Infraestrutura computacional para avaliação da similaridade funcional composta entre microRNAs baseada em ontologias

Mariana Yuri Sasazaki

DISSERTAÇÃO APRESENTADA
AO
PROGRAMA INTERUNIDADES EM BIOINFORMÁTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Área de Concentração: Bioinformática
Orientador: Prof. Dr. Joaquim Cezar Felipe

– Ribeirão Preto – SP – Junho de 2014 –

Infraestrutura computacional para avaliação da similaridade funcional composta entre microRNAs baseada em ontologias

Esta é a versão original da dissertação elaborada
pela candidata Mariana Yuri Sasazaki,
tal como submetida à Comissão Julgadora
como requisito parcial para a obtenção
do título de Mestre em Ciências.
Área de Concentração: Bioinformática

Comissão Julgadora:

- Prof. Dr. Joaquim Cezar Felipe - FFCLRP - USP (orientador)
- Prof. Dr. _____ - _____ - _____
- Prof. Dr. _____ - _____ - _____

“Nossa maior fraqueza está em desistir.

O caminho mais certo para vencer é tentar mais uma vez.”

Thomas Edison

Aos meus pais, Amélia e Edson
Aos meus irmãos, Michele e Roger
Ao meu namorado, Marcos
Por toda dedicação e amor incondicional.

AGRADECIMENTOS

Antes de mais, agradeço a Deus que ao longo de toda minha vida caminhou comigo, marcando ao longo do caminho duas pegadas andando lado a lado e que, nos momentos mais difíceis de minha caminhada, nos quais pude ver apenas uma pegada no caminho, não me abandonou, mas me “carregou no colo”.

Agradeço, com todo amor e sinceridade, aos meus pais, Amélia e Edson, e aos meus irmãos, Michele e Roger, por terem me ajudado a formar a pessoa que sou, por terem sempre com muito amor me apoiado e me ajudado a ultrapassar cada etapa de minha vida. Não poderia deixar de agradecer pelos conselhos e palavras amigas e reconfortantes que não me deixaram desistir do que quer que fosse e por nunca terem deixado de acreditar.

Agradeço também a todos os meus familiares, que me ensinaram que a família é o bem mais precioso que temos e que sempre com muita união e carinho estiveram presentes em minha jornada.

Ao meu namorado, Marcos, pelo carinho, ternura, companheirismo e incentivo de todos os dias, sem os quais com certeza não teria finalizado este trabalho.

Aos meus amigos Camila Martins, Camila Ottaiano, Gabriela Araujo, Juliana Takitane, Márcio Dantas e Melina Fushimi que sempre foram uma família pra mim, obrigada pela convivência e crescimento que vocês me proporcionaram. Obrigada pela amizade sincera independentemente da distância.

Ao meu orientador, Prof. Joaquim Cezar Felipe, pela apresentação deste tema e que me levou a aprofundar um pouco mais os meus conhecimentos nesta área. Obrigada por toda a atenção, disponibilidade e motivação demonstradas, que permitiram a conclusão deste projeto.

À Patrícia Martorelli pela gentileza, disponibilidade e amizade, sempre com muita boa vontade, nos auxiliando nas questões administrativas e burocráticas.

Muito Obrigada!

RESUMO

MicroRNAs (miRNAs) são pequenos RNAs não codificadores de proteínas que atuam principalmente como silenciadores pós-transcricionais, inibindo a tradução de RNAs mensageiros. Evidências crescentes revelam que tais moléculas desempenham papéis críticos em muitos processos biológicos importantes. Uma vez que não existem anotações de termos de miRNAs na *Gene Ontology* (GO), tampouco um banco de dados de referência com anotações funcionais dos mesmos, o cálculo da medida de similaridade entre miRNAs de forma direta não possui um padrão estabelecido. Por outro lado, a existência de bancos de dados de genes-alvo de miRNAs, como o TarBase, e bases de dados contendo informações sobre associações de miRNAs e doenças humanas, como o HMDD, nos permite inferir a similaridade funcional dos miRNAs indiretamente, por meio da análise de seus genes-alvo na GO ou entre suas doenças relacionadas na ontologia MeSH. Além disso, de acordo com a estrutura da ontologia de miRNAs OMIT, um miRNA também pode ser anotado com outras informações, tais como a sua natureza de atuação como oncogênico ou supressor de tumor, o organismo em que se encontra, o tipo de experimento em que foi encontrado, suas associações com doenças, genes-alvo, proteínas e eventos patológicos. Dessa forma, a similaridade entre miRNAs pode ser inferida com base na combinação de um conjunto de informações contidas nas respectivas anotações, de forma que possamos obter um aproveitamento de várias informações existentes, definindo assim um cálculo de similaridade funcional "composta". Assim, neste trabalho, propomos a criação e aplicação de um método chamado *CFSim*, aplicado sobre a OMIT e que utiliza a ontologia de doenças, MeSH, e a ontologia de genes, GO, para calcular a similaridade entre dois miRNAs, juntamente com informações contidas em suas anotações. A validação de nosso método foi realizada por meio da comparação com a similaridade funcional inferida considerando diferentes famílias de miRNAs e os resultados obtidos mostraram que nosso método é eficiente, no sentido de que a similaridade entre miRNAs pertencentes à mesma família é maior que a similaridade entre miRNAs de famílias distintas. Ainda, em comparação com os métodos de similaridade funcional já existentes na literatura, o *CFSim* obteve melhores resultados. Adicionalmente, para tornarmos viável a utilização do método proposto, foi projetado e implementado um ambiente contendo a infraestrutura necessária para que pesquisadores possam incluir dados obtidos de novas descobertas e consultar as informações sobre um determinado miRNA, assim como calcular a similaridade entre dois miRNAs, baseada no método proposto.

Palavras-chave: MicroRNAs, Similaridade funcional, MeSH, Gene Ontology, OMIT

ABSTRACT

MicroRNAs (miRNAs) are small non-coding RNA that mainly negatively regulate gene expression by inhibiting translation of target RNAs. Increasing evidences show that such molecules play critical roles in many important biological processes. Since there are no terms of miRNAs annotations in Gene Ontology (GO), nor a database with microRNAs functional annotations, directly calculating the functional similarity between miRNAs does not have an estabilished pattern aproach. However, the existence of miRNAs target genes database, such as TarBase, and a miRNAs-disease associations database, such as HMDD, allow us to indirectly infer functional similarity of miRNAs through the analysis of their target genes in GO or between their related diseases in MeSH. Moreover, according to the structure of the ontology of miRNAs OMIT, a miRNA can also be annotated with other information, such as if it acts as an oncogene or a tumor suppressor, the organism that it belongs, the experiment in which it was found, its associations with diseases, target genes, proteins and pathological events. Thus, miRNAs similarity can be inferred based on the combination of a broad set of information contained in their annotations, indeed, we can use all available information defining the calculation of a "composed" functional similarity. In this study, we propose the creation and application of *CFSim* method applied to the OMIT using the diseases ontology, MeSH, and gene ontology, GO, to compute miRNAs similarity based on different information in their annotations. We validated our method by comparing with functional similarity inferred by miRNA families and the results showed that our method is efficient in sense that the functional similarity between miRNAs in the same family was greater compared to other miRNAs from distinct families. Furthermore, in comparison with existing methods of functional similarity in the literature until the present day, the *CFSim* showed better results. Finally, to make feasible the use of the proposed method, an environment was designed and implemented, containing the necessary infrastructure so that researchers can include data from new discoveries and see information about a particular miRNA, as well as calculate the similarity between two miRNAs, based in the proposed method.

Keywords: MicroRNAs, functional similarity, MeSH, Gene Ontology, OMIT

SUMÁRIO

LISTA DE FIGURAS I

LISTA DE TABELAS II

LISTA DE EQUAÇÕES III

LISTA DE ABREVIAÇÕES IV

CAPÍTULO 1. INTRODUÇÃO 1

1.1. CONTEXTUALIZAÇÃO 1

1.2. MOTIVAÇÃO E OBJETIVOS 1

1.3. ORGANIZAÇÃO DO TEXTO 3

CAPÍTULO 2. FUNDAMENTOS TEÓRICOS 4

2.1. MICRORNAS 4

2.1.1. DEFINIÇÃO 4

2.1.2. BIOGÊNESE 5

2.1.3. FUNÇÕES 6

2.1.4. FAMÍLIAS DE MIRNAS 7

2.1.5. NOMENCLATURA 7

2.2. ONTOLOGIAS 8

2.2.1. DEFINIÇÃO 8

2.2.2. ONTOLOGIAS BIOMÉDICAS 9

2.3. SIMILARIDADE SEMÂNTICA EM ONTOLOGIAS 11

2.3.1. COMPARAÇÃO ENTRE TERMOS 11

2.3.2. COMPARAÇÃO ENTRE PRODUTOS GÊNICOS 13

2.3.3. MEDIDAS DE SIMILARIDADE ENTRE TERMOS 14

2.3.4. MEDIDAS DE SIMILARIDADE ENTRE PRODUTOS GÊNICOS 19

CAPÍTULO 3. TRABALHOS CORRELATOS 21

3.1. MISIM 21

3.2. GOSEMSIM 22

CAPÍTULO 4. MATERIAIS 23

4.1. ONTOLOGIAS 23

4.1.1. *ONTOLOGY FOR MICRORNA TARGET (OMIT)* 23

4.1.2. *GENE ONTOLOGY (GO)* 24

4.1.3. *MEDICAL SUBJECT HEADINGS (MESH)* 25

4.2. BASES DE DADOS 26

4.2.1. *MIRBASE* 26

4.2.2. *HUMAN MICRORNA DISEASE DATABASE – HMDD* 27

4.2.3. *TARBASE* 27

4.2.4. *GENETIC ASSOCIATION DATABASE (GAD)* 27

4.2.5. *HUMAN MICRORNA ONCOGENES AND TUMOR SUPPRESSORS* 28

4.3. CFSIM DATABASE 29

4.3.1. AMBIENTE DE DESENVOLVIMENTO 31

5. MÉTODO PROPOSTO 32

5.1. Proposta **32**
5.2. Implementação **34**
5.2.1. Determinação das Categorias 34
5.2.2. Cálculo das Similaridades das Categorias 35
5.2.3. Montagem do Vetor 37
5.2.4. Combinação das Similaridades das Categorias 37
5.3. Ambiente de Desenvolvimento **40**

CAPÍTULO 6. FRAMEWORK WEB **41**
6.1. Proposta **41**
6.2. Implementação **41**
6.2.1. Ambiente de Desenvolvimento 42

CAPÍTULO 7. RESULTADOS **43**
7.1. Método de Similaridade Funcional Composta CFSim **43**
7.2. Web Framework **48**

CAPÍTULO 8. CONCLUSÕES E TRABALHOS FUTUROS **50**
8.1. Conclusões **50**
8.2. Trabalhos Futuros **52**

REFERÊNCIAS BIBLIOGRÁFICAS **53**

APÊNDICE **62**

INTERFACES E FUNCIONAMENTO DO SISTEMA WEB **62**

LISTA DE FIGURAS

Figura 1. Crescimento do número de associações miRNA-doença no HMDD ao longo dos anos 2
Figura 2. Processamento do miRNA.....6
Figura 3. Diferentes tipos de Ontologias e seus relacionamentos.....9
Figura 4. Principais abordagens utilizadas na literatura para comparação entre termos 11
Figura 5. Principais abordagens utilizadas na literatura para comparação entre produtos gênicos 13
Figura 6. Possíveis configurações a, ß e γ num grafo acíclico dirigido.....18
Figura 7. Conceito de MiRNA e seus relacionamentos na OMIT.....24
Figura 8. Parte da GO mostrando as três sub-ontologias (função molecular, processo biológico e componente celular) e alguns de seus termos descendentes.....25
Figura 9. Grafo Acíclico Direcionado para Neoplasma de Mama.....26
Figura 10. Esquematização do método proposto.....33
Figura 11. Pseudo-Código do módulo Determinar Categorias.....35
Figura 12. Representação da matriz de genes-alvo (MG) associados a M^1 e M^2 e da matriz de termos (MT) que representam cada par de genes-alvo de M^1 e M^2. 37
Figura 13. Fator de correlação entre a similaridade de genes-alvo e a similaridade de doenças utilizando dados da base de dados GAD.....38
Figura 14. Fator de correlação entre a similaridade de genes-alvo e a similaridade de vias metabólicas utilizando dados da base de dados KEGG.....44
Figura 15. Página inicial "*Home*" do framework web.....62
Figura 16. Página inicial "*Home*" do framework web com busca sobre miRNA hsa-let-7b...63
Figura 17. Exemplo de cálculo do *CFSim* aplicado sobre os miRNAs hsa-let-7d e hsa-let-7g representando o caso em que os dois miRNAs pertencem a uma mesma família e ambos possuem anotações sobre a categoria de genes-alvo e também sobre a categoria de doenças.64
Figura 18. Exemplo de cálculo do *CFSim* aplicado sobre os miRNAs hsa-let-7d e hsa-mir-221 representando o caso em que os miRNAs pertencem a diferentes famílias, mas ambos possuem anotações sobre a categoria de genes-alvo e também sobre a categoria de doenças.65
Figura 19. Exemplo de cálculo do *CFSim* aplicado sobre os miRNAs hsa-miR-29b-3p e hsa-miR-29c-3p representando o caso em que os dois miRNAs pertencem a uma mesma família e ambos possuem anotações somente sobre a categoria de genes-alvo.....65
Figura 20. Exemplo de cálculo do *CFSim* aplicado sobre os miRNAs hsa-miR-29b-3p e hsa-mir-148a-3p representando o caso em que os miRNAs pertencem a diferentes famílias, mas ambos possuem anotações somente sobre a categoria de genes-alvo.....66
Figura 21. Exemplo de cálculo do *CFSim* aplicado sobre os miRNAs hsa-mir-128-1 e hsa-mir-128-2 representando o caso em que os dois miRNAs pertencem a uma mesma família e ambos possuem anotações somente sobre a categoria de doenças.....66
Figura 22. Exemplo de cálculo do *CFSim* aplicado sobre os miRNAs hsa-mir-128-1 e hsa-mir-301a representando o caso em que os miRNAs pertencem a diferentes famílias, mas ambos possuem anotações somente sobre a categoria de doenças.....67
Figura 23. Aba "Insert / Delete" do framework web.....68
Figura 24. Janela de inserção de dados de associação entre o miRNA e doenças humanas...68
Figura 25. Janela na qual o usuário pode escolher uma doença já existente no *CFSim Database* para associar a um miRNA.....69
Figura 26. Janela com informações detalhadas sobre uma publicação.....69
Figura 27. Janela com informações sobre o método *CFSim* e o framework apresentadas na aba "*About*".....70

LISTA DE TABELAS

Tabela 1. Exemplos de MiRNAs agrupados por família, com suas respectivas naturezas oncogenica e/ou supressora de tumor e quantidades de genes-alvo e de doenças associados 45
Tabela 2. Análise comparativa da média das similaridades funcionais entre miRNAs obtidas pelos métodos MISIM, GOSemSim, pela combinação de ambos e pelo método proposto *CFSim*.....47

LISTA DE EQUAÇÕES

Equação 1. Conteúdo da Informação (IC, do inglês "*Information Content*").....12
Equação 2. Similaridade definida por Resnik.....14
Equação 3. Similaridade definida por Lin.....14
Equação 4. Similaridade definida por Jiang e Conrath.....14
Equação 5. Similaridade Relativa definida por Schlicker.....15
Equação 6. Definição de Ancestrais Disjuntos de um termo.....15
Equação 7. Definição de Ancestrais Disjuntos Comuns de dois termos.....15
Equação 8. GraSM - Informação Compartilhada definida por Couto.....16
Equação 9. Similaridade definida por Pekar e Staab.....16
Equação 10. Especificidade (α) do ancestral comum mais baixo de dois termos.....17
Equação 11. Distância (β) ao nó folha mais próximo.....17
Equação 12. Distância (γ) ao ancestral comum mais baixo.....17
Equação 13. Similaridade Específica Relativa (RSS).....18
Equação 14. Contribuição Semântica de um ancestral para um termo.....18
Equação 15. Valor Total de Contribuições Semânticas de todos ancestrais de um termo.....19
Equação 16. Similaridade Semântica entre dois termos definia por D. Wang.....19
Equação 17. Índice de Jaccard.....20
Equação 18. Contribuições Semânticas de doenças similares.....21
Equação 19. *MiRNA Similarity* (MISIM) definida por J. Z. Wang.....21
Equação 20. *Gene Ontology Semantic Similarity* (GOSemSim)definida por Yu.....22
Equação 21. *Composed Functional Similarity* (*CFSim*).....38
Equação 22. Similaridade entre Naturezas de miRNAS (simN).....39
Equação 23. *Composed Functional Similarity* (*CFSim*) com pesos definidos.....40
Equação 24. Similaridade entre vias metabólicas associadas a dois genes.....43

LISTA DE ABREVIAÇÕES

CFSim = Composed Functional Similarity
DAG = *Directed Acyclic Graph*
DCA = *Disjoint Commom Ancestors*
DNA = *Deoxyribonucleic Acid*
FMA = *Foundational Model of Anatomy*
GAD = *Genetic Association Database*
GO = *Gene Ontology*
GOSemSim = *Gene Ontology Semantic Similarity*
HMDD = *Human MicroRNA Disease Database*
IC = *Information Content*
KEGG = *Kyoto Encyclopedia of Genes and Genomes*
MeSH = *Medical Subject Headings*
MICA = *Most Informative Commom Ancestor*
miRBase = *MicroRNA Database*
miRNA = *Micro Ribonucleic Acid*
MISIM = *MicroRNA Similarity*
MRCA = *Most Recent Common Ancestor*
NLM = *National Library of Medicine*
NTO = *Normalized Term Overlap*
OMIT = *Ontology for MiRNAs Target*
RNA = *Ribonucleic Acid*
RNAse = *Ribonuclease*
RISC = *RNA-Induced Silence Complex*
RSS = *Relative Specific Similarity*
Tarbase = Target Database
TO = *Term Overlap*
UMLS = *Unified Medical Language System*

CAPÍTULO 1. Introdução

Esta seção contém uma breve contextualização sobre o tema abordado neste projeto, assim como a motivação e o objetivo do mesmo.

1.1. Contextualização

MicroRNAs (miRNAs) constituem uma classe de pequenos RNAs não codificadores de proteínas, com aproximadamente 20 a 25 nucleotídeos de comprimento, encontrados em diversos organismos, como animais, plantas, vírus, etc. São moléculas conservadas ao longo da evolução que atuam como silenciadores pós-transcricionais, sendo por isso capazes de regular a expressão gênica através da degradação de moléculas-alvo de RNA mensageiro ou da repressão da tradução dos mesmos [Ambros-2004; Filipowicx-2008]. Os miRNAs desempenham papéis críticos em muitos processos biológicos importantes, tais como o crescimento celular, diferenciação dos tecidos, proliferação celular, desenvolvimento embrionário e apoptose. Sendo assim, a mutação, a desregulação e a disfunção da biogênese destas moléculas podem resultar em várias doenças, tal como o câncer [Esquela-Kersher-2006].

Com o intuito de facilitar a aquisição de conhecimentos a partir de fontes já existentes no contexto de miRNAs, foram criados bancos de dados que relacionam miRNAs e doenças, como o *Human MicroRNA Disease Database* (HMDD) [Lu-2008], e bancos que possuem registrados os sítios-alvo de miRNAs, como o *Tarbase* [Sethupathy-2006]. Além disso, recentemente foi criada a primeira ontologia no domínio dos miRNAs, a *Ontology for MiRNAs Target* (OMIT) [Huang-2010]. Todos estes avanços na pesquisa de miRNAs têm motivado ainda mais os estudos para identificação e caracterização da influência dos miRNAs em diversas doenças humanas, com destaque ao câncer, e a previsão de genes-alvos dos miRNA, tornando este ramo de pesquisa cada vez mais ativo e relevante na área de bioinformática e em especial aos pesquisadores que desenvolvem trabalhos relacionados ao câncer.

1.2. Motivação e Objetivos

Recentemente, vários estudos têm sido conduzidos na tentativa de medir a similaridade funcional entre genes, principalmente genes codificadores de proteínas. Uma das metodologias utilizadas para inferir a similaridade funcional entre os genes codificadores de proteínas é por meio da similaridade de sequencia. Uma segunda forma de se obter avaliações

de similaridade entre genes é por meio da similaridade de expressão gênica [Horvath-2006; Lin-2007]. Outra forma é através do uso das anotações da GO [Du-2009; Lee-HK-2004; Wang-2007; Lord-2003; Pesquita-2008; Pesquita-2009].

Apesar de métodos de similaridade entre genes codificadores de proteínas baseados na GO terem alcançado bons resultados [Du-2009], o cálculo da medida de similaridade entre miRNAs de forma direta não é muito eficiente, pois não existem anotações de termos de miRNAs na GO, tampouco um banco de dados com anotações funcionais dos mesmos, uma vez que a maioria das funções dos miRNAs ainda não estão bem definidas [Wang-2010b]. Por outro lado, uma grande quantidade de genes-alvo dos miRNAs já estão registrados em bancos de dados, como o *TarBase* [Sethupathy-2006], por exemplo. Isso nos permite inferir a similaridade funcional dos miRNAs indiretamente, por meio da análise de seus genes-alvo na GO [Ren-2009; Roubelakis-2009].

Por outro lado, muitas vezes as informações encontradas sobre miRNAs consistem em relacioná-los diretamente com doenças afetadas pelos mesmos. Recentemente, a quantidade de registros de associações miRNA-doença tem crescido consideravelmente [Lu-2008], conforme ilustrado na Figura 1; por exemplo, atualmente, o banco de dados de doenças humanas de miRNAs (HMDD) possui 5076 associações miRNA-doença [HMDD-2014]. Nesse caso, a similaridade entre miRNAs poderia ser inferida a partir da similaridade entre as doenças relacionadas ao mesmo. Esta similaridade, por sua vez pode ser calculada pela posição dos termos em uma ontologia de doenças, como, por exemplo, a MeSH [MESH-2014].

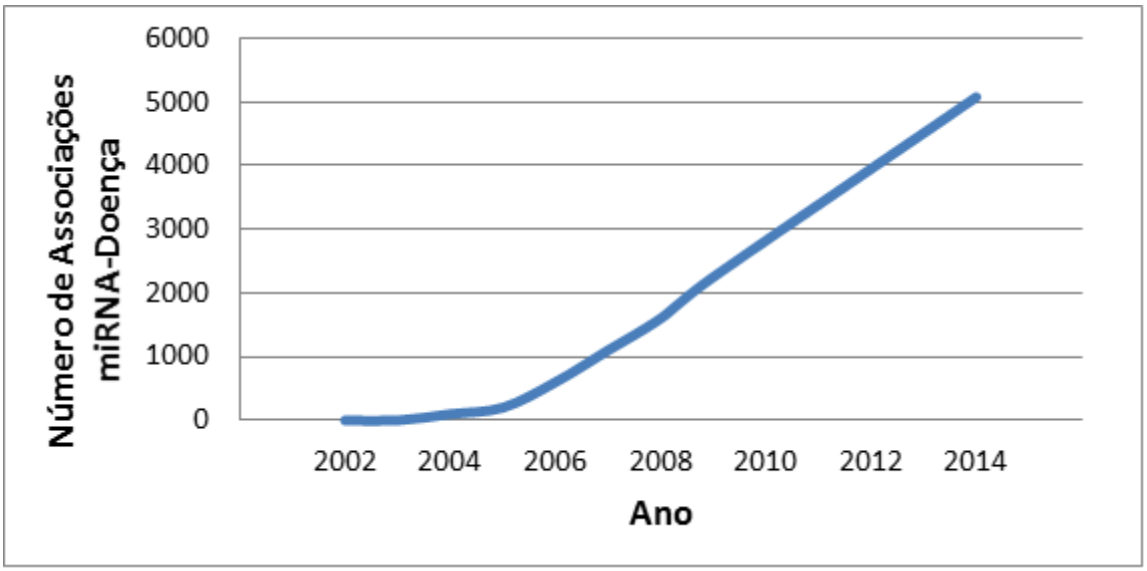

Figura 1. Crescimento do número de associações miRNA-doença no HMDD ao longo dos anos.

Se considerarmos a estrutura da OMIT, um miRNA pode ser anotado também com outras informações, além dos genes-alvo e das doenças relacionadas, tais como a sua natureza de atuação como oncogênico ou supressor de tumor, o organismo em que se encontra, o tipo de experimento em que foi encontrado, suas associações com proteínas e eventos patológicos, sendo que estas anotações podem estar incompletas, ou redundantes. Dessa forma, a similaridade entre miRNAs pode ser inferida baseando-se na similaridade de genes-alvo, ou de doenças associadas, complementada com outras informações contidas nas respectivas anotações, ou então da combinação de informações, de forma que possamos obter um valor de similaridade mais efetivo e preciso, caracterizando assim um cálculo de similaridade funcional "composta".

