UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ERIC AUGUSTO ITO

PLAWSS: POWER LAW SEMANTIC SIMILARITY METODOLOGIA *DATA-DRIVEN* BASEADA EM LEI DE POTÊNCIA PARA O CÁLCULO DE SIMILARIDADE SEMÂNTICA GO

DISSERTAÇÃO

CORNÉLIO PROCÓPIO
2020

ERIC AUGUSTO ITO

PLAWSS: POWER LAW SEMANTIC SIMILARITY DATA-DRIVEN METHODOLOGY BASED ON POWER LAW FOR THE CALCULATION OF SEMANTIC SIMILARITY

PLAWSS: Power Law Semantic Similarity metodologia data-driven baseada em lei de potência para o cálculo de similaridade semântica

Dissertação apresentada como requisito para obtenção do título de Mestre em Pós-Graduação em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador: Fabrício Martins Lopes.

CORNÉLIO PROCÓPIO

2020

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Cornélio Procópio

ERIC AUGUSTO ITO

PLAWSS: POWER LAW SEMANTIC SIMILARITY METODOLOGIA DATA-DRIVEN BASEADA EM LEI DE POTÊNCIA PARA O CÁLCULO DE SIMILARIDADE SEMÂNTICA GO

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Bioinformática.

Data de aprovação: 21 de Outubro de 2020

Prof Fabricio Martins Lopes, Doutorado - Universidade Tecnológica Federal do Paraná

Prof Andre Yoshiaki Kashiwabara, Doutorado - Universidade Tecnológica Federal do Paraná

Prof Fabio Fernandes Da Rocha Vicente, Doutorado - Universidade Tecnológica Federal do Paraná

Prof Mauro Antonio Alves Castro, Doutorado - Universidade Federal do Paraná (Ufpr)

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Cornélio Procópio
Diretoria de Pesquisa e Pós-Graduação
Programa de Pós-graduação Acadêmica em Bioinformática
Mestrado em Pós-graduação em Bioinformática

FOLHA DE APROVAÇÃO

PLAWSS: POWER LAW SEMANTIC SIMILARITY
METODOLOGIA *DATA-DRIVEN* BASEADA EM LEI DE POTÊNCIA PARA O CÁLCULO DE SIMILARIDADE SEMÂNTICA GO

por

ERIC AUGUSTO ITO

Esta Dissertação foi apresentada às 14:00 de 21 de Outubro de 2020 como requisito parcial para a obtenção do título de Mestre(a) em Pós-graduação em Bioinformática, na área de concentração em Bioinformática e na linha de pesquisa em biologia computacional, do Programa de Pós-Graduação em Pós-graduação em Bioinformática. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo citados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

Prof. Dr. Fabrício Martins Lopes
Orientador(a)

Prof(a). Dr(a). Andre Yoshiaki Kashiwabara
Universidade Tecnológica Federal do Paraná

Prof(a). Dr(a). Fabio Fernandes Da Rocha Vicente
Universidade Tecnológica Federal do Paraná

Prof(a). Dr(a). Mauro Antonio Alves Castro
Universidade Federal do Paraná

Prof(a). Dr(a). Alexandre Rossi Paschoal
Coordenador(a) do PGBIOINFO

A Folha de Aprovação assinada encontra-se no
Departamento de Registros Acadêmicos da UTFPR – Câmpus Cornélio Procópio

Dedico este trabalho a minha família, amigos e
ao meu orientador, os quais me deram
incentivos e ferramentas para finalizar este
trabalhos.

AGRADECIMENTOS

Este trabalho não poderia ser terminado sem a ajuda de diversas pessoas e/ou instituições às quais presto minha homenagem. Certamente esses parágrafos não irão atender a todas as pessoas que fizeram parte dessa importante fase de minha vida. Portanto, desde já peço desculpas àquelas que não estão presentes entre estas palavras, mas elas podem estar certas que fazem parte do meu pensamento e de minha gratidão.

A minha família, pelo carinho, incentivo e total apoio em todos os momentos da minha vida.

Ao meu orientador, que me mostrou os caminhos a serem seguidos e pela confiança depositada.

A todos os professores e colegas do departamento, que ajudaram de forma direta e indireta na conclusão deste trabalho.

Enfim, a todos os que de alguma forma contribuíram para a realização deste trabalho.

O sucesso é um professor perverso. Ele seduz as pessoas inteligentes e as faz pensar que jamais vão cair. (GATES, Bill).

RESUMO

ITO, Eric Augusto. **Metodologia *data-driven* baseada em lei de potência para o cálculo de similaridade semântica GO**. 2020. 69 f. Dissertação (Mestrado em Pós-graduação em Bioinformática) – Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2020.

Genes são amplamente estudados pela comunidade científica devido a sua importância em diversas pesquisas, muitas delas relacionadas a saúde. Por conta disto, muitos métodos foram desenvolvidos para calcular a Similaridade Semântica (SS) entre genes. A Similaridade Semântica tem sido usado em várias pesquisas como inferência e validação de redes, dobramento de proteínas, entre outras. Inicialmente proposto por Wang el al. (WANG *et al.*, 2007) e que foi incrementado na ferramenta GOGO (ZHAO; WANG, 2018), a metodologia apresentada por Wang e GOGO não se limita a usar *Information content* (IC) para calcular a similaridade semântica. Wang propôs um método híbrido que calcula a similaridade a partir da topologia do grafo acíclico direcionado GO. GOGO por sua vez propôs usar o número de termos filhos como substituto de IC, visto que o GOGO notou a correlação inversa entre IC e o número de termos filho, dessa forma mesmo sem usar o IC, GOGO consegue ter as vantagens de métodos baseados em IC junto com o método híbrido de Wang. Porém o GOGO propõe um método que depende de variáveis que não se ajustam aos dados de ontologias, por outro o Wang se limita a pesar os termos GO somente dependendo do tipo de ligação entre os termos GO. Este trabalho apresenta um novo método chamado de *Power LAW Semantic Similarity* (PLAWSS) para o cálculo da similaridade semântica em genes utilizando um modelo híbrido para calcular a similaridade semântica utilizando a Ontologia Gênica, o qual é *data-driven* se adaptando aos dados de ontologia utilizando lei de potência para pesar cada termo GO, e que em adição ao tipo de ligação, neste trabalho também é levado em consideração o número de filhos do ancestral para identificar a especificidade do termo GO. Seis conjuntos de dados compostos por vias metabólicas foram clusterizados utilizando as similaridades semânticas calculadas entre cada par de gene, os clusters formados a partir das funções moleculares e processos biológicos apresentaram os melhores resultados, sendo eles, 83,33% e 66,67% respectivamente, corroborando para a provação do método proposto.

Palavras-chave: similaridade semântica. bioinformática. ontologia. redes complexas. genes.

ABSTRACT

ITO, Eric Augusto. **Power law data-driven methodology for calculating semantic similarity GO**. 2020. 69 p. Dissertation (Master's Degree in Pós) – Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2020.

Genes are widely studied by the scientific community due to their importance in various researches, many of them related to health. Because of this, many methods were developed to calculate the SS between genes. Semantic Similarity has been used in several researches such as network inference and validation, protein folding, among others. Initially proposed by Wang el al. (WANG *et al.*, 2007) and which was added to the GOGO tool (ZHAO; WANG, 2018), the methodology presented by Wang and GOGO is not limited to using IC to calculate the semantic similarity. Wang proposed a hybrid method that calculates similarity from the topology of the directed acyclic graph GO. GOGO in turn proposed to use the number of child terms as a substitute for IC, since GOGO noticed the inverse correlation between IC and the number of child terms, so even without using the IC, GOGO manages to take advantage of methods based on IC along with Wang's hybrid method. However, GOGO proposes a method that depends on variables that do not fit the data of ontologies, on the other hand Wang is limited to weighing the terms GO only depending on the type of connection between the terms GO. This work presents a new method named PLAWSS for calculating semantic similarity in genes using a hybrid model to calculate semantic similarity using Genetic Ontology, which is *data-driven* adapting to ontology data using power law to weigh each GO term, and that in addition to the type of connection, this work also takes into account the number of children of the ancestor to identify the specificity of the GO term. Six data sets composed of metabolic pathways were clustered using the semantic similarities calculated between each pair of genes, the clusters formed from the molecular functions and biological processes showed the best results, being 83.33% and 66.67% respectively, corroborating for the testing of the proposed method.

Keywords: semantic similarity. bioinformatic. onthology. complex network. genes.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E ACRÔNIMOS

SIGLAS

A	Adenina
BPO	Processos biológicos
C	Citosina
CCO	Componente celular
DAG	Grafo Direcionado Acíclico
DNA	Ácido Desoxirribonucleico
G	Guanina
GO	Ontologia Gênica
IC	*Information content*
LCA	*Lowest Common Ancestor*
MFO	Função molecular
PLAWSS	*Power LAW Semantic Similarity*
RNA	Ácido Ribonucléico
SGD	*The Saccharomyces Genome Database*
SS	Similaridade Semântica
T	Timina
TAIR	*The Arabidopsis Information Resource*
TopoICSim	*Topological Information Content Similarity*
U	Uracila

SUMÁRIO

1 INTRODUÇÃO

Avanços na área da bioinformática permitiram a criação de novos métodos mais ágeis e precisos para a extração de informações moleculares de experimentos de *DNA microarrays* (SHALON *et al.*, 1996) e o RNA-seq (WANG *et al.*, 2009). Essa evolução trouxe consigo uma imensa geração de dados devido à diminuição de custos de experimentação e também o aumento da velocidade de processamento. Dado este avanço, surgiu o problema de como lidar com esse grande volume de dados, mesmo com a evolução do poder de processamento dos computadores ainda hoje é um desafio para pesquisadores extrair toda a informação contida nos dados. Portanto mesmo com o aumento da quantidade de dados, o conhecimento extraído não acompanhou este crescimento (BLAKE; BULT, 2006).

Como descrito no dogma central da biologia molecular, a partir do DNA são dadas instruções para a transcrição de RNA codificantes e não codificantes e a produção de proteínas, estas que estão presentes em quase todas as funções de uma célula, portanto é evidente que o estudo sobre o DNA é a chave para compreensão de como organismos vivos funcionam (ZAHA *et al.*, 2003). Como já é conhecido, há muitos genes presentes no DNA, estes genes são sequências de nucleotídeos que podem sinalizar a produção de uma proteína e controlar uma característica do indivíduo. Pela sua relevância em diversas áreas como, por exemplo em pesquisas relacionadas a saúde, muitos pesquisadores não medem esforços e dedicam suas vidas em investigar o DNA. Os genes possuem uma dinâmica chamada de regulação genética, esta ação permite que os genes através de suas proteínas interajam uns com outros a fim de inibir ou ativar um gene, a partir destas interações entre os genes é feito o controle genético do organismo. Dessa maneira uma doença ou uma alteração no meio ambiente externo do organismo pode alterar toda a regulação genética, o que pode vir a desencadear alterações nas produções de proteínas (ZAHA *et al.*, 2003).

Estas interações que regulam os genes funcionam como uma rede, na qual os genes podem ser representados pelos vértices e as regulações entre os genes pelas arestas (DÍAZ-MONTAÑA *et al.*, 2017). Muitos trabalhos focam em criar métodos para inferir redes de regulação genética. A inferência de redes gênicas se justifica pelo fato de que mesmo hoje nem todas as espécies possuem suas regulações entre os genes documentadas, o que é um obstáculo para inúmeros trabalhos (LOPES, 2011). Além desta justificativa, inferir redes visa também entender como a dinâmica de ativação e inibição entre os genes funcionam, algo de suma

importância como já mencionado. Contudo, métodos de inferência de redes precisam ser testados e para isso há alguns meios, como comparar com as redes *Gold Standards* as quais possuem a representação real das regulações presente no organismo, porém alguns poucos organismos modelos possuem a *Gold Standard* documentada, em outros casos em que não há a rede *Gold Standard* dos organismos é feito a validação da rede por alguma metodologia, tendo como exemplo o cálculo da similaridade semântica. Por sua vez pode ser de algumas formas, baseados em IC (JIANG; CONRATH, 1997), baseado em nós (RESNIK, 1999) ou arestas (RADA *et al.*, 1989), métodos híbridos (WANG *et al.*, 2007), entre outros.

A partir desta problematização surgiu a necessidade de métodos computacionais mais rápidos e com mais precisão para análise de diversos tipos de dados, como genomas (MCKENNA *et al.*, 2010), redes gênicas (LUSCOMBE *et al.*, 2004), dobramento de proteínas (GEORGE; HERINGA, 2002), sendo um deles o cálculo da similaridade funcional entre genes (DÍAZ-MONTAÑA *et al.*, 2017; ZHAO; WANG, 2018; YU *et al.*, 2010; EHSANI; DRABLØS, 2016), que compara genes com base em características as quais os genes possuem em comum, para então quantificar a sua similaridade. A similaridade semântica é um componente fundamental para pesquisas na área da bioinformática envolvendo diferentes metodologias como clusterização de genes (BRAMEIER; WIUF, 2007; CHO *et al.*, 2009; YANG *et al.*, 2007), predição das funções de proteínas (RADIVOJAC *et al.*, 2013; JIANG *et al.*, 2016) e validação das interações gene-gene (STELZL *et al.*, 2005; CAO; CHENG, 2015).

Com os esforços de pesquisadores, hoje há disponível uma coletânea de informação sobre os genes. Por conta dessa grande quantidade de informação foi criado um consórcio chamado de Ontologia Gênica (GO)(ASHBURNER *et al.*, 2000; CONSORTIUM, 2016). O GO tem o objetivo de padronizar as descrições das características dos genes, para cada característica de um gene há um termo GO relacionado. Os atributos de cada gene são divididos em três ontologias: Função molecular (MFO), Processos biológicos (BPO) e Componente celular (CCO). Pelo fato de que os termos GO especificam as funções dos genes é viável utilizá-los para comparar os genes semanticamente. Termos GO também são amplamente usados em muitas outras aplicações na área da bioinformática, incluindo análise funcional do gene de dados de DNA microarray (OVASKA, 2015), agregação de genes (MENG *et al.*, 2015), similaridade de doenças (MATHUR; DINAKARPANDIAN, 2012), predição e validação de interações proteína-proteína (WU *et al.*, 2006).

As abordagens para o cálculo da similaridade semântica baseada em arestas se funda-

mentam na contagem do número de arestas em um caminho específico entre dois termos, na maioria dos casos, uma função de distância é definida no menor caminho ou na média de todos os caminhos. Medidas baseadas em nó são baseadas no IC dos termos envolvidos. O valor do IC é uma medida de quão específico e informativo é um termo, como descrito no trabalho de Resnik et al.(RESNIK, 1995), conhecida como medida Resnik. A maioria das pesquisas baseadas em nó é derivado da medida de Resnik, que considera apenas o IC de um único ancestral comum e ignora as informações sobre caminhos em subgrafos compostas de ancestrais comuns e pares de termos GO. A partir desta problematização, métodos híbridos foram propostos para explicar os nós e arestas no subgrafo. Por exemplo, Wang et al.(WANG *et al.*, 2007) introduziram uma medida de similaridade combinando a estrutura do grafo GO com os valores IC, integrando a contribuição de todos os termos em um subgrafo GO, que inclui todos os antepassados. Mais recentemente GOGO(ZHAO; WANG, 2018) propôs uma melhoria no método de Wang para não calcular o IC e mesmo assim não perder as vantagens de utilizar o IC para calcular a similaridade funcional entre genes. Além desses, outro método que também propõe um método híbrido é o Jiang et al.(JIANG; CONRATH, 1997) o qual combina a abordagem baseada em arestas usando o IC como um fator de decisão.

Dado a problematização de análise de dados e a importância da similaridade semântica para a validação e inferência de redes gênicas, este trabalho visa utilizar a metodologia híbrida adotada no trabalho do Wang e GOGO que consideram a topologia do grafo para o cálculo da similaridade semântica e que também utilizam o IC. O GOGO propõe uma forma indireta de usar IC mostrando que o número de filhos é correlacionado com o IC, porém o método acaba dependendo de constantes para o cálculo, portanto ele não se ajusta adequadamente dado a Ontologia Gênica que tem constantes atualizações. Neste trabalho é visado um método *data-driven* o qual vai ponderar pesos para termos GO usando a lei de potência dado o número de filhos do termo GO antecessor e também do tipo de ligação que a representa.

2 JUSTIFICATIVAS

A descoberta do DNA deu início aos estudos para compreender o que muitos chamam de código da vida. Foram descobertas as estruturas dos genes, sequências de nucleotídeos que podem exercer uma determinada função no organismo. Muitas pesquisas abordam como calcular a similaridade semântica entre genes e termos GO, afim de clusterização de genes (BRAMEIER; WIUF, 2007; CHO *et al.*, 2009; YANG *et al.*, 2007), predição das funções de proteínas (RADIVOJAC *et al.*, 2013; JIANG *et al.*, 2016) e validação das interações gene-gene (STELZL *et al.*, 2005; CAO; CHENG, 2015). Para ajudar nessa tarefa há Ontologias Gênicas que constroem um grafo de anotações para os genes (ASHBURNER *et al.*, 2000; CONSORTIUM, 2016). Estas anotações são divididas em funções moleculares, processos biológicos e componentes celulares, as quais juntas descrevem as funcionalidades e características dos genes. O conjunto de dados das ontologias é algo que está em constante mudanças, por isso é importante que métodos de similaridade semântica sejam *data-driven* para que se adéquem a estas mudanças.

Medidas para cálculos da similaridade semântica são abordadas em muitos estudos. Esse procedimento visa calcular o grau de relação entre duas entidades. A similaridade semântica fornece a base para a comparação funcional entre genes, por isso tem sido amplamente utilizado em aplicações no campo da bioinformática, como predição de localização sub-nuclear de proteínas (LEI; DAI, 2006), predição de funções genéticas (TAO *et al.*, 2007), análises de agrupamentos de genes (WOLTING *et al.*, 2006; BOLSHAKOVA *et al.*, 2005), caracterização de vias reguladoras humanas (GUO *et al.*, 2006), inferência de redes gênicas relacionadas a doenças humanas (DÍAZ-MONTAÑA *et al.*, 2017), análise de genes causadores de doenças usando similaridade semântica entre termos da Ontologia Gênica (SCHLICKER *et al.*, 2010).

Métodos baseados em nós, como o Resnik, calculam o IC de dois termos GO para encontrar a frequência que um ancestral em comum entres os dois termos ocorre no conjunto de dados do GO, e assim selecionar o maior valor de IC para definir como o valor da similaridade semântica. Métodos como esse acabam não considerando a topologia do grafo. Em um grafo de ontologias, as raízes são termos mais genéricos que não carregam muita informação, pois não são específicos, ao contrario dos nós mais abaixo, perto das folhas, estes são extremamente específicos e tem um peso maior para diferenciar um termo GO de outro.

Métodos baseados em arestas, como o Jiang, se fundamentam na distância entres os

termos GO no grafo, porém estes métodos não consideram a força das ligações e o quanto aquela ligação é importante. Há também métodos chamados de híbridos, por exemplo o GOGO, que combinam métodos baseados em nós e arestas para considerar mais informações para calcular a similaridade, porém o GOGO depende de constantes para fazer o calculo da SS, o que pode trazer valores incorretos, visto que o conjunto de dados das ontologias está em constante mudança e a distribuição dos termos podem mudar.

Logo, esse trabalho apresenta uma nova metodologia que adiciona o número de filhos do ancestral para calcular o peso semântico de cada ligação entre os termos GO, além disso é proposto um método que se ajusta conforme o conjunto de dados de ontologias, utilizando a lei de potência. Dessa maneira o método se torna mais adequado e eficiente para o cálculo da similaridade semântica a partir da ontologia gênica (GO).

3 OBJETIVOS

3.1 OBJETIVOS GERAIS

Desenvolver uma nova metodologia híbrida para o cálculo da similaridade semântica, que seja *data-driven* que mensura a importância e a especificidade do termo GO com base na distribuição do número de filhos do termo ancestral, sendo mais adequado do ponto de vista de modelagem e mais eficiente em termos de assertividade.

