UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA

Exploring the interactions between neuron degeneration and RNA homeostasis through biological network analysis

Marina Luque García-Vaquero

Mestrado em Bioquímica
Especialização em Bioquímica

Dissertação orientada por:
Prof. Doutor Francisco Rodrigues Pinto e
Prof. Doutora Margarida Gama-Carvalho

2016

Abstract

Amyotrophic lateral sclerosis (ALS) and spinal muscular atrophy (SMA) are characterized by motor neuron (MN) degeneration and commonly referred as motor neuron diseases (MND). MN degeneration leads to the loss of muscle innervation and subsequent muscular atrophy. In addition to phenotypic similarity, they also share molecular overlaps. Genes that codify FUS, TDP43, SETX and SOD1 proteins are the best-known causative genes of ALS and SMN dysfunction is the cause of SMA. FUS, TDP43, SMN and SETX (FTSS) proteins are known to physically interact and are involved in similar functions, many of which related to RNA metabolism processes.

This supports the hypothesis that ALS and SMA are different pathophenotypic results derived from related molecular origins, in particular from RNA homeostasis perturbation. However, it is very intriguing how such critical events could specifically induce motor neuron perturbation. Besides, RNA metabolism is not the only function described for MND associated genes, indeed FTSS proteins are highly multifunctional which hinders the identification of the most relevant functions in this context.

In order to solve these questions we followed a systems biology approach exploring the interactomic and functional framework of MN degeneration. Under the hypothesis that FTSS proteins are central elements in MN degeneration, we performed a local network analysis to unravel the most influential functions among FTSS proteins. We constructed a protein-protein interaction (PPI) network constituted by FTSS proteins' common interactors to identify the most over represented functions within this FTSS-focused network.

We also performed a PPI network analysis including all the known MND associated genes. For that purpose we developed a new method, S2B (double specific betweenness) to prioritize nodes specifically linking a pair of diseases. While standard betweenness is measured for all possible shortest paths between any nodes, S2B only considers those shortest paths involving Disease Associated Genes (DAGs) from one disease as initial nodes and DAGs from the other disease as final nodes. Therefore, S2B method only prioritizes proteins linking MND causative genes. Moreover, knowing that highly connected nodes (hubs) are more likely found by chance in a shortest paths involving DAGs, S2B method also performs two network randomization-based statistics to filter out proteins that link MND DAGs non specifically.

Finally we functionally enriched the prioritized candidates and compared against the functional set obtained with FTTS-focused network in other to explore the role of RNA metabolism and other putative molecular mechanisms on MN degeneration. The combined approaches used in this work provided novel biological processes simultaneously involved in ALS and SMA diseases and confirmed the relevance of known related processes.

Globally, our results suggest five pathways in common between ALS and SMA: 1) DNA damage and apoptosis induced by R-loop deregulation, 2) inflammation and neurodegeneration induced by immune hyper-sensitivity, 3) chromatin deregulation and genotoxicity produced by histone biogenesis perturbation, 4) splicing patterns alteration and genotoxicity produced by spliceosome assembly failure and 5) deregulation of microtubule related processes leading to morphological problems in axon and synapse formation.

Besides the new hypothesis of common pathomechanisms in MNDs, our work also supplies a new network-based DAG prioritization method, S2B, to identify disease-disease linking candidates we expect to contribute to the study of various complex diseases.

Keywords: Amyotrophic Lateral Sclerosis, Disease-Associated Gene prioritization, network biology, Spinal Muscular Atrophy, systems biology

Resumo

A esclerose lateral amiotrófica (ALS) e a atrofia muscular espinal (SMA) são caracterizadas pela degeneração dos neurónios motores (MN) e são comummente conhecidas como doenças neuromusculares, ou mais especificamente doenças do neurónio motor (MND). A morte dos neurónios motores está diretamente envolvida na perda da inervação muscular e na consequente atrofia muscular. Para além da convergência fenotípica, estas doenças também partilham grandes semelhanças moleculares. A perda de função dos genes que codificam as proteínas FUS, TDP43, SETX e SOD1 são as causas mais conhecidas de ALS. No caso da SMA, a doença é provocada pela produção de formas não funcionais da proteína SMN. Sabe-se que as proteínas FUS, TDP43, SMN e SETX (FTSS) interagem fisicamente e, além disso, são conhecidas por estar envolvidas num conjunto de funções semelhantes, muitas das quais estão relacionados com os processos de metabolismo do RNA.

Esta observação levou à hipótese de que a ALS e a SMA são fenótipos patológicos que, apesar de diferentes, derivam de mecanismos moleculares semelhantes, possivelmente associados à perturbação da hemóstase do RNA. No entanto , é muito intrigante como eventos transversais a todos os tipos celulares podem induzir a morte especifica dos neurónios motores. A fim de resolver estas questões nós propomos uma abordagem de biologia de sistemas para descrever a estrutura interactómica e funcional da degeneração dos neurónios motores.

A biologia de sistemas (*systems biology*) baseia-se no pressuposto de que "o todo é mais do que a soma das partes". Utiliza um abordagem holística para decifrar a complexidade dos sistemas biológicos e para isso integra muitas disciplinas científicas como a biologia, ciências computacionais, estatística e matemática. A biologia de sistemas concebe as entidades biológicas como sistemas complexos de elementos interrelacionados. Deste modo, uma boa maneira de entender as suas propriedades é representando-as como redes (*networks*). A biologia de redes (*networks biology*) é um subcampo da biologia de sistemas que explora os princípios da teoria de redes para inferir informação biológica. Da mesma forma, as doenças são o resultado fenotípico de perturbações interrelacionadas e assim também podem ser representadas como redes biológicas. A medicina de redes é, por sua vez, focalizada na obtenção de conhecimento biomédico a partir da biologia de redes.

O nosso principal objetivo é, em primeiro lugar, identificar os elementos mais centrais numa rede de interação proteína-proteína contendo os genes associados à ALS e à SMA. Estes elementos serão parte de mecanismos patológicos hipoteticamente envolvidos na degeneração dos neurónios motores. Considerando a hipótese de que as proteínas FTSS são elementos centrais nas MNDs, realizamos primeiramente uma análise exploratória para desvendar as funções mais influentes entre as proteínas FTSS. Para isso foi construída uma rede de interações proteína-proteína (PPI) constituída pelos interactores mais próximos às proteínas FTSS, o que nos permite identificar as funções mais sobre-representadas dentro da rede.

Embora, sabendo que as proteínas FTSS não são as únicas proteínas associadas às MNDs, também realizámos uma exploração mais integrativa incluindo todos os DAGs (genes associados à doença) conhecidos para a ALS e SMA e aplicando um método de priorização de DAGs para prever os elementos mais centrais a ligar as duas patologias. Contudo, depois de fazer extensa uma pesquisa bibliográfica, não encontramos nenhum método com um objectivo semelhante, pelo que construímos um método novo com base em teoria de redes para prever os nós que ligam especificamente os DAGs associados a um par de doenças.

O método S2B foi concebido a partir do pressuposto de que as proteínas que interagem com um DAG são provavelmente relacionadas com a mesma doença (constituindo módulos de doenças na rede) e também que os DAGs são propensos a ser associados a mais do que uma doença (os módulos de doenças podem sobrepor-se). Assim, o método S2B está focado na medição dum tipo particular de medida de centralidade (S2B *betweenness*). O *betweeness* é uma medida de centralidade popular em biologia de redes que conta as vezes que um nó está envolvido num caminho mais curto (*shortest path*) numa rede. Geralmente o *betweeneess standard* é medido para todos os possíveis caminhos mais curtos entre quaisquer nós enquanto que o *S2B betweenness* apenas considera os caminhos mais curtos entre pares de DAGs. Portanto, o método S2B só prioriza a centralidade dos elementos ligando genes causativos de duas doenças. Além disso, sabendo que os nós altamente conectados (*hubs*) são mais propensos a aparecer por acaso num caminho mais curto entre DAGs, o algoritmo do S2B também utiliza dois algoritmos estatísticos baseados em aleatorizações da rede com os quais mede a especificidade dos *hubs* no contexto das doenças em estudo.

As proteínas resultantes da priorização realizada pelo S2B foram enriquecidos funcionalmente. Os resultados da análise de enriquecimento foram comparados com os resultados obtidos no analise da rede particular para as proteínas FTSS para assim, explorar qual é o papel do metabolismo do RNA e outros mecanismos moleculares hipotéticos na degeneração dos neurónios motores.

No conjunto das várias abordagens seguidas, este trabalho levou à descoberta de novos processos biológicos candidatos a mecanismos moleculares comuns entre a ALS e a SMA, mas também confirmou alguns processos já conhecidos simultaneamente envolvidos na ALS e na SMA. Globalmente, os nossos resultados sugerem cinco vias moleculares principais em comum nas duas patologias: 1) danos no DNA e apoptose induzidos pela desregulação da formação de "*R-loops*", 2) inflamação e neurodegeneração induzida por uma hipersensibilidade imunológica, 3) desregulação da cromatina e genotoxicidade produzida pela perturbação da biogénese de histonas, 4) alteração dos padrões de "*splicing*" e genotoxicidade criada pela falha da formação do spliceossoma e 5) desregulação de processos relacionados com microtúbulos que levam a problemas morfológicos na formação de axónios e sinapses.

As vias identificadas sugerem novas hipóteses que podem ser experimentalmente testadas. Assim, esta investigação pode ajudar a melhorar a compreensão dos mecanismos envolvidos na morte dos neurónios motores e também ajudar eventualmente ao desenho de alvos terapêuticos e biomarcadores para as MNDs. Além disso, também fornecemos um novo método para a priorização de DAGs candidatos a ligar os mecanismos moleculares de duas doenças relacionadas. Tal como no caso das MNDs, esperamos que este método ajude a comunidade a estudar outros tipos de doenças complexas.

Palavras-chave: esclerose lateral amiotrófica, priorização de genes associados à doença, atrofia muscular espinal, biologia de sistemas

Table of contents

List of figures and tables

Abbreviations

ALS: Amyotrophic Lateral Sclerosis
BP: Biological process
CUI: Concept Unique Identifier
DAG: Disease Associated Gene
DNA: Deoxyribonucleic Acid
FDR: False Discovery Rate
FTSS: FUS, TDP43, SMN, SETX proteins
GO: Gene Ontology
GOT: Gene Ontology Term
GTLinker: Gene Term Linker
hbn: High Betweenness node
hnRNP: Heterogeneous Nuclear Ribonucleoprotein
MN: Motor Neuron
MND: Motor Neuron Disease
mRNA: Messenger RNA
pol II: RNA Polymerase II
PPI: Protein-Protein Interaction
pre-mRNA: Precursor Messenger RNA
RBP: RNA Binding Protein
RNA: Ribonucleic Acid
RNP: Ribonucleoprotein
rRNA: Ribosomal RNA
SMA: Spinal Muscular Atrophy
snoRNA: Small Nucleolar RNA
SNP: Single Nuclear Polymorphism
snRNP: Single Nuclear Ribonucleoprotein
UMLS: Unified Medical Language System

Chapter 1: Background

Introduction

Systems biology aims to understand biological processes at the system level. It focuses on understanding the roles of interactions between genes, proteins, biochemical reactions and other components in an organism (Kitano 2002). Biological entities and processes are complex systems and therefore too intricate to analyze in a non-systematic way. These can be easily represented as networks and studied applying varied network theory concepts and methods (Barabási & Oltvai 2004).

Additionally, thanks to the development of high-throughput techniques, such as Next Generation Sequencing, we currently have countless amounts of biological data. The availability of completely annotated genome sequences of several organisms and the accessibility to databases of genomic, interactomic or metabolic information has allowed researchers to explore biological questions from a global perspective.

This is why networks have become a central resource in Systems biology. **Network biology** enables researchers to model, store, report, transmit and interpret molecular interactions (Hiesinger & Hassan 2005). Moreover, network biology has enormous applications in biomedical research. Phenotypes are emergent properties of the interactions among all of the components of a system (Hiesinger & Hassan 2005). Likewise, a disease is a pathologic phenotype caused by complex interactions that cannot be understood in a reductionist way. This led to the emergence of a new biomedical field commonly termed **network medicine** (Barabasi 2007). Network theory techniques can be exploited to describe networks' properties and structure that in turn, can help us understand the networks behavior and identify relevant elements within it. Particularly in biological systems, networks' characteristics can help to identify key genes or functions that may have great impact on biomedical research.

An interesting **object of research** in which network medicine can be very useful is the **exploration of commonalities between Amyotrophic Lateral Sclerosis (ALS) and Spinal Muscular Atrophy (SMA) diseases**.

Both are Motor Neuron degenerative Diseases (MND) and thus, share phenotypic characteristics. Knowing that both have numerous known Disease Associated Genes (DAG) and some of these are related to similar cellular functions, it has been hypothesized that ALS and SMA share molecular pathomechanisms. Nevertheless, it is not clear yet to affirm which are the most central genes and which are the most affected functions inducing this common phenotype.

For further analysis of these putative common pathomechanisms, we have pursued two network-based approaches. First, we used four proteins (SMN, FUS, TDP43 and SETX) as seeds to search the human interactome for common interactors, producing a new disease related network (*FTSS-focused network*). These four proteins where chosen because they are involved with the two diseases, have common molecular functions related with mRNA processing and besides, physically interact with each other. Second, we have developed a method that prioritizes within a human interactome network the proteins that show higher specific betweenness linking ALS and SMA-related genes and thus possibly have key roles in the pathogenesis.

In the remaining sections of this chapter, the main concepts behind the followed network-based approaches will be briefly presented.

Network theory

Networks are graphical representations of relations between discrete objects within complex systems. A **complex system** is characterized by two main properties; *i*) they are composed by many components and *ii*) these components are highly interconnected. The behavior of complex systems is quite different from merely the sum of the properties of its individual parts. Therefore, it is not possible to reliably predict the conduct of a complex system only by the simple extrapolation of the properties of a few components (Anderson P.W. 1972).

Networks are not only a convenient way of representing complex data but also mathematical and computationally easier to handle. Thus, network theory provides a set of techniques to analyze complex systems' structure and behavior. The nature of complex systems is highly diverse and therefore network theory methodologies have been exploited in variety of disciplines such as: communications, engineering, sociology, ecology and biology (Newman 2003).

The basic mathematical concept used to model networks is a ***graph***. It can be defined as a diagram representing a system of connections or interrelations (edges or links) among several units (nodes or vertices). In this work, the terms graph and network will be used with the same meaning.

Network properties

This section will discuss some important network parameters and measures useful for the analysis and understanding of network characteristics' impact on biological function. A comprehensive description of these properties can be found in Diestel 2000; Mason and Verwoerd 2007; Pavlopoulos et al. 2011; Winterbach et al. 2013 among others.

Interactions are the basis for building meaningful network models. According to edges' properties we can distinguish different type of networks. Depending on the existence of directionality in the interactions we can sort them in two broad classes**: directed** or **undirected networks**. Depending on the type of interactions and the scientific question to address, edges may have assigned scalar weights according to the relevance or reliability of the interaction. Thus **unweighted networks** model homogeneous relationships and **weighted networks** non-binary ones.

A **connected network** is a graph wherein any node has a path to reach any other node. When a network is unconnected, subsets of nodes and edges become isolated. These subsets or components can have very different sizes being the biggest one usually called **main component**. On the other hand, an **induced subnetwork** is a graph formed by a subset of nodes within a larger network and all the respective edges connecting that subset. Likewise, a **clique** is an induced subnetwork where every node directly interacts with every other node.

The analysis of network structural parameters allows distinguishing amongst various network topologies that in turn involve interesting biological properties.

The **shortest path**, as the name implies, is the minimum number of links one must traverse to move from node *u* to node *v*. The **distance** from a node *u* to a node *v* is the length of the shortest path from *u* to *v* in the network. The **average** (or characteristic) **path length** is the average distance over all pairs of vertices while **network diameter** is the greatest distance between any pair of connected nodes in a network. Conversely, **degree** is the number of edges incident on a certain node within the network whereas the **degree distribution** of a network is the fraction of nodes in the network with degree *k*. Thus the degree distribution P(*k*) gives the probability that a selected node has exactly *k* links.

Clustering coefficient of a given node is a ratio between the actual number of edges connecting the node's direct neighbors and the theoretically maximum number of edges that could connect them. Therefore, it measures the local density of links around a node. **Modularity** on the other hand quantifies the tendency of a graph to be divided into clusters. Knowing a partition of the network nodes into non-overlapping subsets it is possible to compute a **modularity coefficient** as the fraction of all network edges that connects nodes in the same subset. A **cluster** or **module** is a subset of nodes, which ideally has many more edges within the cluster than edges linking to external nodes. Networks' modularity is not always evident and usually several clustering algorithms need to be applied in order to get an acceptable modularity coefficient.

The problem of identifying the most important nodes in large complex networks is of fundamental importance (will the discussed more extensively in following sections). The relevance of nodes is usually based on centrality measures because it is commonly assumed that the removal of nodes with central positions in the network can lead to the network connectivity failure. There are several centrality measures based on varied assumptions and metrics. We will only refer two classical centrality measures that have been widely used in networks biology.

Degree centrality is based on the assumption that: *an important node is involved in a large number of interactions.* As we referred before, the degree of a node is the number of edges the node has connecting it to other nodes (illustrated Figure 1.1). Thus, the degree centrality of a node is its normalized degree.

Betweenness centrality is based on the assumption that: *an important node will lie on a high proportion of paths between other nodes in the network.* Therefore, the betweenness centrality of a node is the number of shortest paths of the network that include it. Nodes with low degree but high betweenness can be considered *bottlenecks* because their removal can be fatal to the network connectivity (Yu et al. 2007) (illustrated in Figure 1.1).

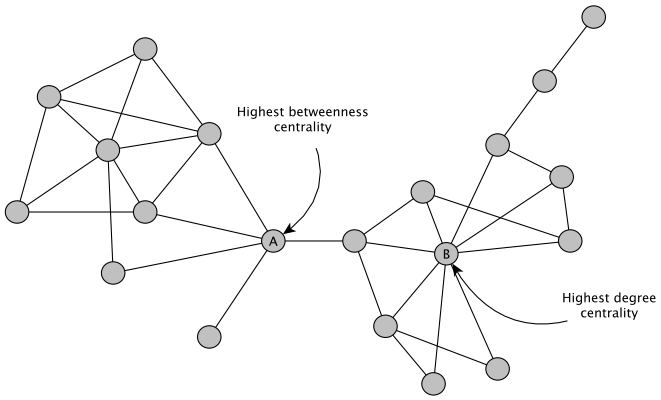

Figure 1.1 Illustration of the differences of centrality measures. There are pointed in the network the most central nodes according to degree (B) or betweenness measure (A) respectively. Node A is involved in the highest number or shortest paths and thus is a bottleneck for global connectivity whereas node B shows the highest number of edges (connections). Figure adapted from (Yu et al. 2007).

Besides the identification of central nodes, we can also be interested in finding the nodes that are more closely related with one or more nodes that are relevant under our subject study. The most direct approach is to consider the distance to the relevant node. Unfortunately, in most biological networks this approach is not sufficiently discriminatory as there will be numerous ties in the distance values. Therefore, other approaches have been used that take the information about the number of paths linking two nodes and the length of those paths.

These approaches can be based in diffusion (Kondor & Lafferty 2002) or in related random-walk algorithms (Can et al. 2005). The former simulate a **diffusion** process along the network edges starting from a seed node. After some time, the amount of the diffusible substance in each node measures the relatedness with the seed node. **Random walks** simulate the iterative trajectory of walkers that decide randomly what is the next node to visit among the direct neighbors of the present node. Different measures based on the trajectories of random walkers (random walk with restarts (Tong et al. 2006) or commute time (Fouss et al. 2007)) can quantify the relatedness between the seed node and other nodes in the network.

The systematic analysis of network parameters, such as degree distribution, characteristic path length or clustering coefficient, allowed researchers distinguish different type of networks models and describe their inherent properties. In this section we will describe three models that had a direct impact on the study of biological networks. For more detailed discussion we suggest several reviews (Newman 2003; Przulj 2011; Raman 2010; Winterbach et al. 2013)

Erdös–Rényi's random graph (Erdös & Rényi 1960) is the earliest random graph model in which edges are drawn between pairs of nodes uniformly at random with the same probability (Figure 1.2A). Random molecular networks can serve as a null model against which to compare biological network data results. The assumption is that the likelihood of an observed feature is determined by considering its distribution in randomly generated networks. In order to be more realistic we can apply some constraints in the randomization process. The network feature most commonly maintained is the nodes' degree.

Scale-free networks (Barabasi & Albert 1999) are characterized by a power-law degree distribution. This means that most of the nodes have relatively low degree while there are other nodes with an unusually high degree (Figure 1.2B). These highly connected nodes are usually called *hubs* and they may have key roles in the network functionality maintenance and hence have a high biological relevance.

Hierarchical networks (Ravasz & Barabasi 2003) have a scale-free topology with additional local modular structure (Figure 1.2C). These features are common in biological networks (Barabási & Oltvai 2004), thus we can focus on the biological features that emerge from the hierarchical networks' topology (Pavlopoulos et al. 2011).

