UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

João Paulo Cassucci dos Santos

Biologia de sistemas e aprendizagem de máquina: novas aplicações de métodos computacionais

São Carlos

2024

João Paulo Cassucci dos Santos

Biologia de sistemas e aprendizagem de máquina: novas aplicações de métodos computacionais

Dissertação apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Mestre em Ciências.

Área de concentração: Física Aplicada

Opção: Física Biomolecular

Orientador: Prof. Dr. Odemir Martinez Bruno

Versão original

São Carlos

2024

Santos, João Paulo Cassucci dos
Biologia de sistemas e aprendizagem de máquina: novas aplicações de métodos computacionais / João Paulo Cassucci dos Santos; orientador Odemir Martinez Bruno -- São Carlos, 2024.
87 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Física Aplicada Biomolecular) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2024.

1. Ciência de redes. 2. Aprendizagem de máquina. 3. Biologia de sistemas. 4. Bioinformática. 5. Biologia molecular. I. Bruno, Odemir Martinez, orient. II. Título.

FOLHA DE APROVAÇÃO

João Paulo Cassucci dos Santos

Dissertação apresentada ao Instituto de Física de São Carlos da Universidade de São Paulo para obtenção do título de Mestre em Ciências. Área de Concentração: Física Biomolecular.

Aprovado (a) em: 08/03/2024

Comissão Julgadora

Dr(a).: Odemir Martinez Bruno

Instituição: (IFSC/USP)

Dr(a).: Thadeu Josino Pereira Penna

Instituição: (UFF/Volta Redonda)

Dr(a).: Tie Koide

Instituição: (FMRP/USP)

"Man has to awaken to wonder - and so perhaps do peoples.
Science is a way of sending him to sleep again."
Ludwig Wittgenstein

RESUMO

SANTOS, J. P. C. **Biologia de sistemas e aprendizagem de máquina:** novas aplicações de métodos computacionais. 2024. 87p. Dissertação (Mestrado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2024.

A ciência de redes nos permite modelar problemas multivariados e complexos de uma forma relativamente simples. Esta vantagem tem se demonstrado bastante promissora dentro do contexto de pesquisas interdisciplinares, pois ela permite caracterizar quantitativamente problemas que antes podiam apenas ser estudados qualitativamente. Um área promissora para a aplicação da ciência de redes é a da biologia molecular, em específico, a biologia de sistemas, onde o contexto em que elementos discretos estão inseridos importa mais do que suas propriedades isoladas. Nesta dissertação, buscamos explorar de duas maneiras distintas as propriedades de redes de modo a averiguar possíveis conclusões biológicas que podem ser extraídas a partir de diferentes experimentos biomoleculares. A primeira abordagem utiliza-se de um novo método de mensurar similaridade entre vetores conhecido como índice de coincidência, que demonstrou ser mais eficiente na extração de informação biológica em redes de interação enzima-enzima do que medidas de correlação tradicionalmente utilizadas para estas modelagens, como o r de Pearson e Spearman. A segunda abordagem aplica novos métodos de extração de características em redes complexas, como o "Life-like Network Automata" e o "Deterministic Tourist Walk", em conjunto com aplicações de algoritmos de aprendizagem de máquina para classificar bancos de dados de redes biológicas que poderão auxiliar na classificação de organismos e na predição de novas vias metabólicas.

Palavras-chave: Ciência de redes. Aprendizagem de máquina. Biologia de sistemas. Bioinformática. Biologia molecular.

ABSTRACT

SANTOS, J. P. C. **Systems biology and machine learning:** new applications of computational methods. 2024. 87p. Dissertation (Master in Science) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2024.

Network Science allows us to model multivariate and complex problems in a relatively simple way. This advantage has been shown to be very promising in the context of interdisciplinary researches because it allows us to characterize problems quantitatively that before could only be studied qualitatively. One promising research area for the application of network science is molecular biology, in specific, systems biology, where the context in which the discrete elements belong is more important than their isolated properties. In this dissertation, we intended to explore in two distinct ways the network properties in order to investigate possible biological conclusions that can be extracted from different biomolecular experiments. The first approach uses a new way to measure similarity between vectors known as coincidence index, which was shown to be more effective in the extraction of biological information from enzyme-enzyme interaction networks than the more common correlation measurements traditionally used in these types of modelings, like Pearson's and Spearman's r. The second approach applies new complex network feature extraction techniques, such as *Life-Like Network Automata* and the *Deterministic Tourist Walk*, together with machine learning algorithms to classify biological networks datasets that can help in the classification of species and in the prediction of new metabolic pathways.

Keywords: Network science. Machine learning. Systems biology. Bioinformatics. Molecular biology.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LLNA	Life-Like Network Automata
DTEP	Density Time Evolution Pattern
SDTEP	State-Density Time Evolution Pattern
HCCA	Heuristic Clustering Chieseling Algorithm
HRR	Highest Reciprocal Ranking
Coindex	Coincidence Index
ML	Machine Learning
NN	Neural Network
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
GSEA	Gene Set Enrichment Analysis
RNA-Seq	RNA sequencing
NCBI	National Center for Biotechnology Information
ATP	Adenosine Triphosphate
ADP	Adenosine Diphosphate
FDR	False Discovery Rate
NADH	Nicotinamide Adenine Dinucleotide + Hydrogen
NAD+	Nicotinamide Adenine Dinucleotide
ETC	Electron Transport Chain
FAD	Flavin Adenine Dinucleotide
SVM	Support Vector Machine
LDA	Linear Discriminant Analysis
GNN	Graph Neural Network
DTWB	Deterministic Tourist Walk Bifurcated

LISTA DE SÍMBOLOS

σ	Desvio Padrão
ρ	Coeficiente de Correlação de Pearson
Z	Z valor
θ	Função de acumulação de probabilidade inversa
R	Função de ranqueamento de vetores
Σ	Somatório
$\in$	Pertence
$\{\}$	Conjunto
$<$	Menor que
Ω	Vetor característico do LLNA
$\cap$	Intersecção
$\|\|$	Módulo
$\langle\rangle$	Média
β	Razão entre número de bifurcações e grau
μ	Memória do algoritmo DTWB
ϕ	Função kernel do algoritmo SVM

SUMÁRIO

1 INTRODUÇÃO

1.1 A Biologia Molecular

As origens da biologia molecular são geralmente atribuídas a James Watson e Francis Crick pela sua descoberta da estrutura de dupla-hélice da molécula de DNA através de experimentos de cristalografia por raio-x. Porém, esta descoberta, assim como todas as outras grandes revoluções científicas, só foi possível graças ao refinamento de conhecimentos anteriores que tiveram suas técnicas aprimoradas o suficiente para abrir espaço para esta nova ciência nascer.

Antes de ter sua estrutura elaborada, por exemplo, a molécula de DNA já havia sido descoberta pelo químico Friedrich Miescher no final da decada de 1860 (2). Décadas depois, Phoebus Levene e Erwin Chargaff foram capazes de isolar os monômeros que compunham o polímero da molécula de DNA, além das ligações estabelecidas entre eles (3, 4). Além disso, Phoebus foi também capaz de descobrir as componentes da molécula de RNA e a possível hibridização que ela poderia exercer com a molécula de DNA.

Chargaff expandiu sobre a pesquisa de Levene ao buscar verificar se havia diferenças na arranjo do DNA em diferentes espécies de seres vivos (3). Ele baseou-se em um famoso paper de Oswald Avery que demonstrou que as unidades hereditárias dos seres vivos (os genes) eram compostas de DNA. Através de uma cromatografia de papel, Chargaff foi capaz de descobrir que a composição dos nucleotídeos das espécies varia, ou seja, a ordem com que os nucleotídeos apareciam não era constante entre as espécies. Ele também foi capaz de descobrir que a quantia de adenina (A) no DNA era igual a de timina (T), e a quantia de citosina (C) igual a de guanina (G), embora ele não tenha concluído que isto era devido ao pareamento destes nucleotídeos.

Foi apenas através destas descobertas, e dos experimentos cristalográficos de Rosalind Franklin, que Watson e Crick foram capazes de elaborar a primeira estrutura tridimensional do DNA (Figura 1B) e conceber o "Dogma Central da Biologia Molecular". Este dogma propõe, nas palavras de Fracis Crick que "uma vez que informação tenha passado para as proteínas ela não pode mais sair. Em mais detalhes, a transferência de informação de ácido nucleico para ácido nucleico ou de ácido nucleico para proteína pode ser possível, mas a transferência de proteína para proteína ou de proteína para ácido nucleico é impossível" (Figura 1A).

Desde esta descoberta, o restante do século XX foi um período de grandes avanços nesta área do conhecimento que acabava de nascer. Começando pela elucidação do código genético por Har Gobind Khorana, depois para o surgimento da técnica de primers, que permitia a amplificação de um segmento de DNA em uma solução combinando moléculas

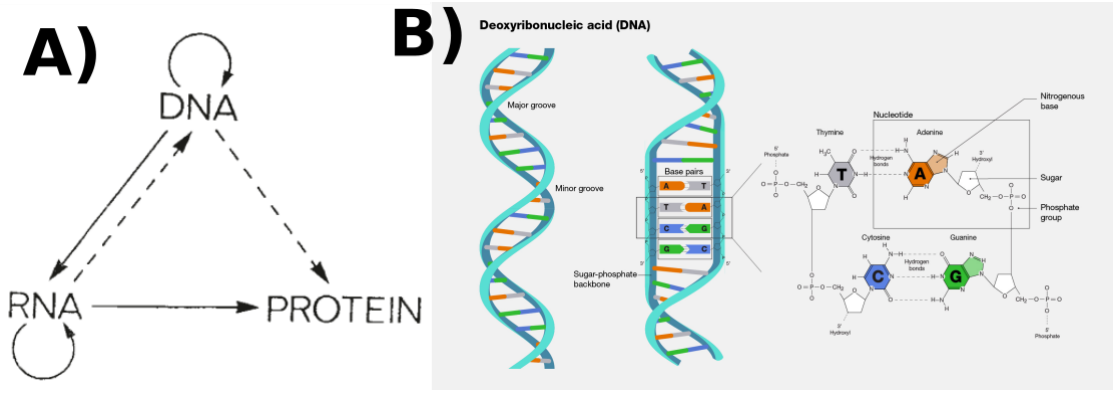

Figura 1 – A) O Dogma Central da Biologia Molecular como determinado por Francis Crick em 1958. As setas inteiras mostral a transferência mais provável de informação de um elemento para outro enquanto que as tracejadas trânsferencias possíveis mas menos prováveis. B) Nesta imagem está sendo mostrado a estrutura tridimentsional da molécula de DNA, o pareamento entre os ácidos nucleicos que a compõe e também as estruturas químicas destes ácidos nucleicos.

Fonte: Adaptada de CRICK. (5)

precursoras dos resíduos do polímero final junto com uma enzima DNA-Polimerase modificada derivada da bactéria *Thermus aquaticus*. Esta enzima possuia uma resistência a altas temperaturas, podendo manter sua funcionalidade em temperaturas próximas a da denaturação das fitas de DNA.

Em 1984, o primeiro experimento de reação em cadeia de polimerase (Polymerase Chain Reaction) foi executado resultando na ampliação de uma porção específica de 110 pares de base em um DNA. Esta técnica foi uma segunda revolução na biologia molecular, pois possibilitou a investigação e sequenciamento de ácidos nucleicos de maneira extremamente eficiente e prática.

A partir daí, observou-se uma diminuição exponencial do custo destas técnicas. Para se ter uma ideia, o projeto do genoma humano chegou a custar 2,7 bilhões de dólares no final dos anos 90, e hoje, o custo de um sequenciamento é de aproximadamente 600 dólares. Esta redução imensa do custo do sequenciamento do DNA gerou uma quantia enorme de dados de experimentos biomoleculares, o que atraiu a área da ciência de dados e da computação dando origem a bioinformática.

1.2 A Ciência de Redes

A análise de redes é uma área do conhecimento recente que têm se mostrado promissora para o entendimento de fenômenos multivariados e caóticos (7). Embora o objeto de estudo desta ciência (grafos) já exista a mais de 2 séculos, apenas no século XXI foi possível analisar suas propriedades mais a fundo com o aumento da capacidade de armazemento de dados possibilitado pelos computadores. Além disso, com o surgimento da

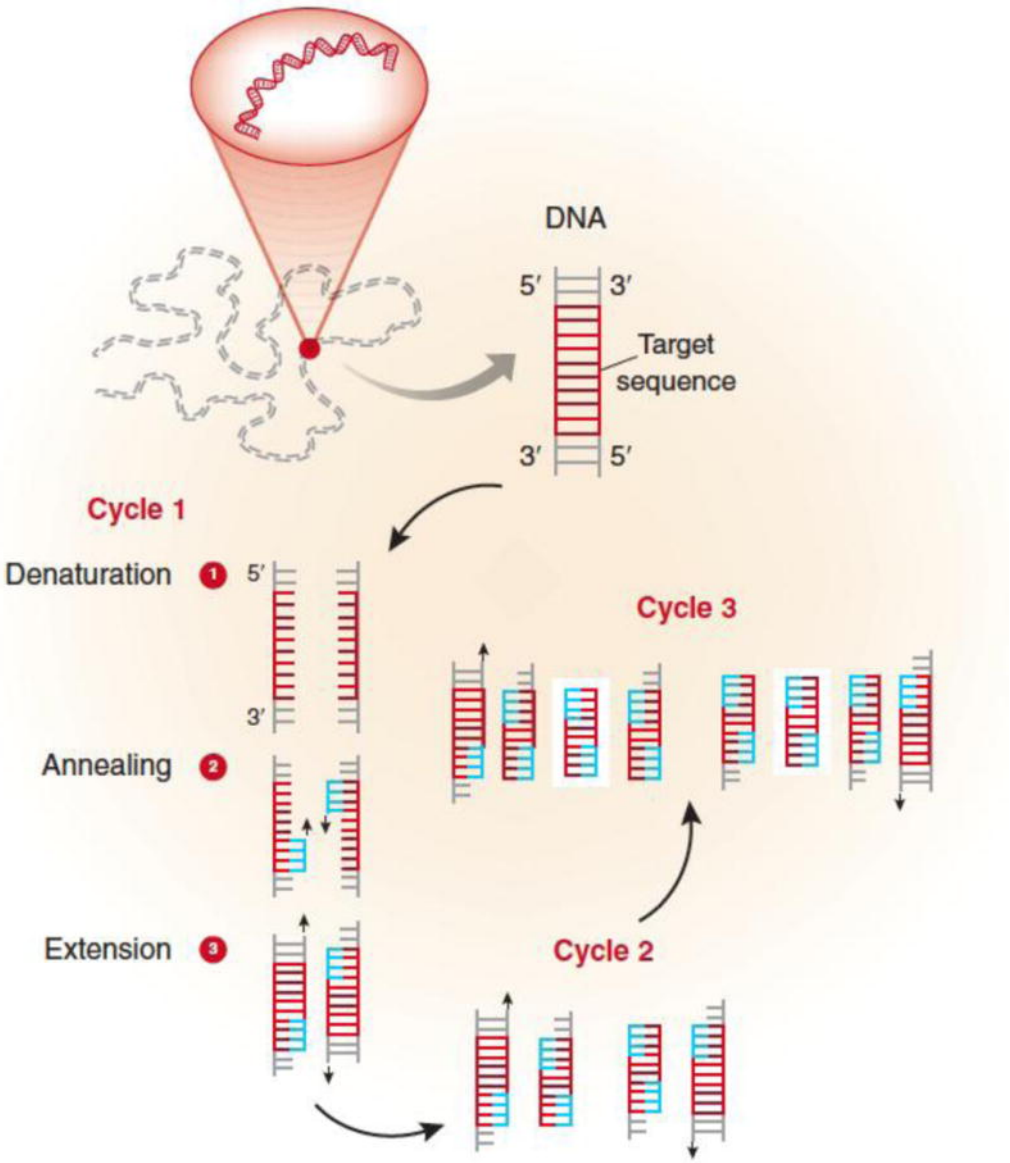

Figura 2 – Esquema mostrando o processo de amplificação de uma porção determinada do DNA por meio da técnica de PCR.

Fonte: Adaptada de GARIBYAN *et al.* (6)

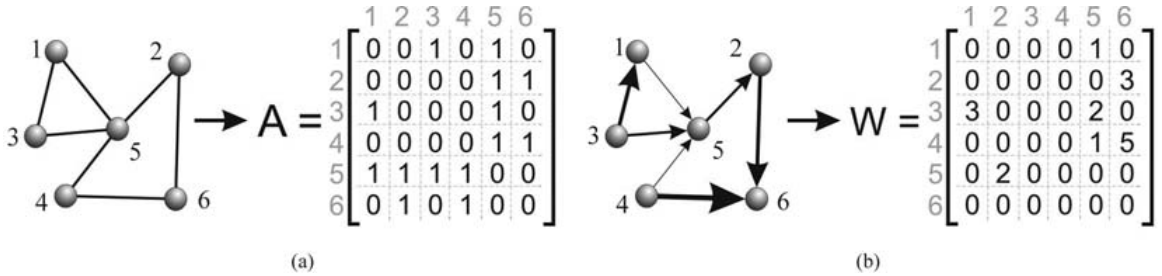

Figura 3 – Em (a) temos a representação em matriz de adjacência de um grafo sem peso ou direção, em (b) a representação de um grafo ponderado e direcionado.

Fonte: Adaptada de COSTA *et al.* (8)

internet, o compartilhamento de informação se tornou mais fácil e permitiu a transferência entre computadores da enorme quantidade de dados necessária para a construção das redes.

Redes, também conhecidas como grafos, são estruturas matemáticas compostas por vértices conectados por arestas. Definidas matematicamente como um par ordenado $G = (V, E)$, onde $V = \{1, 2, i, ..., N\}$ é o conjunto de nós/vértices e $E = \{e_1, e_2, e_3, ..., e_M\}$ o conjunto de arestas. Elas podem ser direcionadas (as arestas são dotadas de sentido) e também podem ter um peso. Quando redes possuem peso em suas arestas, elas são chamadas "redes ponderadas"(*weighted networks*), e quando suas arestas possuem sentido elas são chamadas de "redes direcionadas"(*directed networks*).

Além disso, redes podem possuir uma representação na forma de matrizes de adjacência, onde cada entrada na matriz corresponde a uma possível conexão entre seus nós, com 1 e 0 representando a existência ou não dele respectivamente. Nesta representação, a matriz é sempre quadrada pois todos os nós podem teoricamente se ligar a todos os outros inclusive a si mesmo, mas raramente este é caso, logo, estas matrizes tendem a ser bem esparsas. Outras propriedades das matrizes, como a direcionalidade e peso das arestas, podem ser representadas permitindo que os valores das matrizes sejam números reais, indicando o peso, e também que haja quebra de simetria para representar direcionalidade (neste caso, a aresta (n,m) existiria mas (m,n) não, por exemplo).

Redes não-sintéticas tendem a conter uma quantia muito numerosa de nós e arestas e uma organização topológica de difícil modelagem, no entanto, existem na literatura diversas caracterizações de grafos cujo intuito é facilitar a compreensão das diferenças entre suas topologias e gerar redes sintéticas com propriedades similares. Exemplos de tais modelos são: o modelo Erdos-Rényi aplicado para redes aleatórias onde arestas são estabelecidas entre os nós de forma a seguir uma distribuição normal de probabilidade;(9) o modelo de Watts-Strogatz que também caracteriza uma rede aleatória mas com uma característica conhecida como pequeno-mundo, onde todos os nós estão a poucos vizinhos em relação a todos os outros;(10) o modelo de Barabasí-Albert para redes de livre-escala

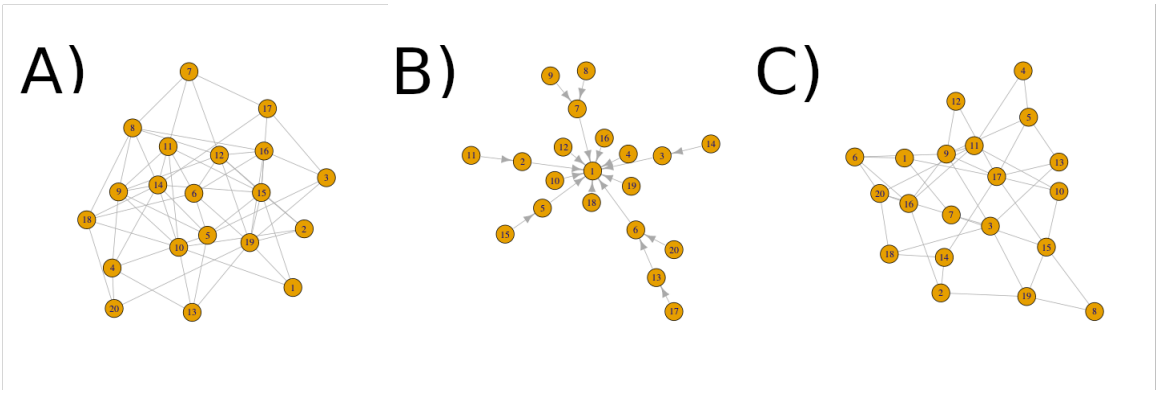

Figura 4 – Três tipos de modelagens distintas de grafos. A) Modelo de grafo randômico. B) Modelo de grafo livre-escala. C) Modelo de grafo randômico pequeno-mundo.

Fonte: Elaborada pelo Autor

onde a distribuição da quatia de nós com um determinado grau segue um decaimento exponencial conforme ele aumenta.(11)

Embora estes modelos tenham sido úteis para analisar uma gama de problemas distintos, distinguir entre diferentes redes complexas ainda é uma questão em aberto, pois muitas redes que fazem parte de um mesmo modelo topológico ainda podem possuir propriedades muito diferentes. A grande maioria das redes não-sintéticas podem ser entendidas como sendo de alta complexidade e difícil caracterização, e requerem métodos e algoritmos robustos para uma classificação satisfatória.(8)

1.3 Aplicações das Redes na Biologia Molecular

Assim como a biologia molecular e a ciência de redes dependeram de diversos avanços técnicos para se consolidarem como suas próprias ciências, a intersecção entre estas duas áreas do conhecimento tem rendido descobertadas interessantes, em especial na biologia estrutural de proteínas e ácidos nucleicos e na biologia de sistemas.

Existem vários exemplos de redes construídas a partir de experimentos pertencentes à biologia molecular. Elas podem ir de redes metabólicas dos processos químicos intracelulares (12) a redes de interação proteína-proteína, redes de estruturas terciárias de proteínas, redes de co-expressão genética, e redes de interação enzima-enzima por metabólitos (13,14). Redes também exerceram um papel fundamental na construção do algoritmo de predição de estruturas AlphaFold2, ao modelar a estrutura tridimensional de proteínas na forma de grafos para utilizá-las como arcabouço na predição de novas estruturas ao comparar a distância par-a-par entre estruturas homólogas (15).

Neste mestrado, foram utilizados dois modos distintos de explorar a topologia de redes complexas com o propósito de modelar dados biológicos de forma mais acurada.

A primeira abordagem buscou ser uma análise de uma nova medida de coincidência

entre vetores contendo valores reais que tem demonstrado grande potencial na obtenção de informações em redes complexas, o índice de coincidência (Coindex). A arquitetura de redes ponderadas fornece vantagens na área de biologia de sistemas, pois ela permite agregar evidências biológicas derivadas de fontes distintas fortalecendo as conclusões biológicas retiradas a partir destes experimentos individualmente.

