Universidade Federal da Bahia

Dérick Gabriel Fernandes Borges

Estudo de estrutura modular em redes multiplex: Uma contribuição a análises filogenéticas

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós–graduação em Física do Instituto de Física da Universidade Federal da Bahia.

Orientador: Prof. Dr. Roberto Fernandes Silva Andrade
Coorientador: Prof. Dr. Gilberto Cafezeiro Bomfim

Salvador
junho de 2018

Dérick Gabriel Fernandes Borges

Estudo de estrutura modular em redes multiplex: Uma contribuição a análises filogenéticas

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós–graduação em Física do Instituto de Física da Universidade Federal da Bahia. Aprovada pela comissão examinadora abaixo assinada.

Prof. Dr. Roberto Fernandes Silva Andrade
Orientador
Instituto de Física — UFBA

Prof. Dr. Gilberto Cafezeiro Bomfim
Coorientador
Instituto de Biologia — UFBA

Profa. Dra. Suani Tavares Rubim de Pinho
Instituto de Física — UFBA

Prof. Dr. Pablo Ivan Ramos
Instituto Gonçalo Moniz — FIOCRUZ

Prof. Dr. Frederico Vasconcellos Prudente
Coordenador do programa de Pós Graduação em Física do Instituto de Física — UFBA

Salvador, 12 de junho de 2018

Dérick Gabriel Fernandes Borges

Possui graduação - bacharelado e licenciatura - em Física pela Universidade Federal da Bahia (2016). Tem experiência na área de Física, com ênfase em Física Estatística e Sistemas Complexos, atuando principalmente nos seguintes temas: modelagem computacional, estrutura modular de redes multiplex e sistemas biológicos.

Ficha Catalográfica

Borges, Dérick Gabriel Fernandes

Estudo de estrutura modular em redes multiplex: Uma contribuição a análises filogenéticas / Dérick Gabriel Fernandes Borges; orientador Roberto Fernandes Silva Andrade; coorientador: Gilberto Cafezeiro Bomfim. — Salvador : UFBA, Instituto de Física, 2018.

v., 88 f: il. ;

1. Dissertação (Mestrado em Física) - Universidade Federal da Bahia, Instituto de Física.

Inclui referências bibliográficas.

1. Física – Dissertação. 2. Redes Complexas, Multiplex, Comunidade, Filogenia, Mitocôndria. I. Andrade, Roberto Fernandes Silva. II. Bomfim, Gilberto Cafezeiro . III. Universidade Federal da Bahia. Instituto de Física. IV. Título.

CDD:

Agradecimentos

À minha mãe Maria Jose Fernandes Boa Sorte e meu avô Francisco de Assis Boa sorte, pela estrutura e suporte necessários para que pudesse concluir mais este nível em minha formação.

Ao meu orientador Roberto Fernandes Silva Andrade por todo seu apoio, dedicação e trabalho, pelas discussões, por estar sempre presente, por tudo que me ensinou e pela atenção e paciência durante meu mestrado.

Ao meu coorientador Gilberto Cafezeiro Bomfim, pela atenção e suporte durante processo de análise dos resultados.

Aos colegas do Fesc. Em especial Daniel Fernandes e Robson Pessoa, pelas inúmeras ajudas, principalmente na parte de programação.

Aos meus amigos, Weliton Fernandes, Edson Ferreira, Ivaniton Oliveira, Maurino Nascimento, Vanuzia Santos e Elisângela Neves pelos anos de amizade e inúmeras histórias vividas.

A todos os amigos e colegas, que vivenciaram um pouco desses dois anos de trabalho. Em especial a Vanuzia Santos e Vivianne Vieira, pelas sugestões e revisão do texto. E a Vagner Santos e Yulo Augusto por compartilhar dos momentos vividos dentro da sala de aula durante esse período.

A CAPES pelo apoio financeiro.

Resumo

Borges, Dérick Gabriel Fernandes; Andrade, Roberto Fernandes Silva; Bomfim, Gilberto Cafezeiro . **Estudo de estrutura modular em redes multiplex: Uma contribuição a análises filogenéticas**. Salvador, 2018. 88p. Dissertação de Mestrado — Instituto de Física, Universidade Federal da Bahia.

No âmbito da ciência de redes, apresentamos um novo método para exploração de estruturas modulares em multiplex, baseado no método desenvolvido por Newman e Girvan. Para testar o método proposto, trabalhamos com multiplex compostos por redes de proteínas codificadas em genes mitocôndriais, com o propósito de fazer uma inferência sobre o grupo irmão do ancestral mitocondrial. Os multiplex foram estabelecidos em termos de um limiar ótimo de similaridade entre as proteínas homólogas dos diversos organismos, de acordo com uma medida de distância entre redes que indica uma mudança significativa de sua estrutura modular. Nestas condições se obtém a melhor relação entre o ruído e o sinal correspondente a informação filogenética recuperável. O método se mostrou confiável através de comparação de seus resultados com os obtidos pela generalização do método de Louvain. Concluímos que o método é um bom candidato para análises de modularidade em multiplex, incluindo estudos de inferência filogenética, apresentando resultados sólidos consistentes com os princípios biológicos.

Palavras–chave

Redes Complexas, Multiplex, Comunidade, Filogenia, Mitocôndria .

Abstract

Borges, Dérick Gabriel Fernandes; Andrade, Roberto Fernandes Silva (advisor); Bomfim, Gilberto Cafezeiro . **Modular structure study in multiplex networks: A contribution to phylogenetic analyzes** . Salvador, 2018. 88p. MsC Dissertation — Instituto de Física, Universidade Federal da Bahia.

Within the network science framework, we present a new method for the exploration of modular structures in multiplex, based on the method developed by Newman and Girvan. To test the proposed method, we work with multiplex composed of networks of proteins encoded in mitochondrial genes, in order to make an inference about the sister group of the mitochondrial ancestor. Multiplex were established in terms of an optimal similarity threshold among homologous proteins of the distinct organisms, according to a network distance measure that indicates a significant change in its modular structure. Under these conditions it is obtained the best relation between the intrinsic noise resulting from the process of establishing the protein structure and the signal corresponding to recoverable phylogenetic information. The method proved to be reliable by comparing its results with those obtained by the generalization of the Louvain method. We conclude that the method is a good candidate for modularity analysis in multiplex, including phylogenetic inference studies, presenting solid results consistent with biological principles.

Keywords

Complex Networks, Multiplex, Community, Phylogeny, Mitochondria .

Sumário

1
Introdução

A física ocupa-se em explicar o comportamento de sistemas naturais sob a ação de forças entre seus constituintes, o que abrange tanto partículas elementares como planetas e galáxias. Neste contexto, não é muito imediato como a física poderia também acrescentar algo relacionado à classificação evolutiva dos seres vivos. A ideia de aplicar uma modelagem matemática ou estatística em análises filogenéticas tem gerado grandes contribuições às Ciências Biológicas. Nas últimas décadas, viu-se o desenvolvimento de um novo ramo na ciência, a teoria de Redes Complexas, hoje conhecida sob o nome de teoria de Ciência de Redes. Este processo envolveu o esforço interdisciplinar de físicos, matemáticos e outros cientistas, com o objetivo de se estabelecer conceitos e métodos precisos que têm aplicação imediata na análise de grandes conjuntos de dados (Big Data) e permitem a construção de representações adequadas para as interações entre os componentes de sistemas complexos.

Historicamente, a humanidade interessa-se em organizar e classificar os elementos do mundo à sua volta. Com relação à história evolutiva, diferentes sistemas foram empregados para compor métodos de organização e classificação de organismos, utilizando-se de critérios naturais ou artificiais. Desde a Escala Natural de Platão (período pré-Darwiniano) até a sistemática filogenética ou Cladística (período pós-Darwiniano), os métodos baseavam-se essencialmente no fenótipo dos organismos, ou seja, em suas características físicas claramente discerníveis. Entretanto, com o advento dos métodos de sequenciamento de aminoácidos, iniciado por Sanger em 1954 [1], abriu-se caminho para que proteínas de uma mesma classe, em diferentes organismos, pudessem ser comparadas quanto às suas origens evolutivas. Em 1977 [2], Sanger decodificou a primeira longa sequência do ácido desoxirribonucleico (DNA), dando início à era do sequenciamento de ácidos nucleicos, permitindo então, a comparação de genes em maior escala.

Com a revolução da informática aumentou-se de forma significativa a quantidade de bancos de dados bem como o poder de computação, fatos que deram origem à era pós-genômica [3], período que seguiu-se ao sequenciamento do genoma humano. Assim, para processar e analisar essa grande quantidade de informações houve a necessidade de novas ferramentas, produzindo uma profunda mudança na forma de pensar a ciência, havendo a necessidade cada vez maior de ir além das abordagens reducionistas e tentar entender

o comportamento dos sistemas como um todo, ou seja, uma abordagem sistêmica.

As informações, provenientes de sistemas reais são bem representadas através do ferramental matemático comumente conhecido como grafo ou rede complexa. Paralelamente aos avanços da biologia molecular e da informática, no final da década de 1990, físicos e matemáticos fizeram uma série de novas contribuições importantes para a teoria de redes complexas. Desta forma, problemas complexos em diversas áreas do conhecimento passaram a ser tratados dentro deste renovado conceito, que originalmente foi densenvolvido por Euler no século XVIII.

A física estatística tem como objetivo estudar sistemas compostos por um grande número de elementos em interação e, prevendo o comportamento macroscópico (ou coletivo) do sistema considerado a partir das leis microscópicas que governam a dinâmica do sistema [4]. Por estas razões, o desenvolvimento da teoria de sistemas complexos incorporou muitos dos conceitos surgidos no âmbito da física estatística, sendo que a mesma observação é válida para o caso das redes complexas. Conceitos como invariância de escala, dimensão fractal, transporte, expoentes críticos, sincronização, agrupamentos (clustering), com origem na física estatística, foram incorporados na teoria de redes complexas.

Dessa forma, os dados obtidos pelos projetos genomas mostraram-se aptos a serem representados por redes complexas, em que a caracterização das interações das unidades básicas de uma célula, como os genes, proteínas, metabólitos etc. é evidenciada em seu aspecto global.

Como um processo biológico não é executado somente por um elemento, mas pela interação de múltiplas unidades que, numa interação complexa, executam as atividades mantenedoras da vida [3], as redes complexas são, de fato, uma boa representação para processos biológicos como um todo.

Uma característica relevante em sistemas reais, modelados com redes é a presença de comunidades ou módulos. Esses módulos podem ser considerados compartimentos, ou grupos independentes em uma rede, desempenhando papel similar, como os tecidos, ou órgãos, do corpo humano [5]. Portanto, a detecção de comunidades é de grande importância em várias áreas da ciência, tendo vasta aplicação, em particular nas Ciências Biológicas [6].

É sabido que cada função biológica é executada por um conjunto de elementos (genes, proteínas etc.), que configuram-se como um módulo funcional, isto é, configuram-se como uma entidade discreta cuja função é separável de outros módulos funcionais [7]. Assim, podemos nos fazer o seguinte questionamento: os módulos funcionais corresponderiam aos módulos

estruturais na rede? De fato, a topologia de redes biológicas indicam a presença de uma estrutura hierárquica de módulos [8]. Portanto, a detecção desses módulos é de suma importância, contribuindo efetivamente para estudos com fins biológicos. Por exemplo em redes de interação proteína-proteína, as comunidades provavelmente agruparão proteínas com a mesma função específica dentro da célula [9]; nas redes metabólicas as comunidades podem estar relacionadas a módulos funcionais como ciclos e vias [10]; e em redes de sequências proteicas a detecção de comunidades provavelmente identificará corretamente grupos de organismos que pertencem ao mesmo táxon [11].

No contexto de inferência de filogenias (relações evolutivas entre organismos), a classificação filogenética mitocondrial tem gerado um intenso debate. Desde que a teoria da endossimbiose foi estabelecida [12], muitos estudos têm buscado definir qual grupo de bactérias seria o mais próximo do ancestral mitocondrial. Diversos estudos filogenéticos sugerem que o ancestral das mitocôndrias situa-se dentro da classe das Alfaproteobactérias [13–19]. Entretanto, não existe unanimidade sobre qual grupo de Alfaproteobactérias tem parentesco evolutivo mais proximo das mitocôndrias [20]. Nas tentativas de solucionar essa questão em relação à origem mitocondrial, diversas árvores filogenéticas foram construídas a partir da aplicação de diferentes métodos e dados de sequências, tanto proteicas quanto nucleotídicas [13–20].

A determinação de comunidades em redes complexas representa um desafio teórico e computacional. Embora tenha havido importantes contribuições ao longo dos últimos 20 anos, a questão não é ainda totalmente resolvida, principalmente para os casos em que a rede apresenta uma estrutura modular pouco evidente, ou quando o sistema real pode ser representado por mais de um conjunto de dados. Por isso, novos formalismos para ciência de redes continuam a ser desenvolvidos, tal como o desenvolvimentos de novos métodos para identificação da estrutura de comunidades em redes multicamadas [21, 22] (*multilayer networks* em inglês). Isto deve-se ao fato da abordagem tradicional de redes complexas aplicada à natureza, ocasionalmente ser incapaz de capturar completamente os detalhes presentes em alguns sistemas reais, podendo levar à descrições incorretas dos mesmos [23].

Esta dissertação, encaixa-se dentro deste cenário. Aqui apresentamos uma proposta de um novo método para determinação de comunidades em multiplex (um subconjunto das redes multicamadas), ao tempo em que fazemos detecções de estruturas modulares em várias redes biológicas de forma simultânea, levando em conta informações de distintas proteínas. Essa abordagem multiplex para modelagem de sistemas em rede permite explicitamente a incorporação da multiplicidade (diferentes tipos de

relacionamentos) e outras características de sistemas reais. Permite associar diferentes relações estruturais codificando-as em um objeto matemático conveniente, e também, unir diferentes processos dinâmicos sobre essa estrutura interconectada.

O problema de natureza biológica que buscamos responder com o método aqui proposto é a obtenção de uma classificação filogenética sem pressupostos anteriores de natureza biológica, ou seja, sem o uso de modelos evolutivos, buscando contribuir para o entendimento de qual grupo atual de bactérias é mais próximo do ancestral mitocondrial.

2
Ciência de redes

2.1
Redes Complexas

A teoria das redes complexas nasceu da aplicação de medidas desenvolvidas por conceitos provenientes da mecânica estatística, física não-linear, sistemas complexos e principalmente pela teoria dos grafos. A origem da teoria grafos retoma à solução de Euler do enigma das pontes de Konigsberg em 1736 [24]. Desde então, o conhecimento sobre grafos e suas propriedades matemáticas consolidaram [25].

Tradicionalmente, o estudo de redes complexas tem sido o território da teoria dos grafos [26]. Embora a teoria dos grafos tenha concentrado-se inicialmente em grafos regulares, desde a década de 1950 as redes de larga escala, sem princípios aparentes de design, foram descritas como grafos aleatórios, conhecidos como grafos de Erdös e Rényi. Esse paradigma começou a mudar com o trabalho de Watts e Strogatz publicado em 1998 [27] sobre rede de mundo pequeno e com o trabalho de Barabási e Albert publicado um ano depois [28] abordando redes livres de escala. A partir daí o número de publicações em revistas de física sobre redes complexas disparou, passando a ser um dos tópicos mais populares na mecânica estatística. A Fig. 2.1 mostra o número de artigos publicados no ArXiv na categoria "Physics" com a palavra chave "complex networks", o gráfico mostra o surgimento desse novo tópico e sua crescente popularidade nos anos posteriores às publicações mencionadas acima.

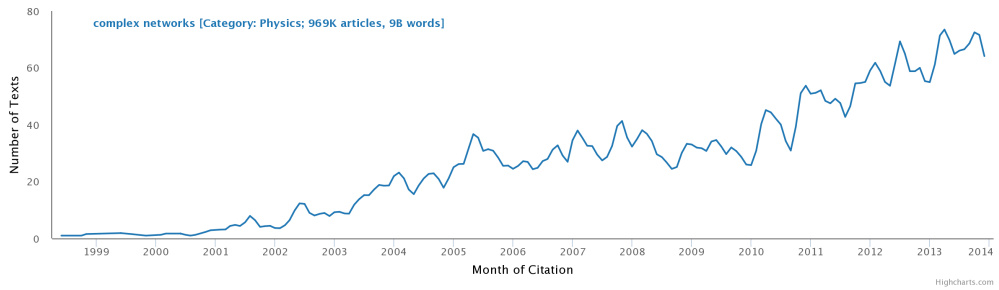


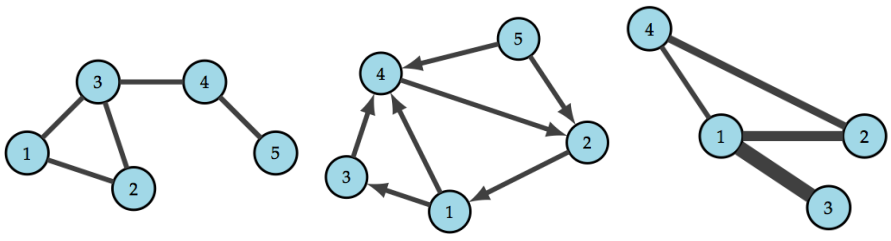
Figura 2.1: Número de artigos publicados no ArXiv dentro da categoria "Physics", contendo as palavras "complex networks". O gráfico foi gerado com ferramenta Bookworm do site http://arxiv.culturomics.org

No entanto, há dificuldade de encontrar na literatura uma conceituação clara e universal de redes complexas. Geralmente chama-se de redes complexas, grafos com propriedades topológicas não triviais [29, 30]. É importante salientar que nem todo grafo pode ser considerado uma rede complexa, pois algumas das propriedades estruturais presentes em redes complexas não são encontradas em grafos simples. Para generalização, no presente trabalho, não mais nos referiremos ao termo grafo, mas somente ao termo rede, utilizando-os como sinônimos. Da mesma forma, usaremos os termos nó e vértice indistintamente. Ao nos referirmos à topologia da rede, apenas gostaríamos de nos referir à estrutura de conexões entre os nós, sem nenhum outro compromisso maior com a teoria matemática de topologia.

Definimos uma rede R como um par ordenado de conjuntos disjuntos (X, E) tal que E é um subconjunto do conjunto de pares não ordenados dos n elementos de X. Este par não ordenado representa a existência do relacionamento entre dois vértices, e é chamado de aresta. Assim, denomina-se X de conjunto dos vértices e E de conjunto das arestas. Além disso, dois vértices são ditos adjacentes quando estão ligados a uma mesma aresta.

$$E = \{(i, j) \ ; \ i, j \in X\} \tag{2-1}$$

As ligações entre os vértices podem ter uma direção, ou seja, uma ligação aponta de um vértice para o outro, caso em que dizemos que a rede é direcionada ou orientada. No entanto, a maioria dos estudos ainda hoje consideram redes não direcionadas. Também podemos ter casos onde as ligações têm um peso. Por exemplo, ligações na internet poderiam ter um peso que representaria a quantidade de dados que passam por uma ligação. Além disso, algumas redes podem apresentar ligações múltiplas, onde há mais de uma ligação entre dois vértices e até auto-ligações, onde um vértice está ligado com ele mesmo [31]. A Fig. 2.2 mostra alguns tipos de redes.



2.2(a): Rede regular 2.2(b): Rede direcionada 2.2(c): Rede ponderada

Figura 2.2: Alguns tipos de redes.

Uma das formas de se representar matematicamente uma rede é através

da matriz de adjacência A, que para uma rede regular, é definida como:

$$A_{ij} = \begin{cases} 1 & \text{se i e j são adjacentes} \\ 0 & \text{caso contrário} \end{cases} \tag{2-2}$$

Esta matriz contém toda a informação sobre os vértices e seus relacionamentos, ou seja, guarda informação sobre todas as relações de adjacência de uma rede. Tomamos como exemplo a Fig. 2.2(a), a sua matriz de adjacência é dada por:

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{2-3}$$

Uma outra representação matricial bastante útil é a matriz de vizinhança V [32, 33], que para uma rede regular, é definida como:

$$V_{i,j} = \sum_{l=1}^{D} l \ \tilde{A} \tag{2-4}$$

onde D é o diâmetro da rede e $\tilde{A}$ é definida sobre a distância d entre nós adjacentes

$$d(i,j) = 1$$

logo

$$\tilde{A}_{ij} = \begin{cases} 1 & \text{d(i, j)} = l \\ 0 & \text{caso contrario} \end{cases} \tag{2-5}$$

A matriz de vizinhança condensa a informação contida nas $\tilde{A}$, ao tempo em que o fator l explicita a ordem de vizinhança em V, facilitando a visualização e estudo de propriedades associadas à distância entre vértices em uma rede. Tomamos como exemplo a Fig. 2.2(a), a sua matriz de vizinhança

é dada por:

$$V = 1\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} + 2\begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} + 3\begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$$V = \begin{pmatrix} 0 & 1 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 & 3 \\ 1 & 1 & 0 & 1 & 2 \\ 2 & 2 & 1 & 0 & 1 \\ 3 & 3 & 2 & 1 & 0 \end{pmatrix} \tag{2-6}$$

Formalizado a definição de rede, vamos então definir matematicamente algumas de suas propriedades estruturais.

Centralidade de grau e sua distribuição: A medida de centralidade de grau consiste no número de ligações que o vértice i realiza com vertices vizinhos j, ou seja, é definido como o conjunto da cardinalidade do conjunto de vértices adjacentes a i, definido por:

$$k_i = \sum_{ij} A_{ij} = \sum_{ij} A_{ji} \tag{2-7}$$

Desta medida podemos extrair três quantidades. A primeira o grau médio, que constitui da média aritmética da sequência de grau, dado por:

$$< k >= \frac{1}{n}\sum_{i} k_i = \sum_{ij} A_{ji} \tag{2-8}$$

sendo n o número de vértices e o k_i grau do vértice i. Como o somatório de k_i em todos os vértices da rede é igual ao dobro de ligações, assim, podemos calcular o número de arestas da rede, dado por:

$$m = \frac{1}{2}\sum_{i} k_i = \frac{1}{2}\sum_{ij} A_{ji} \tag{2-9}$$

Outra medida relevante é a distribuição de grau, uma função de distribuição probabilística que indica a probabilidade de um determinado vértice ter grau fixo. Uma maneira de quantificar essa distribuição é por meio

de uma função de distribuição cumulativa dada por:

$$P(k) = \sum_{k'=k} f(k') \tag{2-10}$$

onde $f(k')$ é a fração de nós da rede com grau k'. Esta distribuição possibilita de forma simples, quantificar o comportamento das ligações na estrutura da rede [34] e, também, é muito utilizada para caracterizar redes complexas.

Coeficiente de aglomeração: O coeficiente de aglomeração é a probabilidade de que o vértice i esteja conectado com um adjacente de j, sendo i e j adjacentes. Em outras palavras, é uma medida que quantifica o número de triângulos da rede. Sendo calculada como:

$$C_i = \frac{2m_i}{k_i(k_i - 1)} \tag{2-11}$$

sendo m_i o número de arestas e k_i grau do vértice i. Também podemos definir o coeficiente de aglomeração médio da rede, dada por:

$$< C >= \frac{1}{n} \sum_i C_i \tag{2-12}$$

Dessa forma, o coeficiente de aglomeração local é uma medida de transitividade [35], e pode ser interpretado como a densidade da vizinhança do nó local.

Métricas - Caminhos e Distâncias: Um caminho entre dois vértices i e j de uma rede R é uma sequência de k vértices $v_1, ... , v_k$, em que cada vértice é visitado apenas uma vez.

O comprimento de um caminho entre dois vértices i e j equivale ao número de arestas que conectam todos vértices pertencentes a este caminho. Posto isto, é essencial esclarecer que a comunidade de redes complexas abusa um pouco destas definições, utilizando "caminho" para referir-se ao seu comprimento. Portanto, também nesta dissertação, toda vez que discutimos alguma noção relacionada a caminho entre vértices, estaremos nos referindo a seu comprimento, um número e não a uma sequência bem definida de vértices.

A distância entre dois vértices i e j é o comprimento do caminho geodésico (menor caminho) entre estes vértices, ou seja, é o comprimento associado à menor sequência de vértices entre i e j.

A distância média também conhecida como comprimento do caminho característico L de uma rede, é a média das distâncias entre todos os pares de

vértices. Ou seja, se $d(i,j)$ é a distância entre os vértices i e j e n o número de vértices da rede, dado por:

$$L = \frac{1}{n(n-1)} \sum_{i \neq j \in R} d(i,j) \tag{2-13}$$

Outro conceito relacionado a distância é a eficiência média de uma rede R, apresentado por Latora e Marchiori [36] é definido, como:

$$E(R) = \frac{1}{n(n-1)} \sum_{i \neq j \in R} \frac{1}{d(i,j)} \tag{2-14}$$

A eficiência tem uma parte importante na caracterização estrutural, porque é um indicador da capacidade do tráfico na rede.

O diâmetro D da rede é comprimento do caminho geodésico finito mais longo em qualquer lugar da rede.

Centralidade de intermediação: A centralidade de intermediação (*betweenness centrality* em inglês) é uma medida de influência de um vértice. Esta medida de centralidade, quantifica o número de vezes que um vértice age como ponte ao longo dos caminhos geodésicos entre dois outros nós [37]. Se $\sigma(i,j)$ é o número de caminhos geodésicos entre os vértices i e j e $\sigma_v(i,j)$ é o número destes caminhos que passam pelo vértice v que intermedia os vértices i e j, definimos essa medida como:

$$g_v = \sum_{i \neq j} \frac{\sigma_v(i,j)}{\sigma(i,j)} \tag{2-15}$$

Estrutura modular: De forma simplificada, a estrutura modular ou de comunidades é definida como conjunto de partições ou módulos na rede, que apresentam uma estrutura, onde muitas arestas conectam vértices do mesmo módulo e comparativamente poucas arestas conectam vértices de módulos diferentes. Uma função qualidade Q [38, 39], definida sob um modelo nulo, mede a qualidade de uma possível comunidade, ou seja, determina se a divisão da rede é ou não significativa. A função é dada por:

$$Q = \sum_{ij} \left[A_{ij} - P_{ij} \right] \delta(c_i, c_j) \tag{2-16}$$

onde $\delta(c_i, c_j) = 1$ se as atribuições da comunidade c_i e c_j dos vértices i e j são iguais e 0 caso contrário. P_{ij} é o peso esperado da aresta entre i e j sob um modelo nulo especificado.

Newman e Girvan [40] propuseram um modelo nulo definido pelo peso $\frac{k_i k_j}{2m}$, onde k_i e k_j é o grau dos nós i e j e m o número de arestas ligando nós da comunidades i com nós da comunidade j, tornado-se um dos modelos nulos mais populares na literatura.

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{2-17}$$

A função de qualidade construída sob o modelo nulo de Newman e Girvan equação 2-17 será discutida com mais detalhes na seção 2.3.

Definido as principais características estruturais de redes complexas. Vamos agora apresentar conceitos utilizados na exploração de banco de dados de proteínas, importantes para a análise filogenética [41].

Matriz de similaridade: Com base em uma sequência de genes é possível construir uma matriz de similaridade S, cujos elementos S_{ij} representam a similaridade entre os organismos i e j, calculadas como a razão

$$S_{ij} = \frac{g_{ij}}{n_g} \ , \tag{2-18}$$

em que g_{ij} é o número de unidades correspondentes em posições iguais nas sequências gênicas dos organismos i e j, enquanto n_g é o número total de unidades. Note que S é uma matriz simétrica, uma vez que todas as espécies possuem o mesmo número total de unidades, aparecendo na cadeia de acordo com a mesma sequência. Naturalmente, pode-se pensar na matriz de similaridade como a representação de uma rede ponderada. Esta definição pode ser também adotada nos casos em que se consideram proteínas e as correspondentes sequências de aminoácidos.

