UNIVERSIDADE DE SÃO PAULO
PROGRAMA INTERUNIDADES DE PÓS GRADUAÇÃO EM BIOINFORMÁTICA
INSTITUTO DE FÍSICA DE SÃO CARLOS

Cynthia Martins Villar Couto

Estruturas de Redes em Ossos ao longo do desenvolvimento

São Carlos - SP

Cynthia Martins Villar Couto

Estruturas de Redes em Ossos ao longo do desenvolvimento

Tese apresentada ao Programa Interunidades de Pós Graduação em Bioinformática da Universidade de São Paulo para obtenção do título de Doutor em Ciências.

Orientador: Prof. Dr. Luciano da Fontoura Costa.

Durante a elaboração deste trabalho, a aluna recebeu apoio financeiro da FAPESP (Fundação de Apoio à Pesquisa do Estado de São Paulo).

São Carlos - SP

Agradecimentos

Sou grata por muitas coisas que aconteceram durante o período do meu doutorado, mas como é costume citar somente o nome de pessoas na seção de agradecimentos de uma dissertação ou tese, quero agradecer a Deus, que manteve todas essas coisas que me referi acima sob Seu controle. Muito obrigada, Senhor.

Agradeço ao meu marido, Leandro, pelo amor incondicional. Agradeço também por ter sido um grande incentivador para que eu encarasse o desafio de mudar de área de conhecimento. Seu apoio foi fundamental durante todo esse período.

Agradeço meu pai, minha mãe e minha irmã também pelo amor incondicional, apoio nos momentos difíceis, comemorações com as conquistas, suporte financeiro, psicológico, emocional e espiritual,... enfim, por tudo.

Agradeço o amor e apoio da minha “família do lado de lá”, meus sogros e minha cunhada.

Agradeço ao Prof. Dr. Luciano da Fontoura Costa pela orientação, ensinamentos, aulas particulares, e pela compreensão durante o doutorado. Me sinto privilegiada pelo fato de que a minha formação foi embasada, direcionada e orientada por um dos melhores pesquisadores da área.

Agradeço ao Prof. Dr. César Henrique Comim por ter abraçado a função de me dar suporte no laboratório. Você fez muito mais do que isso, meu amigo.

Agradeço ao colega de laboratório, Dr. Filipi Nascimento Silva pela colaboração com as análises de dimensionalidade em redes vasculares de ossos de camundongos.

Agradeço meus outros colegas de sala e da USP, Thomas, Alessandra, Henrique, Carol, Paulo, e Davi, por todo o apoio também. Conhecer e desenvolver uma amizade com vocês foi certamente um dos pontos altos da minha trajetória durante o doutorado.

Agradeço ao Prof. Dr. Marcelo Emílio Beletti pela colaboração durante este Doutorado, e pelas amostras de camundongo que compuseram o Capítulo sobre a caracterização do sistema Haversiano. Agradeço também pelo auxílio na discussão sobre esses dados.

Agradeço a Embrapa ao Prof. Dr. Paulo Cruvinel pela colaboração em relação ao uso do microtomógrafo nas dependências da Embrapa Instrumentação de São Carlos, para aquisição dos canais de Havers e Volkmann das amostras de camundongo.

Agradeço ao Presbítero Dr. Paulo Lasso, técnico da Embrapa Instrumentação de São Carlos, pelo auxílio antes, durante e depois da aquisição das imagens de microCT.

Agradeço aos membros da Igreja Presbiteriana de São Carlos, especialmente meus amigos da União de Mocidade Presbiteriana, por terem sido verdadeiramente uma família por todo o tempo que residi em São

Carlos. Citar nomes aqui seria uma injustiça, porque eu provavelmente não conseguiria me lembrar em tempo hábil de todos os queridos que estiveram comigo durante esses 10 anos de membresia.

Agradeço meus colegas do Programa Interunidades de Pós Graduação em Bioinformática pela amizade. Vocês sempre me receberam muito bem durante minhas idas à São Paulo, e vou guardar esses momentos no coração. Quero citar aqui as secretárias de pós graduação que estiveram no Programa durante meu doutorado, Patricia Martoneli e Cristiane de Fátima Braulino, e agradecer por tudo! Tudo mesmo! O trabalho de vocês é fundamental para o bom funcionamento da pós graduação. Muito obrigada.

Agradeço a cidade de São Carlos por me acolher, e por se tornar o meu lar.

Agradeço ao IFSC-USP pelo tempo de uso de sua infraestrutura e dependências.

Agradeço a CAPES pelo período em que me forneceu bolsa de estudos.

Agradeço a FAPESP por ter fomentado minha pesquisa (processo: 2013/19082-7).

Sola Gratia
Sola Fide
Solus Christus
Sola Scriptura
Soli Deo Gloria

Resumo

Uma das possíveis razões do sucesso da área de Redes Complexas decorre da flexibilidade destas estruturas para representação e modelagem de inúmeros sistemas complexos, incluindo em biologia. Entretanto, existem alguns aspectos do uso destes conceitos ainda pouco detalhados, como a questão da limiarização de relacionamentos graduados de forma a se obter uma rede binária de conexões. Uma outra questão interessante, ainda em aberto, refere-se a como redes complexas derivadas de sistemas diversos assemelham-se ou não umas às outras. Em biologia, esta questão aparece com particular interesse no que se refere às escalas das estruturas e sistemas biológicos, motivando a busca de analogias estruturais e funcionais. O presente trabalho de doutorado situa-se na interseção destes dois problemas. Em primeiro lugar, utilizamos a importante questão da limiarização de redes de co-expressão gênica como laboratório para desenvolver e comparar cinco métodos deste tipo, com fundamentações diferentes. Verificamos que dependendo da natureza do banco de dados, o impacto da limiarização nas propriedades topológicas pode ser grande, e sugerimos diretrizes de como utilizar os métodos diante do comportamento dos dados. Em seguida, abordamos a representação dos canais do sistema Haversiano dos ossos, com o objetivo de estudar este problema em particular e compará-lo com as redes de co-expressão na busca de analogias topológicas. As análises mostraram que os ossos são indistinguíveis em relação às propriedades topológicas das redes, mas nota-se uma variação mais pronunciada em relação às propriedades geométricas. Isso sugere que a arquitetura topológica do sistema vascular pode ser independente do tipo ósseo, mas que a demanda biológica de transporte pode variar em relação à posição no mesmo osso, e entre ossos diferentes. Como as redes do sistema Haversiano possuem pesos relacionados à espessura dos canais, utilizamos e comparamos os métodos de limiarização aqui propostos como forma de validação dos resultados. Concluindo estes desenvolvimentos, realizamos uma comparação estrutural dos dois tipos de redes obtidas, ou seja, de co-expressão gênica e de canais Haversianos.

Palavras-chave: Redes Complexas, limiarização, limiar, expressão gênica, rede de co-expressão gênica, ossos, sistema Haversiano, canais de Havers, canais de Volkmann

Abstract

One of the possible reasons for the success of Complex Networks arises from the flexibility of these structures for representation and modeling of numerous complex systems, including in biology. However, there are still some aspects of the use of these concepts, such as the question of the thresholding of graduated relationships in order to obtain a binary network of connections. Another interesting question, still open, concerns how complex networks derived from different systems are similar to another or are not. In biology, this question appears with particular interest in the scales of biological structures and systems, motivating the search for structural and functional analogies. The present PhD work lies at the intersection of these two problems. First, we used the important question of the thresholding of gene co-expression networks as a laboratory for development and to compare five methods of this type, with different foundations. We have found that depending on the nature of the database, the impact of thresholding on topological properties may be large, and we suggest guidelines on how to use the methods in face of the data's behavior. Then, we discuss the characterization of the channels of the Haversian system of bones, with the aim of studying this particular problem and comparing it with the networks of co-expression in the search for topological analogies. The analyzes showed that the bones are indistinguishable in relation to the topological properties of the networks, but a more pronounced variation in relation to the geometric properties is noticed. This suggests that the topological architecture of the vascular system may be independent of the bone type but that the biological demand for transport may be varying relatively to the position in the same bone and between different bones. As the networks of the Haversian system have weights related to the thickness of the channels, we used and compared the thresholding methods proposed here for the validation of the results. Concluding these developments, we performed a structural comparison of the two types of networks obtained, the gene co-expression network and the Haversian channels network.

Keywords: Complex networks, threshold, threshold, gene expression, genetic co-expression network, bones, Haversian system, Havers channels, Volkmann channels

Conteúdo

1	Introdução	1
1.1	Motivação e Objetivos	3
1.2	Organização da Tese	4
2	Conceitos Básicos	5
2.1	Medidas estatísticas	5
2.1.1	Média aritmética	5
2.1.2	Mediana	5
2.1.3	Desvio padrão	6
2.1.4	Entropia	6
2.1.5	Coeficiente de Correlação de Pearson (CCP)	6
2.1.6	Análise de Componentes Principais (do inglês, <i>Principal Component Analysis</i> - PCA)	7
2.2	Nomalização <i>z-score</i>	8
2.3	Medidas topológicas de Rede	8
2.3.1	Estatísticas de grau	9
2.3.2	Transitividade local	9
2.3.3	Centralidade de interposição (do inglês, <i>Betweenness centrality</i>)	9
2.3.4	Tamanho do componente gigante	10
2.3.5	Número de arestas do componente gigante	10
2.3.6	Coeficiente aglomeração (do inglês <i>Clustering coefficient</i>)	11
2.3.7	Assortatividade de grau	11
2.3.8	Comprimento médio do caminho mais curtos (do inglês <i>Average shortest path length</i>)	11
2.3.9	Diâmetro	11
2.3.10	Coeficiente <i>Rich club</i>	11
2.3.11	<i>“Matching index”</i>	12
2.4	A Lei de Potência	12

3	Redes Complexas	14
3.1	Introdução às Redes Complexas	14
3.2	Propriedades das Redes	15
3.3	Modelos de Redes Complexas	18
3.4	Estrutura de dados	21
3.5	Aplicações	21
4	Efeitos da seleção de limiar na topologia das redes de co-expressão gênica	23
4.1	Introdução	23
4.1.1	Redes de co-expressão gênica	24
4.1.2	Limiarização de redes de co-expressão gênica	25
4.2	Material e Métodos	27
4.2.1	Geração e análise das redes	27
4.2.2	Métodos de seleção de limiar	29
4.3	Resultados e Discussão	34
4.3.1	Aplicação dos critérios de Limiarização	37
4.3.2	Comparação entre métodos	41
4.3.3	Discussão Geral	53
4.4	Conclusões	53
5	Caracterização da rede de canais vasculares de ossos longos de camundongo	56
5.1	Introdução ao tecido ósseo	56
5.1.1	Classificação	57
5.1.2	Processos de ossificação	59
5.1.3	O Sistema Harvesiano	62
5.1.4	Alguns aspectos genéticos da osteogênese	63
5.1.5	Estrutura de redes em ossos	64
5.2	Material e Métodos	66
5.2.1	Aquisição de imagens tomográficas de osso de camundongo	68
5.2.2	Processamento das imagens tomográficas de osso de camundongo	70
5.2.3	Reconstrução do Esqueleto 3D a partir das Imagens de microCT	73
5.2.4	Reconstrução e aquisição da estrutura 3D	73
5.2.5	Visualização da Redes	74
5.2.6	Medidas Topológicas	76
5.2.7	Medidas Geométricas	77

5.2.8	Medidas de Dimensionalidade	78
5.3	Resultados e Discussão	80
5.3.1	Análise Topológica	80
5.3.2	Análises Geométricas	96
5.3.3	Análise de Dimensionalidade	96
5.3.4	Osso 5A dupla medula	101
5.4	Conclusões	103
6	Integração das Abordagens	104
6.1	Material e Métodos	104
6.2	Resultados e Discussão	105
6.2.1	Comparação entre métodos	107
7	Conclusões	109
7.1	Trabalhos Futuros	110

Lista de Figuras

1.1	Organização dos Capítulos da presente Tese.	4
2.1	Exemplos hipotéticos de correlação entre variáveis x_i e y_i . (a) Correlação linear perfeita e negativa ($\rho = -1$). (b) Correlação linear perfeita e positiva ($\rho = 1$). (c) Variáveis não correlacionadas ($\rho = 0$). Fonte: Elaborado pela autora.	7
2.2	Gráfico hipotético para demonstração da maximização da variância dos dados pelo PCA, e dos eixos ortogonais PCA1 e PCA2. Fonte: Elaborado pela autora.	8
2.3	Grafo hipotético mostrando uma estrutura de “triângulo”, ou seja, três vértices conectados entre si. A linha tracejada em azul claro mostra uma possível ligação que formaria uma estrutura de “triângulo”, mas que não existe, e portanto não “fecha” essa “tríade”. Uma miniatura do grafo sem essas possíveis ligações encontra-se à direita da figura. Os números dentro dos nós correspondem aos seus respectivos graus. Note que uma estrutura de triângulo não necessariamente ocorre somente na ponta de um grafo, nem ao menos somente no meio dele. Fonte: Elaborado pela autora.	9
2.4	Matriz com pesos (a) e o resultado de aplicar um limiar relativamente baixo sobre ela (T_1) (b) e um limiar relativamente alto (T_2) (c). Quando um limiar suficientemente alto é empregado, a rede pode quebrar-se em grupos de vértices que são desconectados (nós destacados em (c)). Fonte: Elaborado pela autora.	10
2.5	Gráficos hipotéticos para representação do comportamento de dados em distribuição em lei de potência (a) diante da mudança da escala linear para a logarítmica (b). Fonte: Elaborado pela autora.	13
3.1	Porção de um grafo hipotético contendo nove vértices e nove arestas. Fonte: Elaborado pela autora.	14
3.2	Exemplo do grafo randômico de Erdős e Renyi com $n = 40$ vértices, e probabilidade de criação de nós $p = .3$. Fonte: Elaborado pela autora.	18

3.3	Exemplo de rede de Watts e Strogatz com $n = 100$ vértices, cada nó conectado com os $k = 5$ vizinhos, probabilidade de embaralhamento de cada vértice $p = 0.4$. Fonte: Elaborado pela autora.	19
3.4	Exemplo de rede livre de escala de Barabási e Alberts, $n = 100$ vértices, e número de arestas a se conectar com outros vértices, a partir de um nó, igual a $m = 3$. Fonte: Elaborado pela autora.	20
3.5	(a) Exemplo de rede aleatória com estrutura de comunidades formada por 400 vértices com 4 comunidades. Fonte: Elaborado pela autora.	21
4.1	Etapas da metodologia proposta para seleção de limiar. Um conjunto de experimentos é selecionado a partir dos bancos de dados. Cada experimento contém os respectivos dados de expressão gênica (b). O coeficiente de correlação de Pearson é então calculado para todos os pares de genes em cada experimento (c), gerando uma respectiva matriz com pesos (d). Em seguida, um conjunto de limiares são aplicados aos valores de correlação (e), resultando em um conjunto de respectivas redes (f). Os critérios de determinação de limiar são aplicados às redes (g), para escolher a alternativa mais apropriada (h). T_1 , T_2 e T_3 representam valores candidatos de limiar. Fonte: Elaborado pela autora.	27
4.2	Ilustração da variação esperada de medidas de uma rede, à medida que o limiar aumenta. Valores menores de limiar devem resultar em valores mais estáveis, enquanto valores mais altos de limiar devem resultar em topologias mais instáveis. M representa uma medida hipotética, e T o valor de limiar. Fonte: Elaborado pela autora.	30
4.3	Grafos hipotéticos e sua entropia esperada. Os grafos (a) e (c) apresentam baixa entropia devido ao excesso e falta de arestas, respectivamente. A rede (b) pode ser vista como tendo uma estrutura mais complexa, refletindo um maior valor de entropia. Fonte: Elaborado pela autora.	32
4.4	Gráfico da distribuição de grau de redes hipotéticas (a). A linha preta tracejada representa uma função em lei de potência. O gráfico (b) representa a distância em que as diferentes distribuições de grau hipotético têm de uma distribuição em lei de potência. T representa o valor de limiar. Fonte: Elaborado pela autora.	33
4.5	Grau médio para diferentes valores de limiar. Quanto maior o grau médio da rede, mais conectados serão os nós. O grau médio é representado por $\bar{K}$. Um grau médio escolhido f_k possui um limiar correspondente T_k . Fonte: Elaborado pela autora.	34

4.6	Distribuições dos coeficientes de correlação de Pearson para cada conjunto de dados (linha azul). As linhas verticais correspondem aos limiares selecionados de acordo com cada critério adotado. " <i>Pearson dist.</i> "corresponde à distribuição de correlação de Pearson. Fonte: Elaborado pela autora.	35
4.7	Projeções de PCA de propriedades de rede obtidas usando valores de limiar distintos. O nome na parte superior de cada gráfico corresponde ao primeiro autor do conjunto de dados em que as redes foram geradas. Os pontos verdes indicam a rede correspondente ao limiar mais baixo. Pontos vermelhos indicam a transição detectada de propriedades estáveis a instáveis. Fonte: Elaborado pela autora.	37
4.8	Variação entre pontos consecutivos do PCA da Figura 4.7. O nome na parte superior de cada gráfico corresponde ao primeiro autor do conjunto de dados. Pontos vermelhos indicam a transição detectada das porções estáveis às instáveis. Fonte: Elaborado pela autora.	38
4.9	Valores de entropia obtidos para redes geradas por valores de limiar diferentes. Os pontos vermelhos indicam a entropia máxima alcançada por cada conjunto de dados. Fonte: Elaborado pela autora.	38
4.10	Tamanho do componente gigante da rede em função do valor de limiar. O tamanho é normalizado pelo número total de genes estudados em cada experimento. Os pontos coloridos indicam os tamanhos de componentes gigantes usados na análise, representados como f_s . S representa o tamanho do componente gigante. Fonte: Elaborado pela autora.	40
4.11	Estatística Kolmogorov-Smirnov comparando a distribuição de grau da rede a uma função em lei de potência. Fonte: Elaborado pela autora.	41
4.12	Grau médio das redes obtidas por diferentes valores de limiar. Os pontos indicam o grau médio escolhido para analisar este critério, representado como f_k . Fonte: Elaborado pela autora.	42

- 4.13 (a) Grafo com pesos representando a correlação de Pearson entre os resultados dos diferentes critérios. O tamanho do nó é dado pela média do logaritmo da correlação de Pearson entre o critério e todos os outros critérios. As arestas mais grossas correspondem a valores de correlação mais elevados. Grafos com pesos representando (b) $\tilde{\alpha}$ e (c) $\tilde{\beta}$ obtidos também são mostrados. *T variation* representa o critério de Máxima variação das medidas topológicas; *T entropy* é o critério de máxima entropia; *T comp. 0.4* representa o critério de Tamanho do componente gigante fixado em 40% do número total de nós, com um significado análogo para *T comp. 0.6* e *T comp. 0.8*; *T deg. 100* representa o grau médio específico (100), e *T deg. 1000* grau médio específico (1000); *T power – law* representa o critério de distribuição de grau em lei de potência (*power-law*). As bordas vermelhas indicam critérios que divergiram mais entre si, portanto, com uma correlação Pearson baixa ou alta $\tilde{\alpha}$ ou $\tilde{\beta}$. O nó *T variation* foi excluído em (b) e (c) para não mascarar as diferenças entre os outros métodos. Fonte: Elaborado pela autora. 42
- 4.14 Gráficos de dispersão entre os valores de limiar escolhidos pelos diferentes critérios. Cada ponto corresponde a um conjunto de dados. Consulte o texto principal para o significado dos nomes dos eixos. ρ representa o coeficiente de correlação de Pearson entre os valores do gráfico, α corresponde ao coeficiente angular, e β o coeficiente linear. Consulte a legenda da Figura 4.13 para o significado de cada nome. Fonte: Elaborado pela autora. 45
- 4.15 Correlação entre os critérios: *Máxima entropia versus Tamanho do componente gigante 60%* (a), e *Distribuição em lei de potência versus Grau médio específico 1000* (b). A letra ρ corresponde ao coeficiente de correlação de Pearson, α ao coeficiente angular e β ao coeficiente linear. A linha vermelha indica reta que melhor se ajusta aos quadrados mínimos. Fonte: Elaborado pela autora. 46
- 4.16 Significado dos coeficientes angular (α) e linear (β). A linha vermelha representa uma regressão linear dos pontos azul. β é dado pelo valor de y onde a linha vermelha intercepta o eixo y . θ é o ângulo entre a linha vermelha e o eixo x . α é calculado como $\alpha = \tan(\theta)$. Fonte: Elaborado pela autora. 47
- 4.17 Diferença entre os valores de limiar escolhidos por dois dados critérios. Esta comparação considerou os critérios que apresentaram maiores divergências na Figura 4.13. (a) *máxima entropia X tamanho do componente gigante 40%*, (b) *máxima entropia X distribuição em lei de potência*, (c) *tamanho do componente gigante 40% X tamanho do componente gigante 80%*, (d) *tamanho do componente gigante 80% X distribuição em lei de potência*, (e) *tamanho do componente gigante 80% X grau médio específico (1000)*, (f) *grau médio específico (1000) X grau médio específico (100)*. Fonte: Elaborado pela autora. 49

4.18	(a) PCAs das medidas obtidas para cada critério de limiarização. Note que cada ponto corresponde a uma respectiva base de dados. Consulte a legenda da Figura 4.13 para o significado de cada nome. (b) Matriz contendo as posições relativas dos experimentos ao grupo conservado (mais próximo a mais, da esquerda para a direita), cada linha corresponde a um critério. Fonte: Elaborado pela autora.	52
5.1	Esquema do centro de ossificação primária do processo de ossificação endocondral. A cartilagem hialina é um molde para o osso longo que será formado. (b) A proliferação dos condrócitos e sua conseguinte hipertrofia, iniciam a formação do centro de ossificação primário. (c) Esses condrócitos hipertróficos são responsáveis pela secreção do fator de crescimento endotelial vascular, para induzir o brotamento dos vasos sanguíneos a partir do pericôndrio. Depois desses eventos, ocorre a calcificação da matriz do centro de ossificação e a indução da apoptose dos condrócitos hipertróficos. Os vasos sanguíneos que chegaram a partir do pericôndrio formam o broto periosteal, que se ramifica em direção às diáfises. Fonte: elaborada pela autora, com base em Kierszenbaum [1].	60
5.2	Esquematização do processo de ossificação secundária dos ossos longos. (a) O colar periosteal (Figura 5.1b) estende-se ao longo da diáfise óssea e vasos sanguíneos infiltram-se nas epífises para formar os centros de ossificação secundários. (b) Esses centros desenvolvem-se e vão preenchendo as epífises. (c) As placas epifisárias de crescimento são substituídas pela linha epifisária, marcando o final do crescimento em comprimento do osso. Fonte: elaborada pela autora, com base em Kierszenbaum [1].	61
5.3	Esquematização do sistema Harvesiano. Fonte: elaborada pela autora.	62
5.4	Representação da conversão da rede dos canais vasculares adquiridos por microtomografia (a) para a rede complexa (b). As medidas ($m_{1,1}, m_{1,2}, m_{1,3}...$) representam o conjunto de medidas obtidas para um determinado nó. As reticências ao redor da figura mostram que esses conjuntos de medidas serão obtidos para todos os nós da rede. Fonte: elaborada pela autora.	67
5.5	Workflow com os passos de aquisição de imagens, processamento, e aquisição da estrutura 3D. Fonte: elaborada pela autora.	67
5.6	Legenda das partes dos fêmures (a) e úmeros (b) usada nas demais seções. As partes "b"s correspondem necessariamente à porção <i>distal</i> e as respectivas partes opostas correspondem à porção <i>proximal</i> . Fonte: elaborada pela autora.	69

5.7	Projeção lateral do osso úmero 3B. Essa figura mostra as partes nas quais o osso foi dividido nas análises. Destaca-se na imagem a <i>tuberosidade deltóide</i> , estrutura que auxiliou no posicionamento e orientação das amostras. A estrutura branca na parte superior da figura é a massa de sustentação para o apoio no microCT. Fonte: elaborada pela autora.	69
5.8	Exemplificação do método de segmentação pela média. O pixel em questão está destacado em azul, e tem valor de 56 dentro da escala de cinza. A média da vizinhança (destacada em laranja) é de 109,375 $((7 + 78 + 156 + 59 + 48 + 198 + 255 + 74)/8)$, destacado em roxo), maior que 56. Portanto, o pixel de valor 56 passa a ter valor 0. O novo valor de pixel (0) é armazenado em uma matriz binária, e a "janela" de 3x3 <i>pixels</i> move-se para o próximo pixel, para que sua vizinhança seja analisada. Fonte: elaborada pela autora.	71
5.9	Etapas de processamento para todas as imagens obtidas por microCT. Fonte: elaborada pela autora.	72
5.10	Uma forma simples (a) e seu respectivo SAT (b). Fonte: elaborada pela autora, adaptado de Costa e Marcondes [2]).	73
5.11	Representação dos pixels de uma imagem e seus vizinhos. Na figura, os quadrados em branco representam os pixels do fundo da imagem, e os pixels em preto representam o objeto da imagem. O número em branco informa quantos vizinhos um determinado pixel possui. Fonte: elaborada pela autora.	74
5.12	Grafo representado à esquerda da figura, com seus nós numerados de 1 a 8. Respectiva matriz de adjacência à direita. Como o grafo não é direcionado, essa matriz é simétrica. Fonte: elaborada pela autora.	74
5.13	Visualização da rede de canais vasculares (osso 1A femur, parte b). Note que as posições geográficas dos canais estão mantidas. Existe uma concentração interessante de <i>hubs</i> na parte lateral do osso que foi evidenciada por esse tipo de visualização. As cores dos nós e arestas representam os graus dos nós (conforme a legenda ao lado de cada figura). Fonte: elaborada pela autora.	75
5.14	Visualização da rede de canais vasculares. Essas partes de osso correspondem à continuação da peça mostrada na Figura 5.13. Fonte: Elaborado pela autora.	76
5.15	Representação esquemática de uma porção da rede de canais vasculares e o respectivo esqueleto no interior dela. O círculo e azul representa uma área de seção transversal do canal vascular. Os traços representam arestas e os círculos representam nós. Obs.: nesse esquema, os nós estão representados por círculos para caráter didático; no esqueleto original os nós são bifurcações das arestas. Fonte: elaborada pela autora.	78

5.16	Representação esquemática da rede de canais vasculares e o respectivo esqueleto no interior dela. As medias d_{max} e d_{min} correspondem ao diâmetro máximo e mínimo de arestas; d_t corresponde à distância topológica; d_e é a distância entre extremidades; e s é o comprimento de arco. Fonte: elaborada pela autora.	78
5.17	Gráficos com as distribuições de graus das redes vasculares. Fonte: Elaborado pela autora. .	81
5.18	Ilustração da ponta de um grafo hipotético com a discriminação do grau do nó. Fonte: Elaborado pela autora.	82
5.19	Gráficos com as distribuições de transitividade das redes vasculares. Fonte: Elaborado pela autora.	84
5.20	Gráficos com as distribuições da centralidade de interposição das redes vasculares. Fonte: Elaborado pela autora.	85
5.21	Distribuição das medidas topológicas dos fêmures e tíbias. Em ordem, as medidas em questão nos gráficos são: número de vértices, número de arestas, razão <i>vertices/arestas</i> , transitividade indireta, transitividade média local, assortatividade, média dos menores caminhos, diâmetro, coeficiente <i>Rich Club</i> , <i>índice matching</i> . Fonte: Elaborado pela autora.	91
5.22	Distribuição das medidas topológicas dos úmeros. Vide legenda da Figura 5.21 para o significado das siglas. Fonte: Elaborado pela autora.	92
5.23	PCA das medidas topológicas das redes vasculares. Fonte: Elaborado pela autora.	93
5.24	PCA da Figura 5.23 com os pares de ossos esquerdos e direitos unidos por uma reta. Os ossos sem par estão discriminados, assim como as tíbias da análise. Fonte: Elaborado pela autora.	94
5.25	PCA da Figura 5.23 com os pontos maiores e cores diferentes: Em verde, são marcados os fêmures; em amarelo, estão os úmeros; e em vermelho, as tíbias. Fonte: Elaborado pela autora. .	95
5.26	PCA com as cores da Figura 5.25 com destaque para os grupos dos fêmures, úmeros e tíbias. Fonte: Elaborado pela autora.	95
5.27	Gráficos com as distribuições do tamanho das arestas das redes vasculares. Fonte: Elaborado pela autora.	97
5.28	Gráficos com as distribuições do tamanho médio dos raios das arestas das redes vasculares. Fonte: Elaborado pela autora	98
5.29	Gráfico da distribuição da dimensionalidade das redes sistema Harvesiano. Cada linha corresponde a uma das partes dos ossos (destacadas na Figura 5.7). O eixo y contém as dimensões que caracterizam cada conjunto de nós (eixo x). Fonte: Elaborado pela autora.	99
5.30	Divisão dos nós que tem inicialmente duas e três dimensões (2D e 3D). Fonte: Elaborado pela autora.	100

5.31	Projeção lateral da tomografia do osso Tíbia 5A. Fonte: Elaborado pela autora.	101
5.32	Visualização da parte “b” do osso 5A. Fonte: Elaborado pela autora.	102
6.1	Valores de entropia obtidos para redes geradas por valores de limiar diferentes. Os pontos vermelhos indicam a entropia máxima alcançada por cada conjunto de dados. Fonte: Elaborado pela autora.	105
6.2	Tamanho do componente gigante da rede em função do valor de limiar. O tamanho é normalizado pelo número total de genes estudados em cada experimento. Os pontos coloridos indicam os tamanhos de componentes gigantes usados na análise, representados como f_s . S representa o tamanho do componente gigante. Fonte: Elaborado pela autora.	106
6.3	Grau médio das redes obtidas por diferentes valores de limiar. Os pontos vermelhos indicam o grau médio escolhido para analisar este critério. Fonte: Elaborado pela autora.	106
6.4	Estatística Kolmogorov-Smirnov comparando a distribuição de grau da rede a uma função em lei de potência. Fonte: Elaborado pela autora.	107

Capítulo 1

Introdução

As Redes Complexas mostraram-se ser estruturas interessantes para representação, modelagem e caracterização de diversos sistemas complexos. Essa teoria encontrou aplicações em diversas áreas do conhecimento, como física, química, engenharia, ciências sociais, biologia, entre outras [3]. Por se tratar de uma teoria relativamente nova, ainda existem alguns aspectos do uso dos conceitos de Redes Complexas pouco elucidados, como a questão da limiarização de relacionamentos graduados (com peso) para obtenção de uma rede binária de conexões. O processo de limiarizar uma rede é desejável para manter-se somente as conexões mais relevantes de acordo com um dado critério.

Outra temática interessante refere-se à comparação entre redes complexas derivadas de sistemas diversos, em busca de encontrar ou não semelhança entre elas. Na grande área da biologia, esta questão é particularmente importante, porque é base da busca de analogias estruturais e funcionais diante dos mais variados sistemas biológicos. O presente trabalho de doutorado situa-se na interseção destes dois problemas.

Na primeira parte do trabalho, abordamos o problema da limiarização de redes de co-expressão gênica, sugerindo e comparando cinco abordagens diferentes para seleção de limiar. Abordagens semelhantes foram sistematicamente utilizadas na ciência de Redes Complexas, incluindo-se redes caracterizadoras de clima [4], transporte [5], interações sociais [6], entre outras. Cada um desses métodos considera um critério biologicamente motivado para eleger um limiar potencialmente adequado. Um conjunto de 21 experimentos de microarray de diferentes grupos biológicos foi utilizado para investigar o efeito da aplicação dos cinco critérios de limiarização propostos. Para cada experimento, utilizamos o coeficiente de correlação de Pearson para mensurar a relação entre cada par de genes, e cada um dos cinco critérios propostos foi então aplicado para selecionar um valor de limiar respectivo, transformando essas redes graduadas em binárias. Cada critério gera uma rede a partir de uma base de dados ($5 * 21 = 105$). Os efeitos dessas abordagens de limiarização na topologia das redes resultantes foram comparados usando várias medidas.

