BioNetStat: Uma ferramenta para análise diferencial de redes biológicas

Vinícius Jardim Carvalho

Dissertação apresentada
ao Instituto de matemática e estatística
da Universidade de São Paulo para a
obtenção do título de mestre em ciências.

Programa: Interunidades em Bioinformática

Orientador: Marcos Silveira Buckeridge

Coorientador: André Fuijta

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES.

São Paulo, novembro 2017

Agradecimentos

Sabemos que um trabalho científico nunca é feito sozinho. Portanto, é de extrema importância lembrar e agradecer a contribuição de todos aqueles que participaram, de diversas formas, na produção desse trabalho.

É primordial agradecer a importância da Universidade de São Paulo, mais especificamente do Instituto de Biociências e do Instituto de Matemática e Estatística no desenvolvimento do trabalho, pela formação de pessoal muito bem capacitado (professores, funcionários e alunos) e infraestrutura essenciais para o desenvolvimento científico. Quero agradecer aos funcionários da Universidade e desses institutos que fazem com que tudo ocorra dentro da normalidade. Especialmente gostaria de agradecer a Cris e a Patricia que sempre me ajudaram não só com as burocracias mas também foram amigas ao longo dos anos de mestrado.

Gostaria de agradecer também à CAPES pela concessão da bolsa, a FAPESP em projeto conjunto com a Microsoft Research (Processo nº 2011/52065-3) e ao FAPESP-BioEn (INCT do Bioetanol - Processo nº 2008/57908-6) pelo financiamento dos projetos que participei. Além disso, é importante agradecer ao grupo eScience do IME (apoio financeiro FAPESP processo #2011/59761-2, do CNPq, da CAPES e do NAP e Science – PRP – USP) pela rede de computadores usada na realização das execuções.

Agradeço enormemente ao meu orientador, Marcos Buckeridge, que foi essencial não só no desenvolvimento desse trabalho mas principalmente na minha formação como cientista. É em razão de muitas conversas que tivemos que fui apresentado à teoria de redes e hoje enxergo o mundo a minha volta pelas lentes da teoria de sistemas e pela redes.

Tão essencial quanto, foi também a co-orientação de André Fujita. Apreendi muito com ele, não só sobre teoria de redes e técnicas estatísticas, mas também no comportamento e atitudes em relação ao produção de ciência.

Quero fazer uma agradecimento especial a Suzana Siqueira Santos que foi a inspiração e base para o desenvolvimento deste projeto. Sua ajuda, paciência, dedicação e constante disponibilidade certamente garantiram que eu aprendesse muito e me empenhasse cada vez mais nesse trabalho.

É importante agradecer também ao LAFIECO e toda a equipe que nele está ou esteve presente.

Nele, conheci pessoas que me ajudaram e ajudam na minha formação como pesquisador e como pessoa. Quero agradecer em especial à Amanda Pereira de Souza, minha orientadora na Iniciação Científica e que me ensinou muito no final da graduação, Adriana Grandis, que me ajudou muito na iniciação e também no desenvolvimento deste trabalho, e Amanda Rusiska, que foi grande parceira da bioinformática ao longo desse mestrado. Agradeço muito também o valioso trabalho das técnicas Eglee e Viviane que fazem o LAFIECO sempre funcionar em perfeito estado e nos estimulam a trabalhar em um local prazeroso.

Outro grupo importante no desenvolvimento desse trabalho é o grupo de pesquisa do professor Fujita, onde aprendi muito sobre redes e técnicas de estatística. Quero agradecer especialmente a Adéle, que foi muito importante em conselhos e ajudas dadas neste trabalho.

Não posso deixar de agradecer às Professoras Julia Pavan e Marie Anne que, na banca de qualificação, fizeram importantes modificações e deram conselhos essenciais para a melhoria deste trabalho.

Gostaria de agradecer também aos meus amigos de graduação, da Rep Aligator, aos colegas do Joga no Pagode e do Bonde com os quais aprendi e vivi muitas histórias importantes até hoje.

Quero agradecer especialmente à Karine que tem sido uma grande amiga, companheira e amante em todos os momentos da minha vida, me inspirando sempre a fazer um trabalho cada vez melhor.

Por fim, não posso deixar de agradecer aos meus Avós, Terezinha, Enéas, Bela e Carlito que sempre me ensinaram a ter amor e respeito pelo próximo, aos meus tios e tias, à minha irmã que sempre foi uma companheira e amiga durante toda a minha vida, e principalmente aos meus pais Leonardo e Márcia que me ensinaram a viver a vida com alegria, otimismo e muita dedicação para enfrentar os desafios.

Este trabalho e eu não existiríamos se não tivéssemos a participação de todas essas pessoas e ainda de muitas outras que não foram citadas aqui.

Obrigado a todos.

Resumo

A diversidade de interações que ocorre dentro de sistemas biológicos, considerando desde as organelas de uma célula até toda a biosfera, pode ser modelada por meio da teoria de redes. A dinâmica das interações entre os elementos é uma propriedade intrínseca desses sistemas. Diversas ferramentas foram propostas para comparar redes, que representam os muitos estados assumidos por um sistema. Porém, nenhuma delas é capaz de comparar características estruturais de mais de duas redes simultaneamente. Devido à grande quantidade de estados que um sistema pode assumir, construímos uma ferramenta estatística para comparar duas ou mais redes e indicar variáveis chave no processo estudado. A principal proposta deste trabalho foi comparar redes de correlação usando medidas baseadas nos espectros dos grafos (conjunto de autovalores das matrizes de adjacência), como a distribuição espectral. Essa medida está associada a diversas características estruturais das redes como o número de caminhos, diâmetro e cliques. Além da distribuição espectral, também comparamos as redes por entropia espectral, distribuição dos graus e pelas centralidades dos nós. Usamos dois diferentes conjuntos de dados biológicos (expressão gênica de células tumorais e metabolismo vegetal) para realizar os testes de desempenho da ferramenta e para os estudos de caso. O método proposto está implementado em um pacote do programa R, chamado *BioNetStat*, com interface gráfica para o usuário leigo em programação. Constatamos que os testes são eficientes em diferenciar mais de duas redes. Além disso, o aumento do número de redes comparadas e a queda dos números de unidades amostrais, diminui o poder estatístico do teste. Mostramos ainda que ocorre uma economia de tempo significativa ao realizarmos uma única análise para comparar muitas redes ao invés de compará-las par-a-par. Além disso, o método apontou grupos de variáveis com papel central nos sistemas biológicos estudados que não foram encontrados nas análises onde apenas a expressão ou concentração dos elementos foi estudada. Foi possível assim diferenciar células de tipos cancerígenos ou órgãos de organismos vegetais através das centralidades das redes. As variáveis levantadas possibilitam ao usuário gerar hipóteses sobre seus papéis nos processos em estudo. O *BioNetStat* pode assim ajudar a detectar possíveis novas descobertas associadas a mecanismos de funcionamento de sistemas.

Palavras-chave: Análise de redes, Análise diferencial de redes, Redes de correlação, Redes de co-expressão, Biologia de sistemas, Teoria de redes.

Abstract

The diversity of interactions, which are among elements of the biological systems, can be studied based on the networks theory. Moreover, the dynamic of these interactions is an inherent trait of those systems. In this sense, several tools have been proposed to compare networks, in that each network represents a state assumed by the system. However, the biological systems generally can assume much more than two biological states and none of the tools are able to compare structural characteristics among more than two networks simultaneously. To solve this issue, we developed a statistical tool to compare two or more networks and highlight key variables of a system. Here we describe the new method, called *BioNetStat*, that is able to compare correlation networks using traits that are based on graph spectra (the group of eigenvalues of the adjacency matrix), such as the spectral distribution. This measure is associated with several structural characteristics of networks such as the number of walks, diameter, and cliques. In addition to the spectral distribution, *BioNetStat* can also compare networks to the node centralities. We used two different biological datasets, tumoral cells genes expressions and plant metabolism, to evaluate the performance of *BioNetStat* and as case studies. The tool is implemented in an R package, and it also has a user-friendly interface. We showed that *BioNetStat* is efficient in distinguishing more than two networks. In comparison with a similar tool (*GSCA*), the increase in the number of compared networks reduces less the statistical power of the *BioNetStat* than the *GSCA*. Furthermore, *BioNetStat* is able to find signaling pathways in a bigger proportion than the *GSCA*, complementing tools proposed in the literature. In the case studies, the method pointed out variables, and sets of variables, with a central role in biological systems, which were not highlighted when only gene expression pattern or metabolomics were studied. For instance, *BioNetStat* allowed us to differentiate among cancer types and plant organs. The *BioNetStat* results bring new findings on what differentiate the states, giving us a systemic view of our study subject and affording the proposition of new hypotheses about the studied processes.

Keywords: Network analysis, Differential network analysis, Coexpression network, Correlation network, Systems biology, Networks theory.

Sumário

- Agradecimentos.....	2
- Resumo.....	4
- Abstract.....	5
1 - Introdução.....	1
1.1 - Objetivo.....	11
1.1.1 - Metas.....	11
2 - Material e métodos.....	12
2.1 - Estruturas de redes complexas.....	12
2.1.1 - Notação.....	12
2.1.2 - Medidas de centralidade.....	12
2.1.3 - Medidas de segregação funcional.....	14
2.1.4 - Medidas de resiliência.....	15
2.1.5 - Medidas baseadas na caracterização espectral do grafo.....	16
2.1.5.a - Distribuição do espectro em modelos conhecidos.....	17
2.1.5.b - Entropia.....	19
2.1.5.c - Entropia de grafos.....	20
2.1.5.d - Divergência de Kullback-Leibler.....	23
2.1.5.e - Divergência de Jensen-Shannon.....	24
2.2 - Teste estatístico de comparação de múltiplos grafos.....	24
2.2.1 - Enunciado do problema.....	25
2.2.2 - Construção do grafo de correlação.....	25
2.2.3 - Testes estatísticos.....	29
2.2.4 - Estatística dos testes.....	29
2.2.4.a - Análise diferencial de redes.....	29
2.2.4.b - Análise diferencial dos vértices.....	31
2.2.5 - Teste de permutação.....	32
2.3 - Conjuntos de dados.....	33
2.3.1 - Gliomas.....	33
2.3.2 - Metabolismo de Sorgo.....	33
2.4 - Avaliação de desempenho do método.....	34
2.4.1 - Simulações.....	34
2.4.1.a - Controle da taxa de falsos positivos.....	35
2.4.1.b - Poder estatístico.....	36
2.4.1.c - Curvas ROC.....	38
2.4.2 - Teste de tempo de execução.....	38
3 - <i>BioNetStat: Biological Network Statistical analysis</i>	39
3.1.1 - Descrição.....	39
3.1.2 - Implementação.....	45
4 - Desempenho dos métodos.....	47
4.1 - Controle da taxa de falsos positivos.....	47
4.2 - Poder estatístico.....	47
4.3 - Comparação do tempo de execução com testes por pares.....	50

4.4 - Discussão.....	51
5 - Estudos de Caso.....	52
5.1 - Gliomas.....	53
5.1.1 - Análise diferencial de redes baseada em características globais.....	55
5.1.2 - Análise diferencial de redes baseada em características locais.....	57
5.1.3 - Análise diferencial dos vértices.....	59
5.1.4 - Visualização das variáveis em vias metabólicas do KEGG.....	61
5.1.5 - Discussão.....	63
5.2 - Metabolismo de sorgo.....	64
5.2.1 - Análise diferencial de redes baseada em características globais.....	68
5.2.2 - Análise diferencial de redes baseada em características locais.....	69
5.2.3 - Análise diferencial dos vértices.....	71
5.2.4 - Visualização das variáveis em vias metabólicas do KEGG.....	72
5.2.5 - Discussão.....	74
6 - Conclusões gerais.....	76
A - Poder estatístico.....	78
B - Teste de tempo de execução.....	80
C - Apêndice de glioblastoma.....	80
C.A - Tabelas de análise de características globais.....	80
C.B - Tabelas de análise de características locais.....	81
C.C - Análise diferencial de vértices.....	84
C.D - Tabela resultado de ANOVA.....	86
D - Apêndice do estudo de caso de sorgo.....	88
- Referências Bibliográficas.....	92

A Lista de figuras

Figura 1 - Distribuição dos graus dos vértices.....	13
Figura 2 - Distribuição do espectro de três modelos geradores de grafos.....	16
Figura 3 - Grafos gerados pelo modelo de Erdős-Rényi.....	18
Figura 4 - Entropias espectrais de três modelos geradores de grafos.....	20
Figura 5 - Gráfico de dispersão dos pares de observação das variáveis aleatórias X e Y.....	23
Figura 6 - Esquema resumido do experimento de De Souza et al. (2015).....	30
Figura 7 - Curvas ROC utilizadas para os cálculos tanto da taxa de falsos positivos quanto do poder estatístico.....	35
Figura 8 - Esquema de funcionamento do BioNetStat.....	37
Figura 9 - Curvas ROC sob a hipótese nula para cada um dos métodos estatísticos.....	41
Figura 10 - Heatmaps de áreas sob a curva ROC (do inglês, AUC), mostrando os resultados dos experimentos de simulações para o teste diferencial de redes baseado nas oito características das redes..	42
Figura 11 - Resultados dos experimentos de simulações para verificar o poder estatístico da análise diferencial de redes baseada na distribuição espectral, em redes com peso (a) e redes sem peso (b)....	44
Figura 12 - Comparação do tempo de execução entre o teste de múltiplas redes (tBNS) com os testes par a par (tpar).....	46
Figura 13 - Redes de correlação construídas com os 134 genes alvos do fator de transcrição NFκB para os quatro tipos cancerígenos (Astrocitoma, Oligodendroglioma, Oligoastrocitoma e Glioblastoma)....	50
Figura 14 - Gráfico de resumo dos resultados dos testes locais (centralidades e coeficiente de agrupamento).....	53
Figura 15 - Mapa da via Pathways in Cancer apresentando a expressão dos genes que formaram as redes analisadas.....	57
Figura 16 - Mapa da via Pathways in Cancer apresentando a Centralidade de betweenness dos genes que formaram as redes analisadas.....	58
Figura 17 - Redes de correlação dos cinco órgãos, folha, colmo, raiz, raiz escora e grãos, com 73 metabólitos do metabolismo primário de sorgo divididos em cinco grupos.....	62
Figura 18 - Gráficos de resumo dos resultados dos testes locais (centralidades e coeficiente de agrupamento).....	64
Figura 19 - Mapa metabólico da via de Fixação de carbono em organismos fotossintetizantes, retirado do KEGG.....	68
Figura 20 - Mapa metabólico da via do Ciclo do citrato, retirado do KEGG.....	69
Figura 21 - Heatmaps de áreas sob a curva ROC (do inglês, AUC), mostrando os resultados dos experimentos de simulações para a análise diferencial de redes baseada na distribuição espectral, em redes com peso (a) e redes sem peso (b).....	74

B Index of Tables

Tabela 1: Ferramentas para a comparação de redes biológicas.....	8
Tabela 2: Funções implementadas no pacote BioNetStat e qual o seu papel na análise diferencial de redes apresentada na figura esquemática do programa (Figura 8).....	39
Tabela 3 - Funções utilizadas em R para a implementação dos testes estatísticos entre as redes de correlação disponíveis no BioNetStat.....	40
Tabela 4 - Análise diferencial de redes baseada na distribuição espectral.....	51
Tabela 5 - Tabela de resumo da análise diferencial de redes considerando as características globais (Distribuição espectral, Entropia espectral e Distribuição dos graus).....	52
Tabela 6 - Resultados dos testes de análise diferencial das redes dos genes de Pathways in Cancer para cinco características locais (centralidades de grau, autovetor, closeness, betweenness e coeficiente de agrupamento).....	54
Tabela 7 - Resultado da análise diferencial do vértice com centralidade de grau, para os 20 genes da rede de Pathways in cancer.....	55
Tabela 8 - Resultado da análise diferencial do vértice com centralidade de betweenness, para os 20 genes da rede de Pathways in cancer.....	56
Tabela 9 - Análise diferencial de redes baseada na distribuição espectral.....	63
Tabela 10 - Tabela de resumo da análise diferencial de redes considerando apenas características globais (distribuição espectral, entropia espectral e distribuição dos graus).....	64
Tabela 11 - Análise diferencial de redes baseada na centralidade de grau.....	65
Tabela 12 - Análise diferencial de redes baseada na centralidade de closeness.....	65
Tabela 13 - Resultado da análise diferencial do vértice com Centralidade de grau, para os 18 metabólitos do grupo Carboidrato.....	66
Tabela 14 - Resultado da análise diferencial do vértice com Centralidade de grau, para os 13 metabólitos do grupo Ácido orgânico.....	67
Tabela 15: Apresentação dos valores obtidos na comparação do tempo de execução entre o teste de múltiplas redes (tBNS) com os testes par a par (tpar).....	75
Tabela 16: Entropia espectral.....	76
Tabela 17: Distribuição dos graus.....	76
Tabela 18: Centralidade de Betweenness.....	77
Tabela 19: Centralidade de Closeness.....	77
Tabela 20: Centralidade de Grau.....	77
Tabela 21: Centralidade de Autovetor.....	78
Tabela 22: Coeficiente de Agrupamento.....	78
Tabela 25: Coeficiente de Agrupamento.....	79
Tabela 24: Centralidade de Autovetor.....	79
Tabela 23: Centralidade de Closeness.....	79
Tabela 24: ANOVA para 139 genes.....	80
Tabela 25: 79 metabólitos discriminados em cinco grupos (Aminoácido, Carboidrato, Ácido Orgânico, Nucleotídeo, Poliol) de acordo com sua estrutura bioquímica.....	82
Tabela 26: Análise diferencial de redes baseada na entropia espectral.....	83
Tabela 27: Distribuição dos graus.....	83
Tabela 28: Centralidade de autovetor.....	83
Tabela 29: Centralidade de betweenness.....	84
Tabela 30: Coeficiente de agrupamento.....	84

1 Introdução

Complexidade biológica e Biologia de Sistemas

As características biológicas, de forma uma geral, surgem das interações entre os inúmeros elementos das células, como metabólitos, proteínas, RNAs e DNAs (Barabási e Oltvai 2004). No decorrer das duas últimas décadas, a produção de dados em larga escala, de diversos níveis biológicos (por exemplo, genes, transcritos, proteínas e metabólitos), aumentou consideravelmente com técnicas como os sequenciamentos de última geração (RNA-seq) e as análises de espectrometria de massas para detecção de proteínas e metabólitos (McKenzie et al. 2016; Zhu, Zhang, e Schadt 2008). A complexidade das interações entre esses elementos e a grande produção de dados fazem com que a análise e interpretação de sistemas biológicos seja um grande desafio. As alterações que acontecem na estrutura dos sistemas ao longo do tempo e em diferentes condições ambientais conferem uma dinâmica nas interações dos elementos (Kitano 2002) e comparar os estados assumidos por um mesmo sistema torna ainda mais complexa a análise desses dados.

É, portanto, comum e fundamental comparar os diferentes estados (condições) que um sistema assume como, por exemplo, pacientes saudáveis e doentes (McKenzie et al. 2016) ou diferentes estágios de crescimento de uma planta (Grandis, Godoi, e Buckeridge 2010; de Souza et al. 2008). As comparações nos permitem, por meio das modificações observadas, entender quais são os elementos chave na diferença entre os sistemas. Considerando também que a biologia de sistemas utiliza os conjuntos de dados biológicos para revelar as relações funcionais entre os elementos (Mutwil et al. 2010), ferramentas foram desenvolvidas para analisar grandes conjuntos de dados, baseados na abordagem sistêmica e permitindo comparar diferentes condições biológicas (Mutwil et al. 2010; Subramanian et al. 2005). Muitas dessas ferramentas se baseiam em teoria de redes, pela qual é possível representar as interações entre os elementos de um sistema complexo (Barabasi e Albert 1999; Koch e Laurent 1999).

Teoria de redes

Uma rede ou grafo é um objeto matemático composto por vértices e arestas que representam, respectivamente, elementos e suas conexões (Junker e Schreiber 2008). Um dos primeiros estudos na área data de 1736, quando Leonard Euler demonstrou, usando grafos, que não era possível percorrer

todos os pontos da cidade de Kchronenberg (Prússia, na época) atravessando todas as suas pontes uma única vez (Diestel 2000). Só mais de dois séculos depois, em 1969, as redes foram usadas para modelar a relação entre genes (Kauffman 1969), e finalmente, no início dos anos 2000, esse uso se difundiu para o estudo de redes metabólicas e redes de interação proteína-proteína (Jeong et al. 2000, 2001). As redes biológicas são geralmente caracterizadas como redes livres de escala por conterem muitos vértices com poucas arestas e poucos vértices com muitas arestas, sendo estes últimos conhecidos como *hubs*, vértices que são importantes por estarem altamente conectados (Barabási e Oltvai 2004).

A abordagem de redes está muito ligada à biologia de sistemas, justamente por trazer uma visão holística e integrada do objeto de estudo (Choi et al. 2005; Emmert-Streib e Dehmer 2011) permitindo explorar o papel de cada elemento e as suas relações (Barabási e Oltvai 2004). A identificação de elementos com papéis centrais em um sistema pode ser feita por meio do estudo de centralidades das redes. Estas últimas são medidas de importância de cada vértice com relação à sua posição na rede (Junker e Schreiber 2008). Os vértices com maior centralidade indicam elementos chave na interação das partes e no funcionamento do sistema. As interações entre os elementos geram novas características, não observáveis no estudo individual dos mesmos. Essas características são chamadas de propriedades emergentes do sistema e podem ser encontradas por meio da análise de centralidades das redes (Jeong et al. 2001). Considerando que as redes podem modelar sistemas complexos, onde as interações são representadas pelas conexões (Barabasi e Albert 1999; Koch e Laurent 1999), a comparação dos estados assumidos por um determinado sistema pode ser feita por meio das diferenças entre as redes de cada um dos estados.

Comparação entre redes

Diversos estudos utilizam as redes como ferramenta de comparação entre diferentes estados biológicos. Caldana et al. (2011) estudaram o metabolismo da planta *Arabidopsis* sp. em diferentes condições de intensidade luminosa e temperatura. A comparação de padrões de conectividade (concentração de arestas) dos aminoácidos entre as redes permitiu levantar hipóteses sobre a degradação de proteínas para o fornecimento de energia em plantas no escuro (Caldana et al. 2011). Em outro trabalho, Weston et al. (2011), estudaram as respostas ao estresse por temperatura em três espécies vegetais. Para cada espécie construíram uma rede de co-expressão dos genes e as subdividiram em módulos. Para cada módulo, a correlação entre os autovalores das matrizes de adjacência (representação da rede em forma de matriz) com medidas bioquímicas do sistema antioxidante permitiu

concluir que as espécies têm mecanismos regulatórios espécie-específicos em resposta ao estresse oxidativo (Weston et al. 2011). Em ambos os trabalhos, Caldana et al. (2011) e Weston et al. (2011), apenas quantificaram as características estruturais das redes, sem aplicar métodos formalizados de comparação dos tratamentos.

Outros autores, ao compararem as redes de diferentes tratamentos, formalizaram métodos para essa comparação. Sun et al. (2013) propuseram a identificação da rede de expressão diferencial. Essa rede resulta da união da análise de expressão diferencial de genes (diferença estatística na expressão média de um gene entre duas condições) com a co-expressão diferencial (diferença estatística na correlação dos níveis de expressão de um par de genes entre duas condições). Ao comparar tecidos de ratos com e sem câncer, a rede de expressão diferencial permitiu encontrar genes e interações gênicas relacionados à doença (Sun et al. 2013). Em outro trabalho, construiu-se uma rede exclusiva para o estudo do desenvolvimento da madeira de *Populus* sp. (Zhang e Yin 2016). A rede exclusiva foi composta apenas por arestas que estavam presentes exclusivamente na rede de caule (madeira), quando comparado com as redes de folha e raiz. Dessa forma, os autores puderam inferir quais genes eram importantes para os processos de desenvolvimento da madeira. A comparação de redes de correlação de diferentes condições pode apontar elementos importantes na distinção dessas condições. Por exemplo, a comparação de redes de co-expressão (rede de correlação de genes) entre indivíduos saudáveis e doentes permite identificar grupos de genes disfuncionais em câncer (Li, Li, e Li 2016). Apesar dos métodos de comparação serem formalizados, a ausência de métodos estatísticos para comparar redes pode levar a afirmações equivocadas sobre diferenças entre os estados biológicos.

O uso de testes de permutação (testes baseados em permutações aleatórias dos dados) permitiu encontrar uma rede de co-expressão diferencial que apontou genes de importante associação com a obesidade (Walley et al. 2012). A técnica de permutação dos dados permite calcular o nível de confiabilidade para o valor de diferença de co-expressão, ou seja, produz-se um p-valor (probabilidade de significância) para cada par de genes. Em outro trabalho, com o uso da técnica de permutação para comparar redes, os autores encontraram diferenças no metabolismo da uva *Vitis vinifera* entre dois tratamentos: controle e déficit hídrico (Hochberg et al. 2013). Técnicas como permutação e reamostragem dos dados permitem uma interpretação estatística das comparações que são realizadas, já que associamos ao resultado uma probabilidade de erro. Considerando a necessidade de se comparar redes entre diversas condições ou tratamentos, foram propostas muitas ferramentas para realizar essa tarefa, algumas delas são abordadas a seguir.

Ferramentas de comparação de redes

As ferramentas descritas neste trabalho realizam o teste de análise diferencial de redes e estão apresentadas na Tabela 1. Algumas ferramentas podem ser agrupadas por testarem se os pares de genes são diferencialmente co-expressos, verificando se as correlações entre os pares de vértices são diferentes entre uma condição e outra (Dawson, Ye, e Kendzierski 2012; Fukushima 2013; Liu et al. 2010). Todas essas ferramentas são implementadas em linguagem de programação R: DCGL (Liu et al. 2010), EBcoexpress (Dawson et al. 2012) e DiffCorr (Fukushima 2013). Outra abordagem é procurar por subgrupos de genes que são diferencialmente co-expressos entre condições diferentes. Nesse sentido, algumas ferramentas agrupam os vértices e verificam se esse agrupamento se mantém entre os estados biológicos, como o DiffCoEx (Tesson, Breitling, e Jansen 2010) e coXpress (Watson 2006).

