



Guilherme Ribeiro Germek

Caracterização topológica dos motivos
das Redes Biológicas Integradas da
Escherichia coli e da *Saccharomyces
cerevisiae*

Botucatu
2008

Guilherme Ribeiro Germek

Caracterização topológica dos motivos
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

Germek, Guilherme Ribeiro.

Caracterização topológica dos motivos das redes biológicas integradas da
Escherichia coli e da *Saccharomyces cerevisiae* / Guilherme Ribeiro Germek.
- Botucatu [s.n], 2008.

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

1. Bioinformática 2. Biologia molecular

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

Agradecimentos

Agradeço primeiramente aos meus pais e a minha vó pelo apoio, carinho e atenção que sempre me deram.

Agradeço a minha família por estarem sempre comigo.

Agradeço ao meu orientador Ney Lemke por acreditar em mim, mesmo quando as outras pessoas não acreditavam que eu chegaria onde cheguei.

Agradeço aos meus colegas do LBBC (Laboratório de Bioinformática e Biofísica Computacional): Tiago F. Andrade, Marcio L. Acencio e Luiz A. Bovolenta, que sempre me ajudaram durante todo esse tempo de estágio.

Agradeço aos meus amigos da faculdade, que estiveram comigo durante todo este tempo de curso, principalmente as pessoas que moraram comigo durante este período.

E por fim, mas não menos importante agradeço a CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pelo apoio financeiro que me deu durante minhas pesquisas.

*“Um sonho se transforma em objetivo a partir do momento
em que você estabelece uma data para realizá-lo.”*

Resumo

O método reducionista 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 reducionismo. Para suplantarmos esses limites, pesquisadores têm usado um conjunto de métodos conhecido como biologia sistêmica, nova área da biologia cujo objetivo é a compreensão das interações 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 arcos que conectam pares de nodos. Dentro desses grafos é possível encontrar subgrafos, ocorrendo em redes complexas em uma quantidade significativamente maior que a quantidade em redes aleatórias, eles são definidos como motivos. Como motivos em redes biológicas podem representar as unidades estruturais do processo biológico, suas detecções são importantes. Portanto, o objetivo deste presente trabalho foi detectar, contar e classificar motivos presentes em redes biológicas integradas da bactéria *Escherichia coli* e da levedura *Saccharomyces cerevisiae*. Para tanto, implementamos códigos em ambiente *Mathematica*[®] e Python para detectar, contar e classificar os motivos dessas redes. A composição e os tipos de motivos detectados nessas redes integradas indicam que tais redes são organizadas em três módulos pontes principais compostos por motivos em que todas as arestas são do mesmo tipo. A ponte de conexão é composta pelos motivos nos quais os tipos de arestas são diferentes.

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

Abstract

The reductionism method has helped in the clarification of functioning of many biological process. However, such process are extremely complex and have emergent properties that can not be explained or even predicted by reductionism methods. To overcome these limits, researchers have been used a set of methods known as systems biology, a new area of biology aiming to understand the interactions between the multiple components of biological processes. These interactions can be represented by a mathematical object called graph or network, where the interacting elements are represented by a vertex and the interactions by edges that connect a pair of vertexes. Into graphs it is possible to find subgraphs, occurring in complex networks at numbers that are significantly higher than those in randomized networks, they are defined as motifs. As motifs in biological networks may represent the structural units of biological processes, their detection is important. Therefore, the aim of this present work was detect, count and classify motifs present in biological integrated networks of bacteria *Escherichia coli* and yeast *Saccharomyces cerevisiae*. For this purpose, we implemented codes in *Mathematica*[®] and Python environments for detecting, counting and classifying motifs in these networks. The composition and types of motifs detected in these integrated networks indicate that such networks are organized in three main bridged modules composed by motifs in which edges are all the same type. The connecting bridges are composed by motifs in which the types of edges are different.

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

Sumário

	Página
Resumo	4
Abstract	5
1 Introdução	7
2 Objetivos	8
3 Materiais e Métodos	9
3.1 Distribuição de Conectividades	9
3.2 Coeficiente de Clusterização	9
3.3 Banco de Dados	9
3.4 Algoritmos	10
4 Resultados e Discussão	12
5 Conclusão	26

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]. Para suplantar 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 [1].