Um exemplo de como isto é útil fica evidente na análise de co-expressão em séries temporais. Usualmente, este processo utiliza-se de correlações entre vetores de expressões proteicas no espaço euclidiano cuja dimensão é dada pela quantia de pontos temporais N, como no método GSEA (Gene Set Enrichment Analysis).(16, 17) A hipótese é que genes que possuam uma alta correlação em sua expressão entre si ao longo do tempo estão também correlacionados em sua função biológica, e elucidam possíveis novas vias metabólicas, ou novos componentes de um complexo proteico por exemplo. Porém, este espaço é esparso com diversas correlações sendo espúrias, dificultando não só a computação do problema para situações onde a quantidade de proteínas é muito grande mas abrindo também a possibilidade de gerar falsos positivos que comprometem as conclusões obtidas.

A inserção destas informações biológicas em um espaço representado por grafos onde vértices são as proteínas e arestas as relações biológicas conhecidas a priori sobre tais proteínas diminui a quantia de correlações a serem calculadas. Além disso, isto também fortalece a confiança sobre as conclusões biológicas retiradas a partir delas.

Esta abordagem, então, buscou verificar se a medidade de coincidência (coindex) possuiria um bom desempenho sobre topologias descritas por grafos ao invés de topologias "livres" de espaços euclidianos quando comparada a medidas de correlação entre vetores mais tradicionais, como o r de Pearson e o r de Spearman.(18) Aplicamos métodos já bem consolidados de exploração de grafos de interação enzima-enzima em conjunto com vetores de experimentos de microarray da arqueia *Halobacterium salinarum*.

A segunda abordagem explorou bancos de dados de redes mais extensivas, onde uma visão global dos grafos é necessária para a caracterização das redes ao invés de um exploração mais direcionada para as relações entre os nós, como foi o primeiro caso. Para situações como esta, métodos de extração de característica são indicados pois eles permitem uma modelagem heurística das características das redes, circundando a ausência de parâmetros analíticos que descrevam suficientemente bem as propriedades das redes.

As redes biológicas que buscamos realizar a classificação através da extração de características globais foram obtidas de duas maneiras distintas com dois intuitos diferentes ao realizar suas classificações. O primeiro tipo de rede é conhecida como uma rede de interações proteína-proteína, e ela agrega diversos tipos de informações distintas sobre as relações entre duas proteínas. Tais redes foram retiradas do banco de dados público STRINGdb.(19) O segundo tipo são redes de co-expressão genética montadas utilizando o algoritmo de obtenção de clusters PLaNET, um método já estabelecido de montagem de

grafos utilizando vetores de expressão genética.(20) Tais redes utilizam-se de experimentos de expressão genética, assim como os experimentos transcricionais de microarray da arqueia *H. salinarum* da primeira abordagem, porém eles são correlacionados em um espaço livre de informações a priori e, feito isto, são "clusterizados" buscando remover todas as correlações espúrias e mantendo apenas as mais significativas.

Como métodos de extração de características são métodos heurísticos (não possuem garantia analítica de que são a melhor forma de modelar um problema) realizamos uma comparação entre vários métodos distintos para averiguar o quão bom eles são em distinguir entre topologias globais de rede. Para isso, utilizamos também bancos de dados sintéticos e outros bancos de dados biológicos já utilizados em outros projetos do grupo de pesquisa.

Para o banco de dados de redes de interação proteína-proteína buscamos extrair **informações evolutivas** a partir das características topológicas das redes. Fizemos isso comparando como os métodos de extração de características diferenciam entre diferentes clados da árvore da vida e o quão bem segregados eles ficam no espaço de características com base em sua distância evolutiva.

Para o banco de dados de co-expressão proteica, utilizamos um banco de dados de experimentos de RNA-Seq do fungo *Aspergillus fumigatus* e montamos redes com base na variação temporal do experimento. A utilização do método PlaNET gerou então dois tipos de clusters, um contendo vias metabólicas conhecidas e outro sem informação biológica válida. Buscamos distinguir entre estes dois tipos utilizando os métodos de extração de característica, obtendo assim uma modelagem de quais redes podem conter vias metabólicas com base na topologia.

2 MÉTODOS E BANCOS DE DADOS PARA A APLICAÇÃO DA MEDIDA ÍNDICE DE COINCIDÊNCIA EM REDES BIOLÓGICAS

O uso de experimentos transcriptômicos para melhor compreender o metabolismo de seres vivos tornou-se prática comum conforme o custo para a realização destes experimentos decresceu. A enorme quantidade de dados biológicos disponível publicamente passou a requerer então tratamentos estatísticos mais robustos que melhor extraem padrões de tais dados. A primeira abordagem utilizada buscou descobrir genes correlacionados com base em métodos de "guilt-by-association" onde perfis de expressão similares indicam uma função molecular ou biológica em comum.(21) Tais perfis de expressão quantificam o nível de expressão de um determinado gene com base na quantidade de moléculas de RNA sequenciadas (nos experimentos RNA-Seq) ou com base na emissão de luz a partir de moléculas de RNA hibridizadas com a sua sequência de DNA codificante e marcadas com fluorófonos (nos experimentos de tipo "microarray").

Estes dados foram usados em conjunto com grafos de interação enzima-enzima contruídos a partir de reações enzimáticas conhecidas e descritas em experimentos de reconstruções metabólicas.(22) O uso de tais informações ajuda a fortalecer a evidência de correlação entre os genes quando aplicadas em conjunto com os experimentos de perfis de transcrição além de reduzir a quantidade de correlações que devem ser calculadas ou eliminar pares de genes que não possuem metabólitos coincidentes. Isto é feito utilizando **medidas de correlação ou coincidência** que atribuem pesos as arestas, combinando os dois experimentos em um grafo com arestas ponderadas.(23)

A medida de correlação de Pearson é a métrica mais usual utilizada para avaliar a similaridade entre dois vetores, mas ela pode ser ineficaz ao lidar com dados ruidosos devido a maior presença de "outliers" que mudam o resultado significativamente. A medida de correlação de Spearman busca corrigir os efeitos dos "outliers" ranqueando os elementos dos vetores, mas isto têm efetividade limitada.(24)

O **índice de coincidência** é uma nova medida de similaridade entre vetores contendo valores reais definida pela combinação dos índices de Jaccard e interioridade.(18) Estes índices tradicionalmente são usados para comparar elementos discretos, mas aplicando o conceito de multi-conjutos, é possível generalizá-los para serem aplicáveis a valores reais.

Nesta primeira abordagem, buscou-se realizar uma comparação entre os três métodos de comparação de vetores mencionados. O objetivo é mostrar a diferença de aplicabilidade de cada um e as conclusões biológicas que podem ser retiradas através deles.

2.0.1 Dados Transcricionais e Metabolômicos

Os dados transcricionais foram obtidos a partir de um experimento envolvendo H. salinarum NRC-1 que foi exposto continuamente à luz em um ambiente anaeróbico por 72 horas. O experimento, obtido do banco de dados NCBI, está indicado pelo número de acesso GEO GSE7712. Ele possui quatro "spots" para 2400 sequências únicas de genes e aplica uma técnica de dye-swap para corrigir o viés do pigmento de coloração (25). Amostras foram coletadas a cada 3 horas, resultando em 25 pontos temporais na série. Para obter resultados mais confiáveis, a mediana dos valores dos 4 spots foi calculada para cada um dos 468 genes contidos na rede para evitar o efeito de pontos fora da curva. Genes presentes no experimento que não faziam parte da rede foram desconsiderados. Os pontos de dados nos experimentos incluiam um sinal indicando a confiabilidade da medida: "I" para medidas confiáveis, "J" para intermediárias, e "K" para inconfiáveis. "Spots" marcados pelo sinal "K" foram desconsiderados também.

Para este experimento, foram utilizadas as redes de interação enzima-enzima já mencionadas, utilizando-se como base de evidência de interação entre elas o compartilhamento de metabólitos iguais. Nestas redes, portanto, os nós representam as enzimas do ser vivo e as arestas entre tais nós são atribuídas de acordo com os metabólitos que elas compartilham. Supondo, por exemplo, duas enzimas hipotéticas ENZ_1 e ENZ_2, que possuem como metabólitos as moléculas A,B,C e B,C,D respectivamente. Como estas enzimas compartilham entre si o metabólito B e C podemos estabelecer entre elas uma aresta (Figura 5).

A condição mínima para o estabelecimento de uma aresta entre duas enzimas é que elas possuam ao mínimo 1 metabólito em comum. Para a montagem dessa rede, utilizou-se um experimento de reconstrução do reatoma do *Halobacterium salinarum*. Esta reconstrução usou um total de 559 reações envolvendo 468 ORFs, todas sendo codificadoras para a síntese de enzimas (22).

A rede resultante deste experimento, utilizando todos os metabólitos contidos no banco de dados, é altamente interconectada com o caminho mais curto médio entre todos os nós sendo menor que 2. Há um total de 41266 arestas e grau médio é de 176.35. A maior parte destas arestas são estabelecidas a partir do par ATP/ADP, os cofatores redox NAD+/NADH e o par H2O/H+. Estas moléculas são fundamentais para diversas reações metabólicas, elas são referidas na biologia molecular como "moedas de energia" por participarem em diversas vias metabólicas com funções distintas e que requerem uma fonte de energia útil para ocorrerem. Elas podem influenciar drasticamente no resultado final da extração de vias co-expressas ao conectar enzimas biologicamente não-relacionadas através de sua alta presença. Por esta razão, é importante experimentar com quais metabólitos trabalhar para obter os resultados mais coerentes.

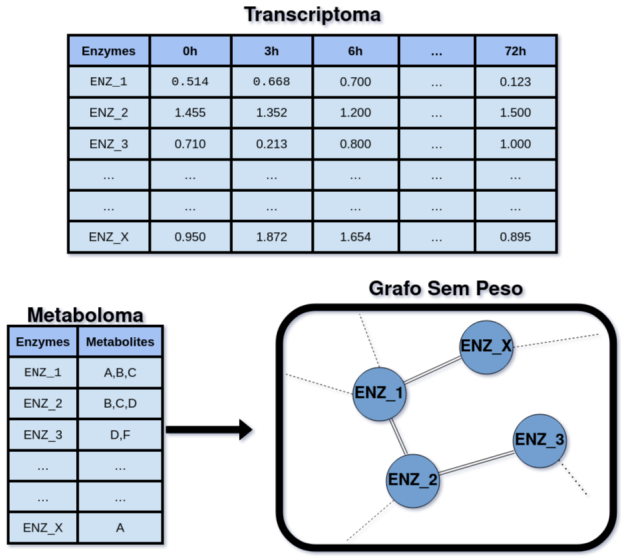

Figura 5 – Tipos de Bancos de Dados utilizados no projeto. Na figura estão mostrados exemplos de dois experimentos: um de transcriptoma e outro de metaboloma. A partir do experimento de metaboloma é montado o grafo de interação enzimática sem peso, e com o experimento de transcriptoma atribui-se peso às arestas desse grafo com base nas enzimas que compartilham metabólitos.

Fonte: Elaborada pelo autor

2.0.2 Métodos de Correlação de Vetores

A medida utilizada para comparar dois vetores de expressão no método descrito por Patil e Nielsen (26) é o coeficiente de correlação de Pearson, também conhecido como r de Pearson. Isto é uma prática tradicional, pois o r de Pearson é o coeficiente de correlação mais utilizado e fácil de ser calculado. Porém, ele possui limitações. Ele pode ser altamente suscetível aos efeitos de ruídos nas medidas devido a sua sensibilidade a pontos fora da curva, e dados de microarray tendem a ser bem ruidosos.(Schober et al., 2018) O método calcula o valor do r de Pearson para os vetores de expressão das enzimas/nós que compartilham um metabólito/aresta, dado pela equação 2.1.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{2.1}$$

O r de Pearson produz valores que vão de -1 a 1, com -1 sendo uma correlação inversa perfeita e 1 uma correlação direta perfeita. Para avaliar a significância destas correlações, uma função de distribuição acumulativa inversa é aplicada ao valor absoluto do r de Pearson (equação 2.2) para obter os Z valores das arestas (Z_{es}) em uma distribuição normal. Estes Z valores, que variam de infinito negativo a infinito positivo, são usados para determinar a quantos desvios padrões uma medida particular está da média.

$$Z_{es} = \theta^{-1} |\rho_e| \tag{2.2}$$

Outra medida de correlação tradicional, no entanto mais computacionalmente custosa, é o r de Spearman. Esta medida faz um ranqueamento das instâncias nos vetores ($\vec{X}$, $\vec{Y}$) e então realiza o mesmo cálculo do r de Pearson. Este método contorna a exigência do r de Pearson de que a correlação entre as instâncias seja linear para obter um grau de correlação maior. Ao invés disso, o r de Spearman quantifica a correlação entre qualquer crescimento ou decrescimento monotônico entre vetores.

$$r_s = \rho_{R(X),R(Y)} = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \tag{2.3}$$

A inovação deste trabalho está na aplicação do índice de coincidência (coindex). Este indíce é um coeficiente de similaridade recentemente desenvolvido que tem mostrado resultados promissores na análise de redes complexas. Ele é o produto entre os índices de Jaccard e Interioridade (o índice de interioridade é também conhecido como índice de sobreposição).(18) Tradicionalmente, estes índices são aplicados utilizando a propriedade de conjuntos, cujos elementos são todos distintos e comparáveis apenas em relação a seus equivalentes. No entanto, ao aplicar o conceito de multi-conjuntos (multisets), é possível generalizar os índices de Jaccard e de Interioridade para vetores contendo valores reais.(18)

$$J_R(\vec{x}, \vec{y}) = \frac{\sum_{i \in S} s_{x_i y_i} min\left\{s_{x_i} x_i, s_{y_i} y_i\right\}}{\sum_{i \in S} max\left\{s_{x_i} x_i, s_{y_i} y_i\right\}} \tag{2.4}$$

$$I_R(\vec{x}, \vec{y}) = \frac{\sum_{i \in S} min\left\{s_{x_i} x_i, s_{y_i} y_i\right\}}{min\left\{\sum_{i \in S} s_{x_i} x_i, \sum_{i \in S} s_{y_i} y_i\right\}} \tag{2.5}$$

$$C_R(\vec{x}, \vec{y}) = J_R(\vec{x}, \vec{y}) I_R(\vec{x}, \vec{y}) \tag{2.6}$$

Nas equações (2.4) e (2.5), x_i e y_i correspondem à componente i dos vetores ($\vec{x}$,$\vec{y}$) de tamanho S. s_{x_i} e s_{y_i} correspondem ao sinal de x_i e y_i e $s_{x_i y_i} = s_{x_i} s_{y_i}$. Estes indíces de valor real são então multiplicados juntos para formar o índice de coincidência na equação (2.6).

O coindex mostrou-se capaz de prover uma medida mais estrita e detalhada da similaridade entre vetores comparado a medidas mais tradicionais.(18) No entanto, é importante notar que a interpretação do seu "output" difere do r de Pearson e r de Spearman. Os valores do "output" também estão limitados por -1 e 1, mas o mínimo do valor representa uma total dissimilaridade entre os vetores, ao invés de uma correlação inversa perfeita. Isto torna a equação (2.2) inapropriada para esta medida, pois ela pode agregar vetores tidos como altamente dissimilares com aqueles altamente similares entre si.

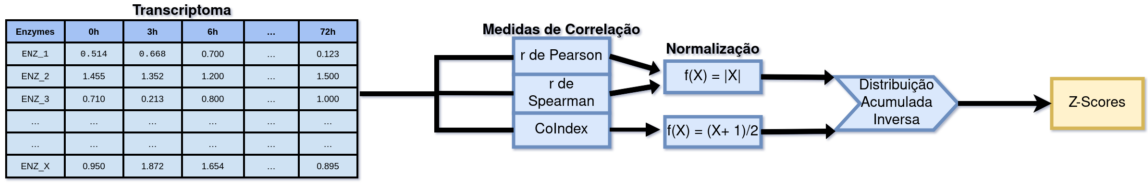

Figura 6 – Fluxograma da atribuição de peso às arestas do grafo de interação enzima-enzima. Nota-se a diferença entre a normalização clássica utilizada por Patil e Nielsen, e a normalização utilizada para o método do índice de coincidência.

Fonte: Elaborada pelo autor

Ele pode também distorcer a distribuição dos valores para 0, gerando uma distribuição gaussiana assimétrica que distorce o significado dos Z valores. Por estas razões, nós escolhemos mapear os "outputs" do índice de coincidência da seguinte forma:

$$Z_e = \theta^{-1}\left(\left(C_{Re} + 1\right)/2\right) \tag{2.7}$$

O fluxograma da Figura 6 resume o processo de atribuição de pesos às arestas do grafo descrito até aqui, mostrando como foram utilizados os três métodos de correlação e similaridade entre os vetores dos dados transcriptomicos do *Halobacterium salinarum*.

Na figura 7, demonstramos como o índice de coincidência é mais estringente em diversos cenários do que o r de Pearson e o r de Spearman. Esta demonstração foi feita utilizando-se de vetores gerados artificialmente e adicionando ruído aleatório aos valores. O resultado final mostra que o coindex não só é mais sensível a ruído do que os outros, como também mostra maior sensibilidade para não-linearidade (mais ainda do que o r de Pearson) e a variações na escala de variáveis lineares, um aspecto que as outras medidas não são capazes de levar em consideração.

2.0.3 Algoritmo de Anelamento Simulado

Para averiguar o quão bem cada medida de correlação performa sobre os dados transcricionais, buscou-se encontrar subredes altamente similares/correlacionadas que poderiam providenciar esclarecimentos sobre os mecanismos sendo co-expressos pela arqueia, elucinando as modificações do seu metabolismo a partir de sua exposição à luz. Para obter isto, realizou-se os mesmos cálculos demonstrados em Patil e Nielsen (26) para pontuar uma subrede de tamanho k:

$$Z_s^k = \frac{1}{\sqrt{k}_{e \in s}} \sum Z_{ei} \tag{2.8}$$

Na equação (2.8), Z_{ei} representa o Z valor de cada aresta na subrede conectada. Porém, este valor precisa ser ajustado de acordo com a distribuição de fundo do valor Z de

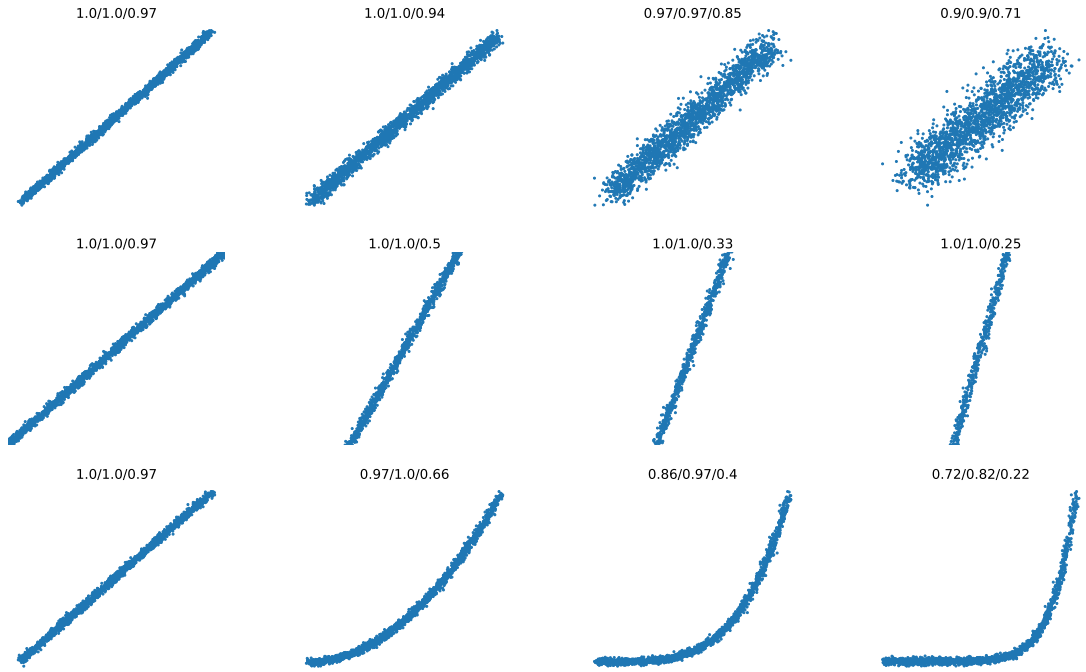

Figura 7 – Comparação das 3 medidas usando dois vetores exemplo, X e Y, com 2000 valores cada. Os valores no topo de cada curva representam o r de Pearson, r de Spearman, e índice de coincidência respectivamente. A primeira linha mostra a relação linear entre X e Y com mais ruído aleatório em cada iteração. A segunda linha mostra uma correlação linear muito forte mas com um coeficiente angular maior em cada iteração. A terceira e última linha mostra um relação cada vez menos linear entre os vetores.

Fonte: Elaborada pelo autor

outras subredes conectadas de mesmo tamanho k. Este ajuste é necessário para garantir que subredes maiores não acabem com Z valores maiores devido a sua maior prevalência. Este ajuste pode ser feito da seguinte maneira:

$$Z_{normalized\ s} = \frac{Z_s^k - \bar{Z^k}}{\sigma^k} \tag{2.9}$$

Com esta medida, é possível determinar se uma subrede resultante é afetada pelas condições experimentais mais do que o esperado sem um viés em relação ao tamanho. A média $\bar{Z^k}$ e o desvio padrão σ^k podem ser obtidos pela amostragem de subredes conectadas dentro da rede principal. Desta forma, nós podemos obter uma determinada quantia de amostras (em torno de 200) com as quais podemos calcular a média e o desvio padrão dos Z valores de subredes de tamanho k para normalizar em relação a variações do tamanho das subredes. Assumindo que uma rede possui N_{max} nós no total na sua maior componente, há N_{max} - 1 possíveis tamanhos de subredes conectadas, e então, N_{max} médias e desvios padrões. Subredes de tamanho k=1 não possuem arestas, portanto elas não podem ser pontuadas através deste método. O fluxograma apresentado na Figura 8 exemplifica como a média e o desvio padrão foram obtidos através de uma amostragem de componentes

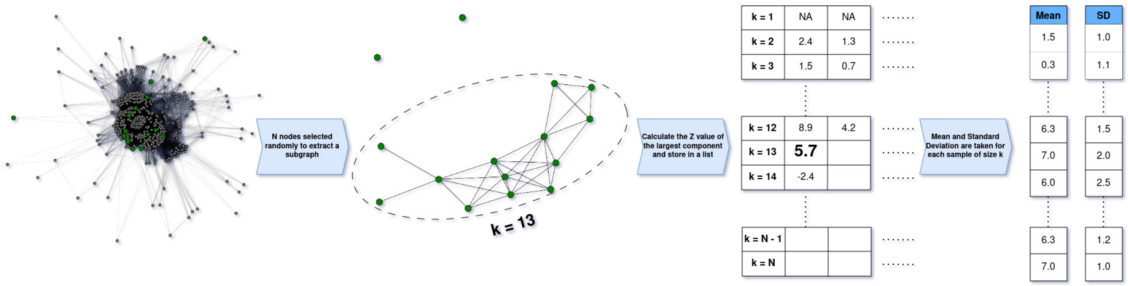

Figura 8 – Método de amostragem usado para realizar a normalização dos Z valores das redes com base no seus tamanhos. Primeiro, N nós são selecionados aleatoriamente, depois disso, a maior componente conectada destes nós é extraída e seu Z valor é calculado. O Z valor é então armazenado em uma lista junto com outras amostras de mesmo tamanho "k". Com as amostras coletadas, a média e o desvio padrão de todas as listas é calculado, gerando duas listas que serão usadas para a normalização das redes. Isto é feito 200 vezes para cada valor possível de N, indo de 2 até N_{max} .

