Análise de redes biológicas: estudo comparativo de medidas de dependência e uma ferramenta computacional para discriminar grafos

Suzana de Siqueira Santos

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Ciência da Computação

Orientador: Prof. Dr. André Fujita

Durante o desenvolvimento deste trabalho a autora recebeu auxílio financeiro da FAPESP
processo 2012/25417-9

São Paulo, abril de 2015

Análise de redes biológicas: estudo comparativo de medidas de dependência e uma ferramenta computacional para discriminar grafos

Esta é a versão original da dissertação elaborada pela candidata Suzana de Siqueira Santos submetida à Comissão Julgadora.

Agradecimentos

Este trabalho é resultado do empenho de diversas pessoas e instituições. Assim, não poderia deixar de agradecê-las e de reconhecer o quanto elas foram importantes para o meu desenvolvimento e para a produção deste trabalho.

Sou muito grata ao IME e à USP por viabilizarem um ensino de qualidade a tantas pessoas, o qual tive o privilégio de usufruir. Nada disso seria possível sem o empenho de tantos funcionários e professores e sem o apoio da sociedade para a manutenção da Universidade de São Paulo. Em especial gostaria de agradecer ao Geraldo, por trabalhar com tamanha boa vontade e ser extremamente gentil com todos. Agradeço à Lucileide não só por ajudar-me com todas as questões burocráticas que precisei, mas, principalmente, por ser sempre tão gentil e prestativa. Não poderia deixar de mencionar a Márcia, a Adenilza, a Ana Carla e a Edna, que estão sempre prontas para ajudar.

Agradeço à FAPESP (processo # 2012/25417-9) pela bolsa de estudos concedida e ao grupo de eScience do IME (apoio financeiro da FAPESP processo # 2011/50761-2, do CNPq, da CAPES e do NAP eScience - PRP - USP) pela rede de computadores que possibilitou as execuções em larga escala que precisei realizar. Em especial agradeço ao Prof. Hirata por se importar genuinamente com os alunos e não medir esforços para viabilizar e aperfeiçoar a rede eScience. Não poderiam faltar agradecimentos aos administradores da rede, David Pires e Jorge, que me auxiliaram com grande disposição.

Todo o desenvolvimento do trabalho foi acompanhado de perto pelo Prof. André Fujita, meu orientador do mestrado, a quem sou profundamente grata por toda a ajuda fornecida e pela enorme paciência e dedicação. Agradeço a ele, sobretudo, por me fazer acreditar em mim mesma e despertar meu interesse por pesquisa, ensinando-me o primeiros passos da vida acadêmica. Não tenho dúvidas de que as tarefas designadas a mim pelo Prof. André trouxeram ensinamentos imensuráveis que levarei para toda a vida.

Graças ao meu ingresso no mestrado, tive contato com a Profa. Suely Marie e a Profa. Sueli Oba da Faculdade de Medicina da USP, que me deram apoio nas questões biológicas do meu trabalho, sempre com muita paciência e disposição. Agradeço imensamente pela confiança que depositaram em mim e por me receberem com tanta consideração em seu laboratório LIM 15. Nas minhas visitas ao laboratório, contei com o apoio do Antonio, da Miyuki, da Paula, do Rodrigo, e da Thais.

Gostaria de agradecer ao Prof. Ronaldo, ao Prof. Carlinhos e ao Prof. Hirata pelas sugestões dadas no exame de qualificação e por também serem professores excelentes em

diversos aspectos: na pesquisa, na didática e, sobretudo, no apoio dado aos alunos.

O desenvolvimento deste trabalho não teria sido tão prazeroso sem o nosso grupo de pesquisa. Em particular, a Adèle, o Fernando, a Gabriela, o Gustavo, o Juan, o Maciel e o Paulo me fizeram grande companhia nas atividades de pós-graduação. Mais recentemente, conheci o Davi, o Grover e a Taiane, com quem tive o prazer de trocar conhecimentos. Agradeço também ao Eduardo e ao Abner pelos conhecimentos compartilhados. Fico feliz de ter acompanhado um pouco do trabalho de iniciação científica do Yuri, a quem admiro pela grande vontade de ajudar. Também agradeço à Carolina, à Stéphannye e ao Allan, pelos ótimos momentos quando fui monitora.

Agradeço a todos do grupo por deixarem o meu cotidiano tão divertido e alegre e por estarem sempre dispostos a ajudar. Em especial, agradeço ao Fernando por revisar o capítulo de preliminares da dissertação. Agradeço ao Maciel pelas diversas dicas sobre a rede eScience e sobre os pacotes do R, que foram extremamente úteis para mim. Ao Gustavo, agradeço por me ajudar inúmeras vezes com as burocracias da FAPESP e, sobretudo, pelos maravilhosos milk-shakes, pela preocupação comigo e pelos conselhos que me deu e desabafos que ouviu.

Agradeço aos colegas do laboratório de eScience, Amanda Rusiska, Amanda Sayuri, Anderson, Caio, David, Éric, Gesiele, Igor, Jihan, Jorge, Leandro, Lucy, Lulu, Sérgio, Silvia e Urpy, por serem tão gentis comigo e deixarem os meus dias mais alegres. Em especial, agradeço ao Anderson pelas conversas sobre óperas e outros espetáculos musicais e ao Sérgio pelo interesse neste trabalho e pelos conselhos dados. Agradeço ao meu colega de pós-graduação, Milson, por sua simpatia fora do comum e seu interesse neste trabalho.

Agradeço aos amigos que conheci durante a graduação, Brócolis, Celso, Coelho, Felipe, Goroba, Haruki, Henrique, Jackson, Jefferson, Jéssica, Katague, Manzo, Miojo, Mônica, Ná, Omar, Paulo Haddad, Renato Vieira, Samu, Wallace e Wilson, por tornarem o BCC tão divertido. Não posso deixar de agradecer de forma especial à Ná, pelo imenso carinho, por continuar tão presente na minha vida e me levar para tomar milk-shake. Agradeço especialmente à Mônica e ao Celso por compartilharem suas histórias tão inspiradoras comigo. À minha veterana Susanna, sou muito grata pelo apreço e por me convidar para participar de atividades sociais tão enriquecedoras.

Gostaria de agradecer aos RCs, Aninha, André Yai, João, Lucas Dario, Ludmila, Nathan, Renato Cordeiro e Victor, por tornarem a organização do Encontro do BCC algo tão divertido. Sou muito grata ao Renato Cordeiro, por sempre demonstrar interesse pelas minhas atividades acadêmicas, fazendo com que eu me sinta ainda mais motivada. Além disso, sinto-me inspirada por sua grande dedicação ao BCC.

Gostaria de agradecer ao Jackson por ser um dos melhores amigos que eu poderia ter. Agradeço por sua lealdade e dedicação e por estar sempre presente, dando apoio nos momentos difíceis. Também sou grata pelas grandes contribuições que ele deu ao BCC e pelas inúmeras conversas enriquecedoras.

Agradeço às minhas amigas da escola, Loly, Nébs, Amanda, Keyla e Ju, pelos momentos de alegria e pela oportunidade de crescer ao lado delas. Agradeço à Tânia, amiga da minha

família, pela afeição e por seus trabalhos de caridade tão inspiradores.

Faço um agradecimento especial ao Samuel que, com amor, paciência e compreensão, me apoiou de todas as formas possíveis. Sua presença me faz mais alegre e mais forte e é fundamental na minha vida. Não poderia deixar de agradecer à sua família, ao Depa, à Luiza e à Jacque, pelo imenso apoio e carinho.

Devo tudo o que sou ao empenho incansável dos meus pais. Sou grata a eles por todo amor e apoio que recebi, tão fundamentais para o meu desenvolvimento como ser humano. Em particular, agradeço à minha mãe Juliana pelo grande amor e amizade, sua participação em minha vida e por infalivelmente zelar pelo meu bem. Sou muito grata ao meu pai Toninho por sua alegria contagiante, seu grande coração e pela dedicação e incentivos que foram fundamentais na minha vida. Apesar de infelizmente ter falecido antes mesmo de eu ingressar na pós, ele inspirou, e continua inspirando, a minha vontade de estudar doenças como o câncer, o que me fez chegar a este trabalho. O seu amor à vida e sua fé, que jamais se abalaram, mesmo nas circunstâncias mais difíceis, são inestimáveis exemplos que levarei sempre comigo.

Sou profundamente grata às minhas irmãs Gabi e Cris, que são minhas melhores amigas. Agradeço pelo carinho, companhia diária e momentos de profunda alegria. Com elas, sempre estive à vontade para compartilhar as minhas inesgotáveis dúvidas, especialmente, durante minha infância e adolescência. Agradeço pelas dicas sobre os mais variados assuntos e sobre o vestibular, que tanto me ajudaram.

Não poderia deixar de agradecer à minha cachorra Lady, que é uma das melhores companheiras que eu poderia ter. Agradeço pela afeição, por ser uma fonte inesgotável de alegria e por estar infalivelmente ao meu lado, inclusive nas horas de estudo. Não tenho dúvidas de que cuidar dela me ajuda a ser uma pessoa melhor.

Agradeço à minha avó Toninha, que desempenhou um papel fundamental na minha formação moral, pelos seus deliciosos almoços de domingo e por seu imenso carinho, cuidado e dedicação. Agradeço, *in memoriam*, aos meus avós Alberto, José e Maria Georgina, pelos grandes exemplos que deram e pelo inesgotável carinho.

Sou profundamente grata à minha tia Lolinha, que é uma segunda mãe para mim. Ela jamais mediu esforços para ajudar a família de todas as formas possíveis, doando o seu tempo e carinho. Agradeço à tia Inês pelo grande zelo e preocupação comigo e por enriquecer a minha vida com tantos momentos de alegria. Ao tio Chico e à tia Sandra agradeço pelo imenso carinho e incentivo que sempre me deram.

Agradeço por todas as alegrias proporcionadas pelas minhas primas Luiza, Julia e Mariana, pelos primos da minha mãe e pelos primos do meu pai. Agradeço aos familiares de Mococa e de Cajuru por me receberem sempre tão bem e ajudarem a renovar minhas forças nos períodos de descanso.

Esta dissertação não existiria não fosse a presença de tantas pessoas boas. Agradeço a Deus pela vida de cada uma delas e pelas responsabilidades e desafios que encontrei e que me fazem hoje uma pessoa melhor.

Resumo

SANTOS, S. S. **Análise de redes biológicas: estudo comparativo de medidas de dependência e uma ferramenta computacional para discriminar grafos**. 2015. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2015.

Resumo: Redes complexas de interações moleculares descrevem o fenótipo celular. Assim, identificar as propriedades de redes que diferenciam o estado doente e saudável de uma célula pode trazer esclarecimentos sobre os mecanismos envolvidos em uma doença. Para estudar esse tipo de rede, são utilizados dados de apenas parte da população. Assim, métodos de inferência estatística são fundamentais no estudo de redes biológicas. Neste trabalho, nos focamos no estudo de grafos de coexpressão, em que os vértices correspondem a genes e as arestas indicam associações estatísticas entre os níveis de expressão genética. Na primeira parte do trabalho, realizamos um estudo comparativo entre medidas de dependência estatística utilizadas para construir grafos de coexpressão. Por meio de simulações e aplicações das medidas de dependência em dados de microarranjos de DNA oriundos de tecidos tumorais, identificamos potencialidades e limitações dos métodos estudados (o coeficiente de correlação de Pearson, o coeficiente de correlação de Spearman, o coeficiente de correlação de Kendall, a correlação de distância, a medida de Heller-Heller-Gorfine, a medida D de Hoeffding, a informação mútua e o coeficiente de informação máxima). Na segunda parte do trabalho, desenvolvemos testes estatísticos para comparar propriedades estruturais de grafos de coexpressão. Nesses testes utilizamos medidas de redes complexas para caracterizar os grafos, como a centralidade de grau, a centralidade de *betweenness*, a centralidade de proximidade, a centralidade de autovetor e o coeficiente de *clustering* e duas medidas recentemente propostas que se baseiam no espectro do grafo (conjunto de autovalores da matriz de adjacência). A escolha do espectro se baseou no fato de ele descrever diversas propriedades estruturais do grafo, sendo considerado uma caracterização mais completa do que as principais medidas de redes complexas. As medidas baseadas no espetro utilizadas neste trabalho são: a entropia espectral (medida de aleatoriedade de um grafo) e a divergência de Jensen-Shannon entre as distribuições dos espectros dos grafos. Os testes desenvolvidos foram disponibilizados em um pacote do R chamado CoGA (*Co-expression Graph Analyzer*). Uma aplicação do CoGA é ilustrada em dados de microarranjos de DNA de dois tipos de câncer no cérebro. Nós mostramos com simulações que os testes propostos controlam a taxa de falsos positivos e

que o poder estatístico cresce à medida que aumentamos a proporção de arestas modificadas na rede. Nossos resultados sugerem que a ferramenta apresentada (CoGA) pode ser útil na identificação de conjuntos de genes associados a uma doença.

Palavras-chave: medidas de dependência estatística, grafos, coexpressão, redes de regulação genética, redes complexas.

Abstract

SANTOS, S. S. **Analysis of biological networks: comparative study of statistical dependence measures and a computational tool to discriminate graphs**. 2015. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2015.

Complex networks of molecular interactions describe the cellular phenotype. Therefore, identifying network properties that are different between healthy and diseased cellular state may elucidate the mechanisms that are involved in a disease. Studies of that kind of network usually analyze data from part of the population. Thus, statistical inference methods are fundamental to study biological networks. In this work, we focus on the analysis of co-expression graphs, in which the vertices correspond to genes and the edges indicate statistical associations between the gene expression levels. In the first part of this work, we present a comparative study of statistical dependence measures used to construct co-expression graphs. We have performed simulation experiments and applications of the methods on microarray data from tumor tissues to evaluate the strengths and limitations of the studied measures (the Pearson's correlation coefficient, the Spearman's correlation coefficient, the Kendall's correlation coefficient, the distance correlation, the Heller-Heller-Gorfine measure, the Hoeffding's D measure, the mutual information, and the maximum information coefficient). In the second part of the work, we have developed statistical tests to compare structural properties of co-expression graphs. To characterize a graph, we used complex network measures, such as the degree centrality, the betweenness centrality, the closeness centrality, the eigenvector centrality and the clustering coefficient, and two recently proposed measures that are based on the graph spectrum (set of eigenvalues of the graph adjacency matrix). A motivation to use the spectrum of a graph is based on the fact that it describes several structural properties of a graph and is considered a more complete graph characterization than the usual complex network measures. The spectrum-based measures used in this work are the spectral entropy (measure of the graph randomness), and the Jensen-Shannon divergence between the distributions of the graph spectra. To make the proposed methods available, we have developed an R package called CoGA (Co-expression Graph Analyzer). We illustrate an application of the CoGA package on microarray data from two types of brain tumor. We show by simulation experiments that the proposed tests control the false positive rate and that their power is proportional to the number of changes in the network. Our results suggest that

the CoGA package may be useful for the identification of gene sets associated with a disease.

Keywords: statistical dependence measures, graphs, co-expression, gene regulatory networks, complex networks.

Sumário

Lista de Figuras

Lista de Tabelas

Introdução

Doenças complexas, como Câncer, Alzheimer, Parkinson e hipertensão, são doenças causadas por uma combinação de diversos fatores, como condições ambientais e a ação de múltiplos genes. Uma das formas de estudar, em nível molecular, mecanismos envolvidos nessas doenças, é analisando a diferença de atividades de milhares de genes entre organismos saudáveis e doentes. Trata-se fundamentalmente de um problema de inferência estatística: a partir de algumas amostras biológicas, deseja-se identificar atividades genéticas envolvidas com uma doença, que possam explicá-la em toda a população de organismos doentes. Assim, métodos estatísticos para analisar as atividades genéticas de grupos de indivíduos são de grande importância no estudo de doenças complexas.

Para quantificar o nível de atividade de milhares de genes em amostras biológicas são utilizados dados de expressão genética que podem ser obtidos, por exemplo, a partir de experimentos de microarranjos de DNA. Uma das abordagens mais utilizadas para analisar esse tipo de dado é o teste da igualdade do valor da expressão média de cada gene entre indivíduos doentes e saudáveis, por exemplo. Contudo, devido ao grande número de testes realizados (geralmente na ordem de milhares, sendo um para cada gene), esse método apresenta pouco poder estatístico. Além disso, quando genes são testados isoladamente, a interpretação biológica dos resultados pode ser difícil.

Uma estratégia adotada para agregar informação biológica às análises dos dados de microarranjos de DNA e aumentar seu poder estatístico é testar a igualdade da expressão média em conjuntos de genes funcionalmente relacionados, ao invés de considerar genes isoladamente (Efron e Tibshirani, 2007; Irizarry *et al.*, 2009; Jiang e Gentleman, 2007; Subramanian *et al.*, 2005). Esse tipo de teste é baseado na ideia de que uma doença complexa é raramente consequência de alterações em um único gene, mas resultado de mudanças em um conjunto de genes relacionados. Apesar de apresentarem aplicações bem sucedidas, os métodos que comparam a expressão genética média de conjuntos de genes (vias), não identificam classes importantes de vias diferencialmente reguladas, como os grupos de genes diferencialmente coexpressos.

A *coexpressão* de dois genes é a correlação entre seus níveis de expressão. Um grafo em que os vértices representam os genes e as arestas indicam a dependência estatística entre os níveis de expressão dos genes é chamado de *grafo de coexpressão*. Assim, um grafo de coexpressão guarda informações sobre as relações entre as atividades genéticas e pode ser utilizado para modelar o sistema de regulação dos genes. Se a estrutura do grafo de

coexpressão de um grupo de genes é diferente entre dois conjuntos de dados, dizemos que esse grupo é *diferencialmente coexpresso*.

Podemos dividir a análise da diferença de coexpressão de conjuntos de genes em duas partes: (i) a construção do grafo de coexpressão a partir de dados de expressão genética e (ii) métodos estatísticos para comparar grafos de coexpressão. Na parte (i), é utilizada uma medida de dependência para identificar associações entre os níveis de expressão dos genes. Como existem diversas medidas, com diferentes finalidades e características, realizamos um estudo comparativo a fim de facilitar a escolha do método mais adequado para realizar a parte (i). As medidas de dependência estatística incluídas no estudo são o coeficiente de correlação de Pearson (Pearson, 1920), o coeficiente de correlação de Spearman (Spearman, 1904), o coeficiente de correlação de Kendall (Kendall, 1938), a correlação de distância (Székely *et al.*, 2007), o método de Heller-Heller-Gorfine (HHG), a medida D de Hoeffding (Hoeffding, 1948), (Heller *et al.*, 2013), a informação mútua (Shannon, 1948) (IM) e o coeficiente de informação máxima (Reshef *et al.*, 2011) (CIM).

Nossos experimentos de simulação e a aplicação dos métodos em dados obtidos a partir de microarranjos de DNA sugerem que, quando supomos que as relações entre duas variáveis aleatórias (níveis de expressão genética) são lineares ou monotônicas, como é usual no estudo de associações entre produtos genéticos, as medidas de Pearson, Spearman e Kendall apresentam maior poder estatístico, sendo que as duas últimas são mais robustas à presença de *outliers*. A correlação de distância, as medidas de Hoeffding, HHG, IM, e CIM são apropriadas quando se buscam formas mais gerais de associação, como as relações não-monotônicas. Para amostras pequenas com, por exemplo, menos que 30 observações, a medida de HHG apresentou o maior poder estatístico. Já, na parte (ii), desenvolvemos métodos que testam a igualdade entre propriedades estruturais de dois grafos de coexpressão.

Uma das abordagens existentes para identificar grupos de genes diferencialmente coexpressos é o GSCA (Choi e Kendziorski, 2009), que testa se a distância euclidiana entre a coexpressão de pares genes de dois conjuntos de dados é zero. Outros métodos se baseiam na ideia de que sistemas biológicos são mais suscetíveis a mudanças nas atividades de genes "importantes" do que alterações isoladas de coexpressão (Barabási e Oltvai, 2004). O método GSNCA (Rahmatallah *et al.*, 2014), por exemplo, considera que a "importância" de um gene v_i é proporcional à soma da "importância" dos demais genes ponderada pela sua coexpressão com v_i. Para identificar vias disfuncionais, o GSNCA testa a igualdade da "importância" dos genes entre duas condições biológicas. A classe de conjuntos de genes diferencialmente coexpressos detectada pelo GSNCA é diferente da detectada pelo GSCA (Rahmatallah *et al.*, 2014).

A medida de centralidade utilizada pelo GSNCA, também conhecida como centralidade de autovetor é apenas uma entre diversas medidas de propriedades estruturais de grafos, como a centralidade de grau, a centralidade de *betweenness*, a centralidade de proximidade e o coeficiente de *clustering*. Essas medidas são utilizadas para analisar grafos de coexpressão por ferramentas como o WGCNA (Langfelder e Horvath, 2008) e o Cytoscape

(Shannon *et al.*, 2003). Contudo, diferentemente do GSNCA, essas ferramentas não realizam testes estatísticos para a identificação de diferenças entre os grafos de coexpressão.

Procurar por uma estrutura exatamente igual entre dois grafos não é efetivo para comparar o funcionamento de sistemas biológicos, como eles podem variar dentro de indivíduo ou entre indivíduos de um mesmo grupo biológico. Para lidar com essa variabilidade, uma abordagem natural seria comparar propriedades estatísticas que são compartilhadas por grafos de uma mesma classe biológica, mas são diferentes entre classes distintas.

O espectro de um grafo, definido como o conjunto de autovalores de sua matriz de adjacência, descreve diversas propriedades estruturais de um grafo, como o seu diâmetro, número de passeios e cliques. Segundo Takahashi *et al.* (2012), a distribuição do espectro do grafo é uma caracterização mais completa de uma classe de grafos quando comparada com medidas como o número de arestas, a média dos comprimentos dos caminhos mais curtos e o coeficiente de *clustering*. A partir da descrição do grafo pelo seu espectro, Takahashi *et al.* (2012) introduziram conceitos de Teoria da Informação para grafos, como a entropia espectral e a divergência de Jensen-Shannon entre as funções de densidade de probabilidade do espectro (densidades espectrais). A primeira mede a quantidade de incerteza (aleatoriedade) associada a um grafo e a segunda é utilizada para diferenciar classes de grafos.

As medidas propostas por Takahashi *et al.* (2012) identificaram mudanças estruturais em redes cerebrais construídas com dados de ressonância magnética funcional. Os testes estatísticos dessas medidas são aplicáveis em conjuntos de dados com diversos grafos (por exemplo, um grafo por indivíduo). Neste trabalho, nós adaptamos os testes da entropia espectral e da divergência de Jensen-Shannon para comparar apenas dois grafos (um de cada condição biológica), como é feito na análise de grafos de coexpressão, em que apenas um grafo é construído a partir da observação das atividades genéticas de diversos indivíduos de uma mesma condição biológica.

Para disponibilizar os testes desenvolvidos, criamos uma ferramenta com licença de software livre chamada CoGA (*Co-expression Graph Analyzer*). O programa constrói grafos de coexpressão e, a partir de uma coleção pré-definida de conjuntos de genes, identifica os conjuntos diferencialmente coexpressos (cada um representando um subgrafo). A análise de diferença de coexpressão é feita por meio de testes da igualdade de propriedades estruturais entre dois grafos, como a centralidade de grau, a centralidade de *betweenness*, a centralidade de proximidade, a centralidade de autovetor, o coeficiente de *clustering* a distribuição do grau, a distribuição do espectro e a entropia espectral. Um dos diferenciais do CoGA em relação às ferramentas disponíveis é que ele realiza testes estatísticos para uma grande variedade de medidas de grafos, além de incluir ferramentas para analisar cada conjunto de genes testado, como a visualização da coexpressão dos genes do conjunto, um ranking dos genes de acordo com as medidas de centralidade e a análise clássica de diferença de expressão média. Nós mostramos com simulações de Monte Carlo que os testes propostos controlam a taxa de falsos positivos e que o poder estatístico cresce à medida que aumentamos a proporção de arestas modificadas no grafo. Ademais, ilustramos uma aplicação do CoGA em dados de

microarranjos de DNA de dois tipos de câncer no cérebro. Nossos resultados sugerem que a ferramenta apresentada (CoGA) pode ser útil na identificação de conjuntos de genes associados a uma doença, que, muitas vezes, não são detectados pelas análises clássicas baseadas na diferença da expressão média dos genes.

Objetivos

Os principais objetivos deste trabalho são resumidos nos três itens abaixo:

1. Estudar as características e avaliar o desempenho de medidas de dependência utilizadas para construir grafos de coexpressão de genes.

2. Desenvolver testes estatísticos para identificar diferenças entre grafos de coexpressão de genes.

3. Desenvolver um software livre que realiza testes entre grafos de coexpressão de conjuntos pré-definidos de genes funcionalmente relacionados (vias).

Organização do Trabalho

O trabalho está dividido em duas partes: uma sobre medidas de dependência estatística (Parte I) e outra sobre a análise diferencial de grafos de coexpressão (Parte II). Antes da Parte I, introduzimos, em Preliminares, a tecnologia de microarranjos de DNA e alguns conceitos básicos de biologia molecular para, finalmente, explicarmos o processo de regulação genética e definirmos um grafo de coexpressão.

Na Parte I, apresentamos o estudo comparativo entre as medidas de dependência estatística. No Capítulo 1, é apresentada a formulação matemática de cada medida estudada. Os experimentos de simulação e uma aplicação dos métodos em dados de microarranjos de DNA são descritos e discutidos no Capítulo 2.

A segunda parte do trabalho (Parte II) trata da análise diferencial entre grafos de coexpressão. Nesse tipo de análise são utilizadas medidas de propriedades estruturais de grafos, que são apresentadas no Capítulo 3. Os testes estatísticos propostos para comparar grafos de coexpressão são descritos no Capítulo 4. Nesse capítulo, também são descritos e apresentados os resultados de experimentos de simulação com os testes desenvolvidos e de uma aplicação dos métodos em dados de expressão genética de dois tipos de tumor cerebral. A ferramenta desenvolvida para a análise diferencial de grafos de coexpressão é apresentada no Capítulo 5.

Finalmente, no Capítulo 6, resumimos as principais contribuições deste trabalho e discutimos suas limitações e direções futuras.

Preliminares

Neste trabalho, estudamos redes de coexpressão, que são utilizadas para modelar o sistema de regulação genética. Para introduzir esse conceito, definiremos expressão genética e outros conceitos básicos de biologia molecular. Ademais, apresentaremos a tecnologia de microarranjos de DNA, que é utilizada para medir simultaneamente a expressão de milhares de genes e para a construção de redes de coexpressão.

Fundamentos de Biologia Molecular

Uma *célula* representa a menor unidade de vida. Nessa estrutura, estão contidas as características morfológicas e fisiológicas dos organismos vivos. Assim, as propriedades de um dado organismo dependem de suas células individuais, cuja continuidade ocorre por meio de seu material genético.

A célula é constituída de componentes moleculares, como as moléculas de DNA e as proteínas, que podem ser vistos como estruturas 3D de diversos formatos. O *DNA* é uma molécula na forma de dupla-hélice constituída de duas fitas antiparalelas de nucleotídeos, como ilustrado na Figura 1. Existem quatro tipos de nucleotídeos no DNA que correspondem às letras A (adenina), T (timina), C (citosina) e G (guanina). O DNA é usualmente representado por sequências desses quatro elementos, considerando apenas uma das fitas. A segunda fita pode ser sempre derivada a partir da primeira, pareando 'A's com 'T's e 'C's com 'G's e vice-versa. Um *gene* é um segmento contíguo de uma das fitas do DNA, composto por éxons e íntrons. Os *éxons* são os fragmentos responsáveis por codificar aminoácidos de uma proteína, enquanto os *íntrons* correspondem às partes não-codificadoras. As *proteínas* são sintetizadas a partir do DNA por meio de três operações ou transformações chamadas de transcrição, *splicing* e tradução. A fase de *splicing* ocorre em eucariotos, mas não acontece na maioria dos procariotos.

O DNA é capaz de se replicar. Os componentes celulares responsáveis por essa tarefa são chamados de *DNA-polimerase*. Genes são *transcritos* em pré-RNA por componentes celulares chamados de *RNA-polimerase*. Na molécula de pré-RNA, o nucleotídeo T (timina) é substituído por outro designado pela letra U (uracila). O pré-RNA pode ser representado por alternâncias de segmentos de sequência que correspondem aos éxons e íntrons após a transcrição. A operação de *splicing* consiste em concatenar os *éxons* e eliminar os *íntrons*

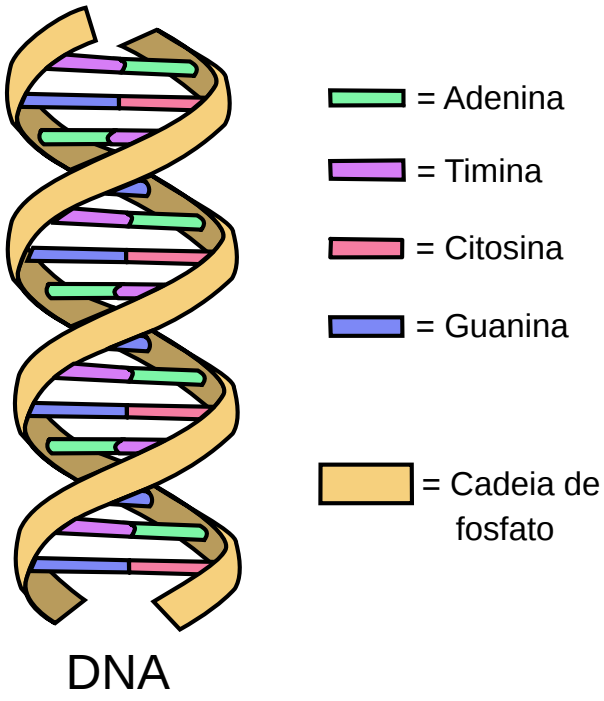


Figura 1: *Molécula de DNA em forma de dupla-hélice. Figura adaptada de http://commons.wikimedia.org/wiki/File:DNA_simple2.svg.*

para formar o que é conhecido como mRNA. Além do *splicing*, outros mecanismos de regulação pós-transcricional ocorrem em uma célula, como o processo de edição e a regulação da estabilidade do mRNA. A fase final da síntese de proteínas, conhecida como *tradução*, é realizada por moléculas complexas que são chamadas de *ribossomos* (um conjunto de RNA e proteínas). Nessa fase, o RNA é lido de três em três nucleotídeos. Cada tripleta de nucleotídeos (*códon*) corresponde a um aminoácido específico que vai se juntar à estrutura da proteína. Um esquema do processo de síntese da proteína pode ser visualizado na Figura 2. Finalmente, definimos a *expressão genética* como o processo pelo qual a informação de um gene é utilizada na síntese de um produto genético funcional, como as proteínas e as moléculas de RNA. Uma das tecnologias utilizadas para medir a expressão de diversos genes simultaneamente em amostras biológicas é o microarranjo de DNA, que veremos em mais detalhes na próxima seção.

Em um organismo, alguns genes são expressos continuamente (por exemplo, genes que sintetizam proteínas envolvidas em funções metabólicas básicas), enquanto outros são expressos apenas em situações específicas. Em eucariotos, as diferenças entre as diversas células que existem em um organismo são determinadas pela expressão de diferentes conjuntos de genes. Por exemplo, uma célula da pele e um neurônio são diferentes devido aos genes expressos nessas células. Além da especialização dos diferentes tipos de célula, outros fatores, como respostas de células a estímulos recebidos pelo ambiente, ocorrem por meio de mudanças da expressão genética.

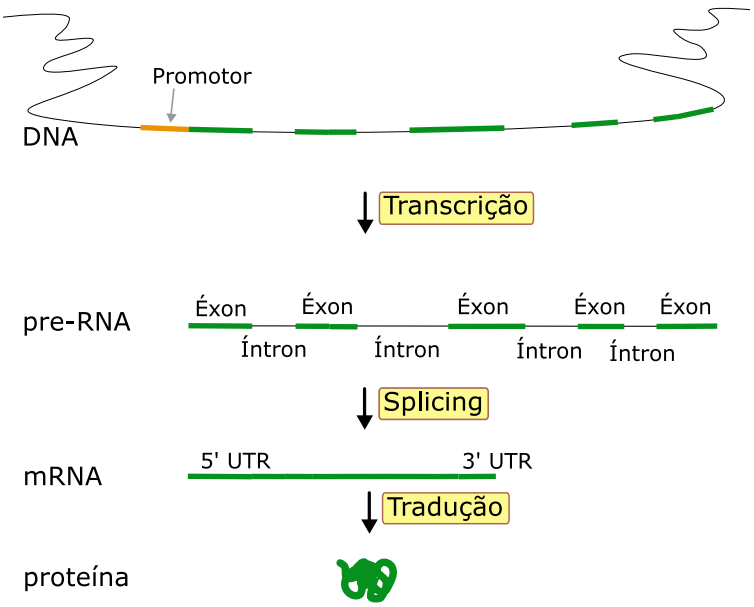


Figura 2: *Operações de transcrição,* splicing *e tradução. Na primeira transformação, o DNA é transcrito em pré-RNA, que é formado por alternâncias de éxons e íntrons. No* splicing*, os éxons são concatenados e os íntrons são eliminados, resultando em uma molécula de mRNA. Por fim, na tradução, o mRNA é utilizado para a síntese de proteína. Figura adaptada de http://commons.wikimedia.org/wiki/File:Gene2-plain-norsk.svg..*

Normalmente os genes são regulados para se tornarem expressos quando seus produtos funcionais são necessários. O momento e o local em que os genes são ativados e a quantidade de proteína e RNA produzidos dependem de um conjunto complexo de interações entre genes, moléculas de RNA e proteínas.

O mecanismo de regulação dos genes é formado por diversos sinais que atuam em momentos específicos. Para que a expressão de um gene seja ativada, a enzima RNA-polimerase se liga a um promotor de um gene. Os promotores são regiões do DNA próximas à extremidade 5' do gene, que atuam como locais de ligação para a enzima RNA-polimerase e fatores de transcrição (proteínas que auxiliam a interação entre a RNA-polimerase e o DNA). Como vimos anteriormente, após a transcrição, o mRNA é processado durante o *splicing* e a proteína é produzida no fim do processo de tradução.

As proteínas, por sua vez, formam a maior parte do contingente de moléculas efetoras das células, ou seja, são essas moléculas que atuam na estrutura, no metabolismo (como enzimas, principalmente) e no fluxo de informação celular (tanto dentro da célula, quanto entre células). Parte desse complexo sistema de interações moleculares pode ser modelado em uma rede de regulação genética, conforme abordaremos nas seções seguintes.

Microarranjo de DNA

Um *microarranjo de DNA* é um arranjo pré-definido de moléculas de DNA ligadas à uma lâmina, que é utilizado para medir os níveis de expressão de diversos genes simultaneamente.

O material fixado na lâmina pode consistir em fragmentos de DNA genômico, cDNAs (DNA complementar, que é sintetizado a partir de uma molécula de mRNA) ou oligonucleotídeos (fragmentos curtos de uma cadeia simples de DNA). Em cada microarranjo, há milhares de posições dispostas em forma de matriz, onde são fixados esses materiais. O fragmento em uma posição da lâmina é chamado de *sonda*.

Para que ácidos nucleicos (mRNA na forma de cDNA ou DNA genômico) provenientes de amostras biológicas sejam detectados e quantificados, as amostras são hibridizadas com o DNA fixado no arranjo (hibridização por complementariedade de bases). A detecção do material proveniente das amostras coletadas é possível pois essas são "marcadas" com fluorocromos. Por fim, é gerada uma imagem da hibridização, a partir da qual os transcritos de cada gene são quantificados. Em geral, a imagem é obtida por meio de leitores (*scanners*) a laser (para os fluorocromos).

Em microarranjos da Affymetrix, que é uma das plataformas de microarranjo existentes, cada gene é representado por um conjunto (usualmente de 11 a 20 pares) de sondas de sequências curtas de oligonucleotídeos . Cada par contém uma sonda com sequência nucleotídea igual ao gene (chamada de *perfect match* ou PM) e outra com uma alteração nucleotídea na décima terceira base (chamada de *mismatch*, MM). Usualmente, os valores de intensidade observados para cada gene são combinados em uma única medida para expressar a quantidade de transcritos de RNA. Um exemplo é a média das diferenças entre PM e MM para cada gene.

Em resumo, os passos (esquematizados na Figura 3) utilizados para quantificar a expressão genética a partir de um microarranjos da plataforma Affymetrix são:

1. O RNA é extraído de uma amostra biológica.
2. A partir de um processo de transcrição reversa, é sintetizado o DNA complementar (cDNA).
3. O cDNA produzido é transcrito *in vitro* para cRNA marcado com biotina.
4. O cRNA produzido é fragmentado e hibridizado.
5. A parte não hibridizada é removida do arranjo.
6. O arranjo passa por processos de lavagem e coloração.
7. É gerada uma imagem a partir do microarranjo de DNA produzido, que é utilizada para quantificar a expressão genética.

