Ministério da Educação
Universidade Tecnológica Federal do Paraná
Pró-Reitoria de Pesquisa e Pós-Graduação

Relatório Final de Atividades

Caracterização de redes gênicas utilizando medidas de redes complexas

vinculado ao projeto
Integração de métodos de identificação de redes gênicas (GRNs) a partir de dados de expressão

Jonathan S. de S. R. da Silva
Bolsista Fundação Araucária
Engenharia de Computação
Data de ingresso no programa: 08/2011
Prof. Dr. Fabrício M. Lopes

Área do Conhecimento: Ciências Exatas e da Terra

CAMPUS Cornélio Procópio, 2012

JONATHAN S. DE S. R. DA SILVA
FABRÍCIO M. LOPES

CARACTERIZAÇÃO DE REDES GÊNICAS UTILIZANDO MEDIDAS DE REDES COMPLEXAS

Relatório Pesquisa do Programa de Iniciação Científica da Universidade Tecnológica Federal do Paraná.

CORNÉLIO PROCÓPIO, 2012

SUMÁRIO

INTRODUÇÃO

Existe ainda muito a ser descoberto sobre as relações funcionais dos mecanismos de controle, níveis de transcrição e proteínas, no sistema regulatório dos organismos. Um caminho que pode levar a um melhor entendimento desses mecanismos de controle regulatório é considerar a evolução temporal dos níveis de expressão gênica.

Sobre os perfis de expressão gênica, a motivação da pesquisa em inferência de GNs é o grande número de variáveis (genes) quando comparados com poucos experimentos disponíveis. Portanto recuperar uma GN representa um grande desafio de pesquisa com significativa importância em bioinformática [1].

A descoberta dessas funções de regulação pode levar à caracterização de uma diversidade de funções biológicas e também à descoberta da dinâmica das atividades moleculares. É de suma importância entender como muitos dos processos biológicos ocorrem e como prevenir que eles aconteçam, como é o caso das doenças.

A inferência, também denominada de engenharia reversa, de GNs (*Gene Networks)* a partir de perfis de expressão é fundamentada no dogma central da biologia molecular, no qual existe a premissa de que o estado funcional de um organismo é amplamente determinado pela sua expressão gênica. Então, se a variação dos níveis de expressão ao longo do tempo for mapeada por meio de GNs, é possível indicar informações como: diferentes vias regulatórias, ciclo celular, mapeamento de alterações provocadas por estímulos e como modelo de representação da atividade molecular.

Historicamente, o estudo de redes tem sido principalmente no domínio de um ramo da matemática discreta conhecido como teoria dos grafos. Desde o seu nascimento em 1736, quando o matemático suíço Leonhard Euler publicou a solução para o Königsberg (problema da ponte que consiste em encontrar um caminho de ida e volta que percorre cada uma das pontes da cidade prussiana de Königsberg exatamente uma vez), desde então a teoria dos grafos passou por muitos desenvolvimentos interessantes e providenciou respostas para uma série de questões práticas tais como: qual a maneira de colorir regiões de um mapa usando o número mínimo de cores de modo que regiões vizinhas recebam cores diferentes, ou como encaixar empregos para pessoas com utilidade total máxima, entre outras. Além dos desenvolvimentos matemáticos na teoria dos grafos, o estudo das redes também passou por evolução em alguns contextos especializados, como por exemplo, nas ciências sociais [3].

Este novo interesse pelo uso de redes em outros domínios pode ser atribuído a dois artigos seminais, por Watts e Strogatz [11] para redes *small-world* e por Barabási e Albert [10] em redes livres de escala (*scale-free*), as quais têm sido induzidas por sua relevância em várias áreas do conhecimento e pela possibilidade de se estudar as propriedades de uma variedade de grandes bancos de dados contendo redes reais. Estes incluem redes de transporte, redes de chamadas de telefone, da Internet e da World Wide Web, rede dos atores que possuem colaboração em bases de dados de cinema, coautoria científica e também aos sistemas de interesse em biologia e medicina, como redes neurais, genéticas, metabólicas e redes de proteínas.

A análise desse volume massivo de dados e a possibilidade de comparação das redes envolvendo campos diferentes, produziu uma série de situações inesperadas e resultados que impulsionaram ainda mais o crescimento da pesquisa na inferência de redes. A primeira questão que tem sido enfrentada é certamente estrutural. A pesquisa em redes complexas começou com o esforço de definição de novos conceitos e medidas para caracterizar a topologia de redes reais. O principal resultado foi a identificação de uma série de princípios unificadores e propriedades estatísticas comuns à maioria das redes [2].