Existem casos onde os grupos de vértices (variáveis) já estão pré-definidos. Isso ocorre quando as variáveis são combinadas por pertencerem à mesma via metabólica ou grupo funcional. Uma ferramenta que compara os genes em redes produzidas por GGM (Gaussian Graphical Model) é a DINGO (Ha, Baladandayuthapani, e Do 2015). A ferramenta DECODE (Lui et al. 2015) combina o uso de expressão diferencial com co-expressão diferencial para apontar grupos de genes importantes no processo estudado. A dCoxS (Cho, Kim, e Kim 2009), por sua vez, relaciona as variáveis (conecta os vértices) pelo método de entropia relativa. A medida de distância entre as matrizes de entropia relativa, que representam as redes, é feita, então, pelo método de correlação. De maneira semelhante ao dCoxS, o método GSCA (Choi e Kendzierski 2009) constrói as matrizes de correlação e as compara utilizando a distância euclidiana. O programa GSCA (Choi e Kendzierski 2009) é o único que apresenta uma generalização para mais de duas condições. A generalização é feita com a média das distâncias par a par das matrizes de correlações de cada condição. Apesar do bom desempenho e efetividade dessas ferramentas (DINGO, DECODE, dCoxS e GSCA), buscar por estruturas exatamente iguais entre dois grafos não é efetivo, já que essa estrutura pode sofrer flutuações ao longo do tempo e do espaço que não representam, necessariamente, uma mudança sistêmica (Santos et al. 2015).

As principais alterações em um sistema ocorrem nos vértices mais importantes da rede (Rahmatallah, Emmert-Streib, e Glazko 2014). Com base nessa afirmação, a ferramenta GSNCA (Rahmatallah et al. 2014) testa se existe diferença entre os vetores de pesos dos vértices, dados pela centralidade de autovetor. Em outras palavras, o método testa se ocorre uma mudança estatisticamente significativa nos vértices mais importantes das redes, que são definidos por essa medida de

centralidade.

Ao comparar duas redes, considerando a característica observada pelo GSCA, diferença entre matrizes de correlação, e pelo GSNCA, diferença entre os autovetores, padrões biológicos distintos são apontados pelas ferramentas. No caso de todos os genes de duas redes serem diferencialmente co-expressos de maneira uniforme, o teste GSCA (Choi e Kendzierski 2009) tem maior poder estatístico que o GSNCA (Rahmatallah et al. 2014). Já quando o tamanho do conjunto de genes é relativamente grande e a co-expressão não muda uniformemente entre os genes (isto é, a co-expressão de apenas alguns poucos genes sofre alteração) o GSNCA tem maior poder estatístico. Em suma, os métodos são complementares por detectarem modificações diferentes nas redes. Para exemplificar as diferenças entre os métodos, Rahmatallah et al. (2014) comparam linhagens celulares normais com linhagens com a mutação do gene p53. O GSNCA encontrou diferenças significativas principalmente nas vias de sinalização, destacando os genes de controle dessas vias. Por outro lado, GSCA encontrou diferenças nas redes com genes diretamente ligados a vias metabólicas, como de glicólise e citocromo oxidase. As vias metabólicas são alteradas indireta e generalizadamente com a mutação do gene p53. Como as alterações na co-expressão ocorrem na maioria dos genes das vias metabólicas, essas diferenças foram melhor percebidas pelo GSCA. Portanto, a busca por mudanças estruturais das redes permite a detecção de alterações em elementos chave de um sistema, que não são encontradas por outras abordagens (Rahmatallah et al. 2014). Além disso, como todas as conexões de um vértice estão sendo levadas em consideração, a estrutura multivariada dos dados é analisada (Rahmatallah et al. 2014).

A medida de autovetor, usada pelo GSNCA, é uma entre as várias medidas estruturais de grafos (por exemplo, coeficiente de agrupamento, medida dos caminhos mais curtos e outras medidas de centralidades) que estão disponíveis em programas como WGCNA (Langfelder e Horvath 2008) ou Cytoscape (Shannon et al. 2005) e que podem ajudar a encontrar vértices importantes na comparação das redes. Porém, para a maioria dessas medidas estruturais não havia testes estatísticos associados a elas que identificassem diferenças entre redes. Além dessas medidas, Takahashi et al. (2012) propuseram o uso do espectro do grafo, que é uma característica estrutural ainda pouco explorada. O espectro do grafo é definido como o conjunto de autovalores da matriz de adjacência que representa o grafo. Além disso, grafos que são formados por um mesmo modelo de grafo (processo probabilístico que define como o grafo é construído) possuem espectros semelhantes e, em geral, grafos gerados por modelos diferentes possuem espectros diferentes. O espectro do grafo descreve muitas de suas propriedades, como o número de caminhos, diâmetro e cliques (Mieghem 2010). Takahashi et al.

(2012) propõem, então, que a distribuição do espectro do grafo é uma melhor caracterização de determinada classe de grafo quando comparada com outras características estruturais mais comumente usadas (por exemplo, número de links, tamanho do caminho médio e coeficiente de agrupamento). Com base nos estudos de Takahashi et al. (2012), Santos et al. (2015) propuseram um método estatístico para comparar o espectro do grafo em redes de co-expressão.

Considerando a distribuição espectral como uma boa característica para identificar a classe do grafo, seria possível usar o teste de Kolmogorov-Smirnov (vide box 1) como estatística de comparação das distribuições. Esse teste tem sob hipótese nula que duas distribuições provêm do mesmo conjunto amostral. Porém, uma de suas premissas é a independência na amostragem do conjunto de dados, fato que não é atendido pela distribuição do espectro, já que os autovalores de uma matriz são altamente dependentes. Outras duas maneiras de comparar distribuições são as divergências de Kullback-Leibler e de Jensen-Shannon (explicadas com detalhes na seção 2.1.5.d). Ambas as medidas já são muito usadas em teoria da informação e foram introduzidas por Takahashi et al. (2012) para comparar distribuições espectrais. Santos et al. (2015), baseados nessas medidas, desenvolveram um teste estatístico para comparar grafos de co-expressão e propuseram o software CoGA (*Co-expression Graph Analyzer*) como uma ferramenta para análise desses grafos. Assim como a maioria dos programas descritos aqui, o CoGA é um pacote em linguagem R que constrói redes de co-expressão (correlação) e identifica se o grupo de genes que formam essa rede é diferencialmente co-expresso. O programa testa a igualdade entre as distribuições espectrais de duas redes. O CoGA também realiza a comparação entre duas redes por meio de outras medidas estruturais, como a entropia espectral (isto é, quantidade de incerteza associada ao grafo) (Takahashi et al. 2012) centralidades, coeficiente de agrupamento e distribuição dos graus dos vértices. Os grupos gênicos modificados entre dois estados que são apontados nos resultados de CoGA, GSNCA e GSCA diferem entre os programas. Isso mostra que o CoGA complementa análises feitas pelas outras duas ferramentas.

Box 1 – Teste de Kolmogorov-Smirnov

Teste que verifica se as maiores diferenças entre as distribuições cumulativas de duas amostras (ou dois estados, como chamamos neste trabalho) é tão pequena que podemos dizer que as distribuições são da mesma população. Esse é um teste não paramétrico e é usado não apenas para comparar a média entre dois grupos, mas também a dispersão e o formato de duas distribuições de probabilidade. O teste é recomendado quando se quer comparar uma mesma variável em duas populações diferentes ou então quando se quer verificar se uma amostra possui a distribuição semelhante a alguma distribuição já conhecida (LEGENDRE; LEGENDRE, 2003).

Motivação e justificativa

A maioria dos programas que compara redes, com exceção de DiffCor e GSCA, não é capaz de

testar a igualdade de mais de duas condições biológicas (Figura Error: Reference source not found). Como mostrado na seção Comparação de redes (Introdução), existem vários estudos, em diversas áreas, nos quais mais de duas condições são abordadas (Caldana et al. 2011; Hochberg et al. 2013; Weston et al. 2011; Zhang e Yin 2016). Para resolver o problema de comparar mais de duas condições biológicas, Fujita et al. (2017) generalizaram a estatística do CoGA (divergência de Jensen-Shannon entre densidades espectrais). Contudo, o teste proposto, chamado de ANOGVA (*Analysis of Graph Structure Variability*), compara populações de grafos. Essa ferramenta foi desenvolvida para séries temporais, onde cada condição possui vários grafos. Assim, o teste não é aplicável para redes de correlação de dados não temporais, já que usualmente há apenas uma rede para cada condição.

No presente trabalho, nós adaptamos o teste proposto por Fujita et al. (2017) para comparar redes de co-expressão, generalizando, assim, o teste estatístico do CoGA para mais de duas condições. Apresentamos aqui o método de comparação entre duas ou mais redes, chamado *BioNetStat (BNS)*. Ademais, diferente de todas as outras ferramentas que são voltadas, mas não exclusivas, para a comparação de redes de co-expressão, o presente trabalho pode ser aplicado a dados de qualquer natureza onde o objetivo seja comparar redes de correlação. Por esse motivo, em vez de usarmos os termos redes de co-expressão e genes diferencialmente co-expressos abordaremos como redes de correlação e vértices ou variáveis diferencialmente correlacionadas. A generalização do teste de igualdades de redes é feita também para as centralidades (de grau, *closeness*, autovetor e *betweenness*), coeficiente de agrupamento e distribuição dos graus. Além disso, o *BioNetStat* realiza, para as estruturas locais (centralidades e coeficiente de agrupamento) testes estatísticos para verificar se o vértice está diferencialmente correlacionado entre as redes das condições estudadas. Nenhuma das ferramentas até agora disponibilizou essa análise realizando comparações estatísticas.

Ferramentas	Elemento comparado	Grupo de nós pré definidos	Estatística de distância entre as redes/arestas	Número de redes comparadas
DCGL (Liu et al., 2010)	Arestas	Não	Cinco métodos de dCs	2
EBcoexpress (Dawson et al., 2012)	Arestas	Não	Inferência Bayesiana	2
DiffCor (Fukushima, 2013)	Arestas	Não	teste T	2
DiffCoEx (Tesson et al., 2010)	Grupos de nós	Não	Medida de sobreposição topológica	2+
CoXpress (Watson, 2006)	Grupos de nós	Não	Estabilidade do grupo	2
DINGO (Ha, Baladandayuthapani, e Do 2015)	Grupos de nós	Sim	Inferência Bayesiana	2
DECODE (Lui et al., 2015)	Grupos de nós	Sim	Expressão and coexpressão diferenciais	2
dCoxS (Cho, Kim, e Kim 2009)	Grupo de arestas	Sim	Correlação	2
GSCA (Choi e Kendzierski, 2009)	Grupos de nós	Sim	Euclideana	2+
GSNCA (Rahmatallah et al., 2014)	Grupos de nós	Sim	Autovetor	2
CoGA (Santos et al., 2015) *	Grupos de nós	Sim	Distribuição espectral	2
GANOVA (Fujita et al., 2017)	Grupos de nós	Sim	Distribuição espectral	2+p
BioNetStat *	Grupos de nós	Sim	Distribuição espectral	2+

Tabela 1: Ferramentas para a comparação de redes biológicas. O métodos proposto aqui é mostrado em **negrito**. Métodos capazes de comparar duas ou mais redes simultaneamente são indicados com o sinal "2+", na coluna "Número de redes comparadas". O sinal "2+p" indica que o método é capaz de comparar mais de duas populações de redes. O sinal "dC" significa coexpressão diferencial do gene e os métodos aplicados por Lui et al. (2015) são a apresentados no trabalho. Por fim, o sinal "*" indica os pacotes que implementaram uma interface gráfica para realização das análises.

Organização do trabalho

O presente trabalho é formado por cinco seções (Material e métodos, BioNetStat: Biological Network Statistical analysis, Desempenho dos métodos, Estudos de Caso e Conclusões gerais). Em Material e métodos apresentamos formalmente o conceito e as estruturas de redes, os testes estatísticos para a construção de grafos de correlação, os conjuntos de dados que serão usados nas análises e como essas análises serão realizadas. Na seção seguinte (BioNetStat: Biological Network Statistical analysis) apresentamos o funcionamento e implementação do programa que disponibiliza a análise diferencial de redes, proposta neste trabalho. Em seguida são apresentados os resultados dos testes de desempenho dos métodos (Desempenho dos métodos) e duas aplicações desses métodos em dados biológicos reais (Estudos de Caso), afim de demonstrar sua aplicabilidade. Por fim, na última seção (Conclusões gerais) resumimos as contribuições deste trabalho avaliando suas limitações e possíveis contribuições futuras.

1.1 Objetivo

Desenvolver e disponibilizar testes que realizem análise diferencial de redes de correlação e aponte propriedades emergentes no sistema estudado.

1.1.1 Metas

- I. Desenvolver testes estatísticos que permitam a comparação de múltiplas redes baseados em suas

características estruturais.

- II. Desenvolver testes estatísticos que permitam comparar os vértices das redes e apontar quais vértices caracterizam as principais diferenças entre elas.
- III. Implementar a ferramenta em software livre para que os testes fiquem disponíveis para a comunidade científica.

2 Material e métodos

2.1 Estruturas de redes complexas

2.1.1 Notação

No presente trabalho, abordaremos somente *grafos não-dirigidos*. Um grafo $G=(V,E)$ não-dirigido é definido como um conjunto V com n_v vértices (podem ser chamados de nós ou pontos) e outro conjunto de arestas E (arcos ou links) que conectam os vértices de V . Denotamos uma aresta $e \in E$, que conecta os vértices v_1 e $v_2 \in V$ e $v_1 \neq v_2$, pelo par não-ordenado $\{v_1, v_2\}$. Dizemos, então que v_1 e v_2 são incidentes em e e que são adjacentes (ou vizinhos) (Junker e Schreiber 2008). Os grafos podem ser com ou sem peso nas arestas. Para facilitar o tratamento, grafos sem peso terão associados, a cada aresta, peso igual a um. Já o grafo com peso nas arestas tem associado a cada elemento $e \in E$ um valor w real. Neste trabalho, w satisfaz $0 \leq |w| \leq 1$. O tamanho do grafo G é igual ao número de vértices (n_v) pertencentes à V . A matriz de adjacência que representa G é chamada de matriz $\mathbf{A}$ e possui n_v linhas e n_v colunas. Nela, definimos que $\mathbf{A}_{ij} = |w_{ij}|$, se existe aresta que conecta v_i e v_j , onde v_i e $v_j \in V$. $\mathbf{A}_{ij} = 0$, se não existe aresta que conecta v_i e v_j . Em grafos com peso nas arestas, para possibilitar o cálculo das medidas de centralidade dos vértices, a matriz de adjacência é composta apenas pelos valores absolutos dos pesos das arestas.

2.1.2 Medidas de centralidade

Em redes de correlação, os vértices representam as variáveis estudadas em um determinado sistema (por exemplo, genes, proteínas, metabólitos, parâmetros fenotípicos ou morfológicos) e as arestas representam correlações (dependência estatística) entre os valores assumidos por essas variáveis (por exemplo, expressão gênica, concentração de metabólitos ou proteínas, assimilação luminosa ou

tamanho do órgão). Nos sistemas biológicos é comum encontrar variáveis que desempenham papéis importantes no sistema em que estão inseridas. Existem diversas formas de identificá-las por meio da posição do vértice que representa aquela variável na rede. As medidas de importância de um vértice na rede são chamadas de centralidades.

A primeira medida de centralidade abordada neste trabalho é o *grau* do vértice, definido como o número de arestas conectadas a ele. Para grafos com pesos nas arestas, o valor final da centralidade é a soma dos módulos dos pesos das arestas conectadas ao vértice. Definimos, então, o *grau* do vértice (C_g) v_i pertencente ao grafo G com n_v vértices como:

$$C_g(v_i) = \sum_{j=1, j \neq i}^{n_v} |A_{ij}|. \quad (1)$$

Como exemplo, em redes de co-expressão gênica (correlações entre expressões gênicas) os vértices que possuem alta centralidade de grau podem indicar genes que estão relacionados a diversos processos metabólicos, podendo ser entendidos como genes chave na regulação do processo estudado.

Outros tipos de centralidade que consideraremos são aquelas baseadas nas distâncias entre os vértices das redes, como o as centralidades de *betweenness* e de *closeness*. A distância entre dois vértices i e j é a menor soma dos pesos das arestas percorridas para ir de i a j . Para o cálculo da distância, os pesos das arestas serão considerados como o inverso do valor de associação.

Mede-se a importância de um vértice na comunicação da rede por meio do número de caminhos mais curtos que passam por esse vértice. Determinamos, assim, o valor da sua centralidade de *betweenness* (Freeman 1978). O número de caminhos mais curtos entre os vértices v_j e v_k que passam por v_i é definido por $g(v_j, v_i, v_k)$. Desta forma, a centralidade de *betweenness* de v_i é

$$C_b(v_i) = \sum_{j=1, j \neq i}^{n_v} \sum_{k=j+1, k \neq i}^{n_v} g(v_j, v_i, v_k). \quad (2)$$

A centralidade de *closeness* está relacionada à distância que um vértice está dos outros nós da rede. É definida como o inverso do comprimento dos caminhos mais curtos entre v_i e os demais vértices da rede (Freeman 1978). Portanto, se $d(v_i, v_j)$ denota o comprimento do caminho mais curto entre v_i e v_j , a centralidade *closeness* (C_c) do vértice v_i é

$$C_c(v_i) = \frac{1}{\sum_{j=1, j \neq i}^{n_v} d(v_i, v_j)} \quad (3)$$

Outra maneira de calcular a “importância” de um vértice é medindo a importância dos vizinhos com os quais ele faz ligações. Neste sentido, a centralidade de *autovetor* (*eigenvector centrality*) de um vértice foi proposta por Bonacich (1972), como sendo um valor proporcional à soma das centralidades dos vértices adjacentes. Caso as arestas tenham peso, ponderam-se as centralidades usando os valores associados às arestas correspondentes (Rahmatallah et al. 2014). Sendo x_i e N_i a centralidade de autovetor do vértice v_i e o número de vizinhos de v_i , respectivamente, temos que

$$\begin{aligned} x_i &= \frac{1}{\lambda} \sum_{j \in N_i} w_{ij} x_j \\ &= \frac{1}{\lambda} \sum_{j=1}^{n_v} A_{ij} x_j. \end{aligned} \quad (4)$$

Considerando A a matriz de adjacência e λ um escalar denominado *autovalor*, em notação vetorial, definimos $\mathbf{x} = (x_1, x_2, \dots, x_{n_v})$ como um vetor que satisfaz a equação

$$A\mathbf{x} = \lambda \mathbf{x}. \quad (5)$$

Existem muitos autovalores λ para os quais existe um autovetor $\mathbf{x}$ que satisfaça a equação 5. Para evitar componentes negativos, ou seja, garantir que cada componente x_i satisfaça $x_i \geq 0$, utilizamos como centralidade o autovetor associado ao maior autovalor (Bonacich 1987). A centralidade do vértice v_i será, então, a i -ésima componente do autovetor associado ao maior autovalor da matriz de adjacência A .

2.1.3 Medidas de segregação funcional

Considerando redes biológicas, grupos altamente interconectados de vértices geralmente estão associados a funções comuns no sistema estudado. As medidas de segregação podem ajudar a identificar esses grupos dentro da rede. O *coeficiente de agrupamento* (Watts e Strogatz 1998) de um vértice v_i é uma medida de segregação que é definida como o número de pares adjacentes a v_i conectados entre si dividido pelo número de arestas que poderiam existir entre eles. Sendo k_i o número de vizinhos de v_i , então, $k_i (k_i - 1) / 2$ é o número de arestas possíveis entre os vizinhos de v_i . O *coeficiente de agrupamento* para grafos com peso é obtido pela razão entre a soma dos pesos das

arestas conectando os vizinhos de v_i e $k_i (k_i - 1) / 2$ (Lopez-Fernandez, Robles, e Gonzalez-Barahona 2004). Sendo N_i a vizinhança de v_i , o *coeficiente de agrupamento* de v_i é dado por

$$C(v_i) = \frac{\sum_{v_j, v_k \in N_i} A_{jk}}{k_i(k_i - 1)}. \quad (6)$$

A medida de segregação estará entre 0 e 1, quando $|A_e| \leq 1$, sendo que zero significa que os vizinhos do vértice não estão conectados. Para grafos sem peso nas arestas, quando todos estão conectados entre si, o valor do *coeficiente de agrupamento* é um. Esta afirmação é válida também para grafos com peso, caso todos os pesos das arestas que conectam os vizinhos forem iguais a um. Para redes de correlação de genes (co-expressão) ou de metabólitos, um *coeficiente de agrupamento* próximo a um pode indicar que aquele grupo de vértices estão relacionados a uma mesma via metabólica ou que desempenham funções semelhantes (Roy, Bhattacharyya, e Kalita 2014).

2.1.4 Medidas de resiliência

Medidas de resiliência de um sistema permitem caracterizá-lo quanto a sua capacidade de sofrer alterações ou ataques e se manter em funcionamento. Em redes aleatórias, onde as arestas são adicionadas aleatoriamente com igual probabilidade para todos os pares de vértices, a retirada de um número crítico de vértices pode desconectar grupos inteiros e causar mal funcionamento do sistema. Já em redes complexas, como redes biológicas, existe uma estrutura mais robusta e menos sensível à ocorrência de falhas ao acaso (Barabási e Oltvai 2004). Uma característica estrutural que está associada à robustez do sistema é a distribuição dos graus dos vértices da rede (Barabasi e Albert 1999). Por exemplo, redes com distribuição dos graus com cauda pesada (onde muitos vértices possuem poucas arestas e poucos vértices possuem muitas arestas) tendem a ser robustas a falhas aleatórias (Figura 1). Por outro lado, são vulneráveis a falhas em vértices centrais (hubs) (Barabasi e Albert 1999).

Figura 1: Distribuição dos graus dos vértices. O eixo x representa o grau dos vértices (número de arestas conectadas ao vértice) e o eixo y a frequência com que os vértices com determinado grau aparecem na rede. Nesse gráfico está representada uma rede onde muitos vértices tem poucas arestas (baixo grau) e poucos vértices tem muitas arestas (alto grau).

2.1.5 Medidas baseadas na caracterização espectral do grafo

Considerando A a matriz de adjacência do grafo $G = (V, E)$ não-dirigido com n_v vértices, temos que o espectro de G é o conjunto de autovalores da matriz de adjacência A . A matriz A é uma matriz simétrica, já que G é um grafo não-dirigido, e assim, todos os autovalores da matriz são reais. Como já dito na Introdução, o espectro do grafo é o conjunto de autovalores da matriz de adjacência. O espectro do grafo descreve várias propriedades estruturais desse grafo, como diâmetro, número de caminhos e cliques (Mieghem 2010). Dessa forma, dizemos que um grafo aleatório é uma família de grafos gerados de acordo com determinada lei de probabilidade. E, então, definimos o *modelo de grafo aleatório* como um algoritmo que descreve como a família de grafos é gerada. Representaremos por g a família de grafos gerada por um mesmo modelo de grafo aleatório. O espectro de um grafo com n_v vértices, será composto por n_v autovalores reais $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{n_v}$. Dessa forma, a *densidade espectral de g* é a *função de densidade de probabilidade* dos autovalores dos grafos de g . Representamos o delta de Dirac por δ , que possui as seguintes propriedades: (i) $\delta(x)=0, x \in \mathbb{R}^*$; (ii) $\delta(0)=\infty$; e (iii)

$\int_{-\infty}^{+\infty} \delta(x) dx = 1$. Os colchetes ' $\langle \rangle$ ' serão usados para denotar a esperança com respeito à lei de

probabilidade de g . Por fim, a *densidade espectral* (Takahashi et al. 2012) da família de grafos g é definida como

$$f(\lambda) = \lim_{n_v \rightarrow \infty} \left\langle \frac{1}{n_v} \sum_{j=1}^{n_v} \delta(\lambda - \lambda_j / \sqrt{n_v}) \right\rangle. \quad (7)$$

A seguir mostraremos alguns modelos geradores de grafos que nos permitam modelar redes complexas e definir as classes de grafos. Nessas classes são compartilhadas muitas características estruturais que ocorrem em sistemas reais. Alguns exemplos de modelos geradores já muito conhecidos são de Erdős e Rényi (1959), Barabási e Albert (1999) e Watts e Strogatz (1998). Apresentaremos esses modelos nas próximas seções, mostrando suas densidades espectrais e medidas baseadas na distribuição do espectro, como a entropia espectral e a divergência Kullback-Leibler entre as densidades espectrais.

2.1.5.a Distribuição do espectro em modelos conhecidos

Modelo de Erdős e Rényi

O modelo mais simples, já citado anteriormente, é o modelo de grafo aleatório de Erdős e Rényi (1959). Os autores propuseram um algoritmo que gera um grafo com n_v vértices acrescentando uma aresta a um par de vértices com probabilidade p .

Modelo de Barabási-Albert

No modelo Barabási-Albert o grau do vértice define a probabilidade de inserção de uma aresta no grafo. Barabási e Albert (1999) descreveram o algoritmo de formação do grafo onde, dado um número inicial pequeno (n_{v0}) de vértices, é adicionado um novo vértice a cada passo, com $m_1 (\leq n_{v0})$ arestas que conectam o novo vértice a m_1 vértices diferentes, já presentes na rede. O expoente de escala (p_s) define a ordem de proporcionalidade que o grau de v_i (número de arestas conectadas a v_i) terá na probabilidade do vértice v_i se conectar a um novo vértice. Bollobás et al. (2001) demonstraram que, nos grafos gerados por esse modelo, a lei de potência rege, de maneira assintótica, a distribuição dos graus. Sendo assim, a proporção de vértices com grau d será aproximadamente igual a d elevado a um fator $-\gamma$ (Bollobás et al. 2001).

Modelo de Watts e Strogatz

Por sua vez, Watts e Strogatz (1998) propuseram um modelo onde ocorrem tanto inserções de

arestas determinadas por uma estrutura em forma de anel com os vértices conectados aos $K/2$ vértices mais próximos de cada lado da estrutura, quanto inserções aleatórias. Como descrito em Santos (2015), o modelo segue os seguintes passos: (i) dados n_v vértices, construa uma estrutura anelar e conecte cada vértice aos K vértices mais próximos ($K/2$ de cada lado do anel); (ii) escolha um vértice v_i e a aresta e que o conecta ao seu vizinho mais próximo no sentido horário; (iii) com probabilidade p_r substitua a aresta e por uma aresta que ligue v_i a um vértice escolhido aleatoriamente entre todos os vértices do anel; (iv) repita os passos (ii) e (iii) para cada vértice do anel em sentido horário; (v) repita os passos (ii-iv), alterando o passo (ii) de forma que a aresta escolhida conecte o vértice v_i ao segundo mais próximo no sentido horário; (vi) repita o processo considerando o terceiro vértice mais próximo e assim por diante, até que todas as arestas tenham sido analisadas.

