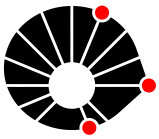
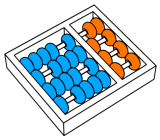




Redes Complexas Aplicadas à Biologia de Sistemas para Estudo do Câncer de Tireoide

Mylena Roberta dos Santos *André Santanchè*

Relatório Técnico - IC-PFG-23-57
Projeto Final de Graduação
2023 - Dezembro

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Redes Complexas Aplicadas à Biologia de Sistemas para Estudo do Câncer de Tireoide

Mylena Roberta dos Santos André Santanchè*

Resumo

A biologia de sistemas, jovem área interdisciplinar de importância ascendente, dedica-se a estudar sistemas biológicos complexos. A descoberta de novos elementos, processos e fenômenos pela biologia promove uma amplificação da complexidade que está envolvida nos análises realizadas por essa área, o que evidencia a necessidade de criação ou adoção de novas técnicas e ferramentas para tal. Recentemente, devido à intersecção entre os conceitos fundamentais da biologia de sistemas e de redes complexas, a comunidade científica passou a notar o potencial relacionado ao uso de abordagens baseadas nessas redes nos estudos desenvolvidos pela área em questão.

Definido o contexto, desenvolvemos um experimento baseado na tradução sob a ótica de redes complexas de um estudo sobre câncer de tireoide e associado à biologia de sistemas. Essa pesquisa, partindo de uma abordagem tabular de análise e do panorama pouco explorado da regulação gênica por microRNAs, resultou em uma rede de regulação pós-transcricional no contexto de câncer de tireoide. Nosso objetivo consistiu em elucidar a importância da abordagem de redes complexas em pesquisas envolvendo biologia de sistemas. Para tal, fizemos uso das ferramentas miRWalk e Neo4j, além de métricas topológicas de centralidade e detecção de comunidades.

1 Introdução

A biologia de sistemas, área interdisciplinar de funcionamento cíclico que é impulsionada por tecnologia e computação, ocupa-se do estudo dos sistemas biológicos complexos [1]. Majoritariamente, esses sistemas são de natureza discreta e podem ser representados como redes [2]. Em paralelo, a área de redes complexas estuda grafos com propriedades não-triviais, ou seja, atributos que não estão presentes em grafos reticulados ou aleatórios. Essas redes são capazes de representar sem perdas uma ampla gama de sistemas complexos, como redes sociais e a internet.

A primeira área analisa desde fenômenos biológicos extensivamente estudados, como transcrição e tradução, até processos recém-identificados, como a regulação de expressão gênica por microRNAs (miRNAs). A descoberta de novos elementos, como o próprio miRNA, resulta em uma potencialização da complexidade envolvida nas investigações realizadas pela biologia de sistemas, o que reforça a necessidade de criar ou adotar novas técnicas e ferramentas para tal.

*Instituto de Computação, Universidade Estadual de Campinas, 13083-852, Campinas, SP.

A intersecção recentemente apontada entre as áreas citadas define o interesse central deste trabalho. Uma vez que elas compartilham conceitos-chave - emergência, robustez e modularidade -, abordagens para análise de sistemas biológicos baseadas em redes complexas passaram a ser consideradas promissoras pela comunidade científica [2]. Em específico, tais abordagens são guiadas pelas propriedades topológicas de redes complexas, que referem-se aos seus padrões estruturais e organizacionais.

Como as doenças, em grande sua maioria, alteram o funcionamento dos sistemas biológicos, torna-se evidente que a área inicial pode contribuir para as pesquisas focadas nas mais diversas enfermidades. Ademais, essa contribuição é capaz de abarcar até mesmo doenças classificadas como complexas e que, hoje, representam grandes desafios para o campo da saúde de modo abrangente, como doenças autoimunes, e.g., esclerose múltipla, lúpus e artrite, e as muitas variações do câncer [3].

A título de exemplo, usando como referência o câncer e a rede do tipo interação proteína-proteína, ou *protein-protein interaction* (PPI), existem estudos originários da biologia de sistemas que geraram evidências relevantes a partir de abordagens baseadas em redes complexas [4, 5]. Jonsson & Bates (2006) e Sun & Zhao (2010) analisaram características topológicas de redes PPI e concluíram que as proteínas vinculadas ao câncer têm topologia de rede diferente das outras. Essa descoberta é capaz, dentre muitos outros aspectos, de auxiliar no entendimento da etiologia do câncer em nível de sistema.

Assim, em razão do fato de ainda haver grande espaço para investigação das abordagens baseadas em redes complexas para a biologia de sistemas, fomos motivados a desenvolver este trabalho. Como base para a metodologia, adotamos a pesquisa de Geraldo & Kimura (2015), estudo que objetivou-se a construir uma rede de regulação pós-transcricional para o câncer de tireoide a partir do panorama pouco explorado da regulação gênica desempenhada por miRNAs. O propósito deste trabalho esteve voltado para a tradução dos passos da pesquisa mencionada, sendo que esse processo seguiu uma ótica computacional apoiada em redes complexas.

A respeito dos resultados obtidos, modelamos uma rede que relaciona os miRNAs vinculados ao câncer de tireoide de acordo com as interações miRNA-alvo e, após isso, aplicamos métricas topológicas (centralidade e detecção de comunidades) sobre ela. Com base nos resultados das medidas e no embasamento teórico, verificamos que essas moléculas estão fortemente associadas e levantamos algumas suposições que atrelam o que resultou da análise com a biologia envolvida.

O restante do texto está organizado da seguinte maneira: a Seção 2 apresenta os pontos-chave da nossa fundamentação teórica; a Seção 3 descreve resumidamente a metodologia de acordo com o estudo de Geraldo & Kimura; a Seção 4 expõe a discussão e resultados associados à tradução que desenvolvemos; e a Seção 5 exibe as conclusões que obtivemos.

2 Fundamentação Teórica

Nesta seção, destacamos conceitos e definições com base em estudos associados ao contexto descrito e que foram de grande relevância para o desenvolvimento deste trabalho. Os tópicos apresentados a seguir permeiam da biologia até a computação.

2.1 Biologia de Sistemas

A biologia de sistemas encarrega-se de estudar os sistemas biológicos complexos através da análise quantitativa e entendimento das interações funcionais ao longo do tempo de seus componentes [1]. Os estudos conduzidos por essa área requerem equipes interdisciplinares, uma vez que combinam conceitos provenientes de diversas áreas, indo da biologia até a computação. A Figura 1 ilustra o comportamento cíclico que caracteriza o funcionamento da biologia de sistemas, que, sob regência da biologia, provoca o desenvolvimento de novas tecnologias e ferramentas computacionais com capacidade de revolucionar a ciência biológica.

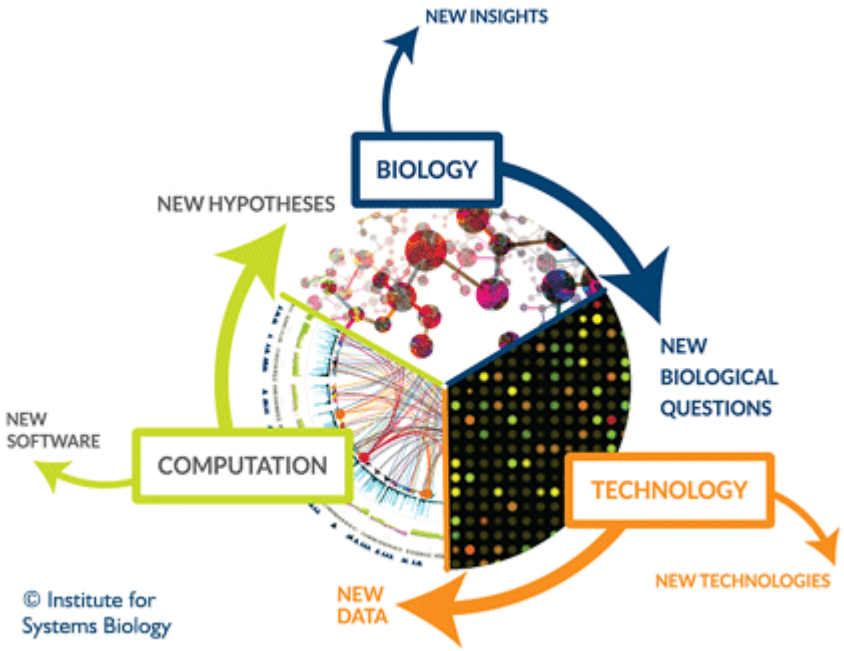


Figura 1: "Motor de inovação" que ilustra o funcionamento cíclico da biologia de sistemas, de acordo com o Institute for Systems Biology.

