INSTITUTO FEDERAL DO ESPÍRITO SANTO
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

JOÃO CARLOS PANDOLFI SANTANA

ANÁLISE DE ROBUSTEZ DO MÉTODO NERI BASEADO NAS HIPÓTESES DA *NETWORK MEDICINE* PARA PRIORIZAÇÃO GÊNICA

SERRA
2017

JOÃO CARLOS PANDOLFI SANTANA

ANÁLISE DE ROBUSTEZ DO MÉTODO NERI BASEADO NAS HIPÓTESES DA *NETWORK MEDICINE* PARA PRIORIZAÇÃO GÊNICA

Trabalho de Conclusão de Curso apresentado à Coordenadoria do Curso de Bacharelado em Sistemas de Informação do Instituto Federal do Espírito Santo, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Orientador:
Prof. Dr. Sérgio Nery Simões

SERRA

2017

P149e João Carlos Pandolfi Santana
ANÁLISE DE ROBUSTEZ DO MÉTODO NERI BASEADO NAS HIPÓTESES DA *NETWORK MEDICINE* PARA PRIORIZAÇÃO GÊNICA/ João Carlos Pandolfi Santana. – Serra, 2017-
62 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Sérgio Nery Simões
Monografia (Graduação) – Instituto Federal do Espírito Santo ,
Coordenadoria de Informática, Curso Bacharelado em Sistemas de Informação, 2017.

1. 2. 3. I. II. Instituto Federal do Espírito Santo. III. Título.

JOÃO CARLOS PANDOLFI SANTANA

ANÁLISE DE ROBUSTEZ DO MÉTODO NERI BASEADO NAS HIPÓTESES DA *NETWORK MEDICINE* PARA PRIORIZAÇÃO GÊNICA

Trabalho de Conclusão de Curso apresentado à Coordenadoria do Curso de Bacharelado em Sistemas de Informação do Instituto Federal do Espírito Santo, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovado em 13 de Julho de 2017.

COMISSÃO EXAMINADORA

Prof. Dr. Sérgio Nery Simões
Instituto Federal do Espírito Santo
Campus Serra

Prof[a] Dr[a] Helena Paula Brentani
Universidade de São Paulo
Instituto de Psiquiatria

Prof. Dr. David Correa Martins-Jr
Universidade Federal do ABC
Centro de Matemática, Computação e Cognição

Prof. MSc. Thiago Meireles Paixão
Instituto Federal do Espírito Santo
Campus Serra

Declaro, para fins de pesquisa acadêmica, didática e técnico-científica, que este Trabalho de Conclusão de Curso pode ser parcialmente utilizado, desde que se faça referência à fonte e ao autor.

Serra, 29 de Julho de 2017.

João Carlos Pandolfi Santana

Aos meus pais, por acreditarem e me suportarem durante esta caminhada.
A minha companheira de conversas quânticas e de vida Carine e ao meu orientador que perdeu sábados e horas de sono comigo.

Agradecimentos

São tantas coisas a dizer em um pequeno pedaço de papel, portanto de minhas mãos e meus pensamentos neste instante, humildemente agradecerei aqueles que for capaz de fazê-lo.

Meus sinceros agradecimentos aos meus professores durante a faculdade, me mostraram o valor do conhecimento e me instigaram a sempre buscar mais para entregar o melhor de mim.

Muito obrigado a minha tia Ieda, que durante os meus primeiros períodos de faculdade, no qual eu não possuía uma boa condição financeira, me apoiou de todas as formas. Neste mesmo sentimento, um sincero agradecimento aos meus amigos de faculdade, que tornaram os momentos difíceis em "só mais uma".

Não poderia deixar de fora minha companheira de conversas filosóficas, discussões sobre física quântica e a intimidade da matéria, essa pessoa que além de ser minha amiga é minha companheira de vida, Carine Ribeiro.

Meus velhos pais e sua incansável garra, seus conselhos e "puxões de orelha". Minha mãe e seu cappuccino às 2h da manhã enquanto eu estava mais uma vez acordado. Meu pai com o seu "vai dormir para não me dar trabalho", que com certeza era sua maior demonstração de carinho.

Minha segunda família, onde fui adotado e alimentado com churrasco às 23h, conselhos e viagens muito peculiares. Abraços muito fortúitos em horas extremamente necessárias e por me ensinarem outra forma de ver a vida. Sim, é a família da Carine.

Agradecimento imenso ao meu orientador Sérgio Nery, que perdeu algumas boas horas de sono e talvez alguns cabelos comigo durante esse tempo. As coisas que fez e que deixou de fazer por mim e principalmente por acreditar que valeria a pena investir. Você é ninja... opa, ninja não Samurai!

O maior bem do Homem é uma mente inquieta.

Isaac Asimov

Resumo

Um dos grandes problemas enfrentados pelos pesquisadores é o estudo das doenças complexas, pois elas são poligênicas e multifatoriais, fazendo com que diferentes estudos apresentem baixa replicabilidade. Recentemente, avanços significativos tem sido obtidos por métodos que realizam integração de dados entre expressão gênica e dados de rede PPI (*Protein Protein Interaction Network*). Dentre eles destaca-se o método NERI que obteve bons resultados de replicabilidade. Esse método baseia-se nas hipóteses da *Network Medicine* combinadas com métodos de importância relativa em redes complexas. A importância relativa é uma forma de inferir a relevância topológica dos nós da rede baseado em um conjunto de nós conhecidos como sementes. Entretanto, esse método carece de uma análise de robustez, que avalie o quanto seus resultados são dependentes dos genes sementes. Neste trabalho, analisamos a robustez do método NERI com relação aos genes sementes visando avaliar o impacto da remoção progressiva destes. Realizamos experimentos mantendo fixos a rede PPI e os dados de expressão, mas removendo progressivamente parte dos nós sementes do conjunto original (de forma similar as técnicas de avaliação de classificadores *leave-one-out* e validação cruzada), e comparando a interseção entre seus resultados com o resultado original. Variamos o percentual de genes sementes excluídos entre 10 e 40% e, em seguida, comparamos os primeiros elementos das listas resultantes com os primeiros da lista original. Considerando o melhor cenário (remoção de 10% das sementes), as listas resultantes apresentaram em média 90% de interseção com a lista original, e mesmo no pior cenário (remoção de 40% das sementes), a interseção foi de 60% em média. Além disso, observamos também que quanto maior a lista dos primeiros genes comparados, menor é a variância das interseções das listas resultantes com a lista original. Portanto, o método NERI pode ser considerado robusto com relação aos genes sementes, o que indica replicabilidade mesmo em situações onde os genes sementes são variados.

Palavras chaves: Network Medicine; Rede PPI; Importância Relativa; Método NERI; Robustez das Sementes.

Abstract

An importante problem faced by researchers is the study of complex diseases, because they are polygenic and multifactorial, implying small replicability among different studies. Recently, significant advances have been reached by methods that perform data integration between gene expression and PPI (Protein-Protein Interaction Network) data. Among them, we highlight the NERI method which achieved good replicability results. This method is based on Network Medicine hypotheses combined with relative importance analyses in complex networks. Relative importance assesses the topological relevance of network nodes based on a set of important nodes known as seeds. However, until this date no robustness analysis was conducted for the NERI method, in order to evaluate how much its results are dependent on the seed genes. In this work, we analyzed the robustness of the NERI method with regard to the seed genes in order to evaluate the impact of their progressive removal. We performed experiments fixing the PPI network and expression data, but progressively removing parts of the seed nodes from the original set (analogous to the classification assessment techniques, such as cross-validation), and comparing the intersection between their results and the original result. We excluded between 10%–40% of the seed genes and compared the top elements of the resulting lists with the top ones from the original list. Considering the best scenario (removal of 10% of seeds), the resulting lists averaged 90% of intersection with the original list, and even in the worst case scenario (removal of 40% of seeds), the intersection was 60% in average. In addition, we also note that the larger the list of the first genes compared, the smaller is the variance of the intersections of the resulting lists with the original list. Therefore, the NERI method can be considered robust with respect to the exclusion of seed genes, also presenting good replicability in such a case.

Keywords: Network Medicine; PPI Network; Relative Importance; NERI Method; Seed Robustness.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Doenças complexas são poligênicas e multifatoriais, ou seja, além de serem causadas por mutações em mais de um gene, também são influenciadas por fatores ambientais. Alguns exemplos de doenças complexas são: esquizofrenia, transtorno do espectro autista, hipertensão, asma, Diabetes Melitus, doença de Parkinson e esclerose múltipla. Quanto aos fatores genéticos, devido ao fato destas doenças serem poligênicas, as mutações podem levar a uma propagação não natural de informação e sinais, de forma que afete outros genes e/ou mecanismos dependentes dos que sofreram determinada mutação. Uma forma de estudar este tipo de doença, é analisar os transcritos gerados pela transcrição dos genes, de forma a buscar uma relação de co-expressão, tendo como objetivo encontrar genes que influenciam na doença em questão. Para utilizar estes dados, é possível modelar em forma de rede, onde cada nó representa um gene, as arestas representam a co-expressão genica, e ao utilizar a topologia de redes com pesos, o fator de co-expressão torna-se então o peso, determinando assim o grau de relacionamento entre dois nós, com essa abordagem, é possível aplicar conceitos e propriedades de grafos no problema, devido ao fato de ele estar modelado em rede. Outra forma de estudar as doenças poligênicas, é analisar as interações entre proteínas (*PPI – Protein-Protein Interaction*), onde também é aplicada a abordagem de redes para investigação da doença, no qual é chamada de hipótese da *Network Medicine* (BARABASI; GULBAHCE; LOSCALZO, 2011). Este modelo leva em conta o nível de interação entre as proteínas e quais foram os genes responsáveis por gerá-las, podendo assim ter um mapeamento gênico e proteico ao mesmo tempo.

Estas duas abordagens citadas englobam conceitos de redes complexas, onde têm-se a representação de dados e relações entre eles em forma de grafos, sejam eles com pesos ou não (em sua grande maioria são utilizados grafos direcionados e com peso), esta abordagem permite utilizar conceitos fundamentados sobre teoria de grafos e algoritmos consolidados para análise do problema, ganhando-se assim mais ferramentas para tratamento do modelo em questão.

De acordo com os conceitos apresentados, existem diversas abordagens para tratar doenças poligênicas, dentre elas, destaca-se o método NERI (SIMÕES et al., 2015) que apresentou bons resultados de replicabilidade. Este é um método que baseia-se em importância relativa, ou seja, fundamenta-se em nós sementes para o seu funcionamento, onde estes são genes sabidamente reconhecidos como importantes. Em vista desta abordagem, este método carece de uma análise de robustez, o que significa analisar o quão dependente dos nós sementes o método é, de forma a encontrar um coeficiente de confiança, para assim gerar uma segurança na utilização da ferramenta, porém, para encontrar estes coeficientes é necessário observar o comportamento do método quando há retiradas de nós sementes

da rede de entrada.

Ao analisar o código fonte do programa em questão, foi identificada a necessidade de reestruturação do mesmo, de forma que fique mais modularizado, facilitando a manutenção e adição de novas funcionalidades futuras. Como a ferramenta foi desenvolvida para ser utilizada por biólogos, foi identificado também, a necessidade de desenvolver uma interface gráfica para facilitar e disseminar o uso. Atualmente a utilização é feita somente por linha de comando no terminal, o que requer um nível de conhecimento mínimo. A interface gráfica tem como objetivo viabilizar o uso do programa para biólogos não familiarizados em utilizar programas por linhas de comando em terminal, potencializando assim o alcance da ferramenta e incentivando o uso de métodos computacionais por biólogos.

1.1 Objetivos

1.1.1 Objetivo Geral

1. Analisar robustez do método NERI avaliando o impacto da retirada de alguns genes sementes.

1.1.2 Objetivos Específicos

1. Refatorar o código para facilitar a manutenção, utilização e contribuição externa da comunidade de usuários e desenvolvedores.

2. Implementar os algoritmos de validação *Leave-one-out Cross-validation* e *Repeated K-Fold Cross-validation* aplicados ao método NERI.

3. Analisar a robustez do método NERI verificando sua dependência em relação aos nós sementes.

4. Iniciar o projeto da implementação de uma interface gráfica para facilitar o uso da ferramenta, visando torná-la mais intuitiva e amigável aos pesquisadores da área biológica.

1.2 Organização do trabalho

No capítulo 2, apresentamos alguns conceitos relacionados a grafos e como estes podem ser utilizados para representar redes biológicas. Também são apresentados conceitos relativos as hipóteses da *Network Medicine*, da redes PPI, de importância relativa através de sementes, e como tais conceitos são utilizados pelo método de integração de dados NERI. No capítulo 3 apresentamos a metodologia que utilizamos para avaliar a robustez

do método através da remoção de um único e de vários genes sementes. No capítulo 4 apresentamos os resultados das experimentos e comparamos com relação aos primeiros genes priorizados no experimento original. No capítulo 5, fazemos algumas considerações finais em relação à robustez do NERI e sugerimos alguns trabalhos futuros.

2 Referencial Teórico

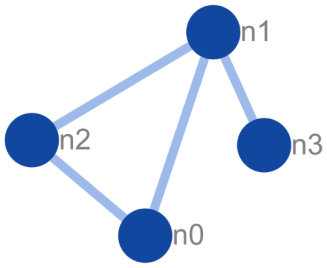
Para melhor compreensão do conteúdo apresentado neste trabalho, este capítulo tem como objetivo explicar os fundamentos conceituais necessários para garantir a boa compreensão e evolução do conteúdo.

2.1 Fundamentos Matemáticos

2.1.1 Grafos

Grafos são formas de estruturação de dados ligados pertencentes ao mesmo conjunto. Um elemento recebe a denominação de nó ou vértice. As relações entre os nós são definidas por *arestas*. Para exemplificação, tome como nós, cidades *n1*,*n2*, *n3* e *n4*, onde as estradas que fazem conexão entre estas cidades representam as arestas 2.1. Com este conceito, pode-se modelar estas ligações em forma de grafo. (STUMPP, 2013)

Figura 2.1 – Representação em grafo



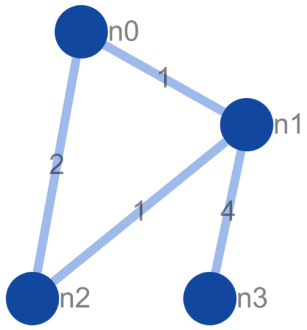
Fonte: Produzido pelos autores.

2.1.2 Grafos com pesos

São grafos que possuem um **grau de importância** (também chamado de peso) em cada **aresta**, este **grau de importância** tem significado apenas em nível de abstração, ou seja, não carrega uma interpretação predefinida. Geralmente, exprime o quão **relacionado** um nó está com outro, onde esta informação é interpretada de acordo com o contexto no qual está inserido.(STUMPP, 2013)

Figura representativa 2.2.

Figura 2.2 – Grafo com peso

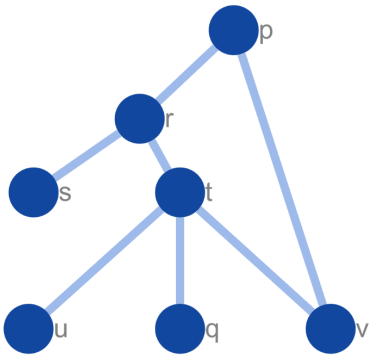


Fonte: Produzido pelos autores.

2.1.3 Passeio

É uma sequência específica de nós ligados, partindo de ***p*** e chegando em ***g*** (PAVLOPOULOS et al., 2011). Onde o **comprimento** do *passeio* é determinado pelo número de *arestas* percorridas. Tomando o *grafo* representado pela Figura 2.3, um dos **passeios** possíveis de p a q é o conjunto A, formado pelos nós visitados $A = \{p,r,t,v,t,q\}$. O **comprimento do passeio** A é 6, como também pode ser definido como o tamanho do *conjunto* de nós visitados *menos 1*.

Figura 2.3 – Grafo



Fonte: Produzido pelos autores.

2.1.4 Caminho e distância

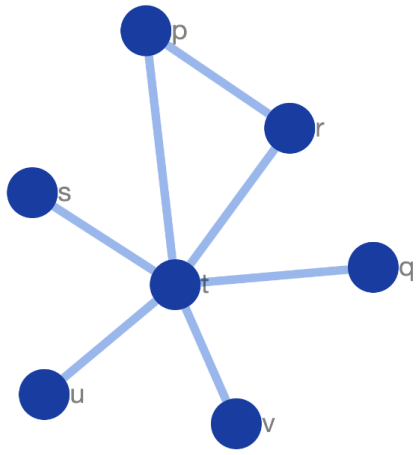
Assim como o *passeio*, o *caminho* é uma sequência específica de nós ligados. Porém este não possui vértices repetidos, ou seja, não passa duas vezes pelo mesmo vértice (PAVLOPOULOS et al., 2011). Se tomarmos como exemplo o grafo da imagem 2.3, um **caminho** de p a q é a sequencia de nós $\{p,r,t,q\}$. A **distância** do *caminho* é definida pela soma dos pesos em suas arestas em *grafos com pesos*. Para os *grafos sem peso*, a **distância** é definida pela quantidade de arestas presentes no *caminho*, implicitamente definindo o

peso de cada *aresta* como *1* e executando a soma das mesmas. No **caminho** *{p,r,t,q}* a **distância** entre *p* e *q* é *4*.