Uma propriedade relevante é o grau de um nó, que é o número das suas ligações diretas aos outros nós. Em redes reais, o grau de distribuição P(k), definido como a probabilidade de que um nó escolhido de forma uniforme ao acaso tem grau k ou, equivalentemente, como a fração de nós no gráfico tendo grau k, significativamente desvia da distribuição Poisson esperada para um gráfico de forma aleatória e, em muitos casos, exibe uma lei de potência com um expoente, tendo um valor entre 2 e 3 [2].

Além disso, as redes *small-world* são caracterizadas por correlações nos graus dos nós, por ter em caminhos relativamente curtos entre quaisquer dois nós (principal propriedade da rede *small-world*) e, pela presença de um grande número de ciclos curtos.

Estas descobertas iniciaram a retomada da modelagem de redes, uma vez que os modelos matemáticos propostos na teoria dos grafos acabaram se distanciando das necessidades reais. Os cientistas tiveram a tarefa de desenvolver novos modelos para modelar o crescimento de uma rede e para reproduzir as propriedades estruturais observadas, em tempo real. A estrutura de uma rede real é o resultado da evolução contínua das forças que a formaram, e, certamente, afeta a função do sistema.

Este trabalho foi motivado pela expectativa de que a compreensão e modelagem da estrutura de uma rede complexa levariam a um melhor conhecimento dos seus mecanismos de evolução e, para uma melhor previsão no seu comportamento dinâmico e funcional.

Mais especificamente, é esperado que a arquitetura das GNs tenham consequências importantes na rede funcional, robustez e resposta a perturbações externas, como falhas aleatórias, ou ataques direcionados. Ao mesmo tempo surge também a possibilidade de estudar o comportamento dinâmico de conjuntos grandes de sistemas dinâmicos interagindo através de topologias complexas, como os observados empiricamente. Isso levou a uma série de evidências apontando para o papel crucial desempenhado pela topologia de rede na determinação do comportamento dinâmico de emergência coletiva, tais como a sincronização, ou no governo as principais características dos processos relevantes que ocorrem em redes complexas, tais como a propagação de epidemias, informação e boatos.

No campo de implementação, Kauffman [12] foi pioneiro em representar matematicamente as redes gênicas. Ele propôs programar genes binários e funções booleanas para descrever o comportamento das redes gênicas. E este padrão de determinar conexão ou não entre dois vértices sobre algum estado específico é usado no modelo de GNs adotado neste trabalho [13], definindo diferentes topologias e características de redes.

O estudo de redes complexas desenvolvido até o presente momento encontrou diferentes topologias, que são definidas pela maneira de como os nós se conectam e também por sua dinâmica. Dentre as topologias consideradas neste trabalho estão as topologias *scale-free* [10], *small-world* [11] e o aleatória [14].

Em resumo, o objetivo deste trabalho é quantificar a semelhança entre as topologias de uma rede inferida e uma rede já conhecida, em termos de medidas de redes complexas. A análise das redes inferidas é uma etapa essencial na pesquisa dos métodos de inferência. Espera-se que este estudo seja fundamental na descoberta de informações e o desenvolvimento de estruturas novas como métodos para inferência de redes de regulação gênica (GRNs).

MATERIAIS E MÉTODOS

Rede *scale-free* de Barabási e Albert [10]. Uma rede é denominada *scale-free* se a sua distribuição de grau, isto é, a probabilidade de que um nó selecionado ao acaso tenha certo número de ligações (grau), seguir uma função matemática chamada lei de potência. A lei de potência é definida pela curva característica de probabilidade $P(k) \sim k^{-\gamma}$, com expoentes variando entre 2 e 3. Em uma rede *scale-free* com um distribuição livre de escala, alguns vértices possuem um grau na ordem de magnitude muito maior do que a média, esses vértices são frequentemente chamados de *hubs*, enquanto que a grande maioria dos vértices possui poucas conexões [3].

Algumas redes com uma distribuição dos graus dos nós que segue uma lei de potência (e específicos de outros tipos de estrutura) podem ser altamente resistentes à supressão aleatória de vértices, isto é, a grande maioria dos vértices permanece ligada em um componente conexo. Essas redes também podem ser bastante sensíveis a ataques direcionados objetivo de romper a rede rapidamente.

Grau de distribuição da scale-free. Um caso comum na ciência até poucos anos atrás era o de redes homogêneas. Homogeneidade na estrutura de interação significa que quase todos nós são topologicamente equivalentes, como em "regular lattices" ou em grafos aleatórios. Nesses últimos, por exemplo, cada N (N - 1) /2 conexões possíveis se apresenta com probabilidade igual, e deste modo o grau de distribuição é binomial ou Poisson [2] no limite de um grafo grande. Não é surpreendente então quando os cientistas se aproximaram do estudo de redes reais ao banco de dados disponíveis, foi considerado razoável encontrar distribuições localizadas em torno de um valor médio. Em contraste com todas as expectativas, foi encontrado que a maioria das redes reais apresenta distribuição de grau na forma lei de potência (*power law*) $P(k) \sim k^{-\gamma}$, com expoentes variando entre 2 e 3 [2].