Existem três conceitos-chave para o entendimento de sistemas biológicos complexos: emergência (i), robustez (ii) e modularidade (iii) [1]. Vale mencionar que, devido ao caráter naturalmente discreto desses sistemas, eles podem ser representados por redes [2].

(i) Para que seja possível compreender as propriedades dos sistemas complexos, é necessário adotar uma perspectiva holística, em nível de sistema. Uma abordagem reducionista, focada apenas nas partes individuais e isoladas, não é capaz de promover entendimento completo dessas "propriedades emergentes". Exemplificando, ainda que sejam compreendidas por completo as propriedades de ambos hidrogênio e oxigênio, não há como prever as propriedades da água.

(ii) Os sistemas biológicos são inerentemente capazes de manter estabilidade frente às perturbações que podem ser impostas pelo ambiente, variações genéticas ou eventos estocásticos. Um exemplo evidente dessa característica é a capacidade de defesa do corpo humano representada pelo sistema imunológico.

(iii) Considerando sistemas complexos modelados por redes, para biólogos, os módulos são conjuntos de nós que têm fortes interações e uma função comum. Essa característica ainda contribui para a propriedade de robustez, uma vez que limita o dano a partes isoladas e, assim, diminui o risco de falha do sistema. Por exemplo, a segmentação do cérebro humano em diversas áreas funcionais ilustra a modularidade presente nesse órgão do sistema nervoso central.

É fundamental destacarmos que esses conceitos também têm sido amplamente explorados nas pesquisas relacionadas às redes complexas [2], portanto, há uma considerável intersecção entre as propriedades dos sistemas biológicos complexos e dessas redes. Assim sendo, torna-se evidente a motivação orgânica para a emergência da utilização de abordagens orientadas a redes complexas nos estudos promovidos pela biologia de sistemas.

Este domínio envolve o estudo de fenômenos biológicos como:

- *transcrição*, cópia da informação contida em um gene do DNA para a produção de um RNA mensageiro (mRNA);
- *pós-transcrição*, modificações e processamentos capazes de transformar o mRNA transcrito em uma molécula madura e funcional, pronta para a tradução; e
- *tradução*, uso da informação do mRNA para produzir uma cadeia específica de aminoácidos, ou seja, uma proteína.

O foco deste trabalho está voltado para a pós-transcrição, especialmente para a regulação de expressão gênica desempenhada por microRNAs (miRNAs), pequenos RNAs que não codificam proteínas. A função regulatória dos miRNAs é cumprida a partir da conexão por reconhecimento de mRNA-alvo, fundamentada na complementaridade de sequências de ambas moléculas [6]. Também é válido salientarmos que, no contexto em questão, os termos "gene" e "mRNA" surgem como sinônimos.

Em sua maioria, os miRNAs têm efeito inibitório na expressão gênica, porém há interações em que a molécula conecta-se à região promotora do gene, facilitando a sua tradução proteica [6]. Ainda destacamos que cada miRNA pode interagir com um grande número de mRNAs, e cada mRNA pode ser alvo de diversos miRNAs, resultando em uma rede complexa de interações miRNA-mRNA [6]. Para finalizar, ressaltamos também o envolvimento dos miRNAs com doenças, por exemplo, como o câncer, tema que tem emergido como foco de pesquisas desde o início dos anos 2000 [6].

2.2 Redes Complexas

As redes complexas, extensão da teoria dos grafos de importância ascendente, consistem em grafos com propriedades topológicas não-triviais, i.e., características que não são apresentadas por grafos reticulados (*lattice*) ou grafos aleatórios. O crescente avanço e popularização da pesquisa relacionada a essas redes decorre do potencial intrínseco que elas possuem para representar virtualmente qualquer sistema composto por elementos discretos, como sistemas biológicos [2].

Algumas das mais principais características topológicas dessas redes são estas:

- *small-word* (ou mundo pequeno), em sua maioria, os nós estão relativamente próximos entre si, de modo que a distância entre quaisquer dois nós é curta até mesmo em redes consideradas grandes;
- *scale-free* (ou livre de escala), a distribuição de grau dos nós segue, de maneira aproximada, uma lei de potência, isso significa que um pequeno número de nós concentra muitas conexões, enquanto a maioria dos nós tem poucas ligações; e
- *clusterização*, os nós da rede tendem a formar *clusters*, ou agrupamentos, onde encontram-se altamente interconectados.

Além disso, a fim de analisar e compreender a estrutura das redes complexas, foram desenvolvidas métricas topológicas. A título de exemplo, estes são alguns dos principais grupos de medidas:

- *centralidade*, quantificam a importância de cada um dos nós da rede de acordo com diversos critérios, como número de conexões, proximidade com outros nós ou importância dos nós aos quais um nó está conectado;
- *caminho e distância*, descrevem a proximidade (perto ou distante) e a facilidade de comunicação entre dos nós da rede; e
- *coeficiente de clusterização*, avaliam a presença de *clusters* na rede, podendo ser estabelecidas de modo local ou global.

A respeito das redes biológicas, i.e., redes complexas que representam sistemas biológicos, estas são as mais adotadas em relação ao controle dos sistemas celulares: regulação transcricional, interação proteína-proteína (PPI) e metabólicas [2]. Essas redes compartilham diversos atributos globais, e.g., distribuição de conectividade *scale-free*; propriedade *small-world*; natureza dissassortativa; organização modular; e robustez estrutural e dinâmica [2]. Em relação à modelagem, por exemplo, a rede de regulação transcricional é representada por um grafo em que os dois tipos de nó, fator de transcrição e gene-alvo, são conectados por uma interação direcionada [2].

3 Metodologia

Este trabalho consistiu na tradução dos métodos empregados para a construção da rede biológica resultante da pesquisa documentada pelo artigo "*Integrated Analysis of Thyroid Cancer Public Datasets Reveals Role of Post-Transcriptional Regulation on Tumor Progression by Targeting of Immune System Mediators*", desenvolvido por Murilo Geraldo e Edna Kimura. Como fundamentação, os autores escolheram dois subtipos de tumores de tireoide, *papillary thyroid carcinoma* (PTC) e *anaplastic thyroid carcinoma* (ATC). Eles exploraram a regulação mediada por miRNAs e se basearam na vasta quantidade de dados de expressão gênica disponíveis publicamente. A fim de construir e analisar uma rede de regulação pós-transcricional para o câncer de tireoide, eles alinharam os dados de expressão gênica de PTC e ATC com a previsão de alvos de miRNAs.

A análise da rede biológica em questão revelou que as interações miRNA-mRNA podem contribuir para a desregulação de mediadores tumorais-chave, o que pode levar a um comportamento mais agressivo e à progressão do tumor de tireoide. Portanto, utilizando dados de expressão gênica disponíveis em repositórios públicos e de algumas ferramentas computacionais, os pesquisadores elucidaram o panorama da regulação pós-transcricional exercida pelos miRNAs no câncer de tireoide. A estratégia adotada para a construção da rede de regulação pós-transcricional foi ilustrada pelos autores através do diagrama apresentado na Figura 2.