2.1.5 Nó Hub

Hub é um nó que possui muitas arestas, ou seja, um nó que se liga a muitos outros nós (PAVLOPOULOS et al., 2011). Na imagem 2.4 o *hub* é o nó *t*, ou seja, é o nó que possui mais arestas *6* do grafo.

Figura 2.4 – Representação de *Hub*



Fonte: Produzido pelos autores.

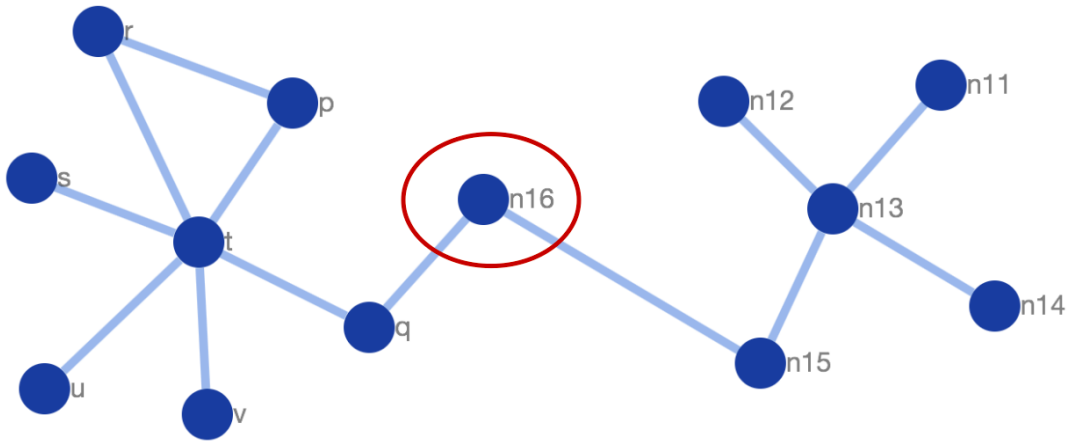
2.1.6 Nó Bridge

Nós *bridge* ou ponte são nós que conectam duas comunidades ou módulos de rede, ou seja, é uma medida que determina o quanto um nó liga dois grandes agrupamentos (HWANG et al., 2006). Na Figura 2.5 o nó ***n16*** é um definido como *bridge* e os nós ***t*** e ***n13*** são *Hubs*, mesmo o nó ***n16*** não estando diretamente ligado nos nós *Hubs*, ele apresenta o comportamento de *Bridge* por estar conectando os dois grandes agrupamentos.

2.1.7 Menor caminho ou caminho mínimo

Quando se trata de grafos o **caminho mínimo** é aquele que possui a menor distância entre dois nós (*p* e *g*) (DIJKSTRA, 1959). No grafo com pesos representado pela Figura 2.2, o caminho mínimo entre os nós *n2* e *n3* é dado pelo conjunto de nós *{n2,n1,n3}*, sendo a distância entre *n2* e *n3* igual a *5*. Um outro caminho válido mas que não é mínimo entre *n2* e *n3* é dado pelo conjunto de nós conectados *{n2,n0,n1,n3}*, no

Figura 2.5 – Representação de um nó *bridge*



Fonte: Produzido pelos autores.

qual a distância entre *n2* e *n3* é igual a *7*. Ambos são caminhos válidos no mesmo grafo, porém o menor caminho possível entre *n2* e *n3* neste grafo é *{n2,n1,n3}*.

2.1.8 Redes complexas

O estudo de redes complexas é uma área relativamente recente e ativa da pesquisa científica e foi inspirada em grande parte pelo estudo empírico de redes do mundo real, como redes de computadores, redes tecnológicas, redes cerebrais e redes sociais. Uma rede complexa pode ser representada por um grafo características topológicas não triviais, ou seja, características que não ocorrem em redes simples tais como reticulados, e nem ocorrem em gráficos aleatórios, mas geralmente ocorrem em modelos de sistemas reais.

O conceito de redes complexas é muito importante para representação de sistemas complexos, sendo que os mesmos podem ser representados em forma de grafos e consequentemente em rede (STROGATZ, 2001). Porém, redes complexas devem apresentar estruturas topológicas não triviais, ou seja, possuírem um conjunto de vértices que sejam interconectados por arestas (BARABÁSI, 2003).

Uma característica importante das redes complexas, são as suas propriedades, como por exemplo, as medidas de centralidade que apontam comportamentos de um determinado nó ou um conjunto deles. Isto faz com que a análise destes dados representem algo direto no problema, visto que a rede é uma abstração do problema real (METZ, 2007). Tais características auxiliam na abstração do problema, onde o pesquisador interpreta cada

comportamento encontrado na rede, uma aplicação no "*mundo real*"estudado. Ao permitir que um problema seja modelado e representado por uma rede complexa, pode-se estudá-la utilizando os ferramentais desenvolvidos para grafos e redes complexas. Isto auxilia a resolução do problema, pois permite ao pesquisador utilizar conceitos já validados.

2.2 Fundamentos biológicos

2.2.1 Co-expressão de transcritos

A correlação de duas variáveis significa o quanto o comportamento de ambas está relacionado, ou seja, se a variação de uma variável A acompanha a variação da variável B. Este fator funciona tanto para variações diretamente proporcionais quanto para variáveis inversamente proporcionais, sendo as diretamente proporcionais quando as duas variáveis possuem variações de mesmo sinal e a inversamente proporcional quando as variações são em relação a sinais opostos. De forma mais objetiva, correlação é uma medida que varia de -1 a 1, onde representando o valor -1 as variáveis analisadas são inversamente proporcionais, ou seja, quando o valor uma variável aumentou as outras diminuíram e vice versa. Quando o fator é 0 as variáveis não tem relação nenhuma, significando que há variação em uma ou mais variáveis, não necessariamente há variação nas outras. E por fim, quando o se apresenta com o valor 1, significa que as variáveis em questão variam juntas, onde quando há aumento do valor de uma ou mais variáveis, todas as outras também apresentaram um aumento em seus valores, o mesmo comportamento se mantém se houver a diminuição do valor em uma ou mais variável, todas as outras irão acompanhar apresentando diminuição nos seus respectivos valores.

O conceito de co-expressão é a correlação de transcritos gênicos, onde determina-se o quanto a variação de um gene está relacionado com outro. Definindo, desta forma valores de co-expressão transcriptômica entre os genes (GAITERI et al., 2014).

2.3 Rede PPI

Redes PPI (Protein-protein Interaction), são redes que representam dados de interação entre diferentes proteínas que indicam como as mesmas ativam ou não processos biológicos dentro das células (PAVLOPOULOS et al., 2011). Apesar das interações entre diferentes proteínas e suas respectivas sequencias gênicas estarem praticamente todas descobertas, ainda não se sabe completamente suas funções moleculares. Através da modelagem das interações em rede, é possível inferir funções proteicas com a interação entre outras biomoléculas.

Assim como as interações entre proteínas podem ser mapeadas em rede, surge um novo conceito denominada *Network Medicine* por (BARABASI; GULBAHCE; LOSCALZO,

2011), onde o autor aborda doenças humanas mapeando-as em redes complexas. Desta forma, pode-se estudar não só os fatores de complexidade molecular relacionados a uma determinada doença, mas também é possível analisar os fenótipos relacionados, levando à possibilidade de encontrar uma via biológica relacionada a doença.

2.4 Redes Biológicas

2.4.1 Relação de menor caminho

A relação de menor caminho em uma rede biológica determina o quão próximo dois elementos estão do outro, sendo assim uma possível via biológica de interferência direta de um elemento a outro. Como por exemplo na rede gênica, o caminho entre os genes pode ser entendido como uma via biológica de ativação, apresentando um ponto para estudo e validação por um pesquisador.

2.4.2 Redes de Co-expressão

As redes podem ser representadas como grafos sem peso e grafos com peso. No caso das redes que utilizam dados de co-expressão, os valores encontrados são utilizados como pesos entre os genes em representados na rede. Isto se aplica pelo fato do peso entre dois nós representar o quão relacionados eles estão. Neste mesmo contexto, os dados de co-expressão determinam o quanto a transcrição de um gene está relacionada coma transcrição de outro gene. Por este motivo, mapeia-se os valores de co-expressão entre os genes para os pesos relativos na rede, determinando o grau de proximidade entre eles com valores reais obtidos por análise (GAITERI et al., 2014).

2.4.3 Importância relativa

Em grafos grandes e complexos, as relações entre os nós e suas conexões geralmente exprimem um significado. O estudo destas relações é importante para análise dos dados que o grafo representa, assim sendo, surge a necessidade de identificar quais são os nós mais importantes da rede em relação aos nós sementes. A necessidade de determinar esta importância é denominada *Importância Relativa*.

Este é um tema que gerou publicações importantes no estudo de redes complexas. Um destes estudos é do (WHITE; SMYTH, 2003), que propuseram diferentes maneiras de eleger os nós mais importantes de uma rede. Assim como a pesquisa no qual este trabalho propõe a validação o (SIMÕES et al., 2015), utiliza destes conceitos para eleger os genes mais importantes relacionados a doença em estudo.

Importância relativa é essencial na abordagem de redes complexas, devido ao fato de elencar os nós mais importantes baseados em um conhecimento prévio. Este conhecimento

prévio é modelado em nós sementes, ou seja, alguns nós que previamente são conhecidos por sua importância, são utilizados como ponto de partida e elementos chaves para a busca de outros nós importantes na rede. Este estudo baseia-se na importância relativa individual, ou seja, a importância de um nó em relação ao conjunto.

Outro modelo de importância relativa é a definição dos caminhos mais importantes na rede, ou seja, o melhor caminho entre em dado nó p ao nó q. Este conceito também é aplicado pelo *Método NERI*, onde os caminhos são identificados como vias biológicas da doença em estudo. Este conceito se popularizou com o (BARABASI; GULBAHCE; LOSCALZO, 2011) denominado *Network Medicine*, abrindo caminho para diversas pesquisas na área e criando um conceito novo de estudo para doenças complexas e multifatoriais.

Um ponto que deve ser ressaltado é no *Método NERI*, onde o autor aplica dois escores para priorização gênica. Estes escores são utilizados como fator de ranqueamento dos genes mais influentes para uma determinada doença, são eles Δ' e X. O score de ranqueamento baseado em importância relativa Δ', que prioriza a maior alteração entre as pontuações em cada condição, e o score X que privilegia a maior soma das pontuações definidas. Estes scores são responsáveis por selecionar os genes mais relacionados a doença estudada.

2.5 Análise de robustez

2.5.1 Método K-Fold Cross-Validation

O método *K-Fold Cross-Validation*, também chamado de estimativa de rotação, é uma técnica desenvolvida para avaliar a capacidade de generalização de um determinado modelo em relação a um conjunto de dados. Este modelo analisa os resultados estatísticos de um agrupamento de dados definido, onde tem sido amplamente empregado em problemas no qual o objetivo da modelagem é predição de dados, isto se dá por seu conceito principal consistir no particionamento dos dados de entrada em subconjuntos mutualmente exclusivos, onde uma parte destes serão revezados na alimentação do modelo a ser validado (grupo de treinamento), e a outra parte utilizados na validação.

Esta separação é feita de forma que o um conjunto seja divido K subconjuntos de tamanhos iguais ou quase iguais. Após a criação dos K conjuntos, o modelo em questão, é treinado com K - *1* conjuntos. De forma que o conjunto que sobrou seja utilizado como teste dos resultados gerados pela etapa de treinamento. Comumente, os dados presentes nos conjuntos são estratificados para que cada subconjunto represente da melhor forma possível o conjunto total (MUDRY; TJELLSTRÖM, 2011).

2.5.2 Método Leave-one-out Cross-Validation

O método *Leave-one-out Cross-Validation* é um caso especial do *K-Fold Cross-Validation*, onde a quantidade de subconjuntos gerados é do tamanho do conjunto total de dados. O conjunto de dados original é separado de forma que cada subconjunto esteja faltando um elemento, ou seja, dado um conjunto P de tamanho K, devem ser gerados K subconjuntos de tamanho K - *1*, onde todos os subconjuntos devem ser diferentes.

Após a separação dos subconjuntos, uma única execução do modelo é feita, assim pode-se analisar resultado daquele elemento faltante. Este modelo geralmente é aplicado em situações onde o volume de dados não é muito significante, uma área que utiliza bastante este método é a *Bioinformática*, onde poucos dados da amostra estão presentes (MUDRY; TJELLSTRÖM, 2011).

2.5.3 Método Repeated K-Fold Cross-Validation

É uma variação do modelo *K-Fold Cross-Validation*, onde o objetivo é executar várias vezes os subconjuntos gerados na etapa de separação de dados visando um resultado mais confiável (MUDRY; TJELLSTRÖM, 2011). Assim sendo, surgem variações deste conceito, neste trabalho, mais especificamente, no Capítulo 3 apresentamos uma variação baseada neste método para análise da robustez do *Método NERI*.

2.6 Método NERI

Para entender doenças complexas, é necessário encontrar os genes que se relacionam com a mesma. Com a evolução em larga escala das tecnologias de sequenciamento do genoma e das medições de transcritos, assim como o conhecimento da interação presente entre proteína-proteína (PPI – *Protein Protein Interaction*), a pesquisa sobre doenças complexas vêm se tornando cada vez mais comum. Ao basear-se no paradigma do *Network Medicine*, as redes de interação proteína-proteína têm sido utilizadas para enfatizar os genes relacionados à doenças complexas levando em conta fatores topológicos.

O método NERI (SIMÕES et al., 2015) procurou resolver o problema da replicabilidade através da integração de dados biológicos e as hipóteses da *Network Medicine*. Neste projeto, analisamos a robustez deste método com relação aos genes sementes. Porém este método é afetado diretamente pela literatura disponível, onde proteínas mais estudadas tendem a ter mais conexões na rede, fazendo com que diminua a qualidade dos resultados. Sendo assim, métodos que utilizam somente redes PPI não fornecem dados dinâmicos e específicos, dado que a topologia da rede não é exclusiva para uma única doença. No trabalho em questão, foi desenvolvido um método que prioriza genes e vias biológicas relacionados a uma dada doença complexa, através da abordagem de não somente redes

PPI mas também transcrissômica e genômica, sendo os dados integrados em uma única rede. Após a integração e construção da rede, aplicou-se o conceito da *Network Medicine*, encontrando caminhos mínimos que possuam maior co-expressão entre seus genes. Com este modelo foi desenvolvido dois escores de ranqueamento, onde um prioriza genes com maior alteração entre suas pontuações em cada condição, e o outro privilegia os genes com a maior soma destas pontuações. Desta forma a aplicação do método em a três estudos envolvendo de expressão da doença esquizofrenia, recuperou com sucesso genes diferencialmente co-expressos em duas condições diferentes, e juntamente evitou os erros de literatura presentes na rede PPI. Em paralelo, melhorou substancialmente a replicação de resultados pelo método aplicado aos três estudos, onde por métodos convencionais, não atingiam uma replicabilidade satisfatória.

3 Metodologia

Para análise de robustez do Método NERI foram utilizados conceitos de validação baseados na alteração dos genes sementes. Para que seja feita a análise foram utilizados os resultados de priorização gênica gerada pela ferramenta. De forma a ser possível verificar e mapear os impactos causados devido a remoção ou falta de genes sementes no experimento.

3.1 Materiais

Este trabalho utilizou a base de dados ***KATO*** que consiste em expressões gênicas de pessoas portadoras da doença *Esquizofrenia*. Estes dados podem ser encontrados no *Stanley Neuropathology Consortium Integrative Database (SNCID)*[1]. Esta base de dados em específico, foi selecionada devido ao fato de ter sido utilizada na tese de doutorado **??** no qual este trabalho se baseia e se referencia. Assim sendo, os dados obtidos podem ser comparados com os encontrados e apresentados pelo autor, evitando o enviesamento do resultado por diferença de experimentação.

O Método NERI também recebe dados de rede de integração proteína proteína *(Protein Protein Interaction – PPI)*, onde os mesmos também foram mantidos os originais utilizados pelo autor, sendo formada pelas pelas bases de dados ***HPRD***[2] (*Human Protein Reference Database*), ***MINT***[3] (*Molecular INTeraction database*) e ***IntAct***[4] (*IntAct molecular interaction database*). Como entrada do sistema, também são definidos os ***Genes Sementes***. Estes no qual, são os genes onde há certeza da sua relação com a doença analisada em questão, a base de dados original pode ser encontrada no apêndice A.1.

3.2 Estratégia utilizada para análise de robustez

Neste trabalho, foi escolhida a remoção de alguns genes sementes como parâmetro de análise de robustez do Método NERI. Onde o objetivo é identificar o impacto gerado na rede de integração gênica e no resultado final (lista de priorização gênica). Desta forma, podendo calcular a dependência e sensibilidade do método em relação a qualidade e quantidade de genes sementes. Os métodos escolhidos para análise de robustez baseiam-se nos métodos de avaliação de classificadores *Leave-one-out Cross-Validation* e *Repeated K-Fold Cross-Validation*.