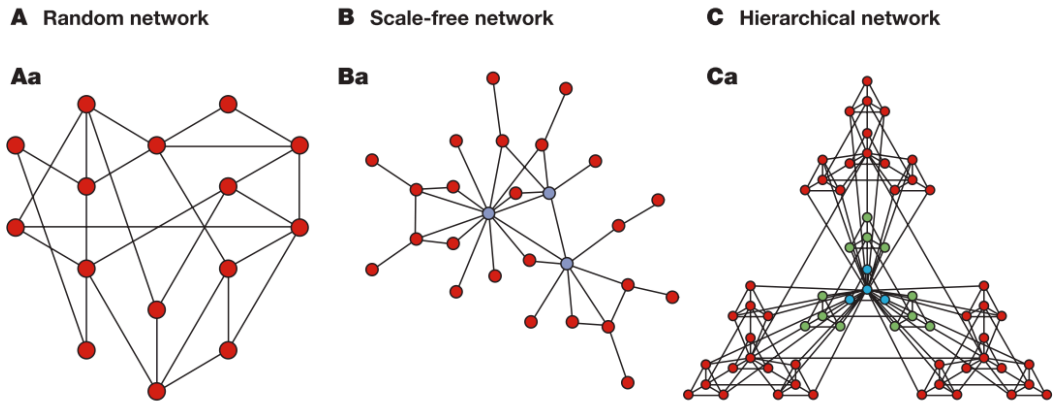

Figure 1.2 Network models that have a direct impact on understanding biological networks. The figure illustrates Erdös–Rényi (ER) model of a random network (part A) in which edges are drawn between pairs of nodes uniformly at random with the same probability. Then Scale-free networks (part B) are characterized by a power-law degree distribution. Finally hierarchical networks (part C) have a scale-free topology with additional local modular structure. Figure adapted from (Barabási & Oltvai 2004).

Network biology

As Emile Zuckerkandl and Linus Pauling stated in 1962, "*life is a relationship between molecules, not a property of any molecule*" (Zuckerkandl & Pauling 1962). This means today that biological characteristics only arise from intricate interactions between the cells' constituents, such us DNA, RNA, proteins and small molecules.

Network biology is the combination of systems biology and network theory principles with computational and statistical analyses in which the topology of the graphs representing molecular interaction networks themselves became the subjects of study (Barabási & Oltvai 2004)**.** The main objective of network biology is the modeling of biological networks and identification of relevant structures (Mason & Verwoerd 2007). As potential applications we can refer drug target identification, protein's or gene's function prediction, or the identification of new biomarkers to provide early disease diagnosis (Pavlopoulos et al. 2011).

In biological networks, nodes can represent molecules like genes, proteins, drugs or conditions like phenotypes or diseases. Edges represent interactions or associations between two nodes. According to the type of biological data that we integrate, there are several kinds of biological networks (Laukens et al. 2015; Pavlopoulos et al. 2011; Winterbach et al. 2013) .

Metabolic networks are formed by macromolecules; enzymes, cofactors and other metabolites needed for catalyzing biochemical reactions. These usually are weighted and directed graphs. Moreover, they are typically scale-free networks.

Signaling networks represent series of interactions between different bioentities such as proteins, chemicals or macromolecules. They are also usually directed graphs and are characterized by the presence of many feedback loops.

Co-expression networks are constructed with genes as nodes that are linked by edges when a similar co-expression pattern is found. They are directed networks and are not necessarily weighted.

Transcriptional regulatory networks contain information concerning the control of gene expression in cells. These networks are imperatively directed graphs. Moreover, they are typically sparsely connected graphs because genes are usually regulated by few transcription factors (Leclerc 2008).

Protein interaction networks hold the information of how proteins operate in coordination with others to enable the biological processes within the cell. These are typically unweighted and undirected graphs.

Network structure and robustness

Biological robustness is a property that allows a system to maintain its functions despite external and internal perturbations. It is considered to be a fundamental feature of complex evolvable systems (Kitano 2004). In the context of network theory, robustness is the resilience against node removal. Thus intuitively, a network is robust if its basic functionality (connectivity) is maintained even after the lost of some of its components.

Many biological networks have been described as scale-free or hierarchical networks (Khanin & Wit 2006; Barabasi & Albert 1999; Ravasz & Barabasi 2003). This apparent universality of topological features in biological networks does not seem a merely matter of chance. It is more likely the result of a high degree of internal order governing the cells' molecular organization in pursuit of a close balance between robustness and evolvability (Jeong et al. 2000; Barabási & Oltvai 2004; Kirschner & Gerhart 1998). **Evolvability** is the capacity of biological systems to generate a heritable phenotypic variation, that can potentially disrupt system's robustness (Kirschner & Gerhart 1998).

Therefore, identification of the basic architecture of a robust system and its associated trade-offs is essential for understanding the strengths and weaknesses of a given network. In a biological system these weaknesses can be translated as nodes' **essentiality**. In random networks if a critical fraction of nodes is removed, the network disintegrates into multiple non-communicating islands of nodes. On the contrary, scale-free networks do not have this critical threshold for disintegration. Even if 80% of randomly selected nodes fail, networks' integrity is maintained (Albert et al. 2000). This happens because the random failure, by chance, mainly affects low degree nodes. However this simultaneously means that the targeted removal of a few hubs could produce a catastrophic effect.

Robustness against perturbations not only depends on nodes' degree but also in the modules architecture and components' dispensability. Besides, modularity itself seems to be a mechanism for limiting the effects of local perturbations in cellular networks. The failure of a cluster blocks a particular function but avoids the general breakdown (Kitano 2004). Furthermore, it has been observed that protein complexes generally are composed of uniformly essential or non-essential molecules (Dezso et al. 2003). Therefore, the functions' dispensability is usually determined by the whole cluster essentiality (Barabási & Oltvai 2004).

It is also intuitive that those nodes with central localization in the network have a greater impact in the network connectivity and then, can be defined as essential nodes for networks robustness maintenance.

Node prioritization, that consists in identifying or ranking the most central nodes in the network is one of the most challenging aims in network theory (Mason & Verwoerd 2007; Freeman 1978). The key difficulty is the identification of a representative measure of essentiality. Furthermore, some nodes may not be necessarily central but hold crucial roles for particular functions on the networks (Mason & Verwoerd 2007).

Some authors proposed that node degree and essentiality might be related (Jeong et al. 2001) however relationship between centrality and essentiality is still under discussion (Zotenko et al. 2008). Besides, node degree alone seems to be a poor measure of essentiality due to the fact that available data sets are biased toward essential proteins because they are more frequently studied (Przulj 2011).

Protein-protein interaction networks

Among the multiple types of biological networks, protein-protein interaction (PPI) networks combine the availability of abundant data and the potential to uncover novel molecular mechanisms. Proteins are the main catalysts, structural elements, signaling messengers and molecular machines of biological tissues. Thus, protein-protein interactions are extremely important in orchestrating the events in a cell. In other words, PPI networks provide a simplified overview of the web of interactions that take place inside the cell (Raman 2010).

A **protein-protein interaction** usually refers to a binary relationship between one protein and another. However, the term "protein interaction" can include a great range of events such as transient and stable complexes, as well as physical and functional interactions. And likewise, a wide variety of approaches have been developed to detect protein-protein interactions. These associations may be direct physical interactions identified by experimental methods or indirect functional linkages predicted on the basis of computational analyses. For detailed information about the PPI prediction techniques we suggest two reviews (Raman 2010; Xie & Nice 2014).

Eventually, all this information is stored in numerous publicly available databases. We can classify the PPI databases into three main categories based on the methods used to collect or generate the data:

i) Repositories of experimental data collected through manual curation, computational extraction or direct deposit by the authors [**MINT** (Licata et al. 2012), **IntAct** (Orchard et al. 2014)]

ii) Predicted PPI interactions [**I2D** (new version of **OPHID**) (Brown & Jurisica 2005)]

iii) Portals that provides unified access to a variety of PPI databases [**STRING** (Snel et al. 2000), **GeneMANIA** (Warde-Farley et al. 2010) or **mentha** (Calderone et al. 2013)]

Here we describe those that were used in this thesis work. For further information about available databases we suggest (Ooi et al. 2010; Raman 2010; Tranchevent et al. 2010).

IntAct (Orchard et al. 2014) is a molecular interaction database that contains manually curated data from public literature or direct user submission. Most of the interaction data is PPI but also captures non-protein molecular interactions such as DNA and RNA. The detection methods and type of interactions are described together with all binary interaction and have assigned a weight to estimate its relevance. Moreover, IntAct is updated frequently and can also be downloaded (http://www.ebi.ac.uk/intact/).

mentha (Calderone et al. 2013) is a Search Tool that integrates high confidence interaction information curated by IMEx (Orchard et al. 2012) databases. These primary databases are manually curated but its literature coverage is not complete. Besides, mentha is focused on experimentally determined direct protein interactions, avoiding information extracted from text mining. These facts limit the number of results obtained from mentha when comparing to other search tools such as STRING (Snel et al. 2000) that is more focused on retrieving information with predictive tools from heterogeneous databases. Results are also returned as weighted interactions, which provide a confidence measure (http://mentha.uniroma2.it/).

GeneMANIA (Warde-Farley et al. 2010) is a gene function prediction server that uses a gene function prediction algorithm to reconstruct a composite network from several PPI databases. It also returns co-expression interactions and other interaction types that can be merged with PPI data in a unique network (http://genemania.org/).

CCSB Database (Rolland et al. 2014) is a compendium of high-throughput datasets that are focused on the prediction of high-quality binary PPIs. Most interactions gathered on these datasets were mainly mapped in a systematic approach and thus unbiased by the preferential study of proteins with biomedical or technological interest (Rual et al. 2005; Yu et al. 2011; Venkatesan et al. 2009; Rolland et al. 2014). These datasets are complemented with high confidence binary associations from the literature (Rolland et al. 2014). Therefore, CCSB results in a non-biased high confidence database that homogeneously covers the human interactome.

In all the cases, it should be pointed that the already identified PPI collection forms a small portion of the assumed total interactome (Rolland et al. 2014). Besides, the relation of PPI data quality with biological significance is not direct. It should also be remembered that PPI databases contain lots of interactions of proteins with ubiquitins, chaperons, ribosomal proteins and other similarly sticky proteins that might not be biological meaningful for the specific problem under study (Ooi et al. 2010).

Network medicine

The subcellular intricate connectivity implies that the impact of a specific genetic perturbation is not restricted to the activity of its gene product but can spread along the links of the network. Thus, a disease phenotype is rarely a consequence of an abnormality of a single gene but reflects the perturbations of the complex intracellular network (Barabási et al. 2011).

Furthermore, this complexity causes a deviation of the correlation between genotype and phenotype. There exist pleiotropic genes that can produce multiple phenotypes and there are numerous environmental factors that can also influence the gene expression patterns in similar genetic backgrounds (Kann 2007). Therefore, integrative analysis of interactions such as network-based approaches can help us to understand the organizing principles that govern cellular networks and their role in disease.

Network medicine is then the application of networks-based approaches into biomedical problems. Is the set of methodologies that aims to understand diseases through theirs underlying molecular interactions (Barabási et al. 2011).

Biological functions are accomplished by the coordinated participation of biological components (metabolites, proteins or genes) (Zuckerkandl & Pauling 1962), thus as *"guilt by association"* principle claims, those proteins that physically interact are strongly suspected to be involved in similar functions (Gillis & Pavlidis 2011; Oliver 2000). Likewise, proteins involved in the same disease-phenotype show a high propensity to interact with each other (Goh et al. 2007; Gandhi et al. 2006; Oti & Brunner 2007).

According to this, we can distinguish three types of interrelated modules:

i) ***topological modules*** that gather highly connected and close nodes,
ii) ***functional modules*** that are conformed by nodes with similar functions on the neighborhood of the network (Spirin & Mirny 2003) and
iii) ***disease modules*** that represents groups of nodes that together contribute to a cellular function disruption that results in a particular phatophenotype (Barabási et al. 2011).

Thus, it is accepted that gene products related to a particular pathophenotype tends to be located in a close neighborhood within the protein-protein interaction network and these are usually related to similar functions (Gandhi et al. 2006; Oliver 2000) and vice-versa. Conversely, if a same gene is linked to two different disease phenotypes this linkage is often an indication that the two diseases have a common genetic origin (Barabási et al. 2011).

Furthermore, there is evidence from many sources that similar phenotypes are the result of functionally related clusters of genes and this is even more obvious in the case of genetically heterogeneous diseases (Oti & Brunner 2007). Moreover, genes associated with a disease preferentially interact with other disease-causing genes over those without any known disease association (Aravind 2000). Thus, different disease modules are more prone to overlap, so that perturbations caused by one Disease Associated Gene (DAG) can affect other disease modules (Goh et al. 2007).

The presence of *hubs* in biological networks suggests that these highly interconnected nodes may play essential roles in biological functions. However, *hubs*' essentiality suggest that these nodes cannot be directly associated to disease causative genes. *Hubs* inactivation could cause the network systemic failure and possibly the early death of the individual. Thus, it is improbable that such genetic alterations can persist in the population (Barabási et al. 2011). Consequently, disease causative genes are expected to be found on the network periphery where they can be more easily tolerated and inherited by the progeny (Goh et al. 2007). Additionally, disease genes seem to be tissue-specific while essential genes are expressed in multiple tissues (Barabási et al. 2011) .

All these observations can be exploited to predict novel DAGs on the basis of the particular location of candidate genes within the studied networks. As such, DAGs prediction (or prioritization) has become one of the most popular applications of network medicine.

Computational methods for gene prioritization are necessary to effectively translate the high-throughput derived experimental data into legible disease-gene associations (Moreau & Tranchevent 2012).

DAGs prioritization methods typically use networks where the nodes are genes or proteins and the aim is to prioritize the nodes that are more important for a given disease (Bromberg 2013). In these networks data is integrated by attributing weights to nodes and by defining the edges and associated weights. Edges can reflect PPI, co-expression patterns or other molecular interaction information. Centrality measures, distances between nodes or diffusion-based algorithms can then be used to rank all the nodes in the network, prioritizing the nodes most relevant for the disease that are not part of the initial set of DAGs (Tényi et al. 2016; Köhler et al. 2008; Simões et al. 2015; Wu et al. 2015; Calderone et al. 2016).

Others build **phenotypic networks** in which the nodes are diseases. In particular, Goh and colleagues constructed a highly detailed *diseasome network* (Goh et al. 2007). Disease networks can helps us comprehend why and how certain groups of diseases arise together, share molecular mechanisms or have common phenotypic properties (Piro 2012). In disease networks, edges can represent varied type of data such as comorbidities (disease co-occurrences), common phenotypic features or common DAGs.

Network based disease-gene prioritization methods require interactomic data and also a minimum set of known DAGs. There are numerous collections of human DAGs (Kann 2007) but we will only describe the two that were used in this work.

OMIM database (Hamosh et al. 2002) is knowledge based, manually curated and frequently updated. Initially focused on monogenic disorders, nowadays includes information of complex and multifactorial diseases. As part of the NCBI Entrez database, OMIM is freely available and contains over 15 000 genes with known sequence and over 6000 phenotypes. As other comparable databases, it does not constitute a standard library for describing disease phenotypes. The lack of a controlled vocabulary and consistent annotations hampers the information retrieval.

DisGeNET (Pinero et al. 2015) is one of the largest repositories currently available of its kind. It integrates expert-curated databases with text-mining data covering information of monogenetic and complex diseases. Besides, it has implemented a score based on evidence to prioritize gene-disease associations. It provides standardized annotations of entities (genes and diseases) and their relationships (ontologies), which helps in the retrieval and analysis of information.

Motor neuron degeneration-related diseases (MND)

Spinal Muscular Atrophy (SMA) and Amyotrophic Lateral Sclerosis (ALS) are both **Motor Neuron Diseases** (MNDs), a group of progressive neurological disorders that destroy motor neurons. These cells control essential voluntary muscle activity such as speaking, walking, breathing, and swallowing. Normally, messages from nerve cells in the brain (upper motor neurons) are transmitted to nerve cells in the brain stem and spinal cord (lower motor neurons) and from them to particular muscles. Upper motor neurons coordinate the lower motor neurons to produce movements such as walking or chewing whereas lower motor neurons control movement in the arms, legs, chest, face, throat, and tongue.

In general, affected individuals lose strength and the ability to move their arms and legs, and to hold the body upright. When muscles of the diaphragm and chest wall fail to function properly, individuals lose the ability to breathe without mechanical support. In the most severe cases of SMA, children never sit or stand and the vast majority usually dies of respiratory failure before the age of 2. On the other hand, ALS patients usually die within 3 to 5 years from the onset of symptoms (typically on the third decade of life).

The molecular causes of motor neurons' degeneration are still unclear, limiting therapeutic options and consequently patients' life expectancy.

Spinal Muscular Atrophy (SMA) is a childhood onset disease characterized by the degeneration of lower motor neurons (localized in the spinal chord). It is an autosomic recessive disease caused by mutations in the Survival Motor Neuron (SMN1) gene (Lefebvre et al. 1995).

Besides SMN1, SMN2 gene also codifies SMN protein. SMN1 and SMN2 share more than 99% nucleotide identity, and both are capable of encoding a functional SMN protein. How ever, due to a silent substitution in exon 7, ~75-90% of SMN2 transcripts encode for a truncated and unstable splicing isoform of SMN, and thus a very small amount of full length-active protein is actually being produced from this gene (Lefebvre et al. 1995; Kashima et al. 2007; Wirth 2000). Henceforth we will only refer to SMN as a conjunction of both genes.

The SMN protein generates the core machinery for varied RNA-metabolism related functions pathways including pre-mRNA splicing, histone mRNA 3'-end processing and cytoplasmic mRNA decay (Li et al. 2014). The SMN1 gene product forms the SMN complex together with Gemin proteins. SMN complex in turn plays a critical role in the assembly of small nuclear ribonucleoproteins (snRNPs) that constitute the spliceosome machinery (Lefebvre et al. 1995). The spliceosome is a macromolecular ribonucleoprotein (RNP) complex responsible for intron removal from pre-mRNAs (a process commonly known as splicing). This process is critical for the production of correct mRNAs and also for generating transcripts' diversity through alternative splicing events.

Among the snRNPs, SMN low levels also affects to U7snRNP biogenesis (Tisdale et al. 2013). This snRNP is crucial for the histone mRNA 3'end processing. Most histone transcripts are not poly-

Adenylated and their 3' ends are produced by an endonucleolytic cleavage mediated by U7snRNP. Besides, SMN is also knwon to recognize methylation marks in histones (Sabra et al. 2013). On the other hand SMN also is involved in axonal mRNA transport. It regulates the mRNA vesicles assembly (Zhang et al. 2006) and interacts with axonal transport machinery by direct association with acting-binding proteins and neuronal RBPs (Fallini et al. 2012).

Amyotrophic Lateral Sclerosis (ALS) in the other hand is an adult onset disease characterized by the degeneration of both lower and upper motor neurons (localized in the spinal chord and brain respectively). In 90% of the cases it is a sporadic disease but there are also familial subtypes caused by mutations in numerous genes related to diverse cell functions as oxidative stress (SOD1), RNA metabolism (TARDBP, FUS, Senataxin, Ataxin2, HNRNPA2/B1, ELP3, HNRNPA1), vesicle trafficking (Alsin, FIG4, OPTN, VABP, CHMP2B) and proteasomal function (UBQLN2, VCP) (Siddique & Siddique 2008; Carrì et al. 2015).

SOD1 was the first and one of the best-known familial ALS-causative genes (Rosen et al. 1993). SOD1 gene encodes the superoxide dismutase 1 protein, which is responsible for destroying free superoxide radicals produced by the oxidative metabolism. Therefore when it is mutated, cells suffer oxidative stress that induce mitochondrial and endoplasmic reticulum stress that usually results in protein aggregation and apoptosis pathway activation (Carrì et al. 2015). However, should be listed that SOD1 mutation only causes the 1% of ALS cases (Marangi & Traynor 2015).

TDP43 (the TARDBP gene product) was firstly associated to ALS in 2008 (Sreedharan et al. 2008; Rutherford et al. 2008). It is a DNA and RNA biding multifunctional protein that can be localized in the nucleus or cytoplasm according to the function in which it is involved. It has been related to several steps of the gene expression pathway including transcription, splicing, RNA transport and localization, mRNA decay (stabilization) and translation (Lagier-Tourenne et al. 2010).

FUS also identified as ALS-causal gene (Kwiatkowski Jr et al. 2009) is as well a nucleocytoplasmic-shuttling multifunctional protein that not only binds to RNA but also to DNA. Thus it is involved in DNA and RNA metabolism including DNA repair, regulation of transcription, RNA splicing and RNA export to cytoplasm(Lagier-Tourenne et al. 2010).

