

**UNIVERSIDADE DE CAXIAS DO SUL
ÁREA DO CONHECIMENTO DE CIÊNCIAS EXATAS E
ENGENHARIAS**

ANGÉLICA LUÍSA FACCHIN

**DESENVOLVIMENTO DE UMA FERRAMENTA COMPUTACIONAL
PARA AUXILIAR NA COMPARAÇÃO DE REDES DE PROTEÍNAS**

CAXIAS DO SUL

2021

ANGÉLICA LUÍSA FACCHIN

**DESENVOLVIMENTO DE UMA FERRAMENTA COMPUTACIONAL
PARA AUXILIAR NA COMPARAÇÃO DE REDES DE PROTEÍNAS**

Trabalho de Conclusão de Curso
apresentado como requisito parcial
à obtenção do título de Bacharel em
Engenharia de Computação na Área
do Conhecimento de Ciências Exatas e
Engenharias da Universidade de Caxias
do Sul.

Orientador: Prof. Dr^a. Helena Gra-
ziottin Ribeiro

CAXIAS DO SUL

2021

ANGÉLICA LUÍSA FACCHIN

**DESENVOLVIMENTO DE UMA FERRAMENTA COMPUTACIONAL
PARA AUXILIAR NA COMPARAÇÃO DE REDES DE PROTEÍNAS**

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Engenharia de Computação na Área do Conhecimento de Ciências Exatas e Engenharias da Universidade de Caxias do Sul.

Aprovado em 02 de Julho de 2021

BANCA EXAMINADORA

Prof. Dr^a. Helena Graziottin Ribeiro
Universidade de Caxias do Sul - UCS

Prof. Dr. Guilherme Holsbach Costa
Universidade de Caxias do Sul - UCS

Prof. Dr^a. Scheila de Avila e Silva
Universidade de Caxias do Sul - UCS

RESUMO

A Biologia de Sistemas é uma área interdisciplinar crescente devido a importância em análise de grandes quantidades de dados. Um ramo muito importante dessa área é o estudo das interações entre as proteínas, com o objetivo de compreender melhor os processos que ocorrem entre esses sistemas buscando informações cruciais para o desenvolvimento de novos tratamentos e medicamentos. Essa interação proteica pode ser representada através de redes e o estudo sobre essas estruturas envolve diversas áreas como a teoria dos grafos, matemática discreta e processamento computacional. Um software bastante utilizado na área da Biologia é o Cytoscape que oferece várias operações para interação e análise de redes, mas essas operações possuem algumas restrições. O objetivo desse trabalho foi o desenvolvimento de uma ferramenta que possibilitasse aprimorar as operações de união, intersecção e diferença entre duas ou mais redes de proteínas, analisando-as ao mesmo tempo através de uma técnica de simplificação de redes em conjuntos de dados, com foco principal no desenvolvimento da operação de diferença simétrica. Essa ferramenta foi implementada em C# para ser utilizada via desktop que, através de uma interface gráfica, permite selecionar os arquivos de texto que contém as redes de proteínas e também selecionar as operações a serem aplicadas sobre essas redes, oferecendo respostas com mais informações aos usuários e gerando como resultado um arquivo de texto para cada operação realizada.

Palavras-chaves: Biologia de Sistemas, Redes de Proteínas, Estrutura de Dados, Operações em Conjuntos.

ABSTRACT

Systems Biology is a growing interdisciplinary area due to its importance in analyzing large amounts of data. A very important branch of this area is the study of interactions between proteins, in order to better understand the processes that occur between these systems, seeking crucial information for the development of new treatments and drugs. This interaction of proteins can be represented through networks and the study of these structures involves several areas such as graph theory, discrete mathematics and computational processing. A widely used software in Biology is Cytoscape, which offers several interaction and network analysis operations, but these operations have some restrictions. The objective of this work was the development of a tool that could improve the union, intersection and difference operations between two or more protein networks, analyzing them at the same time through a technique of simplification of networks in data sets, with a focus main in the development of the symmetric difference operation. This tool was implemented in C# to be used via the desktop which, through a graphical interface, allows selecting the text files that contain the protein networks and also selecting the operations to be applied in these networks, offering answers with more information to users and generating as a result, a text file for each operation performed.

Keywords:Systems Biology, Protein Networks, Data Structure, Set Operations.

LISTA DE FIGURAS

Figura 1	– Representação de um grafo	17
Figura 2	– (1) Rede de proteínas coletadas do Cytoscape e (2) Proteínas coletadas salvas em arquivo de texto (formato CSV)	19
Figura 3	– Exemplo de interação proteína-proteína	22
Figura 4	– A figura mostra um pequeno subconjunto de (a) Internet, onde há roteadores conectados uns aos outros; (b) a rede de atores de Hollywood, onde dois atores estão conectados se atuaram no mesmo filme; (c) uma rede de interação proteína-proteína, onde duas proteínas são conectadas se houver evidência experimental de que elas podem se ligar uma à outra na célula. Embora a natureza dos nós e das arestas seja diferente, essas redes têm a mesma representação gráfica mostrada em (d).	23
Figura 5	– Representação de Rede não direcionada (a) X Rede direcionada (b).	24
Figura 6	– Estrutura de adjacência	24
Figura 7	– Diagrama que representa a união de A e B	25
Figura 8	– Diagrama que representa a intersecção de A e B	26
Figura 9	– Diagrama que representa a diferença de $A - B$	26
Figura 10	– Diagrama que representa a diferença simétrica de A e B	27
Figura 11	– Tela padrão do Cytoscape	28
Figura 12	– Diagrama de caso de uso para o desenvolvimento da ferramenta de comparação de redes de proteínas	29
Figura 13	– Sequência das etapas do processo para a utilização da ferramenta desenvolvida neste trabalho.	30
Figura 14	– Informações das redes de proteínas exportadas do Cytoscape para arquivos no formato CSV	31
Figura 15	– Tela de entrada de dados. (1) Botão para seleção de arquivos, (2) Lista de arquivos selecionados, (3) Botão para seleção de diretório destino dos arquivos, (4) Diretório selecionado e (5) Botão "próximo"	32
Figura 16	– Modelo utilizado para armazenamento dos dados de teste	33
Figura 17	– Armazenamento dos dados de teste	34
Figura 18	– Elementos na tela para geração dos conjuntos de proteínas resultantes. (1) Botões para geração dos arquivos de união, intersecção e diferença simétrica e (2) Campos para informar os nomes desejados para os arquivos destino	35
Figura 19	– Elementos na tela para geração dos conjuntos de proteínas personalizados	36
Figura 20	– Tela completa da ferramenta desenvolvida	37
Figura 21	– Arquivo resultante da operação de união	37
Figura 22	– Arquivo resultante da operação de intersecção	38

Figura 23 – Arquivo resultante da operação de diferença simétrica 38
Figura 24 – Arquivo resultante da operação seleção 38

LISTA DE ALGORITMOS

Algoritmo 1	Processa as linhas dos arquivos e salva as proteínas no dicionário	34
Algoritmo 2	Verifica as proteínas do conjunto intersecção	36
Algoritmo 3	Verifica as proteínas do conjunto diferença	36

LISTA DE ABREVIATURAS E SIGLAS

TCC	<i>Trabalho de conclusão de curso</i>
PPI	<i>Protein-Protein Interaction</i>
CSV	<i>Comma-separated-values</i>
\in	<i>Símbolo matemático - pertence</i>
\notin	<i>Símbolo matemático - não pertence</i>
\cup	<i>Símbolo matemático - união</i>
\cap	<i>Símbolo matemático - intersecção</i>
$-$	<i>Símbolo matemático - diferença</i>
Δ	<i>Símbolo matemático - diferença simétrica</i>
DNA	<i>Ácido Desoxirribonucleico</i>
SIF	<i>Windows Setup Information Format</i>
GML	<i>Graph Modeling Language</i>
XML	<i>eXtensible Markup Language</i>
UCS	<i>Universidade de Caxias do Sul</i>
EDUCS	<i>Editores da Universidade de Caxias do Sul</i>

SUMÁRIO

1	INTRODUÇÃO	17
2	FUNDAMENTAÇÃO TEÓRICA	21
2.1	REDES DE PROTEÍNAS	21
2.2	TEORIA DOS GRAFOS	21
2.2.1	Principais terminologias	22
2.2.2	Estruturas de Dados para Grafos	24
2.3	CONJUNTOS	25
2.3.1	Notação e Operações	25
2.4	A FERRAMENTA CYTOSCAPE	27
3	DESENVOLVIMENTO	29
3.1	ENTRADA DE DADOS	31
3.2	ARMAZENAMENTO / ESTRUTURAS DE DADOS	32
3.3	PROCESSAMENTO	35
3.4	RESULTADOS	37
4	CONSIDERAÇÕES FINAIS	41
4.1	CONCLUSÕES	41
4.2	TRABALHOS FUTUROS	42
	REFERÊNCIAS	43

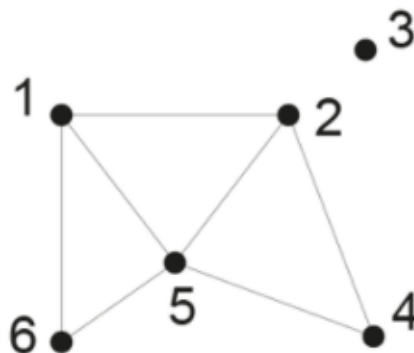
1 INTRODUÇÃO

A Biologia de Sistemas é uma área interdisciplinar nova e que cresce gradativamente devido a importância em análise de dados, que cada vez mais, estão sendo gerados em quantidade e qualidade (SILVA; NOTARI; DALL'ALBA, 2020). Essa área concentra-se no estudo sistemático de interações complexas em sistemas biológicos, buscando compreender melhor os processos que ocorrem entre esses sistemas (BARBOSA; FELISBERTO; HESPANHOL, 2017). Esse estudo sobre dados de sistemas biológicos em larga escala possibilitou entender como ocorrem as interações entre os diferentes tipos de moléculas de origem biológica e compreender de forma mais ampla, a função exercida por genes, proteínas e outras moléculas em um determinado contexto biológico (BARABASI; OLTVAI, 2004).

