



Tiago Felipe Andrade

Análise de propriedades topológicas das
redes biológicas integradas da
Escherichia coli e da *Saccharomyces
cerevisiae*

Botucatu

2008

Tiago Felipe Andrade

Análise de propriedades topológicas das
redes biológicas integradas da
Escherichia coli e da *Saccharomyces
cerevisiae*

Monografia apresentada ao Instituto de Biociências da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Botucatu, para obtenção do título de Bacharel em Física Médica.

Orientador: Prof. Dr. Ney Lemke

Botucatu

2008

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉCNICA DE AQUISIÇÃO E TRATAMENTO
DA INFORMAÇÃO
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: SELMA MARIA DE JESUS

Andrade Tiago Felipe.

Análise de propriedades topológicas das redes biológicas integradas da *Escherichia coli* e da *Saccharomyces cerevisiae* / Tiago Felipe Andrade. - Botucatu [s.n], 2008.

Trabalho de conclusão (bacharelado – Física médica) – Universidade Estadual Paulista, Instituto de Biociências de Botucatu, 2008
Orientador: Ney Lemke

1. Física médica 2. Bioinformática

Palavras-chave: Bioinformática; Propriedades topológicas; Redes biológicas integradas

Agradecimentos

Primeiramente, quero agradecer a Deus, por ter me dado o dom da vida, e pelas oportunidades que me propiciou, pois sem a Sua graça e compaixão, sei que seria impossível chegar até aqui. Agradeço aos meus pais, Theodoro e Lília, pela educação que me concederam, pelo apoio intelectual, espiritual e financeiro, pelo amor a mim dedicado e pela atenção dada a cada dia da minha vida. As minhas irmãs, Débora e Ana, que me apoiaram em todos os momentos da vida, ajudando a solidificar a minha personalidade e o meu caráter. Agradeço a minha família pelo amor, respeito, confiança e credibilidade que depositaram em mim, bem como pelo auxílio que me foi dado nas horas mais difíceis desses quatro anos. Aos meus amigos de faculdade, em especial: Diogo Saraiva e Cyro Mode. Vocês realmente são os irmãos que não tive. Agradeço também aos amigos que fiz no decorrer desses anos, e aos colegas pela diversão e entretenimento. Impossível esquecer, meus amigos da República CB, que se tornaram minha família nesses quatro anos: Guilherme, Henrique, Rafael, Matheus, Brunno, Alexandre, Fábio, Daniel, Dudu e Edu, muito obrigado por terem me acompanhado nesses anos, por terem me agüentado todos os dias em que estive mal humorado e por terem compartilhado das minhas alegrias. Quero também agradecer ao amor a mim dedicado pela minha namorada e companheira Nabila, o qual me fez vencer muitas batalhas no meu dia a dia, bem como sua amizade, companheirismo e confiança. Agradecimentos especiais ao meu orientador, Ney Lemke, por ter confiado em mim e me dado credibilidade para que esse trabalho pudesse ser feito, colaborando assim com a minha carreira profissional e porque sem ele a concretização desse trabalho não seria possível. Aos companheiros do Laboratório de Bioinformática, Marcio e Luiz Augusto, pela ajuda no dia a dia de trabalho. E finalmente não poderia deixar de agradecer ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) por ter financiado durante um ano esse projeto. A todos que, de alguma forma, contribuíram para a realização desse trabalho, o meu "Muito Obrigado"!

*“Direi do Senhor:
Ele é o meu Deus,
o meu refúgio e a minha fortaleza e
Nele confiarei.”*

Sl. 91.2

Resumo

Processos biológicos são complexos e possuem propriedades emergentes que não podem ser explicadas ou previstas através do reducionismo. Para suplantar esses limites, pesquisadores têm usado um conjunto de métodos conhecido como biologia sistêmica, cujo objetivo é a compreensão das interações não-lineares entre os múltiplos componentes dos processos biológicos. Essas interações podem ser representadas por um objeto matemático chamado grafo ou rede, onde os elementos interagentes são representados por nodos e as interações por ligações que conectam pares de nodos. As redes podem ser classificadas segundo sua topologia, podendo ser aleatórias ou livres de escala, sendo que nessa última o número de conexões de um nodo segue a lei de potência, existindo poucos nodos altamente conectados e muitos nodos com poucas ligações. Elas também são classificadas como hierárquicas ou não, e segundo essa classificação respeita-se uma ordem de conexão entre os nodos e seus graus de conectividade, que define essa hierarquia. Neste trabalho, analisamos as redes moleculares integradas da bactéria *Escherichia coli* e da levedura *Saccharomyces cerevisiae*, que incluem interações físicas entre proteínas, interações metabólicas e interações de regulação transcricional. Através de ferramentas computacionais, como *Mathematica*[®], e dos dados obtidos de bancos de dados públicos, realizamos a medição de quatro parâmetros topológicos: a distribuição de conectividades $P(k)$, o coeficiente de agrupamento médio de todos os nodos com k conexões $C(k)$, o grau de proximidade $C_C(k)$ e o grau de intermediação $C_B(k)$. O $P(k)$ é uma função que calcula o número total de nodos com grau k de conexão e serve para classificar a rede como aleatória ou livre de escala. O $C(k)$ mostra se uma rede é hierárquica, isto é, se o agrupamento dos nodos depende da conectividade dos mesmos. O $C_C(k)$ mede quanto um nodo particular está próximo de todos os outros nodos da rede e o $C_B(k)$ mede quanto um nodo está nos menores caminhos entre outros nodos na rede. Na análise comparativa das redes verificamos que as redes biológicas integradas dos organismos *Escherichia coli* e *Saccharomyces cerevisiae* são, respectivamente, livre de escala e exponencial. Além disso, a dependência do $C(k)$ em relação aos valores de k de ambas redes integradas sugere que elas são hierárquicas, mas com uma estrutura hierárquica diferente das redes que contêm apenas um tipo de interação.

Palavras-chave: bioinformática, redes biológicas integradas, propriedades topológicas.