Senataxin (SETX) was firstly related to a juvenile form of Amyotrophic Lateral Sclerosis (ALS4) (Chen et al. 2004). It belongs to the superfamily I of DNA-RNA helicases and it is involved in diverse aspects of RNA metabolism and genome integrity maintenance.
Senataxin coordinates the binding of RNA polymerase II (pol II) to chromatin and therefore, regulates the transcription in all steps. The nascent RNA is coated by RBPs that protect it from the template DNA-strand. This potential DNA/RNA hybrid structure is called R-loop and pauses pol II progression allowing the correct termination of transcription. In absence of R-loop structure, transcription continues leading to read-through intergenic products (Richard & Manley 2016). Additionally, for the proper ending of transcription it is also necessary the R-loop resolution conducted by Senataxin. Lack of SETX activity induces genome damage and instability (Skourti-Stathaki et al. 2011). To simplify we refer SETX both to gene and Senataxin protein.

RNA-metabolism events are ubiquitous and necessary for all cells' survival thus, we are still missing the molecular link between RNA related functions and MNs specific degeneration. Within this set of functions associated to MND-DAGs, splicing seems to be the central piece. SMN is the main causal gene of SMA disease and is directly involved in splicing function maintenance (Lefebvre et al. 1995). On the other hand, ALS related gene products of FUS and TDP43 are also involved in splicing related functions. Furthermore, nervous tissue has an extremely high alternative splicing activity, which is believed to have a great relevance for neural development and differentiation (Madgwick et al. 2015). However, these DAG-products are multifunctional proteins, thus it is not clearly evident which is the key function that, when perturbed, leads to the MND phenotype.

Moreover, there are further non-RNA metabolism related functions directly associated to the MN death. Some authors believe that, due to the high sensitivity of the nervous tissue to reactive oxygen species (Friedman 2011), oxidative stress could be the main disturbing agent. The oxidative stress may be produced by SOD1 dysfunction that leads to protein aggregation, mitochondrial and endoplasmic reticulum stress and eventually induces apoptosis (Carrì et al. 2015).

Others claim that the main cause is the genotoxic stress induced by deregulation of histone expression as a consequence of SMN dysfunction (Tisdale et al. 2013)This hypothesis is also plausible due to the critical role of chromatin remodeling in nervous tissue development and differentiation (Pattaroni & Jacob 2013). Therefore, although several hypotheses about MN degeneration initiation causes have been proposed, the specific pathomechanisms are still unknown.

Phenotypic and genetic similarity between ALS and SMA suggest the presence of closely related disease modules and opens the door to the study of their common molecular pathomechanisms using network-based approaches.

Objectives

ALS and SMA are diseases with different causes but common complex phenotypes. In both cases the molecular mechanisms that link the causes of disease with motor neuron degeneration are not completely known. **This thesis work aims to apply network biology approaches to characterize common molecular pathways involved in ALS and SMA**. In particular this thesis will:

1 – Analyze a protein interaction network around proteins known to be involved in both diseases and in a common molecular pathway – RNA processing (Chapter 2)

2 – Develop a network-based prioritization method to identify relevant proteins specifically connecting genes associated with both diseases (Chapter 3)

Chapter 2: FTSS-focused network

Introduction

ALS and SMA show great phenotypic and molecular similarities. FUS, TARDBP (TDP43) and SETX are ALS-associated genes (Rutherford et al. 2008; Sreedharan et al. 2008; Kwiatkowski Jr et al. 2009; Chen et al. 2004) while loss of SMN activity is the main cause of SMA (Lefebvre et al. 1995).

These four genes (FTSS) are known to physically interact (Sun et al. 2015; Tsuiji et al. 2013; Yamazaki et al. 2012; Skourti-Stathaki et al. 2011; Zhao et al. 2016; Suraweera et al. 2009; Bennett & La Spada 2015). Furthermore, all four proteins are involved in RNA processing, which may have direct implications on MND common phenotype (Cooper et al. 2009).

Gems are nuclear structures that contain SMN complex components but not snRNPs. Both TDP43 and FUS have been found to bind SMN and accumulate in Gems thus seem that they are required for the maintenance of these structures (Ishihara et al. 2013; Tsuiji et al. 2013; Yamazaki et al. 2012). Gems' function is still unknown but probably they are directly related to snRNP maturation processes (Liu & Dreyfuss 1996; Cioce & Lamond 2005; Clelland et al. 2009). Additionally, it was found that Gems assembly is disrupted in ALS model SOD1 deficient mutant mice (Kariya et al. 2012). All these evidences place Gems structures at the core of ALS and SMA-related events, being the snRNP immaturity a potential key factor in MN degeneration.

Together with SMN, FUS has been found to interact with U7snRNP and therefore, it is suspected to be involved in histone mRNA biogenesis (Raczynska et al. 2015). This fact is relevant in MND context because it is known that transient histones modifications can produce lasting cellular changes that influence the synaptic plasticity (Levenson & Sweatt 2005). Moreover, histone biogenesis perturbation can also lead to global chromatin changes that eventually induce cellular genotoxicity (Tisdale et al. 2013).

Additionally, FUS is also involved in DNA damage response events (Rulten et al. 2014) together with SMN that at the same time acts as intermediary between SETX and RNA Pol II to coordinate R-loop formation in RNA elongating complexes (Zhao et al. 2016). In the same way, SETX has been directly related to the immune response depression and when mutated to a hyper-sensitivity to infections (Miller et al. 2015). This fact may be key in MND because SETX is already associated to ALS4 (ALS type 4) and ataxia with oculomotor apraxia (AOA2) (Chen et al. 2004). Besides it is known that the prolonged immune response and expression of inflammatory mediators leads to cell death and neurodegeneration (Amor et al. 2010; Friedman 2011).

On the other hand, it has been observed in mice that the increase of SMN levels improves the neuromuscular function altered by the oxidative stress produced by the lack of SOD1 activity,

typical in ALS (Turner et al. 2014). Likewise, SETX and FUS are involved in DNA damage response generated by oxidative stress (Suraweera et al. 2007; Rulten et al. 2014) and thus, their concurrent failure could lead to fatal perturbations.

These evidences strongly suggest that FTSS proteins are tightly interrelated and physically interact to perform functions mainly related to RNA metabolism. Though RNA metabolism seems to have a central role, it has not been explained yet how RNA-related ubiquitous functions may lead to the death of MN in particular. Furthermore, some authors suggest that the widespread changes in splicing are an indirect feature of the delayed neuronal development (Garcia et al. 2013) caused by the perturbation of synaptogenesis (Zhang et al. 2013). On the same theme, the evidences of perturbations in varied functions non-related to RNA-metabolism raises further doubts about the initial steps on MND.

To answer these questions, we have constructed a PPI network focused only on FTSS genes and its common interactors (*FTSS-focused network*) and performed a functional enrichment analysis to characterize this protein set. We expect that, the overrepresented functions resulting from this analysis may bring clues about MND-related pathomechanisms.

Methodology

PPI data retrieval

FUS, TDP43, SMN and SETX (FTSS) protein physical interactors were retrieved from mentha (Calderone et al. 2013) (http://mentha.uniroma2.it/), IntAct (Orchard et al. 2014) (http://www.ebi.ac.uk/intact/), GeneMANIA (Warde-Farley et al. 2010) (http://genemania.org/) and literature references (Sun et al. 2015; Tsuiji et al. 2013; Yamazaki et al. 2012; Skourti-Stathaki et al. 2011; Zhao et al. 2016; Suraweera et al. 2009; Bennett & La Spada 2015). The referred databases where accessed during November 2015.

To facilitate the retrieval of SMN PPIs, the physical interactions found with both SMN1 and SMN2 genes were included. We only selected PPIs identified in *Homo sapiens* and described as physical interactions (Figure 2.1-1). From the initial set of interactor proteins, only the proteins that interacted with at least two of the FTSS proteins were retained (Figure 2.1 step 1). All raw and curated datasets are available in Supplementary data S-2.1.

FTSS-focused network functional enrichment and functional clustering

Functional enrichment analysis and functional clustering were performed simultaneously using Gene Term Linker (GTLinker) algorithm (Aibar et al. 2015) available within the FGNet R-package and through http://gtlinker.cnb.csic.es/.
The functional enrichment was constrained to Biological Processes (BP) Gene Ontology Terms (GOTs) annotated in *Homo sapiens* proteins. GOT enrichment was considered statistically significant with a False Discovery Rate (FDR) adjusted p-value less than 0.05 (Figure 2.1-2).

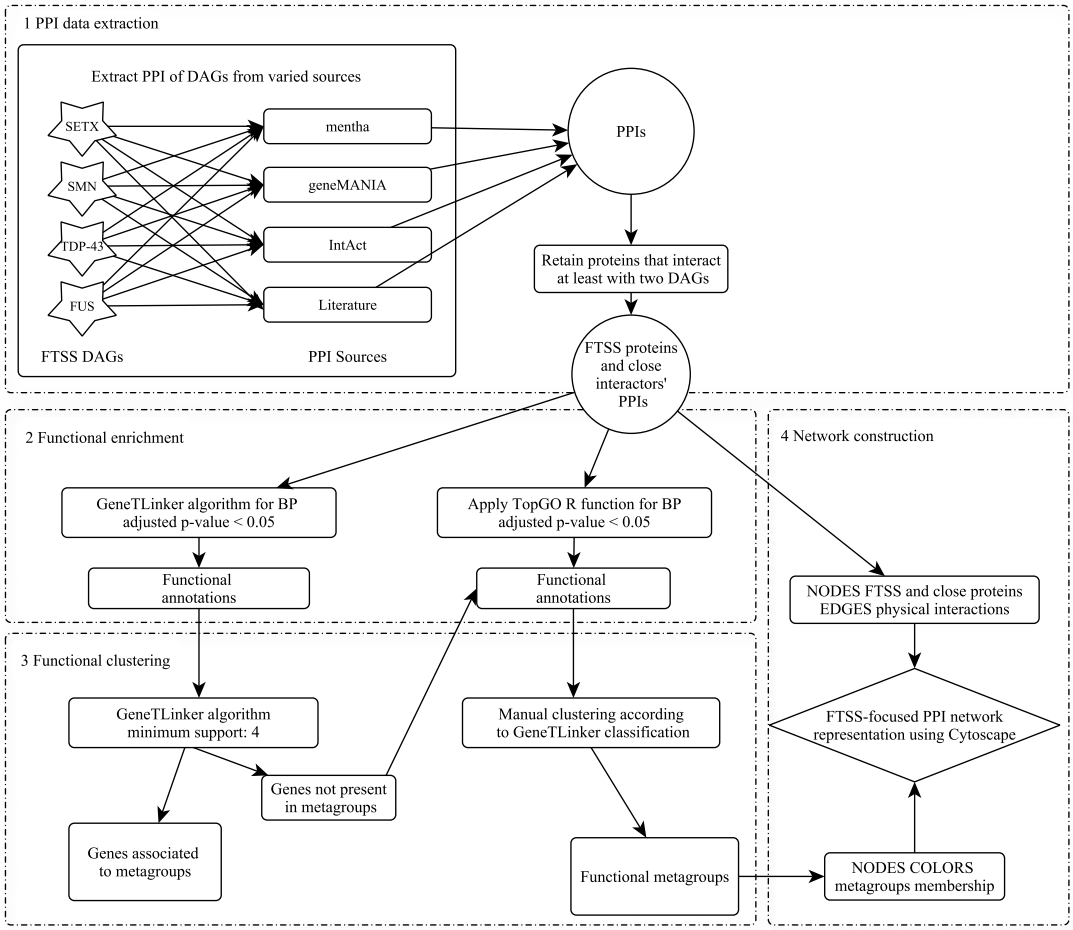

Figure 2.1 Flowchart describing FTSS-focused network construction methodology. FTTS-focused network is constructed using PPI related to FTSS proteins. Proteins were maintained if interact with at least two of the FTSS proteins (step 1). Then, there was performed a functional enrichment analysis using GTLinker and TopGO algorithms (step 2). Metagroups created by GTLinker algorithm were manually edited (step 3) and finally FTSS-network with functional annotation was constructed (step 4).

The GTLinker clustering algorithm was applied to annotations with at least 4 associated proteins (minimum support) (Figure 2.1-3). Since GTLinker metagroups are formed by varied GOTs, we manually assigned general titles to each cluster. Besides, GTLinker does not return

functional information from proteins excluded by the clustering algorithm so, we also performed an additional functional enrichment analysis using topGO R function (Alexa & Rahnenfuhrer 2010) (Figure 2.1-2). We used this functional data to, when possible; manually include the remaining proteins into the metagroups suggested by GTLinker result (Figure 2.1-3). All data relative to functional enrichment and clustering analyses are available in Supplementary data S-2.2.

FTSS-focused network construction

The PPI data and functional information collected in previous steps was finally plotted as a network using Cytoscape software version 3.4.0 (Shannon et al. 2003) (Figure 2.1-4).

Results and Discussion

PPI data retrieval

The initial sets of interactions originated a network with 636 proteins and 1007 interactions with the FTSS proteins. Only proteins that interact with at least two of the four studied proteins were retained, resulting in a final set of 136 proteins and 377 interactions.

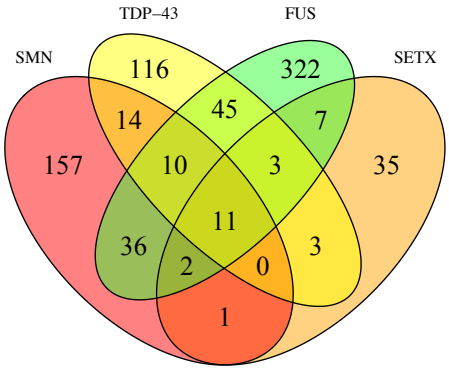

Figure 2.2 Venn diagram describing retrieved FTSS-related PPIs: Each overlapping area describes the number of intermediary proteins interacting simultaneously with FUS, TDP43, SMN and/or SETX. Those non-overlapping areas describe the number proteins that interact only with a single FTSS protein and thus were removed from the FTSS-focused network.

According to Figure 2.2, FUS is the most promiscuous protein having described a total of 436 PPIs. Besides it is the protein that more PPI losses (322) when filtering which reflects FUS multifunctionality and plus demonstrates that it is also involved in many functions non-related to those associated to FTSS proteins. SMN and TDP43 show the same behavior as FUS but in a

lesser extent. Finally, SETX despite of being the protein with less PPIs described (62), is also the protein that less PPI losses relatively. This suggests that SETX is more specialized in functions simultaneously related to those of FTSS proteins.

On the other hand, looking to the Venn's diagram intersecting areas, there are 11 proteins interacting with all FTSS proteins (*FTSS clique*). The existence of this clique underlines the close interactomic relationship among FUS, TDP43, SMN and SETX and evidences their common role in certain functions.

Now comparing SMN (as SMA-associated protein) against FUS, TDP43 and SETX (as ALS-related proteins) we can see that SMN and FUS are the most closely related proteins sharing the highest number of interacting proteins (there are 59 proteins intersecting SMN and FUS areas in Figure 2.2).

The existence of many common interactors between SMN and FUS reinforce the evidences of their common role on Gems maintenance (Ishihara et al. 2013; Tsuiji et al. 2013; Yamazaki et al. 2012) and therefore, their impact on snRNPs maturation (Liu & Dreyfuss 1996; Cioce & Lamond 2005; Clelland et al. 2009). Furthermore, the close relationship of FUS and SMN also sustain their combined impact on snRNPU7 biogenesis and on histone mRNA processing (Raczynska et al. 2015; Tisdale et al. 2013).

SMN and TDP43 also share a high number of interacting proteins (35 proteins intersecting SMN and TDP43 areas in Figure 2.2). This observation could correspond to the fact that TDP43 has been also found interacting with SMN and FUS on Gems (Ishihara et al. 2013; Tsuiji et al. 2013; Yamazaki et al. 2012) and thus, TDP43 might be also necessary for Gems stability and snRNPs maturation.

Finally SMN and SETX share 14 interactors (Figure 2.2), despite being a lower number it represents a great fraction of total interactions associated to SETX.
SETX is known to have an active part on R-loops formation. SMN has also been found as direct interactor between RNA pol II and SETX (Zhao et al. 2016). Thus, their close relationship highlights the additional role of SMN on transcription termination and therefore, SMN lower activity might not only lead to splicing pattern alterations (Skourti-Stathaki et al. 2011) but also to DNA damage and genomic instability.

FTSS-focused network

From the initial set of 136 proteins, GTLinker enrichment and clustering algorithm created 8 functional metagroups including 70 proteins. The remaining proteins were manually classified according to the results of TopGO enrichment algorithm. It was needed to create an additional metagroup to englobe proteins related to cell death and apoptosis. There were 18 proteins without known GOT that were not included in any metagroup.

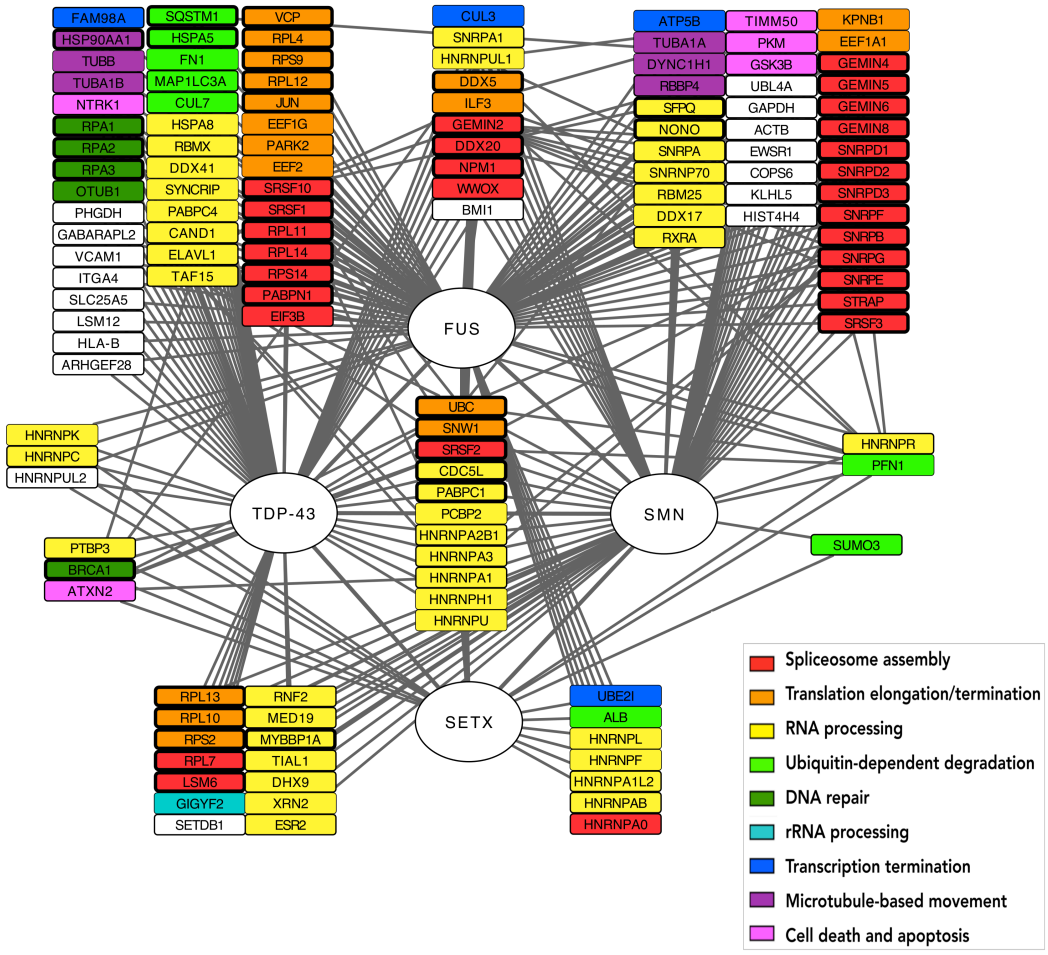

Figure 2.3 FTTS-focused network: Nodes represent FUS, TDP43, SMN, SETX proteins and direct interactors. **Edges** describe physical interactions. Annotations of the proteins in the network where functionally enriched using TopGo and clustered using GTLinker algorithm to simplify the functional characterization. Functional clusters are identified by **node color** and summarized in the attached legend. With the exception of FTSS proteins, nodes without color are proteins without known functions. Proteins that were included in more than one cluster are assigned to the cluster with the lowest enrichment p-value and identified with a **bold border**.

It is noteworthy that the interactors of FTSS proteins are related to a variety of functions (Figure 2.3), which again emphasizes the FTSS proteins' multifunctionality.

Since RNA processing is a very general concept, it is expected that this cluster embraces the highest number of proteins (Figure 2.3 yellow nodes). Moreover, they are homogeneously distributed in the network. Furthermore, the majority of proteins that constitute the FTSS clique (that interact simultaneously with all FTSS proteins) are related to RNA processing and thus, possibly form a protein complex. Most of these proteins are heterogeneous nuclear ribonucleoproteins (hnRNPs) known by their involvement in pre-mRNA processing and mRNA transport.

As expected, splicing related functions are over-represented in FTSS-focused network (Figure 2.3 red nodes). In particular, proteins involved in splicing are more closely related to FUS, SMN and TDP43 than to SETX. Likewise, the translation elongation and termination cluster (Figure 2.3 orange nodes) shows a similar distribution to spliceosome assembly cluster. Thus, FTSS proteins seem to participate in translation as well.