A partir da matriz de similaridade é possível construír uma matriz de adjacência. Estabelecendo-se um valor de similaridade σ, tal que

$$A_{ij} = \begin{cases} 1 & \text{quando } s_{ij} \geqslant \sigma \\ 0 & \text{caso contrário} \end{cases} \tag{2-19}$$

Distância entre redes: Esta medida da distância euclideana $d(\alpha, \xi)$ entre duas redes α e ξ e definida pela soma das diferenças positivas entre os elementos da matriz das duas matrizes de vizinhança correspondentes $V(\alpha)$ e $V(\xi)$ [42],

ou seja

$$d^2(\alpha,\xi) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{V_{ij}(\alpha)}{D_\alpha} - \frac{V_{ij}(\xi)}{D_\xi} \right)^2 , \tag{2-20}$$

onde D_α e D_ξ representam os diâmetros de rede correspondentes. Se $D_\alpha = D_\xi$, um fator comum é removido dos dois denominadores. Para as situações comuns, onde $D_\alpha \neq D_\xi$, a definição mostrou-se bastante adequada, pois normaliza todos os termos na soma. É importante salientar que a definição 2-20 requer que as redes tenham exatamente o mesmo número de nós n. No entanto, o procedimento pode ser estendido a pares de redes que possuem aproximadamente o mesmo número de nós, descartando alguns nós na maior rede, de forma não arbitrária.

Uma aplicação direta dessa medida é sua utilização para medir a distância entre diferentes redes obtidas a partir de uma mesma matriz de similaridade [41]. Neste caso, α e ξ indicam duas redes obtidas da matriz S pela escolha de dois valores próximos do limiar de similaridade σ: $\alpha = \sigma$, $\xi = \alpha + \delta\sigma$, onde $\delta \ll 1$. Com esta análise, é possível identificar valores de σ onde o caracter modular da rede é explicitado de forma otimizada. Ou seja, é possível identificar o conjunto mínimo de arestas que estão incluídos na rede para preservar as informações biológicas relevantes necessárias para revelar o seu caráter modular.

O tópico de redes complexas é interessante por si só. No entanto, boa parte da motivação do estudo deste tópico foi apresentar de forma sucinta os principais conceitos para direcionar a discussão sobre redes multicamadas.

2.2 Multiplex

À medida que a pesquisa em sistemas complexos amadureceu, tornou-se essencial ir além dos grafos e investigar estruturas mais complicadas, mais realistas. Sistemas naturais, são nada mais que o resultado da emergente organização da dinâmica de processos que, por sua vez, envolvem uma multiplicidade de constituintes básicos, ou entidades, interagindo uns com os outros através de padrões complicados que podem, ou não, abranger múltiplos tipos de relacionamentos. Nesse contexto, para estudar sistemas reais com precisão, foi necessária uma generalização da teoria de redes complexas (tradicional) desenvolvendo-se uma base sólida e, consequentemente, novas ferramentas a fim de estudar, de maneira abrangente, sistemas com múltiplas correlações.

O formalismo de redes complexas descrito na seção 2.1, foi desenvolvido

para redes isoladas. Neste caso, não se considera a interação entre duas, ou mais redes. Boccaletti et al. [21] trazem exemplos de redes reais, na qual a abordagem tradicional apresenta limitações.

Um dos exemplos é uma rede biológica, construída a partir do mapa neural de um nematódeo específico, o *Caenorhabditis elegans* ou *C. elegans*, que consiste de 281 neurônios e cerca de duas mil conexões. Por sua vez, os neurônios podem ser conectados por uma ligação química ou por um canal iônico e os dois tipos de conexões têm dinâmicas completamente diferentes. Logo, a maneira apropriada de descrever tal rede é uma rede multiplex com duas camadas com 281 nós em cada, uma das camadas destinada às sinápticas químicas e uma outra para as interações por canal iônico. A consequência mais importante é que cada neurônio pode desempenhar um papel muito diferente nas duas camadas. Assim, uma classificação adequada deve ser capaz de distinguir os casos em que um mesmo nó pode ser de alta centralidade em uma camada, e simplesmente marginal na outra, ao passo que são informações relevantes para estudos de estruturas modulares.

Define-se uma rede multicamadas $\mathcal{M}$ como um par (G, C), sendo $G = \{G_\alpha; \alpha \subset \{1 \ldots m\}\}$, onde G_α corresponde a um grafo qualquer das m camadas (direcionado ou não direcionado, ponderado ou não ponderado) [21]. Portanto, $G_\alpha(X_\alpha, E_\alpha)$ terá a forma usual apresentada na seção 2.1. O outro conjunto C, que faz parte da definição, é formado pelos relacionamentos entre os nós das camadas distintas. Chamaremos, respectivamente, E_α de arestas intracamadas (*intralayer* em inglês) e $E_{\alpha\beta}$ de arestas intercamadas (*interlayer* em inglês). Os vértices ligados dentro da mesma camada por uma mesma aresta serão ditos adjacentes e vértices ligados em camadas distintas por uma mesma aresta serão ditos incidentes. Desta forma, temos

$$C = \{E_{\alpha\beta} \subseteq X_\alpha \times X_\beta;\ \alpha, \beta \in \{1 \ldots m\}; \alpha \neq \beta\}. \tag{2-21}$$

Além disso, o conjunto de nós da camada G_α será denotado por $X_\alpha = \{x_1^\alpha, \ldots, x_{n_\alpha}^\alpha\}$ e a matriz de adjacência de cada camada G_α será denotada por $A^{[\alpha]} = (a_{ij}^\alpha) \in \mathbb{R}^{n^\alpha \times\ n^\alpha}$ onde:

$$A_{ij}^\alpha = \begin{cases} 1 & \text{se } (x_i^\alpha, x_j^\alpha) \in E_\alpha\ , \\ 0 & \text{caso contrario,} \end{cases} \tag{2-22}$$

para $1 \leqslant i, j \leqslant n_\alpha$ e $1 \leqslant \alpha \leqslant m$. A matriz de adjacência entre camadas

correspondente a $E_{\alpha\beta}$ é a matriz $A^{[\alpha\beta]} = (a_{ij}^{\alpha\beta}) \in \mathbb{R}^{n^{\alpha\beta} \times\ n^{\alpha\beta}}$ dada por:

$$A_{ij}^{\alpha\beta} = \begin{cases} 1 & \text{se } (x_i^\alpha, x_j^\beta) \in E_{\alpha\beta}\ , \\ 0 & \text{caso contrario.} \end{cases} \tag{2-23}$$

Podemos escrever a projeção da rede $\mathcal{M}$, como $proj(\mathcal{M}) = (X_\mathcal{M}, E_\mathcal{M})$, onde:

$$X_\mathcal{M} = \bigcup_{\alpha=1}^{m} X_\alpha, \qquad E_\mathcal{M} = \left(\bigcup_{\alpha=1}^{m} E_\alpha \right) \bigcup \left(\bigcup_{\substack{\alpha\beta=1 \\ \alpha\neq\beta}}^{m} E_{\alpha\beta} \right). \tag{2-24}$$

Analogamente, podemos escrever uma matriz de adjacência para a projeção de $\mathcal{M}$. Denotaremos a matriz de adjacência de $proj(\mathcal{M}) = (X_\mathcal{M}, E_\mathcal{M})$ por $\overline{A_\mathcal{M}}$.

$$\overline{A}_{ij} = \begin{cases} 1 & \text{se existe } \alpha \mid a_{i,j}^\alpha = 1, \\ 0 & \text{caso contrario.} \end{cases} \tag{2-25}$$

O formalismo de redes multicamadas inclui vários modelos tais, como: redes multiplex [43, 44], redes temporais [45], redes interconectadas [46], redes multidimensionais [47], redes multiníveis [48], etc. Cada um destes tipos de rede multicamadas tem suas características e aplicações próprias.

Vamos centralizar nossa discussão nas redes multiplex. Uma revisão de artigos recentes neste campo, pode ser encontrada nas referências [21, 22].

Uma rede multiplex M é um tipo especial de rede multicamada em que $X_1 = X_2 = ... X_m = X$. O único tipo possível de conexões entre camadas são aquelas em que um dado nó é conectado apenas ao seu equivalente em cada uma das outras camadas, conforme o exemplo na fig.2.3. Assim, definimos um conjunto de arestas como:

$$E_{\alpha\beta} = \{(x, x); x \in X\} \text{ para cada } \alpha, \beta \in \{1, ..., m\}, \alpha \neq \beta \tag{2-26}$$

A Fig.2.3 ilustra um multiplex, formado por três camadas com 18 nós, 6 em cada camada onde cada nó tem duas conexões intercamada com seus correspondentes em cada uma das outras duas camadas. Este fato estende-se para o caso geral, onde temos um multiplex formado por n camadas e suas conexões intercamadas definida por $E_{\alpha\beta} = (x_i^\alpha, x_i^\beta)$ para $\alpha \neq \beta$. Uma primeira abordagem para o conceito de redes multiplex poderia sugerir que esses novos objetos são, na verdade, redes monoplex [22]. É claro que, se tomarmos um multiplex M, podemos associar a uma rede monoplex $\tilde{M} = (\tilde{X}, \tilde{E})$, onde $\tilde{X}$ é

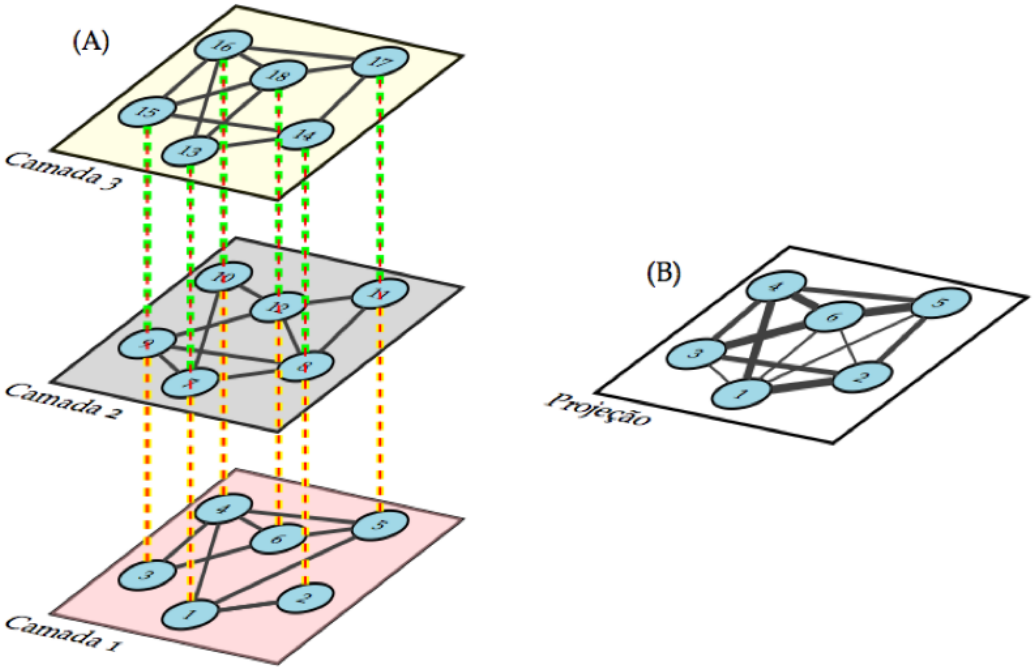


Figura 2.3: Ilustração de um multiplex e monoplex. (A) Rede multiplex com 6 nós em cada camada. As ligações em amarelo ligam os nós das camadas 1 a 2, as em verde 2 a 3 e as em vermelho 1 a 3. (B) Rede projeção da rede multiplex apresentada em (A).

a união disjunta de todos os nós de $G_1, ..., G_m$, dado por:

$$\tilde{X} = \bigcup_{1 \leqslant \alpha \leqslant m} X_\alpha = \{x^\alpha \ ; \ x \in X_\alpha\}, \tag{2-27}$$

e $\tilde{E}$ é dado por:

$$\tilde{E} = \left(\bigcup_{\alpha=1}^{m} \{(x_i^\alpha, x_j^\alpha); (x_i^\alpha, x_j^\alpha)\} \right) \bigcup \left(\bigcup_{\substack{\alpha\beta=1 \\ \alpha \neq \beta}}^{m} \{(x_i^\alpha, x_i^\beta); x_i \in X\} \right). \tag{2-28}$$

É importante reparar que o multiplex $\tilde{M}$ contêm $n \times m$ nós, e que suas matrizes de representação (adjacência, vizinhança) são quadradas e de dimensão nm.

Uma das representações matemáticas para redes multicamadas é a matriz supra-adjacente (*supra-adjacency matrix* em inglês). Nesta representação, pode-se explorar as inúmeras ferramentas, métodos e resultados teóricos que foram desenvolvidos para matrizes.

Para uma rede multiplex, a matriz supra-adjacente, tem a seguinte

estrutura de blocos

$$\tilde{A} = \begin{pmatrix} A_1 & I_n & ... & I_n \\ I_n & A_2 & ... & I_n \\ \vdots & \vdots & \ddots & \vdots \\ I_n & I_n & ... & A_m \end{pmatrix} \in \mathbb{R}^{nm \times nm}, \tag{2-29}$$

onde I_n é a matriz de identidade n-dimensional.

Uma outra maneira de representar matematicamente redes multicamadas, é baseada nos tensores de adjacência (*adjacency tensors* em inglês) [49]. De Domenico et al. definem o tensor de adjacência W sendo representado como uma combinação linear de tensores na base canônica, dado por:

$$W_{\beta}^{\alpha} = \sum_{i,j=1}^{N} w_{ij} e^{\alpha}(i) e_{\beta}(j) = \sum_{i,j=1}^{N} w_{ij} E_{\alpha}^{\beta}(ij), \tag{2-30}$$

onde, $E_{\alpha}^{\beta}(ij)$ indica o tensor na base canônica que corresponde ao produto tensorial dos vetores canônicos atribuídos aos nós n_i e n_j.

É possível escrever o tensor de adjacência para redes multicamada, como o produto tensorial entre os tensores de adjacência $C_{\beta}^{\alpha}(\tilde{h}\tilde{k})$ e os tensores canônicos $E_{\tilde{\delta}}^{\tilde{\gamma}}(\tilde{h}\tilde{k})$, dado por:

$$\begin{aligned} M_{\beta\tilde{\delta}}^{\alpha\tilde{\gamma}} &= \sum_{\tilde{h}\tilde{k}=1}^{L} C_{\beta}^{\alpha}(\tilde{h}\tilde{k}) E_{\tilde{\delta}}^{\tilde{\gamma}}(\tilde{h}\tilde{k}) \\ &= \sum_{\tilde{h}\tilde{k}=1}^{L} \left(\sum_{i,j=1}^{N} w_{ij}(\tilde{h}\tilde{k}) E_{\alpha}^{\beta}(ij) \right) E_{\tilde{\delta}}^{\tilde{\gamma}}(\tilde{h}\tilde{k}) \\ &= \sum_{\tilde{h}\tilde{k}=1}^{L} \sum_{i,j=1}^{N} w_{ij}(\tilde{h}\tilde{k}) E_{\beta\tilde{\delta}}^{\alpha\tilde{\gamma}}(ij\tilde{h}\tilde{k}), \end{aligned} \tag{2-31}$$

onde $w_{ij}(\tilde{h}\tilde{k})$ são números reais que indicam a intensidade do relacionamento (que pode não ser simétrico) entre nós n_i da camada $\tilde{h}$ e nós n_j da camada $\tilde{k}$ e $E_{\beta\tilde{\delta}}^{\alpha\tilde{\gamma}}(ij\tilde{h}\tilde{k}) \equiv e^{\alpha}(i) e_{\beta}(j) e^{\tilde{\gamma}}(\tilde{h}) e_{\tilde{\delta}}(\tilde{k})$ indica os tensores de quarta ordem da base canônica no espaço $\mathbb{R}^{N\times N\times L\times L}$.

O tensor de adjacência multicamada $M_{\beta\tilde{\delta}}^{\alpha\tilde{\gamma}}$ é muito geral, e pode ser usado para representar uma riqueza de relações entre os nós. Mais informações sobre formalismo tensorial aplicado à rede multicamadas podem ser encontradas nos seguintes artigos Ref.[21, 22, 49, 50].

Entretanto, uma vantagem adicional em se usar matrizes de

supra-adjacência ao invés de tensores de adjacência é que elas fornecem uma maneira natural de representar redes multicamadas que não estão alinhadas por nós, sem ter que anexar nós vazios [22].

Formalizado a definição de rede multiplex e suas representações, vamos então definir matematicamente suas propriedades estruturais.

Centralidade de grau: Uma das principais medidas de centralidade é o grau de cada nó: quanto mais arestas um nó possui, mais relevante é ele. O grau de um nó $i \in X$ de uma rede multiplex $M = (G, C)$ é o vetor [44] dado por:

$$\vec{k_i} = (k_i^{[1]}, ... k_i^{[m]}), \tag{2-32}$$

onde $k^{[\alpha]}$ é o grau do nó i na camada α, isto é $k^{[\alpha]} = \sum_j a_{ij}^{[\alpha]}$. Este *grau vetor* (*vector-type degree* em inglês) é a extensão natural da definição estabelecida do grau do nó em uma rede monoplex.

Coeficiente de aglomeração: O coeficiente de aglomeração de grafos introduzido por Watts e Strogatz [27] pode ser estendido para redes multicamadas de muitas maneiras. Este coeficiente quantifica a tendência dos nós para formar triângulos, seguindo o ditado popular "o amigo do seu amigo é meu amigo". Lembre-se de que, dada uma rede $G = (X, E)$, o coeficiente de aglomeração de um dado nó i definido na Equação 2-11 significa

$$C_g = \frac{\text{número de arestas entre vizinhos de } i}{\text{número total de possíveis arestas entre os vizinhos de } i} \tag{2-33}$$

Ou seja, se pensarmos em três pessoas $\{i, j, k\}$ com relações mútuas entre i e j, bem como entre i e k, o coeficiente de agrupamento de i representa a probabilidade de que j e k também estejam relacionados entre si, como definido na seção 2.1.

Generalizando o conceito de coeficiente de aglomeração para o contexto das redes multicamadas, é necessário considerar não apenas as arestas intracamadas, mas também as intercamadas. Essa generalização é obtida de forma direta identificando cada camada G_α da rede multiplex correspondente $M = (G, C)$. Antes de apresentar uma definição do coeficiente de agrupamento de um nó $i \in X$ dentro de uma rede multiplex M, vamos introduzir algumas notações.

Para cada nó $i \in X$, temos $\mathcal{N}(i)$ é o conjunto de todos os vizinhos de i na rede de projeção $proj(M)$. Para cada $\alpha \in \{1, ... , m\}$, temos que $\mathcal{N}_\alpha(i) = \mathcal{N}(i) \cap X_\alpha$ o é conjunto de todos os vizinhos do nó i da rede na

camada α. Por fim, definimos $\overline{\mathcal{S}}(i)$ como um subgrafo da camada G_α induzido por $\mathcal{N}(i)$, de forma que $\overline{\mathcal{S}}_\alpha(i) = (\mathcal{N}_\alpha(i), \overline{E}_\alpha(i))$ onde:

$$\overline{E}_\alpha(i) = \{(k,j) \in E_\alpha \ ; \ k,j \in \mathcal{N}(i). \tag{2-34}$$

Em outras palavras, com o conjunto de nós $\mathcal{N}_\alpha(i)$ e arestas $\overline{E}_\alpha(i)$ dos vizinhos de um nó i em uma camada α, constrói-se um subgrafo $\overline{\mathcal{S}}_\alpha(i)$. Similarmente, pode-se definir $\overline{\mathcal{S}}_\alpha(i)$ como o subgrafo da rede de projeção $proj(M)$ induzida por $\mathcal{N}(i)$. Com esta notação, podemos definir o coeficiente de aglomeração de um dado nó i em M como:

$$C_M(i) = \frac{2\sum_{\alpha=1}^{m} |\overline{E}_\alpha(i)|}{\sum_{\alpha=1}^{m} |\mathcal{N}_\alpha(i)| \ (|\mathcal{N}_\alpha(i)| - 1)}. \tag{2-35}$$

Então, o coeficiente de aglomeração de M pode ser definido como a média de todos $C_M(i)$.

$$< C_M >= \frac{1}{n}\sum_i C_M(i) \tag{2-36}$$

Métricas - Caminho e distância: A estrutura métrica de uma rede multicamada, assim como a de uma rede tradicional, está relacionada à distância topológica entre nós, escrita em termos de caminhos e caminhadas, como apresentado na seção 2.1. Dada uma rede multiplex $M = (G, C)$, consideramos o conjunto dado por:

$$E(M) = \{E_1, ... \ , E_m) \bigcup C. \tag{2-37}$$

Uma caminhada de comprimento $q-1$ em M, é uma sequência alternada não vazia, dada por:

$$\{x_1^{\alpha_1}, l_1, x_2^{\alpha_2}, l_2, \ ... \ , l_{q-1}, x_q^{\alpha_q}\} \tag{2-38}$$

de nós e arestas com $\alpha_1, \alpha_2, \ ... \ , \alpha_q \in \{1, ..., m\}$, de modo que para todo $r \leqslant q$ existe uma sequência $\mathcal{E} \in E(M)$, onde:

$$l_r = (x_r^{\alpha_r}, x_{r+1}^{\alpha_{r+1}}) \in \mathcal{E} \tag{2-39}$$

Se as arestas l_1, l_2, l_{q-1} forem ponderadas, o comprimento da caminhada pode ser definido como a soma do inverso dos pesos correspondentes, bem

como, se $x_1^{\alpha_1} = x_q^{\alpha_q}$ diz-se que a caminhada está fechada, ou seja, é um caminho fechado (ciclíco) começando e terminando no mesmo nó.

De forma análoga à seção 2.1 um caminho $\omega = \{x_1^1, x_2^2, \dots, x_q^q\}$ entre dois nós $x_1^{\alpha 1}$ e $x_q^{\alpha_q}$ em M, é uma caminhada através dos nós de M em que cada nó é visitado apenas uma vez.

O comprimento de um caminho é o número de arestas desse caminho. É claro que, em uma rede multiplex M existem dois tipos de arestas, nomeadas de arestas intracamadas e arestas intercamadas, por definição. Desta forma, a geodésica entre dois nós i e j em M é o caminho mais curto que liga os mesmos. A distância d_{ij} entre i e j é o comprimento de qualquer geodésica entre i e j. A distância máxima entre qualquer dois vértices em M é chamado de diâmetro $\hat{D}$.

O comprimento do caminho característico é definido, como:

$$L(M) = \frac{1}{n(n-1)} \sum_{\substack{i,j \in X_M \\ i \neq j}} d_{ij} \tag{2-40}$$

onde, $|X_M = n|$. Também, o conceito de eficiência é estendido para redes multicamadas. Assim, a eficiência de uma rede multiplex M é definida, como:

$$E(M) = \frac{1}{n(n-1)} \sum_{\substack{i,j \in X_M \\ i \neq j}} \frac{1}{d_{ij}} \tag{2-41}$$

Estruturas modulares: Um método essencial de análise de rede é a detecção de estruturas modulares conhecidas como comunidades [21]. Apesar da multiplicidade de pesquisas sobre a estrutura da comunidade em redes monoplex, o desenvolvimento de métodos de detecção de comunidades para redes multicamadas é ainda incipiente. Até agora, apenas alguns métodos de detecção de comunidades foram generalizados e desenvolvidos para redes multicamadas [51–57].

Dessas generalizações, a de maior relevância foi formulada por Mucha et al. [51] que generalizaram a função qualidade, conhecida como modularidade definida sob o modelo nulo de Newman e Girvan [40], para o estudo da estrutura de comunidade em uma configuração muito geral, abrangendo redes multicamadas. Mucha et al. definem a modularidade multifatia (*multislice modularity* em inglês), como:

$$Q_{multifatia} = \frac{1}{2\mu} \sum_{ijsr} \left[\left(A_{ij} - \gamma_s \frac{k_{is} k_{js}}{2m_s} \right) \delta_{sr} + \delta_{ij} C_{jsr} \right] \delta(C_{i,s}, C_{j,r}), \tag{2-42}$$

onde as variáveis i, j, s, r referem-se aos nós i e j, nas camadas s e r. Assim como na equação original 2-17, o termo entre parêntese dentro do colchete, contendo a matriz de adjacência A_{ij}, o parâmetro de resolução $\gamma = 1$ e grau dos nós k_{is} e k_{js} permanece inalterado e tem significado idêntico ao da equação original. Da mesma forma, na equação original, o parâmetro referente ao total de arestas μ normaliza o valor de $Q_{multifatia}$ para o intervalo $[-1, 1]$. O delta de Kronecker $\delta(s, r)$ garante que apenas os nós na mesma comunidade tenha efeito sobre o cálculo. Isso marca o final do primeiro termo da equação, que só contribui para o cálculo se os nós considerados estiverem na mesma camada.

O segundo termo da fómula corresponde à extensão da equação original, onde as diferentes camadas são incluídas. $\delta(i, j)$ é 1, quando o mesmo nó é comparado e seu impacto no cálculo final é escalado por C_{jsr}. C_{jsr} é um parâmetro fornecido pelo usuário no intervalo unitário $[0, 1]$, que indica o quanto as diferentes camadas contribuem para a modularidade.

Obviamente, pode-se explorar os métodos para a detecção de comunidade em redes monoplex, uma vez que a partir da projeção de uma rede multiplex gera-se uma rede monoplex. Trabalhos recentes vem utilizando-se dessa abordagem [58–61]. Na próxima seção aprofundaremos a discussão dos conceitos sobre Comunidades.

2.3 Comunidades

Uma das características relevantes em redes complexas que representam sistemas reais é a estrutura modular ou estrutura de comunidades [5, 26, 62, 63]. Embora esta característica já tenha sido brevemente mencionada nas duas seções anteriores, apresentamos a seguir uma discussão mais aprofundada sobre este tema.

Esta organização de estrutura modular, que desempenha um papel central na compreensão da topologia e da dinâmica da rede, indica um conjunto de vértices que interagem muito mais fortemente entre si do que com o resto da rede. Tais módulos ou comunidades, podem ser considerados como compartimentos, grupos ou clusters bastante independentes na rede, que desempenham um papel semelhante ou apresentem forte relacionamento.

A estrutura modular está presente em várias redes reais, por exemplo, as redes sociais são exemplos paradigmáticos de redes com comunidades. A própria palavra comunidade traz uma clara referência a um contexto social. As pessoas naturalmente tendem a formar grupos dentro de seu ambiente de trabalho, família e amigos.

Uma rede bem conhecida, e que é bastante usada como referência para

testar algoritmos de detecção de comunidade, é a rede de membros do karate club de Zachary [64], representada na Fig. 2.4. A rede consiste em 34 vértices e 78 arestas. Os vértices representam os membros de um clube de karatê nos Estados Unidos e as arestas conectam os indivíduos que foram observados interagindo fora das atividades do clube, durante um período de três anos. Em algum momento, um conflito entre o presidente do clube e o instrutor levou à uma divisão do clube em dois grupos, um apoiando o instrutor e o outro o presidente. A questão é se, a partir da estrutura original da rede, é possível inferir a composição dos dois grupos. De fato, observando a Fig. 2.4 pode-se identificar duas agregações, uma em torno dos vértices 33 e 34 (34 é o presidente), a outra em torno do vértice 1 (o instrutor). Pode-se também perceber alguns vértices situados entre as duas estruturas principais. A estrutura gerada apresenta uma valor de modularidade Q igual a $0,418$.

