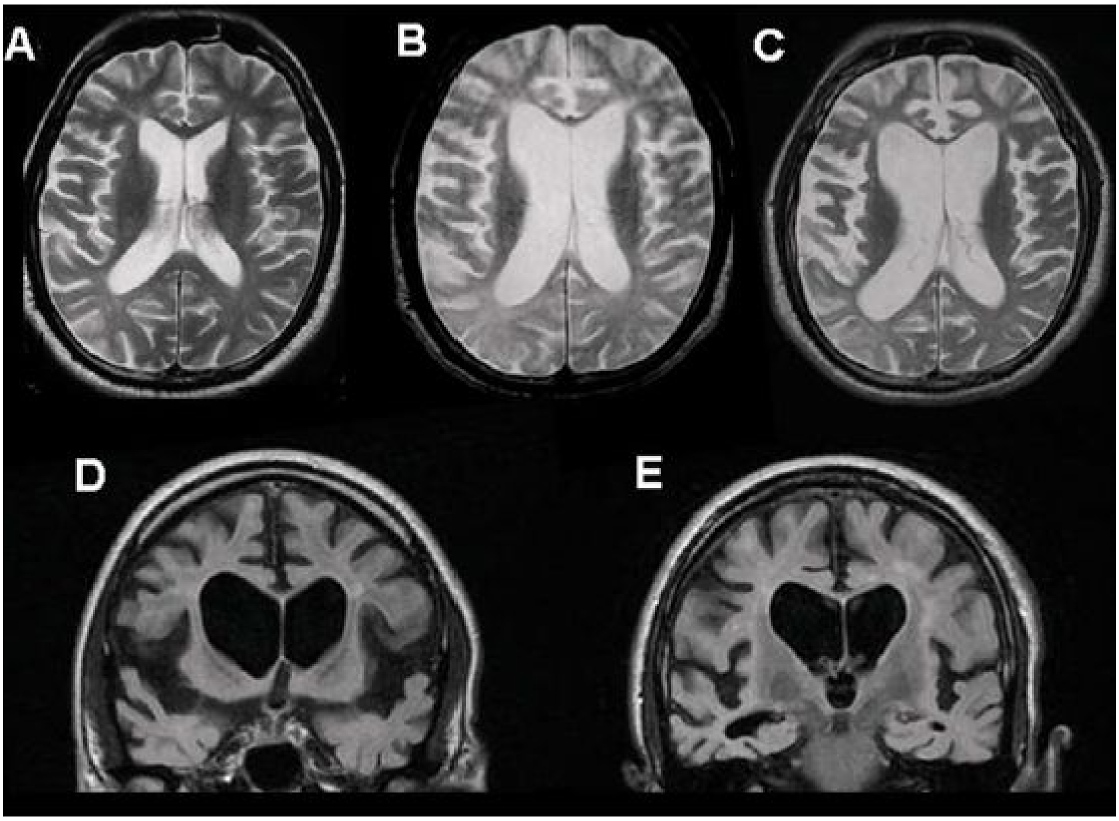


Figure 5: brain magnetic resonance imaging.

Axial T2-weighted images in the early clinical stage of AD **(A)**, after three years **(B)**, and after eight years from onset **(C)**. Images after eight years from onset **(D, E).** The scans show the progression of the cortico-subcortical atrophy and subcortical white matter changes.

1.3 Materials and Methods

1.3.1 The theory of light diffusion

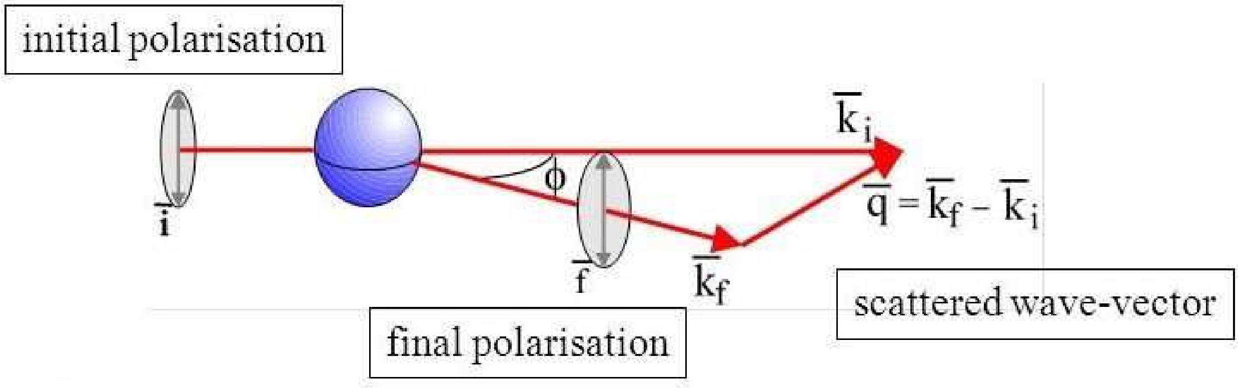
When light passes through any material medium, part of it is diffused. This observation was first described in 1871 by Lord Rayleigh. The Rayleigh's calculus is referred to a diluted gas of molecules with a small dimensions in relation to the wavelength of light λ and where a molecule can diffuses light independently from each others. In 1910, Einstein tried to address the problem from a different point of view, considering the medium as a continuum in which the diffusion of light is caused by local fluctuations of the constant dielectric ε. This demonstrated that light diffusion occurs for those fluctuations that retain the wave vector of the diffusion process. In 1944, Debye extended Einstein's considerations to the case of macromolecular solutions by interpreting dielectric fluctuations in terms of local concentration fluctuations of macromolecules. This is the work that paved the way for the systematic use of light diffusion to determine the molecular weights of macromolecules in solution (colloidal particles, polymers, proteins, viruses, etc.).

With the introduction of laser radiation as a source, light diffusion technology has become an indispensable tool in the study of biological macromolecules and nanoparticles in solution. In addition to facilitate traditional measurements, the laser makes possible new types of measurements thanks to its monochromatic and temporal coherence characteristics. Indeed, the temporal dependence of dielectric constant fluctuations can also be measured, and not just their average quadratic value as a measure of intensity. This allows you to obtain information about the microscopic dynamics of the system.

It should be noted that in the case of macromolecules, the Rayleigh's model based on single-particle can still be valid for very diluted solutions, which can be theoretically considered to be a rarefied gas. For dimensions not more negligible than the wavelength of the incident radiation, there is an angular distribution of the diffused intensity not more isotropic. In fact, in this case, there are phenomena of light interference diffused by different points of the same molecule.

1.3.2 Experiment of light diffusion

It is possible represent schematically an experiment of light diffusion with the following image.



A linearly polarized monochrome light beam (k_i) crosses a medium. A detector placed in P at a distance R from the origin of the reference system measures an intensity of light propagating in a direction other than that of the reflective or refractive beam: this is what is meant by diffuse radiation. An intuitive explanation of the diffusion of light can be due to the fact that all polarizable molecules, excited by an electromagnetic wave, irradiate in the space a wave of the same frequency as the incident ray. If the diffusing particles are present in a large number, an interference phenomenon occurs which, if the polarizability of the medium is homogeneous, can prove to be destructive in all directions other than that of the refracted beam. Thus, it is necessary to have local polarizability fluctuations to have the diffusion of light, as in the case of particles dispersed in a solvent.

The diffused field Es (R, t) is a random function of time because it reflects the statistical nature of dielectric constant fluctuations. For the most part, the field statistic is gaussian, because the diffusion process is determined by a large number of independent diffusions; this leads to a simplification of the diffuse field characterization. In fact, it is enough to define the correlation function to the first order of the field:

$$G1\ (t1,\ t2) = <Es\ (t1)\ Es\ (t2)>$$

That is, the signal is measured at time t1, multiplied by the time t2 for all the possible values of the delays t2-t1 = τ and the products obtained on a high number of samples are mediated, delayed by delay; so you get the correlation function of the signal.
In the particular case where t1 = t2, we have that:

$$G1\ (t1,\ t1) = <|Es\ (t1)|^2>$$

corresponding to the average intensity of the diffused light. In the case of Gaussian field statistic, there is a simple relationship between the G2 (t) correlation function, experimentally easier access, and the field G1 (t) directly related to the diffusion coefficients:

$$G2\ (t1,\ t2) = <I>^2\ (1+ <|G1(t1,t2)|^2>)$$

The properties of a laser experiment are depicted in **figure 6**.

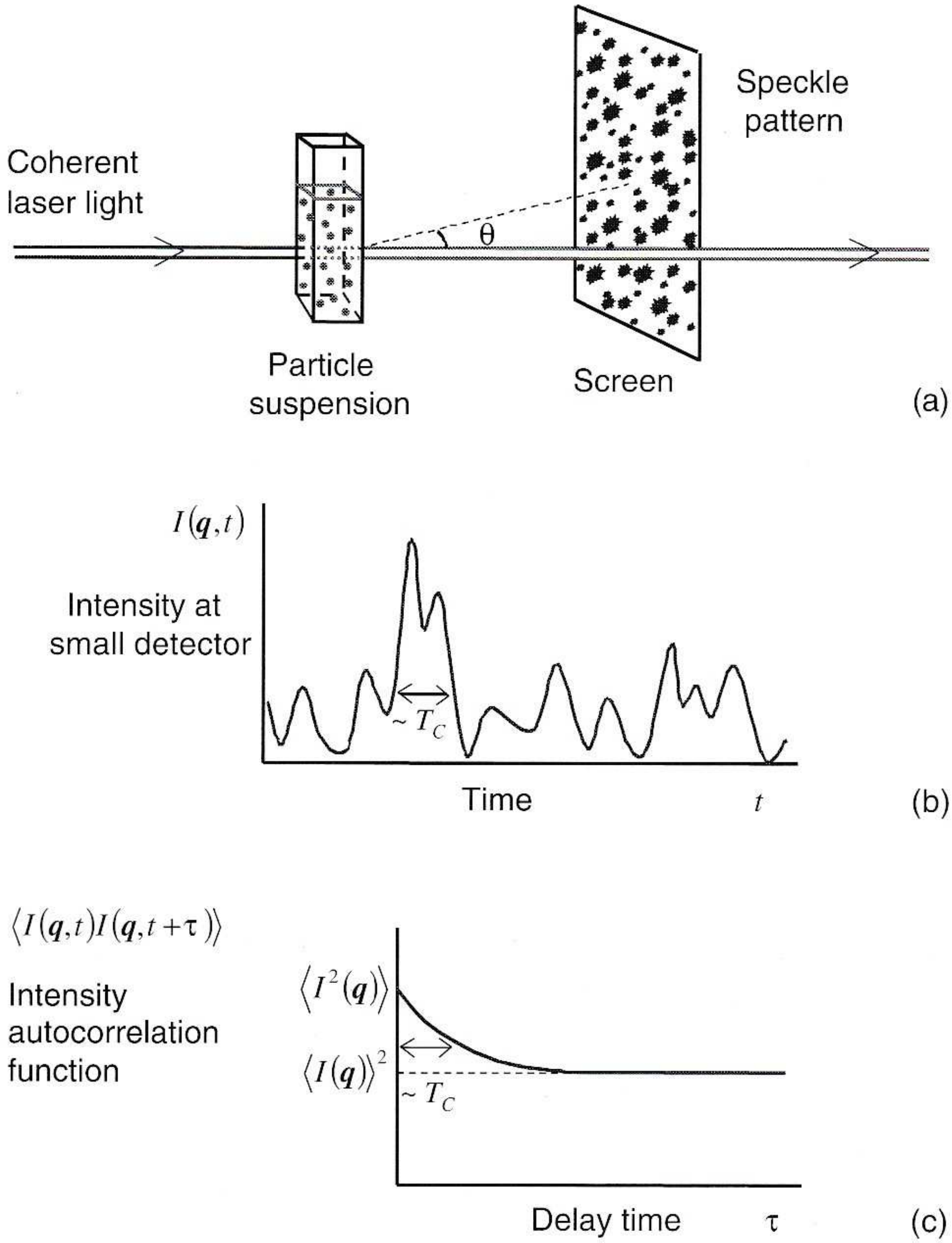

Coherent laser light
Particle suspension
θ
Screen
Speckle pattern
(a)
$I(\boldsymbol{q},t)$
Intensity at small detector
~ T_C
Time
t
(b)
$\langle I(\boldsymbol{q},t)I(\boldsymbol{q},t+\tau)\rangle$
Intensity autocorrelation function
$\langle I^2(\boldsymbol{q})\rangle$
$\langle I(\boldsymbol{q})\rangle^2$
~ T_C
Delay time
τ
(c)

Figure 6: (a) Coherent light diffused by a particle suspension generates a diffraction pattern. (b) The intensity fluctuates over time with a characteristic time T_C. (C) The intensity self-correlation function tends to decrease from $<I^2>$ to $<I>^2$

In the case of macromolecules in solution, the dynamic part of the diffusion, that of the correlation function, is closely related to the temporal fluctuations of the dielectric constant, or from the temporal evolution of spontaneous concentration fluctuations in the medium. For a simplified discussion, one can imagine that a microscopic concentration fluctuation obeys on average to the macroscopic equation for translational diffusion, namely Fick's law, which exponentially decays the concentration fluctuation δc (k, t):

$$\delta c\ (k,t) = \delta c\ (k,0)\ \exp\ (-k^2Dt)$$

where D is the translational diffusion coefficient of the macromolecule. Under stationary conditions, with t2 - t1 = τ and taking into account the relationship between the Intensity correlation function and that of the field highlighted above, it is possible to obtain the following equation:

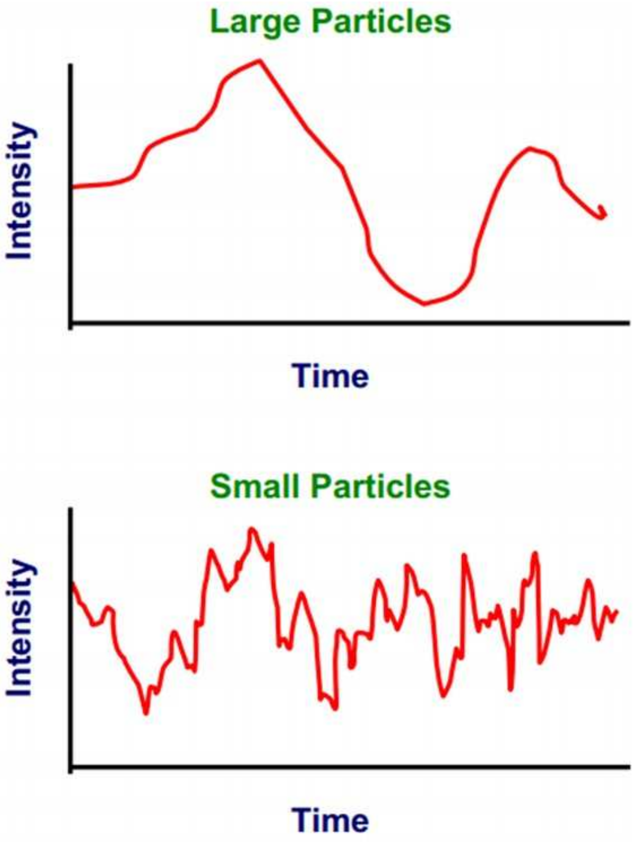
$$G2\ (\tau) = <Is(0)\ Is(\tau)> = <I>^2\ (1 + \exp\ (-2Dk^2\tau)$$

The translational diffusion coefficient D is connected to the hydrodynamic radius R_H of the macromolecule by Stokes-Einstein's law:

$$D = k_BT\ /\ 6\ \pi\eta R_H$$

with: constant Boltzman kB, absolute temperature T and η solvent viscosity. For polydispersed suspensions in size, the intensity correlation function will be due to the overlap of multiple contributions, that is a sum of exponential decreasing.

The following figure illustrates the analysis of the characteristic time of intensity fluctuation. Smaller particles cause the intensity to fluctuate more rapidly than large particles.



The diffused intensity <I>^{2} provides information on the static properties of the system: the size and shape of the particles in solution.

If the particle size is not negligible compared to the incident wavelength, the diffused intensity contains the form factor P(k). Note that P(k) = 1 for k = 0 and within point-to-point diffusers.

To directly compare the intensity of light intensity diffused by different macromolecular solutions and/or their solvents, it is useful to use the Rayleigh ratio R’, which expresses the fraction of incident power diffused in the solid angle unit per unit of the volume of diffusion volume.

Called Ir-1, the fraction of incident intensity diffused by a solution, purified by the contribution of water and normalized, has:

$$I_r - 1 = \frac{I_s / I_0}{I_w / I_0} = \frac{R'}{R'_w} = \frac{(2\pi^2 n^2 / \lambda^4 N_A)(\frac{dn}{dc})^2 cM}{R'_w}$$

Where I_0 is the accident intensity, R' is Rayleigh's ratio for macromolecules and R'_w for water, c is the sample concentration and M is the molecular mass of the molecules in solution.

The measure of water intensity is quite difficult because we should be sure to have to deal with absolutely pure water; for this reason, it is useful to use a calibrator. We used a $C_{12}E_8$ solution at 2% concentration, which, at 20° C, has an intensity exactly 114 times that of water.

1.3.3 Experimental equipment

Experiments were carried out on a non-commercial apparatus equipped with a laser (λ = 532 nm), a digital correlator, and a thermostated cell (Lago et al., 1993). High sensitivity is reached with four optical channels at 90°, displaced 5° above or below the scattering plane, that allow independent parallel measurements of the intensity scattered from the same very dilute sample. Data reported in the results section are obtained by averaging the signal collected by all of the four independent optical channels. **Figure 7** illustrates the entire laser light scattering apparatus.

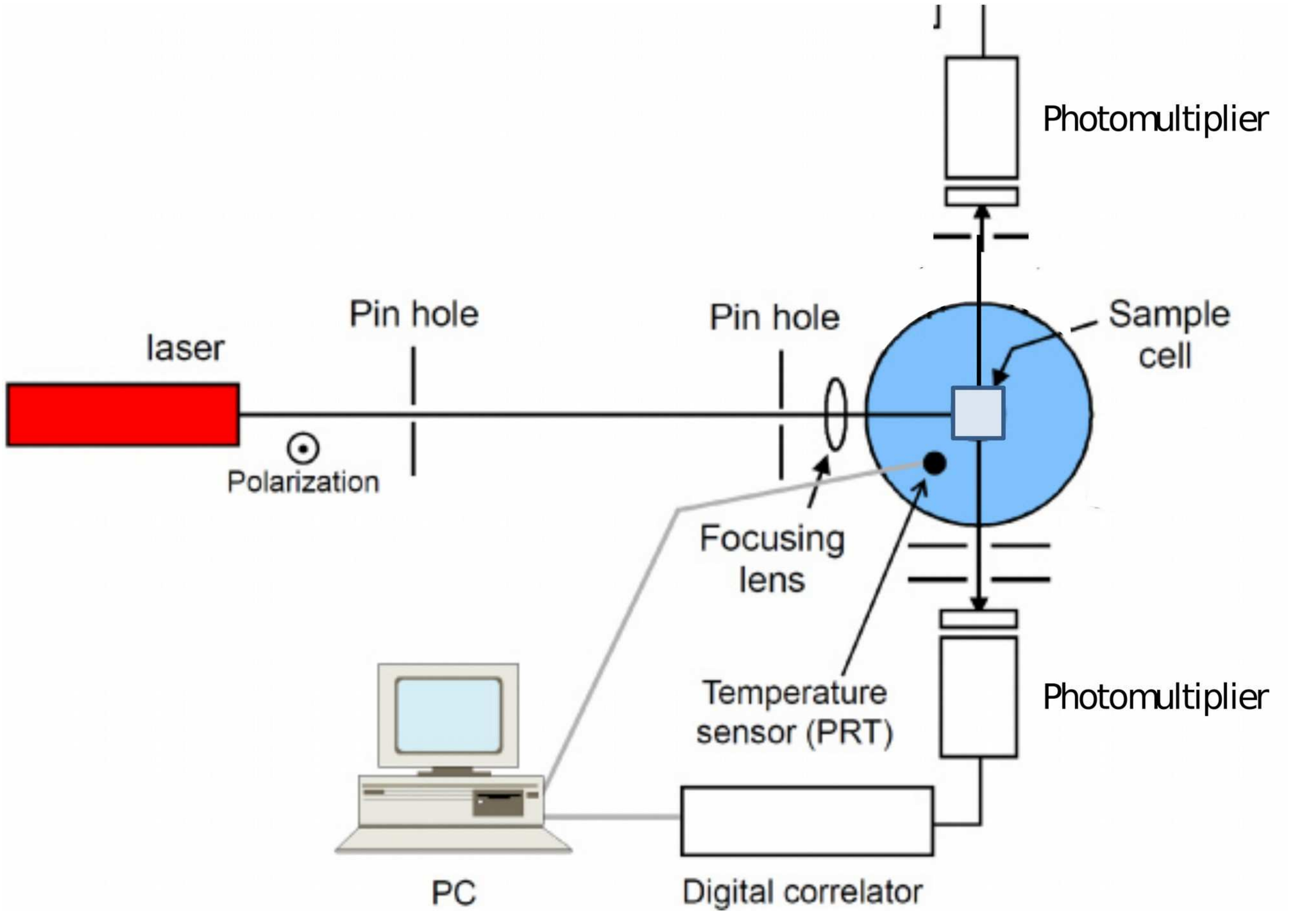


Figure 7: a schema of the laser light scattering apparatus.

1.3.4 Protocol of the experiment

We applied laser light scattering techniques to study the aggregation process of the Aβ peptides, wild type and two different punctual mutations. This technique is very sensitive to the aggregation of monomers into structures with higher molecular mass. In particular, by static laser light scattering technique we measured the average scattered intensity at 90°. The intensity value is directly proportional to the mass of the particles in solution.

This technique is very sensitive, but require very pure samples. In fact the presence of pre-aggregated species or impurities may affect dramatically the results as usually the intensity value diffused by large particles is much greater than that diffused by small one (intensity ÷ number of particles x mass^2), thus making the results so ambiguous. For that reason we used Aβ 1–42 depsi-peptides synthesized at the Mario Negri Institute. The depsi-peptide method (**figure 8**) is a specific technique of synthesis used for amyloidogenic sequences to obtain a batch with a low degree of aggregation, free of either highly folded structures or fibrils and aggregates (seeds free), and as close as possible to monomeric condition (Taniguchi et al., 2009, Beeg et al., 2011). In the Aβ1-42 sequence an O-acyl isopeptide structure is inserted (Gly25–Ser26), stable at low pH. The peptide in this condition is not prone to self aggregation. We stored the peptides, dissolved in acidic solution (water:trifluoroacetic acid, 0.02%) after clarification o/n (16-18 hours) at 55.000 rpm and filtration (centrifugal filter devices, c.o. 10 kDa, Millipore) in aliquots at 200 µM concentration.