3.2 OBJETIVOS ESPECÍFICOS

Para alcançar o objetivo principal serão considerados os seguintes objetivos específicos:

- Obter os conjuntos de dados das Ontologias Gênicas da *Arabidopsis thaliana* e *Saccharomyces cerevisiae*.
- Construir o conjunto de dados de genes e seus termos GO associados.
- Estruturar a Ontologia Gênica como um grafo.
- Investigar a distribuição de filhos no GO.
- Formular uma equação para ponderar a especificidade dos termos GO.
- Criar um pipeline para mensurar a similaridade semântica.
- Investigar métodos de clusterização
- Clusterizar os conjuntos de genes, assim como criar as suas matrizes de similaridade.
- Comparar os resultados com outros métodos, tais como: Jiang, Resnik, GOGO, Wang.

4 REVISÃO BIBLIOGRÁFICA

O objetivo desta seção é apresentar alguns conceitos importantes para o entendimento do trabalho. Na subseção 4.1 é exposto informações sobre a sequência de DNA, como ela funciona, sua importância para organismos e algumas características importantes para este trabalho em questão. Em seguida, na subseção 4.2 é descrito algumas características da Ontologia Gênica, bem como características, estatísticas, mantenedores, estrutura dos dados, entre outras. A subseção 4.3 apresenta resumidamente a teoria de redes complexas e as características da distribuição pela lei da potência, e também a importância das redes *scale-free*. Por último, na seção 4.4 serão tratados trabalhos que são relacionados e que foram utilizados como base para o desenvolvimento deste trabalho.

4.1 DNA

A partir de 1940 com as pesquisas realizadas nos fungos, foi-se descoberto as moléculas denominadas Ácido Desoxirribonucleico (DNA), um ácido nucleico presente em todos os organismo procariontes, eucariontes e em alguns vírus são biologicamente muito importantes para todos os organismos por determinar quais proteínas sintetizar e em quais quantidades (ZAHA *et al.*, 2003). O DNA carrega informações genéticas do indivíduo, sendo responsável por armazenar e transmitir o material genético através de uma sequência de nucleotídeos localizadas nos cromossomos, estas macromoléculas constituem o gene. Os cromossomos são o mais alto nível de condensação do DNA, a justificativa para a condensação do material genético é evitar erros e mutações nos descendentes ao longo dos processos de replicações do DNA (SNUSTAD; SIMMONS, 2012).

Há dois tipos de ácidos nucleicos, Ácido Desoxirribonucleico e o Ácido Ribonucléico (RNA), que como mencionado são formados por nucleotídeos, cada nucleotídeo é composto por um fosfato, um açúcar(pentose) e uma base nitrogenada(púrica ou pirimídica) unidos por ligações covalentes. Em ambos os tipos de ácidos nucleicos, há bases nitrogenadas chamadas de Adenina (A), Guanina (G) e Citosina (C), contudo no DNA é possível encontrar a base Timina (T), enquanto no RNA é encontrado a Uracila (U). Outra diferença entre os ácidos nucleicos é o tipo de açúcar, no DNA o açúcar presente é o desoxirribose, já no RNA é a ribose (ZAHA *et al.*, 2003). O DNA possui uma estrutura de dupla hélice como mostra a Figura 1, duas fitas

compostas por nucleotídeos se conectam por pontes de hidrogênios entre as bases nitrogenadas, Adenina sempre se liga com a Timina, já a Guanina com a Citosina.

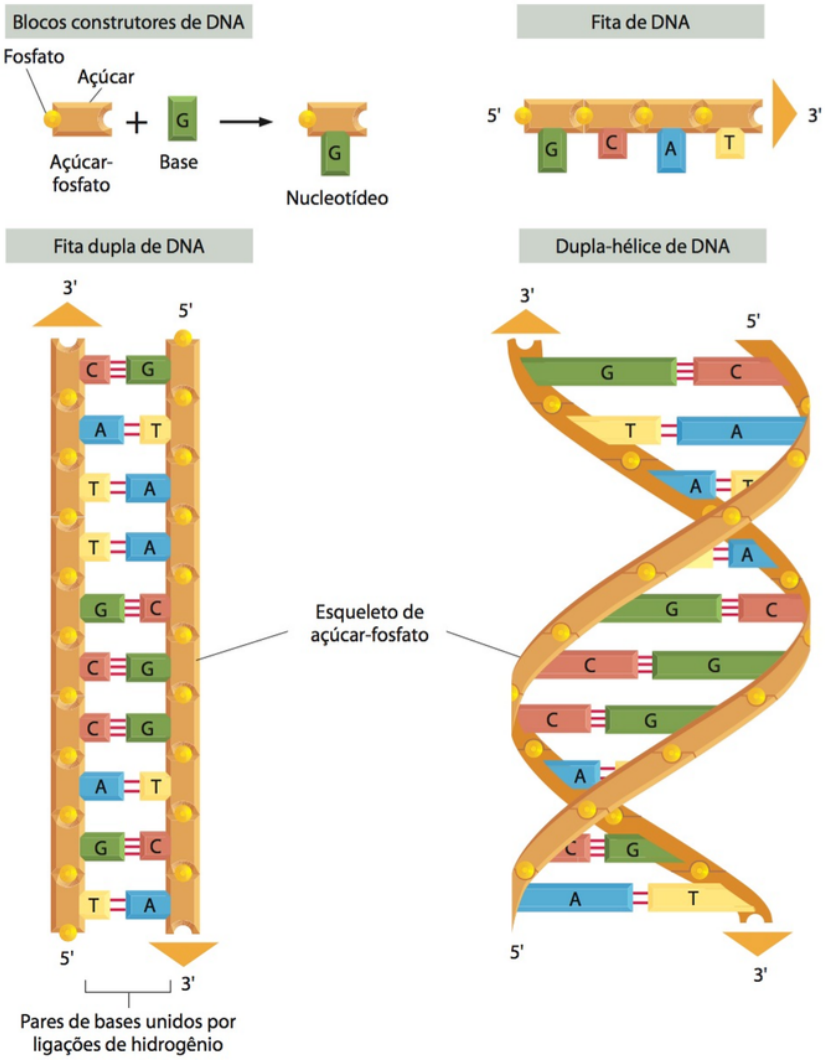


Figura 1 – Estrutura do DNA (LEWIS *et al.*, 2009).

A molécula de DNA passa por um processo chamado de transcrição para originar a molécula de RNA a partir de um gene localizado em uma posição do DNA chamada de *locus*. O RNA é formada por um único filamento de nucleotídeos (ZAHA *et al.*, 2003). A principal função do RNA é fazer a transferência da informação genética contida no DNA para os ribossomos, aonde ocorre a síntese de proteínas. A síntese de proteína se desenvolve a partir de um processo chamado de tradução, nesta etapa o RNA é lido em trincas(códons), dessa maneira cada trinca de nucleotídeos corresponde a um tipo de aminoácido que deve ser sintetizado, a sequência de aminoácidos forma uma proteína.

O fluxo de informação do material genético é explicado pelo dogma central da biologia molecular, assim como mostra a Figura2. Dessa maneira, todo produto genético se origina

do DNA, e cada produto pode estar ligados a funções moleculares, processos biológicos e componentes celulares distintos, que estão ligados a outras atividades e assim consecutivamente. E esta rede de atividades podem ser usadas para calcular a similaridade semântica entre genes.

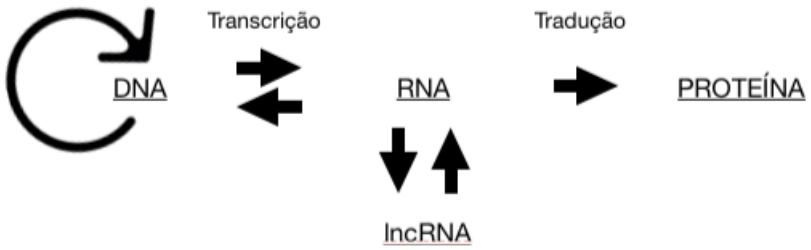


Figura 2 – Dogma central da biologia.

4.2 ONTOLOGIA GÊNICA

Consórcio fundado em 1998, Ontologia Gênica(GO) foi amplamente adotada pela comunidade científica como base para conhecer as funções genéticas, com o passar dos anos, aproximadamente duas décadas, houve várias melhorias tanto em quantidade como em qualidade, as descobertas conquistadas nessas ultimas décadas tornaram as anotações mais precisas, por muitas vezes foram reinterpretadas ou substituídas. Com o decorrer dos anos o seu time de pesquisadores sempre estiveram envolvidos em novas descobertas científicas e estão sempre buscando o estado mais atual do conhecimento biológico para fornecer dados precisos (ASHBURNER *et al.*, 2000; CONSORTIUM, 2016).

Manter consistência e a confiabilidade das anotações durante 20 anos não é fácil, ainda mais levando em consideração que uma enorme quantidade de dados sempre é adicionado e/ou modificada constantemente. Para enfrentar este desafio os bio-curadores da GO se reúnem periodicamente para estabelecimento de diretrizes de anotação e revisão coordenada de áreas específicas da biologia. Além disso as informações são tratadas tanto computacionalmente, para garantir que as anotações sejam válidas, como também manualmente, para assegurar que representam com precisão as descobertas experimentais. Foi descoberto pela equipe da GO que uma das abordagens mais poderosas ao controle e consistência da qualidade é a abordagem filogenética, pois genes que são de um mesmo filo, que possuem ancestral em comum, podem possuir funções parecidas, com essa informação é possível perceber discrepâncias nos dados (CARBON *et al.*, 2019).

Utilizando as abordagens mencionadas, foram reavaliados 2500 anotações manualmente, as quais 70% - 80% foram modificadas para um termo mais apropriado ou removidas. Para as principais espécies é previsto uma queda da adição de novas anotações, visto que os curadores estão focando esforços em rever e revisar anotações antigas (CARBON *et al.*, 2019).

Os termos GO foram estruturados de forma que seja possível realizar análises computacionais, algo indispensável para pesquisas biológicas modernas. A estrutura formada pela Ontologia Gênica define classes de funções genéticas e as correlacionam dados as suas relações, como é visto na Figura 3. A ontologia criada pelo consorcio foi construída de forma a abranger três aspectos distintos das funções dos genes. A primeira são as funções moleculares que são atividades bioquímicas de um produto gênico que podem ser realizadas por uma única máquina macromolécula. Estas funções descrevem apenas o que é feito sem especificar onde ou quando o evento realmente ocorre. Exemplos de termos funcionais são enzima, transportador ou ligante, adenilato ciclase.

A segunda são os processos biológicos que são uma série reconhecida de eventos ou funções moleculares com início e fim definidos, os quais referem-se a um objetivo biológico no qual o gene ou produto gênico contribui. Nesta categoria se encaixam processos que geralmente envolvem uma transformação química ou física, no sentido de que algo entra em um processo e algo diferente sai dele. Exemplos de termos de processos biológicos são crescimento e manutenção de células, tradução, metabolismo de pirimidina (ASHBURNER *et al.*, 2000).

A terceira é o componente celular que trata do local da ocorrência da máquina macromolecular quando desempenha uma função molecular. Exemplos de componentes celulares são ribossoma, proteassoma, membrana nuclear. Cada uma destas três categorias de anotação são estruturadas como na forma de grafo acíclico direcionado. Este tipo de grafo possui uma estrutura semelhante a uma árvore com um nó raiz único, os relacionamentos entre nós são dirigidos (orientado), e a estrutura não é recursiva, ou seja, sem ciclos.

Recentemente o consorcio de ontologias genéticas disponibilizou um grande atualização nos dados, atualmente há 45.000 termos ligados por aproximadamente 134.000 relações, dentre elas, 29.698 são de processos biológicos, 11.147 funções moleculares e 4.201 de componentes celulares(ASHBURNER *et al.*, 2000). Dentre as dezenas de informações contidas nas ontologias de cada termo, as mais importantes para caracterizar ligações entre GO termos são a *"is_a"* que caracteriza se um termo é um subtipo de outro, por exemplo o ciclo celular mitótico é um(*is_a*) ciclo celular, ligações do tipo *"is_a"* permitem fazer agrupamentos de anotações. Se um produto

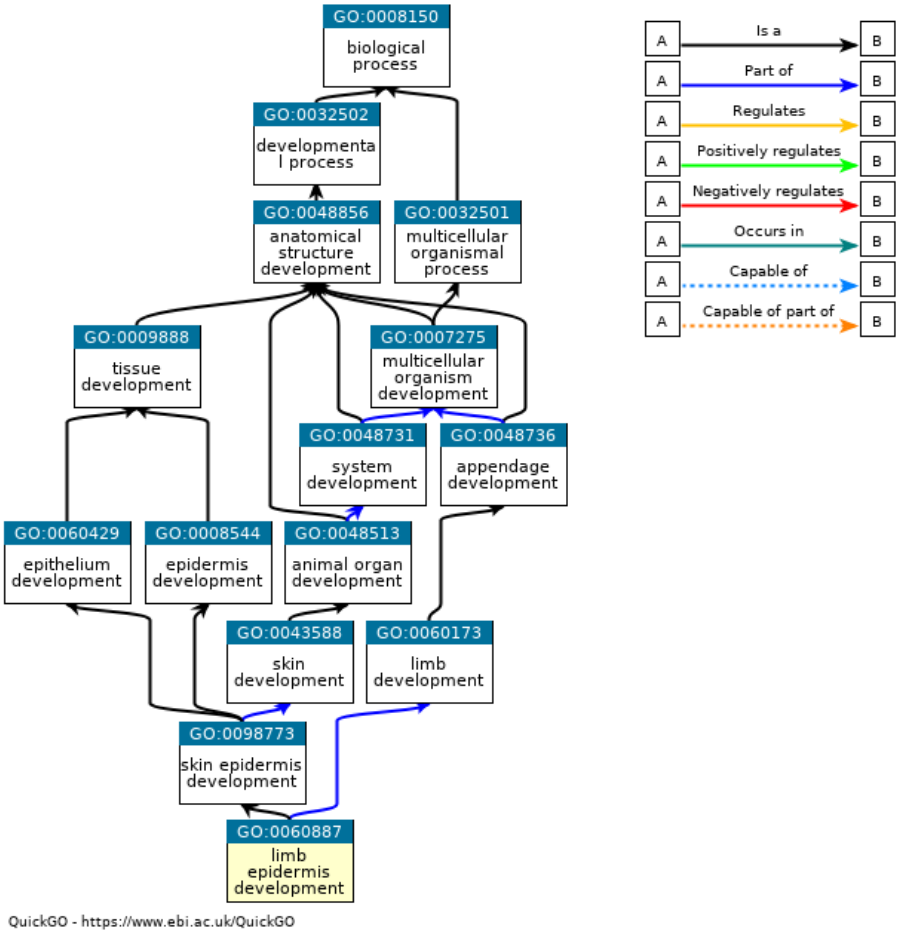


Figura 3 – Estrutura da Ontologia Gênica disponível em https://www.ebi.ac.uk/QuickGO/term/GO: 0060887.

genético *A* tem ligação com uma atividade *Y*, que por sua vez tem uma ligação do tipo *"is_a"* com outra atividade *K*, pode-se concluir que o produto genético *A* também possui a mesma atividade que *K*. A relação *"part_of"* define que dois termos estão tendo um relação, de maneira que se um termo *A* é parte(*part_of*) de *B*, *B* implica na presença de *A*, portanto *"part_of"* também consegue fazer agrupamentos, por exemplo se um produto genético A esta localizado em *X*, e *Y* é parte de *X*, consequentemente A esta localizado em *Y*. Outra ligação é a *"occurs_in"*, a qual denota a localização da ocorrência. As mais importantes são as *"is_a"* e *"part_of"*, por isso recebem um peso semântico maior nas modelagens computacionais, assim como a que estamos propondo aqui.

4.3 REDES COMPLEXAS

Redes complexas é um campo interdisciplinar que engloba áreas como teoria dos grafos e mecânica estatística, segundo Barabasi as redes complexas são grafos com topologias

não trivias, que possuem um conjunto de vértices(nós) que são ligados através de arestas (BARABÁSI, 2003). Há vários modelos teóricos de redes complexas descritas na literatura, entre as mais conhecidas estão as redes aleatórias (ERDÖS; RÉNYI, 1959), redes *small-world* (WATTS; STROGATZ, 1998) e redes *scale-free* (BARABÁSI; ALBERT, 1999). A teoria de redes complexas tem contribuído para estudos de diversas áreas, como a biológica, (BARABáSI *et al.*, 2011; NEWMAN, 2003; LIMA *et al.*, 2019; ITO *et al.*, 2018; BOCCALETTI *et al.*, 2006; COSTA *et al.*, 2007; VICENTE; LOPES, 2014), em particular as redes *scale-free* tem chamado atenção da literatura (SHIRAI *et al.*, 2020; ALBERT, 2005; TIMÁR *et al.*, 2016). Além disso, redes do tipo *scale-free* também tem se mostrado relevante para trabalhos biológicos de metabolismo, proteínas e redes de interações gênicas, mesmo considerando diferentes organismos (ALMAAS; BARABÁSI, 2006; RAVASZ *et al.*, 2002; BARABáSI, 2009; KHANIN; WIT, 2006; JEONG *et al.*, 2000; LOPES *et al.*, 2014).

Em particular, as redes *scale-free* receberam muita atenção na literatura por modelar problemas que possuem uma distribuição na forma de lei de potência na ligação entre os seus vértices. Estas ligações não aleatórias descrevem uma topologia, sendo possível abstrair medidas topológicas, tais como, desvio padrão do número de ligações, média do número de conexões por nó, número máximo e mínimo de conexões presentes na rede, coeficiente de cluster, assortatividade entre os vértices, motifs, entre outras. As redes complexas tem sido aplicadas na literatura com sucesso em modelagens de problemas reais, como pode ser visto na ferramenta BASINET(ITO *et al.*, 2018), o qual propôs aplicar medidas topológicas em redes complexas, formadas a partir de códons de RNA, visando classificar RNA codificantes e não codificantes, e alcançando ótimos resultados.

As redes chamadas de *scale-free*, foram propostas por Albert-László Barabási, que buscava uma representação melhor para redes reais ao invés de aceitar que as arestas são ligadas aleatoriamente. Diferentemente de outros tipos, redes *scale-free* não possuem números de nós fixos, pois há redes como a Web a qual cresce desenfreadamente, rede a qual começou como um único vértice. Além desta característica Barabási também estava preocupado em compreender os *hubs* nas redes, que são nós muito conectados. As redes aleatórias, como o proprio nome diz, criam conexões aleatoriamente entre os nós de forma uniforme, todos possuem a mesma chance de ganhar uma nova conexão. Esta teoria não é válida para redes reais. Nas redes *scale-free* há uma característica descrita como ordem preferencial, um exemplo seria a Web, sites que são mais conectados possuem preferência e acabam recebendo mais conexões ao invés de sites com

poucas conexões. A probabilidade de um nó da rede se conectar com outro nó pode ser descrita pela lei da potência Equação 1, onde *k* é o número de ligações de vértices e γ é a constante de decaimento.

$$P(k) \sim k^{-\gamma}. \tag{1}$$

4.4 SIMILARIDADE SEMÂNTICA

A similaridade semântica é uma abordagem utilizada para comparar objetos a partir de atributos que os mesmos compartilham, esta técnica é amplamente aplicada em muitas áreas de pesquisa, como psicologia, recuperação de informação, biomedicina e inteligência artificial (AKMAL *et al.*, 2014; GARLA; BRANDT, 2012). Na área da bioinformática é abordada na clusterização de genes (BRAMEIER; WIUF, 2007; CHO *et al.*, 2009; YANG *et al.*, 2007), predição das funções de proteínas (RADIVOJAC *et al.*, 2013; JIANG *et al.*, 2016) e validação das interações gene-gene (STELZL *et al.*, 2005; CAO; CHENG, 2015). Obter similaridades entre genes com precisão é o foco de muitos trabalhos acadêmicos, os quais na maioria utilizam como fonte de informação funcional a ontologia GO.