Redes aleatórias de Erdös e Rényi [14]. O grafo aleatório desenvolvido por *Rapoport* e após por *Erdös e Rényi* pode ser considerado o modelo mais simples de redes complexas. No artigo de *Erdös e Rényi* em 1959 [14], foi desenvolvido um modelo para gerar grafos aleatórios consistindo em N vértices e M arestas. Começando com N vértices desconectados, a rede é construída pela adição aleatória de M arestas, evitando múltiplas conexões e laços [3].

Outro modelo similar define N vértices e uma probabilidade p de conexão entre cada par de vértices. Esse modelo é amplamente conhecido como modelo de *Erdös-Rényi* (ER). Para o modelo ER, em uma rede de grande tamanho com $N \to \infty$, o número médio de conexões em cada vértice (k) pode ser obtido pela Equação 1.

$$(k) = p\,(N-1)$$

Equação 1: grau 'k' de conectividade média.

Em vez disso, p é escolhido como uma função de N para manter (k) constante:

$$p = \frac{k}{N-1}$$

Equação 2: probabilidade de um nó estar conectado.

Para este modelo, o grau de distribuição P(k) é uma distribuição de *Poisson* [3].

Redes *small-world* [11]. O estudo de vários processos dinâmicos sobre redes reais tem apontado para a existência de atalhos, conexões que ligam diferentes áreas das redes, acelerando assim a comunicação de nós outrora distantes.

No "hypercubic lattices" [15], o número médio de vértices que têm de passar para alcançar um nó escolhido aleatoriamente, assume a forma de $N^{1/d}$ (sendo N o número de nós e d a dimensão no espaço cartesiano, o hipercubo tem dimensão 4). Reciprocamente, na maioria de redes reais, apesar de possuírem frequentemente um grande tamanho, há relativamente um caminho curto entre quaisquer dois nós. Essa característica é conhecida como "small-world property", e é matematicamente escrita por um caminho curto médio de tamanho L, como definido na Equação 3.

$$L = \frac{1}{N(N-1)} \sum_{i,j \in N, i \neq j} d_{ij}$$

Equação 3: caminho curto médio.

Essa propriedade foi investigada pela primeira vez, em um contexto social, por Milgram [16] na década de 60 em uma série de experimentos para estimar o real número de passos em uma sequência de conhecidos.

Nesse primeiro experimento, Milgram pediu a pessoas selecionadas aleatoriamente em Nebraska que mandassem cartas a um indivíduo alvo em Boston, identificado somente pelo seu nome, profissão e localização superficial. As cartas só poderiam ser mandadas para alguém que o remetente conhecesse o primeiro nome, e que estava presumivelmente próximo ao destinatário final. Milgram rastreou a trajetória das cartas e as características demográficas de seus manipuladores. Embora a hipótese comum fosse de que as cartas pudessem tomar centenas de caminhos para alcançar seu destino final, o surpreendente resultado de Milgram foi que o número de conexões necessárias para alcançar a pessoa alvo teve um valor médio de apenas 6. Mais recentemente um experimento semelhante realizado por Dodds com trocas de e-mails reproduziu com sucesso o experimento de Milgram. As trocas de mensagens, de fato, completaram as sequências suficientes para permitir caracterizá-las estatisticamente segundo a propriedade da rede *small-world*.

A propriedade da *small-world* foi observada em um gama de redes reais, incluindo biológicas e tecnológicas, e é uma óbvia propriedade matemática em alguns modelos de rede, como por exemplo, em grafos aleatórios. Em divergência com grafos aleatórios, a propriedade da *small-world* em redes reais é frequentemente associada com a presença de agrupamentos, denotando altos valores de coeficiente de agrupamento [2].

$$C = \langle c \rangle = \frac{1}{N} \sum_{i \in N} c_i$$

Equação 4: coeficiente de agrupamento.

Por esta razão, Watts e Strogatz [11], propuseram definir redes small-world como redes que possuem um pequeno valor de L, como grafos aleatórios, e um alto coeficiente de agrupamento C (Equação 4). Além disso, possuem alto valor de eficiência global (Equação 5) e local (Equação 6), as quais passam a ser redes extremamente eficientes em trocas de informações em escalas global e local.

$$E = \frac{1}{N(N-1)} \sum_{i,j \in N, i \neq j} \frac{1}{d_{ij}}$$

Equação 5: Eficiência global da rede small-world.

$$E(local) = \frac{1}{N} \sum_{i,j \in N, i \neq j} E(G_i)$$

Equação 6: Eficiência local da rede small-world.

