UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO
DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA

CARLA FERNANDES DA SILVA

Uma abordagem de integração de dados públicos sobre comorbidade para a predição de associação de doenças complexas

Ribeirão Preto–SP

2019

CARLA FERNANDES DA SILVA

Uma abordagem de integração de dados públicos sobre comorbidade para a predição de associação de doenças complexas

Versão Corrigida

Dissertaçãoo apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) da Universidade de São Paulo (USP), como parte das exigências para a obtenção do título de Mestre em Ciências.

Área de Concentração: Computação Aplicada.

Orientador: Prof. Dr. Evandro Eduardo Seron Ruiz
Coorientador: Prof. Dr. Kuruvilla Joseph Abraham

Ribeirão Preto–SP

2019

Carla Fernandes da Silva
Uma abordagem de integração de dados públicos sobre comorbidade para a predição de associação de doenças complexas. Ribeirão Preto–SP, 2019.
82p. : il.; 30 cm.

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras
de Ribeirão Preto da USP, como parte das exigências para
a obtenção do título de Mestre em Ciências,
Área: Computação Aplicada.

Orientador: Prof. Dr. Evandro Eduardo Seron Ruiz

1. Predição de comorbidades. 2. Integração de dados. 3. Predição de links.

Carla Fernandes da Silva

Uma abordagem de integração de dados públicos sobre comorbidade para a predição de associação de doenças complexas

Modelo canônico de trabalho monográfico acadêmico em conformidade com as normas ABNT.

Trabalho aprovado. Ribeirão Preto–SP, 02 de maio de 2019:

Prof. Dr. Evandro Eduardo Seron Ruiz
Orientador

Prof. Dr. Renato Tinós

Prof. Dr. Ivan Torres Pisa

Prof. Dr. Ivan Rizzo Guilherme

Ribeirão Preto–SP
2019

Agradecimentos

Agradeço primeiramente a Deus, por mais essa vitória.

Agradeço aos meus pais, Benedito Marques da Silva e Creusa Fernandes da Silva, por me conceder a vida. Ao meu namorado Lucas Fernandes pelo carinho e apoio na reta final desse trabalho.

Agradeço a meu orientador Dr. Evandro Eduardo Seron Ruiz que com todo seu conhecimento técnico e acadêmico me apoiou, incentivou e me mostrou o caminho correto.

Aos professores Dr. Joseph Abraham, Dra. Maria Eugenia Gazzaroni e Dr. Rafael Silva Rocha pelos ensinamentos, pelas opiniões sinceras e exigentes.

Agradeço também a todos os amigos do grupo Branca de Neve e os Sete Anões pelos desabafos e conselhos, em especial a Simone Gomes.

À Universidade de São Paulo, seu corpo docente e discente pela experiência única de aprendizado técnico e acadêmico.

Ao IFSULDEMINAS, pelo apoio financeiro e pelo afastamento para capacitação durante essa caminhada. E por me ensinar que obstáculos são impostos para ser vencidos e nos fazer mais fortes.

A todos, o meu muito obrigada!

"Existem muitas hipóteses em ciência que estão erradas.
Isso é perfeitamente aceitável, eles são a abertura para achar as que estão certas."
(Carl Sagan)

Resumo

Comorbidade é a coocorrência de dois ou mais distúrbios em uma pessoa. Identificar quais fatores genéticos ou quais são os mecanismos subjacentes à comorbidade é um grande desafio da ciência. Outra constatação relevante é que muitos pares de doenças que compartilham genes comuns não mostram comorbidade significativa nos registros clínicos. Vários estudos clínicos e epidemiológicos têm demonstrado que a comorbidade é uma situação médica universal porque pacientes com vários transtornos médicos são a regra e não a exceção. Neste trabalho, é proposta uma metodologia de predição de associação doença-doença por meio da integração de dados públicos sobre genes e sobre doenças e suas comorbidades. Analisando as redes formadas pelos genes e pelas doenças, a partir da utilização de cinco métodos de predição de links: Vizinhos Comuns, Adamic-Adar, Índice de Conexão Preferencial, Índice de Alocação de Recursos e *Katz*, a fim de encontrar novas relações de comorbidade. Como resultados foram criadas duas redes: uma rede epidemiológica chamada de $rede\_DATASUS$ com 1.941 nós e 248.508 arestas e uma rede gênica, $rede\_KEGG$, com 288 nós e 1.983 arestas. E a predição em cima da $rede\_KEGG$, e dentre as associações de doenças preditas e analisadas encontramos 6 associações preditas que estão presentes na $rede\_DATASUS$ e relatos na literatura. Acreditamos que as associações entre genes podem elucidar as causas de algumas comorbidades.

Palavras-chave: Predição de comorbidades, Integração de dados, Predição de links.

Abstract

Comorbidity is the co-occurrence of two or more heath disturbances in a person. Identify which genetic factors or what are the biological mechanisms underlying the comorbidity is a big challenge in science. Another relevant finding is that many pairs of diseases that share common genes do not show significant comorbidity clinical records. Several clinical and epidemiological studies have shown that comorbidity is a universal medical situation because patients with various medical disorders are the rule and not the exception In this work, a methodology of prediction of disease-illness is provided through the integration of data on genes and on diseases and their comorbidities. Analyzing how to redesign genes and diseases, using five link prediction methods: Common Neighbours, Adamic-Adar, Preferential Attachment Index, Resource Allocation Index and emph Katz, an end to find new relationships of comorbidity. As a redesigned network: an epidemiological network called $network\_DATASUS$ network with 1,941 nodes and 248,508 edges and a genetic network, $network\_KEGG$, with 288 nodes and 1,983 edges. And the prediction over $network\_KEGG$, and among the predicted and analyzed combinations are 6 predicted classes that are present in $network\_DATASUS$ and reports in the literature. We believe that the associations between genes can elucidate the causes of some comorbidities.

Keywords: Comorbidity prediction, Data integration, Link Prediction.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

AIH	Autorização de Internação Hospitalares
AA	Adamic Adar
AUC	Área sob a Curva (AUC), do inglês Area Under the Curve
CID	Classificação Internacional de Doenças
CN	Vizinhos Comuns, do inglês Common Neighbors
DATASUS	Departamento de Informática do SUS da Secretaria Executiva do Ministério da Saúde
GO	*Gene Ontology*
HDN	*Human Disease Network*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
MeSH	*Medical Subject Headings*
RA	Índice de Alocação de Recursos
PA	Índice de Conexão Preferencial
PPI	Rede de Interação de Proteína–Proteína
SIH	Sistema de Internações Hospitalares do SUS
SNC	Sistema Nervoso Central

Sumário

1

Introdução

Os crescentes avanços genéticos e tecnológicos têm proporcionado uma melhor compreensão das doenças humanas no âmbito geral. Hoje já se sabe que muitas doenças cujas causas eram atribuídas anteriormente apenas a fatores ambientais e comportamentais possuem também fatores genéticos associados. As doenças cujos fatores protagonizantes são os genéticos, aliados aos fatores ambientais e comportamentais são conhecidas pelo nome de *doenças complexas*. Salientamos que é de extrema importância a identificação dos genes relevantes associados a essas doenças, bem como os demais fatores. Este trabalho aborda o desafio de propor uma metodologia para predizer associações entre doenças (comorbidades) que eventualmente tenham como base genes comuns.

Atualmente, podemos notar um grande interesse nas pesquisas relacionadas à predição gênica, ou seja, a identificação funcional de genes em associação a uma doença. Essas pesquisas são facilitadas pela enorme quantidade e variedade de dados genéticos que estão disponíveis em bases públicas de dados. Muitas destas pesquisas consideram informações já existentes de relacionamentos de genes associados a doenças para descobrir novos relacionamentos entre associações de doenças (GOH et al., 2007; DUARTE; BECKER, 2007; LEE et al., 2008; LEE et al., 2011; VIDAL; CUSICK; BARABÁSI, 2011; RITCHIE et al., 2015; MENCHE et al., 2015).

As doenças complexas são poligênicas e multifatoriais, ou seja, são causadas por uma combinação de fatores genéticos, ambientais e comportamentais. Enfatizamos que uma doença raramente é consequência de anormalidades num único gene e, muitas vezes, reflete perturbações da complexa rede intracelular (BARABÁSI; GULBAHCE; LOSCALZO, 2011). Cada vez mais é aceita a visão de que doenças humanas resultam de perturbações de sistemas celulares, especialmente das redes moleculares. Genes associados a doenças similares geralmente localizam-se numa mesma rede molecular. Tais observações têm construído a base para uma grande coleção de abordagens computacionais a fim de encontrar genes desconhecidos, e associados com determinadas doenças. Mesmo com o aumento constante na descoberta de associações gene-doença, ainda há uma grande fração

de doenças sem uma base molecular conhecida (WANG; GULBAHCE; YU, 2011).

Apesar da observação de que a comorbidade[1] pode levar-nos a encontrar genes essenciais, e eventualmente responsáveis, para ambas as doenças, muitos pares de doença que compartilham genes não mostram comorbidade significativa (LEE et al., 2008). Essa ausência de comorbidade pode ocorrer, em parte, porque diferentes mutações no mesmo gene podem ter efeitos diferentes sobre a função do produto do gene e sua expressão no órgão, portanto, diferentes consequências patológicas são dependentes de um contexto biológico (BARABÁSI; GULBAHCE; LOSCALZO, 2011). Além disso, estes mesmos estudos de Barabási, Gulbahce e Loscalzo (2011) também podem nos ajudar a compreender comorbidades, tais como os estudos da equipe de Goh et al. (2007) os quais também encontraram resultados que sugestivos de que através das comorbidades podemos tentar identificar os principais genes associados a duas doenças.

Dentre os trabalhos que procuram predizer a associação de genes específicos com doenças destaca-se o desenvolvido por Néto (2014). Esse trabalho compreendeu a integração dos dados epidemiológicos e genéticos para a realização da predição de genes causadores de doença, através do estudo de comorbidades. Ou seja, usando dados de comorbidades gerados pelo SUS em 12 anos, além de dados genéticos do *Online Mendelian Inheritance in Man* (OMIM), foi possível identificar genes associados a doenças para as quais ainda não havia sido relatada essa associação.

Análises desse tipo conduziram à descoberta de associações entre genes e fenótipos relacionados a doenças, permitindo evidenciar comorbidades e associações entre doenças, fornecendo ferramentas potencialmente importantes para diagnósticos e ações preventivas. Sabemos que alterações nas vias metabólicas podem gerar desordens num ou em mais genes. Conforme o trabalho de Garcia-Albornoz e Nielsen (2015), a integração de dados de genes, doenças e vias metabólicas pode ser relevante para compreender as doenças complexas e até a descoberta de novos genes relacionados a uma determinada doença, podendo assim serem úteis na descoberta de medicamentos e aplicações terapêuticas.

Descobrir as associações entre as doenças genéticas e seus genes causadores é um objetivo fundamental da genética humana. No entanto, apesar da recente revolução genômica, este objetivo ainda continua a ser um grande desafio devido à pleiotropia de genes[2], o limitado número de associações de doença-gene, a heterogeneidade genética de doenças, bem como outras complicações (YANG et al., 2011).

Apesar dos avanços impressionantes no mapeamento do interatoma [3] e da iden-

[1] Comorbidade refere-se a dois ou mais distúrbios que coocorrem em uma pessoa simultaneamente, não por acaso, geralmente isso indica o pior prognóstico e causa danos mais graves (HIDALGO et al., 2009).

[2] Pleiotropia refere-se aos múltiplos efeitos de um gene, ou seja, quando um gene controla diversas características do fenótipo e estas características eventualmente não estão relacionadas.

[3] Interatoma é todo o conjunto de interações moleculares que ocorre em uma determinada célula.

tificação de genes em doenças mendelianas[4], em ambos os casos o conhecimento ainda permanece incompleto.

A disponibilidade de bases de dados de comorbidades públicas e dados biológicos multi-omics[5] oferece a oportunidade de explorar essas fontes de dados para novas associações de doenças. A busca pelo mecanismo biológico subjacente que liga as condições comórbidas pode ajudar a elucidar os mecanismos moleculares subjacentes a duas doenças, e também pode ser útil para a genotipagem e identificação de novos alvos de drogas (LIN et al., 2016).

A partir de trabalhos anteriores, há razões para acreditar que muitas conexões de doenças ainda precisam ser descobertas. Este trabalho foca na integração de dados para a descoberta de novas associações doença-doença. Como consequência dessa falta de conhecimento sobre associações de doenças gênicas, muitas comorbidades que realmente existem não estão presentes nas bases de dados disponíveis publicamente. Como por exemplo, na base de dados KEGG (Kyoto Encyclopedia of Genes e Genomes) (KANEHISA; GOTO, 2000) e no DATASUS (Departamento de Informática do SUS da Secretaria Executiva do Ministério da Saúde) (DATASUS, 2018). Com base nesse problema, o presente trabalho apresenta uma proposta metodológica que integra dados sobre genes associados à doenças com o objetivo predizer associações entre elas por intermédio do relacionamento entre uma rede gênica e uma rede epidemiológica.

1.1 Objetivo Geral

Essa pesquisa tem como objetivo principal propor uma metodologia para a predição de relações doença-doença a partir da integração de dados públicos dispersos em diferentes bases de dados. Neste trabalho, é proposta uma metodologia de predição de associação doença-doença por meio da integração de dados públicos sobre genes e sobre doenças e suas comorbidades. Analisando as redes formadas pelos genes e pelas doenças, a partir da utilização de métodos de predição de links, a fim de encontrar novas relações de comorbidade. Como objetivos específicos, temos:

1. Analisar e filtrar os dados de comorbidade disponíveis nas Autorizações de Internação Hospitalares (AIHs) do Sistema Único de Saúde (SUS) do Brasil, desde início do ano de 1998 até abril de 2017;

2. Criar e analisar redes gênicas e rede de comorbidades a partir da integração com os dados disponíveis no KEGG e DATASUS;

[4] Doenças genéticas monogênicas são assim chamadas porque alguma mutação que ocorre na sequência do DNA de um único gene.

[5] Omics: termo relacionado à biologia de sistemas que está relacionado a identificação, quantificação e caracterização todos os componentes de um sistema biológico

3. Elaborar equivalência de nós da rede gênica para o código de Classificação Internacional de Doenças (CID-10) para possibilitar a padronização com rede epidemiológica;
4. Construir e analisar as redes hipergrafos para inferir possíveis associações;
5. Implementar e analisar de métodos de predição de links;
6. Analisar as medidas AUC e precisão e um comparativo entre os métodos Vizinhos Comuns (CN), Adamic Adar (AA), Índice de Alocação de Recursos (RA), Índice de Conexão Preferencial (PA) e Katz;
7. Possibilitar a predição das associações doença-doença.

1.2 Perspectiva de Originalidade

Como originalidade neste trabalho temos a predição de associações de doenças por intermédio do relacionamento entre uma rede gênica e uma rede epidemiológica utilizando métodos de predição de links. Ambas as redes foram construídas com dados públicos. Essa rede epidemiológica é um grande conjunto de associações de doenças de uma população muito diversificada que serve de base para escolha das melhores associações com maiores probabilidades de ligação real.

1.3 Organização

Este trabalho está organizado da seguinte maneira: o Capítulo 2 apresenta o estado da arte sobre os temas norteadores deste trabalho que são a predição de comorbidades, as doenças complexas e a predição de links. No Capítulo 3 apresentamos os conjuntos de dados utilizados bem como as fundamentações teóricas e metodológicas. Assim, neste capítulo apresentamos os bancos de dados públicos utilizados, as métricas usadas em redes complexas e os métodos de predição de links. O Capítulo 4 contém os resultados alcançados e as análises dos resultados. Por último o Capítulo 5 apresentamos nossas conclusões e considerações finais.

2

ESTADO DA ARTE

Neste capítulo são apresentados os principais trabalhos da literatura para a compreensão desta dissertação. Inicialmente apresentamos o conceito de comorbidade e trabalhos relacionados a predição de comorbidades. Depois elencamos os conceitos de doenças complexas. Por fim, os trabalhos e aplicações da predição de links.

2.1 Comorbidades

Pacientes que possuem comorbidades possuem um risco muito maior de serem hospitalizados ou uma taxa de mortalidade maior que pacientes que apresentam apenas uma doença. Por exemplo, foi encontrado que a taxa de mortalidade de pessoas com doença cardiovascular e diabetes tipo 2 foi 7,5 vezes maior do que aqueles com apenas doença cardiovascular. Portanto, a detecção precoce de comorbidade pode ajudar a projetar um tratamento mais eficiente (HE et al., 2017).

Existe uma relação de comorbidade entre duas doenças sempre que elas se manifestarem simultaneamente em um paciente com mais frequência do que a manifestação de uma delas apenas. A comorbidade representa a coocorrência de duas ou mais doenças que acontece simultaneamente em um paciente (HIDALGO et al., 2009)

O estudo das comorbidades é amplamente investigados pelos cientistas e médicos, devido ao seu impacto na expectativa de vida, qualidade de vida e custo em saúde. A disponibilidade de registros eletrônicos de saúde para mineração de dados oferece a oportunidade de descobrir associações de doenças e padrões de comorbidade da história clínica de pacientes reunidos durante o atendimento médico de rotina. Estudos dessa natureza abrem a necessidade de ferramentas analíticas para a detecção de comorbidades de doenças, incluindo a investigação de suas bases genéticas subjacentes (GIANNOULA et al., 2018).

A comorbidade tem sido estudada extensivamente no passado. Contudo, apesar

de progressos consideráveis no sentido de aprofundar nossa compreensão da comorbidade, algumas questões cruciais permanecem sem resposta (CRAMER et al., 2010).

Durante as últimas duas décadas, vários estudos clínicos e epidemiológicos têm demonstrado que a comorbidade ou multimorbidade [1]é uma situação médica universal, porque pacientes com vários transtornos médicos são a regra e não a exceção (VALDERAS et al., 2009). Alguns pesquisadores estimam que nove entre dez pacientes terão simultaneamente mais de um problema de saúde crônica em 2020 (TABARÉS-SEISDEDOS et al., 2011).

Ainda com relação às comorbidades, outro desafio que tem emergido nas ciências biomédicas é determinar se uma doença deve ser pensada como uma comorbidade ou como uma "complicação" de outra doença. No trabalho de Gijsen et al. (2001), foi realizado uma pesquisa na literatura para identificar e resumir as informações existentes sobre causas e consequências de comorbidade de doenças somáticas crônicas. Foram levantadas algumas causas de comorbidades tais como: demográficas, genéticas, fatores de risco biológicos, estilo de vida, ambiente social, ambiente físico e cuidados de saúde.

No artigo de revisão de Goh e Choi (2012), os autores relatam que muitas doenças complexas são doenças poligênicas como, por exemplo: diabetes, câncer e doenças cardíacas. No entanto, diferentes mutações em um único gene podem causar vários fenótipos de doença. Devido à complexidade intrínseca das associações genótipo–fenótipo, a causa e o efeito das doenças tornam-se mais ambíguas, de modo a dificultar a elucidação dos mecanismos subjacentes. Essa revisão bibliográfica também mostrou que duas doenças ligadas por vias metabólicas apresentam maior comorbidade, ou seja, eles são mais prováveis de ocorrer no mesmo paciente do que o esperado ao acaso, sugerindo uma fisiopatologia compartilhada entre essas doenças. Esse é um exemplo de um cenário mais geral, em que componentes e fatores não puramente genéticos podem desempenhar um papel-chave em alguns processos relacionados a doenças. Além disso também demonstra a importância das vias metabólicas no desenvolvimento de comorbidades.