[1] SNCID: (http://sncid.stanleyresearch.org)
[2] HPRD: (http://www.hprd.org/)
[3] MINT: (http://mint.bio.uniroma2.it/)
[4] IntAct: (http://www.ebi.ac.uk/intact/)

3.2.1 Remoção de um único gene semente

Foi utilizado um modelo baseado no método validação *Leave-one-out Cross-Validation*, onde consiste na remoção de um gene semente por vez. Desta forma, formando um ***conjunto diferente*** para cada ***gene semente*** removido, garantindo que sempre haja a mesma quantidade de genes em cada conjunto e que todos ***genes sementes*** tenham ficado de fora pelo menos uma vez. Assim sendo, o número final de conjuntos formados será a quantidade total de genes sementes.

Com este método, é possível descobrir se a falta de um único gene semente é responsável por alterar significantemente a priorização gênica gerada pelo Método NERI. Desta forma, podendo analisar se o método em questão é sensível a retiradas de ***genes sementes***. O método de ***remoção de um único gene*** também permite a análise de importância relativa dos ***genes sementes***, onde aquele que causar maior impacto no resultado final indica uma importância relativa maior em relação aos outros.

Também é possível observar se existe relação entre as medidas de centralidade do gene representado na rede (neste trabalho, utilizaremos somente o grau do nó), com o impacto causado. Este é um ponto importante a ser analisado, pelo fato de poder estimar a importância relativa de um gene semente, observando o seu grau na rede de integração. Esta é uma medida que poderá impactar no resultado da priorização gênica resultante do Método NERI, por isso a importância da análise.

3.2.2 Remoção de vários genes sementes

Foi utilizado um modelo baseado no *Repeated K-Fold Cross-Validation*, que consiste na remoção de mais de um gene semente por experimento. Este método foi adotado pela sua característica principal, organização de conjuntos de genes sementes com tamanhos variados. Com esta característica chave, buscou-se estudar o impacto dos resultados de priorização do método quando há a remoção de vários genes sementes do conjunto original.

Com isso, é possível a observação dos conjuntos no qual foram removidos genes que causaram alto impacto em sua remoção na etapa anterior. Estes conjuntos informarão se o impacto na remoção de um único gene é acumulativo, ou seja, se houver a remoção de mais de um gene semente com alto impacto em sua remoção, se o resultado do experimento em questão será proporcional ao estudo da análise anterior.

3.3 Definição dos genes sementes utilizados

Em seu trabalho, os autores (SIMÕES et al., 2015) utilizaram um conjunto de 38 genes sementes obtidos de um estudo de associação de esquizofrenia, apresentados na Tabela A.1 presente no Apêndice A. No entanto, como o Método NERI utiliza a rede

Tabela 3.1 – Tabela representativa. Experimentos resultantes do método de remoção apenas um gene

Conjunto	***Elementos***
S_1	2,3,4,5
S_2	1,3,4,5
S_3	1,2,4,5
S_4	1,2,3,4

Subconjuntos gerados pelo método de remoção de apenas um gene no conjunto de genes sementes ***{1,2,3,4,5}***, sendo S_i o i-ésimo subconjunto gerado.
Fonte: Tabela gerada pelo autor.

PPI, durante a integração dos genes sementes, somente ***30*** genes foram integrados a rede. Estes nos quais foram utilizados neste trabalho, em vista que os genes sementes que não integraram não afetam o resultado final, ficando de fora dos conjuntos de genes sementes gerados para validação.

3.3.1 Aplicação do método similar ao *Leave-one-out Cross-Validation*

Para a validação utilizando o Método similar ao *Leave-one-out Cross-Validation*, em que neste trabalho chamaremos de ***Método de remoção de apenas um gene***, a preparação do experimento consistiu em gerar conjuntos de genes sementes para o Método NERI de forma que cada conjunto tenha um gene a menos em relação ao experimento original. Foi criado para cada gene semente, um conjunto sem o mesmo, onde exista um conjunto para cada gene removido, ou seja, em uma amostra de ***30*** genes sementes, temos ***30*** experimentos possíveis.

A aplicação direta no sistema consiste em cada execução independente do programa. Onde cada experimento seja um conjunto formado da remoção de um gene semente da amostra original. Deve existir um experimento para cada gene semente removido, garantindo que cada gene semente tenha ficado de fora uma vez em relação a todos os experimentos.

Para ilustração do experimento, segue o exemplo: Suponha que exista um conjunto de {***5*** genes sementes {***1,2,3,4,5***}, suponha que deseja-se gerar todos os possíveis conjuntos de genes sementes utilizando o método de ***remoção de apenas um gene***. A Tabela 3.1 representa os conjuntos resultantes, onde S_i é o i-ésimo subconjunto gerado.

3.3.2 Aplicação do método similar ao *Repeated K-Fold Cross-Validation*

Para a validação utilizando o Método similar ao *Repeated K-Fold Cross-Validation*, em que neste trabalho chamaremos de ***Método de remoção de mais de um gene***, inicialmente, foi definido o tamanho dos grupos de genes sementes de entrada a serem removidos para cada bateria de execuções. Levando em consideração a quantidade de

Tabela 3.2 – Tabela representativa. Porcentagem de genes sementes removidos em relação aos ***30*** genes sementes originais e suas respectivas porcentagens

Quantidade de genes removidos	***Porcentagem de remoção***	**Quantidade de Experimentos**
3	10%	50
6	20%	50
9	30%	50
12	40%	50

Fonte: Tabela gerada pelo autor.

genes de entrada total do experimento, ***30*** após a integração com a rede *PPI*, definiu-se que as remoções seriam feitas em relação a porcentagem da amostra original, sendo as porcentagens definidas ***10%*** (3 genes), ***20%*** (6 genes),***30%*** (9 genes) e ***40%*** (12 genes), conforme pode ser observado na tabela 3.2. Também representada na tabela, determinou-se a quantidade de agrupamentos de dados de entrada para cada porcentagem de remoção.

Para uma maior diversidade dos resultados, o método foi implementado de forma que cada grupo fosse diferente do outro. Isso visa garantir que o conjunto de genes removidos não se repita dentro de cada etapa, para evitar executar o mesmo experimento mais de uma vez. Assim, inicialmente são geradas todas as ***50*** combinações dos genes a serem removidos e em seguida são realizados os ***50*** experimentos. O script que gera todas as combinações de N genes em M experimentos foi desenvolvido em ***Python 3.x*** e pode ser encontrado no Anexo B.

A Figura 3.1 ilustra o o processo de criação dos experimentos. Para uma melhor compreensão, segue o exemplo abaixo. Suponha que exista um conjunto de com 10 genes sementes: {***0,1,2,3,4,5,6,7,8,9***}, e suponha que deseja-se gerar 5 conjuntos de genes sementes com um percentual de remoção em ***20%***. Assim, dado ***20%*** de ***10*** elementos, temos ***fator de remoção = 2*** elementos. Seja S_i o i-ésimo subconjunto gerado cujo o respectivo conjunto de genes removidos é R_i Portanto, um exemplo dos 5 subconjuntos está abaixo:

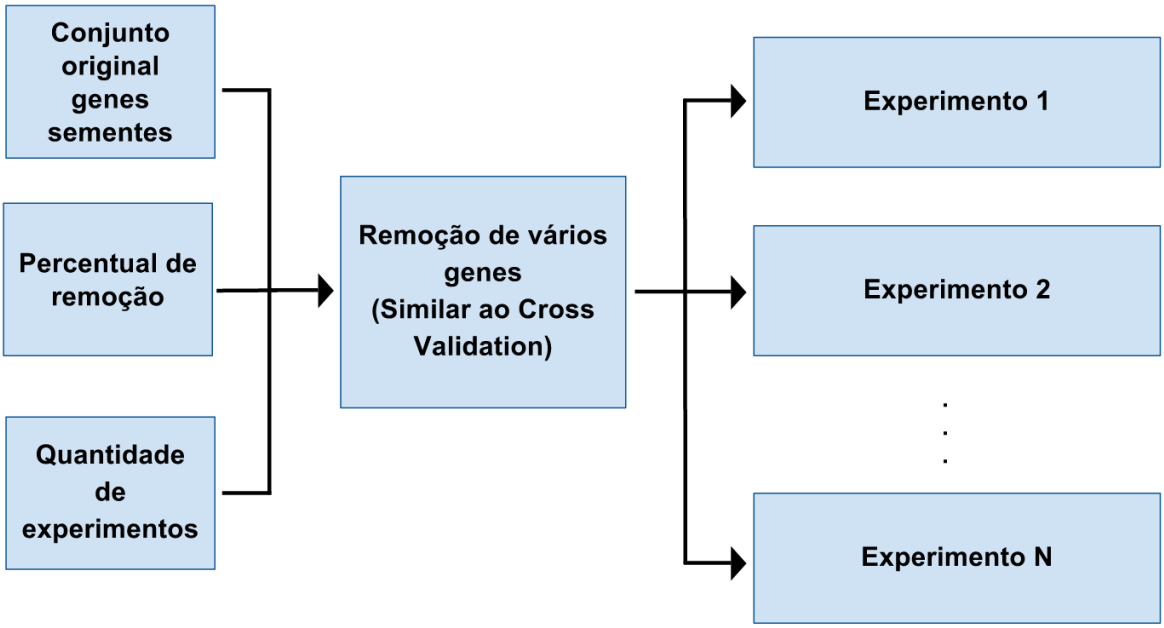
```
1    S1 = {0,1,2,3,4,5,6,7}
2    S2 = {0,1,2,3,4,5,6,8}
3    S3 = {0,1,2,3,4,5,6,9}
4    S4 = {1,2,3,4,5,6,7,8}
5    S5 = {2,3,4,5,6,7,8,9}
```

Sendo R_i os conjuntos dos elementos removidos:

```
1    R1 = {8,9}
2    R2 = {7,9}
```

```
3       R3 = {7,8}
4       R4 = {0,9}
5       R5 = {0,1}
```

Figura 3.1 – Fluxograma de funcionamento da remoção de vários genes sementes.



Fonte: Produzido pelos autores.

Após os agrupamentos de dados de entrada serem preparados, o programa principal é executado individualmente para cada agrupamento. Totalizam-se 200 execuções individuais do programa que implementa o Método NERI.

3.4 Execução dos experimentos

Nesta etapa, os experimentos encontram-se preparados para execução direta no programa principal que implementa o Método NERI. Primeiramente, realizamos experimentos com a remoção de um único gene, para cada um dos ***30*** genes, o que totalizou ***30*** experimentos. Em seguida, realizamos ***50*** experimentos para cada percentual de remoção (***10%***, ***20%***, ***30%***, ***40%***), totalizando ***200*** experimentos nesta etapa.

Devido ao fato de o programa realizar cálculos demorados, houve a necessidade de automatização do processo de execução, onde foi desenvolvido um *Script* em *Shell* (linha de comando Linux), para efetuação do trabalho, podendo ser encontrado no apêndice B. Além disso, houve uma padronização nos diretórios utilizados pelo método NERI, a fim de identificar os experimentos realizados, conforme pode ser observado na Figura C.1 presente no Apêndice C.

Juntamente com as alterações estruturais, também foi desenvolvida uma ***interface gráfica*** e uma interface em ***linha de comando*** (*CLI*), para facilitar a utilização por pessoas que não possuem familiaridade com este modelo de execução de programas. Essa interface está em fase de finalização, faltando a documentação e alguns ajustes, e assim que estiver finalizada será disponibilizada livremente na web.

3.5 Validação

Para garantir que não houvesse erro nos resultados devido a algum erro proveniente a configuração do ambiente de testes ou em relação as bases de dados utilizadas, foi executado o experimento original ***KATO***. O resultado apresentado foi praticamente o mesmo, apresentando pequenos erros nas casas decimais 10^{-20} provenientes por algum erro de arredondamento causado pela arquitetura do computador em questão, também pode ser levado em consideração atualizações das bibliotecas utilizadas pelo ***programa NERI***. Estes erros foram muito baixos e não influenciaram no resultado final gerado pelo programa.

3.6 Metologia de análise dos resultados

Após as execuções dos experimentos definidos, tivemos que definir as métricas para análise dos resultados. O Método NERI realiza uma priorização gênica, produzindo como saída dois escores de ranqueamento (Δ' e X) para cada gene. Assim, são produzidas duas listas (uma lista para o escore X e outra para o escore Δ'), cada uma contendo os genes que participaram dos caminhos mínimos selecionados.

Assim, após a remoção dos genes sementes, realizamos uma comparação de ambas as listas resultantes com a lista original. Essa comparação foi realizada tomando-se os primeiros genes N genes obtidos na lista do experimento e comparando-os com a respectiva lista original. O valor N dos primeiros genes da lista foi variado em {***10***, ***20***, ***50***, ***100***}, e para cada um foi feita a comparação da interseção.

Para comparar duas listas ordenadas, geralmente utiliza-se a ***Correlação de Spearman***. Em nossos testes, inicialmente utilizamos essa correlação para compararmos as listas geradas pelos ***experimentos deste trabalho*** com a lista gerada pelo ***experimento original***. Observamos que, apesar de haver uma interseção razoável dos primeiros genes encontrados em ambas as listas, a comparação utilizando a ***Correlação de Spearman*** apresentou valores próximos de zero, visto que esta medida penaliza listas que não estiverem na mesma ordem. Assim, a ***Correlação de Spearman*** não mostrou-se um bom fator de comparação, pelo fato de a ordem dos genes em um grupo não ser a preocupação principal, mas sim se os genes foram escolhidos para estarem naquele

agrupamento.

Muitas vez um biólogo pode estar mais preocupado com a ***replicabilidade*** de um experimento do que com a ordem dos genes priorizados. Isto é, se os genes recuperados em um método são, em sua maioria, os mesmos recuperados em outro método. Desta forma, algumas vezes pode ser mais importante comparar a interseção das listas do que a ordem das mesmas. Assim, para a análise dos resultados, comparamos as interseções dos $\boldsymbol{N}$ primeiros genes de cada lista, variando $\boldsymbol{N}$ em ***{10,20,50,100,200}***.

Foi utilizada a ***comparação por interseção***, onde consiste em verificar a porcentagem de elementos presentes nas duas listas, de forma que o resultado indique o quão parecido os conjuntos são em relação a presença dos elementos iguais. Sendo definido matematicamente por: $I = \frac{A \cup B}{|A|}$

Este modelo de comparação beneficia as listas que possuírem mais elementos iguais, porém não penaliza as listas seccionadas que tiveram a ordem relativa dos elementos alterada. Este é um fator importante, pois o objetivo da análise é verificar a ***replicabilidade*** do experimento em condições diferenciadas, logo ligeiras variações de posicionamento dos elementos não são problemáticas, desde que os elementos das listas comparadas sejam iguais ou o possuírem uma grande interseçã com o resultado otiginal.

4 Resultados dos experimentos e discussão

Neste capítulo, avaliamos os impactos resultantes da remoção dos genes sementes nas listas de priorização resultante em comparação com a lista original. Ou seja, o quanto a lista de genes resultantes foram recuperados nos experimentos com remoção das sementes em relação a lista resultante original.

Conforme mencionado no capítulo anterior, realizamos experimentos mantendo fixos a rede PPI e os dados de expressão, mas removendo progressivamente parte dos nós sementes do conjunto original (de forma similar as técnicas de avaliação de classificadores *leave-one-out* e validação cruzada), e comparando a interseção entre seus resultados com o resultado original. Após isso, variamos o percentual de genes sementes excluídos entre 10% e 40% e, em seguida, comparamos os primeiros elementos das listas resultantes com os primeiros da lista original.

4.1 Medidas de centralidade dos genes sementes

Primeiramente, uma importante questão a ser verificada é se o impacto causado nos resultados devido à remoção de alguns genes sementes está correlacionado com alguma medida de centralidades dos respectivos genes na rede PPI. Em outras palavras, tais medidas podem informar se o impacto da remoção dos genes sementes deve-se predominantemente a fatores topológicos. A Tabela 4.1 apresenta as medidas de centralidade que os genes sementes utilizados possuem na rede PPI. Os genes estão apresentados ordenados pelo grau decrescente. Observamos que os três primeiros genes: *TP53* (333), *AKT1* (138) e *DISC1* (91), possuem grau destacadamente maior que os demais, que possuem grau abaixo de 50.