A maior parte do interesse no assunto ultimamente tem mudado para investigar o comportamento dinâmico das redes, com ênfase especial sobre como a estrutura da rede afeta as propriedades de sua dinâmica. Um exemplo em destaque é estudar o aparecimento de dinâmicas sincronizadas em redes complexas, a partir do ponto de vista da relação entre a propensão de sincronização de uma rede para a interação entre topologia e propriedades locais dos sistemas dinâmicos. Esse fenômeno, na verdade, representa uma característica crucial em muitas circunstâncias relevantes. Por exemplo, existe evidência de que algumas doenças cerebrais são causadas por uma anormal e, algumas vezes, sincronização abrupta de um grande número de populações neuronais, de modo que a investigação sobre os mecanismos de rede envolvidos na geração, manutenção e propagação dos distúrbios epilépticos é uma questão atual na vanguarda da neurociência.

Estruturas comunitárias é uma importante propriedade das redes complexas. Por exemplo, grupos intimamente ligados de nós em uma rede social representam indivíduos pertencentes a comunidades sociais, grupos fortemente conectados de nós em World Wide Web, muitas vezes correspondem a páginas sobre temas comuns, enquanto as comunidades em redes celulares e genéticas são de alguma forma relacionadas com módulos funcionais.

Consequentemente, encontrando as comunidades dentro de uma rede é uma poderosa ferramenta para a compreensão do funcionamento da rede, bem como para a identificação de uma hierarquia de ligações dentro de uma arquitetura complexa.

Medidas de redes complexas. Extrair características de diferentes redes topológicas é uma tarefa que exige direção de experientes pesquisadores em todo o processo. Caracteriza-se por ser uma tarefa em expansão embora ainda se encontrem poucos guias de como usar as funções para extração de medidas nos diferentes ambientes.

Centralidade alfa ou Centralidade de autovetor. A medida centralidade alfa pode ser considerada uma generalização de centralidade do autovetor para grafos direcionados. Foi proposto por Bonacich em 2001 [17]. Mede a importância de um nó em uma rede. Atribui valores aos nós com base no princípio: “Conexões com nós que tem um valor alto contribuem mais para o valor do nó do que outras conexões com nós que tem valor baixo”. O PageRank (usado para priorizar as páginas) do Google é uma variação de centralidade de autovalor. O papel central alfa dos vértices de um grafo é definida como a solução da seguinte equação: $x = x \,.\, \alpha \,.\, A^T + e$, na qual A é a matriz de adjacência (não necessariamente simétrica) do gráfico, 'e' é o vetor de fontes exógenas de estado dos vértices e 'α' é a importância relativa dos endógenos versus fatores exógenos.

Pontos de articulação. Pontos de articulação ou vértices de corte são vértices cuja remoção aumenta o número de componentes conectados em um grafo.

Centralidade de intermediação. Intermediação é uma medida de centralidade de um nó dentro da rede. Vértices que aparecem em muitos caminhos mínimos entre outros vértices apresentam altos valores de intermediação. A centralidade de intermediação de um vértice 'v' é definida se obtendo o caminho curto de todos os pares de vértice (o vértice de origem deve ser diferente do vértice de chegada) no grafo não incluindo o vértice 'v', então se faz a razão do número de caminhos curtos que passam por ele pelo número total de caminhos curtos. Analogamente é calculado para as arestas.

Componentes biconectados. Um grafo é biconectado se a remoção de qualquer único vértice (com suas arestas adjacentes) ainda o manterem conexo. Um componente biconectado de um grafo é um subgrafo biconectado máximo do mesmo. Os componentes biconectados de um grafo podem ser dados pela partição de suas arestas: cada aresta é um membro de exatamente um componente biconectado. Note-se que isso não é verdade para os vértices: o mesmo vértice pode ser parte de muitos componentes biconectados.

Centralidade Bonacich de Potência. A centralidade Bonacich de potência é definida pelo CBP (α, β) = α (I - βA) A1, onde β é um parâmetro de atenuação (definido na função utilizada no iGraph por expoente) e A é a matriz de adjacência do grafo. O coeficiente α atua como um parâmetro de escalonamento, e é definido aqui (P. Bonacich, 1987) como a soma das raízes das medidas é igual ao número de vértices. Isto permite que 1 possa ser utilizado como um valor de referência para o meio da faixa de centralidade. Quando β → 1/λA1 (o inverso da maior autovalor de A), ele é um múltiplo constante da medida padrão de centralidade do autovetor; para outros valores de β, o comportamento da medida é bastante diferente. Em particular, β dá pesos positivos e negativos para passeios pares e ímpares, respectivamente, como pode ser visto a partir da expansão da série descrita na Equação 7.

$$C_{BP}(\alpha, \beta) = \alpha \sum_{k=0}^{\infty} \beta^k \mathbf{A}^{k+1}\mathbf{1}$$

Equação 7: Série constituinte da Centralidade Bonacich de Potência.

