UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

Milena Giglioli

Determinação de comunidades em redes biológicas integradas.

Botucatu
2009

Milena Giglioli

Determinação de comunidades em redes biológicas integradas.

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

2009

Agradecimentos

Primeiramente, gostaria de agradecer a Deus pela oportunidade de estar concluindo uma etapa muito importante da minha vida e por ter me proporcionado uma família maravilhosa, a qual sempre esteve ao meu lado em todos os momentos e depositou confiança em mim até quando eu mesma já não a tinha.

Agradeço a todos os alunos da turma III de Física Médica da Unesp-Botucatu, pois apesar de todos os acontecimentos, sejam estes bons ou ruins, cada um contribuiu a sua maneira para a minha formação. Em especial, agradeço a minha mãe-irmã Káthia Takeda da Silva pelo apoio, carinho e cuidado durante todos esses anos.

Agradeço a todos os meus amigos, tanto os que estão sempre junto a mim quanto os que, infelizmente, a distância não permite a companhia, pois todos são muito especiais e sem os quais eu não poderia seguir em frente.

Agradeço, também, ao meu orientador Dr. Ney Lemke, que meu deu oportunidade de realizar este trabalho e sempre esteve disposto a me auxiliar nas dificuldades, sem esquecer-se de me ensinar a ser independente; ao professor Dr. Joel Mesa Hormaza e a todos os meus companheiros do LBBC, que sempre estiveram dispostos a me ajudar e proporcionar um ótimo ambiente de trabalho, com dias muito felizes e engraçados.

Agradeço ao CNPq pelo suporte financeiro que possibilitou o desenvolvimento das pesquisas.

Por fim, mas não menos importante, agradeço ao meu pai e a minha irmã, pelo carinho, apoio, atenção e pela confiança depositada em mim, e a minha mãe, pois apesar de ela não ter tido a oportunidade de acompanhar essa etapa da minha vida, posso sentir que esteve sempre perto o suficiente para me dar forças quando necessário. Amo vocês.

Resumo

A fim de compreender como fenômenos biológicos ocorrem, as interações não lineares entre os componentes envolvidos nestes e seus respectivos elementos, tais como genes, proteínas, etc., podem ser representados por um objeto matemático denominado grafo ou rede, onde os elementos interagentes são representados por nodos que são conectados por linhas. A análise de várias propriedades relacionadas a grafos de redes biológicas revela informações sobre os processos biológicos, dentre estas se pode citar a estrutura dos grafos, que se organizam em grupos de nodos densamente conectados entre si, mas esparsamente interligados a outros grupos, sendo denominada comunidade ou módulo. A detecção de comunidades pode ser realizada através de diversos algoritmos, mas a maioria se baseia na alta densidade de ligações entre os nodos de uma mesma comunidade. Até o momento, a detecção e análise das comunidades em redes biológicas foram realizadas apenas para redes compostas com apenas um tipo de interação (ex.: interações proteína-proteína ou interações metabólicas). Como uma rede biológica real é composta, simultaneamente, por interações proteína-proteína, metabólicas e de regulação transcricional, seria interessante investigar como as comunidades são organizadas neste tipo de rede. Para este propósito, nós detectamos as comunidades em redes biológicas integradas da bactéria *Escherichia coli* e da levedura *Saccharomyces cerevisiae* através do *Clique Percolation Method* e verificamos, através do cálculo da frequência de cada tipo de interação e sua respectiva entropia, se os componentes dessas tendem a ser interligados com um tipo preferencial de interação. Posteriormente, comparamos estes resultados com obtidos de redes aleatórias para determinar se essas comunidades têm relevância biológica. Por fim, foi obtida a entropia para os *clusters* agrupados devido a similaridades no perfil de expressão na estrutura da rede da *Saccharomyces cerevisiae* gerada através de dados de *microarray*, e os resultados comparados com os valores obtidos para as comunidades encontradas anteriormente.

Palavras-chave: bioinformática, redes biológicas integradas, comunidades.

Abstract

To understand how biological phenomena emerge, the nonlinear interactions among the components envolved in these and the correspondent connected elements, like genes, proteins, etc., can be represented by a mathematical object called graph or network, where interacting elements are represented by edges connecting pairs of nodes. The analysis of various graph-related properties of biological networks has revealed many clues about biological processes. Among these properties, the community structure, i.e. groups of nodes densely connected among themselves, but sparsely connected to other groups, are important for identifying separable functional modules within biological systems for the comprehension of the high-level organization of the cell. Communities' detection can be performed by many algorithms, but most of them are based on the density of interactions among nodes of the same community. So far, the detection and analysis of network communities in biological networks have only been pursued for networks composed by one type of interaction (e.g. protein-protein interactions or metabolic interactions). Since a real biological network is simultaneously composed by protein-protein, metabolic and transcriptional regulatory interactions, it would be interesting to investigate how communities are organized in this type of network. For this purpose, we detected the communities in an integrated biological network of the *Escherichia coli* and *Saccharomyces cerevisiae* by using the *Clique Percolation Method* and we verified, by calculating the frequency of each type of interaction and its related entropy, if components of communities tend to be interconnected with a preferred type of interaction. Later, we compared these results with those obtained from randomized networks to determine if these communities have biological relevance. Finally, entropy was obtained for clusters grouped for similarities in the profile of expression for *S cerevisiae*'s integrated network, that it was created by microarray data, and results were compared with values obtained for previously identified communities.

Key-words: bioinformatics, integrated biological networks, communities.

Índice

Página

Resumo **3**

Abstract **4**

1 Introdução **6**

1.1 Teoria de Redes 6

1.2 Biologia de Sistemas 11

2 Objetivos **13**

2.1 Objetivos Gerais 13

2.2 Objetivos Específicos 13

3 Metodologia **14**

3.1 Etapas 14

3.1.1 Obtenção de dados 14

3.1.2 Divisão em comunidades 15

3.1.3 Cálculo da entropia 15

3.1.4 Representação 16

3.2 Banco de dados 16

3.3 Algoritmos 17

3.3.1 `Clauset` 17

3.3.2 *Clique Percolation* 17

3.3.3 *MinCut* 18

3.4 Implementação 19

4 Resultados e Discussão **21**

5 Conclusões **28**

1 Introdução

Uma questão importante na fusão das áreas conhecidas por biologia sistêmica e teoria de redes é como e porque integrar dados biológicos, gerar redes de dados e, mais especificamente neste trabalho, por que identificar comunidades presentes nestas redes.

Considerando o estudo das células de um organismo, estas são compostas por muitos tipos distintos de moléculas e, apesar das informações celulares estarem contidas nas moléculas de DNA, as características celulares dependem das interações entre vários constituintes e destes com o meio. Desta forma, a aplicação da teoria de redes na biologia sistêmica visa estudar, dentre outras coisas, como as propriedades celulares se relacionam com as interações moleculares e gênicas que as constituem e qual a importância da determinação dos grupos de interações formados em cada rede.

1.1 Teoria de Redes

Alguns fenômenos biológicos são propriedades emergentes de interações não lineares entre múltiplos componentes de processos biológicos, tais como genes, proteínas, metabólitos e outros. Para compreender como estas propriedades se originam, elementos interagentes e suas respectivas conexões são representados por objetos matemáticos denominados grafos ou redes, nos quais as interações são representadas por linhas (*edges*) que conectam pares de nodos [Barabási e Oltvai 2004].

Uma rede pode ser representada de diversas maneiras dependendo da análise a ser realizada. A Figura 1 ilustra algumas formas de representação utilizando um exemplo de rede com quatro tipos diferentes de nodos identificados por suas posições estruturais. Na parte a) da figura, pode-se visualizar as interações entre os grupos de nodos: os nodos dos tipos A e C são densamente conectados internamente, já os nodos do tipo B tem conexões com os tipos A e C, mas não possuem conexões entre si, são responsáveis por interligar os grupos A e C; os nodos do tipo D não possuem conexões entre si, interagindo apenas com nodos do grupo C e, deste modo, formam a periferia dos nodos tipo C. As partes b) e c) da figura mostram outras duas possíveis maneiras de representação dos grafos, sendo que em b) a rede esta dividida em grupos, nos quais os nodos interagem mais intensamente, enquanto em c) a representação é feita através dos tipos de nodos. Já as partes d) e e) ilustram uma forma de representação denominada matriz de adjacência, considerando os dados em grupos ou tipos de nodos, assim como nos itens b) e c). A matriz de adjacência trata-se de uma matriz de dimensão $n \times n$, na qual as linhas e colunas representam os nodos e a intersecção dos elementos dessa matriz representa a existência ou não de interações entre os estes. A matriz de adjacência é binária: $M(i, j) = 1$, caso exista interação entre os nodos i e j; $M(i, j) = 0$, caso não exista interação [Pinkert, Schultz e Reichardt 2008].