Pode-se chamar de rede um conjunto dessas interações e cada rede é representada por um arranjo composto por elementos e conexões, sendo sua principal característica as interações entre diversos elementos, através de interação de um par de elementos. Grande variedade de sistemas na natureza e na sociedade podem ser vistos dessa forma, tais como redes sociais, redes celulares, redes de computadores, redes de interação de proteínas, a internet e entre outros exemplos.

Redes com interações proteína-proteína (PPI) são valiosas para todos os processos biológicos, pois as funções celulares das proteínas ocorrem por meio de interações combinadas com outras proteínas. O estudo das interações entre os componentes de um sistema biológico, como as redes de proteínas, envolve Biologia Computacional que é uma área de estudo interdisciplinar que se utiliza de técnicas para representação e análise destas entidades biológicas através de sistemas de computador, envolvendo diversas áreas como a teoria dos grafos, matemática discreta e processamento computacional (STERTZ, 2004).

Figura 1 – Representação de um grafo



Fonte: SZWARCFITER (2018) - adaptada

Um grafo (Figura 1) é uma estrutura abstrata que representa um conjunto de elementos denominados vértices com interações diretas entre eles chamadas de arestas (GOLDBARG; GOLDBARG, 2012). Apesar de uma rede biológica ser uma estrutura um tanto complexa, pode-se utilizar recursos de modelagem buscando simplificar as redes para que fiquem próximas a um conjunto de elementos e assim facilitar as análises a serem realizadas. Por exemplo, uma rede de proteínas pode ser vista como um conjunto de proteínas, considerando os elementos como os seus respectivos nomes.

Para a manipulação computacional dessas redes existem alguns softwares disponíveis gratuitamente. Um desses softwares que possibilita diversas representações gráficas é o Cytoscape¹. Desenvolvido em uma parceria entre o Instituto Nacional de Ciências Médicas Gerais dos EUA (NIGMS) e a Fundação Nacional de Ciências dos EUA (NSF), esse software é uma plataforma global para a visualização e análise de complexas redes de interação moleculares.

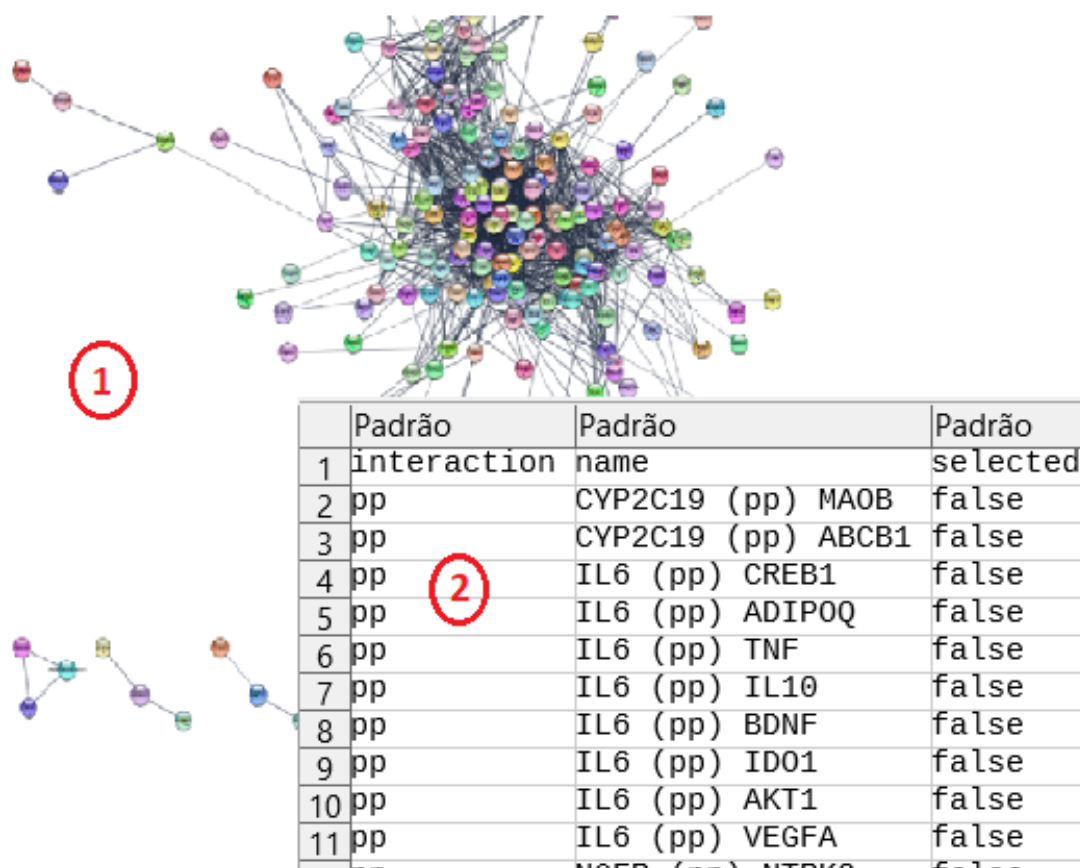
Através desse software, é possível buscar e representar informações de sistemas biológicos como uma rede de proteínas, por exemplo, por meio de mineração de texto. Além disso, o Cytoscape também permite a entrada e saída de dados em outros formatos sem perder as relações entre as redes tal como arquivos de textos.

A ferramenta possui mecanismos de análise de redes como selecionar os nós envolvidos em uma interação criando sub-redes, identificar clusters (regiões altamente interconectadas) e buscar informações de similaridades e diferenças entre as redes através de operações de união, intersecção e diferença. Essas operações são realizadas no Cytoscape de acordo com a definição delas na teoria dos conjuntos, desempenhando-as sempre entre duas redes. Porém, algumas aplicações têm a necessidade de realizar essas operações entre mais redes, exigindo que elas sejam solicitadas duas a duas, gerando resultados distintos dependendo da ordem em que as redes forem solicitadas.

O objetivo principal desse trabalho é aprimorar as operações de união, intersecção e diferença entre duas ou mais redes de proteínas, aplicando uma técnica de simplificação de redes em conjuntos de dados com foco principal no desenvolvimento da operação de diferença simétrica. Essa técnica consiste em converter as redes de proteínas coletadas do Cytoscape (Figura 2 (1)) e salvas em arquivos de texto (Figura 2 (2)) em uma lista de proteínas. A lista gerada apresentará a informação dos nomes das proteínas coletadas e a quais arquivos analisados elas pertencem. Assim, torna-se possível a visualização e a realização de operações entre mais de dois arquivos ao mesmo tempo oferecendo resultados com mais informações aos usuários.

¹ <https://cytoscape.org/>

Figura 2 – (1) Rede de proteínas coletadas do Cytoscape e (2) Proteínas coletadas salvas em arquivo de texto (formato CSV)



Fonte: o autor (2021)

O presente trabalho está organizado da seguinte forma. No Capítulo 2 são apresentados os principais conceitos de redes de proteínas, grafos, conjuntos e o software Cytoscape. No Capítulo 3 são apresentadas as etapas e os recursos utilizados para o desenvolvimento desse trabalho, sendo descrito como as operações foram realizadas e os testes feitos para chegar nos resultados. Por fim, no Capítulo 4 são apresentadas as considerações finais do trabalho realizado.

2 FUNDAMENTAÇÃO TEÓRICA

Nesse capítulo são apresentados os principais conceitos e características das redes de proteínas, assunto principal desse trabalho, além de definições sobre grafos e conjuntos, utilizados para representar as redes de proteínas e suas interações, e a ferramenta Cytoscape.

2.1 REDES DE PROTEÍNAS

As redes de proteínas fazem parte de uma grande área chamada Biologia de Sistemas, que busca compreender como as interações entre os componentes de um sistema biológico influenciam na função e comportamento do sistema, através de abordagens que envolvam teoria e modelos matemáticos e computacionais. É possível denotar as interações entre os diferentes componentes do sistema através da representação matemática por grafos que será apresentada na seção 2.2.

Quando se busca entender algum sistema, primeiro é preciso compreender como os seus elementos interagem entre si. Uma rede é um conjunto de elementos, denominados nós ou vértices e as interações diretas entre eles, chamadas de links ou arestas (BARABASI *et al.*, 2016). A Figura 3 representa um exemplo de redes de proteínas onde ocorrem interações proteína-proteína.