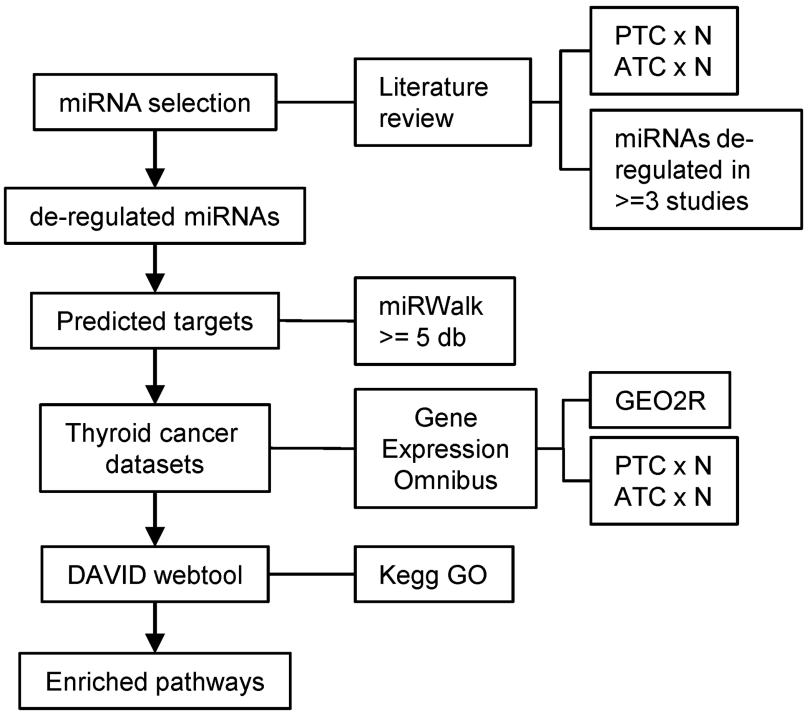


Figura 2: Diagrama que representa a estratégia adotada por Geraldo & Kimura (2015) para a construção da rede de regulação pós-transcricional no câncer de tireoide.

Na sequência, expomos brevemente descrições a respeito das etapas que fazem parte do desenvolvimento da rede biológica citada. Convém frisar que os pesquisadores adotaram uma abordagem tabular para investigar e lidar com os passos envolvidos no estudo. Na seção seguinte, será abordada a tradução que desenvolvemos para as etapas sob uma ótica vinculada ao formalismo de grafos e redes complexas.

3.1 Seleção de MicroRNA

Esta etapa compreende os estágios denominados *miRNA selection* e *de-regulated miRNAs* na estratégia mencionada. A princípio, os autores realizaram uma revisão crítica da literatura orientada pela busca dos termos "*miRNA*" e "*thyroid cancer*" de maneira a identificar os miRNA mais frequentemente desregulados para os subtipos de câncer de tireoide investigados. A análise resultou na seleção de 15 estudos em língua inglesa, 11 para PTC e 4 para ATC, que seguiam estes critérios:

(i) perfila a expressão gênica de pelo menos cinco miRNAs;

(ii) tumores derivados de células foliculares da tireoide humana;

(iii) comparação entre tumores e tecido tireoidiano não tumoral; e

(iv) amostras frescas, congeladas ou FFPE.

Após a filtragem seguindo as regras listadas, a seleção dos miRNAs desregulados a serem empregados na construção da rede ocorreu considerando um padrão de expressão concordante entre três ou mais estudos. Ainda houve a preocupação de corrigir a nomenclatura dos miRNAs de acordo com a versão 19.0 do miRBase (`https://mirbase.org`), o arquivo para sequências e anotações de miRNA.

Assim, como desfecho e com base nos 15 estudos aludidos, os pesquisadores selecionaram estes 15 miRNAs: *miR-221-3p*, *miR-146b-5p*, *miR-222-3p*, *miR-181b-5p*, *miR-155-5p*, *miR-34a-5p*, *miR-26a-5p*, *miR-224-5p*, *miR-138-5p*, *miR-187-3p*, *miR-31-5p*, *miR-125b-5p*, *let-7c*, *miR-30a-5p* e *miR-30d*.

3.2 Predição Computacional dos Alvos dos MicroRNAs

Para a predição dos alvos de cada miRNA diferencialmente expresso selecionado, i.e., a fim de compreender o papel de regulação pós-transcricional exercido por essas moléculas desreguladas no câncer de tireoide, os pesquisadores utilizaram o banco de dados miRWalk (`http://mirwalk.umm.uni-heidelberg.de/`).

No momento de desenvolvimento do estudo, esse programa previa interações miRNA-mRNA utilizando oito algoritmos, permitindo a seleção de genes-alvo previstos simultaneamente por dois ou mais algoritmos. Visando aumentar a confiabilidade dos resultados, os autores optaram por selecionar os mRNAs previstos por pelo menos cinco desses algoritmos.

3.3 Conjuntos de Dados de Expressão Gênica

Nesta etapa, os pesquisadores tiraram proveito do repositório Gene Expression Omnibus (GEO) (`https://www.ncbi.nlm.nih.gov/geo/`) para pesquisar por conjuntos de dados de expressão gênica de tumores tireoidianos derivados de células foliculares, focando nos estudos que faziam comparações entre PTC ou ATC e tecido normal. Eles ainda utilizaram o programa GEO2R (`https://www.ncbi.nlm.nih.gov/geo/geo2r/`) para calcular a expressão gênica diferencial entre tecido tumoral e não-tumoral, sendo que o resultado foi considerado válido para os genes com padrão de expressão concordante e p-valor ajustado $<0{,}05$. Em conclusão, foram eleitos cinco estudos que compreendiam, ao todo, dados de 203 amostras, incluindo PTC, ATC e tecido tireoidiano normal.

3.4 Construção de Redes Regulatórias

Para esta etapa, inicialmente, os autores buscaram nos conjuntos de dados por interações miRNA-alvo com padrões de expressão anti-correlacionadas, i.e., casos com miRNA aumentado e mRNA diminuído e vice-versa, e compararam os resultados com os dados de expressão de amostras tumorais. A lista de genes-alvo com o perfil desejado foi submetida para análise na ferramenta *web* Database for Annotation, Visualization and Integrated Discovery (DAVID) (`https://david.ncifcrf.gov/`). De maneira que a análise em questão era a de enriquecimento de conjunto de genes, ou *gene set enrichment analysis* (GSEA), realizada com o intuito de identificar assinaturas gênicas enriquecidas.

4 Discussão e Resultados

A seguir, apresentaremos em detalhes o processo de tradução do artigo de Geraldo & Kimura (2015). É relevante ressaltar que buscamos reproduzir as etapas da pesquisa explorando ao máximo o formalismo associado a grafos e redes complexas. Frisamos ainda que não foi possível traduzir todas as etapas descritas na seção anterior; trabalhamos nos passos *Seleção de MicroRNA* e *Predição Computacional dos Alvos de MicroRNA*, e criamos a etapa *Construção e Análise de Redes de Interação* para explorar os resultados alcançados pela tradução parcial desenvolvida.

4.1 Seleção de MicroRNA

Visto que a tradução que propomos envolve apenas processos computacionais, uma revisão da literatura em biologia foge do nosso escopo. Assim, tomamos como base os 15 miRNAs previamente citados para prosseguir com o desenvolvimento do trabalho. Dentre essas moléculas, destacam-se duas, *let-7c* e *miR-30d*, por não possuírem sufixo, *-3p* ou *-5p*. A identificação do sufixo de um miRNA é essencial, pois representa a região de conexão com o mRNA-alvo, determinando a consequência biológica da interação [6]. Ademais, a próxima etapa, que envolve a busca por alvos preditos dos miRNAs no miRWalk, requer que as moléculas sejam acompanhadas do sufixo.

Em decorrência disso, foi preciso encontrar um modo de definir quais seriam os sufixos adequados para os dois miRNAs. Adotando como fundamentação o fato de que essas moléculas conectam-se aos seus alvos principalmente na região não traduzida 3' (3' UTR), ou seja, a partir de 5' UTR [6], torna-se viável supor que estamos lidando com *let-7c-5p* e *miR-30d-5p*. A fim de reforçar a inferência, buscamos no banco de dados miRBase pelas moléculas sob discussão e analisamos os histogramas relacionados ao número de leituras experimentais por sequência de cada uma delas.