Abstract

Biological processes are complex and possess emergent properties that can not be explained or predict by reductionism methods. To overcome the limitations of reductionism, researchers have been used a group of methods known as systems biology, a new interdisciplinary field of study aiming to understand the non-linear interactions among components embedded in biological processes. These interactions can be represented by a mathematical object called graph or network, where the elements are represented by nodes and the interactions by edges that link pair of nodes. The networks can be classified according to their topologies: if node degrees follow a Poisson distribution in a given network, i.e. most nodes have approximately the same number of links, this is a random network; if node degrees follow a power-law distribution in a given network, i.e. small number of high-degree nodes and high number of low-degree nodes, this is a scale-free network. Moreover, networks can be classified as hierarchical or non-hierarchical. In this study, we analysed *Escherichia coli* and *Saccharomyces cerevisiae* integrated molecular networks, which have protein-protein interaction, metabolic and transcriptional regulation interactions. By using computational methods, such as *Mathematica*[®], and data collected from public databases, we calculated four topological parameters: the degree distribution $P(k)$, the clustering coefficient $C(k)$, the closeness centrality $C_C(k)$ and the betweenness centrality $C_B(k)$. $P(k)$ is a function that calculates the total number of nodes with k degree connection and is used to classify the network as random or scale-free. $C(k)$ shows if a network is hierarchical, i.e. if the clusterization coefficient depends on node degree. $C_C(k)$ is an indicator of how much a node it is in the lesse way among others some nodes of the network and the $C_B(k)$ is a pointer of how a particular node is among several other nodes of the network. In the comparative analysis, we verified that the integrated biological networks of organisms *Escherichia coli* and *Saccharomyces cerevisiae* are, respectively, scale-free and exponential. Moreover, the dependence of $C(k)$ on k observed for both networks suggests they are hierarchical, althought their structures appear to be distinct from the hierarchical structure of networks containing only one type of interaction.

Key-words: bioinformatics, integrated biological networks, topological properties.

Sumário

	Página
Resumo	4
Abstract	5
1 Introdução	7
2 Objetivos	11
3 Metodologia	12
3.1 Bancos de dados	12
3.2 Definições das propriedades topológicas de redes	13
3.3 Implementação	14
4 Resultados e Discussão	16
4.1 Rede integrada da <i>E. coli</i>	16
4.2 Rede integrada da <i>S. cerevisiae</i>	19
4.3 Análise comparativa do grau de proximidade e do grau de intermediação para as redes integradas da <i>S. cerevisiae</i> e da <i>E. coli</i>	21
5 Conclusão	25

1 Introdução

O método reducionista de dissecação dos sistemas biológicos em suas partes constituintes tem ajudado no esclarecimento do funcionamento de muitos processos biológicos. Porém, tais processos são extremamente complexos e possuem propriedades emergentes que não podem ser explicadas ou mesmo previstas através do estudo de suas partes individuais [1]. Tal limitação imposta pelo método reducionista é um dos fatores preponderantes na falta de uma melhor compreensão e desenvolvimento de terapias eficazes para doenças complexas [2].

Para suplantarmos esses limites do reducionismo, pesquisadores têm usado um conjunto de métodos que tratam os processos biológicos de forma integrada. Esta nova área da biologia é conhecida como biologia sistêmica. A biologia sistêmica objetiva a compreensão das interações não-lineares entre os múltiplos componentes dos processos biológicos [2].

Tais interações são geralmente representadas por um objeto matemático chamado grafo ou rede [3], que é um conjunto de nós G (componentes) e um conjunto de arestas A (interações) que conectam cada dois componentes no conjunto G .

Só recentemente começamos a desvendar a arquitetura das redes naturais que controlam as interações entre os componentes de um sistema complexo, como por exemplo, a rede de computadores na internet, a rede de reações químicas em uma célula (metabolismo), as redes de distribuição elétrica, a rede de colaborações entre cientistas, atores de cinema, parceiros sexuais e outras [4, 5, 6]. As redes formam o esqueleto dos sistemas complexos que permeiam todos os ramos do conhecimento humano e são um tema genuinamente transdisciplinar [7].

O estudo de redes naturais pertence à teoria de grafos aleatórios. Um grafo aleatório é definido como um conjunto de N nós conectados por n arestas que são escolhidas aleatoriamente entre as $N(N-1)/2$ possíveis arestas. Existe um total de:

$$C_{N(N-1)/2}^n = \frac{(N(N-1)/2)!}{n!(N(N-1)/2 - n)!} \quad (1)$$

grafos diferentes que podem ser criados com N nós e n arestas, todos equiprováveis estatisticamente. A Figura 1(Aa) mostra a ilustração de um grafo aleatório simples.

Um grafo aleatório também pode ser definido pelo modelo binomial, no qual um par de nós qualquer é conectado com uma probabilidade p . Desta maneira, o número esperado de arestas no grafo será:

$$E(n) = p[N(N-1)/2]. \quad (2)$$

A distribuição de conectividades de um grafo aleatório obedece uma distribuição bi-

nomial, mas para N grande, ela é bem aproximada por uma distribuição de Poisson:

$$P(k) \simeq e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}. \quad (3)$$

Onde $\langle k \rangle$ é a conectividade média dos nodos no grafo. Ver Figura 1(Ab).

Algumas características importantes das redes naturais são muito diferentes das propriedades conhecidas para grafos aleatórios como o mundo pequeno, o agrupamento e a distribuição de conectividades. A característica de mundo pequeno é conhecida de grafos aleatórios que, apesar do grande tamanho que uma rede aleatória possa ter, o comprimento médio de caminho entre dois nodos quaisquer do grafo cresce muito lentamente (logaritmicamente) com o seu número de nodos, implicando que sempre há alguma trajetória curta unindo dois nodos quaisquer no grafo. O coeficiente de agrupamento está relacionado à formação de módulos altamente conectados no interior de um grafo. Para redes naturais, a formação destes módulos parece ser freqüente, enquanto que para redes aleatórias eles são raros. Já a distribuição de conectividades para um bom número de redes naturais tende a seguir uma lei de potência que caracteriza a ausência de uma escala típica no sistema, diferentemente da distribuição para grafos aleatórios que obedece uma distribuição de Poisson.

A representação de sistemas biológicos por grafos aleatórios direcionados e não direcionados e a aplicação das propriedades topológicas desses grafos no estudo desses sistemas foram iniciadas por Jeong et al. (2000). Nesse trabalho pioneiro, as vias metabólicas de vários organismos foram organizadas em forma de rede e as propriedades topológicas dessas redes metabólicas foram analisadas e biologicamente interpretadas. A partir desse trabalho, a representação dos sistemas biológicos em redes e suas respectivas análises vêm sendo feitas com maior freqüência para elucidação de propriedades emergentes de processos biológicos complexos [8, 9, 10].

Estudos recentes têm sido baseados em um novo conceito sobre redes biológicas, o qual as defini como sendo integradas, ou seja, redes biológicas que possuem mais de um tipo de interação física. Em uma rede biológica integrada[11], os genes representam os nodos, e considerando que os genes $g1$ e $g2$ codificam as proteínas $p1$ e $p2$, estes são conectados na rede, se ocorre:

- a) interação proteína-proteína: $p1$ e $p2$ interagem fisicamente;
- b) regulação transcricional: $g1$ regula a transcrição do gene $g2$ ou
- c) interação metabólica: um produto gerado pela reação catalisada pela proteína $p1$ é consumido na reação catalisada pela proteína $p2$.

