

# Análise de Redes Complexas

Barbara Reis dos Santos e Mayara Lessnau de Figueiredo Neves

CI1054 - Algoritmos em Grafos e Otimização, 2º Semestre de 2024

## 1 Introdução

O objetivo deste trabalho é realizar uma análise detalhada de uma rede complexa utilizando a biblioteca *NetworkX* e a linguagem Python. A análise inclui o cálculo de diversas métricas estruturais da rede, como componentes conexas, raio, diâmetro, distribuição de graus, centralidade dos vértices e cliques. Para que com isso possamos identificar o modelo de grafo que melhor representa a rede estudada.

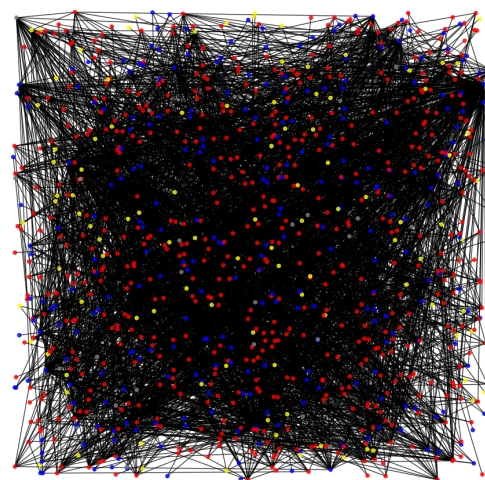


Figura 1: Visualização da rede completa gerada pelo código. Os nós estão coloridos para indicar diferentes tipos: drogas (D) em **vermelho** e os alvos genéticos (P) em **azul**, (Q) em **amarelo** e (O) em **cinza**.

## 2 Descrição da Rede

A rede escolhida para este trabalho é a *Drug-target interaction network*[1], que representa interações entre drogas disponíveis no mercado norte-americano e seus alvos moleculares, geralmente proteínas codificadas por genes. Essa rede é amplamente utilizada em pesquisas farmacêuticas, facilitando a descoberta de novos alvos de medicamentos, o design de drogas, a previsão de metabolismo e interações medicamentosas, entre outros.

A *Drug-target interaction network* modela as interações entre drogas e alvos genéticos em forma de um grafo bipartido. Os vértices da rede são divididos em dois grupos principais: vértices representando drogas (fármacos) identificados por rótulos que começam com a letra D, e vértices representando alvos biológicos, indicados pelas letras P, Q ou O. Na Figura 1 é possível visualizar a rede e sua complexidade.

As arestas da rede representam interações conhecidas entre uma droga e seu respectivo alvo. Essas interações refletem dados experimentais ou preditivos, amplamente utilizados em estudos de biologia computacional e farmacologia. A escolha de representar a rede como um grafo possibilita analisar propriedades estruturais relevantes para a identificação de padrões, hubs (nós altamente conectados), e agrupamentos que podem ter implicações farmacológicas.

Esse modelo de rede é particularmente útil para explorar relações entre drogas e múltiplos alvos, bem como para identificar alvos compartilhados por diferentes medicamentos.

A rede *Drug-target interaction network* foi obtida a partir do site <https://snap>.

stanford.edu[1], uma base de dados mantida pelo grupo *Stanford Network Analysis Project (SNAP)* da Universidade de Stanford. Essa base de dados é amplamente reconhecida por fornecer conjuntos de dados para análise de redes em diversas áreas, como biologia, computação, ciências sociais e engenharia. Os dados disponibilizados incluem tanto a estrutura da rede em formato de lista de arestas quanto metadados detalhados sobre a origem e relevância do conjunto.

## 3 Métodos

Nesta seção, descrevemos os métodos utilizados para a análise da rede. Dividimos a descrição em dois aspectos principais: as métricas analisadas e os códigos desenvolvidos para realizar os testes.