Tais interações são geralmente representadas por um objeto matemático chamado grafo [2], que pode ser definido como um conjunto de nodos e um conjunto de arestas, que conectam pares de nodos. No caso das redes biológicas integradas, os nodos representam os componentes e as arestas representam as conexões (interações) entre os nodos. As arestas de um grafo podem ser direcionadas, indicando uma fonte (ponto de partida) e um alvo (ponto terminal) ou não direcionadas. O número total de conexões de um nodo é chamado de grau de conectividade k [3]. Os nodos que estão diretamente conectados podem ser chamados de vizinhos.

Os grafos possuem algumas propriedades, dentre elas podemos citar a distribuição de conectividades $P(k)$ e o coeficiente de clusterização $C(k)$. Ambos serão definidos na seção Materiais e Métodos.

Um subgrafo é definido por um conjunto de nodos e um conjunto de arestas de um grafo. Os exemplos mais simples de subgrafos são ciclos, triângulos (subgrafo de ordem 3), quadrados (subgrafo de ordem 4) e assim por diante.

Motivos são subgrafos que ocorrem significativamente mais vezes na rede real do que em redes aleatorizadas [4], que preservam as características dos nodos da rede real, ou seja, o número de ligações dos nodos deve permanecer o mesmo. A detecção de motivos em redes biológicas é importante. Por exemplo, foram encontrados três principais motivos nas redes de regulação transcricional da bactéria *Escherichia coli* e da levedura *Saccharomyces cerevisiae* ([4], [5]). Dentre eles temos o:

SIM (*single-input module*), que foi demonstrado teoricamente [5] e experimentalmente ([6], [7], [8]) por gerar programas temporais de expressão, tal ordem temporal pode ser útil em processos que requerem vários estágios para se completarem.

FFL (*feedforward loop*), ocorre quando um sinal externo causa uma resposta rápida em várias sistemas, tal como a repressão do sistema de utilização de açúcar em reação a glicose [5].

Esse tipo de detecção não foi feita em redes integradas, ou seja, as detecções dos motivos, até o presente momento foram feitas em redes que somente possuem apenas um tipo de interação. Portanto, nós propomos neste trabalho a busca de motivos nas redes biológica integrada da *E. coli* e *S. cerevisiae*, sem levar em consideração as direções das ligações entre os nodos.

2 Objetivos

1. Propor e implantar um algoritmo de detecção de motivos em redes integradas
2. Determinar os motivos de ordem 3 na rede integrada da *E. coli* e da *S. cerevisiae*

3 Materiais e Métodos

Nesta seção serão descritas algumas propriedades dos grafos, as quais foram utilizadas no estudo, os bancos de dados utilizados para a formação das redes biológicas integradas e os algoritmos criados e utilizados para o estudo dos motivos.

3.1 Distribuição de Conectividades

O grau de um nodo é definido como k , e representa quantas ligações este nodo faz com outros nodos. A distribuição de conectividades $P(k)$ é uma função que descreve o número total de nodos de um grafo que possuem um mesmo grau. Formalmente, a distribuição de conectividade é:

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

onde v é um nodo que pertence aos nodos do grafo V , e $\deg(v)$ é o grau do nodo v .

3.2 Coeficiente de Clusterização

O coeficiente de clusterização caracteriza a "densidade" das conexões em torno de um nodo. Suponha que uma rede é indireta, e um de seus nodos tem z vizinhos. A máxima ligação neste grupo é aproximada quando todas as $z(z + 1)/2$ possíveis ligações no grupo estão presentes. Isto também significa que todas as $z(z - 1)/2$ possíveis ligações entre os vizinhos mais próximos existem. Convencionalmente ([9], [10], [11]), o coeficiente de clusterização C de um nodo é taxa entre o número total y de ligações conectando seus vizinhos mais próximos e o número total de todas as possíveis ligações entre todos esses vizinhos mais próximos [3].

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

3.3 Banco de Dados

Os dados para a construção das redes biológicas integradas da *S. cerevisiae* e da *E. coli* foram obtidos dos bancos de dados mostrados respectivamente na Tabela 1 e na Tabela 2:

Tabela 1: Lista dos bancos de dados biológicos utilizados para o processo de aquisição dos dados da *S. cerevisiae*.

Banco de dados	Tipo de interação	Referência
BIGG	Metabólica	http://bigg.ucsd.edu/home.pl
YEASTRACT	Regulação transcricional	http://www.yeasttract.com
BIOGRID	Proteína-proteína	http://www.thebiogrid.org

Tabela 2: Lista dos bancos de dados biológicos utilizados para o processo de aquisição dos dados da *E. coli*.