Assim, neste trabalho, com o objetivo de auxiliar pesquisadores da área de Biomedicina a compreender os importantes papéis dos miRNAs, suas funções e suas associações a diversas doenças humanas, e principalmente o inter-relacionamento entre os mesmos, apresentamos a criação e aplicação de um método para calcular a similaridade entre miRNAs, baseado nas diferentes informações contidas em suas anotações. Este método é aplicado sobre a ontologia de miRNAs OMIT utilizando a ontologia de doenças MeSH e a ontologia de genes *Gene Ontology*.

Para tornar viável a utilização do método proposto, foi projetado e implementado um ambiente contendo a infraestrutura necessária para que pesquisadores possam incluir dados obtidos de novas descobertas e consultar as informações sobre um determinado miRNA, assim como calcular a similaridade entre dois miRNAs, baseada no método proposto.

1.3. Organização do texto

Esta monografia está organizada da seguinte forma: a Seção 2 apresenta os fundamentos teóricos relacionados ao contexto deste projeto. A Seção 3 apresenta uma breve descrição de alguns trabalhos relacionados. A Seção 4 relata os materiais que foram utilizados no desenvolvimento das atividades propostas. A Seção 5 descreve a metodologia proposta que foi desenvolvida. A Seção 6 apresenta a descrição do framework web desenvolvido. A Seção 7 contém os resultados obtidos com nosso método. A Seção 8 possui as coclusões e trabalhos futuros, e por fim, a Seção 9 contém as referências bibliográficas utilizadas.

CAPÍTULO 2. Fundamentos Teóricos

Este capítulo apresenta os principais conceitos teóricos abordados no trabalho. São descritos os conceitos de MicroRNA e de Ontologias assim como alguns métodos de similaridade baseados em ontologias já existentes na literatura.

2.1. MicroRNAs

MicroRNAs (miRNAs) são pequenas moléculas de RNA, conservadas ao longo da evolução, que regulam a expressão gênica. Sua recente descoberta tem despertado grande interesse de pesquisadores da área biológica e biomédica, uma vez que a compreensão de sua funcionalidade pode ajudar a elucidar diferentes processos biológicos aos quais estão relacionados. Assim, nesta seção serão descritos com maiores detalhes a definição, a biogênese e o funcionamento desta classe de pequenos RNAs.

2.1.1. Definição

Os miRNAs constituem uma classe de pequenos RNAs não codificadores de proteínas, com aproximadamente 20 a 25 nucleotídeos de comprimento, encontrados em diversos organismos, como animais [Lagos-Quintana-2001; Lau ET al. 2001; Lee-2001; Mourelatos-2002], plantas [Llave-2002; Park et al. 2002; Reinhart-2002] e até mesmo vírus. São moléculas conservadas ao longo da evolução, que atuam principalmente como silenciadores pós-transcricionais, sendo por isso capazes de regular a expressão gênica através da degradação de moléculas-alvo RNA mensageiro ou da repressão da tradução dos mesmos [Ambros-2004; Filipowicx-2008].

O primeiro miRNA, lin-4 (do inglês *lineage-deficient-4*), foi descoberto no ano de 1993 por Victor Ambros e seus colaboradores Rosalind Lee e Rhonda Feinbaum, sendo nesta época associado à regulação do desenvolvimento larval em *Caenorhabditis elegans* [Lee-1993]. Após sete anos, o miRNA let-7 também foi descoberto em *C. elegans*, relacionado a transição da fase larval a fase adulta [Reinhert-2000].

Posteriormente, foram encontradas ocorrências de miRNAs em diversos organismos como vermes, moscas, vírus e seres humanos [Lagos-Quintana et al., 2003]. Isto sugere que essas moléculas representam uma família de genes que evoluíram a partir de um antigo gene ancestral de um pequeno RNA.

Até os dias atuais, foram registrados 24521 precursores de miRNAs, 30424 miRNAs maduros de 206 espécies diferentes no miRBase (base de dados de miRNAs) [MIRBASE-2014].

2.1.2. Biogênese

A biogênese dos miRNAs, mostrada na Figura 2, inclui sua transcrição no núcleo celular, exportação para o citoplasma e subsequente processamento e maturação [Singh-2008].

O processo inicia-se com a ação da polimerase II, uma enzima responsável pela transcrição de genes que codificam proteínas [Cowland-2007; Sun-2008], realizando a transcrição de seu gene em um longo miRNA primário (pri-miRNA) [Lee-Y-2004]. Os nucleotídeos dos transcritos primários dos miRNAs formam estruturas secundárias, como as regiões "*stem*", em que dois segmentos de RNA com bases complementares são pareados, e as regiões "*loop*", nas quais os pares de bases não são complementares, constituindo, assim, alças circulares [Sun-2008]. O pri-miRNA é clivado ainda no núcleo pela ribonuclease (RNAse) III, Drosha, gerando uma molécula precursora de miRNA com aproximadamente 60 a 90 nucleotídeos, chamada de pré-miRNA [Lee-2003]. A proteína de exportação nuclear exportina-5 transporta este pré-miRNA para fora do núcleo [Llund-2003]. No citoplasma, o pré-miRNA é processado pela RNAse III, Dicer, que remove a alça na estrutura *stem-loop* [Schmittgen-2008; Sun-2008; Zhang-2007], resultando na formação de um RNA de fita dupla com aproximadamente 22 nucleotideos [Schmittgen-2008; Singh-2008]. Este dúplex de RNA é incorporado ao complexo de silenciamento induzido por RNA (RISC, do inglês *RNA-induced silence complex*), no qual as duas fitas de RNA são separadas pela ação da Helicase [Bartel-2009; Sun-2008; Zhang et al., 2007]. Uma destas fitas permanece associada ao RISC e constitui o miRNA maduro, enquanto que a fita complementar sofre degradação [Dalmay-2008; Schmittgen-2008; Sun-2008].

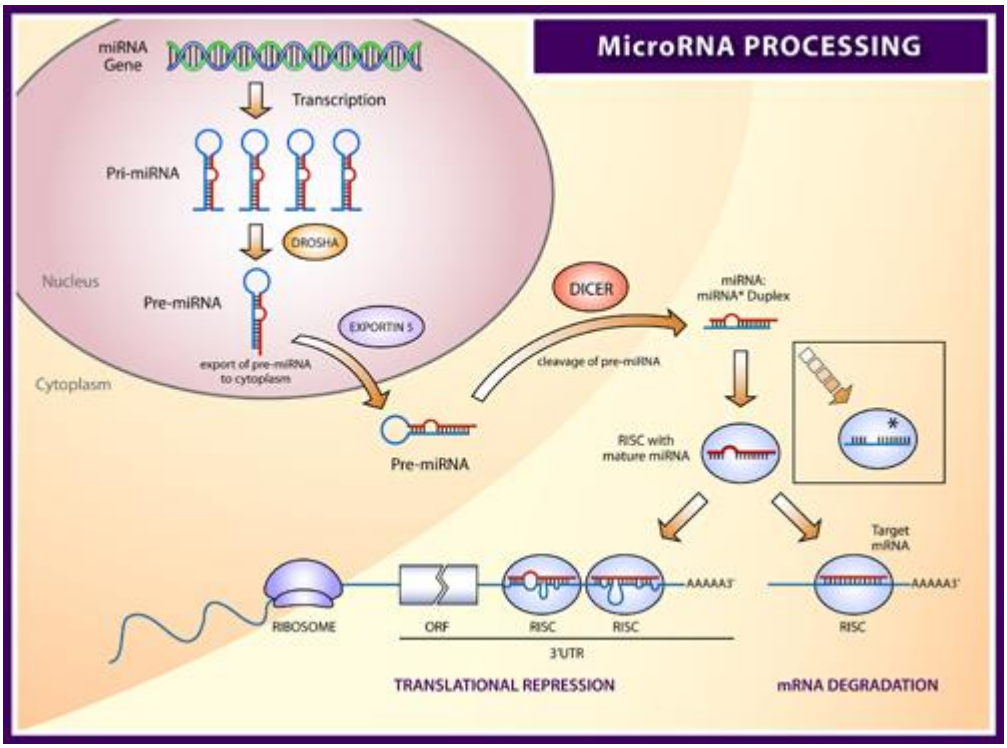

Figura 2. Processamento do miRNA.
(http://biologiaecologia1globalwarming.wordpress.com/tag/microrna/).

2.1.3. Funções

Através do pareamento com sítios 3'UTR da sequência de RNAm (região que não é traduzida em proteína, mas que contém bases de regulação da transcrição e tradução), os miRNAs são responsáveis pela regulação gênica por meio da degradação do RNAm ou repressão de sua tradução em proteínas. Caso o grau de complementaridade entre as bases do miRNA e do RNAm seja de 100%, haverá a clivagem do RNAm. Entretanto, se o pareamento for parcial, ocorrerá somente a repressão da tradução do RNAm [Sassen-2008; Sun-2008].

Entretanto, apesar de serem menos comuns, há ainda outras formas de regulação dos miRNAs. Por exemplo, existem miRNAs que agem positivamente em nível pós-transcricional, ativando a tradução dos genes ao invés de reprimi-la [Iwasaki-Tomari-2009]. Além disso, há miRNAs que realizam o pareamento com sequências codificadoras de proteínas (*Coding Sequence* - CDS), outros que se ligam a regiões promotoras do gene em DNAs e ainda, aqueles se pareiam com sequências de outros RNAs não codificadores [Duursma-2008; Weinmann-2009; Zisoulis-2012; Piovezani-2013].

Há evidências crescentes que revelam que os miRNAs desempenham papéis críticos em muitos processos biológicos importantes, tais como o crescimento celular, diferenciação dos tecidos, proliferação celular, desenvolvimento embrionário e apoptose. A mutação, a desregulação e a disfunção em sua biogênese podem resultar em várias doenças, tais como o

câncer. Dependendo do tipo celular e da forma em que são expressos, um mesmo miRNA pode exibir atividade oncogênica ou supressora tumoral. Os miRNAs oncogênicos encontram-se superexpressos em tumores malignos e atuam estimulando a proliferação celular e inibindo a ação de genes supressores tumorais e genes que controlam a apoptose. Por outro lado, alguns miRNAs apresentam expressão reduzida em tumores malignos, sendo considerados supressores tumorais. Estes miRNAs atuam contra o desenvolvimento de neoplasias através da inibição de oncogenes [Cowland-2007; Gartel-2008; Zhang-2007].

2.1.4. Famílias de miRNAs

Os miRNAs são agrupados em famílias de acordo com suas sequências precursoras, ou seja, miRNAs que pertencem a uma mesma família possuem sequências idênticas ou muito semelhantes [Griffiths-Jones-2010; Xu-2013].

Além disso, as regiões *seed* de sequências de miRNA da mesma família são idênticas. Essas regiões são compostas por normalmente 2 a 8 nucleótidos a partir da extremidade 5' do miRNA e são amplamente utilizadas como as regiões-chave para reconhecimento do gene-alvo do miRNA [Bartel-2009; Xuan-2013; Xu-2014].

Sendo assim, os membros de uma mesma família provavelmente regulam um conjunto comum de genes-alvo, portanto apresentam alta similaridade funcional e desempenham um papel-chave nos processos biológicos semelhantes, e consequentemente, é mais provável que estejam associados a doenças similares [Xuan-2013; Wang-2010-b; Yu-2007].

2.1.5. Nomenclatura

A atribuição do nome para um miRNA ocorre depois que este miRNA já foi validado experimentalmente e possui uma publicação relacionada [Piovezani-2013]. Os miRNAs são nomeados utilizando "miR" seguido de um único número identificador, por exemplo, miR-1, miR-2, ...miR-100, etc). Este número identificador é atribuído sequencialmente, sendo que miRNAs idênticos possuem o mesmo número, independentemente do organismo [Ambros-2003]. Para facilitar o reconhecimento pelo tipo de organismo, é atribuído um prefixo de 3 letras para identificar o mesmo. Por exemplo, no caso de *Homo Sapiens*, temos hsa-miR-121.

Enquanto os miRNAs maduros são nomeados com "R" maiúsculo, os genes que possuem como transcritos estes miRNAs, são referenciados com o "r" minúsculo, como por exemplo hsa-mir-121. Sequências precursoras distintas e diferentes *loci* genômicos que transcrevem sequências de miRNAs idênticas possuem nomes com sufixos numéricos, como hsa-mir-121-1 e hsa-mir-121-2. Já os sufixos em letras denotam sequências de miRNAs

estritamente relacionadas - por exemplo hsa-miR-121a e hsa-miR-121b seriam expressos a partir dos precursores hsa-mir-121a e hsa-mir-121b, respectivamente [Griffiths-Jones-2006].

Inicialmente, utilizava-se o sufixo asterisco (*) para indicar que as duas sequências que compõem a fita dupla do miRNA eram funcionais. Posteriormente, passaram a adotar os sufixos 3p e 5p para indicar se o miRNA foi originado da extremidade 3' ou 5' da fita, respectivamente.

Os miRNAs let-7 e lin-4 são exceções óbvias ao esquema de nomeação, sendo estes nomes mantidos devido a razões históricas.

2.2. Ontologias

Ontologia é um termo de origem filosófica e, embora tenha sido bastante confinada à Filosofia no passado, atualmente tem ganhado importante papel em áreas como Inteligência Artificial, Linguística Computacional e Teoria de Banco de Dados. Nesta seção serão apresentadas sua definição e sua ampla utilização no domínio biológico e biomédico.

2.2.1. Definição

De acordo com Gruber, ontologia é uma especificação explícita de uma conceitualização [Grubber-1995]. A conceitualização é uma visão abstrata simplificada do mundo que se deseja representar para algum propósito [Grubber-1995] e uma especificação explícita significa que o tipo de conceitos usados e restrições sobre o seu uso devem ser explicitamente definidos. Em 1997, Borst complementou a definição de Grubber, dizendo que ontologia é uma especificação formal de uma conceitualização compartilhada. Por "formal", Borst se referia ao fato de que uma ontologia deve ser passível de ser processada por uma máquina, e por "compartilhada" que a ontologia deve refletir o conhecimento consensual, aceito por um grupo e não restrito por alguns indivíduos [Borst-1997]. Um ano depois, Studer, Benjamins, Fensel [Studer-1998] unem ambas as definições anteriores, afirmando que uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada.

Em resumo, uma ontologia possui definições que associam nomes de entidades no universo do discurso (por exemplo, classes, relações, funções etc. com textos que descrevem o que os nomes significam e os axiomas formais que restringem a interpretação e o uso desses termos).

Ontologias têm sido aplicadas em diversas áreas de pesquisa, como Engenharia do Conhecimento, Integração Banco de Dados, Extração e Recuperação de Informações

[Guarino-1998]. Seu uso apresenta muitos benefícios [Noy McGuinnes-2001; Marietto-2002; Freitas-2003], tendo em destaque a possibilidade de comunicação entre pessoas, agentes e sistemas, já que permite reuso, mapeamento de formalismos, compartilhamento de conhecimentos, dentre outras funcionalidades.

De acordo com o nível de generalização, as ontologias são divididas em 4 classificações distintas [Guarino-1998], conforme ilustrado na Figura 3:

- *Ontologias de Alto Nível*: descrevem conceitos muito gerais, como por exemplo espaço, tempo, evento, os quais são independentes de um problema ou domínio particular.

- *Ontologias de Domínio*: descrevem o vocabulário relacionado a um determinado domínio, como medicina, veículos, documentos, entre outros.

- *Ontologias de Tarefas*: descrevem o vocabulário associado a uma determinada tarefa ou atividade, como diagnóstico, venda, etc.

- *Ontologias de Aplicação*: descreve conceitos que dependem tanto de domínios quanto de tarefas. Estes conceitos normalmente correspondem a regras aplicadas a entidades de domínio enquanto executam determinada tarefa.

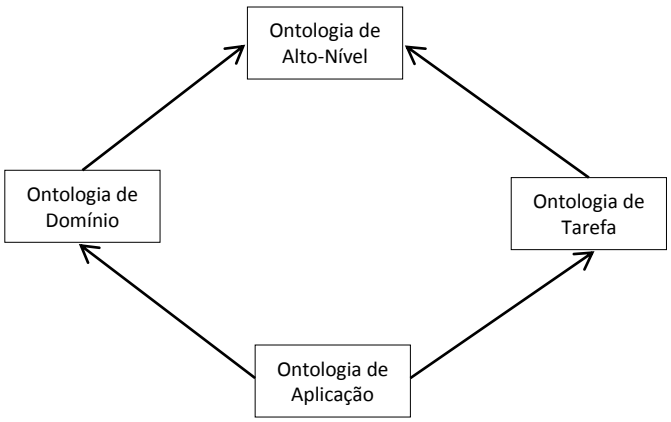

Figura 3. Diferentes tipos de Ontologias e seus relacionamentos.

Na próxima seção será descrito o uso das ontologias e seus benefícios na área de Biomedicina, assim como alguns exemplos das ontologias biomédicas mais famosas e mais amplamente utilizadas.

2.2.2. Ontologias Biomédicas

As bases de dados em ciências biomédicas têm crescido constantemente, [Vogt-2012] e isto faz com que seja difícil para os pesquisadores "ficarem a par" do conhecimento biomédico atual e compreenderem e interligarem essa quantidade maciça de informação *online*. Paralelamente a este crescimento foi surgindo a importância de organizar e padronizar

o conteúdo destas informações. Assim, as ontologias foram sendo reconhecidas dentro da comunidade de bioinformática [Schulze-Kremer-1998] por desempenharem papel importante neste processo de normalização, já que podem fornecer vocabulários controlados com explícitas definições e denominações em uma sintaxe altamente formalizada. Estas ontologias da área biomédica são chamadas bio-ontologias.

A primeira e mais conhecida bio-ontologia é a *Gene Ontology* (GO) criada pelo *consortium* de bioinformatas, liderados por Michael Ashburner e Judith Blake [Consortium, 2000]. A GO é composta por um conjunto de vocabulários controlados, estruturados e bem definidos para descrever domínios chave de biologia molecular, incluindo atributos de produtos gênicos e sequencias biológicas. Portanto, a GO possui três ontologias que representam diferentes domínios: componente celular, composto pelas partes de uma célula ou seu ambiente extracelular; função molecular, composto por atividades elementares de um produto gênico ao nível molecular, tais como ligação ou catálise; e processo biológico, composto pelas operações ou conjuntos de eventos moleculares com começo e fim definidos, pertinentes para o funcionamento de unidades vivas integradas, ou seja, células, tecidos, órgãos e organismos. A interação entre os termos de uma mesma ontologia e entre ontologias diferentes é feito por meio de um conjunto de 6 diferentes relacionamentos: *"is-a", "part_of", "regulates", "positively-regulates", "negatively-regulates"* e *"has_part"* [CONSORTIUM, 2010].

Devido ao sucesso da GO, novas ontologias foram subsequentemente propostas e desenvolvidas para cobrir outros domínios da biologia, tais como anatomia, estrutura, doença, fenótipo e caminhos.

O *Unified Medical Language System* (UMLS) foi criado pela *National Library of Medicine* dos EUA, sendo um tipo de meta-terminologia, subsumindo o conteúdo de outras terminologias. O objetivo da UMLS é construir uma fonte de conhecimento para auxiliar os profissionais da saúde a recuperar e a integrar informações biomédicas, amenizando os problemas de recuperação devido as diferenças de terminologias e dispersão das informações relevantes em diversos bancos de dados existentes.

Outro exemplo de ontologia é o Modelo Fundamental de Anatomia (FMA, do inglês *Foundational Model of Anatomy*), desenvolvido por Cornelius Rosse da Universidade de Washington, é uma ontologia que representa o conhecimento sobre a anatomia humana [Rosse-2003]. O FMA foi desenvolvido especificamente para fornecer conceitos e relações pertencentes a estruturas anatômicas humanas, com a intenção de expandir o conteúdo anatômico do UMLS [Cimino-2006].

O aumento rápido no número de ontologias foi criando obstáculos ao processo de integração das mesmas. Assim, foi criado o *Open Biological and Biomedical Ontologies* (OBO) *Foundry* para fornecer um conjunto de princípios para o desenvolvimento de ontologias, com o objetivo de criar um conjunto de ontologias ortogonais de referência interoperáveis e logicamente bem formadas na área biomédica [Smith et al., 2007; OBO]. A maioria das ontologias em OBO *Foundry* é conduzida por biólogos que tentam compreender as propriedades e as funções das diversas entidades biológicas [Mi-2011].

2.3. Similaridade Semântica em Ontologias

Com uso de ontologias é possível comparar suas entidades por meio de suas anotações. Esta comparação geralmente é feita implicitamente, como por exemplo, encontrando os termos em comum entre as entidades comparadas. Entretanto, uma comparação explícita pode ser feita utilizando medidas de similaridade semântica entre as entidades de uma ontologia.

2.3.1. Comparação entre Termos

Há uma série de abordagens disponíveis para quantificar a similaridade semântica entre os termos ou as entidades anotadas em uma ontologia representada por um grafo acíclico dirigido (DAG, do inglês "*Directed Acyclic Graph*"), como a GO, por exemplo. Com destaque para dois tipos de abordagens de comparação de termos: a baseada em arestas (*Edge-based*) e a baseada em nós (*Node-based*). Diferentes métricas, mostradas na Figura 4, proposta por Pesquita [Pesquita-2009], podem ser aplicadas nestas abordagens e serão descritas a seguir.

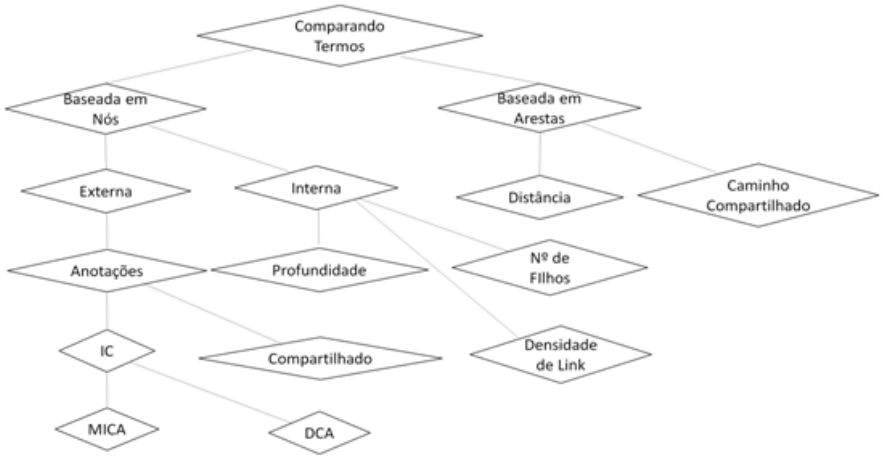

Figura 4. Principais abordagens utilizadas na literatura para comparação entre termos [Pesquita-2009].

- *Baseada em Arestas*: realiza a contagem do número de arestas existentes no caminho do grafo entre os dois termos que estão sendo comparados [Rada-1889]. A técnica de distância, mais comumente utilizada, seleciona o caminho mais curto ou a média de todos os caminhos, quando mais de um caminho existe. Já a técnica do caminho em comum calcula a similaridade diretamente pelo comprimento do caminho a partir do ancestral em comum mais próximo dos dois termos até o nó raiz [Wu-1994].

Ambas as técnicas citadas acima são baseadas no pressuposto de que arestas de mesmo nível na ontologia correspondem à mesma distância semântica entre os termos. Apesar destas duas técnicas serem bastante intuitivas, o pressuposto no qual são baseadas, raramente é verdadeiro no contexto das ontologias biológicas. Assim, várias estratégias foram propostas na tentativa de atenuar estas questões, tais como arestas de ponderação diferente, consoante a sua profundidade hierárquica, ou usando nó densidade e tipo de link [Richardson-1994].

- *Baseada em Nós:* compara as propriedades dos termos envolvidos, os quais podem ser relacionados com eles mesmos, seus antepassados ou seus descendentes. Nesta abordagem, geralmente é calculado o conteúdo da informação (IC, do inglês "*Information Content*"), o qual fornece a medida de quão específico e informativo é um termo. O IC pode ser calculado pela equação:

$$IC = -\log p(c) \qquad (1)$$

em que $p(c)$ é a probabilidade de ocorrência de um termo c em um banco de dados específico, sendo normalmente estimado pela sua frequência de anotação. Em uma taxonomia, a função $p(c)$ é monotônica crescente no sentido das folhas em direção à raiz, ou seja, se um termo c_1 'é-um' termo c_2, então $p(c_1) \leq p(c_2)$. Mais ainda, se a taxonomia possui um único nó topo, então sua probabilidade é igual a 1 e então seu IC é igual a zero. O uso do log negativo da probabilidade se deve ao sentido de que conforme a probabilidade aumenta, o conteúdo da informação diminui, sendo assim, quanto mais abstrato o termo, menor será seu conteúdo de informação [Resnik-1995]. O IC também pode ser representado pelo número de filhos que um termo possui na estrutura de uma ontologia [Seco-2004].

Conforme mostrado na Figura 4, na abordagem baseada em nós também podem ser aplicadas técnicas que utilizando o número de anotações compartilhadas [Budanistky-2001], o número de ancestrais em comum dentro da estrutura da ontologia e outras tipos de informações como a profundidade e densidade do nó [Othman-2007].