Neste trabalho analisamos as redes biológicas integradas da *Saccharomyces cerevisiae*

e da *Escherichia coli* observando as seguintes propriedades topológicas: o grau de conectividade $P(k)$, o coeficiente de agrupamento médio $C(k)$, o grau de proximidade $C_C(k)$ e o grau de intermediação $C_B(k)$. Tais propriedades topológicas serão definidas no item *Metodologia* e suas análises objetivam encontrar a essencialidade de genes, e assim definir sua importância para o organismo em estudo.

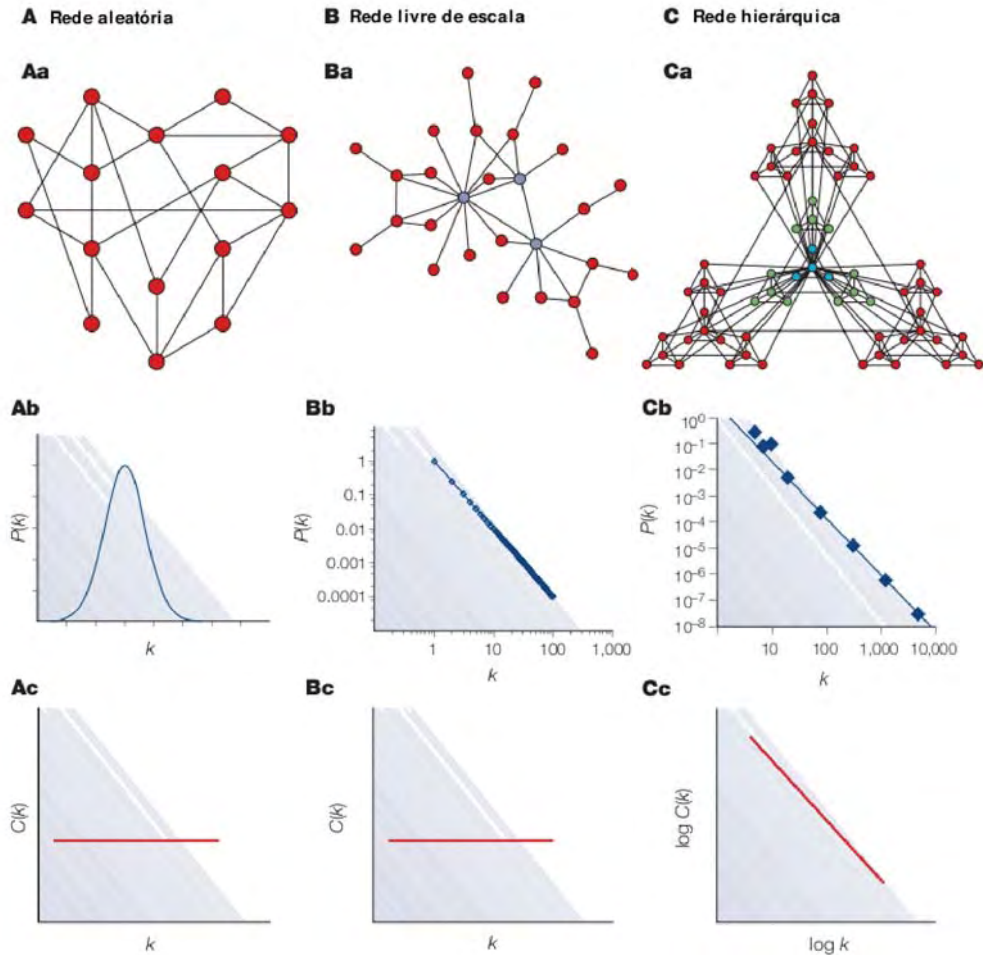


Figura 1: Modelos de redes. (Aa) Rede aleatória. (Ab) Distribuição de Poisson das conectividades de uma rede aleatória. (Ac) Coeficiente de agrupamento $C(k)$ independente de k em uma rede aleatória. (Ba) Rede com distribuição de conectividades obedecendo lei de potência. (Bb) A distribuição de conectividades representada em um gráfico Log-Log é linear. (Bc) Coeficiente de agrupamento $C(k)$ independente da conectividade k em uma rede com distribuição de conectividades obedecendo uma lei de potência. (Ca) Rede hierárquica: possui, simultaneamente, distribuição de conectividades obedecendo uma lei de potência e $C(k)$ dependente de k . (Cb) A distribuição de conectividades representada em um gráfico Log-Log é linear. (Cc) $C(k)$ tem dependência em k . Extraído de [4].

Parte dos resultados desse trabalho foi apresentado em 4 eventos, sendo eles:

1. 7th International Symposium on Mathematical and Computational Biology, 2007, Búzios-RJ, com trabalho intitulado: Network topology-based *in silico* discovery of gene essentiality in *Saccharomyces cerevisiae*. Foi apresentado sob a forma de pôster pelo aluno de doutorado Marcio Luis Acencio, e seu resumo consta nos anais desse evento.
2. VII Workshop da Pós-Graduação, 2008, Botucatu-SP, com o trabalho intitulado: Estudos comparativos das propriedades das redes integradas da *E. coli* e da *S. cerevisiae*. Apresentado sob a forma de pôster pelo aluno de iniciação científica Tiago Felipe Andrade e cujo resumo consta nos anais do evento.
3. XXXI Encontro Nacional de Física da Matéria Condensada, 2008, Águas de Lindóia-SP, com trabalho intitulado: Motivos topológicos na rede integrada da bactéria *E. coli* e da levedura *S. cerevisiae*, apresentado em sessão oral pelo professor doutor Ney Lemke, e seu resumo está incluído nos anais desse evento.
4. 3rd International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2007, São Paulo-SP, com trabalho intitulado Towards the *In Silico* Construction of the Anchorage-Independent Cell Cycle Start Network, apresentado sob a forma de pôster pelo aluno de doutorado Marcio Luis Acencio, e seu resumo consta nos anais desse evento.

2 Objetivos

Construir as redes biológicas integradas da *Escherichia coli* e *Saccharomyces cerevisiae*, e analisá-las através da medição e interpretação de quatro parâmetros topológicos: a distribuição de conectividades, o coeficiente de agrupamento médio de todos os nodos com k conexões, o grau de proximidade e o grau de intermediação.

3 Metodologia

O trabalho foi desenvolvido em etapas, devido a grande demanda de tempo, as quais estão descritas detalhadamente abaixo.