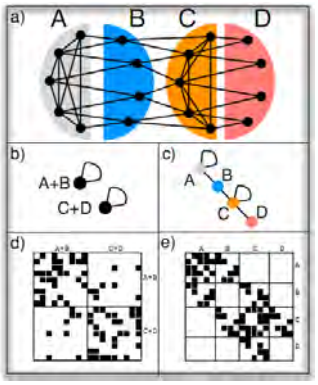
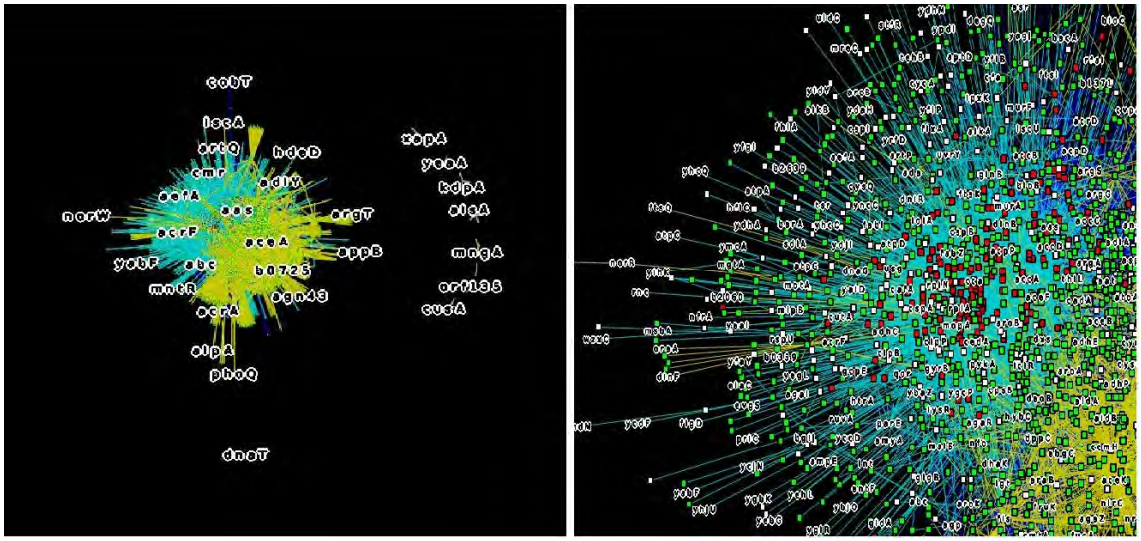


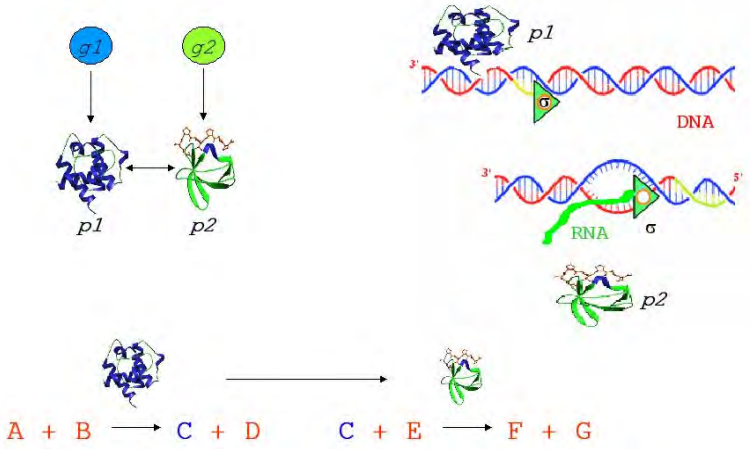
Figura 1: Formas de representação de redes [Pinkert, Schultz e Reichardt 2008].

Em uma rede biológica integrada, figura 2 (figura cedida por Ney Lemke), considerando que os genes (*nodos*) *g1* e *g2* codificam as proteínas *p1* e *p2*, estes são conectados na rede, se ocorre:

1. interação proteína-proteína: *p1* e *p2* interagem fisicamente;
2. regulação transcricional: *g1* regula a transcrição do gene *g2* ou
3. interação metabólica: um produto gerado pela reação catalisada pela proteína *p1* é consumido na reação catalisada pela proteína *p2*.



(a) Rede integrada da bactéria *Escherichia 295 coli*.



(b) Esquema dos tipos de interações.

Figura 2: Rede biológica integrada e os tipos de interações.

A análise das propriedades dos grafos relacionadas às redes biológicas revela muitas informações sobre processos biológicos e segue alguns passos [Joyce e Palsson 2006]:

- Identificar a rede através das interações entre os seus componentes;
- Desenvolver modelos de sistemas que simulem o comportamento da rede diante de situações distintas;
- Dividir a rede em subredes, módulos ou comunidades visando compreender a estrutura geral.

Dentre as várias propriedades analisadas, pode-se citar a estrutura de grafos, os quais se organizam em grupos de nodos densamente conectados entre si, mas esparsamente

interligados a outros grupos, sendo esta formação conhecida por módulo ou comunidade. Pode-se citar diversos exemplos de redes organizadas em módulos, assim como a *worldwide web* formada por grupos de páginas na internet divididos pelo assunto dos mesmos, redes sociais que são organizadas em unidades sociais, dentre outros [Newman 2006]. Redes biológicas possuem, em geral, estrutura modular.

A importância da estrutura modular em uma rede biológica é sustentada por diversos fatores [Hartwell et al. 1999]:

- Ao se identificar uma comunidade e atribuir a esta uma função, pode-se estudar a função e a relação desta com cada constituinte deste módulo;
- A habilidade celular de integrar informações vindas de fontes distintas do organismo pode ser descrita pelos padrões das interações entre os módulos identificados;
- A estrutura modular facilita mudanças evolutivas, já que a função principal de um módulo é resistente a mudanças, mas as propriedades e funções de uma célula (fenótipo desta) podem sofrer modificações evolutivas através de alterações nas interações entre os módulos distintos. Por exemplo, se a função de uma proteína afetasse diretamente todas as propriedades de uma célula, provavelmente esta proteína não sofreria mudanças, mas se a função de uma proteína está restrita a um grupo funcional, apenas, e as interações entre os grupos ocorrem através de proteínas individuais, será muito mais fácil modificar e suprimir interações desta proteína com outros módulos funcionais;
- O conhecimento de um membro de uma comunidade pode auxiliar no estudo de um membro de outra comunidade relacionada a primeira através das interações;
- Do mesmo modo que a comparação de sequencias protéicas de espécies relacionadas pode sugerir suas funções, a comparação entre módulos de funções comuns, em organismos distintos, pode auxiliar na compreensão da função desses módulos.

Um outro exemplo de propriedade analisada que é importante e de fácil compreensão é a resistência. Esta propriedade indica o quão resistente é determinada rede em relação à remoção de nodos e à manutenção destes com funcionamento anormal. Uma rede ideal, de acordo com esta propriedade, é aquela que é resistente à remoção de nodos, mas frágil ao mau funcionamento destes. Embora redes ideais não existam, uma rede que representa bem este tipo é a rede da *Internet*: se um roteador é desligado e, desta forma, removido da rede, a rede global não é afetada, apenas alguns usuários dependentes deste roteador seriam, por isso, pode-se dizer que é resistente à remoção de nodos; mas, se um vírus é enviado constantemente à rede ou esta recebe muitos e-mails *spams*, toda a estrutura

global da rede é ameaçada, o que mostra ser frágil ao mau funcionamento dos nodos [Kitano 2007].