Tabela 4.1 – Medidas de centralidade dos genes sementes utilizados no experimento

GENE	**Degree**	**Betweenness**	**Closeness**	**Clustering**	**Brokering**	**Bridgeness**
TP53	333	1017443.366745	0.401408	0.027968	0.034839	135.635614
AKT1	138	260150.217669	0.374562	0.044219	0.014196	206.018081
DISC1	91	131642.895006	0.333142	0.016606	0.009632	130.539882
FEZ1	43	39961.948306	0.315902	0.024363	0.004515	196.253855
ERBB4	40	21172.397911	0.325737	0.144872	0.003682	188.374839
GRIN2B	33	23614.446663	0.317337	0.090909	0.003229	362.909288
APOE	29	19099.496744	0.318884	0.088670	0.002845	346.398103
HP	21	22059.197543	0.324588	0.047619	0.002153	519.171834
DRD2	17	15283.833355	0.301050	0.014706	0.001803	580.045729
HTR2A	16	5847.333317	0.292464	0.008333	0.001708	212.879317
IL1B	10	8530.925938	0.304214	0.066667	0.001005	1381.703074
RGS4	10	2039.400569	0.292253	0.066667	0.001005	463.743523
GAD1	10	7993.469172	0.317012	0.222222	0.000837	888.443883
DRD1	8	9964.681533	0.282170	0.071429	0.000800	827.421927
PPP3CC	8	10512.488952	0.289196	0.035714	0.000830	985.201738
NRG1	7	349.958475	0.272072	0.238095	0.000574	126.388843
COMT	5	3676.951004	0.273329	0.000000	0.000538	1028.239964
SLC6A4	5	565.928439	0.283911	0.000000	0.000538	197.019715
DRD4	5	2814.994398	0.298228	0.000000	0.000538	1209.842385
PLXNA2	4	9316.180826	0.264610	0.000000	0.000431	1746.023442
TPH1	4	574.162877	0.312481	0.333333	0.000287	7943.713227
RELN	4	43.733810	0.240987	0.333333	0.000287	13.904311
GRM3	4	248.812705	0.284468	0.000000	0.000431	597.497122
GABRB2	4	555.585080	0.261145	0.000000	0.000431	288.445351
DAO	3	768.669783	0.283496	0.000000	0.000323	676.604177
OPCML	1	0.000000	0.253044	0.000000	0.000108	0.000000
ZNF804A	1	0.000000	0.258773	0.000000	0.000108	0.000000
MTHFR	1	0.000000	0.238457	0.000000	0.000108	0.000000
RPGRIP1L	1	0.000000	0.270930	0.000000	0.000108	0.000000
GRIK4	1	0.000000	0.229498	0.000000	0.000108	0.000000

Fonte: Tabela gerada pelo autor.

4.2 Remoção de um único gene semente

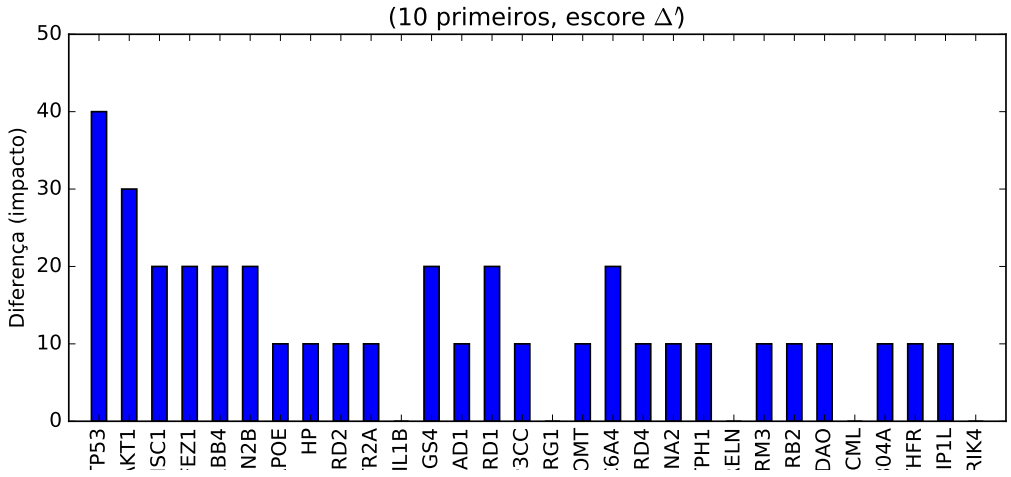
Inicialmente, avaliamos o impacto da remoção de um único gene semente na lista resultante. Esta avaliação foi realizada de forma similar ao método *Leave-One-Out Cross-Validation*. A ideia deste experimento foi avaliar o impacto individual de cada gene no resultado final, com relação aos dois escores X e Δ', obtidos durante a análise da rede diferencial no método *NERI*. Em seguida, comparamos o impacto de tais resultados com as medidas de centralidade de redes para avaliar se há alguma correlação. Desta forma, esta informação pode ser utilizada para lançar luz sobre o impacto das remoções de múltiplos genes sementes.

4.2.1 Estudo dos gráficos em relação ao escore Δ'

4.2.1.1 Análise dos 10 primeiros elementos

A figura 4.1 apresenta um gráfico comparativo dos experimentos utilizando o método de *remoção de um único gene*, onde o eixo *Horizontal* representa o gene removido em relação a amostra original, e o eixo *Vertical*, por sua vez, representa a diferença percentual dos genes ranqueados em relação ao experimento original. Desta forma, comparando os 10 primeiros genes ranqueados relativos a remoção de cada gene apresentado, em relação aos 10 primeiros apresentados na amostra original, sendo o fator de ranqueamento o escore Δ'.

Figura 4.1 – Análise dos 10 primeiros elementos ordenados por Δ'.



Fonte: Produzido pelos autores.

Podemos observar que os genes de maior grau ***TP53*** (333) e ***AKT1*** (138), apresentaram impactos um pouco maior que os demais que foram respectivamente de ***40%*** e ***30%***. No entanto, ao comparar com o impacto dos demais genes que não têm graus tão altos, observamos que, de uma forma geral, os impactos não foram diretamente proporcionais aos graus dos genes.

Observamos também que os genes ***IL1B, RELN, NRG1, GRIK4*** e ***OPCML*** não apresentaram mudanças no resultado em relação ao escore analisada (Δ'), assim como o gene ***MTHFR*** e os outros que apresentaram ***10%*** de diferença dos genes selecionados, comparado ao resultado original do experimento. Devido a isto, podemos presumir de que tais genes não apresentam uma importância significativa para o método em estudo em relação aos 10 primeiros selecionados utilizando o fator de ranqueamento o escore Δ'.

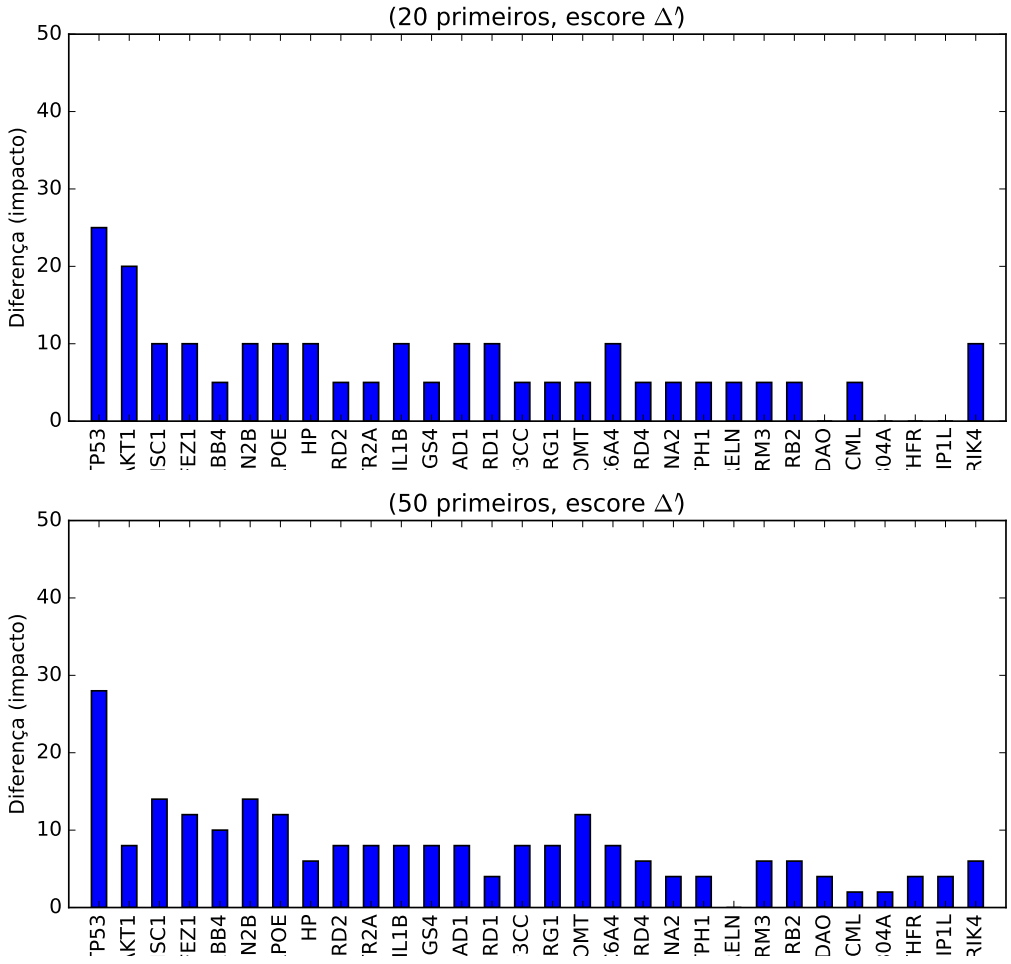
Os genes ***CHRNA7, DAOA, DTNBP1, MUTED, NPAS3, OFCC1, PRODH*** e ***SLC18A1*** não foram integrados com a rede PPI. Ou seja, durante a integração de

dados, tais genes não possuíam um nó correspondente na rede PPI e, portanto, não foram utilizados.

Ao analisar os genes mencionados anteriormente ***TP53*** e ***AKT1***, ambos possuem um alto grau na rede gerada pelo método NERI. E isto pode sugerir que a remoção de um gene com alto grau influencia diretamente no resultado. Por outro lado, os demais genes deram valores bastante parecidos, oscilando entre 0% à 10%. Por exemplo, os genes ***ZNF804A***, ***MTHFR*** e ***RPGRIP1L*** possuem grau ***1***, e no entanto, causaram *10%* de impacto em relação ao experimento original, onde apresenta-se semelhante a remoção do terceiro gene, o ***DISC1*** com grau *91*. Isso demonstra que a correlação entre o grau e o impacto observado na lista resultante final não é direta.

4.2.1.2 Análise dos 20 e 50 primeiros elementos

Figura 4.2 – Análise dos 20 e 50 primeiros elementos ordenados por Δ'.



Fonte: Produzido pelos autores.

A figura 4.2 apresenta dois gráficos de forma comparativa, de modo que o de cima representa os ***20*** primeiros genes ranqueados resultantes em relação ao escore Δ' e o gráfico

de baixo apresenta os **50** primeiros. Sendo eixo *Vertical* a similaridade com o resultado original e o eixo *Horizontal* o gene removido no experimento em questão.

A remoção do gene **MTHFR** apresentou baixo impacto: ***0%*** de diferença nos primeiros ***20*** elementos e ***5%*** em relação aos primeiros ***50*** elementos da lista original. Este mesmo efeito aconteceu na remoção do gene ***RELN***, variando de ***0%*** de diferença percentual, dos genes selecionados em relação ao experimento original, nos primeiros ***20*** elementos para ***5%*** em relação aos primeiros ***50***.

Podemos observar também uma diminuição de experimentos com ***10%*** ou menos de impacto. Caindo de ***36*** experimentos ao todo e ***28*** válidos, para ***31*** ao todo e ***23*** válidos (Os experimentos não válidos para análise são os que não integraram com a base de dados ***GWAS***, totalizando ***8*** *genes/experimentos*).

O gene ***TP53*** que causa o maior impacto na similaridade em sua remoção, variou de ***25%*** para ***29%*** nos respectivos agrupamentos ***20*** e ***50*** primeiros genes selecionados. Isso implica que ao analisar ***30*** elementos a mais, houve um aumento do impacto de ***4%*** no pior caso. Sugerindo uma boa robustez do método em relação a remoção de um único gene semente.

Outro ponto importante de observação, é o gene ***AKT1*** que possui um alto grau na rede gerada pelo método NERI, apresentar uma variação de impacto no resultado bem alta. O impacto apresentado vai de ***20%*** para ***8%***, nos gráficos representantes dos ***20*** e ***50*** primeiros genes, respectivamente. Este fator aponta, mais um vez, que o grau do gene relativo a rede não causa um impacto diretamente proporcional ao grau.

4.2.1.3 Análise dos 100 e 200 primeiros elementos

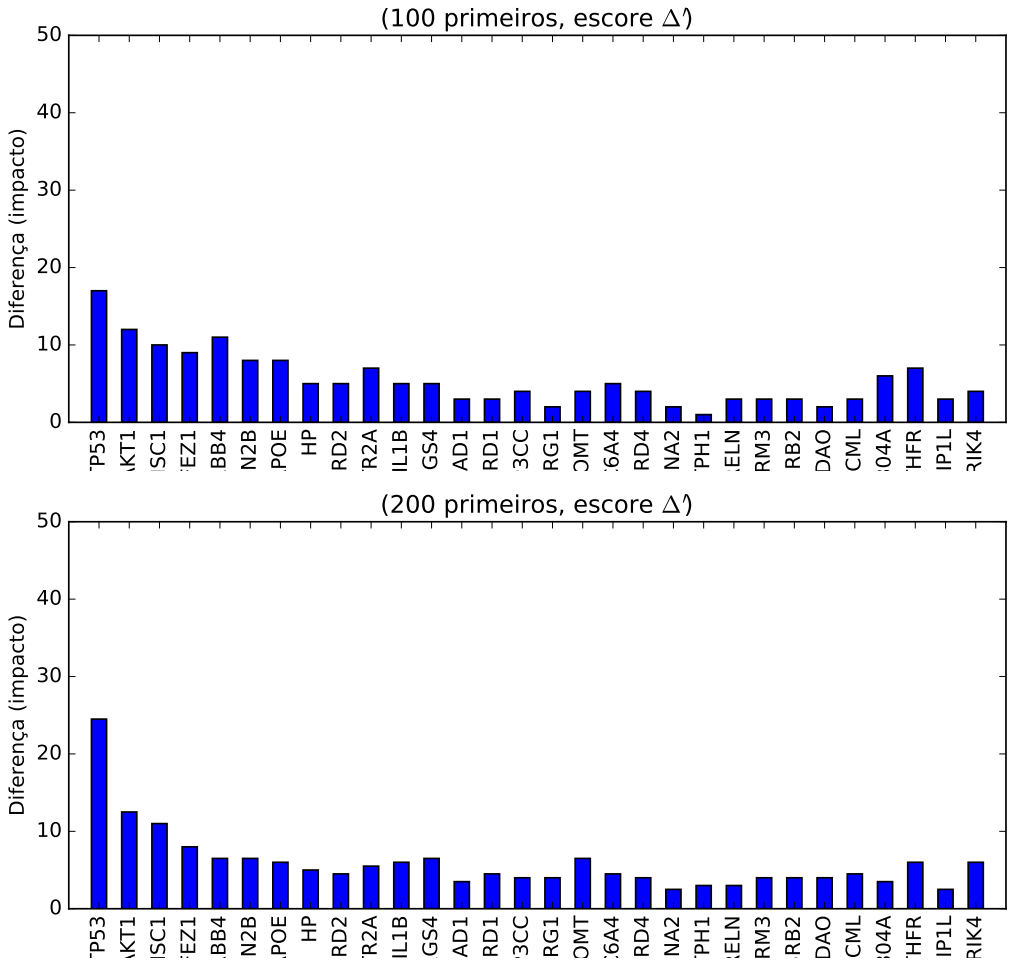
A imagem 4.3 também apresenta dois gráficos comparativos em relação a similaridade do ranqueamento gênico baseado na quantidade de elementos selecionados. O gráfico de cima, apresenta a diferença em percentual da similaridade dos ***100*** primeiros genes e o de baixo os primeiros ***200***.

Podemos observar em ambos os gráficos, que neste ponto de análise não houve remoção de gene que não causou impacto na similaridade dos resultados em relação a amostra original.

Um fato importante que ambos os gráficos demonstram, é a mediana dos impactos, onde apresentaram um valor abaixo de ***10%***. Isso demonstra que o impacto gerado pela remoção de um único gene semente varia em torno de ***10%***, o que é muito aceitável em vista da assertividade obtida.

Ao observarmos o gene ***TP53***, que durante todo o experimento foi o que apresentou maior impacto no resultado, podemos notar que o mesmo apresentou um aumento de ***17%*** do impacto para ***24%*** ao compararmos os ranqueamentos de ***100*** e ***200*** genes. Este valor

Figura 4.3 – Análise dos 100 e 200 primeiros elementos ordenados por Δ'.



Fonte: Produzido pelos autores.

pode ser considerado baixo, em vista que o numero de elementos da lista analisada foi dobrado e a diferença do impacto gerado foi de $\mathcal{7\%}$.

