UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA ANIMAL

Double Specific Betweenness Variants For Cross Disease Network Analysis

Sofia Isabel Rodrigues da Conceição

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:

Professor Doutor Francisco Rodrigues Pinto

2020

Acknowledgements

A big thanks to professor Francisco Pinto for being an awesome mentor with lots of patience, always available to explain the hardest things like they are simple. Also, for being optimistic and helping me to pursuit my career goals.

To the Gene Expression and Regulation group for the shared knowledge and sympathy.

To my friends, with a special mention to Marta for being a ruthless python teacher and although she believes a toddler can program better than me, she always had time (and sometimes patience) to help.

Resumo

Genes associados a doenças (GAD) tendem em agrupar-se em redes celulares definindo módulos na rede. As observações indicam que estes módulos têm grande interação de proteína-proteína e partilham caminhos comuns. É importante identificar estes módulos de doenças, uma vez que poderá ajudar a desvendar os mecanismos moleculares que fazem parte da doença, mas também descobrir novos genes alvos para fármacos. Contudo, achar estes módulos não é fácil, uma vez que o conhecimento atual está em constante progressão, o que leva a redes de doenças incompletas.

Medidas da teoria dos grafos podem caracterizar quantitativamente e comparar redes, os seus vértices ou grupos de vértices. Utilizando estas medidas, alguns métodos foram desenvolvidos de modo a melhor identificar os módulos de doenças nas redes. Um desses métodos é o *Double Specific-Betweenness* (S2B), que prevê GADs partilhados com base nos genes que são achados frequentemente, e, especificamente, nos caminhos mais curtos entre os módulos das duas doenças relacionadas. O S2B prevê genes comuns entre doenças na sobreposição de ambos os módulos de doenças numa rede não dirigida usando GAD conhecidos. Com isto, o método prevê que vértices presentes na rede têm mais probabilidade de fazer parte de ambos os módulos. Este método difere de outros já existentes, porque prevê vértices que estão associados simultaneamente a duas doenças, enquanto os outros métodos focam-se apenas em priorização de proteínas, apenas para uma doença. A utilização de duas doenças simultaneamente é uma vantagem, uma vez que é possível inferir mecanismos subjacentes a comorbidades em doenças ou identificar vértices, e/ou caminhos que não estavam previamente associados a comorbidade, dando assim a possibilidade de tratar a doença ou sintoma em ambas em vez de atuar numa só.

Neste trabalho desenvolvemos variantes mais flexíveis do S2B para redes não dirigidas e dirigidas. Estas variantes foram comparadas em três tipos de redes distintas e foram avaliados em termos de robustez e estabilidade. Os resultados obtidos mostram que o S2B, em conjunto com as variantes que usam passeios aleatórios (SRWR e SLB) são os métodos com melhor performance. Estes métodos têm uma melhor performance na rede de sinalização não dirigida. Os resultados também demonstram que os métodos são robustos a mudanças iniciais nas sementes (variação aleatória de sementes ou no número de sementes usadas). O rácio do módulo de doença completo para o numero de sementes de *input* está correlacionado com a performance do método. Resumindo, este trabalho providencia um guia compreensivo para a aplicação de métodos de previsão da sobreposição com a máxima performance.

Palavras Chave: Biologia de Redes; Priotização de Genes de Doenças; Interações Proteína-Proteína; Redes de sinalização

Abstract

Disease associate genes (DGs) tend to cluster in cellular networks defining network modules. It has been observed that these modules have many protein-protein interactions and share common pathways. It is important to identify these disease modules, since it can help to unveil the molecular mechanisms that contribute to disease, as well as discovering new candidate drug target genes. However, finding these modules is not an easy task, and, as the current knowledge is still expanding, most known disease network modules are incomplete.

Graph theory measures can quantitatively characterize and compare networks, their nodes or group of nodes. Using these measures, some methods were developed to better identify network disease modules. One method is the Double Specific-Betweenness (S2B), that predicts shared DGs based on the genes that are found frequently and specifically in shortest paths between two related disease modules. S2B predicts cross-disease genes in the overlap of both disease modules in an undirected graph using known DGs. With this, the method predicts the nodes present in the graph that are most likely to be part of both disease modules. This method differs from previously existing methods because it predicts nodes that are simultaneously associated with two diseases, while other existing methods focus in protein prioritization only for one disease. Using two diseases at the same time is an advantage, since we can infer new mechanisms underlying comorbid diseases or identify nodes and/or pathways that were not assigned to a comorbidity before, giving the possibility to treat the disease or symptom for both of them instead of a single one.

In this work we developed more more flexible S2B variants for undirected and directed networks. These variants were compared in three distinct network types and were evaluated regarding robustness and stability. The obtained results show that S2B, together with variants using random walks (SRWR and SLB) are the best performing methods. These methods perform better with an undirected signalling network. Also, the results show that the methods are robust to changes in initial seeds (random variation of the seeds or the number of seeds used), and to module connectivity. The ratio of complete module size to the number of input seeds is correlated with method performance. In summary, this work provides a comprehensive guide for the application of module overlap prediction methods with maximal performance.

Keywords: Network Biology; Disease Genes Prioritization; Protein-Protein Interaction; Signalling Network

Resumo Alargado

O aumento de estudos de associação genómica (GWAS), o desenvolvimento de novas tecnologias ómicas e de um grande número de métodos de diagnóstico molecular, fazem com que, nos últimos anos, a descoberta de novos genes associados a doenças esteja a aumentar. Toda a nova informação gerada é importante para uma melhor compreensão das doenças humanas e também para o desenvolvimento de novas terapias. Os genes associados a doenças (GAD) tendem a agrupar-se em redes biomoleculares definindo módulos na rede. Os módulos conhecidos partilham um grande número de interações de proteína-proteína e processos celulares associados. É importante identificar estes módulos de doenças uma vez que poderão ajudar a desvendar os mecanismos moleculares que originam a doença, mas também descobrir novos genes candidatos a alvos terapêuticos. Contudo, encontrar estes módulos a partir de evidências experimentais é um processo demorado e dispendioso. O desenvolvimento de métodos computacionais de previsão de GAD pode optimizar o processo de descoberta dos módulos de doença e reduzir os custos associados. As redes biomoleculares são uma ferramenta útil para o estudo de doenças, uma vez que conseguem representar as interações conhecidas, sendo os intervenientes (genes, RNAs, proteínas, metabolitos) representados por vértices e as interações por arestas. A topologia da rede, ou seja, como os vértices e as arestas estão organizadas na rede, podem ter padrões que se traduzem em tipos de comportamentos no sistema que estamos a representar. Por isso, várias propriedades topológicas são utilizadas para caracterizar os vértices na rede. Uma das propriedades topológicas mais utilizadas é o grau de um vértice. Esta propriedade dá informação sobre quantas ligações tem o vértice. Um vértice com um elevado grau tem influência em vários outros vértices. Outros tipos de medidas topológicas também bastante utilizadas são os caminhos mais curtos que representam a sequência mais pequena de arestas conectadas entre dois vértices, sendo um indicador de relação entre vértices.

A previsão de vértices associados a doenças neste tipo de sistema é bastante importante uma vez que permite não só uma melhor compreensão das doenças, mas também a identificação de possíveis bio marcadores e alvos para tratamento. É neste contexto que surge a medicina de redes, que integra o conhecimento acumulado sobre redes biomoleculares com os GAD conhecidos de modo a construir novos modelos de redes, focando-se nos vértices mais importantes das doenças e nas suas interações. Há medidas que usam a topologia da rede para caracterizar quantitativamente ou comparar redes, os seus vértices ou grupos de vértices. Utilizando estas medidas, alguns métodos foram desenvolvidos de modo a melhor identificar os módulos de doenças em redes celulares. No entanto, a maioria destes métodos apenas

utiliza uma doença de cada vez, ou utilizam dados independentes externos, tal como dados fenotípicos e usam redes não dirigidas que não têm em conta a direção da propagação da interação.

Um método inovador que apenas utiliza a informação da rede e duas doenças que tenham semelhanças fenotípicas é o *Double Specific-Betweenness* (S2B). O S2B prevê GAD associados às duas doenças que são especificamente e frequentemente encontrados nos caminhos mais curtos entre as doenças. Ao prever vértices associados a duas doenças, este método permite identificar novos caminhos na rede, ou vértices, que não estavam diretamente ligados a ambas as patologias, possibilitando assim uma via de tratamento que tenha efeito em ambas as doenças. Embora o S2B seja um método inovador, continua a não ter em conta as redes dirigidas, usando apenas redes não dirigidas. Versões do S2B para redes dirigidas foram elaboradas, no entanto, não foi explorado de um modo detalhado o seu desempenho. Tendo em conta as limitações do S2B, estas serviram como motivação para esta dissertação. Um dos objetivos desta dissertação passa pela criação de variações do S2B, variações que usem outros tipos de medidas topológicas para a previsão de vértices, tanto em redes dirigidas, como em não dirigidas e a comparação entre elas. Para esta tarefa utilizou-se a base de dados DisGeNET para obtenção dos pares de doenças. Estes pares de doenças foram categorizados de acordo com o número de vértices na rede, dando origem às categorias A para pares com 200 a 400 vértices, B com 400 a 600 vértices, C de 600 a 800 vértices, D para 800 a 1000 vértices e E para mais de 1000 vértices. Para cada par foram identificados os vértices presentes na sua intersecção. Os vértices identificados nesta fase correspondem aos verdadeiros positivos da análise. A rede utilizada para ambas as formas de teste (dirigido e não dirigido) foi uma rede de sinalização combinada das bases de dados OmniPath e DoRohEA. Cinco variações do S2B foram desenvolvidas, *Specific Neighbors* (SN), *Specific Closeness* (SC), *Specific Random Walk Closeness* (SRWC), *Specific Random Walk with Restart* (SRWR) e *Specific Lenght weighted Betweenness* (SLB). Cada um destes métodos foi aplicado aos vários pares de doenças e o seu desempenho foi avaliado tendo em conta a precisão e sensibilidade. Os três melhores métodos da rede não dirigida foram adaptados em três versões distintas para a versão dirigida. Uma versão que procura caminhos sem direção entre os módulos (V1), uma segunda versão que procura caminhos bidirecionais que convergem na intersecção dos módulos (V2), e uma última versão que procura caminhos bidirecionais que divergem da intersecção (V3). Os melhores métodos na versão não dirigida foram S2B, SLB e SRWR conseguindo uma precisão mediana de mais de 80% na classe E. Os resultados na rede dirigida foram bastante abaixo dos obtidos com a rede não dirigida, conseguindo o método S2B_V1 uma mediana de 76% na classe E. Estes resultados indicam que este tipo de métodos tem um melhor desempenho em redes não dirigidas.

Também foi avaliada a estabilidade e robustez dos métodos. Para isto, testou-se a variação do número de vértices de *input* inicial, a selecção aleatória dos inputs iniciais, diferentes tamanhos dos módulos de doenças e conectividade do módulo. Uma vez que foram obtidos melhores resultados com a rede não dirigida, esta foi utilizada para estes testes. Foi criada uma classe assimétrica AE com um dos pares da doença com 200 a 400 vértices na rede e outro par com mais de 1000. Os resultados de precisão nesta nova classe foram intermédios às classes A e E. Foram também utilizados diferentes tamanhos de *inputs* como 50, 100 e 150. Ao variar aleatoriamente as proteínas dos módulos usadas como input, o desvio padrão da precisão alcançada foi entre 2% e 4% dos valores medianos, indicando que a variabilidade no desempenho não é muito afetada pela selecção aleatória dos nós de input. Verificou-se se a conectividade inicial dos módulos de doença tinha correlação com a precisão do método. Este teste mostrou que estes métodos não são significativamente afectados pela proximidade na rede entre os nós do mesmo módulo. Por último, foi avaliado o desempenho dos métodos em três tipos de rede, rede de interação proteína – proteína (APID), rede de sinalização (OmD) e uma rede composta (PCNET) que reúne vários tipos de interacção. Os métodos S2B e SLB tiveram um bom desempenho entre todas as redes, no entanto a rede com os melhores resultados foi a OmD. Isto sugere que a rede de sinalização seja mais informativa em termos de mecanismos partilhados entre doenças.

Pela primeira vez o S2B e variantes foram testados com módulos de doença reais. Os resultados desta dissertação mostram que estes métodos têm um melhor desempenho em redes não dirigidas, atingindo o melhor desempenho na rede de sinalização testada e em doenças com mais vértices na rede. Demonstrou-se também a robustez dos métodos em relação ao diferente número de *inputs*, tamanhos de classes e vértices adicionados. Estes resultados combinados demonstram que tanto o S2B como algumas das suas variantes podem ser utilizadas para prever GAD conseguindo assim obter precisões médias a elevadas em alguns cenários.

Contents

1 Introduction **1**

1.1 Theory Fundamentals 1

1.2 Network Medicine 3

1.3 Gene Prioritization Algorithms 4

1.4 Objectives 6

2 General Methods **7**

2.1 Method Inputs: Networks And Gene-Disease Associations 7

2.2 Selection of Disease Pairs for Method Validation 8

2.3 S2B Variants 9

2.3.1 SN - Specific Neighbors 9

2.3.2 SC - Specific Closeness 10

2.3.3 SRWC - Specific Random Walk Closeness 10

2.3.4 SRWR - Specific Random Walk with Restart 11

2.3.5 SLB - Specific Length weighted Betweenness 11

2.4 Directed Network Variants 11

2.5 Performance Evaluation 13

3 Undirected *versus* Directed Network Comparison **15**

3.1 Introduction 15

3.2 Workflow 16

3.2.1 Network 17

3.2.2 Directed S2B Variants 17

3.2.3 Wilcoxon test 17

3.3 Results 18

3.3.1 Undirected versus Directed network 18

3.4 Discussion 22

4 Evaluation of method robustness 25

4.1 Introduction 25

4.2 Workflow 26

4.2.1 Variation of complete module sizes 26

4.2.2 Variation of input seed sets 26

4.2.3 Variation of Disease Module connectivity 26

4.3 Results 27

4.3.1 Variation of complete module sizes 27

4.3.2 Variation of input seed sets 28

4.3.3 Variation of disease module connectivity 28

4.4 Discussion 29

5 Comparison of methods performance across different undirected networks 31

5.1 Introduction 31

5.2 Workflow 32

5.2.1 Networks 32

5.2.2 Computational Workflow 32

5.3 Results 33

5.4 Discussion 37

6 General Discussion **39**

7 Future Prospects **41**

References **43**

Supplementary Results

List of Figures

1.1 Example of networks and adjacency matrices from [Newman, 2010] (**A**) Undirected network. (**B**) Adjacency matrix of network A. (**C**) Directed network. (**D**) Adjacency matrix of network B. 2

2.1 Directed S2B versions from [Ramos, 2018]. (**A**) Version 1. (**B**) Version 2. (**C**) Version 3. 12

2.2 Overhall worflow. 14

3.1 Methods precision distribution per tops on all versions. (**A**) Class A undirected version. (**B**) Class E undirected version. (**C**) Class A directed version. (**D**) Class E directed version 20

3.2 Methods recall distribution per tops on all versions. (**A**) Class A undirected version. (**B**) Class E undirected version. (**C**) Class A directed version. (**D**) Class E directed version . 21

4.1 Class AE metrics boxplots across the different methods and tops on the undirected version. (**A**) Precision. (**B**) Recall. 27

4.2 S2B Precision values for the top candidates across size classes. 28

4.3 Standard variation of precision values obtained with independent random sampling of input seed sets for disease pairs of class A. 29

5.1 Precision scores for class A and E in all networks. (**A**) APID class A. (**B**) APID class E. (**C**) OmD class A. (**D**) OMD class E. (**E**) PCNET class A. (**F**) PCNET class E. 35

5.2 Recall scores for class A and E in all networks. (**A**) APID class A. (**B**) APID class E. (**C**) OmD class A. (**D**) OMD class E. (**E**) PCNET class A. (**F**) PCNET class E. 36

List of Tables

3.1 Pairs of diseases per class used to perform the evaluations. 17

3.2 Wilconxon test based on precision resume between pairs of methods. 19

5.1 Network characteristics. 33

5.2 Pairs of diseases per class used to perform the evaluations. 33

List of Acronyms

APID	Agile Protein Interactomes DataServer
Cardigan	ChARting DIsease Gene AssociatioNs
Dgs	Disease assocciate genes
DIAMOnD	DIseAse MOdule Detection
DoRoThEA	Discriminant Regulon Expression Analysis
FN	False negatives
FP	False positives
GDAS	Gene-disease associations
GLADIATOR	Global Approach for Disease AssociaTed mOdule Reconstruction
OmD	OmniPath and DoRothEA
PCNet	Parsimonious Composite Network
PPI	Protein-protein physical interactions
RWR	Random Walk with Restart
S2B	Double specific-betweenness
SC	Specific Closeness
SLB	Specific Length weighted Betweenness
SN	Specific Neighbors
SRWC	Specific Random Walk Closeness
SRWR	Specific Random Walk with Restart
STRING	Search Tool for Retrieval of Interacting Genes/Proteins
TF	Transcription factor
TN	True negatives
TP	True positives

Chapter 1

Introduction

This chapter serves as an introductory chapter summarizing the theory fundamentals of network theory, while also presenting state of the art and objectives of this dissertation.