A switch to basic pH (switching procedure) is sufficient to convert the Aβ1-42 peptide into the native sequence. The switching procedure of depsi-Aβ was carried out at basic pH: a mix of sodium hydroxide (NaOH) and ammonium hydroxide (NH4OH) (ratio3:1) was added to the peptide solutions (final pH of~10) and incubated on ice for 10-15 minutes.

Finally a proper amount of PBS buffer was added to each Aβ solution directly in the cell for laser light scattering measurements (final concentration 25 µM).

The depsi-peptide structure is the following and is illustrated in **figure 8**:

WT   DAEFRHDSGYEVHHQKLVFFAEDV**GSN**KGAIIGLMVGGVVIA

The early stages of aggregation are deeply influenced by several parameters, such as concentration and temperature. For this reason, in order to compare the effect of either A-to-V or A-to-T mutations on the aggregation pathway of Aβ, we applied the light scattering technique in the same experimental conditions to all systems. Samples were at 25 μM peptide concentration in PBS buffer 50 mM, at pH 7.4. The scattered intensity at 90° was measured at 22°C and the time evolution of the intensity was followed for at least 100 hours from switching.

I performed parallel measurements of the intensity correlation function (dynamic laser light scattering) on the same samples at different time delays from dissolution. This technique is suitable to study the size distribution of particles in solution and its evolution as a function of incubation time. DLS data analysis was carried out using both the cumulants method, suitable to detect the evolution of the weight-average hydrodynamic size of particles in solution, and the NNLS method (Lawson et al., 1995), suitable to determine their size distribution.

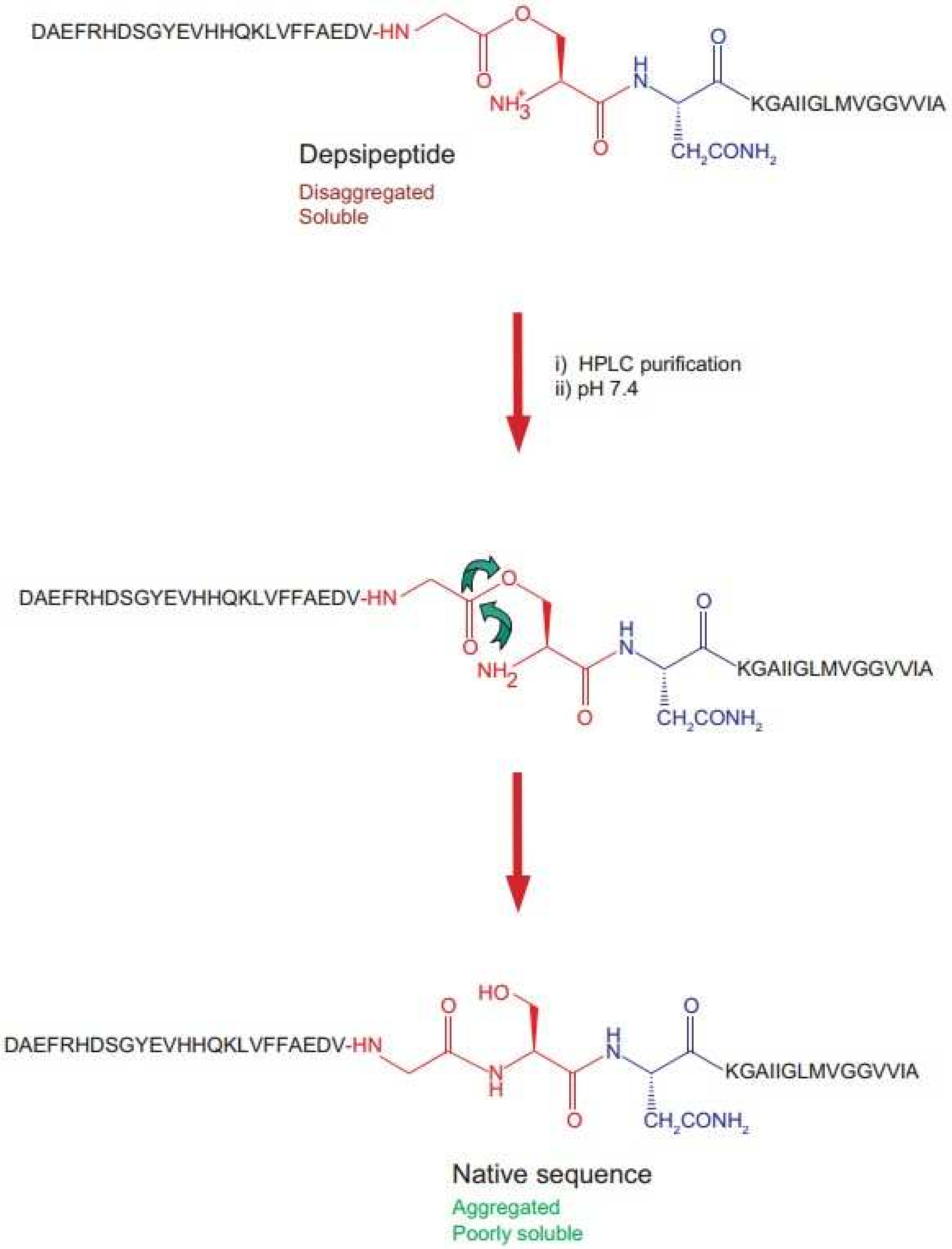


Figure 8: the depsipeptide method with the conformation of depsipeptide before and after the switching procedure.

1.4 Results

As previously described, three different forms of Amyloid β peptide 1-42 have been analyzed, the wild type and two mutated peptides (A2V and A2T), differing only in one amino acid respect to not mutated form. The use of the techniques of dynamic and static light scattering allows to detect and to study *in vitro* the aggregation propensity of the three Aβ peptides. These two techniques, used in parallel, result particular sensitive because they allow to follow on-line the aggregation process. Moreover those techniques are non invasive and do not interfere with the process of aggregation. By static laser light scattering we measure the average scattered intensity by the particles in solution which is proportional to their mass. In this way, it is possible to calculate immediately the mass of the peptides according to their aggregation state, starting from the first steps of the process, when monomers and oligomers are present. Aggregation can be observed, measuring the increase of the intensity, until soluble forms are present in solution. When big aggregates of fibrils start to form, a sedimentation process superimposes to the growing one. The intensity may decreases and the results are no more useful to describe the system.

To describe the kinetics of the aggregation of the three different peptides, we want to acquire the intensity for long time, about 100 h. For this reason it is necessary to work with a stable instrument and to set the appropriate frequency of measures and total acquisition time, in accordance with the behavior of the molecules in the sample.

The three peptides are synthesized in the form of depsi-peptide in order to keep the sample in monomeric state and then they are stocked at -80° C in order to block any reaction of aggregation.

As described in materials and methods section, the structure of depsi-peptide is more rigid respect the original peptide conformation, for this reason the molecule is not prone to the aggregation.

A suitable procedure (as described in the materials and methods section) allows to switch the Aβ from the rigid to the original conformation, that immediately begins to self aggregate.

The final step of this “switch” procedure is performed when the sample is positioned in the measuring cell in the light scattering instrument, ready to be measured. Only in this way it is possible to observe the peptides in their monomeric state and to see the formation of the first oligomeric aggregates. The process of aggregation is influenced by the concentration of the peptides and by the temperature, the lowest the temperature, the slowest the process. The temperature of the apparatus can be regulated and it is set at 22°C, a temperature lower than the biological temperature of the brain that is around 37° C. The temperature is maintained constant for all the duration of the experiment. This particular setting allows a precise observation of different stages of aggregation and to perform comparisons with results obtained on the same systems with other experimental techniques. Another very important parameter is the concentration of sample, because high concentrated solutions of peptide are more prone to aggregate. We use 25 µM peptide solutions in phosphate buffer 50 mM, pH 7.4. This is a very low concentration, but thanks to the hight sensibility of the instrument it is possible to make measurements also in these conditions, when the excess scattered signal of the particles is only a fraction of the signal of the water.

The combination of depsi-peptide switching and radiation scattering observation has already proven to be well suited to follow the kinetics of aggregation of amyloidogenic peptides starting from monomeric condition (Di Fede et al., 2009; Stravalaci et al., 2012; Diomede et al., 2014). In previous studies a more pronounced propensity to aggregate have been observed for the mutated A2V peptide as compared to the wild type (Messa et al., 2014).

In this analysis parallel measurements of static and dynamic light scattering are applied for at least 100 h from the procedure of switching, when the sedimentation of particles in the sample overwhelms the aggregation. Data are obtained by averaging the signal collected by four independent optical channels at 90°, each of them with an inclination of 5° respect the horizontal plane.

Results obtained for the three systems are reported in **figure 9** for comparison.

The frequency of acquisitions in the first range of observation (t < 20h) is quite high to describe the evolution of the scattered intensity when initial aggregates form in solution. In particular we set the frequency at one measure every 5 s for the first hour and then we slow the frequency to one measure every 10 s for the first 20 hours.

Parallel independent dynamic laser light scattering measurements are performed on the same sample. By dynamic light scattering technique we measure the intensity correlation function, connected to the average diffusion coefficient of particles in solution. By the programs directly provided by the correlator, we can calculate the mean hydrodynamic radius of particles and their size distribution. The acquisition of the intensity correlation function, in the case of very diluted samples, requires a duration time of at least one minute to reach good statistics. The frequency of measurements is high at the beginning (every two minutes) and the progressively slow down (every ten minutes). The size distribution are calculated by the Brookhaven instrument software with the “Non Negatively Least Squares” (NNLS) method (Lawson et al., 1995). In order to better compare results obtained on different samples, both the same acquisition algorithm of inversion are used for all the samples.

The initial intensity scattered from all samples is very low, < 1.5 times that of the pure solvent, that is water with the buffer. This is the value of intensity expected for small particles in a very dilute solutions (about 0.1 mg/ml). The low scattered intensity confirms that at the beginning of the experiment the peptides are in monomeric state. Looking at **figure 9,** it is possible to understand immediately that the process of aggregation is very different among the three considered Aβ 1-42. This is confirmed by both the dynamic and the static techniques, in extent and in the kinetics of aggregation, as it is described in detail in the following sections.

1.4.1 Static light scattering

Results obtained from static light scattering are presented in the top row of **figure 9**. For Aβ A2T and A2V we observed an evident growth of the scattered intensity, while the increase in intensity of wild-type system is very low. Nevertheless, looking at 40 h from the beginning of the experiment, a small rise of intensity occurs also for the wild-type, visible in the zoomed image reported in **figure 9** first column, first row.

Interestingly, the Aβ A2V displays the highest intensity growth, with a maximum value 100 times greater than the wild type. The A2T show an intermediate growth and maximum value, respect the other analyzed peptides.

Another very important aspect to consider is the kinetics of the process. Both the lag time (before the beginning of aggregation) and the evolution of the aggregation are very different in the three samples. As we can see from **figure 9**, the aggregation of Aβ A2T begins after 13 h, similar to A2V, whereas in wild type it begins after 40 h. It is difficult to estimate exactly the starting point of the aggregation of A2V sample because it grows exponentially and there are some aggregates immediately after the beginning of experiment. All experiment are stopped at 100 h because at about 90 h the sedimentation of large particles occurs, indicated by the decreasing of intensity. The observed intensity of Aβ A2T follows a two-steps growth. The first step starts at 13 h, the second at 50 h. It is possible to draw the fitting curves of the first step for wild type and A2T. These curves are obtained by

$$I(t) = I_{final} \left(1 - e^{-t/\tau}\right)$$

where I_{final} is the asymptotic value of the scattered intensity growth and τ is the characteristic time. The characteristic time τ is 9 h for the wild-type and 6.5 h for A2T.

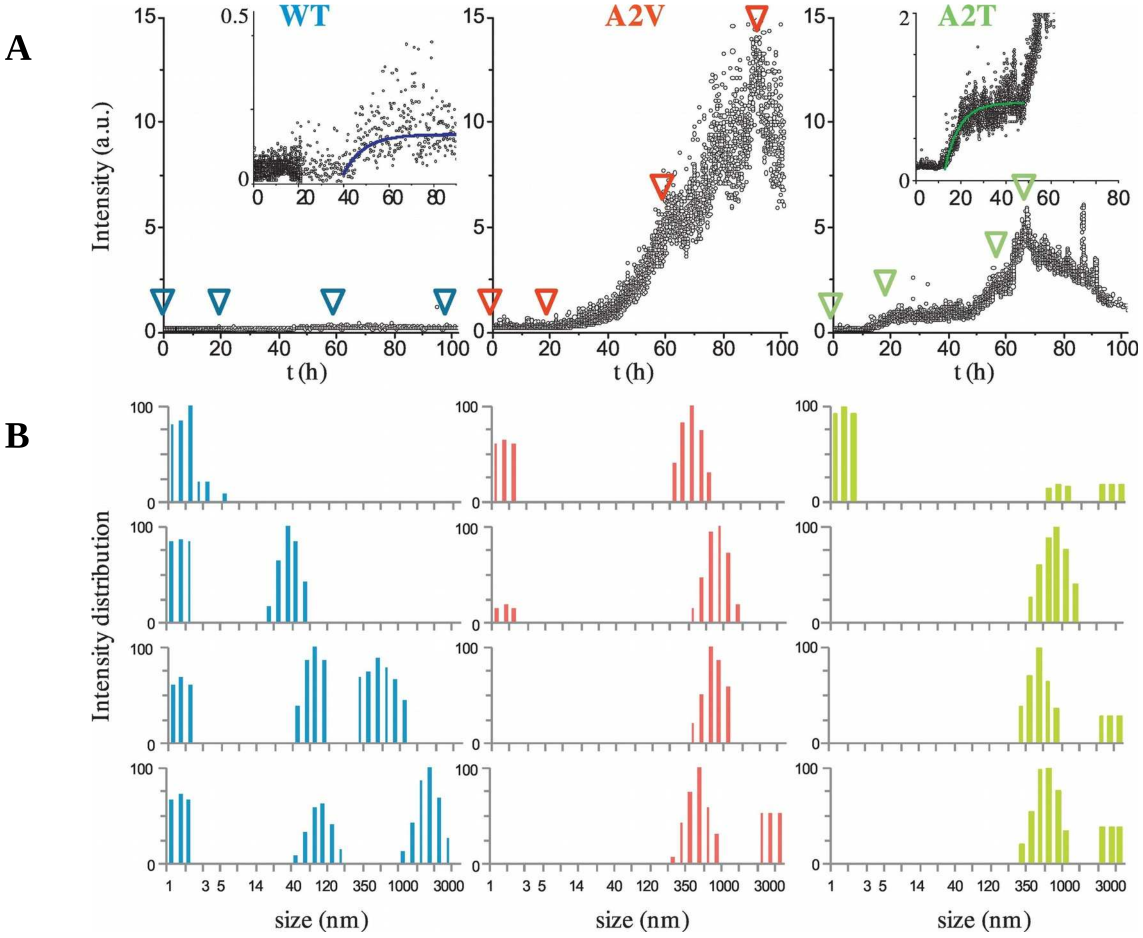


Figure 9: panel **A**, the static light scattering (SLS) results; panel **B**, the dynamic light scattering (DLS) results.

1.4.2 Dynamic light scattering

To examine the differences found with static method in deeper detail, we compared them with the results obtained simultaneously by dynamic light scattering, which give information on the size distributions of the aggregates. In **figure 9** we report data taken at four particularly interesting delays from the beginning for each of the three peptides. The **figure 9** (panel **B**) reports on the rows the size distributions taken at 0-1 h, 20 h, 60h and at the end of the observation time range. The bar charts illustrate the size of the populations in nm in relative importance calculated respect to the relative scattered intensity (from 0 to 100).

At the beginning, the substantially predominant specie is the population of monomers with a size between 2 and 3 nm. These particles are highly present in all three systems, but in the A2V sample is possible to detect also a population of 350-400 nm. This result indicates that A2V displays a very high propensity to aggregate.

The amyloid beta wild type

Regarding the sample with Aβ wild type, only the monomers are detectable at the beginning of experiment. After 20 h it is possible to reveal the formation of intermediate aggregates of size 40 nm, in coexistence with the population of monomers.

At 60 h delay, an increase in intensity is accompanied by the observation of two populations of aggregates with size ~60 nm and ~400 nm. The size of the largest aggregates slowly increases above 1000 nm, but seemingly their number remain very low, as the population of small monomers is still detectable at 100 h. This characteristic of wild type sample is very different from the other two systems, where the monomers in late phases of experiment are no more detectable, probably because the signal of other populations is too high.

The amyloid beta A2V

About the A2V sample, it is possible to observe that a population of size about 500 nm is always detectable, from the first hours until the end (100 h). The aggregation starts immediately as a unique process that leads to the formation of an increasing number of large aggregates (500 nm). At 20 h monomers are still detectable but with an intensity of 10% respect the second population. At 60 h monomers are no more detectable and at 90 h sedimentation of fibrils occurs, with size bigger than 2000 nm, as we can see looking at the final decrease of the scattered intensity.

The amyloid beta A2T

In the Aβ A2T, monomers disappear at 20 h, that is earlier respect the A2V sample. A population of 500 nm growths after a lag-time of 13 h with a characteristic time of 6.5 h. The importance of this population and its mean size stays constant until 40 h from dissolution, when a second different aggregation process begins, with the formation of elongated fibrils with size greater than 200 nm, prone to sedimentation after 60 h from dissolution.

Fibrils appear faster in A2T with respect to A2V, but the number is very different, being lower in the case of A2T.

1.5 Discussion

During the present study, we investigated the effects of two mutations that affect the Alanine in position 673 of the amyloid precursor protein (APP). Mutations are punctual, meaning that only one amino acid (Alanine 673) is changed in the whole protein. After the cleavage of amyloid precursor protein by the beta-secretase enzyme, Aβ 1-42 is produced, carrying the mutation that results in position two of the new peptide. The two mutations regard a substitution of the alanine in position two with valine or threonine, named A2V and A2T respectively. Those mutations have opposite effects on the onset and severity of Alzheimer Disease. As shown in a previous study, A2V induces early onset of the disease in homozygous individuals, while A2T prevents the pathology.

In a parallel study also the activity of beta-secretase is investigated, finding that the A2V mutation enhances the processing of amyloid precursor protein, while the A2T mutation results in a quite similar level of Aβ production as compared to the wild type.

Still we find that this is not the only difference. We applied a multi-technique approach to study the structural and morphological features of the Aβ wild type and the two mutations (A2T and A2V) on different length-scales.

Static and dynamic laser light scattering techniques are used to study the aggregation process of the three peptides with a particular interest for the early stages. These correspond to the soluble-oligomers/structured-oligomers regime, where aggregates are claimed to be essential for the toxicity in vivo and then for the onset and progression of Alzheimer disease. We take advantage of the “depsi peptide method” for the synthesis of Aβ. This method enables obtaining seed-free batches of monomeric peptides.

Comparative results indicate distinctive pathways of monomers aggregation that differed in the kinetics and the extent of the process, in the size of the aggregates, in the evolution of the secondary structure of the peptides. This means that at a given concentration, at the same delay from production within a certain time lag, the three peptides display a different propensity to establish and stabilize interactions with similar molecules. This suggests that oligomeric species with different reactivity towards their surrounding medium may occur, with different outcomes.

Notably, Aβ wild type has the lowest propensity for aggregation, and effectively monomers were still observed at long delay from dissolution. Differently, the mutation A2V increases the number and size of Aβ aggregated structures significantly, leading to fibril formation. In accordance with a parallel study of secondary structure folding, it is found that the A2V peptide evolves faster to a beta-sheet conformation.

The behavior of Aβ A2T is complex. A double-step kinetics lead to the formation of intermediate aggregates before the appearance of a small number of fibrils. In parallel we also observe a slower kinetics of folding: the distribution of secondary structures do not evolve rapidly towards an increasing amount of beta-sheets. Thus, the A2T aggregation is found to be different from the usual pre-fibrillar ones, observed in A2V, on different length-scales. The present findings are also in agreement with recent studies on monomers conformations, that show how the mutation A2T can dramatically alter the beta-hairpin population and switch the equilibrium towards alternative structures. Moreover, the observed evolution of the collection of aggregates during the time suggests that the intermediate aggregates cannot evolve simultaneously towards the progressive formation of pre-fibrillar and fibrillar structures, but spare fibrils are formed randomly in solution. These fibrils immediately precipitate, as observed by laser light scattering, and therefore are no longer present in the solution to give their characteristic contribution. Early spare fibril sedimentation can also affects the controversial results obtained by thioflavin T fluorescent dye method, that follows the fibrillation process by detecting the beta-sheet signal. Other interesting results are obtained evaluating the interaction with membrane of the three peptides and the effects on neuroblastoma N2a cell culture. Both A2V and A2T interact with model membranes, leading to a decrease in the lipid core density and an increase of the lipid chain disorder. This is associated with data on the toxicity on cell culture, where both A2V and A2T result to be more toxic than the wild type. In particular, A2V shows to be the most dangerous for cell viability, in accordance with other results (Colombo et al., 2017).

Chapter two

2.1 Abstract

Background and rationale

In the last decades, many mutated genes have been identified as causative of human diseases. The modified proteins coded by these genes are directly responsible for the symptoms and signs observed in patients.

Proteins can suffer a dramatic change in their functions that affects the normal cellular activities and that is reflected, at organism level, in the expressed phenotype. The complex path that connects a protein with a phenotype, however, is not easy to identify, and remains in many cases unknown. The aim of our study is to develop a method useful to bridge this gap. To achieve this goal, we focused on a particular condition that occurs when different proteins give rise to very similar (almost identical) phenotypes, also if the proteins are apparently unrelated. It is supposed that, in this situation, the proteins share a similar function or co-participate in a more general process that is related to the phenotype.

Experimental design

Very similar phenotypes have been recently grouped together into clusters that have been named “Phenotypic Series”. In this way, a Phenotypic Series is a list of different variants of the same disease. Comparing all these diseases, we searched for similarities among either proteins or phenotypes. In order to quantify the similarities, we considered terms that describe proteins and phenotypes, and that are retrieved from specific databases, structured in ontological way. Thanks to the present availability of this detailed and curated databases, we developed an algorithm for calculate a similarity

coefficient based on the shared terms of annotation.