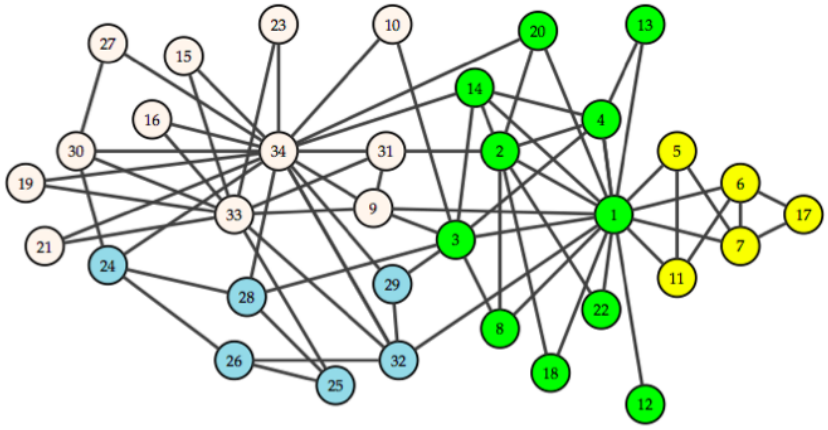


Figura 2.4: Rede Zachary karate club, uma referência padrão na detecção da comunidade, as diferentes cores indicam as comunidades detectadas na rede. (Adaptação da figura 2a Ref. [5])

Outro exemplo de redes reais com estrutura modular são as redes biológicas. A Fig. 2.5 ilustra uma rede de interação proteína-proteína (PPI) do proteoma de rato [65]. Cada interação é derivada por homologia a partir de interações observadas experimentalmente em outros organismos. No exemplo, as proteínas interagem muito frequentemente entre si, uma vez que pertencem a células metastáticas, que têm alta mobilidade e invasividade em relação às células normais. As comunidades correspondem a grupos funcionais, ou seja, grupo de proteínas com a mesma função ou funções semelhantes e que estejam envolvidas nos mesmos processos. As comunidades são rotuladas pela função global ou pela classe de proteína dominante. A maioria das comunidades detectadas está associada ao câncer e à metástase, o que mostra indiretamente,

como é importante a detecção de comunidades em redes PPI. Além desses, existem inúmeros exemplos de redes reais com estrutura modular.

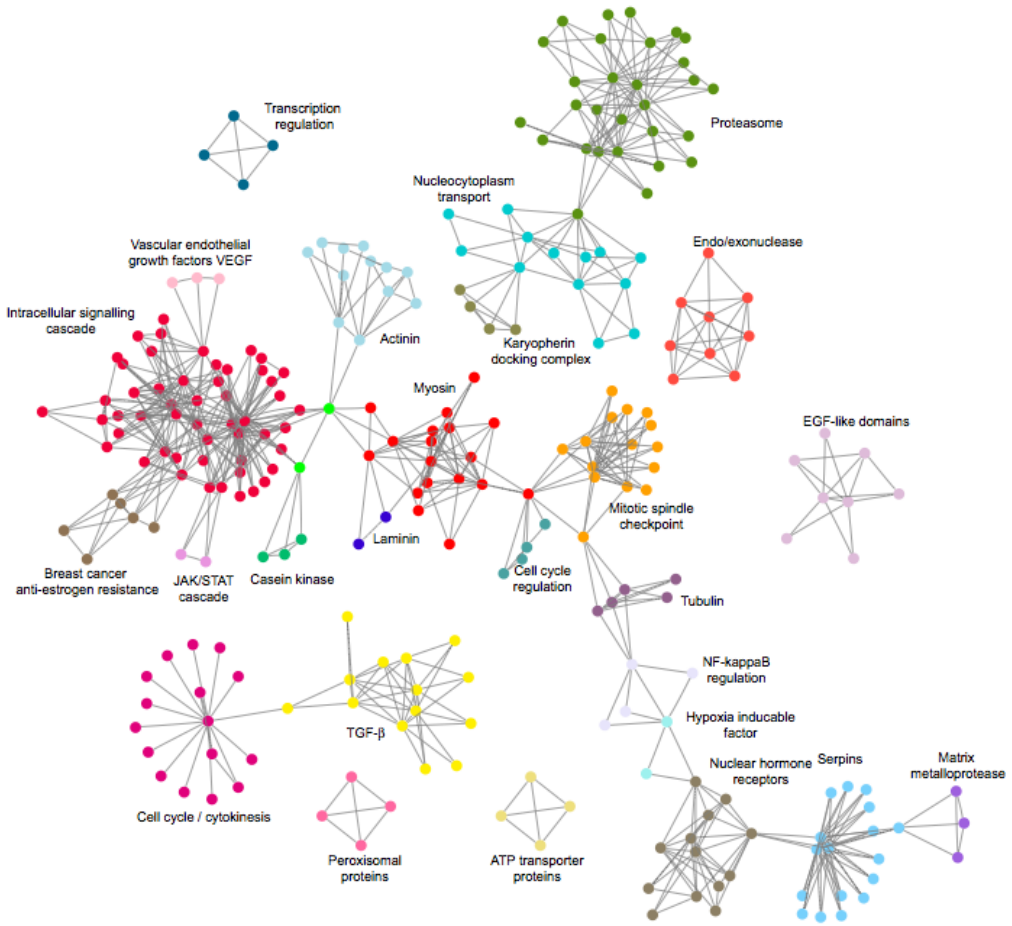


Figura 2.5: Estrutura de comunidade detectada na rede PPI, as diferentes cores representam as comunidades encontradas. Figura retirada do artigo Ref. [65].

As comunidades têm aplicações diretas no cotidiano, a exemplo do o agrupamento de clientes da web que tem interesses semelhantes e são geograficamente próximos uns dos outros. A identificação dessa comunidade pode melhorar o desempenho dos serviços prestados na rede mundial de computadores (*World Wide Web* em inglês), na medida em que cada grupo de clientes pode ser provido por um servidor específico [66]. Identificar grupos de clientes com interesses semelhantes em uma rede de relações de compra entre clientes e produtos de varejistas on-line (por exemplo, www.amazon.com) permite, também estabelecer sistemas de recomendação eficientes [67], que melhor orientam os clientes através da lista de itens do varejista e aumentam as oportunidades de negócios.

Até agora só falamos e exemplificamos as comunidades como algo vago, onde um grupo de vértices apresenta alguma semelhança ou algum

relacionamento interno. Isto acontece por não existir uma definição unívoca de comunidade.

Embora o problema de agrupamento ou detecção de comunidade seja intuitivo à primeira vista, não está bem definido. Os principais elementos do próprio problema, os conceitos de comunidade e partição não são rigorosamente definidos e exigem algum grau de arbitrariedade e subjetividade.

Na verdade, algumas ambiguidades estão encobertas e assim existem muitas maneiras igualmente legítimas para resolvê-las. Portanto, não é surpresa que haja vários métodos na literatura propondo identificação de comunidades em redes [5].

O principal objetivo da detecção de comunidades em redes é identificar os módulos e, possivelmente, sua organização hierárquica, usando apenas as informações codificadas na topologia da rede. Esse problema tem uma longa tradição e aparece de diversas formas em várias áreas das ciências.

Historicamente, a primeira análise de estrutura da comunidade foi realizada por Weiss e JaCobson [68], que procuram identificar grupos de trabalho dentro de uma agência governamental. Os grupos de trabalho foram detectados removendo os membros que trabalham com pessoas de diferentes grupos, que atuavam como conectores entre eles. Essa ideia de retirar as "pontes"entre grupos é bastante relevante e está na base de alguns dos algoritmos modernos de detecção de comunidade. A partir da década de 70 efetivamente, começaram a surgir os primeiros algoritmos de particionamento de redes [5].

Vamos supor uma sub-rede C de uma rede G, com $C = N_c$ e $G = N$ vértices, respectivamente. Definimos o grau interno e externo do vértice $v \subset C$, k_v^{int} e k_v^{ext}, como o número de arestas conectando v a outros vértices de C ou para o resto da rede, respectivamente. Se $k_v^{ext} = 0$, o vértice tem vizinhos apenas dentro de C, que é provavelmente um bom agrupamento para v. Se $k_v^{int} = 0$, e o vértice v é separado de C e ele deve ser melhor atribuído a uma comunidade diferente. O grau interno k_c^{int} de C é a soma dos graus internos de seus vértices. Da mesma forma, o grau externo k_c^{int} de C é a soma dos graus externos de seus vértices. O grau total k_c é a soma dos graus dos vértices de C. Por definição, $k_c = k_c^{ext} + k_c^{int}$.

Assim, podemos definir uma densidade intra-agrupamento $\delta_{int}(C)$ da sub-rede C como a razão entre o número de arestas internas de C e o número de todas as arestas internas possíveis, logo

$$\delta_{int}(C) = \frac{arestas\ internas\ de\ C}{n_c(n_c - 1)/2}. \tag{2-43}$$

Da mesma forma, a densidade inter-cluster $\delta_{ext}(C)$ é a relação entre o número de arestas que vão dos vértices de C para o restante da rede e o número máximo de arestas inter-cluster possíveis, logo

$$\delta_{ext}(C) = \frac{arestas\ inter-cluster\ de\ C}{n_c(n-n_c)}. \tag{2-44}$$

Fortunato [5] propõe que, para C ser uma comunidade, é necessário que $\delta_{int}(C)$ seja apreciavelmente maior que a densidade média de ligação $\delta(G)$ de G, que é dada pela razão entre o número de arestas de G e o número máximo de possíveis arestas $n(n-1)/2$. Por outro lado, $\delta_{ext}(C)$ tem que ser muito menor do que $\delta(G)$. Procurar a melhor proporção entre um grande $\delta_{int}(C)$ e um pequeno $\delta_{ext}(C)$ é implicitamente ou explicitamente o objetivo da maioria dos algoritmos de particionamento.

Sabe-se que existem vários algoritmos para identificação de estrutura de comunidade, e que levam a resultados de muito boa qualidade quando aplicados a situações em que as comunidades são previamente bem identificadas tanto em redes aleatórias geradas artificialmente, quanto em exemplos reais. No entanto, em situações práticas, os algoritmos são normalmente utilizados em redes para as quais as comunidades não são conhecidas. Isso levanta um problema: como sabemos quando as comunidades encontradas pelo algoritmo são boas? Os algoritmos sempre produzem alguma divisão da rede, mesmo em redes completamente aleatórias que não possuem uma estrutura comunitária expressiva. Por isso, seria útil ter alguma maneira de aferir o quão boa é a estrutura encontrada.

Uma maneira de quantificar as comunidades é por uma função de qualidade que compara o número de arestas intracomunitárias (arestas dentro da comunidade) com o que poderia-se esperar ao acaso. A função de qualidade mais popular é a modularidade de Newman e Girvan [40], definida pela equação 2-17.

A definição baseia-se na ideia de que uma rede aleatória não tenha uma estrutura modular, de modo que a possível existência de comunidades é revelada pela comparação entre a densidade real das arestas da sub-rede e a densidade que espera-se ter na sub-rede, se os vértices estiverem reunidos independentemente da estrutura modular.

3
Detecção de comunidades

3.1
Métodos de detecção de comunidades

Com a crescente popularização de detecção de comunidade na ciência de redes, houve investimento em desenvolvimento de métodos específicos para estudos de propriedades modulares. Desde então, muitos algoritmos para detectar comunidades foram desenvolvidos. Eles podem ser agrupados em categorias, com base em critérios específicos, dentre os métodos temos: i) métodos espectrais, ii) métodos baseados na inferência estatística, iii) métodos baseados na otimização, iv) métodos baseados na dinâmica.

Neste trabalho, no entanto, discutiremos apenas dois algoritmos, bastante utilizados na ciência de rede. Mais informações de métodos sobre detecção de comunidades em rede podem ser encontradas nos artigos de revisão Refs. [5, 69].

3.2
Algoritmo Newman e Girvan

O algoritmo proposto por Newman e Girvan [40, 70] historicamente é importante, pois marcou o início de uma nova era no campo da detecção de comunidades [5]. Aqui as arestas são selecionadas de acordo com os valores de medidas de centralidade de aresta, estimando a importância de arestas de acordo com alguma propriedade ou processo percorrendo toda a rede. As etapas do algoritmo são:

1. Calcula a centralidade para todas as arestas;
2. Remove a aresta com maior centralidade: no caso de empate com outras arestas, uma delas é mantida aleatoriamente;
3. Iteração do ciclo a partir do passo 1, até todas as arestas serem removidas.

O método proposto por Newman e Girvan se baseia no conceito de intermediação (*betweenness* em inglês), que expressa a frequência da participação da aresta no processo. Eles propõem três definições alternativas para tal medida: Intermediação de aresta (*Edge betweenness* em inglês),

Intermediação de caminhada aleatória (*Random-walk betweenness* em inglês) e Intermediação de corrente-fluxo (*Current-flow betweenness* em inglês).

A primeira proposta é a mais simples das medidas de intermediação, é baseada nos caminhos mais curtos (geodésicos): encontra-se os caminhos mais curtos entre todos os pares de vértices e conta-se quantos correm ao longo de cada aresta. Para descobrir quais arestas em uma rede são mais percorridas entre todos os pares de vértices, Newman e Girvan generalizaram para as arestas a medida de centralidade de intermediação introduzido por Freeman [37]. Historicamente, esta medida sobre as arestas foi introduzida pela primeira vez por Anthonisse em um relatório técnico nunca publicado em 1971 [71]. Anthonisse chamou-o de "apressar"(*rush* em inglês), porém, Newman e Girvan denominaram o termo como intermediação de aresta. É intuitivo que as arestas intercomunitárias tenham um grande valor da intermediação de arestas, porque muitos caminhos mais curtos conectando vértices de comunidades diferentes passarão por elas, como na Fig. 3.1. Como no cálculo de distribuição de intermediação, se houver dois ou mais caminhos geodésicos com os mesmos pontos de extremidade que percorrem uma aresta, a contribuição de cada um deles para a intermediação de aresta deve ser dividida pela multiplicidade dos caminhos, pois assume-se que o sinal ou informação se propaga igualmente ao longo de cada caminho geodésico.

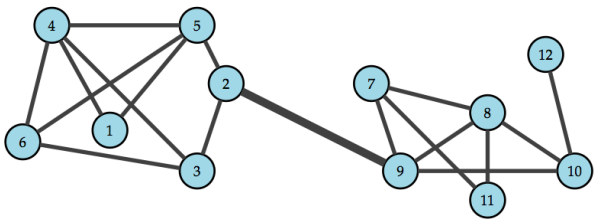


Figura 3.1: Na figura, a aresta entre os nós 2 e 9 (destacada) tem um valor de intermediação muito maior do que todas as outras arestas, porque todos os caminhos mais curtos que conectam os vértices das duas comunidades passam por ela. (Adaptação da figura 10 Ref. [5])

A outra medida de intermediação proposta por Newman e Girvan assemelha-se bastante à primeira. A intermediação de aresta pode ser pensada em termos de sinais que viajam através de uma rede. Se os sinais viajam de uma origem para um destino específico ao longo de caminhos geodésicos da rede e todos os vértices enviam sinais à taxa constante para todos os outros, então a intermediação é uma medida da taxa com que os sinais passam ao longo de cada aresta. Suponha, agora que os sinais não viajam ao longo de caminhos geodésicos, mas em vez disso, apenas faça uma caminhada aleatória

sobre a rede até chegarem ao seu destino. Isso nos dá outra medida com relação às arestas, denominada de intermediação de caminhada aleatória: calcula-se o número líquido esperado de vezes que uma caminhada aleatória entre um par particular de vértices passará por uma aresta particular, e soma-se sobre todos os pares de vértices. Ou seja, escolhe-se um par de vértices ao acaso, v_1 e v_2, o caminhante começa em v_1 e continua se movendo até atingir v_2, onde ele para. É calculada a probalidade de cada uma das arestas, que tenha sido atravessada pelo caminhante levando em conta todas as opções de caminhos possíveis entre os vértices v_1 e v_2.

Por fim, a última medida de intermediação, proposta por Girvan e Newman, foi motivada por idéias da teoria do circuito elementar. Considere uma rede com arestas com resistência unitária, se aplicarmos uma diferença de tensão entre dois vértices, cada aresta carrega alguma quantidade de corrente, que pode ser calculada pela resoluções das equações de Kirchoff. O procedimento é repetido para todos os possíveis pares de vértices. Assim, Newman e Girvan definiram a intermediação de corrente-fluxo de uma aresta como o valor médio da corrente transportada pela mesma. É possível mostrar que esta medida é equivalente à intermediação de caminhada aleatória, uma vez que as diferenças de tensão e os movimentos aleatórios (caminhante) através das arestas satisfazem às mesmas equações [72].

A grande desvantagem deste algoritmo é o tempo computacional gasto na sua execução. O cálculo da intermediação de aresta é o mais rápido dentre os apresentados, estimado-se um tempo de $O(mn)$, ou $O(n^2)$ para redes ponderadas, para a intermediação de todas as arestas da rede, onde m é o número de arestas e n é o número de vértices. Além disso, em aplicações práticas, o algoritmo de Newman e Girvan baseado na intermediação de arestas resulta em melhores resultados do que as outras medidas de intermediação [40]. A característica mais importante do algoritmo é o passo 1, o recálculo da centralidade de intermediação, que é essencial para a obtenção de resultados satisfatórios [40]. Isso introduz um fator adicional m no tempo de execução do algoritmo: conseqüentemente, a versão de intermediação de aresta passa a ter tempo computacional de $O(m^2n)$, ou $O(n^3)$ em redes ponderadas.

3.3 Algoritmo Louvain

O algoritmo proposto por Blondel et al. [73] é uma proposta ambiciosa para a otimização da função modularidade definida pela equação 2-17. Diferente da otimização proposta por Newnam [74], Blondel propõe uma otimização local da modularidade. O algoritmo é dividido em duas fases que se

repetem de forma iterativa, até que não exista mais aumento da modularidade. No início, cada nó pertence a uma comunidade formada apenas por ele e, em seguida, a cada iteração o algoritmo constrói um novo nível em um dendrograma, onde cada um desses níveis é uma partição da rede. A raiz do dendrograma é o particionamento final com a maior modularidade alcançada. As fases do algoritmo são:

1. Calcula-se o ganho de modularidade nos movimentos dos nós;

2. Agrega-se os nós das comunidades, formando um nova rede.

A primeira fase consiste em uma varredura sequencial sobre todos os vértices, atribuindo-se inicialmente uma comunidade diferente para cada vértice da rede. Dado um vértice i, calcula-se a variação da modularidade ΔQ que vai colocar i na comunidade de seu vizinho j , sendo que a escolha da comunidade do vizinho é determinada por aquela que produz o maior aumento de ΔQ, definido como

$$\Delta Q = \frac{1}{m}\left[e_{ic} - \frac{k_i k_c}{2m}\right], \tag{3-1}$$

onde e_{ic} é o grau do nó i levando em conta exclusivamente a comunidade para a qual ele se deslocar, e k_i e k_c respectivamente grau de i e da comunidade c.

No final da varredura, obtém-se a partição de primeiro nível, quando os máximos locais da modularidade são alcançados, ou seja, quando nenhum movimento individual pode melhorar a variação da modularidade.

A segunda fase consiste na construção de uma nova rede cujos nós são agora as comunidades encontradas durante a primeira fase. Para fazer isso, os pesos das arestas entre os novos vértices são dados pela soma do peso das arestas entre vértices nas duas comunidades correspondentes [75] formando uma rede ponderada. Assim, os novos vértices são agora representados pela soma dos pesos das arestas entre os vértices que constituem a comunidade e as arestas entre os novos vértices são a soma dos pesos das arestas entre vértices das comunidades.

Uma vez que esta segunda fase é completada, então é possível voltar a aplicar a primeira fase do algoritmo para a rede resultante. Por construção, o número de comunidades diminui em cada passagem e, como consequência, a maior parte do tempo de cálculo é usada na primeira fase. As fases são iteradas até que não haja mais alterações e um máximo de modularidade é atingida. Veja o exemplo na figura 3.2.

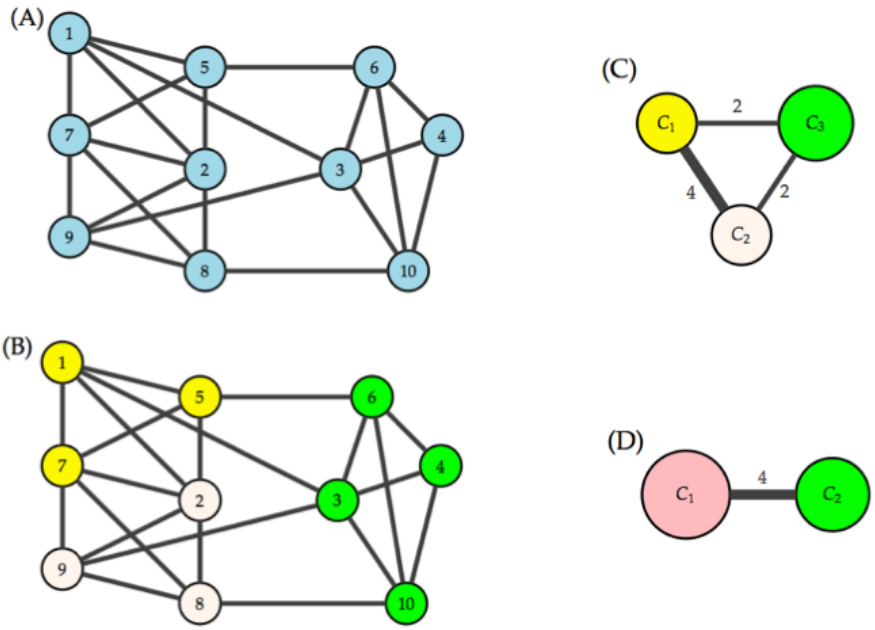


Figura 3.2: Ilustração das fases do algoritmo Louvain. (A) Rede original, na qual será executado o algoritmo. (B) Resultado após a primeira fase, as diferentes cores representam as comunidades encontradas. (C) Resultado após a segunda fase, como podemos ver, onde as comunidades encontradas são agregadas para construir uma nova rede. (D) Resultado final. As fases são repetidas iterativamente até que não seja possível aumentar a modularidade.

3.4 Generalização dos métodos para multiplex

Como já mencionado, nos últimos anos o estudo de redes multicamadas vem tornando-se uma das direções relevantes na ciência de redes. Também ficou evidente a importância da ciência de redes em problemas de detecção de comunidades, em particular em redes baseadas em dados biológicos. Este trabalho, direcionado para detecção de comunidades em multiplex, apresenta contribuições de caráter metodológico e também resultados para um problema de relevância na biologia evolutiva. Apresentamos a seguir uma discussão sobre nossas contribuições, ligadas à generalização do algoritmo de Newman e Girvan, sendo marco deste trabalho, e a utilização da generalização do algoritmo Louvain [52], ambos para detectar comunidades em multiplex.

Dentre os algoritmos encontrados na literatura para detecção de comunidades em redes multicamadas, temos: Genlouvain [52], Molti [53], Sidmod [54], LART [55], multiplex Infomap [56] e Mux-licod [57]. Até o momento, não foi encontrado na literatura a generalização do algoritmo Newman e Girvan, tal fato, talvez deva-se ao custo computacional do algoritmo, uma vez que seu tempo de execução é da ordem $O(m^2n)$. Entretanto,

apesar do custo computacional, este algoritmo é bastante útil pois permite identificar as sequências dos eventos de ramificação, levando a dendrogramas úteis e bem definidos. Além disso, nosso algoritmo permite uma representação da matriz de vizinhança em código de cores, que indica claramente como os nós são agrupados em módulos, assim como a existência de sub-módulos dentro dos módulos e a distância média entre nós em módulos distintos.

Vale salientar, que quase todos os programas utilizado nesse trabalho são de autoria própria do grupo de Física Estatística e Sistemas Complexos da UFBA, exceto o GenLouvain, que está disponível para transferência na pagina *http://netwiki.amath.unc.edu/GenLouvain*. Os programas foram escritos em linguagem Fortran.

3.5 Algoritmo MultiNG

Propomos um algoritmo geral de detecção de comunidade, o MultiNG, para a detecção de comunidades que são compartilhadas por todas as camadas no multiplex.

O algoritmo é baseado na heurística do algoritmo Newman e Girvan, utilizando-se do conceito já conhecido, a intermediação de arestas, discutido na seção 3.2. Em uma caminhada geodésica aleatória no multiplex, as arestas intracamadas com maior valor de intermediação são retidas, mantendo-se as arestas intercamadas, ou seja, na escolha aleatória de dois nós i e j, um caminhante para ir do nó i ao nó j pode "andar" de uma camada a outra ou na própria camada, sem restrições, determinando qual aresta intracamada tem maior valor de intermediação e removendo-a. As arestas intracamadas são removidas uma a uma, e a cada iteração, calcula-se a intermediação para todas as arestas de todas as camadas. Naturalmente, as arestas intercamadas podem apresentar maior valor de intermediação. Entretanto, para preservar a definição de multiplex elas nunca são removidas no processo. Quando uma aresta intracamada é removida, independentemente à camada da qual ela é retirada, o algoritmo força um nó e todos os seus correspondentes nas outras camadas a estarem na mesma comunidade.

O algoritmo opera a partir de uma matriz supravizinhança (generalização da matriz vizinhança para multiplex), e o processo de iteração é encerrado depois que todas as arestas intracamadas do multiplex tenham sido removidas. Como resultado final, o algoritmo gera um dendrograma, representando as etapas de remoção de cada aresta intracamada e um diagrama de cores com informações das distâncias médias entre nós em módulos distintos.

A análise do dendrograma tem certo grau de arbitrariedade, pois o

número de comunidades encontradas depende da posição do corte tomado no eixo horizontal, que representa o número de conexões removidas. Para saber a efetividade, pode-se calcular o valor da modularidade em relação ao número de comunidade encontradas.

Vale ressaltar que o algoritmo Newman e Girvan foi um marco histórico na ciência de redes, dando início a uma nova era no campo da detecção de comunidades. Dentro desse ponto de vista, sua generalização para redes multiplex é bastante relevante.

3.6 Algoritmo GenLouvain

Com o propósito de se obter uma maneira de aferir quantitativamente a qualidade de uma partição multiplex [51], é possível então calcular a correspondente da função modularidade. Dentro dessa pespectiva foi proposta uma generalização do algoritmo Louvain [73], por Mucha et al. [51, 52].

Algoritmo escrito originalmente dentro do ambiente MATLAB, o GenLouvain, é baseado na heurística do algoritmo Louvain. Ele permite que o usuário defina uma função de qualidade em termos de uma estrutura de modelo nulo-modularidade e, posteriormente, segue um processo iterativo de duas fases semelhante ao método Louvain . Uma distinção importante é que o algoritmo opera a partir de uma matriz de modularidade e não da matriz de adjacência, conforme o método do Louvain original. A matriz de modularidade é definida por

$$B_{ij} = A_{ij} - P_{ij}, \tag{3-2}$$

onde P_{ij} é o peso esperado da aresta entre i e j definido por um modelo nulo especificado.

Como resultado o algoritmo fornece as comunidades encontradas compartilhadas em todas as camadas e um valor de modularidade máximo, como demostrado na Fig.3.3. Contudo, o mesmo não fornece um processo histórico da separação das comunidades, ou seja, não mostra quais nós se separaram antes dos outros.