Nesta seção são resumidamente apresentados os principais métodos disponíveis da literatura atual que abordam a similaridade semântica, sendo alguns deles com o propósito de também validar e inferir redes gênicas.

4.4.1 GFD-NET

Devido a clara importância do estudo da genética para a evolução da ciências diversos trabalhos foram elaborados com a finalidade de validar redes regulatórias, dentre metodologias criadas esta a ferramenta GDF-NET(DÍAZ-MONTAÑA *et al.*, 2017), que além da validação de redes também realiza a inferência utilizando a Ontologia Gênica. Esse processo de validação tem como inovação o uso de informações topológicas da rede, além disso o artigo deixa claro que é o primeiro a validar redes utilizando o conceito de similaridade semântica para analisar redes de genes.

A metodologia criada no trabalho GFD-NET tem como base a premissa de que genes que se relacionam compartilham características similares. Para comparar a similaridade funcional entre os genes é construído uma árvore com base na Ontologia Gênica, o primeiro passo para

construir a árvore é atribuir como entrada os genes que serão analisados e verificar se estes genes são encontrados na lista de genes com suas anotações, caso não encontrado é realizado uma busca por sinônimos e assim é feita uma nova busca na lista de genes com as suas anotações, caso o gene não seja encontrado o gene é descartado. Após isto, o segundo passo é buscar pelos produtos dos genes, ou seja, proteínas.

A terceira etapa é a filtragem dos genes dado os três domínios da Ontologia Gênica. Esta etapa se faz necessária pois para cada domínio da ontologia sera feito os próximos dois passos de criação da árvore. O quarto passo é a busca pelas anotações GO relacionadas com as proteínas. Para o quinto e último passo é feita a construção de um grafo utilizando as relações *"is_a"*, *"part_of"* e *"occurs_in"* dos GO termos, em casos aonde os termos GOs dos genes possuem as mesmas relações, estes termos sofrem uma junção, com isso são interpretados como um único nó que representara estes genes, ao fim o grafo é transformado em uma estrutura de árvore e é adicionado as folhas da árvores os genes que estão relacionados. Todas as etapas citadas acima podem ser visualizadas na pipeline do GFD, Figura 4.

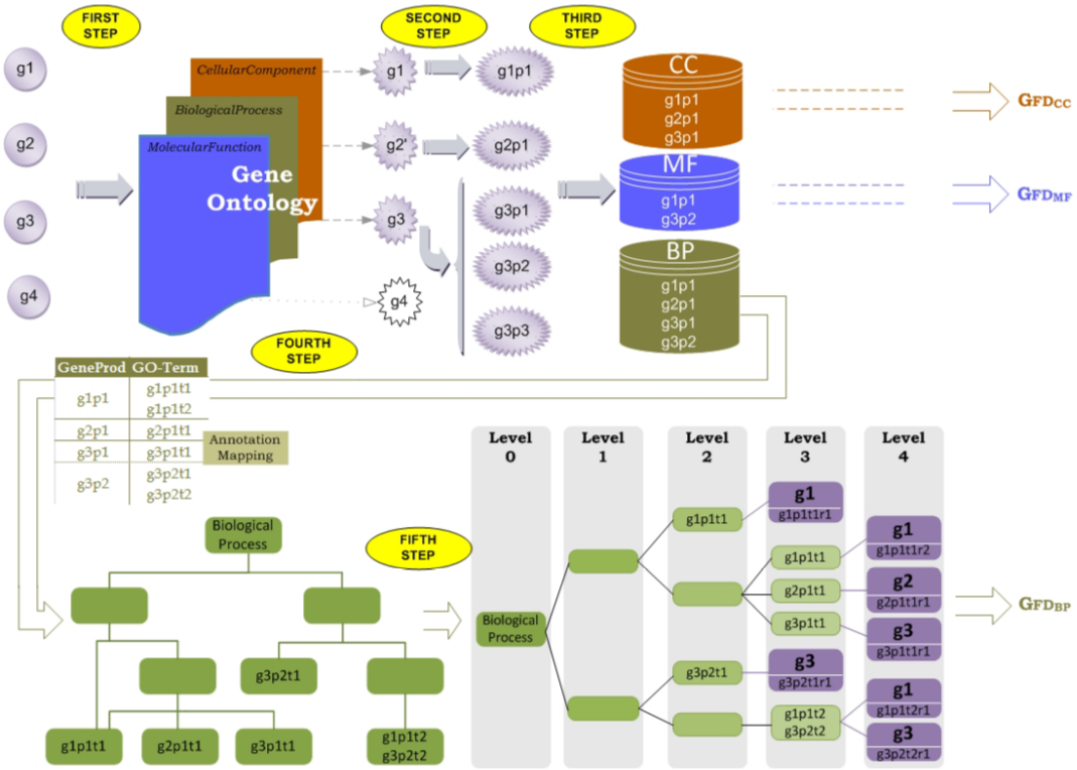


Figura 4 – Pipeline do GFD-NET(DÍAZ-DÍAZ; AGUILAR-RUIZ, 2011).

Com as árvores construídas com base na Ontologia Gênica, é possível calcular a similaridade gênica com bases na Equação 2, onde $distance(t_\alpha,t_\beta)$ é o número de nós que separa os termos GO t_α e t_β, $profundidade(LCA(t_\alpha,t_\beta))$ é a profundidade do *Lowest Common*

Ancestor (LCA) de ambos os termos GO analisados. Esta equação resulta em valores que variam de 0 a 1, valores próximos de 0 indicam similaridade e valores próximos de 1 indicam dissimilaridade. Após calcular a similaridade entre todos os genes é calculada a média dos valores de similaridade. Como saída o GFD-NET gera a rede de dissimilaridade e a rede enriquecida com as informações obtidas.

$$R(t_\alpha,t_\beta) = \frac{distancia(t_\alpha,t_\beta)}{2 * profundidade(LCA(t_\alpha,t_\beta)) + distancia(t_\alpha,t_\beta)} \quad (2)$$

4.4.2 GOGO e Wang

Outro trabalho que é validar redes de genes é proposto por Zhao et. al (ZHAO; WANG, 2018), chamado de GOGO. Esse é baseado em outra metodologia proposto por Wang et al. (WANG *et al.*, 2007), frequentemente utilizado como base para muitos trabalhos que calculam a similaridade semântica com termos GO. A abordagem adotada pelo GOGO e Wang se baseia na Ontologia Gênica para calcular a similaridade genética entre um par de genes, este cálculo é feito com base na similaridade entre os termos GO.

O primeiro passo é a formação de um Grafo Direcionado Acíclico (DAG), descrito na Seção 4.2, a partir das relações *"is_a"* e *"part_of"* dos termos GO. O nó folha do grafo representa o termo GO que será comparado, os demais nós são os termos GO ancestrais até chegar no nó raiz que representa a classe do termo GO, seja ela, componente celular, função molecular ou processo biológico. Uma outra DAG também será gerada para o outro termo GO, visto que o cálculo de similaridade de termos GO acontece aos pares. Com os grafos montados é dado início ao processo de calcular a similaridade entre dois termos GO. Em ambos os métodos é feito o cálculo do *S-value* para cada nó da DAG, esta variável quantifica a contribuição do termo *t* para a semântica do termo A. O cálculo do *S-value* é feito em duas etapas tanto no GOGO quanto no Wang, primeiro é necessário encontrar o peso semântico e em seguida usar o peso encontrado de cada gene para descobrir o *S-value*.

Na metodologia adotada por Wang o peso semântico (w_e) dado para as ligações depende somente do tipo de ligação entre o termo *t* e o ancestral, caso a ligação seja *"is_a"* o valor adotado foi de 0,8; para ligações do tipo *"part_of"* o valor foi de 0,6. Estes valores foram encontrados após testes que o Wang realizou com todas as via metabólicas resgatadas do *The Saccharomyces Genome Database* (SGD), foram clusterizados os genes variando o peso entre 0,5 e 0,9 a fim de obter qual valor resultava em uma melhora clusterização. Para encontrar o valor para o

peso do *"is_a"* foi considerado funções moleculares, visto que o *"is_a"* não esta presente em processos biológicos ou componentes celulares. No caso do *"part_of"* estes foram testados com os processos biológicos ou componentes celulares onde há várias dessas ligações. O melhor resultado para o *"is_a"* foi de 0,8; para *"part_of"* a conclusão foi que o peso atribuído deve ser 0,6 ou 0,7 (WANG *et al.*, 2007).

Por outro lado, o método GOGO propôs mensurar o peso semântico(w_e) utilizando o número de filhos, de modo que a equação adotada pelo GOGO definida pela Equação (3), na qual o *nc(t)* representa o número de filhos do termo GO *t*. Já *c* e *d* são constantes de parâmetro, *d* varia dependendo do tipo de ligação: 0,4 para *"is_a"*; 0,3 para *"part_of"*. *c* é uma constante de valor 0,67. Este valor é uma constante que foi estipulado de forma arbitrária para que o peso varie de 0 a 1 (ZHAO; WANG, 2018).

$$w_e = \frac{1}{c + nc(t)} + d \tag{3}$$

Ambos os métodos utilizam a Equação (4) para definir o valor de *S-value*, onde *t* é o termo GO, *t'* é o termo GO filho, *A* é o termo GO alvo da comparação, ou seja, a raiz do grafo é *A*. Em seguida é realizado a somatória dos *S-values* de cada termo GO presente no grafo, como define a equação (5). Em seguida é usado a Equação (6) para encontrar o valor da similaridade entres os termos GO dividindo os *S-values* dos ancestrais em comum pela soma de todos os *S-values*, o objetivo é penalizar a similaridade de acordo com a diferença entre os caminhos percorridos de cada gene até o nó raiz, levando em consideração que podem existir caminhos diferentes a serem percorridos a partir dos termos até a raiz.

$$\begin{cases} S_A(t) = 1 & \text{if } t = A \\ S_A(t) = Max\{w_c \times S_A(t') | t' \in children(t)\} & \text{if } t \neq A \end{cases} \tag{4}$$

$$SV(A) = \sum_{t \in T_a} S_a(t) \tag{5}$$

$$S_{GO}(A, B) = \frac{\sum_{t \in T_A \cap T_B} (S_A(t) + S_B(t))}{SV(A) + SV(B)} \tag{6}$$

Para o cálculo da similaridade entre genes, para cada termo GO de um gene é encontrado a melhor combinação que maximiza o valor da similaridade, como define a Equação 7. Por exemplo considerando que um gene G_1 possui dois termos GO e um gene G_2 que possui três termos GO, são gerados 5 resultados de similaridade, uma similaridade para cada um dos termos

GO. A Equação (8), soma apenas as melhores combinações de similaridade de cada termo GO e faz a média para encontrar o valor de similaridade entre os genes.

$$Sim(go, G_1) =_{1 \leq i \leq m}^{max} (S_{go}(go, go_{1i})) \tag{7}$$

$$Sim(G1, G2) = \frac{\sum_{1 \leq i \leq m} Sim(go_{1l}, G_2) + \sum_{1 \leq j \leq n} Sim(go_{2j}, G_1)}{m + n} \tag{8}$$

A Figura 5 exibe um exemplo de DAG, as etapas a seguir apresentam como é realizado o cálculo de similaridade pelo método do GOGO entre termos GO:0005975 e GO:1901135 e posteriormente entre genes.

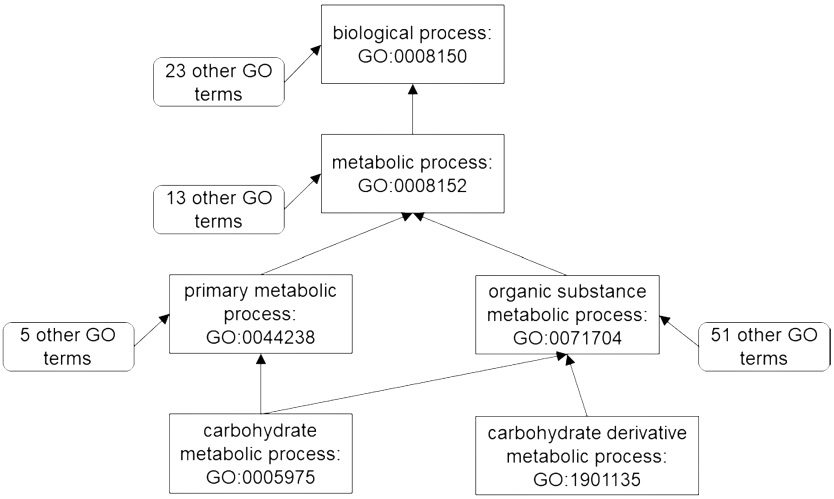


Figura 5 – Exemplo de DAG(ZHAO; WANG, 2018).

1. Cálculo do peso da contribuição semântica utilizando a Equação (3) para todos os ancestrais do termo GO. GOGO atribui um peso de 0,4 para ligações do tipo *"is_a"* e 0,3 para *"part_of"*. A Equação (9) calcula o peso para o termo GO: 0044328, que possui seis filhos e possui uma ligação do tipo *"is_a"* com o termo GO: 0008152. Os pesos para os demais termos GO da Figura 5 podem ser vistos nas Tabelas 1 e 2.

$$w_e = \frac{1}{0{,}67 + 6} + 0{,}4 = 0{,}55 \tag{9}$$

2. GOGO calcula um *S-value* para todos os termos GO presenta na DAG utilizando a Equação (4). A Equação (10) demonstra o cálculo do *S-value* para o GO:0008152 sabendo que o

S_A do GO:0044238 é 0,55; S_A de GO: 0071704 é 0,419 e o w_e de GO: 0008152 é 0,464. Os *S-value* para os demais termos GO da Figura 5 podem ser vistos nas Tabelas 1 e 2.

$$S_A(t) = max(0{,}464 \times 0{,}55; 0{,}464 \times 0{,}419) = 0{,}255 \tag{10}$$

3. Enfim para o cálculo da similaridade entre termos GO são utilizadas as Equações (5) e (6). A Equação (5) resulta no valor semântico do termo GO *A*, que é calculado pela soma de todos *S-value* da DAG do termo *A*. A Equação (6) é a soma dos *S-values* dos ancestrais comuns de *A* e *B* divido pela soma dos valores semânticos de *A* e *B*. As Equações (11) e (12) exemplificam o cálculo da similaridade semântica entre os termos GO:0005975 e GO:1901135 da Figura 5

$$S_{GO}(A, B) = \frac{(0{,}550 + 0{,}255 + 0{,}112) + (0{,}419 + 0{,}194 + 0{,}086)}{(1 + 0{,}55 + 0{,}419 + 0{,}255 + 0{,}112) + (1 + 0{,}419 + 0{,}194 + 0{,}086)} \tag{11}$$

$$S_{GO}(A, B) = 0{,}364 \tag{12}$$

4. A similaridade entre genes é feita utilizando o método *Average Best-Matches(ABM)*, o qual seleciona a melhor pontuação de similaridade entre um termo GO quando comparado com todos os termos GO de outro gene como define a Equação (7). Após isso é realizado a média da soma de todas as melhores combinações de similaridade entre os genes G_1 e G_2, como mostra a Equação (8).

Tabela 1 – Pesos e *S-value* calculados usando o GOGO para todos os termos GO descendentes do termo GO:0005975.

Go term	GO:0005975	GO:0044238	GO:0071704	GO:0008152	GO:0008150
w_e		1 / (0.67 + 6) + 0.4 = 0.55	1 / (0.67 + 53) + 0.4 = 0.419	1 / (0.67 + 15) + 0.4 = 0.464	1 / (0.67 + 24) + 0.4 = 0.441
S-value	1	0.55	0.419	0.255	0.112

Tabela 2 – Pesos e *S-value* calculados usando o GOGO para todos os termos GO descendentes do termo GO:1901135.

Go term	GO:1901135	GO:0071704	GO:0008152	GO:0008150
We		1 / (0.67 + 53) + 0.4 = 0.419	1 / (0.67 + 15) + 0.4 = 0.464	1 / (0.67 + 24) + 0.4 = 0.441
S-value	1	0.419	0.194	0.086

Usando também a Figura 5 como exemplo de DAG, as etapas a seguir apresentam como é feito o cálculo de similaridade pelo método do Wang entre termos GO:0005975 e GO:1901135 e posteriormente entre genes.

1. Como citado anteriormente, o método Wang considera somente o tipo de ligação entre os termos, o peso da contribuição semântica recebe 0,8 para "is_a"e 0,6 para "part_of". A Equação (13) calcula o peso para o termo GO: 0044328 sabendo que ele possui uma ligação do tipo "is_a"com o GO:0008152. Os pesos para os demais termos GO da Figura 5 podem ser vistos nas Tabelas 3 e 4.

$$w_e = 0{,}8 \tag{13}$$

2. Wang calcula um *S-value* para todos os termos GO presenta na DAG utilizando a Equação (4). A Equação (14) demonstra o cálculo do *S-value* para o GO:0008152 sabendo que o S_A do GO:0044238 é 0,8; S_A de GO: 0071704 é 0,8 e o w_e de GO: 0008152 é 0,8. Os *S-value* para os demais termos GO da Figura 5 podem ser vistos nas Tabelas 3 e 4

$$S_A(t) = max(0{,}8 \times 0{,}8; 0{,}8 \times 0{,}8) = 0{,}64 \tag{14}$$

3. Enfim para o cálculo da similaridade entre termos GO são utilizadas as Equações (5) e (6). A Equação (5) resulta no valor semântico do termo GO *A*, que é calculado pela soma de todos *S-value* da DAG do termo *A*. A Equação (6) é a soma dos *S-values* dos ancestrais comuns de *A* e *B* divido pela soma dos valores semânticos de *A* e *B*. A Equação (15) exemplifica o cálculo da similaridade entre os termos GO:0005975 e GO:1901135 da Figura 5

$$S_{GO}(A, B) = \frac{(0{,}8 + 0{,}64 + 0{,}512) + (0{,}8 + 0{,}64 + 0{,}512)}{(1 + 0{,}8 + 0{,}8 + 0{,}64 + 0{,}512) + (1 + 0{,}8 + 0{,}64 + 0{,}512)} = 0{,}582 \tag{15}$$

4. A similaridade entre genes é feita utilizando o método *Average Best-Matches(ABM)*, o qual seleciona a melhor pontuação de similaridade entre um termo GO quando comparado com todos os termos GO de outro gene como define a Equação (7). Após isso é realizada a média da soma de todas as melhores combinações de similaridade entre os genes G_1 e G_2, como define a Equação (8).

Tabela 3 – Pesos e *S-value* calculados usando o Wang para todos os termos GO descendentes do termo GO:0005975.

Go term	GO:0005975	GO:0044238	GO:0071704	GO:0008152	GO:0008150
w_e		0.8	0.8	0.8	0.8
S-value	1	0.8	0.8	0.64	0.512

Tabela 4 – Pesos e *S-value* calculados usando o Wang para todos os termos GO descendentes do termo GO:1901135.

Go term	GO:1901135	GO:0071704	GO:0008152	GO:0008150
w_e		0.8	0.8	0.8
S-value	1	0.8	0.64	0.512

4.4.3 GOSemSim

GOSemSim(YU *et al.*, 2010) é um pacote computacional, desenvolvido usando o projeto R que implementa a similaridade semântica entre termos GO, conjuntos de termos GO, produtos gênicos e clusters gênicos. Por ser um pacote R GOSemSim se destaca pela flexibilidade e também facilidade para integrar o pacote em outras pipelines de alto rendimento. GOSemSim utiliza como anotação o Ontologia Gênica para quantizar as semelhanças entre genes e grupos genéticos. A ferramenta GOSemSim implementa cinco métodos clássicos para o cálculo de similaridade, sendo eles, quatro métodos de conteúdo informático(IC) e um para análise topológica de um grafo.