As Figuras 3 e 4 ilustram *screenshots* com o resultado da busca no miRBase pelo identificador, respectivamente, *hsa-let-7c* e *hsa-mir-30d*, de maneira que o prefixo *hsa-* refere-se à espécie *Homo sapiens*. Para ambos os casos, o histograma tem frequências absolutamente maiores para as sequências dispostas à esquerda, que representam a região 5' UTR das moléculas, destacando que essa região é mais examinada e debatida do que a outra. Por fim, levando em consideração a fundamentação teórica mencionada e os resultados oferecidos pelo banco de dados consultado, admitimos que estamos tratando especificamente dos miRNAs nas versões *let-7c-5p* e *miR-30d-5p*.

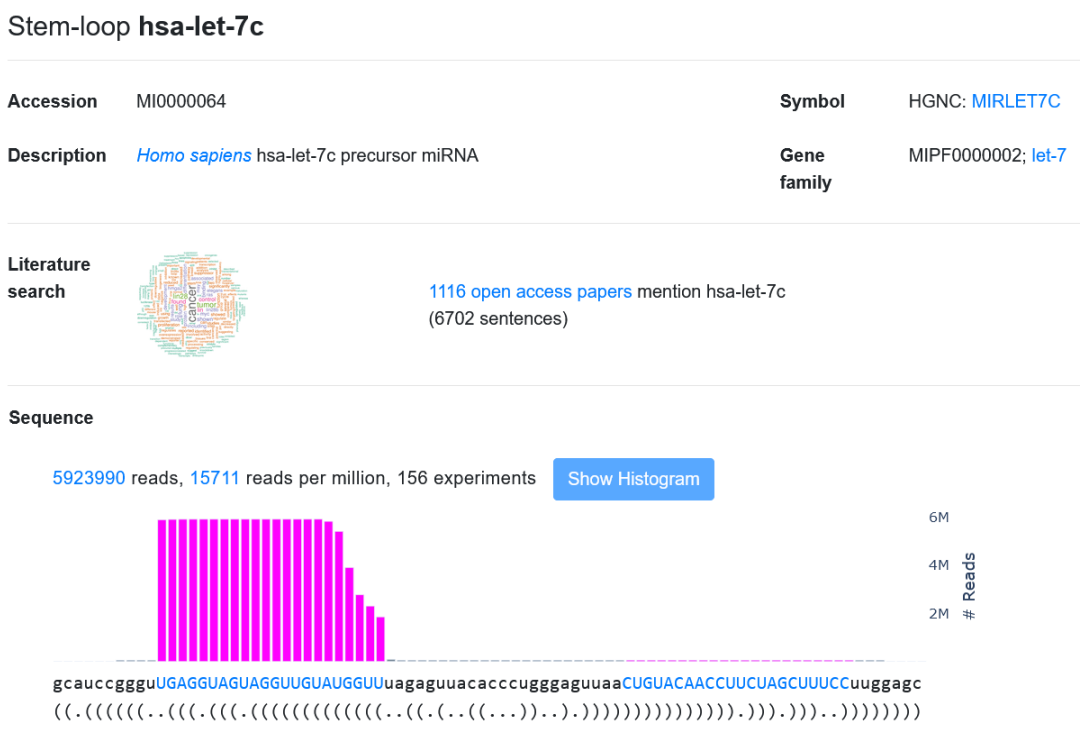


Figura 3: Resultado da busca por *hsa-let-7c* no miRBase, as sequências destacadas à direita e esquerda do histograma representam, respectivamente, 3' UTR e 5' UTR.

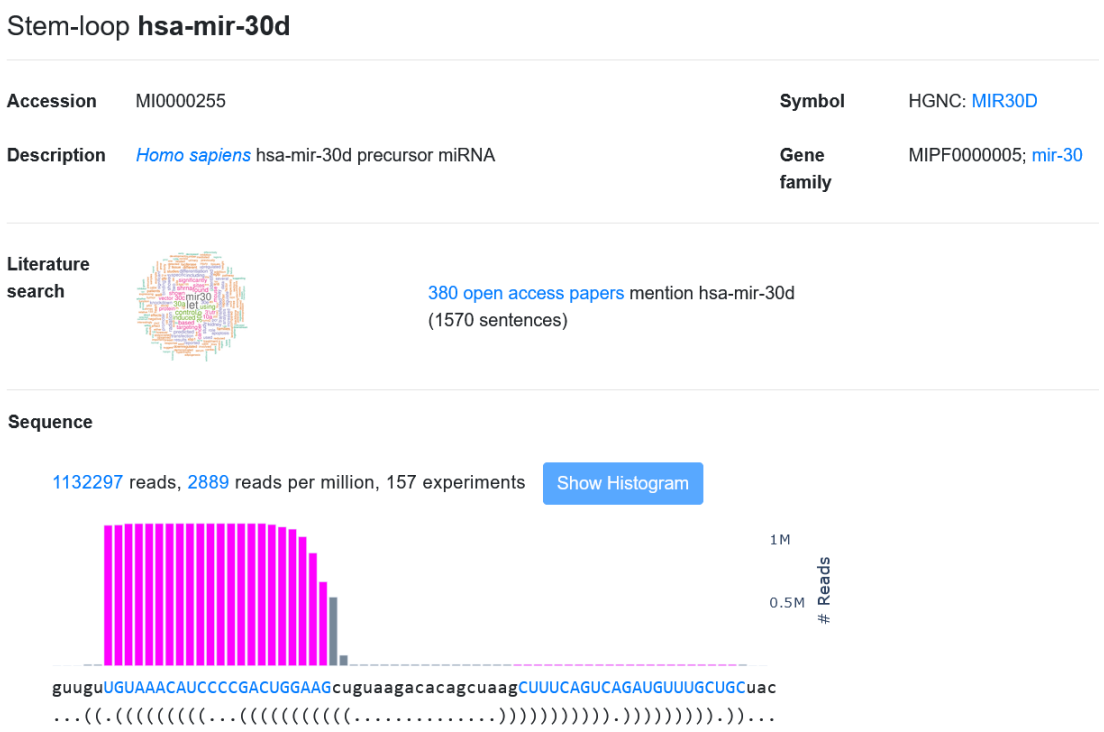


Figura 4: Resultado da busca por *hsa-mir-30d* no miRBase, as sequências destacadas à direita e esquerda do histograma representam, respectivamente, 3' UTR e 5' UTR.

4.2 Predição Computacional dos Alvos dos MicroRNAs

O miRWalk armazena informações produzidas por um algoritmo de aprendizado de máquina baseado em *random forest* (TarPmiR), além de dados gerados por terceiros (TargetScan, miRDB e miRTarBase). As duas primeiras fontes externas, TargetScan (`https://www.targetscan.org/vert_80/`) e miRDB (`https://mirdb.org/`), são bases de dados que hospedam interações miRNA-alvo também resultantes de algoritmos preditivos. A terceira fonte, miRTarBase (`https://awi.cuhk.edu.cn/~miRTarBase`), por outro lado, diz respeito a um banco de dados de interações miRNA-mRNA validadas experimentalmente e documentadas na literatura.

Atualmente, o miRWalk está na versão 3 e consiste em uma aplicação *web* que permite buscar por interações miRNA-alvo a partir da espécie e miRNA, ou gene-alvo (mRNA). A molécula, miRNA ou mRNA, a ser usada como parâmetro para a busca deve ser identificada por alguma das convenções de nomenclatura suportadas pelo miRWalk. Para miRNAs, por exemplo, são aceitas entradas no formato de nome ou número de acesso de acordo com a versão atual do miRBase. Ademais, existem opções avançadas de busca baseadas em listas de miRNAs ou genes, vias biológicas (*pathways*) ou doenças. A Figura 5 mostra o *screenshot* da *homepage* do miRWalk, nela estão dispostas todas as alternativas de busca mencionadas.