Ubiquitin dependent function cluster also forms a small and widely distributed group (Figure 2.3 light green nodes). It is worth mentioning that ubiquitin-dependent degradation can be involved in varied regulatory processes such us mRNA nuclear transport, DNA repair, cell cycle, immunity and signal transductions (Muratani & Tansey 2003; Peng et al. 2003). Furthermore, protein degradation has many critical roles in synaptic plasticity (Diantonio & Hicke 2004). Nevertheless we should be careful and recall that ubiquitin proteins are intrinsically promiscuous and thus these results may be unspecific.

Conversely, there are few proteins associated to microtubule related functions (Figure 2.3 dark pink nodes) that may have a critical role in MN degeneration. It is known that microtubules are involved in fast-axonal transport and then are required for axons' and synapses' structural and regulatory maintenance (Griffin & Watson 1988; Poulain & Sobel 2010).

Moreover, SMN has been directly involved in the axonal transport failure in two main ways. The first hypothesis is that SMN; together with other RNA binding proteins is key to form the mRNA vesicles that will be transported through the axons (Zhang et al. 2006). Therefore it has a key role in the maintenance of axons and in turn synapses. The second striking evidence is that the assembly dynamics of the cytoskeleton and particularly microtubules is mediated by tissue-specific alternative splicing events that in turn have direct impact on nervous tissue development (Madgwick et al. 2015). Strikingly, FUS is the common interactor between microtubule-associated proteins in the network (Figure 2.3) thus, FUS appears again as a central element in MN degeneration.

Despite FUS and SETX are both involved in DNA repair (Rulten et al. 2014; Suraweera et al. 2009) proteins associated to the DNA repair cluster (Figure 2.3 dark green nodes) are mainly interacting with FUS. We might hypothesize that although both being related to DNA repair, only FUS and its interactors are involved in the DNA repair failure in MND while SETX-mediated DNA repair could possibly be activated through other non-related pathways.

Finally, transcription termination cluster (Figure 2.3 dark blue nodes) and rRNA processing (Figure 2.3 light blue nodes) clusters are the smallest groups. However, it is important to highlight that many proteins were related to more than one cluster (Figure 2.3 nodes with bold border). Thus, RNA processing, spliceosome assembly or translation elongation and termination clusters grew at the expense of more specific clusters such as transcription termination cluster or rRNA processing.

Conclusions

FTSS-focused network architecture demonstrates that FUS, TDP43, SMN and SETX are very closely interacting. This is a very important fact because they are directly related to ALS and SMA pathomechanisms and thus may serve as a bridge to link both MNDs.

Despite of being multifunctional proteins, FTSS-focused network functional characterization suggests that RNA-metabolism has a central role among these proteins. In broad terms, the RNA processing cluster is the most over-represented functional group followed by the expected spliceosome assembly related cluster. In particular, hnRNP complex dominates the FTSS clique. Since this clique represents the closest interactomic and functional relationship among FTSS proteins and hypothetically the ALS and SMA phenotypes, pre-mRNA processing and mRNA transport functions acquire great relevance in MNDs.

On the other hand, FTSS-focused network also brought more insights into the understanding of motor neuron degeneration mechanism. Namely, microtubule-related functions acquire new relevance in the degeneration of motor neurons. Microtubule-associated proteins act as important regulators in the development of neurite, axon and dentrite formations (Poulain & Sobel 2010). Besides, cytoskeleton dynamics seems to be highly dependent of tissue-specific splicing activity. Additionally to SMN's role in spliceosomal assembly, it is also involved in mRNA vesicles microtubule-dependent transport through axons. Furthermore, in FTSS-focused network FUS is closely interacting with microtubule-assembly related proteins. Thus, the functional link between SMN (as SMA-DAG and FUS (as ALS-DAG) sets microtubule-based axonal transport in the center of the board as the phenomenon that possibly links the observed genetic and phenotypic features of MNDs.

FUS seems to be a very influential protein not only due to the large number of PPIs but also because of its interactors-associated functions. Together to the prior evidences of its involvement in Gems structure maintenance or histones mRNA processing, we show that FUS is also interacting with proteins related to splicing, microtubule-based movement and DNA repair among others. Furthermore, FUS promiscuity (the existence of a high number of other interactors not related to FTSS) might be an interesting fact considering that its perturbation could lead to the alteration of a number of varied functions within the cell as observed in MND phenotypes.

The perturbation of histone mRNA biogenesis induced by U7snRNP dysfunction (produced by FUS and SMN mutations) can have a great impact on synapse plasticity. Conversely, mutated SETX is also directly involved in immune hyper-response and therefore in neurodegeneration. Despite being directly related to neuron survival, these perturbations are still too general to address motor neuron death specifically. Histone biogenesis changes can produce global perturbations in chromatin patterns and therefore, overall changes in gene expression and cell physiology. Besides, these proteins are also essential for damaged DNA repair and thus, when altered can produce fatal changes in the genotype of any cell type.

We are aware of many other MND-DAGs that were not considered and thus, the over representation of RNA-metabolism related functions could be a result of this bias. Besides, this research is based on the existence of PPIs and ignores the MND diseasome structure.

Despite the referred limitations, we have shown that this simple approach to construct a FTSS-focused network is able to retrieve relevant information about motor neuron degeneration and provide novel cues for further MNDs research.

Chapter 3: S2B method

Introduction

Biological entities are extremely complex systems and biological properties only arise from intricate interactions among the cells' constituents, such as DNA, RNA, proteins and small molecules (Barabási & Oltvai 2004). Likewise, diseases are the phenotypic result of genetic or environmental alterations that perturb the interactome (Barabási et al. 2011).

Thanks to high throughput technologies, we currently have at our disposal a large number of disease-gene associations. However they need to be computationally analyzed to effectively obtain new biological knowledge. Conversely, due to the intrinsic properties of biological data, one of the best ways of representing it is in the form of complex networks. Furthermore, biological networks present a highly conserved architecture that enables the use of network theory principles to infer biological information and associations (Newman 2003; Przulj 2011; Raman 2010; Winterbach et al. 2013).

Amongst the applications of networks theory in biomedical research, Disease-Associated Genes (DAGs) prioritization is one of the most common goals particularly because it is necessary to filter the large lists of DAG candidates returned by high-throughput derived experiments (Moreau & Tranchevent 2012). There are varied types of network-based DAG methods being the most usual those that represent disease molecular details through Protein-Protein Interaction (PPI) networks.

Generally they work under *"guilt by association"* assumption, which states that network neighbors of disease genes tend to cause similar diseases (Oliver 2000). The majority is focused on the integration of heterogeneous networks to better characterize the disease context and therefore obtain the best results. **Chain Rank** method (Tényi et al. 2016) for example allows the implementation of user-defined scores to integrate signaling, regulation or interactomic data .

Typically, DAG prioritization methods explore the topology of networks in order to identify nodes that have more impact on connectivity and therefore, may have more biological relevance. These in turn can be classified according to the type of centrality measures they use. The method developed by **Shunyao Wu and colleagues** (Wu et al. 2015) exploit global topology analyzing how a particular flow propagates through network. It works under the assumption that DAGs are preferentially not well connected to essential proteins (Goh et al. 2007). Then, the method classifies and assigns weights to nodes according to if they are associated to disease genes (positive) or on the contrary are essential genes (negative weight). Eventually the algorithm performs a network propagation analysis to identify the most recurrent paths.

On the other side we find methods focused on exploiting local topological characteristics such as nodes' degree, length of shortest paths or nodes' betweenness centrality. **NERI method** (Simões

et al. 2015) for example constructs a heterogeneous network with PPI and gene expression data. Then, it prioritizes DAGs' neighbors according to their presence in shortest path and the existence of similar gene expression patterns to the observed for DAGs between control and disease contexts.

DAGs can also be used to assess similarities between diseases. These methods study disease-disease relationships at genotypic level representing diseases as nodes in a network linked by common DAGs or comorbidities. The first of its type was the "***human disease phenome***" constructed by **Goh and colleagues** (Goh et al. 2007). They used a systematic approach to represent the diseasome that links all the genetic disorders of known molecular basis.

There is however a scarcity of methods to explore the networks of DAGs of two phenotypically similar diseases. This exploration could identify common molecular mechanisms and expand in this way the current knowledge about the two diseases. The method constructed by **Calderone and coleagues** (Calderone et al. 2016) follows this approach and tries to identify functional commonalities between Alzheimer and Parkinson diseases. First, it extracts topological communities in PPI networks containing each disease DAGs. Then it performs Gene Ontology (G0) functional enrichment in each community. Finally, it compares the enriched functions detected in communities from the Alzheimer network with the ones found in communities from the Parkinson network. Although it is initially based on topological communities extraction, it is more focused on function similarity analysis and does not attempt to build a unique network that integrates DAGs from both diseases.

In this work **we propose a new method, called Double Specific Betweenness (S2B)** that builds an interaction network connecting DAGs from two diseases and prioritizes proteins that specifically link the network modules of both diseases.

S2B method relies on the assumptions that **1)** proteins involved in the same disease phenotype tend to interact with each other (disease modules) (Oti et al. 2006), **2)** disease modules are prone to overlap (Goh et al. 2007) and **3)** hub nodes are less likely related to phatophenotypes (Goh et al. 2007). To exploit these principles we implemented a variant of betweenness centrality specific for cross DAG shortest paths filtering out unspecific highly central nodes by randomization-based statistics.

The standard betweenness count computes the number of times a node is part of a shortest path between any two nodes in the network. In S2B, the count is measured using only shortest paths that link DAGs from one disease to DAGs from the other disease. In this way, it can estimate nodes' relevance connecting the pathomechanisms of both diseases. Conversely, knowing that nodes with high degree are more frequently found in any shortest path, we have implemented two scores to filter those nodes that may be relevant but are not specifically related to the two diseases under study.

As a study-case, **we applied the S2B method to study two Motor Neuron Diseases (MNDs)**, known to share phenotypic and genotypic properties. Both Amyotrophic Lateral Sclerosis (ALS) and Spinal Muscular Atrophy (SMA) are characterized by the Motor Neuron (MN) degeneration and successive muscular atrophy. Besides, several of their DAGs are involved in common cellular activities, being RNA metabolism an outstanding example. However, RNA homeostasis is essential for the survival of any type of cell. Thus we are still missing the key elements that eventually induce the particular motor neuron death. This suggests there are more disease-associated functions non-related to RNA processing that may have relevant roles in MN degeneration.

Therefore, ALS and SMA are ideal study subjects on which to apply S2B method and evaluate the performance of the implemented algorithms. Additionally, it will allow us to obtain further knowledge about MN degeneration mechanisms.

Methodology

S2B method

MND-focused network construction

We retrieved all the ALS and SMA-DAGs described on OMIM (Hamosh et al. 2002) (http://www.omim.org/) and DisGeNET (Pinero et al. 2015) (http://www.disgenet.org/) databases (MND DAGs sets) in September 2015 (Figure 3.1–1). DisGeNET collects not only manually curated DAGs from experimental evidences but also predicted associations and DAGs described in other animal models (*Mus musculus* and *Ratus norvegicus*). We took all DAG data without any score filtering to retrieve the maximum amount of information. OMIM gathers all the information about genotypic-phenotypic relationships for all the disease subtypes together under the classical disease names.

However, DisGeNET makes distinctions among the disease subtypes so we retrieved from MeSH Browser (https://www.nlm.nih.gov/mesh/MBrowser.html) all the UMLS (Unified Medical Language System) CUI (Concept Unique Identifier) identifiers corresponding to ALS and SMA disease subtypes and used them to query DisGeNET. DAGs related with disease subtypes were joined in the corresponding ALS and SMA general sets. We used PPI datasets (Rolland et al. 2014; Rual et al. 2005; Venkatesan et al. 2009; Yu et al. 2011) available in CCSB-Human binary Interactome as a model of the human interactome network (Figure 3.1–1). We manipulated the interactome network using the Igraph R-package (Csárdi & Nepusz 2006). Particularly, we eliminated loop (when a protein interacts with itself) and multiple (when there are more than one interaction describes for the same pair of proteins, only one is retained) edges and selected only the network's main component. Finally, DAGs were labeled as ALS and/or SMA seed nodes. We supply the disease subtypes CUI list and the used PPI network as supplementary data S-3.1

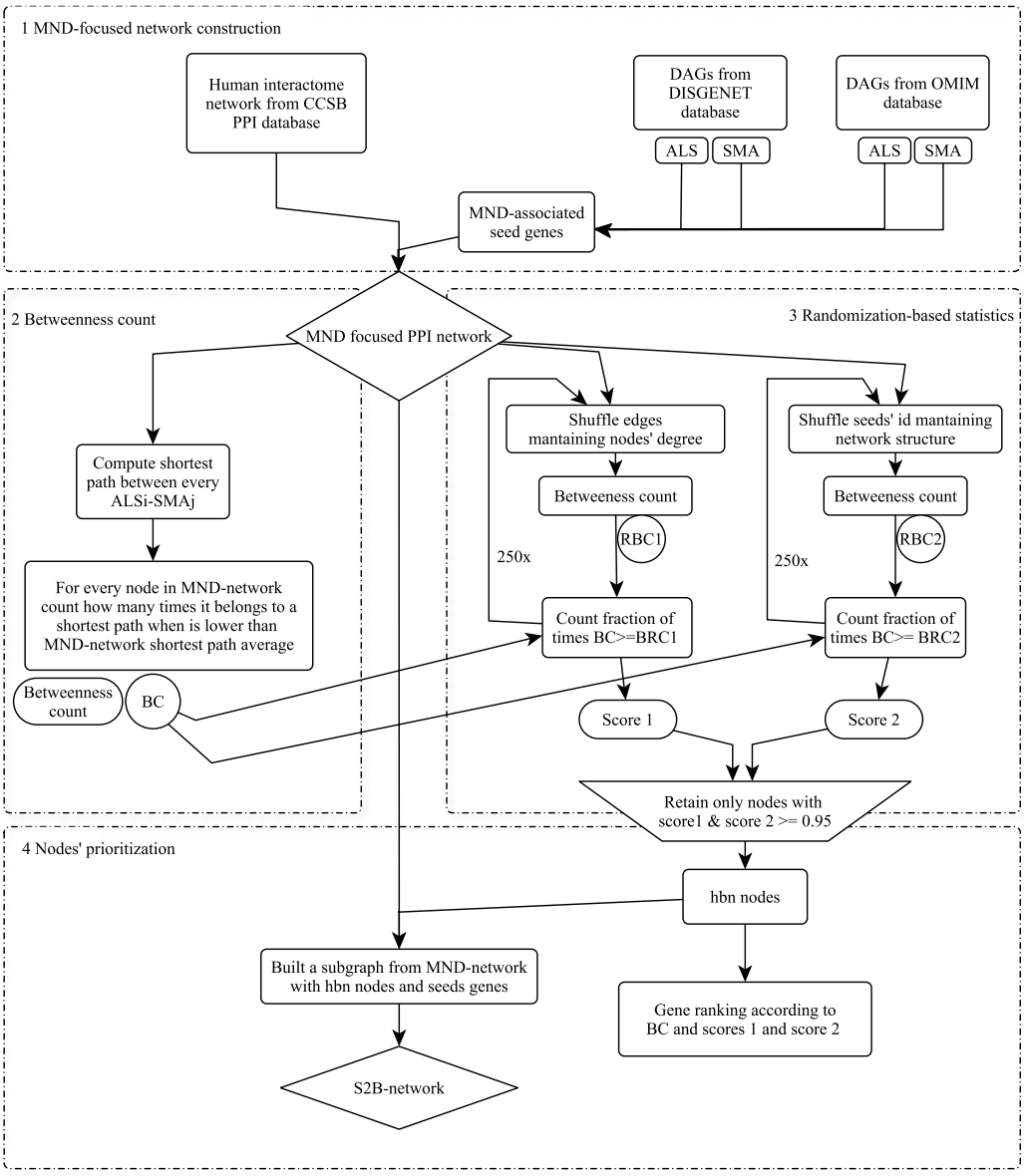

Figure 3.1 Flowchart describing the workflow of S2B method. MND-focused network is constructed using CCSB derived PPI data and MND-DAGs retrieved from DisGeNET and OMIM (step 1). Following S2B method is applied. Betweenness count is measured (step 2) and network is shuffled using BRC1 and BRC2 algorithms (step 3) to obtain the statistical relevance needed to return the protein ranking and construct a S2B network using the prioritized nodes (step 4).

Betweenness count (BC)

We searched the network for the shortest paths linking every possible pair of ALS and SMA seed nodes. In cases where there were multiple shortest paths between two seed nodes, all of them were considered. Whereas shortest paths that had a higher length than the whole network average shortest path length were discarded. For each node in the network we counted the number of times it was part of the shortest paths between ALS and SMA seed nodes. Eventually, we did not count the presence of the actual seed nodes (initial and terminal DAGs nodes) of each shortest path. The resulting count was called the **Betweenness count (BC)** (Figure 3.1-2).

Randomization-based statistics (BRC)

Following, we implemented two algorithms to filter out nodes with unspecific high BC (Figure 3.1-3). Both algorithms generate random networks that are used to compute **betweenness random counts (BRC)** for every node. As for BC, BRCs are computed from the shortest paths between each ALS and SMA seed but according to the new random network. In the first algorithm, network edges are shuffled maintaining their nodes' degree (Figure 3.2). We performed 250 randomizations and determined for each node the fraction of times that BC (computed with the original network) is higher than BRC1. This fraction was named **Score 1**. In the second algorithm, only seeds' identity is shuffled. We also performed 250 randomizations and for each node **Score 2** measures the fraction of times BC is higher than BRC2.
Score 1 and Score 2 values for the same nodes are not highly correlated. The first kind of randomization creates new shortest paths, but maintains the degree of the seed nodes. On the other hand, the second kind of randomization varies the degrees of seed nodes but maintains the network shortest paths, and consequently each node global betweenness. The R script of BC, BRC1 and BRC2 algorithms is available in supplementary data S-3.2

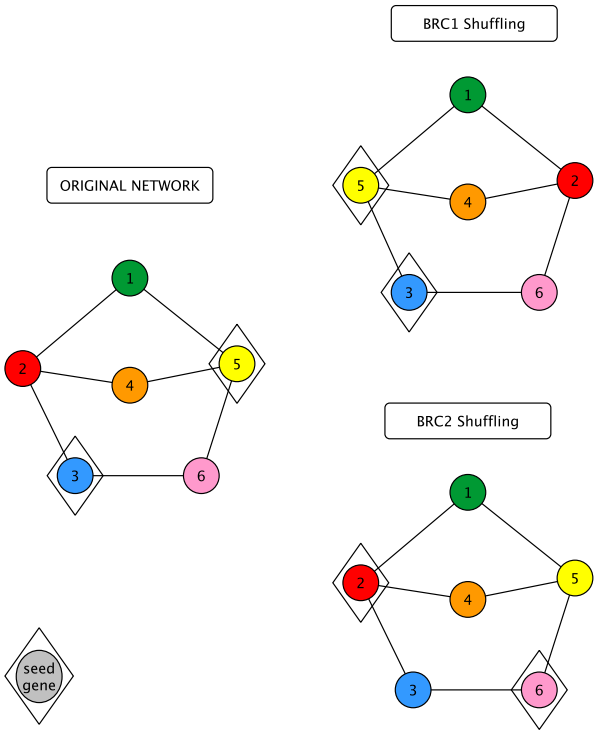

Figure 3.2 Illustration of BRC1 and BRC2 shuffling procedure. Comparing to the original network in the left, the top right network nodes change their edges but maintain their degree (BRC1 shuffling) while in bottom right network, seed nodes (encircled with a diamond) change but the network structure does not change (BRC2 shuffling).

Node prioritization

We selected the nodes that were present at least in one shortest path (BC higher or equal to 1) and showed in both scores values higher or equal to 0.95. Selected proteins were called *hbn nodes* (high betweenness nodes). Finally, we constructed the *S2B-network* extracting the induced subgraph of hbn nodes and MND DAGs from the interactome network (Figure 3.1-4) Besides this network, S2B method returns a list with all nodes' betweenness count (BC), score 1 (BRC1) and score 2 (BRC2) values. The results obtained from S2B method for MND-focused network are available in supplementary data S-3.2.

Functional enrichment comparisons

To evaluate the results obtained with the S2B method, we compared the biological processes enriched in the set of hbn nodes with the ones enriched 1) simultaneously in the initial sets of ALS and SMA DAGs and 2) in a manually curated PPI network built around four proteins (FUS, TDP43, SMN and SETX) known to be associated with ALS and SMA (*FTSS-focused network*). To facilitate these comparisons we developed a common workflow that removes the redundancy in enriched Gene Ontology Terms (GOTs) clustering them by annotated gene sets co-occurrence and by semantic similarity (Figure 3.3).

Functional enrichment
The comparison workflow receives as inputs two gene sets A and B (Figure 3.3-1). Both gene sets are functionally enriched using EnrichGO R function available in ClusterProfiler R-package (Yu et al. 2012). The functional enrichment is constrained to Biological Processes (BP) GOTs described in *Homo sapiens*. An FDR adjusted p-value smaller than 0.05 is considered statistically significant (Figure 3.3-2).