### 3.1 Métricas Analisadas

As seguintes métricas foram calculadas para compreender as propriedades estruturais da rede:

#### 3.1.1 Componentes Conectados

Identificamos o número de componentes conectados na rede e o tamanho do maior componente.

#### 3.1.2 Raio e Diâmetro

Calculamos o raio (menor distância máxima) e o diâmetro (maior distância mínima) para redes conectadas.

#### 3.1.3 Centralidades

Analisamos centralidades importantes, como:

- Centralidade de grau
- Centralidade de intermediação (betweenness)
- Centralidade de proximidade (closeness)

#### 3.1.4 Distribuição de Graus

Estudamos a distribuição de graus para entender a conectividade dos nós na rede.

#### 3.1.5 Ciclos e Cliques

Identificamos a quantidade de ciclos e cliques da rede. Além disso, avaliamos o tamanho das cliques encontradas.

### 3.2 Códigos Utilizados para Teste

A implementação desenvolvida para a análise da rede utiliza amplamente a biblioteca *NetworkX* [2], que fornece diversas funções para manipulação e análise de grafos. Além disso, a biblioteca *Matplotlib* [3] foi empregada para a visualização de dados e métricas e a biblioteca *Argparse* [4] para receber arquivos diretamente pelo terminal. Abaixo, destacamos as principais funções utilizadas e seus respectivos papéis no código:

#### 3.2.1 Manipulação de Grafos com *NetworkX*

- **`nx.read_edgelist`**: Essa função foi usada para carregar o grafo a partir de um arquivo no formato de lista de arestas. Cada linha do arquivo representa uma interação entre dois nós, e a função transforma esses dados em um grafo manipulável.
- **`nx.number_connected_components` e `nx.connected_components`**: As componentes conexas da rede foram identificadas com essas funções. Enquanto a primeira retorna o número total de componentes conexas, a segunda gera conjuntos contendo os nós de cada componente, permitindo análises detalhadas da maior componente.
- **`nx.subgraph`**: Essa função foi usada para extrair subgrafos, como a maior componente conexa, a fim de calcular métricas específicas, como raio e

diâmetro, que só fazem sentido em componentes conexas.

- **`nx.find_cliques`**: Para identificar grupos com forte conexão entre si, utilizamos esta função, que retorna todos os cliques (subgrupos de nós que são grafos completos - todos os vértices se conectam com todos os outros do grupo).
- **`nx.cycle_basis`**: Foi empregada para calcular uma base de ciclos elementares na maior componente conexa, permitindo avaliar a redundância estrutural da rede.

### 3.2.2 Cálculo de Métricas com *NetworkX*

- **`nx.radius` e `nx.diameter`**: Ambas as funções foram aplicadas à maior componente conectada para calcular o raio (menor distância máxima de um nó a qualquer outro) e o diâmetro (a maior distância mínima entre dois nós), métricas que fornecem informações sobre a conectividade global da rede.
- **`nx.degree_centrality`, `nx.closeness_centrality` e `nx.betweenness_centrality`**: Essas funções calcularam, respectivamente, as centralidades de grau, proximidade e intermediação de cada nó, identificando hubs e pontos críticos na rede.

### 3.2.3 Visualização com *NetworkX* e *Matplotlib*

- **`nx.draw`**: A função `nx.draw` foi usada para visualizar o grafo de forma básica, fornecendo uma representação gráfica da rede. Em combinação com layouts como o **`nx.spring_layout`**, os vértices e arestas foram organizados em um estilo baseado em forças, garantindo uma distribuição visual clara e compreensível.
- ***Matplotlib***: Além de auxiliar na visualização do grafo, a biblioteca

*Matplotlib* foi utilizada para gerar histogramas representando a distribuição de graus da rede. A funcionalidade de ajuste de eixos foi aplicada para garantir uma visualização mais informativa, maximizando o tamanho dos gráficos e evidenciando a dispersão dos dados.