Cada família de grafos possui, então, características estruturais decorrentes do seu algoritmo gerador. O modelo Barabási-Albert gera grafos caracterizados pela distribuição dos graus que segue a Lei da Potência. Já o modelo de Watts-Strogatz gera grafos com *coeficientes de agrupamento* altos e comprimentos de caminhos curtos. Embora tais características estejam associadas ao modelo gerador do grafo elas não são efetivas na discriminação de grafos gerados por modelos diferentes. Grafos de Watts-Strogatz se assemelham aos gerados pelo modelo de Erdős e Rényi no sentido de terem altos *coeficientes de agrupamento*, formato anelar com cada vértice conectado aos vértices mais próximos e com pequenas distâncias médias entre os vértices. Levando em consideração que as características citadas acima não são eficazes na caracterização de um grafo, Takahashi et al. (2012) propuseram usar os espectros dos grafos para compará-los. Segundo esses autores, o espectro descreve propriedades mais gerais do grafo se comparado a outras estruturas geralmente utilizadas, como o número de arestas, o *coeficiente de agrupamento* e o comprimento médio dos caminhos mais curtos.

Na Figura 2, retirada de (Takahashi et al. 2012) é possível observar o histograma dos autovalores das matrizes de adjacência geradas pelos modelos Erdős-Rényi (Figura 2a), Barabási-Albert (Figura 2b) e Watts-Strogatz (Figura 2c). Para o caso de grafos de Erdős e Rényi, com um número grande de vértices a densidade espectral pode ser aproximada para

$$f(\lambda) \sim \frac{\sqrt{4p(1-p) - \lambda^2}}{2\pi p(1-p)}, \quad (8)$$

se $0 < |\lambda| < 2\sqrt{p(1-p)}$. Caso contrário, $f(\lambda) = 0$ (Takahashi et al. 2012).

Figura 2 - Distribuição do espectro de três modelos geradores de grafos. Na parte superior da figura estão exemplificados os grafos gerados a partir dos modelos de (a) Erdős-Rényi, (b) Barabási-Albert e (c) Watts-Strogatz. Na parte inferior, estão os respectivos histogramas dos autovalores das matrizes de adjacência desses grafos. A figura foi adaptada de Takahashi et al.(2012).

Considerando a relação entre a distribuição espectral do grafo e suas características estruturais, Takahashi et al. (2012) apresentaram os conceitos de entropia para medir a quantidade de incerteza associada ao grafo e também a divergência de Kullback-Leibler entre densidades espectrais de grafos. Esses conceitos serão descritos a seguir.

2.1.5.b Entropia

Shannon (1948) introduziu o conceito de entropia em teoria da informação. A entropia de uma variável aleatória mede a quantidade de incerteza associada ao seu valor. Para exemplificar o conceito vamos considerar que duas variáveis discretas X e Y possam assumir os valores entre 0 e 1. Supomos, então, que $P(X=0)=0,50$, $P(X=1)=0,50$, $P(Y=0)=0,95$ e $P(Y=1)=0,05$. Ou seja, o valor da variável Y será “quase certamente” 0, com 95% de probabilidade, enquanto o valor de X é incerto pois, existe 50% de chance de ser 1 ou 0. Dizemos, portanto que a variável X está associada a uma quantidade de incerteza maior do que a variável Y . Adicionando uma variável discreta Z , com $P(Z=0)=0,75$ e $P(Z=1)=0,25$, vemos que Z tem uma quantidade de incerteza associada intermediária entre as

incertezas de X e Y . A entropia pode representar, de maneira intuitiva, essa incerteza associada a uma variável discreta X , a qual pode assumir n valores com probabilidade $p_1, p_2, \dots, p_n$ e pode ser definida como

$$H(X) = - \sum_{i=1}^n p_i \log p_i, \quad (9)$$

onde $p_i \log p_i = 0$ quando $p_i = 0$ e o logaritmo tem uma base aleatória fixa (Shannon 1948).

Note que quando $H(X)$ vale zero existe total previsibilidade dos valores de X . Isto porque, apenas neste caso, um dos números de $p_1, p_2, \dots, p_n$ é igual a um (ou seja, $p_i=1$ e o $\log 1 = 0$) e todos os outros valem zero, sendo assim, existe 100% de chance de X ser um determinado valor quando sua entropia $H(X)$ valer zero. Em qualquer outro caso, a entropia terá valor positivo. A maior incerteza possível será associada ao valor de X se $p_1=p_2=\dots=p_n=1/n$. Onde, neste caso, a entropia assumirá o seu maior valor (Khinchin, 1957), ou seja, quando $H(X)$ vale

$$- \sum_{i=1}^n \frac{1}{n} \log \frac{1}{n}.$$

Define-se de maneira semelhante a entropia de variáveis contínuas, ou entropia diferencial. Substitui-se o somatório da equação 9 por uma integral no suporte da função de densidade de probabilidade. A entropia diferencial é capaz de assumir valores negativos, o que não ocorre na entropia definida para distribuições discretas.

2.1.5.c Entropia de grafos

Quando a entropia é definida para um grafo ela quantificará a aleatoriedade da estrutura desse grafo, sendo definida como se segue. Seja g uma família de grafos gerados de acordo com uma lei de probabilidade e f a densidade espectral de g . A *entropia espectral* de g é

$$H(f) \leq - \int_{-\infty}^{+\infty} f(\lambda) \log f(\lambda) d\lambda \quad (10)$$

onde $0 \log 0 = 0$ (Takahashi et al. 2012). A partir das fórmulas (8) e (10) a entropia de um grafo do modelo de Erdős-Rényi pode ser aproximada para

$$H(f) \sim \frac{1}{2} \ln(4\pi^2 p(1-p)) - \frac{1}{2} \quad (11)$$

Sendo que p é a probabilidade de uma aresta conectar um par de vértices (Takahashi et al. 2012). Por

meio da equação 11 podemos notar que o valor de entropia mais alto ocorrido no modelo do grafo de Erdős-Rényi se refere ao valor de $p = 0,5$. O valor corrobora o que já é esperado para esse modelo, pois quando $p = 0,5$ temos o cenário de maior incerteza, isto é, o cenário em que é mais difícil prever se dois vértices estão ou não conectados. Por outro lado, a incerteza será mínima quando $p = 0$ e $p = 1$. Nestes casos os grafos serão vazios (se $p = 0$) e completos (se $p = 1$), como exemplificado na Figura 3. Propriedades que também são satisfeitas na equação 10 quando consideramos esse modelo de grafo (Takahashi et al. 2012), como mostrado nas Figuras 3 e 4a.

Figura 3: Grafos gerados pelo modelo de Erdős-Rényi. Na parte superior da figura estão exemplos grafos gerados pelo modelo de Erdős-Rényi utilizando três diferentes valores do parâmetro p (0, 0,5 e 1). Os grafos a esquerda e direita foram construídos com $p = 0$ e $p = 1$ e possuem o menor valor de entropia possível (menor incerteza), já que não há nenhuma conexão entre os vértices ou todos estão conectados. Por outro lado quando o valor de $p = 0,5$ a entropia é máxima já que a chance de um par de vértices estarem conectados é a mesma de não estarem.

A noção intuitiva de quantidade de incerteza também é satisfeita pelos valores obtidos pela entropia espectral (equação 10), considerando os modelos de Barabási-Albert e de Watts-Strogatz. Takahashi et al. (2012) utilizaram diferentes parâmetros para estimar as entropias a partir de distribuições empíricas do espectro dos grafos de modelos como de Barabási-Albert e Watts-Strogatz. A Figura 4b mostra que a entropia do grafo de Barabási-Albert é inversamente proporcional ao

expoente de escala (p_s), parâmetro apresentado na seção 2.1.5.a. A aleatoriedade do grafo é maior se (p_s) for pequeno. Isso porque o grau do vértice tem pouca influência sobre a probabilidade de conectar um vértice a outro. Por outro lado, quando o p_s fica maior, o grau do vértice passa a definir com maior intensidade onde serão adicionadas as novas arestas, diminuindo o grau de incerteza. Por fim, para o modelo de Watts-Strogatz (Figura 4c), notamos que a entropia aumenta com o aumento do parâmetro de probabilidade p_r . O parâmetro p_r , como dito na seção 2.1.5.a, é a probabilidade de substituir a aresta recém-inserida no grafo, que conecta um vértice v_i a outro vértice que está próximo a ele na estrutura de anel, por uma aresta que conecta v_i a um vértice escolhido aleatoriamente. Dessa forma, se $p_r = 1$, o grafo é construído aleatoriamente, como um grafo de Erdős-Rényi e com a entropia máxima. Já, quando $p_r = 0$, o grafo construído será determinado pela estrutura do anel, sendo que cada vértice estará conectado aos K vértices mais próximos. Neste caso, a entropia do grafo será mínima.

A definição da entropia espectral nos leva a definir a entropia espectral cruzada, entre duas densidades espectrais f_1 e f_2 (quantidade de incerteza quando f_2 é usada para estimar f_1), como

$$H(f_1, f_2) = - \int_{-\infty}^{+\infty} f_1(\lambda) \log f_2(\lambda) d\lambda$$

onde $0 \log 0 = 0$.

Figura 4 - Entropias espectrais de três modelos geradores de grafos. Na parte superior da figura estão representados os grafos gerados por cada um dos modelos. Na parte inferior, estão as respectivas entropias espectrais estimadas de acordo com distribuições empíricas dos espectros gerados pelos modelos. (a) Entropia do espectro do modelo de Erdős-Rényi, onde variou-se a probabilidade p de conectar dois pares de vértices. (b) Entropia do espectro do modelo de Barabási-Albert onde se usou o expoente de p_s . (c) Entropia do espectro do modelo de Watts-Strogatz, onde o parâmetro usado foi p_r , a probabilidade de substituir uma aresta por outra que conecta um vértice escolhido aleatoriamente. Os expoentes e modelos geradores estão apresentados na seção 2.1.5.a. A figura foi adaptada de Takahashi et al. (2012).

2.1.5.d Divergência de Kullback-Leibler

Ao passo que a entropia mostra o quanto de incerteza está associada a uma variável aleatória, a divergência de Kullback-Leibler (KL) mede a informação perdida quando uma distribuição de probabilidade é usada para aproximar outra (Santos 2015). A fim de comparar grafos e obter o modelo de grafo que melhor descreve o grafo em estudo, podemos usar a divergência de KL para discriminar distribuições de probabilidade. A divergência de **Kullback-Leibler** entre grafos é formalmente descrita a seguir.

Definimos g_1 e g_2 como duas famílias de grafos com densidades espectrais f_1 e f_2 , respectivamente. Caso o suporte de f_2 contenha o suporte de f_1 , a divergência de KL entre f_1 e f_2 é definida como

$$KL(f_1|f_2) = H(f_1, f_2) - H(f_1)$$

$$= \int_{-\infty}^{+\infty} f_1(\lambda) \log \frac{f_1(\lambda)}{f_2(\lambda)} d\lambda,$$

onde $0 \log 0 = 0$ e f_2 é chamada de medida de referência (Takahashi et al. 2012). Se o suporte de f_2 não contém o suporte de f_1 , então $KL(f_1|f_2) = +\infty$.

A divergência de KL é não negativa e vale zero se, e somente se, f_1 e f_2 são iguais. KL é uma medida assimétrica, pois em diversos casos $KL(f_1|f_2)$ e $KL(f_2|f_1)$ são diferentes quando $f_1 \neq f_2$.

2.1.5.e Divergência de Jensen-Shannon

Uma alternativa simétrica a divergência de Kullback-Leibler é a divergência de Jensen-Shannon (JS). Definimos a *divergência de JS* entre as densidades espectrais f_1 e f_2 como

$$JS(f_1|f_2) = \frac{1}{2}KL(f_1|f_M) + \frac{1}{2}KL(f_2|f_M)$$

onde $f_M = \frac{1}{2}(f_1 + f_2)$ (Takahashi et al. 2012).

A divergência de Jensen-Shannon pode ser interpretada como uma medida de diferenças estruturais entre dois grafos. A raiz quadrada da *divergência de JS* é uma métrica, ou seja, satisfaz as propriedades: (i) é zero, se e somente se f_1 e f_2 são iguais, (ii) é simétrica, (iii) é não negativa e (iv) satisfaz a desigualdade triangular.

2.2 Teste estatístico de comparação de múltiplos grafos

Na seção 2.1, Estruturas de redes complexas, apresentamos propriedades estruturais que são utilizadas para analisar redes complexas. Neste trabalho, vamos abordar apenas grafos de correlação, que podem apresentar características de redes complexas, como distribuição dos graus dos vértices com cauda pesada e alto coeficiente de *clustering*. A correlação entre as variáveis é inferida a partir dos valores assumidos por essas variáveis em um dado experimento. Portanto, a análise diferencial de redes de correlação requer métodos de inferência estatística. Nas próximas seções vamos apresentar técnicas estatísticas para testar se dois ou mais grafos apresentam as mesmas propriedades estruturais, utilizando

as medidas definidas na seção 2.1.

2.2.1 Enunciado do problema

Consideremos um conjunto de elementos $V = \{v_1, v_2, \dots, v_{nv}\}$ e r populações $P_1, P_2, \dots, P_r$. Seja n_i o número de observações da população P_i , para $i=1, 2, \dots, r$. O problema consiste em verificar se a estrutura de interações entre os itens de V é igual entre as r populações. Se existe diferença nas interações entre os itens de V , queremos identificar, se possível, quais desses itens tem interações diferentes com os demais, comparando as populações. Os itens serão as variáveis nas redes de correlação (por exemplo, genes, proteínas, metabólitos ou variáveis fenotípicas). Cada observação será correspondente a um valor assumido por uma variável em uma amostra experimental. Por exemplo, nível de expressão gênica (para genes), concentração (para proteínas e metabólitos) e um valor numérico que uma variável fenotípica possa assumir (por exemplo, massa ou tamanho).

2.2.2 Construção do grafo de correlação

O grafo de correlação é um grafo não dirigido, onde cada vértice corresponde a uma variável e cada aresta representa uma correlação (medida de dependência estatística) entre duas variáveis. Neste trabalho iremos abordar duas principais medidas de dependência estatística. As correlações de Pearson e Spearman.

A primeira medida que descreveremos é a medida de correlação de Pearson. Essa correlação mede o quanto uma associação pode ser descrita como uma função linear. Ou seja, a equação de uma reta com coeficiente angular não nulo. O coeficiente de correlação de Pearson entre duas variáveis aleatórias X e Y é calculado como a razão entre a covariância e o produto dos desvios padrão. Considerando n o número de unidades amostrais e as médias dos n valores assumidos por X e por Y ,

$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ e $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$, respectivamente. O coeficiente de correlação de Pearson (Pearson 1920) entre X e Y é dado por

$$r_p(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}.$$

O coeficiente pode assumir valores entre -1 (relação linear decrescente perfeita) e 1 (relação linear crescente perfeita). Um coeficiente com valor zero indica que não há dependência linear.

Podemos ver na Figura 5 que o coeficiente não é robusto à presença de *outliers*. Na Figura 5a existe uma forte associação linear, porém, quando inserimos *outliers* (Figura 5b), o coeficiente de correlação de Pearson (r_p) cai bruscamente de 0,98 para 0,5.

Figura 5 - Gráfico de dispersão dos pares de observação das variáveis aleatórias X e Y . As variáveis são linearmente dependentes em (a) porém em (b) existe uma inserção de outliers que desfaz essa relação. Podemos verificar as medidas do coeficiente de correlação de Pearson (r_p) e de Spearman (r_s) para os dois casos. O coeficiente de correlação de Pearson é cai de 0,98 para 0,5 com a inserção dos outliers, enquanto o coeficiente de spearman cai apenas de 0,97 para 0,83.

A estatística do teste de independência é

$$t = \frac{r_p \sqrt{n-2}}{\sqrt{1-r_p^2}}.$$

Sob a hipótese nula, se X e Y seguirem distribuição normal conjunta, t segue uma distribuição de t de Student com $(n-2)$ graus de liberdade.

O coeficiente de correlação de Spearman é considerada uma boa alternativa ao coeficiente de correlação de Pearson quando se deseja utilizar um método robusto à presença de outliers. Além disso, o coeficiente de Spearman detecta não só relações lineares, mas também associações monotônicas não-lineares. Vamos chamar de *posto* a posição de um elemento em um vetor ordenado (em ordem crescente). Realizando o teste de correlação de Pearson nos dados transformados em postos, obtemos o

coeficiente de correlação de Spearman (Spearman 1904). O valor do coeficiente também varia entre -1 (relação monotônica decrescente perfeita) e 1 (relação monotônica crescente perfeita). Ainda, similarmente à medida de Pearson, o valor do *coeficiente de Spearman* igual a zero indica que as variáveis são monotonicamente independentes.

Com a presença de *outliers* nos dados, geralmente, damos preferência a aplicação do coeficiente de correlação de Spearman. Ao serem convertidos em postos, os *outliers* não provocam grandes mudanças no comportamento do coeficiente de Spearman. Por este motivo, podemos afirmar que a correlação de Spearman é robusta à presença de *outliers*. No exemplo apresentado na Figura 5, diferente da correlação de Pearson, a correlação de Spearman muda de 0,97 para 0,83.

A estatística do teste de independência é

$$t = \frac{r_s \sqrt{n-2}}{\sqrt{1-r_s^2}}.$$

Sob H_0 , t segue uma distribuição de t de *Student* com $n-2$ graus de liberdade.

Escalas da força de associação

A partir da escolha de uma medida de dependência, consideramos três escalas para medir a “força” de associação: (i) o valor absoluto da medida de dependência, (ii) um menos o p-valor do teste de independência e (iii) um menos o q-valor obtido pelo método de Benjamini e Hochberg (1995) para controlar a taxa de falsos positivos dos testes de independência. De acordo com o que é definido para a medida de dependência, podemos dizer que a escala (i) quantifica a “força” de associação entre os valores assumidos pelas variáveis. Na escala (ii), considera-se o resultado de teste de independência. p-valores próximos de zero indicam associações “fortes”, enquanto p-valores próximos de um indicam associações “fracas”. Portanto, um menos o p-valor será maior quanto mais “forte” for a associação entre as variáveis. A ocorrência de falsos positivos dos testes de independência realizados com todos os pares de vértices do grafo pode ser controlada utilizando-se a escala (iii).

Para esclarecer melhor o procedimento aplicado na escala (iii), supomos que a probabilidade de ocorrer um erro do tipo I (rejeição incorreta de uma hipótese nula verdadeira, ou falsos positivos) seja fixada em $\alpha = 0,05$. Em 100 testes independentes entre si, onde as hipóteses nulas são verdadeiras, a probabilidade de não cometermos o erro do tipo I é $(1-0,05)^{100}=0,006$. Desse modo, a probabilidade de rejeitar incorretamente pelo menos uma hipótese nula é de $1-0,006 = 0,994$, um valor muito superior ao

α desejado. Usamos o método FDR (*False Discovery Rate*) de Benjamini e Hochberg (1995) com o objetivo de controlar a taxa de ocorrências do erro tipo I (taxa de falsos positivos) em múltiplos testes.

1. Considere as hipóteses testadas $H_1, H_2, \dots, H_h$, sendo h o número total de hipóteses testadas, e os respectivos p-valores $p_1, p_2, \dots, p_h$.
2. Sejam $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(h)}$, os p-valores ordenados.
3. Seja k o maior i tal que $p_{(i)} \leq \frac{i}{h} \alpha$.
4. Rejeitamos $H_{(i)}$, para $i = 1, 2, \dots, k$.

Equivalentemente, podemos “corrigir” os p-valores dos testes, com o seguinte procedimento:

$$\begin{aligned}\tilde{p}_{(h)} &= p_{(h)} \\ p_{(h-1)} &= \min(\tilde{p}_{(h)}, \frac{h}{h-1} p_{(h-1)}) \\ &\vdots \\ \tilde{p}_{(1)} &= \min(\tilde{p}_{(2)}, h p_{(1)})\end{aligned}$$

Chamamos de q-valor ou “p-valor corrigido” cada um dos $\tilde{p}_{(i)}$ obtidos pelo procedimento descrito. De maneira semelhante ao p-valor, q-valores próximos de um indicam altas probabilidades da correlação observada entre as variáveis ocorrer mesmo que não exista correlação entre elas. Se o q-valor é próximo de zero, então há uma probabilidade pequena dos valores de correlação ocorrerem sem que exista de fato uma correlação entre as variáveis. Neste trabalho usaremos a escala (iii), um menos o q-valor, como a força de associação entre os valores assumidos pelas variáveis.

2.2.3 Testes estatísticos

Depois de construídos os grafos de correlação, nosso objetivo é comparar, por meio de testes estatísticos, as estruturas de dois ou mais grafos ou os valores assumidos por um vértice entre dois ou mais grafos. Dada uma estatística θ que quantifica as diferenças estruturais entre múltiplos grafos, aplicaremos testes estatísticos descritos pelas seguintes hipóteses nula e alternativa:

$$\begin{aligned}H_0 &: \theta = 0, \\ H_1 &: \theta > 0.\end{aligned}$$

2.2.4 Estatística dos testes

2.2.4.a Análise diferencial de redes

Considere os grafos $G_1, G_2, \dots, G_r$, cada um com n_v vértices. A estatística θ , apresentada na seção anterior (seção 2.2.3), é medida da seguinte forma. Calcula-se para cada grafo G_i , com $1 \leq i \leq r$, sua propriedade estrutural, que será denotada por S_i . No contexto deste trabalho, chamamos de propriedade estrutural qualquer medida que caracterize alguma estrutura do grafo, como descrevemos na seção 2.1 (centralidades, coeficientes de agrupamento, distribuição dos graus, entropia e distribuição do espectro). A partir das propriedades estruturais de cada um dos r grafos ($S_1, S_2, \dots, S_r$) calculamos uma propriedade estrutural média (M), definida como

$$M = \frac{\sum_{i=1}^r S_i}{r}.$$

Calculamos, então, a distância entre a propriedade estrutural do grafo i (S_i) e a propriedade estrutural média como segue

$$D_i = \text{dist}(S_i, M).$$

A estatística θ de diferença entre os grafos é calculada como a média das r distâncias D_i :

$$\theta = \frac{\sum_{i=1}^r D_i}{r}.$$

Por fim, para sabermos o quanto cada rede influenciou na estatística θ , usamos a razão de D_i por θ obtendo o B_i que representa o peso (em porcentagem) da rede i na estatística final

$$B_i = 100 \frac{D_i}{\theta}.$$

As estatísticas usadas para calcular as distâncias D_i entre a propriedade estrutural da rede i (S_i) e a média (M) são:

- *Distância Euclidiana entre medidas de redes complexas.* Dados, o vetor de propriedades estruturais dos vértices do i -ésimo grafo $s_i = \{s_i^1, \dots, s_i^{n_v}\}$ e o vetor com as médias de cada elemento dos r vetores de propriedades estruturais $\mathbf{m} = \{m^1, \dots, m^{n_v}\}$, a distância euclidiana, ajustada pelo tamanho do grafo (n_v) é dada por

$$D_i = \sqrt{\frac{\sum_{j=1}^{n_v} (s_i^j - m^j)^2}{n_v}}.$$

As medidas comparadas por distância euclidiana serão as centralidades de grau, *closeness*, *betweenness*, autovetor e o coeficiente de agrupamento.

- *Quadrado da diferença entre medidas de redes complexas.* Sejam s_i e m a propriedade estrutural da rede i e a propriedade estrutural média, respectivamente. O quadrado da diferença entre s_i e m é

$$D_i = (s_i - m)^2.$$

Neste caso, utilizamos a medida de entropia espectral como propriedade estrutural de cada grafo. Especialmente, para a entropia espectral a estatística final é calculada por

$$\theta = \sqrt{\frac{\sum_{i=1}^r D_i}{r}}.$$

- *Divergência de Kullback-Leibler entre densidades.* Seja f_i a função de densidade de probabilidade associada ao grafo G_i , onde $1 \leq i \leq r$. Chamamos de f_m a média entre $f_1, f_2, \dots, f_r$. A distância D_i entre f_i e f_m é

$$D_i = KL(f_i \| f_m),$$

onde KL é a divergência de Kullback-Leibler entre duas funções de densidade de probabilidade. No presente trabalho, são obtidas funções de densidade de probabilidade do espectro do grafo (distribuição espectral) e dos graus dos vértices (distribuição dos graus)

2.2.4.b Análise diferencial dos vértices

Considere os grafos $G_1, G_2, \dots, G_r$, cada um com n_v vértices. Calculamos a estatística θ_v , para cada vértice v , da seguinte forma. Para cada grafo G_i , onde $1 \leq i \leq r$, e vértice v , onde $1 \leq v \leq n_v$, obtém-se a propriedade estrutural desejada, que é denotada por S_i^v . No caso de comparações entre os vértices as propriedades estruturais usadas serão só aquelas que podem ser consideradas para cada vértice como centralidades e coeficientes de agrupamento. A partir das propriedades estruturais de cada vértice v ($S_1^v, S_2^v, \dots, S_r^v$) calculamos a propriedade estrutural média, M^v , que é definida como

$$M^v = \frac{\sum_{i=1}^r S_i^v}{r}.$$

Calculamos, então, o módulo da diferença entre a propriedade estrutural do grafo i e vértice v (S_i^v) e a propriedade estrutural média do vértice v (M^v) como segue

$$D_i^v = |S_i^v - M^v|.$$

Por fim, a estatística θ^v de diferença dos valores assumidos pelo vértice v nos grafos $G_1, G_2, \dots, G_r$ é calculada como a média das distâncias D_i^v entre a propriedade estrutural do vértice v , no grafo i , e a propriedade estrutural média do vértice v , da seguinte forma

$$\theta^v = \frac{\sum_{i=1}^r D_i^v}{r}.$$

2.2.5 Teste de permutação

Por simplificação, nesta seção utilizaremos θ para denotar qualquer estatística que meça diferença entre os grafos, seja de análise diferencial de redes (seção 2.2.4.a) ou análise diferencial dos vértices (seção 2.2.4.b). O cálculo do p-valor do teste de hipóteses deve ser feito a partir da obtenção da distribuição de probabilidade da estatística do teste sob a hipótese nula (Legendre e Legendre 2003). A distribuição pode ser obtida ou por aproximação assintótica ou por permutações aleatórias dos dados. Para uma aproximação assintótica é necessária uma forma analítica da distribuição, que é desconhecida para algumas estatísticas. Outro problema é que muitas vezes são necessárias suposições sobre a distribuição dos dados que não podem ser verificadas. Como os testes de permutação são realizados sob os próprios dados, por meio do método de Monte Carlo, não são feitas suposições sobre a distribuição dos dados. Em geral, supomos que as observações são independentes e identicamente distribuídas.