We hypothesize that clinical similarities among Phenotypic Series reflect biological similarities in the underlying mechanisms of disease. To clearly illustrate the correlation among all diseases we use network and cluster analysis. A network is a graph formed by nodes linked by edges, and in our case each node is a Phenotypic Series, whereas edges represent the similarity among them. In particular, two types of network have been generated: the biological and clinical similarity network, based on protein and phenotype similarities, respectively. The assembled networks are fully connected, representing all diseases and every possible similarities among them. However, we are interested in the highest similarities only. For this reason, we proceeded gradually by increasing the threshold of similarity coefficient, a procedure that removes the majority of the connections among nodes.

Results and conclusions

By progressively increasing the threshold, the networks become fragmented into islands that are not mutually linked. These islands are composed of diseases that are similar to each other from the biological and/or clinical point of view. Then, using cluster analysis on the diseases of the fragments, we obtained results similar to the previous fragmentation.

Moreover, the obtained clusters allow defining subsets of Phenotypic series, that are characterized by varying degrees of clinical and biological similarity.

As hypothesized, groups of phenotypically similar diseases show an interesting high level of biological similarity as well, suggesting that the relevant proteins perform similar activities in the cell. As examples, we proposed two types of disease clusters: with high biological but low clinical similarity and with low biological but high clinical similarity.

To conclude, our analysis of the similarities, as expected, can propose different modalities of correlation (or lack of correlation) between clinical manifestations and biological mechanisms of diseases.

2.2 Introduction

2.2.1 Overview

One of the most important improvement in the world of biology is probably the recent introduction of computer and algorithm for the analysis of biological data. New computational approach results in some cases mandatory, in first instance for the intrinsic complexity of biological systems, and in second instance in order to manipulate the big number of data provided by new approaches and in particular the high throughput technologies as the next generation sequencing, micro array, and mass spectroscopy. So many data can provide an equivalent quantity of information if we are able to organized and analyzed them, unveiling the hidden properties they bring. Frequently, more contributions for the comprehension arise from the comparison of many data respect to a deep analysis of the singlets.

Nowadays, many computational studies are focused on biological data that principally regards genes and the proteins encoded by them. In the past decades, thanks to the improvement of fast-sequencing technology, entire genomes have become available. Therefore, In the post-genomic era, the new challenge is no more to obtain the sequence of genes, whereas to understand the role of these genes, meaning reveal all the process in which gene products are involved. Particularly, we are interested in the effects of a mutation or suppression of one gene product in the cell, and also the consequences that the change have on the whole organism. This can be very important for a better understanding of the phenotype that is observed in the contest of a disease with a genetic evidence. It would be also interesting, as we can obtain an about complete map of coding genes and their products, obtain an equally detailed map of all functions of these gene products. Nevertheless, among the incredible number of discovered genes, many of them still remain “orphan”, meaning that, at present we are not able to assign a specific role to that gene in the cell. It would be also very important to be able to move in both directions: from a gene to the corresponding phenotype, as from an observed phenotype to the causative gene or genes. In the study of genetic diseases, the method proposed in this thesis, should be an important tool to fill the gaps in present knowledge, thus resulting useful for the possible development of treatments.

2.2.2 The OMIM database and Phenotypic Series

Today exists a big effort in characterizing genetic diseases on their molecular basis. For this reason several databases have been developed with the aim to catalog the available knowledge. Probably the principal resource about genetic diseases is the "On-line Mendelian Inheritance in Man", known as OMIM. The OMIM database enlist all known diseases with a recognized gene causing them, and is organized using different entries, where each entry provides information of different types regarding one disease, such as the responsible gene, the chromosomal localization, the inheritance mode. The two principal data that we are interested in are the disease name and the gene product causing it, this last called Disease Gene Product (DGP).

The OMIM web site recently added also another feature to the database, called Phenotypic Series (Amberger et al., 2015). Phenotypic Series (PS) are groups of OMIM diseases that are associated based on their clinical similarity. This procedure of clustering is not automated, but depend by the clinical judgment of experts in those disease fields and is based on many evidences. For these reasons we consider the grouping procedure made by OMIM team very accurate. The Phenotypic Series clusters are very useful for our analysis because they reduce the number of diseases to be consider, that is, very similar diseases are considered as only one disease (one Phenotypic Series).

In addition to OMIM database, many other databases have been developed in medical and biological field. We take advantage of them, and consider the usage of these databases for our analysis.

2.2.3 Ontologies

The importance of a common language

Some databases result particularly useful because are organized with an ontological structure. To understand what is an ontology and its importance for the organization of knowledge, we can consider the following example. We imagine to be interested in searching information on one known gene that participates in a particular cellular process. During our research, we can be overwhelmed by a vast amount of biological data available on-line, such as books and journal articles, information on protein structures, genome maps, biochemical pathways, drug efficacy studies, and much more. Thus, often, the problem is not a lack of data, but data are so many that peruse all of them to find relevant information can be a lengthy process. Moreover, obtaining the principal information can be complicated because starting from our detailed query we can retrieves undesired results or even very generic descriptions. We might also miss some interesting results because, within different databases, the same process might be referred to in different way. As we can understand, filtering the relevant results is impossible for a computer algorithm, as it is true it is able to search for words and compare them, it is not able to understand the meaning of these words. The human mind can understand and chose the right results among the proposed, but this operation really requires a lot of time. All of these problematics have been emerged into the need to have an universal information easy to access.

Classification of the information is the first step to make scientific data easier to find, and for this purpose a common language is mandatory. This language describes the entities of a particular field of knowledge, for instance, in biological field, it is used to describe genes, proteins, diseases, or any of other biological entities.

A universal language allows to a program to browse the world wide web, to automatically extract relevant information, and to compare them with similar data from other databases. Thus, it is very important that the language utilized by different sources has a common dictionary of terms and that it is recognized by scientific community. Nevertheless, a controlled vocabulary is not sufficient if we do not have any information on how different concepts are mutually related. A logical relationship should be present among the terms of the language. This last point is very important because regards the structure in which the terms are organized and also the type of relation that link the terms. Such type of structured language presents several advantages in its usage and is properly called ontology. At the beginning of the study of ontologies it was a topic for philosopher, but then ontologies began to be used also in math, physics, and recently, in biology.

Probably, all the effort put into biological classification is inspired by the taxonomic system introduced by Linnaeus, that in ancient times represented a revolution in the world of biology, providing classification of different species and their relationship. Now, ontologies are widespread and can be used to describe potentially anything, even if one single ontology is normally developed to provide information on a restricted and well defined area of knowledge.

Gene Ontology and Human Phenotype Ontology

Starting from 1998 the Gene Ontology Consortium have developed the “Gene Ontology project” with the aim to bridge the gap between different biological communities (Gene Ontology Consortium, 2000). Gene Ontology (GO) is divided into three domains, called: Biological Process, Molecular Function and Cellular Component. The three aforementioned parts are actually three independent ontologies, each of them providing a different type of information about the role and localization of gene products in the cell. In particular, Molecular Function classifies gene products by what they do in the cell, Biological Process by what general processes these gene products are part of and Cellular Component by where they act in the cell. In addition to the Gene Ontology, there are also many other ontologies that classify concepts for different biological domains, such as organism anatomy, development, experimental details, phenotype, the environment, and more. All these ontologies are compiled by a consortium of leading biologists by means of a cyclic process that continuously correct errors and update the information.

In addition to Gene Ontology, another ontology result particularly useful for the scope of our study, that is the Human Phenotype Ontology (HPO). This ontology catalogs the observed phenotypes of a disease in humans, providing in this way a very rich and standardized terminology to describe clinical manifestations of pathologies in patients. Human Phenotype Ontology is developed to describe potentially any disease, but in particular it is referred to OMIM diseases. About all of the entries present in OMIM are described with terms from this ontology.

Structure of ontologies

The structure of an ontology is very important, because it expresses the semantic relationships between entities. A very useful structure that can be used is the “directed acyclic graph” (DAG). A graph is formed by nodes linked by edges, where in our case the nodes are the terms of the ontology and edges are relationships that connect the terms.

The relationships used in an ontology are not predetermined, so any real-world relationship can be logically defined and used to connect terms and reflect the reality. This makes ontologies a flexible framework for modeling many different types of data. However, the two basic relationships used by many ontologies are “is a” and “part of” (Smith *et al.*, 2005). The “is a” relationship allows for simple, hierarchical connections between terms. The “part of” relationship is used for describing how the components of a living system fit together and it can be referred to physical parts, but can also apply to processes, such as those modeled by the Biological Process of Gene Ontology. Just to give an example, prophase, anaphase, metaphase, and telophase are all terms from the Biological Process Ontology and they are “part of” the mitotic cell cycle.

It is called directed acyclic graphs because relationships are directed, meaning that they are only true in one direction, for instance, a nucleus is part of a cell, but a cell is not part of a nucleus. All the relations imply that each node is linked to a more general node, named “parent”, as in the previous example the “cell” term is more general respect to the “nucleus” term. This is a fundamental properties of ontologies, that result hierarchical in structure but differing from “strict hierarchies”, because the directed acyclic graph allows a node to have multiple parent nodes and not only a parent to have multiple descendants. However, being the graph “acyclic”, loops are strictly forbidden, meaning that every route proceeds linearly. In this way, following the linear directions of edges each term is thus child of another more general term, by one of the semantic relations mentioned above, and so on. At the end, the common ancestor of all of the terms in the ontology is reached, which is often called “root”. The root is the most generic term and does not provide any specific information (at least, within that knowledge domain). Conversely, the specificity of the provided information increases as we move downward, away from the root, reaching the last terms that are the most detailed.

Development and maintenance cycle

The maintenance of an ontology is really an enormous work, and collaboration is essential. At the beginning, the initial version of an ontology is created by experts in the field. Every ontology is a work-in-progress, never reaching completeness, reflecting the present level of knowledge. Thus the structure is dynamic, growing and evolving as users found areas that need improvement. Some terms definitions and relationships can be initially wrong, or not properly correct, with either lack or surplus of details, or missing synonyms. It is also possible to encounter unbridgeable gaps in the ontology due to a lack of knowledge or due to an on-going or incomplete research. Errors are find also with the help of the users. All the problems are reported by the community to the experts who maintain the ontology, and the necessary check and changes are made. Then a new version of the ontology is released to the users, incorporating the required changes, and the process of maintenance go ahead in a cyclic improvements. Gene Ontology and other ontologies represent a good example of the collaborative nature of ontology development. Effectively, GO Consortium organizes also workshops where problems and relevant topics can be discussed by the scientific community.

After the build of an ontology it can be used to describe entities, for example genes and diseases. The process of assign terms chosen from an ontology to provide a description to entities is called annotation, or also biocuration when it is referred to the biological field. Genes can be annotated using terms taken from Gene Ontology and diseases can be annotated with terms from HPO. The biocuration is carried out by expert scientist. The curators read recent findings browsing scientific literature and, also with their personal knowledge, annotate newly discoveries with ontological terminology (Howe *et al.*, 2008). The advantage of the entire work is that the entities are tagged with a universal language concerning all the benefits previously described.

One of the most challenging task is probably the annotation project of Gene Ontology, that try to relates each genes with the corresponding Biological Process, Molecular Function and Cellular Component (Hill *et al.*, 2008). Curators of Gene Ontology need to identify important results from genetic, molecular, and biochemical laboratory experiments that are reported in the published literature and then annotate the gene with the appropriate ontological terms. The combination of these annotations, together with the knowledge contained in the ontology, allows striking results to be achieved.

A particular effort is required to encode scientific findings in a structured, knowledge-enhanced way using ontologies, but then, the obtained data can be used for discovering novel connections, which maximizes the overall scientific impact of all research efforts. Unveil existing relations between diseases and genes is the aim of our study.

2.2.4 Setting of the study

The discovery of all molecular mechanism in which gene products are involved is undoubted useful for the understanding of how the whole cell works in normal condition. Although, in our study, we restricted the area of analysis on the genetic diseases, because the understanding of the underling mechanism in a presence of a reported pathological phenotype can be more feasible. As previously discussed, several databases and ontologies are available generated by the effort put in the characterization of diseases and genes.

Today exists a big effort in characterizing diseases on their molecular basis and for this reason diseases with their corresponding phenotypes are collected in databases that provide us with a rich and detailed information, essentially as starting point of our study. The only analyzed diseases are associated with a gene that is clearly identify to cause the disease upon a demonstrated evidence. We can classify the source of information for setting up our study in two type: at microscopic level, regarding the biological function of genes, and at macroscopic level, regarding the observed phenotype. The **figure 10** tries to give an overview of these two level, the phenotypical and biological, provided by Human Phenotype Ontology (HPO) and Gene Ontology (GO), respectively. The terms provided by the two ontologies can be referred to each OMIM entries, that is formed by the association of two elements, the disease (D) and its causative gene (G). A Phenotypic Series is formed by a varying number of OMIM diseases.

With the benefit of availability of comprehensive databases it is possible establish comparative study among all the information related to diseases and set a “multi-level” study. The immediate benefit of this type of analysis is the identification of common elements between diseases (or Phenotypic Series) that apparently may be very distant. This systemic analysis span from biological up to clinical level, and offers a new opportunity for gaining deep understanding of the underlying mechanisms of several diseases. Nevertheless, it is foreseeable that a comprehensive study of all the molecularly-defined diseases will facilitate the identification of recurrent biological mechanisms that are dysfunctional in diseases. Diseases under examination can be clinically or genetically unrelated, but some common elements are identifiable. For example, mutations in different disease gene products (DGP) affect the same biological mechanism and therefore result in the same clinical phenotype. This is an attractive relation that produce the hypothesis we followed and expand in our analysis. The previous contention is supported by the frequent occurrence of clinical similarity even among diseases that are caused by mutations in apparently unrelated genes (Hoehndorf et al., 2015).

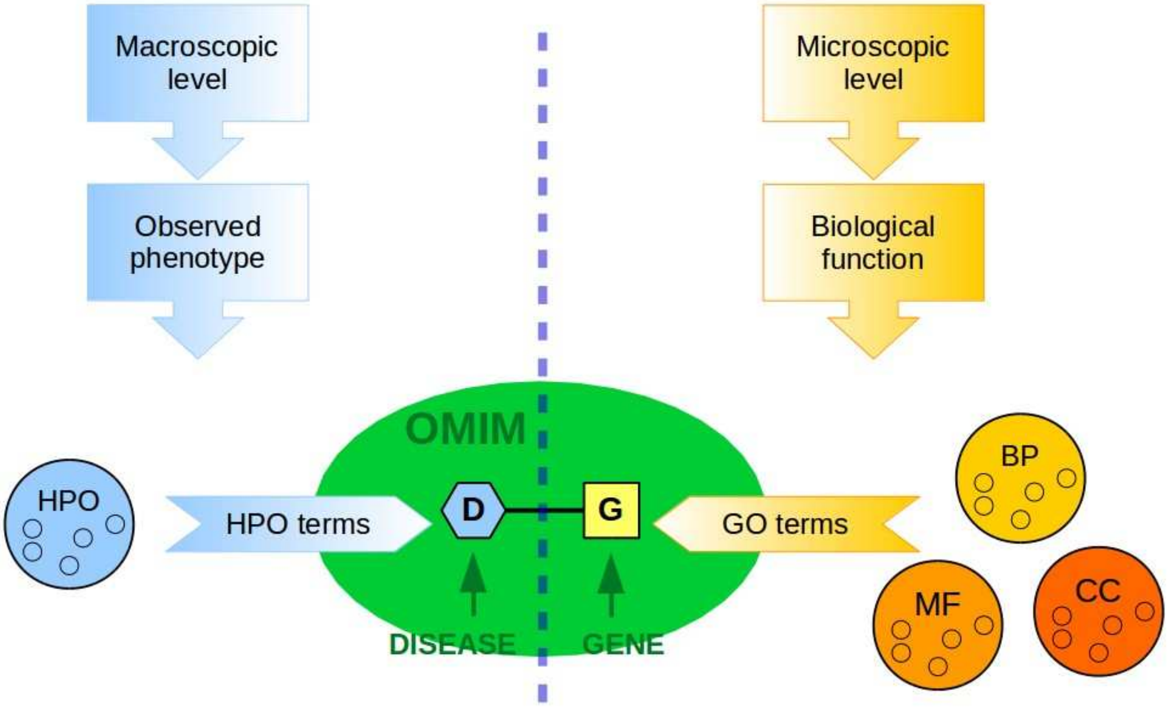


Figure 10: representation of one generic disease of the OMIM database, formed by two part, the disease (**D**) and the associated gene (**G**) that is causative of the genetic disease. Moreover, it is possible to associate more information to this central part, that are taken from external sources (in this case ontologies). On the left side we have phenotypes, that can be referred to the disease part using terms from HPO. On the right side we have biological information, that can be referred to the gene part using terms from GO. In this way we can describe both part of the OMIM disease. Note: phenotypic and biological levels are no actually connected, the image wants to represent the aim of our work, that is to put in communication these two different domains.

2.2.5 Networks

In our study, after the collection of all information from different databases, we need a powerful tool to extrapolate properties of interest from new build system. Probably network analysis is the most appropriate tool to achieve this goal, not only for the comparison of elements but also for the visualization of the results. Networks simply is formed by two types of data: nodes and edges. Graphically, nodes are positioned in a two dimensional space and are linked by a line called edge. While the nodes represent the elements of interest, such as genes, proteins or diseases, the edges represent the relation that occurs between two nodes. In this way, a network can allows a comprehensive display of complex systems, also if it is composed of a large number of elements with the respective interactions among them (Barabási et al., 2011). In our particular case, networks structure results very useful to study diseases at different levels of complexity and thus to link conceptually the molecular determinants with the clinical manifestations of disease (Vidal et al., 2011). Specifically, each Phenotypic Series can be thought of as a meta-node in the network, representing a group of similar diseases. The meta-nodes are an assumption that greatly facilitates the analysis of the interactions among all the elements involved (i.e., genes, gene products, biological characteristics, phenotypes and diseases). In this way, we have generated two networks, the Clinical Similarity Network (CSN) and the Biological Similarity Network (BSN), based on HPO and GO, respectively. In both networks, each node represents a Phenotypic Series (PS), whereas each edge linking a pair of nodes represents the degree of similarity between two PS. The similarity coefficient of a pair of PS is calculated with a procedure that take into consideration the terms of annotation assigned to gene (GO) or disease (HPO) and their information content. For the part regarding Biological Similarity Network, it is important to note that GO is divided into three domain and provide, in such way, three different type of annotation, regarding Biological Process (BP), Cellular Component (CC) and Molecular Function (MF). Such we generated three Biological Similarity Networks, BSN-BP, BSN-CC and BSN-MF. After this procedure, we also assembled a more general BSN, in which the strongest edge only (among the three sub-ontology-related BSN) is used to link two PS.

The formed networks are fully connected, at the beginning of the analysis. This means that each PS is connected in the networks to all the other PS, even though many links represent shared ancestors of low specificity. Thus, to focus on the strongest links only, we assigned a weight to each edge, which is proportional to the similarity coefficient. To remove the weaker links, we increased progressively the similarity thresholds for the clinical and the biological similarity coefficients. This way, we can analyze the two similarity networks (clinical and biological) and define groups of PS, which are characterized by different degrees of clinical and/or biological similarity. Finally, by directly comparing the two types of similarity coefficients for each PS-PS pair, we can define different modes of correlation (or lack of correlation) between the clinical and biological levels of disease.

2.3 Materials and methods

2.3.1 Data and databases

The PS, together with the corresponding diseases and the associated DGP, are retrieved from OMIM (Amberger et al., 2015). The clinical phenotypes that annotate the OMIM-encoded diseases are retrieved from Human Phenotype Ontology (HPO) (Köhler et al., 2017). The biological features that annotate the disease-related gene products are retrieved from each of the three sub-ontologies, precisely, the biological processes, the cellular components and the molecular functions of Gene Ontology (GO) (Gene Ontology Consortium, 2015). The disease categories that annotate the diseases are retrieved from Disease Ontology (Kibbe et al., 2015). In addition, HPO, GO and DO are accessed for retrieving not only the annotating terms mentioned above, but also the complete ontologies, in particular, our work requires the sets of hierarchical relations that link each term to its parent, ancestor, child and descendant terms.

2.3.2 Disease classification

In our networks, the nodes were color-coded according to the Disease Ontoloy terms annotating the majority of diseases within each PS (Kibbe et al., 2015). Specifically, the DO terms used were the eight child terms of the root 'disease' (that are: 'syndrome', 'genetic disease', 'physical disorder', 'disease by infectious agent', 'disease of metabolism', 'disease of mental health', 'disease of cellular proliferation' and 'disease of anatomical entity'), as well as the eleven child terms of 'disease of anatomical entity' (that are: thoracic disease, as well as immune, urinary, integumentary, musculoskeletal, reproductive, respiratory, cardiovascular, endocrine, gastrointestinal and nervous system disease).

2.3.3 Algorithms used and scripting language

Most of the work presented in chapter two of the present thesis was performed using computer programs. Except for network visualization and cluster analysis, all other jobs have been carried out by algorithms developed with the scripting language Python. The development of these programs is part of the work of my thesis. Python is a very easy to learn and simple to use programming language (in my opinion the most easy). It was used not only for the analysis but also for the initial data manipulation, helping in a study with so many data.

2.3.4 Network visualization and analysis

Cytoscape version 2.8.2 (Yeung et al., 2008) and Network Analyzer (Assenov et al., 2008) were used to display and analyze the two networks of this study (CSN and BSN). To select the strongest PS-PS associations only, a threshold for the edge weights was set, as described (Hidalgo et al., 2009).

2.3.5 Cluster analysis

For cluster analysis, the binary similarity coefficients between any PS_i and PS_j pair were assembled in a symmetrical matrix M composed of m rows and m columns (each row and each column corresponding to one PS), such that M_{ij} denotes the (HPO- or GO-based) similarity between the i^{th} and the j^{th} PS (self-similarity excluded). Then, hierarchical cluster analysis was performed using Multiple Experiment Viewer, setting the following parameters: Pearson correlation uncentered (as distance metrics) and average-linked (as linkage method).

2.3.6 Parameters of the networks considered

The first parameter we assessed is the connectivity *k* of a node *i*, that represents the number of neighbors of the given node. Neighbors of node *i* are only the nodes directly linked to *i*. In the **figure 11** (panel A) the node called “a” is connected to 3 other nodes, “b”, “c” and “d”, and thus its connectivity is 3, whereas the connectivity of node “e” is 4 (panel B). In our networks, connectivity represents the number of PS similar to a given PS. The connectivity does not consider if neighbors are linked with other nodes. We consider also the average connectivity of the network <*k*>, that is the average among the connectivities calculated for each node. A hight value of <*k*> indicates that, on average, each PS in the network is well connected to other PS. Here dashed lines indicate possible connection.