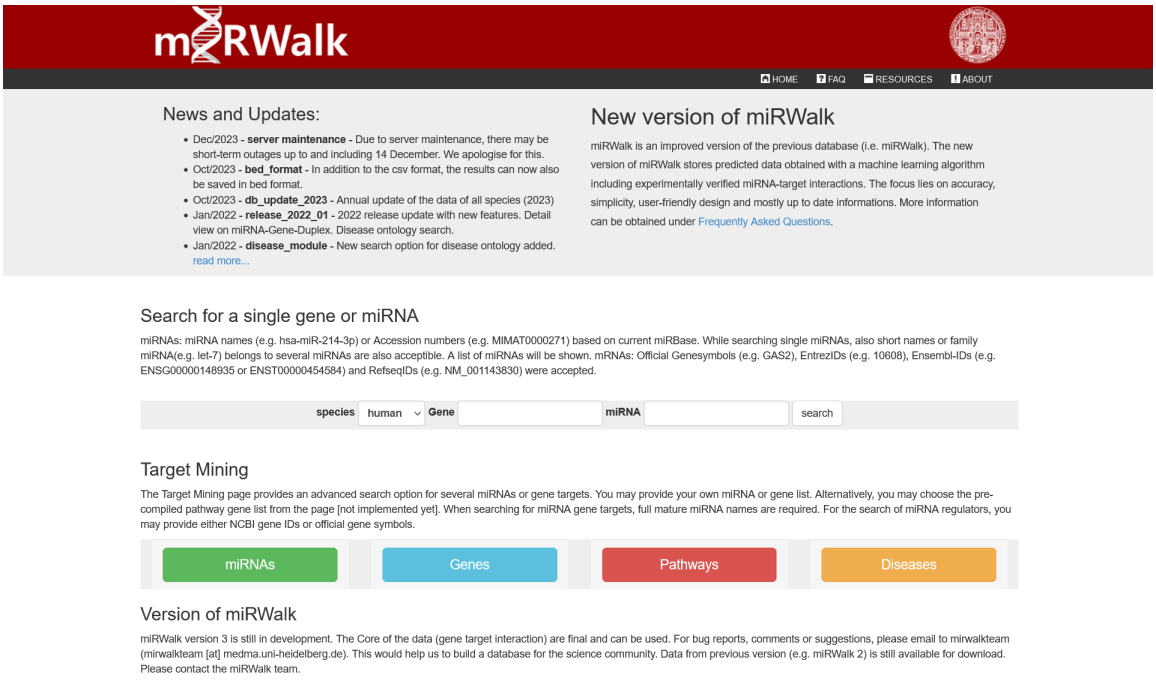
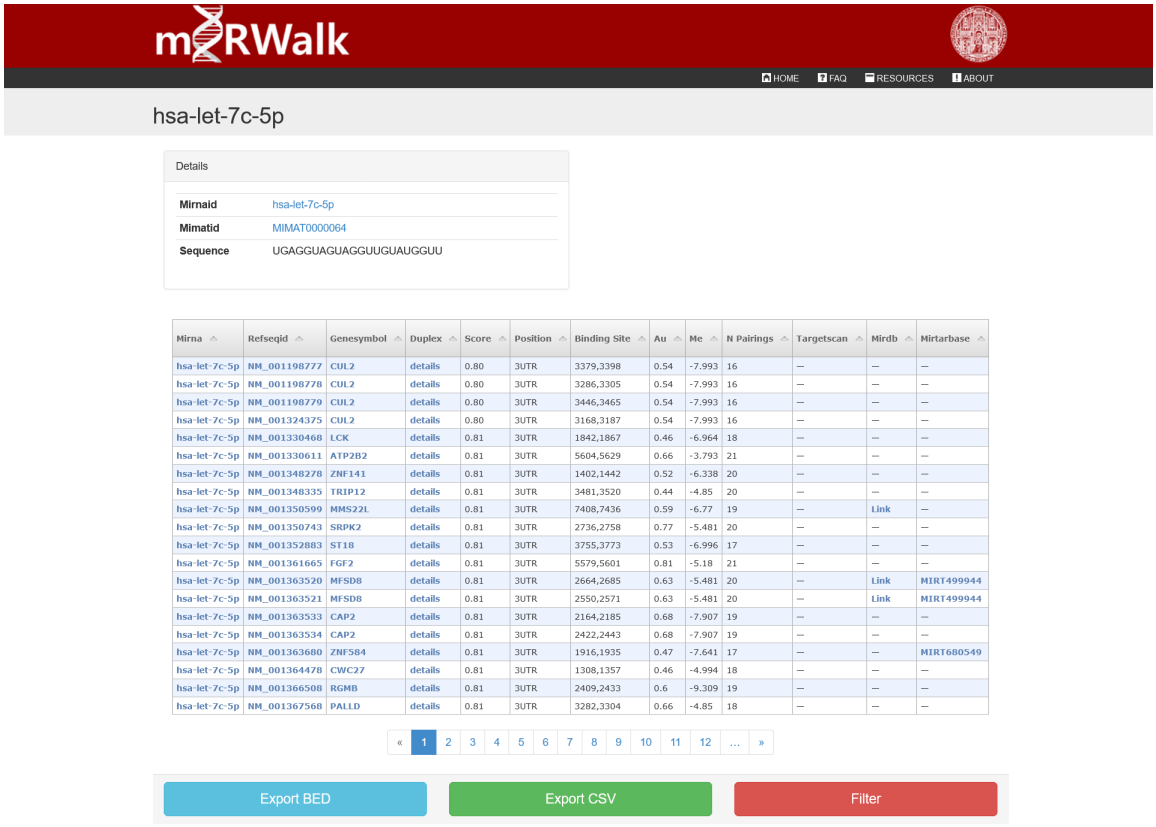


Figura 5: Página inicial da versão atual do miRWalk. As opções de busca simples e avançadas encontram-se, respectivamente, nas seções *Search for a single gene or miRNA* e *Target Mining*.

Portanto, partindo das informações descritas no artigo, caracteriza-se que os parâmetros para a busca no miRWalk são a espécie humana e a lista de 15 miRNAs citada anteriormente, lembrando que admitimos que *let-7c* e *miR-30d* referem-se, na verdade, aos miRNAs *let-7c-5p* e *miR-30d-5p*, respectivamente. Embora haja no banco de dados uma opção de busca avançada baseada em uma lista de miRNAs, ela é incapaz de buscar pelos alvos preditos de todas as 15 moléculas em uma única *query*. Em decorrência disso, optamos por desenvolver um fluxo automatizado de extração de dados para obter as interações miRNA-mRNA de interesse.

Dado que o banco de dados não possui uma função de API até o momento atual, foi preciso desenvolvermos um processo computacional baseado em *web scraping* para viabilizar a mineração de dados. É importante ressaltar que, em relação ao *download* manual, a extração de dados automatizada tende a ser mais eficiente, veloz, escalável e menos propensa a erros. Em relação aos recursos empregados para tal, fizemos uso da linguagem de programação Python (`https://www.python.org/`) e dos pacotes Selenium (`https://pypi.org/project/selenium/`) e Pandas (`https://pypi.org/project/pandas/`). Esses pacotes consistem em, de modo respectivo, um conjunto de ferramentas para automatizar navegadores *web* e um *kit* de ferramentas para análise e manipulação de dados. O *notebook* que implementa esse processo computacional está disponível neste repositório? `https://github.com/MylenaRoberta/PFG`.

De forma concisa, o processo computacional mencionado consiste, inicialmente, em fazer o *download* da predição de alvos para cada um dos miRNAs listados e, depois, concatenar todos os dados em uma única tabela. Para baixar os dados de interação preditos pelo miRWalk, tomamos como referência a busca simples a partir de espécie e nome do miRNA, operação que pode ter os resultados exportados em um único arquivo CSV. A título de exemplo, na Figura 6, apresentamos o resultado imediato da operação de busca pelos genes-alvo da molécula *let-7c-5p* para a espécie humana. Ao final, após concatenarmos o conteúdo de todos esses arquivos em uma única tabela, geramos um único arquivo CSV como a base de dados consolidada resultante desta etapa.