3.1 Bancos de dados

O trabalho teve seu início com a localização dos bancos de dados biológicos públicos na *web*, que comportam os dados de interações físicas entre proteínas, interações metabólicas e interações de regulação transcricional, referente aos organismos estudados. Os dados em forma de texto foram adquiridos por meio de *download*, o que demandou grande parte do tempo desse projeto, pois muitos dados encontram-se esparsos, o que dificulta a coleta dos mesmos. A Tabela 1 apresenta os bancos de dados utilizados para a aquisição dos dados.

Tabela 1: Lista dos bancos de dados biológicos utilizados para o processo de aquisição dos dados.

Banco de dados	Organismo	Tipo de interação	Referência
BIGG	<i>S. cerevisiae</i>	Metabólica	http://bigg.ucsd.edu/home.pl
YEASTRACT	<i>S. cerevisiae</i>	Regulação transcricional	http://www.yeasttract.com
BIOGRID	<i>S. cerevisiae</i>	Proteína-proteína	http://www.thebiogrid.org
KEGG	<i>E. coli</i>	Metabólica	http://www.genome.jp/kegg
RegulonDB	<i>E. coli</i>	Regulação transcricional	http://regulondb.ccg.unam.mx
Bacteriome.org	<i>E. coli</i>	Proteína-proteína	http://www.bacteriome.org

Feita a coleta dos dados, a segunda etapa foi normalizar e integrar os dados. O objetivo da normalização é colocar os dados em um formato padrão único de forma que os mesmos possam ser integrados. Essa etapa foi realizada manualmente, normalizando os nomes dos genes para um formato único.

Com o processo de normalização concluído, executou-se o próximo passo que foi a integração dos dados. Para a execução dessa etapa fez-se necessário a utilização de uma abordagem de integração, e a abordagem utilizada nesse trabalho foi a abordagem materializada, onde os dados são extraídos dos bancos de dados e posteriormente centralizados em um único arquivo, de forma a permitir o estudo e análise destes dados em um processo minucioso.

A terceira etapa realizada foi a limpeza dos dados, que objetiva eliminar ou corrigir dados incorretos ou imprecisos. Esta etapa foi executada por um especialista da área

biológica.

3.2 Definições das propriedades topológicas de redes

a) Distribuição de conectividades

A distribuição de conectividades, $P(k)$, é uma função que calcula o número total de nodos com determinado grau de conexão, em um dado grafo. Formalmente, a distribuição de conectividades é:

$$P(k) = \sum_{v \in V | \deg(v)=k} 1 \quad (4)$$

onde v é um nodo do grupo de nodos V pertencentes ao grafo, e $\deg(v)$ é o grau do nodo v . Essa mesma informação é freqüentemente representada como uma distribuição de conectividades cumulativa:

$$P(k) = \sum_{k'=k} p(k'). \quad (5)$$

Existem diferentes tipos de redes complexas, e cada uma delas possui diferentes características em relação à distribuição de conectividades. Por exemplo, a rede aleatória de Erdős-Rényi possui uma distribuição binomial, enquanto redes livres de escala, possuem uma distribuição representada por uma lei de potência [4]:

$$P(k) \sim k^{-\gamma} \quad (6)$$

o que indica que existem poucos nodos altamente conectados, denominados “*hubs*”.

b) Coeficiente de agrupamento médio

O coeficiente de agrupamento médio, $C(k)$, caracteriza a densidade de conexões próxima a um determinado nodo. Supondo uma rede não direcionada, e que um de seus nodos possuem z nodos vizinhos próximos, a conexão máxima desse nodo é aproximada quando todas $z(z+1)/2$ ligações no nodo estão presentes. Isso também significa que todas $z(z+1)/2$ possíveis ligações entre os nodos vizinhos próximos existem. Convencionalmente [12, 13, 14], o coeficiente de agrupamento, $C(k)$, de um nodo é a relação entre o número total y de ligações que conectam nodos vizinhos próximos e o número total de todas as possíveis ligações entre todos os vizinhos

próximos, representado matematicamente por:

$$C = \frac{2y}{z(z-1)}. \quad (7)$$

Através da análise do coeficiente de agrupamento, pode-se definir se uma rede biológica integrada é hierárquica, isto é, se o agrupamento dos nodos depende da conectividade dos mesmos.

c) Grau de proximidade

Grau de proximidade, $C_C(k)$, mede quanto um nodo particular está próximo de todos os outros nodos da rede. Para calcular o grau de proximidade soma-se a distância geodésica do nodo em relação a todos os demais nodos do grafo e depois inverte-se, uma vez que quanto maior a distância menor a proximidade, e esse valor ainda é normalizado em relação ao nodo de menor valor. Matematicamente, ela é representada por:

$$C_C(v) = \frac{1}{\sum_{t \in V/v} d_G(v, t)} \quad (8)$$

onde $d_G(v, t)$ é a distância entre v e t .

d) Grau de intermediação

Grau de intermediação, $C_B(k)$, mede quanto um nodo está nos menores caminhos entre outros nodos na rede, ou seja, a fração dos trajetos mais curtos que incluem o nodo i . Também é considerado uma medida de relevância do nodo, e é representado matematicamente por:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (9)$$

onde $\sigma_{st}(v)$ é o número de trajetos mais curtos de s até t que inclui v e σ_{st} é o número de trajetos de s até t .

3.3 Implementação

Concluídos todos esses passos e com os dados disponíveis, analisamos as redes biológicas integradas da *E. coli* e da *S. cerevisiae* realizando a medição de quatro parâmetros topológicos: a distribuição de conectividades $P(k)$, o coeficiente de agrupamento médio $C(k)$, o grau de proximidade $C_C(k)$ e o grau de intermediação $C_B(k)$, os quais encontram-se definidos no itens seguintes. Para o cálculo dos parâmetros topológicos, implementamos através dos programas *Mathematica*[®] e *Python* e de pacotes de funções como *Combina-*

torica, *GraphUtilities* e *NetworkX*. Os *notebooks* estão disponíveis no *site* [15] e descritos abaixo:

- a) gernet.nb: gera uma rede através da importação de um arquivo com os dados;
- b) pk.nb: aplicativo que calcula os dados de $P(k)$ em função de k para a *E. coli* e para a *S. cerevisiae*;
- c) ck.nb: aplicativo que calcula os dados de $C(k)$ em função de k para a *E. coli* e para a *S. cerevisiae*;
- d) cent.nb: aplicativo que calcula os dados de $C_C(k)$ em função de k para a *E. coli* e para a *S. cerevisiae*;
- e) bet.nb: aplicativo que calcula os dados de $C_B(k)$ em função de k para a *E. coli* e para a *S. cerevisiae*.