Fonte: Elaborada pelo autor

conectadas randômicas de tamanho k.

Para identificar subredes enriquecidas dentro do grafo principal de interações enzima-enzima, nós usamos o algoritmo de anelamento simulado. (23) Este método faz uma busca heurística por máximos locais dentro da rede principal, encontrando grupos de genes conectados com um $Z_{\text{normalized s}}$ maior do que a média. O método inicialmente atribui um estado de 1 ou 0 para os nós aleatoriamente, e então "oculta" os nós com estado 0 e busca a componente conectada de maior pontuação (ou subrede conectada) dentro dos nós com estado igual a 1. Uma vez que a componente com a pontuação mais alta é encontrada, o algoritmo muda aleatoriamente o estado de um nó da rede como um todo e faz a busca da componente conexa de maior pontuação novamente. Se a pontuação for maior que a componente anterior, o estado alterado se mantém; se ela for menor, a probabilidade do estado alterado ser mantido é determinado pela função $e^{\frac{Z_{i+1}-Z_i}{T}}$, onde T é a "temperatura" atual do sistema. Esta função retorna um valor entre 0 e 1. Se o valor retornado for menor que um número aleatório do intervalo [0,1], então o nó é desalterado; caso contrário, ele permanece alterado. Este processo ajuda o algoritmo a evitar "máximos falsos", permitindo-o realizar uma busca mais aprofundada na rede.

Esta busca é realizada N vezes, com o valor da temperatura T decaindo geometricamente de T_i até $T_f = 0.01$. Depois disso, uma nova busca é conduzida com T = 0 para garantir que o máximo local foi encontrado (um processo conhecido como "simulated quenching").

Nós usamos algumas das heurísticas descritas em Ideker et al. para obter resultados melhores. Uma destas heuristicas envolve buscar por um número M de componentes conectadas simultaneamente, maximizando a soma das pontuações das subredes ao invés

dos Z valores individuais delas. Este método tem mostrado resultados melhores. (23)

Outra heurística envolve uma exceção para nós considerados “hubs”. Estes nós podem acabar conectando muitas componentes juntas que, quando separadas são significativas, mas têm uma pontuação menor quando conectadas devido a relações insignificantes que o “hub” acaba introduzindo. Para evitar estas limitações, se um nó excede um determinado grau mínimo (d_{min}), apenas os vizinhos pertencentes à componente conectada de maior pontuação são mantidos “ligados” quando o nó hub é “ligado”. Os vizinhos que não satisfazerem esta condição são colocados em “off”. Isto minimiza o efeito que hubs podem ter no processo de anelamento simulado restringindo-o apenas à componente de maior pontuação, evitando assim penalizar as componentes menores.

Otimização dos parâmetros T_i , N, M e d_{min} foi realizada para conseguir obter o melhor desempenho possível através de diferentes simulações com valores variados para cada um. Além disso, como este método é heurístico, ele não garante o resultado ótimo, então 10 processos foram rodados em paralelo depois que os melhores parâmetros foram escolhidos. O processo que rendeu o maior Z valor teve as componentes de maior pontuação selecionadas e analisadas.

Para determinar se os genes que compõem cada componente conectada contém qualquer tipo de informação biológica, uma análise de enriquecimento de termos ontológicos de genes foi conduzida através do banco de dados STRING para todas as subredes com as maiores pontuações. O enriquecimento foi feito utilizando apenas os genes pertencentes à rede principal enzima-enzima como plano de fundo para as análises. Desta forma evitamos fazer conclusões erradas com base em um conhecimento a priori inválido.(27, 28)

O STRINGdb inclui 445 dos 468 genes na rede principal. Estes genes foram correspondidos aos devidos IDs do banco de dados e uma análise de termos GO com FDR (taxa de descobrimento falso) foi realizada para todas as componentes identificadas pelo algoritmo do anelamento simulado. O banco de dados KEGG também foi utilizado para o enriquecimento de vias metabólicas.(29)

O processo da busca dos máximos locais da rede de interação enzima-enzima, agora com pesos atribuídos as arestas, seguido da obtenção dos termos GO e KEGG presentes em seus genes, está descrito na Figura 9.

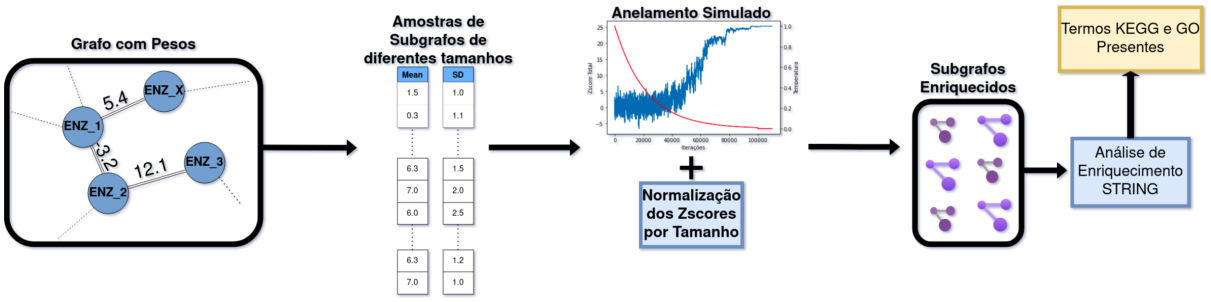

Figura 9 – Fluxograma mostrando a obtenção dos máximos locais do grafo, seguido da normalização e do enriquecimento de termos GO e KEGG. Os valores das médias e dos desvios padrões mostrados foram obtidos através do método descrito na Figura 8

. Fonte: Elaborada pelo autor

3 RESULTADOS E DISCUSSÕES SOBRE A MEDIDA ÍNDICE DE COINCIDÊNCIA

Neste capítulo serão apresentados os resultados obtidos para a aplicação do índice de coincidência sobre os grafos construídos utilizando os experimentos de quantificação de transcrição e também de metabolômica da arqueia *Halobacterium salinarum*.

3.0.1 Otimizações para o Algoritmo de Anelamento Simulado

Para obter o melhor resultado possível, o algoritmo de anelamento simulado precisa ser otimizado, pois ele é heurístico, não-determinístico, e multivariável. Além do mais, como estamos experimentando com uma nova medida de similaridade, é importante determinar quais parâmetros melhor se adequam a cada um deles. Para fazer isto, nós estabelecemos parâmetros "padrões" para o algoritmo ($N = 10^5$, $m = 20$, $T_i = 1$, $d_{min} = 100$) e variamos um de cada vez. O parâmetro que resultou no maior Z valor foi então usado para buscar subredes enriquecidas. O resultado pode ser visto na **Tabela 2**.

A tabela indica que os parâmetros ótimos para o r de Pearson e Spearman são os mesmo ($T_1 = 1$, $d_{min} = 100$, $N = 10^6$, $m = 20$) e os Z valores resultantes não são significativamente diferentes também. O coindex, por outro lado, geralmente gera Z valores maiores, e o valor do grau mínimo ótimo para considerar um nó um "hub" é $d_{min} = 200$. A variação no valor ótimo do d_{min} implica que subredes enriquecidas do coindex são mais conectadas e têm mais nós comparada as subredes do r de Pearson e Spearman.

A busca por máximos locais também foi conduzida em um segundo grafo de interação enzima-enzima que excluí os metabólitos mais comuns. Já foi demonstrado que alguns metabólitos podem atrapalhar a busca de vias metabólicas ao conectar genes não relacionados juntos, estabelecendo um excesso de conexões triviais.(21) Isto também leva a criação de muitos "hubs", necessitando uma busca mais cautelosa e, consequentemente, uma busca mais computacionalmente intensiva.

Para determinar quais metabólitos remover, nós contamos quantas enzimas possuem reações com um metabólito particular. Se o número de enzimas que contém tal metabólito exceder 100 (ao redor de $\frac{1}{5}$ de todas as enzimas no banco de dados), ele foi considerado comum e foi removido do processo de construção de redes (mostrado na Figura 10). Existem métodos mais sofisticados que buscam contornar este procedimento pois ele tende a ser muito arbitrário, como busca de grupos químicos. No entanto, nós estamos primariamente interessados em avaliar a performance de diferentes medidas de similaridade em uma rede de topologia distinta, então este método será suficiente.

Como esperado, os metabólitos mais comuns são primariamente "tokens de energia" como ATP/ADP. Participantes de reações de fosforilação e hidrogenização também estavam

Tabela 1 – Valores das somas dos Z-valores das componentes conectadas encontradas com a respectiva mudança das variáveis. O maior valor está destacado em negrito.

Parâmetros		**Z valor total das Componentes**		
Variáveis	**Valor**	**Pearson**	**Spearman**	**Coindex**
T_1	0.01	11 ± 5	13 ± 4	16 ± 7
	1	**28 ± 1**	**27.1 ± 0.8**	**32.0 ± 0.9**
	2	28 ± 1	27 ± 1	32 ± 1
	5	28 ± 1	26.9 ± 0.9	32 ± 1
	10	27 ± 1	26 ± 1	30 ± 2
d_{min}	25	13 ± 1	14 ± 2	16.9 ± 0.7
	50	23.2 ± 0.6	23.4 ± 0.7	25.7 ± 0.9
	100	**28 ± 0.8**	**27.1 ± 0.8**	32 ± 1
	200	25 ± 6	27 ± 6	**42 ± 1**
	300	14 ± 3	18 ± 5	34 ± 3
N	10000	16 ± 2	15 ± 1	19 ± 3
	100000	26 ± 1	25 ± 2	29 ± 2
	1000000	28.4 ± 0.9	27.1 ± 0.8	32 ± 1
	10000000	**29.1 ± 0.5**	**27.9 ± 0.1**	**33 ± 1**
M	1	7 ± 1	7 ± 1	5 ± 1
	3	14.3 ± 0.5	13.4 ± 0.4	13.1 ± 0.8
	5	18.6 ± 0.3	17.6 ± 0.5	20.1 ± 0.4
	10	27.2 ± 0.9	24.7 ± 0.7	28.7 ± 0.6
	20	**29.1 ± 0.6**	**27.3 ± 0.9**	**31.4 ± 0.9**

Fonte: Elaborada pelo autor.

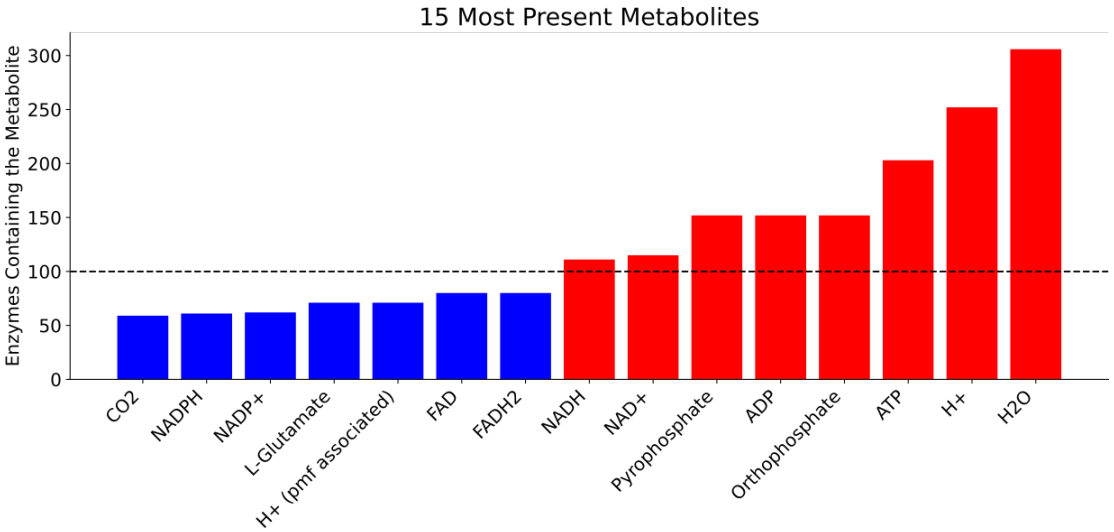

Figura 10 – Histograma mostrando os top 15 metabólitos mais presentes nas reações enzimáticas do *Halobacterium salinarum*. Em vermelho, estão destacados aqueles que estão presentes em mais de 100 enzimas distintas e, por isso, foram considerados comuns.

Fonte: Elaborada pelo autor

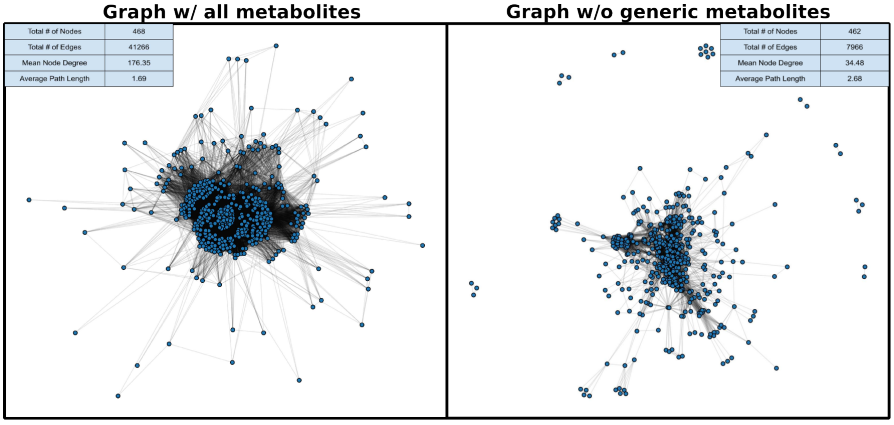

Figura 11 – Mudança na estrutura do grafo com a remoção das enzimas mais comuns. Nós que acabaram ficando sem nenhuma aresta foram removidos.

Fonte: Elaborada pelo autor

Tabela 2 – Valores das somas dos Z-valores das componentes conectadas encontradas com a respectiva mudança nos valores das variáveis. O maior valor está destacado em negrito.

Parâmetros		**Z valor total das Componentes**		
Variáveis	**Valor**	**Pearson**	**Spearman**	**Coindex**
d_{min}	5	33.2 ± 0.3	34.1 ± 0.2	38.2 ± 0.3
	10	40.8 ± 0.9	42.8 ± 0.8	46 ± 1
	25	**52.8 ± 0.8**	**51.6 ± 0.4**	72 ± 1
	100	43 ± 4	44 ± 4	**76 ± 5**
M	5	22 ± 2	22 ± 1	42 ± 5
	10	36 ± 2	35 ± 2	58 ± 6
	20	54.0 ± 0.4	54.0 ± 0.7	79 ± 4
	40	68.9 ± 0.7	70.2 ± 0.2	**91 ± 3**
	80	**69.0 ± 0.4**	**70.7 ± 0.7**	91 ± 3

Fonte: Elaborada pelo autor.

entre eles. Com a remoção deles, o novo grafo tornou-se bem menos conectado, com o número total de arestas caindo para menos que um quarto do tamanho original e com alguns nós sendo removidos devido ao fato que eles não possuem nenhum tipo de relação com nenhum outro nó (mostrado na Figura 11). O tamanho do caminho médio aumentou (de 1.69 para 2.68), mas o novo grafo ainda está altamente conectado.

Como a topologia da rede mudou, suas propriedades, como o número de "hubs" e de componentes conectadas também mudaram. Por conta disso, precisamos reotimizar o algoritmo. Desta vez, porém, nós precisamos apenas de um afinamento dos parâmetros afetados pela mudança de arestas, pois a mudança no número de nós não foi significativa. Estes parâmetros são os números de componentes conectadas buscadas simultaneamente (M) e o valor de grau mínimo para considerar um nó um "hub" (d_{min}). Os Z valores resultantes estão mostrados na **Tabela 2**.

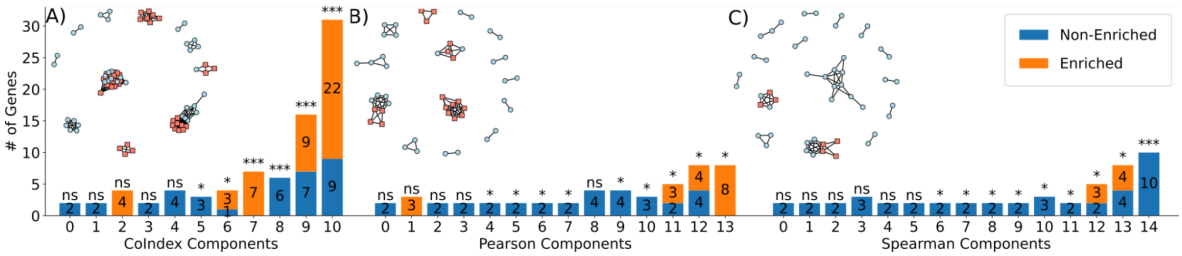

Figura 12 – Componentes obtidas para a rede com todos os metabólitos. A) Mostra as componentes do Índice de Coincidência, B) as componentes do rho de Pearson e C) as componentes do rho de Spearman. P-valores foram obtidos através do Z-score de cada componente e representados por *: ns > 0.05; 0.05<*<0.001; 0.001<**<0.0001;***<0.0001.

Fonte: Elaborada pelo autor

3.0.2 Subgrafos Altamente Correlacionados

Depois da otimização dos parâmetros, o algoritmo foi rodado novamente para todas as 3 medidas, desta vez, escolhendo o processo que obteve o maior Z valor possível. Para a rede contendo todos os metabólitos, todas as três medidas renderam em torno de 10 subredes enriquecidas, mas nem todas elas continham informações biológicas relevantes ou tinham um Z valor significativo. A significância das subredes individuais foi calculada através de um teste de p-valor unidirecional e utilizou-se também de correções para falsos descobrimentos (False Discovery Rate). É importante mencionar que, como a normalização das pontuações é feita através de uma amostragem randômica de todas as redes possíveis de um certo tamanho, o valor pode variar em outras simulações. No entanto, para o nosso caso onde 200 amostras foram coletadas por tamanho a variação dos Z valores das redes significativas foi menor que 5%.

A Figura 12 ilustra quantas componentes cada medida foi capaz de descobrir, e quais incluem um grupo de genes que possuem informações biológicas associadas a eles. Além disso, as configurações das subredes para cada uma das medidas também estão apresentadas logo acima do gráfico de barras.

Podemos ver que o coindex conseguiu identificar as maiores componentes conectadas e também é o que tem mais componentes enriquecidas com informação biológica entre as três medidas. O r de Pearson é o segundo melhor neste quesito, com o r de Spearman sendo o menos efetivo. Embora o algoritmo rodou para buscar 20 componentes simultaneamente, coindex encontrou apenas 11, Pearson encontrou 14, e Spearman 15. A maioria destas componentes compõem um número bem pequeno de genes que não esclarecem nenhum tipo de informação em relação as funções das enzimas co-expressas. Estas componentes pequenas são também as que contém os menores Z valores que resultaram em P-valores insignificantes após a correção por FDR.

Após examinar os genes presentes nas componentes de cada uma das medidas, nós

Tabela 3 – Termos ontológicos enriquecidos mais significativos separados por seus respectivos componentes de origem. O p-Valor foi ajustado usando FDR.

Measurement/ Component	**Enriched GO term**	**Genes**	**p Value**
coindex/2	Cobalamin binding	VNG_0481G,VNG_-0653G,mcmA2	0.0154
	Valine, leucine and isoleucine degradation	VNG_0478C,VNG_-0481G,VNG_0653G, mcmA2	0.00024
	Carbon fixation pathways in prokaryotes	VNG_0478C,VNG_-0481G,VNG_0653G, mcmA2	0.00036
coindex/6	Cobalamin biosynthetic process, and photosynthesis	cbiT,cobH,cbiJ	0.0375
coindex/7	Mo-molybdopterin cofactor biosynthetic process	moaE,moeA1,gdb,mobB	0.0063
	Folate biosynthesis	moaE,moeA1,gdb,mobB	0.0014
	Sulfur relay system	moaE,moeB,mobB	0.0095
coindex/9	Kinase activity	mvk,argB,aroK,suk,thiL, prsA,lysC,thiD, glcK	7.43e-06
coindex/10	Oxidoreductase activity	serA3,yafB,VNG_-0468C,inb,cyb,petA, nolC,nuoL,nuoM,VNG_-0730C,yajO, adh3,trxB2,fabG, serA2,aad,etfA,adh1, txrB3,glcD,bchP	3.44e-05
Pearson/1	Valine, leucine and isoleucine degradation	VNG_0478C,VNG_-0481G,mcmA2	0.0056
	Carbon fixation pathways in prokaryotes	VNG_0478C,VNG_-0481G,mcmA2	0.006
Pearson/11	Succinate dehydrogenase activity	sdhA,sdhB,sdhC	0.0122
Pearson/12	Cobalamin biosynthetic process, and photosynthesis	cbiT,cbiF,cobH,cbiJ	0.02
Pearson/13	Mo-molybdopterin cofactor biosynthetic process	moaE,moeA1,gdb,mobB	0.0124
	Folate biosynthesis	moaE,moeA1,gdb,mobB	0.0028
	Sulfur relay system	moaE,moeB,mobB	0.015
Spearman/12	Aminotransferase	aspB1,aspC2,VNG_-1524C	0.0109
Spearman/13	Cobalamin biosynthetic process, and photosynthesis	cbiT,cobH,cbiC,cbiJ	0.02

Fonte: Elaborada pelo autor.

podemos checar que há um certo overlap entre eles (Tabela 3). Uma via que está presente em todas as três medidas é a via de síntese de cobalamina. Cobalamina é um co-fator proteico mais comumente conhecida como vitamina B12.(30) Este co-fator está associado com enzimas responsáveis pela degradação de aminoácidos e fixação de carbono,(31) outra via que foi demarcada por coindex e Pearson, mas não o Spearman.

Um aspecto notável do índice de coincidência é evidente na sua componente 2, onde 3 genes ligantes em cobalamina estiveram presentes (VNG0481G, VNG0653G, mcmA2). Este componente é análogo ao componente 1 da medida de Pearson, mas ele não incluiu o termo ontológico de ligação a cobalamina devido ao seu tamanho menor que incluí apenas dois dos três genes ligantes a cobalamina; No entanto, tanto o coindex (componente 2) como o r de Pearson (componente 1) obtiveram Z valores bem baixos. Em uma situação onde esta função biológica não é conhecida (em uma investigação sem informação biológica conhecida previamente), isto pode ser descartado como estatisticamente insignificante.

A medida coindex também apresenta certas desvantagens. Suas duas maiores e mais significativas componentes, 9 e 10, contém genes relacionados a atividade de quinase e oxiredutase respectivamente. Estas reações são dependentes de metabólitos como os fosfatos e NADH/NAD+, que são metabólitos extremamente comuns e estão presentes em muitas vias distintas com diferentes funções biológicas. Isto significa que a informação obtida a partir destas enzimas é muito genérica, e não providencia um entendimento mais claro das possíveis vias presentes no metabolismo do *Halobacterium salinarum.* Os experimentos feitos na rede metabólica com os metabólitos mais comuns removidos irá mostrar como cada medida foi afetada e como eles podem revelar fenômenos biológicos mais específicos.