3.7 Outras generalizações

Além da generalização do algoritmo Newman e Girvan, como contribuições adicionais deste trabalho, foram ampliados os conceitos de matrizes de vizinhança e similaridade, definindo respectivamente as matrizes

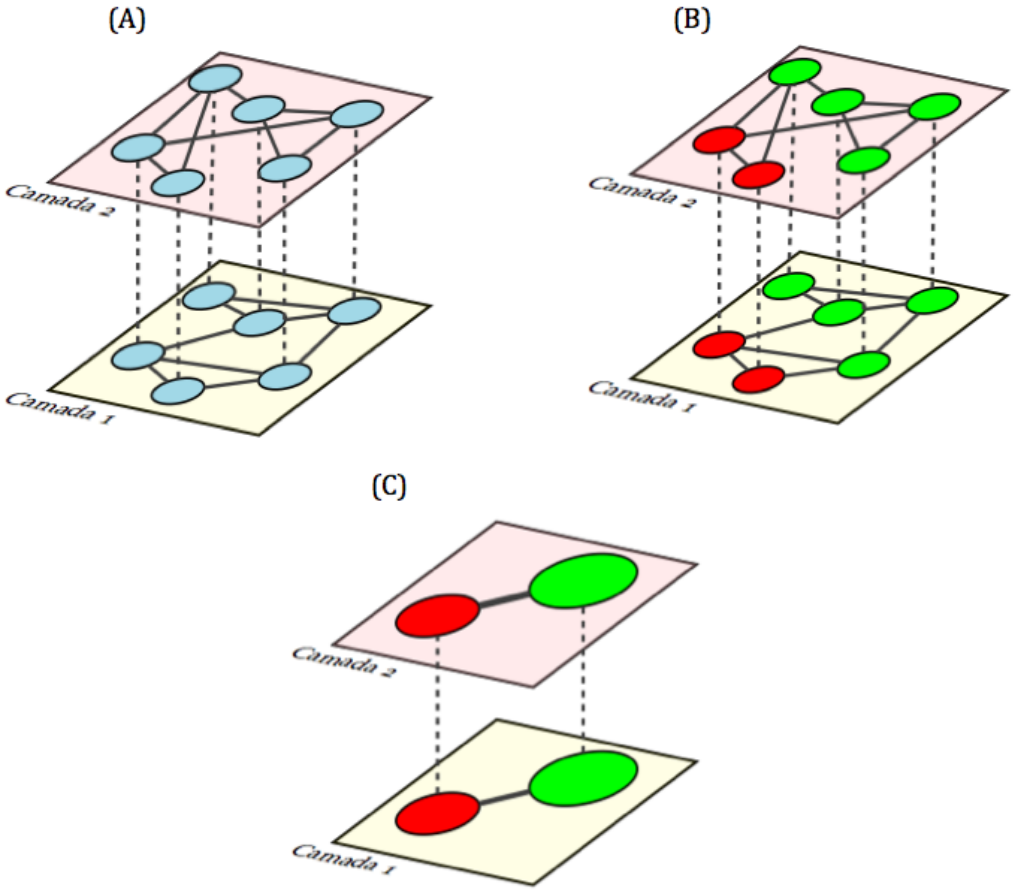


Figura 3.3: Ilustração das fases do algoritmo GenLouvain. (A) Rede original, na qual será executado o algoritmo. (B) Resultado após a primeira fase, as diferentes cores representam as comunidades encontradas. (C) Resultado após a segunda fase, como podemos ver, onde as comunidades encontradas são agregadas para construir uma nova rede. As fases são repetidas iterativamente até que não seja possível aumentar a modularidade.

de supravizinhança e suprasimilaridade. Além disso, foi proposto o cálculo de distância entre redes sobre multiplex.

Análogo à matriz de vizinhança, a matriz supravizinhança fornece informações úteis sobre distância entre vértices, dado pela matriz de blocos

$$\tilde{V} = \begin{pmatrix} V_1 & \hat{V}_n & \dots & \hat{V}_n \\ \hat{V}_n & V_2 & \dots & \hat{V}_n \\ \vdots & \vdots & \ddots & \vdots \\ \hat{V}_n & \hat{V}_n & \dots & V_m \end{pmatrix} \in \mathbb{R}^{nm \times nm} \tag{3-3}$$

onde a matriz de vizinhança V carrega a informação da distância entre os nós da mesma camada e a matriz de vizinhança $\hat{V}$ carrega a informação da distância entre os nós das camadas.

Da definição 3-3 pode se explorar o cálculo de distância entre redes [42] no multiplex. Essa medida ajuda na identificação do caracter modular do mesmo de maneira objetiva. A distinção no cálculo de $d(\alpha, \xi)$ está na utilização da matriz supravizinhança e não de uma matriz de vizinhança.

$$d^2(\alpha, \xi) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\tilde{V}_{ij}(\alpha)}{D} - \frac{\tilde{V}_{ij}(\xi)}{D} \right)^2 \tag{3-4}$$

Além disso, generalizamos a definição da matriz de similaridade, definida por

$$\tilde{S} = \begin{pmatrix} S_1 & I_n & ... & I_n \\ I_n & S_2 & ... & I_n \\ \vdots & \vdots & \ddots & \vdots \\ I_n & I_n & ... & S_m \end{pmatrix} \in \mathbb{R}^{nm \times nm} \tag{3-5}$$

A matriz de suprasimilaridade $\tilde{S}$, é uma matriz de blocos, composta por matrizes de similaridades S e identidades I.

3.8 Validação dos algoritmos

Como já mencionado, na literatura é comum utilizar a rede Zachary karate club [64], apresentada na seção 2.3, para testar algoritmos de detecção de estrutura modular, sendo uma rede referência. Assim, validamos o algoritmo MultiNG e testamos o funcionamento do GenLouvain (em Matlab e em Fortran) a partir da mesma.

Para essa validação, construiu-se um multiplex de duas camadas, denominado Mplex2zach, onde cada camada é uma rede Zachary. Este multiplex Mplex2zach, representado matematicamente por uma matriz supra-adjacência é formado por 68 vértices e 190 arestas, das quais 156 são arestas intracamadas e 34 são arestas intercamadas.

Como os resultados da rede Zachary estão estabelecidos na literatura, esperava-se do algoritmo MultiNG para a estrutura de comunidades a obtenção de um dendrograma que bem reproduzisse as comunidades equivalentes ao resultado conhecido, já demostrado na Fig. 2.4, forçando os nós das duas camadas a estarem nas mesmas comunidades, tendo uma duplicação na escala de ruptura entre os nós. Já do algoritmo GenLouvain, esperava-se a indicação de que os nós correspondentes de cada camada, estivessem nas mesmas comunidades, comunidades estas já conhecidas como demostra a Fig. 2.4,

atingindo uma otimização da modularidade Q, próximo à $0,418$, que é o valor reportado em diversos estudos anteriores.

Executado os algoritmos sobre o multiplex Mplex2zach, conseguiu-se reproduzir satisfatoriamente os resultados esperados, descritos na Fig. 3.4 e Tab. 3.1.

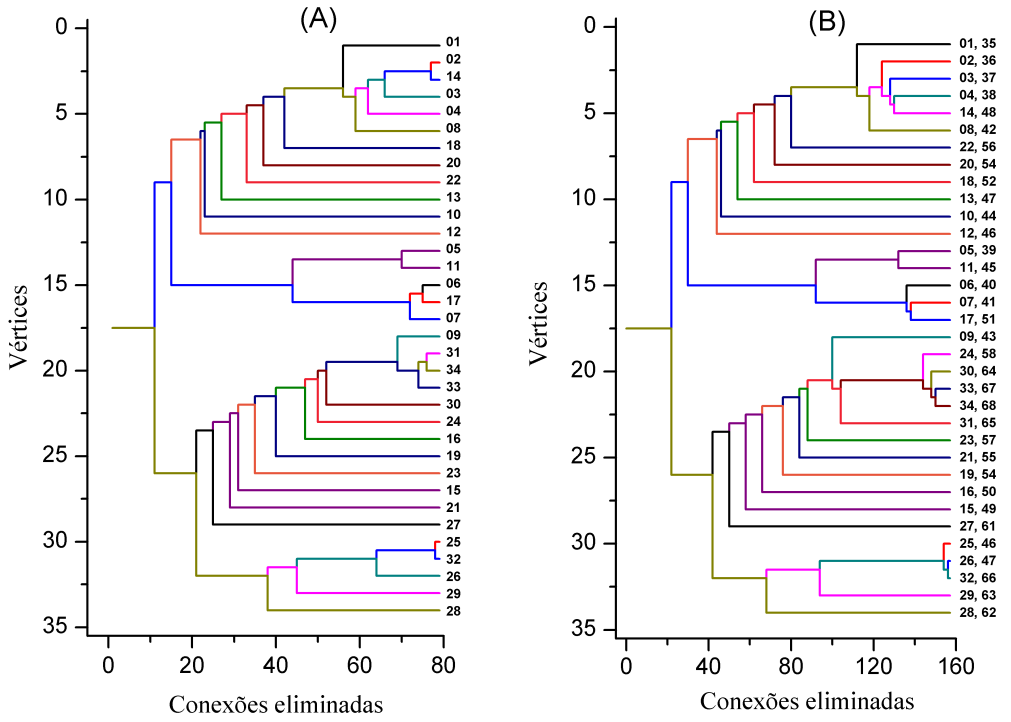


Figura 3.4: Resultados em dendrogramas. (A) Resultado utilizando algoritmo tradicional do Newman e Girvan em uma rede mono camada. (B) Resultado utilizando algoritmo MultiNG. A identificação da posição dos nós em cada ramo do dendrograma é indicada à direita.

Tabela 3.1: Comparação entre as classificações das rede Zachary com o multiplex Mplexzach

(a) Classificação original (Louvain) $Q = 0,418$

Comunidades	Vértices
C1	1, 2, 3, 4, 8, 12, 13, 14, 18, 20, 22
C2	5, 6, 7, 11, 17
C3	9, 10, 15, 16, 19, 21, 23,27, 31, 33, 34
C4	24, 25, 26, 28, 29, 32

(b) Classificação com GenLouvain $Q = 0,519$

Comunidades	Vértices
C1	1, 2, 3, 4, 8, 12, 13, 14, 18, 20, 22, 35, 36, 37, 38, 42, 46, 47, 52, 64, 56
C2	5, 6, 7, 11, 17, 39, 40, 41, 45 ,51
C3	9, 10, 15, 16, 19, 21, 23, 27, 31, 33, 34, 43, 44, 49, 50, 53, 55, 57, 61, 65, 67 ,68
C4	24, 25, 26, 28, 29, 32, 58, 59, 60, 62, 63, 66

Da mesma forma, foram realizados testes para multiplex com 3 e 4 camadas, na qual os resultados indicaram as mesmas comunidades descritas na Fig. 3.4 e Tab. 3.1, acrescidos dos nós das camadas 3 e 4.

4
Aplicação na Biologia

A abordagem de ciências de redes tem sido empregada com sucesso para descobrir princípios organizacionais, evolução e funcionamento de diversos sistemas complexos distintos em todas as áreas da ciência, em particular nas Ciências Biológicas. Nesse sentido propomos aqui um novo método descrito na seção 3.4 para identificar comunidades em estruturas interconectadas de redes e aplicá-las à análise filogenética para abordar o tema da origem evolutiva das mitocôndrias.

4.1
Classificação filogenética

Desvendar a história evolutiva dos organismos vivos é um dos maiores desafios da ciência moderna [76]. Estudos demostram que o início da vida na Terra surgiu entre 3,8 e 4 bilhões de anos atrás [77], e as primeiras tentativas importantes de reconstruir esse processo datam no século XIX, quando a teoria da evolução de Darwin foi proposta, tornando-se uma das teorias mais bem aceitas na ciência para a questão evolutiva. Historicamente, teorias e hipóteses sobre evolução estão intimamente relacionadas com a configuração de árvores filogenéticas [77–79], que reúnem diferentes espécies existentes de acordo com uma medida satisfatória de proximidade, e com o estudo de fósseis, que evidenciou as primeiras espécies vivas ao longo da história evolutiva, incluindo muitas delas hoje extintas.

Com a elucidação da estrutura do DNA e, algumas décadas depois, com o advento dos métodos de sequenciamento, tanto protéico quanto genômico, os dados moleculares foram se tornando importantes nas análises evolutivas de ancestralidade. Neste aspecto, a história evolutiva passou de um ponto de vista macroscópico para um ponto de vista molecular de análise.

O constante aumento no número de genomas disponíveis nos bancos de dados públicos exigiu um grande aumento na capacidade computacional de armazenamento e o desenvolvimento de técnicas de processamento adequadas para a análise destes dados. Algoritmos de análise tiveram de ser criados e aperfeiçoados e, dentre estes, destacam-se as técnicas de alinhamento de sequências, que tornaram-se ferramentas fundamentais na análise de sequências biológicas. Por exemplo, o programa BLAST [80] (Ferramenta de Busca por Alinhamento Local Básico, ou *Basic Local Alignment Search Tool* em inglês)

é um algoritmo capaz de realizar buscas baseadas em alinhamento local que, apesar de não serem exatas, são confiáveis e muito rápidas, sendo estas as suas vantagens em relação a outros métodos [81].

A reconstrução filogenética, ou seja, a reconstrução da história evolutiva de organismos, tendo como base as informações de suas sequencias proteicas, é um complexo processo que envolve uma série de etapas e existem vários métodos igualmente legítimos para resolvê-la. O alinhamento, além de ser o primeiro passo, é um importante ponto para a inferência de filogenias [82]. Um alinhamento preciso, além de garantir maior confiabilidade nas análises posteriores, geralmente é requerido por todos os métodos de inferência filogenética para construção da árvore. Depois que o alinhamento foi realizado, diversos métodos podem ser usados para estimar a filogenia das sequências. Dos métodos tradicionalmente empregados na inferência de filogenias, destacam-se os métodos qualitativos: Distância, Máxima Parcimônia, Máxima Verossimilhança e Inferência Bayesiana [83].

Além dessas, abordagens computacionais baseadas em matrizes de similaridade e análises de módulos ou clusters para a exploração de bases de dados de proteínas são ferramentas importantes para análise filogenética [41]. Atualmente, algumas abordagens estão sendo utilizadas para inferir relações evolutivas entre proteínas, por exemplo: o Algoritmo de Markov Cluster (MCL) [84, 85] é um algoritmo de cluster não supervisionado que foi aplicado à análise de grafos em vários domínios diferentes, principalmente em bioinformática. O Algoritmo MCL foi utilizado, por exemplo, para a detecção de famílias de proteínas [86], principal objetivo de pesquisa em genômica estrutural e funcional. O MCL também foi utilizado para desenvolver análises filogenômicas de taxa específica, como o Ascomycota [87]. Outro método desenvolvido para detecção automática e não supervisionada de famílias de proteínas e anotação do genoma é o Algoritmo Global Super Paramagnetic Clustering (SPC), que mostrou maior precisão, especificidade e sensibilidade de agrupamento do que o MCL [88].

Recentemente, novas abordagens utilizando-se do formalismo da teoria de redes complexas, vem contribuindo de forma direta para a questão da inferência filogenética. No trabalho de Goes-Neto et al. [11] para fins de estudos filogenéticos, foi utilizada a comparação tradicional de similaridade entre sequências e abordagens de redes complexas para se definir um conceito de rede critica, que exibiu a natureza modular da rede com base em um limiar de similaridade relacionado às propriedades de vizinhança na rede. Analisando dados de sequências proteicas relacionadas à via metabólica da quitina, foi demonstrado que o método proposto pode, de fato, recuperar informações

filogenéticas mesmo na ausência de conhecimento prévio sobre os sistemas em estudo. Dentro desta mesma direção, Andrade et al. [41] apresentaram uma nova abordagem para determinar a rede crítica proposta Goes-Neto et al., utilizando do conceito de distância entre redes [42] e determinando uma rede ótima, obtida através de um limiar crítico de similaridade. O método demostra quando a topologia da rede muda abruptamente, determinando uma rede ótima, que revela de maneira clara os módulos (ou comunidades) distintos relacionados aos conjuntos de organismos aos quais as proteínas pertencem. Os resultados foram obtidos a partir do mesmo conjunto de dados do trabalho de Goes-Neto et al.. Fazendo comparações entre métodos qualitativos clássicos como: Máxima Parcimônia, Máxima Verossimilhança e Inferência Bayesiana, os autores demostraram que o método é tão confiável quanto esses métodos comumente usados. Anos depois, Carvalho et al. [18] aplicaram o mesmo método [41] em um conjunto de dados de três subunidades que compõem a porção F0 da ATP sintase mitocondrial (4, 6 e 9) e suas subunidades homólogas em alfa-proteobactérias (b, a e c), e fizeram uma inferência sobre o o grupo irmão do ancestral mitocondrial.

Além disso, estudos demostraram que a utilização de redes complexas nos campos da genômica e da proteômica contribuíram para um melhor conhecimento da estrutura e dinâmica de sistemas biológicos de interações de células vivas [89, 90], sendo que vários tipos de redes biológicas relevantes foram estudados nos últimos anos, principalmente interação proteína-proteína e metabólicas [91]. Assim, a utilização da ciência de redes aplicada a Biologia, mostra-se promissora, contribuindo de forma direta para novos avanços em temas biológicos, em particular na filogenia.

4.2 Origem mitocondrial

A aquisição de mitocôndrias por parte das células eucarióticas foi um marco na história da vida. Atualmente, é largamente aceito que as mitocôndrias evoluiram de bactérias que viviam dentro de suas células hospedeiras, provavelmente há cerca de dois bilhões de anos atrás, em um processo conhecido como endossimbiose.

A teoria da endossimbiose, que é a melhor explicação para a origem de mitocôndrias e, também, de cloroplastos, foi desenvolvida e popularizada por Lynn Margulis em seu livro Origem das Células Eucarióticas (*Origin Of Eukaryotic Cells* em inglês), publicado em 1970 [92]. Essa teoria foi desenvolvida a partir do trabalho de Konstantin Merezhkovskij, botânico russo que utilizou o termo "endossimbiose" para explicar a relação entre os

cloroplastos e a célula vegetal [93].

Segundo essa teoria, todas as organelas da célula eucariótica teriam uma origem endossimbiótica. No entanto, a teoria só passou a ser bem aceita com o trabalho publicado na revista Science em 1978, por Robert Schwartz e Margaret Dayhoff [12]. Este artigo não somente reforçou a teoria, como também trouxe a hipótese de que os cloroplastos compartilham ancestralidade comum com as cianobactérias e as mitocôndrias com as proteobactérias da família *Rhodospirillaceae*. Desde então, muitos estudos têm buscado estabelecer qual grupo atual de bactérias é o mais próximo do ancestral mitocondrial.

Estudos sobre o grupo-irmão das mitocôndrias têm situado, filogeneticamente, o ancestral das mitocôndrias dentro da classe das Alfaproteobactérias. Entretanto, apesar das várias tentativas para solucionar essa questão, não existe consenso sobre qual grupo de Alfaproteobactérias tem parentesco evolutivo mais próximo das mitocôndrias [20].

Fitzpatrick et al. [14] utilizaram-se do método de máxima verossimilhança e apontaram os organimos da ordem *Rickettsiales* como o grupo mais próximo das mitocôndrias. Em outro estudo, Atteia et al. [15] usaram o mesmo método e construíram uma árvore filogenética comparando o proteoma completo de mitocôndrias de *Chlamydomonas* com homólogos das proteínas em bactérias. Os resultados indicaram que as mitocôndrias possuem um conjunto de proteínas mais próximas daquelas das ordens *Rhizobiales* e *Rhodobacterales*. Em 2010, Chang et al. [16] estudaram a ancestralidade das mitocôndrias por meio da análise de redes metabólicas mitocondriais, utilizando um cálculo da distância de matrizes entre essas redes e sugeriram o gênero *Rickettsia* como o mais próximo dessa organela. O mesmo resultado foi obtido por Rodríguez-Ezpeleta e Embley [13], a partir de sequências proteicas de genes mitocondriais conservados, utilizando-se dos métodos máxima verossimilhança e inferência bayesiana. Recentemente, Ferla et al. [17] reconstruíram árvores filogenéticas a partir de sequências de RNA ribossômico 16S e 23S através do método de máxima verossimilhança e também sustentaram a ordem *Rickettsiales* como mais proxima das mitocôndrias. Em 2015, Carvalho et al. [18] propuseram uma classificação de inferência filogenética baseada em redes complexas, para investigar origens evolutivas da mitocôndria a partir da análise de módulos (ou comunidades) de sequências proteicas em uma rede crítica obtida através de um limiar de similaridade. Utilizando sequências proteicas de três subunidades que compõem a porção F0 da ATP sintase mitocondrial (4, 6 e 9) e suas subunidades homólogas em Alfaproteobactérias (b, a e c), os autores sustentam a hipótese de que as mitocôndrias compartilham um ancestral comum com o ancestral das ordens

de Alfaproteobactérias, com exceção da ordem *Rickettsiales*. No mesmo ano, Wang e Wu [19], aumentaram a representação taxonômica de genomas de Alfaproteobactérias da sua base de dados com o sequenciamento genômico de 18 novos organismos. Com este maior conjunto de dados, que incluiu 5 *Rickettsiales* e 4 *Rhodospirillales* novos e um conjunto de 29 genes nucleares, derivados de mitocôndrias de evolução lenta que são menos tendenciosos do que genes codificados nas próprias mitocôndrias. A árvore filogenética construída através de uma abordagem filogenômica integrada sugeriu que as mitocôndrias originaram-se de uma endossimbiose de *Rickettsiales* e não dos *Rhodospirillales* de vida livre. Agora em 2018, Martijn et al. [20], com base também em novos sequenciamentos genômicos, reavaliaram o posicionamento filogenético do ancestral das mitocôndrias. A partir de um conjuntos de dados de metagenoma oceânico, os autores aumentaram a amostragem genômica das Alfaproteobactérias em doze ramos divergentes, e um ramo representando um grupo irmão para todas as Alfaproteobactéria. Os resultados sugerem que as mitocôndrias não evoluíram a partir de *Rickettsiales* ou de qualquer outra linhagem alfaproteobacterial atualmente conhecida. Em vez disso, os resultados indicaram que as mitocôndrias evoluíram a partir de uma linhagem proteobacteriana que se ramificou antes da divergência de todas as Alfaproteobactérias.

Pelo exposto, percebe-se que definir precisamente a ancestralidade das mitocôndrias tem importantes implicações para a Biologia. Esclarecerá a evolução das mitocôndrias e, consequentemente, das células eucarióticas, bem como, permitirá usar métodos comparativos para obter informações sobre a biologia do último ancestral comum das mitocôndrias e da Alfaproteobacteria.

A seguir apresentaremos o conjunto de dados usados neste trabalho.

5
Resultados

Neste capítulo apresentamos resultados para a detecção de comunidades em redes multiplex baseadas em dados genômicos com os métodos descritos anteriormente. Na primeira sub-seção fazemos uma descrição dos dados utilizados e dos métodos de alinhamento usados para se medir a similaridades entre proteínas homólogas de diferentes organismos. Em seguida, apresentamos os resultados de nossas análises, analisando em um primeiro momento as comunidades obtidas pelos diversos multiplex, sem enfatizar o significado biológico de nossos achados. Esta discussão será conduzida em um segundo momento, quando faremos uma analise de cunho filogenético. Por simplicidade, vamos limitar nossas discussões apenas às comunidades que contêm sequências referentes aos organismos eucariotos. Como já mencionado anteriormente, a análise filogenética baseada em redes multiplex aqui desenvolvida leva em conta simultaneamente as informações de similaridades protéicas oriundas de diversas proteínas homólogas. De um modelo geral, diferentes filogenias podem ser obtidas por redes geradas por cada uma das proteínas homólogas, gerando dúvidas sobre qual delas é mais apta a representar a classificação mais precisa. Esta é a maior vantagem do método aqui desenvolvido.

5.1
Conjunto de dados e análise comparativa

Neste trabalho, foi utilizado o mesmo conjunto de dados da Ref. [19] para a análise de inferência filogenética mitocondrial, por apresentar novos sequenciamentos de genomas de Alfaproteobactérias. O mesmo foi publicado na revista Scientific Reports em 2015, e o conjunto de dados encontra-se disponível para transferência na pagina *https://www.nature.com/articles/srep07949*, na seção de Informações Suplementares.

O conjunto de dados é composto por sequências proteicas codificadas tanto por genes mitocondriais quanto de genes nucleares oriundos do genoma mitocondrial. No entanto, para esse trabalho, foram utilizadas apenas as sequências codificadas por genes mitocondriais.

Das sequências proteicas disponíveis no conjunto de dados, foram escolhidas 4 sequências codificadas a partir de genes mitocondriais correspondentes nos seguintes peptídeos: *NADH desidrogenase subunidade 1* (Nad1), *NADH desidrogenase subunidade C* (Nad9), *Citocromo B* (Cob)

e *Citocromo C oxidase subunidade 2* (Cox2). Estes genes foram escolhidos pois apresentam baixa taxa de substituição nas suas sequências, sendo assim, bem conservadas [19] e além disso, por estarem representadas nos 6 organismo eucariotos em estudo: *Phytophthora infestans Hetero* (oomiceto), *Marchantia polymorpha Virid* (briófita), *Mesostigma viride Virid* (alga verde), *Reclinomonas americana Louk* (flagelado heterotrófico de vida livre), *Hemiselmis andersenii Crypt* (alga unicelular) e *Rhodomonas salina Crypt* (alga unicelular).

As sequências peptídicas estão presentes em organismos das classes Alfaproteobactérias, dois grupos externos representados por Gamaproteobactérias e Betaproteobactérias, além dos seis organismos do domínio Eukarya. A Tab. 5.1 traz as informações detalhadas dos correspondentes organismos e taxa de substituição nas sequências.

Tabela 5.1: Detalhamento dos organismos representados nas respectivas proteínas.

	Número de	Grupos:			
Proteínas	organismos	Alfa	Gama/Beta	Eukarya	Taxa de substituição
Nad1	84	70	8	6	1,20648209706513
Cox2	70	58	6	6	1,28554931826072
Nad9	78	66	6	6	1,53723701464102
Cob	80	68	6	6	1,59116671508914

A partir do conjunto de dados, foram geradas matrizes de similaridade para cada proteína, ou seja, biologicamente temos as relações de identidade em porcentual entre todos os pares de sequências que representam a proteína. Naturalmente podemos pensar em termos de rede, onde os nós são os organismos e as arestas são os valores de similaridade entre as sequências peptídicas presentes nos organismos.

Para o processo de alinhamento usou-se o programa de alinhamento de sequências, o *Clustal Omega* (versão 2.1) [94], que realiza um alinhamento global entre as sequências de aminoácidos dos organismos representados. O programa é disponibilizado virtualmente pelo Instituto Europeu de Bioinformática (EMBL-EBI), podendo ser acessado na pagina *https://www.ebi.ac.uk/Tools/msa/clustalo/*.

Os elementos das matrizes de similaridade são geradas a partir da razão aritmética entre o número de aminoácidos correspondentes em posições iguais nas sequências gênicas dos organismos e o número total de aminoácidos. As matrizes são quadradas e simétricas, e cada elemento de matriz fornece a informação de similaridade da sequência de um organismo e de outro. Por

exemplo, o elemento a_{13} da matriz fornece um valor de similaridade entre o par de sequências da proteína nos organismos 1 e 3, equivalente para o elemento a_{31} da matriz, por simetria.

Como dito anteriormente, as proteínas estão representadas em diversos grupos de organismos, demonstrado na Tab. 5.1. Dentre esses grupos, o de maior representatividade é o das Alfaproteobactérias (Alfa), subdividido em 11 grupos referentes as seguintes ordem: *Caulobacterales* (Caul), *Kopriimonadales* (Koprii), *Kordiimonadales* (Kordii), *Parvularculales* (Parvula), *Rhizobiales* (Rhizob), *Rhodobacterales* , *Rhodospirillales* (Rhodosp), *Rickettsiales*, *Sphingomonadales* (Sphing) e *Sneathiellales* (Sneath). Dentre estes subgrupos, *Rickettsiales* e *Rhodospirillales* estão entre os mais representados. A Tab. 5.2 indica o número de organismos representados por suas respectivas sequências proteicas pertencentes aos subgrupos *Rickettsiales* e *Rhodospirillales*.