4 Resultados e Discussão

As redes integradas contendo interações físicas entre proteínas, interações metabólicas e interações de regulação transcricional dos organismos *E. coli* e *S. cerevisiae* construídas possuem 2.349 nodos e 19.908 interações e 6.115 nodos e 82.994 interações, respectivamente. Com o intuito de caracterizar a estrutura topológica dessas redes, ou seja, classificá-las em aleatórias ou livres de escala e determinar se são hierárquicas, nós analisamos quatro parâmetros: a distribuição de conectividades, o coeficiente de agrupamento médio, o grau de proximidade e o grau de intermediação.

4.1 Rede integrada da *E. coli*

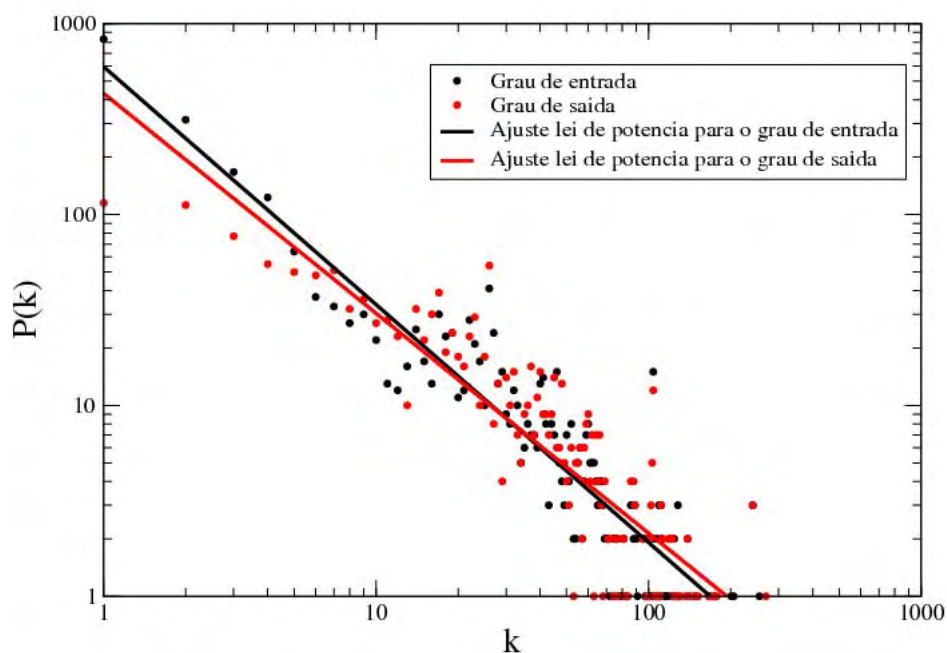


Figura 2: Distribuição de conectividades da rede integrada de interações moleculares entre genes da *E. coli* construída neste projeto. Essa distribuição segue uma lei de potência, o que a caracteriza como uma rede livre de escala.

Os ajustes realizados nas distribuições de conectividades da rede integrada de interações moleculares entre genes da *E. coli* seguem uma lei de potência representada por:

$$y = \alpha x^{-\beta} \quad (10)$$

onde α e β são constantes.

Os ajustes utilizados nas distribuições de conectividades da rede integrada de interações moleculares mostrados na Figura 2 foram:

Grau de entrada:

$$P(k)_{in} = 595.09k^{-1.24} \quad (11)$$

Grau de saída:

$$P(k)_{out} = 430.8k^{-1.15} \quad (12)$$

Na análise desse resultado, verificamos que a rede integrada da *E. coli* é livre de escala, o que indica que ela segue a lei de potência, assim como foi observado nas redes contendo somente um tipo de interação, ou seja, redes somente com interações proteína-proteína ou interações metabólicas ou interações de regulação transcricional [5].

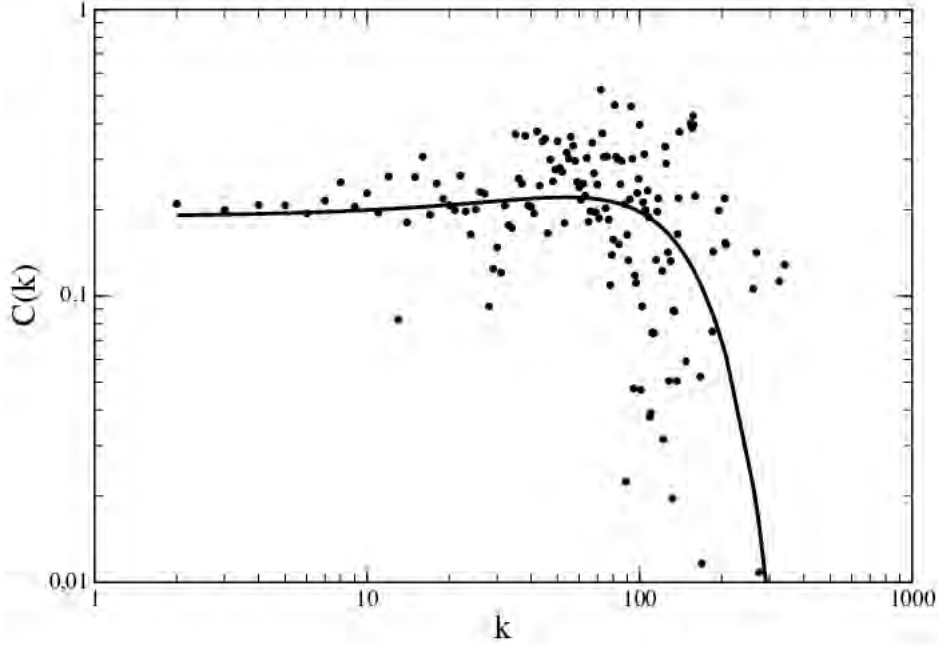


Figura 3: Dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k . Pode-se observar que o $C(k)$ é ajustado por um modelo linear quadrático, o que indica que a rede integrada de interações moleculares entre genes da *E. coli* construída neste projeto é linear quadrática.

O ajuste realizado na dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k dos dados da *E. coli* segue um modelo linear quadrático, descrito por:

$$y = \alpha e^{\beta x - \gamma x^2} \quad (13)$$

onde α , β , γ são constantes.

O ajuste utilizado na análise da dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k mostrado na Figura 3 foi:

$$C(k) = 0.19e^{0.0058k - 0.036k^2} \quad (14)$$

Esse modelo linear quadrático para $C(k)$ indica que a rede integrada da *E. coli* não segue o modelo clássico de classificação proposto por Ravasz et al. [16], onde nodos escassamente conectados fazem parte de áreas altamente conectadas com comunicação entre as diferentes áreas vizinhas altamente conectadas mantidas por alguns nodos altamente

conectados denominados *hubs* [5]. Nessa rede integrada, podemos observar que para baixos valores de k , existe uma fraca dependência em relação ao $C(k)$ e também que o $C(k)$ aumenta gradativamente a medida que o grau de conectividade aumenta até um valor de aproximadamente 100 conexões. Acima desse valor, os valores de $C(k)$ decaem abruptamente a medida que o grau de conectividade aumenta. Classificamos esse modelo como linear quadrático devido ao ajuste utilizado na dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k dos dados da *E. coli*.