Mirna	Refseqid	Genesymbol	Duplex	Score	Position	Binding Site	Au	Me	N Pairings	Targetscan	Mirdb	Mirtarbase
hsa-let-7c-5p	NM_001198777	CUL2	details	0.80	3UTR	3379,3398	0.54	-7.993	16	–	–	–
hsa-let-7c-5p	NM_001198778	CUL2	details	0.80	3UTR	3286,3305	0.54	-7.993	16	–	–	–
hsa-let-7c-5p	NM_001198779	CUL2	details	0.80	3UTR	3446,3465	0.54	-7.993	16	–	–	–
hsa-let-7c-5p	NM_001324375	CUL2	details	0.80	3UTR	3168,3187	0.54	-7.993	16	–	–	–
hsa-let-7c-5p	NM_001330468	LCK	details	0.81	3UTR	1842,1867	0.46	-6.964	18	–	–	–
hsa-let-7c-5p	NM_001330611	ATP2B2	details	0.81	3UTR	5604,5629	0.66	-3.793	21	–	–	–
hsa-let-7c-5p	NM_001348278	ZNF141	details	0.81	3UTR	1402,1442	0.52	-6.338	20	–	–	–
hsa-let-7c-5p	NM_001348335	TRIP12	details	0.81	3UTR	3481,3520	0.44	-4.85	20	–	–	–
hsa-let-7c-5p	NM_001350599	MMS22L	details	0.81	3UTR	7408,7436	0.59	-6.77	19	–	Link	–
hsa-let-7c-5p	NM_001350743	SRPK2	details	0.81	3UTR	2736,2758	0.77	-5.481	20	–	–	–
hsa-let-7c-5p	NM_001352883	ST18	details	0.81	3UTR	3755,3773	0.53	-6.996	17	–	–	–
hsa-let-7c-5p	NM_001361665	FGF2	details	0.81	3UTR	5579,5601	0.81	-5.18	21	–	–	–
hsa-let-7c-5p	NM_001363520	MFSD8	details	0.81	3UTR	2664,2685	0.63	-5.481	20	–	Link	MIRT499944
hsa-let-7c-5p	NM_001363521	MFSD8	details	0.81	3UTR	2550,2571	0.63	-5.481	20	–	Link	MIRT499944
hsa-let-7c-5p	NM_001363533	CAP2	details	0.81	3UTR	2164,2185	0.68	-7.907	19	–	–	–
hsa-let-7c-5p	NM_001363534	CAP2	details	0.81	3UTR	2422,2443	0.68	-7.907	19	–	–	–
hsa-let-7c-5p	NM_001363680	ZNF584	details	0.81	3UTR	1916,1935	0.47	-7.641	17	–	–	MIRT680549
hsa-let-7c-5p	NM_001364478	CWC27	details	0.81	3UTR	1308,1357	0.46	-4.994	18	–	–	–
hsa-let-7c-5p	NM_001366508	RGMB	details	0.81	3UTR	2409,2433	0.6	-9.309	19	–	–	–
hsa-let-7c-5p	NM_001367568	PALLD	details	0.81	3UTR	3282,3304	0.66	-4.85	18	–	–	–

Figura 6: Resultado imediato para a busca baseada na espécie humana e no miRNA *let-7c-5p* no miRWalk, a opção de exportar os resultados em CSV (*Export CSV*) está na parte inferior.

Evidenciada a simplicidade associada ao funcionamento do fluxo de extração de dados, cabe ressaltar que a dificuldade do processo esteve vinculada em compreender como funciona a aplicação *web* que dá forma ao miRWalk. A técnica de *web scraping* exige o entendimento do código fonte do *site*, em especial dos elementos e eventos contidos nele, e, por consequência, a implementação de um algoritmo de acordo com as especificações da operação que deseja-se executar na página.

O Código 1 ilustra as principais instruções da função-chave para a mineração de dados no miRWalk, *export_mirna_targets*. A função, dado um miRNA, conduz a busca pelos genes-alvo preditos para a molécula e armazena o arquivo CSV com o que resulta da operação. A *export_mirna_targets* apresenta o fluxo requerido de interações com os elementos e eventos do código fonte da página e evidencia o grau de customização vinculado ao desenvolvimento desse fluxo automatizado de extração.

```python
def export_mirna_targets(driver, mirna):
  driver.get(URL)  # Access the provided URL

  # Find and fill in the species selector
  species_input = Select(driver.find_element(By.NAME, 'species'))
  species_input.select_by_visible_text(SPECIES)

  # Find and fill in the microRNA input
  mirna_input = driver.find_element(By.NAME, 'mirna')
  mirna_input.send_keys(mirna)

  # Find and click in the search button
  search_btn = driver.find_element(By.XPATH, '//button[text()="search"]')
  search_btn.click()

  # Find and click in the result export link
  export_link = driver.find_element(By.LINK_TEXT, 'Export CSV')
  export_link.click()

  ...
```

Código 1: Definição parcial da função *export_mirna_targets*, conjunto de instruções capaz de realizar uma busca simples por genes-alvo de um determinado miRNA no miRWalk e fazer o download dos resultados no formato de um arquivo CSV.

4.3 Construção e Análise de Redes de Interação

Com os dados de interações miRNA-alvo consolidados e acessíveis, identificamos a oportunidade de explorar a modelagem e análise de redes a partir dessa base. Para tal, empregamos o Neo4j (`https://neo4j.com/`), um banco de dados nativo de grafos que utiliza uma linguagem de consulta declarativa similar ao SQL, o Cypher (`https://neo4j.com/developer/cypher/`). As *queries* e perspectivas empregadas na geração dos resultados expostos a seguir estão documentadas neste repositório: `https://github.com/MylenaRoberta/PFG`.

Ao considerarmos o conjunto completo de dados provenientes do miRWalk para construir as redes, nos deparamos com diversos gargalos que dificultavam ou impediam por completo as modelagens e análises que projetamos. Assim, tomando as limitações de processamento impostas pela versão do Neo4j em uso, identificamos a necessidade de selecionar uma amostra dos dados de interações miRNA-mRNA para servir como fundamentação para a construção das redes.

Retomando o funcionamento do miRWalk, é importante relembrar que, além de armazenar os dados gerados pelo TarPmiR, ele hospeda informações do TargetScan, miRDB e miRTarBase. Devido às diferentes técnicas, computacionais ou experimentais, que originaram esses dados, surgem interações que foram apontadas por todas, algumas ou somente uma dessas fontes. A Tabela 1 apresenta a volumetria vinculada aos dados extraídos do miRWalk de acordo com a fonte associada. Assim, com base somente nos volumes de dados e com intuito de garantir o desempenho mais eficiente do Neo4j, optamos por empregar o subconjunto de interações originárias do miRTarBase como a base de dados a ser utilizada para a construção das redes.

Inicialmente, partindo da base de dados mencionada, modelamos uma rede bipartida com dois tipos de nós, *MicroRNA* e *MessengerRNA*, e um relacionamento, *INTERACTS_WITH*, de maneira que os nós *MicroRNA* e *MessengerRNA* representam, em ordem respectiva, a origem e destino da aresta direcionada que define *INTERACTS_WITH*. É válido destacar que essas denominações acompanham as convenções da comunidade do Neo4j. A Figura 7 ilustra a visualização absoluta do grafo em discussão, destacando que os nós *MicroRNA* revelam-se como *hubs*, uma vez que estão altamente conectados a nós *MessengerRNA*, e evidenciando a complexidade associada a essa rede bipartida.

A fim de viabilizar a realização de análises, adotamos um método para gerar uma nova perspectiva a fim de tratar a rede de forma homogênea. Tal método consistiu na projeção dos nós *MicroRNA*, i.e., a construção de uma nova rede em que os nós *MicroRNA* estão conectados por arestas bidirecionais que representam o relacionamento *IS_RELATED_TO*. A conexão entre quaisquer dois nós *MicroRNA* é constituída somente se eles compartilham ligações *INTERACTS_WITH* com pelo menos um nó *MessengerRNA*. Ainda planejávamos realizar a projeção dos nós *MessengerRNA*, mas não foi possível por limitações de processamento impostas pela versão em uso do banco de dados Neo4j.