Tabela 5.2: Informação da quantidade de sequências proteicas (organismos) dos pertencentes subgrupos Rickett e Rhodosp em cada proteína.

Proteínas	Rickett	Rhodosp
Nad1	16	13
Cox2	12	10
Nad9	15	12
Cob	15	13

A partir destes dados, foi calculada também a representatividade percentual de todos os grupos de organismos, em especial dos *Rickettsiales* e *Rhodospirillales* em cada uma das proteínas, como mostra a Tab. 5.3.

Tabela 5.3: Informação de representatividade dos organismos em percentual.

	Eukarya	Alfa %			Gama/Beta
Proteínas	%	Rickett	Rhodosp	Outros	%
Nad1	7,14	19,05	15,48	48,44	9,52
Cox2	8,57	17,14	14,28	51,44	8,57
Nad9	7,69	19,23	15,38	49,56	7,69
Cob	7,50	18,75	16,50	49,75	7,50

Observa-se, em todas as proteínas, que a soma da representatividade desses 2 subgrupos é quase equivalente à representatividade total de todos os outros 9 subgrupos de Alfaproteobactérias representados. Além disso, a representatividade é similar para todas as proteínas, o que dá uniformidade para a amostra e, consequentemente, consistência para os resultados aqui apresentados.

Os resultados mostrados à seguir darão destaque nos organismos desses dois subgrupos juntamente com os organismos do grupo dos Eukarya.

5.2 Construção do multiplex a partir do conjunto de dados

Geradas as matrizes de similaridade de cada proteína, podemos então inicialmente gerar as redes baseadas nas informações de cada uma delas separadamente, e em seghida construir os multiplex. A partir das 4 matrizes de similaridade, foram construídos de 22 multiplex com a combinação das mesmas, como descrito na Tab. 5.4.

Tabela 5.4: Número de multiplex construídos a partir da combinação das 4 redes

Combinação de matrizes	Número de Multiplex
2	12
3	8
4	2

É importante ressaltar que estes números resultam da adoção de dois critérios distintos para construir os multiplex:

Critério 1 Baseado no grau de similaridade individual de cada rede de proteína.

Critério 2 Baseado no grau de similaridade das combinações das redes de proteínas.

Para determinar o valor ótimo do limiar de similaridade σ na rede, e proceder a determinação de comunidades, gera-se um gráfico que relaciona a distância $d(\sigma, \sigma + \delta\sigma)$ entre as matrizes de vizinhança de duas redes obtidas para valores próximos de σ, de acordo com a equação 2-20 na seção 2.1.

Ao traçamos $d(\sigma, \sigma + \delta\sigma)$ como função de σ, verifica-se que o gráfico é caracterizado pela presença de picos agudos. Esses picos indicam uma mudança significativa na estrutura modular da rede, que é a região com melhor relação entre o ruído e o sinal correspondente à informação filogenética recuperável [41], como mostram as figuras 5.1 5.2, 5.3 e 5.4.

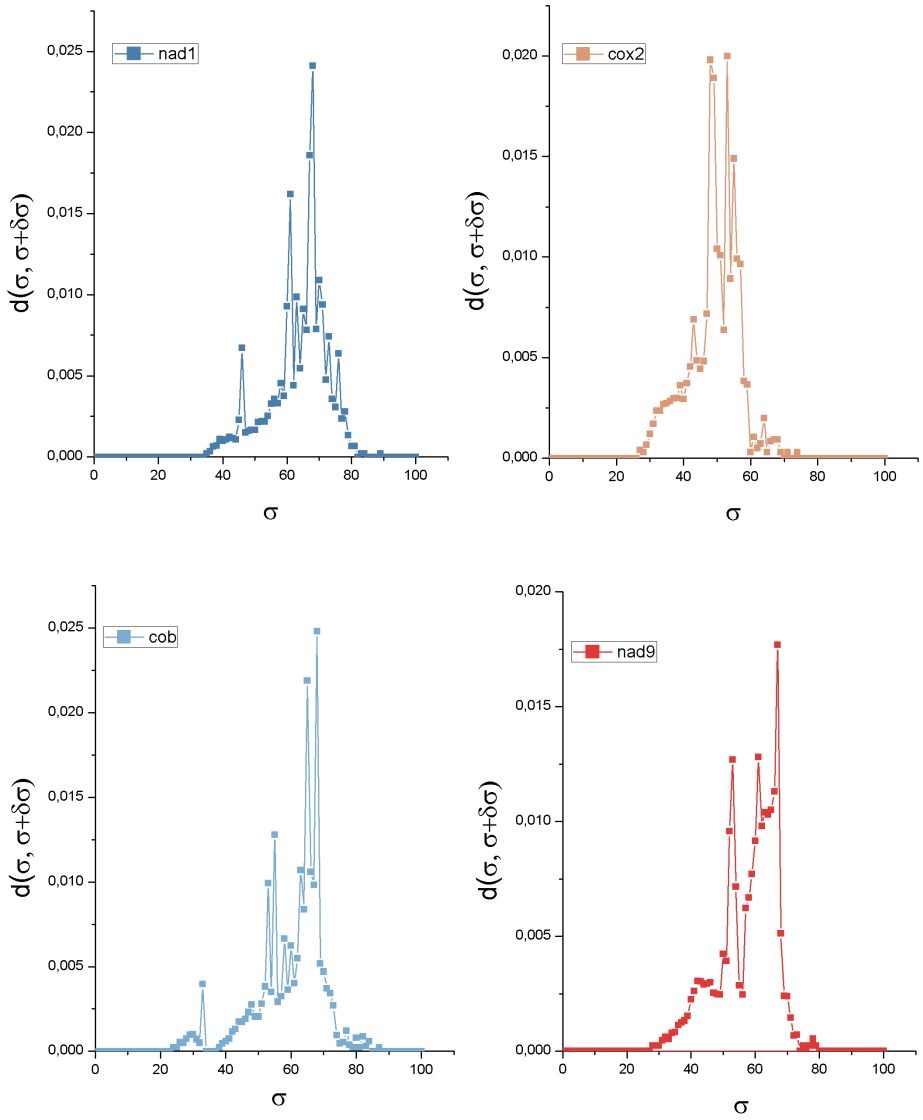


Figura 5.1: Gráficos dos picos calculado sobre cada rede indivudualmente.

A partir destes gráficos, analisou-se em qual pico o grupo dos organismos Eukarya estão correlacionadas e relacionados a alguns outros organismos, por meio de detecção de comunidade. Para este estudo não é interessante assumir um pico no qual os organismos Eukarya estejam isolados dos demais.

Determinamos então os melhores limiares de similaridades para cada rede de proteína individualmente e conjuntamente, como mostrado nas tabelas 5.5 e 5.6, e então geramos as respectivas matrizes de adjacência.

Um passo importante para a montagem do multiplex consiste no alinhamento dos organismos em cada camada. Por definição, um nó i qualquer

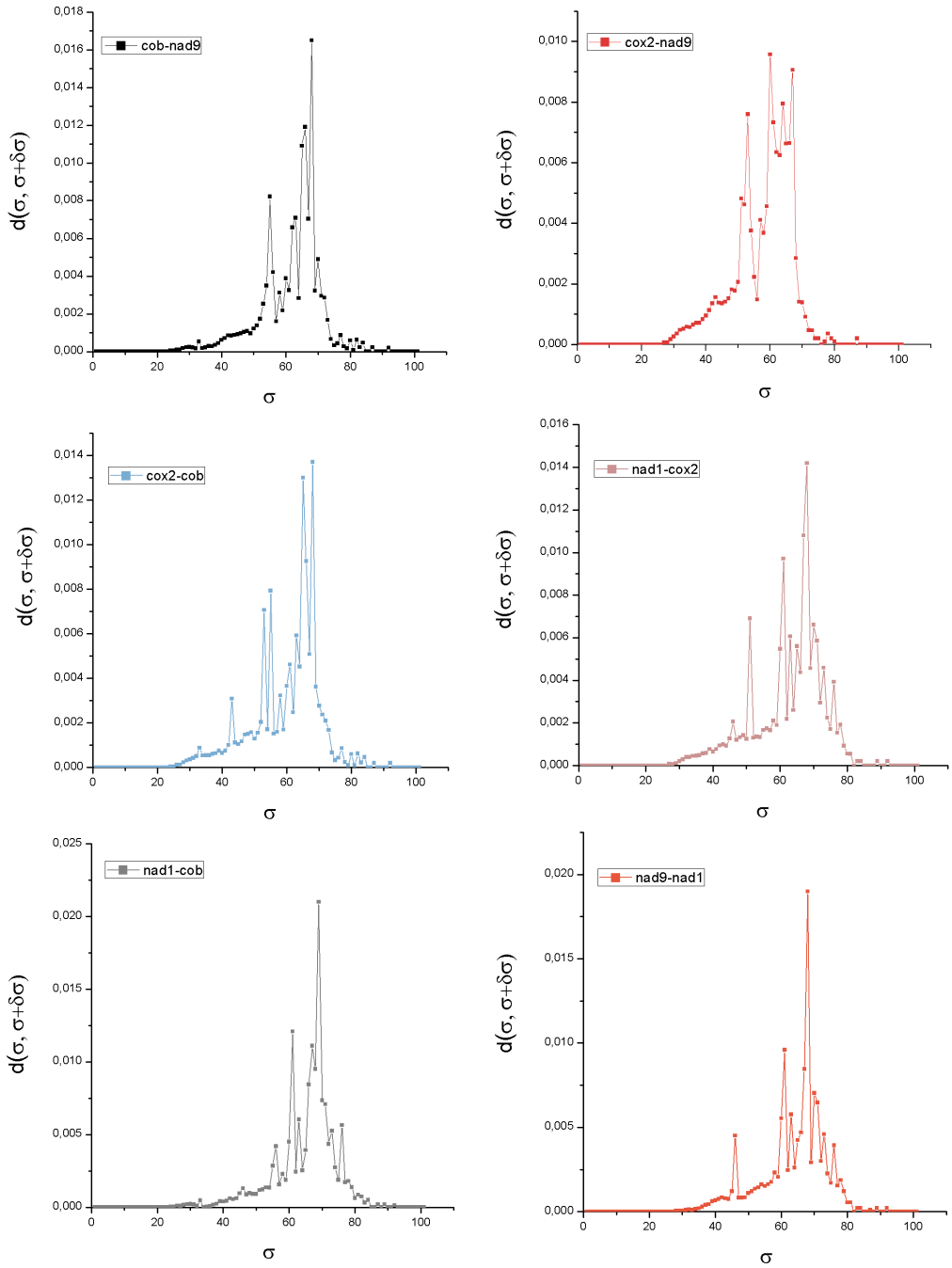


Figura 5.2: Gráficos dos picos calculado sobre multiplex de duas camadas.

do multiplex deve ter seu correspondente na mesma posição em todas as camadas α, na ordem de $i_\alpha = i + n(\alpha - 1)$, para todo $\alpha = 1, 2, 3 \ldots l$, sendo n número de nós. Por exemplo, se supusemos um multiplex composto de 3 redes de 10 nós ($n = 10$), o organismo X representado pelo nó 1 da camada 1, tem que ser representado pelo nó 11 na camada 2, e pelo nó 21 na camada 3.

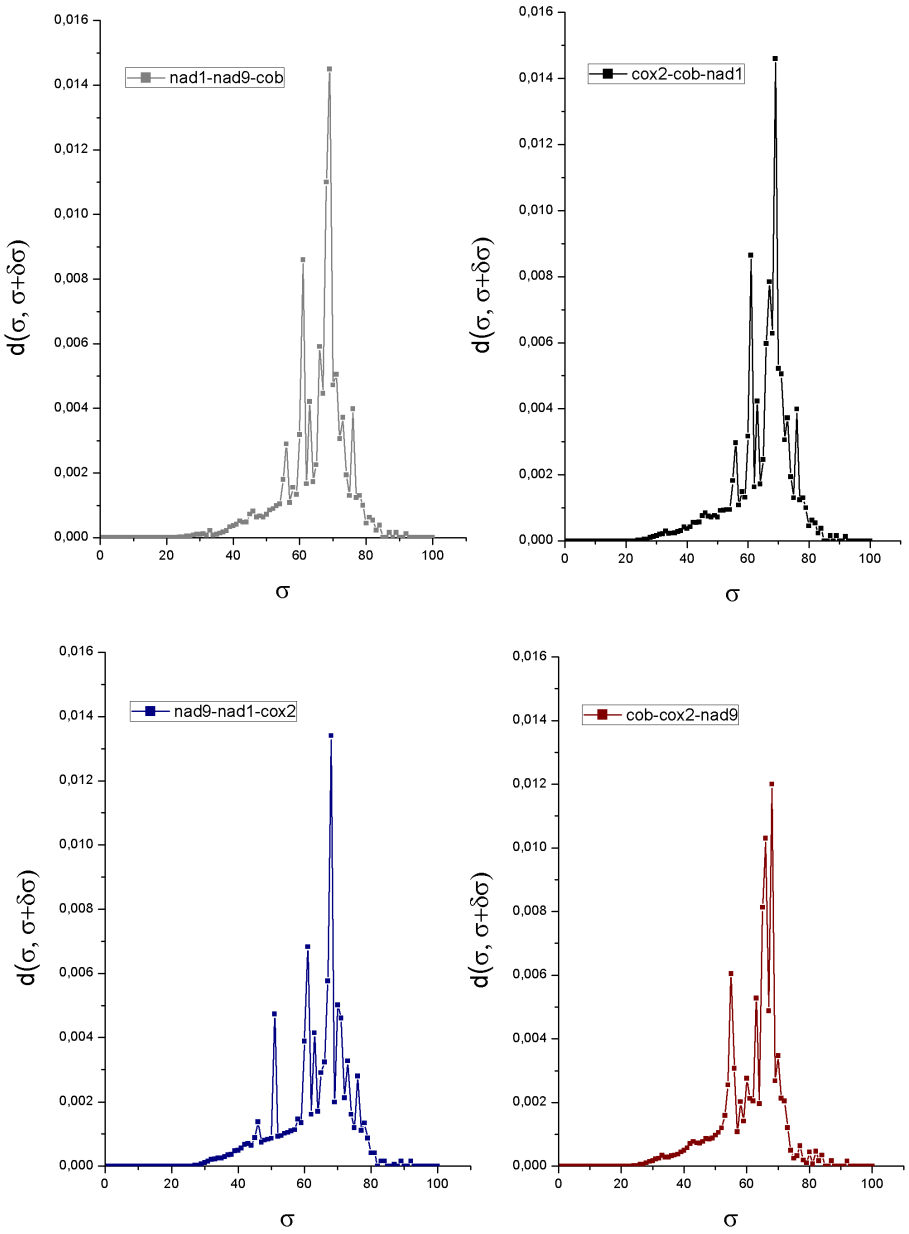


Figura 5.3: Gráficos dos picos calculado sobre multiplex de três camadas.

Isto requer que o organismo X esteja na posição 1 em todas as camadas. Vale salientar que cada rede tem uma topologia diferente.

Por isso é necessário uma comparação entre todos os organismos da 4 redes de proteínas para que todos os nós que tenham correspondentes fiquem nas mesmas posições em todas as camadas. Como os dados de cada proteína não são representados pelos mesmos organismos, é natural que o número total de organismos no multiplex seja maior que o número de organismos

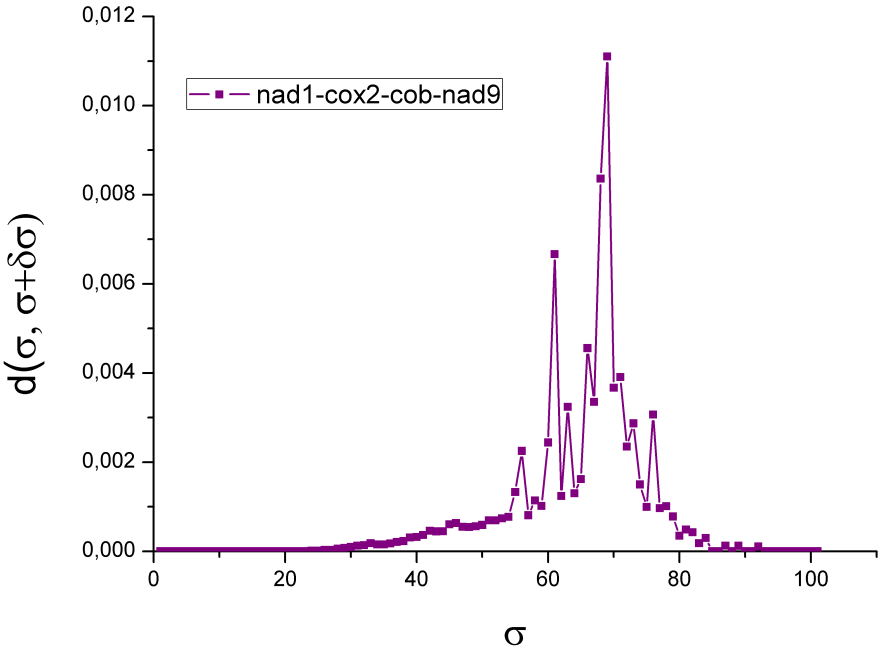


Figura 5.4: Gráficos dos picos calculado sobre multiplex de quatro camadas.

Tabela 5.5: Informações dos limiares de similaridade de cada proteína individualmente, utilizados nos métodos Louvain e Newman Girvan.

	Louvain	Newman-Girvan
Redes	Similaridades (%)	
Nad1	66.5	69
Nad9	53	53.5
Cob	58	64
Cox2	48	51

presentes nos dados de cada proteína. Para resolver este problema, em cada conjunto de dados de proteínas introduzimos os organismos representados em alguma das outras proteínas, de forma a que o multiplex possa contar com o mesmo número de nós em cada camada. Cada nó introduzido neste processo permanece isolado naquela camada, uma vez que não foi disponibilizada qualquer informação genética. No entanto, nas matrizes de similaridade, de adjacência e de vizinhança, são acrescentadas linhas e colunas com elementos identicamente nulos, que não interferem nos resultados feitos com uma única camada, mas que podem influenciar os resultados baseados na análise do multiplex. Após este procedimento, cada rede baseada em uma única proteína teve o seu numero original de vértices aumentado para 86, correspondente ao número de diferentes organismos entre as quatro proteínas. Por outro lado, as

Tabela 5.6: Informações dos limiares de similaridade para as proteínas agrupadas, utilizados nos métodos Louvain e Newman-Girvan.

	Louvain	Newman Girvan
Redes	Similaridades (%)	
Cob-Nad9	59	62
Cox2-Nad9	53	53
Cob-Cox2	59	62
Nad1-Cox2	65	67
Nad1-Cob	66	67
Nad9-Nad1	66	68
Nad1-Nad9-Cob	66	68
Cox2-Cob-Nad1	62	67
Cob-Cox2-Nad9	56	62
Nad9-Nad1-Cox2	65	68
Nad1-Nad9-Cob-Cox2	66	68

quatro bases apresentam 64 organismos em comum, número este que varia se comparamos as camadas duas a duas ou de três em três.

Por fim, com as matrizes de adjacências alinhadas e estabelecidas em termos dos limiares de similaridade, foram construídas as matrizes supra-adjacentes, que representam matematicamente os multiplex.

5.3 GenLouvain

Aplicando o GenLouvain nas 4 redes individuais, investigamos quais os melhores indices de similaridade para cada rede como mostra a Tab. 5.5, de modo a obter um conjunto de comunidades para cada uma das 4 redes. A Tab. 5.7 mostra as comunidades detectadas.

Por uma questão de simplicidade, usaremos uma nomenclatura para as comunidades C com a seguinte identificação: E contém organismos do grupo Eukarya, K do subgrupo *Rickettsiales* e H do subgrupo *Rhodospirillales*.

Tabela 5.7: Informação do número de organismos nas comunidades e seus respectivos grupos, nas redes individuais.

	Comunidade C_{EKH} e C_{EK}			Comunidade C_E		
Rede	Rickett	Rhodosp	Eukarya	Rickett	Rhodosp	Eukarya
Nad1	6	10	5	0	0	1
Cox2	5	8	6	0	0	0
Nad9	9	0	5	0	0	1
Cob	10	9	6	0	0	0

Nessa primeira análise, todos os organismos do grupo Eukarya para as proteínas Cox2 e Cob pertencem apenas uma comunidade C_{EKH}. Já para Nad1 e Nad9, estes organismos aparecem em duas comunidades distintas C_{EKH} e C_E, onde um dos eucariotos aparece isolado na comunidade C_E.

Resultados utilizando o Critério 1 de construção Foram construídos 11 multiplex fazendo todas as combinações entre as 4 redes, com seus limiares de similaridade estabelecidos individualmente. Assim, os multiplex apresentam um similaridade ótima mista, utilizando as similaridades ótimas das 4 redes de proteína, como demostra a Tab. 5.5.

Aplicando o GenLouvain nos 11 multiplex individualmente, foram obtidas as seguintes comunidades das quais os organismos Eukarya fazem partes, como é mostrado nas tabelas, 5.8, 5.9, 5.10.

Tabela 5.8: Informação do número de organismos nas comunidades e seus respectivos grupos, nos multiplex de duas camadas.

	Comunidade C_{EKH}			Comunidade C_E		
Rede	Rickett	Rhodosp	Eukarya	Rickett	Rhodosp	Eukarya
Nad1-Cob	11	13	6	0	0	0
Cox2-Nad1	7	12	6	0	0	0
Nad9-Nad1	6	12	2	8	0	3
Cob-Cox2	10	8	6	0	0	0
Cob-Nad9	13	19	6	0	0	0
Cox2-Nad9	5	8	6	0	0	0

* Na classificação Nad9-Nad1, além das comunidades já demostradas, um Eukarya único constitui uma comunidade C_E isolada

Tabela 5.9: Informação do número de organismos nas comunidades e seus respectivos grupos, nos multiplex de três camadas.

	Comunidade C_{EKH}			Comunidade C_{EK}		
Rede	Rickett	Rhodosp	Eukarya	Rickett	Rhodosp	Eukarya
Nad1-Nad9-Cob	11	12	6	0	0	0
Cox2-Cob-Nad1	13	14	6	0	0	0
Cob-Cox2-Nad9	11	9	6	0	0	0
Nad9-Nad1-Cox2	9	12	6	0	0	0

Tabela 5.10: Informação do número de organismos nas comunidades e seus respectivos grupos, no multiplex de quatro camadas.

	Comunidade C_{EKR}			Comunidade C		
Rede	Rickett	Rhodosp	Eukarya	Rickett	Rhodosp	Eukarya
Nad1-Cob-Cox2-Nad9	12	13	6	0	0	0

Resultados utilizando o Critério 2 de construção Para esse critério, também foram construídos 11 multiplex fazendo todas as combinações entre a 4 matrizes de adjacência. Contudo, aqui não foram considerados os valores dos limiares ótimos de similaridade de cada rede individual, como feito anteriormente. Os limiares de similaridade foram estabelecidos sobre cada multiplex a partir de sua supra-matriz de similaridade. Para cada um deles, foi investigado o melhor limiar de similaridade, demostrado na Tab. 5.6.

Operando o GenLouvain a partir desses 11 multiplex, obtemos os seguintes resultados, descritos nas tabelas 5.11, 5.12 e 5.13.

Tabela 5.11: Informação do número de organismos nas comunidades e seus respectivos grupos, nos multiplex de duas camadas.

	Comunidade C_{EKR} e C_{EK}			Comunidade C_E		
Rede	Rickett	Rhodosp	Eukarya	Rickett	Rhodosp	Eukarya
Nad1-Cob	11	12	6	0	0	0
Cox2-Nad1	6	13	3	0	0	2
Nad9-Nad1	5	12	6	0	0	0
Cob-Cox2	10	9	6	0	0	0
Cob-Nad9	11	7	6	0	0	0
Cox2-Nad9	9	0	6	0	0	0

* Na classificação Cox2-Nad1, além das comunidades já demostradas, um Eukarya único constitui uma comunidade C_E isolada

Tabela 5.12: Informação do número de organismos nas comunidades e seus respectivos grupos, nos multiplex de três camadas.

	Comunidade C_{EKR}			Comunidade C_E		
Rede	Rickett	Rhodosp	Eukarya	Rickett	Rhodosp	Eukarya
Nad1-Nad9-Cob	5	13	5	0	0	1
Cox2-Cob-Nad1	13	13	6	0	0	0
Cob-Cox2-Nad9	13	9	6	0	0	0
Nad9-Nad1-Cox2	5	11	5	0	0	1

Tabela 5.13: Informação do número de organismos nas comunidades e seus respectivos grupos, no multiplex de quatro camadas.

	Comunidade C_{EKR}			Comunidade C_E		
Rede	Rickett	Rhodosp	Eukarya	Rickett	Rhodosp	Eukarya
Nad1-Cob-Cox2-Nad9	5	13	5	0	0	1

Numa análise global, observamos a consistência dos resultados para os dois critérios estabelecidos. Entretanto, ao estabelecer o limiar de similaridade no multiplex (critério 2), perdemos parte das informações subjacentes de identidade entre pares de sequências de uma mesma proteína. Isto acontece, pelo fato das proteínas terem níveis de relações de identidade distintos entre as suas sequências. Ao contrário, o limiar ótimo misto (critério 1) preserva mais as relações de identidade entre os pares de sequências de uma mesma proteína.

Ao comparamos as tabelas 5.10 e 5.13, observamos que o número de organismos representantes no subgrupo Rickett diminui de 12 para 5, o que reflete maoir grau de rigidez no agrupamento, quando utilizamos o critério dois. Entretanto, mesmo que o número total de organismos difere, os organismos presentes nas comunidades descritas são os mesmo, todos os organismos do critério 2 estão no critério 1. A Tab. 5.14 e a Fig. 5.5 mostra essa equivalência.

Tabela 5.14: Comparação da detecção de comunidade entre os multiplex Nad1-Cob-Cox2-Nad9, sob os critérios 1 e 2 (A correspondência vale para todos os elementos das quatro camadas).

(a) Classificação com critério 1

Comunidades	Vértices
C_{EKR}	1, 2, 3, 4, 5, 6, 12, 13, 14, 15, 16, 46, 47, 48, 49, 50, 55, 56, 61, 62, 63, 64, 70, 72, 80, 81, 82, 83, 85, 86

(b) Classificação com critério 2

Comunidades	Vértices
C_{EKR}	2, 3, 4, 5, 6, 46, 47, 48, 49, 50, 55, 56, 61, 62, 63, 64, 80, 81, 82, 83, 85, 86
C_E	1

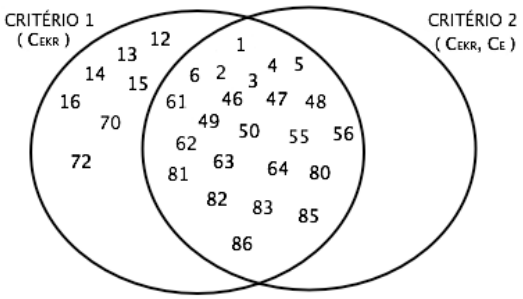


Figura 5.5: Diagrama de Venn: Comparação entre Critério 1 e Critério 2 demostrados na Tab. 5.14.

5.4 MultiNG

Nesta seção, iremos fazer as mesmas análises que foram realizadas com o GenLouvain, agora utilizando o nosso algoritmo MultiNG. Não apresentaremos todos os resultados obtidos, apenas aqueles mais interessantes, descritos durante o decorrer do texto.