O segundo experimento identificou em torno de 3 vezes mais componentes que o primeiro (Figura 13). Porém, muitos deles ainda são muito pequenos. Como o corretor FDR está diretamente ligado com o número de instâncias, muitos deles não tiveram um Z valor alto o suficiente para obter um p-valor significativo. Isto resultou em todas as componentes identificadas pela medida de correlação de Pearson como sendo consideradas não significativas (Figura 13B). Embora Pearson tenha obtido um Z valor similar total parecido com o Spearman, cada componente individualmente não foi capaz de passar no teste FDR. Spearman conseguiu obter algumas componentes significativas, mas o coindex mostrou o melhor desempenho, com 7 subredes grandes e enriquecidas que conseguiram elucidar mais claramente as vias químicas da arqueia.

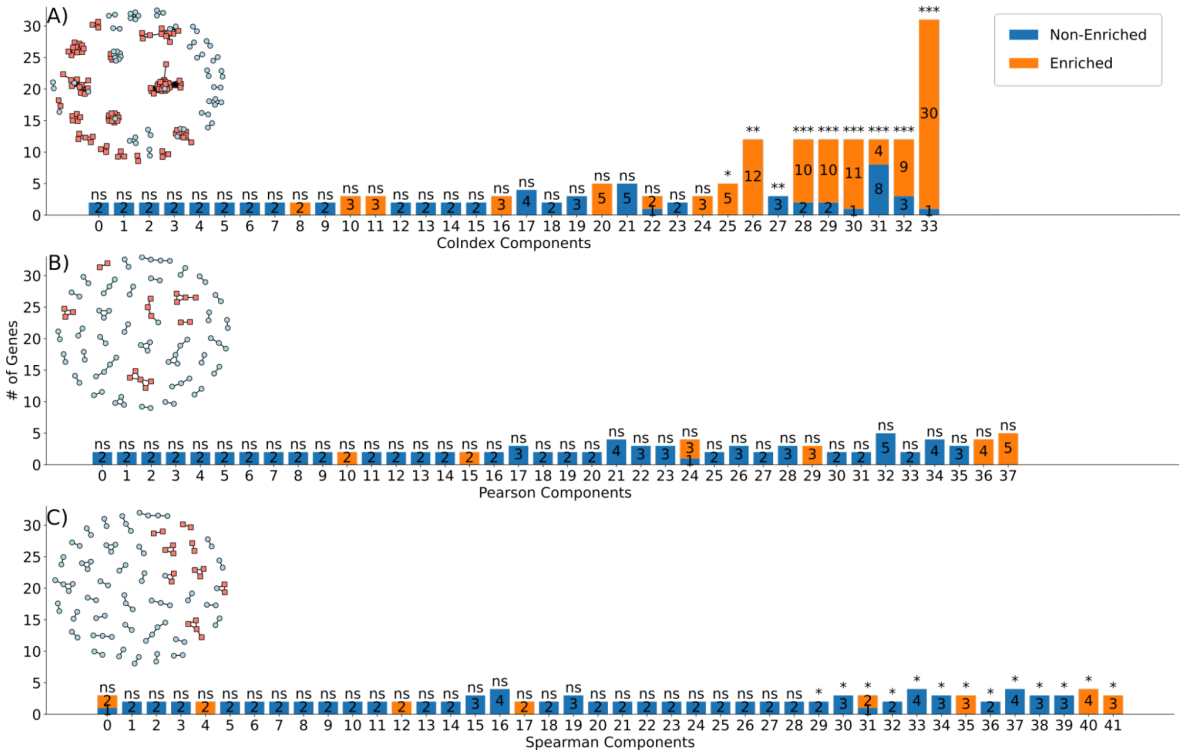

Figura 13 – Componentes obtidas para a rede sem os metabólitos mais comuns. A) Mostra as componentes do Índice de Coincidência, B) as componentes do rho de Pearson e C) as componentes do rho de Spearman. P-valores foram obtidos através do Z-score de cada componente e representados por *: ns $>$ 0.05; 0.05$<$*$<$0.001; 0.001$<$**$<$0.0001;***$<$0.0001.

Fonte: Elaborada pelo autor

Nossa hipótese é que o coindex foi capaz de obter os melhores resultados por conta de sua estringência ao invés de apesar dela. Pearson e Spearman podem obter resultados maiores para relações individuais entre genes, mas isto acaba favorecendo clusters pequenos e altamente correlacionados que não permitem uma interpretação razoável das vias que estão sendo co-expressas. Coindex penaliza dados com ruído, não-linearidade, e variações na escala mais fortemente, e isso diminui a pontuação de todas as relações significativamente, evitando a criação de "picos" com poucos genes. Ao invés disso, esta nova medida identifica locais máximos que são mais abrangentes, que incluem mais genes e permitem construir uma imagem mais completa das reações enzimáticas.

Ao examinar os termos do segundo grafo (Tabela 4), nós agora podemos ver de que modo ele realça as vias coexpressas do *Halobacterium salinarum*. Nós agora temos a via de fosforilação oxidativa. Ela é responsável pela aquisição de metabólitos energéticos através da cadeia transportadora de elétrons (ETC), uma combinação de complexos transmembrana que geram um gradiente de protons e formam NADH. O gradiente é então usado pela ATP sintetase para sintetizar ATP.(32, 33) Além do mais, se nós levarmos em consideração as condições experimentais que a arqueia foi exposta, isto é, a contínua exposição anaeróbica

Tabela 4 – Termos ontológicos enriquecidos mais significativos separados por seus respectivos componentes de origem. Como o segundo experimento resultou em diversas componentes, apenas aqueles que obtiveram um Z-valor significativo e foram enriquecidos com vias metabólicas do KEGG estão sendo mostrados. O p-Valor foi ajustado usando FDR.

Measurement/ Component	**Enriched GO term**	**Genes**	**p Value**
coindex/25	Folate biosynthesis	moaE,moeA1,gdb,mobB	0.00021
coindex/26	Amino sugar and nucleotide sugar metabolism	ugd,galE2,gmd,galE1,udg1, pmu1,pmm	4.63e-06
	Streptomycin biosynthesis	graD2,graD6,graD3,graD1, graD4,pmu1	1.06e-05
coindex/28	Carbon fixation pathways in prokaryotes	VNG_0478C, VNG_0481G, mcmA2,mdhA,can	0.0121
coindex/29	Porphyrin and chlorophyll metabolism	cbiT,cbiF,cobH,cbiC, cbiJ,uroM	0.0028
coindex/30	Aminoacyl-tRNA biosynthesis	metS,thrS,lysS,serS, ileS,trpS2, tyrS,pheS,valS	4.81e-08
coindex/33	Oxidative phosphorylation	petA,nolD,nuoJ1, VNG_0642C,nolC,nuoL,nuoM, coxA2,coxC,coxB1,sdhA, sdhD,sdhC, atpB,atpF,atpC,atpE, atpH,coxB2	3.08e-11
Spearman/31	Streptomycin biosynthesis	pmu1,glcK	0.0457
Spearman/35	Quorum sensing	dppB1,dppF,dppC1	0.0092
Spearman/40	Folate biosynthesis	moaE,moeA1,gdb,mobB	4.41e-05
Spearman/41	Amino sugar and nucleotide sugar metabolism	ugd,galE2,udg1	0.0031

Fonte: Elaborada pelo autor.

a luz, é seguro assumir que fotofosforilação anoxigênica está ocorrendo.(32) Em outras palavras, um elétron, que não foi derivado de um hidrólise com formação de oxigênio (como ocorre nas vias fotossintéticas mais comuns), está sendo excitado por luz em uma rodopsina e entrando no ETC.

Esta forma particular de fotofosforilação oxidativa permite a *H. salinarum* extrair eletrons de moléculas com potenciais químicos mais altos como o succinato e o acetil.(32) Estas moléculas podem ser obtidas através da degradação de aminoácidos, uma via que que está sendo co-expressa junto da fotofosforilação oxidativa, fortalecendo a hipótese de que esta forma de obtenção de energia está sendo expressa.

Alguns componentes que contém termos já presentes no primeiro experimento, como o coindex/29, incluem mais genes biologicamente correlacionados e produzem um resultado mais estatisticamente significativo. É também imteressante notar que cobalamina e folato

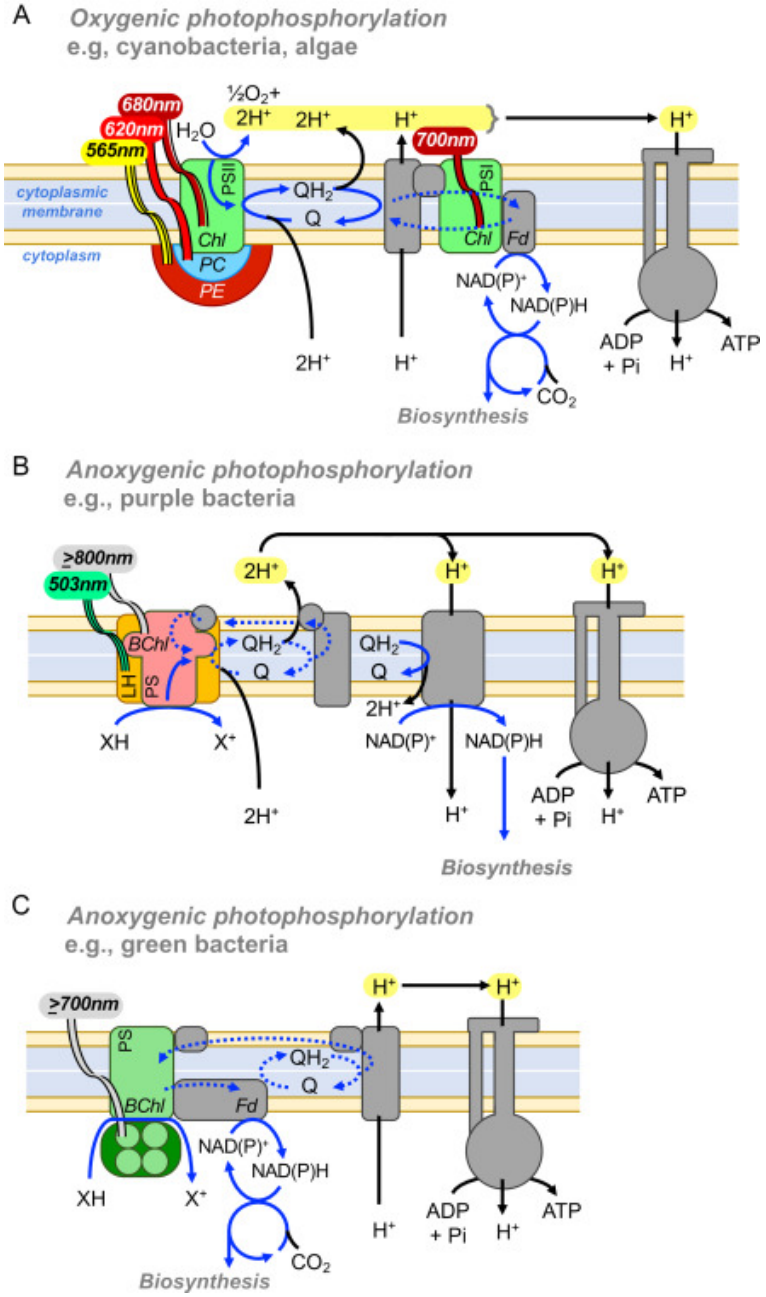

Figura 14 – Esquemas simplificados mostrando 3 tipos de fotofosforilação conhecidos que ocorrem em seres unicelulares como bactérias, arqueias e algas. O esquema análogo à via fotofosforilante presente na *Halobacterium salinarum* está demonstrado em B, onde pode-se observar uma rodopsina, uma enzima transmembrana de cor roxa, utilizando luz para realizar a hidrólise de um composto hipotético XH. Nota-se a diferença em relação a via mais tradicional presente em A, onde a obtenção de prótons se dá através da quebra de moléculas de água, com a liberação de oxigênio.

Fonte: Retirado de MCKINLAY *et al.* (32)

compartilham funções biológicas e uma deficiência em um deles resulta na deficiência do outro através de um fenômeno conhecido por “armadilha metil folato”.(34) Isto pode explicar o porquê das duas vias sintéticas estarem sendo expressas juntas, porque um desbalanço na concentração de uma delas pode resultar em uma deficência nos dois.

Das quatro componentes da medida de Spearman, apenas uma (Spearman/35) contém um termo único em relação ao coindex. Esta componente contém três genes que codificam para a formação de subunidades de um transportador de dipeptídeos com homólogos conhecidos em procariotos famosos por suas funções “quorum-sensing”.(35) Spearman/40 contém os mesmos genes de biossíntese de folato que o coindex/25, mas com um resultado mais significativo, porque ele contém menos genes que não estão relacionados a este processo. As outras duas componentes (Spearman/31 e Spearman/41) contém, separadamente, muitos dos genes presentess no coindex/26, com a exceção de glcK. Isto também demonstra a capacidade do coindex de agregar componentes menores em um grupo mais coesivo e compreensivo.

Entre as componentes que não possuem termos biológicos enriquecidos, mas cujas enzimas estão correlacionadas de maneira maior do que o esperado, há uma oportunidade de se realizar experimentos de bancada que busquem investigar se as relações enzimáticas presentes apontam para novas reações no organismo do *H. salinarum*. Entre componentes exemplares disto estão a CoIndex/31 e CoIndex/27 para o grafo sem os metabólitos mais comuns, e a componente Spearman/14 do grafo com todos os metabólitos. Todas elas possuem p-valor menor que 0.001, dando a elas uma alta confiabilidade estatística de correlação e duas delas (CoIndex/31 e Spearman/14) possuem 10 ou mais genes, o que dá um grupo de tamanho razoável de enzimas que podem ser exploradas em conjunto.

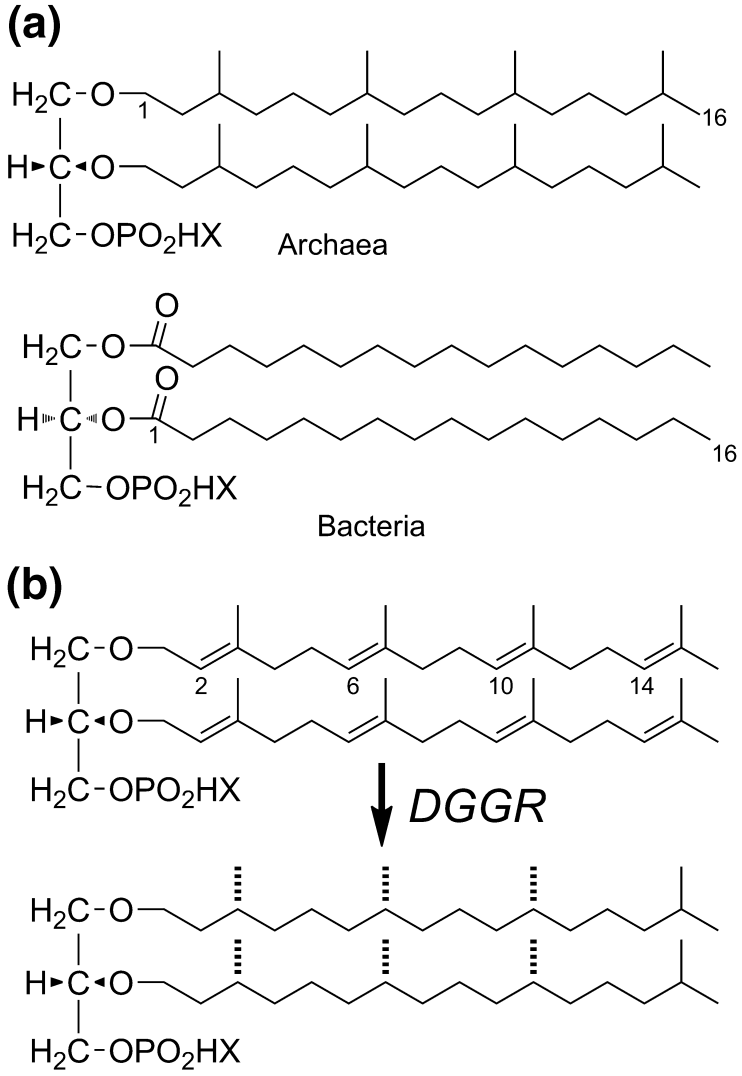

Figura 15 – A) Esquema simplificado mostrando as diferenças entre fosfolipídeos presentes na membrana celular de arqueias e bactérias. Nota-se que os fosfolipídeos das arqueias são ligados ao grupo fosfato através de um grupo quimíco éter, enquanto nas bactérias esta ligação se dá através de um grupo éster. Além disso, o fosfolipídeo das arqueias possuem insaturações ao longo de sua cadeia apolar que não estão presentes nos fosfolipídeos das bactérias. B) Mostra-se a ação enzimática da geranilgeranil redutase na hidrogenação das insaturações da cadeia apolar dos fosfolipídeos das arqueias.

Fonte: Adaptada de XU. *et al.* (36)

A componente CoIndex/31, por exemplo, possui como informação biológica enriquecida apenas que 4 de seus 12 genes possuem domínios de ligação a FAD. Mas, realizando uma investigação mais a fundo, é possível descobrir que entre estes genes estão enzimas hipotéticas com um domínio ortólogo a redutases de geranilgeranil (36, 37), uma molécula importante para o metabolismo de fosfolipídeos da membrana celular de arqueias, além de ser um metabólito intermediário na síntese de vitamina E e também carotenóides, importantes moléculas utilizadas na proteção contra radiação solar e oxidação (38). Há

evidências também que a enzima VNG0439C é uma proteína associada a membrana celular da arqueia (39). Como a condição experimental da arqueia é de exposição contínua à luz, podemos estabelecer como hipótese que esta via está sendo expressa de modo a garantir a ela uma maior proteção à radiação da luz emitida sobre ela.

4 MÉTODOS DE EXTRAÇÃO DE CARACTERÍSTICAS PARA APLICAÇÃO UTILIZAÇÃO DE APRENDIZAGEM DE MÁQUINA

O uso de novas medidas de similaridade/correlação para obter informações biológicas de grafos demonstrou-se promissora para o *Halobacterium salinarum*, mas tal abordagem não é tão adequada quando precisa-se realizar uma comparação entre dois grafos inteiros ao invés das componentes conectadas que os compõem. Para obter conclusões relevantes a partir desta abordagem, é necessário aplicar metódos mais computacionalmente custosos devido a maior quantidade e complexidade dos dados utilizados.

Optou-se então por aplicar métodos de aprendizagem de máquina em conjunto com novos algoritmos de extração de características desenvolvidos pelo grupo de computação interdisciplinar para averiguar sua eficácia em bancos de dados de experimentos de biologia molecular. Dois conjuntos de dados distintos foram utilizados: o primeiro derivado do banco de dados STRING (que contém múltiplas informações distintas sobre as relações entre os genes de seres vivos), e o segundo construído através de experimentos de RNA-Seq realizados sobre o fungo *Aspergillus fumigatus*.

4.1 Algoritmos de Extração de Características

Existem na ciência de grafos diversas medidas para a caracterização de grafos que podem ser utilizadas em algoritmos de aprendizagem de máquina para realizar uma classificação. Porém, grafos complexos costumam possuir características topológicas complexas demais cujas medidas tradicionais não são capazes de distinguir entre. Para solucionar este problema, diversas técnicas de caracterização de grafos foram desenvolvidas que auxilíam na extração de características e permite a aplicação de técnicas estatísticas mais robustas que lidam com muitas variáveis.(40–42)

Dentre os métodos de extração de características, aplicamos principalmente o "Life-Like Network Automata".(41) Este algoritmo é baseado na regra matemática de John Conway do jogo da vida, generalizado para ser aplicado em tecelagens irregulares de grafos e funcionar em regras análogas à do jogo da vida. O algoritmo para tecelagens regulares evolui um padrão inicial seguindo uma regra específica onde espaços, que estão determinados como "vivos" (1) ou "mortos" (0), podem sobreviver (mantém-se no estado 1) ou nascer (mudam do estado 0 para 1).Caso nenhuma destas condições se aplique o estado é alterado para, ou mantido, no estado 0. Em uma tecelagem regular, aplica-se a regra de evolução desejada para a vizinhança de Moore de cada espaço.

Como a vizinhança de Moore é composta por 8 unidades, existem 9 possíveis configurações que ela pode assumir e isto permite fazer uma generalização para seguir outras regras além da famosa B3S23. No total, o número de regras "life-like" possíveis são

2^9x2^9 ou 262144.

Para aplicá-las em uma tecelagem irregular, onde o número de vizinhos de 1º grau são considerados como análogos a vizinhança de Moore, aplicou-se os seguintes passos: o número de vizinhos de 1º grau são contados e calculamos a razão T = (# de vizinhos de valor 1)/(# total de vizinhos é calculada). Este valor é aplicado então a seguinte função:

$$s(c_i, t+1) = \begin{cases} 1 \text{ se } s(c_i,t) = 0 \text{ e } x_x/r <= T(c_x,t) < (x_x+1)/r \rightarrow \text{Born (B)} \\ 1 \text{ se } s(c_i,t) = 1 \text{ e } y_y/r <= T(c_y,t) < (y_y+1)/r \rightarrow \text{Survive (S)} \\ 0, \text{ caso contrário} \end{cases} \quad (4.1)$$

A função s é o estado de um determinado nó que, como foi dito anteriormente, pode ser 1 ou 0. O valor de T determinará se o vértice irá "sobreviver" ou "nascer" dependendo do seu estado inicial e da regra B{X}S{Y} escolhida. x_x e y_y são valores pertencentes aos conjuntos {X} e {Y} respectivamente, podendo conter os números inteiros de 0 a 8. Se o vértice estiver na condição "morta" e possui uma quantidade de vizinhos vivos que mapeia para um valor presente no conjunto X, então ele transiciona para o estado "vivo". Caso ele já esteja no estado "vivo" e seus vizinhos mapeiam para um número presente no conjunto Y então ele se mantém "vivo". Caso nenhuma das duas condições anteriores se aplique, então o nó permanece/transiciona para "morto".

Em conjunto com o LLNA, utilizou-se novos métodos de extração de características dos padrões gerados por ele conhecidos como LLNA-DTEP (Density Temporal Evolution Pattern) e LLNA-SDTEP (State Density Temporal Evolution Pattern). (1) No LLNA-DTEP, ao invés de registrar apenas o padrão binário gerado a partir da evolução do LLNA sobre um grafo, como é o caso do método de extração LLNA-BP (Binary Pattern),(43) registra-se também o valor de T de cada nó em cada iteração. O LLNA-SDTEP, por outro lado, combina o valor T de um nó com o seu estado atual multiplicando T por -1, caso o estado do nó seja 0, ou 1, caso ele seja 1. Desta forma, a informação extraída a partir dos padrões é bem mais detalhada.