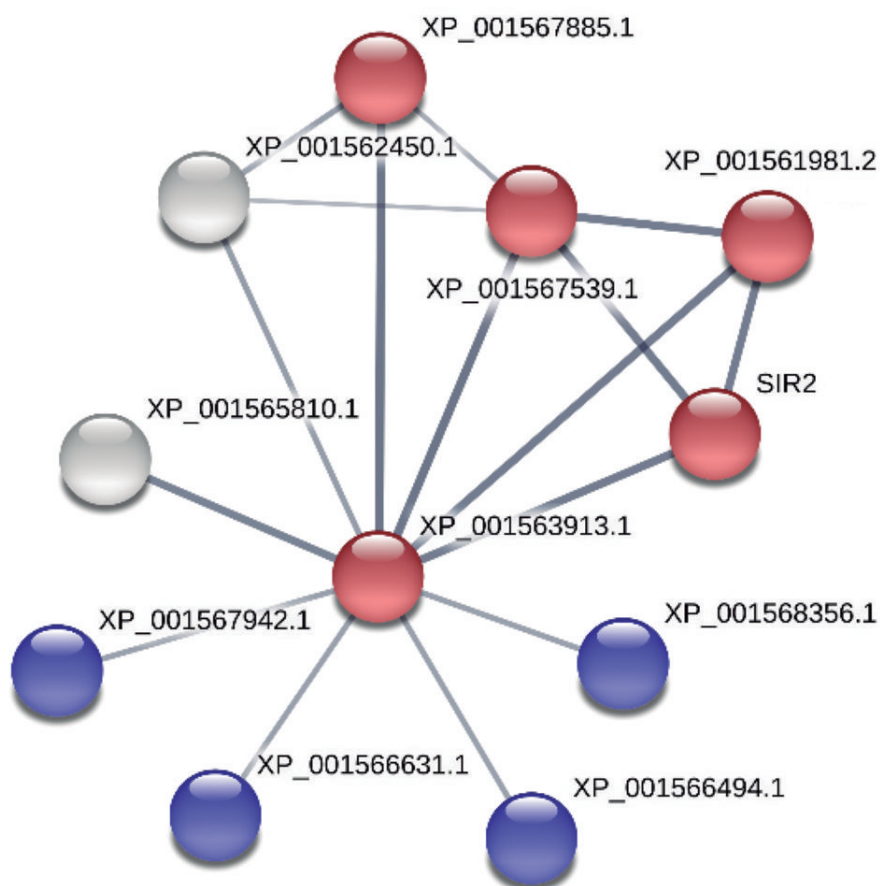
As interações proteína-proteína (PPI) são valiosas para todos os processos biológicos, pois as funções celulares das proteínas ocorrem por meio de interações combinadas com outras proteínas. Além disso, os membros da rede PPI são cruciais fontes de informações para o desenvolvimento de novos medicamentos. Portanto, é importante estudar e compreender a natureza específica dessas interações (ORTIZ-JOYA; CONTRERAS-RODRÍGUEZ; RAMÍREZ-HERNÁNDEZ, 2019).

2.2 TEORIA DOS GRAFOS

Um grafo é uma estrutura abstrata que representa um conjunto de elementos denominados vértices com interações diretas entre eles chamadas de arestas. Essa técnica é bastante útil na representação e solução de numerosos tipos de problemas (GOLDBARG; GOLDBARG, 2012). Atualmente a teoria de grafos é estudada e aplicada em diversas áreas da ciência e embora um grafo seja uma estrutura de dados muito parecida com uma árvore, ao contrário desta ele não apresenta hierarquia.

Para exemplificar, pode-se citar as redes sociais. Em geral, as redes sociais são representadas em grafos, exprimindo graficamente os indivíduos e suas relações sociais, e diversos estudos e tipos de análises sobre as interações são realizadas a partir do uso de algoritmos que operam sobre grafos.

Figura 3 – Exemplo de interação proteína-proteína



Fonte: ORTIZ-JOYA; CONTRERAS-RODRÍGUEZ; RAMÍREZ-HERNÁNDEZ (2019)

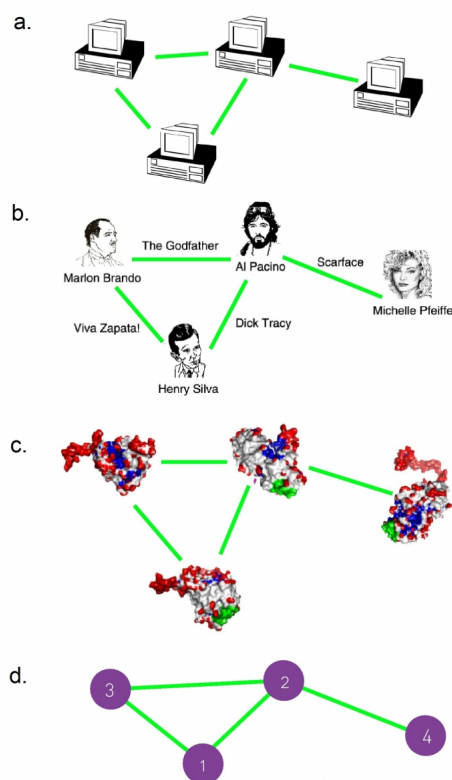
Esse conceito oferece uma linguagem comum para estudar sistemas que podem diferir muito em natureza, aparência ou escopo. De fato, como apresentado na Figura 4, três sistemas bastante diferentes têm exatamente a mesma representação de rede (BARABASI *et al.*, 2016).

Em muitos casos, grafos e redes são usados como sinônimos. Pode-se dizer que um grafo é um modelo matemático, enquanto uma rede pode ser modelada matematicamente por um grafo, porém possui um contexto e informações ligadas a elas (HEBERLE, 2015).

2.2.1 Principais terminologias

Para compreender melhor os conceitos sobre grafos, pode-se considerar a seguinte notação matemática: um grafo G é representado por um conjunto finito de nós N e um conjunto de arestas A , onde as arestas representam os caminhos ou as ligações entre os nós. Sendo assim, temos $G = (N, A)$. O exemplo da Figura 4 (d), traz um grafo G_1 com quatro nós, rotulados de 1 a 4, e quatro relações binárias entre eles, sendo assim, $N = \{1, 2, 3, 4\}$ e $A = \{(1, 2), (1, 3), (2, 3), (2, 4)\}$.

Figura 4 – A figura mostra um pequeno subconjunto de (a) Internet, onde há roteadores conectados uns aos outros; (b) a rede de atores de Hollywood, onde dois atores estão conectados se atuaram no mesmo filme; (c) uma rede de interação proteína-proteína, onde duas proteínas são conectadas se houver evidência experimental de que elas podem se ligar uma à outra na célula. Embora a natureza dos nós e das arestas seja diferente, essas redes têm a mesma representação gráfica mostrada em (d).

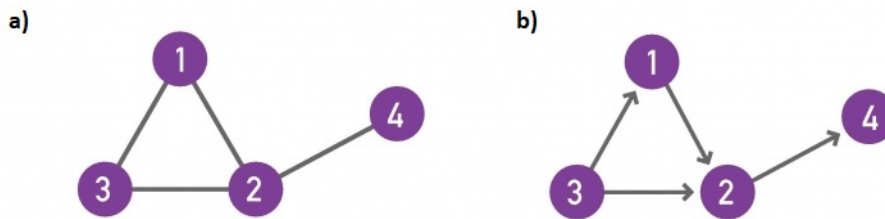


Fonte: BARABASI *et al.* (2016)

Dependendo do tipo de informação que se busca representar na forma de grafos, pode-se atribuir informações às arestas ou não. Por exemplo, se os nós representam cidades e as arestas representam estradas entre essas cidades, é possível associar a cada aresta a distância entre essas cidades e a rodovia que as conecta.

Uma rede é chamada de direcionada (ou dígrafo) se todas as suas arestas forem direcionadas, havendo um fluxo dirigido entre elas, e ela é chamada de não direcionada se todas as suas arestas forem não direcionadas (arestas bidirecionais) (BARABASI *et al.*, 2016). É possível visualizar a representação gráfica dessas duas redes através da Figura 5. No contexto desse trabalho cada ligação na rede, entre duas proteínas, é representada por "pp" e essas ligações não possuem direcionamento e nem há informações associadas a elas.

Figura 5 – Representação de Rede não direcionada (a) X Rede direcionada (b).

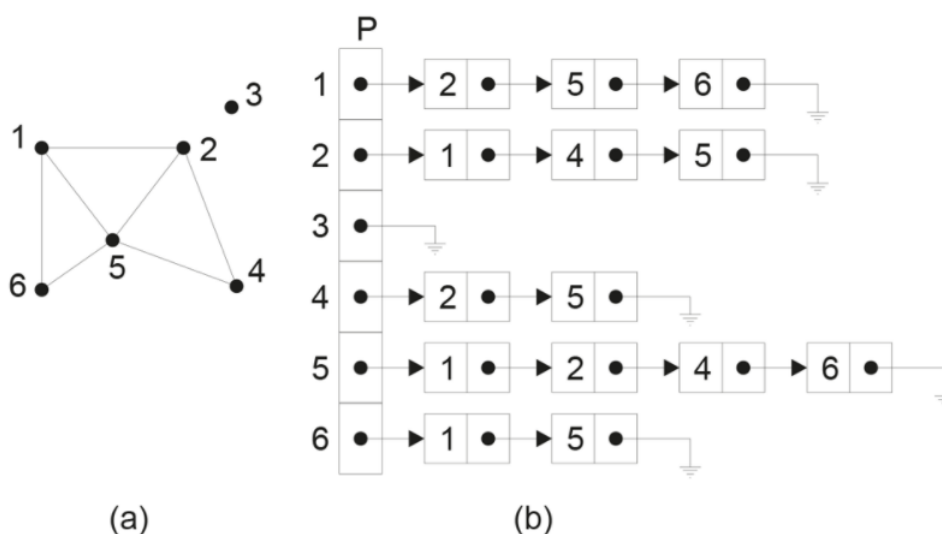


Fonte: BARABASI *et al.* (2016), adaptada

2.2.2 Estruturas de Dados para Grafos

Para o ser humano, a representação de grafos visualmente é conveniente, contudo computacionalmente essa representação precisa ser feita através de estruturas de dados, para que elas possam ser manipuladas e, até mesmo, representadas de forma gráfica. Dentre os diferentes tipos de representações adequadas ao computador, destacam-se as representações matriciais e as por listas (SZWARCFITER, 2018). Essas estruturas são amplamente utilizadas por diferentes operações de manipulação de grafos e facilmente representadas em diferentes linguagens de programação. Por exemplo, a estrutura de adjacência, que é uma representação de grafos por listas. A Figura 6(b) ilustra as listas de adjacências do grafo da Figura 6(a). A lista de adjacência é composta por um vetor P de dimensão n . Cada elemento de P contém dois campos: a identificação de um vértice e um ponteiro para uma lista encadeada contendo os vizinhos do vértice correspondente (GOLDBARG; GOLDBARG, 2012).