Executando o MultiNG nas 4 redes individualmente, investigamos os melhores limiares de similaridade para análise. As figuras 5.6, 5.7, 5.8 e 5.9, mostram os dendrogramas gerados a partir de uma rede "ótima", ou seja, no melhor valor de similaridade, descrito na Tab. 5.5. Mesmo com certo grau de arbitrariedade, a identificação das comunidades contendo organismos Eukarya é bastante clara e consistente.

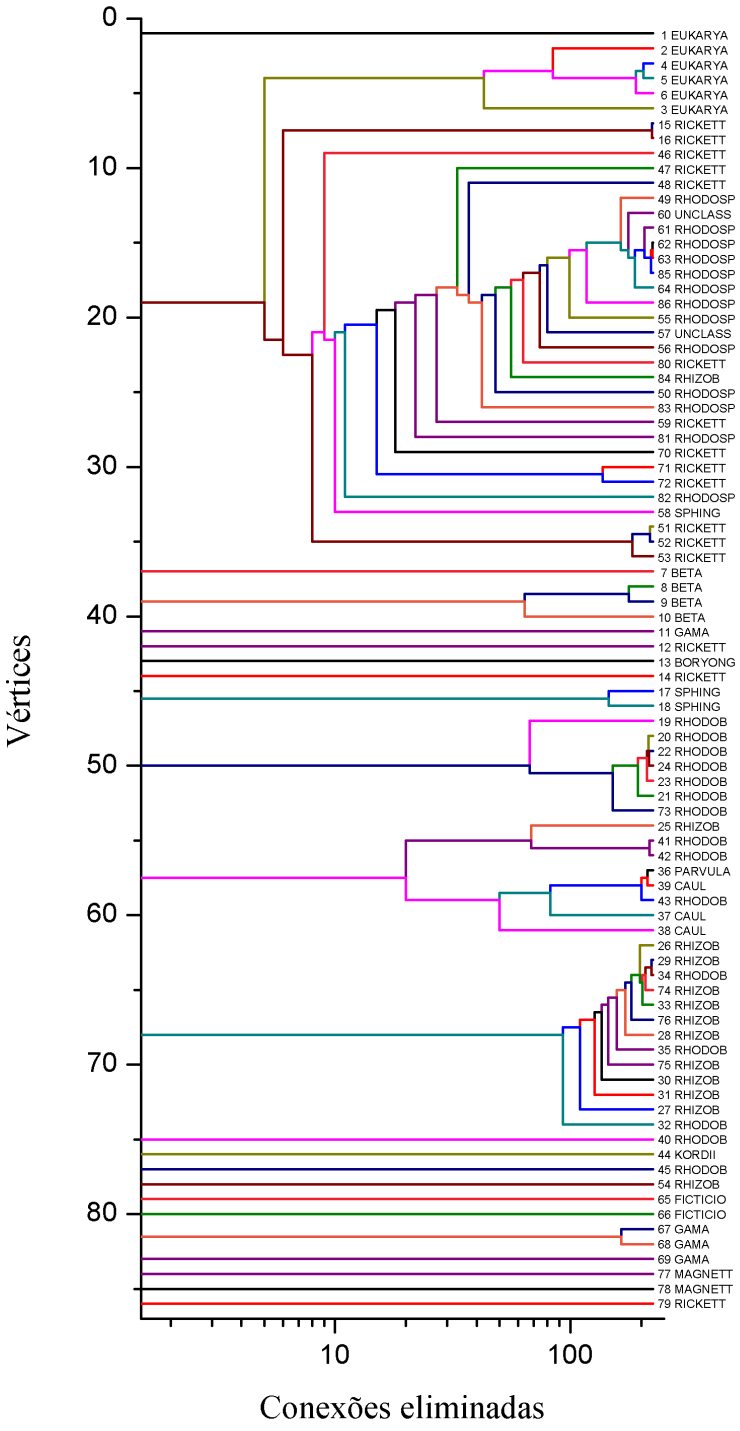


Figura 5.6: Dendrograma da rede Nad1, com valor ótimo de similaridade de 69%.

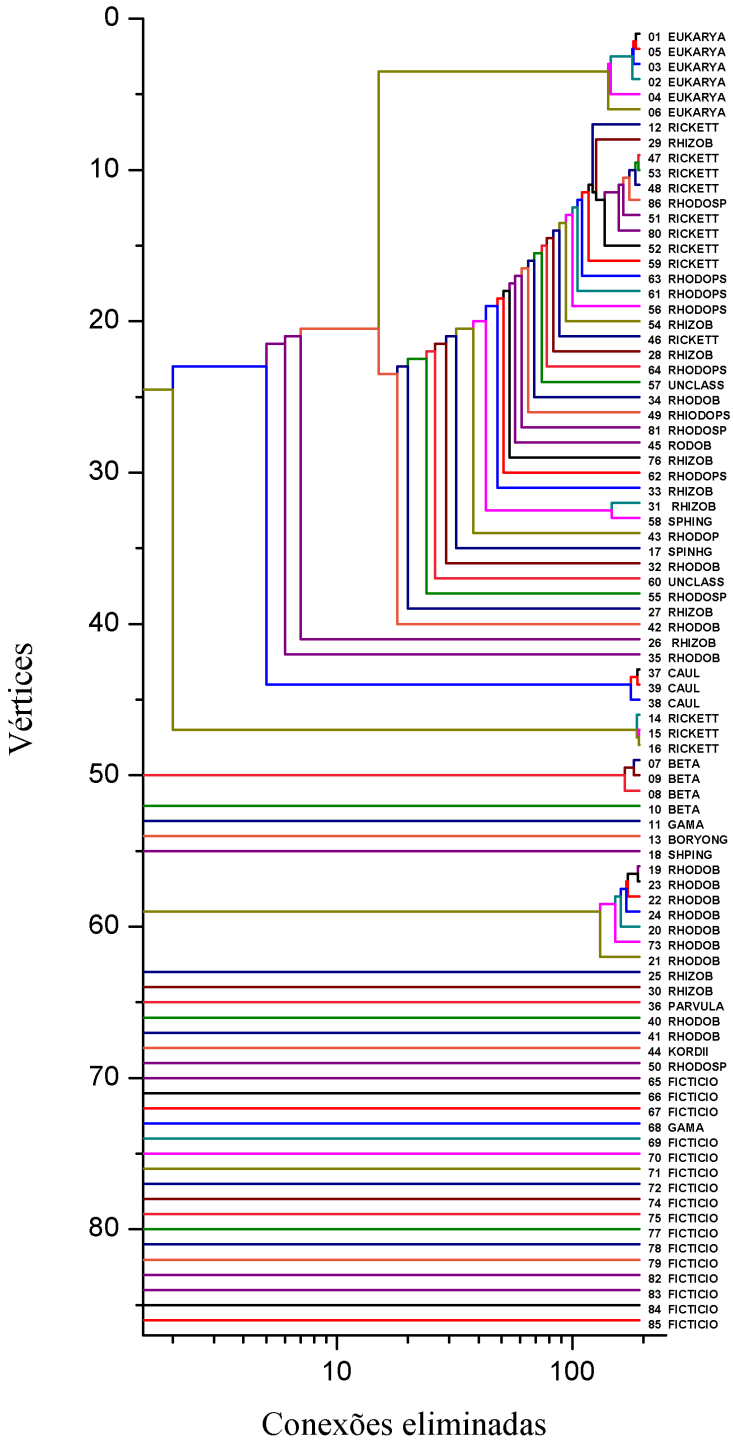


Figura 5.7: Dendrograma da rede Cox2, com valor ótimo de similaridade de 51%.

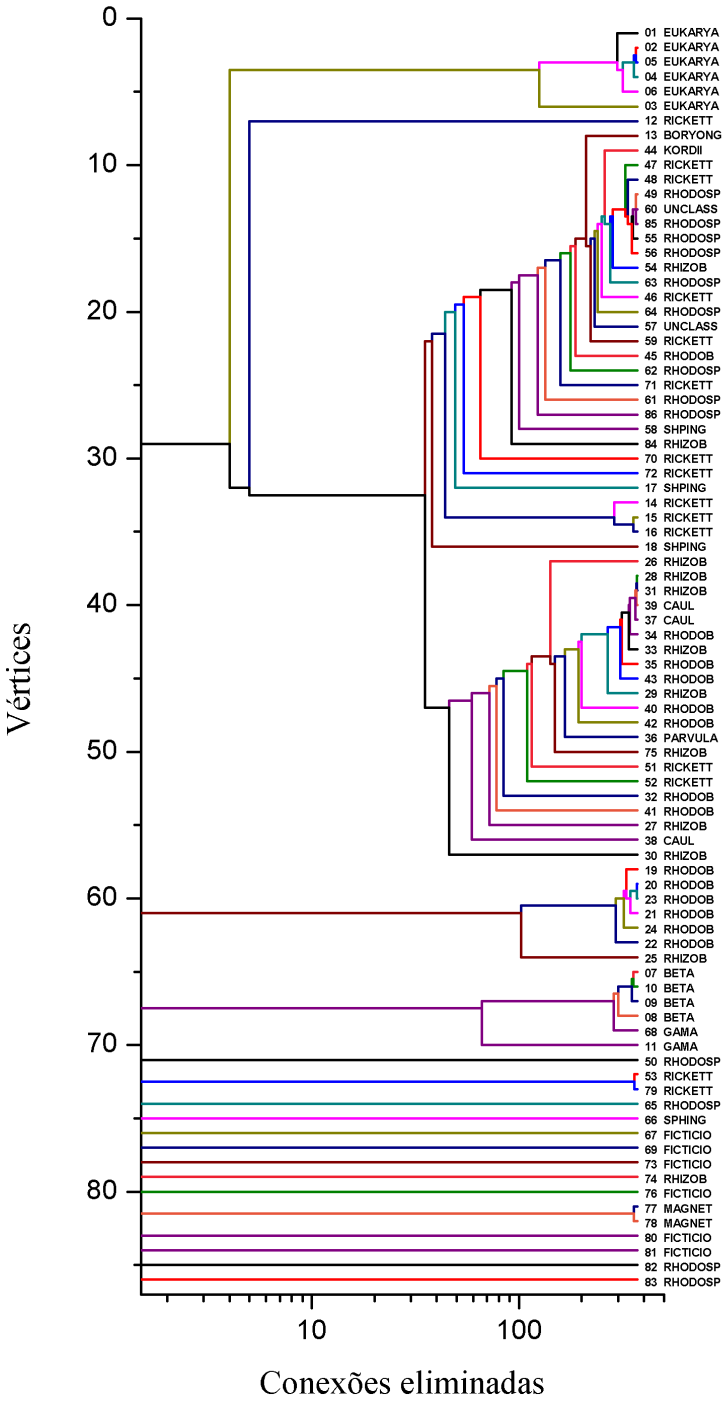


Figura 5.8: Dendrograma da rede Cob, com valor ótimo de similaridade de 64%.

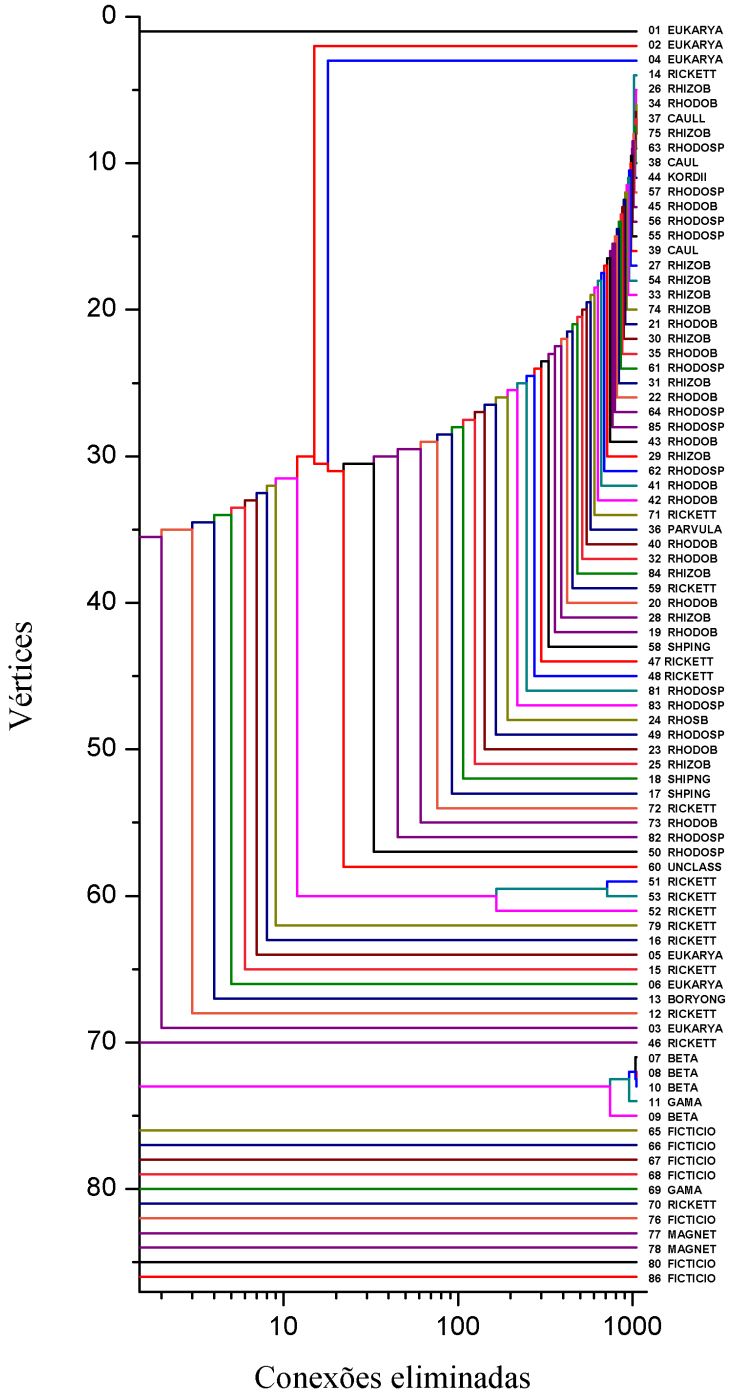


Figura 5.9: Dendrograma da rede Nad9, com valor ótimo de similaridade de 53.5%.

Resultados utilizando o Critério 1 de construção Foram construídos 11 multiplex fazendo todas as combinações entre a 4 matrizes de adjacência com os limiares de similaridades estabelecidos individualmente como mostra a Tab. 5.5, ou seja, os multiplex não apresentam um limiar de similaridade

bem estabelecido. As figuras 5.10, 5.11 e 5.12 mostram alguns dos resultados obtidos.

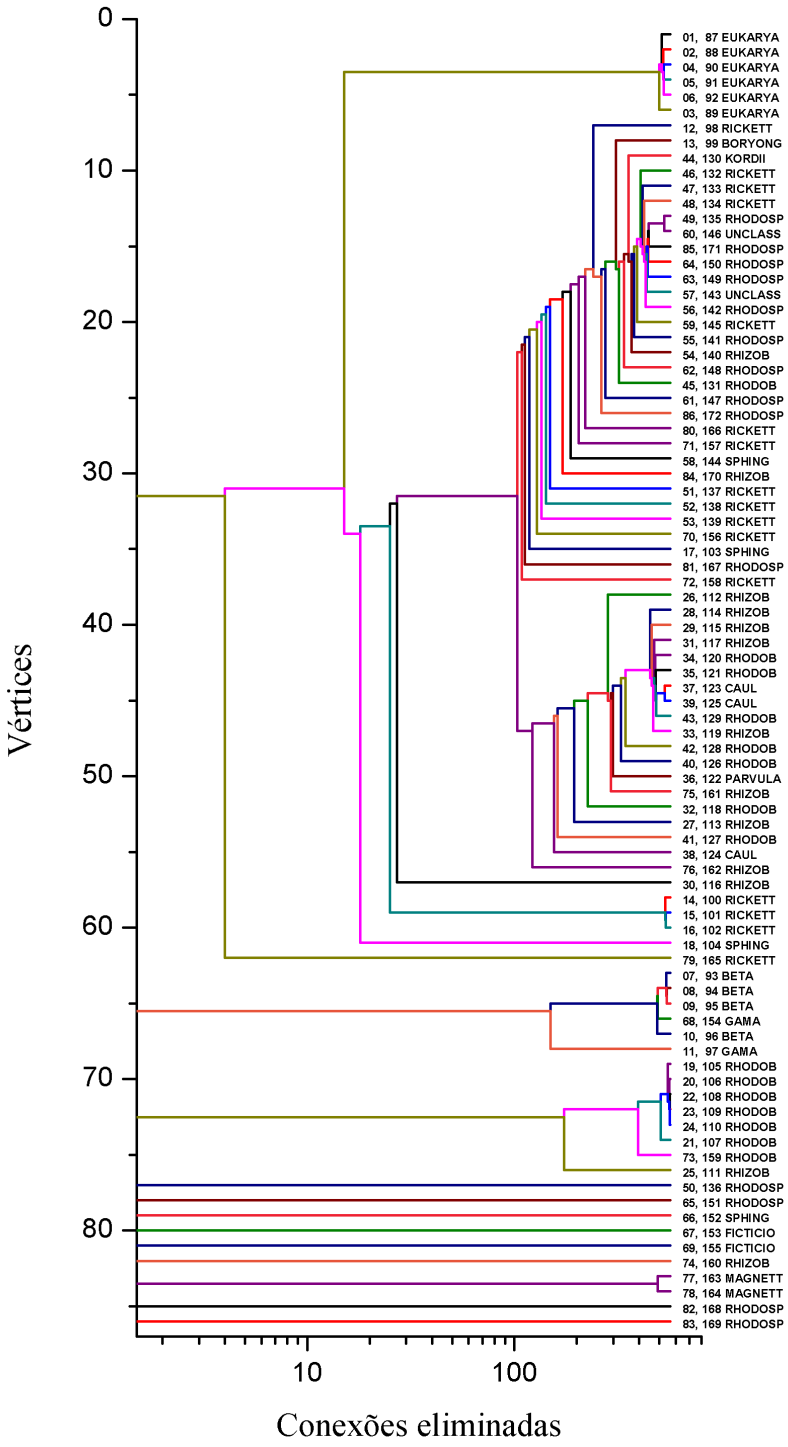


Figura 5.10: Dendrograma do multiplex composto por Cob e Cox2, com valor ótimo de similaridade mista. (critério 1)

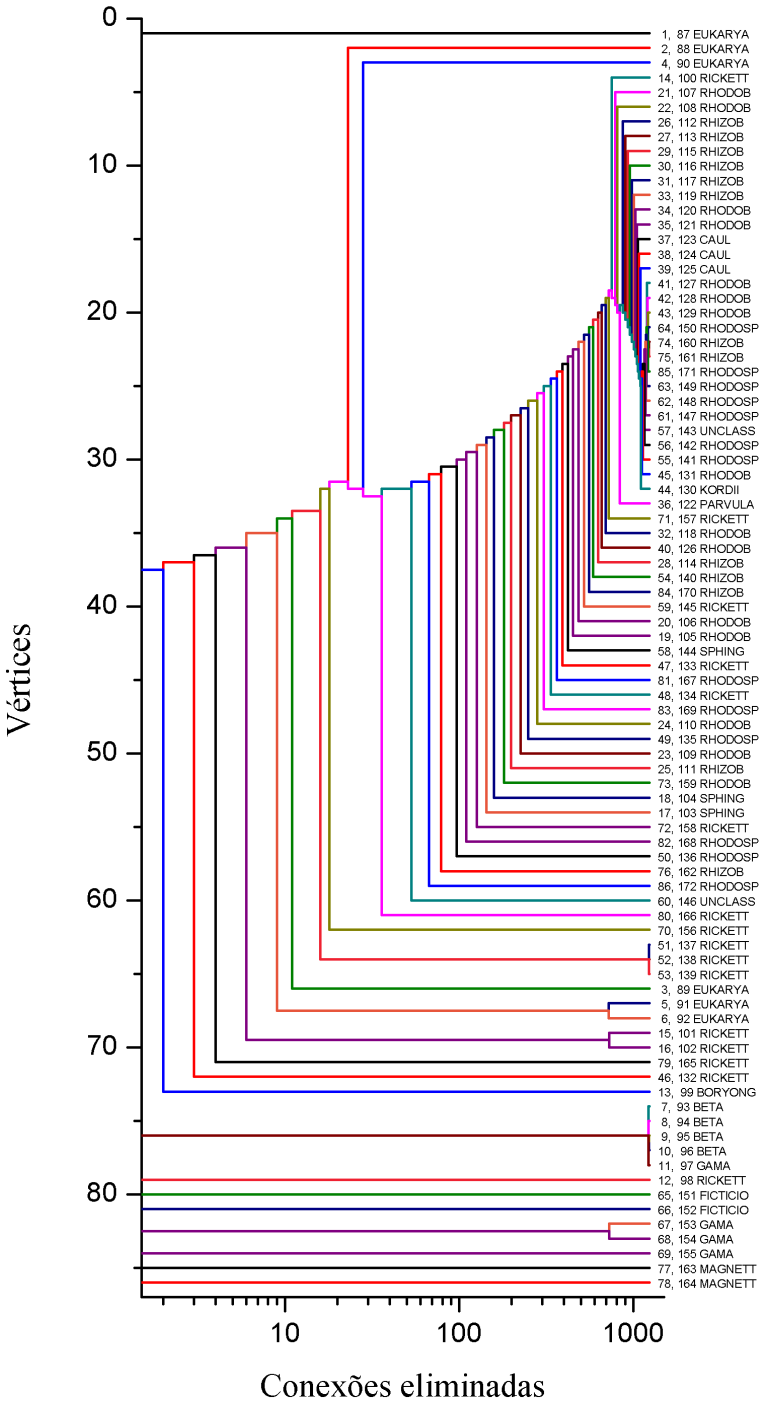


Figura 5.11: Dendrograma do multiplex composto por Nad1 e Nad9, com valor ótimo de similaridade mista. (critério 1)

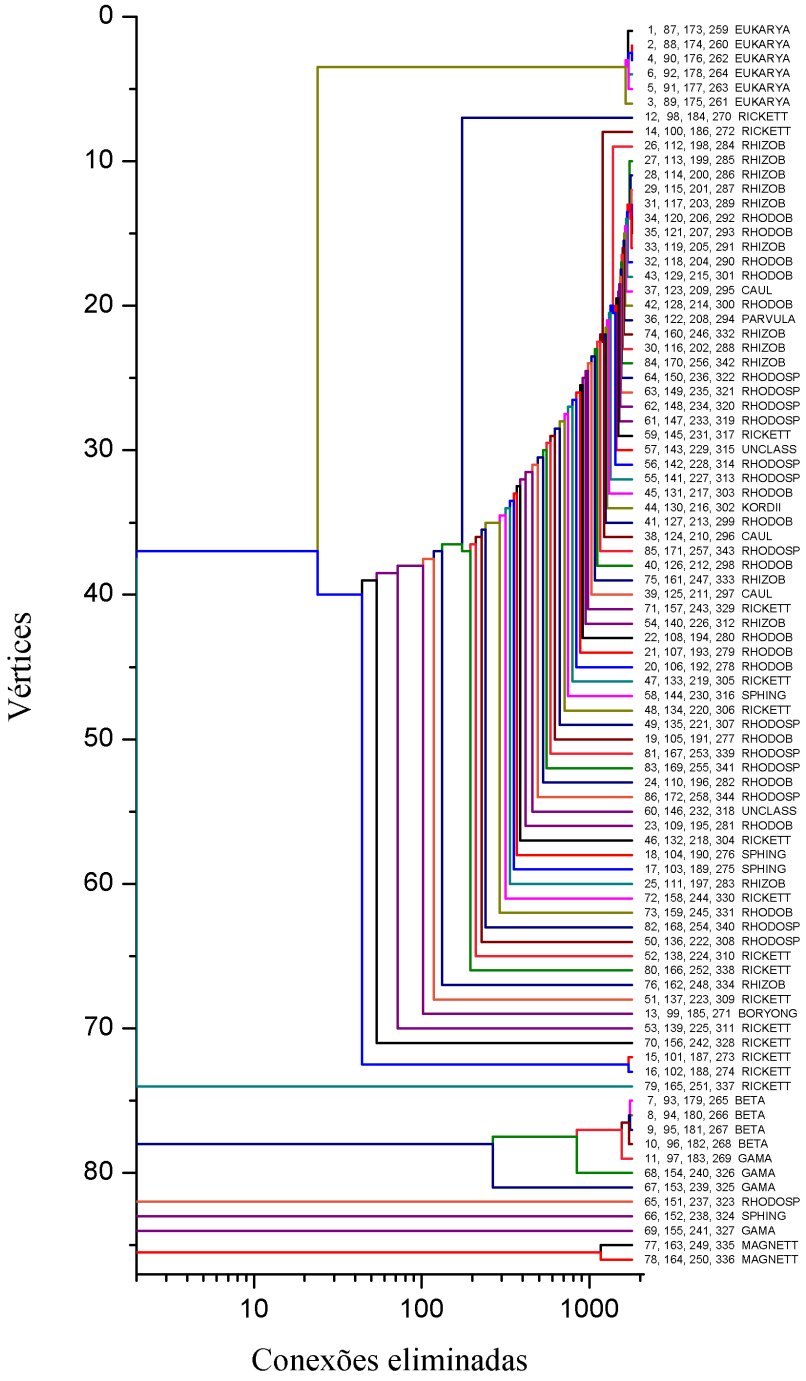


Figura 5.12: Dendrograma do multiplex composto por Nad1, Cob, Cox2 e Nad9, com valor ótimo de similaridade mista. (critério 1)

Resultados utilizando o Critério 2 de construção Analogamente, foram construidos 11 multiplex fazendo todas as combinações entre a 4 matrizes de adjacência, seguindo o procedimento descrito na discussão sobre a aplicação do método GenLouvain, e usando os valores indicados na tabela 5.6. A seguir,

apresentamos alguns dos resultados obtidos.

