UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS NATURAIS E EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

Augusto Loose Bellinaso

UM MODELO PARA A EVOLUÇÃO DE REDES DE PROTEÍNAS CONTEMPLANDO A TRANSFERÊNCIA HORIZONTAL DE GENES

Santa Maria, RS
2022

Augusto Loose Bellinaso

UM MODELO PARA A EVOLUÇÃO DE REDES DE PROTEÍNAS CONTEMPLANDO A TRANSFERÊNCIA HORIZONTAL DE GENES

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Física, Área de Concentração em Física computacional e sistemas complexos, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Mestre em Física**. Defesa realizada por videoconferência.

ORIENTADOR: Prof. José Carlos Merino Mombach

Santa Maria, RS
2022

This study was financied in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001

Bellinaso, Augusto Loose
UM MODELO PARA A EVOLUÇÃO DE REDES DE PROTEÍNAS CONTEMPLANDO A TRANSFERÊNCIA HORIZONTAL DE GENES / Augusto Loose Bellinaso.- 2022.
95 p.; 30 cm

Orientador: José Carlos Merino Mombach
Dissertação (mestrado) - Universidade Federal de Santa Maria, Centro de Ciências Naturais e Exatas, Programa de Pós-Graduação em Física, RS, 2022

1. Física computacional 2. Sistemas complexos 3. Teoria de Grafos 4. Biofísica 5. Redes complexas I. Mombach, José Carlos Merino II. Título.

Sistema de geração automática de ficha catalográfica da UFSM. Dados fornecidos pelo autor(a). Sob supervisão da Direção da Divisão de Processos Técnicos da Biblioteca Central. Bibliotecária responsável Paula Schoenfeldt Patta CRB 10/1728.

Declaro, AUGUSTO LOOSE BELLINASO, para os devidos fins e sob as penas da lei, que a pesquisa constante neste trabalho de conclusão de curso (Dissertação) foi por mim elaborada e que as informações necessárias objeto de consulta em literatura e outras fontes estão devidamente referenciadas. Declaro, ainda, que este trabalho ou parte dele não foi apresentado anteriormente para obtenção de qualquer outro grau acadêmico, estando ciente de que a inveracidade da presente declaração poderá resultar na anulação da titulação pela Universidade, entre outras consequências legais.

Augusto Loose Bellinaso

UM MODELO PARA A EVOLUÇÃO DE REDES DE PROTEÍNAS CONTEMPLANDO A TRANSFERÊNCIA HORIZONTAL DE GENES

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Física, Área de Concentração em Física computacional e sistemas complexos, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Mestre em Física**.

Aprovado em 31 de março de 2022:

_____

José Carlos Merino Mombach, Dr. (UFSM)
(Presidente/Orientador)

_____

Mateus Schmidt, Dr. (UFSM) (videoconferência)

_____

Fernando Lucas Metz, Dr. (UFRGS) (videoconferência)

Santa Maria, RS
2022

NUP: 23081.031315/2022-65 **Prioridade:** Normal

Homologação de ata de banca de defesa de pós-graduação

134.332 - Bancas examinadoras: indicação e atuação

COMPONENTE

Ordem	Descrição	Nome do arquivo
1	Ata de defesa de dissertação/tese (134.332)	Ata Augusto APROVADO.pdf

Assinaturas

31/03/2022 16:25:37

JOSE CARLOS MERINO MOMBACH (PROFESSOR DO MAGISTÉRIO SUPERIOR)

02.32.00.00.0.0 - DEPARTAMENTO DE FÍSICA - DFSC

31/03/2022 16:40:34

MATEUS SCHMIDT (PROFESSOR DO MAGISTÉRIO SUPERIOR)

02.32.00.00.0.0 - DEPARTAMENTO DE FÍSICA - DFSC

31/03/2022 16:53:14

FERNANDO LUCAS METZ (Pessoa Física)

Usuário Externo (805.***.***-**)

Universidade Federal de Santa Maria 1960

Código Verificador: 1281954

Código CRC: 2816f24

Consulte em: https://portal.ufsm.br/documentos/publico/autenticacao/assinaturas.html

AGRADECIMENTOS

Agradeço aos meus pais, Ademar e Lisiane, por todo o suporte, todo amor, carinho, incentivo e por sempre terem me mostrado o valor e o poder da educação. Agradeço profundamente pelos sacríficios que precisaram fazer para que pudesse chegar aqui e finalizar mais um ciclo.

Agradeço à minha irmã, Susana, pela paciência e pela ajuda em várias situações para que o presente trabalho pudesse ser feito e concluído com sucesso.

Agradeço à minha namorada, Nicole, por estar ao meu lado, me incentivando e me cobrando, e me auxiliando sempre que fosse preciso.

Agradeço aos amigos e colegas, pelas inúmeras conversas e trocas de ideias e pelos bons momentos de descontração que tivemos.

Agradeço ao meu orientador, professor José Carlos Mombach, pela oportunidade de poder realizar esse trabalho e pela ajuda sempre que precisei ao longo desses vários anos de trabalho em conjunto.

Agradeço aos professores com os quais tive contato ao longo do mestrado, por fornecerem uma base sólida de conhecimentos que foram fundamentais para esse trabalho e também para a vida.

Agradeço à CAPES, por ter me agraciado com uma bolsa de estudos para a realização desse trabalho.

Agradeço a todos os demais que, de alguma forma contribuíram para que o presente trabalho pudesse ser realizado e concluído.

Be curious. Read widely. Try new things. What people call intelligence just boils down to curiosity.

(Aaron Swartz)

RESUMO

UM MODELO PARA A EVOLUÇÃO DE REDES DE PROTEÍNAS CONTEMPLANDO A TRANSFERÊNCIA HORIZONTAL DE GENES

AUTOR: Augusto Loose Bellinaso
ORIENTADOR: José Carlos Merino Mombach

A nossa volta encontramos diversos exemplos de sistemas complexos, que são sistemas compostos por um grande número de elementos que interagem entre si de maneiras bastante complexas. Esses sistemas vão desde a *World Wide Web* até redes de distribuição de energia e redes de interações entre proteínas, que é o objeto de estudo do presente trabalho. É possível representar as redes de interações entre proteínas, bem como os demais sistemas complexos, através de um conjunto de nós e *links*, formando um grafo a partir do qual poderemos extrair as propriedades de interesse dessas redes. Dentre as mais variadas propriedades das redes, o presente trabalho foca em tentar entender como ocorreram os processos evolutivos que deram origem a essas redes da maneira como as conhecemos hoje. Assim, foram selecionadas algumas redes e a partir delas determinamos a topologia e o espectro de autovalores das mesmas, de modo que seja possível ter ferramentas de comparação. Após a escolha das redes, buscamos na literatura modelos que já foram desenvolvidos com o intuito de reproduzir tais redes, onde encontramos os modelos de Barabási e Albert (BA) e o modelo *Duplication-Divergence* (DD). Apesar de serem modelos bem conceituados, os mesmos não são capazes de reproduzir todas as propriedades relevantes das redes de interações entre proteínas, de modo que passou-se a buscar novas evidências para o desenvolvimento de um modelo que fosse mais preciso e produzisse resultados mais próximos do esperado. Alguns estudos e evidências sugerem que além da duplicação de genes, existe outro mecanismo por trás da evolução das redes de interações entre proteínas, e esse mecanismo é conhecido como Transferência Horizontal de genes (TH). No entanto, não foram encontradas evidências que indicassem como esse mecanismo realmente funciona e como poderia ser implementado para criar novas redes. Assim, partindo da hipótese de que as novas proteínas inseridas num organismo por meio da TH teriam uma maior preferência de se conectarem e interagirem com as proteínas com um grande número de interações já existentes na rede, optou-se por utilizar uma versão modificada do mecanismo do modelo BA para representar a TH. Dessa forma, desenvolveu-se um novo modelo para a evolução das redes de interações entre proteínas que engloba esses dois mecanismos: duplicação e transferência horizontal. Como poderá ser visto ao longo do trabalho, ao analisarmos a topologia, o espectro de autovalores e outras propriedades das redes geradas pelo novo modelo, veremos que esses resultados são mais próximos do que vemos para as redes reais quando comparado ao que é gerado pelo modelo DD, fortalecendo, assim, a ideia de que a duplicação não é o

único mecanismo por trás da evolução das redes de interações entre proteínas.

Palavras-chave: Redes de interações entre proteínas. Topologia. Espectro de autovalores. Duplicação. Transferência Horizontal.

ABSTRACT

A MODEL FOR PROTEIN INTERACTION NETWORKS EVOLUTION CONTEMPLATING HORIZONTAL GENE TRANSFER

AUTHOR: Augusto Loose Bellinaso
ADVISOR: José Carlos Merino Mombach

Around us we find several examples of complex systems, which are composed of a large number of elements that interact with each other in very complex ways. These systems range from the World Wide Web itself to energy distribution networks and protein interaction networks (PIN), which are the focus of this work. It is possible to represent the PINs, as well as other complex systems, through a set of nodes and links, forming a graph from which we can extract the properties of interest to study these networks. Among the wide variety of properties that we can use to study PINs, we focused in trying to understand the evolutionary processes that are responsible for originating the networks as we know them today. Thus, some networks were selected and from them we determined their topology and eigenvalue spectrum, in order to have comparison tools. After choosing the networks, we searched the literature for models that had already been developed in order to reproduce such networks, where we found the Barabási and Albert (BA) model and the Duplication-Divergence (DD) model. Despite being well-regarded models, they are not capable of reproducing all the relevant properties of the PINs, so that new evidence was sought for the development of a model that was more accurate and produced better results. Some studies suggests the existence of another mechanism that alongside gene duplication is responsible for the evolution of PINs: the horizontal gene transfer (HT). However, we were not able to find evidences about how this mechanism really works and how we can implement it. Thus, starting from the hypothesis that the new proteins added to an organism through HT will have a greater preference to connect and interact with the most connected proteins already existing in the network, we decided to use a modified version of the BA model to represent the HT mechanism. We then developed a new model for the evolution of PINs which is based in these two mechanisms: duplication and horizontal transfer. As we can see throughout this work, when we analyze the topology, eigenvalue spectrum and other properties of the networks generated with the new model, the results are closer to the real networks when comparing them to the results generated with the DD model, which supports and strengthen the idea that gene duplication is not the only mechanism behind the evolutionary proccess of protein interaction networks.

Keywords: Protein Interaction Networks. Topology. Eigenvalue Spectrum. Duplication. Horizontal Transfer.

LISTA DE FIGURAS

Figura 1.1 – Estrutura na forma de grafo para a rede de interações entre proteínas da *Caenorhabditis elegans* 16
Figura 1.2 – As sete pontes de Königsberg 17
Figura 1.3 – Desenho esquemático do problema 18
Figura 1.4 – Exemplos de grafos com suas respectivas distribuições de conectividades 20
Figura 1.5 – Representação de uma rede através de grafo e sua matriz de adjacência 21
Figura 1.6 – Coeficiente de clusterização local de diferentes grafos 22
Figura 1.7 – Três realizações de uma rede aleatória com os mesmos parâmetros 23
Figura 1.8 – Topologia de três redes aleatórias com $N = 100000$ nós e probabilidades p indicadas na imagem 24
Figura 1.9 – Comparação entre a topologia de três redes reais e a sua topologia caso fossem redes aleatórias 26
Figura 1.10 – Ilustração do modelo de Barabási e Albert para uma rede com $m_0 = 3$ e $m = 2$ 27
Figura 1.11 – Distribuição de probabilidades de redes com $N = 100000$ nós geradas pelo modelo de Barabási e Albert 28
Figura 1.12 – Ilustração da evolução do modelo *Duplication-Divergence* 30
Figura 2.1 – Representação da topologia de redes de proteínas utilizando a PDF e a CDF. 37
Figura 2.2 – Topologia das redes experimentais. Os triângulos pretos representam a curva da topologia da rede real, e a linha em vermelho é a curva que melhor representa esses dados 37
Figura 2.3 – Topologia de todas as redes num gráfico de $p(k)$ *versus* $log(k/k_{max})$ 40
Figura 2.4 – Espectro de autovalores dos organismos estudados 41
Figura 2.5 – Topologia das redes geradas pelo modelo *Duplication-Divergence*. Os triângulos vermelhos representam as curvas da topologia das redes reais e a curva composta pelos triângulos azuis destaca a topologia da rede experimental do organismo em questão. A curva formada pelos pontos e pelas barras de erro em preto representa a topologia da rede gerada pelo modelo DD para o conjunto de parâmetros descrito no gráfico. 45
Figura 2.6 – Espectro de autovalores das redes geradas pelo modelo *Duplication-Divergence* 47
Figura 3.1 – Topologia das redes geradas pelo novo modelo desenvolvido. Os triângulos vermelhos representam as curvas da topologia das redes reais e a curva composta pelos triângulos azuis destaca a topologia da rede experimental do organismo em questão. A curva formada pelos pontos e pelas barras de erro em preto representa a topologia da rede gerada pelo novo modelo para o conjunto de parâmetros descrito no gráfico. 53
Figura 3.2 – Espectro de autovalores das redes geradas pelo novo modelo 56
Figura 3.3 – Efeitos da transferência horizontal sobre a topologia da rede da *M. tuberculosis* 59
Figura 3.4 – Efeitos da transferência horizontal sobre o espectro de autovalores da rede da *M. tuberculosis* 61
Figura B.1 – Topologia das redes experimentais. Os triângulos pretos representam a curva da topologia da rede real, e a linha em vermelho é a curva que

melhor representa esses dados 68
Figura B.2 – Espectro de autovalores dos organismos estudados 70
Figura C.1 – Topologia das redes geradas pelo modelo de Barabási e Albert. Em preto temos a curva da topologia da rede gerada pelo modelo, com triângulos azuis temos a curva da topologia da rede real e as demais curvas representam a topologia das outras redes. 72
Figura C.2 – Espectro de autovalores das redes geradas pelo modelo de Barabási e Albert 75
Figura D.1 – Topologia das redes geradas pelo modelo *Duplication-Divergence*. Os triângulos vermelhos representam as curvas da topologia das redes reais e a curva composta pelos triângulos azuis destaca a topologia da rede experimental do organismo em questão. A curva formada pelos pontos e pelas barras de erro em preto representa a topologia da rede gerada pelo modelo DD para o conjunto de parâmetros descrito no gráfico. 78
Figura D.2 – Espectro de autovalores das redes geradas pelo modelo *Duplication-Divergence* 80
Figura E.1 – Topologia das redes geradas pelo novo modelo. Os triângulos vermelhos representam as curvas da topologia das redes reais e a curva composta pelos triângulos azuis destaca a topologia da rede experimental do organismo em questão. A curva formada pelos pontos e pelas barras de erro em preto representa a topologia da rede gerada pelo modelo DD para o conjunto de parâmetros descrito no gráfico. 82
Figura E.2 – Espectro de autovalores das redes geradas pelo novo modelo 84
Figura F.1 – Efeitos da transferência horizontal sobre a topologia da rede da *E. coli* . . 86
Figura F.2 – Efeitos da transferência horizontal sobre a topologia da rede da *B. subtilis* 87
Figura F.3 – Efeitos da transferência horizontal sobre o espectro de autovalores da rede da *E. coli* 88
Figura F.4 – Efeitos da transferência horizontal sobre o espectro de autovalores da rede da *B. subtilis* 89

LISTA DE TABELAS

Tabela 2.1 – Relação de parâmetros da lei de potência com *cutoff* exponencial para cada organismo 39
Tabela 3.1 – Valor estimado do parâmetro m para cada organismo, calculado a partir da quantidade de *links* e nós de cada rede 52
Tabela 3.2 – Probabilidade de duplicação estimada para cada um os organismos do reino *Monera* de acordo com a probabilidade de ocorrer transferência horizontal 58

LISTA DE QUADROS

Quadro 2.1 – Tamanho da rede de proteínas (N) e do componente gigante do grafo (N_{CG}) de cada organismo estudado, para um *score* de 0.9. 35
Quadro 2.2 – Estimativa da probabilidade de ocorrer a duplicação de nós em cada organismo 43
Quadro 2.3 – Estimativa da probabilidade de ocorrer a duplicação de nós nas redes geradas pelo modelo DD para cada organismo 46
Quadro 3.1 – Estimativa da probabilidade de ocorrer a duplicação de nós em cada organismo para as redes geradas pelo novo modelo 55
Quadro A.1 – Tamanho das redes e dos Componentes Gigantes para diferentes níveis de intervalo de confiança do STRING 67

SUMÁRIO

1 INTRODUÇÃO 15
1.1 TEORIA DE GRAFOS 17
1.2 MODELOS PARA A REPRODUÇÃO DE GRAFOS E REDES REAIS 22
1.2.1 Modelo de Erdős-Rényi para redes aleatórias 23
1.2.1.1 Redes reais não podem ser completamente aleatórias 25
1.2.2 Modelo de Barabási e Albert (BA) 26
1.2.3 Modelo *Duplication-Divergence* (DD) 28
1.3 TRANSFERÊNCIA HORIZONTAL DE GENES 30
1.4 MOTIVAÇÃO E OBJETIVOS 31
2 METODOLOGIA 33
2.1 ESCOLHA DOS ORGANISMOS 33
2.1.1 STRING - *Search Tool for the Retrieval of Interacting Genes/Proteins* 33
2.2 DEFINIÇÃO DO *SCORE* E CONSTRUÇÃO DAS REDES DE INTERESSE 34
2.2.1 Topologia das redes experimentais 36
2.2.2 Espectro de autovalores das redes reais 40
2.2.3 Probabilidade de duplicação das redes reais 42
2.3 MODELANDO AS REDES REAIS COM O MODELO *DUPLICATION-DIVERGENCE* 44
2.4 DESENVOLVIMENTO DE UM NOVO MODELO PARA REPRODUZIR REDES DE INTERAÇÕES ENTRE PROTEÍNAS 48
3 RESULTADOS E DISCUSSÃO 51
3.1 TOPOLOGIA DAS REDES GERADAS PELO NOVO MODELO 52
3.2 COMPARANDO A PROBABILIDADE DE DUPLICAÇÃO DAS REDES GERADAS PELO NOVO MODELO COM AS REDES EXPERIMENTAIS 54
3.3 ESPECTRO DE AUTOVALORES DAS REDES GERADAS PELO NOVO MODELO 55
3.4 ANÁLISE DO IMPACTO DA TRANSFERÊNCIA HORIZONTAL EM REDES DOS ORGANISMOS PERTENCENTES AO REINO *MONERA* 57
3.4.1 Efeito da transferência horizontal sobre a topologia 58
3.4.2 Efeito da transferência horizontal sobre o espectro de autovalores 60
4 CONCLUSÕES 62
REFERÊNCIAS BIBLIOGRÁFICAS 65
APÊNDICE A – TAMANHO DAS REDES E DO COMPONENTE GIGANTE PARA DIFERENTES NÍVEIS DE CONFIANÇA (*SCORE*) DO STRING 67
APÊNDICE B – TOPOLOGIA E ESPECTRO DE AUTOVALORES DAS REDES REAIS 68
APÊNDICE C – COMPARAÇÃO ENTRE AS REDES GERADAS PELO MODELO DE BARABÁSI E ALBERT E AS REDES EXPERIMENTAIS 72
APÊNDICE D – TOPOLOGIA E ESPECTRO DE AUTOVALORES DAS REDES GERADAS PELO MODELO *DUPLICATION-DIVERGENCE* 78
APÊNDICE E – TOPOLOGIA E ESPECTRO DE AUTOVALORES DAS REDES GERADAS PELO NOVO MODELO 82
APÊNDICE F – EFEITOS DA TRANSFERÊNCIA HORIZONTAL SOBRE AS REDES DA *E. COLI* E DA *B. SUBTILIS* 86
APÊNDICE G – ALGORITMOS DESENVOLVIDOS PARA A REPRODUÇÃO DAS REDES COM OS MODELOS ESTUDADOS 90

ANEXO A – DERIVANDO A DISTRIBUIÇÃO DE *POISSON* A PARTIR DA DISTRIBUIÇÃO BINOMIAL 94

1 INTRODUÇÃO

É difícil negar que estamos no chamado século da complexidade, onde nos encontramos rodeados por sistemas cada vez maiores e mais complexos, compostos por milhares, milhões ou até bilhões de elementos que interagem entre si de diferentes maneiras. Para onde quer que a gente olhe, é possível visualizar ou ao menos relacionar o que estamos vendo com um conceito que vem tomando bastante forma e importância recente: sistemas complexos.

De maneira simplificada, um sistema complexo é uma rede, ou um sistema, formada(o) por um grande número de elementos que interagem entre si. Além disso, em virtude do seu tamanho e da complexidade de suas interações e relacionamentos internos, torna-se extremamente complicado modelar tais sistemas. No entanto, diversos autores já encararam o desafio e se propuseram a estudar e a modelar algumas dessas redes complexas. Mas afinal, onde podemos encontrar tais redes?

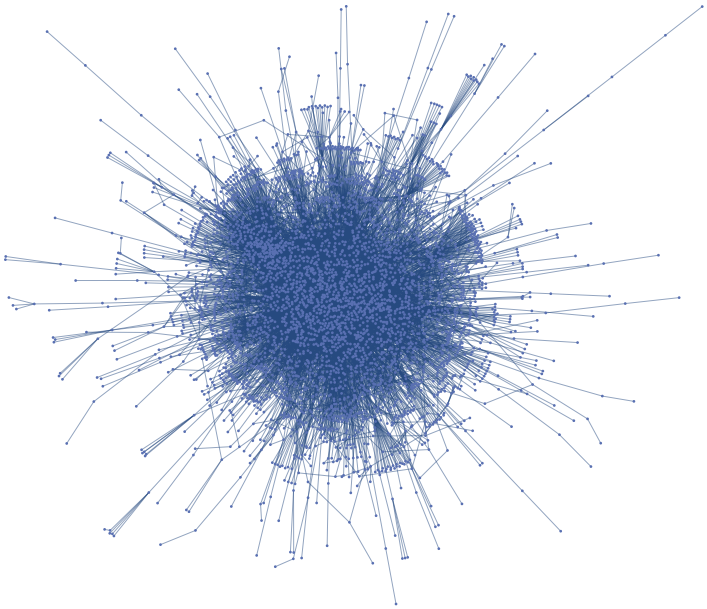
Basta olhar ao redor que já é possível encontrar uma série de redes complexas, entre as quais podemos citar algumas como: a estrutura da *World Wide Web* (WWW), a nossa sociedade cujas interações sociais formam uma rede complexa, redes de distribuição de energia, redes de interações entre proteínas, redes metabólicas e assim por diante. Como podemos ver, em qualquer área e em qualquer lugar que olharmos, poderemos encontrar alguma estrutura que possa ser representada por uma rede complexa.

É importante destacar que apesar de levarem o nome de redes complexas e de serem redes das mais variadas origens e tamanhos, no final podemos representar a estrutura das mesmas utilizando uma única ferramenta matemática: a teoria de grafos. Abordaremos com mais detalhes a teoria de grafos na sessão a seguir, mas de maneira simples, com a teoria de grafos podemos reduzir as redes complexos a um conjunto de nós e de *links* que fazem a conexão entre esses nós, de forma que é possível analisar e estudar cada uma dessas redes de maneira independente da sua origem, utilizando somente uma ferramenta matemática.

No entanto, como também veremos mais adiante, embora seja a base para representar as redes reais, a teoria de grafos não é suficiente para explicar todas as propriedades que podem ser encontradas em redes complexas reais, o que levou muitos pesquisadores a estudarem mais afundo essas redes e com isso proporem modelos para reproduzir essas redes. Vale destacar, no entanto, que apesar de obterem sucesso ao conseguir reproduzir alguns tipos de redes complexos, tais modelos nem sempre conseguem reproduzir todas as propriedades das redes, o que pode ser um forte indício de que apesar de poderem ser representadas com a mesma ferramenta, existem diferenças nos processos evolutivos por trás dessas redes que faz com que não seja possível, até o momento, encontrar um modelo universal para reproduzir redes complexas.

Dentre a enorme gama de redes complexas que existem, voltaremos nossa atenção para as redes de interações entre proteínas. Uma rede de interações entre proteínas de um organismo representa, de maneira simplificada, o conjunto de todas as proteínas expressadas por genes daquele organismo bem como as interações físicas que existem entre essas proteínas. Ao fazer esse mapa de proteínas e interações, é possível obter uma estrutura na forma de rede (ou grafo) para cada organismo, e com essa estrutura podemos calcular uma série de propriedades de interesse para o organismo em questão. Essa rede é resultado de uma série de eventos bioquímicos (forças eletrostáticas, ligações de hidrogênio e o efeito hidrofóbico, por exemplo) que ocorreram durante a evolução de um organismo e deram origem à interações e ligações físicas entre duas ou mais proteínas dentro do organismo, formando, assim, o que conhecemos como uma rede de interações entre proteínas. Embora, num primeiro momento, se imagine que em virtude dessas ligações e dessas propriedades a estrutura dessas redes pode oferecer alguma complicação maior para o estudo da mesma, é importante destacar o fato de que com a teoria de grafos podemos remover a natureza da rede e considerar a mesma como sendo somente um conjunto de nós e *links* entre esses nós, como podemos observar na Figura 1.1.

Figura 1.1 – Estrutura na forma de grafo para a rede de interações entre proteínas da *Caenorhabditis elegans*



Fonte: O autor.

Compreender a estrutura e como ocorreu o processo evolutivo dessas redes é importante pois através dessas informações é possível compreender melhor como algumas doenças funcionam, compreender como, em nível sistêmico, os organismos funcionam, identificar possíveis alvos para drogas e medicamentos que possam auxiliar no tratamento de doenças que ainda não temos cura, dentre outras inúmeras possíveis aplicações. Mo-

tivados por esses e outros fatores, diversos autores se propuseram a estudar as redes de interações entre proteínas e começaram a desvendar as suas características evolutivas. Alguns propuseram modelos matemáticos e computacionais que buscassem produzir redes que ao final teriam as mesmas propriedades das redes reais, o que pode auxiliar na compreensão mais aprofundada de como ocorreu a evolução dessas redes.

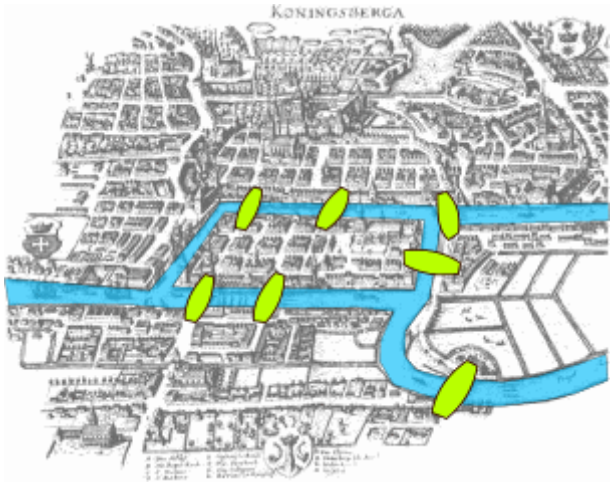
Para uma melhor compreensão dos conceitos e das ferramentas utilizadas ao longo do presente trabalho, trazemos a seguir uma breve introdução à teoria de grafos e às principais características e propriedades de uma rede complexa.

1.1 TEORIA DE GRAFOS

A teoria de grafos é a parte da matemática responsável pelo estudo de estruturas, os chamados grafos, que são formados por um conjunto de nós (ou vértices) ligados por arestas (ou *links*). A origem desta teoria vem da primeira metade do século XVIII, quando Euler a utilizou para resolver um problema que estava sendo bastante discutido na época e que hoje é conhecido como "as sete pontes de Königsberg".

O problema em questão, representado visualmente na Figura 1.2, consistia na seguinte pergunta: é possível cruzar todas as pontes e nunca passar duas vezes pela mesma?

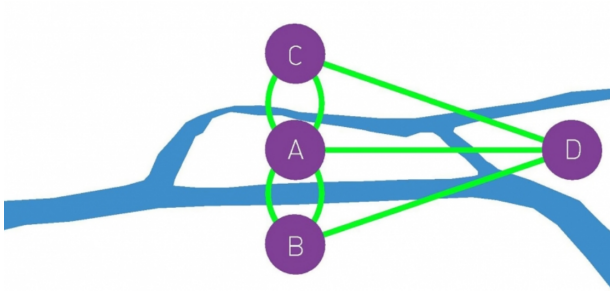
Figura 1.2 – As sete pontes de Königsberg



Fonte: Wikipedia.

Para resolver tal problema, Euler utilizou uma representação bastante simples, nomeando cada parte de terra com uma letra e representado-as por pontos. As pontes foram representadas por linhas unindo os respectivos pontos. Na Figura 1.3 podemos ver uma reprodução da visualização do problema.

Figura 1.3 – Desenho esquemático do problema



Fonte: Albert-László Barabási, *Network Science*.

O que temos na Figura 1.3 nada mais é do que um grafo, composto por 4 nós (que são as áreas de terras A, B, C, e D) e 7 arestas (que são as pontes entre as áreas de terras). A partir deste grafo, Euler chegou a conclusão de que se realmente existir um caminho que cruzasse todas as pontes sem passar duas vezes pela mesma, o ponto inicial e final da caminhada devem ser nós da rede que possuem um número ímpar de ligações.

Como só deve existir um ponto de saída e um ponto de chegada, é necessário que o grafo tenha somente dois nós com um número ímpar de arestas, correspondentes a esses pontos. No entanto, o grafo da Figura 1.3 possui 4 nós com um número ímpar de arestas, de modo que é impossível existir um caminho que cruze todas as pontes sem passar pela mesma duas vezes.

Ao utilizar um grafo para resolver esse problema, Euler mostrou que em muitas situações é muito mais fácil reduzir o problema a uma estrutura simples como a de um grafo, pois permite uma solução mais fácil e rápida. Além disso, ele mostrou que a resposta desse problema em questão é uma propriedade do grafo, pois independentemente de quantas maneiras tentar, não será possível encontrar o caminho pressuposto no problema.

Motivados pelo exemplo acima, podemos introduzir agora os conceitos e propriedades dos grafos (ou redes, como geralmente são chamados). As primeiras quantidades necessárias para se construir um grafo são o número de nós, N, e o número de *links* (arestas), L, de modo que podemos definir um grafo G como sendo um par ordenado $G = (N, L)$. No exemplo anterior, o grafo era formado por $N = 4$ nós e $L = 7$ *links*. Outra propriedade, que inclusive foi usada por Euler para sustentar seu raciocínio, é a conectividade (ou grau) de um nó, k_i, que corresponde ao número de *links* que o nó i tem para outros nós do grafo. No grafo da Figura 1.3, por exemplo, temos as seguintes conectividades: $k_A = 5$, $k_B = 3$, $k_C = 3$ e $k_D = 3$.

Outro conceito importante na análise de uma rede é fazer a diferenciação entre uma rede dirigida e uma rede não-dirigida. Em uma rede não-dirigida, não existe diferença na direção de um *link*, ou seja, um *link* que vai do nó A até o nó B é igual a um link que vai do

nó B ao nó A. Já na rede dirigida, esses dois *links* são diferentes, uma vez que possuem direções diferentes. Na rede não-dirigida, fala-se somente na conectividade k_i do nó i, enquanto que na rede dirigida é preciso diferenciar entre a conectividade de saída do nó i, k_i^{out}, e a conectividade de entrada, k_i^{in}. Temos, ainda, a conectividade média, $\langle k \rangle$, que é uma característica da rede, e fornece o número médio de conexões entre os nós.

Para as redes não-dirigidas, calculamos as propriedades já citadas da seguinte maneira:

- Conectividade total do nó i:

$$k_i$$

- Número total de *links*:

$$L = \frac{1}{2} \sum_{i=1}^{N} k_i \tag{1.1}$$

- Conectividade média:

$$\langle k \rangle \equiv \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2L}{N} \tag{1.2}$$

E para as redes dirigidas, essas propriedades são dadas por:

- Conectividade total do nó i:

$$k_i = k_i^{in} + k_i^{out} \tag{1.3}$$

- Número total de *links*:

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out} \tag{1.4}$$

- Conectividade média:

$$(k^{in}) = \frac{1}{N} \sum_{i=1}^{N} k_i^{in} = (k^{out}) = \frac{1}{N} \sum_{i=1}^{N} k_i^{out} = \frac{L}{N} \tag{1.5}$$

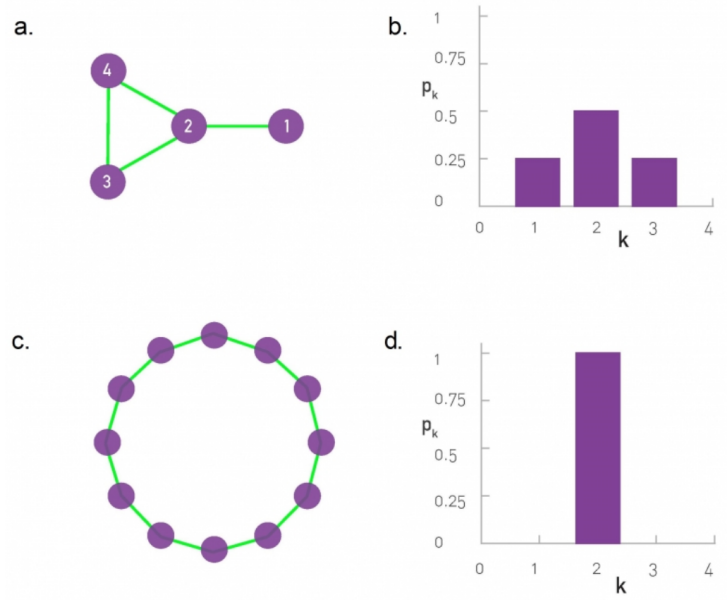
Além das propriedades já citadas, temos também a distribuição de conectividades (ou distribuição de grau), p_k, que nos fornece a probabilidade de que um nó escolhido aleatoriamente na rede possua uma conectividade k. Ao longo deste trabalho, vamos nos referir à distribuição de conectividades de uma rede como sendo a sua topologia. Por ser uma probabilidade, p_k deve ser normalizada, isto é, $\sum_{k=1}^{N} p_k = 1$. Para uma rede de N nós, a distribuição de conectividades é dada pela razão:

$$p_k = \frac{N_k}{N}, \tag{1.6}$$

onde N_k é o número de nós com conectividade k. A distribuição de conectividades é uma propriedade importante das redes, visto que cada tipo de rede possui uma forma carac-

terística dessa distribuição, conforme será visto mais adiante. Na Figura 1.4 podemos visualizar dois grafos diferentes junto com o histograma de suas distribuições de conectividades.