Banco de dados	Tipo de interação	Referência
KEGG	Metabólica	http://www.genome.jp/kegg
RegulonDB	Regulação transcricional	http://regulondb.ccg.unam.mx
Bacteriome.org	Proteína-proteína	http://www.bacteriome.org

Nas redes biológicas integradas [12], os genes representam os nodos, e considerando que os genes *g1* e *g2* codificam as proteínas *p1* e *p2*, estes são conectados na rede, de acordo com:

Interação proteína-proteína: *p1* e *p2* interagem fisicamente;

Regulação: *g1* regula a transcrição do gene *g2* ou

Metabolismo: um produto gerado pela reação catalisada pela proteína *p1* é consumido na reação catalisada pela proteína *p2* (nesta análise foram excluídos os compostos mais usados, tais como, ATP, ADP, NAD, etc.)

3.4 Algoritmos

Através de ferramentas estatísticas e computacionais foram desenvolvidos algoritmos que, a partir de uma dada rede biológica integrada, geram um conjunto de redes aleatórias respeitando as propriedades estatísticas da rede fornecida e, depois, fazem a procura dos motivos. Para tanto, utilizamos o pacote de funções NetworkX (<https://networkx.lanl.gov/>) e, em seguida o programa *Mathematica*[®], para implementar o algoritmo de contagem dos motivos.

Os algoritmos que fazem a procura pelos motivos são devididos em três partes: criação das redes aleatórias, procura pelos triângulos e, por último, a classificação das ligações desses triângulos.

A primeira parte do algoritmo que cria as redes aleatórias possui dois métodos:

Método I: utiliza a lista de entrada com as interações e seus respectivos tipos, armazena os primeiros nodos da lista, depois os tipos de interações e, em seguida, os segundos nodos da lista. Com estes conjuntos de listas, é feita uma escolha aleatória de um termo em cada lista; estas escolhas tornam-se os novos nodos com seus novos tipos de ligações. Assim podemos garantir que os nodos mantenham os seus graus de conectividades. Para observar se os graus de conectividade não mudaram, foi feito o gráfico de $P(k)$ em função de k .

Método II: utiliza a lista de entrada com as interações e seus respectivos tipos, armazena os tipos de interações em uma lista. Em seguida, é feita uma escolha aleatória de um termo da lista do tipo de interações, mantendo-se os nodos das antigas interações. Com isso apenas as interações são aleatorizadas.

Após a criação das redes aleatórias, faz-se a procura pelos nodos que formam um triângulo utilizando uma função do NetworkX chamada *neighbors*. Com isso, conseguimos gerar um arquivo de saída com os triângulos existentes e seus respectivos nodos.

Utilizando uma função do NetworkX chamada *get_edge*, fizemos a procura do tipo de ligação entre cada nodo dos triângulos e ficamos com um arquivo de saída que contém todos os triângulos com suas três ligações caracterizadas.

Finalmente, foi feito um gráfico de $C(k)$ em função de k para comparar as redes reais com as redes aleatórias e também foi feita a contagem dos motivos encontrados.

Todos os algoritmos podem ser encontrados no *website* <http://www.esyurl.com/574>.

(a) tri_Rede2_rede.py

Algoritmo que gera as redes aleatórias(Método I) e encontra os triângulos e seus respectivos tipos de interações.

(b) tri_Rede2_rede3.py

Algoritmo que gera as redes aleatórias(Método II) e encontra os triângulos e seus respectivos tipos de interações.

(c) tri_cerevisiae_rede.nb

Algoritmo que utiliza a função *Count{}* e mede a porcentagem de cada tipo de triângulo nas redes aleatórias da *S. cerevisiae*.

(d) tri_ecoli_rede.nb

Algoritmo que utiliza a função *Count{}* e mede a porcentagem de cada tipo de triângulo nas redes aleatórias da *E. coli*.

(e) cdk.py

Algoritmo que mede $C(k)$. Para tanto faz-se o uso de duas funções: *clustering()* e *degree()*, do pacote NetworkX.

(f) pdk.py

Algoritmo que mede $P(k)$. Para tanto faz-se o uso de uma função chamada *degree_histogram()* do pacote NetworkX.