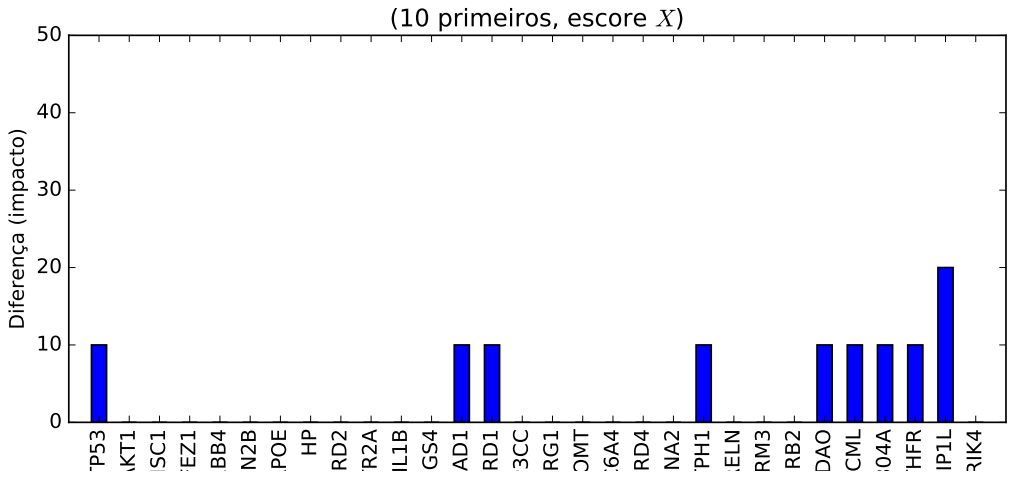
Um comportamento que aparece em ambos os gráficos, que não havia aparecido com tanta nitidez nos anteriores, é a decrescência do impacto da esquerda para direita. Onde os experimentos estão ordenados com relação ao seu grau na rede interna gerada pelo método NERI. Isso implica que mesmo não tendo uma relação direta com o impacto o grau do gene semente apresenta correlação ao analisar uma quantidade maior de genes priorizados do resultado final.

4.2.2 Estudo dos gráficos em relação ao escore X

4.2.2.1 Análise dos 10 primeiros elementos

A figura 4.4 apresenta um gráfico comparativo referente aos resultados dos experimentos com remoção de apenas *um gene semente* por vez. O eixo **Horizontal** representa

Figura 4.4 – Análise dos 10 primeiros elementos ordenados por X.



Fonte: Produzido pelos autores.

cada experimento com seu respectivo **gene semente** removido do agrupamento original, estando estes ordenados pelo grau que representa na rede gerada pelo método NERI. Sendo esta organização, do maior para o menor no sentido esquerda para direita. O eixo *Vertical*, por sua vez, representa a *diferença percentual* dos genes ranqueados em relação ao experimento original, tendo como fator de ordenação o escore $\boldsymbol{X}$. Por conseguinte, apresentando a comparação dos *10* primeiros genes ranqueados, sendo estes em relação a remoção do respectivo *gene semente* apresentado, com os 10 primeiros ranqueados pelo *experimento original*.

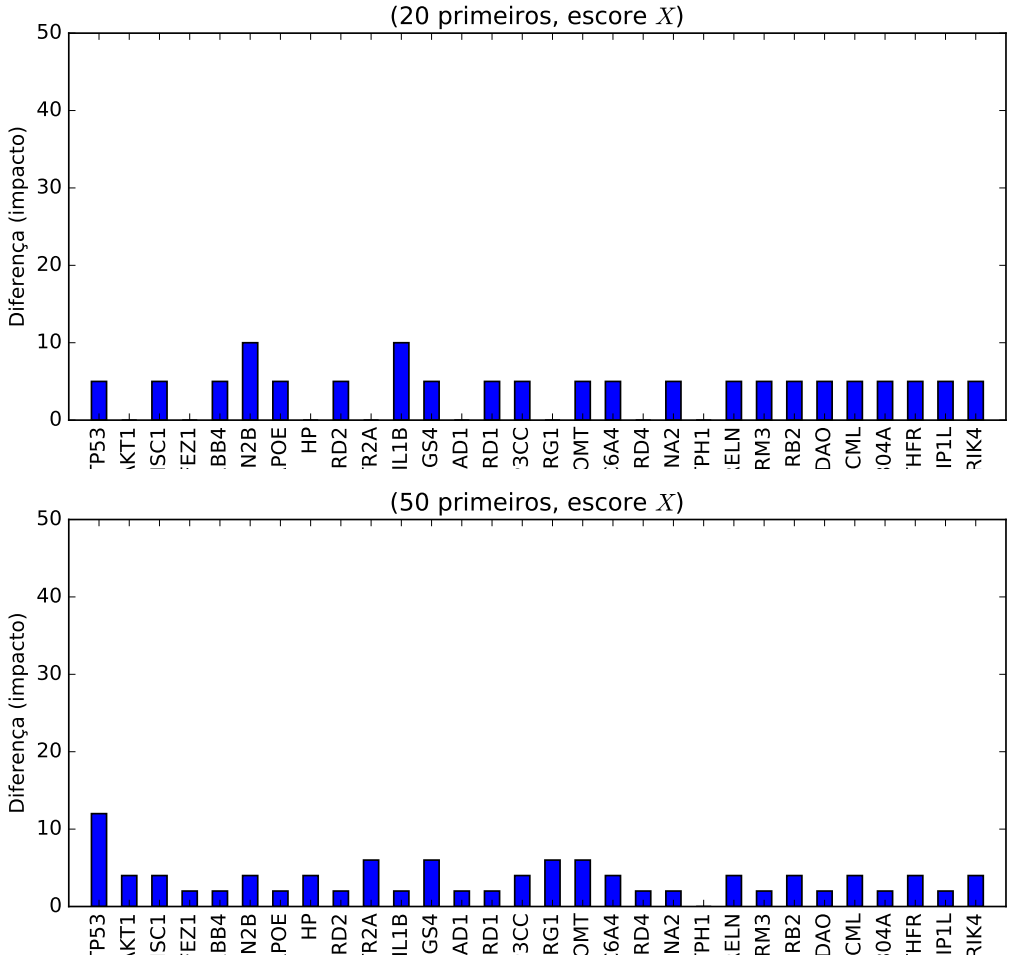
Em primeira análise, podemos perceber que o gene semente ***RPGRIP1L*** causou o maior impacto em sua remoção, apresentando o percentual de ***20%*** de diferença dos ***genes ranqueados*** em relação a amostra original. Este impacto apresenta um comportamento de ***outlier*** em relação ao agrupamento de experimentos, em vista que a mediana dos impactos foi de ***0%***, ou seja, a maior parte dos experimentos em questão não apresentaram diferença entre os *genes ranqueados* com o *experimento original*.

Podemos observar também que apenas ***9*** genes apresentaram impacto em sua remoção. Sendo entre estes, a mediana do impacto ***10%***, o que é um bom indicador da robustez do método NERI.

Um ponto importante para observação é a mudança do gene causador de maior impacto em relação aos ***fatores de ranqueamento***. Quando o ranqueamento foi feito pelo escore Δ' (analisado na seção anterior), o gene semente causador de maior impacto foi ***TP53***, este no qual aprestou apenas ***10%*** de impacto em relação ao escore de ranqueamento $\boldsymbol{X}$. Isso indica que ambos os genes são impactantes no resultado final do experimento, porém, a abordagem adotada para ranqueamento gênico influencia diretamente na análise.

4.2.2.2 Análise dos 20 e 50 primeiros elementos

Figura 4.5 – Análise dos 20 e 50 primeiros elementos ordenados por X.



Fonte: Produzido pelos autores.

A figura 4.5 apresenta dois gráficos comparativos, demonstrando o impacto causado no ranqueamento gênico pelo método NERI ao remover determinados genes sementes. Cada gráfico representa o impacto relativo a quantidade dos genes priorizados analisados, de forma que o eixo *Horizontal* represente os genes removidos em cada experimento, estando ordenados da esquerda para direita levando em conta grau do gene semente. Assim sendo, o eixo *Vertical* indica o impacto causado no resultado final em relação a amostra original, este impacto se dá pela diferença percentual de genes presentes no agrupamento experimento e agrupamento original. Estão representados no gráfico de cima, o impacto causado em relação ao ranqueamento dos ***20*** primeiros genes, e os ***50*** primeiros no gráfico de baixo.

Podemos observar em primeira observação, que a quantidade de genes que causaram impacto no ranqueamento gênico. Onde o impacto aumentou consideravelmente logo nos primeiros ***20*** genes analisados, sendo ***22*** experimentos impactantes. Diferente dos ***10***

primeiros, como pode ser visto no gráfico da figura 4.4, onde apenas ***9*** experimentos apresentaram impacto. Este mesmo comportamento pode ser observado, ao compararmos o gráfico dos ***20*** primeiros genes, com o gráfico dos ***50*** primeiros. Este ultimo apresenta apenas ***1*** experimento que não causou impacto em sua remoção, sendo ele a remoção do gene ***TPH1***.

Quando olhamos para o experimento com a remoção do gene ***TPH1***, podemos notar um comportamento atípico. O mesmo apresentou um impacto no ranqueamento gênico de ***10%*** nos primeiros ***10*** genes analisados. Porém, nos dois gráficos subsequentes, representando respectivamente ***20*** e ***50*** primeiros genes ranqueados, o mesmo não apresentou impacto. Ao observarmos este comportamento, podemos inferir que este gene não é de grande importância para o ranqueamento gênico feito pelo método NERI.

O gráfico que representa os ***20*** primeiros genes ranqueados em relação a ***X***, apresenta uma mediana de impacto de apenas ***5%***, valor este, considerado muito baixo, em vista que estes são os genes considerados mais importantes pelo método NERI.

Uma outra ótica que podemos incutir aos experimentos é a observação em relação ao grau dos genes sementes removidos. Os gráficos demonstram que em relação ao escore ***X***, o grau é um fator altamente impactante no resultado final, em vista que os genes com maior grau ***TP53***, ***AKT1*** e ***DISC1*** não apresentaram em sua maioria os maires impactos no resultado final. Ficando esta métrica salva apenas para o ***TP53*** na análise dos ***50*** primeiros genes ranqueados, onde seu impacto é de ***22%*** em relação a amostra original.

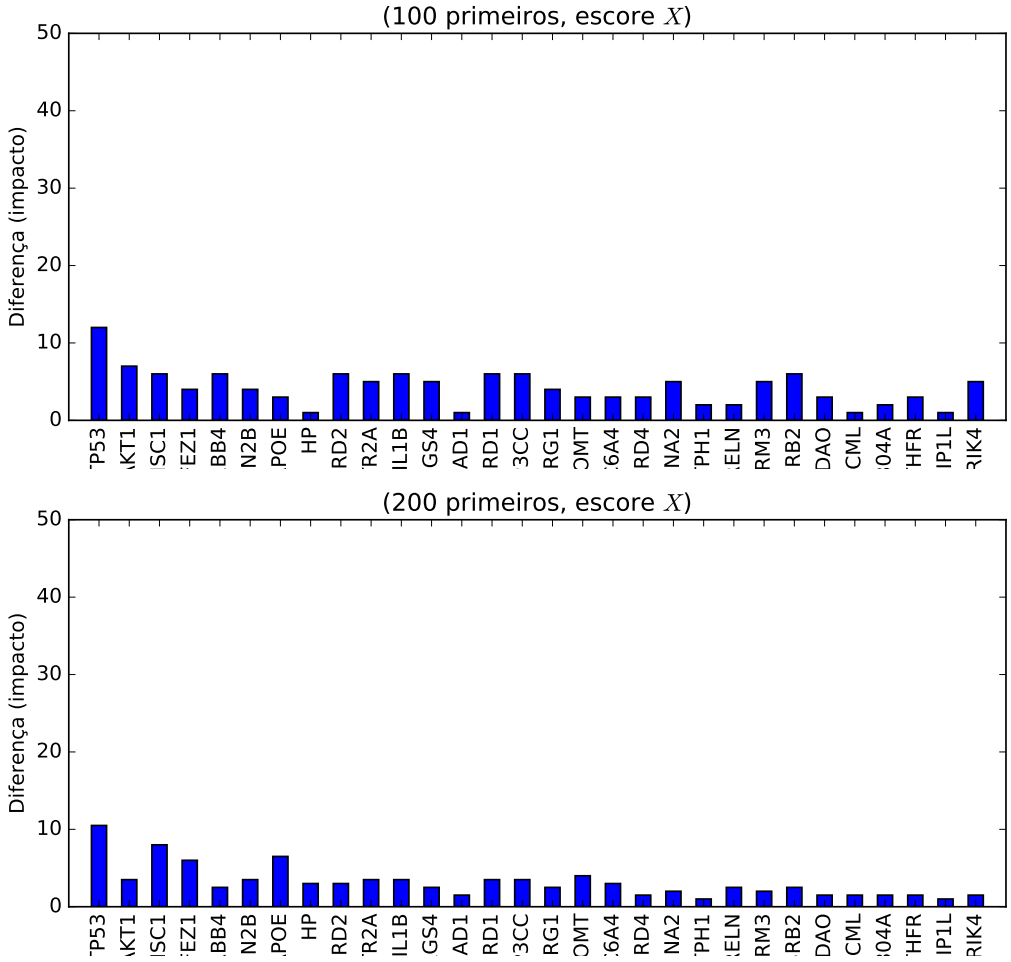
4.2.2.3 Análise dos 100 e 200 primeiros elementos

A Figura 4.3 apresenta dois gráficos comparativos em relação ao impacto representado pela remoção do ***gene semente respectivo***. Estes gráficos podem ser lidos da mesma forma que os apresentados anteriormente nesta sessão, onde o de cima representa os primeiros ***100*** genes ranqueados e o de baixo os ***200*** primeiros.

Um comportamento que podemos observar ao analisar os dois gráficos apresentados, é a diminuição do impacto geral causado pela remoção dos genes sementes encontrados nos primeiros ***200*** genes ranqueados. Esta diminuição no impacto se dá pelo aumento da lista de genes ranqueados, de forma que mais genes em comum estejam nas listas geradas pelos experimentos e na lista do experimento original. Este comportamento indica uma tendência de diminuição do impacto em relação ao tamanho da lista analisada, porém, o gene semente causador de maior impacto em sua remoção ainda se mantém, sendo ***TP53*** com ***12%*** e ***10%*** de impacto nos gráficos de ***100*** e ***200*** genes ranqueados respectivamente.

Também podemos notar um limiar de impacto abaixo de ***10%***, ou seja, a maioria dos experimentos causaram um impacto igual ou inferior a ***10%*** no resultado final em relação

Figura 4.6 – Análise dos 100 e 200 primeiros elementos ordenados por X.



Fonte: Produzido pelos autores.

a remoção do gene semente respectivo. Este comportamento é um bom indicativo para a robustez do método, em vista que o mesmo se mostra com baixa variação no resultado em relação a remoção de um único gene semente, quando o fator de ranqueamento é o escore X.

4.2.3 Observações

Em relação ao método de validação de *remoção de um único gene*, o método NERI apresentou bons resultados de robustez. De forma que o maior impacto encontrado pela remoção de um único gene semente foi de ***40%*** em relação ao escore Δ'. Porém a mediana das correlações com o mesmo escore de ranqueamento, foi de ***20%*** em relação aos ***10*** primeiros elementos. Estes valores melhoram ao observar o ranqueamento em relação ao escore X, apresentando o maior valor de impacto em ***20%*** com o gene ***GRIP1L***. Fato este que chama atenção por ser um gene semente diferente do maior causador de impacto em relação ao escore Δ', o gene semente ***TP53***. A mediana de impacto nos primeiros ***10***

genes ranqueados pelo escore X foi de ***0%***, ou seja, a remoção de mais da metade dos genes sementes individualmente não causou impacto no resultado final. Isso significa que os na maioria dos casos os genes ranqueados tanto nos experimentos quanto na amostra original foram os mesmos.

Também deve-se levar em conta que o maior impacto encontrado relação aos ***200*** primeiros genes selecionados pelo escore Δ', foi de ***24%*** apresentando melhora em relação a análise dos **10** primeiros. A medina apresentou uma queda significativa de **20%** para ***5%***, o que indica uma convergência de genes selecionados da amostra original com os experimentos. Esse tipo de convergência é esperado com o aumento da quantidade de elementos ranqueados, pois a probabilidade de um gene ser selecionado aumenta proporcionalmente ao tamanho do agrupamento de seleção final. Porém, esta premissa não invalida a eficiência do método em questão, em vista que a quantidade de genes possíveis a serem selecionados é muito maior que a lista dos genes ranqueados. Assim sendo, podemos concluir que o método é robusto em relação a retirada de um único gene semente. Porém, para determinar melhor a robustez do método em análise, há a necessidade de estudar os resultados dos outros modelos de validação empregadas neste trabalho.

4.3 Remoção de vários genes sementes

Nesta etapa analisaremos os resultados do método de *Remoção de vários gene sementes*, onde como principal meio de apresentação de dados será o estudo dos gráficos gerados e a discussão de suas interpretações.