Figura 1.4 – Exemplos de grafos com suas respectivas distribuições de conectividades



Fonte: Albert-László Barabási, *Network Science*.

Dando sequência às propriedades dos grafos, introduzimos agora a matriz de adjacência. A matriz de adjacência é uma matriz A com N linhas e N colunas, e seus elementos são:

$$A_{ij} = \begin{cases} 1 & \text{se existe um } link \text{ entre os nós } i \text{ e } j, \\ 0 & \text{se os nós } i \text{ e } j \text{ não estão conectados entre si.} \end{cases}$$

Para o caso de uma rede não-dirigida, temos uma matriz simétrica ($A_{ij} = A_{ji}$), uma vez que um mesmo *link* é representado por duas entradas. A partir da matriz de adjacência podemos calcular a conectividade de cada nó i da rede. Para redes não-dirigidas teremos:

$$k_i = \sum_{j=1}^{N} A_{ij} = \sum_{i=1}^{N} A_{ij}. \tag{1.7}$$

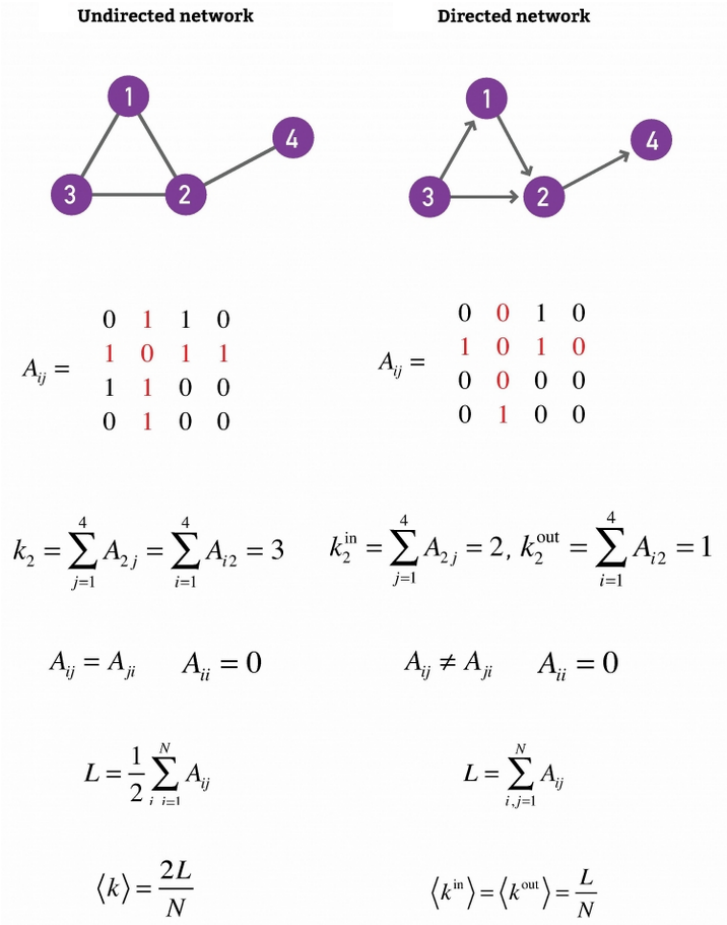
E para redes dirigidas teremos:

$$k_i^{in} = \sum_{j=1}^{N} A_{ij}, \tag{1.8}$$

$$k_i^{out} = \sum_{i=1}^{N} A_{ij}. \tag{1.9}$$

Na Figura 1.5 podemos visualizar um grafo não-dirigido (à esquerda) junto com sua matriz de adjacência e um grafo dirigido (à direita) com sua respectiva matriz de adjacência. Como citado anteriormente, a matriz de adjacência de um grafo não-dirigido é simétrico, ao passo que para um grafo dirigido a matriz não será simétrica.

Figura 1.5 – Representação de uma rede através de grafo e sua matriz de adjacência



Fonte: Albert-László Barabási, *Network Science*.

Dentre as demais propriedades dos grafos, vamos citar ainda o coeficiente de clusterização local, que pode ser definido como uma espécie de densidade local do grafo, uma vez que fornece informações a respeito da maneira como os vizinhos de um dado nó estão conectados entre si. O coeficiente de clusterização local de um nó i com conectividade k_i

é dado por:

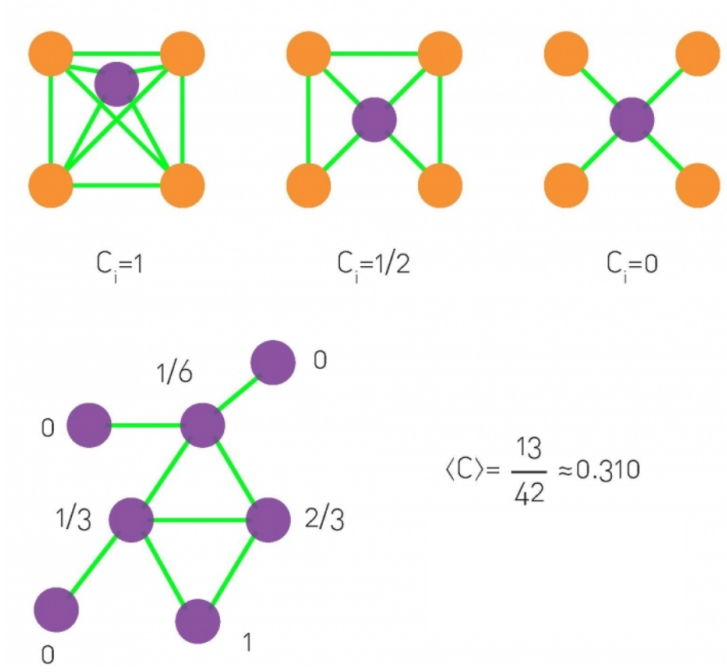
$$C_i = \frac{2L_i}{k_i(k_i - 1)}, \tag{1.10}$$

onde L_i é o número de *links* entre os k_i vizinhos do nó i. Também podemos definir o coeficiente de clusterização médio da rede, $\langle C \rangle$, que é dado por:

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{1.11}$$

Na Figura 1.6 é possível visualizar diferentes tipos de grafos e os respectivos valores dos coeficientes de clusterização local.

Figura 1.6 – Coeficiente de clusterização local de diferentes grafos



Fonte: Albert-László Barabási, *Network Science*.

1.2 MODELOS PARA A REPRODUÇÃO DE GRAFOS E REDES REAIS

A partir da base matemática fornecida pela teoria de grafos, começaram a surgir modelos que buscassem reproduzir as redes encontradas no mundo real. O primeiro destes modelos, que será apresentado a seguir, é o modelo de redes aleatórias, desenvolvido por Paul Erdős e Alfréd Rényi (ERDŐS; RÉNYI, 1959). Nas sessões a seguir também veremos que as redes reais não podem ser completamente aleatórias, e por isso foram desenvolvidos novos modelos para reproduzir as mesmas.

1.2.1 Modelo de Erdős-Rényi para redes aleatórias

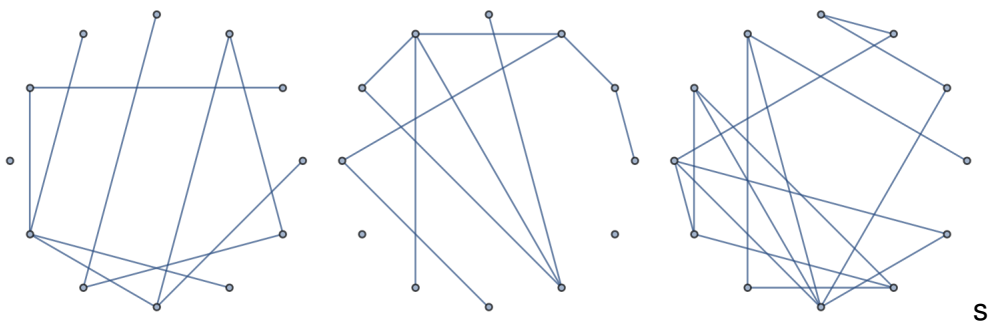
Este foi o primeiro modelo desenvolvido com o objetivo de reproduzir uma rede real, em um período onde ainda acreditava-se que as redes reais eram construídas de maneira aleatória, sem ter um princípio evolutivo claro.

De maneira simples, podemos definir uma rede aleatória como uma *rede com N nós devidamente identificados onde cada par de nós é conectado com a mesma probabilidade* p (BARABÁSI, 2016). Um algoritmo simplificado para gerar uma rede aleatória $G(N, p)$ é o seguinte:

1. Crie uma rede com N nós isolados.
2. Selecione um par de nós desta rede e gere um número aleatório entre 0 e 1. Quando este número for maior que p, cria-se um *link* entre os dois nós.
3. Repita o segundo passo para cada um dos $\frac{N(N-1)}{2}$ pares de nós.

Por ser uma construção aleatória, mesmo que usarmos os mesmos valores de N e p para gerar diferentes redes, é pouco provável que elas sejam iguais, como podemos ver na Figura 1.7.

Figura 1.7 – Três realizações de uma rede aleatória com os mesmos parâmetros



Fonte: O autor.

Conforme vemos na Figura 1.7, o número de *links* em cada uma das realizações é diferente, mesmo tendo usado os mesmo parâmetros para gerar todas elas. Como são redes aleatórias, podemos calcular o número esperado de *links* ($\langle L \rangle$) para uma distribuição com N nós e probabilidade p:

$$\langle L \rangle = p\frac{N(N-1)}{2} \tag{1.12}$$

Além disso, substituindo a equação 1.12 na equação 1.2 podemos obter a conectividade média desta mesma rede:

$$\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1) \tag{1.13}$$

Por fim, falaremos sobre a distribuição de probabilidades das redes aleatórias, propriedade esta, que como já foi mencionado anteriormente, é uma característica da própria rede. A probabilidade de encontrarmos em uma rede aleatória um nó i que tenha exatamente k *links* será dada pelo seguinte produto (BARABÁSI, 2016):

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}, \tag{1.14}$$

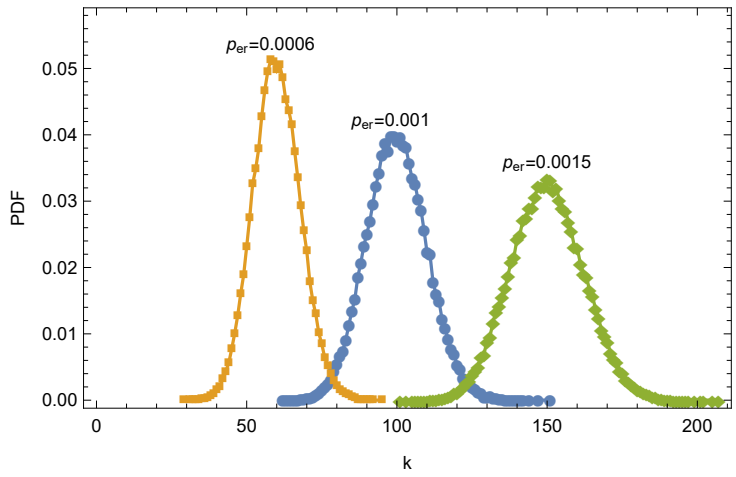
onde p^k é a probabilidade de que esses k *links* existam; $(1-p)^{N-1-k}$ é a probabilidade de que os demais *links* estejam faltando e $\binom{N-1}{k}$ é o número de maneiras que podemos selecionar k *links* dentre as $N-1$ possibilidades. A equação 1.14 é a equação da distribuição binomial, de modo que a distribuição de probabilidades (ou seja, a topologia da rede aleatória), vai ter a forma de uma função binomial.

No entanto, ao aplicarmos a equação 1.14, onde geralmente lidamos com redes com um número muito grande de nós, teremos que $\langle k \rangle \ll N$, de modo que ao tomarmos este limite, a distribuição dada por 1.14 poderá ser aproximada pela distribuição de *Poisson* (a derivação desta expressão encontra-se no Anexo A):

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \tag{1.15}$$

As distribuições dadas pelas equações 1.14 e 1.15 são usadas para representar a topologia das redes aleatórias, pois possuem uma série de propriedades em comum, e a principal diferenças entre elas está relacionada com o tamanho da rede. Na Figura 1.8 podemos visualizar a distribuição de probabilidades de três redes aleatórias. Cada uma delas possui $N = 100000$ nós e as probabilidades utilizadas foram $p = 0.0006$, $p = 0.001$ e $p = 0.0015$.

Figura 1.8 – Topologia de três redes aleatórias com $N = 100000$ nós e probabilidades p indicadas na imagem



Fonte: O autor.

Como pode ser visto pela Figura 1.8, o tamanho da rede é muito maior do que a conectividade média da mesma, de modo que a forma da topologia é a distribuição de *Poisson*, conforme Equação 1.15.

No entanto, é contraintuitivo pensar que as redes reais sejam completamente aleatórias, e mostraremos isso na próxima seção.

1.2.1.1 Redes reais não podem ser completamente aleatórias

Quando paramos para pensar em algumas redes reais, tais como redes de colaborações científicas, redes de atores, a rede da internet, redes de interações entre proteínas, e assim por diante, é complicado pensarmos nestas redes como sendo construídas através de um comportamento completamente aleatório.

Considerando por exemplo uma rede metabólica ou então uma rede de interações entre proteínas, fica claro que as interações entre duas proteínas dentro de um organismo devem seguir leis biológicas e/ou bioquímicas que resultam em funções específicas dentro daquele organismo, de modo que isso não seria possível se fosse um comportamento aleatório.

O mesmo vale para uma rede de colaboração científica, elaborada com as citações dos trabalhos: ao elaborar um artigo referente à uma área X, utilizam-se como fontes e citações outros artigos que também foram publicados na mesma área, de modo que os trabalhos estejam relacionados entre si. Sendo assim, a rede gerada por essas interações não pode ser aleatória.

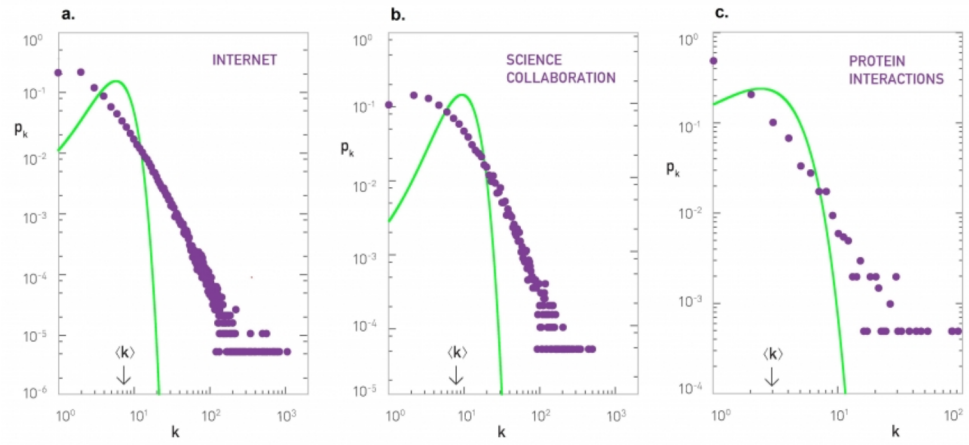
Como visto na seção anterior, o que caracteriza uma rede aleatória é a sua topologia (distribuição de probabilidade), que assume a forma de uma distribuição binomial, ou de *Poisson*, quando a rede é muito grande.

Na Figura 1.9 podemos observar a topologia de três redes reais: a) internet, b) colaboração científica e c) interações entre proteínas. Os pontos roxos correspondem à topologia da rede real, construída com base em informações reais. A curva verde em cada um dos gráficos representa a topologia de uma rede aleatória gerada com o mesmo tamanho da rede real e com o valor da conectividade média de cada uma das redes.

A curva em verde da Figura 1.9 corresponde à distribuição de *Poisson*, com a única diferença de que o gráfico encontra-se em escala logarítmica, por isso a diferença no formato quando comparada com a Figura 1.15. Uma simples análise dessas comparações já basta para observar que a topologia de uma rede real é bem diferente da topologia de uma rede aleatória, fornecendo mais indícios de que as redes reais devem obedecer a mecanismos evolutivos bem definidos, seguindo leis biológicas, físicas, matemáticas ou químicas.

Assim, a partir dessas novas evidências a respeito das redes reais não apresen-

Figura 1.9 – Comparação entre a topologia de três redes reais e a sua topologia caso fossem redes aleatórias



Fonte: Albert-László Barabási, *Network Science*.

tarem um comportamento aleatório, foram desenvolvidos novos modelos que buscassem reproduzir melhor as redes do mundo real.

1.2.2 Modelo de Barabási e Albert (BA)

Embora não tenha sido desenvolvido utilizando como base uma rede de interações entre proteínas, mas sim a rede da *World Wide Web* (WWW), o modelo desenvolvido por Albert-László Barabási e Réka Albert foi logo sendo utilizado como base para representar redes complexas das mais variadas origens. O grande trunfo desse modelo foi o fato de ter sido o primeiro a incorporar padrões de crescimento em redes reais, eliminando, assim, o caráter completamente aleatório das mesmas.

O modelo BA baseia-se em dois pilares

i) **crescimento das redes**: a rede inicia com uma pequena quantidade m_0 de nós, e a cada novo passo, um novo nó é adicionado a rede. Esse novo nó, no entanto, poderá se conectar a um número máximo de m nós diferentes já existentes na rede, onde o valor de m obedece a condição $m \leq m_0$;

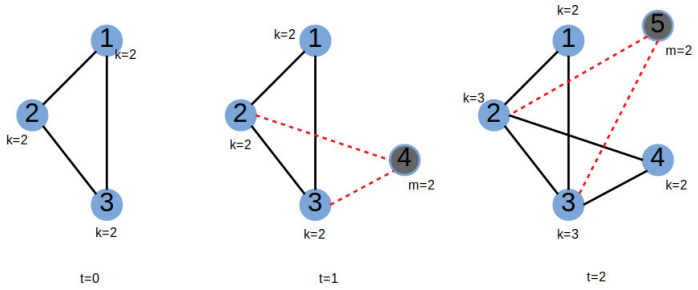
ii) ***Preferential Attachment***: cada novo nó que será adicionado a rede terá uma probabilidade Π de se conectar a um nó i já existente na rede. O valor de Π depende da conectividade, k_i, do nó i, e é calculada pela expressão $\Pi(k_i) = {}^{k_i}/_{\sum_j k_j}$

Dessa forma, após t instantes (ou passos), o modelo produzirá uma rede de $N = m_0 + t$ nós e com mt *links*, evoluindo para uma rede em um estado de escala livre de

escala[1] onde, na rede resultante, a probabilidade de que um nó i tenha uma conectividade k segue uma lei de potência.

Na Figura 1.10 podemos visualizar, de maneira simplificada, como funciona o modelo BA. Nela podemos observar os 3 passos iniciais para uma rede que seguirá o processo evolutivo definido no modelo de BA.

Figura 1.10 – Ilustração do modelo de Barabási e Albert para uma rede com $m_0 = 3$ e $m = 2$



Fonte: O autor.

No primeiro passo, $t = 0$, temos a configuração inicial da rede, que começa com $m_0 = 3$ nós. Foi definido também que $m = 2$ para essa rede em questão. No próximo passo, $t = 1$, o nó 4 será adicionado a rede, e pela definição anterior, ele se conectará a 2 outros nós já existentes na rede. Os nós aos quais ele se conectará serão "escolhidos" pelo mecanismo do *Preferential Attachment*. Nesse caso temos $k_1 = k_2 = k_3 = 2$, então todos os nós têm a mesma preferência na hora de serem "escolhidos" pelo novo nó, e as conexões foram feitas com os nós 2 e 3. Em $t = 2$, temos o nó 5 que também se conectará a dois outros nós. No entanto, nesse passo temos nós com conectividades diferentes: $k_1 = k_4 = 2$ e $k_2 = k_3 = 3$, ou seja, os nós 2 e 3 têm maior preferência sobre os outros na hora que o nó 5 escolhe suas conexões. Embora essa escolha de conexões seja um processo aleatório, de modo que seria plausível que o nó 5 se conectasse também aos nós de menor conectividade, para fim de ilustração e simplificação, nesse exemplo o nó 5 se conecta aos nós 2 e 3 por serem os que tem maior conectividade e por consequência uma maior preferência no mecanismo do *Preferential Attachment*. Assim, ao final do terceiro passo teremos uma rede com 5 nós e com conectividades $k_1 = k_4 = k_5 = 2$ e $k_2 = k_3 = 4$, e o processo ilustrado aqui seguirá até que a rede atinja seu tamanho final N.

Uma das principais características das redes geradas pelo modelo BA é que a topologia dessas redes segue uma lei de potência, ou seja, temos

$$p_k \sim k^{-\gamma}, \tag{1.16}$$

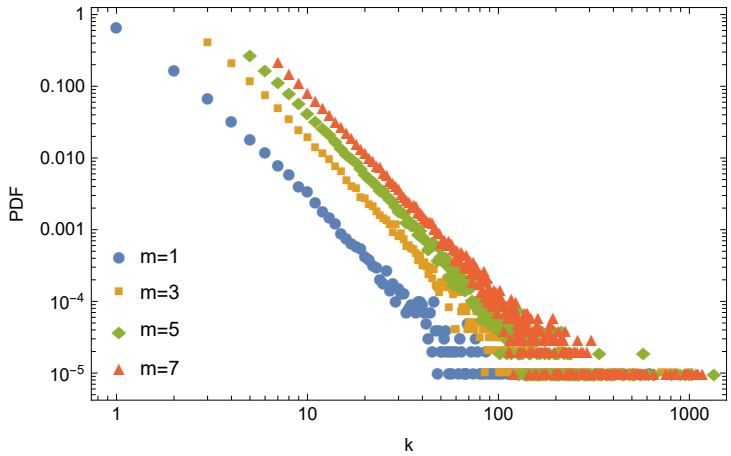
[1]Uma rede livre de escala (do inglês, *scale-free network*) é uma rede cuja distribuição de probabilidades segue uma lei de potência (BARABÁSI; ALBERT, 1999)

onde γ é o grau da distribuição. Ao tomarmos o logaritmo da Equação 1.16,

$$log(p_k) \sim -\gamma log(k), \tag{1.17}$$

teremos, em uma escala logarítmica, uma reta cuja inclinação é dada por γ. Esse comportamento pode ser observado na Figura 1.11, onde temos o gráfico em escala logarítmica de p_k *versus* k, ilustrando o comportanto da Equação 1.17, para 4 realizações de uma rede gerada com o modelo BA com $N = 100000$ nós e diferentes valores de m, descritos na imagem.

Figura 1.11 – Distribuição de probabilidades de redes com $N = 100000$ nós geradas pelo modelo de Barabási e Albert



Fonte: O autor.

Outra característica marcante das redes geradas pelo modelo BA é a formação de *hubs* na rede, isto é, nós com uma conectividade bastante alta. Pela Figura 1.11 é possível verificar a presença desses *hubs*, uma vez que os mesmos se concentram na cauda da distribuição de probabilidades, que é, justamente, a região onde temos a probabilidade de encontrar nós com conectividade alta.

1.2.3 Modelo *Duplication-Divergence* (DD)

Enquanto o modelo BA foi desenvolvido com o intuito de ser utilizado como um modelo universal para explicar a evolução de redes complexas das mais variadas origens, Vázquez e colaboradores propuseram um modelo que fosse capaz de reproduzir os mecanismos evolutivos por trás das redes de interações entre proteínas. Assim, basearam seu modelo no fato de que as proteínas evoluíram a partir de um ancestral em comum através da duplicação de genes, e posteriormente foram sofrendo pequenas mutações ao longo do seu processo evolutivo, de modo que a rede resultante gerada por esse modelo possa ser

vista como sendo uma impressão digital da história da evolução do genoma (VÁZQUEZ et al., 2003).

No modelo DD, cada nó da rede é uma proteína expressada por um gene, de modo que logo após a duplicação de um gene, as proteínas resultantes desse processo terão as mesmas interações. Esse processo de duplicação corresponde à adição de um novo nó a rede, e esse nó terá exatamente as mesmas conexões que seu ancestral. Após ocorrer a duplicação, existe a possibilidade de que a proteína que foi duplicada seja auto-interagente, isto é, poderá se conectar diretamente à seu ancestral. Por fim, ao longo desse processo evolutivo é possível que ocorram mutações na redes, fazendo com que algumas interações entre as proteínas sejam removidas, resultando em nós diferentes na rede. Logo, por mais que o mecanismo básico por trás do crescimento da rede seja a duplicação de nós já existentes, em virtude da possibilidade de ser uma proteína autointeragente e de ocorrerem mutações, a rede resultante não será uma rede composta somente por nós duplicados.

De modo a capturar esses processos que ocorrem, o modelo DD inicia com 2 nós conectados entre si por um *link* entre eles. Os processos de duplicação e mutação serão repetidos por N vezes, onde N é o tamanho final da rede, e esses processos ocorrem da seguinte maneira:

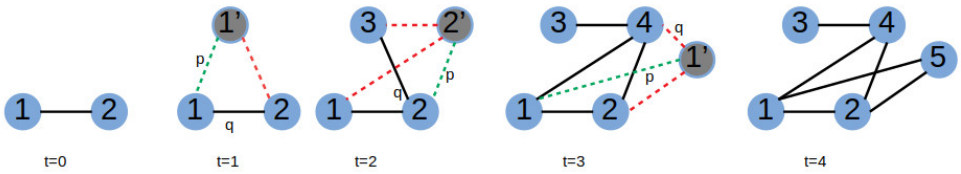
i) **duplicação**: um nó i, já existente na rede, é escolhido aleatoriamente. Esse nó será duplicado, dando origem ao nó $i^{'}$, que terá os mesmos vizinhos de i. Após a duplicação, com uma probabilidade p ($0 \leq p \leq 1$) poderá ser criado um *link* entre i e $i^{'}$;

ii) **mutação**: ocorre logo após o processo de duplicação. Para cada um dos j nós conectados aos nós i e $i^{'}$, escolhe-se aleatoriamente um dos dois links (i, j) ou $(i^{'}, j)$ que será removido com uma probabilidade q ($0 \leq q \leq 1$).

Nesse modelo temos dois parâmetros: p e q. O parâmetro p representa a probabilidade de ocorrer a criação de um link entre a proteína pai e a proteína filha, representando a possibilidade de existirem proteínas auto-interagentes na rede. Já o parâmetro q representa a perda de interações entre os duplicados e seus vizinhos devido a divergência dos duplicados (como consequência de processos de mutação ao longo da evolução da rede). Na Figura 1.12 podemos visualizar, de maneira simplificada, como funciona o mecanismo do modelo DD.

Na Figura 1.12 temos ilustrado os 5 primeiros passo de uma rede que será construída seguindo o modelo DD. Em $t = 0$ temos a configuração inicial da rede, com 2 nós conectados entre si. Em $t = 1$, será adicionado o primeiro novo nó, que irá duplicar, após uma escolha aleatória, um dos nós já existentes na rede. Nesse caso, o nó escolhido foi o 1, de modo que o nó $1^{'}$ irá se conectar ao nó 2 (linha pontilhada vermelha). Além disso, com uma probabilidade p, ele poderá ou não ter o link representado pela linha pontilhada verde. Aqui não foi criada a interação entre o nó pai e o nó filho, então o nó $1^{'}$ passou

Figura 1.12 – Ilustração da evolução do modelo *Duplication-Divergence*



Fonte: O autor.

a ser o nó 3 no passo $t = 2$. O processo de duplicação continua, e dessa vez o nó que será duplicado é o 2. Logo, o nó $2^{'}$ irá se conectar aos nós 1 e 3, que são as mesmas conexões que o nó pai tem. Após a duplicação, com uma probabilidade p foi criado o *link* entre o nó pai e o nó filho, $(2, 2^{'})$. Após esse processo, existe a possibilidade de ocorrer a mutação em algum dos links. O *link* que poderá sofrer a mutação, com probabilidade q, é o *link* $(2, 3)$. Como vemos em $t = 3$, não houve a exclusão desse link. Também em $t = 3$, o nó 1 será duplicado novamente, de modo que o nó $1^{'}$ se conectará aos nós 2 e 4. Como podemos ver em $t = 4$, com a probabilidade p foi criado o link $(1, 5)$ entre o nó pai e o nó filho, e com probabilidade q ocorreu a remoção do link $(4, 5)$, representando a ocorrência de um processo de mutação nessa proteína. Esse mesmo processo que foi representado nos passos de $t = 0$ até $t = 4$ irá se repetir até que a rede atinja seu tamanho final de N nós.

1.3 TRANSFERÊNCIA HORIZONTAL DE GENES

Embora a duplicação de genes (ou nós na rede) seja um fator importante para o processo evolutivo de uma rede de interações entre proteínas, esse pode não ser o único mecanismo responsável por tal processo e pelas propriedades interessantes que serão apresentadas mais a frente para as redes de interações entre proteínas. Um possível mecanismo que pode estar por trás desse processo, ao lado da duplicação, é a Transferência Horizontal de Genes (THG).

De modo simples, a THG é um processo que resulta na aquisição de novas informações genéticas em um organismo, de modo que os processos genômicos evolutivos que existem no organismo passem a atuar não somente nos genes pré-existentes, mas também nesses novos genes (TREANGEN; ROCHA, 2011), o que pode levar à aquisição de novas funções no organismo. Essa aquisição de genes que posteriormente pode levar à novas funções dentro do organismo é de fundamental importância pois é através da aquisição de novas funções que um organismo pode evoluir e dar origem a outros processos e até mesmo outros organismos.

Estimar a real ocorrência da THG em um organismo não é fácil, pois existem uma série de fatores que devem ser considerados na hora de determinar a prevalência de tal processo durante a evolução de um organismo. A real prevalência da THG nos organismos não é algo bem consolidado, uma vez que alguns estudos sugerem que até 96% dos genes seriam oriundos da THG e outros indicam que ela seria responsável por, no máximo, 25% das expansões das famílias de proteínas (TREANGEN; ROCHA, 2011). Apesar de não haver um consenso na real participação da THG na evolução genômica, sabe-se que esse mecanismo, ao lado da duplicação de genes, é responsável por uma série de características que podem ser observadas nos mais diversos organismos, de modo que inserir ele em um modelo que busque reproduzir as redes de interações entre proteínas seja importante para obter resultados o mais próximo do real possível.

1.4 MOTIVAÇÃO E OBJETIVOS

Como apresentado nas sessões anteriores, já foram propostos uma série de modelos que tentaram reproduzir e explicar as principais características das redes de interações entre proteínas. No entanto, como poderá ser visto com mais ênfase no capítulo a seguir, esses modelos utilizaram como base de comparação somente a topologia de uma rede, deixando de lado outras propriedades e características, tais como o espectro de autovalores e a probabilidade de ocorrer a duplicação dentro da rede, que revelam informações extremamente relevantes sobre a evolução dessas redes. Além disso, esses modelos foram desenvolvidos utilizando-se como base um único mecanismo: ou a duplicação de genes (como apresentado no modelo DD) ou o crescimento com *preferential attachment* (conforme apresentado no modelo BA). Mas como apresentamos acima, existe um outro mecanismo (a Transferência Horizontal) que está presente na evolução das redes de interações entre proteínas e que deve ser considerado na hora de desenvolver um modelo que tente reproduzir com a maior fidelidade possível essas redes.

Assim, definimos como o objetivo geral deste trabalho desenvolver um novo modelo para a reprodução de redes de interações entre proteínas que englobe a transferência horizontal de genes e que consiga reproduzir com boa precisão não só a topologia das redes, mas também o seu espectro de autovalores e a probabilidade de ocorrer a duplicação de nós dentro da mesma. A importância de um modelo que consiga reproduzir adequadamente esses aspectos de uma rede está no fato de que com isso teremos em nossas mãos mais uma ferramenta que nos ajude a identificar proteínas de interesse e também ajude a explicar e desvendar os mecanismos evolutivos dos organismos.