### 3.2.4 Integração entre Ferramentas

A integração entre *NetworkX* e *Matplotlib* foi essencial para a implementação do código. Enquanto o *NetworkX* gerencia a manipulação e análise do grafo, o *Matplotlib* complementa o processo, permitindo a apresentação visual de dados e métricas, facilitando a interpretação e comunicação dos resultados.

### 3.2.5 Mais Detalhes

Mais informações sobre os códigos utilizados para esta análise incluindo a descrição completa das implementações, estão disponíveis no repositório público do GitHub: <https://github.com/brareisds/analise-redes-metabolicas.git>[6].

## 4 Análise da Rede

### 4.1 Estrutura Geral da Rede

A rede *Drug-target interaction network* contém 7341 nós e 15138 arestas, conforme resumido na Tabela 1. A maior componente conexa da rede abrange 6621 nós, o que corresponde a aproximadamente 90% do total, evidenciando que a maioria dos fármacos e alvos biológicos está conectada de forma direta ou indireta. Além disso, a rede possui 204 componentes conexas menores, que representam interações mais isoladas. A Figura 2 apresenta a distribuição das componentes conexas, onde os clusters centrais destacam a maior componente, composta pelas interações mais densas e complexas, enquanto os clusters periféricos representam componentes menores e mais isoladas.

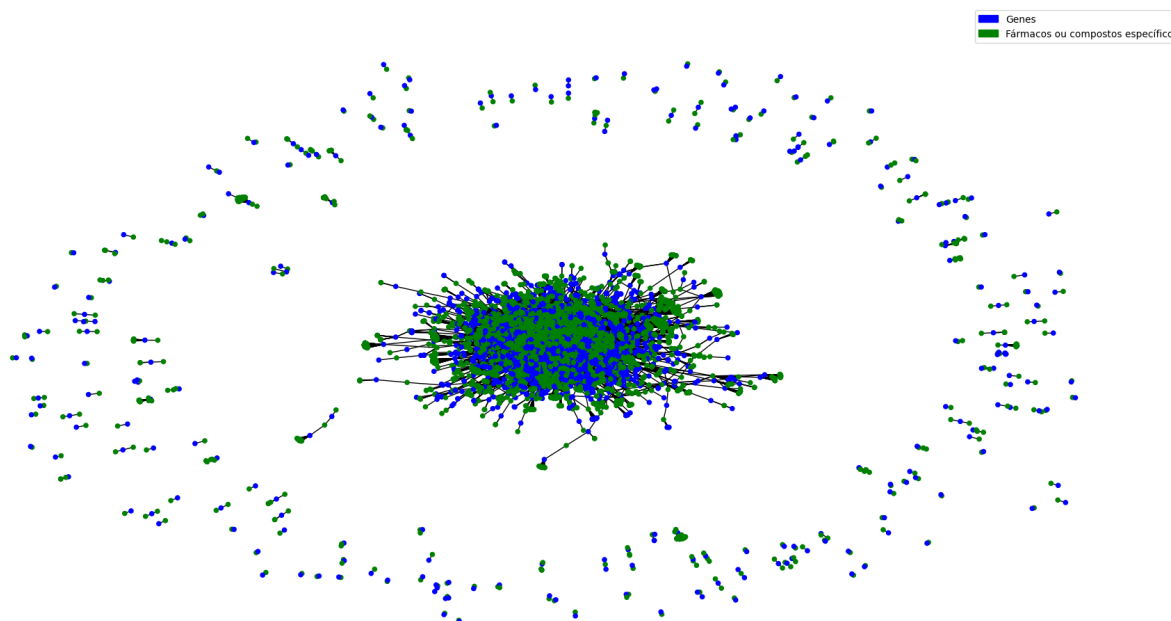


Figura 2: Visualização da rede completa gerada pelo código. Os nós estão coloridos de acordo com suas categorias: *fármacos e compostos químicos* são representados em verde, enquanto os *genes ou alvos biológicos* estão em azul. A estrutura central da figura evidencia a maior componente conexa da rede, caracterizada por interações densas e múltiplas conexões. Já as áreas periféricas representam componentes menores, que correspondem a interações específicas ou isoladas.