4.3.1 Esperado

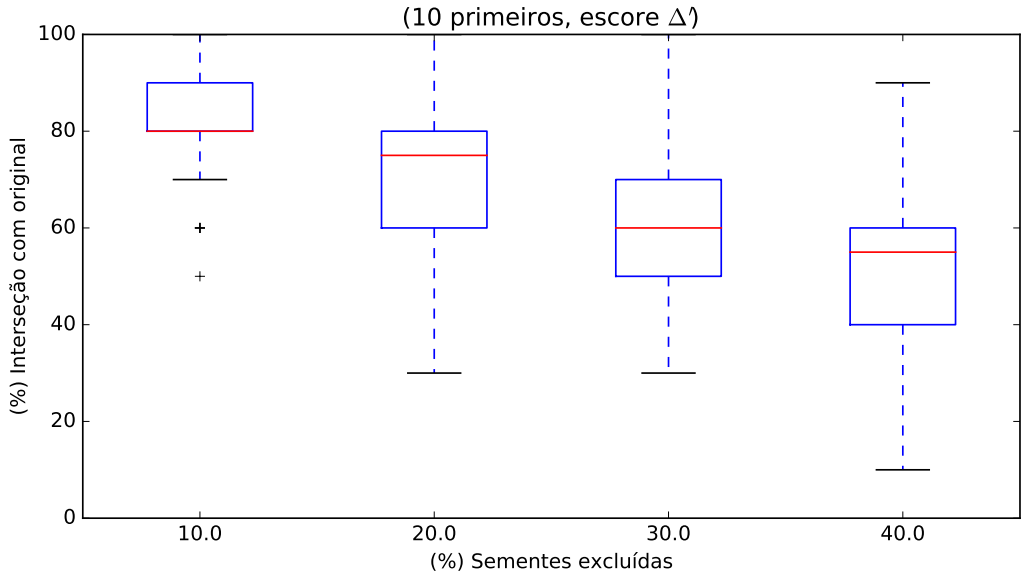
O esperado na utilização do método da *Remoção de mais de um gene* é a capacidade de mapear o impacto causado no resultado final baseado na identificação da quantidade de genes sementes removidos em relação a amostra original. Desta forma observar o impacto causado, a medida que conjuntos de tamanhos diferentes são testados. Com este estudo, podemos aproximar um limiar de confiança no método NERI. Para podermos ter uma análise mais precisa dos resultados, cruzaremos os dados encontrados com os dados gerados pela *Remoção de um único gene*, de forma que consigamos entender melhor o comportamento dos resultados apresentados.

4.3.2 Estudo dos gráficos em relação ao escore Δ'

4.3.2.1 Análise dos 10 primeiros elementos

A Figura 4.7 apresenta um gráfico comparativo dos experimentos utilizando o método *Remoção de mais de um gene*, onde o eixo *Horizontal* representa a porcentagem de

Figura 4.7 – Análise dos 10 primeiros elementos ordenados por Δ'.



Fonte: Produzido pelos autores.

sementes excluídas em relação a amostra original, e o eixo *Vertical*, por sua vez, representa a porcentagem da interseção dos resultados dos ***10*** primeiros genes em relação aos ***10*** primeiros apresentados na amostra original, tendo como fator de ranqueamento ao escore Δ'.

Conforme podemos observar, a medida que aumentamos o percentual de sementes excluídas, ocorre uma redução gradual na mediana do percentual de interseção de genes selecionados. Conforme esperado, isso demonstra que a quantidade de genes sementes excluídas influencia diretamente no resultado do experimento.

Podemos observar também que, a mediana do experimento com a maior porcentagem de remoção apresenta o valor aproximado ***50%*** de interseção, esta métrica é um bom indicativo da robustez do método ao informar que mesmo removendo ***40%*** dos genes sementes, os genes ranqueados pelo método ainda se mantém acima de ***50%*** iguais aos ranqueados pelo método com todos os genes sementes.

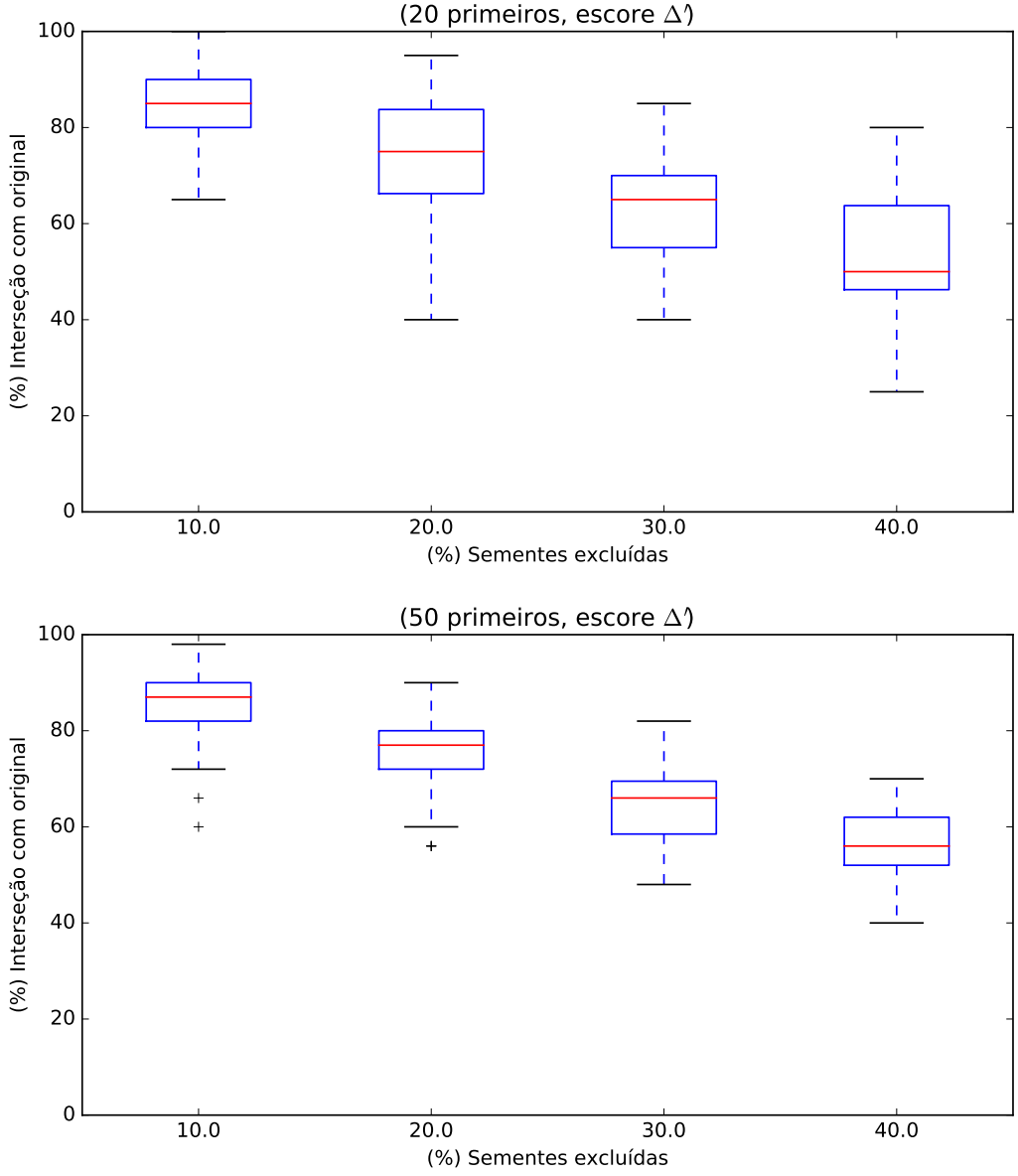
Por este gráfico representar somente os ***10*** primeiros genes selecionados, esperava-se um impacto no resultado final consideravelmente alto devido ao fato do mesmo apresentar poucos genes em relação ao tamanho da rede ***9.554*** Genes (nós). Porém, ao contrário do que imaginávamos, os genes selecionados foram muito próximos da amostra original. Porém a sua precisão varia consideravelmente de modo que a podemos observar que as diferenças entre os limites superiores e inferiores dos experimentos são altas. Isto se dá devido ao tamanho da lista de genes priorizados analisada.

Um ponto que chama bastante atenção ao analisar este gráfico, é o boxplot que

representa ***40%*** dos genes sementes removidos. O mesmo, apresenta uma variação entre o lime inferior e o limite superior de ***60%***, ou seja, a bateria de experimentos representados pelo gráfico possui experimentos de similaridade variada entre ***20%*** a ***80%***. Fato este implica em uma não confiança nos dados representados por este, o comportamento apresentado é reforçado ao fazer uma análise dos ***outliers***. Encontrando um experimento com ***90%*** de similaridade e em contrapartida, um experimento com apenas ***10%*** sendo o menor de todo o estudo em relação aos ***10*** primeiros genes removidos.

4.3.2.2 Análise dos 20 e 50 primeiros elementos

Figura 4.8 – Análise dos 20 e 50 primeiros elementos ordenados por Δ'.



Fonte: Produzido pelos autores.

A Figura 4.8 apresenta dois gráficos comparativos, sendo o de cima representando a correlação com os ***20*** primeiros genes selecionados e de baixo com os primeiros **50**.

Podemos observar no primeiro gráfico a baixa variação da mediana, porém houve uma diminuição na *amplitude interquartílica*. Isto demonstra uma possível convergência de resultados em relação aos dois gráficos. Este fator pode ser observado no *boxplot* referente a ***20%*** do gráfico de cima, onde este representa ***20*** primeiros genes selecionados. Neste caso, *amplitude interquartílica* varia de *68%* a *82%*, totalizando *14%* de faixa de variação. No gráfico de baixo, representando os ***50*** primeiros genes selecionados. Neste caso, há uma variação na *amplitude interquartílica* de ***73%*** a ***80%***, totalizando uma faixa de variação de ***7%***, valor este que apresenta-se metade do valor do gráfico anterior. Esta queda de amplitude remete ao comportamento de convergência, assim representando uma segurança nos resultados apresentados, partindo do princípio de que quanto menor a variação dos resultados, maior é a precisão da medição.

Podemos observar também alguns experimentos que ficaram fora do agrupamento, este comportamento é definido como *outlier*. Para entender o porque destes experimentos terem sido apresentados unanimamente com menores resultados do que o corpo amostral, cruzamos os seus dados com os obtidos pela etapa de *remoção de um único gene*. Com este cruzamento de dados, observamos se os genes removidos nestes experimentos contém um ou mais genes que possuem os maiores *graus de impacto* no resultado.

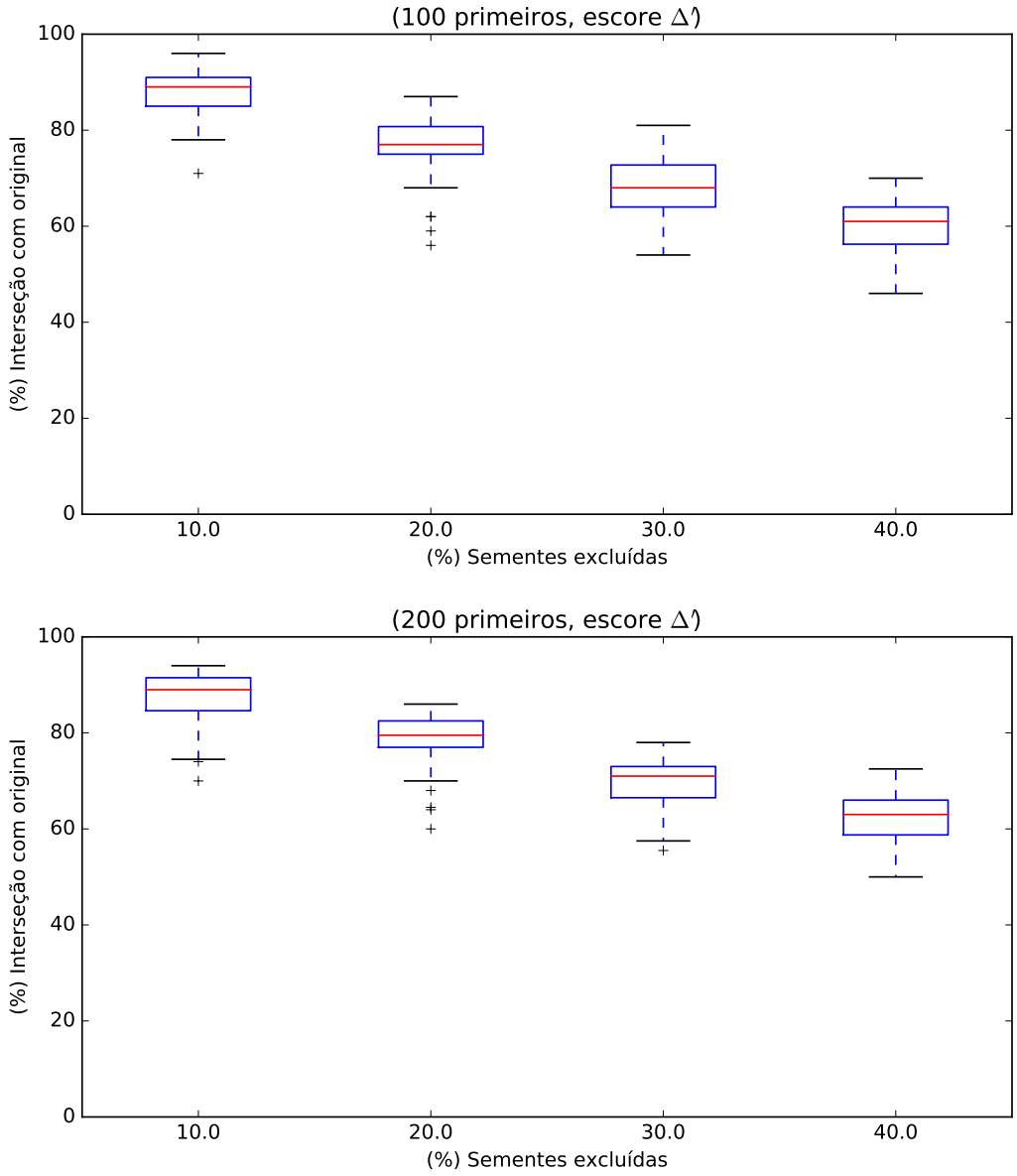
Em 10 experimentos essa premissa apresentou-se verdadeira, resultando menores correlações, onde nestes casos observou-se a falta dos genes sementes ***TP53*** e ***AKT1***. Onde ambos causaram o maior impacto no resultado final ao serem removidos sozinhos do experimento (como foi mencionado na sessão anterior). Estes casos apontam uma correlação do impacto acumulativo da remoção de genes sementes no resultado final.

4.3.2.3 Análise dos 100 e 200 primeiros elementos

A figura 4.9 apresenta dois gráficos comparativos entre agrupamentos de experimentos com variação na quantidade de genes sementes. O gráfico *de cima*, representa a comparação dos ***100*** primeiros genes ranqueados em relação ao escore Δ', onde o eixo *Horizontal* define os *boxplots* correspondentes as suas determinadas porcentagens de retirada dos ***genes sementes***. Assim sendo, o eixo *Vertical* define a porcentagem de interseção dos genes ranqueados pelos experimentos em relação ao experimento original. Seguindo esta mesma organização, o gráfico *de baixo* representa os ***200*** primeiros genes ranqueados.

Podemos notar claramente, que o decrescimento correlacional está presente nos dois gráficos. Ou seja, a correlação dos ***genes ranqueados*** cai em mesma proporção nos dois gráficos conforme a quantidade de ***genes sementes*** são reduzidas. Porém, podemos enxergar que a ***amplitude interquartílica*** respectiva entre os gráficos apresenta uma diminuição. Este aspecto representa bons resultados, pois indica que há uma convergência

Figura 4.9 – Análise dos 100 e 200 primeiros elementos ordenados por Δ'.



Fonte: Produzido pelos autores.

de resultados conforme o aumento dos ***genes ranqueados*** observados.

Nesta comparação, podemos observar novamente comportamentos de ***outlier*** presentes nos gráficos. Como na análise anterior, os agrupamentos que apresentaram menor correlação, foram os que não tinham em seu agrupamento de ***genes sementes*** os mais impactantes observados na etapa de ***retirada de uma único gene semente***, sendo eles ***TP53*** e ***AKT1***.

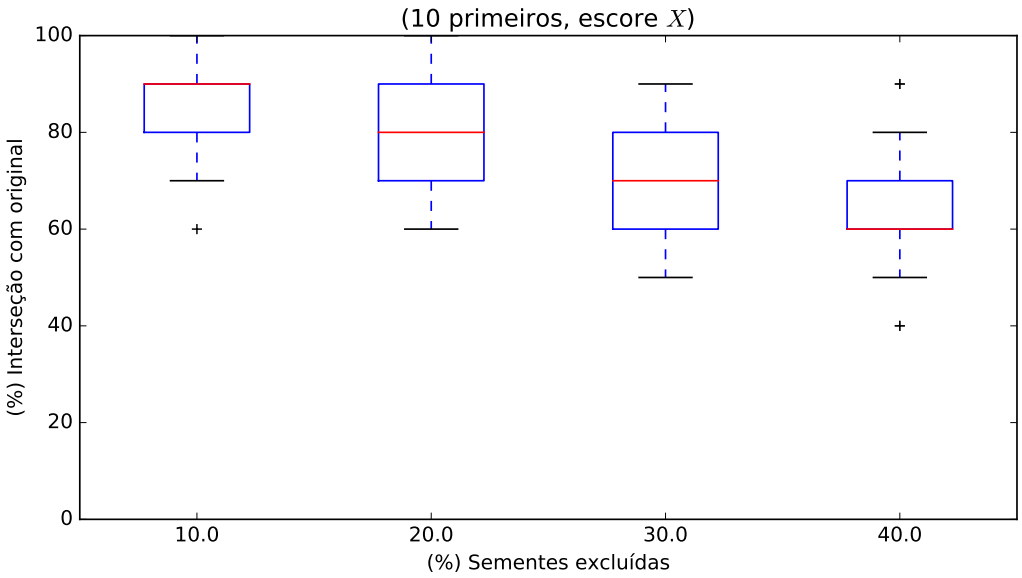
Um forte fator de análise é a comparação entre o gráfico dos ***10*** primeiros genes ranqueados (4.7) com o gráfico dos ***100*** (4.9). Podemos observar que a mediana subiu de ***80%*** de correlação com a amostra original, para ***90%*** ao comparar os ***boxplots***

pertencentes a ***10%*** de remoção. Este fato aponta para uma robustez do método analisado, em vista que fortalece ainda mais o efeito de convergência observado anteriormente. Ao comparar com o gráfico dos ***200*** genes ranqueados, notamos que a ***mediana*** se mantém a mesma em relação a dos ***100***, indicando que esta convergência ocorre entre nos primeiros ***100*** genes ranqueados, sendo este um número muito bom. O mesmo pode ser observado ao comparar os outros ***boxplots*** respectivos, os valores apresentados não são os mesmos, mas apresentam um padrão muito próximo de variação.