Figura 6 – Estrutura de adjacência



Fonte: SZWARCFITER (2018)

Apesar de não considerar as informações das arestas para a análise de dados desse

trabalho, uma estrutura semelhante à Estrutura de Adjacência da figura 6 foi construída. Desse modo, a estrutura permitiu que uma proteína P tenha claramente a informação de todos os arquivos onde ela está inserida e com essa informação, torna-se possível executar as operações necessárias e gerar os conjuntos de proteínas almejados.

2.3 CONJUNTOS

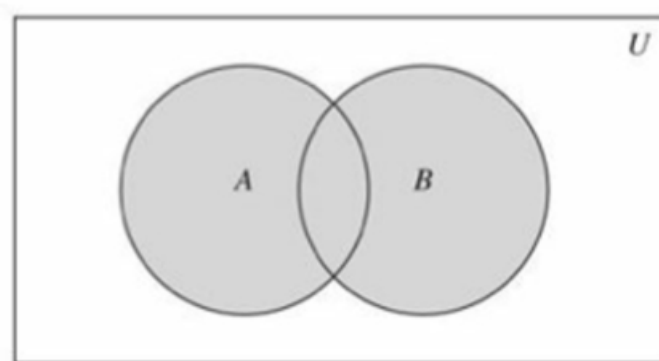
A teoria dos conjuntos é a teoria matemática capaz de agrupar objetos. Esses objetos são denominados elementos e, na teoria dos conjuntos, não importa do que os elementos são formados. Um conjunto pode ser conceituado como uma coleção de elementos agrupados, sem repetição e não ordenada (SCHEINERMAN, 2003). Segundo MENEZES; MENEZES, o conceito de conjunto é fundamental e importante na computação, pois a maioria das aplicações são baseadas em construções sobre conjuntos e a partir dessa estrutura é possível construir estruturas mais complexas.

2.3.1 Notação e Operações

No geral, os conjuntos são representados por letras maiúsculas e seus elementos por letras minúsculas. Para indicar que um elemento a pertence a um conjunto A , utiliza-se a seguinte representação: $a \in A$. Para representar um elemento que não pertence: $a \notin A$.

A **união** entre dois conjuntos A e B , $A \cup B$, é o conjunto que contém todos os elementos de A e de B . Portanto, se $A = \{a, e, i, o\}$ e $B = \{a, e, o, u\}$, então $A \cup B = \{a, e, i, o, u\}$.

Figura 7 – Diagrama que representa a união de A e B



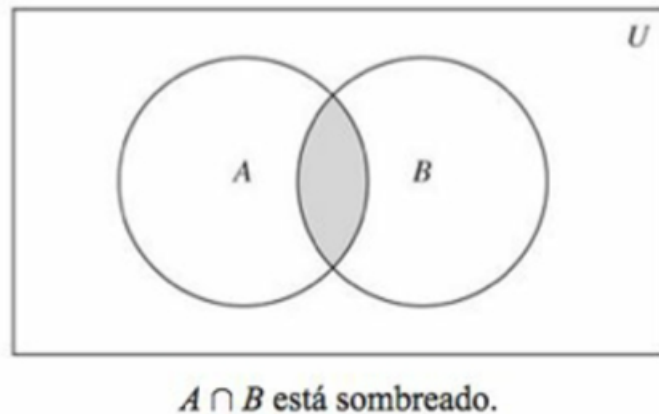
$A \cup B$ está sombreado.

Fonte: ROSEN (2009)

A figura 7 representa a união de dois conjuntos A e B . A área que representa a $A \cup B$ é a área sombreada dos círculos que representam ou A ou B .

A **intersecção** entre dois conjuntos A e B , $A \cap B$, é o conjunto que contém os elementos comuns entre eles. Considerando os mesmos conjuntos A e B descritos acima, então $A \cap B = \{a, e, o\}$, pois somente esses três elementos aparecem em ambos os conjuntos.

Figura 8 – Diagrama que representa a intersecção de A e B

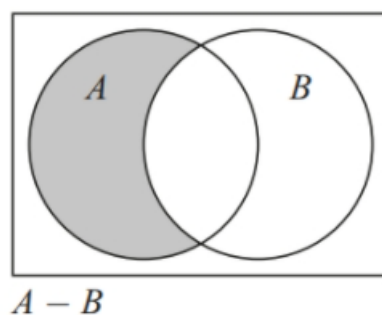


Fonte: ROSEN (2009)

A figura 8 representa a intersecção de dois conjuntos A e B . A área que representa $A \cap B$ é a área sombreada entre os dois círculos.

A **diferença** entre dois conjuntos A e B , $A - B$, é o conjunto que contém os elementos que estão em A e não estão em B . Considerando os mesmos conjuntos A e B , então $A - B = \{i\}$. Para o caso da diferença, a operação não é comutativa, ou seja, $A - B$ é diferente de $B - A$.

Figura 9 – Diagrama que representa a diferença de $A - B$

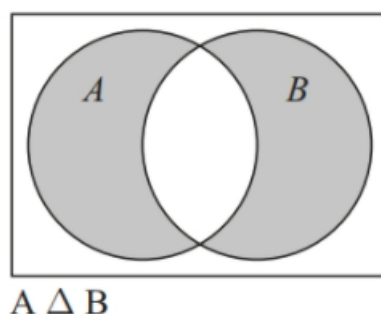


Fonte: SCHEINERMAN (2003)

A figura 9 representa a diferença de dois conjuntos A e B . A área que representa $A - B$ é a área sombreada no círculo A .

Um caso especial de diferença entre conjuntos, é a **diferença simétrica**, $A \Delta B$, que é o conjunto de todos os elementos que estão em A , mas não estão em B , ou estão em B , porém não em A . Isto é, $A \Delta B = (A - B) \cup (B - A)$.

Figura 10 – Diagrama que representa a diferença simétrica de A e B



Fonte: SCHEINERMAN (2003)

A figura 10 representa a diferença simétrica de dois conjuntos A e B . Quando são analisados dois conjuntos de elementos, pode-se considerar essa operação como sendo o resto da intersecção desses conjuntos.

As definições e notações apresentadas nesta subseção, foram retiradas do livros de SCHEINERMAN (2003) e ROSEN (2009).

2.4 A FERRAMENTA CYTOSCAPE

Cytoscape¹ é uma ferramenta criada com o objetivo de integrar, analisar e visualizar redes de interação molecular e vias biológicas. Quando combinada a bancos de dados proteína-proteína, proteína-DNA e interações genéticas essa ferramenta pode ser bastante poderosa.

Embora o Cytoscape seja mais comumente usado para fins de pesquisa biológica, essa ferramenta também pode ser utilizada para visualizar e analisar grafos de redes de qualquer tipo envolvendo nós e arestas (por exemplo, redes sociais).

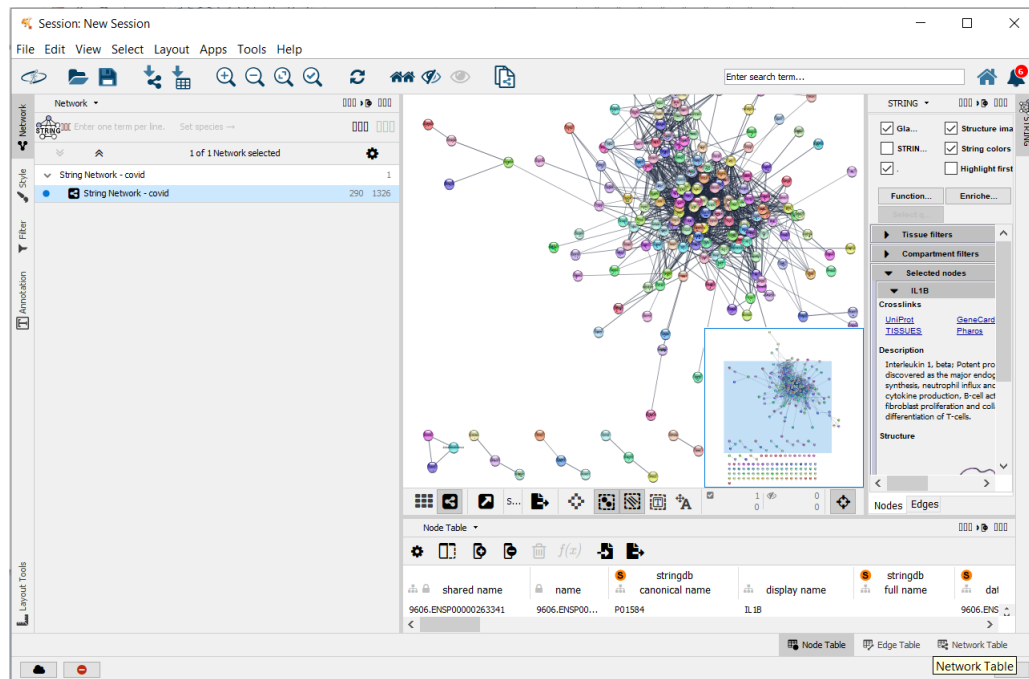
Por apresentar uma arquitetura aberta, é possível introduzir ao software recursos adicionais através de aplicativos que são desenvolvidos pelos principais desenvolvedores e pela grande comunidade de usuários, utilizando a linguagem de programação JAVA. Uma grande parte desses aplicativos encontra-se disponível gratuitamente na Cytoscape App Store e são verificados para rodar em Linux, Windows e MacOS X.