The second parameter examined is the clustering coefficient *C*, that differently from k take into consideration the neighbors of a given node. C simply considers if the neighbors of one node are mutually connected or not. This is expressed as the ratio between actual and maximal possible number of edges that link mutually the neighbors of one node. In the **figure 11** on panel A, the node “a” has 3 neighbors: “b”, “c” and “d”, but only nodes “c” and “d” are connected, so the ratio is one link among 3 possible, and C is equal to 1/3 (or 0.33). It is possible calculate *C* for each node in the network and then obtain the average *C* (or <*C*>), that is in the range 0 to 1. A hight <*C*> indicates that a hight percentage of the PS, which are clinically similar to a given PS, are similar to each other as well.

Next parameter is Density, that is a general parameter referred to entire network taking into consideration how many edges are present respect the possible number of edge, so is the ratio between the given network and a fully connected network. Density can be in the range between zero and one, and in the **figure 11** panel A, being maximal number of possible edges equal to 6 and the present edges 4, it is 4 divided 6.

The last parameter considered is the length (*l*) that is the shortest path between two nodes. Given two nodes *i* and *j*, *l* is the minimal number of edges that put in connection the two nodes (**figure 11**, panel B). The average *l* (or <*l*>) is calculated between each possible pair of nodes, and a low value of *l* means that network is also well connected.

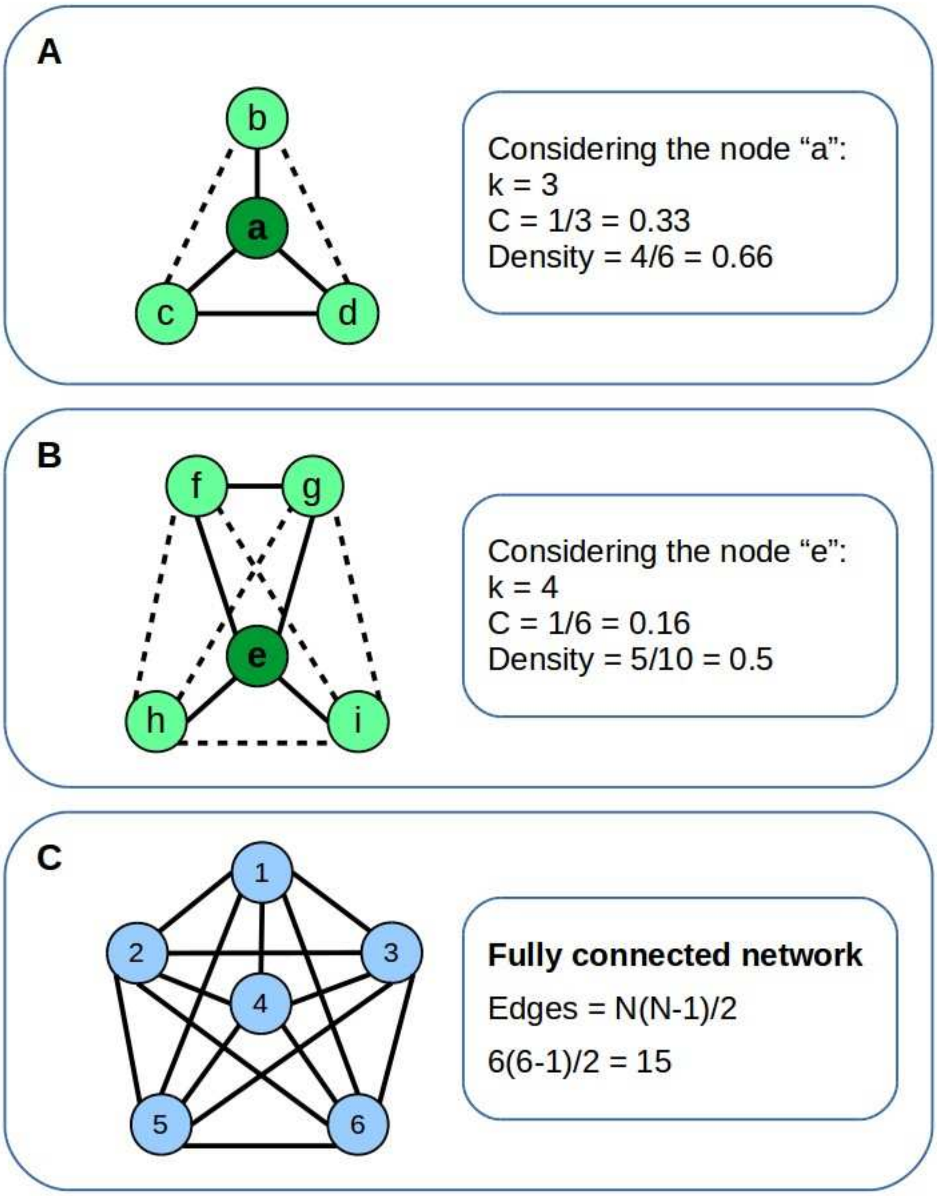


Figure 11: examples of the parameters considered and network properties.

2.4 Results

2.4.1 Similarity coefficient

The method developed for comparing phenotypic series is part of the result of this thesis, and for this reason it is not discussed into methods section.

As we can understand, a key feature of entire study is the analysis of the similarity between phenotypic series. Being a value, similarity can be calculated and expressed as a number using our algorithm. Different methods are today available for calculating the similarity between either two diseases or two genes, but not for phenotypic series. In fact, phenotypic series are not single diseases or genes, but groups of them, which have been clustered using several criteria related to a similar phenotype. Therefore, we needed a different method for calculating the similarity between phenotypic series. In general, one utilizes the annotation terms, which are referred to either the phenotypes of a disease or the biological functions of a gene (see schema in **figure 10**). In our specific case, however, we could not use the annotation terms of the single entity, but rather we needed, first, a method to obtain general annotations for the entire group of entities. Only after this step, it was possible to calculate the similarity between phenotypic series with the same strategy that other investigators have used for calculating the similarity between individual diseases or genes.

2.4.2 Information content

In general, the information content of a given term is a numerical value indicating how much information is contained within that term (for instance, the information on either a gene or a disease annotated by that term). The concept of information content can also be applied to ontological terms as well. In the latter case, the information content indicates the specificity of the term within the ontology. The information content is calculated based on the frequency of the given term, and the frequency is defined as the proportion of objects that are annotated by the term or any of its descendant terms of the ontology (see **figure 12**).

For example, if a disease is annotated with the term “atrial septal defect” in Human Phenotype Ontology, it is also implicitly annotated with all the ancestors of the term, because the term has the meaning also of its ancestors, as “atrial septal defect” is an “abnormality of the atrial septum”, that, at a more general level, is also an “organ abnormality” (see **figure 12**). Always each of the terms converges at the root of its ontology, and therefore the root is implicitly always present in any annotation. It results that the information given by the root term is null and takes the value of zero.

Different terms can be required for the description of either a gene or a disease. Actually, each entity is normally annotated with many terms. Conversely, it is possible to have a situation of very poor or incomplete provisional annotations (this happens in particular for some genes of not completely known function or role).

A high value of information content calculated for a term reflects its high quality in information. Terms of this kind are normally located far away from the root. However, the algorithm for the calculation of information content does not take into consideration only the position of the term in the hierarchy but also its frequency within all annotated entities. The information content is defined as the negative natural logarithm of the frequency of usage (Cover et al., 1991).

Schematically, the procedure is depicted in **figure 13, 14** and **15.**

For example, the information content of a given term is calculated on the basis of its presence within the whole database of annotation. Specifically, the Human Phenotype Ontology provides terms that annotate 4,813 different OMIM diseases. The terms “atrioventricular block” is used to annotate three diseases among the total number of diseases (4813), so that its information content is:

$-\log(3/4813) = 7.38$

The term “abnormality of the musculoskeletal system” is more general, as we can easily understand from its name. Actually, it is used to annotate 2,352 diseases. The information content in this case is:
-log(2352/4813) = 0.72.
We can briefly see, with this last examples, how the information content tends to increase, as we move from the root term of the ontology down to more specific descendant terms.

In the past decades, several similarity measures have been proposed (Resnik, 1995; Jiang et al.,1997; Lin, 1998) and they can find an application also for biomedical ontologies. We chose to calculate the similarity between two terms as described by Resnik (Resnik, 1995). Briefly, similarity can be calculated as the information content of their most informative common ancestor (MICA). The “most informative” in general corresponds to the first common ancestor found. For example, in the **figure 12**, “abnormality of the cardiac septa” meets “abnormality of the cardiac atria” exactly at the term named “cardiac malformation” and their similarity will be the information content of this term, that corresponds to the most informative common ancestor.

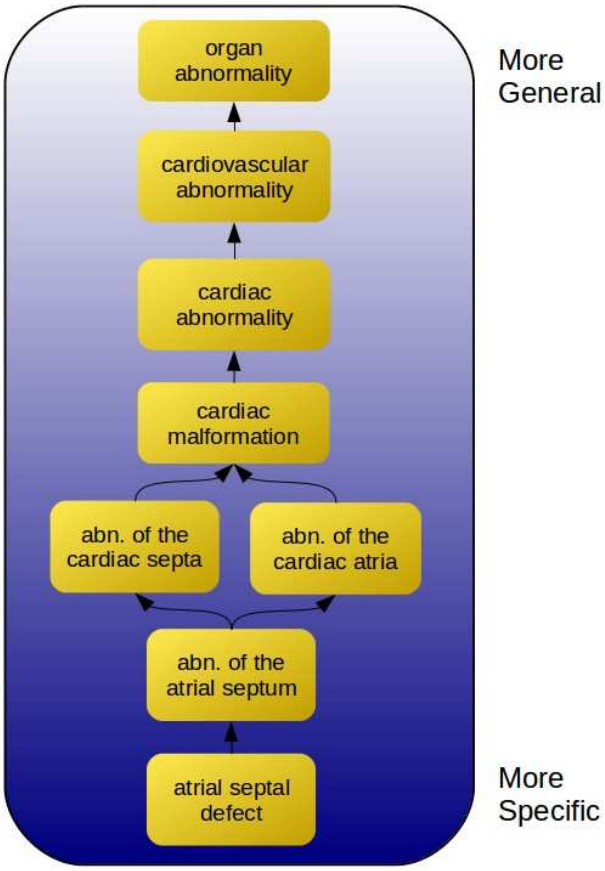
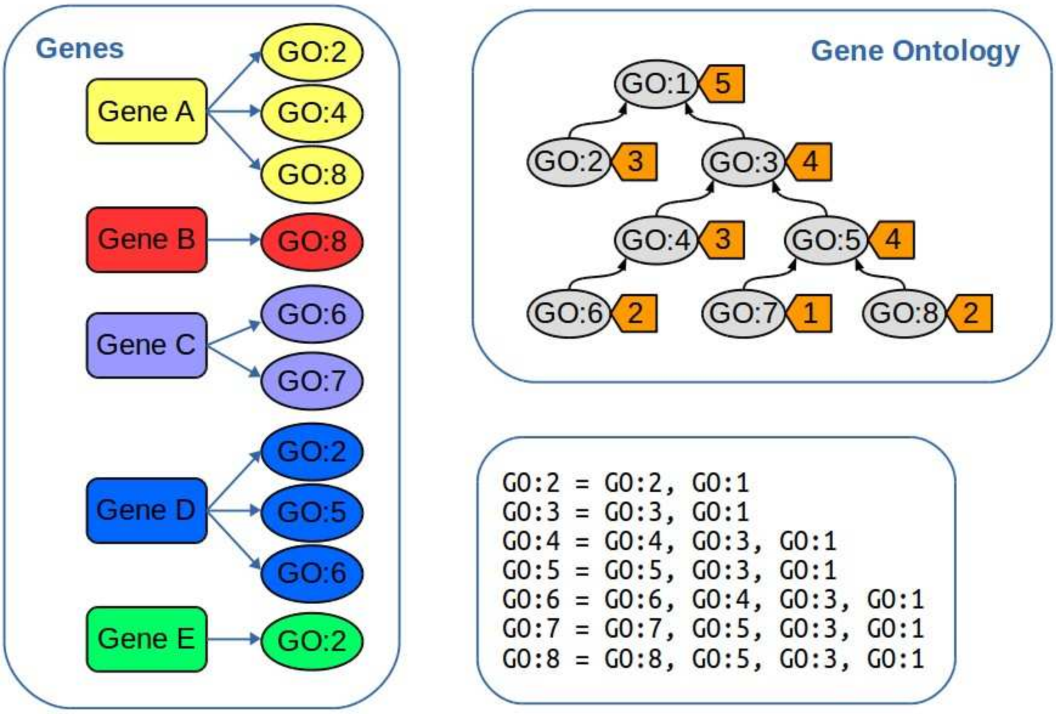
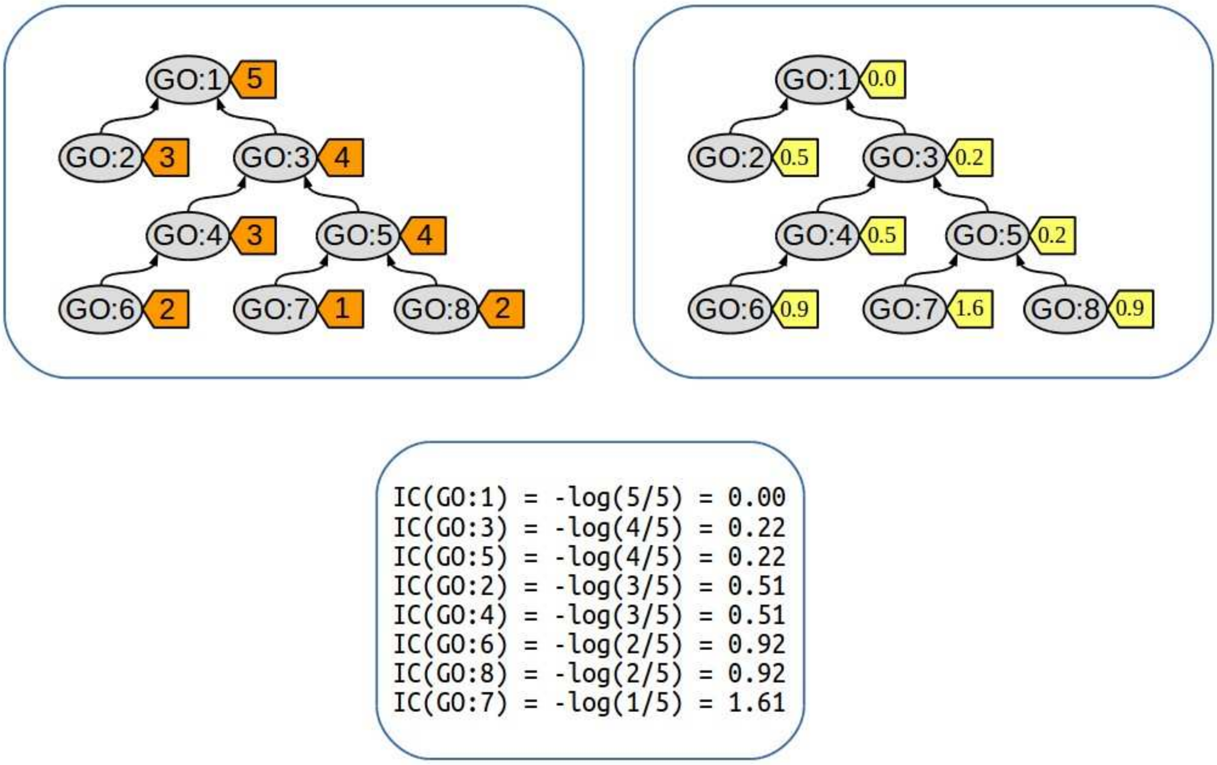

organ abnormality
cardiovascular abnormality
cardiac abnormality
cardiac malformation
abn. of the cardiac septa
abn. of the cardiac atria
abn. of the atrial septum
atrial septal defect
More General
More Specific

Figure 12: a small part of the HPO. Example from the specific term “atrial septal defect” to a more general term. Arrows indicates the direction of the “is a” relationship.



```
GO:2 = GO:2, GO:1
GO:3 = GO:3, GO:1
GO:4 = GO:4, GO:3, GO:1
GO:5 = GO:5, GO:3, GO:1
GO:6 = GO:6, GO:4, GO:3, GO:1
GO:7 = GO:7, GO:5, GO:3, GO:1
GO:8 = GO:8, GO:5, GO:3, GO:1
```

Figure 13: in this example we start in the calculus of information content for the terms of annotation. In our library (that means all genes) we have **five genes**: A, B, C, D and E (in real world we have thousands genes, but it is a an example very simplified). Each gene has **terms** of Gene Ontology associated (normally ten or more, but in the example just 1, 2 or 3). The terms are taken from an **ontology** and are used to describe the genes. Then we proceed counting the **usage** of these terms. For example the term GO:8 means not only GO:8, but also GO:5, GO:3, GO:1 (all its ancestors). Due to this property, the root term (GO:1) is always present (in all five genes) and its count is **5** (**orange tag**), as the count of GO:7 is only one (only one gene: precisely the gene C is the only with that term). The procedure continue in image 14.



```
IC(GO:1) = -log(5/5) = 0.00
IC(GO:3) = -log(4/5) = 0.22
IC(GO:5) = -log(4/5) = 0.22
IC(GO:2) = -log(3/5) = 0.51
IC(GO:4) = -log(3/5) = 0.51
IC(GO:6) = -log(2/5) = 0.92
IC(GO:8) = -log(2/5) = 0.92
IC(GO:7) = -log(1/5) = 1.61
```

Figure 14: the information content of each term of an ontology is calculated as the **negative natural logarithm** of its **frequency** (see also the image 13 and the text for details).

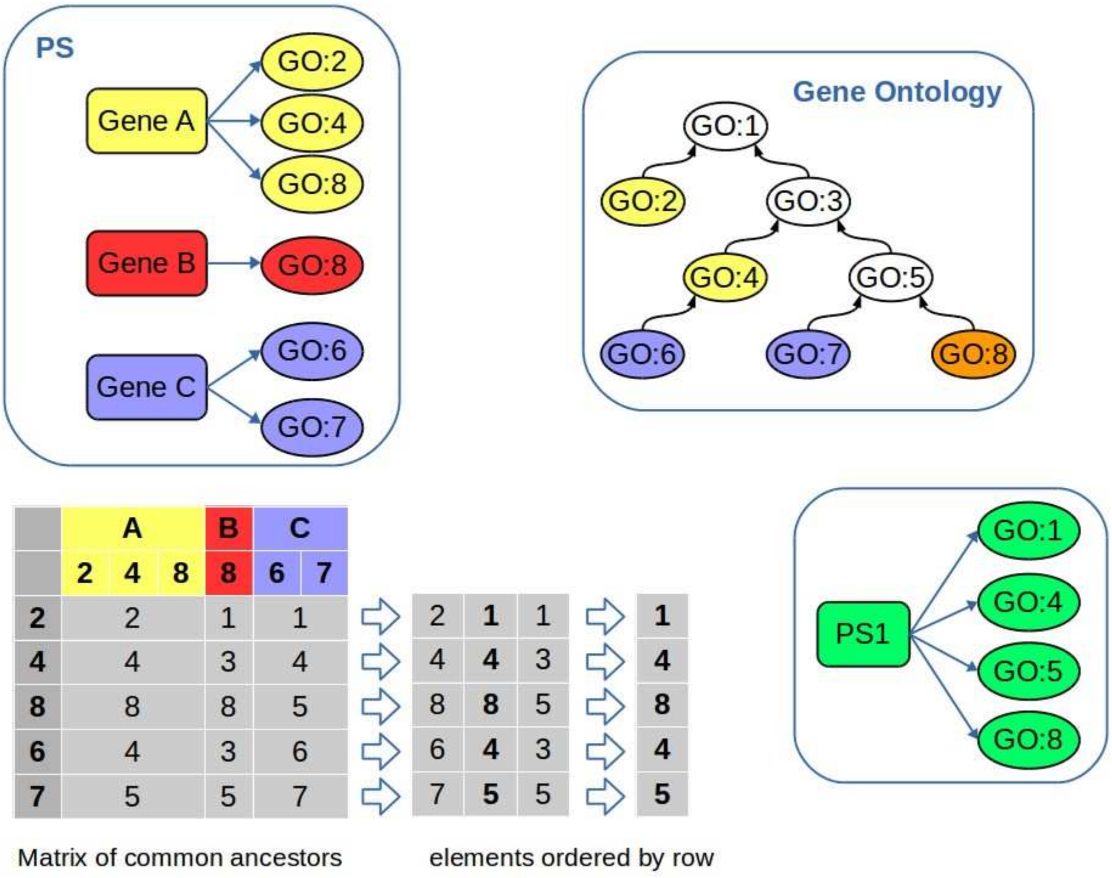
2.4.3 Find annotation terms for the PS

By describing in detail how the Phenotypic Series are compared to each other, we will show how the algorithm works. As previously anticipated, to achieve this goal, it is first essential to identify the annotation terms that are referred to each PS. As mentioned above, each PS is composed of diseases (identified in the OMIM database) and, in turn, each OMIM-defined disease is associated with one gene, whose mutation reportedly causes the disease. Therefore, starting from the annotations of genes and diseases that form the PS we can get new annotations. Obtained annotations will be intermediate terms with respect to the original genes and diseases.

By using the annotations, the method described here can be applied to either diseases or genes, and to any other type of entity with annotations in an ontology.

The procedure is depicted in **figure 15** schematically and explained here in detail. As an example, we have a PS with three associated genes (A, B and C), each of which has a given number of Gene Ontology (GO) annotations. Here, the gene A, B and C are annotated with 3, 1 and 2 number of terms, respectively. Each of the annotation terms (referred to each gene in the PS) is compared with all the annotation terms of the first gene, until a common ancestor is found, which is the closest to both terms under exam. It should be noted that the closest, and not the most informative, term is selected (however the procedure is the same as used in Resnik similarity). Then, the same procedure is performed between the terms of the second gene, and so on, until all the genes in the PS have been analyzed. The selected terms are then sorted, using the same principle adopted for selecting them, i.e., their distance from the root. At the end, the 'median term' (that is, the term in the middle) is used as new annotation term for the PS. In a list of ordered terms, the central term is used, in case of an odd number of annotations, whereas the term immediately before the middle one is used, in case of an even number of annotations. In the specific example reported in the **figure 15**, in which we have three genes and an ordered list of three annotations, the central annotation is used (indicated with bold character).

By and large, this method selects new terms that are more general than the original terms (i.e., the ones used for annotating the single entities), even though it can also select the same original term. Actually, even highly specific terms can be retained, if they are present in at least half of the entities that form the PS, without loss of information. In other cases, the closer intermediate term is selected. In contrast to a small loss of information, the advantage of the procedure is to produce annotations that are more representative of a group of heterogeneous elements. Moreover, we can remove terms of annotation that are highly specific but very uncommon in the PS, thus avoiding the risk of overestimating terms that are not representative of the general characteristics of the PS.