O teste de permutação é um dos métodos, baseados em reamostragens, mais utilizados para fazer testes de hipóteses. A fim de simular a hipótese nula, de que conjuntos de dados vieram da mesma distribuição, o método mistura os conjuntos e constrói novos conjuntos a partir de sorteios sem reposição. O mesmo resultado pode ser obtido permutando os rótulos que indicam a que conjunto cada observação pertence. Inicialmente Fisher (1935) descreveu a ideia de teste de permutação, propondo realizar o teste de hipóteses utilizando todas as permutações possíveis dos rótulos das observações. Porém, Dwass (1957) propôs o uso de um subconjunto obtido aleatoriamente a partir do conjunto de

todas as permutações. A proposta de Dwass facilita a realização do teste de hipóteses, já que, para grandes quantidades de amostras, o número de permutações totais pode tornar o processamento custoso e inviabilizá-lo. Esse subconjunto é geralmente obtido pelo método de Monte Carlo, onde simula-se computacionalmente a geração aleatória (com distribuição uniforme) das permutações, gerando amostras pseudoaleatórias. No presente trabalho utilizaremos sempre o método de Monte Carlo.

Dados uma estatística θ que mede a diferença entre múltiplos grafos de correlação, um número N de permutações e r conjuntos de amostras de cada estado biológico, que formará cada rede, os testes de permutação utilizados neste trabalho consistem nos seguintes passos.

1. Para cada estado biológico P_i , onde $1 \leq i \leq r$, construa o grafo G_i utilizando um método para medir correlação.
2. Calcule o valor de $\theta(G_1, G_2, \dots, G_r)$.
3. Faça N permutações de Monte Carlo dos rótulos dos experimentos (população a que o experimento pertence) e, para cada permutação, repita os passos 1 e 2 com os novos dados obtidos.
4. O p-valor do teste é $(1+E/1+N)$, onde E é o número de estatísticas de Monte Carlo maiores ou iguais à estatística observada originalmente.

2.3 Conjuntos de dados

Usaremos dois conjuntos de dados ao longo deste trabalho. O primeiro será usado tanto nas simulações (seção 4) quanto no estudo de caso (seção 5.1). Já o segundo, usaremos apenas como estudo de caso (seção 5.2), a fim de mostrar a diversidade de dados que podem ser analisados pelo teste estatístico proposto.

2.3.1 Gliomas

Os dados de expressão gênica foram obtidos a partir de quatro tecidos tumorais da classe glioma que são chamados de Oligodendroglioma, Astrocitoma, Oligoastrocitoma e Glioblastoma. O glioma é um tipo de tumor de cérebro, que ocorre em células da glia, tecido responsável por proteger e nutrir os neurônios. Os dados são provenientes do trabalho de Kinker et al. (2016), onde foram selecionados 134 genes alvo do fator de transcrição NFκB. A família dos fatores de transcrição NFκB

tem papel chave em diversos processos como inflamação, imunidade inata, proliferação celular e apoptose celular (Kinker et al. 2016). Os níveis de expressão desses genes foram medidos em 194 pacientes identificados com Astrocitoma, 189 com Oligodendroglioma, 129 com Oligoastrocitoma (todos com graus II ou III) e 146 de Glioblastoma multiforme primário, (GBM) de grau IV. A fim de aproximar a distribuição dos valores de expressão dos genes de uma distribuição normal, transformamos esses valores pelo respectivo log na base 2.

2.3.2 Metabolismo de Sorgo

Um segundo exemplo é o estudo de caso com dados de metabolômica de Sorgo, um organismo vegetal com importância agrícola. Os dados foram produzidos no trabalho de De Souza et al. (2015). Utilizamos dados de um grupo de plantas crescidas sob efeito de seca (com a redução da capacidade de campo) obtidos em duas coletas, de três plantas cada, aos 90 e 120 dias após o plantio. De Souza et al. (2015) mediram a concentração de 79 metabólitos do metabolismo primário do sorgo, em cada coleta, de cinco órgãos da planta, folha, colmo, raiz, raiz escora e grãos. A Figura 6, adaptada de De Souza et al. (2015) mostra uma foto do experimento realizado em laboratório e um resumo esquemático do mesmo.

Figura 6 - Esquema resumido do experimento de De Souza et al. (2015). Resumo esquemático do experimento realizado por De Souza et al. (2015). Aos 60 dias foi dado início ao tratamento de seca em todas as plantas reduzindo de 1,5L a 0,45L a quantidade de água fornecida. As coletas foram feitas aos 90 e 120 dias após o plantio.

2.4 Avaliação de desempenho do método

Com o objetivo de avaliar o desempenho dos métodos propostos para análise diferencial de redes realizamos simulações e testamos o tempo de execução, com dados de expressão gênica de gliomas (seção 2.3.1).

2.4.1 Simulações

Utilizamos a técnica de *bootstrap* para simular novos conjuntos de dados a partir dos dados sob estudo, com a hipótese de que tanto os dados de gliomas (astrocitoma, oligoastrocitoma, oligodendroglioma e glioblastoma) quanto os dados dos cinco órgãos de sorgo (folha, colmo, raiz, raiz escora e grãos) são oriundos da mesma distribuição (hipótese nula). Efron (1979) propôs a técnica de *bootstrap* para estimar a distribuição amostral de uma estatística a partir de reamostragens aleatórias com reposição dos dados originais. Existe uma gama de aplicações para essa técnica. Por exemplo, ela é usada como estimador para o erro padrão de uma estatística e intervalos de confiança para parâmetros de uma população, além da construção de testes de hipóteses.

A técnica do *bootstrap* se baseia na ideia de que, em geral, a amostra coletada é a melhor estimativa da população que a originou. Se assumirmos que as variáveis aleatórias são independentes, podemos implementar o *bootstrap* construindo reamostragens aleatórias do conjunto de observações, de mesmo tamanho do conjunto original. Construimos cada reamostragem a partir do sorteio com reposição dos elementos do conjunto de observações.

Usamos o procedimento de *bootstrap* para gerar novos conjuntos de dados sob a hipótese nula semelhante ao realizado por Santos (2015). Onde:

1. Sejam P um conjunto com n_C unidades amostrais, r o número de conjuntos a serem sorteados e n_j o número de unidades amostrais da população P_j , para $1 \leq j \leq r$, tais que $n_j \leq n_C$.
2. Dados um número $N_{boot} = 1.000$ de reamostragens *bootstrap* e um conjunto V com n_V variáveis presentes nas amostras de P .
3. Construa o conjunto P_j^i sorteando aleatoriamente, com reposição, n_j amostras de P , para $1 \leq i \leq N_{boot}$ e $1 \leq j \leq r$.

Para os dados de gliomas, P é o conjunto de 658 unidades amostrais de tecidos gliais e para o conjunto de dados de metabolismo de sorgo, P é o conjunto de 30 unidades amostrais. Note que os dados dos quatro tipos de células cancerígenas estão misturados nos cenários construídos bem como os dados dos cinco órgãos de sorgo. O resultado esperado é que todos os P_j^i representem amostras de uma mesma população (P). Assim, o cenário apresentado simula a hipótese nula dos testes propostos.

As reamostragens obtidas pelo procedimento de *bootstrap* permitiram avaliar o controle da taxa de falsos positivos dos testes propostos, como mostramos na seção 2.4.1.a. Realizamos também simulações para avaliar poder estatístico dos métodos, onde construímos conjuntos de dados sobre H_1 , conforme descrito na seção 2.4.1.b.

2.4.1.a Controle da taxa de falsos positivos

Nas simulações realizadas, a taxa de falsos positivos é a proporção de vezes que a hipótese nula é rejeitada quando ela é verdadeira. Dada uma estatística θ para comparar a estrutura de múltiplos grafos, testamos $H_0 : \theta = 0$ versus $H_1 : \theta > 0$. Verificamos se os oito métodos (distribuição e entropia espectrais, distribuição dos graus, centralidades de grau, *closeness*, *betweenness* e autovetor e coeficiente de agrupamento) controlam a taxa de falsos positivos, tanto com os dados de glioma (seção 2.3.1) quanto com os dados de metabolismo de sorgo (seção 2.3.2). Para cada conjunto de dados gerado pelo procedimento bootstrap, nós aplicamos o teste estatístico proposto (descrito da seção 2.2.5) usando 1000 permutações aleatórias.

Com os dados de glioma, o método de análise diferencial de redes foi testado para uma coleção de quatro conjuntos P ($r = 4$), com $n_1=194$ unidades amostrais (pacientes com Astrocitoma), $n_2=189$ (Oligodendroglioma), $n_3=129$ (Oligoastrocitoma), $n_4=146$ (Glioblastoma), e $n_v=134$ genes. As redes obtidas a partir dos dados de glioma são grafos completos em que o peso de cada aresta corresponde a 1 menos o p-valor do teste de correlação de Spearman. Para os dados de metabolismo de sorgo testamos o método com uma coleção de cinco conjuntos P ($r = 5$), com $n_j=6$, para $1 \leq j \leq 5$, unidades amostrais de cada órgão em cada conjunto P^i , e $n_v=73$ metabólitos. No caso do metabolismo de sorgo, como o conjunto de dados é relativamente pequeno, utilizamos a correlação de Pearson no lugar da correlação de Spearman, já que o poder estatístico da primeira é ligeiramente superior ao da segunda quando a associação entre os dados é linear (Santos et al. 2013). Além disso, foram removidas as amostras com peso menor do que 0,75, pois para garantir relações fortes entre as variáveis (Caldana et al. 2011; Ding et al. 2015; Jeong et al. 2001). Já que os valores das correlações foram usados como limiar para a construção das arestas, utilizamos o próprio valor de correlação como peso nas arestas da rede, ao invés de 1 menos o p-valor dessa correlação. Dessa maneira, geramos amostras com os mesmos tamanhos dos dados originais e podemos verificar o controle da taxa de falsos positivos em condições semelhantes.

2.4.1.b Poder estatístico

A fim de avaliar o poder estatístico dos métodos, isto é, a taxa de verdadeiros positivos (proporção de vezes que a hipótese nula é rejeitada corretamente), construímos conjuntos de dados sob a hipótese alternativa (H_1), conforme descrito a seguir:

Dados um número n_v de genes, um número N_{boot} de reamostragens e um parâmetro γ :

1. Calcule a correlação entre os vértices de V .
2. Ordene os pares de genes pela correlação.
3. Inspeção os pares dos mais correlacionados para os menos correlacionados, escolhendo um vértice de cada par de forma a obter, no fim do processo, γn_V vértices.
4. Com os vértices selecionados construa o conjunto V' .
5. Faça N_{boot} vezes:
 - (a) Gere r conjuntos P_j , onde $1 \leq j \leq r$, com n amostras cada, conforme cenário construído sob a hipótese nula.
 - (b) Permute aleatoriamente as unidades amostrais de cada variável (vértice) do conjunto V' , em apenas um dos conjuntos P_j .

Assim como no cenário para verificar o controle da taxa de falsos positivos, o procedimento acima, no passo 5(a), gera r conjuntos sob a hipótese nula. Contudo, ao permutarmos as unidades amostrais das variáveis com as maiores correlações do conjunto de dados, no passo 5 (b), várias correlações são “quebradas” em um dos conjuntos P_j , fazendo com que uma das redes seja diferente das demais.

Para os testes de poder estatístico, diferentes proporções de genes (parâmetro γ) foram alteradas. Sob H_1 , utilizamos $\gamma = 0, 0,05, 0,1, 0,2, 0,3$ e $0,5$. Com $\gamma = 0$ temos os resultados sob H_0 , onde todos os conjuntos são provenientes da mesma população, não alterando nenhum dos conjuntos de P , o que é equivalente ao teste de controle da taxa de falsos positivos. Em cada conjunto criado, aplicamos 1000 permutações, segundo os testes de permutação descritos na seção 2.2.5.

Para avaliação do poder estatístico (seção 4.2), as análises foram realizadas com o objetivo de modelar a rede de co-expressão de duas formas. Na primeira, o grafo será não-dirigido, sem peso nas arestas e, na segunda, será não-dirigido e com peso nas arestas. Seja p_{sp} o p-valor do teste de Spearman (1904) entre os níveis de expressão de dois genes. Para redes sem peso, todas as arestas com o peso $1 - p_{sp}$ menor ou igual a 0,95 (isto é, testes com menos de 5% de significância) são removidas e as arestas remanescentes no grafo recebem peso igual a 1. Já para as redes com peso nas arestas, o peso de cada uma será $1 - p_{sp}$.

O método de análise diferencial de redes baseado na distribuição espectral foi testado em 14 cenários diferentes. Inicialmente construímos seis cenários, fixando o tamanho das redes em $n_v = 50$ genes e o número de unidades amostrais em $n = 120$, variando apenas a quantidade de conjuntos usados na coleção P ($r = 2, 3, 5, 10, 15, 20$). Em seguida fixamos o número de conjuntos de P ($r = 3$) e o

tamanho das redes ($n_v = 50$) variando o número de unidades amostrais ($n = 3, 6, 15, 40, 120$). Por fim, mantivemos o número de conjuntos P ($r = 3$) e $n = 120$ unidades amostrais, variando o tamanho das redes ($n_v = 10, 50, 130$).

Os outros sete métodos de análise diferencial de redes baseados nas medidas estruturais restantes (entropia espectral, distribuição dos graus, centralidades de grau, *closeness*, *betweenness* e autovetor e coeficiente de agrupamento) foram testados apenas no cenário com três conjuntos dentro da coleção P ($r = 3$), $n = 120$ unidades amostrais e $n_v = 50$ genes.

2.4.1.c Curvas ROC

A fim de avaliar, tanto o controle da taxa de falsos positivos quanto o poder estatístico dos métodos, construímos curvas ROC (*Receiver Operating Characteristic* ou em português: Característica de Operação do Receptor) com os p-valores dos testes aplicados nos $N_{boot} = 1.000$ conjuntos de dados simulados. As curvas construídas indicam, no eixo x, o nível de significância dos testes e no eixo y, a proporção de vezes que a hipótese nula foi rejeitada (Figura 7). Sob H_0 , o eixo x corresponde ao número esperado de falsos positivos e o eixo y corresponde a taxa observada de falsos positivos. Já sob H_1 , o eixo y corresponde ao poder estatístico empírico para cada nível de significância. Para sumarizar os resultados das curvas ROC, utilizamos a área sob cada curva ROC (AUC, do inglês *Area Under Curve*), que deverá ser próxima ou menor do que 0,5 quando os testes controlam a taxa de falsos positivos (Figura 7 a esquerda). Já para a análise do poder estatístico, áreas próximas de 0,5 equivalem a decisões aleatórias de rejeitar ou aceitar a hipótese nula. Quanto mais próxima a área sob a curva ROC estiver de 1 (Figura 7 a direita), maior o poder estatístico do teste.

Figura 7: Curvas ROC utilizadas para os cálculos tanto da taxa de falsos positivos quanto do poder estatístico. A esquerda vemos uma curva ROC com Area Sob a Curva próximo a 0,5, resultado esperado no teste de controle da taxa de falsos positivos. Já o gráfico à direita representa uma curva ROC com AUC próximo a um, esse resultado significa que o método testado teve bom desempenho em rejeitar corretamente a hipótese nula.

2.4.2 Teste de tempo de execução

Para verificar se existe ganho no tempo de execução do teste de múltiplas redes, se comparado com o tempo total de execução do teste das redes par a par, construímos conjuntos P_j (número de redes), onde $1 \leq j \leq r$, de diferentes tamanhos para serem analisados. Utilizamos o método descrito na seção 2.4.1, para construir os conjuntos simulados de dados. Cada conjunto contém $n_v=50$ vértices e $n=120$ amostras. Foram consideradas diferentes quantidades de conjuntos: $r = \{2, 3, 5, 7, 10, 13, 15, 18, 20\}$. Comparamos o tempo de execução do teste de múltiplas redes, chamado de t_{BNS} , com o tempo de execução de x testes para comparar os conjuntos par a par, chamado de t_{par} . Nesse caso, x é o número de combinações, onde o total de elementos é r e o número de elementos escolhidos é dois

$$x = \binom{r}{2} = \frac{r(r-1)}{2}.$$

A razão R_a entre t_{BNS} e t_{par} é a proporção de vezes que o teste de múltiplas redes é mais rápido que a realização dos testes par a par.

3 *BioNetStat: Biological Network*

Statistical analysis

O R (<http://cran.r-project.org/>) é uma linguagem de programação e um ambiente de desenvolvimento integrado muito utilizado para analisar dados e para criar programas (conhecido como pacotes) para análise estatística. Devido à grande variedade de métodos estatísticos para analisar dados biológicos implementados em R e disponíveis como software livre em grandes repositórios como o CRAN (<http://cran.r-project.org/>) e o Bioconductor (<http://www.bioconductor.org/>), o ambiente é amplamente utilizado pela comunidade de Bioinformática.

Assim, a fim de disponibilizar os testes estatísticos deste trabalho para a comunidade científica, especialmente de Bioinformática, desenvolvemos um pacote de R chamado *BioNetStat* (*BNS*) sob a licença de software livre GNU GPL (*GNU General Public License*) v3. O software *BNS* pode ser utilizado por meio de linhas de comando escritas em linguagem R. Os arquivos de instalação, tutorial, e o guia de usuário estarão disponíveis no repositório git no endereço <http://github.com/jardimViniciusC/BioNetStat>. Nas próximas seções descreveremos as funcionalidades do pacote, suas principais funções e sua instalação.

3.1.1 Descrição

O *BioNetStat* (*BNS*) é um pacote criado para o estudo de redes de correlação. A principal funcionalidade do *BNS* é a análise diferencial de redes. Para tanto, a ferramenta precisa receber um conjunto de dados que contenha medidas de variáveis biológicas (por exemplo, expressão gênica, concentração de proteínas ou metabólitos, medidas fenotípicas, fisiológicas, ecológicas, entre outras). Além disso, é necessário informar ao programa quais são os grupos de variáveis (por exemplo, genes, proteínas ou metabólitos que participam de uma mesma via metabólica ou parâmetros que estejam ligados ao mesmo processo fisiológico ou ecológico). A partir desses dados, a ferramenta é capaz de comparar estatisticamente as redes de correlação formadas por um conjunto de variáveis. O pacote também realiza a análise diferencial de vértices, descrita na seção 2.2.4.b. Além disso, é possível visualizar as redes na forma de grafos ou então visualizar os genes, proteínas e metabólitos em mapas metabólicos do Banco de Dados KEGG. Nos próximos parágrafos vamos descrever as entradas, as funcionalidades e as saídas do pacote proposto que estão apresentadas e resumidas na Figura 8.

Figura 8 - Esquema de funcionamento do BioNetStat. Sequências de passos realizados para chegar à análise diferencial de redes e posteriormente análise diferencial de vértices, desde a obtenção da tabela de dados até a tabela de saída com o valor da estatística θ . Após a obtenção da lista de variáveis diferencialmente correlacionadas, é possível obter também os mapas metabólicos com as respectivas variáveis em destaque. Os nomes em *itálico* dentro das caixas se referem às funções do BioNetStat que realizam as tarefas mostradas na figura.

Entrada e Funções

O *BioNetStat* recebe dois arquivos como entrada, o arquivo *valores das variáveis* (Figura 8) e o arquivo *grupos de variáveis*. O formato desses arquivos e as funções construídas para que eles sejam lidos e analisados estão descritos na Tabela 2. Uma explicação mais aprofundada do funcionamento e utilização do pacote estão no repositório do Bioconductor no link a seguir: <http://bioconductor.org/packages/BioNetStat/>.

É importante destacar que o pré-tratamento dos dados inseridos não são realizados pelo pacote *BNS*, ou seja, a verificação das premissas necessárias para a realização dos testes de correlação deve ser realizada pelo usuário. Como exemplo, a correlação de Pearson tem como premissa que a distribuição dos dados é normal. Para que essa premissa seja atendida, muitas vezes, é necessário, aplicar uma transformação logarítmica nos valores de expressão para que eles adquiram uma distribuição próxima à normal.

Tabela 2: Funções implementadas no pacote *BioNetStat* e qual o seu papel na análise diferencial de redes apresentada na figura esquemática do programa (Figura 8).

Função	Papel
<code>readVarFile</code>	Lê o arquivo <i>valores das variáveis</i> e seleciona as colunas caracterizadas como <i>numeric</i> . Define os nós (variáveis) da rede
<code>doLabels</code>	Lê o arquivo <i>valores das variáveis</i> e seleciona as colunas caracterizadas como <i>factor</i> . Define o número de redes (estados) comparadas
<code>readGmtFile</code>	Lê o arquivo <i>grupo de variáveis</i> e seleciona quais variáveis (nós) estarão nas redes comparadas
<code>adjacencyMatrix</code>	Função que constrói a matriz de adjacência de acordo com os parâmetros definidos
<code>diffNetAnalysis</code>	Função que realiza comparação de redes e usa umas das funções <i>test</i> abaixo
<code>nodeDegreeCentralityTest</code>	Realiza a comparação de redes de acordo com a centralidade de grau
<code>nodeBetweennessCentralityTest</code>	Realiza a comparação de redes de acordo com a centralidade de <i>betweenness</i>
<code>nodeClosenessCentralityTestVector</code>	Realiza a comparação de redes de acordo com a centralidade de <i>closeness</i>
<code>nodeEigenvectorCentralityTestVector</code>	Realiza a comparação de redes de acordo com a centralidade de autovetor
<code>nodeClusteringCoefficientTestVector</code>	Realiza a comparação de redes de acordo com o coeficiente de agrupamento
<code>genePathPlot</code>	Gera uma figura que mostra, no mapa metabólico do KEGG, os valores de expressão ou centralidade dos genes
<code>cdpPathPlot</code>	Gera uma figura que mostra, no mapa metabólico do KEGG, os valores de concentração ou centralidade dos metabólitos

3.1.2 Implementação

O pacote *BioNetStat* requer os seguintes pacotes para ser executado: *igraph* para calcular as propriedades estruturais das redes; *WGCNA* para converter a matriz de expressão genética em uma matriz com uma linha por gene; *ggplot2*, *pheatmap* e *RcolorBrewer* para produzir os gráficos; *pathview* para a visualização dos dados em vias metabólicas; e *Hmisc* e *psych* para a construção dos grafos. Esses pacotes são instalados concomitantemente a instalação do *BioNetStat*.

A descrição das funções e pacotes utilizados para implementar as estatísticas dos testes entre as

redes de correlação são mostradas na Tabela 3. A estimativa das funções de densidade de probabilidade em estatísticas baseadas na distribuição do espectro, na entropia espectral e na distribuição dos graus é feita utilizando o Kernel Gaussiano implementado pela função *density* do pacote *stats* do R, com largura de banda dada por $h = (\max(x) - \min(x))/k$ (Sain et al. 1996), onde $k = [\log_2 n_v + 1]$ (Sturges 1926) e x é um vetor de características do grafo. Se considerarmos as características espectrais, distribuição e entropia, o vetor x representa os autovalores da matriz de adjacência, já para a distribuição dos graus, x corresponde aos graus dos vértices do grafo.

Podemos interpretar o estimador da função de densidade de probabilidade baseado no Kernel Gaussiano como uma versão suavizada de um histograma. Dadas n_v observações, cada observação x_i contribui para estimar a função em um ponto x_0 de acordo com a diferença entre x_i e x_0 . A contribuição é ponderada pela função do Kernel (K) e depende de um parâmetro conhecido como largura de banda (h), que controla o tamanho da vizinhança ao redor de x_0 . Formalmente, a função de densidade estimada em um ponto qualquer x é:

$$\hat{f}(x) = \frac{1}{n_v} \sum_{i=1}^{n_v} K\left(\frac{x - x_i}{h}\right),$$

onde

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}.$$

Tabela 3 - Funções utilizadas em R para a implementação dos testes estatísticos entre as redes de correlação disponíveis no BioNetStat.

Estatística	Grafos com peso	Grafos sem peso	Funções do R
Divergência de Kullback-Libler entre densidades espectrais	Sim	Sim	Função <i>density</i> do pacote <i>base</i>
Diferença absoluta entre as entropias espectrais	Sim	Sim	Função <i>density</i> do pacote <i>base</i>
Divergência de Kullback-Libler entre densidades dos graus	Sim	Sim	Função <i>density</i> do pacote <i>base</i> função <i>graph.strength</i> do pacote <i>igraph</i>
Distância euclidiana entre as centralidades de grau	Sim	Sim	Função <i>graph.strength</i> do pacote <i>igraph</i>
Distância euclidiana entre as centralidades de <i>betweenness</i>	Sim	Sim	Função <i>betweenness</i> do pacote <i>igraph</i>
Distância euclidiana entre as centralidades de <i>closeness</i>	Sim	Sim	Função <i>closeness</i> do pacote <i>igraph</i>
Distância euclidiana entre as centralidades de autovetor	Sim	Sim	Função <i>evcent</i> do pacote <i>igraph</i>
Distância euclidiana entre os coeficientes de agrupamento	Sim	Sim	Função <i>transitivity</i> do pacote <i>igraph</i>

Note que assumimos $h > 0$. Pelo critério de Sturges, adotado neste trabalho para definir a largura de banda, h vale zero sempre que o grafo for vazio. Desta forma, quando são gerados grafos vazios no teste estatístico, o usuário é notificado pelo programa de que o teste não poderá ser realizado com o critério de Sturges para a escolha da largura de banda.

4 Desempenho dos métodos

4.1 Controle da taxa de falsos positivos

A partir dos experimentos de simulação realizados com os oito métodos de análise diferencial de redes (seção 2.4.1.a) verificamos que a taxa de falsos positivos é controlada tanto no conjunto de dados de glioma quanto no metabolismo de sorgo. Isto pode ser verificado na Figura 9, onde as curvas ROC na diagonal indicam valores de área embaixo da curva ROC (AUC) próximos a 0,5 para todos os testes.

Figura 9 - Curvas ROC sob a hipótese nula para cada um dos métodos estatísticos. Os testes foram realizados em grafos com peso para as oito medidas de características da rede, discriminadas na legenda, e para os dois tipos de dados (Glioma e Metabolismo de sorgo).

4.2 Poder estatístico

As áreas embaixo das curvas (AUC) ROC para a análise diferencial de redes com e sem peso

estão apresentadas nas Figuras 10. Os resultados nos permitem concluir que todos os métodos controlam a taxa de falsos positivos ao compararmos três redes, 50 vértices e 120 unidades amostrais em nossos experimentos de simulação, apresentando o AUC próximo a 0,5 na coluna onde $\gamma = 0$ (Figura 10). Além disso, as áreas sob a curva (AUC) crescem com o aumento de γ (proporção de genes alterados). Isto é, o poder estatístico aumenta com a diferença entre as redes. Os testes que apresentaram maior poder estatístico foram os das centralidades de grau e autovetor. Já os testes com menor poder estatístico foram os da centralidade de *closeness* e o da entropia.