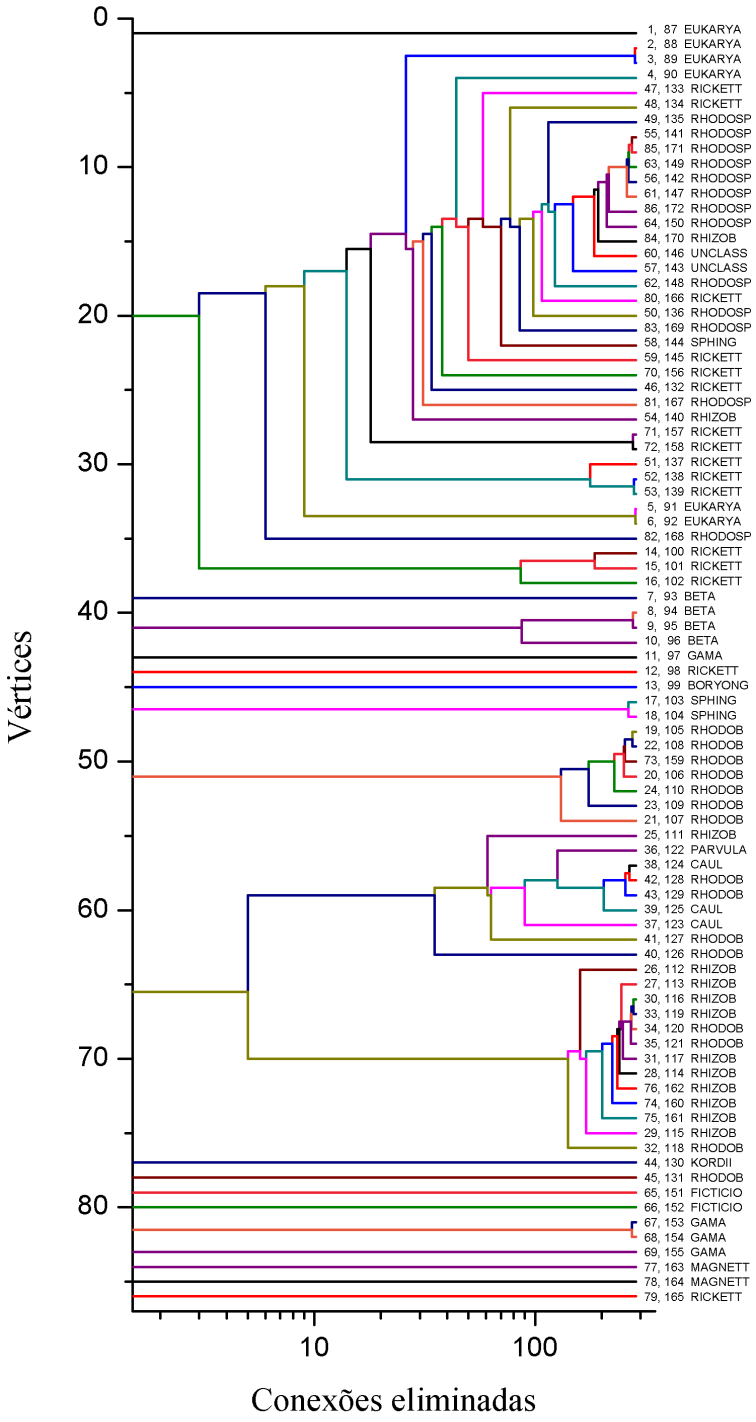


Figura 5.13: Dendrograma do multiplex composto por Cox2 e Nad1, com valor ótimo de similaridade de 67%. (critério 2)

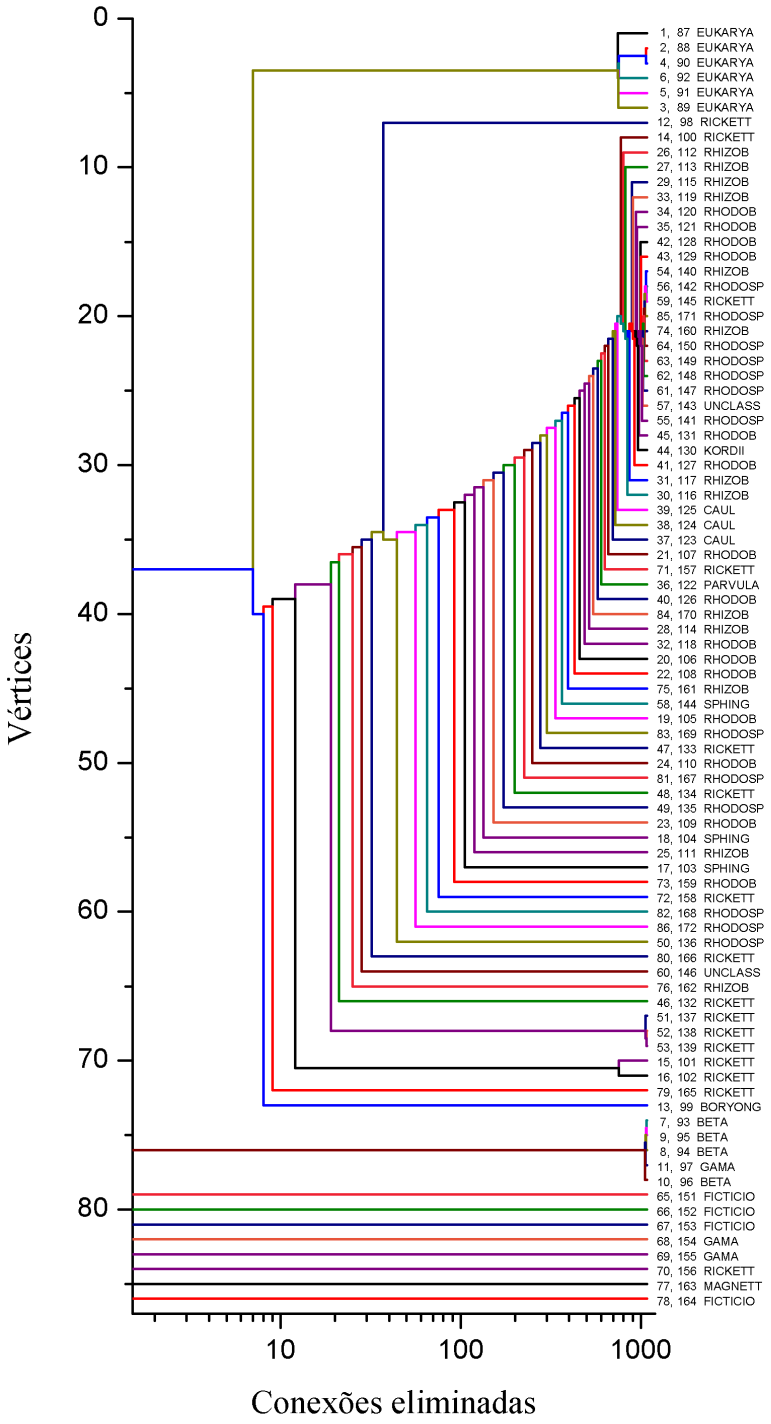


Figura 5.14: Dendrograma do multiplex composto por Cox2 e Nad9, com valor ótimo de similaridade de 53%. (critério 2)

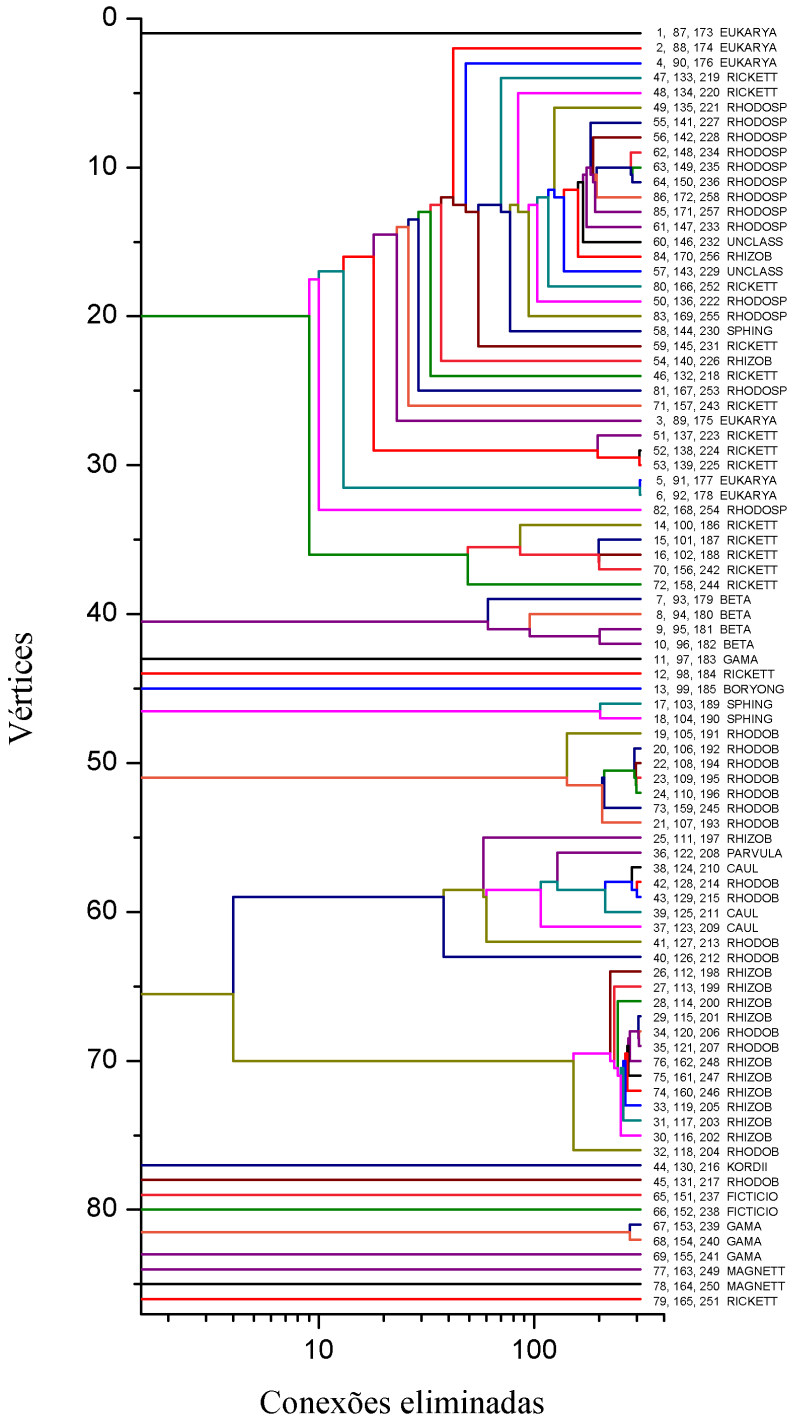


Figura 5.15: Dendrograma do multiplex composto por Nad9, Nad1 e Cox2, com valor ótimo de similaridade de 67%. (critério 2)

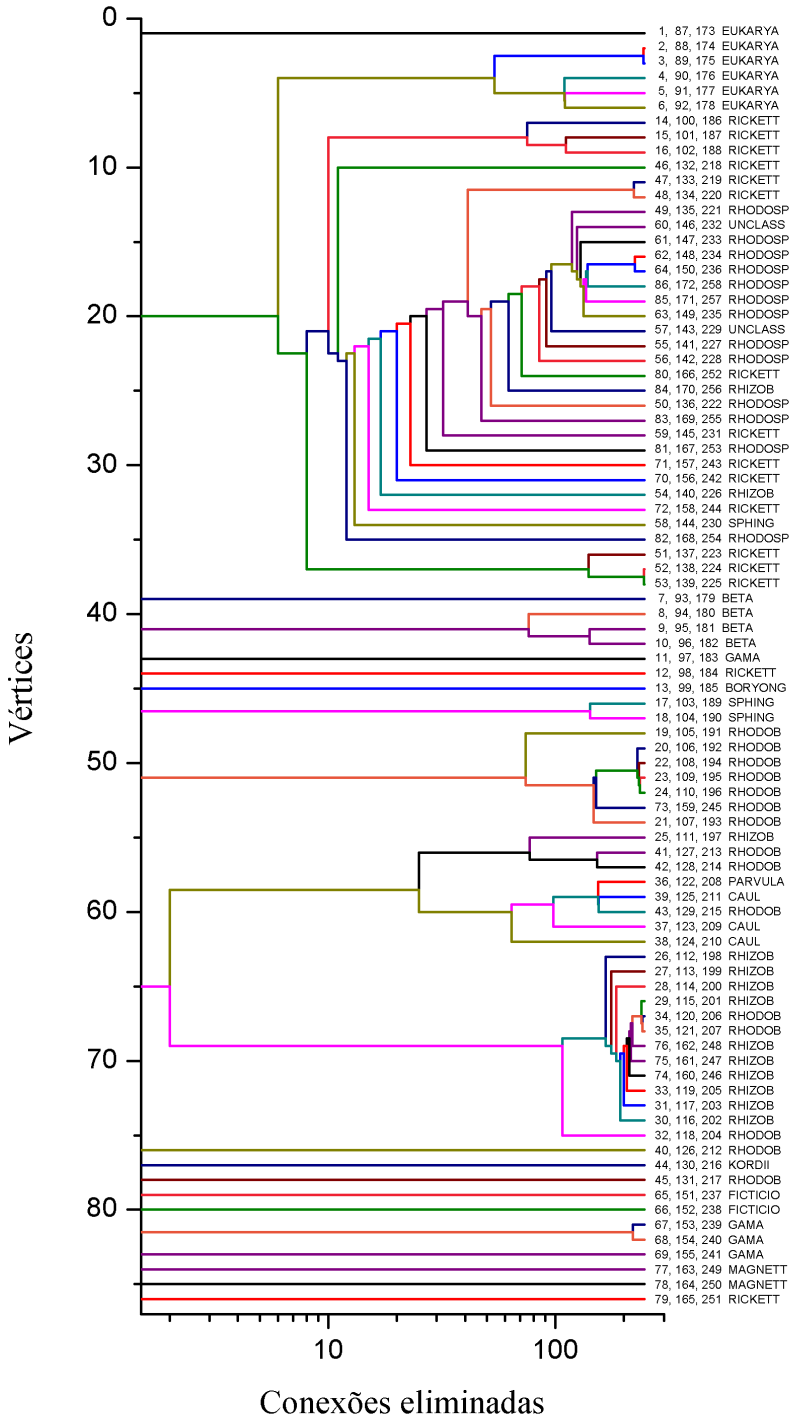


Figura 5.16: Dendrograma do multiplex composto por Nad9, Nad1 e Cox2, com valor ótimo de similaridade de 68%. (critério 2)

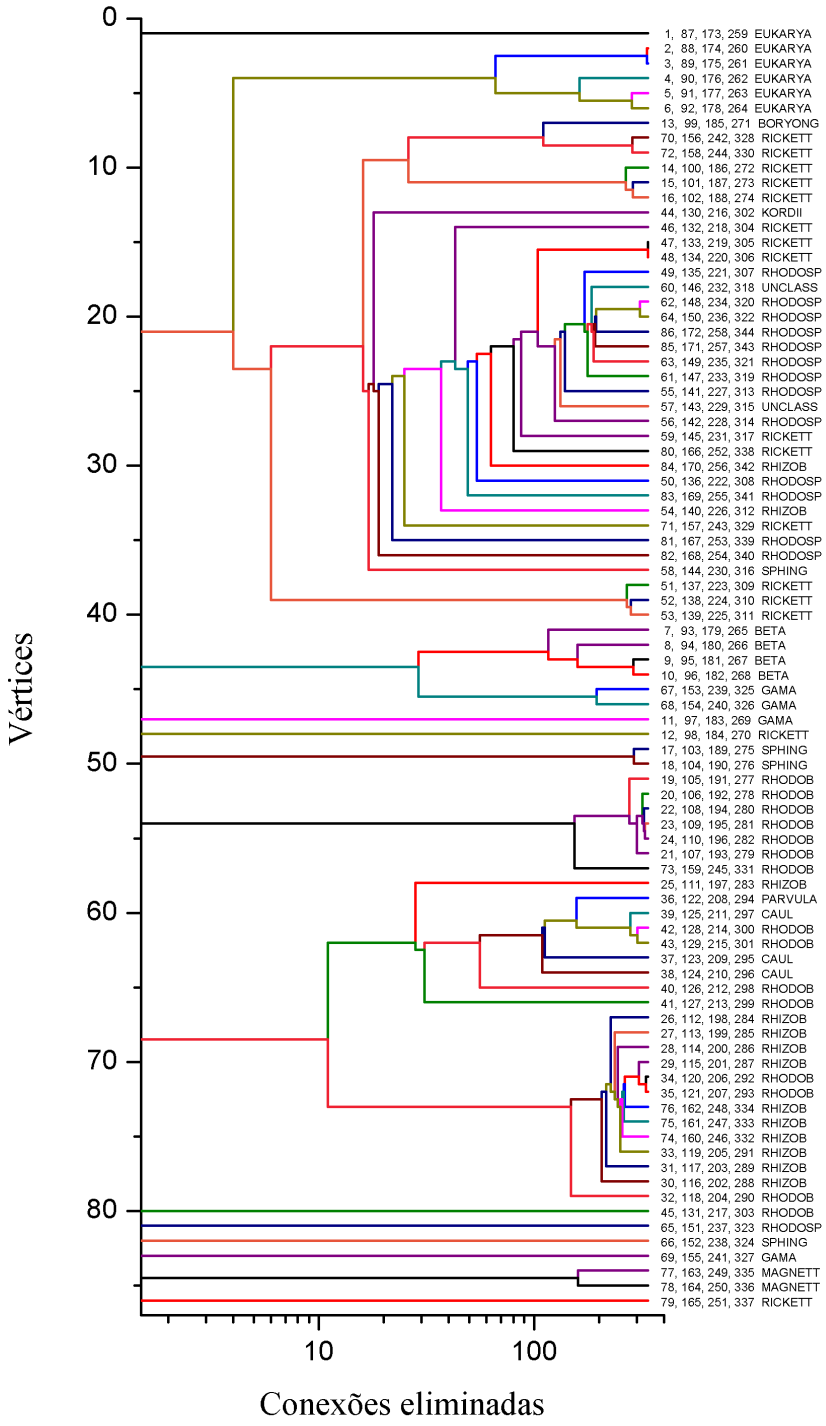


Figura 5.17: Dendrograma do multiplex composto por Nad1, Cob, Cox2 e Nad9, com valor ótimo de similaridade de 68%. (critério 2)

Comparamos também os resultados obtidos com o MultiNG, para os dois critérios. Observamos um maior grade liberdade para o agrupamento, quando o algoritmo MulitNG opera sobre o multiplex estabelecidos a partir do limiar ótimo misto (critério 1), agrupando um maior número de organismos

nas comunidades, motivo já discutido anteriormente. Em geral, os organismos agrupados nas comunidades são sempre os mesmos, independentemente do critério usado.

5.5 Comparação do resultados

Observa-se, mesmo com os limiares de similaridade específicos para cada análise, de forma geral a convergência dos resultados obtidos com os algoritmos GenLouvain e MulitNG.

Comparando os resultados apresentados para as redes individuais, indicados na Tab. 5.7 e nas figuras 5.6, 5.7, 5.8 e 5.9, nota-se que os mesmos se mostram satisfatórios. Dentre estes resultados, os obtidos com a proteína Nad9 apresenta um resultado bastante peculiar: o GenLouvain detecta 2 comunidades nas quais os eucariotos fazem parte, sendo uma dela C_{EK} composta de 5 organismos do domínio Eukarya e 9 *Rickettsiales* e outra C_E composta por um único organismo do domínio Eukarya. Ao comparamos com o resultado do MultiNG, Fig. 5.9, um dos organismos Eukarya já vem isolado dos demais organismos, ramo 1, sendo equivalente à C_E. Se rompermos um ramo em específico, como demostrado na Fig. 5.23, obtemos o resultado equivalente a C_{EK}.

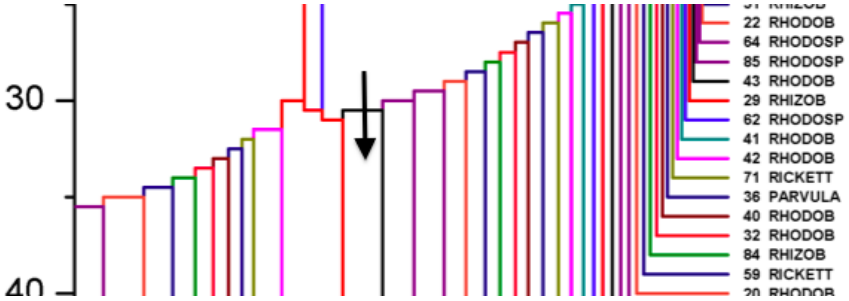


Figura 5.18: Seta indica o corte no dendrograma da rede Nad9.

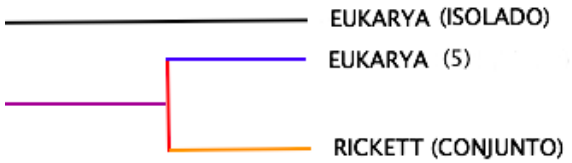


Figura 5.19: Síntese do resultado depois do corte no dendrograma Nad 9.

A comparação dos resultados para os multiplex também mostra que eles são bastante equivalentes. Dentro do critério 1 de construção, vamos analisar

o multiplex composto por Nad1 e Nad9. Os resultados estão descritos na Tab. 5.8 e na Fig. 5.11.

O algoritmo GenLouvain detectou 3 comunidades, que incluem os eucariotos. Temos a comunidade C_{EKH}, composta de 2 organismos do grupo Eukarya, 6 *Rickettsiales* e 12 *Rhodospirillales*, a C_{EK}, composta de 3 organismos do grupo Eukarya, 8 *Rickettsiales* e, por fim a comunidade C_E, composta por um único eucarioto. Ao compararmos com o resultado gerado no MultiNG, 5.11, obtemos as respectivas comunidades ao fazer um corte como mostra a imagem abaixo.

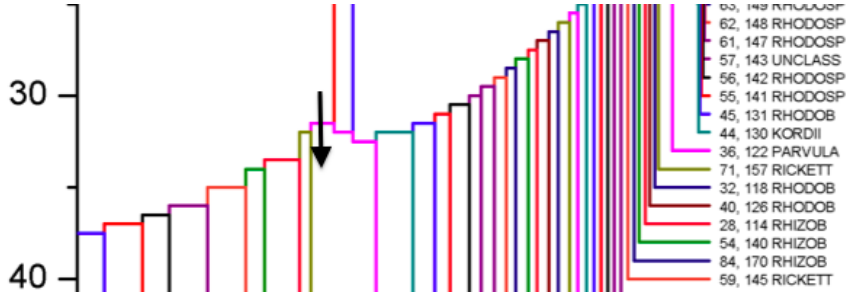


Figura 5.20: Seta indica o corte no dendrograma do multiplex Nad1-Nad9.

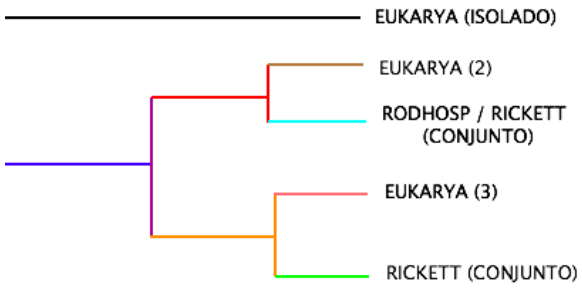


Figura 5.21: Síntese do resultado depois do corte no dendrograma Nad1-Nad9.

A comunidade C_{EKH} corresponde ao grupo de organismos que permanecem conectados após o corte, ao passo que os organismos da comunidade C_{EK} foram separados antes do corte. Se percebe a respectiva da comunidade C_E, que corresponde ao primeiro ramo do dendrograma, já se encontra isolado antes de começar o processo de eliminação de conexões.

Vamos comparar agora o multiplex Cox2-Nad9, dentro do critério 2 de construção. O GenLouvain, detecta uma comunidade C_{EK} que contém todos os 6 eucariotos e 9 organismos *Rickettsiales*. Ao fazermos um corte especifico no dendrograma Fig. 5.14, como demostrado na Fig. 5.22, observa-se que os nós da comunidade C_{EK} obtida pelo programa GenLouvain foram separados antes do corte.

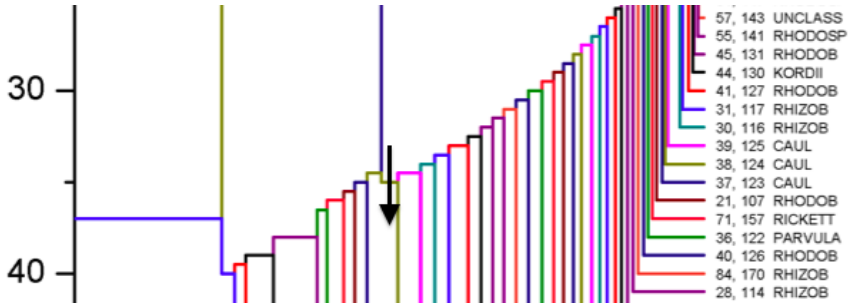


Figura 5.22: Seta indica o corte no dendrograma do multiplex Nad9-Cox2.

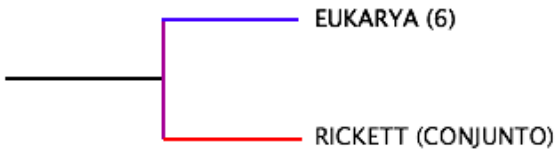


Figura 5.23: Síntese do resultado depois do corte no dendrograma Nad9-Cox2.

Investigando nossos resultados que foram obtidos com o maior número de informações, o multiplex composto pelas 4 redes (Nad1, Cob, Cox2 e Nad9), percebe-se também uma convergência nos resultados. Vamos comparar os resultados da Tab. 5.13 com o dendrograma da Fig. 5.17.

As comunidades detectadas pelo algoritmo Genlouvain, C_{EKH} e C_E, estão claras no dendrograma. Abaixo segue a ampliação dos ramos que formam as respectivas comunidades citadas.

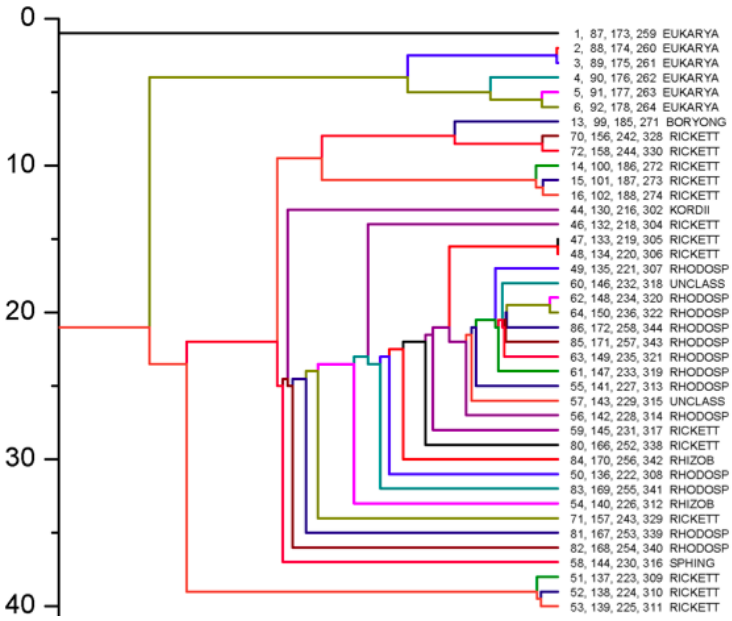


Figura 5.24: Ampliação do dendrograma Nad1-Cob-Cox2-Nad9. (Critério 2)

Em síntese, em uma comparação geral é possível verificar que as comunidades obtidas, tanto pelo GenLouvain como pelo MultiNG, são compostas sempre pelos mesmos grupos de organismos. Em geral, os eucariotos formam comunidades com os organismos *Rickettsiales* e *Rhodospirillales*, os subgrupos *Rhizobiales* e *Rhodobacterales* formam uma comunidade e o grupo externo Gamaproteobactérias e Betaproteobactérias compõem outra comunidade. Vale salientar que os organismos agrupados são sempre os mesmos em relação à posição na rede para os dois métodos.

Diante do apresentado, percebe-se uma equivalência dos resultados gerados a partir dos dois algoritmos utilizados. Portanto, a generalização do método NG implementado no programa MultiNG se mostra bastante eficaz e promissora para análises de comunidades em redes do tipo multiplex.

Para estudos de evolução biologia, o MultiNG mostra ser capaz de gerar bons resultados, utilizando um grande número de dados simultaneamente. Vimos que, a partir do dendrograma, ele leva a uma descrição do sistema evolutivo de forma consistente, que viabiliza uma melhor compreensão e facilita a análise para esse tipo de questão.

5.6 Análise biológica

Na literatura, atualmente, persiste o debate sobre qual grupo das Alfaproteobactérias é o ancestral mais próximo das mitocôdrias. Embora a maioria dos estudos apóie a ideia de que as mitocôndrias evoluíram de um ancestral relacionado a ordem *Rickettsiales*, alguns trabalhos sugerem que as mitocôndrias evoluíram a partir de um grupo de vida livre relacionado à família *Rhodospirillaceae*. Portanto, o presente estudo focalizou a investigação sobre esses dois subgrupos de organismos e sua relação como os eucariotos representados.