1.1 Theory Fundamentals

Systems biology tries to understand the behavior of diverse biological systems as whole entities, preserving the role of the interactions between system components [Ingalls, 2013].

The behavior of the system can be studied using different methodologies and approaches, one of them is through the use of networks. Networks provide an insightful representation of interactions between different components of biological systems [Newman, 2010; Liu et al., 2019, 2020]. These networks reveal patterns and the structure of the networks gives us hints of what might influence the system or how perturbations spread through the system [Newman, 2010]. The components of the networks are represented by nodes (also referred as vertices) and their interactions represented by edges [Newman, 2010]. The edges can be directed , i.e. , having specific direction from one node to the other (A-B different from B-A), or undirected (A-B equal to B-A) [Newman, 2010]. Edges and nodes can also be characterized by quantitative weights, which may reflect the abundance of each node in the system or the strength of each interaction. A network (referred as graph in the mathematical notation) can be mathematically represented using an adjacency matrix. In a simple undirected network (Fig. 1.1a) (does not have self-edges) the interaction between the vertices i and j is described in the matrix element Aij by a binary that is 1 when there is interaction and by 0 when there is not (Fig.1.1b) [Newman, 2010]. Also in these types of networks the diagonal elements are zero and the matrix is symmetric (Aij = Aji)

[Newman, 2010]. For simple directed networks the representation is still binary but it goes from j to i and the resulting matrix is asymmetrical (Fig. 1.1c, 1.1d) [Newman, 2010].

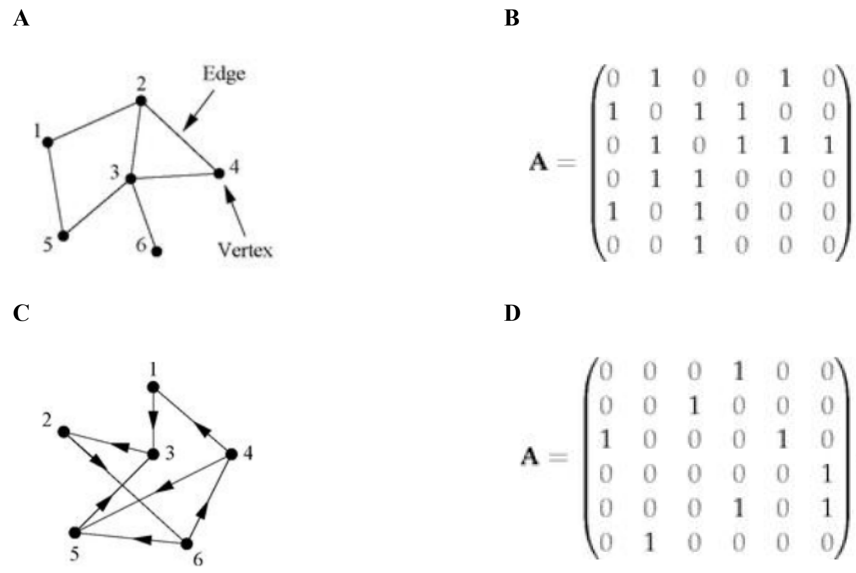

Figure 1.1: Example of networks and adjacency matrices from [Newman, 2010] (**A**) Undirected network. (**B**) Adjacency matrix of network A. (**C**) Directed network. (**D**) Adjacency matrix of network B.

An important topological proprieties of the nodes in the network is the **degree**, which describes the number of edges connected to a node [Newman, 2010; Liu et al., 2020]. In undirected networks it is obtained by the sum of rows or columns of the adjacency matrix (eq. 1.1).

$$ki = \sum_{j=1}^{n} Aij \tag{1.1}$$

For directed networks there are two types of degree. Out-degree corresponds to the edges that diverge from the node while the in-degree counts the edges that converge into the node. In the adjacency matrix the out-degree is the sum of the row (eq. 1.2) and the in-degree the sum of the column (eq. 1.3) [Newman, 2010].

$$kj^{out} = \sum_{i=1}^{n} Aij \tag{1.2}$$

$$ki^{in} = \sum_{j=1}^{n} Aij \tag{1.3}$$

These proprieties are very important since it is the most direct measure to infer the importance or influence of a node in the network, as nodes with high degree have the ability to influence more nodes [Newman, 2010]. Network metrics that try to capture the relative importance of each node are called centrality measures. As such, the **degree** is often referred as **degree centrality**. On directed networks, degree centrality creates two types of important nodes, nodes with many neighbor nodes that flow into them (nodes with an high in-degree, authorities) and nodes with an high out-degree (hubs) [Newman, 2010].

Nodes in a network communicate through paths, a sequence of connected vertices [Newman, 2010]. For directed networks, paths must follow edges with coherent directions: considering two nodes i and j, the direction of the path is very important since the path from i to j might not be he same from j to i. A **Shortest path** between two nodes is a path with the smallest possible number of edges that connects the two nodes. The shortest path length is the number of edges that are part of the shortest path. It defines the distance between two nodes in the network and is a basic measure of similarity or relatedness between nodes. The complete network structure may be characterized by the **average path length**($<$ d $>$) that represents the average of the shortest paths lengths between all pairs of nodes [Newman, 2010].

One special type of path is a **random walk**, which is basically a path generated by repeated steps (t), initiating from a starting node and going from one node to the next by a process of random selection, where each neighbor of the current node can be visited according to a certain probability (p) (k_{j}) [Newman, 2010]. A general random walk in a undirected network can be described by (eq. 1.4).

$$p_{\mathrm{i}}(t) = \sum_{j} \frac{Aij}{kj} p_{\mathrm{j}}(t-1) \tag{1.4}$$

Building upon these basic concepts, a variety of measures and metrics have been developed that use information from the network structure and try to quantify relevant aspects of the network topology. Network topology represents how the nodes and edges are arranged [Liu et al., 2019, 2020]. Among these, closeness centrality and betweenness centrality are also very used [Newman, 2010].

Closeness centrality, specifies the mean distance from one node to all the other nodes in the network. Nodes that present a low closeness centrality receive quickly the information from other nodes or efficiently exert influence on them [Newman, 2010; Cáceres and Paccanaro, 2019; Liu et al., 2020].

On the other hand, **betweenness centrality** measures the frequency with which a node is part of shortest paths between other nodes. An high betweenness indicates that the node might control the information flow across the network. Removal of these nodes can cause a significant disruption of the communication between other nodes [Newman, 2010; Liu et al., 2019].

1.2 Network Medicine

With the recent emergence of genome wide association studies, omic technologies and high throughput molecular screening methods, the discovery of new genes related with diseases has increased [Dozmorov, 2018]. This information is important for the understanding of human diseases and for the development of new therapies [Dozmorov, 2018].

Disease associate genes (DGs) tend to cluster in cellular networks defining network modules [Goh et al., 2007]. Known DGs of the same disease (a disease module) tend to have shared genomic variants, to have protein-protein interactions and be involved in common pathways [Dozmorov, 2018]. It is important to identify these disease modules, since it would help to unveil molecular mechanisms that lead to the disease, as well as discovering new candidate genes. However, finding these modules is not easy, since the current knowledge is still expanding, leading to incomplete disease network modules [Menche et al., 2015].
In this context, Network Medicine has developed as a new scientific discipline that explores biomolecular networks to understand disease etiology, identification of biomarkers and discovery of new drug targets and therapeutic approaches [Sonawane et al., 2019]. Network medicine deeply relies on large data sets to accurately build network models and subsequently perform new predictions [Sonawane et al., 2019]. The core of network medicine starts by the identification of the most important disease nodes and their respective interactions [Sonawane et al., 2019].
A standard approach in this field is understanding human diseases from the perspective of the interactome. This involves trying to figure out the origin of the perturbation in the network that leads to disease, search for new drug repurposing candidates or propose new drug targets, all this using the network topology to infer the relationship between disease genes [Sonawane et al., 2019].
For this purpose, some methodologies to unveil critic nodes associated with diseases were developed.

1.3 Gene Prioritization Algorithms

Using the measures referred on section 1.1, some algorithms were developed to better identify network disease modules. Most of the known algorithms use only one disease at the time to predict new genes/proteins associated with it. The DIseAse MOdule Detection (DIAMOnD) algorithm [Ghiassian et al., 2015] identifies a full disease module associate from an input set of known disease proteins. DIAMOnD works by using the hypergeometric distribution with base on initial seed proteins (s_0). Having two proteins with the same degree (k), the one that has a highest number of links to the s_0 (ks) will retrieve the lowest p-value. The inverse is also true. In each interaction k and ks are assessed. After the candidate proteins are ranked according to the k and then the ones with an highest ks are selected and hypergeometric p-values are calculated. The candidate with the lowest p-value is merged with the seed group. In this way the module grows up one protein element at each interaction, beginning from s_0.

A different algorithm is the GLobal Approach for DIsease AssociaTed mOdule Reconstruction (GLADIATOR) [Silberberg et al., 2017]. This method tries to infer disease modules for multiple diseases simultaneously taking into account a gold-standard of phenotypic similarity between diseases. Similarly to

DIAMOnD, this method also relies on initial seed proteins. It uses a simulated annealing approach to find the set of connected disease modules with an overlap that best matches the known phenotypic similarities.

A more recent algorithm is the ChARting DIsease Gene AssociatioNs (Cardigan) [Cáceres and Paccanaro, 2019] that uses semi-supervised learning. Cardigan initiates by calculating the phenotypic similarity between the query disease and other diseases. On the next step it gives a weight to each gene according to which diseases it is associated. Maximum score is obtained if the gene is associated with the query disease, but associations with other diseases also contribute with a score that is higher if the disease is phenotipically similar with the query disease. The resulting scores are propagated over the network using a semi-supervised method, leading to a final vector with the probability of being associated with the query disease for every gene in the network.

DIAMOnD follows the most common approach that tries to predict new disease genes using a gene/ protein network and genes previously known to be associated with the query disease. GLADIATOR and Cardigan attempt to improve performance through the borrowing of information from phenotypically similar diseases. These methods have to use additional independent data sources to compute the similarity between diseases.

An unique method that approaches two diseases at the same time and relies only on the network information is the **Double Specific-Betweenness (S2B)**, that predicts shared DGs based on the genes that are found frequently and specifically in shortest paths between DGs of two related diseases [Garcia-Vaquero et al., 2018]. S2B predicts cross-disease genes in the overlap of both disease modules in an undirected graph using genes known to be associated to either of the two diseases (or seeds). With this, the method predicts nodes present in the graph that are most likely part of both disease modules [Garcia-Vaquero et al., 2018].
This method differs from previously existing methods because it predicts nodes that are simultaneously associated with two diseases [Garcia-Vaquero et al., 2018]. Using two diseases at the same time is an advantage, since we can infer new mechanisms underlying co-morbid diseases or identify nodes and/or pathways that were not assigned to a co-morbidity before, giving the possibility to treat the disease or symptom for both of them instead of a single one.

S2B is computed from network shortest paths, which may neglect other short (but not shortest) paths connecting two disease modules. This method is calculated according to equation 1.5. Where G is the undirected graph, a and b are the seeds of disease A and B respectively, $sp(k, i, j, G)$ is an indicator function that equals 1 if the length of the shortest path between i and j includes k and $t(i, j, G)$ is an indicator function that equals to 1 if the length of the shortest path between i and j is equal or lower than the average shortest path length of G.

$$S2B(k,G,a,b) = \frac{\sum_{i}^{i\in a, i\neq j} \sum_{j}^{j\in b, j\neq k} sp(k,i,j,G) \times t(i,j,G)}{\sum_{i}^{i\in a, i\neq j} \sum_{j}^{j\in b, j\neq k} t(i,j,G)} \tag{1.5}$$

Additionally, an S2B version adapted to directed networks has been developed [Ramos et al., 2019]. Here, the motivation was to better extract the knowledge encoded in signaling and transcriptional regulatory networks through the correct use of interaction directions. However, its predictive performance has not been completely evaluated. Particularly a comparison between undirected and directed S2B methods using exactly the same network has not been performed. Both versions of S2B have been evaluated through the use of simulated disease modules. Although this has the advantage of defining objectively all predictions as true or false, there is a need to evaluate these methods using a validation with known disease genes, as simulated modules may not accurately represent true disease modules. These S2B limitations have been addressed in the present thesis, where we have developed variants of the S2B method, both for undirected and directed networks, and evaluated their predictive performance with validation tests for multiple disease pairs with known overlapping sets of disease genes.

1.4 Objectives

The goals of this dissertation are to:

- Develop more flexible S2B variants, both for directed and undirected networks;
- Compare directed and undirected methods applied with the same network;
- Evaluate the robustness and stability of the developed methods to variations in input seeds, module size, module connectivity and network type.

Chapter 2

General Methods

This chapter presents the methodology used on the later chapters on this thesis.

2.1 Method Inputs: Networks And Gene-Disease Associations

The methods developed and evaluated in this work use networks as one of its inputs. Different biological networks constitute different perspectives of a biological system. In particular, we were interested in evaluating the impact of using directed networks on the performance of the methods. Therefore, we used several sources of molecular interactions: (1) OmniPath: literature curated mammalian signaling pathways [Türei et al., 2016], (2) Discriminant Regulon Expression Analysis (DoRoThEA) [Garcia-Alonso et al., 2018], (3) Agile Protein Interactomes DataServer (APID) [Alonso-Lopez et al., 2016; Alonso-López et al., 2019], and (4) Parsimonious Composite Network (PCNet) [Huang et al., 2018]. OmniPath is a collection of literature curated human and rodent signalling pathways. DoRothEA is a manually curated database of human regulons, including estimates of single transcription factor (TF) activities inferred from gene expression data, consensus TF-target DNA binding sites and ChIP-seq data. Not all TF-target gene interactions are supported by the same amount and type of evidences. To express the degree of quality of the supporting evidence, DoRothEA interactions have a confidence score that ranges from A to E (A- most confident, E- less confident). In this work, it was chosen to use interactions with scores from A (curated / high confidence) to C (medium confidence). Additionally, OmniPath (only human interactions) was merged with DoRothEA as a unique network and will be referred as OmD. Interactions retrieved from both OmniPath and DoRothEA are directed. OmD was used both as a directed and as an undirected network.

APID is a compilation of protein interactomes based in the integration of known experimentally validated protein-protein physical interactions. Protein-protein physical interactions are undirected. APID also provides different quality levels according to the amount of evidence supporting each interaction. In this work, the interactome used was of Quality Level 1, composed of interactions supported by at least 2 independent experimental assays for the *Homo sapiens* organism.

PCNET is a parsimonious composite network that is a result of the integration of twenty one distinct networks. Each PCNET interaction needs to be present at least on two of the twenty one supporting networks. Although some of the supporting networks are directed, PCNET is an undirected network.

The studied methods also use two lists of genes as inputs. Each list should contain genes associated associated with a disease. In this work, these lists were retrieved from DisGeNET [Pinero et al., 2015; Piñero et al., 2016]. The DisGeNET database is a collection of information obtained from various repositories of human gene-disease associations (GDAs) and variant-disease associations from various repositories. All GDAs present in DisGeNet were used, independently of the types of evidences supporting each association. This maximized the size of the disease modules mapped on the different networks used.