Figura 10 - Heatmaps de áreas sob a curva ROC (do inglês, AUC), mostrando os resultados dos experimentos de simulações para o teste diferencial de redes baseado nas oito características das redes. Foram comparadas três redes ($r=3$) com $n_v=50$ vértices e $n=120$ amostra cada rede, em redes com peso (A) e em redes sem peso (B). Os resultados das simulações sob a hipótese nula para γ (proporção de variáveis alteradas) igual a zero estão na primeira coluna de cada heatmap. As demais colunas, com γ variando entre 0,05 e 0,5 apresentam os resultados das simulações sob a hipótese alternativa. Nas linhas dos heatmaps apresentamos as características que foram usadas para realizar os testes de distribuição espectral, entropia espectral, distribuição de grau, centralidade de grau, autovetor, betweenness e closeness e coeficiente de agrupamento.

Alterando os parâmetros das simulações realizadas com o método de distribuição espectral observamos que conforme aumenta o número de redes testadas o poder estatístico é reduzido ao fixarmos um determinado valor de γ , para redes com peso (Figura 11A(a)) e sem peso (Figura 11A(b)). Quando aumentamos o número de unidades amostrais, de 3 até 120, e fixamos o número de vértices e de redes observamos o crescimento do poder estatístico para todos os valores de γ (Figuras 11B(a) e 11B(b)). É importante ressaltar que com três unidades amostrais o poder empírico do teste é equivalente a aceitar ou rejeitar a hipótese nula ao acaso (proporção esperada de rejeições da hipótese nula feitas ao acaso), ou seja, ele não é capaz de perceber a diferença entre as redes (Apêndice 21B). Considerando conjuntos de dados com seis unidades amostrais a área debaixo da curva ROC se aproxima de 0,6 quando $\gamma=0,5$, ou seja, com $n = 6$ é possível identificar apenas grandes diferenças

entre as redes (Apêndice 21B). Por fim, podemos dizer que o aumento da quantidade de vértices, de dez para 50, aumenta o poder estatístico do método. Porém, quando aumentamos o número de vértices de 50 para 130 o poder estatístico se mantém, indicando uma convergência a partir de um determinado número de vértices (Figura 11C).

Figura 11 - Resultados dos experimentos de simulações para verificar o poder estatístico da análise diferencial de redes baseada na distribuição espectral, em redes com peso (a) e redes sem peso (b). Os gráficos mostram as áreas sob a curva ROC (do inglês, AUC) obtidas as simulações sob a hipótese alternativa com γ (proporção de variáveis alteradas) variando entre 0,05 e 0,5. Em **A** são apresentados os resultados para os testes variando o número r de conjuntos (redes), em 2, 3, 5, 10, 15 e 20, mantendo fixos os parâmetros de número de vértices $n_v=50$ vértices e número de amostras $n=120$. Em **B** estão os resultados para os testes com três redes ($r=3$) e 50 vértices, variando o número de amostras em 3, 6, 15, 40 e 120. Para as figuras **C**, são apresentados os resultados dos testes com três redes ($r=3$) e 120 amostras, variando o número de vértices em 10, 15 e 130.

4.3 Comparação do tempo de execução com testes por pares

A comparação entre o tempo de execução do teste de múltiplas redes (t_{BNS}) e os testes par a par (t_{par}) nos permite dizer que, fixando o número de vértices ($n_v=50$) e amostras ($n = 120$ amostras por rede), o tempo de execução da análise par a par cresce quadraticamente com o número de redes (Figura 12). Por exemplo, quando aumentamos cinco vezes o número de redes (de duas para dez redes comparadas), são realizadas 45 comparações, conseqüentemente, o tempo de execução par a par fica 46 vezes maior. Já para o teste de múltiplas redes, o aumento é linear. Com um aumento de cinco vezes no número de redes o tempo de execução fica 5,2 vezes maior. A análise de múltiplas redes realiza o teste de comparação entre r redes, $r-1$ vezes mais rápido que o teste par a par. Os valores dos tempos de execução estão apresentados também na Tabela 15.

Figura 12 - Comparação do tempo de execução entre o teste de múltiplas redes (t_{BNS}) com os testes par a par (t_{par}). Os círculos indicam o tempo de execução total dos testes par a par (t_{par}) para comparar r redes. Os asteriscos indicam o tempo de execução do teste de múltiplas redes (t_{BNS}) para comparar as mesmas r redes. Os segmentos ligando os círculos e os asteriscos e os números seguidos por “x” indicam quantas vezes o tempo de execução do teste par a par é maior que o teste de múltiplas redes, para o mesmo número de redes.

4.4 Discussão

Considerando que as simulações realizadas com os métodos de análise diferencial de redes controlam as taxas de falsos positivos para dados com muitas (Gliomas) e poucas (Sorgo) unidades amostrais, podemos dizer que em nenhum dos casos a hipótese alternativa, de que pelo menos uma das redes é diferente, é aceita com uma porcentagem de erro acima da esperada. Consideramos então que os métodos podem ser aplicados também em conjuntos de dados com poucas (seis) unidades amostrais. Apesar de controlar a taxa de falsos positivos, o poder estatístico dos métodos com conjuntos de dados que possuem poucas (seis) unidades amostrais é menor, sendo que com três unidades o método não é capaz de detectar diferenças entre as redes. Assim como é comum conjuntos de dados relacionados à saúde humana (dados de gliomas) serem dados com grandes quantidades de unidades amostrais, em estudos de plantas é comum encontrar pequenas quantidades de unidades amostrais, como seis ou até mesmo três unidades. Portanto, os métodos apresentados aqui não são restritos aos dados com grandes quantidades de amostras, podendo ser usados em dados com pelo menos seis unidades amostrais.

Outro fator que limita o poder estatístico do método é o número de vértices da rede, usando poucos vértices as diferenças entre as redes tendem a ser menores e mais difíceis de serem identificadas pelos métodos. Apesar de ser um caso mais raro, a comparação de uma grande quantidade de redes, por volta de 15, também dificulta que o método encontre diferenças entre as redes.

Os testes de tempo de execução deixaram claro o ganho de tempo na análise diferencial de múltiplas redes em relação ao teste par-a-par. Além do tempo de execução, existe uma diferença grande no tempo de manipulação das tabelas de saída das duas análises. Os resultados de comparações par a par geram uma tabela para cada um dos pares comparados. Desse modo, utilizando o exemplo em que comparamos dez redes, teríamos 45 tabelas geradas a partir dessa análise, o que teria um custo considerável de tempo para analisar, principalmente sem o auxílio de programação. Por outro lado, o *BNS* resultaria em apenas uma tabela indicando se existe, ou não, pelo menos uma rede diferente dentre as dez comparadas, associando um peso para cada rede. Além disto, o *BNS* ainda aponta quais são as principais variáveis nas diferenças entre as redes, através da análise diferencial de vértices.

Em suma, podemos concluir que os testes realizados pelo *BNS* satisfazem as propriedades fundamentais de um teste estatístico: o poder estatístico aumenta com o tamanho amostral e a taxa de erro do tipo I (falsos positivos) é controlada. Além disso, é uma ferramenta que torna mais eficiente a comparação de mais de duas redes por reduzir consideravelmente o tempo de análise.

5 Estudos de Caso

Em ambos os conjuntos de dados (Gliomas e Metabolismo de sorgo) fizemos os testes de análise diferencial de redes com as oito medidas de características de redes (distribuição e entropia espectrais, distribuição dos graus, quatro medidas centralidades e o coeficiente de agrupamento). Para este trabalho, propomos uma separação dessas oito medidas em dois grupos, a fim de facilitar a descrição das análises. Aplicamos o termo “características globais da rede” para nos referirmos a medidas que são atribuídas à rede como um todo e que não podem ser dadas para cada vértice individualmente. As características globais são: distribuição espectral, entropia espectral e a distribuição dos graus. Já as medidas que podemos atribuir um valor a cada vértice foram chamadas de “características locais da rede”. Neste trabalho, essas medidas são quatro centralidades (grau, *closeness*, *betweenness* e autovetor) e o coeficiente de agrupamento. Nas seções seguintes nós exemplificaremos o uso do software *BioNetStat* utilizando os conjuntos de dados descritos na seção 2.3.

5.1 Gliomas

Gliomas são tumores cerebrais infiltrativos que ocorrem nas células da glia, células que compõem o sistema nervoso junto com os neurônios. A glia é o tecido responsável pela manutenção da homeostase do tecido neural para o bom funcionamento dos neurônios (Purves et al. 2001). Dentre os tumores classificados como Gliomas, os de baixo grau de agressividade (II e III) são Oligodendroglioma, Oligoastrocitoma e Astrocitoma. O Glioblastoma é o mais agressivo, de grau IV. De acordo com a patologia molecular, os gliomas são identificados por características histológicas em conjunto com marcadores genéticos (Louis et al. 2016). A patologia molecular (estudo do diagnóstico de doenças utilizando dados moleculares) ajuda, ainda, a tornar mais específico o prognóstico dos pacientes e leva ao desenvolvimento de terapias mais personalizadas para cada tipo de tumor (tipo cancerígeno). A organização Mundial da Saúde aceita o diagnóstico de gliomas e a classificação do tipo de glioma baseados em patologia molecular (Louis et al. 2016). O status do gene IDH e a co-deleção das regiões cromossômicas *1p/19q* são os fatores moleculares que levam a uma classificação mais acurada do tipo cancerígeno. Nesse sentido, é importante que mais características moleculares pudessem auxiliar na definição do subtipo celular e de outros tipos mais específicos que possibilitem tratamentos mais direcionados e eficazes (Louis et al. 2016).

Os dados usados nas próximas seções estão descritos na seção 2.3.1. Agrupamos os 134 genes segundo uma coleção de 1329 vias canônicas do MsigDB v5.0 (Molecular signature Database) (Subramanian et al. 2005). Cada via corresponde a um conjunto de genes, que usualmente representa um processo biológico. Foram analisadas apenas as vias que continham pelo menos 10 genes representados na lista de 134 genes analisados. Dessa forma, 16 grupos, dos 1329, foram selecionados para teste. Construímos a rede de correlação (co-expressão) entre os genes realizando testes de correlação de Spearman entre os valores de expressão dos pares de genes, para cada um dos quatro tipos celulares. Escolhemos a medida de Spearman, pois ela apresenta elevado poder estatístico em associações lineares e monotônicas, que são as mais frequentemente investigadas nos estudos de expressão genética. Além disso, essa medida é robusta à presença de *outliers*, que são comuns em dados biológicos. O q-valor, p-valor ajustado por FDR para a correção por múltiplos testes (Benjamini e Hochberg 1995), de 0,05 foi usado como limiar máximo na formação das conexões entre os vértices. Utilizamos o valor de um menos o p-valor nominal do teste de correlação como peso da aresta que liga dois vértices.

As redes de cada um dos tipos cancerígenos (Astrocitoma, Oligoastrocitoma, Oligodendroglioma e Glioblastoma Multiforme), considerando todos os 134 genes, estão representadas na Figura 13. O elevado número de conexões entre os genes inviabiliza a comparação manual entre os quatro tipos de glioma. Vemos, assim, que ferramentas computacionais são essenciais para a análise do sistema estudado.

Nas próximas seções, apresentamos a análise diferencial de múltiplas redes, realizada para cada um dos 16 grupos de genes, comparando a conectividade das redes dos quatro tecidos cancerígenos. Em cada teste realizamos 1.000 permutações aleatórias dos rótulos dos tecidos cancerígenos. Primeiramente, a análise foi feita com base nas características globais e, posteriormente, foram feitos os testes baseados nas características locais. Realizamos, ainda, a análise diferencial de vértices, para características locais. A fim de aprofundar o estudo sobre o papel dos genes investigados e, principalmente, daquele grupo de genes que tiveram co-expressão diferencial em, pelo menos, um dos tecidos, o BioNetStat apresenta a possibilidade de visualizar a expressão ou a centralidade dos genes nas redes, representadas por cores em mapas metabólicos do KEGG.

Figura 13 - Redes de correlação construídas com os 134 genes alvos do fator de transcrição NF κ B para os quatro tipos cancerígenos (Astrocitoma, Oligodendroglioma, Oligoastrocitoma e Glioblastoma). Cada gene está representado por um vértice da rede e as arestas (conexões) são as correlações com q-valores associados menores que 0,05.

5.1.1 Análise diferencial de redes baseada em características globais

Os resultados das estatísticas θ para as comparações da distribuição do espectro entre os quatro tecidos cancerígenos para os 16 grupos de genes estão na Tabela 4. Os grupos de genes da base de dados NABA (*Ecm regulators*, *Secreted factors* e *Matrisome associated*), os quatro primeiros da Tabela

4 (q-valores < 0,05), podem auxiliar a diferenciar o tipo celular Oligoastrocitoma dos outros tecidos, segundo sua distribuição e entropia espectrais (Tabela 5). Já no teste para o grupo *CK CK Recp interaction*, os tecidos que apresentaram maior peso na diferença entre as redes foram o oligodendroglioma e o GBM, ambos com pesos bastante semelhantes. O tecido cancerígeno Astrocitoma é o que mais se diferencia dos outros pelo grupo *FAS Signaling pathway*, com 45% de peso na estatística θ . Por fim, a rede de GBM apresenta grande diferença das demais pelo grupo de genes *Apoptosis (KEEG)*.

Tabela 4 - Análise diferencial de redes baseada na distribuição espectral. A tabela apresenta os resultados dos testes estatísticos para os 16 grupos de genes analisados, mostrando o nome do grupo, a base de dados que definiu aquele grupo de genes, o número de redes, o número de vértices, estatística θ , o p-valor nominal e o q-valor. Nas colunas seguintes estão os pesos de cada uma das quatro redes na estatística θ : Astro (Astrocitoma), Oligoast (Oligoastrocitoma), Oligoden (Oligodendroglioma) e GBM (Glioblastoma Multiforme).

Nome do grupo	Base de dados	Redes	Vértices	Estatística θ	p-valor	q-valor	Astro	Oligoast	Oligoden	GBM
<i>ECM Regulators</i>	NABA	4	11	0,584	0,001	0,004	12,1	57,7	13,5	16,8
<i>Secreted factors</i>	NABA	4	24	0,060	0,001	0,004	17,8	42,2	15,2	24,9
<i>Matrisome associated</i>	NABA	4	43	0,031	0,001	0,004	22,2	33,3	21,5	22,9
<i>Matrisome</i>	NABA	4	44	0,030	0,001	0,004	23,9	31,1	21,0	24,0
<i>CK-CK Recp interaction</i>	KEGG	4	47	0,030	0,002	0,005	22,7	24,8	26,2	26,4
<i>FAS Signaling pathway</i>	ST	4	10	0,049	0,002	0,005	45,0	20,1	17,6	17,3
<i>Apoptosis</i>	KEGG	4	15	0,034	0,017	0,039	8,0	12,2	19,4	60,4
<i>Death pathway</i>	BIOCARTA	4	11	0,036	0,032	0,064	26,5	20,4	9,1	44,1
<i>Small cell lung cancer</i>	KEGG	4	12	0,021	0,081	0,144	13,8	16,5	21,7	47,9
<i>HIVNEF pathway</i>	BIOCARTA	4	10	0,022	0,186	0,297	15,3	3,9	7,3	73,4
<i>P53 Downstream pathway</i>	PID	4	10	0,023	0,219	0,304	51,4	23,0	7,9	17,7
<i>Apoptosis</i>	REACTOME	4	10	0,016	0,246	0,304	13,5	19,6	37,4	29,5
<i>Pathways in cancer</i>	KEGG	4	20	0,017	0,247	0,304	15,6	37,9	27,2	19,3
<i>Immune system</i>	REACTOME	4	16	0,010	0,603	0,690	19,9	9,3	8,4	62,4
<i>Tumor necrosis factor pathway</i>	ST	4	12	0,006	0,766	0,786	20,2	25,5	38,0	16,4
<i>Chemokine signaling pathway</i>	KEGG	4	10	0,008	0,786	0,786	32,6	10,7	11,7	45,0

Outros dois testes de características globais da rede (entropia espectral e distribuição dos graus) nos permitem explorar os 16 conjuntos de genes a partir de perspectivas diferentes. Os resultados detalhados dos testes de análise diferencial de redes para essas outras duas características estão disponíveis no apêndice E.A e sumariados na Tabela 5. A distribuição e entropia espectrais apresentaram resultados semelhantes, apontando diferenças em cinco grupos iguais, com maiores pesos para os mesmos tipos de câncer (Tabela 5). Considerando os quatro grupos de genes em que todas as três características apresentaram diferenças (*ECM Regulators*, *Secreted factors*, *Matrisome associated* e *FAS Signaling pathway*) a distribuição dos graus diverge das medidas espectrais (distribuição e entropia) com relação a qual tipo cancerígeno tem maior peso na estatística θ , ou seja, mais se diferencia dos demais. Além disso, diferentemente das outras características o grupo *Small cell lung cancer* apresentou diferença principalmente para rede GBM.

Tabela 5 - Tabela de resumo da análise diferencial de redes considerando as características globais (Distribuição espectral, Entropia espectral e Distribuição dos graus). Nas primeira e segunda colunas estão o nome do grupo testado e a base de dados de origem, respectivamente. O “X” marca em qual característica global o grupo teve q-valor <0,05. Entre parênteses está o tipo cancerígeno que obteve o maior peso no teste, caso os dois maiores pesos sejam muito semelhantes (diferenças de até dois pontos) os dois tipos cancerígenos são colocados. Astro (Astrocitoma), Oligoa (Oligoastrocitoma), Oligoden (Oligodendrogloma) e GBM (Glioblastoma Multiforme).

Nome do grupo	Base de dados	Distribuição espectral	Entropia espectral	Distribuição dos graus
<i>ECM Regulators</i>	NABA	X (Oligoa)	X (Oligoa)	X (GBM)
<i>Secreted factors</i>	NABA	X (Oligoa)	X (Oligoa)	X (Oligoa/GBM)
<i>Matrisome associated</i>	NABA	X (Oligoa)	X (Oligoa)	X (Oligoa/Astro)
<i>Matrisome</i>	NABA	X (Oligoa)	-	X (Oligoa/Astro)
<i>CK-CK Recp interaction</i>	KEGG	X (Oligoden/GBM)	-	X (Oligoa/GBM)
<i>FAS Signaling pathway</i>	ST	X (Astro)	X (Astro)	X (Oligoden)
<i>Apoptosis</i>	KEGG	X (GBM)	X (GBM)	-
<i>Death pathway</i>	BIOCARTA	-	-	-
<i>Small cell lung cancer</i>	KEGG	-	-	X (GBM)
<i>HIVNEF Pathway</i>	BIOCARTA	-	-	-
<i>P53 Downstream pathway</i>	PID	-	-	-
<i>Apoptosis</i>	REACTOME	-	-	-
<i>Pathways in cancer</i>	KEGG	-	X (Oligoden/GBM)	-
<i>Immune system</i>	REACTOME	-	-	-
<i>Tumor necrosis factor pathway</i>	ST	-	-	-
<i>Chemokine signaling pathway</i>	KEGG	-	-	-

5.1.2 Análise diferencial de redes baseada em características locais

Os resultados dos cinco testes com características locais (Centralidades de grau, autovetor, *closeness*, *betweenness* e coeficiente de agrupamento) estão resumidos na Figura 14. O grupo *Pathways in cancer* foi o único que apresentou diferenças para todas as cinco características testadas.

Figura 14 - Gráfico de resumo dos resultados dos testes locais (centralidades e coeficiente de agrupamento). Para cada um dos 16 grupos de genes testados, os valores zero à cinco indicam quantos dos cinco testes tiveram uma estatística θ significativamente diferentes de zero (p-valor nominal $<0,05$).

Apesar dos cinco testes de *Pathways in cancer* detectarem diferenças entre as redes dos quatro tipos cancerígenos, a rede que possui maior peso nessa estatística não é a mesma para todas as características. O tecido cancerígeno GBM (Glioblastoma Multiforme) teve o maior peso na diferença entre as redes para os testes de autovetor e *betweenness* (Tabela 6). O GBM, com Oligodendroglioma, obtiveram os maiores pesos para o teste de centralidade de grau. Oligodendroglioma, por sua vez, teve o maior peso na estatística de coeficiente de agrupamento, enquanto o tecido Oligoastrocitoma foi o que mais contribuiu para a diferença da centralidade de *closeness* entre os grupos.

Tabela 6 - Resultados dos testes de análise diferencial das redes dos genes de Pathways in Cancer para cinco características locais (centralidades de grau, autovetor, closeness, betweenness e coeficiente de agrupamento). Nas colunas estão: o número de redes comparadas (Redes), número de vértices (Vértices), estatística θ , o p-valor nominal e o q-valor com a correção do p-valor para 16 testes feitos para os 16 grupos. Nas quatro últimas colunas estão os pesos associados a cada um dos quatro tipos cancerígenos: Astro (Astrocitoma), Oligoast (Oligoastrocitoma), Oligoden (Oligodendroglioma) e GBM (Glioblastoma Multiforme).

Pathways in Cancer	Redes	Vértices	Estatística θ	p-valor	q-valor	Astro	Oligoast	Oligoden	GBM
Centralidade de Autovetor	4	20	0,216	0,001	0,001	24,0	24,7	19,1	32,1
Centralidade de <i>betweenness</i>	4	20	5,812	0,001	0,008	24,4	20,1	19,2	36,3
Centralidade de grau	4	20	2,351	0,007	0,014	20,6	22,9	29,0	27,5
Coeficiente de agrupamento	4	20	0,081	0,024	0,077	13,8	25,6	36,2	24,4
Centralidade de <i>closeness</i>	4	20	0,008	0,004	0,064	19,0	42,5	27,6	11,0

5.1.3 Análise diferencial dos vértices

Como o grupo de genes *Pathways in cancer* apresentou diferenças estatísticas entre os tipos cancerígenos para todos os testes baseados em características locais, consideramos esse grupo um bom candidato para análise de diferencial dos vértices (variáveis). Em redes de co-expressão usamos o termo análogo à análise diferencial do vértice que é análise de co-expressão diferencial dos genes.

O gene CDK2 é o único gene diferencialmente co-expresso (q-valor<0,05) segundo o teste baseado na centralidade de grau (Tabela 7). Esse mesmo gene aparece significativamente alterado nos testes de centralidade de *closeness* e de autovetor (Apêndice E.C). Comparando com a análise de expressão diferencial, feita pelo teste de ANOVA (Apêndice E.D), este gene também está diferencialmente expresso nos quatro tipos celulares analisados (q-valor <0,001).

Tabela 7 - Resultado da análise diferencial do vértice com centralidade de grau, para os 20 genes da rede de Pathways in cancer. Para cada gene existe uma estatística θ que representa a diferença da centralidade do vértice nas quatro redes, um p-valor nominal, um q-valor de correção por múltiplos testes (por FDR) e os valores da centralidade de grau (em valor absoluto) em cada uma das redes dos tipos cancerígenos. Astro (Astrocitoma), Oligoast (Oligoastrocitoma), Oligoden (Oligodendroglioma) e GBM (Glioblastoma Multiforme).

				Valor de centralidade na rede			
	Estatística θ	p-valor	q-valor	Astro	Oligoast	Oligoden	GBM
CDK2	2,873	0,001	0,020	9,924	8,968	12,916	2,942
VEGFA	2,994	0,006	0,060	8,938	0	7,968	4,930
MYC	2,237	0,012	0,080	3,99	3,956	9,948	6,946
CCNE1	2,239	0,030	0,124	6,976	1,981	6,954	2,994
MMP2	2,234	0,031	0,124	9,910	3,964	8,926	5,937
NFKB1	1,755	0,041	0,137	8,957	7,919	11,933	5,953
VEGFC	2,453	0,058	0,152	0,991	0	4,918	5,887
HIF1A	2,235	0,061	0,152	7,906	6,981	8,971	1,993
XIAP	2,730	0,079	0,175	6,947	4,952	10,877	1,950
CCND1	1,766	0,106	0,212	4,960	4,956	9,982	5,903
NFKB2	1,865	0,121	0,220	8,989	8,962	12,956	5,998
BCL2L1	2,503	0,163	0,271	3,972	6,960	10,962	3,936
NOS2	2,219	0,198	0,304	6,929	1,979	5,919	1,995
MMP9	1,503	0,233	0,333	7,951	4,926	8,891	5,906
TRAF2	1,512	0,292	0,371	7,957	6,923	8,985	3,972
IL6	1,255	0,297	0,371	8,986	4,994	6,957	7,983
BCL2	1,253	0,409	0,481	7,935	5,927	6,931	3,926
TRAF1	1,993	0,462	0,513	9,918	3,907	5,918	7,878
NFKBIA	1,001	0,647	0,681	5,977	4,944	6,963	7,962
VEGFB	1,003	0,853	0,853	6,946	3,979	5,000	3,955

Por outro lado, o fator de transcrição NfκBIA, que não está diferencialmente expresso segundo o teste do ANOVA, tem a maior mudança no valor da centralidade de *betweenness* (q-valor <0,02) (Tabela 8). Isso que indica que embora haja mudanças significativas de co-expressão do gene, a sua expressão média se mantém.

Tabela 8 - Resultado da análise diferencial do vértice com centralidade de *betweenness*, para os 20 genes da rede de *Pathways in cancer*. Para cada gene existe uma estatística θ que representa a diferença da centralidade do vértice nas quatro redes, um p-valor nominal, um q-valor de correção por múltiplos testes (por FDR) e os valores da centralidade de grau (em valor absoluto) em cada uma das redes dos tipos cancerígenos. Astro (Astrocitoma), Oligoast (Oligoastrocitoma), Oligoden (Oligodendroglioma) e GBM (Glioblastoma Multiforme).

	Estatística θ	p-valor	q-valor	Valor de centralidade na rede			
				Astro	Oligoast	Oligoden	GBM
NFKBIA	11,875	0,001	0,020	5	0	2	34
MMP2	8,625	0,002	0,020	1	2	0	24
VEGFB	5,750	0,015	0,077	0	2	0	16
TRAF2	4,250	0,021	0,077	17	23	9	14
HIF1A	4,500	0,026	0,077	20	11	7	6
MYC	5,750	0,026	0,077	1	15	10	19
BCL2	5,000	0,027	0,077	13	0	1	8
CDK2	5,500	0,057	0,142	10	25	14	7
CCND1	4,250	0,079	0,174	0	9	16	9
IL6	6,750	0,087	0,174	2	0	1	19
BCL2L1	5,000	0,223	0,405	0	2	11	11
NFKB1	3,750	0,429	0,684	14	5	6	1
TRAF1	4,625	0,445	0,684	14	1	0	4
CCNE1	4,000	0,568	0,812	10	4	0	10
NOS2	1,500	0,687	0,916	5	2	0	3
NFKB2	3,500	0,803	1	21	26	26	17
VEGFA	1,500	0,902	1	5	0	1	2
XIAP	1,750	0,934	1	0	4	1	4
MMP9	0	1	1	0	0	0	0
VEGFC	0	1	1	0	0	0	0

5.1.4 Visualização das variáveis em vias metabólicas do KEGG

Utilizando a via *Pathways in cancer* é possível visualizar os 20 genes analisados pelos testes apresentados na seção anterior (5.1.3), colocando-os nas vias metabólicas. Dessa maneira, podemos ver os valores de expressão dos genes, em código de cores, para cada um dos tipos cancerígenos (Figura 15). Observamos que o gene CDK2 (destacado na Figura 15 por um círculo azul), é um gene ligado à proliferação das células cancerígenas. Sua expressão em Astrocitoma é próxima da média, já em Oligoastrocitoma e Oligodendroglioma ocorre uma queda e em Glioblastoma sua expressão passa a ser alta.