A partir do estudo das comorbidades podemos tentar identificar os principais genes associados a duas doenças. Espera-se que as ligações induzidas por vias metabólicas partilhadas sejam mais relevantes do que as ligações baseadas em genes partilhados. Em apoio a essa hipótese, Lee et al. (2008) construíram uma rede de doenças metabólicas na qual dois distúrbios estão ligados se as enzimas associadas a eles catalisarem reações adjacentes. Em seu estudo, os autores ressaltam a descoberta de que doenças metabólicas conectadas através de caminhos compartilhados tendem a mostrar comorbidade significativa, sugerindo que a informação codificada na estrutura da rede metabólica torna-se perceptível em nível de população como padrões de comorbidade.

[1] Multimorbidade é a presença de duas ou mais doenças no mesmo indivíduo

A busca de genes candidatos relacionados a comorbidades de asma e hipertensão pode ajudar a elucidar os mecanismos moleculares subjacentes à comorbidade dessas duas doenças, e também pode ser útil para identificação de novos alvos de drogas. No trabalho dos autores Saik et al. (2018) uma metodologia foi desenvolvida para a análise das comorbidades entre asma e hipertensão. Primeiro os autores construíram duas redes gênicas: uma rede de asma e a rede de hipertensão, a primeira possui 755 genes/proteínas e 62.603 interações, enquanto a segunda possui 713 genes/proteínas e 45.479 interações. Essas duas redes compartilham 205 genes/proteínas e 9.638 interações foram compartilhados entre asma e hipertensão. Depois foi definido nove critérios para classificar os genes pela sua importância, incluindo métodos de priorização de genes, bem como critérios originais que levam em conta as características de uma rede de genes associativos e a presença de polimorfismos de genes conhecidos. O método proposto conseguiu predizer 10 genes que desempenham um papel fundamental no desenvolvimento da condição comórbida das duas doenças e os genes IL10, TLR4 e CAT tiveram a maior prioridade no desenvolvimento da comorbidade de ambas.

Além desses, o trabalho de Kann (2009) ressalta na sua revisão bibliográfica que o impacto de processos causadores de doenças muitas vezes não é limitado aos produtos de um gene mutante, mas ocorre graças a interações entre os componentes moleculares, os quais também podem afetar outras funções celulares, resultando em efeitos potenciais de comorbidade. Em contrapartida, muitos pares de doenças que compartilham genes não apresentam comorbidade significativa. Esta falta de comorbidade pode ocorrer, em parte, devido a diferentes mutações ou por causa da pleiotropia genética.

Em geral, a visão de multimorbidade pode ajudar a refinar nosologias atuais das doenças e eventualmente levar ao desenvolvimento de uma nova abordagem interdisciplinar, que por sua vez pode fornecer novas estratégias terapêuticas (VALDERAS et al., 2009).

Complementarmente os pesquisadores Ibáñez et al. (2014) definiram comorbidade inversa, na sua pesquisa, como a diminuição de risco de determinados tipos de câncer (pulmão, próstata, colorretal) em pacientes com Síndrome de Down com relação a algumas doenças do Sistema Nervoso Central (SNC), por exemplo, Alzheimer, Parkinson e Esquizofrenia. Os autores testaram a hipótese de que a comorbidade inversa, uma comorbidade que inibe uma determinada doença, é orientada por processos moleculares comuns aos distúrbios do sistema SNCs e câncer, e que são desregulados em direções opostas. A conclusão dessa pesquisa sugere que a comorbidade inversa é influenciada por fatores ambientais, tratamentos com drogas e outros aspectos relacionados com o diagnóstico da doença.

Outro trabalho que também usa a abordagem de redes gênicas para estudar comorbidades foi realizado por Zanzoni, Chapple e Brun (2015), que trabalharam numa rede de interação de proteína–proteína (PPI) para identificação em larga escala de proteínas

moonlighting, ou seja, uma subclasse especial de proteínas multifuncionais que executam múltiplas funções autônomas. Nesse trabalho eles estabeleceram que 3% do interatoma humano atual é composto de proteínas *moonlighting* previstas. Os resultados obtidos sugerem que comorbidades entre doenças fenotipicamente diferentes podem ocorrer devido a uma proteína compartilhada envolvida em processos biológicos não relacionados.

No trabalho de Park et al. (2011), foi investigado o relacionamento entre as proteínas associadas à doença e suas localizações subcelulares. Com base na suposição de que os pares de proteínas associadas à doenças fenotipicamente semelhantes são mais prováveis de estarem na mesma localização subcelular, os autores construíram, pela primeira vez, uma matriz de proteínas associadas a doenças e sua localização subcelular, que descreve a inter-relação entre os dois. A partir dessa matriz, descobriram que as proteínas associadas a uma mesma doença são provavelmente enriquecidas em determinadas localizações subcelulares na célula. Também observaram que doenças fenotipicamente semelhantes agrupadas nas mesmas classes de doenças estão associadas a localizações subcelulares particulares. Além disso, foi encontrada uma correlação positiva entre a similaridade da localização subcelular dos pares de doenças e as medidas de comorbidade, o que explica as conexões moleculares entre pares de doenças comórbidas conectadas via localização subcelular. Outro resultado é a descoberta de conexões moleculares entre 7.584 pares de doenças até então desconhecidas.

Na pesquisa de Roque et al. (2011) foram utilizados registros eletrônicos de pacientes que permanecem bastante inexplorados para este fim de correlacionar gene e doença, os quais são uma fonte de dados potencialmente rica para descobrir correlações entre as doenças. Os autores descreveram uma abordagem geral para coletar descrições fenotípicas dos pacientes de registros médicos de forma sistemática. Extraindo informações sobre fenótipos em texto livre dos registros clínicos dos pacientes, os autores demonstraram que podemos estender as informações contidas nos dados de registro para identificar novos genes que podem estar relacionados à associação de doenças.

O estado de saúde dos pacientes geralmente não é caracterizado por uma única doença, mas por múltiplas condições médicas. Essas comorbidades podem estar ligadas fortemente à idade e ao sexo (CHMIEL; KLIMEK; THURNER, 2014). No estudo de Liu et al. (2016), analisou-se um conjunto de dados abrangente e geograficamente distribuído da população chinesa para encontrar as ocorrências e padrões de comorbidades da hipertensão. Com esses dados foi construída uma rede de comorbidades com base na frequência das relações de coocorrência entre essas comorbidades. Por meio dessa rede os autores investigaram as comorbidades comuns da hipertensão com relação ao sexo e idade do paciente. Nesse estudo identificaram as 20 principais comorbidades que apresentavam fortes correlações de coocorrência com hipertensão, dentre elas as quatro principais foram doença cardíaca coronária, diabetes, hiperlipemia e arteriosclerose. Também descobriram

que as pacientes do sexo feminino com hipertensão eram mais propensas a sofrer de osteoporose, enquanto os pacientes do sexo masculino eram mais propensos a desenvolver insuficiência renal.

Esses estudos sobre comorbidade formam uma base teórica sobre os eventuais mecanismos que ligam duas ou mais doenças. A seguir partimos do estudo das relações entre proteínas que são precursores para desvendar os eventuais mecanismos ou agentes que ligam duas doenças. Iniciaremos com os estudos sobre os produtos gênicos mas avançaremos para estudos mais recentes que consideram as vias metabólicas como esses canais de comunicação, de ligação entre doenças.

2.2 Doenças Complexas

Descobrir novos genes causadores de doenças humanas é uma tarefa desafiadora na pesquisa biomédica. Nos últimos anos, várias abordagens computacionais foram propostas para priorizar genes candidatos a doenças. A maioria desses métodos é baseada principalmente em redes de PPI. No entanto, essas redes PPI contêm falsos positivos e cobrem menos da metade dos genes humanos conhecidos. Sua confiabilidade e cobertura são muito baixas. Portanto, é altamente necessário integrar múltiplos dados genômicos para construir uma rede confiável de similaridade de genes e inferir genes de doenças em toda a escala genômica. Os autores Tian et al. (2017) propuseram um novo método, denominado RWRB, para inferir genes causais de doenças. Construíram cinco redes individuais de similaridade de genes (proteínas) com base em múltiplos dados genéticos de humanos, conduziram um estudo de caso abrangente para a doença de Alzheimer no qual conseguiram prever alguns novos genes da doença que são apoiados pela literatura.

A identificação de genes responsáveis por doenças pode fornecer conhecimentos para o desenvolvimento de novos diagnósticos e tratamentos clínicos. Já se sabe que uma doença não é necessariamente consequência de uma mutação em um único gene, mas pode refletir a ação de vários processos biológicos, frutos da interação de uma complexa rede gênica. Essa hipótese foi levantada nos trabalhos de revisão de Loscalzo, Kohane e Barabási (2007) e de Ideker e Krogan (2012) que demonstraram o potencial das abordagens baseadas em redes complexas para desvendar as eventuais relações gênicas em doenças humanas e para interrogar de forma mais abrangente os sistemas biológicos em uma variedade de organismos.

Diversos trabalhos na literatura como Goh et al. (2007), Wu et al. (2008), Li e Agarwal (2009), Barabási (2007), Barrenas et al. (2009), Zhou et al. (2014) utilizaram as abordagens de redes complexas para um melhor entendimento dos mecanismos que servem como base para as doenças complexas. As redes complexas são estruturas relacionais que representam sistemas complexos, tais como redes sociais, redes de informação e

redes de tecnologia, redes biológicas (NEWMAN, 2003). Em um de seus trabalhos pioneiros, Goh et al. (2007) criou uma rede de doenças humanas (*Human Disease Network* - HDN) conectando todas as doenças hereditárias que compartilham um gene causador de doença, de acordo com o banco de dados OMIM. Na rede HDN, dois genes da doença estão conectados se eles estão associados com o mesmo distúrbio. Dado que as ligações significam associação fenotípica relacionada entre dois genes, elas representam uma medida de sua relação fenotípica, que poderia ser usada em estudos futuros, em conjunto com reações metabólicas. Foi possível concluir que apenas as mutações que afetam genes funcionalmente e topologicamente periféricos podem ser considerados responsáveis por doenças hereditárias.

Nos últimos anos, houve um aumento no interesse do uso de ferramentas baseadas em redes complexas para ganhar uma melhor compreensão nas relações gênicas, nas relações entre doenças, redes de interação proteína–proteína, redes metabólicas (LEE et al., 2008). Como exemplo temos o trabalho de Barabási, Gulbahce e Loscalzo (2011) em que foi verificada a hipótese de que uma relação genética compartilhada tem consequências epidemiológicas na ocorrência de doenças na população. Os autores concluíram que um paciente tem o dobro de probabilidade de desenvolver uma nova doença (comorbidade) se a doença atual compartilha um gene com a doença que se manifestou primeiramente. No entanto, sabemos que muitos pares de doenças que compartilham genes não apresentam comorbidade significativa. Essa falta de relacionamento na comorbidade pode ocorrer, em parte, porque diferentes mutações no mesmo gene podem ter efeitos diferentes sobre a função do produto gênico e sobre sua expressão orgânica, portanto, apresentam diferentes consequências patológicas dependentes do contexto.

Conceitos de redes podem ser usados para representar vários sistemas biológicos, como por exemplo, as redes de interação entre proteínas, rede de via metabólica, rede de associação de doenças e genes. A análise de redes biológicas é uma abordagem para refinar grandes conjuntos de dados clínicos transformando em conhecimentos para diagnóstico, prognóstico e tratamento da doença. A ideia de analisar redes biológicas tem como fundamento a concepção de que as doenças são consequência de perturbações nestas redes biológicas. As redes servem como um paradigma de integração de dados e análise, proporcionando uma compreensão do mecanismo subjacentes a doenças (FURLONG, 2013).

Outros dois trabalhos que realizam uma revisão sobre redes biológicas são os trabalhos de Vidal, Cusick e Barabási (2011) e Kann (2009) nos quais foram revisados conceitos básicos de rede biológica e discutidos os diferentes tipos de redes interatoma. Vidal, Cusick e Barabási (2011) comentam a importância dos mapas da rede metabólica, referente a sua abrangência em relação a outras redes biológicas. Os autores discorrem ainda sobre como as células podem consequentemente ser vistas como teias complexas

de interações macromoleculares, e também sobre o fato de que as interações das redes celulares podem ser a base de relações entre genótipos e fenótipos.

Apesar de muitas descobertas inovadoras durante o último século, estamos longe de ter uma compreensão completa da complexa rede de processos moleculares envolvidos em doenças, e ainda estamos buscando as curas para doenças mais complexas. Com o objetivo de compreender as doenças complexas, os pesquisadores passaram a integrar diversos tipos de bases de dados biológicos, e várias metodologias têm sido desenvolvidas para analisar redes de doenças. Li e Agarwal (2009) utilizaram-se da técnica de mineração de dados da literatura científica para extrair os genes que são associados significativamente com termos de doenças humanas, usando termos do *Medical Subject Headings* (MeSH) associados com resumos do *Medline*. Desse modo eles conseguem relacionar doenças a vias biológicas mediante a verificação de genes sobrepostos. Essa equipe ainda gerou uma rede de doenças ligando-as quando elas compartilham vias biológicas em comuns. A principal conclusão desse trabalho é que para muitas doenças vários genes foram identificados coletivamente como responsáveis por fenótipos clínicos. Uma outra conclusão é que os genes não operam sozinhos, eles coordenam as suas atividades sob a forma de complexos gênicos ou vias biológicas. Além disso verificou-se que mais de 50% dos genes associados a uma doença estão também mapeados para vias biológicas. Este trabalho mostra que é possível utilizar caminhos, vias biológicas, para representar a biologia subjacente às doenças.

No trabalho de Zhou et al. (2014) foi desenvolvida uma rede de doenças humanas baseadas em seus sintomas. O objetivo era investigar a conexão entre as manifestações clínicas das doenças e suas interações moleculares subjacentes. Foi utilizada uma grande base de dados de literatura biomédica, a Medline, via MeSH. Os dados utilizados foram os que continham um ou mais termos relacionando doença–sintoma. Essa pesquisa apresentou como um dos resultados a forte associação entre semelhança de sintomas de doenças e genes compartilhados.

O objetivo do trabalho de Menche et al. (2015) foi apresentar uma estrutura baseada em rede de doenças para identificar a localização dos módulos de doença dentro do interatoma e usar a sobreposição entre módulos para prever relacionamentos à doença. Os módulos de doenças são um ou vários subgrafos formados por proteínas da doença (os produtos dos genes da doença) que não são dispersas aleatoriamente, mas tendem a interagir umas com as outras. Uma das conclusões desse trabalho é que quanto maior o grau de aglomeração das proteínas da doença dentro do interatoma, maior a similaridade biológica e funcional dos genes correspondentes. Além disso os genes associados com a mesma doença tendem a aglomerar-se na mesma vizinhança do interatoma.

Barrenas et al. (2009) avaliaram as propriedades topológicas e funcionais de genes de doenças complexas no interatoma humano comparando-as com genes essenciais e de doenças monogênicas. Essa mesma equipe verificou que genes de doenças complexas são

menos centrais do que os genes essenciais e os relacionados às doenças monogênicas. Eles também observaram que genes responsáveis por doenças complexas associados a mais de uma doença tendem com mais frequência a compartilhar uma interação de proteína–proteína e, pelo menos, um termo do processo biológico da *Gene Ontology* (GO) em comparação com os genes associados com doenças diferentes.

Decifrar a base genética das doenças humanas e identificar genes responsáveis por doenças específicas é um objetivo importante da pesquisa biomédica. Com base na suposição de que doenças fenotipicamente semelhantes são causadas por genes funcionalmente relacionados, o trabalho de Wu et al. (2008) propôs um modelo computacional que integra interações de proteína–proteína, semelhanças fenotípicas de doenças e associações gene-fenótipo para explicar a similaridade fenotípica pela proximidade dos genes (proximidade topológica na rede de interação molecular). Nesse trabalho, foi desenvolvido uma ferramenta chamada *CIPHER (Correlating protein Interaction network and PHEnotype network to pRedict disease genes)* para priorizar genes candidatos e explorar o comportamento cooperativo de genes em doenças humanas. O resultado dessa pesquisa demonstrou que a correlação entre semelhanças fenotípicas e a proximidade dos genes é um forte e robusto indicador para a predição de genes em doenças.

Apesar de múltiplas doenças coocorrerem, seus mecanismos moleculares comuns subjacentes permanecem desconhecidos. A identificação de comorbidades, considerando as interações entre os componentes moleculares, é fundamental para entender os mecanismos subjacentes à doença. No trabalho de Ko et al. (2016) foi desenvolvida uma nova abordagem, utilizando genes comuns causadores de doenças e vias moleculares subjacentes para identificar doenças comórbidas. Esses pesquisadores combinaram relações funcionais entre genes codificadores de proteínas e módulos biológicos associados a proteínas para investigar a etiologia da comorbidade inexplicada e para elucidar as origens moleculares ou mecanismos subjacentes essas doenças comórbidas. Essa abordagem permitiu a análise de patologias comuns compartilhadas por comorbidades por meio de redes de interação molecular. Descobriram que as doenças relacionadas a neoplasias mostraram altos padrões de comorbidade entre si e com outras doenças, indicando complicações graves. Esse estudo demonstrou que as informações de vias moleculares podem ser usadas para descobrir a comorbidade e fornecer uma nova visão sobre a patologia da doença.

Outros resultados que ajudam a compreender a metodologia de rede complexa aplicada a doenças humanas podem ser apreciados no trabalho de Park et al. (2009). Nesse projeto foram combinadas interações celulares, associações entre gene e doença, além das informações de padrões de doença no âmbito populacional em relação aos da base de dados do *Medicare*. Esse modelo mostrou correlações estatisticamente significativas entre a estrutura subjacente às redes celulares e os padrões de comorbidade na população humana. Os resultados indicaram que a combinação de dados em nível populacional e informações

da rede celular podem ajudar a construir novas hipóteses sobre os mecanismos de doenças. Eles mostraram ainda que existem correlações estatisticamente significativas entre padrões de comorbidade e interações celulares, e que pares de doença com correlações mais elevadas tendem a ser ligados mais fortemente na rede celular.

Uma das características importantes quanto às doenças complexas consiste no fato de elas apresentarem comorbidades. Nesse sentido, descobrir quais genes estão associados a uma determinada doença é importante para conhecermos as relações entre causa e efeito nela. Por exemplo, o trabalho de Garcia-Albornoz e Nielsen (2015) incitou a criação de uma rede de genes, vias metabólicas e doenças. Nessa pesquisa, para realizar a análise de rede e associações entre genes e vias metabólicas, foi realizada a integração das bases: HGNC para padronização dos símbolos de genes, códigos de CID-10, vias metabólicas do banco KEGG e as associações entre gene e, finalmente, menções de doenças foram extraídas da base de dados *Online Mendelian Inheritance in Man* (OMIM). Foram mapeados 880 códigos CID-10 exclusivos para os 4.315 fenótipos de doenças extraídos do OMIM e 3.083 genes com mutação causadora de fenótipos extraídos, e padronizado utilizando HGNC. A partir desse trabalho, um total de 705 códigos de doenças CID-10 foram ligados a 1.587 genes com mutações causadoras de fenótipos e 801 vias KEGG, criando uma rede tripartite composta por 15.455 interações de vias gene e CID-10. Com a rede de doenças obtida, os autores realizaram o estudo de inclusão entre as doenças e uma análise do método hipergeométrico enriquecido entre genes e vias metabólicas compartilhadas por diferentes doenças. Uma das conclusões da pesquisa é o forte potencial das associações gene-doença-via metabólica para a predição de novas interações gene-doença.