A Equação 7 converge depois de muito tempo a | β | < 1/λA1. A magnitude de β controla a influência de atores distantes na pontuação ego centralidade, com magnitudes maiores, indicando taxas mais lentas de decadência. Taxas elevadas, portanto, implicam uma maior sensibilidade aos efeitos das arestas.

Interpretativamente, a medida Bonacich de potência corresponde à noção de que o poder de um vértice é recursivamente definido pela soma da potência dos seus vizinhos. A natureza da recursão envolvida é então controlada pelo expoente de poder: valores positivos significam que vértices se tornam mais poderosos como seus vizinhos mais poderosos (como ocorre nas relações de cooperação), enquanto valores negativos implicam que os vértices se tornam mais poderosos apenas com seus vizinhos sendo mais fracos (como ocorre no estilo competitivo ou em relações antagônicas). A magnitude do expoente indica a tendência do efeito de decaírem; magnitudes mais altas implicam uma queda mais lenta. Uma característica interessante desta medida é a sua instabilidade em relação às mudanças de magnitude expoente (em particular no caso negativo).

Cliques. Um clique em um grafo G é um subgrafo de G que é completo. O número de cliques de um grafo é a quantidade de vértices envolvendo o maior (es) clique(s).

Centralidade de proximidade. A centralidade de proximidade mede a distância média entre um nó a todos os outros vértices da rede (que ele pode alcançar). Usada como uma medida do tempo que leva para a informação (ou vírus) se espalhar de um nó até todos os nós que ele alcança na rede. A centralidade de proximidade de um vértice é definida pelo inverso do comprimento médio do mais curto caminho para todos os outros vértices do grafo, como definido pela Equação 8.

$$\frac{1}{\sum_{i \neq v} d_v i}$$

Equação 8: Cálculo da centralidade de proximidade para cada nó.

Se não houver um caminho (dirigido) entre o vértice 'v' e 'i', em seguida, o número total de vértices é usado na fórmula em vez de o comprimento do percurso. Há como calcular a centralidade de proximidade dos caminhos curtos que saem do vértice 'v' e que chegam, assim como calcular a centralidade de ambas na mesma medida.

A centralidade de proximidade tem aplicação direta em bioinformática determinando vias metabólicas rápidas de acordo com a pontuação de cada gene em relação aos seus vizinhos no quesito de velocidade do fluxo de qualquer fluído ou impulsos elétricos.

Agrupamentos. Esta medida extrai o número de agrupamentos em cada grafo, em uma rede há vértices que são mais conectados entre si e acabam por formar vários aglomerados de arestas cada um tendo pouca conexão com os agrupamentos adjacentes.

Coligação e Acoplamento bibliográfico. Dois vértices são coligados se houver outro vértice conectando os dois. A instrução 'cocitation' simplesmente conta quantos tipos de dois vértices são coligados. O acoplamento bibliográfico de dois vértices é o número de outros vértices em que ambos se ligam, a instrução 'bibcoupling' é responsável pelo cômputo desse valor.

Grau. O grau de um vértice é sua mais simples propriedade: representa o número de arestas adjacentes a ele.

Diâmetro. O diâmetro de um grafo é a excentricidade máxima de qualquer vértice do grafo. Ou seja, ele é a maior distância entre qualquer par de vértices. Para achar o diâmetro de um grafo encontre o caminho mínimo entre cada par de vértices. O maior comprimento de qualquer um desses caminhos é o diâmetro do grafo.

Díade dos nós. Existem três classes de ligação entre os vértices: ligação mútua, cada um dos nós possui uma aresta de origem e destino em relação ao outro nó; ligação assimétrica, o nó que é origem na aresta não recebe outra aresta como destino também; não existente, inexistem arestas entre os vértices analisados.

Conectividade de aresta e Adesão do grafo. A conectividade de aresta de um par de vértices é o número mínimo de arestas necessário para remover a eliminar todos os caminhos da origem para o destino. A adesão de um grafo é o número mínimo de arestas necessárias para remover a obter um grafo que não é fortemente ligado.

Circunferência. A circunferência de um grafo é o tamanho do menor círculo nele. Para o cálculo da circunferência laços e arestas múltiplas são ignorados. Caso o grafo seja uma floresta a circunferência é zero, pois ele é acíclico.

Transitividade. A transitividade mede a probabilidade de que os vértices adjacentes de um vértice estão ligados. Isto é às vezes também chamado de coeficiente de clusterização. O coeficiente de clusterização é dado pela razão: 3 vezes o número de triângulos pelo número de tripletes (conjunto de 3 vértices centrado em 1 dos vértices. A clusterização de um vértice é o número de arestas entre seus vizinhos dividido pelo número total de vértices possíveis entre eles. O coeficiente de clusterização é calculado como a média de clusterização entre os vértices do grafo. Se um determinado grafo representa uma rede social cujas relações entre os indivíduos sejam de amizade, é natural esperar um coeficiente de clusterização alto, uma vez que é provável que os amigos de alguém sejam amigos entre si.