Os métodos LLNA-DTEP e SDTEP gera os vetores característicos extraindo histogramas a partir dos padrões gerados. Três tipos diferentes de histogramas são construídos:

- Histograma Global: Os valores são extraídos diretamente do padrão gerado e divido em bins de tamanho 1/L. Os valores de T são salvos no "bin" caso ele caia dentro do intervalo correspondente do "bin" $\frac{l}{L} < T < \frac{l+1}{L}$, com $l \in [0, 1]$ para o padrão DTEP e $l \in [-1, 1]$ para o padrão SD-TEP.
- Histograma Temporal: O padrão é divido em intervalos "temporais" (isto é, cada linha do padrão global passa a ser considerado um "subpadrão") e um histograma característico é gerado utlizando o mesmo procedimento

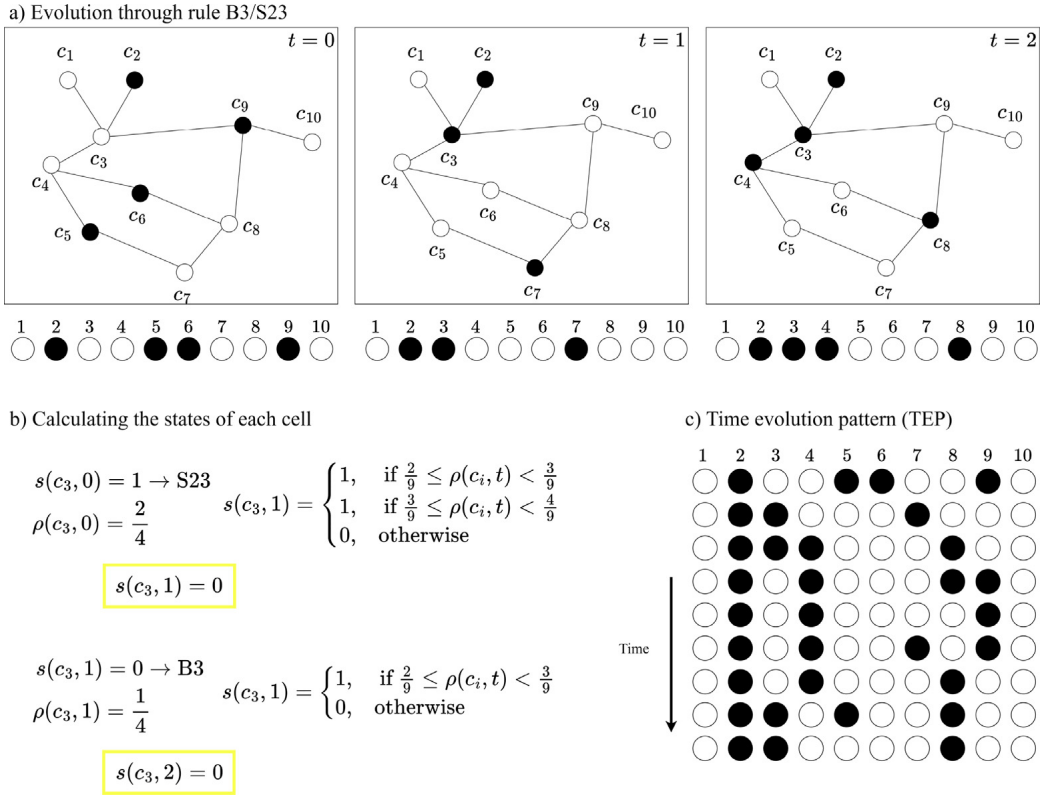

Figura 16 – Detalhamento de como funciona a generalização do do jogo da vida para tecelagens irregulares de grafos.A) Mostra como os grafos evoluem, comparando o estado inicial $t = 0$ com os dois tempos seguintes. B) Mostra como o cálculo é realizado para a mudança dos estados dos nós em uma tecelagem irregular. C) Padrão resultante a partir da evolução do estado inicial $t = 0$.

Fonte: Adaptada de ZIELINSKI *et al.* (1)

descrito para o histograma global. Feito isto, a média dos histogramas de cada intervalo temporal é realizada, obtendo o vetor característico temporal.

- Histograma por Grau: O padrão também é dividido em um "subpadrão", mas desta vez ele é gerado selecionando apenas os nós que possuem o mesmo grau. Assim como no histograma temporal, a extração é feita igual ao global com um uma média final de todos os vetores obtidos para cada grau.

No final deste processo, 6 vetores característicos são obtidos contendo L variáveis cada. Estes 6 vetores são então concatenados em um vetor Ω_{L}, que será utilizado para a classificação das redes dos bancos de dados.

O método LLNA ainda é bastante exploratório e, por isso, é necessário realizar ajustes tanto da melhor regra "life-like" quanto do número de bins mais adequado. Para isso, primeiro calculamos o padrão temporal DTEP de todas as 262144 regras utilizando

apenas o histograma global para uma semente inicial aleatória com L = 50. Após isto, selecionamos as 10 melhores regras "life-like" com estes parâmetros e, com elas, montamos o vetor $\Omega_{\{L\}}$ para diversos valores de L distintos obtidos a partir do padrão cuja acurácia foi a melhor dentro de um grupo de 100 padrões gerados por sementes aleatórias distintas.

Para obter um tempo computacional viável, utilizou-se um algoritmo em CUDA para aplicar a evolução dos padrões em GPUs. Geralmente, os padrões são evoluidos por um número de iterações N = 350, e as 20 primeiras são descartadas pois elas estão próximas demais do estado inicial e longe do estado "estacionário" do padrão, que é mais caótico e possui informações mais relevantes a serem extraídas.(41) Este valor foi mantido para o banco de dados do *A. fumigatus*, mas para o STRINGdb reduzimos N para 100, pois os grafos são muito maiores e muito mais conectados, necessitando de um tempo computacional muito maior. Embora N = 100 diminua a extração de características do ponto de vista temporal, isto é profundamente compensado pelo lado do número de nós e graus, que é muito maior neste banco.

4.1.1 Turista Bifurcado

O Método do turista bifurcado é uma modificação do método já conhecido de caminhada do turista para extração de características em grafos. Este método utiliza os mesmos princípios do algoritmo da caminhada do turista, onde realiza-se um deslocamento ao longo de todos os nós de uma rede de acordo com uma regra topológica específica e, após o turista entrar em um estado estacionário (atrator) ou alcançar o valor máximo de deslocamento l, ele interrompe o seu percurso. O turista pode seguir 4 regras determinísticas distintas para realizar o seu percurso, mas ele pode se encontrar em situações onde mais de um nó vizinho cumpre estas regras. O método do Turista Bifurcado busca solucionar este problema gerando clones do turista toda vez que ele se encontra em um destes impasses. O deslocamento dos clones, junto com o número total de bifurcações realizadas, são então utilizados como parâmetros para extrair características dos grafos.(44)

As quatro regras possíveis que o turista pode seguir sobre a topologia de uma rede são:

- **B-R1:** O turista caminha em direção ao vértice que possui o mesmo grau k que o vértice em que ele está.

- **B-R2:** O turista vai para um vértice com grau distinto em relação ao vértice em que ele está.

- **B-R3:** O turista move-se para o vértice que minimiza a diferença entre o grau do vértice atual e o próximo.

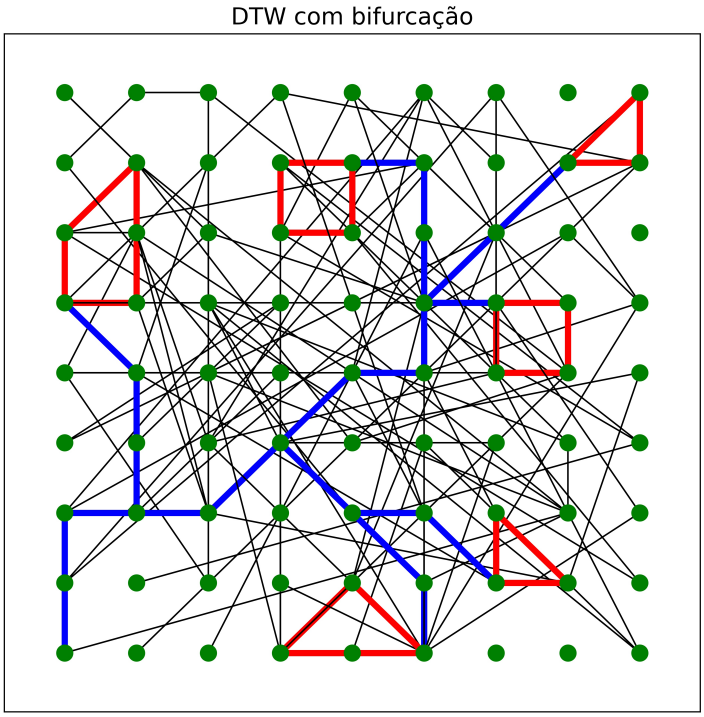

Figura 17 – Exemplo de caminhada do turista com bifurcações (DTWB) sobre uma rede. Em azul temos a caminhada transiente e em vermelho a estacionária (atrator). Observa-se também que quando dois vértices satisfazem a regra do turista uma bifurcação é criada.

Fonte: Adaptada de MERENDA. (44)

- **B-R4:** O turista move-se para o vértice que maximiza a diferença entre o grau do vértice atual e o próximo.

Para a construção do vetor característico deste método, quatro medidas foram realizadas sobre a caminhada do turista a partir de um nó e também seus clones:

- **Grau de Bifurcação (B):** Número de bifurcações que ocorre quando o turista encontra-se em um impasse. Possui valor mínimo dado por $B = \#$ de impasses $- 1$ para uma memória $\mu \geq 1$, mas pode ser generalizado por $|X_i^r| - |X_i^r \cap Y_i^\mu|$. Onde X_i^r é o conjunto de vizinhos empatados no i-ésimo vértice de acordo com a regra r, e Y_i^μ o conjunto de vizinhos do vértice i na memória do turista.

- **Bifurcação/Vizinhos (β):** Dado pela razão entre o grau de bifurcação de um nó (B) e o seu grau topológico (k).

- **comprimento do transiente (t):** Número de arestas que compõem a trajetória transiente do turista.
- **comprimento do atrator (a):** Número de arestas que compõem a trajetória atratora do turista.

Cada nó possuirá um valor de B e β, e cada trajetória irá possuir um valor de t e a. Para a construção do vetor característico foram tomadas medidas estatísticas de todos estes valores, sendo elas a média ($\langle\rangle$), o desvio padrão (σ), e a entropia de Shannon (H). No final, o vetor utilizado irá possuir 4 variáveis, como mostra a equação 4.2.

$$\phi_{\mu}^{r} = [\langle B\rangle, \sigma_B, H_B, \langle\beta\rangle, \sigma_\beta, H_\beta, \langle t\rangle, \sigma_t, H_t, \langle a\rangle, \sigma_a, H_a] \tag{4.2}$$

4.1.2 Redes Neurais Convolucionais em Grafos

O método de redes neurais convolucionais em Grafos aplica os fundamentos de apredizagem de máquina de redes neurais para a extração de características em bancos de dados caracterizados na forma de grafos. Redes neurais convolucionais são constituídas por diversas camadas de "neurônios", que são matematicamente expressos na forma de vetores, com conexões estabelecidas entre si por arestas, representadas na forma de uma matriz que transforma os valores dos vetores para compor a próxima camada (vetor). Estas iterações vão sendo aplicadas sobre um banco de dados $\mathbf{X}$, contendo os vetores característicos x_i, até chegar a camada de decisão que comumente possui o mesmo número de nós que a quantia de classes do vetor $\mathbf{Y}$, que se refere às classes verdadeiras das instâncicas do banco $\mathbf{X}$.

A camada final irá tentar predizer a qual classe o vetor característico x_i pertence, geralmente ativando o nó da camada de decisão com maior valor. Este nó irá representar uma possível classe de $\mathbf{Y}$. A passagem pelas camadas é realizada para todos os vetores do banco $\mathbf{X}$, onde então uma função perda é aplicada sobre o conjunto de valores preditos pela camada de decisão e a classificação verdadeira. Esta função perda pode ser, por exemplo, a distância euclidiana entre o classificação verdadeira e predita. (45)

No caso da utilização de redes neurais para a extração de características de um grafo, a camada final será composta por nós cuja intensidade de ativação são as características utilizadas na classificação. Na arquitetura utilizada neste projeto, a rede possuia além da camada obtida a partir da topologia dos grafos, uma camada escondida com 128 nós e uma camada final de extração de caracterísiticas contendo 30 nós.

4.1.3 Laplace

O método de Laplace se baseia na montagem de uma matriz laplaciana normalizada para realizar a extração de features do grafo. Isto é feito tomando a matriz de adjacência do grafo ($\mathbf{W}$), realizando sua diagonalização através de uma análise espectral ($\mathbf{D}$) e, algumas

vezes também, realizando a normalização de **W**. A matriz laplaciana do grafo é então calculada através da diferença entre **D** e **W** (**L** = **D** - **W**).(42)

A partir da matriz laplaciana, um histograma espectral normalizado do grafo é calculado. Este histograma tem a quantidade de entradas dependente do tamanho do grafo inserido mas, para padronizar as características, as 30 primeiras entradas apenas foram levadas em consideração, gerando um vetor característico para cada grafo com 30 variáveis.

4.1.4 Medidas Estruturais Tradicionais

Extração de características globais de grafos que são tradicionalmente aplicadas em grafos para caracterização. As medidas em questão são as mesmas descritas em (43), sendo elas:

- $\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i$ grau médio dos vértices;
- $\langle H_{k_h} \rangle = \frac{1}{N} \sum_{j=1} k_j^h$, grau hierárquico dos vértices. Quando $h = 2$ obtém-se a soma do graus dos vizinhos de um vértice v_j, quando $h = 3$ obtém-se a soma do grau dos vizinhos dos vizinhos de um vértice v_j.
- $\langle cc \rangle = \frac{1}{N} \sum_j cc_j$ onde cc_j é definido como $cc_j = \frac{e_j}{k_j(k_j-1)}$ e e_j é definido como o número de pares conectados a cada um dos vizinhos de v_j;
- $\langle l \rangle$ é o comprimento médio do caminho mais curto que pode-se estabelecer entre quaisquer dois vértices;
- ρ_P, medida de assortatividade da rede utilizando-se de medidas de correlação entre vértices.

O vetor característico das medidas estruturais fica então sendo dado por 4.3:

$$\theta = [\langle k \rangle, \langle H_{k_h} \rangle, \langle cc \rangle, \langle l \rangle, \rho_P] \tag{4.3}$$

5 MÉTODOS COMPUTACIONAIS DIVERSOS E BANCOS DE DADOS

Neste capítulo serão expostas as ferramentas utilizadas na classificação e montagem dos grafos.

5.0.1 Algoritmos de Apredizagem de Máquina

Após a extração dos vetores característicos dos bancos de dados, foi utilizado o algoritmo "Support Vector Machine" para a classificação das classes. Este algoritmo é eficiente para bancos de dados com poucas instâncias e possui garantias teóricas de aprendizado.(46) Seu funcionamento se dá a partir do ajuste de hiperplanos (dim(N-1) com N sendo o número de variáveis do banco) de modo a separar da melhor maneira as classes do banco de dados. Ele também possui diversos hiperparâmetros a serem ajustados. Utilizou-se neste projeto o pacote sci-kit learn da linguagem de programação Python para realizar a classificação, que possui os parâmetros descritos em sua documentação.(47)

O princípio do algoritmo está na minimização da função 5.1, onde $(\boldsymbol{w}, b)$ são os parâmetros do hiperplano que separará as classes, $(\boldsymbol{x}_i, y_i)$ são as "features" de cada instância e sua classe respectivamente, e C aumenta/diminui a penalidade da função perda (neste caso, mantivemos $C = 1$).

$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{n} \max(0, 1 - y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b)) \tag{5.1}$$

Como os métodos de extração de características geram um número de variáveis bastante extenso, especialmente o LLNA, buscamos favorecer bancos com características linearmente separáveis, que são de mais fácil interpretação e classificação além de poderem ser mais facilmente aplicados em algoritmos de redução de dimensionalidade, como o LDA, que facilitam a compreensão do banco de dados passando ele para um espaço mais manejável (2D ou 3D). Por isso, selecionamos o classificador SVM com um "kernel" linear, e reduzimos a tolerância de convergência da função perda para 10^{-9}. Todos os outros hiper-parâmetros foram mantidos como está no padrão do pacote.

5.0.2 Algoritmos para Redução de Dimensionalidade

O banco de dados STRING é extremamente extenso na quantiidade de variáveis que podem ser extraídas a partir dele. Portanto, optamos por utilizar o método LDA (*Linear Discriminant Analysis*) para servir como redutor de dimensionalidade de modo supervisionado, ou seja, utilizaremos as classes conhecidas do problema para auxiliar na redução da dimensão do banco de dados.

Este método utiliza-se de estatística bayesiana para criar regiões C que segregam as classes do banco de dados desejado. Geralmente, ele assume uma distribuição normal dos dados e classes que compartilham a mesma variância entre si. Dada estas condições, um função de densidade de probabilidade $f_k(\boldsymbol{x})$ para cada classe k é definida. Para calculá-la, primeiro constrói-se um vetor de médias das variáveis pertencentes a $\boldsymbol{x}$ para cada classe k, depois, calcula-se uma matriz covariância Σ. A função $f_k(\boldsymbol{x})$ para a ser definida então por 5.2.

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mu_k)^t \Sigma_k^{-1}(\boldsymbol{x} - \mu_k)\right) \tag{5.2}$$

Caso o método LDA fosse utilizado aqui como um classificador de instâncias, utilizariamos a função densidade sobre o teorema de Bayes para determinar a probabilidade de cada instância no conjunto teste pertencer a uma determinada classe. No entanto, estamos utilizando o LDA apenas como um redutor de dimensionalidade, então basta selecionar as componentes que obtiveram a melhor capacidade discriminativa para compor o espaço transformado ao qual o banco de dados irá ser exibido.

5.0.3 Método de Validação Cruzada

Para estimar a capacidade classificativa de um determinado modelo, métodos de validação cruzada são importantes pois ajudam a diagnosticar problemas como viéses em relação a uma determinada classe ou *overfitting* (quando os parâmetros ajustados se tornam demasiadamente enviasados em função do conjunto treino, prejudicando predições futuras em outros possíveis bancos de dados). Estes métodos dividem os bancos de dados utilizados para o ajuste do classificador em conjuntos teste e treino. No conjunto treino, as instâncias são utilizadas para ajuste dos parâmetros do classificador. Com os parâmetros ajustados, o conjunto teste é então utilizado para averiguar a capacidade classificativa, evitando assim que o resultado seja superdeterminado pelas características dos dados de treino, simulando uma possível situação "inesperada".(48)

Utilizou-se o método de validação cruzada *K-fold* com $K = 5$ para todos os experimentos deste projeto. Neste método, o banco de dados é divido de 5 maneiras distintas de modo a variar as instâncias do conjunto treino e teste, evitando um possível viés inerente a uma forma específica de separação dos dados. Como buscamos manter a classificação o mais balanceado possível, cada conjunto teste/treino constituem 20%/80% do banco de dados respectivamente, com a mesma quantidade de instância para cada classe quando possível.

5.1 Bancos de Dados

5.1.1 Bancos de Dados para validação e comparação dos métodos

Para obter uma melhor compreensão da habilidade classificativa dos métodos de extração de características dos grafos, realizou-se uma comparação com bancos de dados já utilizados pelo nosso grupo de pesquisa para otimização dos novos métodos.(1, 43)

Entre os bancos de dados estão:

1. *Bancos de Dados de Redes Sintéticas*: O primeiro banco de dados é constituído por redes geradas de forma sintética. Ele possui 4 classes, sendo elas 4 modelos de geração de redes diferentes: Rede aleatória de Erdós e Rényi (9), que estabelece uma aresta entre dois vértices com probabilidade igual a $p = \frac{\langle k \rangle}{n}$; rede de Pequeno-mundo com probabilidade de reconexão igual a $p = 0.1$;(10) rede de livre escala tanto com modelo de ligação preferencial linear e não-linear;(49) rede geográfica.(50) As redes de cada um dos modelos foram geradas a partir dos parâmetros: $\langle k \rangle$: 4, 6, 8, 10, 12, 14, 16 e N: 500, 1000, 1500, 2000. Cada uma das 28 possíveis combinações destes dois parâmetros possuem 100 redes. Logo, o número de redes para cada modelo é 2800, resultando em um total de 11.200 redes no banco de dados.

2. *Banco de Dados Sintético com Ruído*: Composto pelas mesmas redes do banco sintético mas com a adição de ruído em diferentes intensidades dada por ρ. Os valores de ρ são quantias percentuais do número total de arestas, sendo $\rho/2$ a quantia adicionada aleatoriamente e $\rho/2$ a quantia removida aleatoriamente. Com o aumento de ρ, mais distante a rede fica do modelo que a gerou.(41) Os valores utilizados foram 10%, 20% e 30%.

3. *Banco de Dados de Redes Sociais*: Composto de bancos de dados da plataforma SNAP (Stanford Network Analysis Project).(51) Cada rede social corresponde às relações sociais ou amizades de um usuário específico que não está representado na rede. Este banco de dados contém 100 amostras dividida em duas classes (Google+ e Twitter) com 50 amostras cada.

4. *Banco de Dados de Redes Metabólicas*: Composto de redes metabólicas que foram construídas o modelo substrato-produto.(52) Ou seja, os vértices são metabólitos onde produtos (metabólito resultantes de uma reação) e substratos (metabólitos anteriores a uma reação) estão ligados por uma aresta caso eles pertençam a uma mesma reação conhecida. As reações de diversos organismos foram obtidos do banco de dados KEGG.(29) Este banco de dados for dividido em 6 conjuntos:

    a) *Banco de dados de Reinos*: Contém diversas espécies do domínio *Eukaryota* com quatro classes distintas: animalia, plantae, fungi, protist, cada um contendo 40 redes.

b) *Banco de dados de Animais*: Possui quatro classes de animais: mamíferos, aves, peixes, e insetos, cada um contendo 14 amostras por classe.

c) *Banco de dados de Plantas*: Contém três classes: monocotiledôneas, algas verdes, e Eudicotiledôneas, cada classe possuindo 9 organismos.

d) *Banco de dados de Protistas*: Este banco é composto de 4 classes: Amebozoas, Alveolados, Stramenopiles e Euglenozoa, cada uma contendo 5 organismos.

e) *Banco de dados Firmicutes-Bacillis*: Este banco contém quatro classes: Bacillus, Staphylococcus, Streptococcus e Lactobacillus, contendo instâncias desbalenceadas de 122, 76, 133, e 83 espécies respectivamente.

f) *Banco de dados de Actinobactérias*: Outro banco de dados desbalanceado, contendo três classes: Mycobacterium, Corynebacterium, e Streptomyces com 60, 86, e 53 espécies respectivamente.

5.1.2 Banco de dados STRING

Nesta parte do projeto, foram utilizados dois tipos de bancos de dados distintos. O primeiro deles foi extraído diretamente do site STRINGdb. Neste banco, há disponível redes de intereração proteína-proteína que podem ser selecionadas com base em espécies específicas. Nestas redes, os nós representam os genes codificantes de proteínas conhecidas dos seres vivos, e as arestas representam evidências que os correlacionam de diversas formas distintas. Tais evidências podem ser divididas em 3 tipos distintos:

- **Predições computacionais baseadas em características genômicas como:** proximidade dos genes ao longo do arcabouço, evidência de fusão entre genes e co-ocorrência dos mesmos ao longo de sua evolução
- **Evidências obtidas a partir de experimentos como:** co-expressão gênica em experimentos de transcriptômica, e experimentos importados de outros bancos de dados, como KEGG.
- **Conhecimentos gerais obtidos das relações entre os genes como:** descobertas ontológicas descritas em outros bancos de dados (GO, KEGG, FUNCAT), e "text-mining" de artigos publicados que demonstram evidências de correlação entre as proteínas.