2.2 Selection of Disease Pairs for Method Validation

To evaluate the performance of the methods in an extensive way, we selected a large number of disease pairs with overlapping network modules. In the selection process we controlled the module size of both diseases and the relative size of the overlap. The selection workflow described bellow was repeated for all the different networks used. First, the proteins from genes associated to diseases obtained from DisGeNET were mapped to the network nodes to define disease modules. Only diseases that correspond to the disease type ”Disease” were used (excluding the types ”Group” and ”Phenotype”). For diseases that did not had a connected module, the minimum Steinner tree algorithm was used to create connected disease modules. This algorithm finds the connected subnetwork with the minimum number of nodes that contains all input nodes. To achieve this, it computes a minimum spanning tree [Prim, 1957] of an auxiliary fully connected network with the input nodes, where each edge is weighted by the shortest path length between the two nodes in the original network. Modules that had more than 30% of nodes added by this process were excluded.

Next, modules were divided into size classes according to the number of nodes: class A, 200 to 400 nodes, class B, 400 to 600 nodes, class C, 600 to 800 nodes, class D, 800 to 1000 and class E more than 1000 nodes.

Within each class, all disease pairs were screened to identify their overlap. Besides nodes in the intersection between the two modules, nodes of one disease module that were direct neighbors of nodes from the other disease module have been considered part of the overlap. Pairs of disease modules that share up to 50% of nodes (relatively to the smaller module) were selected. Pairs of diseases where one disease is from class A and other is from class E were similarly selected.

Finally, for each selected pair of disease modules, random samples of 50, 100 and 150 nodes from each disease were generated. To evaluate the possible influence of the sampling stochasticity, five independent random samples of size 100 were also generated for all disease pairs. The random sampling excluded nodes that were part of the overlap or that were added by the Steiner tree algorithm. The random samples of each pair of diseases were used as inputs to the methods under evaluation. The presence of the known overlapping nodes in the top ranking nodes of each method was used to evaluate their performance (see Performance Evaluation section 2.5).

2.3 S2B Variants

To the best of our knowledge, S2B [Garcia-Vaquero et al., 2018] is the only published method that ranks network nodes by their likelihood of belonging simultaneously to two disease modules, using as inputs a graph G and two sets of seed nodes (a and b). These sets should be part of two overlapping connected subgraphs A and B, respectively, which correspond to the network modules of two related diseases. Disease modules are partially known, and the seed nodes are the presently known module members.

S2B is based on betweenness centrality, which depends on shortest paths. To study if these dependencies could be limiting the performance of the method, we developed a set of alternative methods, using the same inputs and producing a similar output. These S2B variants intend to cover different approaches used in network based single-disease gene prioritisation methods, adapting them to the disease module overlap problem. Each of the alternative methods is described in the following subsections.

2.3.1 SN - Specific Neighbors

SN follows an approach analogous to the DIAMoND method [Ghiassian et al., 2015]. It considers that the best candidates to be part of the disease module are nodes which have a set of neighbors specifically enriched in previously known module members. To evaluate this enrichment, it performs an hypergeometric test, selecting the node with the lowest p-value as the best candidate. To adapt this approach to the module overlap prediction, we perform two hypergeometric tests with the set of neighbors of each candidate node, one for each of the two input seed sets ($phyper\_a$ and $phyper\_b$). We then rank all the candidate nodes by the maximum of the two p-values. The nodes with the lower maximum p-value are the best candidates, as they have low p-values for both input seed sets.

2.3.2 SC - Specific Closeness

SC is based on closeness centrality. General closeness centrality is the inverse of the average distance of the node to all other nodes in the network, where distance is defined as the length of the shortest path between two nodes. Instead of computing closeness for the whole network, we compute the inverse of the average distance from the candidate node to each set of seed inputs separately. SC is the product of both inverse average distances. High scoring nodes will have small distances to both input disease modules simultaneously. This method is calculated according to equation 2.1 where a is the set of input seeds a (associated with disease A), b is the set of input seeds b (associated with disease B) and $dist(i,j)$ is the length of the shortest path between nodes i and j.

$$SC(i) = \frac{\|a\|}{\sum\limits_{j \in a} dist(i,j)} \times \frac{\|b\|}{\sum\limits_{j \in b} dist(i,j)} \tag{2.1}$$

2.3.3 SRWC - Specific Random Walk Closeness

SRWC is similar to SC, but the definition of distance between two nodes is based on short random walks (instead of shortest path lengths). SRWC starts by calculating the probabilities with which each node is reached by random walks starting in each input seed set with lengths from 1 to 5. These probabilities are used to compute a weighted average of the length of random walks that reach the node starting from an input seed set. Lower weighted averages will be obtained for nodes with higher probabilities for shorter random walks. Multiplying the inverse of the average random walk lengths of both input sets produces the SRWC score. The equation for this method is 2.2 where $pRW(x,i,l)$ is the probability that a random walk of length l, starting from a node of set x, reaches node i.

$$SRWC(i) = \frac{1}{\sum\limits_{l=1}^{5} l \times pRW(a,i,l) \times \sum\limits_{l=1}^{5} l \times pRW(b,i,l)} \tag{2.2}$$

2.3.4 SRWR - Specific Random Walk with Restart

Random Walk with Restart (RWR) is a successful algorithm applied to single disease gene prioritisation. Given a set of initial nodes, it computes the stationary probability of a node being visited by a random-walker. At each time step of the random walk, the walker may return to one of the initial nodes with a probability defined by a parameter p. A high RWR stationary probability implies that the node is easily reachable from the initial nodes by short paths. This also can be interpreted as a strong relationship between that node and the initial node set, which is the rationale to select good candidates for disease association (if the initial nodes are previously known disease associated genes). In the module overlap prediction problem, we aim to identify nodes that have high RWR probabilities for the two sets of input seeds (RWR_a and RWR_b). To quantify this aim, SRWR is the average of the two RWR probabilities divided by the absolute value of their difference. The calculation of this method is performed by equation 2.3 where $pRWR(x, i, p)$ is the probability that node i is visited by a random walk with restart, starting from a node of set x and with restart probability p.

$$SRWR(i) = \frac{pRWR(a,i,p) + pRWR(b,i,p)}{2 \times |pRWR(a,i,p) - pRWR(b,i,p)|} \tag{2.3}$$

2.3.5 SLB - Specific Length weighted Betweenness

SLB evaluates for each candidate node, the probability that it is part of a random-walk of length l that connects a seed from set a to a seed from set b. This probability is evaluated for L values from 2 to 5. The SLB score is a weighted sum of these probabilities, where higher L probabilities have smaller weights. SLB is an attempt to generalize S2B, using short random walks instead of shortest paths. This method is represented by equation 2.4 where $pRWB(a, b, i, l)$ is the probability that a random walk of length l, starting from a node of set a and ending in a node of set b, passes through node i.

$$SLB(i) = \sum_{l=1}^{5} \frac{pRWB(a,b,i,l)}{l} \tag{2.4}$$

2.4 Directed Network Variants

S2B has been previously adapted to directed networks [Ramos et al., 2019]. Paths in directed networks have to be composed of edges with concordant directions, and consequently, a path from node i to node j cannot be a path that goes from node j to node i. The most straightforward version of directed S2B (S2B_V1) computes specific betweenness by considering shortest paths that go from a seed in a to a seed in b, but also shortest paths going form a seed in b to a seed in a. However, it is also interesting to find nodes that are influenced simultaneously by both seed sets, or that influence both seed sets simultaneously. These types of nodes are the aim of S2B_V2 and S2B_V3, respectively. S2B_V2 prioritizes nodes that frequently connect both seed sets through converging short paths, while S2B_V3 favours nodes that frequently connect both seed sets through divergent paths (Fig. 2.1). In this work we have developed analogous adaptations of SRWR and SLB to directed networks. Only these two methods were adapted to directed networks because they presented better predictive performances (together with S2B) when using undirected networks.

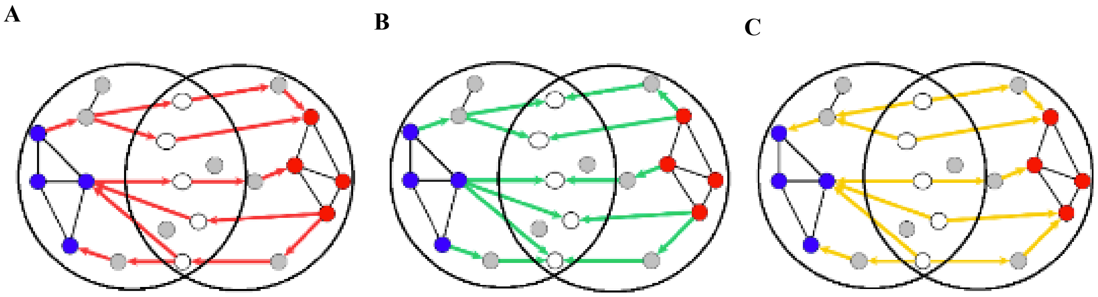

Figure 2.1: Directed S2B versions from [Ramos, 2018]. (**A**) Version 1. (**B**) Version 2. (**C**) Version 3.

The equations to calculate the adaptations for SRWR are represented on equation 2.5 for version 1 where $pRWR(x \rightarrow i + i \rightarrow x, p)$ is the probability of node i being visited by a random walk with restart starting from a node from set x plus the probability of a random walk with restart starting from node i visits a node from set x and with a probability p.

$$SRWR_{v1}(i) = \frac{pRWR(a \rightarrow i + i \rightarrow a, p) + pRWR(b \rightarrow i + i \rightarrow b, p)}{2 \times |pRWR(a \rightarrow i + i \rightarrow a, p) - pRWR(b \rightarrow i + i \rightarrow b, p)|} \tag{2.5}$$

For version 2, the equation that calculates SRWR is 2.6 where $pRWR(x \rightarrow i, p)$ represents the probability of node i being visited by a random walk with restart that started in node from set x with a restart probability of p.

$$SRWR_{v2}(i) = \frac{pRWR(a \rightarrow i, p) + pRWR(b \rightarrow i, p)}{2 \times |pRWR(a \rightarrow i, p) - pRWR(b \rightarrow i, p)|} \tag{2.6}$$

Regarding version 3 the equation for SRWR is presented on equation 2.7 where $pRWR(i \rightarrow x, p)$ is the probability of a random walk to visit a node from set x starting in node i.

$$SRWR_{v3}(i) = \frac{pRWR(i \to a, p) + pRWR(i \to b, p)}{2 \times |pRWR(i \to a, p) - pRWR(i \to b, p)|} \tag{2.7}$$

In adapted SLB version 1, equation 2.8 calculates this method where $pRWB(a \to i \to b + b \to i \to a, l)$ is probability that a random walk of length l starts in a node from set a passes by i to reach a node of set b and also starts at a node from set b to a node from set a passing by i.

$$SLB_{v1}(i) = \sum_{l=1}^{5} \frac{pRWB(a \to i \to b + b \to i \to a, l)}{l} \tag{2.8}$$

Version 2 of SLB is described by equation 2.9 where $pRWB(a \to i \leftarrow b, l)$ is the probability of a random walk of length l starting from a node from set a and another from set b, if it reaches node i.

$$SLB_{v2}(i) = \sum_{l=1}^{5} \frac{pRWB(a \to i \leftarrow b, l)}{l} \tag{2.9}$$

For the third version of SLB, in equation 2.10, $pRWB(a \leftarrow i \to b, l)$ describes the probability of a random walk that starts in node i reach a node from bot sets of a and b.

$$SLB_{v3}(i) = \sum_{l=1}^{5} \frac{pRWB(a \leftarrow i \to b, l)}{l} \tag{2.10}$$

2.5 Performance Evaluation

All the evaluated methods produce a ranked list of nodes as an output. The higher the score (lower rank), the higher is the likelihood that the node is part of the overlap between the two disease modules. The sets of top 50, 100, 200, 300, 400 and 500 candidates (nodes with higher scores) were evaluated through precision and recall metrics. These measures take into account the confusion matrix $\begin{bmatrix} Tp & Fp \\ Fn & Tn \end{bmatrix}$. In which Tp corresponds to true positives (nodes known to be part of the overlap and present in the set of top candidates), Fp to false positives (nodes known to be out of the overlap and present in the set of top candidates), Fn to false negatives (nodes known to be part of the overlap and absent in the set of top candidates) and Tn to true negatives (nodes known to be out of the overlap and absent in the set of top candidates).

Precision (P, eq. 2.11) indicates the fraction of relevant elements from all the retrieved by the method.

$$P = \frac{Tp}{Tp + Fp} \tag{2.11}$$

While recall (R, eq. 2.12) is the true positive rate, that is, the fraction of all relevant elements that are retrieved by the method.

$$R = \frac{Tp}{Tp + Fn} \tag{2.12}$$

Figure 2.2 resumes the workflow presented in the analysis performed in this thesis.

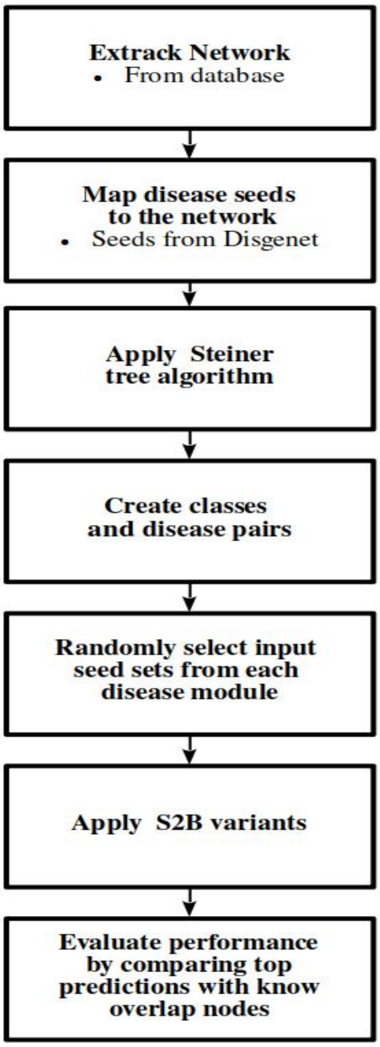
Extrack Network
• From database
Map disease seeds to the network
• Seeds from Disgenet
Apply Steiner tree algorithm
Create classes and disease pairs
Randomly select input seed sets from each disease module
Apply S2B variants
Evaluate performance by comparing top predictions with know overlap nodes

Figure 2.2: Overhall worflow.

Chapter 3

Undirected *versus* Directed Network Comparison

This chapter presents a performance comparison of S2B and variant methods in the undirected and directed versions of a Human signaling/transcriptional regulatory network.

3.1 Introduction

Biological systems can be represented by networks. The information available to build these networks is continuously expanding with the growing number of published studies using both low and high throughput methods to discover molecular interactions. Besides the presence or absence of interactions between two molecules, networks can encode more information through the use of weights (in the edges or in the nodes) or through the definition of a direction for the edges. The specific directions of the interaction can have a substantial impact on the network since it provides functional information [Newman, 2010; Piraveenan et al., 2012]. Some biological networks do not have interaction direction (undirected networks) such as physical protein-protein interaction, while others naturally have it (directed networks) such as signaling and transcriptional regulatory networks [Newman, 2010; Piraveenan et al., 2012; Wang et al., 2014].

One of the main focus of network medicine is the identification of relevant nodes associated with diseases, with various methods being developed to retrieve this information (see [Liu et al., 2019] for a review about different computational methods). Although directed networks are potentially more informative, the majority of algorithms to predict new proteins or genes associated with diseases were developed for undirected networks [Vanunu et al., 2010; Ghiassian et al., 2015; Silberberg et al., 2017; Cáceres and Paccanaro, 2019; Liu et al., 2019]. Directed networks can still be applied, as transforming a directed network into an undirected is simply done by ignoring the directions of edges. However, this approach might introduce some inconsistencies on the network structure [Newman, 2010].

Trying to adapt methods for directed networks is a hard task, due to the high variability of the strategies of existing algorithms. For example, they can be based on centrality measures, connectivity patterns [Ghiassian et al., 2015] or random walks [Köhler et al., 2008]. If the method uses the degree centrality, we have to consider two types of degrees, the in-degree, which is the number of edges pointing to the node, and the out-degree, which is the number of edges exiting from the node [Newman, 2010]. The distinction between these types of degrees are important according to the context. An high out-degree value might be an indicator of a node that is an important regulator on the network [Piraveenan et al., 2012]. For calculations depending on paths and their lengths in directed versions, the path from node A to node B might not be the same path from node B to A. It is even possible that one of the inverse paths might not exist at all [Newman, 2010], adding a layer of complexity that is not present on undirected methods.