Figura 15 - Mapa da via Pathways in Cancer apresentando a expressão dos genes que formaram as redes analisadas. Os 20 genes presentes nos dados de expressão genética do grupo Pathways in Cancer estão coloridos com código de cores representando o nível de expressão em cada tecido cancerígeno, na seguinte ordem: Atrocitytoma, Oligoastrocitoma, Oligodendrogloma e Glioblastoma. A expressão gênica está normalizada pela média e desvio padrão. Genes com expressão abaixo da média tem valores próximos de -1 e estão coloridos em verde. Genes com expressão próxima a média tem valores próximos de zero e estão coloridos em cinza. Genes com expressão acima da média tem valores próximos de 1 e estão coloridos em vermelho. O círculo azul destaca o gene CDK2.

É possível, também, visualizar o valor de centralidade de cada gene na rede, em cada um dos tipos celulares (Figura 16). O gene NFκBIA (destacado na Figura 16 por um círculo azul) tem seu maior valor de *betweenness* no tecido de Glioblastoma, indicado pela cor vermelha, o que mostra sua importância na comunicação dessa rede de *Pathways in cancer* em Glioblastoma.

Figura 16 - Mapa da via Pathways in Cancer apresentando a Centralidade de betweenness dos genes que formaram as redes analisadas. Os 20 genes presentes nos dados de expressão genética do grupo Pathways in Cancer estão coloridos com código de cores representando o valor de betweenness dos vértices na rede de cada tecido cancerígeno, na seguinte ordem: Atrocioma, Oligoastrocitoma, Oligodendrogloma e Glioblastoma. O valor de betweenness varia de 0 a 34. Genes com pequena centralidade estão coloridos em cinza. Genes com alta centralidade estão coloridos em vermelho. O círculo azul destaca o gene NFkBIA que apresentou co-expressão diferencial pelo teste de betweenness e não apresentou expressão diferencial pelo ANOVA.

5.1.5 Discussão

No estudo de caso de gliomas (Seção 5.1) identificamos grupos de genes que podem caracterizar diferenças entre os quatro tipos de células cancerígenas (Astrocioma, Oligoastrocitoma, Oligodendrogloma e Glioblastoma Multiforme). Os resultados podem gerar hipóteses sobre como os grupos de genes estão envolvidos na formação desses tipos celulares ou como poderíamos diferenciar

esses tumores pelas redes de co-expressão. Os resultados vão de acordo com Brösick e Faissner (2015), que afirmam que componentes da matriz extracelular (*ECM, Extracellular Matrix*) influenciam no comportamento de células do glioma. É importante ressaltar que os tipos celulares divergem uns dos outros a partir de diferentes características das redes, mostrando que cada fenótipo sofre modificações específicas na sua rede de correlação. No mesmo sentido, a análise das características locais também permite diferenciar os tipos cancerígenos. Cada característica de rede destaca estruturas específicas e, conseqüentemente, apontará grupos de genes diferencialmente co-expressos que as outras medidas não apontaram, levantando diferentes hipóteses sobre o funcionamento do sistema em determinada condição.

O *BNS* indica genes candidatos para posteriores estudos, mais aprofundados, que podem levar à melhor compreensão do funcionamento dos gliomas. Como mostrado nos resultados, não só grupos de genes podem ser apontados como candidatos, mas também genes individualmente. Muitas vezes as análises de co-expressão e expressão diferencial identificam alterações nos mesmos genes, como no caso do gene *CDK2*, envolvido com proliferação do câncer (Ronchini e Capobianco 2001), que está tanto diferencialmente expresso como diferencialmente co-expresso. Por outro lado, a co-expressão pode complementar a expressão diferencial, como no caso do gene *NfkBIA*, importante na regulação da formação do câncer (Kinker et al. 2016), que não está diferencialmente expresso pelo resultado do ANOVA, mas foi identificado pela centralidade de *Betweenness* como um gene diferencialmente co-expresso. Situações onde o gene está diferencialmente co-expresso sem ter sua expressão média estatisticamente alterada são mostradas em outros trabalhos como o de Santos (2015) e de La Fuente (2010). Nesses casos existem alterações das relações entre os genes, sem haver alteração da sua expressão média.

Dessa forma, com uma análise de expressão diferencial de genes e com testes que posteriormente comprovem *in situ* a importância dos genes apontados como chaves na regulação de processos, o *BNS* pode levar a descobertas de genes ou grupos de genes que possibilitem uma compreensão mais aprofundada da formação e desenvolvimento de gliomas.

5.2 Metabolismo de sorgo

Sabe-se que o metabolismo e enchimento do grão são afetados pelo metabolismo de todos os outros órgãos da planta (de Souza et al. 2015). Entretanto, os trabalhos realizados com metabolismo

vegetal, em geral, são focados nas respostas de apenas um único órgão (de Souza et al. 2015). Sicher e Barnaby (2012), por exemplo, estudaram o metabolismo da folha de milho sem considerar o efeito da resposta do colmo e da raiz aos tratamentos. Por esse motivo, um dos objetivos do trabalho de De Souza et al. (2015) foi (i) comparar as diferenças nos perfis metabólicos de cinco órgãos de Sorgo (Folha, Colmo, Raiz, Raiz escora e Grãos). O Sorgo (*Sorghum bicolor*) é uma espécie muito usada como matéria prima da ração para o gado bovino em países desenvolvidos, como os EUA. Por ter o genoma completamente sequenciado e ser uma espécie com diversidade genética e geográfica, o sorgo é um bom modelo em gramíneas de metabolismo fotossintético C_4 (de Souza et al. 2015).

Os dados produzidos por De Souza et al. (2015) estão descritos na seção 2.3.2. Ao construir as redes de metabolismo de sorgo retiramos os metabólitos que não foram identificados em pelo menos uma das unidades amostrais, restando 73 dos 79 metabólitos originais. Os metabólitos foram divididos em cinco grupos segundo critérios bioquímicos: aminoácidos, carboidratos, ácidos orgânicos, nucleotídeos e polióis. Na Tabela 25 estão discriminados os 73 metabólitos e à qual grupo cada um pertence. Cada vértice da rede representa um dos 73 metabólitos e as arestas representam a relação estatística entre os níveis dos metabólitos. As arestas são formadas apenas com correlações de Pearson acima de 0,75 para garantir relações fortes entre as variáveis (Caldana et al. 2011; Ding et al. 2015; Jeong et al. 2001).

O objetivo desse estudo de caso é comparar mais de duas redes e mostrar a diversidade de aplicações da análise diferencial de redes. Por isso, restringimos o estudo de caso ao uso dos dados do tratamento de seca. Dessa forma, comparamos as redes de cada um dos cinco órgãos, apenas em seca, e consideramos cada órgão como uma condição. Queremos entender as diferenças entre as redes dos órgãos do metabolismo primário de sorgo. Para isso, vamos comparar a conectividade das redes dos cinco órgãos, construídas por cinco grupos metabólicos (Carboidratos, Ácido Orgânico, Aminoácido, Nucleotídeo e redes com todos os 73 metabólitos). Como o grupo de polióis possui apenas seis metabólitos não será construída uma rede exclusiva para esse grupo.

Apresentamos as redes de cada um dos cinco órgãos (Folha, Colmo, Raiz, Raiz escora e Grãos) na Figura 17. Observamos que existem diferenças entre as redes de cada órgão, porém para saber em quais estruturas das redes essas diferenças ocorrem, é necessário compará-las de acordo com suas características estruturais.

Para identificar quais mudanças ocorrem entre as redes, realizamos a análise diferencial de

múltiplas redes. As comparações foram feitas de acordo com as características globais e locais das redes. De acordo com os resultados obtidos, realizamos os testes a posteriori de análise diferencial dos vértices (metabólitos). Por fim, os metabólitos foram analisados em vias do KEGG para entendermos o papel de cada um em mapas metabólicos. Os resultados estão descritos nas próximas seções.

Figura 17 - Redes de correlação dos cinco órgãos, folha, colmo, raiz, raiz escora e grãos, com 73 metabólitos do metabolismo primário de sorgo divididos em cinco grupos. São 24 aminoácidos, 18 carboidratos, 13 ácidos orgânicos, 12 nucleotídeos e 6 polióis. As redes foram construídas usando correlação de pearson acima de 0,75 entre as seis amostras do tratamento Ambiente. As correlações positivas estão representadas por linhas inteiras e as correlações negativas por linhas tracejadas.

5.2.1 Análise diferencial de redes baseada em características globais

De acordo com a distribuição espectral, as redes metabólicas dos grupos *Carboidrato* e *Todos* apresentam diferenças entre os cinco órgãos (Tabela 9). Analisando os pesos de cada rede na estatística final, temos que a rede dos Grãos é a que mais se diferencia das demais, tanto para *Carboidrato* quanto para *Todos*.

Tabela 9 - Análise diferencial de redes baseada na distribuição espectral. A tabela apresenta os resultados dos testes estatísticos dos cinco grupos metabólicos analisados. Para cada grupo, está descrito o nome, o número de redes comparadas (Redes), o número de vértices (Vértices), a estatística θ , p-valor nominal (p-valor) e o q-valor de correção por FDR para os cinco testes (q-valor). Nas colunas seguintes estão os pesos, em porcentagem, das redes na estatística θ (Folha, Colmo, Raiz, Raiz escora e Grãos). A linha em branco separa os grupos com resultados estatisticamente significativos ($q\text{-valor} \leq 0,05$) daqueles com resultados não significativos.

Nome de grupo	Redes	Vértices	Estatística	p-valor	q-valor	Folha	Colmo	Raiz	Raiz escora	Grãos
Carboidrato	5	18	0,281	0,001	0,005	10,8	7,6	9,8	13,0	58,8
Todos	5	73	0,087	0,010	0,025	14,9	12,2	11,1	18,6	43,4
Ácido Orgânico	5	13	0,222	0,078	0,130	45,7	16,2	11,0	12,9	14,3
Aminoácidos	5	24	0,088	0,297	0,315	12,5	16,9	16,5	30,6	23,5
Nucleotídeo	5	12	0,172	0,315	0,315	16,0	14,3	16,5	15,9	37,4

O mesmo padrão foi observado na comparação das cinco redes segundo a entropia espectral (Tabela 10) onde os grupos *Carboidrato* e *Todos* tiveram diferenças significativas, principalmente para a rede dos Grãos. Da mesma forma que na análise das redes de glioma, a característica distribuição dos graus teve resultados diferentes se comparada às características espectrais (Tabela 10). A distribuição dos graus apresentou diferenças significativas entre as redes dos órgãos para os grupos *Ácido Orgânico* e *Carboidrato*. Para *Ácido orgânico*, os maiores pesos estão em folha e colmo e para *Carboidrato*, em Grãos. A rede de grãos é a que mais se diferencia das outras, principalmente para o grupo *Carboidrato*. Os resultados detalhados dos testes de entropia espectral e distribuição dos graus estão nas tabelas 26 e 27 no apêndice.

Tabela 10 - Tabela de resumo da análise diferencial de redes considerando apenas características globais (distribuição espectral, entropia espectral e distribuição dos graus). Nas primeira coluna está o nome do grupo testado. O “X” marca em qual característica global o grupo teve q-valor <0,05. Entre parênteses está o órgão que obteve o maior peso no teste, caso os dois maiores pesos sejam muito semelhantes (diferenças de até dois pontos) os dois órgãos são colocados.

Nome do grupo	Distribuição espectral	Entropia espectral	Distribuição dos graus
Todos	X (Grãos)	X (Grãos)	-
Aminoácido	-	-	-
Carboidrato	X (Grãos)	X (Grãos)	X (Grãos)
Ácido Orgânico	-	-	X (Folha/Colmo)
Nucleotídeo	-	-	-

5.2.2 Análise diferencial de redes baseada em características locais

Ao compararmos os órgãos pelas centralidades (grau, autovetor, *betweenness* e *closeness*) e coeficiente de agrupamento podemos ver que existem grandes diferenças nos resultados de cada método (Figura 18). Podemos observar que a centralidade de grau resultou em diferenças estatisticamente significativas (q-valor < 0,05) em todos os 5 grupos, isso pode ser visto com mais detalhes na Tabela 11.

Figura 18 - Gráficos de resumo dos resultados dos testes locais (centralidades e coeficiente de agrupamento). A) Para cada um dos cinco testes realizados, os valores zero a cinco indicam quantos grupos de metabólitos tiveram diferenças estatisticamente significativas (q-valor<0,05). B) Para cada um dos cinco grupos de metabólitos testados, os valores zero à cinco indicam em quantos testes as redes tem diferenças estatisticamente significativas (q-valor<0,05).

Segundo a centralidade de grau, o órgão com maior peso na estatística θ é diferente entre os 5 grupos (Tabela 11). Por exemplo, para o grupo *Todos*, com os 73 metabólitos, os maiores pesos são para Folha e Grãos enquanto para *Carboidrato* o maior peso está em Raiz escora e Grãos. É importante destacar que em todos os testes, baseados na centralidade de grau, a rede de grãos é a que mais se

distancia das demais.

Tabela 11 - Análise diferencial de redes baseada na centralidade de grau. A tabela apresenta os resultados dos testes estatísticos dos cinco grupos metabólicos analisados. Para cada grupo, está descrito o nome, o número de redes comparadas (Redes), o número de vértices (Vértices), a estatística θ , p-valor nominal (p-valor) e o q-valor de correção por FDR para os cinco testes (q-valor). Nas colunas seguintes estão os pesos, em porcentagem, das redes na estatística θ (Folha, Colmo, Raiz, Raiz escora e Grãos).

Nome de grupo	Redes	Vértices	Estatística	p-valor	q-valor	Folha	Colmo	Raiz	Raiz escora	Grãos
Todos	5	73	17,167	0,001	0,002	25,4	16,7	16,9	17,3	23,7
Carboidrato	5	18	3,857	0,001	0,002	20,0	13,2	11,1	25,6	30,1
Ácido Orgânico	5	13	3,482	0,001	0,002	26,2	19,8	16,2	14,2	23,8
Aminoácidos	5	24	5,152	0,003	0,004	23,3	21,1	16,1	17,4	22,1
Nucleotídeo	5	12	3,041	0,006	0,006	22,9	15,3	23,7	15,5	22,6

Ao contrário da centralidade de grau, a de autovetor (Tabela 28) e *betweenness* (Tabela 29) não apresentaram nenhuma diferença entre as redes dos órgãos para nenhum dos grupos formados, como pode ser visto na Figura 18.

A centralidade de *closeness* apresentou apenas uma diferença estatisticamente significativa entre os órgãos (Figura 18). O grupo *Ácido orgânico*, para a centralidade de *closeness*, foi o único que apresentou um q-valor < 0,05, com maior peso para a rede de Grãos (Tabela 12). Por fim, o coeficiente de agrupamento apresentou três grupos estatisticamente diferentes (*Ácido orgânico*, *Todos* e *Carboidrato*) (Tabela 30). Nos três casos a rede dos Grãos tem o maior peso na estatística θ com Folha e Colmo para *Ácido orgânico*, Folha para *Todos* e Raiz escora para *Carboidrato*.

Tabela 12 - Análise diferencial de redes baseada na centralidade de closeness. A tabela apresenta os resultados dos testes estatísticos dos cinco grupos metabólicos analisados. Para cada grupo, está descrito o nome, o número de redes comparadas (Redes), o número de vértices (Vértices), a estatística θ , p-valor nominal (p-valor) e o q-valor de correção por FDR para os cinco testes (q-valor). Nas colunas seguintes estão os pesos, em porcentagem, das redes na estatística θ (Folha, Colmo, Raiz, Raiz escora e Grãos). A linha em branco separa os grupos com resultados estatisticamente significativos (q-valor≤0,05) daqueles com resultados não significativos.

Nome de grupo	Redes	Vértices	Estatística	p-valor	q-valor	Folha	Colmo	Raiz	Raiz escora	Grãos
Ácido Orgânico	5	13	0,020	0,004	0,020	13,5	18,4	16,6	13,7	37,7
Nucleotídeo	5	12	0,016	0,073	0,183	12,0	34,2	26,8	12,6	14,4
Todos	5	73	0,001	0,132	0,215	12,5	24,5	36,8	13,2	13,0
Carboidrato	5	18	0,007	0,172	0,215	12,0	9,5	17,2	32,9	28,4
Aminoácidos	5	24	0,005	0,375	0,375	27,7	21,8	10,9	19,6	20,1

Em resumo, podemos ver que os resultados de características locais e globais diferem entre si. As características globais evidenciam as diferenças que ocorrem principalmente entre a rede de Grãos e os demais órgãos. Já as locais, apontam diferenças dos grupos para outras redes além dos Grãos, como para a Raiz escora e Folha.

5.2.3 Análise diferencial dos vértices

Na análise diferencial dos vértices, consideramos apenas os grupos *Carboidrato* e *Ácido Orgânico* pois foram os que apresentaram mais diferenças entre os as redes dos órgãos (Figura 18). Como encontrado na análise baseada em centralidade de grau (Tabela 11), a rede de Raiz escura tem o segundo maior peso de diferença para o grupo *Carboidrato*. Esse resultado pode ser explicado observando a Tabela 13, onde o carboidrato Glicose não possui nenhuma conexão na rede de Raiz escura. Enquanto nas outras redes ele tem um valor de centralidade maior. Esse padrão também ocorre com a Sacarose, Sacarose-6P e Frutose. Enquanto isso, Frutose-1,6bP (F-1,6bP) tem grau baixo em Folha, Colmo, Raiz e Grãos, onde é o terceiro menor da rede. Já na rede Raiz escura o grau desse metabólito é o terceiro maior.

Tabela 13 - Resultado da análise diferencial do vértice com Centralidade de grau, para os 18 metabólitos do grupo Carboidrato. Para cada metabólito existe uma estatística θ que representa a diferença da centralidade do vértice nas cinco redes, um p-valor nominal, um q-valor de correção por múltiplos teste (por FDR) e os valores da Centralidade de grau em cada uma das redes dos órgãos (Folha, Colmo, Raiz, Raiz escura e Grãos).

Metabólito	Estatística θ	p-valor	q-valor	Valor de centralidade na rede				
				Folha	Colmo	Raiz	Raiz escura	Grãos
Glicose	4,573	0,001	0,004	12,565	10,696	5,734	0	13,921
Sacarose	4,226	0,001	0,004	11,494	8,841	4,974	0	13,540
Deoxixilulose-5P	5,411	0,001	0,004	12,432	0,856	1,579	4,045	14,434
Pentose-P	3,672	0,001	0,004	6,440	1,556	2,466	4,228	14,360
Galactose-P	3,590	0,002	0,007	4,007	0	4,022	0	13,226
E4P	2,843	0,003	0,009	9,241	0	2,606	0	0
6PG	3,173	0,005	0,013	0	7,752	7,663	5,881	13,237
Trealose-6P	3,472	0,006	0,013	0,840	10,856	9,732	7,863	14,167
Sacarose-6P	3,303	0,007	0,014	6,899	8,840	10,185	0,904	13,325
Frutose	3,242	0,008	0,014	7,059	9,748	9,168	0,904	13,293
F-1,6bP	3,505	0,015	0,025	0	0,812	0	7,981	7,165
Glicose/Manose-P	3,306	0,017	0,025	10,23	8,544	8,527	0	14,026
UDPGlc	3,251	0,040	0,055	11,688	10,002	8,847	6,017	0,795
Ribulose-1,5bP	2,746	0,051	0,066	8,306	4,905	9,674	4,278	12,956
Manose-6P	2,988	0,105	0,126	9,040	6,699	9,730	2,467	13,651
S7P	2,028	0,189	0,212	10,464	9,446	9,465	5,216	14,072
Frutose-6P	2,254	0,230	0,243	6,836	10,746	12,135	8,611	14,377
Glicose-6P	1,721	0,494	0,494	10,215	8,631	8,822	8,757	14,432

O outro grupo que merece destaque é o *Ácido orgânico* por ser o que mais apresentou diferenças estatisticamente significativas (Figura 18). Observando o comportamento dos ácidos orgânicos nas cinco redes, vemos que um grupo de oito ácidos diferencialmente co-expressos tem comportamento semelhante (os oito primeiros da Tabela 14). Esse grupo de ácidos orgânicos tem altos valores de centralidade de grau (sempre por volta de 9) nas redes de Folha e Grãos, enquanto nas outras redes, Colmo, Raiz e Raiz escura, os valores de centralidade são menores (sempre abaixo de 5,5).

Tabela 14 - Resultado da análise diferencial do vértice com Centralidade de grau, para os 13 metabólitos do grupo Ácido orgânico. Para cada metabólito existe uma estatística θ que representa a diferença da centralidade do vértice nas cinco redes, um p-valor nominal, um q-valor de correção por múltiplos teste (por FDR) e os valores da Centralidade de grau em cada uma das redes dos órgãos (Folha, Colmo, Raiz, Raiz escora e Grãos).

Metabólito	Estatística θ	p-valor	q-valor	Valor de centralidade na rede				
				Folha	Colmo	Raiz	Raiz escora	Grãos
Piruvato	4,219	0,001	0,003	10,328	1,906	0,765	0,821	9,579
Mevalonato	4,215	0,001	0,003	9,930	0	0,838	0	8,191
cis-Aconitato	3,582	0,001	0,003	9,361	0	0,872	0,821	6,693
AKG	3,499	0,001	0,003	9,474	2,641	0,872	5,110	10,854
2/3PGA	3,862	0,001	0,003	10,412	0,805	5,216	0	9,695
Chiquimato	3,523	0,002	0,004	8,970	0	0,913	2,517	7,994
Malato	3,029	0,003	0,005	10,374	1,917	0,765	5,206	7,376
Isocitrato	2,782	0,003	0,005	9,588	1,872	4,654	5,316	9,898
Citrato	2,631	0,009	0,013	10,019	1,857	2,702	6,104	7,156
PEP	2,205	0,064	0,083	9,393	1,863	3,583	5,111	2,529
Fumarato	2,387	0,089	0,105	4,018	1,845	3,505	0,829	10,010
trans-Aconitato	2,712	0,097	0,105	0	1,842	0,913	3,304	9,835
Succinato	2,115	0,141	0,141	8,871	1,840	4,567	6,220	7,738

5.2.4 Visualização das variáveis em vias metabólicas do KEGG

Os metabólitos da via *Carbon fixation in photosynthetic organisms* do banco de dados KEGG estão apresentados na Figura 19. Desta via fazem parte apenas alguns carboidratos e ácidos orgânicos do total de 73 dos metabólitos estudados, o que explica a presença de apenas onze metabólitos destacados nas figuras. Podemos observar que os compostos S7P, Rib-1,5bP, 2/3PGA, PEP e Piruvato (destacados por círculos verdes) estão abaixo da média em todos os órgãos, com exceção do colmo onde a concentração desses metabólitos é maior (Figura 19). Em oposição a isso a centralidade de grau desses compostos é maior, principalmente nos órgãos folha e grãos, onde suas concentrações são menores (Tabela 13). Esse fato mostra que a importância de um metabólito na rede não está particularmente atrelada só à sua concentração média, mas também à sua correlação com os outros metabólitos.

Figura 20 - Mapa metabólico da via do Ciclo do citrato, retirado do KEGG. O mapa apresenta a centralidade de grau de nove dos treze metabólitos que formaram as redes do grupo Ácidos orgânicos. Foram utilizados apenas nove metabólitos dos treze pois esses são os únicos que aparecem na via escolhida. Os nove metabólitos estão coloridos com código de cores representando o valor de centralidade de grau dos vértices na rede de cada órgão, na seguinte ordem: Folha, Colmo, Raiz, Raiz escora e Grãos. O valor de grau varia de 0 à 10.9. Metabólitos com pequena centralidade estão coloridos em branco. Metabólitos com alta centralidade estão coloridos em vermelho.

5.2.5 Discussão

As análises realizadas na seção 5.2.1, baseadas nas características globais (Distribuição espectral, Entropia espectral e Distribuição dos Graus) nos mostram que o órgão que mais se destaca são os grãos. A rede de grãos, com todos os 73 metabólitos, se diferencia das demais principalmente pela rede do grupo dos carboidratos. Esse resultado concorda com De Souza et al. (2015) que encontraram altas concentrações de alguns carboidratos, aminoácidos e polióis no metabolismo dos grãos. É esperado que o metabolismo dos grãos seja o que mais se diferencia dos outros órgãos, por ser o principal dreno da planta durante seu enchimento (Schnyder 1993). Ao analisarmos as redes segundo as características locais, outros órgãos também foram importantes na estatística de diferença para os

grupos metabólicos. Considerando o grupo *Carboidratos*, o Grão e a Raiz escura são as redes com o maior peso na diferença da centralidade de grau dos seus vértices. Esse resultado se confirma com a Tabela 13 que mostra os carboidratos pouco conectados na rede de Raiz escura sendo que na rede Grãos os vértices estão altamente interligados. É importante ressaltar que os carboidratos, sacarose, sacarose-6P, glicose e frutose, destacados na seção 5.2.3, são os mesmos que De Souza (2015) mostrou estarem em altas concentrações na Raiz escura em todos os tratamentos. Além disso, De Souza et al. (2015) sugerem que a Raiz escura sirva como um fornecedor de carboidratos para o enchimento dos Grãos. Podemos concluir, então, que os resultados reforçam a ideia de que a Raiz escura tem um metabolismo específico de fornecimento de carboidratos, e os Grãos, por sua vez, armazenam carboidratos, sendo um forte dreno.

O *Ácido orgânico* foi outro grupo destacado por apresentar diferenças entre os órgãos. Em Folha, os ácidos orgânicos com maiores centralidades são o 2/3PGA, malato, piruvato e citrato. Esses metabólitos estão presentes nas vias de fixação de carbono, glicólise e respiração. Já em Grãos as duas maiores centralidades de grau são AKG (α -Ketoglutarate) e Fumarato. O metabólito AKG é um dos componentes do Ciclo de Krebs e precursor de 4 vias metabólicas de síntese de aminoácidos. Isso justifica sua alta centralidade, não só na rede do grupo de Ácidos Orgânicos como também na rede com todos os 73 metabólitos, onde é a centralidade de grau mais alta da rede. No trabalho de De Souza et al. (2015), observando apenas as concentrações dos metabólitos nos órgãos, não foi levantada a importância desse metabólito no metabolismo dos grãos e da folha. Apesar disso, verificou-se uma alta concentração de aminoácidos nos grãos, o que corrobora a ideia de que o metabolismo de aminoácidos está em destaque e que o AKG, sendo um precursor de diversas vias, é uma propriedade emergente nos sistemas de folha e grãos. Uma possível explicação para uma alteração nos grãos é o fato de que houve, segundo De Souza et al. (2015) uma alteração no teor de proteínas. No caso das folhas, é importante lembrar que estas também apresentam síntese de proteínas diferente de outros órgãos, uma vez que precisam sintetizar as proteínas do aparato fotossintético.