Além do objetivo geral, os objetivos específicos do trabalho são:

- propor um mecanismo para a Transferência Horizontal de Genes;

- analisar os efeitos da Transferência Horizontal de Genes sobre a topologia e o espectro de autovalores das redes de interações entre proteínas.

2 METODOLOGIA

2.1 ESCOLHA DOS ORGANISMOS

Dentre os mais diversos organismos existentes na natureza, foi necessário delimitar o escopo do presente trabalho ao escolher um número reduzido de organismos para reproduzirmos as redes de interação entre proteínas dos mesmos. A escolha dos 18 organismos que foram estudados nesse trabalho se deu em virtude de alguns fatores, entre os quais podemos citar: tamanho das redes, divisão dos mesmos nos diferentes reinos biológicos, proximidade com processos que são comuns no nosso dia a dia (como é o caso da *Saccharomyces cerevisiae*, por exemplo), importância do organismo em estudos e também a disponibilidade de dados e referências sobre cada organismo.

Assim, com esses 18 organismos selecionados, temos pelo menos dois representantes de cada reino da biologia. Os 18 organismos, dividos de acordo com seu reino biológico são:

- ***Monera***: *M. tuberculosis*, *E. coli* e *B. subtilis*;
- ***Protista***: *P. falciparum* e *T. cruzi*;
- ***Fungi***: *S. pombe* e *S. cerevisae*;
- ***Plantae***: *O. sativa* e *A. thaliana*;
- ***Animalia***: *A. mellifera*, *C. elegans*, *D. melanogaster*, *G. domesticus*, *X. laevis*, *B. taurus*, *D. rerios*, *M. musculus* e *H. sapiens*.

Outro ponto bastante importante na escolha dos organismos é a escolha de um banco de dados de interações entre proteínas confiável e que esteja em constante atualização e melhoria, de modo que os resultados obtidos possam ser os mais acurados possíveis. Para a construção das redes de proteínas dos organismos nesse trabalho, utilizou-se o banco de dados *STRING: functional protein association networks*[1].

2.1.1 STRING - *Search Tool for the Retrieval of Interacting Genes/Proteins*

De maneira resumida, o String é um banco de dados biológico onde estão armazenadas diversas interações entre proteínas, sejam elas conhecidas e comprovadas cientificamente ou interações previstas com base em algoritmos e programas computacionais. De

[1] Disponível em *https://string-db.org/*

acordo com informações obtidas no portal do STRING, no presente momento encontram-se cadastrados mais de 14000 organismos, o que resulta em um banco de dados com mais de 20 bilhões de interações entre proteínas, sejam elas documentadas e verificadas experimentalmente ou então previstas computacionalmente.

Para atingir tal feito e obter essa grande quantidade de informação, os contribuidores utilizam diferentes fontes para obter informações e evidências sobre as interações entre proteínas de cada organismo. Dentre essas fontes de coleta de dados, encontram-se i) mineração de textos da literatura científica, ii) análise e coleta de informações de bancos de dados com informações sobre interações experimentais, iii) simulações e previsões computacionais de interações e iv) transferência sistemática de interações conhecidas de um organismo para outro (SZKLARCZYK et al., 2021).

Outro ponto bastante importante sobre os dados presentes no STRING é o conceito de *score* que utilizam para indicar se uma interação, ou associação, entre proteínas tem significado biológico com base em todas as evidências encontradas. Esse *score* é um valor entre 0 e 1, onde quanto mais próximo de 1, maior a importância biológica daquela interação (SZKLARCZYK et al., 2021). Assim, é possível filtrar os dados obtidos no STRING com base no *score*, de modo que podemos ter informações bastante precisas e o mais próximas da realidade possível.

2.2 DEFINIÇÃO DO *SCORE* E CONSTRUÇÃO DAS REDES DE INTERESSE

Como o objetivo do trabalho é reproduzir a rede de interações entre proteínas de um organismo da maneira mais próxima da rede real do mesmo, optou-se por utilizar um intervalo de confiança (*score*) elevado para filtrar e obter os dados no STRING. Assim, o *score* escolhido após uma série de análises e comparações das informações e redes geradas com os diferentes valores, foi de 0.9. Desse modo, a rede de interações entre proteínas para cada organismo será formada na sua maior parte por interações que possuem um elevado grau de importância biológica, nos fornecendo informações precisas e o mais próximas da realidade possíveis.

Assim sendo, os organismos escolhidos para estudo no presente trabalho encontram-se dispostos no Quadro 2.1, junto com o tamanho da sua rede de interações entre proteínas para um *score* de 0.9. No Quadro 2.1 temos duas informações relevantes para cada organismo apresentado: na coluna N encontra-se a quantidade de proteínas da rede completa gerada para cada organismo e na coluna N_{CG} temos a quantidade de proteínas presente no Componente Gigante (CG) da rede de cada organismo. O CG nada mais é do que o maior componente conectado da rede, ou seja, cada proteína presente no CG está conectada a pelo menos uma outra no mesmo componente.

Observamos uma diferença no tamanho das redes quando consideramos a rede

completa ou somente o CG pois na rede completa é comum a presença de diversas redes menores, geralmente formadas por 2 ou 3 proteínas conectadas entre si somente. Assim, como o CG é formado, em média, por 80% a 90% das proteínas da rede, utiliza-se somente ele para calcular e obter as propriedades da rede de interações entre proteínas.

Quadro 2.1 – Tamanho da rede de proteínas (N) e do componente gigante do grafo (N_{CG}) de cada organismo estudado, para um *score* de 0.9.

Organismo	N	N_{CG}	N/N_{CG}
M. tuberculosis	1861	1231	0.65
E. coli	2791	1863	0.67
P. falciparum	2353	2111	0.90
B. subtilis	2881	2267	0.79
T. cruzi	3232	2958	0.92
S. pombe	3771	3606	0.96
A. mellifera	4218	3719	0.88
S. cerevisae	5213	4769	0.91
C. elegans	6170	5679	0.92
O. sativa	7000	6339	0.90
D. melanogaster	7003	6410	0.92
G. domesticus	7943	7258	0.91
X. laevis	8820	8279	0.94
A. thaliana	11385	9775	0.86
B. taurus	10771	10167	0.94
D. rerios	11426	10471	0.92
M. musculus	11141	10730	0.96
H. sapiens	12396	11919	0.96

Fonte: O autor.

No Apêndice A é possível visualizar o tamanho das redes e dos componentes gigantes para outros valores de *score* do STRING.

O STRING nos fornece uma lista de pares ordenados, onde cada par corresponde a uma interação entre duas proteínas. Sendo assim, a partir dessa lista de pares ordenados podemos construir o grafo que representa cada rede de interações entre proteínas. Para a construção dos grafos e a análise dos resultados, foi utilizado o *software* Mathematica® na versão 12.1. No Apêndice G podem ser vistos os algoritmos que foram utilizados para reproduzir os modelos que serão apresentados ao longo do presente trabalho.

No entanto, apesar de o grafo nos fornecer a estrutura completa da rede, não é de muita utilidade fazer uma análise baseando-se na sua estrutura visual. Para realizar uma análise mais detalhada das redes, utilizaremos as ferramentas matemáticas que foram explicadas nas seções anteriores.

2.2.1 Topologia das redes experimentais

Como já mencionado anteriormente, uma das maneiras mais fáceis de se verificar o comportamento de uma rede complexa e de analisar suas propriedades é através de sua topologia. A topologia de uma rede pode ser representada de duas formas: através da i) Função Distribuição de Probabilidade (ou *Probability Distribution Function*, PDF) ou da ii) Função de Distribuição Acumulada (*Cumulative Distribution Function, CDF*).

A PDF representa a probabilidade de que uma variável aleatória X assuma um determinado valor, ou seja, temos $P(X = k)$, onde k é um valor qualquer. Em outras palavras, a partir da PDF poderemos estimar qual a probabilidade P de encontrarmos um valor de X que seja exatamente igual à k. Se aplicarmos esse conceito nas redes de interações entre proteínas, através da PDF poderemos estimar qual a probabilidade de encontrar na rede um nó que tenha exatamente k *links*, isto é, a probabilidade de encontrar um nó com conectividade k na rede de interesse. Na Figura 1.11 temos um exemplo de como é a topologia de uma rede representada utilizando-se a PDF.

Podemos perceber que a PDF apresenta um comportamento bem característico na parte final da distribuição, isto é, na sua cauda. Essa região corresponde aos *hubs* que estão presentes na rede, que nada mais são do que nós com alta conectividade. Os *hubs* se encontram na cauda da PDF pois como são nós muito conectados, existem poucos com tal conectividade, de modo que a probabilidade de escolhermos aleatoriamente um nó da rede e este ser um *hub* é extremamente baixa.

Embora a PDF seja uma maneira aceitável de representar a topologia de uma rede, modelar essa curva não é uma tarefa simples principalmente em virtude desse platô que observamos na cauda da distribuição. Assim, uma forma de visualizar a topologia da rede de modo que possamos ter uma curva mais fácil de modelar é utilizando a CDF.

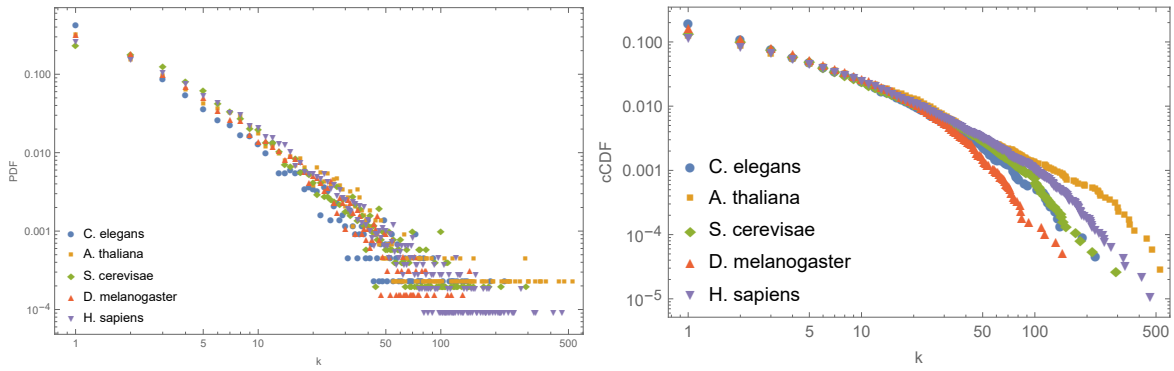
Na CDF, estamos interessados em estimar a probabilidade P de que uma variável aleatória qualquer X assuma um valor maior ou igual à k, isto é, $P(X \geq k)$. Traduzindo isso para as redes de proteínas, a CDF nos mostrará a probabilidade de encontrarmos na rede um nó que tenha uma conectividade maior ou igual à k. Uma das grandes vantagens da CDF é que com ela podemos eliminar o platô formado na cauda da PDF, de modo que teremos, assim, uma curva que pode ser facilmente modelada e analisada.

Na Figura 2.1 podemos visualizar melhor essa diferença entre a topologia representada pela PDF e pela CDF. Os dados utilizados na Figura 2.1 foram retirados de uma outra fonte de dados[2], pois o objetivo aqui é somente ilustrar as diferenças entre as duas formas de se analisar a topologia de uma rede.

Na Figura 2.1b podemos ver com clareza que ao utilizar a CDF, o platô que observamos na Figura 2.1a "desaparece", dando origem a uma curva que pode ser facilmente

[2]Os dados dessas redes em questão foram obtidos no banco de dados do *High-quality Interactomes*, desenvolvido e mantido pelo *Yu Lab* da Universidade de Cornell. Os dados estão disponíveis em http://hint.yulab.org/

Figura 2.1 – Representação da topologia de redes de proteínas utilizando a PDF e a CDF.

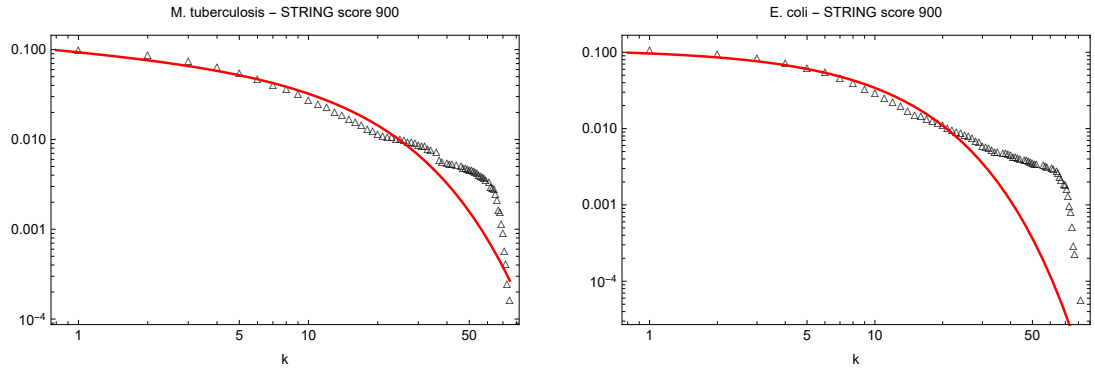


(a) *Função distribuição de probabilidades* (b) *Função de distribuição acumulada*

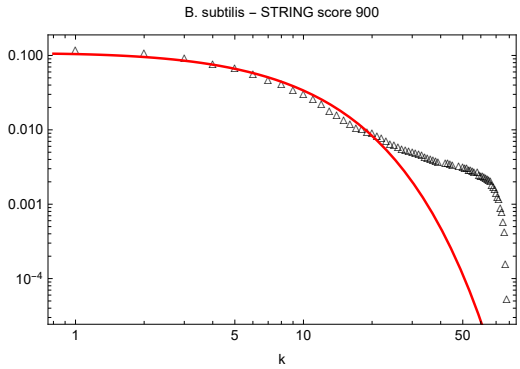
Fonte: O autor.

modelada. Assim, em virtude desse comportamento aqui apresentado, para a análise da topologia das redes que serão estudadas nesse trabalho foi utilizada somente a CDF, pois é a função que fornece melhores detalhes para analisarmos a topologia. Na Figura 2.2 podemos visualizar a CDF das redes de proteínas estudadas e vemos que, assim como era esperado, temos uma curva que pode ser modelada facilmente pois não existe o platô na cauda de sua distribuição. A topologia das demais redes reais, que não foram apresentadas na Figura 2.2 encontram-se na Figura B.1.

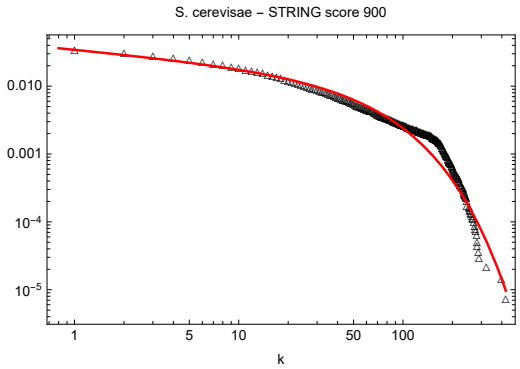
Figura 2.2 – Topologia das redes experimentais. Os triângulos pretos representam a curva da topologia da rede real, e a linha em vermelho é a curva que melhor representa esses dados



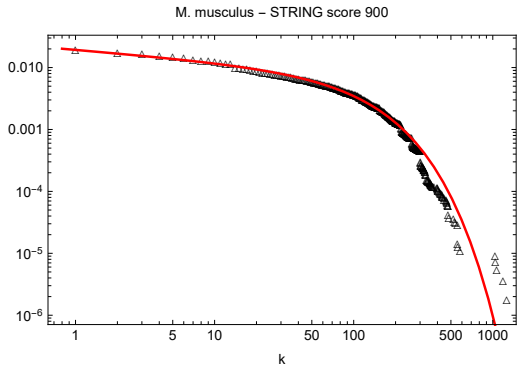
(a) *M. tuberculosis* (b) *E. coli*



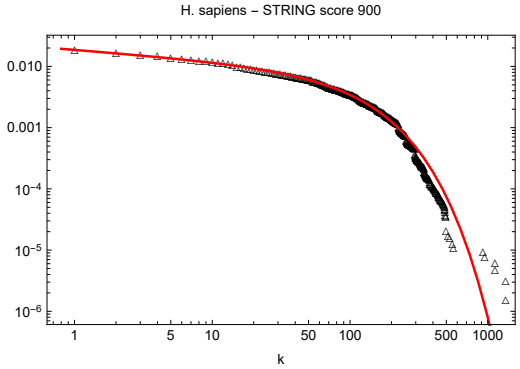
(c) *B. subtilis*



(d) *S. cerevisae*



(e) *M. musculus*



(f) *H. sapiens*

Fonte: O autor.

Algo que podemos observar na Figura 2.2 é que existe, de certa forma, um comportamento semelhante na forma da topologia de todas as redes, independente do seu tamanho e da origem do organismo em questão. Também é possível visualizar que essas redes não seguem uma lei de potência, pois a curva de suas topologias não se aproxima de uma reta em escala logarítmica. De fato, a topologia dessas redes parece seguir uma lei de potência com *cutoff* exponencial, isto é, a distribuição de probabilidades segue a seguinte expressão:

$$p_k \sim k^{-\gamma} e^{-\alpha k}, \tag{2.1}$$

onde γ e α são os parâmetros da distribuição. A curva vermelha presente em cada imagem da Figura 2.2 corresponde à função distribuição de probabilidades calculada para cada rede seguindo a forma dada pela equação 2.1. As curvas foram calculadas e reproduzidas utilizando a função *NonlinearModelFit*[3] do Mathematica®, onde foram obtidos para os parâmetros α e γ os valores apresentados na Tabela 2.1.

Podemos observar na Tabela 2.1 que o valor de γ para a grande maioria dos organismos se mantém próximo de 0.24, ao passo que o valor de α encontra-se numa faixa

[3]A partir de um conjunto de dados, forneido como uma lista de pares ordenados, e de uma equação matemática com parâmetros desconhecidos, retorna o valor dos parãmetros que resultam na equação que melhor representa os dados que estão sendo modelados

Tabela 2.1 – Relação de parâmetros da lei de potência com *cutoff* exponencial para cada organismo

Organismo	γ	α
M. tuberculosis	0.1964	0.0676
E. coli	0.0089	0.1131
P. falciparum	0.2680	0.0261
B. subtilis	0.0788	0.1456
T. cruzi	0.3051	0.0041
S. pombe	0.2472	0.0147
A. mellifera	0.2873	0.0448
S. cerevisae	0.2329	0.0160
C. elegans	0.2346	0.0127
O. sativa	0.2268	0.0056
D. melanogaster	0.2362	0.0094
G. domesticus	0.2026	0.0122
X. laevis	0.1748	0.0099
A. thaliana	0.3044	0.0046
B. taurus	0.1836	0.0091
D. rerios	0.2100	0.0080
M. musculus	0.1881	0.0086
H. sapiens	0.1771	0.0088

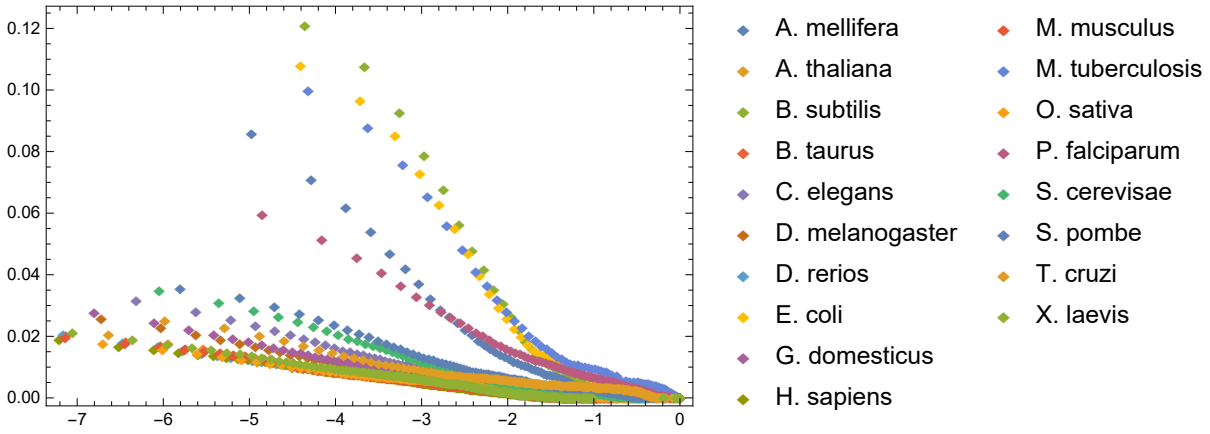
Fonte: O autor.

próxima de 0.01. Embora tenhamos valores próximos entre os organismos, vale destacar que por uma análise visual na Figura 2.2 vemos que para alguns organismos a curva gerada pelo *fit* da lei de potência com *cutoff* exponencial deixou de fora uma quantidade significativa de pontos.

Seria possível, no entanto, fazer um ajuste um pouco melhor nas curvas ao tentar ajustar manualmente as curvas, mas devido ao grande número de combinações possíveis para os valores de α e γ, optou-se por manter o que foi calculado por software e buscar novas formas de, posteriormente, comparar as redes experimentais com as redes geradas por diferentes modelos.

Assim, buscou-se uma outra forma de se comparar a topologia das redes reais. Para fazer essa comparação, no lugar de plotar a CDF das redes em função somente do valor da conecitividade k, a mesma foi plotada em função de $log(k/k_{max})$, onde k_{max} é o valor da conectividade máxima de cada rede. Ao fazer o gráfico de $p(k)$ *versus* $log(k/k_{max})$, é possível visualizar conjuntamente a topologia de todas as redes, e o resultado disso está na Figura 2.3, onde podemos visualizar a topologia de todas as redes de uma só vez.

Figura 2.3 – Topologia de todas as redes num gráfico de $p(k)$ *versus* $log(k/k_{max})$



Fonte: O autor.

Um resultado interessante que a Figura 2.3 nos mostra é que existem 3 regiões distintas no gráfico:

1. uma região superior, onde se encontram *E. coli*, *B. subtilis* e *M. tuberculosis*, que são 3 da menores redes e são as únicas bactérias dentre os organismos estudados;

2. uma região intermediária, composta por *A. mellifera* e *P. falciparum*;

3. e uma região inferior onde se encontram os demais organismos.

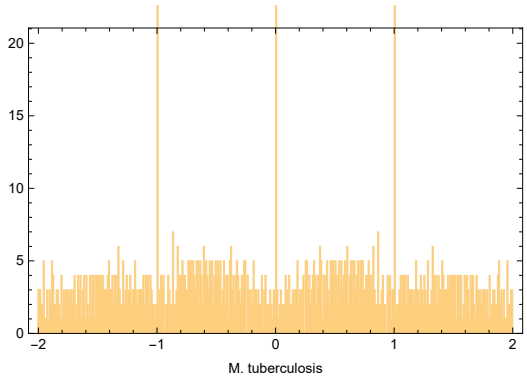
Além de termos essas regiões definidas no gráfico, é possível observar que as curvas que compõem cada região possuem um comportamento semelhante com as demais que estão naquela mesma região. O gráfico apresentado na Figura 2.3 servirá como base para a posterior comparação entre a rede experimental e as geradas por modelos.

2.2.2 Espectro de autovalores das redes reais

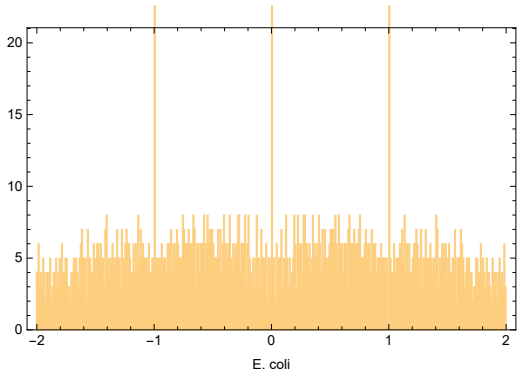
Como as redes de proteínas são representadas por grafos, e um grafo pode ser representado por uma matriz quadrada de ordem NxN onde N é o tamanho da rede, também podemos fazer uma análise do espectro de autovalores das redes para encontrar propriedades das mesmas. Assim, uma vez que as redes que estamos trabalhando são redes não-dirigidas, a sua matriz de adjacência será simétrica e não-negativa, e seus autovalores ($\{\lambda_j\}, j = 1, ..., N$) serão reais. Os autovalores são solução da equação $det(\mathbf{A} - \lambda\mathbf{I}) = 0$ (KAMP; CHRISTENSEN, 2005), onde $\mathbf{A}$ é a matriz de adjacência e $\mathbf{I}$ a matriz identidade de mesma ordem de $\mathbf{A}$.

Como podemos ver na Figura 2.4, apesar de os organismos serem de diferentes reinos biológicos, os seus espectros de autovalores possuem formatos parecidos e alguns pontos em comum que nos revelam informações importantes sobre a estrutura da sua rede de interação entre proteínas. O espectro de autovalores das demais redes reais encontra-se na Figura B.2.

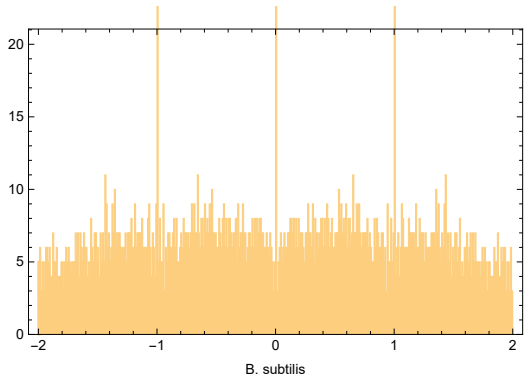
Figura 2.4 – Espectro de autovalores dos organismos estudados



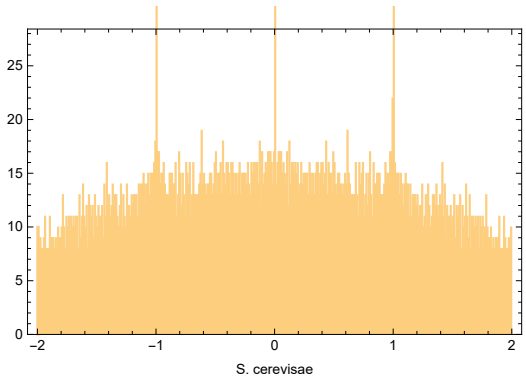
(a) *M. tuberculosis*



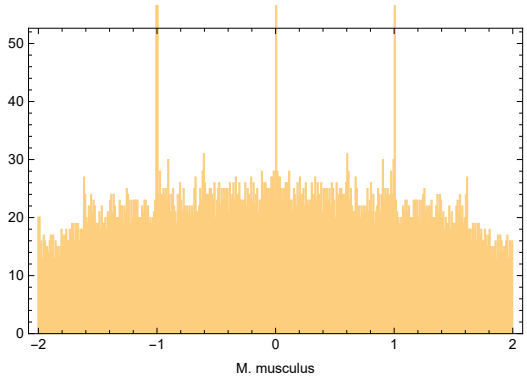
(b) *E. coli*



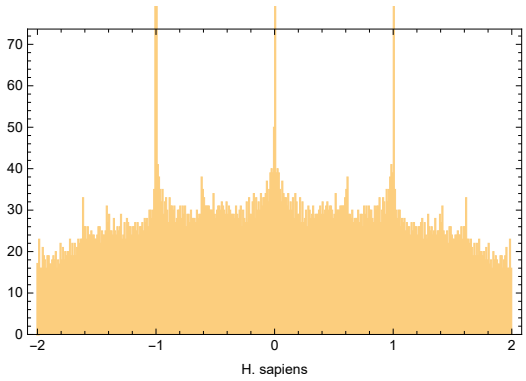
(c) *B. subtilis*



(d) *S. cerevisae*



(e) *M. musculus*



(f) *H. sapiens*

Fonte: O autor.

Os pontos que mais chamam a atenção são a presença de alguns picos de autovalores no espectro, mais especificamente os que vemos nos autovalores -1, 0 e 1, que estão presentes em praticamente todos os organismos. O pico em 0 nos revela que há

uma grande quantidade de nós duplicados na rede (YADAV; JALAN, 2015), pois a cada linha duplicada em uma matriz teremos um autovalor 0. E, como cada linha da matriz representa um nó e suas conexões, linhas iguais indicam a presença de proteínas que possuem exatamente as mesmas conexões na rede. Nós duplicados são aqueles que possuem exatamente os mesmos vizinhos. A degenerescência em 0 também é fruto de duplicações parciais, onde os nós possuem somente alguns vizinhos em comum. Assim, essa grande degenerescência em 0 indica que a duplicação é um fator importante durante o processo evolutivo de uma rede (RAI et al., 2017).

Além da degenerescência em 0, também podemos observar uma alta degenerescência nos autovalores 1 e -1, e essa degenerescência é observada em todos os organismos estudados. A degenerescência no autovalor -1 indica que a rede pode conter diversos sub-grafos completos dentro de si. Esses sub-grafos, por sua vez, são estruturas responsáveis por tornar a rede mais robusta e estável. Como as redes de interações entre proteínas são compostas por diversas sub-redes de ciclos metabólicos e caminhos que constituem um papel fundamental na estrutura da rede (RAI et al., 2017), a degenerescência em -1 pode estar relacionada diretamente à essa característica marcante das redes de interações entre proteínas.

Também podemos visualizar uma degenerescência no autovalor 1 que, embora não seja possível atribuir sua origem à uma estrutura específica dentro da rede (KAMP; CHRISTENSEN, 2005), acredita-se que esse autovalor também esteja relacionado à presença de estruturas na forma de *loop* dentro da rede. Pela simetria que pode ser vista no espectro de autovalores, tal explicação parece razoavél para a origem desses autovalores.

Embora essas evidências sejam importantes, de modo geral é uma tarefa complexa determinar a origem de uma degenerescência em um autovalor em específico (RAI; SHINDE; JALAN, 2018). No entanto, alguns estudos que focaram sua atenção nos autovetores associados à esses autovalores degenerados apresentaram novas evidências nas estruturas internas das redes que contribuem para a degenerescência (MARREC; JALAN, 2017). Entender a origem dessa degenerescência é importante pois os nós que participam dessas estruturas modulares, ou desses padrões nas redes, podem ser identificados através dos seus autovetores associados e são bons candidatos como alvos para a aplicação de drogas em doenças complexas (RAI; SHINDE; JALAN, 2018).

2.2.3 Probabilidade de duplicação das redes reais

Como mencionado na seção anterior, pelo espectro de autovalores vemos que a duplicação de nós é um fator extremamente presente na evolução e na estrutura da rede. Dessa forma, através da matriz de adjacência podemos obter uma estimativa da probabilidade de ocorrer a duplicação em cada rede. Tratamos aqui como uma estimativa pois

estaremos calculando somente a probabilidade de encontrar linhas exatamente iguais na matriz, uma vez que também é possível obter autovalores 0 quando uma linha da matriz for uma combinação linear de duas ou mais outras linhas, o que é chamado de duplicação parcial. No entanto, estimar a probabilidade de duplicação levando em conta essa possibilidade de ocorrerem duplicações parciais é um processo extremanente dispendioso computacionalmente (YADAV; JALAN, 2015).

Dessa forma, a melhor maneira de estimar a probabilidade de duplicação de uma rede a partir da sua matriz é por meio do posto da matriz. Na álgebra linear, o posto da matriz é definido como sendo o número de linhas (ou colunas) linearmente independentes na matriz. Em outras palavras, o posto pode ser entendido como o número de linhas não-nulas após o processo de escalonamento da matriz. Assim, através de uma operação algébrica simples, podemos estimar a probabilidade de duplicação em cada rede:

$$p_{dupl} = \frac{N - rank(A)}{N}, \tag{2.2}$$

onde p_{dupl} é a probabilidade de duplicação, N é o tamanho da rede ou a dimensão da matriz e $rank(A)$ é o posto da matriz de adjacência A. No Quadro 2.2 a seguir estão expostos os valores encontrados para p_{dupl} de cada rede, que serão tomados como referência na hora de fazer a comparação com as redes geradas pelos modelos estudados.

Quadro 2.2 – Estimativa da probabilidade de ocorrer a duplicação de nós em cada organismo

Organismo	p_{dupl}
M. tuberculosis	0.09227
E. coli	0.05421
P. falciparum	0.05306
B. subtilis	0.04676
T. cruzi	0.53482
S. pombe	0.03189
A. mellifera	0.12019
S. cerevisae	0.02852
C. elegans	0.06832
O. sativa	0.19514
D. melanogaster	0.07379
G. domesticus	0.09644
X. laevis	0.12656
A. thaliana	0.13995
B. taurus	0.15570
D. rerios	0.12368
M. musculus	0.15023
H. sapiens	0.12023

Fonte: O autor.

Pelo Quadro 2.2 podemos ver que para a maior parte dos organismos a probabilidade de ocorrer a duplicação é relativamente baixa, o que, de certa forma contrasta com a grande quantidade de autovalores 0 que vemos no espectro de autovalores para esses organismos. No entanto, é importante ressaltar novamente que conseguimos calcular somente uma estimativa para que essa duplicação ocorra, o que pode explicar essa grande diferença entre o que encontramos e o que esperávamos encontrar a partir do espectro de autovalores.