Tabela 1: Métricas Gerais da Rede *Drug-target interaction network*

Métrica	Valor
Quantidade de nós	7341
Quantidade de arestas	15138
Quantidade de componentes conexas	205
Tamanho da maior componente conexa	6621
Raio da maior componente conexa	10
Diâmetro da maior componente conexa	18
Quantidade de ciclos	7961
Número de cliques encontrados	15138
Tamanho máximo de clique	2

## 4.2 Raio e Diâmetro

Na maior componente conexa, o raio da rede é 10, enquanto o diâmetro é 18. Isso significa que a distância máxima entre dois nós quaisquer da maior componente é 18 arestas, enquanto o nó mais central tem uma distância máxima de 10 arestas até qualquer outro nó. Esses valores sugerem que a rede possui uma estrutura compacta e eficiente, caracterizada por caminhos relativamente curtos entre os vértices, uma propriedade típica de redes do tipo *pequeno mundo*.

## 4.3 Ciclos e Cliques

Foram identificados 7961 ciclos na maior componente conexa, indicando redundância e robustez estrutural. Ciclos são importantes em redes biológicas, pois podem representar redundâncias funcionais ou múltiplas vias de interação entre drogas e alvos. Quanto aos cliques, foram encontrados 15138, todos de tamanho 2, o que é consistente com a natureza bipartida da rede, onde drogas e alvos se conectam, mas não há conexões diretas entre vértices do mesmo tipo.

## 4.4 Centralidade de Vértices

Os nós com maior centralidade de grau, proximidade e intermediação foram identificados. Esses nós são considerados hubs ou pontos críticos na rede:

- **Centralidade de Grau:** O nó P08684 apresentou a maior centralidade de grau (0.0796), seguido por P10635 (0.0439) e P11712 (0.0392). Esses nós são hubs diretamente conectados a muitos outros vértices, sendo cruciais para a conectividade da rede.

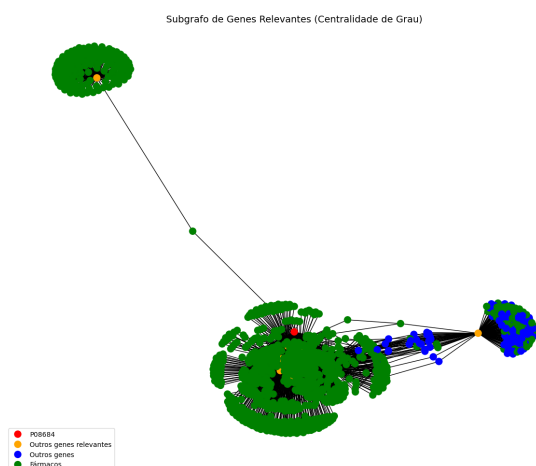


Figura 3: Subgrafo destacando os nós com maior centralidade de grau. O nó em **vermelho** representa o gene P08684, os nós em **laranja** representam outros genes relevantes, os nós em **azul** são genes gerais, e os nós em **verde** representam os fármacos.

- **Centralidade de Proximidade:** O nó P08684 também lidera em centralidade de proximidade (0.2508), seguido por P11712 (0.2313) e P05177 (0.2299). Esses vértices estão mais próximos da maioria dos outros na rede, facilitando a propagação de efeitos ou informações.