2.3.2. Comparação entre Produtos Gênicos

No contexto da GO, seus produtos gênicos são anotados por vários termos contidos em todas as suas três categorias, ou seja, um produto gênico tem suas funções descritas por diferentes termos de funções moleculares, participam de vários processos biológicos e são alocados em diferentes componentes celulares. Dessa forma, é necessário utilizar a comparação entre conjuntos de termos do que entre termos únicos para obter a similaridade funcional entre dois produtos gênicos. Para isso, existem duas abordagens possíveis de serem utilizadas: em pares (*Pairwise)* e em grupos (*Groupwise)*. Diferentes técnicas podem ser aplicadas em cada uma dessas abordagens, conforme mostra a Figura 5.

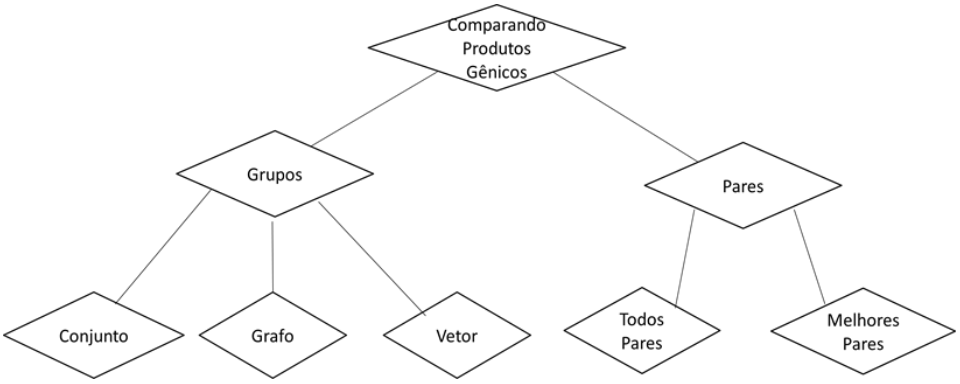

Figura 5. Principais abordagens utilizadas na literatura para comparação entre produtos gênicos [Pesquita-2009].

- *Em Pares*: Nesta abordagem, a similaridade funcional entre dois produtos gênicos é feita por meio da combinação das similaridades semânticas entre seus termos, já que cada produto gênico é representado pelo seu conjunto de anotações. Algumas técnicas utilizam todas as combinações entre os pares de termos dos dois conjuntos de anotações, enquanto outras consideram apenas o par de melhor correspondência para cada termo. Posteriormente, combinando essas similaridades semânticas por meio do calculo da média, máximo ou soma, é obtido um valor global de similaridade semântica.

- *Em Grupo*: Nesta abordagem, a similaridade funcional entre produtos gênicos pode ser calculada utilizando-se uma de três diferentes técnicas. A primeira é a técnica de conjuntos, na qual são consideradas somente as anotações diretas dos termos a serem comparados. A segunda é a técnica de grafos, a qual representa os produtos gênicos como subgrafos da GO que correspondem as suas anotações diretas e indiretas. E a última técnica é a de vetores, representando os produtos gênicos de interesse no espaço vetorial, com cada termo sendo representando por uma dimensão.

2.3.3. Medidas de Similaridade entre Termos

2.3.3.1 Baseada em Nós

As técnicas de medidas de similaridade mais comumente utilizadas na GO são: MICA (do inglês "*Most Informative Commom Ancestor*") e DCA (do inglês, "*Disjoint Commom Ancestors*"). Ambas quantificam a informação que dois termos compartilham, ou seja, calculam a similaridade semântica aplicando o conceito de IC ao ancestral comum entre estes termos. Entretanto, na técnica MICA somente o ancestral comum com o maior valor de IC é considerado [Resnik-1995], enquanto que na técnica DCA todos os ancestrais comuns disjuntos são considerados. Dois ancestrais comuns são disjuntos se forem caminhos independentes de ambos os ancestrais até o termo, ou seja, caminhos que utilizam pelo menos um termo da ontologia não utilizado pelos demais caminhos [Couto-2005].

Em ambas as técnicas MICA e DCA, o valor do IC é enviesado devido ao fato de que os termos relacionados às áreas de maior interesse são mais frequentemente anotados que outros termos. No entanto, a utilização do IC ainda faz sentido a partir de um ponto de vista probabilístico: é mais provável que dois termos compartilhem um termo mais comumente usado do que um termo raramente usado, independentemente de que o termo é comum por ser genérico ou porque está relacionado a um tema de pesquisa relevante [Pesquita-2009].

Resnik, Lin e Jiang e Conrath [Resnik-1995; Lin-1998; Jiang-Conrath-1997] calculam a similaridade utilizando medidas baseadas em nós e no valor de IC. Resnik utiliza somente o valor de IC do MICA como medida de similaridade entre dois termos:

$$sim_{Res}(c_1, c_2) = IC(c_{MICA}) = \max_{c \in S(c_1,c_2)} [-\log(p(c))], \quad (2)$$

onde $S(c_1, c_2)$ é o conjuto de termos que incluem ambos c_1e c_2.

Já Lin, Jiang e Conrath, [Lin-1998; Jiang-Conrath-1997] a fim de levar em consideração não só a informação compartilhada entre dois termos, mas também quão distantes estes termos estão de seu ancestral comum, relacionam o IC do MICA com o IC dos termos que estão sendo comparados:

$$\mathrm{sim_{Lin}}(\mathrm{c_1}, \mathrm{c_2}) = \frac{2\mathrm{xIC}(\mathrm{c_{MICA}})}{\mathrm{IC}(\mathrm{c_1}) + \mathrm{IC}(\mathrm{c_2})} \quad (3)$$

$$\mathrm{sim_{JC}}(\mathrm{c_1}, \mathrm{c_2}) = 1 - \mathrm{IC}(\mathrm{c_1}) + \mathrm{IC}(\mathrm{c_2}) - 2\mathrm{xIC}(\mathrm{c_{MICA}}) \quad (4)$$

A desvantagem de $\mathrm{sim_{Lin}}$ e $\mathrm{sim_{JC}}$ é que ambas medidas são deslocadas do grafo, ou seja, são proporcionais às diferenças de IC entre os termos e seu ancestral comum,

independentemente do valor de IC deste ancestral. A fim de proporcionar o posicionamento no grafo e levar em consideração a relevância da informação, uma vez que termos genéricos não possuem alta relevância para a comparação de diferentes produtos gênicos, Schlicker [Schlicker -2006] utiliza a medida de Lin [Lin-1998] juntamente com a probabilidade de anotação de MICA como um peso.

$$\mathrm{sim_{Rel}}(\mathrm{c_1}, \mathrm{c_2}) = \mathrm{sim_{Lin}}(\mathrm{c_1}, \mathrm{c_2}) \text{ x } (1 - \mathrm{p}(\mathrm{c_{MICA}})) \quad (5)$$

Couto [Couto-2005] propôs uma abordagem de medida de similaridade chamada GraSM, a qual é melhor adaptada à GO por considerar não somente um único ancestral em comum (MICA), mas sim os vários ancestrais disjuntos (DCAs) que os termos da GO podem conter. Ou seja, o fato de que dois ancestrais disjuntos de um termo representam duas interpretações distintas deste termo faz com que o cálculo da similaridade entre dois termos usando apenas o ancestral comum mais informativo represente apenas uma destas interpretações. Sendo assim, medidas de similaridade devem também contribuir para as outras interpretações de ambos os termos. Desta forma, GraSM considera que a_1 e a_2 representam os ancestrais disjuntos de um termo c se existe um caminho de a_1 para c não passando por a_2 e um caminho de a_2 para c não passando por a_1:

$$DisjAnc(c) = \{ (a_1, a_2) |$$

$$\Big(\exists p : \big(p \in Paths(a_1, c)\big) \wedge (a_2 \notin p)\Big) \wedge \Big(\exists p : \big(p \in Paths(a_2, c)\big) \wedge (a_1 \notin p)\Big) \} \quad (6)$$

Dados dois termos c_1 e c_2, seus ancestrais disjuntos comuns são o ancestral comum mais informativo dos ancestrais disjuntos de c_1 e c_2, ou seja, a_1 é um ancestral comum disjunto de c_1 e c_2 se, para cada ancestral a_2 mais informativo do que a_1, a_1 e a_2 são ancestrais disjuntos de c1 ou c2:

$$CommonDisjAnc(c_1, c_2) = \{a_1 |$$

$$a_1 \in CommonAnc(c_1, c_2) \wedge$$

$$\forall a_2 : \big[\big(a_2 \in CommonAnc(c_1, c_2)\big) \wedge \big(IC(a_1) \leq IC(a_2)\big)\big] \rightarrow$$

$$[(a_1, a_2) \in (DisjAnc(c_1) \cup DisjAnc(c_2))]\} \quad (7)$$

Assim, GraSM define a informação compartilhada como a média dos ICs do os ancestrais disjuntos comuns:

$$Share_{GraSM}(c_1, c_2) = \overline{\{\, IC(a) | a \in CommonDisjAnc(c_1, c_2)\}} \quad (8)$$

Ou seja, a abordagem GraSM pode ser aplicada a qualquer um dos métodos já acima descritos, substituindo o IC do MICA pelo IC médio de todos DCAs.

Bodenreider [Bodenreider, et, al., 2005] representa cada termo da GO como um vetor de todos os produtos gênicos anotados pelo termo e então, calcula a similaridade entre dois termos através do produto escalar de seus vetores. Desta forma, Bodenreider utiliza não somente a teoria da informação contida no IC, mas também utiliza os dados de anotações da ontologia.

2.3.3.2. Baseada em Arestas

Yu [Yu-2005] aplicou sobre a GO uma medida baseada no maior caminho entre o ancestral comum mais baixo entre dois termos e a raiz (Profundidade Máxima do Ancestral Comum), e no maior caminho entre cada um dos termos e aquele ancestral comum. Esta medida foi inicialmente proposta por Pekar e Staab [Pekar-Staab-2002] e é dada pela seguinte expressão:

$$\mathrm{sim}_{\mathrm{PS}}(\mathrm{c}_1, \mathrm{c}_2) = \frac{\delta(\mathrm{c_a}, \mathrm{root})}{\delta(\mathrm{c_a}, \mathrm{root}) + \delta(\mathrm{c}_1, \mathrm{c_a}) + \delta(\mathrm{c}_2, \mathrm{c_a})} \quad (9)$$

onde $\delta(\mathrm{c}_1, \mathrm{c}_2)$ é o numero de arestas do caminho mais longo entre termo c_1 e c_2, $\mathrm{c_a}$ é o ancestral comum mais baixo de c_1 e c_2, ou seja, o ancestral em comum de c_1 e $\mathrm{c}_2$que fica mais distante da raiz (root). A medida $\mathrm{sim}_{\mathrm{PS}}$ é diretamente proporcional ao número de arestas de um ancestral comum mais baixo até o nó raiz, o que concorda com a intuição de que um dado número de arestas entre duas classes específicas significa maior semelhança do que o mesmo número de arestas entre duas classes mais gerais.

A medida de Profundidade Máxima do Ancestral Comum também foi utilizada por Cheng [Cheng-2004], entretanto, ele adicionou pesos para cada aresta para refletir a profundidade, uma vez que, quanto mais perto um nó está da raiz, mais geral é o seu conceito e, portanto, maior quantidade de informação é transmitida pelas arestas de alto nível. Já Wu [Wu-2005] propôs a mesma medida, porém, sem pesar as arestas. Além disso, a fim de considerar a especificidade dos termos, Wu também introduziu na medida de Profundidade Máxima do Ancestral Comum a distancia (β) ao nó folha mais próximo e a distancia (γ) ao ancestral comum mais baixo (do inglês *Most Recent Common Ancestor*, MRCA) [Wu-2006]. Wu definiu a especificidade do ancestral comum mais baixo de dois termos c_1 e c_2(α) como sendo:

$$\alpha = \max_{\substack{\text{path}_m \in \text{Path}(c_1), \\ \text{path}_n \in \text{Path}(c_2)}} \{\text{n}^{\text{o}} \text{ termos comuns entre path}_m \text{ and path}_n\} - 1 \quad (10)$$

onde Path(c_1) e Path(c_2) correspondem aos conjuntos de caminhos do termo c_1 e c_2, respectivamente, até o termo raiz da ontologia.

De acordo com Wu [Wu-2006], *dist*(u,v) é o número de arestas do menor caminho entre o termo u e o termo v. Sendo assim, propôs que a generalidade de um termo é definida pela distancia mínima entre o termo e todas as suas descendentes folhas, ou seja, quanto maior a distancia, mais geral é o termo. Assim, β mede quão relativamente gerais são os termos c_1 e $c_2$na ontologia.

$$\beta = \max\left\{\min_{u \in U}\{dist(c_1, u)\},\ \min_{v \in V}\{dist(c_2, v)\}\right\} \quad (11)$$

onde U = {todos os nós folhas descendentes de c_1}, V = {todos os nós folhas descendentes de c_2}.

Wu definiu a distancia entre dois termos e seu MRCA) como:

$$\gamma = dist\ (\text{MRCA}, c_1) + dist(\text{MRCA}, c_2) \quad (12)$$

Se o valor de γ for baixo, isto significa que c_1 e c_2 compartilham muita similaridade localmente relativa ao MRCA.

A figura abaixo (Figura 6) exemplifica num grafo acíclico dirigido as três possíveis configurações α, β e γ, que descrevem dois termos a serem comparados em um determinado DAG. Considerando termi e termj como os termos que estão sendo comparados, MRCA é seu o ancestral comum mais baixo, representando o mais específico de todos os ancestrais comuns de termi e termj. Na configuração A, os dois termos se sobrepõem, ou seja, termi é o próprio termj e sendo assim, também é o próprio MRCA. Em B, termj é um descendente de termi; então termi é também o seu MRCA. Finalmente, na configuração C não existe um caminho entre os dois termos. As linhas tracejadas indicam que pode haver mais do que um caminho entre dois termos. Da mesma forma, nós tracejados representam um ou vários “termos folha” descendentes de um “termo pai”.

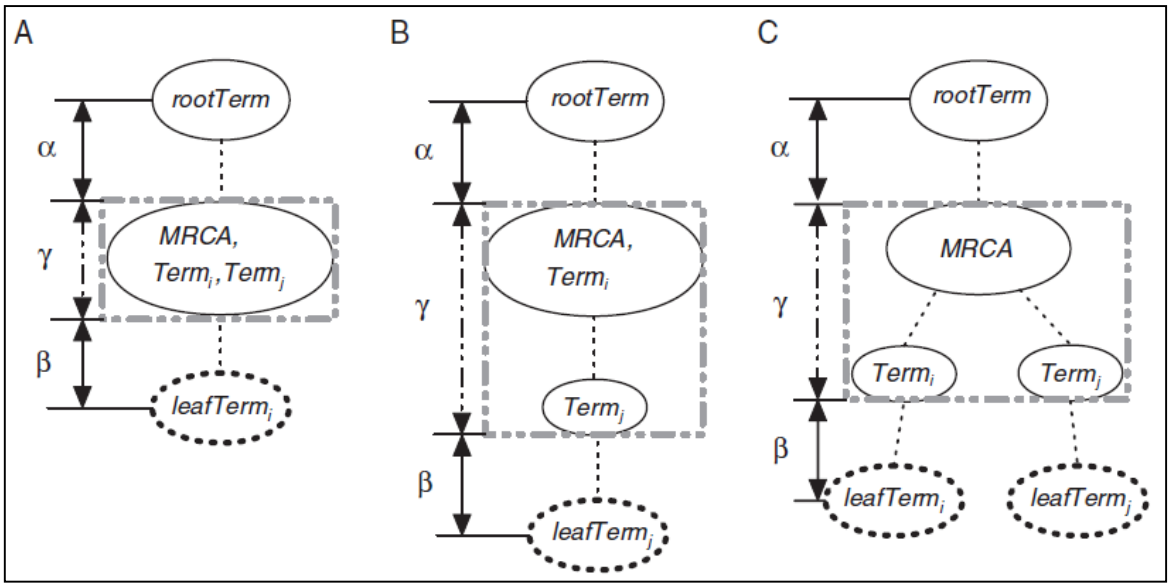

Figura 6. Possíveis configurações α, β e γ num grafo acíclico dirigido.

Assim, combinando as fórmulas α, β e γ, Wu calculou a Similaridade Específica Relativa (RSS):

$$RSS(term_i, term_j) = \frac{maxDepth^{GO}}{maxDepth^{GO}\,\gamma} \cdot \frac{\alpha}{\alpha + \beta} \quad (13)$$

onde $maxDepth^{GO}$ é a distância máxima da raiz até os termos folhas.

2.3.3.3. Híbrido

Um exemplo de medida de similaridade híbrida foi desenvolvido por Wang [Wang-2007] na qual a cada aresta é dado um peso de acordo com o tipo de relacionamento. Wang representou um termo A de uma ontologia como um grafo:

$DAG_A=(A,T_A,E_A)$, onde T_A é o conjunto de todos nós ancestrais de A, incluindo o próprio nó A e E_A é o conjunto de arestas correspondentes.

Considerando um determinado termo A e seu ancestral t, a contribuição semântica de t para A, chamado de valor S, é o produto de todos os pesos das arestas no melhor caminho de t a A, ou seja, o caminho que maximiza o produto.

$$\begin{cases} S_A(A) = 1 \\ S_A(t) = \max\{w_e * S_A(t') \mid t' \in filhosde(t)\}\ se\ t \neq A \end{cases}, \quad (14)$$

onde w_e é o fator de contribuição semântica (peso) para a aresta *e* que liga o termo t aos seus filhos t'.

Wang calcula o valor total de contribuições semânticas de cada ancestral de um determinado termo ao próprio termo pela equação:

$$SV(A) = \sum_{t \in T_A} S_A(t) \quad (15)$$

onde T_A é o conjunto de todos os ancestrais do termo t.

Assim sendo, a similaridade semântica entre dois termos é calculada pela soma das contribuições semânticas de todos ancestrais comuns a ambos os termos e a divisão pelo total de contribuições semânticas de cada ancestral do termo ao próprio termo:

$$S(A,B) = \frac{\sum_{t \in T_A \cap T_B} (S_A(t) + S_B(t))}{SV(A) + SV(B)} \quad (16)$$

Desta forma, Wang calcula a similaridade semântica de dois termos baseada na posição de ambos estes termos dentro de uma ontologia e nas relações semânticas dos mesmos com seus ancestrais.

2.3.4. Medidas de Similaridade entre Produtos Gênicos

- Em Pares: O uso de medidas de similaridade funcional entre produtos gênicos por meio de abordagens baseadas em pares é mais comumente aplicado juntamente com medidas baseadas em nós, e foi primeiramente utilizado por Lord [Lord-2003], por meio do cálculo da média das similaridades entre todos pares de nós. Couto e Azuaje [Couto-2005; Azuaje-2005] também optaram pelo uso da média entre as similaridades entre pares, entretanto, consideraram somente os melhores pareamentos (*best-matching*). Schliker [Schliker-2006] aplicou a estratégia de Couto e Azuje combinando similaridade semântica entre ambas subárvores da GO: *função molecular* e *processo biológico*. Tao [Tao-2007] também aprimorou a similaridade baseada somente nos melhores pareamentos a fim de reduzir ruídos, assim, fez-se do uso de um limiar de similaridade mínima para selecionar os melhores pareamentos. Diferentemente dos autores citados que usaram a média, Sevilla [Sevilla-2006/5] e Riensche [Riensche-2007] levaram em consideração o valor máximo dentre as similaridades entre os pares.

Abordagens baseadas em pares também são aplicadas a medidas baseadas em arestas: Wang e Pozo [Wang-2007; Pozo-2008] utilizaram a estratégia de média dos melhores pareamentos, enquanto que Lei [Lei-2006] testou diferentes combinações de estratégias, como média, máximo, soma de todos os pares, melhores pares e somente combinações exatas.

- Em Grupo: Abordagens puramente baseadas em grupo não são comuns, uma vez que apenas poucas medidas consideram somente anotações diretas. Entretanto, várias abordagens baseadas em grafos utilizam técnicas de similaridade em grupo para simplificar o problema de pareamento de grafos.

Lee [Lee-SG-2004] foi o primeiro a aplicar medidas baseadas em grafos sobre a GO. Para isso, definiu a similaridade entre dois produtos gênicos como sendo equivalente ao número de termos que eles compartilham (em inglês, *Term Overlap* - TO). Mais recentemente, Mistry [Mistry-2008] propôs uma versão normalizada da medida de Lee (NTO, do inglês *Normalized Term Overlap*), na qual o número de termos compartilhados é dividido pelo tamanho do conjunto de anotações para o gene com o menor número de anotações.

Gentleman [Gentleman-2005] criou medidas de similaridade baseadas em grafos como simLP e simUI. A primeira estende o conceito de profundidade máxima do ancestral comum para os produtos gênicos de forma que dois produtos gênicos são tão similares quanto a profundidade do termo que é o ancestral comum mais baixo para anotação direta de todos os termos. Já o simUI considera os produtos gênicos como sendo o conjunto de termos nos seus subgrafos de anotação e utiliza o índice de Jaccard para calcular a similaridade entre eles:

$$J_{(A,B)} = \frac{|A \cap B|}{|A \cup B|} \quad (17)$$

Com base neste mesmo índice de Jaccard, Pesquita [Pesquita-2007] propôs a medida simGIC, na qual cada termo GO é pondereado pelo seu IC. Ye [Ye-2005], por sua vez, propôs uma versão normalizada de simLP, levando em conta as profundidades mínima e máxima dentro de cada categoria da GO.

Cho [Cho-2007] desenvolveu uma abordagem baseada em grupo mais simplificada, calculando a similaridade semântica entre dois produtos gênicos através do IC do termo mais informativo que estes produtos compartilham.

CAPÍTULO 3. Trabalhos Correlatos

Nesta seção serão apresentados dois trabalhos recentes sobre a similaridade entre miRNAs encontrados na literatura.

3.1. MISIM

Wang e seus colaboradores [Wang-2010b] criaram um método de similaridade de miRNAs chamado MISIM (*MicroRNA Similarity*), utilizando a ontologia MeSH e a base de dados de associações de miRNAs-doenças humanas HMDD. Dessa forma, para comparar dois miRNAs MA e MB, por exemplo, Wang seguiu quatro passos principais.

Primeiramente, identificou os conjuntos de doenças, denotados como DA e DB, associadas aos dois miRNAs de interesse. Em seguida, calculou o valor semântico das doenças baseado no DAG das mesmas. Após isso, foi calculada a similaridade semântica para cada par de doenças entre DA e DB no valor semântico da doença. Todas estas etapas foram realizadas com os métodos propostos num trabalho anterior de Wang [Wang-2007], já descritos como métodos híbridos, na seção 2.3.3.3.

Finalmente, a similaridade funcional entre MA e MB foi calculada com base na similaridade semântica entre os conjuntos DA e DB. Entretanto, Wang também levou em consideração as contribuições de doenças similares que são associadas com estes miRNAs. Para isso, a similaridade semântica entre uma doença (dt) e um grupo de doenças ($DT=\{dt_1, dt_2, ...dt_k\}$) foi definida como sendo a similaridade máxima entre aquela doença e cada doença pertencente ao grupo:

$$S(dt, DT) = \max_{1 \le i \le k}(S(dt, dt_i)) \quad (18)$$

Desta forma, considerando que um miRNA MA possui um grupo de doenças relacionadas DT_1 contendo *m* doenças, e um segundo miRNA MB possui outro grupo de doenças relacionadas DT_2, contendo *n* doenças, a similaridade funcional entre MA e MB é dada por:

$$MISIM(MA, MB) = \frac{\sum_{1 \le i \le m} S(dt_{1i}, DT_2) + \sum_{1 \le j \le n} S(dt_{2j}, DT_1)}{m+n} \quad (19)$$

Wang e seus colaboradores validaram seu método, comparando-o com outros métodos capazes de inferir a similaridade funcional dos miRNAs, como famílias e clusteres de miRNAs e similaridade de expressão de miRNAs. Os resultados obtidos mostraram que o método proposto por Wang é confiável.

3.2. GOSemSim

Em um trabalho mais recente, Yu [Yu-2011] e seus colaboradores, desenvolveram um pacote de rotinas escritas na linguagem R chamado GOSemSim (*Gene Ontology Semantic Similarity*), com o intuito de medir a similaridade funcional entre miRNAs. Este pacote possui a implementação dos métodos propostos por Resnik [Resnik-1998], Lin [Lin-1998], Jiang [Jiang-1997], Schlicker [Schliker-2006] e Wang [Wang-2007], além de conter 19 diferentes espécies de miRNAs, como humana, bovina, canina, galinácea, de chimpanzés, de ratos, de porcos, entre outras.