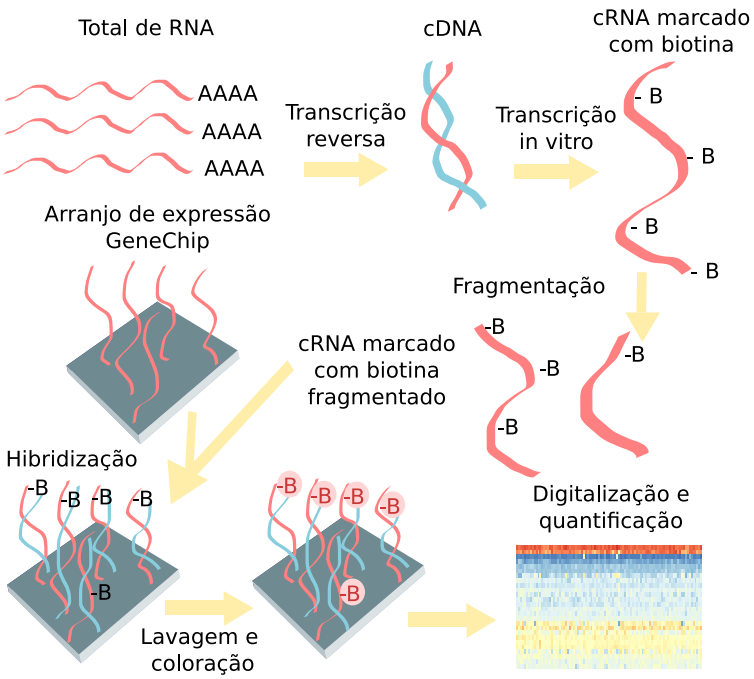


Figura 3: *Esquema ilustrativo da técnica de microarranjo de DNA da plataforma Affymetrix.*

Por fim, para analisar dados de expressão genética oriundos de microarranjos de DNA, é preciso considerar que existem diversas fontes de variação indesejáveis no processo de hibridização e até mesmo imprecisões do sistema de captura de imagens. Assim, algumas técnicas foram desenvolvidas para reduzir o efeito desses ruídos nos resultados das análises. As técnicas adotadas neste trabalho são detalhadas no Apêndice A.

Rede de regulação genética

Uma rede de regulação de genética é um modelo abstrato em que os vértices representam os genes e as arestas representam influências causais ou correlações (dependências) entre as atividades dos genes. Uma influência causal direta $A \rightarrow B$ significa que a atividade do gene B muda como consequência da atividade do gene A, sem a mediação da atividade de nenhum outro gene (de la Fuente, 2010). Um exemplo desse tipo de influência é a de um fator de transcrição (A) sobre um gene alvo (B). Já, quando a influência causal não é direta, a atividade do gene B muda como consequência da atividade do gene A com a mediação de um ou mais genes. Nas arestas que representam correlação ou alguma outra forma de dependência, as atividades de dois genes estão associadas, mas não há informação explícita de causalidade.

A estrutura de regulação entre as atividades genéticas pode mudar de acordo com o contexto, que pode ser o tipo celular, o ambiente, o genótipo, o estado da doença, entre outros.

Rede de coexpressão de genes

Inferir as influências causais de uma rede de regulação genética usualmente envolve experimentos de séries temporais, que permitem obter, por exemplo, a causalidade de Granger (Granger, 1969); experimentos que estudam o efeito de "desativação"/"ativação" de um gene; ou estudos que analisam simultaneamente dados de genótipo e expressão genética (de la Fuente, 2010). Os aspectos técnicos desses experimentos podem ser bastante complicados e custosos e, em geral, este tipo de dado não está disponível.

A maioria dos dados de expressão genética em doenças não são de séries temporais e são coletados sem nenhuma intervenção experimental (como "desativação"/"ativação" de genes), nem dados pareados de genótipo, assim como os dados considerados neste trabalho. Aqui utilizamos dados de expressão genética oriundos de microarranjos de DNA de diferentes indivíduos. Essas características, usualmente, dificultam a inferência das influências causais da rede de regulação genética. Dessa forma, as associações entre as atividades genéticas são muitas vezes inferidas sem causalidade, considerando apenas as correlações entre os níveis de expressão genética.

Em uma rede de coexpressão de genes (ou grafo de coexpressão), pares de genes são conectados por uma aresta sem direção se suas atividades (níveis de expressão) se comportam similarmente em uma série de medições de expressão genética. Usualmente essa similaridade é quantificada por alguma medida de correlação (de la Fuente, 2010).

Parte I

Medidas de dependência estatística

Capítulo 1

Definições

Medidas de dependência estatística quantificam a "força" com a qual duas variáveis aleatórias estão relacionadas. Elas desempenham um papel central na modelagem de redes reguladoras de genes, quantificando a "força" de associação entre os níveis de expressão genética. Uma das medidas de dependência mais utilizadas em Bioinformática é a correlação de Pearson.

Apesar de ser um dos conceitos mais tradicionais em Biologia Molecular moderna, a correlação de Pearson é frequentemente mal compreendida. Parte da confusão se deve ao uso da palavra "correlação" para se referir a qualquer tipo de dependência. Em Estatística, a correlação de Pearson mede apenas uma forma particular de dependência, a associação linear. Outros métodos foram desenvolvidos para identificar formas diferentes de associação, como relações não-lineares e não-monotônicas.

Neste capítulo, apresentaremos as ideias matemáticas por trás de diversos métodos para quantificar e identificar dependência estatística. São eles, o coeficiente de correlação de Pearson (Pearson, 1920), o coeficiente de correlação de Spearman (Spearman, 1904), o coeficiente de correlação de Kendall (Kendall, 1938), a correlação de distância (Székely *et al.* , 2007), o método de Heller-Heller-Gorfine, a medida D de Hoeffding (Hoeffding, 1948), (Heller *et al.*, 2013), a informação mútua (Shannon, 1948) e o coeficiente de informação máxima (Reshef *et al.*, 2011).

1.1 Notação

Neste capítulo, utilizaremos $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$ para representar observações conjuntas de duas varáveis aleatórias contínuas X e Y. No contexto de redes de coexpressão, cada variável aleatória corresponde ao nível de expressão de um gene.

1.2 Dependência estatística

Quando duas varáveis aleatórias são dependentes, o conhecimento sobre uma variável traz informações sobre a outra. Caso contrário, isto é, se as variáveis são independentes, não existe relação entre seus valores. Neste caso, a distribuição de uma variável se mantém a mesma para qualquer valor em que fixarmos a outra variável. Para ilustrar a ausência e a presença de dependência estatística, mostramos os gráficos de dispersão da Figura 1.1. No primeiro gráfico (Figura 1.1a), enquanto os níveis de uma variável (eixo y) são sempre os mesmos, os níveis da segunda variável (eixo x) mudam. No segundo gráfico (Figura 1.1b), os níveis de ambas as variáveis mudam sem que possamos visualizar uma tendência na disposição dos pontos. Já, na figura 1.1c, vemos que os pontos tendem para cima, se aproximando de uma reta com coeficiente angular diferente de zero, o que indica uma dependência linear entre as variáveis.

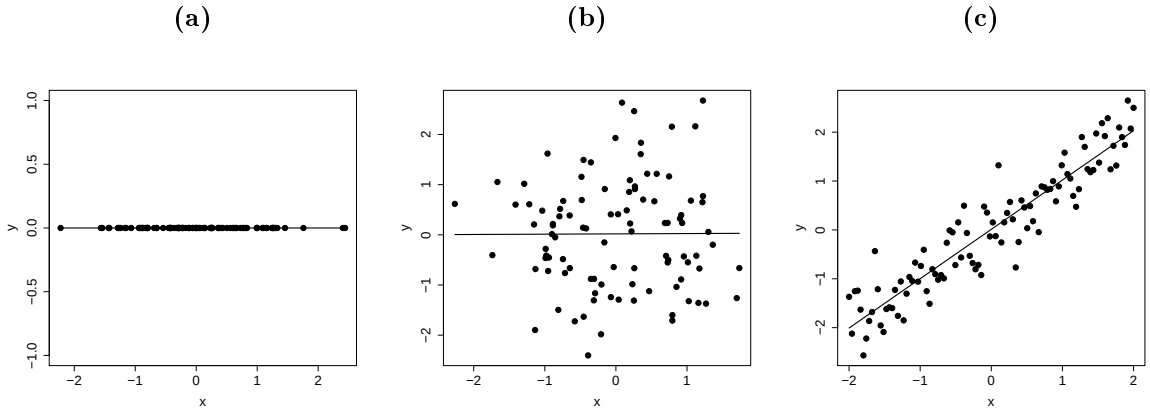


Figura 1.1: *Gráficos de dipersão entre duas variáveis aleatórias. Em (a), enquanto o nível de uma variável é fixo, o da outra muda (independência estatística). Em (b), os níveis de ambas as variáveis mudam independentemente (independência estatística). Em (c), os valores observados seguem uma tendência para cima, se aproximando de uma reta com coeficiente angular diferente de zero (dependência linear).*

Formalmente, duas variáveis aleatórias X e Y, com funções de densidade de probabilidade $f_X(x)$ e $f_Y(y)$, respectivamente, são *independentes* se e, somente se, $f_{XY}(x, y) = f_X(x)f_Y(y)$, onde $f_{XY}(x, y)$ é a função de densidade de probabilidade conjunta de X e Y. Quando $f_{XY}(x, y) \neq f_X(x)f_Y(y)$, dizemos que X e Y são *dependentes*.

1.3 Teste de independência

Os métodos descritos neste capítulo (i) detectam e (ii) quantificam a dependência estatística entre duas variáveis aleatórias. Para realizar (i), é feito um *teste de independência*, descrito pelas seguintes hipóteses nula e alternativa:

H_0 : X e Y são "independentes"

H_1 : X e Y são "dependentes"

Veremos nas seções seguintes que a formulação das hipóteses acima é apresentada de diferentes formas. Em todas as formulações, se a hipótese nula é verdadeira, as variáveis aleatórias são independentes conforme definição apresentada na Seção 1.2. Contudo, em métodos que medem formas mais específicas de dependência, como os testes de Pearson, de Spearman e de Kendall, a hipótese nula é formulada de maneira que ela pode ser válida mesmo quando as variáveis não são independentes. Já, nos testes da correlação de distância, de Heller-Heller-Gorfine, da medida D de Hoeffding, da informação mútua e do coeficiente de informação máxima, a hipótese nula é verdadeira se, e somente se, as variáveis X e Y são independentes.

Nos testes de Pearson, de Spearman, de Kendall e de Hoeffding, a distribuição da estatística do teste sob H_0 é aproximada para distribuições com funções de densidade de probabilidade conhecidas. Já, os testes da correlação de distância, de Heller-Heller-Gorfine, da informação mútua e do coeficiente de informação máxima obtêm a distribuição da estatística do teste sob H_0 empiricamente a partir de permutações aleatórias dos dados, conforme descrevemos a seguir.

Teste de independência entre X e Y usando permutações aleatórias dos dados

1. Calcule a estatística do teste com os dados originais $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$.
2. Construa um conjunto de dados $(x_1, y_1^*), (x_2, y_2^*) \dots (x_n, y_n^*)$ fixando x_i e permutando y_i.
3. Calcule a estatística do teste para este novo conjunto de dados $(x_1, y_1^*), (x_2, y_2^*) \dots (x_n, y_n^*)$.
4. Repita os passos 2 e 3 até alcançar o número desejado de permutações.
5. O p-valor do teste é a proporção de estatísticas obtidas maiores ou iguais do que a estatística observada nos dados originais.

A hipótese nula é rejeitada caso o p-valor seja menor do que o nível de significância (α) desejado. Esse tipo de teste em que são feitas permutações aleatórias dos dados é chamado de *teste de permutação* e será discutido em maiores detalhes na segunda parte do trabalho, na Seção 4.3.2.

Uma característica desejável de um teste de hipóteses é que seu poder estatístico (probabilidade de rejeitar a hipótese nula quando ela é falsa) cresce à medida que aumentamos o tamanho amostral. Quando um teste satisfaz esta característica, diremos que ele é *consistente*. Em particular, diremos que um teste de independência é *consistente contra todas as alternativas de dependência*.

1.4 "Força" de associação entre duas variáveis

Nas seções anteriores, discutimos a ausência ou presença de dependência estatística. Mas, se existir associação entre duas variáveis aleatórias, ela pode ser forte ou fraca. Para ilustrar o conceito de "força" de associação, apresentamos dois exemplos dados por Casella e Berger (2001), na página 169:

1. Considere um experimento em que a variável X mede o peso de uma amostra de água e a variável Y mede o volume da mesma amostra. Claramente, existe uma relação forte entre as duas variáveis. Se observarmos os pares $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$ de um número grande de amostras de água e colocarmos os dados em um gráfico de dispersão, esperamos que os pontos fiquem sobre uma reta. Em um experimento real, que é sujeito a erros de medida e impurezas na água, os pontos não ficam exatamente sobre uma reta, mas, com técnicas laboratoriais adequadas, podem ficar bem próximos de uma.

2. Agora, considere um experimento em que X mede o peso de uma pessoa e Y mede a altura da mesma pessoa. Esperamos que X e Y estejam associadas, mas não tão fortemente quanto no exemplo anterior, já que existem diversos fatores associados ao peso de uma pessoa, além da altura. Se coletarmos dados de diversos indivíduos, não esperamos que os pares $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$ fiquem sobre uma reta, mas que haja alguma tendência dos pontos para cima.

Embora exista associação entre as variáveis em ambos os exemplos, o grau de dependência das variáveis é claramente diferente. *Medidas de dependência estatística* quantificam essas diferenças de "força" de associação entre variáveis aleatórias. Neste capítulo, apresentamos diferentes métodos para quantificar diversas formas de dependência estatística. Os métodos foram agrupados em dependência linear (coeficiente de correlação de Pearson), dependência monotônica (coeficiente de correlação de Spearman, e coeficiente de correlação de Kendall) e dependência monotônica e não-monotônica (correlação de distância, medida de Heller-Heller-Gorfine, medida D de Hoeffding, informação mútua e coeficiente de informação máxima).

1.5 Medidas de dependência linear

Nesta seção descreveremos a correlação de Pearson, que mede o quanto uma associação pode ser descrita como uma função linear (isto é, a equação de uma reta com coeficiente angular não nulo). Para ilustrar esse tipo de relação, construímos a Figura 1.2 mostrando um gráfico de dispersão entre duas variáveis, onde os pontos se aproximam de uma reta.

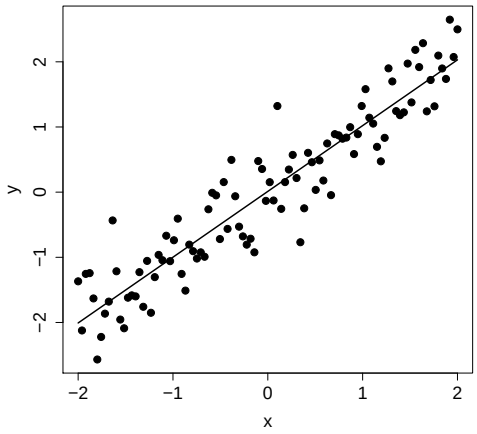


Figura 1.2: *Gráfico de dispersão entre duas variáveis com associação linear. Os pontos correspondem aos pares de observação* (x_i, y_i) *e a linha à reta da qual os pontos se aproximam.*

1.5.1 Coeficiente de correlação de Pearson

O coeficiente de correlação de Pearson é uma medida da dependência linear entre duas variáveis aleatórias definida como a razão entre a covariância e o produto dos desvios padrão.

Sejam $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$. 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 valor do coeficiente está entre -1 (relação linear decrescente perfeita) e 1 (relação linear crescente perfeita). Um coeficiente de valor zero indica que não há dependência linear. O coeficiente não é robusto à presença de *outliers*. Por exemplo, na Figura 1.3a podemos ver uma forte associação linear, contudo, após a inserção de *outliers* (Figura 1.3b), o coeficiente de correlação de Pearson cai bruscamente de 0,99 para 0,01.

A estatística do teste de independência é

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

Se X e Y seguem distribuição normal conjunta, então, sob H_0, t segue uma distribuição t de Student com $(n-2)$ graus de liberdade.

1.6 Medidas de dependência monotônica

Existe uma *relação monotônica crescente* entre duas variáveis aleatórias se cada incremento de uma variável corresponde a um incremento da segunda. Quando cada incremento da primeira variável corresponde a um decremento da segunda, dizemos que há uma *relação*

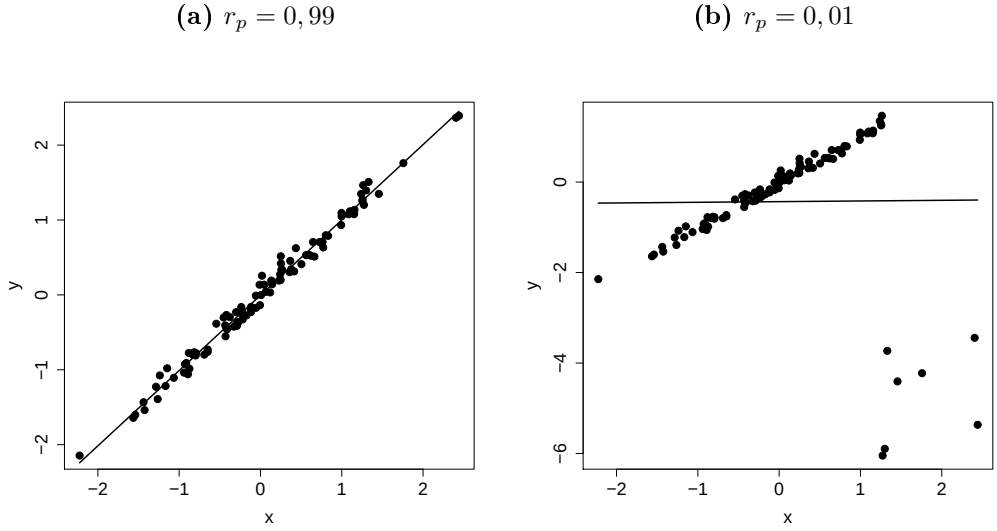


Figura 1.3: *Gráficos de dispersão de pares de observação de duas variáveis aleatórias com dependência linear antes (a) e após (b) a introdução de* outliers.

monotônica descrescente. A relação linear, por exemplo, é um tipo de relação monotônica. Se o coeficiente angular da equação da reta é positivo, então a equação descreve uma relação monotônica crescente. Já, se o coeficiente angular for negativo, então teremos uma relação monotônica decrescente. Para ilustrar relações monotônicas que não são lineares, consideramos as relações $y = x^3$ (crescente) e $y = -2^x$ (descrescente), exibidas nas figuras 1.4a e 1.4b, respectivamente.

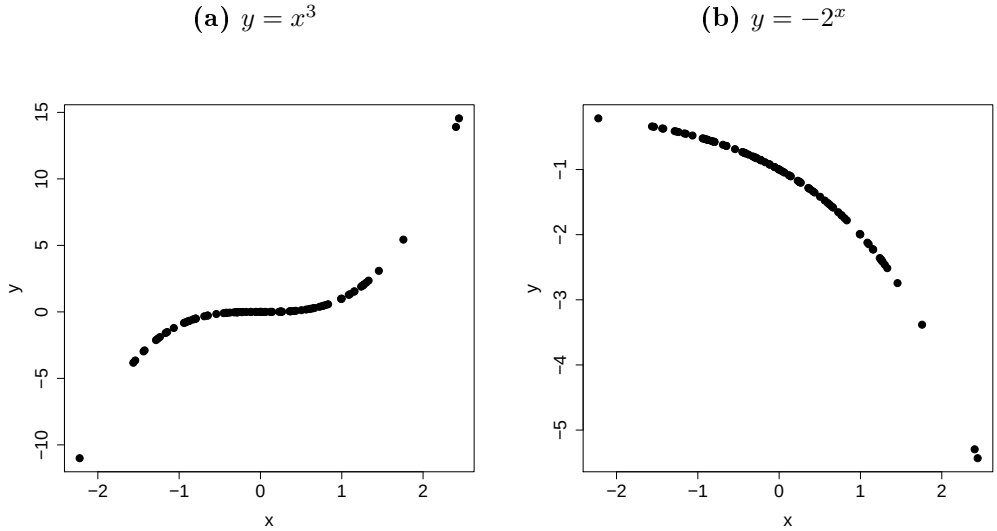


Figura 1.4: *Gráficos de dispersão de pares de observação de duas variáveis aleatórias com dependência monotônica. Em (a), a relação é descrita por $y = x^3$ (relação monotônica crescente). Em (b), a relação é descrita por $y = -2^x$ (relação monotônica decrescente).*

Medidas de dependência monotônica quantificam a "força" de uma relação monotônica (crescente ou decrescente) entre duas variáveis. Nesta seção, descreveremos algumas das medidas de dependência monotônica mais utilizadas em Estatística, como os coeficientes de

correlação de Spearman e de Kendall.

1.6.1 Coeficiente de correlação de Spearman

Chamamos a posição de um elemento em um vetor ordenado (em ordem crescente) de *posto*. O *coeficiente de correlação de Spearman* (Spearman, 1904), denotado por r_s, é obtido aplicando-se o coeficiente de correlação de Pearson nos dados convertidos em postos. O valor do coeficiente está entre -1 (relação monotônica decrescente perfeita) e 1 (relação monotônica crescente perfeita). Um coeficiente de valor zero indica que as variáveis são monotonicamente independentes.

Esta medida é muitas vezes utilizada no lugar da medida de Pearson quando há *outliers* nos dados. Quando os dados são convertidos em postos, os *outliers* não provocam grandes mudanças no comportamento do coeficiente de Spearman. Assim, dizemos que a correlação de Spearman é robusta à presença de *outliers*. No exemplo apresentado na Seção 1.5.1, o coeficiente de correlação de Pearson de duas variáveis com forte dependência linear cai de 0,99 para 0,01 após a introdução de *outliers*. Já, a correlação de Spearman muda de 0,99 para 0,6 nesse mesmo exemplo (Figura 1.4).

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 t de Student com $n-2$ graus de liberdade.

1.6.2 Coeficiente de correlação de Kendall

O coeficiente de correlação de Kendall (Kendall, 1938) é uma medida alternativa à correlação de Spearman, isto é, ele também detecta relações monotônicas e é robusto à presença de *outliers*.

Dois pares de observação quaisquer (x_i, y_i) e (x_j, y_j) são *concordantes* se $x_i > x_j$ e $y_i > y_j$ ou $x_i < x_j$ e $y_i < y_j$; e *discordantes* se $x_i > x_j$ e $y_i < y_j$ ou $x_i < x_j$ e $y_i > y_j$. Sejam C o número de pares concordantes, D o número de pares discordantes e N o número total de pares (se não há valores repetidos nas amostras, $N = n(n-1)/2$). O *coeficiente de correlação Kendall* é definido como

$$\tau(X, Y) = \frac{C-D}{N}.$$

Se dois pares de elementos são sorteamos aleatoriamente, então τ pode ser interpretado como a diferença entre a probabilidade de esses objetos estarem na mesma ordem e a probabilidade de eles estarem em ordem diferente. A estatística τ, sob H_0, para n suficientemente grande, segue distribuição normal com média 0 e variância

$$\sigma^2 = \frac{2(2n+5)}{9n(n-1)}.$$

1.7 Medidas de dependência monotônica e não-monotônica

Quando duas variáveis estão associadas, mas o incremento de uma variável corresponde ora a um incremento ora a um decremento da segunda variável, dizemos que existe uma relação não-monotônica entre as duas variáveis. Por exemplo, a relação descrita pela função quadrática ilustrada na Figura 1.5a não é monotônica, pois a função descresce no intervalo $[-2, 0]$, mas cresce em $[0, 2]$. Mas nem toda forma de dependência é descrita por uma função. A *dependência não-funcional* ocorre quando duas variáveis estão estatisticamente associadas, mas a relação entre elas não pode descrita por uma função matemática. Um exemplo desse tipo de dependência, é uma relação em forma de circunferência (Figura 1.5b).

(a) $y = x^3$ **(b)** *Circunferência*

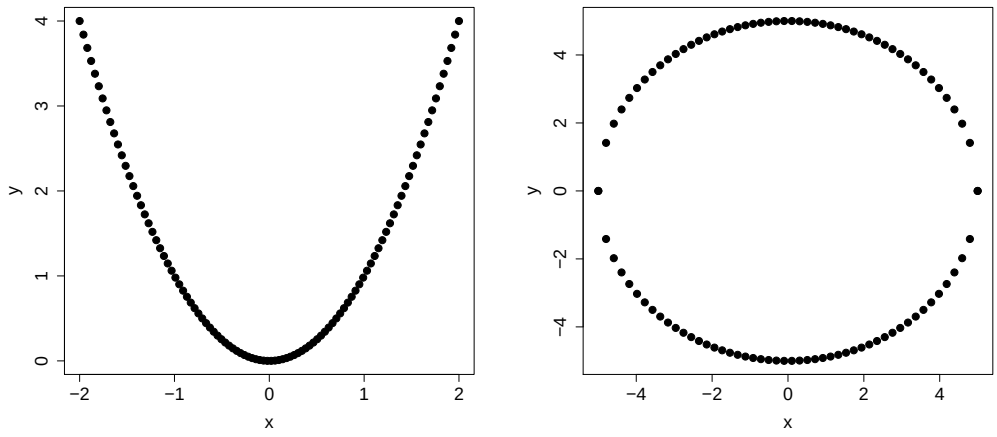


Figura 1.5: *Gráfico de dispersão entre duas variáveis com (a) associação quadrática e (b) associação descrita por uma circunferência (dependência não-funcional).*

Medidas de dependência monotônica e não-monotônica quantificam a dependência estatística entre duas variáveis, sendo esta monotônica ou não. Exemplos deste tipo de medida incluem a correlação de distância, a medida D de Hoeffding, a medida de Heller-Heller-Gorfine, a informação mútua e o coeficiente de informação máxima, que serão apresentados a seguir.

1.7.1 Correlação de distância (Dcor)

A correlação de distância (Székely *et al.*, 2007) mede a dependência estatística entre dois vetores aleatórios (vetores cujos elementos são variáveis aleatórias) de quaisquer dimensões. Neste trabalho, todo vetor aleatório tem dimensão 1, que corresponde a uma variável aleatória. O nome da medida vem do fato de ela ser baseada no conceito de distâncias de energia (distância estatística entre distribuições de probabilidade). A correlação de distância entre as variáveis aleatórias (ou vetores aleatórios de dimensão 1) X e Y é a covariância de distância entre X e Y dividida pelo produto dos desvios padrão de distância de X e Y.

Em outras palavras, definimos $a_{k,l} = ||x_k - x_l||$ e $b_{k,l} = ||y_k - y_l||$, para $k, l = 1, 2, ..., n$, onde $||.||$ denota a norma euclidiana. Chamaremos a matriz

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{pmatrix}$$

de matriz de distância de X e a matriz

$$\begin{pmatrix} b_{1,1} & b_{1,2} & \dots & b_{1,n} \\ b_{2,1} & b_{2,2} & \dots & b_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n,1} & b_{n,2} & \dots & b_{n,n} \end{pmatrix}$$

de matriz de distância de Y. Utilizaremos $\bar{a}_{k.}$, $\bar{a}_{.l}$ e $\bar{a}$ para denotar os valores médios, respectivamente, da k-ésima linha, da l-ésima coluna e da matriz de distância completa de X. Analogamente, os símbolos $\bar{b}_{k.}$, $\bar{b}_{.l}$ e $\bar{b}$ representam os valores médios da matriz de distância de Y. Sejam $A_{j,k} = a_{j,k} - \bar{a}_{j.} - \bar{a}_{.k} + \bar{a}$ e $B_{j,k} = b_{j,k} - \bar{b}_{j.} - \bar{b}_{.k} + \bar{b}$ as distâncias centralizadas de X e Y, respectivamente. A covariância de distância entre X e Y, a variância de distância de X e a variância de distância de Y são definidas, respectivamente, como $dCov(X,Y) = \sqrt{\frac{1}{n^2}\sum_{k,l} A_{k,l}B_{k,l}}$, $dVar(X) = dCov(X,X)$ e $dVar(Y) = dCov(Y,Y)$. Usando as definições anteriores, a *correlação de distância* entre X e Y é definida como

$$dCor(X,Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X)dVar(Y)}}.$$

O valor de $dCor$ está entre 0 (variáveis independentes) e 1 (variáveis fortemente dependentes).

Usualmente, para testarmos se $dCor = 0$, é feito um teste de permutação, conforme descrito na Seção 1.3. O teste da correlação de distância é consistente contra todas as alternativas de dependência.

1.7.2 Medida de Heller, Heller e Gorfine (HHG)

Heller *et al.* (2013) propuseram um teste de independência consistente contra todas as alternativas de dependência entre dois vetores aleatórios de quaisquer dimensões. Contudo, neste trabalho, consideraremos apenas vetores de dimensão 1. Assim como a correlação de distância, o teste de independência de Heller, Heller e Gorfine (HHG) entre duas variáveis aleatórias (ou vetores aleatórios de dimensão 1) X e Y é baseado nas distâncias entre os valores de X e nas distâncias entre os valores de Y. Conforme definimos na Seção 1.7.1, $a_{k,l}$ é a distância Euclidiana entre x_k e x_l e $b_{k,l}$ é a distância euclidiana entre y_k e y_l, para

$k, l = 1, 2, ..., n$.

Note que, se X e Y são dependentes, então existe um ponto (x_0, y_0) no espaço de amostras de (X, Y) e raios R_{x_0} e R_{y_0} ao redor de x_0 e y_0, respectivamente, tais que a distribuição conjunta de X e Y é diferente do produto das distribuições marginais no produto Cartesiano das bolas ao redor de (x_0, y_0). Como não conhecemos (x_0, y_0), R_{x_0} e R_{y_0} previamente, para cada $1 \leq i, j \leq n$, com $i \neq j$, definimos $(x_0, y_0) = (x_i, y_i)$, $R_{x_0} = a_{i,j}$ e $R_{y_0} = b_{i,j}$.

Sejam $I\{.\}$, a função indicadora e $d(.,.)$ a distância Euclidiana. Dados i e j, com $i \neq j$, consideramos as seguintes variáveis:

$$I_X(i,j) = I\{d(x_i, X) \leq a_{i,j}\},$$
$$I_Y(i,j) = I\{d(y_i, Y) \leq b_{i,j}\}.$$

Podemos interpretar $I_X(i,j)$ como uma variável que indica se um ponto está na bola ao

Tabela 1.1: *Tabela de contingência de $I_X(i,j)$ e $I_Y(i,j)$ utilizada no teste de Heller-Heller-Gorfine.*

		$I_Y(i,j)$		Total
		1	0	
$I_X(i,j)$	1	$c_{1,1}(i,j)$	$c_{1,2}(i,j)$	$\sum c_{1.}(i,j)$
	0	$c_{2,1}(i,j)$	$c_{2,2}(i,j)$	$\sum c_{2.}(i,j)$
Total		$\sum c_{.1}(i,j)$	$\sum c_{.2}(i,j)$	$n-2$

redor de x_i com raio $a_{i,j}$ (neste contexto, temos um segmento de reta onde x_i está no centro, já que nossas variáveis aleatórias são escalares). Interpretamos $I_Y(i,j)$ de forma análoga. Construindo a tabela de contingência de $I_X(i,j)$ e $I_Y(i,j)$ (Tabela 1.1) para $(n-2)$ observações isto é, todas as n observações exceto (x_i, y_i) e (x_j, y_j), temos:

$$c_{1,1}(i,j) = \sum_{k=1, k\neq i, k\neq j}^{n} I\{a_{i,k} \leq a_{i,j}\} I\{b_{i,k} \leq b_{i,j}\},$$
$$c_{1,2}(i,j) = \sum_{k=1, k\neq i, k\neq j}^{n} I\{a_{i,k} \leq a_{i,j}\} I\{b_{i,k} > b_{i,j}\},$$
$$c_{2,1}(i,j) = \sum_{k=1, k\neq i, k\neq j}^{n} I\{a_{i,k} > a_{i,j}\} I\{b_{i,k} \leq b_{i,j}\},$$
$$c_{2,2}(i,j) = \sum_{k=1, k\neq i, k\neq j}^{n} I\{a_{i,k} > a_{i,j}\} I\{b_{i,k} > b_{i,j}\},$$
$$\sum c_{m.}(i,j) = c_{m,1}(i,j) + c_{m,2}(i,j),$$
$$\sum c_{.m}(i,j) = c_{1,m}(i,j) + c_{2,m}(i,j),$$

onde $m \in \{1, 2\}$. A estatística do teste de Chi-quadrado de Pearson para a tabela de con-

tingência 2 por 2 de $I_X(i,j)$ e $I_Y(i,j)$ (Tabela 1.1) é

$$S(i,j) = \frac{(n-2)\{c_{1,2}(i,j)c_{2,1}(i,j) - c_{1,1}(i,j)c_{2,2}(i,j)\}^2}{\sum c_{1.}(i,j) \sum c_{2.}(i,j) \sum c_{.1}(i,j) \sum c_{.2}(i,j)}.$$

Somando a estatística de Chi-quadrado para todo $1 \leq i,j \leq n$, com $i \neq j$, temos

$$T = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} S(i,j).$$

Quando $T = 0$, X e Y são independentes. O p-valor do teste de independência pode ser obtido por um teste de permutação (como descrito na Seção 1.3), utilizando T como estatística.

1.7.3 Medida D de Hoeffding

A medida D de Hoeffding (Hoeffding, 1948) é uma medida de dependência estatística entre duas variáveis aleatórias X e Y baseada na distância entre a distribuição conjunta sob a hipótese nula (produto da distribuições marginais de X e Y) e a distribuição conjunta empírica.

Sejam R_i o posto de x_i, S_i o posto de y_i e Q_i o número de pares (x_j, y_j) tais que $x_j < x_i$ e $y_j < y_i$. Chamaremos $\sum_{i=1}^{n} Q_i(Q_i - 1)$ de D_1, $\sum_{i=1}^{n} (R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2)$ de D_2 e $\sum_{i=1}^{n} (R_i - 2)(S_i - 2)Q_i$ de D_3. A *medida D de Hoeffding* é dada por

$$D = \frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)}.$$

Seja ρ_n o menor valor satisfazendo a desigualdade

$$P\{D > \rho_n | F_{XY}(x,y) = F_X(x)F_Y(y)\},$$

onde P é a distribuição de probabilidade de D. Este valor satisfaz

$$30\rho_n \leq \sqrt{\frac{2(n^2 + 5n - 32)}{9n(n-1)(n-3)(n-4)\alpha}},$$

onde α é o nível de significância do teste. Rejeitamos H_0 se, e somente se, $D > \rho_n$.

1.7.4 Informação mútua (IM)

A informação mútua (Shannon, 1948) entre duas variáveis aleatórias X e Y mede o quanto uma variável aleatória informa sobre a outra. Sejam $f_{XY}(x,y)$, a função de densidade de probabilidade conjunta de X e Y e $f_X(x)$ e $f_Y(y)$ as funções de densidade de probabilidade marginais de X e Y, respectivamente. A *informação mútua* entre duas variáveis contínuas

X e Y é dada por

$$IM(X,Y) = \int_Y \int_X f_{XY}(x,y) \log \frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} dxdy.$$

A informação mútua vale zero quando as duas variáveis são independentes. O teste estatístico é consistente contra todas as alternativas de dependência e é realizado com um teste de permutação, conforme descrito na Seção 1.3.

1.7.5 Coeficiente de Informação Máxima (CIM)

O Coeficiente de Informação Máxima (CIM) (Reshef *et al.*, 2011) é uma medida de dependência que se baseia na ideia de que, se duas variáveis são dependentes, pode-se desenhar, no gráfico de dispersão de X e Y, uma grade que encapsula a relação entre as duas variáveis. Definimos uma grade a-por-b como um par que contém uma partição dos dados bivariados nos valores de X formada por a partes (x-partição) e uma partição nos valores de Y composta por b partes (y-partição). Chamaremos a intersecção entre uma parte da x-partição e uma parte da y-partição de célula.

Podemos estimar a função de distribuição de probabilidade em um ponto (x_i, y_i) pela proporção dos pontos que estão na mesma célula de (x_i, y_i). A partir da distribuição estimada, obtemos $IM^*(a,b)$, que é a maior informação mútua (conforme definimos na Seção 1.7.4) entre todas as grades a-por-b. O valor normalizado de $IM^*(a,b)$ é dado por $M_{a,b} = \frac{IM^*(a,b)}{\log \min\{a,b\}}$.

Dado $1 < B(n) \leq n^{0,6}$, o *Coeficiente de Informação Máxima* é definido como

$$CIM(X,Y) = \max_{ab<B(n)} M_{a,b}.$$

O valor de CIM está entre 0 e 1. Quanto mais próximo o valor de CIM está de 1, mais forte é a associação entre X e Y. O teste estatístico é realizado por um teste de permutação conforme descrito na Seção 1.3.

Capítulo 2

Estudo comparativo

A fim de elucidar as características de cada medida de dependência considerada neste trabalho e facilitar a escolha da mais adequada em diferentes cenários (em particular, na construção de redes de coexpressão de genes), realizamos simulações computacionais e aplicamos as medidas em dados reais de expressão genética. Na seções seguintes, explicaremos o método utlizado para avaliar o desempenho de cada medida (curva ROC), os cenários simulados e os resultados obtidos.

2.1 Curva ROC

A curva ROC (*Receiver Operating Characteristics*) é um método utilizado para avaliar o desempenho de quaisquer classificadores binários. Neste trabalho, em particular, ela será utilizada para avaliar o desempenho de testes de hipóteses.

Em Estatística, o erro do tipo I ocorre ao se rejeitar a hipótese nula quando ela é verdadeira. Comete-se esse tipo de erro quando o resultado do teste de hipóteses apresenta significância estatística, sendo que na verdade ele ocorreu por acaso. Assim, o erro do tipo I também é conhecido como *falso positivo*. Já, se a hipótese nula não é rejeitada quando na verdade ela é falsa, comete-se um erro do tipo II, conhecido como *falso negativo*. Um *verdadeiro positivo* ocorre ao se rejeitar a hipótese nula quando ela é falsa. Se a hipótese nula não é rejeitada e ela é de fato verdadeira, obtém-se um *veradeiro negativo*.