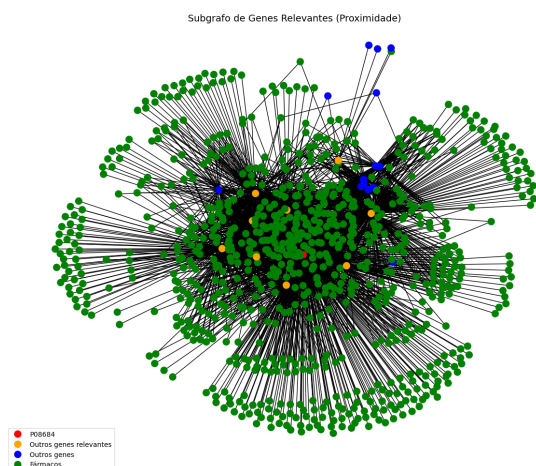


Figura 4: Subgrafo destacando os nós com maior centralidade de proximidade. O nó em **vermelho** representa o gene P08684, os nós em **laranja** representam outros genes relevantes, os nós em **azul** são genes gerais, e os nós em **verde** representam os fármacos.

- **Centralidade de Intermediação:** Novamente, P08684 aparece como o mais influente (0.2371), seguido por P05177 (0.0663) e P10635 (0.0609). Esses nós desempenham papéis críticos em conectar diferentes partes da rede.

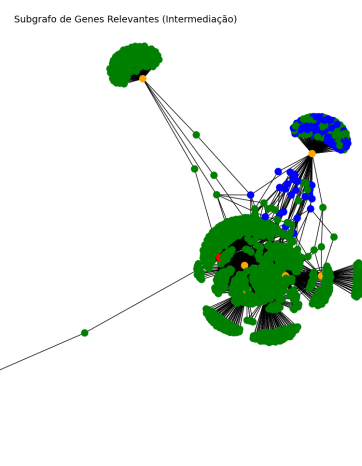


Figura 5: Subgrafo destacando os nós com maior centralidade de intermediação. O nó em **vermelho** representa o gene P08684, os nós em **laranja** representam outros genes relevantes, os nós em **azul** são genes gerais, e os nós em **verde** representam os fármacos.

Os genes ocupam posições de destaque na rede, evidenciadas pelos altos valores de centralidade apresentados por nós como P08684, P10635 e P11712. Destaca-se o gene P08684, que consistentemente aparece como o mais relevante em todas as métricas analisadas (grau, proximidade e intermediação). Essa centralidade reflete a natureza da rede, na qual múltiplos medicamentos compartilham os mesmos alvos genéticos, tornando os genes pontos de convergência essenciais. Eles desempenham um papel crucial na arquitetura da rede, conectando diferentes fármacos e representando alvos terapêuticos altamente explorados na farmacologia.

As figuras exemplificam claramente a importância dos genes identificados como relevantes dentro do contexto de cada métrica. Sua posição na rede evidencia o papel vital que desempenham na conexão e interação com múltiplos medicamentos. Esses genes são fundamentais não apenas para compreender as dinâmicas da rede, mas também para análises biológicas mais amplas, reforçando sua importância no estudo de interações medicamentosas e no desenvolvimento de estratégias terapêuticas.

## 4.5 Distribuição de Graus

A distribuição de graus da rede revela uma característica importante: a maioria dos

nós possui poucas conexões, enquanto um pequeno número de nós concentra um alto número de conexões. Podemos visualizar isso melhor na Figura 6. Esse padrão é típico de redes complexas e indica que a rede segue uma tendência próxima à de uma lei de potência.

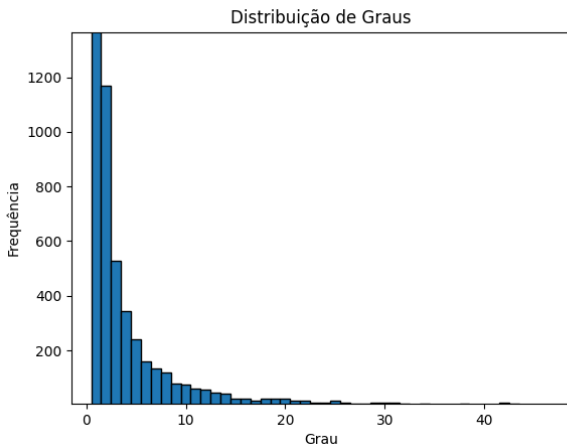


Figura 6: Gráfico que mostra a distribuição de graus da rede. No eixo y temos a frequência com que o grau aparece e no eixo x temos o valor do grau.