Um recurso importante do Cytoscape é que, a partir de palavras-chave é possível realizar um busca global na rede encontrando em artigos relações proteína-proteína. Após essa mineração de texto, o software disponibiliza uma visualização tridimensional da rede de proteínas encontrada, por meio de grafos.

Na Figura 11 pode-se ver a tela padrão do Cytoscape após a realização de uma busca na rede, onde se destacam três áreas importantes. Na parte mais a esquerda, a ferramenta nos mostra

¹ Fonte bibliográfica utilizada em toda a seção 2.4: Cytoscape User Manual <<https://manual.cytoscape.org/en/stable/>>

Figura 11 – Tela padrão do Cytoscape



Fonte: o autor (2020)

as redes pesquisadas e as ações realizadas entre elas. Na direita superior, pode-se visualizar as redes de proteínas e interagir com elas, por exemplo, movimentando os nodos. Por fim, na direita inferior, a ferramenta traz uma tabela com as informações contidas na rede de proteínas, e a partir dessa tabela é possível aplicar filtros as redes, manipular os nodos e definir estilos de formatação dos elementos apresentados.

Além de criar redes através de mineração de texto, o cytoscape também permite a entrada e saída de dados em outros formatos sem perder as relações entre as redes. Por exemplo, os dados podem ser carregados e salvos em diversos tipos de arquivos como: formato SIF (contendo listas de pares de interação proteína-proteína e / ou proteína-DNA), GML, XML e arquivos de texto delimitados por tabulação ou espaço (formato CSV).

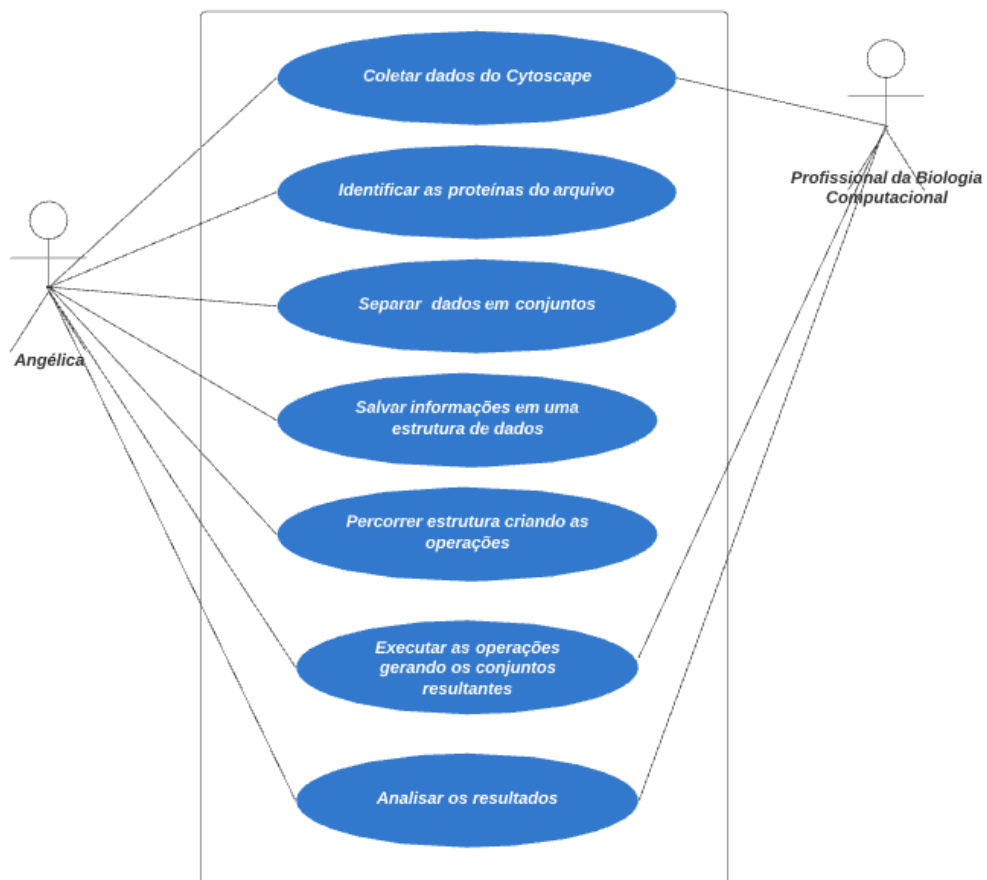
O software permite que várias redes sejam carregadas ao mesmo tempo. Uma rede armazena todos os nós e arestas que são carregados pelo usuário e uma visualização os exibe. A partir dos dados carregados, pode-se realizar operações para análise das redes.

Uma opção é filtrar a rede para selecionar subconjuntos de nós e / ou interações. Por exemplo, os usuários podem selecionar nós envolvidos em uma interação. Também pode-se identificar clusters (regiões altamente interconectadas), que em uma rede de similaridade de proteínas representam famílias de proteínas. Outra opção de análise é realizar operações matemáticas de união, intersecção e diferença sobre as redes, porém o Cytoscape permite que essas operações sejam realizadas apenas duas a duas (entre duas redes) e seguem a semântica dessas operações na teoria dos conjuntos, inclusive a diferença sendo realizada como operação não simétrica.

3 DESENVOLVIMENTO

O objetivo deste trabalho constituiu-se em construir uma ferramenta computacional para comparação de redes de proteínas, buscando contribuir com processo de comparação das redes, analisando dois ou mais arquivos ao mesmo tempo. A função dessa ferramenta é encontrar semelhanças e diferenças entre as redes utilizando as propriedades de conjuntos e, principalmente, a operação de diferença simétrica que o Cytoscape não realiza. A construção da ferramenta foi realizada através do desenvolvimento de um programa em linguagem C# (C Sharp) utilizando o ambiente de desenvolvimento integrado (IDE) da Microsoft, Visual Studio, para a plataforma Windows. Para ser executado, o programa precisa ser compilado gerando um programa executável, que ao ser acionado permite a utilização da ferramenta em desktop. O processo de desenvolvimento dessa ferramenta está representado pelo diagrama de caso de uso (Figura 12) a seguir:

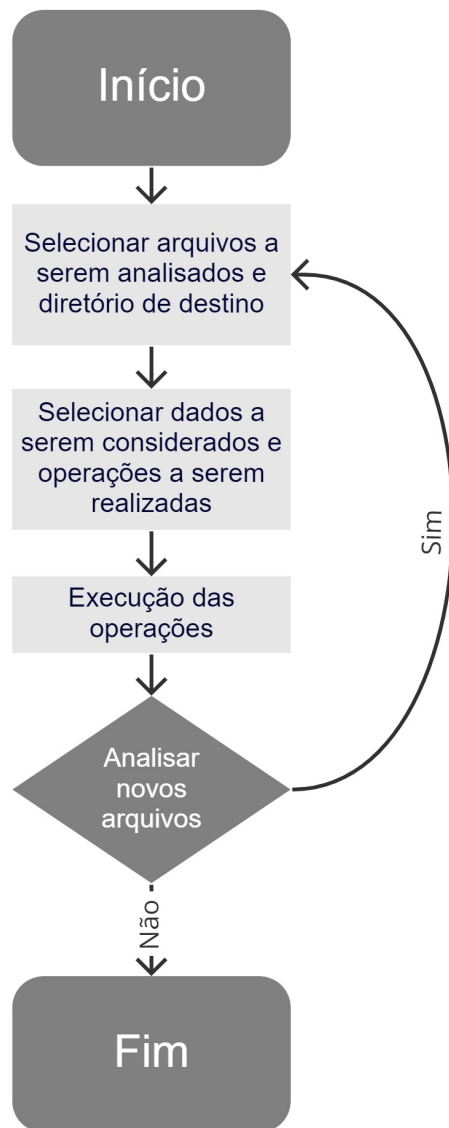
Figura 12 – Diagrama de caso de uso para o desenvolvimento da ferramenta de comparação de redes de proteínas



Fonte: O Autor (2021)

O fluxograma apresentado na imagem 13 representa a sequência das etapas do processo para a utilização da ferramenta desenvolvida neste trabalho. O primeiro passo é selecionar os arquivos a serem analisados e o diretório de destino, após a seleção dos arquivos que se deseja analisar é possível selecionar as operações que se pretende executar, são elas: união, intersecção, diferença simétrica e seleção. Para a opção da operação seleção também é necessário selecionar os dados a serem considerados. Por fim, as operações são executadas e os resultados guardados em arquivos. Caso seja necessário analisar outros arquivos, é possível voltar a primeira etapa de seleção e selecionar novos arquivos.

Figura 13 – Sequência das etapas do processo para a utilização da ferramenta desenvolvida neste trabalho.



Fonte: O Autor (2021)

3.1 ENTRADA DE DADOS

As redes de proteínas trabalhadas foram coletadas por profissionais da área da Biologia Computacional através do Cytoscape e exportadas para arquivos no formato CSV. “Os arquivos CSV (do inglês "Character-separated values" ou "valores separados por um delimitador") servem para armazenar dados tabulares (números e texto) em texto simples” (PIRES, 2015).

A Figura 14 apresenta um exemplo de dados coletados do software e transferidos para o arquivo. Dentro de cada arquivo pode-se identificar as redes representadas por grafos sendo que a segunda coluna do arquivo (a coluna "name") apresenta a informação de duas proteínas (vértices) unidas por uma aresta representada por "pp" (proteína-proteína). Nesse caso, as arestas não possuem informações e indicam apenas as conexões entre os nodos dos grafos analisados.