Para realizar o cálculo da similaridade, Yu utilizou anotações da GO, mais especificamente, o sub-grafo de funções moleculares, de genes-alvo dos miRNAs de interesse e o método de Wang. Yu optou por este método por ser baseado na estrutura de grafo da GO e por utilizar outros algoritmos baseados no conteúdo da informação, sendo assim, mais consistente com as perspectivas humanas [Wang-2007]. Yu realizou a combinação das similaridades entre vários genes-alvo para medir a similaridade entre dois miRNAs com base no método de Wang, conforme a equação 20. As similaridades entre estes dois conjuntos de genes associados aos miRNAs de interesse formam uma matriz, e a similaridade entre os miRNAs é calculada, então, pela média entre a similaridade máxima por linha e a similaridade máxima por coluna.

$$Sim = \frac{\sum_{1 \le i \le m} \max(Score_{[i,]}) + \sum_{1 \le j \le n} \max(Score_{[,j]})}{m+n} \quad (20)$$

Yu e seus colaboradores validaram seu método por meio da comparação com a similaridade de expressão de miRNAs, a qual mostrou que clusteres de miRNAs possuem parentesco funcional quando existe cofunção entre seus genes-alvo. As similaridades obtidas pelo método de Yu são consistentes com o conhecimento biológico de relacionamentos funcionais entre miRNAs.

CAPÍTULO 4. Materiais

Nesta seção serão descritos os materiais utilizados como dados para este projeto e a fonte de onde os mesmos foram retirados. Também será destacado o nosso próprio banco de dados unificado (*CFSim Database*) composto da integração dos materiais utilizados.

4.1. Ontologias

Aqui serão brevemente descritas as ontologias utilizadas neste trabalho.

4.1.1. *Ontology for MicroRNA Target (OMIT)*

Em um trabalho realizado recentemente, em 2010, Huang e seus colaboradores [Huang-2010] criaram a primeira ontologia sobre o domínio de miRNAs, a OMIT (do inglês *Ontology for MicroRNA Target)*. Com esta representação formal do conhecimento, é possível facilitar a aquisição de conhecimento e o compartilhamento de dados de fontes já existentes, de forma a padronizar informações heterogêneas oriundas de diferentes bases de dados de miRNAs. A OMIT, conforme mostra a Figura 7, possui diversos conceitos, principalmente aqueles que pertencem ao contexto de miRNAs, como "*MiRNA*", "*OncoGeneratingMiRNA*", "*TumorSuppressingMiRNA*", "*Gene*", "*Disease*", "*Protein*", "*Process*", "*ExperimentalValidation*", "*PathologicalEvent*", "*Organism*", entre outros. Além disso, há também os relacionamentos estabelecidos entre tais conceitos, como por exemplo "*is-a*", "*hasTarget*", "*regulateDisease*", "*regulateProtein*", "*inOrganism*", "*involvedInEvent*", "*experimentContainsMiRNA*", etc.

Com a OMIT, a equipe de Huang tem a finalidade de auxiliar biólogos em desvendar os importantes papéis dos miRNAs em cânceres humanos e, portanto, ajudar os médicos e especialistas na tomada de decisões no tratamento de pacientes com câncer.

Atualmente, a OMIT ainda não se encontra instanciada, estando somente disponível a sua estrutura no site do laboratório da equipe de Huang.

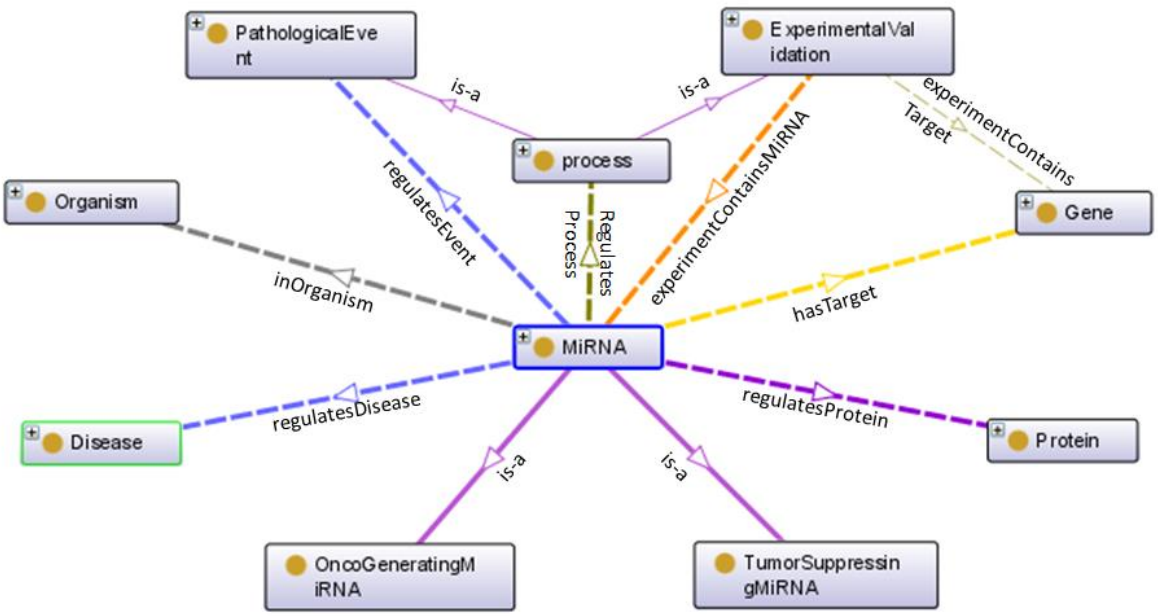

Figura 7. Conceito de MiRNA e seus relacionamentos na OMIT.

4.1.2. *Gene Ontology (GO)*

O Projeto GO, produzido pelo *Gene Ontology Consortium*, teve início em 1998, como uma colaboração entre bancos de dados de três organismos: *FlyBase* (Drosophila), *Saccharomyces Genome Database* (SGD) e *Mouse Genome Informatics* (MGI). Desde então, o GO *Consortium* tem crescido para incluir outras bases de dados de genomas de plantas, animais e micróbios [Consortium,2004].

A GO é composta por um conjunto de vocabulários controlados, estruturados e bem definidos para descrever domínios chave de biologia molecular, incluindo atributos de produtos gênicos e sequencias biológicas. Um dos principais objetivos da GO é proporcionar um recurso público centralizado permitindo o acesso universal para as ontologias, os conjuntos de anotações de dados e as ferramentas de software desenvolvidas para serem utilizadas juntamente com a GO [Consortium,2004].

A GO possui três sub-ontologias que representam diferentes domínios: componente celular, composto pelas partes de uma célula ou seu ambiente extracelular; função molecular, composto por atividades elementares de um produto gênico ao nível molecular, tais como ligação ou catálise; e processo biológico, composto pelas operações ou conjuntos de eventos moleculares com começo e fim definidos, pertinentes para o funcionamento de unidades vivas integradas, ou seja, células, tecidos, órgãos e organismos. A interação entre os termos de uma mesma ontologia e entre ontologias diferentes é feito por meio de um conjunto de 6 diferentes

relacionamentos: *"is-a", "part_of", "regulates", "positively-regulates", "negatively-regulates"* e *"has_part"* [CONSORTIUM, 2010].

A GO é considerada a ontologia padrão para a anotação de produtos gênicos, estruturada como um grafo acíclico dirigido, conforme mostra a Figura 8, no qual cada termo define relações com um ou mais outros termos no mesmo domínio, e às vezes para outros domínios [GO-2014]. Esta estrutura da GO a torna viável para comparações quantitativas semânticas [Yu-2011].

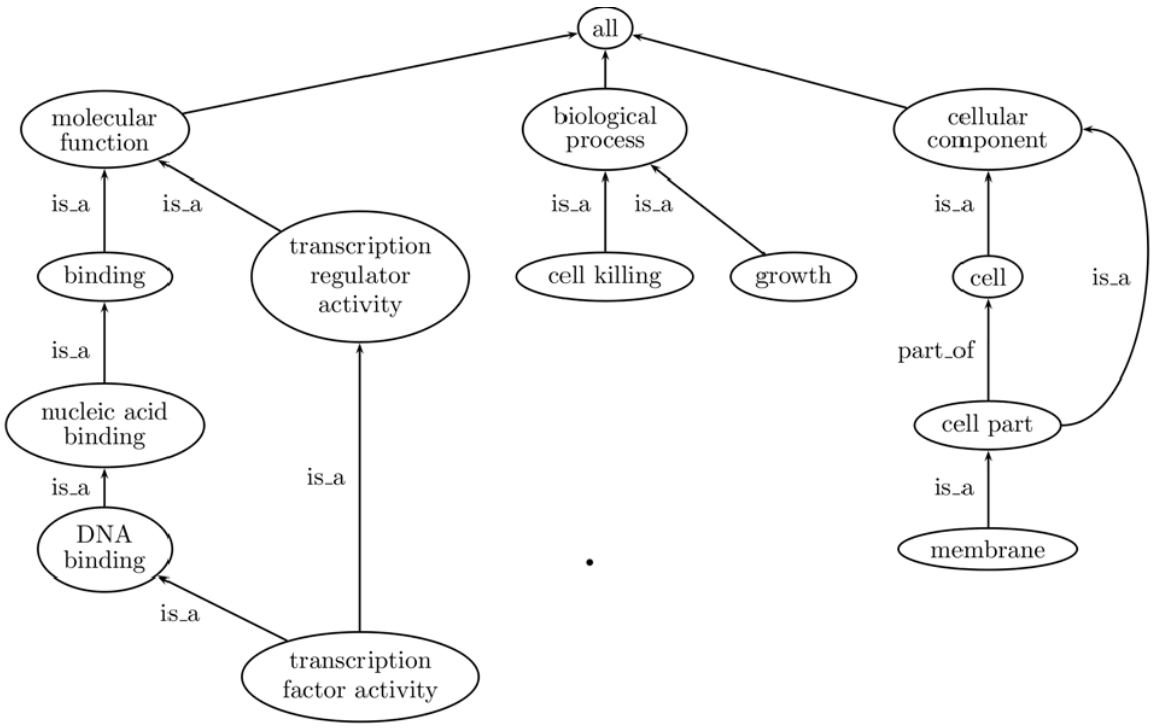

Figura 8. Parte da GO mostrando as três sub-ontologias (função molecular, processo biológico e componente celular) e alguns de seus termos descendentes. O fato da GO ser um DAG em vez de uma árvore é ilustrado pelo termo "*transcription factor activity"*, o qual tem dois nós-pais diretos.

4.1.3. *Medical Subject Headings (MeSH)*

O MeSH, desenvolvido, publicado e disponível online pela *National Library of Medicine* (NLM) [NLM], é um vocabulário controlado cujo domínio de estudo e atuação está delimitado às Ciências da Saúde. É utilizado como instrumento para indexação, classificação, busca e recuperação de informações e de documentos biomédicos e relacionados a saúde [MESH-2014].

No MeSH, um descritor representa uma classe de conceitos, o qual, por sua vez, representa uma classe de termos sinônimos. A organização do MeSH se dá em 16 categorias, por exemplo Categoria A para termos anatômicos, Categoria B para organismos, Categoria C

para doenças, e assim por diante. Cada uma dessas categorias se divide em subcategorias, nas quais os descritores subordinados são organizados hierarquicamente numa relação do mais genérico para o mais específico. Dessa forma, os descritores MeSH formam uma estrutura de um grafo acíclico dirigido, no qual cada nó consiste de um descritor com um identificar único e com números da árvore, que consistem de uma lista dos números da árvore dos nós-pai deste nó, separados por ';'. A vantagem é que a partir de qualquer nó, podemos facilmente analisar um único número de árvore, sem ter de consultar todo o DAG relacionado a ele quando precisamos de acesso a todos os seus ancestrais [Du-2009].

Uma vez que, neste DAG, os nós representam as doenças e as arestas representam os relacionamentos "é-um" entre as doenças, o Mesh fornece um sistema de classificação de doenças que auxilia no estudo dos relacionamentos entre as mesmas. A Figura 9 apresenta uma parte do DAG de neoplasma de mama do MeSH.

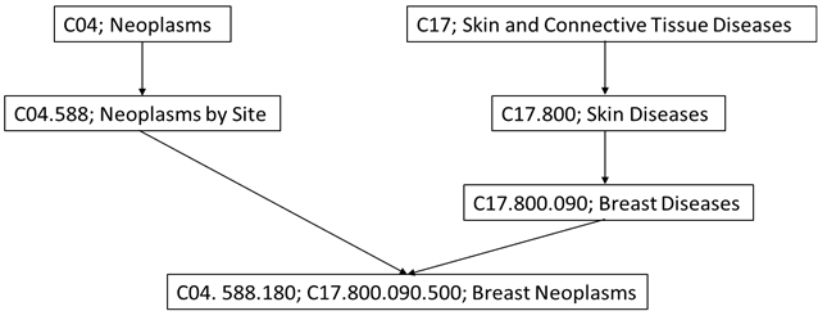

Figura 9. Grafo Acíclico Direcionado para Neoplasma de Mama [Du-2009].

4.2. Bases de Dados

Aqui serão brevemente descritas as bases de dados utilizadas neste trabalho.

4.2.1. m*iRBase*

Base de dados de miRNAs amplamente conhecida e utilizada pelos pesquisadores da área biomédica. Contém informações como sequencias e anotações de miRNAs já publicados, assim como a espécie e família da qual pertencem. Além disso, esta base permite o registro de nomes de novos miRNAs, levando a padronização da nomenclatura [MIRBASE-2014].

As informações do miRBase são frequentemente atualizadas, sendo que sua última versão contém registros de 24521 precursores de miRNAs, 30424 miRNAs maduros de 206 espécies diferentes.

4.2.2. *Human MicroRNA Disease Database – HMDD*

Embora as atuais associações miRNAs-doenças ainda estejam longe de serem completas, com o intuito de auxiliar no estudo das mesmas e oferecer uma plataforma para dissecar os padrões dos miRNAs em doenças, Lu e seus colaboradores [Lu-2008] realizaram uma análise em larga escala dessas associações entre os miRNAs e as doenças humanas, as quais foram manualmente coletadas a partir de literaturas e publicações. Em novembro de 2007, eles construíram um banco de dados de doenças humanas relacionadas a miRNAs (em inglês, *Human MicroRNA Disease Database* - HMDD), o qual contém nomes dos miRNAs, nomes das doenças, evidências de disfunções e número de identificação PubMed das referentes literaturas.

Atualmente, esta base de dados contém 10177 registros, o que inclui 617 miRNAs e 438 doenças padronizadas de acordo com o MeSH, retirados de um total de 3266 publicações.

4.2.3. *TarBase*

Os miRNAs são responsáveis pela regulação da expressão de proteínas por meio da ligação a um ou mais sítios-alvo em um transcrito de RNAm a fim de inibir sua tradução. Para melhor compreender este processo de regulação da expressão dos miRNAs, em 2006, pesquisadores da Universidade da Pensilvânia criaram um banco de dados de genes-alvo de miRNAs, chamado de *TarBase* [Sethupathy-2006]. Este repositório de dados abriga uma coleção de alvos de miRNAs experimentalmente testados em diversas espécies animais de interesse científico centrais, plantas e vírus. Embora existam vários programas computacionais para predição de alvos de miRNA, há uma necessidade de um coleção e descrição compreensiva de alvos miRNA com suporte experimental.

A versão atual do *TarBase* inclui mais de 1300 alvos experimentalmente testados. Cada um destes registros de sítio-alvo de miRNA é descrito pelo miRNA que se liga ao mesmo, o gene em que ocorre, a natureza dos experimentos que foram conduzidos para testá-lo, a suficiência do sítio para induzir a repressão e/ou de clivagem e a referência a partir da qual todos estes dados foram extraídos. Além disso, o banco de dados TarBase está funcionalmente ligado a várias outras bases de dados relevantes e úteis, tais como Ensembl, HUGO, UCSC e SwissProt.

4.2.4. *Genetic Association Database (GAD)*

O aumento da disponibilidade de dados de polimorfismos tem permitido mais estudos de associações genéticas e o número de publicações sobre as mesmas tem crescido

rapidamente [Zhang-2010]. Com o objetivo de permitir que, no contexto de nomenclatura padronizada, o usuário identifique rapidamente polimorfismos medicinalmente relevantes, a partir de um grande volume de polimorfismos e dados mutacionais, Zhang e seus colaboradores criaram um banco de dados que contém estudos de associações de genes humanos com muitos tipos de doenças comuns. Este banco é focado principalmente no arquivamento de informações sobre a doenças humanas comuns, ao invés de raros distúrbios mendelianos, como na *Online Mendelian Inheritance in Man* (OMIM) [McKusick-2007].

Atualmente, o GAD contém aproximadamente 40000 registros de doenças genéticas retiradas de 23000 publicações independentes. Parte de seus registros de doenças, mais especificamente 10324 registros, foi anotada seguindo a padronização do MeSH.

4.2.5. *Human MicroRNA Oncogenes and Tumor Suppressors*

Evidências crescentes têm revelado que os miRNAs podem atuar como oncogênicos ou como supressores tumorais, e apesar do grande interesse na pesquisa destas moléculas associadas ao câncer, ainda há pouca informação registrada sobre estes miRNAs. Assim, D. Wang e seus colaboradores, em 2010, realizaram um estudo sobre miRNAs humanos oncogênicos e supressores de tumor [Wang-2010a].

Neste trabalho, os pesquisadores identificaram os miRNAs como *p-oncomirs* ou como *p-mirsupp*, através da revisão manual de aproximadamente 5000 publicações. Para tanto, seguiram as seguintes regras: se um artigo (ou mais) sugere claramente que um determinado miRNA é um oncogene ou supressor tumoral, este miRNA é então identificado como *p-oncomir ou p-mirsupp*, respectivamente. Por outro lado, se nenhum artigo claramente especifica um dado miRNA como oncogene ou supressor tumoral, este miRNA será identificado como *p-oncomir* somente se ao menos 3 diferentes artigos oferecerem fortes evidências de que esta molécula em questão estava associada a super-expressão, regulação positiva e/ou crescimento do tumor; ou então será identificado como *p-mirsupp* se for associado a baixa-expressão, regulação negativa e/ou inibição do crescimento tumoral. Ainda, para que um miRNA fosse classificado como pertencente a um destes dois grupos, não poderia haver informações conflitantes sobre o mesmo na literatura. Portanto, foram desconsiderados aqueles miRNAs publicados como oncogênicos ou supressores tumorais dependendo do contexto.

4.3. CFSim Database

Sendo o objetivo deste trabalho a criação de um ambiente que permita calcular a similaridade funcional entre miRNAs baseada em ontologias, utilizamos a ontologia de miRNAs OMIT como base estrutural para a criação de nosso banco de dados de miRNAs. Assim, as classes e os relacionamentos da OMIT que envolvem o conceito miRNA, descritos a seguir, foram implementados em uma base de dados relacional.

Tabelas:

- *miRNA*: identificador, nome, sequencia, se possui natureza oncogenica e/ou supressora de tumor e referencia para família e organismo;
- miRNA*Family*: identificador, código e nome;
- *Organism*: identificador, gênero, espécie e símbolo;
- *Experiment*: identificador e nome;
- *Gene*: identificador, nome, símbolo e código Ensembl;
- *Disease*: identificador e nome;
- *PathologicalEvent*: identificador e descrição;
- *Protein*: identificador, nome, símbolo e código SwissProt.

Relacionamentos:

- *hasTarget*: associação entre miRNA e seu gene-alvo;
- *regulateDisease*: associação entre miRNA e doença;
- *regulateProtein*: associação entre miRNA e proteína;
- *inOrganism*: relacionamento entre miRNA e o organismo no qual se encontra;
- *regulatesPathologicalEvent*: relacionamento entre miRNA e o evento patológico que regula;
- *experimentContains*: relacionamento entre o miRNA e o experimento que o contém;
- *isOncogene*: quando um miRNA atua como oncogene;
- *isTumorSupressor*: quando um miRNA atua como supressor tumoral.

Após a realização do levantamento bibliográfico e estudo sobre diversos banco de dados de miRNAs existentes na literatura, certas bases de dados, como miRBase, HMDD, *TarBase*, GAD, *Human MicroRNAs Oncogenes and Tumor Suppressors*, por corresponderem aos interesses de nosso objetivo, foram selecionadas para popular nosso próprio banco de dados. Os demais bancos de dados encontrados na literatura não foram utilizados por não possuírem padronização de nomenclatura dos miRNAs, ou dos genes ou das doenças. Por outro lado, dentre as bases de dados mencionadas acima, todas seguem a nomenclatura de

miRNAs de acordo com o mirBase; as bases *Tarbase* e GAD seguem a nomenclatura de genes de acordo com a HUGO [HGNC-2014]; e por sua vez, HMDD e GAD respeitam a nomenclatura de doenças estipulada pelo MeSH. Assim, a padronização dos registros oriundos de bases de dados heterogêneas facilitou sua integração num único banco de dados a ser utilizado neste projeto.

Todas as bases de dados selecionadas foram obtidas em sua última versão publicada e disponibilizada e algumas delas tiveram seus registros pré-processados. Por exemplo, no banco de dados GAD havia instancias de algumas doenças repetidas, mudando somente a pluralidade da palavra ("*Lung Neoplasm*" versus "*Lung Neoplasms*") ou então a aparição versus a omissão da letra 's' após uma apóstrofe ("*Alzheimer Disease*" versus "*Alzheimer's Disease*", "*Parkinson Disease*" versus "*Parkinson's Disease*", entre outras). Além disso, existiam também alguns erros de ortografia em alguns registros.

Informações dos miRNAs, como nome, espécie e sequência foram retiradas do miRBase, sendo que inicialmente foram considerados somente miRNAs humanos, representando o relacionamento *inOrganism*. As classificações dos miRNAs como oncogênicos ou supressores tumorais, para popular os relacionamentos *isOncogene* e *isTumorSuppressor*, foram extraídas do banco de *Human MicroRNAs Oncogenes and Tumor Suppressors*. A partir dos dados do HMDD, conseguimos obter as associações entre miRNAs e as doenças humanas, representadas pelo relacionamento *regulateDisease*, e, por sua vez, as associações diretas entre genes e doenças humanas, utilizadas para verificação de existência de redundâncias, foram retiradas do GAD. Por fim, a partir do *TarBase*, foi possível coletar informações sobre a maioria dos relacionamentos que envolvem os miRNAs em nosso banco de dados, como por exemplo, *hasTarget*, *regulateProtein*, *regulatePathologicalEvent* e *experimentContains*.

Além disso, apesar da ontologia OMIT não possuir em sua estrutura a indicação da família à qual um determinado miRNA pertence e dada a importância desta informação, foi implementada na *CFSim Database* uma tabela para armazenar informações sobre as famílias, assim como um relacionamento entre os registros de miRNAs e suas famílias. As informações necessárias para a população desta tabela foram oriundas do miRBase.

As ontologias GO e MeSH foram obtidas por meio de seus respectivos *websites*, que disponibilizam gratuitamente tais dados, e integradas ao *CFSim Database*. Ainda, através do *site* do miRBase foi possível encontrar os dados necessários para a implementação da ontologia de organismos, uma vez que o site contém uma árvore das espécies das quais já

foram descobertos registros de miRNAs. Esta ontologia também foi integrada à nossa base de dados.

Com o intuito de facilitar a recuperação das fontes de cada registro pertencente aos relacionamentos de nosso banco de dados (por exemplo, registro de associação entre miRNAs e doenças, ou entre miRNAs e genes-alvo, ou ainda entre genes e doenças), foi criada uma tabela contendo as informações das respectivas publicações, tais como, PubMed ID, título, autores, ano de publicação e observação.

Desta forma, o *CFSim Database* contém todas as informações sobre miRNAs e seus relacionamentos necessários para o desenvolvimento projeto.

4.3.1. Ambiente de Desenvolvimento

Para o pré-processamento das informações contidas nas distintas bases de dados e ontologias utilizadas e para criação e população do banco de dados unificado, foi utilizada a linguagem de programação *Java* através do *Net Beans IDE*[1], juntamente com o Sistema Gerenciador de Banco de Dados *PostgreSQL*[2].

[1] http://netbeans.org/downloads/
[2] http://www.postgresql.org.br/

5. Método Proposto

Esta seção contém a descrição do método proposto neste projeto, chamado *CFSim* (*Composed Functional Similarity*), da implementação e da forma de validação do mesmo.

5.1. Proposta

De forma geral, as métricas descritas na seção 3.1 têm por objetivo calcular a similaridade entre dois termos de uma ontologia. A proposta do presente trabalho consiste em determinar e avaliar um método de cálculo que leve em consideração a possibilidade de haver um conjunto de informações a respeito dos dois miRNAs que estão sendo comparados, de forma que estas informações sejam consideradas no cálculo da similaridade. Isto pode envolver diferentes valores de similaridade, obtidos para diferentes categorias de informação. Por isto, a similaridade funcional é caracterizada neste trabalho como "composta". Exemplificando: sobre os dois miRNAs pode haver anotações sobre genes-alvo e também sobre doenças associadas. Nesse caso, o método deverá calcular a similaridade entre os dois cojuntos de genes-alvo (a partir da GO) e também a similaridade entre os dois cojuntos de doenças (a partir da MeSH) e calcular um valor de similaridade que leve em consideração ambas as similaridades específicas, levando em conta também a possível existência de associação direta entre o gene-alvo e a doença, o que poderia caracterizar uma redundância. Genericamente, é criado um vetor de similaridade, o qual conterá como elementos os valores de similaridades específicas sobre as diferentes informações existentes a respeito dos dois miRNAs a serem comparados. O valor final da similaridade será então obtido a partir da combinação dos elementos do vetor de similaridade.