Definimos a *especificidade* como (número de verdadeiros negativos)/(número de verdadeiros negativos + número de falsos positivos) e a *sensibilidade* como (número de verdadeiros positivos)/(número de verdadeiros positivos + número de falsos negativos). Usualmente, a curva ROC de uma classificador binário que rotula em positivo e negativo é construída em um gráfico bidimensional, onde o eixo x corresponde a um menos a especificidade, isto é, a taxa de falsos positivos, e o eixo y à sensibilidade (ou taxa de verdadeiros positivos). Cada ponto da curva está associado a um limiar para a classificação em positivo e negativo (supondo que o resultado devolvido pelo classificador é um valor utilizado para discriminar as duas classes).

Neste trabalho, realizamos testes estatísticos entre duas variáveis aleatórias, com a hipótese nula de que as variáveis são independentes e a hipótese alternativa de que elas são dependentes. Para avaliar o poder estatístico e o controle da taxa de falsos positivos dos testes de independência estudados, adaptamos a curva ROC de forma que no eixo x temos o nível de significância (α) dos testes e no eixo y temos a proporção de rejeições da hipótese nula de que as variáveis são independentes, ou seja, a proporção de associações detectadas pela medida de dependência. Note que o nível de significância α é o limiar do p-valor para rejeição da hipótese nula (isto é, a hipótese nula é rejeitada se o p-valor do teste for menor do que α) e a proporção de rejeições da hipótese nula é o poder empírico do teste estatístico. Assim, a área sob a curva ROC construída neste trabalho está entre 0 e 1 e quantifica o poder estatístico dos testes realizados, que é a probabilidade de rejeitar a hipótese nula quando ela é falsa.

Em outras palavras, quanto mais perto a área sob a curva ROC está de 1, maior o poder estatístico do teste. Uma área próxima de 0,50 é equivalente a decisões aleatórias. Quando os dados são gerados sob a hipótese nula, esperamos que a curva ROC fique na diagonal, resultando em uma área de 0,5. Esta propriedade é consequência da própria definição do α, que é a probabilidade de ocorrer um falso positivo, ou seja, a probabilidade do p-valor do teste ser menor que α quando H_0 é verdadeira. Assim, $\Pr[\text{p-valor} < \alpha | H_0] = \alpha$. Isto é, sob H_0, o p-valor segue distribuição uniforme. Temos que a taxa esperada de rejeições da hipótese nula (eixo y) é igual ao α (eixo x). Na Figura 2.1, a linha tracejada na diagonal ilustra a curva ROC esperada sob a hipótese nula, a linha vermelha é a curva ROC de um teste com alto poder estatístico (método 1) e a linha verde é a curva de um teste com baixo poder estatístico (método 2).

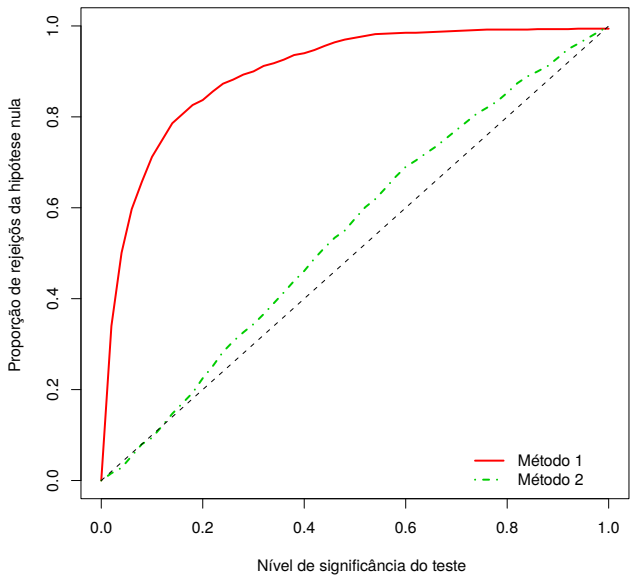


Figura 2.1: *Curvas ROC construídas a partir de dois métodos. A linha tracejada na diagonal ilustra a curva ROC esperada sob a hipótese nula. A linha vermelha é a curva ROC de um teste com alto poder estatístico (método 1) e a linha verde é a curva de um teste com pouco poder estatístico (método 2).*

2.2 Simulações

Ilustramos o poder e as limitações de cada método sistematicamente a partir de simulações de Monte Carlo, isto é, experimentos em que a geração aleatória de dados é simulada pelo computador. Com os cenários simulados, mostramos os efeitos do número de observações, do tipo de dependência (linear, não-linear, monotônica, não-monotônica e não-funcional) e da presença de *outliers* sobre o desempenho dos métodos.

Para avaliar o efeito do tamanho amostral e do tipo de dependência no desempenho dos métodos, simulamos dados de variáveis independentes (Figura 2.2a); dados com dependência linear (Figura 2.2b); dados com dependência monotônica não-linear, como a relação exponencial (Figura 2.2c); dados com dependência não-monotônica, como as relações quadrática (Figura 2.2d) e senoide (Figura 2.2e); dados com relações não funcionais, isto é, que não podem ser escritas como funções, como a circunferência (Figura 2.2f), a relação em forma de "X" (Figura 2.2g) e o quadrado (Figura 2.2h); e dados com dependência linear local, isto é, dados em que apenas parte dos dados estão correlacionados, enquanto o restante não (Figura 2.2i).

Para avaliar o efeito da introdução de *outliers*, consideramos dois cenários (Figura 2.3), um com dados independentes (Figura 2.3a) e outro com dados linearmente dependentes (Figura 2.3b). Em cada experimento foram inseridos *outliers* em 7% das amostras.

Cada cenário descrito na Figura 2.2 e na Figura 2.3 foi simulado 1000 vezes no ambiente de desenvolvimento `R` (http://cran.r-project.org/) para diferentes números de observações, variando de 10 a 140. Para comparar o desempenho de cada medida, construímos curvas ROC, onde o eixo x corresponde ao nível de significância do teste (limiar do p-valor para rejeitar o teste) e o eixo y à proporção de vezes que H_0 é rejeitada (poder empírico do teste). Em seguida, calculamos os valores das áreas debaixo das curvas para, assim, quantificar o poder dos testes considerando diferentes níveis de significância. A tabela 2.1 exibe os valores das áreas debaixo das curvas ROC construídas.

Tabela 2.1: *Área da região abaixo da curva ROC gerada para cada medida, com amostras de tamanho n*

Tipo de associação	n	Pearson	Spearman	Kendall	Dcor	HHG	Hoeffding	IM	CIM
Independente	10	0,5	0,48	0,45	0,5	0,5	0,5	0,34	0,34
	30	0,51	0,49	0,5	0,51	0,49	0,57	0,48	0,50
	50	0,51	0,5	0,51	0,5	0,51	0,57	0,5	0,49
Independente com outliers	10	0,87	0,56	0,52	0,89	0,57	0,48	0,03	0,33
	30	0,76	0,51	0,55	0,98	0,85	0,56	0,66	0,50
	50	0,71	0,54	0,53	1	0,95	0,6	0,89	0,52
Linear	10	0,8	0,75	0,72	0,76	0,61	0,69	0,4	0,52
	30	0,91	0,89	0,89	0,89	0,76	0,87	0,62	0,70
	50	0,96	0,94	0,94	0,94	0,83	0,94	0,69	0,77
Linear com outliers	10	0,86	0,72	0,75	0,95	0,86	0,78	0,06	0,64
	30	0,75	0,97	0,98	1	1	0,99	0,69	0,94
	50	0,72	1	1	1	1	1	0,94	0,98

Exponencial	10	0,88	0,94	0,93	0,99	0,99	0,97	0	0,73
	30	0,96	1	1	1	1	1	0,86	1,00
	50	0,99	1	1	1	1	1	0,9	1,00
Quadrática	10	0,16	0,16	0,14	0,71	0,97	0,9	0,54	0,52
	30	0,2	0,17	0,21	0,99	1	1	1	1,00
	50	0,21	0,18	0,23	1	1	1	1	1,00
Quadrática com outliers	10	0,14	0,28	0,23	0,16	0,78	0,7	0,08	0,36
	30	0,06	0,31	0,31	0,41	0,99	0,96	0,14	0,97
	50	0,05	0,31	0,32	0,64	1	0,99	0,94	0,99
Senoide	10	0,28	0,32	0,24	0,29	0,34	0,51	0,33	0,19
	30	0,35	0,38	0,35	0,89	0,98	0,96	0,93	0,99
	50	0,4	0,42	0,42	0,98	1	0,99	1	1,00
Circunferência	10	0,09	0,24	0,2	0,1	0,71	0,64	0,51	0,10
	30	0,09	0,15	0,18	0,18	0,96	0,88	0,74	0,71
	50	0,09	0,15	0,18	0,38	0,99	0,95	0,96	0,94
X	10	0,09	0,14	0,11	0,03	0,66	0	0,42	0,02
	30	0,11	0,11	0,11	0,46	1	0,42	0,96	0,57
	50	0,12	0,11	0,11	0,77	1	0,85	1	0,97
Quadrado	40	0,25	0,27	0,26	0,18	0,9	0,27	0,33	0,38
	140	0,25	0,26	0,25	0,45	1	0,58	0,7	0,45
Correlação Local	100	0,29	0,43	0,41	1	1	0,99	1	1,00

Observamos que, quando supomos independência entre X e Y (Figura 2.2a), isto é, sob a hipótese nula, todos os métodos apresentam áreas próximas de 0,50. Dessa forma, os métodos controlaram a taxa de falsos positivos conforme o esperado. É importante observar, contudo, que a área sob a curva ROC da medida D de Hoeffding foi um pouco acima de 0,50 devido ao fato de a medida superestimar o número de falsos positivos para $\alpha > 0,40$, conforme discutido por Fujita *et al.* (2009). Como usualmente apenas p-valores menores do que 0,05 são considerados como estatisticamente significativos, essa propriedade da medida D de Hoeffding não afeta as aplicações práticas do teste estatístico.

Sob a hipótese alternativa (dados dependentes), a maioria dos métodos tiveram desempenhos consistentes com o tamanho amostral. Como esperado, quanto maior o número de observações, maior a área sob a curva ROC (poder estatístico do teste). Contudo, as correlações de Pearson, Spearman e Kendall não tiveram esse comportamento para relações não-monotônicas (quadrática e senoide) e não-funcionais (circunferência, "X" e quadrado). Assim, independentemente do número de observações, esses métodos não identificaram esses tipos de dependência.

No caso de dependência linear e monotônica não-linear (exponencial), as medidas de Pearson, Spearman, Kendall, Dcor, HHG e Hoeffding tiveram áreas parecidas (próximas de 1). Já, as medidas IM e CIM apresentaram menor poder. Apesar de, em teoria, a correlação de Pearson identificar apenas relações lineares, ela apresentou um bom desempenho para identificar relações monotônicas não-lineares, como a relação exponencial. Esse comportamento pode ser explicado pelo fato de relações monotônicas não-lineares serem usualmente bem ajustadas por funções lineares.

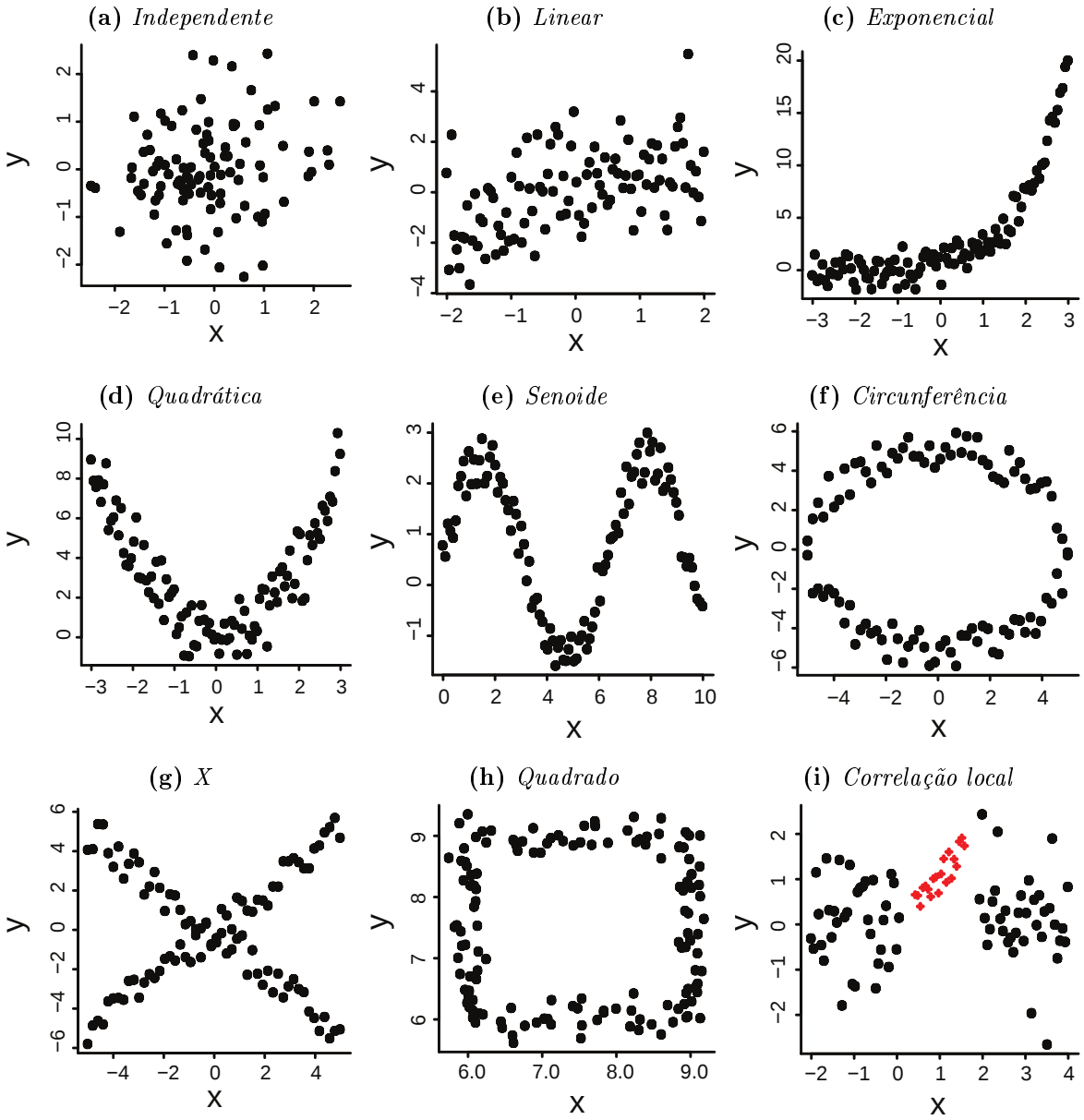


Figura 2.2: *Gráficos de dispersão dos diferentes cenários simulados. Em (a), (b) e (c) temos pares de observação de variáveis independentes, variáveis com dependência linear e variáveis com dependência monotônica não-linear, respectivamente. As figuras (d) e (e) ilustram associações não-monotônicas funcionais. Já as figuras (f), (g) e (h) ilustram associações não-funcionais. Na figura (i), temos uma dependência local, com os pontos em vermelho representando os pares de observação correlacionados e os pontos em preto os pares independentes. Figura adaptada de Santos et al. (2014)*

.

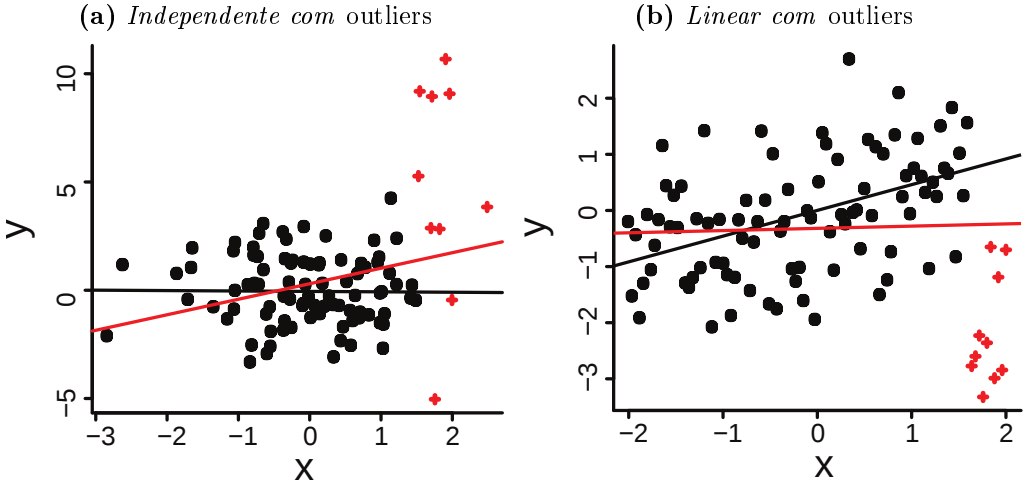


Figura 2.3: *Gráficos de dispersão de cenários com a presença de* outliers *(pontos em vermelho).*

Quando geramos dados com associações não-monotônicas (quadrática e senoide), apenas a correlação de distância, as medidas de HHG e de Hoeffding, a informação mútua e o CIM apresentaram áreas próximas de 1. Nas relações não funcionais (circunferência, "X" e quadrado) a medida de HHG teve melhor desempenho, seguida pelas medidas D de Hoeffding, IM, MIC e correlação de distância. Quando há correlação local (dependência linear em apenas parte dos dados), todos os métodos, exceto as medidas de Pearson, Spearman e Kendall, tiveram bom desempenho.

O objetivo do experimento em dados independentes com *outliers* era verificar a robustez dos métodos para controlar a taxa de falsos positivos. Na Figura 2.3a, a presença de *outliers* gera uma tendência dos pontos para cima, que desejamos ignorar. A correlação de Pearson, a correlação de distância, o teste de HHG e a informação mútua não controlaram a taxa de falsos positivos na presença de *outliers*, como podemos observar na Tabela 2.1 (curvas ROC com áreas menores do que 0,50). Já as correlações de Spearman e Kendall, a medida D de Hoeffding e o CIM foram robustos à introdução dos dados discrepantes. Essas medidas, com exceção da medida D de Hoeffding, apresentaram áreas próximas de 0,50. Como vimos anteriormente, a medida D de Hoeffding, apresenta área ligeiramente superior a 0,50 sob H_0 independentemente da presença de *outliers*, pois ela superestima o número de falsos positivos para $\alpha > 0,40$ (Fujita *et al.*, 2009).

Nos dados com dependência linear, todos os métodos detectaram dependência (área maior do que 0,5) mesmo na presença de *outliers*. Contudo, a área da correlação de Pearson reduziu consideravelmente após a introdução das observações discrepantes, o que é esperado. Como podemos ver na Figura 2.3b, a presença de *outliers* cria um viés que diminui a inclinação da reta que aproxima os pontos no gráfico de dispersão.

2.3 Aplicação em dados de microarranjo de DNA

A fim de ilustrar a utilização dos métodos estudados para identificar interações entre os produtos genéticos, aplicamos as medidas de dependência em dados de microarranjos de DNA de 168 amostras de câncer de pulmão no estágio I. Os dados (Okayama *et al.* , 2012; Yamauchi *et al.*, 2012) estão disponíveis no *Gene Expression Omnibus* - GEO (http://www.ncbi.nlm.nih.gov/geo/) com número de acesso GSE31210.

A correção do ruído do fundo e a normalização dos microarranjos foram realizadas com o método RMA (*Robust Multichip Average*) (Irizarry *et al.*, 2003b) do pacote `affy` do `Bioconductor` (http://www.bioconductor.org/), que é descrito no Apêndice A. Para agrupar as sondas de forma que cada gene seja representado por um único conjunto de sondas, utilizando informação atualizada do genoma, foi adotado o CDF (*Chip Description File*) do `Brainarray` (http://brainarray.mbni.med.umich.edu/brainarray/default.asp).

Selecionamos o gene WNT5A como referência para esta aplicação, pois ele tem grande associação com o câncer de pulmão e vias bem conhecidas na literatura (Mazieres *et al.*, 2005). Realizamos testes de independência entre os níveis de expressão do WNT5A e genes pertencentes a dois grupos:

1. 62 conjuntos de sondas de controle presentes nos microarranjos cujos níveis de expressão são independentes dos demais genes dos microarranjos.

2. 81 genes que fazem parte da via do WNT5A (que já são conhecidos como associados ao WNT5A). A lista completa dos 81 genes pode ser verificada no Apêndice C.

Os conjuntos de sondas de controle foram utilizados para verificar o controle da taxa de falsos positivos dos métodos. Já os genes pertencentes à via do WNT5A foram utilizados para avaliar o poder estatístico. Assim como fizemos nas simulações descritas na Seção 2.2, calculamos as áreas debaixo das curvas ROC para comparar o desempenho das medidas. Para os testes realizados com os conjuntos de sondas de controle, a curva ROC mostra, no eixo y, a proporção de falsos positivos obtidos nos testes de independência entre o WNT5A e os 62 conjuntos de sondas. Quando consideramos os genes da via do WNT5A, a curva ROC mostra a proporção de relações identificadas entre o WNT5A e os 81 genes associados ao WNT5A. A tabela 2.2 mostra as áreas debaixo das curvas ROC construídas com amostras (de microarranjos) de tamanhos 12, 25, 50, 100 e 168 (número total de microarranjos).

Observamos que, quando o tamanho da amostra de microarranjos é grande, todos os métodos apresentam área próxima ou inferior a 0,50 nos testes realizados com os conjuntos de sondas de controle, indicando que a taxa de falsos positivos foi controlada, apesar das hipóteses de alguns testes estatísticos eventualmente não serem satisfeitas em dados biológicos (muitas vezes não é possível verificar a validade das hipóteses feitas). Já nos testes realizados com os genes da via do WNT5A, as áreas sob as curvas ROC são maiores do que 0,50, sugerindo que as medidas de dependência descritas neste trabalho podem inferir

Tabela 2.2: *Área sob a curva ROC calculada para cada medida com os dados de microarranjo de DNA*

	Número de microarranjos	Pearson	Spearman	Kendall	Dcor	HHG	Hoeffding	IM	CIM
Conjuntos de sondas de controle	12	0,67	0,66	0,64	0,58	0,42	0,59	0,3	0,40
	25	0,58	0,66	0,68	0,65	0,52	0,65	0,55	0,60
	50	0,34	0,33	0,34	0,4	0,45	0,52	0,48	0,50
	100	0,45	0,41	0,41	0,39	0,37	0,47	0,46	0,38
	168	0,34	0,33	0,34	0,4	0,45	0,52	0,48	0,50
Genes da via do WNT5A	12	0,53	0,54	0,5	0,5	0,48	0,52	0,35	0,39
	25	0,6	0,59	0,57	0,63	0,56	0,62	0,59	0,59
	50	0,66	0,66	0,66	0,68	0,63	0,72	0,59	0,56
	100	0,63	0,61	0,61	0,66	0,61	0,67	0,56	0,57
	168	0,66	0,66	0,66	0,68	0,63	0,72	0,59	0,56

pelo menos parte da rede reguladora de genes. Em concordância com os resultados obtidos nas simulações, o poder do IM e do CIM foi menor do que o poder dos outros métodos e o aumento do poder dos métodos foi proporcional ao aumento do número de observações.

Para verificar a sobreposição de associações identificadas pelos métodos, contamos o número de associações em comum identificadas entre o gene WNT5A e os 81 genes participantes da sua via. A Tabela 2.3 mostra o número de associações identificadas com diferentes níveis de significância ($\alpha = 0,01, 0,05, 0,1$).

Tabela 2.3: *Número de associações identificadas em comum para diferentes níveis de sinificância (α) pelos testes de dependência entre os níveis de expressão do gene WNT5A e outros 81 genes que participam da via Wnt.*

	α	Pearson	Spearman	Kendall	Dcor	HHG	Hoeffding	IM	CIM
Pearson	0,01	12	10	10	8	4	10	3	2
	0,05	19	16	16	18	8	14	4	6
	0,1	28	23	23	25	14	23	10	9
Spearman	0,01		15	15	11	5	14	4	2
	0,05		24	24	17	10	21	7	6
	0,1		30	30	25	14	27	11	11
Kendall	0,01			15	11	5	14	4	2
	0,05			24	17	10	21	7	6
	0,1			30	25	14	27	11	11
Dcor	0,01				11	5	10	4	1
	0,05				22	11	16	5	6
	0,1				32	17	28	10	12
HHG	0,01					7	5	4	1
	0,05					15	11	5	3
	0,1					20	18	6	7
Hoeffding	0,01						14	4	2
	0,05						22	5	6
	0,1						31	10	12
IM	0,01							4	1
	0,05							8	1
	0,1							15	5
CIM	0,01								3
	0,05								8
	0,1								14

Como, no estudo de redes reguladoras de genes, pesquisadores usualmente buscam identificar relações monotônicas ou lineares, nós verificamos, no nosso conjunto de dados, o número de genes identificados pela correlação de distância, HHG, medida de Hoeffding, IM e CIM que não foram identificados pelas medidas que detectam apenas relações lineares ou monotônicas (isto é, as medidas de Pearson, Spearman e Kendall). A Tabela 2.4 exibe os valores observados. Notamos que o número de genes não identificados pelas medidas de Pearson, Spearman ou Kendall é relativamente pequeno, o que sugere que a maior parte das associações testadas no nosso conjunto de genes são lineares.

Tabela 2.4: *Número de associações identificadas pelo Dcor, medida de Hoeffding, HHG, IM e CIM que não foram identificadas pelas medidas de Pearson, Spearman ou Kendall. O valor entre parênteses indica o número total de associações identificadas pela medida.*

Nível de significância	Dcor	HHG	Hoeffding	IM	CIM
0,01	0 (11)	2 (7)	0 (14)	0 (4)	1 (3)
0,05	2 (22)	4 (15)	1 (22)	1 (8)	2 (8)
0,1	3 (32)	4 (20)	3 (31)	3 (15)	3 (14)

2.4 Conclusões

A escolha de uma medida de dependência depende essencialmente do tipo dos dados sendo analisados, do tipo de associação que se deseja identificar e do número de observações. Para identificar formas mais gerais de dependência, a correlação de distância, a medida D de Hoeffding, o teste de HHG, a IM, e o CIM são apropriados. Se o conjunto de dados for pequeno (em nossas simulações, se tiverem menos do que 30 observações), o teste de HHG é recomendado por apresentar maior poder estatístico.

Para detectar relações lineares ou monotônicas, como acontece muitas vezes em biologia, as medidas de Spearman e Kendall podem ser vantajosas em relação à medida de Pearson, pois são robustas à presença de *outliers*, além de também apresentarem elevado poder estatístico tanto para identificar relações lineares quanto monotônicas não-lineares. Contudo, para conjuntos de dados pequenos com associação linear, a correlação de Pearson pode apresentar maior poder estatístico.

Dentre os métodos estudados, os únicos que podem ser diretamente aplicados em vetores aleatórios com mais de uma dimensão são a correlação de distância e o teste de HHG. Vale ressaltar que apenas os testes da correlação de distância, da medida D de Hoeffding, do HHG e da informação mútua foram demonstrados matematicamente como consistentes contra todas as alternativas de dependência (na teoria, eles podem, assintoticamente, detectar qualquer tipo de dependência estatística).

Parte II

Análise diferencial de grafos de coexpressão

Capítulo 3

Medidas de redes complexas

Redes complexas são grafos que apresentam características estruturais não-triviais, isto é, características que não são totalmente aleatórias nem totalmente regulares. Esse tipo de grafo é usualmente utilizado para modelar sistemas reais. Um exemplo de rede complexa é o grafo de coexpressão de genes, que apresenta distribuição dos graus dos vértices com cauda pesada e grupos de vértices altamente conectados entre si (Carter *et al.*, 2004).

As características estruturais de uma rede complexa influenciam fortemente a dinâmica do processo que ocorre dentro dela (Costa *et al.*, 2007). Assim, medidas de características estruturais são amplamente utilizadas para comparar sistemas complexos que apresentam funcionamento normal (por exemplo, indivíduos saudáveis) com aqueles que apresentam anormalidades (por exemplo, indivíduos portadores de uma doença). Neste capítulo, apresentaremos medidas de redes complexas bem conhecidas na literatura, como medidas de centralidade (*betweenness*, proximidade, grau e autovetor), medidas de segregação funcional (coeficiente de *clustering*) e medidas de resistência (distribuição do grau). Além disso, apresentaremos medidas baseadas na caracterização espectral de um grafo recentemente propostas por Takahashi *et al.* (2012), como a entropia espectral para medir a aleatoriedade de um grafo e a divergência de Jensen-Shannon, para medir a distância entre dois grafos.

3.1 Notação

Os grafos que apresentaremos neste trabalho são não-dirigidos, podendo ter peso nas arestas ou não. Para facilitar o tratamento dos dois casos (com e sem peso), nos referiremos a grafos sem peso como grafos em que cada aresta tem peso um. Um *grafo não-dirigido* é um par ordenado de conjuntos $G = (V, E)$, onde V é um conjunto de vértices e E é um conjunto de arestas que conectam os vértices de V. Cada elemento $e \in E$ é um par não-ordenado $e = \{v_1, v_2\}$, onde $v_1, v_2 \in V$ e $v_1 \neq v_2$. A cada aresta $e \in E$, associamos um valor real positivo w_e, que, no contexto deste trabalho, usualmente satisfaz $w_e \leq 1$. O número de vértices de G, isto é, o tamanho do conjunto V, será denotado por n_V. A matriz de

adjacência de G é uma matriz A com n_V linhas e n_V colunas onde

$$A_{ij} = w_{ij}, \text{ se existe aresta que conecta } v_i \text{ e } v_j,$$
$$A_{ij} = 0, \text{ se não existe aresta conectando } v_i \text{ e } v_j.$$

3.2 Medidas de centralidade

Em redes de regulação genética, os vértices representam os genes e as arestas representam influências causais ou correlações (dependências) entre as atividades dos genes. Esse tipo de rede usualmente contém genes "importantes", conhecidos como *hubs*, que desempenham um papel central na dinâmica do processo biológico que acontece na rede. Medidas de centralidade quantificam a "importância" de cada gene no funcionamento do sistema.

O *grau* de um vértice, definido como o número de arestas conectadas ao vértice, é uma das medidas de centralidade mais utilizadas. Se as arestas tiverem peso, nós obtemos o grau de um vértice pela soma dos pesos das arestas incidentes ao vértice. Assim, podemos escrever o grau do vértice v_i de um grafo G com n_V vértices como

$$C_G(v_i) = \sum_{j=1, j\neq i}^{n_V} A_{ij}.$$

Em uma rede de regulação genética, um gene com grau alto interage funcionalmente com uma grande quantidade de genes na rede. Outras medidas de centralidade, como as centralidades de *betweenness* e de proximidade, se baseiam na ideia de que genes centrais fazem parte de muitos caminhos curtos na rede.

A centralidade de *betweenness* de um vértice v_i é o número de caminhos mais curtos entre v_i e todos os demais vértices da rede (Freeman, 1978). Seja $g(v_j, v_i, v_k)$ o número de caminhos mais curtos entre os vértices v_j e v_k que passam por v_i. Podemos expressar a centralidade *betweenness* de v_i como

$$C_B(v_i) = \sum_{j=1, j\neq i}^{n_V} \sum_{k=j+1, k\neq i}^{n_V} g(v_j, v_i, v_k).$$

Uma medida relacionada é a centralidade de *proximidade*, definida como o inverso do comprimento médio dos caminhos mais curtos entre um vértice e os demais vértices na rede (Freeman, 1978). Em outras palavras, se $d(v_i, v_j)$ denota o comprimento do caminho mais curto entre os vértices v_i e v_j, a centralidade de proximidade do vértice v_i é

$$C_P(v_i) = \sum_{j=1, j\neq i}^{n_V} d(v_i, v_j).$$

Embora existam generalizações dessas medidas para grafos com peso (Brandes, 2001;

Newman, 2001; Opsahl *et al.*, 2010), neste trabalho, consideramos apenas o caso de grafos sem peso. Nas generalizações das centralidades de *betweenness* e de proximidade, os pesos das arestas são convertidos em distâncias. Contudo, neste trabalho, os pesos representam níveis de associações estatísticas entre os vértices, o que dificulta a sua interpretação como distância.

A "importância" de um vértice também pode ser influenciada pela importância de seus vizinhos. Baseado nessa ideia, Bonacich (1972) propôs a centralidade de *autovetor* de um vértice, que é proporcional à soma das centralidade de seus vizinhos. Se as arestas têm peso, então as centralidades dos vizinhos são ponderadas pelos valores associados às arestas correspondentes (Rahmatallah *et al.*, 2014). Equivalentemente, se x_i e N_i denotam, respectivamente a centralidade de autovetor do vértice v_i e a vizinhança de v_i, temos que

$$x_i = \frac{1}{\lambda} \sum_{j \in N_i} w_{ij} x_j = \frac{1}{\lambda} \sum_{j=1}^{n_V} A_{ij} x_j.$$

Em notação vetorial, definimos $\mathbf{x} = (x_1, x_2, ..., x_{n_V})$ como um vetor que satisfaz a equação

$$A\mathbf{x} = \lambda \mathbf{x}. \tag{3.1}$$

Além disso, cada componente x_i deve satisfazer $x_i \geq 0$. Em geral, podem existir diversos autovalores λ para os quais existe uma solução de autovetor $\mathbf{x}$ que satisfaça (3.1). Contudo, para que não haja componentes negativos, apenas o maior autovalor resultará na centralidade desejada. Assim, a centralidade do vértice v_i é a i-ésima componente do autovetor associado ao maior autovalor da matriz de adjacência A.

3.3 Medidas de segregação funcional

Em uma rede biológica, processos com funções especializadas ocorrem usualmente dentro de um grupo densamente interconectado de vértices. Medidas de segregação são utilizadas para identificar esse tipo de grupo. Um exemplo de medida de segregação é o *coeficiente de clustering* (Watts e Strogatz, 1998) de um vértice v_i, definido como o número de pares de vizinhos de v_i conectados entre si dividido pelo número de arestas que poderiam existir entre eles. Se k_i é o número de vizinhos de v_i, então o número total de arestas que poderiam existir entre os vizinhos de v_i é $k_i(k_i - 1)/2$. Quando as arestas têm pesos, o coeficiente de *clustering* 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 *et al.*, 2004). Equivalentemente, podemos expressar o coeficiente de *clustering* de v_i por

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

onde N_i denota a vizinhança de v_i. Assim, quando $w_e \leq 1$, essa medida de segregação está entre 0 e 1, sendo que zero indica que não há conexões entre os vizinhos do vértice e 1 indica que todos os vértices da vizinhança estão conectados entre si. Em um grafo de coexpressão, um coeficiente de *clustering* alto, isto é, próximo de 1, indica a presença de um grupo de genes densamente conectados que têm funções semelhantes ou participam de um mesmo processo biológico (Roy *et al.*, 2014).

3.4 Medidas de resistência

Medidas de resistência caracterizam a habilidade de um sistema de se adaptar e sobreviver diante de ameaças. Em uma rede aleatória, a remoção de um número crítico de vértices usualmente particiona o grafo em grupos de vértices que não se comunicam, prejudicando a funcionalidade do sistema. Contudo, redes complexas, como as redes biológicas, tendem a ser muito resistentes a falhas de seus componentes (Barabási e Oltvai, 2004).

Albert *et al.* (2000) mostraram que a estrutura de uma rede pode influenciar fortemente a sua tolerância a falhas. Uma característica topológica que está relacionada à propriedade de resistência é *distribuição do grau* do vértice (Barabási e Albert, 1999). Por exemplo, redes com distribuição dos graus com cauda pesada tendem a ser robustas a falhas aleatórias, mas vulneráveis a falhas em vértices centrais.

3.5 Medidas baseadas na caracterização espectral de um grafo

Seja $G = (V, E)$ um grafo não-dirigido com n_V vertices. O *espectro* de G é o conjunto de autovalores de sua matriz de adjacência (A). Note que, como G é não-dirigido, então A é simétrica e todos seus autovalores são reais. O conjunto dos autovalores de A descreve muitas propriedades estruturais do grafo, como o diâmetro e o número de passeios e cliques (Van Mieghem, 2010).

Um *modelo de grafo* é um algoritmo que gera grafos de acordo com uma lei de probabilidade. Dado um modelo que gera grafos com n_V vértices, representaremos por g a família de todos os grafos construídos pelo modelo. O espectro de um grafo com n_V vértices é formado por n_V autovalores reais $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n_V}$. A *densidade espectral* de g é a função de densidade de probabilidade dos autovalores dos grafos de g.

Para definirmos formalmente a densidade espectral, representaremos o delta de Dirac por δ. Trata-se de uma distribuição com as seguintes propriedades: (i) $\delta(x) = 0, x \in \mathbb{R}^*$; (ii) $\delta(0) = \infty$; e (iii) $\int_{-\infty}^{+\infty} \delta(x)dx = 1$. Além disso, utilizaremos os colchetes "$\langle\rangle$" para denotar a esperança com respeito à lei de probabilidade de g.

A *densidade espectral* (Takahashi *et al.*, 2012) da família de grafos g é definida como

$$f(\lambda) = \lim_{n_V \to \infty} \langle \frac{1}{n_V} \sum_{j=1}^{n_V} \delta(\lambda - \lambda_j)/\sqrt{n_V} \rangle.$$

Vimos que as medidas definidas nas seções anteriores descrevem características importantes de redes complexas, como a centralidade dos nós, a segregação funcional e a resistência a falhas. Para modelar redes complexas e definir classes de grafos que compartilham características estruturais que ocorrem em muitos sistemas reais, alguns modelos de grafos foram propostos. Neste trabalho, consideraremos os modelos bem conhecidos de Erdös e Rényi (1959), Barabási e Albert (1999) e Watts e Strogatz (1998) como exemplos de classes de grafos. A seguir ilustraremos a densidade espectral desses modelos e apresentaremos algumas medidas baseadas na distribuição do espectro, como a entropia espectral, e as divergências de Kullback-Leibler e Jensen-Shannon entre densidades espectrais.