Verificamos que dependendo da natureza do banco de dados, o impacto nas propriedades topológicas pode ser grande. No entanto, um grupo de bancos de dados foi verificado como sendo afetado de forma semelhante pela maioria dos critérios considerados. Com base em tais resultados, sugerimos que quando as redes geradas apresentam comportamento similar nas medidas, o método de limiar pode ser escolhido com maior liberdade. Se as redes geradas são marcadamente diferentes, o método de limiar que melhor se adequa aos interesses de cada pesquisa específica representa uma escolha razoável.

A segunda parte do trabalho destina-se à caracterização da rede do sistema Haversiano de ossos de camundongo. Essa rede vascular é responsável pela comunicação intracelular no tecido ósseo, transportando os elementos que fazem a manutenção dos ossos. Ela é formada durante o processo de ossificação endocondral, quando vasos sanguíneos invadem o espaço ocupado por cartilagem, ramificam-se, espalham-se e finalmente, adquirem a estrutura de uma rede, sendo posteriormente delimitados por canais formados por matriz óssea [1]. A estrutura matricial extracelular óssea vem sendo extensivamente estudada, no entanto, há uma falta de conhecimento quantitativo acerca dos seus canais vasculares, chamados Sistema de Havers ou Sistema Haversiano. Utilizando abordagens de aquisição e processamento de imagens, analisamos o sistema Haversiano em três tipos diferentes de ossos longos de camundongos (*Mus musculus*) utilizando conceitos robustos e medidas quantitativas derivadas da Teoria das Redes Complexas. Esse estudo visa a avaliação da eficiência organizacional dessa rede biológica em relação à comunicação e transporte. As análises mostraram que a rede do sistema Haversiano de ossos longos de camundongo são regularmente organizadas, e não há uma diferença considerável entre as porções distintas de um mesmo osso. Notamos que os ossos são indistinguíveis em relação às propriedades topológicas das redes, sugerindo que a arquitetura topológica do sistema vascular pode ser independente do tipo ósseo. Em relação às propriedades geométrica das redes, nota-se uma variação mais pronunciada, sugerindo que a demanda biológica de transporte pode variar em relação à posição no mesmo osso, e entre ossos diferentes. Os resultados aqui apresentados sugerem um mecanismo universal sob a organização do sistema Haversiano. Devido ao fato de as redes do sistema Haversiano possuírem pesos relacionados à espessura dos canais (neste estudo representado pelo raio médio das arestas), utilizamos e comparamos, nas redes vasculares de ossos, os métodos de limiarização aqui propostos para as redes de co-expressão gênica.

Concluindo estes desenvolvimentos, realizamos uma comparação estrutural dos dois tipos de redes obtidas, ou seja, de co-expressão gênica e do Sistema Haversiano, e os resultados obtidos indicam que a rede vascular deve ser analisada com todos os seus canais, ou seja, sem sofrer o processo de limiarização.

1.1 Motivação e Objetivos

O volume de dados em Bancos de Dados (BDs) públicos é grande dentro do campo da Genética e Biologia Molecular, e o uso da teoria das Redes Complexas para analisar problemas reais é incentivado por essa disponibilidade de dados. Aplicar um limiar em uma rede com peso é desejável para conservar-se as conexões que são significativas de acordo com algum critério razoável, eliminando as ligações que não cumprem esse propósito. Além disso, não existe uma definição estabelecida para diversas medidas topológicas de redes com peso [7]. Dessarte, a limiarização de redes complexas é uma abordagem importante no estudo dos sistemas. No entanto, ainda há pouca atenção ao desenvolvimento de métodos matemáticos para selecionar um valor de limiar apropriado ao estudar redes de co-expressão de genes.

O tecido ósseo é o principal componente do esqueleto, servindo de suporte para as porções “moles” do corpo e como proteção para órgãos vitais, por exemplo, órgãos contidos na caixa craniana, torácica e canal raquidiano. Além disso, o esqueleto aloja a medula óssea hematopoiética, proporciona apoio aos músculos esqueléticos, e constitui um sistema de alavancas que amplia as forças geradas na contração muscular [8]. A disposição dos canais vasculares que percorre os ossos é de extrema importância para a manutenção desse tecido, e se o transporte sanguíneo não for eficiente, as células podem não ter sua demanda nutricional suprida. O estudo dos processos de crescimento, remodelagem e organização estrutural é importante em aplicações clínico-terapêuticas [9], e a estrutura dos canais vasculares dos ossos possibilita sua descrição também segundo a teoria das Redes Complexas. Com essa ferramenta, visamos investigar diferentes aspectos topológicos, geométricos e dimensionais da estrutura vascular dos ossos.

Em vista disso, os objetivos desta Tese de Doutorado são:

- A proposição de uma abordagem empírica, mas bem embasada, ao problema da limiarização de redes de co-expressão gênica, acompanhada de análises estatísticas e validação. Em particular, buscamos caracterizar redes distintas de co-expressão gênica e comparar o comportamento de um grupo de medidas topológicas de redes, sob diferentes valores de limiar escolhidos por métodos guiados por diferentes critérios.
- A caracterização do sistema Haversiano de ossos de camundongo, visando o detalhamento topológico, geométrico e dimensional dessa estrutura, sob a luz da Teoria das Redes Complexas. Também pretende-se verificar como esta importante rede biológica se comporta frente aos métodos de limiarização desenvolvidos na primeira parte deste trabalho.

1.2 Organização da Tese

A presente Tese de Doutorado é organizada da seguinte maneira descrito no parágrafo abaixo, e o grafo da Figura 1.1 ilustra a disposição dos Capítulos.

O documento inicia-se com um Capítulo destinado à apresentação de conceitos básicos que serão utilizados ao longo da tese. Nele são expostas medidas estatísticas e medidas topológicas de rede. O Capítulo seguinte trata-se de uma introdução às Redes Complexas, com conceituação de propriedades dessas estruturas, explanação dos seus modelos mais utilizados, assim como estruturação de dados e alguns exemplos de aplicações. Em seguida, inicia-se o Capítulo intitulado “Efeitos da seleção de limiar na topologia das redes de co-expressão gênica”, o qual vai tratar da caracterização de algumas redes de co-expressão gênica e da comparação do comportamento de suas medidas, após a limiarização dessas redes por diferentes valores, escolhidos por métodos guiados por diferentes critérios. O próximo Capítulo tem por título “Caracterização da rede de canais vasculares de ossos longos de camundongo”, e ele é o responsável pela caracterização do sistema Haversiano ósseo, sob o ponto de vista da Teoria das Redes Complexas. Um consequente Capítulo integrador tem a função final de assimilar as duas partes da Tese. Por fim, esta Tese de Doutorado é encerrada com o Capítulo de conclusões acerca do trabalho desenvolvido, e perspectivas para trabalhos futuros.

Figura 1.1: Organização dos Capítulos da presente Tese.

Capítulo 2

Conceitos Básicos

Este capítulo destina-se à apresentação de conceitos básicos que serão utilizados ao longo da tese.

2.1 Medidas estatísticas

Para o critério de limiarização **Máxima variação das propriedades topológicas**, descrito no Capítulo 4, Seção 4.2.2, e para as **análises topológicas** das redes vasculares de ossos, no Capítulo 5 Seção 5.3.1, nós consideramos as propriedades topológicas de cada rede descritas na Seção 2.3. As medidas de *grau*, *transitividade* e *entropia* são calculadas para todos os nós da rede, então para elas, consideramos algumas estatísticas desse número, como o mínimo, máximo, média, mediana, desvio padrão e entropia, as quais serão detalhadas nos itens a seguir.

O coeficiente de correlação de Pearson e a análise de componentes principais são os últimos procedimentos estatísticos tratados na Seção 2.1.

2.1.1 Média aritmética

A média aritmética de determinado conjunto de valores é obtida dividindo-se a soma dos valores pelo número dos mesmos. A equação da média aritmética é mostrada abaixo.

$$\bar{x} = \frac{1}{n} * \sum_{i=1}^n x_i \quad (2.1)$$

2.1.2 Mediana

A mediana é o valor que separa em dois a metade maior e a metade menor de uma amostra.

2.1.3 Desvio padrão

O desvio padrão é uma medida de dispersão em torno da média populacional de uma determinada variável. Para uma população finita, o desvio padrão é calculado segundo a equação abaixo:

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.2)$$

2.1.4 Entropia

Na Teoria de Informação, sistemas incorporam um transmissor, um canal e um receptor. O primeiro destes produz mensagens que são transmitidas através do canal, que então modifica a mensagem, e então a mensagem chega ao receptor. A entropia de Shannon é o valor médio de informação esperado nas mensagens [10].

Uma forma de implementar tal medida para uma rede, é calcular a entropia de um conjunto de parâmetros de vértice. Isso pode ser feito da seguinte maneira. Dado um conjunto de M medidas, definidas para cada nó da rede, um histograma M -dimensional é calculado. Esse procedimento define a probabilidade da função de massa $P(x_1, x_2, \dots, x_n)$, onde x_n representa a n -ésima medida considerada. Então, a entropia de Shannon dessa função é calculada, e pode ser escrita como:

$$H(X) = - \sum_{i=1}^N P(x_i) \log_b P(x_i) \quad (2.3)$$

onde b é a base usada no logaritmo, $P(x)$ representa a função de massa de probabilidade de x , e N é o número de possíveis valores de x_i .

2.1.5 Coeficiente de Correlação de Pearson (CCP)

O Coeficiente de Correlação de Pearson (CCP) é um método estatístico amplamente utilizado no meio científico. Esse coeficiente foi elucidado por Karl Pearson, a partir de uma ideia introduzida por Francis Galton [11]. O CCP mede a relação linear entre duas variáveis aleatórias, além da direção dessa correlação (positiva ou negativa) [12].

O CCP é calculado a partir da equação:

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} * \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{cov(X, Y)}{\sqrt{var(X) * var(Y)}} \quad (2.4)$$

onde $x_1, x_2, \dots, x_n$ e $y_1, y_2, \dots, y_n$ são os valores das variáveis que estão sendo correlacionadas, e

$$\bar{x} = \frac{1}{n} * \sum_{i=1}^n x_i \quad (2.5)$$

e

$$\bar{y} = \frac{1}{n} * \sum_{i=1}^n y_i \quad (2.6)$$

são as médias aritméticas dessas variáveis e n é o número de variáveis.

Esse coeficiente fornece um valor entre -1 e 1 , sendo que -1 indica uma anticorrelação linear perfeita, ou seja, se uma aumenta a outra diminui; 1 corresponde a uma sobreposição linear perfeita positiva entre as duas variáveis; e 0 caracteriza ausência de correlação, isto é, as duas variáveis independem linearmente uma da outra. No entanto, uma correlação $\rho = 0$ não descarta a possibilidade de dependência não linear.

Figura 2.1: Exemplos hipotéticos de correlação entre variáveis x_i e y_i . (a) Correlação linear perfeita e negativa ($\rho = -1$). (b) Correlação linear perfeita e positiva ($\rho = 1$). (c) Variáveis não correlacionadas ($\rho = 0$). Fonte: Elaborado pela autora.

Quando o CCP é usado para correlacionar os valores de expressão gênica de um dado experimento global, os valores de ρ calculados podem ser entendidos como uma rede de co-expressão com peso contendo $N(N - 1)/2$ relações entre os N genes considerados.

2.1.6 Análise de Componentes Principais (do inglês, *Principal Component Analysis - PCA*)

A análise de Componentes Principais (do inglês *Principal Component Analysis - PCA*) é um procedimento estatístico que usa uma transformação ortogonal para converter um conjunto de variáveis, que são possivelmente correlacionadas, em um conjunto em menor dimensão de variáveis não correlacionadas chamadas *componentes principais* [13].

A Figura 2.2 mostra um gráfico de $y = f(x)$ com uma dispersão de pontos. Os eixos originais do gráfico em questão não são suficientes para maximizar a distribuição da variação dos pontos, contudo, a inserção

do eixo denominado *PCA1* é disposta de modo a promover essa maximização. O primeiro componente principal guarda a maior variância possível dos dados, e cada componente sucessivo tem a maior variância possível, limitado a ser ortogonal ao componente anterior (eixo *PCA2* da Figura 2.2)

Figura 2.2: Gráfico hipotético para demonstração da maximização da variância dos dados pelo PCA, e dos eixos ortogonais PCA1 e PCA2. Fonte: Elaborado pela autora.

2.2 Normalização *z-score*

A normalização por *z-score* de um valor bruto é representada por:

$$z = \frac{x - \mu}{\sigma} \quad (2.7)$$

onde μ é a média e σ é o desvio padrão da população. O valor absoluto de z representa a distância entre o valor bruto x e a média da população em unidades do desvio padrão. z é negativo quando a pontuação em bruto está abaixo da média, positiva quando acima.

2.3 Medidas topológicas de Rede

Para o critério de limiarização **Máxima variação das propriedades topológicas**, descrito no Capítulo 4, Seção 4.2.2, e para as **análises topológicas** das redes vasculares de ossos, no Capítulo 5 Seção 5.3.1, nós consideramos as seguintes propriedades topológicas de cada rede:

2.3.1 Estatísticas de grau

O grau de um vértice é o seu número de arestas incidentes [14]. Consideramos algumas estatísticas desse número, como o mínimo, máximo, média, mediana, desvio padrão e entropia.

2.3.2 Transitividade local

A transitividade local mede a fração de todos os “triângulos” entre os possíveis (Figura 2.3). Esses possíveis “triângulos” são identificados pelo número de “tríades” (duas arestas com um vértice compartilhado). Portanto, a transitividade local mede a probabilidade de dois vizinhos de um vértice serem conectados também [15]. Também consideramos algumas estatísticas desse número, como o mínimo, máximo, média, mediana, desvio padrão e entropia descritos no item anterior (Seção 2.3.1).

Figura 2.3: Grafo hipotético mostrando uma estrutura de “triângulo”, ou seja, três vértices conectados entre si. A linha tracejada em azul claro mostra uma possível ligação que formaria uma estrutura de “triângulo”, mas que não existe, e portanto não “fecha” essa “tríade”. Uma miniatura do grafo sem essas possíveis ligações encontra-se à direita da figura. Os números dentro dos nós correspondem aos seus respectivos graus. Note que uma estrutura de triângulo não necessariamente ocorre somente na ponta de um grafo, nem ao menos somente no meio dele. Fonte: Elaborado pela autora.

2.3.3 Centralidade de interposição (do inglês, *Betweenness centrality*)

Para calcular esta medida para um determinado vértice, primeiro contamos o número de caminhos mais curtos entre dois vértices que passam pelo vértice em questão. Então, o resultado é dividido pelo número total de caminhos mais curtos entre os dois vértices. Em seguida, este cálculo é repetido para todos os pares de vértices na rede, a soma dos valores obtidos define a *Centralidade de interposição* do vértice [16, 17, 14]. Consideramos as estatísticas: número mínimo, máximo, média aritmética, mediana, desvio padrão e

entropia.

2.3.4 Tamanho do componente gigante

O número de vértices no componente gigante tende a aumentar com o grau médio da rede e fornece uma indicação interessante sobre a uniformidade das conexões ao longo de sua estrutura [3].

Uma rede com peso (Figure 2.4a) é usada para representar a relação entre variáveis: a largura das arestas indica a força da associação entre cada par de variáveis. Após aplicar um certo valor de limiar (T_1 e T_2 na Figura 2.4), algumas arestas serão removidas, e o grafo resultante se torna sem peso (binário). Se o valor de limiar é relativamente baixo, somente as ligações mais fracas serão deletadas (T_1 resultando na rede 1 na Figura 2.4b). De maneira oposta, se o limiar escolhido é relativamente alto, somente as arestas mais fortes serão preservadas (T_2 resultando na rede 2 na Figura 2.4c). Quando o limiar aplicado é alto suficiente, algumas arestas que conectam grupos de vértices podem desaparecer, fazendo com que a rede “quebre-se” em grupos de vértices que não são conectados (grupos destacados na Figura 2.4).

Figura 2.4: Matriz com pesos (a) e o resultado de aplicar um limiar relativamente baixo sobre ela (T_1) (b) e um limiar relativamente alto (T_2) (c). Quando um limiar suficientemente alto é empregado, a rede pode quebrar-se em grupos de vértices que são desconectados (nós destacados em (c)). Fonte: Elaborado pela autora.

2.3.5 Número de arestas do componente gigante

Número de arestas que conectam os vértices no componente gigante [3].

2.3.6 Coeficiente aglomeração (do inglês *Clustering coefficient*)

O coeficiente de aglomeração é uma medida da tendência de vértices para se agruparem em uma rede [18] [19]. Assim como para a transitividade local (Seção 2.3.2), essa medida é baseada em “triângulos”(Figura 2.3). Um triângulo, portanto, inclui três “tríplas”conectadas, uma centrada em cada um dos nós. O coeficiente de aglomeração é o número de “tríplas”conectadas dividido pelo número total de “tríplas”(abertas e fechadas). Esta medida pode ser vista como uma transitividade global da rede.

2.3.7 Assortatividade de grau

Essa medida relaciona-se com o conceito de “homofilia”. Assim como a definição em termos sociais de “homofilia” é a tendência de um pessoa associar-se à outrem com propriedades similares, sua definição sob a Teoria das redes complexas é: “tendência de um nó se conectar a outro com propriedades similares”, seja em relação ao grau, coeficiente de agrupamento, localização etc.

Na presente tese, usaremos a *assortatividade de grau*, ou seja, a que expressa a preferência de vértices em uma rede ligar-se a outros que têm grau similar [20].

2.3.8 Comprimento médio do caminho mais curtos (do inglês *Average shortest path length*)

O comprimento médio dos caminhos mais curtos é um conceito em topologia de rede que é definido como o número médio de passos ao longo dos caminhos mais curtos para todos os possíveis pares de vértices da rede [21]. Essa é uma medida da eficácia do transporte de informações em uma rede [22].

2.3.9 Diâmetro

O diâmetro de uma rede é o maior dos menores número de vértices que devem ser percorridos para viajar de um vértice para outro. Os caminhos com retorno (“*backtrack*”), desvio ou “*loop*” não são considerados [23].

2.3.10 Coeficiente *Rich club*

Assim como a assortatividade mencionada anteriormente, o coeficiente “*rich-club*” trata de homofilia dos graus de uma rede. Esse coeficiente é a medida do quanto os vértices “bem conectados” ligam-se uns com os outros [24]. As redes que têm um coeficiente “*rich-club*” relativamente alto têm muitas conexões entre vértices de alto grau.

2.3.11 “Matching index”

O “Matching index” pode ser atribuído a cada aresta em uma rede para quantificar a semelhança entre a conectividade dos dois vértices adjacentes a essa aresta [25].

2.4 A Lei de Potência

Uma lei de potência é uma função matemática $y = f(x)$ que expressa uma relação onde uma variável é a potência da outra [26]. Ou seja, lei de potência entre dois escalares x e y é tal que a relação pode ser escrita na forma:

$$y = ax^{-k} \quad (2.8)$$

onde o coeficiente de proporcionalidade a , e o expoente k , são constantes [27].

Um exemplo prático sobre a lei de potência é a área de um quadrado em termos do comprimento do seu lado. Se duplicarmos esse comprimento, multiplicamos a área por um fator de quatro. Semelhantemente, se duplicarmos o comprimento da lateral de um cubo, o volume do cubo é multiplicado por um fator de oito [26].

Quando um gráfico de distribuição em lei de potência tem seus eixos convertidos para uma escala logarítmica, a curva passa a ser representada por uma reta (Figura 2.5), pois:

$$\log(y) = k\log(x) + \log(a) \quad (2.9)$$

é análoga à equação da reta:

$$\bar{y} = k\bar{x} + \bar{a} \quad (2.10)$$

e portanto, é uma reta.

As leis de potência são importantes na ciência porque elas revelam regularidades subjacentes nas propriedades dos sistemas. Os sistemas complexos apresentam propriedades correlacionadas em diferentes escalas [26].

As leis de potência são amplamente usadas no campo da física, biologia, ciências planetárias, economia e finanças, informática, demografia e ciências sociais. Por exemplo, as distribuições dos tamanhos de cidades, terremotos, incêndios florestais, alargamentos solares, crateras da lua e fortunas pessoais das pessoas parecem seguir as leis de potência [28, 29, 30, 31].

Figura 2.5: Gráficos hipotéticos para representação do comportamento de dados em distribuição em lei de potência (a) diante da mudança da escala linear para a logarítmica (b). Fonte: Elaborado pela autora.

Capítulo 3

Redes Complexas

3.1 Introdução às Redes Complexas

Uma rede pode ser definida informalmente como uma "coleção de vários pares de pontos unidos por linhas". Usando termos próprios da área de Redes Complexas para aprimorar essa definição, temos que "redes são uma coleção de *nós* (também chamados de *vértices*) unidos par a par por meio de *arestas*", e para representar esses sistemas, podemos utilizar uma estrutura matemática denominada *grafo*.

Figura 3.1: Porção de um grafo hipotético contendo nove vértices e nove arestas. Fonte: Elaborado pela autora.

As redes são representações simplificadas das interações dos componentes de um sistema, reduzindo-o a uma estrutura abstrata para capturar o básico de padrões de conexão. Os vértices e arestas em uma rede podem ser nomeados com informação adicional, como nomes ou pesos, para capturar mais detalhes do sistema, contudo bastante informação é frequentemente perdida no processo de redução de um sistema completo para a representação na forma de rede. Essa característica apresenta vantagens e desvantagens,

mas é amplamente aceito que as redes são formas poderosas de representação de padrões de conexões ou interações entre as partes de um dado sistema [3].

Ao longo dos anos, muitos campos da ciência contribuíram com o desenvolvimento de ferramentas (matemáticas, computacionais e estatísticas) para o estudo de redes. Muitas dessas ferramentas partem de uma representação simples de rede, que após alguns cálculos podem prover alguma informação útil sobre a mesma, como qual é o vértice melhor conectado, ou o comprimento de um caminho de um vértice a outro. Outras ferramentas mais complexas podem modelar a rede para fazer previsões matemáticas sobre processos que envolvem essa rede, tal qual a forma que uma doença espalhar-se-á numa comunidade. Por aplicar-se às formas abstratas das redes, essas ferramentas podem ser virtualmente aplicadas a quase qualquer sistema representado por uma rede, contudo, as medidas e cálculos que melhor se aplicam a uma determinada rede são particulares ao sistema em questão, e ainda dependem das perguntas que se desejam responder. [3].

As Redes Complexas podem ser estudadas sob alguns aspectos diferentes:

1. A natureza dos componentes individuais;
2. A natureza das conexões ou interações;
3. O padrão das conexões entre os componentes.

O presente trabalho tem maior interesse no terceiro aspecto, pois focaremos no comportamento dos sistemas representados pelas redes, traduzido por aspectos de sua topologia

3.2 Propriedades das Redes

Como será detalhado em seções posteriores, muitos sistemas podem ser representados por redes. Então, quando se escolhe um sistema a ser estudado, parte-se para a etapa de coleta de dados sobre o objeto de interesse. A partir desse momento, deve-se pensar no que esses dados podem dizer sobre a forma e função do sistema que a rede representa, quais são as suas propriedades que podem ser medidas ou modeladas e como essas propriedades relacionam-se aos problemas práticos que nos importamos [3].

"Desenhar" a rede geralmente é o primeiro passo para análise de sua estrutura, mas quando as redes tem muitos vértices e arestas, a simples análise visual torna-se inviável. Para essa necessidade, dentro da Teoria das Redes foram desenvolvidos conjuntos de ferramentas de medida para ajudar na compreensão dos dados de redes, mesmo quando a visualização é impossível[3].

Um gráfico G pode ser definido como o par V, E onde V é o conjunto de vértices representando os nós, e E é o conjunto de arestas (em inglês, *edges*) que representam as conexões entre os nós. Uma única conexão entre os nós i e j é definida como:

$$E = (i, j) | i, j \in V \quad (3.1)$$

Nesse caso, dizemos que i e j são **vizinhos**. Dois vértices i e j não adjacentes podem ser conectados por uma sequência de m arestas, as quais são chamadas de *caminho* (*walk*) entre i e j , sendo m o *comprimento* do caminho [14]. Essas estruturas descritas são especialmente importantes para redes nas quais dois elementos podem ser ligados por mais de uma conexão [32].

Um *loop* é definido quando um caminho começa e termina no mesmo vértice i , passando somente uma vez por cada vértice. Se todos os vértices e arestas são diferentes ao longo de um caminho, o mesmo é um *path* [14].

Um exemplo de classe importante para medidas em redes são as medidas de *centralidade*, as quais quantificam quão importantes são os vértices ou arestas em uma rede. Um exemplo desse tipo de medida é o *grau*, por exemplo, o grau de um nó é a representação do número de arestas ligadas a ele, e é definido como $\deg(i) = k(i) = |N(i)|$, onde $N(i)$ é o número de vizinhos de um nó (i). Outra notação para essa medida é:

$$k(i) = \sum_j a_{ij} \quad (3.2)$$

Onde a_{ij} corresponde a um elemento da matriz de adjacência [14].

O *grau médio* de um nó na rede é a média dos *graus médios* para todos os nós da rede, e pode ser representado matematicamente por

$$\langle k \rangle = \frac{1}{N} \sum_i k(i) = \frac{1}{N} \sum_{ij} a_{ij} \quad (3.3)$$

Em muitos casos, os vértices com o maior número de conexões (maior grau) em uma rede tem papéis importantes no funcionamento do sistema, de forma que o grau pode ser um guia para direcionar esforços para os elementos cruciais do sistema. Por exemplo, muitas redes contêm um número pequeno, porém significativo, de *hubs*, vértices que contêm um grau atipicamente alto [3, 32].

Um grafo direcionado (*directed*) é definido com uma tripla ordenada $G = (V, E, f)$, onde f é a função que mapeia cada elemento em E a um par ordenado de vértices em V . Os pares ordenados de vértices são chamados **arestas direcionadas**, **arcos** ou **flechas**, e para uma aresta $E = (i, j)$, considera-se que sua direção vai de i para j . Esse tipo de grafo é mais usado para representações de esquemas descrevendo rotas biológicas, interação sequencial de elementos em um ou múltiplos pontos, ou fluxo de informação ao longo da rede [32]. Em redes não direcionadas (*undirected*), o grau corresponde a somente um valor, contudo, em redes direcionadas, os vértices possuem dois graus diferentes, o *grau-in* e o *grau-out*, correspondendo ao

número de arestas que apontam para, ou a partir do nó em questão [3, 14], e podem ser representadas por

$$k_i^{out} = \sum_j a_{ij} \quad (3.4)$$

$$k_i^{in} = \sum_j a_{ji} \quad (3.5)$$

Ainda em grafos não direcionados, se os vértices i e j são conectados, e os vértices j e k são conectados, então i e k são conectados também. Esses conjuntos conectados são chamados de *clusters* ou *componentes conectados* [14].

Um grafo com pesos (*weighted*) é definido como o grafo $G = (V, E)$ onde V é o conjunto de vértices e E é o conjunto de arestas entre os vértices $E = (u, v) | u, v \in V$ associados com uma função peso $w : E \rightarrow R$, onde R denota o conjunto de todos os números reais. Muitas vezes, o peso W_{ij} da aresta entre os nós i e j representa a relevância da conexão. Esse tipo de grafo é o mais usado dentro do campo da Bioinformática [32].

Outro conceito comum no estudo das redes é o *efeito de small-world*. Para defini-lo, é importante elucidar primeiro o conceito de *distância geodésica*, que corresponde ao número mínimo de arestas entre dois vértices diferentes de uma rede. O efeito de *small-world* em uma rede dá-se por poucos graus de separação entre seus nós. Na cultura popular, conhece-se a expressão "seis graus de separação" para redes sociais ("*six degrees of separation*"), um conceito que surgiu primeiramente no meio acadêmico, mas que se disseminou entre leigos depois do lançamento de um filme que levou esse nome. "Seis graus de separação" diz que qualquer pessoa no mundo pode chegar a qualquer outra pessoa, passando por uma sequência de não mais que cinco conhecidos intermediários (totalizando seis passos) [3, 33].

Um terceiro conceito de redes é dado pelos *clusters* ou comunidades, os quais correspondem a nós com fortes conexões entre si, e poucas e raras entre os nós de fora da comunidade. O *clustering coefficient* é uma forma de caracterizar a presença de *loops* de ordem três. Um coeficiente muito usado é conhecido como *transitividade*:

$$C = \frac{3N_{\Delta}}{N_3} \quad (3.6)$$

onde N_{Δ} é o número de triângulos na rede e N_3 é o número de triplas conectadas ¹ [3].

¹Três vértices ligados entre si.

3.3 Modelos de Redes Complexas

Diversos modelos de redes foram propostos com o intuito de estudar as propriedades topológicas de redes reais. Alguns desses modelos tornaram-se objeto de estudo por si próprios (ex: grafos randômicos, modelo *small world*, redes Barabási-Albert), e outros foram aplicados ao estudo da topologia de redes de característica específica, como redes geográficas e redes com estrutura de comunidade [14].

O **grafo aleatório de Erdős e Rényi** (*random graph of Erdős and Rényi* - ER) pode ser considerado o modelo de rede complexa mais básico. Começando com N vértices desconectados, a rede é construída pela adição de M arestas ao acaso, evitando conexões múltiplas e conexões de um nó com ele mesmo. A probabilidade de conexão de cada par de vértices é dada por p . A Figura 3.2 mostra um grafo aleatório de Erdős e Rényi [14, 33].

Figura 3.2: Exemplo do grafo randômico de Erdős e Rényi com $n = 40$ vértices, e probabilidade de criação de nós $p = .3$. Fonte: Elaborado pela autora.

Uma forma comum de estudar redes reais é comparar suas características com valores esperados para redes aleatórias similares. Como os graus dos vértices são características importantes para a rede, é interessante fazer comparações com redes com mesma distribuição de graus [14].

Muitas redes de mundo real exibem uma propriedade chamada de "mundo pequeno" (*small world*), ou seja, muitos vértices podem ser alcançados a partir de outros, através de um número pequeno de arestas [14].

Algumas redes apresentam um grande número de *loops* de três nós (se o vértice 1 conecta-se com os

vértices j e k , há uma alta probabilidade de os vértices j e k serem conectados também), ou seja, o coeficiente de aglomeração (do inglês *clustering coefficient*) é alto. Redes ER apresentam propriedade de mundo pequeno², mas um *clustering coefficient* baixo, em contrapartida, é fácil construir redes regulares de alto *clustering coefficient*, mas elas terão grandes distâncias médias. O *modelo de mundo pequeno de Watts-Strogatz* (*Watts-Strogatz small-world model* - WS) é o modelo mais popular de rede aleatória com características de mundo pequeno (Figura 3.3) [14, 33].

A construção da rede WS começa com uma grade regular de N vértices, cada um conectado aos k vizinhos mais próximos em cada direção, totalizando $2k$ conexões, onde: $N \gg k \gg \log(N) \gg 1$. Cada aresta é então randomicamente embaralhada com probabilidade p de ligação com um vértice diferente. Quando $p = 0$, temos uma grade ordenada resultante, com grande número de *loops*, mas grandes distâncias entre nós. Quando $p \rightarrow 1$, a rede torna-se um grafo aleatório com pequenas distâncias entre os nós, mas alguns poucos *loops*. Com um p intermediário, há pequenas distâncias entre os nós e alto número de *loops*.

Figura 3.3: Exemplo de rede de Watts e Strogatz com $n = 100$ vértices, cada nó conectado com os $k = 5$ vizinhos, probabilidade de embaralhamento de cada vértice $p = 0.4$. Fonte: Elaborado pela autora.