4.3.3 Estudo dos gráficos em relação ao escore X

4.3.3.1 Análise dos 10 primeiros elementos

Figura 4.10 – Análise dos 10 primeiros elementos ordenados por X.



Fonte: Produzido pelos autores.

A Figura 4.10 apresenta um gráfico comparativo dos experimentos utilizando o método *Remoção de mais de um gene*, onde o eixo *Horizontal* representa a porcentagem de sementes excluídas em relação a amostra original, e o eixo *Vertical*, por sua vez, representa a porcentagem da interseção dos resultados dos ***10*** primeiros genes em relação aos ***10*** primeiros apresentados na amostra original, tendo como fator de ranqueamento o escore X. Em primeira análise, fica claro que, assim como o ranqueamento pelo escore de ranqueamento Δ', quanto maior a quantidade de genes sementes removidos nos experimentos, a correlação das listas de priorização gênica diminui. Fator este esperado, em vista que o método NERI utiliza os genes sementes para realizar a priorização gênica.

Ao compararmos este gráfico com o apresentado na análise do escore Δ' como pode ser vista na Figura 4.7, podemos notar que a variação dos resultados dos experimentos

é muito menor, indicando que o escore X tende a ser mais robusto. O boxplot que intuitivamente apresentaria maior diferença de limite superior e inferior, o referente a ***40%*** de remoção de genes sementes, não apresentou uma grande variação. Este comportamento é o contrário do observado anteriormente, onde a variação anterior apresentou-se em ***60%***, diferentemente do gráfico em relação ao escore X apresentando ***30%***, metade do valor anterior. Sugerindo mais uma vez a robustez do escore de ranqueamento X superior ao escore Δ'.

Podemos notar que a mediana do pior caso ficou em ***60%*** de similaridade com a lista de genes ranqueados em relação ao experimento original. O pior caso sendo determinado intuitivamente pelo conjunto de experimentos com ***40%*** de remoção dos genes sementes em relação ao experimento original. A variação entre a menor e maior mediana, sendo elas respectivamente ***60%*** e ***90%***, apresenta-se em ***30%***. Um valor muito bom se levarmos em consideração que no pior caso foram removidos ***40%*** dos genes sementes da amostra original, ou seja, a variação do impacto proporcional causado foi menor que o fator de remoção de genes sementes em relação a amostra original. Isto indica uma boa robustez do método NERI em relação ao fator de ranqueamento X.

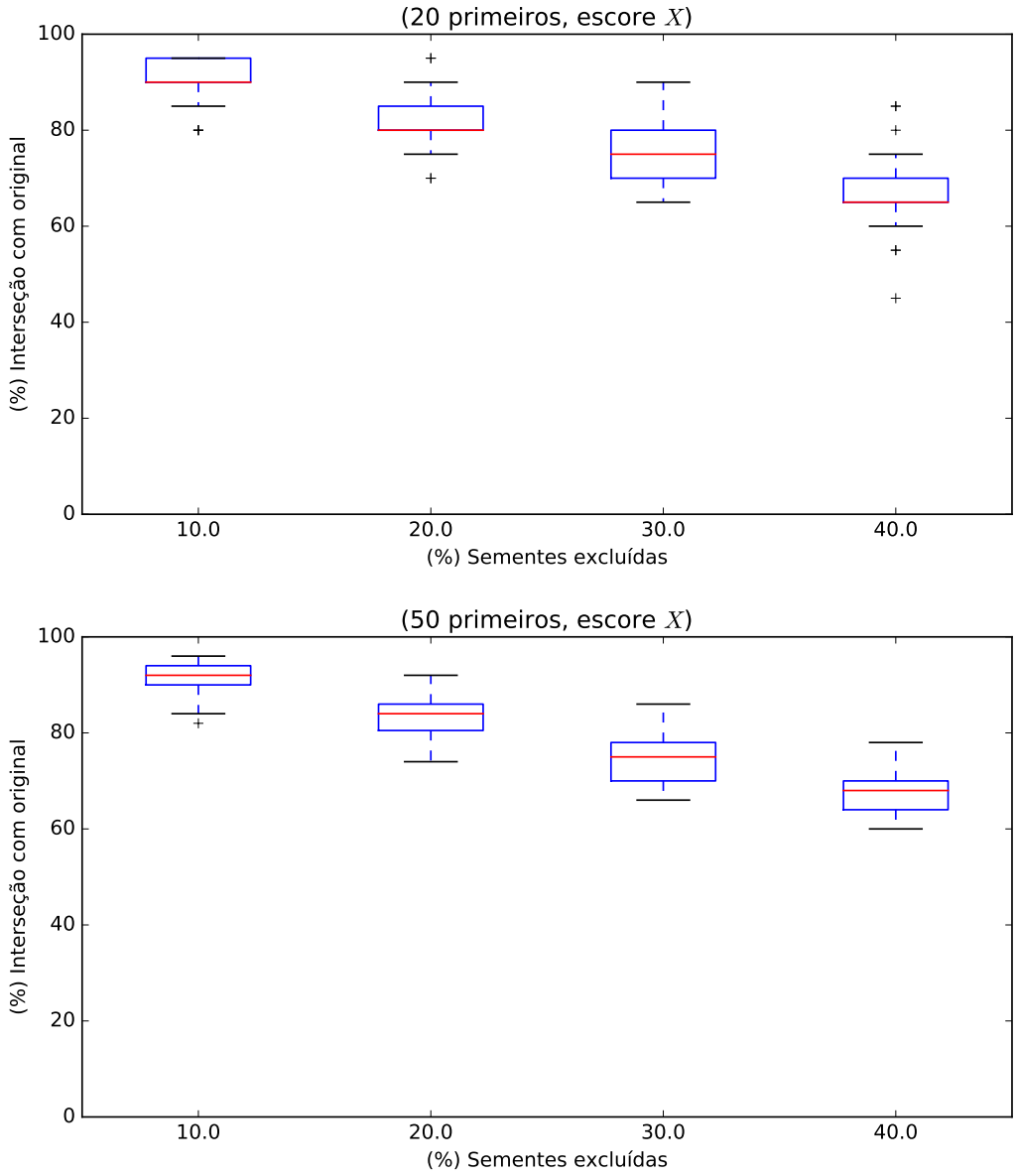
4.3.3.2 Análise dos 20 e 50 primeiros elementos

A Figura 4.11 apresenta dois gráficos comparativos dos experimentos utilizando o método *Remoção de mais de um gene*, onde o gráfico de cima representa os ***20*** primeiros genes ranqueados pelos experimentos e o de baixo os primeiros ***50***. Ambos no eixo *Horizontal* apresentam as porcentagens de genes sementes removidos em relação a amostra original e no eixo *Vertical*, a porcentagem de similaridade do resultado dos experimentos com o resultado original, ou seja, a similaridade das listas dos experimentos em relação a lista de ranqueamento gênico original.

Pode-se notar que, as amplitudes amostrais diminuíram em ambos os casos, isso demonstra uma menor variação dos experimentos em relação a análise feita dos ***10*** primeiros genes. Este comportamento, era intuitivamente esperado, em vista que ao aumentar a quantidade de genes selecionados na lista de ranqueamento, a probabilidade da variação dos resultados diminuírem aumenta. Porém, como são muitos genes na rede, o aumento de ***10*** e ***40*** genes ranqueados em relação a análise anterior, foi o suficiente para identificação. Apesar de ter sido esperado, representa um bom sinal de robustez, de modo que a baixa variação do resultado seja um fator para a mesma.

Nota-se também que no gráfico que representa os ***20*** primeiros genes ranqueados, quando analisou-se os experimentos que tiveram ***40%*** dos genes sementes removidos em relação a amostra original, apresentaram experimentos ***ouliers*** onde alguns representavam uma boa correlação e outros uma má correlação. Isto indica que devemos observar os genes presentes nestes experimentos, para assim tentarmos entender o comportamento

Figura 4.11 – Análise dos 20 e 50 primeiros elementos ordenados por X.



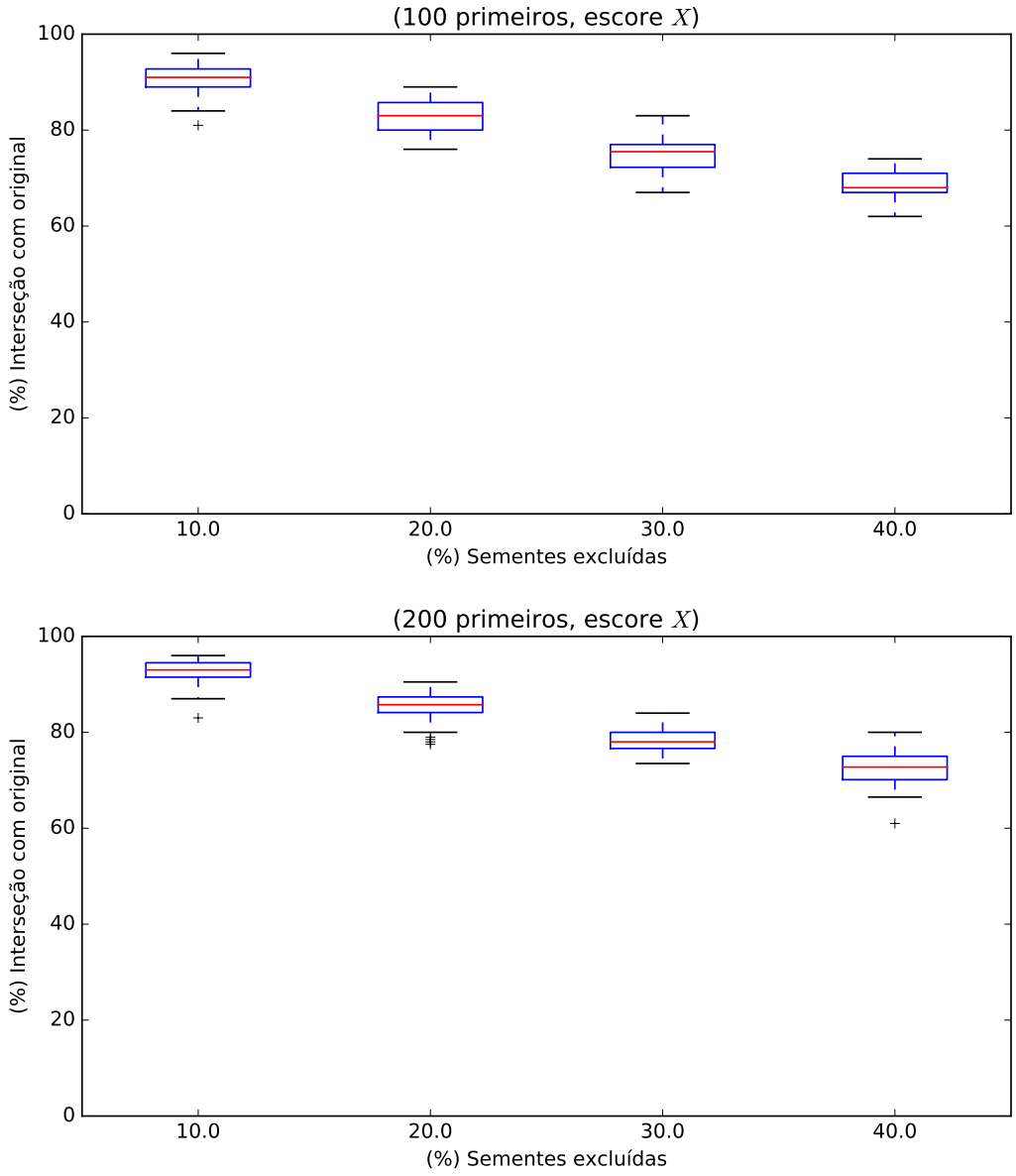
Fonte: Produzido pelos autores.

diferenciado. Os experimentos ***outliers*** que apresentaram uma má correlação, tiveram entre os seus genes sementes removidos os seguintes elementos ***TP53*** e ***RPGRIP1L***. Estes genes sementes, são os que apresentaram um maior impacto em sua remoção única na etapa de *Remoção de um único gene*, onde podemos correlacionar que, o impacto mostra-se acumulativo, ou seja, se um gene com alto impacto em sua remoção é removido juntamente com outro gene causador de um alto impacto, ambos aumentam o impacto total da amostra em questão. Já os experimentos que apresentaram boa correlação, apresentaram a remoção de genes sementes que não causaram grande impacto em sua remoção na etapa de *Remoção de um único gene*, sendo exemplo destes os elementos ***GAD1*** e ***HP***, estes que apresentaram um impacto sempre abaixo de ***10%***.

Ao observar as medianas dos experimentos, pode-se enxergar que há uma diminuição conforme aumenta a quantidade de genes sementes removidos em relação ao experimento original. Este aspecto indica uma dependência do método NERI aos genes sementes, fato este já sabido previamente, devido ao mesmo valer-se destes genes para o ranqueamento gênico. O que chama a atenção é o fato da proporção de remoção ser menor que a proporção de impacto causado, ou seja, ao remover ***40%*** dos genes sementes, não impacta o resultado em ***40%***, mas sim em menos, no caso dos ***20*** e ***50*** primeiros genes ranqueados, este impacto fica em torno dos ***35%***.

4.3.3.3 Análise dos 100 e 200 primeiros elementos

Figura 4.12 – Análise dos 100 e 200 primeiros elementos ordenados por X.



Fonte: Produzido pelos autores.

A Figura 4.12 apresenta dois gráficos comparativos em relação aos experimentos que foram gerados através do método *Remoção de mais de um gene*, onde o gráfico de cima representa os ***100*** primeiros genes ranqueados pelos experimentos e o de baixo os primeiros ***200***. Ambos no eixo *Horizontal* apresentam as porcentagens de genes sementes removidos em relação a amostra original e no eixo *Vertical*, a porcentagem de similaridade do resultado dos experimentos com o resultado original, ou seja, a similaridade das listas dos experimentos em relação a lista de ranqueamento gênico original.

Podemos notar que mesmo os *outliers* presentes em ambos os gráficos apresentam uma boa correlação com o experimento original. Isto afirma que todos experimentos executados apresentaram um bom resultado de replicabilidade ao analisar os primeiros ***100*** e ***200*** genes priorizados pelo método NERI. Este aspecto indica uma forte robustez do método, em vista que mesmo os experimentos que apresentaram comportamento diferente do conjunto no qual estão inseridos obtiveram uma boa correlação com o experimento original.

Outro aspecto importante para se observar é o comportamento das correlações dos experimentos, quanto maior a quantidade de genes sementes removidos da amostra original, menor a correlação obtida com o experimento original. Este comportamento esteve presente em todas as análises feitas, tanto no escore X quanto no Δ', comprovando a dependência do método NERI em relação aos genes sementes. Porém mesmo assim, apresentou-se robusto a remoção dos genes sementes, indicando bons resultados de replicabilidade. Fato este que torna a utilização do método analisado mais confiável.

4.3.4 Comparação do escore X com o escore Δ'

Podemos notar que o escore X é mais robusto em relação a remoção de genes sementes se comparado com o escore Δ', apresentando menor variação nos resultados e um menor impacto no resultado final em relação a amostra original. Um fator impactante é a comparação das Figuras 4.10 (***10*** primeiros genes ordenados pelo score X) e 4.7 (***10*** primeiros genes ordenados pelo score Δ'), onde podemos observar a diferença das correlações nos *boxplots* referentes a ***40%*** de remoção dos genes sementes em relação a amostra original. O *boxplot* que representa a ordenação pelo score Δ' apresenta uma amplitude amostral onde o valor mínimo representado é de aproximadamente ***10%*** de similaridade com a amostra original, valor este que apresenta-se muito baixo se comparado com o *boxplot* que representa a ordenação pelo score X onde o valor mínimo apresentado é por um experimento com comportamento *outlier* e corresponde a ***40%*** de similaridade com a amostra original. Outra métrica observada é a amplitude amostral de ambos, ainda observando o pior caso (***40%*** de remoção dos genes sementes em relação a amostra original), onde o score X apresenta ***30%*** de variação contra ***80%*** apresentado pelo score Δ'.