4 Resultados e Discussão

As ligações foram classificadas de acordo com a Tabela 3:

Tabela 3: Legenda das interações

Letra	Tipo de interação
p	proteína-proteína
m	metabólica
r	regulatória
pm	proteína-proteína e metabólica
pr	proteína-proteína e regulatória

Antes da detecção dos motivos com 3 nodos nas redes integradas da *E. coli* e da *S. cerevisiae*, nós geramos, primeiramente, dois grupos de 100 redes aleatórias: um grupo no qual somente os tipos de ligação entre os nodos foram aleatorizados e outro grupo no qual tanto o tipo de ligação quanto os pares de nodos ligados foram aleatorizados. Nos dois casos, os graus de conectividade dos nodos nas redes reais foram preservados nas redes aleatórias. De acordo com as Figuras 1 e 2, o algoritmo utilizado para a geração dos dois grupos de 100 redes aleatórias foi bem sucedido, já que tanto para a *E. coli* (Figura 1) quanto para a *S. cerevisiae* (Figura 2), os ajustes dos pontos referentes aos valores de $P(k)$ em função de k das redes aleatórias mostraram-se semelhantes aos ajustes das respectivas redes reais, indicando que as distribuições de conectividades foram preservadas.

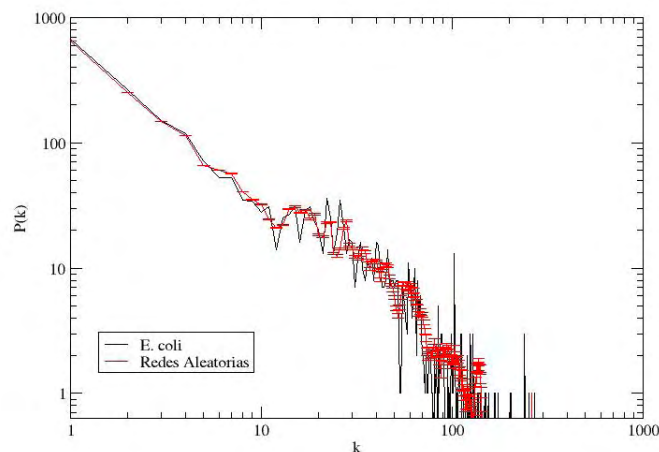


Figura 1: Gráfico de $P(k)$ em relação k da *E. coli* e de 100 redes aleatórias(Método I) da mesma. Neste gráfico é possível observar que o $P(k)$ da *E. coli* e das Redes Aleatórias seguem juntos indicando, portanto, que o grau de conectividade k dos nodos foram mantidos

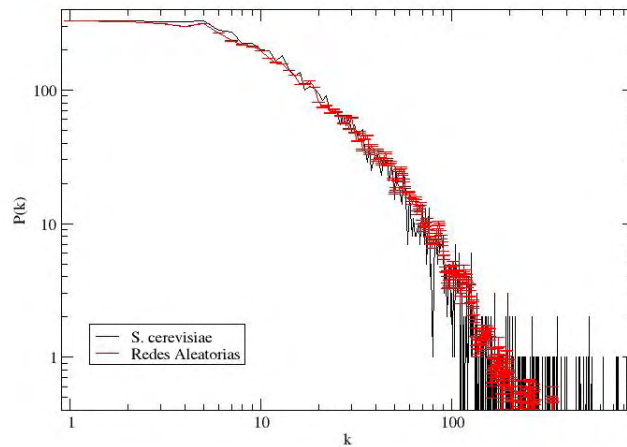


Figura 2: Gráfico de $P(k)$ em relação a k da *S. cerevisiae* e de 100 Redes Aleatórias(Método I) da mesma. Neste gráfico é possível observar que o $P(k)$ da *S. cerevisiae* e das redes aleatórias seguem juntos indicando, portanto, que o grau de conectividade k dos nodos foram mantidos

As aleatorizações dos tipos de ligações e dos pares de nodos ligados das redes integradas da *E. coli* e da *S. cerevisiae* afetaram os valores de $C(k)$ dos nodos e a dependência desses valores em relação aos seus respectivos valores de k . Como pode ser observado na Figura 3 e na Figura 4, os valores de $C(k)$ desse grupo de redes aleatórias variaram pouco a medida que os valores de k aumentaram, comportamento distinto ao das redes reais, onde os valores de $C(k)$ mantêm-se com pouca variação até próximo de $k=100$ e caem abruptamente após esse valor. Como o coeficiente de agrupamento representa, de fato, o número de triângulos aos quais um determinado nodo pertence, então os valores de $C(k)$ das redes aleatórias indicam que, quando o tipo de ligação e os pares de nodos conectados foram aleatorizados, o número de triângulos diminuiu, mesmo com a manutenção da distribuição de conectividades. De fato, enquanto o número de triângulos encontrados na rede real da *E. coli* foi de 72572, foram contados uma média de 42541,7 triângulos nas suas respectivas redes aleatórias onde o tipo de ligação e os pares de nodos conectados foram aleatorizados (Tabelas 4 e 8). O mesmo pode ser observado para *S. Cerevisiae* (Tabelas 10 e 14).