Os resultados do GenLouvain, revelam que, de todos subgrupos dentre Alfaproteobactérias, os mais próximos dos eucariotos, e também entre si, são *Rickettsiales* e *Rhodospirillaceae*. Diante da heurística do método, a maximização da modularidade, ocorreu sempre com o agrupamento dos organimos pertencentes a esses dois subgrupos. Porém, nos diferentes limiares analisados o método não deixou clara a separação entre *Rickettsiales*, *Rhodospirillaceae* e eucariotos, não é detectada a posição relativa dos mesmos nas comunidades detectadas. No entanto, o uso do MultiNG, possibilitou um refinamento dos resultados permitindo separar eucariotos, *Rickettsiales* e *Rhodospirillaceae*, bem como estabelecer suas posições relativas nos dendrograma gerados.

A medida que aumentamos a quantidade de informação genética no multiplex, verificamos que as sequências proteicas mitocondriais foram separadas das sequências proteicas das Alfaproteobactéria antes da diversificação das suas linhagens atualmente conhecidas, conforme claramente descrito nos dendrogramas, figuras 5.12 e 5.17. As figuras 5.25 e 5.26 sintetizam esses resultados, a partir dos critérios 1 e 2 respectivamente.

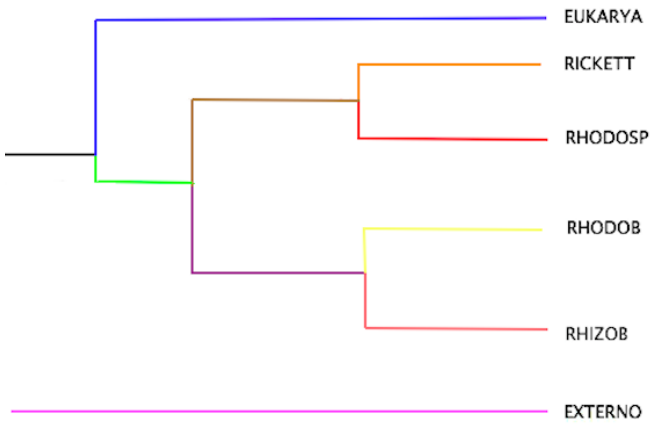


Figura 5.25: Síntese da posição relativa dos grupos dentro do dendrograma 5.12, com critério 1.

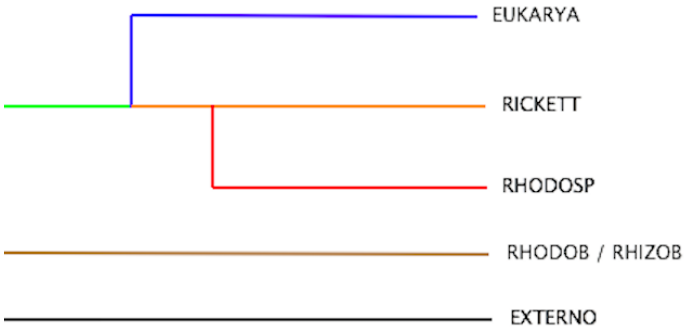


Figura 5.26: Síntese da posição relativa dos grupos dentro do dendrograma 5.17, com critério 2.

Os resultados mostraram que a construção de multiplex, a partir do percentual de identidade entre sequências proteicas de diferentes organismos, permite recuperar informações úteis na inferência de relações filogenéticas que são coerentes com aquelas propostas na literatura a partir da utilização de métodos convencionais, como já demostrado em outros trabalhos [18, 41] para redes monocamadas. Além disso, consideramos que o método aqui proposto

confere maior robustez aos resultados ao definir as relações entre os organismos a partir da análise simultânea de mais de uma proteína.

Nossos resultados podem ser comparados diretamente com outras árvores filogenéticas, que também sugerem uma posição relativa para um ancestral das mitocôndrias evidenciado pela Fig. 5.27. A comparação indica que os mesmos sustentam a hipótese levantada Martijn et al. [20], em trabalho publicado recentemente, segundo a qual as mitocôndrias evoluíram a partir de uma linhagem proteobacteriana que se ramificou antes da divergência de todas as Alfaproteobactérias.

Esta hipótese está em contraste direto com todas as hipóteses anteriores, que sugerem um ancestral para as mitocôndrias dentro das Alfaproteobactérias [13–19], e indica que, possivelmente, o evento de endossimbiose que deu origem às mitocôndrias é mais antigo do que se presume.

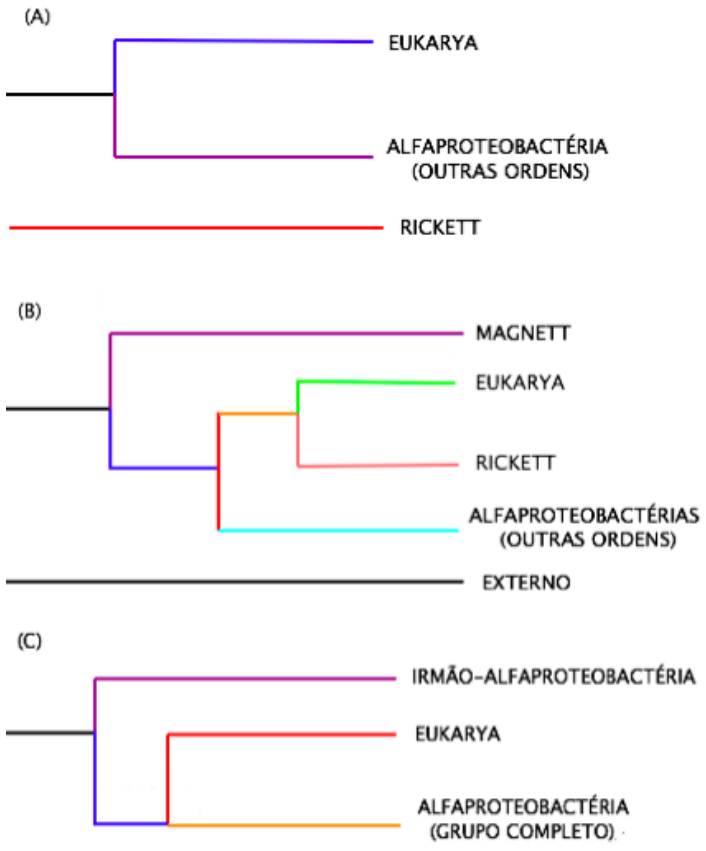


Figura 5.27: Síntese da posição relativa dos grupos dentro das árvores dos artigos Refs. [18–20]. (A) Síntese da árvore construída por Carvalho et al. [18] (B) Síntese da árvore construída por Wang e Wu [19] (C) Síntese da árvore construída por Martijn et al. [20].

6
Conclusão

As redes multicamadas são um modelo emergente que tem sido ultimamente proposto para lidar com a complexidade dos sistemas do mundo real. Essa abordagem vem exigindo a redefinição e ampliação de conceitos existentes para a análise de redes complexas, bem como a adaptação da maioria dos algoritmos até então desenvolvidos. Nesta dissertação abordamos o problema da detecção de comunidades, aplicado ao cenário de redes multiplex, propondo algumas generalizações que inclui algoritmo MultiNG.

Implementamos e validamos nosso algoritmo MultiNG, utilizando para isto a comparação de nossos resultados com os obtidos com o algoritmo GenLouvain. Neste processo consideramos inicialmente uma rede padrão (o clube de karatê de Zachary), e em seguida consideramos um conjunto de dados muito mais complexos, ligados a sequências proteicas de um conjunto de organismos.

As proteínas em questão são representadas por genes mitocondriais, e o problema que foi tratado pelo nosso método leva a resultados para um problema relevante dentro das Ciências Biológicas: a identificação do grupo de organismos procariotos que deram origem às organelas em células de organismos eucariotos, como as próprias mitocôndrias.

Esta generalização incorporou outros protocolos desenvolvidos anteriormente, que estão relacionados com a obtenção de uma classificação filogenética a partir de redes complexas baseadas em uma matriz de similaridade entre as proteínas dos diversos organismos. A análise é feita sobre redes obtidas para valor de um limiar de similaridade ótimo entre os organismos a partir da medida de distância entre redes. Este procedimento, desenvolvido anteriormente para uma rede monocamada, foi também generalizado neste trabalho para redes multiplex.

Desta forma concluímos que o método proposto constitui um avanço metodológico para a ciência de redes, no campo de estudos de estruturas modulares para redes multicamadas e bastante eficaz e promissor para análises de inferência filogenética.

Referências Bibliográficas

[1] A. P. Ryle, F. Sanger, L. F. Smith, and Ruth Kitai, "The disulphide bonds of insulin", Biochemical Journal, vol. 60, no. 4, pp. 541–556, 08 1955.

[2] F. Sanger, S. Nicklen, and A. R. Coulson, "Dna sequencing with chain-terminating inhibitors", Proceedings of the National Academy of Sciences of the United States of America, vol. 74, no. 12, pp. 5463–5467, 12 1977.

[3] David Eisenberg, Edward M. Marcotte, Ioannis Xenarios, and Todd O. Yeates, "Protein function in the post-genomic era", Nature, vol. 405, pp. 823 EP –, 06 2000.

[4] Eric Bertin, Statistical Physics of Complex Systems: A Concise Introduction, Springer International Publishing, 2 edition, 2016.

[5] Santo Fortunato, "Community detection in graphs", Physics Reports, vol. 486, no. 3, pp. 75–174, 2010.

[6] Albert-László Barabási and Zoltán N. Oltvai, "Network biology: understanding the cell's functional organization", Nature Reviews Genetics, vol. 5, pp. 101 EP –, 02 2004.

[7] Leland H. Hartwell, John J. Hopfield, Stanislas Leibler, and Andrew W. Murray, "From molecular to modular cell biology", Nature, vol. 402, pp. C47 EP –, 12 1999.

[8] Zhi Wang and Jianzhi Zhang, "In search of the biological significance of modular structures in protein networks", PLoS Computational Biology, vol. 3, no. 6, pp. e107, 06 2007.

[9] Jingchun Chen and Bo Yuan, "Detecting functional modules in the yeast protein–protein interaction network", Bioinformatics, vol. 22, no. 18, pp. 2283–2290, 09 2006.

[10] Roger Guimerà and Luís A. Nunes Amaral, "Functional cartography of complex metabolic networks", Nature, vol. 433, pp. 895 EP –, 02 2005.

[11] Aristóteles Góes-Neto, Marcelo V. C. Diniz, Leonardo B. L. Santos, Suani T. R. Pinho, José G. V. Miranda, Thierry Petit Lobao, Ernesto P. Borges,

Charbel Niño El-Hani, and Roberto F. S. Andrade, "Comparative protein analysis of the chitin metabolic pathway in extant organisms: A complex network approach", Biosystems, vol. 101, no. 1, pp. 59–66, 2010.

[12] R. M. Schwartz and M. O. Dayhoff, "Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts", Science, vol. 199, no. 4327, pp. 395, 01 1978.

[13] Naiara Rodríguez-Ezpeleta and T. Martin Embley, "The sar11 group of alpha-proteobacteria is not related to the origin of mitochondria", PLOS ONE, vol. 7, no. 1, pp. e30520–, 01 2012.

[14] David A. Fitzpatrick, Christopher J. Creevey, and James O. McInerney, "Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the rickettsiales", Molecular Biology and Evolution, vol. 23, no. 1, pp. 74–85, 01 2006.

[15] Ariane Atteia, Annie Adrait, Sabine Brugière, Marianne Tardif, Robert Van Lis, Oliver Deusch, Tal Dagan, Lauriane Kuhn, Brigitte Gontero, William Martin, Jérôme Garin, Jacques Joyard, and Norbert Rolland, "A proteomic survey of chlamydomonas reinhardtii mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the alpha-proteobacterial mitochondrial ancestor", Molecular Biology and Evolution, vol. 26, no. 7, pp. 1533–1548, 07 2009.

[16] Xiao Chang, Zhuo Wang, Pei Hao, Yuan-Yuan Li, and Yi-Xue Li, "Exploring mitochondrial evolution and metabolism organization principles by comparative analysis of metabolic networks", Genomics, vol. 95, no. 6, pp. 339–344, 2010.

[17] Matteo P. Ferla, J. Cameron Thrash, Stephen J. Giovannoni, and Wayne M. Patrick, "New rrna gene-based phylogenies of the alphaproteobacteria provide perspective on major groups, mitochondrial ancestry and phylogenetic instability", PLOS ONE, vol. 8, no. 12, 12 2013.

[18] Daniel S. Carvalho, Roberto F. S. Andrade, Suani T. R. Pinho, Aristóteles Góes-Neto, Thierry C. P. Lobão, Gilberto C. Bomfim, and Charbel N. El-Hani, "What are the evolutionary origins of mitochondria? a complex network approach", PLOS ONE, vol. 10, no. 9, pp. e0134988–, 09 2015.

[19] Zhang Wang and Martin Wu, "An integrated phylogenomic approach toward pinpointing the origin of mitochondria", Scientific Reports, vol. 5, pp. 7949 EP –, 01 2015.

[20] Joran Martijn, Julian Vosseberg, Lionel Guy, Pierre Offre, and Thijs J. G. Ettema, "Deep mitochondrial origin outside the sampled alphaproteobacteria", Nature, vol. 557, no. 7703, pp. 101–105, 2018.

[21] S. Boccaletti, G. Bianconi, R. Criado, C. I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, "The structure and dynamics of multilayer networks", Physics Reports, vol. 544, no. 1, pp. 1–122, 2014.

[22] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter, "Multilayer networks", Journal of Complex Networks, vol. 2, no. 3, pp. 203–271, 09 2014.

[23] Manlio De Domenico, Clara Granell, Mason A. Porter, and Alex Arenas, "The physics of spreading processes in multilayer networks", Nature Physics, vol. 12, pp. 901 EP –, 08 2016.

[24] Leonard Euler, "Solutio problematis ad geometriam situs pertinentis", Commentarii academiae scientiarum Petropolitanae, vol. 8, pp. 128–140, 1741.

[25] Bela Bollobas, Modern Graph Theory, Springer, 07 1998.

[26] Réka Albert and Albert Barabási, "Statistical mechanics of complex networks", Reviews of Modern Physics, vol. 74, no. 1, pp. 47–97, Jan 2002.

[27] Duncan J. Watts and Steven H. Strogatz, "Collective dynamics of 'small-world'networks", Nature, vol. 393, pp. 440 EP –, 06 1998.

[28] Albert-László Barabási and Réka Albert, "Emergence of scaling in random networks", Science, vol. 286, no. 5439, pp. 509–512, 10 1999.

[29] Mark Newman, Albert-Laszlo Barabasi, and Duncan J Watts, The structure and dynamics of networks, Princeton University Press, 2006.

[30] Jonathan Gross and Jay Yellen, Handbook of Graph Theory (Discrete Mathematics and Its Applications), CRC Press, 12 2003.

[31] M. E. J. Newman, Networks: An Introduction, Oxford University Press, 03 2010.

[32] Roberto F. S. Andrade, José G. V. Miranda, and Thierry Petit Lobão, "Neighborhood properties of complex networks", Physical Review E, vol. 73, no. 4, pp. 046101–, 04 2006.

[33] R. F. S. Andrade, J. G. V. Miranda, S. T. R. Pinho, and T. P. Lobão, "Characterization of complex networks by higher order neighborhood properties", The European Physical Journal B, vol. 61, no. 2, pp. 247–256, 2008.

[34] F. N. Silva, M. P. Viana, B. A. N. Travençolo, and L. F. Costa, "Investigating relationships within and between category networks in wikipedia", Journal of Informetrics, vol. 5, no. 3, pp. 431–438, 2011.

[35] R. Duncan Luce and Albert D. Perry, "A method of matrix analysis of group structure", Psychometrika, vol. 14, no. 2, pp. 95–116, 1949.

[36] Vito Latora and Massimo Marchiori, "Efficient behavior of small-world networks", Physical Review Letters, vol. 87, no. 19, pp. 198701–, 10 2001.

[37] Linton Freeman, "A set of measures of centrality based on betweenness", Sociometry, vol. 40, no. 1, pp. 35–41, 03 1977.

[38] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices", Phys. Rev. E, vol. 74, pp. 036104, Sep 2006.

[39] Jörg Reichardt and Stefan Bornholdt, "Statistical mechanics of community detection", Phys. Rev. E, vol. 74, pp. 016110, Jul 2006.

[40] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks", Physical Review E, vol. 69, no. 2, pp. 026113, Feb 2004.

[41] Roberto F. S. Andrade, Ivan C. Rocha-Neto, Leonardo B. L. Santos, Charles N. de Santana, Marcelo V. C. Diniz, Thierry Petit Lobão, Aristóteles Goés-Neto, Suani T. R. Pinho, and Charbel N. El-Hani, "Detecting network communities: An application to phylogenetic analysis", PLOS Computational Biology, vol. 7, no. 5, pp. e1001131–, 05 2011.

[42] Roberto F. S. Andrade, José G. V. Miranda, Suani T. R. Pinho, and Thierry Petit Lobão, "Measuring distances between complex networks", Physics Letters A, vol. 372, no. 32, pp. 5265–5269, 2008.

[43] Luis Solá, Miguel Romance, Regino Criado, Julio Flores, Alejandro García del Amo, and Stefano Boccaletti, "Eigenvector centrality of nodes in multiplex networks", Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 23, no. 3, pp. 033131, 2018/01/23 2013.

[44] Federico Battiston, Vincenzo Nicosia, and Vito Latora, "Structural measures for multiplex networks", Physical Review E, vol. 89, no. 3, pp. 032804–, 03 2014.

[45] Petter Holme and Jari Saramäki, "Temporal networks", Physics Reports, vol. 519, no. 3, pp. 97–125, 2012.

[46] J. F. Donges, H. C. H. Schultz, N. Marwan, Y. Zou, and J. Kurths, "Investigating the topology of interacting networks", The European Physical Journal B, vol. 84, no. 4, pp. 635–651, 2011.

[47] M. Coscia, Multidimensional network analysis, PhD thesis, Universitá Degli Studi Di Pisa, 2012.

[48] Regino Criado, Julio Flores, Alejandro García del Amo, Jesús Gómez-Gardeñes, and Miguel Romance, "A mathematical model for networks with structures in the mesoscale", International Journal of Computer Mathematics, vol. 89, no. 3, pp. 291–309, 02 2012.

[49] Manlio De Domenico, Albert Solé-Ribalta, Emanuele Cozzo, Mikko Kivelä, Yamir Moreno, Mason A. Porter, Sergio Gómez, and Alex Arenas, "Mathematical formulation of multilayer networks", Physical Review X, vol. 3, no. 4, pp. 041022–, 12 2013.

[50] Emanuele Cozzo, Mikko Kiveläand Manlio De Domenico, Albert Solé-Ribalta, Alex Arenas, Sergio Gómez, Mason A Porter, and Yamir Moreno, "Structure of triadic relations in multiplex networks", New Journal of Physics, vol. 17, no. 7, pp. 073029, 2015.

[51] Peter J. Mucha, Thomas Richardson, Kevin Macon, Mason A. Porter, and Jukka-Pekka Onnela, "Community structure in time-dependent, multiscale, and multiplex networks", Science, vol. 328, no. 5980, pp. 876, 05 2010.

[52] Inderjit S. Jutla Lucas G. S. Jeub, Marya Bazzi and Peter J. Mucha, "A generalized louvain method for community detection implemented in matlab", (2011-2017).

[53] Gilles Didier, Christine Brun, Anaïs Baudot, and Shawn Gomez, "Identifying communities from multiplex biological networks", PeerJ, vol. 3, pp. e1525, 2015.

[54] Laura Bennett, Aristotelis Kittas, Gareth Muirhead, Lazaros G. Papageorgiou, and Sophia Tsoka, "Detection of composite communities in multiplex biological networks", Scientific Reports, vol. 5, pp. 10345 EP –, 05 2015.

[55] Zhana Kuncheva and Giovanni Montana, "Community detection in multiplex networks using locally adaptive random walks", CoRR, vol. abs/1507.01890, 2015.

[56] Manlio De Domenico, Andrea Lancichinetti, Alex Arenas, and Martin Rosvall, "Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems", Physical Review X, vol. 5, no. 1, pp. 011027–, 03 2015.

[57] Manel Hmimida and Rushed Kanawati, Community detection in multiplex networks: A seed-centric approach, vol. 10, 03 2015.

[58] Matteo Barigozzi, Giorgio Fagiolo, and Giuseppe Mangioni, "Identifying the community structure of the international-trade multi-network", Physica A: Statistical Mechanics and its Applications, vol. 390, no. 11, pp. 2051–2066, 2011.

[59] Michele Berlingerio, Michele Coscia, and Fosca Giannotti, "Finding redundant and complementary communities in multidimensional networks", in Proceedings of the 20th ACM International Conference on Information and Knowledge Management, New York, NY, USA, 2011, CIKM '11, pp. 2181–2184, ACM.

[60] Lei Tang, Xufei Wang, and Huan Liu, "Community detection via heterogeneous interaction analysis", Data Mining and Knowledge Discovery, vol. 25, no. 1, pp. 1–33, 2012.

[61] Magnani; M., Micenková; B., and Rossi; L., "Combinatorial analysis of multiple networks", CoRR, vol. abs/1303.4986, 2013.

[62] M. Newman, "The structure and function of complex networks", SIAM Review, vol. 45, no. 2, pp. 167–256, 2018/01/25 2003.

[63] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang, "Complex networks: Structure and dynamics", Physics Reports, vol. 424, no. 4-5, pp. 175–308, 02 2006.

[64] W. W. Zachary, "An information flow model for conflict and fission in small groups", Journal of Anthropological Research, vol. 33, pp. 452–473, 1977.

[65] Pall F. Jonsson, Tamara Cavanna, Daniel Zicha, and Paul A. Bates, "Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis", BMC Bioinformatics, vol. 7, pp. 2–2, 2006.

[66] B. Krishnamurthy and J. Wang, "On network-aware clustering of web clients", SIGCOMM Comput. Commun. Rev., vol. 30, no. 4, pp. 97–110, August 2000.

[67] P. K. Reddy, M. Kitsuregawa, P. Sreekanth, and S. S. Rao, "A graph based approach to extract a neighborhood customer community for collaborative filtering", in Proceedings of the Second International Workshop on Databases in Networked Information Systems, London, UK, UK, 2002, DNIS '02, pp. 188–200, Springer-Verlag.

[68] Robert S. Weiss and Eugene Jacobson, "A method for the analysis of the structure of complex organizations", American Sociological Review, vol. 20, no. 6, pp. 661–668, 1955.

[69] Santo Fortunato and Darko Hric, "Community detection in networks: A user guide", Physics Reports, vol. 659, pp. 1–44, 2016.

[70] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks", Proceedings of the National Academy of Sciences, vol. 99, no. 12, pp. 7821–7826, 06 2002.

[71] J. M. Anthonisse, "The rush in a directed graph", Tech. Rep., Amsterdam, 10 1971.

[72] M. E. J. Newman, "A measure of betweenness centrality based on random walks", Social Networks, vol. 27, no. 1, pp. 39–54, 01 2005.

[73] 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, vol. 2008, no. 10, pp. P10008, 2008.

[74] M. E. J. Newman, "Fast algorithm for detecting community structure in networks", Physical Review E, vol. 69, no. 6, pp. 066133, June 2004.

[75] A. Arenas, J. Duch, A. Fernández, and S. Gómez, "Size reduction of complex networks preserving modularity", New Journal of Physics, vol. 9, no. 6, pp. 176–176, 06 2007.

[76] V. H. Heywood and J. Mcneill, "Phenetic and phylogenetic classification", Nature, vol. 203, pp. 1220 EP –, 09 1964.

[77] James P. Ferris, Iris Fry, The Emergence of Life on Earth - A Historical and Scientific Overview, vol. 31, 06 2001.

[78] L. L. Cavalli-Sforza and A. W. F. Edwards, "Phylogenetic analysis. models and estimation procedures", American Journal of Human Genetics, vol. 19, no. 3 Pt 1, pp. 233–257, 05 1967.

[79] Francesca D. Ciccarelli, Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel, and Peer Bork, "Toward automatic reconstruction of a highly resolved tree of life", Science, vol. 311, no. 5765, pp. 1283, 03 2006.

[80] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman, "Basic local alignment search tool", Journal of Molecular Biology, vol. 215, no. 3, pp. 403–410, 1990.

[81] Scott McGinnis and Thomas L. Madden, "Blast: at the core of a powerful and diverse set of sequence analysis tools", Nucleic Acids Research, vol. 32, no. Web Server issue, pp. W20–W25, 07 2004.

[82] A. M. Lesk, Introdução à bioinformática, Artmed, 2008.

[83] Stephen Wooding, "Inferring phylogenies.", American Journal of Human Genetics, vol. 74, no. 5, pp. 1074–1074, 05 2004.

[84] S. Van Dongen, "Graph clustering by flow simulation", May 2000.

[85] S. Van Dongen, "Graph clustering via a discrete uncoupling process", SIAM Journal on Matrix Analysis and Applications, vol. 30, no. 1, pp. 121–141, 2018/01/19 2008.

[86] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families", Nucleic Acids Research, vol. 30, no. 7, pp. 1575–1584, 04 2002.

[87] Barbara Robbertse, John B Reeves, Conrad L Schoch, and Joseph W Spatafora, "A phylogenomic analysis of the ascomycota", Fungal genetics and biology : FG & B, vol. 43, no. 10, pp. 715–725, 2006.

[88] Igor V. Tetko, Axel Facius, Andreas Ruepp, and Hans-Werner Mewes, "Super paramagnetic clustering of protein sequences", BMC Bioinformatics, vol. 6, no. 1, pp. 82, 2005.

[89] Anne-Claude Gavin, Patrick Aloy, Paola Grandi, Roland Krause, Markus Boesche, Martina Marzioch, Christina Rau, Lars Juhl Jensen, Sonja Bastuck, Birgit Dümpelfeld, Angela Edelmann, Marie-Anne Heurtier, Verena Hoffman, Christian Hoefert, Karin Klein, Manuela Hudak, Anne-Marie Michon, Malgorzata Schelder, Markus Schirle, Marita Remor, Tatjana Rudi, Sean Hooper, Andreas Bauer, Tewis Bouwmeester, Georg Casari, Gerard Drewes, Gitte Neubauer, Jens M. Rick, Bernhard Kuster, Peer Bork, Robert B. Russell, and Giulio Superti-Furga, "Proteome survey reveals modularity of the yeast cell machinery", Nature, vol. 440, pp. 631 EP –, 01 2006.

[90] Charles Boone, Howard Bussey, and Brenda J. Andrews, "Exploring genetic interactions and networks with yeast", Nature Reviews Genetics, vol. 8, pp. 437 EP –, 06 2007.

[91] Eric de Silva and Michael P. H Stumpf, "Complex networks and simple models in biology", Journal of The Royal Society Interface, vol. 2, no. 5, pp. 419, 12 2005.

[92] Lynn Margulis, Origin of Eukaryotic Cells, Yale University Press, 1970.

[93] W. Martin and Klaus V. Kowallik, "Annotated english translation of mereschkowsky's 1905 paper 'über natur und ursprung der chromatophoren im pflanzenreiche"', vol. 34, no. 3, pp. 287–295, 1999.

[94] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D Thompson, and Desmond G Higgins, "Fast, scalable generation of high quality protein multiple sequence alignments using clustal omega", Molecular Systems Biology, vol. 7, no. 1, 01 2011.