As variações dos resultados e a amplitude interquartílica dos *boxplots* apresentam-se maiores em Δ' em todas as análises comparativas entre aos dois escores. Isto indica novamente uma maior confiabilidade em termos de variação de resultado no score X. Este comportamento se dá pela natureza dos scores, onde o score X é baseado na soma de todas as medidas de centralidade calculadas pelo método NERI e o score Δ' é baseado nas pontuações condicionais da rede, fato este que ao faltar um determinado gene semente que representaria algum papel na rede, o score condicional apresenta variação. Sendo assim, uma forma de ranqueamento gênico menos confiável que o embasamento no score X. Este fato não invalida a utilização da mesma, em vista que esta apresentou bons resultados em linhas gerais, onde o impacto no resultado final em pouquíssimos casos ficou acima de ***50%*** (em casos de extremo estresse, como na remoção de ***40%*** dos genes sementes em relação a amostra original e observando o apenas os ***10*** primeiros genes ranqueados).

4.4 Desempenho computacional

Os fatores envolvidos no processo de execução dos experimentos, foram as configurações da máquina no qual foi executada e a disponibilidade de tempo execução. As configurações da maquina no qual foram executados os experimentos são as descritas abaixo:

- **Processador:** *i7* geração 5.
- **Memória:** 16 GB - 2 pentes 8GB DDR3 1600Ghz.
- **Armazenamento:** 50 GB HD disponíveis.

Pelo fato do programa que implementa o método NERI ainda não utilizar paralelismo (utilização de mais de um núcleo de processamento), foram executadas 4 instâncias separadas ao mesmo tempo, durante todo a etapa de execução do experimento.

4.4.1 Consumo de Processamento

Cada instância ocupou 100% de processamento de um núcleo físico presente no processador, como a máquina utilizada possui 4 núcleos físicos e foram executadas 4 instâncias simultaneamente, o consumo de cpu foi para 100% do total presente.

4.4.2 Consumo de Memória

Cada instância em execução consumiu em média 1,5 GB de memória, totalizando aproximadamente 6 GB alocados por todas as 4 instâncias executantes.

4.4.3 Utilização de disco

Devido ao fato de os experimentos serem executados em modo *Debug*, a escrita em arquivo dos *logs* foi realizada durante boa parte do tempo de execução. Isto significa que, se os modo *Debug* for desabilitado, os tempos podem ser um pouco menores. Os 30 experimentos ocuparam aproximadamente 6,5 MB; para a remoção de um único gene foram realizados 30 experimentos (um para cada semente); e para a remoção de vários genes foram realizados 50 experimentos para cada percentual de remoção (10%, 20%, 30% e 40%), totalizando $50x4 = 200$ experimentos de remoção de vários genes. Desta forma, os 230 experimentos realizados ocuparam aproximadamente 1,2 GB de armazenamento no disco rígido.

5 Conclusão

Neste trabalho apresentamos a análise de robustez do método NERI, que integra dados biológicos de expressão gênica com dados de redes PPI para priorização de genes relacionados a doenças complexas. Para explorar a rede PPI, o método parte de alguns nós sementes da rede – conhecidos por estarem associadas à doença – e utiliza princípios de importância relativa para explorar a vizinhança das sementes. Para explorar a rede PPI o método baseia-se nas hipóteses da *Network Medicine* combinadas com coexpressão e com isto realiza a priorização gênica através de duas pontuações (escores X e Δ'). O método NERI obteve bons resultados de replicabilidade em 3 estudos diferentes, mas faltava analisar o quão robusto o método seria caso uma ou algumas sementes fossem removidas.

Para realizar a análise de robustez, alteramos parcialmente o conjunto de 30 genes sementes mas mantivemos os demais dados (ex: expressão KATO inalterados), e em seguida aplicamos o método e comparamos as listas resultantes com as listas originais. As alterações realizadas no conjunto de gene sementes foram basicamente duas: remoção de um único gene e remoção de vários genes (3, 6, 9 e 12, ou em termos percentuais 10%, 20%, 30%, 40%) do conjunto de sementes original.

Em nossos resultados, observamos que a remoção de um único gene semente apresentou maior impacto no score Δ', ao passo que no escore X apresentou pouco impacto. Também observamos que os impactos causados em ambos os escores não estão relacionados diretamente ao grau do gene semente removido. Este é um comportamento importante de citar, pois, a intuição inicial era de que genes com maior grau impactavam mais o resultado em sua remoção. De fato genes com maior grau impactaram no resultado final, mas genes com grau *1* causaram impacto igual ou muito próximo tornando então inconclusiva a hipótese pré afixada. Desta forma, descobrimos que o grau do gene semente removido não está diretamente relacionado ao impacto causado.

Em relação à remocão de vários genes, considerando o melhor cenario (remocao de *10%* das sementes), as listas resultantes do escore X apresentaram em media *90%* de intersecao com a lista original, e mesmo no pior cenario (remocao de *40%* das sementes), a intersecao foi de *60%* em media, em relação ao score X. Alem disso, observamos que quanto maior a lista dos primeiros genes comparados, menor e a variancia das intersecoes das listas resultantes com a lista original. Também notamos que os genes sementes com alto grau na rede não influenciam diretamente o resultado final em sua remoção.

Conforme apresentado na Seção 4.3.4, o escore Δ' sofreu maior impacto que o escore X. Por exemplo, para a comparação dos 10 primeiros genes com remoção de 10%

das sementes, o escore Δ' atingiu uma interseção mínima de 50% e mediana de 80%, ao passo que o score X atingiu mínima de 60% e mediano de 90%. Estes valores aumentam ao analisar os resultados com 40% de remoção dos genes sementes, onde o score X apresenta interseção mínima de 40% e mediana de 60%, enquanto o score Δ' sofreu mais impacto atingindo a interseção mínima de 10% e mediana de 55%.

5.1 Trabalhos Futuros

Como trabalhos futuros, indicamos a realização das seguintes tarefas:

1. Analisar a robustez – em relação as sementes – de outros métodos, tais como: *Random Walk with Restart* e comparar com os resultados obtidos neste trabalho.
2. Pesquisar como integrar novas fontes de dados ao sistema, tais como: dados de epigenética, dados clínicos, etc.
3. Concluir interface gráfica adicionando documentação com tutorial de utilização.
4. Disponibilização do código fonte na web, e possivelmente uma publicação em um *Application Notes*.
5. Criar um serviço web para utilização do método NERI.

Referências

BARABÁSI, A.-L. *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. 2003. 294 p. Disponível em: <http://books.google.com/books?id=rydKGwfs3UAC>. Citado na página 20.

BARABASI, A.-L.; GULBAHCE, N.; LOSCALZO, J. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, Nature Publishing Group, v. 12, n. 1, p. 56–68, 2011. ISSN 1471-0064 (Electronic). Disponível em: <http://dx.doi.org/10.1038/nrg2918>. Citado 3 vezes nas páginas 14, 22 e 23.

DIJKSTRA, E. W. A Note on Two Probles in Connexion with Graphs. *Numerische Mathematik*, v. 1, n. 1, p. 269–271, 1959. ISSN 0029-599X. Citado na página 19.

GAITERI, C. et al. Beyond modules and hubs: The potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. *Genes, Brain and Behavior*, v. 13, n. 1, p. 13–24, 2014. ISSN 16011848. Citado 2 vezes nas páginas 21 e 22.

HWANG, W. et al. Bridging Centrality : Identifying Bridging Nodes In Scale-free Networks. *Kdd*, p. 20–23, 2006. Disponível em: <www.cse.buffalo.edu/tech-reports/2006-05.pdf>. Citado na página 19.

METZ, J. Instituto de Ciências Matemáticas e de Computação ISSN - 0103-2569 Redes Complexas: conceitos e aplicações. 2007. Citado na página 20.

MUDRY, A.; TJELLSTRÖM, A. Historical background of bone conduction hearing devices and bone conduction hearing aids. *Advances in Oto-Rhino-Laryngology*, v. 71, p. 1–9, 2011. ISSN 00653071. Citado 2 vezes nas páginas 23 e 24.

PAVLOPOULOS, G. a. et al. Using graph theory to analyze biological networks. *BioData mining*, v. 4, n. 1, p. 10, 2011. ISSN 1756-0381. Disponível em: <http://www.biodatamining.org/content/4/1/10>. Citado 3 vezes nas páginas 18, 19 e 21.

SIMÕES, S. N. et al. NERI: network-medicine based integrative approach for disease gene prioritization by relative importance. *BMC Bioinformatics*, BioMed Central Ltd, v. 16, n. Suppl 19, p. S9, 2015. ISSN 1471-2105. Disponível em: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-16-S19-S9>. Citado 4 vezes nas páginas 14, 22, 24 e 27.

STROGATZ, S. H. Exploring complex networks. *Nature*, v. 410, n. 6825, p. 268–276, 2001. ISSN 0028-0836. Citado na página 20.

STUMPP, J. Graph Theory. n. November, 2013. ISSN 1098-6596. Citado na página 17.

WHITE, S.; SMYTH, P. Algorithms for Estimating Relative Importance in Networks. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 266–275, 2003. ISSN 1581137370. Citado na página 22.

Apêndice A – Tabelas

Tabela A.1 – Tabela com os genes sementes do experimento original

Code 1	**GENE**	**Description**	**Code 2**
4524	**MTHFR**	5,10-methylenetetrahydrofolate reductase (NADPH)	1p36.3
5999	**RGS4**	regulator of G-protein signaling 4	1q23.3
5362	**PLXNA2**	plexin A2	1q32.2
27185	**DISC1**	disrupted in schizophrenia 1	1q42.1
7166	**TPH1**	tryptophan hydroxylase 1	11p15.3-p14
1815	**DRD4**	dopamine receptor D4	11p15.5
2900	**GRIK4**	glutamate receptor, ionotropic, kainate 4	11q22.3
1813	**DRD2**	dopamine receptor D2	11q23
9638	**FEZ1**	fasciculation and elongation protein zeta 1 (zygin I)	11q24.2
4978	**OPCML**	opioid binding protein/cell adhesion molecule-like	11q25
2904	**GRIN2B**	glutamate receptor, ionotropic, N-methyl D-aspartate 2B	12p12
1610	**DAO**	D-amino-acid oxidase	12q24
3356	**HTR2A**	5-hydroxytryptamine (serotonin) receptor 2A	13q14-q21
207	**AKT1**	v-akt murine thymoma viral oncogene homolog 1	14q32.32|14q32.32
23322	**RPGRIP1L**	RPGRIP1-like	16q12.2
3240	**HP**	haptoglobin	16q22.1
7157	**TP53**	tumor protein p53	17p13.1
6532	**SLC6A4**	solute carrier family 6, member 4	17q11.1-q12
348	**APOE**	apolipoprotein E	19q13.2
3553	**IL1B**	interleukin 1, beta	2q14
2571	**GAD1**	glutamate decarboxylase 1 (brain, 67kDa)	2q31
91752	**ZNF804A**	zinc finger protein 804A	2q32.1
2066	**ERBB4**	v-erb-a erythroblastic leukemia viral oncogene homolog 4	2q33.3-q34
1312	**COMT**	catechol-O-methyltransferase	22q11.21-q11.23|22q11.21
2561	**GABRB2**	gamma-aminobutyric acid (GABA) A receptor, beta 2	5q34
1812	**DRD1**	dopamine receptor D1	5q35.1
2913	**GRM3**	glutamate receptor, metabotropic 3	7q21.1-q21.2
5649	**RELN**	reelin	7q22
3084	**NRG1**	neuregulin 1	8p12
5533	**PPP3CC**	protein phosphatase 3 (formerly 2B), catalytic subunit	8p21.3
267012	*DAOA	D-amino acid oxidase activator	13q33.2|13q34
64067	*NPAS3	neuronal PAS domain protein 3	14q12-q13
1139	*CHRNA7	cholinergic receptor, nicotinic, alpha 7	15q14
5625	*PRODH	proline dehydrogenase (oxidase) 1	22q11.21
84062	*DTNBP1	dystrobrevin binding protein 1	6p22.3
266553	*OFCC1	orofacial cleft 1 candidate 1	6p24.3
63915	*MUTED	muted homolog (mouse)	6p25.1-p24.3
6570	*SLC18A1	solute carrier family 18 (vesicular monoamine), member 1	8p21.3

Obs: durante a integração, apenas os 30 primeiros genes sementes possuíam correspondentes na rede ***PPI***. Os 8 últimos genes (com o marcador *******) não apresentaram integração com a rede ***PPI***.
Fonte: Tabela gerada pelo autor.

Apêndice B – Scripts

Código B.1 – Script em Python para geração dos experimentos com remoção de mais de um gene semente.

```
1
2  class CrossValidationBased():
3          def __init__(self):
4                  self.LEntrada = []  #Ex: [1,2,4,5,6]
5                  self.lenEntrada = 0 #Tamanho entrada
6                  self.LSaida = []    #Lista de saida
7                  self.k = 0          #iteracoes
8                  self.remocoes = 0   #Qtd Remocoes
9                  self.newLen = 0
10                 self.removidas = []
11
12
13         '''
14          @params list {[]}      (Lista semente)
15          @params rem {int}      (% da lista a ser removida)
16          @params it {int}       (quantidade de listas a serem geradas)
17          '''
18         def setData(self,list,rem,it):
19                 self.LEntrada = list
20                 self.k = it
21                 self.lenEntrada = len(self.LEntrada)
22                 self.remocoes = int(rem*self.lenEntrada)
23                 self.newLen = self.lenEntrada - self.remocoes
24                 self.result = []
25                 self.removidas = []
26
27         def generateResult(self):
28                 self.result = []
29                 self.removidas = []
30                 while(self.k>0):
31                         self.LSaida = []
32                         LNRemovidas = self.LEntrada[:]
33                         novoLen = self.lenEntrada
34                         while(novoLen > self.remocoes):
35                                 r = random.randrange(0,novoLen)
36                                 self.LSaida.append(LNRemovidas[r])
37                                 LNRemovidas.pop(r)
38                                 novoLen -= 1
39
40                         self.result.append(self.LSaida)
41                         self.removidas.append(LNRemovidas)
42                         self.k -=1
43
44     '''
45     Retorna lista de conjuntos resultantes
46     @returns self.result {[[]]}
47     '''
48         def getResult(self):
49                 return self.result
50
51     '''
```

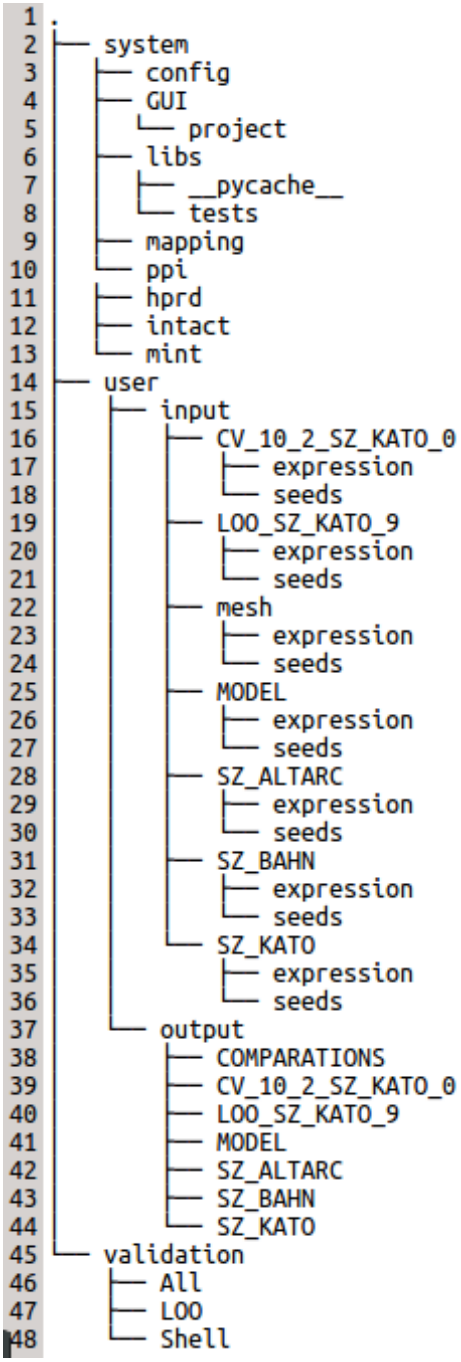
```
52      Retorna lista de conjuntos removidos em cada agrupamento
53      @returns self.removidas {[]}
54      '''
55          def getRemovidas(self):
56                  return self.removidas
```

Código B.2 – Script em *Shell* para execução automatizada dos experimentos.

```
1  #! /bin/bash
2  program="neri.py"
3  call_command="python"
4  params="--nodisplay run"
5  temp_archive_map="mapIn.txt"
6  processed="processed.txt"
7
8  while read line; do
9           $call_command $program $params $line
10         echo $line >> $processed
11 done < $temp_archive_map
```

Apêndice C – Configuração do ambiente

Figura C.1 – Árvore de arquivos do programa *NERI*



Árvore de arquivos
Fonte: Produzido pelos autores.