Nesse estudo de caso, apesar do baixo número de unidades amostrais, vimos que os métodos controlam a taxa de falsos positivos e apontam diferenças entre as redes. Porém, as centralidades de autovetor e *betweenness* não apresentaram diferenças estatísticas em nenhum dos grupos. Esses resultados podem ter ocorrido por dois principais fatores. O primeiro é que essas características realmente não se alterem entre os órgãos. Porém, é possível que devido à falta de unidades amostrais o *BioNetStat* não teve poder estatístico para encontrar diferenças significativas entre as redes, como

vimos nos testes de poder estatístico (seção 4.2).

Assim como no estudo de caso de Glioblastoma (seção 5.1) chegamos a conclusões que corroboram com os achados da literatura e trazem novas hipóteses para o estudo desses dados. Além disso, mostramos que a análise permite observar os dados a partir das correlações entre as variáveis, encontrando padrões que não são percebidos apenas pelas medidas de concentração dos metabólitos. Na seção 5.1.5 dissemos que as diferenças entre as redes dos tipos de gliomas e os principais genes responsáveis por essas diferenças poderiam servir como norteadores para diagnósticos de câncer. Neste mesmo sentido, abordando o estudo de caso de metabolismo de sorgo, podemos entender quais estruturas de rede e metabólitos caracterizam o metabolismo de cada órgão e inferir o papel que o órgão exerce na planta. O *BNS* pode ajudar no apontamento de variáveis ou grupos de metabólitos que são chaves para que determinado órgão desempenhe sua função na planta. Portanto, o *BNS* é uma ferramenta que permite levantar hipóteses, posteriormente testadas em laboratórios e abordadas com outras ferramentas de bioinformática, a fim de apontar grupos de genes candidatos na construção de plantas modificadas.

6 Conclusões gerais

No presente trabalho desenvolvemos e aplicamos um método que realiza a análise diferencial de múltiplas redes, complementando outros testes de comparação de redes já existentes. O pacote estatístico *BioNetStat* permite comparar múltiplas redes de correlação com base em estruturas de redes complexas. A análise diferencial de múltiplas redes junto à análise diferencial de vértices (variáveis) e às visualizações de redes e vias metabólicas são ferramentas implementadas no *BioNetStat*. Tais ferramentas otimizam a busca por grupos de variáveis que podem ser importantes candidatas no processo estudado e norteiam estudos posteriores que se aprofundem nos papéis dessas variáveis.

Desde que os devidos cuidados estatísticos sejam tomados no pré-tratamento dos dados, como normalização da distribuição das variáveis ou retirada de outliers, o *BNS* possibilita que diversos tipos de dados sejam analisados. Apesar de sua ampla aplicação, o *BNS* é restrito à construção de grafos sem direção nas arestas. Portanto, não é possível inserir a informação de direcionalidade na relação das variáveis estudadas. A implementação de direcionalidade para algumas centralidades é relativamente simples, porém, para as medidas espectrais dos grafos, são necessários mais estudos para a aplicação nesse contexto. Esse fato se deve à matriz de adjacência do grafo dirigido geralmente não ser simétrica, o que geraria autovalores não reais. Outra limitação, é que utilizando um pequeno número de unidades amostrais o *BNS* tem poder estatístico relativamente baixo. Esse problema atinge principalmente estudos na área de botânica, onde dificilmente são construídos experimentos com mais de dez unidades amostrais. Por fim, o *BNS* ainda não possui uma interface gráfica, o que limita o uso do teste estatístico a programadores de linguagem R ou ao uso tutorado, do leigo em programação. Essa implementação em interface será realizada em passos posteriores ao término deste trabalho.

Podemos concluir que o *BioNetStat* dará apoio às pesquisas com redes de correlação, facilitando a comparação entre diferentes redes, principalmente quando mais de dois estados biológicos são estudados. Espera-se também que o uso da ferramenta auxilie na difusão da teoria de sistemas, possibilitando abordagens que consideram a interação entre as partes do sistema, tornando necessário o desenvolvimento de novas ferramentas que preencham outras lacunas. Essas novas abordagens possibilitarão a descoberta de mais propriedades emergentes dos sistemas.

Apêndices

C Poder estatístico

A Figura 21 contém os resultados detalhados das simulações de poder estatístico do método de distribuição espectral. Foram construídos diversos cenários afim de avaliar o desempenho do método proposto.

Redes com peso (a)

Redes sem peso (b)

Testes com 50 vértices e 120 amostras

A(a)

0,51	0,64	0,867	0,974	0,994	0,997
0,503	0,614	0,843	0,966	0,991	0,997
0,505	0,599	0,833	0,962	0,989	0,996
0,488	0,573	0,784	0,941	0,974	0,989
0,504	0,572	0,759	0,922	0,956	0,977
0,496	0,563	0,756	0,899	0,952	0,97
0	0,05	0,1	0,2	0,3	0,5

A(b)

2	0,496	0,599	0,82	0,958	0,988	0,998	2
3	0,507	0,597	0,813	0,959	0,991	0,998	3
5	0,496	0,591	0,796	0,94	0,981	0,996	5
10	0,504	0,558	0,746	0,923	0,97	0,991	10
15	0,5	0,564	0,709	0,892	0,95	0,983	15
20	0,477	0,539	0,689	0,867	0,933	0,974	20
0	0,05	0,1	0,2	0,3	0,5		

Redes

Testes com 3 redes e 50 vértices

B(a)

0,5	0,484	0,477	0,487	0,479	0,485
0,493	0,497	0,51	0,537	0,538	0,616
0,507	0,509	0,543	0,628	0,708	0,821
0,501	0,53	0,64	0,811	0,895	0,959
0,503	0,614	0,843	0,966	0,991	0,997
0	0,05	0,1	0,2	0,3	0,5

B(b)

3	0,505	0,488	0,495	0,5	0,486	0,498	3
6	0,505	0,506	0,508	0,532	0,544	0,588	6
15	0,511	0,543	0,579	0,632	0,686	0,793	15
40	0,513	0,527	0,625	0,795	0,884	0,949	40
120	0,507	0,597	0,813	0,959	0,991	0,998	120
0	0,05	0,1	0,2	0,3	0,5		

Amostras

Testes com 3 redes e 120 amostras

C(a)

0,491	0,583	0,57	0,701	0,794	0,884
0,503	0,614	0,843	0,966	0,991	0,997
0,517	0,631	0,829	0,93	0,965	0,985
0	0,05	0,1	0,2	0,3	0,5

C(b)

10	0,496	0,626	0,651	0,764	0,836	0,901	10
50	0,507	0,597	0,813	0,959	0,991	0,998	50
130	0,503	0,577	0,759	0,903	0,96	0,991	130
0	0,05	0,1	0,2	0,3	0,5		

Vértices

γ

γ

Figura 21: Heatmaps de áreas sob a curva ROC (do inglês, AUC), mostrando os resultados dos experimentos de simulações para a análise diferencial de redes baseada na distribuição espectral, em redes com peso (a) e redes sem peso (b). Os resultados das simulações sob a hipótese nula para γ igual a zero estão na primeira coluna de cada heatmap. As demais colunas, com γ variando entre 0,05 e 0,5 apresentam os resultados das simulações sob a hipótese alternativa. Em A são apresentados os resultados para os testes variando o número r de conjuntos (redes), em 2, 3, 5, 10, 15 e 20, mantendo fixos os parâmetros de número de vértices $n_v=50$ vértices e número de amostras $n=120$. Nos heatmaps em B estão os resultados para os testes com três redes ($r=3$) e 50 vértices, variando o número de amostras em 3, 6, 15, 40 e 120. Para a figura C, são apresentados os resultados dos testes com três redes ($r=3$) e 120 amostras, variando o número de vértices em 10, 15 e 130.

D Teste de tempo de execução

Tabela 15: Apresentação dos valores obtidos na comparação do tempo de execução entre o teste de múltiplas redes (t_{BNS}) com os testes par a par (t_{par}).

Número de redes	t_{BNS} (min)	t_{par} (min)	Razão t_{par}/t_{BNS}
2	0,483	0,473	0,979
3	0,697	1,391	1,996
5	1,224	4,974	4,064
7	1,7	10,051	5,912
10	2,513	22,068	8,782
13	3,318	39,343	11,857
15	3,905	53,215	13,627
18	4,755	78,876	16,588
20	5,456	98,934	18,133

E Apêndice de glioblastoma

E.A Tabelas de análise de características globais

Nas tabelas a seguir apresentamos as análises diferenciais de redes baseada características globais das redes. As tabelas apresentam os resultados dos testes estatísticos para os 16 grupos de genes analisados, mostrando o nome do grupo, a base de dados que definiu aquele grupo de genes, o número de redes usadas, o número de vértices de cada rede, estatística θ com o respectivo p-valor associado e o q-valor corrigido para os 16 testes. Nas colunas seguintes estão os pesos de cada uma das quatro redes na estatística θ (Astro-Astrocitoma, Oligoast-Oligoastrocitoma, Oligoden-Oligodendroglioma e Gliob-Glioblastoma).

Tabela 16: Entropia espectral.

Nome do grupo	Base de dados	Redes	Vértices	Estatística θ	p-valor	q-valor	Astro	Oligoast	Oligoden	GBM
<i>ECM Regulators</i>	NABA	4	11	2,223	0,001	0,008	11,5	68,8	14	4,7
<i>Secreted factors</i>	NABA	4	24	0,122	0,001	0,008	17,2	46,5	3,7	32,7
<i>Apoptosis</i>	KEGG	4	15	0,156	0,003	0,012	8,3	13	23,4	55,4
<i>FAS Signaling pathway</i>	ST	4	10	0,133	0,003	0,012	59,5	15,4	14,4	10,7
<i>Matrisome associated</i>	NABA	4	43	0,067	0,006	0,019	3,9	44,5	19,1	32,6
<i>Death pathway</i>	BIOCARTA	4	11	0,103	0,02	0,053	13,1	12,1	15,9	58,9
<i>Small cell lung cancer</i>	KEGG	4	12	0,074	0,026	0,058	28,4	0,3	1,8	69,5
<i>Matrisome</i>	NABA	4	44	0,058	0,029	0,058	0,118	47,5	18,8	33,5
<i>CK-CK Recp interaction</i>	KEGG	4	47	0,067	0,088	0,156	12,9	15,4	24,6	47,1
<i>HIVNEF Pathway</i>	BIOCARTA	4	10	0,087	0,12	0,192	22,2	9,9	1,5	66,4
<i>Immune system</i>	REACTOME	4	16	0,078	0,25	0,363	22,8	11,4	7,3	58,5
<i>P53 Downstream pathway</i>	PID	4	10	0,084	0,367	0,489	54,3	24,5	8,6	12,6
<i>Apoptosis</i>	REACTOME	4	10	0,039	0,61	0,736	20,5	2,7	19,3	57,5
<i>Chemokine signaling pathway</i>	KEGG	4	10	0,044	0,682	0,736	38,2	1,5	3,8	56,6
<i>Pathways in cancer</i>	KEGG	4	20	0,036	0,69	0,736	19,7	55,8	3,9	20,6
<i>Tumor necrosis factor pathway</i>	ST	4	12	0,028	0,875	0,875	18,3	3,6	65,3	12,7

Tabela 17: Distribuição dos graus.

Nome do grupo	Base de dados	Redes	Vértices	Estatística θ	p-valor	q-valor	Astro	Oligoast	Oligoden	GBM
<i>Pathways in cancer</i>	KEGG	4	20	0,19	0,001	0,005	17,8	16,7	35,1	30,4
<i>ECM Regulators</i>	NABA	4	11	0,294	0,001	0,005	14,4	15,3	15,1	55,2
<i>Secreted factors</i>	NABA	4	24	0,23	0,001	0,005	7,5	46,5	7,4	38,6
<i>Matrisome associated</i>	NABA	4	43	0,096	0,002	0,006	32,5	51,4	12,4	3,7
<i>Matrisome</i>	NABA	4	44	0,104	0,002	0,006	35,2	48,2	12,8	3,8
<i>FAS Signaling pathway</i>	ST	4	10	0,301	0,006	0,016	16,9	24,6	37,1	21,4
<i>CK-CK Recp interaction</i>	KEGG	4	47	0,102	0,01	0,023	7,5	54,1	7,4	30,9
<i>Small cell lung cancer</i>	KEGG	4	12	0,37	0,016	0,032	10,8	8,5	40	40,7
<i>Apoptosis</i>	KEGG	4	15	0,086	0,065	0,115	4	16,3	50,6	29,1
<i>Death pathway</i>	BIOCARTA	4	11	0,17	0,083	0,133	30,7	34,4	7,6	27,3
<i>Apoptosis</i>	REACTOME	4	10	0,181	0,104	0,151	6,3	17,1	53,1	23,5
<i>P53 Downstream pathway</i>	PID	4	10	0,075	0,275	0,366	6,3	44,1	12,3	37,3
<i>Tumor necrosis factor pathway</i>	ST	4	12	0,029	0,453	0,542	11,3	14	56,9	17,9
<i>HIVNEF Pathway</i>	BIOCARTA	4	10	0,037	0,475	0,542	17,7	16,7	17,8	47,8
<i>Chemokine signaling pathway</i>	KEGG	4	10	0,04	0,552	0,589	30,5	15,2	22,6	31,8
<i>Immune system</i>	REACTOME	4	16	0,033	0,731	0,731	32,9	6,9	19,7	40,6

E.B Tabelas de análise de características locais

Nas tabelas a seguir apresentamos as análises diferenciais de redes baseada características locais das redes. As tabelas apresentam os resultados dos testes estatísticos para os 16 grupos de genes analisados, mostrando o nome do grupo, a base de dados que definiu aquele grupo de genes, o número de redes usadas, o número de vértices de cada rede, estatística θ com o respectivo p-valor associado e o q-valor corrigido para os 16 testes. Nas colunas seguintes estão os pesos de cada uma das quatro redes na estatística θ (Astro-Astrocitoma, Oligoast-Oligoastrocitoma, Oligoden-Oligodendroglioma e Gliob-Glioblastoma).

Tabela 18: Centralidade de Betweenness.

Nome do grupo	Base de dados	Redes	Vértices	Estatística θ	p-valor	q-valor	Astro	Oligoast	Oligoden	GBM
<i>Pathways in cancer</i>	KEGG	4	20	5,813	0,001	0,008	24,4	20,1	19,2	36,3
<i>P53 Downstream pathway</i>	PID	4	10	3,445	0,001	0,008	39,1	17,1	16,8	27
<i>Small cell lung cancer</i>	KEGG	4	12	5,007	0,002	0,011	25,3	14	18,7	42
<i>CK-CK Recp interaction</i>	KEGG	4	47	11,351	0,004	0,016	20,9	25,7	25,3	28
<i>Apoptosis</i>	REACTOME	4	10	3,583	0,009	0,029	24,9	22	26,6	26,6
<i>Matrisome associated</i>	NABA	4	43	15,461	0,012	0,032	19,7	30,7	22,1	27,5
<i>Matrisome</i>	NABA	4	44	14,885	0,018	0,041	19,1	28,7	24,5	27,7
<i>Chemokine signaling pathway</i>	KEGG	4	10	3,16	0,035	0,066	23,4	30,8	24,7	21,2
<i>Immune system</i>	REACTOME	4	16	2,769	0,037	0,066	28,1	22,8	19,6	29,5
<i>Apoptosis</i>	KEGG	4	15	3,415	0,065	0,097	26,1	23,7	21	29,2
<i>Death pathway</i>	BIOCARTA	4	11	3,693	0,067	0,097	26	27,8	21,8	24,4
<i>Secreted factors</i>	NABA	4	24	6,513	0,08	0,107	20,4	18,7	34	26,9
<i>FAS Signaling pathway</i>	ST	4	10	2,303	0,152	0,187	27	25,3	17,1	30,6
<i>HIVNEF Pathway</i>	BIOCARTA	4	10	2,199	0,256	0,292	30,5	26,3	20,8	22,5
<i>Tumor necrosis factor pathway</i>	ST	4	12	1,876	0,3	0,32	20,2	19	20,5	40,3
<i>ECM Regulators</i>	NABA	4	11	1,996	0,364	0,364	28,5	14,6	37,7	19,2

Tabela 19: Centralidade de Closeness.

Nome do grupo	Base de dados	Redes	Vértices	Estatística θ	p-valor	q-valor	Astro	Oligoast	Oligoden	GBM
<i>Pathways in cancer</i>	KEGG	4	20	0,008	0,004	0,064	19	42,5	27,6	11
<i>Small cell lung cancer</i>	KEGG	4	12	0,013	0,029	0,165	13,8	9,7	30,3	46,2
<i>Secreted factors</i>	NABA	4	24	0,009	0,031	0,165	11,1	38,8	14,9	35,3
<i>Apoptosis</i>	REACTOME	4	10	0,016	0,06	0,195	17,6	13	32	37,5
<i>ECM Regulators</i>	NABA	4	11	0,007	0,061	0,195	15,6	20,6	35,6	28,2
<i>Immune system</i>	REACTOME	4	16	0,004	0,082	0,196	21,9	22,6	18,6	36,8
<i>Apoptosis</i>	KEGG	4	15	0,008	0,086	0,196	20,2	14,3	20,9	44,6
<i>HIVNEF Pathway</i>	BIOCARTA	4	10	0,019	0,106	0,212	14,8	18,3	21,6	45,3
<i>Tumor necrosis factor pathway</i>	ST	4	12	0,006	0,125	0,222	22	19	24	35
<i>P53 Downstream pathway</i>	PID	4	10	0,016	0,153	0,245	13,7	25,4	19,6	41,3
<i>CK-CK Recp interaction</i>	KEGG	4	47	0,004	0,191	0,278	21,1	24,5	6,2	48,3
<i>FAS Signaling pathway</i>	ST	4	10	0,012	0,299	0,374	26	20,5	22,6	30,9
<i>Death pathway</i>	BIOCARTA	4	11	0,012	0,304	0,374	23,7	16,9	23,2	36,3
<i>Chemokine signaling pathway</i>	KEGG	4	10	0,009	0,618	0,707	19,3	20,2	27,5	33
<i>Matrisome associated</i>	NABA	4	43	0,001	0,747	0,76	16,9	24	36,8	22,3
<i>Matrisome</i>	NABA	4	44	0,001	0,76	0,76	17	24	36,6	22,3

Tabela 20: Centralidade de Grau.

Nome do grupo	Base de dados	Redes	Vértices	Estatística θ	p-valor	q-valor	Astro	Oligoast	Oligoden	GBM
<i>CK-CK Recp interaction</i>	KEGG	4	47	6,534	0,001	0,003	19,2	29,7	16,1	35
<i>FAS Signaling pathway</i>	ST	4	10	1,503	0,001	0,003	27	23,1	21,4	28,5
<i>ECM Regulators</i>	NABA	4	11	1,374	0,001	0,003	28,1	18,1	24,8	29,1
<i>Secreted factors</i>	NABA	4	24	4,183	0,001	0,003	15,2	31,8	14,5	38,5
<i>Matrisome associated</i>	NABA	4	43	4,352	0,001	0,003	30	27,4	16,6	26
<i>Matrisome</i>	NABA	4	44	4,502	0,001	0,003	31,4	27,2	16,6	24,8
<i>P53 Downstream pathway</i>	PID	4	10	1,214	0,005	0,011	21,7	24,1	18,9	35,3
<i>Pathways in cancer</i>	KEGG	4	20	2,351	0,007	0,014	20,6	22,9	29	27,6
<i>Immune system</i>	REACTOME	4	16	1,771	0,012	0,021	21,8	22,6	18,3	37,4
<i>Apoptosis</i>	KEGG	4	15	1,613	0,029	0,046	24,4	22,5	23,4	29,7
<i>Chemokine signaling pathway</i>	KEGG	4	10	1,327	0,037	0,052	21,8	18,1	25,5	34,6
<i>HIVNEF Pathway</i>	BIOCARTA	4	10	1,258	0,04	0,052	17,1	24	26,9	32
<i>Apoptosis</i>	REACTOME	4	10	1,378	0,042	0,052	22,4	18,7	33,2	25,7
<i>Death pathway</i>	BIOCARTA	4	11	1,286	0,047	0,054	20,5	25,5	21	33,1
<i>Tumor necrosis factor pathway</i>	ST	4	12	1,261	0,053	0,056	23,1	22,3	23,8	30,9
<i>Small cell lung cancer</i>	KEGG	4	12	1,541	0,077	0,077	15,1	14,3	37,8	32,9

Tabela 21: Centralidade de Autovetor.

Nome do grupo	Base de dados	Redes	Vértices	Estatística θ	p-valor	q-valor	Astro	Oligoast	Oligoden	GBM
<i>CK-CK Recp interaction</i>	KEGG	4	47	0,172	0,001	0,001	21,5	27,3	17,5	33,7
<i>Pathways in cancer</i>	KEGG	4	20	0,216	0,001	0,001	24	24,7	19,1	32,1
<i>HIVNEF pathway</i>	BIOCARTA	4	10	0,17	0,001	0,001	20,4	23,6	25,8	30,2
<i>Tumor necrosis factor pathway</i>	ST	4	12	0,146	0,001	0,001	24,9	22	20,9	32,3
<i>P53 Downstream pathway</i>	PID	4	10	0,188	0,001	0,001	27,5	23,2	20,4	28,9
<i>Apoptosis</i>	REACTOME	4	10	0,217	0,001	0,001	24,9	23,9	25	26,2
<i>Immune system</i>	REACTOME	4	16	0,144	0,001	0,001	23,7	21,3	16,6	38,4
<i>ECM Regulators</i>	NABA	4	11	0,236	0,001	0,001	18,1	21	18	42,9
<i>Secreted factors</i>	NABA	4	24	0,209	0,001	0,001	20,6	29	17,5	32,9
<i>Matrisome associated</i>	NABA	4	43	0,181	0,001	0,001	24,2	25,4	18,4	32,1
<i>Matrisome</i>	NABA	4	44	0,183	0,001	0,001	24,5	25,1	19,7	30,7
<i>Small cell lung cancer</i>	KEGG	4	12	0,22	0,002	0,002	25,2	18,2	24,1	32,6
<i>FAS Signaling pathway</i>	ST	4	10	0,168	0,002	0,002	26,5	19,8	15,2	38,5
<i>Chemokine signaling pathway</i>	KEGG	4	10	0,173	0,006	0,006	16,5	20,6	23,1	39,8
<i>Apoptosis</i>	KEGG	4	15	0,135	0,006	0,006	29,9	21,6	21,4	27,1
<i>Death pathway</i>	BIOCARTA	4	11	0,17	0,012	0,012	15,2	22,6	28	34,2

Tabela 22: Coeficiente de Agrupamento.

Nome do grupo	Base de dados	Redes	Vértices	Estatística θ	p-valor	q-valor	Astro	Oligoast	Oligoden	GBM
<i>FAS Signaling pathway</i>	ST	4	10	0,131	0,001	0,008	23,2	24,8	26,7	25,4
<i>ECM Regulators</i>	NABA	4	11	0,105	0,001	0,008	25,4	18,4	24,8	31,4
<i>Secreted factors</i>	NABA	4	24	0,13	0,002	0,011	6,2	41,7	8,169	43,9
<i>CK-CK Recp interaction</i>	KEGG	4	47	0,079	0,022	0,077	10,9	44,6	5,4	39,1
<i>Pathways in cancer</i>	KEGG	4	20	0,081	0,024	0,077	13,8	25,6	36,2	24,4
<i>P53 Downstream pathway</i>	PID	4	10	0,073	0,064	0,17	10,3	26,4	18,2	45,1
<i>Apoptosis</i>	REACTOME	4	10	0,1	0,104	0,23	10,9	16,5	40,6	32,1
<i>Small cell lung cancer</i>	KEGG	4	12	0,097	0,115	0,23	4,4	6	46,9	42,7
<i>Matrisome</i>	NABA	4	44	0,05	0,135	0,24	47,4	44,6	5	3,1
<i>Matrisome associated</i>	NABA	4	43	0,047	0,184	0,294	47,1	47,5	2,1	3,2
<i>Tumor necrosis factor pathway</i>	ST	4	12	0,057	0,274	0,398	16	21,6	33,8	28,7
<i>HIVNEF Pathway</i>	BIOCARTA	4	10	0,07	0,341	0,454	9,6	15,7	30,3	44,4
<i>Chemokine signaling pathway</i>	KEGG	4	10	0,066	0,411	0,505	23,7	23,1	25,1	28,1
<i>Death pathway</i>	BIOCARTA	4	11	0,056	0,482	0,55	29,5	37,6	15,7	17,2
<i>Apoptosis</i>	KEGG	4	15	0,034	0,709	0,757	12,7	37	12,2	38
<i>Immune system</i>	REACTOME	4	16	0,024	0,885	0,885	20,1	20,9	32,9	26,1

E.C **Análise diferencial de vértices**

Tabela 23: Coeficiente de Agrupamento.

				Valor do coeficiente na rede			
	Estatística θ	p-valor	q-valor	Astro	Oligoast	Oligoden	GBM
BCL2L1	0,083	0,014	0,055	0,401	0,243	0,441	0,266
TRAF1	0,083	0,015	0,055	0,366	0,261	0,471	0,243
VEGFB	0,083	0,016	0,055	0,384	0,261	0,476	0,266
VEGFC	0,089	0,017	0,055	0,418	0,284	0,477	0,255
IL6	0,085	0,019	0,055	0,372	0,255	0,465	0,243
NFKBIA	0,089	0,02	0,055	0,389	0,255	0,465	0,243
MMP9	0,08	0,025	0,055	0,378	0,255	0,453	0,255
MYC	0,085	0,027	0,055	0,401	0,261	0,447	0,249
CCND1	0,083	0,033	0,055	0,395	0,255	0,447	0,255
BCL2	0,082	0,034	0,055	0,378	0,249	0,465	0,266
TRAF2	0,08	0,036	0,055	0,378	0,244	0,453	0,266
NFKB2	0,079	0,039	0,055	0,372	0,232	0,43	0,254
NFKB1	0,079	0,039	0,055	0,372	0,238	0,436	0,255
NOS2	0,076	0,04	0,055	0,384	0,273	0,471	0,278
XIAP	0,073	0,041	0,055	0,384	0,255	0,442	0,278
MMP2	0,076	0,044	0,055	0,366	0,261	0,453	0,255
HIF1A	0,077	0,048	0,056	0,378	0,243	0,453	0,278
CCNE1	0,076	0,053	0,058	0,383	0,273	0,465	0,272
CDK2	0,073	0,057	0,058	0,366	0,232	0,43	0,272
VEGFA	0,072	0,058	0,058	0,372	0,284	0,459	0,26

E.D **Tabela resultado de ANOVA**

Tabela 24: ANOVA para 139 genes.