3.5.1 Distribuição do espectro em modelos conhecidos

O modelo de Erdös-Rényi, também conhecido como modelo de grafo aleatório, é um dos mais simples em termos de construção. Erdös e Rényi (1959) geram um grafo com n_V vértices acrescentando uma aresta a um par de vértices com probabilidade p.

Já, no modelo de Barabási-Albert, a probabilidade de inserir uma aresta no grafo depende do grau do vértice. No procedimento descrito por esse modelo (Barabási e Albert, 1999), dado um número inicial pequeno (n_{V0}) de vértices, a cada passo, é adicionado um novo vértice com $m_1 (\leq n_{V0})$ arestas que conectam o novo vértice a m_1 vértices diferentes já presentes na rede. A probabilidade de um novo vértice ser conectado ao vértice v_i é proporcional ao grau de v_i (número de arestas com ponta em v_i) com ordem de proporcionalidade dada pelo expoente de escala p_s. Bollobás *et al.* (2001) demonstraram que, nos grafos gerados por este modelo, a distribuição dos graus é, assintoticamente, regida pela Lei da Potência, isto é, a proporção de vértices com grau d será aproximadamente igual a d elevado a um fator $-\gamma$.

No modelo de Watts-Strogatz, a construção do grafo envolve tanto inserções aleatórias de arestas quanto inserções determinadas por uma estrutura em forma de anel em que os vértices são conectados aos $K/2$ vértices mais próximos de cada lado da estrutura. O procedimento descrito pelo modelo (Watts e Strogatz, 1998) é (i) dados n_V vértices, construa uma estrutura de anel 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 seja a segunda mais próxima no sentido horário; (vi) repita o processo considerando a terceira mais próxima e assim por diante, até que todas as arestas tenham sido analisadas.

Grafos gerados pelo modelo de Barabási-Albert são caracterizados pela sua distribuição dos graus que segue a Lei da Potência. Já o modelo de Watts-Strogatz gera grafos com coeficientes de *clustring* altos e comprimentos de caminhos curtos. Apesar de serem propriedades essenciais dessas classes de grafos, essas características podem não ser efetivas para discriminar grafos gerados por diferentes modelos. Por exemplo, grafos gerados pelo modelo de Watts-Strogatz têm coeficiente de *clustering* alto como um grafo em forma anel com cada vértice conectado aos vértices mais próximos e comprimentos de caminhos curtos como o os grafos de Erdös-Rényi. Takahashi *et al.* (2012) propuseram, então, comparar grafos em termos de seus espectros. Essa caracterização descreve propriedades mais gerais de um grafo quando comparada com outras medidas comumente utilizadas, como o número de arestas, o coeficiente de *clustering* e o comprimento médio dos caminhos mais curtos .

Na Figura 3.1, mostramos o histograma dos autovalores das matrizes de adjacência geradas pelos modelos de Erdös-Rényi (Figura 3.1a), Barabási-Albert (Figura 3.1b) e Watts-Strogatz (Figura 3.1c), de acordo com simulação realizada por Takahashi *et al.* (2012). No caso dos grafos de Erdös-Rényi, quando o número de vértices é suficientemente grande, a densidade espectral pode ser aproximada para

$$f(\lambda) \sim \frac{\sqrt{4p(1-p)-\lambda^2}}{2\pi p(1-p)}, \tag{3.2}$$

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

Baseados na relação entre a distribuição do espectro do grafo e suas características estruturais, Takahashi *et al.* (2012) introduziram os conceitos de entropia para medir a quantidade de incerteza associada ao grafo, e da divergência de Jensen-Shannon entre densidades espectrais de grafos. Apresentaremos a definição formal desses conceitos a seguir.

3.5.2 Entropia

Em Teoria da Informação, a entropia de uma variável aleatória mede a quantidade de incerteza associada ao seu valor. Para ilustrar o conceito de entropia introduzido por Shannon (1948), consideremos duas variáveis discretas X e Y que podem assumir os valores 0 e 1.

Vamos supor que $P(X = 0) = 0,50, P(X = 1) = 0,50$, $P(Y = 0) = 0,95$ e $P(Y = 1) = 0,05$. Intuitivamente, a variável X está associada a uma quantidade de incerteza maior do que a variável Y. Enquanto é difícil prever o valor de X, o valor da variável Y será "quase certamente" 0. Se considerarmos uma variável discreta Z, com $P(Z = 0) = 0,75$ e $P(Z = 1) = 0,25$, a quantidade de incerteza esperada de Z é um valor intermediário entre a incerteza de X e Y. Uma medida que representa bem a noção intuitiva de quantidade de incerteza associada a uma variável discreta X que pode assumir n valores com probabilidade

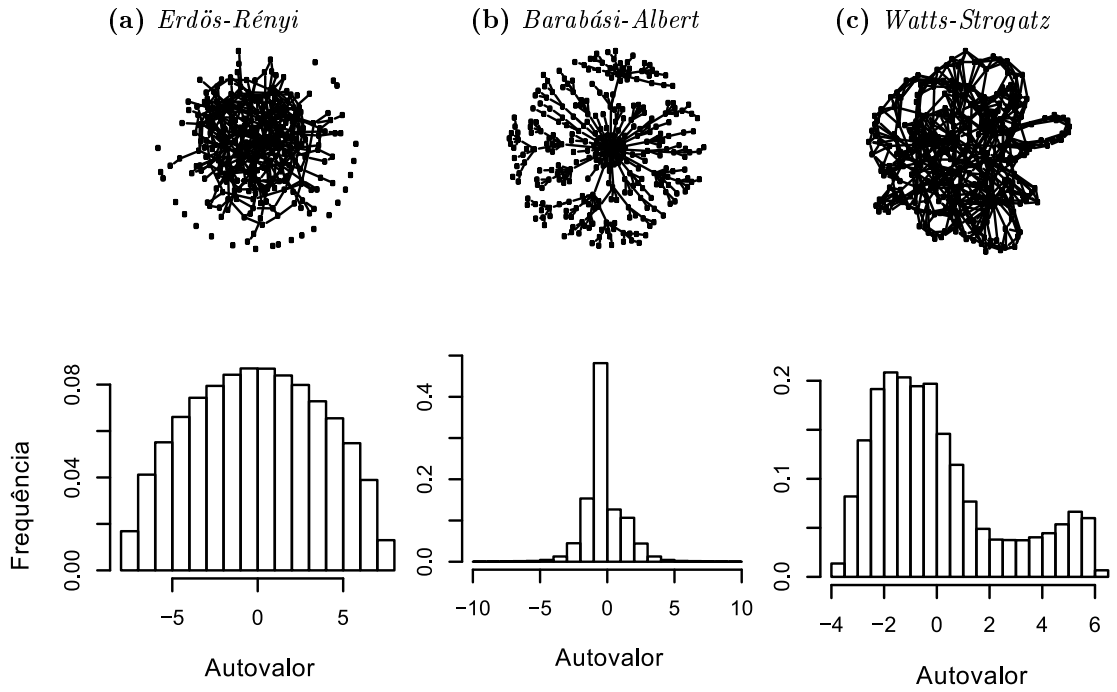


Figura 3.1: *Distribuição espectral de diferentes modelos de grafo. Na parte superior, são exibidos desenhos de grafos gerados pelos modelos de (a) Erdös-Rényi, (b) Barabási-Albert e (c) Watts-Strogatz. Na parte inferior, temos os histogramas dos autovalores dos grafos gerados pelos modelos de (a) Erdös-Rényi, (b) Barabási-Albert e (c) Watts-Strogatz. A figura foi adaptada de Takahashi et al. (2012).*

$p_1, p_2, \ldots, p_n$, é a entropia, definida como

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i, \tag{3.3}$$

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

Observe que H(X) vale zero se, e somente se, um dos números de $p_1, p_2, \ldots, p_n$ é um e todos os demais valem zero. Nesse caso, os valores de X podem ser preditos com absoluta certeza. Para todos os demais casos, a entropia será positiva. Intuitivamente, o valor de X terá maior incerteza se $p_1 = p_2 = \cdots = p_n = 1/n$. De fato, podemos verificar que a entropia será máxima nesse caso, isto é, quando $H(X)$ vale

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

Se f é uma função convexa contínua, então ela satisfaz a propriedade

$$f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \leq \frac{1}{n}\sum_{i=1}^{n} f(x_i),$$

onde $x_1, x_2, \ldots, x_n$ são números reais positivos (Khinchin, 1957). Como $f = x \log x$ é convexa contínua nos reais positivos, segue que

$$\begin{aligned}\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} &= n \left(\frac{1}{n} \log \frac{1}{n} \right) \\ &= n \left(\frac{1}{n} \sum_{i=1}^{n} p_i \right) \log \left(\frac{1}{n} \sum_{i=1}^{n} p_i \right) \\ &\leq n \left(\frac{1}{n} \sum_{i=1}^{n} p_i \log p_i \right) \\ &= \sum_{i=1}^{n} p_i \log p_i = -H(X).\end{aligned}$$

Temos que

$$H(X) \leq -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n},$$

isto é, a entropia é máxima quando $p_1 = p_2 = \cdots = p_n = 1/n$.

A entropia de variáveis contínuas, também conhecida como entropia diferencial, é definida similarmente. O somatório em (3.3) é substituído por uma integral no suporte da função de densidade de probabilidade. Contudo, diferentemente da entropia definida para distribuições discretas, a entropia diferencial pode assumir valores negativos.

Para um grafo, a entropia quantifica a aleatoriedade da sua estrutura. Formalmente, a entropia espectral de um grafo é definida como segue. Sejam 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

$$H(f) = -\int_{-\infty}^{+\infty} f(\lambda) \log f(\lambda) d\lambda, \tag{3.4}$$

onde $0 \log 0 = 0$ (Takahashi *et al.*, 2012).

Utilizando as fórmulas em (3.2) e (3.4), podemos aproximar a entropia do grafo de Erdös-Rényi para

$$H(f) \sim \frac{1}{2} \ln(4\pi^2 p(1-p)) - \frac{1}{2}, \tag{3.5}$$

onde p é a probabilidade de um par de vértices ser conectado por uma aresta (Takahashi *et al.*, 2012). Podemos ver pela Equação (3.5) que o valor máximo da entropia do grafo de Erdös-Rényi é alcançado quando $p = 0,50$, o que é consistente com a ideia intuitiva de que é mais difícil prever se dois vértices do grafo serão conectados quando $p = 0,50$, e, portanto, a quantidade de incerteza é alta. Intuitivamente, quando $p = 0$ e $p = 1$, a quantidade de incerteza associada ao modelo será menor, pois os grafos serão sempre vazios (quando $p = 0$) e completos (quanto $p = 1$). Essa propriedade também é satisfeita em (3.4) para esse modelo de grafo (Takahashi *et al.*, 2012), conforme mostra a Figura 3.2a.

Nos modelos de Barabási-Albert e de Erdös-Rényi, também podemos observar que os

valores obtidos pela entropia espectral (3.4) satisfazem a noção intuitiva de quantidade de incerteza. A partir da distribuição empírica do espectro das grafos de Barabási-Albert e Erdös-Rényi, Takahashi *et al.* (2012) observaram as entropias estimadas a partir de diferentes parâmetros. Como podemos ver na Figura 3.2b, a entropia do grafo de Barabási-Albert é inversamente proporcional ao expoente de escala (p_s). Quando p_s é baixo, a construção do grafo se torna mais aleatória, pois a influência do grau dos vértices sobre a probabilidade de conectar um vértice a outro é pequena. Já, quando p_s é alto, o grau dos vértices tem um peso maior na escolha dos pares de vértices a serem conectados e, assim, a quantidade de incerteza será pequena. Finalmente, na Figura 3.2c, vemos que a entropia cresce à medida que aumentamos o parâmetro p_r. Lembramos que o parâmetro p_r 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. Assim, quando $p_r = 1$, temos um grafo construído de forma aleatória, como o grafo de Erdös-Rényi, e quando $p_r = 0$, temos um grafo determinado pela estrutura de anel, onde cada vértice está conectado aos K vértices mais próximos.

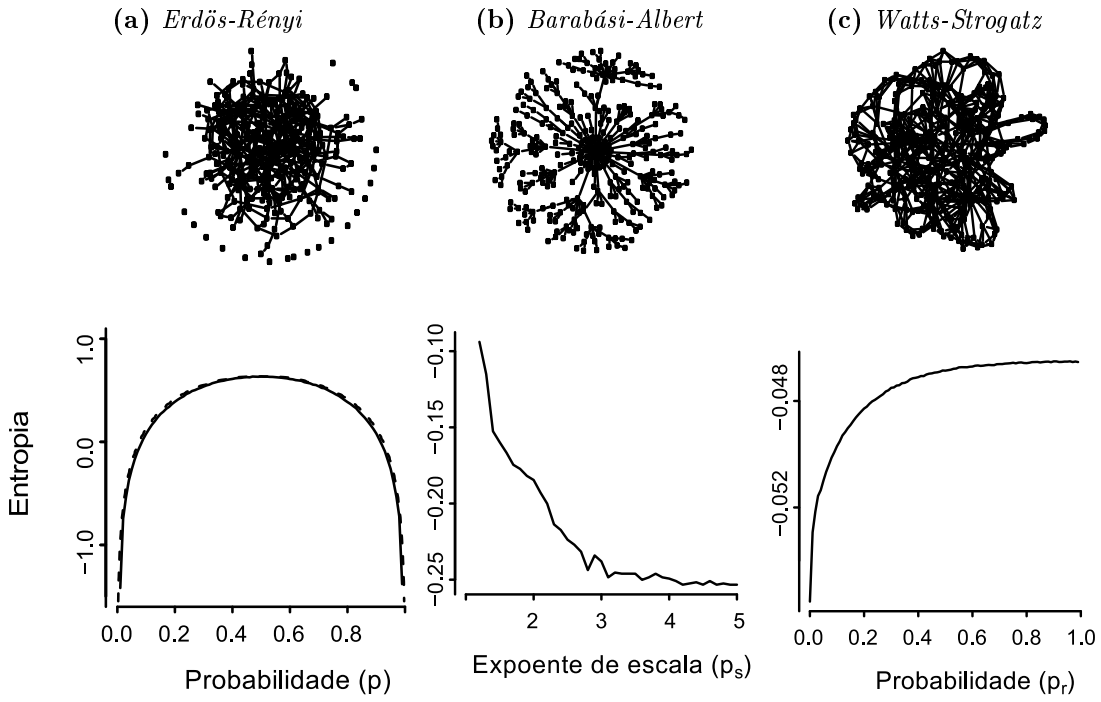


Figura 3.2: *Entropias espectrais de diferentes modelos. Na parte superior, são exibidos desenhos de grafos gerados pelos modelos de (a) Erdös-Rényi, (b) Barabási-Albert e (c) Watts-Strogatz. Os gráficos na parte inferior mostram entropias espectrais estimadas a partir das distribuições empíricas dos espectros de grafos gerados pelos modelos de (a) Erdös-Rényi, (b) Barabási-Albert e (c) Watts-Strogatz. Para a construção das curvas foram considerados diferentes valores dos parâmetros de cada modelo. Em (a), variou-se a probabilidade p de conectar dois pares de vértices, em (b), o parâmetro utilizado foi o expoente de escala p_s e, em (c), temos a probalilidade p_r de substituir uma aresta por outra que conecta um vértice escolhido aleatoriamente. Na figura (c), a linha tracejada mostra o valor da entropia obtido a partir da distribuição teórica do espectro. A figura foi adaptada de Takahashi et al. (2012).*

Uma vez definida a entropia espectral, a entropia cruzada entre duas densidades espectrais f_1 e f_2 (quantidade de incerteza quando f_2 é usada para estimar f_1) é definida como

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

onde $0 \log 0 = 0$.

3.5.3 Divergência de Kullback-Leiber

Enquanto a entropia quantifica a incerteza associada a uma variável aleatória, a divergência de Kullback-Leibler (KL) mede a informação perdida quando uma distribuição de probabilidade é utilizada para aproximar outra. Para grafos, a divergência de KL pode ser utilizada para para discriminar distribuições de probabilidade e para selecionar o modelo de grafo que melhor descreve o grafo observado. Formalmente, a divergência de Kullback-Leibler entre grafos é descrita a seguir.

Sejam g_1 e g_2 duas famílias de grafos com densidades espectrais f_1 e f_2, respectivamente. Se o suporte de f_2 contém o suporte de f_1, então a *divergência de* KL entre f_1 e f_2 é

$$\begin{aligned} KL(f_1|f_2) &= H(f_1, f_2) - H(f_1) \\ &= \int_{-\infty}^{+\infty} f_1(\lambda) \log \frac{f_1(\lambda)}{f_2(\lambda)} d\lambda, \end{aligned}$$

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. Em muitos casos, $KL(f_1|f_2)$ e $KL(f_2|f_1)$ são diferentes quando $f_1 \neq f_2$, isto é, KL é uma medida assimétrica.

3.5.4 Divergência de Jensen-Shannon

A divergência de Jensen-Shannon (JS) é uma alternativa simétrica à divergência de Kullback-Leibler. A *divergência de* JS entre as densidades espectrais f_1 e f_2 é definida 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).

Nós podemos interpretar a divergência de Jensen-Shannon como uma medida de diferenças estruturais entre dois grafos. A raiz quadrada da medida é uma métrica, isto é, 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.

Capítulo 4

Testes estatísticos entre grafos de coexpressão

As medidas apresentadas no capítulo anterior descrevem propriedades estruturais utilizadas para analisar redes complexas. Neste trabalho, estamos particularmente interessados em analisar grafos de coexpressão, que compartilham características de redes complexas, como distribuição dos graus com cauda pesada e alto coeficiente de *clustering*.

Um grafo de coexpressão é inferido a partir de uma amostra dos níveis de expressão genética de uma população. Assim, a análise diferencial de grafos de coexpressão requer métodos de inferência estatística. Neste capítulo, apresentaremos técnicas estatísticas para testar se dois grafos apresentam as mesmas propriedades estruturais, utilizando as medidas definidas no Capítulo 3. Nas seções seguintes enunciamos o problema e descrevemos os experimentos realizados para a análise diferencial de redes de coexpressão.

4.1 Enunciado do problema

Considere um conjunto de itens $V = \{v_1, v_2, \ldots v_{nV}\}$ e duas populações P_1 e P_2. O problema consiste em verificar, a partir de n_1 observações da população P_1 e n_2 observações de P_2, se a estrutura das interações entre os itens de V é igual em P_1 e P_2. No contexto de grafos de coexpressão, os itens correspondem a genes e cada observação corresponde ao nível de atividade de um gene em um experimento.

4.2 Construção do grafo de coexpressão

Um grafo de coexpressão é um grafo não-dirigido, em que cada vértice corresponde a um gene e uma aresta conectando um par de vértices indica uma associação ente os níveis de atividade dos genes correspondentes. Neste contexto, a relação representada por uma aresta indica a dependência estatística entre os níveis de atividade de dois genes (coexpressão).

Na primeira parte do trabalho, estudamos diferentes medidas para identificar e quantificar a dependência estatística entre os níveis de atividades dos genes, como as correlações de Pearson, de Spearman, de Kendall e de distância, a medida de HHG, a medida D de Hoeffding, a informação mútua e o coeficiente de informação máxima. Dada uma medida de dependência, consideramos três diferentes escalas para medir coexpressão: (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 ocorrência de falsos positivos dos testes de independência.

Pela definição de medida de dependência, vemos que a escala (i) quantifica a "força" de associação entre os níveis de expressão de dois genes. Na escala (ii), o resultado do teste de independência é levado em conta. P-valores baixos (próximos de zero) indicam associações "fortes", enquanto p-valores altos (próximos de 1) indicam associações "fracas". Assim, um menos o p-valor será maior quanto mais "forte" a coexpressão entre dois genes. 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 entendermos o procedimento utilizado na escala (iii), vamos supor que fixamos a probabilidade de ocorrer um erro tipo I em $\alpha = 0,05$. Então, em 100 testes independentes entre si, onde as hipóteses nulas são verdadeiras, a probabilidade de não cometermos o erro tipo I é $(1 - 0,05)^{100} = 0,006$. Desse modo, a probabilidade de rejeitar incorretamente pelo menos uma hipótese nula é $1 - 0,006 = 0,994$, um valor muito superior ao α desejado.

Um método utilizado para controlar a proporção de ocorrências do erro tipo I, ou a taxa de falsos positivos, em múltiplos testes de hipóteses, é o método FDR (*False Discovery Rate*) de Benjamini–Hochberg (Benjamini e Hochberg, 1995) descrito a seguir.

1. Considere as hipóteses sendo testadas H_1, H_2, ..., H_m e os respectivos p-valores p_1, p_2, ..., p_m.

2. Sejam $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$ os p-valores ordenados.

3. Seja k o maior i tal que $p_{(i)} \leq \frac{i}{m}\alpha$

4. Rejeitamos $H_{(i)}$, para $i = 1, 2, ..., k$.

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

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

Cada valor $\tilde{p}_{(i)}$ obtido pelo procedimento acima é chamado de q-valor ou "p-valor corrigido". Um q-valor próximo de um indica que os genes não estão significativamente coexpressos. Se o q-valor é próximo de zero, então há uma probabilidade pequena de os genes não estarem coexpressos. Assim, a escala (iii), isto é, um menos o q-valor quantifica a "força" de associação entre as atividades genéticas.

4.3 Testes estatísticos

Uma vez construídos os grafos de coexpressão, nosso objetivo é comparar, por meio de testes estatísticos, as estruturas de dois grafos. Dada uma estatística θ que quantifica as diferenças estruturais entre dois grafos, aplicaremos testes estatísticos descritos pelas seguintes hipóteses nula e alternativa:

$H_0 : \theta = 0$,

$H_1 : \theta > 0$.

4.3.1 Estatísticas dos testes

Sejam G_1 e G_2 dois grafos de coexpressão, cada um com n_V vértices. As estatísticas θ utilizadas nos testes estatísticos para comparar os grafos G_1 e G_2 são

- *Distância Euclidiada entre medidas de redes complexas.* Dados $\mathbf{x} = \{x_1, \ldots, x_{n_V}\}$ e $\mathbf{y} = \{y_1, y_2, \ldots, y_{n_V}\}$, a distância euclidiana, ajustada pelo tamanho do grafo (n_V) é dada por

$$\theta = \sqrt{\frac{\sum_{i=1}^{n_V} (x_i - y_i)^2}{n_V}}.$$

  Neste contexto, $\mathbf{x}$ e $\mathbf{y}$ correspondem a medidas de redes complexas obtidas para cada vértice dos grafos G_1 e G_2, respectivamente. As medidas utilizadas neste trabalho são as *centralidades de grau, de betweenness, de proximidade e de autovetor e o coeficiente de clustering.*

- *Diferença absoluta entre medidas de redes complexas.* Sejam x e y medidas de redes complexas obtidas para os grafos G_1 e G_2, respectivamente. A diferença absoluta entre x e y é

$$\theta = |x - y|.$$

  Utilizamos a *entropia espectral* como medida da estrutura de cada grafo.

- *Divergência de Jensen-Shannon entre densidades.* Sejam f_X e f_Y funções de densidade de probabilidade de características dos grafos G_1 e G_2, respectivamente. A estatística do teste entre as densidades f_X e f_Y é

$$\theta = JS(f_X, f_Y),$$

onde JS é a divergência de Jensen-Shannon entre duas funções de densidade de probabilidade. Neste trabalho, são obtidas as funções de densidade de probabilidade do espectro do grafo (*distribuição do espectro*) e do grau do vértice (*distribuição do grau*).

Detalhes sobre a implementação de cada estatística testada estão descritos na Seção 5.2.

4.3.2 Teste de permutação

Para calcular o p-valor do teste de hipóteses, é preciso obter a distribuição de probabilidade da estatística do teste sob a hipótese nula, o que é feito usualmente a partir de uma aproximação assintótica ou de reamostragens aleatórias dos dados. No primeiro caso, utiliza-se uma fórmula analítica da distribuição, que, para algumas estatísticas, é desconhecida. Além disso, nesse tipo de teste, muitas vezes é preciso fazer suposições sobre a distribuição dos dados que não podem ser verificadas. No segundo caso, os testes são usualmente baseados em reamostragens aleatórias obtidas pelo método de Monte Carlo, em que a geração aleatória das amostras (com distribuição uniforme) é simulada pelo computador. Nesse tipo de teste não fazemos suposições sobre a distribuição dos dados. Em geral, supomos apenas que as observações são independentes e identicamente distribuídas.

Um dos métodos baseados em reamostragens mais utilizados para fazer testes de hipóteses é o teste de permutação. Nesse método, para simular a hipótese nula de que dois conjuntos de dados vieram da mesma distribuição, os conjuntos são misturados e dois novos conjuntos são construídos a partir de sorteios sem reposição. Equivalentemente, podemos permutar os rótulos que indicam a que conjunto cada observação pertence.

A ideia de teste de permutação foi descrita inicialmente por Fisher (1935), que propôs fazer o teste de hipóteses a partir de todas as permutações possíveis dos rótulos das observações. Posteriormente, Dwass (1957) propôs utilizar um subconjunto obtido aleatoriamente a partir do conjunto de todas as permutações. Esta abordagem é útil quando o número de amostras é grande e a enumeração de todas as permutações é inviável ou muito custosa. Usualmente, o subconjunto de permutações é obtido pelo método de Monte Carlo, em que a geração aleatória (com distribuição uniforme) das permutações é simulada computacionalmente, resultando em amostras pseudoaleatórias. Neste trabalho, todos os testes de permutação utilizam o método de Monte Carlo.

Dados uma estatística θ que mede a diferença entre dois grafos de coexpressão, um número N de permutações e duas amostras dos níveis de expressão genética de experimentos das populações P_1 e P_2, os testes de permutação utilizados neste trabalho consistem nos seguintes passos:

1. Construa o grafo G_1 a partir dos experimentos de P_1 e G_2 a partir dos experimentos de P_2, utilizando um método para inferir redes de coexpressão.

2. Calcule o valor de $\theta(G_1, G_2)$.

3. Faça N permutações 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 $\frac{1+B}{1+N}$, onde B é o número de estatísticas Monte Carlo maiores ou iguais à estatística observada originalmente.

4.4 Conjunto de dados

Neste trabalho, os testes apresentados são aplicados em simulações e em dados de expressão genética obtidos a partir de tecidos tumorais de dois tipos de câncer de cérebro que são morfologicamente muito semelhantes, mas que apresentam prognósticos e tratamentos diferentes: o oligodendroglioma grau II e o astrocitoma grau II. Ambos os tumores são classificados como gliomas, pois são oriundos de células gliais, que são responsáveis por proteger e nutrir os neurônios.

O conjunto de dados que utilizamos é composto por 65 microarranjos de astrocitoma grau II e 30 microarranjos de oligodentroglioma grau II, todos da plataforma *Affymetrix Human Genome U133 plus 2.0* e obtidos na base de dados REMBRANDT (NCI) (https://caintegrator.nci.nih.gov/rembrandt). Os dados brutos (aquivos CEL) foram pré-processados com o método RMA (*Robust Multichip Average*) (Irizarry *et al.*, 2003b) para o ajuste do fundo, normalização e sumarização. Esses precedimentos são detalhados no Apêndice A. Para agrupar as sondas a partir de informações atualizadas do genoma e do transcriptoma, foram utilizados arquivos CDF (*Chip Description File*) do Brainarray (Dai *et al.*, 2005) (versão 18.0.0, ENTREZG). Após o pré-processamento, 19.674 genes foram identificados no conjunto de dados.

Para testar se há diferenças de coexpressão de conjuntos de genes, utilizamos uma coleção de 1.320 vias canônicas do MSigDB v4.0 (Molecular Signatures Database) (Subramanian *et al.* , 2005). Cada via corresponde a um conjunto de de genes, que usualmente representa um processo biológico. Para cada conjunto de genes, aplicamos um teste estatístico entre os microarranjos de oligodendroglioma grau II e astrocitoma grau II.

4.5 Resultados e discussões

A fim de avaliar o desempenho dos métodos propostos para comparar duas redes (grafos), realizamos simulações e uma aplicação dos métodos em dados de microarranjos de DNA. Nas análises descritas a seguir, a rede de coexpressão de genes é modelada de duas formas:

(i) grafo não-dirigido sem peso nas arestas e (ii) grafo não-dirigido com peso nas arestas. Seja q_{sp} o p-valor do teste de Spearman (1904) entre os níveis de expressão de dois genes corrigido para múltiplos testes pelo método de Benjamini e Hochberg (1995). Em (i), o peso de cada aresta é $1-q_{sp}$, e em (ii) todas as arestas com peso $1-q_{sp}$ menor ou igual a 0,95 (isto é, testes com menos de 5% de significância após correção de múltiplos testes) são removidas e as arestas remanescentes no grafo recebem peso peso 1.

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.

Tanto as simulações quanto a aplicação em dados de expressão genética utilizam 95 microarranjos da base de dados do REMBRANDT (NCI) oriundos de 65 experimentos de astrocitoma grau II (AII) e 30 experimentos de oligodendroglioma grau II (ODII). A seguir descrevemos as simulações e testes realizados.

4.5.1 Simulações

A fim de simular novos conjuntos de dados a partir dos dados sob estudo, com a hipótese de que os dados de AII e ODII são oriundos da mesma distribuição (hipótese nula), utilizamos uma técnica de reamostragem conhecida como *bootstrap*. O *bootstrap* foi proposto por Efron (1979) para estimar a distribuição amostral de uma estatística a partir de reamostragens aleatórias com reposição dos dados originais. Esta técnica pode ser utilizada em diversas aplicações, como estimadores para o erro padrão de uma estatística e intervalos de confiança para parâmetros de uma população, além da construção de testes de hipóteses.

A ideia básica do *bootstrap* é que a amostra coletada é usualmente a melhor estimativa da população da qual a amostra foi retirada. No caso em que as variáveis aleatórias podem ser assumidas como independentes e igualmente distribuídas, a técnica pode ser implementada construindo-se reamostragens aleatórias do conjunto de observações, de mesmo tamanho do conjunto original. Cada reamostragem é construída a partir do sorteio com reposição dos elementos do conjunto de observações.

Neste trabalho, realizamos o seguinte procedimento *bootstrap* para gerar novos conjuntos de dados sob a hipótese nula:

Sejam M um conjunto com n_M microarranjos e n_1 e n_2 dois inteiros positivos tais que $n_1 + n_2 \leq n_M$. Dados um número N_{boot} de reamostragens *bootstrap* e um conjunto V com n_V genes presentes nos microarranjos de M, faça, para $1 \leq i \leq N_{boot}$:

1. Construa o conjunto M_1^i sorteando aleatoriamente, com reposição, n_1 microarranjos de M.

2. Construa o conjunto M_2^i sorteando aleatoriamente, com reposição, n_2 microarranjos de M.

No procedimento acima, $N_{boot} = 1.000$ e M é o conjunto de 95 microarranjos de tecidos gliais. Note que os dados de astrocitoma grau II e oligodendroglioma grau II estão misturados no cenário construído. O resultado esperado é que M_1^i e M_2^i representem amostras de uma mesma popução (M). Assim, o cenário apresentado simula a hipótese nula dos testes propostos.

A partir das reamostragens obtidas pelo procedimento *bootstrap*, aplicamos testes para avaliar o controle da taxa de falsos positivos. Além disso, construímos conjuntos de dados sob H_1 para avaliar o poder estatístico dos métodos.

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 dois grafos, testamos $H_0 : \theta = 0$ *versus* $H_1 : \theta > 0$. Para verificar se o testes controlam a taxa de falsos positivos nos dados de microarranjos de DNA, geramos dados conforme procedimento descrito no início da Seção 4.5.1, com V (conjunto de genes) sorteado aleatoriamente (de tamanho $n_V = 20, 40, 100$), $n_1 = 65$ microarranjos (número de experimentos com tecidos de AII) e $n_2 = 30$ microarranjos (número de experimentos com tecidos de ODII). Dessa forma, geramos amostras com o mesmo tamanho dos dados originais, e podemos verificar o controle da taxa de falsos positivos em circunstâncias semelhantes.

Para avaliar o controle de falsos positivos, construímos curvas ROC com os p-valores dos testes aplicados nos $N_{boot} = 1.000$ conjuntos de dados simulados, considerando 1.000 permutações em cada teste de permutação (uma descrição detalhada dos testes de permutação realizados está na Seção 4.3.2). As curvas construídas indicam, no eixo x, o nível de significância dos testes (taxa esperada de falsos positivos) e, no eixo y, a proporção de vezes que a hipótese nula foi rejeitada (taxa observada de falsos positivos). Para sumarizar os resultados das curvas ROC, utilizamos a área sob cada curva ROC, que deverá ser próxima ou menor do que 0,5 quando os testes controlam a taxa de falsos positivos. Maiores detalhes sobre as curvas ROC podem ser consultados na Seção 2.1.

Na Tabela 4.1 mostramos as áreas sob as curvas ROC para cada estatística utilizada nos testes. Podemos observar que todas as áreas sob as curvas ROC foram menores ou próximas de 0,5, indicando que a proporção observada de falsos positivos foi controlada pelo nível de significância (α) dos testes. Em particular, verificamos que as taxas de falsos positivos com nível de significância (α) de 0,01, 0,05 e 0,1 (Tabela 4.2 e Tabela 4.3) têm de fato valores próximos ou menores do que α (proporção esperada de falsos positivos).

Quando é calculado um p-valor exato, como no teste de permutação, espera-se que a taxa

Tabela 4.1: *Áreas debaixo das curvas ROC sob a hipótese nula para cada estatística utilizada nos testes de permutação. Os testes foram realizados para grafos com e sem peso nas arestas e para diferentes tamanhos de grafos ($n_V = 20, 40, 100$).*

Estatística	Sem peso			Com peso		
	n_V			n_V		
	20	40	100	20	40	100
Distribuição espectral	0,098	0,484	0,492	0,494	0,498	0,494
Entropia espectral	0,095	0,486	0,491	0,493	0,496	0,494
Distribuição do grau	0,091	0,483	0,490	0,495	0,500	0,496
Centralidade de grau	0,489	0,491	0,495	0,499	0,495	0,497
Centralidade de *betweenness*	0,501	0,518	0,495			
Centralidade de proximidade	0,498	0,486	0,491			
Centralidade de autovetor	0,487	0,491	0,495	0,511	0,489	0,502
Coeficiente de *clustering*	0,489	0,486	0,492	0,499	0,498	0,496

Tabela 4.2: *Proporção de falsos positivos sob H_0 para grafos sem peso nas arestas. Proporção de falsos positivos sob a hipótese nula (rejeições de H_0) para cada estatística utilizada nos testes de permutação, com diferentes níveis de significâcia ($\alpha = 0,01, 0,05, 0,10$). Foram considerados grafos de diferentes tamanhos ($n_V = 20, 40, 100$) e sem peso nas arestas.*

Estatística	$\alpha = 0,01$			$\alpha = 0,05$			$\alpha = 0,10$		
	n_V			n_V			n_V		
	20	40	100	20	40	100	20	40	100
Distribuição espectral	0,005	0,009	0,01	0,013	0,047	0,053	0,021	0,092	0,109
Entropia espectral	0,003	0,009	0,013	0,013	0,043	0,055	0,017	0,095	0,109
Distribuição do grau	0,003	0,01	0,007	0,009	0,042	0,051	0,016	0,091	0,111
Centralidade de grau	0,011	0,008	0,009	0,056	0,051	0,054	0,098	0,088	0,11
Centralidade de *betweenness*	0,014	0,013	0,015	0,053	0,062	0,059	0,106	0,112	0,109
Centralidade de proximidade	0,011	0,007	0,008	0,052	0,035	0,054	0,097	0,086	0,102
Centralidade de autovetor	0,01	0,007	0,008	0,046	0,05	0,052	0,091	0,102	0,112
Coeficiente de *clustering*	0,008	0,007	0,011	0,04	0,047	0,056	0,084	0,098	0,109

Tabela 4.3: *Proporção de falsos positivos sob H_0 para grafos com peso nas arestas. Proporção de falsos positivos sob a hipótese nula (rejeições de H_0) para cada estatística utilizada nos testes de permutação, com diferentes níveis de significâcia ($\alpha = 0,01, 0,05, 0,10$). Foram considerados grafos de diferentes tamanhos ($n_V = 20, 40, 100$) e com peso nas arestas.*

Estatística	$\alpha = 0,01$			$\alpha = 0,05$			$\alpha = 0,10$		
	n_V			n_V			n_V		
	20	40	100	20	40	100	20	40	100
Distribuição espectral	0,008	0,01	0,01	0,042	0,041	0,057	0,089	0,101	0,118
Entropia espectral	0,009	0,009	0,005	0,047	0,043	0,059	0,079	0,097	0,117
Distribuição do grau	0,01	0,013	0,009	0,054	0,044	0,063	0,095	0,089	0,113
Centralidade de grau	0,011	0,012	0,009	0,045	0,05	0,056	0,102	0,101	0,112
Centralidade de autovetor	0,018	0,016	0,008	0,048	0,046	0,043	0,105	0,103	0,09
Coeficiente de *clustering*	0,013	0,011	0,009	0,049	0,044	0,064	0,094	0,099	0,113

de falsos positivos observada seja próxima do nível de significância do teste, o que acontece quando a área debaixo da curva ROC é 0,50. Contudo, algumas simulações apresentaram

taxas de falsos positivos menores do que o esperado, isto é, áreas sob as curvas ROC menores do que 0,50. Note que, mesmo não sendo o comportamento esperado de um teste exato, a taxa de falsos positivos continua controlada pelo nível de significância dos testes.