Os pacientes com a doença pulmonar obstrutiva crônica (DPOC) frequentemente sofrem de distúrbios concomitantes que pioraram significativamente o estado de saúde e o prognóstico vital. Os mecanismos patogênicos subjacentes à DPOC e suas multimorbidades não são completamente compreendidas, daí a exploração do potencial molecular e das ligações biológicas entre a DPOC e suas doenças associadas é de grande interesse. No estudo de Grosdidier et al. (2014), os autores testaram a hipótese do componente compartilhado na DPOC. Sobe essa hipótese, as multimorbidades da DPOC mais frequentemente observadas nas clínicas estariam relacionadas entre si e com a DPOC no nível molecular por genes, proteínas ou vias biológicas comuns. Para esse fim, foi utilizado uma abordagem de rede que incluiu: (1) mineração de dados da doença (ou rede de doenças), o interatoma (definido por uma rede de interação proteína–proteína, PPI) e a exposição à fumaça do tabaco (representando a exposição a substâncias químicas da fumaça do tabaco, o principal fator de risco para a DPOC); e, (2) uma análise funcional para identificar as vias biológicas possivelmente envolvidas na multimorbidade da DPOC. Os resultados desse estudo indicaram que todas as multimorbidades da DPOC estudadas estão relacionadas em nível biológico, compartilhando genes, proteínas e vias biológicas. Os autores identificaram algumas vias biológicas associadas a DPOC, como inflamação,

disfunção endotelial ou apoptose, servindo como prova de conceito da metodologia. Além disso, também observaram semelhanças entre multimorbidades da DPOC no âmbito das vias biológicas, sugerindo mecanismos biológicos comuns para diferentes multimorbidades. Finalmente os produtos químicos contidos na fumaça do tabaco atingem uma média de 69% das proteínas que participam das multimorbidades da DPOC.

As doenças são geralmente definidas por um conjunto de fenótipos e associadas a vários processos patobiológicos e suas interações mútuas. Recentemente, tem havido impressionante progresso na compreensão de vários tipos de relações entre fenótipos de doenças e os processos moleculares subjacentes comuns. Chmiel, Klimek e Thurner (2014) propuseram uma rede de comorbidade de doenças humanas baseada em dados médicos de toda a população da Áustria. Os autores demonstraram que a rede passa por mudanças estruturais dramáticas ao longo da vida dos pacientes. As redes de doenças para crianças consistem em um único cluster fortemente interconectado, e durante a adolescência e a idade adulta surgem outros surtos de doenças relacionadas a classes específicas, como distúrbios circulatórios, mentais ou geniturinários. Enquanto para pessoas com mais de 65 anos, esses clusters começam a se fundir e os hubs altamente conectados dominam a rede. Esses hubs estão relacionados à hipertensão, doenças cardíacas crônicas e doenças pulmonares obstrutivas crônicas. O resultado desse trabalho foi capaz de mostrar que os pacientes desenvolvem predominantemente doenças que estão estreitamente próximas com distúrbios que eles já sofrem.

Na seção subsequente destacamos como os métodos de predição de ligações (links) tem o potencial de serem usados para ampliar nosso conhecimento sobre as associações de doenças.

2.3 Predição de Links

As doenças estão intimamente relacionadas aos genes, indicando que anormalidades genéticas podem estar diretamente evolvidas com doenças. O reconhecimento de genes relacionados com doenças tem sido um objetivo na biologia, o que pode contribuir para a melhoria da assistência médica e para a compreensão das funções dos genes, caminhos e interações. Vários métodos e técnicas de alta produtividade têm sido usados para reconhecer genes relacionados a doenças. No entanto, geralmente esses métodos fornecem centenas de candidatos, e identificar genes relacionados a doenças entre os genes candidatos usando metodologia experimentais estritamente em bancadas bilógicas é muito demorado e caro. Para lidar com as questões acima, os métodos baseados em redes complexas, associações entre genes ou doenças foram propostos. Muitos métodos baseados em redes foram criados para reconhecer genes de doenças. No artigo de Zou et al. (2014), é apresentado um resumo de cinco algoritmos típicos baseados em redes, como *CIPHER*,

PRINCE, *RWRH*, *Katz* e *CATAPULT* (*Combining data Across species using Positive-Unlabeled Learning Techniques*). O resultado apresentado é que o *Katz* e o *CATAPULT* oferecem melhores resultados que os demais métodos quando aplicados a uma rede gênica.

Um grande desafio na era pós-genômica é propor uma representação computadorizada completa da célula e do organismo, o que permitirá a predição computacional de alto nível dos processos celulares e do comportamento do organismo a partir da informação genômica, incluindo assim a predição de comorbidades. Para esse fim, a comunidade científica tem desenvolvido uma abordagem baseada no conhecimento para a predição de redes gênicas relacionadas a doenças, que é prever, dado um conjunto completo de genes no genoma, as redes de interação de proteínas que são responsáveis por vários processos celulares. A exemplo, o KEGG é uma a base de referência de conhecimento biológico que integra informações funcionais genômicas, químicas e sistêmicas (KANEHISA et al., 2014).

Estudos de redes complexas também foram aplicados em RNAs. Estudos recentes descobriram que os microRNAs eram altamente relevantes a várias doenças, incluindo vários tipos de câncer, diabetes, doença de Alzheimer, Síndrome da imunodeficiência adquirida (AIDS) e hipertrofia cardíaca. No entanto, poucas obras investigaram os padrões ou caminhos pelos quais existem as associações microRNA–doença. Vários métodos de reconstrução de redes de microRNA–doença têm sido usados para descrever a associação a partir de uma perspectiva de biologia de sistemas. No artigo de Zou et al. (2016), foi feita uma revisão dos principais métodos de cálculo de similaridade para a inferência de associações microRNA–doença dentre eles KNN (*k–nearest neighbors*), algoritmo caminhada aleatória e aprendizado semi-supervisionado.

Experimentos biológicos realizados sobre doenças humanas mostram que os microRNAs que causam doenças similares frequentemente interagem uns com os outros diretamente ou indiretamente. Assim, redes de associações doença-doença são semelhante a redes sociais, e os autores de (ZOU et al., 2015) usaram métodos de redes sociais para prever associações. Em destaque, foram a descoberta de associações desconhecidas a partir de associações conhecidas, incluindo associações microRNA-microRNA, uma pequena quantidade de associações de microRNA-doença e associações doença-doença. Os autores demonstraram que os métodos *Katz* e *CATAPULT* são capazes de propor muitas associações potenciais, o que é de grande valor para futuros estudos.

As predições de ligações físicas e funcionais entre componentes celulares são frequentemente baseadas em correlações entre medidas experimentais, como a expressão gênica. No entanto, as correlações são afetadas por caminhos diretos e indiretos, confundindo nossa capacidade de identificar interações verdadeiras entre pares. Barzel e Barabási (2013) exploram as propriedades fundamentais de correlações dinâmicas em redes para desenvolver um método para silenciar efeitos indiretos. O método desenvolvido recebe como entrada as correlações observadas entre os pares de nós e transforma a matriz de

correlação em uma matriz silenciada altamente discriminativa, que aumenta os termos associados a ligações diretas.

Apesar do considerável progresso na descoberta de genes de doenças, estamos longe de descobrir os mecanismos celulares subjacentes das doenças, uma vez que possuem características complexas. Mesmo muitas doenças Mendelianas não podem ser explicadas por relacionamentos simples genótipo–fenótipo. Mais recentemente, uma visão cada vez mais aceita é que as doenças humanas resultam de perturbações de sistemas celulares, especialmente de redes moleculares. Os genes associados a doenças semelhantes ou similares residem comumente no mesmo sítio de redes moleculares. Tais observações construíram a base para uma grande coleção de abordagens computacionais para encontrar genes anteriormente desconhecidos associados a certas doenças. A maioria dos métodos baseia-se em redes de interações de proteínas, com a integração de outros dados genômicos em larga escala ou informações de fenótipo da doença para inferir quão provável é que um gene esteja associado a uma doença. (WANG; GULBAHCE; YU, 2011) analisaram métodos de última geração, baseados em rede, usados para priorizar os genes da doença. Os autores destacam três categorias principais: métodos de distâncias locais, métodos de distâncias globais e outros métodos de agrupamento para medir a proximidade de pares de proteínas em uma rede para priorizar genes candidatos dentre esses métodos estão caminho mínimo, algoritmo propagação, caminhada aleatória e outros.

Identificar corretamente associações de genes com doenças tem sido um objetivo na biologia. Os autores Singh-Blom et al. (2013) apresentam dois métodos para prever associações gene-doença com base em associações de genes funcionais e associações gene-fenótipo em organismos modelo. O primeiro método, a índice de *Katz*, é motivado pelo seu sucesso na predição de links de redes sociais, e está intimamente relacionado com alguns dos recentes métodos propostos para a inferência de associação gene-doença. O segundo método, chamado CATAPULT, é um método supervisionado de aprendizado de máquina que usa uma máquina de vetor de suporte polarizado, onde as características são derivadas de caminhadas em uma rede de traços genéticos heterogêneos. Para analisar o desempenho desses dois métodos foram usados dois conjuntos de dados distintos, ou seja, fenótipos OMIM e interações droga-alvo. Embora ambos os métodos tenham um bom desempenho, a índice de *Katz* mostrou-se mais apropriado para identificar associações entre traços e genes pouco estudados, enquanto CATAPULT é mais adequada para identificar genes corretamente e reunir as associações em geral.

A predição de links visa a descobrir ligações ausentes ou prever o surgimento de relacionamentos futuros a partir da estrutura de rede atual. Diversos algoritmos foram desenvolvidos para a predição de links em redes não ponderadas, mas apenas alguns foram estendidos a redes ponderadas. O trabalho de Zhao et al. (2015) tem como objetivo prever links faltantes e seus pesos usando apenas informações locais. Os autores se inspiraram

no algoritmo de roteamento *Reliable Routes* e propuseram um método para generalizar índices de similaridade não ponderados. Assumindo que o índice de similaridade entre dois nós desconectados reflete sua força de interação e, usando a correlação linear entre os scores de similaridade e os pesos de enlace em redes empíricas, definiram pesos de elos perdidos proporcionais aos scores de similaridade. Demonstraram que usando esses índices é possível prever a existência de links e seus pesos e obter maior precisão do com que outros algoritmos testados.

Algoritmos de predição de links visam a estimar a tendência da existência de um link entre dois nós, com base em vínculos observados, atributos de nós ou correlações dinâmicas. A aplicabilidade da predição de links é enorme em diversos tipos de redes, dentre elas redes sociais.

A predição de links em redes complexas têm atraído cada vez mais atenção de comunidades físicas e informáticas. Os algoritmos podem ser usados para extrair informações, identificar interações espúrias, avaliar a evolução da rede mecanismos, e assim por diante. Valverde-Rebaza e Lopes (2014) apresentou uma abordagem de predição de links em redes sociais em que propôs três medidas baseadas no algoritmo Vizinhos Comuns para a tarefa de previsão de links que levam em conta todas as diferentes comunidades às quais os usuários pertencem. Para comparação utilizou os seguintes algoritmos Vizinhos Comuns, Adamic Adar, Coeficiente de Jaccard, Índice de Alocação de Recursos e Índice de Conexão Preferencial. Os resultados mostram que essa abordagem supera a previsão de link não supervisionada de última geração e ajuda a melhorar a tarefa de previsão de link abordada como uma estratégia supervisionada porém o desempenho de nossas propostas são influenciadas pela estrutura topológica da rede analisada e pelo coeficiente de aglomeração.

Usar métodos de predição para guiar os experimentos laboratoriais em vez de verificar cegamente todas as interações possíveis reduzirá muito os custos experimentais (PAN et al., 2016). Com intuito de descobrir associações de doenças propomos uma metodologia que integra bases de dados epidemiológicas e genéticas para prever futuras comorbidades.

3

Contextualização

Este projeto envolve o estudo bibliográfico recente sobre o tema, a descrição da metodologia, a implementação de um modelo computacional de predição de links e, finalmente, os testes e a validação. Nas seções seguintes, apresentaremos uma visão ampla da metodologia, bem como as ferramentas que poderão ser utilizadas para o desenvolvimento. Uma visão global da nossa proposta metodológica é apresentada. A seguir descreveremos os recursos informacionais que foram úteis no desenvolvimento deste projeto. Seguem as descrições.

3.1 Visão Geral

Neste trabalho, é proposta uma metodologia de predição de associação de doença-doença através da integração de dados públicos de genes, de doenças e suas comorbidades, visando a descoberta de associações de comorbidades. A Figura-1 apresenta uma visão geral do fluxo básico da metodologia. Inicialmente serão extraídas as relações genes-doenças do KEGG *Disease* e criamos duas redes: rede gene-doença e rede doença-doença. A primeira servirá como base para a construção da segunda rede, doença-doença. Na rede doença-doença KEGG, uma doença estará conectada com outra se compartilhar pelo menos um gene, essa rede servirá para identificar possíveis comorbidades. Para garantir uma padronização e integração dessa pesquisa com diversas bases de dados, foi realizado o mapeamento dos códigos de doenças do KEGG *Disease* para o código de CID-10. Ambas redes construídas são ponderadas e não direcionadas.

Em paralelo foi construída a rede epidemiológica do DATASUS a partir dos pares de diagnóstico principal e secundário presentes nas AIHs extraídos do DATASUS. Está também é uma rede doença-doença. Cada registro de internação tem até duas doenças informadas: a) o diagnóstico principal que é a principal causa de internação e; b) um diagnóstico secundário opcional que descreve uma condição que coexistiu no momento da admissão, ou que mais tarde se desenvolveu. Somente os registros que continham os dois diagnósticos que foram extraídos para formar a rede de associação de doença-doença

DATASUS.

Com a construção das duas redes doença-doença aplicamos na rede KEGG métodos de predição de links para encontrar possíveis associações de doenças. Para a validação da predição utilizamos métricas de medição e com base nessas métricas foi criado um ranking de futuras associações doença-doença da rede KEGG. Com os pares mais bem colocados nesse ranking realizamos uma busca desses pares na rede DATASUS e somente os pares com número de ocorrências maior que 100 são considerados possíveis pares de comorbidades.

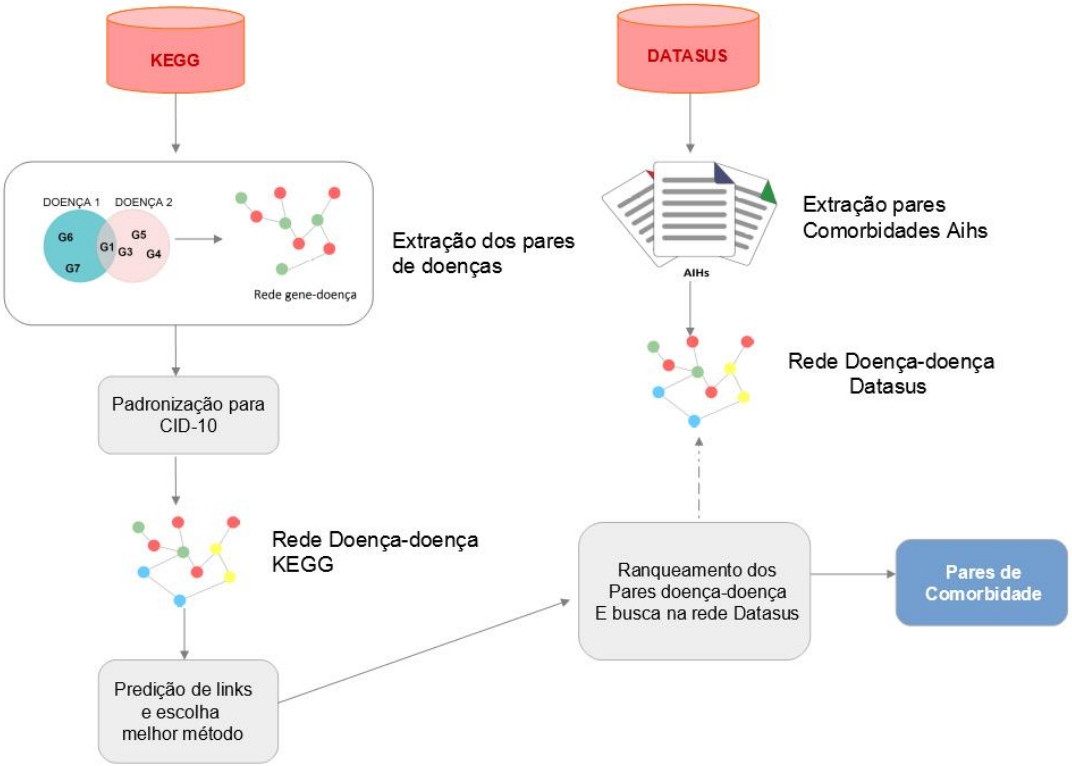


Figura 1 – Fluxo básico da metodologia do projeto.

3.2 SIH/SUS

O Sistema de Internações Hospitalares do SUS (SIH/SUS), criado em agosto 1981, é um banco de dados administrativo, que tem como a sua principal finalidade o registro de internações hospitalares custeadas pelo SUS (PORTELA et al., 1997; DATASUS, 2018). A base de dados do SIH/SUS reúne informações sobre as internações efetuadas através das AIHs que é o documento utilizado para o reembolso dos serviços prestados sob regime de internação com vinculação com o SUS (CAMPOS et al., 2000).

Os registros das AIHs são enviados ao DATASUS[1] e mensalmente repassados a todos os estabelecimentos de saúde pública. Este sistema, permite que o gestor possa fazer uma análise do perfil epidemiológico da morbidade, da mortalidade hospitalar e serve de base para o planejamento e adoção de ações de controle de doenças entre outras (DATASUS, 2018; LOYOLA FILHO et al., 2004).

Os registros de comorbidade provenientes do DATASUS foram utilizados neste projeto como fonte de informação para extração de associações de doenças para criação da rede de comorbidades.

3.3 KEGG

O Kyoto Encyclopedia of Genes and Genomes (KEGG) é uma base de conhecimento e de referência para as funções moleculares de alto nível dos sistemas biológicos. O KEGG mostra-se como uma coleção de base de dados sobre genomas, caminhos metabólicos, doenças, drogas e diversas substâncias químicas importantes para funções moleculares de alto nível, ou seja, funções celulares e do organismo (KANEHISA et al., 2014).

O KEGG tornou-se um dos bancos de dados biológicos mais utilizados no mundo. Ele é curado manualmente com base na literatura publicada. Todos os dados em KEGG e ferramentas de software associados são disponibilizados <http://www.kegg.jp/>. O banco de dados é composto por 16 principais bases de dados que são classificadas em sistemas genômicos, químicos e informação de saúde (KANEHISA et al., 2016). Neste projeto será enfatizada a análise a respeito das associações de doenças associadas a um conjunto de genes. Para isso utilizamos o (KEGG *BRITE*) e o (KEGG *Disease*) que serão apresentados abaixo:

3.3.1 KEGG *BRITE*

O KEGG *BRITE* é um banco de dados de ontologias que representa hierarquias funcionais de vários objetos biológicos, incluindo moléculas, células, organismos, doenças e drogas, bem como relações entre eles. Alguns arquivos de hierarquia *BRITE* contêm atributos delimitados por tabulação que são adicionados manualmente ou computacionalmente. Dentre suas anotações estão classificações de doenças, incluindo: doenças infecciosas, doenças humanas e classificação de doenças no código CID-10. (KANEHISA et al., 2009).