O método proposto é representado na Figura 10. Os dois miRNAs a serem comparados, M^1 e M^2, são pesquisados na OMIT, a fim de serem levantadas as anotações contidas sobre as informações relativas a ambos em todas as suas possíveis categorias de informação (Ca, Cb, ..., Cn), por meio de um módulo chamado Determinação das Categorias. O que chamamos aqui de categorias de informações são os relacionamentos existentes na OMIT para uma determinada instância de miRNA, que caracterizam as informações anotadas a respeito daquele miRNA. Por exemplo, a categoria Doenças (Ca^1) para o miRNA M^1 representa todo o conjunto de doenças associadas ao miRNA M^1, obtidas pelas associações da instância M^1 da classe miRNA com as instâncias da classe Disease na OMIT; a categoria Organismo (Cb^1) representa o relacionamento entre o miRNA M^1 e o organismo no qual se encontra; e assim por diante. Sempre que existirem informações de uma determinada

categoria para ambos os miRNAs M^1 e M^2, o módulo Cálculo das Similaridades das Categorias realizará o cálculo da similaridade específica de cada categoria, $Sim(Ca^1,Ca^2)$, $Sim(Cb^1,Cb^2)$, e assim sucessivamente, com base em suas respectivas ontologias. Por exemplo, se houver anotações para a categoria Doenças associadas aos miRNAs, a similaridade nesta categoria será calculada a partir da aplicação de uma métrica (como aquelas descritas na seção 3.1) sobre a ontologia MeSH; e assim para cada categoria de anotação existente. Com os valores de similaridade calculados para cada categoria, o módulo Montagem do Vetor irá montar o vetor de similaridade, levando em consideração casos de redundância, assim como especificidades, tais como a informação de que o miRNA é oncogênico ou supressor da doença. Por fim, a combinação dos elementos do vetor de similaridade irá gerar o valor final da similaridade funcional composta entre os miRNAs $CFSim(M^1,M^2)$. Esta combinação poderá ser implementada no módulo Combinação das Similaridades das Categorias, por exemplo, pela média ponderada entre os diferentes valores de similaridade obtidos para cada categoria.

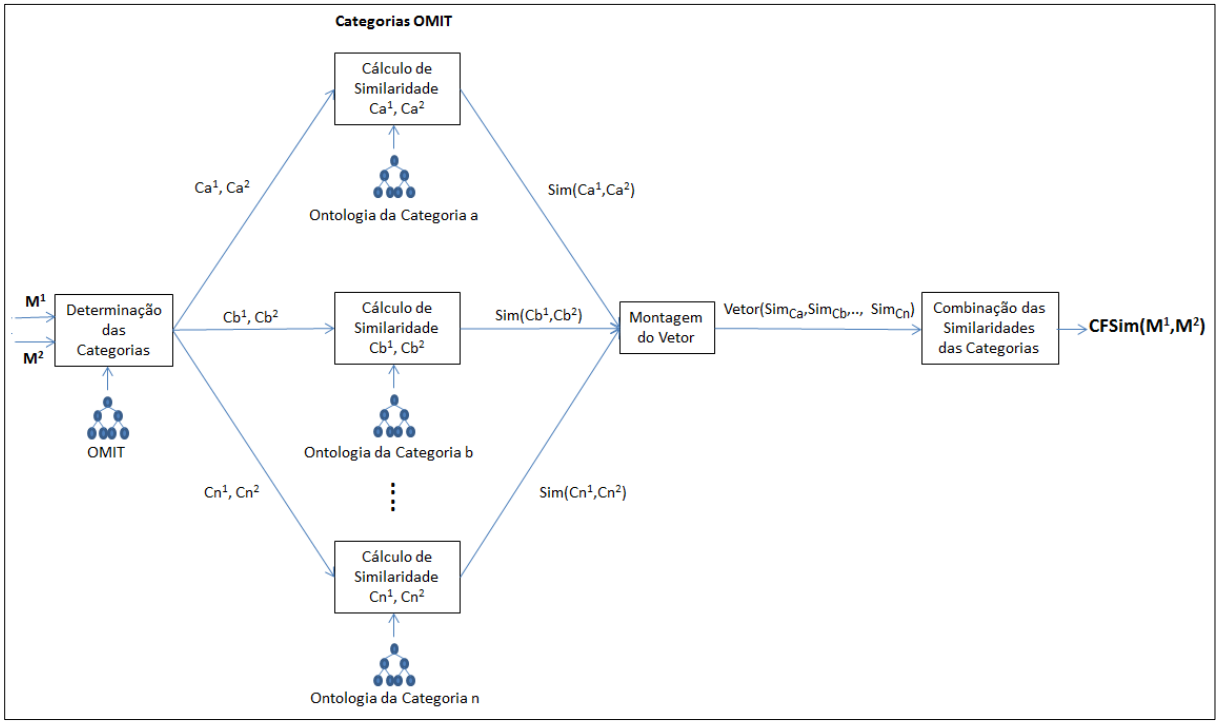

Figura 10. Esquematização do método proposto.

5.2. Implementação

Para o desenvolvimento da metodologia proposta acima, foram implementados algoritmos para os quatro módulos: Determinação das Categorias, Cálculo das Similaridades das Categorias, Montagem do Vetor e Combinação das Similaridades das Categorias. Estes módulos estão representados por retângulos no esquema da Figura 10 e serão detalhados a seguir.

5.2.1. Determinação das Categorias

O módulo Determinar Categorias verifica a existência de três categorias de anotações para os miRNAs a serem comparados: genes-alvo, doenças e natureza (oncogênico/supressor). O funcionamento deste módulo é mostrado no pseudo-código da Figura 11. Caso existam anotações sobre ambas as categorias genes-alvo e doenças para M^1 e M^2, será necessário verificar se já não existe associação direta entre os genes-alvo e as doenças, o que indicaria uma redundância. Se existirem informações redundantes, consideraremos somente a categoria de genes-alvo e portanto a *CFSim* será igual à similaridade funcional entre os genes-alvo; caso contrário, levaremos em consideração ambas as categorias, fazendo uma combinação entre as similaridades funcionais entre genes-alvo e doenças. Por outro lado, se existirem anotações para ambos M^1 e M^2 somente sobre a categoria de genes-alvo, a similaridade funcional composta *CFSim* será igual a similaridade funcional entre os genes-alvo e caso haja anotações somente sobre a categoria de doenças, a *CFSim* será equivalente a similaridade funcional entre as doenças. Além disso, todos os casos descritos acima também são combinados com o valor de similaridade para a categoria de Natureza, ou seja, baseado na propriedade dos miRNAs serem oncogenicos ou supressores de tumor.

```
ALGORITMO DeterminaçãoDasCategorias
INÍCIO
    SE existem anotações para M1 e M2 sobre categoria de Genes-Alvo ENTÃO
        SE existem anotações para M1 e M2 sobre categoria de Doenças ENTÃO
            SE existe relacionamento direto entre Gene-Alvo e Doença ENTÃO
                Uso da categoria de Genes-Alvo e de Natureza;
            SENÃO
                Uso da combinação entre categorias de Genes-Alvo, de Doenças e de Natureza;
            FIM SE;
        SENÃO
            Uso da categoria de Genes-Alvo e de Natureza;
        FIM SE;
    SENÃO
        SE existem anotações para M1 e M2 sobre categoria de Doenças ENTÃO
            Uso da categoria de Doenças e de Natureza;
        SENÃO
            Uso da categoria de Natureza;
        FIM SE;
    FIM SE;
FIM.
```

Figura 11. Pseudo-Código do algoritmo para o módulo Determinação das Categorias.

5.2.2. Cálculo das Similaridades das Categorias

Para a implementação do cálculo de similaridade para cada uma das categorias determinadas pelo módulo anterior utilizamos o método híbrido de similaridade funcional baseado em ontologias definido por J. Z. Wang [Wang-2007], que posteriormente foi aprimorado por D. Wang [Wang-2010-b], descritos nas seções 2.3.3.3 e 3.1, respectivamente. Os motivos que nos levaram a esta escolha são que o método híbrido de J. Z. Wang calcula a similaridade funcional entre dois termos levando em consideração tanto a posição de ambos dentro de uma ontologia quanto as relações semânticas dos mesmos com seus ancestrais; e o aprimoramento dado por D. Wang permite aplicar o método híbrido em comparações em grupo, ainda, levando em consideração as contribuições de termos similares dentro de um mesmo grupo.

Feita a escolha do método local, o próximo passo foi implementa-lo, replicando-o e adaptando suas variáveis para os termos de interesse, ou seja, este método foi replicado para calcular a similaridade funcional entre genes com base na GO, posteriormente foi replicado para a similaridade funcional entre doenças com base na MeSH e, finalmente, foi replicado para a similaridade funcional entre organismos na ontologia de organismos retirada do miRBase. Como estes métodos fazem buscas pelos termos de interesse em suas respectivas ontologias, também foi necessário desenvolver métodos que pudessem acessar a GO, a MeSH e a ontologia de organismos integradas ao nosso banco de dados unificado, *CFSim Database*, e nos retornar os termos de interesse. Inicialmente trabalhamos somente com dados de

miRNAs humanos, de forma que o método para similaridade aplicado sobre a ontologia de organismos não foi utilizado. E no caso da GO, o nosso método busca por seus termos especificamente na sub-ontologia de funções moleculares, sendo assim, em nosso método somente são considerados os genes-alvos que possuem termos nesta sub-ontologia da GO.

Após a definição de quais categorias serão utilizadas para compor o valor final da *CFSim*, é realizado o cálculo de similaridade para cada uma destas categorias independentemente, utilizando o método de J. Z Wang. e D. Wang Para exemplificar, vamos supor que ambos miRNAs M^1 e M^2 possuem informações para a categoria de genes-alvo, ou seja, fazendo uma busca em nosso banco *CFSim Database*, encontramos seus respectivos conjuntos de genes-alvo G1={g_{11}, g_{12}, ..., g_{1m}} e G2={g_{21}, g_{22}, ..., g_{2n}}. Primeiramente, montamos uma matriz de genes-alvo MG na qual um dos eixos representa o conjunto G1 e o outro eixo representa o conjunto G2. Em seguida, para cada elemento da matriz MG, ou seja, para cada par de genes-alvo, como por exemplo (g_{11}, g_{21}), realizamos a busca dos termos relacionados a g_{11} e a g_{21} na sub-ontologia de funções moleculares da GO, obtendo os conjuntos de termo T_{G1}{t_{11A}, t_{11B}, ..., t_{11K}} e T_{G2}={t_{21A}, t_{21B}, ..., t_{21L}}, respectivamente. Assim, geramos uma matriz de termos MT na qual um dos eixos representa o conjunto T_{G1} e o outro eixo representa o conjunto T_{G2}. Feito isso, para cada elemento da matriz MT, ou seja, para cada par de termos, como por exemplo (t_{11A}, t_{21A}), calculamos a similaridade semântica entre os mesmos, como mostrado na Figura 12. A similaridade semântica entre dois termos é calculada pela soma das contribuições semânticas de todos os ancestrais comuns a ambos os termos e a divisão pelo total de contribuições semânticas de cada ancestral ao próprio termo, Equação 16. Para isso, antes foi necessário obter o valor total da contribuição semântica de cada ancestral ao próprio termo, dado pela Equação 15, para t_{11A} e para t_{21A}. O valor da contribuição semântica de um termo a ele mesmo é 1, enquanto que a contribuição semântica de um acentral do termo ao próprio termo é o produto de todos os pesos das arestas no melhor caminho entre o ancestral e o termo em questão, ou seja, o caminho que maximiza o produto, definido na Equação 14. Consideramos os pesos das arestas como sendo 0.8 para relacionamentos "*is-a*" e 0.6 para relacionamentos "*part-of*" da GO. Feito o cálculo da similaridade semântica para cada par de termos da matriz MT, o próximo passo é calcular a similaridade funcional entre o par de genes-alvo (g_{11}, g_{21}) da matriz MG. Esta etapa é mostrada nas Equações 18 e 19, que de forma simplificada realizam a soma dos máximos de cada linha e dos máximos de cada coluna da matriz MT e a divisão pelo número total de termos nos conjuntos T_{G1} e T_{G2}. Este processo é repetido para todos os elementos da matriz MT e depois para todos os elementos da matriz MG, obtendo assim, o valor final de

similaridade entre M^1 e M^2 para a categoria de genes-alvo. E, posteriormente, toda esta metodologia é replicada para a categoria de doenças, caso exista.

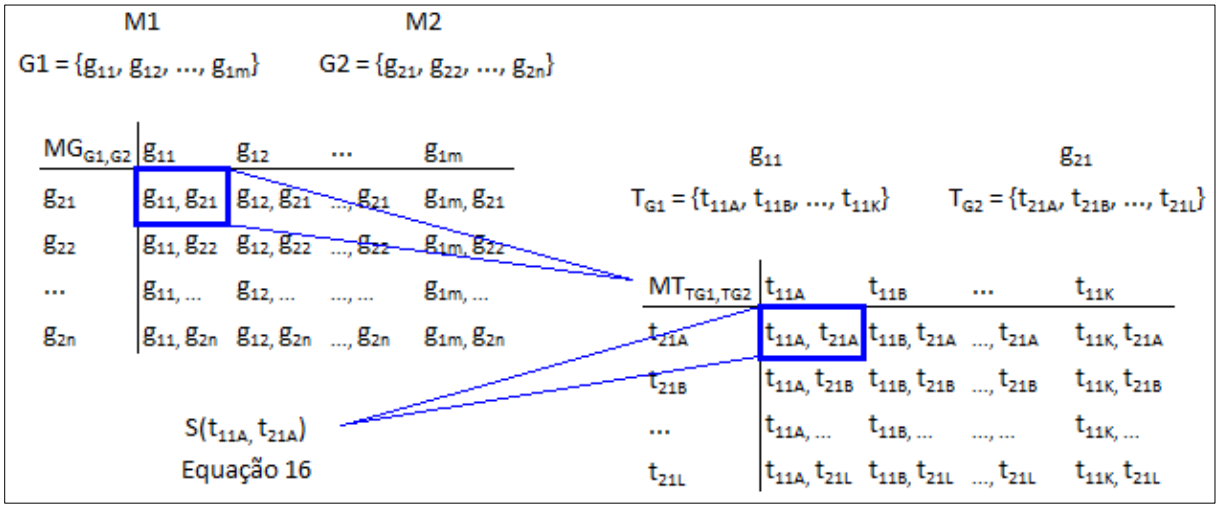

Figura 12. Representação da matriz de genes-alvo (MG) associados a M^1 e M^2 e da matriz de termos (MT) que representam cada par de genes-alvo de M^1 e M^2.

5.2.3. Montagem do Vetor

Após a definição dos valores de similaridades independentes para cada categoria (genes-alvo, doenças e natureza) é montado um vetor com os valores obtidos, de forma a facilitar o desenvolvimento da próxima etapa, a qual consistiu em definir como seria feita a combinação entre estas similaridades de cada categoria.

5.2.4. Combinação das Similaridades das Categorias

Duas abordagens foram estudadas objetivando determinar a forma de combinação dos valores de similaridade calculados para cada uma das categorias existentes.

Abordagem 1: fator de correlação genes x doenças associadas

A primeira abordagem consistiu em tentar estabelecer um fator de correlação entre genes e doenças, de forma que, na combinação, o valor da similaridade de doenças pudesse ser convertido em similaridade de genes, uma vez aplicado o fator de correlação. Dessa forma, o conjunto de doenças poderia ser considerado - em termos de similaridade - como se fosse um conjunto de genes e então as doenças poderiam entrar na tabela MG juntamente com os genes, e teríamos apenas um valor de similaridade final. Para tanto, foi desenvolvido um algoritmo no qual as informações utilizadas foram retiradas da base de dados GAD, a qual possui registros de associações diretas entre genes e doenças. Para cada par de associações, calculamos a similaridade entre os dois genes com base na GO e a similaridade entre as duas doenças com base na MeSH, ambos utilizando o método de J. Z. Wang e D. Wang.

Posteriormente, plotamos um gráfico do tipo *scatterploot* no qual o eixo X representa as similaridades de genes e o eixo Y representa as similaridades das doenças associadas àqueles genes, com o objetivo de buscar por uma equação de correlação, obtido a partir de uma regressão sobre esses dados. Entretanto, o resultado obtido não foi compatível com nossa hipótese inicial, ou seja, não existe a correlação entre as similaridades de genes e as similaridades das doenças associadas aos mesmos. Para um mesmo valor de similaridade de gene, há uma gama completamente aberta de similaridade de doenças, e isto ocorre para todos os valores de similaridades de gene, conforme ilustrado na Figura 13. Assim, concluímos que a similaridade de genes baseada em funções moleculares da GO não pode ser usada para avaliar similaridade entre doenças associadas e vice-versa.

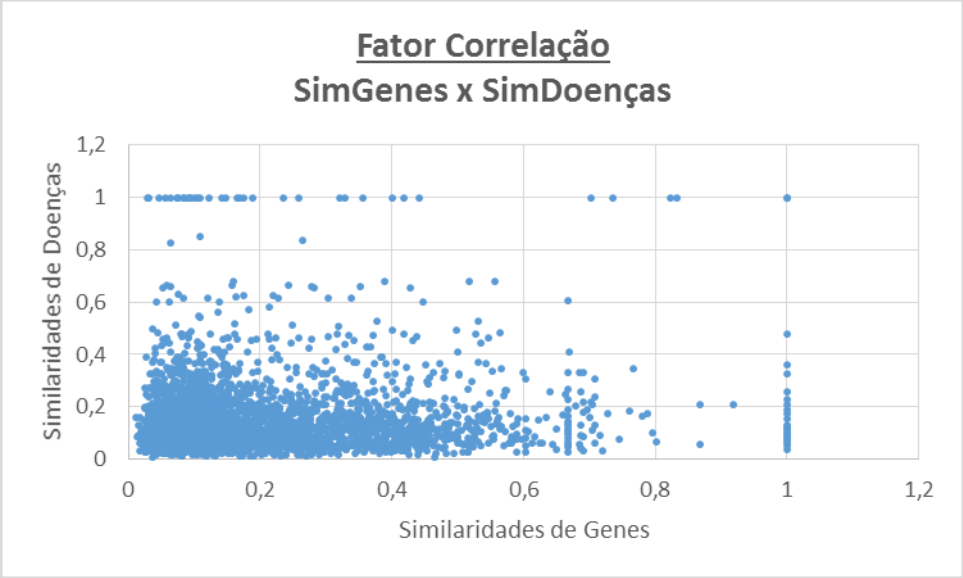

Figura 13. Fator de correlação entre a similaridade de genes-alvo e a similaridade de doenças utilizando dados da base de dados GAD.

Abordagem 2: ponderação supervisionada

Outra abordagem consistiu em desenvolver e executar um algoritmo para encontrar pesos (*pg*, *pd e pn*) adequeados para as similaridades para as categorias de genes-alvo (*simG*), doenças (*simD*) e natureza (*simN*), de acordo com a Equação 21.

$$CFSim(M^1, M^2) = pg * simG + pd * simD + pn * simN \quad (21)$$

E para isso, os valores de *simG* e *simD* continuaram respeitando os métodos propostos por Wang J. Z. e Wang D., enquanto que o valor da similaridade para a categoria de natureza (*simN*) para dois miRNAs M^1 e M^2 foi definido conforme a equação 22. Caso ambos M^1 e M^2

sejam de mesma natureza (se ambos são oncogenicos ou se ambos são supressores de tumor) $simN$ = 1; caso sejam de naturezas opostas (M^1 oncogenico e M^2 supressor ou M^1 supressor e M^2 oncogenico) $simN$ = 0; e se não existirem anotações sobre esta categoria em nosso banco de dados $simN$ = 0.5.

$$simN = \begin{cases} 1, se\ M^1\ e\ M^2\ são\ de\ mesma\ natureza \\ 0, \text{se}\ M^1\ e\ M^2\ \text{são}\ de\ naturezas\ opostas \\ 0.5, se\ M^1\ e/ou\ M^2\ são\ de\ natureza\ indeterminada \end{cases} \quad (22)$$

Em seguida, calculamos as similaridades de cada categoria de forma independente para cada par de miRNAs com anotações sobre as categorias de genes-alvo e doenças, mas sem relacionamento direto entre o gene-alvo e a doença, e desta vez, sem pesos. Lembrando que novamente agrupamos 100 miRNAs de acordo com suas famílias, totalizando 22 famílias distintas. Após a execução do algoritmo, calculamos a média da similaridade de cada categoria para os miRNAs de mesma família (média-intra) e para os miRNAS de famílias distinas (média-inter), para cada par de famílias de miRNAs. Posteriormente, calculamos a diferença entre as médias-intra e as médias-inter e, por fim, obtivemos a média das diferenças para cada categoria. As médias obtidas foram 0,159 para a categoria de genes-alvo, 0,175 para a categoria de doenças e 0,346 para a categoria de natureza. Analisando tais médias, notamos que, convertendo esta proporção para 100%, seria possível utilizar estes valores como pesos para a combinação das categorias em nosso método. Desta forma obtivemos os seguintes valores de pesos: pg = 0,23, pd = 0,26 e pn = 0,51 para casos em que ambos miRNAs que estão sendo comparados possuem anotações sobre as categorias de genes-alvo e doenças.

O mesmo processo foi realizado para dados de miRNAs com anotações somente na categoria de doenças e depois para aqueles com somente anotações na categoria de genes-alvo. No primeiro caso, os resultados obtidos foram: a média das diferenças para categoria de doenças igual a 0,48 e para a categoria de natureza igual a 0,09. Sendo assim, fazendo a proporção para 100%, obtivemos pd = 0,84 e pn = 0,16 para serem utilizados como pesos em nosso método para casos em que ambos miRNAs que estão sendo comparados possuem anotações somente sobre a categoria doenças. No segundo caso, ou seja, anotações somente de genes-alvo, verificamos que nenhum dos miRNA que possuíam apenas anotações de genes-alvo possuía anotação de natureza e, portanto, não foi possível realizar o estudo para determinação dos pesos. Entretanto, analisando os valores de pesos obtidos para miRNAs com anotações sobre genes-alvo e doenças (pg = 0,23, pd = 0,26 e pn = 0,51), notamos que as influências das categorias de genes-alvo (pg) e de doenças (pd) são semelhantes, uma vez que

os valores de seus pesos são muito próximos. Sendo assim, optamos por adotar os pesos obtidos para os miRNAs com anotações somente na categoria de doenças (*pd* = 0,84 e *pn* = 0,16) para os miRNAs possuem anotações somente sobre a categoria de genes-alvo também. Finalmente, a definição da equação de *CFSim* está descrita na Equação 23.

$$CFSim(M1, M2) = \begin{cases} 0{,}23 * simG + 0{,}26 * simD + 0{,}51 * simN \\ 0{,}84 * simG + 0{,}16 * simN \\ 0{,}84 * simD + 0{,}16 * simN \end{cases} \quad (23)$$

Apesar de existirem outros tipos de anotações na estrutura da ontologia OMIT, e portanto integradas a nossa base de dados, estas não puderam ser adicionadas como parâmetros em nosso método de similaridade funcional composta. Por exemplo, em relação aos dados de eventos patológicos associados aos miRNAs, somente a versão 5 *TarBase* (com apenas 1300 associações entre genes-alvo e miRNAs experimentalmente validadas) continha estas informações, porém em muito pequena quantidade. Além disso, recentemente foi disponibilizada a versão 6 (com mais de 65.000 associações) e sem apresentar os dados de eventos patológicos. Outro exemplo, é o caso das associações entre miRNAs e proteínas, o qual também foi removido na versão 6 do *TarBase*, mas de qualquer forma, também não poderia ser utilizado na metodologia proposta pois representaria uma redundância com o relacionamento entre miRNAs e genes-alvo. E por fim, as informações sobre os experimentos também foram desconsideradas no cálculo da similaridade pois são apenas formas distintas que foram utilizadas para validação experimental dos miRNAs, não podendo, portanto, influenciar na similaridade funcional entre os mesmos.

5.3. Ambiente de Desenvolvimento

Toda a implementação para o desenvolvimento do método, citada acima, foi feita utilizando a linguagem de programação *C#* através do *Visual Studio*[3], de modo a facilitar a criação do framework web que oferecerá a infraestrutura para utilização do método proposto. E cada método que realizava o acesso às anotações contidas em nosso banco de dados *CFSim Database*, assim como suas inserções, atualizações e deleções possuíam instruções escritas em linguagem SQL para o gerenciador *PostgreSQL*[4].

[3] http://www.microsoft.com/visualstudio
[4] http://www.postgresql.org.br/

CAPÍTULO 6. Framework Web

Esta seção contém a descrição funcional do framework web desenvolvido neste projeto, o qual pode ser acessado através da URL *dcm.ffclrp.usp.br/caib/?pg=cfsim*.