Em redes biológicas e farmacológicas, essa característica sugere a presença de *hubs*, ou seja, vértices altamente conectados que desempenham papéis centrais na conectividade e funcionalidade da rede. Esses hubs, como o vértice P08684, que apresenta alta centralidade de grau, são críticos para a estrutura da rede. Sua remoção pode fragmentar a rede ou reduzir significativamente sua eficiência. Por outro lado, a presença de muitos nós com poucos graus contribui para a robustez geral, tornando a rede resiliente a falhas aleatórias.

Esse comportamento aproxima a rede de um modelo de grafo livre de escala (*scale-free network*), onde a distribuição de graus segue aproximadamente uma lei de potência, caracterizada pela equação:

$$P(k) \sim k^{-\gamma}, \quad (1)$$

em que  $P(k)$  é a probabilidade de um nó ter grau  $k$ , e  $\gamma$  é o expoente da lei de potência. Apesar de a rede não ser perfeitamente ajustada a esse modelo, a presença de hubs e a disparidade na distribuição de

graus são consistentes com as propriedades observadas em redes biológicas.

## 5 Comparação com Modelos de Redes

Para avaliar qual modelo teórico melhor representa a *Drug-target interaction network*, foram selecionados quatro modelos para comparação: **Erdős-Rényi**, **Barabási-Albert**, **Watts-Strogatz** e **Regular**. As métricas analisadas para cada modelo incluem o *clustering coefficient* (coeficiente de agrupamento), o caminho médio na maior componente conexa (*average shortest path length in LCC*) e o grau médio (*average degree*). A Tabela 2 apresenta os resultados obtidos da comparação entre eles e a rede original que está sendo avaliada neste trabalho.

### 5.1 Construção dos Modelos de Redes

Nesta seção, explicamos como cada modelo foi construído e configurado para refletir as características do grafo original:

#### 5.1.1 Erdős-Rényi (ER)

O modelo de Erdős-Rényi cria grafos aleatórios onde cada par de nós tem uma probabilidade fixa  $p$  de serem conectados. Este modelo é definido por dois parâmetros principais:

- $n$ : número de nós (igual ao número de nós do grafo original);
- $p$ : probabilidade de conexão entre pares de nós, calculada como:

$$p = \frac{\text{Número de Arestas}}{\text{Número Total de Possíveis Arestas}} = \frac{E}{N \cdot (N - 1)}. \quad (2)$$

Este modelo é útil para representar redes completamente aleatórias, sem hubs ou agrupamentos significativos. Apesar de sua simplicidade, é eficiente para entender propriedades gerais de conectividade e comportamento estatístico de redes aleatórias.



Tabela 2: Comparação entre a rede original e modelos teóricos de redes.

Rede/Modelo	Coef. Agrup.	Caminho Médio (LCC)	Grau Médio
Original	0.000	6.153	4.123
Erdős-Rényi	0.000281	11.456	2.080
Barabási-Albert	0.005687	4.905	3.999
Watts-Strogatz	0.375369	12.246	4.000
Regular	0.000409	7.446	4.000

### 5.1.2 Barabási-Albert (BA)

O modelo de Barabási-Albert é baseado no mecanismo de *ligação preferencial*, no qual novos nós são mais propensos a se conectar a nós que já possuem um grande número de conexões (hubs). Este modelo é amplamente utilizado para representar redes com distribuição de graus em lei de potência (*scale-free networks*), característica comum em redes biológicas e sociais.