Semelhança ponderada log-inversa. A semelhança ponderada log-inversa de dois vértices é o número de seus vizinhos comuns, ponderada pelo logaritmo inversa dos seus graus. Baseia-se no pressuposto de que dois vértices devem ser considerados mais semelhantes se eles partilharem um vizinho de baixo grau comum, vizinhos comuns por um vértice de alto grau são mais propensos a aparecer mesmo por puro acaso. Vértices isolados vão ter zero semelhança com qualquer outro vértice. Auto-semelhanças não são calculadas.

$$similarity(A,B) = \sum_{shareditems} \frac{1}{\log[frequency(shareditem)]}$$

Equação 9: Cálculo da semelhança ponderada log-inversa.

Softwares de medidas de redes complexas. Com o objetivo de analisar diferentes características em redes topológicas escolhidas para serem estudadas foi escolhido a linguagem de programação R [18] devido ao pacote iGraph [6] que possui uma implementação robusta e com grande quantidade de referências bibliográficas para as medidas dos grafos.

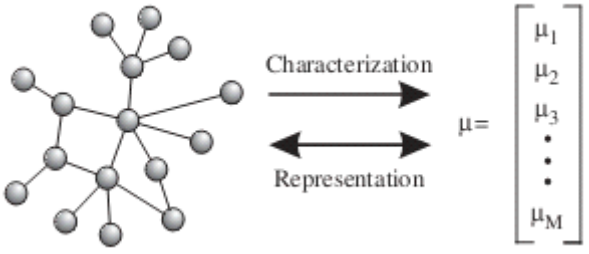

Figura 1: Representação das redes e extração do vetor de características. Extraída de [3].

O pacote iGraph [6] contém funções para geração de gráficos regulares e aleatória, manipulação de gráficos, atribuindo atributos para vértices e arestas. Pode calcular várias propriedades estruturais, o isomorfismo gráfico, inclui heurísticas para detecção de estrutura da comunidade.

Também foram analisadas outras ferramentas para obtenção das medidas de rede complexas. Inicialmente a foi utilizado o pacote NetworkX [9] da linguagem de programação Python, este pacote possui uma série de funções sobre grafos. Estas funções que criam, modelam, extraem características de grafos são de extrema

importância para o estudo das redes gênicas, pois os grafos são os modelos das redes gênicas: as diferentes redes topológicas são representadas de forma ideal pelos diversos modos de formação de um grafo. O uso do pacote NetworkX não obteve sucesso porque apesar da grande quantidade de funções, as funções para extrair características dos grafos, que é a meta de estudo, não eram em boa quantidade e não funcionavam simultaneamente para todas as redes topológicas: rede aleatória, rede livre de escala, rede *small-world*.

A segunda tentativa de uma biblioteca que analisasse de forma eficiente os grafos para inferir GNs foi a API Jung [8] desenvolvido em linguagem Java. A API Jung possui um robusto desempenho e inúmeras utilidades inclusive para visualização de grafos em três dimensões. Mas como se trata de uma API nova ainda não possui tutoriais de como utilizar os métodos disponibilizados, e as poucas funções que conseguiram ser utilizadas por seus parâmetros serem descobertos não foram suficientes porque para distinguir as redes topológicas.

Uma abordagem mais eficiente sobre o estudo das características aconteceu quando foram testadas as funções da biblioteca iGraph [6] no programa estatístico R [18], o qual foi adotado para o desenvolvimento deste trabalho. Além disso, a abundância de características que poderiam ser analisadas proporcionou um estudo muito mais abrangente aumentando as chances de serem encontradas medidas de redes que diferenciam topologias e assim serviriam para guiar o processo de inferência de redes.

RESULTADOS E DISCUSSÕES

Os resultados da extração de medidas de GNs começaram a ser obtidos quando o pacote iGraph [6] da linguagem estatística R foi estudado, com muitos tutoriais e exemplos das funções que foram utilizadas para inserção de dados no vetor de características para cada rede topológica, variando o número de nós (genes) e o grau médio de cada nó. Como resultado da extração de características foi gerada uma matriz, e em cada característica foram feitas duas análises: a primeira buscava diferenciar as redes Erdös-Rényi (ER), Barabási-Albert (BA) e Watts-Strogatz (WS) em cada característica; por sua vez, a segunda calculava a correlação de uma característica variando não só as topologias de rede, mas também o tamanho (número de nós). As figuras a seguir exibem algumas características estruturais que apresentaram baixa correlação, i.e., características não correlacionadas que podem ser utilizadas para identificar as topologias consideradas.