Despite these difficulties, exploring the extra information of edge directions can pay off. Wang and collaborators [Wang et al., 2014] proposed a new method to predict important nodes on directed networks using motifs and developed a complementary index score to infer node importance. This method performed better than degree and other topological methods.
Exploratory analysis using S2B directed versions was done by Ramos and collaborators [Ramos, 2018; Ramos et al., 2019], where three directed versions of this method were developed, one that search for undirectional paths across disease modules, another searching for intermediate nodes receiving convergent paths from both disease modules and a last one that searches for intermediate nodes with divergent paths reaching both disease modules.

In this chapter we will test S2B and other variants, both with undirected and directed versions, using the OmD directed network (signalling and transcriptional regulatory interactions retrieved from OmniPath and Dorothea). The aim is to find the best performing variants and to test if they achieve and increased performance through the use of the directed version of OmD.

3.2 Workflow

In this section the procedures specifically used for this chapter are presented. Procedures not presented here were performed as described in chapter 2. Information about the used network and formation of disease classes pairs is present on subsection 3.2.1. The methodology used for directed variants is referred on subsection 3.2.2. Finally, the used statistical test is reported on subsection 3.2.3.

3.2.1 Network

For this chapter we used the OmD network since it is a directed network. To avoid large computational times, only 500 disease pairs were used to evaluate method performance using diseases from class A and B. This number is sufficient to give a representative sampling of the methods performance with these disease classes. The 500 disease pairs were randomly sampled form all possible pairs in each class. For classes C, D and E sampling was not necessary due to the small number of possible disease pairs. The number of disease pairs used for method evaluation from each class are presented on Table 3.1.

Table 3.1: Pairs of diseases per class used to perform the evaluations.

Network	Class A	Class B	Class C	Class D	Class E
OmD	500	500	308	15	122

3.2.2 Directed S2B Variants

S2B and all the undirected variants presented in section 2.4 were evaluated using the undirected version of the OmD network. The best performing methods were then adapted to directed networks. Similarly to the directed versions of S2B, already developed by Ramos and collaborators [Ramos et al., 2019], each of the remaining best performing methods originated three directed versions. Version 1 that searches for shortest paths without defined direction between both diseases (S2B_V1, SLB_V1 and SRWR_V1), version 2 that searches for convergent bi-directed paths starting from both disease modules (S2B_V2, SLB_V2 and SRWR_V2) and version 3 that searches for divergent paths from the overlap (S2B_V3, SLB_V3 and SRWR_V3).

See section 2.3 for implementation details.

3.2.3 Wilcoxon test

Wilcoxon signed-rank test is a non-parametric statistical test. Considering two paired samples of repeated measurement, it tests if the null hypothesis of the samples arise from the same distribution [Wilcoxon, 1945]. The alternative hypothesis used was two-sided. In this case in each repeated measurement we evaluate the performance of a different method (S2B or variant). A small p-value indicates that the methods have statistically different performances.

3.3 Results

3.3.1 Undirected versus Directed network

Since OmD is a signalling network, therefore, naturally a directed network, it was used for a method validation test using the same pairs of diseases in the same network, but in two alternative scenarios: considering or not considering the direction of the edges. Random samples of 100 representative nodes of both disease modules were used has as input, while the known overlap of the complete modules was used to evaluate the method's predictions. Performance was assessed with precision and recall measurements for different cutoffs including the top 50, 100, 200, 300, 400 and 500 candidate nodes retrieved by each method, for each disease pair and across all disease classes. As a control, the same performance measurements were obtained for a random classifier (which generates a ranking of all nodes in the network through a random permutation).

Using the undirected version of OmD, S2B, SLB and SRWR presented the predictions with best precision performance in all classes and cutoffs (Fig. 3.1).
In Class A top 50, the methods S2B, SLB and SRWR had a median precision of 0.54, 0.6 and 0.58 respectively. The SC method had an average performance with 0.34. The remaining methods, SWRC and SN had a median of 0.1 and 0.14. The data emulating a random classifier used as control had a median of 0 and a mean of 0.009. The best precision results were obtained in Class E top 50 where S2B had a median of 0.82, SLB of 0.88 and SRWR had 0.86. In this class the random data had a median of 0.06 and a mean of 0.05. For the remaining classes, with exception of SN, the performance improves with greater module sizes. In class A, all methods perform better than the random classifier, while in class E, SN performs worse than the random control. As expected, precision decreases with less stringent cutoffs for all methods, except for the random classifier. This tendency confirms that the proposed methods tend to concentrate nodes that are part of the disease modules overlap in the top of their rankings.

To check if the methods are retrieving a significant fraction of all known overlap nodes we used the recall metric. The higher values were obtained in class A top 500 with a median of 0.54 for S2B, 0.62 for SLB and 0.58 for SRWR (Fig. 3.2). Again, all methods perform better than the random classifier, except for SN with the E class. Contrary to precision, recall increases with less stringent cutoffs, as a larger number of candidates favours higher recall, even for the random classifier. All boxplots regarding undirected precision and recall data are available on the supplementary data (Figures S1 and S2).

The best three methods in the undirected version were selected to be adapted to directed versions. Compared with the undirected version the directed results regarding precision had a significant decrease in performance. The best methods for this version were S2B_V1, SLB_V1 and SLB_V3 with a median of 0.42, 0.34 and 0.42 for class A top 50 and a median of 0.76, 0.52 and 0.6 for class E top 50, respectively

(Fig. 3.1).
Similarly with the methods behaviour from undirected tests, the methods had better performance with the module size increment, except for all SRWR versions that start to decrease on class C and further (see figures in appendix S2).

In terms of recall, the methods adapted to directed networks also show a lower performance when compared with the respective undirected versions (Fig. 3.2).

A non-parametric Wilcoxon test was used to infer the differences among all three top methods on top 50 Class A and Class E precision and recall, for undirected and best directed version of each methods. According to tables 3.2 and S1, which resumes all the statistical tests, the best methods of undirected version (S2B, SLB and SRWR) when compared with each others present a significant p <0.001, indicating that the methods are statistically different. The same methods when paired with their analogous best directed version also lead to p <0.001 (except for S2B compared with S2B_V1 on Class E for precision, where p is 0.041) indicating that the methods performances are statistically different. This statistical results showed that S2B, SLB and SRWR are different from each other, and each one them is better that their corresponding directed versions.

Table 3.2: Wilconxon test based on precision resume between pairs of methods.

		Top 50 Class A			Top 50 Class E		
Method A	Method B	Median A	Median B	p-value	Median A	Median B	p-value
S2B	SLB	0.54	0.6	<0.001	0.82	0.88	<0.001
S2B	SRWR	0.54	0.58	<0.001	0.82	0.86	<0.001
SLB	SRWR	0.6	0.58	<0.001	0.88	0.86	0.041
S2B	S2B_V1	0.54	0.42	<0.001	0.82	0.76	<0.001
SLB	SLB_V3	0.6	0.42	<0.001	0.88	0.6	<0.001
SRWR	SRWR_V1	0.58	0.12	<0.001	0.86	0.08	<0.001

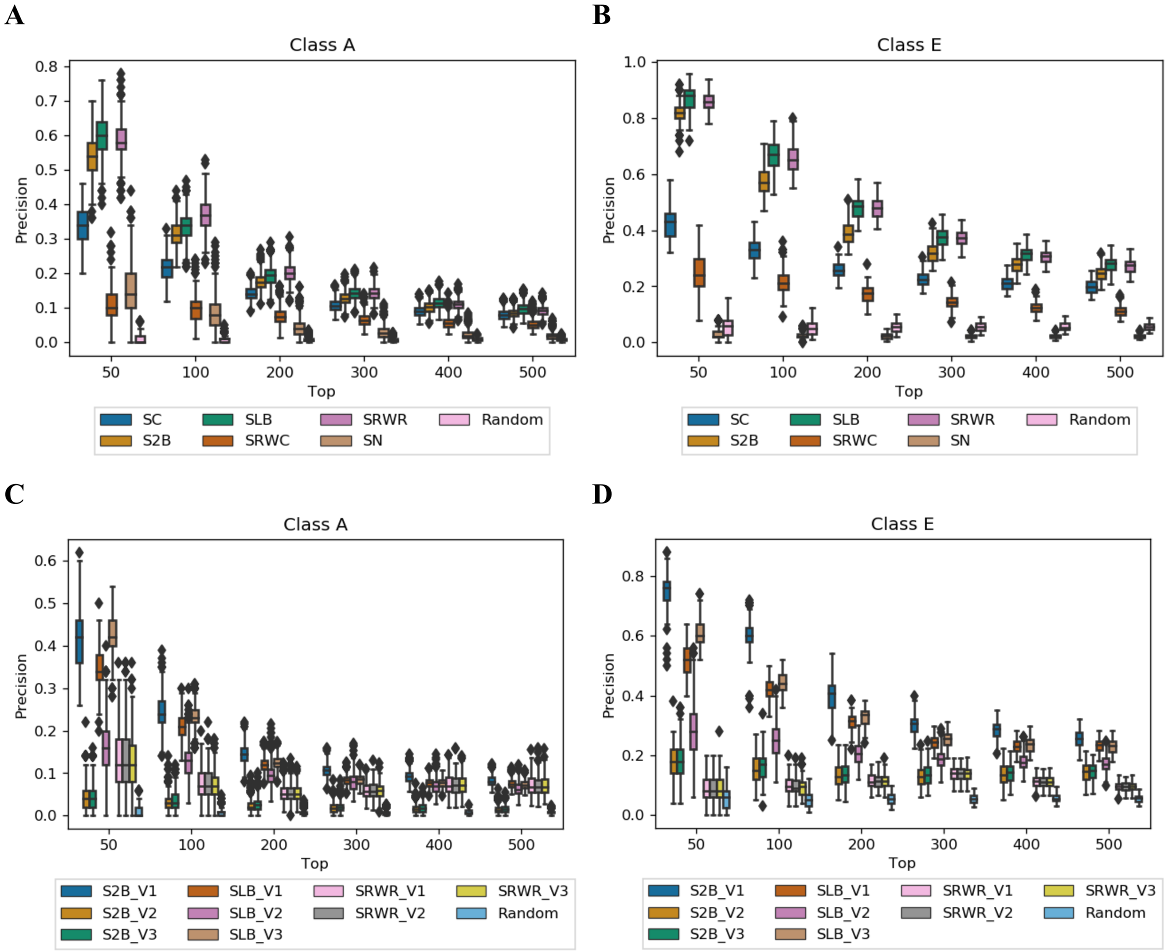

Figure 3.1: Methods precision distribution per tops on all versions. (**A**) Class A undirected version. (**B**) Class E undirected version. (**C**) Class A directed version. (**D**) Class E directed version

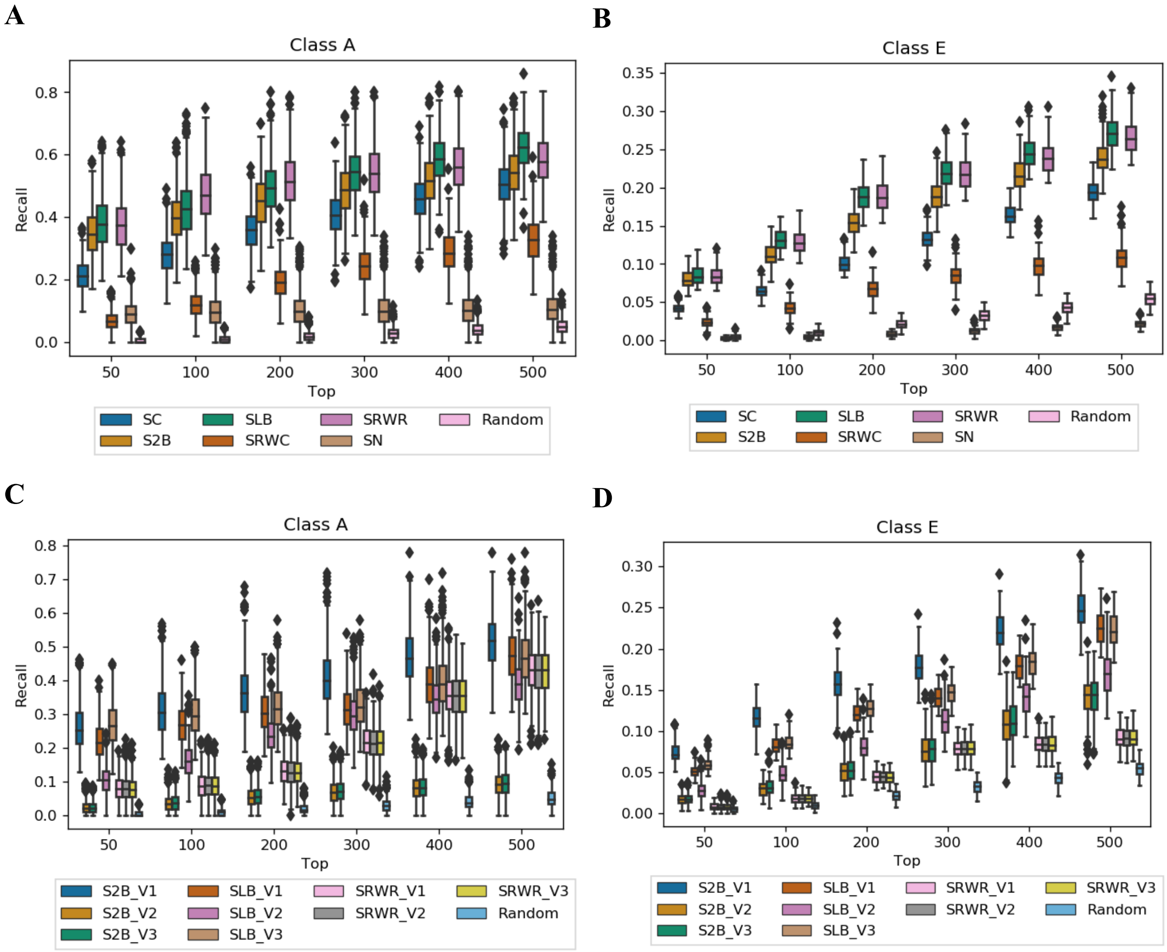

Figure 3.2: Methods recall distribution per tops on all versions. (**A**) Class A undirected version. (**B**) Class E undirected version. (**C**) Class A directed version. (**D**) Class E directed version

3.4 Discussion

The hypothesis explored on this chapter was that the use of directed networks, by encoding more information about the system, might lead to better predictive performance. Surprisingly, our results show an opposite effect. In the comparison of undirected versus directed network methods, we obtain better precision scores with the undirected versions. Since the methods compared (S2B, SLB and SRWR) are based on shortest paths and random walks, when applied with directed networks, they depend on paths with coherent directions. This requirement reduces the number of available paths to be explored by the methods when comparing with the undirected network scenario. This might be the reason for the lower performance in directed versions.

Although, the directed versions were not an improvement over the undirected ones, the three best methods are still able to provide useful information about candidate overlap nodes . The best methods on the directed version were S2B_V1, SLB_V1 and SLB_V3, being the first two, methods that search for unidirectional paths connecting both disease modules (but the paths can go either from A to B or from B to A). The last one searches for divergent paths from the overlap that reach both disease modules. Ramos and collaborators [Ramos, 2018; Ramos et al., 2019] used artificial modules to test the directed S2B's versions and verified that these methods had the potential to prioritize nodes with specific roles in the disease module, such as modifiers or causal nodes, since they capture the direction of cause-effect relationships between nodes. SLB_V3 being one of the top methods can generate predictions enriched in modifier genes. The fact that it performs better may suggest that these modifier genes are over-represented in disease module overlaps.

An parallel result from this work is the comparison of S2B and its variant methods. Regarding the undirected versions, the best methods were S2B, SLB and SRWR. These methods are robust since they kept their superior performance across different cutoffs and module size classes, reaching a median precision greater than 80% on Class E. Among the three methods, SLB and SRWR are slightly better, both in terms of precision and recall. The remaining methods had a poor performance compared with the previous ones, although some, such as the SN approach, have been successfully applied to single disease gene prioritization [Ghiassian et al., 2015].

Across all methods, there is a general trend of decrease in precision and increase in recall as the module sizes increase. This trend maybe caused by the fact that we use a constant size for the input seed sets for each disease, while the known disease module size is varying. The known overlaps of bigger disease modules are also bigger. The greater abundance of positive nodes to classify favours higher precisions and lower recalls. The application of the methods to disease pairs of class E is possibly analogous to a situation were the present knowledge of both disease modules is only a small fraction of the real disease modules. In this scenario, we expect that our best methods will have high precision but lower recall. On the other hand, our tests with class A pairs resembles the situation where the two diseases under study are very well known, and the input seed sets cover a large fraction of the real disease modules. In this scenario, we expect that our methods will have lower precision but higher recall.