Algumas das metodologias que abordam conteúdo informativo são: Resnik (RESNIK, 1999), Lin(LIN *et al.*, 1998), Jiang e Conrath (JIANG; CONRATH, 1997), Schlicker (SCHLICKER *et al.*, 2006). Como mencionado uma metodologia que analisa a semelhança semântica através da topologia da estrutura do grafo GO também foi introduzida no GOSemSim, esta analise considera que a especificidade de um termo GO é geralmente determinado por sua localização no gráfico GO, a metodologia utilizada GOSemSim foi apresentada por Wang (WANG *et al.*, 2007). Cada um desses métodos podem ser selecionados no pacote R para cálculo da similaridade semântica.

O IC é específico de cada espécie e é calculado a partir dos pacotes de anotação que o Bioconductor disponibiliza, chamado de AnnotationDbi (PAGèS *et al.*, 2019). AnnotationDbi é um pacote mantido pelo Bioconductor que contém anotações de genomas e genes que foram modificados, afim de torna-los mais acessíveis para os pesquisadores. Os conjuntos suportados pelo GOSemSim são: org.Hs.eg.db, org.Rn.eg.db, org.Mm.eg.db, org.Dm.eg.db e org.Sc.sgd.db para humanos, ratos, mouse, mosca e fermento, respectivamente.

4.4.4 TopoICSim

Topological Information Content Similarity (TopoICSim) tem como diferencial examinar todos os ancestrais comuns para um par de GO termos, e não apenas o último (ou mais profundo) antepassado em comum, como é o caso de outros métodos. TopoICSim possui uma metodologia que combina IC com topologia do grafo GO, como pode ser visto na metodologia abaixo. TopoICSim recebe um conjunto de anotações GO as quais descrevem duas entidade biológicas, e retorna um valor numérico que descreve o quão próximo funcionalmente estas duas entidades estão. A ferramenta TopoICSim foi implementada em R e está disponível no Bioconductor.

A metodologia do TopoICSim como dito acima se apoia no cálculo da IC com a topologia do grafo GO. A equação 21 calcula a distância entre dois termos, ela deve ser feita para cada ancestral em comum existente, e assim selecionado o menor valor. O numerador da Equação 21 é o menor caminho de cada gene com o seu antecessor comum, esse cálculo de menor caminho e realizado com base nas Equações 18 e 19, aonde P da Equação 16 e 17 é o caminho entre o gene e o antecessor, t da Equação 18 é o termo GO. Já o denominador da Equação 21 é a maior distância do antecessor com a raiz, calculado com as Equações 17 e 18.

$$SP(t_i, t_j) = argmin IIC(P) \tag{16}$$

$$LP(t_i, t_j) = argmax IIC(P) \tag{17}$$

$$IIC(P) = \sum_{a}^{t \in P} \frac{1}{IC(t)} \tag{18}$$

$$wSP(t_i, t_j) = SP(t_i, t_j) \times len(P) \tag{19}$$

$$wLP(t_i, t_j) = LP(t_i, t_j) \times len(P) \tag{20}$$

$$D(t_i, t_j, x) = \frac{wSP(t_i, t_j, x)}{wLP(x, root)} \tag{21}$$

Com o cálculo de da distância feita para todos os caminhos dado os ancestrais comuns, é escolhido o menor. Com isso é obtido o valor da similaridade entre os dois genes utilizando a Equação 22.

$$S(t_i, t_j) = 1 - \frac{Arctan(D(t_i, t_j))}{\frac{\pi}{2}} \tag{22}$$

4.4.5 Resnik

Em seu trabalho Resnik cria um método que utiliza *information-content* para calcular a similaridade semântica de maneira a não apresentar ambiguidade sintática e semântica. Resnik aborda que genes que compartilham mais informações possuem uma similaridade maior, visto isso, é utilizado o conceito de *information-content* para calcular a similaridade semântica, *information-content* pode ser descrito como o negativo do logaritmo da probabilidade, como mostra a Equação 23, com isso o conceito de *information-content* acaba sendo formado de forma intuitiva, visto que conforme a probabilidade aumenta, a informação diminui, em outras palavras, quanto menos especifico a informação, menos valor ela carrega. Para o cálculo da similaridade Resnik considera o valor máximo de IC de todas os atributos que os objetos análisados possuem em comum, como mostra a Equação 24. Em experimentos Resnik demonstra que métodos baseados em arestas são menos efetivos, por apresentar resultados espúrios de similaridade.

$$IC = -\log p(c) \tag{23}$$

$$sim(w1, w2) = \max_{c \in S(c1,c2)} [-log(-p(c))] \tag{24}$$

4.4.6 Jiang

Jiang apresenta uma nova abordagem para o cálculo da similaridade semântica entre palavras e conceitos, esta nova abordagem herda o conceito apresentado em métodos baseados em arestas, porém Jiang aprimora o método adicionando abordagem de nós usando o *information-content*, e assim Jiang combina a estrutura da taxonomia lexical com um corpus estatístico de informações. Jiang mensura pesos para as arestas(*LS*) que ligam o nó filho e o nó pai utilizando

Tabela 5 – Comparação entre as correlações segundo Jiang com o julgamento humano(JIANG; CONRATH, 1997)

Similarity Method	Correlation (r)
Julgamento humano	0,8848
Baseado em nó(IC)	0,7941
Baseado em aresta(Contagem de aresta)	0,6004
Combinando o modelo de distância	0,8282

a diferença de IC, como define a Equação 25.

$$LS(c_i, p) = IC(c_i) - IC(p) \tag{25}$$

Para o cálculo da similaridade Jiang também leva em consideração outros fatores além do IC para o cálculo do peso das arestas(*wt*), como profundidade do nó *d(p)*, quantidade de arestas que o nó filho possui *E(p)*, a média de ligações em toda a hierarquia $\overline{E}$ e também o tipo de ligação entre os nós *T(c, p)*, como define a Equação 26.

$$wt(c,p) = \left(\beta + (1-\beta)\frac{\overline{E}}{E(p)}\right)\left(\frac{d(p)+1}{d(p)}\right)^{\alpha}[IC(c) - IC(p)]T(c,p) \tag{26}$$

Por fim o cálculo da similaridade é a soma dos pesos das arestas ao longo do caminho mais curto ligando dois nós, Equação 27.

$$Dist(w_1, w_2) = \sum_{c \in \{path(c_1,c_2) - LCA(c_1,c_2)\}} wt(c,parent(c)) \tag{27}$$

Em seus experimentos Jiang comparou a correlação de seu método e de métodos baseados em aresta com o julgamento humano para a similaridade, e seu método apresentou o valor mais alto de correlação, os resultados podem ser vistos na Tabela 5.

5 MATERIAIS E MÉTODOS

5.1 CONJUNTO DE DADOS

Afim de realizar testes na metodologia e comparar com outros trabalhos semelhantes da literatura, foram adotados alguns conjuntos de dados contendo genes e seus termo GO de duas espécies distintas, sendo elas *Arabidopsis thaliana* e *Saccharomyces cerevisiae*.

Foram separados três grupos de genes pertencentes a vias metabólicas distintas presentes na *Arabidopsis thaliana*, sendo elas: *mevalonate*, *methylerythritol phosphate* e *stachyose biosynthesis*, para comporem um conjunto de dados, descritos na Tabela 6. A escolha desses três pathways se deu pelo motivo que os pathways *mevalonate*, *methylerythritol phosphate* são independentes, visto que estes pathways estão relacionadas com a produção de proteínas diferentes e possuem localização celular diferentes, o terceiro pathway *stachyose biosynthesis* foi adicionado para ser um grupo de controle, para ter um resultado de clusterização mais confiável. Dados dos genes presentes em cada via metabólica foram auferidos do *The Arabidopsis Information Resource* (TAIR) (LAMESCH *et al.*, 2012) que mantém um banco de dados da genética e biologia molecular da espécie *Arabidopsis thaliana*. Dados disponibilizados pelo TAIR incluem a sequência completa do genoma, juntamente com a estrutura do gene, informações sobre o produto, expressão do gene, estoques de DNA e sementes, mapas do genoma, marcadores físicos e genéticos, publicações e informações sobre a comunidade de pesquisa da *Arabidopsis thaliana*. Dados presentes no TAIR utilizam pesquisas recentes e informações da comunidade para semanalmente atualizar o banco de dados.

Foram adotados também cinco conjuntos de dados da *Saccharomyces cerevisiae*, cada qual correspondendo a uma via metabólica: *mevalonate, phenylalanine degradation, removal of*

Tabela 6 – Conjunto de genes e respectivas vias metabólicas da *Arabidopsis thaliana*.

Mevalonate	Methylerythritol phosphate	Stachyose biosynthesis
AACT1	CLA1	GolS1
ACAT2	DXPS1	GolS2
HMGS	DXPS3	GolS3
HMG1	DXR	GolS4
HMG2	ISPD	GolS5
MK	CDPMEK	GolS6
At1g31910	ISPF	GolS7
MVD1	HDS	GATL10
MDD2	HDR	GATL4

Tabela 7 – Conjunto de genes e respectivas vias metabólicas da *Saccharomyces cerevisiae*.

Mevalonate	Phenylalanine degradation	Removal of superoxide radicals	Valine degradation	Mannose degradation
ERG10	ADH1	SOD1	BAT1	GLK1
MVD1	ADH2	SOD2	BAT2	HXK1
IDI1	ADH3	CTA1	PDC1	HXK2
ERG13	ADH4	CTT1	PDC5	PMI40
HMG1	ADH5		PDC6	
HMG2	SFA1		SFA1	
ERG12	PDC1		ADH4	
ERG8	PDC5		ADH5	
	PDC6			
	ARO10			
	ARO8			
	ARO9			

superoxide radicals, valine degradation e *mannose degradation*, descritos na Tabela 7. Todos esses pathways foram o mesmos usado no trabalho do GOGO (ZHAO; WANG, 2018), por esse motivo foi separado estes conjuntos de dados para poder fazer uma comparação entre os resultados. Estas vias metabólicas foram auferidos no banco de dados SGD (CHERRY *et al.*, 2012) o qual provê informações biológicas integradas abrangentes para a levedura, além disso fornece ferramentas de pesquisa e análise para exploração dos dados, permitindo assim a descoberta de relações funcionais entre sequências e produtos gênicos em fungos e organismos superiores.

A Ontologia dos genes foi obtida do consórcio da Ontologia Gênica (ASHBURNER *et al.*, 2000; CONSORTIUM, 2016), o qual tem como objetivo fornecer informações disponíveis sobre as funções dos genes e produtos gênicos de uma forma padronizada. Toda a ontologia é contida em um arquivo único nomeado de "go-basic.obo", ele possui uma estrutura que agrupa as informações de cada termo GO e é utilizado uma palavra reservada '[Term]' entre colchetes para representar um outro termo GO. O arquivo é único pra todas as espécies, ou seja, não um arquivo específico para a espécie. A versão da ontologia usada é a versão de 02/05/2020. A Ontologia Gênica foi utilizada junto ao pacote R chamado "ontologyPlot"e "ontologyIndex"para desenhar as DAG, assim como ler as ontologias e criar listas para manipular o conjunto de termos. Além do arquivo de ontologia, também foram colhidas as anotações dos genes separadas por espécie, nesses arquivos são descritos os genes, seus sinônimos, termos GO e também a categoria da ontologia, como mostra a Figura 6. Esses arquivos que descrevem os genes da *Arabidopsis thaliana* e *Saccharomyces cerevisiae* foram usados para seleção dos termos GO presentes em cada gene.

5.2 PRÉ PROCESSAMENTO DOS DADOS

Para cada um dos genes presentes nas Tabelas 6 e 7 foi necessário automatizar a busca pelos termos GO associados a cada um deles, para tal tarefa foi desenvolvido um algoritmo na linguagem R, o qual usa as anotações de cada espécie, descritas na seção 5.1, para encontrar todos os termos GO associados aos genes, de forma a retornar uma lista de genes e seus termos GO, como mostra a Figura 7.

```
25  TAIR    locus:2031476   ENO1        GO:0000015  TAIR:AnalysisReference:501756966    IEA
    InterPro:IPR000941  C   AT1G74030   AT1G74030|ENO1|enolase 1|F2P9.10|F2P9_10    protein
    taxon:3702  20190907    InterPro        TAIR:locus:2031476
26  TAIR    locus:2043067   ENOC        GO:0000015  TAIR:AnalysisReference:501756966    IEA
    InterPro:IPR000941  C   AT2G29560   AT2G29560|ENOC|ENO3|cytosolic enolase|enolase
    3|F16P2.6|F16P2_6 protein taxon:3702  20190408    InterPro        TAIR:locus:2043067
27  TAIR    locus:2044851   LOS2        GO:0000015  TAIR:AnalysisReference:501756966    IEA
    InterPro:IPR000941  C   AT2G36530   AT2G36530|LOS2|ENO2|LOW EXPRESSION OF OSMOTICALLY
    RESPONSIVE GENES 2|enolase 2|F1O11.16|F1O11_16    protein taxon:3702  20190408    InterPro
            TAIR:locus:2044851
28  TAIR    locus:2032970   AT1G25260       GO:0000027  TAIR:AnalysisReference:501756966    IEA
    InterPro:IPR033867  P   AT1G25260   AT1G25260|F4F7.35|F4F7_35   protein taxon:3702  20190404
        InterPro        TAIR:locus:2032970
```

Figura 6 – Exemplos de anotação dos genes da *Arabidopsis thaliana*

```
1  AACT1 GO:0005737 GO:0009507 GO:0047634 GO:0003985 GO:0006635 GO:0003988
2  ACAT2 GO:0009507 GO:0009793 GO:0005886 GO:0005777 GO:0005829 GO:0003985 GO:0009846 GO:0009860
   GO:0016125 GO:0009536 GO:0003988 GO:0006635
3  HMGS GO:0005739 GO:0019287 GO:0005829 GO:0009506 GO:0004421 GO:0010142 GO:0006084
4  HMG1 GO:0015936 GO:0016126 GO:0005515 GO:0005789 GO:0019287 GO:0042282 GO:0008299 GO:0016020
   GO:0005783 GO:0005778 GO:0004420
5  HMG2 GO:0005634 GO:0015936 GO:0016104 GO:0016126 GO:0008299 GO:0042282 GO:0043231 GO:0004420
   GO:0005789 GO:0005778
6  MK GO:0019287 GO:0005829 GO:0005737 GO:0004496 GO:0016310
```

Figura 7 – Exemplo de saída dos dados após o pré-processamento

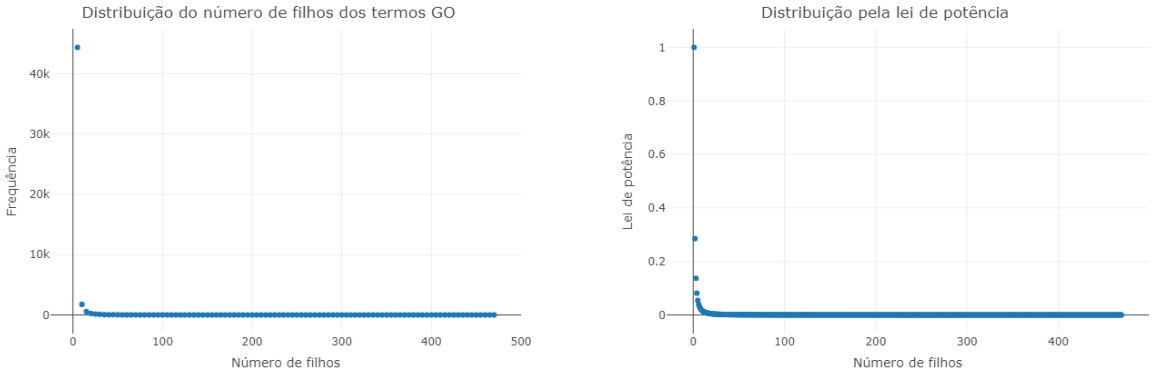
5.3 METODOLOGIA PROPOSTA

Esta seção apresenta a metodologia proposta para calcular a similaridade semântica entre termos GO e genes. A metodologia proposta neste trabalho apresenta uma nova abordagem para o cálculo da similaridade semântica, trazendo uma metodologia *data-driven* que diferentemente de outros métodos, o modelo se ajusta conforme o conjunto de dados das ontologias para calcular a similaridade semântica.

Inicialmente foi feito uma análise exploratória da Ontologia Gênica, um dos gráficos gerados foi a Figura 8(a), o qual exibe como é a distribuição do número de filhos dos termos GO. Sabendo que a média de filhos de cada termo GO é 1,81; foi realizado uma distribuição usando a lei de potência, Equação 1, definindo o k como o número de filhos; e o γ como 1,81; o resultado pode ser visto na Figura 8(b). Visto isso, foi identificado uma semelhança entre as

duas distribuições, indicando que a distribuição do número dos filhos do termos é uma lei de potência. Afim de validar esta afirmação, foi aplicado a função Log na base 10 sobre ambos os gráficos 8(a) e 1. A figura 9(b) mostra que quando aplicado um log sobre os eixos X e Y de uma distribuição definida pela lei de potência, o resultado é uma reta. A Figura 9(a) mostra que o log da distribuição do número de filhos tende a uma reta também.

Com base no histograma das distribuições de probabilidade da lei de potência com média 1,81 e da probabilidade de ocorrência do número de filhos da ontologia (GO), foi realizado o teste de Wilcoxon (HEUMANN *et al.*, 2016), que é um método não-paramétrico para comparação de duas amostras pareadas, dessa forma foi avaliado a diferença entre esses dois histogramas para estabelecer se eles são estatisticamente diferentes um do outro. O p-valor obtido com a aplicação do teste de hipóteses foi de 0,7847, sendo assim maior que 5% de significância, portanto não é possível rejeitar a hipótese nula de que os valores dos histogramas de probabilidades são semelhantes.



(a) Distribuição do número de filhos na Onotologia Gênica.

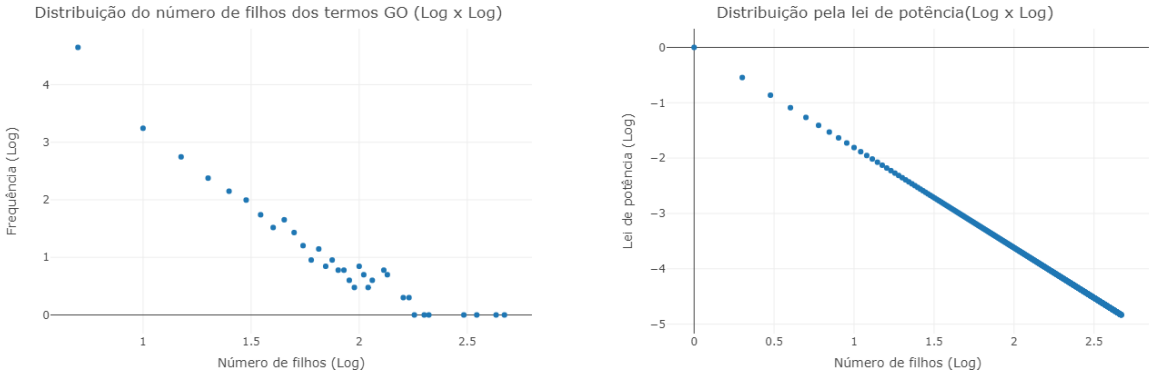
(b) Distribuição dos dados pela lei de potência.

Figura 8 – Distribuição do número de filhos na Ontologia Gênica comparado com a distribuição de dados descrita pela lei de potência.

Fonte: Autoria própria.