[1] Departamento de Informática do SUS da Secretaria Executiva do Ministério da Saúde.

3.3.2 KEGG *Disease*

A base de dados KEGG *Disease* é uma coleção de informações sobre doenças. Esta coleção captura os conhecimentos sobre perturbações genéticas e ambientais. As doenças são vistas como estados perturbados do sistema molecular. As informações sobre doença são disponibilizadas em duas formas: via mapas e listas de gene/molécula (KANEHISA et al., 2009).

O KEGG mantém seu próprio sistema de códigos. Cada doença é identificada pela letra H seguida de uma sequência de números, por exemplo, H00056 que corresponde a doença Alzheimer. Cada doença também contém uma lista de fatores genéticos conhecidos (genes relacionados a doença), categoria que pertencem a doença, fatores ambientais, vias metabólicas e outros (KANEHISA et al., 2009).

3.4 Medidas de Redes Complexas

Redes Complexas surgiram como uma área multidisciplinar da Ciência que abrange várias áreas de conhecimentos para resolver diversos problemas tais como redes sociais, redes biológicas, Internet e entre outros. Redes complexas visam a estudar um tipo de grafo que apresenta propriedades topográficas bastante específicas, não encontradas em grafos mais simples, tais como elevado número de entidades ou ligações. São formados por um conjunto de vértices (nós) que são interligados por meio de arestas, e os vértices e arestas podem representar objetos diferentes (BARABÁSI et al., 2002).

Redes complexas descrevem uma ampla gama de sistemas na natureza e na sociedade e possuem características topológicas não triviais, não totalmente regulares e nem totalmente aleatória. Modelam a interação entre entidades, como por exemplo, relações entre pessoas, redes de epidemias, rede de proteínas e a Internet (NEWMAN, 2003).

Uma rede complexa pode ser definida por um grafo $G(V, E)$ composta por dois conjuntos finitos V e E, onde V é o conjunto de vértices (nós) e E o conjunto de arestas (ligações) que conectam pares de vértices.

A literatura apresenta diversas medidas e conceitos usados para caracterizar a estrutura das redes complexas. As medidas são utilizadas para analisar as propriedades estatísticas que descrevem a estrutura e o comportamento da rede. A seguir, apresentaremos algumas dessas medidas:

Grafo não direcionado: Define-se como grafo não direcionado aquele que suas arestas apresentam uma relação de adjacência simétrica. Sejam u e v vértices. Nos grafos não direcionados as arestas (u, v) e (v, u) são consideradas como únicas.

Grafo ponderado: Se suas arestas possuem pesos, valores numéricos, associados

a elas.

Hipergrafo: é uma generalização de um grafo, com suas arestas ligando quaisquer quantidades positivas de vértices.

Grafo bipartido: grafo não direcionado $G = (V, E)$ no qual V pode ser particionado em dois conjuntos V_1 e V_2 tais que toda aresta conecta um vértice em V_1 também conecta um vértice em V_2, ou seja, V_1 e V_2 são conjuntos independentes.

Grau: Seja A uma aresta entre dois vértices u, v. O grau do um vértice u denominado k_u, é o número de arestas conectadas a ele (NEWMAN, 2010). Apesar da simplicidade do conceito é uma das métricas base para várias outras.

$$\mathrm{k}_u = \sum_{j=1}^{n} a_{u,v}$$

Grau médio: é média aritmética do grau de todos os seus vértices.

$$\mathrm{k} = \frac{1}{n} \sum_{i=1}^{n} k_i$$

Em que n é o número de vértices.

Diâmetro: é o comprimento, em número de arestas, do caminho geodésico mais longo entre quaisquer dois vértices.

Coeficiente de aglomeração: expressa a presença de ciclos mínimos em uma rede, ciclos com apenas três vértices formando triângulos. Quando um vértice U está conectado a um vértice V, e o vértice V a um vértice X. Essa medida verifica se o vértice U está conectado ao vértice X. O coeficiente aglomeração C de uma rede é obtido conforme mostrado abaixo:

$$\mathrm{C} = \frac{3 \times \triangle}{v}$$

Em que $\triangle$ refere-se ao número de triângulos na rede e, v representa o número de vértices com arestas não direcionadas para o outro par de nós. O fator 3 no numerador é para garantir que o coeficiente seja um valor entre 0 e 1 .

3.5 Métodos de Predição em Links

O problema de predição do link atrai muita atenção de diferentes comunidades de pesquisa. Atribui-se essa atenção principalmente à sua ampla aplicabilidade. Para algumas redes, especialmente redes biológicas, como a redes de interação proteína–proteína, redes metabólicas e redes alimentares, a descoberta de links ou as interações são dispendiosas

no laboratório ou no campo. A predição de links pode ser usados para extrair informações faltantes, avaliar o mecanismo de evolução da rede e assim por diante (LU; ZHOU, 2010).

Dada uma rede $G = (V, E)$, onde V e E são conjuntos de nós e arestas respectivamente, onde múltiplas arestas e auto–conexões não são permitidos. Sendo G uma rede não direcionada, o número de arestas possíveis denominado U é:

$$|U| = \frac{|V|(|V|-1)}{2}$$

Em que $|V|$ denota o número vértices ou elementos do conjunto V. O conjunto de links inexistente da rede é $U - E$. A tarefa de predição de links visa descobrir os links faltantes ou links futuros no conjunto $U - E$ atribuindo uma pontuação, ou probabilidade de ocorrência, para cada link neste conjunto. Quanto maior a pontuação, maior a probabilidade de conexão, e vice-versa. Para a validação dos métodos de predição podemos utilizar a validação de sub-amostragem aleatória e a validação cruzada que eram apresentadas a seguir.

3.6 Validação

Na validação sub-amostragem aleatória, dividimos aleatoriamente o conjunto de links E, em duas partes: o conjunto de treinamento E^t, é tratado como os links conhecidos, e o conjunto de teste E^p. Nenhuma informação do conjunto de teste pode ser usada para predição. Tal que, $E^t \cup E^p = E$ e $E^t \cap E^p = \emptyset$. A vantagem validação de sub-amostragem é que a proporção do conjunto de treinamento utilizada não depende do número de iterações.

Na validação cruzada o objetivo é definir um conjunto de dados para testar o modelo durante a fase de treinamento. A técnica consiste em dividir o conjunto de links em k subconjuntos aleatoriamente. Cada vez que um subconjunto é selecionado como conjunto de teste, os demais $k - 1$ dados constituem o conjunto de treinamento. Esse processo de validação é repetido k vezes com cada um dos k subconjuntos sendo usados com conjunto de teste. Duas métricas padrão são usadas para quantificar a precisão dos algoritmos de predição, são elas: AUC e precisão. Uma breve descrição dessas duas métricas é apresentada a seguir.

3.7 Métricas de Avaliação

Inicialmente, o algoritmo de predição de links fornece uma lista ordenada de todos os links não observados $(U - E^t)$. Para cada link não observado, $x, y \, \epsilon \, U - E^t$, em que a pontuação $S_{x,y}$ é somada para quantificar a probabilidade de sua existência. A área sob a curva (AUC) avalia o desempenho do algoritmo de acordo com a lista inteira, enquanto a precisão

só se concentra nos links com as pontuações mais altas. Uma introdução detalhada dessas duas métricas é a apresentada:

AUC: Dada a lista das pontuações de todos os links não observados, o valor da AUC pode ser interpretado como a probabilidade que um link faltante escolhido aleatoriamente (E^p) ter uma pontuação mais alta em comparação à escolha aleatória de um link não existente ($U - E$). Na implementação dos algoritmos de predição de links, usualmente é calculada a pontuação de cada link não observado em vez de criar uma lista ordenada, pois isso pode necessitar de maior tempo de processamento. A cada vez que escolher aleatoriamente um link faltante e um link não existente, serão comparadas suas pontuações (ZHOU; LÜ; ZHANG, 2009). Se entre n comparações independentes, há n' comparações, onde os links faltantes têm uma maior pontuação e n'' comparações onde ambos têm a mesma pontuação, então, o valor da AUC é:

$$\text{AUC} = \frac{n' + 0.5n''}{n}$$

Todas as pontuações são geradas considerando que são idependente e identicamente distribuída. Portanto, o valor da AUC deve-se aproximar de 0,5. Então, a pontuação em que o valor excede 0,5 indica quanto o desempenho do algoritmo de predição avaliado é melhor em comparação com um de predição por simples aleatoriedade.

Precisão: dada a lista ordenada das pontuações dos links não observados, a precisão é definida como a proporção de links escolhidos que foram selecionados para o número de links selecionados. Isto é, se tomamos os L links do topo da lista de como os links preditos, entre os quais Lr, são os links preditos corretamente (Lr links pertencem ao conjunto de avaliação E^P), então a precisão é dada por:

$$\text{Precisão} = \frac{Lr}{L}$$

Uma maior precisão significa uma maior desempenho na predição.

Segundo (LU; ZHOU, 2010), muitas técnicas de predição de links foram desenvolvidas e elas podem ser divididas em três grupos: a) índices baseados na similaridade; b) métodos baseados na máxima verossimilhança e; c) modelos probabilísticos. Os índices baseados na similaridade baseiam-se nas informações estruturais, topológicas, da rede. Esse índices podem ser do tipo local, nos quais são utilizadas somente as informações de um par de vértices, ou tipo global, nos quais são utilizadas as informações de toda a rede (LIBEN-NOWELL; KLEINBERG, 2007)

3.7.1 Índices de Similaridade Local

Os índices de similaridade local caracterizam-se por usarem informações só de um par de vértices.

Vizinhos Comuns (CN): dois vértices x e y, em que $\Gamma(x)$ é conjunto de vizinhos do vértice x, têm uma maior probabilidade de um futuro relacionamento se eles tem muitos vizinhos em comum. A pontuação é calculada conforme a fórmula abaixo:

$$\mathrm{S}_{x,y}^{CN} = |\Gamma(x) \cap \Gamma(y)| \tag{3.1}$$

Newman (2001) calculou o índice de CN no contexto de redes de colaboração, verificando uma correlação positiva entre o número de vizinhos comuns e a probabilidade de que dois cientistas possam colaborar no futuro.

Adamic-Adar (AA): Esse índice refina a simples contagem dos vizinhos comuns, atribuindo um maior peso aos vizinhos com menor número de conexões (ADAMIC; ADAR, 2003). Esse índice de similaridade é definido como:

$$\mathrm{S}_{x,y}^{\mathrm{AA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z} \tag{3.2}$$

Índice de Conexão Preferencial (PA): proposto por Newman (2001), Barabási et al. (2002), que verificaram que a probabilidade de um link entre dois vértices é proporcional ao produto do número de vizinhos que cada um possui. Pode-se notar que o índice PA não requer nenhuma informação da vizinhança. Sendo assim, esse índice tem a menor complexidade computacional. Definido como:

$$\mathrm{S}_{x,y}^{\mathrm{PA}} = k_x \times k_y \tag{3.3}$$

Índice de Alocação de Recursos (RA): considerando um par de nós x e y, não conectados diretamente, o nó x pode enviar algum recurso para y utilizando seus vizinhos comuns como transmissores. No caso mais simples, cada transmissor tem uma unidade de recurso e vai distribuí-la igualmente aos seus vizinhos. A semelhança entre x e y pode ser definida como a quantidade de recursos que y tem recebido de x (ZHOU; LÜ; ZHANG, 2009). Definido como:

$$\mathrm{S}_{x,y}^{\mathrm{RA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \tag{3.4}$$

Esta medida é simétrica, ou seja, $S_x y = S_y x$. O índice AA e o índice RA possuem forma similar.

3.7.2 Índices de Similaridade Global

Os índices de similaridade global caracterizam-se por envolver o uso de toda ou quase toda a informação topológica presente na rede.

Katz: esta heurística computa o resultado da soma do tamanho de todos os caminhos possíveis entre dois vértice. Conceitualmente o índice de *Katz* baseia-se na expectativa de que quanto mais caminhos houver entre dois vértices e quanto mais curtos forem esses caminhos, mais forte será a conexão (KATZ, 1953).

$$\mathrm{S}_{x,y}^{katz} = \sum_{xy}^{\infty} \beta^l \cdot \left|\mathrm{paths}_{xy}^{(l)}\right| = \beta A_{x,y} + \beta^2 A_{x,y}^2 + \beta^3 A_{x,y}^3 + ... \tag{3.5}$$

para paths_{xy} o conjunto de todos os caminhos de tamanho l conectando x e y. Sendo que $\beta \in (0, 1)$ é um parâmetro livre que controla os pesos dos caminhos de acordo com seus tamanhos. O alto custo computacional para o cálculo do índice *Katz* pode ser minimizado através do cálculo da matriz de similaridade que pode ser escrita, conforme apresentado abaixo:

$$\mathrm{KATZ} = (I - \beta A)^{-1} - I$$

Para A e I as matrizes adjacente e de identidade do grafo analisado. β é calculado de forma que $\beta < \frac{1}{\|A\|^2}$.

4

Método para Predição de Associação de Doenças

Neste capítulo são apresentados os métodos adotados para o desenvolvimento deste projeto tais como: a coleta de dados na base KEGG e DATASUS para a geração das redes, a predição de links aplica para a descoberta de associações de doenças.

4.1 Pipeline Etapas do Projeto

O fluxo do processo é dividido em três etapas e dentro de cada etapa existem operações. A Figura 2 ilustra as três etapas que foram realizadas para a execução desse projeto. Cada uma dessas etapas e operações serão descritas a seguir:

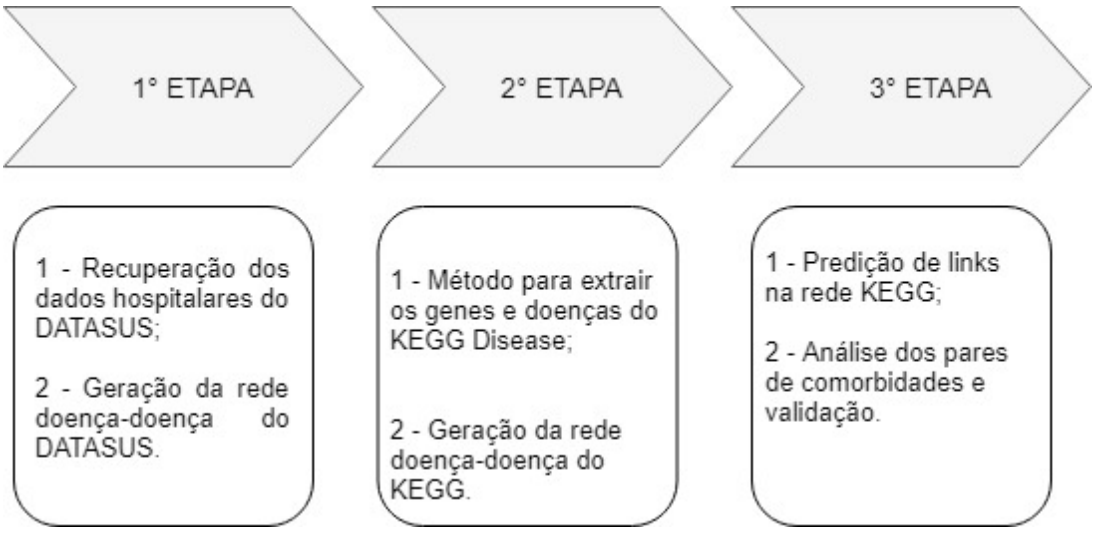


Figura 2 – Pipeline de execução das etapas do projeto.

1. Na primeira etapa construímos uma rede doença-doença a partir dos pares de diagnóstico principal e secundário, presentes nas AIHs, extraídos do Sistema de

Internações Hospitalares (SIH) do DATASUS. O SIH contém informações sobre as internações de todos os hospitais públicos ou conveniados do SUS. As AIHs incluem informações de internações realizadas em todo país, tais como valor dos procedimentos realizados, datas de entrada e outros, assim como diagnóstico primário e secundário, com seus respectivos códigos CID-10. Cada registro de internação tem até duas doenças informadas: a) a principal causa de internação, e; b) um diagnóstico secundário opcional que descreve uma condição que coexistiu no momento da admissão, ou que mais tarde se desenvolveu. Essas informações foram extraídas para formar a rede de associação de doença-doença do DATASUS;

2. A segunda etapa consiste em extrair as associações de doença–gene do banco de dados KEGG *Disease* que possui em torno de 1.704 tipos de doenças humanas[1].

   Para cada doença foi criada uma lista de genes relacionados com determinada doença. A partir dessa lista gene-doença foram extraídas a sobreposição de genes relacionados com duas doenças ou mais doenças.

   Para uma melhor compreensão a Figura 3 representa a sobreposição de genes de duas doenças, D_1 e D_2. Seus respectivos genes extraídos do KEGG *Disease* também estão representados. Para esses destacamos que os genes G_1, G_2 e G_3 representam a relação gene-doença e são possíveis explicações para esta comorbidade.

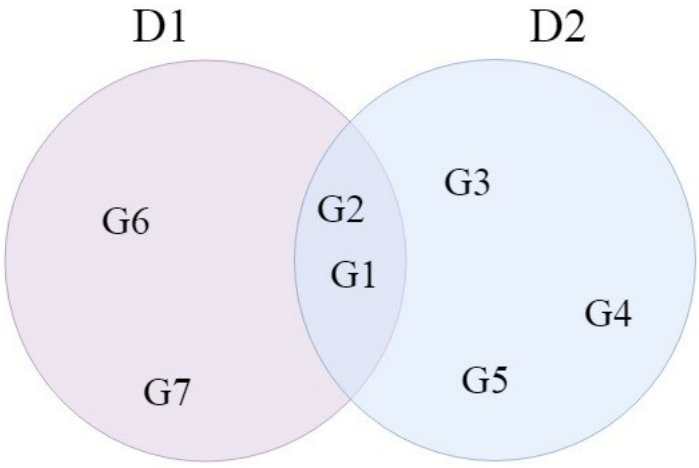


Figura 3 – Sobreposição de genes entre duas doenças.

A rede gene-doença serviu como base para a rede doença-doença. Na rede doença-doença, uma doença está conectada com outra se compartilhar pelo menos um gene. Para garantir uma padronização e integração dessa pesquisa com diversas bases de dados, foi realizada uma padronização, o mapeamento dos códigos de doenças

[1] Última atualização do KEGG *Disease* em 8 janeiro de 2017.

do KEGG *Disease* para o código de CID-10, esse mapeamento é feito utilizando uma lista do KEGG *Disease* e nessa etapa verificamos que o KEGG *Brite* agrupa algumas doenças do KEGG na transformação para CID-10. Isso se deve ao que trabalhamos apenas com as categorias do CID-10 que é uma forma mais generalizada. As categorias do CID-10 é grupo de doenças e sintomas que tem a mesma etimologia;

3. Na terceira etapa foi realizada a predição utilizando os cinco métodos de predição de links:CN, AA, PA, RA e *Katz*.Esses métodos foram aplicados à rede KEGG para descobrir elos de doença-doença ausentes. Cada método atribui uma pontuação (score) a cada aresta potencial. Uma análise foi realizada para verificar o nível de concordância entre os métodos e com base nessa análise foi escolhido um método para predizer novas arestas a rede KEGG. E finalmente as comorbidades sugeridas pelo método de predição de links foram avaliadas contra a rede DATASUS.