Para construir o grafo Barabási-Albert:

- O número de nós  $n$  é o mesmo do grafo original;
- O parâmetro  $m$  (número médio de arestas adicionadas por nó) foi definido como:

$$m = \max(1, \text{round}(\frac{\text{Número de Arestas}}{\text{Número de Nós}})).$$

Esse modelo captura bem a existência de hubs, característica observada na rede original, onde alguns genes (nós) possuem um número desproporcionalmente alto de conexões.

### 5.1.3 Watts-Strogatz (WS)

O modelo de Watts-Strogatz gera redes do tipo *pequeno mundo* (*small-world networks*), caracterizadas por:

- Alto coeficiente de agrupamento (*clustering coefficient*);
- Baixo caminho médio (distância média entre pares de nós).

Para construir o grafo Watts-Strogatz:

- O número de nós  $n$  é o mesmo do grafo original;

- O grau médio  $k$  (número de conexões por nó) é definido como:

$$k = \max(2, \text{round}(\frac{2 \times \text{Número de Arestas}}{\text{Número de Nós}})).$$

Garantimos que  $k$  seja par, como exigido pelo modelo;

- A probabilidade de reconexão  $p$  foi fixada em 0.1, permitindo a introdução de atalhos na rede regular inicial, reduzindo o caminho médio.

O modelo Watts-Strogatz é particularmente útil para redes onde o clustering é significativo. No entanto, no caso da *Drug-target interaction network*, o coeficiente de agrupamento é baixo, limitando a representatividade deste modelo.

### 5.1.4 Regular

O modelo Regular constrói um grafo onde todos os nós possuem exatamente o mesmo grau, distribuído uniformemente. Este modelo é útil para explorar redes homogêneas, onde não há hubs nem variação significativa no número de conexões. Ele é definido pelos seguintes parâmetros:

- $n$ : número de nós (igual ao número de nós do grafo original);
- $k$ : grau uniforme para todos os nós, definido como:

$$k = \text{round}(\frac{2 \times \text{Número de Arestas}}{\text{Número de Nós}}).$$

Caso  $k$  não seja um número par, ele é ajustado para satisfazer a restrição do modelo Regular.

No modelo Regular, todos os nós possuem o mesmo número de conexões, resultando em um coeficiente de agrupamento muito baixo. Apesar disso, o caminho médio tende a ser relativamente curto. No entanto, esse modelo não captura a heterogeneidade observada em redes biológicas, como a *Drug-target interaction network*, onde hubs desempenham um papel fundamental.

## 6 Resultados

### 6.1 Análise da Rede

A análise da rede resumida pela Tabela 1 revela que a rede é dominada por uma grande componente conexa, apresentando alta redundância devido ao número significativo de ciclos. A ausência de cliques maiores que 2 confirma que a rede é bipartida, refletindo a estrutura subjacente de interações droga-alvo. O nó P08684 surge como um hub central em todas as métricas de centralidade, indicando sua importância funcional e estrutural. A robustez e a conectividade geral da rede sugerem que ela pode resistir a falhas aleatórias, mas é sensível à remoção de hubs.

Esses resultados são relevantes para a biologia e farmacologia, destacando potenciais alvos terapêuticos e possíveis caminhos para o reposicionamento de medicamentos. A estrutura da rede pode também ser utilizada para prever interações medicamentosas e explorar sinergias entre drogas.