5.3.1 Similaridade entre termos GO

A visão geral do novo método cálculo da similaridade semântica entre genes considerando a Ontologia Gênica é apresentada na Figura 10. Mais especificamente, é proposto a aplicação de um peso semântico(W_c) entre as ligações dos termos GO com seus ancestrais, para quantificar a relevância do termo para o cálculo da similaridade. A relevância que um termo GO possui é ligada a quão específico e único é o termo, posto isso, foi realizado um estudo sobre a distribuição de filhos de cada termo GO presente na Ontologia Gênica. Usando a lei de potência,



(a) Log da distribuição do número de filhos. **(b) Log da distribuição pela lei de potência**

Figura 9 – Função log aplicada sobre as distribuições de dados da Figura 8(a) e 8(b)
Fonte: Autoria própria.

foi formulada a Equação 28 com o expoente recebendo um valor negativo que varia conforme o tipo de relação semântica, se a ligação com o termo descendente for do tipo *"part_of"* o valor de w_s será de 0,6; caso seja *"is_a"* o valor será de 0,8. Tais valores foram definidos no artigo (WANG *et al.*, 2007). O *k* é o número de filhos que o termo pai possui, dessa forma quanto mais filhos o termo pai possui menos especifico é o termo GO, levando a uma penalização dada pela lei de potência, de acordo com a distribuição dos graus indetificada na árvore GOGO, i.e., sem a definição de parâmetros, adotando apenas o peso semântico w_s.

$$w_c = k^{(w_s - 1)} \tag{28}$$

Utilizando o peso semântico de cada termo, a próxima etapa é o cálculo do valor semântico (S_A) de cada termo, que é de fato quanto de informação cada termo carrega, pois é levado em consideração a topologia do grafo, quanto mais longe o termo GO esta da folha menos relevante ele é, e o seu valor semântico diminuirá. Tal valor é definido na Equação 29, onde *t* é um termo GO presente no grafo, o *t'* é o termo filho do termo *t*, *A* é o termo alvo, o que está sendo calculado a similaridade.

$$\begin{cases} S_A(t) = 1 & \text{if } t = A \\ S_A(t) = Max\{w_c \times S_A(t') | t' \in children(t)\} & \text{if } t \neq A \end{cases} \tag{29}$$

Por fim, para encontrar o valor de similaridade entre os termos GO é aplicada a Equação 30, usada para definir um valor que varia de 0 a 1 e que mostra quão diferente é o caminho dos termos GO até a raiz e também o quão especifico é este caminho. O numerador é a somatória dos S_A de todos os termos que são ancestrais em comum dos termos A e B, já o denominador é

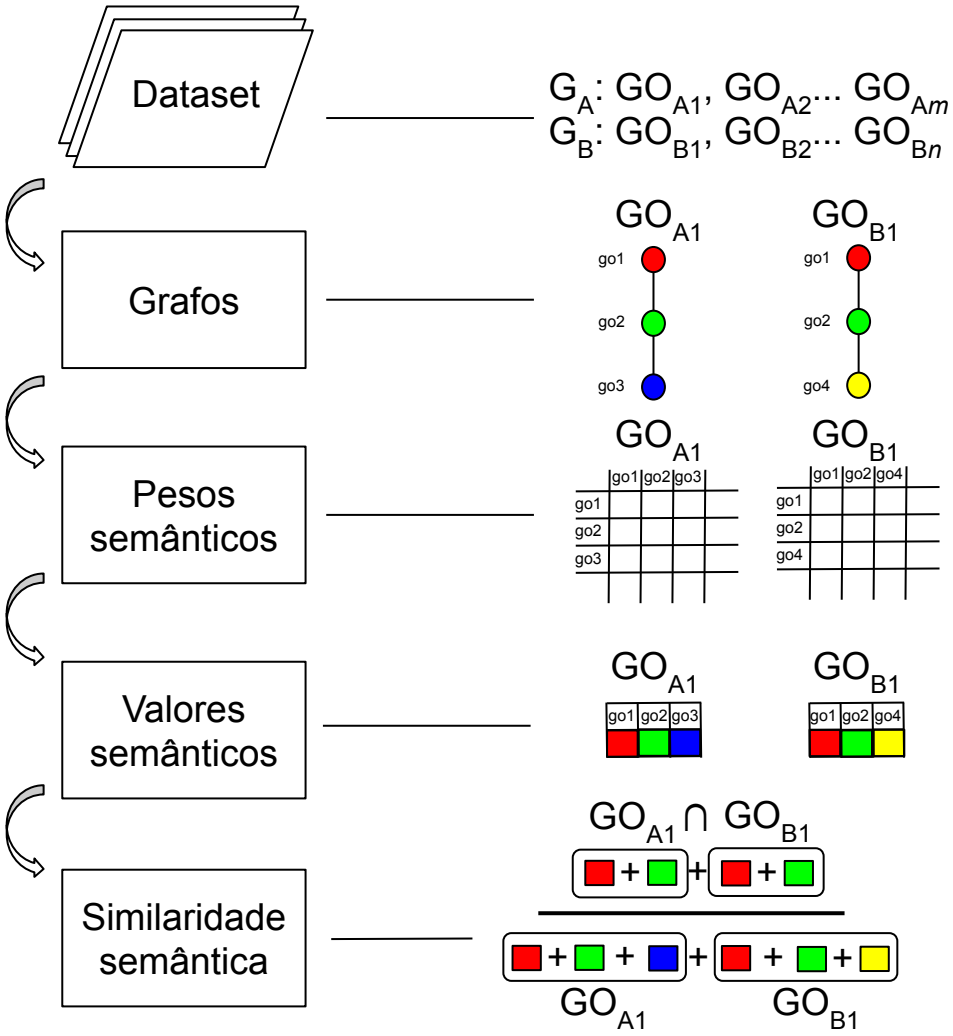


Figura 10 – Pipeline do cálculo da similaridade semântica entre termos GO

a somatorio de todos os S_A definidos para cada termo GO A e B.

$$S_{GO}(A,B) = \frac{\sum\limits_{t \in T_A \cap T_B} (S_A(t) + S_B(t))}{\sum\limits_{t \in T_A} S_A(t) + \sum\limits_{t \in T_A} S_B(t)} \tag{30}$$

5.3.2 Similaridade entre genes

O cálculo da similaridade entre um par de genes considera a média das melhores combinações de similaridade entre os termos GO. Para cada termo GO presente em cada gene, é selecionado a combinação com um termo GO do outro gene que maximiza o valor de similaridade, como define a Equação 31, essa Equação é realizada para cada um dos genes. Os resultados obtidos da Equação 31 são somados, em seguida é realizado uma média de similaridade entre os termos GO, como define a Equação 32. A Figura 11 define o pipeline para o cálculo da

similaridade semântica entre genes, aonde *BmS* são as melhores combinações de similaridade semântica. O denominador da Equação 32 é o número de termos GO presentes em cada gene, sendo assim, na Figura 11 *m* é igual a dois e *n* também é igual a 2.

$$Sim(go, G_1) = \overset{Max}{_{1 \leq i \leq m}} S_{GO}(go, go_{1i}) \tag{31}$$

$$Sim(G_1, G_2) = \frac{\sum_{1 \leq i \leq m} Sim(go_{1i}, G_2) + \sum_{1 \leq j \leq n} Sim(go_{2j}, G_1)}{m + n} \tag{32}$$

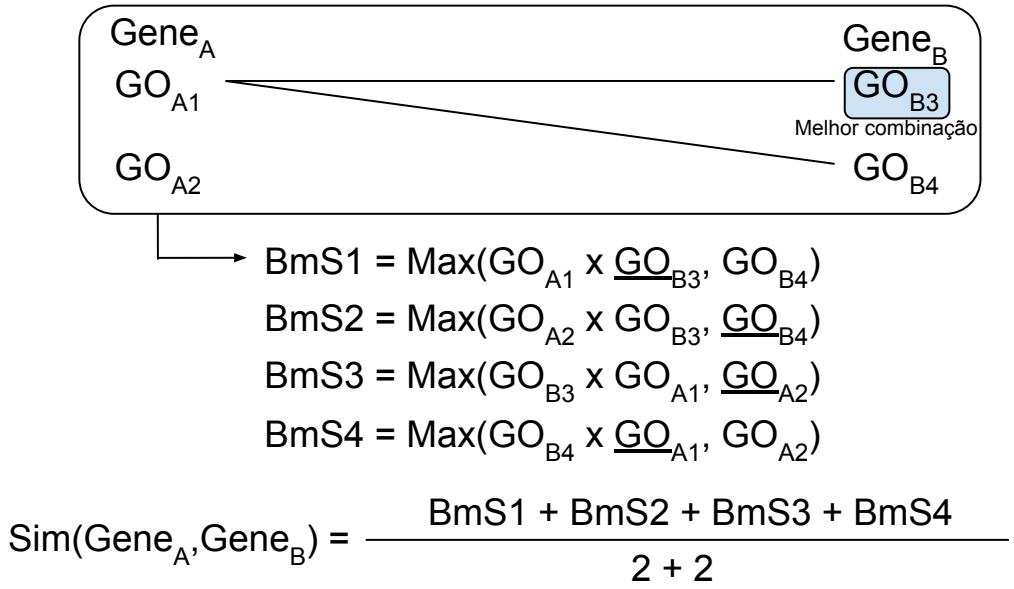


Figura 11 – Pipeline do cálculo da similaridade semântica entre termos genes

6 RESULTADOS E DISCUSSÃO

Nesta seção é descrito todos os testes e resultados obtidos ao decorrer do projeto, além disso serão levantadas observações em relação aos resultados.

Para cada um dos 6 conjuntos de dados descritos na seção 5.1, foi aplicado um produto cartesiano entre todos os genes presentes no conjunto de dados, a partir disso foi construída uma matriz de similaridade, como mostra a Tabela 8, que representa a matriz criada usando funções moleculares para o via metabólica *Valine degradation* para a espécie *Saccharomyces cerevisiae*, outro exemplo de matriz de similaridade, mas usando processos biológicos é a Tabela 9, que representa a via metabólica *Removal of superoxide radicals* da espécie *Saccharomyces cerevisiae*. A via metabólica Valine degradation é divido em 3 grupos, como mostra a Figura 12, o primeiro grupo é formado pelos genes BAT1 e BAT2, os quais estão relacionados com o *branched-chain amino acid*, o segundo grupo tem os genes PDC6, PDC5 e PDC1, cuja estão ligados com a *decarboxylase*, por fim o último grupo é composto pela SFA1, ADH5 e ADH4, quais estão relacionados com a *alcohol dehydrogenase*. A tabela 10 mostra que as relações dos termos GO presentes em cada grupo possuem termos semelhantes, alguns exemplo de cálculo da SS entre esses genes da via metabólica Valine degradation podem serem vistos na Tabela 11, a qual mostra os resultados de SS entre alguns termos GO presentes nos genes ADH4 e ADH5, foram considerados os termos GO associados a processos biológicos, pode-se observar que genes semelhantes que eh o caso do termo GO:0000947 resulta em uma alta similaridade, enquanto que termos GO associados a diferentes processo biológicos possuem baixa similaridade, por exemplo o GO:0006116 e o GO:0000947. Cada um desses grupos foram identificados com altas similaridades, cada grupo foi destacado com cores na Figura 12. O que mostra adequação do método proposto em mensurar as similaridades semânticas entres os genes.

Tabela 8 – Matriz de similaridade do método proposto para a via metabólica Valine degradation

	BAT1	**BAT2**	**PDC1**	**PDC5**	**PDC6**	**SFA1**	**ADH4**	**ADH5**
BAT1	1.000	1.000	0.052	0.052	0.051	0.042	0.031	0.031
BAT2	1.000	1.000	0.052	0.052	0.051	0.042	0.031	0.031
PDC1	0.052	0.052	1.000	1.000	0.799	0.036	0.025	0.025
PDC5	0.052	0.052	1.000	1.000	0.799	0.036	0.025	0.025
PDC6	0.051	0.051	0.799	0.799	1.000	0.038	0.028	0.028
SFA1	0.042	0.042	0.036	0.036	0.038	1.000	0.601	0.601
ADH4	0.031	0.031	0.025	0.025	0.028	0.601	1.000	1.000
ADH5	0.031	0.031	0.025	0.025	0.028	0.601	1.000	1.000

Tabela 9 – Matriz de similaridade do método proposto para a via metabólica Removal of superoxide radicals

	SOD1	SOD2	CTT1	CTA1
SOD1	1.000	0.390	0.105	0.197
SOD2	0.390	1.000	0.187	0.408
CTT1	0.105	0.187	1.000	0.563
CTA1	0.197	0.408	0.563	1.000

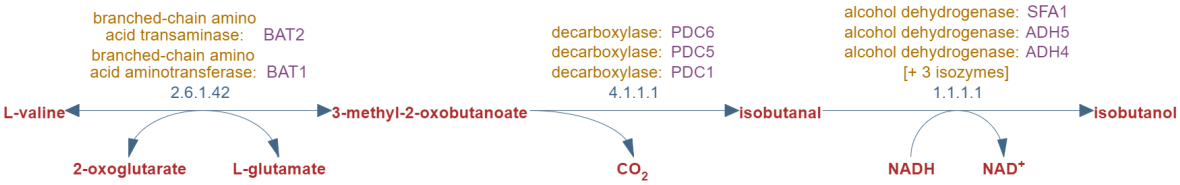


Figura 12 – Clusters da via metabólica Valine degradation. Imagem retirada do website do SGD.

Tabela 10 – Relação entre as atividades dos termos GO de cada gene para a via metabólica Valine degradation.

Gene	Termo GO	Relação
BAT1	GO:0009082	*branched-chain amino acid biosynthetic process*
	GO:0009083	*branched-chain amino acid catabolic process*
	GO:0004084	*branched-chain-amino-acid transaminase activity*
	GO:0005759	*mitochondrial matrix*
	GO:0005739	*mitochondrion*
BAT2	GO:0009082	*branched-chain amino acid biosynthetic process*
	GO:0009083	*branched-chain amino acid catabolic process*
	GO:0004084	*branched-chain-amino-acid transaminase activity*
	GO:0005737	*cytoplasm*
	GO:0005634	*nucleus*
PDC1	GO:0000949	*aromatic amino acid family catabolic process to alcohol via Ehrlich pathway*
	GO:0006090	*pyruvate metabolic process*
	GO:0006559	*L-phenylalanine catabolic process*
	GO:0006569	*tryptophan catabolic process*
	GO:0004737	*pyruvate decarboxylase activity*
	GO:0047433	*branched-chain-2-oxoacid decarboxylase activity*
	GO:0005634	*nucleus*
	GO:0005829	*cytosol*
	GO:0046809	*replication compartment*
	GO:0005737	*cytoplasm*
PDC5	GO:0000949	*aromatic amino acid family catabolic process to alcohol via Ehrlich pathway*
	GO:0006090	*pyruvate metabolic process*
	GO:0006559	*L-phenylalanine catabolic process*
	GO:0006569	*tryptophan catabolic process*
	GO:0004737	*pyruvate decarboxylase activity*
	GO:0047433	*branched-chain-2-oxoacid decarboxylase activity*
	GO:0005634	*nucleus*
	GO:0005737	*cytoplasm*
PDC6	GO:0000949	*aromatic amino acid family catabolic process to alcohol via Ehrlich pathway*
	GO:0006067	*ethanol metabolic process*
	GO:0006559	*L-phenylalanine catabolic process*
	GO:0006569	*tryptophan catabolic process*
	GO:0004737	*pyruvate decarboxylase activity*
	GO:0005737	*cytoplasm*
	GO:0000947	*amino acid catabolic process to alcohol via Ehrlich pathway*
	GO:0033859	*furaldehyde metabolic process*
	GO:0046294	*formaldehyde catabolic process*
	GO:0004022	*alcohol dehydrogenase (NAD+) activity*

SFA1	GO:0033833	*hydroxymethylfurfural reductase (NADH) activity*
	GO:0051903	*S-(hydroxymethyl)glutathione dehydrogenase activity*
	GO:0005737	*cytoplasm*
	GO:0005739	*mitochondrion*
ADH4	GO:0000947	*amino acid catabolic process to alcohol via Ehrlich pathway*
	GO:0006113	*fermentation*
	GO:0004022	*alcohol dehydrogenase (NAD+) activity*
	GO:0005739	*mitochondrion*
ADH5	GO:0000947	*amino acid catabolic process to alcohol via Ehrlich pathway*
	GO:0006116	*NADH oxidation*
	GO:0043458	*ethanol biosynthetic process involved in glucose fermentation to ethanol*
	GO:0004022	*alcohol dehydrogenase (NAD+) activity*
	GO:0005634	*nucleus*
	GO:0005737	*cytoplasm*

Tabela 11 – Similaridade entres os termos GO presentes nos genes ADH4 e ADH5 relacionados a processos biológicos.

	GO:0000947	**GO:0006113**
GO:0000947	1,000	0,084
GO:0006116	0,071	0,261
GO:0043458	0,264	0,394

A partir das matrizes de similaridade foram clusterizados os resultados utilizando o método k-metoids, sendo este um método mais robusto em relação ao k-means. A Tabela 12 apresenta os resultados de clusterização para a via metabólica Valine degradation e a Tabela 13 apresenta os resultados de clusterização para a via Removal of superoxide radicals, ambas pertencentes a *Saccharomyces cerevisiae*.

Tabela 12 – Clusters gerados para a via metabólica Valine degradation com termos GO relacionados a funções moleculares.

Funções moleculares					
PLAWSS	**Resnik**	**Wang**	**Jiang**	**GOGO**	**SGD**
BAT1	BAT1	BAT1	BAT1	BAT1	BAT1
BAT2	BAT2	BAT2	BAT2	BAT2	BAT2
PDC1	PDC1	PDC1	PDC1	PDC1	PDC1
PDC5	PDC5	PDC5	PDC5	PDC5	PDC5
PDC6	PDC6	PDC6	PDC6	PDC6	PDC6
SFA1	SFA1	SFA1	SFA1	SFA1	SFA1
ADH4	ADH4	ADH4	ADH4	ADH4	ADH4
ADH5	ADH5	ADH5	ADH5	ADH5	ADH5

Nenhum dos métodos analisados conseguiu clusterizar corretamente os genes da *Arabidopsis thaliana*, como mostras as Tabelas 14, 15 e 16. Dentre as vias metabólicas, a Stachyose biosynthesis foi a que teve os melhores resultados de clusterização.

Quando clusterizados os genes com base na ontologia componente celular, foram obtidos

Tabela 13 – Clusters gerados para a via metabólica Removal of superoxide radicals com termos GO relacionados aos processos biológicos.

Processos biológicos					
PLAWSS	**Resnik**	**Wang**	**Jiang**	**GOGO**	**SGD**
SOD1	SOD1	SOD1	SOD1	SOD1	SOD1
SOD2		SOD2		SOD2	SOD2
	CTT1		CTT1		
CTA1	CTA1	CTA1	CTA1	CTA1	CTA1
CTT1	SOD2	CTT1	SOD2	CTT1	CTT1

Tabela 14 – Clusters formados para a espécie *Arabidopsis thaliana* utilizando os termos GO associados a funções moleculares.