   A metodologia aqui adotada assume que as redes doenças são representadas como grafos não direcionados que não permitem que os nós se autorrelacionem.

4.2 Recuperação de dados hospitalares do DATASUS e geração da rede doença-doença DATASUS

As AIHs registram até dois diagnósticos por internação: o diagnóstico principal e o diagnóstico secundário, ambos codificados pelo sistema CID-10. De acordo com o Manual do SUS, o diagnóstico principal é o motivo responsável pela admissão do paciente no hospital. E o diagnóstico secundário são todas as condições que se desenvolvem durante o período de internamento ou que afetem o tempo de uma rede doença-doença a partir dos pares de diagnóstico principal e secundário presentes nas AIHs extraídos do DATASUS. Além disso, as AIHs contém outros dados, tais como: datas de entrada e saída do paciente, os procedimentos realizados pelo paciente durante a sua internação, idade, sexo, etnia e endereço (MINISTÉRIO da SAÚDE, 2013).

Apenas em 2017, foram processadas no total de 11.673.773 milhões de AIHs processadas em todo país (DATASUS, 2018). O SIH/SUS é uma grande fonte de dados para estudos epidemiológicos e para a área de investigação em serviços da saúde (LOYOLA FILHO et al., 2004).

Com base nos diagnósticos principal e secundário e seus respectivos código CID-10 anotados nas AIHs, extraímos os pares de comorbidades. Nessa extração eliminamos todos os registros que não estavam completos, isto é, que não possuíam diagnóstico primário e secundário. Essa coleta foi realizada e registrada no período entre 1998 e abril/2017. Todas as AIHs que foram extraídas estão disponíveis no DATASUS pelo link <http://www2.datasus.gov.br/DATASUS/index.php?area=0901&item=1&acao=25>. Os

arquivos das AIHs são em formato específicos e para a leitura/conversão desses arquivos foi utilizado o programa *TabWin*. No final dessa etapa conseguimos extrair o total de 18.956.582 milhões de registros que possuíam o diagnóstico principal e secundário. Esses pares de comorbidades foram utilizados para a criação da segunda rede, a rede DATASUS, que se baseia na associação de duas doenças que aparecem nas AIHs. A rede DATASUS também é ponderada, onde os pesos das arestas foi quantificado, usando coocorrência da associação dos pares de doenças extraídos.

4.3 Método para extrair genes e doenças do *KEGG Disease*

A coleta de genes é baseada em uma busca na base de dados KEGG, mais especificamente no KEGG *Disease*. O banco KEGG *Disease* não relaciona doenças entre si. No KEGG *Disease*, todas as doenças estão relacionadas a um gene ou mais. E é com base nessa informação que extraímos pares de doenças que compartilham um ou mais genes.

Para a extração dos genes e das doenças do KEGG, utilizaremos o framework `BioServices`, desenvolvido em *Python* que fornece acesso aos principais serviços Web de bioinformática que está disponível em <http://pypi.python.org/pypi/bioservices> sob uma licença *GPL-v3*. Entre os serviços disponíveis está a interface para KEGG, com alguns métodos implementados para que o usuário possa utilizar os recursos desses bancos de dados. O pacote `BioServices` fornece interface para o serviço KEGG e também fornece uma interface para a `API REST KEGG` (COKELAER et al., 2013). Por meio dessa interface uma lista de doenças e seus respectivos genes foram extraídas.

4.4 Geração da rede doença-doença

A partir da lista doença–gene da etapa anterior foi construído uma grafo da rede de doença–gene, em que os nós são doenças ou genes, e duas doenças estão conectadas, se compartilham, pelo menos, um gene comum. E as arestas são ponderadas de acordo com número de genes que duas doenças compartilham.

A segunda fase dessa etapa foi criar um grafo bipartido de doença-doença com base no grafo doença–gene. Para essa tarefa contamos com o apoio do software NetworkX. O NetworkX é um pacote desenvolvido na linguagem *Python* para a criação, manipulação e estudo da estrutura, dinâmica e funções de redes complexas (HAGBERG; SCHULT; SWART, 2008). O pacote oferece uma intefarce para implementação de grafos e ferramentas para o estudo de estrutura de redes sociais, biológicas e de infraestrutura (HAGBERG; SCHULT; SWART, 2008). O pacote está disponível em <https://networkx.github.io>

Com o auxílio deste pacote e de bibliotecas do *Python* como *Numpy* e *Scipy*

desenvolvemos um método que tem como objetivo produzir uma segunda lista doença-doença.

Entretanto o KEGG *Disease* possui uma classificação específica para as suas doenças. Assim as doenças foram extraídas com os códigos do KEGG e, com a ajuda do KEGG *BRITE*, desse modo foi realizado um mapeamento para a codificação CID-10. O banco de dados KEGG trabalha com um código único para cada doença com uma nomenclatura do próprio banco. Para uma melhor padronização os códigos de doenças do KEGG foram mapeados para CID-10. Para esse mapeamento foi utilizado um arquivo do KEGG *BRITE* onde as doenças são classificadas de acordo com uma hierarquia. As doenças do KEGG Disease estão mapeadas em categorias do CID-10 que correspondem ao código de uma letra e dois dígitos do CID-10 no KEGG *BRITE*. Assim realizamos o mapeamento de cada código do KEGG *Disease* para o código do CID-10. Devido ao fato de o mapeamento ser por categorias, essa padronização acarretou alguns agrupamentos de doenças, pois em alguns casos duas ou mais doenças possuem o mesmo código no CID-10, mas isso gerou alguns auto-loops na rede. Com o mapeamento das doenças para código CID-10 no final dessa etapa obtivemos como resultado uma lista doença-doença.

E assim foi construída a primeira rede doença-doença não direcionada e ponderada baseada no banco de dados KEGG, onde os nós são as doenças padronizadas pelo código CID-10 e arestas ponderadas pela quantidade de genes que duas doenças compartilham. Para a visualização dessa primeira rede obtida utilizamos o software *Gephi* (BASTIAN; HEYMANN; JACOMY, 2009). O *Gephi* é uma plataforma de software de código aberto para a visualização e análise de grafos e serve para todos os tipos de rede.

4.5 Predição de links na rede KEGG

Depois que as redes foram criadas, podemos então iniciar o processo de predição de links. A metodologia aqui adotada assume que a rede KEGG, $K = (V, E)$ é representada como grafo não direcionado, não ponderado e que não permite que os nós se autorrelacionem. A predição de links foi aplicada na rede KEGG sem os pesos das arestas e sem os auto-loops, pois temos a intenção de predizer associações de pares de comorbidades, isto é, se uma doença está relacionada a outra e para isso foram descartados os pesos das arestas que representam a quantidade de genes. Para a predição os métodos recebem como entrada uma rede que possui arestas não observadas que é definida como $U - E$, onde U é número de aresta em potenciais, e retorna um ranking de tais arestas com valores de score que indica a qualidade de tais arestas. Este trabalho utiliza cinco métodos de predição: Vizinhos Comuns (NC), Adar Adâmico (AA), Índice de Alocação de Recursos (RA), Índice de Conexão Preferencial (PA) e *Katz*, todos descritos no capítulo anterior. Para a escolha do melhor método foi executada a métrica AUC e precisão e o método com melhor AUC e

precisão foi adotado.

4.6 Análise dos pares de comorbidades e Validação

A partir deste ponto, os pares são ordenados de acordo com a similaridade calculada entre seus nós. Apenas os pares que atingirem um limiar aceitável de similaridade serão propostos como links futuros. E esses pares foram buscados na rede DATASUS e foram selecionados os pares com maiores números de ocorrências. Para a validação da metodologia, buscamos esses pares de comorbidades na literatura.

5

Resultados

Neste capítulo, serão apresentados os resultados obtidos pela metodologia apresentada nesta dissertação e algumas discussões sore os resultados obtidos. Também apresentamos duas redes doenças geradas, a metodologia de validação e associações de doenças encontradas.

5.1 Rede DATASUS

A rede epidemiológica de doença-doença, $rede\_DATASUS$, também foi construída a partir dos registros de todos os pacientes admitidos em todos os hospitais do Brasil nos últimos 18 anos. Cada registro admissão de internação AIHs tem até duas doenças informadas: a) a principal causa de internação, e; b) um diagnóstico secundário opcional que descreve uma condição que coexistiu no momento da admissão, ou que mais tarde se desenvolveu. Todos as doenças registradas nas AIHs são feitas utilizando o código CID-10.

Utilizando apenas os registros contendo diagnósticos primário e secundário, exatamente 18.956.582 registros, foi construída a rede epidemiológica $rede\_DATASUS$, existe um grande número de registro que não possuem os dois diagnósticos, isso pode ser a não ocorrência de complicação na internação do paciente ou devido ao falto do diagnóstico não ser obrigatório o seu preenchimento.

Esta rede é ponderada porque as arestas contêm a frequência das duas doenças concomitantes. A rede $rede\_DATASUS$ tem 1.941 nós e 248.508 arestas. A própria $rede\_DATASUS$ é seu principal componente fortemente conectado. A Figura 4 ilustra a $rede\_DATASUS$, para uma melhor visualização $rede\_DATASUS$ foi representada pelos 22 capítulos do CID-10 devido ao grande número de nós. Os capítulos do CID-10 é uma hierarquia acima das categorias, cada capítulo abrange um grupo de categorias, por exemplo, o Capítulo 2 Neoplasia abrange as categorias de $C00$ à $D48$. Atualmente no CID-10 temos o total de 1.943 categorias e 22 capítulos, na $rede\_DATASUS$ temos o total de 1.941 abrangendo quase todo as categorias e seus respectivos 22 capítulos.

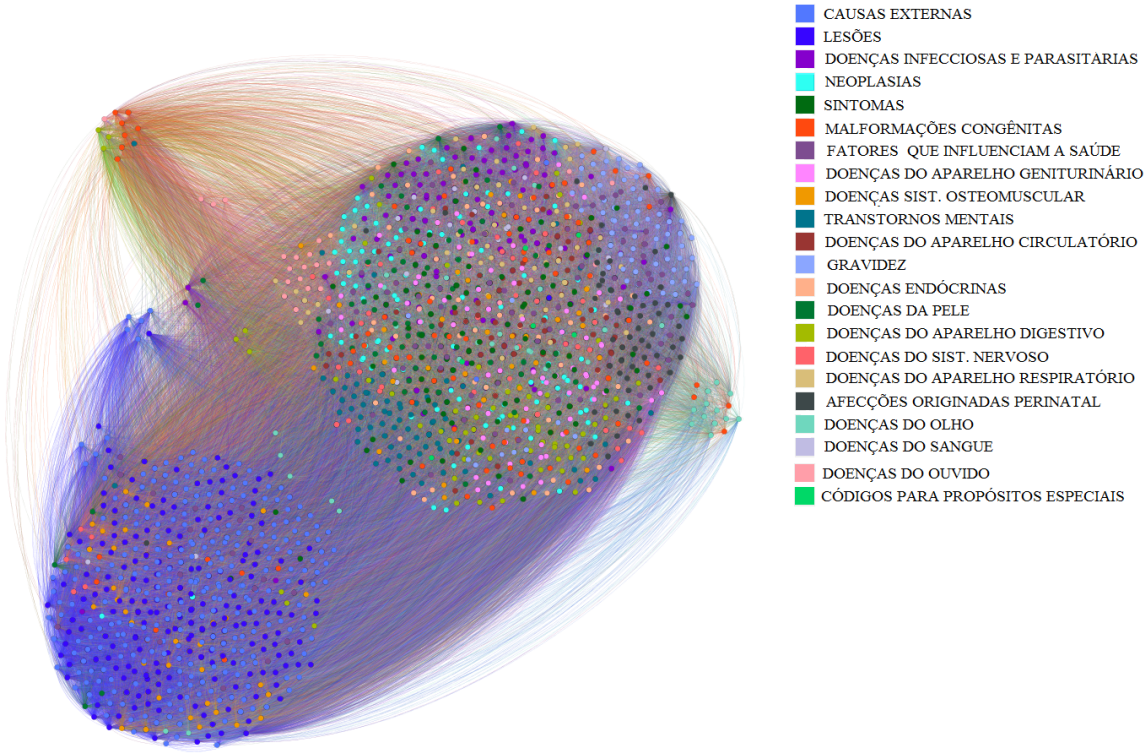


Figura 4 – Rede DATASUS para CID-10 representada pelos 22 capítulos.

Os seis maiores *hubs* são doenças que não possuem fundos genéticos, isso ocorre devido a $rede\_DATASUS$ ser uma rede epidemiológica, a seguir apresentamos os *hubs*:

1. J18 – Pneumonia por microorganismo não especificada
2. I10 – Hipertensão essencial (primária)
3. N39 – Outros transtornos do trato urinário
4. A41 – Outras septicemias
5. J46 – Estado de mal asmático
6. A49 – Infecção bacteriana de localização não especificada

A Figura 5 abaixo mostra o percentual dos nós (doenças) distribuídos pelos 22 capítulos. Os 4 capítulos que se destacam com maior número de doenças são: Causas externas de morbidade e mortalidade (Capítulo 21), Lesões (Capítulo 19), Algumas doenças infecciosas e parasitárias (Capítulo 1) e Neoplasias (Capítulo 2). Isso se dá pelas das condições de saúde da população brasileira, segundo Parreira e Lenea (2012) países em desenvolvimento tem a ter uma predominância de doenças emergentes ou “doenças

negligenciadas e da pobreza", que são caracterizadas pelas doenças infecciosas. Importante ressaltar o capítulo de Neoplasia que representa 7% dos nós da rede.

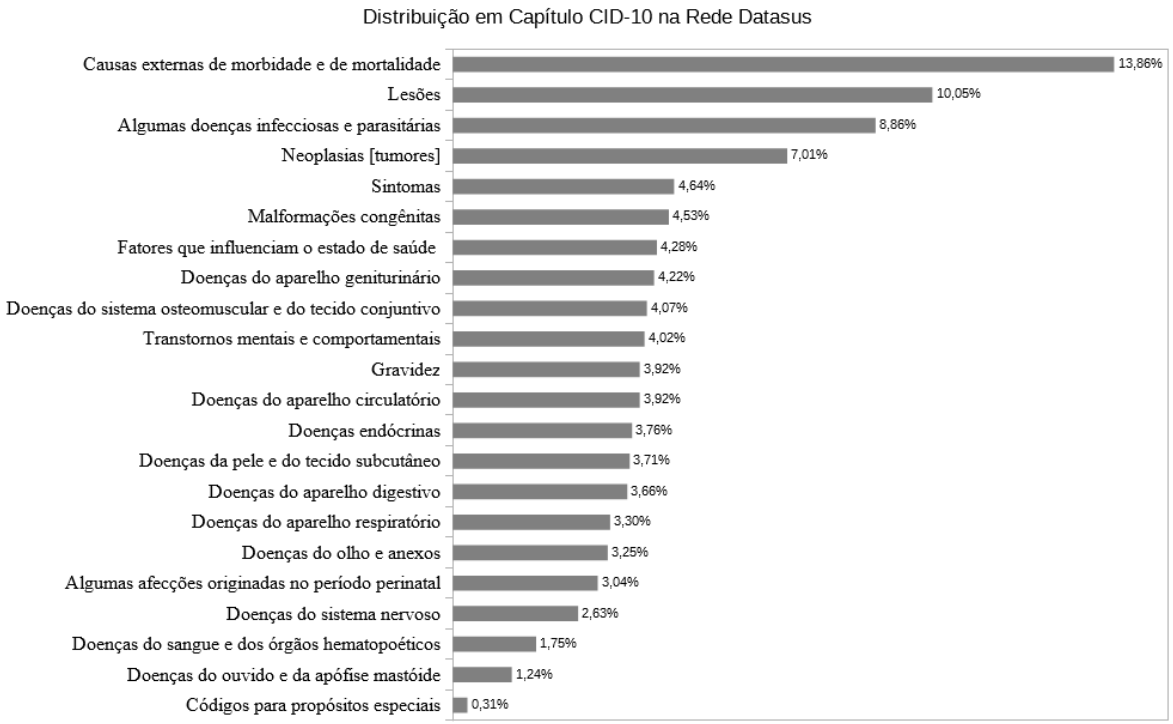


Figura 5 – Distribuição de capítulos da Rede DATASUS totalizando 22 capítulos.

5.2 Rede KEGG

A rede doença–gene extraída possui 1.279 vértices correspondentes a doenças, 2.692 vértices correspondentes a genes e 4.432 arestas. A rede possui o total de 418 componentes conexos sendo que o maior componente com contém 2.778 nós. É possível perceber um grande número de doenças, no total de 297, só com 2 vértices. Esses componentes são doenças ligadas a apenas 1 gene e não estão conectadas a nenhuma outra doença e foram totalmente excluídos. A partir do maior componente da rede doença–gene, foi gerada a rede doença-doença, essa rede não direcionada e ponderada, aqui chamada de G, possui 905 vértices (doenças) e 3.717 arestas. A rede doença-doença apresenta no total 44 componentes sendo que o maior componente conectado do G, apresenta 824 nós e 3.575 arestas.

A Figura 6 ilustra a rede doença-doença gerada a partir do maior componente conexo. A tonalidade de azul dos vértices está relacionado com o grau de cada doença, ou seja, quanto mais conectada uma doença, mais escuro é a tonalidade de azul.