Gene	Estatística F	p-valor	q-valor	Gene	Estatística F	p-valor	q-valor
MYCBP	246,58	0	0	COX20	26,381	0	0
VEGFA	226,009	0	0	TNFRSF9	26,294	0	0
MMP9	214,922	0	0	CXCL14	25,718	0	0
BCL2L12	214,316	0	0	NFKB2	25,449	0	0
TNFRSF12A	206,809	0	0	MYCBPAP	24,317	0	0
PLAUR	206,574	0	0	NOS2	24,308	0	0
CDK2	183,542	0	0	CXCL1	24,066	0	0
TNFAIP6	182,678	0	0	TNFSF11	23,664	0	0
PLAU	165,858	0	0	IMMP2L	23,1	0	0
TNFRSF1A	155,185	0	0	C1QTNF3	21,182	0	0
TNFAIP8	148,418	0	0	C1QTNF8	20,554	0	0
MYCBP2	147,747	0	0	ACOX2	20,51	0	0
TNFRSF11B	145,065	0	0	CDK2AP2	20,416	0	0
BCL2L2	130,386	0	0	VEGFB	19,184	0	0
CXCL10	127,768	0	0	TNFRSF6B	18,541	0	0
C1QTNF6	119,941	0	0	TNFSF15	18,173	0	0
HIF1AN	115,548	0	0	TRAF1	17,673	0	0
TNFRSF14	110,883	0	0	CDK20	17,243	0	0
SOD2	110,294	0	0	NFKBID	16,828	0	0
TNFSF12.TNFSF13	97,086	0	0	C1QTNF1	15,182	0	0
NFKBIZ	92,296	0	0	BCL2L13	14,881	0	0
TNFSF14	87,697	0	0	TRAF2	14,711	0	0
CXCL11	85,945	0	0	BCL2L1	14,473	0	0
CCNE1	80,711	0	0	TNFRSF21	14,396	0	0
ICAM1	80,109	0	0	NFKBIE	14,292	0	0
MMP2	78,644	0	0	C1QTNF4	13,625	0	0
TNFRSF4	77,563	0	0	PPIL6	12,068	0	0
C1QTNF2	76,963	0	0	TNFRSF10A	12,015	0	0
BCL2A1	76,943	0	0	BCL2L14	11,821	0	0
TNFRSF19	73,632	0	0	BCL2L15	10,401	0	0
MMP24	73,34	0	0	TNFRSF17	9,938	0	0
TNFAIP8L2	70,662	0	0	CDK2AP1	9,388	0	0
NFKBIB	66,908	0	0	TNFRSF25	9,206	0	0
TNFSF8	65,828	0	0	MYCT1	8,558	0	0
TNFAIP3	64,335	0	0	TNF	8,534	0	0
CFLAR	64,2	0	0	GADD45B	8,136	0	0
NFKBIL1	63,641	0	0	HIF1A	7,976	0	0
VCAM1	59,975	0	0	TNFAIP8L3	7,802	0	0
IL6	59,879	0	0	NANOS2	7,795	0	0
TNFAIP2	57,812	0	0	CCND1	7,641	0	0
TERT	54,391	0	0	IL6R	6,267	0	0
CXCL16	52,123	0	0	C1QTNF9	6,257	0	0
TNFRSF10D	51,875	0	0	MMP23A	6,074	0	0,001
TNFRSF1B	51,428	0	0	TNFSF9	4,59	0,003	0,004
XIAP	50,494	0	0	MMP23B	4,528	0,004	0,004
BCL2	48,966	0	0	SELE	4,372	0,005	0,006
TNFSF13B	47,298	0	0	IL6ST	4,139	0,006	0,008
TNFRSF10B	45,069	0	0	TNFAIP1	4,028	0,007	0,009
TNFRSF10C	43,604	0	0	MMP21	3,646	0,013	0,014
TNFSF10	41,019	0	0	TICAM1	3,503	0,015	0,017
TNFRSF18	39,602	0	0	TNFSF18	3,199	0,023	0,026
TNFRSF11A	39,115	0	0	WHAMMP2	3,125	0,025	0,029
MYCNOS	38,344	0	0	MMP28	3,056	0,028	0,031
NANOS1	37,907	0	0	C1QTNF9B	2,723	0,043	0,048
NFKB1	37,891	0	0	MYC	2,704	0,045	0,049
MMP25	33,864	0	0	NFKBIA	2,546	0,055	0,06
MYCL	33,08	0	0	SELENBP1	2,152	0,092	0,1
MYCN	32,787	0	0	CXCL17	1,882	0,131	0,141
TNFRSF13C	32,562	0	0	BCL2L11	1,639	0,179	0,19
TNFSF4	32,535	0	0	TNFRSF13B	1,601	0,188	0,198
C1QTNF9B.AS1	32,058	0	0	CXCL12	1,564	0,197	0,206
TNFSF13	29,858	0	0	VEGFC	1,507	0,211	0,22
TNFAIP8L1	29,386	0	0	MMP20	1,37	0,251	0,259
TNFRSF8	29,005	0	0	BCL2L10	1,192	0,312	0,319
CXCL13	27,761	0	0	MMP27	0,529	0,662	0,672
IL1B	26,649	0	0	C1QTNF7	0,355	0,786	0,792
TNFSF12	26,529	0	0	MMP26	0,03	0,993	0,993

F Apêndice do estudo de caso de sorgo

Tabela 25: 79 metabólitos discriminados em cinco grupos (Aminoácido, Carboidrato, Ácido Orgânico, Nucleotídeo, Poliol) de acordo com sua estrutura bioquímica.

Metabólito	Grupo
Alanina	Aminoácido
Aspartato	
Arginina	
Asparagina	
Citrulina	
Cisteína	
GABA	
Glutamato	
Glutamina	
Glicina	
Histidina	
Homoserina	
Isoleucina	
Leucina	
Lisina	
Metionina	
Ornitina	
Fenilalanina	
Prolina	
Serina	
Treonina	
Triptofano	
Tirosina	
Valina	
6PG	Carboidrato
Deoxixilulose-5P	
Frutose	
Frutose-6P	
F-1,6bP	
Galactose-P	
Glicose	
Glicose-6P	
Glicose/Manose-P	
Manose-6P	
Pentose-P	
Ribulose-1,5bP	
Sacarose	
Sacarose-6P	
Trealose-6P	
UDP-Glc	
S7P	
E4P	

Metabólito	Grupo
cis-Aconitato	Ácido Orgânico
trans-Aconitato	
Citrato	
Fumarato	
Isocitrato	
Malato	
Piruvato	
Succinato	
Mevalonato	
Chiquimato	
AKG	
2/3PGA	
PEP	
AMP	Nucleotídeo
ADP	
ATP	
CMP	
CDP	
CTP	
GMP	
GDP	
GTP	
IMP	
UMP	
UDP	Poliol
UTP	
Eritritol-treitol	
Glicerol	
Glicerol-P	
Inositol	
Manitol-Galactitol	
Sorbitol	
Pentitol	

Análise diferencial de redes

Tabela 26: Análise diferencial de redes baseada na entropia espectral. A tabela apresenta os resultados dos testes estatísticos dos cinco grupos de metabólitos analisados. Para cada grupo, está descrito o nome do grupo, o número de redes testadas (Redes), o número de vértices em cada rede (Vértices), a estatística θ (Estatística) com o respectivo p-valor associado (p-valor) e o q-valor de correção por FDR para os cinco testes (q-valor). Nas colunas seguintes estão os pesos de cada uma das redes na estatística θ (Folha, Colmo, Raiz, Raiz escora e Grãos).

Nome de grupo	Redes	Vértices	Estatística	p-valor	q-valor	Folha	Colmo	Raiz	Raiz escora	Grãos
Carboidrato	5	18	0,741	0,004	0,015	0,388	3,3	10,9	2,6	83
Todos	5	73	0,298	0,006	0,015	22,6	8,9	7,5	15,9	45,2
Ácido Orgânico	5	13	0,487	0,102	0,17	76,1	5,3	6,8	3,9	7,9
Nucleotídeo	5	12	0,413	0,203	0,229	21,5	14,4	5,1	8,8	50,1
Aminoácidos	5	24	0,193	0,229	0,229	28,1	7,8	3,5	23,1	37,5

Tabela 27: Distribuição dos graus

Nome de grupo	Redes	Vértices	Estatística	p-valor	q-valor	Folha	Colmo	Raiz	Raiz escora	Grãos
Ácido Orgânico	5	13	0,653	0,006	0,025	28	27,1	14,6	12	18,2
Carboidrato	5	18	0,381	0,01	0,025	6,5	6,5	7,7	20,2	59,1
Todos	5	73	0,322	0,044	0,073	16,7	14,7	17,5	17,6	33,4
Nucleotídeo	5	12	0,555	0,138	0,173	11,1	11,5	34,3	11	32,1
Aminoácidos	5	24	0,172	0,705	0,705	28,6	10,4	4	37,6	19,5

Tabela 28: Centralidade de autovetor.

Nome de grupo	Redes	Vértices	Estatística	p-valor	q-valor	Folha	Colmo	Raiz	Raiz escora	Grãos
Todos	5	73	0,333	0,123	0,271	22,1	18,5	19	19,2	21,2
Aminoácidos	5	24	0,311	0,217	0,271	20,9	21,6	18	19,1	20,3
Carboidrato	5	18	0,298	0,469	0,469	21,9	15,5	12,2	27,9	22,5
Ácido Orgânico	5	13	0,383	0,107	0,271	19	23,7	20,4	17,9	19
Nucleotídeo	5	12	0,355	0,167	0,271	22,9	13,3	29,1	15,7	19

Tabela 29: Centralidade de betweenness.

Nome de grupo	Redes	Vértices	Estatística	p-valor	q-valor	Folha	Colmo	Raiz	Raiz escora	Grãos
Nucleotídeo	5	12	2,067	0,914	1	11,7	31	11,8	22,3	23
Carboidrato	5	18	4,016	0,959	1	21,5	24,7	25,5	10,7	17,5
Ácido Orgânico	5	13	1,505	0,981	1	22,5	11,7	19,6	21,2	25,2
Todos	5	73	38,923	0,986	1	16	18,4	19,7	25,1	20,7
Aminoácidos	5	24	6,473	1	1	15,2	15,2	27,9	21,3	20,2

Tabela 30: Coeficiente de agrupamento.

Nome de grupo	Redes	Vértices	Estatística	p-valor	q-valor	Folha	Colmo	Raiz	Raiz escora	Grãos
Ácido Orgânico	5	13	0,245	0,001	0,005	25,8	22,6	16,5	10,9	24,1
Todos	5	73	0,135	0,002	0,005	26,5	16,2	18,2	15,6	23,5
Carboidrato	5	18	0,119	0,003	0,005	5,5	8,2	4,9	36,2	45,3
Nucleotídeo	5	12	0,174	0,098	0,123	19	8,7	35,7	13,1	23,5
Aminoácidos	5	24	0,085	0,33	0,33	30,2	23,6	6	21	19,3

Referências Bibliográficas

- Barabasi, Albert-Laszlo e Reka Albert. 1999. “Emergence of Scaling in Random Networks”. *Science (New York, N.Y.)* 286(October):509–12.
- Barabási, Albert-László e Zoltán N. Oltvai. 2004. “Network biology: understanding the cell’s functional organization.” *Nature reviews. Genetics* 5(2):101–13.
- Benjamini, Yoav e Yosef Hochberg. 1995. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. *Journal of the Royal Statistical Society* 57(1):289–300.
- Bollobás, Béla, Oliver Riordan, Joel Spencer, e Gábor Tusnády. 2001. “The degree sequence of a scale-free random graph process”. *Random Structures and Algorithms* 18(3):279–90.
- Bonacich, Phillip. 1987. “Power and Centrality: A Family of Measures”. *The University of Chicago Press* 92(5):1170–82.
- Bonacich, Phillip. 1972. “The Journal of Mathematical Sociology Factoring and weighting approaches to status scores and clique identification”. *Journal of Mathematical Sociology* © Gordon and Breach Science Publishers 2:113–20. Recuperado (<http://www.tandfonline.com/loi/gmas20>).
- Brosicke, Nicole e Andreas Faissner. 2015. “Role of tenascins in the ECM of gliomas”. *Cell Adhesion and Migration* 9(1–2):131–40.
- Caldana, Camila et al. 2011. “High-density kinetic analysis of the metabolomic and transcriptomic response of Arabidopsis to eight environmental conditions”. *Plant Journal* 67(5):869–84.
- Cho, Sung Bum, Jihun Kim, e Ju Han Kim. 2009. “Identifying set-wise differential co-expression in gene expression microarray data”. *BMC Bioinformatics* 10(1):109.
- Choi, Jung Kyoon, Ungsik Yu, Ook Joon Yoo, e Sangsoo Kim. 2005. “Differential coexpression analysis using microarray data and its application to human cancer”. *Bioinformatics* 21(24):4348–55.
- Choi, Younjeong e Christina Kendzierski. 2009. “Statistical methods for gene set co-expression analysis”. *Bioinformatics*.
- Dawson, John A., Shuyun Ye, e Christina Kendzierski. 2012. “R/ebcoexpress: An empirical bayesian framework for discovering differential co-expression”. *Bioinformatics* 28(14):1939–40.
- Diestel, Reinhard. 2000. *Graph Theory (Graduate Texts in Mathematics)*. Recuperado (<http://www.math.uni-hamburg.de/home/diestel/books/graph.theory/%5Cnhttp://www.amazon.com/Graph-Theory-Graduate-Texts-Mathematics/dp/3540261834>).
- Ding, Yuduan et al. 2015. “Network Analysis of Postharvest Senescence Process in Citrus Fruits Revealed by Transcriptomic and Metabolomic Profiling”. *Plant Physiology*

- 168(May):pp.114.255711. Recuperado
(<http://www.plantphysiol.org/lookup/doi/10.1104/pp.114.255711>).
- Dwass, Meyer. 1957. “Modified Randomization Tests for Nonparametric Hypotheses”. *The Annals of Mathematical Statistics* 28(1):181–87. Recuperado
(<http://projecteuclid.org/euclid.imsc/1207580085>).
- Efron, B. 1979. “Bootstrap Methods: Another Look at the Jackknife”. *The Annals of Statistics* 7(1):1–26. Recuperado
(<http://projecteuclid.org/euclid.aos/1176344552>)(<https://projecteuclid.org/euclid.aos/1176344552>).
- Emmert-Streib, Frank e M. Dehmer. 2011. “Networks for systems biology: conceptual connection of data and function”. *IET systems biology* 5(3):185–207.
- Erdős, P. e a Rényi. 1959. “On random graphs”. *Publicationes Mathematicae* 6:290–97.
- Fisher, Sir Ronald A. 1935. “The Design of Experiments”. *Hafner Publishing Company* 1–27.
- Freeman, Linton C. 1978. “Centrality in social networks conceptual clarification”. *Social Networks* 1(3):215–39.
- Fujita, André, Maciel C. Vidal, e Daniel Y. Takahashi. 2017. “A Statistical Method to Distinguish Functional Brain Networks”. *Frontiers in Neuroscience* 11(February). Recuperado
(<http://journal.frontiersin.org/article/10.3389/fnins.2017.00066/full>).
- Fukushima, Atsushi. 2013. “DiffCorr: An R package to analyze and visualize differential correlations in biological networks”. *Gene* 518(1):209–14. Recuperado
(<http://dx.doi.org/10.1016/j.gene.2012.11.028>).
- Grandis, Adriana, Simone Godoi, e Marcos Silveira Buckeridge. 2010. “Respostas fisiológicas de plantas amazônicas de regiões alagadas às mudanças climáticas globais”. *Revista Brasileira de Botânica* 33(1):1–12.
- Ha, Min Jin, Veerabhadran Baladandayuthapani, e Kim Anh Do. 2015. “DINGO: Differential network analysis in genomics”. *Bioinformatics* 31(21):3413–20.
- Hochberg, Uri et al. 2013. “Metabolite profiling and network analysis reveal coordinated changes in grapevine water stress response”. *BMC Plant Biology* 13(1):184. Recuperado
(<http://dx.plos.org/10.1371/journal.pone.0115581>)(<http://www.biomedcentral.com/1471-2229/13/184>).
- Jeong, H., S. P. Mason, a L. Barabási, e Z. N. Oltvai. 2001. “Lethality and centrality in protein networks”. *Nature* 411(6833):41–42.
- Jeong, Hawoong, B. Tombor, Reka Albert, Zoltan N. Oltvai, e Albert-Laszlo Barabasi. 2000. “The

- large-scale organization of metabolic networks”. *Nature* 407(6804):651–54. Recuperado (<http://www.hubmed.org/display.cgi?uids=11034217>).
- Junker, Björn H. e Falk Schreiber. 2008. *Analysis of biological networks*. organizado por B. H. Junker e F. Schreiber. John Wiley & Sons, Inc. All.
- Kauffman, S. A. 1969. “Metabolic stability and epigenesis in randomly constructed genetic nets.” *Journal of theoretical biology* 22(3):437–67.
- Kendall, M. G. 1938. “A New Measure of Rank Correlation”. *Biometrika* 30(1):81–93.
- Kinker, Gabriela Sarti, Andrew Maltez Thomas, Vinicius Jardim Carvalho, Felipe Prata Lima, e André Fujita. 2016. “Deletion and low expression of NFKBIA are associated with poor prognosis in lower-grade glioma patients”. *Scientific Reports* 6(April):24160. Recuperado (<http://www.nature.com/articles/srep24160>).
- Kitano, Hiroaki. 2002. “Systems biology: a brief overview.” *Science (New York, N.Y.)*.
- Koch, C. e G. Laurent. 1999. “Complexity and the nervous system”. *Science* 284(April):96–98. Recuperado (<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.42.727&rep=rep1&type=pdf>).
- de la Fuente, Alberto. 2010. “From ‘differential expression’ to ‘differential networking’ - identification of dysfunctional regulatory networks in diseases”. *Trends in Genetics* 26(7):326–33.
- Langfelder, Peter e Steve Horvath. 2008. “WGCNA: an R package for weighted correlation network analysis.” *BMC bioinformatics* 9:559.
- Legendre, Pierre e Louis Legendre. 2003. *Numerical Ecology*. 2nd ed organizado por Elsevier.
- Li, Junyi, Yi Xue Li, e Yuan Yuan Li. 2016. “Differential Regulatory Analysis Based on Coexpression Network in Cancer Research”. *BioMed Research International* 2016(4241293):8.
- Liu, Bao Hong et al. 2010. “DCGL: An R package for identifying differentially coexpressed genes and links from gene expression microarray data”. *Bioinformatics* 26(20):2637–38.
- Lopez-Fernandez, L., G. Robles, e J. M. Gonzalez-Barahona. 2004. “Applying Social Network Analysis to the Information in CVS Repositories”. *International Workshop on Mining Software Repositories Edinburgh Scotland United Kingdom 25th May 2004* 2004(917):101–5. Recuperado (<http://gsync.escet.urjc.es/~jgb/libresofteng/sna-repositories-2004/sna-repositories-2004.pdf>).
- Louis, David N. et al. 2016. “The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary”. *Acta Neuropathologica* 131(6):803–20.
- Lui, Thomas W. H. et al. 2015. “DECODE: an integrated differential co-expression and differential expression analysis of gene expression data.” *BMC bioinformatics* 16(1):182. Recuperado (<http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0582-4>).

- McKenzie, Andrew T., Igor Katsyv, Won-Min Song, Minghui Wang, e Bin Zhang. 2016. “DGCA: A comprehensive R package for Differential Gene Correlation Analysis.” *BMC systems biology* 10(1):106. Recuperado (<http://www.ncbi.nlm.nih.gov/pubmed/27846853><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5111277>).
- Mieghem, Piet Van. 2010. “Graph spectra of complex networks”. *Cambridge University Press* 1–364. Recuperado (<http://medcontent.metapress.com/index/A65RM03P4874243N.pdf>).
- Mutwil, Marek et al. 2010. “Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm.” *Plant physiology* 152(1):29–43.
- Pearson, Karl. 1920. “Notes on the history of correlation.” *Biometrika* 2(1):209–13.
- Purves, Dale et al. 2001. *Neuroscience*. 2nd ed Sunderland (MA): Sinauer Associates.
- Rahmatallah, Yasir, Frank Emmert-Streib, e Galina Glazko. 2014. “Gene Sets Net Correlations Analysis (GSNCA): A multivariate differential coexpression test for gene sets”. *Bioinformatics* 30(3):360–68.
- Ronchini, Chiara e Anthony J. Capobianco. 2001. “Induction of Cyclin D1 Transcription and CDK2 Activity by Notch ic : Implication for Cell Cycle Disruption in Transformation by Notch ic Induction of Cyclin D1 Transcription and CDK2 Activity by Notch ic : Implication for Cell Cycle Disruption in Transfor”. *Molecular and cellular biology* 21(17):5925–34.
- Roy, Swarup, Dhruba K. Bhattacharyya, e Jugal K. Kalita. 2014. “Reconstruction of gene co-expression network from microarray data using local expression patterns.” *BMC bioinformatics* 15 Suppl 7(Suppl 7):S10. Recuperado (<http://www.biomedcentral.com/1471-2105/15/S7/S10>).
- Sain, Stephan R., David W. Scott, Stephan R. Sain, e David W. Scott. 1996. “On Locally Adaptive Density Estimation”. *Journal of the American Statistical Association* 91(436):1525–34. Recuperado (<http://www.jstor.org/stable/2291578>).
- Santos, Suzana de Siqueira. 2015. “Analise de redes biologicas: estudo comparativo de medidas de dependencia e uma ferramenta computacional para dicriminar grafos”. Universidade de Sao Paulo.
- Santos, Suzana de Siqueira et al. 2015. “CoGA: An R package to identify differentially co-expressed gene sets by analyzing the graph spectra”. *PLoS ONE*.
- Santos, Suzana de Siqueira, Daniel Yasumasa Takahashi, Asuka Nakata, e André Fujita. 2013. “A comparative study of statistical methods used to identify dependencies between gene expression signals”. *Briefings in Bioinformatics* 15(6):906–18.
- Schnyder, Hans. 1993. “The role of carbohydrate storage and redistribution in the source-sink relations of wheat and barley during grain filling -a review”. *Nezc Phytol* 123(November):233–45.

- Shannon, Claude Edmund. 1948. "A Mathematical Theory of Communication". *Bell System Technical Journal* 5(3):3. Recuperado (<http://portal.acm.org/citation.cfm?doid=584091.584093%5Cnhttp://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6773024>).
- Shannon, Paul et al. 2005. "Cytoscape: a software environment for integrated models of biomolecular interaction networks". *American Association for Cancer Research Education Book* (Karp 2001):12–16.
- Sicher, Richard C. e Jinyoung Y. Barnaby. 2012. "Impact of carbon dioxide enrichment on the responses of maize leaf transcripts and metabolites to water stress". *Physiologia Plantarum* 144(3):238–53.
- de Souza, Amanda P. et al. 2008. "Elevated CO₂ increases photosynthesis, biomass and productivity, and modifies gene expression in sugarcane". *Plant, Cell and Environment* 31(8):1116–27.
- de Souza, Amanda P., Jean-Christophe Cocuron, Ana Carolina Garcia, Ana Paula Alonso, e Marcos S. Buckeridge. 2015. "Changes in Whole-Plant Metabolism During Grain-Filling Stage in Sorghum Bicolor L. (Moench) Grown Under Elevated Co₂ and Drought". *Plant Physiology* (July 2016):pp.01054.2015. Recuperado (<http://www.plantphysiol.org/lookup/doi/10.1104/pp.15.01054>).
- Spearman, C. 1904. "The Proof and Measurement of Association between Two Things". *The American Journal of Psychology* 15(1):72–101.
- Sturges, Herbert A. 1926. "The Choice of a Class Interval". *Journal of the American Statistical Association* 21(153):65–66. Recuperado (<http://dx.doi.org/10.1080/01621459.1926.10502161>).
- Subramanian, Aravind et al. 2005. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102(43):15545–50. Recuperado (<http://www.ncbi.nlm.nih.gov/pubmed/16199517>).
- Sun, Shao-Yan, Zhi-Ping Liu, Tao Zeng, Yong Wang, e Luonan Chen. 2013. "Spatio-temporal analysis of type 2 diabetes mellitus based on differential expression networks." *Scientific reports* 3:2268. Recuperado (<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3721080&tool=pmcentrez&rendertype=abstract>).
- Takahashi, Daniel Yasumasa, Joao Ricardo Sato, Carlos Eduardo Ferreira, e Andre Fujita. 2012. "Discriminating Different Classes of Biological Networks by Analyzing the Graphs Spectra Distribution". *PLoS ONE*.
- Tesson, Bruno M., Rainer Breitling, e Ritsert C. Jansen. 2010. "DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules." *BMC bioinformatics* 11(1):497. Recuperado (<http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-497>).

- Walley, A. J. et al. 2012. "UKPMC Funders Group Differential co-expression analysis of obesity-associated networks in human subcutaneous adipose tissue". *International Journal* 36(1):137–47.
- Watson, Michael. 2006. "CoXpress: differential co-expression in gene expression data." *BMC bioinformatics* 7:509. Recuperado (<http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?dbfrom=pubmed&id=17116249&retmode=ref&cmd=prlinks> %5Cnpapers3://publication/doi/10.1186/1471-2105-7-509).
- Watts, D. J. e S. H. Strogatz. 1998. "Collective dynamics of 'small-world' networks". *Nature* 393(6684):440–42.
- Weston, David J. et al. 2011. "Comparative physiology and transcriptional networks underlying the heat shock response in *Populus trichocarpa*, *Arabidopsis thaliana* and *Glycine max*". *Plant, Cell and Environment* 34(9):1488–1506.
- Zhang, Huanping e Tongming Yin. 2016. "Identifying candidate genes for wood formation in poplar based on microarray network analysis and graph theory". *Tree Genetics & Genomes* 12(3):61. Recuperado (<http://link.springer.com/10.1007/s11295-016-1016-9>).
- Zhu, Jun, Bin Zhang, e Eric E. Schadt. 2008. "A Systems Biology Approach to Drug Discovery". *Advances in Genetics* 60(07):603–35.