Figura 14 – Informações das redes de proteínas exportadas do Cytoscape para arquivos no formato CSV

	Padrão	Padrão	Padrão	Padrão	Padrão	Padrão
1	interaction	name	selected	shared interaction	shared name	stringdb::coexpression
2	pp	CYP2C19 (pp) MAOB	false	pp	CYP2C19 (pp) MAOB	0.063
3	pp	CYP2C19 (pp) ABCB1	false	pp	CYP2C19 (pp) ABCB1	0.062
4	pp	IL6 (pp) CREB1	false	pp	IL6 (pp) CREB1	
5	pp	IL6 (pp) ADIPOQ	false	pp	IL6 (pp) ADIPOQ	
6	pp	IL6 (pp) TNF	false	pp	IL6 (pp) TNF	0.123
7	pp	IL6 (pp) IL10	false	pp	IL6 (pp) IL10	0.077
8	pp	IL6 (pp) BDNF	false	pp	IL6 (pp) BDNF	
9	pp	IL6 (pp) ID01	false	pp	IL6 (pp) ID01	0.076
10	pp	IL6 (pp) AKT1	false	pp	IL6 (pp) AKT1	0.055
11	pp	IL6 (pp) VEGFA	false	pp	IL6 (pp) VEGFA	0.063
12	pp	NGFR (pp) NTRK2	false	pp	NGFR (pp) NTRK2	0.072
13	pp	NGFR (pp) IGF1	false	pp	NGFR (pp) IGF1	
14	pp	NGFR (pp) NGF	false	pp	NGFR (pp) NGF	
15	pp	NGFR (pp) INS	false	pp	NGFR (pp) INS	

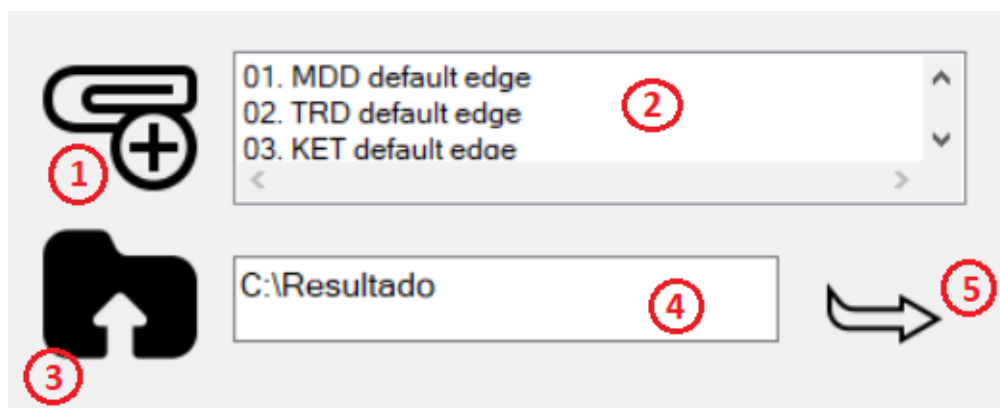
Fonte: O Autor (2020)

Apesar de ser uma área complexa, a Biologia de Sistemas busca simplificar os dados para que fiquem próximos a um conjunto de elementos. Uma rede de proteínas pode ser vista como um conjunto de proteínas, considerando os elementos como os seus respectivos nomes (HEBERLE, 2015).

A ferramenta desenvolvida nesse trabalho realiza a leitura de dois ou mais arquivos contendo redes de proteínas geradas pelo Cytoscape, nos quais identificou-se cada uma das proteínas presentes nesses arquivos, armazenando-as em conjuntos, sendo um conjunto para cada arquivo. Cada arquivo representa um conjunto diferente de proteínas e cada proteína presente no arquivo é um elemento do conjunto.

Para facilitar a entrada dos dados pelos arquivos, desenvolveu-se uma interface gráfica que tem a parte da tela de entrada de dados representada na figura 15.

Figura 15 – Tela de entrada de dados. (1) Botão para seleção de arquivos, (2) Lista de arquivos selecionados, (3) Botão para seleção de diretório destino dos arquivos, (4) Diretório selecionado e (5) Botão "próximo"



Fonte: O Autor (2021)

Na figura 15 (1) é possível selecionar os arquivos exportados do Cytoscape. A ferramenta permite selecionar vários arquivos ao mesmo tempo e depois de selecionados são descritos no campo da figura 15 (2). Na figura 15 (3) é possível selecionar o diretório de destino que se deseja gerar os arquivos com o resultado das operações e depois de selecionado é descrito no campo da figura 15 (4). Ao clicar no botão da figura 15 (5), a ferramenta valida se os arquivos e o diretório de destino foram selecionados, bloqueia esses campos para edição e passa para as próximas fases que são armazenamento de dados e processamento de dados, liberando em tela as opções para geração dos arquivos destino.

3.2 ARMAZENAMENTO / ESTRUTURAS DE DADOS

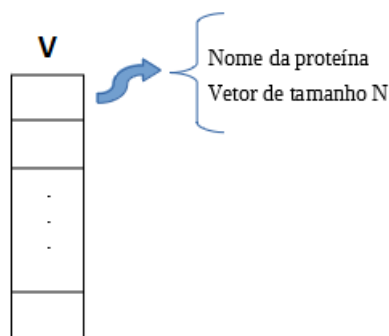
Conforme os testes realizados, a melhor maneira de obter os resultados esperados é explorando os arquivos ao mesmo tempo, salvando todas as informações das redes a serem analisadas em uma única estrutura de dados, por exemplo, vetor, matriz ou estrutura de adjacência (vista na seção 2.2.2).

Realizou-se testes na linguagem de programação C, utilizando uma estrutura do tipo vetor (V), onde cada posição do vetor corresponde a um struct¹ (Figura 16). Nesse caso, cada struct continha uma variável com o nome da proteína e um vetor de tamanho N (número de conjuntos analisados) que indica se a proteína pertence, ou não, a aquele determinado conjunto.

Também foram realizados testes na linguagem de programação C# em duas estruturas diferentes. A primeira forma analisada, foi criando uma lista de objetos, onde cada objeto continha o nome da proteína e um vetor com a informação de quais arquivos aquela determinada

¹ Struct: Estrutura que contém um conjunto de uma ou mais variáveis, agrupadas sob um único nome, buscando facilitar a sua referência. Essas variáveis podem ser qualquer tipo de dados válidos na Linguagem C de programação (tipos básicos, vetores, strings, ponteiros ou mesmo outras estrutura) (DAMAS, 2007).

Figura 16 – Modelo utilizado para armazenamento dos dados de teste



Fonte: O Autor (2020)

proteína pertencia. A segunda forma analisada foi através da criação de um dicionário² e nesse caso, as chaves são os nomes das proteínas e os valores são vetores com a informação de quais arquivos aquela determinada proteína pertence.

Pela agilidade no processamento e facilidade na manipulação dos dados, optou-se por desenvolver a ferramenta em linguagem de programação C# utilizando a classe "Dictionary" para armazenar e manipular os conjuntos de proteínas. A linguagem de programação C# é uma linguagem que trabalha com objetos e classes pré-definidas, com operações já desenvolvidas para elas e com processamento rápido, acelerando bastante o processo de desenvolvimento de programas.

A Figura 17 está representando como os dados de três arquivos analisados simultaneamente foram armazenados. Os nomes das proteínas representam as chaves do dicionário e o restante das informações representam o valor associado a cada chave, ou seja, o vetor de arquivos que possui a informação se uma determinada proteína pertence a um determinado arquivo. Sendo assim, 0 representa que a proteína não pertence (\notin) ao conjunto e 1 que pertence (\in). Em nenhum momento pode haver um vetor onde todos os valores sejam 0, pois se uma proteína está no dicionário então, obrigatoriamente, ela pertence a algum arquivo.

O algoritmo 1 processa os dados de cada um dos arquivos e guarda as informações no dicionário. Este algoritmo funciona da seguinte forma: para cada linha do arquivo ele busca a informação presente na segunda coluna (linha 3 do algoritmo), que contém as informações da interação entre duas proteínas, ao obter essa informação o algoritmo separa as duas proteínas (linha 4 do algoritmo) e verifica a existência de cada uma delas no dicionário. Caso a proteína já exista, o código apenas informa 1 na posição do vetor correspondente ao arquivo analisado e caso contrário insere a proteína no dicionário e informa 1 na posição do vetor correspondente ao arquivo analisado.

² Dicionário: Classe abstrata composta de uma coleção de pares (chave, valor), de modo que cada chave possível apareça no máximo uma vez na coleção. Através de cada chave fornecida, chega-se ao seu valor armazenado. (MEHLHORN; SANDERS, 2008).

