

INSTITUTO FEDERAL DO ESPÍRITO SANTO  
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

JOÃO CARLOS PANDOLFI SANTANA

**ANÁLISE DE ROBUSTEZ DO MÉTODO DE INTEGRAÇÃO  
DE DADOS NERI**

SERRA

2017

JOÃO CARLOS PANDOLFI SANTANA

**ANÁLISE DE ROBUSTEZ DO MÉTODO DE INTEGRAÇÃO  
DE DADOS NERI**

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 DE INTEGRAÇÃO DE DADOS NERI/

João Carlos Pandolfi Santana. – Serra, 2017-

33 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.

\AM@currentdocname .png

.png

\AM@currentdocname .png

.png

*Aos meus pais*

*Aos xxx*

# Acknowledgements

Agradecimientos

*Texto motivador*  
Winston Churchill



# Abstract

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. Esse problema tem sido abordado 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. O método NERI baseia-se nas hipóteses da *Network Medicine* combinadas com métodos de importância relativa e obteve bons resultados de replicabilidade. A importância relativa é uma forma de inferir a importância dos nós da rede a partir de um conjunto de nós conhecidos como sementes. Entretanto, este 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 destes durante a análise. Utilizamos as técnicas de *leave-one-out* e validação cruzada na qual removemos alguns nós sementes e comparamos cada resultado com o resultado original. Com isso, avaliamos a similaridade das listas de transcritos utilizando o método estatístico da correlação de postos de Spearman. Observamos que a correlação variou de 0,75 até 0,99 para o *leave-one-out* e de... para a validação cruzada com 5 grupos de 6 genes cada. Portanto, o método é considerado robusto (ou não) e recomendamos que...

Palavras chaves: Network Medicine, Validação Cruzada, Leave-one-out, Robustez.

# Abstract

Traduzir o resumo

## Lista de ilustrações

## Lista de tabelas

# Contents

<b>1</b>	<b>Introdução</b>	<b>14</b>
1.1	Objetivos	15
1.1.1	Objetivo Geral	15
1.1.2	Objetivos Específicos	15
1.2	Organização do trabalho	16
<b>2</b>	<b>Referencial Teórico</b>	<b>17</b>
2.1	Redes	17
2.1.1	Grafos	17
2.1.2	Grafos com pesos	17
2.1.3	Passeio	17
2.1.4	Caminho	17
2.1.5	Distância	18
2.1.6	Hub	18
2.1.7	Bridge	18
2.1.8	Menor caminho ou caminho mínimo	18
2.1.9	Redes complexas	18
2.2	Fundamentos biológicos	18
2.2.1	Transcrição	18
2.2.2	Coexpressão de transcritos	18
2.2.3	Doenças multifatoriais	18
2.3	Redes Biológicas	19
2.3.1	Representação de genes em rede	19
2.3.2	Relação de menor caminho	19
2.3.3	Co-expressão como peso	19
2.3.4	Conceito de genes e nós sementes	19
2.4	Métodos de análise de robustez	19
2.4.1	Conceito de robustez	19
2.4.2	Importância da análise	19
2.4.3	Método de validação cruzada	19
2.4.4	Método Leave-one-out	20
2.5	Trabalhos correlatos	20
2.5.1	Teses	20
2.5.1.1	Tese de doutorado Sérgio Nery Simões	20
2.5.2	Redes complexas	21
2.5.2.1	Exploring complex networks	21