No modelo de Barabási e Albert (*Barabási-Albert model* (BA)) alguns vértices são altamente conectados, enquanto outros possuem menos conexões, com ausência de um grau característico, diferentemente dos modelos de ER e WS, nos quais os vértices das redes possuem um padrão aleatório de conexões de grau

²maioria dos vértices pode ser alcançada a partir de outros, através de um pequeno número de arestas

característico. Nesse modelo, a distribuição de graus segue uma lei de potência para um k grande (Figura 3.4) [14, 33].

$$P(k) \propto k^{-\gamma} \quad (3.7)$$

Começando com N vértices iniciais conectados, novos vértices são adicionados à rede um de cada vez, com probabilidade p de conectar-se aos nós previamente existentes. p é proporcional ao número de ligações que os nós já possuem

Uma característica marcante desse tipo de rede é a existência de *hubs* (vértices que são ligados a uma significativa fração do número total de arestas da rede). O modelo de rede BA é baseado em duas regras básicas: *crescimento* e *ligação preferencial* ("the rich gets richer").

Figura 3.4: Exemplo de rede livre de escala de Barabási e Alberts, $n = 100$ vértices, e número de arestas a se conectar com outros vértices, a partir de um nó, igual a $m = 3$. Fonte: Elaborado pela autora.

Algumas redes reais apresentam estrutura *modular*, ou seja, são formadas por conjuntos ou *comunidades* de vértices, de modo que a maioria das conexões são formadas entre vértices dentro da mesma comunidade, enquanto vértices de diferentes comunidades não conectam com muita frequência (Figura 3.5) [14] [33].

Figura 3.5: (a) Exemplo de rede aleatória com estrutura de comunidades formada por 400 vértices com 4 comunidades. Fonte: Elaborado pela autora.

3.4 Estrutura de dados

Dado um grafo $G = (V, E)$ a *matriz de adjacência* ($|V| * |V| = n * n$) consiste de uma matriz $A = (a_{ij})$, tal que $a_{ij} = 1$ se $(i, j) \in E$ ou $a_{ij} = 0$ se não. Para **grafos não direcionados**, a matriz é simétrica porque $a_{ij} = a_{ji}$, contudo, para **grafos direcionados** essa regra não se aplica, porque os valores da diagonal superior e inferior da matriz revelam a direção dos vértices.

3.5 Aplicações

A notação das Redes Complexas é útil na descrição de problemas que envolvem um grande número de agentes que interagem entre si por meio de regras. Como citado no início dessa introdução, a estrutura matemática *grafo* pode ser utilizada para representar esses sistemas.

Muitas áreas do conhecimento, como Física, Biologia e Ciências Sociais, possuem objetos de interesse que podem ser traduzidos como Redes, uma vez que existem muitos sistemas (sejam naturais ou artificiais) que são compostos por partes individuais, ligados entre si de alguma forma [3, 32]. Dentro do grande campo da Biologia, temos exemplos de aplicação da teoria das Redes Complexas para descrever intera-

ções gênicas [34, 35], interações entre proteínas [36, 37, 38, 39, 40, 41], redes metabólicas [42, 43], redes de interatoma [44], redes estruturais [45], redes neurais [46], e cadeias alimentares [47].

Um trabalho inspirador do presente projeto foi o de Viana et al [9]. Utilizando técnicas de análise de imagem e computação gráfica, os autores fizeram a reconstrução 3D de um osso inteiro de galinha (falange) e representaram-no como uma rede complexa, para a qual cada canal foi associado a uma aresta, e cada confluência de três ou mais canais gerava um nó. A representação da estrutura dos canais dos ossos como uma rede complexa permitiu a aplicação de vários métodos para caracterizar e analisar a organização desse sistema, e sua robustez. A distribuição dos graus dos nós confirmou que a rede dos canais ósseos segue uma lei de potência, e portanto, apresentam nós altamente conectados, os *hubs*. Essa rede também se mostrou particionada entre comunidades, as quais exibiam propriedades topológicas distintas, provavelmente relacionadas com suas funções específicas.

Capítulo 4

Efeitos da seleção de limiar na topologia das redes de co-expressão gênica

O presente Capítulo está organizado da seguinte maneira: Primeiro, introduzo a problemática da seleção de limiar em redes de co-expressão gênica e a motivação deste trabalho. Em seguida, descrevo a seleção de um grupo representativo de base de dados biológicos e públicos para a investigação das metodologias de limiarização (do inglês *thresholding*). A partir de cada uma dessas bases de dados, algumas redes de co-expressão são construídas, variando-se o valor de limiar entre cada uma delas. Diferentes métodos para escolha de limiar são então aplicados nesse conjunto de redes. As redes finais são comparadas para verificação da influência do critério de seleção de limiar sobre os dados. Os resultados obtidos em cada critério são então apresentados, e mostro como a Análise de Componente Principal (PCA) pode ser usada para comparar os efeitos dos métodos de escolha de limiar. O conteúdo desse Capítulo motivou um respectivo artigo publicado na revista científica *Molecular Biosystems* (aceito em 21 de julho de 2017) [48].

4.1 Introdução

O estudo de sistemas complexos é uma tarefa consideravelmente árdua, porque não se pode entender completamente um fenômeno simplesmente o reduzindo a seus componentes isolados [32]. Uma abordagem mais precisa seria considerar todo o cenário de interação, mas ao mesmo tempo, essa é uma tarefa muito complicada, quiçá inviável atualmente. Apesar disso, os estudos de estruturas complexas se beneficiarão grandemente com a consideração do conjunto de interações par a par, examinando não apenas os componentes individuais, mas também como eles se relacionam entre si. Tais estudos são inerentemente parte da grande área de Redes Complexas, uma extensão da Teoria dos Grafos, cuja pesquisa visa caracterizar, analisar, modelar e simular sistemas complexos [49]. A relação entre teoria das Redes Complexas e o estudo de

sistemas biológicos tornou-se ainda mais importante com a ciência de que redes de interações biológicas não são aleatórias, e seguem vários princípios organizacionais topológicos e estruturais [50]. Estes princípios são significativos para caracterizar a rede, e revelam características importantes do sistema estudado. Os complexos princípios da Teoria de Redes podem ser aplicados em um campo específico de pesquisa biológica, a genômica.

A era genômica preocupou-se principalmente com a análise específica do código genético. Atualmente, o desafio mudou para a interpretação da função desses genes no genoma, de modo a esclarecer como eles interagem uns com os outros para executar determinadas funções biológicas [51]. O objetivo da era pós-genômica da Biologia Sistêmica coloca mais ênfase na modelagem das interações biológicas (genes, *small* RNAs e RNAs de interferência, proteínas, metabólitos, entre outros), para unir esses componentes isolados em redes de relação resultantes [52, 49].

4.1.1 Redes de co-expressão gênica

Um tipo particular de rede genética é a rede de co-expressão gênica, que pode ser usada para analisar os dados de expressão gênica em escala global. Esta abordagem baseia-se no fato de que os genes que codificam proteínas que fazem parte de algum complexo, ou que estão envolvidos na mesma via, geralmente são co-regulados e podem exibir um padrão de expressão correlacionado [51]. A construção de uma rede de co-expressão gênica começa com um banco de dados de expressão de genes, que contém diferentes perfis de expressão que mostram como a mesma é influenciada por alguma atividade biológica específica (por exemplo: doença, estágio de desenvolvimento, condição de crescimento etc). As intensidades da expressão gênica são normalizadas, um índice de associação é aplicado entre cada par de genes e suas correlações são classificadas. Alguns índices utilizados para medir a associação entre moléculas biológicas são: índice de Jaccard, índice de Simpson, índice Geométrico, índice cosseno, coeficiente de correlação de Pearson, índice Hipergeométrico e índice de conexão específica [53]. Posteriormente, um método probabilístico é usado para avaliar a chance de observar uma determinada configuração ao acaso, e um corte (limiar) deve ser escolhido para indicar se dada correlação implica em associação (acima do valor de limiar) ou não (abaixo do valor de limiar). Entre os índices mencionados anteriormente, um dos mais frequentemente usados é o coeficiente de correlação de Pearson [11, 54], uma maneira direta e objetiva de medir a correlação entre os perfis de expressão [32, 55, 56, 57, 58, 59, 60, 61, 62, 63].

Vários estudos em Biologia Sistêmica basearam-se na análise de redes de co-expressão gênica *com peso* para revelar genes que desempenham um papel importante na regulação de alguma disfunção biológica, como uma maneira de mostrar potenciais marcadores de genes [64, 65]. Um uso típico de uma rede de co-expressão é a notação funcional de genes, com base em semelhanças de perfil de interação, associando genes de função conhecida a genes de função até então desconhecida [53]. O estudo de Zhang e Horvart

[61] propôs um método de “semi-limiarização” (“*thresholding*”) de redes de co-expressão, o “*soft thresholding*”. Esse método de certo modo filtra os pesos das arestas, ainda mantendo a informação das forças de conexão entre cada par de genes, mantendo um critério de motivação biológica, com base no fato de que muitas redes biológicas tem a distribuição de graus que segue uma lei de potência [30, 66, 67, 68].

Em geral, a Teoria dos Grafos é uma poderosa ferramenta para descrever as relações entre os genes com base em sua co-expressão em cenários biológicos distintos [66]. Em Stuart et al. [51], uma rede de co-expressão de genes de múltiplas espécies foi construída para elucidar a função de genes em uma escala global. Os dados de *microarray* de seres humanos, moscas, vermes e levedura foram unidos em grupos de genes ortólogos¹ que foram atribuídos a um único metagene e, em seguida, foi gerada uma rede de co-expressão de genes dessas múltiplas espécies. A rede mostrou várias relações de co-expressão, conservadas ao longo do tempo evolutivo, entre metagenes recentemente evoluídos e antigos, implicando que esses grupos de genes podem estar funcionalmente relacionados. Com base na co-regulação e conservação evolutiva, a função biológica de alguns dos genes foi confirmada pela associação a genes de função conhecida [51].

4.1.2 Limiarização de redes de co-expressão gênica

Como mencionado anteriormente, após o cálculo da correlação entre cada par de genes, é necessário definir um valor — ou limiar — que indica que um par de vértices está conectado ou não. A escolha do limiar constitui um passo importante, e até mesmo crítico, durante a construção da rede, porque um valor muito alto implicaria poucas conexões entre os vértices, possivelmente resultando em redes não conectadas. Alternativamente, um valor de limiar demasiadamente baixo implicaria em várias conexões entre os vértices, tornando a rede inespecífica e menos significativa.

Limiarizar uma rede, que a princípio tem pesos nas arestas, é desejável para manter as conexões que são significativas de acordo com algum critério razoável, e rejeitar as que não são. A abordagem foi sistematicamente utilizada na ciência de Redes Complexas, incluindo redes de clima [4], transporte [5] e social [6]. Além disso, muitos trabalhos que envolvem redes biológicas [51, 69, 70, 71, 72, 73, 74] empregaram metodologias de limiarização, dado que ela apresenta várias vantagens, como menor custo computacional e eliminação de relacionamentos menos relevantes. Ainda mais importante, não existe uma definição estabelecida de muitas medidas topológicas em redes com peso [7]. Por exemplo, Saramaki et al [7] analisam quatro definições alternativas para *coeficiente de agrupamento* em tais redes.

A questão de qual seria o valor de limiar ideal é tão importante quanto difícil e, em última instância, só pode ser respondida levando-se em consideração os mecanismos biológicos completos subjacentes às redes estudadas. Como isso é atualmente impossível, precisamos de abordagens alternativas capazes de

¹*genes ortólogos* são aqueles originados de um processo de especiação, ou seja, encontram-se em espécies diferentes, mas partilham um ancestral comum.

iluminar esse problema. De fato, mesmo se os mecanismos biológicos completos fossem conhecidos, eles seriam representados como um grafo com peso. Este tipo de grafo é de análise relativamente complexa quando comparado aos gráficos binários, que podem ser caracterizados por um conjunto mais amplo de medidas. O limiar de um grafo com peso é desejável como uma forma de simplificá-lo, e também enfatizar propriedades topológicas dependendo do foco de interesse. Ao mesmo tempo, foi dado relativamente pouca atenção ao desenvolvimento de métodos matemáticos para selecionar um valor de limiar apropriado ao estudar redes de co-expressão de genes. Um dos principais propósitos do presente trabalho é propor uma abordagem empírica a esse problema, acompanhada de algumas análises estatísticas e validação. Em particular, buscamos caracterizar redes distintas de expressão gênica e comparar o comportamento de medidas topológicas de redes sob diferentes valores de limiar escolhidos por diferentes abordagens.

Cinco metodologias foram propostas e caracterizadas para definir um valor de limiar adequado para redes de co-expressão gênicas. Cada método considera um critério distinto para definir o limiar e, dependendo do tipo e finalidade da pesquisa, o critério proposto pode ser mais ou menos desejável. Se o objetivo da análise da expressão gênica global é identificar as relações mais relevantes ao mesmo tempo em que simplifica a rede (eliminando redundâncias que não contribuem necessariamente para o ganho de informações), pode-se recorrer à *maximização da entropia* de determinadas propriedades de rede [75]. Se o objetivo do estudo é analisar relacionamentos genéticos entre genes de uma mesma “vizinhança” [53], como os componentes conectados da rede [3], seria interessante definir o limiar de acordo com o *tamanho desejado do componente gigante da rede*. Uma vez que muitas redes biológicas são livres de escala [30, 66, 67, 68], impondo um limite onde a *distribuição de grau da rede resultante é a mais próxima possível de uma lei de potência* pode gerar uma rede que tenha um significado biológico mais claro [30]. Outra abordagem interessante é restringir o número médio de relacionamentos que os genes podem ter na rede, o que imediatamente se traduz na *configuração do grau médio da rede*. Também consideramos uma situação em que o limiar é definido na *configuração topológica mais estável*, ou seja, um limiar no qual pequenas mudanças no seu valor dificilmente alteram qualquer propriedade topológica da rede.

O objetivo desta parte do trabalho portanto é identificar como a topologia resultante da rede de co-expressão gênica muda de acordo com o método de escolha. Tal análise pode levar a resultados interessantes. Por exemplo, casos em que os limiares obtidos pela maioria dos métodos acima mencionados são semelhantes evitam a necessidade de escolher um método específico. Por outro lado, no caso de obtenção de valores de limiar bastante diferentes, é importante considerar cuidadosamente qual aspecto da pesquisa deve ser enfatizado, o que implica na escolha da metodologia de limiarização.

4.2 Material e Métodos

O conjunto de dados desse estudo consiste em 21 experimentos de *microarray* obtidos das bases de dados *Gene Expression Omnibus* [76] e *Array Express* (www.ebi.ac.uk/arrayexpress). As informações dos conjuntos de dados estão resumidas na Tabela 4.1, incluindo: o grupo biológico do experimento, um código de referência, o número de acesso da base de dados, uma referência bibliográfica chave, o número de genes e réplicas biológicas, e a organização dos dados.

4.2.1 Geração e análise das redes

Cada experimento foi selecionado de uma base de dados de expressão gênica em larga escala, como ilustrado na Figura 4.1a, contendo um respectivo perfil de expressão gênica (Figura 4.1b). Os dados foram normalizados seguindo a transformação *z-score*. Para cada experimento, o Coeficiente de Correlação de Pearson (CCP) foi utilizado para quantificar a coexpressão entre cada par de genes [51, 64, 72].

Figura 4.1: Etapas da metodologia proposta para seleção de limiar. Um conjunto de experimentos é selecionado a partir dos bancos de dados. Cada experimento contém os respectivos dados de expressão gênica (b). O coeficiente de correlação de Pearson é então calculado para todos os pares de genes em cada experimento (c), gerando uma respectiva matriz com pesos (d). Em seguida, um conjunto de limiares são aplicados aos valores de correlação (e), resultando em um conjunto de respectivas redes (f). Os critérios de determinação de limiar são aplicados às redes (g), para escolher a alternativa mais apropriada (h). T_1 , T_2 e T_3 representam valores candidatos de limiar. Fonte: Elaborado pela autora.

A matriz de pesos resultante (Figura 4.1d) é submetida a diferentes valores de limiar (Figura 4.1e), e as respectivas matrizes de adjacência são uma forma alternativa de representação das suas correspondentes redes de coexpressão gênica (Figura 4.1f). Dessa forma, os diferentes critérios de determinação de limiar podem ser aplicados nessas redes (Figura 4.1g) para escolha do mais apropriado (Figura 4.1h). As estatísticas utilizadas para a seleção de limiar são calculadas sobre o componente gigante da rede, ou seja, o maior grupo de vértices conectados após aplicação do limiar.

Tabela 4.1: Conjuntos de dados utilizados e seus respectivos números de acesso e referências. A última coluna da tabela indica os grupos nos quais os dados estão organizados (ex.: controle, condição 1, condição 2).

	Número de acesso	Ref. (Código)	Tamanhos dos dados	Classes
Matriz óssea	GDS928	[77] Pardo (B1)	10499 genes e 6 réplicas	controle (1), grupos temporais 1, 2, 3, 4, 5 (5)
	E-MEXP-1618	[78] Reppe (B2)	54675 genes e 84 réplicas	grupo 1 (84)
	GSE30322	[79] Zhang (B3)	24747 genes e 31 réplicas	grupo temporal 1 (5) e controle 1 (5), grupo temporal 2 (5) e controle 2 (5), temporal group 3 (5) and control 3 (5)
	GSE47552	[80] Lopes (B4)	33297 genes e 99 réplicas	controle (5), grupo 1 (41), grupo 2 (33), grupo 3 (20)
Células cancerígenas	GDS5059	[81] Abdul (C1)	54675 genes e 30 réplicas	grupo temporal 1 (8) e controle (2), temporal group 2 (8) and control (2), temporal group 3 (8) and control (2)
	GDS5071	[82] Kobayashi (C2)	33297 genes e 13 réplicas	controle (6), grupo 1 (7)
	GDS5010	[83] Postnikov (C3)	45101 genes e 18 réplicas	controle 1 (3) and grupo 1 (3), controle 2 (3) e grupo 2 (3), control 3 (3) and group 3 (3)
	GDS5070	[84] Ramsey (C4)	54675 genes e 12 réplicas	controle 1 (3) and grupo 1 (3), controle 2 (3) e grupo 2 (3)
	GSE45016	[85] Satake (C5)	54675 genes e 11 réplicas	controle (1), grupo 1 (10)
Desenvolvimento em <i>Drosophila</i>	GSE9149	[86] Baker (D1)	14010 genes e 29 réplicas	grupo 1 (6), grupo 2 (6), grupo 3 (6), grupo 4 (6), group 5 (5)
	GSE35240	[87] Baumbach (D2)	18952 genes e 15 réplicas	controle 1 (3), controle 2 (3), grupo 1 (3), grupo 2 (3), group
	GSE31564	[88] Chung (D3)	18952 genes and 24 réplicas	2 controls (3) and 6 grupo temporal(3)
	GSE18208	[89] Kong (D4)	18952 genes and 16 réplicas	controle (2), grupo 1 (7), grupo 2 (7)
	GSE34400	[90] Lundberg (D5)	18952 genes and 44 réplicas	controle (6), grupo 1 (38)
Células neuronais	GSE28146	[91] Blalock (N1)	54657 genes e 30 réplicas	controle (7), grupo 1 (7), grupo 2 (8), grupo 3 (8)
	GSE43578	[92] Paschaki (N2)	35557 genes e 18 réplicas	controle 1 (5), grupo 1 (4), controle 2 (3), grupo 2 (6)
Células vegetais	GDS2895	[93] Ehrling (P1)	28800 genes e 20 réplicas	grupo 1 (4), grupo 2 (4), grupo 3 (4), grupo 4 (4), grupo 5 (4)
	GSE14304	[94] Fujita A (P2)	42537 genes e 16 réplicas	grupo 1 (2), grupo 2 (2), grupo 3 (2), grupo 4 (2), grupo 5 (2), grupo 6 (2), grupo 7 (2), grupo 8 (2)
	GSE14304	[94] Fujita B (P3)	57381 genes e 98 réplicas	grupo 1 (98)
	GDS2114	[95] Nagpal (P4)	22810 genes e 12 réplicas	controle 1 (1) e grupo 1(3), controle 2 (1) e grupo 2 (3), controle 3 (1) e grupo 3 (3)
	GDS3244	[96] Wang (P5)	61170 genes e 160 réplicas	40 grupo (4)

4.2.2 Métodos de seleção de limiar

Os critérios a serem usados e comparados são:

1. Máxima variação das propriedades topológicas (4.2.2.a)
2. Máxima entropia (4.2.2.b)
3. Tamanho do componente gigante (4.2.2.c)
4. Distribuição em lei de potência (4.2.2.d)
5. Grau médio específico (4.2.2.e)

4.2.2.a Máxima variação das propriedades topológicas

Para esse critério de limiarização, nós consideramos as seguintes propriedades topológicas de cada rede (descritas em detalhe no Capítulo 2, Seção 2.3):

1. *Estatísticas de grau*;
2. *Transitividade local* [15];
3. *Centralidade de interposição (do inglês, "Betweenness centrality")* [16, 17, 14];
4. *Tamanho do componente gigante* [3];
5. *Número de arestas do componente gigante* [3];
6. *Coefficiente aglomeração (do inglês "Clustering coefficient")* [18] [19];
7. *Assortatividade de grau* [20];
8. *Comprimento médio do caminho mais curtos (do inglês "Average shortest path length")* [21, 22];
9. *Diâmetro* [23];
10. *Coefficiente "Rich club"* [24];
11. *"Matching index"* [25].

Para cada experimento de expressão, 20 limiares linearmente espaçados foram aplicados à respectiva matriz de correlação de Pearson (com os valores de Pearson correspondendo ao peso da correlação). As propriedades topológicas descritas acima foram medidas em cada rede gerada, e a Análise de Componentes Principais (do inglês *Principal Component Analysis* - PCA) foi usada para projetar as propriedades em

um espaço de poucas dimensões, definido pelas duas características mais representativas (dado pelo dois primeiros eixos principais).

Para valores baixos de limiar, espera-se que as propriedades da rede não vão mudar significativamente quando o limiar variar, dado que a rede tenderá a ser altamente conectada. Conforme o valor de limiar aumenta, a rede começará a quebrar em componentes conectados menores, e as propriedades dessas redes podem mudar drasticamente entre si. O ponto no qual a rede vai de um estado estável ² para um estado instável ³ deve representar uma boa escolha de limiar para os valores de correlação. Uma ilustração desse raciocínio é mostrada na Figura 4.2.

Figura 4.2: Ilustração da variação esperada de medidas de uma rede, à medida que o limiar aumenta. Valores menores de limiar devem resultar em valores mais estáveis, enquanto valores mais altos de limiar devem resultar em topologias mais instáveis. M representa uma medida hipotética, e T o valor de limiar. Fonte: Elaborado pela autora.

As redes geradas por cada valor de limiar podem ser representadas como pontos em um espaço bidimensional (2D) pelo PCA. Para encontrar o limiar no qual a topologia da rede vai do estado estável para o instável, a distância euclidiana entre os pontos consecutivos foi medida. O primeiro limiar para o qual a distância é maior que $d_{\max}/f$ é escolhido como o mais apropriado para o conjunto de dados, onde:

- $d_{\max}$ é a distância euclidiana máxima entre os pontos de PCA observado para o experimento;
- f é a constante usada para definir a quantidade de variação no PCA, relativa ao valor máximo, onde a rede é considerada instável.

Nas presentes análises, foi considerado $f = 1.7$, valor que providenciou limiares razoáveis para todos os conjuntos de dados.

²Poucas mudanças nas propriedades topológicas enquanto o limiar varia.

³Grande variação nessas propriedades com a mudança de limiar.

Esse critério preserva as características de uma rede com uma estrutura relativamente robusta, porque o valor de limiar ótimo será selecionado antes que as medidas sofram variação maiores, resultantes de, por exemplo, uma quebra no componente gigante em outros componentes menores. Espera-se que essas redes “quebradas” resultarão em pontos relativamente esparsos no PCA, dado que variações topológicas maiores são mais comuns em redes com menor grau médio [97, 98].

O critério de *Máxima variação nas propriedades topológicas* será útil para mostrar os genes que têm alta correlação de expressão, e que também contribuem para a integridade de um sistema de produtos de co-expressão gênica. Essas características permitem a “clusterização” de genes baseada em perfis de similaridade de interação, e pode ser usado para anotar genes funcionalmente, associando genes desconhecidos com genes de função conhecida [53].

4.2.2.b Máxima entropia

Esse critério baseia-se na hipótese de que o limiar ótimo criará uma rede com a maior quantidade de informação possível, para que a estrutura seja particularmente informativa sobre o sistema que ela representa. Uma forma de medir a quantidade de informação carregada pela rede é a entropia de Shannon [99], que foi detalhada no Capítulo 2, Seção 2.1.4. Foram aplicados 20 limiares diferentes nas matrizes de peso para verificação de qual valor resultou na medida de vértice com a maior entropia.

Quando o limiar utilizado para gerar a rede é baixo, a estrutura da rede tende a um grafo completo, no qual todas as medidas têm o mesmo valor para todos os nós. Para limiares muito altos, a rede apresentará muitos nós isolados, e novamente, os nós terão propriedades similares. Portanto, espera-se que para alguns valores intermediários de limiar, os nós apresentarão um máximo de heterogeneidade para as medidas, o que deverá refletir na entropia. Esse conceito é ilustrado na Figura 4.3: a primeira e a terceira rede (Figuras 4.3a e 4.3c) apresentam baixa entropia devido ao excesso e falta de conexões, respectivamente; a segunda rede (Figura 4.3b) é mais informativa, refletindo uma estrutura e topologia mais complexas. Para análise desse critério, as medidas da *transitividade local* e a *centralidade de interposição* foram usadas como medidas de nó.

O critério de *Máxima entropia* pode ser escolhido com o propósito de analisar as relações mais relevantes entre os genes, em termos de maximização da quantidade de informação sobre a estrutura da rede, enquanto simplificando a mesma (eliminando redundâncias). Essa propriedade pode ser utilizada para analisar o fluxo de informação ao longo da rede, baseado na propriedade que os genes com perfis de expressão de mRNA comparáveis têm uma alta chance de serem regulados pelo mesmo mecanismo genético [100, 101, 102, 75].

Figura 4.3: Grafos hipotéticos e sua entropia esperada. Os grafos (a) e (c) apresentam baixa entropia devido ao excesso e falta de arestas, respectivamente. A rede (b) pode ser vista como tendo uma estrutura mais complexa, refletindo um maior valor de entropia. Fonte: Elaborado pela autora.

4.2.2.c Tamanho do componente gigante

Esse critério é baseado no fato que quanto maior o limiar escolhido para construir a rede, mais esparsa será sua estrutura, dado que um menor número de relações entre genes será mantido na rede final. Enquanto a rede torna-se mais esparsa, seus vértices serão separados em componentes conexos distintos. O maior desses componentes, ou seja, o componente que apresenta o maior número de nós, é chamado de *componente gigante da rede*. O tamanho do componente gigante é fortemente influenciado pelo limiar utilizado para a binarização das conexões. A Figura 2.4 do Capítulo 2, Seção 2.3.4 exemplifica isso. Para analisar esse critério, foram escolhidos limiares que geram redes com três tamanhos de componente gigante: aproximadamente 40, 60 e 80 por cento do número original de vértices.

O critério *Tamanho do componente gigante*, assim como o critério *Máxima variação das propriedades topológicas* (Seção 4.2.2) vai de encontro com o propósito de analisar genes que têm expressão altamente correlacionada. Esse critério pode também ser usado para escolher genes para ser funcionalmente anotados [53]. Quanto menor o tamanho escolhido do componente gigante, maior será a correlação média entre os genes (vértices) conectados.

4.2.2.d Distribuição em lei de potência

Outra maneira de escolher um limiar consiste em determinar um modelo *a priori* ao qual a estrutura da rede deve se adaptar. A maioria das redes biológicas tende a apresentar uma estrutura livre de escala, ou seja, sua distribuição de grau é aproximadamente uma lei de potência [30, 29]:

$$P(k) \propto k^{-\gamma} \quad (4.1)$$

onde γ é o expoente e o símbolo $\propto$ indica “proporcional a”.

Adotamos um critério de limiar com base na imposição de uma distribuição de grau de lei de potência, ou seja, o limiar ótimo seria aquele que tornaria a distribuição de nós da rede mais próxima a uma lei

de potência [61]. Um exemplo do procedimento é mostrado na Figura 4.4. Na parte a desta figura, há um gráfico da distribuição de grau de nós para diferentes redes hipotéticas geradas por valores de limiar diferentes. O gráfico à direita ilustra a distância que uma distribuição de rede hipotética tem de uma distribuição de lei de potência (Figura 4.4b). A fim de medir a semelhança da distribuição de grau com uma lei de potência, primeiro ajusta-se uma função de lei de potência à distribuição usando o procedimento descrito por Clauset et al [103] (ex.: um estimador de máxima verossimilhança do expoente γ). Então, o valor do teste estatístico de Kolmogorov-Smirnov [104] ao qual a distribuição melhor se ajusta à lei de potência é usada como medida de similaridade.

Figura 4.4: Gráfico da distribuição de grau de redes hipotéticas (a). A linha preta tracejada representa uma função em lei de potência. O gráfico (b) representa a distância em que as diferentes distribuições de grau hipotético têm de uma distribuição em lei de potência. T representa o valor de limiar. Fonte: Elaborado pela autora.

O critério de *Distribuição em lei de potência* baseia-se em redes de co-expressão gênica modeladas como rede livre de escala de genes [30, 66], portanto, esse método vai se encaixar com o objetivo de alcançar uma rede que segue esse modelo o mais próximo possível.

4.2.2.e Grau médio específico

Esse critério consiste em escolher valores de limiar que geram redes com graus médios específicos, como ilustrado na Figura 4.5. Usando esse critério, pode-se determinar o número médio de interações que um gene vai ter na rede. Foram considerados dois valores de grau médio, 100 e 1000, e os limiares escolhidos foram aqueles que geraram redes com o grau médio mais próximo ao desejado. Esse critério pode ser usado para notação funcional de genes [53], assim como assistir na identificação de uma família de transcritos envolvidos em um processo biológico similar [102, 66].

Figura 4.5: Grau médio para diferentes valores de limiar. Quanto maior o grau médio da rede, mais conectados serão os nós. O grau médio é representado por $\bar{K}$. Um grau médio escolhido f_k possui um limiar correspondente T_k . Fonte: Elaborado pela autora.

4.3 Resultados e Discussão

A Figura 4.6 descreve a distribuição de todos os valores dos coeficientes de correlação de Pearson (correspondente a cada par de expressões de genes, produzindo os pesos das arestas) obtidos para todos os conjuntos de dados originais. Também estão incluídos os limiares dados pela respectiva aplicação dos vários critérios considerados. Observe que em 71.4% das bases, o limiar escolhido está entre 0.70 e 1.0. Interessantemente, a maioria das distribuições de peso nesta figura mostram um perfil semelhante, parecido com a porção direita de uma curva gaussiana, indicando uma predominância de valores relativamente pequenos, que estão sendo filtrados pelos critérios propostos como originalmente desejado.

A Tabela 4.2 mostra uma análise complementar utilizando o Coeficiente de Correlação de Pearson (CCP), correspondente aos valores absolutos obtidos em relação à força (soma dos pesos das arestas conectados a cada nó) em relação ao grau. Os resultados indicam uma relação significativa entre força e grau, o que justifica que o critério de limiar removeu os links mais fracos entre os genes na maioria dos casos. Considerando todos os critérios, pode-se verificar que as correlações de Pearson que são maiores que 0.5 (células verdes da tabela) correspondem a 66.5% do total. Excluindo o critério de *Máxima variação da propriedades topológicas*, a percentagem de células verdes vai para 71.42%. Estas percentagens são substancialmente maiores do que a hipótese nula de 50%, corroborando com o fato de que as relações com peso alto tendem a ser incorporadas nas conexões após limiar.