Os maiores *hubs* da rede são todos relacionados a Câncer, isso deve ao fato de ser uma rede genética onde os genes sobre câncer são bem mais estudados pela comunidade

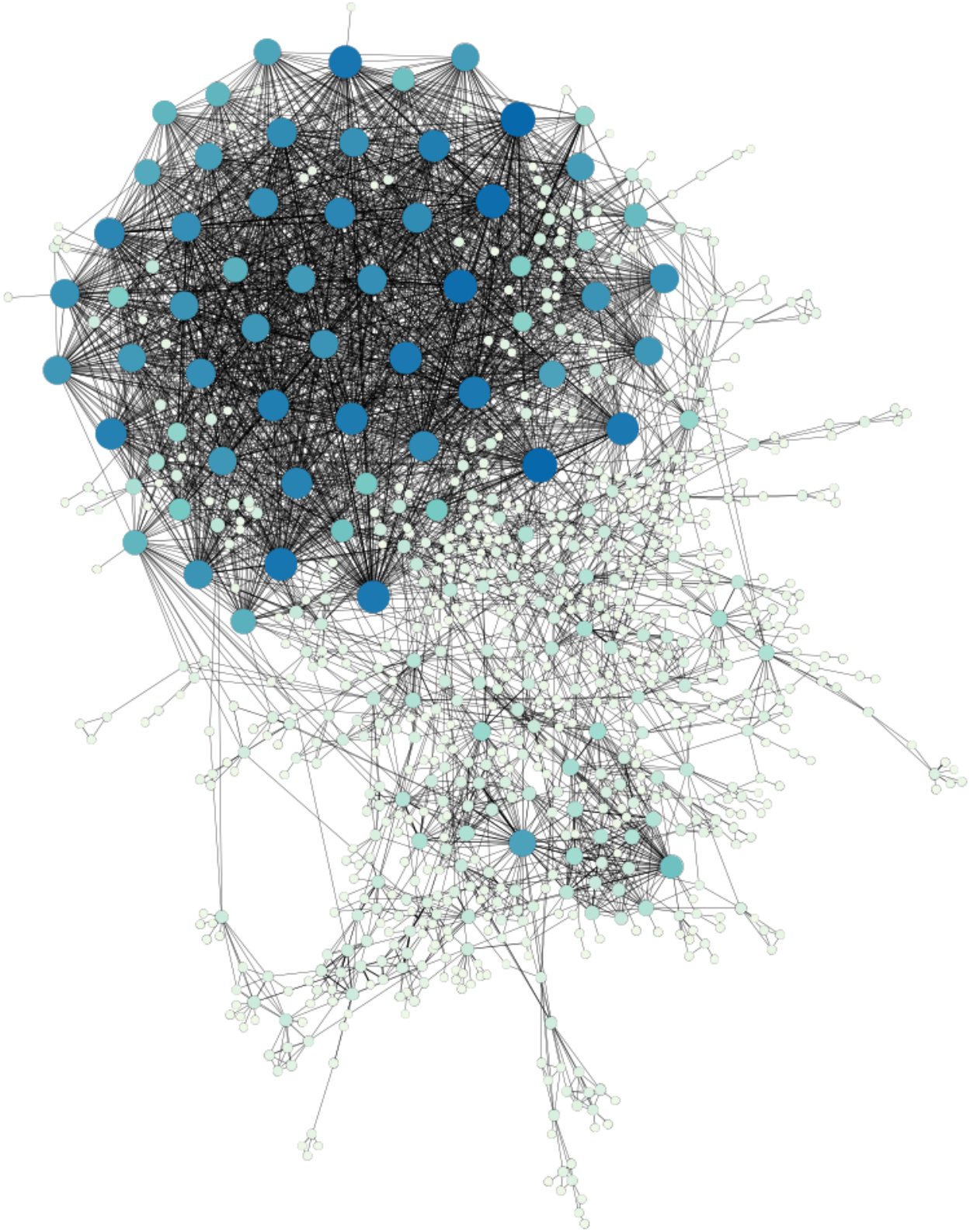


Figura 6 – Rede representando as relações genéticas do KEGG não padronizada para CID-10 entre 824 doenças.

científica principalmente em países desenvolvidos, como por exemplo, no Japão onde é desenvolvido a base de dados *KEGG*.

Os seis *hubs* são apresentados na respectiva ordem:

1. H00016 – Câncer de boca (oral)
2. H00048 – Carcinoma hepatocelular
3. H00040 – Carcinoma de células escamosas
4. H00038 – Melanoma maligno
5. H00025 – Câncer de pênis
6. H00046 – Colangiocarcinoma

A fim de realizar uma comparação com uma rede epidemiológica real, os nomes dos nós de G foram padronizados, conforme apresentado na Seção 4.4, para o sistema de codificação da CID-10.

A nova rede denominada $rede\_KEGG$ padronizada, tem 288 nós e 1.983 arestas. O maior componente tem 279 nós e 1.977 arestas. Abaixo são apresentados os *hubs* da seguinte maneira, primeiro é apresentado o nome do nó em CID-10, por exemplo, Neoplasia maligna do encéfalo ($C71$) e em seguida as doenças que foram agrupadas nessa categoria com os respectivos código do *KEGG*, H00042 referente a doença Glioma. Os seus 6 maiores *hubs* da rede são todos pertencentes a categorias de Neoplasia, apresentados a seguir:

1. C71 – Neoplasia maligna do encéfalo
    - H00042 – Glioma
    - H01667 – Meduloblastoma
2. C91 – Leucemia linfóide
    - H00001 – Leucemia / linfoma linfoblástico B
    - H00002 – Leucemia linfoblástica T / linfoma
    - H00005 – Leucemia linfocítica crônica (CLL)
    - H00006 – Leucemia de células pilosas
    - H00009 – Leucemia de células T para adultos
3. C25 – Neoplasia maligna do pâncreas
    - H00019 – Câncer de pâncreas
    - H00045 – Tumor neuroendocrino pancreático

4. C22 – Neoplasia maligna do fígado e das vias biliares intra–hepáticas

    - H00048 – Carcinoma hepatocelular
    - H01557 – Angiosarcoma hepatico
    - H01666 – Angiosarcoma
    - H00046 – Colangiocarcinoma

5. C44 – Outras neoplasias malignas da pele

    - H00039 – Carcinoma basocelular
    - H00040 – Carcinoma de células escamosas
    - H01555 – Carcinoma de células de Merkel
    - H01666 – Angiosarcoma

6. C00 – C06 – Neoplasia maligna do lábio, Neoplasia maligna da base da língua, Neoplasia maligna de outras partes e de partes não especificadas da língua, Neoplasia maligna da gengiva, Neoplasia maligna do assoalho da boca, Neoplasia maligna do palato e Neoplasia maligna de outras

    - H00016 – Câncer Boca/Oral

A Figura 7 ilustra a rede doença-doença padronizada $rede\_KEGG$ essa rede está representada pelos capítulos do CID-10, dos 22 capítulos do CID-10 a $rede\_KEGG$ possui 19 deles. Para uma melhor apresentação da $rede\_KEGG$ padronizada os nós da rede foi mapeado para os capítulos dos CID-10.

Com a padronização a $rede\_KEGG$ passou a ter alguns auto-loops devido ao agrupamento de doenças. E esse agrupamento acontece porque o banco de dados *KEGG* organiza as doenças com base nos processos biológicos de âmbito celular enquanto o código CID-10 organiza as doenças com base nos fenótipos apresentados. Um exemplo é as 46 doenças que foram agrupadas a categoria Transtornos primários dos músculos $G71$ que estão vinculada a doenças do sistema nervoso e aparece destacada na Figura 7 de cor rosa. Esses loops foram eliminados para aplicação dos métodos de predição de links.

Vale a pena mencionar que, mesmo após a padronização dos nomes das doenças, os seis *hubs* principais permanecem os mesmos que estavam sob a nomenclatura KEGG. A rede_KEGG ainda contínua possuem seus maiores *hubs* relacionado a câncer.

A Figura 8 abaixo mostra o percentual dos nós (doenças) distribuídos pelos capítulos. A rede $rede\_KEGG$ tem nós que estão representados nos 19 capítulos do CID-10, isso demonstra uma rede muito heterogênea.

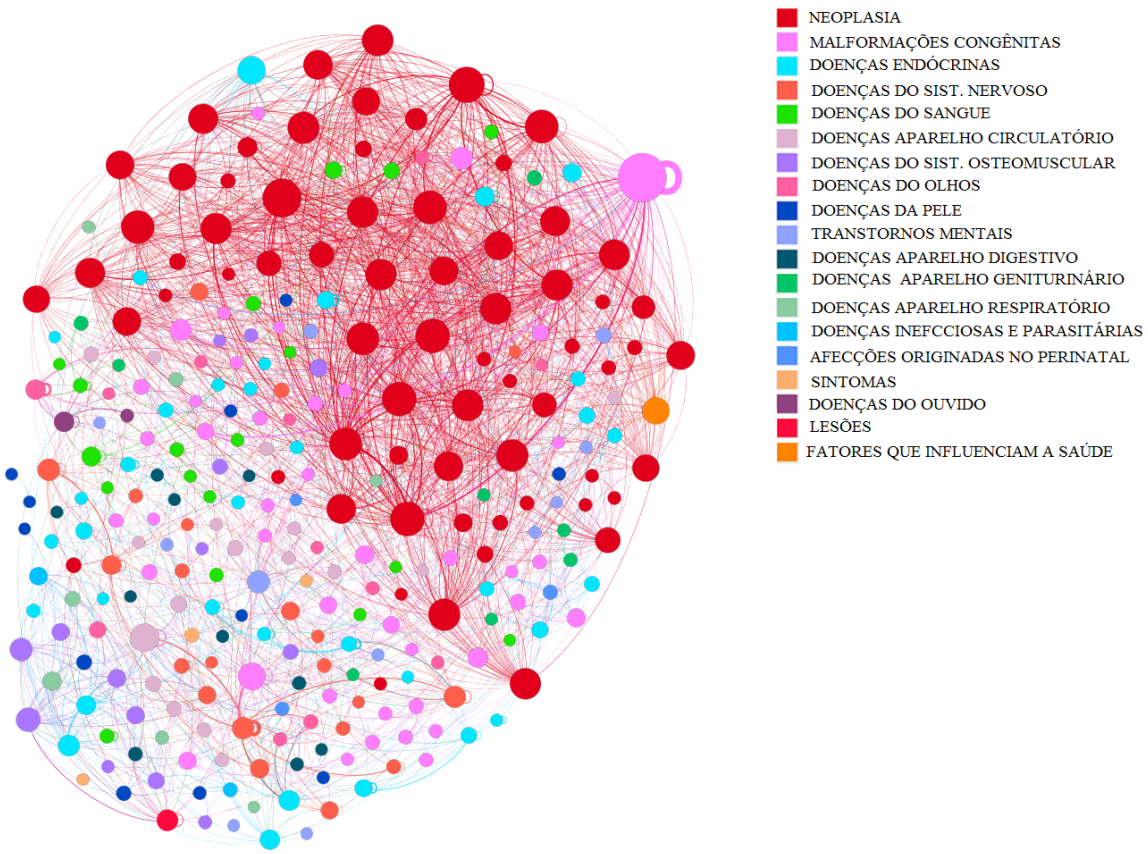


Figura 7 – Rede KEGG padronizada para CID-10 representada pelos 19 capítulos.

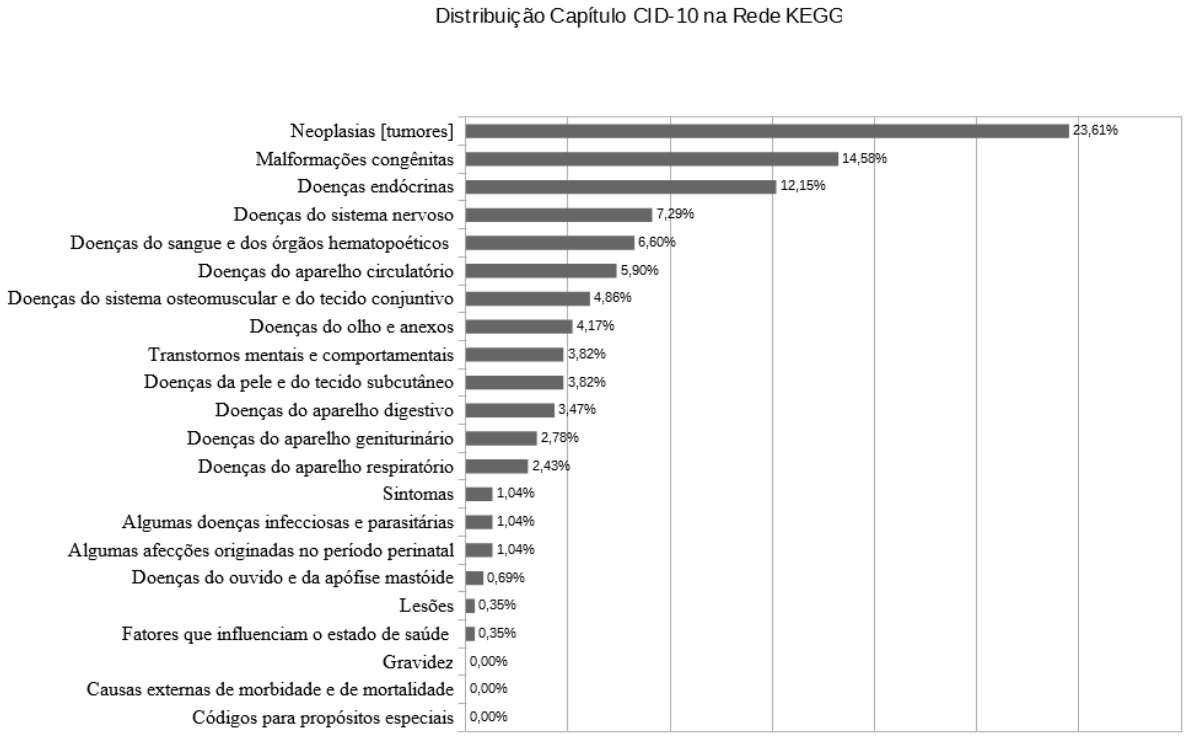


Figura 8 – Distribuição de capítulos da Rede KEGG totalizando 19 capítulos.

5.3 Análise da Rede KEGG e rede DATASUS

A Tabela 1 apresenta uma análise comparativa das principais propriedades topológicas de ambas as redes. Nessa tabela, para cada rede, é apresentada o número de vértices e aresta, grau médio, diâmetro e o coeficiente de aglomeração médio.

Tabela 1 – Propriedades topológicas das redes de doença-doença.

Propriedade	KEGG	DATASUS
Número de vértices	288	1.941
Número de arestas	1.983	248.508
Grau médio	33,431	256,062
Diâmetro	6	4
Coeficiente de aglomeração médio	0,588	0,499

Como pode ser visto na Tabela 1 acima, a $rede\_KEGG$ e a $rede\_DATASUS$ possuem características muito diferentes. A $rede\_DATASUS$ é maior em número de vértices e arestas porém seu diâmetro é menor o que significa um menor caminho entre duas doenças mas possui um coeficiente de aglomeração menor. Enquanto a $rede\_KEGG$ possui uma maior coeficiente de aglomeração. Além das diferenças mencionadas acima, a $rede\_KEGG$ contém vários componentes conexos enquanto a $rede\_DATASUS$ contém apenas um componente conexo.

Acredita-se que na $rede\_KEGG$ existam várias arestas ausentes devido ao que diversas relações genéticas ainda não forma estudadas. Está é uma motivação adicional para considerar a vantagem métodos de predição apresentado anteriormente. Verificou-se também que 1.136 arestas da $rede\_KEGG$ estão presente na rede DATASUS, isso é equivalente a 57% da $rede\_KEGG$. Em se tratar de encontrar arestas de uma rede gênica numa rede epidemiológica esses 57% é um índice elevado. É possível que esse índice possa ser maior, pois nos registros das AIHs o diagnóstico secundário não é obrigatório e pelo também fato dos registro não serem identificados não sabemos se um paciente em um intervalo de tempo teve uma nova complicação.

É instrutivo comparar a distribuição dos pesos das arestas que são comuns à $rede\_KEGG$ e $rede\_DATASUS$ e àquelas que aparecem em nossa $rede\_KEGG$, mas não na $rede\_DATASUS$. Essa comparação é mostrada na Figura 9, onde sobrepomos os dois histogramas, um histograma contém arestas comuns a KEGG e DATASUS e o outro apenas DATASUS. O eixo vertical representa o peso da aresta normalizado, e o eixo horizontal se refere ao número de arestas. Essas arestas que são comuns a ambas as redes (mostradas em cinza) tendem a ter pesos maiores (número de genes comuns) do que aquelas que aparecem apenas na $rede\_KEGG$ (mostrada em preto). Como os pesos de arestas são o número de genes que são comuns às doenças ligadas pelas arestas,

uma explicação plausível para essa discrepância é que comorbidades são mais prováveis de serem observadas quando o número de genes que ligam ambas as doenças é maior.

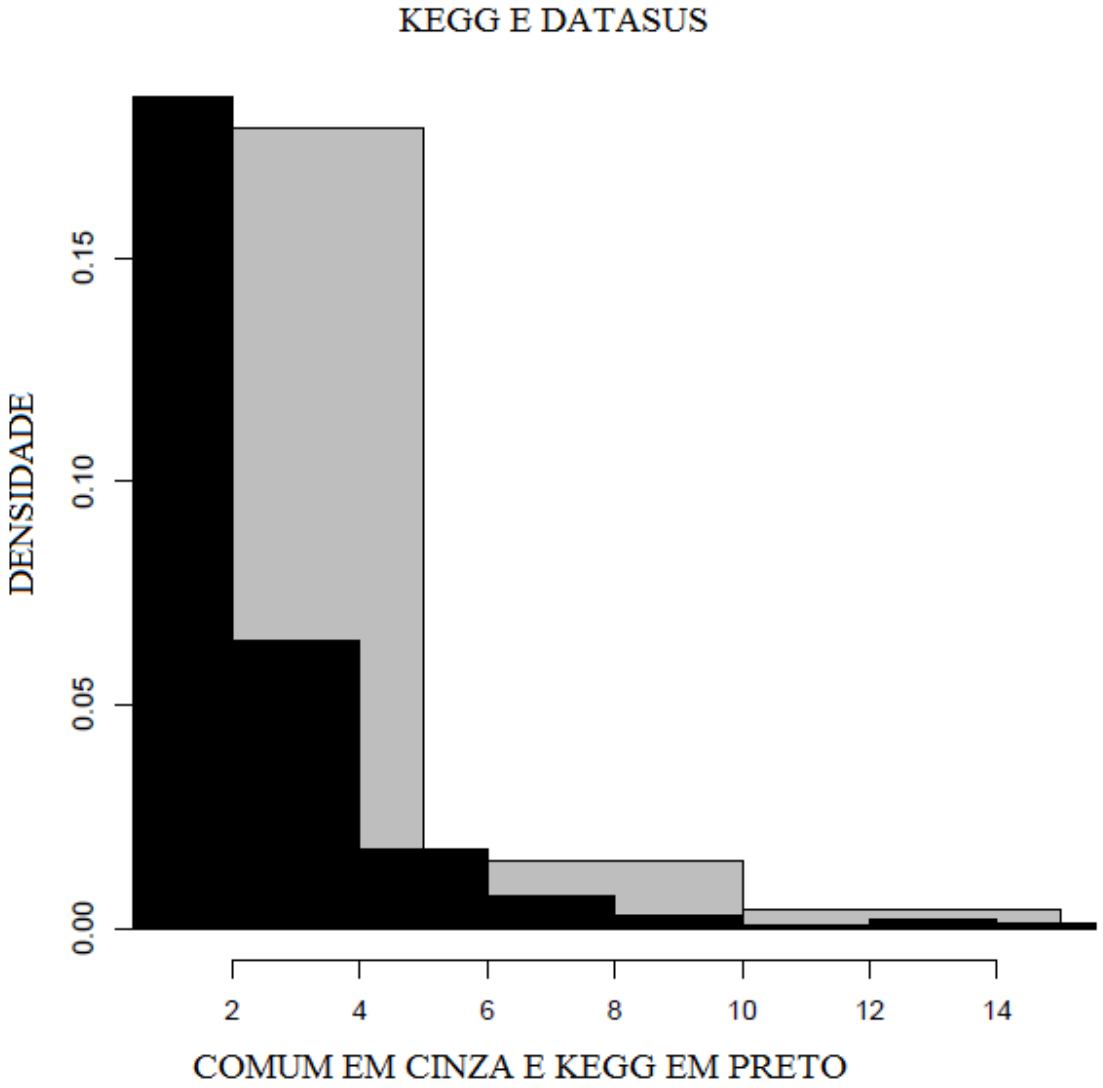
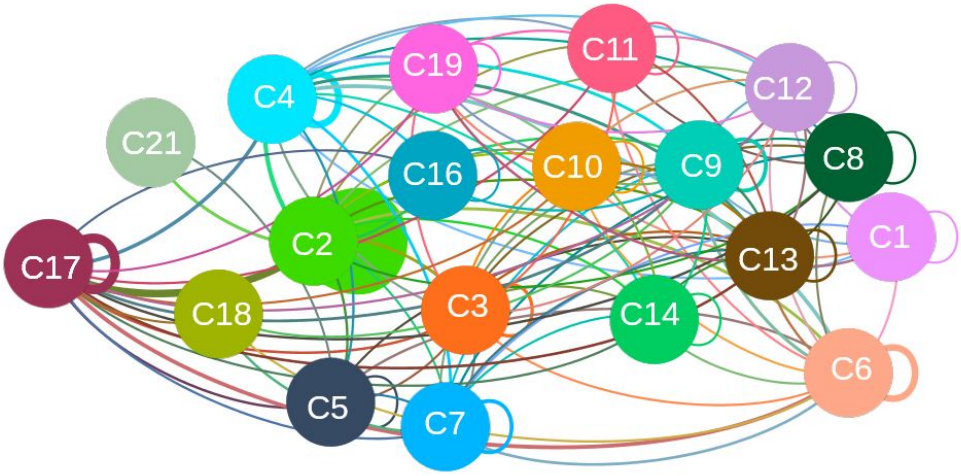


Figura 9 – Comparação de pesos entre rede gênica (KEGG) e a rede epidemiológica (DATASUS).