2.5.2.2	Algorithms for Estimating Relative Importance in Networks . . . . .	21
2.5.2.3	Linked . . . . .	21
2.5.3	Biologia . . . . .	21
2.5.3.1	DNA methylation: a form of epigenetic control of gene expression . . . .	21
2.5.3.2	DNA methylation and its basic function . . . . .	21
2.5.4	Redes Biológicas . . . . .	21
2.5.4.1	Using graph theory to analyze biological networks . . . . .	21
2.5.4.2	An Integrative Systems Medicine Approach to Mapping Human Metabolic Diseases . . . . .	21
2.5.4.3	Exploring the human diseasome: The human disease network . . . . .	21
2.5.4.4	Network Medicine . . . . .	21
<b>3</b>	<b>Metodologia . . . . .</b>	<b>22</b>
3.0.1	Materiais . . . . .	22
3.0.2	Escolha da variação dos genes sementes . . . . .	22
3.0.3	Escolha dos métodos de validação . . . . .	22
3.0.3.1	Leave one out . . . . .	23
3.0.3.2	Cross Validation . . . . .	23
3.0.4	Preparação dos experimentos . . . . .	23
3.0.4.1	Aplicação do método Leave one out . . . . .	24
3.0.4.2	Aplicação do método Cross Validation . . . . .	24
3.0.5	Execução dos experimentos . . . . .	25
<b>4</b>	<b>Experimentos, Resultados e Discussão . . . . .</b>	<b>27</b>
4.1	Análise dos resultados . . . . .	27
4.1.1	Escolha dos métodos de análise dos resultados . . . . .	27
4.1.1.1	Correlação de postos de spearman . . . . .	27
4.2	Resultados . . . . .	27
4.2.1	Dados computacionais . . . . .	27
4.2.1.1	Consumo de CPU: . . . . .	27
4.2.1.2	Consumo de Memória: . . . . .	27
4.2.1.3	Uso de disco: . . . . .	28
<b>5</b>	<b>Considerações Finais . . . . .</b>	<b>29</b>
5.1	Trabalhos Futuros . . . . .	29
	<b>A – bbbb . . . . .</b>	<b>30</b>
	<b>B – aaaaa . . . . .</b>	<b>31</b>
	<b>C – aaaa . . . . .</b>	<b>32</b>

D – Configuração do ambiente . . . . .	33
--	----

# 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 (??). Como título de informação, alguns exemplos de doenças complexas são 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. Uma forma de utilizar estes dados, é fazer a modelagem 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* (??). 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. Como por exemplo, algoritmos de caminho mínimo, onde visam encontrar o menor caminho entre dois nós, na genética, cada aresta é uma relação entre os genes (*nós*), portanto, quanto o menor caminho entre dois genes (*nós*), mais próxima é a sua relação.

De acordo com os conceitos apresentados, existem diversas abordagens para tratar doenças poligênicas, dentre elas, destaca-se o método NERI 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, identificamos a necessidade de reestruturação do mesmo, de forma que fique mais modularizado, facilitando a manutenção e adição de novas funcionalidades futuras, também temos como objetivo facilitar e incentivar a colaboração de pesquisadores no desenvolvimento futuro da ferramenta, pelo fato de o código ser aberto, ou seja, qualquer um que estiver disposto a contribuir terá acesso ao código fonte. Como a ferramenta foi desenvolvida para ser utilizada por biólogos, identificamos 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 para avaliar 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 cruzada e *leave-one-out* aplicados ao método NERI.
3. Implementar interface gráfica para facilitar o uso da ferramenta, visando torná-la mais intuitiva e amigável aos pesquisadores da área biológica.
4. Desenvolver um módulo de *Template Matching*.
5. Analisar a robustez do método NERI verificando sua dependência em relação aos nós sementes.

## 1.2 Organização do trabalho

Escrever

## 2 Referencial Teórico

Para melhor compreensão do conteúdo apresentado neste trabalho, este capítulo tem como objetivo explicar os fundamentos conceituais apresentados, de forma que os conceitos fundamentais possam ser compreendidos para a evolução do conteúdo.

### 2.1 Redes

#### 2.1.1 Grafos

Grafos são formas de estruturação de dados ligados, onde um dado elemento é denominado nó ou vértice, a sua relação com outro elemento é chamada de aresta. Para exemplificação, tome como nós, duas cidades A e B, as estradas que ligam estas cidades representam as arestas, desta forma pode-se modelar as ligações entre as cidades como um grafo.

#### 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, o que significa que não carrega nenhum significado predefinido, geralmente, exprime o quão relacionado um nó está com outro.

#### 2.1.3 Passeio

É uma sequência específica de nós ligados, partindo de p e chegando em g. O comprimento do passeio é determinado pelo número de arestas.

#### 2.1.4 Caminho

Assim como o passeio, é uma sequência específica de nós ligados, porém este não possui vértices repetidos, ou seja, não passa duas vezes pelo mesmo vértice. A distância do caminho é definida pela soma dos pesos em suas arestas, para 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.

### 2.1.5 Distância

Em grafos com peso, é definida pela soma dos pesos das arestas em um determinado caminho. Em grafos sem peso, é definida como a quantidade de arestas (aresta peso 1) em um determinado caminho.

### 2.1.6 Hub

Hub é um nó que possui muitas arestas, ou seja, um nó que se liga a muitos outros.