Figura 4.6: Distribuições dos coeficientes de correlação de Pearson para cada conjunto de dados (linha azul). As linhas verticais correspondem aos limiares selecionados de acordo com cada critério adotado. “*Pearson dist.*” corresponde à distribuição de correlação de Pearson. Fonte: Elaborado pela autora.

Tabela 4.2: Coeficientes de correlação de Pearson absolutos obtidos para as relações entre a força do nó (soma dos pesos das bordas) e o grau. As células verdes indicam valores maiores que 0.5. Os nomes das linhas referem-se ao primeiro nome do conjunto de autores responsáveis pela publicação da correspondente base de dados.

	T variation	T entropy	T comp. 0.4	T comp. 0.6	T comp. 0.8	T deg. 100	T deg. 1000	T power-law
Pardo	0.65	0.65	0.31	0.47	0.47	0.75	0.93	0.31
Reppe	0.16	0.63	0.41	0.58	0.77	0.46	0.84	0.70
Zhang	0.08	0.65	0.36	0.50	0.61	0.61	0.87	0.58
Lopes	0.06	0.82	0.52	0.64	0.75	0.53	0.82	0.33
Abdul	0.19	0.89	0.62	0.77	0.84	0.51	0.75	0.28
Kobayashi	0.28	0.70	0.50	0.56	0.61	0.62	0.81	0.67
Postnikov	0.34	0.65	0.43	0.51	0.55	0.67	0.91	0.91
Ramsey	0.84	0.73	0.57	0.64	0.66	0.50	0.73	0.44
Satake	0.13	0.30	0.16	0.19	0.22	0.40	0.77	0.15
Baker	0.51	0.55	0.40	0.46	0.64	0.31	0.60	0.42
Baumbach	0.77	0.83	0.72	0.77	0.80	0.70	0.80	0.80
Chung	0.62	0.22	0.61	0.60	0.43	0.30	0.94	0.82
Kong	0.56	0.53	0.63	0.78	0.83	0.51	0.78	0.58
Lundberg	0.11	0.92	0.61	0.78	0.87	0.67	0.92	0.78
Blalock	0.09	0.59	0.38	0.42	0.46	0.60	0.89	0.71
Paschaki	0.13	0.63	0.41	0.51	0.59	0.61	0.86	0.68
Ehlting	0.20	0.75	0.49	0.57	0.64	0.67	0.90	0.76
Fujitaa	0.42	0.49	0.38	0.42	0.53	0.35	0.66	0.29
Fujitab	0.00	0.16	0.07	0.03	0.12	0.07	0.55	0.07
Nagpal	0.34	0.71	0.44	0.56	0.63	0.69	0.90	0.58
Wang	0.10	0.72	0.60	0.77	0.89	0.53	0.76	0.38

4.3.1 Aplicação dos critérios de Limiarização

4.3.1.a Máxima variação das propriedades topológicas.

As projeções de PCA são mostradas na Figura 4.7. Cada ponto de cada um dos gráficos representa um conjunto de medidas de uma rede, a qual foi gerada por um respectivo valor de limiar. Essas redes consecutivas são geradas a partir de limiares de valores também consecutivos, de modo que a ligação desses pontos representa a evolução das redes em função do valor de limiar. A Figura 4.8 mostra a mensuração da variação entre pontos consecutivos nas projeções de PCA. Em ambas as Figuras, os pontos vermelhos indicam a transição detectada das porções estáveis às instáveis.

Existe uma tendência de estabilização no início de cada trajetória e uma região flutuante e instável no final das mesmas. As redes no início dessas trajetórias foram geradas por limiares mais baixos (perto de 0.85) e, portanto, têm uma maior concentração de arestas, uma vez que um limiar inferior permite que um maior número de níveis de expressão gênica seja classificado como correlacionado.

Figura 4.7: Projeções de PCA de propriedades de rede obtidas usando valores de limiar distintos. O nome na parte superior de cada gráfico corresponde ao primeiro autor do conjunto de dados em que as redes foram geradas. Os pontos verdes indicam a rede correspondente ao limiar mais baixo. Pontos vermelhos indicam a transição detectada de propriedades estáveis a instáveis. Fonte: Elaborado pela autora.

4.3.1.b Máxima entropia.

A Entropia de Shannon foi aplicada nos valores de transitividade local e centralidade de interposição. O resultado é mostrado na Figura 4.9. Os pontos vermelhos indicam a entropia máxima alcançada por cada experimento.

Figura 4.8: Variação entre pontos consecutivos do PCA da Figura 4.7. O nome na parte superior de cada gráfico corresponde ao primeiro autor do conjunto de dados. Pontos vermelhos indicam a transição detectada das porções estáveis às instáveis. Fonte: Elaborado pela autora.

Figura 4.9: Valores de entropia obtidos para redes geradas por valores de limiar diferentes. Os pontos vermelhos indicam a entropia máxima alcançada por cada conjunto de dados. Fonte: Elaborado pela autora.

4.3.1.c Tamanho do componente gigante.

A Figura 4.10 mostra o tamanho do componente gigante como uma função dos limiares considerados. Observe que os tamanhos são normalizados pelo número total de genes de cada experimento. Os pontos em vermelho, verde e roxo indicam redes com componentes gigantes que compõem aproximadamente, respectivamente, 40%, 60% e 80% do número total de genes. A distância entre esses pontos destacados também pode indicar a resiliência da estrutura da rede diante das variações de limiar. Alguns deles (ex.: Postnikov [83], Ramsey [84] e Blalock [91]) tiveram o componente gigante quebrado em componentes menores muito rápido, enquanto outros (Reppe [78], Abdul [81] and Wang [96]), mostraram uma transição mais gradual entre as marcas destacadas.

Figura 4.10: Tamanho do componente gigante da rede em função do valor de limiar. O tamanho é normalizado pelo número total de genes estudados em cada experimento. Os pontos coloridos indicam os tamanhos de componentes gigantes usados na análise, representados como f_s . S representa o tamanho do componente gigante. Fonte: Elaborado pela autora.

4.3.1.d Distribuição em Lei de Potência.

A Figura 4.11 mostra as estatísticas de Kolmogorov-Smirnov (K-S) da hipótese de que a distribuição de grau das redes segue uma função de lei de potência. Os pontos vermelhos indicam o valor mínimo obtido para cada conjunto de dados. Note que, em muitos casos, a estatística K-S mostra grandes flutuações à medida que o limiar é variado.

Figura 4.11: Estatística Kolmogorov-Smirnov comparando a distribuição de grau da rede a uma função em lei de potência. Fonte: Elaborado pela autora.

4.3.1.e Grau médio específico.

A Figura 4.12 mostra os graus médios das redes como uma função de limiar. Os pontos vermelho e verde indicam os limiares para os quais o grau médio é, respectivamente, 100 e 1000.

4.3.2 Comparação entre métodos

A Tabela 4.3 mostra nove propriedades relevantes obtidas para as redes geradas para cada método de limiarização. Pela Tabela 4.3, está claro que a medida de *número de nós do componente gigante* (N_{giant}) de uma dada base de dados varia substancialmente entre os critérios. Para algumas bases, a limiarização manteve a maioria dos genes no componente gigante, como a base de Postnikov [83], que manteve todos seus genes na rede final quando limiarizada pelo critério *Grau médio específico* (1000) e pelo *Distribuição em lei de potência*. Além desta, a base de Ramsey [84] também manteve todos os genes considerando o critério de *Máxima variação das propriedades topológicas*. Por fim, a base de Satake [85] comportou-se similarmente sob o critério de *Grau médio específico* (100).

Como cada critério é usado para selecionar limiares apropriados para os 21 conjuntos de dados, uma

Figura 4.12: Grau médio das redes obtidas por diferentes valores de limiar. Os pontos indicam o grau médio escolhido para analisar este critério, representado como f_k . Fonte: Elaborado pela autora.

abordagem útil para comparar os valores selecionados por cada critério é calcular o coeficiente de correlação de Pearson entre os limiares obtidos. O resultado pode ser representado como um grafo ponderado, onde cada nó é um critério e as arestas indicam a correlação de Pearson entre pares de nós. Esse grafo é mostrado na Figura 4.13a. O tamanho dos nós está associado à correlação média de Pearson entre o nó e todos os outros nós, ou seja, indica quão fortemente o respectivo critério está relacionado aos outros. Os gráficos de dispersão usados para calcular esses valores de correlação são mostrados na Figura 4.14, e os valores calculados são mostrados nas Tabelas 4.4, 4.5 and 4.6.

Figura 4.13: (a) Grafo com pesos representando a correlação de Pearson entre os resultados dos diferentes critérios. O tamanho do nó é dado pela média do logaritmo da correlação de Pearson entre o critério e todos os outros critérios. As arestas mais grossas correspondem a valores de correlação mais elevados. Grafos com pesos representando (b) $\tilde{\alpha}$ e (c) $\tilde{\beta}$ obtidos também são mostrados. *T variation* representa o critério de Máxima variação das medidas topológicas; *T entropy* é o critério de máxima entropia; *T comp. 0.4* representa o critério de Tamanho do componente gigante fixado em 40% do número total de nós, com um significado análogo para *T comp. 0.6* e *T comp. 0.8*; *T deg. 100* representa o grau médio específico (100), e *T deg. 1000* grau médio específico (1000); *T power-law* representa o critério de distribuição de grau em lei de potência (*power-law*). As bordas vermelhas indicam critérios que divergiram mais entre si, portanto, com uma correlação Pearson baixa ou alta $\tilde{\alpha}$ ou $\tilde{\beta}$. O nó *T variation* foi excluído em (b) e (c) para não mascarar as diferenças entre os outros métodos. Fonte: Elaborado pela autora.

Tabela 4.3: Medidas obtidas para cada conjunto de dados ao considerar os diferentes critérios de limiar. A caixa abaixo da tabela indica a posição de cada medida nas células da tabela. *N giant* é o número de nós do componente gigante, *Avg. degree* representa o grau médio, *Std. degree* é o desvio padrão do grau dos nós, *Clust. coef.* É o coeficiente de agrupamento. Nós também mostramos estatísticas associadas à estrutura da comunidade encontrada para cada rede (encontrado usando o método apresentado em [105]): *R1* indica o número de nós na maior comunidade, dividido por *N giant*, *R2* é o número de nós na segunda maior comunidade, dividido por *N giant*, *Tam. med. comun* é o número médio de nós contidos nas comunidades. Consulte a legenda da Figura 4.13 para o significado de cada nome.

	T variation			T entropy			T comp. 0.4			T comp. 0.6			T comp. 0.8			T deg. 100			T deg. 1000			T power-law		
Pardo	10303	33	32.2	10305	33	32.7	2152	5	6.4	7492	8	9.3	7492	8	9.3	10498	111	91.2	10499	1040	520.9	2152	5	6.4
	0.47	25	0.73	0.47	25	0.73	0.63	53	0.87	0.52	51	0.85	0.52	51	0.85	0.48	14	0.65	0.54	4	0.43	0.63	53	0.87
	0.12	0.09	468.3	0.14	0.11	448.0	0.07	0.06	55.2	0.07	0.07	170.3	0.07	0.07	170.3	0.19	0.14	954.4	0.21	0.21	1749.8	0.07	0.06	55.2
Reppe	1171	42	52.0	37161	394	556.6	22939	146	257.8	34396	316	470.6	44480	715	887.9	27365	190	319.1	49756	1060	1234.3	40664	527	695.9
	0.60	12	0.52	0.52	14	0.66	0.59	16	0.71	0.53	12	0.68	0.46	12	0.59	0.57	15	0.71	0.43	11	0.52	0.49	12	0.63
	0.48	0.31	234.2	0.21	0.17	1238.7	0.22	0.17	573.5	0.21	0.15	1146.5	0.19	0.19	2471.1	0.21	0.16	829.2	0.33	0.19	1658.5	0.20	0.17	1506.1
Zhang	117	6	6.0	21191	165	215.6	9520	26	41.8	16180	62	94.5	20256	132	179.1	20144	128	175.0	24566	1018	879.4	19272	106	149.4
	0.46	11	0.49	0.46	14	0.68	0.46	34	0.78	0.46	19	0.74	0.46	17	0.69	0.46	18	0.70	0.46	9	0.49	0.46	18	0.71
	0.19	0.18	16.7	0.24	0.18	1324.4	0.20	0.16	207.0	0.22	0.18	622.3	0.25	0.15	964.6	0.25	0.14	1060.2	0.39	0.22	1889.7	0.24	0.14	837.9
Lopes	106	8	7.1	28928	1120	1763.3	14186	201	428.2	20392	386	762.0	25632	713	1248.4	14585	209	444.5	28928	1120	1763.3	5370	60	123.5
	0.43	6	0.42	0.57	11	0.23	0.55	16	0.28	0.57	15	0.26	0.57	12	0.24	0.56	20	0.28	0.57	11	0.23	0.43	26	0.41
	0.31	0.24	21.2	0.32	0.25	1257.7	0.21	0.19	322.4	0.26	0.20	485.5	0.31	0.25	732.3	0.20	0.19	331.5	0.32	0.25	1257.7	0.23	0.17	141.3
Abdul	2354	10	15.0	52335	3246	4510.3	22336	706	1044.6	34157	1773	2477.8	44441	2527	3560.2	16060	303	470.2	32175	1600	2242.1	4901	24	37.0
	0.38	19	0.68	0.65	15	0.19	0.52	17	0.31	0.60	13	0.24	0.63	13	0.21	0.45	18	0.37	0.59	14	0.24	0.36	19	0.61
	0.25	0.16	102.3	0.41	0.21	951.5	0.26	0.24	698.0	0.25	0.24	509.8	0.35	0.22	389.8	0.28	0.25	446.1	0.25	0.24	527.5	0.21	0.19	175.0
Kobayashi	2168	4	3.6	32385	303	546.3	13045	25	41.1	20015	56	105.0	26014	107	206.1	26298	110	213.6	33297	1007	1427.9	30769	201	380.7
	0.31	28	0.82	0.41	18	0.34	0.27	29	0.53	0.30	23	0.46	0.34	23	0.41	0.34	22	0.40	0.48	8	0.27	0.38	21	0.36
	0.10	0.09	57.1	0.21	0.17	2944.1	0.20	0.13	109.6	0.19	0.19	171.1	0.18	0.16	292.3	0.18	0.15	226.7	0.34	0.18	5549.5	0.23	0.20	732.6
Postnikov	6845	29	68.4	44800	79	197.1	15662	34	91.9	30790	36	111.4	37931	41	126.3	44993	96	223.1	45101	960	896.2	45101	960	896.2
	0.54	40	0.48	0.48	12	0.52	0.52	28	0.50	0.52	24	0.52	0.51	25	0.53	0.47	10	0.51	0.37	3	0.39	0.37	3	0.39
	0.19	0.15	152.1	0.42	0.17	5600.0	0.22	0.18	127.3	0.21	0.14	311.0	0.17	0.13	972.6	0.37	0.25	6427.6	0.51	0.17	11275.2	0.51	0.17	11275.2
Ramsey	54675	2387	3782.0	54191	1051	2131.8	22713	626	1031.7	35320	742	1404.2	43166	767	1553.5	16757	395	638.9	54191	1051	2131.8	13084	241	390.5
	0.70	5	0.20	0.66	18	0.21	0.58	23	0.27	0.62	34	0.24	0.63	33	0.23	0.54	23	0.31	0.66	18	0.21	0.50	26	0.34
	0.67	0.17	18225.0	0.68	0.14	9031.8	0.33	0.27	172.1	0.52	0.18	353.2	0.50	0.17	553.4	0.27	0.26	174.6	0.68	0.14	9031.8	0.30	0.28	165.6
Satake	12403	24	66.2	53595	52	129.4	21734	25	74.0	32091	27	82.1	42580	31	92.1	54665	110	205.8	54675	1048	814.6	20499	25	72.9
	0.66	39	0.63	0.59	19	0.70	0.65	34	0.67	0.64	30	0.69	0.62	31	0.70	0.55	10	0.69	0.43	3	0.53	0.65	38	0.67
	0.22	0.08	206.7	0.13	0.13	3828.2	0.22	0.09	255.7	0.22	0.10	324.2	0.20	0.09	519.3	0.14	0.13	4969.5	0.16	0.14	5467.5	0.23	0.12	230.3
Baker	8924	960	879.9	9786	1123	1012.3	3462	410	426.7	7862	756	711.8	11535	1512	1298.5	2084	275	278.6	10922	1361	1193.2	3831	449	481.2
	0.77	22	0.50	0.79	24	0.50	0.68	17	0.31	0.74	24	0.50	0.82	17	0.50	0.67	12	0.23	0.81	22	0.50	0.68	30	0.29
	0.47	0.36	175.0	0.46	0.34	191.9	0.49	0.34	123.6	0.47	0.38	178.7	0.45	0.33	349.5	0.35	0.32	148.9	0.45	0.33	254.0	0.47	0.32	182.4
Baumbach	11322	2314	2494.3	18287	2004	2686.8	7610	2736	2210.1	11468	2294	2495.9	15300	1964	2544.3	7064	2794	2123.9	15356	1962	2545.4	15356	1962	2545.4
	0.99	24	0.02	0.96	14	0.08	0.99	33	0.01	0.99	24	0.02	0.98	18	0.04	0.99	35	0.01	0.98	18	0.04	0.98	18	0.04
	0.43	0.21	808.7	0.31	0.30	3047.8	0.52	0.11	380.5	0.42	0.21	955.7	0.34	0.27	1530.0	0.55	0.11	353.2	0.34	0.28	1535.6	0.34	0.28	1535.6
Chung	3425	88	92.1	11702	130	109.3	6745	107	95.5	7905	104	96.3	10634	112	99.0	11418	122	104.9	13577	1033	763.5	13575	577	468.5
	0.98	57	0.87	0.80	20	0.89	0.97	49	0.94	0.96	33	0.94	0.89	20	0.92	0.83	20	0.90	0.52	4	0.40	0.54	6	0.53
	0.20	0.09	142.7	0.17	0.16	417.9	0.22	0.06	249.8	0.27	0.07	304.0	0.25	0.12	393.9	0.29	0.12	346.0	0.19	0.19	1357.7	0.18	0.17	646.4
Kong	7382	458	577.4	6912	357	460.8	8107	798	920.2	11746	1680	1921.5	14814	1789	2264.8	6438	324	412.4	11380	1667	1877.6	7621	563	686.9
	0.61	34	0.28	0.58	38	0.31	0.67	26	0.22	0.81	28	0.10	0.86	22	0.08	0.57	41	0.32	0.80	29	0.11	0.63	30	0.26
	0.24	0.21	388.5	0.26	0.21	406.6	0.24	0.23	405.4	0.24	0.20	783.1	0.34	0.17	1139.5	0.27	0.22	306.6	0.20	0.19	711.2	0.24	0.21	401.1
Lundberg	85	6	5.9	18680	984	1233.3	7533	138	209.1	11310	399	513.5	15728	621	828.5	8670	205	287.2	18680	984	1233.3	11084	382	493.5
	0.56	9	0.46	0.57	9	0.29	0.56	13	0.47	0.53	16	0.40	0.55	12	0.34	0.55	21	0.45	0.57	9	0.29	0.53	17	0.40
	0.32	0.25	17.0	0.36	0.26	3736.0	0.23	0.22	342.4	0.28	0.25	418.9	0.25	0.25	1048.5	0.21	0.21	433.5	0.36	0.26	3736.0	0.29	0.24	481.9
Blalock	38	8	6.7	54152	92	198.1	24906	28	77.2	33214	33	91.1	41658	40	108.8	54331	100	209.2	54675	1048	903.5	54675	245	371.1
	0.57	4	0.19	0.38	11	0.61	0.42	19	0.66	0.41	22	0.66	0.40	20	0.65	0.38	11	0.61	0.31	3	0.42	0.36	6	0.55
	0.34	0.24	9.5	0.15	0.11	3610.1	0.15	0.12	191.6	0.16	0.11	281.5	0.13	0.10	344.3	0.18	0.11	3395.7	0.16	0.15	6075.0	0.25	0.13	4970.5
Paschaki	583	23	26.7	31465	149	266.7	14083	42	91.5	22071	68	140.6	28757	111	211.2	30089	127	235.9	35556	1008	1031.9	33897	215	353.4
	0.60	18	0.54	0.46	21	0.59	0.52	25	0.61	0.49	21	0.61	0.47	21	0.60	0.46	18	0.59	0.45	5	0.48	0.45	15	0.57
	0.25	0.23	36.4	0.31	0.14	533.3	0.21	0.21	260.8	0.27	0.20	315.3	0.30	0.21	364.0	0.31	0.15	518.8	0.30	0.17	5926.0	0.31	0.14	997.0
Ehlting	438	5	5.9	25175	221	387.6	10640	32	70.3	15956	54	119.4	20439	88	186.4	22382	114	232.6	25739	1015	1065.1	25346	243	414.8
	0.30	11	0.62	0.44	13	0.40	0.37																	

Tabela 4.4: Valores de correlação de Pearson entre os métodos mostrados na Figura 4.14. Consulte a legenda da Figura 4.13 para o significado de cada nome.

	T variation	T entropy	T comp 0.4	T comp 0.6	T comp 0.8	T deg. 100	T deg. 1000	T power-law
T variation	1	0.05	0.09	0.05	0.02	0.02	-0.03	0.14
T entropy	0.05	1	0.92	0.93	0.90	0.80	0.75	0.61
T comp 0.4	0.09	0.92	1	0.97	0.93	0.89	0.79	0.64
T comp 0.6	0.05	0.93	0.97	1	0.99	0.80	0.71	0.56
T comp 0.8	0.02	0.90	0.93	0.99	1	0.73	0.66	0.51
T deg. 100	0.02	0.80	0.89	0.80	0.73	1	0.95	0.83
T deg. 1000	-0.03	0.75	0.79	0.71	0.66	0.95	1	0.88
T power-law	0.14	0.61	0.64	0.56	0.51	0.83	0.88	1

Tabela 4.5: Valores de alpha ($\tilde{\alpha}$) para os métodos mostrados na Figura 4.14. Consulte a legenda da Figura 4.13 para o significado de cada nome.

	T variation	T entropy	T comp. 0.4	T comp. 0.6	T comp. 0.8	T deg. 100	T deg. 1000	T power-law
T variation	1.00	0.02	0.04	0.02	0.01	0.01	-0.01	0.04
T entropy	0.16	1.00	1.39	1.09	0.87	1.06	0.70	0.57
T comp. 0.4	0.19	0.60	1.00	0.75	0.60	0.77	0.48	0.39
T comp. 0.6	0.12	0.78	1.25	1.00	0.81	0.89	0.56	0.45
T comp. 0.8	0.06	0.93	1.47	1.20	1.00	0.99	0.63	0.50
T deg. 100	0.04	0.61	1.02	0.71	0.53	1.00	0.67	0.59
T deg. 1000	-0.10	0.81	1.29	0.90	0.69	1.34	1.00	0.88
T power-law	0.49	0.65	1.04	0.71	0.53	1.17	0.87	1.00

Tabela 4.6: Valores de beta ($\tilde{\beta}$) para os métodos mostrados na Figura 4.14. Consulte a legenda da Figura 4.13 para o significado de cada nome.

	T variation	T entropy	T comp. 0.4	T comp. 0.6	T comp. 0.8	T deg. 100	T deg. 1000	T power-law
T variation	0.00	0.94	0.91	0.93	0.94	0.94	0.95	0.91
T entropy	0.64	0.00	-0.45	-0.13	0.09	-0.11	0.30	0.31
T comp. 0.4	0.71	0.41	0.00	0.25	0.41	0.23	0.55	0.56
T comp. 0.6	0.73	0.22	-0.27	0.00	0.19	0.08	0.45	0.47
T comp. 0.8	0.75	0.07	-0.50	-0.21	-0.00	-0.04	0.35	0.39
T deg. 100	0.82	0.38	-0.05	0.25	0.43	0.00	0.38	0.36
T deg. 1000	0.80	0.07	-0.44	-0.06	0.16	-0.44	-0.00	-0.03
T power-law	0.38	0.32	-0.09	0.23	0.41	-0.17	0.22	0.00

Figura 4.14: Gráficos de dispersão entre os valores de limiar escolhidos pelos diferentes critérios. Cada ponto corresponde a um conjunto de dados. Consulte o texto principal para o significado dos nomes dos eixos. ρ representa o coeficiente de correlação de Pearson entre os valores do gráfico, α corresponde ao coeficiente angular, e β ao coeficiente linear. Consulte a legenda da Figura 4.13 para o significado de cada nome. Fonte: Elaborado pela autora.

Os resultados indicam que, tirando-se o critério *Variação máxima de propriedades topológicas*, a correlação entre todos os métodos é semelhante. O referido critério mostrou baixa correlação com os outros métodos, o que significa que as redes geradas pelo método *Variação máxima de propriedades topológicas* possuem propriedades topológicas altamente distintas das redes geradas pelos outros métodos. Um exemplo dessa divergência pode ser ilustrada pela grande diferença dos valores de *número de nós do componente gigante* (N_{giant}) para esse critério. Um desses casos é o da base de Lundberg [90], a qual vai de 85 (critério *Variação máxima de propriedades topológicas*) a 18.000 nós (critério *Grau médio específico (1000)*) (Tabela 4.3). Além disso, o critério *Variação máxima de propriedades topológicas* tem tendência de produzir redes com valores de N_{giant} particularmente baixos (respectivos às bases de Blalock [91], Paschaki [92], Ehltig [93], e Wang [96]) (Tabela 4.3).

No entanto, é importante notar que o resultado da correlação (Figura 4.13) não tem uma interpretação tão direta. Para explicar o porquê, a Figura 4.15 mostra dois casos diferentes de comparação de limiar entre os resultados obtidos: *Máxima entropia versus Tamanho do componente gigante 60%* (Figura 4.15a), e *Distribuição em lei de potência versus Grau médio específico 1000* (Figura 4.15b). A linha vermelha indica o ajuste linear dos quadrados mínimos dos pontos. O fato de que ambos os casos tem uma alta correlação (ρ) não significa que os limiares escolhidos serão os mesmos. Por exemplo, o resultado mostrado na Figura 4.15b indica que os limiares selecionados pelo critério *Distribuição em lei de potência* tende a ser maior que aqueles selecionados pelo critério *grau médio específico (1000)*, mesmo que eles sejam fortemente correlacionados.

Figura 4.15: Correlação entre os critérios: *Máxima entropia versus Tamanho do componente gigante 60%* (a), e *Distribuição em lei de potência versus Grau médio específico 1000* (b). A letra ρ corresponde ao coeficiente de correlação de Pearson, α ao coeficiente angular e β ao coeficiente linear. A linha vermelha indica reta que melhor se ajusta aos quadrados mínimos. Fonte: Elaborado pela autora.

Para comparar adequadamente os limiares selecionados, os coeficientes angulares e lineares das relações entre os métodos também precisam ser levados em consideração. Esses coeficientes são obtidos aplicando uma regressão linear de quadrados mínimos aos dados [106] (ver Figura 4.16 para a definição).

Sobre o coeficiente angular, α , quando $\alpha \approx 1$ os limiares selecionados tenderão a variar de maneira semelhante. Portanto, definimos a relação angular entre dois critérios como $\tilde{\alpha} = 1 - \alpha$. Valores baixos de $\tilde{\alpha}$ significam que os métodos podem estar relacionados (a relação real ainda depende do coeficiente de correlação), enquanto valores grandes indicam ausência de relação. Por sua vez, o valor absoluto do coeficiente linear, $\tilde{\beta} = |\beta|$, também está relacionado à similaridade entre métodos, dado que $\tilde{\beta} \approx 0$ significa que os métodos podem estar relacionados. Os valores de $\tilde{\alpha}$ and $\tilde{\beta}$ podem ser diferentes dependendo da escolha da variável explicativa. Em outras palavras, aplicar o método linear dos quadrados mínimos a um gráfico de y em função de x pode dar um $\tilde{\alpha}$ e $\tilde{\beta}$ diferentes do que quando aplicando o mesmo método a um *plot* de x em função de y . O anterior considera x como uma variável explanatória, enquanto no segundo, a variável explanatória é y . Portanto, para cada par de critérios sendo comparados, escolhemos o mínimo $\tilde{\alpha}$ e $\tilde{\beta}$ obtidos quando consideramos as variáveis explanatórias os limiares de cada critério.

Figura 4.16: Significado dos coeficientes angular (α) e linear (β). A linha vermelha representa uma regressão linear dos pontos azul. β é dado pelo valor de y onde a linha vermelha intercepta o eixo y . θ é o ângulo entre a linha vermelha e o eixo x . α é calculado como $\alpha = \tan(\theta)$. Fonte: Elaborado pela autora.

As Figuras 4.13b e 4.13c mostram uma representação gráfica das relações entre os critérios considerando, respectivamente, $\tilde{\alpha}$ e $\tilde{\beta}$. As arestas destacadas em vermelho indicam os pares de métodos que mais divergiram. Para melhor interpretar os resultados, calculamos a diferença entre os valores limiar (usando a Tabela 4.7) obtidos por esses pares de métodos destacados, o resultado é mostrado na Figura 4.17. Observamos que esses critérios selecionaram valores de limiar diferentes para muitos conjuntos de dados, embora a diferença possa variar amplamente entre os casos. Por exemplo, a comparação entre o critério *máxima entropia* e o critério *Distribuição em lei de potência* resultou em 11 de 21 (11/21) valores de limiar com uma diferença menor que 0.05 (Figura 4.17b) (bases de Lopez [80], Abdul [81], Kong [89], Blalock [91], Wang [96]), e a comparação entre os critérios *grau médio específico (100)* e *grau médio específico (1000)* resultaram em todos os limiares (21/21) com uma diferença maior que 0.05 (Figura 4.17f).

Tabela 4.7: Limiares escolhidos pelos diferentes critérios para cada experimento. Consulte a legenda da Figura 4.13 para o significado de cada nome.

	T variation	T entropy	T comp. 0.4	T comp. 0.6	T comp. 0.8	T deg. 100	T deg. 1000	T power-law
Pardo	0.974	0.974	0.995	0.990	0.990	0.949	0.808	0.995
Reppe	0.892	0.590	0.711	0.616	0.521	0.677	0.475	0.557
Zhang	0.974	0.813	0.911	0.869	0.827	0.828	0.667	0.839
Lopes	0.923	0.535	0.711	0.648	0.584	0.707	0.535	0.801
Abdul	0.974	0.667	0.869	0.784	0.721	0.909	0.798	0.964
Kobayashi	0.974	0.857	0.942	0.921	0.900	0.899	0.788	0.875
Postnikov	0.903	0.779	0.869	0.837	0.821	0.768	0.606	0.606
Ramsey	0.800	0.879	0.942	0.921	0.911	0.960	0.879	0.971
Satake	0.964	0.909	0.953	0.942	0.932	0.879	0.737	0.954
Baker	0.933	0.920	0.963	0.948	0.884	0.980	0.899	0.960
Baumbach	0.933	0.830	0.984	0.932	0.890	0.990	0.889	0.889
Chung	0.944	0.838	0.932	0.916	0.869	0.848	0.525	0.624
Kong	0.974	0.978	0.963	0.916	0.884	0.980	0.919	0.971
Lundberg	0.974	0.545	0.811	0.706	0.627	0.778	0.545	0.711
Blalock	0.964	0.662	0.763	0.742	0.721	0.657	0.495	0.600
Paschaki	0.974	0.818	0.900	0.869	0.837	0.828	0.677	0.794
Ehlting	0.954	0.742	0.869	0.837	0.806	0.788	0.616	0.736
Fujitaa	0.964	0.943	0.974	0.963	0.932	0.980	0.879	0.989
Fujitab	0.974	0.755	0.848	0.806	0.763	0.848	0.687	0.898
Nagpal	0.964	0.880	0.953	0.932	0.911	0.889	0.737	0.927
Wang	0.974	0.703	0.784	0.653	0.521	0.818	0.667	0.874

Figura 4.17: Diferença entre os valores de limiar escolhidos por dois dados critérios. Esta comparação considerou os critérios que apresentaram maiores divergências na Figura 4.13. (a) *máxima entropia X tamanho do componente gigante 40%*, (b) *máxima entropia X distribuição em lei de potência*, (c) *tamanho do componente gigante 40% X tamanho do componente gigante 80%*, (d) *tamanho do componente gigante 80% X distribuição em lei de potência*, (e) *tamanho do componente gigante 80% X grau médio específico (1000)*, (f) *grau médio específico (100) X grau médio específico (1000)*. Fonte: Elaborado pela autora.