2.3 MODELANDO AS REDES REAIS COM O MODELO *DUPLICATION-DIVERGENCE*

Como o foco do presente trabalho é a modelagem de redes de interações entre proteínas, optou-se por utilizar como base de comparação para o desenvolvimento de um novo modelo somente as redes geradas pelo modelo DD, uma vez que este modelo foi desenvolvido especificamente para redes de interações entre proteínas[4].

Assim, não foram realizadas simulações utilizando-se como base o modelo de BA pois embora seja um modelo que foi desenvolvido pensando em reproduzir qualquer tipo de rede, já vimos que a topologia das redes geradas por ele difere significativamente da topologia das redes de interações entre proteínas.

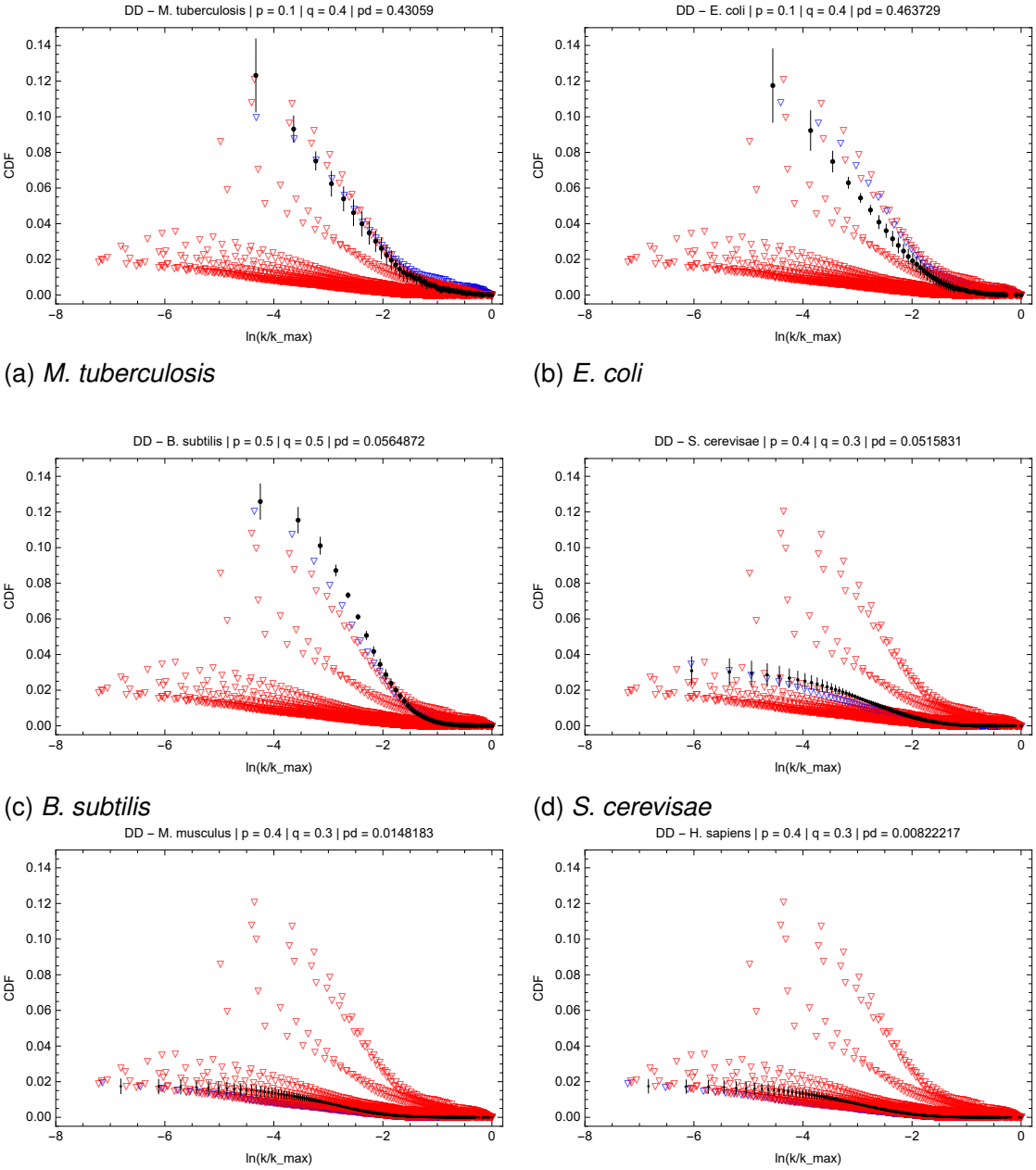
O modelo DD, além do tamanho final da rede (N), necessita de dois parâmetros para ser executado: p e q. Como torna-se inviável testar todas as possíveis combinações desses parâmetros, as simulações foram realizadas utilizando-se valores nos intervalos de $0.1 \leq p \leq 0.9$ e $0.1 \leq q \leq 0.9$ com um incremento de 0.1 em cada parâmetro, o que resultou em 81 possíveis combinações desses parâmetros para cada rede estudada.

Para cada uma dessas 81 combinações, foram feitas 10 realizações da rede, de modo que fosse possível obter um resultado mais significativo. Além disso, obteve-se a partir das simulações a topologia das redes, o espectro de autovalores e a estimativa para a probabilidade de ocorrer a duplicação em cada rede.

Na Figura 2.5 podemos visualizar a topologia das redes geradas pelo modelo DD e comparar a mesma com a topologia da rede experimental. Ao fundo, representadas pelas curvas com os triângulos em vermelhos temos a topologia de cada uma das redes experimentais, e em destaque, com triângulos azuis, a topologia da rede que está sendo simulada. Por fim, os pontos em preto representam a topologia da rede gerada pelo modelo DD. As barras de erro foram determinadas após fazer a média para 10 realizações de cada rede. A topologia das demais redes geradas pelo modelo DD encontra-se na Figura D.1.

[4]Como o modelo de Barabási e Albert não é exclusivo para redes de proteínas, o mesmo não será utilizado no trabalho para comparação. No entanto, no Apêndice C é possível visualizar a topologia (Figura C.1) e o espectro de autovalores (Figura C.2) das redes geradas por ele

Figura 2.5 – Topologia das redes geradas pelo modelo *Duplication-Divergence*. Os triângulos vermelhos representam as curvas da topologia das redes reais e a curva composta pelos triângulos azuis destaca a topologia da rede experimental do organismo em questão. A curva formada pelos pontos e pelas barras de erro em preto representa a topologia da rede gerada pelo modelo DD para o conjunto de parâmetros descrito no gráfico.



(a) *M. tuberculosis*

(b) *E. coli*

(c) *B. subtilis*

(d) *S. cerevisae*

(e) *M. musculus*

(f) *H. sapiens*

Fonte: O autor.

Pelos resultados apresentados na Figura 2.5, podemos observar que o modelo DD consegue capturar com bastante precisão a topologia das redes experimentais, uma vez que as curvas geradas pelo modelo se aproximam bastante da curva experimental. Além disso, podemos ver que para a grande maioria dos organismos, os valores de p e q que resultaram na melhor topologia foram bem próximos, mantendo-se na média de 0.3 e 0.4

tanto para p quanto para q.

No entanto, como apresentamos mais cedo, a topologia não é a única ferramenta que temos para comparar os resultados gerados pelo modelo. Na Figura 2.5 podemos ver no título de cada gráfico o valor da propriedade pd, que representa a probabilidade de duplicação calculada para cada uma das redes. Assim, podemos comparar esse valor com o valor apresentado no Quadro 2.2, e essa comparação pode ser observada no Quadro 2.3 a seguir.

Quadro 2.3 – Estimativa da probabilidade de ocorrer a duplicação de nós nas redes geradas pelo modelo DD para cada organismo

Organismo	$p_{dupl,exp}$	$p_{dupl,DD}$	Erro (%)
M. tuberculosis	0.09227	0.43059	366.7
E. coli	0.05421	0.46373	755.4
P. falciparum	0.05306	0.01990	62.5
B. subtilis	0.04676	0.05649	20.8
T. cruzi	0.53482	0.00778	98.5
S. pombe	0.03189	0.03134	1.7
A. mellifera	0.12019	0.12261	2.0
S. cerevisae	0.02852	0.05158	80.8
C. elegans	0.06832	0.04332	36.6
O. sativa	0.19514	0.46939	140.5
D. melanogaster	0.07379	0.39610	436.8
G. domesticus	0.09644	0.01116	88.4
X. laevis	0.12656	0.03246	74.3
A. thaliana	0.13995	0.09493	32.2
B. taurus	0.15570	0.01482	90.5
D. rerios	0.12368	0.00822	93.4
M. musculus	0.15023	0.01482	90.1
H. sapiens	0.12023	0.00822	93.2

Fonte: O autor.

Pelo Quadro 2.3 podemos observar que apesar de a topologia das redes geradas pelo modelo DD se aproximarem bastante da topologia das redes reais (Figura 2.5), a probabilidade de ocorrer a duplicação nessas redes apresenta um grande percentual de erro quando comparada com as redes experimentais. A taxa de erro foi calculada através da seguinte expressão:

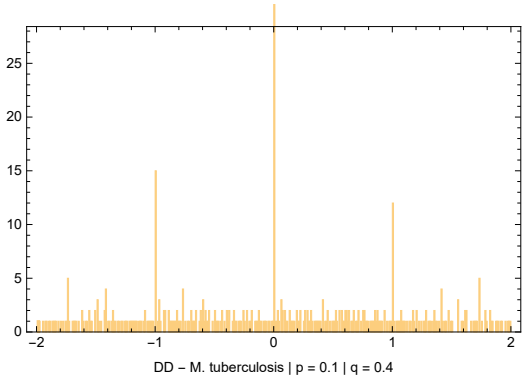
$$erro(\%) = \left(\frac{|p_{dupl,exp} - p_{dupl,DD}|}{p_{dupl,exp}} \right) * 100, \tag{2.3}$$

resultando nos valores apresentados na última coluna do Quadro 2.3.

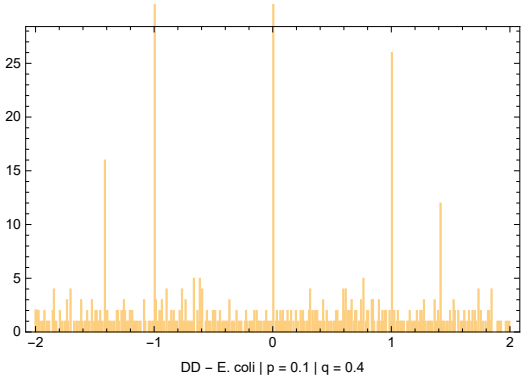
Além de compararmos a topologia e a probabilidade de ocorrer duplicação, também é necessário comparar o espectro de autovalores das redes geradas pelo modelo DD com o que foi apresentado na Figura 2.4. Na Figura 2.6 podemos observar o espectro de

autovalores das redes geradas pelo modelo DD para os mesmos parâmetros em que a topologia foi apresentada. Na Figura D.2 encontra-se o espectro de autovalores para as demais redes que não foram apresentadas aqui na Figura 2.6.

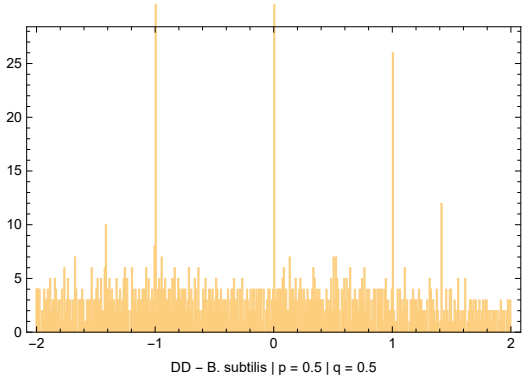
Figura 2.6 – Espectro de autovalores das redes geradas pelo modelo *Duplication-Divergence*



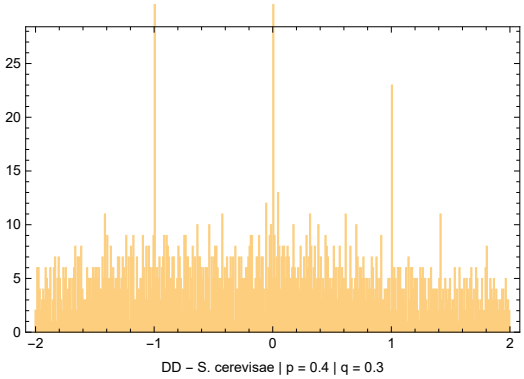
(a) *M. tuberculosis*



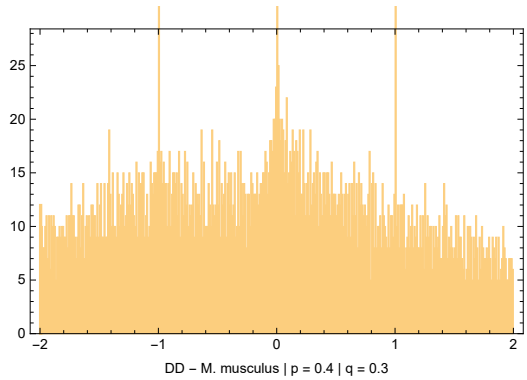
(b) *E. coli*



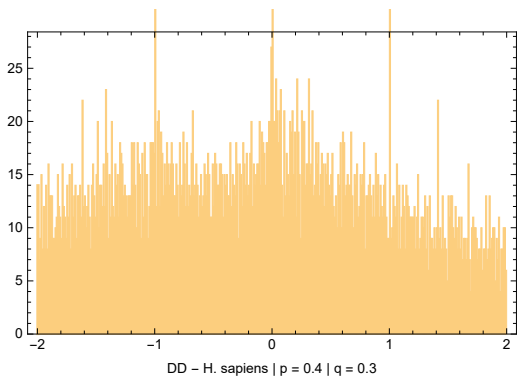
(c) *B. subtilis*



(d) *S. cerevisae*



(e) *M. musculus*



(f) *H. sapiens*

Fonte: O autor.

De primeira vista já percebe-se que, salvo algumas exceções, os espectros de autovalores das redes geradas pelo modelo DD apresentam os mesmos picos que podem ser observados nas redes reais, conforme Figura 2.4. No entanto, é possível ver que esses picos são menos proeminentes do que aqueles que encontramos nas redes experimentais. Além disso, parece ocorrer um maior espalhamento dos autovalores, de modo que a concentração dos mesmos na região de interesse seja bem menor do que o esperado. Também é possível perceber a falta de simetria que era esperada para o espectro de autovalores das redes reais.

Assim, com o que foi apresentado nas Figuras 2.5 e 2.6 e no Quadro 2.3, podemos observar que apesar de o modelo DD representar com acurácia significativa a topologia das redes, o mesmo falha em capturar as demais propriedades dessas redes, o que sugere a existência de outros mecanismos por trás da evolução das redes de interações entre proteínas.

2.4 DESENVOLVIMENTO DE UM NOVO MODELO PARA REPRODUZIR REDES DE INTERAÇÕES ENTRE PROTEÍNAS

Apesar de os resultados obtidos para o modelo DD serem promissores, como já destacamos, as redes geradas por ele só são capazes de reproduzir a topologia das redes reais, deixando de lado aspectos como o espectro de autovalores e a probabilidade de encontrar nós duplicados na rede. Essa observação nos mostra que apesar de a duplicação ser um fator importante na evolução de uma rede de interações entre proteínas, deve haver algum outro mecanismo por trás desses processo que, em conjunto com a duplicação de nós, irá gerar redes o mais próximo possível das redes reais.

Assim, como a duplicação de nós é um dos processos por trás do processo evolutivo e como os resultados encontrados para o modelo DD foram satisfatórios em alguns aspectos, optou-se por utilizar no novo modelo o mesmo mecanismo de duplicação que foi proposto por Vázquez (VÁZQUEZ et al., 2003).

Definido um dos mecanismos que serão utilizados no desenvolvimento de um novo modelo, é preciso definir como será o segundo mecanismo e como ocorrerá essa "escolha" por um ou por outro mecanismo para cada novo nó que será adicionado à rede.

Como já mencionado anteriormente, o processo de Transferência Horizontal de Genes (TH) é responsável pela aquisição de novos nós (proteínas) que acaba resultando em novas informações genéticas (TREANGEN; ROCHA, 2011), e pode ser o responsável por algumas das propriedades que encontramos ao analisar a topologia e o espectro de autovalores das redes de interações entre proteínas.

Dessa forma, apesar de termos definidos quais serão os dois mecanismos que serão utilizados no modelo, ainda falta definir como vamos reproduzir a TH e como será

feita a escolha entre esses mecanismos. Algumas evidências sugerem que a TH teria um papel mais predominamente do que a duplicação de genes na evolução da rede de proteína de argumentos, o que resultaria em um modelo que daria maior preferência para a TH do que a duplicação. Assim, optou-se por utilizar diferentes probabilidades de ocorrer a TH com o objetivo de verificar o papel desse mecanismo na evolução das redes que estão sendo estudadas. Serão realizadas simulações com as seguintes probabilidades de ocorrer a TH: 0.01, 0.05, 0.10, 0.25 e 0.50, de modo que será possível observar o impacto da mesma na evolução das redes.

Definido como será feita a escolha entre os dois mecanismos, ainda é necessário definir como será implementada dentro do modelo a TH. Como trata-se de um processo de aquisição de novos nós, isto é, será adicionado um nó completamente novo na rede (sem ser uma cópia de algum já existente), a primeira possibilidade que surge é de utilizar ao lado do modelo DD, o modelo BA, uma vez que este incorpora o crescimento das redes no seu algoritmo através da criação de um nó completamente novo e diferente do que já existe na rede naquele passo. No entanto, foi feita uma pequena modificação na forma como o modelo BA incorpora o novo nó na rede. Enquanto que no modelo BA cada novo nó se conectará sempre a um número fixo m de nós já existentes na rede, adicionou-se uma pequena modificação nessa parte ao permitir que cada novo nó que será adicionado por esse mecanismo possa se conectar a um valor diferente de m, isto é, um nó pode se conectar a $m = 2$ nós já existentes na rede, enquanto que outro nó pode se conectar a $m = 5$ nós já existentes, por exemplo. Essa alteração foi feita com o objetivo de reduzir a possibilidade de que a distribuição de probabilidades assuma uma lei de potência pura e também para favorecer a criação de nós com uma baixa conectividade na rede.

Assim, o novo modelo é composto por dois mecanismos: duplicação e transferência horizontal, e a probabilidade de ocorrer cada um será controlado por um parâmetro d, que poderá assumir os valores definidos acima (0.01, 0.05, 0.10, 0.25 e 0.50). Temos, então, um novo modelo que engloba os seguintes mecanismos:

i) com probabilidade $(1-d)$, **duplicação**: o mecanismo de duplicação será o mesmo definido pelo modelo *Duplication-Divergence*, e ocorrerá com as mesmas etapas (duplicação e mutação), e serão utilizados os mesmos parâmetros p e q do modelo original;

ii) com probabilidade d, **transferência horizontal**: esse mecanismo funcionará de maneira semelhante ao modelo de Barabási e Albert, mas aqui, ao invés de cada novo nó se conectar sempre ao mesmo valor de m nós já existentes na rede, m será um parâmetro do modelo que será gerado aleatoriamente para cada novo nó que for adicionado na rede. Ainda será considerado, entretanto, o *Preferential Attachment* para fazer a escolha desses m nós aos quais o novo nó se conectará.

E para poder gerar as redes utilizando esse modelo, precisamos de 4 parâmetros:

- d: que é o parâmetro responsável por determinar qual mecanismo será seguido, e aqui serão utilizados os 5 valores já mencionados anteriormente: 0.01, 0.05, 0.10, 0.25 e 0.50;

- p: que terá a mesma função que é definida no modelo DD, e assumirá valores entre $0.1 \leq p \leq 0.9$, com incrementos de 0.1;

- q: assim como p, terá o mesmo significado que foi apresentado no modelo DD e também assumirá os valores $0.1 \leq q \leq 0.9$, com incrementos de 0.1;

- m: que representará a quantidade máxima de *links* que um novo nó poderá ter quando adicionado via transferência horizontal, e para fins de simulações serão considerados valores inteiros entre 1 e 5;

onde teremos 2025 combinações possíveis de parâmetros para analisar após as simulações. Foram realizadas simulações com todas essas combinações e para cada conjunto de parâmetros, foram feitas 10 realizações da rede, de modo que fosse possível obter um resultado mais significativo. Posteriormente foram analisados visualmente os resultados obtidos, de modo que fosse possível obter os melhores resultados, que serão apresentados no capítulo a seguir. O algoritmo desenvolvido e utilizado para gerar as redes está presente no Apêndice G.

3 RESULTADOS E DISCUSSÃO

São necessários 4 parâmetros diferentes para gerar as redes utilizando o novo modelo. Seria interessante se fosse possível determinar experimentalmente algum dos parâmetros necessários, o que reduziria significativamente a quantidade de combinações que precisam ser feitas, bem como poderia produzir resultados mais realistas. No entanto, não encontramos evidências que sugiram maneiras de calcular cada parâmetro.

Mensurar e estimar os valores de p e q é um tarefa bastante complexa e como vimos pelos resultados das redes geradas pelo modelo DD, apesar de encontrar valores próximos, cada organismo possui a sua própria combinação desses valores, o que não possibilita a definição de um valor global para nenhum desses parâmetros. O único parâmetro que talvez seja possível estimar experimentalmente é o parâmetro m, que no caso do modelo BA é dado pela razão entre o número de *links* e o número de nós na rede, isto é:

$$m \approx \frac{L}{N}, \tag{3.1}$$

onde L é o número de *links* da rede e N o número de nós. Na Tabela 3.1 podemos visualizar os valores calculados de m para cada um dos organismos. No entanto, simulações iniciais logo revelaram que ao aumentarmos significativamente o valor de m, a topologia e o espectro da rede começava a divergir rapidamente do que era esperado. Sendo assim, optou-se por delimitar os valores de m no intervalo de 1 até 5, já explicitado anteriormente.

Por fim, sobre o valor de d, isto é, da probabilidade de ocorrer a transferência horizontal ou a duplicação, também não foi possível encontrar fontes que fornecessem uma maneira de calcular essa probabilidade, de modo que para facilitar as simulações foram considerados os valores já apresentados.

Tabela 3.1 – Valor estimado do parâmetro m para cada organismo, calculado a partir da quantidade de *links* e nós de cada rede

Organismo	m
M. tuberculosis	5
E. coli	5
P. falciparum	8
B. subtilis	4
T. cruzi	25
S. pombe	14
A. mellifera	6
S. cerevisae	15
C. elegans	16
O. sativa	30
D. melanogaster	20
G. domesticus	19
X. laevis	24
A. thaliana	25
B. taurus	26
D. rerios	25
M. musculus	27
H. sapiens	27

Fonte: O autor.

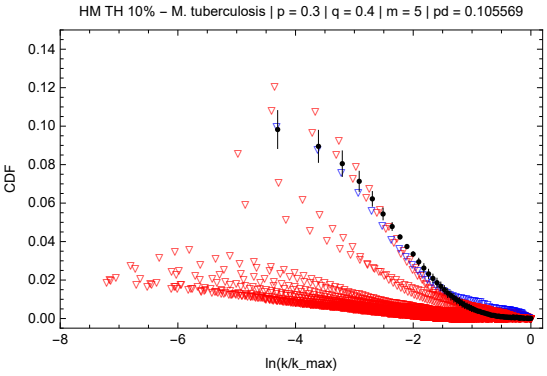
Nas seções a seguir apresentaremos os resultados encontrados para as redes geradas pelo novo modelo, e ao longo da apresentação dos resultados traremos algumas discussões pertinentes sobre os mesmos.

3.1 TOPOLOGIA DAS REDES GERADAS PELO NOVO MODELO

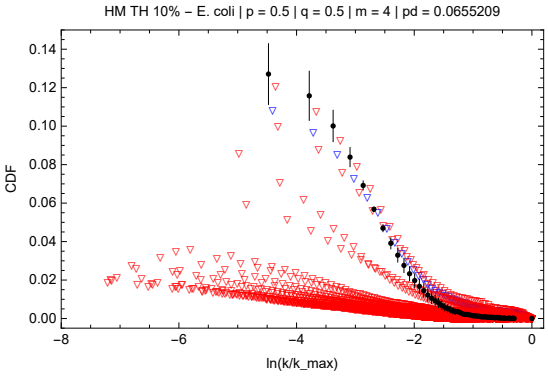
Na Figura 3.1 encontram-se os resultados obtidos para a topologia das redes geradas pelo novo modelo. Aqui temos, novamente, representada pelas pontos pretos com as barras de erro a topologia da rede gerada pelo modelo, a curva com os triângulos azuis representa a topologia da rede real questão e as demais curvas com triângulos vermelhos representam a topologia das outras redes reais. Os demais resultados para a topologia das redes geradas pelo novo modelo encontram-se na Figura E.1.

Analisando-se unicamente a topologia das redes, podemos observar que os resultados gerados por esse modelo são muito próximos da topologia das redes reais. Destaca-se, também, a similaridade dessas curvas com as curvas geradas pelo modelo DD na Figura 2.5, uma vez que os novos resultados aqui apresentados são resultados em que a transferência horizontal esteve presente em, no máximo, 10% dos novos nós que foram

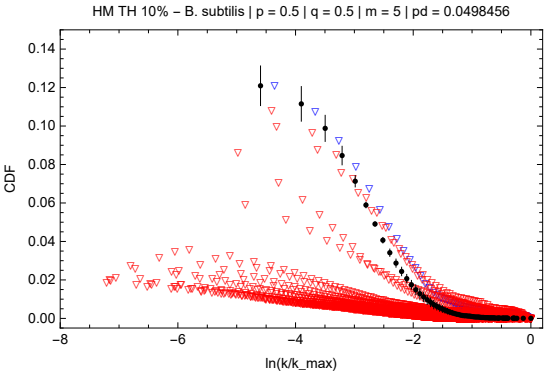
Figura 3.1 – Topologia das redes geradas pelo novo modelo desenvolvido. Os triângulos vermelhos representam as curvas da topologia das redes reais e a curva composta pelos triângulos azuis destaca a topologia da rede experimental do organismo em questão. A curva formada pelos pontos e pelas barras de erro em preto representa a topologia da rede gerada pelo novo modelo para o conjunto de parâmetros descrito no gráfico.



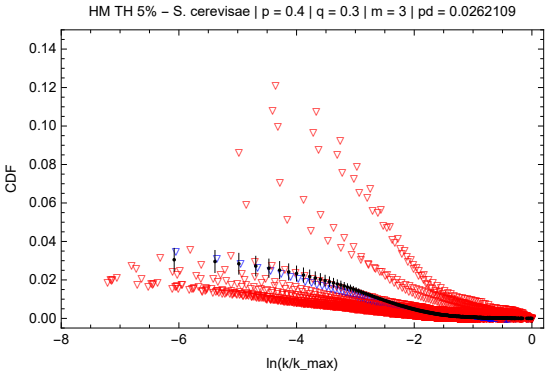
(a) *M. tuberculosis*



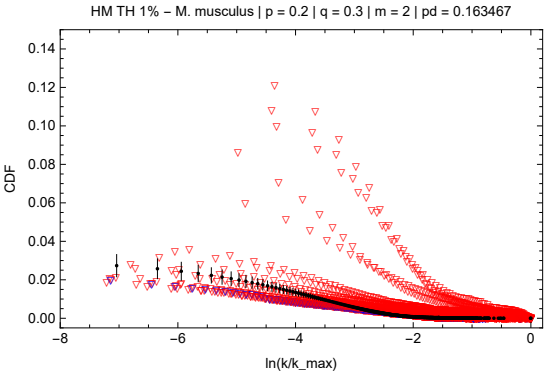
(b) *E. coli*



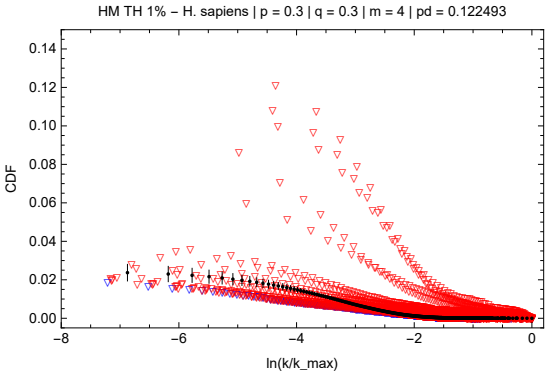
(c) *B. subtilis*



(d) *S. cerevisae*



(e) *M. musculus*



(f) *H. sapiens*

Fonte: O autor.

adicionados às redes. Assim, há uma predominância da duplicação de nós no resultado final, reforçando o fato de que a duplicação é um fenômeno importante na evolução de uma rede de interações entre proteínas.

Também é importante destacar que para as redes do reino *Monera*, além de todas pertencerem à mesma região no gráfico apresentado na Figura 2.3, podemos ver pelas Figuras 3.1a, 3.1b e 3.1d que para as três redes, a probabilidade de ocorrer a transferência horizontal foi a mesma: 10%, além do fato de os valores dos demais parâmetros não variar muito entre uma rede e outra.

Para as demais redes, infelizmente não foi possível observar esse comportamento em comum entre as redes de um mesmo reino. Vale destacar, no entanto, que salvo algumas poucas exceções, para os organismos que não pertencem ao reino *Monera*, o valor do parâmetro q manteve-se quase constante em $q = 0.3$ (variando nesses casos de exceção em 0.1 para mais ou para menos, mas nunca indo além dessa variação). Embora pareça pouco significativo, conseguir encontrar um parâmetro cujo valor mantenha-se quase sempre constante para diferentes redes é importante pois ajuda a reduzir a quantidade de combinações possíveis para determinar o melhor conjunto de parâmetros que representa aquelas redes.

Os parâmetros p e m apresentaram uma grande taxa de variação entre os organismos, não sendo possível determinar um comportamento universal para os mesmos. No entanto, é importante poder variar esses parâmetros pois assim é possível controlar a topologia e as propriedades das redes resultantes do novo modelo.

Com exceção das redes menores (redes do reino *Monera* e a rede do *P. falciparum*, a transferência horizontal parece ter pouca contribuição na topologia das redes, pois a probabilidade de ocorrer a TH ficou entre 1% e 5%, sendo que o valor que dominou foi 1%. No entanto, apesar de a chance de ocorrer a TH ser pequena, ela ainda está presente e parece influenciar diretamente nos resultados quando levamos em conta a probabilidade de duplicação de cada rede, que será discutida na próxima seção.

3.2 COMPARANDO A PROBABILIDADE DE DUPLICAÇÃO DAS REDES GERADAS PELO NOVO MODELO COM AS REDES EXPERIMENTAIS

Assim como para as redes geradas pelo modelo DD, para as redes geradas pelo novo modelo foi estimada a probabilidade de ocorrer a duplicação de nós, e o resultado encontra-se no Quadro 3.1, com a comparação entre o valor da rede real e o valor obtido para a rede gerada pelo modelo. Na última coluna temos o erro percentual, que foi obtido utilizando a Equação 2.3. Numa primeira olhada no Quadro 3.1 já é possível perceber que para as redes geradas pelo novo modelo temos um percentual de erro bem menor no valor encontrado para o probabilidade de duplicação quando comparado com o que foi obtido

para o modelo DD e apresentado no Quadro 2.3.

Quadro 3.1 – Estimativa da probabilidade de ocorrer a duplicação de nós em cada organismo para as redes geradas pelo novo modelo

Organismo	$p_{dupl,exp}$	$p_{dupl,HM}$	Erro (%)
M. tuberculosis	0.09227	0.10557	14.4
E. coli	0.05421	0.06552	20.9
P. falciparum	0.05306	0.05590	5.4
B. subtilis	0.04676	0.04985	6.6
T. cruzi	0.53482	0.48749	8.8
S. pombe	0.03189	0.03134	1.7
A. mellifera	0.12019	0.13606	13.2
S. cerevisae	0.02852	0.02621	8.1
C. elegans	0.06832	0.07624	11.6
O. sativa	0.19514	0.16958	13.1
D. melanogaster	0.07379	0.07176	2.8
G. domesticus	0.09644	0.10609	10.0
X. laevis	0.12656	0.14313	13.1
A. thaliana	0.13995	0.11478	18.0
B. taurus	0.15570	0.14262	8.4
D. rerios	0.12368	0.09494	23.2
M. musculus	0.15023	0.16347	8.8
H. sapiens	0.12023	0.12249	1.9

Fonte: O autor.

Como já mencionado na seção anterior, apesar de a probabilidade de ocorrer a TH ser baixa e a duplicação ser o fator evolutivo predominante nas redes geradas pelo modelo desenvolvido, quando analisamos os valores obtidos para a probabilidade de duplicação dessas redes podemos ver a influência que a TH teve pois foi possível encontrar valores bem próximos daqueles encontrados para as redes reais.