6.1. Proposta

Após o desenvolvimento do método de similaridade funcional composta *CFSim* entre miRNAs, implementamos um framework para um ambiente web que tem como objetivo permitir que o usuário acesse e atualize as anotações contidas em nossa base de dados integrada , e execute o cálculo da similaridade entre dois miRNAs.

6.2. Implementação

A implementação do ambiente web foi realizada contemplando os seguintes requisitos funcionais:

- Consulta sobre informações pertinentes aos miRNAs (nome, sequencia, família, organismo e natureza);
- Consulta sobre as funções de miRNAS;
- Consulta sobre relacionamentos entre miRNAS e genes-alvo;
- Consulta sobre relacionamentos entre miRNAS e doenças humanas;
- Inserção de novos miRNAS;
- Inserção de novas funções de miRNAS;
- Inserção de novas associações entre miRNAs e genes-alvo;
- Inserção de novas associações entre miRNAs e doenças humanas;
- Inserção de novos genes-alvo;
- Inserção de novas doenças humanas;
- Alteração ou deleção de miRNAS;
- Alteração ou deleção de funções de miRNAS;
- Alteração ou deleção de associações entre miRNAS e genes-alvo;
- Alteração ou deleção de associações entre miRNAS e doenças humanas;
- Cálculo da similaridade funcional composta entre dois miRNAs por meio do *CFSim*.

Além disso, também foram identificados alguns requisitos não funcionais:

- Interface Amigável: a interface ambiente Web deve apresentar suas funcionalidades de maneira intuitiva para o usuário;

- Tempo de resposta: o ambiente deve apresentar um tempo de resposta de curto quando da execução de suas funcionalidades;
- Usabilidade: o ambiente deve possibilitar que o usuário realize suas tarefas sem nenhum treinamento e com maior qualidade e satisfação.
- Portabilidade: o ambiente possibilitará o acesso de qualquer máquina contanto que esteja conectada à uma rede, não sendo necessária nenhuma etapa de instalação.

Sendo assim, o princípio da implementação do framework web se iniciou com sua integração com o *CFSim Database* de modo a permitir o acesso e a manipulação das anotações contidas no banco. Para realizar a busca das anotações a serem exibidas ao usuário, quando este realiza uma consulta, foram implementdos métodos de busca de miRNAs e todos os seus relacionamentos existentes no *CFSim Database*. Por outro lado, para a inserção, alteração e deleção de anotações foram criados métodos que executassem tais funções em nosso banco de dados. Finalmente, para o cálculo do *CFSim* foi utilizada toda a metodologia citada na seção anterior.

As interfaces criadas para a implementação desses requisitos estão apresentadas no Apêndice.

6.2.1. Ambiente de Desenvolvimento

Este framework web foi desenvolvido utilizando a linguagem de programação *C#* através do *Visual Studio*[5]. O acesso às anotações contidas no de dados *CFSim Database*, assim como suas inserções, atualizações e deleções são definidos em linguagem SQL para o gerenciador *PostgreSQL*[6].

[5] http://www.microsoft.com/visualstudio
[6] http://www.postgresql.org.br/

CAPÍTULO 7. Resultados

Esta seção contém os resultados obtidos por meio do método proposto e desenvolvido neste projeto, assim como uma visualização da infraestrutura do framework web criado para o usuário poder utilizar nosso banco de dados *CFSim Database* e o método de similaridade funcional composta *CFSim*.

7.1. Método de Similaridade Funcional Composta CFSim

Após a definição do método de similaridade funcional composta *CFsim* para dois miRNAs, com base na combinação das similaridades de cada categoria de informação sobre os dois miRNAs que estão sendo comparados, conforme a Equação 23, o próximo passo consitiu em aplicar nosso método sobre os dados e avaliar os resultados obtidos.

Buscando por uma forma de avaliação da eficácia do método proposto, inicialmente, optamos por utilizar como referência de similaridade as informações sobre as vias metabólicas das quais os genes-alvo dos miRNAS participam, considerando que os miRNAs são similares na proporção da quantidade de vias metabólicas que os seus genes-alvo possuem em comum.. Tais informações foram retiradas da base de dados KEGG (*Kyoto Encyclopedia of Genes and Genomes*) [Kanehisa-2000; Kanehisa-2004], a qual abriga registros de relacionamentos entre genes e vias metabólicas. Sendo assim, tentamos encontrar um fator de correlação entre os genes e suas vias metabólicas. Para tanto, implementamos um algoritmo que realiza o cálculo da similaridade entre os dois genes com base na GO e a similaridade entre as duas vias metabólicas para cada par de relacionamentos. A similaridade entre os genes foi calculada com base no método de J. Z. Wang e D. Wang, enquanto que a similaridade entre vias metabólicas foi definida da seguinte forma: dados dois conjuntos de vias metabólicas $V1=\{v_{11}, v_{12}, ..., v_{1m}\}$ e $V2=\{v_{21}, v_{22}, ..., v_{2n}\}$, a similaridade de vias metabólias (*simV*) entre *V1* e *V2* é dada pela intersecção sobre a união dos conjuntos, em termos de quantidade de vias, como mostrado na Equação 24.

$$simV = \frac{V1 \cap V2}{V1 \cup V2} \quad (24)$$

Posteriormente, plotamos um gráfico do tipo *scatterplot* no qual o eixo X representa o valor de similaridade entre genes e o eixo Y representa o valor de similaridade entre as respecitivas vias metabólicas. Entretanto, o resultado obtido não foi compatível com nossa ideia inicial, ou seja, nos deparamos com a inexistência de correlação entre as similaridades de genes e as similaridades de vias metabólicas. Para um mesmo valor de similaridade de

gene, a similaridade de vias metabólicas varia muito, e isto ocorre para todos os valores de similaridades de gene, sem apresentar nenhuma tendência de correlação,, conforme mostrado na Figura 14. Portanto, concluímos que a similaridade de vias metabólicas das quais participam os genes- não pode ser usada para avaliar similaridade entre microRNAs.

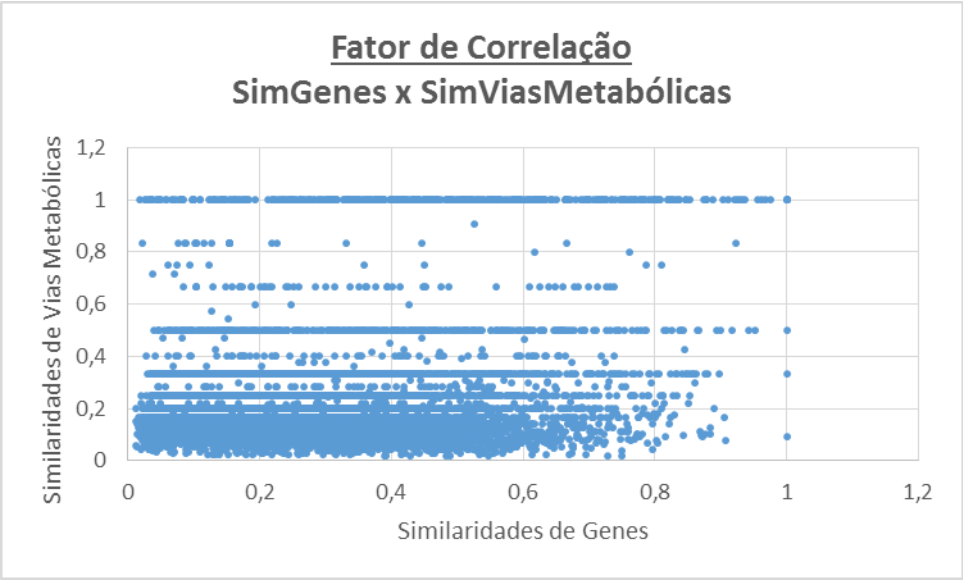

Figura 14. Fator de correlação entre a similaridade de genes-alvo e a similaridade de vias metabólicas utilizando dados da base de dados KEGG.

Outra referência de similaridade entre dois miRNAs, frequentemente utilizada na literatura, baseia-se na pertinência ou não dos mesmos à mesma família. Há muitas evidências que revelam que miRNAs de uma mesma família são susceptíveis de ter funções semelhantes, uma vez que suas sequencias *seed* regulam um conjunto comum de genes-alvo, e consequentemente, é mais provável que também estejam associados a doenças similares [Yu-2007; Lu-2008; Kaczkowski-2009; Wang-2010-b, Xu-2011]. Portanto, a validação de métodos de similaridade funcional entre miRNAs é comumente efetuada com base no fato de as instâncias sendo comparadas pertencerem ou não à mesma família [Wang-2010b; Sun-2013; Xuan-2013]. Sendo assim, os miRNAs utilizados foram agrupados por família. Para exemplificar, selecionamos algumas famílias distintas de miRNAs com dois ou mais membros por família e com grande diversidade em relação à categoria de natureza, ou seja, havia miRNAs oncogenicos, miRNAs supressores e outros com esta informação sendo desconhecida, conforme mostrado na Tabela 1.

Tabela 1. Exemplos de MiRNAs agrupados por famílias, com suas respectivas naturezas oncogenica e/ou supressora de tumor e quantidades de genes-alvo e doenças associados.

Família	MiRNA	Oncogenico	Supressor	Qtd Genes-Alvo	Qtd Doenças
let-7	hsa-let-7b	-	Sim	512	37
let-7	hsa-let-7c	-	Sim	5	42
let-7	hsa-let-7d	-	Sim	83	35
let-7	hsa-let-7e	-	Sim	5	33
let-7	hsa-let-7g	-	Sim	21	34
let-7	hsa-mir-98	-	Sim	515	11
mir-146	hsa-mir-146a	Sim	-	60	70
mir-146	hsa-mir-146b	Sim	-	1	30
mir-17	hsa-mir-106a	Sim	-	6	28
mir-17	hsa-mir-106b	Sim	-	78	32
mir-17	hsa-mir-17	Sim	-	9	65
mir-17	hsa-mir-18a	Sim	-	54	46
mir-17	hsa-mir-20a	Sim	-	6	50
mir-17	hsa-mir-20b	Sim	-	5	16
mir-17	hsa-mir-93	Sim	-	248	20
mir-221	hsa-mir-221	Sim	-	9	58
mir-221	hsa-mir-222	Sim	-	9	53
mir-27	hsa-mir-27a	Sim	-	4	23
mir-27	hsa-mir-27b	-	-	3	15
mir-29	hsa-mir-29a	-	Sim	8	46
mir-29	hsa-mir-29b	-	-	4	36
mir-29	hsa-mir-29c	-	Sim	35	31
mir-30	hsa-mir-30a	-	-	397	24
mir-30	hsa-mir-30b	-	-	59	16
mir-30	hsa-mir-30d	-	-	3	14
mir-8	hsa-mir-141	-	Sim	3	27
mir-8	hsa-mir-200a	-	Sim	4	43
mir-8	hsa-mir-200b	-	Sim	2	49
mir-8	hsa-mir-200c	-	Sim	5	52
mir-8	hsa-mir-429	-	-	21	25

Aqui serão apresentados exemplos de resultados obtidos por meio do cálculo do *CFSim* para miRNAs de uma mesma família e para miRNAs de famílias distintas para cada um dos seguintes seguintes casos: quando ambos miRNAs possuem anotações sobre a categoria de genes-alvo e de doenças, quando possuem anotações somente sobre a categoria de genes-alvo e por último, somente sobre a categoria de doenças. Vale lembrar que em todos estes casos a categoria de natureza é levada em consideração, uma vez que trata situações em que os dois miRNAs que estão sendo comparados são oncogenicos ou supressores de tumor,

quando são de naturezas opostas e até mesmo quando as informações sobre suas naturezas ainda são desconhecidas.

Os miRNAs hsa-let-7d e hsa-let-7g, ambos da família let-7, possuem anotações sobre as categorias de genes-alvo e doenças e apresentaram um valor de *CFSim* igual a 0,83. Agora, comparando um deles, como o hsa-let-7d, com um miRNA de outra família, como o hsa-mir-221 pertencente a família mir-221, o valor de *CFSim* foi igual a 0,28. Já os miRNAs hsa-miR-29b-3p, hsa-miR-29c-3p e hsa-mir-148a-3p possuem anotações somente sobre a categoria de genes-alvo. Comparando os dois primeiros, os quais são da família mir-29, obtivemos uma *CFSim* igual a 0,78, enquanto que o valor de *CFSim* para hsa-miR-29c-3p e hsa-mir-148a-3p, das famílias mir-29 e mir-148 respecticamente, foi de 0,37. Por fim, utilizando os miRNAs hsa-mir-128-1 e hsa-mir-128-2 da família mir-128 e o miRNA hsa-mir-301a, pertencente a família mir-130, todos com anotações somente na categoria de doenças, o *CFSim* entre os dois primeiros foi de 0,82 e o *CFSim* entre o primeiro e o último foi igual a 0,14. Os exemplos citados acima estão ilustrados na seção 7.2, com informação mais detalhada sobre as anotações que foram utilizadas em cada caso.

Para facilitar a análise dos resultados obtidos, executamos o cálculo de *CFSim* para todos os pares de miRNAs de todas as famílias. Posteriormente, foi realizado o cálculo da média dos valores obtidos pelo *CFSim* para os miRNAs de mesma família (média-intra) e para os miRNAS de famílias distinas (média-inter) para cada par de famílias de miRNAs. Em seguida, calculamos a média das médias-intras e das médias-inter para todas as famílias. Assim, obtivemos uma média de *CFSim* entre miRNAs de mesma família igual a 70%, enquanto a média *CFSim* entre miRNAs de diferentes famílias foi igual a 44%. Portanto, a diferença foi de 26% entre os valores de *CFSim* obtidos entre miRNAs de mesma família e entre miRNAs de famílias distintas.

Para efeito de comparação, realizamos o cálculo de similaridade utilizando o método original de J. Z. Wang e D. Wang para cada categoria de informação independemente, reproduzindo os trabalhos já existentes na literatura, apresentados na seção 3. Primeiramente, utilizamos somente a categoria de doenças, assim como Wang [Wang-2010b] em seu método MISIM, descrito na seção 3.1; e posteriormente, utilizamos somente a categoria de genes-alvo, assim como Yu [Yu-2011] em seu método GOSemSim, descrito na seção 3.2. Assim como feito na análise de nossos resultados, também foi relizado o cálculo da média dos valores obtidos para os miRNAs de mesma família (média-intra) e para os miRNAS de famílias distinas (média-inter) para cada par de famílias de miRNAs, seguido, do cálculo da média das médias-intras e das médias-inter para todas as famílias, para os resultados obtidos

pela reprodução de MISIM e GOSemSim. Os resultados obtidos por MISIM, apresentaram uma média entre miRNAs de mesma família igual a 63%, enquanto a média entre miRNAs de diferentes famílias foi igual a 46%. Neste caso, a diferença entre os valores de MISIM obtidos entre miRNAs de mesma família e entre miRNAs de famílias distintas foi de 17%. Por outro lado, os resultados obtidos por GOSemSim, apresentaram uma média entre miRNAs de mesma família igual a 52%, enquanto a média entre miRNAs de diferentes famílias foi igual a 36%. Sendo assim, a diferença entre os valores de GOSemSim obtidos entre miRNAs de mesma família e entre miRNAs de famílias distintas era equivalente a 16%.

Além disso, também realizamos a execução de um método que levasse em consideração a categoria de informação sobre genes-alvo e de doenças, como MISIM e GOSemSim, respectivamente. Desta vez, os resultados obtidos, apresentaram uma média entre miRNAs de mesma família igual a 58%, enquanto a média entre miRNAs de diferentes famílias foi igual a 41%. Sendo assim, a diferença entre os valores de MISIM combinado com GOSemSim obtidos entre miRNAs de mesma família e entre miRNAs de famílias distintas era equivalente a 17%. A Tabela 2 mostra a comparação entre as médias dos valores resultantes para o cálculo da similaridade utilizando s métodos MISIM, GOSemSim, a combinação de ambos e o método proposto neste estudo *CFSim*.

Tabela 2. Análise comparativa da média das similaridades funcionais entre miRNAs obtidas pelos métodos MISIM, GOSemSim, pela combinação de ambos e pelo método proposto *CFSim*.

Método	Intra-Família	Inter-Famílias	Diferença
MISIM	0,63	0,46	0,17
GOSemSim	0,52	0,36	0,16
MISIM+GOSemSim	0,58	0,41	0,17
CFSim	0,70	0,44	0,26

Portanto, analisando os nossos resultados por meio do método *CFSim* em comparação com os demais métodos existentes na literatura, verificamos que o cálculo da similaridade funcional entre dois miRNAs utilizando a combinação das diversas categorias de informação simultaneamente, apresentou melhores resultados. O uso da combinação das informações propostas em trabalhos já existentes, ou seja sobre as categorias de genes-alvo e de doenças, combinadas com anotações sobre a natureza oncogenica ou supressora de tumor dos miRNAs que estão sendo comparados acabam por elevar o valor da similaridade funcional entre miRNAs pertencentes a uma mesma família e reduzir o valor da similaridade funcional entre

miRNAs de famílias distintas. Estes resultados são compatíveis com as evidencias de que os miRNAs pertencentes a uma mesma família, os quais possuem sequencias muito semelhantes e regiões *seed* idênticas, apresentam funcionalidades similares.

7.2. Web Framework

Para alcançar os objetivos da proposta deste framework web de forma eficiente e amigável ao usuário, suas funcionalidades foram organizadas na interface em duas seções, "*Home*" e "*Insert / Delete*".

A seção "*Home*" é acessada através da primeira aba presente na interface, compondo a página inicial do framework assim que este é aberto. Nesta aba o usuário é capaz de pesquisar informações sobre o miRNA de interesse e cálcular a similaridade funcional entre 2 miRNAs por meio do método *CFSim*. Ao realizar uma busca por anotações de um miRNA, o framework retorna ao usuário os dados sobre o miRNA de interesse contidos no banco de dados unificado *CFSim Database*, tais como nome, família, sequencia, organismo, natureza oncogenica ou supressora de tumor e funções. O framework web também mostra ao usuário informações sobre as associações do miRNA de interesse com genes-alvo, como nome e símbolo do gene-alvo, código *ensembl*, tipo de experimento utilizado na validação, evento patológico relacionado e pubmedID de onde estas informações foram retiradas. Além disso, também são apresentadas ao usuário as associações entre o miRNA pesquisado e doenças humanas, com informações como nome da doença, pubmedID de onde foi retirado este dado e algumas anotações. Por outro lado, quando o usuário realiza o cálculo de similaridade *CFSim* entre 2 miRNAs, é retornado o valor numérico da similaridade funcional composta e algumas informações mais detalhadas de ambos os miRNAs, como nomes, famílias, genes-alvo e doenças associadas e se são de natureza oncogenica ou supressora de tumor.

A segunda aba "*Insert / Delete*" permite que o usuário atualize as anotações no banco de dados *CFSim Database*, por meio da inserção de novas informações ou da alteração e deleção das informações já existentes. Para isso, ele deve primeiramente buscar por informações sobre o miRNA de interesse, garantindo que este ainda não exista no banco de dados. Caso este miRNA não esteja contido no *CFSim*, ele poderá inserir novas informações. Por outro lado, se já existir informações sobre este miRNA no *CFSim*, o usuário poderá alterá-las ou removê-las. A especificação do pubmedID da publicação de onde as informações foram retiradas é obrigatória, de forma a garantir a integridade dos dados no *CFSim Database*. Juntamente com o pubmedID, o usuário também pode inserir o nome do autor

principal e o ano da publicação (em forma de citação) e algumas anotações. Consequentemente, há também uma funcionalidade na qual o usuário pode visualizar tais informações mais detalhadas sobre a publicação. A cada nova inserção de associações entre genes-alvo para um miRNA de interesse, o framework permite que o usuário inclua um novo gene-alvo ou escolha um gene-alvo já existente em nosso banco. O mesmo acontece para as associações entre o miRNA de interesse e doenças humanas. No caso das deleções, sempre que um usuário optar por remover alguma informação, aparecerá uma mensagem para que ele confirme a deleção. Assim, garantimos que não ocorra uma remoção de dados não intencional.

Há também uma terceira aba, "*About*", que contém a descrição do framework e do método *CFSim*, assim como o contato dos responsáveis pelo projeto.

Algumas figuras apresentando o *layout* de cada umas das abas do web framework juntamente com alguns dos resultados obtidos pelo método *CFSim* aplicado em miRNAs de mesma famílias e de famílias distintas encontram-se no Apêndice deste documento.

CAPÍTULO 8. Conclusões e Trabalhos Futuros

Esta seção contém as conclusões e alguns trabalhos futuros para o método proposto e desenvolvido neste projeto.

8.1. Conclusões

Atualmente, existem diversas bases de dados contendo informações sobre miRNAs. Entretanto, a falta de padronização e organização de muitas delas acaba por dificultar o estudo de pesquisadores da área. Sendo assim, para podermos concretizar nossa proposta de um novo método de similaridade funcional entre miRNAs, primeiramente tivemos de estudar diferentes bases de dados existentes na literatura para posteriormente criar um banco de dados integrado, chamado *CFSim Database*, com as anotações de miRNA provenientes de fontes de dados heterogêneas já existentes, respeitando a padronização do miRBase, a nomenclatura de genes definida pelo HUGO e a nomeclaruta de doenças do MeSH. Além disso, o *CFSim Database* foi organizado respeitando a estrutura da primeira ontologia de miRNAs OMIT.

Conforme apresentado na seção 3, os trabalhos existentes na literatura sobre o cálculo de similaridade funcional entre miRNAs ou utilizam a similaridade funcional entre seus genes-alvos com base na GO ou a similaridade funcional entre suas doenças relacionadas com base no MeSH. O presente estudo propõe um novo método, chamado *CFSim*, que utiliza diferentes tipos de anotações de miRNAs simultaneamente para calcular similaridade funcional composta entre dois miRNAs, inclusive suas associações com genes-alvo e doenças. Ou seja, se um miRNA possui somente informações sobre suas doenças associadas, um método baseado em somente seus genes-alvo não conseguiria utilizá-lo para calcular um valor de similaridade com outro miRNA, já o nosso método consegue. E se um miRNA só possui informações sobre seus genes-alvo, um método baseado somente em seus relacionamentos com doenças não conseguiria calcular a similaridade entre este miRNA e um outro, e mais uma vez o *CFSim* consegue. Portanto, o nosso método é mais abrangente e flexível que os demais já existentes na literatura. O *CFSim* considera a similaridade de genes-alvo com base na GO, a similaridade de doenças com base na MeSH e a similaridade da natureza oncogênica ou supressora de tumor dos miRNAs que estão sendo comparados, e então combina essas similaridades a fim de obter um valor final para miRNA a similaridade funcional composta.

Considerando-se o fato de que miRNAs pertencentes a uma mesma família possuem sequencias muito semelhantes e regiões seed idênticas, regulando um conjunto comum de

genes-alvo e consequentemente apresentam alta similaridade funcional, utilizamos as famílias como forma de validação. Sendo assim, os resultados obtidos por meio do *CFSim*, levando em consideração uma maior gama de anotações sobre miRNAs, apresentaram uma melhora em relação a metodologias utilizadas nos trabalhos já existentes na literatura. O método GOSemSim que utiliza somente associações com genes-alvo resulta em uma média de similaridade entre miRNAs da mesma família de 52% e o método MISIM que utiliza somente associações com doenças resulta em uma média de similaridade entre miRNAs da mesma família de 63%. E a combinação entre estes dois métodos MISIM e GOSemSim resulta em uma similaridade entre miRNAs da mesma família de 58%. Por outro lado, o método proposto neste estudo, o qual utiliza não só a combinação das associações com genes-alvo e doenças dos miRNAs comparados mas também as anotações sobre sua natureza oncogênica ou supressora de tumor dos miRNAs obteve uma média similaridade entre miRNAs pertencentes a mesma família de 70%. Desta forma, a diferença entre os valores de *CFSim* obtidos entre miRNAs de mesma família e entre miRNAs de famílias distintas foi equivalente a 26%, sendo maior que a diferença de 17% para os valores de MISIM, de 16% para os valores de GOSemSim e de 17% para valores da combinação destes dois últimos. Portanto, nosso método mostrou ser mais eficiente para calcular a similaridade funcional entre dois miRNAS, corroborando com o fato de que miRNAs de mesma família são funcionalmente mais similares.

Além disso, neste estudo também foi desenvolvido um framework web no qual o usuário pode pesquisar, inserir, atualizar e excluir informações do *CFSim Database* e calcular a similaridade funcional composta entre dois miRNAs usando o método proposto *CFSim*.

Em conclusão, o método proposto tem o potencial de auxiliar os pesquisadores de biomedicina a compreender melhor os importantes papéis dos miRNAs em processos biológicos, suas funções e suas associações com várias doenças humanas, especialmente considerando-se a similaridade funcional entre eles. Além disso, o *CFSim Database*, o qual é um banco de dados contendo diversas anotações de miRNAs oriundas de bancos de dados diferentes integradas, padronizadas e bem organizadas estruturalmente, assim como o framework web com uma interface amigável que permite a atualização destas anotações a qualquer momento facilitam a pesquisa e o estudo de miRNAs por parte de especialistas na área.