Funções moleculares					
PLAWSS	**Resnik**	**Wang**	**Jiang**	**GOGO**	**TAIR**
AACT1	AACT1	AACT1	AACT1	AACT1	AACT1
ACAT2	ACAT2	ACAT2	ACAT2	ACAT2	ACAT2
HMGS	HMGS	HMGS	HMGS	HMGS	HMGS
MK	MK	MK	HMG2	MK	HMG1
AT1G31910	CLA1	AT1G31910	MK	AT1G31910	HMG2
CLA1	DXPS1	ISPD	AT1G31910	CLA1	MK
DXPS1	DXPS3	CDPMEK	MVD1	DXPS1	AT1G31910
DXPS3	ISPD	GolS1	MDD2	DXPS3	MVD1
ISPD	CDPMEK	GATL10	DXR	ISPD	MDD2
CDPMEK			ISPD	CDPMEK	
GolS1	HMG1	HMG1	CDPMEK	GolS1	CLA1
GATL10	HMG2	HMG2	ISPF	GATL10	DXPS1
	AT1G31910	MVD1	HDS		DXPS3
HMG1	MVD1	MDD2	HDR	HMG1	DXR
HMG2	MDD2	DXR	GolS1	HMG2	ISPD
MVD1	DXR	ISPF		MVD1	CDPMEK
MDD2	ISPF	HDS	HMG1	MDD2	ISPF
DXR	HDS	HDR	CLA1	DXR	HDS
ISPF	HDR		DXPS1	ISPF	HDR
HDS		CLA1	DXPS3	HDS	
HDR	GolS1	DXPS1		HDR	GolS1
	GolS2	DXPS3	GolS2		GolS2
GolS2	GolS3	GolS2	GolS3	GolS2	GolS3
GolS3	GolS4	GolS3	GolS4	GolS3	GolS4
GolS4	GolS5	GolS4	GolS5	GolS4	GolS5
GolS5	GolS6	GolS5	GolS6	GolS5	GolS6
GolS6	GolS7	GolS6	GolS7	GolS6	GolS7
GolS7	GATL10	GolS7	GATL10	GolS7	GATL10
GATL4	GATL4	GATL4	GATL4	GATL4	GATL4

resultados inferiores quando comparado com a funções moleculares e processos biológicos, como mostra a Tabela 20. Demonstrando que somente a localização aonde o gene exerce a sua função molecular, não dita as características do gene, pois só a localização é uma informação muito genérica que não diz muito sobre o gene e a sua função.

Os resultados dos clusters para cada conjunto de dados pode ser visto na Tabela 17, 18 e 19, aonde o resultado da clusterização foi atribuído como "Correto"se todos os grupos de genes foram gerados corretamente, caso contrário recebe "Errado". Clusters construídos a

Tabela 15 – Clusters formados para a espécie *Arabidopsis thaliana* utilizando os termos GO associados a processos biológicos.

Biological Process					
PLAWSS	**Resnik**	**Wang**	**Jiang**	**GOGO**	**TAIR**
HMGS	AACT1	AACT1	AACT1	AACT1	AACT1
HMG1	ACAT2	ACAT2	ACAT2	ACAT2	ACAT2
HMG2	HMG2	DXPS1	HMG1	DXPS1	HMGS
MK	CLA1	GATL4	HMG2	GolS5	HMG1
AT1G31910	DXPS1		MVD1	GolS6	HMG2
MVD1	DXPS3	HMGS	CLA1	GolS7	MK
MDD2	DXR	HMG1	DXPS3	GATL4	AT1G31910
	ISPD	HMG2	DXR		MVD1
CLA1	CDPMEK	MK	ISPD	HMGS	MDD2
DXPS3	ISPF	AT1G31910	CDPMEK	HMG1	
DXR	HDS	MVD1	ISPF	HMG2	CLA1
ISPD	HDR	MDD2	HDS	MK	DXPS1
CDPMEK		CLA1	HDR	AT1G31910	DXPS3
ISPF	HMGS	DXPS3		MVD1	DXR
HDS	HMG1	DXR	HMGS	MDD2	ISPD
HDR	MK	ISPD	MK	CLA1	CDPMEK
	AT1G31910	CDPMEK	AT1G31910	DXPS3	ISPF
AACT1	MVD1	ISPF	MDD2	DXR	HDS
ACAT2	MDD2	HDS	DXPS1	ISPD	HDR
DXPS1		HDR		CDPMEK	
GolS1	GolS1		GolS1	ISPF	GolS1
GolS2	GolS2	GolS1	GolS2	HDS	GolS2
GolS3	GolS3	GolS2	GolS3	HDR	GolS3
GolS4	GolS4	GolS3	GolS4		GolS4
GolS5	GolS5	GolS4	GolS5	GolS1	GolS5
GolS6	GolS6	GolS5	GolS6	GolS2	GolS6
GolS7	GolS7	GolS6	GolS7	GolS3	GolS7
GATL10	GATL10	GolS7	GATL10	GolS4	GATL10
GATL4	GATL4	GATL10	GATL4	GATL10	GATL4

partir de funções moleculares tiveram os melhores resultados para todas os conjuntos de dados analisado. O número de clusters formados corretamente é exibido na Tabela 20, o novo método conseguiu alcançar outros métodos já consolidados, como Wang e GOGO, mostrando assim a sua aplicabilidade.

O método proposto mostrou também ter resultados mais definidos em relação aos clusters, por exemplo para a via metabólica mannose degradation, todos os métodos geraram os clusters esperados usando termos relacionados a processos biológicos, Tabela 21, contudo se analisar as matrizes de similaridades criadas, o método proposto conseguiu diferenciar melhor os clusters, como pode ser visto nas Tabelas 22, 23, 24, 25, 26. O gene PMI40 teve resultados significativamente mais baixos de similaridade no método proposto em comparação aos métodos Jiang, Resnik e Wang, o que é esperado, visto que ele pertence a outro grupo, como mostra a Figura 13. O Método de Resnik apresentou resultados mais baixos entre os genes GLK1, HXK1

Tabela 16 – Clusters formados para a espécie *Arabidopsis thaliana* utilizando os termos GO associados a componente celular.

Componente celular					
PLAWSS	**Resnik**	**Wang**	**Jiang**	**GOGO**	**TAIR**
AACT1	AACT1	AACT1	AACT1	AACT1	AACT1
CLA1	CLA1	ACAT2	ACAT2	ACAT2	ACAT2
DXPS1	DXPS1	AT1G31910	MK	CLA1	HMGS
DXPS3	DXPS3	CLA1	AT1G31910	DXPS1	HMG1
DXR	DXR	DXPS1	CLA1	DXPS3	HMG2
ISPD	ISPD	DXPS3	DXPS1	DXR	MK
CDPMEK	CDPMEK	DXR	DXPS3	ISPD	AT1G31910
ISPF	ISPF	ISPD	DXR	CDPMEK	MVD1
HDS	HDS	CDPMEK	ISPD	ISPF	MDD2
HDR	HDR	ISPF	CDPMEK	HDS	
GATL10	GATL10	HDS	ISPF	HDR	CLA1
GATL4	GATL4	HDR	HDS	GATL10	DXPS1
		GolS3	HDR	GATL4	DXPS3
ACAT2	ACAT2	GATL10	GolS3		DXR
HMGS	HMGS	GATL4	GATL10	HMGS	ISPD
HMG1	HMG1		GATL4	HMG1	CDPMEK
MK	HMG2	HMGS		HMG2	ISPF
AT1G31910	MK	HMG1	HMGS	MK	HDS
MVD1	AT1G31910	HMG2	HMG1	AT1G31910	HDR
MDD2	MVD1	MK	HMG2	MVD1	
GolS1	GolS3	MVD1	MVD1	MDD2	GolS1
GolS3		MDD2	MDD2	GolS2	GolS2
GolS5	MDD2	GolS2	GolS2	GolS3	GolS3
	GolS1	GolS4	GolS4	GolS4	GolS4
HMG2	GolS2	GolS6	GolS6	GolS6	GolS5
GolS2	GolS4	GolS7	GolS7	GolS7	GolS6
GolS4	GolS5				GolS7
GolS6	GolS6	GolS1	GolS1	GolS1	GATL10
GolS7	GolS7	GolS5	GolS5	GolS5	GATL4

Tabela 17 – Comparação dos resultados para as clusterizações usando ontologia ligadas a funções moleculares.

Funções moleculares					
Conjunto de dados	**PLAWSS**	**Resnik**	**Wang**	**Jiang**	**GOGO**
Mannose degradation	Correto	Correto	Correto	Correto	Correto
Mevalonate	Correto	Correto	Correto	Errado	Correto
Phenylalanine degradation	Correto	Correto	Correto	Correto	Correto
Removal of superoxide radicals	Correto	Correto	Correto	Correto	Correto
Valine degradation	Correto	Correto	Correto	Correto	Correto
Arabidopsis thaliana	Errado	Errado	Errado	Errado	Errado

e HXK2, os quais estão no mesmo grupo.

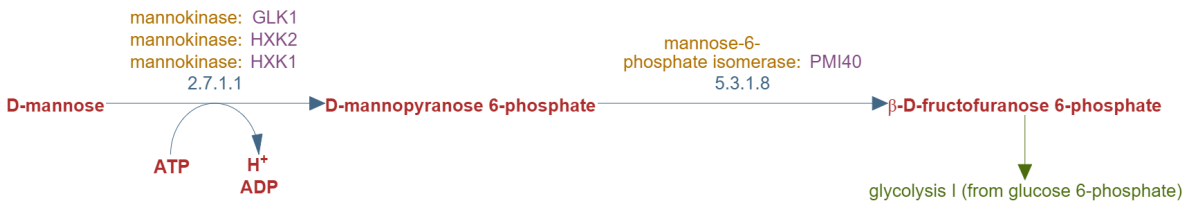


Figura 13 – Clusters da via metabólica Mannose degradation. Imagem retirada do website do SGD.

Tabela 18 – Comparação dos resultados para as clusterizações usando ontologia ligadas a processos biológicos.

Processos biológicos					
Conjunto de dados	**PLAWSS**	**Resnik**	**Wang**	**Jiang**	**GOGO**
Mannose degradation	Correto	Correto	Correto	Correto	Correto
Mevalonate	Errado	Errado	Errado	Errado	Errado
Phenylalanine degradation	Correto	Correto	Correto	Correto	Correto
Removal of superoxide radicals	Correto	Errado	Correto	Errado	Correto
Valine degradation	Correto	Correto	Correto	Correto	Correto
Arabidopsis thaliana	Errado	Errado	Errado	Errado	Errado

Tabela 19 – Comparação dos resultados para as clusterizações usando ontologia ligadas a componentes celulares.

Componente celular					
Conjunto de dados	**PLAWSS**	**Resnik**	**Wang**	**Jiang**	**GOGO**
Mannose degradation	Errado	Correto	Errado	Errado	Errado
Mevalonate	Errado	Errado	Errado	Errado	Errado
Phenylalanine degradation	Errado	Errado	Errado	Errado	Errado
Removal of superoxide radicals	Errado	Errado	Errado	Errado	Errado
Valine degradation	Errado	Errado	Errado	Errado	Errado
Arabidopsis thaliana	Errado	Errado	Errado	Errado	Errado

Tabela 20 – Porcentagem de clusters gerados corretamente.

Ontologia	**PLAWSS**	**Resnik**	**Wang**	**Jiang**	**GOGO**
Função molecular	83.33%	83.33%	83.33%	66.67%	83.33%
Processos biológicos	66.67%	50.00%	66.67%	50.00%	66.67%
Componente celular	0.00%	16.67%	0.00%	0.00%	0.00%

Tabela 21 – Clusters formados para a via metabólica Mannose degradation

Biological Process				
PLAWSS	**Resnik**	**Wang**	**Jiang**	**GOGO**
GLK1	GLK1	GLK1	GLK1	GLK1
HXK1	HXK1	HXK1	HXK1	HXK1
HXK2	HXK2	HXK2	HXK2	HXK2
PMI40	PMI40	PMI40	PMI40	PMI40

Tabela 22 – Matriz de similaridade criado com o método proposto.

	GLK1	**HXK1**	**HXK2**	**PMI40**
GLK1	1.000	0.906	0.702	0.114
HXK1	0.906	1.000	0.805	0.103
HXK2	0.702	0.805	1.000	0.085
PMI40	0.114	0.103	0.085	1.000

Tabela 23 – Matriz de similaridade criado com o método Jiang.

	GLK1	**HXK1**	**HXK2**	**PMI40**
GLK1	1.000	0.916	0.712	0.310
HXK1	0.916	1.000	0.806	0.241
HXK2	0.712	0.806	1.000	0.192
PMI40	0.310	0.241	0.192	1.000

Foram realizadas a medição da precisão e recall para os clusters formados para todos os conjuntos de dados, comparando entre cada método analisado aqui. Em conjuntos de dados cons-

Tabela 24 – Matriz de similaridade criado com o método Resnik.

	GLK1	HXK1	HXK2	PMI40
GLK1	1.000	0.638	0.511	0.285
HXK1	0.638	1.000	0.605	0.250
HXK2	0.511	0.605	1.000	0.220
PMI40	0.285	0.250	0.220	1.000

Tabela 25 – Matriz de similaridade criado com o método Wang.

	GLK1	HXK1	HXK2	PMI40
GLK1	1.000	0.949	0.761	0.217
HXK1	0.949	1.000	0.826	0.210
HXK2	0.761	0.826	1.000	0.260
PMI40	0.217	0.210	0.260	1.000

Tabela 26 – Matriz de similaridade criado com o método GOGO.

	GLK1	HXK1	HXK2	PMI40
GLK1	1.000	0.890	0.690	0.109
HXK1	0.890	1.000	0.805	0.104
HXK2	0.690	0.805	1.000	0.098
PMI40	0.109	0.104	0.098	1.000

tituídos de mais de duas classes, foram feitas analises para cada uma das classes separadamente, cada classe pode ser vista na Tabela 27. Os resultados da precisão são exibidos nas Tabela 28, 30, 32, 34, 36, 38. Já os resultados do recall são mostrados nas Tabelas 29, 31, 33, 35, 37, 39.

Tabela 27 – Clusters de genes para cada conjunto de dados.

Dataset	C1	C2	C3	C4
Mevalonate	ERG10, ERG13	HMG1,HMG2	ERG12, ERG8	MVD1, IDI1
Arabidopsis thaliana	AACT1,ACAT2, HMGS,HMG1, HMG2,MK, AT1G31910,MVD1, MDD2	CLA1,DXPS1, DXPS3,DXR, ISPD,CDPMEK, ISPF,HDS, HDR	GolS1,GolS2, GolS3,GolS4, GolS5,GolS6, GolS7,GATL10, GATL4	
Phenylanine degradation	ARO8, ARO9	ARO10, PDC1, PDC5, PDC6,	SFA1, ADH1, ADH2,ADH3, ADH4, ADH5	
Valine degradation	BAT1, BAT2	PDC1, PDC5 PDC6	SFA1, ADH4, ADH5	

A Figura 14 mostra a DAG dos termos GO:0032787, GO:0072329. Os nós na cor verde são termos GO os quais não são ancestrais do termo GO:0032787, sendo assim, esses termos são a diferença entre GO:0032787 e GO:0072329. A Tabela 40 compara os resultados da similaridade semântica entre os termos para diferentes métodos, os métodos Wang e Jiang apresentaram valores de similaridade altos, o que vai contra a intuição, já que há vários termos GO que diferem eles. Além desse exemplo, foi feito também o cálculo para os termos GO:0044282 e GO:0009056, os quais estão próximos da raiz e termos GO em leveis altos possui uma alta dissimilaridade, por não serem específicos. O método proposto foi o que definiu a menor similaridade entres

Tabela 28 – Precisão da via metabólica Mannose degradation

	BPO	CCO	MFO
PLAWSS	1,0	1,0	1,0
GOGO	1,0	1,0	1,0
Jiang	1,0	1,0	1,0
Resnik	1,0	1,0	1,0
Wang	1,0	1,0	1,0

Tabela 29 – Recall da via metabólica Mannose degradation

	BPO	CCO	MFO
PLAWSS	1,0	0.33	1,0
GOGO	1,0	0.33	1,0
Jiang	1,0	0.33	1,0
Resnik	1,0	1,0	1,0
Wang	1,0	0.33	1,0

Tabela 30 – Precisão da via metabólica Mevalonate

	BPO				CCO				MFO			
Classes	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
PLAWSS	0,67	0,33	0,0	1,0	0,25	0,0	0,0	1,0	1,0	1,0	1,0	1,0
GOGO	0,67	0,33	0,0	1,0	0,25	0,0	0,0	1,0	1,0	1,0	1,0	1,0
Jiang	0,5	1,0	1,0	1,0	0,67	1,0	0,5	1,0	0,5	1,0	0,0	1,0
Resnik	0,67	0,33	0,0	1,0	0,5	0,0	0,0	1,0	1,0	1,0	1,0	1,0
Wang	0,67	0,33	0,0	1,0	0,4	1,0	0,0	1,0	1,0	1,0	1,0	1,0

Tabela 31 – Recall da via metabólica Mevalonate

	BPO				CCO				MFO			
Classes	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
PLAWSS	1,0	0,5	0,0	0,5	0,5	0,0	0,0	0,5	1,0	1,0	1,0	1,0
GOGO	1,0	0,5	0,0	0,5	0,5	0,0	0,0	0,5	1,0	1,0	1,0	1,0
Jiang	1,0	0,5	1,0	0,5	1,0	1,0	0,5	0,5	1,0	1,0	0,0	0,5
Resnik	1,0	0,5	0,0	0,5	0,5	0,0	0,0	0,5	1,0	1,0	1,0	1,0
Wang	1,0	0,5	0,0	0,5	1,0	0,5	0,0	0,5	1,0	1,0	1,0	1,0

Tabela 32 – Precisão da espécie *Arabidopsis thaliana*

	BPO			CCO			MFO		
Classes	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**
PLAWSS	0,28	0,53	1,0	0,08	0,0	0,8	0,42	0,5	1,0
GOGO	0,28	0,53	1,0	0,15	0,0	1,0	0,42	0,5	1,0
Jiang	0,38	0,2	1,0	0,25	0,0	1,0	0,53	0,75	1,0
Resnik	0,25	0,0	1,0	0,08	0,0	0,86	0,44	0,44	1,0
Wang	0,5	0,53	1,0	0,2	0,0	1,0	0,56	0,5	0,7

Tabela 33 – Recall da espécie *Arabidopsis thaliana*

	BPO			CCO			MFO		
Classes	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**
PLAWSS	0,22	0,89	0,56	0,11	0,0	0,44	0,56	0,44	0,78
GOGO	0,22	0,89	0,56	0,22	0,0	0,22	0,56	0,44	0,78
Jiang	0,56	0,11	1,0	0,44	0,0	0,22	0,89	0,33	0,89
Resnik	0,33	0,0	1,0	0,11	0,0	0,67	0,44	0,44	1,0
Wang	0,22	0,89	0,89	0,33	0,0	0,22	0,56	0,44	0,78

Tabela 34 – Precisão da via metabólica Phenylanine degradation

	BPO			CCO			MFO		
Classes	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**
PLAWSS	1,0	1,0	1,0	0,2	0,5	1,0	1,0	1,0	1,0
GOGO	1,0	1,0	1,0	0,2	0,5	1,0	1,0	1,0	1,0
Jiang	1,0	1,0	1,0	0,33	0,43	1,0	1,0	1,0	1,0
Resnik	1,0	1,0	1,0	0,2	0,5	1,0	1,0	1,0	1,0
Wang	1,0	1,0	1,0	0,25	0,5	1,0	1,0	1,0	1,0

Tabela 35 – Recall da via metabólica Phenylanine degradation

	BPO			CCO			MFO		
Classes	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**
PLAWSS	1,0	1,0	1,0	0,5	0,5	0,5	1,0	1,0	1,0
GOGO	1,0	1,0	1,0	0,5	0,5	0,5	1,0	1,0	1,0
Jiang	1,0	1,0	1,0	0,5	0,75	0,33	1,0	1,0	1,0
Resnik	1,0	1,0	1,0	0,5	0,5	0,5	1,0	1,0	1,0
Wang	1,0	1,0	1,0	0,5	0,75	0,33	1,0	1,0	1,0

Tabela 36 – Precisão da via metabólica Removal of superoxide radicals

	BPO	CCO	MFO
PLAWSS	1,0	0,67	1,0
GOGO	1,0	0,67	1,0
Jiang	1,0	0,67	1,0
Resnik	1,0	0,67	1,0
Wang	1,0	0,5	1,0

Tabela 37 – Recall da via metabólica Removal of superoxide radicals

	BPO	CCO	MFO
PLAWSS	1,0	1,0	1,0
GOGO	1,0	1,0	1,0
Jiang	0,5	1,0	1,0
Resnik	0,5	1,0	1,0
Wang	1,0	0,5	1,0