Para uma melhor visualização das ambas as redes foi criado dois hipergrafos. A Figura 10 representa o hipergrafo da $rede\_KEGG$ que foi criado através da $rede\_KEGG$ somando os pesos das arestas que pertenciam ao mesmo capítulo. Observamos através dos pesos das arestas do hipergrafo da Figura 11 as seguintes ligações $C2 - C2$ onde é apresentado uma forte ligação no auto-loop entre Neoplasias, em seguida temos o segundo auto-loop representado pelo capítulo de Malformações congênitas $C17 - C17$. Outras arestas que merecem destaques são: $C2 - C17$ – Neoplasias e Malformações Congênitas, $C4 - C4$ – Doenças endócrinas, $C6 - C6$ – Doenças do sistema nervoso, $C2 - C4$ – Neoplasia e Doenças endócrinas e $C6 - C17$ – Doenças do sistema nervoso e Malformações congênitas. Essas ligações apresentam uma forte conexão.



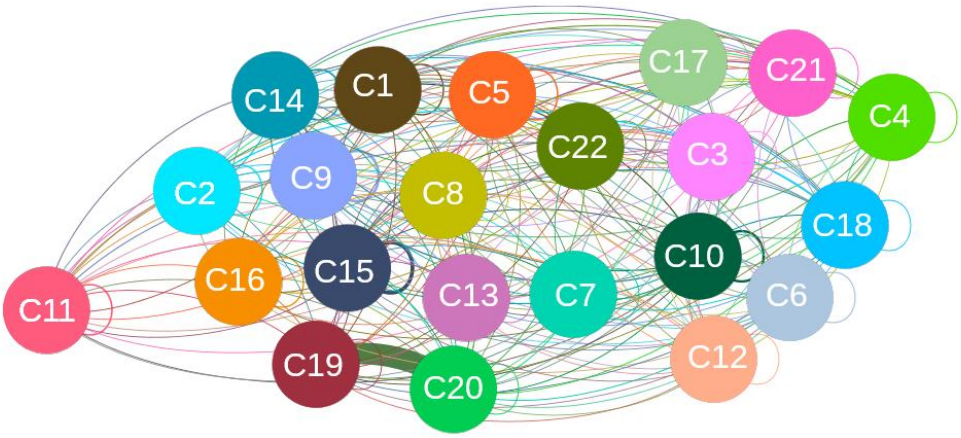
C1 - Algumas doenças infecciosas e parasitárias
C2 - Neoplasias
C3 - Doenças do sangue e dos órgãos hematopoéticos
C4 - Doenças endócrinas
C5 - Transtornos mentais e comportamentais
C6 - Doenças do sistema nervoso
C7 - Doenças do olho e anexos
C8 - Doenças do ouvido e da apófise mastóide
C9 - Doenças do aparelho circulatório
C10 - Doenças do aparelho respiratório
C11 - Doenças do aparelho digestivo
C12 - Doenças da pele e do tecido subcutâneo
C13 - Doenças do sistema osteomuscular
C14 - Doenças do aparelho geniturinário
C16 - Algumas afecções originadas no período perinatal
C17 - Malformações congênitas
C18 - Sintomas
C19 - Lesões
C21 - Fatores que influenciam o estado de saúde

Figura 10 – Hipergrafo da Rede KEGG.

A Figura 11 representa o hipergrafo da $rede\_DATASUS$ que apresenta fortes ligações entre $C19-C20$ – Lesões e Causas externas de morbidade e mortalidade, $C15-C25$ – Gravidez, $C9-C9$ – Doenças do aparelho circulatório, $C10-C10$ – Doenças do aparelho respiratório. A $rede\_DATASUS$ apresenta maiores ligações com ligações de auto-loop entre os capítulos, isso deve ser pelo fato de ser doenças com mesma etimologia.

Buscando encontrar uma correlação entre o hipergrafo da Figura 10 e o hipergrafo da Figura 11, resolvemos procurar associações dos 4 capítulos que possuem maiores ligações na $rede\_KEGG$ também na $rede\_DATASUS$. Os 4 capítulos são: Neoplasia (capítulo 2), Malformações congênitas (Capítulo 17), Doenças endócrinas Capítulo (4) e Doenças do sistema nervoso (Capítulo 6).

Após a realização da busca observamos que as associações entre os 4 capítulos da rede KEGG também acontecem no hipergrafo da rede DATASUS, apesar dos pesos das arestas serem diferentes. Com base nessas associações podemos extrair correlações entre doenças que compartilham genes as quais têm uma grande chance de serem comorbidades na rede DATASUS. A Figura 12 mostra os dois hipergrafos e a relação dos 4 capítulos e seus pesos. Outra informação importante que podemos extrair destes hipergrafos é a relação entre doenças endócrinas e neoplasias que deverá ser abordada detalhadamente em trabalhos futuros.



C1 - Algumas doenças infecciosas e parasitárias
C2 - Neoplasias
C3 - Doenças do sangue e dos órgãos hematopoéticos
C4 - Doenças endócrinas
C5 - Transtornos mentais e comportamentais
C6 - Doenças do sistema nervoso
C7 - Doenças do olho e anexos
C8 - Doenças do ouvido e da apófise mastóide
C9 - Doenças do aparelho circulatório
C10 - Doenças do aparelho respiratório
C11 - Doenças do aparelho digestivo
C12 - Doenças da pele e do tecido subcutâneo
C13 - Doenças do sistema osteomuscular
C14 - Doenças do aparelho geniturinário
C15 - Gravidez, parto e puerpério
C16 - Algumas afecções originadas no período perinatal
C17 - Malformações congênitas
C18 - Sintomas
C19 - Lesões
C20 - Causas externas de morbidade e de mortalidade
C21 - Fatores que influenciam o estado de saúde
C22 - Códigos para propósitos especiais

Figura 11 – Hipergrafo da Rede DATASUS.

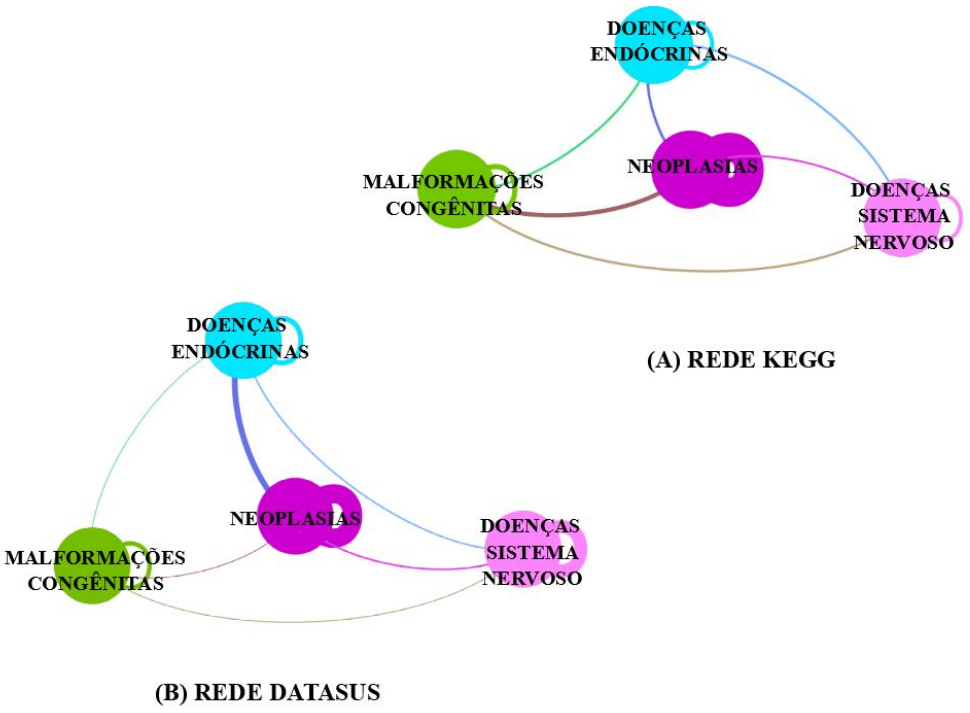


Figura 12 – A – Hipergrafo da Rede KEGG com 4 capítulos CID-10, e B – Hipergrafo da Rede DATASUS com 4 capítulos CID-10.

5.4 Predição e Análise

Para analisar o desempenho dos cinco índices Vizinhos Comuns (NC), Adar Adâmico (AA), Índice de Alocação de Recursos (RA), Índice de Conexão Preferencial (PA) e *Katz* apresentados na Seção 3.5, foi realizado um experimento apresentado a seguir.

A predição foi realizada na *rede_KEGG* sem auto-loops e para isso dividimos a rede em dois conjuntos de forma aleatória e independentes: o conjunto de treinamento e o conjunto de teste. O conjunto de treinamento possui 90% das arestas e o conjunto de teste 10% das arestas. Esse processo foi repetido por 1.000 interações. Para cada interação foi calculado as medidas AUC e Precisão de cada algoritmo. A média dessas interações é apresentado na Tabela 2.

Tabela 2 – Resultados da Precisão e AUC.

Algoritmo	Média AUC	Média Precisão
CN	0,769	0,807
AA	0,797	0,817
RA	0,810	0,778
PA	0,817	0,366
KATZ	**0,877**	**0,881**

Após a escolha do algoritmo *Katz*, aplicamos o método na *rede_KEGG* com intuito de reduzir falsos positivos utilizamos os scores atribuídos a cada aresta para fazer uma seleção preliminar deles. Se assumirmos que as arestas ausentes estão entre aquelas com maiores pesos, então a proporção de arestas que precisamos encontrar é muito menor do que a que começamos com 39.421. Por uma questão de concruente, nós escolhemos os maiores scores para estar entre os top 10% de todas as arestas preditas pelo *Katz*.

No total foram selecionadas 3.941 arestas e com essa lista de arestas realizamos uma busca na *rede_Datasus* e para simplificar nosso procedimento de avaliação, focamos em arestas que possuem pelo menos um nó dos quatros capítulos com as doenças mais citadas na *rede_KEGG*, que são: Capítulo 6 – Doenças do sistema nervoso; Capítulo 4 – Doenças endócrinas, Capítulo 2 – Neoplasias e; Capítulo 17 – Malformações congênita, selecionamos todas as arestas que continha pelo menos um nó envolvendo um dos quatro capítulo e isso totalizou as 1.400 das arestas previstas pelo algoritmo de *Katz* foram observadas neste subgrupo. Dentre elas selecionamos um ranking de comorbidades que possuem número de ocorrências maior que 100. No total foram selecionadas 35 pares de comorbidades apresentado na Tabela 3.

Para uma maior validação da metodologia, buscamos na literatura os pares de associações de doenças que fazem parte do ranking das trinta e cinco comorbidades

apresentada na Tabela 3. Onde foram encontrados as seguintes associações apresentadas abaixo.

A primeira associação apresentada é o par de doenças leiomioma do útero (D25) e endometriose (N80) que possui 4215 ocorrências verificadas na $rede\_DATASUS$. A endometriose é uma doença inflamatória dependente de estrogênio que afeta 5 a 10% das mulheres em idade reprodutiva nos Estados Unidos (BULUN, 2009). Enquanto o leiomioma uterino, também conhecido como tumores fibroides uterinos é uma condição comum em mulheres em idade reprodutiva. A sua taxa de incidência anual do Leiomioma é estimada em 9,2 casos por 1.000 em mulheres com idade entre 25 a 44 anos nos Estados Unidos, e de 12,7 casos em 100.000 mulheres com até 65 anos na Alemanha (SOLIMAN et al., 2015). No estudo de Gallagher et al. (2018), os autores identificaram 27 novos locos genômicos associados ao leiomioma e no qual quatro dos 17 locos identificados e replicados nessas análises também foram associados ao risco de endometriose. O trabalho de Gallagher et al. (2018) apresentou evidências dessa conexão entre essas duas doenças pela primeira vez.

Outro quatro pares de associação que mostra evidências em relatos clínicos têm relação com epilepsia. Segundo Neligan et al. (2011) pessoas com epilepsia têm um risco aumentado de morte. Uma das associações ligadas a epilepsia é a associação neoplasia maligna do encéfalo ($C71$) e epilepsia ($G40$) e epilepsia ($G40$) e neoplasia benigna do encéfalo ($D33$) que apresentou um número total de 845 e 136 ocorrências registrados na $rede\_DATASUS$, respectivamente, conforme a Tabela 3. Onde estudos mostram que pacientes com epilepsia possuem anormalidades estruturais no cérebro e revelam duas patologias principais: displasia cortical focal e tumores cerebrais (MORGAN; KERR, 2002; ARONICA; CRINO, 2014). As associações de epilepsia ($G40$) e neoplasia maligna de brônquios e pulmões ($C34$) com 107 ocorrências e epilepsia ($G40$) com outras doenças cerebrovasculares ($I67$) com 210 ocorrências registrados na rede_DATASUS foram relatados na literatura (MORGAN; KERR, 2002; NELIGAN et al., 2011).

No estudo de Onitilo et al. (2014) realizou-se uma análise temporal entre diabetes e risco de câncer de mama e concluíram que o risco de câncer de mama parece aumenta quando mulheres possem diabetes. E a associação diabetes mellitus tipo 2 ($E11$) e câncer de mama ($C50$) teve 227 ocorrências na $rede\_DATASUS$.

Assim, podemos inferir que as associações preditas a partir da $rede\_KEGG$ apresenta uma confiabilidade quando comparamos com dados clínicos previamente relatados, demonstrando que as predições utilizando uma rede gênica e uma rede epidemiológica são eficaz e efetiva.

Tabela 3 – Ranking de 35 comorbidades.

Doença	Doença	Número Ocorrências
Leiomioma do útero (D25)	Endometriose (N80)	4215
Esquizofrenia (F20)	Epilepsia (G40)	1893
Isquêmia do coração (I20)	Hipoparatireoidismo (E10)	1135
Epilepsia (G40)	Neoplasia maligna do encéfalo (C71)	845
Neoplasia maligna de outras localizações (C76)	História familiar de neoplasia maligna (Z80)	592
Neoplasia maligna da próstata (C61)	Neoplasia maligna da bexiga C67)	485
Neoplasia maligna do tecido conjuntivo (C49)	Outras neoplasias malignas da pele (C44)	459
Neoplasia maligna da próstata (C61)	Neoplasia maligna dos testículos (C62)	345
Neoplasia de comportamento incerto (D48)	Neoplasia maligna de outras localizações (C76)	329
Neoplasia maligna do estômago (C16)	Anemia por deficiência de ferro (D50)	292
Leucemia linfóide (C91)	Outras anemias aplásticas (D61)	284
Diabetes mellitus insulino-dependente (E10)	Outros transtornos da secreção pancreática (E16)	281
Leucemia mielóide (C92)	Outras anemias aplásticas (D61)	260
Neoplasia maligna da mama (C50)	Diabetes mellitus não-insulino-dependente (E11)	227
Epilepsia (G40)	Outras doenças cerebrovasculares (I67)	210
Neoplasia maligna de outras localizações (C76)	Neoplasia maligna do esôfago (C15)	205
Epilepsia (G40)	Outros transtornos da secreção pancreática (E16)	184
Neoplasia maligna de outras localizações (C76)	Neoplasia maligna da mama (C50)	181
Neoplasia maligna do intestino delgado (C17)	Neoplasia maligna do estômago (C16)	172
Neoplasia maligna do estômago (C16)	Neoplasia maligna de outras localizações (C76)	163
Epilepsia (G40)	Neoplasia benigna do encéfalo (D33)	136
Diabetes mellitus não-insulino-dependente (E11)	Epilepsia (G40)	133
Leucemia mielóide (C92)	Falha e rejeição de órgãos (T86)	132
Neoplasia maligna da mama (C50)	Epilepsia (G40)	126
Neoplasia maligna do tecido conjuntivo (C49)	Melanoma maligno da pele (C43)	125
Leucemia linfóide (C91)	Outros defeitos da coagulação (D68)	122
Diabetes mellitus não-insulino-dependente (E11)	Neoplasia maligna da próstata (C61)	115
Outras neoplasias benignas da pele (D23)	Outras neoplasias malignas da pele (C44)	114
Neoplasia maligna de outras localizações (C76)	Neoplasia maligna do colo do útero (C53)	113
Epilepsia (G40)	Facomatoses não classificadas (Q85)	112
Epilepsia (G40)	Neoplasia maligna dos brônquios e dos pulmões (C34)	107
Neoplasia maligna da mama (C50)	Outros distúrbios metabólicos (E88)	105
Neoplasia maligna dos brônquios e dos pulmões (C34)	Distúrbios do metabolismo de minerais (E83)	103
Neoplasia benigna do encéfalo (D33)	Neoplasia benigna das meninges (D32)	102
Púrpura e outras afecções hemorrágicaS (D69)	Linfoma não-Hodgkin difuso (C83)	102

6

Conclusão

Este capítulo apresenta as considerações finais do projeto, as principais contribuições, limitações da pesquisa e sugestões para trabalhos futuros.

6.1 Considerações Finais

A partir do levantamento bibliográfico realizado, há razões para acreditar que muitas associações de doenças ainda não foram descobertas. Encontrar comorbidades é um desafio enorme e com base nesse desafio que neste trabalho foi proposta uma metodologia para predição de associações de doença-doença através da integração de dados públicos sobre genes, doenças e suas comorbidades. Relembramos que o objetivo deste trabalho foi predizer associações entre doenças por intermédio do relacionamento entre uma rede gênica e uma rede epidemiológica. Utilizamos a rede gênica KEGG para inferir possíveis associações de doenças faltantes. Para realizar a predição utilizamos métodos de predição de links. Os pares de comorbidades preditas são comparadas com rede epidemiológica DATASUS. Dentre as contribuições do projeto destacam-se:

1. Construção das duas redes doenças: uma a partir dos dados de doenças e genes, a rede gênica KEGG. E a rede epidemiológica DATASUS a partir dos dados epidemiológicos;

2. Padronização dos códigos da rede KEGG para códigos da rede DATASUS;

3. Construção dos hipergrafos baseados nas redes gênicas e de comorbidades;

4. Avaliação das associações de doenças com base na rede epidemiológica;

5. Predição das associações doença—doença.

6. Análise das associações de doenças com base na rede epidemiológica

Como um resultado relevante dessa dissertação temos a predição de 6 associações: leiomioma do útero e endometriose, epilepsia e neoplasia benigna do encéfalo, neoplasia maligna do encéfalo e epilepsia, epilepsia e neoplasia maligna de brônquios e pulmões, epilepsia e com outras doenças cerebrovasculares, diabetes mellitus tipo 2 e câncer de mama. Todas essas associações possuem relatos clínicos comprovando as comorbidades, as seis associações também foram encontradas na rede DATASUS dentre as associações podemos destacar leiomioma do útero e endometriose que possui 4.215 ocorrências na rede DATASUS. Esse resultado demonstra que a comparação entre as duas redes é de grande importância.