Volumetria	**miRWalk**	**TargetScan**	**miRDB**	**miRTarBase**
# miRNA	15	15	15	15
# mRNA (genes)	19.313	4.600	5.238	2.966
# interações	575.985	32.645	33.915	15.234

Tabela 1: Volumetria dos dados de interação miRNA-mRNA obtidos a partir da extração dos resultados das buscas no banco de dados miRWalk.

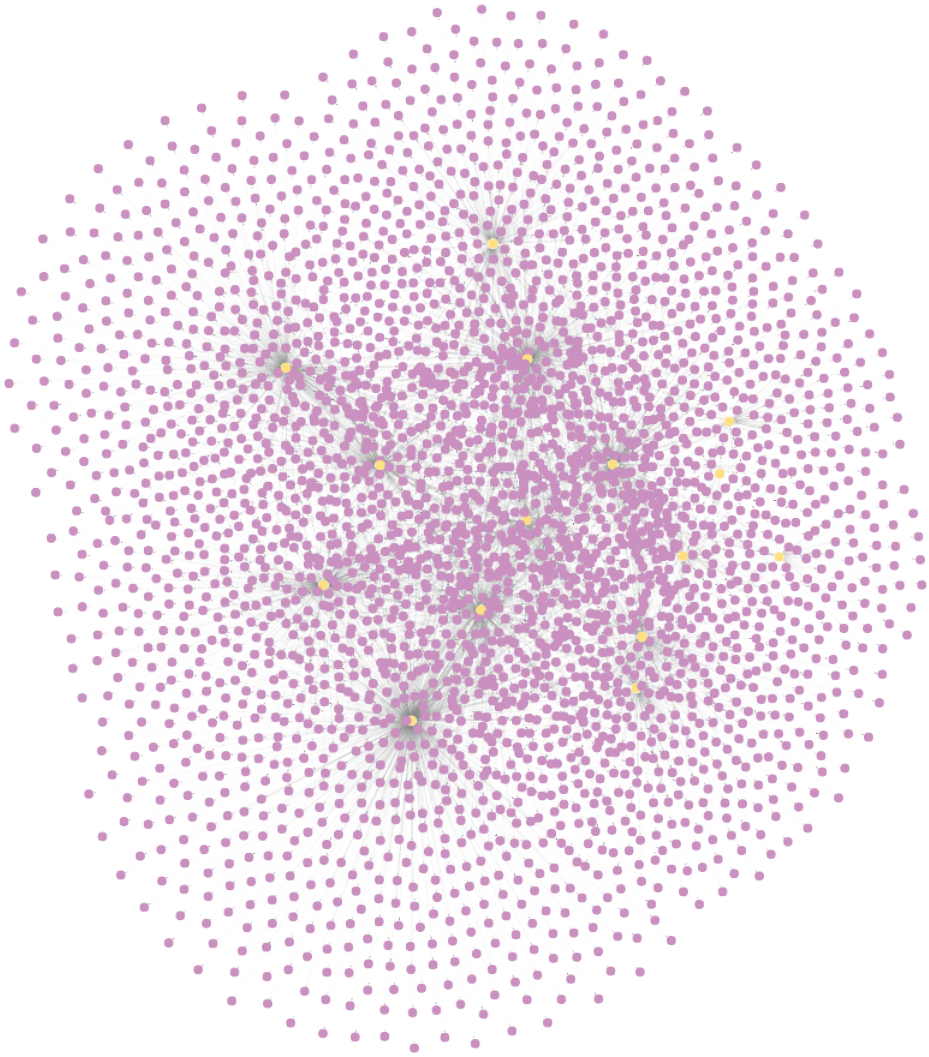


Figura 7: Visualização completa da rede de interações entre miRNAs e mRNAs, sendo as moléculas representadas, respectivamente, pelos nós de cor amarela e rosa.

Para analisar a rede resultante da projeção dos nós *MicroRNA*, empregamos as métricas topológicas de centralidade e detecção de comunidades instantaneamente acessíveis no Neo4j. Ao todo, consideramos três medidas de centralidade - grau, intermediação (*betweenness*) e autovetor (*eigenvector*) - e outras três de detecção de comunidades - propagação de rótulos (*label propagation*), método de Louvain e componentes fracamente conectados (*weakly connected components*). Ainda poderíamos ter avaliado a centralidade de Page Rank, porém optamos por descartá-la em razão de sua semelhança com a centralidade de autovetor.

Iniciando pelas métricas de centralidade, apresentamos, na Tabela 2, os valores de cada medida mencionada para os 15 nós da rede. Com estes dados, torna-se evidente que, em sua maioria, estamos tratando de nós altamente conectados, uma vez que, em 80% dos casos, eles conectam-se com pelo menos 13 nós. É curioso observar também que, para nós com o mesmo grau, podemos identificar variações nas centralidades de intermediação e autovetor. Focando nos nós de grau 13, é evidente que dois deles têm *betweenness* mais de cinco vezes maior e *eigenvector* minimamente menor (cerca de 3%). Isto é, *hsa-miR-26a-5p* e *hsa-miR-224-5p* são substancialmente mais importantes para a comunicação entre os nós da rede do que os outros de grau 13.

miRNA	Centralidade de Grau (*Degree*)	Centralidade de Intermediação (*Betweenness*)	Centralidade de Autovetor (*Eigenvector*)
hsa-miR-187-3p	5	0	0,106
hsa-miR-146b-5p	10	0	0,213
hsa-miR-138-5p	12	0	0,254
hsa-miR-221-3p	13	0,3	0,270
hsa-miR-222-3p	13	0,3	0,270
hsa-miR-181b-5p	13	0,3	0,270
hsa-miR-155-5p	13	0,3	0,270
hsa-miR-34a-5p	13	0,3	0,270
hsa-miR-125b-5p	13	0,3	0,270
hsa-let-7c-5p	13	0,3	0,270
hsa-miR-26a-5p	13	1,6	0,262
hsa-miR-224-5p	13	1,6	0,262
hsa-miR-31-5p	14	2,233	0,278
hsa-miR-30a-5p	14	2,233	0,278
hsa-miR-30d-5p	14	2,233	0,278

Tabela 2: Valores das três métricas de centralidade (grau, intermediação e autovetor) para a rede resultante da projeção dos nós *MicroRNA*.

Prosseguindo com as métricas de detecção de comunidades, duas delas, *label propagation* e *weakly connected components*, identificaram somente uma comunidade na rede em questão. Todavia, Louvain foi capaz de identificar duas comunidades de nós na rede. A Figura 8 mostra uma visualização da rede de projeção personalizada de acordo com duas métricas, o tamanho dos nós está de acordo com *betweenness* e as cores definem as comunidades detectadas pelo método de Louvain. Convém também frisar que, devido ao Neo4j ser capaz de representar somente relacionamentos direcionados, fez-se necessário, como pode ser verificado visualmente, empregar duas arestas em direções opostas para expressar *IS_RELATED_TO* entre quaisquer dois nós.

Retomando os resultados do método de Louvain, denominaremos os grupos de nós como *Comunidade A* (cor rosa, com seis nós) e *Comunidade B* (cor amarela, com os nove nós que restam) para facilitar a referência. A *Comunidade A* contém tanto o nó menos importante (*hsa-miR-187-3p*) quanto os cinco mais importantes, de acordo com as centralidades de grau e intermediação. A *Comunidade B* engloba todos os que restam. É provável que o nó *hsa-miR-187-3p* tenha sido incluído na *Comunidade A* por ter conexões apenas com os outros integrantes do grupo, apesar de ser destoante já que ele tem baixo grau de centralidade e seu *betweenness* é nulo.