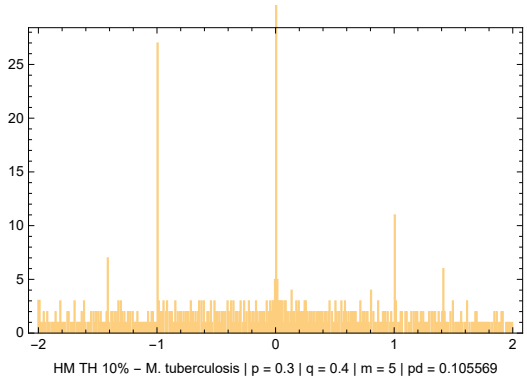
Além da topologia e da probabilidade de duplicação, também é preciso analisar o espectro de autovalores das redes geradas. Os resultados encontrados para o espectro de autovalores serão apresentados na seção a seguir.

3.3 ESPECTRO DE AUTOVALORES DAS REDES GERADAS PELO NOVO MODELO

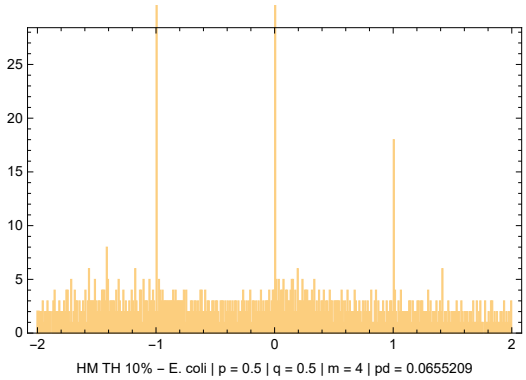
A última propriedade que levamos em consideração na hora de analisar as redes de interações entre proteínas é o espectro de autovalores das mesmas. Na Figura 3.2 podemos visualizar o espectro de autovalores para cada rede gerada pelo modelo que foi desenvolvido. Na Figura E.2 é possível visualizar o espectro de autovalores das demais redes.

Aqui, apesar de os resultados estarem bem parecidos com o que foi apresentado na Figura 2.6 para as redes geradas pelo modelo DD, podemos observar algumas diferenças, tais como uma maior predominância de autovalores no intervalo analisado, picos de interesse mais marcados e uma presença um pouco maior de simetria no espectro, algo que é visto no espectro das redes reais e apresentado na Figura 2.4.

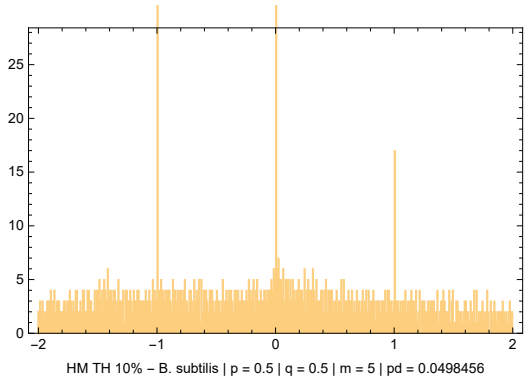
Figura 3.2 – Espectro de autovalores das redes geradas pelo novo modelo



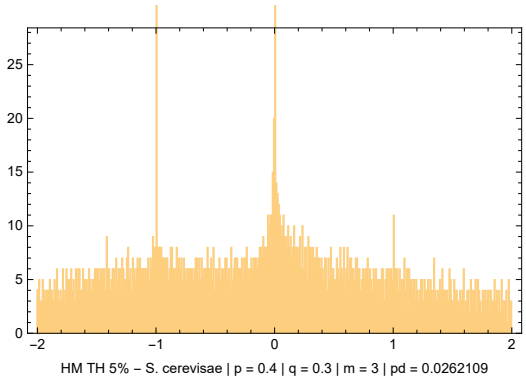
(a) *M. tuberculosis*



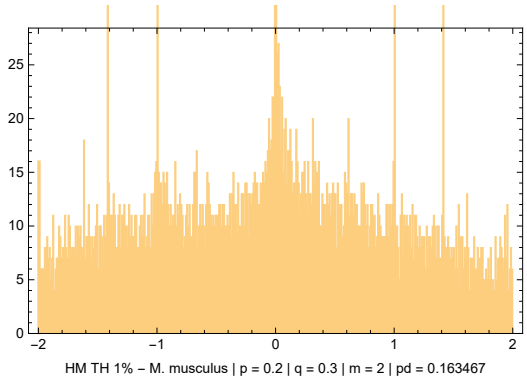
(b) *E. coli*



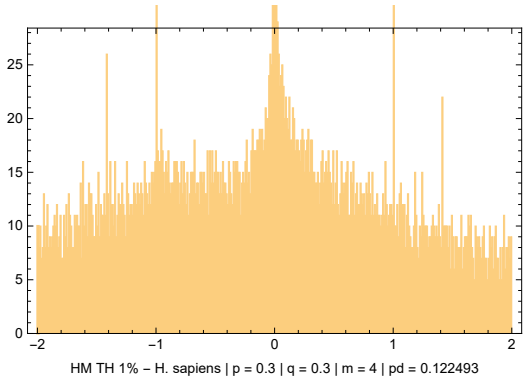
(c) *B. subtilis*



(d) *S. cerevisae*



(e) *M. musculus*



(f) *H. sapiens*

Fonte: O autor.

É importante destacar que em uma comparação direta com os espectros de autovalores das redes reais (Figura 2.4), os resultados obtidos aqui não são muito satisfatórios ao falharem em capturar todas as propriedades de interesse que estão presentes no espectro das redes reais. Mesmo assim, ao analisar em conjunto com as demais propriedades que são relevantes para compreender e reproduzir as redes de interação entre proteínas, podemos dizer que o modelo que foi desenvolvido é capaz de reproduzir com grande fidelidade as redes reais.

Além de analisar as propriedades de todas as redes, serão analisados os efeitos da transferência horizontal na topologia e no espectro de algumas redes, de modo que seja possível compreender um pouco melhor esse mecanismo. Na seção a seguir serão apresentados os resultados para as redes do reino *Monera*, uma vez que essas redes apresentam propriedades e comportamentos similares entre si.

3.4 ANÁLISE DO IMPACTO DA TRANSFERÊNCIA HORIZONTAL EM REDES DOS ORGANISMOS PERTENCENTES AO REINO *MONERA*

Como apresentado nos resultados anteriores, apesar de pouco presente, a transferência horizontal possui um papel importante na definição de algumas propriedades das redes de interações entre proteínas, principalmente quando olhamos para as propriedades internas das redes, como é o caso da probabilidade de duplicação.

Por isso, traremos uma breve discussão sobre a influência desse mecanismo em algumas redes. Foram escolhidos as redes do reino *Monera* pois temos 3 redes que além de serem de organismos que pertencem ao mesmo reino, apresentam tamanho similares, topologias parecidas e alguns parâmetros em comuns ao serem geradas com o novo modelo.

Para fazer essa análise do impacto da TH sobre cada rede, os demais parâmetros necessários para reproduzir a rede pelo novo modelo foram mantidos constantes, utilizando-se os valores apresentados na Figura 3.1, de modo que temos os seguintes parâmetros para as redes em questão:

- *M. tuberculosis*: $p = 0.3$, $q = 0.4$ e $m = 5$;
- *E. coli*: $p = 0.5$, $q = 0.5$ e $m = 4$;
- *B. subtilis*: $p = 0.5$, $q = 0.5$ e $m = 5$;

3.4.1 Efeito da transferência horizontal sobre a topologia

O primeiro aspecto que iremos analisar é como a transferência horizontal influencia na topologia de uma rede. Nas Figuras 3.3, F.1 e F.2 podemos visualizar o efeito da variação na probabilidade de ocorrer a transferência horizontal nas redes dos organismos já mencionados. Nas figuras temos novamente representada pelos pontos pretos a curva que corresponde à topologia das redes geradas pelo novo modelo, nos triângulos azuis a curva da topologia da rede real em questão e as demais curvas formadas pelos triângulos vermelhos representam a topologia das demais redes experimentais.

Pelas Figuras 3.3, F.1 e F.2 podemos observar que conforme variamos a probabilidade de ocorrer a TH, ocorre um deslocamento horizontal na topologia da rede gerada pelo novo modelo, de modo que ela se afasta ou se aproxima mais da topologia da rede real. Além disso, também é possível visualizar que ocorrem leves mudanças no formato da curva da topologia gerada pelo modelo, onde a TH acaba fazendo com que essa curva se aproxime de uma forma de "S" conforme é aumentada a sua probabilidade de ocorrência.

Também podemos observar que ocorre uma variação significativa na probabilidade de duplicação dessas redes conforme vamos variando a probabilidade de ocorrência da TH, e podemos acompanhar essa variação pela Tabela 3.2.

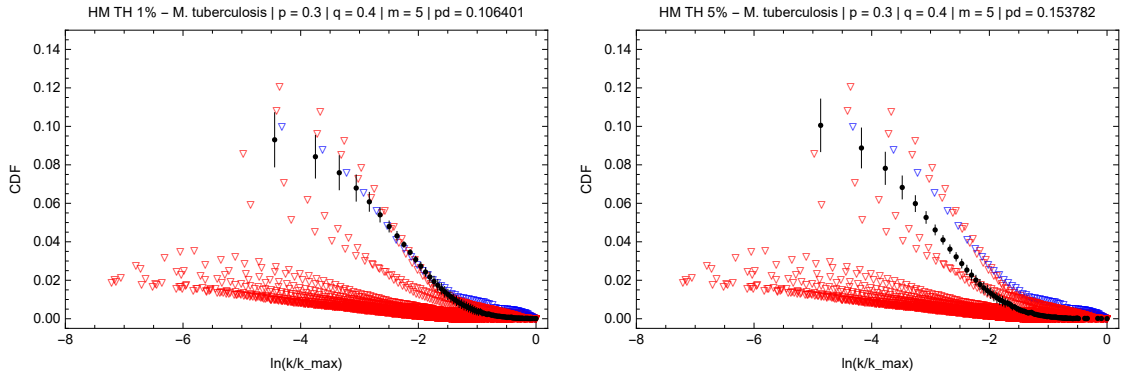
Tabela 3.2 – Probabilidade de duplicação estimada para cada um os organismos do reino *Monera* de acordo com a probabilidade de ocorrer transferência horizontal

Organismo	p_{dupl}	$p_{dupl}^{TH=0.01}$	$p_{dupl}^{TH=0.05}$	$p_{dupl}^{TH=0.10}$	$p_{dupl}^{TH=0.25}$	$p_{dupl}^{TH=0.50}$
M. tuberculosis	0.09227	0.10640	0.15378	0.10557	0.09476	0.07315
E. coli	0.05421	0.101449	0.06978	0.06552	0.06069	0.06763
P. falciparum	0.05306	0.07411	0.06970	0.04984	0.04590	0.04279

Fonte: O autor.

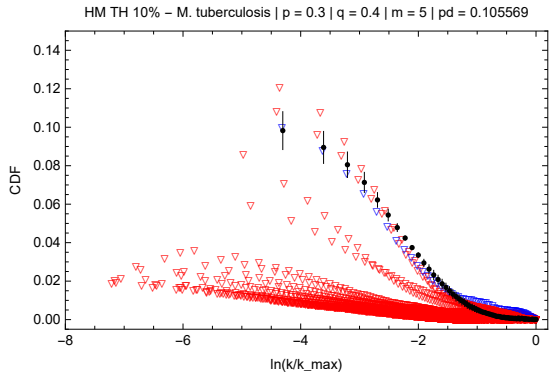
Pelas figuras já mencionadas acima e pelo que está apresentado na Tabela 3.2 podemos ver a que a transferência horizontal tem, de fato, uma influência significativa na estrutura da rede como um todo, visto que ela consegue deslocar a topologia e alterar a probabilidade de ocorrer a duplicação de nós nas redes.

Figura 3.3 – Efeitos da transferência horizontal sobre a topologia da rede da *M. tuberculosis*

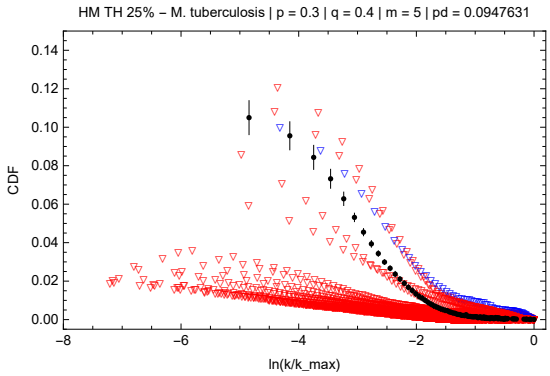


(a) $p_{TH} = 0.01$

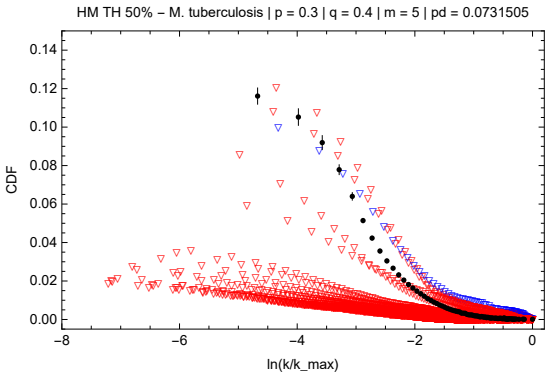
(b) $p_{TH} = 0.05$



(c) $p_{TH} = 0.10$



(d) $p_{TH} = 0.25$



(e) $p_{TH} = 0.50$

Fonte: O autor.

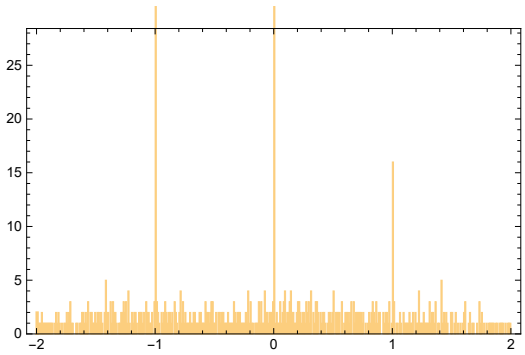
Além da topologia e da probabilidade de duplicação de nós, na próxima seção vamos verificar qual o efeito da transferência horizontal sobre o espectro de autovalores das redes geradas pelo novo modelo.

3.4.2 Efeito da transferência horizontal sobre o espectro de autovalores

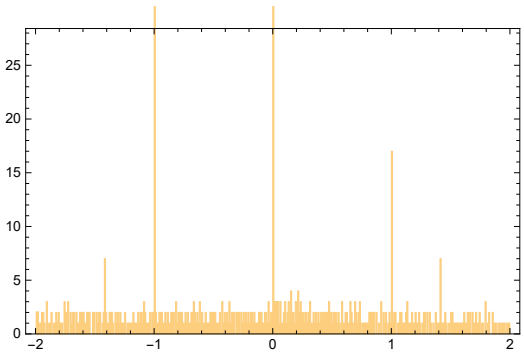
Voltando nossa atenção para o espectro de autovalores das redes de organismos do reino *Monera* podemos perceber pelas Figuras 3.4, F.3 e F.4 que o impacto da transferência horizontal é menos vísivel no espectro de autovalores do que quando paramos para analisar a topologia das redes e a probabilidade de ocorrer a duplicação nas mesmas. No entanto, algumas diferenças podem sim ser observadas no espectro de autovalores.

Para as três redes, podemos observar que conforme aumentamos a probabilidade de ocorrer a TH na rede, a quantidade de autovalores 1 presente na rede vai diminuindo, o que indica que há uma menor concentração de subgrafos na rede, ou seja, a TH não favorece a criação de pequenos grupos conectados dentro da rede. Para os demais picos, não percebemos uma mudança significativa assim como o que é visto para o autovalor 1, com exceção da rede da *M. tuberculosis* (Figura 3.4) onde é possível ver que assim como há uma redução no número de autovalores 1 com o aumento da TH, também ocorre uma redução na quantidade de autovalores -1 nessa rede.

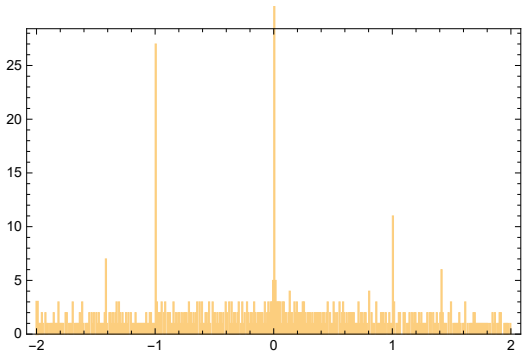
Figura 3.4 – Efeitos da transferência horizontal sobre o espectro de autovalores da rede da *M. tuberculosis*



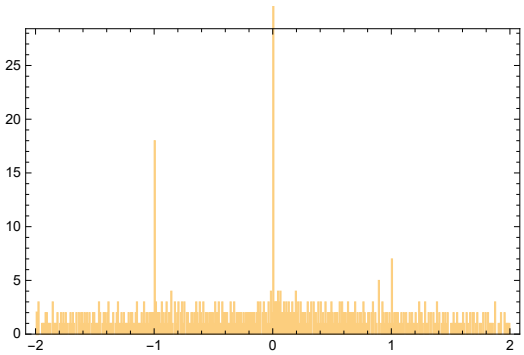
(a) $p_{TH} = 0.01$



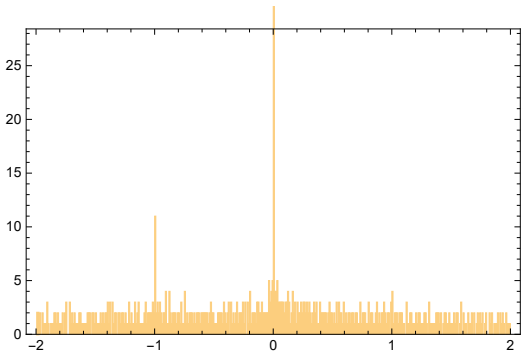
(b) $p_{TH} = 0.05$



(c) $p_{TH} = 0.10$



(d) $p_{TH} = 0.25$



(e) $p_{TH} = 0.50$

Fonte: O autor.

4 CONCLUSÕES

O objetivo principal desta dissertação foi o desenvolvimento de um novo modelo evolutivo para a reprodução das redes de interações entre proteínas que englobasse a Transferência Horizontal de Genes. Compreender como ocorrem os processos evolutivos de uma rede de interações entre proteínas é de fundamental importância para podermos identificar propriedades de interesse e também possíveis alvos de novos tratamentos, por exemplo. Para que isso seja possível, é necessário que tenhamos em mãos ferramentas que nos auxiliem a entender como alguns processos ocorrem e também que nos permitam reproduzir e simular com elevada precisão as redes de nosso interesse.

Como mostramos no Capítulo 2, ao apresentarmos as características de redes geradas pelo modelo *Duplication-Divergence* (VÁZQUEZ et al., 2003) nas Figuras 2.5, D.1 (topologia), 2.6 e D.2 (espectro de autovalores) e também os dados apresentados no Quadro 2.3 (estimativa de ocorrer a duplicação de nós dentro da rede), percebemos que a duplicação de nós tem sim um fator de impacto alto nas propriedades das redes de interações entre proteínas, mas pelo que vimos nos resultados apresentados, ela não pode ser o único mecanismo por trás da evolução das mesmas redes. A análise dos resultados gerados pelo modelo DD nos motivou a buscar explicações para as divergências no comportamento que observamos quando comparamos as redes reais com as redes geradas pelo modelo em questão, e essa busca nos levou a questionarmos se existiria algum outro mecanismo que, aliado com a duplicação de nós, seria responsável por explicar as principais características das redes de interações entre proteínas.

Esse questionamento nos levou a encontrar referências sobre um outro mecanismo em questão, que é a Transferência Horizontal de Genes. Como mencionado ao longo do trabalho, a existência da TH é algo concreto e já foi estudado, sendo atribuído à ela algumas características que permitiram a evolução de diversos organismos. Por ser um mecanismo que trás novas proteínas para uma rede, não há dúvidas de que ela é uma possibilidade real para atuar junto com a duplicação de nós na dinâmica e na evolução de uma rede de interações entre proteínas, uma vez que essa aquisição de novas proteínas custuma estar relacionada com o desenvolvimento de novas funções dentro de um organismo e também com algumas mutações que podem ocorrer no mesmo. Reforça-se, no entanto, assim como foi apresentado ao longo do texto, que não existe um consenso entre a real prevalência da TH na evolução de uma rede de interações entre proteínas, sendo que em alguns trabalhos trata-se a mesma como sendo responsável por até 96% dos novos genes de uma rede, ao passo que em outros trabalhos, acredita-se que a mesma seja responsável por, no máximo, 25% dos novos genes (TREANGEN; ROCHA, 2011).

Dessa forma, tínhamos em nossas mãos dois mecanismos distintos e possivelmente complementares para explicarmos e tentarmos reproduzir os processos evolutivos

das redes de interações entre proteínas. No entanto, apesar de sabermos da existência da TH, não foram encontradas referências e nem mesmo hipóteses de como esse mecanismo realmente funciona e atua na evolução de uma rede de interações entre proteínas. Sendo assim, para podermos implementar em um algoritmo computacional a TH, partimos da hipótese de que novas proteínas que são adicionadas ao organismo via TH terão uma probabilidade maior de se ligarem às proteínas mais "promíscuas" que já se encontram na rede, isto é, aquelas proteínas que têm uma conectividade mais elevada dentro da rede (os chamados *hubs*), uma vez que essas proteínas possivelmente são proteínas mais antigas dentro do organismo e que já possuem suas funções bem definidas e estáveis, o que pode fazer com que as novas proteínas não causem tantos impactos negativos (como por exemplo, aumentar a chance de ocorrerem mutações malignas) quando adicionadas ao organismo.

Tomando como base a hipótese acima utilizamos o algoritmo do modelo desenvolvido por Barabási e Albert (BARABÁSI; ALBERT, 1999) para incorporar a TH, fazendo pequenos ajustes na parte do crescimento proposto pelo modelo BA. Sobre a duplicação de nós, a mesma foi utilizada exatamente igual à forma como definida no modelo DD pois não podemos desconsiderar os resultados gerados por esse modelo, uma vez que a topologia das redes geradas por ele é muito próxima da topologia das redes reais e as bases biológicas que o mesmo modelo fornece para a duplicação foram bem validadas. Assim, ao combinarmos os dois mecanismos e as suas respectivas implementações, desenvolvemos um novo modelo para a reprodução das redes de interações entre proteínas.

Como foi apresentado ao longo do Capítulo 3, podemos observar que as redes geradas pelo novo modelo apresentam uma topologia (Figuras 3.1 e E.1) bastante próxima da topologia vista para as redes reais, algo que já era esperado em função do mecanismo de duplicação escolhido para o novo modelo ser o utilizado no modelo DD e que produz bons resultados para a topologia das redes. Além disso, pelo que foi apresentado no Quadro 3.1 observamos que as estimativas para a probabilidade de duplicação das redes geradas pelo novo modelo encontram-se muito mais próximas dos valores encontrados para as redes reais do que as estimativas obtidas para o modelo DD (Quadro 2.3), o que é um resultado bastante importante pois revela que a TH, apesar de pouco presente nas redes, tem um grande impacto nas propriedades internas das mesmas. Por fim, ao analisarmos o espectro de autovalores (Figuras 3.2 e E.2), constatamos que o mesmo ainda se encontra um pouco distante daquele que é observado para as redes reais (Figuras 2.4 e B.2), mas mesmo assim consegue reproduzir os principais pontos de interesse que são as degenerescências nos autovalores -1, 0 e 1.

Como podemos ver no Capítulo 3, os resultados apresentados para as redes de interesse apresentaram uma participação de certa forma pouco significativa da TH (variando de 1% até 10%, no máximo), contrastando com as informações obtidas na literatura de que a mesma poderia estar presente em até 25% dos novos genes que são adquiridos

por um organismo. No entanto, apesar da baixa presença, foi possível notar que a TH tem sim uma influência na evolução da rede de interações entre proteínas de um organismo, principalmente no que diz respeito às propriedades internas da mesma. Essa influência ficou mais clara quando analisamos o impacto da TH sobre as redes do reino *Monera*, que foi apresentada na Seção 3.4 e com continuação no Apêndice F. Nesses resultados ficou nítida a influência que a TH tem nas redes, uma vez que ao aumentarmos a ocorrência da mesma, conseguimos observar um deslocamento horizontal na topologia das redes, uma diminuição na estimativa da probabilidade de ocorrer a duplicação de nós na rede e também um menor favorecimento à criação de subgrafos dentro da rede, conforme podemos ver pela diminuição das degenerescências nos autovalores -1 e 1 (Figuras 3.4, F.3 e F.4) que estão ligados à essas estruturas.

Com os resultados apresentados aqui, podemos observar que o novo modelo desenvolvido e proposto, é promissor e surge como uma primeira tentativa de explicar algumas características por trás da evolução das redes de interações entre proteínas que até então não foram explicadas. Além disso, também mostramos que ao tentarmos reproduzir uma rede de interações entre proteínas, não podemos voltar nossa atenção somente para a topologia da rede, é necessário que um modelo que busque reproduzir tais redes seja capaz de capturar outras propriedades das mesmas, uma vez que o espectro de autovalores, por exemplo, nos fornece informações preciosas a respeito da estrutura interna da rede. Outro ponto que vale destacar é a forma como foi implementada a TH. Uma vez que não foram encontradas sugestões e nem tentativas de como implementar a mesma, partimos de uma hipótese sobre como ela funcionaria e ao final, constatamos que fomos na direção correta, pois os resultados aqui apresentados foram extremamente satisfatórios. Assim, como já mencionado previamente, temos em nossas mãos mais uma ferramenta para auxiliar no estudos da evolução das redes de interações entre proteínas.

REFERÊNCIAS BIBLIOGRÁFICAS

AGUIAR, M. A. M. de; BAR-YAM, Y. Spectral analysis and the dynamic response of complex networks. **Physical Review E**, v. 71:016106, 2005.

AMARAL, L. A. N. et al. Classes of small-world networks. **PNAS**, v. 97, p. 11149–11152, 2000.

BARABÁSI, A.-L. **Network Science**. Cambridge: Cambridge University Press, 2016. 475 p.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, v. 286, p. 509–512, 1999.

BARABÁSI, A.-L.; OLTVAI, Z. N. Network biology: Understanding the cell's functional organization. **Nature Reviews: Genetics**, v. 5, p. 101–114, 2004.

ERDŐS, P.; RÉNYI, A. On random graphs, I. **Publicationes Mathematicae**, v. 6, p. 290–297, 1959.

GOGARTEN, J. P.; DOOLITTLE, W. F.; LAWRENCE, J. G. Prokaryotic evolution in light of gene transfer. **Molecular Biology and Evolution**, v. 19(12), p. 2226–2238, 2002.

INC., W. R. **Mathematica, Version 12.1**. Champaign, IL, 2020.

JEONG, H. et al. The large-scale organization of metabolic networks. **Nature**, v. 407, p. 651–654, 2000.

KAMP, C.; CHRISTENSEN, K. Spectral analysis of protein-protein interactions in *Drosophila melanogaster*. **Physical Review E**, n. 71:041911, 2005.

LERAT, E. et al. Evolutionary origins of genomic repertoires in bacteria. **PLoS Biology**, v. 3(5), 2005.

LIMA-MENDEZ, G.; HELDEN, J. van. The powerful law of the power law and other myths in network biology. **Molecular BioSystems**, v. 5, p. 1482–1493, 2009.

MARREC, L.; JALAN, S. Analysing degeneracies in networks spectra. **Europhysics Letters**, v. 117(4):48001, 2017.

MOMBACH, J. C. M. Redes complexas e aplicações. In: ANDRADE, E. X. L. de; SAMPAIO, R.; SILVA, G. N. (Ed.). **Modelagem em biomatemática**. São Carlos: SBMAC, 2004. cap. 13, p. 34–53.

OCHMAN, H.; LAWRENCE, J. G.; GROISMAN, E. A. Lateral gene transfer and the nature of bacterial innovation. **Nature**, v. 405, 2000.

PELLEGRINI, M.; HAYNOR, D.; JOHNSON, J. M. Protein interaction networks. **Expert Review of Proteomics**, v. 1(2), p. 239–249, 2004.

RAI, A. et al. Understanding cancer complexome using networks, spectral graph theory and multilayer framework. **Scientific Reports**, v. 7:41676, 2017.

RAI, A.; SHINDE, P.; JALAN, S. Network spectra for drug-target identification in complex diseases: new guns against old foes. **Applied Network Science**, v. 3:51, 2018.

SZKLARCZYK, D. et al. The string database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. **Nucleic acids research**, v. 49, p. 605–612, 2021.

TREANGEN, T. J.; ROCHA, E. P. C. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. **PLoS Genetics**, v. 7(1), 2011.

UETZ, P. et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. **Nature**, v. 403(6770), p. 623–627, 2000.

VÁZQUEZ, A. et al. Modeling of protein interaction networks. **ComPlexUs: Original Research Paper**, n. 1, p. 38–44, 2003.

WEISSTEIN, Eric W. **MATRIX RANK**. From *MathWorld* - A Wolfram Web Resource. Acesso em 06 jun. 2019. Disponível em: <http://mathworld.wolfram.com/MatrixRank.html>.

YADAV, A.; JALAN, S. Origin and implications of zero degeneracy in networks spectra. **CHAOS: An Interdisciplinary Journal of Nonlinear Science**, v. 25:043110, 2015.

APÊNDICE A – TAMANHO DAS REDES E DO COMPONENTE GIGANTE PARA DIFERENTES NÍVEIS DE CONFIANÇA (*SCORE*) DO STRING

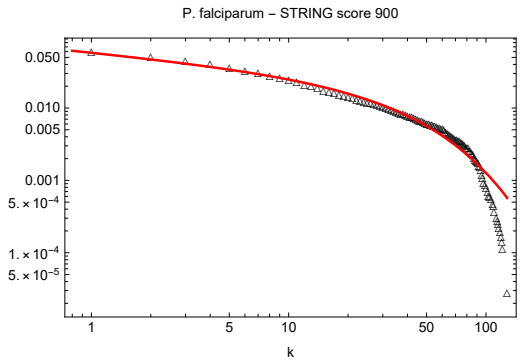
Quadro A.1 – Tamanho das redes e dos Componentes Gigantes para diferentes níveis de intervalo de confiança do STRING

	0.000		0.700		0.750		0.800	
Organismo	N	N_{CG}	N	N_{CG}	N	N_{CG}	N	N_{CG}
M. tuberculosis	4034	4032	3304	2708	3093	2356	2824	1907
E. coli	4602	4592	3946	3392	3716	3025	3437	2454
P. falciparum	5201	3751	3710	3373	3295	3093	2984	2111
B. subtilis	4181	4175	3816	3690	3707	3518	3557	3294
T. cruzi	8106	7865	5355	5156	4719	4520	4294	4013
S. pombe	5076	5074	4598	4587	4481	4456	4298	4227
A. mellifera	10226	10136	6907	6680	6064	5710	5583	5119
S. cerevisae	6574	6568	6157	6108	6075	6007	5847	5676
C. elegans	18181	18108	10750	10519	9861	9419	9031	8293
O. sativa	25106	25014	14208	13731	12592	11955	10542	9635
D. melanogaster	13046	13015	10431	10229	9888	9604	9301	8854
G. domesticus	14683	14681	12856	12788	12321	12145	10755	10364
X. laevis	17286	17250	13644	13540	13123	12946	11792	11412
A. thaliana	25490	25449	17990	17737	16844	16367	15396	14556
B. taurus	19845	19841	16387	16191	15043	14532	13889	13138
D. rerios	24681	24679	17696	17550	16548	16220	15542	15055
M. musculus	21291	21236	16690	16227	15035	14424	13655	13036
H. sapiens	19354	19354	17185	16957	15928	15561	14711	14319

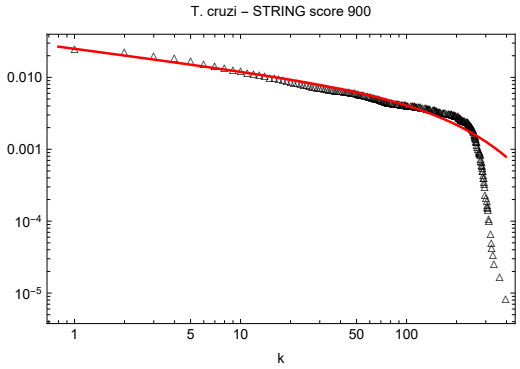
Fonte: O autor.