	A			B	C	
	2	4	8	8	6	7
2	2			1	1	
4	4			3	4	
8	8			8	5	
6	4			3	6	
7	5			5	7	

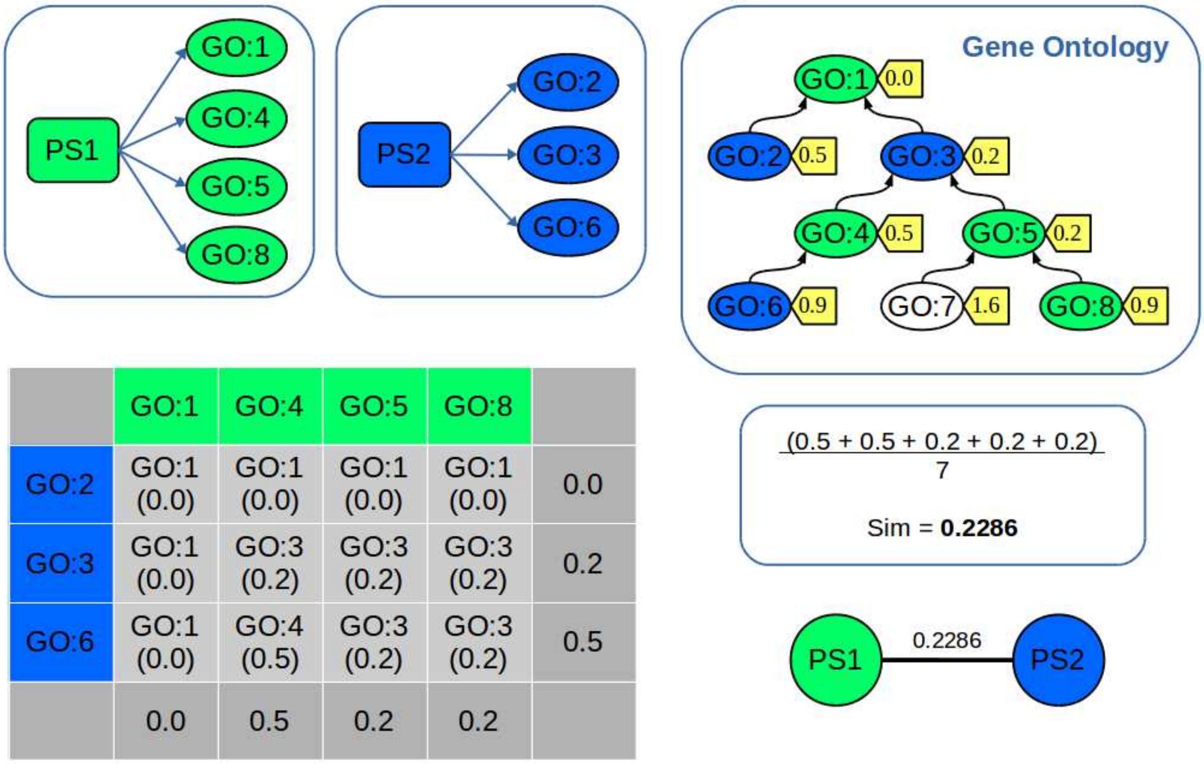
2	**1**	1	**1**
4	**4**	3	**4**
8	**8**	5	**8**
6	**4**	3	**4**
7	**5**	5	**5**

Figure 15: see the text for details.

2.4.4 Similarity between PS

After generating new annotation terms for each individual PS, it is possible to proceed with the comparison between PS. Similarity is calculated between only a pair of different PS and the direction is not important (the similarity between A and B is exactly the same as the similarity between B and A). Among different methods proposed for the calculation of similarity between two elements using their annotations, we chose the Best-Match Average (BMA) method, that, in our opinion, is the most complete (Wang et al.). Briefly, it consists of an average of the maximal values between all the most common ancestor found for each terms of the first and second PS compared.

Looking at the **figure 16**, we can imagine to build a big matrix. We denote the two PS to compare as PS1 and PS2, and with color green and blue, respectively. On the matrix, we put all the terms of the former PS along one axis and all the terms of the latter PS along the other axis. Then, we take the maximal most informative common ancestor of each row and each column of the matrix. At this point, the arithmetic mean of these values is calculated. This value correspond to the similarity coefficient.



	GO:1	GO:4	GO:5	GO:8	
GO:2	GO:1 (0.0)	GO:1 (0.0)	GO:1 (0.0)	GO:1 (0.0)	0.0
GO:3	GO:1 (0.0)	GO:3 (0.2)	GO:3 (0.2)	GO:3 (0.2)	0.2
GO:6	GO:1 (0.0)	GO:4 (0.5)	GO:3 (0.2)	GO:3 (0.2)	0.5
	0.0	0.5	0.2	0.2	

$$\frac{(0.5 + 0.5 + 0.2 + 0.2 + 0.2)}{7}$$

Sim = **0.2286**

Figure 16: similarity calculated between two generic PS.

PS1 has **4 terms** of annotation from Gene Ontology (GO): 1, 4, 5 and 8.

PS2 has **3 terms** of annotation from Gene Ontology (GO): 2, 3 and 6.

Each of the 4 terms of PS1 (**green**) is matched with each of the 3 terms of PS2 (**blue**). In this way we obtain values from **4 columns** plus **3 rows** (7 values) that are then averaged, obtaining the similarity coefficient (edge weight).

2.4.5 Assembling the networks

First, we obtain from the OMIM web-site all the diseases forming the PS. We retain only the OMIM diseases that are molecularly defined, assembling in this way a database that enlists a starting set of 319 PS.

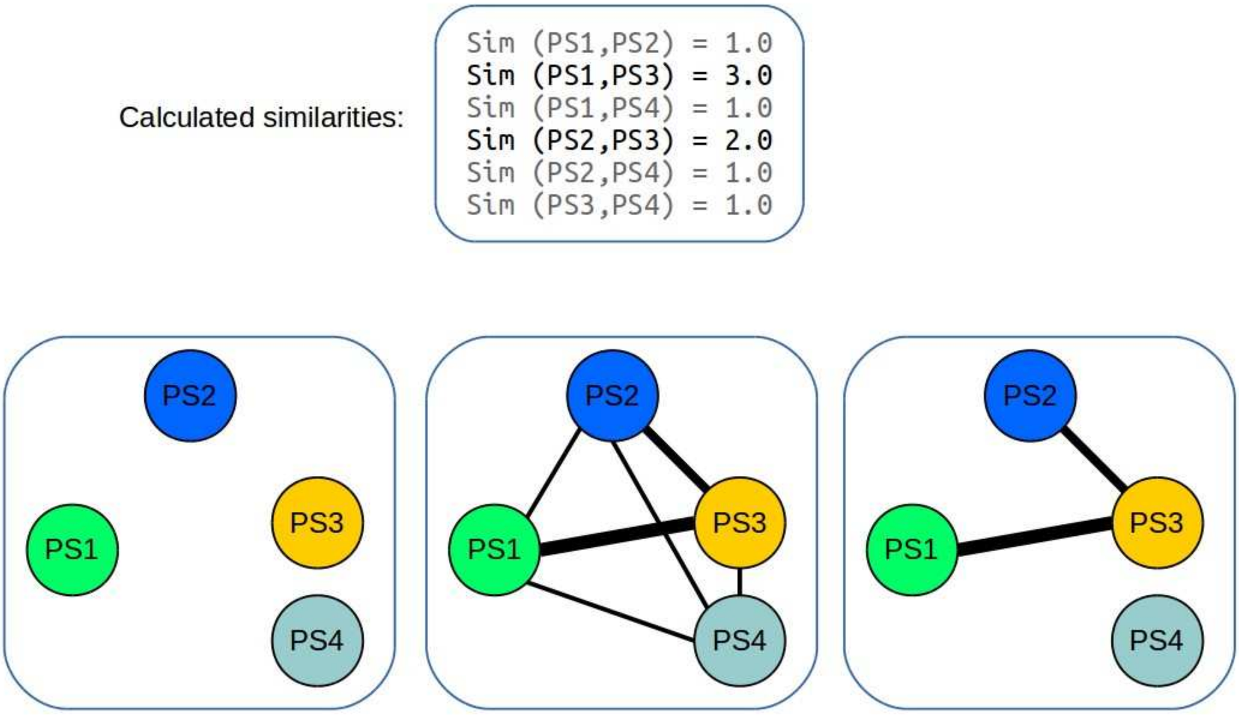
Then, being each PS formed by an association of different diseases with respective genes, we retrieve the annotation terms referred to each of these diseases and genes, from HPO and GO, respectively.

At this point, starting from original terms of annotation, we can calculate new terms for the annotation of the whole PS. We use the procedure illustrated in the following section.

Then, using the newly obtained terms, it is possible begin the calculation of similarity score between PS, two by two. In order to have a complete view we want obtain all possible similarity, eliminating then the not significants of them. For this reason we perform, using our algorithm, the calculus of similarity between all possible pairs of PS. This is an enormous number, because with a given number of PS, “N”, we have a number of PS-PS pairs that is expressed as: N(N-1)/2. For example with 319 PS the calculated similarities are 50,721.

Now, it is possible assemble a network in which PS are the nodes, whereas calculated similarities are the edges. The resulting network is fully connected, meaning that each node is linked to each other. An example of a fully connected network is presented in **figure 17**, as we can see the number of edge can be calculated with the formula N(N-1)/2.

We calculate not just one but different type of similarity, and in particular, it is possible assemble two type of network that we named **“Biological Similarity Network”** (BSN) and **“Clinical Similarity Network”** (CSN). The Biological Similarity Network is based on similarity calculated with GO annotations, whereas the Clinical Similarity Network is based on HPO annotations. Moreover, the BSN comprehends three networks. Effectively, GO is divided into three independent domains, BP, CC and MF, and thus we are



```
Sim (PS1,PS2) = 1.0
Sim (PS1,PS3) = 3.0
Sim (PS1,PS4) = 1.0
Sim (PS2,PS3) = 2.0
Sim (PS2,PS4) = 1.0
Sim (PS3,PS4) = 1.0
```

Figure 17: procedure to build a generic (very small) network of Phenotypic Series (PS). In the example 4 nodes are present: PS1, PS2, PS3 and PS4. After the calculation of similarities among all pairs of PS, only the edges with a weight 2 and 3 are retained, the others are eliminated. Then, also the node PS4 is removed from the network, because it is not connected to the others.

able to build three independent networks, based on these three type of annotations. We call them BSN-BP, BSN-CC and BSN-MF. Moreover we decide to generate an additional network starting from the three BSN. This last is simply called BSN and among the three similarity calculated (BP, CC and MF) use the one with the highest value.

It is worth remember that also two completely different PS have a value of similarity, because every ontological term meets anyone of the other terms at the ancestor term, called root of ontology. So, every term has at least one ancestor in common with another, and for this reason it is always possible calculate the similarity. As previously observed, the information content provided by the root is null and in this case the calculated similarity is zero. By the way, we take all similarities, also if equal to zero and this results in a network where every possible connection is present. Fully connected network are not useful because they are not very informative. Obviously, we want remove very weak edges, retaining only strong connections, in order to carry out our analysis. We proceed assigning a weight *w* to each edge, where *w* is equal to the similarity. Then we gradually increase the threshold for *w*, in order to remove the weaker similarities, while retaining the edges that link the most similar PS (**Figure L1** panel **A**). We found that for both CSN and BSN the majority of edges have a very low value of similarity, as described in detail in the next section.

2.4.6 Analysis of the Clinical Similarity Networks (CSN)

At the beginning CSN is a fully connected network composed of 293 nodes linked by 42,778 edges (**table 1**).

Each PS is connected to another PS but the majority of these connections represent a very weak PS-PS similarity. Nevertheless, by gradually increasing the threshold for *w*, it is possible retaining the most significant edges only. Effectively, setting a threshold > 1.0, the 87% of the least specific edges were removed, but the number of nodes remains practically unchanged, with just 2% of them removed. Specifically, the CSN retains 287 nodes but only 5,660 edges and was still composed of one giant connected component, meaning that network is not fragmented in more part and every node is connected through a path to another node (indirectly) (see the **table 1** and **2**).

The average connectivity <*k*> of CNS is 39.4 indicating that, on average, each PS in the CSN is connected to about 39 other PS. Interestingly, the connectivity is not homogeneous in whole network but differs depending by the clinical class, that is assigned based on Disease Ontology.

PS in the class “Syndrome” have the highest <*k*> (66 ± 33), which likely reflects the phenotypic variety of these diseases. Effectively, syndromic diseases are more connected because often interests more organs and have multiple symptoms and this results in have similarity with different PS. Then we considered the average clustering coefficient <*C*> of CSN, which is 0.635 indicating that 63.5% of the PS, which are clinically similar to a given PS, are similar to each other as well. As last, the average *lenght* (<*l*>) is equal to 2.29 indicating that, even for those pairs of PS that are not similar to each other (and thus are not linked directly in the CSN), there is at least a third PS that is similar to the two unconnected PS. Taken together, the high <*k*>, the high <*C*> and the low <*l*> indicate that, on average and at a threshold > 1.0, most PS already display a significant degree of clinical similarity to each other.

2.4.7 Islands and clusters of high clinical similarity within the CSN

To identify examples of the high clinical similarities, the threshold was further increased until, at a value of 2.46, about 20% of PS, but only 0.2% of PS-PS similarities, are retained (i.e., 58 nodes and 63 edges; **figure L1** and **L2**). It is important to note that at this threshold the number of edges is similar to the number of nodes. This great reduction, not only removes most of the edges, but also a good part of the nodes. In our opinion this is important, because effectively some PS do not show any similarity (or at least strong similarity) with other PS, and are correctly excluded by the analysis.

Moreover with this elimination the CSN becomes much more disconnected and result divided into 13 fragments. These are called islands, and represent sub-network not communicating with other islands. The PS inside an island are highly similar with respect to the phenotypes but the data are also in accord with the classification of Diseases Ontology (DO), represented by *node colors*. The highest similarity coefficient measured is 4.394 and connects two PS named “Seizures, familial febrile” and “Epilepsy, generalized, with febrile seizures, plus”, corresponding to the ID PS121210 and PS604233, respectively. These PS provide an example of how the coefficient reflects the sharing of highly informative phenotypes, including both identical terms (e.g., “Febrile seizures”) and common ancestors (e.g., “Dialeptic seizures”, which is the ancestor of both “Focal seizures with impairment of consciousness or awareness” and “Absence seizures” in PS604233 and PS121210, respectively).

However, CSN fragmentation does not imply that disconnected PS are completely dissimilar, but merely that their similarity is below the threshold. It remains possible that even some disconnected PS islands are similar (to each other) and dissimilar (to the other PS). To this purpose, the CSN is analyzed also by hierarchical cluster analysis, using the same threshold (**figure L7**). In this way, we identify eight clusters (*red areas along the diagonal* in **figure L7**), many of which are mostly composed of clinically similar groups of PS islands (*dotted lines* in **figure L2**), e.g., convulsive (**cluster 1**), renal (**cluster 4**), inflammatory (**cluster 5**), proliferative (**cluster 7**) and cardiovascular (**cluster 6**) PS. In addition, clustering also allow grouping similar PS that, at this threshold used, are disconnected. For instance, seven PS, scattered in three small islands, could be grouped into **cluster 8**, which is characterized by ocular involvement (even though the PS affect different components of the eye, such as cornea, iris and retina).

The case of "Inflammatory Bowel disease" and "Colorectal cancer"

It is worth noting that the similarity among the PS may extend beyond the boundaries of the previously considered clusters (*red areas off the diagonal*). For instance, "Inflammatory Bowel Disease" (PS266600) inserted in cluster 5 and "Colorectal cancer, hereditary non-polyposis" (PS120435) in cluster 7, have a similarity coefficient of 2.17, which is 4.2-fold higher than the average similarity in the whole CSN (0.52 ± 0.72), but lower than the threshold of 2.46. for this reason the two diseases are not connected, and result also to be associated to different clusters. However, their hight value of similarity is mostly due to the shared phenotype "Abnormality of the large intestine". In effect, Inflammatory Bowel Disease is annotated with term "Rectal abscess", whereas Colorectal cancer, is annotated with term "Colon cancer".
Inflammatory Bowel Disease is principally characterized by inflammatory process at intestinal level, fistula, abscess also in colon-rectal trait, including ulcer, folliculitis, proctitis.

2.4.8 The Biological Similarity Networks (BSN)

Using the same procedure previously described for the assemblage of CSN, we build the Biological Similarity Networks (BSN) that is not unique, but divided into three networks representing the fact that they are derived from the similarity calculated based on the three ontologies.

We also assembled an additional network, simply called BSN, in which each edge (linking any two PS) is the strongest (i.e., the one with the highest *w*) among the three sub-ontology-related BSN.

Like the CSN, also the BSN-BP, BSN-CC, BSN-MF and BSN are fully connected networks, even though most edges represent weak PS-PS biological similarities (**table 1**). Thus, the threshold was gradually increased, to focus on the most significant similarities only (**figure L1** *panels B-E*).

A threshold of 1.0 was used for network analysis. At this threshold, all the four biological networks are still composed of one GCC, despite the loss of many edges. However, they differ in the fraction of nodes and edges that are retained (with respect to the initial conditions, i.e., at a zero threshold). As reported in **table 2**, more edges are retained in the BSN-BP than in either the BSN-CC or (to an even greater extent) the BSN-MF. Consequently, at the threshold of 1.0, more nodes remain connected to the GCC, thus indicating that BP generates more informative edges than the other two sub-ontologies. In addition, compared with the BSN-CC and the BSN-MF, the BSN-BP is also a more dense networks, whose nodes have more connections and lie at closer distances to each other. Finally, all these parameters are even greater in the BSN, which is not unexpected, as each edge in the BSN corresponds to the most informative edge among the three sub-ontology-restricted BSN.

2.4.9 Islands and clusters of high biological similarity within the BSN

The threshold is further increased until, at the values indicated in **figure L1** (panels **B-E**), about 20% of the PS are retained and the networks are fragmented into islands of biologically similar PS. As for the CSN, hierarchical cluster analysis is applied to identify clusters of PS and PS-containing islands within the BSN-BP (**figure L3**), BSN-CC (**figure L4**), BSN-MF (**figure L5**) and BSN (**figure L6**). Most of the clusters carry a clear biological meaning, which can be interpreted by searching for the GO terms that most significantly annotate the PS-related DGP. These terms are

indicated in the Figures as labels for the clusters (or, for subsets of nodes within large clusters). For instance, most PS clusters in the BSN-CC (**figure L4**) correspond, according to this criterion, to diseases (e.g., the ciliopathies) that are specifically due to defects in a defined sub-cellular structure (e.g., the cilium; **figure L4**, *cluster 7*).
The cilium constitute an organelle very important for eukariotic cell. They are involved in many disease that involves motile cilia.

Notably, in the largest cluster of the BSN-BP (**figure L3**, *cluster 8*), the majority of PS are diseases of cardiovascular system, and in particular ion-channel disorders (Cerrone et al., 2012), which are characterized clinically by an arrhythmic phenotype. Among them, the highest similarity coefficient (4.44) links the “Brugada syndrome” with the “Short QT syndrome”, because the two PS share numerous and highly informative processes, including the BP annotation “Potassium ion export from cell”. Surprisingly, however, some PS in the same cluster are not cardiologic diseases, but rather convulsive (e.g., “Seizures, benign familial infantile”), respiratory (“Bronchiectasis”), renal (“Bartter syndrome”) and other syndromic (“Familial episodic pain syndrome”) disorders. The same PS cluster together also in the BSN-CC (**figure L4**, *cluster 8*) and the BSN-MF (**figure L5**, *cluster 5*). The reason behind such heterogeneity is that most of the DGP participate in the activity of transporting ion across the membrane, at the level of voltage-gated (cardiac, convulsive and episodic pain-related disorders) and ligand-gated (bronchiectasis) sodium and chloride channels (Bartter syndrome). In the GO terminology, a defined MF takes place in distinct CC to carry out specific BP. Thus, the different ontology-related BSN, albeit at different levels of informational contents, converge in identifying a broad community of clinically diverse (but biologically related) disorders.

Table 1: all five networks fully-connected (all possible edges are present for the given number of nodes)

	Nodes	Edges
CSN	293	42,778
BSN-BP	316	49,770
BSN-CC	314	49,141
BSN-MF	305	46,360
BSN	319	50,721

Table 2: all five networks in which only similarities > 1 are retained (edge threshold = 1). Also four network parameters are calculated (as described in Materials and Methods section).

	Nodes	Edges	<*k*>	<*C*>	<*l*>	Density
CSN	287	5,660	39.440	0.635	2.294	0.138
BSN-BP	298	11,223	75.320	0.708	1.950	0.254
BSN-CC	258	2,821	21.870	0.653	2.486	0.085
BSN-MF	209	1,633	15.630	0.691	2.829	0.075
BSN	311	13,282	85.420	0.654	1.799	0.276

2.4.10 Correlations of the CSN and the BSN

So far, the clinical and biological similarities have been considered separately. Next, however, we compared the clinical and biological similarity coefficients of all the 42,782 PS-PS pairs, for which both coefficients were available. The resulting scatter plot indicates that the vast majority of PS pairs have low clinical and low biological similarity (**figure L9** panel **A**). However, there are other instances of potential interest.

First, many PS pairs have high clinical and high biological similarity (both coefficients being ≥ 2 in 57 PS pairs). For example, many of the ion channel-related cardiac arrhythmias display the highest similarity in both clinical and biological terms (**figure L9** panel **B**, *red diamonds*). Similarly, few convulsion-related PS display medium-high levels of clinical and biological similarity (*green diamonds*). The graphical equivalent of these correlations are the *thicker edges* in the high biological similarity view of the BSN shown in **figure L6** (cluster 8), where edge thickness is proportional to the clinical similarity.

Second, many pairs have high biological but low clinical similarity, their similarity coefficients being ≥ 2 (biological) and < 2 (clinical) in 500 pairs. Interestingly, many heterogeneous PS pairs composed of one cardiac arrhythmias and one convulsion-related PS are examples of this condition (*yellow diamonds* in **figure L9 B** and *thinner edges* in **figure L6**, cluster 8).