Pode-se, também, realizar análise através de mais de uma propriedade, como, por exemplo, identificar a resistência de cada comunidade formada na rede. Este tipo de estudo é muito utilizado para encontrar grupos que seriam possíveis alvos para aplicação de drogas, através da identificação de grupos resistentes à droga ou grupos que estejam com funcionamento anormal, os quais podem ter seu funcionamento normal restabelecido se forem alvos para drogas em pesquisa. O sucesso de muitas drogas no passado deu-se a estas atingirem o ponto de fragilidade, regulando processos biológicos através da modificação de apenas um ou poucos grupos identificados previamente [Kitano 2007]. Em alguns casos, pode ser necessária a inativação de determinada função e isto pode ser alcançado a partir da remoção de um grupo funcional inteiro [Gulbahce e Lehmann 2008].

Uma outra área na qual a identificação de módulos é de grande importância é a biologia molecular, mais exatamente, a área relacionada ao estudo da função dos genes. Um método utilizado para estudar a função gênica é a deleção de genes e posterior observação das mudanças ocorridas no fenótipo. Porém, a deleção de um único gene não revela muita informação, enquanto a deleção de grupos de genes, pelo contrário, pode gerar muitas informações importantes. Como organismos eucariontes possuem uma grande quantia de genes, a combinação destes para deleção de grupos se torna inviável, portanto, uma aproximação computacional promissora pode ser utilizada para identificar os grupos.

Se as comunidades estão claramente separadas no sistema, a maioria dos métodos de aglomeração pode identificá-las. De fato, muitos métodos têm sido introduzidos recentemente para identificar módulos de redes variadas [Newman 2006] [Adamcsek et al. 2006] [Clauset, Newman e Moore 2004] [Palla et al. 2005] [Kumpula et al. 2008], usando ainda a descrição topológica das redes ou combinando a topologia com os dados genômicos funcionais integrados. Entretanto, métodos diferentes predizem divisões diferentes entre módulos que não estão claramente separados. Atualmente, não há critério matemático objetivo para decidir que uma divisão é melhor que outra, assim, alguns parâmetros internos controlam o tamanho típico de módulos não conhecidos e a mudança destes gera as diferenças entre métodos distintos de divisão em comunidades. Uma solução, portanto, é evitar procurar uma separação em um grupo absoluto de módulos, mas sim visualizar uma relação hierárquica entre comunidades de tamanhos diferentes.

Devido a esta diversidade de métodos de divisão de redes em comunidades, seguem duas linhas de pesquisa: partição de grafos e detecção de estrutura de comunidade. A primeira é adotada, particularmente, nas áreas de ciência da computação e campos relacionados, com aplicação em computação paralela e elaboração de circuitos integrados, dentre outras áreas. A segunda é utilizada por sociólogos e, recentemente, por físicos, biólogos e matemáticos, com aplicação, especialmente, em redes sociais e biológicas

[Newman 2006]. Ambas as linhas de pesquisa descritas são direcionadas a mesma questão, embora por meios um pouco diferentes. Há, entretanto, diferenças entre os objetivos dos dois campos que originam a diferença técnica das aproximações desejadas.

Um exemplo de aplicação típica da partição de grafos é a divisão de tarefas entre os processadores na computação paralela para minimizar a quantidade necessária de comunicação entre os processadores. Em cada cálculo, o número de processadores e a quantia de tarefas que cada um destes pode suportar são, normalmente, conhecidos e, deste modo, pode-se identificar em quantos grupos a rede será dividida e o tamanho de cada grupo. Além disso, o objetivo é, normalmente, encontrar a melhor divisão da rede independente se uma boa divisão existe [Newman 2006].

A detecção de estrutura de comunidades, pelo contrário, é, talvez, o melhor método para identificar a estrutura de redes de larga escala, como redes sociais, da internet e dados da *web*, ou redes biológicas. Este método assume que a rede de interesse está dividida naturalmente em subgrupos e o trabalho dos pesquisadores é identificá-los. O número e o tamanho dos grupos são determinados pela própria rede e não pelo pesquisador. Além disso, este método pode, explicitamente, admitir a possibilidade de que não haja uma boa divisão para a rede [Newman 2006].

1.2 Biologia de Sistemas

A biologia de sistemas abrange o estudo de redes biológicas complexas, como metabólicas, gênicas, proteicas, dentre outras, e visa, a partir de fenômenos complexos, prover uma estrutura abstrata que aumente a compreensão da dinâmica de redes biológicas [Ferrell 2009]. Em um senso limitado, a biologia de sistemas pode ser considerada uma antítese ao reducionismo, pois, apesar da importância de identificar e conhecer cada parte de um sistema biológico, essas informações não são suficientes para a compreensão da complexidade intrínseca de um sistema. Portanto, deve-se considerar não só a estrutura de um sistema, mas também a dinâmica deste, ou seja, as informações sobre cada componente e o comportamento destes sob determinadas condições [Ferrell 2009] [Kitano 2002].

Muitas pesquisas são, atualmente, realizadas a partir do uso de dados de *microarray* para a obtenção de estrutura de redes biológicas. No caso de redes de interações gênicas, obtém-se o perfil de expressão, a análise de *cluster* é realizada e, assim, os genes são agrupados em *clusters* de acordo com suas similaridades de expressão. Este tipo de aglomeração é utilizada devido à hipótese de que genes co-expressos estão, provavelmente, relacionados a funções biológicas [Ge et al. 2001]. No trabalho realizado por [Ge et al. 2001], os dados de interações proteína-proteína foram obtidos a partir de diversas fontes, descriminadas na metodologia do artigo, e foi feita a análise de *clustering*, formando grupos de genes envolvidos na regulação do ciclo celular, regulação de meiose e

regulação do stress celular. Os 30 grupos obtidos nessa análise são utilizados no presente trabalho.

Existem, ainda, outras formas de construir redes biológicas, neste caso, através de informações sobre as interações entre os diferentes genes que compoem um organismo. Estas interações incluem interações regulatórias, metabólicas e físicas, entre outras possibilidades. Combinando essas informações com algoritmos de decisão que tem como base a entropia, hipóteses sobre a estrutura e dinâmica de redes são sugeridas com menor ambiguidade [Kitano 2002] [Ferrell 2009]. Espera-se que as entre redes construídas utilizando estas abordagens devam possuir similiaridades ou até idealmente equivalências. Esta questão, contudo, ainda não foi suficientemente explorada. Neste trabalho iremos comparar as redes construídas por estas abordagens.

Até agora, a detecção e análise de comunidades são realizadas em redes biológicas envolvendo apenas um tipo de interação, por exemplo, interação proteína-proteína ou interação metabólica. Como uma rede biológica real é composta, simultaneamente, por interações proteína-proteína, metabólicas e de regulação transcricional, esta pesquisa abordará como as comunidades são organizadas em redes que possuam os três tipos de interações, identificando-as, pelos métodos de detecção de estrutura e partição de grafos, e analisando-as a partir do cálculo da frequência de cada tipo de interação e sua respectiva entropia.

2 Objetivos

2.1 Objetivos Gerais

1. Determinar a estrutura de comunidades para as redes biológicas integradas da bactéria *Escherichia coli* e da levedura *Saccharomyces cerevisiae*.

2.2 Objetivos Específicos

1. Comparar algoritmos distintos para a divisão de redes biológicas integradas em comunidades;

2. Obter as frequências de cada tipo de interação e suas respectivas entropias, para as comunidades determinadas por cada método.

3 Metodologia

3.1 Etapas

As etapas realizadas podem ser acompanhadas através da figura 3.