A comparação entre os critérios *Máxima entropia* e *distribuição de grau em power-law* (Figura 4.17b) culminou em resultados mais semelhantes para as medidas consideradas, porque os valores de limiar estão mais próximos uns dos outros, e as redes resultantes terão também uma topologia mais comparável. Esta semelhança é mais evidente em medidas relacionadas ao número de nós e grau (*número de nós* e *grau médio*). No entanto, para as medidas relacionadas à conectividade entre nós (*coeficiente de aglomeração*, *diâmetro*, *modularidade*, *tamanho médio da comunidade*) (Tabela 4.3), não notamos essa similaridade, o que sugere que essas redes, embora de tamanho comparável, possuem topologia diferente.

Comparando o critério *grau médio específico 100* com *grau médio específico 1000*, verificamos grandes diferenças entre as redes resultantes (Tabela 4.3), mesmo para medidas que não são diretamente relacionadas ao grau médio, como a *modularidade* e o *coeficiente de aglomeração* das redes.

Análise de Componentes Principais

O espaço definido pelas sete propriedades da Tabela 4.3, para cada critério de limiarização, foi projetado em um respectivo espaço bidimensional, usando o PCA. A medida de *número de nós do componente gigante* (*N giant*) e o *grau médio* não foram incluídos, pois são influenciados pelo limiar de forma trivial. A Figura 4.18a mostra os PCAs obtidos. O PCA é frequentemente usado porque permite a visualização da distribuição de dados, o que de outra forma seria impossível em dimensões superiores. No entanto, o PCA também é particularmente útil para identificar, com várias medidas, quais deles representam a maior parte da variação dos dados. A análise do peso de cada medida no PCA mostra que as medidas que contribuíram mais para a variação de PCA 1 é o *desvio padrão do grau médio*, seguido de *modularidade* e *R1*; Da mesma forma, as medidas com maiores pesos em PCA 2 são o *diâmetro*, *tamanho médio da comunidade* e *coeficiente de aglomeração*.

A medida *coeficiente de aglomeração* na base de Baumbach [87] resultou em valores altos (> 0.98) para todos os critérios. A pequena variação dessa medida para algumas bases de dados (ex.: Lundberg [90] e Fujita a [94]), deve-se provavelmente ao fato de os dados serem fortemente clusterizados, sugerindo uma rede de co-expressão gênica onde os genes influenciam altamente a expressão um do outro. Por outro lado, o *coeficiente de aglomeração* varia grandemente para outras bases de dados (ex.: Blalock [91] e Wang [96]), subentendendo um efeito contrário. Baixos valores de modularidade foram obtidos para as bases de Kong [89] (0.08 to 0.32) e Ramsey [84] (0.20 to 0.34).

Voltando à Figura 4.18 a, os resultados indicam que alguns conjuntos de dados se comportam de forma semelhante, ou seja, sua posição no PCA, em relação a outros conjuntos de dados, é similar para critérios distintos. Para descobrir quais conjuntos de dados possuem essa propriedade, foi selecionada, por inspeção visual, a região em cada PCA contendo a maior concentração de pontos. Então, calculou-se quantas vezes cada conjunto de dados apareceu nas regiões selecionadas. Os conjuntos de dados que aparecem

mais de quatro vezes foram: *Pardo* [77], *Reppe* [78], *Kobayashi* [82], *Blalock* [91], *Paschaki* [92], *Fujita b* [94] and *Nagpal* [95]. Esses conjuntos de dados pertencem ao que chamamos de *grupo conservado*, uma vez que se comportam de forma semelhante para os diferentes critérios. Em seguida, para cada critério, foi encontrada a distância entre cada conjunto de dados e o grupo conservado. Esse espaço foi calculado como a distância euclidiana mínima, no espaço PCA, entre conjuntos de dados no grupo conservado e o conjunto de dados de interesse. As distâncias foram classificadas para conjuntos de dados não incluídos no grupo conservado e uma matriz contendo os resultados é mostrada na Figura 4.18 b. Nesta matriz, cada linha corresponde a um critério. Para uma determinada linha, a primeira coluna contém o índice do conjunto de dados mais próximo do grupo conservado, a segunda linha contém o índice para o segundo conjunto de dados mais próximo, e assim por diante.

A matriz obtida mostra que as bases de dados que geralmente ficaram próximas ao grupo conservado foram *Satake* [85], seguida de *Ehlting* [93]. A base de *Postnikov* [83] também estava usualmente próxima ao grupo conservado, e a base de *Baumbach* [87] com frequência esteve distante desse grupo.

Figura 4.18: (a) PCAs das medidas obritdas para cada critério de limiarização. Note que cada ponto corresponde a uma respectiva base de dados. Consulte a legenda da Figura 4.13 para o significado de cada nome. (b) Matriz contendo as posições relativas dos experimentos ao grupo conservado (mais próximo a mais, da esquerda para a direita), cada linha corresponde a um critério. Fonte: Elaborado pela autora.

4.3.3 Discussão Geral

Considerando o grupo conservado da Figura 4.18, notamos que para cada critério, algumas medidas tiveram uma variação considerável, e outros seguiram um padrão. Para o critério *Tamanho do componente gigante 0.6*, a *modularidade* e o *diâmetro* foram as medidas que variaram menos, enquanto as medidas de *número de nós do componente gigante* e *tamanho médio de comunidade* tiveram uma variação alta. Para o critério *Grau médio específico (100)*, a *modularidade* teve uma flutuação pequena, e a medida de *número de nós do componente gigante* teve uma alta variação, assim como o *coeficiente de aglomeração* e *diâmetro*. Apesar de *Grau médio específico (100)* ser um critério que leva em consideração o grau médio da rede, a medida de *grau médio* teve uma variação significativa para as redes geradas com os dados das bases de Reppe [78] e Fujita b [94] sob esse critério. Isso ocorre porque o limite foi especificado de acordo com o grau médio esperado de todo o banco de dados, mas as medidas na Tabela 4.3 foram calculadas apenas para o componente gigante das redes resultantes. Para o critério *Grau médio específico (1000)*, o *grau médio* teve, como esperado, uma variação menor. Entretanto, o *desvio padrão* e outras medidas (*número de nós*, *coeficiente de aglomeração*, *diâmetro*, *modularidade*, *tamanho médio de comunidades*) variável consideravelmente. Para o critério *Entropia máxima*, o *coeficiente de aglomeração* e a *modularidade* foram as medidas que variaram menos.

É interessante observar que todos os grupos biológicos (matriz óssea, células cancerígenas, células neuronais, células vegetais) estão representados no grupo conservado, exceto os bancos de dados de Desenvolvimento em *Drosophila*. Isso pode implicar que os genes tenham um comportamento semelhante nestas experiências, com interações análogas entre os genes expressos.

A base de dados de Baumbach [87] gerou a rede com medidas que divergiram mais do grupo conservado. Esse trabalho investigou um cenário biológico em que os cromossomos do díptero *Drosophila* têm uma perda ou amplificação de centrômero. Embora seja esperado que as anomalias do centrômero induzam algumas mudanças, incluindo as transcricionais, os autores relatam que os defeitos do centrômero causaram poucas mudanças no transcriptoma global [87]. Nas análises aqui relatadas, no entanto, a rede de co-expressão gerada pelo perfil transcriptômico de *Drosophila* mostrou um comportamento distinto quando comparado a outras redes de co-expressão (Figura 4.18a), o que poderia sugerir algum tipo de efeito biológico.

4.4 Conclusões

Este Capítulo da Tese aborda a desafiadora tarefa da seleção de limiar para a construção de redes de co-expressão gênica. Embora seja um passo muito importante em Biologia dos Sistemas, há poucos trabalhos matemáticos relacionados abordando esse problema de forma mais sistemática e propondo possíveis metodologias. O objetivo principal dessa parte da Tese foi, portanto, desenvolver uma abordagem experi-

mental relativamente sistemática ao problema da seleção de limiar em redes de co-expressão, suportadas por avaliações numéricas. Também propusemos cinco métodos diferentes para seleção de limiar, dado um conjunto de dados. Cada um desses métodos tem um critério respectivo para eleger o limiar apropriado para cada conjunto de dados em questão. O impacto nas redes resultantes foi analisado comparando-se a topologia das redes obtidas através de várias medidas em relação às diferentes abordagens de seleção de limiar.

O método de *Máxima variação das propriedades topológicas* (descrito na Seção 4.2.2.a) gerou as redes com as variáveis topológicas mais distintas. Este método resultou em grandes variações no número de nós do componente gigante, grau médio e seu desvio padrão, e o tamanho médio da comunidade. Diferentemente, o método de *Grau médio específico (1000)* (Seção 4.2.2.e) Produziu redes com estrutura semelhante, mostrando um grau médio em torno de 1000, naturalmente, seguido por um coeficiente de aglomeração com baixa variação. Os outros métodos resultaram em redes com variações consideráveis de medidas, especialmente em relação à média e ao desvio padrão do grau.

A base de dados de Baumbach [87] pareceu não depender do método de seleção de limiar. Este conjunto de dados apresentou baixa variação de medidas topológicas para os diferentes critérios estudados neste trabalho. As bases de dados do “grupo conservado” [77, 78, 82, 91, 92, 94, 95] também apresentaram relativa independência do método de limiarização. Por outro lado, as bases de Lopez [80], Zhang [79], Abdul [81], Postnikov [83], Ramsey [84], Satake [85], Chung [88], Lundeborg [90], Fujita a [94] e Wang [96] apresentaram-se mais dependente do método aplicado.

Ao levar em consideração as constatações acima mencionadas, propomos algumas diretrizes que podem ajudar na escolha de limiares para redes de co-expressão gênica de interesse:

1. Nossos resultados mostram que, para alguns tipos de bases de dados, o limiar escolhido pode ter um grande impacto nas propriedades topológicas das redes resultantes. Portanto, uma atenção especial deve ser dada a esse aspecto em cada aplicação específica. Por outro lado, um bom acordo entre os resultados obtidos por critérios alternativos fornece algum suporte intrínseco à rede obtida, de forma semelhante ao conceito de *boosting* da inteligência artificial [107, 108];
2. Dada uma rede específica de interesse, a metodologia aqui descrita pode ser aplicada, considerando também outros bancos de dados relacionados (por exemplo, mesmas espécies, mesmo tipo de experiência, mesmo tecido/órgão, etc.). Isso permitiria algum contexto comparativo para identificar como os dados específicos comportam-se em relação a outras situações semelhantes;
3. No caso de a aplicação de interesse envolver um aspecto ou requisito de importância crítica não incluída nos métodos aqui relatados, seria possível elaborar uma nova variação que visa enfatizar essa característica nos dados originais, de forma semelhante às abordagens de Regularização, adotadas

em áreas como o aprendizado de máquinas e também a biologia [109, 110].

Deve-se observar que orientações mais específicas para limiarização de redes de co-expressão gênica exigiriam o acesso a bases de dados biológicas completas, que poderiam ser usadas como padrão de ouro de validação e abordagem comparativa. Infelizmente, a maioria das bases de dados biológicas, especialmente aquelas com natureza molecular, estão incompletas devido à própria limitação do método. Espera-se que a disponibilidade futura de dados novos e mais completos possa levar a complementações e aprimoramentos das três diretrizes sugeridas acima.

As técnicas matemáticas e computacionais foram progressivamente propostas e usadas para analisar dados biológicos, especialmente devido à crescente complexidade dos sistemas investigados. A abordagem aqui relatada pode colaborar na direção de obter melhores representações de cenários biológicos específicos. Aplicamos os princípios da Teoria das Rede Complexas no estudo de dados transcriptômicos, a fim de compreender melhor os efeitos da seleção de limiar em redes de co-expressão gênica, incentivando o desenvolvimento de novos critérios e métodos para comparação e validação respectivas. Os avanços mais imediatos dos resultados reportados aqui incluem, mas não estão limitados a, incorporação de outras bases de dados e de medidas topológicas adicionais.

Capítulo 5

Caracterização da rede de canais vasculares de ossos longos de camundongo

Este Capítulo está organizado da seguinte maneira: Primeiro, faço uma descrição do tecido ósseo e dos canais de Havers e Volkmann dos vertebrados. Em seguida, descrevo a natureza dos dados amostrais e o tratamento dessas amostras, passando pela análise de imagens até a reconstrução dos canais vasculares em três dimensões e finalmente a análise das redes de canais de osso de camundongo. Os resultados topológicos, geométricos e biológicos obtidos para cada osso são então apresentados e discutidos.

5.1 Introdução ao tecido ósseo

O tecido ósseo é constituído de tecido conjuntivo rígido e inflexível, no qual a matriz extracelular foi impregnada com sais de cálcio e fosfato por um processo de mineralização. Trata-se de um tipo de tecido altamente vascularizado e metabolicamente ativo. O esqueleto vertebrado é composto principalmente por tecido ósseo, servindo de suporte para as porções "moles" do corpo e como proteção para os órgãos vitais, como órgãos contidos na caixa craniana, torácica e canal raquidiano. Além disso, o esqueleto aloja a medula óssea hematopoiética ¹, proporciona apoio aos músculos esqueléticos e constitui um sistema de alavancas que amplia as forças geradas na contração muscular [8, 1].

¹Hematopoiese é o processo formador de novas células do sangue (células vermelhas, maioria das células brancas e plaquetas).

5.1.1 Classificação

O tecido ósseo apresenta classificações morfológicas importantes, considerando-se sua aparência macro ou microscópica. De acordo com seu aspecto macroscópico, duas formas podem ser diferenciadas:

Tecido ósseo compacto

Massa sólida que forma a maior parte dos ossos longos (exemplo, o fêmur) e a *diáfise*, uma haste cilíndrica central que delimita a *cavidade medular*.

Tecido ósseo esponjoso

Formado por uma trama de espículas (ou trabéculas) ósseas que delimitam espaços ocupados pela medula óssea. As extremidades dos ossos longos, também chamadas de *epífises*, constituem-se de *osso esponjoso* revestido por uma fina camada de *osso compacto*. Durante o crescimento, as epífises são separadas da diáfise por uma *placa epifisária cartilaginosa*, conectada à diáfise por osso esponjoso. A placa epifisária e o osso esponjoso adjacente representam a zona de crescimento, responsável pelo aumento (em comprimento) do osso. A *metáfise* é uma região de transição de formato progressivamente afilado que conecta a epífise à diáfise. A *cavidade medular* da diáfise e os espaços no interior do osso esponjoso são revestidos pelo *endósteo*². As *superfícies articulares*, nas extremidades dos ossos longos, são revestidas por *cartilagem hialina* (cartilagem articular). Com exceção das *superfícies articulares* e nos locais de inserção dos tendões e ligamentos, os ossos são revestidos pelo *periósteo*³, uma camada de tecido conjuntivo também especializado com potencial osteogênico.

Com base na organização *microscópica* da matriz extracelular, dois tipos de tecidos ósseos são identificados: o tecido ósseo lamelar primário e o secundário.

Tecido ósseo lamelar primário (ou tecido ósseo imaturo)

Tecido observado no osso em desenvolvimento, o qual é caracterizado por alinhamento irregular das fibras colágenas. Esse osso é mecanicamente fraco, formado rapidamente e substituído posteriormente pelo osso lamelar secundário.

O tecido ósseo é estruturado pela matriz óssea, composta por componentes orgânicos (35%) e inorgânicos (65%); fibras de colágeno do tipo I (90%); proteoglicanos, enriquecidos em condroitinsulfato, queratansulfato e ácido hialurônico; e proteínas não colagenosas. A porção inorgânica do osso compõem-se principalmente pelos depósitos de fosfato de cálcio com características cristalinas de hidroxiapatita. Os

²camada de células pavimentosas e fibras do tecido conjuntivo, servindo de revestimento para as paredes esponjosas que abrigam a medula óssea. Estendem-se para todas as cavidades ósseas, incluindo os *canais haversianos*.

³camada interna de células osteogênicas (*osteoblastos*) em contato direto com o osso, presente durante o crescimento embrionário e pós-natal. No adulto, contém células de tecido conjuntivo que retêm seu potencial osteogênico para reparo ósseo.

cristais são distribuídos ao longo do comprimento das fibras colágenas com auxílio das proteínas não colágenas.

O colágeno do tipo I é a proteína mais abundante da matriz óssea. No osso lamelar secundário, as fibras colágenas possuem um arranjo altamente organizado acompanhando mudanças da sua orientação de acordo com o eixo dos canais haversianos.

O tecido ósseo possui também componentes celulares. Quando encontra-se em crescimento ativo, o osso contém células de duas linhagens diferentes:

Linhagem osteoblástica. Inclui as *células osteoprogenitoras* de origem mesenquimal e possuem potencial para proliferação e capacidade de diferenciação (propriedades básicas de células-tronco), as quais dão origem aos *osteoblastos* através de processos que envolvem fatores de crescimento e de transcrição. Essa camada celular está presente na camada interna do periósteo e do endósteo, e persiste por toda a vida pós-natal, sendo reativadas no adulto durante o reparo de fraturas ósseas e outras lesões.

Os osteoblastos são células cúbicas que revestem todos os locais de formação óssea ativa. São células altamente polarizadas, e depositam *osteóide*, a matriz orgânica não mineralizada do osso, iniciando e controlando a mineralização subsequente do osteóide. Seus produtos específicos são *colágeno do tipo I*, *osteocalcina*, *osteopontina* e *sialoproteína óssea*, além de fatores de crescimento.

Após serem aprisionados dentro de lacunas dentro da matriz mineralizada, os osteoblastos diferenciam-se em *osteócitos*, processo que envolve a participação dos fatores de transcrição *Cbfa1/Runx2* e *osterix*. Portanto, quando o osso está formado, os osteoblastos transformam-se em osteócitos, células ramificadas cujo corpo ocupa pequenos espaços entre as lamelas, as lacunas. As lamelas são percorridas por pequenos canais que interconectam as lacunas vizinhas formando *junções comunicantes*. Os nutrientes difundem-se dos vasos sanguíneos (localizados no interior do canal haversiano) através dos canalículos nas lacunas. O tempo de vida de um *osteócito* depende da difusão dos nutrientes, e a vida da matriz óssea depende dos osteócitos. Os osteócitos podem permanecer vivos por anos se o fornecimento vascular for contínuo.

Linhagem osteoclástica. Os *osteoclastos* derivam da linhagem celular progenitora de monócito - macrófago na medula óssea, diferentemente da via progenitora osteoclástica. Os *monócitos* trazidos ao osso pela circulação sanguínea são as células precursoras dos osteoclastos. Elas fusionam-se em células multinucleadas com até 30 núcleos, formando os osteoclastos por meio de um processo regulado pelos osteoblastos e pelas células do estroma da medula óssea. O osteoclasto resultante é uma célula grande (até 100 µm de diâmetro) e altamente polarizada que ocupa uma concavidade rasa chamada *lacuna de Howship* ou *compartimento subosteoclástico*.

Quando se ligam à matriz óssea-alvo, os osteoclastos produzem um ambiente ácido necessário para a reabsorção óssea, a qual envolve a dissolução dos componentes inorgânicos do osso (desmineralização

óssea) mediada pela H^+ -ATPase (adenosina trifosfatase), seguida pela degradação enzimática da matriz orgânica (colágeno do tipo I e proteínas não colagenosas) pela protease catepsina K.

Os osteoclastos são essenciais para o remodelamento e renovação óssea, processos que envolvem a remoção da matriz óssea em vários locais, seguida pela sua substituição por novo osso pelos osteoblastos.

Tecido ósseo lamelar secundário (ou tecido ósseo maduro)

Consiste em lamelas, amplamente constituídas de *matriz óssea*, uma substância mineralizada depositada em camadas (lamelas), e *osteócitos*, células que ocupam individualmente uma cavidade com canaliculos ramificados e irradiados que penetram nas lamelas das lacunas adjacentes. Também possui alinhamento regular das fibras colágenas. Esse osso é mecanicamente forte e é formado lentamente.

5.1.2 Processos de ossificação

Dois processos guiam o desenvolvimento do esqueleto vertebrado: a *ossificação intramembranosa* e a *ossificação endocondral* [8, 1, 111]. A *ossificação intramembranosa* forma o tecido ósseo esponjoso e ocorre em poucos tipos de ossos (como a porção lateral das clavículas e os ossos chatos do crânio). Seu desenvolvimento dá-se pela condensação direta de células progenitoras mesenquimais ⁴, as quais se diferenciam em osteoblastos, cuja secreção forma a matriz óssea [112]. Numerosos centros de ossificação desenvolvem-se e fundem-se, formando uma rede de trabéculas anastomosadas semelhantes a uma esponja. A conversão desse tecido ósseo primário em tecido ósseo lamelar é feita em estágios subsequentes do desenvolvimento vertebrado [1].

A maior parte do esqueleto mamífero é formada por osso compacto, desenvolvido pelo processo de osteogênese por *ossificação endocondral*, o qual envolve uma série de eventos que se iniciam no estabelecimento de um molde de cartilagem hialina (Figura 5.1a). Células precursoras mesenquimais são recrutadas para a linhagem condrogênica, iniciam a formação de condensados celulares, diferenciam-se em condroblastos e, depois de aprisionados em lacunas na matriz extracelular, são denominados condrócitos. Na segunda fase, os grupamentos de condrócitos recém formados tornam-se mitoticamente ativos até saírem do ciclo celular, para então hipertrofiar e terminar sua diferenciação. Em seu estágio final, os condrócitos sofrem apoptose e deixam uma matriz extracelular calcificada, formando assim a cartilagem hialina. Extremidades cegas de capilares penetram nas cavidades formadas no interior da cartilagem ossificada. Essa cartilagem é substituída pela atividade coordenada de osteoblastos, osteoclastos e células endoteliais trazidos por circulação sanguínea nascente (Figura 5.1c) [8, 1, 113, 114, 115]. As últimas fases de proliferação e diferenciação de condrócitos ocorrem na placa de crescimento epifisária (Figura 5.2b), na interface entre cartilagem e osso, uma região controladora do crescimento longitudinal dos ossos endocondrais por meio

⁴Mesênquima é o tecido mesodérmico embrionário indiferenciado dos vertebrados, que origina os tecidos conjuntivos no adulto.

da multiplicação e hipertrofia de condrócitos [114].

Figura 5.1: Esquema do centro de ossificação primária do processo de ossificação endocondral. A cartilagem hialina é um molde para o osso longo que será formado. (b) A proliferação dos condrócitos e sua consequente hipertrofia, iniciam a formação do centro de ossificação primário. (c) Esses condrócitos hipertróficos são responsáveis pela secreção do fator de crescimento endotelial vascular, para induzir o brotamento dos vasos sanguíneos a partir do pericôndrio. Depois desses eventos, ocorre a calcificação da matriz do centro de ossificação e a indução da apoptose dos condrócitos hipertróficos. Os vasos sanguíneos que chegaram a partir do pericôndrio formam o broto periosteal, que se ramifica em direção às diáfises. Fonte: elaborada pela autora, com base em Kierszenbaum [1].

Em etapas mais avançadas da ossificação endocondral, vasos sanguíneos invadem o espaço ocupado anteriormente por condrócitos hipertróficos, ramificam-se e projetam-se na direção das extremidades desse centro de ossificação (Figura 5.2a). Em momentos posteriores do desenvolvimento ósseo, vasos sanguíneos e mesênquima infiltram-se nas epífises (apontadas na Figura 5.1a e 5.2c) para formar um centro de ossificação secundário (Figura 5.2) primeiro em uma epífise, depois na outra.

A placa epifisária de crescimento é uma estrutura cartilaginosa que fica entre a metáfise e a epífise (Figura 5.2), e a proteína *indian hedgehog* (Ihh) tem o papel de estimular a proliferação dos condrócitos na placa epifisária de crescimento, para impedir a hipertrofia dos condrócitos que lá se encontram. Ao longo do desenvolvimento ósseo, essa placa é eventualmente substituída por matriz óssea mineralizada e torna-se uma parte particularmente densa do osso, a linha epifisária. Esse processo acontece gradualmente ao longo do crescimento do indivíduo (da puberdade à maturidade), e implica no término do crescimento em

Figura 5.2: Esquematização do processo de ossificação secundária dos ossos longos. (a) O colar periosteal (Figura 5.1b) estende-se ao longo da diáfise óssea e vasos sanguíneos infiltram-se nas epífises para formar o centros de ossificação secundários. (b) Esses centros desenvolvem-se e vão preenchendo as epífises. (c) As placas epifisárias de crescimento são substituídas pela linha epifisária, marcando o final do crescimento em comprimento do osso. Fonte: elaborada pela autora, com base em Kierszenbaum [1].

comprimento do osso.

Com esse encontro dos centros de ossificação primário e secundários, os vasos sanguíneos da diáfise e epífise se encontram e intercomunicam-se (Figura 5.2c). As cartilagens epifisária e diafisária, com exceção da superfície articular, serão então substituídas por tecido ósseo. As células osteoprogenitoras iniciam esse processo de substituição, elas são direcionadas ao centro de ossificação por meio do tecido conjuntivo perivascular, e por fim diferenciam-se em osteoblastos que se acumulam nas superfícies de cartilagem calcificada e começam a depositar matriz óssea, formando lamelas dispostas concentricamente ao redor de um canal longitudinal que circunda uma estrutura neurovascular [116] Essas estruturas são chamadas de *Sistemas de Havers* (ou ósteons) (Figura 5.3), as quais se dispõem em torno dos chamados *Canais de Havers* e são conectados entre si por meio de canais transversais, denominados *canais de Volkmann*, cujo conteúdo consiste em vasos sanguíneos da medula e do periósteo (região osteogênica que envolve os ossos)[8, 1, 115].

5.1.3 O Sistema Harvesiano

Figura 5.3: Esquemática do sistema Harvesiano. Fonte: elaborada pela autora.

Semelhantemente aos vasos do sistema linfático, esse sistema vascular presente no interior dos ossos possui topologia em anastomose, diferindo do restante dos sistemas vasculares do corpo, por tratar-se de uma rede de canais que se bifurcam e recombina em vários pontos, ao invés das dicotomias que formam as “árvores vasculares”. Essas redes dos canais vasculares são continuamente “remodeladas” por proces-

sos que incluem a quebra por osteoclastos e reconstrução por osteoblastos, cuja regulação dá-se por fatores genéticos, biológicos e biomecânicos. O último processo citado é crucial durante o desenvolvimento, porque forças físicas aplicadas em tecidos podem causar mudanças significativas no metabolismo celular e expressão gênica [117].

Nos ossos compactos de adultos, novos canais são criados pela atividade dos osteoclastos. Esses canais são mais largos do que o necessário, e em seguida, células osteoblásticas preenchem esses “túneis” com nova matriz óssea, até que o diâmetro da cavidade seja reduzido a um único canal vascular [9, 118].

5.1.4 Alguns aspectos genéticos da osteogênese

Os sucessivos eventos de proliferação e diferenciação celular, que produzem todos os estágios observados na placa de crescimento óssea, são regulados por redes de sinalização envolvendo interações entre matriz intracelular e extracelular. A compreensão sobre essa regulação gênica vem aumentando consideravelmente nos últimos 15 anos [114, 119, 120, 121]. Proteínas de diversas famílias foram relacionadas ao crescimento longitudinal do esqueleto, entre elas incluem-se: fatores de crescimento; moléculas de matriz extracelular; e fatores de transcrição (do inglês, transcription factor com a sigla TF). O TF Sox9 mostrou-se necessário para a condrogênese; o TF ATF-2 controla a progressão do ciclo celular e proliferação; e o TF Cbfa1/Runx2 é envolvido na diferenciação hipertrófica [1, 114, 119].

O desenvolvimento ósseo bem sucedido demanda a coordenação entre vários fatores (condrogênicos e osteogênicos) para haver estabelecimento e manutenção normais da placa de crescimento e subsequente estrutura óssea. A regulação desordenada de genes pertencentes a categorias funcionais foi relacionada a distúrbios de crescimento e condições patológicas, por exemplo, camundongos deficientes em Cbfa1/Runx2 possuem um esqueleto constituído de cartilagem, sem indicação de diferenciação osteoblástica ou de formação óssea. Outro exemplo é a osteomalácia, uma doença caracterizada pelo amolecimento progressivo e arqueamento dos ossos, processo decorrente de um defeito na mineralização da matriz óssea devido à deficiência de vitamina D ou a distúrbios hereditários de ativação dessa vitamina, como a deficiência da enzima 1α -hidroxilase renal, que resulta na não conversão de calciferol na sua forma ativa, o calcitrol [1].

Como citado anteriormente nesta Tese, os ossos são compostos por uma matriz orgânica principalmente colagenosa e mineralizada com cristais inorgânicos de hidroxiapatita. Essa matriz inorgânica é desmineralizada e remineralizada constantemente. A matriz óssea desmineralizada é um biomaterial bioativo e biodegradável, que induz a morfogênese ⁵ do osso em sítios extraesqueléticos ectópicos. Essas características dão ao tecido ósseo a posição de tecido com o maior potencial para regeneração, dentre todos os tecidos do corpo humano [122]. O sinal bioativo responsável por todos esses processos, é a Proteína Óssea Morfogenética.

⁵Desenvolvimento das formas e estruturas características de uma espécie a partir do embrião.

Proteína Óssea Morfogenética (*Bones Morphogenetic Proteins - BMP*)

As proteínas ósseas morfogenéticas (BMPs) são um grupo de fatores de crescimento também conhecidos como citocinas e metabológenos (*metabologens*) [122]. Os metabológenos são moléculas morfogenéticas que podem iniciar, promover e manter o metabolismo e a homeostase. Essas proteínas foram definidas pela sua capacidade de induzir a formação de osso e cartilagem, e atualmente são consideradas como parte de um grupo de sinais morfogenéticos fundamentais, participativas da arquitetura de tecido no corpo todo [123]. A regulação do desenvolvimento é uma integração entre as famílias sinalizadoras de morfogênese, como as BMPs, família Fgf, *Hedgehogs* e Wnts.

Vinte genes do genoma humano codificam BMPs funcionais [124], e a morfogênese óssea é uma reação em cadeia multisequencial, cujos passos chave são a quimiotaxia⁶ de células tronco progenitoras, proliferação celular e diferenciação de formação óssea endocondral

Dentre as funções conhecidas das BMPs temos: desenvolvimento de cartilagem (BMP1, BMP2, BMP5, BMP8a); desenvolvimento ósseo (BMP2, BMP3, BMP4, BMP8a); diferenciação osteoblástica (BMP2, BMP7); desenvolvimento dentário (BMP4); desenvolvimento de membros de origem mesodérmica (BMP4); papel no reparo de fraturas, na formação da epiderme na formação do eixo dorso-ventral, e desenvolvimento do ovário folicular (BMP4); controle da homeostase do ferro (BMP6); papel na integridade das articulações no adulto (BMP6); papel chave no desenvolvimento e reparo renal (BMP7); expressão no hipocampo (BMP8b); trabeculação do coração embrionário (BMP10); controle da padronização do eixo antero-posterior (BMP11); desenvolvimento do oócito e desenvolvimento folicular (BMP15) [124]. Tecidos que sofrem constante renovação celular geralmente apresentam expressão de genes BMP, como por exemplo o estômago [123].

Mutações nas vias das BMPs são a causa da síndrome de polipose herdada (*inherited polyposis syndromes*), bem como o evento de iniciação do desenvolvimento de câncer gastrointestinal esporádico [125, 126]. As vias de sinalização das BMPs também são conhecidas por suprimir oncogênese no intestino delgado e grosso de camundongos e humanos [123], e no colo e reto [127].