Como mostra a Tabela 4.1, apenas testes feitos com grafos sem pesos nas arestas apresentaram comportamento diferente do esperado, isto é, áreas sob as curvas ROC menores do que 0,50. Foi verificado que, nesses casos, alguns grafos vazios foram gerados nas simulações, sendo que os testes baseados em estimadores da função de densidade de probabilidade (distribuição do espectro, distribuição do grau, entropia espectral) supõe que os grafos não são vazios. Maiores detalhes sobre a implementação desses métodos podem ser verificados na Seção 5.2.

Para evitar esse problema, o número mínimo de genes a ser testado em cada conjunto deve ser escolhido cuidadosamente. Testes com conjuntos pequenos (por exemplo, menores do que 20) podem interferir na estimação das propriedades da rede, enquanto testes com conjuntos grandes (por exemplo, com mais de 1.000 genes) podem ter um grande custo computacional. Outra limitação dos testes é o tamanho da amostra, que, além restringir o poder estatístico, pode afetar a inferência do grafo de coexpressão, caso não seja suficientemente grande.

Poder estatístico

Após verificarmos que os dois métodos controlam a taxa de falsos positivos, fizemos simulações para observar a taxa de verdadeiros positivos (proporção de vezes que a hipótese nula é rejeitada corretamente). Para isso, criamos o seguinte cenário:

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

1. Calcule a correlação entre os genes de V.
2. Ordene os pares de genes pela correlação.
3. Inspecione os pares dos mais correlacionados para os menos correlacionados, escolhendo um gene de cada par de forma a obter, no fim do processo, γn_V genes.
4. Com os genes selecionados construa o conjunto V'.
5. Faça N_{boot} vezes:
   (a) Gere um conjunto M_1 com $n_1 = 40$ microarranjos e M_2 com $n_2 = 40$ microarranjos, conforme cenário construído sob a hipótese nula (maiores detalhes estão descritos no início da Seção 4.5.1).
   (b) Permute aleatoriamente os níveis de expressão de cada gene do conjunto V', apenas nos microarranjos do conjunto M_2.

Assim como nas simulações para verificar o controle da taxa de falsos positivos, o procedimento acima, no passo 5 (a), gera dois conjuntos sob a hipótese nula. Contudo, ao permutarmos os níveis de expressão dos genes com as maiores correlações do conjunto de dados, no passo 5 (b), várias correlações são "quebradas" no conjunto M_2, fazendo com que as redes de genes em M_1 e M_2 sejam diferentes.

Para verificar o poder estatístico dos métodos no cenário apresentado, geramos $N_{boot} = 1.000$ conjuntos e aplicamos, em cada conjunto, os testes de permutação propostos com 1.000 permutações. Em seguida, construímos curvas ROC, com γ (proporção de genes que mudam) assumindo os valores 0,05, 0,1, 0,15, 0,2, 0,25, 0,30 e 0,5. Nas curvas construídas, o eixo y corresponde à proporção de testes rejeitados (poder empírico do teste) e o eixo x aos níveis de significância (probabilidade de ocorrer um falso positivo). Para sumarizar a informação da curva ROC, novamente utilizamos a área debaixo da curva. Áreas próximas de 0,50 equivalem a decisões aleatórias de rejeitar ou aceitar a hipótese nula. Quanto mais próxima a área sob a curva ROC estiver de 1, maior o poder estatístico do teste.

Na Tabela 4.4 e na Tabela 4.5 mostramos as áreas debaixo das curvas ROC para grafos com e sem peso, respectivamente. Observamos que, conforme o esperado, todas as áreas crescem com o aumento de γ (proporção de genes que mudam). As curvas ROC ficaram, assim, longe da diagonal quando γ é grande.

Tabela 4.4: *AUC sob H_1 para grafos sem peso nas arestas. Áreas debaixo das curvas ROC sob a hipótese alternativa para cada estatística utilizada nos testes de permutação. Foram considerados diferentes valores de γ (γ = 0,05, 0,10, 0,15, 0,20, 0,25, 0,30, 0,50), onde γ é a proporção de genes que tiveram os níveis de expressão permutados.*

Estatística	γ						
	0,05	0,1	0,15	0,2	0,25	0,3	0,5
Distribuição espectral	0,497	0,578	0,615	0,695	0,781	0,845	0,949
Entropia espectral	0,502	0,575	0,613	0,693	0,778	0,839	0,941
Distribuição do grau	0,496	0,566	0,597	0,669	0,762	0,832	0,958
Centralidade de grau	0,564	0,757	0,789	0,866	0,937	0,971	0,999
Centralidade de *betweenness*	0,482	0,435	0,450	0,469	0,467	0,524	0,755
Centralidade de proximidade	0,585	0,785	0,877	0,955	0,980	0,993	0,999
Centralidade de autovetor	0,550	0,708	0,723	0,770	0,841	0,876	0,932
Coeficiente de *clustering*	0,493	0,516	0,539	0,591	0,645	0,720	0,907

Essas simulações sugerem que as medidas utilizadas neste trabalho podem ser utilizadas para identificar diferenças entre grafos de coexpressão, mesmo considerando-se diferentes níveis de significância. Note, contudo, que os experimentos não permitem concluir o desempenho comparativo entre os métodos, uma vez que a adequação de cada um dependerá das características dos dados a serem analisados e das perguntas que o pesquisador deseja responder.

Tabela 4.5: *AUC sob H_1 para grafos com peso nas arestas. Áreas debaixo das curvas ROC sob a hipótese alternativa para cada estatística utilizada nos testes de permutação. Foram considerados diferentes valores de γ (γ = 0,05, 0,10, 0,15, 0,20, 0,25, 0,30, 0,50), onde γ é a proporção de genes que tiveram os níveis de expressão permutados.*

Estatística	γ						
	0,05	0,1	0,15	0,2	0,25	0,3	0,5
Distribuição espectral	0,505	0,663	0,746	0,869	0,934	0,968	0,999
Entropia espectral	0,507	0,675	0,757	0,873	0,937	0,970	0,999
Distribuição do grau	0,529	0,706	0,802	0,908	0,955	0,980	0,999
Centralidade de grau	0,614	0,838	0,890	0,953	0,982	0,992	0,999
Centralidade de autovetor	0,648	0,881	0,915	0,963	0,986	0,994	0,999
Coeficiente de *clustering*	0,517	0,693	0,785	0,902	0,957	0,982	0,999

4.5.2 Aplicação em dados de microarranjo de DNA

Os métodos apresentados neste trabalho foram aplicados nos dados de microarranjos de DNA oriundos de astrocitoma grau II (65 amostras) e oligodendroglioma grau II (30 amostras), após o pré-processamento descrito na Seção 4.4. Escolhemos comparar as redes de coexpressão de genes do astrocitoma grau II (AII) e do oligodendroglioma grau II (ODII), pois o diagnóstico diferencial entre eles é difícil, já que esses tumores podem ser muito semelhantes morfologicamente. Contudo, o grau de malignidade dos dois tumores e o tratamento indicado aos portadores dessas doenças é diferente. Assim, a comunidade científica tem interesse em investigar as diferenças desses dois tipos de tumor cerebral.

A abordagem adotada neste trabalho foi aplicar os testes estatísticos para identificar vias de genes (grupos de genes relacionados a um processo biológico) que estejam diferencialmente reguladas nos dois tipos de câncer. Para isso, consideramos os conjuntos de vias biológicas canônicas do Molecular Signatures Database (MSigDB) (Subramanian *et al.*, 2005) com pelo menos 20 genes, resultando em 850 vias com 20 a 828 genes. Cada via corresponde a um conjunto de genes que participam de um processo biológico. As interações genéticas foram inferidas pela medida de Spearman (1904), conforme descrito no início da Seção 4.5. Para utilizar informação de todos os pares de genes da rede, consideramos grafos com peso nas arestas em todos os testes, exceto nos testes da centralidade de *betweenness* e da centralidade de proximidade, que, como discutido na Seção 3.2, no contexto deste trabalho, são implementados apenas para grafos sem peso nas arestas.

Para cada conjunto de genes (via), aplicamos os testes baseados nas estatísticas descritas na Seção 4.3.1 com 10.000 permutações Monte Carlo. Na Tabela 4.6, mostramos o número de conjuntos de genes identificados em comum pelas medidas de diferença estrutural de grafos, considerando diferentes níveis de significância ($\alpha = 0,01, 0,05, 0,10$). Uma lista de todos os conjuntos de genes que apresentaram p-valor menor do que $0,10$ nos testes é exibida no Apêndice D. As estatísticas baseadas na distribuição de grau, distribuição de espectro, entropia espectral e centralidade de grau identificaram mais conjuntos, seguidas pela centralidade de proximidade, centralidade de *betweenness* e centralidade de autovetor.

Tabela 4.6: *Número de conjuntos em comum que foram identificados pelos métodos para diferentes níveis de significância ($\alpha = 0,01, 0,05, 0,10$). No total, foram testados 850 conjuntos de genes envolvidos em vias biológicas. Para cada conjunto de genes, as estatísticas utilizadas nos testes medem as diferenças estruturais entre o grafo de coexpressão do astrocitoma grau II e o grafo de coexpressão do oligodendroglioma grau II, baseando-se na distribuição do espectro (DE), entropia espectral (EE), distribuição do grau (DG), centralidade de grau (CG), centralidade de* betweenness *(CP), centralidade de proximidade (CP), centralidade de autovetor (CA) e coeficiente de* clustering *(CoC).*

	α	DE	EE	DG	CG	CB	CP	CA	CoC
DE	0,01	11	11	8	9	0	4	0	9
	0,05	98	94	77	79	10	39	13	81
	0,1	192	179	166	162	34	85	41	162
EE	0,01		11	8	9	0	4	0	9
	0,05		97	76	78	9	39	11	80
	0,1		185	160	153	34	80	42	154
DG	0,01			17	11	0	4	1	12
	0,05			100	78	9	36	11	81
	0,1			205	156	35	85	40	161
CG	0,01				15	0	5	1	12
	0,05				94	11	36	14	90
	0,1				182	34	86	48	171
CB	0,01					8	0	0	0
	0,05					48	3	5	10
	0,1					99	8	14	32
CP	0,01						18	0	5
	0,05						65	7	37
	0,1						124	26	87
CA	0,01							8	1
	0,05							43	14
	0,1							92	44
CoC	0,01								14
	0,05								96
	0,1								179

Podemos notar que os testes baseados nas estatísticas que comparam a distribuição do espectro (DE), a entropia espectral (EE), a distribuição de grau (DG), a centralidade de grau (CG) e o coeficiente de *clustering* (CoC) compartilham grande parte dos conjuntos de genes cujos testes estatísticos apresentaram p-valor menor do que 0,10. Para verificar a semelhança entre os testes estatísticos, calculamos a correlação de Pearson entre os p-valores obtidos pelos testes. Podemos ver na Tabela 4.7 que, de fato, os p-valores dos testes das estatísticas DE, EE, DG, GC e CoC estão fortemente correlacionados entre si (coeficiente de correlação de Pearson maior do que 0,90). Tomando a centralidade de grau como referência, podemos explicar a correlação entre as medidas pelo fato de essas estatísticas estarem fortemente relacionadas ao grau do vétice. No caso das estatísticas DG e CG, há uma relação explícita na própria definiçao. No caso do coeficiente de *clustering* (CoC), podemos observar que as mudanças de conectividade entre os vizinhos de um vértice têm uma forte relação com

alterações nos graus dos vértices. Já as estatísticas DE e EE são baseadas no espectro de um grafo, que, por sua vez, apresenta uma forte relação com o grau dos vértices. Por exemplo, o maior autovalor da matriz de adjacência de um grafo é pelo menos a média entre os graus dos vértices do grafo e no máximo o maior grau do grafo.

Tabela 4.7: *Correlações de Pearson entre os p-valores dos testes estatísitcos. No total, foram testados 850 conjuntos de genes envolvidos em vias biológicas. Para cada conjunto de genes, as estatísticas utilizadas nos testes medem as diferenças estruturais entre o grafo de coexpressão do astrocitoma grau II e o grafo de coexpressão do oligodendroglioma grau II, baseando-se na distribuição do espectro (DE), entropia espectral (EE), distribuição do grau (DG), centralidade de grau (CG), centralidade de* betweenness *(CP), centralidade de proximidade (CP), centralidade de autovetor (CA) e coeficiente de* clustering *(CoC).*

	DE	EE	DG	CG	CB	CP	CA	CoC
DE	1,000	0,964	0,974	0,970	0,030	0,559	0,191	0,981
EE		1,000	0,916	0,910	0,014	0,542	0,169	0,921
DG			1,000	0,965	0,034	0,547	0,144	0,982
CG				1,000	0,021	0,546	0,313	0,991
CB					1,000	-0,079	-0,045	0,023
CP						1,000	0,112	0,549
CA							1,000	0,246
CoC								1,000

Na Tabela 4.7, podemos observar que os testes baseados na centralidade de *betweenness* (CB) e na centralidade de autovetor (CA) não apresentaram resultados semelhantes aos dos demais métodos. Pela definição da centralidade de *betweeness* (Seção 3.2), de fato, intuitivamente não esperamos uma relação direta entre essa estatística e o grau dos vértices. Além disso, é preciso considerar que a centralidade de *betweenness* foi aplicada em grafos sem peso, enquanto as estatísticas DE, EE, DG, GC, CoC e CA foram aplicadas em grafos com peso. Já a centralidade de autovetor de um vértice é calculada em termos de seus vizinhos, o que evidencia sua relação com o grau do vértice. Contudo, a centralidade do vértice depende também fortemente da centralidade dos seus vizinhos. Dessa forma, em muitos casos, as mudanças das centralidades dos demais vértices podem "mascarar" as mudanças do grau do vértice. Assim, os testes baseados na estatística CA não apresentam necessariamente resultados semelhantes aos dos testes baseados na centralidade de grau e nas demais estatísticas relacionadas. Já a centralidade de proximidade apresenta uma correlação de aproximadamente 0,50 com todos os demais métodos, exceto as medidas de CB e CA. Como a centralidade de proximidade é o inverso da média do comprimento do caminho mais curto entre um vértice e os demais vértices do grafo, intuitivamente, há uma relação dessa medida com o grau dos vértices que não é tão "forte" como no caso das estatísticas DE, EE, DG, GC e CoC. A relação também pode ser enfraquecida pelo fato de a centralidade de proximidade ser aplicada apenas em grafos sem peso nas arestas, enquanto as estatísticas DE, EE, DG, GC, CoC e CA foram aplicadas em grafos com peso nas arestas.

Estas análises sugerem grupos de genes que apresentam diferenças de coexpressão entre

AII e ODII. Para que os resultados sejam confirmados por pesquisadores de Biologia Molecular, é preciso também tratar os múltiplos testes. Dessa forma, espera-se que o pesquisador faça uma seleção dos conjuntos de genes de interesse, de forma a fazer o menor número de testes possível e evitar, assim, que p-valores significativos sejam "perdidos" após a correção dos múltiplos testes. Neste trabalho, consideramos uma coleção grande de conjuntos de genes com o objetivo de ilustrar o desempenho das medidas em detrimento de confirmações biológicas. Análises subsequentes dos conjuntos testados serão abordadas no próximo capítulo, em que apresentaremos a ferramenta desenvolvida para comparar grafos de coexpressão.

Capítulo 5

CoGA: *Co-expression Graph Analyzer*

O R (http://www.r-project.org/) é uma linguagem de programação e um ambiente de desenvolvimento integrado muito utilizado para analisar dados e para criar programas (conhecidos como pacotes) estatísticos. 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/Bioconductor) 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 do R chamado CoGA (*Co-expression Graph Analyzer*) sob a licença de software livre GNU GPL (*GNU General Public License*) v3. Além de ser um pacote com funções que podem ser chamadas pela linha de comando do R, o CoGA apresenta uma interface gráfica para que o programa seja mais facilmente utilizado por quem não tem conhecimentos de programação. Os arquivos de instalação, bem como o tutorial e o guia do usuário estão disponíveis na página http://www.ime.usp.br/~suzana/coga. A seguir, descrevemos as funcionalidades do pacote, a sua implementação e um exemplo ilustrativo de análise com o CoGA.

5.1 Descrição

O pacote CoGA é uma ferramenta com uma interface gráfica para analisar grafos de coexpressão de genes. Ele recebe dados de expressão genética e uma coleção pré-definida de conjuntos de genes a partir dos quais são feitos os testes estatísticos para comparar os grafos de coexpressão. O pacote também inclui ferramentas para analisar um conjunto particular de genes, como uma interface para visualização das redes, medidas de "importância" de cada gene do conjunto e a análise clássica de expressão diferencial, em que é verificada a diferença entre a expressão média/mediana dos genes. A Figura 5.1 mostra uma visão geral do programa.

Os arquivos que o CoGA recebe como entrada são: um com os dados de expressão genética

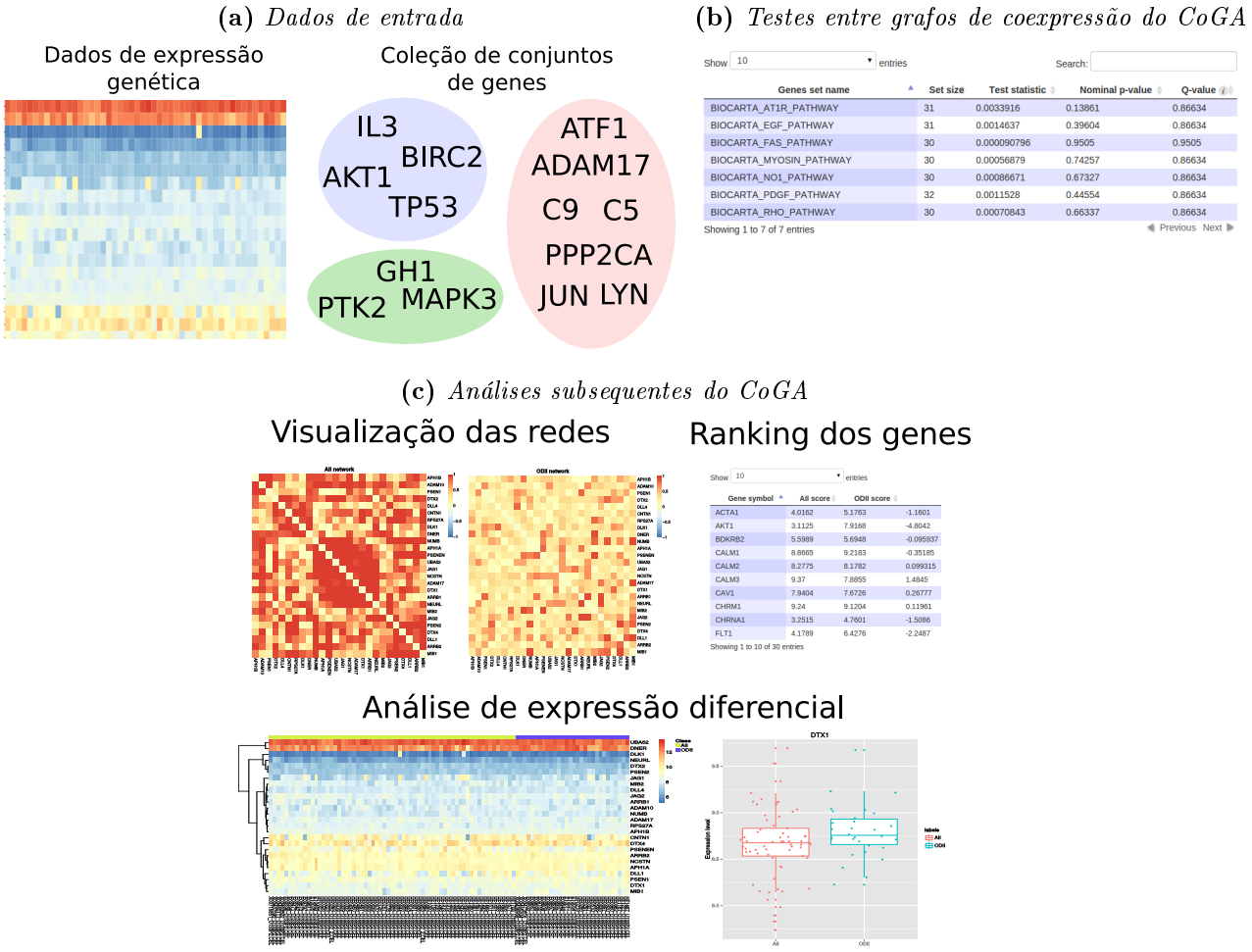


Figura 5.1: *Visão geral do CoGA. O CoGA recebe como entrada uma matriz contendo os dados de expressão genética, os rótulos das amostras e uma coleção de conjuntos de genes (A). O programa constrói um grafo de coexpressão para cada conjunto de genes e cada condição biológica e testa a igualdade entre as características estruturais das condições biológicas (B). O programa permite que o usuário analise cada conjunto de genes (C) a partir da visualização das matrizes de adjacência dos grafos de coexpressão, do ranking dos genes pertencentes ao conjunto e da análise clássica de expressão diferencial.*

já pré-processados, um com rótulos indicando o grupo a que cada experimento de expressão genética pertence e outro contendo uma coleção pré-definida de conjuntos de genes (por exemplo, conjuntos de genes que fazem parte de alguma via ou processo biológico). O primeiro arquivo contém uma matriz de expressão genética, em que as linhas correspondem aos genes e as colunas ao experimentos.

Se um gene é representado por mais de uma linha da matriz de expressão, é preciso converter os dados de forma que cada gene corresponda a apenas uma linha da matriz de expressão genética. Um exemplo em que esse procedimento é necessário são dados de microarranjo de DNA em que duas ou mais linhas representam transcritos diferentes de um mesmo gene. Nesse caso, o usuário deverá fornecer um arquivo adicional contendo dados de anotação, isto é, dados que descrevem a que gene cada linha da matriz de expressão genética corresponde. O programa permite que o usuário escolha um método para converter a matriz de expressão em uma matriz em que cada linha corresponde a um gene. Os métodos disponíveis são detalhados no Apêndice E.

Para obter a coleção pré-definida de conjuntos de genes e os dados de anotação, é possível utilizar bancos de dados públicos. Exemplos de coleções de conjuntos de genes e de dados de anotação no formato que é aceito pelo pacote são o *Molecular Signature Database* (MSigDB) (http://www.broadinstitute.org/gsea/msigdb/index.jsp) e o *Broad ftp site* (ftp://gseaftp.broadinstitute.org/pub/gsea/annotations), respectivamente. Ambos os bancos de dados são gratuitos.

Uma vez fornecidos os dados, o usuário poderá realizar testes entre os grafos de coexpressão, que são inferidos a partir dos dados de dois grupos usualmente representando duas condições biológicas. Sejam M_1 e M_2 dois grupos que rotulam as colunas da matriz de expressão genética recebida como entrada do programa. Para cada conjunto V da coleção pré-definida de conjuntos de genes fornecida pelo usuário, é feito um teste estatístico entre os grafos de coexpressão $G_1 = (V, E_1)$ (grafo inferido a partir dos experimentos de M_1) e $G_2 = (V, E_2)$ (grafo inferido a partir dos experimentos de M_2). Para a construção do grafo, o usuário pode escolher entre três medidas utilizadas para identificar associações monotônicas entre os níveis de expressão do gene, são elas a correlação de Pearson (Pearson, 1920), de Spearman (Spearman, 1904) e de Kendall (Kendall, 1938). Além disso, é possível escolher o tipo do grafo (com peso ou sem peso nas arestas) e a escala utilizada para quantificar a associação entre as atividades genéticas (valor absoluto da correlação, um menos o p-valor ou um menos o q-valor obtido pelo método de Benjamini e Hochberg (1995)).

Os testes realizados pelo programa utilizam estatísticas baseadas na distribuição do espectro, entropia espectral, distribuição do grau, centralidade de *betweenness*, centralidade de proximidade, centralidade de autovetor e coeficiente de *clustering*, conforme procedimento descrito na Seção 4.3. Após executar os testes estatísticos, o pacote devolve uma tabela contendo o nome e o tamanho de cada conjunto de genes, a estatística utilizada nos testes, os p-valores nominais e os q-valores obtidos pelo método FDR (*False Discovery Rate*) (Benjamini e Hochberg, 1995).

Além dos testes estatísticos, o pacote fornece uma interface para visualizar as alterações nos grafos de co-expressão, uma lista das diferenças entre as entre as correlações dos pares de genes, as propriedades de rede de cada conjunto de gene (por exemplo, entropia espectral, centralidade média, coeficiente de *clustering* médio), uma lista de genes ordenada por alguma medida de "importância" do gene e a análise clássica de expressão diferencial.

5.2 Implementação

O CoGA foi implementado em R (http://www.r-project.org/) e utiliza os seguintes pacotes: (i) `shiny`, `shinyBS`, `yaml`, `whisker` e `RJSONIO` para a interface web; (ii) `igraph` para as propriedades estruturais dos grafos; (iii) `WGCNA` para converter a matriz de expressão genética em uma matriz com uma linha por gene; (iv) `ggplot2`, `pheatmap`, e `RColorBrewer` para produzir os gráficos; e (v) `Hmisc` e `psych` para a construção dos grafos. Para algumas funcionalidades da interface gráfica, foram utilizados códigos dos pacotes `rCharts` (https://github.com/ramnathv/rCharts) e `shinyIncubator` (https://github.com/rstudio/shiny-incubator).

A Tabela 5.1 mostra as funções e pacotes utilizados na implementação de cada estatística do teste entre grafos de coexpressão. Para estimar as funções de densidade de probabilidade nas estatísticas baseadas na distribuição do espectro (DE), na entropia espectral (EE) e na distribuição do grau (DG), utilizamos o Kernel Gaussiano implementado pela função `density` do pacote `stats` do R, com largura de banda dada por $h = (\max(\mathbf{x}) - \min(\mathbf{x}))/k$ (Sain e Scott, 1996), onde $k = \lceil \log_2 n_V + 1 \rceil$ (Sturges, 1926) e $\mathbf{x}$ é um vetor de características do grafo. Para as estatísticas DE e EE, o vetor $\mathbf{x}$ representa os autovalores da matriz de adjacência e, para a estatística DG, o vetor representa os graus dos vértices do grafo.

O estimador da função de densidade de probabilidade baseado no Kernel Gaussiano pode ser interpretado como uma versão suavizada de um histograma. Dadas n_V observações $\{x_1, x_2 \dots x_{n_V}\}$, 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. Essa 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

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

Note que assumimos $h > 0$. Pelo critério de Sturges, que foi adotado neste trabalho para calcular a largura de banda, h vale zero sempre que o grafo for vazio. Assim, 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.

Tabela 5.1: *Funções do R utilizadas para implementar os testes entre grafos de coexpressão disponíveis no CoGA.*

Estatística	Grafos com peso	Grafos sem peso	Funções do R
Divergência de Jensen-Shannon entre densidades espectrais (DE)	Sim	Sim	Função `density` do pacote `base`
Diferença absoluta entre as entropias espectrais (EE)	Sim	Sim	Função `density` do pacote `base`
Divergência de Jensen-Shannon as densidades dos graus (DG)	Sim	Sim	Função `density` do pacote `base` e função `graph.strength` do pacote `igraph`
Distância euclidiana entre as médias dos graus (CG)	Sim	Sim	Função `graph.strength` do pacote `igraph`
Distância euclidiana entre as médias das centralidades de *betweenness* (CB)	**Não**	Sim	Função `graph.strength` do pacote `igraph`
Distância euclidiana entre as médias das centralidades de proximidade (CP)	**Não**	Sim	Função `closeness` do pacote `igraph`
Distância euclidiana entre as médias das centralidades de autovetor (CA)	Sim	Sim	Função `evcent` do pacote `igraph`
Distância euclidiana entre as médias dos coeficientes de *clustering* (CoC)	Sim	Sim	Função `transitivity` do pacote `igraph`

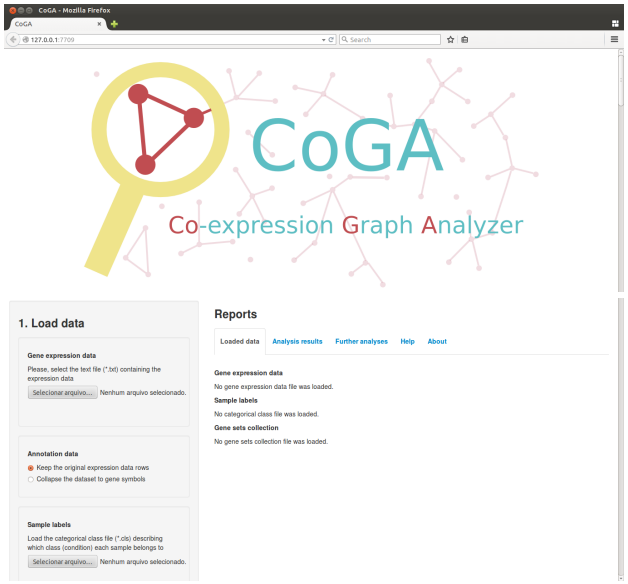
5.3 Exemplo ilustrativo

Nesta seção nós exemplificamos o uso do pacote CoGA com o conjunto de dados descrito na Seção 4.4, que contém dados de microarranjos de DNA oriundos de tecidos gliais de astrocitoma grau II (AII) e oligodendroglioma grau II (ODII). Além dos níveis de expressão genética obtidos pelos microarranjos, o usuário deve fornecer uma coleção de conjuntos de genes que correspondem aos vértices dos grafos a serem testados. Neste exemplo, nós utilizamos as vias canônicas do MSigDB v4.0 (Subramanian *et al.*, 2005), em que os genes estão agrupados de acordo com o processo biológico do qual eles participam. Depois de configurar o tamanho mínimo dos conjuntos de genes para 20, apenas 850 dos 1.320 conjuntos permaneceram nas análises.

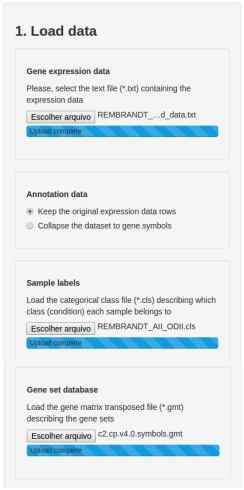
Para realizar os testes entre grafos de coexpressão a partir da interface gráfica do CoGA, seguimos os seguintes passos (ilustrados na Figura 5.2.): (i) após carregar o pacote no ambiente do R, digite o comando `runCoga()`; (ii) espere o navegador abrir e carregar a página inicial do CoGA; (iii) carregue a matriz de expressão genética e a coleção de conjuntos de genes na barra lateral do CoGA; (iv) configure os parâmetros de execução na barra lateral do CoGA; (v) clique no botão "Start analysis" para dar início aos testes entre os grafos de coexpressão; e (vi) espere os testes terminarem de executar.

Neste exemplo, utilizamos grafos com peso nas arestas, em que, a cada par de genes é associado um menos o p-valor do teste de Spearman (Spearman, 1904) corrigido para múltiplos testes pelo método de Benjamini e Hochberg (1995). O valor de cada aresta mede, assim, a dependência estatística entre as atividades dos genes correspondentes. Em seguida aplicamos testes de permutação, com 10.000 permutações aleatórias dos rótulos dos experi-

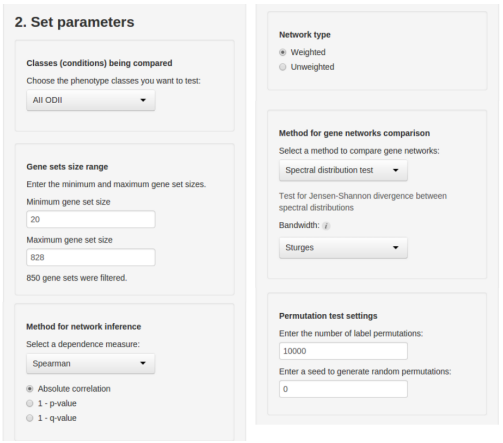
(a) *Página inicial*



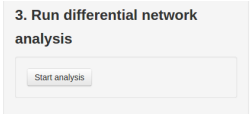
(b) *Barra lateral com as opções para carregar os arquivos*



(c) *Barra lateral com as configurações dos testes*



(d) *Botão que aciona os testes de coexpressão*



(e) *Barra de progresso dos testes*

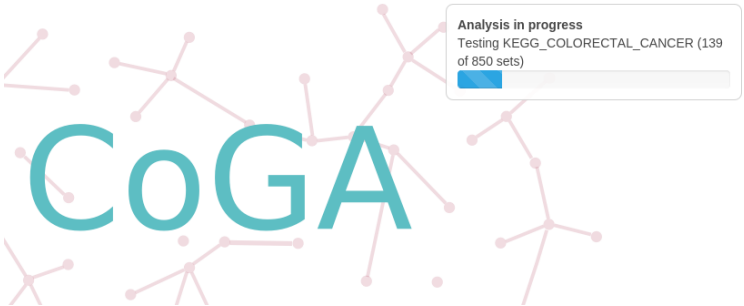


Figura 5.2: *Passos para executar testes entre grafos de coexpressão a partir da interface gráfica do CoGA.*

mentos, para cada uma das estatísticas disponíveis e cada um dos 850 conjuntos de genes da via canônica do MSigDB v4, conforme descrito na Seção 4.5.2.

Para ilustrar as funcionalidades do pacote além dos testes de coexpressão descritos na Seção 4.5.2, aplicamos o CoGA para explorar um dos conjuntos de genes testados. Para isso, escolhemos o conjunto que apresentou o menor p-valor (p-valor = 0,0019) no teste da divergência de Jensen-Shannon entre densidades espectrais, que identifica diferenças relacionadas com diversas propriedades estruturais de grafos. Ademais, esse conjunto apresentou p-valor menor do que 0,05 em todos os testes, exceto nos testes da centralidade de *betweenness* e de autovetor. O conjunto é identificado pelo nome REACTOME ACTIVATED NOTCH1 TRANSMITS SIGNAL TO THE NUCLEUS e será abreviado por RANTSN.

Interessantemente, os genes do conjunto RANTSN desempenham um papel importante no desenvolvimento do sistema nevoso central e influenciam a diferenciação dos astrócitos. Os genes desse cojunto, que fazem parte da via do Notch, estão envolvidos no desenvolvimento de gliomas, atuando na proliferação celular, na apoptose (morte celular programada) (Purow *et al.*, 2005) e na migração e invasão celular (Zhang *et al.*, 2012). Além disso, a via do Notch já foi descrita como alvo terapêutico em gliomas (Stockhausen *et al.*, 2010). Outros tipos de câncer também estão associados com alterações na regulação da via do Notch, como o câcer de pulmão, de mama e de pâncreas (Stockhausen *et al.*, 2010).

Para analisar o grafo de coexpressão e os níveis de atividades dos genes do conjunto RANTSN, utilizamos as seguintes ferramentas do CoGA: (i) visualização da rede (matriz de adjacência), (ii) propriedades estruturais da rede, (iii) medidas da "importância" dos genes do conjunto e (iv) análise da diferença da expressão média/mediana de cada gene do conjunto. Explicamos cada uma dessas ferramentas a seguir.

5.3.1 Visualização da rede

A ferramenta de visualização da rede mostra, para um dado conjunto de genes, a matriz dos graus de associação entre os níveis de expressão genética para cada condição biológica (oligodendroglioma grau II e astrocitoma grau II, neste exemplo). A Figura 5.3 mostra as matrizes construídas com os genes do conjunto RANTSN a partir dos dados de astrocitoma grau II (gráfico da esquerda, Figura 5.3a) e de oligodendroglioma grade II (gráfico da direita, Figura 5.3b).

Outro gráfico gerado pela ferramenta de visualização é a matriz das diferenças absolutas entre os graus de associação das duas condições (Figura 5.3c). A cor vermelha indica os pares de genes que têm as maiores diferenças de associação entre AII e ODII, enquanto a cor azul indica as menores diferenças. Além de visualizar as matrizes de associações, o usuário pode ver o valor de cada entrada das matrizes e ordenar os pares de genes de acordo com os valores associados às arestas. Neste exemplo, nós geramos uma tabela com todos os graus de associação entre os níveis de expressão genética e ordenamos os pares de genes pela diferença absoluta entre o peso das arestas (Apêndice F). O par de genes com maior diferença no

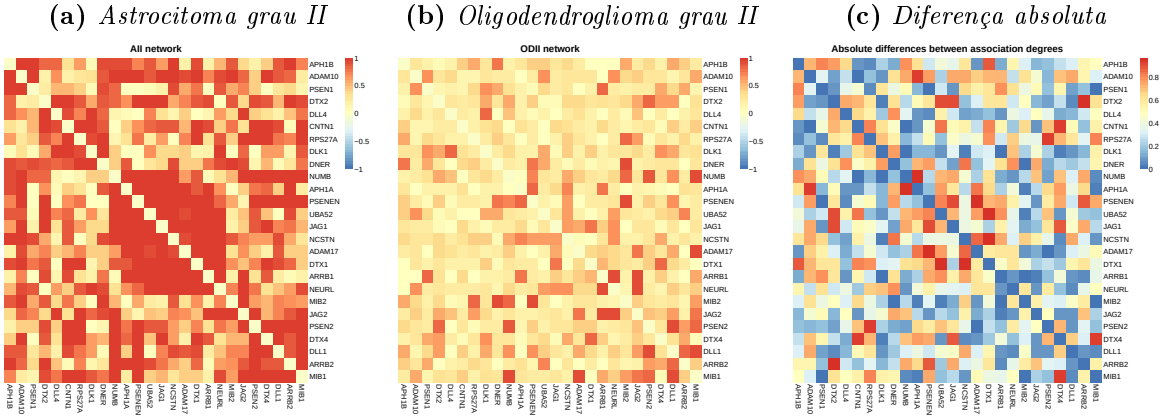


Figura 5.3: *Visualização dos grafos de coexpressão do conjunto REACTOME ACTIVATED NOTCH1 TRANSMITS SIGNAL TO THE NUCLEUS: (A) matriz de adjacência do grafo de coexpressão do astrocitoma grau II, abreviado por AII; (B) matriz de adjacência do grafo de coexpressão do oligodendroglioma grau II, abreviado por ODII; e (C) diferenças absolutas entre as matrizes de AII e ODII. Em (A) e (B) a cor vermelha indica um alto grau de associação entre as atividades dos genes da linha e da coluna, enquanto a cor amarela indica uma associação baixa. Em (C) as cores vermelha, azul e amarela representam, respectivamente diferenças altas, baixas e intermediárias entre as entradas das matrizes de AII e ODII.*

valor associado à aresta é o DTX1-NCSTN (o peso é 0,999 no grafo de coexpressão de AII e 0,033 no grafo de coexpressão de ODII).