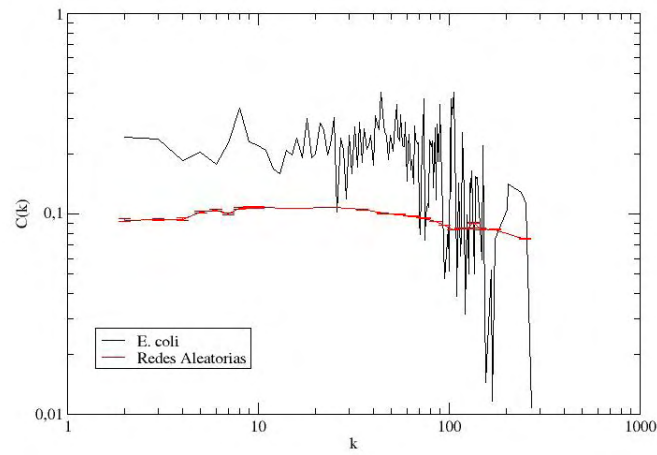


Figura 3: Grafico de $C(k)$ em relação a k da *E. coli* e de 100 Redes Aleatórias(Método I) da mesma. Observa-se que a dependência de $C(k)$ em relação ao grau de conectividade k da *E. coli* fica reduzido após sua rede ser aleatorizada

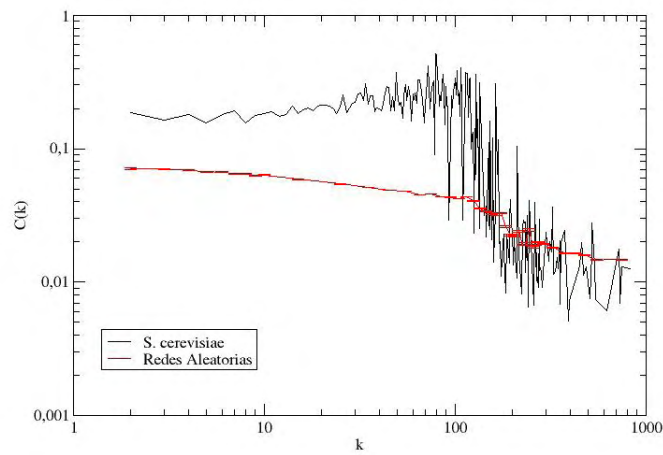


Figura 4: Grafico de $C(k)$ em relação a k da *S. cerevisiae* e de 100 Redes Aleatórias(Método I) da mesma. Pode-se perceber que a dependência do $C(k)$ em relação ao número de conexões k da *S. cerevisiae* fica reduzido após sua rede ser aleatorizada

Tabela 4: Triângulos encontrados na rede biológica integrada da *E. coli*

Tipos de ligações	Porcentagem(%)
p;p;p	81,80
m;m;m	12,47
r;r;r	0,54
p;p;m	1,26
p;p;r	0,0068
p;m;m	0,97
m;m;r	0,15
p;r;r	0,87
m;r;r	0,096
p;m;r	0,0082
p;p;pm	1,36
m;m;pm	0,41
r;r;pm	0,0096
p;m;pm	0,019
p;r;pm	0
m;r;pm	0,0013
Total	72572 triângulos

Tabela 5: Porcentagem dos tipos de interações da *E. coli*

Interação	Porcentagem(%)
um tipo de ligação	94,81
uma ligação de um tipo e duas do outro	3,35
três tipos de ligações	0,0082
ligações do tipo pm	1,8



Figura 5: Porcentagem de triângulos encontrados na rede biológica integrada da *E. coli*

Tabela 6: Triângulos encontrados na média de 100 Redes Aleatórias(Método II) da *E. coli*