No banco de dados, estas evidências são demonstradas através de diferentes colorações e pesos em cada aresta demonstrando a significância destas evidências para correlacionar os genes. Porém, estamos interessados em explorar somente a topologia da rede sem levar em consideração os pesos. Fizemos então, o download de 100 destas redes na sua forma sem peso nas arestas, ou seja, quaisquer das evidências descritas que estão

presentes entre os genes foram consideradas. Estas 100 espécies foram separadas em 10 classes distintas contendo 10 instâncias cada, ou seja, um banco de dados balanceado para facilitar o processo de aprendizagem de máquina.

As 10 classes do banco de dados STRING são: *Vertebrados*, *Angiospermas*, *Artrópodes*, *Protistas*, *Chytridiomycota*, *Basidiomycota*, *Ascomycota*, *Mucoromycotina*, *Bactérias*, e *Arqueias*. Das 10 classes, 8 delas são classificações cladísticas (Protistas e Mucomycotinas não são, isto é, as espécies pertecentes a estas classes podem ser mais diferentes entre si do que entre membros fora do grupo). Elas foram propositalmente selecionadas para compor diferentes distâncias evolutivas na árvore da vida. Desta forma, será possível avaliar o quão bem o extrator de característica consegue determinar a distância entre as classes. As classes não-cladísticas servem também como bons parâmetros para verificar como elas se dispõem no contexto em conjunto com as outras classes.

Já existem evidências que demonstram haver uma relação entre a topologia de grafos metabólicos e a evolução de diferentes filos de vegetais.(53) Nosso intuito com este novo banco de dados, foi averiguar se relações análogas podem ser encontradas em em redes de interação proteína-proteína e em outros domínios da vida, como animais, fungos e procariontes.

5.1.3 Banco de Dados de Co-expressão Proteica

O segundo tipo de banco de dados utilizado foi obtido através de experimentos de RNA-Seq do fungo *Aspergillus fumigatus* exposto ao fungicida Caspofungin ao longo de um período de 8 horas com "data points" coletados nos períodos de 0h, 0.5h, 1h, 4h e 8h após a exposição a Caspofungin.(54) Tais experimentos foram processados através da pipeline usual de bioinformática, realizando o alinhamento com o genoma do fungo com o software *Hisat2*, extraindo o nível de expressão para cada um dos genes alinhados e montando uma normalização das expressões dos genes para Transcritos por Milhão (TPM). Esta pipeline já possui precedentes na literatura para o processamentos de dados que são utilizados em redes de co-expressão.(20)

Para a montagem da rede de co-expressão global, utilizamos o método HRR (HIghest Reciprocal Ranking)(55) que correlaciona as expressões dos genes em diferentes períodos através da correlação de Pearson e, depois, realizou-se a clusterização da rede global em redes menores com o algoritmo HCCA (56) através dos seus parâmetros padrões. Buscamos focar apenas na construção do banco de dados utilizando os experimentos exercidos sobre o *Aspergillus fumigatus* sem mutações (WILD), pois seu metabolismo já é melhor entendido e podemos realizar uma comparação com as informações biológicas conhecidas da literatura de maneira mais adequada evitando falsos positivos.

O método HRR calcula a correlação de cada gene em relação a todos os outros e, após isto, remove as todas as correlações tidas como fracas (< 0.8). Ele então realiza um

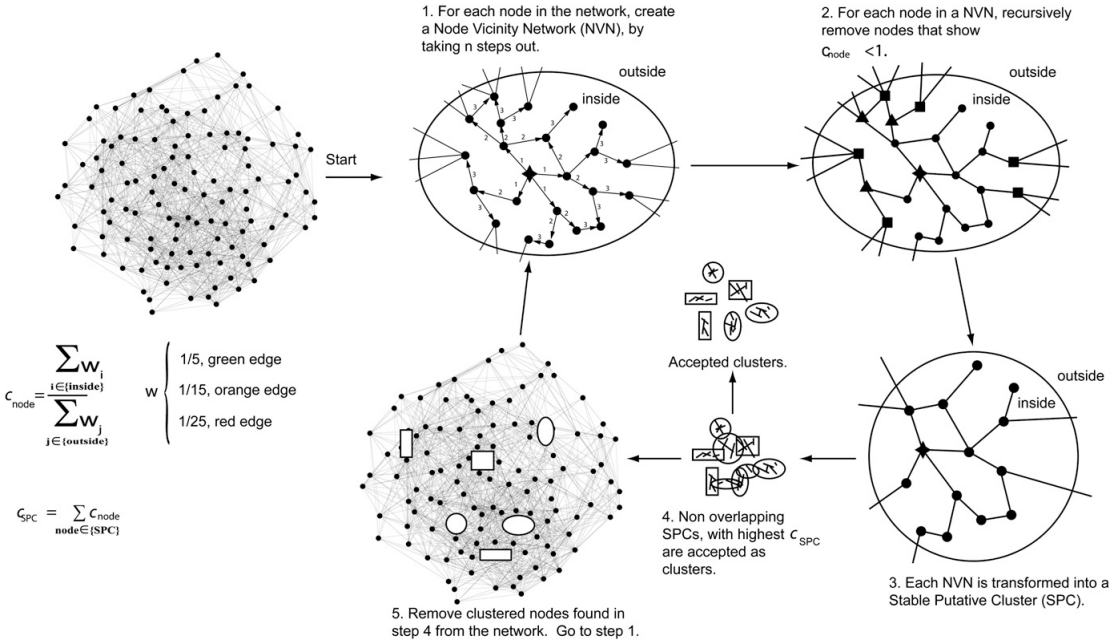

Figura 18 – Fluxograma mostrando os passos do algoritmo HCCA. 1) Para cada nó, o algoritmo gera uma subrede baseada na vizinhança de grau $n = 3$. 2) Os nós que tiverem um valor de $C_{node} < 1$ devem ser removidos da rede de vizinhança. 3) A rede resultante é transformada então em um "cluster putativo estável"(SPC). 4) Todos os SPCs que não se sobrepõem e possuem o C_{SPC} mais alto são considerados clusters. 5) Os nós pertencentes a estes clusters são removidos da rede global e o processo é repetido até que nenhum cluster válido sobre.

Fonte: Adaptada de MUTWIL *et al.*(56)

ranqueamento das arestas entre os genes dando valores maiores para arestas com ranques maiores. Estes valores são então utilizados pelo método HCCA para realizar o segregação de clusters fortemente correlacionados.

Os clusters são construídos selecionando a vizinhança de grau $n = 3$ de cada nó e removendo os nós com a medida $C_{nonde} < 1$, que é calculada através da razão entre a soma dos pesos das arestas do nó em questão que pertencem ao cluster, e as que estão fora do cluster (Figura 18). Este cluster modificado ("Cluster Putativo Estável") tem então um valor global calculado somando todos os C_{node} dos nós pertencentes a ele (C_{SPC}). Este valor global é então usado para discriminar clusters que se sobrepõem separando do resto da rede aqueles que obtém o maior valor entre todas as sobreposições. O processo é então repetido até que toda a rede seja clusterizada ou não haja mais componentes conectadas que satisfaça os parâmetros padrões (no caso, os parâmetros padrões que o HCCA usam como limiar é clusters de no mínimo 40 nós até 200 nós e pertencentes a vizinhança de grau $n = 3$ do nó original do cluster). A Figura 18 mostra o workflow do método HCCA.

Com o banco de dados do *A fumigatus*, geramos 4 redes globais, sendo cada uma delas resultante correlação entre os tempos consecutivos dois a dois, ou seja, obtivemos

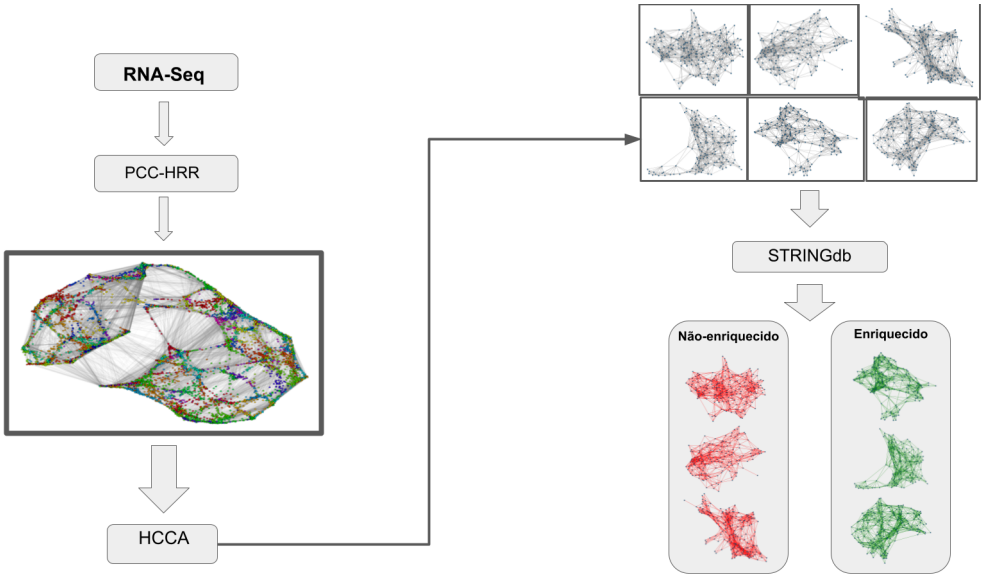

Figura 19 – Pipeline de montagem do banco de dados dos clusters de redes de co-expressão do *Aspergillus fumigatus*

Fonte: Elaborada pelo autor

redes que comparam 0h-0.5h, 0.5h-1h, 1h-4h, e 4h-8h. Os clusters foram então extraídos a partir destas 4 redes.

Nos clusters gerados a partir do método descrito, a classificação se deu a partir do enriquecimento de termos ontológicos enriquecidos presentes. Cada cluster gerado teve seus genes colocados no banco de dados STRINGdb e verificou-se se eles fazem parte mais do que o esperado de vias metabólicas presentes no banco de dados KEGG.(29) A Classe 1, então, ficou determinada como os clusters enriquecidos com vias do banco KEGG e a Classe 0 ficou determinada como aqueles clusters que não obtiveram nenhum enriquecimento do banco KEGG, ou seja, são genes que não estão expressando nenhuma via. A Figura 19 mostra um resumo dos passos tomados para a construção do banco de dados.

No total, obteve-se 59 clusters enriquecidos com termo KEGG (tipo 1). Para balancear o banco de dados, selecionou-se 59 clusters sem nenhum enriquecimento para compor a outra classe (classe 0). A tabela abaixo (5) compara as diferenças topológicas básicas entre o banco de dados STRING e o banco de dados do fungo *A. fumigatus*.

Tabela 5 – Comparação das diferenças e semelhanças entre os bancos de dados explorados com inteligência artificial.

	StringDB	**Clusters *A. fumigatus***
# de Classes	Multiclasse (10)	Binário
Tamanho dos grafos	Grandes, de 2000 a 15000 nós	Pequenos, de 40 a 200
Topologia Básica	Livre escala e Randômicos	Randômicos
O que representa? (Arestas)	Interações diversas entre proteínas	co-expressão entre proteínas
O que representa? (Nós)	Genes codificantes de proteínas	Genes codificantes de proteínas
# de Espécies	100 espécies distintas	Uma única espécie

Fonte: Elaborada pelo autor.

6 RESULTADOS E DISCUSSÕES PARA A APLICAÇÃO DE APRENDIZAGEM DE MÁQUINA SOBRE REDES BIOLÓGICAS

Aqui serão discutidos os resultados obtidos a partir da aplicação dos métodos de extração de característica sobre os bancos de dados sintéticos, obtidos a partir de bancos de dados públicos, e montados a partir de experimentos de biologia molecular.

6.0.1 Acurácia dos Bancos de Dados Diversos

Os resultados obtidos para a aplicação dos diferentes métodos de caracterização de grafos sob os bancos de dados diversos demonstram o potencial que eles possuem e seus pontos fortes em relação a natureza de cada banco (**Tabela 6**). As medidas estruturais tradicionais de grafos obtiveram os melhores resultados para as redes sintéticas, incluindo as que continham ruído. Os vetores característicos $\Omega_{(60,100)}$ e $\Omega_{(40,100)}$ foram os segundos melhores, enquanto os métodos do turista bifurcado com as 4 regras distintas, rede neural de grafos e laplace obtiveram resultados equiparáveis.

Já para os bancos de dados reais, as medidas estruturais começaram a demonstrar suas limitações. Dos 7 bancos de dados reais 5 deles (actinobacteria, animais, firmicutes-bacillus, fungos e plantas) tiveram a acurácia superada pela regra BR1 do turista bifurcado, e todos foram superados pelos vetores $\Omega_{(60,100)}$ e $\Omega_{(40,100)}$. Isto demonstra a flexibilidade destes modelos para modelar e classificar redes que descrevem situações reais.

Tabela 6 – Acurácias obtidas para os diferentes métodos de extração de características de grafos. Note-se que medidas estruturais tradicionais tendem a possuir um nível de acurácia melhor que os outros métodos,mas demonstra limitações em bancos de dados reais. Os resultados dos vetores $\Omega_{(60,100)}$ e $\Omega_{(40,100)}$ foram retirados do trabalho de Kallil *et al.* (1).

dataset	**DTWB-R1**	**-R2**	**-R3**	**-R4**	**GNN**	**laplace**	**estrutural**	$\Omega_{(60,100)}$	$\Omega_{(40,100)}$
sintetica	97.0	95.9	93.1	94.7	93.5	98.9	100.0	100.0	100.0
s (10%)	96.3	95.8	92.6	93.2	91.2	97.4	100.0	100.0	100.0
s (20%)	95.7	95.2	90.6	92.5	89.4	96.8	100.0	100.0	100.0
s (30%)	94.9	95.3	89.1	91.4	88.5	95.5	100.0	99.75	99.75
actbac.	99.4	98.9	83.6	95.8	89.1	93.4	93.16	97.65	97.68
animais	94.6	98.2	82.1	90.5	85.6	93.2	83.71	100.0	99.71
firm.-bac.	97.3	93.4	90.5	85.0	84.2	94.8	95.67	95.73	96.06
fungos	78.8	81.6	53.3	74.6	59.8	78.4	54.9	81.00	80.43
plantas	81.4	92.5	48.1	55.6	61.3	77.1	54.19	79.58	81.33
reinos	94.3	98.1	89.3	95.0	69.2	84.6	96.61	96.24	96.24
sociais	81.0	78.0	66.0	64.0	45.0	77.0	88.0	92.5	92.5

Fonte: Elaborada pelo autor.

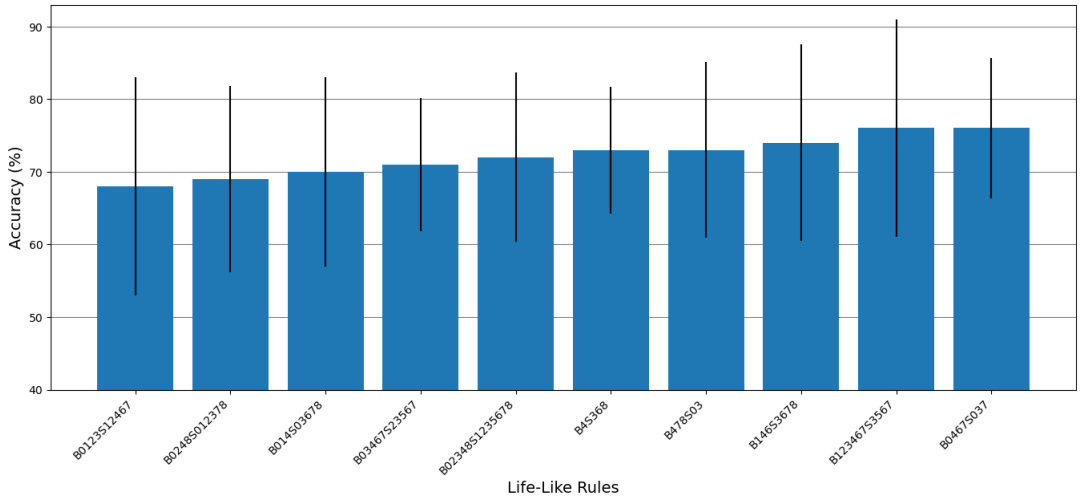

Figura 20 – Valor das acurácias obtidas para as 10 melhores regras dentre as 256.144 regras life-like possíveis. Os valores giram em torno de 70% o que, como será visto adiante, está em torno da acurácia dos métodos de extração tradicionais, mas com a otimização houve uma melhora significativa.

Fonte: Elaborada pelo autor

6.0.2 Resultados no Banco de Dados STRING

Após demonstrarmos que os novos métodos extratores de características possuem uma maior aplicabilidade para bancos de dados reais iremos agora aplicá-los sobre o banco de dados STRING para verificar o quão bem ele consegue distinguir entre os grupos separados por nós. No entanto, este banco de dados apresenta uma estrutura possivelmente muito distinta da estruta das redes metabólicas utilizadas na calibração das regras "life-like". Portanto, uma nova calibração é necessária, fazendo uma seleção de novas melhores regras e, após ter feito isto, a montagem de um novo vetor característico que obtenha a maior acurácia possível.

Foi possível verificar que as regras ótimas para este banco de dados são de fato distintas das regras ótimas aplicadas para o banco KEGG nos trabalhos de (43), demonstrando a diferença topológica das redes do STRINGdb (Figura 8). As três melhores regras "life-like" foram B146S3678, B123467S53567 e B0467S037 e as acurácias obtidas foram 74 $\pm$ 14%, 76 $\pm$ 15% e 76 $\pm$ 10% respectivamente. É possível notar que o desvio padrão é relativamente grande em todas as 10 medidas, muitas vezes alcançando os dois dígitos, evidenciando uma necessidade de mais otimizações.

Realizou-se, então, a montagem do novo vetor característico para adquirir um novo aprimoramento. A montagem se deu utilizando as 10 melhores regras e as variações dos tamanhos de "bins" descritos nos métodos, além de 100 sementes aleatórias que para cada uma das regras para garantir que o máximo de acurácia foi alcançado. É importante notar que não necessariamente as melhores 3 regras obtidas pelo primeiro passo da seleção serão

as 3 melhores regras no próximo passo. Como ainda se desconhece a natureza das regras life-like é possível que as características extraídas por regras menos "eficientes" em um primeiro momento possam adquirir uma melhora considerável após a construção do novo vetor característico e após a seleção da melhor semente.(41, 43)

Como se observa na Figura 9, é exatamente isto que ocorreu. Das 3 melhores regras para o segundo passo de otimização (B123467S3567; B0123S12467; B03467S23567) 2 são distintas em relação ao primeiro (B0467S037; B123467S3567; B146S678). Observa-se também uma melhora considerável na acurácia da classificação, com as acurácias sendo ao redor de 10 pontos percentuais acima do melhor resultado do primeiro passo. Outro ponto importante a ser notado é a melhora no desvio padrão que teve uma redução considerável, indo de 10 pontos percentuais no melhor resultado para 8 no pior resultado.

Como a regra que obteve a melhor acurácia foi B12346S3567, que chamaremos a partir de agora coloquialmente de regra ótima, esta foi então usada para os resultados que virão a seguir.

A Figura 22 faz uma análise comparativa entre as acurácias dos demais métodos. Como se pode observar, o vetor proposto neste projeto com a regra ótima foi capaz de superar as acurácias de todos os métodos de extração de características tradicionais por uma margem considerável de acurácia. Nota-se também que a regra BR1 do turista bifurcado alcançou uma acurácia próxima das medidas estruturais (74 $\pm$3%), mas não conseguiu equiparar-se ao $\Omega_{(25)}$.

Aqui iremos apresentar as melhores classificações de cada um dos métodos, como os classificadores lidaram com cada classe e quais classes foram mais fáceis de serem classificadas corretamente. Para visualizar isto, nós utilizamos matrizes confusão que indicaram quais classes obtiveram a maior quantidade de instâncias fora da diagonal da matriz, que representa a classificação correta. A matriz é uma representação visual da junção dos resultados obtidos para todos os "folds" teste que dividiram o banco de dados. No caso, como usamos o método de "k-fold" com k=5, temos 5 conjuntos teste contendo 20 instâncias, sendo 2 delas de cada uma das 10 classes.

Na Figura 23 fica evidente a vantagem que o $\Omega_{(25)}$ possui sobre os outros métodos, conseguindo classificar razoavelmente bem até mesmo os filos de fungos, algo que os outros métodos possuem uma grande dificuldade em conseguir. As classes "Angiosperma" e "Vetebrados" foram os que obtiveram as melhores classificações na maioria dos métodos, com exceção do GNN. Isto provavelmente se deve a complexidade e também a quantidade de nós que estes organismos possuem. Angiospermas compõe os seres vivos com a maior quantidade de genes conhecida no planeta,(57) gerando uma enorme quantidade de dados que dá a nossos métodos uma capacidade de extração de características excepcionalmente vasta. Vertebrados também são seres bastante complexos e altamente estudados, o que contribui para uma grande quantidade de genes e arestas também.

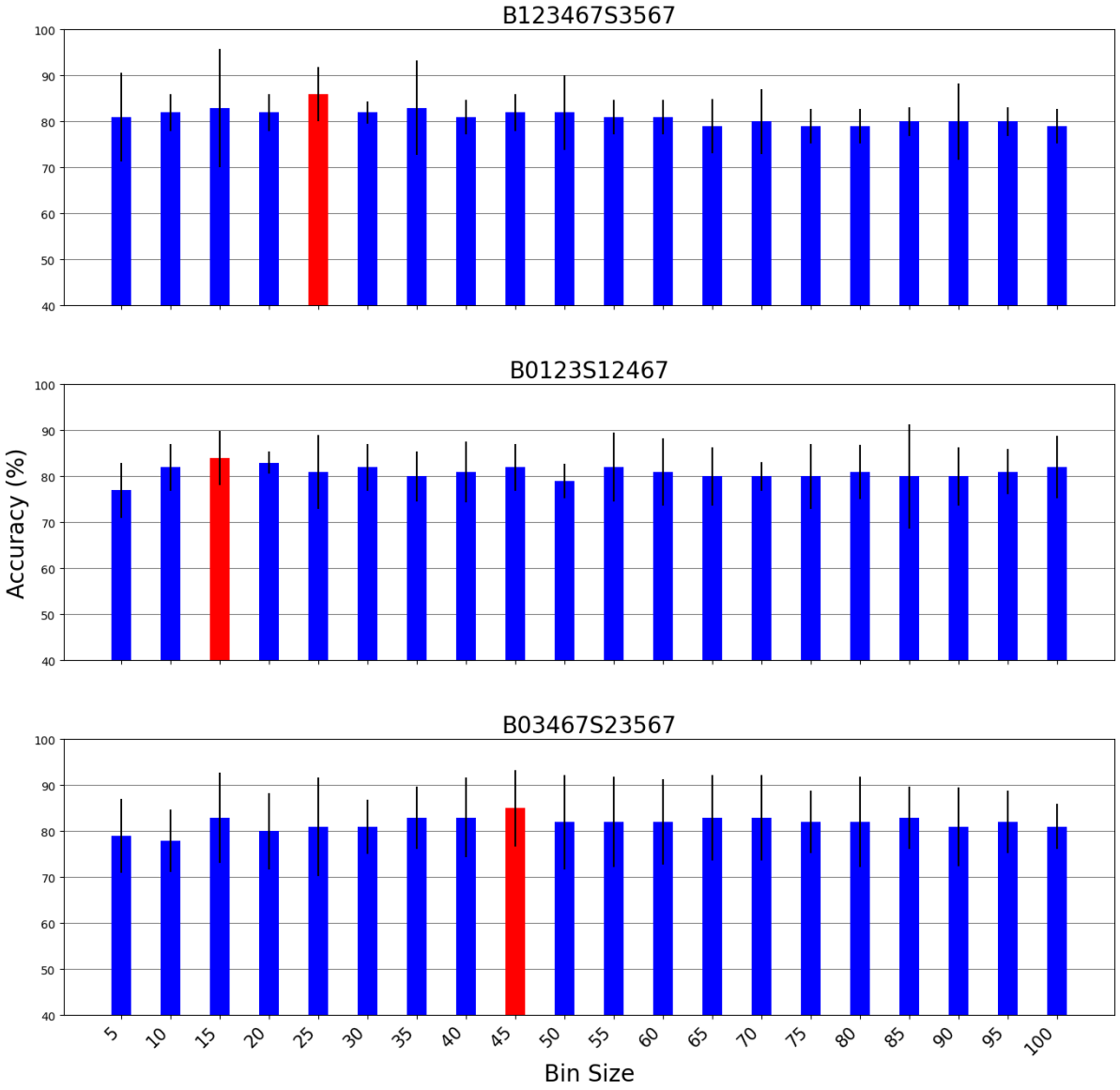

Figura 21 – Otimização do tamanho dos bins dos histogramas que compõem o vetor Ω. Neste caso, estão destacados em vermelho os bins que obtiveram as melhores classificações nas 3 melhores regras entre as 10. Estas 3 regras no caso são B12346S3567 (com a melhor acurácia sendo 86 ± 6%), B0123S12467 (melhor acurácia 84 ± 6%) e B0346S23567 (melhor acurácia 85 ± 8%).