5.3.2 Propriedades da rede

Dado um conjunto de genes, as medidas de propriedades estruturais de grafos de coexpressão com peso nas arestas disponíveis no CoGA para cada condição biológica são (i) entropia espectral, (ii) a média entre as centralidades de grau dos vértices, (iii) a média entre as centralidades de autovetor dos vértices e (iv) as médias entre os coeficientes de *clustering* dos vértices. Por exemplo, o conjunto RANTSN apresentou entropia espectral de 0,751 em AII e de 0,443 em ODII e o p-valor do teste da diferença entre as entropias foi de 0,0024. Na Tabela 5.2, mostramos as médias e os intervalos de confiança de 95%, calculados com 1.000 reamostragens *bootstrap*, para cada medida de centralidade nos grafos de AII e ODII. Podemos notar que não houve intersecção entre os intervalos de confiança de AII e ODII para nenhuma das medidas consideradas.

5.3.3 Ranking dos genes

O CoGA fornece uma lista dos genes pertencentes a um dado conjunto e permite que eles sejam ordenados de acordo com medidas da "importância" dos genes na rede, como as centralidades de grau e de autovetor e o coeficiente de *clustering*, que estão disponíveis para redes com peso nas arestas (para grafos sem peso nas arestas, as centralidades de *betweenness* e de proximidade também estão disponíveis).

Tabela 5.2: *Médias e intervalos de confiança (IC) de 95% das médias da centralidade de grau (CG), da centralidade de autovetor e do coeficiente de* clustering*, calculados para o grafo de coexpressão do conjunto REACTOME ACTIVATED NOTCH1 TRANSMITS SIGNALS TO THE NUCLEUS, em astrocitoma grau II (AII) e oligodendroglioma grau II (ODII).*

Medida	Média (IC)	
	AII	ODII
Centralidade de grau	16,895 (15,987, 17,728)	8,19 (7,709, 8,733)
Centralidade de autovetor	0,849 (0,804, 0,888)	0,739 (0,691, 0,789)
Coeficiente de *clustering*	0,676 (0,673, 0.679)	0.328 (0.326, 0.329)

No conjunto RANTSN, o gene com maior centralidade de grau no grafo de coexpresão do AII é o DTX1 (grau na rede do AII = 19,761, grau na rede do ODII = 7,064), que é um regulador da via de sinalização do Notch. Esse gene também apresentou a maior diferença de grau entre os dois grafos de coexpressão (diferença de 12,697). Interessantemente, a expressão do gene DTX1 tem correlação com a sobrevida de pacientes em gliomas e a superexpressão desse gene pode aumentar a migração e invasão celular em glioblastoma multiforme, que é o glioma de maior grau de agressividade (Huber *et al.*, 2013). O DTX1 também pode induzir vias a protegerem as células tumorais da apoptose e estimular a proliferação celular (Huber *et al.*, 2013). Assim, o gene DTX1 está associado à agressividade da célula tumoral.

No grafo de coexpressão do oligodendroglioma grau II, o gene com maior centralidade de grau é o DLL1 (grau na rede do ODII = 10,862, grau na rede do AII = 17,128), que atua como um ligante para os receptores do Notch. As centralidades de grau dos demais genes do conjunto RANTSN estão disponíveis no Apêndice F.

5.3.4 Análise de expressão genética

O pacote também inclui a análise clássica de expressão diferencial, em que compara-se a expressão média/mediana de cada gene. A ferramenta mostra a matriz da expressão genética de um conjunto de genes com cores representando os níveis de expressão (Figura 5.4) e o resultado do teste t (diferença entre médias) e do teste de Wilcoxon-Mann-Whitney (differença entre medianas). Para o conjunto RANTSN, apenas o gene ARRB2 apresentou p-valor nominal do teste t menor do que 5% e apenas o ARRB2 e o DNER tiveram p-valor menor do que 5% no teste de Wilcoxon-Mann-Whitney. Os resultados dos testes para os demais genes do conjunto RANTSN podem ser encontrados no Apêndice F.

5.4 Conclusões

O CoGA é um pacote do R com interface gráfica que inclui as seguintes funcionalidades (i) testes estatísticos entre grafos de coexpressão, (ii) visualização das redes, (iii) propriedades estruturais das redes, (iv) ranking dos genes e (v) análise da diferença de expressão genética.

As ferramentas (i-iv) auxiliam na identificação de pares e conjuntos de genes diferenci-

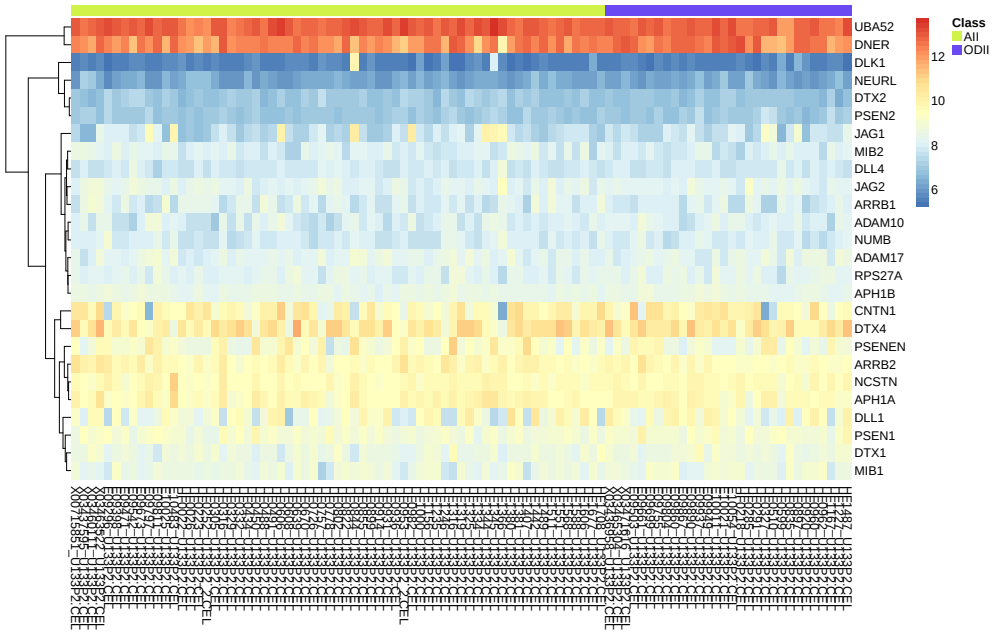


Figura 5.4: *Matriz de expressão genética do conjunto de genes REACTOME ACTIVATED NOTCH1 TRANSMITS SIGNAL TO THE NUCLEUS em astrocitoma grade II (verde) e oligodendroglioma grade II (azul). As cores vermelha e azul na matriz de expressão genética representam, respectivamente, os maiores e menores níveis de expressão. A cor amarela representa níveis intermediários.*

almente coexpressos, enquanto a funcionalidade (v) permite que o usuário teste se os genes de um dado conjunto têm a mesma expressão média/mediana em duas condições biológicas. É importante notar que a coexpressão pode mudar sem que a expressão média de um gene mude significativamente.

No exemplo dado na Seção 5.3, nossos testes estatísticos mostraram diferenças do grafo de coexpressão do conjunto REACTOME ACTIVATED NOTCH1 TRANSMITS SIGNAL TO THE NUCLEUS (RANTSN) entre astrocitoma grau II (AII) e oligodendroglioma grau II (ODII). O gene DTX1, que faz parte desse conjunto, não apresentou diferença significativa de expressão média (p-valor = 0,06) nem de expressão mediana (p-valor = 0,057), mas teve a maior diferença de grau no grafo de coexpressão do RANTSN e faz parte da aresta com maior diferença de coexpressão.

Esse tipo de mudança das atividades genéticas, que não altera a expressão genética média, pode ser ilustrada em um experimento de simulação. Considere dois genes X e Y. Em uma condição A, as atividades dos genes X e Y são representadas por duas variáveis aleatórias independentes, respectivamente, X_A e Y_A. Cada uma dessas variáveis segue uma distribuição normal com média $\mu = 0$ e desvio padrão $\sigma = 1$. Na condição B, nós representamos os níveis de expressão de X e Y pelas variáveis aleatórias X_B e Y_B, que seguem uma distribuição normal bivariada com média $\mu = (0, 0)$, e matriz de covariância

$$\sigma = \begin{pmatrix} 1 & 0,8 \\ 0,8 & 1 \end{pmatrix}.$$

A partir de um experimento realizado no ambiante R, em que são sorteadas 40 observações conjuntas dos níveis de expressão de X e Y para cada uma das condições, realizamos o teste t para testar a igualdade entre as médias dos níveis de expressão e o teste de Pearson para testar se a correlação entre as variáveis é zero. O p-valor do teste t entre A e B foi de 0,714 para o gene X e 0,464 para o gene Y, indicando assim que não há diferença significativa entre as médias, conforme o esperado pelo procedimento utilizado para gerar as amostras. O teste de independência de Pearson entre os níveis de expressão dos genes X e Y teve um p-valor alto (0,619) na condição A e baixo ($1,415 \times 10^{-4}$) na condição B, sugerindo, assim, que as atividades genéticas não estão correlacionadas em A (coeficiente de correlação de Pearson = 0,081), mas associadas em B (coeficiente de correlação de Pearson = 0,64). Assim, a coexpressão de um gene pode mudar mesmo quando a expressão média dos genes não muda significativamente.

Alterações das atividades genéticas que não provocam mudanças na expressão média também podem ser encontradas na literatura (Chan *et al.*, 2000; de la Fuente, 2010; Hudson *et al.*, 2009; Kato *et al.*, 2003; Keller *et al.*, 2008). Por exemplo, no estudo de Hudson *et al.* (2009), dois grupos de bois são comparados, um com e outro sem uma mutação conhecida no regulador transcricional miostatina. De acordo com uma medida de diferença de coexpressão, o gene miostatina apresentou a maior discrepância de coexpressão dentre os outros 920 reguladores transcricionais analisados. Já a expressão média do mesmo gene não apresentou diferenças significativas entre os dois grupos de bois. Note que esperamos que haja diferença nas atividades do gene miostatina, uma vez que apenas um dos grupos apresenta uma mutação nesse gene. Assim, a medida baseada em coexpressão conseguiu detectar mudanças nas atividades genéticas que não foram identificadas pela expressão média de um gene. Outros exemplos e mais discussões sobre esse assunto podem ser consultados na revisão de de la Fuente (2010).

Neste capítulo, nós vimos que as funcionalidades do CoGA podem identificar genes relacionados a uma dada condição biológica e alterações das atividades genéticas além daquelas identificadas pelas análises clássicas de expressão diferencial.

Capítulo 6

Considerações finais

Neste trabalho, estudamos métodos para a construção (Parte I) e a análise diferencial (Parte II) de grafos de coexpressão a fim de contribuir com pesquisas sobre doenças complexas. Os resultados obtidos neste trabalho foram disponibilizados de duas formas: (i) artigo publicado em revista científica de Bioinformática e (ii) software livre para realizar as análises propostas.

Na Parte I, realizamos um estudo comparativo de medidas utilizadas para identificar dependência estatística entre os níveis de expressão dos genes. Os resultados obtidos foram publicados na revista *Briefings in Bioinformatics* com o título *A comparative study of statistical methods used to identify dependencies between gene expression signals* (Santos *et al.*, 2014). A partir de experimentos de simulação e de uma aplicação dos métodos em dados de expressão genética de tecidos tumorais, este trabalho avaliou diferentes características das medidas, como o poder estatístico e o tipo de associação detectada por estas. Dessa forma, o estudo apresentado pode facilitar a escolha do teste de independência mais adequado em um dado contexto.

Para estender o trabalho apresentado na Parte I, outros experimentos de simulação podem ser considerados de acordo com a aplicação de interesse. Este trabalho se limitou a estudar métodos para medir a coexpressão de cada par de genes. Contudo, no contexto de redes de regulação genética, análises multivariadas, que consideram três genes ou mais, também são de grande interesse.

Na Parte II, desenvolvemos um pacote do R chamado CoGA (*Co-expression Graph Analyzer*) para a análise diferencial de grafos de coexpressão. A ferramenta está disponível na página www.ime.usp.br/~suzana/coga e é descrita por um manuscrito que foi submetido para uma revista científica.

A escolha do R se deve ao fato de ele ser um dos ambientes e linguagens de programação mais utilizados em estatística, além de incluir implementações de métodos estatísticos já bem estabelecidos. Como as análises feitas no R exigem um certo grau de conhecimento em programação, muitos pacotes são pouco acessíveis aos pesquisadores que não têm esse conhecimento específico. A fim de disponibilizar um pacote mais acessível, desenvolvemos

uma interface gráfica para o CoGA, a partir da qual o usuário pode fazer testes estatísticos entre grafos e analisar conjuntos de genes.

Os testes estatísticos para comparar grafos de coexpressão implementados pelo programa podem ser úteis na identificação de conjuntos de genes envolvidos em uma doença. Conforme verificamos por experimentos de simulação, os testes controlam a taxa de falsos positivos e apresentam poder estatístico proporcional ao número de mudanças no grafo. As funcionalidades para a visualização das redes e do ranking dos genes de um conjunto são úteis para identificar os genes e as mudanças mais "importantes" no grafo de coexpressão.

Uma das limitações dos métodos apresentados para a análise é que a direção das arestas dos grafos não são consideradas, sendo que existe direcionalidade na regulação genética. Algumas medidas, como as medidas de centralidade, podem ser naturalmente estendidas para grafos com direção nas arestas. Já os métodos baseados no espectro dos grafos envolvem mais estudo para serem aplicados nesse contexto, uma vez que a matriz de adjacência do grafo dirigido, em geral, não é simétrica e, portanto, os autovalores não serão necessariamente valores reais.

Dessa forma, este trabalho dá apoio às pesquisas que estudam coexpressão de genes e, em particular, mudanças de coexpressão em doenças. Além disso, foram agregados métodos estatísticos à análise de grafos, que são estruturas bastante estudadas em computação. Acredita-se ser uma tendência na área da computação utilizar grafos para modelar sistemas reais a partir de amostras de uma população. Esse tipo de aplicação envolve, naturalmente, métodos estatísticos, sendo um indício de que há uma grande demanda pelo estudo de grafos do ponto de vista estatístico.

Apêndice A

RMA (*Robust Multi-array Average*)

Para obter os sinais de expressão a partir de experimentos de microarranjos de DNA é preciso considerar que diversos fatores podem afetar o tratamento das amostras biológicas e a medição das intensidades a partir da imagem resultante do experimento.

Neste trabalho, utilizamos o método RMA (*Robust Multi-array Average*) (Irizarry *et al.*, 2003b) para pré-processar dados de microarranjos de DNA da plataforma Affymetrix. Nossa escolha se deve ao fato de o RMA ser um dos métodos mais utilizados para analisar dados da plataforma Affymetrix (McCall *et al.*, 2010), além de apresentar alta especificidade e sensibilidade quando comparado com outros métodos (Irizarry *et al.*, 2003a). Os passos executados pelo RMA podem ser divididos em (i) correção do ruído do fundo, (ii) normalização e (iii) sumarização das sondas que representam um único transcrito.

Correção do ruído do fundo

Os valores brutos medidos a partir das imagens geradas como resultado de experimentos de microarranjos de DNA podem sofrer influências indesejados, como, por exemplo, o ruído da autofluorescência do própria superfície do arranjo. Os métodos de correção do ruído do fundo estimam a porção do sinal que veio do fundo para subtraí-lo do valor da intensidade. Como não é possível estimar área entre as posições do microarranjo da Affymetrix, já que os arranjos são tão densos que praticamente não há espaço separando duas sondas, é preciso estimar o fundo a partir das próprias intensidades obtidas para cada sonda.

O método do RMA para correção do fundo se baseia na hipótese de que a intensidade observada de uma sonda, denotada por O, é obtida pela combinação de uma componente do fundo, N, que segue distribuição normal, e de uma componente do sinal "verdadeiro" da sonda, S, com distribuição exponencial, tal que

$$O = N + S,$$

onde $N \sim N(\mu, \sigma^2)$ e $S \sim exp(\alpha)$.

O RMA estima os parâmetros α, μ e σ a partir dos dados. Com esse modelo, as in-

tensidades observadas podem ser ajustadas substituindo-as pelo sinal "verdadeiro" esperado $E(s|O = o)$, definido como

$$E(s|O = o) = a + b\frac{\phi(\frac{a}{\sigma}) - \phi(\frac{o-a}{\sigma})}{\Phi(\frac{a}{\sigma}) + \Phi(\frac{o-a}{\sigma}) - 1},$$

onde $a = o - \mu - \sigma^2\alpha$ e ϕ e Φ são as funções de densidade de probabilidade e de distribuição acumulada, respectivamente (Irizarry *et al.*, 2003b).

Normalização

Cada microarranjo de DNA provoca um viés nas intensidades obtidas pelo experimento. Assim, a fim de diminuir os efeitos das variações indesejadas nas análises e tornar as intensidades observadas "comparáveis" entre diferentes microarranjos, são necessários métodos de normalização. Em particular, o RMA aplica a normalização quantílica. Esse método de normalização assume que as intensidades "verdadeiras" em cada microarranjo são oriundas da mesma distribuição. Assim, para reduzir os vieses dos experimentos, a normalização quantílica transforma os dados de forma que cada quantil é o mesmo em todos os microarranjos (Bolstad *et al.*, 2003). A normalização de N microarranjos pode ser descrita pelos seguintes passos:

1. Ordene as intensidades $X_j = \{x_1^{(j)}, \ldots, x_n^{(j)}\}$, para cada microarranjo j contendo n valores de intensidade.
2. Para cada quantil i, calcule a média $m_i = \frac{1}{N}\sum_{j=1}^{N} x_i^{(j)}$.
3. Substitua $x_i^{(j)}$ por m_i em cada microarranjo j.
4. Recupere a ordem original de X_j para cada microarranjo j.

Sumarização

O método utilizado pelo RMA para sumarizar as intensidades das sondas de um único transcrito é o *Medianpolish*, que é baseado no seguinte modelo:

$$\log_2(y_{ij}) = \alpha_i + \mu_j + \epsilon_{ij},$$

onde y_{ij} é a intensidade observada da sonda i no microarranjo j, α_i é o efeito da afinidade da sonda i, $\sum_{i=0}^{n} \alpha = 0$, μ_j representa o nível de expressão do microarranjo j e ϵ_{ij} é um erro independente e identicamente distribuído com média 0 (Irizarry *et al.*, 2003b).

O estimador $\hat{\mu}_j$ é o valor de expressão sumarizado que queremos obter. Os parâmetros do modelo são estimados por meio de um procedimento robusto que iterativamente estima a

matriz de erros $\hat{\epsilon}_{ij}$ por meios de subtrações das medianas das linhas e das colunas de forma alternada, até o alcançar a convergência.

Apêndice B

Cenários simulados

Cenários com diferentes tipos de dependência

Cenários com amostras de tamanhos 10, 30 e 50

(a) Independente:
x_i varia de -3 a 3, em intervalos iguais
$y_i = \varepsilon_i$, onde $\varepsilon \sim N(0,1)$

(b) Linear:
x_i varia de $-1,5$ a $2,5$, em intervalos iguais
$y_i = 0,5x_i + \varepsilon_i$, onde $\varepsilon \sim N(0,1)$

(c) Exponencial:
x_i varia de -30 a 20, em intervalos iguais
$y_i = 0,01e^{x_i} + \varepsilon_i$, onde $\varepsilon \sim N(0,1)$

(d) Quadrática:
x_i varia de -3 a 3, em intervalos iguais
$y_i = x_i^2 + \varepsilon_i$, onde $\varepsilon \sim N(0,1)$

(e) Senóide:
x_i varia de 0 a 10, em intervalos iguais
$y_i = 2seno(x_i) + \varepsilon_i$, onde $\varepsilon \sim U(-1,1)$

(f) Circunferência:
x_i varia de -5 a 5, e de 5 a -5, em intervalos iguais
$y_i = \sqrt{25 - x_i^2} + \varepsilon_i$, para i de 1 a $\frac{n}{2}$
$y_i = -\sqrt{25 - x_i^2} + \varepsilon_i$, para i de $\frac{n}{2} + 1$ a n, onde $\varepsilon \sim N(0,1)$

(g) X:
x_i varia de -5 a 5, e de 5 a -5, em intervalos iguais
$y_i = x_i + \varepsilon_i$, para i de 1 a $\frac{n}{2}$
$y_i = -x_i + \varepsilon_i$, para i de $\frac{n}{2} + 1$ a n, onde $\varepsilon \sim U(-1,1)$

Cenários com amostras de tamanho 40 e 140

(h) Quadrado:
x_i varia de 6 a 9, em intervalos iguais, para i de 1 a $\frac{n}{4}$
x_i vale 9, para i de $\frac{n}{4}+1$ a $\frac{2n}{4}$
x_i varia de 9 a 6, em intervalos iguais, para i de $\frac{2n}{4}+1$ a $\frac{3n}{4}$
x_i vale 6, para i de $\frac{3n}{4}+1$ a n
$y_i = 6 + \varepsilon_i$, para i de 1 a $\frac{n}{4}$
$y_i = w_i + \varepsilon_i$, e w varia de 6 a 9, em intervalos iguais, para i de $\frac{n}{4}+1$ a $\frac{2n}{4}$
$y_i = 9 + \varepsilon_i$, para i de $\frac{2n}{4}+1$ a $\frac{3n}{4}$
$y_i = w_i + \varepsilon_i$, e w varia de 9 a 6, em intervalos iguais, para i de $\frac{3n}{4}+1$ a n
onde $\varepsilon \sim U(-1, 1)$

Cenários com amostras de tamanho 100

(i) Correlação local:
x_i varia de 3 a 6, em intervalos iguais
$y_i = \varepsilon_i$, para i de 1 a 40 e 61 a 100
$y_i = x_i + \varepsilon_i$, para i de 41 a 60
onde $\varepsilon \sim N(0, 1)$

Cenários com introdução de *outliers*

Cenários com amostras de tamanhos 10, 30 e 50

(a) Independente com *outliers*:
x_i varia de -3 a 3, em intervalos iguais
$y_i = \varepsilon_i$, onde $\varepsilon \sim N(0, 1)$
Foram introduzidos *outliers* em 7% da amostra, nas posições finais: $y_i \sim N(0, 100)$

(b) Linear com *outliers*:
x_i varia de -3 a 3, em intervalos iguais
$y_i = 0,5x_i + \varepsilon_i$, onde ε_i é uma observação de $\varepsilon \sim N(0, 1)$
Foram introduzidos *outliers* em 7% da amostra, nas posições finais: $y_i \sim N(0, 100)$

Apêndice C

Via do WNT5A

AES	DVL2	GSK3B	TCF7
APC	EP300	JUN	TCF7L1
AXIN1	FBXW11	KREMEN1	TLE1
BCL9	FBXW2	LEF1	TLE2
BTRC	FBXW4	LRP5	WIF1
CCND1	FGF4	LRP6	WISP1
CCND2	FOSL1	MYC	WNT1
CCND3	FOXN1	NKD1	WNT10A
CSNK1A1	FRAT1	NLK	WNT11
CSNK1D	FRZB	PITX2	WNT16
CSNK1G1	FSHB	PORCN	WNT2
CSNK2A1	FZD1	PPP2CA	WNT2B
CTBP1	FZD2	PPP2R1A	WNT3
CTBP2	FZD3	RHOU	WNT4
CTNNB1	FZD4	SENP2	WNT5A
CTNNBIP1	FZD5	SFRP1	WNT5B
CXXC4	FZD6	SFRP4	WNT6
DAAM1	FZD7	SLC9A3R1	WNT7A
DIXDC1	FZD8	SOX17	WNT7B
DKK1	GSK3A	T	WNT8A
DVL1			

Apêndice D

Testes entre grafos de coexpressão de astrocitoma grau II e oligodendroglioma grau II

Conjunto	DE	EE	DG	CG	CB	CC	CA	CoC
REACTOME ACTIVATED NOTCH1 TRANSMITS SIGNAL TO THE NUCLEUS	0,002	0,002	0,022	0,008	0,966	0,004	0,289	0,008
REACTOME COMPLEMENT CASCADE	0,002	0,001	0,001	0,001	0,955	0,017	0,081	0,001
REACTOME NUCLEAR RECEPTOR TRANSCRIPTION PATHWAY	0,004	0,004	0,003	0,004	0,631	0,028	0,103	0,003
PID EPOPATHWAY	0,004	0,005	0,007	0,001	0,998	0,002	0,010	0,001
PID INTEGRIN CS PATHWAY	0,005	0,008	0,001	0,003	0,890	0,013	0,101	0,002
REACTOME ION TRANSPORT BY P TYPE ATPASES	0,006	0,004	0,003	0,002	0,975	0,000	0,580	0,002
REACTOME AMINO ACID AND OLIGOPEPTIDE SLC TRANSPORTERS	0,006	0,007	0,017	0,010	0,987	0,646	0,628	0,011
REACTOME INNATE IMMUNE SYSTEM	0,008	0,008	0,010	0,009	0,038	0,043	0,663	0,010
REACTOME ION CHANNEL TRANSPORT	0,008	0,009	0,004	0,006	0,127	0,002	0,063	0,003
KEGG SYSTEMIC LUPUS ERYTHEMATOSUS	0,009	0,010	0,009	0,008	0,014	0,019	0,283	0,008
REACTOME GROWTH HORMONE RECEPTOR SIGNALING	0,010	0,010	0,013	0,020	0,722	0,113	0,676	0,018
KEGG ALLOGRAFT REJECTION	0,011	0,026	0,006	0,008	0,159	0,040	0,237	0,007
PID ATF2 PATHWAY	0,011	0,012	0,017	0,019	0,972	0,067	0,637	0,017
PID CONE PATHWAY	0,012	0,019	0,008	0,006	0,790	0,015	0,009	0,006
REACTOME DEGRADATION OF THE EXTRACELLULAR MATRIX	0,013	0,012	0,019	0,011	0,837	0,022	0,036	0,010
REACTOME AMINO ACID TRANSPORT ACROSS THE PLASMA MEMBRANE	0,013	0,011	0,020	0,028	0,669	0,102	0,123	0,026
KEGG COMPLEMENT AND COAGULATION CASCADES	0,013	0,015	0,007	0,009	0,427	0,009	0,341	0,008
REACTOME TRANSPORT TO THE GOLGI AND SUBSEQUENT MODIFICATION	0,015	0,017	0,029	0,015	0,543	0,006	0,062	0,017
REACTOME APOPTOTIC CLEAVAGE OF CELLULAR PROTEINS	0,015	0,023	0,006	0,026	0,481	0,013	0,639	0,022
REACTOME KERATAN SULFATE KERATIN METABOLISM	0,015	0,017	0,016	0,013	0,791	0,009	0,651	0,011
KEGG LEISHMANIA INFECTION	0,015	0,015	0,013	0,011	0,049	0,084	0,044	0,014
REACTOME IMMUNOREGULATORY INTERACTIONS BETWEEN A LYMPHOID AND A NON LYMPHOID CELL	0,015	0,016	0,004	0,019	0,432	0,321	0,334	0,011
PID AMB2 NEUTROPHILS PATHWAY	0,015	0,015	0,035	0,009	0,631	0,023	0,080	0,010
PID TCR PATHWAY	0,016	0,016	0,010	0,025	0,099	0,816	0,451	0,019
KEGG INTESTINAL IMMUNE NETWORK FOR IGA PRODUCTION	0,016	0,026	0,032	0,014	0,205	0,022	0,007	0,014
PID GMCSF PATHWAY	0,017	0,018	0,009	0,016	0,186	0,071	0,295	0,013
PID PS1PATHWAY	0,017	0,020	0,095	0,033	0,891	0,036	0,040	0,038
REACTOME O LINKED GLYCOSYLATION OF MUCINS	0,017	0,018	0,011	0,009	0,615	0,012	0,125	0,011
REACTOME EXTRACELLULAR MATRIX ORGANIZATION	0,019	0,018	0,022	0,017	0,121	0,241	0,056	0,019
REACTOME NEUROTRANSMITTER RELEASE CYCLE	0,019	0,022	0,033	0,043	0,472	0,790	0,279	0,039
PID CD8TCRPATHWAY	0,019	0,020	0,020	0,032	0,287	0,053	0,454	0,027
REACTOME FGFR LIGAND BINDING AND ACTIVATION	0,020	0,029	0,020	0,023	0,888	0,353	0,156	0,017
REACTOME NUCLEAR SIGNALING BY ERBB4	0,020	0,018	0,023	0,015	0,805	0,058	0,114	0,013
REACTOME KERATAN SULFATE BIOSYNTHESIS	0,021	0,028	0,033	0,027	0,739	0,066	0,699	0,025
REACTOME G ALPHA S SIGNALLING EVENTS	0,021	0,020	0,028	0,023	0,183	0,028	0,170	0,023
KEGG TYPE I DIABETES MELLITUS	0,022	0,022	0,010	0,027	0,026	0,300	0,385	0,019
BIOCARTA INTRINSIC PATHWAY	0,023	0,022	0,021	0,020	0,948	0,466	0,409	0,019
PID ALK1PATHWAY	0,023	0,026	0,019	0,010	0,988	0,017	0,029	0,011
KEGG JAK STAT SIGNALING PATHWAY	0,023	0,022	0,034	0,026	0,012	0,887	0,113	0,026
REACTOME INTERFERON GAMMA SIGNALING	0,023	0,024	0,022	0,022	0,109	0,053	0,031	0,023
PID DELTANP63PATHWAY	0,024	0,023	0,014	0,035	0,952	0,160	0,132	0,032
KEGG GLYCOSPHINGOLIPID BIOSYNTHESIS LACTO AND NEOLACTO SERIES	0,025	0,036	0,011	0,014	0,998	0,001	0,237	0,013

REACTOME GLUCAGON TYPE LIGAND RECEPTORS	0,025	0,027	0,054	0,053	0,918	0,029	0,367	0,061
PID IL3 PATHWAY	0,026	0,029	0,057	0,013	0,903	0,461	0,232	0,017
REACTOME SMAD2 SMAD3 SMAD4 HETEROTRIMER REGULATES TRANSCRIPTION	0,026	0,036	0,090	0,128	0,638	0,367	0,628	0,124
KEGG TOLL LIKE RECEPTOR SIGNALING PATHWAY	0,026	0,026	0,026	0,032	0,039	0,012	0,323	0,036
REACTOME FORMATION OF FIBRIN CLOT CLOTTING CASCADE	0,027	0,027	0,017	0,024	0,996	0,442	0,224	0,020
KEGG BUTANOATE METABOLISM	0,027	0,031	0,018	0,033	0,876	0,080	0,191	0,034
BIOCARTA IL12 PATHWAY	0,028	0,026	0,062	0,090	0,990	0,618	0,531	0,105
REACTOME TRANSMEMBRANE TRANSPORT OF SMALL MOLECULES	0,029	0,029	0,026	0,032	0,077	0,887	0,806	0,030
KEGG VIRAL MYOCARDITIS	0,029	0,036	0,043	0,024	0,425	0,029	0,401	0,023
REACTOME AMINE COMPOUND SLC TRANSPORTERS	0,029	0,024	0,036	0,061	0,361	0,700	0,317	0,046
REACTOME TOLL RECEPTOR CASCADES	0,030	0,031	0,035	0,031	0,124	0,019	0,328	0,030
KEGG HISTIDINE METABOLISM	0,031	0,030	0,020	0,029	0,571	0,288	0,292	0,025
PID IL2 1PATHWAY	0,031	0,033	0,034	0,032	0,821	0,016	0,817	0,028
PID RHODOPSIN PATHWAY	0,032	0,027	0,042	0,039	0,547	0,009	0,064	0,048
KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION	0,032	0,031	0,040	0,043	0,045	0,093	0,040	0,035
REACTOME POST TRANSLATIONAL PROTEIN MODIFICATION	0,033	0,032	0,037	0,030	0,077	0,019	0,250	0,036
REACTOME POTASSIUM CHANNELS	0,033	0,035	0,028	0,035	0,200	0,800	0,351	0,038
REACTOME SHC MEDIATED CASCADE	0,033	0,032	0,051	0,040	0,689	0,351	0,313	0,036
PID IL1PATHWAY	0,034	0,029	0,060	0,056	0,797	0,060	0,621	0,057
REACTOME TRANSPORT OF GLUCOSE AND OTHER SUGARS BILE SALTS AND ORGANIC ACIDS METAL IONS AND AMINE COMPOUNDS	0,034	0,034	0,017	0,038	0,169	0,020	0,589	0,031
PID IL12 2PATHWAY	0,035	0,034	0,024	0,039	0,138	0,059	0,371	0,034
KEGG O GLYCAN BIOSYNTHESIS	0,035	0,051	0,058	0,033	0,810	0,709	0,049	0,044
REACTOME TRANSPORT OF INORGANIC CATIONS ANIONS AND AMINO ACIDS OLIGOPEPTIDES	0,035	0,035	0,071	0,029	0,344	0,135	0,371	0,031
REACTOME FANCONI ANEMIA PATHWAY	0,036	0,030	0,046	0,015	0,951	0,022	0,063	0,016
REACTOME NEGATIVE REGULATION OF FGFR SIGNALING	0,036	0,038	0,029	0,026	0,459	0,069	0,071	0,023
PID CXCR4 PATHWAY	0,036	0,037	0,046	0,044	0,068	0,127	0,402	0,049
REACTOME AMYLOIDS	0,037	0,039	0,013	0,028	0,991	0,005	0,633	0,027
REACTOME INSULIN SYNTHESIS AND PROCESSING	0,037	0,020	0,160	0,065	0,747	0,035	0,099	0,083
KEGG LYSOSOME	0,038	0,037	0,042	0,042	0,040	0,145	0,076	0,036
BIOCARTA TPO PATHWAY	0,038	0,035	0,016	0,031	0,796	0,851	0,909	0,026
KEGG HEMATOPOIETIC CELL LINEAGE	0,038	0,040	0,052	0,041	0,015	0,346	0,056	0,038
KEGG MAPK SIGNALING PATHWAY	0,039	0,039	0,036	0,033	0,074	0,118	0,609	0,036
REACTOME NCAM SIGNALING FOR NEURITE OUT GROWTH	0,039	0,040	0,055	0,057	0,976	0,027	0,372	0,050
REACTOME POST TRANSLATIONAL MODIFICATION SYNTHESIS OF GPI ANCHORED PROTEINS	0,040	0,034	0,033	0,049	0,936	0,235	0,075	0,051
REACTOME REGULATION OF BETA CELL DEVELOPMENT	0,041	0,040	0,142	0,072	0,924	0,604	0,024	0,093
KEGG GNRH SIGNALING PATHWAY	0,041	0,040	0,039	0,060	0,748	0,044	0,852	0,056
REACTOME PHOSPHOLIPASE C MEDIATED CASCADE	0,041	0,042	0,043	0,056	0,138	0,114	0,639	0,050
KEGG NATURAL KILLER CELL MEDIATED CYTOTOXICITY	0,041	0,043	0,044	0,057	0,262	0,592	0,606	0,054
PID WNT SIGNALING PATHWAY	0,042	0,048	0,036	0,095	0,803	0,035	0,541	0,095
REACTOME STRIATED MUSCLE CONTRACTION	0,042	0,038	0,046	0,024	0,290	0,091	0,254	0,026
BIOCARTA IL6 PATHWAY	0,043	0,046	0,016	0,034	0,616	0,089	0,582	0,030
KEGG LEUKOCYTE TRANSENDOTHELIAL MIGRATION	0,043	0,042	0,053	0,047	0,105	0,181	0,675	0,050
ST MYOCYTE AD PATHWAY	0,043	0,039	0,058	0,104	0,470	0,007	0,382	0,094
PID INTEGRIN2 PATHWAY	0,044	0,031	0,097	0,078	0,471	0,195	0,344	0,068
KEGG MATURITY ONSET DIABETES OF THE YOUNG	0,044	0,047	0,088	0,040	0,937	0,561	0,018	0,046
PID ARF6 TRAFFICKINGPATHWAY	0,045	0,047	0,029	0,043	0,898	0,642	0,790	0,042
REACTOME SLC MEDIATED TRANSMEMBRANE TRANSPORT	0,046	0,045	0,043	0,049	0,065	0,869	0,905	0,048
REACTOME VOLTAGE GATED POTASSIUM CHANNELS	0,047	0,046	0,042	0,059	0,085	0,310	0,153	0,059
REACTOME SIGNALING BY ILS	0,048	0,047	0,032	0,077	0,142	0,061	0,928	0,068
KEGG OOCYTE MEIOSIS	0,048	0,048	0,031	0,053	0,218	0,066	0,907	0,047
BIOCARTA AMI PATHWAY	0,048	0,044	0,057	0,043	0,946	0,671	0,198	0,060
REACTOME TRIF MEDIATED TLR3 SIGNALING	0,049	0,048	0,064	0,070	0,801	0,009	0,884	0,071
REACTOME CELL SURFACE INTERACTIONS AT THE VASCULAR WALL	0,049	0,050	0,057	0,033	0,118	0,085	0,166	0,035
REACTOME PTM GAMMA CARBOXYLATION HYPUSINE FORMATION AND ARYLSULFATASE ACTIVATION	0,049	0,046	0,070	0,036	0,992	0,141	0,171	0,045
REACTOME INTEGRIN CELL SURFACE INTERACTIONS	0,049	0,051	0,032	0,020	0,049	0,788	0,008	0,022
REACTOME CYTOKINE SIGNALING IN IMMUNE SYSTEM	0,049	0,054	0,049	0,081	0,110	0,160	0,906	0,076
PID IL6 7PATHWAY	0,051	0,064	0,055	0,039	0,050	0,277	0,230	0,053
KEGG PROTEIN EXPORT	0,052	0,057	0,084	0,069	0,940	0,559	0,085	0,065
KEGG GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS	0,052	0,065	0,055	0,071	0,842	0,072	0,217	0,075
PID TRKRPATHWAY	0,053	0,052	0,016	0,049	0,175	0,806	0,632	0,038
REACTOME FRS2 MEDIATED CASCADE	0,054	0,053	0,074	0,065	0,213	0,290	0,335	0,061
REACTOME APOPTOTIC EXECUTION PHASE	0,055	0,062	0,031	0,089	0,699	0,364	0,869	0,073
KEGG REGULATION OF ACTIN CYTOSKELETON	0,055	0,054	0,057	0,052	0,205	0,054	0,448	0,050
PID FCER1PATHWAY	0,056	0,058	0,046	0,040	0,361	0,088	0,076	0,044
KEGG FC EPSILON RI SIGNALING PATHWAY	0,057	0,058	0,055	0,067	0,007	0,364	0,461	0,060
REACTOME BILE ACID AND BILE SALT METABOLISM	0,058	0,082	0,053	0,065	0,916	0,334	0,554	0,060