As BMPs estão sendo usadas clinicamente. Essas proteínas são produzidas através da tecnologia de DNA recombinante (BMP humana recombinante - rhBMPs) [128, 129].

5.1.5 Estrutura de redes em ossos

O sistema vascular encerrado nos canais de *Havers* e *Volkman* é de extrema importância para a manutenção do tecido ósseo. Se o transporte sanguíneo não for eficiente, as células podem não ter sua demanda nutricional suprida, e o osso eventualmente entraria em isquemia. Ferramentas matemáticas e computacionais foram utilizadas para o estudo de diferentes aspectos do esqueleto [116, 130, 131]. O trabalho de

⁶Processo de atração de células em direção a um gradiente químico (estímulo químico).

Cooper e colaboradores [116] introduz o uso da microtomografia computadorizada (cuja sigla é microCT ou μ CT) para quantificação de aspectos da arquitetura tridimensional (3D) da rede de canais do sistema Harvesiano, e demonstram que essa rede é uma estrutura complexa que não pode ser apreciada totalmente somente com seções 2D. Esses autores mostraram que parâmetros relativos ao tamanho, espaçamento e volume dos canais são úteis na investigação das propriedades mecânicas do osso. Além disso, eles registraram que as mudanças promovidas pelo remodelamento ósseo podem ser avaliadas por parâmetros descritores da arquitetura 3D da rede de canais, tal qual a conectividade entre canais.

Em trabalhos anteriores do presente grupo de pesquisa⁷ o sistema vascular do esqueleto de falange de galinha mostrou-se com características topológicas interessantes [9] e estudar outros aspectos dessas redes pode responder perguntas que permaneceram abertas no trabalho citado. O trabalho de Viana et al. [9] utilizou visualização científica e análise de imagens para reconstrução 3D do sistema vascular de um osso completo de galinha, o que permitiu a identificação e quantificação de padrões estruturais importantes na organização da conectividade entre os canais. A rede formada mostrou-se altamente modular, com várias comunidades com padrões distintos de conectividade. Os autores notaram que há comunidades progressivamente mais conectadas ao longo do comprimento do osso, o que sugere a presença de um gradiente de fatores de transcrição. Essas estruturas identificadas nos canais do osso foram relacionadas com fatores moleculares do desenvolvimento, anormalidades e patologias ósseas.

Um aspecto interessante do trabalho mencionado [9] é a observação de que as comunidades localizadas na porção posterior do osso tendem a ser mais densas, com grau médio de 2.75, ao passo que as comunidades na porção anterior são mais similares a árvores, com menos ciclos. Essa questão ainda não teve uma resposta do ponto de vista biológico, mas os autores levantaram algumas hipóteses. A região posterior do osso investigado (falange) apresenta duas protusões que são particularmente importantes para as propriedades mecânicas do osso, e é possível que as maiores pressões e tensões sobre essa região podem ser responsáveis pela maior conectividade das comunidades. As diferenças na densidade das conexões observadas ao longo do eixo pósterio-anterior também podem ser uma consequência de diferentes estágios do desenvolvimento, de modo que as comunidades menos cíclicas da região anterior tornem-se mais densas posteriormente.

A caracterização da conectividade dos canais vasculares de ossos é crítica para uma melhor compreensão de como mudanças moleculares e disfunções podem levar ao desenvolvimento anormal e a doenças. Com a associação das duas abordagens citadas anteriormente (análises genética e morfológica), os efeitos finais de alterações moleculares (p.e. mutações) podem ser investigados ao nível microscópico através do uso de abordagens baseadas em redes, pois a topologia do sistema de canais vasculares é imediatamente relacionada à sinalização célula a célula dentro do osso, sendo afetada por sinalização molecular durante

⁷<http://cyvision.ifsc.usp.br/Cyvision/>

a ontogênese e estágios mais avançados de vida [9].

Kollmannsberger e colaboradores [132] também trata de modelagem de estruturas ósseas sob a teoria das redes complexas. Os autores mapearam a rede lacuno-canalicular osteocítica (*osteocyte lacuno-canalicular network* - OLCN) de fêmures de camundongos e ovelhas. Como mencionado anteriormente na Seção 5.1.2, o tecido ósseo mineralizado é densamente populado por osteócitos. Essas células são embebidas na matriz extracelular óssea durante a deposição óssea, e são ligadas umas com as outras e com a circulação sanguínea através de uma rede altamente interconectada [1], comparada em aparência e tamanho com o sistema neuronal [133, 132]. Os corpos celulares dos osteócitos residem em *lacunas* ocas na matriz óssea mineralizada, e extensões dessas células intercomunicam-se através de *canalículos* que ligam as lacunas. Essa rede é responsável por permitir uma comunicação intercelular mais eficiente quando comparada à difusão através da matriz óssea. Entre os processos que ela guia, estão a homeostase mineral, a mecano recepção, e comunicação celular [1, 132].

O trabalho citado acima foi o primeiro a detalhar quantitativamente as propriedades topológicas e estatísticas da OLCN ao nível celular. Os autores utilizaram medidas quantitativas robustas para determinar o quão eficiente é a organização dessas redes com respeito ao transporte intercelular e comunicação. Com suas análises, eles concluíram que a arquitetura topológica da OLCN é idêntica entre os tipos de tecido analisados, sugerindo assim um mecanismo universal de crescimento durante a formação da rede. Esse tipo de estudo provê compreensão sobre quais princípios podem atuar quando células ósseas individuais se auto organizam em uma grande rede durante a formação óssea e a mineralização.

O objetivo dessa parte do trabalho, portanto, é a caracterização do sistema Harvesiano de ossos longos de camundongo, visando o detalhamento topológico e geométrico dessa estrutura, sob a luz da Teoria das Redes Complexas. No presente documento, também nos referimos às redes do sistema Harvesiano como *redes vasculares de osso*.

5.2 Material e Métodos

Os canais vasculares localizados no interior de ossos de camundongos, foram adquiridos por meio de microtomografia computadorizada, e armazenados em computador na forma de imagens digitais de duas dimensões (2D). Como essa rede de ossos é tridimensional (3D), as imagens adquiridas contém o corte transversal do sistema em diferentes profundidades, sendo que cada imagem corresponde a uma seção da estrutura total. Essas imagens podem então ser agrupadas para gerar o volume dos canais vasculares, que passarão por um processo de reconstrução 3D para gerar a representação computacional desse sistema. Para realização dessa conversão 3D - 2D - 3D (Figura 5.4), uma série de passos de processamento de imagens é necessária, os quais serão citados a seguir (Figura 5.5).

Figura 5.4: Representação da conversão da rede dos canais vasculares adquiridos por microtomografia (a) para a rede complexa (b). As medidas ($m_{1.1}, m_{1.2}, m_{1.3}, \dots$) representam o conjunto de medidas obtidas para um determinado nó. As reticências ao redor da figura mostram que esses conjuntos de medidas serão obtidos para todos os nós da rede. Fonte: elaborada pela autora.

Figura 5.5: *Workflow* com os passos de aquisição de imagens, processamento, e aquisição da estrutura 3D. Fonte: elaborada pela autora.

Tabela 5.1: Indivíduos amostrados nesse estudo, e suas respectivas nomeações dos ossos.

Indivíduo 1	1A, 2A	fêmur
	1B, 2B	úmero
Indivíduo 2	3A, 4A	fêmur
	3B, 4B	úmero
Indivíduo 3	5A*, 6A	tíbia, fêmur
	5B, 6B	úmero
Indivíduo 4	7A*, 8A*	tíbia
	7B, 8B	úmero

5.2.1 Aquisição de imagens tomográficas de osso de camundongo

Oito conjuntos de patas dianteiras e traseiras de camundongo (quatro animais) foram disponibilizadas pelo Prof. Dr. Marcelo Emílio Beletti, do Laboratório de Histologia da UFU, e constituem o conjunto de amostras testadas (CEUA-UFU processo 020/16). Durante a manipulação do material biológico, um fêmur quebrou-se e não foi adicionado ao conjunto amostral. O experimento com todas essas amostras foi feito no período de 5 a 23 de outubro de 2015, e todas as análises foram feitas de só uma vez, para que não houvesse variação temporal na fonte de raio X do aparelho de microCT.

Convencionamos que cada animal teria seus ossos numerados da seguinte maneira: os números 1 e 2 representam o par de fêmur e úmero, sendo que as letras "A" e "B" especificam se esse par é traseiro ou dianteiro (fêmur ou úmero, respectivamente). Os números 1 e 2 identificam o primeiro indivíduo, 3 e 4, o segundo indivíduo e assim por diante (Tabela 5.1). Importante notar que essa numeração não identifica se a pata é direita ou esquerda. Algumas amostras de tíbias foram incluídas na análise, e também serão devidamente identificadas.

Um total de 4 fêmures, 3 tíbias e 8 úmeros de camundongos foram escaneados em microtomógrafo (microCT) de raio-X modelo Skyscan 1172 (Aartselaar, Bélgica) na EMBRAPA Instrumentação - São Carlos, em colaboração com o Prof. Dr. Paulo Cruvinel. Os parâmetros utilizados foram: resolução de 2.91 μm por pixel, filtro de alumínio com cobre, *rotation step*⁸ de 0.2, *average framing*⁹ de 3. A distância aproximada entre cada uma das imagens da tomografia é de 3 μm .

As amostras de ossos foram divididas em segmentos de acordo com o tamanho total do osso. A Figura 5.7 contém a legenda identificando essas partes. As partes "b"s correspondem necessariamente à porção *distal*¹⁰ dos ossos, e as respectivas partes extremamente opostas correspondem à porção *proximal*¹¹.

⁸*Rotation step* determina o grau de rotação ao qual a câmera vai capturar a imagem. Esse grau pode ser maior ou igual a 0.5°. Quanto menor o número, mais imagens serão adquiridas, e maior a duração total do escaneamento.

⁹*Average framing* detecta o número de quadros (*frames*) cuja média produzirá uma figura final. Aumentando o *average framing* a imagem final terá maior resolução de detalhes, e a duração total do escaneamento será maior.

¹⁰diáfise da articulação do joelho (fêmur), do cotovelo (úmero) ou pé (tíbia).

¹¹diáfise da inserção na bacia (fêmur), no ombro (úmero) ou da articulação do joelho (tíbia).

Figura 5.6: Legenda das partes dos fêmures (a) e úmeros (b) usada nas demais seções. As partes "b"s correspondem necessariamente à porção *distal* e as respectivas partes opostas correspondem à porção *proximal*. Fonte: elaborada pela autora.

Figura 5.7: Projeção lateral do osso úmero 3B. Essa figura mostra as partes nas quais o osso foi dividido nas análises. Destaca-se na imagem a *tuberosidade deltoíde*, estrutura que auxiliou no posicionamento e orientação das amostras. A estrutura branca na parte superior da figura é a massa de sustentação para o apoio no microCT. Fonte: elaborada pela autora.

Dificuldades durante a aquisição das imagens tomográficas de osso de camundongo

A aquisição das imagens tomográficas de osso de camundongo foi uma etapa particularmente desafiadora no presente doutorado. Nesta Seção, vou descrever esses desafios e como eles foram superados.

O primeiro problema durante a etapa de aquisição das imagens tomográficas aconteceu durante a manipulação do osso número 4A, um fêmur. Esse osso partiu-se no momento de seu posicionamento no aparelho de microCT, inviabilizando as análises do mesmo.

O segundo problema foi relacionado à massa de sustentação da amostra (porção superior da Figura 5.7, identificado como "massinha de sustentação"). Na Figura 5.7, a parte branca superior está identificada com esse nome. Essa massa acrílica tem a função de ancorar e imobilizar o osso que vai ser submetido ao escaneamento em microCT. Seu uso não pode ser dispensado, porque sem ela, o osso poderia se deslocar durante o processo, inviabilizando a análise.

A presença dessa massinha (convencionada na parte "a" dos ossos) modificou a distribuição das cores da imagem, e a imagem final de canais 5.9(h) ficou demasiadamente diferente do restante das partes ("b", "c", "d", "e"). Por esse motivo, optamos por remover as partes "a" das análises, para evitar enviesamento das comparações. Essa decisão não representou uma perda considerável, porque nas partes "as" englobam a região epifisária do osso, local de predominante ossificação trabecular, na qual não conseguimos identificar os canais de Havers e Volkmann. A região final da última parte do osso (região "b", ou "c", ou "d", ou "e") também é removida para retirada da porção epifisária do osso.

5.2.2 Processamento das imagens tomográficas de osso de camundongo

O processamento das imagens adquiridas (exemplo na Figura 5.9(a)) iniciou-se com a segmentação pelo critério local da média. Esse método seleciona o limiar como uma média da distribuição local de escala de cinza. Se o pixel em questão tiver valor maior do que a média dos pixels da vizinhança, então o valor de pixel torna-se 255 (branco); caso contrário, o valor de pixel torna-se 0 (preto). Utilizamos uma variação desse método que usa *média* - C , onde C é uma constante. Com isso, a imagem que antes tinha suas cores distribuídas ao longo de uma escala de cinza, passa a ser binária (valores de 0 para a cor preta, e valores de 1 para a cor branca).

A Figura 5.8 representa o modo de funcionamento da segmentação da imagem pela média dos valores dos pixels da vizinhança. Suponha que cada uma das células do *grid* da imagem representa um *pixel*, e todos esses *pixels* em conjunto compõem uma imagem. Os valores dentro de cada um dos *pixels* equivale a uma cor na escala de cinza, que vai de 0 (preto) a 255 (branco). O pixel destacado em azul passará pelo processo de binarização: se o seu valor for menor ou igual da média dos *pixels* da vizinhança (os quadrados em laranja), seu valor será 0, e portanto, a sua cor será preta; ao contrário, se o seu valor for maior do que a média dos *pixels* da vizinhança (os quadrados em laranja), seu valor será 1, e portanto, a sua cor será

branca. Esse processo é repetido em todos os pixels da imagem, resultando em uma figura binária.

Figura 5.8: Exemplificação do método de segmentação pela média. O pixel em questão está destacado em azul, e tem valor de 56 dentro da escala de cinza. A média da vizinhança (destacada em laranja) é de 109,375 $((7 + 78 + 156 + 59 + 48 + 198 + 255 + 74)/8)$, destacado em roxo), maior que 56. Portanto, o pixel de valor 56 passa a ter valor 0. O novo valor de pixel (0) é armazenado em uma matriz binária, e a "janela" de 3x3 pixels move-se para o próximo pixel, para que sua vizinhança seja analisada. Fonte: elaborada pela autora.

As imagens obtidas por raio X são naturalmente ruidosas, e a etapa de remoção de ruídos é feita após a binarização da imagem (Figura 5.9(b)). Esses ruídos foram detectados por tamanho de componente conexo e posteriormente removidos (Figura 5.9(c)). Existem, no entanto, estruturas dentro da medula óssea que não são de natureza mineralizada, mas que são maiores que 100 pixels estipulado (Figura 5.9(d)). Essas estruturas foram removidas com o auxílio do *software* Fiji [134] com a função de rotulação dos tamanhos de componentes, e eliminação de todos os componentes, exceto o maior deles, o qual corresponde à matriz óssea.

Quando a imagem está livre de ruídos, é possível aplicar uma máscara para separar o objeto do plano de fundo (Figura 5.9(e)), e nessa imagem, a cor negra corresponde ao fundo, branco à matriz óssea, e cinza aos canais vasculares e medula óssea. A partir dessa classificação, é possível remover o componente relativo à matriz óssea, resultando numa imagem somente com a medula óssea e os canais vasculares da matriz (Figura 5.9(f)).

Apesar de a medula óssea representar um grande canal vascular que se conecta com grande parte dos canais menores, nós desejamos reconstruir a rede de canais de Havers e Volkmann somente, por esse motivo removemos a medula das análises (Figura 5.9(g)), um processo feito através da mesma máscara do passo anterior utilizada ao contrário, ou seja, no objeto e não no fundo da figura. Ao final de todas

essas etapas de processamento, são obtidas as imagens com os cortes transversais dos canais de Havers e Volkmann (Figura 5.9(h)).

Figura 5.9: Etapas de processamento para todas as imagens obtidas por microCT. Fonte: elaborada pela autora.

O processamento das imagens, reconstrução do esqueleto e análise de redes foi implementada em Python [135]. Na etapa de processamento de imagem, o *software* Fiji [134] também foi utilizado.

5.2.3 Reconstrução do Esqueleto 3D a partir das Imagens de microCT

As imagens dos canais (Figura 5.9(h)) podem ser agrupadas em sequência ("empilhadas") para reconstrução da estrutura 3D dos canais vasculares do osso, sendo que cada imagem corresponde a um plano. A partir da estrutura reconstruída, a aquisição do esqueleto da imagem é feita usando o algoritmo de afinamento (*thinning*) [136]. O propósito desse tipo de algoritmo é converter um objeto em sua forma mais simples, mantendo a topologia original. O esqueleto está relacionado com a estrutura essencial de cada formato, provendo uma representação mais estreita possível de uma dada forma, e posicionada na região mais central da figura [2].

Um dos conceitos matemáticos fundamentando a definição de esqueleto é conhecida como transformada de eixo medial (do inglês *medial axis transform* — MAT) ou transformada de eixo simétrico (do inglês *symmetry axis transform* — SAT) [137]. A SAT de uma forma contínua específica corresponde a todas as possíveis posições do centro de círculos satisfazendo as seguintes condições (Figura 5.10): (i) x é bitangente à forma (toca a curva em dois pontos distintos); (ii) e e y é completamente envolvida pela forma.

Figura 5.10: Uma forma simples (a) e seu respectivo SAT (b). Fonte: elaborada pela autora, adaptado de Costa e Marcondes [2].

5.2.4 Reconstrução e aquisição da estrutura 3D

O próximo passo é obter a rede complexa (topologia) a partir da estrutura 3D, e o princípio desse método consiste em identificar a vizinhança dos *pixels* do objeto. Cada pixel do esqueleto com três ou mais vizinhos foi classificado como uma bifurcação, e *pixels* com apenas um vizinho são identificados como pontos terminais (Figura 5.11). Essas posições de bifurcação e terminação são identificados como **nós**. Os *pixels* com dois vizinhos são considerados segmentos triviais da rede, e são identificados como **arestas**, conectando um nó ao outro. Segmentos demasiadamente curtos foram eliminados da rede, porque ruídos e irregularidades na estrutura podem gerar pequenos segmentos espúrios.

Depois da reconstrução estrutural da rede de canais vasculares, a mesma pode ser representada como uma matriz de adjacência binária $N \times N$, sendo que N é o número de nós da rede, e consequentemente, o número de linhas e colunas dessa matrix. Essa representação mostra se determinado par de vértices está conectado (representado pelo número 1) ou não (representado pelo número 0) (Figura 5.12).

Figura 5.11: Representação dos pixels de uma imagem e seus vizinhos. Na figura, os quadrados em branco representam os pixels do fundo da imagem, e os pixels em preto representam o objeto da imagem. O número em branco informa quantos vizinhos um determinado pixel possui. Fonte: elaborada pela autora.

Figura 5.12: Grafo representado à esquerda da figura, com seus nós numerados de 1 a 8. Respectiva matriz de adjacência à direita. Como o grafo não é direcionado, essa matriz é simétrica. Fonte: elaborada pela autora.

5.2.5 Visualização da Redes

Depois que a rede foi obtida e convertida para formato apropriado, o resultado imediato é a visualização da rede através do software *Network 3D*®, desenvolvido em 2012 pelo colega Filipi Nascimento Silva. Utilizando o software *Network 3D*, as redes vasculares foram visualizadas e também armazenadas como imagem (Figura 5.13). Com esse procedimento, temos a primeira etapa de verificação de qualidade (visual) e também é possível verificar se existe alguma estrutura interessante, principalmente em relação aos graus dos nós da rede.

Figura 5.13: Visualização da rede de canais vasculares (osso 1A femur, parte b). Note que as posições geográficas dos canais estão mantidas. Existe uma concentração interessante de *hubs* na parte lateral do osso que foi evidenciada por esse tipo de visualização. As cores dos nós e arestas representam os graus dos nós (conforme a legenda ao lado de cada figura). Fonte: elaborada pela autora.

(a) Visualização dos canais vasculares do osso 1A femur, parte c.

(b) Visualização dos canais vasculares do osso 1A femur, parte d. A crista presente nessa rede corresponde aos canais vasculares localizados no terceiro trocânter.

Figura 5.14: Visualização da rede de canais vasculares. Essas partes de osso correspondem à continuação da peça mostrada na Figura 5.13. Fonte: Elaborado pela autora.

5.2.6 Medidas Topológicas

As matrizes de adjacência de cada parte de osso também foram convertidas para o formato do *igraph*, uma coleção de bibliotecas do pacote *Python* para criação e manipulação de grafos, e análise de redes.

Utilizando *scripts* escritos em *Python*, as medidas topológicas das redes de canais vasculares foram adquiridas. As medidas topológicas em questão são as mesmas descritas para as análises de topologia de redes de co-expressão gênica, do Capítulo anterior (Seção 4.2.2.a) (descritas em detalhe no Capítulo 2, Seção 2.3):

1. *Estatísticas de grau*;
2. *Transitividade local* [15];
3. *Centralidade de interposição* (do inglês, “*Betweenness centrality*”) [16, 17, 14];
4. *Tamanho do componente gigante* [3];
5. *Número de arestas do componente gigante* [3];
6. *Coeficiente aglomeração* (do inglês “*Clustering coefficient*”) [18] [19];
7. *Assortatividade de grau* [20];
8. *Comprimento médio do caminho mais curtos* (do inglês “*Average shortest path length*” [21, 22];
9. *Diâmetro* [23];
10. *Coeficiente “Rich club”* [24];
11. “*Matching index*” [25].

Todas essas medidas para todos ossos foram submetidas à Análise de Componentes Principais (do inglês *Principal Component Analysis* - PCA)

5.2.7 Medidas Geométricas

A estrutura formada pelos canais vasculares dos ossos é tanto topológica quanto geométrica, portanto, pode-se também extrair medidas desse tipo. A informação do diâmetro da estrutura original pode ser adicionada ao esqueleto da imagem (mostrado na Figura 5.10). Podemos associar cada ponto do esqueleto ao diâmetro da estrutura original, e para isso, consideraremos que essa medida em um ponto é aproximado pelo diâmetro do maior círculo inscrito na seção transversal naquele dado ponto (Figura 5.15), o que é obtido pela transformada de distância do objeto original [138]. A transformada de distância (nesse caso, euclidiana) basicamente atribui a cada pixel a distância entre ele e o pixel da borda mais próxima.

A Figura 5.16 mostra um esquema de uma parte da rede de canais vasculares, e a representação de uma série de medidas que podem ser obtidas a partir dessa estrutura. As mádias d_{max} e d_{min} são o diâmetro máximo e mínimo dos canais, ou seja, das arestas da rede. A medida de d_t corresponde à *distância topológica*,

Figura 5.15: Representação esquemática de uma porção da rede de canais vasculares e o respectivo esqueleto no interior dela. O círculo azul representa uma área de seção transversal do canal vascular. Os traços representam arestas e os círculos representam nós. Obs.: nesse esquema, os nós estão representados por círculos para caráter didático; no esqueleto original os nós são bifurcações das arestas. Fonte: elaborada pela autora.

caracterizando um par de nós como ligados ou não (caráter binário). A medida de s é o comprimento de arco, que contabiliza a distância do caminho entre um par de nós, contrastando com d_e que é a distância entre as extremidades que dois nós delimitam.

Com essas medidas, poderemos estudar os canais vasculares de ossos como uma figura geométrica (distribuição de pontos), além de considerar a distribuição geométrica ao longo do comprimento do osso.

Figura 5.16: Representação esquemática da rede de canais vasculares e o respectivo esqueleto no interior dela. As medidas d_{max} e d_{min} correspondem ao diâmetro máximo e mínimo de arestas; d_t corresponde à distância topológica; d_e é a distância entre extremidades; e s é o comprimento de arco. Fonte: elaborada pela autora.

5.2.8 Medidas de Dimensionalidade

A dimensão é um conceito matemático que, a princípio, quantifica o número de variáveis independentes e necessárias para descrever uma estrutura e/ou sua dinâmica. Em sistemas complexos, a dimensão também pode ser entendida como os graus de liberdade que um sistema em particular pode atingir [139].

Diversas estruturas abstratas da matemática apresentam dimensão bem definida. Uma curva geralmente apresenta dimensão 1 (ou é unidimensional) pois apenas uma variável é necessária para descrever completamente uma posição de um ponto nesta estrutura. Planos e superfícies são estruturas bidimensionais, já que dois descritores são necessários para identificar uma posição. A mesma ideia vale para sólidos (3-dimensionais) e estruturas de dimensões maiores. Há ainda estruturas matemáticas que apresentam dimensão fractal, nas quais a dimensão pode atingir valores fracionários. Nestes casos, uma estrutura a priori de uma certa dimensão pode ganhar características de dimensões menores ou maiores.

Apesar de bem estabelecida matematicamente em estruturas contínuas ou geradas a partir de auto-similaridade, na prática, a dimensão nem sempre é bem definida para estruturas do mundo real. Redes complexas, assim como qualquer outra representação discreta, formalmente, apresentam dimensão nula, o que impossibilita o cálculo exato da dimensão do sistema original através de técnicas puramente analíticas. Entretanto, algumas técnicas foram desenvolvidas para estimar a dimensão destes tipos de estruturas. Espera-se, portanto que estruturas regulares como grades, apresentem dimensão compatível com seu espaço original. Isto no entanto, não pode ser estendido diretamente a estruturas mais complexas, que inclusive podem apresentar regiões de dimensões diferentes dentro da sua própria estrutura, ou ainda, podem depender da escala em que estão sendo observadas. Esta última característica pode ser exemplificada ao considerarmos uma estrutura tubular muito longa e com espessura pequena. Quando vista de longe, observa-se apenas um comportamento de uma linha, portanto, unidimensional. Ao aproximar-se do objeto, no entanto, é possível identificar uma estrutura mais similar àquela de uma superfície, portanto, bidimensional. Por fim, uma observação em uma escala ainda menor, mostra que o objeto em questão se trata de um sólido 3-dimensional, sendo então necessárias ao menos três. Estes diferentes comportamentos observáveis em algumas estruturas reais requerem que a dimensão seja definida dentro de um arcabouço multiescala [140].

Nos trabalhos de Filipi Nascimento Silva [141, 142] são introduzidos os conceitos de dimensão local e dimensão multiescala em redes, que permitem a caracterização tanto de redes quanto de vértices em termos de padrões de dimensão vinculados a um conjunto de escalas de interesse. O método utiliza o cálculo da estimação da dimensão de capacidade [143], que pode ser fractal, aplicada, independentemente, a cada vértice da rede. A característica multiescala é conseguida ao limitar a estimação da dimensão a uma certa extensão de progressivas vizinhanças centradas no vértice de referência. Como resultado, é atribuído um conjunto de valores de dimensão definidos ao longo dos níveis concêntricos tomados considerando cada um dos vértices da rede como referência.

Em sistemas biológicos, a dimensão pode ter uma participação importantíssima em diversos fenômenos. Por exemplo, sistemas de maior dimensão são mais robustos a ataques aleatórios dado que os caminhos entre os elementos são mais redundantes. Isso pode ser ilustrado ao considerar um sistema de canais

como o Sistema Haversiano. Uma região bidimensional de vasos é mais sujeita à falha do que uma região tridimensional no caso de um ataque aleatório, como no caso de entupimento de vasos, mesmo que ambas as regiões apresentem mesma quantidade de elementos e conectividade média [143].

5.3 Resultados e Discussão

Nesse estudo, é reportada uma análise extensiva da quantificação das propriedades topológicas e geométricas do sistema de Havers em diferentes ossos longos de camundongos - fêmur, tíbia e úmero. As Tabelas 5.2 e 5.3 reúnem todos os valores dessas medidas. As próximas seções serão destinadas ao relato desses resultados e sua discussão matemática e biológica.

5.3.1 Análise Topológica

As medidas topológicas foram analisadas *individualmente* (Seção 5.3.1a) e de modo *global* (Seção 5.3.1b).

5.3.1.a Análise Individual das Medidas

As medidas de *grau dos nós*, *transitividade* e *centralidade de interposição* são feitas para cada nó, e portanto, podem ser dispostas em gráfico na forma de histograma. A Figura 5.17 mostra o histograma da distribuição dos *graus* das redes vasculares de fêmures e tíbias (Figura 5.17(a)) e úmeros (Figura 5.17(b)), note que os gráficos não são representados na forma de barras para evitar que a figura ficasse demasiadamente densa.

Nessas figuras, nota-se um conjunto de pontos que fogem à distribuição em lei de potência, o conjunto de nós com grau 1 (em $x = 10^0$), também conhecidos como *nós folha*. Esse grupo de nós corresponde às bordas da rede, como ilustrado na Figura 5.18. O valor de grau mais frequente dos nós corresponde ao $k = 3$, ou seja, um vértice de bifurcação dicotômica, tal qual uma árvore.

A Figura 5.17 mostra uma propriedade interessante das redes de canais vasculares, a distribuição de graus dos nós segue uma estrutura livre de escala, ou seja, segue aproximadamente uma lei de potência [30, 29]. As redes vasculares de ossos longos, além de serem redes fundamentalmente biológicas, são redes geográficas, e essas redes tem grande tendência a apresentar *hubs*. Os grafos geográficos refletem um caso de redes reais: ligações entre nós distantes geralmente são mais custosas, ou difíceis de construir, o que diminui as chances de eles existirem nesses grafos [144].

Um modelo que explica essa tendência é o de grafos geométricos aleatórios (do inglês, *Random geometric graphs*) [145]. Suponha que existam n pontos em um plano, dispostos ao acaso. Um grafo é formado unindo-se os nós i e j se estes estiverem a uma distância menor que r um do outro [145]. Outro modelo geométrico é o grafo aleatório de Waxman (*Waxman random graph*). Esse grafo inicialmente também é formado por n pontos dispostos de modo uniformemente aleatório. A ligação dos nós i e j é guiada pela

(a) Distribuição de graus das redes vasculares dos fêmures e tíbias.

(b) Distribuição de graus das redes vasculares dos úmeros.

Figura 5.18: Ilustração da ponta de um grafo hipotético com a discriminação do grau do nó. Fonte: Elaborado pela autora.

probabilidade dada por uma função d_{ij} entre eles [146]. Ambos grafos resultantes apresentarão uma estrutura planar, apesar de primordialmente aleatório. Além disso, a imposição da ligação dentro da distância r aumentará a probabilidade de grupos de nós altamente conectados dentro da rede, podendo até formar comunidades bem distintas

Propriedades interessantes na distribuição de grau dos nós já foram relatadas em outros trabalhos. Redes elétricas (*power grid*) são redes de linhas de transmissão de alta voltagem que provêm o transporte de eletricidade dentro e entre países [3]. Na *Survey* sobre redes elétricas de Pagani e colaboradores [147], os autores afirmam que geralmente, sua distribuição de grau dos nós tende a seguir uma distribuição exponencial. Nessas redes, uma medida caracterizadora de suas propriedades é o grau médio. Xu e Harriss [148] estudaram a estrutura da rede aérea de transporte de passageiros dos Estados Unidos da América, e relataram uma distribuição de grau seguindo uma lei de potência. O trabalho de Kollmannsberger [132] citado anteriormente neste Capítulo também relatou a presença de distribuições de grau de uma rede biológica com caudas seguindo uma lei de potência. Esses autores abordaram a rede de canalículos de osteócitos embebidos em matriz extracelular mineralizada. Nessas redes, as lacunas onde coabitam os osteócitos correspondem à maioria dos *hubs*.