4.2 Rede integrada da *S. cerevisiae*

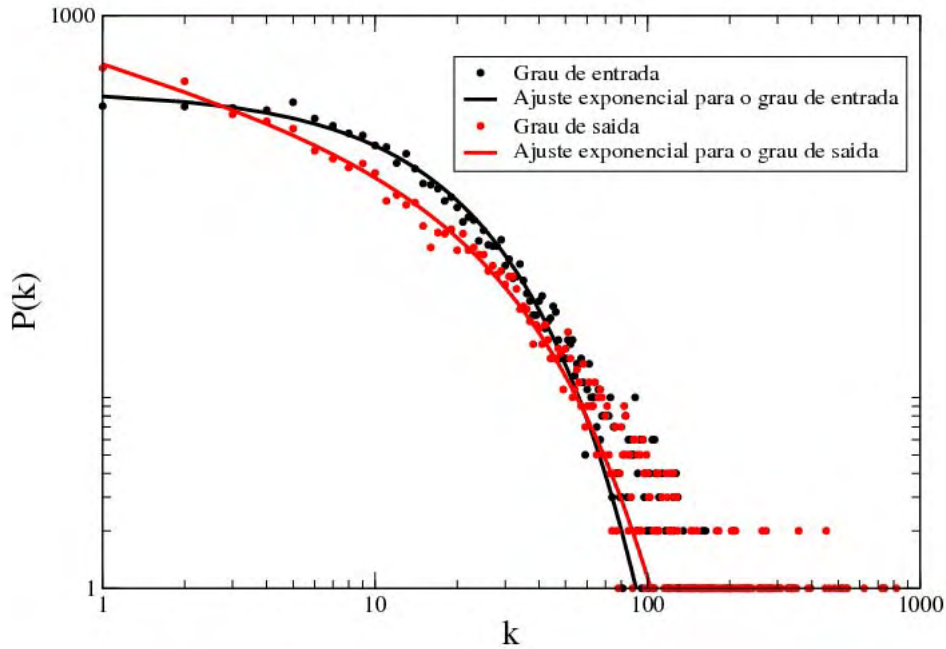


Figura 4: Distribuição de conectividades da rede integrada de interações moleculares entre genes da *S. cerevisiae* construída neste projeto. Essa distribuição é ajustada por um modelo livre de escala diferenciado.

Os ajustes realizados na distribuição de conectividades da rede de interações moleculares entre genes da *S. cerevisiae* seguem um modelo diferenciado representado por:

$$y = \alpha x^\beta e^{-\gamma x} \quad (15)$$

onde α e β são constantes.

Os ajustes utilizados na distribuição de conectividades da rede integrada de interações moleculares mostrado na Figura 4 foram:

Grau de entrada:

$$P(k)_{in} = 1141.67k^{0.037}e^{-0.57k} \quad (16)$$

Grau de saída:

$$P(k)_{out} = 582.089k^{-0.43}e^{-0.043k} \quad (17)$$

Na análise desse resultado, verificamos que a rede da *S. cerevisiae* possui um comportamento distinto do observado em outros grafos, pois a distribuição de conectividades é representada por uma exponencial mas com um pré fator lei de potência. Se esse ajuste fosse puramente exponencial, diríamos que o grafo seria compatível com o modelo de Erdős-Rényi [17], mas neste caso dizemos que o grafo é compatível com o modelo livre de escala, no entanto, com um *cutoff* exponencial.

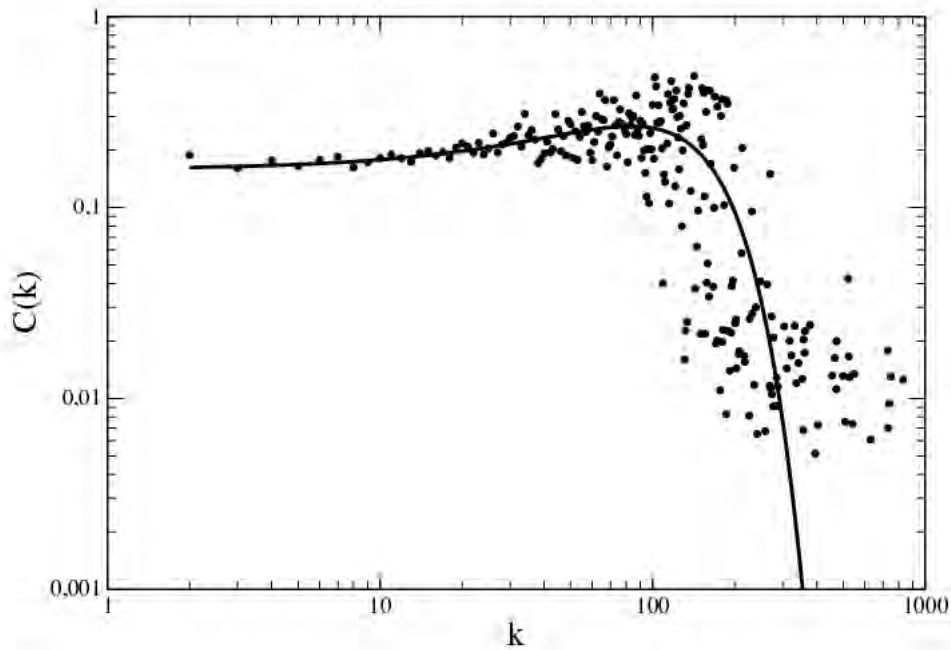


Figura 5: Dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k . Pode-se observar que o $C(k)$ é ajustado por um modelo linear quadrático, o que indica que a rede integrada de interações moleculares entre genes da *S. cerevisiae* construída neste projeto é hierárquica, mas com uma estrutura diferente das redes contendo apenas um tipo de interação.

O ajuste realizado na dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k dos dados da *S. cerevisiae* segue um modelo linear quadrático, descrito por:

$$y = \alpha e^{\beta x - \gamma x^2} \quad (18)$$

onde α , β , γ são constantes.