APÊNDICE B – TOPOLOGIA E ESPECTRO DE AUTOVALORES DAS REDES REAIS

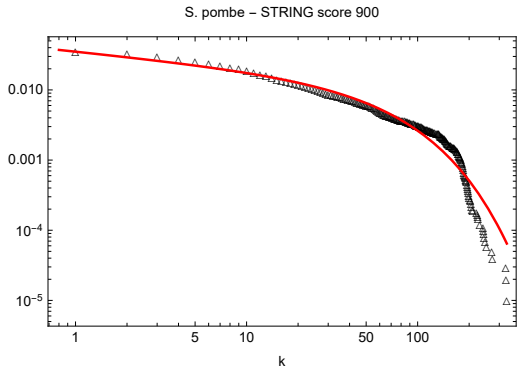
Figura B.1 – Topologia das redes experimentais. Os triângulos pretos representam a curva da topologia da rede real, e a linha em vermelho é a curva que melhor representa esses dados



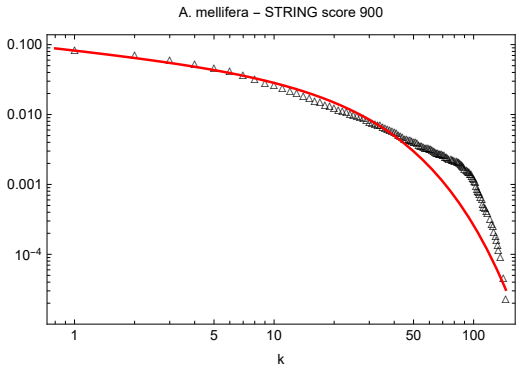
(a) *P. falciparum*



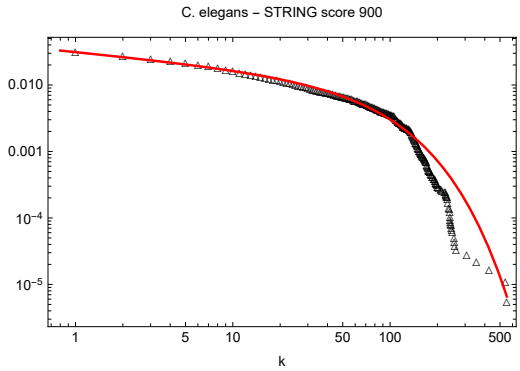
(b) *T. cruzi*



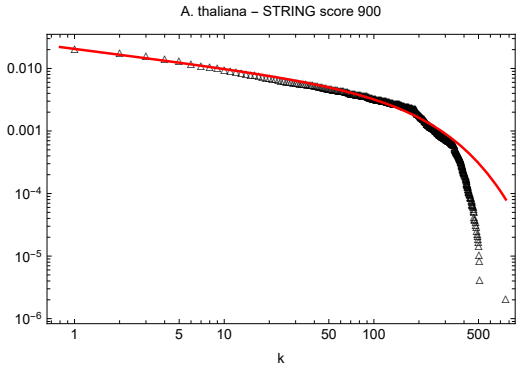
(c) *S. pombe*



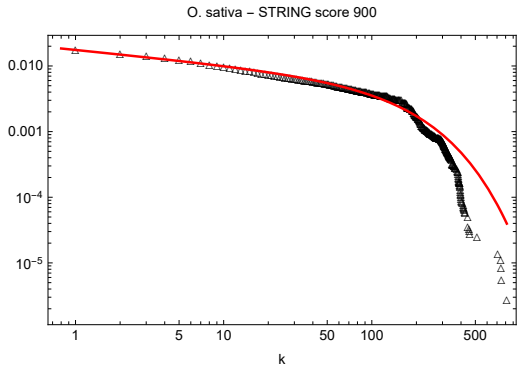
(d) *A. mellifera*



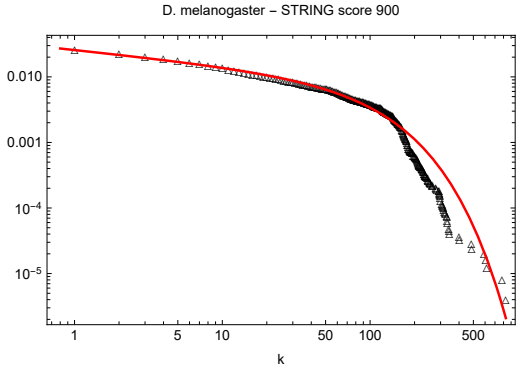
(e) *C. elegans*



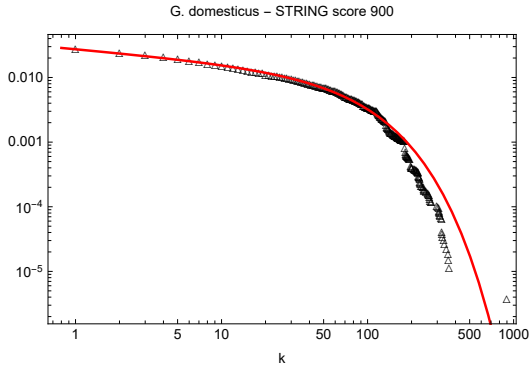
(f) *A. thaliana*



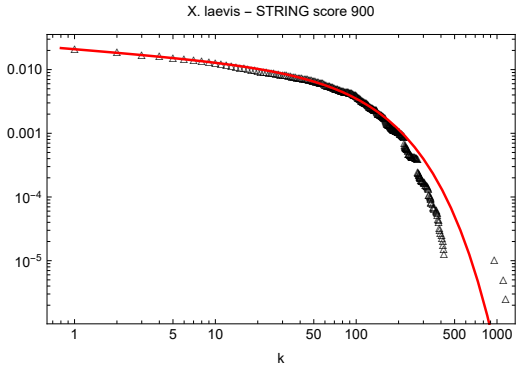
(g) *O sativa*



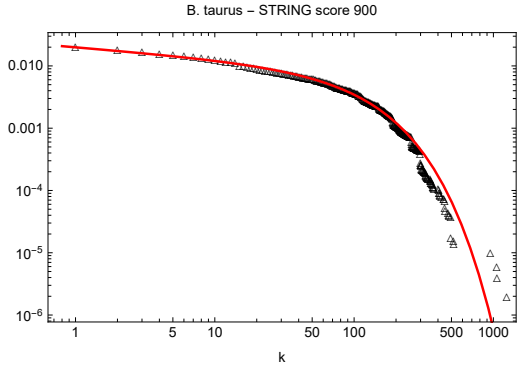
(h) *D. melanogaster*



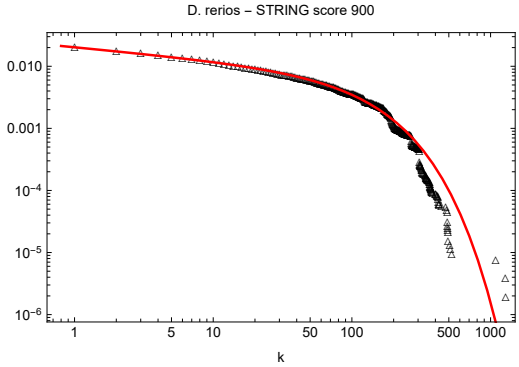
(i) *G. domesticus*



(j) *X. laevis*



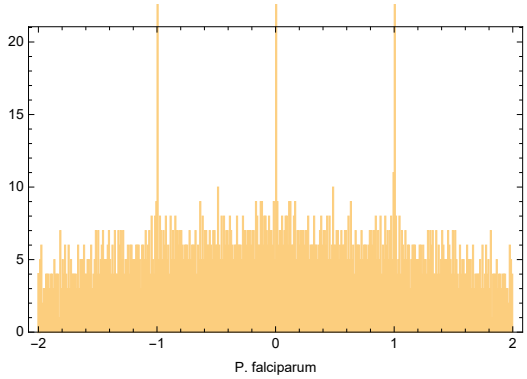
(k) *B. taurus*



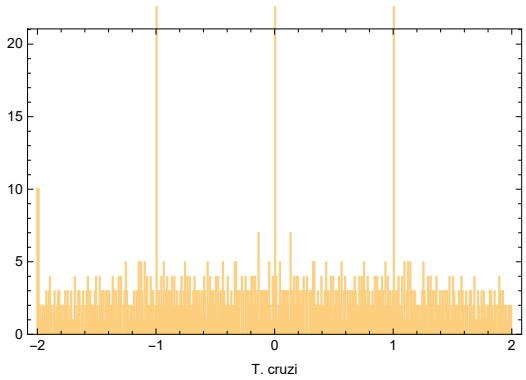
(l) *D. rerios*

Fonte: O autor.

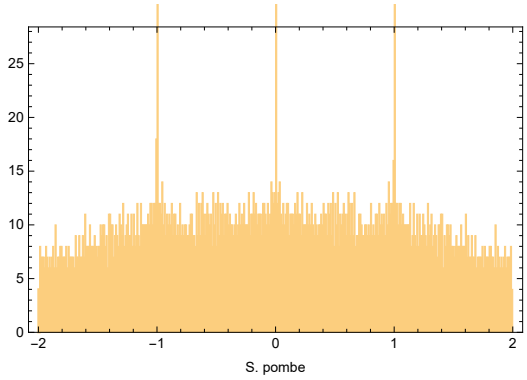
Figura B.2 – Espectro de autovalores dos organismos estudados



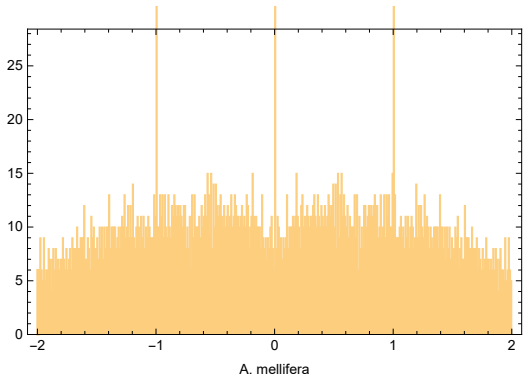
(a) *P. falciparum*



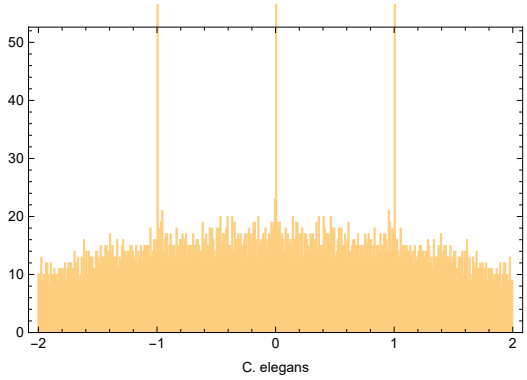
(b) *T. cruzi*



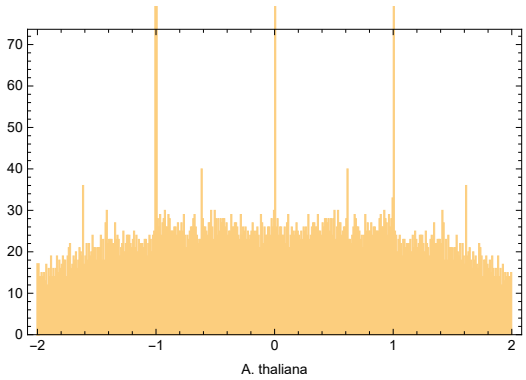
(c) *S. pombe*



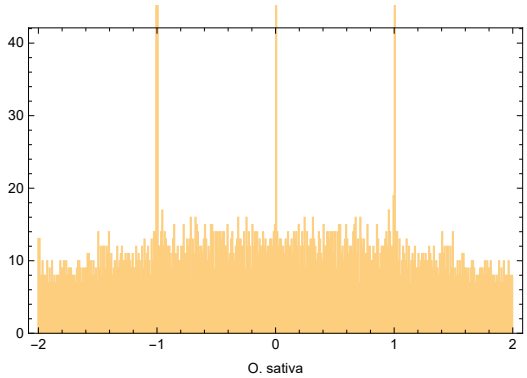
(d) *A. mellifera*



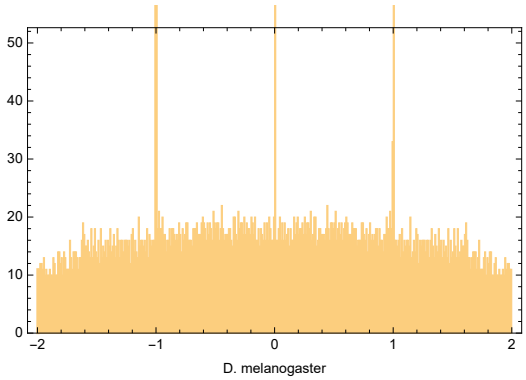
(e) *C. elegans*



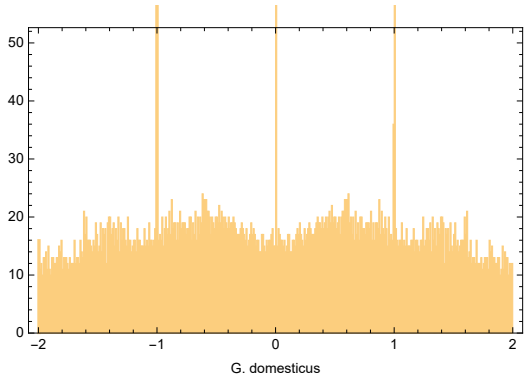
(f) *A. thaliana*



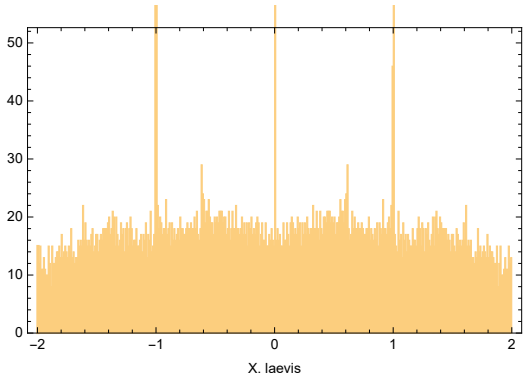
(g) *O sativa*



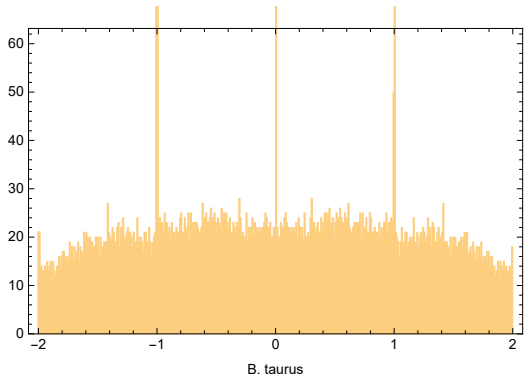
(h) *D. melanogaster*



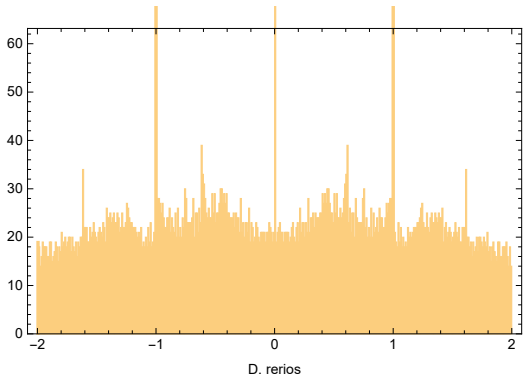
(i) *G. domesticus*



(j) *X. laevis*



(k) *B. taurus*

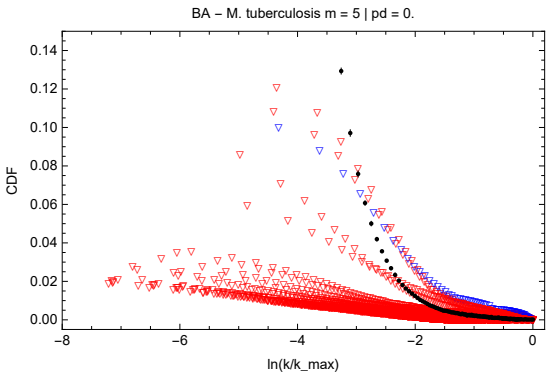


(l) *D. rerios*

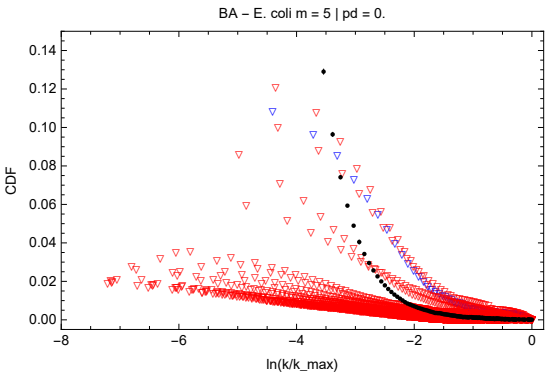
Fonte: O autor.

APÊNDICE C – COMPARAÇÃO ENTRE AS REDES GERADAS PELO MODELO DE BARABÁSI E ALBERT E AS REDES EXPERIMENTAIS

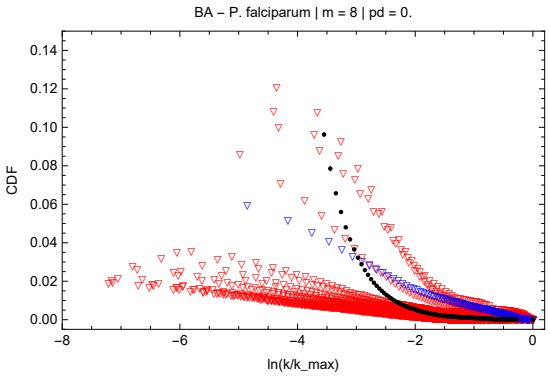
Figura C.1 – Topologia das redes geradas pelo modelo de Barabási e Albert. Em preto temos a curva da topologia da rede gerada pelo modelo, com triângulos azuis temos a curva da topologia da rede real e as demais curvas representam a topologia das outras redes.



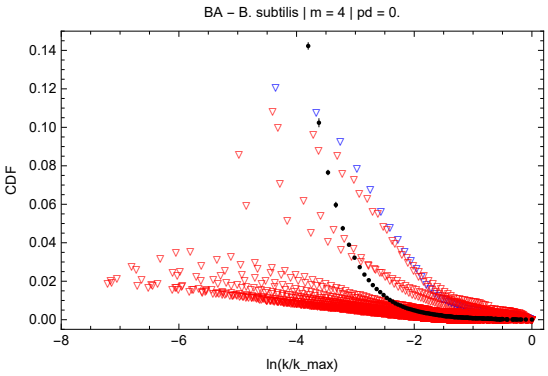
(a) *M. tuberculosis*



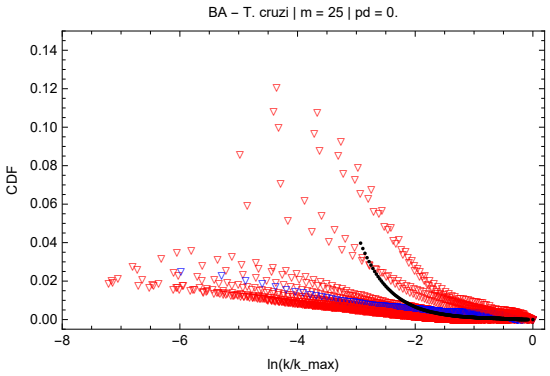
(b) *E. coli*



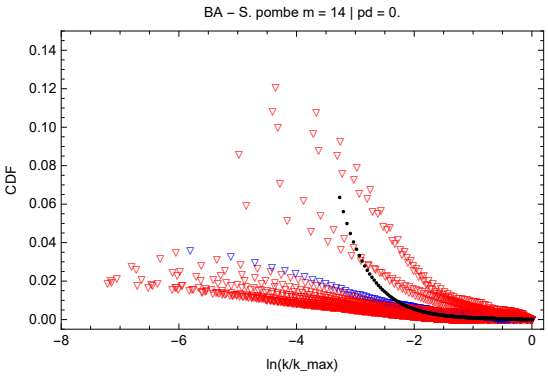
(c) *P. falciparum*



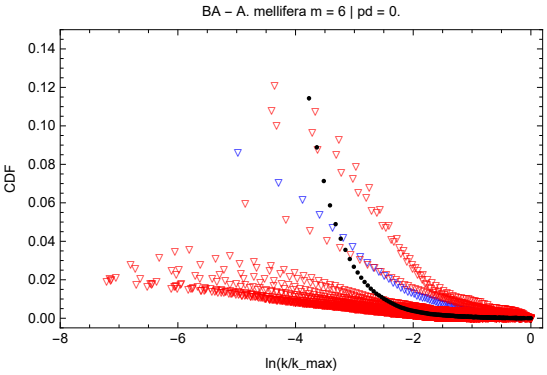
(d) *B. subtilis*



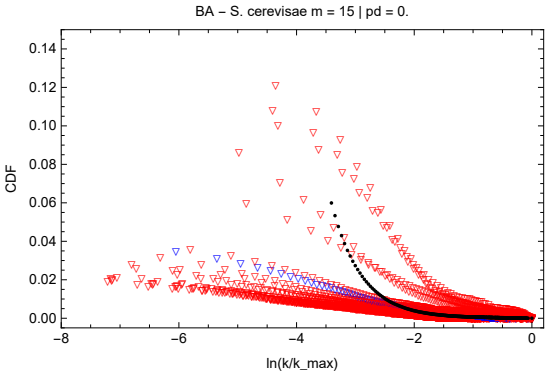
(e) *T. cruzi*



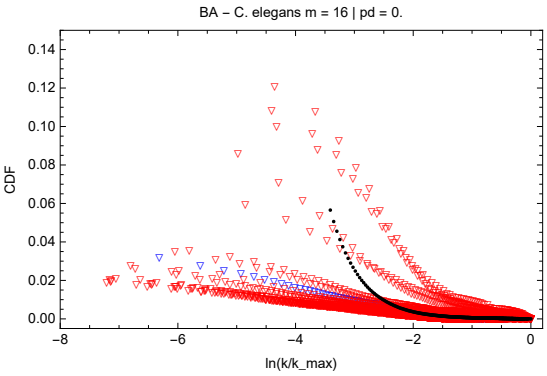
(f) *S. pombe*



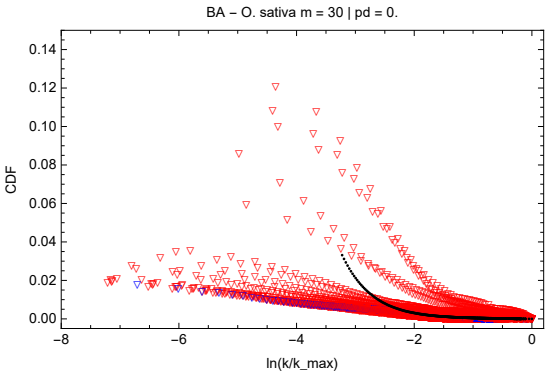
(g) *A. mellifera*



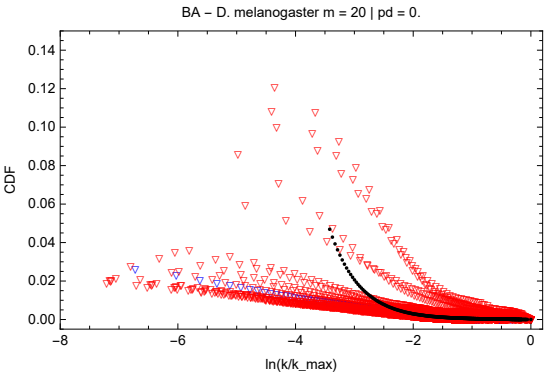
(h) *S. cerevisae*



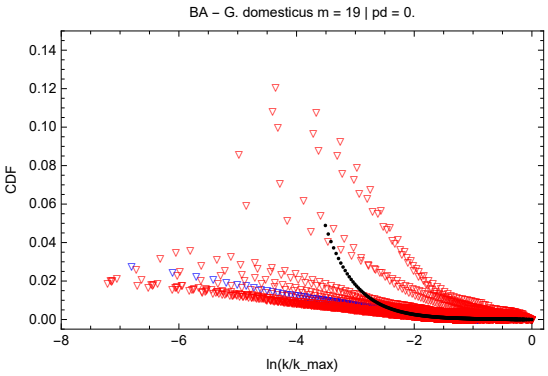
(i) *C. elegans*



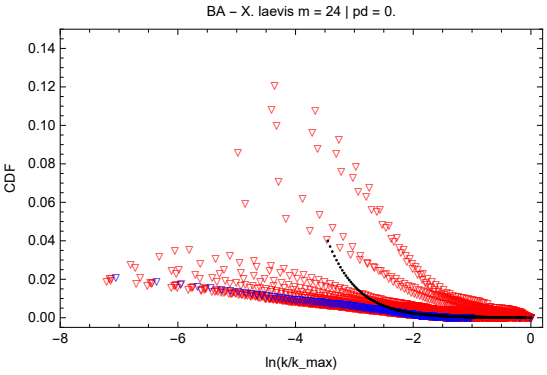
(j) *O sativa*



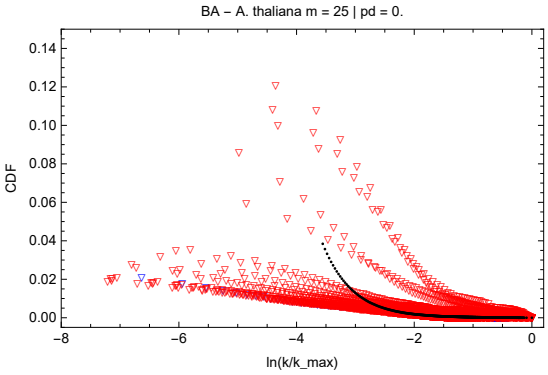
(k) *D. melanogaster*



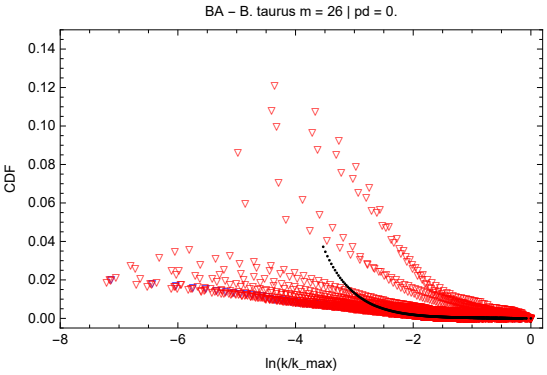
(l) *G. domesticus*



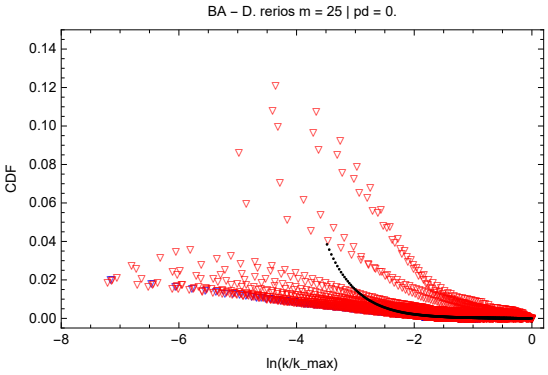
(m) *X. laevis*



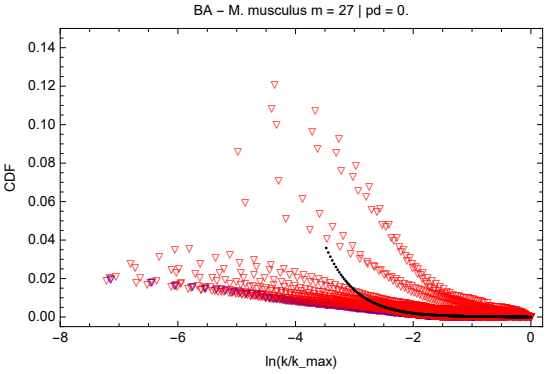
(n) *A. thaliana*



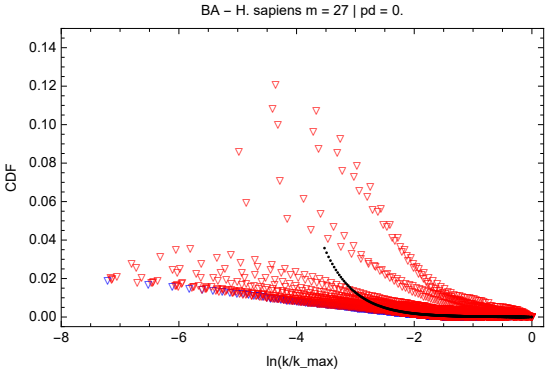
(o) *B. taurus*



(p) *D. rerios*



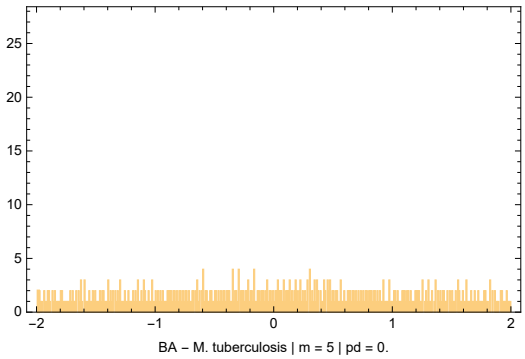
(q) *M. musculus*



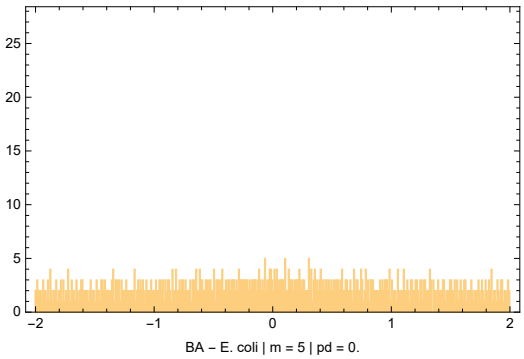
(r) *H. sapiens*

Fonte: O autor.