Tabela 38 – Precisão da via metabólica Valine degradation

	BPO			CCO			MFO		
Classes	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**
PLAWSS	1,0	1,0	1,0	0,5	0,33	0,33	1,0	1,0	1,0
GOGO	1,0	1,0	1,0	0,5	0,33	0,33	1,0	1,0	1,0
Jiang	1,0	1,0	1,0	0,5	0,6	1,0	1,0	1,0	1,0
Resnik	1,0	1,0	1,0	0,33	0,5	0,0	1,0	1,0	1,0
Wang	1,0	1,0	1,0	0,5	0,6	1,0	1,0	1,0	1,0

Tabela 39 – Recall da via metabólica Valine degradation

	BPO			CCO			MFO		
Classes	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**
PLAWSS	1,0	1,0	1,0	0,5	0,33	0,33	1,0	1,0	1,0
GOGO	1,0	1,0	1,0	0,5	0,33	0,33	1,0	1,0	1,0
Jiang	1,0	1,0	1,0	0,5	1,0	0,33	1,0	1,0	1,0
Resnik	1,0	1,0	1,0	0,5	0,67	0,0	1,0	1,0	1,0
Wang	1,0	1,0	1,0	0,5	1,0	0,33	1,0	1,0	1,0

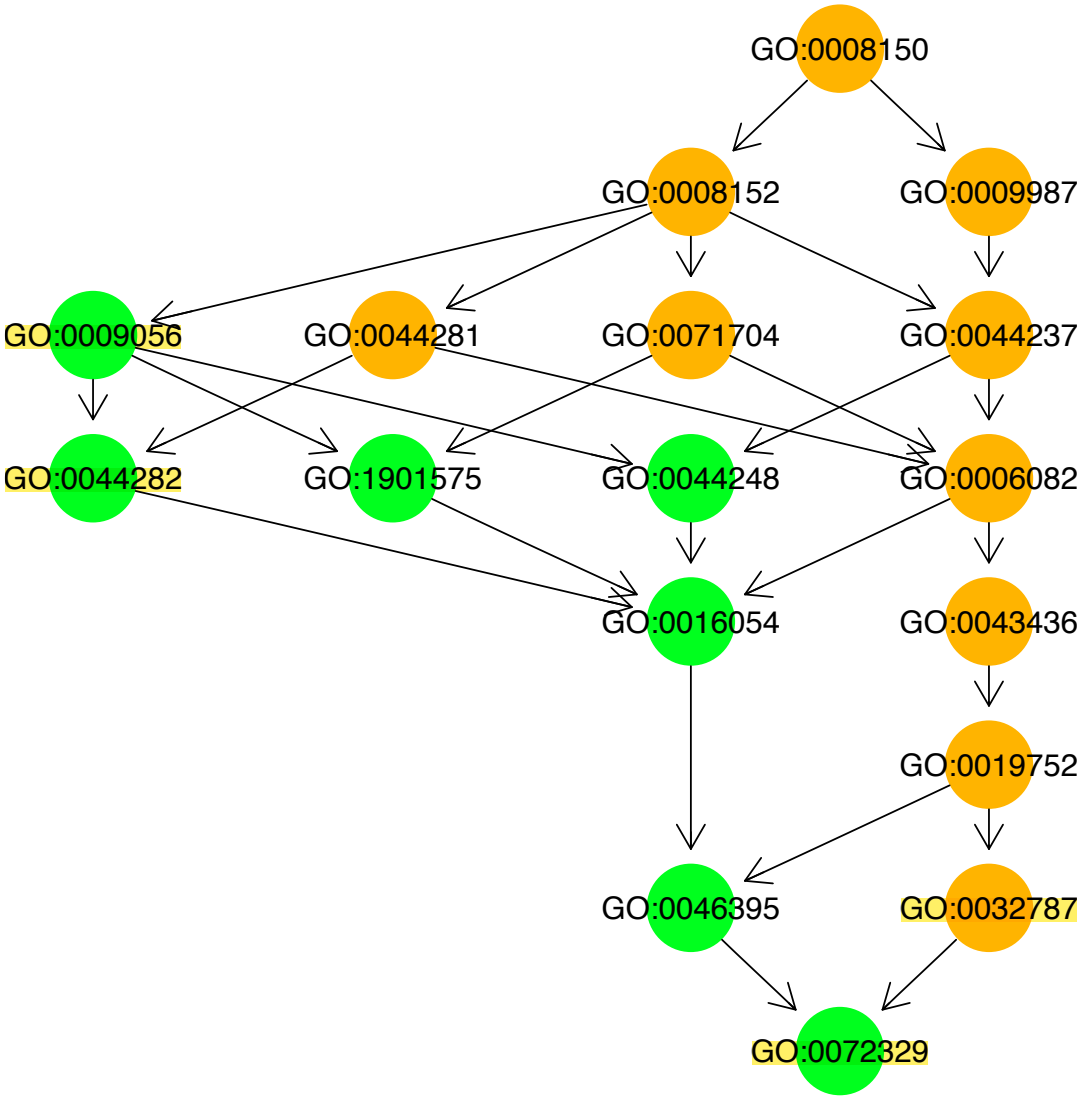


Figura 14 – DAG dos termos GO:0032787 e GO:0072329.

os termos GO. Jiang, Wang e GOGO tiverem como resultado uma similaridade alta, o que é contraintuitivo, pois ambos os genes estão próximo da raíz, sendo então genéricos para afimar uma alta similaridade.

Tabela 40 – Resultados das SS para cada método

Termos GO	Método	SS
GO:0032787, GO:0072329	Resnik	0,431
	Wang	0,712
	Jiang	0,825
	GOGO	0,598
	PLAWSS	0,545
GO:0044282, GO:0009056	Resnik	0,273
	Wang	0,516
	Jiang	0,811
	GOGO	0,484
	PLAWSS	0,182

7 CONCLUSÕES E DIRECIONAMENTOS

Este trabalho tem como objetivo criar um novo método para calcular a similaridade semântica entre termos GO e entre pares de genes, contudo propondo um modelo *data-driven* que não depende de constantes, mas que se ajusta conforme o conjunto de ontologia GO. Foram usados os trabalhos de Wang e GOGO (WANG *et al.*, 2007) e (ZHAO; WANG, 2018) como modelo inicial para criar uma nova abordagem usando lei de potência como medida para mensurar o peso semântico dos termos GO, esse peso leva em consideração não somente o tipo de ligação, como também o número de filhos do nó ancestral(*parent*). Como pode ser visto nos resultados, o método proposto mostrou adequado e funcional, visto que conseguiu alcançar outros métodos, além disso apresentou resultados mais definidos para cada cluster. Considerando apenas a ontologia de componente celular mostrou-se insuficiente para o cálculo da similaridade semântica, tendo resultados muitos inferiores as demais ontologias. Para trabalhos futuros é interessante investigar outras informações que podem ser adicionadas a medida, como informações de outras ontologia, por exemplo a KEGG(do inglês, *Kyoto Encyclopedia of Genes and Genomes*)(KANEHISA *et al.*, 2016; KANEHISA *et al.*, 2015; KANEHISA; GOTO, 2000). Outra abordagem é aplicar a similaridade semântica tratada neste trabalho para a inferência e validação de redes gênicas, as quais exibem como os genes se relacionam, há trabalhos na literatura que utilizam outros métodos de similaridade semântica para tratar esse problema, como o GFD-NET(DÍAZ-MONTAÑA *et al.*, 2017).

REFERÊNCIAS

AKMAL, Suriati; SHIH, Li-Hsing; BATRES, Rafael. Ontology-based similarity for product information retrieval. **Computers in Industry**, Elsevier, v. 65, n. 1, p. 91–107, 2014.

ALBERT, Réka. Scale-free networks in cell biology. **J Cell Sci**, v. 118, n. 21, p. 4947–4957, 2005.

ALMAAS, Eivind; BARABÁSI, Albert-László. Power laws in biological networks. *In*: ____. **Power Laws, Scale-Free Networks and Genome Biology**. Boston, MA: Springer US, 2006. p. 1–11. Disponível em: https://doi.org/10.1007/0-387-33916-7_1.

ASHBURNER, Michael; BALL, Catherine A; BLAKE, Judith A; BOTSTEIN, David; BUTLER, Heather; CHERRY, J Michael; DAVIS, Allan P; DOLINSKI, Kara; DWIGHT, Selina S; EPPIG, Janan T *et al.* Gene ontology: tool for the unification of biology. **Nature genetics**, Nature Publishing Group, v. 25, n. 1, p. 25, 2000.

BARABÁSI, Albert-László. **Linked: The new science of networks**. *[S.l.]*: American Association of Physics Teachers, 2003.

BARABáSI, Albert-Laszlo. Scale-Free Networks: A Decade and Beyond. **Science**, v. 325, n. 5939, p. 412–413, 2009.

BARABÁSI, Albert-Laszlo; ALBERT, Réka. Emergence of scaling in random networks. **Science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999.

BARABáSI, Albert-Laszlo; GULBAHCE, Natali; LOSCALZO, Joseph. Network medicine: a network-based approach to human disease. **Nat Rev Genet**, Nature Publishing Group, v. 12, n. 1, p. 56–68, 2011. ISSN 1471-0056.

BLAKE, Judith A; BULT, Carol J. Beyond the data deluge: data integration and bio-ontologies. **Journal of biomedical informatics**, Elsevier, v. 39, n. 3, p. 314–320, 2006.

BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D. U. Complex networks: Structure and dynamics. **Physics Reports**, v. 424, n. 4-5, p. 175–308, 2006. ISSN 0370-1573.

BOLSHAKOVA, Nadia; AZUAJE, Francisco; CUNNINGHAM, Pádraig. A knowledge-driven approach to cluster validity assessment. **Bioinformatics**, Citeseer, v. 21, n. 10, p. 2546–2547, 2005.

BRAMEIER, Markus; WIUF, Carsten. Co-clustering and visualization of gene expression data and gene ontology terms for saccharomyces cerevisiae using self-organizing maps. **Journal of biomedical informatics**, Elsevier, v. 40, n. 2, p. 160–173, 2007.

CAO, Renzhi; CHENG, Jianlin. Deciphering the association between gene function and spatial gene-gene interactions in 3d human genome conformation. **BMC genomics**, BioMed Central, v. 16, n. 1, p. 880, 2015.

CARBON, S; DOUGLASS, E; DUNN, N; GOOD, B; HARRIS, NL; LEWIS, SE; MUNGALL, CJ; BASU, S; CHISHOLM, RL; DODSON, RJ *et al.* The gene ontology resource: 20 years and still going strong. **Nucleic Acids Research**, 2019.

CHERRY, J Michael; HONG, Eurie L; AMUNDSEN, Craig; BALAKRISHNAN, Rama; BINKLEY, Gail; CHAN, Esther T; CHRISTIE, Karen R; COSTANZO, Maria C; DWIGHT, Selina S; ENGEL, Stacia R *et al.* Saccharomyces genome database: the genomics resource of budding yeast. **Nucleic acids research**, Oxford University Press, v. 40, n. D1, p. D700–D705, 2012.

CHO, Young-Rae; ZHANG, Aidong; XU, Xian. Semantic similarity based feature extraction from microarray expression data. **International journal of data mining and bioinformatics**, Inderscience Publishers, v. 3, n. 3, p. 333–345, 2009.

CONSORTIUM, Gene Ontology. Expansion of the gene ontology knowledgebase and resources. **Nucleic acids research**, Oxford University Press, v. 45, n. D1, p. D331–D338, 2016.

COSTA, L. da F.; RODRIGUES, F. A.; TRAVIESO, G.; VILLAS-BOAS, P. R. Characterization of complex networks: a survey of measurements. **Advances in Physics**, v. 56, n. 1, p. 167–242, 2007.

DÍAZ-DÍAZ, Norberto; AGUILAR-RUIZ, Jesús S. Go-based functional dissimilarity of gene sets. **BMC bioinformatics**, BioMed Central, v. 12, n. 1, p. 360, 2011.

DÍAZ-MONTAÑA, Juan J; DÍAZ-DÍAZ, Norberto; GÓMEZ-VELA, Francisco. Gfd-net: A novel semantic similarity methodology for the analysis of gene networks. **Journal of biomedical informatics**, Elsevier, v. 68, p. 71–82, 2017.

EHSANI, Rezvan; DRABLØS, Finn. Topoicsim: a new semantic similarity measure based on gene ontology. **BMC bioinformatics**, BioMed Central, v. 17, n. 1, p. 296, 2016.

ERDÖS, Paul; RÉNYI, Alfréd. On random graphs publ. **Math. debrecen**, v. 6, p. 290–297, 1959.

GARLA, Vijay N; BRANDT, Cynthia. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. **BMC bioinformatics**, BioMed Central, v. 13, n. 1, p. 261, 2012.

GEORGE, Richard A; HERINGA, Jaap. An analysis of protein domain linkers: their classification and role in protein folding. **Protein Engineering, Design and Selection**, Oxford University Press, v. 15, n. 11, p. 871–879, 2002.

GUO, Xiang; LIU, Rongxiang; SHRIVER, Craig D; HU, Hai; LIEBMAN, Michael N. Assessing semantic similarity measures for the characterization of human regulatory pathways. **Bioinformatics**, Oxford University Press, v. 22, n. 8, p. 967–973, 2006.

HEUMANN, Christian; SCHOMAKER, Michael *et al.* **Introduction to statistics and data analysis**. *[S.l.]*: Springer, 2016.

ITO, Eric Augusto; KATAHIRA, Isaque; VICENTE, Fábio Fernandes da Rocha; PEREIRA, Luiz Filipe Protasio; LOPES, Fabrício Martins. Basinet - biological sequences network: a case study on coding and non-coding rnas identification. **Nucleic Acids Research**, p. gky462, 2018. Disponível em: http://dx.doi.org/10.1093/nar/gky462.

JEONG, H.; TOMBOR, B.; ALBERT, R.; OLTVAI, Z. N.; BARABáSI, Albert-Laszlo. The large-scale organization of metabolic networks. **Nature**, v. 407, p. 651–654, 2000.

JIANG, Jay J; CONRATH, David W. Semantic similarity based on corpus statistics and lexical taxonomy. **arXiv preprint cmp-lg/9709008**, 1997.

JIANG, Yuxiang; ORON, Tal Ronnen; CLARK, Wyatt T; BANKAPUR, Asma R; D'ANDREA, Daniel; LEPORE, Rosalba; FUNK, Christopher S; KAHANDA, Indika; VERSPOOR, Karin M; BEN-HUR, Asa *et al.* An expanded evaluation of protein function prediction methods shows an improvement in accuracy. **Genome biology**, BioMed Central, v. 17, n. 1, p. 184, 2016.

KANEHISA, Minoru; FURUMICHI, Miho; TANABE, Mao; SATO, Yoko; MORISHIMA, Kanae. Kegg: new perspectives on genomes, pathways, diseases and drugs. **Nucleic acids research**, Oxford University Press, v. 45, n. D1, p. D353–D361, 2016.

KANEHISA, Minoru; GOTO, Susumu. Kegg: kyoto encyclopedia of genes and genomes. **Nucleic acids research**, Oxford University Press, v. 28, n. 1, p. 27–30, 2000.

KANEHISA, Minoru; SATO, Yoko; KAWASHIMA, Masayuki; FURUMICHI, Miho; TANABE, Mao. Kegg as a reference resource for gene and protein annotation. **Nucleic acids research**, Oxford University Press, v. 44, n. D1, p. D457–D462, 2015.

KHANIN, Raya; WIT, Ernst. How scale-free are biological networks. **Journal of Computational Biology**, v. 13, n. 3, p. 810–818, 2006.

LAMESCH, Philippe; BERARDINI, Tanya Z; LI, Donghui; SWARBRECK, David; WILKS, Christopher; SASIDHARAN, Rajkumar; MULLER, Robert; DREHER, Kate; ALEXANDER, Debbie L; GARCIA-HERNANDEZ, Margarita *et al.* The arabidopsis information resource (tair): improved gene annotation and new tools. **Nucleic acids research**, Oxford University Press, v. 40, n. D1, p. D1202–D1210, 2012.

LEI, Zhengdeng; DAI, Yang. Assessing protein similarity with gene ontology and its use in subnuclear localization prediction. **BMC bioinformatics**, Springer, v. 7, n. 1, p. 491, 2006.

LEWIS, J; ALBERTS, B; BRAY, D. Biologia molecular da célula. **Porto Alegre: ArtMed**, 2009.

LIMA, Geovana V.L. de; SAITO, Priscila T.M.; LOPES, Fabricio M.; BUGATTI, Pedro H. Classification of texture based on bag-of-visual-words through complex networks. **Expert Systems with Applications**, v. 133, p. 215 – 224, 2019. ISSN 0957-4174. Disponível em: http://www.sciencedirect.com/science/article/pii/S0957417419303483.

LIN, Dekang *et al.* An information-theoretic definition of similarity. *In*: CITESEER. **Icml**. *[S.l.]*, 1998. v. 98, n. 1998, p. 296–304.

LOPES, Fabrício Martins. **Redes complexas de expressão gênica: síntese, identificação, análise e aplicações**. 2011. Tese (Doutorado) — Universidade de São Paulo, 2011.

LOPES, Fabrício M.; JR., David C. Martins; BARRERA, Junior; JR., Roberto M. Cesar. A feature selection technique for inference of graphs from their known topological properties: Revealing scale-free gene regulatory networks. **Information Sciences**, v. 272, n. 0, p. 1–15, 2014. ISSN 0020-0255. Disponível em: http://www.sciencedirect.com/science/article/pii/S0020025514002023.

LUSCOMBE, Nicholas M; BABU, M Madan; YU, Haiyuan; SNYDER, Michael; TEICHMANN, Sarah A; GERSTEIN, Mark. Genomic analysis of regulatory network dynamics reveals large topological changes. **Nature**, Nature Publishing Group, v. 431, n. 7006, p. 308, 2004.

MATHUR, Sachin; DINAKARPANDIAN, Deendayal. Finding disease similarity based on implicit semantic similarity. **Journal of biomedical informatics**, Elsevier, v. 45, n. 2, p. 363–371, 2012.

MCKENNA, Aaron; HANNA, Matthew; BANKS, Eric; SIVACHENKO, Andrey; CIBULSKIS, Kristian; KERNYTSKY, Andrew; GARIMELLA, Kiran; ALTSHULER, David; GABRIEL,

Stacey; DALY, Mark *et al.* The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. **Genome research**, Cold Spring Harbor Lab, v. 20, n. 9, p. 1297–1303, 2010.

MENG, Jun; LI, Rui; LUAN, Yushi. Classification by integrating plant stress response gene expression data with biological knowledge. **Mathematical biosciences**, Elsevier, v. 266, p. 65–72, 2015.

NEWMAN, M. E. J. The structure and function of complex networks. **SIAM Review**, SIAM, v. 45, n. 2, p. 167–256, 2003.

OVASKA, Kristian. Using semantic similarities and csbl. go for analyzing microarray data. *In*: **Microarray Data Analysis**. *[S.l.]*: Springer, 2015. p. 105–116.

PAGèS, Hervé; CARLSON, Marc; FALCON, Seth; LI, Nianhua. **AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor**. *[S.l.]*, 2019. R package version 1.48.0.

RADA, Roy; MILI, Hafedh; BICKNELL, Ellen; BLETTNER, Maria. Development and application of a metric on semantic nets. **IEEE transactions on systems, man, and cybernetics**, IEEE, v. 19, n. 1, p. 17–30, 1989.

RADIVOJAC, Predrag; CLARK, Wyatt T; ORON, Tal Ronnen; SCHNOES, Alexandra M; WITTKOP, Tobias; SOKOLOV, Artem; GRAIM, Kiley; FUNK, Christopher; VERSPOOR, Karin; BEN-HUR, Asa *et al.* A large-scale evaluation of computational protein function prediction. **Nature methods**, Nature Publishing Group, v. 10, n. 3, p. 221, 2013.