GOT filter by specificity
The initial GOT sets are depurated leaving out the terms that show a background frequency equal or higher to 10% (Figure 3.3-3). A **GOT background frequency** is the percentage of genes annotated with that GOT in the background list used in the enrichment analysis (*Homo sapiens* genome in our case). The resulting pair of filtered GOT is merged into a single list and sorted according to GOT fold enrichment in decreasing order. **Fold enrichment** is the ratio between the frequency of the GOT in the gene list and the frequency of the same GOT in the background gene list.

GOT fusion by gene co-occurrence
GOTs in the merged list are analyzed according to their associated gene sets. Two GOTs are fused if the intersection of the associated gene sets is at least 75% of the smaller gene set. The new GOT groups maintain the official ID and descriptor of the GOT that shows higher fold enrichment (Figure 3.3-4). Each GOT is compared with all GOT with lower fold enrichment. Each time two terms are fused, the GOT with higher fold enrichment collects the other GOT and associated gene sets and the GOT with the lower fold enrichment is immediately deleted from the list. In this way, each GOT can only be fused once with a GOT with higher fold enrichment. On the other hand, one GOT can fuse with multiple GOTs with lower fold enrichment.

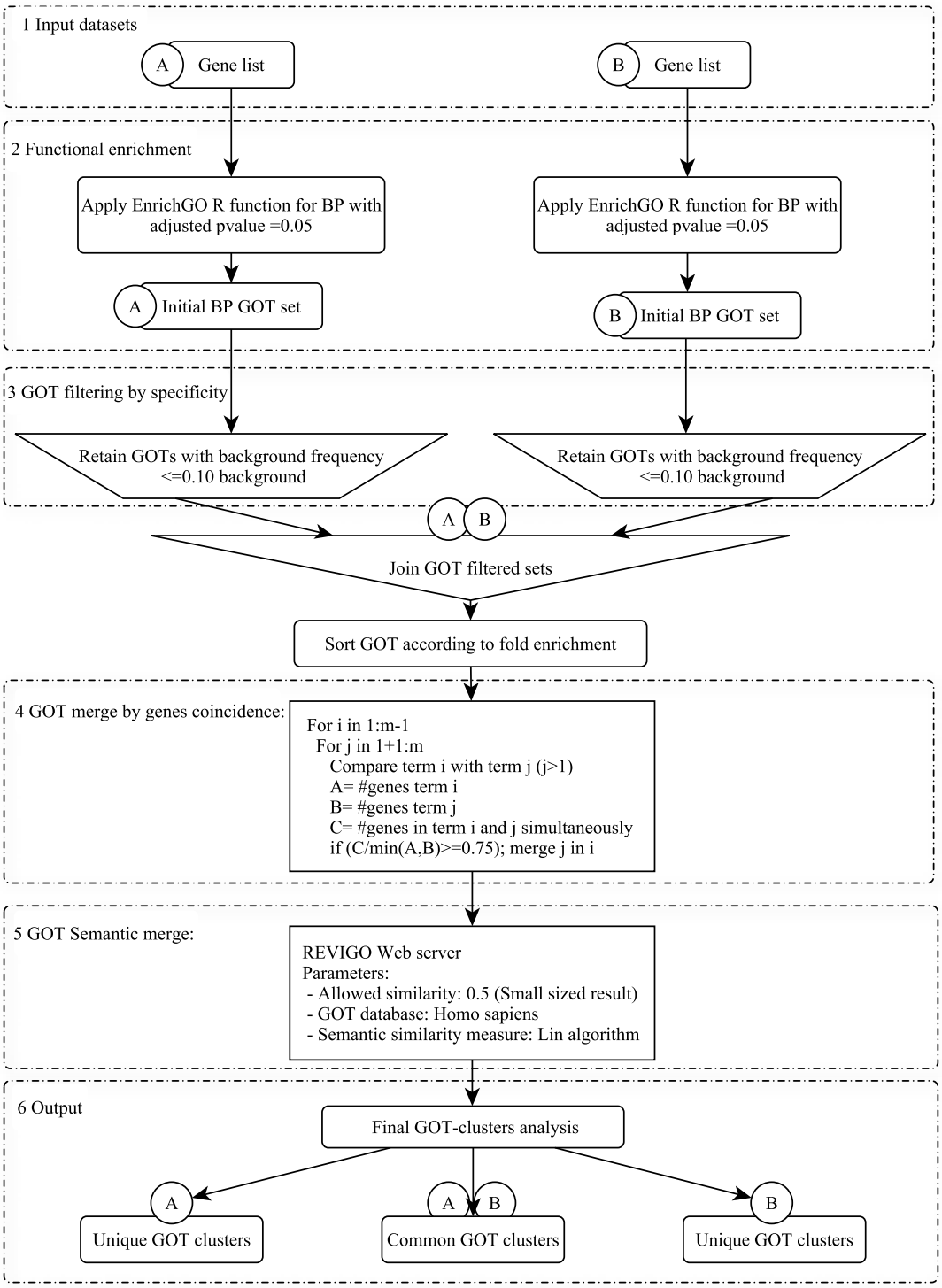

Figure 3.3 Flowchart describing the standard workflow of functional characterization on any pair of protein sets. Two proteins or gene sets (step 1) are functionally enriched (step 2). Then the GOTs are filtered by their specificity (step 3), fused by gene co-ocurrence (step 4) and semantic similarity (step 5), resulting in three final GOT sets: unique GOTs associated to gene set A, unique GOTs associated to gene set B and common GOTs associated to both gene sets (step 6).

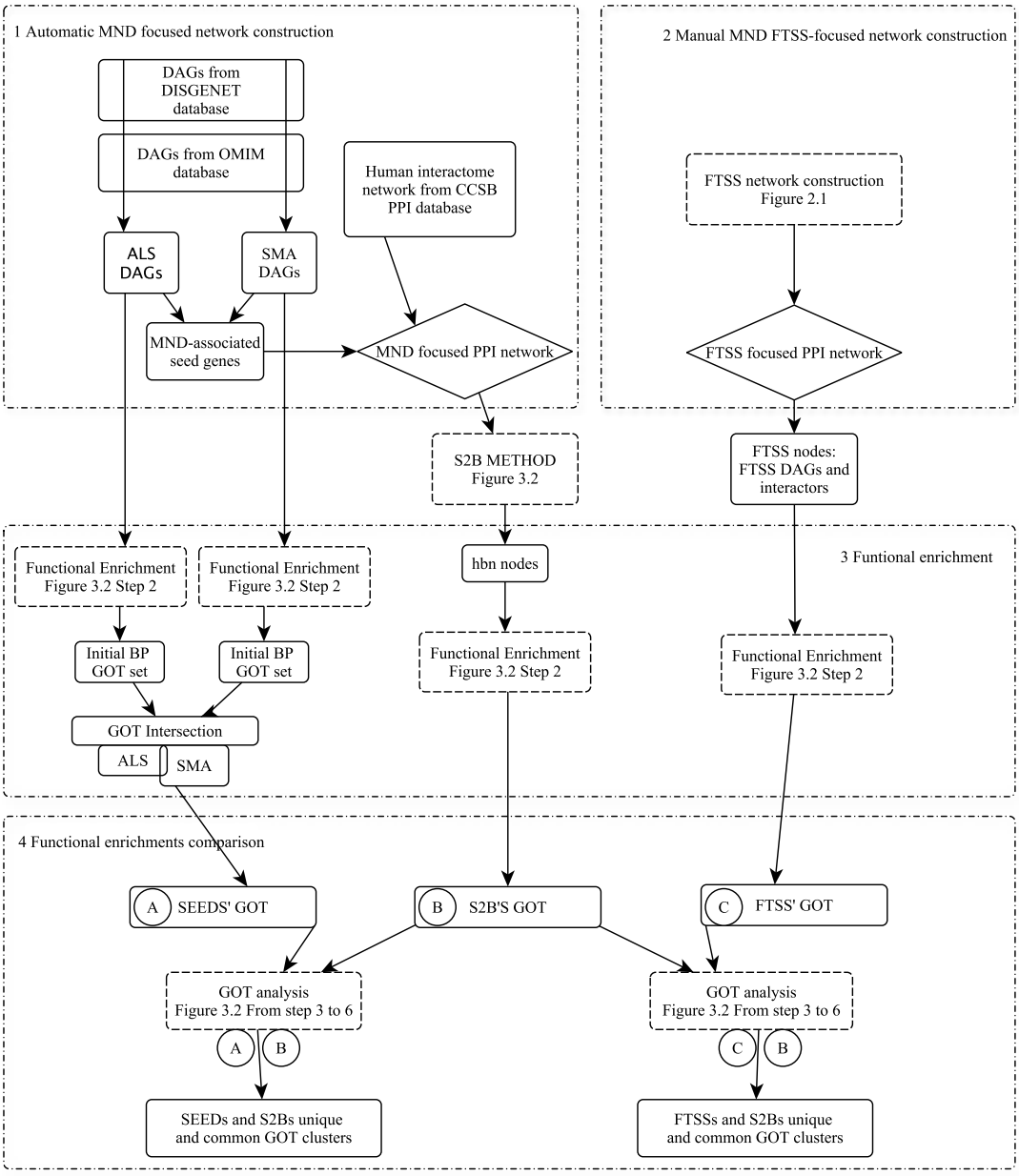

Figure 3.4 Comparison of S2B results with FTSS-focused network and MND-DAGs. Firstly, S2B and FTSS networks were constructed (Steps 1 and 2). Then, S2B (B), FTSS (C) networks' proteins were functionally enriched. Conversely ALS and SMA DAGs were functionally enriched independently so that matching results (intersection) are joined to form the SEEDS functional result (A) (left middle part of figure). These merged GOT sets were compared SEEDS(A) vs S2B(B) and FTSS(C) vs S2B(B) (bottom box) to return the final GOT sets.

GOT fusion by semantic similarity

After fusion by gene co-occurrence, resulting GOTs are fused by semantic similarity using REVIGO (Supek et al. 2011) (http://revigo.irb.hr). The input set is conformed by single GOTs' ID and fold enrichment scores as a significance measure. The semantic similarity measurement is performed using Lin's algorithm (Lin 1998) in *Homo sapiens* GOT database. GOTs are fused with a minimum allowed similarity of 0.5, which returns a small sized result (Figure 3.3-5).

Functional clusters comparison
The fused GOTs returned by REVIGO are separated according to their provenience which results in three final GOT sets: GOTs uniquely enriched in gene set A, GOTs uniquely enriched in gene set B and those GOTs commonly enriched in both sets (Figure 3.3-6).

Comparison of S2B results with FTSS-focused network and MND-DAGs

In order to assess S2B method's success retrieving biologically relevant information we performed two functional enrichment comparison analyses using as controls MND-DAGs (SEEDS) and the FTSS-focused network proteins (Figure 3.4).

With the first comparison we asked if the S2B method retrieves further information than the resulting from the functional enrichment of isolated DAGs (SEEDS). Moreover, we also intended to verify that S2B retrieves biologically significant information. Therefore, we used the FTSS-focused network as a positive control under the premise that it was constructed semi-automatically using previously MND-knowledge.

Results and discussion

MND-focused network

We can see on Table 3.1 that ALS has more related DAGs than SMA (290 against 96 respectively). This imbalance was expected because unlike SMA, ALS is considered a heterogenic disease. DisGeNET holds more DAGs than OMIM for these two diseases. DisGeNET retrieves DAGs predicted from text mining and DAGs identified in animal models such as *Mus musculus* or *Ratus norvegicus* (Pinero et al. 2015), whereas OMIM collects manually curated data about human diseases, which leads to a smaller DAGs set (Hamosh et al. 2002). From 290 total ALS DAGs, only 43 were present in both databases and likewise from 96, 16 were redundant in SMA set (Table 3.1). However, not all DAGs have PPIs described in the CCSB network. This narrowed the final MND DAGs list (commonly referred in this work as *seed genes*) to 163 DAGs associated to ALS, 43 SMA-DAGs and 21 DAGs associated with both diseases (Table 3.1).

Table 3.1 Summary of MND DAGs retrieval.

	ALS	**SMA**	**Common**
DisGeNET	219	35	
OMIM	114	77	
DisGeNET and OMIM (union)	290	96	
DisGeNET and OMIM (intersection)	43	16	
MND-focused network	**163**	**43**	**21**

S2B method results' topological analysis

Once the MND-focused network was constructed we applied S2B method. It identified 211 **hbn proteins** that appeared specifically in shortest paths connecting ALS and SMA DAGs. We constructed a S2B subnetwork using hbn nodes and seed genes (Figure 3.5). All the data returned by S2B is available in supplementary data S-3.2 S2B method.

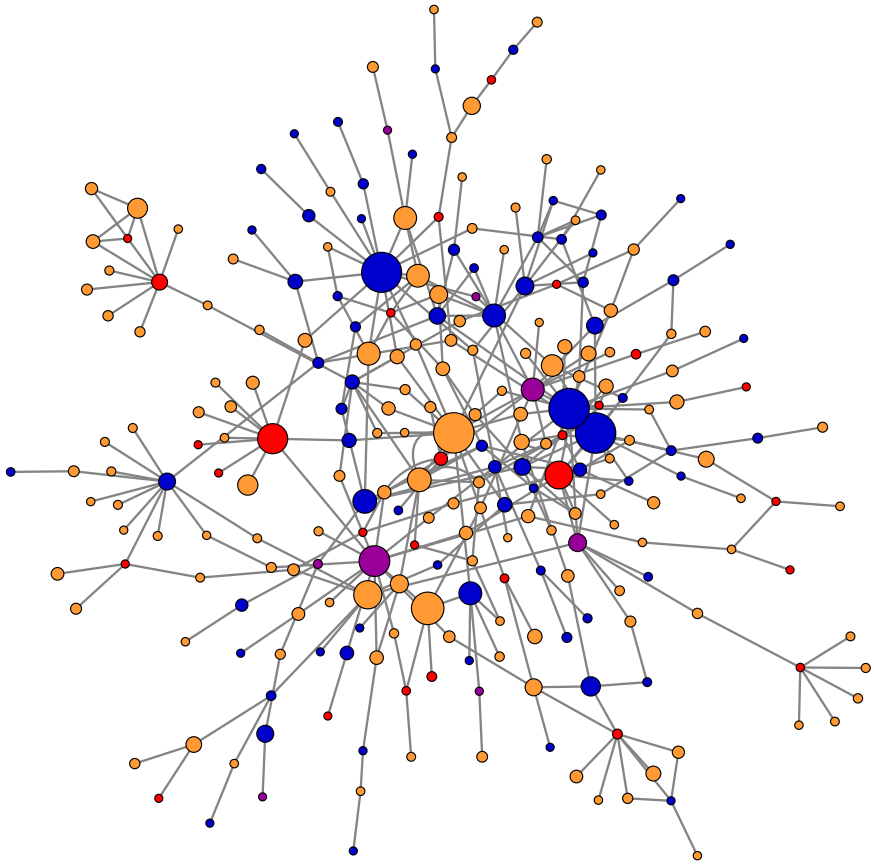

Figure 3.5 Subnetwork resulted from the S2B method prioritization. The network gathers all the hbn nodes whereas blue nodes are seed genes associated to ALS, red nodes are associated to SMA, purple nodes are associated to both ALS and SMA and eventually, orange nodes correspond to hbn nodes without previous disease-causative relation. Node sizes are proportional to S2B count (BC).

As can be seen in Figure 3.5, nodes with higher S2B counts are more central in S2B subnetwork, which may suggest a stronger relation with molecular mechanisms common to ALS and SMA. Therefore S2B count (BC) could be used to rank hbn proteins.

Before exploring the biological processes associated with hbn proteins, we asked if the S2B count and the selection of hbn proteins through specificity scores was not simply reflecting node centrality in the overall interactome network.

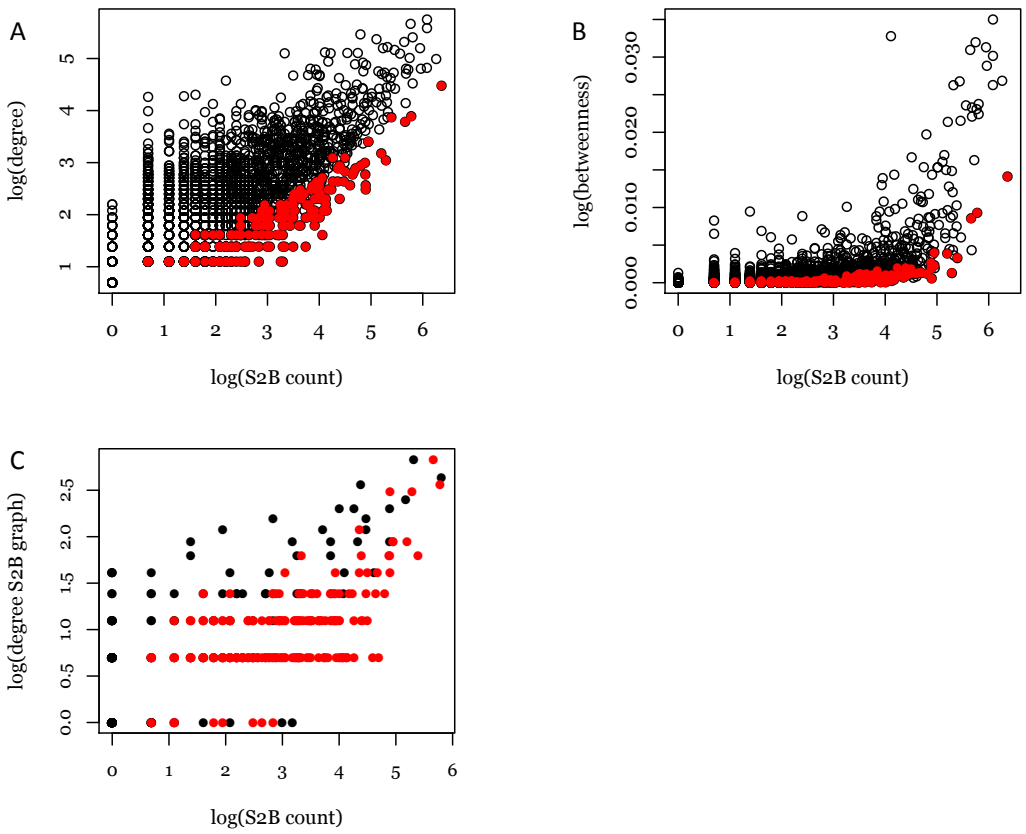

Figure 3.6 Centrality analyses of proteins prioritized by S2B method. We measured the degree, betweenness centrality and S2B count for all the nodes of the CCSB interactome network using Igraph R-package functions and S2B algorithm respectively. Each black empty dot represents a protein/node of the CCSB interactome whereas red filled dots identify those nodes that exceeded S2B thresholds (hbn nodes). **A)** Log-log plot comparing degree and S2B-betweenness count (S2B count) of all nodes in MND-focused network. **B)** Log-log plot comparing general betweenness and S2B count of all nodes in MND-focused network. **C)** Log-log plot comparing degree and S2B count of the nodes retained on the S2B subgraph. In this case, black dots represent those seed genes that did not exceed the S2B thresholds.

In broad terms we can observe a positive correlation between S2B count and degree (Figure 3.6A) and between S2B count and betweenness (Figure 3.6B), weaker for low S2B count values and stronger for higher S2B count values. The correlation between standard betweenness and S2B count is particularly weak for log(S2B count) between 0 and 5 (Figure 3.6B). This happens because S2B count only takes into consideration those shortest paths *from-to* seed nodes, excluding many nodes with high values of standard betweenness.

Nodes that were considered specific by the S2B method (hbn nodes) have generally lower degree and standard betweenness when compared with unspecific nodes. However, the fact that S2B method does not reject automatically all nodes with high degree (Figure 3.6A) nor with high betweenness scores (Figure 3.6B) demonstrates its ability to distinguish those general central nodes with actual MND linking nodes.

Finally, we analyzed the degree centrality within the S2B subnetwork formed by hbn nodes and all seed genes. As can be seen on Figure 3.6C, S2B counts show a wide variation for nodes with similar S2B network degree. Thus, S2B count has a higher discriminatory power, even in the S2B network context.

As expected, proteins with highest S2B count values have also the highest degrees in the S2B network, highlighting their high centrality in the context of ALS and SMA diseases. Moreover, we can observe that S2B method discards many seed genes with high S2B counts presumably because of their high degree.

Conversely, hbn nodes do not have to necessarily be very central in S2B network to be considered specific in the MND-context.

S2B method functional analysis

We performed GOT functional enrichment analysis with the set of hbn proteins selected by the S2B method. Moreover, we compared the enrichment analysis results with similar analysis made with the sets of MND-DAGs (SEEDS) and with the proteins of the FTSS-focused network. To make fair and non-redundant comparisons we applied an analysis workflow that fused GOT associated with many common genes or with a high semantic similarity. All the files resulting from each step are available in Supplementary data: S-3.3 Functional characterization raw results and the numeric summary is shown in figure 3.7.