8.2. Trabalhos Futuros

A versão atual de nosso método de similaridade composta *CFSim* realiza o cálculo de similaridade de miRNAs, mais especificamente, na categoria de genes-alvo, somente sobre a sub-ontologia de funções moleculares de GO. Um ponto a ser investigado é aplicar o método nas demais sub-ontologias da GO: componente celular e processo biológico.

Além disso, nos experimentos realizados, o *CFSim* foi aplicado somente em miRNAs humanos. Como trabalho futuro, pretendemos calcular a similaridade entre miRNAs de outras espécies, quando houver dados suficientes de doenças associadas a miRNAs disponíveis para as demais espécies. Além disso, também iremos permitir a inserção de novas espécies em nosso banco de dados.

Em relação ao framework web que disponibiliza as informações do *CFSim Database* e o cálculo do método de similaridade, pretendemos implementar uma funcionalidade na qual o pesquisador pode requerer a criação de um usuário e senha para poder acessar o framework. Após isso, os administradores do sistema analisarão as informações do requerente e o incluirão em um dentre dois perfis: em um perfil o usuário pode somente visualizar as anotações de miRNA no *CFSim Database* e calcular a similaridade funcional composta, enquanto que no outro perfil o usuário pode realizar as mesmas funcionalidades do primeiro, além de incluir/alterar/deletar anotações de miRNAs em nosso banco de dados unificado. Assim, melhoramos nosso framewrok no quesito segurança.

REFERÊNCIAS BIBLIOGRÁFICAS

[Akleman-Chen-1999] Akleman, E. and J. Chen (1999). Generalized Distance Functions. Intl. Conf. on Shape Modeling and Applications, Aizu-Wakamatsu.

[Amaral-2010] Amaral, Bruna Aguiar do et al. (2010). MicroRNAs: biogênese, funções e seu papel potencial na carcinogênese oral. *Odontol. Clín.-Cient. (Online)* [online]. Vol.9, n.2 [citado 2013-02-05], pp. 105-109 .

[Ambros-2003] Ambros, V., Bartel, B., Barte, D. P. (2003). A Uniform System for MicroRNA Annotation. RNA, Vol. 9, No.3: 277-279.

[Ambros-2004] Ambros, V. (2004). The functions of animal microRNAs. Nature 431, 350–355

[Azuaje-2005] Azuaje F, Wang H, Bodenreider O (2005) Ontology-driven similarity approaches to supporting gene functional assessment. In: Proceedings of the ISMB 2005 SIG meeting on Bio-ontologies.

[Bartel-2009] Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions, Cell, 136,215-233.

[Bodenreider-2005] Bodenreider O, Aubry M, Burgun A (2005) Non-lexical approaches to identifying associative relations in the gene ontology. Pac Symp Biocomput. pp 91–102. URL http://view.ncbi.nlm.nih.gov/pubmed/15759617.

[Borst-1997] Borst, W. N. (1997) Construction of Engeneering Ontologies for Knowkedge Shating and Reuse.. (PhD Thesis). University of Twente, Enschede, The Netherlands.

[Budanitsky-1999] Budanitsky A (1999) Lexical semantic relatedness and its application in natural language processing. URL http://citeseer.ist.psu.edu/budanitsky99lexical.html.

[Budanitsky-2001] Budanitsky A (2001) Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. URL citeseer.ist.psu.edu/budanitsky01semantic.html.

[Cheng-2004] Cheng J, Cline M, Martin J, Finkelstein D, Awad T, et al. (2004) A knowledgebased clustering algorithm driven by gene ontology. Journal of Biopharmaceutical Statistics 14: 687–700.

[Cho-2007] Cho YR, Hwang W, Ramanathan M, Zhang A (2007) Semantic integration to identify overlapping functional modules in protein interaction networks. BMC Bioinformatics 8: 265+. doi:http://dx.doi.org/10.1186/1471-2105-8-265.

[Cimino-2006] Cimino, j.j., Zhu, X. (2006). The practical impact of ontologies on biomedical informatics. Yearbook of medical informatics, Vol. 45, Suppl 1, pp. 124:135

[Consortium-2000] The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nature Genetics

[Consortium-2004] The Gene Ontology Consortium (2004): The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research

[Consortium-2010] The Gene Ontology Consortium (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic acids research*,38, D331-5.

[Couto-2005] Couto FM, Silva MJ, Coutinho PM (2005) Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors. In: Proc. of the ACM Conference in Information and Knowledge Management as a short paper.

[Cowland-2007] Cowland JB, Hother C, Grønbæk K (2007). MicroRNAs and cancer. APMIS; 115: 1090-1106

[Dalmay-2008] Dalmay T. (2008) MicroRNAs and cancer. J Intern Med; 263:366-375.

[Du-2009] Du, Z., Li, L., Chen, C.F., Yu, P.S. and Wang, J.Z. (2009) G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery, Nucleic Acids Res, 37, W345-349.

[Duursma-2008] Duursma, A. M., Kedde, M., Schrier, M. & et al (2008). mir-148 targets human DNMT3b protein coding region. RNA , 872–877.

[Esquela-Kerscher-2006] Esquela-Kerscher A, Slack FJ (2006). Oncomirs - microRNAs with a role in cancer. Nat Rev Cancer 6: 259–269.

[Felipe-2005] Felipe, J. C. (2005) Desenvolvimento de métodos para extração, comparação e análise de características intrínsecas de imagens médicas, visando à recuperação perceptual por conteúdo. Tese de Doutorado.

[Filipowicz-2008] Filipowicz,W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat. Rev. Genet. 9, 102–114.

[Freitas-2003] Freitas, Frederico L. G. (2003) Ontologias e a Web Semântica. Santos: Universidade Católica de Santos. Disponível em: <http://www.inf.unisinos.br/~renata/cursos/topicosv/ontologias-ws.pdf >

[Gartel-2008] Gartel AL, Kandel ES. (2008) miRNAs: Little known mediators of oncogenesis. Semin Cancer Biol; 18: 103-110.

[Gentleman-2005] Gentleman R (2005) Visualizing and Distances Using GO. URL http://www. bioconductor.org/docs/vignettes.html.

[GO] Disponível em: <http://www.geneontology.org/>. Último acesso em: Mar. 2014.

[Goh-2007] Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. and Barabasi, A.L. (2007) The human disease network, Proc Natl Acad Sci U S A, 104, 8685-8690.

[Griffiths-Jones-2006] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 34, D140-144

[Griffiths-Jones-2010] S. Griffiths-Jones (2010), Curr. Protoc. Bioinformatics, ch. 12, Unit 12 19, pp. 11–10.

[Grubber-1995] Gruber, T.R. (1995) Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human Computer Studies, [S.l.]. v.43, n.5-6, p.907.

[Guarino-1998] Guarino, N. (1998) Formal ontology and information systems. In: International Conference on Formal Ontologies in Information Systems, FOIS, 1, Trento, Italy, Proceedings… Amsterdam: IOS Press, 1998. P. 3-15.

[Guo-2006] Guo X, Liu R, Shriver CD, Hu H, LiebmanMN (2006) Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics 22: 967–973. doi:10.1093/bioinformatics/btl042. URL http://www.ncbi.nlm.nih.gov/ entrez/query.fcgi?cmd = Retrieven&db = pubmedn&dopt = Abstractn&list uids=16492685.

[HGNC-2014] HUGO Gene Nomenclature Committee: http://www.genenames.org/ Último acesso em: Junho. 2014.

[Horvath-2006] Horvath, S., Zhang, B., Carlson, M., Lu, K.V., Zhu, S., Felciano, R.M., Laurance, M.F., Zhao, W., Qi, S., Chen, Z., Lee, Y., Scheck, A.C., Liau, L.M., Wu, H., Geschwind, D.H., Febbo, P.G., Kornblum, H.I., Cloughesy, T.F., Nelson, S.F. and Mischel, P.S. (2006) Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target, Proc Natl Acad Sci U S A, 103, 17402-17407.

[Huang-2010] J. Huang, M. Tan, D. Dou, L. He, J. Dang, and P. Hayes (2010). “Ontology for MiRNA Target Prediction in Human Cancer,” Proc. *First ACM International Conference on Bioinformatics and Computational Biology (BCB-10)*, Niagara Falls, NY.

[Iwasaki-Tomari-2009] Iwasaki, S. & Tomari, Y. (2009). Argonaute-mediated translational repression (and activation). Fly 3, 204–206.

[Jiang-1997] Jiang J, Conrath D (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. of the 10th International Conference on Research on Computational Linguistics, Taiwan.

[Kaczkowski-2009] Kaczkowski B, Torarinsson E, Reiche K, Havgaard JH, Stadler PF, et al. (2009) Structural profiles of human miRNA families from pairwise clustering. Bioinformatics 25: 291–294.

[Kanehisa-2000] Kanehisa, M. (2000); "Post-genome Informatics", Oxford University Press.

[Kanehisa-2004] Kanehisa,M. et al. (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res., 32, D277–D280

[Lau-2001] Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294: 858–862.

[Lee-1993] Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14* . *Cell* 75: 843–854.

[Lee-2001] Lee, R.C. and Ambros, V. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* 294: 862–864.

[Lee-2003] Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, et al. The nuclear RNase III Drosha initiates microRNA processing. Nature 2003;425(6956):415-9.

[Lee-HK-2004]Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14: 1085–1094. doi:http://dx.doi.org/10.1101/gr.1910904.

[Lee-SG-2004] Lee, S.G., Hur, J.U. and Kim, Y.S. (2004) A graph-theoretic modeling on GO space for biological interpretation of gene clusters, Bioinformatics, 20, 381-388.

[Lee-Y-2004] Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, et al. (2004) MicroRNA genes are transcribed by RNA polymerase II. Embo J;23(20):4051-60.

[Lei-2006] Lei Z, Dai Y (2006) Assessing protein similarity with gene ontology and its use in subnuclear localization prediction. BMC Bioinformatics 7.

[Lin-1998] Lin D (1998) An information-theoretic definition of similarity. In: Proc. of the 15th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann. pp 296–304.

[Lin-2007] Lin, J., Gan, C.M., Zhang, X., Jones, S., Sjoblom, T., Wood, L.D., Parsons, D.W., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., Parmigiani, G. and Velculescu, V.E. (2007) A multidimensional analysis of genes mutated in breast and colorectal cancers, Genome Res, 17, 1304-1318.

[Llave-2002] Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. (2002). Endogenous and silencing-associated small RNAs in plants. *Plant Cell* 14: 1605–1619.

[Lord-2003] Lord P, Stevens R, Brass A, Goble C (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19: 1275–1283.

[Lu-2008] Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., and Cui, Q. (2008) An analysis of human microRNA and disease associations, PLoS ONE, 3(10): e3420. doi:10.1371/journal.pone.0003420.

[Lund-2004] Lund E, Guttinger S, Calado A, Dahlberg JE, Kutay U. (2004) Nuclear export of microRNA precursors. Science;303(5654):95-8.

[Marietto-2002] Marietto, M. G. B.; David, N.; Sichman, J. S.; Coelho, H. (2002) “Infraestrutura para a Construção de Ontologias”. Working Paper 01-2002. Escola Politécnica da Universidade de São Paulo.

[McKusick-2007] McKusick VA (2007): Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet, 80(4):588-604

[MESH-2014] Disponível em: <http://www.nlm.nih.gov/mesh/>. Último acesso em: Junho. 2014.

[Mi-2011]Mi H and Thomas PD (2011) Ontologies and standards in bioscience research: for machine or for human. Front. Physio. 2:5. doi: 10.3389/ fphys.2011.00005

[MIRBBASE] Disponível em: <http://www.mirbase.org>. Último acesso em: Nov. 2013.

[Mistry-2008] Mistry M, Pavlidis P (2008) Gene ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics 9: URL http://www.biomedcentral. com/1471-2105/9/327.

[Mourelatos-2002] Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B.,Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. (2002). miRNPs: A novel class of ribonucleoproteins containing numerous micro-RNAs. *Genes & Dev.* 16: 720–728.

[NLM] Disponível em: <http://www.nlm.nih.gov/>. Último acesso em: Out. 2013.

[Noy-2001] Noy, F. N.; Guinness, D. L (2001). *Ontology development 101:* a guide to create your first ontology. Disponível em: http://ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.doc>

[OBO] Disponível em: <http://www.obofoundry.org/>. Último acesso em: Set. 2012.

[Othman-2007] Othman R, Deris S, Illias R (2007) A genetic similarity algorithm for searching the gene ontology terms and annotating anonymous protein sequences. J Biomed Inform 23: 529–38.

[Park-2002] Park, W., Li, J., Song, R., Messing, J., and Chen, X. (2002). CARPEL FACTORY, a Dicer Homolog, and HEN1, a Novel Protein, Act in microRNA metabolism in *Arabidopsis thaliana*. *Curr. Biol.* 12: 1484–1495.

[Pekar-2002] Pekar V, Staab S (2002) Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. In: Proceedings of the 19th international conference on Computational linguistics. Morristown, NJ, USA: Association for Computational Linguistics. pp 1–7. doi:http: //dx.doi.org/10. 3115/1072228.1072318.

[Pesquita-2007] Pesquita C, Faria D, Bastos H, Falca˜o AO, Couto F (2007) Evaluating gobased semantic similarity measures. In: ISMB/ECCB 2007 SIG Meeting Program Materials. International Society for Computational Biology.

[Pesquita-2008] Pesquita, C., Faria, D., Bastos, H., Ferreira, A.E., Falcao, A.O. and Couto, F.M. (2008) Metrics for GO based protein semantic similarity: a systematic evaluation, BMC Bioinformatics, 9 Suppl 5, S4.

[Pesquita-2009] Pesquita C, Faria D, Falca˜o AO, Lord P, Couto FM (2009) Semantic Similarity in Biomedical Ontologies. PLoS Comput Biol 5(7): e1000443. doi:10.1371/journal.pcbi.1000443

[PFAM] Disponível em: <http://pfam.sanger.ac.uk/>. Último acesso em: Jun. 2012.

[Piovezani-2013] Piovezani, Amanda Rusiska (2013). SIMTar: Uma ferramenta para predição de SNPs interferindo em sítios alvos de microRNAs. 85 p. Dissertação (Mestrado em Bioinformática) - Interunidades em Bioinformática, Universidade de São Paulo, São Paulo.

[Pozo-2008] Pozo AD, Pazos F, Valencia A (2008) Defining functional distances over gene ontology. BMC Bioinformatics 9: doi:http://dx.doi.org/10.1186/1471-2105-9-50.

[Quintana-2001] Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science* 294: 853–858.

[Rada-1989] Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. In: IEEE Transaction on Systems, Man, and Cybernetics. 19. pp 17–30.

[Reinhart-2000] Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21- nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* 403: 901–906.

[Reinhart-2002] Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. (2002). MicroRNAs in plants. *Genes & Dev.* 16: 1616–1626.

[Ren-2009] Ren, J., Jin, P., Wang, E., Marincola, F., Stroncek, D., (2009) MicroRNA and gene expression patterns in the differentiation of human embryonic stem cells. Journal of Translational Medicine, 7, 20.

[Resnik-1995] Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. In: Proc. of the 14th International Joint Conference on Artificial Intelligence. pp 448–453.

[Resnik-1998] Resnik, P. (1998), Semantic Similarity in a Taxonomy: An Information- Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artifical Intelligence Research, 11, 130, 95.

[Ricarte-2006] Ricarte Filho, Júlio C.M. and Kimura, Edna Teruko (2006). MicroRNAs: nova classe de reguladores gênicos envolvidos na função endócrina e câncer. *Arq Bras Endocrinol Metab* [online]. Vol.50, n.6

[Richardson-1994] Richardson R, Smeaton AF, Murphy J (1994) Using WordNet as a knowledge base for measuring semantic similarity between words. Technical Report CA-1294, Dublin, Ireland. URL citeseer.ist.psu.edu/richardson94using.html.

[Riensche-2007] Riensche RM, Baddeley BL, Sanfilippo AP, Posse C, Gopalan B (2007) Xoa: Web-enabled cross-ontological analytics. In: Services, 2007 IEEE Congress on. pp 99–105.

[Rosse-2003] Rosse C, Mejino JL Jr (2003). A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform; 36(6): 478-500.

[Roubelakis-2009] Roubelakis, M., Zotos, P., Papachristoudis, G., Michalopoulos,I., et al. (2009), Human microRNA target analysis and gene ontology clustering by GOmir, a novel stand-alone application. BMC Bioinformatics, 10, S20.

[Sasse-2008] Sassen S, Miska EA, Caldas C (2008). MicroRNA – implications for cancer. Virchows Arch; 452: 1-10.

[Schlicker-2006] Schlicker, A., Domingues, F. S., Rahnenführer, J., Lengauer, T., (2006). A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics, 7, 302.

[Schmittgen-2008] Schmittgen TD (2008). Regulation of microRNA processing in development, differentiation and cancer. J Cell Mol Med; 12: 1811-1819.

[Schulze-Kremer-1998] Schulze-Kremer, S. (1998) Ontologies for Molocular Biology. Proceedings of the Third Pacific Symposium on Biocomputing. AAAI Press, Hawaii, pp.693-704.

[Seco-2004] Seco N, Veale T, Hayes J (2004) An intrinsic information content metric for semantic similarity in wordnet. In: ECAI. pp 1089–1090.

[Sethupathy-2006] Sethupathy,P., Corda,B. and Hatzigeorgiou,A.G. (2006) TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA, 12, 192–197.

[Sevilla-2005] Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, et al. (2005) Correlation between gene expression and go semantic similarity. In: IEEE/ ACM Transactions on Computational Biology and Bioinformatics.

[Shao-Yao-2007] Shao-Yao Ying, Joseph D. Miller and Shi-Lung Lin. (2007) Non-coding RNAs – development of man-made vector-based intronic microRNAs (miRNAs)

[Singh-2008] Singh SK, Bhadra MP, Girschick HJ, Bhadra U (2008). MicroRNAs – micro in size but macro in function. FEBS Journal; 275: 4929-4944.

[Smith-2007] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J.; OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S. A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255.

[Studer-1998] Studer R, Benjamins VR, Fensel D (1998). Knowledge Engineering: Principles and Methods. IEEE Transactions on Data and Knowledge Engineering 25(1-2):161–197.

[Sun-2008] Sun BK, Tsao H (2008). Small RNAs in development and disease. J Am Acad Dermatol; 59: 725-737.

[Sun-2013] Sun J, Zhou M, Yang H, Deng J, Wang L, et al. (2013) Inferring Potential microRNA-microRNA Associations Based on Targeting Propensity and Connectivity in the Context of Protein Interaction Network. PLoS ONE 8(7): e69719. doi:10.1371/journal.pone.0069719

[Tao-2007] Tao Y, Sam L, Li J, Friedman C, Lussier Y (2007) Information theory applied to the sparse gene ontology annotation network to predict novel gene function. Bioinformatics 23: 529–38.

[UNIPROT] Disponível em: <www.uniprot.org/>. Último acesso em: Mar. 2012.

[Vogt-2012] Vogt L, Grobe P, Quast B, Bartolomaeus T (2012) Accommodating Ontologies to Biological Reality—Top-Level Categories of Cumulative-Constitutively Organized Material Entities. PLoS ONE 7(1): e30004. doi:10.1371/journal.pone.0030004

[Wang-2004] Wang H, Azuaje F, Bodenreider O, Dopazo J (2004) Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In: Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB ’04. Proceedings of the 2004 IEEE Symposium on. pp 25–31. URL http://ieeexplore.ieee.org/xpls/abs all.jsp?arnumber=1393927

[Wang-2007] Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., Chen, C., (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics (Oxford, England); 23, 1274-1281.

[Wang-2010-a] Wang D, Qiu C, Zhang H, Wang J, Cui Q, et al. (2010) Human MicroRNA Oncogenes and Tumor Suppressors Show Significantly Different Biological Patterns: From Functions to Targets. PLoS ONE 5(9): e13067. doi:10.1371/journal.pone.0013067

[Wang-2010-b] Wang, D., Wang, J., Lu, M., Song, F., Cui, Q., (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics, 26, 1644-50.

[Weinmann-2009] Weinmann, L., Hock, J., Ivacevic, T. & et al (2009). Importin 8 is a gene silencing factor that targets argonaute proteins to distinct mRNAs. Cell 136, 496–507.

[Wu-1994] Wu Z, Palmer MS (1994) Verb semantics and lexical selection. In: Proceedings of the 32nd. Annual Meeting of the Association for Computational Linguistics (ACL 1994). pp 133–138. URL http://dblp.uni-trier. de/db/conf/acl/acl94.html#WuP94.

[Wu-2005] Wu H, Su Z, Mao F, Olman V, Xu Y (2005) Prediction of functional modules based on comparative genome analysis and gene ontology application. Nucleic Acids Res 33: 2822–2837. URL http://view.ncbi.nlm.nih.gov/pubmed/ 15901854.

[Wu-2006] Wu X, Zhu L, Guo J, Zhang DY, Lin K (2006) Prediction of yeast proteinprotein interaction network: insights from the gene ontology and annotations. Nucleic Acids Research 34: 2137–2150.

[Xu-2011] Xu J, Li CX, Li YS, Lv JY, Ma Y, et al. (2011) MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. Nucleic Acids Res 39: 825–836.

[Xu-2013] Xu J, Li Y, Li X, Li C, Shao T, Bai J, Chen H, Li X (2013): Dissection of the potential characteristic of miRNA–miRNA functional synergistic regulations. Mol BioSyst, 9:217–224

[Xuan-2013] Xuan P, Han K, Guo M, Guo Y, Li J, et al. (2013) Prediction of microRNAs Associated with Human Diseases Based on Weighted k Most Similar Neighbors. PLoS ONE 8(8): e70204. doi:10.1371/journal.pone.0070204

3-Y-31

[Ye-2005] Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, et al. (2005) Gene function prediction from congruent synthetic lethal interactions in yeast. Molecular Systems Biology 1: msb4100034–E1–msb4100034–E12. doi: http://dx.doi.org/ 10.1038/msb4100034.

[Yu-2005] Yu H, Gao L, Tu K , Guo Z (2005) Broadly predicting specific gene functions with expression similarity and taxonomy similarity. Gene 352: 75–81.

[Yu-2007] Yu F, Yao H, Zhu P, Zhang X, Pan Q, et al. (2007) Let-7 regulates self renewal and tumorigenicity of breast cancer cells. Cell 131: 1109–1123.

[Yu-2011] Guangchuang Yu, Chuan-Le Xiao, Xiaochen Bo, Chun-Hua Lu, Yide Qin, Sheng Zhan, Qing-Yu He (2011). A new method for measuring functional similarity of micrornas.

Journal of Integrated OMICS, 1(1):49–54, ISSN 2182-0287. doi: 10.5584/jiomics.v1i1.21. URL http://www.jiomics.com/index.php/jio/article/view/21.

[Zhang-2007] Zhang B, Pan X, Cobb GP, Anderson TA (2007). MicroRNAs as oncogenes and tumor suppressors. Dev Biol; 302: 1-12.

[Zhang-2010] Zhang et al. (2010): Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. BMC Medical Genomics 3:1

[Zisoulis-2012] Zisoulis, D. G., Kai, Z. S. & Chang, R. K. (2012). Autoregulation of microRNA biogenesis by let-7and Argonaute. Nature 486, 541–544.

APÊNDICE

Interfaces e funcionamento do sistema web

O ambiente web foi desenvolvido com o intuito de permitir que o usuário possa acessar as anotações sobre miRNAs de nosso banco de dados *CFSim Database*, como nome, família, sequencia, se atua como oncogene ou supressor de tumor, organismo ao qual pertence, experimentos realizados para sua validação e seus relacionamentos com genes-alvo e doenças humanas. Ainda, o usuário pode inserir informações sobre novas descobertas de miRNAs ou de associações com genes-alvo e doenças, assim como atualizar ou remover as anotações já existentes. Além disso, por meio deste framework o usuário também é capaz de calcular a similaridade funcional composta *CFSim* entre dois miRNAs proposta neste projeto, com base em todas estas anotações contidas no *CFSim Database*.

A Figura 15 mostra a páginal inicial de nosso ambiente, chamado "*Home*", no qual o usuário pode fazer a pesquisa sobre um miRNA de interesse (botão "*Search*") e realizar o cálculo do *CFSim* entre dois miRNAs (botão "*Calculate*").