RAVASZ, E.; SOMERA, A. L.; MONGRU, D. A.; OLTVAI, Z. N.; BARABÁSI, A.-L. Hierarchical organization of modularity in metabolic networks. **Science**, American Association for the Advancement of Science, v. 297, n. 5586, p. 1551–1555, 2002. ISSN 0036-8075. Disponível em: https://science.sciencemag.org/content/297/5586/1551.

RESNIK, Philip. Using information content to evaluate semantic similarity in a taxonomy. **arXiv preprint cmp-lg/9511007**, 1995.

RESNIK, Philip. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. **Journal of artificial intelligence research**, v. 11, p. 95–130, 1999.

SCHLICKER, Andreas; DOMINGUES, Francisco S; RAHNENFÜHRER, Jörg; LENGAUER, Thomas. A new measure for functional similarity of gene products based on gene ontology. **BMC bioinformatics**, BioMed Central, v. 7, n. 1, p. 302, 2006.

SCHLICKER, Andreas; LENGAUER, Thomas; ALBRECHT, Mario. Improving disease gene prioritization using the semantic similarity of gene ontology terms. **Bioinformatics**, Oxford University Press, v. 26, n. 18, p. i561–i567, 2010.

SHALON, Dari; SMITH, Stephen J; BROWN, Patrick O. A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization. **Genome research**, Cold Spring Harbor Lab, v. 6, n. 7, p. 639–645, 1996.

SHIRAI, Shota; ACHARYA, Susant Kumar; BOSE, Saurabh Kumar; MALLINSON, Joshua Brian; GALLI, Edoardo; PIKE, Matthew D.; ARNOLD, Matthew D.; BROWN, Simon Anthony. Long-range temporal correlations in scale-free neuromorphic networks. **Network Neuroscience**, v. 4, n. 2, p. 432–447, 2020. Disponível em: https://doi.org/10.1162/netn_a_00128.

SNUSTAD, D Peter; SIMMONS, Michael J. **Genetics: international student version**. *[S.l.]*: John & Wiley & Sons, Incorporated, 2012.

STELZL, Ulrich; WORM, Uwe; LALOWSKI, Maciej; HAENIG, Christian; BREMBECK, Felix H; GOEHLER, Heike; STROEDICKE, Martin; ZENKNER, Martina; SCHOENHERR, Anke; KOEPPEN, Susanne *et al.* A human protein-protein interaction network: a resource for annotating the proteome. **Cell**, Elsevier, v. 122, n. 6, p. 957–968, 2005.

TAO, Ying; SAM, Lee; LI, Jianrong; FRIEDMAN, Carol; LUSSIER, Yves A. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. **Bioinformatics**, Oxford University Press, v. 23, n. 13, p. i529–i538, 2007.

TIMÁR, G.; DOROGOVTSEV, S. N.; MENDES, J. F. F. Scale-free networks with exponent one. **Phys. Rev. E**, American Physical Society, v. 94, p. 022302, Aug 2016. Disponível em: https://link.aps.org/doi/10.1103/PhysRevE.94.022302.

VICENTE, Fabio F. R.; LOPES, Fabrício M. SFFS-WS: A feature selection algorithm exploring the small-world properties of GNs. *In*: **Pattern Recognition in Bioinformatics, Proceedings**. *[S.l.]*: Springer Berlin / Heidelberg, 2014. (Lecture Notes in Computer Science, v. 8626), p. 60–71. ISBN 9783319091914. 9th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB), Stockholm, Sweden.

WANG, James Z; DU, Zhidian; PAYATTAKOOL, Rapeeporn; YU, Philip S; CHEN, Chin-Fu. A new method to measure the semantic similarity of go terms. **Bioinformatics**, Oxford University Press, v. 23, n. 10, p. 1274–1281, 2007.

WANG, Zhong; GERSTEIN, Mark; SNYDER, Michael. Rna-seq: a revolutionary tool for transcriptomics. **Nature reviews genetics**, Nature Publishing Group, v. 10, n. 1, p. 57, 2009.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of small-world networks. **Nature**, v. 393, p. 440–442, 1998.

WOLTING, Cheryl; MCGLADE, C Jane; TRITCHLER, David. Cluster analysis of protein array results via similarity of gene ontology annotation. **BMC bioinformatics**, Springer, v. 7, n. 1, p. 338, 2006.

WU, Xiaomei; ZHU, Lei; GUO, Jie; ZHANG, Da-Yong; LIN, Kui. Prediction of yeast protein–protein interaction network: insights from the gene ontology and annotations. **Nucleic acids research**, Oxford University Press, v. 34, n. 7, p. 2137–2150, 2006.

YANG, Da; LI, Yanhui; XIAO, Hui; LIU, Qing; ZHANG, Min; ZHU, Jing; MA, Wencai; YAO, Chen; WANG, Jing; WANG, Dong *et al.* Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant go categories. **Bioinformatics**, Oxford University Press, v. 24, n. 2, p. 265–271, 2007.

YU, Guangchuang; LI, Fei; QIN, Yide; BO, Xiaochen; WU, Yibo; WANG, Shengqi. Gosemsim: an r package for measuring semantic similarity among go terms and gene products. **Bioinformatics**, Oxford University Press, v. 26, n. 7, p. 976–978, 2010.

ZAHA, Arnaldo; FERREIRA, HB; PASSAGLIA, LMP. Biologia molecular básica. 3a edição. **Editora Mercado Aberto Ltda, Porto Algre-RS**, 2003.

ZHAO, Chenguang; WANG, Zheng. Gogo: An improved algorithm to measure the semantic similarity between gene ontology terms. **Scientific reports**, Nature Publishing Group, v. 8, n. 1, p. 15107, 2018.

APÊNDICES

APÊNDICE A – MATRIZES DE SIMILARIDADES

As matrizes de similaridades apresentadas no apêndice A são resultados obtidos pelo PLAWSS, método apresentado nesse trabalho.

Tabela 41 – Matriz de similaridade para a via metabólica Mannose degradation utilizando termos GO associados processos biológicos

Processos biológicos				
	GLK1	**HXK1**	**HXK2**	**PMI40**
GLK1	1.000	0.906	0.702	0.114
HXK1	0.906	1.000	0.805	0.103
HXK2	0.702	0.805	1.000	0.085
PMI40	0.114	0.103	0.085	1.000

Tabela 42 – Matriz de similaridade para a via metabólica Mannose degradation utilizando termos GO associados funções moleculares

Funções moleculares				
	GLK1	**HXK1**	**HXK2**	**PMI40**
GLK1	1.000	0.802	0.802	0.034
HXK1	0.802	1.000	1.000	0.042
HXK2	0.802	1.000	1.000	0.042
PMI40	0.034	0.042	0.042	1.000

Tabela 43 – Matriz de similaridade para a via metabólica Mannose degradation utilizando termos GO associados componente celular

Componente celular				
	GLK1	**HXK1**	**HXK2**	**PMI40**
GLK1	1.000	0.610	0.454	0.415
HXK1	0.610	1.000	0.821	0.748
HXK2	0.454	0.821	1.000	0.702
PMI40	0.415	0.748	0.702	1.000

Tabela 44 – Matriz de similaridade para a via metabólica Mevalonate utilizando termos GO associados processos biológicos

Processos biológicos								
	ERG10	**ERG13**	**HMG1**	**HMG2**	**ERG12**	**ERG8**	**MVD1**	**IDI1**
ERG10	1.000	1.000	0.735	1.000	0.600	0.499	0.500	0.207
ERG13	1.000	1.000	0.735	1.000	0.600	0.499	0.500	0.207
HMG1	0.735	0.735	1.000	0.735	0.874	0.748	0.387	0.349
HMG2	1.000	1.000	0.735	1.000	0.600	0.499	0.500	0.207
ERG12	0.600	0.600	0.874	0.600	1.000	0.874	0.331	0.599
ERG8	0.499	0.499	0.748	0.499	0.874	1.000	0.278	0.507
MVD1	0.500	0.500	0.387	0.500	0.331	0.278	1.000	0.178
IDI1	0.207	0.207	0.349	0.207	0.599	0.507	0.178	1.000

Tabela 45 – Matriz de similaridade para a via metabólica Mevalonate utilizando termos GO associados funções moleculares

Funções moleculares								
	ERG10	**ERG13**	**HMG1**	**HMG2**	**ERG12**	**ERG8**	**MVD1**	**IDI1**
ERG10	1.000	0.151	0.016	0.016	0.049	0.058	0.024	0.027
ERG13	0.151	1.000	0.051	0.051	0.122	0.130	0.061	0.062
HMG1	0.016	0.051	1.000	1.000	0.032	0.039	0.036	0.040
HMG2	0.016	0.051	1.000	1.000	0.032	0.039	0.036	0.040
ERG12	0.049	0.122	0.032	0.032	1.000	0.345	0.044	0.047
ERG8	0.058	0.130	0.039	0.039	0.345	1.000	0.049	0.051
MVD1	0.024	0.061	0.036	0.036	0.044	0.049	1.000	0.051
IDI1	0.027	0.062	0.040	0.040	0.047	0.051	0.051	1.000

Tabela 46 – Matriz de similaridade para a via metabólica Mevalonate utilizando termos GO associados componente celular

Componente celular								
	ERG10	**ERG13**	**HMG1**	**HMG2**	**ERG12**	**ERG8**	**MVD1**	**IDI1**
ERG10	1.000	0.710	0.439	0.430	0.933	1.000	0.732	0.933
ERG13	0.710	1.000	0.496	0.425	0.559	0.710	0.163	0.559
HMG1	0.439	0.496	1.000	0.814	0.397	0.439	0.188	0.397
HMG2	0.430	0.425	0.814	1.000	0.400	0.430	0.253	0.400
ERG12	0.933	0.559	0.397	0.400	1.000	0.933	0.716	1.000
ERG8	1.000	0.710	0.439	0.430	0.933	1.000	0.732	0.933
MVD1	0.732	0.163	0.188	0.253	0.716	0.732	1.000	0.716
IDI1	0.933	0.559	0.397	0.400	1.000	0.933	0.716	1.000

Tabela 47 – Matriz de similaridade para a via metabólica Phenylalanine degradation utilizando termos GO associados funções moleculares

Funções moleculares												
	ARO8	**ARO9**	**ARO10**	**PDC1**	**PDC5**	**PDC6**	**SFA1**	**ADH1**	**ADH2**	**ADH3**	**ADH4**	**ADH5**
ARO8	1.000	0.716	0.081	0.051	0.051	0.051	0.039	0.034	0.028	0.028	0.028	0.028
ARO9	0.716	1.000	0.081	0.051	0.051	0.051	0.039	0.034	0.028	0.028	0.028	0.028
ARO10	0.081	0.081	1.000	0.646	0.646	0.618	0.071	0.059	0.048	0.048	0.048	0.048
PDC1	0.051	0.051	0.646	1.000	1.000	0.799	0.036	0.032	0.025	0.025	0.025	0.025
PDC5	0.051	0.051	0.646	1.000	1.000	0.799	0.036	0.032	0.025	0.025	0.025	0.025
PDC6	0.051	0.051	0.618	0.799	0.799	1.000	0.038	0.032	0.028	0.028	0.028	0.028
SFA1	0.039	0.039	0.071	0.036	0.036	0.038	1.000	0.479	0.601	0.601	0.601	0.601
ADH1	0.034	0.034	0.059	0.032	0.032	0.032	0.479	1.000	0.554	0.554	0.554	0.554
ADH2	0.028	0.028	0.048	0.025	0.025	0.028	0.601	0.554	1.000	1.000	1.000	1.000
ADH3	0.028	0.028	0.048	0.025	0.025	0.028	0.601	0.554	1.000	1.000	1.000	1.000
ADH4	0.028	0.028	0.048	0.025	0.025	0.028	0.601	0.554	1.000	1.000	1.000	1.000
ADH5	0.028	0.028	0.048	0.025	0.025	0.028	0.601	0.554	1.000	1.000	1.000	1.000

Tabela 48 – Matriz de similaridade para a via metabólica Phenylalanine degradation utilizando termos GO associados processos biológicos

Processos biológicos												
	ARO8	**ARO9**	**ARO10**	**PDC1**	**PDC5**	**PDC6**	**SFA1**	**ADH1**	**ADH2**	**ADH3**	**ADH4**	**ADH5**
ARO8	1.000	1.000	0.355	0.327	0.365	0.385	0.255	0.191	0.227	0.242	0.250	0.219
ARO9	1.000	1.000	0.355	0.327	0.365	0.385	0.255	0.191	0.227	0.242	0.250	0.219
ARO10	0.355	0.355	1.000	0.570	0.626	0.691	0.524	0.462	0.510	0.565	0.566	0.519
PDC1	0.327	0.327	0.570	1.000	0.911	0.632	0.320	0.571	0.323	0.349	0.468	0.416
PDC5	0.365	0.365	0.626	0.911	1.000	0.701	0.358	0.415	0.358	0.393	0.532	0.464
PDC6	0.385	0.385	0.691	0.632	0.701	1.000	0.398	0.393	0.616	0.410	0.413	0.448
SFA1	0.255	0.255	0.524	0.320	0.358	0.398	1.000	0.379	0.440	0.479	0.500	0.440
ADH1	0.191	0.191	0.462	0.571	0.415	0.393	0.379	1.000	0.695	0.713	0.503	0.863
ADH2	0.227	0.227	0.510	0.323	0.358	0.616	0.440	0.695	1.000	0.830	0.522	0.804
ADH3	0.242	0.242	0.565	0.349	0.393	0.410	0.479	0.713	0.830	1.000	0.615	0.848
ADH4	0.250	0.250	0.566	0.468	0.532	0.413	0.500	0.503	0.522	0.615	1.000	0.600
ADH5	0.219	0.219	0.519	0.416	0.464	0.448	0.440	0.863	0.804	0.848	0.600	1.000

Tabela 49 – Matriz de similaridade para a via metabólica Phenylalanine degradation utilizando termos GO associados componente celular

Componente celular												
	ARO8	**ARO9**	**ARO10**	**PDC1**	**PDC5**	**PDC6**	**SFA1**	**ADH1**	**ADH2**	**ADH3**	**ADH4**	**ADH5**
ARO8	1.000	0.732	0.889	0.592	0.732	1.000	0.830	0.567	1.000	0.432	0.448	0.732
ARO9	0.732	1.000	0.716	0.794	1.000	0.732	0.768	0.493	0.732	0.470	0.493	1.000
ARO10	0.889	0.716	1.000	0.716	0.716	0.889	0.789	0.587	0.889	0.363	0.350	0.716
PDC1	0.592	0.794	0.716	1.000	0.794	0.592	0.639	0.719	0.592	0.344	0.331	0.794
PDC5	0.732	1.000	0.716	0.794	1.000	0.732	0.768	0.493	0.732	0.470	0.493	1.000
PDC6	1.000	0.732	0.889	0.592	0.732	1.000	0.830	0.567	1.000	0.432	0.448	0.732
SFA1	0.830	0.768	0.789	0.639	0.768	0.830	1.000	0.552	0.830	0.764	0.802	0.768
ADH1	0.567	0.493	0.587	0.719	0.493	0.567	0.552	1.000	0.567	0.276	0.243	0.493
ADH2	1.000	0.732	0.889	0.592	0.732	1.000	0.830	0.567	1.000	0.432	0.448	0.732
ADH3	0.432	0.470	0.363	0.344	0.470	0.432	0.764	0.276	0.432	1.000	0.883	0.470
ADH4	0.448	0.493	0.350	0.331	0.493	0.448	0.802	0.243	0.448	0.883	1.000	0.493
ADH5	0.732	1.000	0.716	0.794	1.000	0.732	0.768	0.493	0.732	0.470	0.493	1.000

Tabela 50 – Matriz de similaridade para a via metabólica Removal of superoxide radicals utilizando termos GO associados funções moleculares

Funções moleculares				
	SOD1	**SOD2**	**CTT1**	**CTA1**
SOD1	1.000	1.000	0.352	0.352
SOD2	1.000	1.000	0.352	0.352
CTT1	0.352	0.352	1.000	1.000
CTA1	0.352	0.352	1.000	1.000

Tabela 51 – Matriz de similaridade para a via metabólica Removal of superoxide radicals utilizando termos GO associados processos biológicos

Processos biológicos				
	SOD1	**SOD2**	**CTT1**	**CTA1**
SOD1	1.000	0.390	0.105	0.197
SOD2	0.390	1.000	0.187	0.408
CTT1	0.105	0.187	1.000	0.563
CTA1	0.197	0.408	0.563	1.000

Tabela 52 – Matriz de similaridade para a via metabólica Removal of superoxide radicals utilizando termos GO associados componente celular

Componente celular				
	SOD1	**SOD2**	**CTT1**	**CTA1**
SOD1	1.000	0.640	0.464	0.423
SOD2	0.640	1.000	0.432	0.766
CTT1	0.464	0.432	1.000	0.338
CTA1	0.423	0.766	0.338	1.000

Tabela 53 – Matriz de similaridade para a via metabólica Valine degradation utilizando termos GO associados funções moleculares

Funções moleculares								
	BAT1	**BAT2**	**PDC1**	**PDC5**	**PDC6**	**SFA1**	**ADH4**	**ADH5**
BAT1	1.000	1.000	0.052	0.052	0.051	0.042	0.031	0.031
BAT2	1.000	1.000	0.052	0.052	0.051	0.042	0.031	0.031
PDC1	0.052	0.052	1.000	1.000	0.799	0.036	0.025	0.025
PDC5	0.052	0.052	1.000	1.000	0.799	0.036	0.025	0.025
PDC6	0.051	0.051	0.799	0.799	1.000	0.038	0.028	0.028
SFA1	0.042	0.042	0.036	0.036	0.038	1.000	0.601	0.601
ADH4	0.031	0.031	0.025	0.025	0.028	0.601	1.000	1.000
ADH5	0.031	0.031	0.025	0.025	0.028	0.601	1.000	1.000

Tabela 54 – Matriz de similaridade para a via metabólica Valine degradation utilizando termos GO associados processos biológicos

Processos biológicos								
	BAT1	**BAT2**	**PDC1**	**PDC5**	**PDC6**	**SFA1**	**ADH4**	**ADH5**
BAT1	1.000	1.000	0.339	0.339	0.326	0.337	0.358	0.309
BAT2	1.000	1.000	0.339	0.339	0.326	0.337	0.358	0.309
PDC1	0.339	0.339	1.000	1.000	0.779	0.389	0.408	0.374
PDC5	0.339	0.339	1.000	1.000	0.779	0.389	0.408	0.374
PDC6	0.326	0.326	0.779	0.779	1.000	0.398	0.413	0.448
SFA1	0.337	0.337	0.389	0.389	0.398	1.000	0.500	0.440
ADH4	0.358	0.358	0.408	0.408	0.413	0.500	1.000	0.600
ADH5	0.309	0.309	0.374	0.374	0.448	0.440	0.600	1.000

Tabela 55 – Matriz de similaridade para a via metabólica Valine degradation utilizando termos GO associados componente celular

Componente celular								
	BAT1	**BAT2**	**PDC1**	**PDC5**	**PDC6**	**SFA1**	**ADH4**	**ADH5**
BAT1	1.000	0.470	0.371	0.482	0.432	0.764	0.883	0.470
BAT2	0.470	1.000	0.823	1.000	0.732	0.768	0.493	1.000
PDC1	0.371	0.823	1.000	0.845	0.526	0.624	0.365	0.823
PDC5	0.482	1.000	0.845	1.000	0.598	0.721	0.503	1.000
PDC6	0.432	0.732	0.526	0.598	1.000	0.830	0.448	0.732
SFA1	0.764	0.768	0.624	0.721	0.830	1.000	0.802	0.768
ADH4	0.883	0.493	0.365	0.503	0.448	0.802	1.000	0.493
ADH5	0.470	1.000	0.823	1.000	0.732	0.768	0.493	1.000

ÍNDICE REMISSIVO