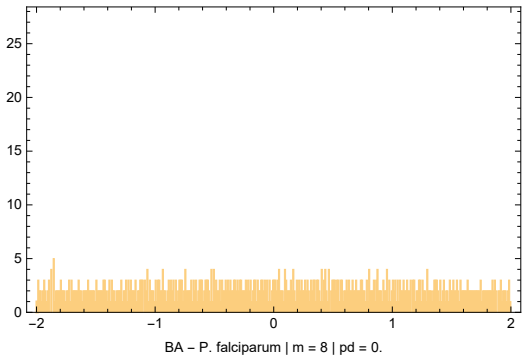
Figura C.2 – Espectro de autovalores das redes geradas pelo modelo de Barabási e Albert



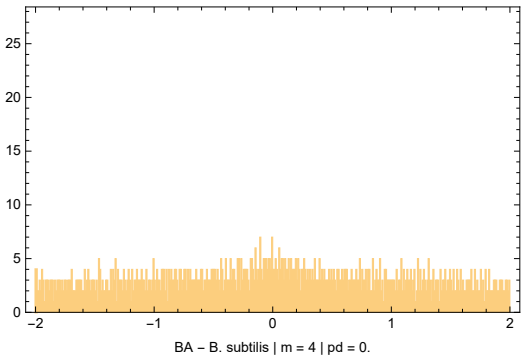
(a) *M. tuberculosis*



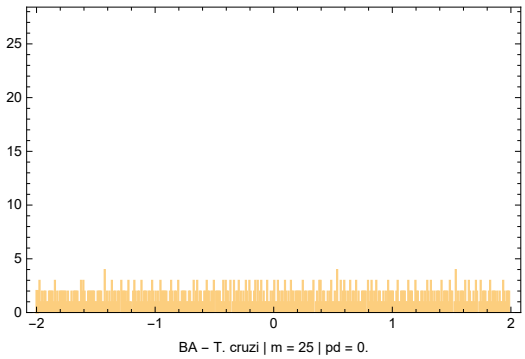
(b) *E. coli*



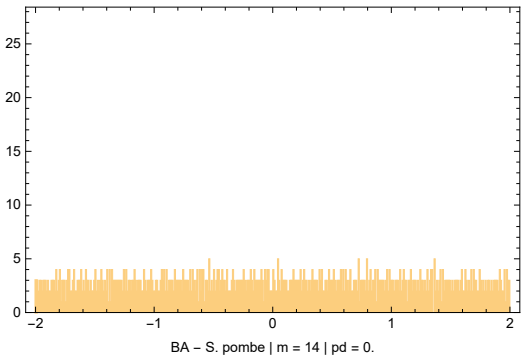
(c) *P. falciparum*



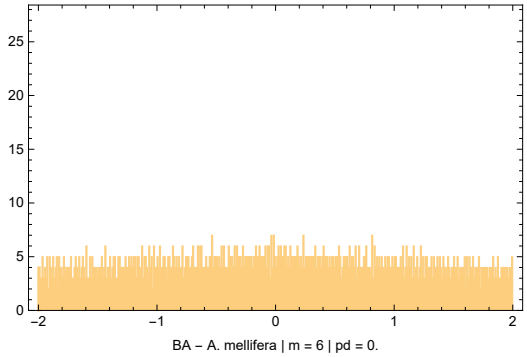
(d) *B. subtilis*



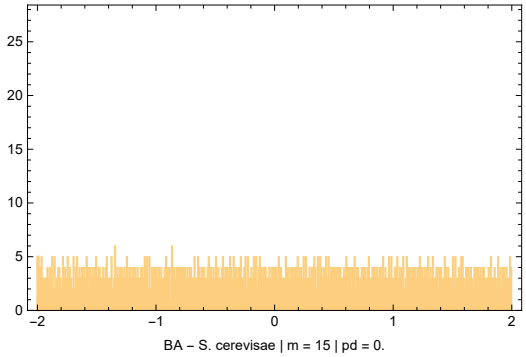
(e) *T. cruzi*



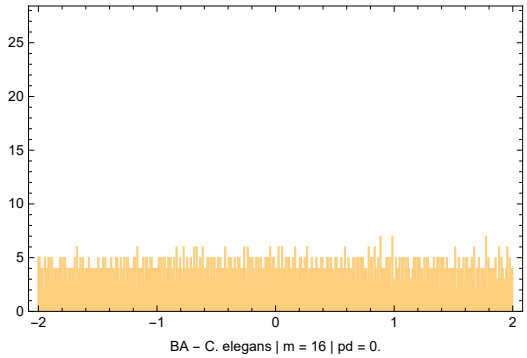
(f) *S. pombe*



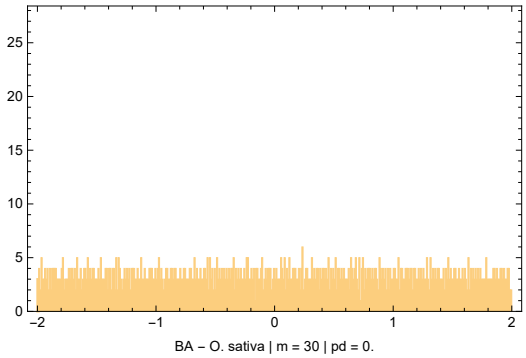
(g) *A. mellifera*



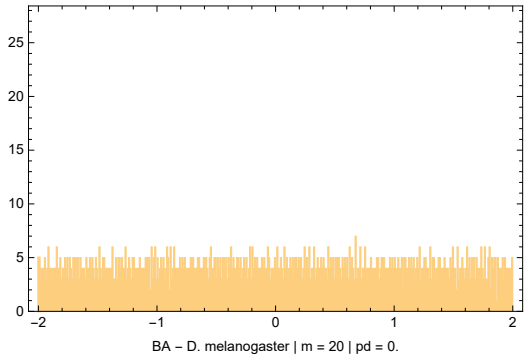
(h) *S. cerevisae*



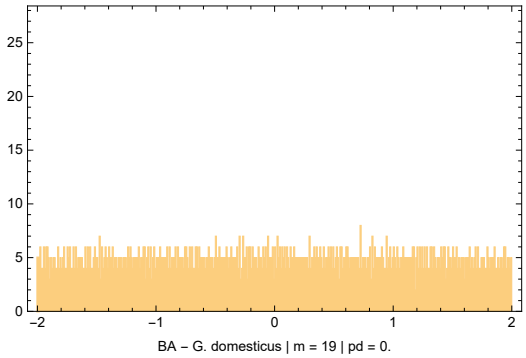
(i) *C. elegans*



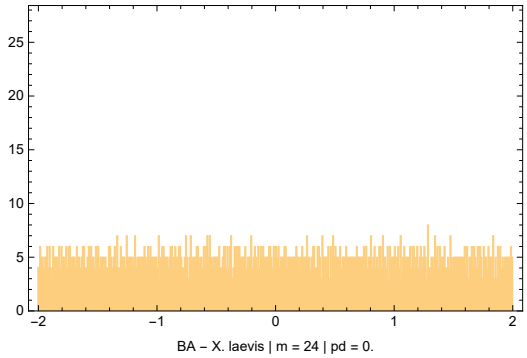
(j) *O sativa*



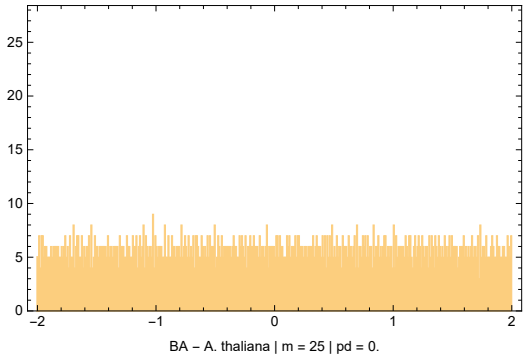
(k) *D. melanogaster*



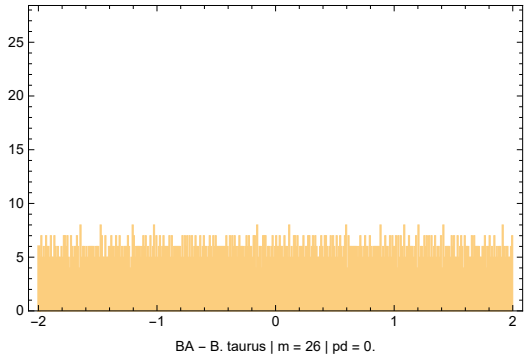
(l) *G. domesticus*



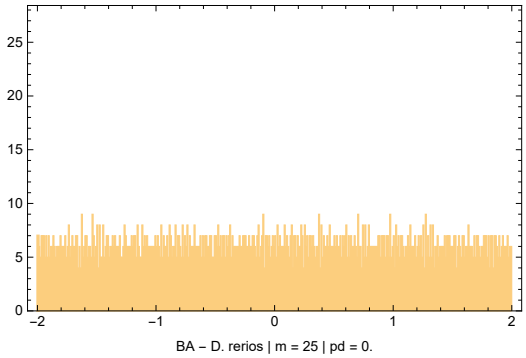
(m) *X. laevis*



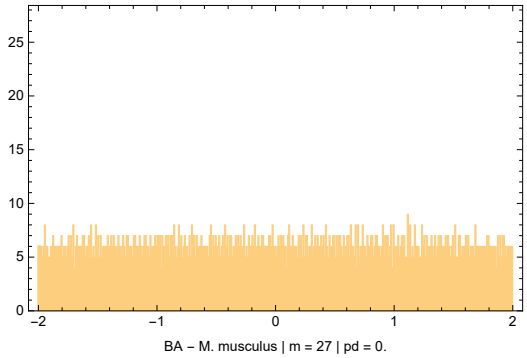
(n) *A. thaliana*



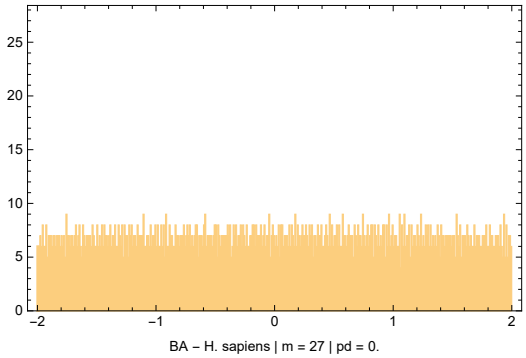
(o) *B. taurus*



(p) *D. rerios*



(q) *M. musculus*

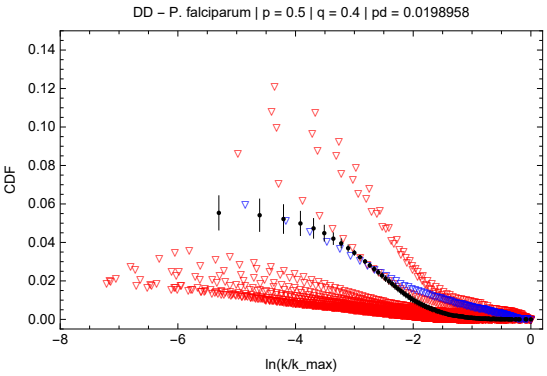


(r) *H. sapiens*

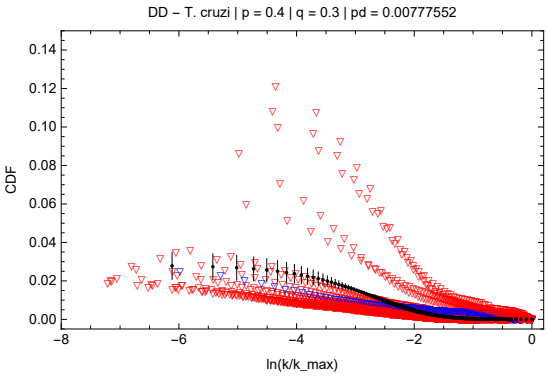
Fonte: O autor.

APÊNDICE D – TOPOLOGIA E ESPECTRO DE AUTOVALORES DAS REDES GERADAS PELO MODELO *DUPLICATION-DIVERGENCE*

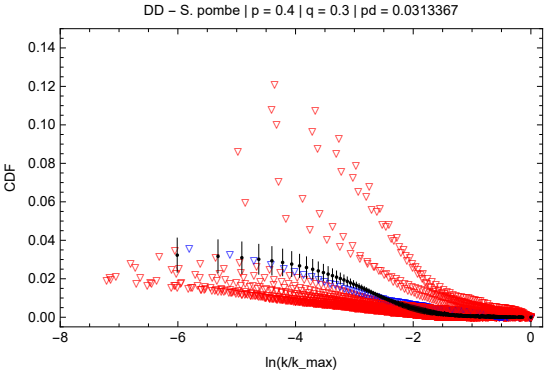
Figura D.1 – Topologia das redes geradas pelo modelo *Duplication-Divergence*. Os triângulos vermelhos representam as curvas da topologia das redes reais e a curva composta pelos triângulos azuis destaca a topologia da rede experimental do organismo em questão. A curva formada pelos pontos e pelas barras de erro em preto representa a topologia da rede gerada pelo modelo DD para o conjunto de parâmetros descrito no gráfico.



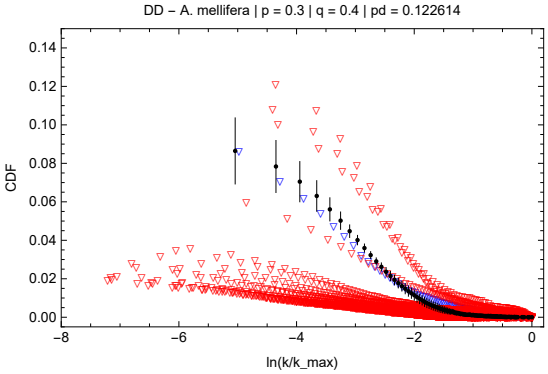
(a) *P. falciparum*



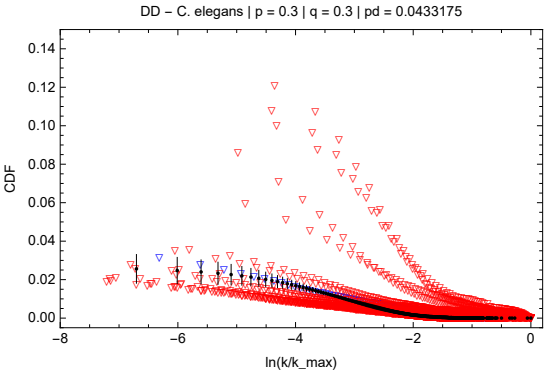
(b) *T. cruzi*



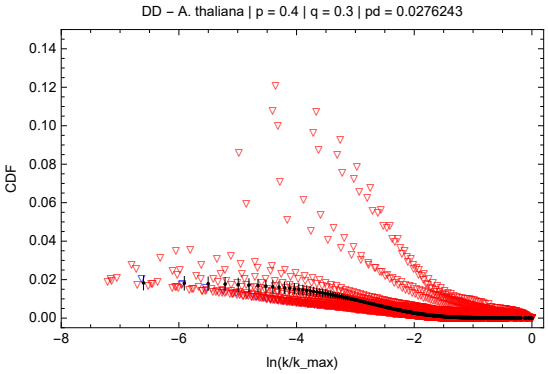
(c) *S. pombe*



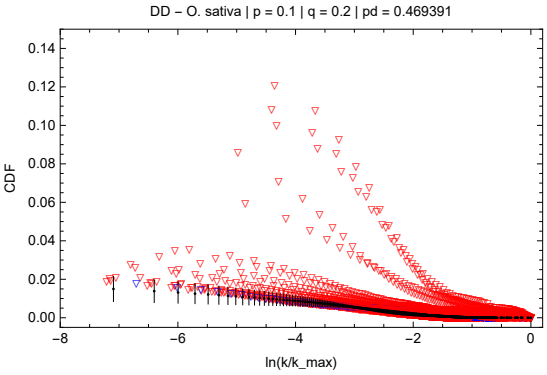
(d) *A. mellifera*



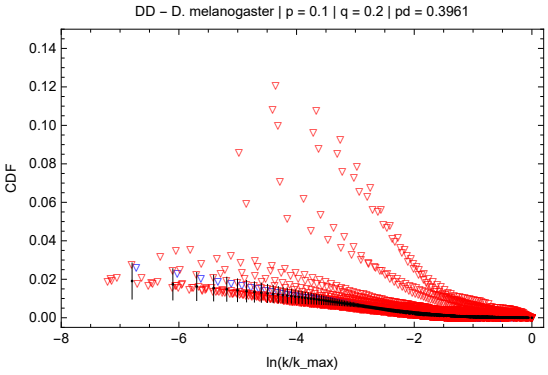
(e) *C. elegans*



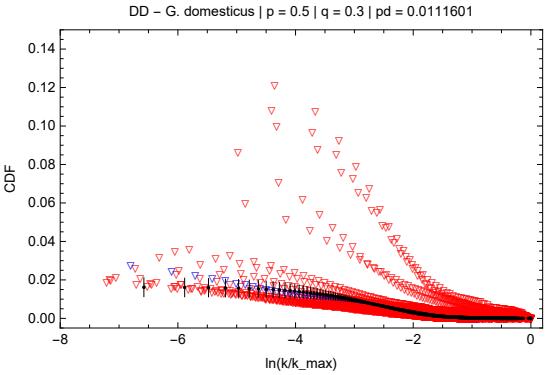
(f) *A. thaliana*



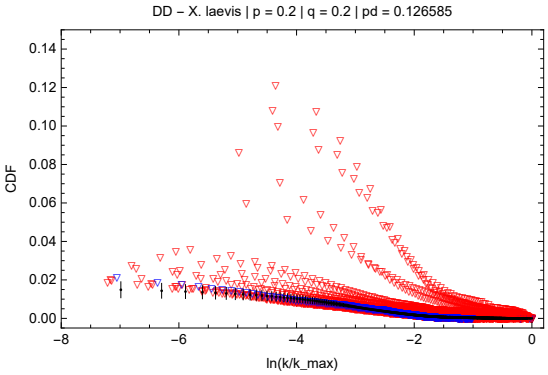
(g) *O sativa*



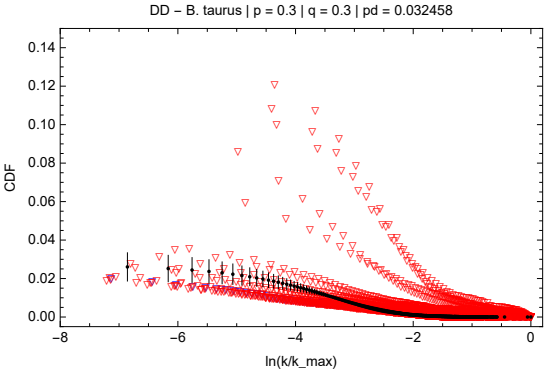
(h) *D. melanogaster*



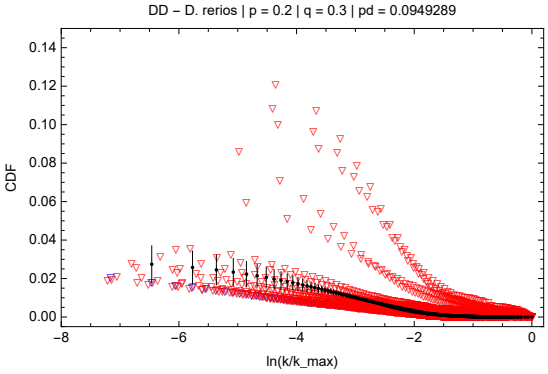
(i) *G. domesticus*



(j) *X. laevis*



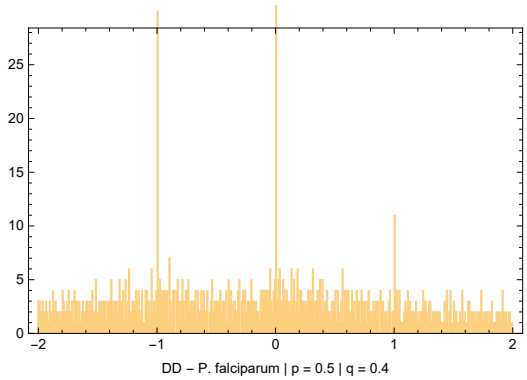
(k) *B. taurus*



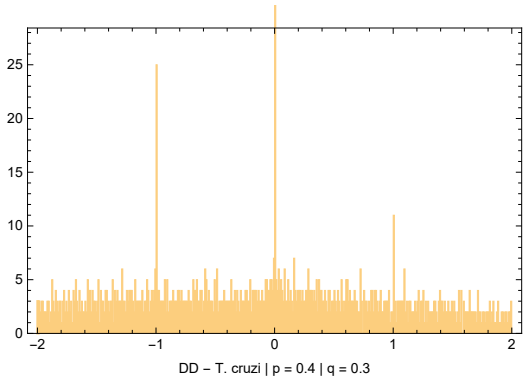
(l) *D. rerios*

Fonte: O autor.

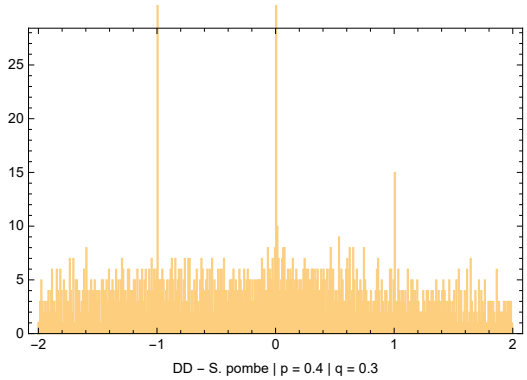
Figura D.2 – Espectro de autovalores das redes geradas pelo modelo *Duplication-Divergence*



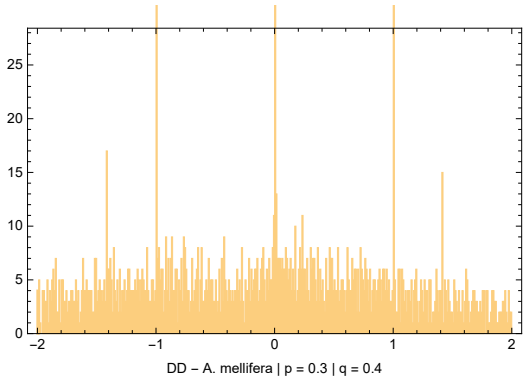
(a) *P. falciparum*



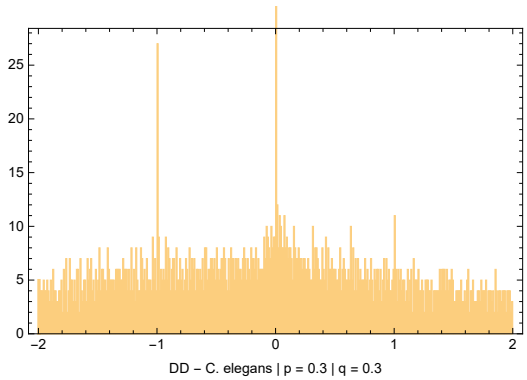
(b) *T. cruzi*



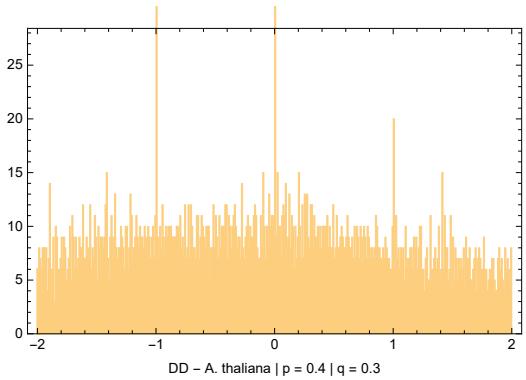
(c) *S. pombe*



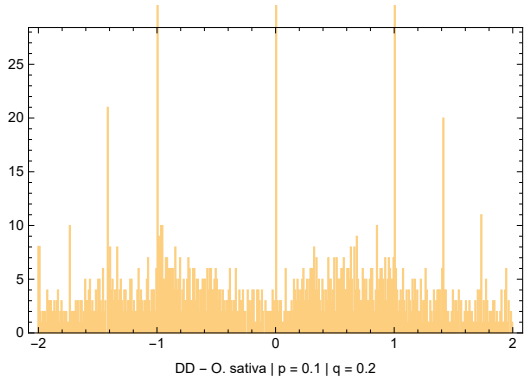
(d) *A. mellifera*



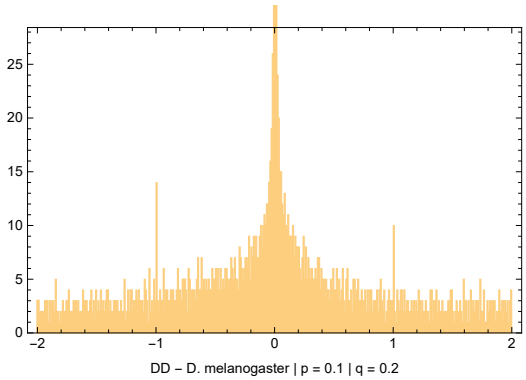
(e) *C. elegans*



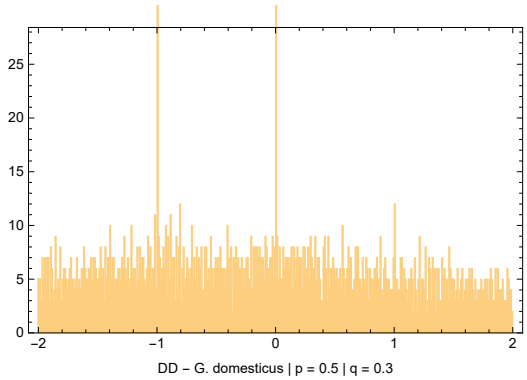
(f) *A. thaliana*



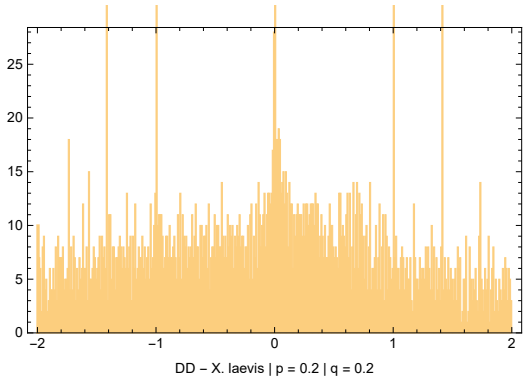
(g) *O sativa*



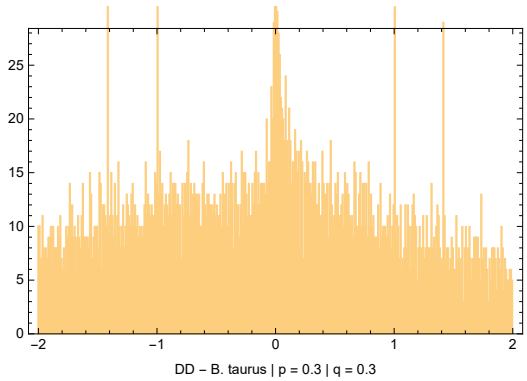
(h) *D. melanogaster*



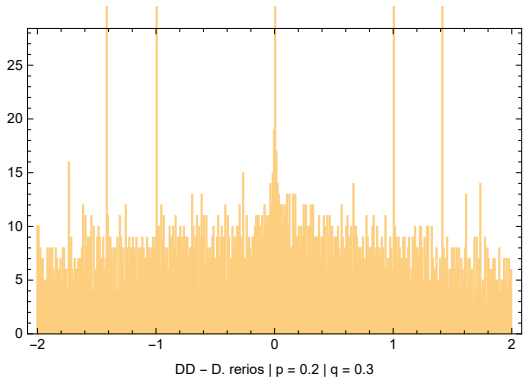
(i) *G. domesticus*



(j) *X. laevis*



(k) *B. taurus*

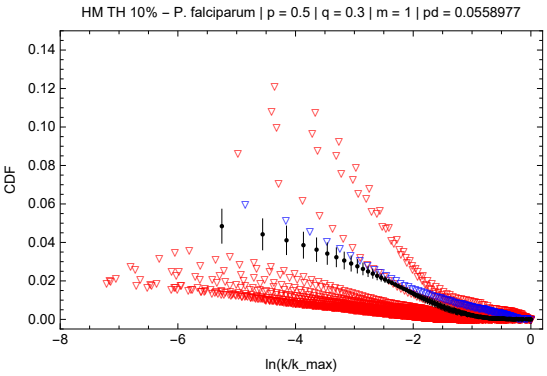


(l) *D. rerios*

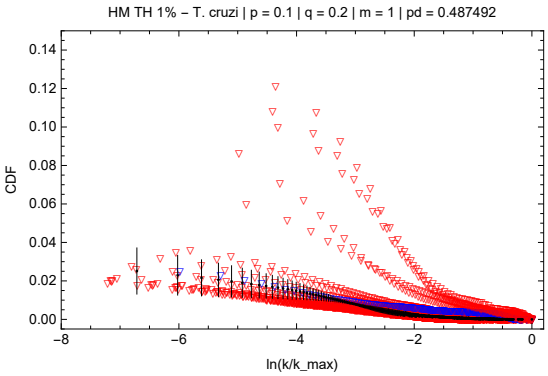
Fonte: O autor.

APÊNDICE E – TOPOLOGIA E ESPECTRO DE AUTOVALORES DAS REDES GERADAS PELO NOVO MODELO

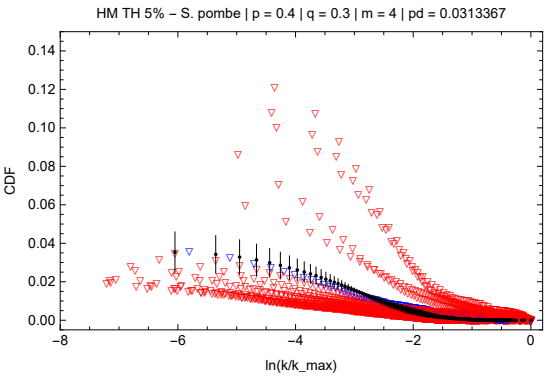
Figura E.1 – Topologia das redes geradas pelo novo modelo. Os triângulos vermelhos representam as curvas da topologia das redes reais e a curva composta pelos triângulos azuis destaca a topologia da rede experimental do organismo em questão. A curva formada pelos pontos e pelas barras de erro em preto representa a topologia da rede gerada pelo modelo DD para o conjunto de parâmetros descrito no gráfico.



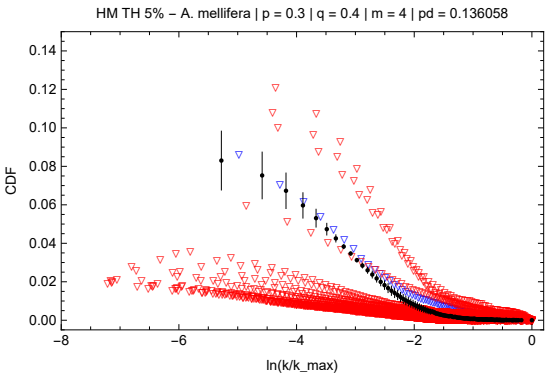
(a) *P. falciparum*



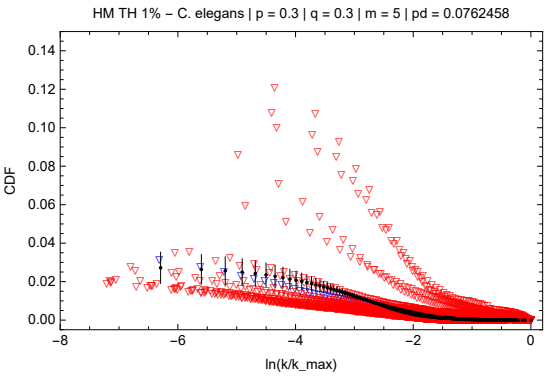
(b) *T. cruzi*



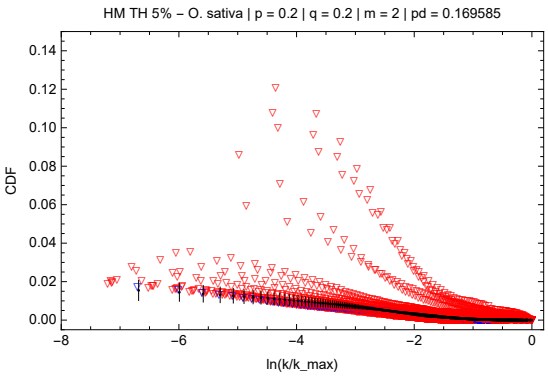
(c) *S. pombe*



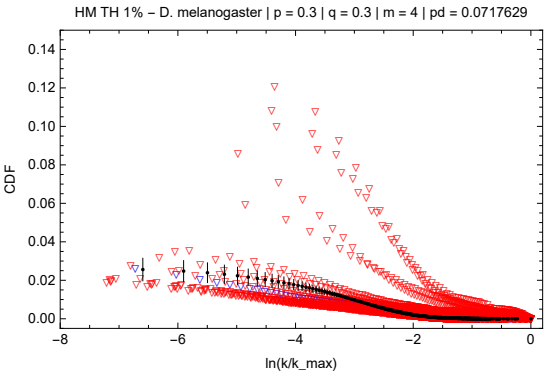
(d) *A. mellifera*



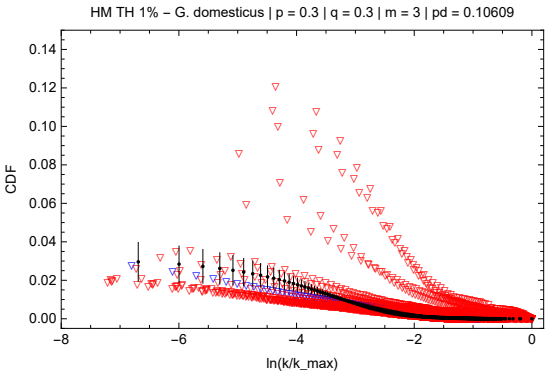
(e) *C. elegans*



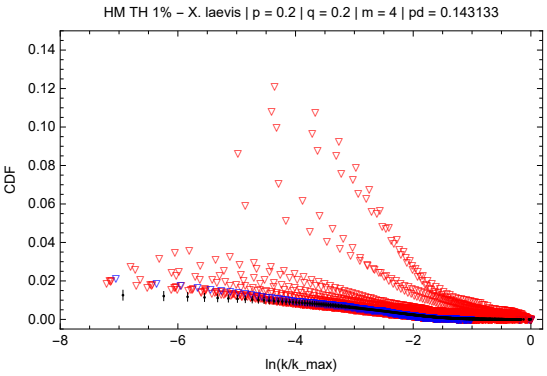
(f) *O sativa*



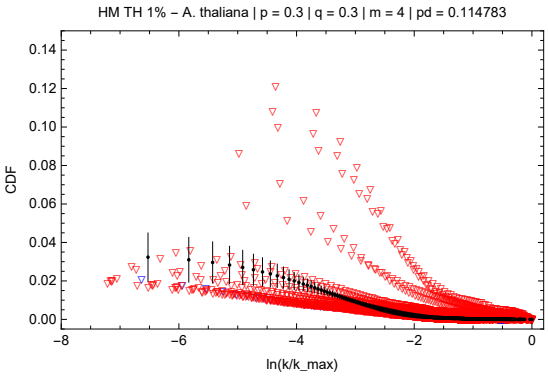
(g) *D. melanogaster*



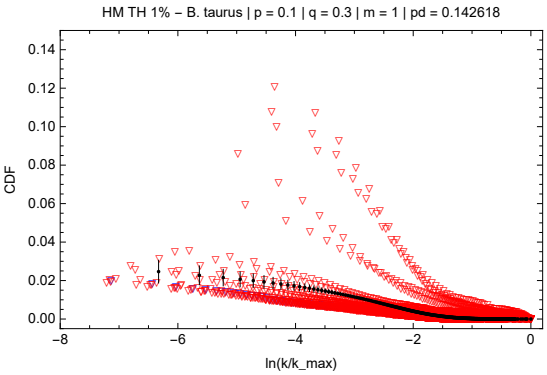
(h) *G. domesticus*



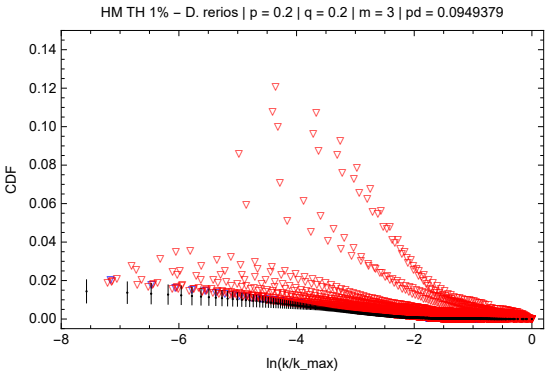
(i) *X. laevis*



(j) *A. thaliana*



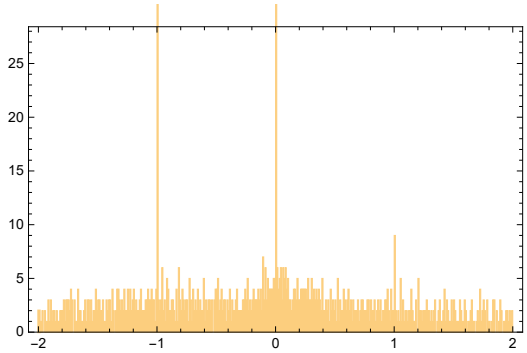
(k) *B. taurus*



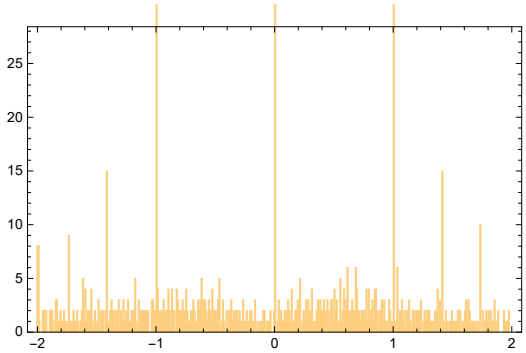
(l) *D. rerios*

Fonte: O autor.