O ajuste utilizado na análise da dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k mostrado na Figura 5 foi:

$$C(k) = 0.16e^{0.012k - 0.051k^2} \quad (19)$$

Esse modelo linear quadrático para $C(k)$ indica que a rede integrada da *S. cerevisiae* não segue o modelo clássico de classificação proposto por Ravasz et al. [16], onde nodos escassamente conectados fazem parte de áreas altamente conectadas com comunicação entre as diferentes áreas vizinhas altamente conectadas mantidas por alguns nodos altamente conectados denominados *hubs* [5]. Nessa rede integrada, podemos observar que para baixos valores de k , existe uma dependência em relação ao $C(k)$ e também que o $C(k)$ aumenta gradativamente a medida que o grau de conectividade aumenta até um valor de aproximadamente 100 conexões. Acima desse valor, os valores de $C(k)$ decaem abruptamente a medida que o grau de conectividade aumenta. Classificamos esse modelo como linear quadrático devido ao ajuste utilizado na dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k dos dados da *S. cerevisiae*.

4.3 Análise comparativa do grau de proximidade e do grau de intermediação para as redes integradas da *S. cerevisiae* e da *E. coli*

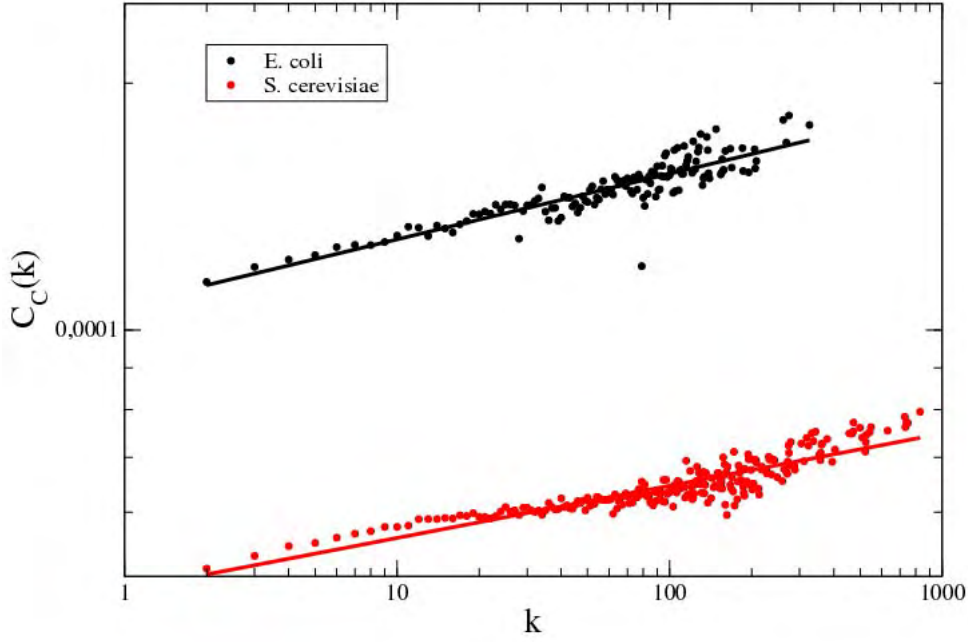


Figura 6: Grau de proximidade dos nodos em função de k para as redes integradas da *E. coli* e da *S. cerevisiae*. Ambos organismos tiveram suas medidas ajustadas por uma lei de potência.

Os ajustes realizados no grau de proximidade da rede integrada de interações moleculares entre genes da *E. coli* e da *S. cerevisiae* seguem uma lei de potência representada por:

$$y = \alpha x^\beta \quad (20)$$

onde α e β são constantes.

Os ajustes utilizados no grau de proximidade da rede integrada de interações moleculares mostrados na Figura 6 foram:

E. coli:

$$C_C(k) = 0.00011k^{0.079} \quad (21)$$

S. cerevisiae:

$$C_C(k) = 0.032k^{0.063} \quad (22)$$

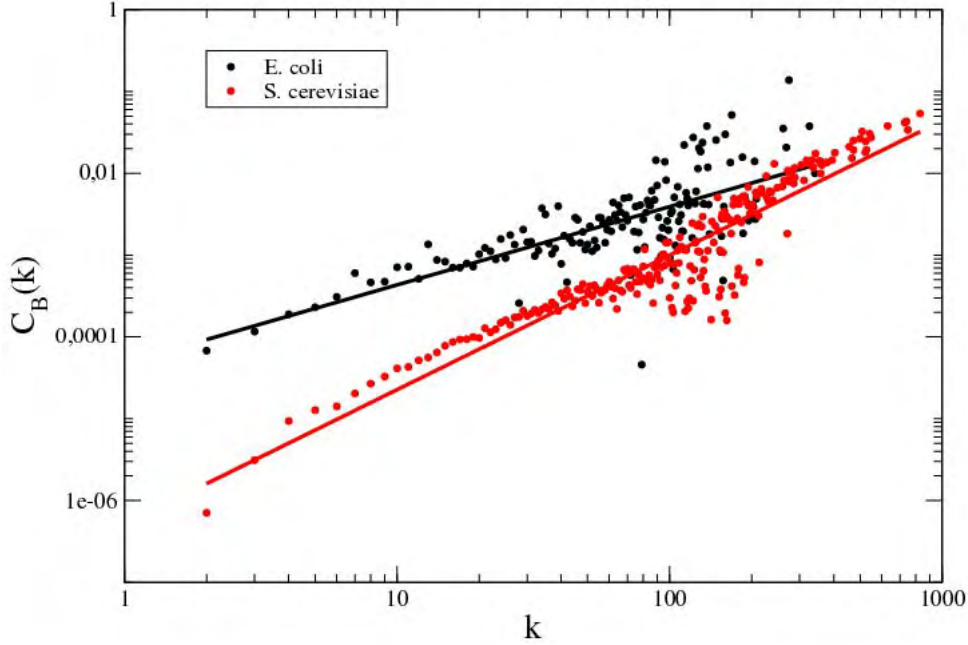


Figura 7: Grau de intermediação dos nodos em função de k para as redes integradas da *E. coli* e da *S. cerevisiae*. O ajuste utilizado para ambos organismos foi feito através de uma lei de potência.

Os ajustes realizados no grau de intermediação da rede integrada de interações moleculares entre genes da *E. coli* e da *S. cerevisiae* seguem uma lei de potência representada por:

$$y = \alpha x^\beta \quad (23)$$

onde α e β são constantes.