### 2.1.7 Bridge

O quão ponte o nó é em relação a dois Hubs, ou seja, se um nó conectar dois Hubs o mesmo é definido como bridge

### 2.1.8 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; Floyd, 1962]

### 2.1.9 Redes complexas

Explicação

## 2.2 Fundamentos biológicos

### 2.2.1 Transcrição

Conceito de transcrição e fundamento da genômica

### 2.2.2 Coexpressão de transcritos

Explicação

### 2.2.3 Doenças multifatoriais

epresentam um fenótipo ou determinam a doença. O que significa, a doença não é composta por um único pedaço de DNA sequenciado, mas sim por vários pedaços de locais separados. (??)

## 2.3 Redes Biológicas

### 2.3.1 Representação de genes em rede

Texto

### 2.3.2 Relação de menor caminho

Texto

### 2.3.3 Co-expressão como peso

Texto

### 2.3.4 Conceito de genes e nós sementes

Texto

## 2.4 Métodos de análise de robustez

### 2.4.1 Conceito de robustez

Texto

### 2.4.2 Importância da análise

Texto

### 2.4.3 Método de validação cruzada

O método de validação cruzada, 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. A definição do método consiste na separação dos dados em subconjuntos, de forma que os elementos sejam diferentes em todos subconjutos, feito o agrupamento, estes são revezados na alimentação do modelo a ser validado, em cada passo faz-se uma análise estatística dos resultados obtidos. <EQ MATEMATICA> <referencia>

### 2.4.4 Método Leave-one-out

Leave-one-out é um modelo de validação cruzada, diferencia-se na formação de agrupamentos, neste modelo a quantidade de subconjuntos é a quantidade de elementos presentes, desta forma cada subconjunto possui somente um elemento. <DESENVOLVER> <EQ MATEMATICA> <REFERENCIA>

## 2.5 Trabalhos correlatos

### 2.5.1 Teses

#### 2.5.1.1 Tese de doutorado Sérgio Nery Simões

(??) Este trabalho é a referencia principal do meu projeto, pelo fato do método no qual analisei a robustez é apresentado e descrito nele.

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. 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 transcritômica 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.

## 2.5.2 Redes complexas

### 2.5.2.1 Exploring complex networks

(??)

### 2.5.2.2 Algorithms for Estimating Relative Importance in Networks

(??)

### 2.5.2.3 Linked

(??)

## 2.5.3 Biologia

### 2.5.3.1 DNA methylation: a form of epigenetic control of gene expression

(??)

### 2.5.3.2 DNA methylation and its basic function

(??)

## 2.5.4 Redes Biológicas

### 2.5.4.1 Using graph theory to analyze biological networks

(??) Este paper contém os conceitos fundamentais de redes biológicas e uso de grafos para sua análise. <Descrever artigo>

### 2.5.4.2 An Integrative Systems Medicine Approach to Mapping Human Metabolic Diseases

(??)

### 2.5.4.3 Exploring the human diseasome: The human disease network

(??)

### 2.5.4.4 Network Medicine

(??)

Neste trabalho é definido o conceito de Network Medicine, este no qual baseia-se o método NERI. <Descrever artigo>

## 3 Metodologia

Para análise de robustez do método NERI foram utilizados conceitos de validação baseados na alteração dos parâmetros de entrada, de forma que sejam analisados os resultados de saída, para entender o impacto causado devido a estas alterações, no qual consistem na remoção ou não inserção de elementos chaves para o input do programa.

O processo de validação foi separado em etapas para melhor entendimento e apresentar a temporalidade e cadenciamento dos processos.

### 3.0.1 Materiais

Este trabalho utilizou a base de dados <BASE DE DADOS AQUI> que consiste em expressões gênicas de pessoas portadoras da doença <ESQUIZOFRENIA>, estes dados podem ser encontrados <AQUI>. Esta base de dados em específico, foi selecionada esta base devido ao fato de ter sido utilizada na tese de doutorado no qual este trabalho se baseia e se referencia, assim os dados obtidos podem ser comparados com os encontrados e apresentados pelo autor, desta forma evitando o enviesamento do resultado por diferença de experimentação.

O programa 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, podendo ser encontradas <AQUI>.

Como entrada do sistema também são definidos os Genes Sementes, onde estes são os genes onde há certeza da sua relação com a doença analisada em questão, a base de dados original pode ser encontrada <AQUI>.

### 3.0.2 Escolha da variação dos genes sementes

Neste trabalho, os genes sementes foram escolhidos para variação como parâmetro de entrada, onde o objetivo é identificar o impacto gerado na rede de correlação gênica e no resultado final exibido pelo programa NERI, assim podendo calcular a dependência e sensibilidade do método em relação a qualidade e quantidade de dados dos nós sementes.