Observando as Tabelas 5.2 e 5.3, temos que o grau médio dos nós das redes vasculares dos ossos das patas traseiras é de 3,49 em média, e o dos úmeros, 3,12. Portanto, de um modo geral, temos que o osso com maior grau médio de nós é a *tíbia* (3,73), seguido do *fêmur* (3,24) e *úmero* (3,12). O menor grau médio para *fêmur* foi registrado para o osso $3A_d$, com o valor de 2,81, e maior grau médio foi encontrado no osso $6A_b$, com o valor de 3,66. O menor grau médio para *tíbia* é o do osso $7A_e$, com o valor de 3,06, e maior grau médio foi registrado para o osso $5A_b$, com o valor de 4,19. Finalmente, o menor grau médio para *úmero* foi encontrado para o osso $2B_d$, com o valor de 2,58, e maior grau médio é o do osso $8B_c$, com o valor de 3,76.

A Figura 5.19 mostra a distribuição da *transitividade* das redes vasculares de fêmures (Figura 5.19(a)) e

úmeros (Figura 5.19(b)), ou seja, a probabilidade de dois vizinhos de um mesmo vértice serem conectados também. Note que os gráficos não são representados na forma de barras para evitar que a figura ficasse demasiadamente densa.

As figuras mencionadas mostram que as redes vasculares dos ossos de um modo geral apresentam uma assinatura similar para essa medida. A alta frequência de nós com alta transitividade (mais próximas de 10^0) indica que os nós dessas redes são altamente interconectados, menos organizados como uma “árvore”. A presença de nós com alto grau, assim como a distribuição em lei de potência do grau dos nós, corrobora a alta transitividade desses nós.

A investigação da *centralidade de interposição* das redes vasculares mostra que o histograma da centralidade de interposição também aparenta ter uma assinatura similar para as redes (Figura 5.20).

Com essas análises, temos que a topologia das redes sugere uma otimização da vascularização devido à alta transitividade das redes. Além disso, a organização interconectada das redes é uma vantagem quando se considera que um tubo por onde passam fluidos pode eventualmente ficar bloqueado (entupido). Uma rede altamente conectada tem *loops* e desvios que podem permitir a nutrição bem sucedida de uma área, mesmo que alguns canais estejam inviabilizados. Todas as redes investigadas parecem exibir uma organização topológica local similar, apesar de diferenças pronunciadas no tamanho total das mesmas.

(a) Distribuição de transitividade das redes vasculares dos fêmures e tíbias.

(b) Distribuição de transitividade das redes vasculares dos úmeros.

(a) Distribuição da centralidade de interposição das redes vasculares dos fêmures e tíbias.

(b) Distribuição de transitividade das redes vasculares dos úmeros.

Tabela 5.2: Reunião das medidas topológicas dos fêmures e tíbias.

	1A			2A			3A				5A				6A		7A				8A			
	b	c	d	b	c	d	b	c	d	e	b	c	d	e	b	c	b	c	d	e	b	c	d	e
V max grau	138	125	123	52	98	163	75	61	45	247	590	248	264	445	225	226	171	133	214	100	1738	92	473	2944
grau médio	3.52	3.28	3.10	3.24	3.32	3.14	3.24	3.15	2.81	2.92	4.19	4.17	4.15	4.14	3.66	3.51	3.66	3.45	3.39	3.06	3.92	3.37	3.63	3.66
desvpad grau	3.07	2.74	2.75	2.54	2.70	2.79	2.36	2.20	1.86	2.41	5.73	4.76	4.52	4.88	3.69	3.26	3.34	2.83	2.88	2.37	7.98	2.66	3.50	12.91
entropia grau	0.46	0.41	0.38	1.24	0.63	0.27	0.92	1.16	1.30	0.08	0.08	0.28	0.27	0.11	0.24	0.20	0.38	0.46	0.18	0.54	0.00	0.64	0.05	0.00
transit média	0.11	0.10	0.09	0.10	0.10	0.09	0.09	0.08	0.07	0.08	0.17	0.16	0.16	0.16	0.13	0.11	0.12	0.10	0.10	0.08	0.16	0.10	0.12	0.15
mediana transit	0	0	0	0	0	0	0	0	0	0	0.07	0.06	0.06	0.06	0	0	0	0	0	0	0.03	0	0	0
desvpad transit	0.18	0.17	0.17	0.17	0.17	0.17	0.15	0.15	0.13	0.15	0.24	0.22	0.21	0.22	0.20	0.18	0.19	0.17	0.17	0.15	0.23	0.17	0.18	0.23
entropia transit	1.51	1.35	1.28	1.36	1.41	1.29	1.29	1.24	1.01	1.12	1.88	1.85	1.83	1.84	1.61	1.50	1.58	1.44	1.42	1.20	1.78	1.40	1.57	1.68
V max centr (*10 ⁸)	5.54	5.66	7.70	2.17	1.81	2.92	5.31	6.03	7.91	9.81	2.14	1.64	19.5	6.83	7.15	4.78	7.55	8.44	6.62	3.63	11.0	3.40	4.20	30.8
centr média (*10 ⁶)	1.91	2.31	3.48	1.56	1.09	1.71	3.72	2.92	3.85	4.38	2.00	2.53	2.49	1.20	1.64	2.05	1.78	2.15	2.85	3.00	8.40	1.43	1.76	1.71
mediana centr (*10 ⁵)	8.52	9.29	1.08	6.63	5.14	6.20	1.55	1.40	1.64	1.13	4.36	5.91	6.46	3.55	6.07	9.86	8.60	1.09	1.31	1.17	2.87	8.48	7.94	3.78
desvpad centr (*10 ⁷)	1.32	1.31	1.97	6.85	5.34	80.8	1.80	1.33	1.98	2.69	2.41	2.50	2.54	1.23	1.32	1.28	1.36	1.33	1.50	1.38	1.10	77.8	1.02	2.79
entropia centr	0.14	0.17	0.19	0.28	0.24	0.24	0.27	0.21	0.19	0.17	0.04	0.07	0.06	0.07	0.10	0.17	0.09	0.10	0.18	0.30	0.03	0.16	0.17	0.02
núm vértices (*10 ³)	86	90	108	61	51	61	118	95	98	102	160	175	166	94	91	99	105	105	120	97	82	76	95	125
núm arestas (*10 ³)	152	147	168	99	84	97	192	150	138	150	335	366	345	196	168	173	192	182	204	148	161	128	172	230
<i>vertices / arestas</i>	0.57	0.61	0.64	0.62	0.60	0.64	0.62	0.63	0.71	0.68	0.48	0.48	0.48	0.48	0.55	0.57	0.55	0.58	0.59	0.65	0.51	0.59	0.55	0.55
transit não direc	0.09	0.09	0.08	0.09	0.09	0.08	0.09	0.09	0.08	0.08	0.07	0.08	0.08	0.08	0.09	0.08	0.09	0.09	0.08	0.09	0.03	0.09	0.08	0.01
transit méd local	0.15	0.13	0.14	0.14	0.14	0.13	0.12	0.12	0.10	0.12	0.21	0.20	0.19	0.19	0.17	0.15	0.16	0.14	0.14	0.12	0.20	0.14	0.16	0.21
assortatividade (*10 ⁻²)	-3.80	-3.21	-5.04	-8.44	-6.25	-2.09	-6.88	-5.73	-4.39	-0.86	-4.57	-6.89	-6.12	-4.10	-4.07	-2.72	-3.93	-3.95	-2.88	-4.12	-1.40	-5.51	-2.20	-2.21
méd men cam (*10 ¹)	4.53	5.24	6.52	5.18	4.36	5.62	6.38	6.25	7.90	8.65	2.59	2.99	3.10	2.62	3.67	4.24	3.48	4.16	4.82	6.28	2.14	3.85	3.81	2.82
diâmetro	106	127	168	120	97	134	140	139	188	255	61	78	84	71	94	111	101	105	116	161	57	93	88	72
coef <i>Rich club</i> (*10 ⁻⁵)	4.08	3.65	2.87	5.27	6.52	5.08	2.74	3.31	2.85	2.86	2.62	2.38	2.50	4.37	3.99	3.54	3.48	3.27	2.81	3.16	4.78	4.42	3.81	2.91
índice <i>matching</i> (*10 ⁻²)	4.76	4.45	4.40	4.55	4.67	4.39	4.25	4.20	3.72	3.99	5.28	5.19	5.18	5.19	4.97	4.72	4.82	4.53	4.53	4.09	5.20	4.56	4.84	5.08

Tabela 5.3: Reunião das medidas topológicas dos úmeros.

	1B			2B		3B			4B		5B		6B			7B			8B		
	b	c	d	b	c	b	c	d	b	c	b	c	b	c	d	b	c	d	b	c	d
V max grau	109	47	30	135	81	5041	1555	20	59	499	47	37	61	47	788	37	123	37	206	1315	217
grau médio	3.34	2.92	2.59	3.53	2.90	3.19	2.98	2.71	3.29	3.36	3.15	3.08	3.31	3.11	3.08	2.73	2.72	2.64	3.70	3.77	3.52
desvio padrão grau	2.51	1.96	1.68	2.93	2.00	17.46	5.40	1.68	2.35	3.33	2.19	2.10	2.41	2.19	3.22	1.76	1.94	1.78	3.31	5.65	3.48
entropia grau	0.59	1.15	1.38	0.49	0.74	0.00	0.00	1.43	1.19	0.03	1.22	1.54	1.22	1.21	0.01	1.39	0.23	1.36	0.29	0.01	0.21
transit média	0.09	0.07	0.05	0.11	0.07	0.09	0.07	0.06	0.09	0.10	0.08	0.08	0.09	0.08	0.08	0.06	0.06	0.06	0.12	0.14	0.11
desvio padrão transit	0.16	0.14	0.13	0.18	0.14	0.17	0.14	0.13	0.16	0.17	0.15	0.15	0.16	0.15	0.16	0.13	0.13	0.13	0.19	0.21	0.19
entropia transit	1.37	1.09	0.85	1.50	1.08	1.29	1.11	0.91	1.34	1.42	1.23	1.19	1.34	1.21	1.21	0.96	0.94	0.87	1.59	1.67	1.52
V max centr (*10 ⁸)	3.39	2.43	1.60	7.99	1.79	16.4	17.4	8.94	1.19	3.03	3.43	2.14	3.27	5.48	13.0	3.33	6.11	2.15	9.28	9.64	6.46
centr média	1.73	2.35	1.67	2.66	1.93	1.89	2.79	9.88	1.14	1.58	2.26	2.14	2.14	2.61	5.13	2.45	3.55	2.37	1.88	2.08	1.68
mediana centr (*10 ⁵)	96.3	1.12	5.85	1.18	86.9	95.3	1.51	47.6	71.8	71.5	1.24	1.00	1.05	1.35	1.82	1.21	1.23	58.2	91.5	69.3	69.2
desvio padrão centr (*10 ⁷)	89.0	1.04	70.2	1.61	81.4	1.09	1.30	48.0	46.4	76.2	1.04	91.7	1.04	1.27	2.68	1.02	2.04	1.18	1.45	1.69	1.23
entropia centr	0.21	0.35	0.37	0.14	0.37	0.04	0.06	0.33	0.35	0.22	0.26	0.35	0.25	0.19	0.17	0.29	0.21	0.34	0.08	0.09	0.10
núm vértices (*10 ³)	78	72	43	119	57	84	94	27	52	70	85	69	85	93	152	72	82	48	107	117	79
núm arestas (*10 ³)	131	105	55	211	84	134	141	37	86	118	134	107	140	144	234	98	112	64	199	220	140
<i>vertices/arestas</i>	0.60	0.69	0.77	0.57	0.69	0.63	0.67	0.74	0.61	0.60	0.64	0.65	0.60	0.64	0.65	0.73	0.73	0.76	0.54	0.53	0.57
transit não direc	0.09	0.09	0.08	0.09	0.08	0.00	0.02	0.08	0.09	0.07	0.09	0.09	0.09	0.09	0.06	0.08	0.08	0.08	0.09	0.05	0.08
transit méd local	0.13	0.11	0.09	0.15	0.11	0.13	0.11	0.09	0.13	0.14	0.12	0.11	0.13	0.12	0.12	0.10	0.10	0.09	0.16	0.18	0.16
assortatividade	-0.04	-0.06	-0.13	-0.06	-0.09	-0.02	-0.01	-0.16	-0.05	-0.02	-0.06	-0.09	-0.07	-0.05	-0.01	-0.10	0.01	-0.06	-0.04	-0.01	-0.03
média menores cam	44.95	65.66	78.76	45.53	67.50	45.61	59.95	72.27	44.60	45.78	53.58	62.35	51.30	56.95	68.37	68.91	86.77	98.21	35.85	36.56	42.98
diâmetro	108	191	185	120	158	116	146	161	106	106	131	160	118	140	179	208	282	344	91	89	112
coef <i>Rich club</i> *10 ⁻⁵)	4.24	4.01	6.01	2.95	5.01	3.77	3.16	9.76	6.28	4.77	3.67	4.42	3.89	3.33	2.02	3.79	3.29	5.41	3.44	3.22	4.40
índice <i>matching</i> (*10 ⁻²)	4.42	3.88	3.36	4.72	3.85	4.19	3.88	3.49	4.42	4.60	4.14	4.07	4.37	4.14	4.12	3.65	3.56	3.32	4.86	5.05	4.76

5.3.1.b Análise Global das Medidas

As vinte e oito medidas topológicas que foram calculadas para cada uma das redes vasculares se encontram nas Tabelas 5.2 e 5.3. As distribuições dessas medidas encontram-se nas Figuras 5.21 e 5.22.

O número de *vértices* varia de aproximadamente 75.000 a 180.000 nos pedaços de tíbias, 25.000 a 155.000 nos pedaços de úmeros e de aproximadamente 50.000 a 120.000 nos fêmures. Já o número de *arestas* vai de cerca de 125.000 a 375.000 nos pedaços de tíbias, cerca de 40.000 a 240.000 nos pedaços de úmeros, e de 75.000 a 200.000 em fêmures. Portanto, de um modo geral, as tíbias alcançam maiores valores de número de vértices e arestas, seguidos dos úmeros e dos fêmures.

Os valores de *transitividade indireta* na sua maioria ficaram entre o intervalo $[0.07 < t < 0.1]$, exceto para as tíbias $8A_b$ e $8A_e$, e os úmeros $3B_b$ e $3B_c$. Já os valores de *transitividade média local* foram mais variáveis, sem uma tendência marcada, assim como a *médias dos menores caminhos*, *diâmetro*, *coeficiente "Rich Club"* e *índice "matching"*.

Retomando os conceitos do Capítulo 2, temos que a medida de *assortatividade* mede se uma rede tem nós com preferência por determinadas conexões ou não. A assortatividade registrada nas redes de ossos são bem próxima de zero, tanto para os ossos das patas traseiras quanto das dianteiras, indicando que os nós tem baixa preferência de ligação em relação ao grau. Contudo, maioria dos ossos apresenta uma assortatividade negativa (entre 0 e -0.16), mostrando que esses ossos tem uma leve tendência a se conectarem com outros nós de graus diferentes (grau alto com grau baixo).

Observando os gráficos da Figura 5.21, temos alguns casos em que a tíbia 5A tem valores bem distintos quando comparados com o restante. Essa tíbia apresenta um número de *vértices* e *arestas* consideravelmente superior às demais redes, e o índice *matching* levemente mais elevado. Contudo, em relação à razão *vertices/arestas*, esse osso tem o menor valor de todos.

Esse conjunto de dados foi submetido à Análise de Componentes Principais (do inglês, *Principal Component Analysis* – PCA), e o resultado pode ser visualizado na Figura 5.23. Juntos, o eixo PCA1 e o eixo PCA2 contém 73,2% da variação total dos dados, sendo que o eixo PCA1 corresponde ao **desvio padrão da transitividade**, e o eixo PCA2 à **centralidade de interposição média**. A Tabela 5.4 mostra o ranqueamento do peso das medidas topológicas no eixo PCA1, e a Tabela 5.5 o do eixo PCA2. De acordo com a Tabela 5.4, *desvio padrão transitividade*, *transitividade média*, e *transitividade média local* tem um peso grande no PCA1 (~ 0.27). Portanto, todas essas medidas são de similar importância na variação desses conjunto de dados.

A análise de PCA mostra que a transitividade em geral é uma medida importante para explicar a variação dos dados. A riqueza dessa medida pode estar relacionada ao fato que uma rede com mais "triângulos"¹² terá mais redundância de caminhos e menos propensão à inviabilização de rotas de transporte.

¹²Estrutura formada três nós ligados entre si.

A Figura 5.24 é o mesmo PCA da Figura 5.23, mas com os pares de fêmures, úmeros e tíbias marcados, com semelhante destaque para a tíbia 5A que será discutida com mais detalhes na Seção 5.3.4

A linha reta evidenciando pareamento dos ossos na Figura 5.24 mostra uma propriedade interessante: esses pares tendem se alinhar ao longo do eixo PCA1 e PCA 2, talvez indicando que esses grupos de pares tenham grau médio de nós o máxima transitividade semelhantes entre os ossos análogos

As Figuras 5.25 e 5.26 mostram uma representação alternativa do PCA da Figura 5.23. Nesta imagem, os fêmures estão na cor verde, úmeros em amarelo, e tíbias em vermelho. A delimitação dos grupos (Figura 5.26) mostra que há uma sobreposição central das medidas topológicas das redes vasculares dos três ossos.

Tabela 5.4: Ranqueamento do peso das medidas topológicas no PCA1.

Medida	peso
desvio padrão transitividade	0.274330355
transitividade média	0.272448203
transitividade média local	0.2717138
entropia da transitividade	0.267885957
grau médio	0.265733514
vértices / arestas	0.262762517
índice matching	0.260706624
entropia centralidade	0.24516056
comprimento médio do caminho mais curtos	0.244336424
vértices	0.23512598
diâmetro	0.2128724
valor mínimo centralidade	0.206177177
entropia grau	0.20608609
arestas	0.203114506
transitividade mediana	0.182000999
desvio padrão grau	0.16437503
desvio padrão centralidade	0.138203586
coeficiente Rich Club	0.129385164
assortatividade	0.12201619
valor máximo grau	0.116178234
mediana centralidade	0.110570499
transitividade indireta	0.105363531
centralidade média	0.054418619

Tabela 5.5: Ranqueamento do peso das medidas topológicas no PCA2.

Medida	peso
centralidade média	0.442098024
desvio padrão centralidade	0.362519229
coeficiente Rich Club	0.361178061
mediana centralidade	0.343526892
arestas	0.265446661
assortatividade	0.235147799
valor mínimo centralidade	0.221787018
diâmetro	0.212119497
comprimento médio do caminho mais curtos	0.195087025
vértices	0.160851652
transitividade indireta	0.140639892
entropia grau	0.127889457
índice matching	0.124574922
entropia da transitividade	0.112534349
grau médio	0.110103818
vértices / arestas	0.109466832
entropia centralidade	0.109453511
valor máximo grau	0.107002
transitividade média	0.097890019
transitividade média local	0.088098825
desvio padrão transitividade	0.074911535
desvio padrão grau	0.06543384
mediana transitividade	0.041295794

Figura 5.21: Distribuição das medidas topológicas dos fêmures e tíbias. Em ordem, as medidas em questão nos gráficos são: número de vértices, número de arestas, razão *vertices/arestas*, transitividade indireta, transitividade média local, assortatividade, média dos menores caminhos, diâmetro, coeficiente *Rich Club*, índice *matching*. Fonte: Elaborado pela autora.

Figura 5.22: Distribuição das medidas topológicas dos úmeros. Vide legenda da Figura 5.21 para o significado das siglas. Fonte: Elaborado pela autora.

Figura 5.23: PCA das medidas topológicas das redes vasculares. Fonte: Elaborado pela autora.

Figura 5.25: PCA da Figura 5.23 com os pontos maiores e cores diferentes: Em verde, são marcados os fêmures; em amarelo, estão os úmeros; e em vermelho, as tíbias. Fonte: Elaborado pela autora.

Figura 5.26: PCA com as cores da Figura 5.25 com destaque para os grupos dos fêmures, úmeros e tíbias. Fonte: Elaborado pela autora.

5.3.2 Análises Geométricas

As Análises Geométricas aqui reportadas estão ilustradas na Figura 5.16. Essa Seção inicia-se com a análise da medida de *tamanho das arestas* em comprimento de arco (medida s da Figura 5.16). Enquanto o padrão de distribuição mostrado pelas medidas topológicas era similar entre os ossos (Seção 5.3.1), as medidas geométricas apresentam assinaturas distintas entre ossos, e essa característica merece nossa atenção.

A distribuição do tamanho das arestas das redes vasculares dos fêmures, tíbias e úmeros encontra-se na Figura 5.27. Entre as patas traseiras, a tíbia 8A foi a mais diferenciada do restante dos ossos (Figura 5.27(a)). Entre os úmeros, o 2B tem um padrão marcadamente diferenciado, seguido pelos 4B, 6B e 8B. O restante das redes tem a distribuição do tamanho de arestas seguindo aproximadamente uma lei de potência.

A Figura 5.16 ilustra como um mesmo canal pode ter raios diferentes (d_{max} e d_{min}), e a próxima medida relaciona-se com o raio médio de um mesmo canal. A distribuição do tamanho médio dos raios das arestas das redes vasculares dos fêmures, tíbias e úmeros pode ser observada na Figura 5.28. Os gráficos dessa figura mostram que de um modo geral, as arestas tem um raio médio menor. Entre as patas traseiras, a tíbia 8A teve um padrão bem diferenciado, e o fêmur 6A também (Figura 5.28(a)). Todos os úmeros seguem um padrão pouco previsível (Figura 5.28(b)).

Essa variação mais pronunciada nas geometrias das redes sugerem que a demanda biológica de transporte pode variar em relação à posição no mesmo osso, e entre ossos diferentes. As variações geométricas também podem ser consequência de particularidades biomecânicas dos ossos, ou podem estar ligadas a uma determinada função [9].

5.3.3 Análise de Dimensionalidade

A análise de dimensionalidade multiescala representa uma novidade na análise de redes sistemas Harvesianos. A Figura 5.29(a) mostra a dimensão dos nós das redes vasculares. Tratando-se de uma análise multiescala, temos que os primeiros nós reportados nesses gráficos correspondem a regiões mais centrais da rede. Conforme o *eixo x* dos gráficos aumenta, temos a adição dos nós da vizinhança desses nós centrais, até que se atinja as bordas da rede em questão. De acordo com a dimensão inicial dos nós ($\sim [0 < x < 5]$), as curvas foram agrupadas em três grupos diferentes, apresentados nos próximos parágrafos.

O primeiro grupo corresponde às redes de dimensão inicial entre 1D e 2D (Figura 5.29(b)), entre eles encontram-se redes vasculares de fêmures ($N = 2 - 3A_d, 3A_e$) e úmeros ($N = 7 - 1B_c, 1B_d, 2B_c, 3B_d, 7B_b, 7B_c, 7B_d$), as tíbias não estão representadas aqui. O gráfico mostra que a dimensão dos nós vai caindo lentamente, todavia se mantendo entre 1D e 2D.

O segundo grupo corresponde às redes de dimensão inicial entre 2D e 2D (Figura 5.29(c)). Este é o maior grupo, com 10 redes vindas de fêmures ($1A_b, 1A_c, 1A_d, 2A_b, 2A_c, 2A_d, 3A_b, 3A_c, 6A_b, 6A_c$), 15 de

(a) Distribuição do tamanho das arestas das redes vasculares dos fêmures e tíbias.

(b) Distribuição do tamanho das arestas das redes vasculares dos úmeros.

(a) Distribuição do tamanho médio dos raios das arestas das redes vasculares dos fêmures e tíbias.

(b) Distribuição do tamanho médio dos raios das arestas das redes vasculares dos úmeros.

úmeros ($1B_b, 2B_b, 2B_d, 3B_b, 3B_c, 4B_b, 4B_c, 5B_b, 5B_c, 6B_b, 6B_c, 6B_d, 8B_b, 8B_c, 8B_d$) e 6 tíbias ($7A_b, 7A_c, 7A_d, 7A_e, 8A_c, 8A_d$). Até ($\sim [x < 40]$) o valor de dimensão está consideravelmente estável, mas a partir desse ponto ela cai acentuadamente para parte das redes.

O terceiro, e menor, grupo corresponde às redes de dimensão inicial maior que 3D (Figura 5.29(d)). Essas redes provêm de ossos de estrutura peculiar, as tíbias 8A ($N = 2 - 8A_b, 8A_e$) e 5A ($N = 4 - 5A_b, 5A_c, 5A_d, 5A_e$). A tibia 8A visualmente não apresenta nenhuma alteração, mas no PCA da Figura 5.23 as partes aqui mencionadas ficaram proximamente posicionadas à tibia 5A (marcado no pca da Figura 5.24). A tibia 5A por sua vez é um osso bem peculiar, e será discutido melhor na próxima Seção (Seção 5.3.4)

Apesar de este grupo iniciar o gráfico com dimensões notavelmente mais altas, elas decrescem mais rapidamente também.

Figura 5.29: Gráfico da distribuição da dimensionalidade das redes sistema Harvesiano. Cada linha corresponde a uma das partes dos ossos (destacadas na Figura 5.7). O eixo y contém as dimensões que caracterizam cada conjunto de nós (eixo x). Fonte: Elaborado pela autora.

O grupo com nós inicialmente entre uma e duas dimensões (1D e 2D) pode ser ainda subdividido entre dois outros grupos (Figura 5.30) de acordo com a "velocidade" do decaimento das dimensões em função dos nós. Segundo essa divisão, as redes da Figura 5.29(c) são separadas em: 4 fêmures ($1A_b, 2A_c, 6A_b, 6A_c$) 7 úmeros ($1B_b, 2B_b, 4B_b, 4B_c, 8B_b, 8B_c, 8B_d$) e 5 tíbias ($7A_b, 7A_c, 7A_d, 8A_c, 8A_d$) e 6 fêmures ($1A_c, 1A_d, 2A_b, 2A_d, 3A_b, 3A_c$) 8 úmeros ($2B_d, 3B_b, 3B_c, 5B_b, 5B_c, 6B_b, 6B_c, 6B_d$) e 1 tibia ($7A_e$).

(a) Distribuição dos nós inicialmente entre duas e três dimensões (2D e 3D).

(b) Grupo 1 proveniente da divisão dos nós que tem inicialmente duas e três dimensões (2D e 3D).

(c) Grupo 2 proveniente da divisão dos nós que tem inicialmente duas e três dimensões (2D e 3D).

5.3.4 Osso 5A dupla medula

Durante a etapa de aquisição das imagens tomográficas dos ossos de camundongo, a projeção lateral das imagens tomográficas da Tíbia 5A mostrou uma estrutura não usual: esse osso apresenta duas medulas (Figura 5.31). A Figura 5.32 mostra a visualização da rede vascular desse osso no software *Network 3D*®, e esses detalhes revelam que essas duas porções separadas do osso são cilíndricas, semelhante a dois ossos emparelhados. Entre esses dois cilindros principais, há diversos canais que interligam essas regiões. Essa estruturação não usual pode ser o motivo pelo qual essas redes apresentam medidas topológicas tão distintas quando comparadas com os demais ossos (Seção 5.3.1, Figura 5.24), e tem mais dimensões (Seção 5.3.3).

Figura 5.31: Projeção lateral da tomografia do osso Tíbia 5A. Fonte: Elaborado pela autora.

(a) Visão frontal do osso 5A.

(b) Visão lateral do osso 5A.

(c) Visão dorsal do osso 5A.

Figura 5.32: Visualização da parte “b” do osso 5A. Fonte: Elaborado pela autora.

5.4 Conclusões

As análises aqui reportadas mostram que os ossos considerados são praticamente indistinguíveis em relação às propriedades topológicas das redes enquanto nota-se uma variação mais pronunciada em relação à sua geometria.

O primeiro efeito pode estar relacionado ao fato de que a estrutura topológica destes ossos é, em grande parte, definida para suprir demandas básicas (por exemplo, nutrição, robustez à eventuais obstruções etc) comuns a todos os tipos considerados de ossos. Dentre as pequenas variações topológicas observadas, destaca-se o número de nós progressivamente menor de tíbia para o úmero, e então do úmero para o fêmur.

Já a variabilidade geométrica pode ser consequência das demandas biomecânicas mais específicas de cada um dos tipos de osso, situados em posições diferentes dos indivíduos, e portanto sujeitos a esforços e cargas diferentes. No que se refere às variações geométricas observadas, ocorrem principalmente de forma não sistemática, a nível individual e, às vezes entre as partes de um mesmo osso.

No desenvolvimento embrionário de um vertebrado, as porções centrais de um osso longo correspondem ao centro de ossificação primário, portanto, inicial no desenvolvimento quando comparado às epífises (Figura 5.1). Os experimentos relatados no presente Capítulo não mostraram diferenciação topológica entre as distintas porções de um mesmo osso, e uma variação geométrica pouco pronunciada, sugerindo assim que no animal adulto não há diferenças morfológicas resultantes desse desenvolvimento embrionário heterogêneo.

Capítulo 6

Integração das Abordagens

Anteriormente nesta Tese, o Capítulo 4 descreveu cinco métodos de detecção de limiar de redes de co-expressão gênica, segundo diferentes critérios. Por sua vez, o Capítulo 5 tratou da caracterização da rede de canais de Havers e Volkmann (o sistema Harvesiano, ou de Havers) segundo a Teoria das Redes Complexas. Como as redes do sistema Harvesiano apresenta valores relacionados à espessura dos canais, essas informações podem ser consideradas como "pesos", e portanto, serem submetidas a um processo de limiarização. Utilizamos assim métodos de limiarização propostos no Capítulo 4 nas redes do sistema de Havers do Capítulo 5

Este capítulo portanto destina-se à análise unificadora entre os Capítulos 4 e 5. Através da aplicação dos métodos de limiarização propostos no Capítulo 4 nas redes do sistema de Havers do Capítulo 5, e comparação do comportamento das redes do sistema Harvesiano após a limiarização com o comportamento das redes de expressão gênica analisadas no Capítulo 4.

6.1 Material e Métodos

No Capítulo 5, Seção 5.3.2 intitulada **Análises Geométricas**, pode-se encontrar o gráfico da distribuição do tamanho médio dos raios das arestas das redes vasculares dos fêmures, tíbias e úmeros (Figura 5.28). Essa informação foi transferida a uma matriz análoga a uma matriz de pesos, e posteriormente submetida aos critérios de limiarização do Capítulo 4, Seção 4.2.2. O raio médio das arestas é calculado numa escala em *pixels*. Por exemplo, $x = 10$ representa um raio médio de 10 *pixels*.

O critério de Máxima variação das propriedades topológicas (Seção 4.2.2.a) não foi incluído nessas análises, porque, como relatado na Seção 4.3.2, esse critério foi o que mais divergiu em relação aos outros.

Para aplicação dos critérios de limiarização, foram utilizados as redes provenientes das porções $1A_c$, $2A_b$, $3A_c$, $1B_c$, $5B_c$ e $6B_c$ dos ossos.

6.2 Resultados e Discussão

A Figura 6.1 mostra os valores médios de entropia para transitividade e centralidade de interposição das redes geradas por diferentes valores de limiar. Essa análise mostra de forma unânime para os conjuntos de dados verificados, que a entropia máxima das redes é obtida quando o limiar mais baixo é aplicado no conjunto de dados.

Figura 6.1: Valores de entropia obtidos para redes geradas por valores de limiar diferentes. Os pontos vermelhos indicam a entropia máxima alcançada por cada conjunto de dados. Fonte: Elaborado pela autora.

Os diferentes valores de limiar geram redes de diferentes tamanhos de componente gigante, e esse resultado é mostrado na Figura 6.2. Os tamanhos aqui reportados são normalizados pelo número total de genes de cada experimento. Os pontos em vermelho, verde e roxo indicam redes com componentes gigantes que compõem aproximadamente, respectivamente, 40%, 60% e 80% do número total de nós.

Esse método de limiarização de redes é análogo ao processo de ataque às arestas de menor raio.