Tipos de ligações	Porcentagem(%)	Desvio($\pm$)
p;p;p	29,27	0,73
m;m;m	1,39	0,097
r;r;r	0,077	0,013
p;p;m	31,85	0,38
p;p;r	12,18	0,39
p;m;m	11,54	0,42
m;m;r	1,60	0,095
p;r;r	1,68	0,10
m;r;r	0,61	0,046
p;m;r	8,82	0,25
p;p;pm	0,41	0,076
m;m;pm	0,054	0,013
r;r;pm	0,0079	0,0034
p;m;pm	0,29	0,057
p;r;pm	0,11	0,025
m;r;pm	0,042	0,010
Total	72572 triângulos	

Tabela 7: Porcentagem dos tipos de interações encontrados na média de 100 Redes Aleatórias(Método II) da *E. coli*

Interação	Porcentagem(%)
um tipo de ligação	30,73
uma ligação de um tipo e duas do outro	59,46
três tipos de ligações	8,82
ligações do tipo pm	0,92



Figura 6: Porcentagem de triângulos encontrados na média de 100 Redes Aleatórias(Método II) da *E. coli*

Tabela 8: Triângulos encontrados na média de 100 Redes Aleatórias(Método I) da *E. coli*

Tipos de ligações	Porcentagem(%)	Desvio($\pm$)
p;p;p	29,30	0,51
m;m;m	1,38	0,085
r;r;r	0,077	0,013
p;p;m	31,89	0,36
p;p;r	12,20	0,33
p;m;m	11,51	0,30
m;m;r	1,58	0,081
p;r;r	1,68	0,11
m;r;r	0,60	0,048
p;m;r	8,81	0,22
p;p;pm	0,41	0,074
m;m;pm	0,054	0,015
r;r;pm	0,0073	0,0047
p;m;pm	0,29	0,054
p;r;pm	0,11	0,024
m;r;pm	0,041	0,012
Total	42541,7 triângulos	

Tabela 9: Porcentagem dos tipos de interações encontrados na média de 100 Redes Aleatórias(Método I) da *E. coli*

Interação	Porcentagem(%)
um tipo de ligação	30,76
uma ligação de um tipo e duas do outro	59,46
três tipos de ligações	8,81
ligações do tipo pm	0,91



Figura 7: Percentagem de triângulos encontrados na média de 100 Redes Aleatórias(Método I) da *E. coli*

Tabela 10: Triângulos encontrados na *S. cerevisiae*

Tipos de ligações	Porcentagem(%)
p;p;p	49,63
m;m;m	23,087
r;r;r	8,93
p;p;m	0,33
p;p;r	1,14
p;m;m	1,10
m;m;r	0
p;r;r	11,33
m;r;r	3,80
p;m;r	0,022
p;p;pm	0,011
m;m;pm	0,12
r;r;pm	0,020
p;m;pm	0,0066
p;r;pm	0
m;r;pm	0
p;p;pr	0,028
m;m;pr	0
r;r;pr	0,34
p;m;pr	0
p;r;pr	0,051
m;r;pr	0,0096
Total	226932 triângulos

Tabela 11: Porcentagem dos tipos de interações da *S. cerevisiae*

Interação	Porcentagem(%)
um tipo de ligação	81,65
uma ligação de um tipo e duas do outro	17,7
três tipos de ligações	0,022
ligações do tipo pm ou pr	0,59

Figura 8: Porcentagem de triângulos encontrados na *S. cerevisiae*

Tabela 12: Triângulos encontrados na média de 100 Redes Aleatórias(Método II) da *S. cerevisiae*

Tipos de ligações	Porcentagem(%)	Desvio(±)
p;p;p	14,46	0,20
m;m;m	0,25	0,013
r;r;r	3,86	0,094
p;p;m	11,26	0,16
p;p;r	27,92	0,20
p;m;m	2,92	0,087
m;m;r	1,87	0,055
p;r;r	17,96	0,22
m;r;r	4,66	0,078
p;m;r	14,48	0,14
p;p;pm	0,028	0,0089
m;m;pm	0,0018	0,0011
r;r;pm	0,011	0,0040
p;m;pm	0,014	0,0050
p;r;pm	0,037	0,010
m;r;pm	0,0094	0,0037
p;p;pr	0,056	0,012
m;m;pr	0,0039	0,0017
r;r;pr	0,024	0,0058
p;m;pr	0,029	0,0069
p;r;pr	0,072	0,015
m;r;pr	0,018	0,0053
Total	226932 triângulos	