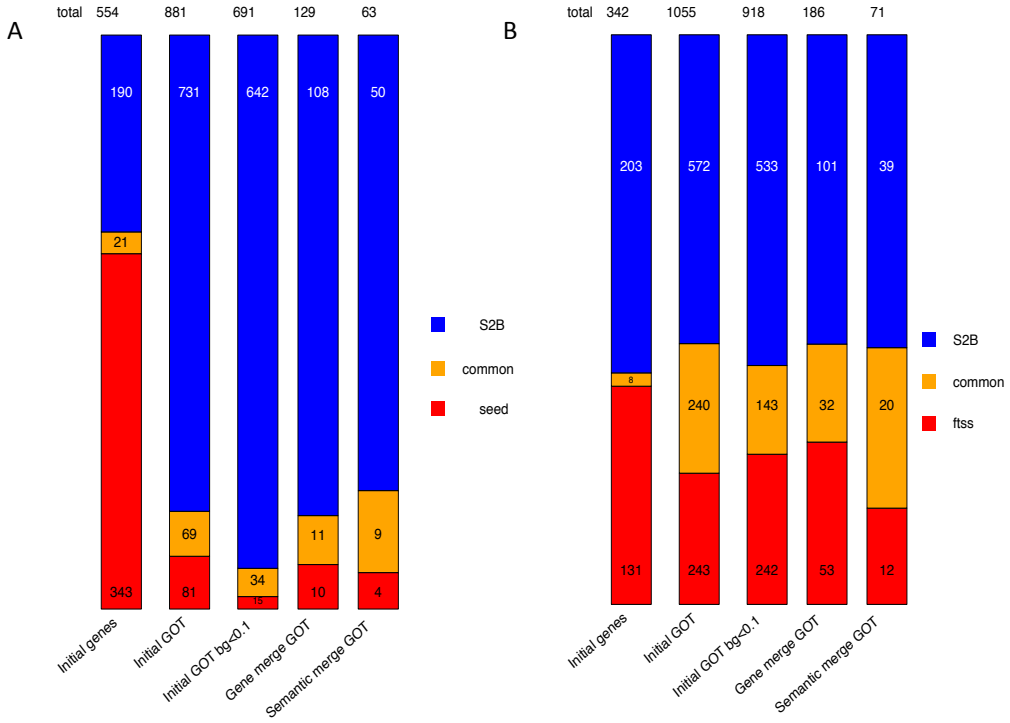

Figure 3.7 Descriptive summary of functional enrichment comparison between S2B and SEEDS (A) and S2B and FTSS (B) results. Each bar describes from left to right each functional filtering or fusion step. The initial set of genes are functionally enriched (initial GOT) and filtered by their specificity retaining only those GOT with a background frequency lower than 10% (initial GOT bg<0.1). Following GOT are fused when they show a gene co-occurrence higher than 75% (Gene merge GOT). Finally, GOTs are merged by semantic similarity using REVIGO resulting in the final GOT sets (semantic merge GOT). At the same time each bar is divided by 1) genes or GOTs uniquely related to S2B (blue), 2) associated to FTSS or SEEDS (in red), and 3) genes or terms common to both sets (in orange).

The first fact to take into consideration is that initial GOT set of SEEDS is constituted only by **GOT intersection between ALS and SMA seeds enrichment**. As such, although the SEEDS set being larger than the FTSS one (364 and 139), initial GOT set was smaller (150 versus 483).

During the GOT fusion process in both cases (Figure 3.7), there is a general trend towards the increase of S2B-related and common GOT sets. Looking to the progression of bars in Figure 3.7B, the vast majority of FTSS-related initial GOTs could be considered specific because with the exception of one term, they show a background frequency smaller than 10%. However these are considerably redundant, because when filtering by gene co-occurrence 78% of terms are lost. At least, the final subset resulting from semantic merge contained 12 functions uniquely related to FTSS.
On the contrary, in Figure 3.7A, we see that SEEDS-related initial GOT set is constituted by more general functions but less redundant. However, SEEDS final subset is even smaller than the respective FTSS-subset.
Conversely, both S2B-related GOT subsets (Figure 3.7A and B) are constituted by specific but redundant GOTs loosing only 12% and 7% when background threshold is applied while 93% of GOT terms are lost after gene co-occurrence and semantic fusions. Common FTSS-S2B and SEEDS-S2B GOT sets gather GOTs with higher background frequencies and gene co-occurrences, thus more general and redundant functions.
Most strikingly, S2B specific GOT are the majority in both comparisons (Figure 3.7A and B), which suggests that the S2B method is able to retrieve many novel biological processes linking these two diseases. It is also reassuring to find a high fraction of GOT in common between the S2B results and the FTSS-focused network, since the latter provided a sample of true positive findings.

Functional comparison between SEEDS and S2B results

The comparison of functional enrichment between S2B results and SEEDS set resulted in 4 GOT clusters specific for SEEDS (Figure 3.8), 9 clusters related to both SEEDS and S2B proteins (Figure 3.9) and 50 clusters only associated to S2B protein set (Figure 3.10).

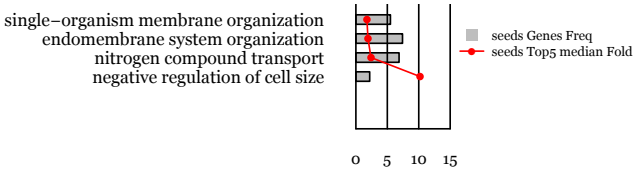

Figure 3.8 Overview unique functional clusters found in S2B comparing to SEED genes' results. This bar plot only describes those functions uniquely associated to SEEDS genes set functional enrichment results. At the same time, each function corresponds to a functional group constituted by the GOTs merged by gene co-occurrence and or semantic similarity and receives the descriptor of the GOT with highest fold enrichment. Bars represent the GOT associated gene frequency within the SEEDS subset. Red lined-points represent the fold enrichment average of (up to) the 5 highest fold enrichment scores' among the GOTs merged in the respective function cluster.

The functions uniquely associated to SEEDS are not very specific (Figure 3.8), which prevents the formulation of clear hypothesis for disease molecular mechanisms.

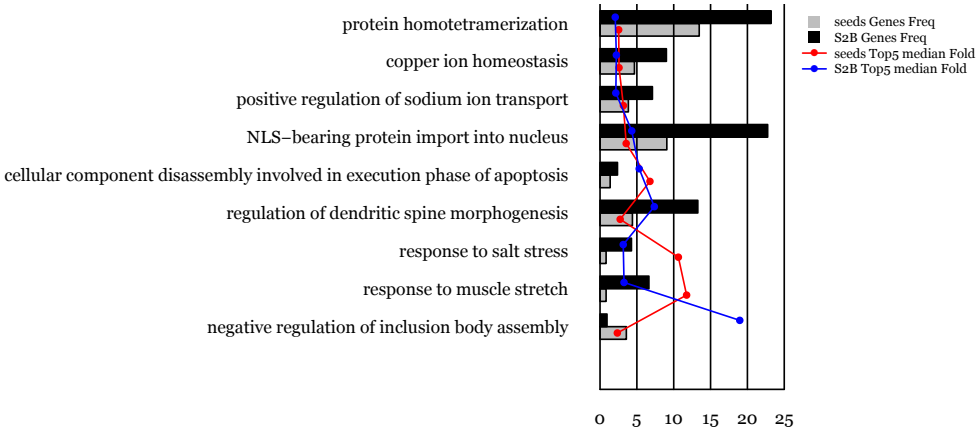

Figure 3.9 Overview of common functional clusters between S2B and SEED genes. This bar plot describes those functions common to S2B and SEEDS genes sets functional enrichment results. At the same time, each function corresponds to a functional group constituted by the GOTs merged by gene co-occurrence and or semantic similarity and receives the descriptor of the GOT with highest fold enrichment. Bars represents relative gene frequency (dark grey for S2B and light grey for SEEDS). Lined-points represent the fold enrichment average of (up to) the 5 highest fold enrichment scores' among the GOTs merged in the respective functional cluster (blue for S2B and red for SEEDS).

Among the GOT clusters that appear in S2B and SEEDS results simultaneously, we can find processes more easily connected with the studied pathologies, such as "regulation of dendritic spine morphogenesis" or apoptosis related processes (Figure 3.9).

In the functions uniquely associated to S2B proteins' set (Figure 3.10) there are also many functions related to neuromuscular functions such as, "skeletal myofibril assembly", "brain morphogenesis "or "membrane repolarization during action potential" which already demonstrates that S2B method prioritizes biologically consistent information.

snoRNAs (small nucleolar RNAs) 3' end processing is a function with simultaneously high gene frequency and high fold enrichment. snoRNAs are known to guide other RNAs' biochemical modifications and thus regulate tRNA, rRNA or snRNAs functionality (Dragon et al. 2006). These snoRNAs are synthetized by RNA pol II thus, their biogenesis could be affected by the lack of activity of RNA pol II-binding proteins such us FUS or SETX (Jorjani et al. 2016). Besides RNA stabilization also shows up in this GOT cluster, a key event that could be also affected by snoRNAs perturbation.

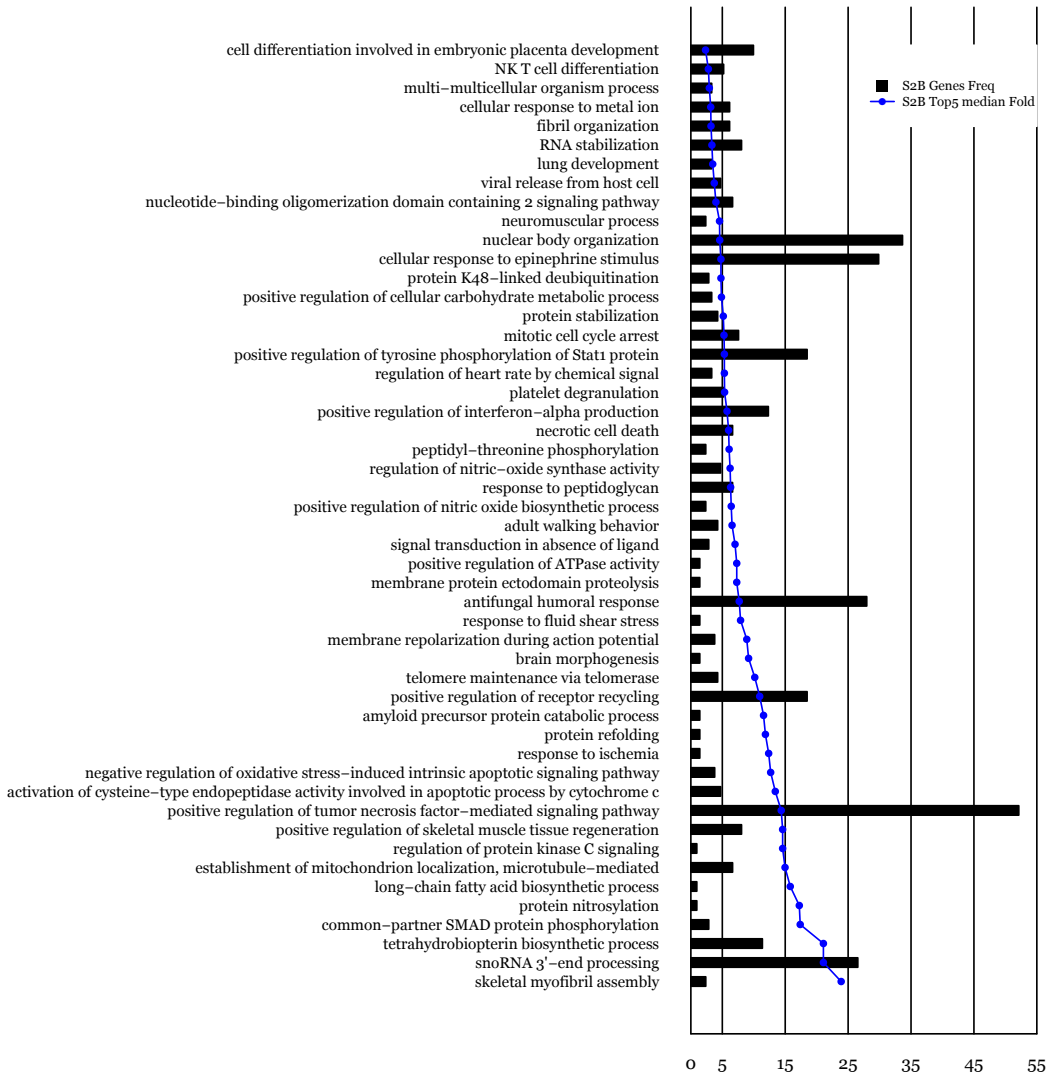

Figure 3.10 Overview unique functional clusters found in S2B comparing to SEED genes' results. This bar plot only describes functions uniquely associated to S2B genes set functional enrichment results. Each bar corresponds to a functional group constituted by the GOTs merged by gene co-occurrence and or semantic similarity and receives the descriptor of the GOT with highest fold enrichment. Bar length represents relative gene frequency. Blue lined-points represent the fold enrichment average of (up to) the 5 highest fold enrichment scores' among the GOTs merged in the respective function cluster.

Surprisingly, there is a great number of functions related to inflammation and host-pathogen responses such as "tumor necrosis factor-mediated signaling", "anti-fungical humoral response", "regulation of interferon" and "NT K cell differentiation". These could be directly related to the immune *hyper-sensitivity* caused by perturbations on SETX activity (Miller et al. 2015). Likewise, it is known that prolonged inflammatory responses lead to neurodegeneration (Amor et al. 2010). Other well-known neurodegeneration inducer is the oxidative stress (Friedman 2011). It can be produced for example by isquemic events or nitric oxide overloads (Friedman 2011; Xiong et al. 2007), two events described in S2B specific GOT clusters (Figure 3.10). The oxidative stress usually induces the protein aggregation and extracellular accumulation, apoptosis or necrotic cell death (Carrì et al. 2015), four processes identified by the S2B method (Figure 3.10).

Interestingly, oxidative stress and the subsequent perturbations are directly related to ALS-related gene SOD1, responsible for the elimination of free superoxide radicals (Rosen et al. 1993). Furthermore, when genes involved in damaged DNA repair such as FUS or SETX are mutated, genome aberrations are accumulated and eventually boost oxidative stress (Lagier-Tourenne et al. 2010; Skourti-Stathaki et al. 2011)

Functional comparison between FTSS-network and S2B results

We also compared functional enrichment results obtained with the FTSS-focused network proteins (FTSS) with the ones obtained with S2B method selected proteins (S2B). It resulted in 12 GOT clusters specific for FTSS (Figure 3.11), 20 clusters related to both FTSS and S2B proteins (Figure 3.12) and 39 clusters only associated to S2B protein set (Figure 3.13).

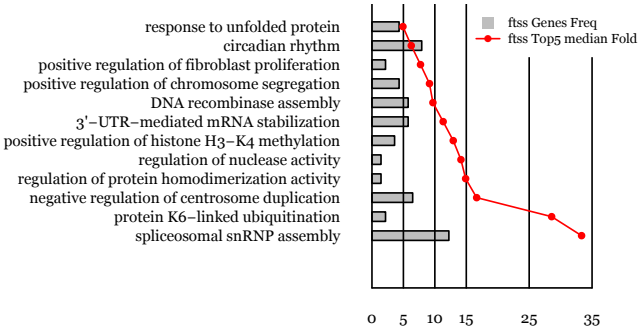

Figure 3.11 Overview unique functional clusters found in S2B comparing to FTSS-network genes' results. This bar plot only describes functions uniquely associated to FTSS genes set functional enrichment results. Each function corresponds to a functional group constituted by the GOTs merged by gene co-occurrence and or semantic similarity and receives the descriptor of the GOT with highest fold enrichment. Bars represent relative gene frequency. Red lined-points represent the fold enrichment average of (up to) 5 highest fold enrichment scores' among the GOTs merged in the respective function cluster.

Unlike SEEDS four unique functions (Figure 3.8), FTSS set specifically gathers more functions with more information content (Figure 3.11). Most are DNA or RNA related functions. Among them, spliceosome assembly is the most over-represented showing the highest gene frequency and fold enrichment. It also identifies histones H3 and H4 methylation regulation and 3' UTR mediated mRNA stabilization.

Due to the fact that FTSS network was constructed using known MND causative genes, we can assume that great part of the returned functions are closely related to this pathomechanisms and likewise, those functions common between S2B and FTSS may certainly be related to MN degeneration.
Surprisingly, despite of the low number of common genes between FTSS and S2B subsets (Figure 3.7), there are more common functions between FTSS and S2B (Figure 3.12) than between SEEDS and S2B (Figure 3.9). This fact shows that even using different type of DAG sets, network-based approaches can reach similar information and therefore more robust biologically insights.

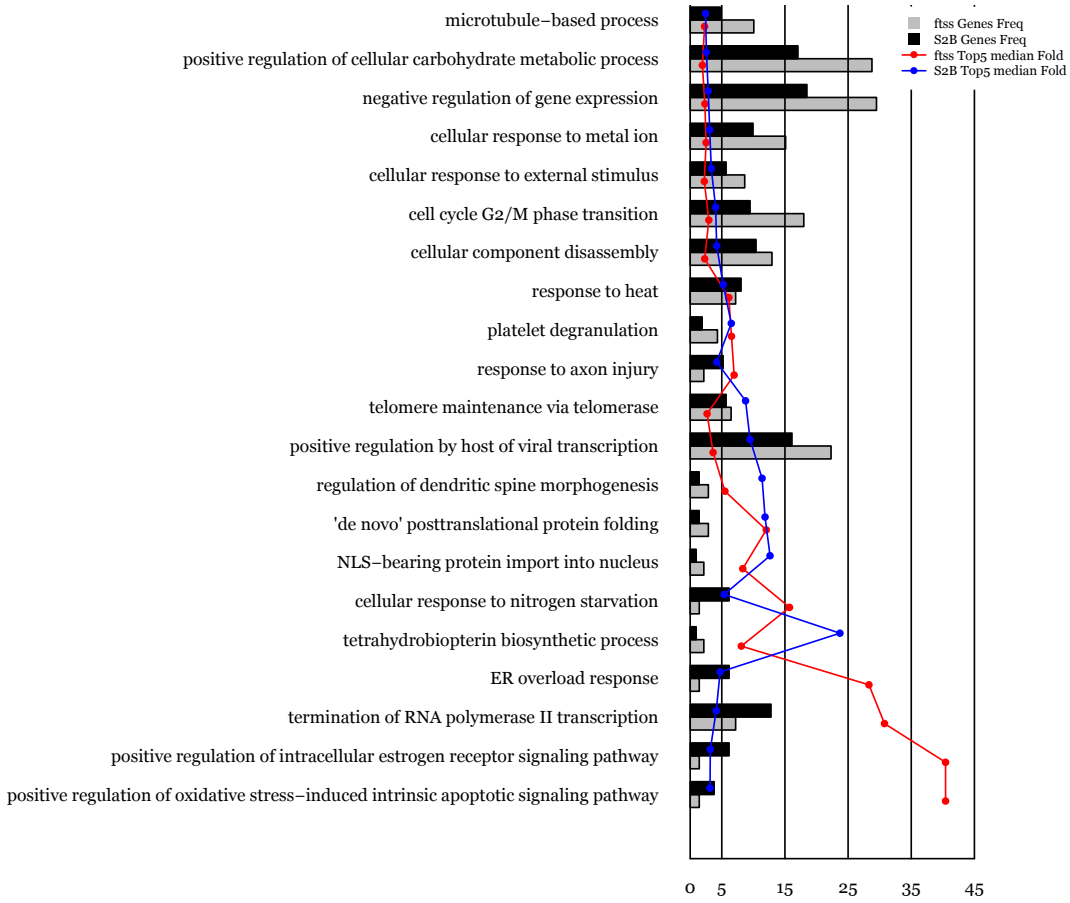

Figure 3.12 Overview of common functional clusters between S2B and FTSS -network genes. This bar plot describes functions common to S2B and FTSS genes sets functional enrichment results. Each function corresponds to a functional group constituted by the GOTs merged by gene co-occurrence and or semantic similarity and receives the descriptor of the GOT with highest fold enrichment. Bars represent relative gene frequency (dark grey for S2B and light grey for FTSS). Lined-points represent the fold enrichment average of (up to) 5 highest fold enrichment scores' among the GOTs merged in the respective function cluster (blue for S2B and red for FTSS).

Amongst these 20 common functions (Figure 3.12) it is noteworthy that neuron related functions appears again and now in more abundance. Among them we can highlight as a novelty tetrahydrobiopterin biosynthetic process. This compound is a critical cofactor for the biosynthesis of serotonin, melatonin, dopamine or nitric oxide among others (Thöny et al. 2000). This can possibly establish a link to oxidative stress and neurodegeneration (Xiong et al. 2007).
Moreover there are also over represented RNA pol II-mediated transcription termination, host viral transcription and microtubule based processes (Figure 3.12). Functions previously highlighted in FTSS-focused network functional enrichment (Chapter 2).