### 6.2 Análise dos Modelos

Com base nos resultados apresentados na Tabela 2, podemos observar as seguintes características de cada modelo:

- **Original:** A rede original apresenta um coeficiente de agrupamento extremamente baixo (0.000), indicando que não há triângulos na rede, e um caminho médio relativamente curto (6.153), o que sugere uma boa conectividade global. O grau médio é 4.123, próximo de uma rede esparsa.
- **Erdős-Rényi:** Este modelo gera redes completamente aleatórias, com coeficiente de agrupamento 0.000281, próximo do valor da rede original, mas apresenta um caminho médio significativamente maior (11.456), indicando baixa eficiência na conectividade. O grau médio (2.080) é muito menor que o da rede original, dificultando sua representatividade.
- **Barabási-Albert:** Este modelo, baseado no princípio de ligação preferencial, apresenta um coeficiente de agrupamento maior que o da rede original (0.005687), mas ainda baixo. O caminho médio é 4.905, o mais próximo da rede original, e o grau médio 3.999 também é bastante similar. Este modelo reflete bem a presença de hubs na rede original.
- **Watts-Strogatz:** O modelo *small-world* apresenta um coeficiente de agrupamento muito alto (0.375369) e um caminho médio (12.246) muito maior que o da rede original, indicando que não é adequado para representar redes com agrupamento baixo como a *Drug-target interaction network*. O grau médio (4.000) é compatível, mas não é suficiente para justificar sua representatividade.
- **Regular:** Este modelo, onde todos os nós possuem o mesmo grau médio (4.000), apresenta um coeficiente de agrupamento muito baixo (0.000409) e um caminho médio de 7.446, mais próximo da rede original em relação ao Erdős-Rényi e Watts-Strogatz, mas ainda distante. A ausência de variação nos graus dos nós limita sua aplicabilidade para redes biológicas, onde a conectividade é altamente heterogênea.

### 6.3 Modelo Melhor Representado

Com base nos resultados, o modelo **Barabási-Albert** é o que melhor representa a



*Drug-target interaction network*. Ele apresenta:

- Um coeficiente de agrupamento (0.005687) baixo, mas ainda compatível com a rede original (0.000);
- O caminho médio mais próximo (4.905) do da rede original (6.153);
- Grau médio (3.999) muito similar ao da rede original (4.123).

Embora outros modelos apresentem algumas métricas próximas da rede original, nenhum deles reflete tão bem a conectividade geral, a distribuição de graus e a estrutura da rede como o modelo Barabási-Albert.

## 7 Conclusão

Concluimos que o modelo **Barabási-Albert** é o mais adequado para representar a *Drug-target interaction network*, capturando propriedades como a presença de hubs, caminhos médios curtos e grau médio compatível com a rede real. A análise revelou que o gene *P08684* [5], que codifica a enzima *Cytochrome P450 3A4* (CYP3A4), desempenha um papel central na rede, refletindo sua importância biológica e farmacológica. Como hub central, este gene conecta várias interações droga-alvo, sendo crítico para a eficácia terapêutica e prevenção de interações medicamentosas adversas. Sua posição na rede reforça como hubs podem ser usados para identificar alvos prioritários e orientar o desenvolvimento de novas terapias.

## Referências

- [1] Marinka Žitnik, Rok Sosič, Sagar Maheshwari, and Jure Leskovec. *BioSNAP Datasets: Stanford Biomedical Network Dataset Collection*. Disponível em: <http://snap.stanford.edu/biodata>. Agosto de 2018.
- [2] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. “Exploring network structure, dynamics, and function using NetworkX.” In: *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds.), Pasadena, CA USA, pp. 11–15, Aug 2008.
- [3] J. D. Hunter. “Matplotlib: A 2D Graphics Environment”. *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
- [4] Python Software Foundation. “argparse — Parser for command-line options, arguments, and sub-commands”. Disponível em: <https://docs.python.org/3/library/argparse.html>. Acesso em: 4 dez. 2024.
- [5] UniProt. *Cytochrome P450 3A4 (CYP3A4)*. Disponível em: <https://www.uniprot.org/uniprotkb/P08684/entry>. Acesso em: 4 dez. 2024.
- [6] Reis dos Santos, B.; Lessnau de Figueiredo Neves, M. *Análise de Redes Metabólicas - Repositório no GitHub*. Disponível em: <https://github.com/brareisds/analise-redes-metabolicas.git>. Acesso em: 8 dez. 2024.