CORRELAÇÃO – mean_degree			
	ER	BA	WS
ER	1	0,3693	0,1173
BA		1	-0,0334
WS			1

Figura 2: Correlação entre as topologias quanto ao grau.

CORRELAÇÃO – mean_graph_coreness			
	ER	BA	WS
ER	1	0,6408	0,3170
BA		1	-0,0721
WS			1

Figura 3: Correlação entre as topologias referente ao número de subgrafos maximais de cada nó.

CORRELAÇÃO – mean_neighborhood			
	ER	BA	WS
ER	1	-0,0300	0,0481
BA		1	0,0111
WS			1

Figura 4: Correlação entre as topologias quanto à vizinhança média.

É importante destacar na Figura 2 que o grau médio (*mean degree*) apresentou baixa correlação (-0,0334) entre as topologias BA e WS. Outra medida que se mostrou relevante, exibida na Figura 3, foi o número de subgrafos maximais (*mean graph coreness*), o qual apresentou baixa correlação (-0,0721) entre as topologias BA e WS.

Além disso, uma medida apresentou baixa correlação entre as três topologias de redes complexas consideradas neste trabalho. A vizinhança média (*mean neighborhood*), exibida na Figura 4, apresentou baixa correlação (-0,03) entre as topologias BA e ER, baixa correlação (0,0481) entre as redes ER e WS e também baixa correlação (0,0111) entre as redes BA e WS, se indicando a possibilidade de ser usada como medida de classificação entre as diferentes topologias consideradas.

Como uma forma de investigar como as medidas de redes complexas poderiam se comportar em uma análise envolvendo não só pares de características, mas também combinações envolvendo um subconjunto de características aplicado na identificação/classificação das diferentes topologias, foi adotado o software DimReduction [5]. O resultado pode ser observado na Figura 5, o qual apresenta a melhor solução encontrada e respectivo espaço de características. Foi identificado que as medidas log-inversa ponderada e transitividade formaram a dupla que apresentou melhor separação entre as topologias (classes) analisadas.

Tais medidas de redes complexas são boas separadoras entre as topologias de rede e devem guiar de maneira mais precisa o processo de inferência de redes como fundamenta em [19].

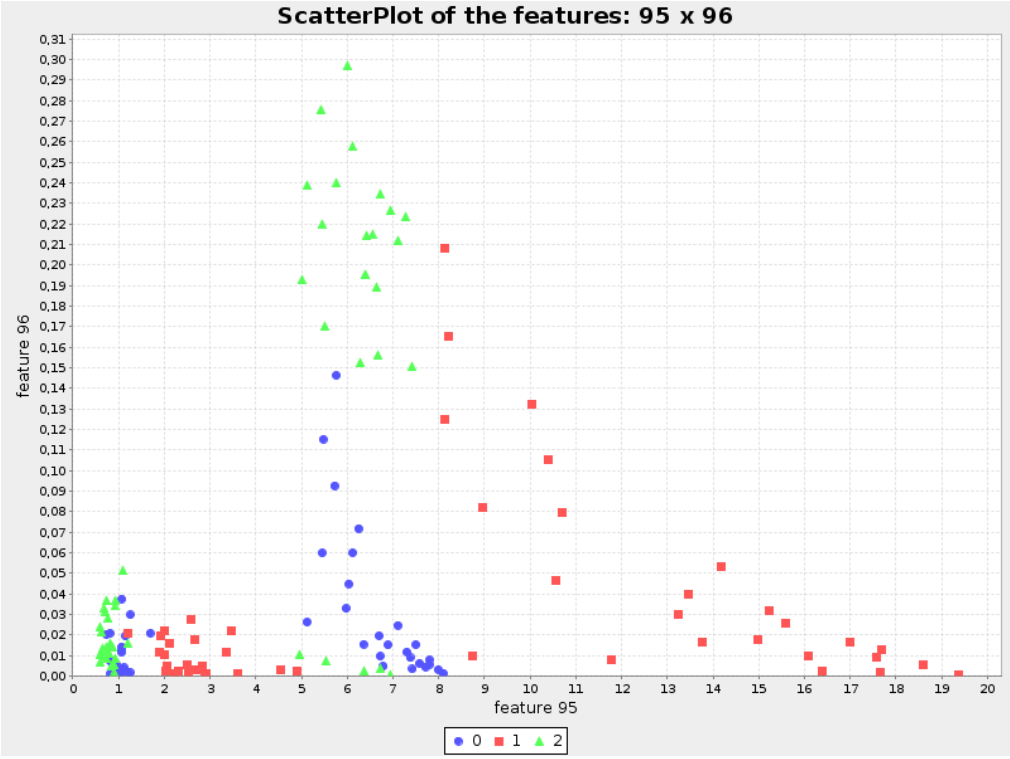

Figura 5: Separação das topologias quanto às medidas de similaridade, considerando as medidas log-inversa ponderada (feature 95) e transitividade (feature 96). As classes são as topologias de redes (0) Aleatória - ER, (1) *Scale-Free* - BA e (2) *Scale-free* - WS.