Since S2B was the first method developed to predict nodes in the overlap between two disease modules, there are no available benchmark tests to directly compare our performance results. But for single disease gene prioritizations, there are several reports available. In a recent benchmark study carried by Picart-Armada and colleagues [Picart-Armada et al., 2019], evaluating various methods for disease target prioritization, the method with better performance retrieved between 11.58 and 11.98 hits in the top 20, corresponding to a precision between 0.579 and 0.599. Comparing this results with our methods using the undirected version and diseases pairs from class A (which has lower precision), the best three methods, for a top 20 candidate cutoff, have precision scores of 0.9 for S2B and 0.95 for SLB and SRWR. This raises the possibility of using our overlap prediction methods to prioritize new single disease genes.

In conclusion, the best performing methods for the prediction of disease module overlap genes are SLB and SRWR, closely followed by S2B, all of them applied to undirected networks.

Chapter 4

Evaluation of method robustness

In chapter it is presented the evaluation of the robustness of performance of the methods.

4.1 Introduction

To completely evaluate a prediction method performance, besides the determination of precision and recall measures, it is also necessary to evaluate how those measures are robust against variations in the used inputs. These variations can be due to different data sources or random errors in data collection. To use the developed methods with confidence it is important to ensure that the observed performance is not only valid for a particular dataset or conditions.

To evaluate S2B and variant methods, random samples of known disease modules are used as input. Some methods may be differently affected by changing the particular random sample that is used, or by using a different number of seeds as inputs. The performance of these methods can also be influenced by the characteristics of the disease modules being considered. In chapter 3 it was already observed that methods performance increased when larger disease modules where tested. As referred in the discussion section of chapter 3 (section 3.4), using larger modules with a constant seed set size is analogous to apply the methods to a situation where only a small fraction of the complete disease module is known. However, method performance was not evaluated when the two diseases used as inputs have significantly different sizes. An additional module property that may influence the quality of predictions is the connectivity among its members. This property can be quantified in diverse ways. Here, we consider that a module is has a high connectivity if the relative number of nodes that one has to add to known disease associated genes in order to get a connected subgraph is small. The evaluated methods assume that disease modules are composed of closely interacting disease associated genes, therefore it may be expected to reach higher performances with disease modules of higher connectivity.

In this chapter we evaluate the robustness of the performance of S2B and variant methods to variations in the referred input properties.

4.2 Workflow

The methodology used in this chapter is described below. Procedures not presented here were performed as described in chapter 2. The methodology for the creation of heterogeneous class is explained in subsection 4.2.1. For the replicate runs the used procedures are described in subsection 4.2.2. Lastly in this section, the connected disease modules are reported on subsection 4.2.3.

4.2.1 Variation of complete module sizes

The heterogeneous Class AE, was created by overlapping diseases with 200 to 400 nodes in the network with diseases with more than 1000 nodes on the networks. So, for each disease pair one of the disease modules is between 200 to 400 and the other disease has to have more than 1000. This new class has a total of 666 disease pairs.

4.2.2 Variation of input seed sets

Besides the pairs of diseases with heterogeneous sizes, this chapter also used the same disease pairs used in subsection 3.2.1. To evaluate the robustness of method performance to random changes in input seed sets, five independent replicate samplings of 50, 100 or 150 seeds from each disease were used as inputs for the evaluation of the different methods.

4.2.3 Variation of Disease Module connectivity

All the disease modules used in this work resulted from a minimal expansion of the original disease associated gene sets through the minimum Steiner tree algorithm (see section section 2.2 of General Methods chapter). The connectivity of a disease module was quantified as the number of nodes added by the minimum Steiner tree expansion divided by the original number of disease associated genes.

4.3 Results

Since in the previous chapter, the best results were obtained with undirected versions, in this chapter we focus only on these versions, testing the robustness of their performance when several input properties are varied.

4.3.1 Variation of complete module sizes

To understand if pairs with significantly different module sizes would influence performance, a new heterogeneous Class AE was created. This class uses pairs with a disease from Class A and other from Class E. The resulting scores appear to be between the values of both original classes, with a precision of 0.62 for S2B, 0.72 for SLB and 0.68 for SRWR (top 50 candidates). As for the recall values S2B has 0.42, SLB 0.5 and SRWR 0.46 (for top 500 candidates) (Fig. 4.1) The comparison with the use of disease pairs of homogeneous size is presented in the representative 4.2, showing the precision values of the S2B method for the different classes of input disease pairs. As previously discussed, as the number of input seeds is kept constant, complete module size correlates with the lack of knowledge about a particular disease module. Therefore, these intermediate performances suggest that searching for the overlap between diseases with different degrees of knowledge completion does not compromise the quality of the results more than expected due to the use of a more deeply studied disease.

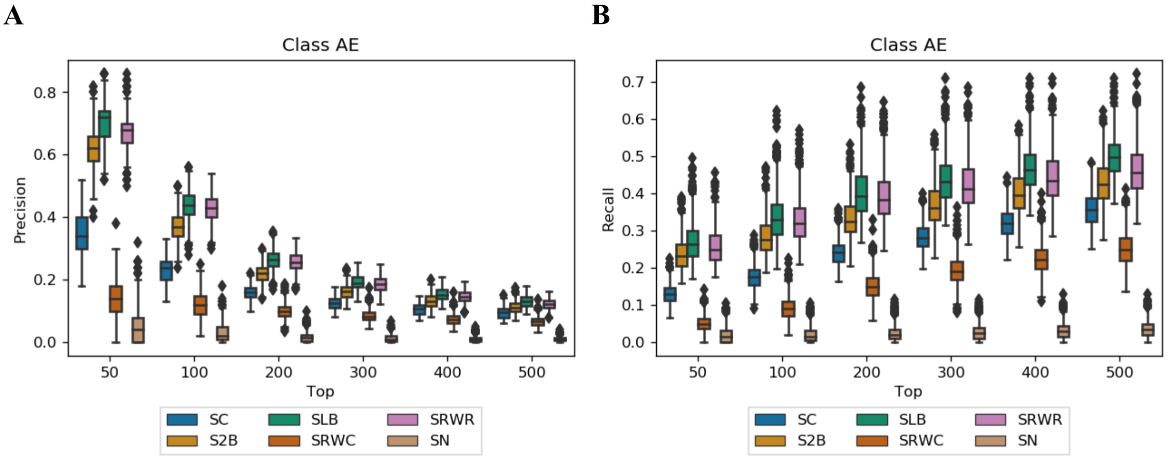

Figure 4.1: Class AE metrics boxplots across the different methods and tops on the undirected version. (**A**) Precision. (**B**) Recall.

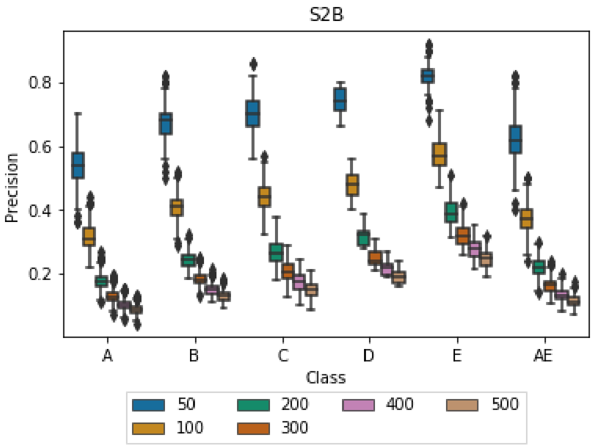

Figure 4.2: S2B Precision values for the top candidates across size classes.

4.3.2 Variation of input seed sets

To further validate if the methods are robust relatively to random changes in the input seed sets, five replicates for groups of 50, 100 and 150 initial seeds were performed.

The results show that the standard deviation of replicate precision values is relatively small across all methods, tending to decrease with the number of initial seeds (Fig. 4.3). Considering a larger number of top candidates as positives (predicted to be part of the overlap between the two disease modules) also stabilizes the method performances. Among the best performing methods, SRWR is the more robust to random changes in input seeds, followed by SLB and S2B.

4.3.3 Variation of disease module connectivity

To infer if the connectivity of the disease modules could influence the precision, a distribution of the average percentage of nodes added by the minimum Steiner tree was correlated with the respective precision of method predictions for each disease pair.

The average percentage of added nodes ranged from 10 to 26 on Class A, 8 to 15 in Class B, 6 to 12 in Class C, 6 to 7.5 in class D and finally 4 to 8 in Class E. These analysis were performed in the three best methods S2B, SLB and SRWR of the undirected version. All plots are presented in the supplementary materials (Figures S3, S4 and S5).

No strong correlation was observed. The average trend, evaluated through a linear regression adjusted

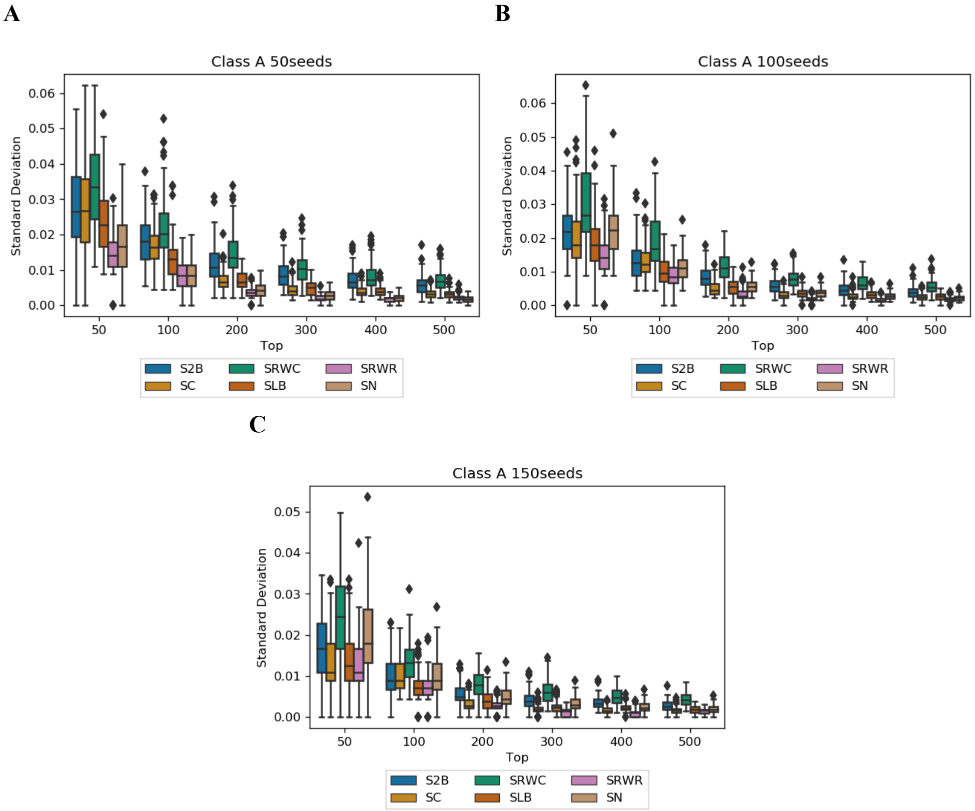

Figure 4.3: Standard variation of precision values obtained with independent random sampling of input seed sets for disease pairs of class A.

to each scatter plot, generally presented low slope values, changing sign across different disease classes. This suggests that, overall, there is no correlation between the disease module connectivity's and the resulting method precisions. This analysis may also assure us that the nodes introduced by the minimum Steiner tree expansion are not significantly affecting the methods precision.

4.4 Discussion

In this chapter, S2B and variant methods were applied under varying scenarios. First, it was possible to observe that applying these methods to disease modules of heterogeneous sizes does not introduce a perturbation greater than expected by changing the module size of both diseases in the pair. The obtained performances are between the ones expected for homogeneous pairs of both sizes. Considering that the ratio between the complete module size and the number of seeds used as input is a measure of the degree of knowledge about the disease, the evaluated methods can be applied to disease pairs with levels of knowledge (one that is very well studied with other that is not well characterized). The expected performance should be higher than if two well studied diseases are used, but lower than if two understudied diseases were being analysed.

Second, the variability in performance scores due to random variation in the input seeds was very low. For the best methods (S2B, SLB, SRWR), the precision standard deviations were between 2% and 4% of its values (evaluated with 50 top candidates, which have the higher standard deviations, and with class A diseases, which have the lower precision scores). As expected, evaluating precision for a larger number of top candidates or using a larger number of seeds as inputs decreased the variability of the estimates.

Finally, the connectivity of the disease modules had a negligible effect on the precision of the best performing methods. These results show that by including new nodes in the disease modules with the minimum Steiner tree method does not seem to include a bias in the analysis. In fact, a previous work that tries to complete disease modules, uses a similar approach [Wang and Loscalzo, 2018]. They use the Seed Connector Algorithm which tries to find modules using seed proteins by adding the minimum possible extra nodes that are linked to a maximum number of nodes already in the module. They validated this method with 70 diseases and were able to demonstrate that the added nodes had biological relevance regarding their function when compared with their seed proteins.

In conclusion, S2B and variant methods have a robust performance, easily coping with noise in the input seeds or with diverse disease module properties.

Chapter 5

Comparison of methods performance across different undirected networks

This chapter presents the comparison of the methods performance on three undirected networks: protein-protein physical interaction network, on a signalling network and on a parsimonious composite network

5.1 Introduction

The origin and type of data used to build a network have a considerable impact on how informative that network is. The input data might have bias on the type interactions or nodes, not covering all existing and functional interactions that occur in biological systems.

In chapter 3, S2B and variant methods for the prediction of nodes in the overlap between two network disease modules were evaluated. It was possible to observe that these methods have better performances with an undirected version of a signalling network, which is originally directed. But there are multiple undirected biological networks that can be used instead. Each different network may focus in a different type of molecular interaction (physical interactions, co-expression, among others). Therefore, it is relevant to ask if the performance of these methods is similar across distinct networks. Network size and topology play a critical role for the efficiency of certain methods that predict genes associated with diseases [Huang et al., 2018; Hwang et al., 2019]. Therefore, it is important to select a network that represents biological systems as completely and accurately as possible, but that it is also adequate for the methods being used.

Huang and collaborators [Huang et al., 2018] developed a benchmark for the selection of molecular networks for human disease research. In their work they tested twenty-one networks and they noticed that larger networks achieve the best performance. They also developed an efficient parsimonious composite network (PCNET) that requires that each interaction is present at least in two of the networks. HumanNet v2 [Hwang et al., 2019] is another network developed to improve the predictive power of disease genes. This network integrates interactions from four distinct types such as co-citation, interolog, functional and protein-protein physical interactions. Both these networks were reported to have an equal or better performance than Search Tool for Retrieval of Interacting Genes/Proteins (STRING), which is reported as the best in recovering disease genes from the literature [Huang et al., 2018; Hwang et al., 2019]. STRING [Szklarczyk et al., 2019] is a well-known database that provides protein - protein interaction information combining curated experimental evidence as well as predicted interactions, computing for each interaction a confidence score according to the type of information supporting the interactions.
The methods studied in this thesis are also susceptible to be affected by the type of network used. To assess to what degree their performance is affected, S2B and S2B variant methods where tested using a signalling networks, protein-protein physical interaction and a composite network.

5.2 Workflow

This section covers the methodology used in this chapter. Subsection subsection 5.2.1 describes the used networks and respective characteristics, while subsection 5.2.2 the process of disease mapping, class creation and evaluation metrics.

5.2.1 Networks

The APID, OmD and PCNET networks were used to compare the performance of S2B and variant methods. These networks represent protein-protein physical interactions (PPI), signalling and a composite networks respectively. PCNET is the biggest network with 2693109 interactions, followed by APID with 135834 interactions and lastly OmD with less interactions of all covering 32442 interactions. Their characteristics regarding number of nodes and edges are summarized on Table 5.1. Detailed information about these networks is available on section 2.1.

5.2.2 Computational Workflow

The workflow performed in this chapter is defined in section 2.2. To directly compare the results obtained, exactly the same disease pairs were used as inputs for the three networks. This has limited the number of

Table 5.1: Network characteristics.