Fonte: Elaborada pelo autor

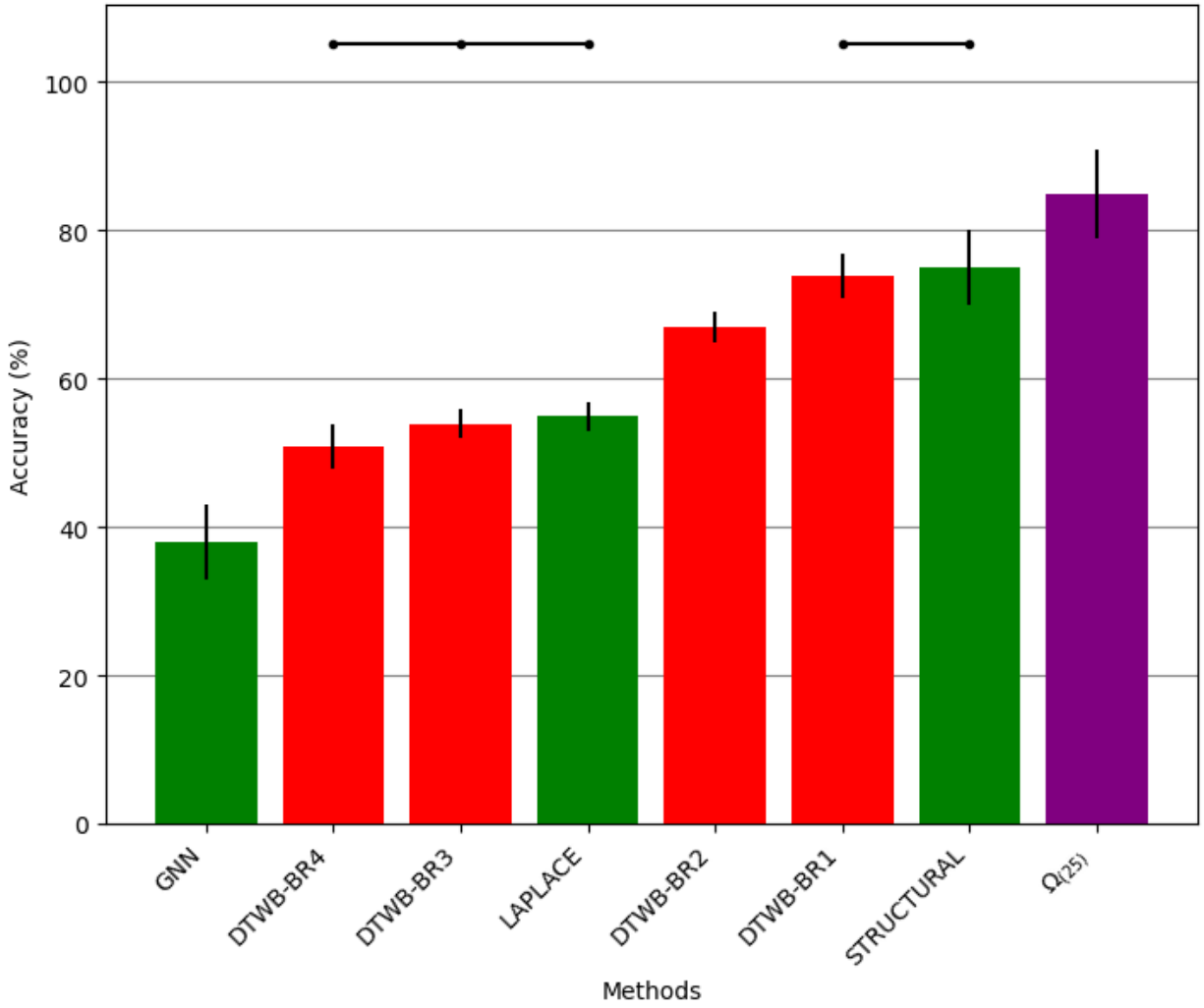

Figura 22 – Barras mostrando a acurácia da classificação do banco de dados STRINGdb para diversos métodos aplicados. Em vermelho estão destacas as 4 diferentes regras do turista bifurcado, em verde os métodos de extração utilizados mais tradicionalmente e, por último, em roxo o vetor característico $\Omega_{(25)}$. Sua vantagem é evidente em relação aos outros, com uma acurácia de $86 \pm 6\%$ com 11 pontos percentuais acima da segunda maior medida ($75 \pm 5\%$). Os traços com pontos acima da barra indicam quais métodos obtiveram resultados pouco significativos através do teste t de Student corrigido para múltiplas comparação com Bonferroni.

Fonte: Elaborada pelo autor

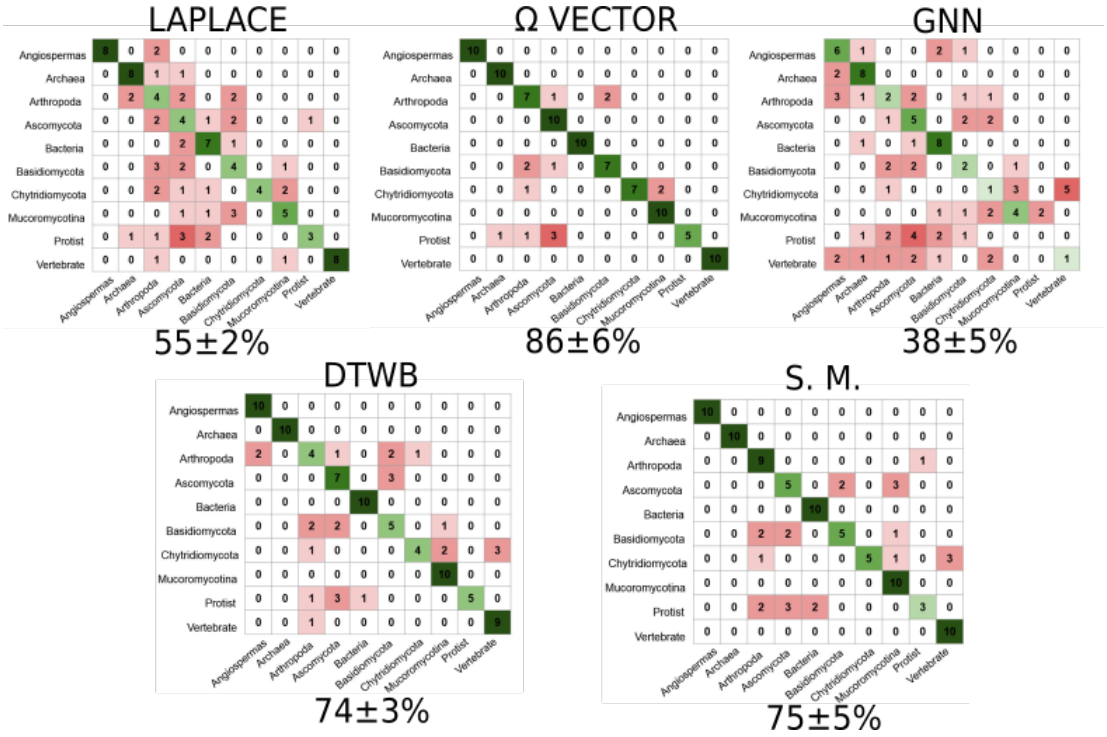

Figura 23 – Matrizes confusão dos melhores classificadores para cada método aplicado no banco de dados do STRINGdb. Nesta representação, podemos observar quais classes estão sendo mais difíceis de serem classificadas corretamente.

Fonte: Elaborada pelo autor.

Protistas foram as instâncias mais difíceis de serem classificadas, com uma acurácia máxima de apenas 50% das 10 instâncias nos bancos de dados dos vetores DTWB-R1 e $\Omega_{(25)}$. Protistas são seres de difícil classificação, até mesmo em métodos biológicos tradicionais, pois eles não formam um grupo cladístico,(58) isto é, 2 membros classificados como protistas podem ser mais distantes evolutivamente entre si do que em relação a um ser vivo fora do grupo em questão, como um fungo por exemplo. O que as matrizes confusão parecem indicar, portanto, é que membros mais distantes evolutivamente entre si tendem a ser classificados incorretamente com maior frequência. Mais evidências em cima desta hipótese estão presentes na próxima análise mostrada na Figura 24 e 25.

Nesta análise, optamos por fazer uma redução nas dimensões dos bancos de dados dos métodos que obtiveram as 3 melhores acurácias: $\Omega_{(25)}$, medidas estruturais e DTWB-R1. A redução das dimesões conseguiu evidenciar aspectos interessantes da disposição dos grafos no espaço em cada um dos bancos de dados. O modelo que obteve a pior separação foi o DTWB-R1, mesmo sua acurácia estando praticamente equivalente ao modelo de medidas estruturais. Isto demonstra que as características extraídas por este método são altamente linearmente independentes, ou seja, não é possível realizar uma redução dimensional do banco de dados e manter sua capacidade discriminatória pois todas as componentes possuem um papel significativo na separação das classes.

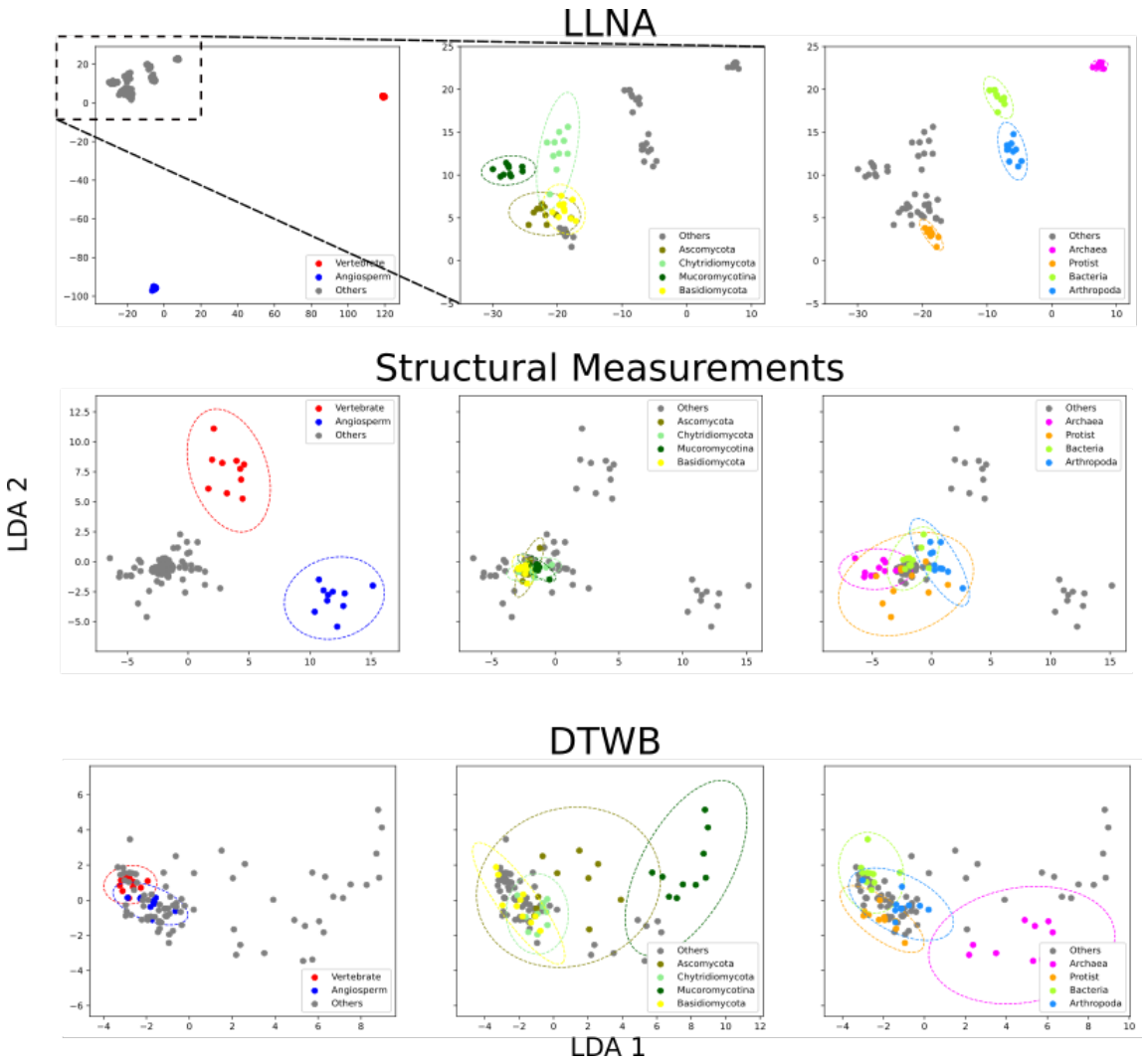

Figura 24 – Redução para duas componentes dos 3 melhores métodos de extração de características dos grafos do banco de dados STRINGdb. Nota-se que o LLNA possui uma performance bem superior na separação das classes. Muitas das classes estão a uma distância de mais de 3 desvios padrões das outras. Isso também demonstra a alta linearidade das características extraídas pelo LLNA.

Fonte: Elaborada pelo autor.

O total oposto desse caso é observado no vetor $\Omega_{(25)}$. Embora ele contenha 150 variáveis no total, sendo de longe o banco de dados com a maior quantidade delas, a redução de dimensionalidade do LDA funciona de forma extremamente eficiente, deixando 6 das 10 classe sem qualquer sobreposição em apenas 2 dimensões. A implicação disto é que, embora o método de extração de características do LLNA construa muitas variáveis, a maioria delas possui uma certa dependência e podem ser combinadas sem prejudicar a capacidade discriminativa do modelo.

É interessante notar também que, entre as classes que acabaram sobrepostas (Ascomicetos, Basidiomicetos, Quitrídiomicetos e Protistas) 3 delas são compostas por filos de fungos. A sobreposição mais significativa se deu entre Basidiomicetos e Ascomicetos,

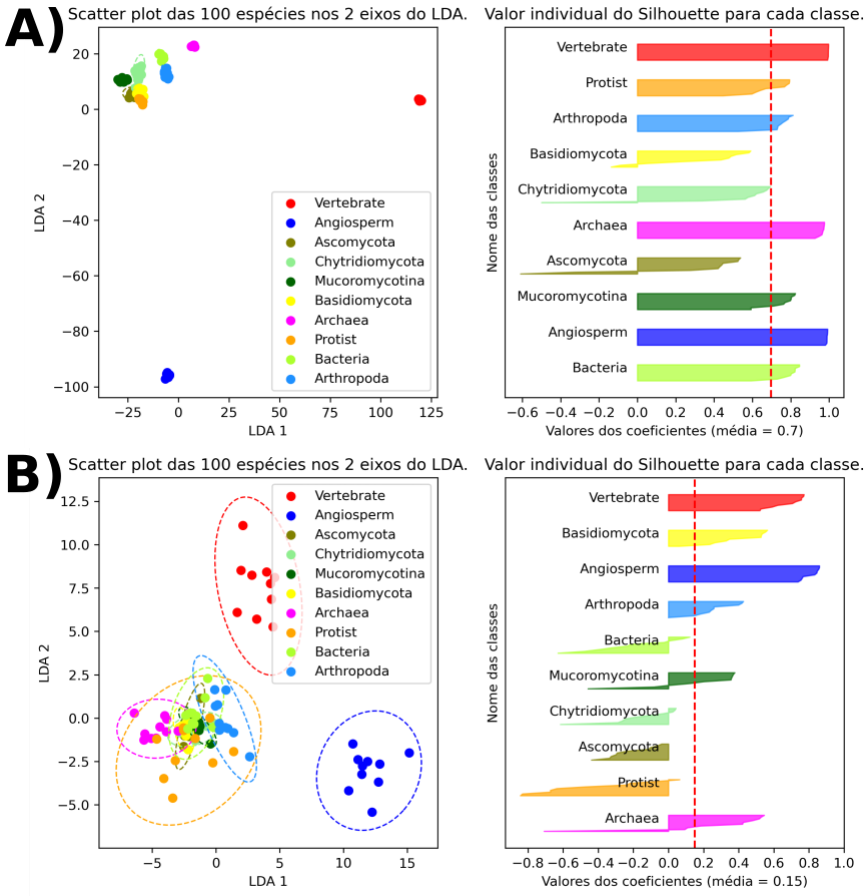

Figura 25 – Análise mais quantitativa da performance de segregação das classes do banco de dados STRINGdb. É possível observar que o coeficiente silhouette obtém uma média de clusterização de 0.7 para as características extraídas utilizando o LLNA (A). Um valor significativamente maior do que a média de clusterízação da segundo melhor banco de dados (0.15), que foram as medidas estruturais dos grafos.

Fonte: Elaborada pelo autor.

dois filos pertencentes ao subreino dos Dikarya, ou "fungos superiores" como são conhecidos coloquialmente.(59) Isto demonstra que há uma similaridade topológica entre redes de seres vivos mais próximos evolutivamente. Experimentos futuros podem buscar extrair padrões evolutivos mais fundamentais conforme o método for sendo otimizado.

6.0.3 Resultados das redes de co-expressão

Para o banco de dados do *A. fumigatus* a mesma busca exaustiva de regras foi realizada, selecionando as 10 melhores regras testadas. O resultado das acurácias pode ser visto na Figura 26.

Verifica-se uma acurácia de mais de 80% de acerto na classificação entre redes que possuem vias metabólicas em comparação com redes que não possuem redes metabólicas enriquecidas para o método do LLNA-DTEP. Ao compararmos ele com os outros métodos

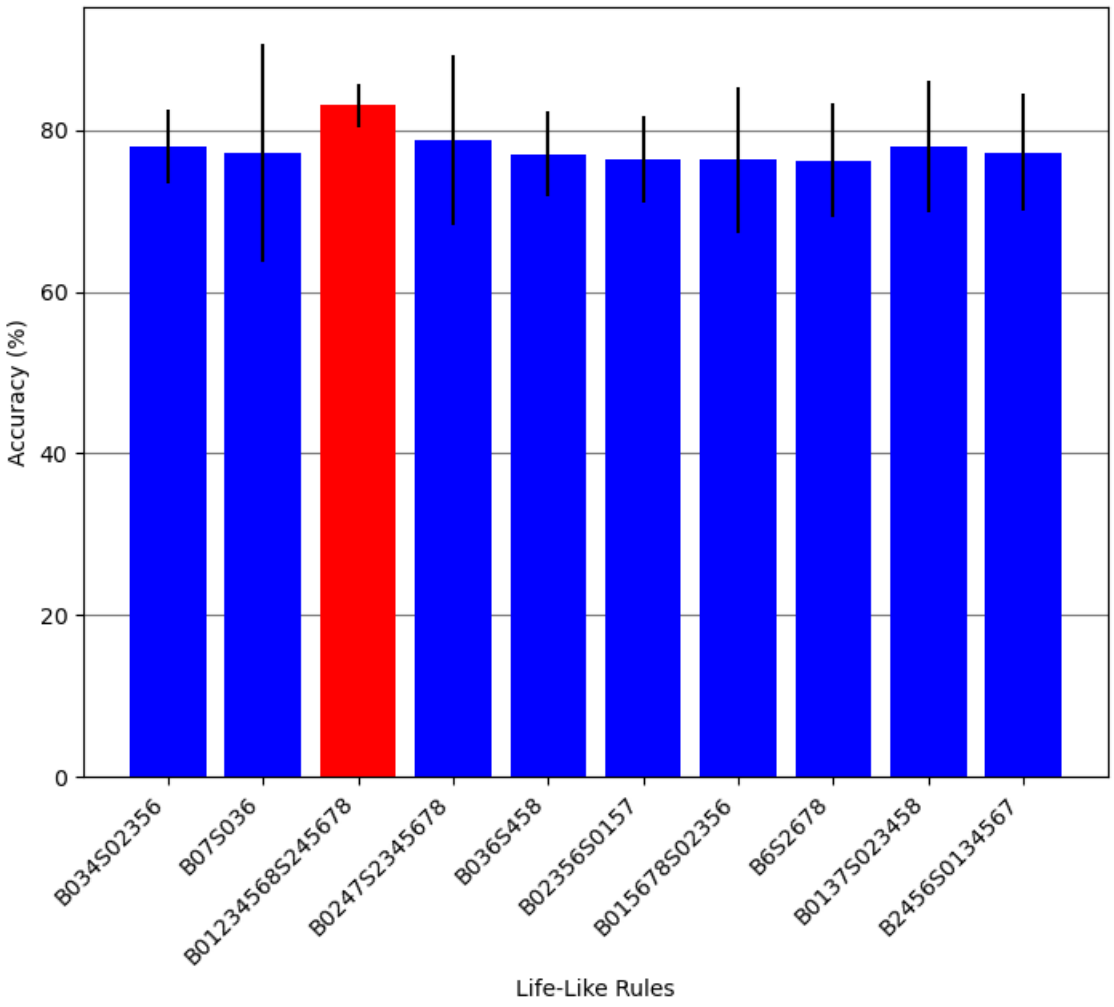

Figura 26 – Acurácias das 10 melhores regras life-like aplicadas ao banco de dados do *Aspergillus fumigatus* selvagem para todos os 4 intervalos de horário. A regra B01234568S245678 foi a que obteve a melhor acurácia com 83 ± 3%.

Fonte: Elaborada pelo autor.

(Figura 27) é possível observar que quase nenhum consegue obter uma acurácia significativamente acima de 50%, o que para um banco de dados binário como este é algo ruim. A única exceção é o método Laplaciano, que obteve uma acurácia de 73 ± 5% e a maior especifidade de todas.