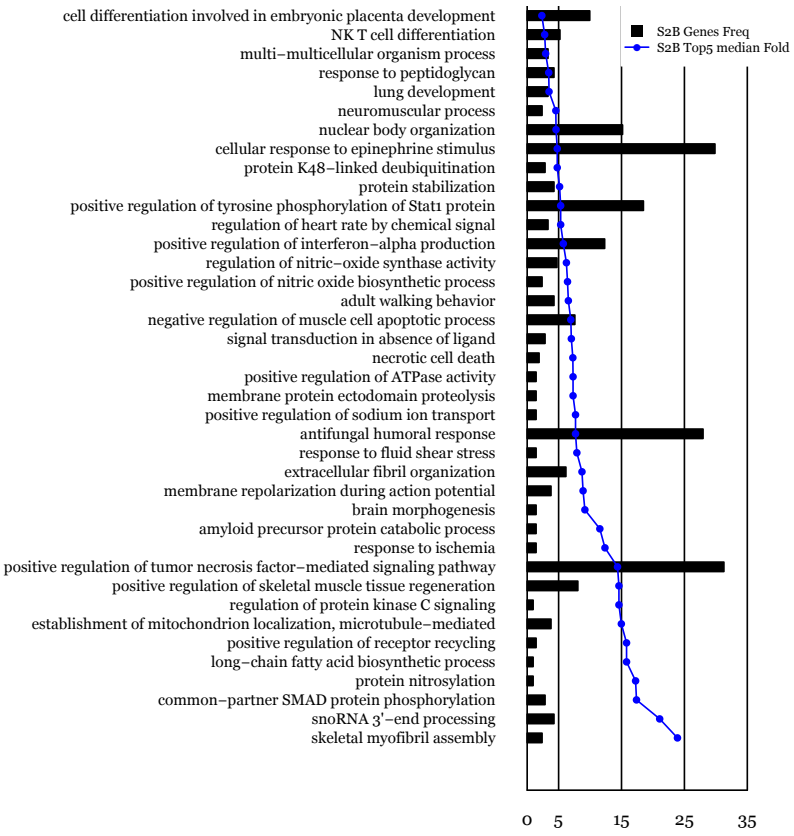

Figure 3.13 Overview unique functional clusters found in S2B comparing to FTSS-network genes' results. This barplot describes functions uniquely associated to S2B genes set functional enrichment results. Each function corresponds to a functional group constituted by the GOTs merged by gene co-ocurrence and or semantic similarity and receives the descriptor of the GOT with highest fold enrichment. Bars represent relative gene frequency. Blue lined-points represent the fold enrichment average of (up to) 5 highest fold enrichment scores' among the GOTs merged in the respective function cluster.

The 39 S2B unique functions found when compared with FTSS (Figure 3.13) are, as expected, very similar to the set of 50 functions uniquely enriched in the S2B set when compared with the SEEDS set (Figure 3.10). Globally, these functions demonstrate the capacity of the S2B method to uncover new processes that potentially link the molecular mechanisms involved in ALS and SMA.

Conclusions

Topological analysis of S2B results has shown that it effectively identifies central nodes within the MND-focused network. It prioritizes nodes specifically linking ALS and SMA and besides filters general hub nodes, even if they are MND-associated. This discriminative power is essential to discover new DAG candidates and furthermore avoid trivial data that could overshadow the biological inference.

Looking to the commonalities observed between FTSS and S2B functional sets, we confirm that S2B method not only retrieves biologically relevant information but also specific functions with MND-related congruence. Furthermore, it is pleasingly surprising that even with very low DAG overlap, FTSS and S2B approaches obtain similar conclusions. This fact together with the poor performance of SEEDS functional characterization reinforces our conviction of the usefulness of network-based approaches towards the study of complex diseases.

This goes along the observation done by Calderone and colleagues, claiming that topological commonalities imply the existence of similar biological processes and not vice-versa (Calderone et al. 2016). Thus, even though S2B and Calderone's methods share the goal of finding similarities between diseases, Calderone's method retrieves distinct functional communities for both diseases and thus cannot extract the common molecular pathomechanisms linking a pair of diseases.

On the other hand, NERI method has a very similar approach to S2B but seeks a different goal. It tries to identify disease modules of a single disorder exploiting the "*guilt by association"* concept. S2B method instead, proposes a new betweenness count specifically constructed to prioritize nodes linking two diseases.

S2B, as the method constructed by Shunyao Wu and colleagues, works under the assumption that essential genes are less prone to appear involved in a particular disease. However, it is also focused on the identification of disease pathways of a single disorder exploiting in this case global topological measures that can return very dissimilar results.

Calderone's method and NERI use DAGs retrieved from specialized databases for the particular diseases studied. This could be troublesome considering that the majority of diseases lack such databases. We instead took all the available DAG data from OMIM (Hamosh et al. 2002)and DisGeNET, which we believe could also be feasible for many other disease examples. PPI data sources can also introduce errors, especially when using predicted PPIs, as in the case of Shunyao Wu and colleagues' method that uses STRING. Besides, it is known that PPI databases have an intrinsic bias toward proteins with biomedical interest. In our case, we selected the CCSB human interactome database due to its non-biased approach complemented with high confidence literature curated PPI source.

As for S2B method functional results, neuromuscular processes appear both in common SEEDS-S2B and FTSS-S2B subsets maintaining naturally a key position in MNDs. We have also identified interesting new functions possibly related with MN degeneration. Host pathogen responses, inflammation or nitric oxide overload arise in unique S2B sets as potential MN degeneration inducers. These functions are highly related to MND causative genes such as SOD1 or SETX and besides are well-known neurodegeneration inducers.

Likewise, knowing the critical role of FTSS proteins in MND, functions specifically described in FTSS-S2B common set gain great relevance because they are also associated to S2B prioritized proteins. Among the most relevant, we find host pathogen responses (again), RNA pol II and microtubule-based functions. Once we have confirmed the S2B method ability to recapitulate most of the FTSS functional context, it is reasonable to presume that those functions found in FTSS but not by S2B could be molecular events more remotely related to MND pathogenesis.

Proteins generally interact with a high number of other macromolecules and thus may have multiple functions. This is one of the reasons why diseases are complex events in which such a high number of elements take part. However, this simultaneously means that each function of a given protein can show different grades of relevance in the disease context. This could be the case of spliceosome assembly and histones H3 and H4 modification regulation processes. Both functions are associated to SMN (MND-causing gene) and are critical for cell survival and thus it is difficult to explain how their perturbation could only lead to MN degeneration.

Conversely, the role of microtubules on axonal transport, together with the involvement of SMN-mediated neuron-specific alternative splicing on cytoskeleton dynamism seems to better match with the particular phenotype of MND diseases.

On balance, we have confirmed that S2B method returns different information to the expected from standard centrality measures such as degree or betweenness. Additionally S2B method was useful to confirm the relevance of previously described processes and supplement new hypotheses about the functions of FTSS genes and proteins prioritized by S2B method in MN degeneration.

Chapter 4. General discussion and conclusions

Amyotrophic Lateral Sclerosis (ALS) and Spinal Muscular Atrophy (SMA) are both Motor Neuron Diseases (MND) and thus show similar phenotypic characteristics. This enabled researchers to hypothesize that common molecular mechanisms are shared between these two diseases.

FTSS-focused network construction was useful to confirm the tight interatomic relationship among FUS, TDP43, SMN and SETX (FTSS) proteins. These proteins are directly associated to ALS and SMA (Chen et al. 2004; Rutherford et al. 2008; Kwiatkowski Jr et al. 2009; Lefebvre et al. 1995) and besides are physically interacting (Sun et al. 2015; Tsuiji et al. 2013; Yamazaki et al. 2012; Skourti-Stathaki et al. 2011; Zhao et al. 2016; Suraweera et al. 2009; Bennett & La Spada 2015) what provides new insights into the putative common pathomechanisms in MNDs. Besides, the functional characterization of FTSS-network has made possible the identification of functions closely related to FTSS proteins and thus possibly related to Motor Neuron (MN) degeneration. Among them we could highlight spliceosome assembly, microtubule-based movement and DNA repair processes.

According to the topological analysis of S2B method results, the method identified nodes with high betweenness between Disease Associated Gene (DAG) seeds and additionally rejected those non-specific hubs (nodes with high degree). Therefore, we can state that it is an efficient method to prioritize nodes linking two disease modules within a Protein-Protein Interaction (PPI) network.

Noteworthy, although MND-focused network was constructed using CCSB human interactome network (Rolland et al. 2014) and DAGs from DisGeNET (Pinero et al. 2015) and OMIM (Hamosh et al. 2002) databases, S2B is a flexible method that can be used with diverse data sources. We consider that CCSB database was an appropriate PPI source because it was constructed using a proteome scale mapping procedure enriched with curated literature knowledge. This provides one of the less biased PPI interactome networks available.

Moreover, the functional characterization of S2B selected proteins has retrieved many functions with biological congruence and therefore possibly causal of MN degeneration. Thus, we can also affirm that S2B method is useful to prioritize proteins linking ALS and SMA disease. Additionally, the comparison of functional results obtained with the analysis of MND-DAGs alone (SEEDS), has proved that network-based methods (FTSS and S2B) are able to extract much more information related to complex pathomechanisms.

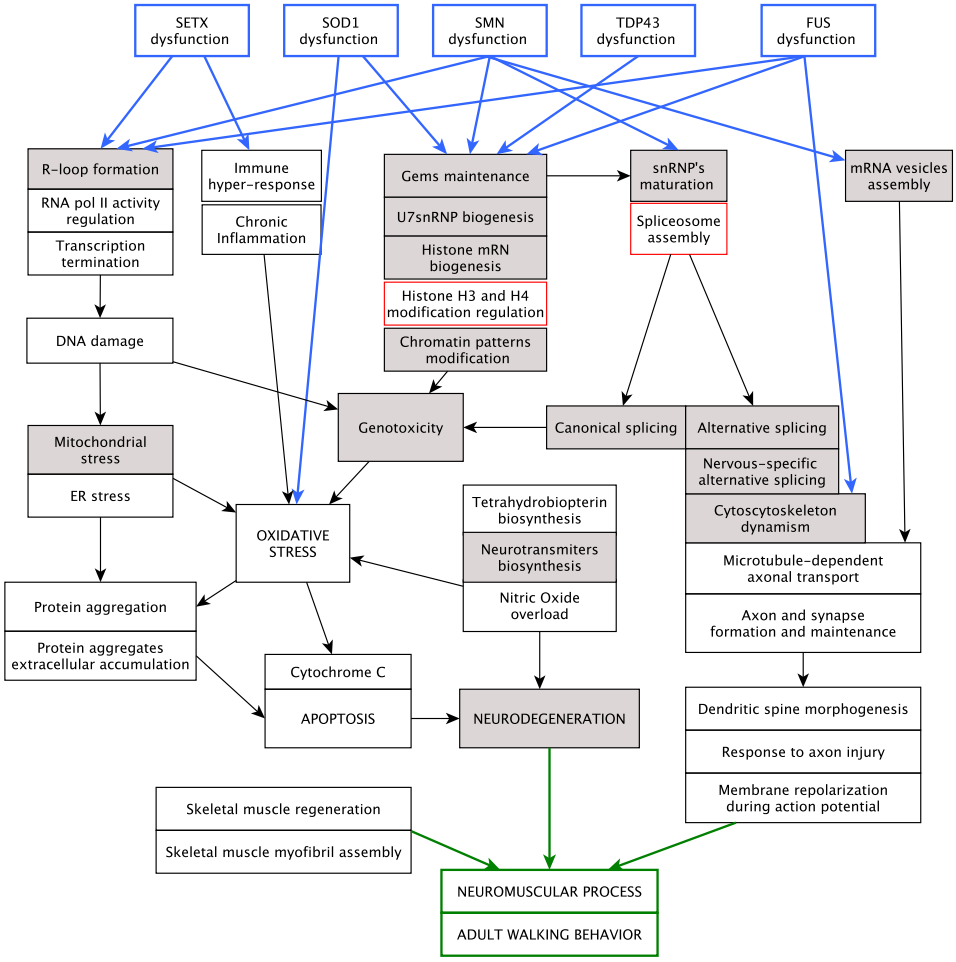

Figure 4.1 MNDs hypothetic mechanisms. FTSS and SOD1 genes dysfunction are considered as MNDs possible causes (boxes with blue border) and final neuromuscular-related functions as phenotypic outputs (boxes with green border). Each box corresponds to a function resulted from this research (white boxes) or described in literature (grey boxes). Functions found only through the analysis of the FTSS-focused network are labeled with a red border.

Figure 4.1 summarizes the main hypothesis retrieved from FTSS and S2B network analysis and related literature knowledge that may have a great impact on MN degeneration. We can divide the general mechanism in **five different pathways** (from left to right in Figure 4.1): **1)** DNA damage and apoptosis induced by R-loop deregulation, **2)** inflammation and neurodegeneration induced by immune hyper-sensitivity, **3)** chromatin deregulation and genotoxicity produced by histone biogenesis perturbation, **4)** splicing patterns alteration and genotoxicity produced by spliceosome assembly failure and **5)** deregulation of microtubule related processes leading to morphological problems in axon and synapse formation.

In the light of these results, splicing and histones perturbation may not be as central as previously was thought. As can be seen on Figure 4.1, these pathways are only based on literature-knowledge (grey colored boxes) and data resulted from FTSS-focused network (described in Chapter 2). Besides, these processes are not specific for MN cells.

On the opposite side, it is well known that oxidative stress is very damaging for nervous tissue (Friedman 2011). Besides, it produces protein aggregation that is directly involved with neurodegeneration as well (Carrì et al. 2015). Within this context these processes can be triggered by mutations in SETX and FUS. These are involved in R-loops formation that when disturbed, induce DNA damage (Rulten et al. 2014; Zhao et al. 2016). FUS and SETX are also involved in DNA repair (Rulten et al. 2014) that, when dysfunctional, lead to genotoxicity. This directly induces mitochondrial stress that in turn also produces endoplasmic reticulum ER stress. These perturbations eventually trigger protein-folding problems, which results in massive protein aggregations that also feedback to oxidative stress. Finally, cytochrome C, localized in mitochondria activates the irreversible apoptosis cascade (Carrì et al. 2015).

Additionally, the functional characterizations done in Chapter 3 identified numerous nitric oxide-related processes. It is an important neurotransmitter that in excess, it also produces oxidative stress (Friedman 2011; Xiong et al. 2007). Furthermore, this hypothesis is greatly supported by the fact that SOD1, the best-known ALS-causative gene (Rosen et al. 1993) is involved in destroying superoxide radicals and thus, when mutated is directly associated to the oxidative stress increase.

SETX is also known to be involved in the immune suppression and thus, when disturbed may lead to a *hyper-sensitivity* to pathogens which causes in turn a excessive inflammatory response (Miller et al. 2015). This fact could be very relevant in MNDs because chronic inflammation is also directly involved in oxidative stress (Amor et al. 2010), protein aggregation and neurodegeneration (Carrì et al. 2015).

Another interesting hypothesis highlighted by our results is that SMN-MN degeneration causal dysfunction is not the spliceosome assembly but its involvement in mRNA transport through axons. Firstly it is known that the cytoskeleton dynamism needed for axonal growth is tightly orchestrated by neuron-specific alternative splicing (Madgwick et al. 2015). Secondly, fast axonal transport is mediated by microtubules (Poulain & Sobel 2010) and in the case of mRNAs, requires the collaboration of SMN for mRNA vesicles formation (Zhang et al. 2006) and is required for synapse maintenance (Griffin & Watson 1988).
Thus when mutated, there is a concurrent alteration of the splicing patterns needed for microtubules' growth and also the mRNA vesicles disassembly which produces an irreparable axon injury. It seems reasonable that when it takes place in motor neurons possibly causes the lack of muscle innervation and thus the general perturbation of neuromuscular processes. As shown on Figure 4.1 this hypothesis is highly supported with functional data obtained simultaneously form FTSS and S2B sets (Chapter 3). Additionally, we also demonstrated in the preliminary FTSS-focused functional analysis (Chapter 2) that FUS interacts with a high number of proteins involved in microtubule-related processes, which reinforces the putative key role of microtubules on MNDs.

Nevertheless, these are predictive results and thus should be confirmed experimentally. In any case, it shows the great value of network-based methods towards the understanding of complex diseases and discovery of associated drug targets and biomarkers.

Future remarks

The proposed S2B method can still be subject to further improvements. Firstly we should implement training sets to evaluate the specificity and sensitivity with which S2B identifies real DAGs. Due to the fact that no disease is fully molecularly described, we must resort to artificial training sets where disease modules would be forced to partially overlap. Then, we would randomly take subsets of these modules to run S2B and compare prioritized nodes with the expected module intersection.

Furthermore, we could also improve the algorithms of S2B method to assign more accurate weights. Similarly to Chain Rank (Tényi et al. 2016) or NERI (Simões et al. 2015) methods, we could integrate into our PPI network gene expression, transcription regulation and/or signaling data. Nevertheless, the integration strategy should be thoroughly designed and at the same time S2B method should be adapted to effectively apply S2B count on a weighted and/or directed PPI network.

Shortest path lengths are biologically relevant and thus, it would be interesting to implement a new score to increase the weight of smaller shortest paths. Currently, when a shortest path is detected, S2B method automatically takes out from S2B count the "*from-to"* seed nodes. This was initially done to avoid DAGs overestimation but it might be excessively dismissing as well. Therefore, we will implement another score that will assign to each *"from-to"* seed node a weight inversely proportional to the length of the detected shortest path.

Thereupon, S2B method could be automatized and applied in all the diseases with known molecular causes, constructing a new "*human diseasome network*". Unlike Goh and colleagues that used gene co-occurrence and comorbidities data (Goh et al. 2007), we could exploit S2B count statistics to link and analyze closeness between diseases.

In a different but related topic, proteins interact with numerous macromolecules within the cell and thus can be considered multifunctional. This is one of the reasons why proteins are usually assigned with varied types of functional terms in Gene Ontology (GO) databases. It does not necessarily mean that these proteins are wrongly annotated. However, one should note that proteins' multifunctionality is restricted to its singular moment in the interactomic context. Thus, when a functional enrichment analysis is performed, retrieved functions can have different relevance depending on the study context. As it is illustrated on Figure 4.1, this fact is highly relevant when the main goal is to discover novel pathomechanisms. Besides, functional enrichment algorithms are known to retrieve noisy results that usually require the use of further depuration and simplifying methods such as REVIGO (Supek et al. 2011).

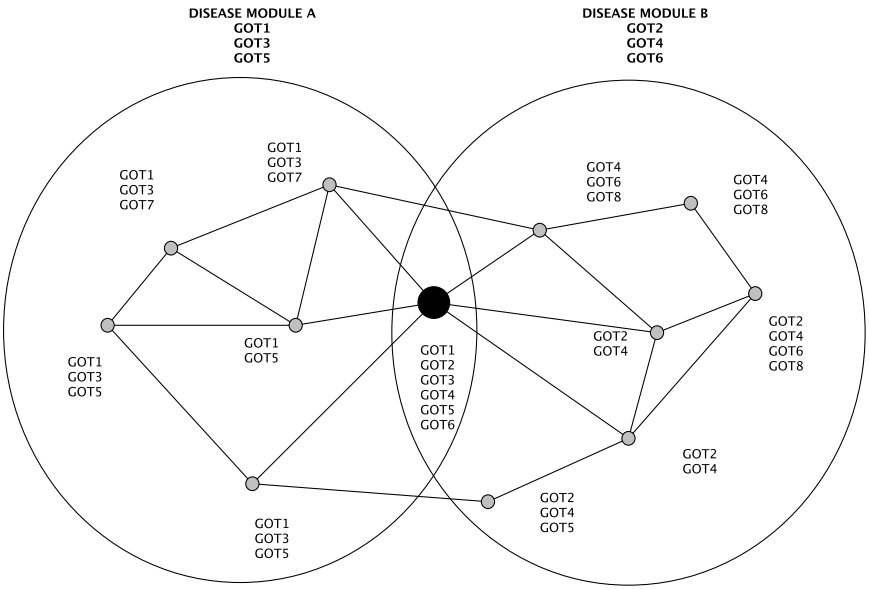

Figure 4.2 Illustration of the differing grades of relevance of functions according to particular diseasome contexts. Nodes represent particular proteins and each one has several associated GO Terms (GOTs). Nodes are linked by PPI data and they are clustered in two overlapping disease modules (A and B). According to the module that is considered, black node's GOTs have different relevance due to the different functions of their interactors. Most relevant GOTs of the black node according to the disease are shown above the respective disease module.

To solve these difficulties, we propose the construction of a network-based method for GO Terms (GOT) prioritization similar to S2B count in order to find which are the GOTs more closely related to the particular studying context. To do that we would firstly implement a simplifying workflow similar to the used in this work (Chapter 3) and secondly we would construct an algorithm to identify which are the functions overrepresented in a particular neighborhood within the PPI network.

Throughout this document we have discussed systems biology principles applied in biomedicine and also demonstrated the usefulness of networks-based approaches in DAGs prioritization in the context of two phenotypically related diseases.

Genes are not isolated entities but interact in complex pathways and mechanisms. Thus, the properties that emerge from these interactions cannot be identified only with reductionist approaches. This fact is reasonably also applied to the progression of all pathologic phenotypes, which is particularly evident in the degeneration of motor neurons on ALS and SMA.

Therefore, we credit network-based approaches applied in biomedicine may bring many clues for MNDs pathomechanisms identification and thus, may contribute towards the development of more effective drugs to improve the quality of life of patients and besides increase their life expectancy.

References

Aibar, S. et al., 2015. Functional gene networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics*, 31(10), pp.1686–1688.