Network	Nodes	Edges	Type
APID	15706	135834	Protein - Protein physical interaction
OmD	9228	32442	Signalling
PCNET	18820	2693109	Parsimonious Composite

disease pairs available, since each disease pair should comply with the module and overlap size conditions simultaneously in the three networks. Classes A, B and E had disease pairs common to the three networks. But for the remaining classes C and D there were not common disease pairs in the three networks. For these classes, disease pairs were created with all diseases mapped on that class size for each network independently. The final numbers of disease pairs per class on each network used for this chapter are shown on Table 5.2.

The outputs were subject to the performance evaluation metrics as described in section 2.5.

Table 5.2: Pairs of diseases per class used to perform the evaluations.

Network	Class A	Class B	Class C	Class D	Class E
APID	213	116	207	185	112
OmD	213	116	308	15	122
PCNET	213	116	298	206	122

5.3 Results

To understand if the type of network has an impact on performance, S2B and variant methods were tested using a protein-protein physical interaction (APID), signalling (OmD) and composite (PCNET) networks. The predictions where evaluated in terms of precision (Fig.5.1) and recall (Fig.5.2). For both parameters, the OmD network allowed the methods to achieve better performances. APID and PCNET produced similarly lower performances, although PCNET showed more variable precisions (across disease pairs) and higher recall values for disease pairs from class A.

Consistently for all networks, number of top candidates considered and module size classes, S2B and SLB where among the best performing methods. SRWR showed also good performances for the APID and OmD networks, but the quality of predictions decayed for PCNET, particularly for disease pairs of class A. SC was close to the best performing methods for the APID and PCNET networks, but got worse results with the OmD network. On the opposite end of the performance scale, SRWC has a poor performing method for all networks. SN was also a poor performing method, except for the PCNET network, specially with diseases from class A. All networks shared the common trend of increasing precision and decreasing recall from class A to class E. All the remaining result for precision and recall for all classes in the three networks are available on the supplementary results (Figure S6, S7 and S8)

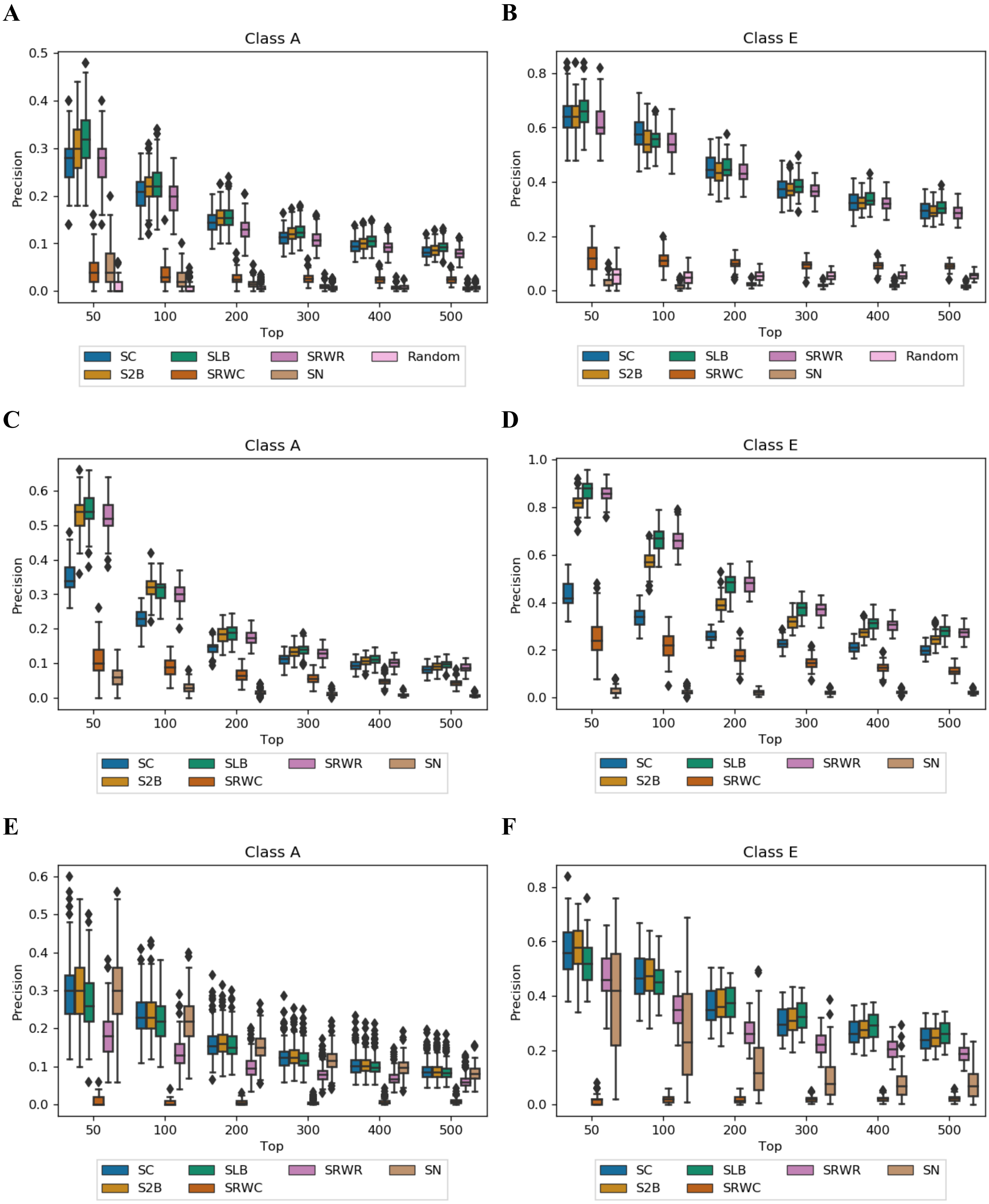

Figure 5.1: Precision scores for class A and E in all networks. (**A**) APID class A. (**B**) APID class E. (**C**) OmD class A. (**D**) OMD class E. (**E**) PCNET class A. (**F**) PCNET class E.

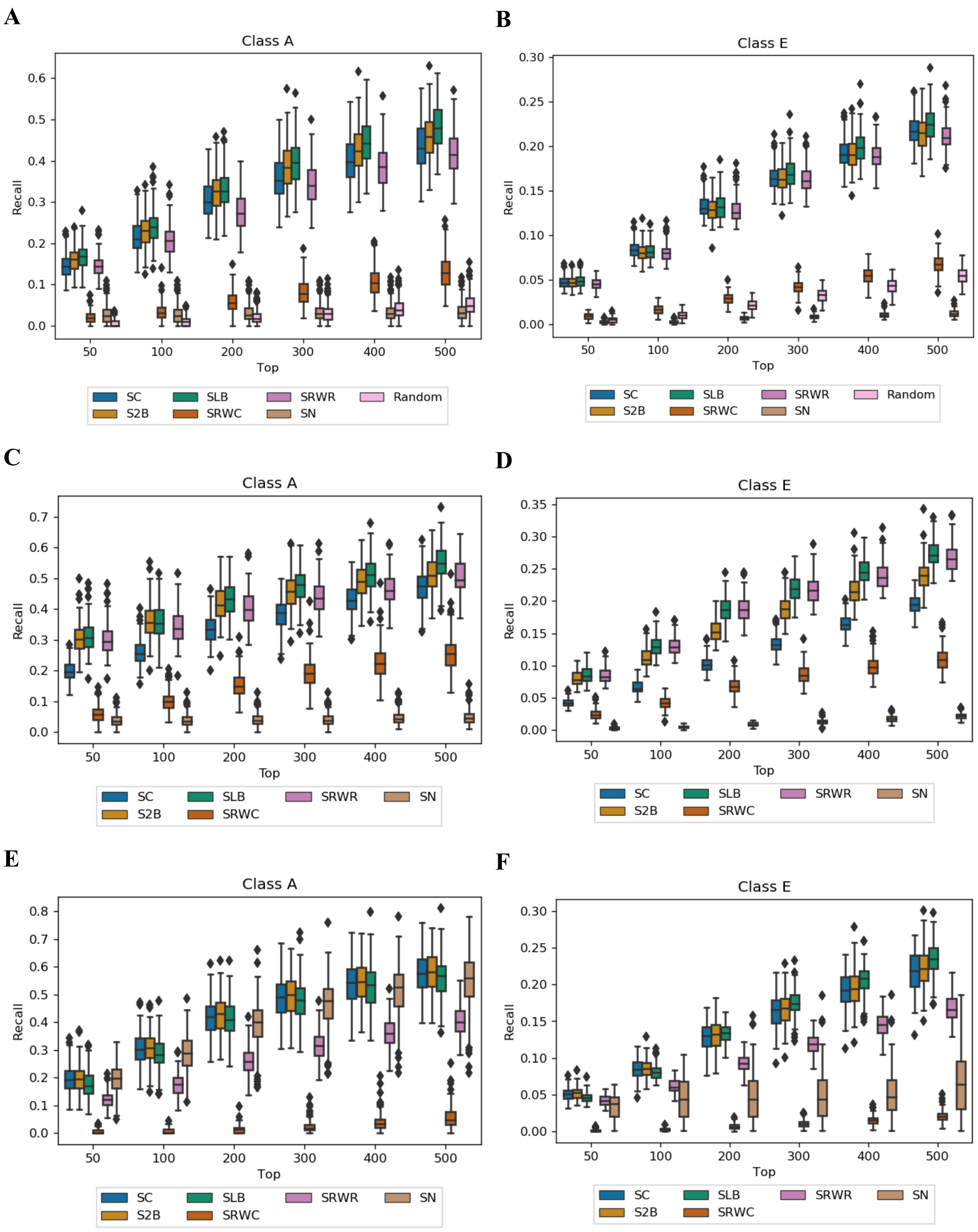

Figure 5.2: Recall scores for class A and E in all networks. (**A**) APID class A. (**B**) APID class E. (**C**) OmD class A. (**D**) OMD class E. (**E**) PCNET class A. (**F**) PCNET class E.

5.4 Discussion

Networks built from heterogeneous data sources are expected to offer a better basis for disease gene prediction than a network built with a single data type. This effect has been observed in a large benchmark comparing the capacity of different networks to recover disease associated genes through a network propagation algorithm [Huang et al., 2018]. However, our results do not agree with these findings, as we have observed better results with a signalling network than with a composite network integrating interactions from several heterogeneous sources. This discrepancy may be explained by the difference between methods that try to expand the list of disease associated genes for a single disease and methods that search for nodes in the overlap between two disease modules. In contrast with single disease methods, signalling pathways may be significantly more informative for overlap prediction methods

The good performance of S2B and SLB across networks indicates that, even with different network topologies, betweeness centrality is a powerful measure for gene prediction and can deliver medium to medium-high precision scores, being a more generally applicable method, although with lower performance on APID and PCNET. Although SLB does not consider shortest paths exclusively, it attributes greater weight to shorter random walk paths connecting both disease modules, using a strategy that is conceptually similar to betweenness. SLB is more flexible than S2B, by giving some score to nodes that are not part of shortest paths, but nevertheless are part of short bridges between the two input disease modules. This flexible strategy leads to slight but consistent performance improvements while using the APID and OmD networks, but introduces misleading noise when PCNET is used. Through the inclusion of different interaction types within the same network, pathways in the PCNET may not always reflect functionally coherent pathways. The SC was also one of the best methods for APID and PCNET networks. Since this method is based on closeness it indicates that in these networks the target nodes present on the overlap are well discriminated by lower network distances to both disease modules.

In PCNET one of the best methods was SN. This method has performed poorly on the others networks. Additionally, PCNET was the only network where the best three methods were not the same for the two module size classes, being SN replaced with SLB in CLass E. This may suggest that when looking at disease modules that are well covered by present knowledge, overlap nodes are enriched in interactions with both disease modules simultaneously. This enrichment is not as evident for diseases with less well studied disease modules. The inclusion of a broader set of interaction types in PCNET may also favour the statistical detection of these interaction enrichments.

In summary, the developed methods seem to predict overlap disease genes more effectively with the signaling network rather than with protein-protein interaction or with the composite network. This suggests that signaling networks are more informative about the common mechanisms shared by different diseases.

Chapter 6

General Discussion

Network medicine is a growing research field that can make important contributions to the development of precision medicine and personalized drug treatments. With the increasing availability of high throughput technologies and the resulting information rich datasets, it is necessary to developed tools capable to efficiently extract useful knowledge from the accumulated observations. With this in mind, new methods to predict disease related knowledge from biological networks were conceived in this dissertation.

The S2B method was previously developed by the host research group. It aims to predict proteins in the overlap between two disease network modules. Its performance was originally evaluated with artificial disease modules. The S2B method was also previously adapted to be applied with directed networks. The work presented in this dissertation advances the referred previous work in several perspectives. First, it develops novel methods to predict proteins in the overlap of disease modules, using strategies that are similar to S2B but trying to score nodes in a more flexible manner (SLB and SRWR), or strategies that have been applied in single disease gene prioritisation methods (SC, SRWC, SN). Second, it evaluates the predictive performance of all these methods using real network disease modules instead of artificial modules. Fractions of the known disease associated genes are used as inputs and the known overlap between the disease modules is used to evaluate and validate predictions. Third, directed and undirected versions of the methods are evaluated using the same signalling network. This comparison allows the objective assessment of the impact of using edge direction information, and the capacity of the directed versions of the methods to capture that extra information layer. Fourth, the different methods are tested in three distinct networks containing different types of interactions between genes and proteins. Finally, the robustness of the prediction performance was evaluated relatively to input factors such as the size of the complete modules, their original connectivity, random variation of selected input nodes and the number of input nodes used.

In chapter 3 we compared how the methods behave using the same network but using the edge directions or using an undirected version of the network. The undirected version obtained the best results achieving a median precision of 0.86 on Class E top 50 for the SRWR method. The results in this chapter suggest that, although directed network are more informative, the requirement of direction coherent paths linking both disease modules reduces the total number of paths used to score candidates. This lower number may compromise the methods performance.

Both in chapter 4 and in chapter 3 it was clear that the factor with a greater impact on the methods performance was the size of the modules. As the number of input seeds was constant, the effect of the module size is related with how completely the disease module is known. Better results are obtained when the ratio of input seeds to complete module is lower. On the other hand, chapter 4 shows that prediction performances are quite stable to changes in module connectivity, number of input seeds or random variations of the input seeds.

For the last chapter, chapter 5, three distinct undirected networks were tested to see in which one the methods would perform better, one representing protein-protein physical interaction, other signalling and the last one a parsimonious composite of various networks. The methods performed better on the OmD network (signalling). It was also observed that some methods were better on some types of networks and not in others (SC, SN and SRWR), while S2B and SLB achieve a good performance on all networks.

Overall it was demonstrated that S2B and some variants have a good performance on real disease modules, particularly using an undirected version of a signalling network and studying diseases with network modules that are not well covered by present knowledge.

Chapter 7

Future Prospects

The work presented in this dissertation achieved its main aims, developing new S2B variant methods and evaluating their performance in multiple conditions. Nonetheless, some enhancements and supplementary approaches might be pursued to improve our ability to predict proteins involved in two diseases simultaneously.

One approach may be using more informative networks. One example is the STRING network, which is reported in some studies to achieve better performance in retrieving disease associated genes [Huang et al., 2018; Hwang et al., 2019]. Alternatively, it is possible to adapt the methods to multiplex networks with several layers. Each layer contains one type of interaction for the same genes or proteins and the layers complement each other giving a more complete representation of the biological systems, which may allow improved performance for our methods.

Confidence in the predictions would also increase if candidate nodes are validated in terms of functional annotations, that should be coherent with the known molecular pathological mechanisms of diseases under study. Further validation could be achieved by comparing predictions with independent databases of gene disease associations, such as Open Targets [Carvalho-Silva et al., 2018], or applying the methods with input seeds that were characterized up to a defined date, and evaluating the predictions by comparing with genes associated with disease after such date.

Finally, after defining which are the best methods with the respective best networks, novel predictions should be analyzed to expand our knowledge of disease molecular mechanisms, helping to explain disease comorbidities or to find new opportunities for drug repurposing.

References

Alonso-López, D., Campos-Laborie, F. J., Gutiérrez, M. A., Lambourne, L., Calderwood, M. A., Vidal, M., and De Las Rivas, J. (2019). Apid database: redefining protein–protein interaction experimental evidences and binary interactomes. *Database*.

Alonso-Lopez, D., Gutiérrez, M. A., Lopes, K. P., Prieto, C., Santamaría, R., and De Las Rivas, J. (2016). Apid interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic acids research*, 44(W1):W529–W535.