Tabela 13: Porcentagem dos tipos de interações encontrados na média de 100 Redes Aleatórias(Método II) da *S. cerevisiae*

Interação	Porcentagem(%)
um tipo de ligação	18,57
uma ligação de um tipo e duas do outro	66,59
três tipos de ligações	14,48
ligações do tipo pm ou pr	0,30



Figura 9: Porcentagem de triângulos encontrados na média de 100 Redes Aleatórias(Método II) da *S. cerevisiae*

Tabela 14: Triângulos encontrados na média de 100 Redes Aleatórias(Método I) da *S. cerevisiae*

Tipos de ligações	Porcentagem(%)	Desvio($\pm$)
p;p;p	14,44	0,24
m;m;m	0,25	0,023
r;r;r	3,85	0,11
p;p;m	11,25	0,17
p;p;r	27,94	0,24
p;m;m	2,94	0,10
m;m;r	1,88	0,067
p;r;r	17,96	0,22
m;r;r	4,67	0,10
p;m;r	14,46	0,18
p;p;pm	0,028	0,0099
m;m;pm	0,0021	0,0018
r;r;pm	0,012	0,0041
p;m;pm	0,014	0,0058
p;r;pm	0,036	0,013
m;r;pm	0,0097	0,0044
p;p;pr	0,057	0,014
m;m;pr	0,0035	0,0019
r;r;pr	0,024	0,0069
p;m;pr	0,029	0,0082
p;r;pr	0,071	0,016
m;r;pr	0,019	0,0062
Total	79476,5 triângulos	

Tabela 15: Porcentagem dos tipos de interações encontrados na média de 100 Redes Aleatórias(Método I) da *S. cerevisiae*

Interação	Porcentagem(%)
um tipo de ligação	18,54
uma ligação de um tipo e duas do outro	66,64
três tipos de ligações	14,46
ligações do tipo pm ou pr	0,30



Figura 10: Porcentagem de triângulos encontrados na média de 100 Redes Aleatórias(Método I) da *S. cerevisiae*

Como, por definição, motivos são subgrafos que ocorrem significativamente mais vezes na rede real do que nas respectivas redes aleatórias, realizamos a detecção dos triângulos tanto nas redes reais da *E. coli* e da *S. cerevisiae* quanto nas respectivas redes aleatórias. Os dois grupos de redes aleatórias puderam ser utilizadas nessa detecção por que, embora os tipos de aleatorizações sejam diferentes, ambas tiveram praticamente a mesma distribuição de triângulos (Tabelas: 6 e 8; 12 e 14).

Como pode ser observado nas Tabelas 4 e 7, foram detectados os seguintes três motivos em ambos os organismos: o motivo p-p-p, que contém somente interações físicas entre proteínas; o motivo r-r-r que contém somente interações de regulação transcricional e o motivo m-m-m que contém somente interações metabólicas. Esses motivos contendo somente um tipo de interação perfazem cerca de 93% dos triângulos na *E. coli* e cerca de 90% dos triângulos na *S. cerevisiae* indicando que, embora essas redes sejam integradas, elas parecem ser altamente modularizadas, ou seja, pares de nodos ligados através de um mesmo tipo de interação tendem a formar módulos dentro das redes integradas. Os triângulos que possuem mais de um tipo diferente de interação correspondem a cerca de 10% dos triângulos e poderiam, por sua vez, representar "pontes" de ligação entre os módulos mencionados acima.

Na rede da *E. coli* foram detectados alguns motivos - p-p-pm e m-m-pm - que poderiam fazer parte dessas pontes de ligação entre os módulos contendo nodos ligados somente

através de um tipo de ligação, mais especificamente, entre os módulos contendo os motivos p-p-p e os motivos m-m-m. Embora virtualmente todos os genes de um organismo sejam regulados transcricionalmente por pelo menos um outro gene, não foram detectados motivos que evidenciassem esse fato na *E. coli*. Porém, há triângulos na rede biológica integrada da *E. coli* contendo, ao mesmo tempo, interações de regulação transcricional e interações proteína-proteína ou metabólicas (Tabela 4), mas tais triângulos apresentaram-se em menor número na rede real do que nas redes aleatórias. Uma possível explicação para tal fato é que o algoritmo utilizado para a detecção dos motivos poderia deixar de considerar alguns triângulos funcionalmente importantes a favor da significância estatística.