Os ajustes utilizados no grau de intermediação da rede de interações moleculares mostrados na Figura 7 foram:

E. coli:

$$C_B(k) = 0,032k^{0.95} \quad (24)$$

S. cerevisiae:

$$C_B(k) = 0,0047k^{1.64} \quad (25)$$

Em relação aos graus de proximidade e de intermediação dos nodos das redes integradas da *E. coli* e da *S. cerevisiae*, pode-se observar nas Figuras 6 e 7 que os valores dessas medições aumentam gradativamente a medida que aumenta k . O comportamento

referente ao grau de proximidade nessas redes é semelhante ao que ocorre em outras redes complexas, biológicas ou não, e significa que os menores caminhos entre os nodos são cada vez menores a medida que aumenta o grau de conectividade dos nodos nas extremidades desses caminhos. O comportamento do grau de intermediação nessas redes também é semelhante ao que ocorre em outras redes complexas e significa que quanto mais conexões um nodo possui, ele participa de mais menores caminhos entre outros nodos.

Em suma, podemos observar os resultados obtidos nas Tabelas 2 e 3 para cada organismo estudado.

Tabela 2: Resultados para a rede integrada da bactéria *E. coli*

Parâmetro	Equação	Coeficientes
$P(k)$	$y = \alpha x^{-\beta}$	$\beta_{in} = -0.89 \pm 0.15$; $\beta_{out} = -0.97 \pm 0.15$
$C(k)$	$y = \alpha e^{\beta x - \gamma x^2}$	$\beta = -0.0071 \pm 0.005$; $\gamma = -9.27 \times 10^{-6} \pm 3.45 \times 10^{-5}$
$C_C(k)$	$y = \alpha x^{\beta}$	$\beta = -0.081 \pm 0.007$
$C_B(k)$	$y = \alpha x^{\beta}$	$\beta = 1.6 \pm 0.4$

Segundo a análise dos parâmetros topológicos da Tabela 2, podemos inferir que a rede integrada da *E. coli* é classificada como livre de escala e segue o modelo linear quadrático.

Tabela 3: Resultados para a rede integrada da levedura *S. cerevisiae*

Parâmetro	Equação	Coeficientes
$P(k)$	$y = \alpha x^{-\beta} e^{\gamma x}$	$\beta_{in} = 0.037 \pm 0.01$; $\beta_{out} = 0.023 \pm 0.005$
$C(k)$	$y = \alpha e^{\beta x - \gamma x^2}$	$\beta = 0.017 \pm 0.005$; $\gamma = 1.5 \times 10^{-4} \pm 4 \times 10^{-5}$
$C_C(k)$	$y = \alpha x^{\beta}$	$\beta = -0.067 \pm 0.003$
$C_B(k)$	$y = \alpha x^{\beta}$	$\beta = 1.7 \pm 0.6$

Segundo a análise dos parâmetros topológicos da Tabela 3, podemos inferir que a rede integrada da *S. cerevisiae* é classificada como livre de escala, mas com um *cutoff* exponencial e segue o modelo linear quadrático.

5 Conclusão

Os resultados obtidos indicam que a rede biológica integrada da *E. coli* é livre de escala e a dependência do $C(k)$ nos valores de k indica um novo modelo de ajuste, o que nos conduziu a um método novo de classificação, definido como linear quadrático. Já para a rede integrada da *S. cerevisiae* os resultados indicam que, ela é classificada como livre de escala, porém, diferenciada, visto que o ajuste utilizado possui uma tendência para o modelo exponencial, devido ao *cutoff* que o ajuste apresenta. Além disso ela também é classificada segundo o novo método de ajuste que propusemos, como linear quadrático.

O novo método de classificação proposto para os ajustes da dependência do $C(k)$ nos valores de k vale apenas para a avaliação de redes biológicas integradas, ou seja, que possuem mais de um tipo de interação física. O mesmo ainda está em processo de avaliação para ser validado como um novo método de classificação e para isso deve-se comprovar sua validade para outros organismos. Como podemos observar nas Tabelas 2 e 3, os valores do intervalo de confiança são válidos para esses dados.

Neste trabalho medimos a relação entre a proximidade, inserção e conectividade de um nodo. Os resultados obtidos indicam que existe uma relação do tipo lei de potência entre estas quantidades. O significado biológico desta relação ainda não foi estudado, mas nos parece razoável supor que esta relação possa explicar por que os nodos com um maior número de conexões tendem a ser mais relevantes para um organismo [4], pois a inserção e a proximidade são propriedades que dependem de todo o grafo e não apenas da vizinhança de um nodo e portanto é natural assumir que elas estejam mais correlacionadas com a relevância de um dado gene.

A perspectiva desse trabalho é continuar com as análises das propriedades topológicas estudadas para outros organismos, como por exemplo *Homo sapiens* e *Mus musculus*. Esse estudo objetiva propor um novo método de classificação para suas redes biológicas integradas, sempre buscando descobrir a essencialidade de novos genes importantes para o organismo em questão.

Referências

- [1] M. H. V. V. Regenmortel, EMBO Rep **5**, 1016 (2004).
- [2] A. C. Ahn, M. Tewari, C.-S. Poon, and R. S. Phillips, PLoS Med **3**, e208 (2006).
- [3] B. Bollobás, *Graph Theory: an introductory course*, 1st ed. (Springer Verlag, New York, 1979), Vol. 3.
- [4] R. Albert and A.-L. Barabási, Reviews of Modern Physics **74**, (2002).
- [5] A.-L. Barabási and Z. N. Oltvai, Nat Rev Genet **5**, 101 (2004).
- [6] A.-L. Barabási, *Linked: The New Science of Networks*, 1st ed. (Basic Books, Cambridge, 2002).
- [7] E. Carvalho and T. Mendonça, *Ensaio de complexidade 2*, 2004.
- [8] B. Perroud, J. Lee, N. Valkova, A. Dhirapong, P.-Y. Lin, O. Fiehn, D. Kültz, and R. H. Weiss, Mol Cancer **5**, 64 (2006).
- [9] A. Ertel, A. Verghese, S. W. Byers, M. Ochs, and A. Tozeren, Mol Cancer **5**, 55 (2006).
- [10] D. R. Rhodes, S. A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyanasundaram, D. Ghosh, A. Pandey, and A. M. Chinnaiyan, Nat Biotechnol **23**, 951 (2005).
- [11] J. P. M. da Silva, M. L. Acencio, J. C. M. Mombach, R. Vieira, J. C. da Silva, N. Lemke, and M. Sinigaglia, *In silico network topology-based prediction of gene essentiality*, 2008.
- [12] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).
- [13] M. E. Newman and D. J. Watts, Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics **60**, 7332 (1999).
- [14] M. E. Newman, S. H. Strogatz, and D. J. Watts, Phys Rev E Stat Nonlin Soft Matter Phys **64**, 026118 (2001).
- [15] Disponível em: <<http://www.esyurl.com/574>>.
- [16] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, Science **297**, 1551 (2002).
- [17] P. Erdős and A. Rényi, *On Random Graphs*, 1st ed., 1959.