Carvalho-Silva, D., Pierleoni, A., Pignatelli, M., Ong, C., Fumis, L., Karamanis, N., Carmona, M., Faulconbridge, A., Hercules, A., McAuley, E., Miranda, A., Peat, G., Spitzer, M., Barrett, J., Hulcoop, D. G., Papa, E., Koscielny, G., and Dunham, I. (2018). Open Targets Platform: new developments and updates two years on. *Nucleic Acids Research*, 47(D1):D1056–D1065.

Cáceres, J. J. and Paccanaro, A. (2019). Disease gene prediction for molecularly uncharacterized diseases. *PLOS Computational Biology*, 15(7):1–14.

Dozmorov, M. G. (2018). Disease classification: from phenotypic similarity to integrative genomics and beyond. *Briefings in bioinformatics*.

Garcia-Alonso, L., Iorio, F., Matchan, A., Fonseca, N., Jaaks, P., Peat, G., Pignatelli, M., Falcone, F., Benes, C. H., Dunham, I., Bignell, G., McDade, S. S., Garnett, M. J., and Saez-Rodriguez, J. (2018). Transcription factor activities enhance markers of drug sensitivity in cancer. *Cancer Research*, 78(3):769–780.

Garcia-Vaquero, M. L., Gama-Carvalho, M., De Las Rivas, J., and Pinto, F. R. (2018). Searching the overlap between network modules with specific betweeness (s2b) and its application to cross-disease analysis. *Scientific reports*, 8(1):1–10.

Ghiassian, S. D., Menche, J., and Barabási, A.-L. (2015). A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol*, 11(4):e1004120.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690.

Huang, J. K., Carlin, D. E., Yu, M. K., Zhang, W., Kreisberg, J. F., Tamayo, P., and Ideker, T. (2018). Systematic evaluation of molecular networks for discovery of disease genes. *Cell systems*, 6(4):484–495.

Hwang, S., Kim, C. Y., Yang, S., Kim, E., Hart, T., Marcotte, E. M., and Lee, I. (2019). Humannet v2: human gene networks for disease research. *Nucleic acids research*, 47(D1):D573–D580.

Ingalls, B. P. (2013). *Mathematical modeling in systems biology: an introduction*. MIT press.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949–958.

Liu, C., Ma, Y., Zhao, J., Nussinov, R., Zhang, Y.-C., Cheng, F., and Zhang, Z.-K. (2020). Computational network biology: Data, models, and applications. *Physics Reports*, 846:1–66.

Liu, X., Hong, Z., Liu, J., Lin, Y., Rodríguez-Patón, A., Zou, Q., and Zeng, X. (2019). Computational methods for identifying the critical nodes in biological networks. *Briefings in Bioinformatics*, 21(2):486–497.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601.

Newman, M. (2010). *Networks: An Introduction Oxford Univ*. Press.

Picart-Armada, S., Barrett, S. J., Willé, D. R., Perera-Lluna, A., Gutteridge, A., and Dessailly, B. H. (2019). Benchmarking network propagation methods for disease gene identification. *PLoS computational biology*, 15(9):e1007276.

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L. I. (2016). Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, page gkw943.

Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., and Furlong, L. I. (2015). Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*.

Piraveenan, M., Prokopenko, M., and Zomaya, A. (2012). Assortative mixing in directed biological networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1):66–78.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6):1389–1401.

Ramos, I. F., García-Vaquero, M. L., Gama-Carvalho, M., and Pinto, F. R. (2019). Cross disease network analysis. In *2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*, pages 1–4. IEEE.

Ramos, I. F. F. (2018). Cross disease network analysis. Master's thesis, Faculdade de Ciêcias, Universidade de Lisboa.

Silberberg, Y., Kupiec, M., and Sharan, R. (2017). Gladiator: a global approach for elucidating disease modules. *Genome medicine*, 9(1):1–14.

Sonawane, A. R., Weiss, S. T., Glass, K., and Sharma, A. (2019). Network medicine in the age of biomedical big data. *Frontiers in Genetics*, 10:294.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., et al. (2019). String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613.

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). Omnipath: guidelines and gateway for literature-curated signaling pathway resources. *Nature methods*, 13(12):966.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol*, 6(1):e1000641.

Wang, P., Lü, J., and Yu, X. (2014). Identification of important nodes in directed biological networks: A network motif approach. *PLOS ONE*, 9(8):1–15.

Wang, R.-S. and Loscalzo, J. (2018). Network-based disease module discovery by a novel seed connector algorithm with pathobiological implications. *Journal of molecular biology*, 430(18):2939–2950.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Supplementary Results

To complete the results presented on 3.1 and 3.2 on chapter 3, figure S1 represents the remaining classes used were the methods were used with the respective precision and recall for the undirected version and figure S2 for the directed version. Table S1 shows the Wilcoxon results regarding the recall scores.

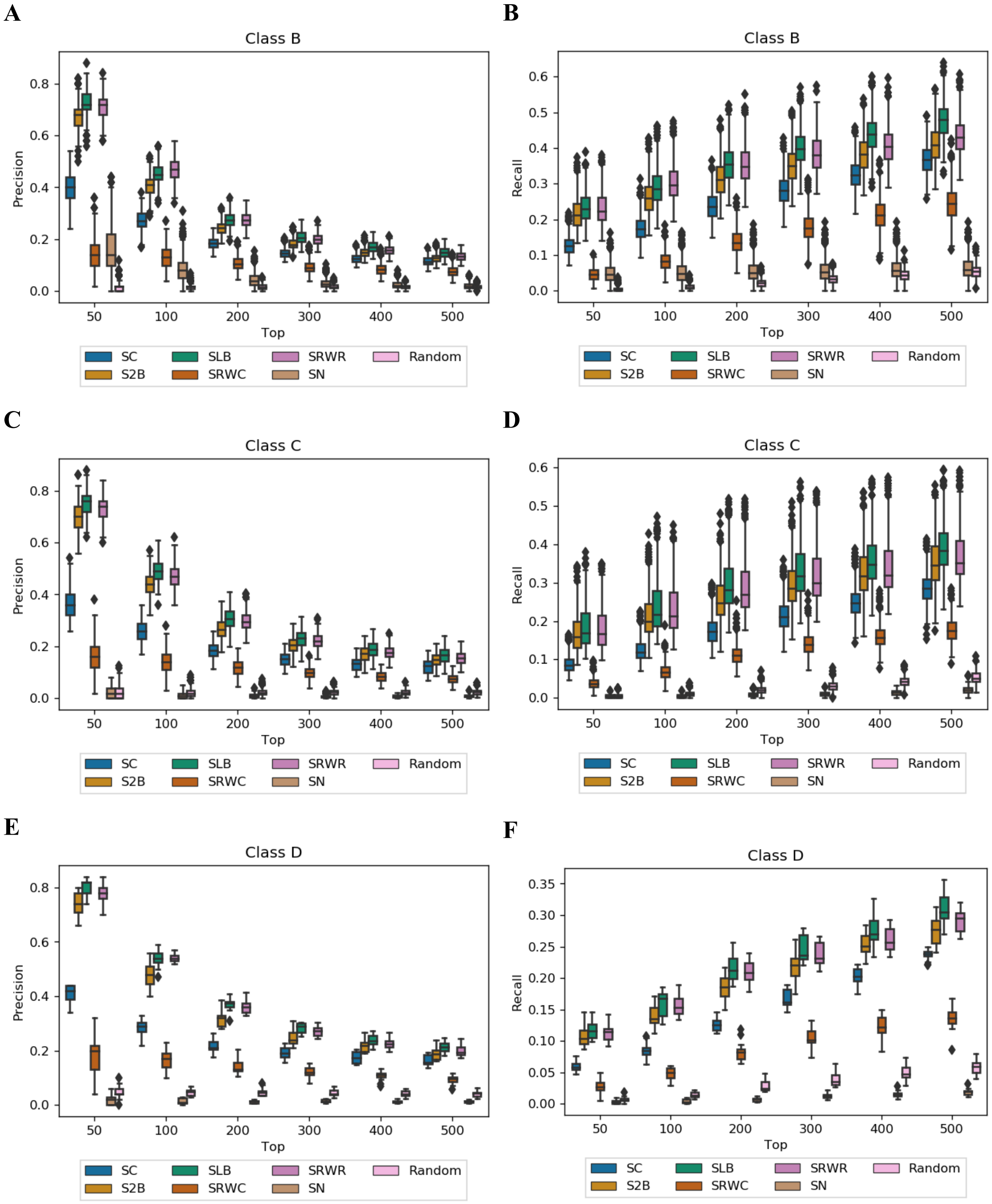

Figure S1: Methods precision and recall distribution per tops on undirected version. (**A**) Class B. (**B**) Class C. (**C**) Class D.

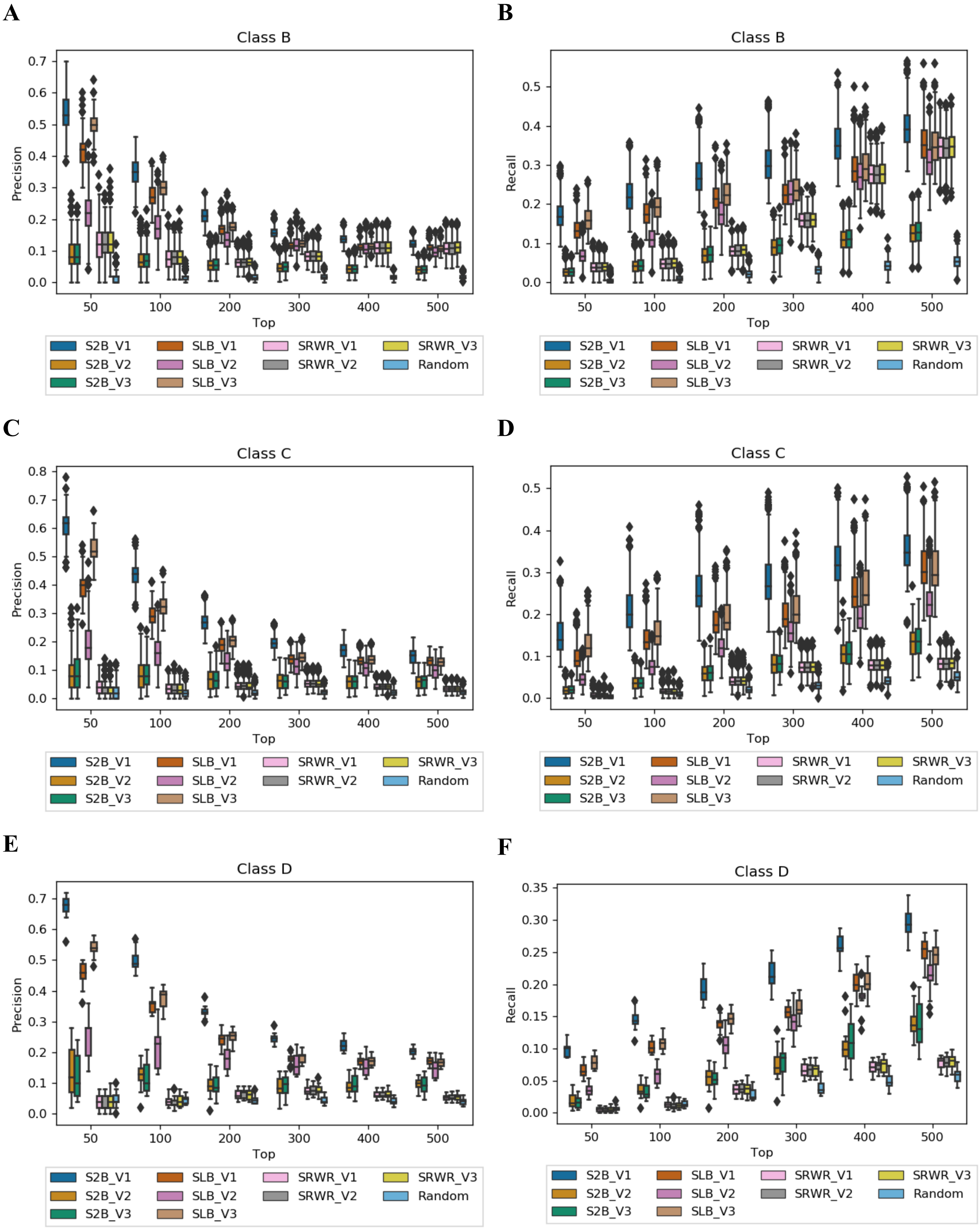

Figure S2: Methods precision and recall distribution per tops on directed version. (**A**) Class B. (**B**) Class C. (**C**) Class D.

Table S1: Wilcoxon test on recall resume between pairs of methods.

		Top 50 Class A			Top 50 Class E		
Method A	Method B	Median A	Median B	p-value	Median A	Median B	p-value
S2B	SLB	0.34	0.38	<0.001	0.08	0.08	<0.001
S2B	SRWR	0.34	0.375	<0.001	0.078	0.08	<0.001
SLB	SRWR	0.37	0.375	<0.001	0.08	0.08	0.031
S2B	S2B_V1	0.34	0.25	<0.001	0.07	0.07	<0.001
SLB	SLB_V3	0.38	0.27	<0.001	0.08	0.06	<0.001
SRWR	SRWR_V1	0.375	0.08	<0.001	0.08	0.008	<0.001

The following scatter plots represent the the average percentage of nodes added with respective precision of each disease pair in the three best methods, S2B (S3), SLB (S4) and SRWR (S5) as described in subsection 4.3.3.

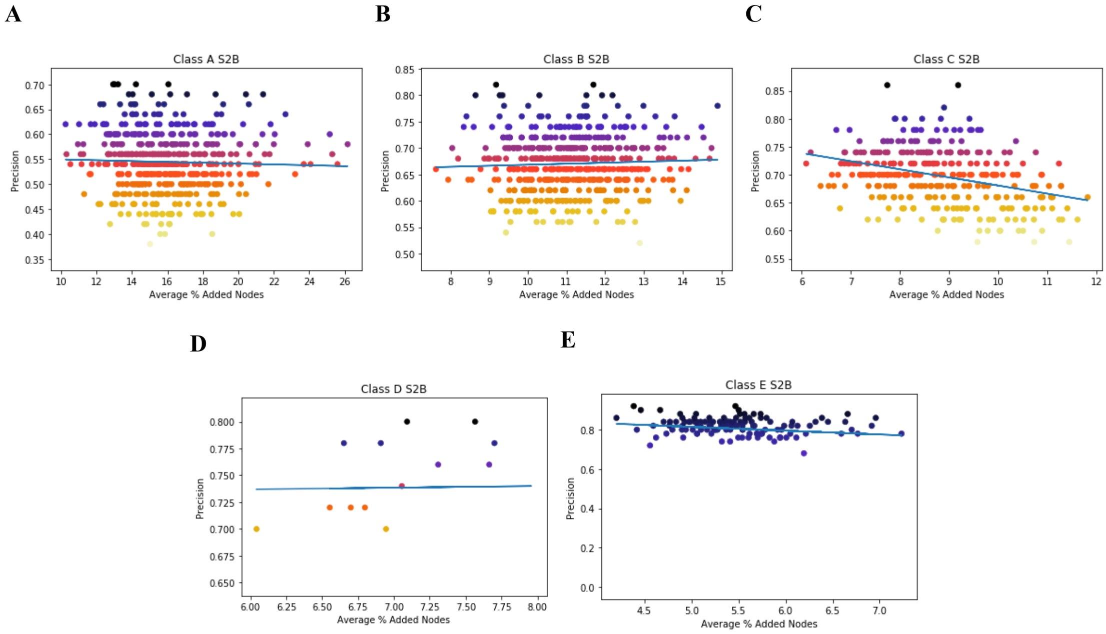

Figure S3: Precision score per percentage of added nodes on all pairs of diseases in S2B top 50 . (**A**) Class A. (**B**) Class B . (**C**) Class C. (**D**) Class D. (**E**) Class E.