Third, many PS pairs have low biological but high clinical similarity, their similarity coefficients being < 2 (biological) but ≥ 2 (clinical) in 178 PS pairs. Among others, the renal diseases in of the CSN (*purple diamonds* in **figure L9** panel **B** and *thinner edges* in **figure L2**, cluster 4) are examples of clinical similarity, which reflect, on one side, the sharing of highly informative phenotypes (e.g., “proximal tubule” nephrolithiasis, bone pain and fractures), but, on the other hand, biological alterations as diverse as defective endocytosis in Dent’s disease (Devuyst and Thakker, 2010), defective glyoxylate metabolism in primary hyperoxaluria (Cochat and Rumsby, 2013) or decreased renal absorption of phosphate in the hypophosphaturic nephrolithisais (Sayer, 2017).

2.5 Discussion

As a result of our analysis, we have obtained three findings.
First, networks, the principal tool used in the study, are well suited to give a general overview of biological and clinical similarity existing among molecularly characterized OMIM diseases, and particularly among Phenotypic Series (PS). Moreover, using weighted connections, represented by the thickness of edges, it is possible to observe, both locally and globally, the level of similarity.
Second, the gradual increase of threshold for the similarity between Phenotypic Series, allows the identification of different clusters with progressively higher similarity, from both the clinical and the biological point of view.
Third, perhaps most importantly, the clinical and the biological similarities are not always directly correlated, allowing the observation of particular situations in which one outnumbers the other one. These types of lack of correlation, lead to the formulation of potential pathogenic mechanisms and the identification of possible molecular targets for pharmacological intervention.

The use of networks in medicine and pharmacology is not unprecedented (Barabási et al., 2011; Vidal et al., 2011, Bazzoni et al., 2015; Hopkins, 2008) , also with purpose of investigating the clinical similarity among diseases (Hoehndorf et al., 2015). The novelty of the present study, however, is three-fold, because, in parallel to the clinical similarity of the diseases, we have also examined the biological similarity of the Disease Gene Products (DGP), and then we have analyzed the correlation between these two types of similarity. Finally, we have based the analysis not on individual diseases but rather on all OMIM-defined Phenotypic Series (Amberger et al., 2015), which can be regarded to as “meta-nodes” in the networks. The use of Phenotypic Series, in place of the individual diseases, is justified by the widespread occurrence of locus heterogeneity in human genetics, whereby mutations in different genes cause similar diseases that are hardly (if at all) distinguishable at the clinical level. Thus, by merging similar diseases into the same Phenotypic Series (or, in graphical terms, by reducing thousands of disease nodes into just few hundreds of Phenotypic Series meta-nodes), the Phenotypic Series have greatly simplified our analysis of the networks. An alternative approach (focused on the individual diseases) would have required comparing millions of disease-associated phenotypes and then identifying clusters of similar diseases. Thus, we argue that the use of the Phenotypic Series is more advantageous, because it is computationally less demanding and relies more on authoritative clinical judgment (rather than on the arbitrary choice of a similarity threshold to define the boundaries of the clusters).

Apart from the Phenotypic Series, our study has also benefited from the availability of similarity metrics that have allowed us to quantify the similarity of both the clinical (Human Phenotype Ontology-defined) phenotype annotations (Hoehndorf et al., 2015) and the biological (Gene Ontology-defined) cellular annotations (Wang et al., 2007) that are shared by Phenotypic Series pairs. As both Human Phenotype Ontology and Gene Ontology are ontologies, in which each term is a specific instance of a more general parent term, the similarity methods identify the shared term within the hierarchical structure of the ontology. Thus, many shared terms are common ancestors endowed with various degrees of information content. At the lowest extreme of the spectrum, even the most dissimilar annotations share, as common ancestor, the root, which is the least specific term of the ontology (and thus, the one with the lowest informational content). The downside of retrieving the informational content of an annotation from an ontology is that, at the end, each Phenotypic Series has some similarity with all the other Phenotypic Series. As a further consequence, the resulting network has limited usefulness, because each of its N nodes is connected to all the other (N-1) nodes by means of all the theoretically possible N(N-1)/2 edges. Nevertheless, assigning a similarity coefficient to each PS pairs allows converting the fully-connected network into a weighted graph (Barrat et al., 2004; Newman, 2001), in which the weight of each edge characterizes the strength of the similarity. Importantly, by increasing the threshold for such weight, it becomes feasible to retain the strongest links only, thereby making the network usable for analytical purposes. Clearly, losing edges results in a progressive fragmentation of the network. Nonetheless, the resulting fragments are those highly related sets of PS that were the initial goal of the study and that have been described throughout the text.

Far from considering the clinical and the biological similarities as separate entities, this study has also attempted at integrating them. Actually, the basic question we had in mind at the inception of this study, was to test whether the altered biological functions, which are caused by gene mutations (at the level of cells and tissues), might account for the clinical phenotypes that are observed in the clinics (at the level of anatomical systems and the whole organism). As a corollary, we had hypothesized that the clinical and biological similarities should be correlated and, specifically, that clinically similar PS should be biologically similar as well. We did identify PS in the Clinical Similarity Network (CSN) that were similar not only clinically (as expected, given the definition of CSN) but also biologically and, conversely, PS in the Biological Similarity Network (BSN) that were similar not only biologically (again, not unexpectedly) but also clinically. For example, similar clusters of cardiac arrhythmia PS were detectable in both CSN and BSN. However, the correlation of the two types of similarity indicates that the picture is more complex. In many instances, missing correlations are likely due to non-specific annotations of either the disease phenotypes or the gene products (in the related ontology). In particular, when performing a binary PS-PS comparison, even a low specificity annotation of just one of the two PS (in one ontology) results in a poorly informative common ancestor being retrieved, which lowers the overall similarity coefficient of the two PS (with regard to that ontology) and ultimately affects the correlation between the two types of similarity. Nevertheless, we propose that several cases reflect not simply defective annotations but rather conditions of potential pathogenic interest. For ease of analysis, we have identified two major categories of lack of biological-clinical correlation.

The former category consists of PS pairs with high biological, but low clinical, similarity. The category is exemplified by those PS, whose normal (i.e., non-mutated) gene products all participate in the channel-mediated transport of cations during the repolarization of excitable cells. Yet, in spite of the high biological similarity, the PS comprise clinical conditions as diverse as cardiac arrhythmia- and seizure-related disorders. A likely explanations for the different clinical manifestations of a similar biological dysfunction is the tissue-specific expression of many DGP. For instance, it is known that several ion channel subunits are specifically expressed in excitable cells of the heart and the brain (Seitter and Koschak, 2017). The latter category consists of PS pairs with high clinical, but low biological, similarity. We propose that this condition highlights how different biological mechanisms converge in a composite response, thus ultimately resulting in clinically similar phenotypes. An interesting example is a cluster of three PS within the CSN (Robinow, van Maldergem and Carpenter syndromes), which are all characterized by severe skeletal dysplasia (with limb shortening and craniofacial anomalies). Although the corresponding DGP annotate apparently unrelated Biological Process (thereby accounting for their low biological similarity), most DGP that are mutated in these syndromes (i.e., WNT5A, DVL1, DVL3 and ROR2 in the Robinow syndrome, FAT4 and DCHS1 in the van Maldergem syndrome and RAB23 in the Carpenter syndrome) participate in the non-canonical Wnt signaling pathway, which underlies the developmental process of planar cell polarity (Butler and Wallingford, 2017).

It is also important to take into account, as a general and final consideration of our study, that most of the analysis suffers from two principal limitations. First, our current biological knowledge is incomplete, thereby resulting in missing information (and, consequently, missing annotation) of a fraction of genes. Second, even for diseases or genes that are known, annotation can be missing. Actually, annotation (of both HPO and GO) is still under progress and thus incomplete. Moreover, also the OMIM database, from which the diseases have been retrieved, has many limitations. For these reasons, actually we considered our primary source of information largely incomplete but it can be improved in the future using our system and also with similar methods.

In conclusion, our findings indicate that, even with the intrinsic limitations of the available databases and of our current biological understanding, it is already possible to rely on semi-automated procedures with the final aim of identifying altered biological responses as likely mechanisms for many inherited diseases with a known molecular basis.

Final conclusions

In the work presented in this thesis, I have focused on complex diseases with a known molecular-genetic basis. Several approaches to the study of these genetic disorders can be used and, during my PhD, I chose two strategies that are very different (albeit not entirely unrelated to each others). The aim is to explore different techniques, that can be complementary in several aspects, helping to face a complex problem. Each of the analysis proposed has particular fields of application, as also advantages and disadvantages that are present and are important to consider.

One of the used methods is purely computational and is based on the information provided by large clinical and biological databases. The big data science makes use of huge quantity of information, that is also as much complete as possible. This method starts from diseases in their general aspects and tries to find the properties that emerge from the system. The shared properties can help to understand the molecular basis of diseases. For this reason, the approach is top-down, moving from general to particular phenomena.

Nevertheless, this type of analysis can be applied only when a big quantity of data is available. For individual diseases, such as Alzheimer, it is important to study particular events, in detail. An event that seems to play a relevant role in Alzheimer disease is the amyloid-beta aggregation. An approach of such type starts from a precise event and tries to rise at the general properties for understanding of entire disease. In this case, the approach is bottom-up, as it moves from particular to general phenomena.

The projects for a future expansion of these methods are the following:

1) Development of computational approaches also for those diseases, such as Alzheimer, for which not so much is known and for which a therapy is not available.

2) Development of a database, freely available, able to collect all the information generated by our computational study, presenting clearly the data and that can results useful for researchers.

3) Use classic methods (experimental) in order to verify the hypothesis that emerged from the computational analysis, that automatically has highlighted an hight number of relations among different diseases.

References

Chapter one

M. **Beeg**, M. Stravalaci, A. Bastone, M. Salmona, M. Gobbi, A modified protocol to prepare seed-free starting solutions of amyloid-β (Aβ)1-40 and Aβ 1-42 from the corresponding depsipeptides. *Anal Biochem* 411 (2011) 297-299

Colombo L, Gamba A, Cantù L, Salmona M, Tagliavini F, Rondelli V, Del Favero E, Brocca P, Pathogenic Aβ A2V versus protective Aβ A2T mutation: Early stage aggregation and membrane interaction. *Biophys Chem.* (2017) Oct; 229:11-18.

Di Fede, G., Catania, M., Morbin, M., Rossi, G., Suardi, S., Mazzoleni, G., Merlin, M., Giovagnoli, A. R., Prioni, S., Erbetta, A., Falcone, C., Gobbi, M., Colombo, L., Bastone, A., Beeg, M., Manzoni, C., Francescucci, B., Spagnoli, A., Cantu, L., Del Favero, E., Levy, E., Salmona, M., and Tagliavini, F. A recessive mutation in the APP gene with dominant-negative effect on amyloidogenesis. *Science* 323, (2009) 1473-1477.

L. **Diomede**, P. Rognoni, F. Lavatelli, M. Romeo, E. Del Favero, L. Cantù, E. Ghibaudi, A. di Fonzo, A. Corbelli, F. Fiordaliso, G. Palladini, V. Valentini, V. Perfetti, M. Salmona, G. Merlini, A Caenorhabditis elegans-based assay recognizes immunoglobulin light chains causing heart amyloidosis, *Blood* 123 (23) (2014) 3543–3552.

J. **Hardy**, D. J. Selkoe, The amyloid hypothesis of Alzheimer’s disease: progress and problems on the road to therapeutics. *Science* 297 (2002) 353–356.

T. **Jonsson**, J.K .Atwal, S. Steinberg, J. Snaedal, P.V. Jonsson, S. Bjornsson, H. Stefansson, P. Sulem, D. Gudbjartsson, J. Maloney, K. Hoyte, A. Gustafson, Y. Liu, Y. Lu, T. Bhangale, R.R. Graham, J. Huttenlocher, G. Bjornsdottir, O.A. Andreassen, E.G. Jönsson, A. Palotie, T.W. Behrens, O.T. Magnusson, A. Kong, U. Thorsteinsdottir, R.J. Watts, K. Stefansson, A mutation in APP protects against Alzheimer's disease and age-related cognitive decline *Nature* 488(7409) (2012) 96-99

P. **Lago**, L. Rovati, L. Cantù, M. Corti, A quasielastic light scattering detector for chromatographic analysis, *Review of Scientific Instruments* 64 (1993) 1797-1802; doi: 10.1063/1.1144013

C.L. **Lawson**, R.J. Hanson (1995) *Solving Least Squares Problems*, Vol 15, SIAM, Philadelphia, PA

J.A. **Maloney**, T. Bainbridge, A. Gustafson, S. Zhang, R. Kyauk, P. Steiner, M. van der Brug, Y. Liu, J. A. Ernst, R. J. Watts, J. K. Atwal, Molecular Mechanisms of Alzheimer Disease Protection by the A673T Allele of Amyloid Precursor Protein *JBC* 289(45) (2014) 30990–31000

M. **Messa**, L. Colombo, E. Del Favero, L. Cantù, T. Stoilova, A. Cagnotto, A. Rossi, M. Morbin, G. Di Fede, F. Tagliavini, M. Salmona, The peculiar role of the A2V mutation in Aβ1–42 molecular assembly, *JBC* 289 (2014) 24143–24152.

Selkoe, D. J., Alzheimer's disease: genes, proteins, and therapy. *Physiol Rev* 81 (2001), 741-766.

M. **Stravalaci**, A. Bastone, M. Beeg, A. Cagnotto, L. Colombo, G. Di Fede, F. Tagliavini, L. Cantù, E. Del Favero, M. Mazzanti, R. Chiesa, M. Salmona, L. Diomede, M. Gobbi, Specific recognition of biologically active amyloid-β oligomers by a new surface plasmon resonance-based immunoassay and an in vivo assay in Caenorhabditis elegans, J. Biol. Chem. 287 (2012) 27796–27805.

A. **Taniguchi**, Y. Sohma, Y. Hirayama, H. Mukai, T. Kimura, Y. Hayashi, K. Matsuzaki, Y. Kiso, "Click peptide": pH-triggered in situ production and aggregation of monomer Abeta1-42. *Chembiochem* 10(2009)710-715

Chapter two

Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. *43*, D789-798.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. Bioinforma. Oxf. Engl. *24*, 282–284.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nat. Rev. Genet. *12*, 56–68.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. Proc. Natl. Acad. Sci. U. S. A. *101*, 3747–3752.

Bazzoni, G., Marengoni, A., Tettamanti, M., Franchi, C., Pasina, L., Djade, C.D., Fortino, I., Bortolotti, A., Merlino, L., and Nobili, A. (2015). The drug prescription network: a system-level view of drug co-prescription in community-dwelling elderly people. Rejuvenation Res. *18*, 153–161.

Butler, M.T., and Wallingford, J.B. (2017). Planar cell polarity in development and disease. Nat. Rev. Mol. Cell Biol. *18*, 375–388.

Cerrone, M., Napolitano, C., and Priori, S.G. (2012). Genetics of ion-channel disorders. Curr. Opin. Cardiol. *27*, 242–252.

Cochat, P., and Rumsby, G. (2013). Primary hyperoxaluria. N. Engl. J. Med. *369*, 649–658.

Cover, T.M., and Thomas, J.A. (1991). Elements of information theory (John Wiley and Sons, Inc.).

Devuyst, O., and Thakker, R.V. (2010). Dent’s disease. Orphanet J. Rare Dis. *5*, 28.

Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. Nucleic Acids Res. *43*, D1049-1056.

Hidalgo, C.A., Blumm, N., Barabási, A.-L., and Christakis, N.A. (2009). A dynamic network approach for the study of human phenotypes. PLoS Comput. Biol. *5*, e1000353.

Hoehndorf, R., Schofield, P.N., and Gkoutos, G.V. (2015). Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases. Sci. Rep. *5*, 10888.

Hopkins, A.L. (2008). Network pharmacology: the next paradigm in drug discovery. Nat. Chem. Biol. *4*, 682–690.

Jiang, J.J., and Conrath, D.W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of International Conference on Research in Computational Linguistics, pp. 19–33.

Kibbe, W.A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C.J., Binder, J.X., Malone, J., Vasant, D., et al. (2015). Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. *43*, D1071-1078.

Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P.N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am. J. Hum. Genet. *85*, 457–464.

Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., et al. (2017). The Human

Phenotype Ontology in 2017. Nucleic Acids Res. *45*, D865–D876.

Lin, D. (1998). An information-theoretic definition of similarity. In ICML ‘98: Proceedings of the Fifteenth International Conference on Machine Learning. (San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.), pp. 296–304.

Newman, M.E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *64*, 016132.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence. 448–453.

Sayer, J.A. (2017). Progress in Understanding the Genetics of Calcium-Containing Nephrolithiasis. J. Am. Soc. Nephrol. JASN *28*, 748–759.

Seitter, H., and Koschak, A. (2017). Relevance of tissue specific subunit expression in channelopathies. Neuropharmacology.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., et al. (2007). The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 25, 1251–1255.

Vidal, M., Cusick, M.E., and Barabási, A.-L. (2011). Interactome networks and human disease. Cell *144*, 986–998.

Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. Bioinforma. Oxf. Engl. *23*, 1274–1281.

Yeung, N., Cline, M.S., Kuchinsky, A., Smoot, M.E., and Bader, G.D. (2008). Exploring biological networks with Cytoscape software. Curr. Protoc. Bioinforma. *Chapter 8*, Unit 8.13.

Large figures

Figure L1:

Variation in number of nodes and edges at different thresholds.

Figure L2:

The Clinical Similarity Network (CSN).

Figure L3:

The Biological Similarity Network – Biological Process (BSN-BP).

Figure L4:

The Biological Similarity Network – Cellular Component (BSN-CC).

Figure L5:

The Biological Similarity Network – Molecular Function (BSN-MF).

Figure L6:

The Biological Similarity Network (BSN). The edge thickness is proportional to the clinical similarity.

Figure L7:

Cluster analysis of the CSN represented in figure L2.

Figure L8:

Cluster analysis of the BSN represented in figure L6.

Figure L9:

Scatter plot of all PS-PS similarities (**A**) and a particular of cardiac, convulsive and renal diseases (**B**).

Figure L1

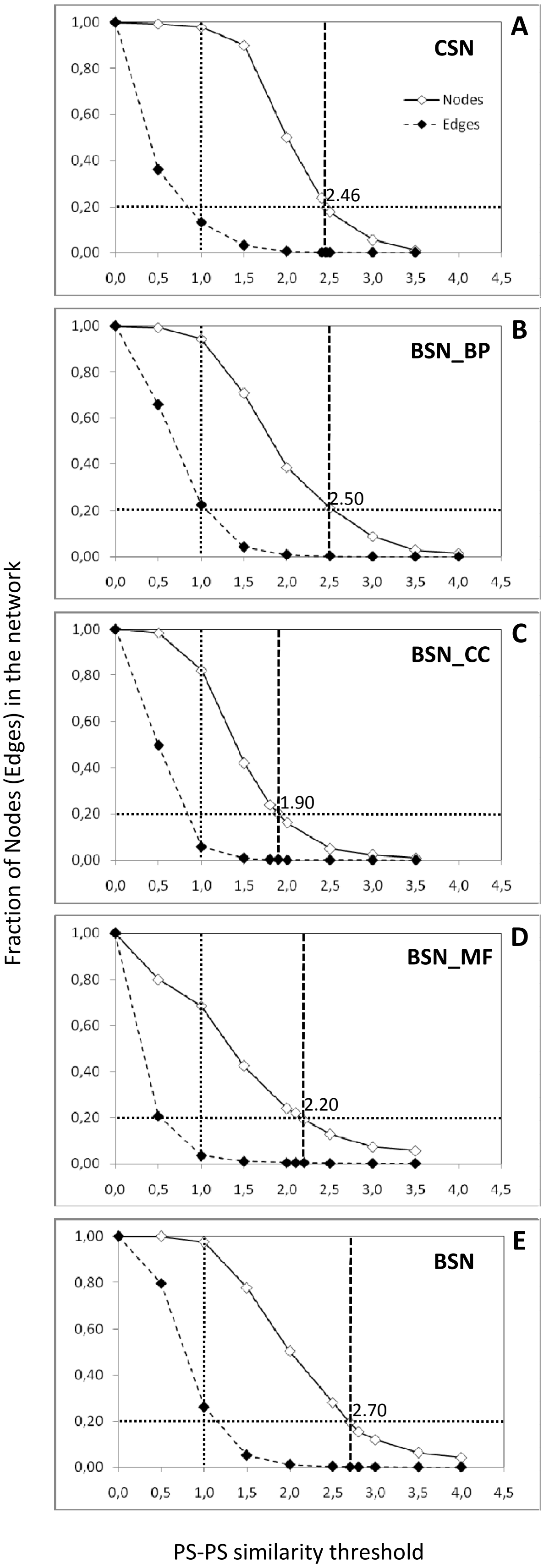

A
CSN
Nodes
Edges
2.46
B
BSN_BP
2.50
C
BSN_CC
1.90
D
BSN_MF
2.20
E
BSN
2.70
1,00
0,80
0,60
0,40
0,20
0,00
0,0
0,5
1,0
1,5
2,0
2,5
3,0
3,5
4,0
4,5
Fraction of Nodes (Edges) in the network
PS-PS similarity threshold