De maneira a conciliar os resultados da análise da rede resultante da projeção dos nós *MicroRNA* e os conceitos biológicos envolvidos no artigo, podemos presumir que esses miRNAs são altamente relacionados porque, muito provavelmente, participam simultaneamente de processos biológicos associados ao câncer de tireoide. Ademais, como esse grupo originou-se de uma revisão crítica da literatura orientada pela busca dos miRNAs mais frequentemente desregulados em PTC e ATC, é bastante plausível que tenhamos identificado uma forte relação entre as moléculas.

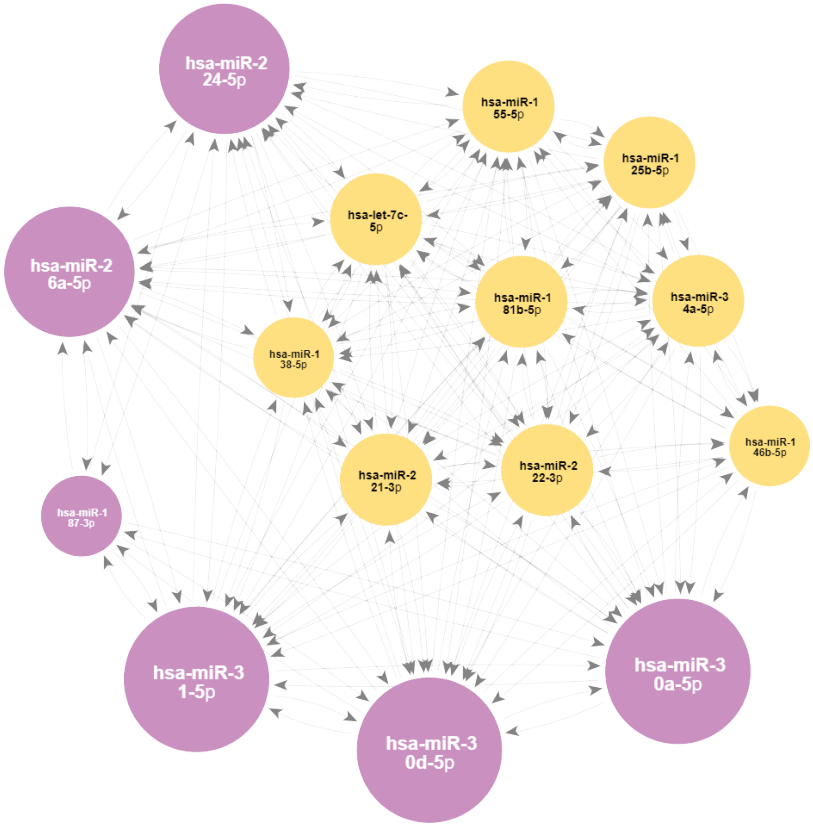


Figura 8: Visualização da rede resultante da projeção dos nós *MicroRNA* configurada a partir de métricas de centralidade e detecção de comunidades. O tamanho dos nós segue a centralidade de intermediação e as cores seguem as comunidades detectadas pelo método de Louvain.

Também não podemos descartar a hipótese de que cada comunidade detectada por Louvain possa estar mais fortemente vinculada a um dos subtipos estudados. Uma vez que ATC é a variação menos frequente [7], é natural que a comunidade científica promova mais estudos e discussões sobre PTC, logo são gerados menos dados sobre o primeiro subtipo. Disso, surge um viés a ser considerado nas análises, pois desbalanço na quantidade de dados a respeito de cada variação pode levar a conclusões que tendem a favorecer as características de PTC. Assim, não é absurdo supor que a comunidade com os nós mais importantes esteja mais associada a PTC do que ATC. Tal suposição deve ser algo de validação em trabalhos futuros.

Por fim, é notável que, para estabelecer conclusões robustas e acuradas a partir da interpretação dos resultados da análise de rede, precisamos consultar especialistas com conhecimentos de domínio vinculados à expressão gênica e câncer de tireoide. Não é possível expressarmos afirmações sobre assuntos tão específicos como vias biológicas e interações miRNA-mRNA, pois eles fogem do escopo compreendido pela computação e deste trabalho. Portanto, alcançamos um estágio em que a inerente interdisciplinaridade da biologia de sistemas manifesta-se e é requerida para desenvolvermos conclusões verossímeis.

5 Conclusão

No que diz respeito aos objetivos definidos para este trabalho, ficamos satisfeitos com os resultados alcançados, ainda que não tenha sido possível traduzir absolutamente todas as etapas incluídas no artigo. Retomando o contexto da biologia de sistemas, acreditamos que a análise de rede desenvolvida tenha demonstrado minimamente a relevância do uso de abordagens baseadas em redes complexas, tal como a inerente interdisciplinaridade envolvida em estudos compreendidos pela área.

Planejando trabalhos futuros, traduziremos as duas etapas faltantes, *Conjuntos de Dados de Expressão Gênica* e *Construção de Redes Regulatórias*, com enfoque mais direcionado para a última, por razões óbvias. Também seria interessante explorarmos outras modelagens de rede, como a projeção dos nós *MessengerRNA* anteriormente comentada, e mais métricas topológicas, tais como distância do caminho mais curto e coeficiente de clusterização.

Cabe ainda destacar que nos deparamos com dificuldades que decorreram, em especial, por estarmos tratando de uma área naturalmente multidisciplinar. O processo de aprofundamento da fundamentação teórica do artigo estudado foi desafiador, devido a alta carga de conceitos e definições oriundos da biologia. Embora não seja nosso papel dominá-los, é de suma importância que alcancemos ao menos uma compreensão básica do contexto holístico em que está inserido o estudo. Esse entendimento é necessário para habilitar a nossa participação ativa em projetos vinculados à biologia de sistemas. Ademais, como abordamos na seção anterior, foi complexo desenvolver suposições integradas com a biologia a respeito da análise de rede.

Enfrentamos obstáculos também com situações mais próximas do nosso domínio de conhecimento, como a extração automatizada de dados do miRWalk e as limitações de processamento do Neo4j. Em relação ao miRWalk, embora a mineração de dados seja efetiva, ela é extremamente sensível a alterações na aplicação, uma vez que, devido a inexistência de uma função de API, o processo computacional baseia-se no código fonte do site. Quanto ao Neo4j, a falta de familiaridade ao lidarmos com as configurações específicas desse banco de dados de grafos nos gerou diversos gargalos e restrições, impedindo a realização de análises com amostras de dados maiores e que exigiam processamentos mais robustos.

Referências

[1] A. Aderem, "Systems biology: its practice and challenges," *Cell*, vol. 121, no. 4, pp. 511–513, 2005.

[2] L. d. F. Costa, F. A. Rodrigues, and A. S. Cristino, "Complex networks: the key to systems biology," *Genetics and Molecular Biology*, vol. 31, pp. 591–601, 2008.

[3] N. Safari-Alighiarloo, M. Taghizadeh, M. Rezaei-Tavirani, B. Goliaei, and A. A. Peyvandi, "Protein-protein interaction networks (ppi) and complex diseases," *Gastroenterology and Hepatology from bed to bench*, vol. 7, no. 1, p. 17, 2014.

[4] P. F. Jonsson and P. A. Bates, "Global topological features of cancer proteins in the human interactome," *Bioinformatics*, vol. 22, no. 18, pp. 2291–2297, 2006.

[5] J. Sun and Z. Zhao, "A comparative study of cancer proteins in the human protein-protein interaction network," *BMC genomics*, vol. 11, pp. 1–10, 2010.

[6] A. L. Jorge, E. R. Pereira, C. S. d. Oliveira, E. d. S. Ferreira, E. T. N. Menon, S. N. Diniz, and J. A. Pezuk, "Micrornas: entendendo seu papel como reguladores da expressão gênica e seu envolvimento no câncer," *einstein (São Paulo)*, vol. 19, 2021.

[7] M. V. Geraldo and E. T. Kimura, "Integrated analysis of thyroid cancer public datasets reveals role of post-transcriptional regulation on tumor progression by targeting of immune system mediators," *PLoS One*, vol. 10, no. 11, p. e0141726, 2015.