### 3.0.3 Escolha dos métodos de validação

Os métodos de validação adotados foram selecionados pelas suas características de estudo do problema em questão, não podendo deixar margem para enviesamento dos resultados e serem capazes de explorar comportamentos diferentes nos resultados do experimento. Os modelos de validação escolhidos foram: *Leave one Out* e *Cross Validation*.



### 3.0.3.1 Leave one out

<VERIFICAR SE FICA AQUI MESMO> Leave one out foi aplicado no agrupamento de genes sementes, onde cada execução do programa está faltando um gene semente diferente, de forma que sempre haja a mesma quantidade de genes em cada execução e garantindo que todos tenham ficado de fora pelo menos uma vez, assim sendo, o número final de execuções e amostras de entrada sejam a quantidade total de genes sementes menos 1 ( $N - 1 \mid N = \text{total de genes}$ ).

Com este método, é possível descobrir, se a falta de um único gene semente é responsável por alterar significativamente o resultado final do método NERI. Desta forma, podendo analisar se o método em questão é sensível a retiradas de nós sementes. O Leave one Out 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.

Porém há outra análise importante que deve ser feita mas o Leave One Out não é capaz de prover, é se a quantidade de nós removidos influenciam diretamente no resultado. Para observar este aspecto, utilizamos o método Cross Validation, no qual, suas características se moldam mais a esta ótica de estudo.

### 3.0.3.2 Cross Validation

Este método foi adotado pela sua característica principal, organização de agrupamentos de dados de entrada. Com esta característica chave, buscou-se estudar o comportamento da rede quando há a remoção de mais de um gene semente do agrupamento original de entrada.

Com a formação de agrupamentos de genes sementes aleatórios e de tamanhos variados, pôde-se observar o comportamento do método analisado em situações variadas, buscando o ponto de ruptura de proximidade ao resultado original, assim estimando um grau de dependência e sensibilidade a uma quantidade ou arranjo de genes sementes como parâmetro de entrada.

O fato de os agrupamentos possuírem arranjos de genes diferentes, abre-se a possibilidade de estudo sobre a eficácia de um ou mais genes semente juntos sobre o resultado final, ou seja, se um arranjo específico promove melhores resultados que os demais, sendo estes de mesmo tamanho.

## 3.0.4 Preparação dos experimentos

Após determinado os métodos de validação e a base de dados a ser aplicada para análise da ferramenta, o próximo passo é a preparação do experimento a ser desenvolvido, no caso, como a base dados utilizada apresenta 38 genes sementes, a regulação dos

parâmetros de remoção para preparação do experimento deve levar em conta diretamente esta quantidade.

Os genes em questão são: <TABELA DE GENES AQUI>

Estes dados são os brutos de entrada, como o método NERI faz a integração gênica com o GWAS, alguns destes genes, apresentados na tabela acima, não tem representação e ficam de fora, resultando em 30 genes de entrada.

Os genes resultantes são <TABELA DOS 30 GENES AQUI>

#### 3.0.4.1 Aplicação do método Leave one out

Para a validação utilizando o método Leave One Out, a preparação do experimento consistiu em gerar entradas para o NERI de forma que cada amostra de entrada tenha um gene a menos do experimento original, onde exista uma entrada para cada gene removido, ou seja, em uma amostra de 38 genes de entrada, temos 37 combinações de entradas possíveis. Assim sendo, dado um conjunto de  $N$  elementos, a quantidade de entradas possíveis é  $N-1$ .

<IMAGEM REPRESENTATIVA>

A aplicação direta no sistema consistem em cada execução independente do programa, o agrupamento de dados de entrada faltar um gene semente diferente, de forma que sempre haja a mesma quantidade de genes em cada execução e garantindo que todos os genes tenham ficado de fora pelo menos uma vez no total de execuções independentes.

<Para preparação destas entradas, foi desenvolvido um script em Python 3.X para automatização do processo e para evitar falha humana.>

<Script Python GERA<sub>LOO</sub> >

Para ilustração do experimento, segue o exemplo abaixo: Temos a amostra original A sendo: 1,2,3,4,5 Os subconjuntos gerados utilizando o conceito de Leave One Out são: As1: 2,3,4,5 As2: 1,3,4,5 As3: 1,2,4,5 As4: 1,2,3,4

<IMAGEM ToyExampleLOO1>

#### 3.0.4.2 Aplicação do método Cross Validation

Para melhor aproveitamento do método escolhido, o primeiro passo a ser dado, é a definição de tamanho dos agrupamentos de dados de entrada para cada bateria de execuções. Levando em consideração a quantidade de genes de entrada total do experimento, 30 após a integração com o GWAS, 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).