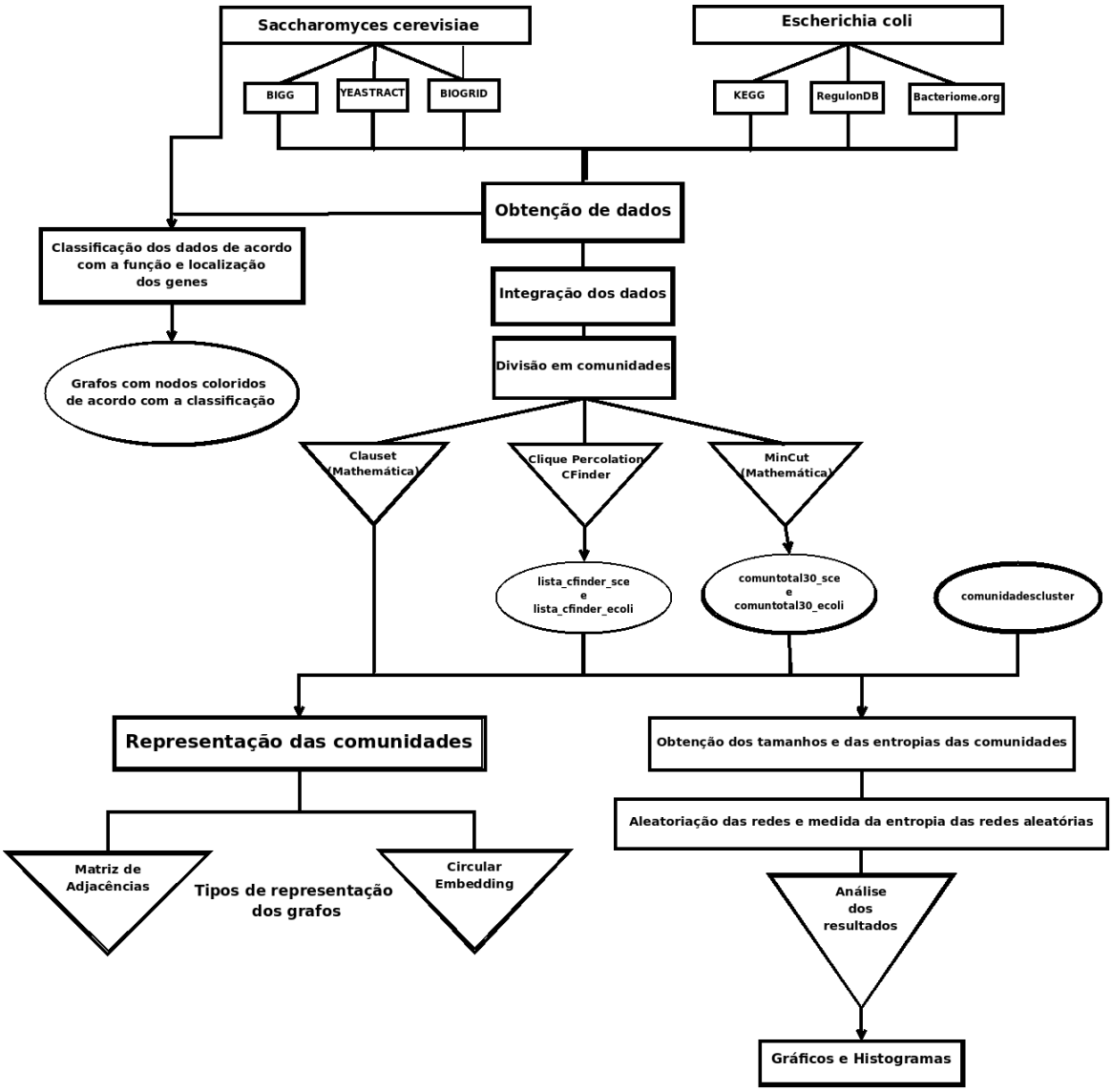


Figura 3: Esquema da metodologia utilizada neste trabalho.

3.1.1 Obtenção de dados

A obtenção dos dados inicia com a busca nos bancos de dados especificados no item 3.2 e estes são utilizados para a obtenção das redes da bactéria *Escherichia coli* e da levedura *Saccharomyces cerevisiae*.

Em seguida, os dados foram analisados, integrados e, os da *S. cerevisiae*, classificados de acordo com suas funções e localizações.

3.1.2 Divisão em comunidades

Ambas as redes foram divididas em comunidades utilizando três métodos distintos, sendo duas funções contidas no *software* `Mathematica`, denominadas *CommunityStructureAssignment*, que aplica o algoritmo proposto por `Clauset`, e *MinCut*, e o utilizado pelo *software* `CFinder`, denominado *Clique Percolation Method*. Os algoritmos utilizados estão especificados no item 3.3 e os códigos no item 3.4.

3.1.3 Cálculo da entropia

Neste trabalho, a entropia representa a medida de incerteza de uma variável aleatória e é utilizada como ferramenta de análise para os resultados obtidos. Seja n uma variável aleatória discreta e a função de probabilidade $p(n)$. Supondo x variáveis $\{n_1, n_2, n_x\}$, as quais aparecem em um sistema independentemente, ou seja, sem dependência dos valores anteriores e com probabilidades $\{p_1, p_2, p_n\}$ [Cover e Thomas 2006] [Shannon 1948].

Sendo a quantidade de informação de cada variável do sistema dada por $-\log_2(p_i)$ em uma única observação, para N observações, cada variável aparece no sistema N vezes a sua probabilidade, isto é, $N \times p_i$ e, assim, para N observações independentes, tem-se a informação total dada pela equação 1 e a informação média por variável pela equação 2 [Shannon 1948].

$$I = -\sum_{i=1}^{x} Np_i \log(p_i) \tag{1}$$

$$\frac{I}{N} = -\frac{I}{N}\sum_{i=1}^{x} Np_i \log(p_i) = -\sum_{i=1}^{x} p_i \log(p_i) \tag{2}$$

Deste modo, a informação é estritamente descrita em termos de probabilidade de eventos. Supondo uma distribuição de probabilidades $P = \{p_1, p_2, p_n\}$, a entropia desta, para valores discretos, é dada pela equação 3. Quando o log é de base 2, a entropia é expressa em bits.

$$H(P) = -\sum_{i=1}^{x} p_i \log_2(p_i) \tag{3}$$

As entropias dos padrões de distribuição dos tipos de interações em cada comunidade da bactéria *Escherichia coli* e da levedura *Saccharomyces cerevisiae* foram obtidas através dos códigos, contidos no item 3.4, que utilizam a equação 3 para medir o teor de informação das distribuições. Um alto valor de entropia, medido na comunidade, indica uma

distribuição homogênea dos tipos de interações, ou seja, presença dos três tipos de interações existentes, sem predominância de nenhum dos três tipos [Butte e Kohane 2000].

Após o cálculo da entropia, as comunidades foram aleatorizadas. Isto é realizado alterando os tipos de interações entre os genes de maneira aleatória. Em seguida, as entropias das comunidades das redes aleatórias foram medidas e foi realizada uma comparação entre os resultados obtidos para as redes reais e as aleatorizadas.

3.1.4 Representação

Os genes de ambas as redes foram reordenados de acordo com a divisão em comunidades utilizando a função, do `Mathematica` que aplica o algoritmo de `Clauset`, *CommunityStructureAssignment*, e as redes representadas através das matrizes de adjacências.

Posteriormente, as redes obtidas através da divisão de comunidades utilizando os outros dois métodos foram representadas através da utilização da função *Circular Embedding*, presente no `Mathematica`, e foram calculados os tamanhos e as entropias, como especificado no item 3.1.3, para estas comunidades. Os resultados destes cálculos estão apresentados nos gráficos e histogramas da seção 4. Para a *Saccharomyces cerevisiae*, foram calculados os tamanhos e as entropias para os *clusters* identificados por [Ge et al. 2001], baseado na coexpressão dos genes e, estes resultados também estão presentes na seção 4 em forma de gráfico e histograma.

3.2 Banco de dados

Os dados utilizados sobre as interações físicas entre proteínas, interações metabólicas e interações de regulação transcricional, referentes à *Escherichia coli* e à *Saccharomyces cerevisiae*, foram obtidos de bancos de dados biológicos públicos na *web*. A Tabela 1 mostra os bancos de dados utilizados na aquisição de dados.

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

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

Após a obtenção dos dados, os mesmo foram normalizados e integrados, ou seja, os nomes dos genes receberam um formato único e os dados centralizados em um único

arquivo, possibilitando o estudo e análise destes em um processo mais detalhado. Posteriomente, foi realizada a limpeza dos dados, isto é, correção de dados incorretos ou imprecisos e esta operação foi realizada por um especialista da área biológica.