A Figura 6.3 mostra o grau médio das redes em função do valor de limiar. Os pontos em vermelho indicam os limiares geradores de grau médio com valor de aproximadamente 2,5.

Figura 6.4 mostra a estatística Kolmogorov-Smirnov (K-S) comparando a distribuição de grau da rede a uma função em lei de potência, quanto mais baixo o valor da estatística K-S, mais próximo a uma lei de potência está a distribuição de graus. Assim como para o método de limiarização de Máxima Entropia (Figura 6.1), a menor estatística K-S das redes é obtida quando o limiares baixos são aplicados no conjunto de dados.

Figura 6.2: Tamanho do componente gigante da rede em função do valor de limiar. O tamanho é normalizado pelo número total de genes estudados em cada experimento. Os pontos coloridos indicam os tamanhos de componentes gigantes usados na análise, representados como f_S . S representa o tamanho do componente gigante. Fonte: Elaborado pela autora.

Figura 6.3: Grau médio das redes obtidas por diferentes valores de limiar. Os pontos vermelhos indicam o grau médio escolhido para analisar este critério. Fonte: Elaborado pela autora.

Figura 6.4: Estatística Kolmogorov-Smirnov comparando a distribuição de grau da rede a uma função em lei de potência. Fonte: Elaborado pela autora.

6.2.1 Comparação entre métodos

A Tabela 6.1 sumariza os limiares elegidos às porções escolhidas por cada método. Diferentemente dos limiares determinados para as redes de co-expressão gênica, houve uma relativa concordância de valores de limiar, para cada critério considerado, entre as redes vasculares de ossos. A variação mais alta ocorreu para o método de grau médio específico (2.5), que elegeu o limiar de $t = 6$ para a rede 1A_c. A concordância dos limiares obtidos, segundo os vários critérios, implica que as redes limiarizadas atendem aos diversos requisitos respectivos.

Considerando-se a variação dos limiares obtidos entre métodos diferentes, observa-se boa concordância entre os valores escolhidos pelo critério de *máxima entropia* e *distribuição em lei de potência*, indicando que a

Tabela 6.1: Limiares escolhidos pelos diferentes critérios para cada experimento. *T variation* representa o critério de Máxima variação das medidas topológicas; *T entropy* é o critério de máxima entropia; *T comp.* 0.4 representa o critério de tamanho do componente gigante fixado em 40% do número total de nós, com um significado análogo para *T comp.* 0.6 e *T comp.* 0.8; *T deg.* representa o grau médio específico (2.5); *T power-law* representa o critério de distribuição de grau em lei de potência (*power-law*)

	T entropy	T comp. 0.4	T comp. 0.6	T comp. 0.8	T deg. 2.5	T power-law
1A_c	0.8	5.5	4.0	2.5	6	1.0
2A_b	0.8	5.0	4.5	3	1.5	1.0
3A_c	0.8	4.5	3.5	2.5	1.5	1.0
1B_c	0.8	3.5	3.0	2.5	1.5	1.0
5B_c	0.8	4.5	3.5	2.5	1.5	1.0
6B_c	0.8	4.5	3.5	2.5	1.5	1.0

rede limiarizada atenderá ambos os critérios. O baixo valor de limiar obtido para esses dois casos sugere que a rede vascular ideal a ser analisada é aquela contendo todos os canais, ou seja, sem sofrer o processo de limiarização.

Capítulo 7

Conclusões

O estudo da limiarização de redes de co-expressão gênica, e a comparação entre redes complexas derivadas de sistemas biológicos diversos foram os temas principais desta Tese de Doutorado. O principal objetivo da primeira parte da Tese foi o desenvolvimento de uma abordagem experimental e sistemática ao problema da seleção de limiar em redes de co-expressão. A abordagem aqui relatada teve a intenção de contribuir para a obtenção de melhores representações de cenários biológicos complexos e específicos. Aplicamos os princípios da Teoria das Rede Complexas no estudo de dados transcriptômicos, a fim de compreender melhor os efeitos da seleção de limiar em redes de co-expressão gênica, incentivando o desenvolvimento de novos critérios e métodos para comparação e validação. As principais contribuições dessa parte incluíram a identificação, dentre um conjunto de várias bases de dados, das tendências de variações dos liminares escolhidos e de como as propriedades topológicas das redes obtidas se diversificaram em consequência da limiarização.

Na segunda parte da Tese usamos a Teoria das Redes Complexas para caracterizar o sistema formado pelos canais de Havers e Volkmann em ossos longos de camundongo. As análises aqui reportadas mostram que ossos são indistinguíveis em relação às propriedades topológicas das redes, mas em relação à sua geometria, nota-se uma variação mais pronunciada. Isso sugere que a arquitetura topológica do sistema vascular pode ser independente do tipo ósseo, mas que a demanda de transporte pode variar em relação à posição no mesmo osso, e entre ossos diferentes. As principais contribuições desta parte da Tese correspondem à caracterização da rede vascular de fêmures, úmeros e tíbias de camundongos, e a comparação entre as redes correspondentes aos respectivos pares de ossos. Além disso, nesta Tese, introduzimos a análise de dimensionalidade multiescala no estudo do sistema de Havers.

Por fim, os métodos de limiarização aqui propostos foram usados como forma de validação dos resultados da caracterização das redes do sistema de Havers. Os resultados obtidos nos levam a sugerir que as redes vasculares sejam analisadas sem sofrer o processo de limiarização.

7.1 Trabalhos Futuros

Dentre os avanços mais imediatos que possam decorrer dos resultados aqui descritos para as análises de limiarização de redes de co-expressão gênica, citamos a adição de outras bases de dados de expressão gênica global nas análises, além da incorporação de medidas topológicas adicionais na análise das redes resultantes, assim como a criação de novos critérios de limiarização com base em propriedades biológicas de interesse.

Um trabalho futuro imediato no estudo da caracterização das redes vasculares de ossos, consiste na análise de robustez das redes vasculares através de ataques a nós e arestas. Também seria interessante considerar a adição de redes vasculares provenientes de outros tipos de ossos, outros vertebrados, animais de sexo e idade diferentes, e animais acometidos por doença óssea. Além disso, a investigação da base molecular das variações encontradas nas redes vasculares de ossos seria um desdobramento muito interessante deste estudo.

Bibliografia

- [1] L Kierszenbaum, A and Laura T. *Histologia e biologia celular*. Elsevier Brasil, 3rd edition edition, 2012.
- [2] Luciano da Fontoura Costa and Roberto Marcond Cesar Jr. *Shape analysis and classification: theory and practice*. CRC press, 2010.
- [3] M Newman. *Networks: an introduction*. Oxford University Press, 1st edition edition, 2010.
- [4] Jonathan F Donges, Yong Zou, Norbert Marwan, and Jürgen Kurths. Complex networks in climate dynamics. *The European Physical Journal-Special Topics*, 174(1):157–179, 2009.
- [5] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1):1–101, 2011.
- [6] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.
- [7] Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertesz. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):027105, 2007.
- [8] L C Junqueira and J Carneiro. *Histologia Básica*. Guanabara Koogan, 11th edition edition, 2008.
- [9] M P Viana, E Tanck, M E Beletti, and L da F C Costa. Musculoskeletal ontogeny, phylogeny, and functional adaptation. *Molecular Biosystems*, 5(3):255–261, 2009.
- [10] Wikipedia. Entropy (information theory), 2016.
- [11] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.
- [12] Joseph Lee Rodgers and W Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.
- [13] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.
- [14] L da F Costa, Francisco A Rodrigues, Gonzalo Travieso, and P R Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

- [15] Norman Biggs. *Algebraic graph theory*. Cambridge university press, 1993.
- [16] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.
- [17] Ana M Martín González, Bo Dalsgaard, and Jens M Olesen. Centrality measures and the importance of generalist species in pollination networks. *Ecological Complexity*, 7(1):36–43, 2010.
- [18] Paul W Holland and Samuel Leinhardt. Transitivity in structural models of small groups. *Comparative Group Studies*, 1971.
- [19] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’ networks. *nature*, 393(6684):440–442, 1998.
- [20] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.
- [21] Guoyong Mao and Ning Zhang. Analysis of average shortest-path length of scale-free network. *Journal of Applied Mathematics*, 2013:1–5, 2013.
- [22] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.
- [23] Douglas Brent West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.
- [24] Shi Zhou and Raúl J Mondragón. The rich-club phenomenon in the internet topology. *Communications Letters, IEEE*, 8(3):180–182, 2004.
- [25] László Lovász and Michael D Plummer. *Matching theory*, volume 367. American Mathematical Soc., 2009.
- [26] Yaneer Bar-Yam. Concepts: Power law, 2016.
- [27] Vincenzo Guerriero. Power law distribution: Method of multi-scale inferential statistics. *Journal of Modern Mathematics Frontier*, 1(1):21–28, 2012.
- [28] Narushige Shiode and Michael Batty. Power law distributions in real and virtual worlds. 2000.
- [29] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.
- [30] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell’s functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

- [31] Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.
- [32] G A Pavlopoulos, M Secrier, C N Moschopoulos, T G Soldatos, S Kossida, J Aerts, R Schneider, and P. G. Bagos. Using graph theory to analyze biological networks. *BioData mining*, 4(10):1–27, 2011.
- [33] A L Barabasi. *Linked: How everything is connected to everything else and what it means*. Plume Editors, 2002.
- [34] N Guelzim, S Bottani, P Bourguine, and F Képès. Topological and causal structure of the yeast transcriptional regulatory network. *Nature genetics*, 31(1):60–63, 2002.
- [35] J I F Bass, A Diallo, J Nelson, J M Soto, C L Myers, and A J Walhout. Using networks to measure similarity between genes: association index selection. *Nature methods*, 10(12):1169–1176, 2013.
- [36] H Jeong, S P Mason, A L Barabási, and Z N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.
- [37] A Vazquez, A Flammini, A Maritan, and A Vespignani. Modeling of protein interaction networks. *Complexus*, 1(1):38–44, 2002.
- [38] S H Yook, Z N Oltvai, and A L Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(1):928–942, 2004.
- [39] B J O’Roak, L Vives, S Girirajan, E Karakoc, N Krumm, B P Coe, R Levy, A Ko, C Lee, J D Smith, E H Turner, I B Stanaway, B Vernot, M Malig, C Baker, B Reilly, J M Akey, E Borenstein, M J Rieder, D A Nickerson, R Bernier, J Shendure, and E E Eichler. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397):246–250, 2012.
- [40] J Seebacher and A C Gavin. Snapshot: Protein-protein interaction networks. *Cell*, 144(6):1000–1000, 2011.
- [41] B Wu, C Li, Z Du, Q Yao, J Wu, L Feng, P Zhang, S Li, L Xu, and E Li. Network based analyses of gene expression profile of lcn2 overexpression in esophageal squamous cell carcinoma. *Scientific Reports*, 4(5403):1–10, 2014.
- [42] H Jeong, B Tombor, R Albert, Z N Oltvai, and A L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.
- [43] D A Fell and A Wagner. The small world of metabolism. *Nature*, 18(1):1121–1122, 2000.
- [44] M Vidal, M E Cusick, and A L Barabási. Interactome networks and human disease. *Cell*, 144(1):986–998, 2011.

- [45] C L Oliveira, J H Bates, and B Suki. A network model of correlated growth of tissue stiffening in pulmonary fibrosis. *New Journal of Physics*, 16(1):065022, 2013.
- [46] C C Hilgetag and M Kaiser. Clustered organization of cortical connectivity. *Neuroinformatics*, 2(1):353–360, 2004.
- [47] J Camacho, R Guimerà, and L A Amaral. Robust patterns in food web structure. *Physical Review Letters*, 88(22):228102, 2002.
- [48] Cynthia Martins Villar Couto, Cesar Henrique Comin, and Luciano da Fontoura Costa. Effects of threshold on the topology of gene co-expression networks. *Molecular BioSystems*, 2017.
- [49] Luciano da F Costa, Francisco A Rodrigues, and Alexandre S Cristino. Complex networks: the key to systems biology. *Genetics and Molecular Biology*, 31(3):591–601, 2008.
- [50] A L Barabási, N Gulbahce, and J Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews - Genetics*, 12(1):56–68, 2011.
- [51] Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim. A gene-coexpression network for global discovery of conserved genetic modules. *science*, 302(5643):249–255, 2003.
- [52] Luciano da Fontoura Costa, Osvaldo N Oliveira Jr, Gonzalo Travieso, Francisco Aparecido Rodrigues, Paulino Ribeiro Villas Boas, Lucas Antiqueira, Matheus Palhares Viana, and Luis Enrique Correa Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 60(3):329–412, 2011.
- [53] Juan I Fuxman Bass, Alos Diallo, Justin Nelson, Juan M Soto, Chad L Myers, and Albertha JM Wahlout. Using networks to measure similarity between genes: association index selection. *Nature methods*, 10(12):1169–1176, 2013.
- [54] Ronald A Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.
- [55] Sudharsana Sundarajan and Mohanapriya Arumugam. Weighted gene co-expression based biomarker discovery for psoriasis detection. *Gene*, 593(1):225–234, 2016.
- [56] Hyo Jung Kang, Yuka Imamura Kawasawa, Feng Cheng, Ying Zhu, Xuming Xu, Mingfeng Li, André MM Sousa, Mihovil Pletikos, Kyle A Meyer, Goran Sedmak, et al. Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370):483–489, 2011.
- [57] Ali Torkamani, Brian Dean, Nicholas J Schork, and Elizabeth A Thomas. Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. *Genome research*, 20(4):403–412, 2010.

- [58] Renuka R Nayak, Michael Kearns, Richard S Spielman, and Vivian G Cheung. Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome research*, 19(11):1953–1962, 2009.
- [59] Laura L Elo, Henna Järvenpää, Matej Orešič, Riitta Lahesmaa, and Tero Aittokallio. Systematic construction of gene coexpression networks with applications to human t helper cell differentiation process. *Bioinformatics*, 23(16):2096–2103, 2007.
- [60] Tova F Fuller, Anatole Ghazalpour, Jason E Aten, Thomas A Drake, Aldons J Lusis, and Steve Horvath. Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome*, 18(6-7):463–472, 2007.
- [61] Bin Zhang, Steve Horvath, et al. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1128, 2005.
- [62] I King Jordan, Leonardo Mariño-Ramírez, Yuri I Wolf, and Eugene V Koonin. Conservation and coevolution in the scale-free human gene coexpression network. *Molecular biology and evolution*, 21(11):2058–2070, 2004.
- [63] M Madan Babu. Introduction to microarray data analysis. *Computational genomics: Theory and application*, pages 225–249, 2004.
- [64] Rong Liu, Yu Cheng, Jing Yu, Qiao-Li Lv, and Hong-Hao Zhou. Identification and validation of gene module associated with lung cancer through coexpression network analysis. *Gene*, 563(1):56–62, 2015.
- [65] X Zheng, C Xue, G Luo, Y Hu, W Luo, and X Sun. Identification of crucial genes in intracranial aneurysm based on weighted gene coexpression network analysis. *Cancer gene therapy*, 2015.
- [66] Kylie K Harrall, Katerina J Kechris, Boris Tabakoff, Paula L Hoffman, Lisa M Hines, Hidekazu Tsukamoto, Michal Pravenec, Morton Printz, and Laura M Saba. Uncovering the liver’s role in immunity through rna co-expression networks. *Mammalian Genome*, 27(9-10):469–484, 2016.
- [67] Sven Bergmann, Jan Ihmels, and Naama Barkai. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*, 2(1):e9, 2003.
- [68] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.
- [69] Francesco Iorio, Marti Bernardo-Faura, Andrea Gobbi, Thomas Cokelaer, Giuseppe Jurman, and Julio Saez-Rodriguez. Efficient randomization of biological networks while preserving functional characterization of individual nodes. *BMC bioinformatics*, 17(1):542, 2016.

- [70] Mrityunjay Sarkar and Aurpan Majumder. Intelligent topological differential gene networks. In *Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015*, pages 79–93. Springer, 2016.
- [71] Feng Ying Yu, Zhi Hao Yang, Nan Tang, Hong Fei Lin, Jian Wang, and Zhi Wei Yang. Predicting protein complex in protein interaction network-a supervised learning based method. *BMC systems biology*, 8(3):S4, 2014.
- [72] Zhiao Shi, Catherine K Derow, and Bing Zhang. Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC systems biology*, 4(1):74, 2010.
- [73] Andy M Yip and Steve Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics*, 8(1):22, 2007.
- [74] Brynn H Voy, Jon A Scharff, Andy D Perkins, Arnold M Saxton, Bhavesh Borate, Elissa J Chesler, Lisa K Branstetter, and Michael A Langston. Extracting gene networks for low-dose radiation using graph theoretical algorithms. *PLoS Comput Biol*, 2(7):e89, 2006.
- [75] Pawel Michalak. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, 91(3):243–248, 2008.
- [76] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.
- [77] S J Pardo, M J Patel, M C Sykes, M O Platt, N L Boyd, G P Sorescu, M Xu, J J W A van Loon, M D Wang, and H Jo. Simulated microgravity using the random positioning machine inhibits differentiation and alters gene expression profiles of 2t3 preosteoblasts. *American Journal of Physiology-Cell Physiology*, 288(6):C1211–C1221, 2005.
- [78] S Reppe, D Sachse, O K Olstad, V T Gautvik, H K Sanderson, P and Datta, J P Berg, and K M Gautvik. Identification of transcriptional macromolecular associations in human bone using browser based in silico analysis in a giant correlation matrix. *Bone*, 53(1):69–78, 2013.
- [79] R Zhang, H Fang, Y Chen, J Shen, H Lu, C Zeng, J Ren, H Zeng, Z Li, D Cai, and Q Zhao. Gene expression analyses of subchondral bone in early experimental osteoarthritis by microarray. *PLoS one*, 7(2):e32356, 2012.
- [80] Lucía López-Corral, Luis Antonio Corchete, María Eugenia Sarasquete, María Victoria Mateos, Ramón García-Sanz, Encarna Ferriñán, Juan José Lahuerta, Joan Bladé, Albert Oriol, Ana Isabel Teruel,

- et al. Transcriptome analysis reveals molecular profiles associated with evolving steps of monoclonal gammopathies. *haematologica*, pages haematol–2013, 2014.
- [81] Anmaar M Abdul-Nabi, Enas R Yassin, Nobish Varghese, Hrishikesh Deshmukh, and Nabeel R Yaseen. In vitro transformation of primary human cd34+ cells by aml fusion oncogenes: early gene expression profiling reveals possible drug target in aml. *PLoS One*, 5(8):e12464, 2010.
 - [82] Gerson Shigeru Kobayashi, Lucas Alvizi, Daniele Yumi Sunaga, Philippa Francis-West, Anna Kuta, Bruno Vinícius Pimenta Almada, Simone Gomes Ferreira, Leonardo Carmo de Andrade-Lima, Daniela Franco Bueno, Cassio Eduardo Raposo-Amaral, et al. Susceptibility to dna damage as a molecular mechanism for non-syndromic cleft lip and palate. *PloS one*, 8(6):e65677, 2013.
 - [83] Yuri V Postnikov, Takashi Furusawa, Diana C Haines, Valentina M Factor, and Michael Bustin. Loss of the nucleosome-binding protein hmgn1 affects the rate of n-nitrosodiethylamine-induced hepatocarcinogenesis in mice. *Molecular Cancer Research*, 12(1):82–90, 2014.
 - [84] Jon E Ramsey and Joseph D Fontes. The zinc finger transcription factor zxdc activates ccl2 gene expression by opposing bcl6-mediated repression. *Molecular immunology*, 56(4):768–780, 2013.
 - [85] Hirofumi Satake, Kenji Tamura, Mutsuo Furihata, Takashi Anchi, Hatsune Sakoda, Chiaki Kawada, Tatsuo Iiyama, Shingo Ashida, and Taro Shuin. The ubiquitin-like molecule interferon-stimulated gene 15 is overexpressed in human prostate cancer. *Oncology reports*, 23(1):11–16, 2010.
 - [86] Dean A Baker, Lisa A Meadows, Jing Wang, Julian AT Dow, and Steven Russell. Variable sexually dimorphic gene expression in laboratory strains of drosophila melanogaster. *Bmc Genomics*, 8(1):454, 2007.
 - [87] Janina Baumbach, Mitchell P Levesque, and Jordan W Raff. Centrosome loss or amplification does not dramatically perturb global gene expression in drosophila. *Biology open*, page BIO20122238, 2012.
 - [88] Yoon-Suk Alexander Chung and Christine Kocks. Phagocytosis of bacterial pathogens. *Fly*, 6(1):21–25, 2012.
 - [89] Eric C Kong, Lorien Allouche, Paul A Chapot, Karen Vranizan, Monica S Moore, Ulrike Heberlein, and Fred W Wolf. Ethanol-regulated genes that contribute to ethanol sensitivity and rapid tolerance in drosophila. *Alcoholism: Clinical and Experimental Research*, 34(2):302–316, 2010.
 - [90] Lina E Lundberg, Margarida LA Figueiredo, Per Stenberg, and Jan Larsson. Buffering and proteolysis are induced by segmental monosomy in drosophila melanogaster. *Nucleic acids research*, 40(13):5926–5937, 2012.

- [91] Eric M Blalock, Heather M Buechel, Jelena Popovic, James W Geddes, and Philip W Landfield. Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient alzheimer's disease. *Journal of chemical neuroanatomy*, 42(2):118–126, 2011.
- [92] Marie Paschaki, Carole Schneider, Muriel Rhinn, Christelle Thibault-Carpentier, Doulaye Dembélé, Karen Niederreither, and Pascal Dollé. Transcriptomic analysis of murine embryos lacking endogenous retinoic acid signaling. *PloS one*, 8(4):e62274, 2013.
- [93] Jürgen Ehrling, Nathalie Mattheus, Dana S Aeschliman, Eryang Li, Britta Hamberger, Ian F Cullis, Jun Zhuang, Minako Kaneda, Shawn D Mansfield, Lacey Samuels, et al. Global transcript profiling of primary stems from arabidopsis thaliana identifies candidate genes for missing links in lignin biosynthesis and transcriptional regulators of fiber differentiation. *The Plant Journal*, 42(5):618–640, 2005.
- [94] Masahiro Fujita, Youko Horiuchi, Yayoi Ueda, Yoko Mizuta, Takahiko Kubo, Kentaro Yano, Shini-chiro Yamaki, Katsutoshi Tsuda, Toshifumi Nagata, Mitsuru Niihama, et al. Rice expression atlas in reproductive development. *Plant and cell physiology*, 51(12):2060–2081, 2010.
- [95] Punita Nagpal, Christine M Ellis, Hans Weber, Sara E Ploense, Lana S Barkawi, Thomas J Guilfoyle, Gretchen Hagen, José M Alonso, Jerry D Cohen, Edward E Farmer, et al. Auxin response factors arf6 and arf8 promote jasmonic acid production and flower maturation. *Development*, 132(18):4107–4118, 2005.
- [96] Hehe Wang, LaChelle Waller, Sucheta Tripathy, Steven K St Martin, Lecong Zhou, Konstantinos Krampis, Dominic M Tucker, Yongcai Mao, Ina Hoeschele, Saghai Maroof, et al. Analysis of genes underlying soybean quantitative trait loci conferring partial resistance to phytophthora sojae. *The Plant Genome*, 3(1):23–40, 2010.
- [97] Sergey N Dorogovtsev, Alexander V Goltsev, and José FF Mendes. Critical phenomena in complex networks. *Reviews of Modern Physics*, 80(4):1275, 2008.
- [98] Roger Guimera, Marta Sales-Pardo, and Luís A Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2):025101, 2004.
- [99] Roger Balian. Entropy, a protean concept. In *Poincaré Seminar 2003*, pages 119–144. Springer, 2004.
- [100] Russ B Altman and Soumya Raychaudhuri. Whole-genome expression analysis: challenges beyond clustering. *Current opinion in structural biology*, 11(3):340–347, 2001.
- [101] Almut Schulze and Julian Downward. Navigating gene expression using microarrays—a technology review. *Nature cell biology*, 3(8):E190–E195, 2001.

- [102] Dominic J Allocco, Isaac S Kohane, and Atul J Butte. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC bioinformatics*, 5(1):1, 2004.
- [103] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
- [104] Frederick James. *Statistical methods in experimental physics*. World Scientific, 2006.
- [105] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
- [106] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2015.
- [107] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.
- [108] Leo Breiman et al. Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics*, 26(3):801–849, 1998.
- [109] M Veltman et al. Regularization and renormalization of gauge fields. *Nuclear Physics B*, 44(1):189–213, 1972.
- [110] Attila Gábor and Julio R Banga. Robust and efficient parameter estimation in dynamic models of biological systems. *BMC systems biology*, 9(1):74, 2015.
- [111] R Cancedda, P Castagnola, F D Cancedda, B Dozin, and R Quarto. Developmental control of chondrogenesis and osteogenesis. *International Journal of Developmental Biology*, 44(6):707–714, 2000.
- [112] G Karsenty. The complexities of skeletal biology. *Nature*, 423(6937):316–318, 2003.
- [113] H M Kronenberg. Developmental regulation of the growth plate. *Nature*, 423(6937):332–336, 2003.
- [114] C G James, C T G Appleton, V Ulici, T M Underhill, and F Beier. Microarray analyses of gene expression during chondrocyte differentiation identifies novel regulators of hypertrophy. *Molecular biology of the cell*, 16(11):5316–5333, 2005.
- [115] C G James, L A Stanton, H Agoston, V Ulici, T M Underhill, and F Beier. Genome-wide analyses of gene expression during mouse endochondral ossification. *PLoS One*, 5(1):e8693, 2010.
- [116] D M L Cooper, A L Turinsky, C W Sensen, and B Hallgrímsson. Quantitative 3d analysis of the canal network in cortical bone by micro-computed tomography. *The Anatomical Record Part B: The New Anatomist*, 274(1):169–179, 2003.

- [117] D R Carter, M Wong, and T E Orr. Musculoskeletal ontogeny, phylogeny, and functional adaptation. *Journal of biomechanics*, 24(1):3–16, 1991.
- [118] A M Parfitt. Osteonal and hemi-osteonal remodeling: The spatial and temporal framework for signal traffic in adult human bone. *Journal of cellular biochemistry*, 55(3):273–286, 1994.
- [119] R V Sugars, E Kärner, U Petersson, B Ganss, and M Wendel. Transcriptome analysis of fetal metatarsal long bones by microarray, as a model for endochondral bone formation. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1763(10):1031–1039, 2006.
- [120] G Karsenty. Transcriptional control of skeletogenesis. *Annual Reviews of Genomics and Human Genetics*, 9(1):183–196, 2008.
- [121] A Javed, H Chen, and F Y Ghorri. Genetic and transcriptional control of bone formation. *Oral and maxillofacial surgery clinics of North America*, 22(3):283–293, 2010.
- [122] A Hari Reddi and Anand Reddi. Bone morphogenetic proteins (bmps): from morphogens to metabologens, 2009.
- [123] Sylvia A Bleuming, Xi C He, Liudmila L Kodach, James C Hardwick, Frieda A Koopman, J Fiebo, Sander JH van Deventer, Daniel W Hommes, Maikel P Peppelenbosch, G Johan Offerhaus, et al. Bone morphogenetic protein signaling suppresses tumorigenesis at gastric epithelial transition zones in mice. *Cancer research*, 67(17):8149–8155, 2007.
- [124] Jesse Even, Mark Eskander, and James Kang. Bone morphogenetic protein in spine surgery: current and future uses. *Journal of the American Academy of Orthopaedic Surgeons*, 20(9):547–552, 2012.
- [125] Mariann Bienz and Hans Clevers. Linking colorectal cancer to wnt signaling. *Cell*, 103(2):311–320, 2000.
- [126] Gijs R van den Brink. Linking pathways in colorectal cancer. *Nature genetics*, 36(10):1038–1039, 2004.
- [127] Liudmila L Kodach, Eliza Wiercinska, Noel FCC De Miranda, Sylvia A Bleuming, Alex R Musler, Maikel P Peppelenbosch, Evelien Dekker, Gijs R van den Brink, Carel JM Van Noesel, Hans Morreau, et al. The bone morphogenetic protein pathway is inactivated in the majority of sporadic colorectal cancers. *Gastroenterology*, 134(5):1332–1341, 2008.
- [128] Hassan G Moghadam, Marshall R Urist, George KB Sandor, and Cameron ML Clokie. Successful mandibular reconstruction using a bmp bioimplant. *Journal of Craniofacial Surgery*, 12(2):119–127, 2001.

- [129] UME Wikesjö, M Qahash, Y-H Huang, A Xiropaidis, G Polimeni, and Cristiano Susin. Bone morphogenetic proteins for periodontal and alveolar indications; biological observations–clinical implications. *Orthodontics & craniofacial research*, 12(3):263–270, 2009.
- [130] E Tanck, G Hannink, R Ruimerman, P Buma, E H Burger, and R Huiskes. Cortical bone development under the growth plate is regulated by mechanical load transfer. *Journal of Anatomy*, 208(1):73–79, 2006.
- [131] G Chen, G Pettet, M Pearcy, and D L S McElwain. Comparison of two numerical approaches for bone remodelling. *Medical Engineering and Physics*, 29(1):134–139, 2007.
- [132] Philip Kollmannsberger, Michael Kerschnitzki, Felix Repp, Wolfgang Wagermaier, Richard Weinkamer, and Peter Fratzl. The small world of osteocytes: Connectomics of the lacuno-canalicular network in bone. *arXiv preprint arXiv:1702.04117*, 2017.
- [133] Pascal R Buenzli and Natalie A Sims. Quantifying the osteocyte network in the human skeleton. *Bone*, 75:144–150, 2015.
- [134] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, et al. Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7):676–682, 2012.
- [135] Guido Rossum. Python reference manual. 1995.
- [136] Kálmán Palágyi and Attila Kuba. A 3d 6-subiteration thinning algorithm for extracting medial lines. *Pattern Recognition Letters*, 19(7):613–627, 1998.
- [137] Harry Bium. A transformation for extracting new descriptions of shape. In *Symposium on Models for the Perception of Speech and Visual Form*.
- [138] Ricardo Fabbri, Luciano Da F Costa, Julio C Torelli, and Odemir M Bruno. 2d euclidean distance transform algorithms: A comparative survey. *ACM Computing Surveys (CSUR)*, 40(1):2, 2008.
- [139] Kerson Huang. Statistical mechanics. Technical report, 1963.
- [140] Gábor Csányi and Balázs Szendrői. Fractal–small-world dichotomy in real-world networks. *Physical Review E*, 70(1):016122, 2004.
- [141] Filipi Nascimento Silva et al. Dimensão e simetria em redes complexas: uma abordagem multiescala. *Teste de Doutorado, Universidade de São Paulo, Instituto de Física de São Carlos*, 2016.
- [142] Filipi Nascimento Silva and Luciano da Fontoura Costa. Local dimension of complex networks. *arXiv preprint arXiv:1209.2476*, 2012.

- [143] Li Daqing, Kosmas Kosmidis, Armin Bunde, and Shlomo Havlin. Dimension of spatially embedded networks. *Nature Physics*, 7(6):481, 2011.
- [144] Matthew Roughan, Jonathan Tuke, and Eric Parsonage. Estimating the parameters of the waxman random graph. *arXiv preprint arXiv:1506.07974*, 2015.
- [145] Robin Chapman. *Surveys in Combinatorics 2011*, volume 392. Cambridge University Press, 2011.
- [146] Bernard M Waxman. Routing of multipoint connections. *IEEE journal on selected areas in communications*, 6(9):1617–1622, 1988.
- [147] Giuliano Andrea Pagani and Marco Aiello. The power grid as a complex network: a survey. *Physica A: Statistical Mechanics and its Applications*, 392(11):2688–2700, 2013.
- [148] Zengwang Xu and Robert Harriss. Exploring the structure of the us intercity passenger air transportation network: a weighted complex network approach. *GeoJournal*, 73(2):87, 2008.