Conseguimos encontrar uma relação entre as duas redes através da construção dos hipergrafos, foi verificado 4 capítulos que possuem as mesmas ligações em ambas redes apenas com pesos diferentes. Os quatro capítulos relacionados são: Neoplasia, Doenças endócrinas, Doenças sistema nervoso e Malformações congênitas. Esses capítulos podem gerar uma nova rede para estudos tendo em vista que eles possuem ligações genéticas encontrada na rede KEGG e ocorrências na rede DATASUS. Outra informação relevante dessa pesquisa é os seis *hubs* relacionados a neoplasias encontrados na rede KEGG uma nova rede pode ser criada apenas com esses *hubs*.

O método proposto confirma que a abordagem de predição de associações de doenças em rede gênica é efetiva e relevante para apontar possíveis comorbidades que tem seus genes desconhecido. Com base nessas associações podemos extrair correlações entre doenças que compartilham genes as quais têm uma grande chance de serem comorbidades na rede DATASUS. Outra informação importante que podemos extrair dos hipergrafos é a relação entre doenças endócrinas e neoplasias que deverá ser abordada detalhadamente em trabalhos futuros.

Por meio do método proposto e dos resultados encontrados e dos estudos de validação, podemos concluir que a predição de comorbidades por meio de rede gênica pode ser relevante para encontrar associações de doença-doença.

Os resultados mostraram que as duas redes possui um nível encorajador de concordância entre elas, mesmo que a rede epidemiológica possua muitas comorbidades de origem não genética que a rede KEGG não pode inferir por ser estritamente genética.

Um desafio encontrado nesse trabalho foi da padronização das doenças do KEGG para código do CID-10 devido a generalização e o agrupamento que essa padronização e auto-relações causados.

Acreditamos que essa nossa metodologia proposta seja original e pode ser replicada com outros bancos de dados de saúde pública. Como um outro trabalho futuro, propomos integrar as vias metabólicas para encontrar quais genes que interligam as associações doenças preditas. Outra opção seria analisar as comorbidades relacionadas a hipertensão

que foi o segundo maior *hub* encontrado na rede DATASUS. Uma outra abordagem seria fazer uma nova rede de comorbidades mas incluindo outros dados das AIHs, como por exemplo: idade e sexo.

Como trabalho futuro pretendemos desenvolver uma metodologia para predição da relação gene-doença através da integração de três tipos de informação: sobre os genes, sobre as doenças e suas comorbidades, e, finalmente, sobre as vias metabólicas. Esperamos que através da integração destes dados possamos desvendar novos genes potencialmente associados a uma doença através das vias metabólicas compartilhadas por diferentes doenças.

Referências

ADAMIC, L. A.; ADAR, E. Friends and neighbors on the Web. *Social Networks*, v. 25, n. 3, p. 211–230, 2003. ISSN 03788733.

ARONICA, E.; CRINO, P. B. Epilepsy related to developmental tumors and malformations of cortical development. *Neurotherapeutics*, v. 11, n. 2, p. 251–268, 2014. ISSN 1878-7479. Disponível em: <https://doi.org/10.1007/s13311-013-0251-0>.

BARABÁSI, A. L. Network medicine — from obesity to the " diseasome". *N Engl J Med*, Massachusetts Medical Society, v. 357, n. 4, p. 404–407, 2007. ISSN 1533-4406. Disponível em: <http://dx.doi.org/10.1056/nejme078114>.

BARABÁSI, A. L.; GULBAHCE, N.; LOSCALZO, J. Network Medicine: A Network-based Approach to Human Disease. *Nature Reviews Genetics*, v. 12, n. 1, p. 56–68, 2011.

BARABÁSI, A. L. et al. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, v. 311, p. 590–614, 2002. ISSN 03784371.

BARRENAS, F. et al. Network Properties of Complex Human Disease Genes Identified through Genome-Wide Association Studies. *PLoS ONE*, v. 4, n. 11, p. 2–7, 2009. ISSN 19326203.

BARZEL, B.; BARABÁSI, A. L. Network link prediction by global silencing of indirect correlations. *Nature Biotechnology*, v. 31, n. 8, p. 720–725, 2013.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*, 2009. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.

BULUN, S. E. Endometriosis. *New England Journal of Medicine*, v. 360, n. 3, p. 268–279, 2009. PMID: 19144942. Disponível em: <https://doi.org/10.1056/NEJMra0804690>.

CAMPOS, M. R. et al. Proposta de Integração de Dados do Sistema de Informações Hospitalares do Sistema Único de Saúde (SIH-SUS) para Pesquisa. *Informe Epidemiológico do SUS*, v. 9, n. 1, p. 51–58, 2000.

CHMIEL, A.; KLIMEK, P.; THURNER, S. Spreading of diseases through comorbidity networks across life and gender Spreading of diseases through comorbidity networks across life and gender. *New Journal of Physics*, IOP Publishing, p. 115013, 2014.

COKELAER, T. et al. BioServices: A common Python package to access biological Web Services programmatically. *Bioinformatics*, v. 29, n. 24, p. 3241–3242, 2013. ISSN 14602059.

CRAMER, A. O. J. et al. Comorbidity: A network perspective. *Behavioral and Brain Sciences*, Cambridge University Press, v. 33, n. 2-3, p. 137–150, 2010.

DATASUS. 2018. Disponível em: <http://datasus.saude.gov.br/sistemas-e-aplicativos/hospitalares/sihsus>. Acesso em: 23 de outubro de 2018.

DUARTE, N.; BECKER, S. A. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, v. 104, n. 6, p. 1777–1782, 2007. ISSN 0027-8424. Disponível em: <http://www.pnas.org/content/104/6/1777.short>.

FURLONG, L. I. Human diseases through the lens of network biology. *Trends in Genetics*, Elsevier Ltd, v. 29, n. 3, p. 150–159, 2013. ISSN 01689525. Disponível em: <http://dx.doi.org/10.1016/j.tig.2012.11.004>.

GALLAGHER, C. S. et al. Genome-wide association analysis identifies 27 novel loci associated with uterine leiomyomata revealing common genetic origins with endometriosis. *bioRxiv*, Cold Spring Harbor Laboratory, 2018. Disponível em: <https://www.biorxiv.org/content/early/2018/05/18/324905>.

GARCIA-ALBORNOZ, M.; NIELSEN, J. Finding directionality and gene-disease predictions in disease associations. *BMC Systems Biology*, BMC Systems Biology, p. 1–8, 2015. ISSN 1752-0509. Disponível em: <http://bmcsystbiol.biomedcentral.com/articles/10.1186/s12918-015-0184-9>.

GIANNOULA, A. et al. comoRbidity: an R package for the systematic analysis of disease comorbidities. *Bioinformatics*, v. 34, n. 18, p. 3228–3230, 04 2018. ISSN 1367-4803. Disponível em: <https://dx.doi.org/10.1093/bioinformatics/bty315>.

GIJSEN, R. et al. Causes and consequences of comorbidity: A review. *Journal of Clinical Epidemiology*, v. 54, n. 7, p. 661–674, 2001. ISSN 08954356.

GOH, K. I.; CHOI, I. G. Exploring the human diseasome: The human disease network. *Briefings in Functional Genomics*, v. 11, n. 6, p. 533–542, 2012. ISSN 20412649.

GOH, K.-I. et al. The human disease network. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 104, n. 21, p. 8685–8690, 2007. ISSN 0027-8424. Disponível em: <https://www.pnas.org/content/104/21/8685>.

GROSDIDIER, S. et al. Network medicine analysis of COPD multimorbidities. *Respiratory Research*, v. 15, n. 1, 2014. ISSN 1465993X.

HAGBERG, A. A.; SCHULT, D. A.; SWART, P. J. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, p. 11–15, 2008. ISSN 1540-9295.

HE, F. et al. PCID: A novel approach for predicting disease comorbidity by integrating multi-scale data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, IEEE, v. 14, n. 3, p. 678–686, 2017. ISSN 15455963.

HIDALGO, C. A. et al. A Dynamic Network Approach for the Study of Human Phenotypes. *PLoS Computational Biology*, v. 5, n. 4, 2009. ISSN 1553734X.

IBÁÑEZ, K. et al. Molecular Evidence for the Inverse Comorbidity between Central Nervous System Disorders and Cancers Detected by Transcriptomic Meta-analyses. *PLoS Genetics*, v. 10, n. 2, p. 1–7, 2014. ISSN 15537390.

IDEKER, T.; KROGAN, N. J. Differential network biology. *Molecular Systems Biology*, Nature Publishing Group, v. 8, n. 565, p. 1–9, 2012. ISSN 1744-4292. Disponível em: <http://dx.doi.org/10.1038/msb.2011.99>.

KANEHISA, M.; GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, v. 28, n. 1, p. 27–30, 2000.

KANEHISA, M. et al. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, v. 38, n. SUPPL.1, p. 355–360, 2009. ISSN 03051048.

KANEHISA, M. et al. Data , information , knowledge and principle : back to metabolism in KEGG. v. 42, n. November 2013, p. 199–205, 2014.

KANEHISA, M. et al. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, v. 44, n. D1, p. D457–D462, 2016. ISSN 13624962.

KANN, M. G. Advances in translational bioinformatics: Computational approaches for the hunting of disease genes. *Briefings in Bioinformatics*, v. 11, n. 1, p. 96–110, 2009. ISSN 14675463.

KATZ, L. A new status index derived from sociometric analysis. *Psychometrika*, Springer, v. 18, n. 1, p. 39–43, 1953.

KO, Y. et al. Identification of disease comorbidity through hidden molecular mechanisms. *Nature Publishing Group*, v. 6, n. December, p. 6–13, 2016. Disponível em: <http://dx.doi.org/10.1038/srep39433>.

LEE, D. S. et al. The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences of the United States of America*, v. 105, n. 29, p. 9880–9885, 2008. ISSN 1091-6490. Disponível em: <http://www.pnas.org/content/105/29/9880>.

LEE, I. et al. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, v. 21, n. 7, p. 1109–1121, 2011. ISSN 10889051.

LI, Y.; AGARWAL, P. A pathway-based view of human diseases and disease relationships. *PLoS ONE*, v. 4, n. 2, p. e4346, 2009. ISSN 1932-6203.

LIBEN-NOWELL, D.; KLEINBERG, J. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, New York, NY, USA, v. 58, n. 7, p. 1019–1031, maio 2007. ISSN 1532-2882. Disponível em: <http://dx.doi.org/10.1002/asi.v58:7>.

LIN, D. et al. An integrative imputation method based on multi-omics datasets. *BMC Bioinformatics*, BMC Bioinformatics, v. 17, n. 1, p. 1–12, 2016. ISSN 14712105. Disponível em: <http://dx.doi.org/10.1186/s12859-016-1122-6>.

LIU, J. et al. Comorbidity Analysis According to Sex and Age in Hypertension Patients in China. *Internacional Journal of Medical Sciences*, v. 13(2), p. 99–107, 2016.

LOSCALZO, J.; KOHANE, I.; BARABÁSI, A. L. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Molecular systems biology*, v. 3, n. 124, p. 124, 2007. ISSN 1744-4292.

LOYOLA FILHO, A. I. et al. Causas de internações hospitalares entre idosos brasileiros no âmbito do Sistema Único de Saúde. *Epidemiologia e Serviços de Saúde 2004*, v. 13, n. 4, p. 229–238, 2004.

LU, L.; ZHOU, T. Link prediction in complex networks: A survey. *Physica A*, v. 390, n. 6, p. 11501170, 2010.

MENCHE, J. et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science*, v. 347, n. 6224, p. 1257601, 2015. ISSN 1095-9203. Disponível em: <http://www.sciencemag.org/cgi/doi/10.1126/science.1257601$\ delimiter"026E30F$npapers3://publication/doi/10.1126/science.1257601>.

MINISTÉRIO da SAÚDE. Manual Técnico e Operacional do Sistema de Informação Hospitalar. *Secretaria de Atenção à Saúde*, n. nível 3, p. 87, 2013.

MORGAN, C.; KERR, M. Epilepsy and mortality: A record linkage study in a u.k. population. *Epilepsia*, v. 43, n. 10, p. 1251–1255, 2002. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1528-1157.2002.38701.x>.

NELIGAN, A. et al. The long-term risk of premature mortality in people with epilepsy. *Brain*, v. 134, n. 2, p. 388–395, 01 2011. ISSN 0006-8950. Disponível em: <https://dx.doi.org/10.1093/brain/awq378>.

NEWMAN, M. E. Clustering and preferential attachment in growing networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, v. 64, n. 2, p. 4, 2001. ISSN 1063651X.

NEWMAN, M. E. J. The Structure and Function of Complex Networks. *SIAM Review*, v. 45, p. 167–256, jan. 2003.

NéTO, K. F. *Análise gênica de comorbidades a partir da integração de dados epidemiológicos.* Dissertação (Mestrado) — Bioinformática, Universidade de São Paulo, 2014. Disponível em: <http://www.teses.usp.br/teses/disponiveis/95/95131/tde-29012015-150351/>. Acesso em: 6 de junho de 2016.

ONITILO, A. A. et al. Breast cancer incidence before and after diagnosis of type 2 diabetes mellitus in women: increased risk in the prediabetes phase. *European Journal of Cancer Prevention*, v. 2, p. 76–83, 2014.

PAN, L. et al. Predicting missing links and identifying spurious links via likelihood analysis. *Scientific Reports*, Nature Publishing Group, v. 6, p. 1–10, 2016. ISSN 20452322. Disponível em: <http://dx.doi.org/10.1038/srep22955>.

PARK, J. et al. The impact of cellular networks on disease comorbidity. *Molecular Systems Biology*, v. 5, n. 262, p. 262, 2009. ISSN 1744-4292. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2683720{&}tool= pmcentrez{&}rendertype=ab>.

PARK, S. et al. Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. *Molecular Systems Biology*, Nature Publishing Group, v. 7, n. 494, p. 1–11, 2011. ISSN 17444292. Disponível em: <http://dx.doi.org/10.1038/msb.2011.29>.

PARREIRA, R. M. S.; LENEA, M. G. C. Doenças da pobreza , negligenciadas e emergentes. *Anais do Instituto de Higiene e Medicina Tropical: Edição Comemorativa*, v. 11, p. 42–43, 2012.

PORTELA, M. C. et al. Algoritmo para a composição de dados por internação a partir do sistema de informações hospitalares do sistema único de saúde (SIH / SUS) – Composição de dados por internação a partir do SIH / SUS. *Cad Saúde Pública*, v. 13, n. 4, p. 771–774, 1997.

RITCHIE, M. D. et al. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, Nature Publishing Group, v. 16, n. 2, p. 85–97, 2015. ISSN 1471-0056.

ROQUE, F. S. et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Computational Biology*, v. 7(8), n. 8, 2011. ISSN 1553734X.

SAIK, O. V. et al. Novel candidate genes important for asthma and hypertension comorbidity revealed from associative gene networks. *BMC Medical Genomics*, v. 11, n. S1, p. 15, 2018. ISSN 1755-8794. Disponível em: <https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-018-0331-4>.

SINGH-BLOM, U. M. et al. Prediction and Validation of Gene-Disease Associations Using Methods Inspired by Social Network Analyses. *PLoS ONE*, v. 8, n. 5, 2013. ISSN 19326203.

SOLIMAN, A. M. et al. The direct and indirect costs of uterine fibroid tumors: a systematic review of the literature between 2000 and 2013. *American Journal of Obstetrics and Gynecology*, v. 213, n. 2, p. 141 – 160, 2015. ISSN 0002-9378. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0002937815002355>.

TABARÉS-SEISDEDOS, R. et al. No paradox, no progress: Inverse cancer comorbidity in people with other complex diseases. *The Lancet Oncology*, v. 12, n. 6, p. 604–608, 2011. ISSN 14702045.

TIAN, Z. et al. Constructing an integrated gene similarity network for the identification of disease genes. *Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016*, v. 8, n. Suppl 1, p. 1663–1668, 2017. ISSN 20411480.

VALDERAS, J. M. et al. Defining comorbidity: implications for understanding health and health services. *Annals of Family Medicine*, v. 7, p. 357–363, 2009. ISSN 1544-1709.

VALVERDE-REBAZA, J.; LOPES, A. A. Link prediction in Online Social Networks Using Group Information. *Iccsa*, p. 31–45, 2014.

VIDAL, M.; CUSICK, M. E.; BARABÁSI, A. L. Interactome networks and human disease. *Cell*, v. 144, n. 6, p. 986–98, 2011. ISSN 1097-4172. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3102045{&}tool=pmcentrez{&}rendertype=ab>.

WANG, X.; GULBAHCE, N.; YU, H. Network-based methods for human disease gene prediction. *Briefings in Functional Genomics*, v. 10, n. 5, p. 280–293, 2011. ISSN 20412649.

WU, X. et al. Network-based global inference of human disease genes. *Molecular Systems Biology*, v. 4, n. 189, p. 189, 2008. ISSN 1744-4292. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424293{&}tool=pmcentrez{&}rendertype=ab>.

YANG, P. et al. Inferring Gene-Phenotype associations via global protein complex network propagation. *PLoS ONE*, v. 6, n. 7, 2011. ISSN 19326203.

ZANZONI, A.; CHAPPLE, C. E.; BRUN, C. Relationships between predicted moonlighting proteins, human diseases, and comorbidities from a network perspective. *Frontiers in Physiology*, v. 6, n. JUN, p. 1–8, 2015. ISSN 1664042X.

ZHAO, J. et al. Prediction of Links and Weights in Networks by Reliable Routes. *Scientific Reports*, Nature Publishing Group, v. 5, p. 1–15, 2015. ISSN 20452322. Disponível em: <http://dx.doi.org/10.1038/srep12261>.

ZHOU, T.; LÜ, L.; ZHANG, Y. C. Predicting missing links via local information. *European Physical Journal B*, v. 71, n. 4, p. 623–630, 2009. ISSN 14346028.

ZHOU, X. et al. Human symptoms–disease network. *Nature Communications*, v. 5, n. May, 2014. ISSN 2041-1723. Disponível em: <http://www.nature.com/doifinder/10.1038/ncomms5212>.

ZOU, Q. et al. Prediction of microRNA-disease associations based on social network analysis methods. *BioMed Research International*, v. 2015, p. 810514, 2015. ISSN 2314-6133.

ZOU, Q. et al. Similarity computation strategies in the microRNA-disease network: A survey. *Briefings in Functional Genomics*, v. 15, n. 1, p. 55–64, 2016. ISSN 20412657.

ZOU, Q. et al. Approaches for recognizing disease genes based on network. *BioMed Research International*, v. 2014, p. 25–29, 2014. ISSN 23146141.