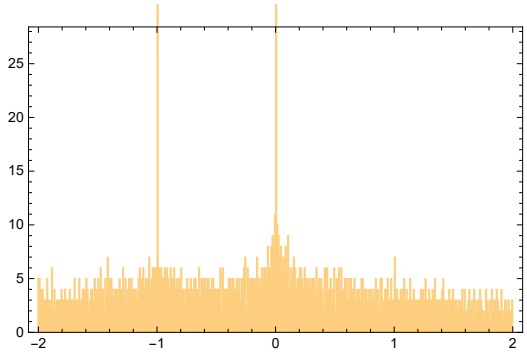
Figura E.2 – Espectro de autovalores das redes geradas pelo novo modelo



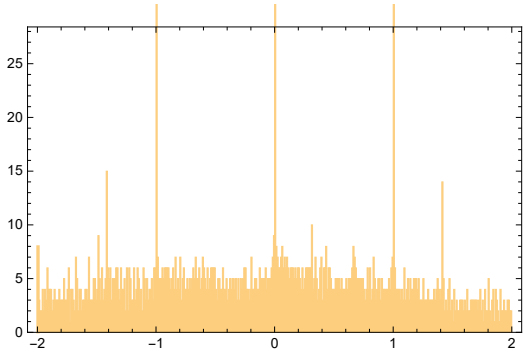
(a) *P. falciparum*



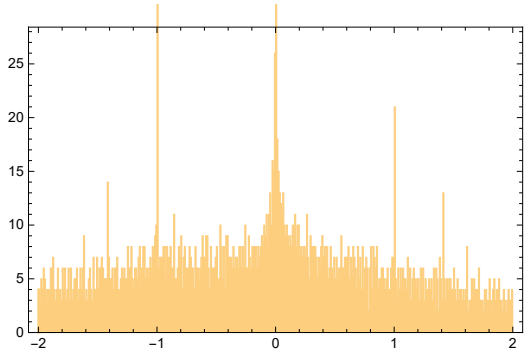
(b) *T. cruzi*



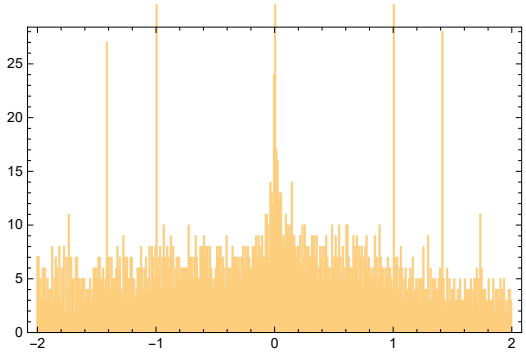
(c) *S. pombe*



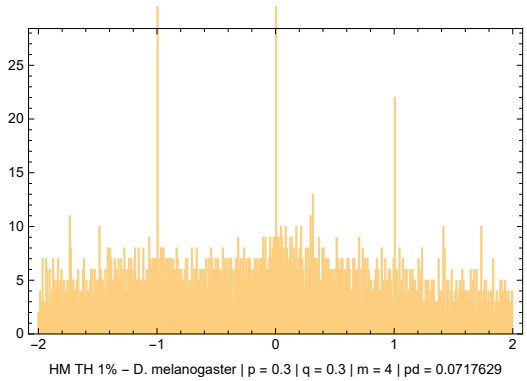
(d) *A. mellifera*



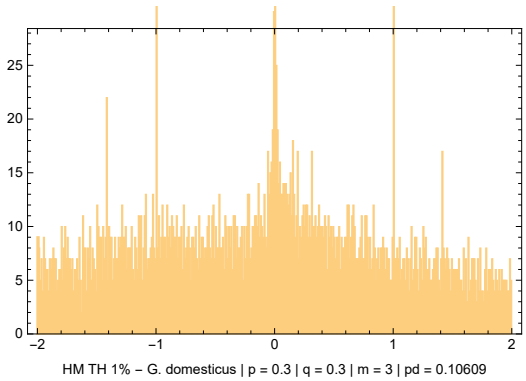
(e) *C. elegans*



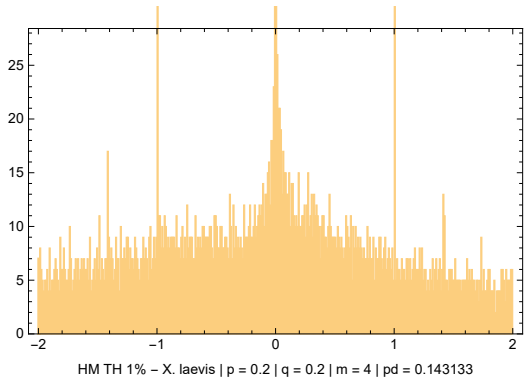
(f) *O sativa*



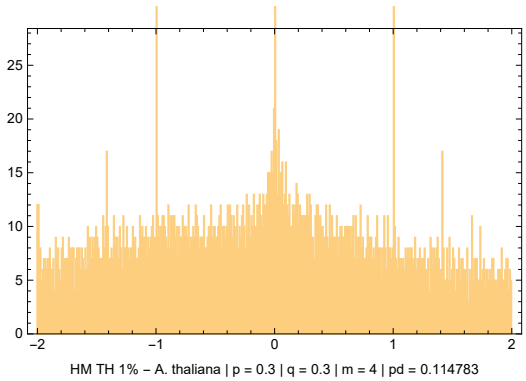
(g) *D. melanogaster*



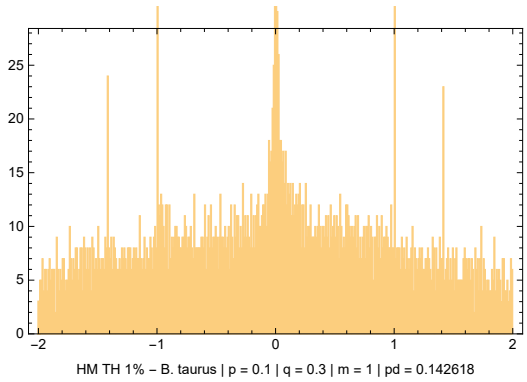
(h) *G. domesticus*



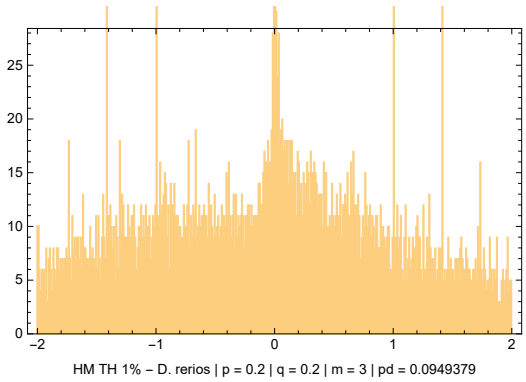
(i) *X. laevis*



(j) *A. thaliana*



(k) *B. taurus*

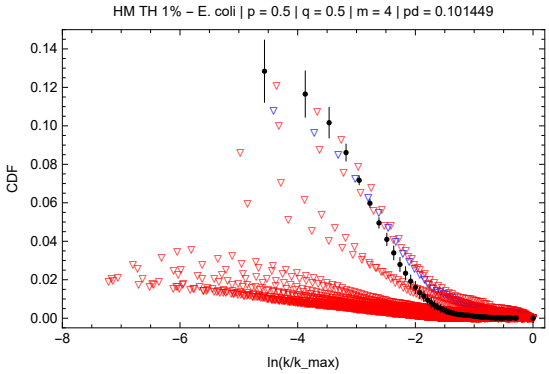


(l) *D. rerios*

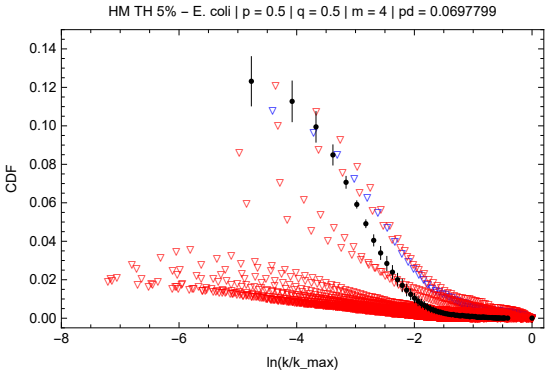
Fonte: O autor.

APÊNDICE F – EFEITOS DA TRANSFERÊNCIA HORIZONTAL SOBRE AS REDES DA *E. COLI* E DA *B. SUBTILIS*

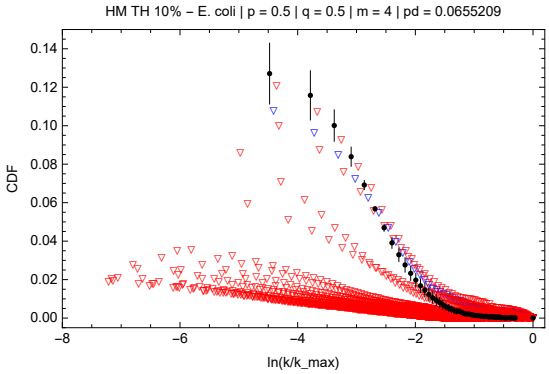
Figura F.1 – Efeitos da transferência horizontal sobre a topologia da rede da *E. coli*



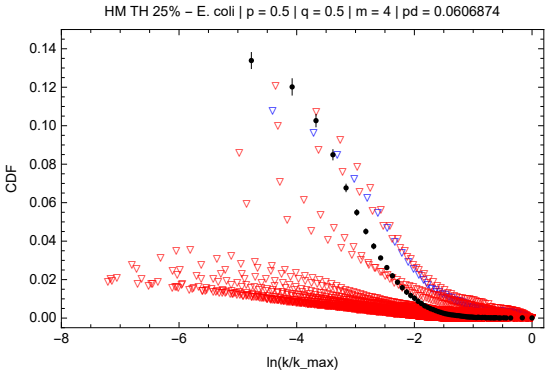
(a) $p_{TH} = 0.01$



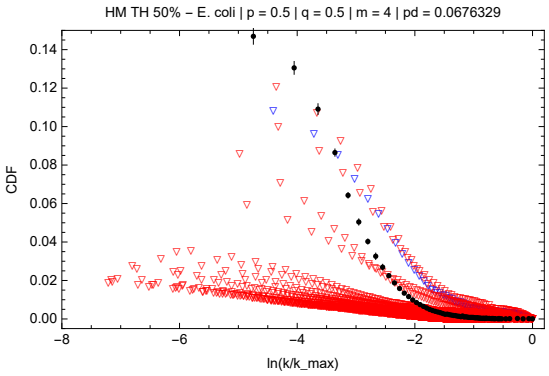
(b) $p_{TH} = 0.05$



(c) $p_{TH} = 0.10$



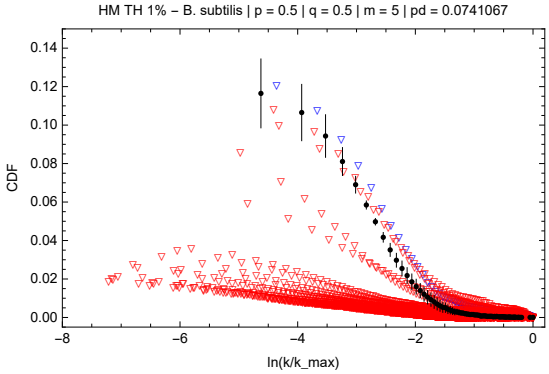
(d) $p_{TH} = 0.25$



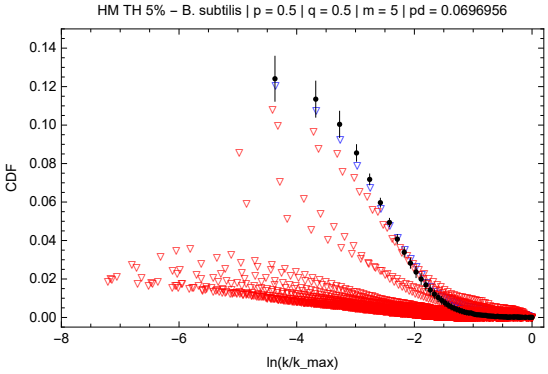
(e) $p_{TH} = 0.50$

Fonte: O autor.

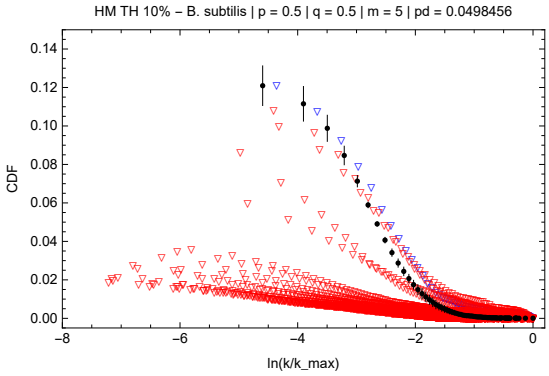
Figura F.2 – Efeitos da transferência horizontal sobre a topologia da rede da *B. subtilis*



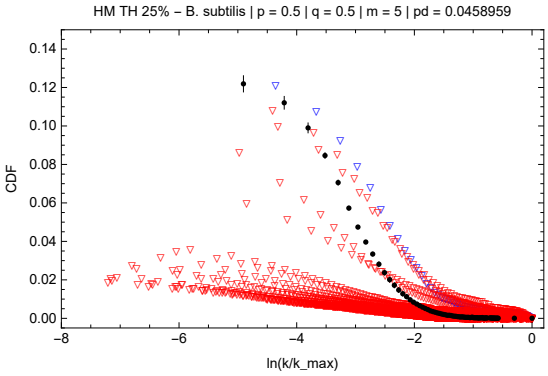
(a) $p_{TH} = 0.01$



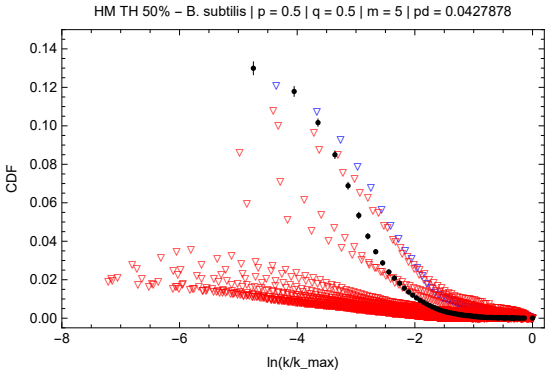
(b) $p_{TH} = 0.05$



(c) $p_{TH} = 0.10$



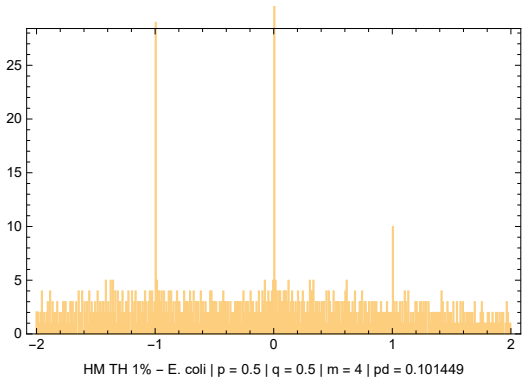
(d) $p_{TH} = 0.25$



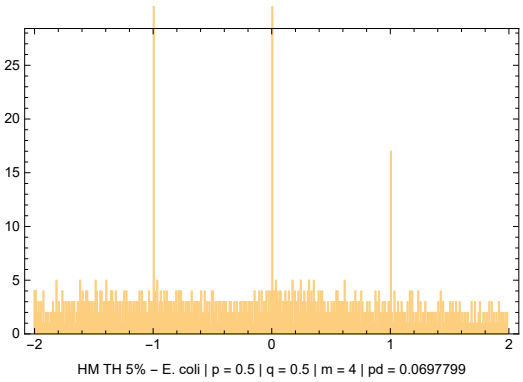
(e) $p_{TH} = 0.50$

Fonte: O autor.

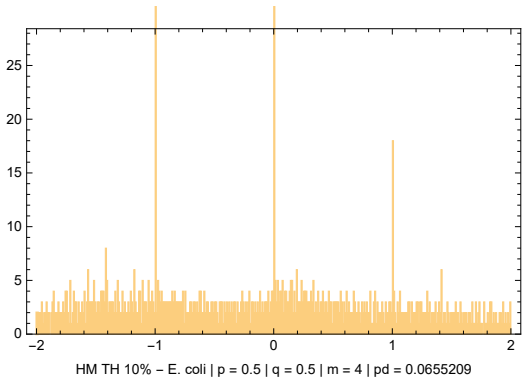
Figura F.3 – Efeitos da transferência horizontal sobre o espectro de autovalores da rede da *E. coli*



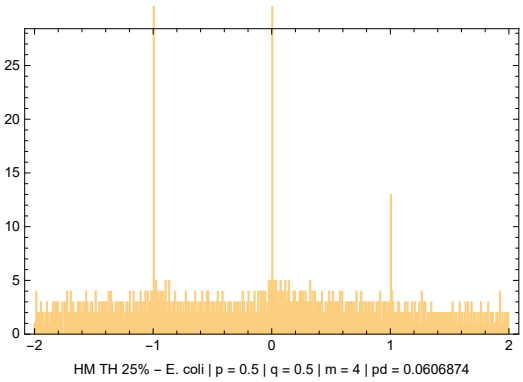
(a) $p_{TH} = 0.01$



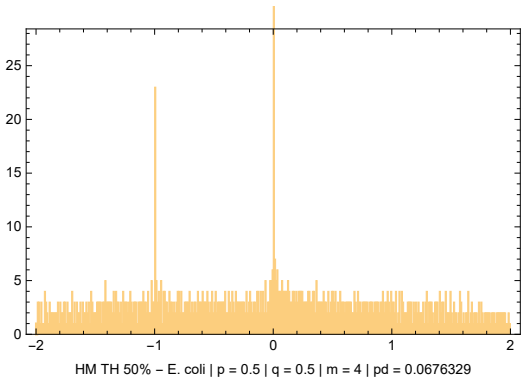
(b) $p_{TH} = 0.05$



(c) $p_{TH} = 0.10$



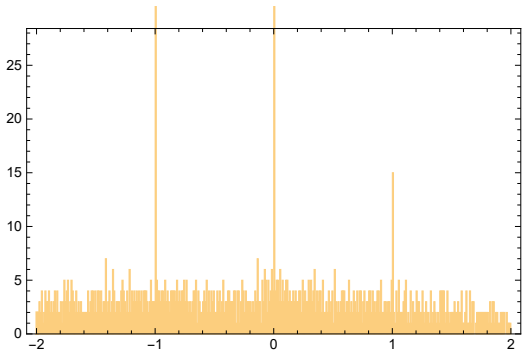
(d) $p_{TH} = 0.25$



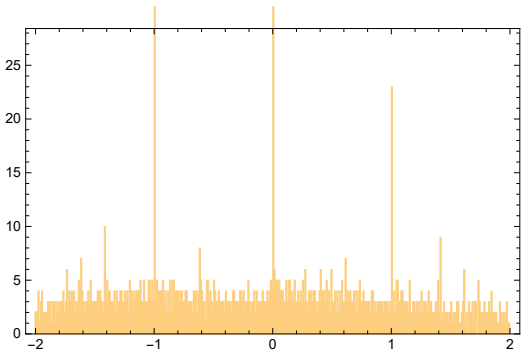
(e) $p_{TH} = 0.50$

Fonte: O autor.

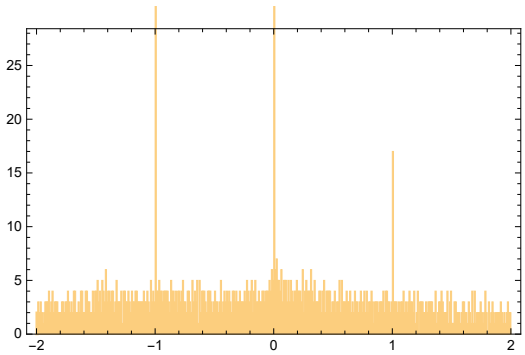
Figura F.4 – Efeitos da transferência horizontal sobre o espectro de autovalores da rede da *B. subtilis*



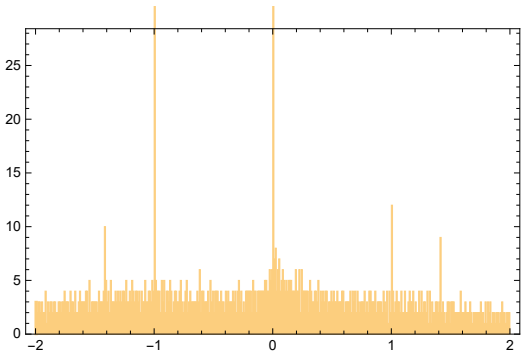
(a) $p_{TH} = 0.01$



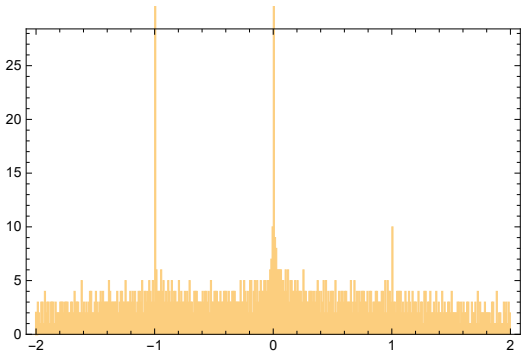
(b) $p_{TH} = 0.05$



(c) $p_{TH} = 0.10$



(d) $p_{TH} = 0.25$



(e) $p_{TH} = 0.50$

Fonte: O autor.

APÊNDICE G – ALGORITMOS DESENVOLVIDOS PARA A REPRODUÇÃO DAS REDES COM OS MODELOS ESTUDADOS

G.1 – ALGORITMO DESENVOLVIDO PARA O MODELO DE BARABÁSI E ALBERT

O algoritmo abaixo foi desenvolvido para podermos reproduzir as redes utilizando o modelo de Barabási e Albert. Ele inicia recebendo dois parâmetros: m e n, onde n é o tamanho final da rede e m é o número de *links* aos quais cada novo nó irá se conectar. Logo de início é inicializado um *array* g composto por n listas vazias, que é onde serão armazenadas os nós da rede. Logo na sequência são adicionados nas posições de 1 até $m + 1$ desse *array* os nós aos quais cada um desses $m + 1$ nós se conectará. Por exemplo, para uma rede com $m = 2$, teremos os elementos 1 a 3 de g da seguinte maneira: $\{\{2, 3\}, \{1, 3\}, \{1, 2\}, ...\}$, indicando que o nó 1 se conectará aos nós 2 e 3, o nó 2 aos nós 1 e 3 e o nó 3 aos nós 1 e 2.

A próxima etapa é a inicialização do vetor $vdeg$, com n posições e em cada posição um 0, que é onde serão armazenadas as conectividades de cada nó da rede. As posições iniciais são inicializadas para o valor de m, uma vez que cada nó da rede inicial terá m conexões. Pro nosso exemplo com $m = 2$, teremos $vdeg = \{2, 2, 2, ...\}$.

Após essa inicialização, começa o *loop* de execução do algoritmo, onde serão feitas as iterações necessárias para se atingir o tamanho final n da rede. Na linha onde é definida a variável $node$, utilizamos a função *RandomSample* para obter um número aleatório m de nós já existentes na rede. No entanto, a escolha desses nós se dá com base em um peso, que é o que temos ao fazer da seguinte maneira: $vdeg[[1;;j-1]]->Range[j-1]$, onde iremos selecionar do retorno da função *Range* os m nós de acordo com o peso dado a cada um deles por $vdeg[[1;;j-1]]$ simulando, assim, o *Preferential Attachment* pois os nós com conectividade maior terão maior preferência na hora de serem escolhidos pelo *RandomSample*.

Após a escolha dos nós, são atualizados os *arrays* g e $vdeg$, e no *loop Do* seguinte, são adicionadas as novas conexões dos nós escolhidos. Esse processo como um todo será executado até que a rede tenha o tamanho final esperado de n nós. Ao final, será retornado um objeto do tipo *Graph*, a partir do qual são extraídas as propriedades de interesse do trabalho.

```
barabasi[n_, m_] /; n >= 3 :=
    Module[{g, vdeg, el, el2, node, newnode},
      g = ConstantArray[{}, n];
      g[[1 ;; m + 1]] = Table[Delete[Range[m + 1], j], {j, m + 1}];
      vdeg = ConstantArray[0, n];
```

```
 vdeg[[1 ;; m + 1]] = m;
 Do[
     node = RandomSample[vdeg[[1 ;; j - 1]] -> Range[j - 1], m];
     g[[j]] = node;
     vdeg[[j]] = m;
     Do[vdeg[[k]] += 1; g[[k]] = Append[g[[k]], j], {k, node}]
, {j, m + 2, n}];

Graph[
   Union[Map[Sort, Flatten[MapIndexed[Thread[{#2[[1]], #1}] &, g], 1]]]
]
```

G.2 – ALGORITMO DESENVOLVIDO PARA O MODELO *DUPLICATION-DIVERGENCE* E O NOVO MODELO

Como no novo modelo utilizamos o mecanismo da duplicação da mesma forma que definido no modelo DD, foi utilizado o mesmo algoritmo, que está apresentado abaixo, para a reprodução das redes com esses dois modelos, pois podemos controlar pelos parâmetros de entrada qual(is) mecanismo(s) seguir.

O modelo recebe 5 parâmetros como entrada: n é o tamanho final da rede, d é a probabilidade de ocorrer a duplicação (e $1 - d$ a probabilidade de ocorrer a transferência horizontal), m é a quantidade máxima de *links* que o nó terá quando se conectar via transferência horizontal, p é a probabilidade de ocorrer uma ligação entre a proteína pai e a proteína filha no modelo DD e q é a probabilidade de ocorrer a mutação, também conforme modelo DD.

Da mesma forma que no algoritmo BA, começamos inicializando o vetor onde serão armazenados os nós da rede e com o vetor que armazena a conectividade de cada nó. A rede aqui, no entanto, inicia somente com dois nós conectados entre si, de modo que os elementos 1 e 2 de $vdeg$ sejam 1. Com o objetivo de otimizar o processamento, são inicializados dois vetores de tamanho n cada: $irand$ e $prand$. O vetor $irand$ será utilizado para determinar qual nó da rede será escolhido para duplicar. Assim, os seus n elementos são números inteiros que variam de 1 até n. Já o vetor $prand$ armazena as probabilidades de 0 a 1 de ocorrer a conexão entre a proteína pai e a proteína filha logo após duplicação. Essa inicialização foi feita para tentar melhorar o desempenho do algoritmo, de modo que não é necessário a cada nova etapa ficar gerando um número aleatório.

Após essa inicialização é que se dá início ao *loop* de execução do algoritmo. O primeiro passo é gerar um número aleatório para ver se o nó será adicionado por dupli-

cação ou por transferência horizontal, conforme o valor de d que definimos ao inicializar o algoritmo. É nesse ponto que se definirmos $d = 1$ poderemos utilizar o algoritmo para gerar redes do modelo DD somente, pois a condição *If[RandomReal[] <= d* será sempre verdadeira, uma vez que a função *RandomReal[]* retorna números aleatórios entre 0 e 1.

Caso o mecanismo escolhido seja a duplicação, seleciona-se o valor na posição j do vetor $irand$ que será o nó escolhido para ser duplicado. Esse nó é armazenado na variável u que será utilizada para pegar o elemento na posição u do vetor g, criando assim o vetor gi que terá as informações referentes aos nós que esse novo nó irá se conectar. Após essa escolha do nó é feita a verificação se haverá ou não a criação do *link* entre nó pai e filho, e após isso é iniciado um novo *loop* que é onde será feita a parte que envolve a mutação. Após a execução desse loop, a conectividade dos nós é alterada de acordo com a necessidade e o que temos no vetor g é atualizado.

Caso o mecanismo escolhido seja a transferência horizontal, o mecanismo utilizado aqui é bem semelhante ao que foi apresentado na seção acima para o modelo de BA. A única diferença é que aqui o novo nó irá se conectar a um número aleatório de nós já existentes na rede, sendo o número máximo definido pelo parâmetro m passado na inicialização do algoritmo. A quantidade de links que esse novo nó terá está representada na variável *links*, que assumirá um valor entre 1 e m (ou quando a rede ainda for menor que m, terá um valor entre 1 e $j - 1$ que é o tamanho da rede naquele instante). Essa parte é feita no primeiro *If* que temos no algoritmo da TH. Logo após, o *Preferential Attachment* e as demais implementações são feitas da mesma forma que já foi explicado para o modelo BA.

```
newModel[n_, d_, m_, p_, q_] /; n >= 3 :=
 Module[{g, vdeg, el, el2, node, newnode, links, lcluster, u, v, gi,
   ginew, prand, qrand, irand, cointosses},
  g = ConstantArray[{}, n];
  g[[1]] = {2};
  g[[2]] = {1};
  vdeg = ConstantArray[0, n];
  vdeg[[1 ;; 2]] = 1;
  irand = 1 + RandomInteger /@ Range[0, n];
  prand = RandomReal[{0, 1}, n];
  Do[If[RandomReal[] <= d,
     (*duplicação*)
     u = irand[[j]];
     gi = g[[u]];
     ginew = gi;
     If[prand[[j]] <= p, AppendTo[gi, j];
      AppendTo[ginew, u];];
```

```
    Do[v = gi[[k]];
     qrand = RandomReal[{0, 1}, Length[gi]];
     cointosses = RandomInteger[{0, 1}, Length[gi]];
     If[qrand[[k]] <= q,
      If[cointosses[[k]] == 1, g[[v]] = g[[v]] /. u -> j;
        gi = Delete[gi, k];, ginew = Delete[ginew, k];];,
      g[[v]] = Append[g[[v]], j];], {k, Length[gi], 1, -1}];
    g[[u]] = gi;
    g[[j]] = ginew;
    vdeg[[1 ;; j]] = Length /@ g[[1 ;; j]];
    ,
    (*transferência horizontal*)
    If[m > (j - 1), links = RandomInteger[{1, j - 1}],
     links = RandomInteger[{1, m}]];
    node = RandomSample[(vdeg[[1 ;; j - 1]] + 1) -> Range[j - 1],
      links];
    g[[j]] = node;
    vdeg[[j]] = links;
    Do[vdeg[[k]] += 1; g[[k]] = Append[g[[k]], j], {k, node}]];
  , {j, 3, n}];
Graph[Union[
  Map[Sort, Flatten[MapIndexed[Thread[{#2[[1]], #1}] &, g], 1]]]]]
```

ANEXO A – DERIVANDO A DISTRIBUIÇÃO DE *POISSON* A PARTIR DA DISTRIBUIÇÃO BINOMIAL

Partimos da equação da distribuição binomial, dada pela Equação 1.14:

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k} \tag{A.1}$$

Vamos analisar separadamente cada um dos três termos do lado direito da Equação A.1. Começando com o primeiro termo, reescrevemos ele como:

$$\binom{N-1}{k} = \frac{(N-1)(N-1-1)(N-1-2)\ldots(N-1-k+1)}{k!} \approx \frac{(N-1)^k}{k!} \tag{A.2}$$

onde foi usado que $k \ll N$. Podemos simplificar o último termo de A.1 tomando o seu logaritmo:

$$ln[(1-p)^{N-1-k}] = (N-1-k)ln(1-p) \tag{A.3}$$

Usando a Equação 1.13 podemos escrever p como:

$$p = \frac{\langle k \rangle}{(N-1)} \tag{A.4}$$

Substituindo a Equação A.4 na Equação A.3 temos:

$$ln[(1-p)^{N-1-k}] = (N-1-k)ln\left(1 - \frac{\langle k \rangle}{N-1}\right) \tag{A.5}$$

Usando a expansão em série para o logaritmo:

$$ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots, \forall |x| \leq 1 \tag{A.6}$$

De modo que é possível escrever:

$$ln[(1-p)^{N-1-k}] \approx -(N-1-k)\frac{\langle k \rangle}{N-1} = -\langle k \rangle \left(1 - \frac{k}{N-1}\right) \approx -\langle k \rangle \tag{A.7}$$

aproximação esta que é válida quando $N \gg k$, que representa a aproximação de pequeno grau. Logo, tomando a exponencial de ambos os lados da Equação A.7, podemos escrever o último termo da Equação A.1 como:

$$(1-p)^{N-1-k} = e^{-\langle k \rangle} \tag{A.8}$$

Por fim, substituindo as Equações A.2, A.4 e A.8 na Equação A.1 teremos:

$$p_k = \frac{(N-1)^k}{k!}\left(\frac{\langle k\rangle}{N-1}\right)^k e^{-\langle k\rangle} \tag{A.9}$$

que pode ser reescrita como:

$$p_k = \frac{\langle k\rangle^k}{k!}e^{-\langle k\rangle} \tag{A.10}$$

que é a expressão apresentada na Equação 1.15.