O método LLNA-DTEP obteve um aumento tanto da especifidade quanto da sensibilidade em sua classificação em relação as medidas estruturais, o terceiro melhor método (de 59% para 86% para especificidade e de 74% para 79% para a sensibilidade). Isto é um bom indicativo que o LLNA-DTEP tem uma capacidade de reconhecimento de grafos não enriquecidos por vias metabólicas maior do que as medidas estruturais tradicionais, tornando ele um bom método para o reconhecimento de grafos com vias enriquecidas com base na topologia, mas é interessante notar que a maior especificidade obtida foi para o método Laplaciano, com apenas um falso negativo. Isto indica que a

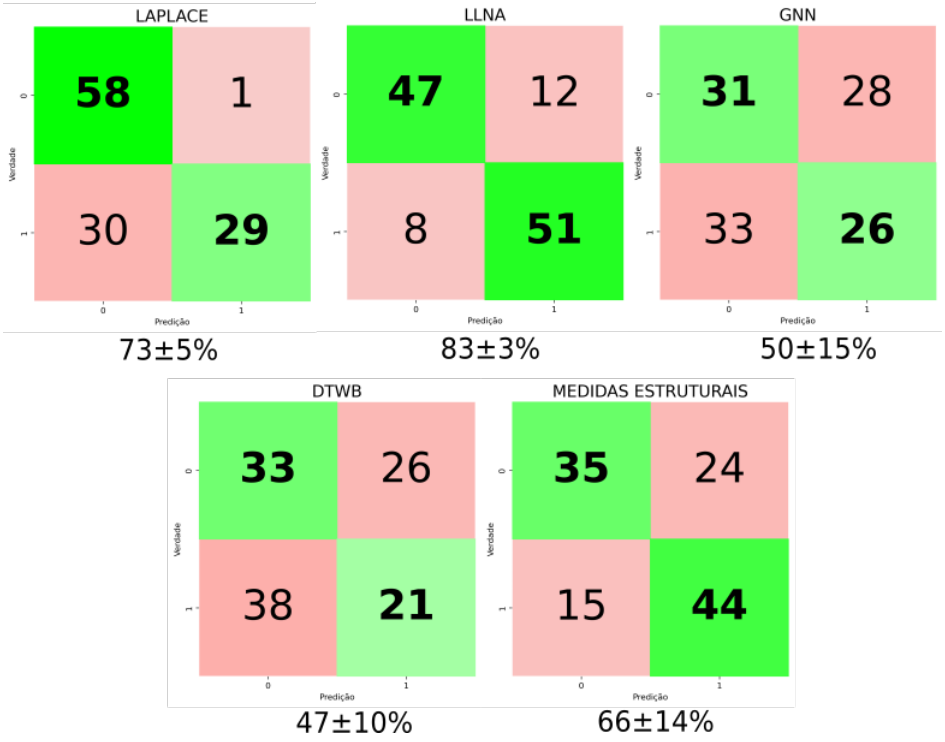

Figura 27 – Matrizes confusão para a classificação do banco de dados do *A. fumigatus*. Abaixo das matrizes encontra-se a acurácia de cada método de extração de característica. A extração de característica do LLNA continua sendo a melhor, porém desta vez é possível observar que a método Laplaciano conseguiu obter uma sensibilidade excepcional (98%).

Fonte: Elaborada pelo autor.

depender da problemática que se deseja resolver com a inteligência artifial (como por exemplo, uma segurança maior de que os grafos extraídos estão de fato enriquecidos com vias metabólicas, mesmo que isso aumente o quantidade de falsos negativos) o método Laplaciano também pode ser de grande utilidade.

7 CONCLUSÃO

Neste mestrado buscou-se utilizar novas técnicas de caracterização de grafos para aplicações biológicas diversas, tanto com a presença de algoritmos de aprendizagem de máquina como sem. Para os métodos aplicados sem o uso de aprendizagem de máquina, o novo método de quantificação de similaridade entre vetores, índice de coincidência, foi utilizado sobre grafos de interação enzimática. A nova medida demonstrou ser capaz de extrair vias de enzimas co-expressas da arqueia Halobacterium salinarum muito mais robustas e ricas em informação biológica do que os métodos de comparação de vetores tradicionalmente utilizados. O índice de coincidência também foi capaz de obter uma maior confiabilidade estatística para os subgrafos enriquecidos extraídos da rede principal.

Através das vias co-expressas extraídas a partir dos experimentos de microarray, pode-se concluir que uma via de fotofosforilação oxidativa sem a presença de oxigênio estava sendo expressa. Além disso, conseguimos montar uma hipótese que moléculas de acetil e succinato derivadas da degradação de proteínas estavam sendo utilizadas como fornecedoras de íons H+ para a construção do potencial eletroquímico gerador de ATP. Outras vias relevantes extraídas foram a da biossíntese de vitamina B12 e folato e também, para a medida de correlação Spearman, extraiu-se a via de detecção de quórum, uma via metabólica pouco analisada em arqueias.

Para o método de extração de subvias utilizado, conhecido como anelamento simulado, o índice de coincidência performou melhor quando houve uma remoção de substratos enzimáticos comuns que superconectam o grafo gerando muitos nós do tipo "hub". Outro aspecto importante é que muitas das subvias extraídas possuem genes cujas funções biológicas ainda são desconhecidas, o que tornam elas uma boa porta de entrada para futuras pesquisas sobre o metabolismo do *Halobacterium salinarum*.

Os resultados obtidos com esta abordagem foram publicados em um pré-print que está disponível no BioArxiv e pode ser acessado pelo DOI: 10.1101/2023.06.26.546540.

Para os métodos com a utilização de aprendizagem de máquina 2 bancos de dados foram explorados com a utilização do novo algoritmo de extração de características conhecido por LLNA-DTEP e LLNA-SDTEP: o banco de dados STRING, com 100 instâncias e 10 classes, e o banco de dados de co-expressão proteica do A. fumigatus com 118 instâncias e 2 classes.

Para o banco de dados STRING, fomos capazes de verificar que o novo método de extração de característica demonstra-se superior a outros métodos encontrados na literatura, além de ser capaz de distinguir entre diferentes clados de espécies com base nas características topológicas das redes. Como o banco de dados possui espécies em diferentes

níveis de proximidade evolutiva, pudemos verificar que o método consegue destinguir os grafos até o nível de filo, abrindo espaço para um novo método de distinção entre espécies com base nas relações proteícas conhecidas da espécie. O método também demonstrou-se mais capaz de gerar vetores característicos linearmente separáveis, facilitando processos de redução de dimensionalidade, como foi demonstrado pelo método LDA.

Para o banco de dados do A. fumigatus, o método teve uma capacidade de distinguir entre "clusters" enriquecidos por vias metabólicas do banco de dados KEGG e aqueles que não possuiam tal enriquecimento, com uma melhora especialmente relevante na sua especificidade. O método, então, pode ter um resultado promissor na distinção entre genes correlacionados devido a relações espúrias e genes correlacionados com uma informação biológica relevente presente que merece ser estudada.

7.1 Próximos Desenvolvimentos

Os resultados da aplicação do índice de coincidência sobre grafos de interação enzima-enzima demonstram diversas possíveis hipóteses biológicas que podem ser exploradas mais a fundo com experimentos de bancada. É possível notar que genes ainda não anotados experimentalmente foram co-expressos em meio a vias não resolvidas completamente, o que demonstra potencial na investigação de complementar o entendimento de reações metabólicas pouco estudadas.

Além disso, o Coindex demonstra ouma capacidade exploratória de dados biológicos representados em forma de grafo que supera o desempenho das medidas mais comuns r de Pearson e r de Spearman. Desenvolvimentos envolvendo outras espécies a serem investigadas, em conjunto com outras medidas de correlação como a correlação de Hellinger, *biweight midcorrelation* e *mutual information coefficient*, que são medidas de correlação mais recentes, podem ajudar a elucidar melhor o potencial do Coindex.

Também seria interessante realizar a aplicação do índice de coincidência sobre experimentos de sequenciamento de RNA (RNA-Seq) que são o estado-da-arte em termos de experimento de transcriptômica e são menos suscetíveis a variações estocásticas do que experimentos de microarray.

Já para os métodos de extração de características, conforme o algoritmo LLNA for sendo aprimorado, será possível utilizar-se de sementes determinísticas que irão dispensar a busca extensiva por um conjunto de regras/sementes para classificação de banco de dados. Isto permitirá o uso do algoritmo para montagem de vetores característicos que não requerem supervisão por meio de uma classe pré-estabelecida e será possível utilizar os métodos aqui demonstrados para classificação de espécies a nível de biologia de sistemas e também para o descobrimento de redes enriquecidas com vias metabólicas ainda não exploradas.

A investigação de redes de interação proteína-proteína permite avaliar informações evolutivas de seres vivos que não estão isoladas somente às sequências das proteínas e dos ácidos nucleicos e também não às propriedades fenotípicas do ser vivo. Ao invés disso, foca-se em estudar a evolução dos sistemas moleculares que compõe os seres vivos. O uso de aprendizagem de máquina e métodos de extração de características mais recentes demonstrou ser capaz de resolver padrões nas grandes redes que não são facilmente visualizados com métodos tradicionais.

Já para as redes de co-expressão proteica, o uso da aprendizagem de máquina permite obter subgrafos que têm maior probabilidade de conter vias metabólicas ou interações proteicas de interesse para o objetivo específico do projeto. Isto é feito treinando o modelo sobre conjuntos contendo classes onde uma delas são "clusters" enriquecidos com vias metabólicas, e outra porção sem nenhum enriquecimento de interesse. A utilizadade de tal modelo classificatório está em sua capacidade de selecionar porções mais manejáveis de uma rede de interação proteína-proteína global que sejam promissoras na elucidação de novas interações, uma abordagem "top-down" dentro da biologia de sistemas.

Com estas redes obtidas, pode-se realizar o processo inverso da biologia de sistemas ("bottom-up") que envolve tentativas de modelagem matemática das possíveis interações proteína-proteína ou experimentos bioquímcos que buscam extrair qual é a atividade exercida pelas proteínas sobre determinados subtratos.

A crescente quantidade de dados biológicos exige análises mais robustas para obter conclusões genuinamente impactantes. Este mestrado buscou analisar novas ferramentas que podem ser utilizadas neste âmbito para chegar a tais conclusões com mais facilidade e segurança estatística.

REFERÊNCIAS

1 ZIELINSKI, K. M. C. *et al.* **A network classification method based on density time evolution patterns extracted from network automata**. 2022. Disponível em: https://arxiv.org/pdf/2211.13000.pdf Acesso em: 23 Jan 2023.

2 DAHM, R. Discovering dna: Friedrich miescher and the early years of nucleic acid research. **Human Genetics**, Springer Science and Business Media LLC, v. 122, n. 6, p. 565–581, Sept. 2007. ISSN 1432-1203. DOI: 10.1007/s00439-007-0433-0.

3 CHARGAFF, E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. **Experientia**, Springer Science and Business Media LLC, v. 6, n. 6, p. 201–209, June 1950. ISSN 1420-9071. DOI: 10.1007/bf02173653.

4 LEVENE, P. The structure of yeast nucleic acid. **Journal of Biological Chemistry**, Elsevier BV, v. 40, n. 2, p. 415–424, Dec. 1919. ISSN 0021-9258. DOI: 10.1016/s0021-9258(18)87254-4.

5 CRICK, F. Central dogma of molecular biology. **Nature**, Springer Science and Business Media LLC, v. 227, n. 5258, p. 561–563, Aug. 1970. ISSN 1476-4687. DOI: 10.1038/227561a0.

6 GARIBYAN, L.; AVASHIA, N. Polymerase chain reaction. **Journal of Investigative Dermatology**, Elsevier BV, v. 133, n. 3, p. 1–4, Mar. 2013. ISSN 0022-202X. DOI: 10.1038/jid.2013.1.

7 NATIONAL RESEARCH COUNCIL. **Network science**. Washington, DC: The National Academies Press, 2005. ISBN 978-0-309-10026-7.

8 COSTA, L. da F. *et al.* Characterization of complex networks: a survey of measurements. **Advances in Physics**, v. 56, n. 1, p. 167–242, Jan. 2007. DOI: 10.1080/00018730601170527.

9 BOLLOBÁS, B. *et al.* Random induced graphs. **Discrete Mathematics**, v. 248, n. 1-3, p. 249–254, Apr. 2002.

10 WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. **Nature**, v. 393, n. 6684, p. 440–442, June 1998. DOI: 10.1038/30918.

11 BARABASI, A.-L.; OLTVAI, Z. Network biology: understanding the cell's functional organization. **Nature reviews. Genetics**, v. 5, n. 2, p. 101–113, Mar. 2004.

12 RAVASZ, E. *et al.* Hierarchical organization of modularity in metabolic networks. **Science**, v. 297, n. 5586, p. 1551–1555, Aug. 2002. DOI: 10.1126/science.1073374.

13 JEONG, H. *et al.* The large-scale organization of metabolic networks. **Nature**, v. 407, n. 6804, p. 651–654, Oct. 2000. DOI : 10.1038/35036627.

14 LIM, J. *et al.* A protein–protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. **Cell**, v. 125, n. 4, p. 801–814, May 2006. DOI: 10.1016/j.cell.2006.03.032.

15 YANG, Z. *et al.* Alphafold2 and its applications in the fields of biology and medicine. **Signal Transduction and Targeted Therapy**, Springer Science and Business Media LLC, v. 8, n. 1, Mar. 2023. ISSN 2059-3635. DOI: 10.1038/s41392-023-01381-z.

16 SUBRAMANIAN, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. **Proceedings of the National Academy of Sciences**, v. 102, n. 43, p. 15545–15550, Sept. 2005. DOI: 10.1073/pnas.0506580102.

17 MOOTHA, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. **Nature Genetics**, v. 34, n. 3, p. 267–273, June 2003. DOI: 10.1038/ng1180.

18 COSTA, L. da F. Coincidence complex networks. **Journal of Physics: Complexity**, v. 3, n. 1, p. 015012, Mar. 2022. DOI: 10.1088/2632-072x/ac54c3.

19 SZKLARCZYK, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. **Nucleic Acids Research**, Oxford University Press (OUP), v. 49, n. D1, p. D605–D612, Nov. 2020. Disponível em: https://doi.org/10.1093/nar/gkaa1074.

20 PROOST, S.; MUTWIL, M. **PlaNet: comparative co-expression network analyses for plants**. New York: Springer, 2016. 213–227 p. DOI: 10.1007/978-1-4939-6658-5_12.

21 GERLEE, P.; LIZANA, L.; SNEPPEN, K. Pathway identification by network pruning in the metabolic network of Escherichia coli. **Bioinformatics**, v. 25, n. 24, p. 3282–3288, Oct. 2009. DOI: 10.1093/bioinformatics/btp575.

22 GONZALEZ, O. *et al.* Reconstruction, modeling & analysis of halobacterium salinarum r-1 metabolism. **Molecular BioSystems**, v. 4, p. 148–159, 2008. DOI: 10.1039/B715203E.

23 IDEKER, T. *et al.* Discovering regulatory and signalling circuits in molecular interaction networks. **Bioinformatics**, v. 18, n. suppl. 1, p. S233–S240, Jul. 2002. DOI: 10.1093/bioinformatics/18.suppl_1.s233.

24 SCHOBER, P.; BOER, C.; SCHWARTE, L. A. Correlation coefficients: Appropriate use and interpretation. **Anesthesia & Analgesia**, Ovid Technologies (Wolters Kluwer Health), v. 126, n. 5, p. 1763–1768, May 2018. Disponível em: https://doi.org/10.1213/ane.0000000000002864.

25 BONNEAU, R. *et al.* A predictive model for transcriptional control of physiology in a free living cell. **Cell**, v. 131, n. 7, p. 1354–1365, Dec. 2007. DOI: 10.1016/j.cell.2007.10.053.

26 PATIL, K. R.; NIELSEN, J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. **Proceedings of the National Academy of Sciences**, v. 102, n. 8, p. 2685–2689, Feb. 2005.

27 ASHBURNER, M. *et al.* Gene ontology: tool for the unification of biology. **Nature Genetics**, v. 25, n. 1, p. 25–29, May. 2000. DOI: 10.1038/75556.

28 Gene Ontology Consortium. The gene ontology knowledgebase in 2023. **Genetics**, v. 224, n. 1, p. iyad031, Mar. 2023. DOI: 10.1093/genetics/iyad031.

29 KANEHISA, M. KEGG: Kyoto encyclopedia of genes and genomes. **Nucleic Acids Research**, v. 28, n. 1, p. 27–30, Jan. 2000. DOI: 10.1093/nar/28.1.27.

30 WARREN, M. J. *et al.* The biosynthesis of adenosylcobalamin (vitamin b12). **Natural Product Reports**, v. 19, n. 4, p. 390–412, June 2002. DOI: 10.1039/b108967f.

31 ALLEN, L. H. Vitamin b-12. **Advances in Nutrition**, v. 3, n. 1, p. 54–55, Jan. 2012. DOI: 10.3945/an.111.001370.

32 MCKINLAY, J. B.; COOK, G. M.; HARDS, K. Microbial energy management—a product of three broad tradeoffs. *In*: POOLE, R. K. (ed.). **Advances in microbial physiology**. New York: Academic Press, 2020, (Advances in Microbial Physiology, v. 77). p. 139–185.

33 TALAUE, C. O. *et al.* Model construction and analysis of respiration in halobacterium salinarum. **PLOS ONE**, v. 11, n. 3, p. e0151839, Mar. 2016. DOI: 10.1371/journal.pone.0151839.

34 GUZZO, M. B. *et al.* Methylfolate trap promotes bacterial thymineless death by sulfa drugs. **PLOS Pathogens**, v. 12, n. 10, p. e1005949, Oct. 2016. DOI: 10.1371/journal.ppat.1005949.

35 CHARLESWORTH, J. C. *et al.* Quorum sensing in archaea: recent advances and emerging directions. *In*: WITZANY, G. (ed.). **Biocommunication of Archaea**. Cham: Springer, 2017. p. 119–132. DOI: 10.1007/978-3-319-65536-9_8.

36 XU, Q. *et al.* Insights into substrate specificity of geranylgeranyl reductases revealed by the structure of digeranylgeranylglycerophospholipid reductase, an essential enzyme in the biosynthesis of archaeal membrane lipids. **Journal of Molecular Biology**, v. 404, n. 3, p. 403–417, 2010.

37 KOK, N. A. W. de; DRIESSEN, A. J. M. The catalytic and structural basis of archaeal glycerophospholipid biosynthesis. **Extremophiles**, v. 26, n. 3, Aug. 2022. DOI: 10.1007/s00792-022-01277-w.

38 GRIVARD, A. *et al.* Archaea carotenoids: natural pigments with unexplored innovative potential. **Marine Drugs**, v. 20, n. 8, p. 524, Aug. 2022. DOI: 10.3390/md20080524.

39 GOO, Y. A. *et al.* Proteomic analysis of an extreme halophilic archaeon, halobacterium sp. nrc-1*. **Molecular & Cellular Proteomics**, v. 2, n. 8, p. 506–524, 2003.

40 KůRKOVá, V. *et al.* **Artificial neural networks and machine learning** - ICANN 2018. Cham: Springer, 2018. (Lecture notes in computer science, v. 11141). ISBN 978-3-030-01423-0.

41 MIRANDA, G. H. B.; MACHICAO, J.; BRUNO, O. M. Exploring spatio-temporal dynamics of cellular automata for pattern recognition in networks. **Scientific Reports**, v. 6, n. 1, Nov. 2016. DOI: 10.1038/srep37329.

42 SCHöLKOPF, B.; PLATT, J.; HOFMANN, T. **Advances in neural information processing systems 19:** proceedings of the 2006 conference. Cambridge: The MIT Press, 2007. DOI: 10.7551/mitpress/7503.001.0001.

43 RIBAS, L. C.; MACHICAO, J.; BRUNO, O. M. Life-like network automata descriptor based on binary patterns for network classification. **Information Sciences**, v. 515, p. 156–168, 2020. DOI: 10.1016/j.ins.2019.09.063.

44 MERENDA, J. V. B. de S. **Reconhecimento de padrões em redes complexas usando caminhadas determinísticas do turista**. 2023: Dissertação (Mestrado em Ciências) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2023. DOI: 10.11606/d.76.2023.tde-12062023-084122.

45 HAN, S. *et al.* **Incremental boosting convolutional neural network for facial action unit recognition**. 2017. Disponível em: https://proceedings.neurips.cc/paper_-files/paper/2016/file/d09bf41544a3365a46c9077ebb5e35c3-Paper.pdf Acesso em: 23 Jan 2023.

46 NOBLE, W. S. What is a support vector machine? **Nature Biotechnology**, v. 24, n. 12, p. 1565–1567, Dec. 2006.

47 PEDREGOSA, F. *et al.* Scikit-learn: machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

48 HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**. 2nd. ed. Berlin: Springer, 2017. (Springer series in statistics).

49 BARABASI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, v. 286, n. 5439, p. 509–512, Oct. 1999. DOI: 10.1126/science.286.5439.509.

50 WAXMAN, B. Routing of multipoint connections. **IEEE Journal on Selected Areas in Communications**, v. 6, n. 9, p. 1617–1622, 1988. DOI: 10.1109/49.12889.

51 LESKOVEC, J.; KREVL, A. **SNAP Datasets:** Stanford large network dataset collection. 2014. http://snap.stanford.edu/data.

52 ZHAO, J. *et al.* Complex networks theory for analyzing metabolic networks. **Chinese Science Bulletin**, Springer Science and Business Media LLC, v. 51, n. 13, p. 1529–1537, July 2006. DOI: 10.1007/s11434-006-2015-2.

53 MACHICAO, J. *et al.* Topological assessment of metabolic networks reveals evolutionary information. **Scientific Reports**, v. 8, n. 1, Oct. 2018. DOI: 10.1038/s41598-018-34163-7.

54 ALTWASSER, R. *et al.* Network modeling reveals cross talk of MAP kinases during adaptation to caspofungin stress in aspergillus fumigatus. **PLOS ONE**, v. 10, n. 9, p. e0136932, Sept. 2015. DOI: 10.1371/journal.pone.0136932.

55 LIESECKE, F. *et al.* Ranking genome-wide correlation measurements improves microarray and rna-seq based global and targeted co-expression networks. **Scientific Reports**, v. 8, n. 1, p. 10885, Sept. 2018.

56 MUTWIL, M. *et al.* Assembly of an interactive correlation network for the arabidopsis genome using a novel heuristic clustering algorithm . **Plant Physiology**, v. 152, n. 1, p. 29–43, Nov. 2009. DOI: 10.1104/pp.109.145318.

57 STERCK, L. *et al.* How many genes are there in plants (... and why are they there)? **Current Opinion in Plant Biology**, v. 10, n. 2, p. 199–203, Apr. 2007. DOI: 10.1016/j.pbi.2007.01.004.

58 SOGIN, M. L.; SILBERMAN, J. D. Evolution of the protists and protistan parasites from the perspective of molecular systematics. **International Journal for Parasitology**, v. 28, n. 1, p. 11–20, Jan. 1998. DOI: 10.1016/s0020-7519(97)00181-1.

59 HIBBETT, D. S. *et al.* Phylogenetic taxon definitions for fungi, dikarya, ascomycota and basidiomycota. **IMA Fungus**, v. 9, n. 2, p. 291–298, Dec. 2018. DOI: 10.5598/imafungus.2018.09.02.05.