Figura 17 – Armazenamento dos dados de teste

Nome	v0	v1	v2
p1	1	1	0
p2	1	1	0
p3	1	0	0
p4	1	0	0
p5	1	1	1
p6	0	1	0
p7	0	0	1
p8	0	1	1
p9	0	0	1
p10	0	0	1

Fonte: O Autor (2020)

Algoritmo 1 – Processa as linhas dos arquivos e salva as proteínas no dicionário

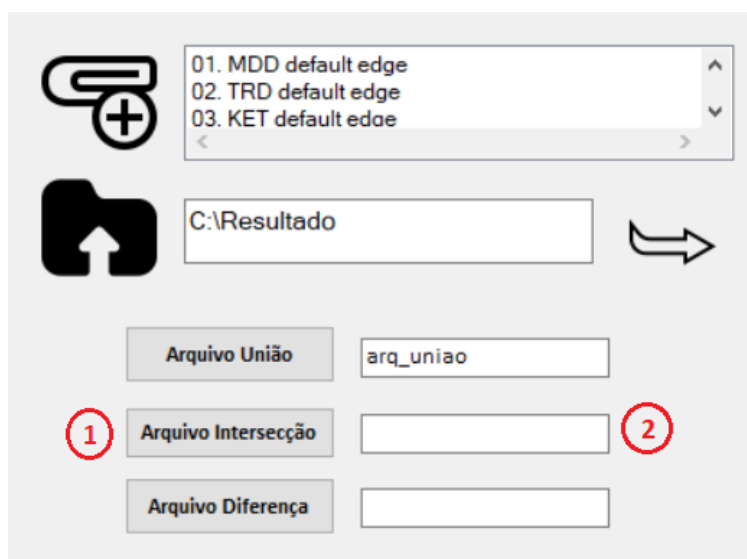
```
1  foreach (string line in lines){
2      string [] fields = line.Split(',');
3      string name = fields[1];
4      string [] proteinas = name.Split("(pp)");
5      string p1 = proteinas[0] + "\"";
6      string p2 = "\" + proteinas[1];
7      if (!dicionario.ContainsKey(p1))
8          dicionario.Add(p1, (int[]) vetorZeros.Clone());
9      dicionario[p1][file] = 1;
10     if (!dicionario.ContainsKey(p2))
11         dicionario.Add(p2, (int[]) vetorZeros.Clone());
12     dicionario[p2][file] = 1;
13 }
```

3.3 PROCESSAMENTO

Após a obtenção dos dados das redes e o armazenamento correto desses dados, pode-se percorrer o dicionário e localizar os conjuntos de proteínas desejados. Ao aplicar operações semelhantes as que vimos na seção 2.3 em conjuntos, pode-se obter os seguintes conjuntos de dados: operação de união (conjunto contendo todas as proteínas encontradas nos arquivos analisados), operação de intersecção (conjunto com as proteínas comuns a todos os arquivos analisados), operação de diferença simétrica (conjunto de proteínas que aparecem somente em um dos arquivos analisados).

Esperando facilitar a geração desses conjuntos de proteínas, foram incluídos na tela os elementos que pode-se observar na imagem 18. Na figura 18 (1) tem-se botões para a geração de cada um dos conjuntos e ao lado de cada um deles (figura 18 (2)), um campo para informar o nome desejado para o arquivo destino.

Figura 18 – Elementos na tela para geração dos conjuntos de proteínas resultantes. (1) Botões para geração dos arquivos de união, intersecção e diferença simétrica e (2) Campos para informar os nomes desejados para os arquivos destino



Fonte: O Autor (2021)

Para o conjunto união, o arquivo gerado é o conjunto de todas as proteínas encontradas nos arquivos analisados e salvas no dicionário. Para o conjunto intersecção, um novo dicionário é gerado a partir do dicionário principal, utilizando o código do algoritmo 2. Esse algoritmo verifica se a soma dos elementos do vetor da proteína é igual ao número de arquivos analisados. Caso essa soma seja igual, significa que a proteína pertence a todos os arquivos e conseqüentemente faz parte do conjunto intersecção, sendo assim, o arquivo gerado contém as proteínas desse novo dicionário. Já para o conjunto diferença, um novo dicionário é gerado a partir do dicionário principal, utilizando o código do algoritmo 3. Esse algoritmo verifica se a soma dos elementos do vetor da proteína é igual a 1, ou seja, a proteína pertence a um único arquivo e conseqüentemente

faz parte do conjunto de diferença simétrica. Assim como para a intersecção, o arquivo gerado contém as proteínas desse novo dicionário.

Algoritmo 2 – Verifica as proteínas do conjunto intersecção

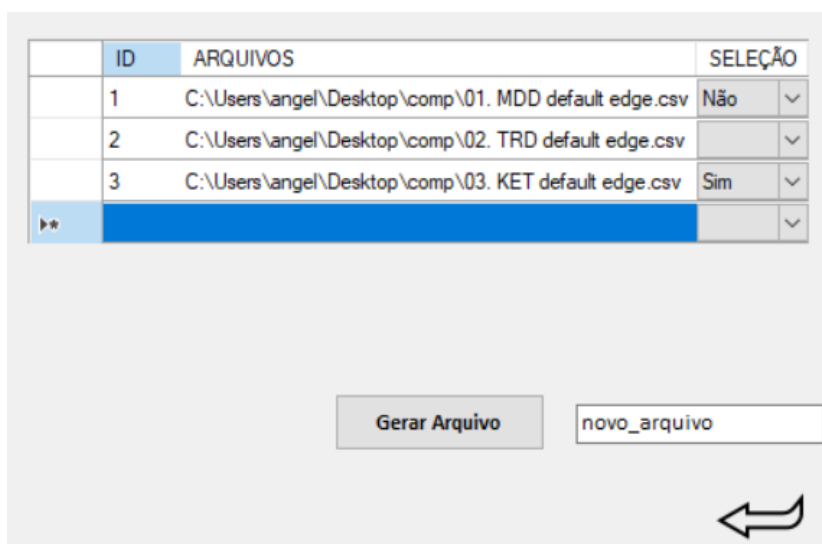
```
1 intersec = dicionario.Where(proteina => proteina.Value.Sum() == qtdFiles);
```

Algoritmo 3 – Verifica as proteínas do conjunto diferença

```
1 diferenca = dicionario.Where(proteina => proteina.Value.Sum() == 1);
```

Além dos conjuntos principais, uma operação extra de arquivos foi inserida na ferramenta. Para realizar a operação de seleção entre os arquivos de origem. Nesta operação, o usuário pode escolher se quer que a operação seja aplicada nas proteínas que fazem parte de um arquivo, ou nas proteínas que não fazem parte dele. Por exemplo, na figura 19 o arquivo 1 está marcado como não, o arquivo 2 não está marcado e o arquivo 3 está marcado como sim. Sendo assim, o arquivo resultante dessa operação apresentará as proteínas que não pertencem ao arquivo 1, indiferente para o arquivo 2, ou seja, tanto faz se pertencem ou não ao arquivo, e que pertencem ao arquivo 3 simultaneamente. Com essa funcionalidade é possível gerar diversas combinações diferentes entre os arquivos, gerando ainda mais informações sobre as redes analisadas aos usuários. Para gerar esse arquivo resultante, o botão Gerar Arquivo precisa ser acionado e assim como para os demais arquivos apresenta o campo ao lado desse botão possibilitando informar o nome desejado ao arquivo destino.

Figura 19 – Elementos na tela para geração dos conjuntos de proteínas personalizados



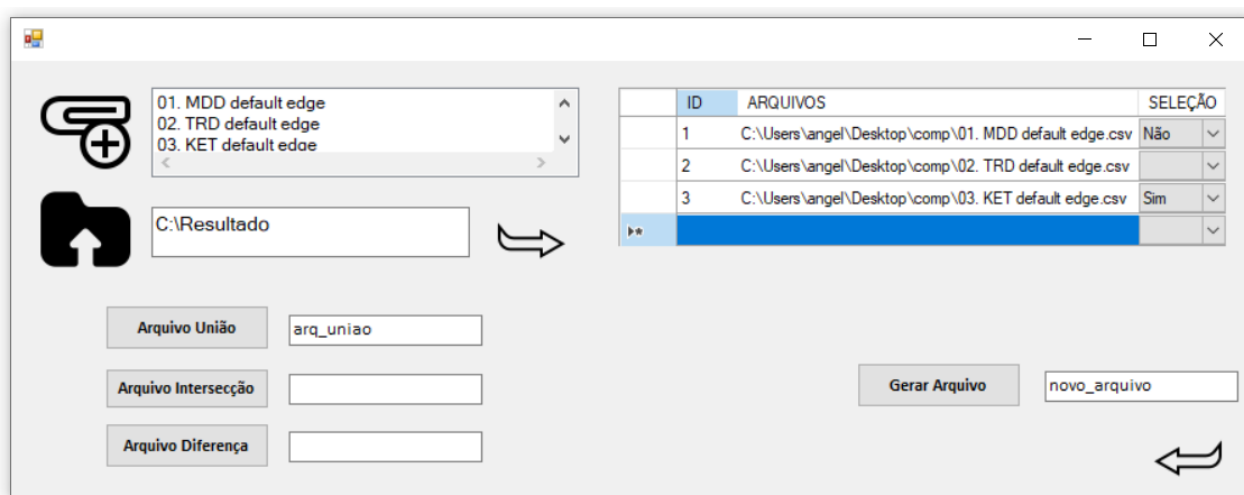
Fonte: O Autor (2021)

No canto inferior direito da figura 19, tem-se o botão voltar. Ao pressionar esse botão os campos serão limpos e escondidos novamente, deixando visível apenas os componentes da tela de entrada de dados (figura 15) e tornando possível selecionar novos arquivos para a análise.

3.4 RESULTADOS

Na Figura 20 pode-se observar a tela completa desenvolvida para a ferramenta. Após a aplicação dos comandos apresentados na seção 3.3 sobre 5 conjuntos de proteínas, obteve-se os resultados presentes nas figuras: Figura 21 (arquivo união), Figura 22 (arquivo intersecção), Figura 23 (arquivo diferença simétrica) e Figura 24 (arquivo seleção, considerando os seguintes valores de seleção {sim,não, , , }). Além do resultado da operação, exibindo apenas as proteínas resultantes de sua execução, cada arquivo mostra também a informação de qual, ou quais arquivos a proteína pertence.