3.3 Algoritmos

3.3.1 `Clauset`

O algoritmo publicado por `Clauset` [Clauset, Newman e Moore 2004] é baseado na variação do valor da modularidade encontrado entre os nodos na tentativa de maximizar essa. Inicialmente, adiciona-se o nodo central (*hub*) à comunidade C e seus vizinhos à U; a cada passo, o algoritmo adiciona à C, os nodos que maximizam a variação do valor de modularidade ou que menos reduzam essa, isto é, para cada vizinho adicionado a U, calcula-se o valor da variação da modularidade através da Equação 4 e os nodos que maximizam este são adicionados à C. Este processo continua até que atinja-se um determinado número de nodos k ou até que uma comunidade seja totalmente identificada.

$$\Delta Q_{ij} = \begin{cases} \frac{1}{2m} - \frac{k_i k_j}{(2m)^2} & \text{se i, j estão interagindo} \\ 0 & \text{se i, j não estão interagindo} \end{cases} \tag{4}$$

Onde, m é o número de ligações de cada comunidade, dado por $m = \frac{1}{2}\sum_{ij} A_{ij}$, sendo que A_{ij} trata-se da matriz de adjacências da rede e k_i e k_j são os graus, ou seja, o número de ligações, dos nodos i e j.

3.3.2 *Clique Percolation*

Este algoritmo possui duas etapas e para otimizar a compreensão destas deve-se, primeiramente, definir *k-clique* e *comunidade k-clique*. *K-clique* é uma rede de k nodos, na qual todos estão ligados entre si; *comunidade k-clique* ou *comunidade-k* trata-se de uma rede de nodos que é obtida através da sobreposição de *k-cliques* de modo que os *k-cliques* compartilhem *k-1* nodos. O algoritmo funciona da seguinte maneira: inicialmente, todos os *links* l são removidos da rede e então inseridos novamente um por um e em qualquer sequência para redes em que as interações não possuam pesos diferentes. Deste modo, a primeira etapa do algoritmo consiste na detecção dos *k-cliques* conforme são adicionados os *links*; na segunda etapa, a determinação das comunidades é dada através da reconstrução da rede com a sobreposição dos *k-cliques* identificados na primeira etapa. A determinação das comunidades é feita através da detecção local de possíveis sobreposições conforme novos *k-cliques* são adicionados à rede que está sendo reconstruída e a estrutura de comunidades é atualizada conforme são identificadas novas sobreposições.

A figura 4 ilustra as etapas para determinação das comunidades, utilizando como

exemplo uma rede simples com $k = 3$.

1. rede original consiste de três 3-cliques representados como A, B e C;

2. representação da rede dividida de acordo com as estruturas dos cliques. Pode-se notar que os 3-cliques B e C, que formam uma comunidade 3-clique, são conectados pelo 2-clique f. O clique A gera outra comunidade 3-clique;

3. 3-cliques identificados na primeira parte do algoritmo conforme os links eram sequencialmente inseridos na rede. Cada novo k-clique é denotado por nodos vermelhos enquanto os nodos associados a estes são representados cinzas;

4. Atualizações da rede $(k-1)-clique$ como resultado da segunda parte do algoritmo. Comunidades $k-clique$ correspondem aos componentes conectados (áreas circuladas) [Palla et al. 2005].

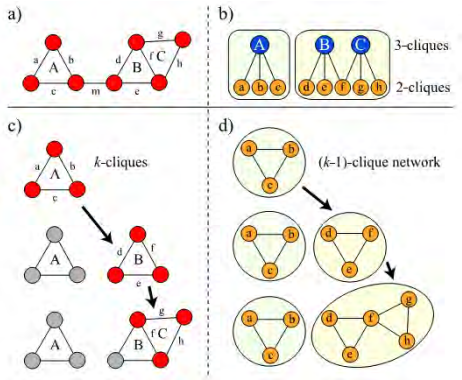


Figura 4: Ilustração das etapas do algoritmo *Clique Percolation*, retirada de [Kumpula et al. 2008].

3.3.3 *MinCut*

A função *MinCut*, contida no `Mathematica`, é um método de divisão em comunidades por partição de grafos, ou seja, o número de comunidades é definido previamente. A função considera a rede como sendo indireta e a divide em k grupos com, aproximadamente, o mesmo número de nodos, minimizando o número de interações entre os grupos.

O número pré definido de comunidades, aplicado neste trabalho, foi de 30 grupos.

Tabela 2: Tabela comparativa para os métodos de divisão de comunidades utilizados.

Método	Tipo de divisão (número de comunidades)	Intersecção entre comunidades	Uso de CPU
Clauset	Detecção de estrutura	Vazia	20 minutos
Clique Percolation	Detecção de estrutura	Não vazia	12 horas
MinCut	Partição de grafos	Vazia	2 minutos

3.4 Implementação

Para a realização da metodologia descrita, foram utilizados o *software* `CFinder` e o `Mathematica` e pacotes de funções como *Combinatorica* e *GraphUtilities*. Os *notebooks* (códigos implementados no `Mathematica`) estão disponíveis no *site*[1] e descritos a seguir.

`comunidades_circulo_cfinder_ecoli.nb.nb` $*$[2] Faz a representação gráfica das comunidades identificadas na rede da *E. coli*, pelo `CFinder`, utilizando *Circular Embedding*.

`comunidades_ecoli_clauset.nb` $*$ Divide a rede da *E. coli* em comunidades usando o algoritmo de `Clauset` [Clauset, Newman e Moore 2004], faz a representação através da matriz de adjacência e calcula a entropia das comunidades.

`entropia_ecoli_cfinder.nb`$*$ Calcula a entropia das comunidades determinadas pelo `CFinder` para a rede da *E. coli*.

`color_nodos_funcao_final_sce.nb`$*$ Gera o grafo da rede da *S. cerevisiae* e colore os nodos de acordo com a função.

`color_nodos_localizacao_final_sce.nb`$*$ Gera o grafo da rede da *S. cerevisiae* e colore os nodos de acordo com a localização.

`comunidades_circulo_cfinder_sce.nb`$*$ Faz a representação gráfica das comunidades identificadas na rede da *S. cerevisiae*, pelo `CFinder`, utilizando *Circular Embedding*.

`comunidades_circulo_sce.nb` Divide a rede da *S. cerevisiae* em comunidades usando *MinCut* e faz a representação gráfica das comunidades utilizando *Circular Embedding*.

[1] Os códigos estão disponíveis no Twiki, http://www.esyurl.com/574.
[2] Todos os códigos com $*$ são os quais tive participação.

`comunidades_Sce_clauset.nb*` Divide as comunidades usando o algoritmo de `Clauset` [Clauset, Newman e Moore 2004], faz a representação através da matriz de adjacência e calcula a entropia das comunidades.

`entropiacluster.nb*` Calcula a entropia dos clusters agrupados de acordo com o perfil de expressão dos genes da *S. cerevisiae*.

`entropiaMinCut30.nb*` Calcula a entropia das comunidades determinadas pelo *MinCut* para a rede da *S. cerevisiae*

`entropia_sce_cfinder.nb*` Calcula a entropia das comunidades determinadas pelo `CFinder` para a rede da *S. cerevisiae*.

`grafolbbc.nb` Possui diversas funções que foram utilizadas nos códigos descritos acima.

4 Resultados e Discussão

Três métodos distintos foram utilizados na determinação das comunidades para as redes biológicas integradas da *E. coli* e da *S. cerevisiae* e como cada algoritmo utiliza um método de identificação, diferenças podem ser notadas e estão apresentadas nas Tabelas 2 e 3 e na Figura 16.

As diferenças mais notáveis entre os tipos de divisões em comunidades são o número de comunidades obtidas por cada método e o tamanho destas. Isto pode ter relação com o fato de haver padrões distintos de divisão: partição de grafos e divisão por estrutura de comunidades (ver Tabela 2).

Tabela 3: Quantia de comunidades obtidas por cada método de divisão de comunidades utilizado.