CONCLUSÕES

Este trabalho trata do estudo das topologias de redes complexas e suas respectivas caracterizações por meio de suas medidas. Foram analisadas as medidas de redes complexas considerando três topologias, aleatórias, livre de escala, pequeno-mundo, com diferentes tamanhos de redes e graus de conectividade, as quais foram geradas a partir de um modelo de redes gênicas artificiais [13].

Foi identificado que as medidas grau médio (*mean degree*), número de subgrafos maximais (*mean graph coreness*) e especialmente a vizinhança média (*mean neighborhood*) apresentaram coeficientes de correlação próximos de zero, representando que essas medidas são boas candidatas para a classificação das diferentes topologias analisadas. Enquanto que as demais medidas se apresentaram correlacionadas entre si e, portanto, não são adequadas para a identificação/classificação entre as topologias de redes analisadas neste trabalho.

Como possibilidade de trabalho futuro, pretende-se que as medidas de redes complexas que se apresentaram com coeficientes de correlação próximos a zero sejam utilizadas para melhorar a inferência das redes gênicas por meio de sua inclusão como informação *à priori* na busca por redes com topologias previamente conhecidas.

Também é esperado que um novo algoritmo de inferência de GNs possa ser criado a partir do uso das informações levantadas neste trabalho e, integrado ao software DimReduction [5].

Um aspecto que pode ser mencionado no desenvolvimento deste trabalho é que foram encontradas dificuldades já que o tópico abordado aqui é novo e não há muitas plataformas que contenham as funções necessárias para a realização da pesquisa, embora a bibliografia forneça ótimas definição e abordagens sobre as diversas características das GNs [1-3].

REFERÊNCIAS

[1] Lopes, F. M. Redes complexas de expressão gênica : síntese, identificação, análise e aplicações. São Paulo : Bioinformática, Universidade de São Paulo, 2011. Tese de Doutorado em Bioinformática. Disponível em: http://www.teses.usp.br/teses/disponiveis/95/95131/tde-27072011-105810/.

[2] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. U. Complex networks: Structure and dynamics. Physics Reports, 424(4-5):175-308, 2006.

[3] Costa, L. d. F., Rodrigues, F. A., Travieso, G., and Villas-Boas, P. R. Characterization of complex networks: a survey of measurements. Advances in Physics, 56(1):167-242, 2006.

[4] Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., Bagos, P. G. Using graph theory to analyze biological networks. BioData Mining 28(4):10, 2011.

[5] Lopes, F. M., Martins-Jr, D. C. e Cesar-Jr, R. M. Feature selection environment for genomic applications. BMC Bioinformatics, 9(1):451, 2008.

[6] iGraph library. Disponível em: http://igraph.sourceforge.net/.

[7] Langfelder, P. e Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9(1):559, 2008.

[8] JUNG - Java Universal Network/Graph Framework. Disponível em: http://jung.sourceforge.net.

[9] NetworkX. Disponível em: http://networkx.lanl.gov/index.html.

[10] Barabási, A. L. and Albert, R. Emergence of scaling in random networks. Science, 286(5439):509-512, 1999.

[11] Watts, D. J. and Strogatz, S. H. Collective dynamics of smallworld networks. Nature, 393:440-442, 1998.

[12] Kauffman, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology, 22(3):437-467, 1969.

[13] Lopes, F. M., Cesar-Jr, R. M., and Costa, L. d. F. Gene expression complex networks: synthesis, identification and analysis. Journal of Computational Biology, 18(10): 1353-1367, 2011.

[14] Erdös, P. and Rényi, A. On random graphs. Publ. Math. Debrecen, 6:290-297, 1959.

[15] Lattice Geometries. Disponível em: http://www.hermetic.ch/compsci/lattgeom.htm.

[16] Milgram, S. The small world problem. Psychology today, 2(1):60-67, 1967.

[17] Bonacich, P. e Paulette, L. Eigenvector-like measures of centrality for asymmetric relations. Social Networks, v. 23, p. 191-201, 2001.

[18] The R Project for Statistical Computing. Disponível em: http://www.r-project.org/.

[19] Lopes, F. M., Martins-Jr, D. C., Barrera, J., and Cesar-Jr, R. M. An iterative feature selection method for GRNs inference by exploring its topology properties. Arxiv preprint arXiv:1107.5000, abs/1107.5000v1:1-10, 2011.

Fabrício M. Lopes

Jonathan S. de S. R. da Silva