Com as porcentagens de remoção definidos, determinou-se a quantidade de agrupamentos de dados de entrada a serem executados para cada etapa, onde estão descritos na tabela abaixo

<COLOCAR EM TABELA> 10% — 50 execuções 20% — 50 execuções 30% — 50 execuções 40% — 50 execuções

Cada agrupamento deve ser diferente do outro, de forma que o conjunto de genes removidos não se repita dentro de cada etapa, em vista que, caso isso aconteça, a análise final será comprometida por possuir resultados iguais provenientes de entradas de dados iguais.

<Para garantir a diferença entre os dados de entrada, foi desenvolvido um script em Python 3.x para automatização da tarefa evitando falha humana no processo.> <Script GERA<sub>C</sub>VV >

Para ilustração do experimento, segue o exemplo abaixo:

Temos a amostra original A, sendo: 0,1,2,3,4,5,6,7,8,9. Determinado o fator de remoção em 20Dado 20Temos os subconjuntos: As1 = 0,1,2,3,4,5,6,7 As2 = 0,1,2,3,4,5,6,8 As3 = 0,1,2,3,4,5,6,9 As4 = 1,2,3,4,5,6,7,8 As5 = 2,3,4,5,6,7,8,9

Sendo os removidos Rem As1 = 8,9 Rem As2 = 7,9 Rem As3 = 7,8 Rem As4 = 0,9 Rem As5 = 0,1

<IMG<sub>TOY</sub>EXAMPLE<sub>C</sub>V >

Após os agrupamentos de dados de entrada serem preparados, o programa principal é executado individualmente para cada agrupamento. Totalizam-se 200 execuções individuais para a aplicação da validação cruzada.

### 3.0.5 Execução dos experimentos

Nesta etapa, os experimentos encontram-se preparados para execução direta no programa principal que implementa o método NERI, a somatória de execuções totais a serem efetuadas com as técnicas de validação escolhidas consiste em 230 chamadas separadas. 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. <REF AO APENDICE> Um outro quesito no qual influenciou diretamente na execução dos experimentos, foi o fato de o programa original não possuir um esquema de diretórios robusto e chamadas pelo terminal preparada para este tipo de utilização em massa, acarretou na alteração estrutural do programa original para organização dos dados de entrada e dos resultados apresentados, onde também foi desenvolvida uma interface gráfica e uma interface em linha de comando (CLI), para facilitar a utilização por pessoas que não tem familiaridade com este tipo de

utilização de programas.

## 4 Experimentos, Resultados e Discussão

### 4.1 Análise dos resultados

#### 4.1.1 Escolha dos métodos de análise dos resultados

##### 4.1.1.1 Correlação de postos de spearman

- Calcular correlção de cada seed com a correlação das listas (Leave One Out) - [Impacto do Seed na priozação] > Tem um cara que calcula medidas de rede (Betwness, Bridgnes e a porra toda) > Intersecção \* Correlação de Spearman
- Definir a medida chamada impacto (O quanto ficou distante o nó), ou seja, o quanto o gene semene fez diferença (Listas priorizadas) > Checar com todas as medidas de rede (Impacto x Medidas de centralidade)

### 4.2 Resultados

#### 4.2.1 Dados computacionais

Os fatores envolvidos no processo de execução dos experimentos, foram as configurações da máquina no qual foi executada e a disponibilidade de tempo de máquina. As configurações da maquina no qual foram executados os experimentos são as descritas abaixo: Processador: i7 5 geração <Olhar numeração> Memória: 16 Gb <Olhar marca e velocidade> Armazenamento: 50 Tb HD.

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

##### 4.2.1.1 Consumo de CPU:

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

##### 4.2.1.2 Consumo de Memória:

Cada instância em execução consumiu em média 1,5 Gb.

#### 4.2.1.3 Uso de disco:

Devido ao fato de os experimentos serem executados em modo Debug, a escrita em arquivo permaneceu constante durante a maior parte do tempo de execução, aumentando somente nos momentos de escrita de resultados. Os valores consistem em: 78kbs e 1Mbps.

## 5 Considerações Finais

### 5.1 Trabalhos Futuros

1. a
2. b
3. c

ApÃndice A – bbbb



ApÃndice B – aaaaa

ApÃndice C – aaaa

# Apêndice D – Configuração do ambiente

aaa

aaa