	Número de comunidades	
Método	*E.coli*	S. cerevisiae
Clauset	127	28
Clique Percolation	220	1691
MinCut	30	30

O algoritmo proposto por `Clauset` [Clauset, Newman e Moore 2004] divide as redes pelo método de estrutura de comunidades, isto é, assume que a rede possui uma estrutura naturalmente dividida em módulos e os identifica. Este método dividiu a rede da *E. coli*, a qual possui 2349 genes e 6115 interações, em 127 comunidades e a da *S. cerevisiae*, que possui 5667 genes e 72802 interações, em 28 comunidades (ver Tabela 3). As comunidades identificadas estão representadas nas Figuras 5 e 10 para *E. coli* e *S. cerevisiae*, respectivamente, através das matrizes de adjacência.

Na Figura 5, referente à *E. coli*, pode-se observar a formação de uma comunidade bem distinta com predominância de interações do tipo proteína-proteína, uma leve aglomeração com predominância de interações metabólicas e as interações regulatórias estão dispersas e são dificilmente visualizadas. Já na Figura 10, referente à *S. Cerevisiae* pode-se verificar a presença de três comunidades maiores e dentro destas a existência de comunidades menores, sendo que na primeira comunidade há presença de interações proteína-proteína e metabólica; já na segunda, há predomínio de ligações proteína-proteína e a última comunidade apresenta interações de todos os tipos e muitas comunidades menores. As interações do tipo regulatórias estão menos concentradas (mais espalhadas), indicando que um mesmo gene pode participar do processo regulatório de vários outros, interagindo com vários genes e originando uma linha horizontal na matriz.

A divisão em comunidades realizada pelo `CFinder`, com o método *Clique Percolation* é, também, feita por estrutura de comunidades. O método também foi aplicado para a *E.*

coli e para a *S. cerevisiae* e identificou, respectivamente, 220 e 1691 comunidades. Pode-se notar uma grande diferença no número de comunidades encontradas pelo algoritmo de `Clauset` e de *Clique Percolation* e isso ocorre devido a aplicação de critérios distintos de detecção de comunidades.

Já a divisão pelo método *MinCut* é realizada por partição de grafos, ou seja, o número de comunidades, no qual a rede será dividida, é previamente definido e o tamanho é aproximadamente o mesmo para todas. Neste trabalho, foi realizada a divisão das redes da *E. coli* e da *S. cerevisiae* em 30 comunidades.

As Figuras 6 e 11 são grafos que representam as comunidades identificadas através dos algoritmos *Clique Percolation* e *MinCut*, sendo que cada cor representa uma comunidade. As diferenças nos tamanhos das comunidades identificadas para a *S. cerevisiae* podem ser observadas a partir da Figura 11, pois na parte *a* desta há o predomínio da cor azul, o que indica a presença de uma grande comunidade, enquanto na parte *b*, é possível a visualização de mais cores, o que indica a presença de comunidade de tamanhos aproximadamente iguais.

Para as comunidades obtidas por todos os métodos utilizados foram medidas as frequências de cada tipo de interação e sua repectiva entropia, para verificar se há predomínio de algum tipo de interação nos módulos determinados. Apesar das diferenças entre os métodos, os resultados apresentaram um certo padrão, pois os valores de entropia medidos foram, em geral, bem próximos de zero, o que indica a existência de um tipo predominante de interação nas comunidades identificadas, independente do método utilizado. A fim de verificar se esses valores de entropia não são independentes da estrutura de comunidades, as redes foram aleatorizadas e a entropia medida novamente; os valores obtidos para as redes aleatórias foram próximos de *um*, portanto, distintos dos obtidos para as redes experimentais, o que indica que o predomínio de um tipo de interação está relacionado a estrutura modular das redes. Estes resultados estão apresentados nas Figuras 7, 8, 9, 12, 13 e 14.

Na monografia apresentada por *Germek* [Germek], foram identificados os motivos de ordem 3 nas redes integradas da *E. coli* e da *S. cerevisiae* e a análise da frequência dos tipos de interações em cada motivo foi realizada. Os resultados apontaram para o predomínio de um tipo de interação nos motivos identificados, com 94,81% dos motivos com presença de apenas um tipo de interação para a *E. coli* e 81,65% para a *S. cerevisiae*. Estes resultados estão de acordo com os apresentados no presente trabalho, pois ambos mostram que há predomínio de um tipo de interação em grupos com alta densidade de interações entre seus componentes. Possivelmente, essa concordância de resultados está relacionada ao fato de alguns algoritmos de detecção de comunidades utilizarem a estrutura de *cliques* na determinação da estrutura da rede e, como os motivos triangulares são *cliques* de ordem 3, podem se sobrepor para formar grupos de motivos, os módulos

ou comunidades.

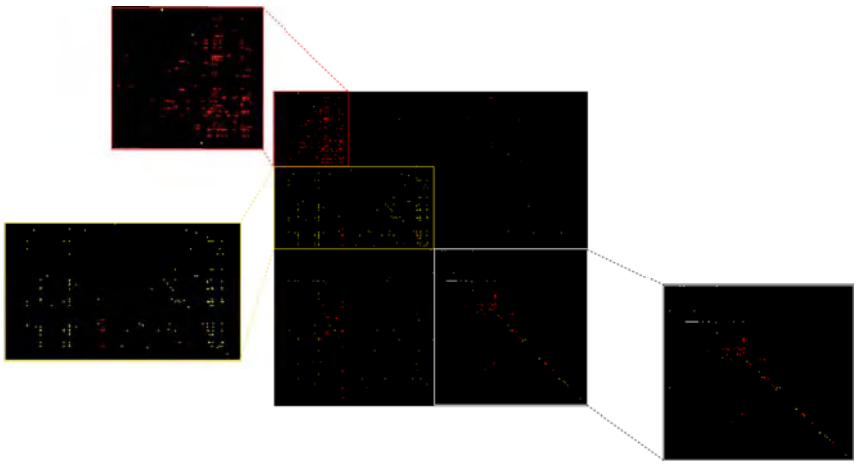
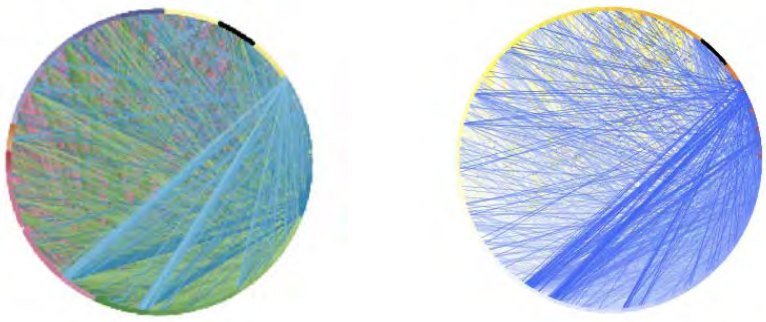


Figura 5: Matriz de Adjacência representando os dados da rede da bactéria *Escherichia coli* após a identificação de comunidades através do algoritmo de `Clauset` [Clauset, Newman e Moore 2004]. As diferentes cores representam os três tipos distintos de interações analisados: em vermelho, interação proteína-proteína; em amarelo, interação metabólica; e, em branco, interação de regulação transcricional.



(a) *Clique Percolation* (b) *MinCut*

Figura 6: Representação gráfica obtida através do `Mathematica` denominada *CircularEmbedding*, com esquema de cores "DarkBands", mostrando as comunidades distintas encontradas na rede da levedura *Escherichia coli* através dos métodos a) *Clique Percolation* e b) *MinCut*.

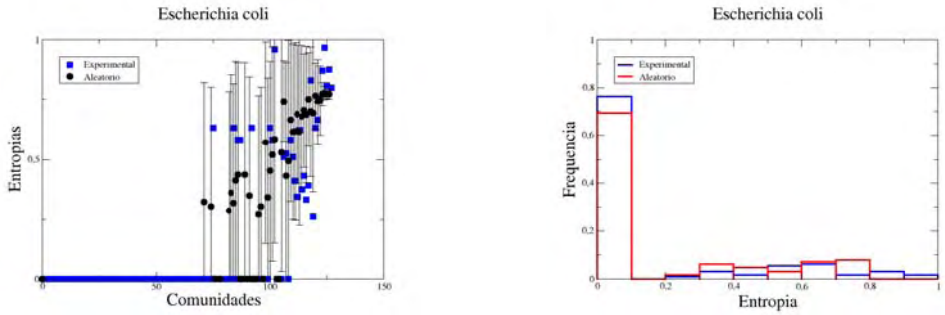


Figura 7: Representação da entropia para as comunidades encontradas pelo método proposto por `Clauset` para a rede integrada da bactéria *Escherichia coli*.