Figura 20 – Tela completa da ferramenta desenvolvida



Fonte: O Autor (2021)

Figura 21 – Arquivo resultante da operação de união

	Padrão	Padrão	Padrão	Padrão	Padrão	Padrão
	Name	File 1	File 2	File 3	File 4	File 5
1	Name	File 1	File 2	File 3	File 4	File 5
2	cCYP2C19	1	0	0	0	0
3	MAOB1	1	1	0	0	0
4	ABCB1	1	1	1	1	0
5	ccIL6	1	0	0	0	0
6	CREB1	1	1	1	1	1
7	cIL6	1	0	0	0	0
8	ADIPOQ	1	1	1	0	0
9	TNF	1	1	1	1	1
10	IL10	1	1	1	1	1
11	BDNF	1	1	1	1	1
12	IL6	1	1	1	1	1
13	IDO1	1	1	1	1	0
14	AKT1	1	1	1	1	1
15	VEGFA	1	1	1	1	1
16	NGFR	1	1	1	0	0
17	NTRK2	1	1	1	1	1
18	IGF1	1	1	1	0	0
19	NGF	1	1	1	1	1
20	INS	1	1	1	1	1
21	IL2	1	1	1	0	1
22	IFNG	1	1	1	1	0
23	IL4	1	1	1	0	1

Fonte: O Autor (2021)

Figura 22 – Arquivo resultante da operação de intersecção

	Padrão	Padrão	Padrão	Padrão	Padrão	Padrão
1	Name	File 1	File 2	File 3	File 4	File 5
2	CREB1	1	1	1	1	1
3	TNF	1	1	1	1	1
4	IL10	1	1	1	1	1
5	BDNF	1	1	1	1	1
6	IL6	1	1	1	1	1
7	AKT1	1	1	1	1	1
8	VEGFA	1	1	1	1	1
9	NTRK2	1	1	1	1	1
10	NGF	1	1	1	1	1
11	INS	1	1	1	1	1
12	CRP	1	1	1	1	1
13	IL1B	1	1	1	1	1
14	ALB	1	1	1	1	1
15	PRL	1	1	1	1	1
16	FOS	1	1	1	1	1
17	CXCL8	1	1	1	1	1
18	MTOR	1	1	1	1	1
19	POMC	1	1	1	1	1
20	GRIN2B	1	1	1	1	1
21	HCRT	1	1	1	1	1
22	GRM5	1	1	1	1	1
23	GRM1	1	1	1	1	1

Fonte: O Autor (2021)

Figura 23 – Arquivo resultante da operação de diferença simétrica

	Padrão	Padrão	Padrão	Padrão	Padrão	Padrão
1	Name	File 1	File 2	File 3	File 4	File 5
2	cCYP2C19	1	0	0	0	0
3	ccIL6	1	0	0	0	0
4	cIL6	1	0	0	0	0
5	aCYP2C19	0	1	0	0	0
6	aABCb1	0	1	0	0	0
7	aIL6	0	1	0	0	0
8	bCYP2C19	0	0	1	0	0
9	bMAOb	0	0	1	0	0
10	bABCb1	0	0	1	0	0
11	bIL6	0	0	1	0	0
12	bCREB1	0	0	1	0	0
13	bADIPOQ	0	0	1	0	0
14	bTNF	0	0	1	0	0
15	bIL10	0	0	1	0	0
16	NTRK1	0	0	0	1	0
17	ITGB3	0	0	0	1	0
18	TACR1	0	0	0	1	0
19	ARHGEF25	0	0	0	1	0
20	UGT1A4	0	0	0	1	0
21	NPY5R	0	0	0	1	0

Fonte: O Autor (2021)

Figura 24 – Arquivo resultante da operação seleção

	Padrão	Padrão	Padrão	Padrão	Padrão	Padrão
1	Name	File 1	File 2	File 3	File 4	File 5
2	cCYP2C19	1	0	1	0	0
3	ccIL6	1	0	0	0	1
4	cIL6	1	0	0	1	0

Fonte: O Autor (2021)

Desse modo, as informações permanecem disponíveis mesmo que o programa seja encerrado e os resultados salvos nos arquivos de texto podem ser consultados a qualquer momento

ou importados para outros programas. Como o Cytoscape permite filtrar os nodos desejados, caso seja necessário, será possível informar manualmente ao software os nós que resultaram dessas operações e visualizar as redes resultantes graficamente.

4 CONSIDERAÇÕES FINAIS

Com o intuito de poder demonstrar todo o conteúdo que foi trabalhado e apresentar resumidamente o conhecimento adquirido com o presente trabalho, no capítulo 4 são apresentadas algumas considerações que visam falar sobre pontos importantes visualizados no decorrer do desenvolvimento.

4.1 CONCLUSÕES

A Biologia de sistemas é uma área interdisciplinar crescente devido a importância em análise de dados em larga escala. Um ramo muito importante dessa área é o estudo das interações entre as proteínas que podem ser representados por grafos através de redes de proteínas. Para simplificar essas interações, pode-se considerar os nós dessas redes como os elementos de um conjunto de proteínas. A partir disso é possível realizar operações entre esses conjuntos buscando similaridades e diferenças entre as proteínas presentes em cada rede.

No presente trabalho foi desenvolvida uma ferramenta que, a partir de redes de proteínas extraídas do Cytoscape e exportadas para arquivos no formato CSV, torna possível a comparação de dois ou mais arquivos ao mesmo tempo para realizar as operações de união, intersecção e diferença simétrica. Para chegar a esses resultados foi necessário analisar as redes de proteínas a serem manipuladas e escolher uma forma adequada de representação que permitisse a execução das operações solicitadas. Essa escolha exigiu definir a linguagem de programação a ser utilizada (C#) e selecionar a estrutura de dados mais apropriada entre as possíveis na linguagem para representar os grafos. Essa estrutura de dados (dicionário) foi customizada de acordo com os dados que teriam que ser apresentados e manipulados pelas operações e para executar essas operações, foi necessário desenvolver os algoritmos que trabalhassem corretamente sobre as estruturas de dados utilizadas.

Essa organização tornou possível ler cada um dos arquivos, que representam conjuntos de proteínas diferentes, com interações diferentes. Esses conjuntos foram armazenados em uma estrutura de dados única e comandos que percorrem essa estrutura foram aplicados construindo novos conjuntos de proteínas com os resultados das operações de união, intersecção, diferença simétrica e da seleção dos arquivos, sendo que essas últimas duas o Cytoscape não realiza.

Após a implementação e testes realizados com diferentes quantidades de arquivos, o objetivo principal foi alcançado gerando informações úteis para contribuir com uma área tão importante e necessária que é a Biologia Computacional.

4.2 TRABALHOS FUTUROS

Para uma possível continuação do trabalho apresentado pode-se pensar na inclusão de outras operações a serem realizadas com proteínas de várias redes, importação/exportação direta de arquivos para outros programas e aperfeiçoamento da interface da ferramenta desenvolvida nesse trabalho. Outra melhoria que pode ser feita é a implementação do software em outras plataformas como, por exemplo, Linux e a criação de um aplicativo para dispositivos móveis ou a implementação em linguagem web.

REFERÊNCIAS

- BARABASI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. **Nature reviews genetics**, Nature Publishing Group, v. 5, n. 2, p. 101–113, 2004.
- BARABASI, A.-L. *et al.* **Network science**. [S.l.]: Cambridge university press, 2016.
- BARBOSA, T. A. P.; FELISBERTO, A. D. dos S.; HESPANHOL, J. M. de M. **Biologia sistêmica e ensino de ciências: Um novo paradigma ou neo-positivismo?** 2017.
- DAMAS, L. **Linguagem C**. FCA – Editora de Informática, 2007. Disponível em: <https://www.ams.eti.br/livros/Luis_Damas.pdf>.
- GOLDBARG, M.; GOLDBARG, E. **Grafos**. [S.l.]: Elsevier Brasil, 2012.
- HEBERLE, H. **Uma abordagem visual para análise comparativa de redes biomoleculares com apoio de diagramas de Venn**. Tese (Doutorado) — Universidade de São Paulo, 2015.
- MEHLHORN, K.; SANDERS, P. **Algorithms and data structures: The basic toolbox**. [S.l.]: Springer Science & Business Media, 2008.
- MENEZES, P. B.; MENEZES, P. B. **Matemática discreta para computação e informática**. [S.l.]: Bookman, 2010. v. 2.
- ORTIZ-JOYA, L.; CONTRERAS-RODRÍGUEZ, L. E.; RAMÍREZ-HERNÁNDEZ, M. H. Protein-protein interactions of the nicotinamide/nicotinate mononucleotide adenylyltransferase of *Leishmania braziliensis*. **Memórias do Instituto Oswaldo Cruz**, scielo, v. 114, 00 2019. ISSN 0074-0276. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0074-02762019000100316&nrm=iso>.
- PIRES, M. T. **Guia de dados abertos**. São Paulo, 2015. Disponível em: <https://ceweb.br/media/docs/publicacoes/13/Guia_Dados_Abertos.pdf>.
- ROSEN, K. H. **Matemática discreta e suas aplicações**. [S.l.]: Grupo A Educação, 2009.
- SCHEINERMAN, E. R. **Matemática Discreta - Uma introdução**. [S.l.]: Cengage Learning Editores, 2003.
- SILVA, S. de Avila e; NOTARI, D. L.; DALL'ALBA, G. **Bioinformática: contexto computacional e aplicações**. [S.l.]: EDUCS, 2020.
- STERTZ, K. **Modelagem e simulação em biologia computacional e bioinformática. Salão de iniciação Científica (16.: 2004: Porto Alegre, RS). Livro de resumos. Porto Alegre: UFRGS, 2004.**, 2004.
- SZWARCFITER, J. L. **Teoria computacional de grafos: Os Algoritmos**. [S.l.]: Elsevier Brasil, 2018.