Na rede da *S. cerevisiae* foram detectados alguns motivos - m-m-pm, r-r-pm e r-r-pr - que poderiam fazer parte dessas pontes de ligação entre os módulos contendo nodos ligados somente através de um tipo de ligação. Diferente da *E. coli*, foram detectados motivos na *S. cerevisiae* contendo, ao mesmo tempo, interações de regulação transcricional e interações proteína-proteína ou metabólicas. Esses motivos evidenciam o fato mencionado acima de que, virtualmente, todos os genes de um organismo são regulados transcricionalmente.

5 Conclusão

A composição e os tipos de motivos detectados nas redes integradas da levedura *Saccharomyces cerevisiae* e da bactéria *Escherichia coli* indicam que tais redes são organizadas em três módulos principais que são interligados entre si. Os módulos principais são constituídos pelos motivos contendo somente um tipo de interação e as interligações entre esses módulos contém os motivos que possuem mais de um tipo diferente de interação.

O processo de aleatorização diminuiu sensivelmente o número de triângulos, isto evidencia que existe pressão seletiva para que as redes integradas seja modularizadas.

O alto grau de modularização possui um efeito no comportamento de $C(k)$. Pois de fato temos três módulos heterogêneos e um único grafo, ligados fracamente por pontes não homogêneas. Este comportamento deve ser investigado com mais profundamente em estudos posteriores.

Este foi um estudo inicial sobre motivos em redes biológicas integradas. Obtivemos alguns resultados, os quais devem ser estudados mais profundamente, e deve-se realizar este estudo em mais organismos além da levedura *Saccharomyces cerevisiae* e da bactéria *Escherichia coli*, para facilitar o entendimento dos resultados, além de se poder encontrar algum padrão de comportamento. Também deve-se refazer o estudo caso haja alguma alteração nos dados de interações, como por exemplo alguma atualização descoberta. Outro fato importante foi que neste estudo as interações foram tomadas como indiretas e os motivos encontrados eram de triângulos que tinham apenas seus nodos conectados entre si, portanto nem todos os motivos foram computados, deixamos aqui proposto que o estudo seja novamente realizado, mas desta vez levando em conta as direções das interações entre os nodos e analisando também os outros motivos não trabalhados neste estudo das redes biológicas integradas.

Parte dos resultados aqui apresentados foram divulgados no XXXI Encontro Nacional de Física da Matéria Condensada em Águas de Lindóia, em maio de 2008. No projeto: Motivos topológicos na rede integrada da bactéria *E. coli* e da levedura *S. cerevisiae*.

Sobre os estudos dos grafos de redes biológicas integradas, também divulgamos três trabalhos: 7th *International Symposium on Mathematical and Computational Biology*, 2007, Búzios, intitulado: *Network topology-based in silico discovery of gene essentiality in Saccharomyces cerevisiae*; Third International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2007, São Paulo, intitulado *Towards the In Silico Construction of the Anchorage-Independent Cell Cycle Start Network* e VII Workshop da Pós-Graduação, 2008, Botucatu, 2008, intitulado: Estudos comparativos das propriedades das redes integradas da *E. coli* e da *S. cerevisiae*.

Referências

- [1] M. H. V. V. Regenmortel, EMBO Rep **5**, 1016 (2004).
- [2] B. Bollobás, *Graph Theory: an introductory course*, 1st ed. (Springer Verlag, New York, 1979), Vol. 3.
- [3] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the internet and WWW*, 1st ed. (Oxford University Press, USA, 2003), Vol. 1.
- [4] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).
- [5] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, Nat Genet **31**, 64 (2002).
- [6] S. Kalir, J. McClure, K. Pabbaraju, C. Southward, M. Ronen, S. Leibler, M. G. Surette, and U. Alon, Science **292**, 2080 (2001).
- [7] M. Ronen, R. Rosenberg, B. I. Shraiman, and U. Alon, Proc Natl Acad Sci U S A **99**, 10555 (2002).
- [8] A. Zaslaver, A. E. Mayo, R. Rosenberg, P. Bashkin, H. Sberro, M. Tsalyuk, M. G. Surette, and U. Alon, Nat Genet **36**, 486 (2004).
- [9] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).
- [10] D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*, 1st ed. (Princeton University Press, Princeton, 1999), Vol. 1.
- [11] S. H. Strogatz, Nature **410**, 268 (2001).
- [12] 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.