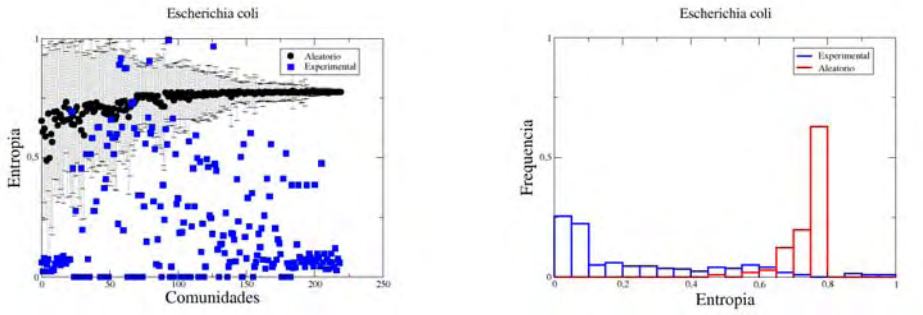


Figura 8: Representação da entropia para as comunidades encontradas pelo método *Clique Percolation* para a rede integrada da bactéria *Escherichia coli*.

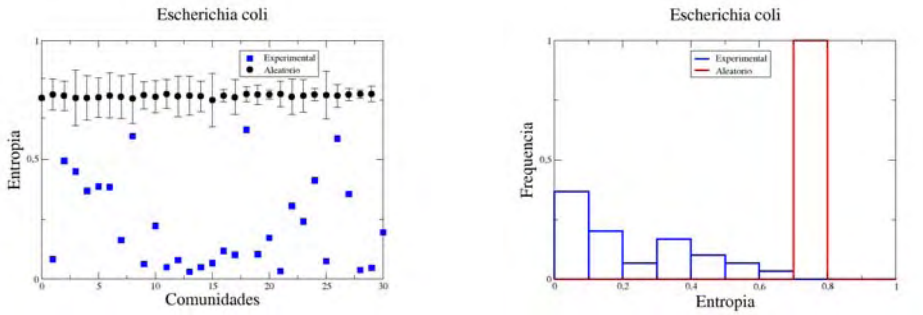


Figura 9: Representação da entropia para as comunidades encontradas pelo método *Min-Cut* para a rede integrada da bactéria *Escherichia coli*.

A fim de investigar uma possível relação entre grupos funcionais e a estrutura de comunidades em redes com predomínio de um tipo de interação, foram medidas as entropias

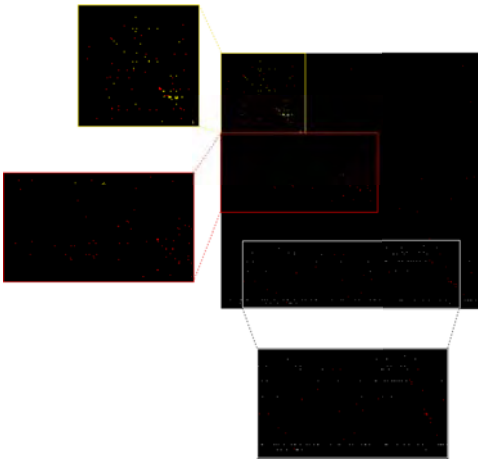
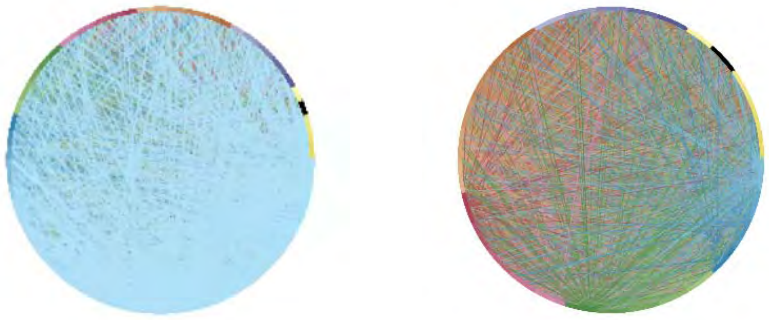


Figura 10: Matriz de Adjacência representando os dados da rede integrada da levedura *Saccharomyces cerevisiae* após a identificação de comunidades através do algoritmo de `Clauset` [Clauset, Newman e Moore 2004]. As diferentes cores representam os três tipos distintos de interações analisados: em vermelho, interação proteína-proteína; em amarelo, interação metabólica; e, em branco, interação de regulação transcricional.



(a) *Clique Percolation* (b) *MinCut*

Figura 11: Representação gráfica obtida através do `Mathematica` denominada *CircularEmbedding*, com esquema de cores "DarkBands", mostrando as comunidades distintas encontradas na Rede integrada da levedura *Saccharomyces cerevisiae* através dos métodos a) *Clique Percolation* e b) *MinCut*.

dos *clusters* de genes identificados a partir da coexpressão na rede integrada da *S. cerevisiae*, em experimento mostrado por [Ge et al. 2001]. Essas medidas foram realizadas porque, em geral, genes coexpressos estão envolvidos em um mesmo processo biológico, seja direta ou indiretamente, e os resultados estão apresentados na Figura 15. Como pode ser observado, os valores obtidos não possuem o mesmo padrão dos obtidos anteriormente

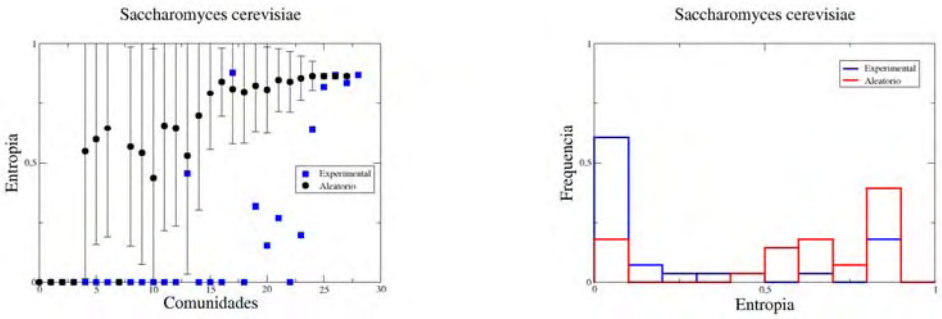


Figura 12: Representação da entropia para as comunidades encontradas pelo método proposto por `Clauset` para a rede integrada da levedura *Saccharomyces cerevisiae*.

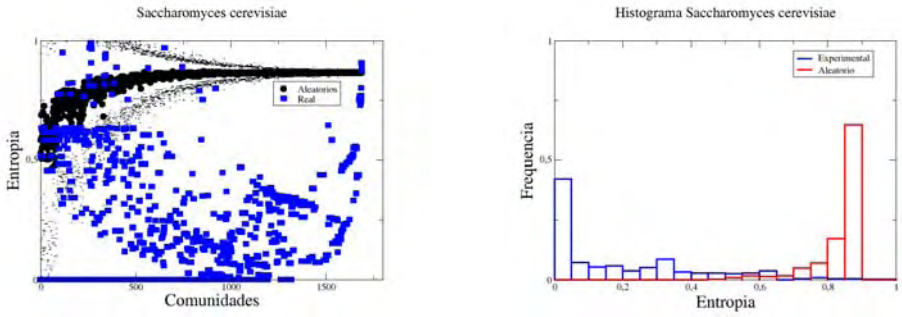


Figura 13: Representação da entropia para as comunidades encontradas pelo método *Clique Percolation* para a rede integrada da levedura *Saccharomyces cerevisiae*.