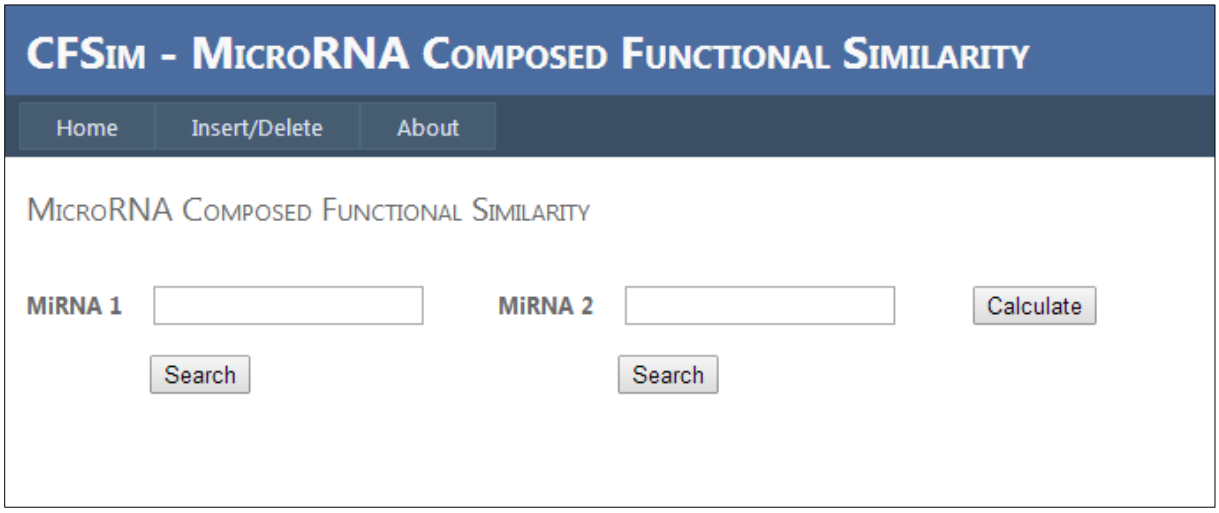

Figura 15. Página inicial "*Home*" do framework web.

Quando o usuário faz uma busca sobre um determinado miRNA na aba "*Home*", suas anotações contidas no *CFSim Database* são mostradas em tabelas, de forma a organizar as informações de forma mais amigável o possível para o usuário, confome a Figura 16. Sendo assim, a primeira tabela apresenta as informações básicas do miRNA, ou seja, nome, família, funções, sequencia, organismo e se atua como oncogênico ou supressor de tumor. A segunda tabela contém informações dos genes-alvo do miRNA de interesse, como nome e símbolo do

gene, sódigo Ensembl, experimento utilizado para sua validação, evento patológico e pubmedID de onde as informações foram retiradas. Por fim, a última tabela consiste dos relacionamentos entre o miRNA e as doenças humanas, apresentando colunas com o nome da doença, o pubmedID que cita a descoberta de tal relacionamento e algumas anotações. Note que as anotações do miRNA buscado na Figura 16 não estão completas, tendo sido reduzidas somente para visualização. O nosso framework disponibiliza ao usuário a botões de busca por anotações de ambos os miRNAs que serão comparados no cálculo de similaridade, facilitando assim a pesquisa e a analise de informações dos mesmos.

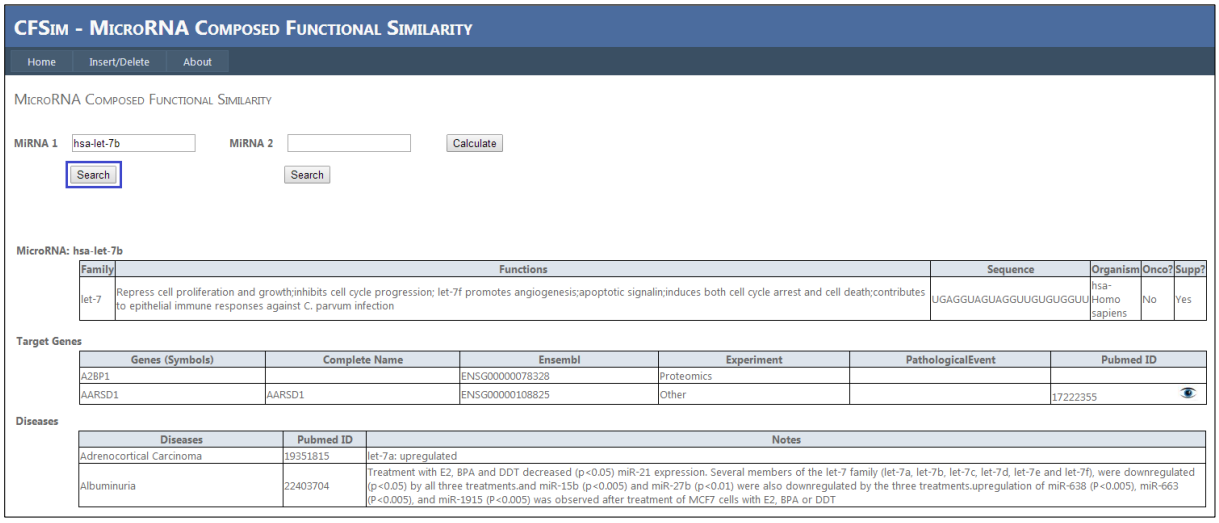

Figura 16. Página inicial "*Home*" do framework web com busca sobre miRNA hsa-let-7b.

Por outro lado, quando o usuário deseja realizar o cálculo do *CFSim* entre dois miRNAs, o valor final da similaridade funcional composta é apresentado ao lado do botão "*Calculate*" e na parte inferior do framework são apresentadas as anotações que foram utilizadas no método. Tais anotações apresentadas sobre os miRNAs comparados no cálculo da *CFSim* podem ser diferentes das anotações apresentadas ao usuário quando feita uma busca pelas anotações destes mesmo miRNAs através do botão "*Search*". Isto ocorre devido ao fato de que nem todas informações existentes no *CFSim Database* são utilizadas no método. Por exemplo, vamos supor que um miRNA possua associações com 10 genes-alvo diferentes, sendo assim, serão mostrados 10 registros de genes-alvo pela busca por meio do clique em "*Search*". Entretanto, no cálculo de *CFSim* somente serão utilizados aqueles genes-alvo que estiverem cadastrados na GO e que possuírem resgitros de funções moleculares associadas a eles. Sendo assim, a quantidade de registros de genes-alvo retornados pela busca

pode ser maior do que a quantidade apresentada como anotações utilizadas pelo método de similaridade funcional composta.

As figuras 17 a 22 a seguir apresentarão os resultados referente aos exemplos citados na seção anterior com mais detalhes, ou seja, os resultados obtidos por meio do cálculo do *CFSim* para miRNAs de uma mesma família e para miRNAs de famílias distintas para miRNAs com anotações sobre a categoria de genes-alvo e de doenças, miRNAs com anotações somente sobre a categoria de genes-alvo e por último, com anotações somente sobre a categoria de doenças.

Figura 17. Exemplo de cálculo do *CFSim* aplicado sobre os miRNAs hsa-let-7d e hsa-let-7g representando o caso em que os dois miRNAs pertencem a uma mesma família e ambos possuem anotações sobre a categoria de genes-alvo e também sobre a categoria de doenças.

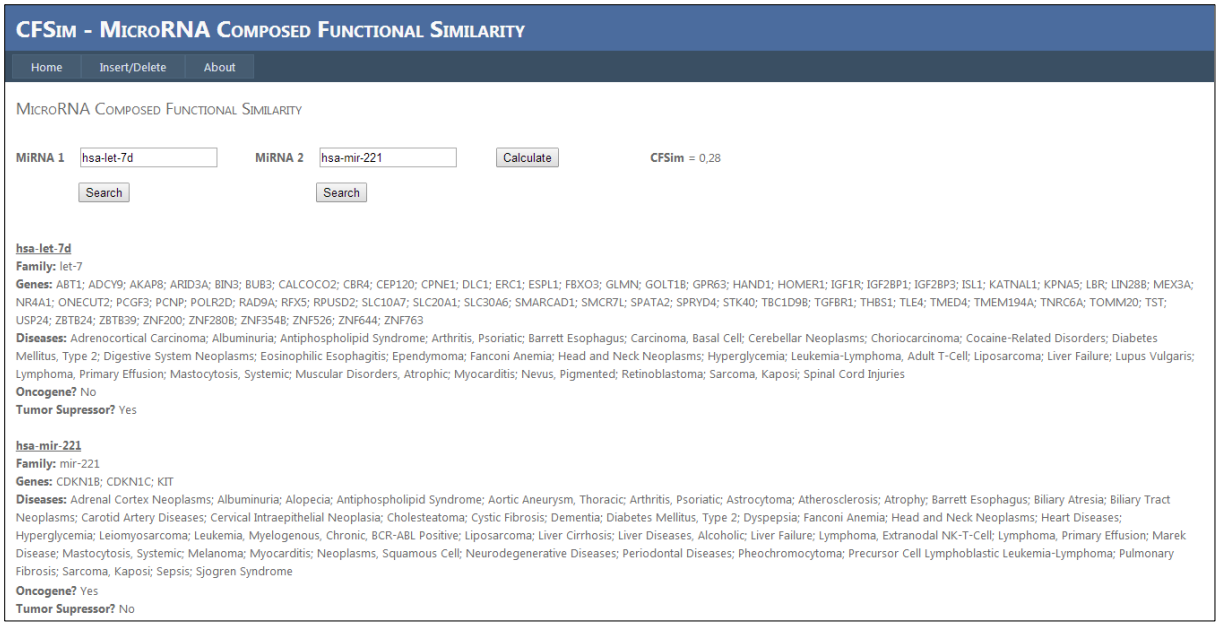

Figura 18. Exemplo de cálculo do *CFSim* aplicado sobre os miRNAs hsa-let-7d e hsa-mir-221 representando o caso em que os miRNAs pertencem a diferentes famílias, mas ambos possuem anotações sobre a categoria de genes-alvo e também sobre a categoria de doenças.

Figura 19. Exemplo de cálculo do *CFSim* aplicado sobre os miRNAs hsa-miR-29b-3p e hsa-miR-29c-3p representando o caso em que os dois miRNAs pertencem a uma mesma família e ambos possuem anotações somente sobre a categoria de genes-alvo.

Figura 20. Exemplo de cálculo do *CFSim* aplicado sobre os miRNAs hsa-miR-29b-3p e hsa-mir-148a-3p representando o caso em que os miRNAs pertencem a diferentes famílias, mas ambos possuem anotações somente sobre a categoria de genes-alvo.

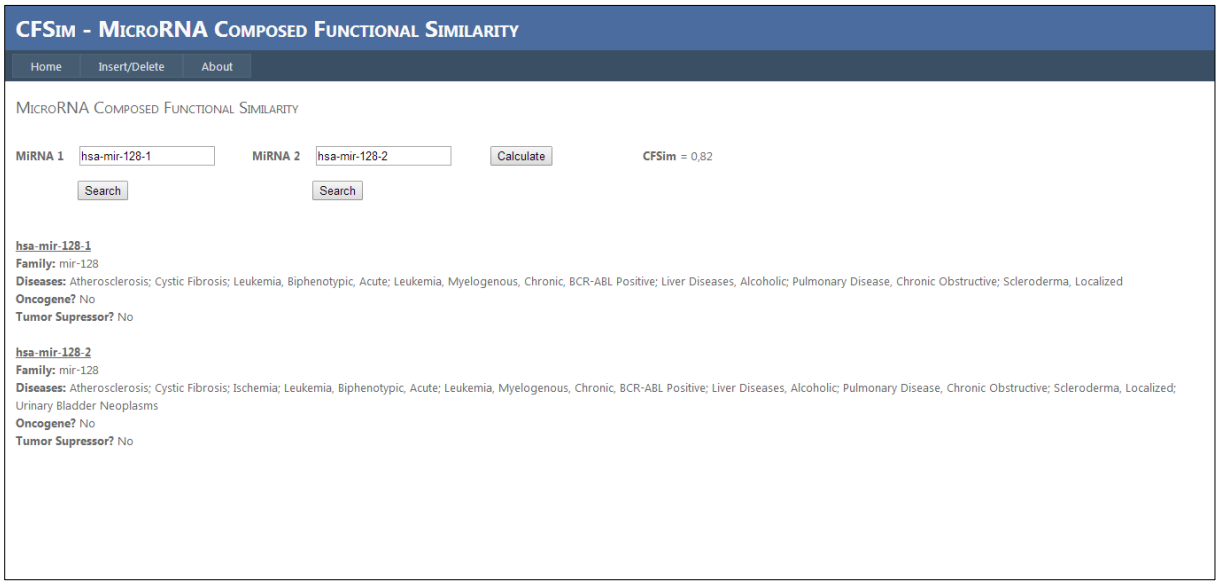

Figura 21. Exemplo de cálculo do *CFSim* aplicado sobre os miRNAs hsa-mir-128-1 e hsa-mir-128-2 representando o caso em que os dois miRNAs pertencem a uma mesma família e ambos possuem anotações somente sobre a categoria de doenças.

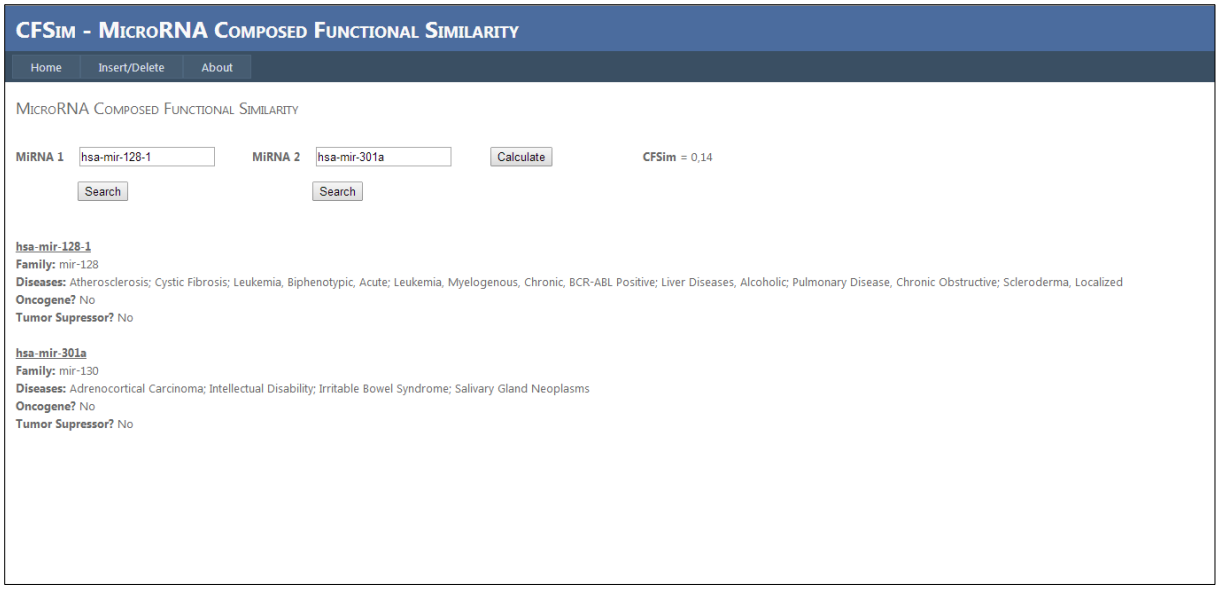

Figura 22. Exemplo de cálculo do *CFSim* aplicado sobre os miRNAs hsa-mir-128-1 e hsa-mir-301a representando o caso em que os miRNAs pertencem a diferentes famílias, mas ambos possuem anotações somente sobre a categoria de doenças.

Na aba "*Insert / Delete*", mostrada na Figura 23, o usuário pode inserir ou atualizar as informações sobre um determinado miRNA através do botão "*Save*". Caso queira remover um miRNA, basta utilizar o botão "*Delete*". Se o usuário desejar inserir novas anotações, deverá clicar no ícone azul e uma nova janela aparecerá e então o usuário poderá incluir novos dados. A Figura 24, por exemplo, é mostra uma janela de inserção de dados de associação entre o miRNA e doenças humanas, onde pode ser digitada uma nova doença ou o usuário cpode clicar no ícone de uma lupa e assim poderá escolher uma doença já existente no *CFSim Database*, conforme ilustrado na Figura 25. Por outro lado, se o usuário desejar remover alguma informação basta clicar no ícone vermelho e uma mensagem de confirmação será apresentada ao usário para garantir que não ocorram remoções indesejadas. Se o usuário desejar limpar todos os campos para realizar uma nova busca ou inserção ou atualização, basta clicar no botão "*Clear*". Ainda, para visualizar mais informações sobre a publicação de referencia de uma anotação, o usuário deverá clicar no ícone que contém um olho e uma nova janela abrirá com informações mais detalhadas sobre a mesma, como mostrado na Figura 26.

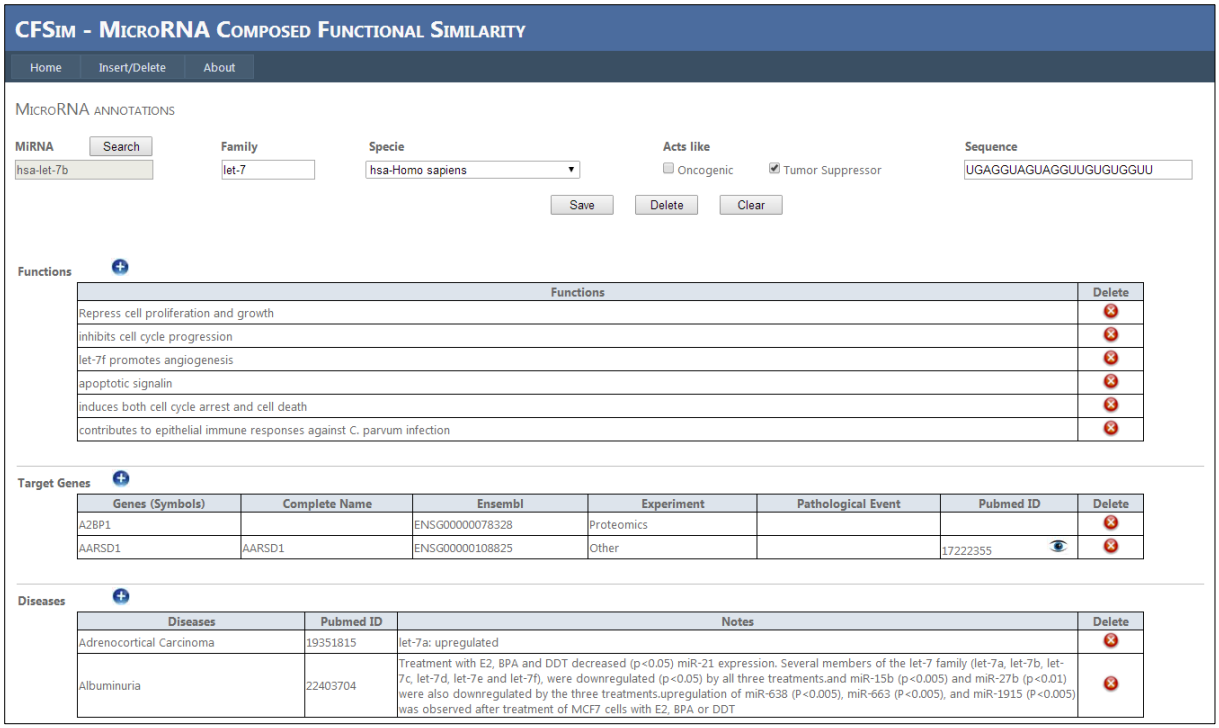

Figura 23. Aba "*Insert / Delete*" do framework web.

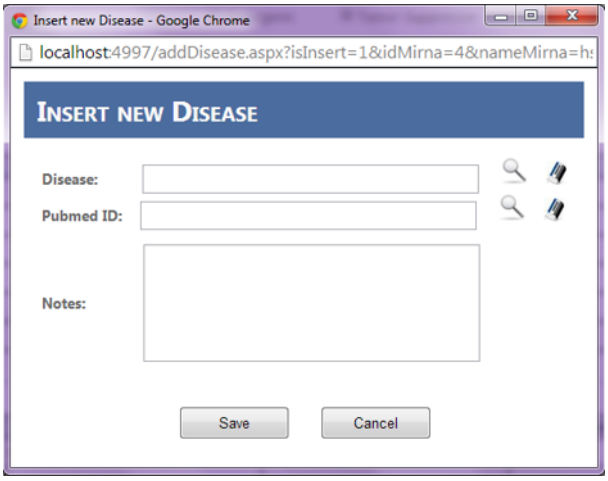

Figura 24. Janela de inserção de dados de associação entre o miRNA e doenças humanas.

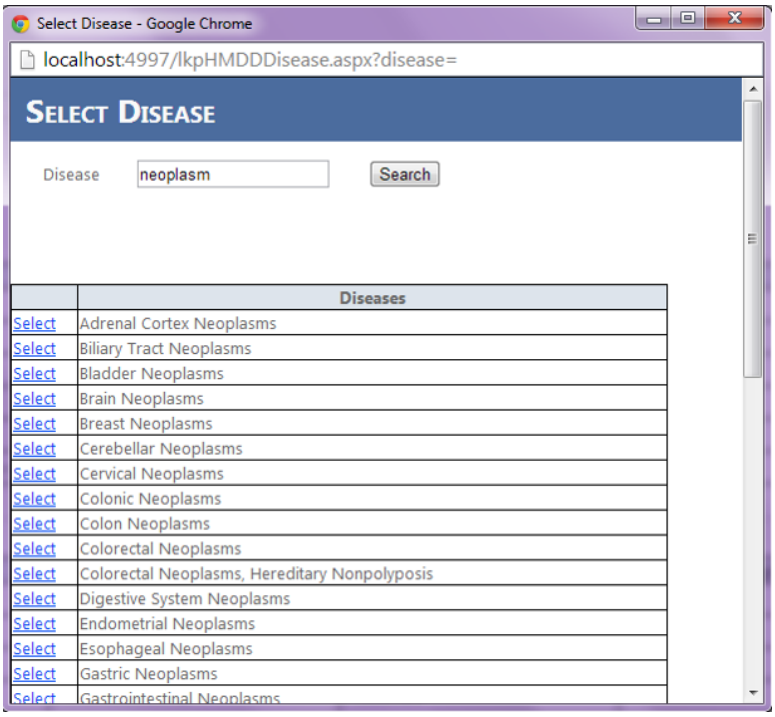

Figura 25. Janela na qual o usuário pode escolher uma doença já existente no *CFSim Database* para associar a um miRNA.

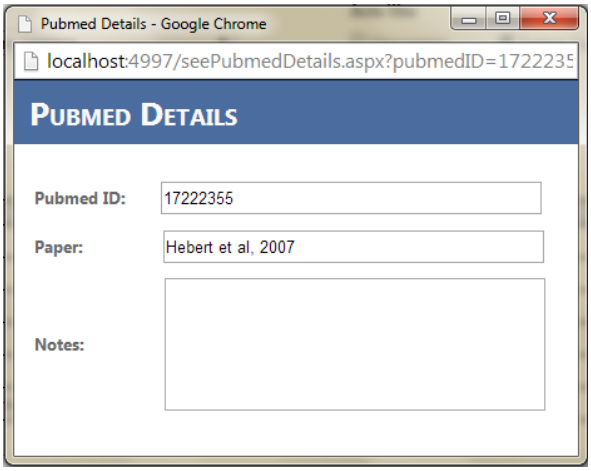

Figura 26. Janela com informações detalhadas sobre uma publicação.

Finalmente, na aba "*About*", mostrada na Figura 27, há uma descrição sobre o framework web e suas funcionalidades disponibilizadas ao usuário.

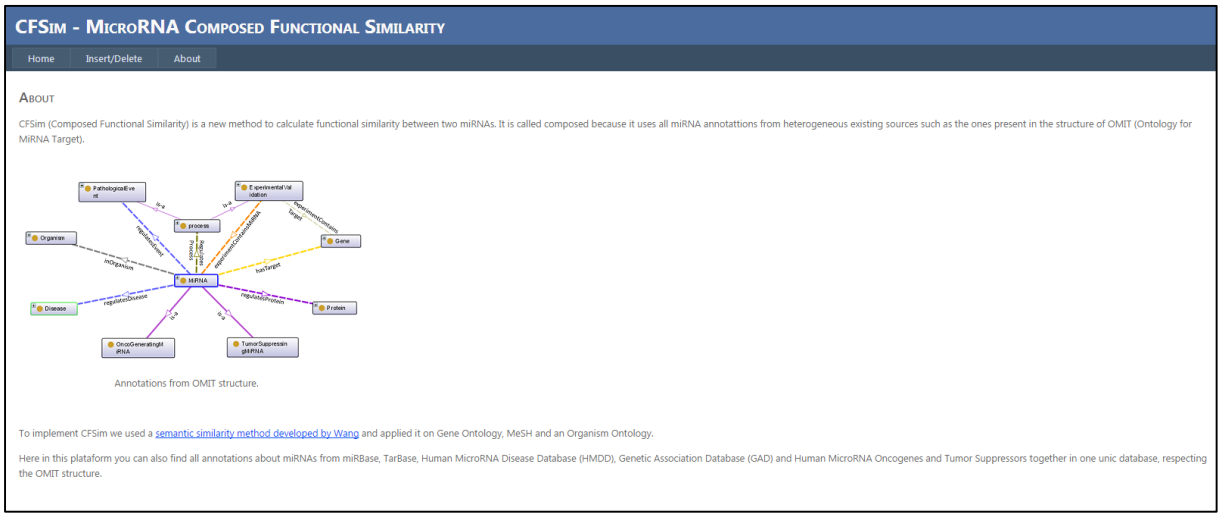

Figura 27. Janela com informações sobre o método *CFSim* e o framework apresentadas na aba "*About*".