Albert, R., Jeong, H. & Barabasi, A.-L., 2000. Error and attack tolerance of complex networks. *Nature*, 406(July), pp.378–381.

Alexa, A. & Rahnenfuhrer, J., 2010. topGO: Enrichment analysis for Gene Ontology. , p.R package version 2.18.0.

Amor, S. et al., 2010. Inflammation in neurodegenerative diseases. *Immunology*, 129(2), pp.154–169.

Anderson P.W., 1972. More is Different. *Science*, 177(4047), pp.393–396.

Aravind, L., 2000. Guilt by association: Contextual information in genome analysis. *Genome Research*, 10(8), pp.1074–1077.

Barabási, A., Gulbahce, N. & Loscalzo, J., 2011. Network Medicine: A Network-based approach to human disease. *Nature Reviews Genetics*, 12(1), pp.56–68.

Barabási, A. & Oltvai, Z.N., 2004. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5, pp.101–113.

Barabasi, A.-L., 2007. Network Medicine — From obesity to the " Diseasome ." *n engl j med*, 357, pp.404–407.

Barabasi, A.-L. & Albert, R., 1999. Emergence of scaling in random networks. *Science*, 286, pp.509–512.

Bennett, C.L. & La Spada, A.R., 2015. Unwinding the role of Senataxin in neurodegeneration. *Discovery medicine*, 19(103), pp.127–136.

Bromberg, Y., 2013. Disease gene prioritization. *PLoS computational biology*, 9(4).

Brown, K.R. & Jurisica, I., 2005. Online predicted human interaction database. *Bioinformatics*, 21(9), pp.2076–2082.

Calderone, A. et al., 2016. Comparing Alzheimer's and Parkinson's diseases networks using graph communities structure. *BMC Systems Biology*, 10(1), p.25.

Calderone, A., Castagnoli, L. & Cesareni, G., 2013. mentha: a resource for browsing integrated protein-interaction networks. *Nature Methods*, 10(8), pp.690–691.

Can, T., Çamoğlu, O. & Singh, A.K., 2005. Analysis of protein-protein interaction networks using random walks. *Proceedings of the 5th international workshop on Bioinformatics - BIOKDD '05*, p.61.

Carrì, M.T. et al., 2015. Oxidative stress and mitochondrial damage: importance in non-SOD1 ALS. *Frontiers in cellular neuroscience*, 9, p.41.

Chen, Y. et al., 2004. DNA / RNA Helicase gene mutations in a form of Juvenile Amyotrophic Lateral Sclerosis (ALS4). *Am. J. Hum. Genet*, 74, pp.1128–1135.

Cioce, M. & Lamond, A.I., 2005. Cajal bodies: A Long History of Discovery. *Annu. Rev. Cell. Dev. Biol.*, 21, pp.105–131.

Clelland, C.D. et al., 2009. A Functional Role for Adult Hippocampal Neurogenesis in Spatial Pattern Separation. *Science*, 325(5937), pp.210–213.

Cooper, T.A., Wan, L. & Dreyfuss, G., 2009. RNA and Disease. *cell*, 136(4), pp.777–793.

Csárdi, G. & Nepusz, T., 2006. The igraph software package for complex network research. *InterJournal Complex Systems*, 1695, p.1695.

Dezso, Z., Oltvai, Z.N. & Barabási, A.-L., 2003. Bioinformatics analysis of experimentally determined protein complexes in the yeast Saccharomyces cerevisiae. *Genome Research*, 13, pp.2450–2454.

Diantonio, A. & Hicke, L., 2004. Ubiquitin-dependent regulation of the synapse. *Annu. Rev. Neurosci*, 27, pp.223–246.

Diestel, R., 2000. *Graph Theory*,

Dragon, F., Lemay, V. & Trahan, C., 2006. snoRNAs: Biogenesis, Structure and Function. *Encyclopedia of Life Sciences*, pp.1–7.

Erdös, P. & Rényi, A., 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5, pp.17–61.

Fallini, C., Bassell, G.J. & Rossoll, W., 2012. Spinal muscular atrophy: the role of SMN in axonal mRNA regulation. *Brain Research*, 1462, pp.81–92.

Fouss, F. et al., 2007. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3), pp.355–369.

Freeman, L.C., 1978. Centrality in social networks conceptual clarification. *Social Networks*, 1(3), pp.215–239.

Friedman, J., 2011. Why Is the Nervous System Vulnerable to Oxidative Stress? In *Oxidative Stress and Free Radical Damage in Neurology*. pp. 19–27.

Gandhi, T.K.. et al., 2006. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38(3), pp.285–293.

Garcia, E.L. et al., 2013. Developmental arrest of Drosophila survival motor neuron (Smn) mutants accounts for differences in expression of minor intron-containing genes. *RNA society*, 19, pp.1510–1516.

Gillis, J. & Pavlidis, P., 2011. The impact of multifunctional genes on guilt "by association "analysis. *PLoS ONE*, 6(2), p.e17258.

Goh, K.-I. et al., 2007. The human disease network. *PNAS*, 104(21), pp.8685–8690.

Griffin, J.W. & Watson, D.F., 1988. Axonal Transport in Neurological Disease. *Ann neurol*, 23, pp.3–13.

Hamosh, A. et al., 2002. Online Mendelian Inheritance in Man (OMIM): a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30(1), pp.52–55.

Hiesinger, P.R. & Hassan, B.A., 2005. Genetics in the age of systems biology. *Cell*, 123(7), pp.1173–1174.

Ishihara, T. et al., 2013. Decreased number of gemini of coiled bodies and U12 snRNA level in amyotrophic lateral sclerosis. *Human Molecular Genetics*, 22(20), pp.4136–4147.

Jeong, H. et al., 2001. Lethality and centrality in protein networks. *Nature*, 411(6833), pp.41–42.

Jeong, H. et al., 2000. The large-scale organization of metabolic networks. *Nature*, 407(6804), pp.651–654.

Jorjani, H. et al., 2016. An updated human snoRNAome. *Nucleic Acids Research*, 44(11), pp.5068–5082.

Kann, M.G., 2007. Protein interactions and disease: Computational approaches to uncover the etiology of diseases. *Briefings in Bioinformatics*, 8(5), pp.333–346.

Kariya, S. et al., 2012. Mutant superoxide dismutase 1 (SOD1), a cause of amyotrophic lateral sclerosis, disrupts the recruitment of SMN, the spinal muscular atrophy protein to nuclear cajal bodies. *Human Molecular Genetics*, 21(15), pp.3421–3434.

Kashima, T. et al., 2007. hnRNP A1 functions with specificity in repression of SMN2 exon 7 splicing. *Human Molecular Genetics*, 16(24), pp.3149–3159.

Khanin, R. & Wit, E., 2006. How Scale-Free Are Biological Networks. *Journal of Computational Biology*, 13(3), pp.810–818.

Kirschner, M. & Gerhart, J., 1998. Evolvability. *Proceedings of the National Academy of Sciences*, 95(15), pp.8420–7.

Kitano, H., 2004. Biological robustness. *Nature Reviews Genetics*, 5, pp.826–837.

Kitano, H., 2002. Systems biology: A brief overview. *Science*, 295(5560), pp.1662–1664.

Köhler, S. et al., 2008. Walking the Interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4), pp.949–958.

Kondor, R.I. & Lafferty, J., 2002. Diffusion kernels on graphs and other discrete input spaces. *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, pp.315–322.

Kwiatkowski Jr, T.J. et al., 2009. Mutations in the FUS/TLS Gene on Chromosome 16 Cause Familial Amyotrophic Lateral Sclerosis. *Science*, 323, pp.1205–1209.

Lagier-Tourenne, C., Polymenidou, M. & Cleveland, D.W., 2010. TDP-43 and FUS/TLS: Emerging roles in RNA processing and neurodegeneration. *Human Molecular Genetics*, 19(R1), pp.46–64.

Laukens, K., Naulaerts, S. & Berghe, W. Vanden, 2015. Bioinformatics approaches for the functional interpretation of protein lists: from ontology term enrichment to network analysis. *Proteomics*, 15(5-6), pp.981–96.

Leclerc, R.D., 2008. Survival of the sparsest: robust gene networks are parsimonious. *Molecular systems biology*, 4(1), p.213.

Lefebvre, S. et al., 1995. Identification and characterization of a spinal muscular atrophy-determining gene. *Cell*, 80(1), pp.155–165.

Levenson, J.M. & Sweatt, J.D., 2005. Epigenetic mechanisms in memory formation. *Nature Reviews Neuroscience*, 6(2), pp.108–18.

Li, D.K. et al., 2014. SMN control of RNP assembly: from post-transcriptional gene regulation to motor neuron disease. *Semin Cell Dev Biol*, 0, pp.22–29.

Licata, L. et al., 2012. MINT, the molecular interaction database: 2012 Update. *Nucleic Acids Research*, 40(D1), pp.857–861.

Lin, D., 1998. An Information-Theoretic Definition of Similarity. *Proceedings of ICML*, pp.296–304.

Liu, Q. & Dreyfuss, G., 1996. A novel nuclear structure containing the survival of motor neurons protein. *The EMBO journal*, 15(14), pp.3555–65.

Madgwick, A. et al., 2015. Neural differentiation modulates the vertebrate brain specific splicing program. *PLoS ONE*, 10(5), pp.1–14.

Marangi, G. & Traynor, B.J., 2015. Genetic causes of amyotrophic lateral sclerosis: New genetic analysis methodologies entailing new opportunities and challenges. *Brain Research*, 1607, pp.75–93.

Mason, O. & Verwoerd, M., 2007. Graph theory and networks in Biology. *IET Syst Biol*, 1(2), pp.89–119.

Miller, M.S. et al., 2015. The helicase senataxin suppresses the antiviral transcriptional response and controls viral biogenesis. *Nature Immunol.*, 16(5), pp.485–494.

Moreau, Y. & Tranchevent, L.-C., 2012. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13(8), pp.523–536.

Muratani, M. & Tansey, W.P., 2003. How the ubiquitin-proteasome system controls transcription. *Nature reviews. Molecular cell biology*, 4(3), pp.192–201.

Newman, M.E.J., 2003. The structure and function of complex networks. *SIAM Review*, 45(2), pp.pp. 167–256.

Oliver, S., 2000. Guilt-by-association goes global. *Nature*, 403(6770), pp.601–603.

Ooi, H.S. et al., 2010. Databases of Protein–Protein Interactions and complexes. In *Data Mining Techniques for the Life Sciences,Methods in Molecular Biology*. pp. 145 – 159.

Orchard, S. et al., 2012. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature Methods*, 9(6), pp.626–626.

Orchard, S. et al., 2014. The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1), pp.358–363.

Oti, M. et al., 2006. Predicting disease genes using protein-protein interactions. *J Med Genet*, 43, pp.691–698.

Oti, M. & Brunner, H., 2007. The modular nature of genetic diseases. *Clinical Genetics*, 71(1), pp.1–11.

Pattaroni, C. & Jacob, C., 2013. Histone methylation in the nervous system: Functions and dysfunctions. *Molecular Neurobiology*, 47, pp.740–756.

Pavlopoulos, G. a et al., 2011. Using graph theory to analyze biological networks. *BioData mining*, 4(1), p.10.

Peng, J. et al., 2003. A proteomics approach to understanding protein ubiquitination. *Nature biotechnology*, 21(8), pp.921–926.

Pinero, J. et al., 2015. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015, p.bav028.

Piro, R.M., 2012. Network medicine: Linking disorders. *Human Genetics*, 131(12), pp.1811–1820.

Poulain, F.E. & Sobel, A., 2010. The microtubule network and neuronal morphogenesis: Dynamic and coordinated orchestration through multiple players. *Molecular and Cellular Neuroscience*, 43(1), pp.15–32.

Przulj, N., 2011. Protein-protein interactions: Making sense of networks via graph-theoretic modeling. *Bioessays*, 33(2), pp.115–123.

Raczynska, K.D. et al., 2015. FUS/TLS contributes to replication-dependent histone gene expression by interaction with U7 snRNPs and histone-specific transcription factors. *Nucleic Acids Research*, 43(20), pp.9711–9728.

Raman, K., 2010. Construction and analysis of protein-protein interaction networks. *Automated experimentation*, 2(1), p.2.

Ravasz, E. & Barabasi, A.-L., 2003. Hierarchical Organization in Complex Networks. *Physical Review E*, 67(2), p.026112.

Richard, P. & Manley, J.L., 2016. R loops and links to human disease. *Journal of Molecular Biology*.

Rolland, T. et al., 2014. Resource a proteome-scale map of the human interactome network. *Cell*, 159, pp.1212–1226.

Rosen, D.R. et al., 1993. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature*, 362(6415), pp.59–62.

Rual, J.F. et al., 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437, pp.1173–1178.

Rulten, S.L. et al., 2014. PARP-1 dependent recruitment of the amyotrophic lateral sclerosis-associated protein FUS/TLS to sites of oxidative DNA damage. *Nucleic Acids Research*, 42(1), pp.307–314.

Rutherford, N.J. et al., 2008. Novel mutations in TARDBP (TDP-43) in patients with familial amyotrophic lateral sclerosis. *PLoS genetics*, 4(9), p.e1000193.

Sabra, M. et al., 2013. The Tudor protein survival motor neuron (SMN) is a chromatin-binding protein that interacts with methylated lysine 79 of histone H3. *Journal of cell science*, 126(Pt 16), pp.3664–77.

Shannon, P. et al., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13, pp.2498–2504.

Siddique, N. & Siddique, T., 2008. Genetics of Amyotrophic Lateral Sclerosis. *Phys Med Rehabil Clin N Am*, 19(3), pp.429–439.

Simões, S.N. et al., 2015. NERI: network-medicine based integrative approach for disease gene prioritization by relative importance. *BMC Bioinformatics*, 16(Suppl 19), p.S9.

Skourti-Stathaki, K., Proudfoot, N.J. & Gromak, N., 2011. Human Senataxin Resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-Dependent termination. *Molecular Cell*, 42(6), pp.794–805.

Snel, B. et al., 2000. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research*, 28(18), pp.3442–4.

Spirin, V. & Mirny, L. a, 2003. Protein complexes and functional modules in molecular networks. *PNAS*, 100(21), pp.12123–12128.

Sreedharan, J. et al., 2008. TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science*, 319(5870), pp.1668–72.

Sun, S. et al., 2015. ALS-causative mutations in FUS/TLS confer gain and loss of function by altered association with SMN and U1-snRNP. *Nature communications*, 6, p.6171.

Supek, F. et al., 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, 6(7), p.e21800.

Suraweera, A. et al., 2009. Functional role for senataxin, defective in ataxia oculomotor apraxia type 2, in transcriptional regulation. *Human Molecular Genetics*, 18(18), pp.3384–3396.

Suraweera, A. et al., 2007. Senataxin, defective in ataxia oculomotor apraxia type 2, is involved in the defense against oxidative DNA damage. *Journal of Cell Biology*, 177(6), pp.969–979.

Tényi, Á. et al., 2016. ChainRank, a chain prioritisation method for contextualisation of biological networks. *BMC Bioinformatics*, 17(1), p.17.

Thöny, B., Auerbach, G. & Blau, N., 2000. Tetrahydrobiopterin biosynthesis, regeneration and functions. *The Biochemical journal*, 347 Pt 1, pp.1–16.

Tisdale, S. et al., 2013. SMN is essential for the biogenesis of U7 Small nuclear ribonucleoprotein and 3'-end formation of Histone mRNAs. *Cell Reports*, 5(5), pp.1187–1195.

Tong, H., Faloutsos, C. & Pan, J.Y., 2006. Fast random walk with restart and its applications. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp.613–622.

Tranchevent, L.C. et al., 2010. A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*, 12(1), pp.22–32.

Tsuiji, H. et al., 2013. Spliceosome integrity is defective in the motor neuron diseases ALS and SMA. *EMBO Molecular Medicine*, 5(2), pp.221–234.

Turner, B.J. et al., 2014. Overexpression of survival motor neuron improves neuromuscular function and motor neuron survival in mutant SOD1 mice. *Neurobiology of Aging*, 35(4), pp.906–915.

Venkatesan, K. et al., 2009. An empirical framework for binary interactome mapping. *Nature methods*, 6(1), pp.83–90.

Warde-Farley, D. et al., 2010. The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38, pp.214–220.

Winterbach, W. et al., 2013. Topology of molecular interaction networks. *BMC Systems biology*, 7(1), p.90.

Wirth, B., 2000. An update of the mutation spectrum of the survival motor neuron gene (SMN1) in autosomal recessive spinal muscular atrophy (SMA). *Human Mutation*, 15(3), pp.228–237.

Wu, S. et al., 2015. Network Propagation with Dual Flow for Gene Prioritization. *Plos One*, 10(2), p.e0116505.

Xie, K. & Nice, E.C., 2014. Interactomics : toward protein function and regulation. *Expert Rev. Proteomics*, 24, pp.1–24.

Xiong, Y., Rabchevsky, A.G. & Hall, E.D., 2007. Role of peroxynitrite in secondary oxidative damage after spinal cord injury. *Journal of Neurochemistry*, 100(3), pp.639–649.

Yamazaki, T. et al., 2012. FUS-SMN protein interactions link the motor neuron diseases ALS and SMA. *Cell Reports*, 2(4), pp.799–806.

Yu, G. et al., 2012. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, 16(5), pp.284–287.

Yu, H. et al., 2011. Next-generation sequencing to generate interactome datasets. *Nature methods*, 8(6), pp.478–480.

Yu, H. et al., 2007. The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, 3(4), pp.713–720.

Zhang, H. et al., 2006. Multiprotein complexes of the Survival of Motor Neuron protein SMN with gemins traffic to neuronal processes and growth cones of Motor Neurons. *J Neuroscience*, 26(33), pp.8622–8632.

Zhang, Z. et al., 2013. Dysregulation of synaptogenesis genes antecedes motor neuron pathology in spinal muscular atrophy. *PNAS*, 110(48), pp.19348–19353.

Zhao, D.Y. et al., 2016. SMN and symmetric arginine dimethylation of RNA polymerase II C-terminal domain control termination. *Nature*, 529(7584), pp.48–53.

Zotenko, E. et al., 2008. Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality. *PLoS Computational Biology*, 4(8).

Zuckerkandl, E. & Pauling, L., 1962. Molecular Disease, Evolution, and Genic Heterogeneity. *Horizons in Biochemistry*, pp.189–222.

Supplementary data description (CD)

S-2.1 PPI data retrieval

FUS, TDP43, SMN and SETX protein physical interactors were retrieved from mentha, IntAct, GeneMANIA and literature references. All raw PPIs with the source information are described in ALL_PPI.csv. Information of the PPIs obtained from literature is expanded in ftss_literature_PPI.xlsx.

After the data depuration, we only retain the proteins that interact with at least two of FTSS proteins resulting in a smaller set gathered in Selected PPI.csv and ftss_NODES.csv.

S-2.2 FTSS-network functional enrichment and functional clustering

FTSS-focused network functional enrichment result of topGO R function is described in topGO_ftss.txt (GTLinker server does not return the functional enrichment results) and the functional clustering performed using GTLinker algorithm is resumed in genetermlinker_metagroups.xlsx.

S-3.1 MND-focused network construction

ALS and SMA diseases subtypes UMLS CUI identifiers according to MeSH Browser used to retrieve DAG data from DisGeNET are gathered in DISGENET_DiseaseSubtypes_nomenclature.docx.

The raw DAGs collected from DisGeNET are described in ALS_0912_DISGENET_raw.zip and those retrieved from OMIM in ALS_1_OMIM_raw.txt and SMA_1_OMIM_raw.txt files. Finally joined results are described in genedt_ALS.csv, SMA_genedt.csv and genedt_ALSSMA.csv respectively. Continuing, PPI datasets provided in CCSB Download Page were joined and cleaned resulting in the CCSB human interactome. Then the MND-focused network summarized in int_file.csv in was constructed using these PPI and MND DAGs retrieved in the previous subsection.

S-3.2 S2B method

S2B method was applied in MND-focused network resulting in the scores summarized in S2B_250_res.csv.

S-3.3 Functional characterization raw results

Here we list all the functional results returned in each step of functional characterization process of FTSS-S2B and SEEDS-S2B sets comparison.

Functional enrichment: Raw functional enrichment results of individual sets performed with EnrichGO R function.

ftss_enrichGOdf.csv
ALS_enrichGOdf.csv
SMA_enrichGOdf.csv

S2B_enrichGOdf.csv

Functions filter by specificity: Raw functional enrichment results of joined sets after the discarding of GOTs with background frequency higher or equal to 10%.

S2Bseeds_GOT_total_RAW.csv
S2Bftss_GOT_total_RAW.csv

Functions merge by gene coincidence: Raw functional enrichment results of joined sets after the merge of GOTs with a gene co-occurrence equal or higher to 75%.

S2Bseeds_GOT_genesmerge_RAW.csv
S2Bftss_GOT_genesmerge_RAW.csv

Functions merge by semantic similarity: Raw functional enrichment results of joined sets after the REVIGO's merge by GOTs by semantic similarity.

S2Bseeds_GOT_semantmerge_RAW.csv
S2Bftss_GOT_semantmerge_RAW.csv