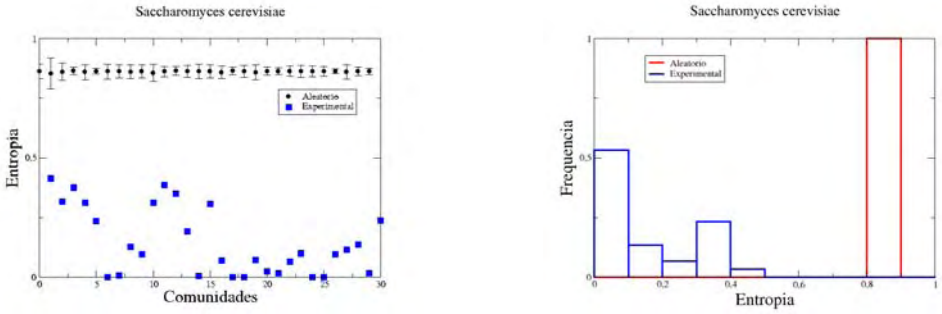


Figura 14: Representação da entropia para as comunidades encontradas pelo método *MinCut* para a rede integrada da levedura *Saccharomyces cerevisiae*.

para as divisões em comunidades, pois o histograma mostra que os *clusters* possuem distribuições distintas das frequências dos tipos de interações, o que indica ausência de um tipo predominante de interação.

Enfim, nota-se o predomínio de um tipo de interação para as comunidades independente do método utilizado neste trabalho, mas não para os *clusters* identificados por co-expressão. Portanto, o fato de uma comunidade possuir um tipo predominante de interação não indica que os genes presentes nesta se agruparam de acordo com a função na qual estão envolvidos e nem que exista formação de comunidades como módulos funcionais.

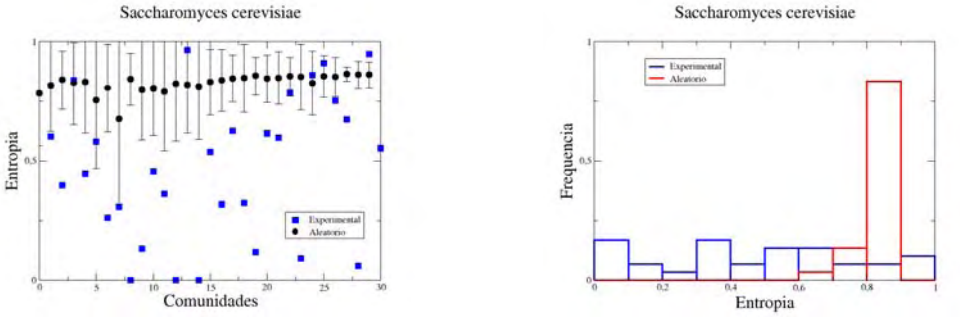


Figura 15: Representação da entropia para os *clusters* identificados, por coexpressão, na rede integrada da levedura *Saccharomyces cerevisiae*.

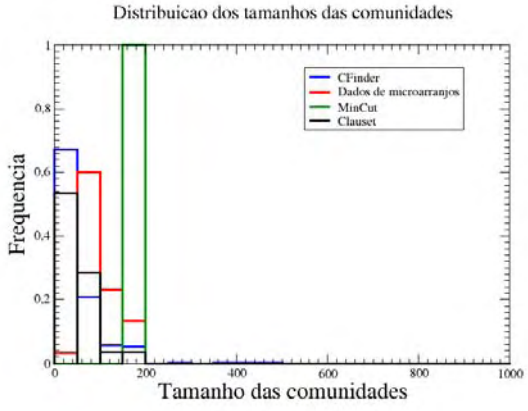


Figura 16: Histograma dos tamanhos das comunidades, obtidas por diversos métodos, para a rede integrada da levedura *Saccharomyces cerevisiae*.

5 Conclusões

Neste trabalho foram determinadas as estruturas de comunidades para as redes integradas da bactéria *Escherichia coli* e da levedura *Saccharomyces cerevisiae*. Podemos concluir que métodos distintos de divisão das redes em comunidades geram divisões diferentes das redes, com números distintos de comunidades e do tamanho destas. Porém, apesar destas diferenças, pode-se afirmar que, para as comunidades identificadas, independentemente do método utilizado, há o predomínio de um tipo de interação entre os genes, mas que este fato não indica que, nestas comunidades, os genes estão agrupados de acordo com as funções as quais estão relacionados.

Como possíveis extensões para este trabalho propomos:

1. Determinar a estrutura de comunidades através de outros algoritmos, como o proposto por Newman [Newman 2006];

2. Calcular a frequência de cada tipo de função das quais os genes das comunidades estão relacionados e sua respectiva entropia.

Trabalhos apresentados

Parte dos resultados desse trabalho foi apresentado em 3 eventos:

1. IV Congresso de Física Aplicada à Medicina, 2008, Botucatu-SP, com o trabalho intitulado *Determinação de comunidades em redes biológicas integradas*. Apresentado sob a forma de pôster pela aluna de iniciação científica Milena Giglioli e cujo resumo consta nos anais do evento.

2. VII Workshop da Pós-Graduação, 2009, Botucatu-SP, com o trabalho intitulado *Análise de comunidades detectadas em redes biológicas integradas*. Apresentado sob a forma de pôster pela aluna de iniciação científica Milena Giglioli e cujo resumo consta nos anais do evento.

3. XXXII Encontro Nacional de Física da Matéria Condensada, 2009, Águas de Lindóia-SP, com o trabalho intitulado *Detection of communities and entropy in integrated biological netwoks*. Apresentado sob a forma de pôster pela aluna de iniciação científica Milena Giglioli e cujo resumo consta nos anais do evento.

Referências

[Adamcsek et al. 2006]ADAMCSEK, B. et al. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, v. 22, p. 1021–1023, Apr 2006.

[Barabási e Oltvai 2004]BARABáSI, A. L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, v. 5, p. 101–113, Feb 2004.

[Butte e Kohane 2000]BUTTE, A. J.; KOHANE, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, p. 418–429, 2000.

[Clauset, Newman e Moore 2004]CLAUSET, A.; NEWMAN, M. E.; MOORE, C. Finding community structure in very large networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 70, p. 066111, Dec 2004.

[Cover e Thomas 2006]COVER, T. M.; THOMAS, J. A. *Elements of information theory*. Second. [S.l.: s.n.], 2006. 13–18 p.

[Ferrell 2009]FERRELL, J. E. A: systems biology. *J. Biol.*, v. 8, p. 2, 2009.

[Ge et al. 2001]GE, H. et al. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat. Genet.*, v. 29, p. 482–486, Dec 2001.

[Germek]GERMEK, G. R. Caracterização topológica dos motivos das Redes Biológicas Integradas da *Escherichia coli* e da *Saccharomyces cerevisiae*.

[Gulbahce e Lehmann 2008]GULBAHCE, N.; LEHMANN, S. The art of community detection. *Bioessays*, v. 30, p. 934–938, Oct 2008.

[Hartwell et al. 1999]HARTWELL, L. H. et al. From molecular to modular cell biology. *Nature*, v. 402, p. 47–52, Dec 1999.

[Joyce e Palsson 2006]JOYCE, A. R.; PALSSON, B. . The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell Biol.*, v. 7, p. 198–210, Mar 2006.

[Kitano 2002]KITANO, H. Systems biology: a brief overview. *Science*, v. 295, p. 1662–1664, Mar 2002.

[Kitano 2007]KITANO, H. A robustness-based approach to systems-oriented drug design. *Nat Rev Drug Discov*, v. 6, p. 202–210, Mar 2007.

[Kumpula et al. 2008]KUMPULA, J. M. et al. Sequential algorithm for fast clique percolation. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 78, p. 026109, Aug 2008.

[Newman 2006]NEWMAN, M. E. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.*, v. 103, p. 8577–8582, Jun 2006.

[Palla et al. 2005]PALLA, G. et al. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, v. 435, p. 814–818, Jun 2005.

[Pinkert, Schultz e Reichardt 2008]PINKERT, S.; SCHULTZ, J.; REICHARDT, J. Protein interaction network - more than mere modules. Dec 2008.

[Shannon 1948]SHANNON, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, 1948.