REACTOME ANTIGEN ACTIVATES B CELL RECEPTOR LEADING TO GENERATION OF SECOND MESSENGERS	0,058	0,051	0,068	0,089	0,836	0,697	0,511	0,074
REACTOME GASTRIN CREB SIGNALLING PATHWAY VIA PKC AND MAPK	0,058	0,058	0,070	0,054	0,017	0,953	0,582	0,053
KEGG AUTOIMMUNE THYROID DISEASE	0,058	0,089	0,044	0,070	0,119	0,498	0,268	0,053
REACTOME SIGNALING BY GPCR	0,058	0,059	0,058	0,052	0,069	0,381	0,129	0,048
REACTOME SIGNALING BY FGFR1 MUTANTS	0,059	0,080	0,016	0,031	0,547	0,070	0,303	0,031
BIOCARTA NO1 PATHWAY	0,059	0,042	0,061	0,013	0,756	0,145	0,054	0,023
BIOCARTA BAD PATHWAY	0,060	0,061	0,002	0,027	0,622	0,049	0,956	0,022
BIOCARTA INFLAM PATHWAY	0,060	0,063	0,031	0,076	0,905	0,082	0,873	0,057
PID INSULIN GLUCOSE PATHWAY	0,061	0,070	0,118	0,146	0,939	0,515	0,741	0,138
KEGG NICOTINATE AND NICOTINAMIDE METABOLISM	0,062	0,059	0,105	0,133	0,441	0,328	0,581	0,154
REACTOME CLASS B 2 SECRETIN FAMILY RECEPTORS	0,062	0,061	0,051	0,062	0,030	0,056	0,436	0,063
REACTOME G ALPHA Q SIGNALLING EVENTS	0,062	0,065	0,066	0,051	0,015	0,776	0,373	0,051
BIOCARTA NOS1 PATHWAY	0,062	0,385	0,024	0,047	0,844	0,010	0,162	0,039
KEGG GLYCOSAMINOGLYCAN DEGRADATION	0,063	0,046	0,091	0,088	0,944	0,362	0,220	0,081
REACTOME A TETRASACCHARIDE LINKER SEQUENCE IS REQUIRED FOR GAG SYNTHESIS	0,063	0,068	0,043	0,075	0,815	0,448	0,651	0,061
REACTOME G ALPHA I SIGNALLING EVENTS	0,063	0,063	0,085	0,061	0,228	0,019	0,211	0,057
KEGG REGULATION OF AUTOPHAGY	0,064	0,058	0,074	0,065	0,904	0,405	0,064	0,069
PID CDC42 REG PATHWAY	0,065	0,064	0,016	0,057	0,903	0,717	0,899	0,050
KEGG ALDOSTERONE REGULATED SODIUM REABSORPTION	0,065	0,076	0,058	0,162	0,561	0,873	0,896	0,146
REACTOME IL 2 SIGNALING	0,065	0,066	0,058	0,077	0,856	0,084	0,606	0,066
REACTOME MEIOTIC RECOMBINATION	0,066	0,064	0,054	0,049	0,926	0,058	0,375	0,055
PID BCR 5PATHWAY	0,066	0,063	0,068	0,095	0,343	0,058	0,300	0,097
REACTOME INTERFERON ALPHA BETA SIGNALING	0,066	0,068	0,054	0,071	0,046	0,231	0,173	0,075
REACTOME GPCR DOWNSTREAM SIGNALING	0,067	0,070	0,072	0,059	0,088	0,702	0,073	0,056
REACTOME GPCR LIGAND BINDING	0,067	0,069	0,068	0,057	0,056	0,442	0,052	0,053
BIOCARTA GH PATHWAY	0,068	0,057	0,097	0,109	0,864	0,740	0,287	0,099
REACTOME MUSCLE CONTRACTION	0,068	0,062	0,042	0,026	0,639	0,121	0,138	0,031
KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION	0,069	0,072	0,081	0,061	0,069	0,166	0,030	0,056
REACTOME BASIGIN INTERACTIONS	0,070	0,082	0,124	0,063	0,220	0,319	0,073	0,075
BIOCARTA NTHI PATHWAY	0,070	0,064	0,042	0,060	0,867	0,071	0,294	0,072
PID IL4 2PATHWAY	0,071	0,077	0,137	0,070	0,069	0,884	0,175	0,082
SA B CELL RECEPTOR COMPLEXES	0,072	0,091	0,065	0,062	0,586	0,066	0,077	0,055
BIOCARTA IL1R PATHWAY	0,073	0,076	0,083	0,061	0,905	0,668	0,649	0,060
REACTOME OPIOID SIGNALLING	0,073	0,073	0,085	0,088	0,062	0,075	0,396	0,081
KEGG CELL ADHESION MOLECULES CAMS	0,073	0,073	0,096	0,122	0,124	0,325	0,289	0,115
PID IL23PATHWAY	0,074	0,073	0,081	0,107	0,150	0,058	0,621	0,105
PID IL27PATHWAY	0,076	0,072	0,043	0,080	0,842	0,015	0,550	0,074
PID ALPHASYNUCLEIN PATHWAY	0,077	0,068	0,084	0,128	0,476	0,110	0,809	0,113
REACTOME SIGNALLING TO RAS	0,078	0,050	0,002	0,011	0,610	0,485	0,567	0,009
BIOCARTA PTDINS PATHWAY	0,078	0,086	0,218	0,083	0,692	0,343	0,013	0,104
REACTOME NFKB AND MAP KINASES ACTIVATION MEDIATED BY TLR4 SIGNALING REPERTOIRE	0,078	0,078	0,096	0,101	0,858	0,016	0,783	0,099
REACTOME ACTIVATED TLR4 SIGNALLING	0,079	0,081	0,073	0,082	0,174	0,036	0,780	0,078
REACTOME LYSOSOME VESICLE BIOGENESIS	0,079	0,109	0,061	0,042	0,911	0,012	0,301	0,047
REACTOME METAL ION SLC TRANSPORTERS	0,080	0,057	0,157	0,123	0,356	0,161	0,545	0,120
BIOCARTA IL2 PATHWAY	0,080	0,052	0,029	0,034	0,606	0,093	0,028	0,030
REACTOME G PROTEIN ACTIVATION	0,081	0,100	0,059	0,081	0,555	0,103	0,326	0,087
PID TOLL ENDOGENOUS PATHWAY	0,081	0,381	0,076	0,068	0,658	0,044	0,112	0,068
REACTOME PEPTIDE LIGAND BINDING RECEPTORS	0,082	0,082	0,103	0,070	0,053	0,660	0,023	0,066
KEGG PRION DISEASES	0,083	0,069	0,048	0,121	0,496	0,441	0,517	0,120
PID VEGFR1 2 PATHWAY	0,084	0,088	0,120	0,079	0,968	0,039	0,316	0,093
KEGG SPHINGOLIPID METABOLISM	0,084	0,084	0,091	0,156	0,733	0,102	0,841	0,144
PID PDGFRBPATHWAY	0,085	0,085	0,054	0,093	0,132	0,233	0,880	0,090
REACTOME GABA RECEPTOR ACTIVATION	0,086	0,099	0,095	0,076	0,005	0,166	0,039	0,090
REACTOME LOSS OF NLP FROM MITOTIC CENTROSOMES	0,086	0,086	0,058	0,134	0,856	0,866	0,947	0,122
KEGG TIGHT JUNCTION	0,087	0,088	0,092	0,100	0,095	0,684	0,790	0,102
REACTOME GLYCOSPHINGOLIPID METABOLISM	0,087	0,085	0,121	0,133	0,927	0,788	0,422	0,139
PID BARD1PATHWAY	0,088	0,092	0,107	0,093	0,926	0,781	0,459	0,094
SIG IL4RECEPTOR IN B LYPHOCYTES	0,089	0,086	0,120	0,110	0,933	0,781	0,237	0,114
REACTOME HEPARAN SULFATE HEPARIN HS GAG METABOLISM	0,089	0,095	0,148	0,112	0,892	0,125	0,463	0,110
REACTOME NEURONAL SYSTEM	0,089	0,089	0,086	0,102	0,163	0,338	0,781	0,101
KEGG ANTIGEN PROCESSING AND PRESENTATION	0,089	0,111	0,078	0,086	0,226	0,125	0,672	0,069
PID ER NONGENOMIC PATHWAY	0,089	0,098	0,143	0,092	0,862	0,099	0,391	0,111
KEGG BETA ALANINE METABOLISM	0,090	0,095	0,041	0,128	0,897	0,955	0,590	0,111
KEGG CALCIUM SIGNALING PATHWAY	0,090	0,090	0,084	0,102	0,274	0,774	0,676	0,093
BIOCARTA PGC1A PATHWAY	0,091	0,237	0,100	0,078	0,643	0,019	0,136	0,081
PID NFKAPPABCANONICALPATHWAY	0,091	0,113	0,184	0,228	0,840	0,748	0,216	0,288
REACTOME IMMUNE SYSTEM	0,092	0,097	0,088	0,148	0,141	0,229	0,954	0,147
KEGG DILATED CARDIOMYOPATHY	0,092	0,094	0,060	0,080	0,004	0,349	0,482	0,079
REACTOME NITRIC OXIDE STIMULATES GUANYLATE CYCLASE	0,092	0,106	0,173	0,090	0,895	0,296	0,345	0,111
REACTOME ASPARAGINE N LINKED GLYCOSYLATION	0,092	0,094	0,108	0,080	0,220	0,065	0,169	0,098
REACTOME HEMOSTASIS	0,093	0,096	0,078	0,103	0,108	0,198	0,881	0,102

REACTOME SIGNALLING TO ERKS	0,095	0,098	0,026	0,056	0,112	0,698	0,613	0,051
REACTOME TRAF6 MEDIATED INDUCTION OF NFKB AND MAP KINASES UPON TLR7 8 OR 9 ACTIVATION	0,096	0,094	0,120	0,128	0,936	0,029	0,836	0,131
REACTOME RNA POL I PROMOTER OPENING	0,096	0,093	0,057	0,071	0,966	0,063	0,070	0,074
ST WNT CA2 CYCLIC GMP PATHWAY	0,097	0,171	0,136	0,101	0,973	0,851	0,581	0,101
BIOCARTA ERK PATHWAY	0,097	0,060	0,033	0,018	0,838	0,386	0,012	0,019
REACTOME LIPID DIGESTION MOBILIZATION AND TRANSPORT	0,097	0,090	0,115	0,171	0,763	0,100	0,762	0,180
PID TAP63PATHWAY	0,097	0,103	0,076	0,041	0,651	0,007	0,242	0,046
KEGG CHEMOKINE SIGNALING PATHWAY	0,097	0,096	0,077	0,128	0,069	0,214	0,710	0,124
KEGG ASTHMA	0,098	0,311	0,157	0,071	0,469	0,371	0,008	0,087
PID ENDOTHELINPATHWAY	0,098	0,099	0,116	0,126	0,487	0,275	0,627	0,133
BIOCARTA GSK3 PATHWAY	0,098	0,117	0,108	0,095	0,233	0,215	0,669	0,095
PID IL2 STAT5PATHWAY	0,099	0,093	0,130	0,160	0,625	0,202	0,465	0,140
PID ECADHERIN STABILIZATION PATHWAY	0,099	0,107	0,073	0,091	0,110	0,778	0,428	0,088
KEGG TYPE II DIABETES MELLITUS	0,100	0,115	0,071	0,104	0,409	0,443	0,902	0,081
PID INTEGRIN A4B1 PATHWAY	0,101	0,094	0,109	0,068	0,912	0,435	0,157	0,075
REACTOME PLC BETA MEDIATED EVENTS	0,101	0,103	0,066	0,137	0,102	0,190	0,963	0,123
REACTOME TRANSMISSION ACROSS CHEMICAL SYNAPSES	0,101	0,104	0,079	0,102	0,089	0,541	0,645	0,103
REACTOME DIABETES PATHWAYS	0,102	0,101	0,076	0,122	0,138	0,118	0,927	0,122
REACTOME SIGNALING BY FGFR MUTANTS	0,102	0,116	0,070	0,051	0,663	0,757	0,224	0,056
REACTOME AQUAPORIN MEDIATED TRANSPORT	0,103	0,104	0,133	0,097	0,138	0,183	0,384	0,120
PID P75NTRPATHWAY	0,104	0,109	0,071	0,082	0,186	0,202	0,734	0,081
REACTOME NCAM1 INTERACTIONS	0,105	0,071	0,201	0,126	0,694	0,101	0,102	0,158
REACTOME L1CAM INTERACTIONS	0,105	0,108	0,154	0,113	0,084	0,183	0,280	0,107
REACTOME PLATELET ACTIVATION SIGNALING AND AGGREGATION	0,105	0,109	0,088	0,131	0,165	0,292	0,932	0,114
REACTOME INTERFERON SIGNALING	0,107	0,114	0,126	0,139	0,072	0,462	0,617	0,137
BIOCARTA DEATH PATHWAY	0,108	0,105	0,077	0,059	0,630	0,737	0,147	0,065
REACTOME P75 NTR RECEPTOR MEDIATED SIGNALLING	0,110	0,109	0,099	0,120	0,119	0,312	0,914	0,117
KEGG MELANOGENESIS	0,110	0,111	0,080	0,125	0,469	0,436	0,821	0,123
REACTOME GLYCOSAMINOGLYCAN METABOLISM	0,111	0,113	0,084	0,110	0,977	0,044	0,776	0,107
REACTOME SHC1 EVENTS IN ERBB4 SIGNALING	0,113	0,077	0,103	0,071	0,662	0,445	0,440	0,082
REACTOME RNA POL I TRANSCRIPTION	0,113	0,120	0,128	0,068	0,776	0,097	0,088	0,076
KEGG STEROID HORMONE BIOSYNTHESIS	0,115	0,106	0,131	0,108	0,004	0,367	0,092	0,100
KEGG HYPERTROPHIC CARDIOMYOPATHY HCM	0,115	0,121	0,097	0,099	0,037	0,406	0,674	0,104
REACTOME CLASS A1 RHODOPSIN LIKE RECEPTORS	0,117	0,120	0,145	0,104	0,107	0,512	0,055	0,103
REACTOME LIGAND GATED ION CHANNEL TRANSPORT	0,118	0,487	0,162	0,131	0,020	0,255	0,259	0,106
KEGG VASCULAR SMOOTH MUSCLE CONTRACTION	0,120	0,121	0,111	0,144	0,065	0,110	0,621	0,154
KEGG ECM RECEPTOR INTERACTION	0,120	0,123	0,106	0,118	0,070	0,752	0,121	0,108
KEGG PROGESTERONE MEDIATED OOCYTE MATURATION	0,120	0,121	0,081	0,152	0,934	0,499	0,612	0,136
SIG CD40PATHWAYMAP	0,121	0,156	0,140	0,080	0,685	0,009	0,202	0,096
REACTOME LIPOPROTEIN METABOLISM	0,122	0,123	0,218	0,166	0,671	0,094	0,565	0,189
KEGG RIG I LIKE RECEPTOR SIGNALING PATHWAY	0,122	0,125	0,113	0,162	0,159	0,046	0,442	0,194
REACTOME PACKAGING OF TELOMERE ENDS	0,122	0,122	0,092	0,091	0,617	0,112	0,063	0,101
REACTOME PERK REGULATED GENE EXPRESSION	0,123	0,127	0,288	0,123	0,706	0,108	0,039	0,147
KEGG LONG TERM POTENTIATION	0,123	0,123	0,087	0,139	0,224	0,141	0,681	0,130
REACTOME CELL DEATH SIGNALLING VIA NRAGE NRIF AND NADE	0,124	0,119	0,095	0,146	0,504	0,243	0,971	0,127
PID EPHRINBREVPATHWAY	0,124	0,132	0,072	0,110	0,962	0,063	0,479	0,094
KEGG GRAFT VERSUS HOST DISEASE	0,125	0,270	0,032	0,157	0,305	0,585	0,849	0,092
BIOCARTA BCR PATHWAY	0,126	0,130	0,133	0,080	0,950	0,030	0,146	0,072
KEGG T CELL RECEPTOR SIGNALING PATHWAY	0,128	0,133	0,112	0,157	0,071	0,121	0,890	0,155
REACTOME IL 3 5 AND GM CSF SIGNALING	0,129	0,147	0,114	0,162	0,051	0,378	0,741	0,143
BIOCARTA CYTOKINE PATHWAY	0,129	0,077	0,205	0,178	0,707	0,105	0,051	0,183
REACTOME GENERATION OF SECOND MESSENGER MOLECULES	0,129	0,061	0,241	0,225	0,489	0,208	0,294	0,244
REACTOME MYD88 MAL CASCADE INITIATED ON PLASMA MEMBRANE	0,130	0,131	0,131	0,158	0,180	0,096	0,886	0,153
REACTOME GLUCAGON SIGNALING IN METABOLIC REGULATION	0,133	0,149	0,067	0,116	0,359	0,213	0,852	0,106
REACTOME REGULATION OF INSULIN SECRETION	0,134	0,137	0,102	0,141	0,055	0,130	0,950	0,135
KEGG SNARE INTERACTIONS IN VESICULAR TRANSPORT	0,134	0,128	0,098	0,228	0,480	0,122	0,995	0,199
PID P38ALPHABETADOWNSTREAMPATHWAY	0,134	0,134	0,198	0,122	0,607	0,236	0,071	0,135
KEGG SELENOAMINO ACID METABOLISM	0,134	0,142	0,013	0,036	0,832	0,545	0,403	0,031
KEGG N GLYCAN BIOSYNTHESIS	0,137	0,155	0,111	0,084	0,228	0,899	0,121	0,110
KEGG LONG TERM DEPRESSION	0,139	0,142	0,089	0,151	0,612	0,111	0,949	0,140
KEGG VIBRIO CHOLERAE INFECTION	0,140	0,151	0,076	0,187	0,140	0,316	0,992	0,159
ST ADRENERGIC	0,142	0,157	0,103	0,191	0,663	0,078	0,937	0,166
KEGG ARRHYTHMOGENIC RIGHT VENTRICULAR CARDIOMYOPATHY ARVC	0,143	0,147	0,145	0,117	0,055	0,404	0,170	0,123
REACTOME DEFENSINS	0,145	0,659	0,097	0,079	0,191	0,143	0,065	0,077
ST ERK1 ERK2 MAPK PATHWAY	0,145	0,155	0,239	0,204	0,927	0,016	0,251	0,287
PID HIVNEFPATHWAY	0,145	0,142	0,200	0,140	0,618	0,079	0,162	0,183
KEGG VALINE LEUCINE AND ISOLEUCINE DEGRADATION	0,147	0,172	0,096	0,132	0,585	0,212	0,284	0,118
REACTOME SEMA4D IN SEMAPHORIN SIGNALING	0,148	0,215	0,073	0,104	0,213	0,215	0,525	0,115
REACTOME INTERACTION BETWEEN L1 AND ANKYRINS	0,149	0,544	0,105	0,129	0,565	0,094	0,442	0,105
BIOCARTA CSK PATHWAY	0,156	0,455	0,042	0,175	0,538	0,377	0,771	0,161

PID CD8TCRDOWNSTREAMPATHWAY	0,156	0,162	0,094	0,146	0,021	0,810	0,261	0,125
KEGG PRIMARY IMMUNODEFICIENCY	0,156	0,130	0,202	0,088	0,329	0,625	0,003	0,109
BIOCARTA STRESS PATHWAY	0,159	0,156	0,251	0,208	0,813	0,016	0,223	0,253
PID SHP2 PATHWAY	0,163	0,172	0,091	0,115	0,758	0,089	0,750	0,112
REACTOME INTEGRATION OF ENERGY METABOLISM	0,167	0,168	0,152	0,193	0,023	0,299	0,950	0,193
REACTOME PI3K CASCADE	0,167	0,180	0,315	0,122	0,815	0,555	0,037	0,176
KEGG OLFACTORY TRANSDUCTION	0,169	0,161	0,172	0,140	0,020	0,643	0,107	0,159
REACTOME EFFECTS OF PIP2 HYDROLYSIS	0,170	0,167	0,101	0,088	0,893	0,645	0,547	0,099
REACTOME LATENT INFECTION OF HOMO SAPIENS WITH MYCOBACTERIUM TUBERCULOSIS	0,172	0,228	0,073	0,252	0,985	0,054	0,712	0,177
PID KITPATHWAY	0,175	0,183	0,168	0,093	0,086	0,937	0,014	0,120
PID ERBB1 RECEPTOR PROXIMAL PATHWAY	0,175	0,231	0,081	0,069	0,418	0,072	0,667	0,077
BIOCARTA NFAT PATHWAY	0,176	0,181	0,090	0,141	0,546	0,059	0,836	0,123
PID REG GR PATHWAY	0,182	0,174	0,170	0,242	0,083	0,263	0,652	0,232
PID HNF3APATHWAY	0,184	0,250	0,073	0,137	0,278	0,929	0,745	0,118
REACTOME THE ROLE OF NEF IN HIV1 REPLICATION AND DISEASE PATHOGENESIS	0,185	0,147	0,330	0,194	0,981	0,003	0,138	0,235
BIOCARTA INSULIN PATHWAY	0,187	0,158	0,030	0,104	0,827	0,496	0,065	0,094
REACTOME SIGNALING BY NOTCH	0,189	0,187	0,249	0,149	0,556	0,691	0,021	0,163
KEGG HEDGEHOG SIGNALING PATHWAY	0,191	0,203	0,171	0,154	0,051	0,235	0,291	0,162
KEGG PPAR SIGNALING PATHWAY	0,193	0,186	0,146	0,310	0,781	0,038	0,511	0,273
BIOCARTA INTEGRIN PATHWAY	0,194	0,217	0,093	0,114	0,389	0,656	0,285	0,109
REACTOME ACTIVATION OF GENES BY ATF4	0,195	0,193	0,278	0,168	0,609	0,161	0,059	0,186
REACTOME ACTIVATION OF NMDA RECEPTOR UPON GLUTAMATE BINDING AND POSTSYNAPTIC EVENTS	0,195	0,346	0,064	0,146	0,408	0,222	0,597	0,129
REACTOME INWARDLY RECTIFYING K CHANNELS	0,197	0,293	0,199	0,086	0,874	0,022	0,016	0,174
REACTOME SIGNAL TRANSDUCTION BY L1	0,199	0,209	0,271	0,245	0,966	0,096	0,286	0,251
KEGG ADHERENS JUNCTION	0,203	0,215	0,153	0,161	0,071	0,245	0,141	0,165
REACTOME PRE NOTCH EXPRESSION AND PROCESSING	0,204	0,204	0,149	0,198	0,625	0,473	0,037	0,175
BIOCARTA WNT PATHWAY	0,207	0,208	0,353	0,256	0,815	0,332	0,038	0,236
KEGG ALANINE ASPARTATE AND GLUTAMATE METABOLISM	0,208	0,185	0,201	0,182	0,683	0,098	0,332	0,175
REACTOME MAP KINASE ACTIVATION IN TLR CASCADE	0,210	0,204	0,189	0,331	0,763	0,059	0,959	0,307
BIOCARTA EGF PATHWAY	0,211	0,214	0,226	0,226	0,844	0,083	0,484	0,214
REACTOME GLUCONEOGENESIS	0,212	0,236	0,158	0,077	0,595	0,325	0,009	0,079
REACTOME INSULIN RECEPTOR SIGNALLING CASCADE	0,213	0,217	0,378	0,165	0,799	0,218	0,006	0,221
REACTOME BMAL1 CLOCK NPAS2 ACTIVATES CIRCADIAN EXPRESSION	0,222	0,226	0,164	0,162	0,962	0,052	0,048	0,163
PID MAPKTRKPATHWAY	0,224	0,239	0,120	0,237	0,042	0,230	0,794	0,217
ST WNT BETA CATENIN PATHWAY	0,228	0,235	0,243	0,334	0,237	0,036	0,905	0,316
PID RB 1PATHWAY	0,229	0,232	0,251	0,223	0,247	0,189	0,051	0,213
REACTOME MAPK TARGETS NUCLEAR EVENTS MEDIATED BY MAP KINASES	0,230	0,253	0,182	0,227	0,935	0,007	0,899	0,224
BIOCARTA PYK2 PATHWAY	0,231	0,300	0,097	0,203	0,185	0,406	0,693	0,156
KEGG TASTE TRANSDUCTION	0,236	0,198	0,224	0,226	0,041	0,308	0,152	0,262
PID RXR VDR PATHWAY	0,243	0,264	0,258	0,192	0,782	0,057	0,207	0,194
BIOCARTA MET PATHWAY	0,244	0,259	0,146	0,145	0,920	0,918	0,015	0,148
KEGG BASAL CELL CARCINOMA	0,244	0,252	0,226	0,222	0,025	0,228	0,477	0,218
REACTOME PLATELET HOMEOSTASIS	0,248	0,249	0,252	0,253	0,008	0,458	0,500	0,298
REACTOME ANTIGEN PRESENTATION FOLDING ASSEMBLY AND PEPTIDE LOADING OF CLASS I MHC	0,250	0,070	0,099	0,131	0,160	0,247	0,099	0,177
SIG REGULATION OF THE ACTIN CYTOSKELETON BY RHO GTPASES	0,253	0,246	0,221	0,399	0,005	0,280	0,966	0,347
KEGG THYROID CANCER	0,257	0,257	0,092	0,145	0,583	0,205	0,016	0,131
KEGG STARCH AND SUCROSE METABOLISM	0,259	0,254	0,193	0,276	0,047	0,441	0,387	0,244
REACTOME SIGNALING BY RHO GTPASES	0,260	0,261	0,216	0,273	0,090	0,509	0,755	0,274
ST INTEGRIN SIGNALING PATHWAY	0,262	0,277	0,182	0,222	0,018	0,669	0,500	0,218
KEGG NOTCH SIGNALING PATHWAY	0,262	0,270	0,290	0,171	0,963	0,091	0,019	0,187
BIOCARTA IGF1 PATHWAY	0,263	0,286	0,113	0,125	0,455	0,702	0,067	0,124
REACTOME INTEGRIN ALPHAIIB BETA3 SIGNALING	0,263	0,294	0,298	0,185	0,414	0,727	0,095	0,217
REACTOME DAG AND IP3 SIGNALING	0,264	0,256	0,266	0,330	0,056	0,190	0,853	0,306
REACTOME PHASE1 FUNCTIONALIZATION OF COMPOUNDS	0,268	0,256	0,336	0,276	0,094	0,528	0,217	0,276
PID SYNDECAN 1 PATHWAY	0,269	0,249	0,329	0,285	0,049	0,777	0,160	0,362
PID AP1 PATHWAY	0,280	0,271	0,253	0,356	0,799	0,067	0,646	0,341
REACTOME DEPOSITION OF NEW CENPA CONTAINING NUCLEOSOMES AT THE CENTROMERE	0,282	0,297	0,264	0,175	0,477	0,310	0,054	0,217
BIOCARTA GLEEVEC PATHWAY	0,284	0,465	0,249	0,142	0,717	0,841	0,005	0,156
KEGG GLYCOLYSIS GLUCONEOGENESIS	0,285	0,284	0,287	0,218	0,484	0,085	0,040	0,234
PID ERA GENOMIC PATHWAY	0,300	0,318	0,317	0,247	0,106	0,810	0,049	0,287
REACTOME TRANS GOLGI NETWORK VESICLE BUDDING	0,310	0,301	0,340	0,349	0,071	0,216	0,531	0,375
KEGG PYRUVATE METABOLISM	0,311	0,296	0,331	0,346	0,577	0,308	0,096	0,341
KEGG GLYCOSAMINOGLYCAN BIOSYNTHESIS CHONDROITIN SULFATE	0,313	0,302	0,390	0,327	0,206	0,477	0,065	0,403
REACTOME ASSOCIATION OF TRIC CCT WITH TARGET PROTEINS DURING BIOSYNTHESIS	0,321	0,472	0,110	0,220	0,023	0,958	0,581	0,158
BIOCARTA AT1R PATHWAY	0,321	0,341	0,197	0,279	0,690	0,094	0,516	0,289

PID PTP1BPATHWAY	0,325	0,307	0,452	0,470	0,044	0,839	0,635	0,452
REACTOME RNA POL I RNA POL III AND MITOCHONDRIAL TRANSCRIPTION	0,327	0,325	0,378	0,235	0,809	0,179	0,037	0,278
REACTOME GAP JUNCTION TRAFFICKING	0,328	0,609	0,245	0,135	0,003	0,940	0,023	0,144
PID TCPTP PATHWAY	0,331	0,323	0,317	0,340	0,396	0,294	0,096	0,394
PID S1P S1P1 PATHWAY	0,332	0,367	0,321	0,372	0,368	0,091	0,613	0,350
PID AR TF PATHWAY	0,346	0,362	0,331	0,345	0,287	0,465	0,081	0,362
REACTOME INHIBITION OF VOLTAGE GATED CA2 CHANNELS VIA GBETA GAMMA SUBUNITS	0,353	0,709	0,325	0,211	0,521	0,283	0,084	0,344
REACTOME RECYCLING PATHWAY OF L1	0,356	0,273	0,481	0,470	0,590	0,060	0,793	0,454
REACTOME INHIBITION OF INSULIN SECRETION BY ADRENALINE NORADRENALINE	0,357	0,634	0,230	0,348	0,018	0,449	0,908	0,309
PID FANCONI PATHWAY	0,358	0,360	0,413	0,368	0,095	0,511	0,602	0,369
REACTOME NEF MEDIATES DOWN MODULATION OF CELL SURFACE RECEPTORS BY RECRUITING THEM TO CLATHRIN ADAPTERS	0,358	0,201	0,486	0,426	0,914	0,009	0,364	0,452
ST GA12 PATHWAY	0,363	0,328	0,378	0,341	0,266	0,218	0,069	0,347
REACTOME GLUCOSE TRANSPORT	0,367	0,400	0,381	0,360	0,016	0,893	0,377	0,356
REACTOME RNA POL II TRANSCRIPTION PRE INITIATION AND PROMOTER OPENING	0,389	0,387	0,334	0,434	0,026	0,273	0,754	0,420
PID RETINOIC ACID PATHWAY	0,400	0,430	0,314	0,257	0,714	0,327	0,052	0,283
REACTOME UNFOLDED PROTEIN RESPONSE	0,407	0,415	0,465	0,404	0,093	0,343	0,449	0,454
KEGG COLORECTAL CANCER	0,419	0,438	0,447	0,414	0,079	0,537	0,265	0,403
KEGG AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM	0,435	0,444	0,421	0,536	0,018	0,452	0,858	0,524
PID AR PATHWAY	0,452	0,445	0,458	0,493	0,035	0,462	0,084	0,485
BIOCARTA AGR PATHWAY	0,458	0,478	0,311	0,448	0,189	0,089	0,819	0,419
PID TCRCALCIUMPATHWAY	0,492	0,372	0,522	0,716	0,876	0,049	0,794	0,678
BIOCARTA VIP PATHWAY	0,492	0,606	0,446	0,463	0,022	0,534	0,617	0,444
BIOCARTA CARM ER PATHWAY	0,504	0,465	0,523	0,435	0,697	0,693	0,028	0,461
PID P38 MKK3 6PATHWAY	0,508	0,528	0,555	0,536	0,016	0,240	0,527	0,519
KEGG TGF BETA SIGNALING PATHWAY	0,511	0,511	0,652	0,443	0,048	0,815	0,135	0,504
REACTOME SIGNALING BY NOTCH1	0,518	0,511	0,577	0,364	0,792	0,567	0,050	0,455
PID S1P S1P3 PATHWAY	0,524	0,602	0,481	0,458	0,076	0,471	0,480	0,473
KEGG RETINOL METABOLISM	0,545	0,522	0,540	0,518	0,076	0,666	0,555	0,538
PID CD40 PATHWAY	0,553	0,630	0,403	0,501	0,090	0,515	0,434	0,550
REACTOME TRANSCRIPTION	0,555	0,551	0,547	0,505	0,495	0,443	0,085	0,548
REACTOME ANTIVIRAL MECHANISM BY IFN STIMULATED GENES	0,595	0,586	0,594	0,659	0,080	0,927	0,955	0,640
PID HEDGEHOG GLIPATHWAY	0,601	0,590	0,519	0,624	0,056	0,603	0,564	0,627
REACTOME ADP SIGNALLING THROUGH P2RY12	0,612	0,980	0,483	0,500	0,014	0,501	0,477	0,512
REACTOME GLOBAL GENOMIC NER GG NER	0,647	0,627	0,691	0,820	0,093	0,655	0,990	0,723
PID REELINPATHWAY	0,648	0,647	0,696	0,741	0,083	0,572	0,668	0,674
PID PI3KCIAKTPATHWAY	0,660	0,639	0,651	0,647	0,607	0,798	0,096	0,630
KEGG VASOPRESSIN REGULATED WATER REABSORPTION	0,662	0,634	0,706	0,638	0,004	0,801	0,316	0,697
KEGG PROSTATE CANCER	0,687	0,693	0,647	0,607	0,848	0,364	0,015	0,641
REACTOME TRANSPORT OF MATURE MRNA DERIVED FROM AN INTRONLESS TRANSCRIPT	0,692	0,879	0,578	0,822	0,076	0,935	0,963	0,693
BIOCARTA RAS PATHWAY	0,705	0,919	0,537	0,456	0,870	0,912	0,049	0,485
REACTOME ADP SIGNALLING THROUGH P2RY1	0,708	0,819	0,733	0,582	0,045	0,353	0,420	0,663
REACTOME CIRCADIAN REPRESSION OF EXPRESSION BY REV ERBA	0,711	0,853	0,501	0,493	0,700	0,092	0,198	0,501
REACTOME THROMBOXANE SIGNALLING THROUGH TP RECEPTOR	0,736	0,759	0,755	0,636	0,058	0,367	0,349	0,728
REACTOME PYRUVATE METABOLISM AND CITRIC ACID TCA CYCLE	0,773	0,713	0,705	0,838	0,070	0,927	0,788	0,794
PID A6B1 A6B4 INTEGRIN PATHWAY	0,774	0,783	0,800	0,806	0,019	0,894	0,682	0,841
REACTOME PKB MEDIATED EVENTS	0,782	0,800	0,815	0,677	0,362	0,533	0,068	0,772
REACTOME NEP NS2 INTERACTS WITH THE CELLULAR EXPORT MACHINERY	0,805	0,931	0,762	0,902	0,020	0,910	0,938	0,795
PID LKB1 PATHWAY	0,840	0,893	0,866	0,642	0,219	0,213	0,038	0,747
REACTOME RORA ACTIVATES CIRCADIAN EXPRESSION	0,867	0,973	0,637	0,526	0,838	0,006	0,167	0,631
REACTOME NOTCH1 INTRACELLULAR DOMAIN REGULATES TRANSCRIPTION	0,868	0,902	0,799	0,682	0,192	0,872	0,089	0,792
REACTOME PI METABOLISM	0,883	0,868	0,913	0,832	0,090	0,926	0,232	0,910
REACTOME PI3K EVENTS IN ERBB4 SIGNALING	0,894	0,876	0,938	0,821	0,927	0,680	0,100	0,896
REACTOME REGULATION OF GLUCOKINASE BY GLUCOKINASE REGULATORY PROTEIN	0,894	0,923	0,898	0,949	0,086	0,890	0,920	0,869
REACTOME SYNTHESIS OF PIPS AT THE PLASMA MEMBRANE	0,917	0,868	0,944	0,856	0,075	0,949	0,203	0,925
REACTOME MITOCHONDRIAL TRNA AMINOACYLATION	0,971	0,978	0,980	0,954	0,086	0,791	0,767	0,945
KEGG INOSITOL PHOSPHATE METABOLISM	0,972	0,963	0,972	0,971	0,063	0,936	0,456	0,981

Apêndice E

Métodos para sumarizar as linhas da matriz de expressão genética

Quando um gene é representado por mais de uma linha na matriz de expressão genética, o pacote CoGA desenvolvido neste trabalho utiliza métodos do pacote `WGCNA` (Langfelder e Horvath , 2008) para resumir as informações da matriz de forma que cada gene seja representado por uma única linha. As funções disponíveis são:

1. `MaxMean`: escolhe a linha com a maior média entre as linhas que representam o mesmo gene.
2. `MinMean`: escolhe a linha com menor média entre as linhas que representam o mesmo gene.
3. `AbsMaxMean`: escolhe a linha com maior valor absoluto médio entre as linhas que representam o mesmo gene.
4. `AbsMinMean`: escolhe a linha com menor valor absoluto médio entre as linhas que representam o mesmo gene.
5. `MaxRowVariance`: escolhe a linha com maior variância entre as linhas que representam o mesmo gene.
6. `ME`: componente principal das linhas que representam o mesmo gene.
7. `Average`: calcula, para cada coluna, uma média dos valores das linhas que representam o mesmo gene.