Figure L2

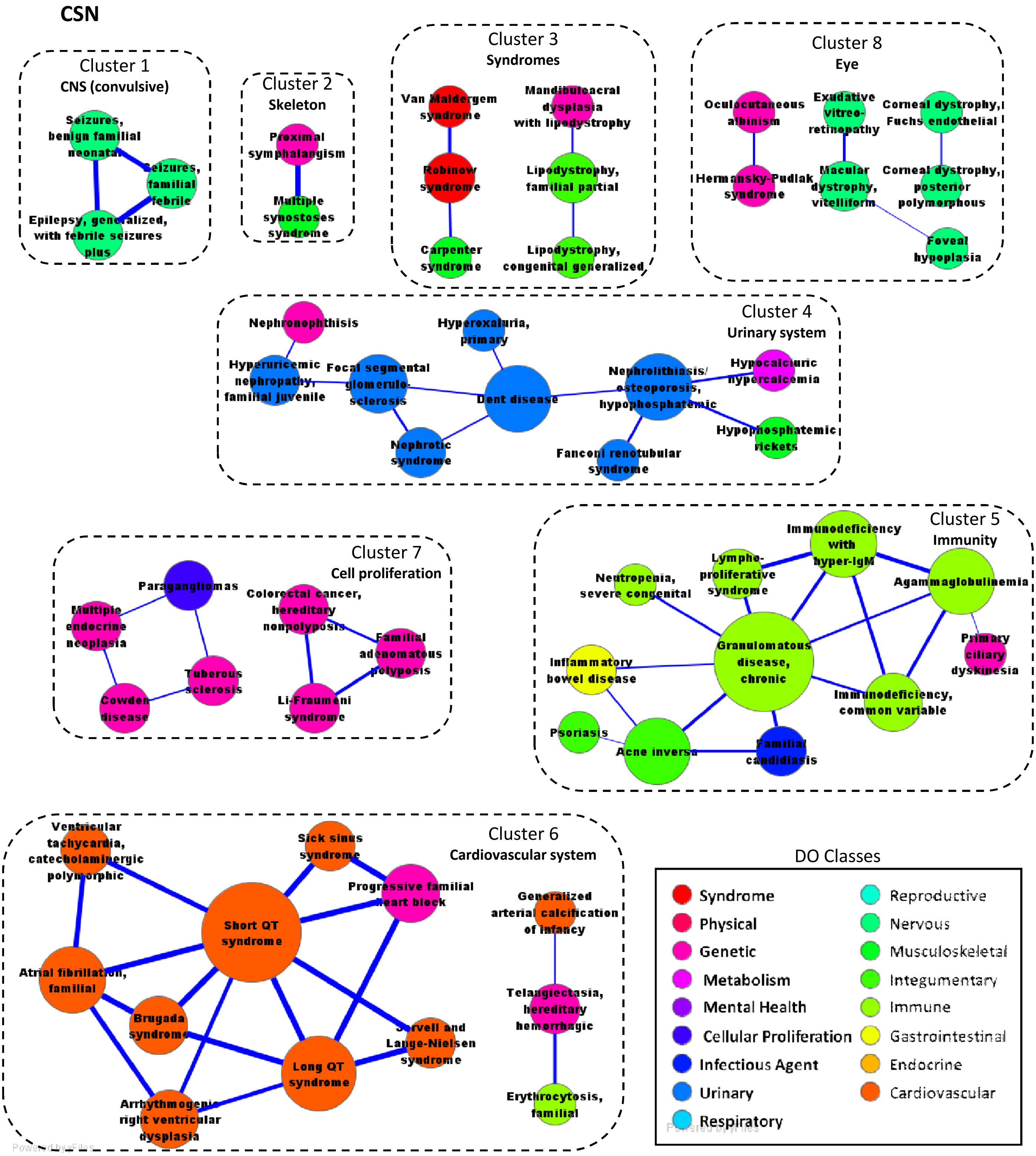

CSN
Cluster 1
CNS (convulsive)
Seizures, benign familial neonatal
Seizures, familial febrile
Epilepsy, generalized, with febrile seizures plus
Cluster 2
Skeleton
Proximal symphalangism
Multiple synostoses syndrome
Cluster 3
Syndromes
Van Maldergem syndrome
Robinow syndrome
Carpenter syndrome
Mandibuloacral dysplasia with lipodystrophy
Lipodystrophy, familial partial
Lipodystrophy, congenital generalized
Cluster 8
Eye
Oculocutaneous albinism
Hermansky-Pudlak syndrome
Exudative vitreo-retinopathy
Macular dystrophy, vitelliform
Corneal dystrophy, Fuchs endothelial
Corneal dystrophy, posterior polymorphous
Foveal hypoplasia
Cluster 4
Urinary system
Nephronophthisis
Hyperuricemic nephropathy, familial juvenile
Focal segmental glomerulo-sclerosis
Hyperoxaluria, primary
Dent disease
Nephrotic syndrome
Nephrolithiasis/ osteoporosis, hypophosphatemic
Hypocalciuric hypercalcemia
Hypophosphatemic rickets
Fanconi renotubular syndrome
Cluster 7
Cell proliferation
Paragangliomas
Multiple endocrine neoplasia
Tuberous sclerosis
Cowden disease
Colorectal cancer, hereditary nonpolyposis
Familial adenomatous polyposis
Li-Fraumeni syndrome
Cluster 5
Immunity
Immunodeficiency with hyper-IgM
Neutropenia, severe congenital
Lympho-proliferative syndrome
Agammaglobulinemia
Primary ciliary dyskinesia
Granulomatous disease, chronic
Inflammatory bowel disease
Immunodeficiency, common variable
Psoriasis
Acne inversa
Familial candidiasis
Cluster 6
Cardiovascular system
Ventricular tachycardia, catecholaminergic polymorphic
Sick sinus syndrome
Progressive familial heart block
Short QT syndrome
Atrial fibrillation, familial
Brugada syndrome
Long QT syndrome
Jervell and Lange-Nielsen syndrome
Arrhythmogenic right ventricular dysplasia
Generalized arterial calcification of infancy
Telangiectasia, hereditary hemorrhagic
Erythrocytosis, familial
DO Classes
Syndrome
Physical
Genetic
Metabolism
Mental Health
Cellular Proliferation
Infectious Agent
Urinary
Respiratory
Reproductive
Nervous
Musculoskeletal
Integumentary
Immune
Gastrointestinal
Endocrine
Cardiovascular

Figure L3

BSN-BP

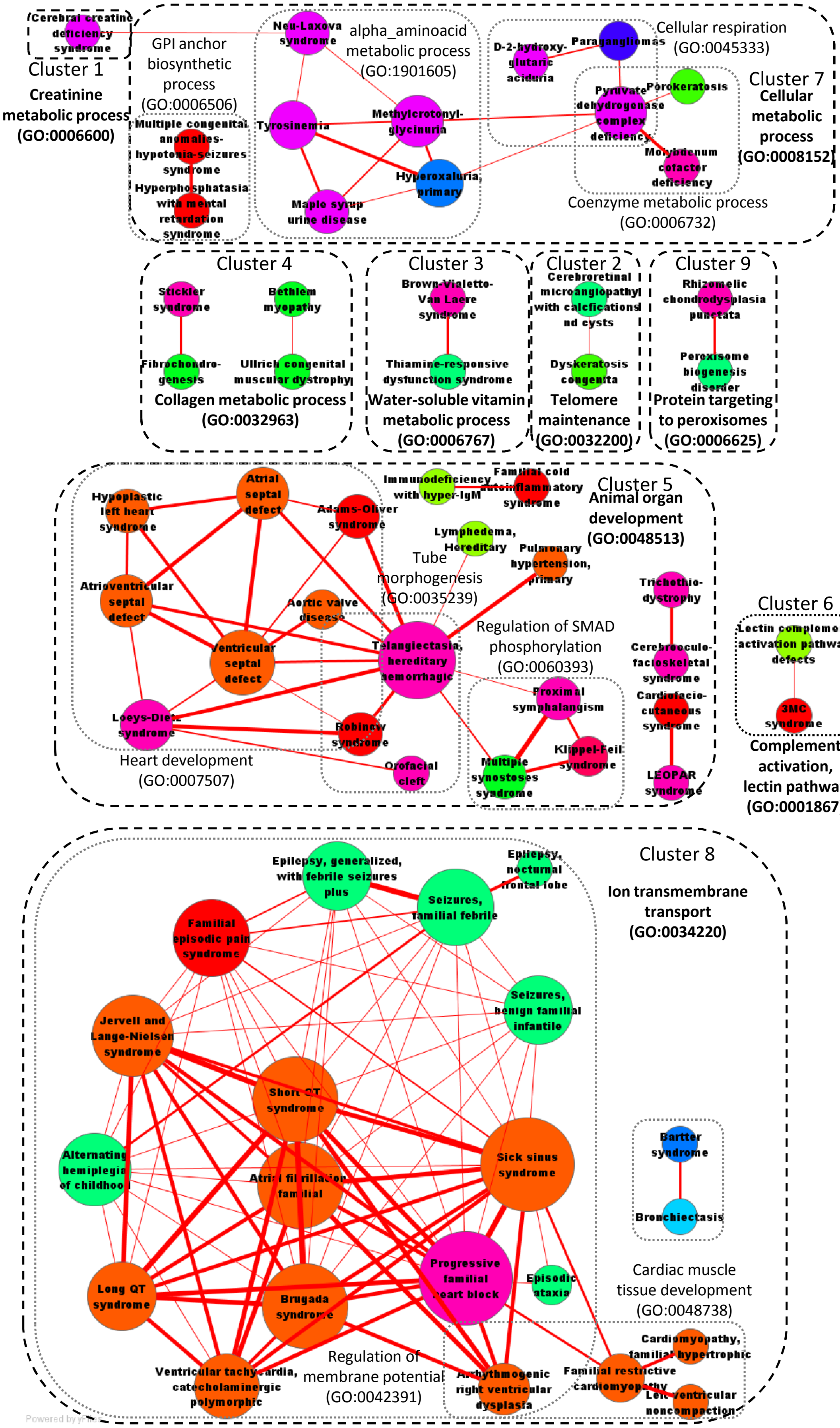

Cerebral creatine deficiency syndrome
Cluster 1
Creatinine metabolic process (GO:0006600)
GPI anchor biosynthetic process (GO:0006506)
Multiple congenital anomalies-hypotonia-seizures syndrome
Hyperphosphatasia with mental retardation syndrome
Neu-Laxova syndrome
alpha_aminoacid metabolic process (GO:1901605)
Tyrosinemia
Methylcrotonyl-glycinuria
Hyperoxaluria, primary
Maple syrup urine disease
D-2-hydroxy-glutaric aciduria
Paragangliomas
Cellular respiration (GO:0045333)
Pyruvate dehydrogenase complex deficiency
Porokeratosis
Molybdenum cofactor deficiency
Coenzyme metabolic process (GO:0006732)
Cluster 7
Cellular metabolic process (GO:0008152)
Cluster 4
Stickler syndrome
Bethlem myopathy
Fibrochondro-genesis
Ullrich congenital muscular dystrophy
Collagen metabolic process (GO:0032963)
Cluster 3
Brown-Vialetto-Van Laere syndrome
Thiamine-responsive dysfunction syndrome
Water-soluble vitamin metabolic process (GO:0006767)
Cluster 2
Cerebroretinal microangiopathy with calcifications nd cysts
Dyskeratosis congenita
Telomere maintenance (GO:0032200)
Cluster 9
Rhizomelic chondrodysplasia punctata
Peroxisome biogenesis disorder
Protein targeting to peroxisomes (GO:0006625)
Hypoplastic left heart syndrome
Atrial septal defect
Adams-Oliver syndrome
Immunodeficiency with hyper-IgM
Familial cold autoinflammatory syndrome
Cluster 5
Animal organ development (GO:0048513)
Lymphedema, Hereditary
Pulmonary hypertension, primary
Tube morphogenesis (GO:0035239)
Atrioventricular septal defect
Aortic valve disease
Ventricular septal defect
Telangiectasia, hereditary hemorrhagic
Regulation of SMAD phosphorylation (GO:0060393)
Trichothio-dystrophy
Cerebrooculo-facioskeletal syndrome
Cardiofacio-cutaneous syndrome
LEOPAR syndrome
Proximal symphalangism
Klippel-Feil syndrome
Multiple synostoses syndrome
Loeys-Dietz syndrome
Robinow syndrome
Orofacial cleft
Heart development (GO:0007507)
Cluster 6
Lectin complement activation pathway defects
3MC syndrome
Complement activation, lectin pathway (GO:0001867)
Cluster 8
Ion transmembrane transport (GO:0034220)
Epilepsy, generalized, with febrile seizures plus
Epilepsy, nocturnal frontal lobe
Seizures, familial febrile
Familial episodic pain syndrome
Seizures, benign familial infantile
Jervell and Lange-Nielsen syndrome
Short QT syndrome
Alternating hemiplegia of childhood
Atrial fibrillation, familial
Sick sinus syndrome
Bartter syndrome
Bronchiectasis
Progressive familial heart block
Episodic ataxia
Long QT syndrome
Brugada syndrome
Cardiac muscle tissue development (GO:0048738)
Regulation of membrane potential (GO:0042391)
Ventricular tachycardia, catecholaminergic polymorphic
Arrhythmogenic right ventricular dysplasia
Familial restrictive cardiomyopathy
Cardiomyopathy, familial hypertrophic
Left ventricular noncompaction
Powered by yFiles

Figure L4

BSN-CC

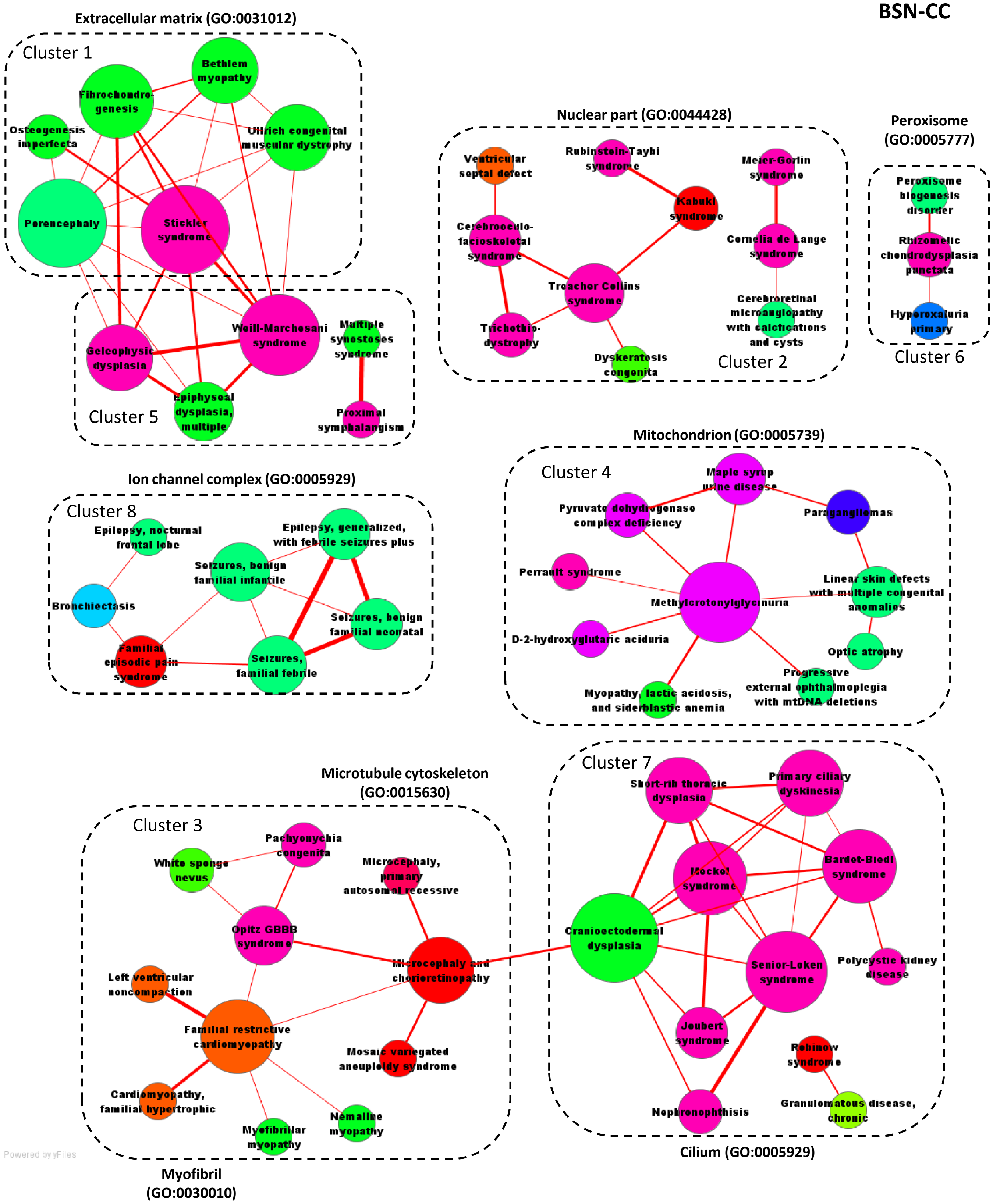

Extracellular matrix (GO:0031012)
Cluster 1
Bethlem myopathy
Fibrochondrogenesis
Osteogenesis imperfecta
Ullrich congenital muscular dystrophy
Porencephaly
Stickler syndrome
Cluster 5
Weill-Marchesani syndrome
Multiple synostoses syndrome
Geleophysic dysplasia
Epiphyseal dysplasia, multiple
Proximal symphalangism
Nuclear part (GO:0044428)
Ventricular septal defect
Rubinstein-Taybi syndrome
Meier-Gorlin syndrome
Kabuki syndrome
Cerebrooculo-facioskeletal syndrome
Cornelia de Lange syndrome
Treacher Collins syndrome
Trichothio-dystrophy
Cerebroretinal microangiopathy with calcifications and cysts
Dyskeratosis congenita
Cluster 2
Peroxisome (GO:0005777)
Peroxisome biogenesis disorder
Rhizomelic chondrodysplasia punctata
Hyperoxaluria primary
Cluster 6
Mitochondrion (GO:0005739)
Cluster 4
Maple syrup urine disease
Pyruvate dehydrogenase complex deficiency
Paragangliomas
Perrault syndrome
Linear skin defects with multiple congenital anomalies
Methylcrotonylglycinuria
D-2-hydroxyglutaric aciduria
Optic atrophy
Progressive external ophthalmoplegia with mtDNA deletions
Myopathy, lactic acidosis, and siderblastic anemia
Ion channel complex (GO:0005929)
Cluster 8
Epilepsy, nocturnal frontal lobe
Epilepsy, generalized, with febrile seizures plus
Seizures, benign familial infantile
Bronchiectasis
Seizures, benign familial neonatal
Familial episodic pain syndrome
Seizures, familial febrile
Microtubule cytoskeleton (GO:0015630)
Cluster 3
Pachyonychia congenita
White sponge nevus
Microcephaly, primary autosomal recessive
Opitz GBBB syndrome
Left ventricular noncompaction
Microcephaly and chorioretinopathy
Familial restrictive cardiomyopathy
Mosaic variegated aneuploidy syndrome
Cardiomyopathy, familial hypertrophic
Nemaline myopathy
Myofibrillar myopathy
Powered by yFiles
Myofibril (GO:0030010)
Cluster 7
Short-rib thoracic dysplasia
Primary ciliary dyskinesia
Meckel syndrome
Bardet-Biedl syndrome
Cranioectodermal dysplasia
Senior-Loken syndrome
Polycystic kidney disease
Joubert syndrome
Robinow syndrome
Nephronophthisis
Granulomatous disease, chronic
Cilium (GO:0005929)

Figure L5

BSN-MF

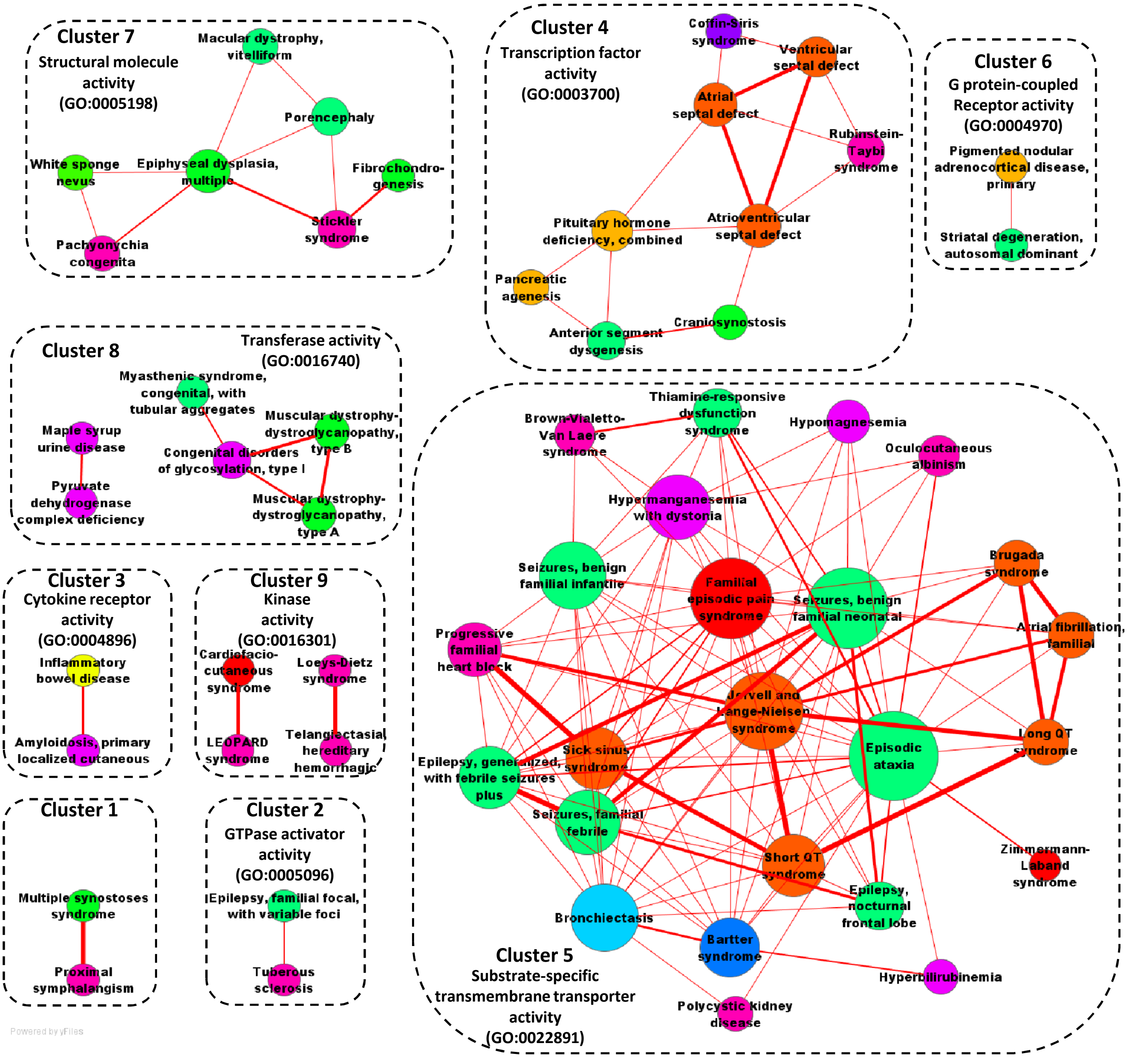

Cluster 7
Structural molecule activity (GO:0005198)
Macular dystrophy, vitelliform
Porencephaly
White sponge nevus
Epiphyseal dysplasia, multiple
Fibrochondrogenesis
Stickler syndrome
Pachyonychia congenita
Cluster 4
Transcription factor activity (GO:0003700)
Coffin-Siris syndrome
Ventricular septal defect
Atrial septal defect
Rubinstein-Taybi syndrome
Pituitary hormone deficiency, combined
Atrioventricular septal defect
Pancreatic agenesis
Craniosynostosis
Anterior segment dysgenesis
Cluster 6
G protein-coupled Receptor activity (GO:0004970)
Pigmented nodular adrenocortical disease, primary
Striatal degeneration, autosomal dominant
Cluster 8
Transferase activity (GO:0016740)
Myasthenic syndrome, congenital, with tubular aggregates
Maple syrup urine disease
Muscular dystrophy-dystroglycanopathy, type B
Congenital disorders of glycosylation, type I
Pyruvate dehydrogenase complex deficiency
Muscular dystrophy-dystroglycanopathy, type A
Cluster 3
Cytokine receptor activity (GO:0004896)
Inflammatory bowel disease
Amyloidosis, primary localized cutaneous
Cluster 9
Kinase activity (GO:0016301)
Cardiofacio-cutaneous syndrome
Loeys-Dietz syndrome
LEOPARD syndrome
Telangiectasia, hereditary hemorrhagic
Cluster 1
Multiple synostoses syndrome
Proximal symphalangism
Cluster 2
GTPase activator activity (GO:0005096)
Epilepsy, familial focal, with variable foci
Tuberous sclerosis
Powered by yFiles
Thiamine-responsive dysfunction syndrome
Brown-Vialetto-Van Laere syndrome
Hypomagnesemia
Oculocutaneous albinism
Hypermanganesemia with dystonia
Seizures, benign familial infantile
Familial episodic pain syndrome
Seizures, benign familial neonatal
Brugada syndrome
Progressive familial heart block
Atrial fibrillation, familial
Jervell and Lange-Nielsen syndrome
Long QT syndrome
Sick sinus syndrome
Episodic ataxia
Epilepsy, generalized, with febrile seizures plus
Seizures, familial febrile
Zimmermann-Laband syndrome
Short QT syndrome
Epilepsy, nocturnal frontal lobe
Bronchiectasis
Bartter syndrome
Hyperbilirubinemia
Polycystic kidney disease
Cluster 5
Substrate-specific transmembrane transporter activity (GO:0022891)