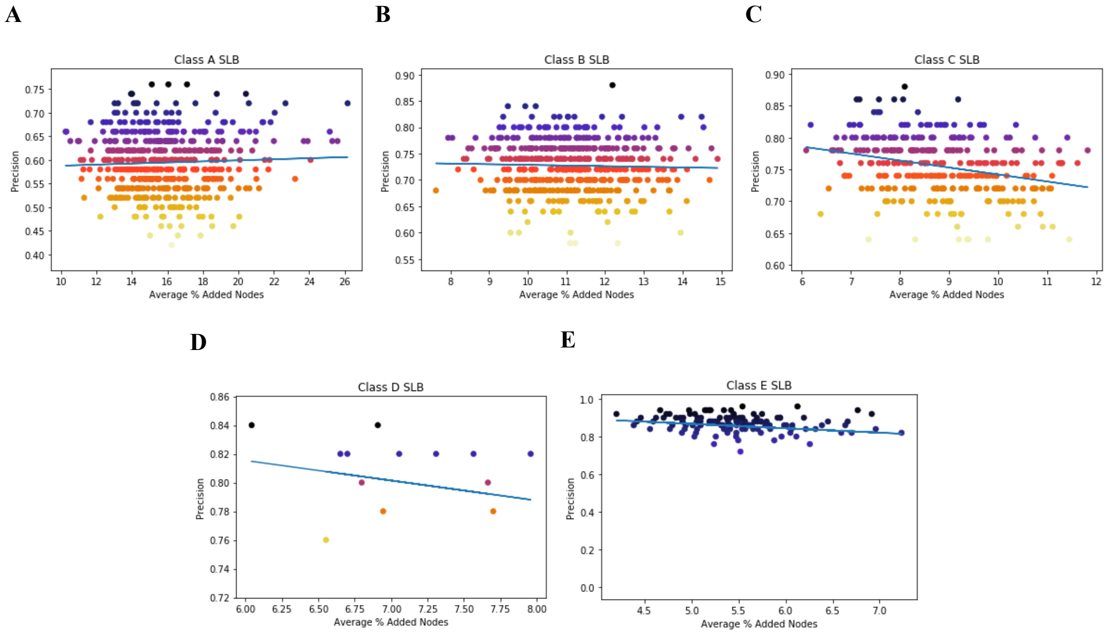

Figure S4: Precision score per percentage of added nodes on all pairs of diseases in SLB top 50. (**A**) Class A. (**B**) Class B. (**C**) Class C. (**D**) Class D. (**E**) Class E.

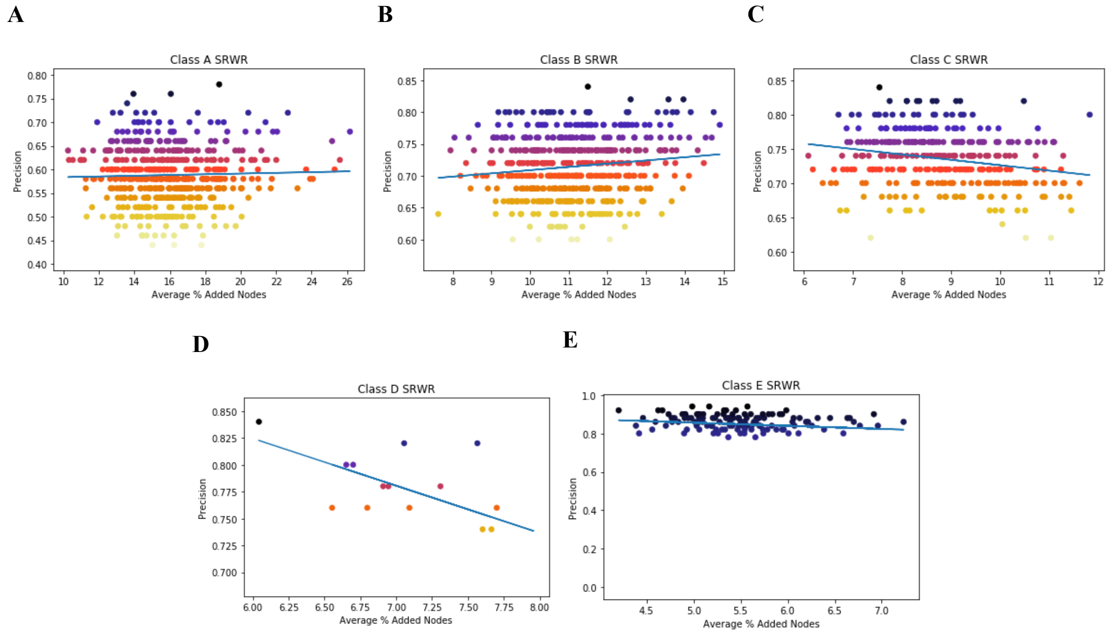

Figure S5: Precision score per percentage of added nodes on all pairs of diseases in SRWR top 50. (**A**) Class A. (**B**) Class B. (**C**) Class C. (**D**) Class D. (**E**) Class E.

This results shown the remaing precision and recall for the remaining classes for APID (S6), OmD (S7) and PCNET (S8) as refered on the results of chapter 5.

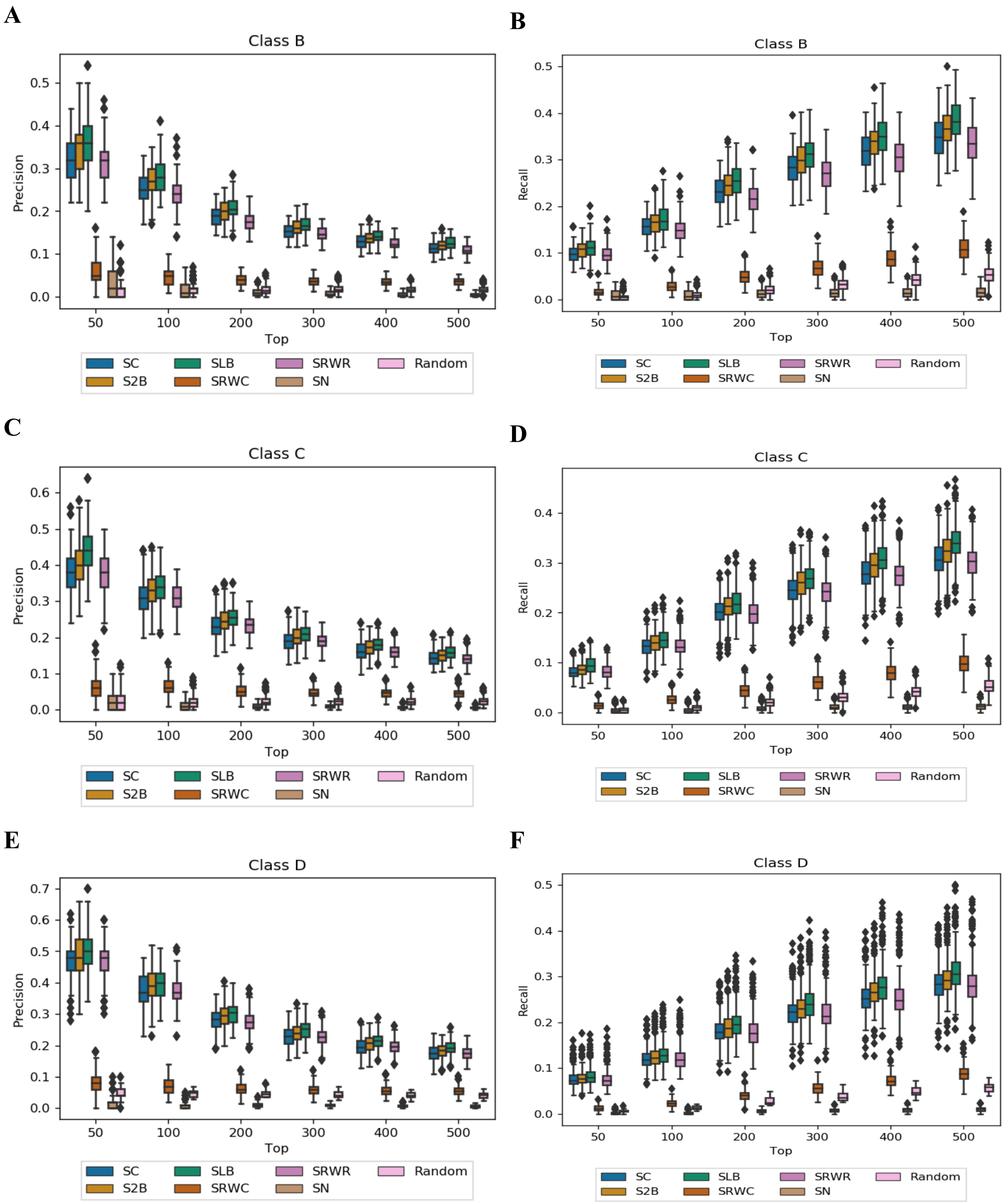

Figure S6: Methods precision and recall distribution per tops on APID. (**A**) Class B. (**B**) Class C. (**C**) Class D.

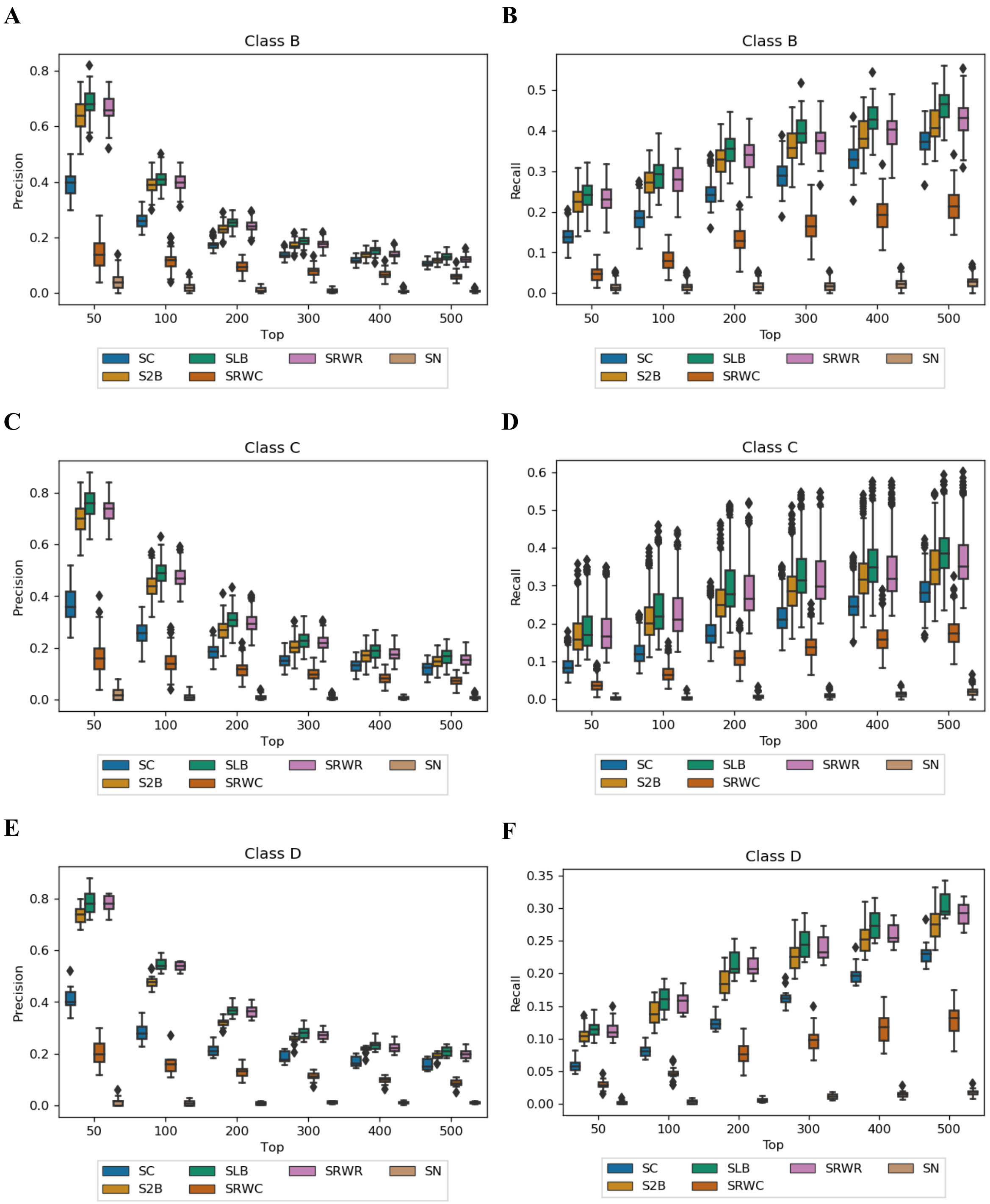

Figure S7: Methods precision and recall distribution per tops on OmD. (**A**) Class B. (**B**) Class C. (**C**) Class D.

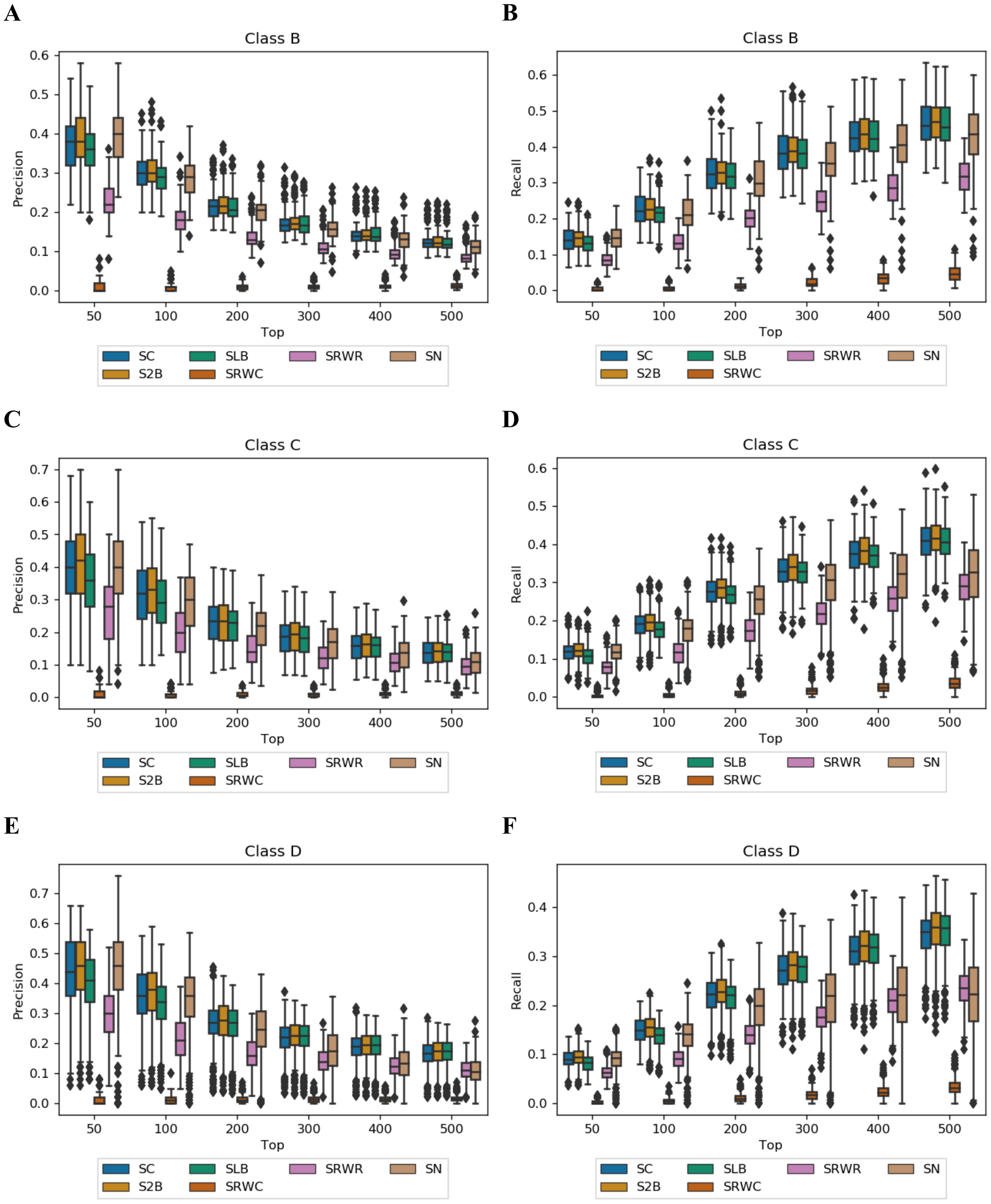

Figure S8: Methods precision and recall distribution per tops on PCNET. (**A**) Class B. (**B**) Class C. (**C**) Class D.