Caso o usuário defina o parâmetro *Connectivity based collapsing* como `TRUE`, é adotado o seguinte procedimento: se um gene tem exatamente duas linhas correspondentes, a informação do gene é sumarizada com um dos métodos descritos acima. Se o gene é representado por três linhas ou mais, o programa calcula as correlações entre todos os pares de linhas que representam o gene e seleciona a linha mais correlacionada.

De acordo com Miller *et al.* (2011), entre os métodos apresentados, o `MaxMean` com o parâmetro *Connectivity based collapsing* definido como `FALSE`, que é a opção padrão do CoGA, mostrou-se o mais robusto na análise de expressão genética de diferentes conjuntos de dados.

Apêndice F

Análise do conjunto REACTOME ACTIVATED NOTCH1 TRANSMITS SIGNAL TO THE NUCLEUS com o pacote CoGA

As tabelas exibidas neste apêndice foram geradas a partir de análises dos dados de expressão dos genes pertencentes ao conjunto REACTOME ACTIVATED NOTCH1 TRANSMITS SIGNAL TO THE NUCLEUS utilizando o pacote CoGA, que foi desenvolvido neste trabalho. Os grafos de coexpressão construídos nestas análises têm peso nas arestas (um menos o p-valor do teste de independência de Spearman corrigido pelo método de Benjamini e Hochberg (1995)).

Pesos das arestas dos grafos de coexpressão

A tabela a seguir mostra o peso da aresta que conecta cada par de genes do conjunto REACTOME ACTIVATED NOTCH1 TRANSMITS SIGNAL TO THE NUCLEUS no grafo de coexpressão do astrocitoma grau II (AII) e no grafo do oligodendroglioma grau II (ODII). As linhas estão ordenadas pela diferença absoluta entre o peso da aresta em AII e ODII.

Gene 1	Gene 2	AII	ODII	Diferença absoluta
DTX1	NCSTN	0,999	0,033	0,966
DTX2	ARRB2	0,993	0,033	0,960
PSENEN	ADAM17	0,991	0,033	0,958
DTX1	UBA52	0,976	0,019	0,956
APH1A	NUMB	0,988	0,033	0,955
CNTN1	DTX4	0,995	0,054	0,941
ADAM10	APH1A	0,973	0,033	0,941
PSEN2	RPS27A	0,999	0,059	0,940
DTX4	MIB1	0,995	0,058	0,937
DTX2	JAG1	0,982	0,063	0,919
JAG1	PSENEN	1,000	0,084	0,915
DTX2	UBA52	0,983	0,073	0,910
DTX1	APH1B	0,912	0,019	0,893
ARRB2	PSENEN	1,000	0,109	0,891
JAG1	DTX4	0,956	0,073	0,883
NCSTN	ADAM17	0,922	0,058	0,864
NCSTN	DNER	0,929	0,084	0,845
APH1A	ADAM17	0,995	0,162	0,834
CNTN1	UBA52	0,837	0,011	0,826
PSENEN	UBA52	0,999	0,178	0,821
MIB1	RPS27A	0,995	0,178	0,817
CNTN1	RPS27A	0,994	0,178	0,816
CNTN1	DTX1	0,975	0,161	0,814
PSEN1	APH1B	0,986	0,178	0,808
DLL4	DNER	0,832	0,039	0,793
APH1B	NUMB	0,949	0,161	0,789
NEURL	DNER	0,923	0,138	0,785
DTX1	RPS27A	0,961	0,178	0,783
ARRB1	PSEN1	0,021	0,803	0,782
ARRB1	UBA52	0,958	0,178	0,779
DTX2	APH1B	0,832	0,058	0,774
ADAM10	PSENEN	0,993	0,223	0,769
DTX4	APH1A	0,837	0,068	0,769
DLK1	DNER	0,992	0,223	0,768
ADAM17	UBA52	0,941	0,178	0,763
PSENEN	DNER	0,015	0,776	0,762
APH1A	UBA52	0,996	0,239	0,757
JAG1	NUMB	0,992	0,239	0,754
NCSTN	DLL1	0,939	0,185	0,754
ADAM10	NUMB	0,999	0,246	0,753
ARRB2	NUMB	0,998	0,246	0,752
ARRB1	PSENEN	0,993	0,246	0,747
JAG2	DNER	0,930	0,185	0,745
JAG1	ARRB2	0,755	0,011	0,744
ADAM10	DLL1	0,884	0,144	0,740
PSENEN	MIB1	0,978	0,246	0,732
ADAM10	NCSTN	1,000	0,268	0,732
DTX2	DLL4	1,000	0,268	0,732
DTX4	PSENEN	0,985	0,254	0,731
NCSTN	ARRB1	0,999	0,268	0,731
ADAM10	JAG1	0,969	0,239	0,731
ADAM10	DTX1	0,998	0,268	0,730
UBA52	NUMB	0,998	0,268	0,730
DTX2	CNTN1	0,975	0,246	0,729
ADAM10	APH1B	0,995	0,268	0,727
DTX4	PSEN2	1,000	0,274	0,726
JAG1	UBA52	1,000	0,274	0,726
ADAM10	ARRB1	0,747	0,025	0,721
DTX1	PSENEN	0,962	0,246	0,716
APH1A	DNER	0,878	0,162	0,716
DTX4	RPS27A	0,976	0,268	0,708
DTX4	NUMB	0,974	0,268	0,706
PSEN2	UBA52	0,878	0,178	0,700
CNTN1	PSEN2	0,973	0,273	0,700
RPS27A	DLK1	0,941	0,244	0,697
JAG2	ARRB2	0,873	0,178	0,694
ADAM10	DTX4	0,957	0,265	0,692
JAG2	NUMB	0,962	0,273	0,688
DLK1	NEURL	0,962	0,274	0,688
DTX4	DLL1	1,000	0,317	0,683
RPS27A	NEURL	0,898	0,216	0,682
MIB2	NCSTN	0,862	0,185	0,677
ADAM10	ADAM17	0,949	0,273	0,676
APH1A	PSEN2	0,759	0,103	0,656
CNTN1	DLL4	0,905	0,249	0,656
DTX1	APH1A	0,894	0,239	0,655
DLL4	APH1B	0,815	0,163	0,652
NCSTN	PSEN1	0,917	0,268	0,649
ARRB2	APH1A	0,983	0,335	0,648
CNTN1	MIB2	0,780	0,139	0,642
PSEN1	NEURL	0,863	0,227	0,636
ARRB1	DNER	0,787	0,154	0,633
DTX2	MIB1	0,949	0,317	0,633
APH1A	APH1B	0,921	0,293	0,628
CNTN1	DLK1	0,962	0,335	0,627
DTX2	DNER	0,888	0,268	0,620
ADAM10	DNER	0,862	0,244	0,619
DTX1	NUMB	0,884	0,268	0,616
ADAM10	MIB2	0,989	0,374	0,616
JAG1	APH1A	1,000	0,390	0,610
NCSTN	APH1B	0,999	0,390	0,609
DTX1	NEURL	0,996	0,390	0,606
JAG1	PSEN2	0,874	0,268	0,606
DTX2	ARRB1	0,920	0,314	0,606
DTX2	DTX1	0,873	0,268	0,605
DTX1	PSEN1	0,781	0,178	0,602
DLL1	RPS27A	0,781	0,178	0,602
ADAM17	NEURL	0,991	0,390	0,601
PSEN1	DLK1	0,055	0,656	0,601
MIB2	NEURL	0,727	0,131	0,595
DTX2	RPS27A	0,862	0,268	0,594
DTX2	PSENEN	0,996	0,404	0,593
ARRB1	ADAM17	0,999	0,407	0,592
DTX1	ARRB2	0,837	0,246	0,591
PSEN2	NEURL	0,791	0,216	0,575
PSENEN	PSEN2	0,998	0,427	0,571
MIB2	ARRB2	0,685	0,114	0,571
DTX1	ADAM17	0,958	0,390	0,568
NCSTN	JAG2	0,731	0,166	0,566
JAG1	ARRB1	1,000	0,434	0,565
PSENEN	NEURL	0,832	0,267	0,565
DLL4	UBA52	0,748	0,185	0,563
DLL4	RPS27A	0,720	0,162	0,559
ARRB2	MIB1	0,999	0,442	0,557
CNTN1	MIB1	0,993	0,442	0,551
PSEN1	ADAM17	0,583	0,033	0,550
RPS27A	DNER	0,992	0,442	0,549
MIB2	JAG2	0,982	0,433	0,549
MIB1	DLK1	0,865	0,317	0,548
CNTN1	NCSTN	0,889	0,343	0,547
DTX4	PSEN1	0,014	0,553	0,539
APH1B	NEURL	0,549	0,010	0,539
ARRB2	APH1B	0,985	0,447	0,538
ADAM10	ARRB2	0,805	0,268	0,537
CNTN1	ARRB1	0,921	0,389	0,532
DTX1	DLL1	0,921	0,389	0,532
MIB1	DNER	0,787	0,268	0,519
PSENEN	RPS27A	0,757	0,246	0,511
DTX2	APH1A	0,683	0,177	0,507
MIB1	ADAM17	0,428	0,935	0,507
PSEN1	NUMB	0,819	0,314	0,504
JAG2	PSEN1	0,746	0,246	0,500
CNTN1	DLL1	0,942	0,442	0,500
CNTN1	JAG1	0,889	0,390	0,499
JAG2	DLL4	0,956	0,461	0,495
JAG2	PSEN2	0,978	0,485	0,493
DLL4	MIB1	0,764	0,273	0,491
DTX2	MIB2	0,875	0,392	0,483
MIB1	NEURL	0,921	0,443	0,477
MIB1	APH1B	0,718	0,246	0,472
DTX1	ARRB1	0,999	0,531	0,467
MIB2	DLL1	0,735	0,268	0,467
CNTN1	DNER	0,853	0,390	0,463
JAG1	RPS27A	0,637	0,178	0,459

Gene 1	Gene 2	AII	ODII	Diferença absoluta
JAG1	DLL1	0,720	0,265	0,455
JAG2	DLK1	0,837	0,390	0,448
NCSTN	NUMB	0,999	0,553	0,446
CNTN1	PSENEN	0,658	0,216	0,442
JAG1	ADAM17	0,995	0,553	0,441
CNTN1	PSEN1	0,549	0,114	0,435
RPS27A	ADAM17	0,727	0,293	0,433
ADAM10	JAG2	0,832	0,404	0,428
DLL1	UBA52	0,014	0,439	0,425
UBA52	APH1B	0,857	0,434	0,423
DTX4	ARRB1	0,862	0,447	0,416
MIB2	DTX1	0,689	0,274	0,415
APH1A	NEURL	0,651	0,249	0,402
DTX1	MIB1	0,431	0,033	0,398
JAG1	NCSTN	1,000	0,602	0,398
ARRB2	ADAM17	0,428	0,033	0,396
PSEN1	DNER	0,947	0,553	0,393
MIB2	RPS27A	0,409	0,803	0,393
NCSTN	RPS27A	0,014	0,407	0,393
MIB2	PSEN1	0,578	0,190	0,389
DLL1	APH1A	0,413	0,025	0,388
DLL4	NEURL	0,633	0,246	0,387
NCSTN	UBA52	1,000	0,615	0,385
NCSTN	PSENEN	0,996	0,615	0,381
DLL4	PSEN2	0,406	0,033	0,373
NCSTN	PSEN2	0,637	0,268	0,369
ARRB2	NEURL	0,637	0,268	0,369
JAG2	RPS27A	0,993	0,625	0,368
UBA52	NEURL	0,969	0,601	0,367
ARRB1	MIB1	0,199	0,566	0,367
JAG1	APH1B	0,633	0,273	0,359
DLL4	PSEN1	0,523	0,166	0,357
MIB2	PSENEN	0,574	0,925	0,352
DTX1	DNER	0,406	0,059	0,347
MIB2	PSEN2	0,614	0,268	0,346
DLL4	NUMB	0,113	0,458	0,345
ARRB2	PSEN2	0,960	0,615	0,345
ADAM10	PSEN1	1,000	0,656	0,344
JAG1	DLK1	0,397	0,054	0,343
ARRB2	RPS27A	0,731	0,390	0,342
JAG1	NEURL	0,988	0,656	0,333
NCSTN	ARRB2	0,596	0,268	0,328
MIB2	DTX4	0,533	0,206	0,327
ADAM10	MIB1	0,231	0,553	0,322
NCSTN	NEURL	0,999	0,678	0,321
DTX4	ARRB2	0,874	0,553	0,321
DLL1	APH1B	0,874	0,553	0,320
MIB2	DLL4	0,369	0,054	0,315
DLL1	ADAM17	0,866	0,553	0,313
MIB2	APH1B	0,865	0,553	0,312
DTX2	ADAM17	0,700	0,389	0,311
DLL1	NUMB	0,995	0,684	0,311
DTX1	DLK1	0,133	0,442	0,309
ARRB2	DLL4	0,483	0,178	0,305
DLL1	PSENEN	0,981	0,678	0,304
PSEN1	UBA52	0,096	0,389	0,293
ARRB1	RPS27A	0,314	0,025	0,289
DTX2	NEURL	0,007	0,293	0,286
PSENEN	APH1B	0,700	0,416	0,284
ADAM17	DLK1	0,445	0,161	0,284
JAG1	DLL4	0,328	0,045	0,283
CNTN1	ADAM17	0,529	0,246	0,283
APH1A	RPS27A	0,314	0,033	0,281
DTX4	APH1B	0,431	0,151	0,280
DTX1	PSEN2	0,708	0,434	0,273
UBA52	DNER	0,280	0,553	0,273
ARRB2	DNER	0,280	0,011	0,269
DTX2	DTX4	0,878	0,612	0,266
ADAM10	NEURL	0,943	0,678	0,265
ADAM10	DLL4	0,651	0,390	0,262
DTX4	DNER	0,523	0,268	0,255
MIB2	NUMB	0,549	0,803	0,254
PSENEN	NUMB	1,000	0,746	0,254
NCSTN	APH1A	1,000	0,746	0,254
JAG2	UBA52	0,498	0,246	0,252
JAG2	APH1A	0,428	0,178	0,251
JAG1	MIB1	0,540	0,293	0,247
DTX2	NUMB	0,805	0,560	0,245
JAG1	DTX1	0,992	0,760	0,233
MIB2	DLK1	0,037	0,267	0,230
PSEN2	DLK1	0,018	0,246	0,228
DTX1	JAG2	0,470	0,244	0,227
ADAM17	NUMB	0,627	0,404	0,223
JAG2	MIB1	0,999	0,776	0,222
ADAM10	UBA52	0,899	0,678	0,221
APH1B	DLK1	0,055	0,273	0,219
ARRB1	APH1A	1,000	0,782	0,217
CNTN1	NEURL	0,425	0,216	0,209
ARRB2	UBA52	0,943	0,741	0,202
DTX1	DTX4	0,997	0,803	0,195
DLL4	DLK1	0,996	0,803	0,194
DLL1	DNER	0,082	0,273	0,191
DTX1	DLL4	0,007	0,193	0,186
PSEN1	PSEN2	0,408	0,223	0,185
NUMB	DNER	0,428	0,246	0,182
RPS27A	UBA52	0,213	0,033	0,180
DTX4	NEURL	0,301	0,121	0,180
ARRB2	PSEN1	0,375	0,553	0,179
ADAM10	RPS27A	0,280	0,103	0,178
APH1A	PSENEN	1,000	0,825	0,175
PSEN2	APH1B	0,199	0,374	0,175
PSEN2	MIB1	1,000	0,825	0,175
ADAM17	DNER	0,413	0,240	0,173
DTX2	NCSTN	0,566	0,394	0,172
JAG1	JAG2	0,440	0,268	0,172
ARRB1	PSEN2	0,349	0,178	0,170
MIB1	UBA52	0,414	0,246	0,168
CNTN1	ARRB2	0,082	0,249	0,167
DLL1	DLL4	0,190	0,025	0,165
PSENEN	DLK1	0,007	0,166	0,159
DTX2	JAG2	0,397	0,553	0,156
ADAM10	PSEN2	0,423	0,268	0,155
APH1A	PSEN1	0,007	0,161	0,154
APH1A	DLK1	0,043	0,196	0,153
RPS27A	NUMB	0,118	0,268	0,150
JAG2	APH1B	0,389	0,246	0,143
ARRB2	DLK1	0,118	0,259	0,142
APH1B	DNER	0,913	0,772	0,141
PSEN2	DNER	0,107	0,246	0,139
MIB2	JAG1	0,156	0,024	0,132
DLL4	ADAM17	0,445	0,314	0,130
ARRB1	NUMB	0,090	0,216	0,126
ARRB1	APH1B	0,158	0,033	0,125
DTX4	DLL4	0,349	0,223	0,125
JAG2	ADAM17	0,658	0,776	0,118
JAG1	DNER	0,363	0,246	0,117
ADAM10	DTX2	0,152	0,268	0,116
PSEN2	NUMB	0,999	0,884	0,115
NCSTN	DLL4	0,133	0,246	0,113
DTX2	PSEN2	0,906	0,803	0,103
DLL1	PSEN1	0,287	0,390	0,103
DTX4	ADAM17	0,174	0,268	0,094
CNTN1	JAG2	0,337	0,244	0,094
CNTN1	APH1A	0,409	0,317	0,093
DLL1	JAG2	0,737	0,645	0,092
DTX2	PSEN1	0,205	0,114	0,091
JAG1	PSEN1	0,155	0,244	0,088
NUMB	NEURL	0,301	0,389	0,087
CNTN1	APH1B	0,347	0,434	0,087
JAG2	PSENEN	0,109	0,194	0,085
UBA52	DLK1	0,656	0,574	0,082
ARRB1	DLL4	0,107	0,025	0,081
NCSTN	MIB1	0,273	0,194	0,079
DLL1	ARRB2	0,882	0,803	0,079

Gene 1	Gene 2	AII	ODII	Diferença absoluta
DTX4	JAG2	0,502	0,423	0,079
DLL1	ARRB1	0,231	0,154	0,077
DTX4	UBA52	0,466	0,390	0,077
ARRB1	NEURL	1,000	0,925	0,075
DLL1	PSEN2	1,000	0,925	0,075
JAG2	NEURL	0,999	0,925	0,074
ADAM17	APH1B	0,155	0,228	0,073
DLL1	DLK1	0,633	0,560	0,072
APH1A	DLL4	0,089	0,161	0,072
ADAM10	DLK1	0,339	0,268	0,071
MIB2	ADAM17	0,339	0,268	0,071
DTX2	DLL1	0,549	0,615	0,066
DLL1	MIB1	1,000	0,935	0,065
ARRB1	ARRB2	0,096	0,033	0,063
DTX4	NCSTN	0,631	0,573	0,058
MIB1	NUMB	0,981	0,925	0,056
DLL1	NEURL	0,470	0,415	0,056
PSEN1	RPS27A	0,444	0,389	0,055
DLL4	PSENEN	0,021	0,073	0,053
RPS27A	APH1B	0,604	0,553	0,051
JAG2	ARRB1	0,974	0,935	0,039
MIB2	UBA52	0,231	0,268	0,037
APH1A	MIB1	0,231	0,268	0,037
NCSTN	DLK1	0,369	0,404	0,035
DTX4	DLK1	0,633	0,666	0,034
DTX2	DLK1	0,599	0,566	0,033
MIB2	APH1A	0,007	0,039	0,032
PSEN2	ADAM17	0,428	0,407	0,021
PSENEN	PSEN1	0,039	0,058	0,020
MIB2	DNER	0,954	0,935	0,019
ADAM10	CNTN1	0,155	0,137	0,019
NUMB	DLK1	0,406	0,390	0,016
PSEN1	MIB1	0,107	0,122	0,015
MIB2	ARRB1	0,155	0,166	0,011
MIB2	MIB1	0,605	0,615	0,010
CNTN1	NUMB	0,021	0,024	0,003
ARRB1	DLK1	0,181	0,178	0,003

Centralidades de grau

Na tabela a seguir mostramos as centralidades de grau obtidas para cada gene do conjunto REACTOME ACTIVATED NOTCH1 TRANSMITS SIGNAL TO THE NUCLEUS utilizando no grafo de coexpressão do astrocitoma grau II (AII) e do oligodendroglioma grau II (ODII).

Gene	AII	ODII	Diferença entre AII e ODII
DTX1	19,761	7,064	12,697
JAG1	18,761	6,925	11,836
ADAM10	19,028	7,540	11,488
CNTN1	17,398	6,237	11,161
NCSTN	19,402	8,813	10,590
APH1A	16,500	6,213	10,287
DTX2	18,576	8,316	10,260
ARRB2	17,402	7,385	10,017
RPS27A	16,675	6,717	9,958
PSENEN	18,582	8,652	9,931
UBA52	17,851	7,990	9,861
DTX4	17,842	8,157	9,685
NEURL	18,776	9,362	9,414
APH1B	16,841	7,529	9,312
ADAM17	16,688	7,994	8,694
DNER	16,365	7,779	8,587
NUMB	17,997	9,867	8,129
PSEN2	17,412	9,283	8,129
JAG2	17,759	10,500	7,258
DLL4	12,581	5,434	7,146
MIB1	17,404	10,518	6,886
ARRB1	15,059	8,416	6,643
DLL1	17,128	10,862	6,266
MIB2	14,300	8,694	5,607
PSEN1	11,503	7,778	3,725
DLK1	11,677	8,911	2,766

Análise de diferença de expressão genética

Na tabela a seguir mostramos, para cada gene do conjunto REACTOME ACTIVATED NOTCH1 TRANSMITS SIGNAL TO THE NUCLEUS, a diferença entre a expressão média do grafo do astrocitoma grau II (AII) e do oligodendroglioma grau II (ODII), o p-valor do teste t entre as médias, a diferença entre as pseudomedianas do AII e do ODII e o p-valor do teste de Wilcoxon-Mann-Whitney.

Gene	Diferença entre as médias (AII – ODII)	P-valor do teste t	Diferença entre as pseudomedianas (AII – ODII)	P-valor do teste de Wilcoxon-Mann-Whitney
ARRB2	0,163	0,030	0,165	0,014
DNER	-0,223	0,050	-0,215	0,032
DTX1	-0,135	0,060	-0,113	0,057
MIB2	0,129	0,076	0,144	0,062
PSENEN	0,267	0,073	0,267	0,065
MIB1	-0,130	0,175	-0,133	0,131
ADAM10	-0,116	0,143	-0,127	0,155
RPS27A	-0,080	0,218	-0,089	0,180
DLL1	-0,247	0,182	-0,250	0,202
DTX2	0,083	0,073	0,068	0,219
ARRB1	0,135	0,164	0,122	0,222
UBA52	0,095	0,125	0,071	0,228
PSEN1	-0,085	0,264	-0,066	0,307
DTX4	-0,150	0,284	-0,111	0,519
APH1A	0,107	0,204	0,054	0,546
JAG1	0,069	0,663	-0,068	0,562
DLL4	0,026	0,573	0,025	0,567
NUMB	-0,029	0,651	-0,038	0,583
NCSTN	0,006	0,914	-0,039	0,589
ADAM17	-0,018	0,814	-0,037	0,645
JAG2	-0,007	0,912	-0,027	0,645
DLK1	0,088	0,342	0,017	0,663
PSEN2	-0,040	0,447	-0,016	0,686
CNTN1	-0,051	0,816	-0,048	0,770
APH1B	-0,017	0,672	-0,010	0,807
NEURL	0,011	0,864	0,010	0,876

Referências Bibliográficas

Albert *et al.* **(2000)** Réka Albert, Hawoong Jeong e Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382. Citado na pág. 40

Barabási e Albert (1999) Albert-László Barabási e Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512. Citado na pág. 40, 41

Barabási e Oltvai (2004) Albert-László Barabási e Zoltán N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113. Citado na pág. 2, 40

Benjamini e Hochberg (1995) Yoav Benjamini e Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 57(1):289–300. Citado na pág. 48, 52, 63, 65, 91

Bollobás *et al.* **(2001)** Béla Bollobás, Oliver Riordan, Joel Spencer e Gábor Tusnády. The degree sequence of a scale-free random graph process. *Random Struct. Alg.*, 18(3):279–290. Citado na pág. 41

Bolstad *et al.* **(2003)** B. M. Bolstad, R. A. Irizarry, M. Åstrand e T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193. Citado na pág. 76

Bonacich (1972) Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *J Math Sociol*, 2(1):113–120. Citado na pág. 39

Brandes (2001) Ulrik Brandes. A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology*, 25:163–177. Citado na pág. 38

Carter *et al.* **(2004)** Scott L. Carter, Christian M. Brechbühler, Michael Griffin e Andrew T. Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–2250. Citado na pág. 37

Casella e Berger (2001) George Casella e Roger L. Berger. *Statistical Inference*. Cengage Learning, Australia ; Pacific Grove, CA, 2nd edition ed. Citado na pág. 16

Chan *et al.* **(2000)** W. Y. Chan, K. K. Cheung, J. O. Schorge, L. W. Huang, W. R. Welch, D. A. Bell, R. S. Berkowitz e S. C. Mok. Bcl-2 and p53 protein expression, apoptosis, and p53 mutation in human epithelial ovarian cancers. *Am. J. Pathol.*, 156(2):409–417. Citado na pág. 71

Choi e Kendziorski (2009) YounJeong Choi e Christina Kendziorski. Statistical methods for gene set co-expression analysis. *Bioinformatics*, 25(21):2780–2786. Citado na pág. 2

Costa *et al.* **(2007)** Luciano da F. Costa, Francisco A. Rodrigues, Gonzalo Travieso e Paulino R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242. Citado na pág. 37

Dai *et al.* **(2005)** Manhong Dai, Pinglang Wang, Andrew D. Boyd, Georgi Kostov, Brian Athey, Edward G. Jones, William E. Bunney, Richard M. Myers, Terry P. Speed, Huda Akil, Stanley J. Watson e Fan Meng. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*, 33(20):e175. Citado na pág. 51

de la Fuente (2010) Alberto de la Fuente. From 'differential expression' to 'differential networking' – identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*, 26(7):326–333. Citado na pág. 9, 10, 71

Dwass (1957) Meyer Dwass. Modified Randomization Tests for Nonparametric Hypotheses. *Ann. Math. Statist.*, 28(1):181–187. Citado na pág. 50

Efron (1979) B. Efron. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.*, 7(1):1–26. Citado na pág. 52

Efron e Tibshirani (2007) Bradley Efron e Robert Tibshirani. On Testing the Significance of Sets of Genes. *Ann. Appl. Stat.*, 1(1):107–129. Citado na pág. 1

Erdös e Rényi (1959) Paul Erdös e Alfréd Rényi. On random graphs. *Publ. Math. Debrecen*, 6:290–297. Citado na pág. 41

Fisher (1935) Sir Ronald Aylmer Fisher. *The Design of Experiments*. Oliver and Boyd. Citado na pág. 50

Freeman (1978) Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239. Citado na pág. 38

Fujita *et al.* **(2009)** André Fujita, João Ricardo Sato, Marcos Angelo Almeida Demasi, Mari Cleide Sogayar, Carlos Eduardo Ferreira e Satoru Miyano. Comparing Pearson, Spearman and Hoeffding's D measure for gene expression association analysis. *J Bioinform Comput Biol*, 7(4):663–684. Citado na pág. 28, 30

Granger (1969) C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438. Citado na pág. 10

Heller *et al.* **(2013)** Ruth Heller, Yair Heller e Malka Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510. Citado na pág. 2, 13, 21

Hoeffding (1948) Wassily Hoeffding. A Non-Parametric Test of Independence. *Ann. Math. Statist.*, 19(4):546–557. Citado na pág. 2, 13, 23

Huber *et al.* **(2013)** Roland M. Huber, Michal Rajski, Balasubramanian Sivasankaran, Gerald Moncayo, Brian A. Hemmings e Adrian Merlo. Deltex-1 Activates Mitotic Signaling and Proliferation and Increases the Clonogenic and Invasive Potential of U373 and LN18 Glioblastoma Cells and Correlates with Patient Survival. *PLoS ONE*, 8(2):e57793. Citado na pág. 69

Hudson *et al.* **(2009)** Nicholas J. Hudson, Antonio Reverter e Brian P. Dalrymple. A Differential Wiring Analysis of Expression Data Correctly Identifies the Gene Containing the Causal Mutation. *PLoS Comput Biol*, 5(5):e1000382. Citado na pág. 71

Irizarry *et al.* **(2003a)** Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs e Terence P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, 31(4):e15. Citado na pág. 75

Irizarry *et al.* **(2003b)** Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf e Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264. Citado na pág. 31, 51, 75, 76

Irizarry *et al.* **(2009)** Rafael A. Irizarry, Chi Wang, Yun Zhou e Terence P. Speed. Gene set enrichment analysis made simple. *Stat Methods Med Res*, 18(6):565–575. Citado na pág. 1

Jiang e Gentleman (2007) Zhen Jiang e Robert Gentleman. Extensions to gene set enrichment. *Bioinformatics*, 23(3):306–313. Citado na pág. 1

Kato *et al.* **(2003)** Kiyoshi Kato, Toshihiko Toki, Motohiko Shimizu, Tanri Shiozawa, Shingo Fujii, Toshio Nikaido e Ikuo Konishi. Expression of replication-licensing factors MCM2 and MCM3 in normal, hyperplastic, and carcinomatous endometrium: correlation with expression of Ki-67 and estrogen and progesterone receptors. *Int. J. Gynecol. Pathol.*, 22(4):334–340. Citado na pág. 71

Keller *et al.* **(2008)** Mark P. Keller, YounJeong Choi, Ping Wang, Dawn Belt Davis, Mary E. Rabaglia, Angie T. Oler, Donald S. Stapleton, Carmen Argmann, Kathy L. Schueler, Steve Edwards, H. Adam Steinberg, Elias Chaibub Neto, Robert Kleinhanz, Scott Turner, Marc K. Hellerstein, Eric E. Schadt, Brian S. Yandell, Christina Kendziorski e Alan D. Attie. A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Res.*, 18(5):706–716. Citado na pág. 71

Kendall (1938) M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1-2): 81–93. Citado na pág. 2, 13, 19, 63

Khinchin (1957) A. Ya Khinchin. *Mathematical Foundations of Information Theory*. Dover Publications, New York, 1st dover edition edition ed. Citado na pág. 44

Langfelder e Horvath (2008) Peter Langfelder e Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.*, 9(1):559. Citado na pág. 2, 89

Lopez-Fernandez *et al.* **(2004)** Luis Lopez-Fernandez, Gregorio Robles e Jesus M. Gonzalez-Barahona. Applying Social Network Analysis to the Information in CVS Repositories. Em *Proceedings 1st International Workshop on Mining Software Repositories*, páginas 101–105. Citado na pág. 39

Mazieres *et al.* **(2005)** Julien Mazieres, Biao He, Liang You, Zhidong Xu e David M. Jablons. Wnt signaling in lung cancer. *Cancer Lett.*, 222(1):1–10. Citado na pág. 31

McCall *et al.* **(2010)** Matthew N. McCall, Benjamin M. Bolstad e Rafael A. Irizarry. Frozen robust multiarray analysis (fRMA). *Biostat*, 11(2):242–253. Citado na pág. 75

Miller *et al.* **(2011)** Jeremy A. Miller, Chaochao Cai, Peter Langfelder, Daniel H. Geschwind, Sunil M. Kurian, Daniel R. Salomon e Steve Horvath. Strategies for aggregating gene expression data: The collapseRows R function. *BMC Bioinform.*, 12(1):322. Citado na pág. 90

NCI () NCI. REMBRANDT home page. http://rembrandt.nci.nih.gov. Acesso em 25/06/2013. Citado na pág. 51, 52

Newman (2001) M. E. J. Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64(1):016132. Citado na pág. 39

Okayama *et al.* **(2012)** Hirokazu Okayama, Takashi Kohno, Yuko Ishii, Yoko Shimada, Kouya Shiraishi, Reika Iwakawa, Koh Furuta, Koji Tsuta, Tatsuhiro Shibata, Seiichiro Yamamoto, Shun-ichi Watanabe, Hiromi Sakamoto, Kensuke Kumamoto, Seiichi Takenoshita, Noriko Gotoh, Hideaki Mizuno, Akinori Sarai, Shuichi Kawano, Rui Yamaguchi, Satoru Miyano e Jun Yokota. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res.*, 72(1):100–111. Citado na pág. 31

Opsahl *et al.* **(2010)** Tore Opsahl, Filip Agneessens e John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3): 245–251. Citado na pág. 39

Pearson (1920) Karl Pearson. Notes on the History of Correlation. *Biometrika*, 13(1): 25–45. Citado na pág. 2, 13, 17, 63

Purow *et al.* **(2005)** Benjamin W. Purow, Raqeeb M. Haque, Martha W. Noel, Qin Su, Michael J. Burdick, Jeongwu Lee, Tilak Sundaresan, Sandra Pastorino, John K. Park, Irina Mikolaenko, Dragan Maric, Charles G. Eberhart e Howard A. Fine. Expression of Notch-1 and its ligands, Delta-like-1 and Jagged-1, is critical for glioma cell survival and proliferation. *Cancer Res.*, 65(6):2353–2363. Citado na pág. 67

Rahmatallah *et al.* **(2014)** Yasir Rahmatallah, Frank Emmert-Streib e Galina Glazko. Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets. *Bioinformatics*, 30(3):360–368. Citado na pág. 2, 39

Reshef *et al.* **(2011)** David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher e Pardis C. Sabeti. Detecting Novel Associations in Large Data Sets. *Science*, 334(6062): 1518–1524. Citado na pág. 2, 13, 24

Roy *et al.* **(2014)** Swarup Roy, Dhruba K. Bhattacharyya e Jugal K. Kalita. Reconstruction of gene co-expression network from microarray data using local expression patterns. *BMC Bioinform.*, 15(Suppl 7):S10. Citado na pág. 40

Sain e Scott (1996) Stephan R. Sain e David W. Scott. On Locally Adaptive Density Estimation. *Journal of the American Statistical Association*, 91(436):1525–1534. Citado na pág. 64

Santos *et al.* **(2014)** Suzana de Siqueira Santos, Daniel Yasumasa Takahashi, Asuka Nakata e André Fujita. A comparative study of statistical methods used to identify dependencies between gene expression signals. *Brief Bioinform*, 15(6):906–918. Citado na pág. xvi, 29, 73

Shannon (1948) C. E. Shannon. A Mathematical Theory of Communication. *Bell Syst. Tech. J.*, 27(3):379–423. Citado na pág. 2, 13, 23, 42, 43

Shannon *et al.* **(2003)** Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski e Trey Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.*, 13(11):2498–2504. Citado na pág. 3

Spearman (1904) C. Spearman. The Proof and Measurement of Association between Two Things. *Am J Psychol*, 15(1):72–101. Citado na pág. 2, 13, 19, 52, 57, 63, 65

Stockhausen *et al.* **(2010)** Marie-Thérése Stockhausen, Karina Kristoffersen e Hans Skovgaard Poulsen. The functional role of Notch signaling in human gliomas. *Neuro Oncol*, 12(2):199–211. Citado na pág. 67

Sturges (1926) Herbert A. Sturges. The Choice of a Class Interval. *J. Am. Statist. Assoc.*, 21(153):65–66. Citado na pág. 64

Subramanian *et al.* **(2005)** Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander e Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102 (43):15545–15550. Citado na pág. 1, 51, 57, 65

Székely *et al.* **(2007)** Gábor J. Székely, Maria L. Rizzo e Nail K. Bakirov. Measuring and Testing Dependence by Correlation of Distances. *Ann. Stat.*, 35(6):2769–2794. Citado na pág. 2, 13, 20

Takahashi *et al.* **(2012)** Daniel Yasumasa Takahashi, João Ricardo Sato, Carlos Eduardo Ferreira e André Fujita. Discriminating Different Classes of Biological Networks by Analyzing the Graphs Spectra Distribution. *PLoS ONE*, 7(12):e49949. Citado na pág. xvi, 3, 37, 41, 42, 43, 44, 45, 46

Van Mieghem (2010) Piet Van Mieghem. *Graph Spectra for Complex Networks*. Cambridge University Press, Cambridge. Citado na pág. 40

Watts e Strogatz (1998) Duncan J. Watts e Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442. Citado na pág. 39, 41

Yamauchi *et al.* **(2012)** Mai Yamauchi, Rui Yamaguchi, Asuka Nakata, Takashi Kohno, Masao Nagasaki, Teppei Shimamura, Seiya Imoto, Ayumu Saito, Kazuko Ueno, Yousuke Hatanaka, Ryo Yoshida, Tomoyuki Higuchi, Masaharu Nomura, David G. Beer, Jun Yokota, Satoru Miyano e Noriko Gotoh. Epidermal growth factor receptor tyrosine kinase defines critical prognostic genes of stage I lung adenocarcinoma. *PLoS ONE*, 7(9):e43923. Citado na pág. 31

Zhang *et al.* **(2012)** Xiaohua Zhang, Tao Chen, Jiannan Zhang, Qin Mao, Shanquan Li, Wenhao Xiong, Yongming Qiu, Qiuling Xie e Jianwei Ge. Notch1 promotes glioma cell migration and invasion by stimulating β-catenin and NF-κB signaling via AKT activation. *Cancer Sci.*, 103(2):181–190. Citado na pág. 67