Figure L6

BSN

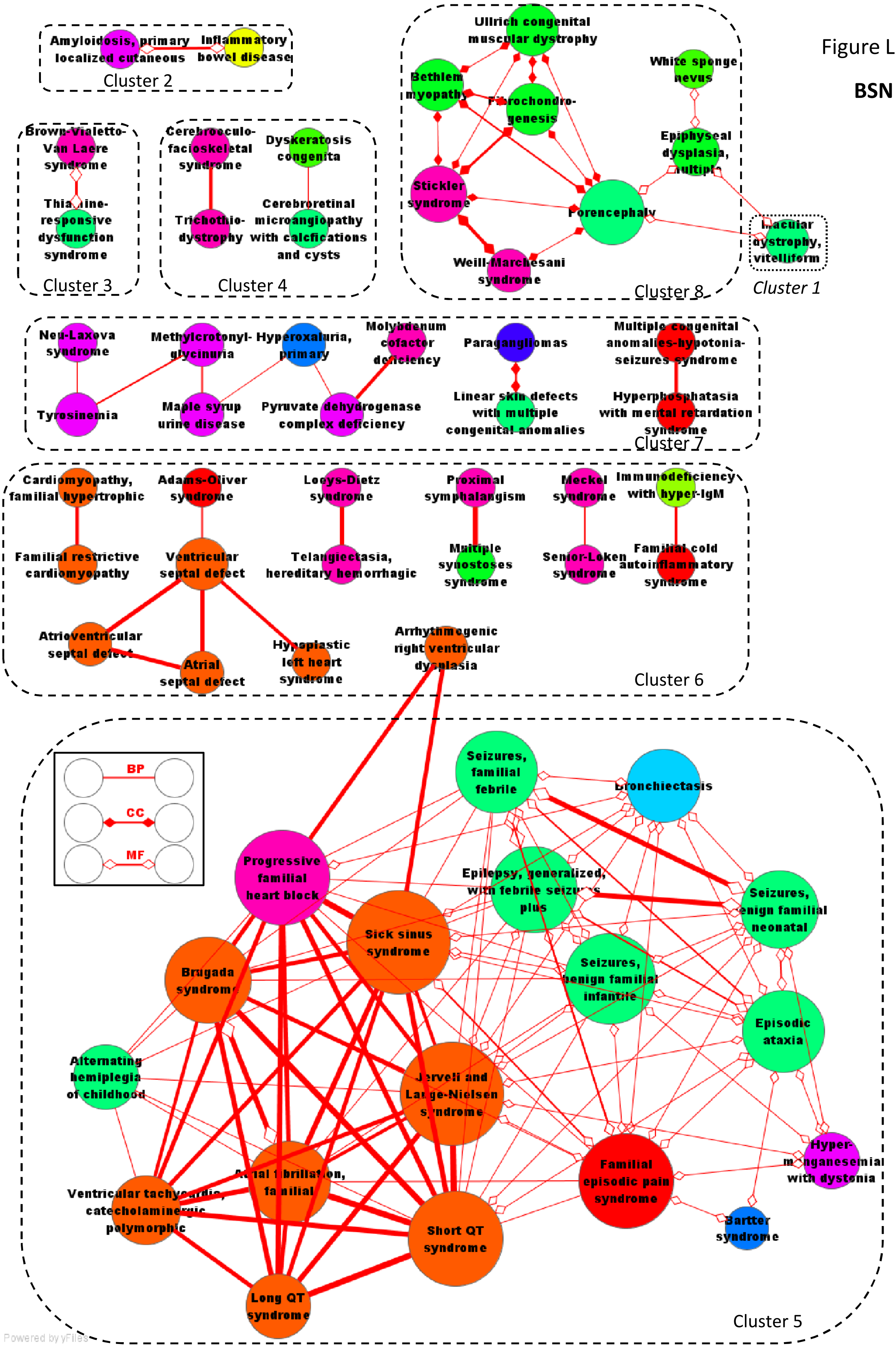
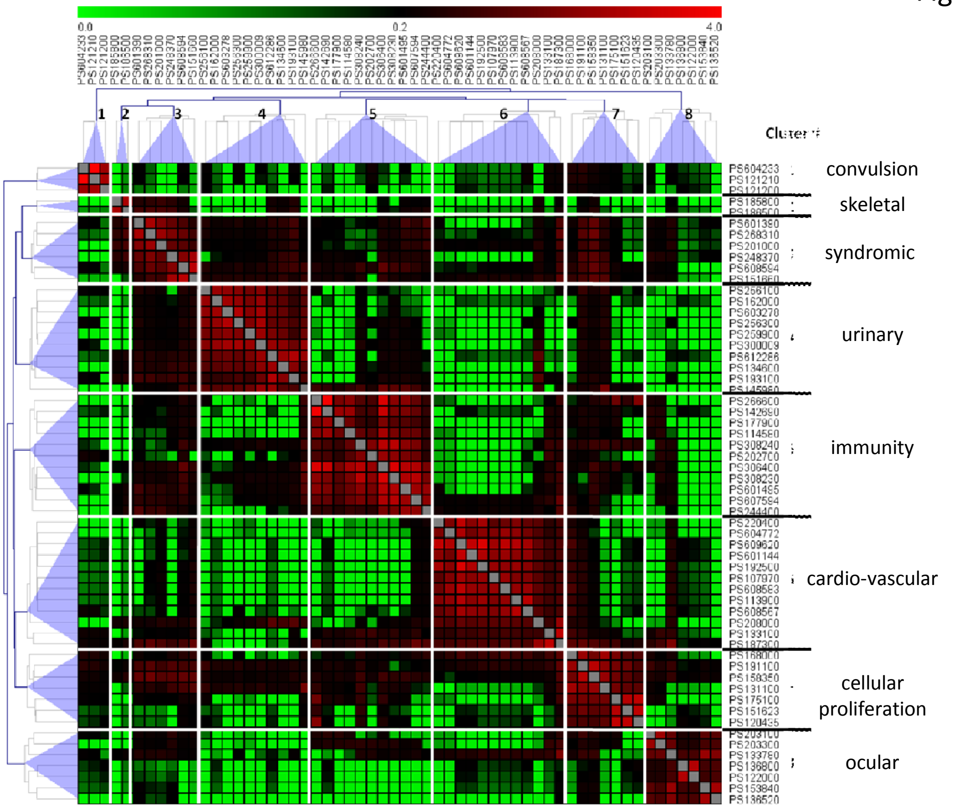

Amyloidosis, primary localized cutaneous
Inflammatory bowel disease
Cluster 2
Brown-Vialetto-Van Laere syndrome
Thiamine-responsive dysfunction syndrome
Cluster 3
Cerebrooculo-facioskeletal syndrome
Trichothio-dystrophy
Dyskeratosis congenita
Cerebroretinal microangiopathy with calcifications and cysts
Cluster 4
Ullrich congenital muscular dystrophy
Bethlem myopathy
Fibrochondro-genesis
Stickler syndrome
Porencephaly
Weill-Marchesani syndrome
White sponge nevus
Epiphyseal dysplasia, multiple
Cluster 8
Macular dystrophy, vitelliform
Cluster 1
Neu-Laxova syndrome
Tyrosinemia
Methylcrotonyl-glycinuria
Hyperoxaluria, primary
Maple syrup urine disease
Molybdenum cofactor deficiency
Pyruvate dehydrogenase complex deficiency
Paragangliomas
Linear skin defects with multiple congenital anomalies
Multiple congenital anomalies-hypotonia-seizures syndrome
Hyperphosphatasia with mental retardation syndrome
Cluster 7
Cardiomyopathy, familial hypertrophic
Familial restrictive cardiomyopathy
Adams-Oliver syndrome
Ventricular septal defect
Atrioventricular septal defect
Atrial septal defect
Hypoplastic left heart syndrome
Loeys-Dietz syndrome
Telangiectasia, hereditary hemorrhagic
Proximal symphalangism
Multiple synostoses syndrome
Meckel syndrome
Senior-Loken syndrome
Immunodeficiency with hyper-IgM
Familial cold autoinflammatory syndrome
Arrhythmogenic right ventricular dysplasia
Cluster 6
BP
CC
MF
Seizures, familial febrile
Bronchiectasis
Progressive familial heart block
Epilepsy, generalized, with febrile seizures plus
Seizures, benign familial neonatal
Sick sinus syndrome
Brugada syndrome
Seizures, benign familial infantile
Episodic ataxia
Alternating hemiplegia of childhood
Jervell and Lange-Nielsen syndrome
Atrial fibrillation, familial
Ventricular tachycardia, catecholaminergic polymorphic
Familial episodic pain syndrome
Hyper-manganesemia with dystonia
Short QT syndrome
Bartter syndrome
Long QT syndrome
Cluster 5
Powered by yFiles

Figure L7

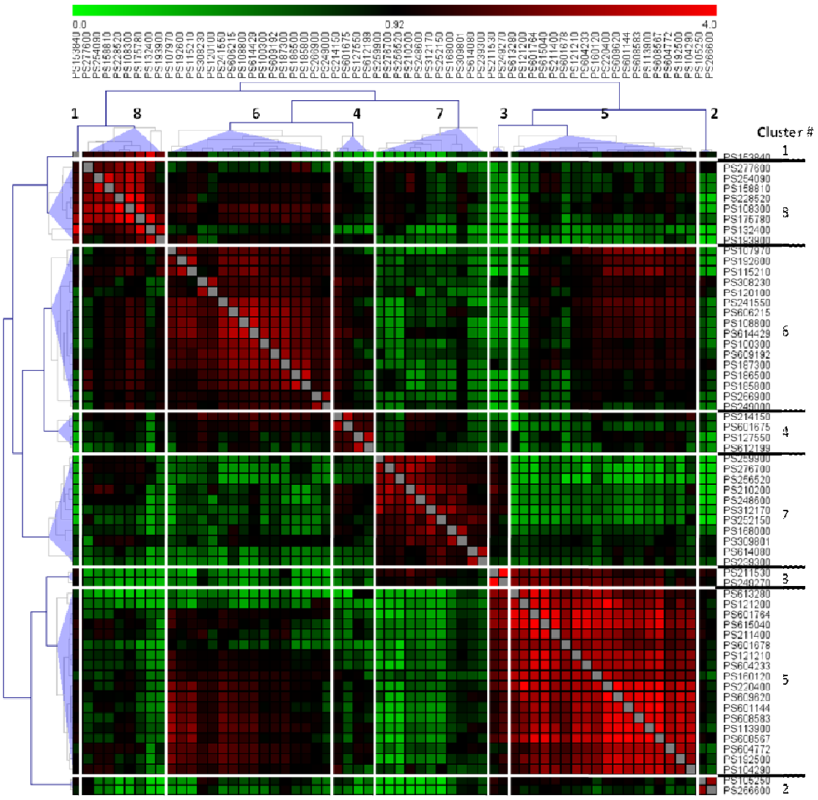
CSN



Cluster	PS id	PS name
1	PS604233	Epilepsy, generalized, with febrile seizures plus
	PS121210	Seizures, familial febrile
	PS121200	Seizures, benign familial neonatal
2	PS185800	Proximal symphalangism
	PS186500	Multiple synostoses syndrome
3	PS601390	Van Maldergem syndrome
	PS268310	Robinow syndrome
	PS201000	Carpenter syndrome
	PS248370	Mandibuloacral dysplasia with lipodystrophy
	PS608594	Lipodystrophy, congenital generalized
	PS151660	Lipodystrophy, familial partial
4	PS256100	Nephronophthisis
	PS162000	Hyperuricemic nephropathy, familial juvenile
	PS603278	Focal segmental glomerulosclerosis
	PS256300	Nephrotic syndrome
	PS259900	Hyperoxaluria, primary
	PS300009	Dent disease
	PS612286	Nephrolithiasis/osteoporosis, hypophosphatemic
	PS134600	Fanconi renotubular syndrome
	PS193100	Hypophosphatemic rickets
	PS145980	Hypocalciuric hypercalcemia
5	PS266600	Inflammatory bowel disease
	PS142690	Acne inversa
	PS177900	Psoriasis
	PS114580	Familial candidiasis
	PS308240	Lymphoproliferative syndrome
	PS202700	Neutropenia, severe congenital
	PS306400	Granulomatous disease, chronic
	PS308230	Immunodeficiency with hyper-IgM
	PS601495	Agammaglobulinemia
	PS607594	Immunodeficiency, common variable
	PS244400	Primary ciliary dyskinesia
6	PS220400	Jervell and Lange-Nielsen syndrome
	PS604772	Ventricular tachycardia, catecholaminergic polymorphic
	PS609620	Short QT syndrome
	PS601144	Brugada syndrome
	PS192500	Long QT syndrome
	PS107970	Arrhythmogenic right ventricular dysplasia
	PS608583	Atrial fibrillation, familial
	PS113900	Progressive familial heart block
	PS608567	Sick sinus syndrome
	PS208000	Generalized arterial calcification of infancy
	PS133100	Erythrocytosis, familial
	PS187300	Telangiectasia, hereditary hemorrhagic
7	PS168000	Paragangliomas
	PS191100	Tuberous sclerosis
	PS158350	Cowden disease
	PS131100	Multiple endocrine neoplasia
	PS175100	Familial adenomatous polyposis
	PS151623	Li-Fraumeni syndrome
	PS120435	Colorectal cancer, hereditary nonpolyposis
8	PS203100	Oculocutaneous albinism
	PS203300	Hermansky-Pudlak syndrome
	PS133780	Exudative vitreoretinopathy
	PS136800	Corneal dystrophy, Fuchs endothelial
	PS122000	Corneal dystrophy, posterior polymorphous
	PS153840	Macular dystrophy, vitelliform
	PS136520	Foveal hypoplasia

Figure L8

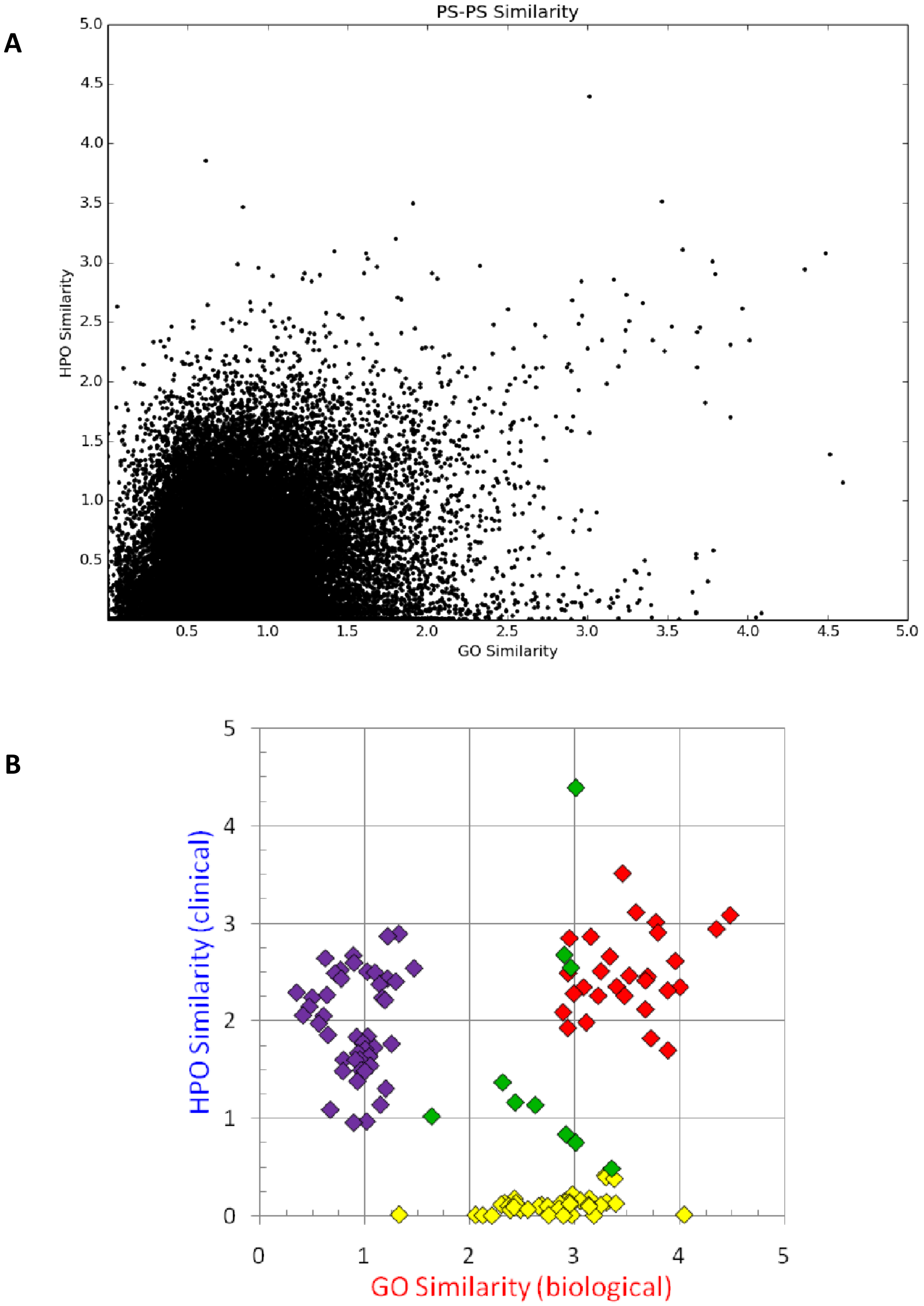
BSN



Cluster	PS id	PS_name
1	PS153840	Macular dystrophy, vitelliform
8	PS277600	Weill-Marchesani syndrome
8	PS254090	Ullrich congenital muscular dystrophy
8	PS158810	Bethlem myopathy
8	PS228520	Fibrochondrogenesis
8	PS108300	Stickler syndrome
8	PS175780	Porencephaly
8	PS132400	Epiphyseal dysplasia, multiple
8	PS193900	White sponge nevus
6	PS107970	Arrhythmogenic right ventricular dysplasia
6	PS192600	Cardiomyopathy, familial hypertrophic
6	PS115210	Familial restrictive cardiomyopathy
6	PS308230	Immunodeficiency with hyper-IgM
6	PS120100	Familial cold autoinflammatory syndrome
6	PS241550	Hypoplastic left heart syndrome
6	PS606215	Atrioventricular septal defect
6	PS108800	Atrial septal defect
6	PS614429	Ventricular septal defect
6	PS100300	Adams-Oliver syndrome
6	PS609192	Loeys-Dietz syndrome
6	PS187300	Telangiectasia, hereditary hemorrhagic
6	PS186500	Multiple synostoses syndrome
6	PS185800	Proximal symphalangism
6	PS266900	Senior-Loken syndrome
6	PS249000	Meckel syndrome
4	PS214150	Cerebrooculofacioskeletal syndrome
4	PS601675	Trichothiodystrophy
4	PS127550	Dyskeratosis congenita
4	PS612199	Cerebroretinal microangiopathy with calcifications and cysts

Cluster	PS id	PS_name
7	PS259900	Hyperoxaluria, primary
7	PS276700	Tyrosinemia
7	PS256520	Neu-Laxova syndrome
7	PS210200	Methylcrotonylglycinuria
7	PS248600	Maple syrup urine disease
7	PS312170	Pyruvate dehydrogenase complex deficiency
7	PS252150	Molybdenum cofactor deficiency
7	PS168000	Paragangliomas
7	PS309801	Linear skin defects with multiple congenital anomalies
7	PS614080	Multiple congenital anomalies-hypotonia-seizures syndrome
7	PS239300	Hyperphosphatasia with mental retardation syndrome
3	PS211530	Brown-Vialetto-Van Laere syndrome
3	PS249270	Thiamine-responsive dysfunction syndrome
5	PS613280	Hypermanganesemia with dystonia
5	PS121200	Seizures, benign familial neonatal
5	PS601764	Seizures, benign familial infantile
5	PS615040	Familial episodic pain syndrome
5	PS211400	Bronchiectasis
5	PS601678	Bartter syndrome
5	PS121210	Seizures, familial febrile
5	PS604233	Epilepsy, generalized, with febrile seizures plus
5	PS160120	Episodic ataxia
5	PS220400	Jervell and Lange-Nielsen syndrome
5	PS609620	Short QT syndrome
5	PS601144	Brugada syndrome
5	PS608583	Atrial fibrillation, familial
5	PS113900	Progressive familial heart block
5	PS608567	Sick sinus syndrome
5	PS604772	Ventricular tachycardia, catecholaminergic polymorphic
5	PS192500	Long QT syndrome
5	PS104290	Alternating hemiplegia of childhood
2	PS105250	Amyloidosis, primary localized cutaneous
2	PS266600	Inflammatory bowel disease

Figure L9



Disorder types		Sim_HPO	Sim_GO_max
Cardiac-cardiac	🟥	2.49 ± 0.43	3.51 ± 0.43
Cardiac-convulsive	🟨	0.12 ± 0.10	2.79 ± 0.48
Convulsive-convulsive	🟩	1.65 ± 1.21	2.72 ± 0.49
Renal-renal	🟪	1.98 ± 0.52	0.92 ± 0.26